



EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

Research and Innovation Action

Call: H2020-ICT-2018-1

Call topic: ICT-29-2018 A multilingual Next generation Internet

Project start: 1 January 2019

Project duration: 36 months

D2.5: Final cross-lingual semantic enrichment technology (T2.1)

Executive summary

The report describes the final version of the systems developed for semantic enrichment in Task T2.1. The systems address the tasks of named entity recognition, named entity linking, and event detection. The experiments show an improvement with respect to the outcomes obtained in Deliverable D2.2. In many cases, the results surpass the performance of state-of-the-art tools.

Partner in charge: ULR

Project co-funded by the European Commission within Horizon 2020
Dissemination Level

PU	Public	PU
PP	Restricted to other programme participants (including the Commission Services)	—
RE	Restricted to a group specified by the Consortium (including the Commission Services)	—
CO	Confidential, only for members of the Consortium (including the Commission Services)	—



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825153

Deliverable Information

Document administrative information	
Project acronym:	EMBEDDIA
Project number:	825153
Deliverable number:	D2.5
Deliverable full title:	Final cross-lingual semantic enrichment technology
Deliverable short title:	Final cross-lingual semantic enrichment technology
Document identifier:	EMBEDDIA-D25-FinalCrosslingualSemanticEnrichmentTechnology-T21-submitted
Lead partner short name:	ULR
Report version:	submitted
Report submission date:	31/12/2020
Dissemination level:	PU
Nature:	R = Report
Lead author(s):	Adrián Cabrera (ULR)
Co-author(s):	Emanuela Boros (ULR), Elvys Linhares-Pontes (ULR), Jose G Moreno (ULR), Antoine Doucet (ULR), Marko Robnik-Šikonja (UL)
Status:	__ draft, __ final, <u>x</u> submitted

The EMBEDDIA Consortium partner responsible for this deliverable has addressed all comments received. Changes to this document are detailed in the change log table below.

Change log

Date	Version number	Author/Editor	Summary of changes made
13/10/2020	v0.1	Adrián Cabrera (ULR)	Structure proposal
5/11/2020	v0.2	Lead and Co-authors (ULR)	First complete version without global conclusion
08/11/2020	v0.3	Jose G Moreno (ULR)	Corrections and comments
10/11/2020	v0.4	Adrián Cabrera (ULR)	Proposal for global conclusion
20/11/2020	v0.5	Emanuela Boros (ULR)	Minor changes in structure
22/11/2020	v0.6	Adrián Cabrera (ULR)	Update Abbreviations list and Executive summary
23/11/2020	v0.7	Adrián Cabrera (ULR)	Addition of Appendixes
27/11/2020	v0.8	Marko Robnik-Šikonja (UL)	Internal review
09/12/2020	v0.9	Andraž Pelicon (JSI)	Internal review
14/12/2020	v1.0	Emanuela Boros (ULR), Elvys Pontes Linhares (ULR), Adrián Cabrera (ULR)	Changes/corrections suggested by internal reviewers
16/12/2020	v1.1	Nada Lavrač (JSI)	Quality control
18/12/2020	final	Emanuela Boros (ULR), Elvys Pontes Linhares (ULR), Adrián Cabrera (ULR)	Minor changes from Quality Control
29/12/2020	submitted	Tina Anžič (JSI)	Report submitted

Table of Contents

1. Introduction.....	6
2. Background	7
2.1 Named Entity Recognition	7
2.2 Named Entity Linking	7
2.3 Event Detection	8
3. Named Entity Recognition (NER)	9
3.1 Previous work.....	9
3.2 Explored Approaches	10
3.2.1 BiLSTM with FastText and Multilingual BERT	10
3.2.2 BiLSTM with FastText and Pseudo-affixes Embeddings	12
3.2.3 BERT with Stacked Transformer Blocks	13
3.2.4 Multitask BERT	14
3.3 Datasets	16
3.4 Experimental Setup	17
3.4.1 Evaluation Metrics.....	17
3.4.2 BiLSTM with FastText and Multilingual BERT Experiments	18
3.4.3 BiLSTM with FastText and Pseudo-affixes Experiments	19
3.4.4 BERT with Stacked Transformer Blocks Experiments.....	19
3.4.5 Multitask BERT Experiments	23
3.4.6 Comparative Results for all the Explored Systems.....	27
3.4.7 Comparative Results with the State of the Art.....	31
3.5 Discussion	33
3.6 Conclusions.....	36
4. Named Entity Linking (NEL)	37
4.1 Previous Work	37
4.2 Cross-lingual Named Entity Linking.....	38
4.3 Multilingual End-to-end Entity Linking.....	38
4.3.1 Building Resources	39
4.3.2 Entity Embeddings	39
4.3.3 Entity Disambiguation	40
4.3.4 Match Corrections	41
4.3.5 Multilingualism	42
4.4 Datasets	42
4.5 Experimental Setup	42
4.5.1 Evaluation Metrics.....	43
4.5.2 Cross-lingual and Multilingual NEL Experiments	43
4.6 Discussion	44
4.7 Conclusions.....	44
5. Event Detection (ED)	45
5.1 Definitions	45

5.2	Challenges	47
5.3	Previous Work	47
5.4	Datasets	49
5.4.1	BSNLP 2019 Dataset	49
5.4.2	ACE 2005 Dataset	50
5.4.3	DAnIEL Dataset	51
5.5	Explored Approaches	53
5.5.1	DAnIEL System	53
5.5.2	Convolutional Neural Network-based Approaches	54
5.5.3	Fine-tuned Language Model-based Approaches	56
5.6	Experimental Setup	56
5.6.1	Evaluation Metrics	57
5.6.2	BSNLP 2019 Experiments	57
5.6.3	DAnIEL Experiments	58
5.6.4	ACE 2005 Experiments	60
5.7	Discussion	62
5.8	Conclusions	63
6.	Conclusions and Future Work	64
7.	Associated Outputs	65
Appendices		76
1.	A Dataset for Multilingual Epidemiological Event Extraction	77
2.	Alleviating Digitization Errors in Named Entity Recognition for Historical Documents	84
3.	CTLR@WiC-TSV: Target Sense Verification using Marked Inputs and Pre-trained Models	96
4.	Dataset for Temporal Analysis of English-French Cognates	103
5.	Entity Linking for Historical Documents: Challenges and Solutions	109
6.	Event Detection with Entity Markers	127
7.	Event Extraction over Digitised and Historical Documents	136
8.	Impact Analysis of Document Digitization on Event Extraction	154
9.	Improving NER systems by marking uppercase tokens, and predicting masked tokens and entities boundaries	167
10.	Linking Named Entities across Languages using Multilingual Word Embeddings	176
11.	Multilingual Epidemic Event Extraction	181
12.	Multilingual Epidemiological Text Classification: A Comparative Study	187
13.	Relation Classification via Relation Validation	200
14.	Robust Named Entity Recognition and Linking on Historical Multilingual Documents	209
15.	TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data	227

List of abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
Bound.	Boundaries
BPEmb	Byte-Pair Embeddings
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
DAnIEL	Data Analysis for Information Extraction in any Language
ED	Event Detection
EE	Event Extraction
EL	Entity Linking
EVT	Event (named entity type)
GH	Ganea and Hofmann
IE	Information Extraction
IOB	Inside-Outside-Beginning
IOBES	Inside-Outside-Beginning-End-Single
KB	Knowledge Base
LOC	Location (named entity type)
MISC	Miscellaneous (named entity type)
NE	Named entity
NEL	Named Entity Linking
NER	Named Entity Recognition
ORG	Organisation (named entity type)
PER	Person (named entity type)
POS	Part-of-Speech
PRO	Product (named entity type)
RNN	Recursive Neural Network
RoBERTa	Robustly optimized BERT approach
Upper.	Uppercase
WP	Work Package
XEL	Cross-lingual Entity Linking
XLNet	Cross-lingual RoBERTa

1 Introduction

The overall objective of WP2 is the embeddings-based semantic enrichment of individual documents and their content. This enrichment is achieved by performing multi and cross-lingual named-entity recognition and disambiguation, and linking the recognised named entities to external knowledge bases such as Wikipedia. Further, based on these cross-lingual semantic descriptors, we will advance event detection techniques to markup potentially breaking events.

Task T2.1 is concerned with the cross-lingual semantic enrichment of text. It provides named entity recognition, linking, and event detection, interacting notably with Task T2.2 on multilingual keyword extraction and matching, and being evaluated as defined in Task T2.4.

The present document, entitled 'Final cross-lingual semantic enrichment', and the corresponding source code compose Deliverable D2.5, which is the final deliverable in Task T2.1 of WP2.

Central to Task T2.1, named entities (NEs) are real-world objects, such as persons, locations and organisations. They are considered important concepts as they often are key descriptors of a text. The first aim of Task T2.1 is named entity recognition (NER), which seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as 'person', 'location' and 'organisation'.

Another aim of T2.1 is named entity linking (NEL) which is the task of assigning an unambiguous identifier to every mention of an NE, for instance using an external knowledge base (KB) such as Wikidata.

In this deliverable, we continue to experiment with the recognition of a NE category 'event', as we did in Deliverable D2.2, used to perform event detection (ED). As well, we explore other styles of defining an event, at sentence or document level, and test new methods to detect them.

The second year of work on Task T2.1 mainly resulted in the following achievements:

- For NER, presented in Section 3, we developed new NER systems for all the languages explored in Embeddia by using and improving tools from the state of the art. The created tools show an improvement with respect to the NER architecture proposed in D2.2, in some cases, improving the performance with respect to the current state of the art. The work related to NER, described in Section 3, is further addressed in Appendices: 2, 9, and 14.
- For NEL, presented in Section 4, we improved our previous cross-lingual model by proposing a multilingual model to disambiguate entities in the EMBEDDIA languages. Most specifically, our new approach analyses multilingual documents and disambiguates their NEs to a common KB (i.e. Wikidata). The work in NEL, described in Section 4, is further addressed in Appendices: 5, 10, and 14.
- ED was processed in the previous deliverable by the proposed NER approach, where events were considered as specific types of named entities. In this deliverable, we explored more advanced approaches for event detection in several datasets. The work described in Section 5 is further addressed in Appendices: 1, 6, 8, 11, and 12.

The work presented in this deliverable uses the collected data presented in D2.1 delivered at M9 (part of Task 2.4 'Data sets and evaluation for NLP technology') to train and evaluate NER and NEL for the languages of the EMBEDDIA project.

The present report is organised as follows: Section 2 introduces a background for the different tasks explored in this deliverable. Section 3 presents our NER approach and its performance on the EMBEDDIA languages over several NE categories and datasets. Section 4 describes our work on multilingual NEL and shows its results over multiple languages. Section 5 presents the different event definitions and the approaches for the event detection task. Finally, the conclusions and future work are set in Section 6.

2 Background

In this section, we present a brief description, as well as some examples, of the three tasks covered in this deliverable: named entity recognition (NER), named entity linking (NEL), and event detection (ED). A more detailed background can be found in Deliverable D2.1.

2.1 Named Entity Recognition

Named entity recognition (NER) is the task addressing the extraction and tagging of a word, or a group of them, that semantically refer to aspects such as locations, persons, organisations, products, genes, and proteins (Luoma, Oinonen, Pyykönen, Laippala, & Pyysalo, 2020; Yu, Bohnet, & Poesio, 2020; J. Li, Sun, Han, & Li, 2020).

In Figure 1a, we present an example of named entities in English, from the corpus (Tjong Kim Sang & De Meulder, 2003), and in Figure 1b, a Croatian example (Ljubešić & Erjavec, 2016). In both instances, the last column makes reference to the named entities annotation.

Only	RB	B-NP	0	Seb	B-PER
France	NNP	I-NP	B-LOC	Bytyci	I-PER
and	CC	I-NP	0	,	0
Britain	NNP	I-NP	B-LOC	izvršni	0
backed	VBD	B-VP	0	ravnatelj	0
Fischler	NNP	B-NP	B-PER	Instituta	B-ORG
's	POS	B-NP	0	za	I-ORG
proposal	NN	I-NP	0	balkansku	I-ORG
.	.	0	0	politiku	I-ORG
				,	0
				slaže	0
				se	0
				s	0
				time	0
				.	0

(a) English

(b) Croatian

Figure 1: Examples of NER annotations in two different languages.

In the examples presented in Figure 1, we can see three different types of named entities: LOC (Location), PER (Person), and ORG (Organisation).

2.2 Named Entity Linking

After recognising the named entities, named entity linking (NEL) aims to disambiguate these entities by linking them to entries of a knowledge base (KB). NEL is a challenging task because named entities may have multiple surface forms, such as its full name, partial names, aliases, abbreviations, and alternate spellings (Shen, Wang, & Han, 2014).

In a nutshell, NEL aims to recover the ground truth entities in a KB referred to in a document by locating mentions, and for each mention accurately disambiguating the referent entity (Figure 2). The EMBED-DIA project aims to link named entities presented in less-resourced languages to a KB. Unfortunately, available corpora for NEL is scarce for these less-resourced languages. In order to overcome this problem, we developed a multilingual approach to better analyse documents in these languages and link their entities to a KB.

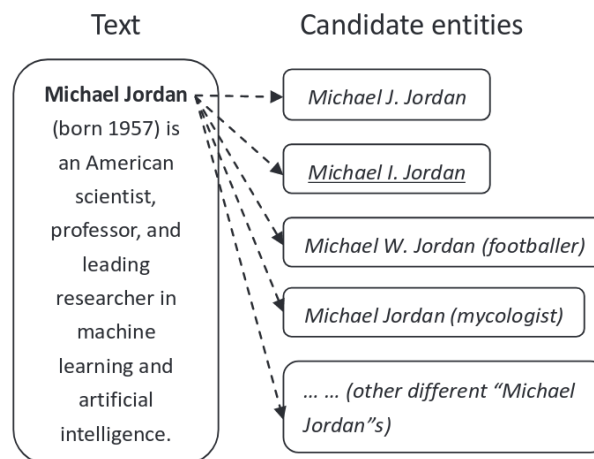


Figure 2: An illustration for the NEL task. The named entity mention detected from the text is in bold face, the correct mapping entity is underlined.

2.3 Event Detection

Event extraction (EE) is an application of information extraction (IE), and it consists on obtaining specific knowledge of certain incidents from textual documents. This task is focused on obtaining event-related information from texts, and, as commonly defined in the field of IE, it consists of two main sub-tasks. The first sub-task involves event detection (ED) that deals with the extraction of critical information regarding an event, that can be represented by a keyword, a phrase, a sentence or a span of text, which evoke that event. For example, an article can talk about a new epidemic outbreak, or about the election of a new president, where the events to be detected are represented by the name of the epidemic, or by the word 'election'. The second sub-task, mostly referred to as event argument extraction, concentrates on obtaining the event extents referring to more details about the events. These extents often refer to elements such as the events' arguments or participants. For example, the location of the epidemic event, the name of the president, the country of the election, are to be detected in this sub-task. After NER and NEL, ED may take advantage from the detected and linked named entities since they can be participants of an event.

Over the years, several event definitions have been proposed, each showing specific strengths and weaknesses. Thus, the event detection task is challenging due to the ambiguous nature of the concept of event. In this deliverable, we continue experimenting with event detection as a named entity recognition task and we analyse two other different annotation styles that are commonly used in the research.

3 Named Entity Recognition (NER)

Named entity recognition (NER) is a fundamental task in the processing of texts that consists of extracting entities that semantically refer to aspects such as locations, persons, or organisations (Luoma et al., 2020). Named entities can be used as a stand-alone output but also to improve other NLP tasks such as automatic text summarisation, question-answering, and machine translation (J. Li et al., 2020).

In this section, we present the work we conducted in Task 2.1 regarding NER systems for the eight languages of EMBEDDIA. In addition to the systems presented in Deliverable 2.2, we present in details several NER architectures separately for every EMBEDDIA language. Furthermore, we experiment with improved versions of the models from the previous deliverable and with new and most current models based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2019). The results show an improvement with respect to the models presented in Deliverable 2.2, and in multiple cases, we can achieve a performance higher than the state of the art.

3.1 Previous work

We present in this section a collection of the most representative works related to NER systems that have been created for the languages explored in EMBEDDIA. These works have been classified into four different classes according to the main approach utilised for recognising named entities: rule-based, CRF-based, BiLSTM-based, and Transformer-based. These four classes cover most of the systems found in the literature and are described below.

Rule-based systems. These kinds of NER systems are, in most cases, the earliest type of NER systems that can be found in the literature, e.g. (Dalianis & Åström, 2001). They consist of systems that search for patterns in a given text to determine the location of named entities. Rules are frequently based on regular expressions or grammars, which can cover syntactical and lexical elements, but also aspects such as detection of uppercase characters, suffix, and quantity patterns (Dalianis & Åström, 2001; Bekavac & Tadić, 2007; J. Li et al., 2020). Furthermore, these rules are, in most cases, created by hand, but there are some systems that can generate these patterns automatically (J. Li et al., 2020). Some examples of NER rule-based systems are: the first version of FiNER (Lindén et al., 2013) (Finnish), (Kapočiūtė & Raškinis, 2005) (Lithuanian), SweNam (Dalianis & Åström, 2001) (Swedish), HFST-SweNER (Kokkinakis, Niemi, Hardwick, Lindén, & Borin, 2014) (Swedish) and Croatian NERC System (Bekavac & Tadić, 2007) (Croatian). Certain NER systems make use of other resources as well, such as lexica or thesauri, in order to improve the performance, such as those presented in (Gareev, Tkachenko, Solovyev, Simanovsky, & Ivanov, 2013) (Russian) or (Kokkinakis, 2003) (Swedish). These resources might contain information regarding the inflection of words, names, and locations.

CRF-based systems. These NER systems make use of a supervised approach which is based on the generation of a sequence tagger using Conditional Random Fields (CRF) (Lafferty, McCallum, & Pereira, 2001). Apart from the original input text, CRF-based systems utilise additional features such as gazetteers, POS tags, lemmas, n-grams and affixes (Glavaš et al., 2012; Mozharova & Loukachevitch, 2016; Tkachenko, Petmanson, & Laur, 2013). Moreover, some systems include post-processing filters to improve the performance, e.g., (Pinnis, 2012). In the literature, NER systems based on CRF are the most frequent for the EMBEDDIA languages: Croatian (Glavaš et al., 2012; Štajner, 2013; Ljubešić, Stupar, Jurić, & Agić, 2013; Fišer, Ljubešić, & Erjavec, 2020), Estonian (Tkachenko et al., 2013; Dembowski, Wiegand, & Klakow, 2017), Latvian (Pinnis, 2012), Lithuanian (Pinnis, 2012; Kapočiūtė-Dzikienė, Nøklestad, Johannessen, & Krupavičius, 2013), Russian (Gareev et al., 2013; Mozharova & Loukachevitch, 2016). It should be indicated that, in occasions, these NER systems use as their

core the well-known CRF architectures, such as Mallet (McCallum, 2002) or the StanfordNER (Finkel, Grenager, & Manning, 2005). Furthermore, some of these CRF-based systems are currently used in larger NLP projects, such as the Janes Project (Fišer et al., 2020) or EstNLTK (Laur, Orasmaa, Sārg, & Tammo, 2020).

BiLSTM-based systems. In these kinds of NER systems, the core architecture is based on a Bidirectional Long Short-Term Memory (BiLSTM) neural network (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997). Specifically, most of these NER systems are based on variations from the NER tagger proposed by (Ma & Hovy, 2016) or by (Qi, Zhang, Zhang, Bolton, & Manning, 2020). For instance, we can name the work presented in our previous Deliverable D2.2, which was published as well in (Moreno, Linhares Pontes, Coustaty, & Doucet, 2019), and that supported Croatian, Russian, Slovene, among other Baltic-Slavic languages. Similarly, we can name for Russian the work of (Tsygankova, Mayhew, & Roth, 2019) and (Qi et al., 2020); for Finnish (Luoma et al., 2020); for Estonian (Kittask, Milintsevich, & Sirts, 2020) and for Latvian (Znotiņš & Cīrule, 2018).

Transformer-based systems. In the last couple of years, new deep learning technologies using Transformers (Vaswani et al., 2017) have become the new standard in the creation of NER systems. For instance, in the literature, we can find two main core transformer-based architectures: BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). The success of these architectures resides in the fact that it is necessary only to fine-tune a pre-trained language model to achieve high performance in multiple NLP tasks, including NER. Some examples regarding NER systems based on Transformers for the languages of EMBEDDIA are: Croatian (Ulčar & Robnik-Šikonja, 2020a), Estonian (Ulčar & Robnik-Šikonja, 2020a; Kittask et al., 2020; Tanvir, Kittask, & Sirts, 2020), Finnish (Ulčar & Robnik-Šikonja, 2020a; Luoma et al., 2020), Latvian (Znotiņš & Guntis Barzdīņš, 2020), Russian (Arkhipov, Trofimova, Kuratov, & Sorokin, 2019) and Swedish (Malmsten, Börjeson, & Haffenden, 2020).

Some other works that can be highlighted are those proposed by (Munro & Manning, 2012) and (Ulčar & Robnik-Šikonja, 2020b). The former explored the creation of an unsupervised multilingual NER system based on a loose alignment of texts from parallel corpora from the European Parliament. The authors explored five different European languages including Finnish, Lithuanian, and Slovene. The latter proposed an NER system using a neural network that uses ELMo contextual embeddings (Peters et al., 2018) that were trained specifically for Croatian, Estonian, Finnish, Latvian, Lithuanian, Slovenian, and Swedish.

3.2 Explored Approaches

In this section, we present four approaches that we explored for the creation of NER systems capable of processing the 8 languages of EMBEDDIA. The first approach, described in detail in Deliverable D2.2, is summarised in Section 3.2.1. An NER system based on Pseudo-affixes is presented in Section 3.2.2. Then, we introduce, in Section 3.2.3 and Section 3.2.4, two NER systems based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), an architecture that uses bidirectional language representations. For each new architecture, we describe in detail the used methodology.

3.2.1 BiLSTM with FastText and Multilingual BERT

In Deliverable D2.2, and published in (Moreno et al., 2019), we proposed a sequence labelling architecture for NER based on the work of (Ma & Hovy, 2016), which has been extended to use BERT as in (Reimers & Gurevych, 2019). In this architecture, depicted in Figure 3, we combine three types of

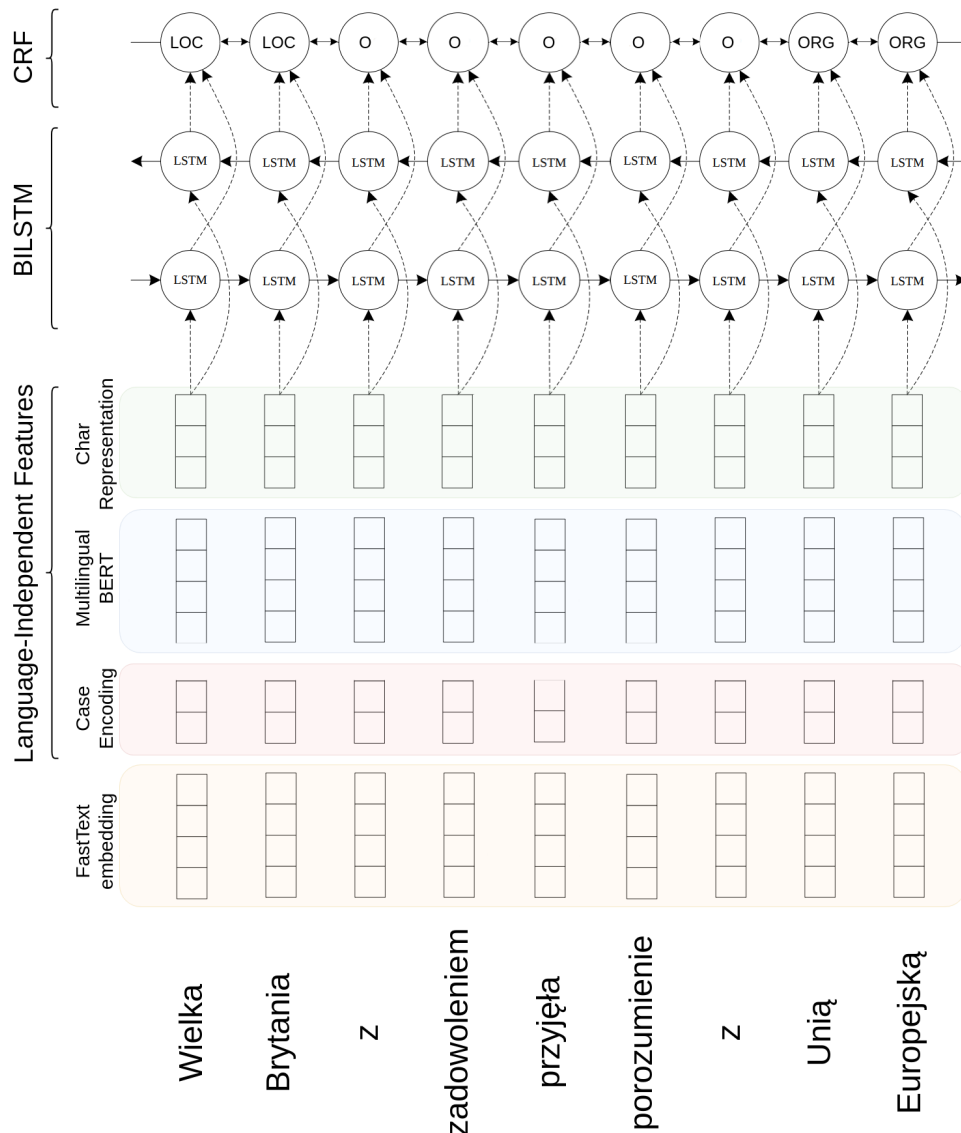


Figure 3: Proposed architecture, including an input example and the expected output.

embeddings: word embeddings, character embeddings and contextual embeddings. The NER system has as core an ensemble of Bidirectional Long Short-Term Memory (BiLSTM) networks (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997) and Conditional Random Fields (CRF) (Lafferty et al., 2001) layers.

The word embeddings come from pre-trained models generated by FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017). The contextual embeddings from the multilingual BERT (Devlin et al., 2019) are introduced using the weighted strategy proposed by (Reimers & Gurevych, 2019), where only the two first layers are combined and used in our model. The character embeddings are trained along with the NER system; it should be noted that all the characters are converted into ASCII to share the same embeddings among the different languages to process.

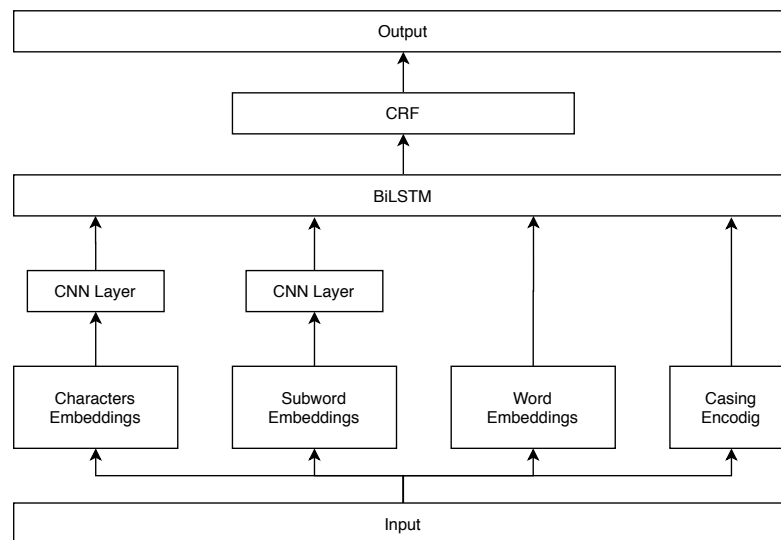


Figure 4: Architecture of the NER system based on FastText Embeddings and Pseudo-affixes.

3.2.2 BiLSTM with FastText and Pseudo-affixes Embeddings

As presented in Section 3.1, multiple works found in the state of the art make use of affixes for the improvement of NER in the languages explored in EMBEDDIA. Affixes are lexical units that can be added to a word to change its meaning or express additional information. For instance, in Russian, the suffixes *-enko* (-енко), *-in* (-ин) or *-ov* (-ов) can make reference to the name of a person (Mozharova & Loukachevitch, 2016), while in Lithuanian, suffixes like *iuose* can represent the locative case, such as in *Šiauliuose* (In Šiauliai) (Kapočiūtė-Dzikienė et al., 2013).

In the state of the art, the use of affixes can be defined in two classes: from lexica or thesauri (Gareev et al., 2013; Dalianis & Åström, 2001; Mozharova & Loukachevitch, 2016) and from n-grams of characters such as in (Tkachenko et al., 2013; Kapočiūtė-Dzikienė et al., 2013).

In this deliverable, we explore whether the inclusion of subword embeddings could be used as pseudo-affixes and therefore improve the performance of the NER systems developed for EMBEDDIA. Specifically, we utilise *BPEmb*, a collection of pre-trained subword embeddings provided by (Heinzerling & Strube, 2018). These embeddings were obtained by applying *Byte-Pair Encoding* (Gage, 1994), a compression algorithm that tries to represent different words using small but frequent units of letters, over Wikipedia and generating embeddings using GloVe (Pennington, Socher, & Manning, 2014). For instance, *BPEmb* splits the name Захаров (Zakharov) into [за, ха, ', ров]. Moreover, the most similar units for ров are [шин, нов, сов, зов, ев], which in turn are similar to suffixes -ин and -ов used in Russian for representing names of people.

The architecture for this NER system, shown in Figure 4, is based on the architecture proposed by (Ma & Hovy, 2016), which consists of a BiLSTM-CNN-CRF structure. In the following paragraphs, we explain the architecture in detail.

The proposed architecture can be divided in four type of inputs: case information, word embeddings, character embeddings and pseudo-affixes embeddings. For the former, we use the approach described in (Reimers & Gurevych, 2019), where for each token, we indicate information regarding its casing. Seven possible types of casing were considered: *numeric*, *mainly numeric*, *lowercase*, *uppercase*, *titlecase*, *contains digit* and *other*. This approach was used as well in our previous Deliverable D2.2.

For the word embeddings, we utilise FastText pre-trained models (Bojanowski et al., 2017), which were trained on multiple languages including the ones analysed in EMBEDDIA. These word embeddings convert each word into a numerical dense vector that can be interpreted by the neural network.

Regarding the characters embeddings, we add at the bottom of the neural network an embeddings layer. These embeddings are trained, along with the NER system, to create dense representation for each character found in the training dataset. As words do not have the same number of characters, we use a convolutional neural network (CNN) layer to combine the multiple character embeddings into a unique representation. This representation has the objective of keeping the most relevant information of each original embedding. These two approaches are based on the ideas of (Ma & Hovy, 2016). Nonetheless, unlike our previous work from Deliverable D2.2, we do not transliterate non-ASCII characters into ASCII ones, in order to keep the greatest amount of information available.

With respect to the pseudo-affixes, we utilise the BPEmb pre-trained models proposed by (Heinzerling & Strube, 2018) and similarly to the character embeddings, they are combined using a CNN layer. The reason for using a CNN is that the number of Pseudo-affixes in each word is variable, therefore, we cannot introduce them to the BiLSTM without doing a processing of the dense vectors.

The information from the four different inputs converge in a BiLSTM layer (Hochreiter & Schmidhuber, 1997; Schuster & Paliwal, 1997). A BiLSTM is a recurrent neural network that processes sequences of text in both directions, left to right and right to left. Moreover, it is capable of processing text sequences of variable length.

The output of the BiLSTM layer is then introduced into a conditional random fields (CRF) (Lafferty et al., 2001) layer. This neural network layer implements a statistical model that takes into account the output context in order to provide structured predictions.

3.2.3 BERT with Stacked Transformer Blocks

As presented in Section 3.1, in the last couple of years, there has been a growth of the number of NER systems based on BERT (Devlin et al., 2019). The main reason is that it is only necessary to fine-tune a pre-trained BERT model in order to achieve high performance. Furthermore, as the popularity of BERT has increased, it is more frequent to find in the literature the NER systems based on BERT models trained on just a few languages, e.g. (Ulčar & Robnik-Šikonja, 2020a; Virtanen et al., 2019; Malmsten et al., 2020), rather than on the multilingual BERT model proposed by (Devlin et al., 2019).

Despite the fact that BERT-based NER systems can reach, in general, good performance, these systems are not perfect and on occasion have to be modified to improve their stability and efficiency (Arkhipov et al., 2019; Boros, Hamdi, et al., 2020; Sun et al., 2020). For instance, in (Sun et al., 2020), the authors observed that BERT is prone to misunderstand the correct meaning of words when the BERT's tokeniser splits a word in unexpected tokens, e.g. due to misspellings mistakes or OCR errors. Furthermore, BERT can have minor issues in setting correctly the boundaries of the entities, thus, a CRF layer, as in (Ma & Hovy, 2016), is necessary after a BERT-based NER system (Arkhipov et al., 2019). For this reason, we present in this deliverable a new NER system, developed by us and published in (Boros, Linhares Pontes, et al., 2020; Boros, Hamdi, et al., 2020), that adds extra Transformer layers (Vaswani et al., 2017) blocks over BERT and a CRF layer. These additions have for objective to alleviate negative effects regarding misspelling mistakes, words out of the vocabulary, i.e. words never seen by a pre-trained model, and also contribute to the learning of the most informative words around the entities.

Figure 5 presents the proposed architecture based on BERT with Stacked Transformer Blocks. Specifically, we utilise a pre-trained BERT model to which we stack on top two encoders based on the Transformer architecture along with a CRF layer charged with producing the predictions.

A Transformer is a deep learning architecture that follows an encoder-decoder structure. In the proposed architecture, we focus on the encoding part, which processes a given input and determines which parts of the provided information are the most relevant; an encoder is composed of two main layers. The first main layer is the multi-head self-attention mechanism, which is followed by a residual connection and normalisation sub-layer. The second main layer is a position-wise fully connected feed-forward

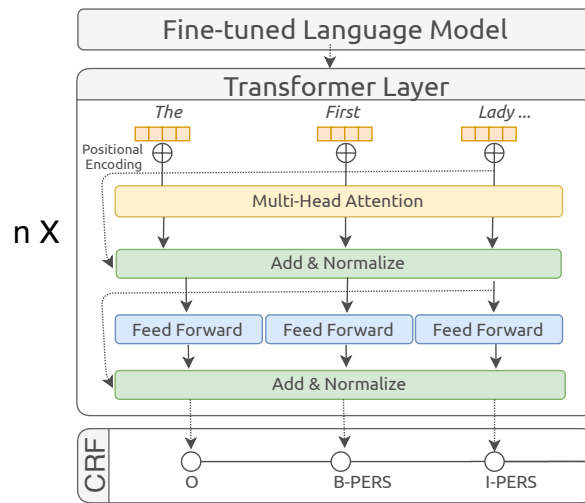


Figure 5: Proposed architecture, including an example of the expected output.

network, that is also followed by a residual connection and a normalisation sub-layer. These elements can be observed in Figure 5. In contrast to the original implementation proposed by (Vaswani et al., 2017), which used sinusoidal positional embeddings at the entry of a Transformer block, we utilise the absolute positional embeddings (Gehring, Auli, Grangier, Yarats, & Dauphin, 2017), which has become a common practice, while providing similar results (Vaswani et al., 2017).

As indicated previously, this work is associated to the publications (Boros, Hamdi, et al., 2020) and (Boros, Linhares Pontes, et al., 2020). These can be found in Appendix 2 and Appendix 14, respectively.

3.2.4 Multitask BERT

As we indicated previously in Section 3.2.3, fine-tuning BERT can produce good NER systems, but it is still necessary to add extra features in order to get the best performance. Therefore, we propose a new NER system that searches to alleviate some issues that we observed in multiple NER systems.

To be precise, we observed that BERT can have issues in analysing tokens that are in capital letters. This is similar to the aspect observed by (Sun et al., 2020), in which certain words can be tokenised by BERT into subwords that do not represent the correct idea. For example, the word *ITALY* is not segmented equally as the word *Italy* or *italy*.

In the second place, NER systems based on BERT can have trouble in determining correctly the boundaries of named entities. Therefore, the task of predicting named entities becomes harder. For instance, in the Croatian dataset HR500k (Ljubešić, Klubička, Agić, & Jazbec, 2016) the prediction of entities boundaries using BERT can be as low as a micro F-score of *0.867*, but as high as *0.937* in the Slovene dataset SSJ500k (Krek et al., 2019).

Finally, the context available for predicting the named entities might not be enough for BERT. For instance, the NER system based on BERT proposed by (Devlin et al., 2019), improved its performance by adding document context. This has been followed by other works where the authors add as many available contiguous sentences in a dataset to simulate a document context, such as in (Virtanen et al., 2019; Luoma et al., 2020; Znotiņš & Guntis Barzdīns, 2020).

Therefore, we proposed three different methods and their combination to alleviate the aforementioned issues. Specifically, we explored the substitution of uppercase tokens, the masking and prediction of

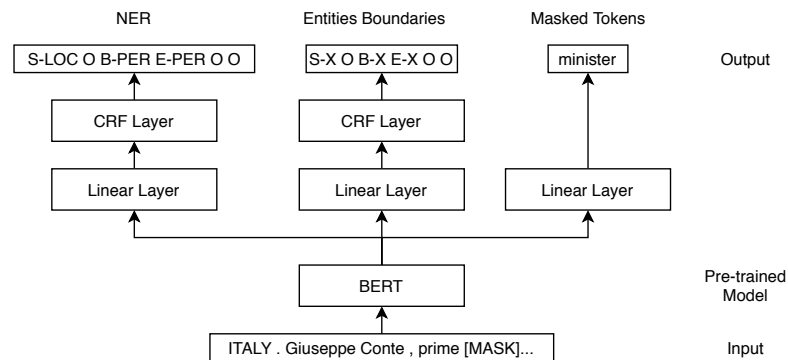


Figure 6: Proposed architecture, including an example of the expected output.

tokens, and the detection of boundaries, these last two techniques implemented in a multitask manner along with the prediction of named entities.

In Figure 6, we present the architecture of Multitask BERT, which as its name indicated, is based on multitask learning. It focuses on three different tasks: NER, prediction of masked tokens, and prediction of entity boundaries. For the NER part, we follow the architecture proposed by (Devlin et al., 2019), where after a pre-trained model, it is added a linear layer. However, following the architecture of multiple BERT-based NER systems, such as (Boros, Linhares Pontes, et al., 2020; Arkhipov et al., 2019), we add a CRF layer to improve the correct prediction of entities.

For the prediction of entities' boundaries, we use the same architecture as the one for the NER part. The only difference is that we reduce the number of possible labels. For instance, the NER part would predict labels of type *B-LOC*, *I-ORG* or *S-PER*; the entities boundaries part would only predict *B-X*, *I-X* and *S-X*. In other words, it does not consider the type, only the boundaries.

Regarding the prediction of masked tokens, we utilise the architecture used by (Devlin et al., 2019) for training a masked language model. This architecture consists of introducing the output of a pre-trained model into a linear layer, which has the same size as the pre-trained vocabulary. This linear layer is expected to predict the masked token.

It should be indicated, that during training, the losses produced by all the previously described tasks, are summed. At prediction time, the neural network only makes use of the NER part.

With respect to the marking of uppercase tokens, we use an approach similar to the one proposed by (Baldini Soares, FitzGerald, Ling, & Kwiatkowski, 2019). To be more specific, we add to BERT's vocabulary two special tokens, *[UP]*, *[up]*, that has for objective to indicate the presence of an uppercase token. Within these special tokens, we introduce three different cases of the word in analysis: the uppercase version, the title-formatted version, and the lowercase version. For instance, the word *ITALY* would be represented as *[UP] ITALY Italy italy [up]*. The reason for including the different casing variants is to give BERT more information about the possible correct casing format. It should be indicated, that only the first token, i.e. *[UP]*, is used in the prediction of the entity type and boundary; this follows the approach used by (Devlin et al., 2019) regarding the prediction of named entities with more than one subword token.

This work is associated to the publication shown in Appendix 9 and which will be submitted to a conference workshop in January 2021.

3.3 Datasets

All the previously described NER systems, in Section 3.2, have been trained and tested over the same collection of data. A more detailed description has been presented in Deliverable 2.1.

Specifically, we utilise the Wikiann collection (Pan et al., 2017). This dataset is an NER collection annotated in 282 different languages and includes all the EMBEDDIA languages. We utilise the splits for training, development, and testing used by (Rahimi, Li, & Cohn, 2019); the statistics regarding these corpora are presented in Table 1.

For certain models, we have trained and tested additional models on specific corpora that have been used previously in the state of the art. For Croatian, we explore the corpus HR500k (Ljubešić et al., 2016); for Slovene we use SSJ500k (Krek et al., 2019). In the case of Finnish, we use the annotated corpus proposed by (Luoma et al., 2020), henceforth known as *Turku*. We trained and tested on the Estonian corpus proposed by (Laur, 2013), which will be called in this deliverable as *Nimeüksuste*.

The statistics for the previously corpora are presented in Table 2 for HR500k, Table 3 for SSJ500k, Table 4 for *Turku*, and Table 5 for *Nimeüksuste*.

Table 1: Statistics regarding each language of the Wikiann corpus.

Language	Type	Train	Development	Test	Total
et	LOC	8,763	5,877	5,888	20,528
	ORG	5,834	3,909	3,875	13,618
	PER	6,194	4,057	4,129	14,380
fi	LOC	10,850	5,437	5,629	21,916
	ORG	8,367	4,194	4,180	16,741
	PER	9,665	4,627	4,745	19,037
hr	LOC	10,026	4,917	4,862	19,805
	ORG	8,374	4,085	4,100	16,559
	PER	8,697	4,467	4,404	17,568
it	LOC	4,983	5,008	4,829	14,820
	ORG	3,620	3,531	3,610	10,761
	PER	3,764	3,785	3,785	11,334
lv	LOC	5,314	4,939	5,223	15,476
	ORG	3,792	3,855	3,749	11,396
	PER	3,739	3,859	3,727	11,325
ru	LOC	9,498	4,852	4,560	18,910
	ORG	7,944	3,892	4,074	15,910
	PER	7,187	3,590	3,543	14,320
sl	LOC	7,622	5,017	5,387	18,026
	ORG	5,369	3,553	3,524	12,446
	PER	5,863	3,886	3,876	13,625
sv	LOC	10,925	4,981	5,143	21,049
	ORG	7,709	3,986	3,926	15,621
	PER	9,285	4,606	4,515	18,406
Total		173,384	104,910	105,283	383,577

Table 2: Statistics regarding the HR500k corpus.

Type	Train	Development	Test	Total
LOC	5,491	606	228	6,325
ORG	5,401	588	365	6,354
PER	5,904	670	117	6,691
Total	16,796	1,864	710	19,370

Table 3: Statistics regarding the SSJ500k corpus.

Type	Train	Development	Test	Total
LOC	1,588	169	210	1,967
MISC	498	57	47	602
ORG	1,120	124	112	1,356
PER	2,408	263	257	2,928
Total	5,614	613	626	6,853

Table 4: Statistics regarding the Finnish corpus Turku.

Type	Train	Development	Test	Total
DATE	1,099	119	114	1,332
EVENT	157	17	7	181
LOC	2,694	288	287	3,269
ORG	2,154	239	208	2,601
PER	2,477	298	310	3,085
PRO	799	102	79	980
Total	9,380	1,063	1,005	11,448

Table 5: Statistics regarding the Estonian corpus Nimeüksuste.

Type	Train	Development	Test	Total
LOC	4,742	533	436	5,711
ORG	3,266	361	311	3,938
PER	4,640	504	618	5,762
Total	12,648	1,398	1,365	15,411

3.4 Experimental Setup

We introduce the evaluation metrics used for assessing the NER systems in Section 3.4.1. Then, in the following subsections, we present the setup and the results for each of the explored models.

3.4.1 Evaluation Metrics

In this deliverable, as we did in Deliverable D2.2 and as presented in Deliverable D2.1, we evaluated the NER systems using precision, recall, and the F-score. This last metric can be averaged using either a micro or a macro approach.

Specifically, the evaluation of all the systems has been done using the *Seqeval*¹, a Python library that implements the evaluation metric used for CoNLL 2003.

In this section, we present individually the results obtained for each NER system explored in this deliverable. At the end of the section, we include also a table that includes the results of all the explored NER systems.

3.4.2 BiLSTM with FastText and Multilingual BERT Experiments

We present in Table 6 the results, obtained in Deliverable D2.2, for the four explored languages, Estonian (et), Finnish (fi), Croatian (hr) and Slovene (sl).² In Table 7, we retrained the models for the original four languages and extended the experiments in order to cover the eight EMBEDDIA languages.

Table 6: Results per entity type for the original four languages explored with the NER system based on BiLSTM with FastText and Multilingual BERT and presented in Deliverable D2.2.

Language	F-score				
	Entity Type			Average	
	LOC	ORG	PER	Macro	Micro
et	0.872	0.770	0.908	0.850	0.855
fi	0.866	0.753	0.924	0.847	0.853
hr	0.863	0.809	0.902	0.858	0.859
sl	0.886	0.855	0.921	0.887	0.888

Table 7: Results per language and per entity type for the NER system based on BiLSTM with FastText and Multilingual BERT.

Language	F-score				
	Entity Type			Average	
	LOC	ORG	PER	Macro	Micro
et	0.874	0.777	0.911	0.854	0.859
fi	0.861	0.761	0.927	0.850	0.855
hr	0.859	0.809	0.899	0.856	0.858
lt	0.836	0.799	0.890	0.842	0.843
lv	0.868	0.850	0.936	0.885	0.883
ru	0.853	0.753	0.924	0.844	0.841
sl	0.894	0.852	0.925	0.890	0.892
sv	0.956	0.831	0.923	0.903	0.909

We can observe in Table 6 and Table 7 that there are some minor differences between the scores for Estonian (et), Finnish (fi), Croatian (hr), and Slovene (sl). However, it should be indicated that these differences are normal due to the retraining of the models. To be specific, certain elements such as the randomisation seed or the initialisation of character embeddings, were not the same between the training done in D2.2 and the one done for D2.5.

It can be observed in Table 7, that in general, this NER system has problems, in the first place, with the correct prediction of *Organisation* entities, and in second place with entities of type *Location*. For the latter, the only exception occurs in Swedish (sv), where we can reach an F-score for *Locations* of 0.956.

¹<https://github.com/chakki-works/seqeval>

²In Deliverable D2.2, there was an error in the columns and the results of Organisation and Person were switched. We present the correct order in this deliverable.

3.4.3 BiLSTM with FastText and Pseudo-affixes Experiments

The subwords embeddings from BPEmb have a dimension of 25 and have a vocabulary size of 3000. In the case of FastText, the word embeddings have a size of 300 and we utilise them through FastText's original implementation which provides special features for words out of the vocabulary.³ Words and subwords embeddings were frozen during training. Regarding the neural network hyperparameters, we present in Table 8 the different values used.

Table 8: Hyperparameters used for the NER system based on a BiLSTM with FastText and Pseudo-affixes.

Hyperparameter	Value
Character Embeddings Dimension	30
Character CNN Output Dimension	30
Affixes Embeddings Dimension	25
Affixes CNN Output Dimension	25
BiLSTM Layers	1
BiLSTM Hidden Layer Dimension	400 (200 per LSTM)
BiLSTM Input Dropout	0.5
BiLSTM Output Dropout	0.5
Maximum Epochs	30
Early Stop Patience	5
Learning rate	0.001
Optimiser	Adam
Mini-batch size	2

We present the results for the BiLSTM with FastText and Pseudo-affixes (Section 3.2.2) in Table 9. For each language, we present the values of F-score obtained by the NER architecture with and without the pseudo-affixes for each language.

As it can be noticed in Table 9, for all the languages, except Russian (ru), the use of pseudo-affixes increments the performance of the proposed NER architecture based on BiLSTM. The difference in Croatian (hr) Swedish (sv), and in lesser degree Slovene (sl), for the BiLSTM with and without pseudo-affixes, is quite small. This contrasts with other languages, such as Estonian (et), Finnish (fi), Lithuanian (lt), Latvian (lv), where the difference was greater.

Based on mean results presented in Table 9, we can determine that, in general, the use of pseudo-affixes improves the performance of our NER system. Furthermore, and more specifically, the pseudo affixes helped to improve to a greater degree the recognition of *Location* entities in Estonian (et), Finnish (fi) and Latvian (lv); in a minor degree in Croatian (hr) and Lithuanian (lt). Regarding the *Organisation* entities, the pseudo-affixes helped in great proportion to the Estonian (et) and Latvian (lv) systems. Regarding the entities of type *Person*, we do not observe a particular increment in the performance of the NER system with and without pseudo-affixes expecting for Lithuanian (lt).

3.4.4 BERT with Stacked Transformer Blocks Experiments

We add a total of two Transformer blocks, each of them has a hidden size of 128 and are composed of 12 self attentions heads; these values were selected empirically. Regarding the fine-tuning hyperparameters, we make use of those proposed by (Devlin et al., 2019); only changing the learning rate to 2×10^{-5} and a mini-batch of size 4. The summary of the utilised hyperparameters can be found in

³FastText can represent words, even if they were not seen during the generation of the model, by averaging vectors of character n -grams.

Table 9: Results for each language using the NER system based on BiLSTM with FastText and Pseudo-Affixes. At the bottom of the table, we present the mean for each column.

Language	Method	F-score				
		Entity Type			Average	
		LOC	ORG	PER	Macro	Micro
et	No Affixes	0.880	0.798	0.932	0.870	0.874
	Affixes	0.905	0.819	0.939	0.888	0.892
fi	No Affixes	0.868	0.801	0.945	0.871	0.875
	Affixes	0.883	0.806	0.945	0.878	0.882
hr	No Affixes	0.880	0.845	0.921	0.882	0.883
	Affixes	0.888	0.839	0.927	0.885	0.886
it	No Affixes	0.863	0.822	0.917	0.867	0.868
	Affixes	0.870	0.829	0.926	0.875	0.876
lv	No Affixes	0.893	0.879	0.960	0.910	0.908
	Affixes	0.904	0.894	0.961	0.920	0.918
ru	No Affixes	0.889	0.821	0.947	0.886	0.884
	Affixes	0.884	0.808	0.947	0.879	0.877
sl	No Affixes	0.920	0.885	0.944	0.916	0.918
	Affixes	0.918	0.891	0.952	0.920	0.921
sv	No Affixes	0.967	0.865	0.952	0.928	0.933
	Affixes	0.967	0.871	0.952	0.930	0.935
Mean						
All	No Affixes	0.895	0.840	0.940	0.891	0.893
	Affixes	0.902	0.845	0.944	0.897	0.898

Table 10. We implemented BERT with Stacked Transformer Blocks using Hugging Face Transformers (Wolf et al., 2019), Python and FastNLP⁴.

We train up to 10 epochs, from which we select the best model based on its performance on the development partition. As well, the words that are considered by BERT as unknown, i.e. *[UNK]*, are added to BERT's vocabulary. Furthermore, sentences longer than BERT's sequence size are split into one or more sentences. During prediction, an alignment is done to produce the correct output.

Regarding the pre-trained models, we use the multilingual BERT for all EMBEDDIA languages. For six languages, we explore additional pre-trained BERT models as shown in Table 11. It is important to remark that the BERT models of FinEst and CroSloEngual (Ulčar & Robnik-Šikonja, 2020a), were generated within WP1.

It should be indicated that the two types of experiments are done. The first one, called *Baseline*, is a BERT-based NER system that consists in the previously described architecture excepting the Transformer blocks. The second experiments, i.e. the *Stacked* ones, is the architecture where we include the stacked Transformer blocks. These experiments are done in order to determine the difference between using and not using the stacked Transformer blocks.

We present in Table 12 the performance, in terms of F-score, of the application of an NER system based on BERT and BERT with Stacked Transformer Blocks (Section 3.2.3) over the corpora HR500k and SSJ500k, datasets covering texts in Croatian and Slovene respectively.

⁴<https://github.com/fastnlp/fastNLP>

Table 10: Hyperparameters used for the NER system based on a BERT with Stacked Transformer Blocks.

Hyperparameter	Value
Epochs	10
Learning Rate	2×10^{-5}
Scheduler	Linear with warm-up
Warm-up Ratio	0.1
Optimiser	AdamW with bias correction
AdamW ϵ	1×10^{-8}
Random Seed	2020
Dropout rate	0.1
Weight decay	0.01
Clipping gradient norm	1.0
BERT's Sequence Size	128
Transformer Blocks	2
Hidden Size	128
Self-Attention Heads	12
Mini-Batch	4

Table 11: Additional pre-trained BERT models explored for certain languages in the development of BERT with Stacked Transformer Blocks.

Language	Pre-trained Models
et	FinEst (Ulčar & Robnik-Šikonja, 2020a)
fi	FinBERT (Virtanen et al., 2019), FinEst (Ulčar & Robnik-Šikonja, 2020a)
hr	CroSloEngual (Ulčar & Robnik-Šikonja, 2020a)
ru	RuBERT (Kuratov & Arkhipov, 2019), Slavic BERT (Arkhipov et al., 2019)
sl	CroSloEngual (Ulčar & Robnik-Šikonja, 2020a)
sv	Swedish BERT (Malmsten et al., 2020)

Table 12: Results for the datasets HR500k and SSJ500k using BERT and BERT with Stacked with Transformer Blocks.

Dataset	Configuration		F-score					
			Entity Type				Average	
	Method	Pre-trained Model	LOC	MISC	ORG	PER	Macro	Micro
HR500k	Baseline	CroSloEngual	0.966	-	0.837	0.847	0.883	0.861
		Multilingual BERT	0.878	-	0.685	0.766	0.776	0.742
	Stacked	CroSloEngual	0.941	-	0.831	0.840	0.871	0.852
		Multilingual BERT	0.864	-	0.677	0.753	0.765	0.733
SSJ500k	Baseline	CroSloEngual	0.925	0.710	0.782	0.968	0.846	0.899
		Multilingual BERT	0.898	0.544	0.643	0.907	0.748	0.827
	Stacked	CroSloEngual	0.933	0.695	0.794	0.968	0.848	0.902
		Multilingual BERT	0.909	0.540	0.686	0.939	0.768	0.850

As presented in Table 12, we can observe that in Croatian corpus HR500k the most performing system is the one based on a BERT with a CRF layer, i.e. without the Stacked Transformer Blocks. This contrasts with the results obtained for the Slovene dataset SSJ500k, where the most performing system is the one based on Stacked Transformer Blocks. Despite that, it is clear that using a pre-trained BERT

model that only focuses on Croatian and Slovene, allows creating a more performing NER system in comparison to use a pre-trained model that was created using multiple languages.

The results for the corpus Turku when trained and tested using BERT and BERT with Stacked Transformer Blocks are presented in Table 13.

Table 13: Results for the dataset Turku using BERT and BERT with Stacked with Transformer Blocks.

Configuration		F-score							
		Entity Type						Average	
		DATE	EVENT	LOC	ORG	PER	PROD	Macro	Micro
Baseline	FinBERT	0.948	0.300	0.830	0.647	0.715	0.444	0.647	0.741
	FinEst	0.941	0.500	0.948	0.856	0.906	0.634	0.798	0.887
	Multilingual BERT	0.928	0.353	0.926	0.788	0.853	0.639	0.748	0.850
Stacked	FinBERT	0.949	0.429	0.825	0.685	0.752	0.435	0.679	0.763
	FinEst	0.957	0.400	0.935	0.907	0.932	0.729	0.810	0.913
	Multilingual BERT	0.948	0.533	0.909	0.793	0.865	0.620	0.778	0.854

We can see in Table 13, that BERT with Stacked Transformer Blocks, performs the best on the Turku corpus. This configuration helped in the prediction of entities of type Date, Organisation, Person, and Product; although it affected negatively, in a minor degree, the prediction of Locations. Furthermore, it is interesting to notice that the performance of FinBERT is much lower than the performance of FinEst and multilingual BERT. Moreover, because the corpus is not balanced in the number of named entities, as seen in Table 4, there is a difference between the micro and macro average F-score. As the macro F-score is lower than the micro F-score, we can determine that these NER systems predict better the most frequent types of entities.

The results for the Nimeüksuste corpus are presented in Table 14. We can observe that using Multilingual BERT produces models, with and without stacked Transformer blocks, that are less performing than when using FinEst. It is interesting to notice that the difference between the baseline and the stacked version does not produce large differences when using FinEst. However, when we use Multilingual BERT along with the stacked Transformer blocks, the performance can decrease in great measure.

In Table 15, we show the results regarding the Wikiann dataset obtained by BERT with and without Transformer Blocks. For each language, we indicate which pre-trained model was used to obtain the scores.

We can see in Table 15, that the addition of the Transformer Blocks provides an improvement with respect to an implementation based uniquely on a pre-trained BERT with a CRF layer. For some languages, i.e. Estonian (et), Latvian (lt), Russian (ru), and Slovene (sl), the difference between a model with and without stacked Transformer Blocks can be 0.01 points.

It is possible to notice in Table 15, as we did in Table 12, that in general pre-trained model based on fewer languages, such as RuBERT and CroSloEngual, provide better performance than pre-trained models on multiple languages as it happens in multilingual BERT. This would provide evidence to observations done by (Ulčar & Robnik-Šikonja, 2020a; Virtanen et al., 2019; Cañete, Chaperon, Fuentes, & Pérez, 2020), in which they noticed that models trained on fewer languages perform better than BERT models trained on a large set of languages. One particular exception occurs in Finnish (fi), we observe that FinBERT, a model trained only in Finnish, performs worse than the multilingual BERT and FinEst, a Finnish-Estonian BERT. This phenomenon can be observed as well in Table 13.

Table 14: Results for the dataset Nimeüksuste dataset using BERT and BERT with Stacked with Transformer Blocks.

Configuration		Entity Type			Average	
Method	Pre-trained Model	LOC	ORG	PER	Macro	Micro
Baseline	FinEst	0.908	0.838	0.954	0.900	0.912
	Multilingual BERT	0.913	0.794	0.931	0.879	0.893
Stacked	FinEst	0.915	0.844	0.956	0.905	0.916
	Multilingual BERT	0.890	0.793	0.917	0.867	0.879

3.4.5 Multitask BERT Experiments

For each language, we train eight different models. The first one, called *Multitask baseline*, is the system that only consists of the NER branch either during training or testing. The resting seven, are a combination of the methods previously described in Section 3.2.4. Following the recommendation of (Mosbach, Andriushchenko, & Klakow, 2020), we train our models up to 20 epochs using AdamW with bias correction and an early stop approach. The early stop is based on the micro F-score and the loss of the development dataset.⁵ In Table 16, we present the hyperparameters used for the training of the Multitask BERT.

The masking of tokens is done by filtering in the first place the sentences in the training partitions that are longer than three tokens. Then, at each epoch, we choose randomly 25% of the filtered sentence's tokens and substitute them with special token *[MASK]*. The sentences in the training partitions that do not fulfill the length are used for training the model but never masked.

As in (Boros, Linhares Pontes, et al., 2020), we encode the tags for the named entities using IOBES⁷, and sentences surpassing BERT's sequence size are split into multiple ones. Furthermore, it should be noted that unlike other works, such as (Devlin et al., 2019; Virtanen et al., 2019; Luoma et al., 2020), we do not introduce to BERT extra contextual information. In Table 17, the pre-trained models used in the training of our Multitask BERT. The models FinEst and CroSloEngual (Ulčar & Robnik-Šikonja, 2020a) are product of the work done within WP1.

The architecture of the explores NER were created using using Python and Huggins's Face Transformers (Wolf et al., 2019).

In Table 18, we present the results obtained from applying our Multitask BERT (Section 3.2.4) on the Croatian corpus HR500k and on the Slovene corpus SSJ500k.

We can observe in Table 18 that for the Croatian dataset, the best configuration is the one based on the prediction of boundaries. While for Slovene, the best method is the one only consisting in predicting masked tokens, although in second place we can find the configuration consisting of masked tokens along with the marking of uppercase tokens.

In Table 19, we show the results obtained from applying the Multitask BERT over the Turku corpus. As it can be seen in Table 19, the best configuration is the Multitask BERT where we predict masked tokens and mark uppercase tokens. We should highlight that, for the detection of named entities of

⁵The early stop waits for some extra epochs if the value of the micro F-score and loss is within a range of the maximum achieved.

⁶The models from the dataset Wikiann for Croatian (hr) and Slovene (sl), as well as the models for the corpus Turku (see Section 3.3), were trained on a GPU with 16GB of VRAM. The rest on a GPU with 10GB of VRAM. Thus, we had to reduce the batch size depending on the model and the capacities of the GPU.

⁷IOBES (Inside-Outside-Beginning-End-Single) is an annotation scheme frequently used for tagging tokens, such as in NER. This scheme allows representing aspects like the beginning and end of a chunk belonging to a named entity.

⁸For the corpus Turku (see Section 3.3), we make use of FinBERT (Virtanen et al., 2019)

Table 15: Results for each language using the NER system based on BERT and BERT with Stacked with Transformer Blocks.

Language	Configuration		F-score				
			Entity Type			Average	
	Method	Pre-trained Model	LOC	ORG	PER	Macro	Micro
et	Baseline	FinEst	0.945	0.901	0.960	0.935	0.937
		Multilingual BERT	0.938	0.882	0.954	0.924	0.927
	Stacked	FinEst	0.953	0.913	0.966	0.944	0.946
		Multilingual BERT	0.941	0.886	0.959	0.929	0.931
fi	Baseline	FinBERT	0.894	0.827	0.939	0.887	0.890
		FinEst	0.933	0.886	0.964	0.928	0.930
		Multilingual BERT	0.922	0.859	0.954	0.912	0.914
	Stacked	FinBERT	0.902	0.848	0.946	0.899	0.901
		FinEst	0.934	0.890	0.964	0.929	0.931
		Multilingual BERT	0.926	0.874	0.954	0.918	0.920
hr	Baseline	CroSloEngual	0.937	0.913	0.958	0.936	0.936
		Multilingual BERT	0.926	0.895	0.952	0.924	0.925
	Stacked	CroSloEngual	0.937	0.919	0.960	0.938	0.939
		Multilingual BERT	0.933	0.902	0.955	0.930	0.930
it	Baseline	Multilingual BERT	0.900	0.867	0.940	0.902	0.903
	Stacked	Multilingual BERT	0.910	0.876	0.946	0.911	0.911
lv	Baseline	Multilingual BERT	0.924	0.910	0.965	0.933	0.932
	Stacked	Multilingual BERT	0.933	0.918	0.972	0.941	0.940
ru	Baseline	Multilingual BERT	0.899	0.825	0.948	0.891	0.888
		RuBERT	0.911	0.849	0.949	0.903	0.901
		Slavic BERT	0.897	0.828	0.946	0.890	0.888
	Stacked	Multilingual BERT	0.908	0.846	0.952	0.902	0.900
		RuBERT	0.917	0.868	0.958	0.914	0.913
		Slavic BERT	0.905	0.843	0.950	0.899	0.897
sl	Baseline	CroSloEngual	0.944	0.919	0.967	0.944	0.944
		Multilingual BERT	0.942	0.913	0.963	0.940	0.941
	Stacked	CroSloEngual	0.954	0.929	0.976	0.953	0.954
		Multilingual BERT	0.946	0.921	0.970	0.946	0.946
sv	Baseline	Swedish BERT	0.975	0.905	0.959	0.947	0.950
		Multilingual BERT	0.978	0.911	0.957	0.949	0.952
	Stacked	Swedish BERT	0.979	0.919	0.967	0.955	0.958
		Multilingual BERT	0.977	0.918	0.963	0.953	0.956

type Event, we can really improve the performance by marking uppercase tokens. As we indicated in Section 3.4.4, a micro F-score greater than the macro F-score, indicates that the system focuses on the most frequent entity types. Nonetheless, by marking uppercase tokens, we can observe that we can improve the macro F-score while keeping a competitive micro F-score.

We present the results for the Nimeüksuste corpus using the multi-task BERT in Table 20. It can be observed in Table 20, that the most performing system is the baseline of the multi-task BERT and in the second place, it is the BERT trained with masked tokens. The difference between these two models is 0.004, however, with respect to other models, we can see that this difference can be around 0.10 points

Table 16: Hyperparameters used for the NER system based on a Multitask BERT.

Hyperparameter	Value
Maximum Epochs	20
Early Stop Patience	3
Learning Rate	2×10^{-5}
Scheduler	Linear with warm-up
Warm-up Ratio	0.1
Optimiser	AdamW with bias correction
AdamW ϵ	1×10^{-8}
Random Seed	12
Dropout rate	0.1
Weight decay	0.01
Clipping gradient norm	1.0
BERT's Sequence Size	128
Masking Percentage	25%
Training Mini-Batch ⁶	
Wikiann dataset	
hr, sl	32
et, fi, sv	16
lt, lv, ru	8
Nimeüksuste corpus	
No Masked tokens	32
Masked tokens	16
Other datasets	32
Testing Mini-Batch	8

Table 17: Additional pre-trained BERT models explored for certain languages in the development of Multitask BERT.

Language	Pre-trained Models
et	FinEst (Ulčar & Robnik-Šikonja, 2020a)
fi ⁸	FinEst (Ulčar & Robnik-Šikonja, 2020a)
hr	CroSloEngual (Ulčar & Robnik-Šikonja, 2020a)
lt	Multilingual BERT (Devlin et al., 2019)
lv	Multilingual BERT (Devlin et al., 2019)
ru	RuBERT (Kuratov & Arkhipov, 2019)
sl	CroSloEngual (Ulčar & Robnik-Šikonja, 2020a)
sv	Swedish BERT (Malmsten et al., 2020)

in most cases. The only exception is when we mask tokens, predict boundaries, and mark uppercase tokens, where the difference is 0.20 points.

The results regarding the Wikiann corpus are presented in Table 21 and Table 22. These results are presented in terms of F-score for each language, entity type, and two different averaging methods, micro and macro.

We can observe in Table 21 and Table 22 that for Latvian (lv) and Russian (ru), there is a larger margin of improvement between the combinations with and without prediction of masked tokens, with respect to other languages. As well, we can notice, that for some languages, Estonian (et), Croatian (hr), Latvian (lv), Russian (ru) and Swedish (sv), the prediction of boundaries and/or the marking of uppercase tokens without the prediction of masked tokens, can decrease, in general, the performance.

Table 18: Results for the datasets HR500k and SSJ500k using Multi-task BERT. All the models were trained using the pre-trained BERT model CroSloEngual.

Dataset	Method	F-score					
		Entity Type				Average	
		LOC	MISC	ORG	PER	Macro	Micro
HR500k	Multitask Baseline	0.954	-	0.791	0.849	0.865	0.836
	Boundaries	0.962	-	0.843	0.881	0.895	0.874
	Uppercase	0.971	-	0.837	0.829	0.879	0.857
	Boun. + Upper.	0.957	-	0.809	0.865	0.877	0.850
	Masked	0.957	-	0.831	0.818	0.869	0.848
	Masked + Boundaries	0.946	-	0.834	0.822	0.867	0.848
	Masked + Uppercase	0.949	-	0.844	0.838	0.877	0.860
	Masked + Boun. + Upper.	0.953	-	0.839	0.824	0.872	0.853
SSJ500k	Multitask Baseline	0.912	0.748	0.818	0.964	0.860	0.903
	Boundaries	0.926	0.695	0.800	0.948	0.842	0.892
	Uppercase	0.926	0.748	0.784	0.950	0.852	0.896
	Boun. + Upper.	0.927	0.781	0.769	0.943	0.855	0.893
	Masked	0.933	0.828	0.831	0.973	0.891	0.924
	Masked + Boundaries	0.926	0.748	0.805	0.955	0.859	0.902
	Masked + Uppercase	0.931	0.854	0.811	0.968	0.891	0.919
	Masked + Boun. + Upper.	0.926	0.772	0.791	0.957	0.862	0.903

Table 19: Results for the dataset Turku using Multi-task BERT.

Method	F-score							
	Entity Type						Average	
	DATE	EVENT	LOC	ORG	PER	PROD	Macro	Micro
Multitask Baseline	0.970	0.471	0.949	0.859	0.937	0.681	0.811	0.905
Boundaries	0.966	0.500	0.937	0.870	0.948	0.699	0.820	0.910
Uppercase	0.974	0.429	0.935	0.895	0.934	0.688	0.809	0.909
Boun. + Upper.	0.966	0.471	0.936	0.876	0.946	0.652	0.808	0.906
Masked	0.965	0.625	0.936	0.865	0.941	0.671	0.834	0.905
Masked + Boundaries	0.957	0.533	0.939	0.887	0.949	0.639	0.817	0.909
Masked + Uppercase	0.969	0.667	0.936	0.877	0.950	0.667	0.844	0.912
Masked + Boun. + Upper.	0.961	0.462	0.935	0.884	0.947	0.693	0.814	0.910

In all the cases, it is the prediction of masked tokens, the approach that globally, improves the performance of the NER system. For instance, we can see in Table 21 and Table 22, that for four languages, Estonian (et), Finnish (fi), Lithuanian (lt), and Russian (ru), the most performing configuration is the Multitask BERT where we predict masked tokens and we predict boundaries during the training. While for Slovene (sl) and Swedish (sv), the best model is the one where we predict masked tokens and mark uppercase tokens. Croatian (hr) is the only dataset that gets the best performance by using all the methods described in Section 3.2.4. And Latvian (lv) is the only language where predicting masked tokens provides the best performance in terms of micro F-score.

Table 20: Results for the dataset Nimeüksuste dataset using Multi-task BERT.

Method	F-score				
	Entity Type			Average	
	LOC	ORG	PER	Macro	Micro
Multitask Baseline	0.932	0.851	0.958	0.913	0.924
Boundaries	0.919	0.838	0.958	0.905	0.916
Uppercase	0.932	0.835	0.947	0.905	0.916
Boun. + Upper.	0.928	0.840	0.950	0.906	0.916
Masked	0.931	0.841	0.954	0.909	0.920
Masked + Boundaries	0.917	0.846	0.952	0.905	0.915
Masked + Uppercase	0.918	0.830	0.956	0.901	0.914
Masked + Boun. + Upper.	0.925	0.829	0.931	0.895	0.904

3.4.6 Comparative Results for all the Explored Systems

We present in Table 23 a summary of all the results, in terms of micro and macro F-score, for all the explored NER systems presented in this deliverable.

Table 21: Results for Estonian, Finnish, Croatian and Lithuanian languages using the NER system based on Multitask BERT.

Language	Method	F-score				
		Entity Type			Average	
		LOC	ORG	PER	Macro	Micro
et	Multitask Baseline	0.954	0.914	0.966	0.945	0.946
	Boundaries	0.949	0.910	0.967	0.942	0.944
	Uppercase	0.950	0.906	0.966	0.941	0.943
	Boun. + Upper.	0.953	0.905	0.967	0.942	0.944
	Masked	0.951	0.904	0.964	0.940	0.942
	Masked + Boundaries	0.957	0.917	0.968	0.947	0.949
	Masked + Uppercase	0.953	0.905	0.964	0.941	0.943
	Masked + Boun. + Upper.	0.957	0.911	0.966	0.945	0.947
fi	Multitask Baseline	0.936	0.894	0.967	0.932	0.934
	Boundaries	0.939	0.894	0.969	0.934	0.936
	Uppercase	0.938	0.901	0.968	0.935	0.937
	Boun. + Upper.	0.935	0.892	0.963	0.930	0.932
	Masked	0.934	0.886	0.965	0.928	0.930
	Masked + Boundaries	0.942	0.904	0.971	0.939	0.941
	Masked + Uppercase	0.938	0.896	0.970	0.934	0.936
	Masked + Boun. + Upper.	0.942	0.899	0.971	0.937	0.939
hr	Multitask Baseline	0.941	0.923	0.962	0.942	0.942
	Boundaries	0.939	0.920	0.963	0.941	0.941
	Uppercase	0.939	0.918	0.960	0.939	0.939
	Boun. + Upper.	0.937	0.919	0.959	0.938	0.939
	Masked	0.944	0.924	0.964	0.944	0.944
	Masked + Boundaries	0.944	0.923	0.965	0.944	0.945
	Masked + Uppercase	0.942	0.922	0.965	0.943	0.944
	Masked + Boun. + Upper.	0.945	0.925	0.965	0.945	0.946
lt	Multitask Baseline	0.905	0.885	0.945	0.912	0.912
	Boundaries	0.912	0.888	0.950	0.917	0.917
	Uppercase	0.909	0.885	0.944	0.913	0.913
	Boun. + Upper.	0.905	0.874	0.942	0.907	0.907
	Masked	0.909	0.888	0.950	0.916	0.915
	Masked + Boundaries	0.919	0.892	0.953	0.921	0.922
	Masked + Uppercase	0.917	0.893	0.953	0.921	0.921
	Masked + Boun. + Upper.	0.916	0.895	0.952	0.921	0.921

Table 22: Results for Latvian, Russian, Slovene and Swedish languages using the NER system based on Multitask BERT.

Language	Method	F-score				
		Entity Type			Average	
		LOC	ORG	PER	Macro	Micro
lv	Multitask Baseline	0.932	0.922	0.970	0.941	0.940
	Boundaries	0.928	0.920	0.968	0.939	0.938
	Uppercase	0.931	0.920	0.970	0.940	0.939
	Boun. + Upper.	0.931	0.921	0.971	0.941	0.940
	Masked	0.941	0.928	0.977	0.949	0.948
	Masked + Boundaries	0.940	0.927	0.976	0.948	0.947
	Masked + Uppercase	0.939	0.926	0.978	0.948	0.947
	Masked + Boun. + Upper.	0.940	0.927	0.976	0.948	0.947
ru	Multitask Baseline	0.915	0.864	0.959	0.913	0.911
	Boundaries	0.913	0.857	0.961	0.910	0.908
	Uppercase	0.916	0.856	0.956	0.909	0.908
	Boun. + Upper.	0.913	0.858	0.955	0.909	0.907
	Masked	0.919	0.865	0.959	0.914	0.913
	Masked + Boundaries	0.920	0.871	0.959	0.917	0.915
	Masked + Uppercase	0.919	0.864	0.960	0.914	0.912
	Masked + Boun. + Upper.	0.916	0.860	0.960	0.912	0.910
sl	Multitask Baseline	0.952	0.929	0.973	0.951	0.952
	Boundaries	0.950	0.935	0.973	0.953	0.953
	Uppercase	0.950	0.928	0.972	0.950	0.951
	Boun. + Upper.	0.953	0.931	0.975	0.953	0.954
	Masked	0.946	0.927	0.977	0.950	0.950
	Masked + Boundaries	0.954	0.930	0.977	0.953	0.954
	Masked + Uppercase	0.953	0.934	0.978	0.955	0.956
	Masked + Boun. + Upper.	0.953	0.928	0.975	0.952	0.953
sv	Multitask Baseline	0.978	0.919	0.963	0.954	0.956
	Boundaries	0.976	0.912	0.961	0.950	0.953
	Uppercase	0.974	0.912	0.959	0.948	0.951
	Boun. + Upper.	0.977	0.921	0.961	0.953	0.956
	Masked	0.978	0.923	0.968	0.956	0.959
	Masked + Boundaries	0.976	0.919	0.967	0.954	0.957
	Masked + Uppercase	0.978	0.924	0.968	0.956	0.959
	Masked + Boun. + Upper.	0.978	0.923	0.966	0.956	0.958

F-score per Language																		
Architecture		Method	et		fi		hr		lt		lv		ru		sl		sv	
		Embeddings	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
BiLSTM	Deliverable D2.2	FastText, Multilingual BERT	0.854	0.859	0.850	0.855	0.856	0.858	0.842	0.843	0.885	0.883	0.844	0.841	0.880	0.892	0.903	0.909
	Baseline	FastText	0.870	0.874	0.871	0.875	0.882	0.883	0.867	0.868	0.910	0.908	0.886	0.884	0.916	0.918	0.928	0.933
BERT	Affixes	FastText, BPEmb	0.888	0.892	0.878	0.882	0.885	0.886	0.875	0.876	0.920	0.918	0.879	0.877	0.920	0.921	0.930	0.935
	Baseline	Multilingual BERT	0.924	0.927	0.912	0.914	0.924	0.925	0.902	0.903	0.933	0.932	0.891	0.888	0.940	0.941	0.949	0.952
		CroSloEngual	-	-	-	0.887	0.890	0.936	0.936	-	-	-	-	-	0.944	0.944	-	-
		FinBERT	-	-	-	0.887	0.890	-	-	-	-	-	-	-	-	-	-	-
		FinEst	0.935	0.937	0.928	0.930	-	-	-	-	-	-	-	0.903	0.901	-	-	-
BERT	Stacked	RuBERT	-	-	-	-	-	-	-	-	-	-	0.903	0.901	-	-	-	-
		Slavic BERT	-	-	-	-	-	-	-	-	-	-	0.890	0.888	-	-	-	-
		Swedish BERT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.947	0.950
		Multilingual BERT	0.929	0.931	0.918	0.920	0.930	0.930	0.911	0.911	0.941	0.940	0.902	0.900	0.946	0.946	0.953	0.956
		CroSloEngual	-	-	-	0.899	0.901	0.938	0.939	-	-	-	-	-	-	0.953	0.954	-
Multitask BERT	Masked	FinBERT	-	-	-	0.929	0.931	-	-	-	-	-	-	-	-	-	-	-
		FinEst	0.944	0.946	0.929	0.931	-	-	-	-	-	-	-	0.914	0.913	-	-	-
		RuBERT	-	-	-	-	-	-	-	-	-	-	-	0.899	0.897	-	-	-
		Slavic BERT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		Swedish BERT	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.955	0.958
Multitask BERT	Masked	Multi-task Baseline	0.945	0.946	0.932	0.934	0.942	0.942	0.912	0.912	0.941	0.940	0.913	0.911	0.951	0.952	0.954	0.956
		Boundaries	0.942	0.944	0.934	0.936	0.941	0.941	0.917	0.917	0.939	0.938	0.910	0.908	0.953	0.953	0.950	0.953
		Uppercase	0.941	0.943	0.935	0.937	0.939	0.939	0.913	0.913	0.940	0.939	0.909	0.908	0.950	0.951	0.948	0.951
		Boun. + Upperc.	0.942	0.944	0.930	0.932	0.938	0.939	0.907	0.907	0.941	0.940	0.909	0.907	0.953	0.954	0.953	0.956
		Masked	0.940	0.942	0.928	0.930	0.944	0.944	0.916	0.915	0.949	0.948	0.914	0.913	0.950	0.950	0.956	0.959
Multitask BERT	Masked + Boundaries	Masked	0.947	0.949	0.939	0.941	0.944	0.945	0.921	0.922	0.948	0.947	0.917	0.915	0.953	0.954	0.954	0.957
		Masked + Uppercase	0.941	0.943	0.934	0.936	0.943	0.944	0.921	0.921	0.948	0.947	0.914	0.912	0.955	0.956	0.956	0.959
		Masked + Boun. + Upperc.	0.945	0.947	0.937	0.939	0.945	0.946	0.921	0.921	0.948	0.947	0.912	0.910	0.952	0.953	0.956	0.958

From Table 23, we can observe that the system proposed in Section 3.2.2, based on a BiLSTM with FastText embeddings, performs in all languages better than the system proposed in Deliverable D2.2. Furthermore, the addition of pseudo-affixes, increases, in most of the languages, the performance of the BiLSTM architecture.

