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

D2.2: Initial cross-lingual semantic enrichment technology (T2.1)

Executive summary

The present report describes the initial version of cross-lingual systems for semantic enrichment of Task T2.1 of WP2. The report presents our cross-lingual systems to address the tasks of named entity recognition, named entity linking and event detection. We combined language-dependent and language-independent features to reduce the dependency on a specific language and be able to recognize named entities in several languages. Additionally, we use multilingual word embeddings to project entities and words from several languages into the same dimensional space. Then, we link mentions of source documents in Croatian, Estonian, Finnish, and Slovenian to the English Wikipedia. Further developments will be presented in a second and final version of this report to be delivered at M24 as Deliverable D2.5.

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:	**D2.2**
Deliverable full title:	**Initial cross-lingual semantic enrichment technology**
Deliverable short title:	**Initial cross-lingual semantic enrichment technology**
Document identifier:	**EMBEDDIA-D22-InitialCrosslingualSemanticEnrichmentTechnology-T21-submitted**
Lead partner short name:	**ULR**
Report version:	**submitted**
Report submission date:	**31/12/2019**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Elvys Linhares Pontes (ULR)**
Co-author(s):	**Jose G Moreno (ULR), Antoine Doucet (ULR)**
Status:	**__ draft, __ final, _x_ 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
23/10/2019	v0.1	Elvys Linhares Pontes (ULR)	Initial version.
05/11/2019	v0.2	Jose G Moreno (ULR)	NER sections were added.
07/11/2019	v0.3	Elvys Linhares Pontes (ULR)	NEL sections were added.
14/11/2019	v0.4	Nicolas Sidere (ULR)	Review and comments.
28/11/2019	v0.45	Hannu Toivonen (UH)	Review and comments.
03/12/2019	v0.5	Elvys Linhares Pontes (ULR)	Corrections.
06/12/2019	v0.6	Jose G Moreno (ULR)	Corrections.
08/12/2019	v0.7	Elvys Linhares Pontes (ULR)	Corrections.
12/12/2019	v0.8	Nada Lavrač (JSI)	Quality control.
21/12/2019	final	Antoine Doucet (ULR)	Final corrections towards submission.
23/12/2019	submitted	Tina Anžič (JSI)	Report submitted.

Table of Contents

List of abbreviations

BILSTM	Bidirectional Long-Short Term Memory
CRF	Conditional Random Fields
ED	Entity Disambiguation
GH	Ganea and Hofmann
KB	Knowledge Base
LSTM	Long-Short Term Memory
MAP	Mean Average Precision
NDCG	Normalised Discounted Cumulative Gain
NE	Named Entity
NEL	Named Entity Linking
NER	Named Entity Recognition
NLP	Natural Language Processing
RNN	Recurrent Neural Networks

1 Introduction

The overall objective of WP2 is the embeddings-based semantic enrichment of individual documents and their content, to be achieved by performing multi- and cross-lingual named-entity recognition and disambiguation and linking the recognized 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 to the project, interacting notably with Task T2.2 on multilingual keyword extraction and matching, and being evaluated as defined in Task T2.4.

Delivered at M12, the present document entitled 'initial cross-lingual semantic enrichment' and the corresponding source code are composing D2.2, the first of two deliverables in Task T2.1 of WP2. The second deliverable is D2.5, to be delivered and shared publicly at M24.

Central to Task T2.1, named entities (NE) are real-world objects, such as persons, locations, organizations, etc. They are important concepts as they often are key descriptors of what a text is about. 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', 'organisation', etc. In this deliverable, we further experimented with the recognition of a NE category 'event', used to perform event detection. 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 such as DBpedia.