In conclusion, we can see in Table 23, that the use of BERT can outperform BiLSTM architectures. In most languages, a simple BERT with a CRF, e.g. *BERT baseline* or *Multi-task Baseline*, can completely boost the performance of an NER system. The only exceptions occur in Russian (ru). For this language, *BERT baseline* using Multilingual and Slavic BERT have a performance that is similar to BiLSTM using pseudo-affixes.

For some languages, the difference between the F-scores of BERT with Stacked Transformer Blocks and Multitask BERT is quite small. Some exceptions are Finnish (fi) and Lithuanian (lt), where the difference is around 0.01 points, and in minor degree Croatian (hr) and Latvian (lv), where the difference is around 0.007 points. For the rest of the languages, the difference is minimal, although it shows a tendency over the Multitask BERT.

3.4.7 Comparative Results with the State of the Art

In this subsection, we compare our NER systems based on BERT against tools found in the state of the art. Specifically, we compare against four baselines for Croatian using the dataset HR500k: CroSloEngual (Ulčar & Robnik-Šikonja, 2020a), Janes-NER (Fišer et al., 2020), Polygot (Al-Rfou, Kulkarni, Perozzi, & Skiena, 2014) and Croatian NERC (Bekavac & Tadić, 2007). These last two systems were tested by (Alves, Thakkar, & Tadić, 2020). Regarding the Slovene, we compare against two baselines using SSJ500k: CroSloEngual (Ulčar & Robnik-Šikonja, 2020a) and Janes-NER (Fišer et al., 2020). It should be indicated that for both datasets the evaluation found in the state of the art is not standardised and we had to recalculate some scores.

In the case of Janes-NER, the system is evaluated using the Babushka-Bench⁹. This evaluation tool does not consider incorrect boundaries, e.g. B-LOC is the same as I-LOC, and counts the predictions of the *Other* type in order to calculate the macro F-score; this last point is infrequent in the state of the art and can inflate the performance of the system, as the prediction of the *Other* type is the easiest to learn, due to their extremely high frequency. Furthermore, it is indicated that the corpus HR500k comprises annotations regarding the *Miscellaneous* type, however, other works in the state of the art do not make reference to this type of entity (Ulčar & Robnik-Šikonja, 2020a; Alves et al., 2020). With respect to CroSloEngual, the authors evaluate their NER system using only three named entities types in the Slovene dataset SSJ500k: *Location*, *Person* and *Organisation*, leaving out the *Miscellaneous* type.

We present in Table 24, the results for the state-of-the-art systems and those presented in this deliverable using the three common named entities types (*Location*, *Person* and *Organisation*). As we can observe in Table 24, we can improve the state of the art for the Croatian dataset HR500k. However, unlike for other datasets and languages, the best-performing system is based on Multitask BERT with boundaries prediction.

In the case of the Slovene dataset, SSJ500k, we can see in Table 24, that we do not surpass the performance of the state of the art set by (Ulčar & Robnik-Šikonja, 2020a). However, our Slovene model was trained over four entity types, while the model of (Ulčar & Robnik-Šikonja, 2020a) was only trained over three. It should be indicated as well that Janes-NER gets an F-score for *Miscellaneous* of 0.270 while our masked model gets 0.828. We can notice also that the performance of the Multitask BERT decays when the combination does not include the prediction of masked tokens.

We show in Table 25 a comparison between our developed tools and those from the state of the art that have used the Turku corpus. As it can be seen in Table 25, the most performing system, in terms of micro F-score, is the one proposed by (Luoma et al., 2020), while in terms of macro F-score is the

⁹<https://github.com/clarinsi/babushka-bench>

Table 24: Recalculation of macro F-scores for the datasets HR500k and SSJ500k, with respect to the common named entities, used along the systems explored in this deliverable and in the literature. All the methods of Deliverable D2.5 were trained using as initial model CroSloEngual (Ulčar & Robnik-Šikonja, 2020).

Dataset	Method	F-score			Macro Average
		Entity Type			
		LOC	ORG	PER	
HR500k	Ulčar & Robnik-Šikonja, 2020				
	Multilingual BERT	NA	NA	NA	0.790
	XLM-RoBERTa	NA	NA	NA	0.817
	CroSloEngual	NA	NA	NA	0.884
	Fišer et al., 2020				
	Janes-NER	0.890	0.850	0.720	0.820
	Alves, Thakkar, & Tadić, 2020				
	Polyglot	NA	NA	NA	0.622
	Croatian NERC	NA	NA	NA	0.640
	Deliverable D2.5				
	BERT Baseline	0.966	0.837	0.847	0.883
	Stacked	0.941	0.831	0.840	0.871
	Multitask Baseline	0.954	0.791	0.849	0.865
	Boundaries	0.962	0.843	0.881	0.895
	Uppercase	0.971	0.837	0.829	0.879
	Boun. + Upperc.	0.957	0.809	0.865	0.877
	Masked	0.957	0.831	0.818	0.869
	Masked + Boundaries	0.946	0.834	0.822	0.867
	Masked + Uppercase	0.949	0.844	0.838	0.877
	Masked + Boun. + Upperc.	0.953	0.839	0.824	0.872
SSJ500k	Ulčar & Robnik-Šikonja, 2020				
	Multilingual BERT	NA	NA	NA	0.897
	XLM-RoBERTa	NA	NA	NA	0.914
	CroSloEngual	NA	NA	NA	0.920
	Fišer et al., 2020				
	Janes-NER	0.890	0.800	0.670	0.786
	Deliverable D2.5				
	BERT Baseline	0.925	0.782	0.968	0.891
	Stacked	0.933	0.794	0.968	0.898
	Multitask Baseline	0.912	0.818	0.964	0.898
	Boundaries	0.926	0.800	0.948	0.891
	Uppercase	0.926	0.784	0.950	0.886
	Boun. + Upperc.	0.927	0.769	0.943	0.879
	Masked	0.933	0.831	0.973	0.912
	Masked + Boundaries	0.926	0.805	0.955	0.896
	Masked + Uppercase	0.931	0.811	0.968	0.903
	Masked + Boun. + Upperc.	0.926	0.791	0.957	0.894

Multi-task BERT with marked uppercase tokens. The biggest difference comes from the performance of two entities types, Events, and Organisations.

Table 25: Comparison between methods from the state of the art and the systems developed in this deliverable for the dataset Turku.

Configuration		F-score							
		Entity Type						Average	
Method	Pre-trained Model	DATE	EVENT	LOC	ORG	PER	PROD	Macro	Micro
Luoma et al., 2020									
FiNER	-	NA	NA	NA	NA	NA	NA	NA	0.740
BiLSTM-CNN-CRF	-	NA	NA	NA	NA	NA	NA	NA	0.815
	FinBERT	0.968	0.435	0.947	0.902	0.952	0.658	0.810	0.916
Deliverable D2.5									
BERT Baseline	FinBERT	0.948	0.300	0.830	0.647	0.715	0.444	0.647	0.741
	Finest	0.941	0.500	0.948	0.856	0.906	0.634	0.798	0.887
	Multilingual BERT	0.928	0.353	0.926	0.788	0.853	0.639	0.748	0.850
Stacked	FinBERT	0.949	0.429	0.825	0.685	0.752	0.435	0.679	0.763
	Finest	0.957	0.400	0.935	0.907	0.932	0.729	0.810	0.913
	Multilingual BERT	0.948	0.533	0.909	0.793	0.865	0.620	0.778	0.854
Multitask Baseline		0.970	0.471	0.949	0.859	0.937	0.681	0.811	0.905
Boundaries		0.966	0.500	0.937	0.870	0.948	0.699	0.820	0.910
Uppercase		0.974	0.429	0.935	0.895	0.934	0.688	0.809	0.909
Boun. + Upper.	FinBERT	0.966	0.471	0.936	0.876	0.946	0.652	0.808	0.906
Masked		0.965	0.625	0.936	0.865	0.941	0.671	0.834	0.905
Masked + Boundaries		0.957	0.533	0.939	0.887	0.949	0.639	0.817	0.909
Masked + Uppercase		0.969	0.667	0.936	0.877	0.950	0.667	0.844	0.912
Masked + Boun. + Upper.		0.961	0.462	0.935	0.884	0.947	0.693	0.814	0.910

Regarding the results for the Nimeüksuste corpus, we present in Table 26 the results obtained by us in this deliverable, as well as other tools from the state of the art. Specifically, we compare our results against the experiments done by (Kittask et al., 2020; Tanvir et al., 2020; Ulčar & Robnik-Šikonja, 2020a). It is important to indicate that the experiments done by (Kittask et al., 2020) and (Tanvir et al., 2020) explored different BERT sequence sizes. Moreover, in the case of (Kittask et al., 2020), they experimented with the addition of consecutive sentences to increase the context as in (Devlin et al., 2019; Luoma et al., 2020). In the case of (Ulčar & Robnik-Šikonja, 2020a), the authors used for all their models a sequence size of 512 tokens.

We can observe in Table 26, that the best system in terms of macro F-score is the BERT model proposed by (Ulčar & Robnik-Šikonja, 2020a), while in terms of micro F-score is our multi-task baseline. Both of these systems are based on FinEst, however, in (Ulčar & Robnik-Šikonja, 2020a), the authors make use of a simple fine-tuning of BERT; in other words, they do not make use of a CRF layer unlike us.

3.5 Discussion

In Table 23 we can observe that most of the languages were improved with the addition of pseudo-affixes, with the Estonian (et), Finnish (fi), Lithuanian (lt) and Latvian (lv) getting an improvement greater of 0.05 points. The exception was Russian (ru), which was affected negatively in comparison with the baseline described in Section 3.2.2. We were expecting to get an improvement for Finnish (fi) at the level of the Estonian (et), i.e. around 0.020 points, as both languages are agglutinant and they make affixes in their grammar. It was interesting to notice that the Lithuanian (lt) and Latvian (lv) were improved by

Table 26: Comparison between methods from the state of the art and the systems developed in this deliverable for the dataset Nimeüksuste.

Configuration		F-score				
		Entity Type			Average	
Method	Pre-trained Model	LOC	ORG	PER	Macro	Micro
Kittask et al., 2020						
CRF	-	NA	NA	NA	NA	0.879
Stanza	-	NA	NA	NA	NA	0.908
Seq. Size 128	Multilingual BERT	NA	NA	NA	NA	0.865
Seq. Size 512		NA	NA	NA	NA	0.883
Seq. Size 512 + Context		NA	NA	NA	NA	0.880
Seq. Size 128	XLM-RoBERTa	NA	NA	NA	NA	0.893
Seq. Size 512		NA	NA	NA	NA	0.891
Seq. Size 512 + Context		NA	NA	NA	NA	0.901
Tanvir et al., 2020						
Seq. Size 128	EstBERT	NA	NA	NA	NA	0.893
Seq. Size 512		NA	NA	NA	NA	0.890
Ulčar & Robnik-Šikonja, 2020						
	Multilingual BERT	NA	NA	NA	0.898	NA
	XLM-RoBERTa	NA	NA	NA	0.908	NA
	FinEst	NA	NA	NA	0.927	NA
	FinBERT	NA	NA	NA	0.876	NA
Deliverable D2.5						
BERT Baseline	FinEst	0.908	0.838	0.954	0.900	0.912
	Multilingual BERT	0.913	0.794	0.931	0.879	0.893
Stacked	FinEst	0.915	0.844	0.956	0.905	0.916
	Multilingual BERT	0.890	0.793	0.917	0.867	0.879
Multitask Baseline	FinEst	0.932	0.851	0.958	0.913	0.924
Boundaries		0.919	0.838	0.958	0.905	0.916
Uppercase		0.932	0.835	0.947	0.905	0.916
Boun. + Upper.		0.928	0.840	0.950	0.906	0.916
Masked		0.931	0.841	0.954	0.909	0.920
Masked + Boundaries		0.917	0.846	0.952	0.905	0.915
Masked + Uppercase		0.918	0.830	0.956	0.901	0.914
Masked + Boun. + Upper.		0.925	0.829	0.931	0.895	0.904

using the pseudo-affixes.

With respect to the low performance of Multilingual BERT and Slavic BERT on Russian, see Table 23, it might be caused due to multiple aspects. In the first place, the vocabulary size of the Slavic BERT and RuBERT is the same (Kuratov & Arkhipov, 2019; Arkhipov et al., 2019), however, in the former, the vocabulary is shared with multiple languages; as well, not all the languages covered by the Slavic BERT are Cyrillic, i.e. Polish and Czech. In the second place, multilingual BERT might not have enough data regarding the Russian language or the Cyrillic alphabet, decreasing the performance of the pre-trained model.

Based on the outputs presented in Table 21, Table 22, and Table 24, we can notice that for most languages, the masking of tokens, improves the performance of an NER system. There can be multiple reasons, for instance, it can help in forcing BERT to focus on the surrounding words in order to determine

whether a word is a named entity or not. In other words, we can be teaching BERT to look for possible named entity *triggers*. Another option is that we improve the BERT language model, as it is possible that certain words or contexts were not seen during the creation of the language model.

With respect to the Croatian dataset HR500k, we can observe in Table 24 that our Multitask BERT with boundaries prediction can improve the results with respect to the fine-tuned CroSloEngual NER system (Ulčar & Robnik-Šikonja, 2020a). Nonetheless, the masking of tokens, using the Multitask BERT, affects negatively the performance of the NER system, specially the prediction of entities of type Person.

Regarding the Slovene Dataset SSJ500k, as we indicated in Section 3.4.7, we did not manage to outperform the score obtained by (Ulčar & Robnik-Šikonja, 2020a). One of the reasons can be due to the fact that we trained a model over four possible named entities, while (Ulčar & Robnik-Šikonja, 2020a) only over three. Although the type *Miscellaneous* in SSJ500k is one of the less frequent named entity types, training a model with it, can induce some additional noise with respect to a model that was not trained over this named entity type.

With respect to the results obtained for the corpus Turku, presented in Table 25, we did not arrive to improve the scores in the state of the art in terms of micro F-score. However, we managed to improve the performance of the macro F-score while keeping a competitive micro F-score. Despite the results, it should be noted that the NER system trained by (Luoma et al., 2020) provided as many surroundings sentences were possible in FinBERT. This was done to replicate, up to a certain degree, the document context approach used by (Devlin et al., 2019) for improving the performance of their fine-tuned BERT over English CoNLL 2003 (Tjong Kim Sang & De Meulder, 2003). In our case, none of the explored methods used this kind of approach, meaning that we had, in some cases, smaller contexts that could be used to predict the named entities. Therefore, based on these aspects, we consider that our approaches could be improved further if the additional context would be provided. Furthermore, we think that they are suitable in cases where training bigger models¹⁰ is not possible or too expensive.

With respect to the results of the Nimeüksuste corpus, presented in Table 26, there are three elements to discuss. It is hard to determine which model is the best for this corpus, either the work proposed by (Ulčar & Robnik-Šikonja, 2020a) or the multi-task baseline proposed in this deliverable. The reason is that (Ulčar & Robnik-Šikonja, 2020a) only indicates the macro F-score, which consider all the types of entities equally important, regardless of their frequency in the corpus. However, as the Nimeüksuste is not a balanced corpus, as seen in Table 5, it is important to present as well the micro F-score to determine whether the NER system is focusing on the less frequent types or not.¹¹ Despite this, there are some factors that could have made the work of (Ulčar & Robnik-Šikonja, 2020a) to be more performing than us. For instance, unlike us, they use a sequence size of 512 and do not add a CRF layer on top of BERT. A larger sequence size means that longer sentences can be used in the model without having to split them, but also, an increased context can be provided to the model. Furthermore, aspects such as the size of the mini-batch or other BERT hyperparameters could have played a big role.

Another point to discuss regarding the results of the NER systems over the Nimeüksuste corpus, see Table 26, is the variability of the F-scores between models using Multilingual BERT. The difference, in terms of micro F-score, can be 0.028 points in some cases when using a sequence size of 128 tokens. This might be due to the fact that BERT, on certain configurations, can produce unstable models (Mosbach et al., 2020; T. Zhang, Wu, Katiyar, Weinberger, & Artzi, 2020). In other words, BERT can generate models that on occasion are outliers due to aspects as the size of the dataset or the number of epochs. In our work, we have tried to prevent this by following the recommendations of (Mosbach et al., 2020) for training the multitask BERT models.

As well, it is interesting to note in Table 26, that all the additional methods explored in the multi-task BERT performed worse than the baseline. We were expecting that at least one of the additional aspects explored in the multi-task BERT would improve the performance with respect to the multi-task baseline.

¹⁰Increasing the size of Bert's sequence size, for supporting more textual information, can produce an increment on BERT's memory usage. See <https://github.com/google-research/bert#out-of-memory-issues>

¹¹We make use of the exact same partitions for testing.

Furthermore, when using Finest and the Stacked BERT, we cannot observe a large difference with respect to the BERT baseline using Finest as well.

Based on the outcomes shown in Section 3.4 and the discussion presented in this section, we can indicate that models based on BERT can outperform other architectures for NER. Furthermore, it is quite simple to generate highly performing NER systems using pre-trained models. However, in order to obtain the best performance, it is necessary to add more elements to BERT. As well, it is necessary to train BERT-based models for a longer number of epochs, as we have done in our experiments, but also as the literature recommends (Mosbach et al., 2020).

3.6 Conclusions

Named entity recognition (NER) is a natural language processing task that aims to extract and classify groups of tokens referring to specific types like locations, persons, and organisations. Although it is a key task for the processing of documents, the creation of NER systems for languages different than English, has been slow, especially for less-resourced languages, as those explored in EMBEDDIA.

Until not so long, most of the NER systems developed for languages such as Croatian, Slovene, or Russian were based on conditional random fields (CRF) classifiers; in some cases, they were based on more advanced architectures like BiLSTM neural networks. However, with the creation of BERT (Devlin et al., 2019) and their multilingual pre-trained models, multiple NER systems have started to use its architecture. The reason is that fine-tuning BERT models can generate NER systems of high quality without having to spend money or time in creating elements such as lexica, thesauri, or rules. Nonetheless, as more NER systems based on BERT have been increased, the scientific community has noticed that a simple fine-tuning of a pre-trained BERT is not enough. For instance, BERT is prone to have a lower performance if the text to analyse contains spelling mistakes, unseen word variants, or even uppercase tokens. Therefore, it is necessary to propose improved BERT architectures in order to achieve the best performance of an NER system.

Therefore, in this deliverable, for the task of NER we presented four different main architectures; two of them based on BiLSTM and two founded on BERT. The results obtained show that BERT is an excellent model to create an NER system for the languages explored in EMBEDDIA. However, there are some aspects that need to be taken into account to get the best performance. For instance, it is necessary to prioritise pre-trained models with fewer languages, and if possible with just one. As well, the process of long sentences is vital for NER systems based on BERT, and an excellent approach is the splitting of long sentences into smaller ones. In addition, we observed that training NER systems along with other tasks might improve the general performance of the NER system. For example, in this deliverable, we explored the prediction of boundaries and masked tokens to improve the performance of the NER systems in multiple languages. We observed as well, that some elements, such as affixes, might be still of relevance in the development of an NER system, however, it is necessary to explore how to include this information in BERT-based systems.

4 Named Entity Linking (NEL)

Named Entity Linking (NEL) aims to map each named entity found in a document to its corresponding named entity in a knowledge base (Shen et al., 2014). NEL approaches can be divided into two classes:

- **Disambiguation approaches** only analyse gold standard named entities in a document and disambiguate them to the correct entry in a given knowledge base (KB).
- **End-to-end approaches** process documents to extract the entities and then disambiguate these extracted entities to the correct entry in a given KB.

Most works in the state of the art are based on three modules: candidate entity generation, candidate entity ranking, and unlinkable mention prediction (Shen et al., 2014). More precisely, the first module aims to retrieve related entity mentions in KB that refer to mention in a document. Several works use name dictionary-based techniques (Guo, Chang, & Kiciman, 2013), surface form expansion from the local document (W. Zhang, Sim, Su, & Tan, 2011), and methods based on search engine (Han & Zhao, 1999).

4.1 Previous Work

Below, we describe the two main approaches used in the literature for dealing with NEL:

Disambiguation approaches. Among the disambiguation-only approaches, the one proposed by (Ganea & Hofmann, 2017) built a deep learning model for joint document-level entity disambiguation. The authors embed entities and words in a common vector space and use a neural attention mechanism to select words that are informative for the disambiguation decision. Then, their model collectively disambiguates the mentions in a document. We describe this approach in further detail in Section 4.2. Motivated by Ganea and Hofmann's approach, (Le & Titov, 2018) analysed relations between mentions as latent variables in their neural NEL model. They rely on representation learning and learn embeddings of mentions, contexts, and relations to reduce the amount of human expertise required to construct the system and make the analysis more portable across domains. (Rosales-Méndez, Hogan, & Poblete, 2020) proposed a fine-grained categorisation scheme for NEL that distinguishes different types of mentions and links. More precisely, they extended five NEL systems with word sense disambiguation and coreference resolution components in order to measure the impact on performance per category.

End-to-end approaches. In the class of end-to-end approaches, (Raiman & Raiman, 2018) developed a system for integrating symbolic knowledge into the reasoning process of a neural network through a type system. They constrained the behavior to respect the desired symbolic structure and automatically designed the type of system without human effort. Their model first uses heuristic search or stochastic optimisation over discrete variables that define a type system informed by an oracle and a learnability heuristic. Based on a joint analysis of the named entity recognition and linking tasks, (Kolitsas, Ganea, & Hofmann, 2018) proposed an end-to-end NEL system that jointly discovers and links entities in a document. They generate all possible spans (mentions) that have at least one possible entity candidate. Then, each mention-candidate pair receives a context-aware compatibility score based on word and entity embeddings (Ganea & Hofmann, 2017) coupled with neural attention and a global voting mechanism. (Broscheit, 2019) proposed a neural model that performs entity linking without any pipeline or any heuristics. Their model analyses the entity linking task as a token classification.

Extending this monolingual analysis, cross-lingual named entity linking (XEL) analyses documents and named entities that are in a different language than those used for the content of the knowledge base.

In this context, (McNamee, Mayfield, Lawrie, Oard, & Doermann, 2011) proposed an XEL approach and examined the importance of transliteration, the utility of cross-language information retrieval, and the potential benefit of multilingual named entity recognition on the XEL task.

(Rijhwani, Xie, Neubig, & Carbonell, 2019) proposed a zero-shot transfer learning method for XEL. Their approach uses phonological representations and a pivot-based method, which leverages information from a high-resource “pivot” language to train character-level neural entity linking models that are transferred to the source low-resource language in a zero-shot manner.

(Zhou, Rijhwani, & Neubig, 2019) extensively evaluated the effect of resource restrictions on existing XEL methods in low-resource settings. They investigated a hybrid candidate generation method, combining existing lookup-based and neural candidate generation methods and proposed a set of entity disambiguation features that are entirely language-agnostic. Finally, they designed a non-linear feature combination method, which makes it possible to combine features in a more flexible way.

As the NEs are recognised by our NER system (Section 3), we focused on disambiguation approaches in this deliverable. In this case, these approaches consider having already identified the named entities in the documents and aim to analyse the context of these entities to disambiguate them in a KB. The following subsections describe our previous work (Section 4.2) and our current multilingual approach (Section 4.3) to disambiguate entities in several languages.

4.2 Cross-lingual Named Entity Linking

In Deliverable D2.2, we described our previous disambiguation method that is based on the approach proposed by (Ganea & Hofmann, 2017) (Figure 7). Our method projects entities and words in a common vector space, which avoids hand-engineered features, multiple disambiguation steps, or the need for additional ad-hoc heuristics when solving the disambiguation task (Linhares Pontes, Doucet, & Moreno, 2020). Entities for each mention are locally scored based on cosine similarity with the respective document embedding. Combined with these embeddings, an attention mechanism over local context windows selects words that are informative for the disambiguation decision. The final local scores are based on the combination of the resulting context-based entity scores and a mention-entity prior. Finally, mentions in a document are resolved jointly by using a conditional random field in conjunction with an inference scheme.

We proposed a cross-lingual extension of this method that can be easily adapted to any source language. We use multilingual word embeddings and a fine-tuning method to represent words and entities in multiple languages into the same dimensional space, and then to disambiguate mentions across languages.

Unfortunately, our cross-lingual approach contains some limitations: the small vocabulary of multilingual word embeddings and the English probability table contains a limited number of mentions in less well-resourced languages. In order to improve the analysis of less-resourced languages, we proposed a multilingual end-to-end approach to analyse and disambiguate NEs in these languages.

This work can be found in Appendix 10.

4.3 Multilingual End-to-end Entity Linking

In the following subsections, we describe our method for dealing with NEL in several less-resourced languages.

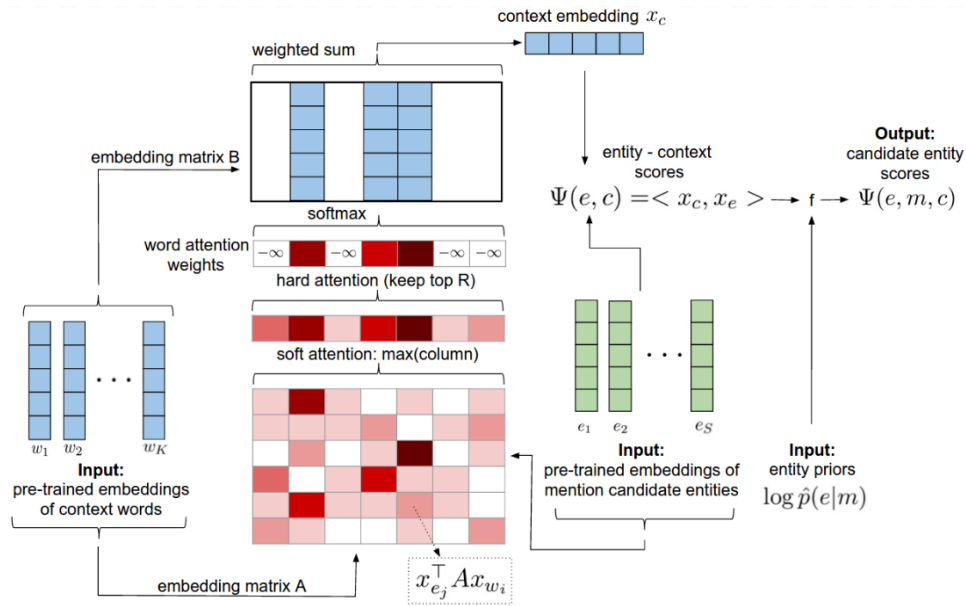


Figure 7: Local model with neural attention. Inputs: context word vectors, candidate entity priors and embeddings. Outputs: entity scores. All parts are differentiable and trainable with back-propagation Ganea et al., 2017.

4.3.1 Building Resources

By definition of the task, NEL systems use knowledge bases (KB) as entry reference but their use is not limited to it. KBs are also used by NEL systems for tasks such as extraction of supplementary contexts or surface names, disambiguation of cases, or linking of entities with a particular website entry.

In this work, we decided to build our own KB consisting of information from Wikipedia. Nevertheless, rather than just focusing on the English Wikipedia, we make use, as well, of the versions for each of EMBEDIA languages: Estonian, Finnish, Swedish, Latvian, Lithuanian, Croatian, Slovene, and Russian. The reasoning behind this is that despite the richness and coverage of the English Wikipedia, on occasion other versions of Wikipedia might contain information that is only found in a specific language. Table 27 lists the number of all Wikipedia pages for all EMBEDIA languages. Moreover, this table also provides the number of Wikipedia pages that are not present in the English version of Wikipedia. As expected, the analysis of target language versions of Wikipedia added a considerable number of entities for each language that not exists in the English Wikipedia.

Table 27: Knowledge base statistics. k and M represent thousands and millions respectively.

Number of Entities	hr	et	fi	lv	lt	ru	sl	sv
All	216k	236k	556k	135k	226k	2.1M	229k	4M
∉ English Wikipedia	71k	91k	166k	41k	93k	1M	68k	3M

4.3.2 Entity Embeddings

Based on the work of (Ganea & Hofmann, 2017), we decided to create entity embeddings for each language by generating two conditional probability distributions. The first one, the positive distribution, is a probability approximation based on word-entity co-occurrence counts, i.e. which words appear

in the context of an entity. The counts were obtained, in first place, from the entity Wikipedia page, and, in second place, from the context surrounding the entity in an annotated corpus using a fixed-length window. The second distribution, the negative one, was calculated by randomly sampling context windows that were unrelated to a specific entity. Both probability distributions were used to change the alignment of word embeddings with respect to an entity embedding. The positive probability distribution is expected to approach the embeddings of the co-occurring words with the embedding vector of the entity, while the negative probability distribution is used to distance the embeddings of words that are not related to an entity. Figure 8 shows an example of entity embeddings representations, e.g., red circles represent the EMBEDDIA partner countries.



Figure 8: Visualisation of entity embeddings in Finnish and Swedish. Only 10,000 entities from the intersection between Swedish and Finnish entity embeddings are visualised.

It should be noted that, unlike some works, where all the possible entities are known beforehand, in our work, the creation of entity embeddings is not directed by a dataset. This is done to prevent bias and low generalisation. In case an entity does not have entity embeddings, the NEL system will propose a NIL¹². Table 28 lists the number of entity embeddings for each language.

Table 28: Entity embeddings statistics.

	hr	et	fi	lv	lt	ru	sl	sv
Number of entities	91k	111k	258k	53k	93k	726k	66k	550k

4.3.3 Entity Disambiguation

The entity disambiguation model is based on the neural end-to-end entity linking architecture proposed by (Kolitsas et al., 2018) (Figure 9). The first advantage of this architecture is that it performs both entity linking and disambiguation. This method can then benefit from simplicity and lack of error propagation. Furthermore, this architecture does not require complex feature engineering, which makes it easily adaptable to other languages.

For recognising all entity mentions in a document, Kolitsas *et al.* utilised an empirical probabilistic table entity–map, defined by $p(e|m)$. Where p is the probability of an entity e to be related to a mention m ; $p(e|m)$ is calculated using the number of times that mention m refers e within Wikipedia. From this probabilistic table, it is possible to find which are the top entities that a mention span refers to.

¹²NEL systems provide a NIL entry to indicate that a mention does not have a ground-truth entity in the KB.

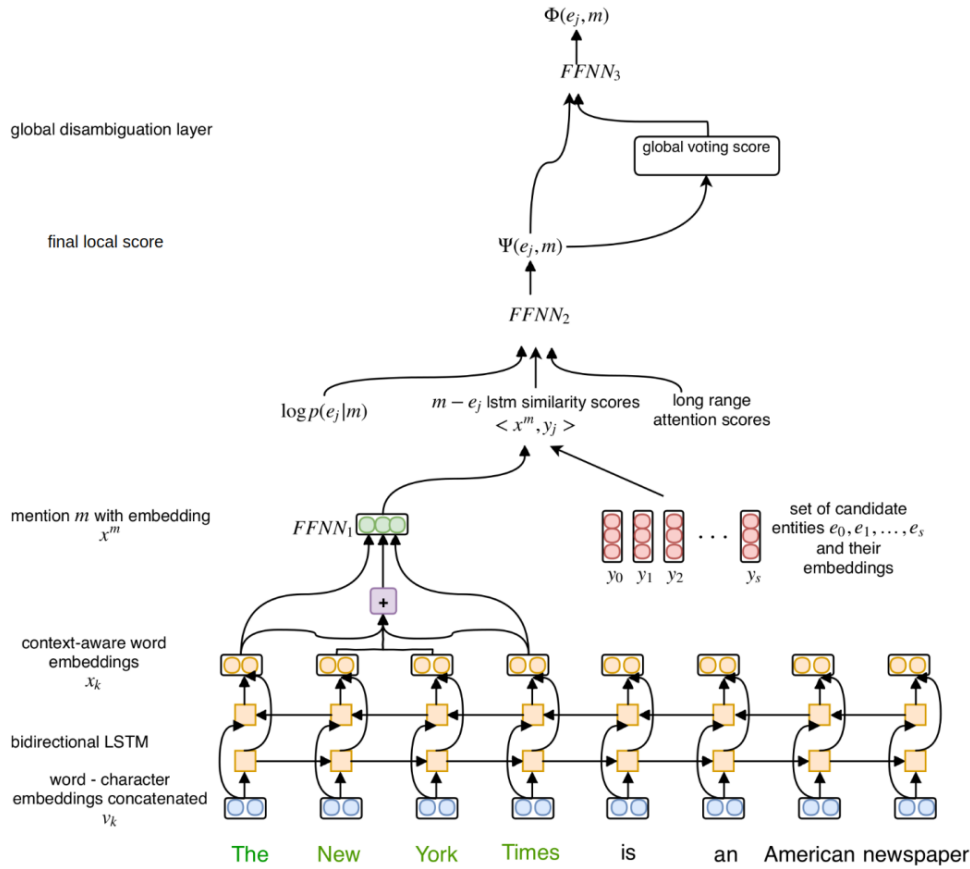


Figure 9: Global model architecture shown for the mention The New York Times. The final score is used for both the mention linking and entity disambiguation decisions Kolitsas et al., 2018.

The end-to-end NEL model starts by encoding every token in the text input by concatenating word and character embeddings and fed into a Bidirectional Long Short Term Memory (BiLSTM) (Hochreiter & Schmidhuber, 1997) network. This representation is used to project mentions of this document into a shared dimensional space with the same size as the entity embeddings. These embeddings are fixed continuous entity representations generated separately, namely in the same manner as presented in (Ganea & Hofmann, 2017), and aforementioned in Subsection 4.3.2. In order to analyse long context dependencies of mentions, the authors utilised the attention mechanism proposed by (Ganea & Hofmann, 2017). This mechanism provides one context embedding per mention based on surrounding context words that are related to at least one of the candidate entities.

The final local score for each mention is determined by the combination of the $\log p(e|m)$, the similarity between the analysed mention and the candidate entity, and the long-range context attention for this mention. Finally, a top layer in the neural network promotes the coherence among disambiguated entities inside the same document.

4.3.4 Match Corrections

Multiple NEL approaches, including the one used in this work, rely on the matching of entities and candidates using a probability table. If an entity is not listed in the probability table, the NEL system cannot disambiguate it and, therefore, it cannot propose candidates.

To increase the matching of entities in the probability table, we propose an analysis that consists of

exploring several surface name variations using multiple heuristics. For instance, we evaluate variations by lower and uppercasing, capitalising words, concatenating surrounding words, removing stopwords, and transliterating special characters, like accentuated letters, to Latin characters. If after applying the previous heuristics, a match is still lacking, we use the Levenshtein distance to overcome more complex cases, such as spelling mistakes.

4.3.5 Multilingualism

News documents may contain words and phrases in a language different from that of the document under analysis. To overcome this problem, we combined the probability tables of several languages in order to identify the surface names of entities in multiple languages.

This work can be found as well in Appendix 5 and Appendix 14.

4.4 Datasets

Wikipedia is a multilingual resource that currently hosts 294 languages and contains annotated markups and rich informational structures through crowd-sourcing. In this resource, name mentions are often labelled as anchor links to their corresponding referent pages (Pan et al., 2017). (Pan et al., 2017) developed an independent language framework to automatically extract name mentions from Wikipedia articles in 282 languages and link them to the English Wikipedia (WikiANN dataset). It is important to note that this dataset is automatically built and that it contains all the types of named entities used in EMBEDDIA.

We used the WikiANN on Croatian, Estonian, Finnish, Latvian, Lithuanian, Slovenian, Russian, and Swedish. We also converted the links of the English Wikipedia of these datasets for their IDs in the Wikidata KB (Table 29). WikiANN datasets have different numbers of available entities for each language version of the Wikipedia KB. Indeed, some entities presented in the English version of the Wikipedia KB do not have a corresponding entity in the Wikidata. When an entity does not exist in a KB, we replace its link with a NIL entry (Sil, Kundu, Florian, & Hamza, 2018).

Table 29: Number of mentions on WikiANN datasets.

Languages	hr	et	fi	lv	lt	ru	sl	sv
train	47,288	44,921	211,546	32,751	45,855	843,312	34,140	1,233,758
dev	11,670	8,137	44,852	7,616	10,117	189,255	14,670	267,405
test	10,113	9,815	47,555	10,476	10,250	200,198	13,840	242,789

The WikiANN data set was split into three separate data sets, 70% of the corpus for training, 15% for dev, and 15% for testing. For the training process, we use the training split to train the NEL system for each EMBEDDIA language.

4.5 Experimental Setup

NEL aims to connect named entities to external knowledge bases. In order to accomplish this task, we first need to recognise these entities in the documents and, then, disambiguate them to a KB. In this deliverable, we analyse the disambiguation approaches that only analyse gold standard named entities in a document and disambiguate them to the correct entries in a given KB, i.e. NEL systems know the offset of all mentions in the documents.

For the Ganea and Hofmann's (GH) approach (Ganea & Hofmann, 2017), we followed a similar procedure described in our previous work (Linhares Pontes et al., 2020). More precisely, we used the pre-trained multilingual MUSE word vectors with 300 dimensions¹³ to train entity embeddings on the Wikipedia (Feb 2014) corpus. Then, we trained their entity disambiguation approach on AIDA training dataset. Finally, we disambiguate NEs on target languages using the word embeddings on their corresponding languages.

For our multilingual NEL approach, we used the pre-trained FastText words embeddings (Bojanowski et al., 2017) with 300 dimensions¹⁴ to train entity embeddings for all EMBEDDIA languages on the Wikipedia (Jan 2020) corpus. Then, we trained the Kolitsas et al.'s approach (Kolitsas et al., 2018) on WikiANN training datasets for each language.

4.5.1 Evaluation Metrics

We used the F1-measure described in Deliverable D2.1 to evaluate the NEL performance. Since knowledge bases contain millions of entities, only mentions that contain a valid ground-truth entry in the KB are analysed. For mentions without corresponding entries in the KB, NEL systems have to provide a NIL entry to indicate that these mentions do not have a ground-truth entity in the KB.

4.5.2 Cross-lingual and Multilingual NEL Experiments

Table 30 compares the NEL performance between our cross-lingual and multilingual approaches on the test WikiANN corpora for all EMBEDDIA languages. Our multilingual approach achieved the best results for almost all languages (except Slovenian). Unfortunately, we did not have the performance of our cross-lingual method for Latvian, Lithuanian, and Russian.

Table 30: NEL system performance on the test WikiANN corpora.

Systems		hr	et	fi	lv	lt	ru	sl	sv	Avg.
Cross-lingual NEL (Ganea and Hofmann)	Precision	87.7	88.4	91.2	–	–	–	95.4	56.5	83.8
	Recall	42.5	32.1	51.7	–	–	–	56.1	39.8	44.4
	F-measure	57.2	47.1	66.0	–	–	–	70.6	46.7	57.2
Multilingual NEL	Precision	93.5	94.0	95.5	94.6	93.0	82.4	94.0	98.9	95.2
	Recall	74.6	65.0	76.4	46.5	65.1	37.6	56.4	88.1	72.1
	F-measure	83.0	76.9	84.9	62.3	76.6	51.6	70.5	93.2	81.7

The WikiANN data set is composed of short sentences with little context information. This characteristic makes the context analysis of NEL systems being less relevant and making the disambiguation process be decided mainly by the pairwise matching between mentions and entities on the $\log p(m|e)$. Both systems achieved high precision for most languages; however, the baseline had poor recall results. The main reason for the poor results of the GH's approach is the poor quality of the probability table $p(e|m)$ generated from the English Wikipedia. Indeed, the probability table extracted from the English version of Wikipedia contains only a limited number of mentions in less well-resourced languages. Another limiting factor is the small MUSE vocabulary. Indeed, the MUSE vocabulary is composed of only 200,000 words, while the FastText vocabulary is composed of two million words.

¹³<https://github.com/facebookresearch/MUSE>

¹⁴<https://fasttext.cc/docs/en/crawl-vectors.html>

4.6 Discussion

The main reasons for the improvements of our multilingual approach are the new probability tables for each language. The probability tables $p(e|m)$ of the EMBEDDIA language versions of Wikipedia provided more information about persons, organisations, and locations for these languages. Indeed, these tables contain a larger number of entities and their surface names than the table generated by the baseline. These tables helped the disambiguation method to find the entities that are more related to a mention. Moreover, our multilingual approach disambiguates NEs to Wikidata that is a common KB for several languages.

Indeed, our multilingual model achieved better results because it contains specific models and probability tables for each language which improved the candidate generation and representations of entities. For instance, the English and the Finnish Wikipedia pages with the title “Paris” do not describe the same entity; in Finnish “Paris” makes reference to Greek mythology while the French capital is known as “Pariisi”. In the English Wikipedia, the page “Pariisi” makes reference to a village in Kadrina Parish, Lääne-Viru County, in northeastern Estonia. In this case, the cross-lingual model will propose the wrong candidates for the mention “Pariisi” which will reduce the precision score. The English probability table does not contain several entities and their surface names for Finnish entities which drastically dropped the recall scores of the cross-lingual system. To illustrate the difference between the probability tables, the English probability contains only the surface name for the entity “London” in English, while the Finnish probability table contains the English (“London”) and Finnish (“Lontoo”) surface names for this entity.

4.7 Conclusions

Named Entity Linking (NEL) is the task of recognising and disambiguating named entities by linking them to entries of a knowledge base. This task can solve problems related to duplicate and ambiguous information about named entities in different contexts and languages. Moreover, NEL is a relevant task to several NLP applications, e.g. information extraction, information retrieval, content analysis, question answering, and knowledge base population.

We evaluated our multilingual and cross-lingual models on the WikiANN corpora for all EMBEDDIA languages. For almost all languages, our multilingual approach outperformed our previous cross-lingual method. As expected, using specific language versions of Wikipedia provided more relevant information, such as surface names and context information. Our multilingual probability tables $p(e|m)$, training using training data and word embeddings on the target language improved the overall performance of our approach.

5 Event Detection (ED)

Event extraction (EE) is an application of information extraction (IE) that implies the extraction of specific knowledge from certain incidents from texts. This task is focused on obtaining event-related information from texts, and, in this section, we describe three datasets with different annotation styles, and the proposed methods for exploring them.

In this derivable, we focus on event detection and we analyse two different ways of events annotations. Over the years, several event definitions have been proposed, each showing specific strengths and weaknesses. Thus, the event detection task is challenging due to the ambiguous nature of the concept of event.

5.1 Definitions

<p>TST1-MUC3-0080</p> <p>BOGOTA, 3 APR 90 (INRAVISION TELEVISION CADENA 1) -- [REPORT] [JORGE ALONSO SIERRA VALENCIA] [TEXT] LIBERAL SENATOR FEDERICO ESTRADA VELEZ WAS KIDNAPPED ON 3 APRIL AT THE CORNER OF 60TH AND 48TH STREETS IN WESTERN MEDELLIN, ONLY 100 METERS FROM A METROPOLITAN POLICE CAI [IMMEDIATE ATTENTION CENTER]. THE ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER HAD LEFT HIS HOUSE WITHOUT ANY BODYGUARDS ONLY MINUTES EARLIER. AS HE WAITED FOR THE TRAFFIC LIGHT TO CHANGE, THREE HEAVILY ARMED MEN FORCED HIM TO GET OUT OF HIS CAR AND GET INTO A BLUE RENAULT.</p> <p>HOURS LATER, THROUGH ANONYMOUS TELEPHONE CALLS TO THE METROPOLITAN POLICE AND TO THE MEDIA, THE EXTRADITABLES CLAIMED RESPONSIBILITY FOR THE KIDNAPPING. IN THE CALLS, THEY ANNOUNCED THAT THEY WILL RELEASE THE SENATOR WITH A NEW MESSAGE FOR THE NATIONAL GOVERNMENT.</p> <p>LAST WEEK, FEDERICO ESTRADA VELEZ HAD REJECTED TALKS BETWEEN THE GOVERNMENT AND THE DRUG TRAFFICKERS.</p>	<table border="1"> <tr><td>0. MESSAGE ID</td><td>TST1-MUC3-0080</td></tr> <tr><td>1. TEMPLATE ID</td><td>1</td></tr> <tr><td>2. DATE OF INCIDENT</td><td>03 APR 90</td></tr> <tr><td>3. TYPE OF INCIDENT</td><td>KIDNAPPING</td></tr> <tr><td>4. CATEGORY OF INCIDENT</td><td>TERRORIST ACT</td></tr> <tr><td>5. PERPETRATOR: ID OF INDIV(S)</td><td>"THREE HEAVILY ARMED MEN"</td></tr> <tr><td>6. PERPETRATOR: ID OF ORG(S)</td><td>"THE EXTRADITABLES" / "EXTRADITABLES"</td></tr> <tr><td>7. PERPETRATOR: CONFIDENCE</td><td>CLAIMED OR ADMITTED: "THE EXTRADITABLES" / "EXTRADITABLES"</td></tr> <tr><td>8. PHYSICAL TARGET: ID(S)</td><td>*</td></tr> <tr><td>9. PHYSICAL TARGET: TOTAL NUM</td><td>*</td></tr> <tr><td>10. PHYSICAL TARGET: TYPE(S)</td><td>"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR" / "ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER" / "SENATOR" / "LIBERAL PARTY LEADER" / "PARTY LEADER")</td></tr> <tr><td>11. HUMAN TARGET: ID(S)</td><td>1</td></tr> <tr><td>12. HUMAN TARGET: TOTAL NUM</td><td>GOVERNMENT OFFICIAL / POLITICAL FIGURE: "FEDERICO ESTRADA VELEZ"</td></tr> <tr><td>13. HUMAN TARGET: TYPE(S)</td><td>*</td></tr> <tr><td>14. TARGET: FOREIGN NATION(S)</td><td>*</td></tr> <tr><td>15. INSTRUMENT: TYPE(S)</td><td>COLOMBIA: MEDELLIN (CITY)</td></tr> <tr><td>16. LOCATION OF INCIDENT</td><td>*</td></tr> <tr><td>17. EFFECT ON PHYSICAL TARGET(S)</td><td>*</td></tr> <tr><td>18. EFFECT ON HUMAN TARGET(S)</td><td>*</td></tr> </table>	0. MESSAGE ID	TST1-MUC3-0080	1. TEMPLATE ID	1	2. DATE OF INCIDENT	03 APR 90	3. TYPE OF INCIDENT	KIDNAPPING	4. CATEGORY OF INCIDENT	TERRORIST ACT	5. PERPETRATOR: ID OF INDIV(S)	"THREE HEAVILY ARMED MEN"	6. PERPETRATOR: ID OF ORG(S)	"THE EXTRADITABLES" / "EXTRADITABLES"	7. PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: "THE EXTRADITABLES" / "EXTRADITABLES"	8. PHYSICAL TARGET: ID(S)	*	9. PHYSICAL TARGET: TOTAL NUM	*	10. PHYSICAL TARGET: TYPE(S)	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR" / "ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER" / "SENATOR" / "LIBERAL PARTY LEADER" / "PARTY LEADER")	11. HUMAN TARGET: ID(S)	1	12. HUMAN TARGET: TOTAL NUM	GOVERNMENT OFFICIAL / POLITICAL FIGURE: "FEDERICO ESTRADA VELEZ"	13. HUMAN TARGET: TYPE(S)	*	14. TARGET: FOREIGN NATION(S)	*	15. INSTRUMENT: TYPE(S)	COLOMBIA: MEDELLIN (CITY)	16. LOCATION OF INCIDENT	*	17. EFFECT ON PHYSICAL TARGET(S)	*	18. EFFECT ON HUMAN TARGET(S)	*
0. MESSAGE ID	TST1-MUC3-0080																																						
1. TEMPLATE ID	1																																						
2. DATE OF INCIDENT	03 APR 90																																						
3. TYPE OF INCIDENT	KIDNAPPING																																						
4. CATEGORY OF INCIDENT	TERRORIST ACT																																						
5. PERPETRATOR: ID OF INDIV(S)	"THREE HEAVILY ARMED MEN"																																						
6. PERPETRATOR: ID OF ORG(S)	"THE EXTRADITABLES" / "EXTRADITABLES"																																						
7. PERPETRATOR: CONFIDENCE	CLAIMED OR ADMITTED: "THE EXTRADITABLES" / "EXTRADITABLES"																																						
8. PHYSICAL TARGET: ID(S)	*																																						
9. PHYSICAL TARGET: TOTAL NUM	*																																						
10. PHYSICAL TARGET: TYPE(S)	"FEDERICO ESTRADA VELEZ" ("LIBERAL SENATOR" / "ANTIOQUIA DEPARTMENT LIBERAL PARTY LEADER" / "SENATOR" / "LIBERAL PARTY LEADER" / "PARTY LEADER")																																						
11. HUMAN TARGET: ID(S)	1																																						
12. HUMAN TARGET: TOTAL NUM	GOVERNMENT OFFICIAL / POLITICAL FIGURE: "FEDERICO ESTRADA VELEZ"																																						
13. HUMAN TARGET: TYPE(S)	*																																						
14. TARGET: FOREIGN NATION(S)	*																																						
15. INSTRUMENT: TYPE(S)	COLOMBIA: MEDELLIN (CITY)																																						
16. LOCATION OF INCIDENT	*																																						
17. EFFECT ON PHYSICAL TARGET(S)	*																																						
18. EFFECT ON HUMAN TARGET(S)	*																																						

Figure 10: Example of MUC-3 *template*

In the first campaign related to event extraction, the message understanding conference (MUC-3) in 1991, the task of EE was seen as a *template* with slots to be automatically filled with participants, time and space details. The articles in the MUC dataset focused on events about terrorist attacks and violent acts with political aims and a motive of intimidation. An article could contain multiple events, from a pre-set list of event types e.g. *bombing*, *attack*, *kidnapping* with multiple arguments, e.g. *human target*, *perpetrator*. The task of event extraction was defined as the extraction of *templates* as shown in the Figure 10 (Chinchor, Lewis, & Hirschman, 1993), where the incident is *kidnapping*, from the incident category *terrorist attack*, with different human targets (e.g. *Federico Estrada Velez*), the date (*03 April 90*) and the location (*Colombia*) of the incident. The MUC campaigns (Grishman & Sundheim, 1996) lasted from 1987 through 1998 under the auspices of the US government (ARPA/DARPA).

While the MUC definition of an event consisted in the extraction of the type of event and the event participants, without making a difference between these two tasks, the automatic content extraction (ACE) competitions were rather different. The event extraction task was defined as two separate sub-tasks: event detection, that implies the identification of instances of specified types of events in text, and event argument extraction, which is the extraction of the arguments associated to them. In the event detection sub-task, each event is represented by a phrase or a sentence, the *event trigger* (most often single verbs or phrasal verbs, but also nouns, phrasal nouns, pronouns and adverbs), which evokes that event. An example is provided in Figure 11. After the detection and classification of the triggers, in the second sub-task, the arguments of the event must be detected and correctly classified in a pre-set list of argument roles. Event arguments are entity mentions or temporal expressions that are involved in an event (as participants) with specific roles for all the event types.

Typically, an ACE event in a text is expressed by the following components:

```

<event ID="APW_ENG_20030520.0757-EV8" TYPE="Conflict" SUBTYPE="Attack" MODALITY="Other"
  <event_argument REFID="APW_ENG_20030520.0757-E18" ROLE="Attacker"/>
  <event_argument REFID="APW_ENG_20030520.0757-E9" ROLE="Place"/>
  <event_mention ID="APW_ENG_20030520.0757-EV8-1">
    <extent>
      <charseq START="1392" END="1477">Osama bin Laden's Al-Qaeda group possibly
launching fresh attacks in the United States</charseq>
    </extent>
    <ldc_scope>
      <charseq START="1305" END="1516">Earlier this week, Saudi and U.S. officials said they had new
intelligence pointing to Osama bin Laden's Al-Qaeda group possibly
launching fresh attacks in the United States or against American
interests overseas</charseq>
    </ldc_scope>
    <anchor>
      <charseq START="1450" END="1456">attacks</charseq>
    </anchor>
    <event_mention_argument REFID="APW_ENG_20030520.0757-E18-42" ROLE="Attacker">
      <extent>
        <charseq START="1392" END="1423">Osama bin Laden's Al-Qaeda group</charseq>
      </extent>
    </event_mention_argument>
    <event_mention_argument REFID="APW_ENG_20030520.0757-E9-44" ROLE="Place">
      <extent>
        <charseq START="1461" END="1477">the United States</charseq>
      </extent>
    </event_mention_argument>
  </event_mention>
</event>

```

Figure 11: Example of ACE 2005 event annotation

- **Event mention:** an occurrence of an event with a particular type. These are usually sentences or phrases that describe an event. The example in the Figure 11 is an **Attack** event mention: the text talks about an attack.
- **Event trigger:** the word that most clearly expresses the event mention. The **Attack** from the figure is revealed by the event trigger word **attacks**.
- **Event argument:** an entity mention or temporal expression (e.g. *Crime*, *Job-Title*) that serves as a participant or attribute with a specific role in an event mention.
- **Argument role:** the relationship between an argument and the event in which it participates. The argument roles that should be extracted in this case are: *Osama bin Laden's Al-Qaeda group* that has the role of an **Attacker** and the **Place** where the event produced is *the United States*. The **Attacker** is an argument role that are specific for the **Conflict.Attack** event type.

The *event mention* and *event trigger* are notions used in ED, and the *event argument* and *argument role* are notions used in the event arguments extraction.

Shortly after, the definition of the event has undergone minor changes, and the ERE (entities, relations, events) scheme has been developed later within the DARPA DEFT program (Aguilar et al., 2014) in order to simplify the ACE event type definition that made the process of annotating data very challenging. ERE and ACE share the same event types and subtypes, but the ERE annotation is simplified by collapsing tags and therefore loosening the event extent and also reducing the annotation features in order to improve coherency and consistency of the dataset.

Throughout the years, MUC, ACE, and TAC initiatives have been of central importance to the IE field since they provided a set of corpora that are available to the research community for the evaluation and comparison of IE systems and approaches.

In the previous deliverable, the notion of *event* was defined as a named entity type, in the context of Balto-Slavic natural language processing (BSNLP 2019) dataset. While we also continued experimenting with this event type definition, in this deliverable, we also experiment with two more commonly used ways of annotating events in literature. Thus, events can be defined:

- At sentence level:
 - in BSNLP 2019, an event is represented as a named entity with the tag EVT;

- As for example, in ACE 2005, several events can be present in a sentence, and we propose different models for detecting the event trigger;
- At document level:
 - We also consider a system and a dataset related to epidemiological events proposed by (Lejeune, Brixstel, Doucet, & Lucas, 2015) where an event of type (disease-location) can be present in an article.

5.2 Challenges

Despite the usefulness and prospective applicability of EE (which implies the ED sub-task), several issues and challenges are to be overcome until an IE system is widely adopted as an effective tool in practice.

- The **annotation cost**, **data scarcity**, and **lack of resources for less-resourced languages**: there are practical issues related to the high cost of manual annotation of texts (e.g. human resources). The human effort needs to be minimised while keeping the quality of an IE system. Data annotation takes advantage of a massive human expertise and this causes labour-intensive work for data interpretation at two levels. Firstly, an IE system may use NLP resources and tools, created using lots of annotated documents and secondly, an IE system needs a higher-level of annotation of relations or events, annotations that can be complex and extremely costly. For this reason, there is a lack of annotated datasets for less-represented languages or languages lacking large monolingual or parallel datasets and manually crafted linguistic resources sufficient for building IE applications. One major drawback when working with data in these languages is that the previously developed tools in languages that have wider monolingual coverage are not directly applicable to other languages, requiring new corpora for every language added.
- The difficult choice of features, also known as **feature engineering**. Features that come from NLP tools and resources (i.e., dependency parsers, part of speech taggers etc.) and the hard decision making in combining them is considered an important issue due to the error propagation issues. The errors from these sources can propagate to the downstream tasks, e.g. an NER system may mistakenly detect the wrong entity needed by a relation extraction system, which downgrades considerably its accuracy.
- The **context of extraction** can be also considered an issue, since the extraction of the needed information is often approached at a local level, as in the case of the detection and extraction of entities, relations or events that are fully expressed within a single sentence. Sentence-level extraction patterns are commonly used in IE systems, but an event can benefit from the global structure of news.

5.3 Previous Work

In order to better generalise the systems developed for the event detection task, one can divide the prior work in: pattern-based systems (Riloff, 1996a, 1996b), machine learning systems based on engineered features (i.e. feature-based) (Liao & Grishman, 2010; Hong et al., 2011), neural-based approaches (Nguyen & Grishman, 2015; Nguyen, Cho, & Grishman, 2016). Recently, there has been a lot of interest in approaching the ED task with external resource-based models which are either feature-based (S. Liu et al., 2016; W. Li, Cheng, He, Wang, & Jin, 2019) or neural-based (S. Liu, Chen, Liu, & Zhao, 2017) combined with resources as in FrameNet¹⁵ (Baker, Fillmore, & Lowe, 1998), or event data generation as

¹⁵FrameNet is a linguistic corpus that defines complete semantic frames and frame-to-frame relations. <https://framenet.icsi.berkeley.edu/fndrupal/>

in (T. Zhang, Ji, & Sil, 2019; Hong, Zhou, Jingli, Zhou, & Zhu, 2018). The neural-based methods hold the state of the art, and the most recent approaches are based on pre-trained Transformer-based language models (Wadden, Wennberg, Luan, & Hajishirzi, 2019; Du & Cardie, 2020; J. Liu, Chen, Liu, Bi, & Liu, 2020).

Pattern-based approaches. Several pattern-based (rule-based) systems have been proposed to speed up the annotation process. The pattern-based approaches first acquire a set of patterns, where the patterns consist of a predicate, an event trigger, and constraints on its local syntactic context. They also include a rich set of ad-hoc lexical features (e.g. compound words, lemma, synonyms, Part-of-Speech (POS) tags), syntactic features (e.g. grammar-level features, dependency paths) and semantic features (e.g., features from a multitude of sources, WordNet¹⁶, gazetteers) to identify role fillers. Earlier pattern-based extraction systems were developed for the MUC conferences (Krupka, Jacobs, Rau, & Iwańska, 1991; Hobbs, Appelt, Tyson, Bear, & Israel, 1992; Riloff, 1996a; Yangarber, Grishman, Tapanainen, & Huttunen, 2000). Many proposed approaches targeted the minimisation of human supervision with a bootstrapping technique for event extraction. The authors of (Huang & Riloff, 2012) proposed a bootstrapping method to extract event arguments using only a small amount of annotated data. After the manual inspection of the patterns, another effort was made for performing manual filtering of resulting irrelevant patterns.

Feature-based approaches. Most recent event extraction frameworks are feature-based approaches applied at the sentence-level or to a larger context (e.g. document-level). Feature-based approaches rely on discriminative features to build statistical models and usually require effort to develop rich sets of features. The feature-based approaches rely mainly on designing large effective feature sets for statistical models, ranging from *local features* (Grishman, Westbrook, & Meyers, 2005; Ahn, 2006; P. Li, Zhu, & Zhou, 2013) to the higher-level structures such as cross-document, cross-sentence and cross-event information e.g. *global features* (Gupta & Sarawagi, 2009; Hong et al., 2011; Ji, Grishman, et al., 2008; J. Li, Luong, & Jurafsky, 2015; Liao & Grishman, 2010; Patwardhan & Riloff, 2009). The discrete local features include: lexical features (e.g. lemma, Part-of-Speech (POS) tags, Brown clusters (Brown, Desouza, Mercer, Pietra, & Lai, 1992)), syntactic features (e.g. dependency paths) and semantic features (e.g., features from a set of sources, WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), gazetteers).

Neural-based approaches. The current state-of-the-art systems for ED involve neural network models. The approaches presented in (Nguyen & Grishman, 2015) and (Chen, Xu, Liu, Zeng, & Zhao, 2015) deal with this task with a model based on CNNs. The CNN models in (Nguyen & Grishman, 2016) improve the previous models (Nguyen & Grishman, 2015) for ED by taking into account the possibility to have non-consecutive n -grams as basic features instead of continuous n -grams.

The system proposed by (Jagannatha & Yu, 2016) extracts event instances from health records with bidirectional recurrent neural networks (Bi-RNNs) while (Nguyen, Cho, & Grishman, 2016) proposes a joint framework with the same type of neural networks for predicting at the same time event triggers and their arguments. Additionally, the model presented in (Nguyen, Cho, & Grishman, 2016) is augmented with discrete local features inherited from (Q. Li, Ji, & Huang, 2013). The authors of (Nguyen & Grishman, 2018) advocate a graph convolution network (GCN) based on dependency trees for exploiting syntactic dependency relations. The papers (Duan, He, & Zhao, 2017) and (Zhao, Jin, Wang, & Cheng, 2018) explore another extension of RNNs by integrating a larger context through a document representation, while (Hong et al., 2018) exploits a generative adversarial network for discarding spurious detections.

¹⁶<https://wordnet.princeton.edu/>

A hybrid neural network (a CNN and an RNN) is developed in (Feng et al., 2016) in order to capture both sequence and chunk information from specific contexts, and use them to train an event detector for multiple languages without any handcrafted features.

External resource-based approaches. Neural-based approaches achieve relatively high performance due to their ability of learning automatic features. However, as we mentioned before, data scarcity in ED limits their further performance. An external resource-based model tackles data scarcity problems by exploiting additional information. The authors of (Bronstein, Dagan, Li, Ji, & Frank, 2015) take the example trigger terms mentioned in the guidelines as seeds, and then applies an event-independent similarity-based classifier for trigger labelling. Thus, a great amount of effort has been put in to overcome the manual annotation of data. The model described in (W. Li et al., 2019) also leverages FrameNet by tackling the challenge of the annotation cost and data scarcity by redefining event schemas based on FrameNet.

Transformer-based approaches. Recently, several approaches for the event detection task that include contextual sub-word representations have been proposed, based generally on BERT. The approach attempted by (Yang, Feng, Qiao, Kan, & Li, 2019) is based on the BERT model with an automatic generation of labeled data by editing prototypes and filtering out the labeled samples through argument replacement by ranking their quality. A similar framework was proposed by (Wang, Han, Liu, Sun, & Li, 2019) where the informative features are encoded by BERT or a CNN, which would suggest a growing interest not only in language model-based approaches, but also in adversarial models. Simultaneously, an integration of a distillation technique to enhance the adversarial prediction was explored in (Lu, Lin, Han, & Sun, 2019). A recent work proposed by (Du & Cardie, 2020) introduced this new paradigm for event extraction by formulating it as a question answering (QA) task, which extracted the event arguments in an end-to-end manner. Another recent paper (J. Liu et al., 2020) also approaches the event extraction task as a question answering task, similar to the (Du & Cardie, 2020) method.

5.4 Datasets

We consider three datasets, two that contain documents in less-resourced languages and one only in English.

5.4.1 BSNLP 2019 Dataset

This dataset was used in the previous deliverable and it consists in four Slavic languages: Bulgarian, Czech, Polish and Russian. As presented in (Tsygankova et al., 2019), there is a large imbalance in the amount of training data by language, with the largest (Polish), containing almost three times as many tokens as the smallest (Russian). The training data is in the form of newswire articles and contains document-level annotations of five different entity types: persons (PER), locations (LOC), organisations (ORG), events (EVT) and products (PRO). The training documents are divided into two topics: one set containing news articles relating to Brexit, and the other with news articles about a Pakistani woman named Asia Bibi. These focused domains suggest that the set of unique entities will be relatively small within each topic. The statistics of the dataset reported in (Tsygankova et al., 2019) are illustrated in Tables 31 and 32.

In our case, we are interested in the entities of type EVT. The high ratio of total to unique mentions for certain tags such as event (EVT) means that the training data contains a small variety of distinct surface forms labeled as EVT, which could lead to potential overfitting to these entities. Given that the test set used for evaluation of our models contains news articles surrounding two distinct topics

Table 31: Training data sizes in BSNLP 2019 dataset. Of the BSNLP 2019 sets, the largest (Polish) is nearly three times the size of the smallest (Russian).

Languages	Docs	Tokens
Bulgarian	699	226,728
Czech	373	84,636
Polish	586	237,333
Russian	271	67,495

Table 32: Entity distribution statistics across all languages in the BSNLP 2019 training set, where the Ratio column refers to the proportion of the Total number of entity type annotations to the Unique annotations.

Tag	Total	Unique	Ratio
PER	9,986	2,851	3.5
LOC	9,563	1,540	6.2
ORG	8,520	1,923	4.4
EVT	2,601	235	11.0
PRO	1,699	739	2.3

(containing documents about Nord Stream, an offshore gas pipeline in Russia, and Ryanair, an Irish low-cost airline), it is also likely that the small number of unique entities could lead to poor domain generalisation results for those tags.

5.4.2 ACE 2005 Dataset

The annotated ACE 2005 corpus is provided by the ACE evaluation¹⁷. The ACE events are restricted to a range of types, each with a set of subtypes. Thus, only the events of an appropriate type are annotated in a document. The ACE dataset contains datasets in multiple languages (Chinese, Arabic, and English) with various types annotated for entities, relations, and events, from various information sources (e.g., broadcast conversations, broadcast news and telephone conversations). The data were created by Linguistic Data Consortium (LDC) with support from the ACE Program. The proposed tasks by ACE are more challenging than their MUC forerunners. In particular, the increased complexity resulted from the inclusion of various information sources and the introduction of more fine-grained entity types (e.g., facilities, geopolitical entities, etc.). In the context of this project, we use only the English ACE 2005 corpus that is composed of 599 articles. For the comparison of both models proposed, this dataset cannot be tested with the DANIEL system, since it is designed only for epidemic related data.

Table 33: English ACE 2005 corpus summary, Newswire (NW), Broadcast Conversation (BC), Broadcast News (BN), Telephone Speech (CTS), Usenet Newsgroups (UN), and Weblogs (WL). The number of documents annotated with one or multiple events is reported in brackets.

Total documents	NW	BN	BC	WL	UN	CTS
599 (553)	106 (104)	226 (211)	60 (60)	119 (93)	49 (47)	39 (38)

The corpus has 8 types of events, with 33 subtypes. These are the types of events:

- **Business:** Start-Org, Merge-Org, Declare-Bankruptcy, End-Org

¹⁷<https://catalog.ldc.upenn.edu/ldc2006t06>

- **Conflict:** Attack, Demonstrate
- **Contact:** Meet, Phone-Write
- **Life:** Be-Born, Marry, Divorce, Injure, Die
- **Movement:** Transport
- **Justice:** Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon
- **Transaction:** Transfer-Ownership, Transfer-Money
- **Personnel:** Start-Position, End-Position, Nominate, Elect

ACE 2005 event definition As presented in the Section 5.1, an ACE event is represented by an *event mention* (a text contains an event of a specific type and subtype), *event trigger* (the word that expresses the event mention), *event argument* (a participant in the event of a specific type), *argument role* (the role that the entity has in the event).

Since the EE task in the context of ACE 2005 has two sub-tasks, the ED represents the detection of the texts that contain an event of a specific type and the extraction of the event trigger from the text that expresses that type of event, and the event argument extraction, that is the detection of entities and their role in the event.

Every document is characterised by multiple events, or no events at all. The annotation of the event is done at the sentence level, and thus, the imbalanced nature of this dataset.

If we consider, for instance, this example from ACE 2005 dataset: *There was the free press in Qatar, Al Jazeera, but its offices in Kabul and Baghdad were **bombed** by Americans.*, an event detection system should output:

- *event mention*: this sentence contains an event of type **Conflict** and subtype **Attack**
- *event trigger*: this event of type **Conflict** and subtype **Attack** is triggered by the word **bombed**

An event argument extraction system should output:

- the *event arguments*: *Kabul* and *Baghdad*, which are entities of type **location**, and *Americans* which are considered an entity of type **person**
- the *event argument roles*: *Kabul* and *Baghdad* are **Places** and *Americans* have the **Attacker** role

5.4.3 DAnIEL Dataset

In this section, we present the dataset that was created for the *DAnIEL* system (Lejeune et al., 2015). The corpus is dedicated to multilingual epidemic surveillance and contains articles on different press threads in the field of *health* (Google News) that focused on epidemic events from different collected documents in different languages, with events simply defined as disease-location-number of victims triplets.

The corpus was built specifically for this system (Lejeune et al., 2015), containing articles from six different languages: English, French, Greek, Russian, Chinese, and Polish. It contains articles on different press threads in the field of *health* (Google News) focused on epidemic events. These documents have lengths that vary substantially, ranging from a short dispatch with one paragraph to a long article with a more detailed structure. Annotators, native speakers of the aforementioned languages, decide whether an article is relevant (speaks about an event) or not and then provide the disease name and location of the event.

A *DAnIEL* event (Lejeune et al., 2015) is defined at document-level, meaning that an article is considered as relevant if it is annotated with a (disease, location, number of victims) triplet, or a (disease, location) pair. An example is presented in Figure 12, where the event is a *listeria* outbreak in *USA* and number of victims is unknown.

Thus, in this dataset the event extraction task is defined as identifying articles that contain an event and the extraction of the disease name, location, number of victims, i.e. the words or compound words that evoke the event. Since the events are epidemic outbreaks, there is no pre-set list of types and subtypes of events, and thus the task of event extraction is simplified to detecting whether an article contains an ongoing epidemic event or not. Throughout the paper, we refer to the disease name or the location as event triggers (considering that these words most clearly express an epidemiological event).

```
"15960": {
  "annotations": [
    [
      "listeria",
      "USA",
      "unknown"
    ]
  ],
  "comment": "",
  "date": "2012-01-12",
  "language": "en",
  "document_path": "doc_en/20120112_www.cnn.
com_48eddc7c17447b70075c26a1a3b168243edcbfb28f0185",
  "url": "http://www.cnn.com/2012/01/11/health/listeria-outbreak/index.html"
}
```

Figure 12: Example of an event annotated in *DAnIEL* dataset.

However, the *DAnIEL* dataset is annotated at the document-level, which differentiates it from typical datasets used in research for the event extraction task. A document is either reporting an event at interest (i.e., disease-place pair, and sometimes the number of victims) or not. We pre-process and transform the dataset from document-level annotation to sentence-level. The annotations provided by *DAnIEL*, at the document-level, are looked-up in the appropriate file and the found offsets are attached to them. We consider as an example an article that has the following annotations at the document level: **malaria** and **worldwide**. The text of the article contains the following mentions: *Malaria*, *worldwide*.

*GENEVA: **Malaria** caused the death of an estimated 655,000 people last year, with 86 percent of victims children aged under five, World Health Organisation figures showed on Tuesday. The figure marked a five percent drop in deaths from 2009. Africa accounted for 91 percent of deaths and 81 percent of the 216 million cases **worldwide** in 2010. In its annual World **Malaria** Report for 2011, the WHO hailed as a "major achievement" a 26 percent fall in mortality rates since 2000 despite being well short of its 50 percent target. The UN health agency aims to eradicate malaria deaths altogether by the end of 2015 and reduce the number of cases by 75 percent on 2000 levels.*

In this case, in the first sentence, "*GENEVA: **Malaria** caused the death of an estimated 655,000 people [...]*", we are able to annotate **Malaria** at offsets 8 – 14. The process is automatic and continues in the same manner for the other annotations.

First, we consider the lemma of an annotated disease name that will further be looked-up in the text. If any disease name or location is found multiple times in the text, we annotate all the present instances. Sometimes, the exact surface form of a disease name cannot be found in the text, as it is the case for Russian, Greek, and Polish articles (morphologically rich languages), we considered the annotation of the grammatical cases of nouns. For example, in Russian, "Простуда" ("prostuda") means "cold", and since this disease name cannot be found in the text article, we used the instrumental case in Russian that can generally be distinguished by the "-ом" ("-om") suffix for most masculine and neuter nouns, the "-ою"/"-ой" ("-oju"/"-oj") suffix for most feminine nouns. The instrumental case for singular "простудой" was annotated in the article text.

In the case of locations, there were 57% of cases where the location could not be found in the text, mainly due to the coarse-grained type of manual annotation at the country-level. For the annotation of the locations at a finer-grained level, we considered the presence of cities or regions in the text. For example, if the document was previously annotated with “France”, and “Corsica” is mentioned in the text, we changed the final annotation to “Corsica”.

Finally, we tokenise the articles at the sentence-level and format them in the IOB (Inside-Outside-Beginning) tagging scheme where each token is given one of the following labels: *DISEASE*, *LOCATION* or *O*. We split the data into training (3,852 documents), test (481 documents), and validation (482 documents) sets, stratified by language. Table 34 presents some statistics for this dataset.

Table 34: Number of relevant tokens and sentences per dataset split per language.

Split	Sentences	Tokens	French	English	Polish	Chinese	Greek	Russian
Training	6,575	197,825	155,816	13,139	12,712	4,831	4,484	6,843
Validation	1,000	31,184	23,283	2,336	1,861	175	2,214	1,315
Test	782	23,930	18,183	1,472	119	366	1,836	1,954

5.5 Explored Approaches

In this section, we present three types of approaches for event detection.

5.5.1 DANIEL System

DANIEL (Lejeune et al., 2015) stands for Data Analysis for Information Extraction in any Language. The approach is at discourse-level, as opposed to the commonly used analysis at sentence-level, by exploiting the global structure of news as defined by the authors of (Lucas, 2009). Entries in the system are news texts, title and body of text, the name of the source when available, and other metadata (e.g date of article). As the name implies, the system has the capability to work in a multilingual setting due to the fact that it is not a word-based algorithm, which are highly language-specific, but rather a character-based one that centers around repetition and position (Lejeune et al., 2015). By avoiding grammar analysis and the usage of other NLP toolkits (e.g Part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style (Hamborg, Lachnit, Schubotz, Hepp, & Gipp, 2018; Lucas, 2009), the system is able to detect crucial information in salient zones that are peculiar to this genre of writing: the properties of the journalistic genre, the style universals, form the basis of the analysis.

Due to the fact that the DANIEL does not rely on any language-specific grammar analysis, and considers text as sequences of strings instead of words, DANIEL can quickly operate on any foreign language and extract crucial information early on and improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more than often, initial medical reports are in the vernacular language where patient zero appears (Lejeune et al., 2015).

DANIEL uses a minimal knowledge base, its central processing chain includes four phases:

- **Article segmentation:** The system first divides the document into stylistic segments: title, header, body and footer. The purpose is to identify salient zones where important information is usually repeated.

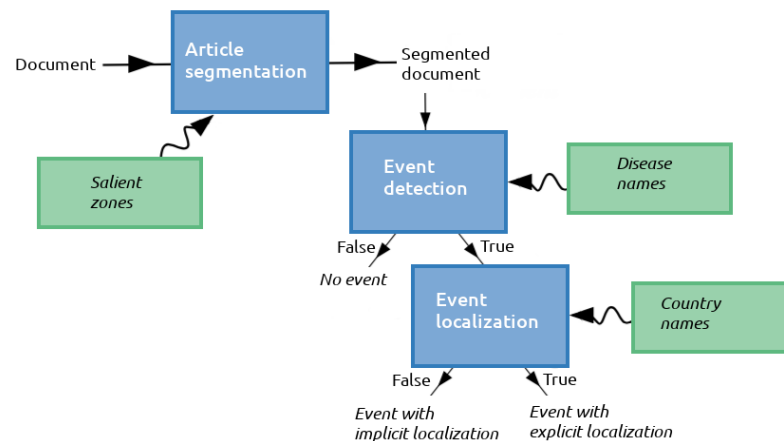


Figure 13: Event Detection pipeline in DANIEL.