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

- For NER, we defined a new cross-lingual approach mixing language-dependent and language-independent features in order to reduce the dependency to a specific language, described in Section 2 and in the appended paper by Moreno et al. (2019), published in the proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing held at ACL 2019.
- Event detection, which is part of Task 2.1, is currently processed by this NER approach, where events are considered as a specific type of named entities. This is intended as an initial baseline, while more advanced approaches for event detection are under development and will be presented at M24.
- For NEL, we added to the well-known model of Ganea & Hofmann (2017) the ability to perform cross-lingual analysis. More precisely, we projected words and entities in several languages into the same dimensional space. Then we used these representations to train the Ganea & Hofmann approach for the monolingual NEL. Finally, we optimised this model on cross-lingual data sets to adapt it to the cross-lingual NEL task (i.e. linking mentions of source documents in Croatian, Estonian, Finnish, and Slovenian to an English Knowledge Base).

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 organized as follows: Section 2 presents our cross-lingual NER approach and its performance on the EMBEDDIA languages over several NE categories including events. Section 3 describes our work on NEL. Section 4 lists the availability of the resources produced in this deliverable. Then, conclusions are set out in Section 5. Finally, the appendix includes our recent paper published at 7th Workshop on Balto-Slavic Natural Language Processing (Moreno et al., 2019).

2 Named Entity Recognition

Named Entity Recognition (NER) aims to extract and type mentions of named entities from raw texts, e.g., identify sequences of words—the mention of an entity—that refer to an entity and assign each one an entity type. In this project, we are interested on some types of entities: person (PER), organization (ORG), location (LOC), miscellaneous (MISC), and events (EVT); the latter type is used in our initial event detection system.

2.1 NER Approach

Our approach is based on recent advances in deep neural methods for NLP. We focus on methods that do not use specialised handcrafted features as they are hard to collect for low-resourced languages. In this context, we opted to ground our method on the LSTM-CNNs-CRF method proposed by Ma & Hovy (2016). In particular, we make an extra effort to include strong and multilingual representation models[1].

This section describes our model which is based on a standard end-to-end architecture for sequence labelling, namely LSTM-CNNs-CRF (Ma & Hovy, 2016). Each level of our model (FastText, Case Encoding, multiBERT, Char representation, BILSTM, and CRF) is developed in the following subsections. We have combined this architecture with contextual embeddings using a weighted average strategy (Reimers & Gurevych, 2019) applied to a pre-trained model for multiple languages (Devlin et al., 2019) (including all EMBEDDIA languages, e.g., Croatian, Estonian, Finnish, and Slovenian, as well as non-EMBEDDIA languages such as Czech, Polish, Russian, Slovak, Ukrainian, and Bulgarian). We trained a NER model per language. As an example, the overall architecture of our model for Polish using the sentence:

"Wielka Brytania z zadowoleniem przyjęła porozumienie z Unią Europejską"

(or "United Kingdom welcomes agreement with the European Union" in English) is depicted in Figure 1. Note that our input is composed of multiple features including the FastText embeddings and the language-indepent features. Then, the input is processed by the BILSTM and the CRF layers to produce the output indicated on the upper circles, where each token has a related type. In our example, the two first tokens ("United Kingdom") where predicted as the LOC type indicating that those tokens are a mention of a location.

For training, we follow a classical strategy for sequence labelling. It means that weights of BILSTM and CRF layers are learned during training time where an annotation collection is processed by the model. Similarly, during testing time the weights are frizzed[2] and the output of the upper layers correspond to the prediction fo the model.

2.1.1 FastText Embedding

In this layer, we used pre-trained embeddings for each language trained on Common Crawl and Wikipedia using fastText (Bojanowski et al., 2017; Grave et al., 2018). These models were trained using the continuous bag-of-words (CBOW) strategy with position weights. A total of 300 dimensions were used with character n-grams of length 5, a window of size 5 and 10 negatives. All the EMBEDDIA languages are included in this publicly available[3] pre-trained embedding (Grave et al., 2018). Besides, the fastText library provides a corresponding vector for every token (also in other alphabets) which avoids out-of-vocabulary tokens.