- **Pattern extraction:** For detecting events, the system will look for repeated substrings at the salient zones aforementioned and determine whether they are maximal or not. A maximal substring is a string that cannot be extended to either its left nor right side (Ukkonen, 2009).
- **Filtering of these patterns:** Substrings that satisfy this condition will be matched to a list of disease/location names that was constructed by crawling from Wikipedia. The reason for using Wikipedia to build the knowledge base is that it is convenient to add lexicons from new languages without the assistance of a native speaker since information on Wikipedia can be easily crawled from one language to another.
- **Detection of disease – location pairs** (in some cases, the number of victims also): The end result of processing a document with DANIEL is one or more events that are described by pairs of disease-location.

5.5.2 Convolutional Neural Network-based Approaches

Word CNN In the Word CNN model, the target token $x^{(0)}$ is surrounded by a context formed by the surrounding words in the sentence, which constitutes the input for the convolution. In order to consider a limited sized context, longer sentences are trimmed and shorter ones are padded with a special token. We consider $2 \times n + 1$ the size of the context window, thus the input for the trigger candidate $x^{(0)}$ is represented as $x = [x^{(-n)}, x^{(-n+1)}, \dots, x^{(0)}, \dots, x^{(n-1)}, x^{(n)}]$. Each context token $x^{(i)}$ is associated with a word embedding an embedding representing its relative position to the trigger candidate $x^{(0)}$. This distance is an important informative feature, as it helps the representation of the context tokens to be focused on the candidate trigger.

That is, each core feature is embedded into a d -dimensional space, and represented as a vector in that space. The feature embeddings (the real values of the vector entries for each feature, in this case, words and distances) are treated as model parameters that need to be trained together with the other components of the network.

- **Word embeddings table** (initialised or not by some pre-trained word embeddings): to capture the hidden semantic and syntactic properties of the tokens
- **Position embeddings table:** to embed the relative distance i of the token $x^{(i)}$ to the current token $x^{(0)}$. Each distance value is associated with a d -dimensional vector (in practice, the table is initialised randomly), and these position embedding vectors are then trained as regular parameters

in the network.

For each token $x^{(i)}$, the vectors $X_{n \times 2+1, mt}$ obtained from the table look-ups above are concatenated into a single vector to represent the token. As a result, the original event trigger x is transformed into a matrix X , where d is the position embedding, w is the word embedding and mt is the size of the concatenated embeddings.

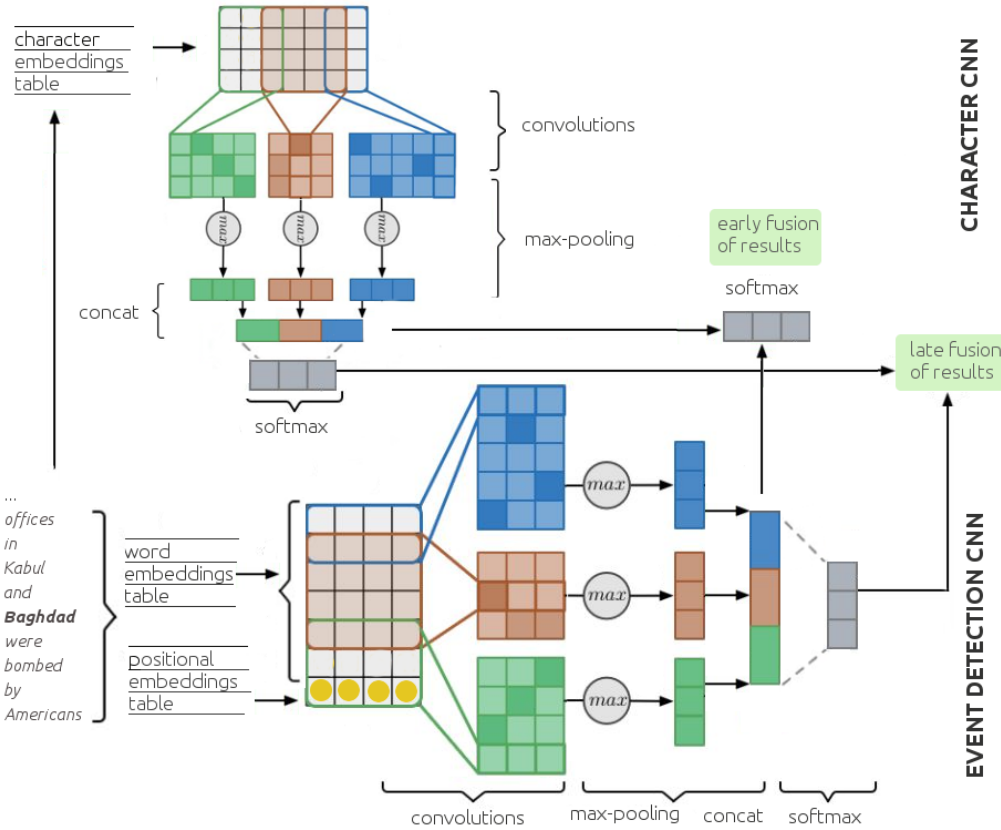


Figure 14: Word + Character CNN

For instance, in the example given in Figure 14 (bottom part), for the candidate trigger *Baghdad*, the word, and position embeddings are concatenated and passed through the convolution layer, where we have a set of feature maps (filters) $\{f_0, f_1, \dots, f_n\}$ for the convolution operation. Each feature map f_i corresponds to some window size k and can be essentially seen as a weight matrix of size $mt \times x$. A max-over-time pooling operation is applied over every feature map and the maximum value is taken as the feature corresponding to this particular filter.

The concatenated output of these filters form the representation of the input (target token and context), that can be given to a fully connected softmax layer for classification.

Character CNN We introduce the Character CNN that is close to the Word CNN and aims at generating a semantically-rich representation of the sequence of characters, by capturing relevant information associated to character n-grams: the different sizes of the filters of the convolution allows considering different sizes of character n-grams.

As in the Word CNN, for each trigger candidate $x^{(0)}$, we associate it with a context window of a maximum length of m characters. The considered characters are starting with the first word in the $2 \times n + 1$ words window considered in the Word CNN until the end of the sentence, all the characters of all the words

before are masked with zeros, and then, if the vector of characters does not exceed m , it is padded with zero. By padding with zeroes before the first word of the window of text, we maintain its position in the sentence, and thus the importance of the position of the trigger candidate. Each character in this context is represented by its embedding. However, we do not add position embeddings, which would not be relevant for characters. Consequently, the Character CNN applies a convolutional layer with a set of feature maps (filters) and a max-over-time pooling on the output as shown in the top model in Figure 14.

As for the Word CNN, the concatenated output of the filters constitutes the representation of the input, that can similarly be fed to a fully connected softmax for classification.

Early Fusion The first type of integration is the early fusion model, in which the two representations of the input sequence produced by the Word and Character CNNs (i.e. the concatenation of the output vectors of the filters) are concatenated before the fully-connected Softmax classification layer. Using this type of integration allows joint learning of the parameters of the two models in the training phase.

Late Fusion The late fusion integration of the Word and Character CNNs relies on combining the decision results of the two models, that are trained separately and therefore learned different characteristics of the candidate trigger. Indeed, the baseline CNN combines word and position embeddings that can capture syntactic and semantic information, and of course, the relative positions of words to the candidate trigger. The character-level CNN learns more local features from character n -grams and can capture morphological information. The late fusion focuses on the individual strength of these two models. The late fusion of the two models is motivated by the fact that the Word model has good coverage whereas the Character model is more focused on precision. More precisely, this late fusion is performed by a type of voting method, implemented as follows:

- if a trigger was detected by *Word CNN* and *Character CNN*, we keep the *Character CNN* label;
- if a trigger was detected by *Word CNN* but not by *Character CNN*, we keep the *Word CNN* label;
- if a trigger was detected by *Character CNN* but not by *Word CNN*, we keep the *Character CNN* label.

5.5.3 Fine-tuned Language Model-based Approaches

First, these models extend the recently introduced BERT (Devlin et al., 2019) model applied on sequential data. BERT itself is a stack of Transformer layers (Vaswani et al., 2017) which takes as input a sequence of subtokens, obtained by the WordPiece tokenization (Wu et al., 2016), and produces a sequence of context-based embeddings of these subtokens. When a word-level task, such as NER, is being solved, the embeddings of word-initial subtokens are passed through a dense layer with softmax activation to produce a probability distribution over output labels. We refer the readers to the original paper for a more detailed description. We modify BERT by adding a CRF layer instead of the dense one, which was commonly used in other works on neural sequence labeling (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016) to ensure output consistency.

5.6 Experimental Setup

We introduce in the following subsections the evaluation metrics and the experiments done regarding event detection.

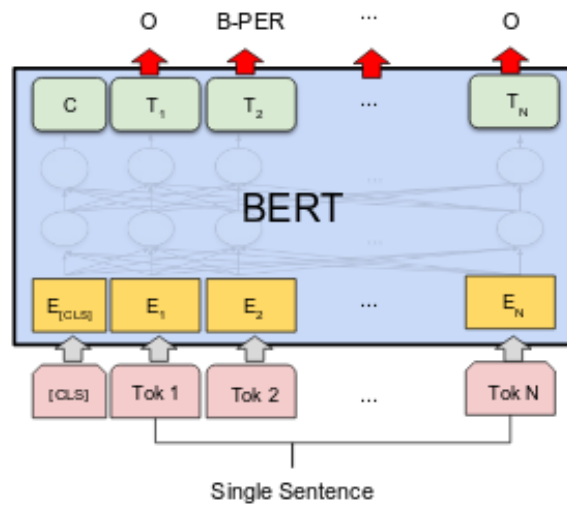


Figure 15: BERT-based classifier for sequential data (e.g. NER).

5.6.1 Evaluation Metrics

The evaluation is performed in a coarse-grained manner, with the entity (not token) as the unit of reference (Makhoul, Kubala, Schwartz, Weischedel, et al., 1999). We compute precision (P), recall (R), and F1 measure (F1) at micro-level, i.e. error types are considered over all documents, in a micro-strict scenario, *micro-strict*, which looks for an exact boundary matching (Ehrmann, Romanello, Bircher, & Clematide, 2020).

5.6.2 BSNLP 2019 Experiments

Methodology. The model used in the previous deliverable was proposed by (Ma & Hovy, 2016), an end-to-end model combining a BiLSTM and a CNN character encoding, in order to take advantage of the word and character features. The word representations are obtained from an unsupervised learning model that yields word embeddings as distributional semantics (Bojanowski et al., 2017; Pennington et al., 2014). Character embeddings are obtained from a CNN which takes as input a sequence of characters of each token, similarly to (Ma & Hovy, 2016). The BiLSTM model adds, therefore, to each word vector a new feature represented as a character-based vector. The character-based features are concatenated with the word embeddings and fed into a BiLSTM. A CRF is used on top to jointly decode labels for the whole sequence of words. A more detailed description of the model can be found in (Ma & Hovy, 2016). We consider that character-level features can capture morphological and shape information (Kanaris, Kanaris, Houvardas, & Stamatatos, 2007; C. D. Santos & Zadrozny, 2014; C. N. d. Santos & Guimaraes, 2015) which can increase the power of representation for words that occur infrequently or misspelled or custom words produced by an OCR tool.

Experimental setup. For this BiLSTM+CNN model presented in the previous deliverable, we used the FastText¹⁸ pre-trained word embedding models (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018)¹⁹.

For this dataset, we compare our previously obtained results (BiLSTM+CNN) with two approaches based on the pre-trained model BERT proposed by (Devlin et al., 2019) and a model based on XLM-

¹⁸<https://fasttext.cc/docs/en/crawl-vectors.html>

¹⁹For a more detailed description of the model and of the hyperparameters can be found in (Ma & Hovy, 2016).

Table 35: BSNLP 2019 results for EVT on the blind test data. * The results for EVT classification in deliverable D2.2.

		Bulgarian	Czech	Polish	Russian
BiLSTM+CNN*	P	–	–	–	–
	R	–	–	–	–
	F1	26.5	0.0	20.1	0.0
SlavicBERT	P	27.8	26.1	36.4	9.1
	R	33.3	26.0	30.8	25.0
	F1	30.3	26.0	33.3	13.3
<i>BERT-multilingual-cased</i>	P	23.8	25.0	39.1	8.3
	R	33.3	26.1	34.6	25.0
	F1	27.8	25.5	36.7	12.5
<i>XLNet-RoBERTa-base</i>	P	41.7	29.6	46.2	33.3
	R	33.3	34.8	46.2	25.0
	F1	37.0	32.0	46.2	28.6

RoBERTa (Conneau et al., 2020). For the BERT-based models, we use a multilingual trained language model *bert-multilingual-cased* and a model trained on the languages present in BSNLP 2019, *slavic-bert*. The XLNet-RoBERTa is a Transformer-based masked cross-lingual language model trained on one hundred languages, using more than two terabytes of filtered CommonCrawl data. We chose this model due to the fact that it significantly outperformed the *bert-multilingual* on a variety of cross-lingual benchmarks.

The results clearly state that the *XLNet-RoBERTa* is more suited for this task considering that it obtained the highest F1 values for all the languages, for the EVT tag. Due to the imbalance of this tag in the documents, the BiLSTM+CNN is not able to capture enough informative features about EVT, and thus, it obtains an F1 of 0 for Czech and Russian, while all the fine-tuned pre-trained language models manage to outperform this model.

5.6.3 DANIEL Experiments

Methodology. Often, approaches for text-based disease surveillance follow a two-step process (Joshi, Karimi, Sparks, Paris, & Macintyre, 2019): document classification and event extraction. First, we perform the document classification into relevant and irrelevant documents, e.g. documents that contain mentions of disease names and locations, and documents that do not. For this, we chose a BERT-based model whose performance on text classification is an F1 of 86.54%. We do not focus on the classification task but rather on the event extraction task: detection and extraction of the disease names and locations. For this step, we compare different state-of-the-art models, first, we experiment with deep learning models, BiLSTM models (Lample et al., 2016; Ma & Hovy, 2016), and further with two architectures based on pre-trained language models. We test on the predicted relevant documents provided by the document classification step. We also evaluate two deep neural architectures based on bidirectional LSTM models proposed by (Lample et al., 2016) and (Ma & Hovy, 2016) that use character and word representations²⁰.

Additionally, we evaluate the pre-trained model BERT proposed by (Devlin et al., 2019) for token classification²¹.

²⁰The hyperparameters for both models are detailed in the papers (Lample et al., 2016) and (Ma & Hovy, 2016).

²¹For this model, we used the hyperparameters recommended in (Devlin et al., 2019).

Table 36: Evaluation of the DAnIEL system on the initial test data for event detection.

		Polish	Chinese	Russian	Greek	French	English	All
DAnIEL	P	30.0	50.0	25.0	58.3	50.4	40.0	42.3
	R	25.0	50.0	28.5	53.8	44.1	40.0	46.1
	F1	27.2	50.0	26.6	56.0	47.1	40.0	44.1
BiLSTM+CNN	P	—	—	—	—	—	—	88.7
	R	—	—	—	—	—	—	50.0
	F1	N/A	67.1	48.1	66.6	69.5	26.0	64.0
BiLSTM+LSTM	P	—	—	—	—	—	—	91.1
	R	—	—	—	—	—	—	49.9
	F1	N/A	67.0	49.0	66.6	69.5	30.0	64.5
BERT-multilingual-cased	P	—	—	—	—	—	—	82.2
	R	—	—	—	—	—	—	53.7
	F1	N/A	68.7	50.0	46.1	75.5	28.1	64.9
BERT-multilingual-uncased	P	—	—	—	—	—	—	84.0
	R	—	—	—	—	—	—	55.9
	F1	N/A	68.4	77.1	57.1	79.5	32.3	67.1
XLM-RoBERTa-base	P	—	—	—	—	—	—	81.7
	R	—	—	—	—	—	—	60.5
	F1	N/A	75.5	49.1	88.8	62.6	22.6	69.5

Experimental setup. Due to the multilingual characteristic of the dataset, we utilise the *bert-base-multilingual-cased* pre-trained and then fine-tuned BERT model. We also experiment with the *XLM-RoBERTa-base* model (hereafter *XLM-RoBERTa*) proposed by (Conneau et al., 2020) that has shown significant performance gains for a wide range of cross-lingual transfer tasks. We consider this model appropriate for our task and dataset due to the multilingual characteristic of the data²².

When we test on the predicted relevant documents, errors are being propagated to the event extraction step. The recall drops significantly since some relevant documents have been discarded by the classifier but we still evaluate by comparing with all the ground-truth of relevant documents. Still, one can notice from the Table 36 the same tendency of obtaining the highest precision with the *BiLSTM-LSTM* model and the highest F1 with the *XLM-RoBERTa-base* model. This model seems to be the most robust since it has the lowest drop in recall among all the models.

Finally, the performance of the models was evaluated for each language on the predicted relevant documents. As presented in Table 36, the models produced highest results for the French language which is not surprising since it is the language with the largest training dataset. The BiLSTM-based and BERT-based models generally performed well for French and Greek languages, while the *XLM-RoBERTa-base* for Chinese language. Interestingly, the worst results are for the Russian dataset. If we look at macro F1-measure, *BERT-multilingual-uncased* is the best performing model with 68.58 (63.8 for *XLM-RoBERTa*).

First, the low precision values when training and testing on all data instances are not surprising, since the amount of negative examples, with potential false positives, rises up to around 90%. We also notice that the results when using the ground-truth documents are balanced in precision and recall, while, when testing on the predicted relevant documents, the recall is lower for the BiLSTM-based models and higher for the transformer-based models. Overall, *XLM-RoBERTa-base* had the best performance in terms of the F1 score. This can be attributed to the robust optimisation and pre-training in a cross-lingual manner of the model on a significantly larger multilingual dataset compared to BERT.

²² *XLM-RoBERTa* was trained on 2.5TB of newly created clean CommonCrawl data in 100 languages.

The analysis of the performance of the models per language reveals that the best model (*XLM-RoBERTa-base*) had the highest scores for the French language. These results could be attributed to the size of French-language texts. For instance, French language tokens constitute 75.98% of all the tokens in the test data for the relevant documents.

This work is associated with Appendixes: 1, 8, 11, and 12.

5.6.4 ACE 2005 Experiments

For comparison purposes, we use the same test set with 40 newswire articles (672 sentences), the same development set with 30 other documents (863 sentences) and the same training set with the remaining 529 documents (14,849 sentences) as in previous studies of this dataset (Ji et al., 2008; Liao & Grishman, 2010; Q. Li et al., 2013; Nguyen & Grishman, 2015; Nguyen, Cho, & Grishman, 2016). Following previous work (Ji et al., 2008; Liao & Grishman, 2010; Hong et al., 2011; Nguyen & Grishman, 2015; Chen et al., 2015), a trigger is correct if its event subtype and offsets match those of a reference trigger.

Experimental setup. The hyperparameters used for both models are depicted in Table 38 and described as follows.

We train both networks (Word CNN and Character CNN) with Adam optimizer (Kingma & Ba, 2014). During the training, we optimise the embedding tables (i.e., word, position, and character embeddings) to achieve the optimal states. Finally, for training, we use the batch size of 256 for the Word CNN and a batch size of 128 for the Character CNN. When they are trained jointly in the early fusion model, we use a batch size of 128. We would also stress the fact that the batch size affects the Adam optimiser (Smith, Kindermans, Ying, & Le, 2017), and thus our different choices of batch size for the models, which was optimised on the validation set.

We compare our model with several neural-based models proposed for the same task, that do not use external resources, namely: the CNN model without any external features in (Nguyen & Grishman, 2015), the dynamic multi-pooling CNN model (Chen et al., 2015), the bidirectional joint RNNs (Nguyen, Cho, & Grishman, 2016), the non-consecutive CNN in (Nguyen, Fu, et al., 2016), the hybrid model proposed by (Feng et al., 2016), the GAIL-ELMo model proposed by (T. Zhang et al., 2019), the Gated Cross-Lingual Attention model presented in (J. Liu et al., 2018), and the Graph CNN proposed by (Nguyen & Grishman, 2018). We do not consider models that are using other external resources such as (Bronstein et al., 2015), (W. Li et al., 2019), or (Yang et al., 2019), since we only rely on the given sentences in our model. We also compare this model with four baselines based on the BERT language model, applied in a similar way to (Devlin et al., 2019) for the named entity recognition (NER) task, with the recommended hyperparameters, a learning rate of 2×10^{-5} and with a maximum length of 128 tokens (longer sentences are split into 128 texts that are also added to the whole process).

The best performance (75.79 F1 on the blind test set) is achieved by combining word and position embeddings with the character-level features using a late fusion strategy. This performance relates to improvements that have been reported on other tasks, when concatenating word embeddings with the output from a character-level CNN, for Part-of-Speech tagging (Dos Santos & Gatti, 2014) and NER (C. N. d. Santos & Guimaraes, 2015). From Table 37, we can also outline that adding character embeddings in a late fusion strategy outperforms all the word-based models, including complex architectures such as the graph CNN, and the models based on the BERT language model. Between the BERT models, it is worth noticing that the *cased* models perform better than the *uncased* ones, which confirms that the character morphology is important for the task, maybe because capitalisation is connected to the recognition of named entities, that are usually considered important to detect event mentions.

However, we can see that the character embeddings are not sufficient on their own: using only the Character CNN, we observe that the recall is the smallest of all the approaches considered. Yet, the

Table 37: Evaluation of our models and comparison with state-of-the-art systems for event detection on the blind test data. ⁺with gold arguments.

Models	P	R	F1
CNN (Nguyen & Grishman, 2015)	71.9	63.8	67.6
CNN ⁺ (Nguyen & Grishman, 2015)	71.8	66.4	69.0
Dynamic multi-pooling CNN (Chen et al., 2015)	75.6	63.6	69.1
Joint RNN (Nguyen, Cho, & Grishman, 2016)	66.0	73.0	69.3
CNN with document context (Duan et al., 2017)	77.2	64.9	70.5
Non-Consecutive CNN (Nguyen, Fu, Cho, & Grishman, 2016)	N/A	N/A	71.3
Attention-based ⁺ (S. Liu et al., 2017)	78.0	66.3	71.7
GAIL (T. Zhang et al., 2019)	74.8	69.4	72.0
Gated Cross-Lingual Attention (J. Liu, Chen, Liu, & Zhao, 2018)	78.9	66.9	72.4
Graph CNN (Nguyen & Grishman, 2018)	77.9	68.8	73.1
Seed-based (Bronstein et al., 2015)	80.6	67.1	73.2
Hybrid NN (Feng et al., 2016)	84.6	64.9	73.4
Attention-based GCN (X. Liu, Luo, & Huang, 2018)	76.3	71.3	73.7
Δ -learning (Lu et al., 2019)	76.3	71.9	74.0
DEEB-RNN3y (Zhao et al., 2018)	72.3	75.8	74.0
BERT-large-uncased+LSTM (Wadden et al., 2019)	N/A	N/A	68.9
BERT-base-uncased (Wadden et al., 2019)	N/A	N/A	69.7
BERT-base-uncased (Du & Cardie, 2020)	67.1	73.2	70.0
BERT-QA (Du & Cardie, 2020)	71.1	73.7	72.3
DMBERT (Wang et al., 2019)	77.6	71.8	74.6
DMBERT+Boot (Wang et al., 2019)	77.9	72.5	75.1
BERT-multilingual-uncased	61.7	67.7	64.6
BERT-multilingual-cased	68.2	70.8	69.5
BERT-base-uncased	71.6	68.4	70.0
BERT-base-cased	71.3	72.0	71.6
BERT-large-uncased	72.0	72.9	72.5
BERT-large-cased	69.3	77.1	73.0
Our CNN (replicated, changed hyperparameters)	68.8	66.1	67.4
Character CNN	71.7	41.1	52.3
Word + Character CNN - early fusion	88.5	61.8	72.8
Word + Character CNN - late fusion	87.1	67.0	75.7

precision achieved is considerably high (71.72), which implies that this model is more sure about the triggers that were retrieved.

Given this observation, we can compare the two integration strategies, early and late fusions:

- in the case of early fusion, where the two models are trained jointly, we notice that the precision is the highest between all the compared models. We assume that in the joint approach, the power of representation of morphological properties provided by the characters is overtaking the influence of the word and positions embedding, and the combination reproduce the imbalance between precision and recall observed for the Character CNN, the recall being the lowest between all the models except for the *Character CNN*;
- in the case of the late fusion, since we have more control on the combination and we can give priority to the Character CNN to establish the labels on the trigger candidates retrieved by the Word CNN, the method takes advantage of the high precision of the Character CNN, allowing an increase of the precision from 71.72 to 87.15, while still having a high recall, also increasing the recall of the *Word CNN* model from 65.88 to 67.05. The late fusion integration is therefore able to

Table 38: Hyperparameters used for the Word and Character CNNs.

Hyperparameter	Value
Maximum Sequence	31
Convolutional window sizes (Word CNN)	{1, 2, 3}
Convolutional window sizes (Character CNN)	{2, 3, 4, 5, 6, 7, 8, 9, 10}
Word Embedding Dimension	300
Character Embedding Dimension	300
Position embeddings dimension	50
Non-linear Layer	ReLU
Weights Initialisation	Orthogonal (Saxe, McClelland, & Ganguli, 2013)
Dropout	0.5
Pre-trained Word Embeddings	Word2vec, (Mikolov, Chen, Corrado, & Dean, 2013)
Learning Rate	0.001
Optimiser	Adam

Table 39: Examples of results correctly found with the Word+Character CNN (late fusion).

Event Type	New triggers correctly found	Trigger words in training data
End-Position	<i>steps</i>	<i>step</i>
Extradite	<i>extradited</i>	<i>extradition</i>
Attack	<i>wiped</i>	<i>wipe</i>
Start-Org	<i>creating</i>	<i>create</i>
Attack	<i>smash</i>	<i>smashed</i>
End-Position	<i>retirement</i>	<i>retire</i>

take into account the complementarity of the two models.

Finally, for more qualitative analysis, we examine the new triggers correctly detected by the Word + Character CNN (late-fusion), in comparison with the Word CNN. We observe that among the 37 new correctly found triggers, some are indeed derivational or inflectional variants of known words in the training data, such as illustrated in Table 39. This seems to confirm that the character-based model can capture some semantic information associated with morphological characteristics of the words and manage to detect new correct event mentions that correspond to inflections of known event triggers (i.e. existing in the training data). Also, the fact that the convolution windows in the Character CNN range from 2 to 10 means that character n -grams in the same range are included in the model and contribute to the model's ability to handle different word variations.

This work is associated with the Appendix 8.

5.7 Discussion

In a transfer learning scenario, where we trained a BERT-based model pre-trained on a large multilingual dataset on the ACE 2005 dataset, we intended to analyse how the model would perform on the EMBEDDIA languages. In the next examples, we also mark the predicted entities from Section 3 in Croatian. Thus, we ran the *BERT-multilingual-cased* model on the Croatian test dataset of Wikiann (Pan et al., 2017), and we present some examples of detected events. The next sentence can be translated as *Namely, the cause of her anger was the sacrifice of her daughter Iphigenia and jealousy for Cassandra*.

Naime, uzrok njezina bijesa bilo je <Die Event> žrtvovanje </Die Event> njezine kćeri <PER> Ifigenije </PER> i ljubomora zbog <PER> Kasandre </PER>.

We notice that the model is able to predict a person's death. Moreover, in the next example, *Veljko Bulajić*

(co-winner) was seriously wounded in one of the fights for the defense of Madrid., the model also manages to predict an attack with an injure consequence.

<PER> Veljko Bulajić </PER> (sudobitnik) U jednoj od <Attack> borbi </Attack> za obranu Madrida bio je teško <Injure> ranjen </Injure>.

Finally, the model seems to be able to detect at sentence-level the presence of several types of events, but due to the lack of annotated data, the evaluation assessment is not possible.

5.8 Conclusions

We experimented with three datasets from which two of them contain documents in low-resourced languages BSNLP 2019 (Bulgarian, Czech, Polish, and Russian) and DANIEL dataset (Polish, Chinese, Russian, Greek and two with a considerable larger amount of documents, English and French). The ACE 2005 dataset contains a widely used set of documents annotated with events in English. We investigated three annotation styles for event detection: a sentence-level annotation where an event defined as a named entity (e.g. *Brexit* represents an event entity); a document-level defined event (e.g. a document talks about an epidemic outbreak and it is annotated with a pair disease-location, e.g. *malaria-U.S.*); and another sentence-level annotation where an event is represented by an event mention (event triggers that represent the most a type of event, e.g. *killing* triggers an event of type *Attack*) and multiple events can be present in the same sentence.

The events-based datasets are characterised by imbalance, and this observation is visible in the obtained scores, as, for example, in BSNLP 2019, the event named entities proved to be difficult to be detected. The previous results showed that the events in Czech and Russian could not be identified, while current pre-trained language models-based approaches managed to increase the scores to around 30% for both languages.

For DANIEL, all the experimented methods obtained higher results than the DANIEL system baseline, with the exception of English due to error propagation from the article classification.

For ACE 2005, the models based on character (CNNs with word and character embeddings) and sub-word contextual embeddings (fine-tuned pre-trained language models) proved to increase the possibility of capturing informative features for morphologically rich languages. At the same time, these models are more difficult to transfer onto other languages, as it also depends on the type of the embeddings used. Meanwhile, the BERT-based models that are based on contextual embeddings and can better models these type of cases, are more adaptable to other languages when pre-trained on multilingual datasets. After analysing the results on several EMBEDDIA languages, we consider that this could be a possible solution for language and domain adaptation to other datasets, regardless of the richness of their morphology.

6 Conclusions and Future Work

In this deliverable, we presented the outcomes regarding three natural language processing (NLP) tasks: named entity recognition (NER), named entity linking (NEL), and event detection (ED). On these three tasks, we explored different methods from the state of the art, that have been improved and modified to work on multiple languages.

Specifically, for the NER task, we explored four different neural network architectures, two based on a BiLSTM and two founded on BERT (Devlin et al., 2019). We trained models on all the eight languages of EMBEDDIA using the dataset Wikiann (Pan et al., 2017). Moreover, for some languages, such as Croatian and Finnish, we trained and tested models on datasets that have been extensively used in the state of the art. The obtained results showed that NER systems based on BERT can outperform in most cases other architectures, even other neural-based approaches. However, in order to achieve the best performance possible with BERT, it is necessary to provide a complex architecture rather than just fine-tuning a pre-trained language model. We observed, as well, that, in most of the cases, BERT models pre-trained on one or just a couple of languages, work better than pre-trained models over hundreds of them.

For the NEL task, our evaluation of the WikiANN corpora showed us that our multilingual model outperformed our cross-lingual approach for almost all languages of the EMBEDDIA project. Building resources and analysing specific language versions of Wikipedia has made it possible to obtain more relevant information, such as surface names and contextual information. Moreover, the multilingual probability tables combined with word/entity embeddings and training data on the target language improved the analysis and overall performance of our approach.

For the ED task, we experimented with three datasets from which two of them contain documents in low-resourced languages BSNLP 2019 (Bulgarian, Czech, Polish, and Russian) and DANIEL dataset (Polish, Chinese, Russian, Greek and two with a considerably larger amount of documents, English and French). The ACE 2005 dataset contains a widely used set of documents annotated with events in English. We investigated three annotation styles for event detection: a sentence-level annotation where an event defined as a named entity (e.g. *Brexit* represents an event entity); a document-level defined event (e.g. a document talks about an epidemic outbreak and it is annotated with a pair disease-location, e.g. *malaria-U.S.*); and another sentence-level annotation where an event is represented by an event mention (event triggers that represent the most a type of event, e.g. *killing* triggers an event of type *Attack*) and multiple events can be present in the same sentence. The experiments reveal that, generally, pre-trained and fine-tuned language models-based methods performed better than the other methods (DANIEL system or CNN-based models).

We will continue working on the improvement of the three aforementioned tasks. In the case of NER, we would like to add stacked Transformer Blocks to our Multitask BERT and see if it can improve the performance. As well, we would add information, such as the pseudo-affixes into BERT and see if we can improve the results in highly inflected languages. Finally, we would continue applying the methods to other datasets and languages, in order to have a more global vision of the performance of our systems.

Concerning the NEL task, a perspective would be to adapt our entity linking approach to automatically generate ontologies for low-resourced data. As well, it would be interesting to propose a post-processing filter to select candidates that have the same entity type as the mentions and exploit the data from other knowledge bases such as DBpedia, Wikidata, or BabelNet.

Future work for the ED task implies investigating language models-based methods due to their ability for low-shot transfer learning, while also using additional informative features brought by named entities.

7 Associated Outputs

The work described in this deliverable has resulted in the following resources:

Description	URL	Availability
NER BiLSTM-CNN-CRF-BERT	https://github.com/EMBEDDIA/bert-bilstm-cnn-crf-ner	Public (MIT License)
NER REST API	https://github.com/EMBEDDIA/URL_NER_REST	Public (MIT License)
Multilingual NEL	https://github.com/EMBEDDIA/multilingual_entity_linking	Public (Apache 2.0 Licence)
Event Detection	https://github.com/EMBEDDIA/event-detection	To become public
NER BERT Multi-task	https://github.com/EMBEDDIA/NER_BERT_Multitask	To become public

Works marked as *To become public* mean that they are available only within the consortium while the associated work is yet to be published. They will be released publicly when the associated work is published.

We present in Table 40 and Table 41, the publications that have been produced between December 2019 and December 2020 and that are related to this deliverable.

Table 40: Publications related to this deliverable.

Citation	Status	Appendix
Mutuvi, S., Doucet A., Lejeune, G. and Odeo, M. (2020). A Dataset for Multi-lingual Epidemiological Event Extraction. Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)	Published	1
Boroş, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N. and Doucet, A. (2020). Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. Proceedings of the 24th Conference on Computational Natural Language Learning (CoNLL)	Published	2
Moreno, J. G., Linhares Pontes, E., and Dias, G. (2020) CTRL@WiC-TSV: Target Sense Verification using Marked Inputs and Pre-trained Models. Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)	Published	3
Frossard, E., Coustaty, M., Doucet, A., Jatowt, A. and Hengchen S. (2020). Dataset for Temporal Analysis of English-French Cognates. Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)	Published	4
Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Boroş, E., Hamdi, A., Sidère, N., Coustaty, M. and Doucet, A. (2020). Entity Linking for Historical Documents: Challenges and Solutions. Proceedings of the 22nd International Conference on Asia-Pacific Digital Libraries (ICADL)	Published	5

Table 41: Publications related to this deliverable.

Citation	Status	Appendix
Boroş, E., Moreno, J. G. and Doucet A. Event Detection with Entity Markers	Accepted at ECIR 2021 (Conf.)	6
Boroş, E., Nguyen, N. K., Lejeune, G. and Doucet, A. Event Detection over digitised and historical documents	Submitted to JOCCH (Journal).	7
Nguyen, N. K., Boroş, E., Lejeune, G. and Doucet, A. (2020). Impact Analysis of Document Digitization on Event Extraction. Proceedings of the 4th Workshop on Natural Language for Artificial Intelligence (NL4AI)	Published	8
Cabrera-Diego, L. A., Moreno, J. G. and Doucet, A. Improving NER systems by marking uppercase tokens, and predicting masked tokens and entities boundaries	To submit to CLEOPATRA Workshop 2021	9
Linhares Pontes, E., Moreno, J. G. and Doucet, A. (2020). Linking Named Entities across Languages using Multilingual Word Embeddings. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)	Published	10
Mutuvi, S. Boroş, E., Lejeune, G., Jatowt, A., Doucet, A. and Odeo, M. Multilingual Epidemic Event Extraction	Submitted to EACL 2021 (Conf.)	11
Mutuvi, S., Boroş, E., Doucet, A., Lejeune, G., Jatowt, A. and Odeo, M. Multilingual Epidemiological Text Classification: A Comparative Study. Proceedings of the 28th International Conference on Computational Linguistics (COLING)	Published	12
Moreno, J. G., Doucet, A., and Grau, B. (2020) . Relation Classification via Relation Validation. Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)	Published	13
Boroş, E., Linhares Pontes, E., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidère, N. and Doucet, A. (2020). Robust Named Entity Recognition and Linking on Historical Multilingual Documents. Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2020)	Published	14
Moreno, J. G., Boroş, E. and Doucet, A. (2020) TLR at the NTCIR-15 FinNum-2 Task: Improving Text classifiers for Numeral Attachment in Financial Social Data. Proceedings of the 15th NTCIR Conference Evaluation of Information Access Technologies	Published	15

References

- Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., & Ellis, J. (2014). A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the second workshop on events: Definition, detection, coreference, and representation* (pp. 45–53).
- Ahn, D. (2006). The stages of event extraction. In *Proceedings of the workshop on annotating and reasoning about time and events* (pp. 1–8).
- Al-Rfou, R., Kulkarni, V., Perozzi, B., & Skiena, S. (2014). POLYGLOT-NER: Massive Multilingual Named Entity Recognition. *CoRR*, *abs/1410.3791*. Retrieved from <http://arxiv.org/abs/1410.3791> (eprint: 1410.3791)
- Alves, D., Thakkar, G., & Tadić, M. (2020, May). Evaluating Language Tools for Fifteen EU-official Under-resourced Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1866–1873). Marseille, France: European Language Resources Association.
- Arhipov, M., Trofimova, M., Kuratov, Y., & Sorokin, A. (2019, August). Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 89–93). Florence, Italy: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3712> doi: 10.18653/v1/W19-3712
- Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 17th international conference on computational linguistics-volume 1* (pp. 86–90).
- Baldini Soares, L., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019, July). Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2895–2905). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1279
- Bekavac, B., & Tadić, M. (2007, June). Implementation of Croatian NERC System. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing* (pp. 11–18). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W07-1702>
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. Retrieved from <https://www.aclweb.org/anthology/Q17-1010> doi: 10.1162/tacl_a_00051
- Boros, E., Hamdi, A., Linhares Pontes, E., Cabrera-Diego, L. A., Moreno, J. G., Sidere, N., & Doucet, A. (2020, November). Alleviating Digitization Errors in Named Entity Recognition for Historical Documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 431–441). Online: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/2020.conll-1.35>
- Boros, E., Linhares Pontes, E., Cabrera-Diego, L. A., Hamdi, A., Moreno, J. G., Sidere, N., & Doucet, A. (2020). Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névél (Eds.), *CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*. CEUR-WS.
- Bronstein, O., Dagan, I., Li, Q., Ji, H., & Frank, A. (2015). Seed-based event trigger labeling: How far can event descriptions get us? In *Acl (2)* (pp. 372–376).
- Broscheit, S. (2019, November). Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd conference on computational natural language learning (conll)* (pp. 677–685). Hong Kong, China: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/K19-1063> doi: 10.18653/v1/K19-1063
- Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. , 18, 467–479.

- Cañete, J., Chaperon, G., Fuentes, R., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.
- Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015). Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing* (Vol. 1, pp. 167–176).
- Chinchor, N., Lewis, D. D., & Hirschman, L. (1993). Evaluating message understanding systems: an analysis of the third message understanding conference (muc-3). *Computational linguistics*, 19(3), 409–449.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2020, July). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/2020.acl-main.747> doi: 10.18653/v1/2020.acl-main.747
- Dalianis, H., & Åström, E. (2001). *SweNam-A Swedish Named Entity recognizer Its construction, training and evaluation* (Tech. Rep. No. TRITA-NA-P0113 - IPLab-189.) Stockholm, Sweden: Department of Numerical Analysis and Computing Science, KTH Royal Institute of Technology.
- Dembowski, J., Wiegand, M., & Klakow, D. (2017). Language Independent Named Entity Recognition using Distant Supervision. In Z. Vetulani & P. Paroubek (Eds.), *Proceedings of the 8th Language & Technology Conference* (pp. 68 – 72). Poznań: Fundacja Uniwersytetu im. Adama Mickiewicza.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. doi: 10.18653/v1/N19-1423
- Dos Santos, C. N., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Coling* (pp. 69–78).
- Du, X., & Cardie, C. (2020). Event extraction by answering (almost) natural questions. *arXiv preprint arXiv:2004.13625*.
- Duan, S., He, R., & Zhao, W. (2017). Exploiting document level information to improve event detection via recurrent neural networks. In *Eighth international joint conference on natural language processing (ijcnlp 2017)* (pp. 352–361). Asian Federation of Natural Language Processing.
- Ehrmann, M., Romanello, M., Bircher, S., & Clematide, S. (2020). Introducing the clef 2020 hipe shared task: Named entity recognition and linking on historical newspapers. In *European conference on information retrieval* (pp. 524–532).
- Feng, X., Huang, L., Tang, D., Ji, H., Qin, B., & Liu, T. (2016). A language-independent neural network for event detection. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 66–71).
- Finkel, J. R., Grenager, T., & Manning, C. (2005, June). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 363–370). Ann Arbor, Michigan: Association for Computational Linguistics. doi: 10.3115/1219840.1219885
- Fišer, D., Ljubešić, N., & Erjavec, T. (2020). The janex project: language resources and tools for slovene user generated content. , 54(1), 223–246. doi: 10.1007/s10579-018-9425-z
- Gage, P. (1994). A new algorithm for data compression. , 12(2), 23–38. (Publisher: McPherson, KS: R & D Publications, c1987-1994.)
- Ganea, O.-E., & Hofmann, T. (2017). Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2619–2629).

- Association for Computational Linguistics. Retrieved from <http://aclweb.org/anthology/D17-1277> doi: 10.18653/v1/D17-1277
- Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., & Ivanov, V. (2013). Introducing Baselines for Russian Named Entity Recognition. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 329–342). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.
- Glavaš, G., Karan, M., Šaric, F., Šnajder, J., Mijic, J., Šilic, A., & Bašic, B. D. (2012). CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. In *Information Society 2012-Eighth Language Technologies Conference* (pp. 73–78).
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (lrec 2018)*.
- Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. Retrieved from <https://www.aclweb.org/anthology/C96-1079>
- Grishman, R., Westbrook, D., & Meyers, A. (2005). Nyu's english ace 2005 system description. *ACE*, 5.
- Guo, S., Chang, M.-W., & Kiciman, E. (2013, June). To link or not to link? a study on end-to-end tweet entity linking. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1020–1030). Atlanta, Georgia: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/N13-1122>
- Gupta, R., & Sarawagi, S. (2009). Domain adaptation of information extraction models. *ACM SIGMOD Record*, 37(4), 35–40.
- Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., & Gipp, B. (2018, 03). Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions.. doi: 10.1007/978-3-319-78105-1_39
- Han, X., & Zhao, J. (1999). Nlpr_kbp in tac 2009 kbp track: A two-stage method to entity linking. In *In proceedings of test analysis conference 2009 (tac 09)*. MIT Press.
- Heinzerling, B., & Strube, M. (2018). BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In N. C. C. chair et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA).
- Hobbs, J. R., Appelt, D., Tyson, M., Bear, J., & Israel, D. (1992). SRI International: Description of the FASTUS system used for MUC-4. In *4th conference on message understanding* (pp. 268–275).
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.
- Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., & Zhu, Q. (2011). Using cross-entity inference to improve event extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 1127–1136).
- Hong, Y., Zhou, W., jingli, Zhou, G., & Zhu, Q. (2018). Self-regulation: Employing a generative adversarial network to improve event detection. In *56th annual meeting of the association for computational linguistics (acl 2018)* (pp. 515–526). Association for Computational Linguistics. Retrieved from <http://aclweb.org/anthology/P18-1048>
- Huang, R., & Riloff, E. (2012). Bootstrapped training of event extraction classifiers. In *13th conference of the european chapter of the association for computational linguistics (eacl 2012)* (pp. 286–295).

- Jagannatha, A. N., & Yu, H. (2016). Bidirectional rnn for medical event detection in electronic health records. In *2016 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 2016, pp. 473–482).
- Ji, H., Grishman, R., et al. (2008). Refining event extraction through cross-document inference. In *Adl* (pp. 254–262).
- Joshi, A., Karimi, S., Sparks, R., Paris, C., & Macintyre, C. R. (2019). Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6), 1–19.
- Kanaris, I., Kanaris, K., Houvardas, I., & Stamatatos, E. (2007). Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06), 1047–1067.
- Kapočiūtė, J., & Raškinis, G. (2005). Rule-based Annotation of Lithuanian Text Corpora. *Information Technology and Control*, 34(3), 290–296.
- Kapočiūtė-Dzikiene, J., Nøklestad, A., Johannessen, J. B., & Krupavičius, A. (2013, May). Exploring Features for Named Entity Recognition in Lithuanian Text Corpus. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (pp. 73–88). Oslo, Norway: Linköping University Electronic Press, Sweden. Retrieved from <https://www.aclweb.org/anthology/W13-5611>
- Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kittask, C., Milintsevich, K., & Sirts, K. (2020). Evaluating multilingual BERT for Estonian. In A. Utka, J. Vaičenonienė, J. Kovalevskaitė, & D. Kalinauskaitė (Eds.), *Human Language Technologies – The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020* (Vol. 328, pp. 19–26). IOS Press. doi: 10.3233/FAIA200597
- Kokkinakis, D. (2003). Swedish NER in the Nomen Nescio project. In H. Holmboe (Ed.), *Nordisk Sprogteknologi – Nordic Language Technology 2002* (pp. 379–398). Copenhagen, Denmark: Museum-Tusculanums Forlag.
- Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., & Borin, L. (2014, May). HFST-SweNER — A New NER Resource for Swedish. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2537–2543). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/391_Paper.pdf
- Kolitsas, N., Ganea, O.-E., & Hofmann, T. (2018). End-to-end neural entity linking. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 519–529). Association for Computational Linguistics. Retrieved from <http://aclweb.org/anthology/K18-1050>
- Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., ... Zajc, A. (2019). *Training corpus ssj500k 2.2*. Retrieved from <http://hdl.handle.net/11356/1210> (Slovenian language resource repository CLARIN.SI)
- Krupka, G., Jacobs, P., Rau, L., & Iwańska, L. (1991). GE: Description of the NLToolset System as Used for MUC-3. In *3rd conference on message understanding* (pp. 144–149).
- Kuraton, Y., & Arkhipov, M. (2019). *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*. (eprint: 1905.07213)
- Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning* (pp. 282–289). Morgan Kaufmann Publishers Inc.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Conference of the north american chapter of the association for computational linguistics: Human language technologies*.
- Laur, S. (2013). *Nimeüksuste korpus*. Center of Estonian Language Resources. doi: 10.1515/1-00-0000-0000-00073L

- Laur, S., Orasmaa, S., Särg, D., & Tammo, P. (2020, May). EstNLTK 1.6: Remastered Estonian NLP Pipeline. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 7152–7160). Marseille, France: European Language Resources Association. Retrieved from <https://www.aclweb.org/anthology/2020.lrec-1.884>
- Le, P., & Titov, I. (2018). Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 1595–1604). ACL.
- Lejeune, G., Brixteel, R., Doucet, A., & Lucas, N. (2015, 07). Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65. doi: 10.1016/j.artmed.2015.06.005
- Li, J., Luong, M.-T., & Jurafsky, D. (2015). A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.
- Li, J., Sun, A., Han, J., & Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. doi: 10.1109/TKDE.2020.2981314
- Li, P., Zhu, Q., & Zhou, G. (2013). Argument inference from relevant event mentions in chinese argument extraction. In *Acl (1)* (pp. 1477–1487).
- Li, Q., Ji, H., & Huang, L. (2013). Joint event extraction via structured prediction with global features. In *Acl (1)* (pp. 73–82).
- Li, W., Cheng, D., He, L., Wang, Y., & Jin, X. (2019). Joint event extraction based on hierarchical event schemas from framenet. *IEEE Access*, 7, 25001–25015.
- Liao, S., & Grishman, R. (2010). Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 789–797).
- Lindén, K., Axelsson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., . . . Silfverberg, M. (2013). HFST — A System for Creating NLP Tools. In C. Mahlow & M. Piotrowski (Eds.), *Systems and Frameworks for Computational Morphology* (pp. 53–71). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Linhares Pontes, E., Doucet, A., & Moreno, J. G. (2020). Linking named entities across languages using multilingual word embeddings. In *Jointed Conference on Digital Libraries (JCDL) 2020*.
- Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020). Event extraction as machine reading comprehension. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 1641–1651).
- Liu, J., Chen, Y., Liu, K., & Zhao, J. (2018). Event detection via gated multilingual attention mechanism. In *Thirty-second aai conference on artificial intelligence (aaai-18)*.
- Liu, S., Chen, Y., He, S., Liu, K., Zhao, J., et al. (2016). Leveraging framenet to improve automatic event detection.
- Liu, S., Chen, Y., Liu, K., & Zhao, J. (2017). Exploiting argument information to improve event detection via supervised attention mechanisms. In *55th annual meeting of the association for computational linguistics (acl 2017)* (pp. 1789–1798). Vancouver, Canada.
- Liu, X., Luo, Z., & Huang, H. (2018). Jointly multiple events extraction via attention-based graph information aggregation. *arXiv preprint arXiv:1809.09078*.
- Ljubešić, N., & Erjavec, T. (2016, May). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1527–1531). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from <https://www.aclweb.org/anthology/L16-1242>
- Ljubešić, N., Klubička, F., Agić, Z., & Jazbec, I.-P. (2016, May). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings*

- of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 4264–4270). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from <https://www.aclweb.org/anthology/L16-1676>
- Ljubešić, N., Stupar, M., Jurić, T., & Agić, Z. (2013, December). Combining available datasets for building named entity recognition models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2), 35–57. doi: 10.4312/slo2.0.2013.2.35-57
- Lu, Y., Lin, H., Han, X., & Sun, L. (2019). Distilling discrimination and generalization knowledge for event detection via delta-representation learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4366–4376).
- Lucas, N. (2009). *Modélisation différentielle du texte, de la linguistique aux algorithmes* (HDR Thesis, Université de Caen). Retrieved from <https://tel.archives-ouvertes.fr/tel-01073406>
- Luoma, J., Oinonen, M., Pyykönen, M., Laippala, V., & Pyysalo, S. (2020, May). A Broad-coverage Corpus for Finnish Named Entity Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 4615–4624). Marseille, France: European Language Resources Association. Retrieved from <https://www.aclweb.org/anthology/2020.lrec-1.567>
- Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1064–1074). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P16-1101> doi: 10.18653/v1/P16-1101
- Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al. (1999). Performance measures for information extraction. In *Proceedings of darpa broadcast news workshop* (pp. 249–252).
- Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv cs.CL*. (eprint: 2007.01658)
- McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. Retrieved from "<http://mallet.cs.umass.edu>"
- McNamee, P., Mayfield, J., Lawrie, D., Oard, D., & Doermann, D. (2011). Cross-language entity linking. In *Proceedings of 5th international joint conference on natural language processing* (pp. 255–263). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *International conference on learning representations (iclr 20013), workshop track*.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4), 235–244.
- Moreno, J. G., Linhares Pontes, E., Coustaty, M., & Doucet, A. (2019, August). TLR at BSNLP2019: A Multilingual Named Entity Recognition System. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 83–88). Florence, Italy: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3711> doi: 10.18653/v1/W19-3711
- Mosbach, M., Andriushchenko, M., & Klakow, D. (2020). *On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines*. (eprint: 2006.04884)
- Mozharova, V., & Loukachevitch, N. (2016). Two-stage approach in Russian named entity recognition. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)* (pp. 1–6). doi: 10.1109/FRUCT.2016.7584769
- Munro, R., & Manning, C. (2012). *State-of-the-Art Multilingual NER Using Loosely Aligned Text* (Tech. Rep.). Stanford University.
- Nguyen, T. H., Cho, K., & Grishman, R. (2016). Joint event extraction via recurrent neural networks. In *Proceedings of naacl-hlt* (pp. 300–309).

- Nguyen, T. H., Fu, L., Cho, K., & Grishman, R. (2016). A two-stage approach for extending event detection to new types via neural networks. *ACL 2016*, 158.
- Nguyen, T. H., & Grishman, R. (2015). Event detection and domain adaptation with convolutional neural networks. In *Acl (2)* (pp. 365–371).
- Nguyen, T. H., & Grishman, R. (2016). Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of emnlp*.
- Nguyen, T. H., & Grishman, R. (2018). Graph convolutional networks with argument-aware pooling for event detection. In *Thirty-second AAAI conference on artificial intelligence (aaai 2018)*.
- Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (2017, July). Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1946–1958). Vancouver, Canada: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P17-1178> doi: 10.18653/v1/P17-1178
- Patwardhan, S., & Riloff, E. (2009). A unified model of phrasal and sentential evidence for information extraction. In *2009 conference on empirical methods in natural language processing (emnlp 2009)* (pp. 151–160).
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014)* (pp. 1532–1543). Association for Computational Linguistics (ACL).
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. doi: 10.18653/v1/N18-1202
- Pinnis, M. (2012, May). Latvian and Lithuanian Named Entity Recognition with TildeNER. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 1258–1265). Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/948_Paper.pdf
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, July). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14
- Rahimi, A., Li, Y., & Cohn, T. (2019, July). Massively Multilingual Transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 151–164). Florence, Italy: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P19-1015>
- Raiman, J., & Raiman, O. (2018). Deeptype: Multilingual entity linking by neural type system evolution. In *Aaai conference on artificial intelligence* (pp. 5406–5413).
- Reimers, N., & Gurevych, I. (2019). *Alternative weighting schemes for ELMo embeddings*. (eprint: 1904.02954)
- Rijhwani, S., Xie, J., Neubig, G., & Carbonell, J. (2019, January). Zero-shot neural transfer for cross-lingual entity linking. In *Thirty-third aaai conference on artificial intelligence (aaai)*. Honolulu, Hawaii.
- Riloff, E. (1996a). Automatically generating extraction patterns from untagged text. In *Aaai'96* (pp. 1044–1049).
- Riloff, E. (1996b). An empirical study of automated dictionary construction for information extraction in three domains. *Artificial intelligence*, 85(1), 101–134.

- Rosales-Méndez, H., Hogan, A., & Poblete, B. (2020). Fine-grained entity linking. *Journal of Web Semantics*, 65, 100600. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1570826820300378> doi: <https://doi.org/10.1016/j.websem.2020.100600>
- Santos, C. D., & Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st international conference on machine learning (icml-14)* (pp. 1818–1826).
- Santos, C. N. d., & Guimaraes, V. (2015). Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*.
- Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. , 45(11), 2673–2681. doi: 10.1109/78.650093
- Shen, W., Wang, J., & Han, J. (2014). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2), 443–460.
- Sil, A., Kundu, G., Florian, R., & Hamza, W. (2018). Neural cross-lingual entity linking. In *AAAI* (pp. 5464–5472). AAAI Press.
- Smith, S. L., Kindermans, P.-J., Ying, C., & Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.
- Štajner, T. (2013). Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0*, 1(2).
- Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., & Xiong, C. (2020). Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT. (eprint: 2003.04985)
- Tanvir, H., Kittask, C., & Sirts, K. (2020). EstBERT: A Pretrained Language-Specific BERT for Estonian. (eprint: 2011.04784)
- Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (pp. 142–147). Retrieved from <https://www.aclweb.org/anthology/W03-0419>
- Tkachenko, A., Petmanson, T., & Laur, S. (2013, August). Named Entity Recognition in Estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing* (pp. 78–83). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W13-2412>
- Tsygankova, T., Mayhew, S., & Roth, D. (2019, August). BSNLP2019 Shared Task Submission: Multisource Neural NER Transfer. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 75–82). Florence, Italy: Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3710> doi: 10.18653/v1/W19-3710
- Ukkonen, E. (2009, 10). Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science*, 410, 4341–4349. doi: 10.1016/j.tcs.2009.07.015
- Ulčar, M., & Robnik-Šikonja, M. (2020a). FinEst BERT and CroSloEngual BERT. In P. Sojka, I. Kopeček, K. Pala, & A. Horák (Eds.), *Text, Speech, and Dialogue* (pp. 104–111). Cham: Springer International Publishing.
- Ulčar, M., & Robnik-Šikonja, M. (2020b). High Quality ELMo Embeddings for Seven Less-Resourced Languages. In N. Calzolari et al. (Eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020* (pp. 4731–4738). European Language Resources Association. Retrieved from <https://www.aclweb.org/anthology/2020.lrec-1.582/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

- Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., ... Pyysalo, S. (2019). *Multilingual is not enough: BERT for finnish*. (eprint: 1912.07076)
- Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.
- Wang, X., Han, X., Liu, Z., Sun, M., & Li, P. (2019). Adversarial training for weakly supervised event detection. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 998–1008).
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2019). Hugging-Face's Transformers: State-of-the-art Natural Language Processing. *ArXiv, abs/1910.03771*.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... others (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019). Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5284–5294).
- Yangarber, R., Grishman, R., Tapanainen, P., & Huttunen, S. (2000). Automatic acquisition of domain knowledge for information extraction. In *18th international conference on computational linguistics (coling 2000)* (pp. 940–946).
- Yu, J., Bohnet, B., & Poesio, M. (2020, July). Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6470–6476). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.577
- Zhang, T., Ji, H., & Sil, A. (2019). Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2), 99–120.
- Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., & Artzi, Y. (2020). *Revisiting Few-sample BERT Fine-tuning*. (eprint: 2006.05987)
- Zhang, W., Sim, Y. C., Su, J., & Tan, C. L. (2011). Entity linking with effective acronym expansion, instance selection and topic modeling. In *Proceedings of the twenty-second international joint conference on artificial intelligence - volume three* (pp. 1909–1914). AAAI Press. Retrieved from <http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-319> doi: 10.5591/978-1-57735-516-8/IJCAI11-319
- Zhao, Y., Jin, X., Wang, Y., & Cheng, X. (2018). Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 414–419).
- Zhou, S., Rijhwani, S., & Neubig, G. (2019, November). Towards zero-resource cross-lingual entity linking. In *Proceedings of the 2nd workshop on deep learning approaches for low-resource nlp (deeplo 2019)* (pp. 243–252). China: ACL.
- Znotiņš, A., & Cīrule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. In K. Muischnek & K. Mūrisep (Eds.), *Human Language Technologies – The Baltic Perspective: Proceedings of the Eighth International Conference Baltic HLT 2018* (Vol. 307, pp. 183 – 189). Tartu, Estonia: IOS Press. doi: 10.3233/978-1-61499-912-6-183
- Znotiņš, A., & Guntis Barzdīns. (2020). LVBERT: Transformer-Based Model for Latvian Language Understanding. In Andrius Utkā, Jurgita Vaičenonienė, Jolanta Kovalevskaitė, & Danguolė Kalinauskaitė (Eds.), *Human Language Technologies – The Baltic Perspective: Proceedings of the Ninth International Conference Baltic HLT 2020* (Vol. 328, pp. 111 – 115). Kaunas, Lithuania: IOS Press. doi: 10.3233/FAIA200610

Appendices

1 A Dataset for Multilingual Epidemiological Event Extraction

A Dataset for Multilingual Epidemiological Event Extraction

Stephen Mutuvi[✉] Antoine Doucet[✉] Gaël Lejeune[◇] Moses Odeo[✉]

[✉]University of La Rochelle, L3i Laboratory, firstname.lastname@univ-lr.fr

[✉]Multimedia University of Kenya, smutuvi@mmu.ac.ke

[◇]Sorbonne University Paris, gael.lejeune@sorbonne-universite.fr

Abstract

This paper proposes a corpus for the development and evaluation of tools and techniques for identifying emerging infectious disease threats in online news text. The corpus can not only be used for information extraction, but also for other natural language processing (NLP) tasks such as text classification. We make use of articles published on the Program for Monitoring Emerging Diseases (PROMED) platform, which provides current information about outbreaks of infectious diseases globally. Among the key pieces of information present in the articles is the uniform resource locator (URL) to the online news sources where the outbreaks were originally reported. We detail the procedure followed to build the dataset, which includes leveraging the source URLs to retrieve the news reports and subsequently pre-processing the retrieved documents. We also report on experimental results of event extraction on the dataset using the Data Analysis for Information Extraction in any Language (DANIEL) system. DANIEL is a multilingual news surveillance system that leverages unique attributes associated with news reporting to extract events: repetition and saliency. The system has wide geographical and language coverage, including low-resource languages. In addition, we compare different classification approaches in terms of their ability to differentiate between epidemic-related and unrelated news articles that constitute the corpus.

Keywords: Epidemiology, corpus creation, event extraction, classification, multilingual NLP

1. Introduction

Web corpora, describing text corpora created from the Web, have recently become popular due to the availability of a vast amount of electronic texts on the Web. The World Wide Web is a valuable source of data, which enables building corpora with wide-ranging attributes such as varying sizes, languages and domains. Such corpora can be analyzed and utilized in key application areas, among which epidemic intelligence.

Epidemic intelligence is an integral component of infectious disease early-warning mechanisms. It involves the collection, analysis, and dissemination of key information related to disease outbreaks, with the objective of detecting outbreaks and providing early warning to public health stakeholders (World Health Organization, 2014). Disease surveillance mechanisms can broadly be classified as either indicator- or event-based surveillance (Huff et al., 2016).

Indicator-based surveillance is the conventional type of surveillance systems which rely predominantly on local health practitioners to identify infectious disease outbreaks, where suspected outbreak cases are subjected to laboratory tests for confirmation. A key determinant of the efficiency of such surveillance methods is the underlying health care infrastructure. Poor infrastructure can result in inaccurate and irrelevant information being disseminated, with a likelihood of significant time delays in the dissemination of key information relevant to disease outbreaks (Zhou et al., 2011). Inadequate health infrastructure directly often leads to incomplete geographical coverage when reporting epidemics (Huff et al., 2016; World Health Organization, 2014). Time delays and incomplete coverage may hinder the deployment of effective health interventions, potentially leading to loss of lives.

Today, using NLP to monitor informal information sources, such as social media, search queries, online news outlets, and blogs has become an essential part of epidemic surveil-

lance (Salathé et al., 2013; Bernardo et al., 2013). Advancements in NLP present an opportunity to efficiently collect, process and analyze large textual data from the Web, to detect disease-related features from the text. Near real-time data-driven surveillance systems, commonly referred to as event-based surveillance (EBS) systems can now be easily developed and deployed. EBS encompasses analyzing textual data mostly generated via the Web, for incidences of events related to disease outbreaks (Huff et al., 2016). A more plausible approach is to utilise a combination of formal and informal sources for the timely and accurate detection of infectious disease outbreak (O’Shea, 2017). As such, it has been determined that event-based surveillance methods can complement traditional surveillance methods, for timely and accurate detection of epidemics (Chunara et al., 2012; O’Shea, 2017).

However, despite the rise in the use of advanced text processing and analysis approaches, such as deep learning, a limited number of corpora exists for the training and evaluation of disease events extraction models. The few available datasets are relatively small in size and predominantly in English language. A key requirement for training deep learning models that give satisfactory results is having large-scale datasets. This is further compounded by the fact that epidemic reports originate from a wide range of sources and languages.

In view of the above, we attempt to address the dearth of data for epidemic event extraction, by creating a corpus that can be used by researchers and practitioners in building and evaluate epidemic event extraction algorithms and applications. We leverage the Program for Monitoring Emerging Diseases (PROMED) reporting platform to create the corpus. PROMED aggregates disease outbreak reports across the world and is open and publicly available. The PROMED articles undergo a review and verification process by experts before being published on the

platform. The aggregation of the reports by subject matter experts makes the articles suitable for use as ground truth to evaluate epidemiological information extraction systems. The multilingual dataset we extracted from PROMED comprises articles in English, French, Portuguese and Spanish languages. To the best of our knowledge, this is among the largest datasets of this nature that is available for developing and evaluating multilingual epidemic surveillance tools and techniques.

The paper is organized as follows. We review related work on event extraction in Section 2., while the methodology used to create the corpus is described in Section 3. The experiments to train the corpus in a text classification task are detailed in Section 4. Additionally, we evaluate event extraction over the corpus using the DANIEL system. The results are discussed in Section 5., before conclusions are drawn and future work presented in Section 6.

2. Related Work

Event extraction (EE) is an important information extraction (IE) task that focuses on identifying an event mention from text and extracting information relevant to the event. Typically, this entails predicting event triggers, the occurrence of events with specific types, and extracting arguments associated with an event.

While event extraction is a crucial sub-task of information extraction, it still remains quite a challenging task due to the difficulties associated with encoding words semantics in various context (Zhan and Jiang, 2019). For instance, the same event might appear in the form of various trigger expressions or expressions might represent different event types in different contexts.

Event extraction methods are classified into three types, namely pattern-based, data-driven and hybrid methods (Hogenboom et al., 2011). Pattern-based methods use rules and templates to extract events from text through representation and exploitation of expert knowledge. On the other hand, data-driven approaches use statistical techniques to discover the relations in text. Recently, methods based on deep learning have gained popularity among researchers in the field (Zhan and Jiang, 2019).

Specific to epidemiological event extraction, there exist a number of empirical works targeted to extract events related to disease outbreaks. Among them is Data Analysis for Information Extraction in any Language (DANIEL), a multilingual news surveillance system that leverages repetition and saliency, properties that are common in news writing (Lejeune et al., 2015). The multilingual nature of the system enables global and timely detection of epidemic events since it eliminates the requirement for translating local news to other languages for subsequent transmission. The system can easily be adapted and scaled to extract events across languages, therefore, being able to have a wider geographical coverage. Reactivity and geographic coverage are of paramount importance in epidemic surveillance (Lejeune et al., 2015).

Similar to DANIEL are BIOCASTER (Collier, 2011; Collier et al., 2008) and PULS (Du et al., 2011) which have produced good results in analysing disease-related news reports and providing a summary of the epidemics. The Eco-

Health Alliance Global Rapid developed the Identification Tool System (GRITS), an application that provides automatic analyses of epidemiological texts. The system extracts important information about a disease outbreak, such as the most likely disease, dates, and countries where the outbreak originates. The pipeline for GRIT entails transforming words to vectors using TF-IDF, extracting features using pattern-matching tools, before applying the binary relevance-based classifier to predict the available disease in the text (Huff et al., 2016).

Internet search data has also been exploited for disease surveillance. In one study, internet searches for specific cancers were found to correlate with their estimated incidence and mortality (Cooper et al., 2005). Monitoring influenza outbreak using data drawn from the Web has also been previously explored. Two different studies, one utilizing GOOGLE (Ginsberg et al., 2009) and the other YAHOO (Polgreen et al., 2008) search queries, analyzed the searches and estimated the number of reported influenza cases. In recent years, a flurry of work has utilized social media data for infectious disease surveillance (Paul et al., 2016; Charles-Smith et al., 2015). Mostly, Twitter data, has been used for disease tracking (Lamb et al., 2013; Collier et al., 2011; Culotta, 2010), outbreak detection (Li and Cardie, 2013; Bodnar and Salathé, 2013; Diaz-Aviles et al., 2012; Aramaki et al., 2011) and predicting the likelihood of individuals falling sick (Sadilek et al., 2012). News media has also been used to give early warning of increased disease activity before official sources have reported (Brownstein et al., 2008). The studies have demonstrated the potential value of harnessing data-driven approaches for epidemic surveillance.

3. Methods

In this section, we describe the procedure followed to create the corpus. We also detail the process for evaluating event extraction and classification models over the corpus.

3.1. Corpus Creation

We retrieved PROMED articles in English, French, Spanish and Portuguese languages, for the period August 1, 2013, to August 31, 2019. PROMED reports global outbreaks of infectious diseases. The articles contain various key meta-data such as title, description, location, date and source URL where the article was originally published. The source URLs present in the PROMED articles were extracted and their corresponding source documents downloaded. Figure 1 shows the percentage of documents still available online for each year in the date range 01-08-2013 to 31-08-2019. The source URLs, together with the other meta-data were formatted and stored in JSON format making corpus¹ easily reusable and reproducible. Therefore, this makes it easy for any interested researcher to process the dataset and use it in modeling epidemiological event extraction or any other related NLP tasks.

Various processing tasks were performed on the extracted Web data to transform it into a clean text corpus. Firstly,

¹Available online at <https://zenodo.org/record/3709617>.

language filtering was performed to ensure that only documents belonging to the languages of interest were retained. The documents were grouped into different clusters using the K-means clustering algorithm. This enabled filtering documents with little to no textual content. The silhouette coefficient was computed to quantify the appropriate number of clusters for each set of data. This coefficient measures how well data is assigned to its own cluster and how far it is from other clusters (Rousseeuw, 1987). A coefficient close to 1 (one) means the data sample is located in the appropriate cluster while -1 (negative one) implies data has been assigned to the wrong cluster. Elimination of boilerplate content from the corpus was among the data cleaning tasks. Content such as navigation links, headers and footers were removed from HTML pages using the JUSTEXT library (Pomikálek, 2011). Removal of boilerplate content is highly desirable, since such content rarely provides useful evidence about the phenomenon being investigated. On the contrary, the high frequency of the boilerplate content could introduce bias into the text data, hence negatively impacting the performance of derived applications (Vogels et al., 2018). The final pre-processing task was deduplication. Deduplication involves eliminating perfect duplicate and near-duplicate content so that only one instance of each text was preserved. The ONION (ONE Instance ONLY tool (Pomikálek, 2011)), which deduplicates text data by measuring the similarity of paragraphs or entire document was used. It is based on a n-gram-based one-pass deduplication algorithm, where for each document all word n-grams are extracted (10-grams by default) and compared with the set of previously seen n-grams (Pomikálek, 2011).

Another dataset was specifically prepared for training a text classification model as described in Section 3.3.2. This dataset is composed of news articles from the News Category Dataset (Misra, 2018), consisting of around 200,000 English news articles. These news articles, which do not have mentions of disease outbreaks, were published on the HuffPost news website between the years 2012 and 2018. The dataset categorizes news articles based on their headlines and short descriptions. The news articles are grouped into various categories such as politics, wellness, travel, entertainment, sports and healthy living, among others. A total of 5,000 articles from the categories politics, entertainment, and sports were randomly selected and downloaded from the HuffPost news platform. They form the set of irrelevant documents, completed by a random selection of 5,000 documents from the PROMED dataset. Together with 444 evaluation documents from DANIEL, this forms the English part of the data set described in Table 3.

3.2. Corpus Statistics

The corpus statistics, PROMED and source documents, are presented on Table 1 and Table 2 respectively.

Table 3 presents statistics for the corpus used for training and evaluating text classification models. The dataset is composed of epidemic relevant articles from ProMED and non-relevant documents from News Category Dataset, described in Section 3.2. A total of 10,000 and 2,996 documents in English and French language, comprising relevant and non-relevant documents formed the training set.

Language	#Documents	#Sentences	#Words
English (en)	19,149	558,448	53,325,455
French (fr)	1,849	28,823	5,593,184
Spanish (es)	3,453	27,918	4,458,533
Portuguese (pt)	3,451	48,591	5,994,583

Table 1: Statistics for Retrieved PROMED Documents

Language	#Documents	#Sentences	#Words
English	13,275	320,613	8,749,272
French	1,395	13,777	439,153
Spanish	1,994	27,751	863,672
Portuguese	1,562	14,424	528,701

Table 2: Statistics for PROMED documents retrieved from their source

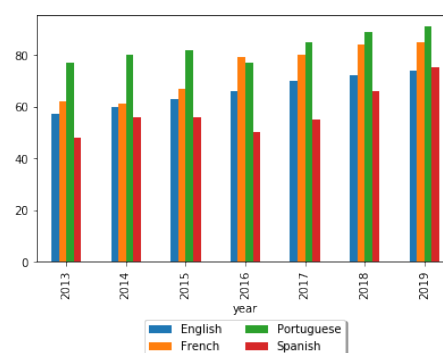


Figure 1: Percentage of ProMED sources accessible by year.

The relevant and non-relevant documents were equally distributed among the two classes.

Human-annotated datasets (Lejeune et al., 2015) in English and French provided the ground truth to evaluate the models. The availability of the annotated dataset for the two languages informed the decision for their consideration in our experiments. The two other languages, Spanish and Portuguese, which currently do not have ground truth data will be considered in our future studies. The annotated datasets comprising 444 and 2,722 documents for English and French languages respectively. The test data had a high degree of imbalance between the classes. The 444 English test documents had 31 relevant and 413 non-relevant documents. The French test corpus comprised 299 relevant and 2,207 non-relevant documents. The highly imbalanced nature of the data was important in helping depict the models' ability to classify the documents into their respective classes.

3.3. Evaluation

In this subsection, we describe the procedure for extracting epidemic events present in the corpus using DANIEL. We use supervised approaches to classify the retrieved documents as either being epidemic-related or not. Naive Bayes, Random Forest and Neural Network classification models

were trained on the extracted text and evaluated on a manually annotated dataset to ascertain the models ability to generalize on unseen data.

3.3.1. The DANIEL System

We evaluated the performance of the DANIEL system in extracting epidemic events present in the corpus. The DANIEL processing pipeline comprises three steps: news article segmentation, event detection, and event localization. DANIEL adopts a discourse-level event extraction approach, where the global structure of news is exploited (Lejeune et al., 2015). The system relies on properties that are common to the journalistic genre regardless of the language. The most useful features are repetition and saliency, which defines the relative importance of prominence of news contents. While the majority of systems extract events at the sentence level, by harnessing the morphological, syntactical and semantic features of a sentence, hence dependent on language-specific modules, DANIEL uses language-agnostic text-level features. It is character-based, hence handles text as a sequence of characters rather than as a sequence of words. Rather than exploiting keywords, the system exploits strings of text, but only if the strings have been repeated in pre-defined salient zones in text. The output of the DANIEL system is a disease-location pair describing an event as a disease outbreak and the place where it occurred. Recall and precision scores were obtained to determine the performance. The results are presented in Section 4.

For further evaluation, subsets of the English and French language datasets were subjected to annotation by three native speakers for each language. These annotators had to judge whether documents presented to them had mentions of an infectious disease outbreak or not. Subsequently, for the relevant documents, the annotators were requested to specify the disease name and location. We measured the inter-annotator agreement using Cohen's kappa coefficient. The inter-annotator agreement determines the extent to which annotators assign the same score to the same variable (McHugh, 2012). Finally, leveraging the generated ground truth, the evaluation was quantitatively measured against the annotators' judgments on the evaluation corpus.

3.3.2. Text Classification Model

We train and evaluate text classification models using datasets described in Table 3. The models classify a news article as either relevant or non-relevant, depending on whether it alerts about a disease outbreak or not. The training data comprised 10,000 and 2,722 news articles in English and French languages respectively, with documents equally distributed among the two classes. Pre-processing of the text input was undertaken which included filtering of stopwords and tokenizing the data.

A human-annotated dataset presented in Table 3 was used as the test set to evaluate the performance of the classification models. With the data ready, we trained multinomial naive Bayes, random forest and neural network classifiers over the created corpus. The naive Bayes classifier was used as the base model. Naive Bayes has been proven to be viable for text classification and information retrieval in general (Le et al., 2019). Parameter tuning was undertaken for the random forest, with the aim of enhancing

Dataset	#Documents	#Sentences	#Words
Train-en	10,000	317,862	9,879,559
Train-fr	2,996	43,257	1,959,584
Test-en	444	4,728	230,353
Test-fr	2,722	75,479	2,058,941

Table 3: Statistics for train and test datasets used in training and evaluating the text classification models

its performance. Finally, the models were evaluated to determine their performance using the human-annotated test data. Due to the imbalanced nature of the data, we considered recall, precision, and F-measure metrics, which are more appropriate if there exists a greater degree of imbalance in the classes (Bunker and Thabtah, 2017)

4. Results

The DANIEL system attained an F-score of 75% for documents both in English and French. For the documents in the English, the system achieved a precision of 60% and was able to correctly identify all the relevant documents. Precision and recall scores of 74% and 83% were obtained on the French language documents.

The results for text classification models trained and evaluated using the datasets built in this study are presented in Table 4 and Table 5 for English and French datasets respectively. The models' F-score on the English test data was 74%, 63% and 53% for Random Forest, Neural Network and Naive Bayes model respectively. For the French documents, an F-score of 67%, 63%, and 50% was obtained for Random Forest, Naive Bayes and Neural Network model respectively.

Classifier	Precision	Recall	F-measure
Naive Bayes	57%	75%	53%
Random Forest	80%	70%	74%
Neural Network	68%	76%	61%

Table 4: Text Classification Report for the English Documents

Classifier	Precision	Recall	F-measure
Naive Bayes	62%	74%	63%
Random Forest	80%	63%	67%
Neural Network	64%	52%	50%

Table 5: Text Classification Report for the French Documents

5. Discussion

We developed and evaluated baseline models on detection and extraction of epidemic events from online news articles. Overall, the random forest model gave the highest prediction, in classifying news articles as either reporting an epidemic event or not. However, for French documents, the model could predict only one class using the default classification threshold of 0.5. This necessitated experimenting

with a lower threshold of 0.3, which produced superior results compared to the other models. This can be attributed to the fine-tuning of the model's parameters and its ability to learn discriminating and reliable features from the text corpus. However, the process of tuning the parameters required significant effort and time.

The neural network model did not give strong results compared to the random forest model. Possible approaches towards improving the performance of the model could include using more advanced model architectures and transfer learning. In natural language processing (NLP), transfer learning is achieved via pre-trained language models, for instance the bidirectional encoder representations from transformers (BERT). Such language models enable to learn contextualized representations which upon fine-tuning for tasks such as classification usually result in significant performance gains. Typically, language models are trained on large text corpora, hence being able to adequately capture linguistic features and representations, which result in improved performance in downstream tasks. The performance of the DANIEL event extraction system was consistently good for both languages. This can be attributed to the fact that DANIEL's rule-based inference engine leverages language and disease text resources, which are readily available for all languages.

6. Conclusion

Early detection of disease outbreaks is critical for the deployment of effective public health interventions. Delayed interventions may result in severe consequences including loss of lives. In addition to reactivity, the coverage of epidemiological event detection systems is of paramount importance, particularly because outbreaks are reported from different parts of the world in different languages. Taking this into account, multi- and cross-lingual computational approaches are relevant solutions, referred to in this paper as event-based surveillance systems. A key requirement for the development of such systems is the availability of large multi-lingual datasets to train and evaluate high-performance machine learning models. Such large and multi-lingual datasets are not readily available especially for epidemiological surveillance settings. In this study, we attempt to contribute towards solving this challenge by developing and making available a large multi-lingual dataset suitable for training and evaluating epidemiological event extraction models. The dataset can also be used for other natural language processing tasks such as text classification or text summarization, among others.

7. Acknowledgments

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 825153 (Embeddia) and 770299 (NewsEye).

8. Bibliographical References

- Aramaki, E., Maskawa, S., and Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.
- Bernardo, T. M., Rajic, A., Young, I., Robiadek, K., Pham, M. T., and Funk, J. A. (2013). Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research*, 15(7):e147.
- Bodnar, T. and Salathé, M. (2013). Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 699–702. Acm.
- Brownstein, J. S., Freifeld, C. C., Reis, B. Y., and Mandl, K. D. (2008). Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, 5(7):e151.
- Bunker, R. P. and Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied computing and informatics*.
- Charles-Smith, L. E., Reynolds, T. L., Cameron, M. A., Conway, M., Lau, E. H., Olsen, J. M., Pavlin, J. A., Shigematsu, M., Streichert, L. C., Suda, K. J., et al. (2015). Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10):e0139701.
- Chunara, R., Freifeld, C. C., and Brownstein, J. S. (2012). New technologies for reporting real-time emergent infections. *Parasitology*, 139(14):1843–1851.
- Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., et al. (2008). Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.
- Collier, N., Son, N. T., and Nguyen, N. M. (2011). Omg u got flu? analysis of shared health messages for bio-surveillance. *Journal of biomedical semantics*, 2(5):S9.
- Collier, N. (2011). Towards cross-lingual alerting for bursty epidemic events. *Journal of Biomedical Semantics*, 2(5):S10.
- Cooper, C. P., Mallon, K. P., Leadbetter, S., Pollack, L. A., and Peipins, L. A. (2005). Cancer internet search activity on a major search engine, united states 2001-2003. *Journal of medical Internet research*, 7(3):e36.
- Culotta, A. (2010). Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*.
- Diaz-Aviles, E., Stewart, A., Velasco, E., Denecke, K., and Nejdil, W. (2012). Epidemic intelligence for the crowd, by the crowd. In *Sixth International AAAI Conference on Weblogs and Social Media*.
- Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., and Yangarber, R. (2011). Building support tools for russian-language information extraction. In *International Conference on Text, Speech and Dialogue*, pages 380–387. Springer.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.
- Hogenboom, F., Frasinca, F., Kaymak, U., and De Jong, F. (2011). An overview of event extraction from text. In *DeRiVE@ ISWC*, pages 48–57. Citeseer.

- Huff, A. G., Breit, N., Allen, T., Whiting, K., and Kiley, C. (2016). Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources. *Interdisciplinary perspectives on infectious diseases*, 2016.
- Lamb, A., Paul, M. J., and Dredze, M. (2013). Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.
- Le, C.-C., Prasad, P., Alsadoon, A., Pham, L., and Elchouemi, A. (2019). Text classification: Naïve bayes classifier with sentiment lexicon. *IAENG International Journal of Computer Science*, 46(2):141–148.
- Lejeune, G., Brixteel, R., Doucet, A., and Lucas, N. (2015). Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65(2):131–143.
- Li, J. and Cardie, C. (2013). Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.
- Misra, R. (2018). News category dataset, 06.
- O'Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics*, 101:15–22.
- Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., and Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific.
- Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448.
- Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora*. Ph.D. thesis, Masarykova univerzita, Fakulta informatiky.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Sadilek, A., Kautz, H., and Silenzio, V. (2012). Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.
- Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F., and Brownstein, J. S. (2013). Influenza a (h7n9) and the importance of digital epidemiology. *The New England journal of medicine*, 369(5):401.
- Vogels, T., Ganea, O.-E., and Eickhoff, C. (2018). Web2text: Deep structured boilerplate removal. In *European Conference on Information Retrieval*, pages 167–179. Springer.
- World Health Organization. (2014). Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version. Technical report, World Health Organization.
- Zhan, L. and Jiang, X. (2019). Survey on event extraction technology in information extraction research area. In *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (IT-NEC)*, pages 2121–2126. IEEE.
- Zhou, X., Ye, J., and Feng, Y. (2011). Tuberculosis surveillance by analyzing google trends. *IEEE transactions on biomedical engineering*, 58(8):2247–2254.

2 Alleviating Digitization Errors in Named Entity Recognition for Historical Documents

Alleviating Digitization Errors in Named Entity Recognition for Historical Documents

Emanuela Boros¹, Ahmed Hamdi¹, Elvys Linhares Pontes¹, Luis Adrián Cabrera-Diego¹,
Jose G. Moreno^{1,2}, Nicolas Sidere¹, and Antoine Doucet¹

¹ University of La Rochelle, L3i, F-17000, La Rochelle, France

{emanuela.boros, ahmed.hamdi, elvys.linhares.pontes, luis.cabrera-diego}@univ-lr.fr
{nicolas.sidere, antoine.doucet}@univ-lr.fr

² University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France

jose.moreno@irit.fr

Abstract

This paper tackles the task of named entity recognition (NER) applied to digitized historical texts obtained from processing digital images of newspapers using optical character recognition (OCR) techniques. We argue that the main challenge for this task is that the OCR process leads to misspellings and linguistic errors in the output text. Moreover, historical variations can be present in aged documents, which can impact the performance of the NER process. We conduct a comparative evaluation on two historical datasets in German and French against previous state-of-the-art models, and we propose a model based on a hierarchical stack of Transformers to approach the NER task for historical data. Our findings show that the proposed model clearly improves the results on both historical datasets, and does not degrade the results for modern datasets.

1 Introduction

With the emergence of large scale archives of digitized contents, the need for efficient preservation and accessibility of historical documents through appropriate technologies increased exponentially. At the same time, there is a growing interest in extracting relevant information from historical sources. In this paper, we address the named entity recognition (NER) task which aims at identifying real-world entities, such as names of people, organizations, and locations within historical documents.

Since most of the state-of-the-art research focuses on NER for modern available datasets, the performance of the NER systems grew at a fast pace, enabled by the representational capacity of neural networks and off-the-shelf pre-trained word embeddings (Ma and Hovy, 2016; Lample et al., 2016; Yadav and Bethard, 2018). More recently,

NER models based on contextual word and sub-word representations provided by ELMo (Peters et al., 2018), Flair (Akbi et al., 2018), or BERT (Devlin et al., 2019), achieved impressive improvements. The Transformer-based (Vaswani et al., 2017) architectures for NER became popular since the release of the BERT (Bidirectional Encoder Representations from Transformers) model.

However, while most NER systems have been developed to generally address contemporary data, NER systems for processing historical documents are less common. To extract entities from historical documents, NER tools face additional challenges. As the majority of these documents are hardcover, they are scanned and processed by an OCR to transcribe the text. However, an OCR tool can occasionally misrecognize letters and improperly identify its textual content. This can be due to the level of degradation of the actual document being scanned, to the digitization artifacts and also to the quality of the OCR tool. This leads to digitization errors in the transcribed text, such as misspelled locations or person names.

Languages evolve through time and certain words can have a different meaning depending on the period of time analyzed (Hamilton et al., 2016). The spelling of words can also change due to new orthographic conventions or cultural tendencies (Scheible et al., 2011). This high level of spelling differences can be incompatible with modern orthography and the produced noise can severely affect modern NLP systems (Lopresti, 2009).

To address these challenges of NER on historical documents, we propose a robust NER model based on a stack of Transformers that includes fine-tuned BERT encoders. We study the impact of such a model, and we conclude that this type of model is suited for the extraction of entities from historical documents.

The remainder of this paper is organized as follows. In Section 2, we present and discuss a selection of works concerning NER in modern and historical documents. Then, in Section 3, the datasets explored in this work are presented. The proposed model is detailed in Section 4. The experiments are described in Section 5. We present and discuss the obtained results in Section 6. Finally, Section 7 concludes this paper and hints at future work.

2 Related Work

NER for modern documents The first end-to-end systems for sequence labeling tasks are based on pre-trained word and character embeddings encoded either by a bidirectional Long Short Term Memory (BiLSTM) network or a Convolutional Neural Network (CNN) (Collobert et al., 2011; Lample et al., 2016; Ma and Hovy, 2016; Aguilar et al., 2017; Chiu and Nichols, 2016), along with a Conditional Random Fields (CRF) decoder. One shortcoming of this type of model is that they were based on a single context-independent representation for each word. This problem has been further attenuated by methods based on language model pre-training that produced context-dependent word representations. These recent large-scale language models methods such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) further enhanced the performance of NER, yielding state-of-the-art performances (Peters et al., 2017, 2018; Baevski et al., 2019).

NER for historical documents Historical documents pose multiple challenges that either depend on the quality of digitization or the historical variations of a language. Studies on how the NER models can be impacted by the digitization process (Miller et al., 2000; Rodriguez et al., 2012; Hamdi et al., 2019; van Strien et al., 2020) have clearly shown that the performance scores of a NER model can significantly decrease when applied on historical documents.

The increased interest in contributing to historical language resources is driven forward by the creation of new gold standards for historical document processing. For example, Hubková (2019) created and annotated a corpus using scanned Czech historical newspapers, and Ahmed et al. (2019) proposed a German gold standard for NER in historical biodiversity literature.

A recent competition organized by the *Identifying Historical People, Places, and other Entities*

(HIPE) lab at CLEF 2020¹, not only that it created a gold standard for German and French historical texts, but also encouraged researchers to participate in two sub-tasks, named entity recognition and classification and entity linking.

Considering the high level of spelling differences between modern and historical documents, variance (inconsistency), and uncertainty (digitization errors) found in historical documents, the recent methods assess these shortcomings differently.

Erdmann et al. (2016) presented a CRF-based model with handcrafted features for Latin historical texts and motivated the choice of Part-of-Speech (POS) tagger by the fact that this NLP tool leverages the highly informative morphological complexity of Latin. The BiLSTM-based model proposed by Hubková (2019) applied a character-based CNN to encode the different spellings of words.

Similar to the latter approach, we also consider that the NER model itself can help in alleviating the historical documents issues, without the use of language-specific engineered features. Differently, we introduce the NER for historical documents to the language model methods based on the Transformer architecture (Vaswani et al., 2017) and BERT (Devlin et al., 2019) methods, that, to our knowledge, have not been approached in previous research, with regard to processing historical documents.

With new needs and resources in the context of historical NER processing, we evaluate our proposed model on the dataset proposed by the HIPE competition, and we also propose a new gold standard for German and French, to assess our assumptions.

3 Datasets

We conduct experiments on two datasets that comprise digitized historical newspapers, HIPE and NEWSEYE datasets in French and German. Additionally, we study how the proposed methods behave in the case of contemporary data, by experimenting on the English CoNLL 2003 dataset (Tjong Kim Sang and De Meulder, 2003).

The HIPE dataset was created by the CLEF 2020 Evaluation Lab HIPE challenge (Ehrmann et al., 2020a). It is composed of articles from several Swiss, Luxembourgish, and American historical newspapers from 1790 to 2010 (Ehrmann et al.,

¹impresso.github.io/CLEF-HIPE-2020/

2020b). More concisely, the German articles were collected from 1790 to 1940, and the French articles, from 1790 to 2010. The corpus was manually annotated by natives following the annotation guidelines derived from the Quaero annotation guide².

We also present the NEWS EYE dataset, composed of historical newspapers in French (1814-1944) and German (1845-1945). The documents were collected through the national libraries of France³ (BnF) and Austria⁴ (ONB), respectively. This dataset was annotated following guidelines derived from the Quaero annotation guide⁵. The annotation process was made by native speakers for each language using the Transkribus tool⁶. In order to compute the inter-annotator agreement (IAA), we used the Kappa coefficient introduced by Cohen (1960). Several pages from each corpus (German and French) have been annotated twice by two groups of annotators. Satisfactory IAA scores were reached for the two corpora (0.90 for French and 0.91 for German). The NewsEye corpus is split into 80% for training and 20% for both validation and testing.

The CoNLL 2003 dataset consists of newswire from the Reuters RCV1 corpus and it includes standard train, development, and test sets.

Table 1 presents the statistics regarding the number and type of entities in the aforementioned datasets. The statistics are divided according to the training, development, and test sets.

4 Model

We based our NER model on the pre-trained model BERT proposed by Devlin et al. (2019). Although original recommendations suggest that unsupervised pre-training of BERT encoders are expected to be sufficiently powerful on modern datasets, we consider that adding extra Transformer layers could contribute to the alleviation of word errors or misspellings.

First, we use a pre-trained BERT model, and second, we stack n Transformer blocks on top, finalized with a CRF prediction layer. We refer to this model as BERT+ $n \times$ Transf where n is a hyper-

²Quaero guidelines

³<https://www.bnf.fr>

⁴<https://www.onb.ac.at/>

⁵The main difference is that several named entities subtypes were ignored. In addition, the TIME type was not included in the annotation of the NEWS EYE dataset.

⁶<https://transkribus.eu/Transkribus/>

	Type	FR			DE		
		train	dev	test	train	dev	test
HIPE	LOC	3,067	664	854	1,747	771	595
	ORG	833	172	130	358	158	130
	PERS	2,513	428	502	1,170	677	311
	PROD	198	53	61	112	48	62
	TIME	273	73	53	118	69	49
NEWS EYE	LOC	4,878	522	698	4,024	525	894
	ORG	1,602	142	229	3,171	307	252
	PERS	5,023	853	788	2,346	424	461
	PROD	185	57	23	43	12	16
	Type	EN					
		train	dev	test			
CoNLL-03	LOC	7,140	1,837	1,668			
	ORG	6,321	1,341	1,661			
	PERS	6,600	1,842	1,617			
	MISC	3,438	922	702			

Table 1: Overview of the HIPE, NEWS EYE, and CoNLL 2003 datasets statistics. LOC = Location, ORG = Organization, PERS = Person, PROD = Product, TIME = Time and MISC = Miscellaneous.

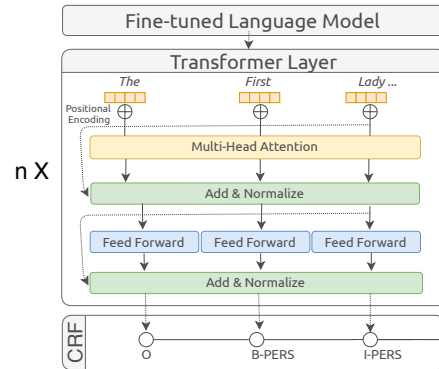


Figure 1: Main architecture of the BERT+ $n \times$ Transf.

parameter referring to the number of Transformer layers. The global architecture of our model is depicted in Figure 1. We used Transformer blocks with parameters that we chose empirically similar to the configuration of the blocks in the fine-tuned model⁷.

The reasons for using BERT models are that they can easily be fine-tuned for a wide range of tasks, but also that they produce high-performing systems (Devlin et al., 2019; Conneau and Lample, 2019; Radford et al., 2018). Nonetheless, despite the major impact of BERT in the NLP community, re-

⁷Note that they can vary as multiple BERT-based models are available for different languages.

searchers question the ability of this model to deal with noisy text (Sun et al., 2020) unless complementary techniques are used (Muller et al., 2019; Pruthi et al., 2019).

More specifically, the built-in tokenizer of BERT first performs simple white-space tokenization, then applies a Byte Pair Encoding (BPE) based tokenization, WordPiece (Wu et al., 2016). For example, word can be split into character n -grams (e.g. compatibility \rightarrow 'com', '##pa', '##ti', '##bilty'), where ## is a special symbol for representing the presence of a sub-word that was recognized.

Between the types of OCR errors that can be encountered in historical documents, the character insertion modification has the minimum influence (Sun et al., 2020), because the tokenization at the sub-word level of BERT would not change much in some cases, such as 'practically' \rightarrow 'practicaally'. Meanwhile, the substitution and deletion errors can hurt the performance of the tokenizer the most due to the generation of uncommon samples, such as 'professionalism' \rightarrow 'pr9fessi9nalism' that is tokenized as 'pr', '##9', '##fes', '##si', '##9', '##nal', '##sm'. BERT has been demonstrated to have a sensitivity to its sub-word segmentation when it comes to such words, as the meaning of the sub-words can diminish the initial meaning of the correctly spelled word (Sun et al., 2020). Thus, these new noisy tokens could influence the performance of BERT-based models⁸.

On top of BERT, we add a stack of Transformer blocks (encoders). A Transformer block (encoder), as proposed in (Vaswani et al., 2017), is a deep learning architecture based on multi-head attention mechanisms with sinusoidal position embeddings. It is composed of a stack of identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by layer normalization. All sub-layers in the model, as well as the embedding layers, produce outputs of dimension 512. In our implementation, we used learned absolute positional embeddings (Gehring et al., 2017) instead, as it is a common practice⁹. Vaswani et al.

⁸To increase the chances for misspelled, non-canonical, or new words to be recognized, we enrich the vocabulary of the tokenizer with these tokens, while allowing not only the BERT encoder but also the added Transformer layers to learn them from scratch.

⁹<https://huggingface.co/>

(2017) found that the two versions produced nearly identical results.

We assume that the additional Transformer layers can alleviate the sensitivity of the built-in tokenizer of BERT towards OOV, OCR errors, or misspellings, and contribute to the learning or finding the proper informative words around entities.

5 Experiments

5.1 Baseline

We chose as a baseline the model proposed by Ma and Hovy (2016), an end-to-end model combining a BiLSTM and a CNN character encoding, in order to take advantage of the word and character features. The character-level features are known to capture morphological and shape information (Kanaris et al., 2007; Santos and Zadrozny, 2014; dos Santos and Guimarães, 2015) that can also offer the possibility of obtaining a representation for misspelled, custom, or abnormal words. For the baseline, we used the FastText¹⁰ pre-trained word embedding models (Grave et al., 2018)¹¹.

Additionally, we analyze the aid that can be brought by an available larger dataset by training the baseline model in two stages in a transfer learning setting, similar to the setting in which the BERT encoder is used in our model:

1. *pre-training*, where the network is trained on a larger-scale available contemporary dataset
2. *fine-tuning*, where the pre-trained network is further trained on the historical datasets

The modern datasets are the following:

- For French, we use the fr-WikiNER¹² dataset that is extracted from Wikipedia articles. It contains about 500k tokens from which around 31k are named entities.
- For German, we use the de-GermEval¹³ dataset generated from German Wikipedia and News Corpora as a collection of citations. The dataset covers over 31k sentences corresponding to over 590k tokens from which around 33k are named entities.

¹⁰<https://fasttext.cc/docs/en/crawl-vectors.html>

¹¹For a more detailed description of the model and of the hyperparameters can be found in Ma and Hovy (2016).

¹²https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500

¹³<https://sites.google.com/site/germeval2014ner/data>

5.2 Metrics

The evaluation of the NER task is done in a coarse-grained manner, with the entity (not token) as the unit of reference (Makhoul et al., 1999). We compute precision (P), recall (R), and F1 measure (F1) at micro-level, i.e. error types are considered over all documents. Two evaluation scenarios were considered: *micro-strict*, which looks for an exact boundary matching, and *micro-fuzzy*, where a prediction is correct when there is at least one token overlap (Ehrmann et al., 2020a). Further, statistical significance is measured through a two-tailed t-test, with an estimated p-value between 0.01 and 0.05.

5.3 Data Pre-processing

The HIPE dataset was initially segmented at the article-level. Since BERT is able to consume only a limited context of tokens as their input (512), we segment the articles at sentence-level. We also reconstruct the original text, including hyphenated words. The reconstructed text was passed through Freeling 4.1 (Padró and Stanilovsky, 2012) to obtain a segmentation based on sentences. We made use of the same segmentation for the baseline model. Moreover, for the BERT+ $n \times$ Transf, we feed the model with batches of same sized inputs.

5.4 Hyperparameters

The hyperparameters used for both models are depicted as follows.

For the German NER, we chose as a pre-trained encoder the `bert-base-german-europeana`. This BERT model has been used in other NER tasks for processing contemporary and historical German documents (Schweter and Baiter, 2019; Riedl and Padó, 2018). It was trained using a large collection of newspapers provided by the Europeana Library.¹⁴

For the French NER, we rely on the large version of the pre-trained CamemBERT (Martin et al., 2020) model, i.e. (`camembert-large`). This model was trained on a large French corpus. CamemBERT proposes some differences with respect to other BERT models. For instance, it uses whole-word masking and SentencePiece tokenization (Kudo and Richardson, 2018) instead of WordPiece tokenization (Wu et al., 2016) as the original BERT.

For the English dataset CoNLL, we experimented with both `bert-base-cased` and

`bert-large-cased`, pre-trained models presented in (Devlin et al., 2019).

We denote the number of layers (i.e., Transformer blocks) as L , the hidden size as H , and the number of self-attention heads as A . `bert-base-cased` has $L=12$, $H=768$, $A=12$, `bert-large-cased` and `camembert-large`, $L=24$, $H=1024$, $A=16$. In all the cases, the top Transformer blocks have $L=1$ for $1 \times$ Transf and $L=2$ for $2 \times$ Transf, $H=128$, $A=12$, chosen empirically. The BERT-based encoders are fine-tuned on the task during training.

For training, we followed the selection of parameters presented in (Devlin et al., 2019). We found that 2×10^{-5} learning rate and a mini-batch of dimension 4 for German and English, and 2 for French, provide the most stable and consistent convergence across all experiments as evaluated on the development set.

6 Results

In this section, we provide experimental results of the baseline model and the proposed method. In order to assess the ability of both models with regard to the presence of errors provided by an OCR, we present several experiments:

- In Table 2, the first two experiments are performed with the baseline model, with and without the pre-training proposed by the transfer learning method on larger contemporary datasets.
- It is necessary to analyze how sensitive the proposed model is to the number of Transformer layers, the hyper-parameter n . Therefore, we conduct two experiments for ablation study with the n value $\in \{0, 1, 2\}$. The values > 2 obtained lower performance results and had a tendency to overfit. Therefore, in the same Table 2, we present next these experiments.
- In Table 3, the results for the baseline model without any transfer learning (as it was unnecessary) are presented, along with the same ablation study for the BERT+ $n \times$ Transf.

From the results in the Table 2, we can see the evidence that the BERT-based models with $n \times$ Transf achieve, for both datasets and languages, higher *micro-fuzzy* and *micro-strict* performance values than the BERT model stand-alone and the baseline

¹⁴<http://www.europeana-newspapers.eu/>

HIPE							NEWSEYE					
DE				FR			DE			FR		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BiLSTM-CNN												
fuzzy	83.3	70.1	76.1	89.9	83.9	86.8	81.2	42.4	55.7	82.2	77.2	79.6
strict	69.4	58.4	63.4	77.7	72.5	75.0	54.8	28.6	37.6	65.5	61.4	63.4
BiLSTM-CNN (transfer learning) [†]												
fuzzy	81.1	75.0	77.9**	87.8	88.8	88.3	76.4	49.4	60.0**	83.6	77.8	80.6*
strict	67.4	62.2	64.7**	77.3	78.2	77.7	48.6	31.4	38.1**	66.9	62.3	64.5*
BERT												
fuzzy	83.4	88.3	85.8**	89.5	91.9	90.7*	60.1	67.0	63.4**	86.1	81.8	83.9**
strict	74.1	78.5	76.2**	81.1	83.3	82.1*	46.8	52.2	49.4**	70.1	66.6	68.3**
BERT+1×Transf												
fuzzy	85.8	87.3	86.5**	91.3	92.9	92.1**	82.3	66.4	73.5**	88.7	82.1	85.3**
strict	77.2	78.6	77.9**	83.5	84.9	84.2**	62.7	50.6	56.0**	74.4	68.9	71.5**
BERT+2×Transf												
fuzzy	87.0	87.2	87.1**	91.5	92.4	91.9**	83.3	64.4	72.6**	89.7	80.1	84.7
												**
strict	78.6	78.7	78.7**	83.4	84.2	83.8**	64.9	50.2	56.6**	75.0	67.0	70.8**

Table 2: NER test results for the HIPE and NEWS EYE datasets in French and German. All models have as a decoder layer a CRF. [†]= with pre-training on larger modern datasets. All metrics are micro. Statistical significance is measured through a two-tailed t-test. * denotes a significant improvement over the BiLSTM model at $p \leq 0.05$, ** denotes $p \leq 0.01$.

models. All models have a statistical significance < 0.01 , thus, adding $n \times \text{Transf}$ can improve model generalizability for NER on historical documents.

Moreover, they generally manage to maintain a balance between recall and precision, while the baseline models vary, depending on the language. We also notice that, while in general, both models obtain a more or less precision-recall balance, there are two cases where there is a large imbalance, more specifically in the NEWS EYE German dataset. Comparing with the baseline models, the BERT+ $n \times \text{Transf}$ only achieves a 20 percentage points difference between precision and recall, while the baseline suffers from 40 points difference.

In the context of transfer learning applied for the baseline models, two performance results, for NEWS EYE in German, and for HIPE in French are higher due to the fine-tuning on these datasets, while the others are not degraded by the pre-training on larger contemporary datasets. This observation confirms the previous studies done on this type of model regarding their robustness to misspellings (Sun et al., 2020; Pruthi et al., 2019). We also notice that for German both datasets, the results for transfer learning from contemporary Ger-

man datasets are statistically significant ($< 0.01\%$), while contemporary datasets the performance difference for both French datasets was minimal (either < 0.5 for French NEWS EYE or < 0.9 for French HIPE).

CoNLL-03						
EN						
	P		R		F1	
BiLSTM-CNN						
micro-fuzzy	91.0		89.7		90.4	
micro-strict	89.2		87.9		88.5	
	P	R	F1	P	R	F1
	bert-base-cased			bert-large-cased		
BERT						
micro-fuzzy	91.7	93.0	92.3	92.4	93.5	92.9
micro-strict	90.3	91.6	90.9	91.1	92.2	91.6
BERT+1×Transf						
micro-fuzzy	92.5	93.2	92.8	92.7	93.4	93.1
micro-strict	91.1	91.8	91.4	91.4	92.1	91.8
BERT+2×Transf						
micro-fuzzy	92.0	93.2	92.6	92.9	93.4	93.1
micro-strict	90.6	91.8	91.2	91.6	92.1	91.8

Table 3: NER test results for the CoNLL 2003 dataset. All models have as a decoder layer a CRF.

Gold Standard	<p> LOC ALLEMAGNE que pas PER iJKeaz Schwietz, le bour-jeu de LOC Br eslasi _Piappi les pliante fameux de PER Reindel _figurjal la femme _JVie & e, l'horrible mégère de LOC Hambourg, qui assassina une vingtaine d'enfents confiés à _ses soins mercenaires. </p>
BERT	<p> LOC Allemagne que pas LOC iJKeaz Schwietz, le bour-jeu de LOC Br eslasi _Piappi les pliante fameux de PER Reindel _figurjal la femme _JVie & e, l'horrible mégère de LOC Hambourg, qui assassina une vingtaine d'enfents confiés à _ses soins mercenaires. </p>
BERT+ $n \times$ transf	<p> LOC Allemagne que pas PER iJKeaz Schwietz, le bour-jeu de LOC Br eslasi _Piappi les pliante fameux de PER Reindel _figurjal la femme _JVie & e, l'horrible mégère de LOC Hambourg, qui assassina une vingtaine d'enfents confiés à _ses soins mercenaires. </p>
Gold Standard	<p> LOC Amiens werde zwar im Augenblick noch gehalten, aber der Entwicklungsangriff von LOC Lille aus, also vo'i dem toten Punkte, der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vor auszusehen. </p>
BERT	<p> LOC Amiens werde zwar im Augenblick noch gehalten, aber der Entwicklungsangriff von LOC Lille aus, also vo'i dem toten Punkte, LOC der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vor auszusehen. </p>
BERT+ $n \times$ transf	<p> LOC Amiens werde zwar im Augenblick noch gehalten, aber der Entwicklungsangriff von LOC Lille aus, also vo'i dem toten Punkte, der leichter und rafcher die Zusammenziehung der Referven gestatte, sei vor auszusehen. </p>

Figure 2: An example of NER predictions on the HIPE dataset in French (top part) and German (bottom part).

In the context of modern data, in the Table 3, the F1 values of the stand-alone BERT model applied on the CoNLL 2003 dataset fairly correspond to the ones reported in (Devlin et al., 2019) (the authors report a F1 of 92.4% for *bert-base-cased* and 92.8% for *bert-large-cased*). While the F1 value has a very small margin difference from the (Devlin et al., 2019), the performance results for the BERT+ $n \times$ Transf slightly increased for both proposed models. We assume that one reason would be that the capacity of representation of extra Transformer layers, even in a context where no misspelling errors are present, can contribute to a modest improvement. While this improvement is more visible for the BERT *bert-base-cased*+1 \times Transf (a difference of a half of percentage point), and 0.3 percentage points for *bert-base-cased*+2 \times Transf, for the *bert-large-cased* BERT+ $n \times$ Transf, the values remain unchanged (with a difference of 0.2 percentage points from BERT).

6.1 Discussion

For more qualitative analysis, we examine the number of unrecognized words by the pre-trained BERT-based models that were added to the specific tokenizers (WordPiece for BERT and SentencePiece for CamemBERT). For NEWSEYE German, 8.84% of the total number of words in the vocabulary needed to be fully trained, while only 0.14% were unknown in the HIPE dataset. Following this observation, we notice that there is a large F1 margin between BERT+CRF and BERT+ $n \times$ Transf (63.4% in comparison with 73.5% and 72.6%, respectively), a fact that could be motivated by the large percentage of unknown words.

Moreover, for German, even though the BERT encoder was pre-trained on a digitized historical dataset (*bert-base-german-europeana*), the proposed model contributed greatly to the coverage of the misspelled or abnormal words present in the NEWSEYE. For French, the results vary of around 1 – 2 percentage F1 points between the stand-alone BERT and the BERT+ $n \times$ Transf models.

Between the two datasets, only HIPE was also annotated with the Levenshtein Ratio between the gold standard entities and the transcribed ones. In Figure 3, we compare BERT and BERT+ $n \times$ Transf by analyzing the number of correct predicted entities with respect to the Levenshtein distance. For the French predictions, for 56.25% of the different values of the distance, the stacked models had relatively more correct predictions. A French example of a misspelled entity that is recognized by both BERT+ $n \times$ Transf but not by BERT is presented in Figure 2, in the upper part. For German, only in 18.75% of the cases, the stacked models have more correctly identified entities that are misrecognized.

We also presume that the introduction by the stacked Transformers of additional hyperparameters can increase the ability of the architecture to better model long-range contexts. Thus, we analyzed the correctly predicted German and French HIPE entities by their length. We noticed that BERT+ $n \times$ Transf is better than BERT at predicting entities composed of multiple tokens (large entities). For example, for French HIPE, from 170 entities with a length equal or higher than five tokens¹⁵, the stand-alone BERT managed to correctly detect 70% of them, while both BERT+ $n \times$ Transf models

¹⁵The length of French HIPE entities ranges from one to 21 tokens.

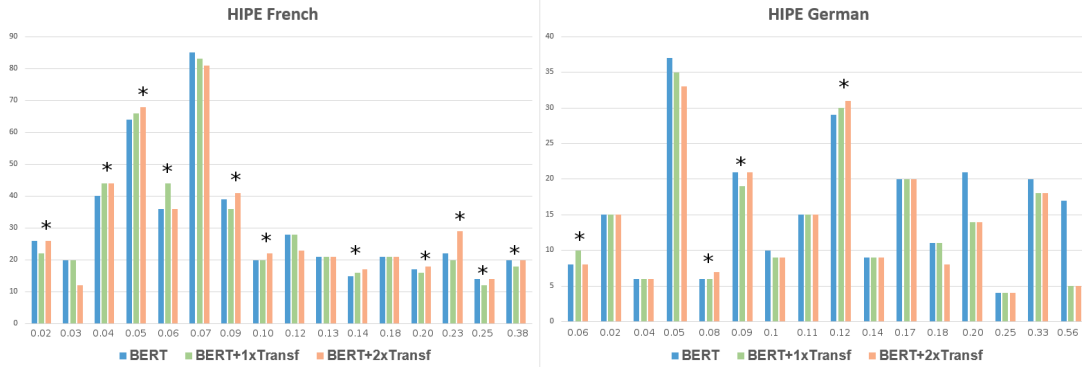


Figure 3: Correct predictions of misspelled entities based on the Levenshtein Ratio.

correctly identified 72.94% of them. German HIPE has less entities longer than five tokens¹⁶, more exactly 97, and while the stand-alone BERT detected 50.51% of them, the BERT+ $n \times$ Transf models correctly detected and classified 55.67% for $n = 1$ and 54.63% for $n = 2$. In the following examples from Table 4, our method correctly predicted the full entity frequently while the stand-alone BERT only predicted a part of it.

Analyzing the French predictions for BERT and BERT+ $n \times$ Transf, we observed that BERT detects on average 75.04% of the entities of size 1 to 10, with other models performing slightly better. However, for entities with more than 10 tokens, there is clear a difference, since BERT detects 55.54% of the entities, while BERT+1 \times Transf detects 57.13%, and BERT+2 \times Transf reaches 82.52%. Examples are given in Table 4.

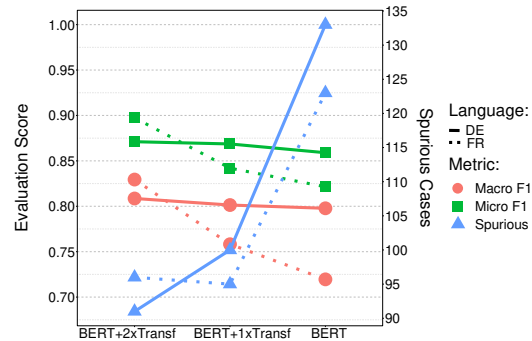
Gold standard			Predicted by	
			BERT	BERT+ $n \times$ Transf
signéKochH,	avocat	, avocat	signéKochH,	avocat
district de Gumbinnen		Gumbinnen	district de	Gumbinnen
Armél Guerne.	Armél Guerne		Armél Guerne.	son adaptateur
M. Javits, sénateur de New York juif et pro- israélien	M. Javits, sénateur de New York		M. Javits, sénateur de New York	juif et pro- israélien

Table 4: Examples of long entities predicted by all models (the entity parts detected by BERT alone are highlighted in bold font under BERT+ $n \times$ Transf).

In the lower part of Figure 2, we present a German example where BERT becomes confused and

¹⁶The length of German HIPE entities ranges from one to 16 tokens.

predicts multiple partial spurious entities in a sentence. One can also observe that these entities are of two of the most common types in the dataset, persons (PERS) and locations (LOC). In this case, there is an overprediction of these types, which leads us to the interpretation that BERT is sensitive to misspellings and might overfit on OCR-related patterns. This observation proves that BERT has unbalanced attention to misspelled or corrupted words when the most informative words contain such errors (Sun et al., 2020).

Figure 4: Number of spurious entities with respect to *micro-fuzzy* and *macro-fuzzy* F1 regarding the HIPE corpus.

To assess these assumptions, in Figure 4, we compare, per model and language, the values of *micro-fuzzy* F1 and *macro-fuzzy* F1 in the HIPE corpus. We include, as well, the number of spurious cases, i.e. tokens that were considered as an entity, despite not belonging to one, such as 'Zusammenziehung' in Figure 2.¹⁷ Due to the difference between *micro* and *macro* metrics, we can

¹⁷We obtained the spurious cases by searching for predicted named entities that did not correspond, partially or totally, to one in the gold standard.

ascertain that the three presented models focused on predicting the most frequent entity types, i.e. PERS and LOC. Moreover, we can see that BERT achieved its result by creating more spurious cases in comparison to BERT+ $n \times$ Transf. This could mean that BERT learned that overpredicting was a straightforward solution to achieve better results. In the case of BERT+ $n \times$ Transf, we can see that the Transformer layers made the models to be more conservative and at the same time more accurate in their predictions.

7 Conclusions and Future Work

We presented a deep learning architecture for NER based on stacked Transformer layers that includes a fine-tuned BERT encoder and several Transformer blocks. Results on two historical datasets in French and German showed the fitness of the proposed model to process noisy digitized text corpora in distinct languages. At the same time, the approach did not degrade the performance over modern data. Thus, this type of model appears to be adapted for the NER of historical document collections.

While the improvements brought by the proposed NER model are clear, our analysis of the results highlighted several factors that could influence the results. Further analysis remains to be done. Thus, hereafter, we will investigate detailed variations of our architecture. In addition, we intend to explore data augmentation techniques, simulating digitized data by adding noise to digitally-born documents. This could be a solution to increase the size and expand the diversity of training datasets for performing NLP tasks over historical documents.

Acknowledgments

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Tamar Solorio. 2017. [A multi-task approach for named entity recognition in social media data](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Sajawel Ahmed, Manuel Stoeckel, Christine Driller, Adrian Pachzelt, and Alexander Mehler. 2019. [BIOfid Dataset: Publishing a German Gold Standard for Named Entity Recognition in Historical Biodiversity Literature](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 871–880, Hong Kong, China. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. [Cloze-driven pretraining of self-attention networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China. Association for Computational Linguistics.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Alexis Conneau and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Matteo Romanello, Stefan Bircher, and Simon Clematide. 2020a. Introducing the clef 2020 hipe shared task: Named entity recognition and linking on historical newspapers. In *European Conference on Information Retrieval*, pages 524–532. Springer.

- Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. 2020b. Language resources for historical newspapers: the impresso collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 958–968.
- Alex Erdmann, Christopher Brown, Brian D Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *COLING 2016: 26th International Conference on Computational Linguistics*, pages 85–93. Association for Computational Linguistics.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML 17*, page 1243–1252. JMLR.org.
- Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334. IEEE.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. *Diachronic word embeddings reveal statistical laws of semantic change*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Helena Hubková. 2019. *Named-entity recognition in Czech historical texts: Using a CNN-BiLSTM neural network model*. Ph.D. thesis.
- Ioannis Kanaris, Konstantinos Kanaris, Ioannis Houvardas, and Efstathios Stamatatos. 2007. Words versus character n-grams for anti-spam filtering. *International Journal on Artificial Intelligence Tools*, 16(06):1047–1067.
- Taku Kudo and John Richardson. 2018. *SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. *Neural architectures for named entity recognition*. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Daniel Lopresti. 2009. Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJ DAR)*, 12(3):141–151.
- Xuezhe Ma and Eduard Hovy. 2016. *End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.
- John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252. Herndon, VA.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. *CamemBERT: a tasty French language model*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. Named entity extraction from noisy input: speech and ocr. In *Proceedings of the sixth conference on Applied natural language processing*, pages 316–324. Association for Computational Linguistics.
- Benjamin Muller, Benoît Sagot, and Djamé Seddah. 2019. Enhancing bert for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306.
- Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, pages 2473–2479, Istanbul, Turkey. ELRA.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

- Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 5582–5591, Florence, Italy.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Martin Riedl and Sebastian Padó. 2018. A named entity recognition shootout for german. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125.
- Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. [Comparison of named entity recognition tools for raw OCR text](#). In *11th Conference on Natural Language Processing, KONVENS 2012, Empirical Methods in Natural Language Processing, Vienna, Austria, September 19-21, 2012*, volume 5 of *Scientific series of the ÖGAI*, pages 410–414. ÖGAI, Wien, Österreich.
- Cícero dos Santos and Victor Guimarães. 2015. [Boosting named entity recognition with neural character embeddings](#). In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China. Association for Computational Linguistics.
- Cícero dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.
- Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. [Evaluating an ‘off-the-shelf’ POS-tagger on early modern German text](#). In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–23, Portland, OR, USA. Association for Computational Linguistics.
- Stefan Schweter and Johannes Baiter. 2019. [Towards robust named entity recognition for historic German](#). In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.
- Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Ksra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks.
- Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-bert: Bert is not robust on misspellings! generating nature adversarial samples on bert. *arXiv preprint arXiv:2003.04985*.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Vikas Yadav and Steven Bethard. 2018. [A survey on recent advances in named entity recognition from deep learning models](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

3 CTLR@WiC-TSV: Target Sense Verification using Marked Inputs and Pre-trained Models

CTRL@WiC-TSV: Target Sense Verification using Marked Inputs and Pre-trained Models

Jose G. Moreno

University of Toulouse
IRIT, UMR 5505 CNRS
F-31000, Toulouse, France
jose.moreno@irit.fr

Elvys Linhares Pontes

University of La Rochelle
L3i
F-17000, La Rochelle, France
elvys.linhares.pontes@univ-lr.fr

Gaël Dias

University of Caen
GREYC, UMR 6072 CNRS
F-14000, Caen, France
gael.dias@unicaen.fr

Abstract

This paper describes the CTRL participation in the Target Sense Verification of the Words in Context challenge (WiC-TSV) at SemDeep-6. Our strategy is based on a simplistic annotation scheme of the target words to later be classified by well-known pre-trained neural models. In particular, the marker allows to include position information to help models to correctly identify the word to disambiguate. Results on the challenge show that our strategy outperforms other participants (+11,4 Accuracy points) and strong baselines (+1,7 Accuracy points).

1 Introduction

This paper describes the CTRL¹ participation at the Word in Context challenge on the Target Sense Verification (WiC-TSV) task at SemDeep-6. In this challenge, given a target word w within its context participants are asked to solve a binary task organised in three sub-tasks:

- *Sub-task 1* consists in predicting if the target word matches with a given *definition*,
- *Sub-task 2* consists in predicting if the target word matches with a given *set of hypernyms*, and
- *Sub-task 3* consists in predicting if the target word matches with a given couple *definition* and *set of hypernyms*.

Our system is based on a masked neural language model with position information for Word Sense Disambiguation (WSD). Neural language models are recent and powerful resources useful for multiple Natural Language Processing (NLP) tasks (Devlin et al., 2018). However, little effort

¹University of Caen Normandie, University of Toulouse, and University of La Rochelle team.

has been made to perform tasks, where positions represent meaningful information. Regarding this line of research, Baldini Soares et al. (2019) include markers into the learning inputs for the task of relation classification and Boualili et al. (2020) into an information retrieval model. In both cases, the tokens allow the model to carefully identify the targets and to make an informed prediction. Besides these works, we are not aware of any other text-based tasks that have been tackled with this kind of information included into the models. To cover this gap, we propose to use markers to deal with target sense verification task.

The remainder of this paper presents a brief background knowledge in Section 2. Details of our strategy, including input modification and prediction mixing is presented in Section 3. Then, unofficial and official results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 Background

NLP research has recently been boosted by new ways to use neural networks. Two main groups of neural networks can be distinguished² on NLP based on the training model and feature modification.

- First, *classical neural networks* usually use pre-trained embeddings as input and models learn their own weights during training time. Those weights are calculated directly on the target task and integration of new features or resources is intuitive. As an example, please refer to the Figure 1(a) which depicts the model from Zeng et al. (2014) for relation classification. Note that this model uses

²We are aware that our classification is arguable. Although this is not an established classification in the field, it seems important for us to make a difference between them as this work tries to introduce well-established concepts from the first group into the second one.

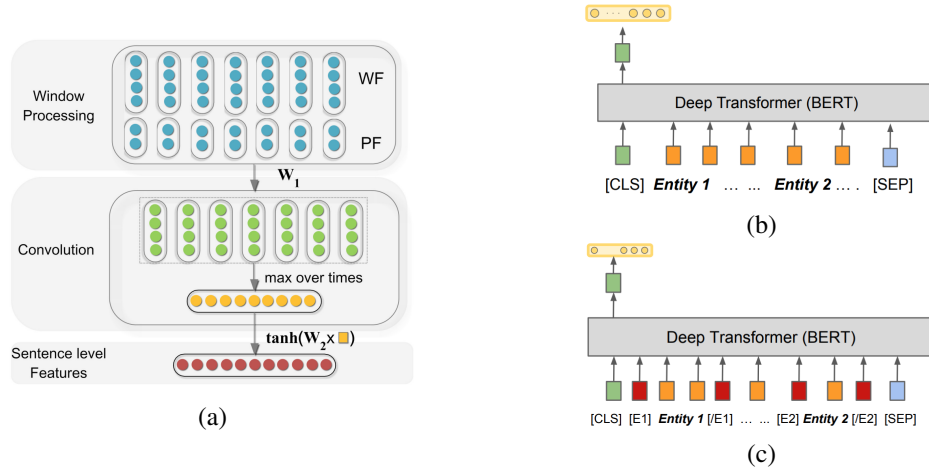


Figure 1: Representation examples for the relation classification problem proposed by Zeng et al. (2014) (a) and Baldini Soares et al. (2019) (b and c).

the positional features (PF in the figure) that enrich the word embeddings (WF in the figure) to better represent the target words in the sentence. In this first group, models tend to use few parameters because embeddings are not fine-tuned. This characteristic does not dramatically impact the model performances.

- The second group of models deals with *neural language models*³ such as BERT (Devlin et al., 2018). The main difference, w.r.t. the first group, is that the weights of the models are not calculated during the training step of the target task. Instead, they are pre-calculated in an elegant but expensive fashion by using generic tasks that deal with strong initialised models. Then, these models are fine-tuned to adapt their weights to the target task.⁴ Figure 1(b) depicts the model from Devlin et al. (2018) for the sentence classification based on BERT. Within the context of neural language models, adding extra features like PF demands re-train of the full model, which is highly expensive and eventually prohibitive. Similarly, re-train is needed if one opt for adding external information as made for recent works such as KNOW+E+E (Peters et al., 2019) or SenseBERT (Levine et al., 2019).

We propose an alternative to mix the best of both worlds by including extra tokens into the input in

³Some subcategories may exist.

⁴We can imagine a combination of both, but models that use BERT as embeddings and do not fine-tune BERT weights may be classified in the first group.

order to improve prediction without re-training it. To do so, we base our strategy on the introduction of signals to the neural language models as depicted in Figure 1(c) and done by Baldini Soares et al. (2019). Note that in this case the input is modified by introducing extra tokens ($[E1]$, $[/E1]$, $[E2]$, and $[/E2]$ are added based on target words (Baldini Soares et al., 2019)) that help the system to point out the target words. In this work, we mark the target word by modifying the sentence in order to improve performance of BERT for the task of target sense verification.

3 Target Sense Verification

3.1 Problem definition

Given a first sentence with a known target word, a second sentence with a definition, and a set of hypernyms, the target sense verification task consists in defining whether or not the target word in the first sentence corresponds to the definition or/and the set of hypernyms. Note that two sub-problems may be set if only the second sentence or the hypernyms are used. These sub-problems are presented as sub-tasks in the WiC-TSV challenge.

3.2 CTRLR method

We implemented a target sense verification system as a simplified version⁵ of the architecture proposed by Baldini Soares et al. (2019), namely $BERT_{EM}$. It is based on BERT (Devlin et al., 2018), where an extra layer is added to make the

⁵We used the *EntityMarkers[CLS]* version.

classification of the sentence representation, i.e. classification is performed using as input the [CLS] token. As reported by Baldini Soares et al. (2019), an important component is the use of mark symbols to identify the entities to classify. In our case, we mark the target word in its context to let the system know where to focus on.

3.3 Pointing-out the target words

Learning the similarities between a couple of sentences (sub-task 1) can easily be addressed with BERT-based models by concatenating the two inputs one after the other one as presented in Equation 1, where S_1 and S_2 are two sentences given as inputs, $t_i^1 (i = 1..n)$ are the tokens in S_1 , and $t_j^2 (j = 1..m)$ are the tokens in S_2 . In this case, the model must learn to discriminate the correct definition and also to which of the words in S_1 the definition relates to.

$$\begin{array}{l} \text{input}(S_1, S_2) = \\ \begin{array}{llll} [\text{CLS}] & t_1^1 & t_2^1 & \dots & t_n^1 \\ [\text{SEP}] & t_1^2 & t_2^2 & \dots & t_m^2 \end{array} \end{array} \quad (1)$$

To avoid the extra effort by the model to evidence the target word, we propose to introduce this information into the learning input. Thus, we mark the target word in S_t by using a special token before and after the target word⁶. The input used when two sentences are compared is presented in Equation 2. S_t is the first sentence with the target word t_i , S_d is the definition sentence, and t_x^k are their respective tokens.

$$\begin{array}{l} \text{input}^{\text{sp1}}(S_t, S_d) = \\ \begin{array}{llllll} [\text{CLS}] & t_1^t & t_2^t & \dots & \$ & t_i^t & \$ & \dots & t_n^t \\ [\text{SEP}] & t_1^d & t_2^d & \dots & t_m^d \end{array} \end{array} \quad (2)$$

In the case of hypernyms (sub-task 2), the input on the left side is kept as in Equation 2, but the right side includes the tagging of each hypernym as presented in Equation 3.

$$\begin{array}{l} \text{input}^{\text{sp2}}(S_t, S_h) = \\ \begin{array}{llllll} [\text{CLS}] & t_1^t & t_2^t & \dots & \$ & t_i^t & \$ & \dots & t_n^t \\ [\text{SEP}] & s_1^h & \$ & s_2^h & \$ & \dots & \$ & s_l^h \end{array} \end{array} \quad (3)$$

3.4 Verifying the senses

We trained two separated models, one for each sub-problem using the architecture defined in Section 3.2. The output predictions of both models are

⁶We used '\$' but any other special token may be used.

used to solve the two-tasks problem. So, our overall prediction for the main problem is calculated by combining both prediction scores. First, we normalise the scores by applying a *softmax* function to each model output, and then we select the prediction with the maximum probability as shown in Equation 5.

$$\text{pred}(x) = \begin{cases} 1, & \text{if } m_1^{\text{sp1}}(x) + m_1^{\text{sp2}}(x) \\ & > m_0^{\text{sp1}}(x) + m_0^{\text{sp2}}(x). \end{cases} \quad (4)$$

$$\text{where } m_i^{\text{spk}} = \frac{\exp(p_i^{\text{spk}})}{\sum_{j=\{0,1\}} \exp(p_j^{\text{spk}})} \quad (5)$$

and p_i^{spk} is the prediction value for the model k for the class i (m_i^{spk}).

4 Experiments and Results

4.1 Data Sets

The data set was manually created by the task organisers and some basic statistics are presented in Table 1. Detailed information can be found in the task description paper (Breit et al., 2020). No extra-annotated data was used for training.

	train	development	test
Positive	1206	198	-
Negative	931	191	-
Total	2137	389	1324

Table 1: WiC-TSV data set examples per class. Positive examples are identified as 'T' and negative as 'F' in the data set.

4.2 Implementation details

We implemented $BERT_{EM}$ of Baldini Soares et al. (2019) using the huggingface library (Wolf et al., 2019), and trained two models with each training set. We selected the model with best performance on the development set. Parameters were fixed as follows: 20 was used as maximum epochs, *Cross Entropy* as loss function, Adam as optimiser, *bert-base-uncased*⁷ as pre-trained model, and other parameters were assigned following the library recommendations (Wolf et al., 2019). The final layer is composed of two neurons (negative or positive).

⁷<https://github.com/google-research/bert>

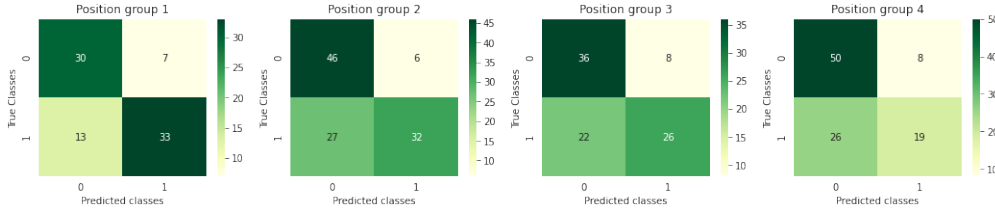


Figure 2: Confusion matrices for different position groups. Group 1 (resp. 2, 3, and 4) includes all sentences for which the target word appears in the first (resp. second, third, and fourth) quarter of the sentence.

4.3 Results

As the test labels are not publicly available, our following analysis is performed exclusively on the development set. Results on the test set were calculated by the task organisers.

We analyse confusion matrices depending on the position of the target word in the sentence as our strategy is based on marking the target word. These matrices are presented in Figure 2. The confusion matrix labelled as position group 1 shows our results when the target word is in the first 25% positions of the S_t sentence. Other matrices show the results of the remaining parts of the sentence (second, third, and fourth 25%, for respectively group 2, 3, and 4).

Confusion matrices show that the easiest cases are when the target word is located in the first 25%. Other parts are harder mainly because the system considers positive examples as negatives (high false negative rate). However, the system behaves correctly for negative examples independently of the position of the target word. To better understand this wrong classification of the positive examples, we calculated the true label distribution depending on the normalised prediction score as in Figure 3. Note that positive examples are mainly located on the right side but a bulk of them are located around the middle of the figure. It means that models m^{sp1} and m^{sp2} were in conflict and average results were slightly better for the negative class. In the development set, it seems important to correctly define a threshold strategy to better define which examples are marked as positive.

In our experiments, we implicitly used 0.5 as threshold⁸ to define either the example belongs to the ‘T’ or ‘F’ class. When comparing Figures 3 and 4, we can clearly see that small changes in the threshold parameter would affect our results with

⁸Because of the condition $m_1^{sp1}(x) + m_1^{sp2}(x) > m_0^{sp1}(x) + m_0^{sp2}(x)$.

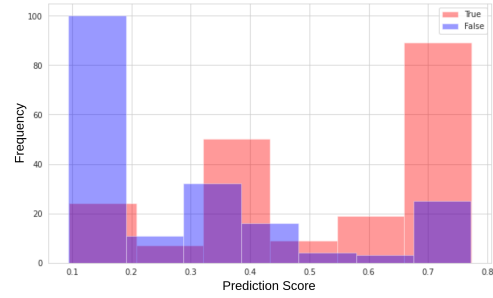


Figure 3: Histograms of predicted values in the dev set.

a larger impact in recall than in precision. This is mainly given to the fact that our two models contradict for some examples.

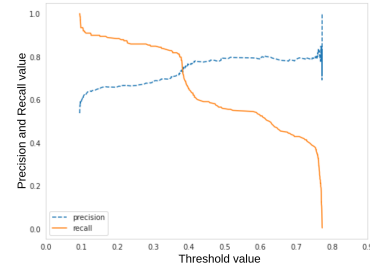


Figure 4: Precision/Recall curve for the development set for different threshold values.

We also considered the class distribution depending on a normalised distance between the target token and the beginning of the sentence. From Figure 5, we observe that both classes are less frequent at the beginning of the sentence with negative examples slightly less frequent than positive ones. It is interesting to remark that negative examples uniformly distribute after the first bin. On the contrary, the positive examples have a more unpredictable distribution indicating that a strategy based on only positions may fail. However, our strategy that combines markers to indicate the target word and a

Run	User	Global				WordNet/Wiktionary				Cocktails				Medical entities				Computer Science			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
run2	CTLR (ours)	78.3	78.9	78.0	78.5	72.1	75.8	70.7	73.2	87.5	82.4	90.3	86.2	85.9	86.7	85.8	86.3	83.3	78.4	88.5	83.1
szte	begab	66.9	61.6	92.5	73.9	70.2	66.5	89.6	76.4	55.1	48.9	96.8	65.0	65.4	60.5	95.3	74.0	70.2	61.3	97.4	75.2
szte2	begab	66.3	61.1	92.8	73.7	69.9	66.2	90.2	76.3	53.7	48.1	96.8	64.3	64.4	59.8	95.3	73.5	69.6	60.8	97.4	74.9
BERT	-	76.6	74.1	82.8	78.2	73.5	76.1	74.2	75.1	79.2	67.8	98.2	80.2	79.8	75.8	89.6	82.1	82.1	73.0	97.9	83.6
FastText	-	53.4	52.8	79.4	63.4	57.1	58.0	74.0	65.0	43.1	43.1	100.0	60.2	51.1	51.5	90.3	65.6	54.0	50.5	67.1	57.3
Baseline (true)	-	50.8	50.8	100.0	67.3	53.8	53.8	100.0	70.0	43.1	43.1	100.0	60.2	51.7	51.7	100.0	68.2	46.4	46.4	100.0	63.4
Human	-	85.3	80.2	96.2	87.4	82.1	-	-	-	92.0	-	-	-	89.1	-	-	-	86.5	-	-	-

Table 2: Accuracy, Precision, Recall and F1 results of participants and baselines. Results were split by type. General results are included in column ‘Global’. All results were calculated by the task organisers (Breit et al., 2020) as participants have not access to test labels. Best performance for each global metric is marked in **bold** for automatic systems.

strong neural language model (BERT) successfully manage to classify the examples.

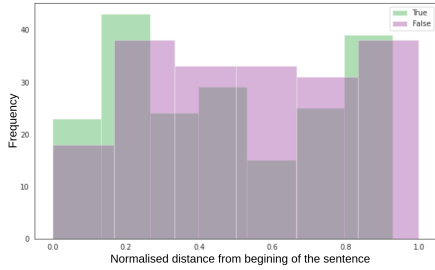


Figure 5: Position distribution based on the target token distances.

Finally, the main results calculated by the organisers are presented in Table 2. The global column presents the results for the global task, including definitions and hypernyms. Our submission is identified as run2-CTLR. In the global results, our strategy outperforms participants and baselines in terms of Accuracy, Precision, and F1. Best Recall performance is unsurprisingly obtained by the baseline (true) that corresponds to a system that predicts all examples as positives. Two strong baselines are included, FastText and BERT. Both baselines were calculated by the organisers with more details in (Breit et al., 2020). It is interesting to remark that the baseline BERT is very similar to our model but without the marked information. However, our model focuses more on improving Precision than Recall resulting with a clear improvement in terms of Accuracy but less important in terms of F1.

Organisers also provide results grouped by different types of examples. They included four types with three of them from domains that were not included in the training set⁹. From Table 2, we can also conclude that our system is able to adapt

⁹More details in (Breit et al., 2020).

to out-of-domain topics as it is clearly shown for the *Cocktails* type in terms of F1, and also for the *Medical entities* type to a less extent. However, our system fails to provide better results than the standard BERT in terms of F1 for the *Computer Science* type. But, in terms of Accuracy, our strategy outperforms for a large margin the out-of-domain types (8.3, 6.1, and 1.2 improvements in absolute points for *Cocktails*, *Medical entities*, and *Computer Science* respectively). Surprisingly, it fails on both, F1 and Accuracy, for *WordNet/Wiktionary*.

5 Conclusion

This paper describes our participation in the WiC-TSV task. We proposed a simple but effective strategy for target sense verification. Our system is based on BERT and introduces markers around the target words to better drive the learned model. Our results are strong over an unseen collection used to verify senses. Indeed, our method (Acc=78, 3) outperforms other participants (second best participant, Acc=66, 9) and strong baselines (BERT, Acc=76, 6) when compared in terms of Accuracy, the official metric. This margin is even larger when the results are compared for the out-of-domain examples of the test collection. Thus, the results suggest that the extra information provided to the BERT model through the markers clearly boost performance.

As future work, we plan to complete the evaluation of our system with the WiC dataset (Pilehvar and Camacho-Collados, 2019) as well as the integration of the model into a recent multi-lingual entity linking system (Linhares Pontes et al., 2020) by marking the anchor texts.

Acknowledgements

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grant 825153 (EMBEDIA).

References

- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *ACL*. 2895–2905.
- Lila Boualili, Jose G. Moreno, and Mohand Boughanem. 2020. MarkedBERT: Integrating Traditional IR Cues in Pre-Trained Language Models for Passage Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, 1977–1980.
- Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2020. WiC-TSV: An Evaluation Benchmark for Target Sense Verification of Words in Context. *arXiv:2004.15016* (2020).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805* (2018).
- Yoav Levine, Barak Lenz, Or Dagan, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some Sense into Bert. *arXiv:1908.05646* (2019).
- Elvys Linhares Pontes, Jose G. Moreno, and Antoine Doucet. 2020. Linking Named Entities across Languages Using Multilingual Word Embeddings. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*. Association for Computing Machinery, 329–332.
- Matthew E. Peters, Mark Neumann, Robert L. Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP*. 43–54.
- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1267–1273.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771* (2019).
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2335–2344.

4 Dataset for Temporal Analysis of English-French Cognates

Dataset for Temporal Analysis of English-French Cognates

Esteban Frossard[♣] Mickaël Coustaty[♣] Antoine Doucet[♣] Adam Jatowt[◇] Simon Hengchen[♣]

[♣]University of La Rochelle, L3i Laboratory, {firstname.lastname}@univ-lr.fr

[◇]Kyoto University, adam@dl.kuis.kyoto-u.ac.jp

[♣]University of Helsinki, simon.hengchen@helsinki.fi

Abstract

Languages change over time and, thanks to the abundance of digital corpora, their evolutionary analysis using computational techniques has recently gained much research attention. In this paper, we focus on creating a dataset to support investigating the similarity in evolution between different languages. We look in particular into the similarities and differences between the use of corresponding words across time in English and French, two languages from different linguistic families yet with shared syntax and close contact. For this we select a set of cognates in both languages and study their frequency changes and correlations over time. We propose a new dataset for computational approaches of synchronized diachronic investigation of language pairs, and subsequently show novel findings stemming from the cognate-focused diachronic comparison of the two chosen languages. To the best of our knowledge, the present study is the first in the literature to use computational approaches and large data to make a cross-language diachronic analysis.

Keywords: Crosslingual semantic change, cognates, temporal analysis, semantic analysis

1. Introduction

Languages, our main tools of communication, evolve constantly: words obtain new and lose old meanings over time, they become popular or fade into obscurity. Because of its importance, language is studied by academics and public alike, as shown by the large number of publications and websites devoted to language evolution, etymology and semantic changes (Cresswell, 2010; Ayto, 2011; Lewis, 2013). Most of these focus on individual words only or are done on a small scale, mainly because the analysis requires manual work to locate occurrences of features in old texts, and then to compare manually their contexts or other characteristics.

In the recent years, large amounts of digitized old books and texts were made available, such as Google's Books initiative (Michel et al., 2010) with 5% of books ever published. Computational approaches have also been conducted to analyze them (Gulordava and Baroni, 2011), proposing novel approaches for understanding lexical semantic change – for an overview, we refer to the survey by (Tahmasebi et al., 2018). However, to the best of our knowledge, no cross-language temporal analysis has been proposed in the literature using computational approaches and large data. In addition, most prior studies focused only on English, whereas comparing two or more languages can shed light on how they actually co-evolved over time.

To study multiple languages over time, we assume the most intuitive approach: we focus on their similar connecting aspects. We use in particular words in both languages that have the same origins and similar meaning, also known as cognate words. We propose to study the temporal characteristics of cognate words as an approach to cross-language diachronic analysis. These cognates, loanwords included (i.e., words that come directly from the other languages) are an important subset of the lexicon and have been frequently studied. Most prior works focused on synchronous analysis of cognates (see for example (Uban et al., 2019)), while we look at their temporal aspects and correlations.

We have used the largest multilingual corpora available on a relatively long time, allowing thanks to its size to set a yearly granularity of analysis. In particular, we used Google Books Ngrams¹ in English and French to conduct the analysis. Despite its inherent problems (Pechenick et al., 2015), it is one of the few corpora of this size available in both French and English. We also prepared a list of English-French cognates based on existing lists and few selection criteria described below.

Cognates are, in linguistics, words that share a common etymological origin (Crystal, 2011), of which loanwords (words borrowed from other languages, e.g. English *communiqué* is borrowed from the French) are particular cases. Both are of great interest in multi-language analysis thanks to the ease of understanding and the identification of links between languages.

Numerous works have focused on either cognates or loanwords. On the one hand there are works for cognate detection harnessing computational methods that propose the first step in a (semi-) automatic analysis of cognates using the vast amount of digitally available data, when manual annotation requires a lot of man-hours (Jäger et al., 2017; List et al., 2018). On the other hand there are semantic analyses of cognates, that manually investigate cognates to look for links between two different languages (List et al., 2018; Aske, 2015). Some recent works cope with the limitations of these two categories by mixing the use of automatic detection of cognates with the semantic analysis (List et al., 2018; Rabinovich et al., 2018).

Nevertheless, to the best of our knowledge, there has been no automatic study of the frequency correlations and patterns of cognates over time across different languages, especially one that uses large datasets. In this paper, we propose a statistical change-oriented analysis of cognates, and focus on English and French.

¹<https://books.google.com/ngrams>, accessed on November 15, 2019

2. Datasets

We started the study of English-French cognate by constructing a large cognate dataset that fits our criteria (see Section 2.1.). First, we created a list of cognates applicable for our study, basing our selection on available English and French lists of cognates (Bergsma and Kondrak, 2007), removing those that did not fit our criteria and adding some other. Each word's "cognateness" was confirmed by investigating its etymology with the Oxford English Dictionary, the on-line etymology dictionary² and the French National Center for Textual and Lexical Resources (FR: *Centre National de Ressources Textuelles et Lexicales*).

We used the 1-gram from the Google Books n-grams, for English and French (Michel et al., 2010) as an underlying dataset. It contains around half a trillion English words and one hundred billion French ones coming from books of varying literature genres. We note that although the dataset is not balanced in terms of document types its strong advantage lies in the very large size in comparison to other similar datasets, both in number of words and periods covered (from the 1500s to the late 2000s).

Finally, we would like to mention that we first focus on the differences in use frequency of words over time, hence we chose Google Books 1-grams. However, the underlying dataset can be easily extended by using larger n-grams such as 5-grams.

2.1. Criteria for Selecting Words

We chose English-French word pairs for constructing the cognates dataset and we based the selection on four criteria as follow. (1) We restricted the time scope to the years from 1800 to 2008, where most of the data is. (2) We chose words that were cognate pairs based on their etymology to make sure they were actual cognates. (3) We discarded verbs as their many inflections in French introduce noise, mostly as shared surface forms with other lexical items. (4) Finally, we chose words that appeared above a minimal frequency threshold (one in two million, or from 35 to 10,000 appearances in a single year, depending on the number of words available for that year) in both English and French to allow a proper analysis and to minimize the chance of an erroneous detection.

Once all words were selected, every inflection of each word was found using dedicated dictionaries. The frequency of all forms of a word were summed for each year to compute the total frequency of the word for that year. We then obtained for each word a time series from 1800 to 2008 representing its frequency. Finally, for each word, the time series, year of the first appearance, the maximum frequency and its year are all stored in a text file.

2.2. Cognates Dataset

Based on the data and the criteria presented above, we built, and release, a cognate dataset with 492 word pairs composed of nouns, adjectives and adverbs³. Each pair has between one and four forms in English, and up to ten in

French. In English, most words have only one form for adjectives and adverbs, while most nouns have two forms (singular and plural). In French, with masculine and feminine, singular and plural forms, most nouns and adjectives can be found in four different surface forms.

The dataset includes 353 (71%) French loanwords (French words used in English) and 15 (3%) English loanwords⁴. These numbers include words taken from Old French and Old English. Note that the words are eclectic, both in meaning, as we aimed not to bias the dataset to any topic, and in frequency, as shown in Figure 1 where we plot median frequency as well as quartiles.

In the end, the dataset contains, for each cognate, both in English and in French, its frequency all inflexions combined in each year from 1800 to 2008 (0 in years before they appear or they are not part of the dataset).

3. Temporal Analysis of Cognates

We present below the preliminary results of the frequency analysis using the constructed cognate dataset.

3.1. Correlation of Cognates

First, we wanted to examine if the level of use of words in each of the languages changed in their own way or, rather, if the cognates shared similar patterns of changes in the intensity of their use over time. We then started by computing the frequency correlation for each pair of cognates. We used Pearson correlation coefficient (Pearson, 1895) on the time series representing cognate use in the concerned period. The frequency of a term in a given year is computed by dividing the number of occurrences of the term (the sum of the number of occurrences of each of its forms) by the total number of summed appearances of all words in this year.

As shown on Figure 2, there was a strong positive correlation for most pairs, with more than half (57%, 281) having a correlation value above 0.5, and over 13% (65) above 0.9. However, the high positive correlation is not true for every pair, as correlations go from -0.87 for the pair *employee* – *employé* to 0.99 for the pair *traditionally* – *traditionnellement*. Nevertheless, the number of pairs with a negative correlation, or close to zero, is rather small, as shown on Figure 2. This suggests that *cognates do not only share a past (etymological roots), but they also share similar usage patterns over time*.

Most of the cognate pairs had correlated changes of frequency over time. On the left of Figure 2, negatively correlated words are quite rare (6%, 31 words below -0.3). This suggests that cases when cognate words have tendencies to change the frequency of their use in an opposite way are quite rare.

If we restrict the analysis to the French loanwords (see the red plot in Figure 2), the positive correlation is similar, 201 loanwords (57%) having a correlation above 0.5 with their counterpart and 46 (13%) having the correlation value above 0.9.

²Available online at <https://www.etymonline.com/>

³Available online at <https://zenodo.org/record/3688087>.

⁴Due to the small number of English loanwords, we will focus only on French loanwords in our analysis.

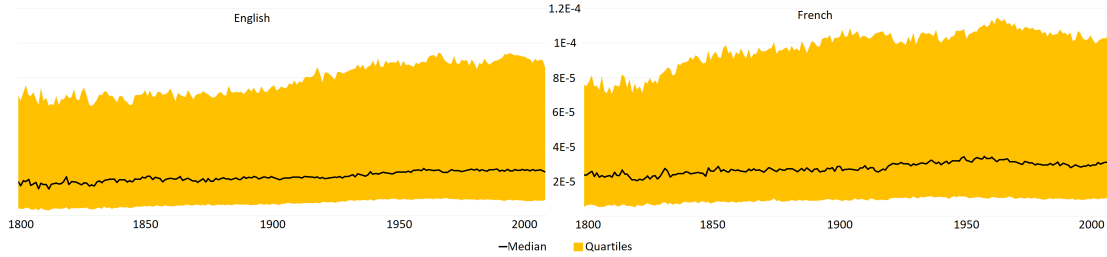


Figure 1: Distribution of the frequencies of cognates pairs, expressed through the quartiles and median.

3.2. Level of Word Use

The correlation of fluctuations in word frequencies over time as studied above still does not tell us whether words were actually used at the similar intensity levels in the same years. One word in a cognate pair could be used very frequently, while its counterpart could be barely used even though their relative frequency changes over time may be correlated.

To compare whether the frequency of a word is similar to its cognate counterpart, we first looked at the ratio between their maximal and mean frequencies. Then, for a cognate pair (w_E, w_F) , with $f_E(w, y)$ and $f_F(w, y)$ denoting the frequency (respectively, in English and French) of the word w in year y , we computed the following formula:

$$\frac{\max(\max_{y \in [1800; 2008]} f_E(w_E, y), \max_{y \in [1800; 2008]} f_F(w_F, y))}{\min(\max_{y \in [1800; 2008]} f_E(w_E, y), \max_{y \in [1800; 2008]} f_F(w_F, y))}$$

This equation gives a real number of one or greater and is based on the comparison of the maximum frequencies of cognates. The closer to one, the greater the similarity between the maximum frequencies of the two cognates, with the limit at one where both the values (maximum frequency in English and maximum frequency in French between 1800 and 2008) being equal. When the resulting value is higher, the two words in a given cognate pair have a less similar use.

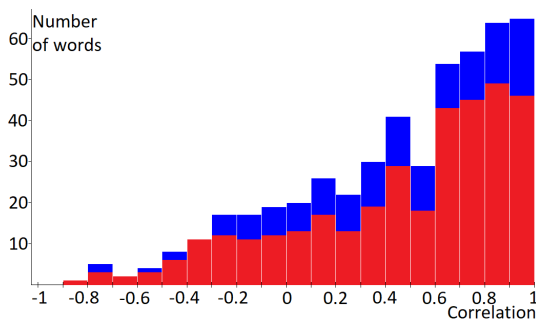


Figure 2: Correlation of English-French cognate pairs (blue) and French loanwords (red), from the first appearance of a word (English or French, depending on the earliest one) to 2008, as including earlier years would artificially increase correlation.

The cognate words not only tend to be correlated in terms of their changes over time, but they also have (for most of them) a similar level of use in their languages. The maximum usage of the most used word in each cognate pair is, for more than half of the words, at most 1.63 times more than its counterpart in the other language.

Moreover, the more we focus on the correlated words, the smaller this median line is (1.53 for correlation above 0.5; 1.49 for correlation above 0.7; 1.48 for correlation above 0.9). If we analyze only the loanwords, the results are similar.

To see if this ratio changes according to the frequency in one or both languages, and if one language has the cognates consistently more used (especially interesting are outliers), their respective mean frequencies seem to follow a linear distribution (see Figure 3). However, there are also cases of high frequency of use of a cognate in one language with low frequency in the other language (even several thousand times more in one language).

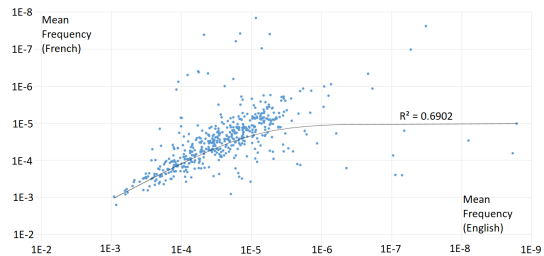


Figure 3: Distribution of the mean frequency in French according to the mean frequency in English (log-log plot). The linear regression $y = 1.1457x + 10^{-5}$ (black) shows the global relation between mean frequencies.

These extremes tend to be as likely to result from higher use in English as in French. As the correlation analysis indicated that the level of use of cognates evolved according to the same pattern across time, the frequency ratio indicates the *cognates have a similar level of use in both languages across time*.

3.3. Language Specificities

As the results show that cognate words are often used similarly at the same time in both the languages, one could be

tempted to say that a cognate, independently of language, performs in general a similar role in both languages and is used in very similar ways over time.

There are several potential reasons that could be proposed behind the differences in use frequencies and their temporal variations over time in both languages. To a certain degree, these could be explained by the subtle differences in the meaning of the cognates in both the languages, which would be used for slightly different purposes or in differing situations. Another driving force behind the observed differences in cognate use could be the existence of a synonym or multiple synonyms in only one of the two languages, which could “drain” the usage of one of the two words of the cognate pair: as per (Saussure, 1916), there is no bijective relationship between words in different languages.

Another explanation could be the occurrence of an additional acquired sense behind a cognate in one language increasing the use of this word with relation to its use in the other language. For example *azote* is barely used in English, in favor of *nitrogen*, while it is the opposite in French (*nitrogène* exists, yet *azote* is more commonly used).

3.4. Impact of External Factors

French and English are not only affected by each other, but by a multitude of external factors which can explain at least some of the correlations between cognates pairs, like the common history of corresponding countries. Analyzing history – i.e., the context around language use – can lead to an understanding of the impact of important events on some words, the most explicit example in our dataset being *bombardment* – *bombardement*, shown in Figure 4, a word which was obviously used more frequently in times of war, or, rather in the case of our corpus, when war-related books were popular. However, such effects are often difficult to determine, especially when the causes are less known.

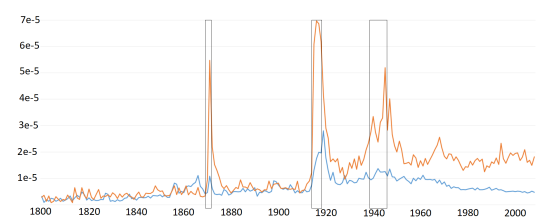


Figure 4: Frequency of *Bombardment* (English, in blue) and *Bombardement* (French, in orange) from 1800 to 2008. Three spikes can be observed (denoted by black rectangles), which correspond to the Franco-Prussian war (1870-1871), World War I (1914-1918) and World War II (1939-1945), showing the effect of the events on the languages.

4. Limitations

The dataset is not exempt from limitations, from its rather small size, as we focused on most-known cognates for the

first analysis, to potential bias coming from the choice of words, even if we did our best to limit it, or from the corpus choice. We also provide the results of preliminary frequency-focused analysis of the cognates based on the created dataset. The analysis itself has some limitations: as it only covers two well-known languages, English and French, and only by not taking into accounts synonyms that made some cognates out of use in one of the two languages.

5. Conclusions & Future Work

In this paper, we describe a dataset of English and French cognates constructed to study their evolution from 1800 to 2008.

Diachronic language analysis and in particular studies of word origins have recently attracted considerable attention. In this paper we also emphasized the idea of studying temporal variability of a language by its synchronized comparison with another language where the synchronization is based on using cognates (serving as a comparative “bridges”) aligned over time. By this, we add a second dimension or an additional investigation axis to the usual diachronic analysis approaches.

In the future, we plan to extend the current study to embrace larger number of cognates and to conduct a semantic analysis of the cognate variation across time and languages. We will also study other language pairs including ones that had less interaction and exchange in the past.

6. Acknowledgments

This work has been supported by the European Union Horizon 2020 research and innovation programme under grants 825153 (Embeddia) and 770299 (NewsEye).