[1] They were not available at the time that LSTM-CNNs-CRF was proposed.
[2] They are not updated any more.
[3] `https://fasttext.cc/docs/en/crawl-vectors.html`

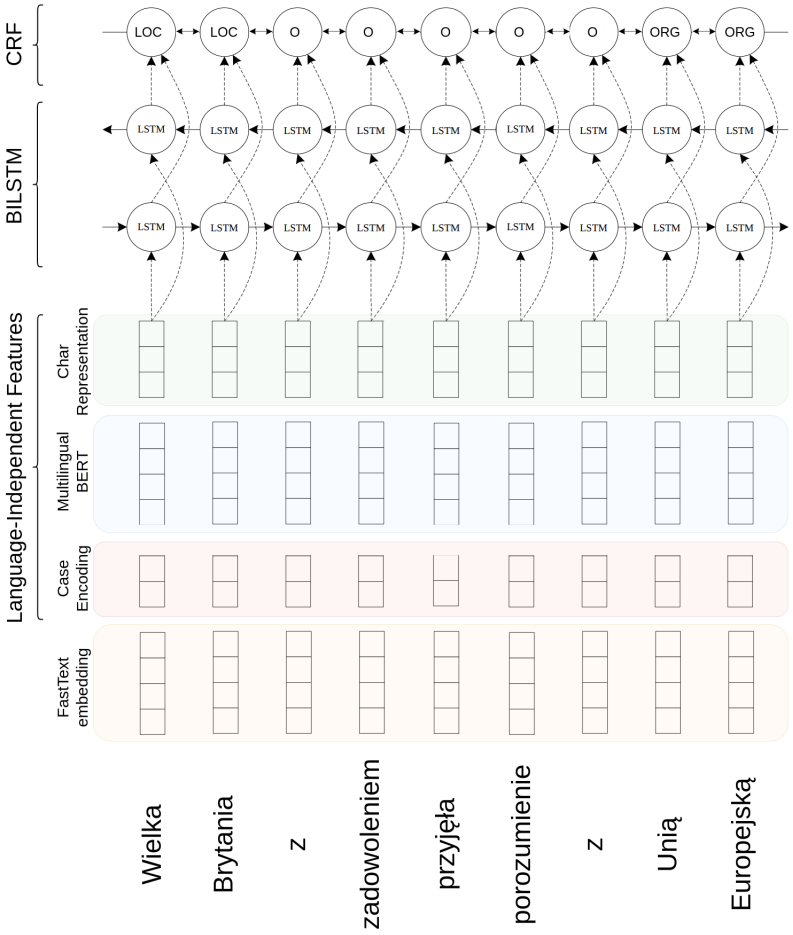

Figure 1: Architecture of a single-language model of our system. Note that for each token we provide a unique NER prediction.

2.1.2 Case Encoding

This layer allows to encode each token based on the case information as proposed by Reimers & Gurevych (2017). We have used the one-hot encoding of the following seven classes: {'numeric', 'mainly_numeric', 'allLower', 'allUpper', 'initialUpper', 'contains_digit', 'other'}.

2.1.3 Multilingual BERT

We also used the multilingual pre-trained embedding of BERT[4]. This embedding has the advantage that a unique representation is used across multiple languages. Also they are context dependant, providing extra information when compared to fastText. In particular, we used the model learned for 104 languages including all EMBEDDIA languages. This model is composed of 12 layers and 768 dimensions in each layer for a total of 110M parameters. Directly using the 12 layers can be hard to compute. To cope with this problem, we used the weighted strategy proposed by Reimers & Gurevych (2019) and combined only the first two layers[5]. As BERT is based on word pieces, a word may be composed of multiple BERT tokens. In this case, we averaged them to obtain a unique vector per word.

[4] https://github.com/google-research/bert/blob/master/multilingual.md
[5] In our implementation, the number of layers is a hyperparameter of the model.

2.1.4 Char Representation

We used the char representation strategy proposed by Ma & Hovy (2016) where char embeddings are combined using a convolutional neural network (CNN). Thus, an embedding vector is learned for each character by iterating trough the entire collection. Note that the four main EMBEDDIA languages[6] include unique characters which make the sharing of patterns between languages harder. To deal with this problem, we transliterated each token to the Latin alphabet using the unidecode library[7] as a preprocessing step. This conversion is only applied at this layer and is not used elsewhere.

2.1.5 BILSTM

In order to capture interrelated features between all the inputs, previous works have been interested in recurrent neural networks (RNN). Many different RNN architectures have been proposed, but BILSTM (Schuster et al., 1997) is one of the most successfully used because of its generalization and convergence properties. In particular, they are well-know for been easy to train. Indeed, BILSTM is a bidirectional (left to right and right to left) model based on Long-Short term memory (LSTM). We opted for a conventional implementation of BILSTM[8] and configure it to generate as output a new representation for each input vector. This new representation benefits of its neighbours to increase the contextualised knowledge. As opposite to all the previous layers, the weights of this layer are learned during training time.

2.1.6 CRF

Conditional Random Fields (CRF) Lafferty et al. (2001) is a learning algorithm widely used for structured prediction. This method produce a discrete prediction for each of the samples in a sequence. It has been widely used in NER, even before the emergence of deep learning techniques. Nowadays, several NER models use CRF as the latest layer to predict the final output. In practice, the algorithm must be optimised using dynamic programming strategies such as the Verbatim algorithm (Sniedovich, 2010). In our model, this algorithm uses as input the vectors obtained by the BILSTM layer and gives as output one of the entity types (LOC, PER, ORG, MISC, EVT) for each input vector.

2.1.7 Language-Dependent and Independent Features

Note that the “char representation”, “multilingual BERT”, and “case encoding” layers are grouped in Figure 1 as language-independent features. This is due to the fact that, in our model, they follow exactly the same process independently of the language. These steps become completely independent in this specific context as the system is not aware of the language that is being processed. So, all the processing steps are applied without considering the language, including the transliteration to the Latin alphabet. It means that some tokens are translated even knowing that they are already in a Latin alphabet. On the other hand, the “fastText embedding” layer is aware of the language making it a language-dependent feature. However, we intentionally reduce the language dependency by using the architecture in Figure 1. Each time a sentence is processed on training or testing time, we switched the “fastText embedding” model for the one corresponding to the sentence to be processed. As aforementioned, all the other layers remain unchanged.

[6] Croatian, Estonian, Finnish, and Slovenian.
[7] https://pypi.org/project/Unidecode/
[8] Available in most deep learning packages.

2.2 Experimental setup

We experimented with three of the collections described in deliverable D2.1, table 1. A short description is included here:

- CoNLL2003 (Sang & De Meulder, 2003) collection in English (13879 train, 3235 dev, and 3422 test sentences). The used metrics include the officially proposed metrics and standard metrics for the CoNLL2003 data set (F1-measure).
- Wikiann collection is a NER collection including 282 languages, including the EMBEDDIA languages. The main drawback of this collection is that it was built semi-automatically by using Wikipedia content.
- BSNLP collection is a NER collection composed by two data set, BSNLP2017 and BSNLP2019. They were built using news articles of filtered topics such as "nord_stream" and "ryanair" across Slavic Languages[9]. These topics include 1100 documents per language. Further details can be found in the shared task overview papers (Piskorski et al., 2019, 2017).

In order to perform our experiments, we transformed all data sets into the CoNLL format following the classical partition into train, dev, and test. Then training is performed epoch by epoch until no improvement is observed in the dev partition. Precision, Recall, and F1-measure are reported as presented in D2.1.

2.3 Experimental Assessment

Comparison to the state-of-the-art. The first experiments using the CoNLL data set are performed to compare our method against the state of the art algorithms. This comparison is done only in English in which the most recent results are provided or simpler to obtain (following experiments show how well our model is capable to adapt to low-resourced languages.)