7. Bibliographical References

- Aske, J. (2015). Spanish-English cognates: An introduction to Spanish linguistics. Open Access eBook (Open Textbook). CC BY-NC-ND 3.0 US. (version: 29 June 2018).
- Ayto, J. (2011). *Dictionary of Word Origins: The Histories of More Than 8,000 English-Language Words*. Arcade Publishing.
- Bergsma, S. and Kondrak, G. (2007). Alignment-based discriminative string similarity. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 656–663.
- Cresswell, J. (2010). *Oxford Dictionary of Word Origins*. Oxford University Press.
- David Crystal, editor. (2011). *A Dictionary of Linguistics and Phonetics (6th ed.)*. David Blackwell Publishing. p. 104, ISBN 978-1-4443-5675-5. OCLC 899159900.
- Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.
- Jäger, G., List, J.-M., and Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the*

- European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain, April. Association for Computational Linguistics.
- Lewis, D. (2013). *Now I Know: The Revealing Stories Behind the World's Most Interesting Facts*. Adams Media.
- List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., and Forkel, R. (2018). Sequence Comparison in Computational Historical Linguistics Phonetic Alignments and Cognate Detection with LingPy 2.6. *Journal of Language Evolution*.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2010). Quantitative analysis of culture using millions of digitized books. *Science*.
- Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of London*, pages 240–242.
- Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.
- Rabinovich, E., Tsvetkov, Y., and Wintner, S. (2018). Native language cognate effects on second language lexical choice. *CoRR*, abs/1805.09590.
- Saussure, F. d. (1916). *Cours de linguistique générale*, ed. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot.
- Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.
- Uban, A., Ciobanu, A. M., and Dinu, L. P. (2019). Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166.

5 Entity Linking for Historical Documents: Challenges and Solutions

Entity Linking for Historical Documents: Challenges and Solutions

Elvys Linhares Pontes¹() , Luis Adrián Cabrera-Diego¹, Jose G. Moreno²,
Emanuela Boros¹, Ahmed Hamdi¹, Nicolas Sidère¹, Mickaël Coustaty¹,
and Antoine Doucet¹

¹ University of La Rochelle, L3i, 17000 La Rochelle, France

{elvys.linhares_pontes,luis.cabrera_diego,emanuela.boros,
ahmed.hamdi,nicolas.sidere,mickael.coustaty,antoine.doucet}@univ-lr.fr

² University of Toulouse, IRIT, UMR 5505 CNRS, 31000 Toulouse, France
jose.moreno@irit.fr

Abstract. Named entities (NEs) are among the most relevant type of information that can be used to efficiently index and retrieve digital documents. Furthermore, the use of Entity Linking (EL) to disambiguate and relate NEs to knowledge bases, provides supplementary information which can be useful to differentiate ambiguous elements such as geographical locations and peoples' names. In historical documents, the detection and disambiguation of NEs is a challenge. Most historical documents are converted into plain text using an optical character recognition (OCR) system at the expense of some noise. Documents in digital libraries will, therefore, be indexed with errors that may hinder their accessibility. OCR errors affect not only document indexing but the detection, disambiguation, and linking of NEs. This paper aims at analysing the performance of different EL approaches on two multilingual historical corpora, CLEF HIPE 2020 (English, French, German) and NewsEye (Finnish, French, German, Swedish), while proposes several techniques for alleviating the impact of historical data problems on the EL task. Our findings indicate that the proposed approaches not only outperform the baseline in both corpora but additionally they considerably reduce the impact of historical document issues on different subjects and languages.

Keywords: Entity linking · Deep learning · Historical data · Digital libraries.

1 Introduction

Historical documents are an essential resource in the understanding of our cultural heritage. The development of recent technologies, such as optical character recognition (OCR) systems, allows the digitisation of physical documents and the extraction of the textual content. Digitisation provides two major advantages in

Digital Humanities: the exponential increase of target audiences, and the preservation of original documents from any damage when accessing them. The recent interest in massive digitisation raises multiple challenges to content providers including indexing, categorisation, searching, to mention a few. Although these challenges also exist when dealing with contemporary text documents, digitised version augments each challenge because of inherent problems associated with the source quality (natural degradation of the documents) and to the digitisation process itself (e.g., image quality and OCR bias).

While the number of works in natural language processing (NLP) and information retrieval (IR) domains concerning contemporary documents has known an important raise during the last decade, it has not been the case for historical documents. One of the main reasons is the additional difficulties that NLP and IR systems have to face regarding historical documents. For instance, tools need to know how to deal correctly with errors produced by OCR systems. Moreover, historical languages may contain a number of spelling variations with respect to modern languages, that might be difficult to recognise, as orthographic conventions can be reformed from time to time. Finally, some historic documents may also contain cases where the name of places is in a language different to the main text one. These particularities have then a significant impact on NLP and IR applications over historical documents.

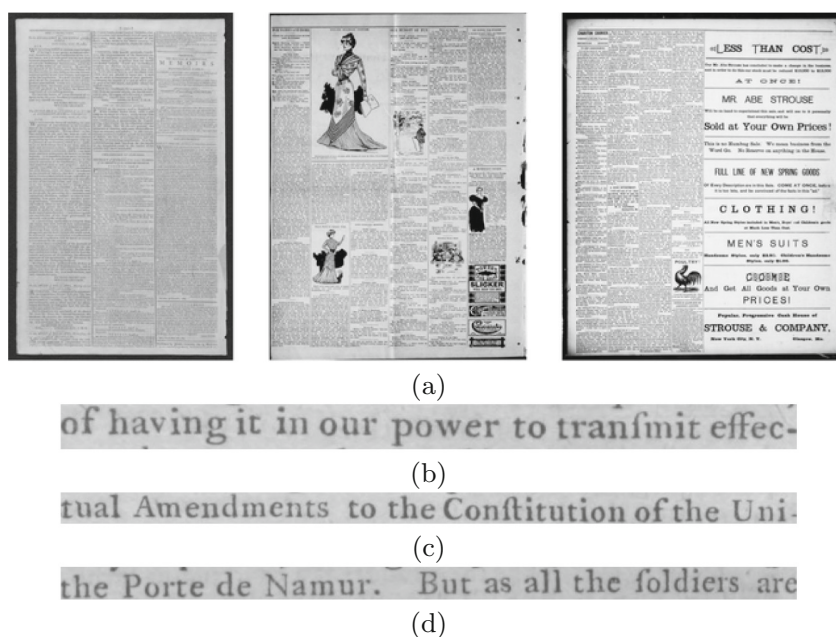


Fig. 1. Examples of historical documents from the Chronicling America newspapers used in CLEF HIPE 2020.

To illustrate some of the aforementioned problems, let us consider Fig. 1(a) which includes some English documents used in the evaluation campaign CLEF

HIPE 2020 [9]. Figure 1(b) and (c) are zoomed and cropped portions of most left document presented in Fig. 1(a). We can observe in these images a common characteristic found in multiple historical documents, the presence of a *Long S* (“*ſ*”), a character that is frequently confused by OCR systems for an “*l*” or “*f*” given its geometrical similarity. Figure 1(b) illustrates a case where the word “*tranſmit*” was recognised as “*tranlinit*” by a state-of-the-art OCR system.¹ Figure 1(c) illustrates a similar case where the word “*Conſtitution*” was recognised as “*Conftitution*”² which makes harder for an automatic system to recognise that this document concerns the *Constitution of the Unites States of America*³. In Fig. 1(d), we observe a case where an article uses the French name “*Porte de Namur*” to make reference to “*Namur Gate*”.⁴

Apart from digitising and recognising the text, the processing of historical documents consists as well on extracting metadata from these documents. This metadata is used to index the key information inside documents to ease the navigation and retrieval process. Among all the possible key information available, named entities are of major significance as they allow structuring the documents’ content [12]. These entities can represent aspects such as people, places, organisations, and events. Nonetheless, historical documents may contain duplicated and ambiguous information about named entities due to the heterogeneity and the mix of temporal references [13, 30]. A disambiguation process is thus essential to distinguish named entities to be further utilised by search systems in digital libraries.

Entity linking (EL) aims to recognise, disambiguate, and relate named entities to specific entries in a knowledge base. EL is a challenging task due to the fact that named entities may have multiple surface forms, for instance, in the case of a person an entity can be represented with their full or partial name, alias, honorifics, or alternate spellings [29]. Compared to contemporary data, few works in the state of the art have studied the EL task on historical documents [3, 4, 13, 16, 23, 28, 30] and OCR-processed documents [20].

In this paper, we present a deep learning EL approach to disambiguate entities on historical documents. We investigate the issues of historical documents and propose several techniques to overcome and reduce the impact of these issues in the EL task. Moreover, our EL approach decreases possible bias by not limiting or focusing the explored entities to a specific dataset. We evaluate our methods in two recent historical corpora, CLEF HIPE 2020 [9], and NewsEye datasets, that are composed of documents in English, Finnish, French, German, and Swedish. Our study shows that our techniques improve the performance of EL systems and partially solve the issues of historical data.

This paper is organised as follows: we describe and survey the EL task on historical data in Sect. 2. Next, the CLEF HIPE 2020 and NewsEye datasets are described in Sect. 3. We detail our multilingual approach in Sect. 4. Then the experiments and the results are discussed in Sects. 5 and 6. Lastly, we provide the conclusion and some final comments in Sect. 7.

¹ [HIPE-data-v1.3-test-masked-bundle5-en.tsv#L45-L53](#).

² [HIPE-data-v1.3-test-masked-bundle5-en.tsv#L56-L61](#).

³ https://en.wikipedia.org/wiki/Constitution_of_the_United_States.

⁴ [HIPE-data-v1.3-test-en.tsv#L1663-L1665](#).

2 Entity Linking for Historical Data

Entity linking (EL) is an information extraction task that semantically enriches documents by identifying pieces of text that refer to entities, and by matching each piece to an entry in a knowledge base (KB). Frequently, the detection of entities is delegated to an external named entity recognition (NER) system. Thus, in the state of the art, EL tools are either *end-to-end systems*, i.e. tools that perform both tasks, or *disambiguation systems* [11,18], i.e. tools that perform only the matching of entities and consider the first task as an input.

End-to-end EL systems were initially defined for contemporary documents [5]. First systems were focused on monolingual corpora and then gradually moved to a multilingual context. Some recent configuration, named Cross-Lingual Named Entity Linking (XEL), consist in analysing documents and named entities in a language different from the one used in the knowledge base. Some recent works proposed different XEL approaches: zero-shot transfer learning method by using a pivot language [27], hybrid approach using language-agnostic features that combine existing lookup-based and neural candidate generation methods [31], and the use of multilingual word embeddings to disambiguate mentions across languages [21].

Regarding the application of end-to-end EL in Digital Humanities, some works have focused on using available EL approaches to analyse historical data [16,23,28]. Other works have concentrated on developing features and rules for improving EL in a specific domain [13] or entity types [3,4,30]. Furthermore, some researchers have investigated the effect of issues frequently found in historical documents on the task of EL [13,20].

Some NER and EL systems dedicated to historical documents have also been explored [16,23,24,28]. For instance, van Hooland *et al.* [16] evaluated three third-party entity extraction services through a comprehensive case study, based on the descriptive fields of the Smithsonian Cooper-Hewitt National Design Museum in New York. Ruiz and Poibeau [28] used DBpedia Spotlight tool to disambiguate named entities on Bentham's manuscripts. Finally, Munnelly and Lawless [24] investigated the accuracy and overall suitability of EL systems in 17th century depositions obtained during the 1641 Irish Rebellion.

Most of the developed end-to-end EL systems are monolingual like the work of Mosallam *et al.* [22]. The authors developed a monolingual unsupervised method to recognise person names, locations, and organisations in digitised French journals of the National Library of France (*Bibliothèque nationale de France*) from the 19th century. Then, they used a French entity knowledge base along with a statistical contextual disambiguation approach. Interestingly, their method outperformed supervised approaches when trained on small amounts of annotated data. Huet *et al.* [17] also analysed the French journal *Le Monde*'s archive, a collection of documents from 1944 until 1986 discussing different subjects (e.g., post-war period, end of colonialism, politics, sports, culture). The authors calculated a conditional distribution of the co-occurrence of mentions with their corresponding entities (Wikipedia article). Then, they linked these Wikipedia articles to YAGO [26] to recognise and disambiguate entities in the archive of *Le Monde*.

Monolingual disambiguation systems have also been studied by focusing on specific types of entities in historical documents, e.g., person and place names. Smith and Crane [30] investigated the identification and disambiguation of place names in the Perseus digital library. They concentrated on representing historical data in the humanities from Ancient Greece to 19th century America. In order to overcome with the heterogeneous data and the mix of temporal references (e.g., places that changed their name through time), they proposed a method based on honorifics, generic geographic labels, and linguistic environments to recognise entities, while they made use of gazetteers, biographical information, and general linguistic knowledge to disambiguate these entities. Another work [3,4] focused on authors' names in French literary criticism texts and scientific essays from the 19th and early 20th centuries. They proposed a graph-based method that leverages knowledge from different linked data sources to generate the list of candidates for each author mention. Then, it crawls data from other linked data sets using equivalence links and fuses graphs of homologous individuals into a non-redundant graph in order to select the best candidate.

Heino *et al.* [13] investigated EL in a particular domain, the Second World War in Finland, using the reference datasets of WarSampo. They proposed a ruled-based approach to disambiguate military units, places, and people in these datasets. Moreover, they investigated problems regarding the analysis and disambiguation of these entities in this kind of data while they proposed specific rules to overcome these issues.

The impact of OCR errors on EL systems, to our knowledge, has rarely been analysed or alleviated in previous research. Thus, the ability of EL to handle noisy inputs continuous to be an open question. Nevertheless, Linhares Pontes *et al.* [20], reported that EL systems for contemporary documents can see their performance decreased around 20% when OCR errors, at the character and word levels, reach rates of 5% and 15% respectively.

Differently from previous works, we propose a multilingual end-to-end approach to link entities mentioned in historical documents to a knowledge base. Our approach contains several techniques to reduce the impact of the problems generated by the historical data issues, e.g., multilingualism, grammatical errors generated by OCR engines, and linguistic variation over time.

3 Historical Datasets

Unlike contemporary data that have multiple EL resources and tools, historical documents face the problem of lacking annotated resources. Moreover, contemporary resources are not suitable to build accurate tools over historical data due to the variations in orthographic and grammatical rules, not to mention the fact that names of persons, organisations, and places could have significantly changed over time.

To the best of our knowledge, there are few publicly available corpora in the literature with manually annotated entities on historical documents. Most EL corpora are composed of contemporary documents. Unfortunately, they do

not contain the distinctive features found in historical documents. In this work, we focus on two corpora that contain historical documents in English, Finnish, French, German, and Swedish.

The first corpus was produced for the CLEF HIPE 2020 challenge⁵ [8]. This corpus is composed of articles published between 1738 and 2019 in Swiss, Luxembourgish, and American newspapers. It was manually annotated by native speakers according to HIPE annotation guidelines [8].

Table 1. Number of entities for the training, development, and test sets in CLEF HIPE 2020 and NewsEye corpora.

Split	CLEF HIPE 2020			NewsEye			
	German	English	French	German	Finnish	French	Swedish
Training	3,505	–	6,885	–	1,326	–	1,559
Development	1,390	967	1,723	–	284	–	335
Test	1,147	449	1,600	7,349	287	5,090	337

The second corpus was produced for the Horizon 2020 NewsEye project⁶ and it is a collection of annotated historical newspapers in French, German, Finnish, and Swedish. These newspapers were collected by the national libraries of France⁷ (BnF), with documents from 1814 to 1944, Austria⁸ (ONB) with documents from 1845 to 1945, and Finland⁹ (NLF), with Finnish and Swedish documents from 1771 to 1910 and 1920, respectively.

Both corpora contain named entities that are classified according to their type and, when possible, linked to their Wikidata ID. Non-existent entities in the Wikidata KB are linked to NIL entries. Table 1 shows the statistics of the datasets for the training, development, and test partitions.

4 Multilingual End-to-end Entity Linking

As aforementioned, historical documents present particular characteristics that make challenging the use of EL. In the following subsections, we describe the methods and techniques we developed for creating an EL system that addresses these challenges.

⁵ <https://impresso.github.io/CLEF-HIPE-2020/>.

⁶ <https://www.newseye.eu>.

⁷ <https://www.bnf.fr>.

⁸ <https://www.onb.ac.at>.

⁹ <https://www.kansalliskirjasto.fi>.

4.1 Building Resources

By definition of the task, EL systems use knowledge bases (KB) as entry reference but their use is not limited to it. KBs are also used by EL systems for tasks such as extraction of supplementary contexts or surface names, disambiguation of cases, or linking of entities with a particular website entry. In the following paragraphs, we present the most representative KBs used in this domain.

Wikipedia¹⁰, a multilingual encyclopedia available in 285 languages, is commonly used as KB in the state-of-the-art. For instance, [11,18] make use of the English Wikipedia to disambiguate entity mentions in newspapers. Agirre *et al.*[1] used Wikipedia not only to disambiguate mentions found in historical documents but also to explore the feasibility of matching mentions with articles on Wikipedia according to their cultural heritage.

Wikidata¹¹ is a KB created by the Wikimedia Foundation¹² to store, in a structured way, data generated and used by the different Wikimedia projects, e.g., Wikipedia and Wiktionary. For instance, it has been used to annotate historical corpora, such as those used on this paper, CLEF HIPE 2020 and NewsEye.

DBpedia [19] is a KB that structures and categorise information collected from different Wikimedia projects, including Wikipedia and Wikidata, while including links to other KBs such as YAGO [26] or GeoNames¹³. For instance, it was used by [6] for annotating mentions of locations in *Historische Kranten*, a historical newspaper corpus. While [23] used DBpedia for annotating historical legal documents. Other examples of EL and DBpedia can be found in the works of [10,16].

In this work, we decided to build our own KB consisting of information from Wikipedia. Nevertheless, rather than just focusing on the English Wikipedia, we make use as well of the versions found in the languages used in the datasets to evaluate: French, German, Finnish, and Swedish. The reasoning behind this is that despite the richness and coverage of the English Wikipedia, on occasion other versions of Wikipedia might contain information that is only found in a specific language. For instance, *Valentin Simond*, owner of the French newspaper *L'Écho de Paris*, has an entry only in the French Wikipedia¹⁴.

4.2 Entity Embeddings

Based on the work of [11], we decided to create entity embeddings for each language by generating two conditional probability distributions. The first one, the “positive distribution”, is a probability approximation based on word-entity co-occurrence counts, i.e. which words appear in the context of an entity. The counts were obtained, in the first place, from the entity Wikipedia page, and,

¹⁰ <https://www.wikipedia.org>.

¹¹ <https://www.wikidata.org>.

¹² <https://www.wikimedia.org>.

¹³ <http://www.geonames.org>.

¹⁴ https://fr.wikipedia.org/wiki/Valentin_Simond.

in second place, from the context surrounding the entity in an annotated corpus using a fixed-length window. The second distribution, the “negative” one, was calculated by randomly sampling context windows that were unrelated to a specific entity. Both probability distributions were used to change the alignment of words embeddings with respect to an entity embedding. The positive probability distribution is expected to approach the embeddings of the co-occurring words with the embedding vector of the entity, while the negative probability distribution is used to distance the embeddings of words that are not related to an entity.

It should be noted that, unlike some works, where all the possible entities are known beforehand, in our work the creation of entity embeddings is not directed by a dataset. This is done to prevent bias and low generalisation. In case an entity does not have an entity embeddings, the EL system will propose a NIL.

4.3 Entity Disambiguation

The entity disambiguation model is based on the neural end-to-end entity linking architecture proposed by Kolitsas et al. [18]. The first advantage of this architecture is that it performs both entity linking and disambiguation. This method can then benefit from simplicity and from lack of error propagation. Furthermore, this architecture does not require complex feature engineering, which makes it easily adaptable to other languages.

For recognising all entity mentions in a document, Kolitsas *et al.* utilised an empirical probabilistic table entity–map, defined by $p(e|m)$. Where p is the probability of an entity e to be related to a mention m ; $p(e|m)$ is calculated using the number of times that mention m refers e within Wikipedia. From this probabilistic table, it is possible to find which are the top entities that a mention span refers to.

The end-to-end EL model starts by encoding every token in the text input by concatenating word and character embeddings and fed into a Bidirectional Long Short Term Memory (BiLSTM) [14] network. This representation is used to project mentions of this document into a shared dimensional space with the same size as the entity embeddings. These embeddings are fixed continuous entity representations generated separately, namely in the same manner as presented in [11], and aforementioned in Subsect. 4.2. In order to analyse long context dependencies of mentions, the authors utilised the attention mechanism proposed by [11]. This mechanism provides one context embedding per mention based on surrounding context words that are related to at least one of the candidate entities.

The final local score for each mention is determined by the combination of the $\log p(e|m)$, the similarity between the analysed mention and the candidate entity, and the long-range context attention for this mention. Finally, a top layer in the neural network promotes the coherence among disambiguated entities inside the same document.

4.4 Match Corrections

Multiple EL approaches, including the one used in this work, rely on the matching of entities and candidates using a probability table. If an entity is not listed in the probability table, the EL system cannot disambiguate it and, therefore, it cannot propose candidates. In historic documents, not matching entities is a frequent problem, due to their inherent nature and processing, as explained in Sect. 1.

To increase the matching of entities in the probability table, we propose an analysis that consists of exploring several surface name variations using multiple heuristics. For instance, we evaluate variations by lower and uppercasing, capitalising words, concatenating surrounding words, removing stopwords, and transliterating special characters, like accentuated letters, to Latin characters. If after applying the previous heuristics, a match is still lacking, we use the Levenshtein distance to overcome more complex cases, such as spelling mistakes or transcription errors generated by the OCR systems.

4.5 Multilingualism

Historical and literary documents may contain words and phrases in a language different from that of the document under analysis. For instance, as shown in Fig. 1(d), an English article uses “Porte de Namur” instead of “Namur Gate”. However, the former only exists in the French probability table while the latter is only found in the English one. To overcome this problem, we combined the probability tables of several languages in order to identify the surface names of entities in multiple languages.

4.6 Filtering

To improve the accuracy of the candidates provided by the EL systems, we use a post-processing filter based on heuristics and DBpedia. Specifically, we utilise DBpedia’s SPARQL Endpoint Query Service¹⁵. This filter uses DBpedia’s hierarchical structure for specifying categories that represent each named entity type. For instance, entities belonging to a location type were associated with categories such as “dbo:Location” and “dbo:Settlement”. The categories associated with each entity type were manually defined. Specifically, after requesting to the EL system the top five candidates for each named entity, the filtering steps are the following:

1. Verify that each candidate is in DBpedia and is associated with the correct categories. Candidates not matching the categories are put at the bottom of the rankings after a NIL;
2. Request to DBpedia the name of the candidates in the language of analysis; if the named entity is of type person, request as well the year of birth;

¹⁵ <https://wiki.dbpedia.org/public-sparql-endpoint>.

3. (Only if available) Remove those candidates that were born 10 years after the document publication;
4. Among the candidates with a retrieved name, find the most similar with respect to the named entity using Fuzzy Wuzzy Weighted Ratio¹⁶;
5. The most similar candidate is ranked at the top;
6. If the ranking does not contain a NIL, add one as the last possible candidate.

Since DBpedia does not always contain the requested candidate or the candidate's name, we rely as well on DBpedia Chapters when available. For instance, "Turku" is categorised in DBpedia¹⁷ but its name in Swedish, "Åbo" is not indexed; nevertheless, its Swedish name can be found in the Swedish DBpedia Chapter¹⁸. Another example is the case of "Luther-Werke", which does not exist in DBpedia, but it does exist in the German DBpedia Chapter¹⁹.

5 Experimental Settings

In the context of multilingual historical newspapers, documents tend to contain local information that is often specific to a language and one or more related geographical areas. The use of KB in the historical newspaper's language is an obvious choice because it reduces problems of data consistency while decreases noise from entities in other languages. For instance, entities can represent different things according to each KB. For example, the English and the Finnish Wikipedia pages with the title "Paris" do not describe the same entity; in Finnish "Paris" make reference to Greek mythology while the French capital is known as "Pariisi". Therefore, we trained our EL model for the corresponding language of historical newspapers.

For the entity embeddings and the entity disambiguation model, we used the pre-trained multilingual MUSE²⁰ word embeddings with of size 300 for all the languages in the corpora. The character embeddings are of size 50. As no historical data is available for English, we used the AIDA dataset [15] and validated on the CLEF HIPE 2020 data. Based on the statistical analysis of the training data, we defined a Levenshtein distance ratio of 0.93 to search for other mentions in the probability table if this mention does not have a corresponding entry in the table²¹.

For the evaluation, we compute precision (P), recall (R), and F-score (F1) measures calculated on the full corpus (micro-averaging). For the mentions without corresponding entries in the KB, EL systems provide a NIL entry to indicate that these mentions do not have a ground-truth entity in the KB.

¹⁶ <https://github.com/seatgeek/fuzzywuzzy>.

¹⁷ <http://dbpedia.org/page/Turku>.

¹⁸ <http://sv.dbpedia.org/page/%C3%85bo>.

¹⁹ <http://de.dbpedia.org/page/Luther-Werke>.

²⁰ <https://github.com/facebookresearch/MUSE>.

²¹ The source code of our EL system is available at: https://github.com/NewsEye/Named-Entity-Linking/tree/master/multilingual_entity_linking.

6 Evaluation

As we previously stated, the semantic textual enrichment of historical documents depends on aspects such as the OCR quality or how a language has evolved. In order to analyse the EL performance on historical data and the impact of our techniques on the disambiguation of entities in historical data, we present in the Tables 2 and 3 a simple EL baseline ($p(e|m)$) and different combinations of our EL approach (henceforth MEL). For the filtering experiments (see Sect. 4.6), we predicted the five best candidate entities for a mention m based on the probability table ($p(e|m)$).

The configuration MEL+ML+MC+F²² achieved the best results for French and German languages in CLEF HIPE 2020 corpora (Table 2).²³ Our model for English was trained on a contemporary dataset which degraded the performance of the MEL model and, consequently, all the variations. Despite the lack of historical training data, our model MEL+MC+F achieved the best results for the English data set (Table 2).

Table 2. Entity linking evaluation on the test CLEF HIPE 2020 data

Methods	English			French			German		
	P	R	F1	P	R	F1	P	R	F1
$p(e m)$	0.595	0.593	0.594	0.586	0.583	0.585	0.532	0.530	0.531
MEL	0.549	0.546	0.547	0.535	0.532	0.533	0.484	0.482	0.483
MEL+F	0.608	0.607	0.607	0.591	0.588	0.590	0.528	0.528	0.528
MEL+ML	0.535	0.533	0.534	0.554	0.551	0.552	0.492	0.490	0.491
MEL+ML+F	0.595	0.593	0.594	0.602	0.600	0.601	0.538	0.537	0.538
MEL+MC	0.559	0.557	0.558	0.556	0.553	0.555	0.500	0.498	0.499
MEL+MC+F	0.613	0.613	0.613	0.621	0.619	0.620	0.538	0.537	0.538
MEL+ML+MC	0.547	0.546	0.547	0.577	0.574	0.576	0.507	0.505	0.506
MEL+ML+MC+F	0.589	0.589	0.589	0.630	0.628	0.629	0.557	0.556	0.557

ML: Multilingualism; MC: Match correction; F: Filter

For the NewsEye corpora, the MEL+MC+F version achieved the best results for all languages (Table 3). Similar to CLEF HIPE 2020, the MEL version generated the worst predictions. The filter increased the F-scores values of all EL versions. The combination of probability tables had almost no changes in the predictions.

Though we generated the embedding representation for the 1.5M most frequent entities in each Wikipedia language, several historical entities are not so

²² The MEL+ML+MC+F model (team 10-run 1) [2] achieved the best performance for almost all metrics in English, French, and German on the CLEF HIPE 2020 shared task [results](#).

²³ The filter used in CLEF HIPE 2020 was modified in this work to improve accuracy and support DBpedia Chapters.

Table 3. Entity linking evaluation on the test NewsEye data

Methods	Finnish			French			German			Swedish		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
$p(e m)$	0.522	0.500	0.511	0.579	0.587	0.583	0.596	0.601	0.599	0.473	0.479	0.476
MEL	0.495	0.471	0.483	0.554	0.556	0.555	0.579	0.575	0.577	0.388	0.392	0.390
MEL+F	0.515	0.490	0.502	0.588	0.601	0.594	0.588	0.601	0.594	0.487	0.494	0.491
MEL+ML	0.505	0.481	0.493	0.555	0.558	0.557	0.575	0.573	0.574	0.392	0.397	0.394
MEL+ML+F	0.486	0.471	0.479	0.586	0.601	0.593	0.586	0.601	0.593	0.491	0.499	0.495
MEL+MC	0.501	0.481	0.491	0.562	0.568	0.565	0.582	0.580	0.581	0.386	0.390	0.388
MEL+MC+F	0.527	0.502	0.515	0.597	0.611	0.604	0.597	0.611	0.604	0.513	0.521	0.517
MEL+ML+MC	0.504	0.486	0.495	0.564	0.570	0.567	0.578	0.577	0.577	0.386	0.392	0.389
MEL+ML+MC+F	0.500	0.481	0.490	0.595	0.611	0.602	0.595	0.611	0.602	0.511	0.519	0.515

ML: Multilingualism; MC: Match correction; F: Filter

frequent on this KB. As our EL approach only disambiguates candidate entities that contain embedding representations, the MEL version achieved worse results than the baseline ($p(e|m)$). The major impact of this limitation was on the CLEF HIPE 2020 corpora where our approach had a drop of 0.05 in the F-score values.

Multilingualism. The combination of probability tables of several languages has slightly improved the results on both corpora. This combination provided different surface names for an entity in different languages. In addition, this combination of probability tables allowed our models to disambiguate entities that are non-existent in some KBs. For example, the Russian politician “Nikolai Alexeïevitch Maklakov” who is mentioned in the Finnish data does not exist in our Finnish KB, but he exists in our English and French KBs.

Despite providing additional surface variations, some surface names (e.g., acronyms) can have different meanings in different languages. Other potential risks are mentions with some OCR mistakes that can make reference to another entity in other languages and the combination of probability tables can increase the number of candidate entities and the ambiguity of mentions.

Match Corrections. Our different analysis to normalise mentions and correct small mistakes generated by the OCR engine improved the performance of our approach. CLEF HIPE 2020 benefited slightly more from this technique than NewsEye. This could be either due to differences in the images quality, type of OCR used or manual correction.

On one hand, the combination of normalisation and Levenshtein distance methods allowed our method to correct mentions like “Londires” and “Toujquet” to “Londres” and “Touquet”, respectively. On the other hand, our method could not find the correct mentions for simple cases. In the example “Gazstte of the Unites States”, our approach did not find corresponding candidates for this mention. The correct answer is “Gazette of the United States”; however, the Levenshtein distance ratio is 0.928 and our threshold to correct a mention is

0.93. Another example of OCR errors is the mention “United Staeres”. In this case, the correct entity is “United States”; however, the candidate mention in the probability with the best Levenshtein distance ratio is “United Stars” which made our approach generated the wrong disambiguation. A lower Levenshtein distance ratio may find more degraded mention; however, this low ratio can generate too many mistakes for entities that not exist in KB. In the future, we will explore whether Fuzzy Wuzzy, an improved Levenshtein distance used in the filter (Sect. 4.6), could alleviate these issues.

Filtering. The use of a post-processing filter for refining the top five most probable candidates, allowed us to achieve the best results, as observed in Table 2 and Table 3. Specifically, with the filter, we prioritised the candidates that not only were the most similar to the named entity but also, those that agreed with the named entity type and publication year. For instance, in an English newspaper published in 1810 the named entity of type person “Mr. Vance”²⁴ had for candidates the following Wikidata IDs: “Q507981” (location), “Q19118257” (person born in 1885), “Q985481” (location), and “Q7914040” (person born in 1930). Thanks to the filter, we observed that most of the candidates belonged to locations, while the proposed people were born long after the journal publication; thus, the best candidate should be a NIL, which in fact was the correct prediction. Despite DBpedia does not support languages such as Finnish, the filter can still improve the results using only the information regarding named entity categories, as seen in Table 3. It should be noticed that the filter is not free of errors. In some cases, the best candidate was positioned at the end of the rankings because DBpedia’s categories did not match the categories defined for the named entity type, e.g., the journal “Le Temps”, a product-type named entity, is not classified as a human work in DBpedia²⁵.

As digital library frameworks tend to provide the top N most probable entities for a mention in a context, we analysed the performance of the best two EL approach versions when we provide the top three candidate entities for each mention. These results are presented in Table 4. The MEL+MC+F method achieved the best average F-score, which is remarkable considering that the issues encountered in multilingual historical data can increase the difficulty of

Table 4. F-scores values for the top three candidate entities on the test data sets.

Methods	CLEF HIPE 2020			NewsEye			
	English	French	German	Finnish	French	German	Swedish
MEL+MC+F	0.726	0.691	0.623	0.598	0.706	0.699	0.594
MEL+ML+MC+F	0.710	0.690	0.645	0.566	0.710	0.700	0.605

ML: Multilingualism; MC: Match correction; F: Filter

²⁴ [HIPE-data-v1.3-test-en.tsv#L4232-L4234](#).

²⁵ [http://dbpedia.org/page/Le_Temps_\(Paris\)](http://dbpedia.org/page/Le_Temps_(Paris)).

this task. Compared to Tables 2 and 3, the results are at least 14% better than the top one prediction.

Based on all the previous results, we can observe that our EL approach outperformed the baseline for both corpora in all languages. Thus, we can conclude that the proposed techniques partially attenuated the impact of historical data issues. As well, the proposition of the best candidates can accelerate the work of librarians and humanities professionals in the analysis of historical documents in several languages and on different subjects. Finally, despite the recent progress, the EL for historical data is still a challenging task due to the multiple constraints. Examples of these limitations are the lack of annotated training data and the existence of multiple missing historical entities in the KBs, which can limit the training of more robust models.

7 Conclusion

Historical documents are essential resources for cultural and historical heritage. Enriching semantically historical documents, with aspects such as named entity recognition and entity linking, can improve their analysis and exploitation within digital libraries. In this work, we investigated a multilingual end-to-end entity linking system created for processing historical documents and disambiguate entities in English, Finnish, French, German, and Swedish. Specifically, we make use of entities embeddings, built from Wikipedia in multiple languages, along with a neural attention mechanism that analyses context words and candidate entities embeddings to disambiguate mentions in historical documents.

Additionally, we proposed several techniques to minimise the impact of issues frequently found in historical data, such as multilingualism and errors related to OCR systems. As well, we presented a filtering process to improve the linking of entities. Our evaluation on two historical corpora (CLEF HIPE 2020 and NewsEye) showed that our methods outperform the baseline and considerably reduce the impact of historical document issues on different subjects and languages.

There are several potential avenues of research and application. Following the idea proposed by [7], entity linking in historical documents could be used to improve the coverage and relevance of historical entities within knowledge bases. Another perspective would be to adapt our entity linking approach to automatically generate ontologies for historical data. As well, it would be interesting to use diachronic embeddings to deal with named entities that have changed of name through the time, such as “Beijing” in English²⁶. Finally, we would like to improve our post-processing filter by including information from knowledge bases such as Wikidata or BabelNet [25].

Acknowledgments. This work has been supported by the European Union’s Horizon 2020 research and innovation program under grant 770299 (NewsEye) and 825153 (EMBEDDIA).

²⁶ Google N-grams in English for “Beijing”, “Peking”, and “Pekin” between 1700 and 2008: books.google.com/ngrams/.

References

1. Agirre, E., Barrena, A., de Lacalle, O.L., Soroa, A., Fernando, S., Stevenson, M.: Matching cultural heritage items to Wikipedia. In: Eight International Conference on Language Resources and Evaluation (LREC) (2012)
2. Boros, E., et al.: Robust named entity recognition and linking on historical multilingual documents. In: Cappellato, L., Eickhoff, C., Ferro, N., Névél, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
3. Brando, C., Frontini, F., Ganascia, J.-G.: Disambiguation of named entities in cultural heritage texts using linked data sets. In: Morzy, T., Valduriez, P., Bellatreche, L. (eds.) ADBIS 2015. CCIS, vol. 539, pp. 505–514. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23201-0_51
4. Brando, C., Frontini, F., Ganascia, J.G.: REDEN: named entity linking in digital literary editions using linked data sets. *Complex Syst. Inf. Model. Q.* **7**, 60–80 (2016). <https://doi.org/10.7250/csimq.2016-7.04>. <https://hal.sorbonne-universite.fr/hal-01396037>
5. Cucerzan, S.: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 708–716. Association for Computational Linguistics, Prague, Czech Republic, Jun 2007. <https://www.aclweb.org/anthology/D07-1074>
6. Wilde, M.: Improving retrieval of historical content with entity linking. In: Morzy, T., Valduriez, P., Bellatreche, L. (eds.) ADBIS 2015. CCIS, vol. 539, pp. 498–504. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23201-0_50
7. Dredze, M., McNamee, P., Rao, D., Gerber, A., Finin, T.: Entity disambiguation for knowledge base population. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 277–285. Coling 2010 Organizing Committee, Beijing, China, August 2010. <https://www.aclweb.org/anthology/C10-1032>
8. Ehrmann, R., Clematide, F.: HIPE - Shared Task Participation Guidelines, January 2020. <https://doi.org/10.5281/zenodo.3677171>
9. Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: named entity recognition and linking on historical newspapers. In: Jose, J.M., et al. (eds.) ECIR 2020, Part II. LNCS, vol. 12036, pp. 524–532. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_68
10. Frontini, F., Brando, C., Ganascia, J.G.: Semantic web based named entity linking for digital humanities and heritage texts. In: Proceedings of the First International Workshop Semantic Web for Scientific Heritage at the 12th ESWC 2015 Conference, vol. 1364, June 2015
11. Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2619–2629. Association for Computational Linguistics (2017). <https://doi.org/10.18653/v1/D17-1277>
12. Gefen, A.: Les enjeux épistémologiques des humanités numériques. *Socio* (2015). <https://doi.org/10.4000/socio.1296>
13. Heino, E., et al.: Named entity linking in a complex domain: case second world war history. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) LDK 2017. LNCS (LNAI), vol. 10318, pp. 120–133. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_10

230 E. L. Pontes et al.

14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
15. Hoffart, J., et al.: Robust disambiguation of named entities in text. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 782–792. Association for Computational Linguistics, Edinburgh, Scotland, UK, July 2011. <https://www.aclweb.org/anthology/D11-1072>
16. van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R.: Exploring entity recognition and disambiguation for cultural heritage collections. *Digit. Sch. Humanit.* **30**(2), 262–279 (2013). <https://doi.org/10.1093/llc/fqt067>
17. Huet, T., Biega, J., Suchanek, F.M.: Mining history with Le Monde. In: *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pp. 49–54. AKBC 2013. Association for Computing Machinery, New York, NY, USA (2013). <https://doi.org/10.1145/2509558.2509567>
18. Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 519–529. Association for Computational Linguistics (2018). <https://doi.org/10.18653/v1/K18-1050>
19. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web J.* **6**(2), 167–195 (2015). <https://doi.org/10.3233/SW-140134>
20. Linhares Pontes, E., Hamdi, A., Sidere, N., Doucet, A.: Impact of OCR quality on named entity linking. In: Jatowt, A., Maeda, A., Syn, S.Y. (eds.) *ICADL 2019*. LNCS, vol. 11853, pp. 102–115. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-34058-2_11
21. Linhares Pontes, E., Moreno, J.G., Doucet, A.: Linking named entities across languages using multilingual word embeddings. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, JCDL 2020*, pp. 329–332. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3383583.3398597>
22. Mosallam, Y., Abi-Haidar, A., Ganascia, J.-G.: Unsupervised named entity recognition and disambiguation: an application to old French Journals. In: Perner, P. (ed.) *ICDM 2014*. LNCS (LNAI), vol. 8557, pp. 12–23. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08976-8_2
23. Munnelly, G., Lawless, S.: Investigating entity linking in early english legal documents. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018*, pp. 59–68. Association for Computing Machinery, New York, NY, USA (2018). <https://doi.org/10.1145/3197026.3197055>
24. Munnelly, G., Pandit, H.J., Lawless, S.: Exploring linked data for the automatic enrichment of historical archives. In: Gangem, A., et al. (eds.) *ESWC 2018*. LNCS, vol. 11155, pp. 423–433. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98192-5_57
25. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* **193**, 217–250 (2012). <https://doi.org/10.1016/j.artint.2012.07.001>
26. Pellissier Tanon, T., Weikum, G., Suchanek, F.: YAGO 4: a reason-able knowledge base. In: Harth, A.A., et al. (eds.) *ESWC 2020*. LNCS, vol. 12123, pp. 583–596. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49461-2_34
27. Rijhwani, S., Xie, J., Neubig, G., Carbonell, J.: Zero-shot neural transfer for cross-lingual entity linking. In: *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*. Honolulu, Hawaii, January 2019. <https://doi.org/10.1609/aaai.v33i01.33016924>

28. Ruiz, P., Poibeau, T.: Mapping the Bentham Corpus: Concept-based Navigation. *J. Data Min. Digit. Humanit. Special Issue: Digital Humanities between knowledge and know-how (Atelier Digit_Hum)*, March 2019. <https://hal.archives-ouvertes.fr/hal-01915730>
29. Shen, W., Wang, J., Han, J.: Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.* **27**(2), 443–460 (2015). <https://doi.org/10.1109/TKDE.2014.2327028>
30. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Constantopoulos, P., Sølvsberg, I.T. (eds.) *ECDL 2001. LNCS*, vol. 2163, pp. 127–136. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44796-2_12
31. Zhou, S., Rijhwani, S., Neubig, G.: Towards zero-resource cross-lingual entity linking. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 243–252. ACL, China, November 2019. <https://doi.org/10.18653/v1/D19-6127>

6 Event Detection with Entity Markers

Event Detection with Entity Markers

Anonymous ECIR submission

Abstract. Event detection involves the identification of instances of specified types of events in text and their classification into event types. In this paper, we approach the event detection task as a relation extraction task. In this context, we assume that the clues brought by the entities participating in an event are important and could improve the performance of event detection. Therefore, we propose to exploit entity information explicitly for detecting the event triggers by marking them at different levels while fine-tuning a pre-trained language model. The experimental results prove that our approach obtains state-of-the-art results on the ACE 2005 dataset.

Keywords: Information Extraction · Event Extraction · Event Detection

1 Introduction

Event detection (ED) aims to identify the instances of specified types of events in text. An event is represented by an *event mention* (a text that contains an event of a specific type and subtype), an *event trigger* (the word that expresses the event mention), an *event argument* (a participant in the event of a specific type), and an *argument role* (the role of the entity in the event). For instance, according to the ACE 2005 annotation guidelines¹, in the sentence “*She’s been convicted of obstruction of justice.*”, an event detection system should be able to recognize the word *convicted* as a trigger for the specific event type **Convict**.

A main challenge intervenes when the same event might appear in the form of various trigger expressions and an expression might represent different event types in different contexts. For example, *transfer* could refer to transferring ownership of an item, transferring money, or transferring personnel from one location to another. Each sense of the word is linked with an event type. In the same manner, *fired* can correspond to an **attack** type of event as in “*an American tank fired on the street*” or it can express the **dismissal** of an employee from a job as in “*Hillary Clinton was fired from the House Judiciary Committee’s Watergate investigation*”.

Therefore, we would assume that, in such cases, significant clues can be given by the context of a candidate trigger and by the presence of the participants at the event in this context, e.g. named entities. For analyzing the importance of these indicators of the existence of an event in a sentence, we adopt a relation extraction model to perform event detection by taking advantage of the participants in the event (event arguments).

¹ <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf>

2 Anonymous ECIR submission

2 Related Work

Most current state-of-the-art systems perform event detection individually [2, 20, 6], where the entities are either ignored or considered helpful in joint models.

Some works made use of gold-standard entities in different manners. Higher results can be obtained with gold-standard entity types [20], by concatenating randomly initialized embeddings for the entity types. A graph neural network (GNN) based on dependency trees [21] has also been proposed to perform event detection with a pooling method that relies on entity mentions aggregating the convolution vectors. Arguments provided significant clues to this task in the supervised attention mechanism proposed to exploit argument information explicitly for event detection [14], while also using events from FrameNet.

Although some joint learning-based methods have been proposed, which tackled event detection and argument extraction simultaneously, these approaches usually only make significant improvements on the argument extraction, but insignificant to event detection. These methods usually combine the loss functions of these two tasks and are jointly trained under the supervision of annotated triggers and arguments. Event triggers and their arguments are predicted at the same time in a joint framework [18] with bidirectional recurrent neural networks (Bi-RNNs) and a convolutional neural network (CNN) and systematically investigate the usage of memory vectors/matrices to store the prediction information during the course of labeling sentence features.

The architecture adopted in [15] was to jointly extract multiple event triggers and event arguments by introducing syntactic shortcut arcs derived from the dependency parsing trees to enhance the information flow in an attention-based graph convolution network (GCN) model. The gold-standard entity types are embedded as features for trigger and argument prediction. The argument information was also exploited in [14] explicitly for event detection by experimenting with different strategies for adding supervised attention mechanisms. The authors exploit the annotated entity information by concatenating the token embeddings with randomly initialized entity type embeddings.

Recently, different approaches that include external resources and features at a sub-word representation level have been proposed. Thus, generative adversarial networks (GANs) have been applied in event detection [28, 8]. Besides, reinforcement learning (RL) is used in [28] for creating an end-to-end entity and event extraction framework. The approach attempted in [27] based on the BERT model with an automatic generation of labeled data by editing prototypes and filtering out the labeled samples through argument replacement by ranking their quality. A similar framework is proposed by [25] but information is encoded by BERT or a CNN suggesting a growing interest in adversarial models. Simultaneously, an integration of a distillation technique to enhance the adversarial prediction was explored in [16].

Although recent advances are focused on multiple techniques, several BERT-based architectures have been proposed [24, 27, 25]. In this work, we demonstrate that the advantages of BERT can be improved by adding extra information by explicitly marking the entities in the input text. We continue with the presenta-

tion of our proposed model in Section 3. The experimental setup and the results are detailed in Section 4 and we finalize with some conclusions and perspectives in Section 5.

3 Approach

We implemented the BERT-based model with *EntityMarkers*² [22] applied for relation classification and we adapted it to perform event detection.

First, our model extends the recently introduced BERT [3] model applied to sequential data. BERT itself is a stack of Transformer layers [23] which takes as input a sequence of subtokens, obtained by the WordPiece tokenization [26], and produces a sequence of context-based embeddings of these subtokens. We refer the readers to the original paper for a more detailed description. We modify BERT by adding a conditional random fields (CRF) layer instead of the dense one, which is commonly used in other works on sequential labeling [10, 17] to ensure output consistency.

Next, the *EntityMarkers* model [22] consists in augmenting the input data with a series of special tokens. Thus, if we consider a sentence $x = [x_0, x_1, \dots, x_n]$ with n tokens, we augment x with two reserved word pieces to mark the beginning and the end of each event argument mention in the sentence.

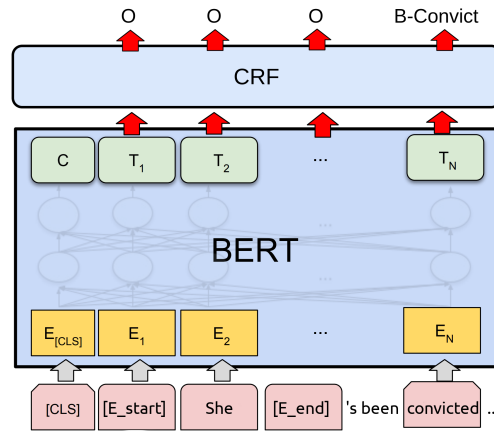


Fig. 1. The BERT-based model with *Entity Position Markers* and a CRF top layer.

We introduce three types of markers: (1) *Entity Position Markers*, e.g. $[E_start]$ and $[E_end]$ where E represents an entity of any type, (2) *Entity Type Markers*,

² We only used the input type representation and consider a more complex output based on tokens which are not considered in [22].

4 Anonymous ECIR submission

e.g. PER_{start} and PER_{end} where PER represents an entity of type Person, and (3) we also test that, in the case of the event argument roles are known beforehand, the *Argument Role Markers*, e.g. $[Defendant_{start}]$, $[Defendant_{end}]$ where Defendant is an event argument role. We modify x to give:

$x = [x_0, x_1, \dots, [MARKER_{start}]x_i \dots x_{j-1}[MARKER_{end}], \dots, x_n]$ and we feed this token sequence into BERT instead of x . We also update the entity indices $E = (i + 1, j + 1)$ to account for the inserted tokens, as shown in Figure 1 for the model with *Entity Position Markers*.

As an example, in the sentence “*She’s been convicted of obstruction of justice.*”, where *She* has the argument role of a **Defendant** and *obstruction of justice* is an argument of type **Crime**, the sentence is augmented as follows:

- (1) $[E_{start}]$ *She* $[E_{end}]$ ’s been convicted of $[E_{start}]$ **obstruction of justice** $[E_{end}]$.
- (2) $[PER_{start}]$ *She* $[PER_{end}]$ ’s been convicted of $[Crime_{start}]$ **obstruction of justice** $[Crime_{end}]$.
- (3) $[Defendant_{start}]$ *She* $[Defendant_{end}]$ ’s been convicted of $[Crime_{start}]$ **obstruction of justice** $[Crime_{end}]$.

For the *Argument Role Markers*, if an entity has different roles in different events that are present in the same sentence, we mark the entity with all the argument roles that it has.

4 Experiments and Results

The evaluation is conducted on the annotated data ACE 2005 corpus. For comparison purposes, we use the same test set with 40 news articles (672 sentences), the same development set with 30 other documents (863 sentences), and the same training set with the remaining 529 documents (14,849 sentences) as in previous studies of this dataset [9, 12, 11, 20, 18]. The ACE 2005 corpus has 8 types of events, with 33 subtypes (e.g. the event type *Conflict* has two subtypes *Attack*, *Demonstrate*) that, along with one class “O” for the non-trigger tokens, constitutes a 34-class classification problem. Following the same line of works, we consider that a trigger is correct if its event type, subtype, and offsets match those of a reference trigger. We use Precision (P), Recall (R), and F-measure (F1) to evaluate the overall performance.

We use the Stanford CoreNLP toolkit³ to preprocess the data, including tokenization and sentence splitting. For fine-tuning the BERT-based models, we followed the selection of parameters presented in [3]. We found that 2×10^{-5} learning rate and a mini-batch of dimension 4 provided stable and consistent convergence across all experiments as evaluated on the development set.

We first consider four baselines based on the BERT language model, applied in a similar way to [4] for the named entity recognition (NER) task, with the recommended hyperparameters. We test four pre-trained widely used English language models, two based on BERT-base and two based on BERT-large, *cased*

³ <http://stanfordnlp.github.io/CoreNLP/>

Table 1. Evaluation of the BERT-based models on the blind test data.

Models	Precision	Recall	F1
BERT-base-uncased	71.6	68.4	70.0
BERT-base-cased	71.3	72.0	71.6
BERT-large-uncased	72.0	72.9	72.5
BERT-large-cased	69.3	77.1	73.0

(trained on the original words) and *uncased* (trained on lowercased words). Between the BERT models, it is worth noticing that the *cased* models perform better than the *uncased* ones, which could confirm that named entities that are usually capitalized are an important clue for the event detection task⁴.

Table 2. Evaluation of our models and comparison with state-of-the-art systems for event detection on the blind test data. ⁺with gold-standard arguments. Change improvements w.r.t. our models are showed in columns “F1 Improvement (%)”. Improvements greater than 10% are highlighted with background color.

Models	Precision	Recall	F1	F1 Improvement (%)		
				(1)	(2)	(3)
CNN [20]	71.9	63.8	67.6	12.72%	16.12%	17.75%
CNN ⁺ [20]	71.8	66.4	69.0	10.43%	13.77%	15.36%
Dynamic multi-pooling CNN [2]	75.6	63.6	69.1	10.27%	13.60%	15.20%
Joint RNN [18]	66.0	73.0	69.3	9.96%	13.28%	14.86%
CNN with document context [6]	77.2	64.9	70.5	8.09%	11.35%	12.91%
Non-Consecutive CNN [19]	N/A	N/A	71.3	6.87%	10.10%	11.64%
Attention-based ⁺ [14]	78.0	66.3	71.7	6.28%	9.48%	11.02%
GAIL [28]	74.8	69.4	72.0	5.83%	9.03%	10.56%
Gated Cross-Lingual Attention [13]	78.9	66.9	72.4	5.25%	8.43%	9.94%
Graph CNN [21]	77.9	68.8	73.1	4.24%	7.39%	8.89%
Seed-based [1]	80.6	67.1	73.2	4.10%	7.24%	8.74%
Hybrid NN [7]	84.6	64.9	73.4	3.81%	6.95%	8.45%
Attention-based GCN [15]	76.3	71.3	73.7	3.39%	6.51%	8.01%
Δ -learning [16]	76.3	71.9	74.0	2.97%	6.08%	7.57%
DEEB-RNN3y [29]	72.3	75.8	74.0	2.97%	6.08%	7.57%
BERT-base-uncased+LSTM [24]	N/A	N/A	68.9	10.60%	13.93%	15.53%
BERT-base-uncased [24]	N/A	N/A	69.7	9.33%	12.63%	14.20%
BERT-base-uncased [5]	67.1	73.2	70.0	8.86%	12.14%	13.71%
BERT-QA [5]	71.1	73.7	72.3	5.39%	8.58%	10.10%
DMBERT [25]	77.6	71.8	74.6	2.14%	5.23%	6.70%
DMBERT+Boot [25]	77.9	72.5	75.1	1.46%	4.53%	5.99%
BERT-large-cased	69.3	77.1	73.0	4.38%	7.53%	9.04%
BERT-large-cased+Entity Position Markers ⁺ (1)	75.9	76.6	76.2	-	3.02%	4.46%
BERT-large-cased+Entity Type Markers ⁺ (2)	79.3	77.8	78.5	-	-	1.40%
BERT-large-cased+Argument Role Markers ⁺ (3)	78.9	80.4	79.6	-	-	-

We compare our proposed models with markers with several state-of-the-art neural-based models proposed for event detection, that do not use external

⁴ An amount of around 30% of the entities and 3% of the event triggers have the first token capitalized.

6 Anonymous ECIR submission

resources, more specifically with the following models based on CNNs and RNNs: the CNN-based model proposed in [20] with and without the addition of gold-standard entities, the dynamic multi-pooling CNN model [2], the bidirectional joint RNNs [18], the non-consecutive CNN in [19], the hybrid model proposed by [7], the GAIL model proposed by [28], the gated cross-lingual attention model presented in [13], and the graph CNN proposed by [21].

We also compare our approach with recent proposed BERT-based models, the fine-tuned baseline BERT-base-uncased in [5], the QA-BERT [5] where the task has been approached as a question answering task, the two models with adversarial training for weakly supervised event detection proposed in [25], and the BERT and LSTMs approaches proposed by [24] that models text spans and captures within-sentence and cross-sentence context.

We first notice that our baselines presented in Table 1 achieve similar results with the BERT-base-uncased in [5] (the same F1 value and similar precision and recall scores) and [24]. Since we could not replicate the results of the BERT-based approach [27]⁵, we did not consider it a comparable approach. Full results of our model and its comparison against state of the art is presented in Table 2. There is a significant gain with the trigger classification of 9.04% higher over the stand-alone BERT-based model and 5.99% to the best reported previous models. These results demonstrate the effectiveness of our method to incorporate the argument information. Moreover, the improvements are consistent regardless of the type of encoder (BERT or other) used to represent the inputs. For our first model (*Entity Position Markers*), where the entities are surrounded by a general marker that does not depend on the entity type, the results are improved with three percentage points revealing that the position of the entities is relevant for the trigger detection task. Furthermore, when we mark the entities with their argument roles (*Argument Role Markers*), the recall and F1 increase with around one absolute percentage point. However, this case is substantially optimistic as it assumes that argument roles were correctly identified and typed.

5 Conclusions and Perspectives

We presented an approach for integrating entity information for the event detection task by adding different levels of entity markers, their positions, their types, and finally, their argument roles. Considering the results, we can conclude that marking entities in a sentence can significantly improve the F1 scores and obtain state-of-the-art values. Further analysis remains to be done in order to understand in which cases the markers bring informative features. As future work, we propose to tackle the drawbacks of our current model by introducing the recognition and typing of the entities in our model.

⁵ To the best of our knowledge, there is no public implementation and our attempt to implement their model did not achieve their results.

References

1. Bronstein, O., Dagan, I., Li, Q., Ji, H., Frank, A.: Seed-based event trigger labeling: How far can event descriptions get us? In: ACL (2). pp. 372–376 (2015)
2. Chen, Y., Xu, L., Liu, K., Zeng, D., Zhao, J.: Event extraction via dynamic multi-pooling convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. vol. 1, pp. 167–176 (2015)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
5. Du, X., Cardie, C.: Event extraction by answering (almost) natural questions. arXiv preprint arXiv:2004.13625 (2020)
6. Duan, S., He, R., Zhao, W.: Exploiting document level information to improve event detection via recurrent neural networks. In: Eighth International Joint Conference on Natural Language Processing (IJCNLP 2017). pp. 352–361. Asian Federation of Natural Language Processing (2017)
7. Feng, X., Huang, L., Tang, D., Ji, H., Qin, B., Liu, T.: A language-independent neural network for event detection. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 66–71 (2016)
8. Hong, Y., Zhou, W., Zhang, J., Zhou, G., Zhu, Q.: Self-regulation: Employing a generative adversarial network to improve event detection. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 515–526 (2018)
9. Ji, H., Grishman, R., et al.: Refining event extraction through cross-document inference. In: ACL. pp. 254–262 (2008)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
11. Li, Q., Ji, H., Huang, L.: Joint event extraction via structured prediction with global features. In: ACL (1). pp. 73–82 (2013)
12. Liao, S., Grishman, R.: Using document level cross-event inference to improve event extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 789–797. Association for Computational Linguistics (2010)
13. Liu, J., Chen, Y., Liu, K., Zhao, J.: Event detection via gated multilingual attention mechanism. In: Thirty-second AAAI Conference on Artificial Intelligence (AAAI-18) (2018)
14. Liu, S., Chen, Y., Liu, K., Zhao, J.: Exploiting argument information to improve event detection via supervised attention mechanisms. In: 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017). pp. 1789–1798. Vancouver, Canada (2017)
15. Liu, X., Luo, Z., Huang, H.: Jointly multiple events extraction via attention-based graph information aggregation. arXiv preprint arXiv:1809.09078 (2018)

8 Anonymous ECIR submission

16. Lu, Y., Lin, H., Han, X., Sun, L.: Distilling discrimination and generalization knowledge for event detection via delta-representation learning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4366–4376 (2019)
17. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)
18. Nguyen, T.H., Cho, K., Grishman, R.: Joint event extraction via recurrent neural networks. In: Proceedings of NAACL-HLT. pp. 300–309 (2016)
19. Nguyen, T.H., Fu, L., Cho, K., Grishman, R.: A two-stage approach for extending event detection to new types via neural networks. ACL 2016 p. 158 (2016)
20. Nguyen, T.H., Grishman, R.: Event detection and domain adaptation with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. pp. 365–371 (2015)
21. Nguyen, T.H., Grishman, R.: Graph convolutional networks with argument-aware pooling for event detection. In: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018) (2018)
22. Soares, L.B., FitzGerald, N., Ling, J., Kwiatkowski, T.: Matching the blanks: Distributional similarity for relation learning. arXiv preprint arXiv:1906.03158 (2019)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
24. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, relation, and event extraction with contextualized span representations. arXiv preprint arXiv:1909.03546 (2019)
25. Wang, X., Han, X., Liu, Z., Sun, M., Li, P.: Adversarial training for weakly supervised event detection. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 998–1008 (2019)
26. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
27. Yang, S., Feng, D., Qiao, L., Kan, Z., Li, D.: Exploring pre-trained language models for event extraction and generation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5284–5294 (2019)
28. Zhang, T., Ji, H., Sil, A.: Joint entity and event extraction with generative adversarial imitation learning. Data Intelligence 1(2), 99–120 (2019)
29. Zhao, Y., Jin, X., Wang, Y., Cheng, X.: Document embedding enhanced event detection with hierarchical and supervised attention. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 414–419 (2018)

7 Event Extraction over Digitised and Historical Documents

Event Extraction over Digitised and Historical Documents

EMANUELA BOROS and NHU KHOA NGUYEN, University of La Rochelle, France

GAËL LEJEUNE, Sorbonne University, France

ANTOINE DOUCET, University of La Rochelle, France

In order to extract relevant semantic information (e.g., named entities, relations, events) from historical documents written in different languages, several challenges have to be overcome. The level of degradation of historical documents and the quality of their optical character recognition (OCR) might hinder the performance of information extraction systems. Moreover, historical spellings and word variations can have further negative impact. In this paper, we aim at approaching the event extraction task by experimenting with language-independent models and we analyse their robustness to OCR noise, and their ability to mitigate problems caused by the low quality of the digitised documents. To the best of our knowledge, there is no prior research addressing this topic. Being also faced with a lack of annotated data, we simulate the existence of transcribed data, synthesised from clean annotated text, by injecting synthetic noise. We observe that the imbalance of the datasets and the richness of the different annotation styles are two important factors that influence the event extraction task. Finally, we conclude that the errors propagated from the digitisation process can affect all the tested systems.

Additional Key Words and Phrases: information extraction, event extraction, event detection

ACM Reference Format:

Emanuela Boros, Nhu Khoa Nguyen, Gaël Lejeune, and Antoine Doucet. 2020. Event Extraction over Digitised and Historical Documents. 1, 1 (October 2020), 17 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Event extraction is an application of information extraction (IE) that implies the extraction of specific knowledge from certain incidents from texts. This task is focused on obtaining event-related information from texts, and, as commonly defined in the field of IE, it consists of two main sub-tasks. The first sub-task involves event detection (ED) that deals with the extraction of critical information regarding an event, that can be represented by a keyword, a phrase, a sentence or a span of text, which evoke that event. For example, an article can talk about a new epidemic outbreak, or about the election of a new president, where the events to be detected are represented by the name of the epidemic, or by the word ‘election’. The second sub-task, mostly referred to as event argument extraction, concentrates on the extraction of event extents referring to more details about the events, such as their arguments. They often refer to the participants in the event. For example, the location of the epidemic event, the name of the president, the country of the election, are to be detected in this sub-task. Despite the usefulness and prospective applicability of EE (which implies the ED sub-task), several issues and challenges are to be overcome until an IE system is widely adopted as an effective tool in practice. For example, there are

Authors’ addresses: Emanuela Boros, emanuela.boros@univ-lr.fr; Nhu Khoa Nguyen, University of La Rochelle, 23 Avenue Albert Einstein, La Rochelle, France, 17000; Gaël Lejeune, Sorbonne University, 15-21 Rue de l’École de Médecine, Paris, France, 75006, gael.lejeune@sorbonne-universite.fr; Antoine Doucet, antoine.doucet@univ-lr.fr, University of La Rochelle, 23 Avenue Albert Einstein, La Rochelle, France, 17000, antoine.doucet@univ-lr.fr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

XXXX-XXXX/2020/10-ART \$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

, Vol. 1, No. 1, Article . Publication date: October 2020.

2 • Boros, et al.

practical issues related to the high cost of manual annotation of texts (e.g. human resources). The human effort needs to be minimised while keeping the quality of an IE system. Data annotation takes advantage of a massive human expertise and this causes labour-intensive work for data interpretation at two levels. Firstly, an IE system may use NLP resources and tools, created using lots of annotated documents and secondly, an IE system needs a higher-level of annotation of relations or events, annotations that can be complex and extremely costly. Also, features that come from NLP tools and resources (i.e., dependency parsers, part of speech taggers etc.) and the hard decision making in combining them is considered an important issue due to the error propagation issues. The errors from these sources can propagate to the downstream tasks, e.g. an NER system may mistakenly detect the wrong entity needed by a relation extraction system, which downgrades considerably its accuracy.

However, while the task of event extraction is already challenging for contemporary data, in the context of Digital Humanities, another issue must also be approached. Since the documents are being digitised using different optical character (OCR) tools, from the historical degraded documents or due to the fact that most of digitised documents are indexed through their transcribed version, errors arise from OCR errors that may hinder the access to them. Though there has been an interest in studying the effect of OCR onto other IE tasks (e.g. NER, NEL), to our knowledge, there is no research done on this impact on event extraction.

Thus, in this paper, we distinguish between different event definitions, and we introduce the evaluation framework based on two datasets: one was created along with the data analysis for information extraction in any language (DAnIEL) [5] system, and the other one is the ACE 2005 corpora provided by the automatic content extraction (ACE) evaluation¹. Both datasets will be used for all the following experiments. The DAnIEL dataset consists of a large number of multilingual collected documents from different press threads in the field of *health* (Google News) focused on epidemic events. ACE 2005 dataset covers the most common events of national and international news, in three languages (Arab, English, and Chinese) from a variety of sources selected from broadcast news programs, newspapers, newswire reports, internet sources or transcribed audio.

Consequently, we present two approaches to event detection, both with the ability of handling multilingual data. The first one is based on the DAnIEL system which is a discourse-level approach that exploits the global structure of news in a newswire. It also tackles the difficulty of language adaptation by its character-based characteristic that uses positions of occurrences in text. We believe that DAnIEL is very adequate for its ability to handle text in any language and because of an algorithm that should be robust to noise. It only requires two occurrences of adequate substrings, regardless of the recognition of the rest of the text. Its weakness is that is tailored for epidemic events, although it should be possible to adapt to other domains. In this report, we will, therefore, experiment with it over epidemic events to decide on the worthiness of its adaptation to other domains.

We also introduce a neural network-based approach based on a convolutional neural network (CNN) applied to a local context, more exactly to a window of text around potential keywords that can represent events (we refer to them as *triggers*). This model automatically learns features from the sequence of tokens (word and/or character) and decides if the middle word of the window of text can trigger an event or not. We chose this model for its ability to learn features automatically, independently of the domain, randomly initialised at first and fine-tuned on the event extraction task. Determining its ability to handle noise is one of the objectives of the present report.

We analyse both models and both datasets systematically. Firstly, for the DAnIEL dataset, we consider both approaches, DAnIEL system and the CNN-based approach. For the ACE 2005 dataset, we consider only the CNN-based approach, since the DAnIEL system holds the specificity of being focused only on epidemic events. We aim at testing the robustness of the models against noise, their ability of treating highly inflected languages and misspelled or unseen words, which can be either due to the low quality of text or the spelling variants. For these experiments, we present separately their evaluation general settings. Furthermore, we create synthetic

¹<https://catalog.ldc.upenn.edu/ldc2006t06>

data starting from the initial datasets in order to study the direct impact of OCR over the performance of both approaches.

2 DATASETS

In this section, we present two datasets. The first one is specific to the DAnIEL system [5] destined for multilingual epidemic surveillance and which contains articles on different press threads in the field of *health* (Google News) focused on epidemic events from different collected documents in different languages, with events simply defined as disease-location-number of victims triples. The second one covers a larger set of predefined events, ACE, which contains documents in several languages for the 2005 Automatic Content Extraction (ACE) evaluation ², with 8 events types, and 33 subtypes covering the most common events of national and international news (from a variety of sources selected from broadcast news programs, newspapers, newswire reports, internet sources and from transcribed audio).

2.1 DAnIEL Dataset

Since the DAnIEL system was designed to tackle specific type of text, a completely new dataset was created to fit its purposes [5]. The corpus consists of health articles from different news source from Google News that concentrated on epidemic events. As six different languages (English, French, Greek, Russian, Chinese, and Polish) were introduced, the data were annotated by native speakers of said languages to decide whether an article has a relevant event or not, and if yes, specify the disease name and location it occurs. Aside from language diversity, length of each document also deviate considerably from each other, varying from just one short paragraph to an article with complete structure.

As mentioned previously, a tuple of disease name-location defines a relevant **DAnIEL event**. In rarer cases, the annotation can include the number of victims affected by the disease, making the event a triplet of disease name-location-victims number. By representing an event this way, the task event extraction happens at document-level with the goal to identify articles that contain events that fit the description above and extract the best representation of the event i.e. single or compound words. Because of the spontaneous and haphazard nature of an epidemic outbreak, there is no pre-define list of types or subtypes of event, thus simplifying the detection process to just whether an article mentions a novel epidemic event or not.

An example is presented in Figure 1, where the number of victims is unknown.

```
"15962": {
  "annotations": [
    [
      "listeria",
      "USA",
      "unknown"
    ]
  ],
  "comment": "",
  "date_collecte": "2012-01-12",
  "language": "en",
  "document_path": "doc_en/20120112_www.businessweek.com_2a21025f6f4dc13c9eb8ebf3d",
  "url": "http://www.businessweek.com/news/2012-01-10/listeria-cantaloupe-outbreak"
},
```

Fig. 1. Example of an event annotated in DAnIEL dataset.

²<https://catalog.ldc.upenn.edu/ldc2006t06>

4 • Boros, et al.

A common characteristic of an event extraction dataset is the lack of balance in distribution. In the case of this corpus, documents that are relevant to epidemic events only occupy about 10% of the total dataset, which is very sparse. The number of documents per language, however, is relatively balance with 352 Polish documents (30 relevant), 446 in Chinese (16 relevant), 390 in Greek (26 relevant), and 475 in English (31 relevant). French is the only exception, having five times more documents than the other, with 2,733 documents, in which 340 of them are relevant. In total, the dataset comprises of 4,822 documents (489 relevant).

The DAnIEL dataset is annotated at document-level, which differentiates it from other datasets used in research for the event extraction task. A document is either reporting an event (disease-place pair, and sometimes the number of victims) or not. In order for us to be able to compare the two different models that we proposed, we transformed this annotation to sentence-level. The annotations provided by DAnIEL at document-level are looked-up in the appropriate file and the found offsets are attached to them. For example, the article below has the following annotations, at document level: **malaria**, **worldwide**, and **655000**.

GENEVA: Malaria caused the death of an estimated 655,000 people last year, with 86 percent of victims children aged under five, World Health Organisation figures showed on Tuesday. The figure marked a five percent drop in deaths from 2009. Africa accounted for 91 percent of deaths and 81 percent of the 216 million cases worldwide in 2010. In its annual World Malaria Report for 2011, the WHO hailed as a "major achievement" a 26 percent fall in mortality rates since 2000 despite being well short of its 50 percent target. The UN health agency aims to eradicate malaria deaths altogether by the end of 2015 and reduce the number of cases by 75 percent on 2000 levels.

In this case, in the first sentence, *GENEVA: **Malaria** caused the death of an estimated 655,000 people [. . .]*, we are able to annotate **Malaria** at positions relative to the the entire article 8 – 14. The process is automatic and continues in the same manner for the other annotations. In the case where one annotation is not found in the article (e.g. **655,000** is not recognised) it is disconsidered with the risk of penalty in evaluation. From a total of 1268 (disease names, place names, and number of patients), 1084 were identified in the DAnIEL dataset.

2.2 ACE 2005 Dataset

We used for our experiments the annotated ACE 2005 corpus provided by the ACE evaluation. ACE events are restricted to a range of types, each with a set of subtypes. Thus, only the events of an appropriate type are annotated in a document. The ACE dataset contains datasets in multiple languages (Chinese, Arabic, and English) with various types annotated for entities, relations, and events, from various information sources (e.g., broadcast conversations, broadcast news and telephone conversations). The data were created by Linguistic Data Consortium (LDC) with support from the ACE Program. The proposed tasks by ACE are more challenging than their MUC forerunners. In particular, the increased complexity resulted from the inclusion of various information sources and the introduction of more fine-grained entity types (e.g., facilities, geopolitical entities, etc.). In the context of this project, we use only the English ACE 2005 corpus that is composed of 599 articles. For the comparison of both models proposed, this dataset cannot be tested with the DAnIEL system, since it is designed only for epidemic related data.

An **ACE event** is represented by an *event mention* (a text contains an event of a specific type and subtype), *event trigger* (the word that expresses the event mention), *event argument* (a participant in the event of a specific type), *argument role* (the role that the entity has in the event).

Since the EE task in the context of ACE 2005 has two sub-tasks, the event detection represents the detection of the texts that contain an event of a specific type and the extraction of the event trigger from the text that expresses that type of event, and the event argument extraction, that is the detection of entities and their role in the event.

Every document is characterised by multiple events, or no events at all. The annotation of the event is done at the sentence level, and thus, the imbalanced nature of this dataset. If we consider, for instance, this example from ACE 2005 dataset:

*There was the free press in Qatar, Al Jazeera, but its offices in Kabul and Baghdad were **bombed** by Americans.,* an event detection system should output:

- *event mention*: this sentence contains an event of type **Conflict** and subtype **Attack**
- *event trigger*: this event of type **Conflict** and subtype **Attack** is triggered by the word **bombed**

An event argument extraction system should output:

- the *event arguments*: *Kabul and Baghdad*, which are entities of type **location**, and *Americans* which are considered an entity of type **person**
- the *event argument roles*: *Kabul and Baghdad* are **Places** and *Americans* have the **Attacker** role

3 APPROACHES

This section describes the approaches that will be evaluated, DANIEL and the CNN-based model.

3.1 DANIEL System

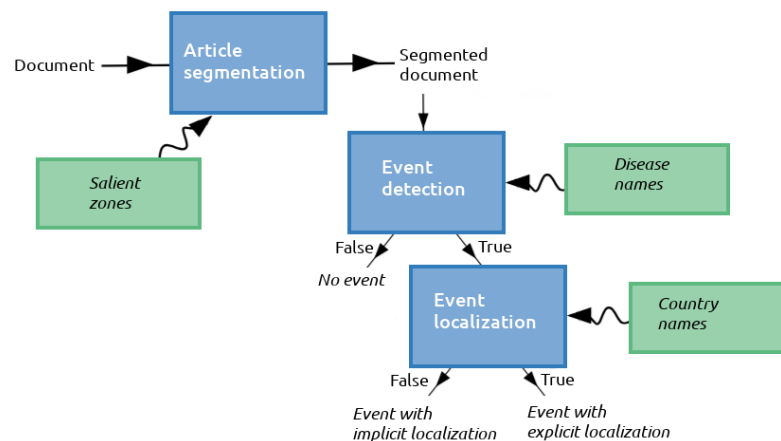


Fig. 2. Event Detection pipeline in DANIEL.

DANIEL [5] stands for Data Analysis for Information Extraction in any Language and it is an approach at discourse-level, as opposed to the commonly used analysis at sentence-level, by exploiting the global structure of news as defined by the authors of [9]. Entries in the system are news texts, including the title and the body of text, the name of the source when available, and other metadata (e.g date of article). As the name implies, the system is capable of working in a multilingual setting due to the fact that it does not utilise any word-based algorithm, which are highly language-specific, but rather a character-based one that centers around repetition and position [5]. By avoiding grammar analysis and the usage of other NLP toolkits (e.g Part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style [2, 9], the system is

6 • Boros, et al.

able to detect crucial information in salient zones that are peculiar to this genre of writing: the properties of the journalistic genre, the style universals, form the basis of the analysis.

Moreover, due to the fact that DAnIEL does not rely on any language-specific grammar analysis, and considers text as sequences of strings instead of words, it can quickly operate on any foreign language and extract crucial information early on and improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more than often, initial medical reports are in the vernacular language where patient zero appears [5].

DAnIEL uses a minimal knowledge base, its central processing chain includes four phases:

- **Article segmentation:** The system first divides the document into stylistic segments: title, header, body and footer. The purpose is to identify salient zones where important information is usually repeated.
- **Pattern extraction:** For detecting events, the system will look for repeated substrings at the salient zones aforementioned and determine whether they are maximal or not. A maximal substring is a string that cannot be extended to either its left nor right side [15].
- **Filtering of these patterns:** Substrings that satisfy this condition will be matched to a list of disease/location names that was constructed by crawling from Wikipedia. The reason for using Wikipedia to build the knowledge base is that it is convenient to add lexicons from new languages without the assistance of a native speaker since information on Wikipedia can be easily crawled from one language to another.
- **Detection of disease – location pairs** (in some cases, the number of victims also): The end result of processing a document with DAnIEL is one or more events that are described by pairs of disease-location.

3.2 Convolutional Neural Network-based Model

We chose a convolutional neural network (CNN) based model proposed by [1, 12] where the event detection (ED) task is modelled as a word classification task.

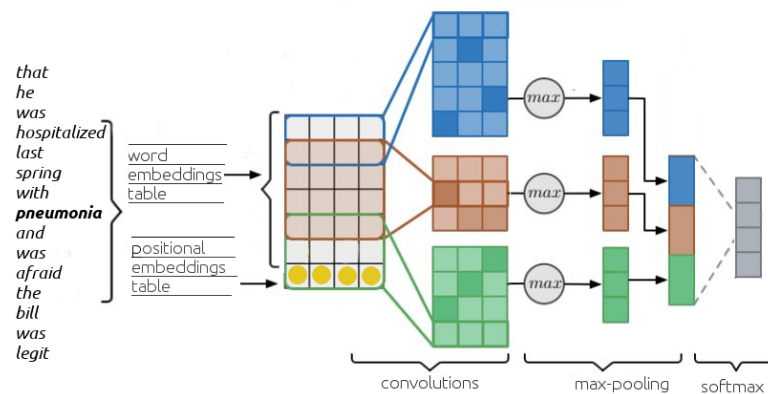


Fig. 3. CNN model for event detection, where *pneumonia* is the current event candidate in a context window of $2 \times 7 + 1$ words. Figure from [1].