As suggested by Reimers & Gurevych (2019), the computational cost of our model can be reduced by excluding some of the layers of the contextualised embeddings (BERT). For this reason, our results on the CoNLL data set are presented in Table 1 using two and six BERT embedding layers (see 2.1.3).

Two observations can be made. First, our model slightly under-performs two strong baselines (Ma & Hovy, 2016; Reimers & Gurevych, 2019). Second, using more BERT layers (six vs. two) improves our results. However, the amount of memory used is also increased manifold. In the following experiments, we set the number of layers (hyperparameter) to two, due to our computation constraints, despite a slightly downgrading performance for English.

Applicability to several languages. We now test how well our multi-lingual model performs on languages other than English. State-of-the-art methods by Ma & Hovy (2016) or Reimers & Gurevych (2019) are not available for other languages, so baselines from other methods are not available.

In the first multi-lingual tests we used the *wikiann data set*. We used partitions suggested for this data set, e.g., train, dev, test and extra. Only the train partition was used for training[10]. Parameters were selected using best performance in the dev partition. We report $F1$ values for dev and test partitions separately in Table 2. Our results are consistent across partitions and languages, e.g., no degradation was observed between dev and test partitions and most performances are around $F1 = 85.5$ with a higher performance for Slovenian. In all cases, the entities typed as PER unperformed w.r.t. other types, showing the particularities of less resourced languages. Despite the fact that direct comparison is not possible, our multi-lingual NER system obtains similar performance as for a data set of a high resourced language (e.g., the CoNLL data set for English).

[9] Croatian,Czech, Polish, Russian, Slovak, Slovene, Ukrainian for 2017; Bulgarian,Czech, Polish, Russian for 2019
[10] Extra partition was ignored.

Method		Metric		
	Set	P	R	F1
BRNN-CNN-CRF	Dev	94.8	94.6	94.7
(Ma & Hovy, 2016)	Test	**91.3**	91.0	91.2
BiLSTM + ElMo	Dev	**95.1**	**95.7**	**95.4**
(Reimers & Gurevych, 2019)	Test	90.9	**92.1**	**91.5**
BiLSTM + MultiBERT2L	Dev	92.3	93.0	92.7
(ours)	Test	88.2	89.7	89.0
BiLSTM + MultiBERT6L	Dev	93.2	93.8	93.5
(ours)	Test	89.3	90.3	89.8

Table 1: Evaluation results on the CoNLL 2003 data set, an English only data set.

Language		Entity Type			
	Set	LOC	PER	ORG	All
Croatian	Dev	85.69	81.20	90.96	86.11
	Test	86.30	80.91	90.20	85.97
Estonian	Dev	87.86	78.28	91.55	86.34
	Test	87.26	77.09	90.80	85.58
Finnish	Dev	85.83	74.00	91.16	84.18
	Test	86.61	75.32	**92.48**	85.35
Slovenian	Dev	**88.28**	**86.10**	**92.43**	**88.96**
	Test	**88.64**	**85.58**	92.18	**88.88**

Table 2: F1 results on the wikiann data set for selected EMBEDDIA languages.

Final experiments on multi-lingual NER were performed using the *BSNLP data sets*. As these data sets do not include the dev partition, the 20% of the training data is used as dev partition. We also create a partition into train and test of the test partition for the BSNLP2017 data set as no train data is publicly available[11]. Results in terms of $F1$ are presented in Table 3.

The most striking result is that the performance on BSNLP2017 is very low compared to BSNLP2019. This is explained by the size of the collection: more data was provided for the more recent BSNLP2019 data set. Indeed, the collection used for BSNLP2017 is composed of 380 documents with around 3700 mentions of entities. BSNLP2019 collections is much larger. In this case, a total of 1932 documents are used for training and 1102 for test, hence 48900 mentions of entities. This makes BSNLP2019 between 5 and 10 times larger than the BSNLP2017 collection. Our results suggest that to increase from $F1 = 27.11$ (average performance for BSNLP2017) to $F1 = 65.44$ (average performance for BSNLP2019) a 10 times larger number of annotated mentions is needed.

Another observation from the results is that in terms of entity type, the MISC[12], EVT and PRO types show the lowest performances. As above, this is due to the collection characteristics as these types are less represented than LOC, PER and ORG[13].

[11] Only test data is available for this data set.

[12] This type was explited into the EVT and PRO types in the BSNLP2019 data set.

[13] For Russian no EVT entities are found for the topic RYANAIR. Only 7, 12, and 4 entities are found for the same topic for Polish, Czech, and Bulgarian, respectively.

Language/data set	Set	Entity Type						All
		LOC	PER	ORG	MISC	EVT	PRO	
BSNLP2017								
Croatian	Dev	**74.81**	7.78	60.14	4.44	-	-	61.68
	Test	38.93	8.81	15.33	1.33	-	-	22.24
Czech	Dev	71.85	28.94	50.29	18.94	-	-	56.67
	Test	42.08	26.87	**28.12**	4.07	-	-	30.12
Polish	Dev	56.11	**48.35**	59.59	7.64	-	-	55.44
	Test	40.58	35.92	20.50	**10.35**	-	-	32.64
Russian	Dev	48.14	15.72	61.83	0.91	-	-	55.12
	Test	28.07	11.46	9.12	0.00	-	-	15.05
Slovak	Dev	63.19	40.22	45.09	**45.45**	-	-	49.92
	Test	35.99	25.30	22.95	1.09	-	-	27.07
Slovene	Dev	49.21	35.71	49.13	19.86	-	-	46.97
	Test	28.10	12.34	19.43	7.82	-	-	18.30
Ukrainian	Dev	68.82	43.74	**68.20**	20.27	-	-	**64.62**
	Test	**61.37**	**40.07**	17.93	0.00	-	-	**44.41**
BSNLP2019								
Bulgarian	Dev	99.32	99.11	**98.67**	-	**99.63**	91.86	98.88
	Test	**90.13**	**84.56**	**63.23**	-	**26.56**	22.04	**74.28**
Czech	Dev	**99.83**	99.46	98.19	-	99.22	**97.87**	**99.15**
	Test	74.29	59.56	43.42	-	0.00	14.57	60.34
Polish	Dev	97.89	96.29	95.23	-	97.08	90.78	96.40
	Test	84.44	67.19	61.44	-	20.18	**24.28**	67.50
Russian	Dev	99.55	**99.73**	97.87	-	99.56	94.05	99.08
	Test	78.84	59.59	45.79	-	0.00	10.61	59.66

Table 3: F1 results on the BSNLP17 and BSNLP19 data sets.

The results indicate that:

- as the number of annotations increases and as the closeness between the training topic and the target topic increases, our model is more likely to make the correct prediction;
- our method nearly reaches reference methods specialised for English;
- our model shows strong performance across different languages and data sets.

3 Named Entity Linking

Extending the analysis of NER that recognises named entities in documents, Named Entity Linking (NEL) aims to disambiguate these entities by linking them to entries of a Knowledge Base (KB). However, some mentions do not have a correspondent entry. Concerning these mentions, they should be

linked to the NIL entry. Besides, NIL mentions that refer to the same entity should be grouped.

So far, our NEL system disambiguates mentions to a KB and provides a NIL entry if a mention does not have a corresponding entry in the KB. In future work, we intend to cluster these NIL mentions by analysing their similarities.

Section 3.1 presents the Ganea and Hofmann's approach to disambiguate the mentions in a document. Then, Section 3.2 describes how we extended their approach to a cross-lingual analysis. Finally, the experimental setup and the evaluation of this approach on the wikiann corpora are presented in Sections 3.3 and 3.4, respectively.