Considering a sentence, we want to predict, for each word of the sentence, if the current token is a possible trigger of an event, thus the classification type is binary. The current token $x^{(i)}$ is surrounded by a context

that constitutes the main input for the CNN. The maximum size of a sentence is established on the training data. In order to consider a limited sized context, longer sentences are trimmed and shorter ones are padded with a special token. Let $x = [x^{(0)}, x^{(1)}, \dots, x^{(N)}]$ be a sentence with words from 0 to N . Given a document, we first generate a set of event candidates \mathcal{T} . For each event candidate $x^{(i)} \in \mathcal{T}$, we associate it a context window. We consider $2 \times n + 1$ the size of the context window, thus a trigger candidate $x^{(0)}$ is represented as $x = [x^{(-n)}, x^{(-n+1)}, \dots, x^{(0)}, \dots, x^{(n-1)}, x^{(n)}]$. Each context token $x^{(i)}$ has as features the word itself and the relative position of the token to the trigger candidate $x^{(0)}$. In this case, the distance 0 will be attributed to the trigger candidate $x^{(0)}$ and $-n, +n$ to the marginal tokens of the window, all the other relative distances in between $-n$ and $+n$ belong to the tokens in between. The position of an event trigger can be informative signal for this prediction task. Each core feature is embedded and represented in a d -dimensional space. Each feature (word, distance) is mapped to a vector retrieved from the following embedding tables:

- Word embedding table: initialised randomly or by pre-trained word embeddings;
- Positional embedding table: to embed the relative distance i of the token $x^{(i)}$ to the current token $x^{(0)}$. The table is initialised randomly, and these distance embedding vectors are then trained as regular parameters in the network [1, 12]

4 EXPERIMENTS

In order to create such an appropriate datasets, raw text from both datasets was extracted and converted into clean images. For the simulation of different levels of degradation on these images, we used DocCreator [4]. The rationale is to simulate what can be found in deteriorated documents due to time effect, poor printing materials or inaccurate scanning processes, which are common conditions in historical newspapers. We used four types of noise: *Character Degradation* adds small ink dots on characters to emulate the age effect on articles, *Phantom Character* appears when characters erode due to excessive use of documents, *Bleed Through* appears in double-paged document image scans where the content of the back side appears in the front side as interference, and *Blur* is a common degradation effect encountered during a typical digitisation process. After contaminating the corpus, all the text was extracted from noisy images using Tesseract optical character recognition (OCR) Engine v4.0³ [13] to produce the digitised documents, for initial clean images (without any adulteration) and the noisy synthetic ones. An example with the degradation levels is illustrated in Figure 4. The hyperparameters are presented in detail in [1].

The experiments were conducted in the following manner: for each noise type, the different intensity is generated to see its relation to the performance of the model. Character error rate (CER) and word error rate (WER) were calculated for each noise level, that can align long noisy text even with additional or missing text with the ground truth, thus enables it to calculate the error rate of OCR process. The experiments are performed under conditions of varying word error rate (WER) and character error rate (CER): Original text (no OCR, 0% WER, 0% CER); OCR from high-quality text images (~1% WER, ~0.5% CER); OCR on degraded text images synthetically produced with DocCreator (2–50% WER, 1–20% CER).

4.1 General Evaluation Setting

For the evaluation of the performance of the event detection task, we use the standard metrics: Precision (P), Recall (R), and F-measure (F1). For measuring the document distortion due to the OCR process, we also report the standard metrics: *character error rate* and *word error rate*.

Character error rate (CER) is defined as: $CER = (i_c + s_c + d_c) / n_c$ where n_c is the ground truth in terms of character, i_c , s_c , and d_c are the number characters that needed to insert, substitute and delete respectively to reconstruct the transcribed text into the ground-truth.

³<https://github.com/tesseract-ocr/tesseract>

8 • Boros, et al.



Fig. 4. Example of types of noise applied on ACE 2005 dataset: clean image, *Phantom Character*, *Character Degradation*, *Bleed Through*, *Blur*, and all mixed together.

Similarly, *Word Error Rate* (WER) is calculated as follows: $WER = (i_w + s_w + d_w)/n_w$ where all the parameters remain the same, except they are counted in words. It is worth noting that WER is generally higher than CER within the same sample, as WER is a stricter evaluation where any character mistake would make a whole word considered as wrong. On the other hand, CER is not as tight as the fore-mention, since the error in character is independent of each other and does not affect any previous or subsequent characters.

4.2 Experiments on DAnIEL Dataset

For the purpose of comparing the two approaches, the data with a total of 4822 documents was split at document level, 3857 documents for training (80%), 482 documents for validation (10%), and the rest of 483 documents for testing (10%), stratified by language, as shown in Table 1.

Evaluation framework. We perform two types of evaluations, both at the document level (specific DAnIEL):

- **Event identification:** a document represents an event if the triggers were found, regardless of their types
- **Event classification:** a document represents an event if the triggers are correctly found and match with the groundtruth ones

, Vol. 1, No. 1, Article . Publication date: October 2020.

Table 1. DAnIEL dataset splits. In (relevant), the number of documents annotated with events is reported.

	total documents	Polish	Chinese	Russian	Greek	French	English
Train	3,857 (377)	281 (22)	357 (13)	341 (28)	312 (16)	2,186 (269)	380 (29)
Validation	483 (51)	35 (3)	45 (2)	42 (6)	39 (5)	274 (33)	48 (2)
Test	482 (61)	36 (5)	44 (1)	43 (7)	39 (6)	273 (38)	47 (4)

Table 2. Evaluation of the CNN-based model and DAnIEL on the initial test data for event identification.

		Polish	Chinese	Russian	Greek	French	English	All languages
DAnIEL (%)	P	80	100	75.0	100	56.67	80	64.71
	R	80	100	85.71	100	89.47	100	90.16
	F1	80	100	80	100	69.39	88.89	75.34
CNN-based (%)	P	100	0	66.67	71.43	68.09	75	69.84
	R	40	0	28.57	83.33	84.21	75	72.13
	F1	57.14	0	40	76.92	75.29	75	70.97

4.2.1 Experiments with clean data. For event identification on clean textual data, one can notice from the Table 2, that usually DAnIEL favours recall instead of precision and tends to suffer from an imbalance between precision and recall, which may be due to the high imbalance of the data, while the CNN-based is more robust to this characteristic of the dataset. We can note also that DAnIEL seems to detect more relevant documents of lower quality, giving a higher cost to false positives and favouring in this way recall over precision, for all the cases. For Russian, Greek, and Polish (high inflectional languages), due to the annotation process (changing the DAnIEL format to the format accepted by the CNN), many of the events/words were not found in the text due to the inflections, and thus the CNN is not able to identify the documents that contain those events.

It is not surprising that the DAnIEL system has 100% rate of event identification for Chinese and Greek, since for Chinese, there is only one relevant document in the test set, and for Greek, there are only six of them. The CNN-based model is not able to detect the only relevant document in Chinese, and we assume that this is due to the lack of training data.

Table 3. Evaluation of the CNN-based model and DAnIEL on the initial test data for event classification.

		Polish	Chinese	Russian	Greek	French	English	All languages
DAnIEL (%)	P	30	50	25	58.33	50.48	40	42.35
	R	25	50	28.57	53.85	44.17	40	46.15
	F1	27.27	50	26.67	56	47.11	40	44.17
CNN-based (%)	P	100	0	66.67	50	60.23	75	60.75
	R	16.67	0	14.29	38.46	50.48	30	41.67
	F1	28.57	0	23.53	43.48	54.92	42.86	49.43

In the case of event classification, we can observe from Table 3, that DAnIEL is more balanced regarding the precision and recall metrics, being able to have higher F1 on the under-represented languages (Chinese, Russian, and Greek) than the CNN-based model. Analysing the results of DAnIEL, we noticed that, in all the cases, DAnIEL does not detect the number of victims. We assume that this is due to the fact that many of the annotated numbers

cannot be found in the text, e.g. 10000 cannot be detected since the original text has the 10,000 form, or it is spelled *ten thousands*. This applies to the CNN-based model since some numbers could not be annotated.

The values of recall for the CNN-based model are in general low. This might be related to the fact that the model is not able to detect some locations due to the fact they are not mentioned in the original text, whether DAnIEL is capable due to the usage of external resources and article metadata. The only relevant Chinese document in the testing data is annotated with a pair disease–location, but the location cannot be found in the text (one of the advantages of DAnIEL of using external resources). The DAnIEL system is able to detect correctly only the disease, but the CNN-based model cannot retrieve any of them correctly, even more, the location. Besides this, the small amount of data greatly affects the performance of the CNN-based model. We assume that the CNN-based model performs better for the French documents, due to the larger amount of data, and for the English documents, due to the fact that, in the annotation process, all the disease–location pairs (in the English documents, no number of victims was annotated) were located in the texts, and thus a higher chance of better performance.

Finally, the CNN-based performed slightly better in total than DAnIEL, with a difference of 5.26 percentage points in F1. We add also that one issue that needs to be further studied is DAnIEL's false positives problem: for instance documents relating vaccination campaigns are usually tagged as non-relevant in the ground truth dataset.

Table 4. Document degradation OCR evaluation on the DAnIEL dataset.

		Clean	CharDeg	Bleed	Blur	Phantom	All
All	CER	2.61	9.55	2.83	8.76	2.65	11.07
	WER	4.23	26.23	5.93	19.05	4.71	27.36
Polish	CER	0.15	5.86	0.19	7.57	0.19	5.51
	WER	0.74	20.66	1.17	13.23	1.17	20.70
Chinese	CER	36.89	41.01	38.24	43.97	36.91	46.97
	WER	–	–	–	–	–	–
Russian	CER	0.93	16.20	1.45	8.13	1.03	10.91
	WER	1.63	28.46	6.61	14.94	2.73	29.72
Greek	CER	3.52	9.04	3.76	13.79	3.54	16.28
	WER	15.86	41.36	17.39	54.02	15.93	54.76
French	CER	1.96	8.37	2.13	7.43	2.0	10.90
	WER	3.33	23.56	4.89	16.31	3.76	26.07
English	CER	0.35	5.75	0.52	4.74	0.44	7.43
	WER	0.66	24.78	2.14	14.72	1.66	20.99

4.2.2 Experiments with noisy data. The results in Table 4 clearly state that *Character Degradation* is the effect that affects the most the transcription of the documents. However, for character-based languages (e.g. Chinese), CER is commonly used instead of WER as the measure for OCR, and, thus, we report only the CER [16].

Also, regarding the Chinese documents, the high values for CER, for every type of noise, might be caused by the existence of the enormous number of characters in the alphabet that, by adding such an effect as *character Degradation* can change drastically the recognition of a character (and in Chinese, one single character can often be a word). Otherwise, while *Character Degradation* noise and *Blur* effect have more impact on the performance of DAnIEL than *Phantom Character* type since it did not generate enough distortion to the images. A similar case applies for the *Bleed Through* noise.

Next, we present the results for event classification for both systems. Results indicated in bold are the best F1 scores given by the system according to the type of degradation. We compute also a δ measure that gives the minimum decrease rate between the F1 given using clean data and the F1 given using noisy data for each type of degradation. This measure represents the perfect system which will give the best F1 for all degradation levels. We also present the evolution of the δ measure according to the types of noise, for both systems, for each language. Due to the high level of detail, and we include here only the evolution of δ for all the languages together.

Table 5. Evaluation of DANIEL results on the noisy test data for event identification.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	64.71	77.61	79.63	78.79	79.31	77.61	82.98
	R	90.16	85.25	70.49	85.25	75.41	85.25	63.93
	F1	75.34	81.25	74.78	81.89	77.31	81.25	72.22
Polish	P	80	80	100	80	75	80	80
	R	80	80	40	80	60	80	80
	F1	80	80	57.14	80	66.67	80	80
Chinese	P	100	100	100	100	100	100	100
	R	100	100	100	100	100	100	100
	F1	100	100	100	100	100	100	100
Russian	P	75.0	66.67	75	75	75	66.67	83.33
	R	85.71	85.71	85.71	85.71	85.71	85.71	71.43
	F1	80	75	80	80	80	75	76.92
Greek	P	100	83.33	75	83.33	100	83.33	100
	R	100	83.33	50	83.33	33.33	83.33	50
	F1	100	83.33	60	83.33	50	83.33	66.67
French	P	56.67	78.05	80.56	78.05	80	78.05	82.76
	R	89.47	84.21	76.32	84.21	84.21	84.21	63.16
	F1	69.39	81.01	78.38	81.01	82.05	81.01	71.64
English	P	80	80	66.67	80	66.67	80	66.67
	R	100	100	50	100	50	100	50
	F1	88.89	88.89	57.14	88.89	57.14	88.89	57.14

Regarding the experiments with the DANIEL system, from the Table 5 we notice, first of all, that the *Character Degradation* effect, *Blur*, and most of all, all the effects mixed together, have indeed an impact or effect over the performance of DANIEL, but with little variability. Meanwhile, *Phantom Degradation* and *Bleed through* had very little to no impact on the quality of detection with DANIEL.

The cause of the decrease in performance of DANIEL is that to detect events, it looks for repeated substrings at salient zones. In the case of many incorrectly recognised words during the OCR process, there may be no repetition anymore, implying that the event will not be detected. However, since DANIEL only needs 2 occurrences of its clues (substring of a disease name and substring of a location), it is assumed to be robust to the loss of many repetitions, as long as 2 repetitions remain in salient zones.

For all the languages, Figure 5, δ can exceed 5% when using noisy data, with WER and CER reaching more than 9 and 25 respectively.

Tables 7 and 8 analyse the effect of applying noise on the document images for the CNN-based model. The decrease in precision and recall is produced similar to the DANIEL system, the impact on the scores being higher for the *Character Degradation*, *Blur*, and all mixed together, also. One drawback of this model is that it is based on

Table 6. Evaluation of DANIEL results on the noisy test data for event classification.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	42.35	52.24	53.7	53.03	53.45	52.24	60.64
	R	46.15	44.87	37.18	44.87	39.74	44.87	36.77
	F1	44.17	48.28	43.94	48.61	45.59	48.28	45.78
Polish	P	30	30	25	30	25	30	40
	R	25	25	8.33	25	16.67	25	33.33
	F1	27.27	20	12.5	27.27	20	20	36.36
Chinese	P	50	50	50	50	50	50	50
	R	50	50	50	50	50	50	50
	F1	50	50	50	50	50	50	50
Russian	P	25	27.78	31.25	31.25	18.75	27.78	33.33
	R	28.57	35.71	35.71	35.71	21.43	35.71	28.57
	F1	26.67	31.25	33.33	33.33	20	31.25	30.77
Greek	P	58.33	41.67	25	41.67	25	41.67	50
	R	53.85	38.46	15.38	38.46	7.69	38.46	23.08
	F1	56	40	19.05	40	11.76	40	31.58
French	P	50.48	63.41	65.28	63.41	67.5	63.41	74.14
	R	44.17	49.52	44.76	49.52	51.43	49.52	41.35
	F1	47.11	55.61	53.11	55.61	58.38	55.61	53.09
English	P	40	40	33.33	40	16.67	40	33.33
	R	40	40	20	40	100	40	20
	F1	40	40	25	40	12.5	40	25

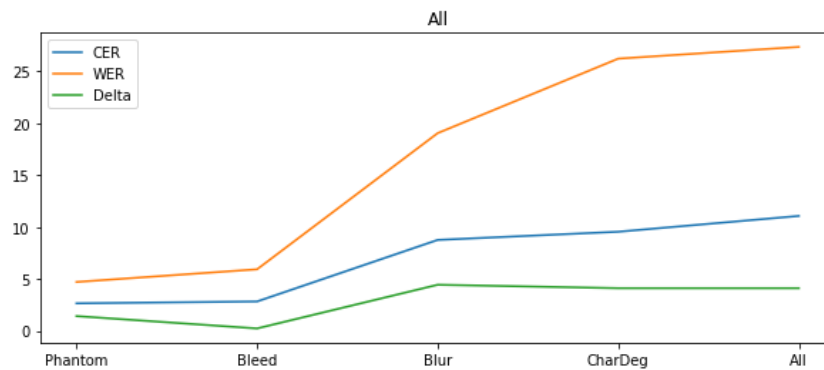


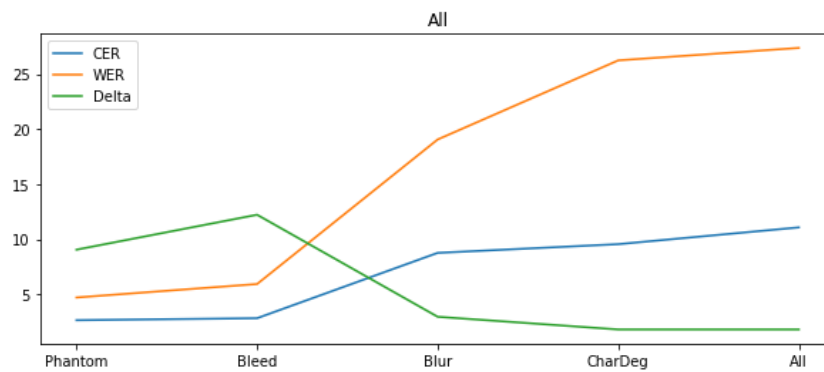
Fig. 5. F1 degradation according to OCR error rates for event classification for the DANIEL system, for all the languages

embeddings at word-level, which can degrade the performance in the case of many modified words in the test set during the OCR process.

Studying the degree of variability of F1-scores for all the effects mixed together for event identification, and for event classification, we notice the CNN-based model is more sensitive to the added effects, as shown in Figure

Table 7. Evaluation results of the CNN-based model on the noisy test data for event identification.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	69.84	68.52	44.26	70.83	82.35	68.52	72.22
	R	72.13	60.66	72.97	55.74	45.9	60.66	42.62
	F1	70.97	64.35	55.1	62.39	58.95	64.35	53.61
Polish	P	100	100	0	100	100	100	0
	R	40	20	0	40	20	20	0
	F1	57.14	33.33	0	57.14	33.33	33.33	0
Chinese	P	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
	F1	0	0	0	0	0	0	0
Russian	P	66.67	66.67	28.57	100	100	66.67	100
	R	28.57	28.57	66.67	14.29	28.57	66.67	28.57
	F1	40	40	40	25	44.44	40	44.44
Greek	P	71.43	100	16.67	100	0	100	0
	R	83.33	50	100	16.67	0	50	0
	F1	76.92	66.67	28.57	28.57	0	66.67	0
French	P	68.09	65.22	60.53	67.44	80.65	65.22	70.59
	R	84.21	78.95	71.88	76.32	65.79	78.95	63.16
	F1	75.29	71.43	65.71	71.6	72.46	71.43	66.67
English	P	75	100	25	100	0	100	0
	R	75	25	100	25	0	25	0
	F1	75	40	40	40	0	40	0



6. We conclude that using representations at word-level in the CNN-based model indeed hurts the performance of the model when evaluated on the text transcribed from degraded images.

Regarding all the results aforementioned, for the DAnIEL system, and the CNN-based model, computing the number of affected event words (disease, location, number of patients), we also notice that a very small number of them have been modified by the OCR process, only 1.98% for all the languages together, for all the effects mixed together, not far from the 1.63% that were affected by the OCR on clean data. This is due to the fact that DAnIEL

Table 8. Evaluation results of the CNN-based model on the noisy test data for event classification.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	60.75	62.5	67.8	65.88	75.44	62.5	74.07
	R	41.67	38.46	25.64	35.9	27.56	38.46	25.81
	F1	49.43	47.62	37.21	46.47	40.38	47.62	38.28
Polish	P	100	100	0	100	100	100	0
	R	16.67	8.33	0	16.67	8.33	8.33	0
	F1	28.57	15.38	0	28.57	15.38	15.38	0
Chinese	P	0	0	0	0	0	0	0
	R	0	0	0	0	0	0	0
	F1	0	0	0	0	0	0	0
Russian	P	66.67	66.67	66.67	100	100	66.67	100
	R	14.29	14.29	14.29	7.14	14.29	14.29	14.29
	F1	23.53	23.53	23.53	13.33	25	23.53	25
Greek	P	50	60	100	100	0	60	0
	R	38.46	23.08	7.69	15.38	0	23.08	0
	F1	43.48	33.33	14.29	26.67	0	33.33	0
French	P	60.23	61.63	66.67	63.29	74.07	61.63	73.08
	R	50.48	50.48	34.29	47.62	38.1	50.48	36.54
	F1	54.92	55.5	45.28	54.35	50.31	55.5	48.72
English	P	75	100	100	100	0	100	0
	R	30	10	100	10	0	10	0
	F1	42.86	18.18	100	18.18	0	18.18	0

dataset is highly imbalanced (only 10.14% of a total of 4,822 documents contain events), and it brings us to the conclusion that the event detection task is not considerably impacted by the degradation of the image documents.

One interesting observation is that the precision or the recall can increase, resulting in a higher F1, despite the higher noise effect applied, for event classification with the CNN-based model, where the δ is decreasing when applying all the degradation types. From our observation, it is because of that with a greater level of noise some false positives disappear. Documents, which were previously classified wrongly due to being too ambiguous to the system (for instance documents relating vaccination campaigns are usually tagged as non-relevant in the ground truth dataset), were given much more distinction due to the noise, thus making them look less like relevant samples to the system. This may seem counter-intuitive but noise can improve classification results, see for instance [6] for a study on the same dataset of the influence of boilerplate removal on results.

4.3 Experiments on ACE 2005 Dataset

For comparison purposes, we use the same test set with 40 newswire articles (672 sentences), the same development set with 30 other documents (863 sentences) and the same training set with the remaining 529 documents (14,849 sentences) as in previous studies of this dataset [3, 7, 8, 11, 12].

The hyperparameters used for the CNN model for event detection are as follows. The window sizes used in the experiments are in the set $\{1, 2, 3\}$ to generate feature maps and 300 feature maps are used for each window size in this set. After each convolutional layer, a *ReLU* nonlinear layer is applied with orthogonal weights initialisation. The window size for triggers is also set to 31 and the dimensionality of the position embeddings is 50 [12]. The size of the batch is set to 256 and we employed also the pre-trained word embeddings *Word2vec* for Google News

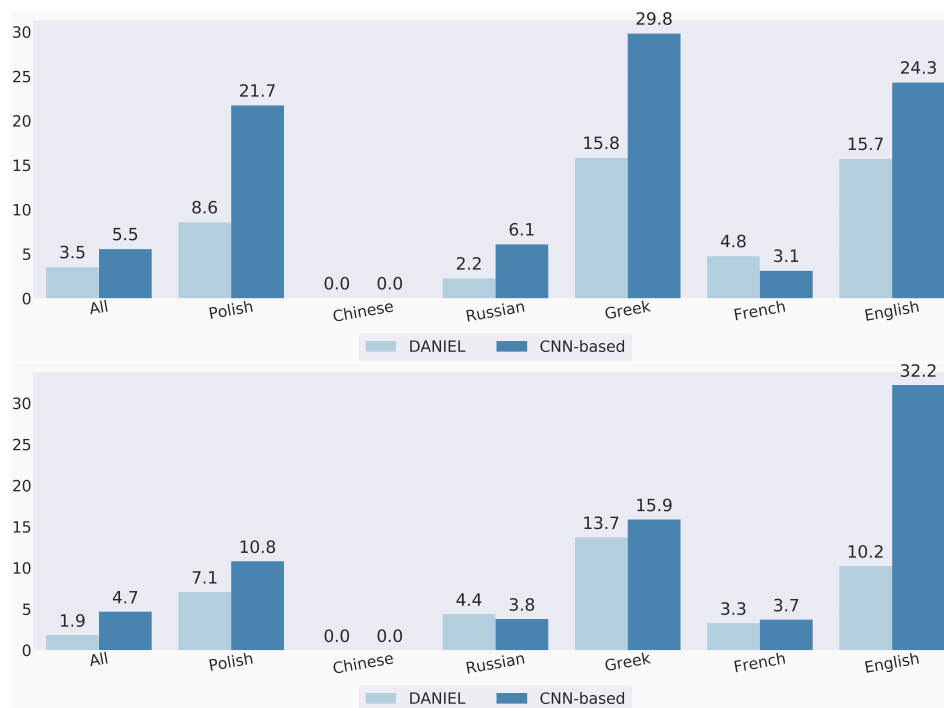


Fig. 6. Standard deviations of the F1-scores for all effects mixed together, per language, for the event identification, upper figure, and for the event classification, respectively, lower figure.

[10]. We would also stress the fact that the batch size affects the Adam optimizer [14], and thus our choice of 256, which performed the best on the validation set. Also, deep learning models are stochastic and use randomness (e.g. random initial weights, random shuffling) while being trained on a dataset and, because of this, a common practice is to run the algorithms several times and to report a measure of variability. Thus we report the precision, recall and F1 in terms of means and standard deviations.

We perform the following evaluation from the ACE 2005 evaluation:

- **Event classification:** a trigger is correct if its event subtype and offsets match those of a reference trigger

4.3.1 Experiments with clean data. For the experiments with clean ACE 2005 data, we replicated the model presented in [12]. However, the source code for [12] has not been published and the reproducibility of their system has proven difficult due to different reasons: the choice of the NLP tools for pre-processing (sentence splitting, tokenization), which can influence considerably such a system, the hyperparameters are vaguely presented and some assumptions and choices are not stated. In order to replicate their results, we tuned the model parameters on the development data and obtained a different configuration of parameters presented in Subsection 4.3.

Table 9. Evaluation of the CNN-based model on the test data for event classification.

	P	R	F1
CNN [12] (reported)	71.9	63.8	67.6
CNN [12] (replicated)	68.88 \pm 0.69	58.45 \pm 1.56	63.18 \pm 0.91
Our CNN (replicated, changed hyperparameters)	68.82 \pm 0.83	66.13 \pm 1.24	67.40 \pm 0.51

4.3.2 *Experiments with noisy data.* The Table 11 illustrates the effect of applying noise on the document images for the CNN-based model. The decrease in precision and recall is produced similar to the DAnIEL system, the impact on the scores being higher for the *Character Degradation*, *Blur*, and all mixed together, also. We recall that one drawback of this model is that it is based on pre-defined set of word embeddings, which can degrade the performance in the case of many wrongly detected words by in the OCR process.

Table 10. Evaluation results on the noisy test data for event classification. CharDeg = character degradation, Bleed = Bleed through, All = CharDeg + Bleed + Phantom + Blur.

	Original	Clean	CharDeg	Bleed	Blur	Phantom	All
CER	0	0.83	4.10	1.34	7.28	0.95	14.81
WER	0	1.13	17.96	5.61	18.49	2.50	35.93
Affected triggers	0	0.94	19.05	2.11	19.05	0.94	41.17

Table 11. Evaluation results on the noisy test data for event classification. CharDeg = character degradation, Bleed = Bleed through, All = CharDeg + Bleed + Phantom + Blur.

	Original	Clean	CharDeg	Bleed	Blur	Phantom	All
P	68.82 \pm 0.83	68.62 \pm 1.23	47.63 \pm 1.09	57.75 \pm 1.05	67.55 \pm 1.30	59.05 \pm 1.52	48.02 \pm 0.79
R	66.13 \pm 1.24	65.51 \pm 0.78	50.54 \pm 0.92	64.37 \pm 0.87	53.77 \pm 1.19	64.94 \pm 1.34	35.48 \pm 0.66
F1	67.40 \pm 0.51	66.97 \pm 0.14	48.97 \pm 0.44	60.82 \pm 0.20	59.80 \pm 0.59	61.72 \pm 0.30	40.77 \pm 0.36
CER	0	0.83	4.10	1.34	7.28	0.95	14.81
WER	0	1.13	17.96	5.61	18.49	2.50	35.93

Analysing the results, we notice that for all the noise effects together, 41.17% of the trigger words were affected as shown in Table 10, which is a large amount of event triggers, and for this reason, a large drop in performance of almost 27 percentage points in F1.

5 CONCLUSIONS

We conclude that, in general, event detection is prone to errors induced by an imperfect OCR, depending on the level of data imbalance. The DAnIEL dataset was highly imbalanced and the variability in results was lower than in the case of the ACE 2005, and thus the probability that the few words annotated as events were affected was quite low. Moreover, ACE 2005 has a much higher number of events and event types in almost every document: 92.32% of the documents are relevant, while in the DAnIEL dataset, only 10.14% are. Comparing the models, the CNN-based model is more impacted by the effects of the noise added to the images than the DAnIEL system.

We believe this is due to the more robust string-level representation used by the DAnIEL system, compared to the word-level representation of other approaches. The lesser impact on the DAnIEL system, meanwhile, can also be explained by the fact that the model uses external resources in order to predict the presence of an event. One disadvantage of this model might be its exclusive applicability to epidemic events, and the amount of effort needed in order to adapt it to other domains (e.g. Wikipedia seeds for different domains need to be provided). An advantage that is common to both models is language independence. In future work, we consider to approach the alleviation of these digitisation errors with the inclusion of data augmentation and perhaps the usage of language models for their ability to represent words and subwords contextually.

6 ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

REFERENCES

- [1] Emanuela Boros. 2018. *Neural Methods for Event Extraction*. Ph.D. Dissertation.
- [2] Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions. https://doi.org/10.1007/978-3-319-78105-1_39
- [3] Heng Ji, Ralph Grishman, et al. 2008. Refining Event Extraction through Cross-Document Inference.. In *ACL*. 254–262.
- [4] Nicholas Journet, Muriel Visani, Boris Mansencal, Kieu Van-Cuong, and Antoine Billy. 2017. DocCreator: A New Software for Creating Synthetic Ground-Truthed Document Images. *Journal of imaging* 3, 4 (2017), 62.
- [5] Gaël Lejeune, Romain Brixel, Antoine Doucet, and Nadine Lucas. 2015. Multilingual Event Extraction for Epidemic Detection. *Artificial intelligence in medicine* 65 (07 2015). <https://doi.org/10.1016/j.artmed.2015.06.005>
- [6] Gaël Lejeune and Lichao Zhu. 2018. A New Proposal for Evaluating Web Page Cleaning Tools. *Computacion y Sistemas* 22, 4 (2018), 1249–1258.
- [7] Qi Li, Heng Ji, and Liang Huang. 2013. Joint Event Extraction via Structured Prediction with Global Features.. In *ACL (1)*. 73–82.
- [8] Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 789–797.
- [9] Nadine Lucas. 2009. *Modélisation différentielle du texte, de la linguistique aux algorithmes*. Ph.D. Dissertation. Université de Caen.
- [10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR 2013), workshop track*.
- [11] Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of NAACL-HLT*. 300–309.
- [12] Thien Huu Nguyen and Ralph Grishman. 2015. Event Detection and Domain Adaptation with Convolutional Neural Networks.. In *ACL (2)*. 365–371.
- [13] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.
- [14] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. 2017. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489* (2017).
- [15] Esko Ukkonen. 2009. Maximal and minimal representations of gapped and non-gapped motifs of a string. *Theoretical Computer Science* 410 (10 2009), 4341–4349. <https://doi.org/10.1016/j.tcs.2009.07.015>
- [16] Peilu Wang, Ruihua Sun, Hai Zhao, and Kai Yu. 2013. A New Word Language Model Evaluation Metric for Character Based Languages. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 315–324.

8 Impact Analysis of Document Digitization on Event Extraction

Impact Analysis of Document Digitization on Event Extraction*

Nhu Khoa Nguyen¹[0000–0003–2751–5349], Emanuela Boros¹[0000–0001–6299–9452],
Gaël Lejeune²[0000–0002–4795–2362], and Antoine Doucet¹[0000–0001–6160–3356]

¹ University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr
<https://www.univ-larochelle.fr>

² Sorbonne University, F-75006 Paris, France
firstname.lastname@sorbonne-universite.fr
<https://www.sorbonne-universite.fr/>

Abstract. This paper tackles the task of event extraction in the epidemiological field applied to digitized documents. Event extraction is an information extraction task that focuses on identifying event mentions from textual data. In the context of event-based health surveillance from digitized documents, several key issues remain challenging in spite of great efforts. First, image documents are indexed through their digitized version and thus, they may contain numerous errors, e.g. misspellings. Second, in this field, it is important to address international news, which would imply the inclusion of multilingual data. To clarify these important aspects of how to extract epidemic-related events, it remains necessary to maximize the use of digitized data. In this paper, we investigate the impact of working with digitized multilingual documents with different levels of synthetic noise over the performance of a specialized event extraction system. This type of analysis, to our knowledge, has not been alleviated in previous research.

Keywords: Information Extraction · Event Extraction · Event Detection · Multilingualism.

1 Introduction

The surveillance of epidemic outbreaks has been an ongoing challenge globally and it has been a key component of public health strategy to contain diseases spreading. While digital documents have been the standard format in the modern days, many archives and libraries still keep printed historical documents and records. Historians and geographers have a growing interest in these documents as they still hold many crucial information and events in the past to analyze, noticeably in health and related to epidemics events in an international context.

* This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

Event extraction (EE) is an important information extraction (IE) task that focuses on identifying event mentions from text and extracting information relevant to them. Typically, this entails predicting event triggers, the occurrence of events with specific types, and extracting arguments associated with an event. In the context of event-based health surveillance from digitized documents, for extracting relevant events, even though the historical documents are in physical form, few of them have been converted into digital form for further storage as records in a database. However, due to the digitization process, several issues can arise, most commonly in the case when the original document is distorted, whether through deterioration due to aging or was damaged in the storing process, which will affect the converted content. Moreover, errors from the digitization process could also be a factor that causes adulteration of the converted documents e.g. word variations or misspellings.

In this article, we propose to experiment with an approach to event extraction with the ability of handling not only multilingual data, but also large amounts of data without relying on any additional natural language processing (NLP) tools. The architecture is based on the DANIEL system [11] which is a discourse-level approach that exploits the global structure of news. It also tackles the difficulty of language adaptation by its character-based approach that uses positions of substring occurrences in text. We believe that DANIEL is adequate for its ability to handle text in any language and that its algorithm should be robust to noise. We aim at testing the robustness of this model against noise, its ability of treating highly inflected languages and misspelled or unseen words, which can be either due to the low quality of text or the spelling variants. For these experiments, we present the evaluation general settings. Furthermore, we create synthetic data starting from the initial dataset in order to study the direct impact of automatic text recognition (ATR) over the performance of both approaches.

The paper is organized as follows: Section 2 briefly overviews the related works on epidemiological event extraction. Section 3 introduces the DANIEL system and its characteristics and in Section 4 the dataset built specifically for the DANIEL system is presented in detail. The Section 5 describes the experiments and an extrinsic evaluation of the results. We conclude and propose possible suggestions for future research in Section 6.

2 Related Work

Specific to epidemiological event extraction, there exist a few of empirical works targeted to extract events related to disease outbreaks. For instance, similar to the chosen system for this paper, DANIEL, there are two other systems, BIOCASTER [2,3] and PULS [5]. These architectures produced adequate results in analyzing disease-related news reports and providing a summary of the epidemics. For example, the BIOCASTER, an ontology-based text mining system, processes and analyzes web texts for the occurrence of disease outbreak in four phases namely, topic classification, named entity recognition (NER), disease/location detection and event extraction.

To our knowledge, there are no works related to the analysis of the impact of documents digitization for event extraction in the epidemiological domain. In return, few studies have been devoted to other information extraction tasks i.e. the extraction of named entities from digitized historical data [1,4]. Dealing with noisy data, several efforts have been devoted to extracting named entities from diverse text types such as outputs of automatic speech recognition (ASR) systems [6,9], informal messages and noisy social network posts [16]. The authors of [8] quantitatively estimate the impact of digitization quality on the performance of named entity recognition. Other studies focused on named entity linking, for example, [15] evaluate the performance of named entity linking over digitized documents with different levels of digitization quality.

3 Approach

DAnIEL [11] stands for Data Analysis for Information Extraction in any Language. The approach is at document-level, as opposed to the commonly used analysis at sentence-level, by exploiting the global structure of news as defined by the authors of [14]. The entries of the system are news texts, title and body of text, the name of the source when available, and other metadata (e.g date of article). As the name implies, the system has the capability to work in a multilingual setting due to the fact that it is not a word-based algorithm, segmentation in words can be highly language-specific, but rather a character-based one that centers around the repetition and position of character sequences.

By avoiding grammar analysis and the usage of other NLP toolkits (e.g Part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style [7,14], the system is able to detect crucial information in salient zones that are peculiar to this genre of writing: the properties of the journalistic genre and the style universals form the basis of the analysis. This combines with the fact that DAnIEL considers text as sequences of characters, instead of words, the system can quickly operate on any foreign language and extract crucial information early on and improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more than often, initial reports where patient zero appears are in the vernacular language.

DAnIEL uses a minimal knowledge base for matching between the extracted possible disease names or locations and the knowledge base entries. Its central processing chain includes four phases. In the *Article segmentation* phase, the system first divides the document into salient positions: title, header, body and footer. In *Pattern extraction*, for detecting events, the system looks for repeated substrings at the salient zones aforementioned. In *Pattern filtering*, the substrings that satisfy this condition will be matched to a list of disease/location names that was constructed by crawling from Wikipedia.

For the string matching between the extracted character sequences and knowledge base entries, the system is parameterized with a ratio. For instance, a small ratio value could offer a perfect recall but with high noise (many irrelevant entries are selected). For a maximum value (1.0), the system will match the exact

extracted substrings which could be detrimental to the morphologically rich languages (e.g. Greek, Russian). There are cases where the canonical disease name cannot be found in the text, as in the case of aforementioned languages, but grammatical cases of nouns. For example, in Russian, “Простуда” (“prostuda”) means “cold”, and since this disease name cannot be found in the text article, we used the instrumental case in Russian that can generally be distinguished by the “-ом” (“-om”) suffix for most masculine and neuter nouns, the “-ою/“-ой” (“-oju/“-oj”) suffix for most feminine nouns. A ratio of less than 1.0 will consider the instrumental case for singular “простудой” as a true positive.

Finally, the *Detection of disease – location pairs* (in some cases, the number of victims also) produces the end result with one or more events that are described by pairs of disease-location.

4 Dataset Description

In this section, we present the dataset that was created for the DANIEL system [11]. The corpus is dedicated to multilingual epidemic surveillance and contains articles on different press threads in the field of *health* (Google News) that focused on epidemic events from different collected documents in different languages, with events simply defined as disease-location-number of victims triplets.

The corpus was built specifically for this system [11,12], containing articles from six different languages: English, French, Greek, Russian, Chinese, and Polish. It contains articles on different press threads in the field of *health* (Google News) focused on epidemic events. These documents have lengths that vary substantially, ranging from a short dispatch with one paragraph to a long article with a more detailed structure. Annotators, native speakers of the aforementioned languages, decide whether an article is relevant (speaks about an event) or not and then provide the disease name and location of the event.

A DANIEL event [11] is defined at document-level, meaning that an article is considered as relevant if it is annotated with a (disease, location, number of victims) triplet, or a (disease, location) pair. An example is presented in Figure 1, where the event is a *listeria* outbreak in *USA* and number of victims is unknown.

Thus, in this dataset the event extraction task is defined as identifying articles that contain an event and the extraction of the disease name, location, number of victims, i.e. the words or compound words that evoke the event. Since the events are epidemic outbreaks, there is no pre-set list of types and subtypes of events, and thus the task of event extraction is simplified to detecting whether an article contains an ongoing epidemic event or not. Throughout the paper, we refer to the disease name or the location as event triggers (considering that these words most clearly express an epidemiological event).

Common to event extraction, the dataset is characterized by imbalance. In this case, only around 10% of these documents are relevant to epidemic events, which is very sparse. The number of documents in each language is rather balanced, except for French, having about five times more documents compared to the rest of the languages. More statistics on the corpus can be found in Table 1.

Fig. 1. Example of an event annotated in DANIEL dataset.

```

"15960": {
  "annotations": [
    [
      "listeria ",
      "USA",
      "unknown"
    ]
  ],
  "comment": "",
  "date": "2012-01-12",
  "language": "en",
  "document_path": "doc_en/20120112_www.cnn.com_48eddc7c17447b70075c26a1a3b168243edcbfb28f0185",
  "url": "http://www.cnn.com/2012/01/11/health/listeria-outbreak/index.html"
}

```

Table 1. Summary of the DANIEL dataset. The number of documents annotated with events is reported in brackets.

total documents	Polish	Chinese	Russian	Greek	French	English
4,822 (489)	352 (30)	446 (16)	426 (41)	390 (26)	2,733 (340)	475 (31)

The DANIEL dataset is annotated at document-level, which differentiates itself from other datasets used in research for the event extraction task. A document is either reporting an event (disease-place pair, and sometimes the number of victims) or not. We will elaborate the evaluation framework in the Section 5.

5 Experiments

In the case of historical newspapers, we are particularly keen on evaluating the performance of the models over texts that were the results of an automatic text recognition (ATR) process, as historical documents are evidently not digitally-born. The focal point of this set of experiments is to observe how the level of noise stemming from the digitization process impacts the performance of the models. However, there is no adequate historical document dataset provided with manually curated event annotation that could directly be used to measure the performance of the models over deteriorated historical documents. Thus, the noise and degradation levels have to be artificially generated into clean documents, so as to measure the impact of ATR over event detection using DANIEL. We shall thus use readily available data sets over contemporary and digitally-born datasets, which are free of any ATR-induced noise.

In order to create such an appropriate dataset, the raw text from the DANIEL dataset was extracted and converted into clean images³. The rationale is to simulate what can be found in deteriorated documents due to time effect, poor printing materials or inaccurate scanning processes, which are common conditions in historical newspapers. We used four types of noise: *Character Degradation* adds small ink dots on characters to emulate the age effect on articles, *Phantom Character* appears when characters erode due to excessive use of documents, *Bleed Through* appears in double-paged document image scans where the content of the back side appears in the front side as interference, and *Blur* is a common degradation effect encountered during a typical digitization process. After contaminating the corpus, all the text was extracted from noisy images⁴, for initial clean images (without any adulteration) and the noisy synthetic ones. An example with the degradation levels is illustrated in Figure 2. The noise levels were empirically chosen with a considerable level of difficulty⁵.

The experiments were conducted in the following manner: for each noise type, the different intensity is generated to see its relation to the performance of the model. Character error rate (CER) and word error rate (WER) were calculated for each noise level, that can align long noisy text even with additional or missing text with the ground truth, thus enables it to calculate the error rate of OCR process. The experiments are performed under conditions of varying word error rate (WER) and character error rate (CER):

- Original text (no OCR, 0% WER, 0% CER)
- OCR from high-quality text images (~1% WER, ~0.5% CER)
- OCR on degraded text images synthetically produced with DocCreator (2–50% WER, 1–20% CER)

5.1 Evaluation framework

For the evaluation of the performance of the event detection task, we use the standard metrics: Precision (P), Recall (R), and F-measure (F1). For measuring the document distortion due to the OCR process, we also report the standard metrics: character error rate (CER) and word error rate (WER).

We perform two types of evaluations, both at the document level (included in the DANIEL system):

- Event identification: a document represents an event if both triggers were found, regardless of their types
- Event identification and classification: a document represents an event if the triggers are correctly found and match exactly with the groundtruth ones

³ For simulating different levels of degradation, we used DocCreator [10]

⁴ The Tesseract optical character recognition (OCR) Engine v4.0 <https://github.com/tesseract-ocr/tesseract> [17] was used to produce the digitised documents

⁵ The following values of DocCreator are: *Character Degradation* (2-6), *Phantom Character* (Very Frequent), *Blur* (1-3), *Bleed Through* (80-80)

Impact Analysis of Document Digitization on Event Extraction

7

nd insisted that the United States would not bow to blackmail , threats or bellicose statements .
 out Pyongyang s bold plan for peace while offering nothing new itself .
 ade no reference to the reports that North Korea had admitted to possessing nuclear weapons .
 oncerns of the DPRK and the US , and the parties concerned with the nuclear issue on the Korean peninsula
 re dialogue without advancing any new proposal at the talks .

nd insisted that the United States would not bow to blackmail , threats or bellicose statements .
 out Pyongyang s bold plan for peace while offering nothing new itself .
 ade no reference to the reports that North Korea had admitted to possessing nuclear weapons .
 oncerns of the DPRK and the US , and the parties concerned with the nuclear issue on the Korean peninsula
 re dialogue without advancing any new proposal at the talks .

nd insisted that the United States would not bow to blackmail , threats or bellicose statements .
 out Pyongyang s bold plan for peace while offering nothing new itself .
 ade no reference to the reports that North Korea had admitted to possessing nuclear weapons .
 oncerns of the DPRK and the US , and the parties concerned with the nuclear issue on the Korean peninsula
 re dialogue without advancing any new proposal at the talks .

nd insisted that the United States would not bow to blackmail , threats or bellicose statements .
 out Pyongyang s bold plan for peace while offering nothing new itself .
 ade no reference to the reports that North Korea had admitted to possessing nuclear weapons .
 oncerns of the DPRK and the US , and the parties concerned with the nuclear issue on the Korean peninsula
 re dialogue without advancing any new proposal at the talks .

nd insisted that the United States would not bow to blackmail , threats or bellicose statements .
 out Pyongyang s bold plan for peace while offering nothing new itself .
 ade no reference to the reports that North Korea had admitted to possessing nuclear weapons .
 oncerns of the DPRK and the US , and the parties concerned with the nuclear issue on the Korean peninsula
 re dialogue without advancing any new proposal at the talks .

nd insisted that the United States would not bow to blackmail , threats or bellicose statements .
 out Pyongyang s bold plan for peace while offering nothing new itself .
 ade no reference to the reports that North Korea had admitted to possessing nuclear weapons .
 oncerns of the DPRK and the US , and the parties concerned with the nuclear issue on the Korean peninsula
 re dialogue without advancing any new proposal at the talks .

Fig. 2. Example of types of noise applied on a dataset: (i) clean image, (ii) *Phantom Character*, (iii) *Character Degradation*, (iv) *Bleed Through*, (v) *Blur*, and (vi) all mixed together.

5.2 Experiments with clean data

Hereafter, we present the experiments performed with the clean data. Considering that the DANIEL system has a ratio parameter for matching the extracted triggers, we test two values for it. For the first experiments, we use a ratio value of 0.8 (the default value of the system) that was empirically chosen in [11] for the best trade-off between recall and precision. Second, we test the maximum ratio value of 1.0 in order to analyze the system's performance when the extracted disease names and locations exactly match with the knowledge base.

For event identification on clean textual data, one can notice from the Table 2, that usually DANIEL favors recall instead of precision and tends to suffer from an imbalance between precision and recall, which may be due to the high imbalance of the data. It is also not surprising that the DANIEL system the highest performance values for event identification for Chinese and Greek, since for Chinese, there are few relevant documents comparing with the other languages (16 documents that report an event), and for Greek, there are 26 of them.

Table 2. Evaluation of DANIEL on the initial dataset for event identification (regardless of the types of the triggers)

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.6842	0.8	0.7115	0.641	0.592	0.4918	0.6052
	R	0.8667	1.0	0.9024	0.9259	0.9088	0.8571	0.9059
	F1	0.7647	0.8889	0.7957	0.7576	0.7169	0.625	0.7256
ratio=1.0	P	0.0	0.0	0.0	0.0	0.9155	0.0	0.9155
	R	0.0	0.0	0.0	0.0	0.5735	0.0	0.3988
	F1	0.0	0.0	0.0	0.0	0.7052	0.0	0.5556

We also can note the large difference between the two chosen ratios. More exactly, an increase in this value comes in the detriment of the languages that are not only morphologically rich, but also in the case where the exact name of the disease is not located in the text.

Table 3. Evaluation of DANIEL for event identification and classification (triggers are correctly found and match with the groundtruth ones)

		Polish	Chinese	Russian	Greek	French	English	All languages
ratio=0.8	P	0.3421	0.35	0.2692	0.4103	0.5211	0.2951	0.4645
	R	0.4	0.4118	0.3146	0.5079	0.5781	0.4737	0.5363
	F1	0.3688	0.3784	0.2902	0.4539	0.5481	0.3636	0.4978
ratio=1.0	P	0.0	0.0	0.0	0.0	0.7934	0.0	0.7934
	R	0.0	0.0	0.0	0.0	0.3592	0.0	0.2666
	F1	0.0	0.0	0.0	0.0	0.4945	0.0	0.3991

In the case of event identification and classification, we observe from Table 3, that DANIEL is balanced regarding the precision and recall metrics, being able to have higher F1 on the under-represented languages (Chinese, Russian, and Greek). We also notice that, in all the cases, DANIEL does not detect the number of victims. We assume that this is due to the fact that many of the annotated numbers cannot be found in the text, e.g. 10000 cannot be detected since the original text has the 10,000 form, or it is spelled *ten thousands*. Generally, for the detection of locations, we recall that DANIEL is capable to detect locations due to the usage of external resources and article metadata.

For the experiments on noisy data, we will use a ratio value of 0.8, since the maximum value for the ratio creates results prone to suffer from word variations or misspellings of words (which is a direct consequence of the digitization process).

5.3 Experiments with noisy data

The results in Table 4 clearly state that *Character Degradation* is the effect that affects the most the transcription of the documents. However, for character-

Table 4. Document degradation OCR evaluation on the DANIEL dataset

		Clean	CharDeg	Bleed	Blur	Phantom	All
All	CER	2.61	9.55	2.83	8.76	2.65	11.07
	WER	4.23	26.23	5.93	19.05	4.71	27.36
Polish	CER	0.15	5.86	0.19	7.57	0.19	5.51
	WER	0.74	20.66	1.17	13.23	1.17	20.70
Chinese	CER	36.89	41.01	38.24	43.97	36.91	46.97
	WER	—	—	—	—	—	—
Russian	CER	0.93	16.20	1.45	8.13	1.03	10.91
	WER	1.63	28.46	6.61	14.94	2.73	29.72
Greek	CER	3.52	9.04	3.76	13.79	3.54	16.28
	WER	15.86	41.36	17.39	54.02	15.93	54.76
French	CER	1.96	8.37	2.13	7.43	2.0	10.90
	WER	3.33	23.56	4.89	16.31	3.76	26.07
English	CER	0.35	5.75	0.52	4.74	0.44	7.43
	WER	0.66	24.78	2.14	14.72	1.66	20.99

based languages (e.g. Chinese), CER is commonly used instead of WER as the measure for OCR, and, thus, we report only the CER [18].

We note also that, regarding the Chinese documents, the high values for CER, for every type of noise, might be caused by the existence of the enormous number of characters in the alphabet that, by adding such an effect as *Character Degradation* can change drastically the recognition of a character (and in Chinese, one single character can often be a word). Otherwise, while *Character Degradation* noise and *Blur* effect have more impact on the performance of DANIEL than *Phantom Character* type since it did not generate enough distortion to the images. A similar case applies for the *Bleed Through* noise.

Regarding the experiments presented in Tables 5 and 6, we notice, first of all, that the *Character Degradation* effect, *Blur*, and most of all, all the effects mixed together, have indeed an impact or effect over the performance of DANIEL, but with little variability. Meanwhile, *Phantom Degradation* and *Bleed through* had very little to no impact on the quality of detection with DANIEL.

The cause of the decrease in performance of DANIEL is that, in order to detect events, the system looks for repeated substrings at salient zones. In the case of many incorrectly recognised words during the OCR process, there may be no repetition anymore, implying that the event will not be detected. However, since DANIEL only needs two occurrences of its clues (substring of a disease name and substring of a location), it is assumed to be robust to the loss of many repetitions, as long as two repetitions remain in salient zones.

Regarding all the aforementioned results for the DANIEL system, computing the number of affected event words (disease, location, number of cases), we also notice that a very small number of them have been modified by the OCR process, only 1.98% for all the languages together, for all the effects mixed together, close to the 1.63% that were affected by the OCR on clean data. This is due to the imbalance in the DANIEL dataset: only 10.14% of a total of 4,822 documents

Table 5. Evaluation of DANIEL results on the noisy data for event identification (regardless of the types of the triggers). PL=Polish, ZH=Chinese, RU=Russian, EL=Greek, FR=French, EN=English.

		Orig	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.61	0.735 (+0.12)	0.755 (+0.14)	0.735 (+0.12)	0.74 (+0.13)	0.731 (+0.12)	0.758 (+0.14)
	R	0.91	0.859 (-0.05)	0.674 (-0.23)	0.862 (-0.04)	0.857 (-0.05)	0.862 (-0.04)	0.718 (-0.19)
	F1	0.73	0.792 (+0.06)	0.712 (-0.01)	0.793 (+0.06)	0.794 (+0.06)	0.791 (+0.06)	0.737 (+0.00)
PL	P	0.68	0.643 (-0.03)	0.656 (-0.02)	0.658 (-0.02)	0.692 (+0.01)	0.643 (-0.03)	0.645 (-0.03)
	R	0.87	0.9 (+0.03)	0.7 (-0.17)	0.9 (+0.03)	0.9 (+0.03)	0.9 (+0.03)	0.667 (-0.20)
	F1	0.76	0.75 (-0.01)	0.677 (-0.08)	0.761 (+0.00)	0.783 (+0.02)	0.75 (-0.01)	0.656 (-0.10)
ZH	P	0.8	0.882 (+0.08)	0.882 (+0.08)	0.789 (-0.01)	0.733 (-0.06)	0.789 (-0.01)	0.857 (+0.05)
	R	1.0	0.938 (-0.06)	0.938 (-0.06)	0.938 (-0.06)	0.917 (-0.08)	0.938 (-0.06)	0.75 (-0.25)
	F1	0.89	0.909 (+0.01)	0.909 (+0.01)	0.857 (-0.03)	0.815 (-0.07)	0.857 (-0.03)	0.8 (-0.09)
RU	P	0.71	0.688 (-0.02)	0.691 (-0.01)	0.688 (-0.02)	0.705 (-0.00)	0.688 (-0.02)	0.727 (+0.01)
	R	0.9	0.805 (-0.09)	0.744 (-0.15)	0.846 (-0.05)	0.795 (-0.10)	0.846 (-0.05)	0.821 (-0.08)
	F1	0.8	0.742 (-0.05)	0.716 (-0.08)	0.759 (-0.04)	0.747 (-0.05)	0.759 (-0.04)	0.771 (-0.02)
EL	P	0.64	0.59 (-0.05)	0.682 (+0.04)	0.59 (-0.05)	0.639 (-0.00)	0.59 (-0.05)	0.667 (+0.02)
	R	0.93	0.852 (-0.07)	0.556 (-0.37)	0.852 (-0.07)	0.852 (-0.07)	0.852 (-0.07)	0.518 (-0.41)
	F1	0.76	0.697 (-0.06)	0.612 (-0.14)	0.697 (-0.06)	0.73 (-0.03)	0.697 (-0.06)	0.583 (-0.17)
FR	P	0.59	0.803 (+0.21)	0.828 (+0.23)	0.806 (+0.21)	0.801 (+0.21)	0.801 (+0.21)	0.816 (+0.22)
	R	0.91	0.849 (-0.06)	0.666 (-0.24)	0.849 (-0.06)	0.849 (-0.06)	0.849 (-0.06)	0.723 (-0.18)
	F1	0.72	0.826 (+0.10)	0.738 (+0.01)	0.827 (+0.10)	0.825 (+0.10)	0.825 (+0.10)	0.767 (+0.04)
EN	P	0.49	0.508 (+0.01)	0.458 (-0.03)	0.508 (+0.01)	0.516 (+0.02)	0.508 (+0.01)	0.52 (+0.03)
	R	0.86	0.943 (+0.08)	0.629 (-0.23)	0.943 (+0.08)	0.943 (+0.08)	0.943 (+0.08)	0.743 (-0.11)
	F1	0.62	0.66 (+0.04)	0.53 (-0.09)	0.66 (+0.04)	0.667 (+0.04)	0.66 (+0.04)	0.612 (-0.00)

contain events. It brings us to the conclusion that the event extraction task is not considerably impacted by the degradation of the image documents.

One interesting observation is that the precision or the recall can increase, resulting in a higher F1, despite the higher noise effect applied. One possible explanation for this phenomenon is that with a greater level of noise, some false positives disappear. Documents, which were previously classified wrongly due to being too ambiguous to the system (for instance documents relating vaccination campaigns are usually tagged as non-relevant in the ground truth dataset), were given much more distinction thanks to the noise, thus making them look less like relevant samples to the system. More formally: let document D be a false positive in its raw format (D_{raw}). Let D_{Noisy} be its noisy version. If the paragraph that triggered both system's misclassifications disappeared in D_{noisy} , there are good chances that it will be classified as non-relevant. In that case, D_{raw} is a false positive but D_{noisy} is a true negative. That may seem counter-intuitive but noise can improve classification results, see for instance [13] for a study on the same dataset of the influence of boilerplate removal on results.

Table 6. Evaluation of DANIEL results on the noisy data for event identification and classification (triggers are correctly found and match with the groundtruth ones). PL=Polish, ZH=Chinese, RU=Russian, EL=Greek, FR=French, EN=English.

		Original	Clean	CharDeg	Bleed	Blur	Phantom	All
All	P	0.46	0.552 (+0.09)	0.548 (+0.08)	0.549 (+0.08)	0.558 (+0.09)	0.548 (+0.08)	0.547 (+0.08)
	R	0.54	0.497 (-0.04)	0.377 (-0.16)	0.496 (-0.04)	0.497 (-0.04)	0.498 (-0.04)	0.4 (-0.14)
	F1	0.5	0.523 (+0.02)	0.447 (-0.05)	0.521 (+0.02)	0.526 (+0.02)	0.521 (+0.02)	0.462 (-0.03)
PL	P	0.34	0.333 (-0.00)	0.328 (-0.01)	0.342 (+0.00)	0.359 (+0.01)	0.333 (-0.00)	0.274 (-0.06)
	R	0.4	0.431 (+0.03)	0.323 (-0.07)	0.431 (+0.03)	0.431 (+0.03)	0.431 (+0.03)	0.262 (-0.13)
	F1	0.37	0.376 (+0.00)	0.326 (-0.04)	0.381 (+0.01)	0.392 (+0.02)	0.376 (+0.00)	0.268 (-0.10)
ZH	P	0.35	0.412 (+0.06)	0.353 (+0.00)	0.342 (-0.00)	0.367 (+0.01)	0.342 (-0.00)	0.464 (+0.11)
	R	0.41	0.412 (+0.00)	0.353 (-0.05)	0.382 (-0.02)	0.423 (+0.01)	0.382 (-0.02)	0.382 (-0.02)
	F1	0.38	0.412 (+0.03)	0.353 (-0.02)	0.361 (-0.01)	0.393 (+0.01)	0.361 (-0.01)	0.419 (+0.03)
RU	P	0.27	0.302 (+0.03)	0.312 (+0.04)	0.302 (+0.03)	0.295 (+0.02)	0.302 (+0.03)	0.273 (+0.00)
	R	0.31	0.326 (+0.01)	0.357 (+0.04)	0.341 (+0.03)	0.306 (-0.00)	0.341 (+0.03)	0.282 (-0.02)
	F1	0.31	0.314 (+0.00)	0.333 (+0.02)	0.32 (+0.01)	0.301 (-0.00)	0.32 (+0.01)	0.278 (-0.03)
EL	P	0.41	0.333 (-0.07)	0.341 (-0.06)	0.333 (-0.07)	0.361 (-0.04)	0.333 (-0.07)	0.357 (-0.05)
	R	0.51	0.413 (-0.09)	0.238 (-0.27)	0.413 (-0.09)	0.413 (-0.09)	0.413 (-0.09)	0.238 (-0.27)
	F1	0.45	0.369 (-0.08)	0.28 (-0.17)	0.369 (-0.08)	0.385 (-0.06)	0.369 (-0.08)	0.286 (-0.16)
FR	P	0.47	0.691 (+0.22)	0.693 (+0.22)	0.69 (+0.22)	0.689 (+0.21)	0.689 (+0.21)	0.675 (+0.20)
	R	0.51	0.527 (+0.01)	0.402 (-0.10)	0.524 (+0.01)	0.527 (+0.01)	0.527 (+0.01)	0.431 (-0.07)
	F1	0.49	0.598 (+0.10)	0.509 (+0.01)	0.596 (+0.10)	0.597 (+0.10)	0.597 (+0.10)	0.526 (+0.03)
EN	P	0.47	0.292 (-0.17)	0.26 (-0.21)	0.292 (-0.17)	0.297 (-0.17)	0.292 (-0.17)	0.31 (-0.16)
	R	0.51	0.5 (-0.01)	0.329 (-0.18)	0.5 (-0.01)	0.5 (-0.01)	0.5 (-0.01)	0.408 (-0.10)
	F1	0.49	0.369 (-0.12)	0.291 (-0.19)	0.369 (-0.12)	0.372 (-0.11)	0.369 (-0.12)	0.352 (-0.13)

6 Conclusions and Perspectives

We conclude that, in our experimental setting, the epidemical event extraction is prone to digitization errors, but, at the same time, the impact on the DANIEL system is not considerable, which makes it a robust solution for health surveillance applications. Nevertheless, while these experiments were performed in an artificial setting with synthetically produced noise effects, the challenges that exist in a more realistic reasonable scenario could generate other tremendous issues due to the digitization process. As a perspective, we consider the annotation of a digitized dataset in order to assess our assumptions.

References

1. Byrne, K.: Nested named entity recognition in historical archive text. In: International Conference on Semantic Computing (ICSC 2007). pp. 589–596. IEEE (2007)
2. Collier, N.: Towards cross-lingual alerting for bursty epidemic events. *Journal of Biomedical Semantics* **2**(5), S10 (2011)
3. Collier, N., Doan, S., Kawazoe, A., Goodwin, R.M., Conway, M., Tateno, Y., Ngo, Q.H., Dien, D., Kawtrakul, A., Takeuchi, K., et al.: Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics* **24**(24), 2940–2941 (2008)

4. Crane, G., Jones, A.: The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries. pp. 31–40 (2006)
5. Du, M., Von Etter, P., Kopotev, M., Novikov, M., Tarbeeva, N., Yangarber, R.: Building support tools for russian-language information extraction. In: International Conference on Text, Speech and Dialogue. pp. 380–387. Springer (2011)
6. Favre, B., Béchet, F., Nocéra, P.: Robust named entity extraction from large spoken archives. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 491–498. Association for Computational Linguistics (2005)
7. Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., Gipp, B.: Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions (03 2018). https://doi.org/10.1007/978-3-319-78105-1_39
8. Hamdi, A., Jean-Caurant, A., Sidère, N., Coustaty, M., Doucet, A.: Assessing and minimizing the impact of ocr quality on named entity recognition. In: International Conference on Theory and Practice of Digital Libraries. pp. 87–101. Springer (2020)
9. Hatmi, M.: Reconnaissance des entités nommées dans des documents multimodaux. Ph.D. thesis, UNIVERSITÉ DE NANTES (2014)
10. Journet, N., Visani, M., Mansencal, B., Van-Cuong, K., Billy, A.: Doccreator: A new software for creating synthetic ground-truthed document images. *Journal of imaging* **3**(4), 62 (2017)
11. Lejeune, G., Brixtel, R., Doucet, A., Lucas, N.: Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine* **65** (07 2015). <https://doi.org/10.1016/j.artmed.2015.06.005>
12. Lejeune, G., Doucet, A., Yangarber, R., Lucas, N.: Filtering news for epidemic surveillance: towards processing more languages with fewer resources. In: Proceedings of the 4th Workshop on Cross Lingual Information Access. pp. 3–10 (2010)
13. Lejeune, G., Zhu, L.: A new proposal for evaluating web page cleaning tools. *Computacion y Sistemas* **22**(4), 1249–1258 (2018)
14. Lucas, N.: Modélisation différentielle du texte, de la linguistique aux algorithmes. Ph.D. thesis, Université de Caen (2009)
15. Pontes, E.L., Hamdi, A., Sidère, N., Doucet, A.: Impact of ocr quality on named entity linking. In: International Conference on Asian Digital Libraries. pp. 102–115. Springer (2019)
16. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1524–1534. Association for Computational Linguistics (2011)
17. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). vol. 2, pp. 629–633. IEEE (2007)
18. Wang, P., Sun, R., Zhao, H., Yu, K.: A new word language model evaluation metric for character based languages. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 315–324. Springer (2013)



9 Improving NER systems by marking uppercase tokens, and predicting masked tokens and entities boundaries

Improving NER systems by marking uppercase tokens, and predicting masked tokens and entities boundaries

Anonymous submission

Abstract. We explore three different methods for improving Named Entity Recognition (NER) systems based on BERT. Specifically, we explore the marking of uppercase tokens for providing extra casing information. We randomly mask tokens, as in a masked language model, and predict them along with the NER task. And we predict entity boundaries using multi-task learning to ameliorate entity detection. The experiments were done over five languages, three of which are low-resourced: English, Spanish, Slovene, Croatian and Finnish. Results show that predicting masked tokens can be beneficial for most languages, while predicting entity boundaries can improve the state of the art on the Croatian dataset HR500k. Marking uppercase tokens can improve the correct detection of entities in sentences that are composed only of capitalized words.

Keywords: Named Entity Recognition · BERT · multi-task

1 Introduction

Named Entity Recognition (NER) is a fundamental task in the processing of texts that consists on extracting entities that semantically refer to aspects such as locations, persons or organizations [15, 25]. In 2019, Devlin et al. presented the model *Bidirectional Encoder Representations from Transformers (BERT)* [8] and demonstrated that pre-trained models based on BERT can be fine-tuned to achieve high performance in multiple tasks including NER.

Although BERT have proved to be an excellent base for generating new NER systems, we have identified on different datasets three aspects that if alleviated, they could improve the performance of an NER system. In first place, BERT can have issues analyzing sentences that are in capital letters. In second place, BERT can have trouble in determining correctly the boundaries of named entities and in consequence affecting the correct recognition of them. For instance, in the Croatian dataset HR500k [14] the prediction of entities boundaries can be as low as a micro F-score of *0.867*, while in English CoNLL 2003 [20] we can reach a micro F-score of *0.954*. Finally, BERT, in order to predict correctly the type of a named entity, might need a larger context than the available one. Therefore, in this paper, we explore three different methods and their combination to alleviate these issues. Specifically, we explored the substitution of uppercase tokens, the masking and prediction of tokens, and the detection of boundaries, these last two techniques implemented in a multi-task manner. Our experiments focus on

2 Anonymous

five different datasets. We improved the state of the art for Croatian (HR500k [14]). While we have interesting results for English (CoNLL 2003 [20]), Spanish (CoNLL 2002 [19]), Slovene (SSJ500k [11]) and Finnish [15].

2 Related Work

Recent multilingual NER systems have opted for BERT-based architectures. For instance, Luoma et al. [15] presented a new dataset in Finnish based on the Universal Dependency Finnish corpus and evaluated it using different NER systems from the state of the art, including FinBERT [23], a Finnish BERT.

For Croatian and Slovene, the Janes Project [10] proposed Janes-NER, an NER system that uses a Conditional Random Fields (CRF) classifier, along with lexica and Brown clusters; it is based on [13]. It was trained and tested on HR500k [14] and SSJ500k [11] using 5 possible entity types: Location, Person, Person-Derived, Organization and Miscellaneous. Both languages have been evaluated¹ using the Babushka-Bench². The work of [21] presented *CroSloEngual*, a multilingual BERT for Croatian, Slovene and English; it was evaluated on NER using the datasets of HR500k [14] and SSJ500k [11]; only entities of type Location, Person and Organization were predicted. Despite the good results for Slovene, the performance for Croatian was poor. In [3], the authors evaluated two NER systems from the state of the art: Polyglot [2] and the Croatian NERC System [5] over the corpus HR500k [14]. Only the entities of type Location, Person and Organization were considered.

Yu et al. [25] used BERT [8], FastText [6] and character embeddings, with a biaffine model [9] in a new NER system. Their results improved state of the art results in multiple datasets including Spanish CoNLL 2002 [19].

In [12], the authors created *BdryBot* a tool for detecting named entities boundaries. It is based on multiple recursive neural networks, a pointer mechanism and BERT. On English CoNLL 2003 [20], they arrived to an F-score of 0.974. Meaning that the detection of named entities boundaries is easier than the prediction of their types.

3 Methodology

BERT [8] is a deep neural network architecture based on multiple transformers [22] used for creating bidirectional language representations. Pre-trained BERT models can be fine-tuned by adding extra layers to solve specific tasks.

The proposed architecture is shown in Figure 1, which it is based on multi-task learning. We follow the NER architecture proposed by [8], which consists on using after BERT a linear layer. However, to improve the correct annotation of entities, we add as well a CRF such as in [16]. For the entities boundaries, we use the same architecture, but with a reduced number of possible labels. Regarding

¹ <https://github.com/clarinsi/janes-ner>

² <https://github.com/clarinsi/babushka-bench>

NER with marked uppercase, and prediction of tokens and boundaries 3

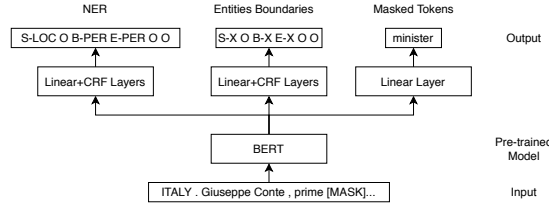


Fig. 1. Proposed architecture, including an example of the expected output.

the prediction of masked tokens, the architecture follows the same propose by [8] for training a masked language model. This consists on introducing the output of a BERT model into a linear layer, which has the same size of the pre-trained vocabulary. The linear layer is expected to predict the masked token. During training, the losses produced by all the tasks, are summed. At prediction time, only the NER part is active.

For the marking of uppercase tokens, we decided to follow an approach similar to [1]. We add two different special tokens to BERT's vocabulary, i.e. *[UP]*, *[up]*, to mark the occurrence of an uppercase token. Inside these special tokens we include the uppercase token, the title-formatted token and the lowercase token. For instance *ITALY* would be represented as *[UP] ITALY Italy italy [up]*. Only the first token is used in the prediction of the entity type and boundary.

4 Experimental Setup

The NER systems explored in this article are based on BERT, using Pytorch, HuggingFace's Transformers [24] and different pre-trained BERT models: for English we make use of *BERT_{BASE}* [8]; for Finnish, *FinBERT* [23]; for Slovene and Croatian, *CroSloEngual* [21] and for Spanish we use *BETO* [7].

For each language, we train 8 different models. The first model, i.e. baseline, is the implementation that only consists on BERT+Linear+CRF. The remaining 7 models, are the different combinations of the approaches described in Section 3 when added to our baseline. Every model is trained up to 20 epochs using an early stop approach and AdamW with bias correction [17]. The early stop is based on the micro F-score and loss of the development dataset. The hyperparameters are: Maximum epochs 20; Learning rate $2e^{-5}$ using a linear schedule with a warm-up ratio of 0.1; Batch size 32 for training, 8 for testing; Adam ϵ $1e^{-8}$; Random seed 12; Dropout rate 0.1; Weight decay 0.01; Clipping gradient norm 1.0; early stop patience of 3 epochs; BERT's token window size 128. For the masking of tokens, we only affect the sentences in the training partitions that are longer than 3 tokens. At each epoch, we select randomly 25% of each sentence's tokens and substitute them with *[MASK]*. We encode the tags for the named entities using BIOES, and we evaluate the systems using *Segeval*³. The assessment of boundaries is done with the *exact* metric provided by Nervaluate⁴.

³ <https://github.com/chakki-works/segeval>

⁴ <https://github.com/ivyleavedtoadflax/nervaluate/>

4 Anonymous

It should be noted that unlike [8, 23, 15], where BERT's input was enriched either with surrounding sentences or document context, our models have for input only the sentence that needs to be analyzed. Moreover, and in contrast with some BERT implementations, the inputs surpassing BERT's token window size are split instead of truncated.⁵ The splitting consists on generating a new input sentence with the rest of the tokens; during prediction, the tokens are aligned to match the original input.

Regarding the datasets, for English, we use CoNLL 2003 [20] and for Spanish CoNLL 2002 [19]. Both corpora have been annotated using 4 types of named entities: Location, Person, Organization and Miscellaneous. For Finnish, we use the corpus proposed by [15]. This corpus has 6 different types of named entities: Location, Date, Person, Event, Organization and Product.

For Croatian and Slovene, we use the corpus HR500k [14] and SSJ500k [11], respectively. According to their respective authors, both corpora have been annotated with 5 types of named entities: Location, Person, Person-derived, Organization and Miscellaneous. However, in the case of HR500k, we did not find entries tagged with the Miscellaneous type, as it happened as well in [3, 21]. Following some previous works [3, 21], we removed the type Person-derived, as it is the less frequent type in both corpora.

5 Results and Discussion

We present in Table 1 the results, in terms of micro and macro F-score, for the different combinations of systems proposed in this work. As well, we present results from the state of the art; in the case of English, we just present some of the most representative. It should be noted in Table 1 that the evaluation of Janes-NER using the Babushka-Bench, does not consider errors in boundaries, and calculates the macro F-score using the performance of 5 named entities types plus the obtained score of predicting the *Other* type.

From Table 1, we can notice that training in parallel for the entities boundaries can be of great help in Croatian and in lesser degree for English, Spanish and Finnish; for Slovene it can be harmful. According to Nerevaluate, the exact prediction of entities boundaries in terms of micro F-score, passed, for each language, from the *baseline* to the *bound.* as follows: Croatian from *0.867* to *0.894*; Slovene from *0.937* to *0.932*; Finnish from *0.928* to *0.925*; Spanish from *0.954* to *0.955*; English from *0.954* to *0.955*. It is interesting to notice that in Finnish, the prediction of boundaries affected the exact metric, with respect to the baseline, but augmented the performance of the NER system as show in Table 1. This means that we predicted incorrectly more boundaries, but we managed to find other entities that we did not find previously.

We can observe, as well in Table 1, that marking the uppercase tokens can improve the performance mainly in English and Croatian; minor improvement in Finnish. Based on an analysis of the results, this is due to the number of

⁵ Some implementations disregard the tokens surpassing the token window or considered these as the type *Other*.

NER with marked uppercase, and prediction of tokens and boundaries 5

Table 1. Values of micro and macro F-score for each experiment, and some results from the state of the art. The best performance is in bold; the second-best is in italics.

	English		Spanish		Croatian		Slovene		Finnish	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
BERT Base [8]	0.924	-	-	-	-	-	-	-	-	-
Cloze-driven [4]	0.935	-	-	-	-	-	-	-	-	-
Seq2seq+BERT [18]	<i>0.929</i>	-	0.888	-	-	-	-	-	-	-
NER Dep.Par. [25]	0.935	-	0.903	-	-	-	-	-	-	-
FinBERT [15]	-	-	-	-	-	-	-	-	0.916	0.810
Fin. BiLSTM-CNN-CRF [15]	-	-	-	-	-	-	-	-	0.815	-
Janes-NER [10]	-	-	-	-	-	0.673	-	0.752	-	-
CroSloEngual [21]	-	-	-	-	-	0.884	-	0.920	-	-
Polyglot [2] [3]	-	-	-	-	-	0.622	-	-	-	-
Croatian NERC [5] [3]	-	-	-	-	-	0.654	-	-	-	-
Baseline	0.912	0.898	0.885	0.870	0.835	0.865	0.902	0.860	0.905	0.811
Bound.	0.916	0.902	0.889	0.868	0.873	0.895	0.891	0.842	0.909	0.820
Upper.	0.916	0.902	0.884	0.864	0.856	<i>0.879</i>	0.896	0.852	0.908	0.809
Bound.-Upper.	0.920	0.906	0.880	0.860	0.849	0.877	0.893	0.854	0.906	0.807
Mask. Baseline	0.920	0.907	0.892	0.874	0.847	0.868	0.923	<i>0.891</i>	0.905	<i>0.834</i>
Mask. Bound.	0.918	0.904	0.892	0.873	0.848	0.867	0.902	0.858	0.909	0.817
Mask. Upper.	0.924	0.910	0.883	0.857	<i>0.859</i>	0.877	<i>0.919</i>	<i>0.891</i>	<i>0.911</i>	0.844
Mask. Bound.-Upper.	0.926	0.912	<i>0.895</i>	0.879	0.852	0.871	0.903	0.861	0.909	0.813

Table 2. Results, in terms of F-score, for each entity type for the Finnish corpus and their average as Macro F-Score (F1).

	PER	LOC	ORG	DATE	EVENT	PROD	Macro F1
FinBERT [15]	0.952	0.947	0.902	0.968	0.435	0.658	0.810
Baseline	0.937	0.949	0.858	0.969	0.470	0.680	0.811
Mask. Upper.	0.950	0.935	0.877	0.969	0.666	0.666	0.844

uppercase tokens found in each dataset. It is possible that the low number of uppercase tokens in the rest of the datasets did not allow BERT to learn correctly the meaning of the special tokens.

In all the cases, it is the masking of tokens the approach that improves the results in general. By masking, marking uppercase tokens and training the boundaries, we can achieve slightly better results than *BERT_{BASE}* [8], although still far from the current state of the art. For Spanish, we can get the second best score in the literature. For Finnish, by masking and marking uppercase tokens, we can get the second best micro F-score, while we can improve the macro F-score. This means that we can ameliorate the prediction of the less-represented entity types in Finnish, see Table 2, although we lose some points in the most frequent ones.

From Table 1, we can notice that masking tokens can be a possible and performing substitute for the addition of context in BERT-based systems. This could be useful in cases where the datasets are not split by documents, we have short texts or the infrastructure is limited either during training or testing.

As the evaluation of NER systems over the Croatian and Slovene datasets is not standard over the state-of-the-art systems, we present in Table 3 the recalculation of the macro F-scores. These are based on the three common types of named entities used in the different NER systems.

With respect to Croatian, we can observe in Table 1 and Table 3 that we can improve the results with respect to CroSloEngual, which is based as well on

6 Anonymous

Table 3. F-score values of the three common named entities for the Croatian and Slovene systems. Best score in bold; second-best in italics. Results of Janes-NER comes from the Babushka-Bench which disregards boundaries.

	Croatian				Slovene			
	PER	LOC	ORG	Macro F1	PER	LOC	ORG	Macro F1
CroSloEngual [21]	NA	NA	NA	0.884	NA	NA	NA	0.920
Janes-NER [10]	0.890	0.850	0.720	0.820	0.890	0.800	0.670	0.786
Polyglot [2] [3]	NA	NA	NA	0.622	-	-	-	-
Croatian NERC [5] [3]	NA	NA	NA	0.640	-	-	-	-
Baseline	0.849	0.954	0.791	0.865	0.963	0.912	0.817	0.897
Bound.	0.881	0.961	0.842	0.895	0.948	0.925	0.800	0.891
Masked Baseline	0.818	0.956	0.831	<i>0.868</i>	0.973	0.933	0.831	<i>0.912</i>

BERT. For Slovene, we were not been able to surpass the performance showed in [21]; however, we have trained a model that includes as well the Miscellaneous entity type. It should be indicated that Janes-NER gets and F-score for Miscellaneous of *0.270* while our masked baseline gets *0.828*.

6 Conclusions and future work

Named Entity Recognition (NER) is a task that aims to extract and classify groups of tokens referring to specific types like locations, persons and organizations. In the last couple of years, with the creation of BERT [8], multiple NER systems made use of its architecture to provide high-performing tools. Nonetheless, we observed that this kind of systems could face some issues, like the bad prediction of uppercase sentences, the wrong detection of entities boundaries and the need of more information to correctly predict entities in short sentences.

Therefore, in this work, we presented three different methods that could alleviate these issues. Experiments were done over five languages, three of them low-resourced ones. We improved the state of the art with a micro F-score of *0.873* in Croatian by predicting entities boundaries along NER. By masking tokens, predicting boundaries and tokens we managed to improve the performance of *BERT_{BASE}* to F-score of *0.926* in English, while getting the second-best performance in Spanish with an F-score *0.895*. In Finnish, we arrived to improve the prediction of the less frequent named entity types, with a macro F-score of *0.844* versus *0.810* in the state of the art, while keeping a comparable micro F-score. And we produced a NER for Slovene that can predict 4 types of named entities, one of which is not frequent, and get comparable results to another tool from the state of the art that only predicts the three most frequent types.

In the future, we will experiment with additional languages. We will also try to capitalize random sentences for improving the marking of uppercase tokens. Finally, we would like to asses whether the addition of some context to the left of the split sentences could improve the performance of the NER.

NER with marked uppercase, and prediction of tokens and boundaries 7

References

1. Matching the Blanks: Distributional Similarity for Relation Learning. Florence, Italy
2. Al-Rfou, R., Kulkarni, V., Perozzi, B., Skiena, S.: POLYGLOT-NER: Massive Multilingual Named Entity Recognition. CoRR **abs/1410.3791** (2014), <http://arxiv.org/abs/1410.3791>, eprint: 1410.3791
3. Alves, D., Thakkar, G., Tadić, M.: Evaluating Language Tools for Fifteen EU-official Under-resourced Languages. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1866–1873. European Language Resources Association, Marseille, France (May 2020)
4. Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M.: Cloze-driven Pre-training of Self-attention Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5360–5369. Association for Computational Linguistics, Hong Kong, China (Nov 2019). <https://doi.org/10.18653/v1/D19-1539>
5. Bekavac, B., Tadić, M.: Implementation of Croatian NERC System. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. pp. 11–18. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), <https://www.aclweb.org/anthology/W07-1702>
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association of Computational Linguistics **5**, 135–146 (2017)
7. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: PML4DC at ICLR 2020 (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). <https://doi.org/10.18653/v1/N19-1423>
9. Dozat, T., Manning, C.D.: Deep Biaffine Attention for Neural Dependency Parsing. CoRR **abs/1611.01734** (2016), <http://arxiv.org/abs/1611.01734>, eprint: 1611.01734
10. Fišer, D., Ljubešić, N., Erjavec, T.: The Janes project: language resources and tools for Slovene user generated content. Language Resources and Evaluation **54**(1), 223–246 (Mar 2020). <https://doi.org/10.1007/s10579-018-9425-z>
11. Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Marko, D., Jezeršek, L., Zajc, A.: Training corpus ssj500k 2.2 (2019), <http://hdl.handle.net/11356/1210>, slovenian language resource repository CLARIN.SI
12. Li, J., Sun, A., Ma, Y.: Neural Named Entity Boundary Detection. IEEE Transactions on Knowledge and Data Engineering pp. 1–1 (2020)
13. Ljubešić, N., Erjavec, T.: Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1527–1531. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), <https://www.aclweb.org/anthology/L16-1242>

8 Anonymous

14. Ljubešić, N., Klubička, F., Agić, , Jazbec, I.P.: New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4264–4270. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), <https://www.aclweb.org/anthology/L16-1676>
15. Luoma, J., Oinonen, M., Pyrkönen, M., Laippala, V., Pyysalo, S.: A Broad-coverage Corpus for Finnish Named Entity Recognition. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4615–4624. European Language Resources Association, Marseille, France (May 2020), <https://www.aclweb.org/anthology/2020.lrec-1.567>
16. Ma, X., Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (Aug 2016). <https://doi.org/10.18653/v1/P16-1101>
17. Mosbach, M., Andriushchenko, M., Klakow, D.: On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines (2020), [_eprint: 2006.04884](https://arxiv.org/abs/2006.04884)
18. Straková, J., Straka, M., Hajic, J.: Neural Architectures for Nested NER through Linearization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5326–5331. Association for Computational Linguistics, Florence, Italy (Jul 2019). <https://doi.org/10.18653/v1/P19-1527>
19. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). <https://www.aclweb.org/anthology/W02-2024>
20. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. pp. 142–147 (2003), <https://www.aclweb.org/anthology/W03-0419>
21. Ulčar, M., Robnik-Šikonja, M.: FinEst BERT and CroSloEngual BERT. In: Sojka, P., Kopeček, I., Pala, K., Horák, A. (eds.) Text, Speech, and Dialogue. pp. 104–111. Springer International Publishing, Cham (2020)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
23. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for Finnish (2019), [_eprint: 1912.07076](https://arxiv.org/abs/1912.07076)
24. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P.v., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv **abs/1910.03771** (2019)
25. Yu, J., Bohnet, B., Poesio, M.: Named Entity Recognition as Dependency Parsing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6470–6476. Association for Computational Linguistics, Online (Jul 2020). <https://doi.org/10.18653/v1/2020.acl-main.577>

10 Linking Named Entities across Languages using Multilingual Word Embeddings

Linking Named Entities across Languages using Multilingual Word Embeddings

Elvys Linhares Pontes
elvys.linhares_pontes@univ-lr.fr
University of La Rochelle
La Rochelle, France

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle
La Rochelle, France

José G. Moreno
jose.moreno@irit.fr
University of Toulouse
Toulouse, France

ABSTRACT

Digital libraries are online collections of digital objects that can include text, images, audio, or videos in several languages. It has long been observed that named entities (NEs) are key to the access to digital library portals as they are contained in most user queries. However, NEs can have different spellings for each language which reduces the performance of user queries to retrieve documents across languages. Cross-lingual named entity linking (XEL) connects NEs from documents in a source language to external knowledge bases in another (target) language. The XEL task is especially challenging due to the diversity of NEs across languages and contexts. This paper describes a XEL system applied and evaluated with several languages pairs including English and various low-resourced languages of different linguistic families such as Croatian, Finnish, Estonian and Slovenian. We tested this approach to analyze documents and NEs in low-resourced languages and link them to the English version of Wikipedia. We present the resulting study of this analysis and the challenges involved in the case of degraded documents from digital libraries. Further works will make an extensive analysis of the impact of our approach on the XEL task with OCRred documents.

KEYWORDS

Cross-Lingual Named Entity Linking, Multilingual Word Embeddings, Digital Library, Indexing

ACM Reference Format:

Elvys Linhares Pontes, Antoine Doucet, and José G. Moreno. 2020. Linking Named Entities across Languages using Multilingual Word Embeddings. In *Proceedings of ACM/IEEE Joint Conference on Digital Libraries*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Digital libraries are composed of a large number of digital contents (e.g. journals, books, magazines, videos, and so on) in several languages about diverse subjects (e.g. history, languages, politics, sciences, philosophy, and so on). Named entities have been demonstrated to be essential to digital library access as they are included in a majority of the search queries submitted to digital library portals [2]. However, the spelling of an entity is language-dependent which impacts the performance of search engines when trying to retrieve all relevant documents with respect to a query. For instance, the entity “United States” has different spelling in other languages: “Estados Unidos” (in Portuguese and Spanish) and “États-Unis” (in French).

Moreover, data from different sources can contain ambiguous, complementary, and duplicate information about named entities. Therefore, they are often not distinctive since one single name may correspond to multiple entities. A disambiguation process is thus essential to distinguish the correct named entities to be indexed in digital libraries. In this case, a monolingual disambiguation analysis cannot disambiguate these entities in several languages for a common knowledge base.

Named Entity Linking (NEL) aims to recognize mentions in a document and link them to their corresponding entries in a Knowledge Base (KB), such as Wikipedia¹, DBpedia², and Freebase³. Additionally, Cross-Lingual Named Entity Linking (XEL) considers documents that are written in a source language that is different from the target language of the KB [18]. In addition to the challenges of NEL such as multiple surface forms of a named entity [16], XEL disambiguates mentions in several languages by analyzing different spellings and contexts related to each language.

Digital libraries often contain the digitised version of old documents that are degraded due to storage conditions, handling of users and inherent vice of the material (e.g. paper naturally deteriorates over time). These problems cause numerous errors at the character and word levels in the OCR of these documents [10]. Linhares Pontes et al. [10] analyzed the impact of OCR quality on the NEL task and achieved satisfying results for NEL. They provided recommendations on the OCR quality that is required for a given level of expected NEL performance. However, their approach is monolingual, restricting the analysis and linking of entities to knowledge bases that are in the same language (in this case, English).

The XEL task is especially challenging due to the diversity of NEs across languages and contexts. This paper describes a XEL system applied and evaluated with several languages pairs including English and various low-resourced languages of different linguistic families such as Croatian, Finnish, Estonian and Slovenian. We tested this approach to analyze documents and NEs in low-resourced languages and link them to the English version of Wikipedia. We present the resulting study of this analysis and the challenges involved in the case of degraded documents from digital libraries.

The remainder of the paper is organized as follows : Section 2 makes a brief overview of the most recent and available NEL and XEL approaches in the state of the art. Section 3 details our approach to extend a monolingual NEL system for the XEL task by using multilingual word embeddings. Then, the experimental setup and

ACM/IEEE Joint Conference on Digital Libraries, 2020, China
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00
<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

¹<http://www.wikipedia.org/>

²<https://wiki.dbpedia.org/>

³<http://www.freebase.com/>

evaluation are respectively described in Sections 4 and 5. Finally, conclusions and future works are set out in Section 6.

2 AN OVERVIEW OF (CROSS-LINGUAL) NAMED ENTITY LINKING

Given a set of documents $D = \{d_1, d_2, \dots, d_l\}$, a set of detected mentions $M^j = \{m_1^j, m_2^j, \dots, m_n^j\}$ in the document d_j for $\forall j \in [1, l]$, and a knowledge base $KB = \{e_1, e_2, \dots, e_s\}$, Named Entity Linking (NEL) aims to map each mention m_i^j with its corresponding entity e_k in the KB [16]. NEL approaches can be divided into two classes: disambiguation (they use M^j as an input) and end-to-end approaches (they do not use M^j as an input, but calculate it). While end-to-end approaches extract candidate entities from documents and then disambiguate them to the correct entries in a given KB [8], disambiguation approaches only disambiguate entities already recognized from documents [5, 9, 14].

Among the only disambiguation approaches, Ganea and Hofmann [5] built a deep learning model for joint document-level entity disambiguation. They embed entities and words in a common vector space and use a neural attention mechanism to select words that are informative for the disambiguation decision. Then, their model collectively disambiguates the mentions in a document (more details in Section 3.1). Motivated by Ganea and Hofmann's approach, Le and Titov [9] analyzed relations between mentions as latent variables in their neural NEL model. They rely on representation learning and learn embeddings of mentions, contexts, and relations to reduce the amount of human expertise required to construct the system and make the analysis more portable across languages and domains.

In the class of end-to-end approaches, Raiman and Raiman [14] developed a system for integrating symbolic knowledge into the reasoning process of a neural network through a type system. They constrained the behavior to respect the desired symbolic structure, and automatically design the type system without human effort. Their model first uses heuristic search or stochastic optimization over discrete variables that define a type system informed by an Oracle and a learnability heuristic. Based on a joint analysis of the named entity recognition and linking tasks, Kolitsas et al. [8] proposed an end-to-end NEL system that jointly discovers and links entities in a document. They generate all possible spans (mentions) that have at least one possible entity candidate. Then, each mention-candidate pair receives a context-aware compatibility score based on word and entity embeddings [5] coupled with neural attention and a global voting mechanism.

Extending this monolingual analysis, Cross-Lingual Named Entity Linking (XEL) analyzes documents and named entities that are in a different language than that used for the content of the knowledge base. In this context, McNamee et al. [11] proposed an XEL approach and examined the importance of transliteration, the utility of cross-language information retrieval, and the potential benefit of multilingual named entity recognition on the XEL task.

Zhou et al. [18] extensively evaluated the effect of resource restrictions on existing XEL methods in low-resource settings. They investigated a hybrid candidate generation method, combining existing lookup-based and neural candidate generation methods and proposed a set of entity disambiguation features that are entirely

language-agnostic. Finally, they designed a non-linear feature combination method, which makes it possible to combine features in a more flexible way.

3 OUR CONTRIBUTION

This section describes our contribution to adapt Ganea and Hofmann's approach for the XEL task. We make a short description of Ganea and Hofmann's approach (Section 3.1), and then we detail how we extended this approach for the XEL task by using multilingual word embeddings (Section 3.2).

3.1 Ganea and Hofmann's approach

Entity Disambiguation (ED) approaches consider having already identified the named entities in the documents. In this case, these approaches aim to analyse the context of these entities to disambiguate them in a KB. In this context, Ganea and Hofmann [5] (GH) proposed a deep learning model for joint document-level entity disambiguation⁴.

They project entities and words in a common vector space, which avoids hand-engineered features, multiple disambiguation steps, or the need for additional ad-hoc heuristics when solving the ED task. Entities for each mention are locally scored based on cosine similarity with the respective document embedding. Combined with these embeddings, they proposed an attention mechanism over local context windows to select words that are informative for the disambiguation decision. The final local scores are based on the combination of the resulting context-based entity scores and a mention-entity prior. This mention-entity prior ($p(e|m)$) is a conditional distribution of the co-occurrence of the mention m with the entity e . In this case, GH collected mention-entity co-occurrence counts from Wikipedia to calculate this distribution.

Finally, mentions in a document are resolved jointly by using a conditional random field in conjunction with an inference scheme.

3.2 Cross-lingual extension

Most data sets for NEL are available only in English. Among them, the AIDA data set is the main data used to train NEL system on the state of the art. Unfortunately, there are few data sets for low-resourced languages, with the notable exception of the WikiANN corpora.

In order to extend GH's system to a cross-lingual setting, we made a number of modifications to their approach. Instead of using the Word2Vec embeddings, we used the pre-trained multilingual MUSE embeddings⁵ [3]. These embeddings are available in 30 languages (including Croatian, Estonian, Finnish, Slovenian, to mention a few) and they are aligned in a single vector space. Therefore, words like "house" and "talo" ("house" in Finnish) have similar word representations. One of the main goals of using these embeddings is to generate multilingual entity embeddings that can provide entity representations for mentions in several languages. Then, GH's approach will be able to analyse documents in the languages of these embeddings and link them to an English KB. Therefore, we generate the entity embeddings using the English version of Wikipedia and train this system on the AIDA data set using the MUSE embeddings.

⁴The code is publicly available: <https://github.com/dalab/deep-ed>

⁵The MUSE embeddings are available at: <https://github.com/facebookresearch/MUSE>

In this scenario, GH's approach analyses English documents and links their mentions to an English KB.

Moreover, we extend the training process for some low-resourced languages by using the previous English model and continue the training process with data on other languages. This tuning procedure optimises our model to analyse better the documents on low-resourced languages and link their mention to an English KB. More precisely, we initialized the weights of the neural network model with the weights of the English model, and we reduced the learning rate to tune our model for the target languages. This process enables our model to adapt the analysis of words and their context for each language (e.g. the order of words and how they are combined to express a same idea in different languages).

4 EXPERIMENTAL SETUP

In order to analyse the impact of using multilingual embeddings on the representation of entity embeddings, we used the entity relatedness data set of Ceccarelli et al. [1] to compare the quality of entity embeddings produced by the WORD2VEC and multilingual embeddings. This data set contains 3319 and 3673 queries for the test and validation sets. Each query consists of one target entity and up to 100 candidate entities with gold standard binary labels indicating if the two entities are related or not. The associated task requires ranking of related candidate entities higher than unrelated ones. Following GH's work, we used the normalised discounted cumulative gain (NDCG) and mean average precision (MAP) measures to evaluate them. We also performed candidate ranking based on cosine similarity of entity pairs.

We then trained and tested GH's approach with the following benchmarks: AIDA-CoNLL [7], AQUAINT [6], ACE2004 [6, 15], WikiANN [13], CWEB [4] and WIKI [15].

The WikiANN data set was split into 2 separate data sets, 70% of the corpus for training and 30% for testing. For the training process, we use AIDA data set to train the NEL system for English using the MUSE embeddings. Then, we use the WikiANN training data set to optimise the English model for each low-resourced language. Finally, we tested our model on the WikiANN test data sets.

Following previous works, we evaluate the performance of our approach by analyzing the precision, recall and F1-measure. Precision is the fraction of correctly linked entity mentions that are generated by a system. Recall considers all entity mentions that should be linked and determines how correct linked entity mentions are concerning the total entity mentions that should be linked. Finally, the F1-measure is defined as the harmonic mean of precision and recall.

Since knowledge bases contain millions of entities, only mentions that contain a valid ground-truth entry in the KB are analysed. For mentions without corresponding entries in the KB, NEL systems have to provide a NIL entry to indicate that these mentions do not have a ground-truth entity in the KB.

5 EXPERIMENTAL ASSESSMENT

Entity embeddings performance: Table 1 shows the entity relatedness results using WORD2VEC and MUSE embeddings for the English data set [1]. Both embeddings have the same dimensional space (300 dimensions) but different vocabulary sizes: WORD2VEC

(3 million tokens) and MUSE (200 thousand tokens). This large difference helps WORD2VEC to achieve the best results for all entity related measures. More precisely, the WORD2VEC embeddings provide a better analysis of the Wikipedia documents because it has less out-of-vocabulary words than the MUSE embeddings and can represent better the meaning of sentences and entities. Despite this performance drop, GH's approach using MUSE embeddings achieved better results than [17] and [12] for all metrics.

Table 1: Entity relatedness quality for English.

Embeddings	NDCG1	NDCG5	NDCG10	MAP
Ganea and Hofmann (WORD2VEC)	0.632	0.609	0.641	0.578
Ganea and Hofmann (MUSE)	0.613	0.568	0.592	0.536
Yamada et al. [17]	0.59	0.56	0.59	0.52
Milne and Witten [12]	0.54	0.52	0.55	0.48

NEL analysis for mono- and multilingual embeddings: Advancing our analysis of GH's system, we compared the F1-measure results for this system on English corpora using the WORD2VEC and MUSE embeddings (Table 2). As expected, the small vocabulary and lower performance in the entity relatedness measures reduced the performance of GH's system in the NEL task. These factors reduced the quality of the attention and the context embeddings, and prioritised the relevance of entity priors ($\log p(e|m)$) to disambiguate the mentions in a document. Despite this drop, GH's system using MUSE achieved identical or very close performance for most data sets.

Table 2: F1-measure results for Ganea and Hofmann's approach on English corpora.

Embed.	AIDA	ACE2004	AQUAINT	CLUEWEB	WIKI
WORD2VEC	92.2	88.5	88.5	77.9	77.5
MUSE	86.6	88.5	87.5	74.9	74.2

NEL analysis: Table 3 presents the F1-measure results for the NEL on four languages of the WikiANN corpora. We tested the NEL system using only the AIDA training data set to train GH's model in order to link mentions to the English version of the Wikipedia; and using the AIDA training data set in a first step and, then, the WikiANN training data set for each language (second line of Table 3). The tuning process on the WikiANN data set improved the performance of GH's for the WikiANN test data sets. Unfortunately, the WikiANN data set is composed of short sentences with little contextual information. This characteristic makes the context analysis of GH's system less relevant and implies that the disambiguation process mainly consists in pairwise matching between mentions and entities using $\log p(m|e)$. Another limiting factor is the small MUSE vocabulary. Finally, the English version of Wikipedia does not have all entities listed on the Croatian, Estonian, Finnish, and Slovenian Wikipedia versions, which reduces the number of entities that can be linked to the KB.

Table 3: F1-measure results for Ganea and Hofmann’s models on the test WikiANN corpora (Croatian, Estonian, Finnish, and Slovenian languages only).

Models	hr	et	fi	sl
AIDA data set (using MUSE)	60.97	57.82	62.51	69.78
pre-trained model on AIDA data set + tuning on WikiANN data set (using MUSE)	61.53	58.47	63.04	70.31

XEL is a fundamental tool for search engines in digital libraries to retrieve documents where their contents (including named entities) are in different languages and contexts. Linhares Pontes et al. [10] showed an analysis of the impact of problems detected in these libraries using Ganea and Hofmann’s and Le and Titov’s systems. In this analysis, these systems had a small reduction in NEL performance despite the errors caused by the deterioration and conservation problems in libraries. In this work, we showed that Ganea and Hofmann’s system using multilingual embeddings achieved satisfactory results for the English NEL task (maximal F1-measure drop of 5.6%). Additionally, the tuning procedure improved the results for XEL in the low-resourced languages. We assume our approach will perform similarly for the XEL task in OCRed documents, but additional experiments are needed to validate this assumption.

6 CONCLUSION

This paper is the first step to analyze the impact of multilingual embeddings to extend monolingual NEL to XEL. The next step is to investigate the impact of degraded documents on this cross-lingual task.

Despite the small multilingual vocabulary on the word embeddings and the poor context quality of training data sets for low-resourced languages, our experiments showed a worst drop of 5.6% on F1-measure on the English test data set (and the same performance of monolingual embeddings in the best case) and a small improvement with the tuning procedure on low-resourced languages. Therefore, we intend to build training data sets on the target languages that are composed of long sentences with rich context information to improve our XEL model.

Further work is under progress to develop and analyze the performance of end-to-end XEL systems on OCRed data sets. More precisely, we want to extend the analysis of multilingual embeddings with language-agnostic features and relations between entities to provide correct predictions in different languages and overcome the problem of OCR degradation. We also intend to analyze and test the performance of these systems using real data in other languages (e.g. Spanish and Chinese) including other low-resourced languages.

ACKNOWLEDGMENTS

This work has been partly supported by the European Union’s Horizon 2020 research and innovation programme under grants 770299 (NewsEye) and 825153 (EMBEDIA).

REFERENCES

- [1] Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2013. Learning Relatedness Measures for Entity Linking. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM ’13)*. ACM, New York, NY, USA, 139–148.
- [2] Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR errors on the use of digital libraries: towards a better access to information. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press, 249–252.
- [3] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *International Conference on Learning Representations (ICLR ’18)*.
- [4] Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0).
- [5] Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2619–2629.
- [6] Zhaochen Guo and Denilson Barbosa. 2014. Robust Entity Linking via Random Walks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM ’14)*. ACM, New York, NY, USA, 499–508.
- [7] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ’11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 782–792.
- [8] Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-End Neural Entity Linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 519–529.
- [9] Phong Le and Ivan Titov. 2018. Improving Entity Linking by Modeling Latent Relations between Mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1595–1604.
- [10] Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidere, and Antoine Doucet. 2019. Impact of OCR Quality on Named Entity Linking. In *Digital Libraries at the Crossroads of Digital Information for the Future*, Adam Jatowt, Akira Maeda, and Sue Yeon Syn (Eds.). Springer International Publishing, Cham, 102–115.
- [11] Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-Language Entity Linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 255–263.
- [12] David Milne and Ian H. Witten. 2008. Learning to Link with Wikipedia. In *CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge mining*. ACM, New York, NY, USA, 509–518.
- [13] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1946–1958.
- [14] Jonathan Raiman and Olivier Raiman. 2018. DeepType: Multilingual Entity Linking by Neural Type System Evolution. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, New Orleans, Louisiana, USA, February 2-7, 2018. 5406–5413.
- [15] Lev Ratnikov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT ’11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1375–1384.
- [16] W. Shen, J. Wang, and J. Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (Feb 2015), 443–460.
- [17] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Berlin, Germany, 250–259.
- [18] Shuyan Zhou, Shruti Rijhwani, and Graham Neubig. 2019. Towards Zero-resource Cross-lingual Entity Linking. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Association for Computational Linguistics, Hong Kong, China, 243–252.

11 Multilingual Epidemic Event Extraction

Multilingual Epidemic Event Extraction

Anonymous EACL submission

Abstract

In this paper, we focus on epidemic event extraction in a multilingual setting and evaluate several methods for this task. The task is defined as the detection and extraction of a disease name and a location from a document. We experiment on an annotated multilingual news dataset comprising 4,815 relevant and irrelevant documents in six different languages. The relevant documents are annotated with a disease-location pair. The dataset is automatically transformed from a document-level to a sentence-level annotation style, a common scheme utilized in recent event extraction systems. The task of extracting epidemic events involves first detecting the relevant documents from a large collection of news reports. Then, the event extraction (disease-location pairs) is performed on the selected relevant documents. We demonstrate that the multi-stage extraction process and multilingualism can introduce new challenges to the epidemic event extraction task. We show the influence of quality of text classification on subsequent event extraction. A baseline score is established and the dataset is made available for further research.

1 Introduction

The ability to detect disease outbreaks early enough is critical in the deployment of measures to limit their spread. While disease surveillance has in the past been a critical component in epidemiology, conventional surveillance methods are limited in terms of both promptness and coverage, while at the same time requiring labor-intensive human input. The large amounts of continuously generated unstructured data, for instance in the ongoing COVID-19 epidemic, are often challenging and difficult to process by humans without leveraging computational techniques. With the advancements in natural language processing (NLP) techniques, processing such data and applying data-

driven methods for epidemic surveillance has become feasible.

Online news data contains critical information about emerging health threats such as what happened, where it happened, when, and to whom it happened (Ng et al., 2020). When processed into a structured and more meaningful form, the information can foster early detection of disease outbreaks, a critical aspect of epidemic surveillance. News reports on epidemics often originate from different parts of the world and events are likely to be reported in other languages than English. Hence, efficient multilingual approaches are necessary for effective epidemic surveillance.

Several works have tackled the detection of events related to epidemic diseases. Some approaches include external resources and features at a sub-word representation level. For example, the Data Analysis for Information Extraction in any Language (DANIEL) system was proposed as a multilingual news surveillance system that leverages repetition and saliency (salient zones in the structure of a news article), properties that are common in news writing (Lejeune et al., 2015). By avoiding the usage of language-specific NLP toolkits (e.g., Part-of-speech taggers, dependency parsers) and by focusing on the general structure of journalistic writing style (Hamborg et al., 2018), the system is able to detect key event information from news articles in multilingual corpora. We consider it a baseline multilingual model.

Recent NLP models make use of neural-based architectures. Lamos et al. (2017) take advantage of the word embeddings representations used widely in various NLP tasks. Word embeddings capture semantic properties of words, and thus the authors use them to compute the distances between relevant concepts for completing the task of flu event detection from texts. Another type of approach is based on long short term memory (LSTM) (Wang

Split	Sentences	Tokens	French	English	Polish	Chinese	Greek	Russian
Training	6,575	197,825	155,816	13,139	12,712	4,831	4,484	6,843
Validation	1,000	31,184	23,283	2,336	1,861	175	2,214	1,315
Test	782	23,930	18,183	1,472	119	366	1,836	1,954

Table 1: Number of relevant tokens and sentences per dataset split per language.

et al., 2017) models that approach the epidemic detection task from the perspective of classification of tweets to extract influenza-related information.

In this study, we formulate the problem of extracting the disease names and locations in the text as a sequence labeling task. We use a multilingual dataset comprising news articles from the medical domain with diverse morphological structures (Chinese, English, French, Greek, Polish, and Russian). In this dataset, an epidemic event is characterized by the disease name and the reported location relevant to the disease. We evaluate the previously described specialized baseline system and experiment with the most recent NER neural architectures. Error propagation from the classification task that affects the event extraction task is also evaluated since the event extraction task is a multi-step task, comprising various sub-tasks (Joshi et al., 2019; Doan et al., 2008).

The remainder of this paper is organized as follows. Section 2 presents the multilingual dataset used for our study. Section 3 discusses our experimental methodology and empirical results. Finally, Section 4 presents the discussion of results, conclusions, and suggestions for future research.

2 Dataset

Due to the lack of dedicated datasets for epidemic event extraction from multilingual news articles, we adapt a freely available epidemiological dataset¹, called DANIEL, that was proposed by (Lejeune et al., 2015). This is a multilingual dataset consisting of news articles in six different languages, namely French, Polish, English, Chinese, Greek, and Russian. An epidemiological event is represented by the disease name and the location of the reported event.

However, the DANIEL dataset is annotated at the document-level, which differentiates it from typical datasets used in research for the event extraction task. A document is either reporting an event at interest (i.e., disease-place pair, and sometimes the number of victims) or not. We pre-

process and transform the dataset from document-level annotation to sentence-level. The annotations provided by DANIEL, at the document-level, are looked-up in the appropriate file and the found offsets are attached to them. We consider as an example an article that has the following annotations at the document level: **malaria** and **worldwide**. The text of the article contains the following mentions: *Malaria, worldwide*. In this case, in the first sentence, “*GENEVA: Malaria caused the death of an estimated 655,000 people [...]*”, we are able to annotate **Malaria** at offsets 8 – 14. The process is automatic and continues in the same manner for the other annotations.

First, we consider the lemma of an annotated disease name that will further be looked-up in the text. If any disease name or location is found multiple times in the text, we annotate all the present instances. Sometimes, the exact surface form of a disease name cannot be found in the text, as it is the case for Russian, Greek, and Polish articles (morphologically rich languages), we considered the annotation of the grammatical cases of nouns. For example, in Russian, “Простуда” (“prostuda”) means “cold”, and since this disease name cannot be found in the text article, we used the instrumental case in Russian that can generally be distinguished by the “-ом” (“-om”) suffix for most masculine and neuter nouns, the “-ою/“-ой” (“-oju/“-oj”) suffix for most feminine nouns. The instrumental case for singular “простудой” was annotated in the article text.

In the case of locations, there were 57% of cases where the location could not be found in the text, mainly due to the coarse-grained type of manual annotation at the country-level. For the annotation of the locations at a finer-grained level, we considered the presence of cities or regions in the text. For example, if the document was previously annotated with “France”, and “Corsica” is mentioned in the text, we changed the final annotation to “Corsica”.

Finally, we tokenize the articles at the sentence-level and format them in the IOB (Inside, Outside, Beginning) tagging scheme where each token is

¹The dataset is available at <http://ANONYMOUS>.

given one of the following labels: *DISEASE*, *LOCATION* or *O*. We split the data into training (3,852 documents), test (481 documents), and validation (482 documents) sets, stratified by language. Table 1 presents some statistics for this dataset.

3 Experiments

Often, approaches for text-based disease surveillance follow a two-step process (Joshi et al., 2019): document classification and event extraction. First, we perform the document classification into relevant and irrelevant documents, e.g. documents that contain mentions of disease names and locations, and documents that do not. For this, we chose a BERT-based model whose performance on text classification is an F1 of 86.54%. We do not focus on the classification task but rather on the event extraction task: detection and extraction of the disease names and locations. For this step, we compare different state-of-the-art models, first, we experiment with deep learning models, BiLSTM models (Lample et al., 2016; Ma and Hovy, 2016), and further with two architectures based on pre-trained language models.

The following types of experiments are carried out: (1) using all data instances (relevant and irrelevant documents), (2) using only the relevant documents in a perfect setting when these have been detected with a 100% accuracy, and (3) testing on the predicted relevant documents provided by the document classification step.

Event extraction evaluation is performed at coarse-grain, with the entity as the reference unit (Makhoul et al., 1999). We compute precision (P), recall (R), and F1-measure (F1) at the micro-level (error types are considered over all documents).

3.1 Compared Models for Event Extraction

We chose DANIEL (Lejeune et al., 2015) as a baseline model for epidemical event extraction. It is a complete pipeline that first detects the relevant documents and then extracts the event triggers.

We also evaluate two deep neural architectures based on bidirectional LSTM models proposed by Lample et al. (2016) and Ma and Hovy (2016) that use character and word representations².

Additionally, we evaluate the pre-trained model BERT proposed by Devlin et al. (2019) for token

²The hyperparameters for both models are detailed in the papers (Lample et al., 2016) and (Ma and Hovy, 2016).

All data instances (relevant and irrelevant)			
Models	P	R	F1
DANIEL	38.97	47.32	42.74
BiLSTM+LSTM	78.17	69.74	73.71
BiLSTM+CNN	75.43	68.87	72
MBERT-CASED	63.24	53.72	76.88
MBERT-UNCASED	67.13	72.72	62.33
XLM-ROBERTA-BASE	72.23	89.20	79.82

Table 2: Evaluation results for the detection of disease names and locations on all languages and all data instances (relevant and irrelevant documents).

classification³. Due to the multilingual characteristic of the dataset, we use the *bert-base-multilingual-cased* pre-trained and then fine-tuned BERT model. We will refer to these models as MBERT-CASED and MBERT-UNCASED.

We also experiment with the XLM-ROBERTA-BASE model (hereafter XLM-ROBERTA) proposed by Conneau et al. (2020) that has shown significant performance gains for a wide range of cross-lingual transfer tasks. We consider this model appropriate for our task and dataset due to the multilingual characteristic of the data⁴.

3.2 Results

We present results of the evaluated models, namely the DANIEL system, BiLSTM-based and Transformer-based models, described in 3.1. Overall, the XLM-ROBERTA-BASE was the best performing model across all the datasets. As shown in Table 2, XLM-ROBERTA-BASE recorded the highest F1 and recall scores with 79.82% and 89.2% respectively, on the dataset comprising both relevant and irrelevant examples. On the other hand, the BiLSTM+LSTM had the highest precision at 78.17%. This was a significant performance difference when compared to the performance of the chosen baseline, the DANIEL system. The baseline system had a precision of 38.97%, recall of 47.32%, and an F1 of 42.74%.

When evaluating the relevant examples only, as shown in Table 3, the task is obviously easier in particular in terms of precision. Overall, XLM-ROBERTA attained the best F1-measure score of 86.59%. The model with the best recall was MBERT-UNCASED (88.57%), while the BiL-

³For this model, we used the parameters recommended in (Devlin et al., 2019).

⁴XLM-ROBERTA was trained on 2.5TB of newly created clean CommonCrawl data in 100 languages.

Only relevant documents			
Models	P	R	F1
BiLSTM+LSTM	89.46	81.56	85.33
BiLSTM+CNN	88.67	81.30	84.82
MBERT-CASED	84.17	87.01	85.57
MBERT-UNCASED	83.78	88.57	86.11
XLM-ROBERTA	88.05	85.17	86.59

Table 3: Evaluation results for the detection of disease names and locations using only the ground-truth relevant documents.

Predicted relevant documents			
Models	P	R	F1
BiLSTM+LSTM	91.10	49.92	64.50
BiLSTM+CNN	88.76	50.08	64.03
MBERT-CASED	82.22	53.71	64.98
MBERT-UNCASED	84.02	55.97	67.18
XLM-RoBERTA	81.75	60.59	69.59

Table 4: Evaluation results for the detection of disease names and locations on the relevant documents found by the classification model.

STM+LSTM model had the highest precision. It is interesting to see that MBERT-UNCASED showed the best improvement with this easier task with +24 percentage points in F1-measure.

When we test on the predicted relevant documents, errors are being propagated to the event extraction step. The recall drops significantly since some relevant documents have been discarded by the classifier but we still evaluate by comparing with all the ground-truth of relevant documents. Still, one can notice from the Table 4 the same tendency of obtaining the highest precision with the BiLSTM-LSTM model and the highest F1 with the XLM-ROBERTA-BASE model. This model seems to be the most robust since it has the lowest drop in recall among all the models.

Finally, the performance of the models was evaluated for each language on the predicted relevant documents. As presented in Table 5, the models produced highest results for the French language which is not surprising since it is the language with the largest training dataset. The BiLSTM-based and BERT-based models performed generally well for French and Greek languages, while the XLM-ROBERTA-BASE for Chinese language. Interestingly, the worst results are for the Russian dataset. If we look at macro F1-measure, MBERT-UNCASED is the best performing model with 68.58

Model	fr	en	zh	el	ru
BiLSTM+LSTM	94.85	49.06	66.67	69.57	30.00
BiLSTM+CNN	94.87	48.15	66.67	69.57	26.09
MBERT-CASED	96.73	50.00	46.15	75.56	28.17
MBERT-UNCASED	96.19	77.15	57.14	79.55	32.35
XLM-ROBERTA	95.65	49.12	88.89	62.65	22.64

Table 5: Evaluation scores of the analyzed models for the relevant documents per language (no predicted relevant documents for Polish).

(63.8 for XLM-RoBERTA).

4 Discussion and Conclusions

First, the low precision values when training and testing on all data instances are not surprising, since the amount of negative examples, with potential false positives, rises up to around 90%. We also notice that the results when using the ground-truth documents are balanced in precision and recall, while, when testing on the predicted relevant documents, the recall is lower for the BiLSTM-based models and higher for the transformer-based models. Overall, XLM-RoBERTa-base had the best performance in terms of the F1 score. This can be attributed to the robust optimization and pre-training in a cross-lingual manner of the model on a significantly larger multilingual dataset compared to BERT.

The analysis of the performance of the models per language reveals that the best model (XLM-RoBERTa-base) had the highest scores for the French language. These results could be attributed to the size of French-language texts. For instance, French language tokens constitute 75.98% of all the tokens in the test data for the relevant documents.

We conclude that classifying the documents before detecting the events provides an advantage in performance in comparison with applying directly the event extraction task, albeit the downstream error propagation. As future work, we propose to continue with the analysis of the influence of different languages onto each other and their ability to transfer learning between languages. We also plan to address the challenges of grouping multiple sources that refer to the same event together and dealing with imperfections in the accuracy of information extraction due to multilingualism.

References

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Son Doan, Quoc Hung Ngo, Ai Kawazoe, and Nigel Collier. 2008. [Global health monitor - a web-based system for detecting and mapping infectious diseases](#). In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.
- Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions. In *Transforming Digital Worlds*, pages 356–366, Cham. Springer International Publishing.
- Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina Macintyre. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6):1–19.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Vasileios Lampsos, Bin Zou, and Ingemar Johansson Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. In *Proceedings of the 26th International Conference on World Wide Web*, pages 695–704.
- Gaël Lejeune, Romain Brixte, Antoine Doucet, and Nadine Lucas. 2015. [Multilingual event extraction for epidemic detection](#). *Artificial intelligence in medicine*, 65.
- Xuezhe Ma and Eduard Hovy. 2016. [End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.
- John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. 1999. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252. Herndon, VA.
- Victoria Ng, Erin E Rees, Jingcheng Niu, Abdelhamid Zaghou, Homeira Ghiasbeglou, and Adrian Verster. 2020. Application of natural language processing algorithms for extracting information from news articles in event-based surveillance. *Canada Communicable Disease Report*, 46(6):186–191.
- Chen-Kai Wang, Onkar Singh, Zhao-Li Tang, and Hong-Jie Dai. 2017. Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 33–38.

12 Multilingual Epidemiological Text Classification: A Comparative Study

Multilingual Epidemiological Text Classification: A Comparative Study

Stephen Mutuvi

Multimedia University
Kenya

smutuvi@mmu.ac.ke

Emanuela Boros

University of La Rochelle
France

emanuela.boros@univ-lr.fr

Antoine Doucet

University of La Rochelle
France

antoine.doucet@univ-lr.fr

Gaël Lejeune

Sorbonne University
France

gael.lejeune@sorbonne-universite.fr

Adam Jatowt

Kyoto University
Japan

jatowt@gmail.com

Moses Odeo

Multimedia University
Kenya

modeo@mmu.ac.ke

Abstract

In this paper, we approach the multilingual text classification task in the context of the epidemiological field. Multilingual text classification models tend to perform differently across different languages (low- or high-resource), more particularly when the dataset is highly imbalanced, which is the case for epidemiological datasets. We conduct a comparative study of different machine and deep learning text classification models using a dataset comprising news articles related to epidemic outbreaks from six languages, four low-resourced and two high-resourced, in order to analyze the influence of the nature of the language, the structure of the document, and the size of the data. Our findings indicate that the performance of the models based on fine-tuned language models exceeds by more than 50% the chosen baseline models that include a specialized epidemiological news surveillance system and several machine learning models. Also, low-resource languages are highly influenced not only by the typology of the languages on which the models have been pre-trained or/and fine-tuned but also by their size. Furthermore, we discover that the beginning and the end of documents provide the most salient features for this task and, as expected, the performance of the models was proportionate to the training data size.

1 Introduction

Monitoring and containment of infectious disease outbreaks have been an ongoing challenge globally. Whether previously with Ebola or today with the Covid-19 pandemic, surveillance has remained a key component of public health strategy to contain the diseases. The ability to detect disease outbreaks in an accurate and timely manner is critical in the deployment of efficient intervention measures. For instance, Ebola cases and outbreaks need to be immediately detected in order to be contained and stopped. A delayed response can have a significant economic and social impact, in addition to increased morbidity and mortality rates. Thus, the detection needs to be done as soon as the first reports appear, and, naturally, as such reports may not be in English, there is a need for effective multilingual surveillance systems.

In recent years, there has been a rapid increase in data generated as a result of the progressive evolution of the Internet. The proliferation of digital data sources provide an avenue for data-driven surveillance, referred to as Epidemic Intelligence. Epidemic intelligence involves the collection, analysis, and dissemination of key information related to disease outbreaks, with the objective of detecting outbreaks and providing early warning to public health stakeholders (World Health Organization, 2014). Natural Language Processing (NLP) techniques have made it possible to analyze data from web sources, such as social media, search queries, blogs, and online news articles for health-related incidents and/or events (Salathé et al., 2013; Bernardo et al., 2013). Data-driven epidemic intelligence can be viewed as a two-step process comprising a classification task followed by the event extraction task (Joshi et al., 2019), which can help to predict the epidemic disease dynamics and where the next outbreak of epidemic would most likely happen. The classification task entails the identification of texts relevant to disease outbreaks

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

from a large collection of data. Considering, for instance, a dataset of online news articles, the articles that report an outbreak of disease are separated from those which do not. Next, the event extraction task takes the identified relevant documents as input and predicts health-related events with arguments such as the disease name and the location where the outbreak was reported. Like any other task that involves text analysis using NLP approaches, ambiguity is a key challenge when dealing with epidemic-related text data. Ambiguity manifests itself where a sentence in a document may have mentioned a disease, but may not necessarily be reporting on an outbreak of a disease. For instance, with the ongoing coronavirus pandemic, there are numerous news articles posted daily reporting on various aspects related to the disease. It becomes a challenge to extract the few relevant news articles that are of interest to the epidemiologist, articles that report on the number and location of new cases. Epidemic reporting is also characterized by news reports from divergent sources and languages which further compounds computational epidemiology. Furthermore, when working in a multilingual setting of real-world data, another challenge arises from the lack of annotated data for low-resource languages. The creation of such data can be expensive, time-consuming, and requires human expertise to annotate, hence it is a labor-intensive task.

In view of these challenges, appropriate NLP approaches are required in order for data-driven epidemic surveillance to be successful. Therefore, we seek to provide a comprehensive quantitative study of low-shot text classification models applied on a dataset comprising news articles about disease outbreaks from several diverse language families namely, English, Greek, French, Russian, Polish, and Chinese.

We seek to compare state-of-the-art approaches for epidemiological text classification from both deep learning and classical machine learning techniques by training a variety of models and evaluating them in several circumstances, in order to analyze their application in a real-world scenario. To the best of our knowledge, this is the first extensive study to specifically evaluate the performance of multilingual epidemiological text classification methods.

The remainder of this paper is organized as follows. Section 2 reviews works related to NLP-based epidemic surveillance systems, Section 3 describes the dataset used in the study, while the experiment setup and results are presented in Section 4. Finally, we provide a discussion of the results in Section 5, and the conclusions and possible suggestions for future research are presented in Section 6.

2 Related Work

There are a number of empirical works targeted at the application of NLP for the detection of disease outbreaks. Among them is Data Analysis for Information Extraction in any Language (DANIEL), a multilingual news surveillance system that leverages repetition and saliency (the beginning and the end of a news text often comprises the salient zones), properties that are common in news writing (Lejeune et al., 2015). By avoiding grammar analysis and the usage of language-specific NLP toolkits (e.g., Part-of-speech tagger, dependency parser) and by focusing on the general structure of journalistic writing style (Hamborg et al., 2018; Lucas, 2009), the system is able to detect crucial information from news articles. Furthermore, the multilingual nature of the system enables global and timely detection of epidemic events since it eliminates the requirement for translating local news to other languages for subsequent transmission. The system can easily be adapted and scaled to extract events across languages, therefore, being able to have a wider geographical coverage. Reactivity and geographic coverage are of paramount importance in epidemic surveillance (Lejeune et al., 2015).

Similar to DANIEL is BIOCASTER (Collier, 2011; Collier et al., 2008) which has produced good results in analyzing disease-related news reports and in providing a summary of the epidemics. The BIOCASTER, an ontology-based text mining system, processes, and analyzes web text for the occurrence of disease outbreak in four phases namely, topic classification, Named Entity Recognition (NER), disease/location detection, and event recognition. The Naïve Bayes algorithm is used for the classification of the reports for topical relevance. The news stories identified to be relevant to disease outbreaks are propagated to the subsequent levels of processing. The major limitation of the BIOCASTER is its inability to detect disease outbreaks beyond the eight languages (Chinese, English, French, Japanese, Korean, Spanish, Thai, and Vietnamese) present in its ontology (Doan et al., 2019). Therefore, scaling the system to work across different languages requires manually updating the ontology with information

for new languages. In addition, the system is not publicly available, except for the ontology.

The EcoHealth Alliance Global Rapid developed the Identification Tool System (GRITS), an application that provides automatic analyses of epidemiological texts. The system extracts important information about a disease outbreak, such as the most likely disease, dates, and countries where the outbreak originates. The pipeline for GRITS entails transforming words to vectors using the term frequency-inverse document frequency (TF-IDF) method, by first extracting features using pattern-matching tools, before applying a binary relevance-based classifier to predict the presence of a disease name in the text (Huff et al., 2016). The system translates non-English documents using the Bing translator, which can potentially introduce errors to subsequent analysis steps if the translation is incorrect (Huff et al., 2016).

Internet search queries have also been exploited for disease surveillance. In one study, internet searches for specific cancers were found to be correlated with their estimated incidence and mortality (Cooper et al., 2005). Monitoring influenza outbreak using data drawn from the Web has also been previously explored. Two different studies, one that analyzes large numbers of Google search queries to track influenza-like illness in a population (Ginsberg et al., 2009) and the other that examines search queries from Yahoo¹ related to the same aforementioned infectious disease (Polgreen et al., 2008) were conducted in this context.

In recent years, various studies have utilized social media data for infectious disease surveillance (Paul et al., 2016; Charles-Smith et al., 2015). Mostly, Twitter data, has been used for disease tracking (Lamb et al., 2013; Collier et al., 2011; Culotta, 2010), outbreak detection (Li and Cardie, 2013; Bodnar and Salathé, 2013; Diaz-Aviles et al., 2012; Aramaki et al., 2011) and predicting the likelihood of individuals falling sick (Sadilek et al., 2012). News media has also been used to give early warning of increased disease activity before official sources have reported (Brownstein et al., 2008). The studies have demonstrated the potential value of harnessing data-driven approaches for epidemic surveillance.

While prior attempts to develop multilingual epidemic surveillance systems have been made, the proposed systems are predominantly ontology-based, which require the ontologies to be updated on an ongoing basis in order to improve their performance and ensure broad coverage of different languages. Recently, there has been a growing interest in exploring the multilingual nature of the data in different domains, which could also be beneficial to the epidemiological domain. Existing multilingual methods use word representations that are either learned jointly using parallel corpora (Gouws et al., 2015; Luong et al., 2015) or via mapping separately trained word embeddings in different languages to a shared space through linear transformations (Artetxe et al., 2018; Mikolov et al., 2013). The embedding spaces of the different languages ought to have a similar structure for the linear mapping from one space to the other to be effective.

More recently, effective cross-lingual representations have been developed, by simultaneously training contextual word embedding models over multiple languages, without requiring mapping to a shared space. Such models learn representations of unlabeled data that generalize across languages. Examples include the cross-lingual language model (XLM) (Lample and Conneau, 2019) and multilingual BERT (Devlin et al., 2019) which are pre-trained on Wikipedia data for different languages. BERT has been shown to allow effective cross-lingual transfer on different downstream tasks. This includes, document classification (Qin et al., 2020; Wu and Dredze, 2019), named entity recognition (NER) (Wu and Dredze, 2019; Pires et al., 2019), sentiment classification (Qin et al., 2020), neural machine translation, (Kudugunta et al., 2019), and dependency parsing (Kondratyuk and Straka, 2019).

3 Dataset

We extend the dataset proposed by Mutuvi et al. (2020) to include additional languages so that it covers news articles from several, diverse language families: Germanic: English (en), Hellenic: Greek (el), Romance: French (fr), Slavic: Russian and Polish, and Chinese that descends from the Sino-Tibetan family. The articles were obtained from different online news sources with articles relevant to disease outbreak being obtained mainly via the Program for Monitoring Emerging Disease (ProMED)² platform,

¹<http://search.yahoo.com>

²<https://promedmail.org/>

which is a program from the International Society for Infectious Diseases that tracks infectious disease outbreaks and acute exposures to toxins, across the world.

Language	#Documents	#Sentences	#Tokens
English (en)	3,562	117,190	2,692,942
French (fr)	2,415	70,893	1,959,848
Polish (pl)	341	9,527	151,901
Russian (ru)	426	6,865	133,905
Chinese (zh)	446	4,555	236,707
Greek (el)	384	6,840	183,373

Table 1: Dataset statistics.

The process of gathering the data involved first, retrieval of ProMED news articles published between August 1, 2013, and August 31, 2019. The articles clearly annotate the title, the description that captures details about the reported disease, location, date, and the source Uniform Resource Locator (URL) where the article was originally published. The source URLs were extracted and their corresponding source documents downloaded to form the relevant documents of the dataset. On the other hand, the irrelevant news articles consist of general health-related news, but without direct or indirect mentions of disease outbreaks (e.g., *plague*, *cholera*, *cough*), as well as general news like politics and sports. Most of the irrelevant documents were obtained from the News Category Dataset (Misra, 2018) comprising HuffPost³ news articles for the period 2012 to 2018. The news articles cover various topics such as culture, politics, wellness, among other topics.

	All	Polish	Chinese	Russian	Greek	French	English
Train	5,074 (10.8)	241 (7.4)	300 (2.6)	296 (9.45)	253 (6.7)	1,593 (10.9)	2,365 (11.7)
Validation	1,250 (10.9)	54 (7.4)	71 (2.8)	60 (10.0)	68 (10.2)	388 (13.4)	583 (12.6)
Test	1,250 (10.5)	46 (13.0)	75 (6)	70 (10.0)	63 (4.7)	434 (12.4)	614 (12.8)

Table 2: The number of documents (percentage of relevant documents) per dataset split.

To simulate the real scenario of news reporting, we set the number of documents reporting disease outbreak (relevant documents) to be no more than 10% of the total dataset. The statistics of the dataset are presented in Table 1.

We split the data, with a total of 7,574 articles, into training, validation, and testing sets. The training set comprises a total of 5,074 documents, while the remaining documents were shared equally between the validation and the testing sets, that is, 1,250 documents for validation, and 1,250 documents for testing, stratified by language, as shown in Table 2. We also present the percent of relevant articles that refer to epidemiological news, which depicts the imbalanced nature of the dataset.

4 Experiments

The metrics considered in the evaluation of the models are precision, recall, and F1-score. Measuring recall is particularly important because of the risk posed by not identifying all the positive cases, with regard to disease outbreaks.

4.1 Models and Hyperparameters

Baseline model: DANIEL As a baseline model, we chose DANIEL⁴ (Lejeune et al., 2015), an unsupervised system that does not rely on any language-specific grammar analysis and considers text as a sequence of strings instead of words. Consequently, DANIEL can be easily adapted to operate on any

³<https://www.huffpost.com/>

⁴<https://github.com/NewsEye/event-detection/tree/master/event-detection-daniel>

language and extract crucial information early on, which can significantly improve the decision-making process. This is pivotal in epidemic surveillance since timeliness is key, and more than often, initial medical reports are in the vernacular language where patient zero appears (Lejeune et al., 2015). We did not evaluate the BIOCASTER because only the ontology is publicly available and covers a limited number of languages, while GRITS is targeted to mostly English text.

Machine Learning models We also investigate three commonly used text classification models as baselines, Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM), using default hyperparameters and the TF-IDF weighting measure⁵.

Deep Learning models Firstly, we consider two models, a CNN and a BiLSTM with FastText (Joulin et al., 2016) word representations for all the languages in the dataset, with an embedding dimension of 300. For the CNN, a sequence of word embeddings is passed through a convolution of kernel size 3 and a filter size of 250. Similarly, the BiLSTM passes the word embeddings through a bi-directional LSTM with a cell size of 128. Other hyperparameters for the models are a batch size of 32, a learning rate of 1×10^{-2} , and 15 epochs with early stopping of 3 to avoid overfitting.

Additionally, we chose to perform experiments with different BERT-based architectures (Devlin et al., 2018) for the sequence classification task. We used the default hyperparameters, a learning rate of 2×10^{-5} , and a maximum length of 512 tokens, with the longer sentences truncated to the defined maximum length. The pre-trained models are the `bert-base-multilingual-cased` and `uncased`. Finally, the CNN/BiLSTM described earlier in this section, but this time utilizing BERT features were also evaluated. We also test a graph convolutional networks (GCN) based-approach that augments BERT with graph embeddings (VGCN+BERT) (Lu and Nie, 2019). A GCN is a multilayer neural network that calculates directly on a graph and induces embedding vectors of nodes based on properties of their neighborhoods. Combining the capabilities of BERT with GCNs has been shown to be effective in capturing both local information and global information.

Models	Precision %	Recall %	F1 %
DANIEL	33.9	60.61	43.48
LR	93.81	68.94	79.48
RF	95.70	67.42	79.11
SVM	91.26	71.21	80
CNN+FastTtext	86.11	70.45	77.5
BiLSTM+FastTtext	77.44	78.03	77.74
BERT (cased) [†]	88.62	82.58	85.49
CNN+BERT (cased) [†]	88.79	71.97	79.5
BiLSTM+BERT (cased) [†]	90.20	69.70	78.63
BERT (uncased) [†]	84.67	87.88	86.25
CNN+BERT (uncased) [†]	82.14	87.12	84.56
BiLSTM+BERT (uncased) [†]	83.72	81.82	82.76
BERT (cased)	80.71	85.61	83.09
CNN+BERT (cased)	86.67	78.79	82.54
BiLSTM+BERT (cased)	75.95	90.91	82.76
BERT (uncased)	88.52	81.82	85.04
CNN+BERT (uncased)	86.07	79.55	82.68
BiLSTM+BERT (uncased)	81.51	73.48	77.29
VGCN+BERT	87.18	77.27	81.93

Table 3: Evaluation scores of the analyzed models for the relevant documents for all languages. The pre-trained BERT models are `base-multilingual`. LR stands for Logistic Regression, RF for Random Forest, and SVM for Support Vector Machines, [†]fine-tuned.

⁵<https://scikit-learn.org/>

4.2 Results

Deep learning transfer approaches such as BERT have demonstrated the ability to outperform the state-of-the-art methods on larger datasets. However, when there exist only a few labeled examples per class, 100 to 1,000 as is the case of the low-resourced languages present in the dataset used in this study, the choice of the most appropriate approach is less clear, with classical machine learning and deep transfer learning presenting plausible options. Results of the experiments for different machine learning and deep learning models, using the dataset splits indicated in Table 2 are presented in Table 3 and discussed below.

Regarding the machine learning methods, we notice, from the results in Table 3, that SVM outperforms by a small margin the LR and RF on precision and recall, while the RF has not only the highest precision (95.70%) among this category of models, but the highest compared to all the models analyzed. We observe that the machine learning models (LR, RF, SVM) are greatly imbalanced, registering the highest values in precision and the lowest in the recall. This can be detrimental to the interests of an epidemiological detection system. Compared with the baseline results provided by DANIEL, this specialized model had a higher recall than precision which proves the specialized nature of such a tool, although its recall is the lowest among all the methods compared.

On the other hand, the models based on either CNN or BiLSTM with FastText embeddings have lower F1 scores than the classical machine learning methods (LR, RF, SVM). This could be explained by the fact that the training data is insufficient to train the models to have the ability to better distinguish between relevant and irrelevant documents.

In the case of deep transfer learning models, one can notice a great difference in the F1 score performance of BERT-based models, compared to all the other models. We can also observe that BERT-based models manage to balance recall and precision (precision remains consistent despite the increase in recall). The models benefit from the pre-trained language models that are either used as features or fine-tuned on the task. BERT relies on Byte Pair Encoding (BPE) based WordPiece tokenization (Wu et al., 2016) which makes it more robust to handle out-of-vocabulary words.

Models	Polish	Chinese	Russian	Greek	French	English
DANIEL	40	80	33.33	33.33	71.43	32.23
LR	0	0	66.67	66.67	84.21	80
RF	0	0	40	66.67	86.84	78.83
SVM	0	0	33.33	0	87.18	81.38
CNN+FastText	0	0	0	0	84.21	81.88
BiLSTM+FastText	0	0	0	0	73.12	85.71
BERT (cased) [†]	50	80	66.67	66.67	94.12	82.89
CNN+BERT (cased) [†]	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (cased) [†]	0	80	40.00	66.67	87.36	86.27
BERT (uncased) [†]	57.14	80	50	100	91.95	86.08
CNN+BERT (uncased) [†]	50	80	66.67	40	86.05	86.75
BiLSTM+BERT (uncased) [†]	0	80	40	66.67	87.36	86.27
BERT (cased)	33.33	80	50	66.67	87.50	85.54
CNN+BERT (cased)	0	0	40	66.67	83.33	86.45
BiLSTM+BERT (cased)	0	80	22.22	28.57	85.11	88.37
BERT (uncased)	0	66.67	85.71	66.67	87.18	86.25
CNN+BERT (uncased)	0	50	40	66.67	82.35	86.45
BiLSTM+BERT (uncased)	0	0	33.33	0	72.94	84.42
VGCN+BERT	71.43	88.89	88.89	80	87.80	78.26

Table 4: F1-micro scores of the analyzed models for the relevant documents per language. The pre-trained BERT models are `base-multilingual`. LR stands for Logistic Regression, RF for Random Forest, and SVM for Support Vector Machines, [†]fine-tuned.

Regarding the difference between the fine-tuned BERT-based models and those that use the BERT encoder for generating features only, the performance is slightly better when BERT is fine-tuned on the task. However, in the case of additional layers on top of the BERT encoder, when fine-tuned, a considerable decrease in performance can be seen. Overall, these results suggest that the deep learning approaches are capable of much deeper and complex representations, such that they can utilize previously learned features for newer documents, even when the language of the document differs.

As observed in Table 4, all the machine learning models (LR, RF, SVM) display similar trends in their unequal performance based on language by not detecting (having the F1 values of zero) the relevant documents in Polish and Chinese. This is likely due to the size of the training data for these particular languages. Similarly, for all the low-resource languages (Polish, Chinese, Russian, and Greek), unsurprisingly, the CNN and BiLSTM -based models with pre-trained FastText embeddings were not able to distinguish relevant documents from irrelevant ones, as indicated by their low F1 scores. This might be due to the low embedding coverage of the languages. The F1 values for Chinese tend to be consistent for all BERT-based models while the performance for Polish varies a lot between models. VGCN+BERT had the highest F1 scores for the low-resourced languages Polish, Chinese, and Russian and the second-highest for Greek.

In order to analyze the influence of the documents with a larger quantity of documents (French and English, around 2,000 news articles) over the classification of low-resource languages, we consider every language as the source language and the other five languages as target languages. At every iteration, the best performing model from the previous experiments is trained on the data in the source language and applied directly to every target language.

Train \ Test	Polish	Chinese	Russian	Greek	French	English
Polish	40	0	66.67	66.67	76.92	85.71
Chinese	0	80	60	0	70.97	81.08
Russian	33.33	0	33.33	66.67	62.86	88.61
Greek	0	0	0	66.67	0	63.05
French	0	66.67	57.14	0	91.95	85.90
English	50	0	33.33	66.67	39.29	84.35

Table 5: Evaluation scores of the BERT (multilingual-uncased)[†] fine-tuned model for the relevant documents in a zero-shot transfer learning setting.

The performance of models trained on the English and French documents is consistently higher than models trained on the other languages, as shown in Table 5. This can mainly be attributed to the larger quantity of annotated data (> 2,000 documents for training) for the two languages compared to the other languages. Also, English typology more closely resembles French typology as it has more recent influence from French and other Romance languages. The two languages share lexical similarities and cognate words. Looking at familial origins of the Slavic languages, Russian and Polish, the languages have typological properties that are intuitively more important for a model based on a language model. However, we noticed that their performance varies greatly in the case of Polish, and less in the case of Russian. Considering the quantity of training data, the difference of only around 50 more documents for Russian in train set compared with Polish seems to influence the performance.

4.2.1 Effect of Article Structure

In the approach presented by Lejeune et al. (2015), the document is considered as the main unit and it has language-independent organizational properties. The assumption is that the document-detectable features at a document granularity offer high robustness at the multilingual scale. The author suggests using the text as a minimal unit of analysis beyond its relation to the genre from which it came. The press article is thus of this type, which has precise rules: the structure of the press article and the vocabulary used are established and there are well-defined communication aims known to the source as well as the

target of the documents. These rules, at a higher level than the grammatical rules, are very similar in different languages, and from the knowledge of these rules, remarkable positions are defined which are independent of languages. To exploit particular zones of news article content, we perform experiments similar to (Lejeune et al., 2015) inspired by the work on genre invariants carried out by Giguët and Lucas (2004) and Lucas (2009). The different areas of texts that we analyze are as follows:

- Beginning of the text: ideally composed of the title of the article
- Beginning of body: containing the first two paragraphs
- End of body (foot): comprising the last two paragraphs
- Rest of body: made up of the rest of the textual elements (e.g., paragraphs)

Text Position	Models	Precision %	Recall %	F1 %
Beginning	VGCN+BERT	87.18	77.27	81.93
	BERT (uncased) [†]	84.67	87.88	86.25
Body	VGCN+BERT	79.83	71.97	75.70
	BERT (uncased) [†]	75.71	80.30	77.94
End	VGCN+BERT	72.93	73.48	73.21
	BERT (uncased) [†]	76.12	77.27	76.69
Beginning+End	VGCN+BERT	86.61	83.33	84.94
	BERT (uncased) [†]	85.61	90.15	87.82

Table 6: Performance based on portions of the documents using the best performing model, BERT (uncased) fine-tuned and the VGCN-based model. The pre-trained BERT models are base-multilingual. All positions of text have a limit of 512 tokens.

The results, as presented in Table 6, indicate that the combination of the beginning and the concluding text in the news documents provided the best features required to classify a document as either relevant or irrelevant to a disease outbreak. The lowest performance score was noted when the body and the conclusion were evaluated independently.

4.2.2 Effect of Training Data Size

Different sizes of the training data were selected at an interval of ten percent and evaluated to ascertain the impact on the overall performance of the best model, in this case, the BERT (multilingual-uncased) fine-tuned model.

We observe that there is a generally positive trend for F1 score performance when trained on increasingly large datasets, as can be seen in Figure 1. When using only 10% of the data, the model achieves an F1 score performance that is comparable to that of the classical machine learning models and plateaus at 30% of the data. It is worth noting that the model achieves an F1 score of 64.03 using 5% of the training data, which is a significant performance for such a minimal amount of data.

5 Discussion

Out of all the models, deep learning BERT-based models were the best performing models, in terms of both F1 score and recall measures. The good performance can be attributed to the deep network architectures and large corpora used to train Transformer-based pre-trained language models (PLMs) such as BERT, which enable learning of rich text representations. Moreover, BERT fine-tuning performed better compared to the feature-based approaches, where FastText and BERT embeddings were used as input features to CNN and BiLSTM classifiers. Essentially, the PLMs end up learning universal language representations that are beneficial to downstream tasks.

The high precision and low recall noted in the machine learning models suggest that the models are unable to detect the relevant class well but are highly reliable when they do. This implies that while the

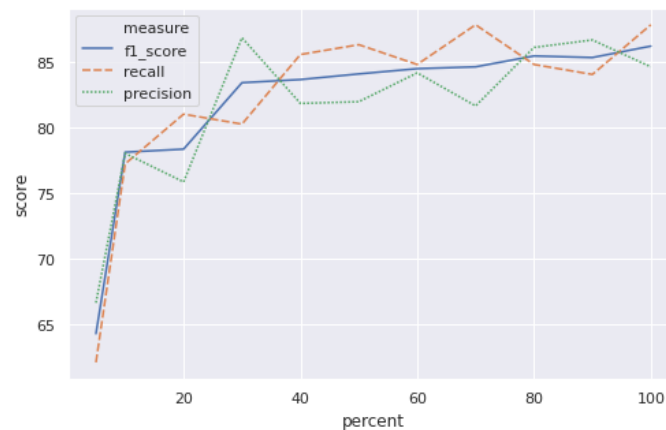


Figure 1: Impact of data size on performance of the best performing model: BERT (multilingual-uncased) fine-tuned.

classifiers returned reliable results, the machine learning models had a high false-negative rate, hence a few of all relevant results were returned. The approaches based on fine-tuned BERT uncased generally struck a good balance between precision and recall.

VGCN+BERT performed particularly well for Polish, Chinese, and Russian. The model utilizes graph embeddings produced by integrating local information captured by BERT and global information from the vocabulary graph that is based on word co-occurrence information. Both the local and global information interact with each other through a self-attention mechanism during the learning process. The interaction introduces useful global information to BERT, which contributes to the improved results across all the languages, including the low-resource languages.

With regard to the contribution of various document segments on performance, it was observed from the results that, the beginning and the end of the text combined had the highest recall and F1 score. This was particularly the case for models based on BERT namely, VGCN+BERT and BERT fine-tuned models. This can be explained by the fact that the beginning paragraphs in an article often capture the most important information, which informs the reader what the story is about. On the other hand, the last part of the article tends to provide a summary of the article.

The performance of the model improved proportionately with training data size. This is in line with neural network models, which require large amounts of data to train and evaluate. The competitive performance even with a small amount of data results from the transfer of knowledge from the pre-trained language model, trained on a large corpus, to the specific task of classifying epidemic text. This demonstrates the extent to which transfer learning can benefit the process of extracting useful information from multilingual epidemiological text.

6 Conclusions

Building effective epidemiological surveillance systems is of high importance these days. Detection of news reports on disease outbreaks is a crucial requirement of such systems. In this paper, we study in detail the performance of different methods on the task of epidemiological news report detection. The evidence presented in this work suggests that the models based on fine-tuned language models and/or graph convolutional networks achieve very good performance ($> 90\%$) on the classification of multilingual epidemiological texts, not only for high-resource languages but also for low-resource languages. In future work, we will consider the perspective of pursuing the task of epidemiological event extraction from news texts in low-resourced languages.

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

References

- Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Theresa Marie Bernardo, Andrijana Rajic, Ian Young, Katie Robiadek, Mai T Pham, and Julie A Funk. 2013. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research*, 15(7):e147.
- Todd Bodnar and Marcel Salathé. 2013. Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 699–702. Acm.
- John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. 2008. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, 5(7):e151.
- Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10):e0139701.
- Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, et al. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.
- Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. 2011. Omg u got flu? analysis of shared health messages for bio-surveillance. *Journal of biomedical semantics*, 2(5):S9.
- Nigel Collier. 2011. Towards cross-lingual alerting for bursty epidemic events. *Journal of Biomedical Semantics*, 2(5):S10.
- Crystale Purvis Cooper, Kenneth P Mallon, Steven Leadbetter, Lori A Pollack, and Lucy A Peipins. 2005. Cancer internet search activity on a major search engine, united states 2001–2003. *Journal of medical Internet research*, 7(3):e36.
- Aron Culotta. 2010. Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv:1007.4748*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Ernesto Diaz-Aviles, Avaré Stewart, Edward Velasco, Kerstin Denecke, and Wolfgang Nejdl. 2012. Epidemic intelligence for the crowd, by the crowd. In *Sixth International AAAI Conference on Weblogs and Social Media*.
- Son Doan, Quoc-Hung Ngo, Ai Kawazoe, and Nigel Collier. 2019. Global health monitor: A web-based system for detecting and mapping infectious diseases. *arXiv preprint arXiv:1911.09735*.
- Emmanuel Giguët and Nadine Lucas. 2004. La détection automatique des citations et des locuteurs dans les textes informatifs. *Le discours rapporté dans tous ses états: Question de frontières*, pages 410–418.
- Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012.

- Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments.
- Felix Hamborg, Soeren Lachnit, Moritz Schubotz, Thomas Hepp, and Bela Gipp. 2018. Giveme5w: Main event retrieval from news articles by extraction of the five journalistic w questions. 03.
- Andrew G Huff, Nathan Breit, Toph Allen, Karissa Whiting, and Christopher Kiley. 2016. Evaluation and verification of the global rapid identification of threats system for infectious diseases in textual data sources. *Interdisciplinary perspectives on infectious diseases*, 2016.
- Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina Macintyre. 2019. Survey of text-based epidemic intelligence: A computational linguistics perspective. *ACM Computing Surveys (CSUR)*, 52(6):1–19.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.
- Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China, November. Association for Computational Linguistics.
- Alex Lamb, Michael J Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Gaël Lejeune, Romain Brixte, Antoine Doucet, and Nadine Lucas. 2015. Multilingual event extraction for epidemic detection. *Artificial intelligence in medicine*, 65(2):131–143.
- Jiwei Li and Claire Cardie. 2013. Early stage influenza detection from twitter. *arXiv preprint arXiv:1309.7340*.
- Zhibin Lu and Jian-Yun Nie. 2019. Raligraph at hasoc 2019: Vgcn-bert: Augmenting bert with graph embedding for offensive language detection.
- Nadine Lucas. 2009. *Modélisation différentielle du texte, de la linguistique aux algorithmes*. Ph.D. thesis, Université de Caen.
- Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.
- Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.
- Rishabh Misra. 2018. News category dataset. <https://www.kaggle.com/rmisra/news-category-dataset>.
- Stephen Mutuvi, Antoine Doucet, Gaël Lejeune, and Moses Odeo. 2020. A dataset for multi-lingual epidemiological event extraction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4139–4144, Marseille, France, May. European Language Resources Association.
- Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.
- Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448.
- Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2020. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*.

- Adam Sadilek, Henry Kautz, and Vincent Silenzio. 2012. Predicting disease transmission from geo-tagged microblog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.
- Marcel Salathé, Clark C Freifeld, Sumiko R Mekaru, Anna F Tomasulo, and John S Brownstein. 2013. Influenza a (h7n9) and the importance of digital epidemiology. *The New England journal of medicine*, 369(5):401.
- World Health Organization. 2014. Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version. Technical report, World Health Organization.
- Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

13 Relation Classification via Relation Validation

Relation Classification via Relation Validation

Jose G. Moreno

University of Toulouse
IRIT, UMR 5505 CNRS
F-31000, Toulouse, France
jose.moreno@irit.fr

Antoine Doucet

University of La Rochelle
L3i
F-17000, La Rochelle, France
antoine.doucet@univ-lr.fr

Brigitte Grau

LIMSI, UPR 3251 CNRS
F - 91405 Orsay, France
bg@limsi.fr

Abstract

Recognising if a relation holds between two entities in a text plays a vital role in information extraction. To address this problem, multiple models have been proposed based on fixed or contextualised word representations. In this paper, we propose a meta relation classification model that can integrate the most recent models by the use of a related task, namely relation validation. To do so, we encode the text that may contain the relation and a relation triplet candidate into a sentence-triplet representation. We grounded our strategy in recent neural architectures that allow single sentence classification as well as pair comparisons. Finally, our model is trained to determine the most relevant sentence-triplet pair from a set of candidates. Experiments on two public data sets for relation extraction show that the use of the sentence-triplet representation outperforms strong baselines and achieves comparable results when compared to larger models.

1 Introduction

Recognising and classifying relations between two entities in a text plays a vital role in knowledge base population (KBP), a major sub-task of information extraction (IE). Some examples of typical relations in knowledge bases (KB) are spouse, CEO, place of birth, profession, etc. Nowadays, there exist large KB that store millions of facts such as DBpedia (Bizer et al., 2009) or YAGO (Hoffart et al., 2013). However, more than 70% of people entities have not associated information for relations such as place of birth or nationality (Dong et al., 2014).

Most approaches model the relation classification (RC) (dos Santos et al., 2015; Nguyen and Grishman, 2015) task as a learning problem where it is required to predict if a passage contains a type of relation (multi-class classification). This setup requires annotated examples of each class, i.e. each

type of relation, which can be difficult to obtain. To overcome this problem, distant supervision has been proposed (Mintz et al., 2009) for automatically annotating texts given relation triplets existing in a KB by projecting triplets into texts to increase the input data. Its main counterpart is that distant supervision models must deal with wrongly annotated examples. The difficulty of the task is shown by the results of the TAC KBP slot filling task¹. For instance, in 2014, the maximum F1-score of the task was 0.3672 (Surdeanu and Ji, 2014). Another trend is trying to collect information directly from the web in an unsupervised setting, i.e. the open IE paradigm (Banko et al., 2007). In these two last settings, one crucial point is to be able to assess the validity of the extracted relations. This point motivated an extra track in TAC KBP 2015 following a divide-and-conquer setup. It consists in validating the relations extracted by relation extraction (RE) systems in order to improve their final scores.

The purpose of relation validation (RV) aims at taking advantage of several hypotheses, provided by one or several systems, for improving the recognition of relations in texts and discarding false ones. Given a candidate relation triplet $(e1, R, e2)$ and a passage, this task can be defined as learning to decide if the passage supports the relation in a binary classification setup. Trigger words and relation patterns are usually modelled in relation validation as features for representing the relation type. In Wang and Neumann (2008), the relation validation setup is modified and presented as an entailment problem, where systems learn whether the text entails the relation based on linguistic features.

In this paper, we propose not only to learn the representation of the relation type, but also to learn the representation of the validation knowledge by using a neural architecture for modelling relation

¹<https://catalog.ldc.upenn.edu/LDC2018T22>

validation, inspired by neural entailment models. We aim to decide whether the text supports the relation by encoding the text and the triplet² in a transformer architecture as in (Baldini Soares et al., 2019; Zhao et al., 2019). Once a model for relation validation is learned, we use it to validate the output of a relation classification model. Our experiments show that our proposal outperforms robust neural models for relation classification but fails to improve most recent works.

The remainder of this paper is structured as follows: Section 2 presents some relevant models for relation classification and validation. Section 3 details our strategy to classify relations based on relation validation. Then, the experimental setup and results are presented in Sections 4. Finally, conclusions are drawn in Section 5.

2 Related Work

Different ensemble models (Viswanathan et al., 2015) have been defined for the relation validation KBP task based on the prediction made by the RE systems. However, Yu et al. (2014) show that relation validation requires considering linguistic features for recognising if a relation is expressed in a text by exploiting rich linguistic knowledge from multiple lexical, syntactic, and semantic levels. In Wang and Neumann (2008), the relation to validate is transformed by simple patterns in a sentence and an alignment between the two texts is performed by a kernel-based approach.

Traditional methods for relation extraction are based on feature engineering and rely on lexical and syntactic information. Dependency trees provide clues for deciding the presence of a relation in unsupervised relation extraction (Culotta and Sorensen, 2004; Bunescu and Mooney, 2005; Fundel et al., 2007). Gamallo et al. (2012) defined patterns of relation by parsing the dependencies in open information extraction. Words around the entity mentions in sentences give clues to characterise the semantics of a relation (Niu et al., 2012; Hoffmann et al., 2011; Yao et al., 2011; Riedel et al., 2010; Mintz et al., 2009). In addition to linguistic information, collective information about the entities and their relations were exploited for RV (Rahman et al., 2018) by adding features based on a graph of entities and for RE by Augenstein

(2016) that integrated global information about the object of a relation. The latter model shows the importance of adding information about the entities in the triplet. The above approaches rely on Natural Language Processing (NLP) tools for syntactic analysis and on lexical knowledge for identifying triggers. Thus, it remains difficult to overcome the lexical gap between texts and relation names when learning relation patterns for different types of relations in an open domain.

Recently, end-to-end neural network (NN) based approaches have been emerged and getting lots of attention for the relation classification task (dos Santos et al., 2015; Nguyen and Grishman, 2015; Vu et al., 2016; Dligach et al., 2017; Zheng et al., 2016; Zhang et al., 2018). However, they do not leverage any triplet representation of a relation for better understanding the relatedness between the text and the triplet. A lot of NN models for evaluating the similarity of two sentences have been proposed. They encode each entry by a CNN or an RNN (e.g., LSTM or BiLSTM), and compute a similarity between the sentence representations (Severyn and Moschitti, 2015) or compute interactions between the texts by an attention layer (Yin et al., 2016).

Most recent models encode one or two sentences by using the pre-trained neural models. Their use in RC has been successfully tested by Baldini Soares et al. (2019) where entities are marked and the sentence representation is used. Then a simple but effective sequence classification is performed using the sentence representation token which encodes the full sentence including the marked tokens. Their performances are boosted by using more documents in an unsupervised fashion. Despite more information being used, Baldini Soares et al. (2019) do not use an explicit relation representation. In an effort to cope with this problem, we explore the use of pre-trained neural models into the RV problem by explicitly using a triplet-sentence representation.

3 Relation classification via relation validation

Our proposal first learns how to validate relations ground on a sentence-triplet representation in order to predict if a relation stands or not in a sentence. To do so, our model is based on a pre-trained BERT model for sequence classification (Devlin et al., 2018). Using pre-trained models to address RC is

²We are aware that our model mainly based its improvements on input modification. However, we strongly believe that this is unfairly underestimated in the field.

a promising strategy as shown by Baldini Soares et al. (2019). In both cases, i.e. RV or RC, a major consideration is the input definition to correctly identify the target entities, mainly because pre-trained models do not include this option by default. In this section, we present the details of the architecture together with the input transformations to correctly feed a sequence classification model such as BERT.

3.1 BERT-based Architecture

We opted for a simplified version³ of the architecture proposed in Baldini Soares et al. (2019) for relation classification, namely $BERT_{EM}$. It is based on fine-tuning of a pre-trained transformer called BERT (Devlin et al., 2018) where an extra layer is added to make the classification of the sentence representation, e.g. a classification task is performed using as input the [CLS] token. As reported by Baldini Soares et al. (2019), an important component is the use of mark symbols to identify the entities to classify.

3.2 Relation Classification

3.2.1 Problem definition

Given a tokenised sentence $S = "t_1 t_2 \dots t_n"$, an origin offset $o_o \in 1, n$, a target offset $o_t \in 1, n$, and a set of k relations $R = \{r_1, r_2, \dots, r_k\}$. The relation extraction problem consists in determining which relation $r_p \in R$ stands in the sentence between the tokens in positions o_o and o_t , respectively.⁴

3.2.2 Input considerations

We follow the input considerations for RC proposed by (Baldini Soares et al., 2019). Thus, to introduce those markers, the original input of RC models

$$\text{input}(S) = [\text{CLS}] \quad t_1 \quad t_2 \quad \dots \quad t_n \quad [\text{SEP}] \quad (1)$$

is modified to include the entities markers

$$\text{input}'(S) = [\text{CLS}] \quad \dots \$ \quad t_{o_o} \quad \$ \quad \dots \# \quad t_{o_t} \quad \# \dots \quad [\text{SEP}] \quad (2)$$

Note that $\text{length}(\text{input}'(S)) = \text{length}(\text{input}(S)) + 4$, because we added the tokens \$ and # twice.

³We used the EntityMarkers[CLS] version. Other configurations were not explored and are left for future work.

⁴Note that a *non-relation* or *other relation* may be part of the set R .

3.3 Relation Validation

3.3.1 Problem definition

Given a tokenised sentence $S = "t_1 t_2 \dots t_n"$, an origin offset $o_o \in 1, n$, a target offset $o_t \in 1, n$, and a triplet $t = \langle t_{o_o}, r, t_{o_t} \rangle$. The relation validation problem consists in determining whether the relation r between t_{o_o} and t_{o_t} is supported by the sentence S or not.

3.3.2 Input considerations

We transform triplets $t = \langle t_{o_o}, r, t_{o_t} \rangle$ into a sequence of its label words. Then we use the sentence S on one side and the triplet t on the other side as input of the model to match the relation validation problem into a text entailment setup as suggested by Wang and Neumann (2008). So, in this case, the input is modified to

$$\begin{aligned} \text{input}''(S) = & [\text{CLS}] \dots \$ \quad t_{o_o} \quad \$ \\ & \dots \# \quad t_{o_t} \quad \# \dots [\text{SEP}] \quad (3) \\ & t_{o_o} \quad t_{o_t} \quad r_{w_1} \quad r_{w_2} \dots \quad r_{w_m} [\text{SEP}] \end{aligned}$$

Note that $\text{length}(\text{input}''(S)) = \text{length}(\text{input}(S)) + 4 + (m + 2)$, because of the tokens \$ and #, and the triplet t is represented by $m + 2$ tokens (m words for the relation r and the two entities tokens). This architecture is possible because of the single or double input capabilities of transformer architectures such as BERT. Our proposed architecture is depicted in Figure 1. As for RC, we add the mark symbols in the sentence but not for the triplet. The final prediction is based on the sentence representation or the [CLS] token.

As our work focuses on relation extraction, a prior stage is needed to transform any relation classification data set into a relation validation one (i.e. as many examples as relations/classes). This transformation consists in generating $|R|$ relation validation examples for each relation extraction one, by considering the correct relation as positive and others as negatives. In this case, if \mathcal{S} is the set of examples for RC, then the set of examples for RV (\mathcal{S}_{RV}) is $|R|$ times larger than \mathcal{S} . However, to prevent imbalance, negative sampling is commonly used. In this case, $|\mathcal{S}_{RV}| = (ns + 1) \times |\mathcal{S}|$ where ns is the number of negative examples used to build \mathcal{S}_{RV} .

3.4 Validation of a classification prediction

Our main contribution is the definition of a new model for RC using RV, namely $BERT+RC+RV$.

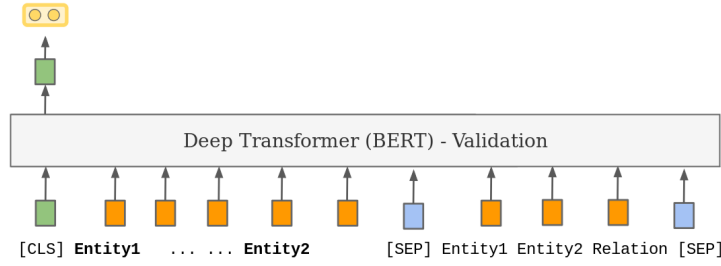


Figure 1: Our relation validation model. Tokens in bold are marked using "\$" for the Entity1 and "#" for the Entity2.

During training time our RV model behaves as described in Algorithm 1. The set \mathcal{S}_{RV} used as input is built as described in Section 3.3.2. `createInput` generates an input such as in Equation 2. The output is a relation validation model (M_{RV}) capable of detecting if the input is valid or not.

Algorithm 1: BERT+RC+RV train

Input: Set of examples \mathcal{S}_{RV} {Sentence (S), triplet (t), label (l_{RV})}

$epoch = 1$

while $epoch < max_{epochs}$ **do**

for $S, t, l_{RV} \in \mathcal{S}_{RV}$ **do**

$input''(S) = createInput(S, t)$
 update with $Loss(input''(S), l_{RV})$

Output: Validation model (M_{RV})

On the other hand, at inference time not all cases are evaluated. Our model can use as input the outputs of multiples RC models⁵ (\mathcal{S}_v) as described in Algorithm 2. Each example in \mathcal{S}_v is composed of a sentence and n_{RC} labels predicted by n_{RC} RC models, i.e. each example has a list (L) of n_{RC} predictions. Thus, our RV model defines the most suitable label based on the sentence and the triplet together instead of a classic RC model that only uses the sentence. `getTriplet` is a function based on a simple dictionary that returns the relation words (r_{w_1}, \dots, r_{w_m}) related to a label l_{rc} and the entities (t_{o_o} and t_{o_t}) in S . This way, our model is not only capable of learning from the same data but also capable of aggregating multiple RC predictions.

⁵In our experiments, we used the outputs of our implementation of a state-of-the-art RC model, $BERT_{EM}$, described in Section 4.2.

Algorithm 2: BERT+RC+RV prediction

Input: Set of examples to validate \mathcal{S}_v
{Sentence (S), labels (L)}, a Validation model (M_{RV})

$l_V = []$

for $S, L \in \mathcal{S}_v$ **do**

$l_{i-valid} = []$

for $l_i \in \text{unique}(L)$ **do**

$t = \text{getTriplet}(l_i, S)$

$input''(S) = \text{createInput}(S, t)$

$confid = \text{predict}(M_{RV}, input''(S))$

$l_{i-valid}.append(l_i, confid)$

$l_V.append(\text{labelMaxConfidence}(l_{i-valid}))$

Output: List of predictions (l_V)

4 Experiments and Results

4.1 Data Sets

In this study, we experimented on two publicly available data set: *SemEval10*⁶ and *TACRED*⁷. Statistics of these standard relation classification data sets are presented in Table 1. We created a relation validation version from both data sets as described in Section 3.3.2. The input of our RV model needs a set of relation words which, originally, are not present in the data sets. Thus, to obtain these words, we used a rather simple strategy that consists of tokenising the relations names and using them as relation words. If needed it considers the relation direction by reversing the position of the tokenised words. Table 2 shows some examples of the selected words.

In both cases, we used the respective official F_1 metric⁸ for evaluation.

⁶Task 8 (Hendrickx et al., 2010) from <http://semEval2.fbk.eu/semEval2.php?location=tasks>

⁷<https://nlp.stanford.edu/projects/tacred/>

⁸Macro-F1-measures are calculated using each script. Both scripts exclude the *other* class during evaluation.

Data set	Train	Dev	Test	# Relations
SemEval10	8000	-	2717	19
TACRED	68124	22631	15509	42

Table 1: Summary of SemEval10 and TACRED data sets for relation classification.

4.2 Implementation details

We implemented $BERT_{EM}$ (EntityMarkers[CLS] version) of Baldini Soares et al. (2019) for RC and adapted it to perform RV⁹. For SemEval10, we used 10% of training data as validation data which allows fair comparison against previous works. A maximum number of epochs was fixed to 5 and the best epoch in validation used for prediction¹⁰. Negative sampling was fixed to 10 where the input sentence remains and the entities remain the same but the words used for the relation representation ($r_{w_1}, r_{w_2}, \dots, r_{w_m}$) are sampled from other classes. *Binary Cross Entropy* was used as loss function, Adam as optimiser, *bert-base-uncased*¹¹ as pre-trained model, and other parameters were assigned following the library recommendations (Wolf et al., 2019).¹² The final layer is composed of as many neurons as classes in each data set for RC and equal to two for RV (negative or positive).

Data set	Relation	Words
SemEval10	Cause-Effect(e1,e2)	Cause, Effect
	Cause-Effect(e2,e1)	Effect, Cause
	Content-Container(e1,e2)	Content, Container
TACRED	org:founded_by	org, founded, by
	per:city_of_death	per, city, of, death
	per:age	per, age

Table 2: Examples of words used per relation.

4.3 Results

Average and best result of 5 runs of our implementation of (Baldini Soares et al., 2019) using the SemEval10 data set are presented in Table 3 ($BERT_{EM}^*$). The reported results are within the values reported in the original paper for this configuration, but we used *bert-base-uncased* instead

⁹Our code is publicly available at <https://github.com/jgmorenof/rcviarv2020>.

¹⁰Our models got the best validation performances at epoch 5, no further epochs were explored.

¹¹<https://github.com/google-research/bert>

¹²We did not perform parameters search.

	SemEval10	TACRED
$BERT_{EM}^*$ - average	87.03	65.50
$BERT_{EM}^*$ - best	87.70	66.02
$BERT+RC+RV$ - average (ours)	88.36	66.20
$BERT+RC+RV$ - best (ours)	88.44	67.48
$BERT_{EM}^*$ - voting	89.02	68.67
$BERT+RC+RV$ - voting (ours)	89.41	69.13
<i>TRE</i> (Alt et al., 2019)	87.1	67.4
<i>BERT-LSTM-base</i> (Shi and Lin, 2019)	-	67.8
<i>C-GCN+PALSTM</i> (Zhang et al., 2018)	-	68.2
<i>C-AGGCN</i> (Guo et al., 2019)	-	68.2
<i>Att-Pooling-CNN</i> (Wang et al., 2016)	88.0	-
<i>Entity-Aware BERT</i> (Wang et al., 2019)	89.0	-
<i>KnowBert-W+W</i> (Peters et al., 2019)	89.1	<u>71.5</u>
<i>R-BERT</i> (Wu and He, 2019)	89.25	-
$BERT_{EM}$ (Baldini Soares et al., 2019)	89.2	<u>70.1</u>
<i>Span-BERT</i> (Joshi et al., 2019)	-	<u>70.8</u>
$BERT_{EM}+MTB$ (Baldini Soares et al., 2019)	89.5	<u>71.5</u>
<i>EPGNN</i> (Zhao et al., 2019)	<u>90.2</u>	-

Table 3: Results of official F_1 metric for the SemEval10 and TACRED data sets. Best result of our tested models is marked in **bold**. Results that outperform our method are underlined. '*' indicates that the result was obtained by our implementation of (Baldini Soares et al., 2019). Other values were taken from referenced papers.

Number of candidates							
		2		3		4	
		Corr.	Incorr.	Corr.	Incorr.	Corr.	Incorr.
BERT+RC+RV	338	154	37	52	2	4	
	68.69%	31.30%	41.57%	58.42%	33.33%	66.66%	

Table 4: Percentage of correct (Corr.) and incorrect (Incorr.) predictions from RV model for the SemEval10 data set grouped by the number of candidates provided by RC.

Epoch					
	1	2	3	4	5
$BERT+RC+RV$	0.8790	0.8807	0.8793	0.8802	0.8831
$BERT_{EM}^*$ - run1	-	-	-	-	0.8760
$BERT_{EM}^*$ - run2	-	-	-	-	0.8683
$BERT_{EM}^*$ - run3	-	-	-	-	0.8688
$BERT_{EM}^*$ - run4	-	-	-	-	0.8770
$BERT_{EM}^*$ - run5	-	-	-	-	0.8614

Table 5: Performances for one run of our method vs $BERT_{EM}$ runs in terms of F_1 using the SemEval10 data set. We calculated our results by epoch after training.

of *bert-large-uncased* due to computational constraints. In both cases, for average and best, our results using the relation validation model outperform their counterparts by a non-negligible margin. In order to understand the cases in which *BERT+RC+RV* makes the right prediction, we have reported the percentage of correct and incorrect predictions grouped by the number of candidates in Table 4. Note that at this stage *BERT+RC+RV* does not consider the number of predictions made for a candidate (as is made by voting) but analyse each candidate independently of its popularity. Although we used 5 runs, none of the examples obtained five candidates as for every test example at least two models predicted the same class. The number of correct predictions made by our validation model is 68.69% when there are only 2 candidates but decreases as the number of candidates increase (down to 33.33% for 4 candidates). However, in most of the cases, the predictions of the relation classification model only get 2 candidates (83.81%). Clearly, this result shows that there is still room for improvement by proposing better RV models.

Following this direction, we apply majority voting¹³ over the predictions of *BERT_{EM}* and *BERT+RC+RV*. Results are included in Table 3. Note that voting benefits our baseline but also our method by a similar margin. The lower part of Table 3 allows comparing our results to those of the most recent RC models. The best result, giving an F_1 score of 0.8941 is obtained based on majority voting of the prediction from the RV model. When compared against results reported in SemEval10, our method achieves the third position slightly behind *BERT_{EM}+MTB*, but quite far from *EPGNN* (Zhao et al., 2019). However, *BERT+RC+RV* remains an easy-to-implement model as no special modification is needed when compared with *BERT_{EM}+MTB* which uses extra auto-supervised training plus a larger model¹⁴ and *EPGNN* which needs graph embeddings. Moreover, we believe that *BERT_{EM}+MTB* can be improved if more robust models are validated.

We also studied the performance of our method by epoch, as reported in Table 5. Results of *BERT_{EM}** are presented for epoch 5 as this epoch got the best validation result. Note that our method

outperforms all individual RC predictions from the first epoch and no underperformance is observed across epochs. This result suggests that our method is an effective way to mixture RC predictions.

Finally, we experimented with our model using the TACRED data set. Results are reported in Table 3. The results follow the same pattern as with the SemEval10 data set, except for one important difference: The performance obtained with *BERT_{EM}** ($F_1 = 65.50$) is much lower than the value reported by the authors ($F_1 = 69.13$). This can be explained from the fact that the number of relations in TACRED is twice as high as in SemEval10. Subsequently, more parameters allowed a richer representation and a better starting point (+4.5 absolute points w.r.t. F_1).

5 Conclusion

In this paper, we presented a new strategy to improve the neural models for relation classification by using relation validation knowledge, i.e. the sentence-triplet representation. Experiments with two public data sets experimentally support our hypothesis. The proposed strategy enables new ways to improve existing methods as it can be easily plugged into more recent (or future) and powerful models. Future work will be focused on the use of this strategy across tasks from different (and far) domains as our relation validation architecture can validate triplets with unseen relations. This opened an interesting research direction for relation classification by focusing more on triplet-sentence representations rather than exclusively on the sentence.

Acknowledgements

This work has been partly supported by the European Union’s Horizon 2020 research and innovation programme under grant 825153 (EMBED-DIA).

We also thank the anonymous reviewers for their careful reading of this paper and their many insightful comments and suggestions.

References

- Christoph Alt, Marc Hübner, and Leonhard Henning. 2019. Improving Relation Extraction by Pre-trained Language Representations. In *Proceedings of AKBC*.
- Isabelle Augenstein. 2016. *Web Relation Extraction with Distant Supervision*. Ph.D. Dissertation. University of Sheffield.

¹³The class that receives the highest number of votes will be chosen.

¹⁴*bert-large-uncased* uses three times more parameters (340 millions) than *bert-base-uncased* (110 millions).

- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the ACL*. 2895–2905.
- Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web.. In *IJ-CAI*, Vol. 7. 2670–2676.
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia-A crystallization point for the Web of Data. *Journal of web semantics* 7, 3 (2009), 154–165.
- Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on HLT and EMNLP*. Association for Computational Linguistics, 724–731.
- Aron Culotta and Jeffrey Sorensen. 2004. Dependency Tree Kernels for Relation Extraction. In *Proceedings of the 42nd Annual Meeting on ACL*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).
- Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural Temporal Relation Extraction. *EACL 2017* (2017), 746.
- Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmman, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 601–610.
- Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying Relations by Ranking with Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of ACL and the 7th International JCNLP*. 626–634.
- Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2007. RelEx-Relation extraction using dependency parse trees. *Bioinformatics* 23, 3 (2007), 365–371.
- Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of the joint workshop on unsupervised and semi-supervised learning in NLP*. 10–18.
- Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention Guided Graph Convolutional Networks for Relation Extraction. In *Proceedings of the 57th Annual Meeting of ACL*. 241–251.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th SemEval*. 33–38.
- Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194 (2013), 28–61.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of ACL*. 541–550.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529* (2019).
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th International JCNLP of the AFNLP*. 1003–1011.
- Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 39–48.
- Feng Niu, Ce Zhang, Christopher Ré, and Jude W Shavlik. 2012. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS 12* (2012), 25–28.
- Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP*.
- Rashedur Rahman, Brigitte Grau, and Sophie Rosset. 2018. Impact of Entity Graphs on Extracting Semantic Relations. In *Information Management and Big Data*. 31–47.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*. Springer, 148–163.
- Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *Proceedings of the 38th International ACM SIGIR*. 373–382.
- Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv preprint arXiv:1904.05255* (2019).

- Mihai Surdeanu and Heng Ji. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. TAC*.
- Vidhoon Viswanathan, Nazneen Fatema Rajani, Yinon Bentor, and Raymond Mooney. 2015. Stacked Ensembles of Information Extractors for Knowledge-Base Population. In *Proceedings of the 53rd Annual Meeting of ACL and the 7th International JCNLP*. 177–187.
- Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016. Combining Recurrent and Convolutional Neural Networks for Relation Classification. In *Proceedings of the 2016 Conference of the NAACL-HTL*. 534–539.
- Haoyu Wang, Ming Tan, Mo Yu, Shiyu Chang, Dakuo Wang, Kun Xu, Xiaoxiao Guo, and Saloni Potdar. 2019. Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers. In *Proc. of the 57th Annual Meeting of ACL*. 1371–1377.
- Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation Classification via Multi-Level Attention CNNs. In *Proceedings of the 54th Annual Meeting of the ACL*. 1298–1307.
- Rui Wang and Günter Neumann. 2008. Relation validation via textual entailment. *Ontology-based information extraction systems (obies 2008)* (2008).
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771* (2019).
- Shanchan Wu and Yifan He. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In *CIKM*.
- Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum. 2011. Structured relation discovery using generative models. In *Proceedings of the Conference on EMNLP*. 1456–1466.
- Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. *Transactions of the Association for Computational Linguistics* 4 (2016), 259–272.
- Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, and et al. 2014. The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding. In *Proceedings of 2014 International CICLING*.
- Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. In *Proceedings of the 2018 Conference on EMNLP*. 2205–2215.
- Yi Zhao, Huaiyu Wan, Jianwei Gao, and Youfang Lin. 2019. Improving Relation Classification by Entity Pair Graph. In *ACML*. 1156–1171.
- Suncong Zheng, Jiaming Xu, Peng Zhou, Hongyun Bao, Zhenyu Qi, and Bo Xu. 2016. A neural network framework for relation extraction: Learning entity semantic and relation pattern. *Knowledge-Based Systems* 114 (2016), 12–23.

14 Robust Named Entity Recognition and Linking on Historical Multilingual Documents

Robust Named Entity Recognition and Linking on Historical Multilingual Documents^{*}

Emanuela Boros¹[0000-0001-6299-9452],
 Elvys Linhares Pontes¹[0000-0002-9571-5193],
 Luis Adrián Cabrera-Diego¹[0000-0002-9881-9799],
 Ahmed Hamdi¹[0000-0002-8964-2135],
 Jose G. Moreno^{1,2}[0000-0002-8852-5797],
 Nicolas Sidère¹, and
 Antoine Doucet¹[0000-0001-6160-3356]

¹ University of La Rochelle, L3i, F-17000, La Rochelle, France

`firstname.lastname@univ-lr.fr`
<https://www.univ-larochelle.fr>

² University of Toulouse, IRIT, UMR 5505 CNRS, F-31000, Toulouse, France

`firstname.lastname@irit.fr`

Abstract. This paper summarizes the participation of the L3i laboratory of the University of La Rochelle in the *Identifying Historical People, Places, and other Entities* (HIPE) evaluation campaign of CLEF 2020. Our participation relies on two neural models, one for named entity recognition and classification (NERC) and another one for entity linking (EL). We carefully pre-processed inputs to mitigate its flaws, notably in terms of segmentation. Our submitted runs cover all languages (English, French, and German) and sub-tasks proposed in the lab: NERC, end-to-end EL, and EL-only. Our submissions obtained top performance in 50 out of the 52 scoreboards proposed by the lab organizers. In further detail, out of 70 runs submitted by 13 participants, our approaches obtained the best score for all metrics in all three languages both for NERC and for end-to-end EL. It also obtained the best score for all metrics in French and German for EL-only.

Keywords: Information Extraction · Named Entity Recognition · Entity Linking

1 Introduction

Identifying historical people, places and other entities is a key task in the automatic understanding of historical newspapers. However, the use of electronic

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

^{*} This work has been supported by the European Union's Horizon 2020 research and innovation program under grants 770299 (NewsEye) and 825153 (Embeddia).

formats for storing text content is relatively new in comparison to the origins of newspapers. For instance, in Europe, the first newspapers appeared at the beginning of the 17th century [25]. Electronic text files started to be widely used since the adoption of operating systems such as MS-DOS in the 1980s [2]. Thus, in the absence of electronic versions of historical newspapers, a common strategy is to recognize the text from digital images of newspapers using optical character recognition (OCR) techniques. In this context, the HIPE 2020 lab at CLEF presented an evaluation campaign with the goal of assessing the recent advances in two major NLP tasks, named entity recognition and classification (NERC) and entity linking (EL), in the context of historical newspapers [5]. This paper presents the participation of the *Laboratoire Informatique, Image et Interaction* (L3i laboratory) at the University of La Rochelle at CLEF HIPE 2020. We developed two new models for NERC and EL. Despite the fact that both models are based on neural networks, there are strong differences between them. Our NERC model is mainly based on the transformer architecture [24] while our EL model is based on a BiLSTM architecture [10]. Our main contributions are three-fold: (1) we propose a pre-processing strategy to mitigate the characteristics of input documents, (2) we extend a transformer-based model for NERC, and (3) we adapt an EL model to a multilingual context. Official results of our participation show the effectiveness of our models over the CLEF HIPE 2020 benchmark.

The remaining of the paper is organized as follows: Section 2 presents the task and the used corpus. Section 3 presents the global architecture of our participation, Section 4.1 presents the pre-processing strategy, while Sections 4 and 5 present individually our NERC and EL systems respectively.

2 HIPE Corpus and HIPE Evaluation

The HIPE corpus [4] is a collection of digitized documents covering three different languages: English, French, and German. The documents come from archives of several Swiss, Luxembourgish, and American newspapers. The dataset was annotated according to the HIPE annotation guidelines [6] which derived from the Quaero³ annotation guide.

The corpus uses the IOB format with hierarchical information and, provides training, development, and test datasets for each language, except for English. In the case of the latter, the organizers provided only partitions for development and test. In Table 1, we present the statistics regarding the number of named entities found in each dataset. See [3] for a more detailed description of the HIPE dataset.

Regarding the HIPE evaluation, it consists in assessing both tasks, NERC and EL, in terms of Precision (P), Recall (R), and F-measure (F1) at macro and micro levels [14, 3]. Two evaluation scenarios are considered: strict (exact boundary matching) and relaxed (fuzzy boundary matching).

³ Quaero guidelines: <http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011>

Table 1: Number of entities for the training, development, and test sets in HIPE 2020 corpora.

Splits	German	English	French
training	3,505	-	6,885
development	1,390	967	1,723
test	1,147	449	1,600

3 L3i NERC-EL Model for Historical Newspapers

In Figure 1, we present the global architecture of our end-to-end NERC-EL model composed of three elements. The first one is a pre-processing module, which reformats the input provided by the organizers. The second element is the NERC module, where we predict the named entities for each language, English, French, and German. The third element is the EL module, where we disambiguate the named entities, and we link them to the Wikidata.

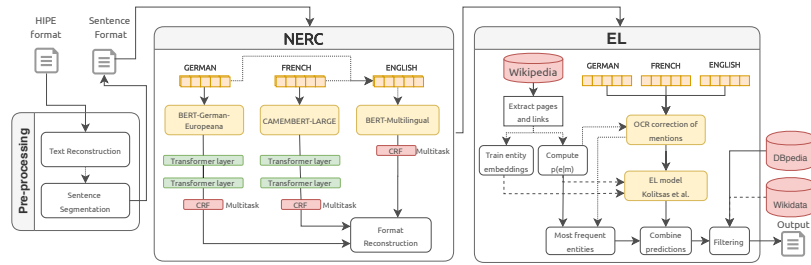


Fig. 1: Global architecture of the NERC and EL proposed models.

In the following sections, we will describe in-depth each of the modules showed in Figure 1.

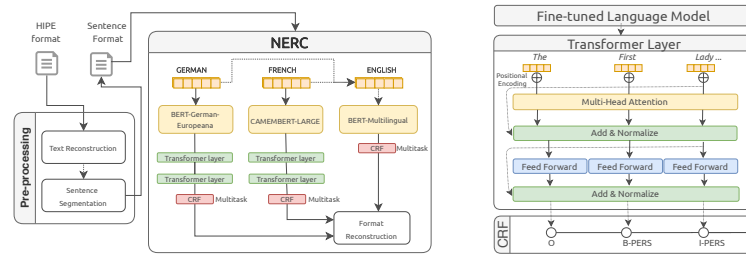
4 Named Entity Recognition and Classification (NERC)

In CLEF HIPE 2020, the NERC task consists in the recognition and classification of entities, such as people and locations, within historical multilingual newspapers. According to the organizers [3], it is composed of two sub-tasks with different levels of difficulty:

- Sub-task 1.1 - NERC coarse-grained: the identification and categorization of entity mentions according to high-level entity types, Person, Location, Organization, Product, and Time.

- Sub-task 1.2 - NERC fine-grained: the recognition and classification of entity mentions at different levels, finer-grained entity types and nested entities, up to one level of depth. It also consists in detecting the components belonging to an entity mention, such as its function, title, honorifics, and name.

Due to the complexity and characteristics of both coarse-grained and fine-grained NERC sub-tasks, we propose the use of a hierarchical, multitask learning approach consisting in a fine-tuned encoder based on *Bidirectional Encoder Representations from Transformers* (BERT) [1]. Our approach includes the use of a stack of Transformer [24] blocks on top of the BERT model for the French and German languages. The multitask prediction layer consists of six separate conditional random field (CRF) layers. The architecture of the model is presented in Figure 2.



(a) NERC architecture for all the languages, (b) Detailed model proposed for each language including the pre-processing step.

Fig. 2: The main architecture of the BERT-based model and the additional Transformers (a) is composed of modules stacked on top of each other multiple times. The transformer encoder module (b) mainly consists of multi-head attention and pointwise feed-forward layers.

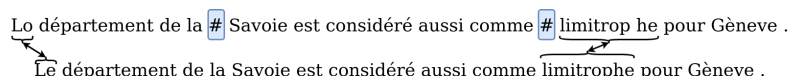
We decided to use BERT not only because it is easy to fine-tune, but it has also proved to be one of the most performing technologies in multiple NLP tasks [1, 12, 20]. However, while BERT had a major impact in the NLP community, its ability to handle noisy inputs is still an open question [23] or at least requires the addition of complementary methods [16, 19]. More specifically, the built-in tokenizer used by BERT first performs simple white-space tokenization, then applies a Byte Pair Encoding (BPE) based WordPiece tokenization [27]. A word can be split into character n-grams (e.g. “compatibility” → “com”, “##pa”, “##ti”, “##bility”), where “##” is a special symbol for representing the presence of a sub-word that was recognized. Between the types of OCR errors that can be encountered, the character insertion modification has the minimum influence [23], because the tokenization at the sub-word level of BERT would not change much in some cases, such as “practically” → “practicaally”, but the sub-

stitution and deletion errors can hurt the performance of the tokenizer the most due to the generation of uncommon samples, as such as “professionalism” → “pr9fessi9nalism”. Thus, these new noisy tokens could influence the performance of BERT-based models⁴.

The added layers consist in a stack of Transformer blocks (Transformer encoders). As proposed in [24], this model is a deep learning architecture based on multi-head attention mechanisms with *sinusoidal position embeddings*⁵. It is composed of a stack of identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by layer normalization. All sub-layers in the model, as well as the embedding layers, produce outputs of dimension 512.

4.1 Data Pre-processing

The HIPE dataset has three different levels of segmentation: article-level, line-level, and newspaper-level. Figure 3 shows an example of the segmentation proposed in the HIPE dataset.



Lo département de la # Savoie est considéré aussi comme # limitrop he pour Genève .
 Le département de la Savoie est considéré aussi comme limitrophe pour Genève .

Fig. 3: An example of a French instance from the training data. The upper sentence shows the provided input and the lower sentence contains no OCR errors. “#” represents the segmentation at line-level in historical newspapers. The arrows indicate the matching between the provided sentence and the correct sentence to highlight the OCR limitations.

Since BERT is able to consume only a limited context of tokens (512) and a line-level context would have been too short to grasp, we segment the articles at sentence level. We reconstructed the original text, including hyphenated words, using the miscellaneous annotated column that indicates if a word is split into two or more text lines. Then, the reconstructed text was passed through Freeling 4.1 [18] which determined the boundaries of each sentence.⁶

4.2 Parameters

For the German NERC, we chose as a pre-trained model the **bert-base-german-europeana**. This BERT model was trained using the open-source corpus Euro-

⁴ To increase the chances for misspelled, non-canonical, or new words to be recognized, we enrich the vocabulary of the tokenizer with these tokens, while allowing not only the BERT encoder but also the added Transformer layers to learn them from scratch.

⁵ In our implementation, we used *learned absolute positional embeddings* [8] instead, as suggested by [26]. [24] found that both versions produced nearly identical results.

⁶ It should be noted, that the segmentation using Freeling was not flawless. For instance, certain abbreviations were unknown by the tool. Thus, in some cases, Freeling oversegmented the sentences. Nonetheless, these errors were ignored.

peana newspapers⁷ [17]. It has been used in other NERC systems for contemporary and historical German texts [22, 21]. Moreover, it has shown an improvement with respect to other NERC systems.

For the French NERC, we relied on a pre-trained CamemBERT [15] model, specifically on the large version, **camembert-large**. Unlike BERT, this French version makes use of a whole-word masking and *SentencePiece* tokenization [11]. Additionally, for **camembert-large**, we found that fine-tuning was sometimes unstable on small datasets, so we ran several random restarts and selected the best model on the development set.

For the English NERC, since no training data was provided, we tackled the task with two approaches. The first one was to train the NERC using the English CoNLL 2003 dataset and the **bert-large-cased** model. The second approach was to use the German and French training data and the pre-trained multilingual BERT model, **bert-base-multilingual-cased**.

We denote the number of layers (i.e., Transformer blocks) as L , the hidden size as H , and the number of self-attention heads as A . **bert-base** has $L=12$, $H=768$, $A=12$, **bert-large** and **camembert-large**, $L=24$, $H=1024$, $A=16$. In all the cases, the top Transformer blocks have $L=1$ for $1 \times \text{Transf}$ and $L=2$ for $2 \times \text{Transf}$, $H=128$, $A=12$, chosen empirically.

The BERT-based encoders are fine-tuned on the task during training. For training, we followed the selection of parameters presented in [1]. We found that a 2×10^{-5} learning rate and a mini-batch of dimension 4 for German and English, and 2 for French, provide the most stable and consistent convergence across all experiments as evaluated on the development set.

4.3 Experiments

The experiments consider two configurations of our previously described model. The first one consists in using only the BERT encoder along with the CRF layers. The second configuration adds the Transformer blocks to the BERT encoder and the CRF layers. In Table 2, we present these experiments per language.

- RUN1: for German, French, and English, the models consist in only the fine-tuning BERT and the CRF layers, with the difference that, for English, we use the CoNLL dataset, and the fine-tuned BERT encoder is the English **bert-large-cased**
- RUN2: for German and French, the models consist in only the fine-tuning BERT, two stacked Transformer blocks, and the CRF layers, while for English, the model is **bert-large-cased**
- RUN3: for English, the model is the one used in RUN1, with the difference that the training data consists of the French and German training data, and the fine-tuned BERT encoder is **bert-base-multilingual-cased**

⁷ <http://www.europeana-newspapers.eu/>

Table 2: The NERC participating COARSE-LIT results for all runs.

Runs	Metrics	German			French			English		
		P	R	F1	P	R	F1	P	R	F1
RUN1	micro-fuzzy	0.838	0.886	0.861	0.909	0.926	0.917	0.775	0.797	0.786
	micro-strict	0.764	0.807	0.785	0.823	0.839	0.831	0.623	0.641	0.632
RUN2	micro-fuzzy	0.87	0.886	0.878	0.912	0.931	0.921	0.774	0.786	0.78
	micro-strict	0.79	0.805	0.797	0.831	0.849	0.84	0.621	0.63	0.625
RUN3	micro-fuzzy	–	–	–	–	–	–	0.794	0.817	0.806
	micro-strict	–	–	–	–	–	–	0.617	0.635	0.626

Table 3: The NERC participating results (all metrics) for the best performing run for each language.

Metrics	German			French			English		
	P	R	F1	P	R	F1	P	R	F1
COARSE-LIT									
micro-fuzzy	0.87	0.886	0.878	0.912	0.931	0.921	0.794	0.817	0.806
micro-strict	0.79	0.805	0.797	0.831	0.849	0.84	0.617	0.635	0.626
macro_doc-fuzzy	0.879	0.876	0.871	0.933	0.939	0.934	0.782	0.797	0.798
macro_doc-strict	0.782	0.781	0.777	0.852	0.859	0.854	0.635	0.64	0.644
COARSE-METO									
micro-fuzzy	0.626	0.78	0.694	0.676	0.67	0.673	1.0	0.12	0.214
micro-strict	0.571	0.712	0.634	0.658	0.652	0.655	0.667	0.08	0.143
macro_doc-fuzzy	0.558	0.678	0.686	0.628	0.732	0.718	1.0	0.075	0.533
macro_doc-strict	0.525	0.637	0.645	0.624	0.73	0.715	0.5	0.05	0.333
FINE-COMP									
micro-fuzzy	0.654	0.768	0.707	0.751	0.827	0.787	0	0	0
micro-strict	0.595	0.698	0.642	0.661	0.728	0.693	0	0	0
macro_doc-fuzzy	0.609	0.719	0.678	0.773	0.833	0.809	0	0	0
macro_doc-strict	0.559	0.649	0.618	0.703	0.757	0.735	0	0	0
FINE-LIT									
micro-fuzzy	0.734	0.813	0.771	0.843	0.869	0.856	0.733	0.817	0.773
micro-strict	0.629	0.697	0.661	0.772	0.797	0.784	0.547	0.61	0.577
macro_doc-fuzzy	0.754	0.813	0.776	0.871	0.883	0.875	0.742	0.798	0.774
macro_doc-strict	0.644	0.694	0.663	0.799	0.81	0.803	0.584	0.614	0.602
FINE-METO									
micro-fuzzy	0.659	0.771	0.711	0.626	0.688	0.655	1.0	0.16	0.276
micro-strict	0.601	0.703	0.648	0.618	0.679	0.647	0.75	0.12	0.207
macro_doc-fuzzy	0.595	0.659	0.705	0.558	0.7	0.687	1.0	0.108	0.522
macro_doc-strict	0.562	0.618	0.664	0.556	0.698	0.686	0.667	0.083	0.389
NESTED									
micro-fuzzy	0.588	0.411	0.484	0.366	0.415	0.389	0	0	0
micro-strict	0.49	0.342	0.403	0.333	0.378	0.354	0	0	0
macro_doc-fuzzy	0.339	0.326	0.413	0.502	0.484	0.521	0	0	0
macro_doc-strict	0.229	0.159	0.252	0.476	0.456	0.491	0	0	0

4.4 Results

From the results in Table 2, we can see the evidence that the BERT-based models with $n \times \text{Transf}$ achieve, for both German and French languages, higher fuzzy and strict performance values than the stand-alone BERT model.

For a more qualitative analysis, we examine the number of unrecognized words by the pre-trained BERT-based models that were added to the specific tokenizers (*WordPiece* for BERT and *SentencePiece* for CamemBERT). Following this observation, we notice that there is a tendency of performance increase of around 1 percentage F1 points for the $n \times \text{Transf}$ models (RUN2 for German and French). In Table 3, the highest values for all the coarse and fine metrics are presented.

In the case of English, when comparing RUN1 and RUN2, where the CoNLL 2003 dataset was used for training, with RUN3, where only HIPE German and French datasets were used, we notice that the F1 values are usually degraded by the use of modern datasets in the training process.

In summary, the methods that performed the best for the NERC task were the BERT-based models with n stacked Transformers for German and French. For English, the transfer learning from these two languages was clearly better than the models trained on modern English data.

5 Entity Linking (EL)

Regarding EL, in CLEF HIPE 2020, the task consists in the disambiguation of named entities using two settings:

- End-to-end EL: We do not have prior knowledge of the named entities. Thus, we rely on the information obtained from the NERC system.
- EL-only: We have access to the ground-truth regarding named entities, i.e. types and boundaries.

In both settings, it is necessary to take into account literal and metonymic senses. Furthermore, all the disambiguated named entities have to be linked to the Wikidata knowledge base (KB).

Our EL system is the composition and improvement of two EL approaches (Figure 4). First, we make use of the methodology proposed by [7] to create entity embeddings. Second, we utilize the EL architecture proposed by [10] to disambiguate the candidates. We have modified both EL approaches to support the multilingual aspect of the CLEF HIPE 2020 task.

More precisely, our approach consists of the following four steps which will be elaborated in the subsequent sections:

1. **Building resources:** the setup of a knowledge base per language.
2. **Entity embeddings:** the creation of entity feature representations based on the model proposed by [7].
3. **Entity disambiguation:** the main end-to-end EL model [10].
4. **Candidates filtering:** the post-processing step where several filtering techniques are proposed and studied.

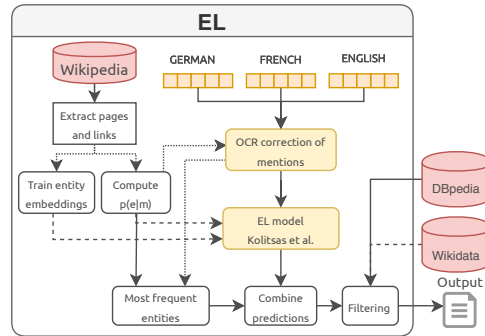


Fig. 4: The proposed model [10] for EL and the post-processing steps.

5.1 Building Resources

We build a KB for English, French, and German, in order to have a richer KB following these steps:

- Retrieve the last language version of the Wikipedia dump.
- Extract titles and ids of Wikipedia pages.
- Extract list of disambiguation pages and redirection pages.
- Calculate the probability entity-map $p(e|m)$ that analyzes how an entity e is related to a mention m based on the number of times that mention refers to that entity.

5.2 Entity Embeddings

We also build a dataset to train entity embeddings for each language, in which case, we use the methodology proposed by [7]. First, we generate two conditional probability distributions per language: the *positive distribution*, which is a probability approximation based on word-entity co-occurrence counts (i.e. which words appear in the context of an entity) and the *negative* one, which was calculated by randomly sampling context windows that were unrelated to a specific entity. Both probability distributions were used for word embeddings alignment with respect to an entity embedding. The *positive distribution* is expected to approach the embeddings of the co-occurring words with the embedding vector of the entity. While the negative probability distribution is used to distance the embeddings of words that are not related to an entity.

5.3 Entity Disambiguation

For the entity disambiguation, our model is based on Kolitsas *et al.*'s work [10], an end-to-end EL model that jointly performs entity linking and entity disambiguation. Besides the simplicity of the model brought by the joint-learning,

the model also takes advantage of the fact that it does not require complex engineered features.

First, for recognizing all entity mentions in a document, Kolitsas *et al.* proposed an empirical probabilistic entity-map⁸ $p(e|m)$ to analyze each span m and select top entities e that might be referred by this mention in $p(e|m)$.

The end-to-end EL model starts by encoding every token in the text input by concatenating word and character embeddings that are fed into a Bidirectional Long Short Term Memory (BiLSTM) network. This representation is used to project mentions of this document into a shared dimensional space with the same size as the entity embeddings. These embeddings are fixed continuous entity representations generated separately, namely in the same manner as presented in [7], and aforementioned in Section 5.2.

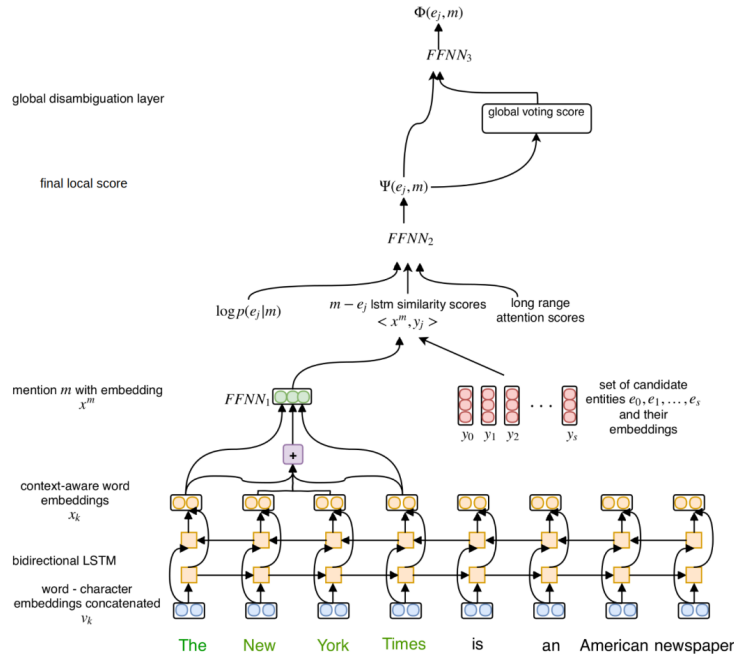


Fig. 5: Global model architecture for the mention “The New York Times”. The final score is used for both mention linking and entity disambiguation decisions (Kolitsas *et al.* [10]).

For analyzing long context dependencies of mentions, the authors used the attention model proposed by [7] that produces one context embedding per men-

⁸ Calculated from the Wikipedia corpora for each language.

tion based on informative context words that are related to at least one of the candidate entities. Next, the local score for each mention is determined by the combination of the $\log p(e|m)$, the similarity between the analyzed mention and each candidate entity embeddings, and the long-range context attention for the target mention. Finally, a top layer in the neural network promotes the coherence among disambiguated entities inside the same document. Additionally, we provide the five best candidate entities for a mention based on the probability entity-map $p(e|m)$.

5.4 Candidates Filtering

To improve the accuracy of the candidates provided by the EL system, we created a filtering tool based on heuristics and the DBpedia hierarchical structure [13].

Specifically, we used the DBpedia structure to manually specify subsets that represented each named entity type. For instance, the entity type location was associated with categories such as “dbo:Location” and “dbo:Settlement”.

These categories were used to determine whether a candidate provided by the EL system had to be positioned at the bottom of the rankings. In other words, candidates that according to DBpedia did not belong to the named entity type were positioned at the bottom of the ranking.⁹ For those candidates matching the named entity type¹⁰, we extracted their name in the language of analysis. This name was compared with the entity entry using Fuzzy Wuzzy Weighted Ratio¹¹. The most similar candidate to the entity entry was considered to be the most suitable candidate and was positioned at the top.

In the case of person-type entities, we requested to DBpedia their date of birth and extract the year if it was possible.¹² Then, we compared the extracted year of birth with the newspaper publication year, which was provided by the organizers, plus ten years more. If the person entity was born ten years after the publication of the newspaper, we removed completely the candidate.

Furthermore, we created a heuristic that consisted in adding NIL as the last possible candidate. This was done for each named entity unless the EL system proposed candidates with a type different from the named entity one. In this last case, a NIL was inserted between the different types of candidates. For example, if the location “Paris, France” had four candidates entries of type LOC, PERS, LOC, the filter would sort them as LOC, LOC, NIL, PERS. When the EL system proposed only candidates that were different from the named entity type, the filter would position on first place a NIL. These heuristics were based on the idea that if the EL system could not provide a candidate of the same type to the named entity, we might be dealing with an entity without an entry in Wikidata.

For RUN3 in the EL-only task, which will be described in Section 5.5, we proposed as well a filter based on DBpedia along with Wikidata. The reason is

⁹ This included candidates that could not be found in DBpedia as well.

¹⁰ In the case the literal and metonymic entities types were discordant, we considered both types as possible.

¹¹ github.com/seatgeek/fuzzywuzzy

¹² Certain person-type entities, such as music bands, do not have a date of birth.

that the former indexes only a subset of the latter. Thus, to improve the filter, we decided to use Wikidata as a backup knowledge base.

To access DBpedia¹³ and Wikidata¹⁴, we utilized their respective SPARQL Endpoint query service.

Table 4: EL results without prior knowledge of mention types and boundaries.

Runs	Metrics	English			French			German		
		P	R	F1	P	R	F1	P	R	F1
Literal										
RUN1	micro-strict	0.514	0.533	0.523	0.592	0.601	0.597	0.508	0.529	0.518
	micro-relaxed	0.514	0.533	0.523	0.612	0.621	0.617	0.53	0.552	0.541
RUN2	micro-strict	0.496	0.506	0.501	0.592	0.602	0.597	0.531	0.538	0.534
	micro-relaxed	0.496	0.506	0.501	0.612	0.622	0.617	0.553	0.561	0.557
RUN3	micro-strict	0.523	0.539	0.531	0.594	0.602	0.598	0.502	0.528	0.515
	micro-relaxed	0.523	0.539	0.531	0.613	0.622	0.617	0.524	0.55	0.537
Metonymic										
RUN1	micro-strict	0.172	0.2	0.185	0.236	0.402	0.297	0.324	0.508	0.396
	micro-relaxed	0.172	0.2	0.185	0.366	0.625	0.462	0.384	0.602	0.469
RUN2	micro-strict	0.062	0.04	0.049	0.217	0.339	0.265	0.324	0.508	0.396
	micro-relaxed	0.062	0.04	0.049	0.343	0.536	0.418	0.384	0.602	0.469
RUN3	micro-strict	0.059	0.04	0.048	0.236	0.402	0.297	0.308	0.508	0.383
	micro-relaxed	0.059	0.04	0.048	0.366	0.625	0.462	0.364	0.602	0.454

5.5 Experiments

Both entity embeddings and the end-to-end EL method used the pre-trained multilingual MUSE¹⁵ word embeddings of size 300 for all languages in the dataset. We chose the size of 50 for the character embeddings. The German and French models were trained on the HIPE split (Table 1). As the HIPE dataset does not contain training data for English, we trained our English model on the AIDA dataset [9].

In order to overcome or reduce OCR problems, we analyzed several mention variations in order to improve the matching with candidates within the probability entity-map. More precisely, we analyze the following variations: concatenation, lowercase, no punctuation, and the Levenshtein distance between a mention and all candidate mentions within the probability table. In the metonymic sense, the approach used was to annotate the corpus consisted in copying the candidates used for the literal sense.

We implemented three configurations of our EL approach for the EL-only task:

¹³ wiki.dbpedia.org/public-sparql-endpoint

¹⁴ query.wikidata.org

¹⁵ <https://github.com/facebookresearch/MUSE>

- RUN1: for German, French, and English, the output is composed of the candidate entities proposed by [10].
- RUN2: for German, French, and English, the output is composed of the five most frequent candidate entities related to a mention.
- RUN3: for German, French, and English, the output is composed of the candidate entities proposed by [10] and the ten most frequent candidate entities related to a mention. For this run, the filter used not only information from DBpedia but also from Wikidata as indicated in Section 5.4.

We also made three configurations of our end-to-end NERC-EL architecture to recognize and disambiguate entities:

- RUN1: for German, French, and English, the output is composed of entities of NERC RUN1 and the disambiguation method of EL RUN1.
- RUN2: for German, French, and English, the output is composed of entities of NERC RUN2 and the disambiguation method of EL RUN1.
- RUN3: for German and French, the output is composed of entities of NERC RUN1 and the disambiguation method of EL RUN2. For English, the output is composed of entities of NERC RUN3 and the disambiguation method of EL RUN1.

All runs analyze the mention variations and use the filter to select the best five candidate entities among all selected candidate entities by each run.

Table 5: EL results with prior knowledge of mention types and boundaries.

Runs	Metrics	English			French			German		
		P	R	F1	P	R	F1	P	R	F1
Literal										
RUN1	micro-strict	0.593	0.593	0.593	0.64	0.638	0.639	0.565	0.564	0.565
	micro-relaxed	0.593	0.593	0.593	0.66	0.657	0.659	0.588	0.587	0.587
RUN2	micro-strict	0.593	0.593	0.593	0.635	0.632	0.633	0.564	0.563	0.564
	micro-relaxed	0.593	0.593	0.593	0.654	0.652	0.653	0.587	0.586	0.586
RUN3	micro-strict	0.58	0.58	0.58	0.633	0.63	0.632	0.581	0.582	0.582
	micro-relaxed	0.58	0.58	0.58	0.653	0.65	0.652	0.601	0.602	0.602
Metonymic										
RUN1	micro-strict	0.286	0.48	0.358	0.303	0.446	0.361	0.443	0.627	0.519
	micro-relaxed	0.286	0.48	0.358	0.461	0.679	0.549	0.515	0.729	0.604
RUN2	micro-strict	0.286	0.48	0.358	0.303	0.446	0.361	0.443	0.627	0.519
	micro-relaxed	0.286	0.48	0.358	0.461	0.679	0.549	0.515	0.729	0.604
RUN3	micro-strict	0.286	0.48	0.358	0.297	0.438	0.354	0.431	0.61	0.505
	micro-relaxed	0.286	0.48	0.358	0.455	0.67	0.542	0.485	0.686	0.568

5.6 Results

For the EL without prior knowledge of mention types and boundaries, our EL approach depends on the performance of our NERC system to recognize and classify the type of entities in historical documents. The results on all the languages

are presented in Table 4. While RUN1 achieved the best results for metonymic, RUN3 outperformed the other configurations on the literal analysis.

For EL with prior knowledge of mention types and boundaries, our system has access to the ground-truth of NERC entities, i.e. correct span and NERC type for all mentions. Table 5 shows the results. As expected, our EL system achieved better results with the ground-truth information (improvement up to 0.09 and 0.31 in the F1 values for literal and metonymic, respectively). All runs achieved similar results for all languages, with the RUN1 being slightly superior to the other runs for literal and metonymic analysis.

The use of the filter based on DBpedia and Wikidata reduced the performance of the EL system in English and French. This might be due to the increment of noise, such as names of disambiguation pages.¹⁶ Our filter analyses all candidate entities for each mention to order the list of candidates based on their NERC types and names. Since RUN1 and RUN2 provide up to five candidate entities for each mention, these runs are more likely than RUN3 to provide a NIL entry for a mention. For RUN3, the filtering process has a higher probability to find a candidate of the same named entity type as the mention and disambiguates this mention to a less frequent candidate entity in a KB.

6 Conclusions

For the participation of our team (L3i) to the HIPE lab at CLEF 2020, we proposed two neural-based methods for the tasks of NERC and EL. We conclude, for NERC, that the proposed models generally performed well, and that the stacked transformer-based model with a BERT fine-tuned model and additional transformer layers better learned the characteristics of the HIPE historical dataset.

For EL, our neural model combined with the filtering process analyzed the historical mentions and disambiguated them to the Wikidata KB. Combining information from Wikipedia, Wikidata, and DBpedia allowed a thorough analysis of the characteristics of the entities and helped our method to correctly disambiguate mentions in historical documents.

¹⁶ DBpedia does not index disambiguation pages.

Bibliography

- [1] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019)
- [2] Duncan, R.: Advanced MS-DOS Programming. Microsoft Press Redmond, WA (1988)
- [3] Ehrmann, M., Romanello, M., Bircher, S., Clematide, S.: Introducing the CLEF 2020 HIPE shared task: Named entity recognition and linking on historical newspapers. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in information retrieval. pp. 524–532. Springer International Publishing, Cham (2020)
- [4] Ehrmann, M., Romanello, M., Clematide, S., Ströbel, P.B., Barman, R.: Language resources for historical newspapers: the impresso collection. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 958–968 (2020)
- [5] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In: Cappellato, L., Eickhoff, C., Ferro, N., Névél, A. (eds.) CLEF 2020 Working Notes. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum. CEUR-WS (2020)
- [6] Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Impresso named entity annotation guidelines (version 2.2.0). <https://doi.org/10.5281/zenodo.3604227> (2020)
- [7] Ganea, O.E., Hofmann, T.: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2619–2629 (2017)
- [8] Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122 (2017)
- [9] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 782–792. Edinburgh, Scotland, UK. (2011)
- [10] Kolitsas, N., Ganea, O.E., Hofmann, T.: End-to-end neural entity linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. pp. 519–529 (2018)
- [11] Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)
- [12] Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)

- [13] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Kleef, P.v., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia **6**(2), 167–195. <https://doi.org/10.3233/SW-140134>
- [14] Makhoul, J., Kubala, F., Schwartz, R., Weischedel, R., et al.: Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop. pp. 249–252. Herndon, VA (1999)
- [15] Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 (2019)
- [16] Muller, B., Sagot, B., Seddah, D.: Enhancing bert for lexical normalization. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). pp. 297–306 (2019)
- [17] Neudecker, C.: An open corpus for named entity recognition in historic newspapers. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). pp. 4348–4352. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), <https://www.aclweb.org/anthology/L16-1689>
- [18] Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12). pp. 2473–2479. ELRA, Istanbul, Turkey (May 2012)
- [19] Pruthi, D., Dhingra, B., Lipton, Z.C.: Combating adversarial misspellings with robust word recognition. In: 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). pp. 5582–5591. Florence, Italy (2019)
- [20] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
- [21] Riedl, M., Padó, S.: A named entity recognition shootout for german. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 120–125 (2018)
- [22] Schweter, S., Baiter, J.: Towards robust named entity recognition for historic german. arXiv preprint arXiv:1906.07592 (2019)
- [23] Sun, L., Hashimoto, K., Yin, W., Asai, A., Li, J., Yu, P., Xiong, C.: Advbert: Bert is not robust on misspellings! generating nature adversarial samples on bert. arXiv preprint arXiv:2003.04985 (2020)
- [24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
- [25] Weber, J.: Strassburg, 1605: The origins of the newspaper in europe. German history **24**(3), 387–412 (2006)
- [26] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: Huggingface’s transformers: State-of-the-art natural language processing. ArXiv **abs/1910.03771** (2019)

- [27] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)

15 TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data

TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data

Jose G. Moreno
jose.moreno@irit.fr
University of Toulouse - IRIT
F-31000, Toulouse, France

Emanuela Boros
emanuela.boros@univ-lr.fr
University of La Rochelle - L3i
F-17000, La Rochelle, France

Antoine Doucet
antoine.doucet@univ-lr.fr
University of La Rochelle - L3i
F-17000, La Rochelle, France

ABSTRACT

This paper presents the TLR participation in the FinNum-2 task. Our system is based on a Transformer architecture improved by a pre-processing strategy for numeral attachment identification. Instead of relying on a vanilla attention mechanism, we focus the attention to specific tokens that are essential for the task. The results in an unseen test collection show that our model correctly generalises the predictions as our best run outperforms all those of other participants in terms of F1-macro (official metric). Further, results show the robustness of our method as well as the experiments with two alternatives (with and without parameter tuning) leading to an additional improvement of 4% over our best run.

ACM Reference Format:

Jose G. Moreno, Emanuela Boros, and Antoine Doucet. 2018. TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/1122445.1122456>

TEAM NAME

TLR

SUBTASKS

Numeral attachment in financial tweets (English)

1 INTRODUCTION

Social media platforms are becoming a main source of information nowadays [5]. News media, politicians, personalities, etc., use microblogs such as Twitter daily to briefly communicate with their target public (followers). As an example, the current President of the United States of America¹ publishes an average of seven to ten tweets per day, approximately totalling 3,000 tweets per year to an audience of 86 millions of followers [9, 10].

However, politicians are not alone in the use of social media. Companies also use social media to publish information about their current and new products, successful histories, or information to their shareholders, including financial information. Similarly,

¹These statistics are based on the @realDonaldTrump Twitter account.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://www.acm.org/permissions).

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

shareholders or the general public also share financial information on social media through the use of special identifiers [2]. Indeed, social media platforms use some special characters to allow users to tag information. One of the commonest is the hashtag token (#) that is used to tag conversations in Twitter [11]. Another special character is the at symbol (@) that is usually used to mention users of the platform. The main difference is that the latter refers to unique elements while the former may be related to ambiguous topics. A similar token of the latter group is the dollar symbol (\$) used on the financial-related information. It is widely promoted by Cashtag² platform. This platform is based on the use of \$cashtags that are unique identifiers for individuals and businesses using Cash App. The \$Cashtags allow an aggregation of the information related to a unique organisation as they are used as identifiers.

The use of \$Cashtags opens promote the exploration of current challenges and techniques in information extraction (IE) applied to the financial domain. In this context, multiple natural language processing (NLP) tasks [2–4, 7] can be addressed automatically to improve user experience when using \$cashtags or to mine vital information from their use. This includes named entity recognition and linking, relation extraction and classification, or numeral attachment identification and classification, to mention a few of tasks that may be associated with the use of \$cashtags. During its 15th edition, NTCIR hosted the numeral attachment challenge, where participants are asked to automatically identify the relatedness of numeral information and \$cashtags within financial tweets. However, several recent models in IE tend to give hardly any attention to this type of information. FinNum-1 [3] and FinNum-2 [4] addressed the problem of the understanding of numbers in financial information, where fine-grained numeral understanding in financial social media data is essential to link \$Cashtags and numeral data.

In this paper, we present the TLR participation in the FinNum-2 task. Our system is based on recent architectures based on neural language models and a simple but effective explicit attention mechanism. Our best official run outperforms other participants on the task with a significant margin. Moreover, improvements over our own runs can be obtained by the use of an ensemble strategy but on a larger number of predictions.

The remainder of this paper is organised as follows. Section 2 presents the background information related to the task, the works that inspired our model and the details of our models. The experimental setup, our official results, and complementary experiments are elaborated in Section 4. Finally, the conclusions are drawn in Section 5.

²<https://cash.app/>

2 BACKGROUND

In this section, we briefly introduce recent neural models based on Transformers [14], such as BERT and RoBERTa, and their use for relation extraction (RE). Our intuition is that the RE task is closely related to the numeral attachment task.

2.1 Neural-based Language Models

Given the strong performance of recent deep architectures trained on variants of language modelling, we chose BERT [6] and RoBERTa [12] models. These architectures have been successfully evaluated in a wide number of NLP tasks [1, 6, 13, 15, 16].

Both use the same architecture based on several layers of the Transformer blocks. A Transformer block [14] is a deep learning architecture based on multi-head attention mechanisms with sinusoidal position embeddings³. It is composed of a stack of identical layers, each layer having two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network. A residual connection is around each of the two sub-layers, followed by a layer normalisation. All sub-layers in the model, as well as the embedding layers, produce outputs of dimension 512.

The RoBERTa model is based on different modifications of the BERT pre-training procedure that improve end-task performance.

A binary text classification based on the Transformer (either BERT or RoBERTa) is depicted in Figure 1.

2.2 Relation Extraction with Transformers

Extracting relations between tokens in sentences is a challenging NLP task. However, a recent work on relation classification Baldini Soares et al. [1] showed that vanilla Transformer-based⁴ sequence classifiers are strong enough to identify relations. This is achieved by the introduction of additional markers that help the model to drive their attention mechanics. Indeed, Transformer-based models are already strong to classify sentences. However, it may struggle when the same entry is considered for multiple classes (positive and negative). To address this problem, Baldini Soares et al. [1] proposed a pre-processing step that is required to indicate entity tokens by using extra tokens in the input sentence. Then, the typical sequence classification strategy proposed by Devlin et al. [6] can be used. This strategy consists of using a feed-forward layer that takes the [CLS] token representation and that is trained to perform the classification task. This simple but powerful architecture is privileged in our work.

3 NUMERAL ATTACHMENT IN FINANCIAL TWEETS

Although detailed information regarding the task can be found in Chen et al. [4], we briefly describe hereafter the task.

³In practice, these models use absolute positional embeddings [8] instead, as a common practice.

⁴The BERT model is used in [1].

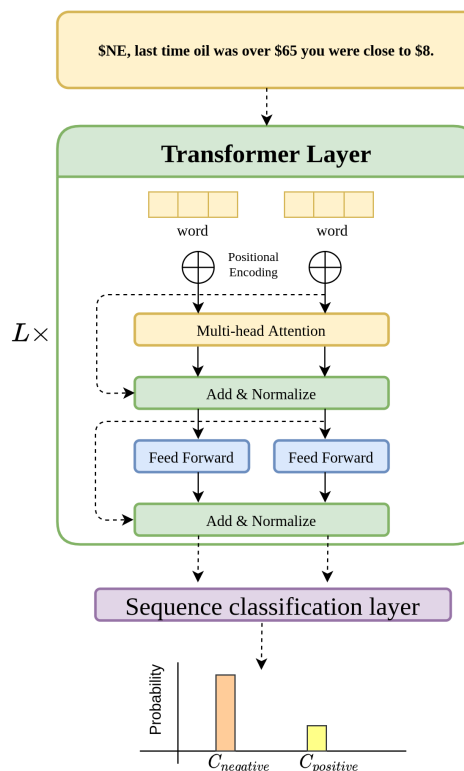


Figure 1: BERT architecture for binary text classification. L is the number of Transformer layers. In our experiments, the BERT and RoBERTa models have 12 Transformer blocks.

3.1 Task Definition

Given a tweet (in text format), a *\$cashtag* offset, and a numeral offset, the FinNum-2 task consists in determining whether a numeral indicated by the numeral offset relates or not to the *\$cashtag* indicated by the *\$cashtag* offset. Two examples of this task are presented in Figure 2 (a), where the \$NE token is positively attached to \$8 and negatively to \$65. Note that in this case the same sentence is associated with two examples depending on the attached numeral. Thus, this is a binary classification task. In the context of the NTCIR, three files are shared by the organisers depending on the task stage. In particular, the train and development sets are provided with labels at the beginning of the task while the test set is provided without labels. Finally, after the official results are published, the labels of the test set are shared with all participants.

3.2 Numeral Attachment Classification

We opted for a Transformer neural-based language model (as described in Section 2) and we introduced a pre-processing technique for the numeral attachment task. In our case, we mainly focus on

the input preparation to facilitate the system identification of the key information for the task.

Regarding the necessary pre-processing step, we first add two reserved words to mark the beginning and the end of the *\$cashtag* mentioned in the text. We introduce the £ and \$ additional reserved tokens, and we mark the words concerned in the sequence. Figure 2 presents the (a) initial input provided by the organisers, (b) the transformed information for our system. Note that the transformed information transcribed in the text formats only the tokens concerned in the classification (Figure 2 (c)).

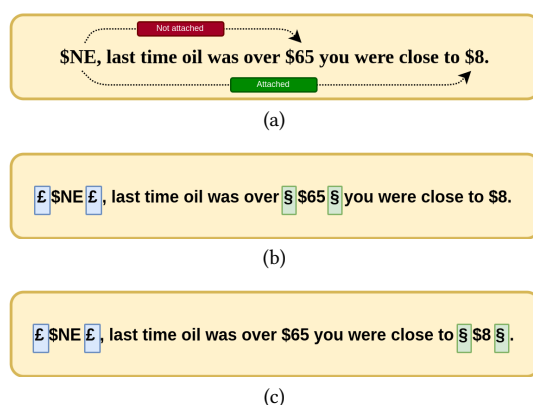


Figure 2: Input pre-processing step with marked key information.

Our predictions for each model are based on the output probability obtained by the model as depicted in Figure 1. Finally, we perform an ensemble strategy based on a *max* or *min* selection to define the final class prediction.

4 EXPERIMENTS AND RESULTS

4.1 Dataset

A manually annotated dataset was provided by the organisers of the FinNum-2 task. This dataset is composed of 7,187 training examples, 1,044 validation examples, and 2,109 test examples. The details of the dataset and the manual annotation process can be found in Chen et al. [2, 4].

4.2 Metrics

The official metric of the task is the F1-macro computed as the harmonic mean between precision and recall.

4.3 Analysis on the Validation Partition

Despite the multiple parameters involved in the architecture based on Transformer layers, we opted for a standard configuration of the models as shown in Table 1.

The main parameter that was selected using the validation partition is related to the number of epochs used to train the model. We explored a total of 20 epochs and selected the model with the highest validation performance between the epochs. Results for BERT

Table 1: Parameters used for our BERT and RoBERTa models.

Name	Value
Weight Initialisation	BERT-base / RoBERTa-base
Batch Size	32
Optimiser	Adam
Learning Rate	3×10^{-5}
Epsilon	10^{-8}
Clipnorm	1
Loss	Sparse Binary Crossentropy

and RoBERTa models are presented in Figure 3. Best performances are obtained in epoch 2 and 4 for BERT and RoBERTa, respectively. From Figure 3, we can also see that (1) for both models, the choice of three epochs seems an inadequate option so this shows the relevance of this parameter, (2) later epochs (after 15) the performances between BERT and RoBERTa are indiscernible.

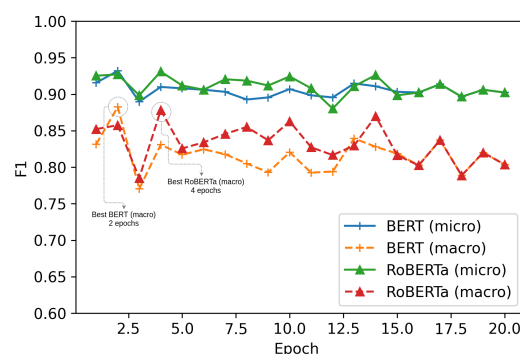


Figure 3: Performance of the model over the validation partition.

Additionally, our ensemble strategy consists in selecting the *max* or *min* function as the predictor for our last run. Results are presented on Table 2 for these two functions. In the validation set, the BERT model outperformed the RoBERTa model while both models are outperformed by the *min* function. Thus, these three models were selected for our participation on the task⁵.

4.4 Official Results

Our team submitted three runs that were calculated as follow:

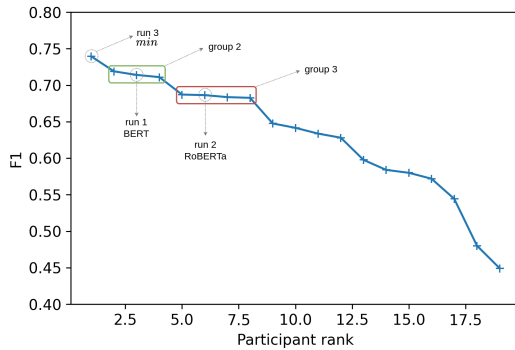
- Run 1: It consists in the BERT predictions of our model.
- Run 2: It consists in the RoBERTa predictions of our model.
- Run 3: It consists in the *min* between the BERT and RoBERTa predictions of our model.

⁵We discarded *max* function as each participation was limited to three runs and it showed the lowest performance.

Table 2: Results using the validation partition for *min* and *max* functions.

	Precision	Recall	F1
BERT	0.9013	0.8667	0.8826
RoBERTa	0.9088	0.8542	0.8781
<i>min</i>	0.8850	0.8926	0.8887
<i>max</i>	0.9385	0.8284	0.8704

The official results ordered by run performance (rank) are presented in Figure 4. Our *run 3 (min)* outperformed all other participants of the task while *run 1 (BERT)* and *run 2 (RoBERTa)* achieved the 3rd and 6th position in terms of F1-macro, respectively. We can note that our *run 2* is part of a group of runs denoted *group 2*. The runs in the *group 2* have very similar performance values suggesting that there are few chances of observing statistical differences between them⁶. A similar situation can be observed in the case of *group 3*. Indeed, after *group 3* it seems to exist more variation between runs. Based on the observation of low intra variance within *groups 2* and *3*, we intuit that combination of any couple of runs from the two groups may deal with similar improvements. This extra exploration is studied in section 4.5.2.

**Figure 4: F1-macro performances ordered by participant rank. Our three runs are identified by dotted circles.**

4.5 Unofficial Results

In order to understand the improvement of our model, we perform additional experiments that were not submitted as official runs.

4.5.1 Impact of *min* ensemble. The combination of our BERT and RoBERTa models was successful when using the *min* function. It clearly outperformed the *group 2* results (second best). Moreover, the *min* function can be interpreted as a higher threshold strategy as a positive classified example must be considered positive by our BERT model as well as our RoBERTa model. This can be individually analysed by making it harder for each system to predict an example as positive. So, instead of considering all examples with a

⁶This is arguable because it depends on the scale. However, it seems fair to assume.

probability greater than 0.5 for the positive class, we variate this threshold⁷ and presented the results in Table 3. Note that most of the F1 performances increased as the threshold value is higher. However, our BERT and the *min*-based strategy models achieve their best performances at 0.7 and decrease after that. This result suggests that this parameter must be carefully tuned.

Table 3: Results of our runs using the test partition. Threshold for the positive class was increased from 0.5 (official runs) to 0.9.

		Precision	Recall	F1
0.5	BERT	0.8015	0.6793	0.7141
	RoBERTa	0.8435	0.6484	0.6864
	<i>min</i>	0.8016	0.7078	0.7395
0.6	BERT	0.7938	0.7184	0.7461
	RoBERTa	0.8301	0.6612	0.6996
	<i>min</i>	0.7866	0.7387	0.7585
0.7	BERT	0.7731	0.7478	0.7592
	RoBERTa	0.8114	0.6716	0.7081
	<i>min</i>	0.7691	0.7632	0.7661
0.8	BERT	0.7445	0.7648	0.7538
	RoBERTa	0.7796	0.6844	0.7147
	<i>min</i>	0.7380	0.7733	0.7528
0.9	BERT	0.7228	0.7841	0.7438
	RoBERTa	0.7646	0.7115	0.7324
	<i>min</i>	0.7169	0.7896	0.7388

4.5.2 Baselines and extra ensemble combinations. Following the same configuration setup and parameters, we train multiple extra models to better understand the real improvements of our models (when compared against original models) and a further understanding of the ensemble *min* function:

- A vanilla BERT with any input modification.
- A vanilla RoBERTa with any input modification.
- A *min* ensemble based on our BERT model (three models).
- A *min* ensemble based on our RoBERTa model (three models).
- A *min* between the ensemble *min* of the BERT and RoBERTa predictions.

Table 4: Unofficial results of our models and baselines using the test partition. We ran three times our models instead of only ones and applied the *min* ensemble function. Parameters remain unchanged w.r.t. our official runs.

		Precision	Recall	F1
Baselines (vanilla models)	BERT	0.8134	0.6638	0.7004
	RoBERTa	0.9182	0.6041	0.6313
<i>min</i> Ensemble (n=3) (our models)	BERT	0.7933	0.6949	0.7267
	RoBERTa	0.8204	0.7090	0.7447
	<i>min</i>	0.7964	0.7489	0.7688

⁷between 0.5 and 0.9

NTCIR 15 Conference: Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, December 8-11, 2020 Tokyo Japan

TLR at the NTCIR-15 FinNum-2 Task: Improving Text Classifiers for Numeral Attachment in Financial Social Data

Woodstock '18, June 03–05, 2018, Woodstock, NY

Results for the validation and test partitions are presented in Table 4. Note that aggregating ensemble methods have a beneficial effect in our model. Indeed, the *min* plus *min* function outperforms all the models including our submissions. Note that this result is obtained without extra parameters nor any special tuning. Despite these positive results, we strongly believe that there is still room for improvement as our model is based on a standard sequence classification strategy and more elaborated representation may be included in the model by using not only the [CLS] representation but also the representation of the £ and \$ tokens.

5 CONCLUSIONS

This paper presents our participation in the FinNum-2 task at NTCIR-15. The proposed model is based on an information extraction strategy that combines Transformer-based models with positional information. Our main finding is that this representation is relevant for the task of numeral attachment identification. Our best run achieved the top performance in the F1-macro, the official metric. As future work, we intend to evaluate the quality of the proposed model into other financial tasks such as fine-grained numeral understanding [3].

ACKNOWLEDGEMENTS

This work has been partly supported by the European Union's Horizon 2020 research and innovation programme under grant 825153 (EMBEDIA).

REFERENCES

- [1] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *ACL*. 2895–2905.
- [2] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Numeral attachment with auxiliary tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1161–1164.
- [3] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Overview of the NTCIR-14 FinNum Task Fine-Grained Numeral Understanding in Financial Social Media Data. In *Proceedings of the 14th NII Testbeds and Community for Information Access Research (NTCIR-14) Conference (NTCIR-14)*.
- [4] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 task: Numeral Attachment in Financial Tweets. In *Proceedings of the 15th NII Testbeds and Community for Information Access Research (NTCIR-15) Conference (NTCIR-15)*.
- [5] Mary J Culnan, Patrick J McHugh, and Jesus I Zubillaga. 2010. How large US companies can use Twitter and other social media to gain business value. *MIS Quarterly Executive* 9, 4 (2010).
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] Ismail El Maarouf, Youness Mansar, Virginie Moulleron, Dialekti Valsamou-Stanislawski, and Fortia Financial Solutions. 2020. The finsim 2020 shared task: Learning semantic representations for the financial domain. In *The Second Workshop on Financial Technology and Natural Language Processing in conjunction with IJCAI-PRICAI 2020*. 81.
- [8] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (Sydney, NSW, Australia) (ICML '17)*. JMLR.org, 1243–1252.
- [9] Panayota Gounari. 2018. Authoritarianism, discourse and social media: Trump as the 'American agitator'. *Critical theory and authoritarian populism* (2018), 207–227.
- [10] Wendy Hall, Ramine Tinati, and Will Jennings. 2018. From Brexit to Trump: Social media's role in democracy. *Computer* 51, 1 (2018), 18–27.
- [11] Jeff Huang, Katherine M Thornton, and Efthimis N Efthimiadis. 2010. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. 173–178.
- [12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [13] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).
- [14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [15] Yu Wang, Yining Sun, Zuchang Ma, Lisheng Gao, Yang Xu, and Ting Sun. 2020. Application of Pre-training Models in Named Entity Recognition. *arXiv preprint arXiv:2002.08902* (2020).
- [16] Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2361–2364.