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 & Hofmann (2017) proposed a deep learning model for joint document-level entity disambiguation[14] (Figure 2).

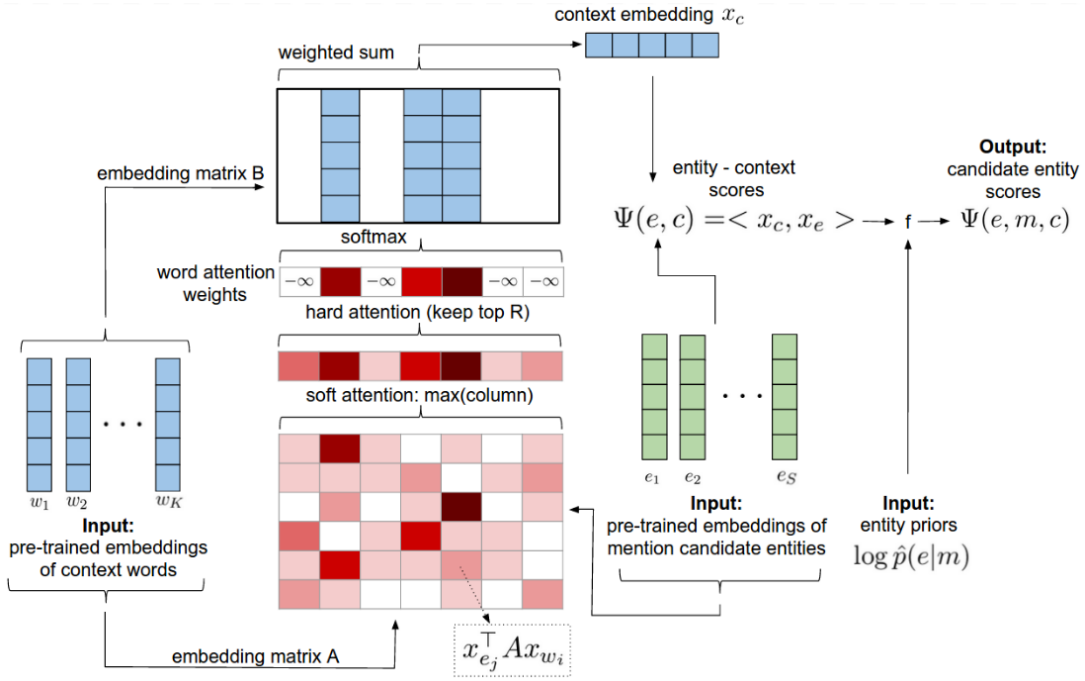

Figure 2: Architecture of the Ganea and Hofmann's approach. Their method uses a local model with neural attention to process context word vectors, candidate entity priors, and embeddings to generate the candidate entity scores (Ganea & Hofmann, 2017).

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.

Finally, mentions in a document are resolved jointly by using a conditional random field in conjunction with an inference scheme.

[14] The code is publicly available: `https://github.com/dalab/deep-ed`

3.2 Our contribution

As described in Deliverable D2.1, 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 EMBEDDIA languages, e.g., wikiann corpora.

In order to extend the Ganea and Hofmann's (GH) system to a cross-lingual setting, we made some modifications to their approach. Instead of using the word2vec embeddings, we used the pre-trained multilingual MUSE embeddings[15] (Conneau et al., 2017). These embeddings are available in 30 languages (including Croatian, Estonian, Finnish, and Slovenian) 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, the Ganea and Hofmann'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. In this scenario, the GH's approach analyses English documents and links their mentions to an English KB. Moreover, we extend the training process for the EMBEDDIA languages by using the previous English model and continue the training process with data on other languages. This post-training will optimise our model to analyse better the documents on the languages of the EMBEDDIA project and link their mention to a English KB.

3.3 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. (2013) 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. The associated task requires ranking of related candidate entities higher than the others. 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 the GH's approach with the benchmarks collected and described in deliverable D2.1. A short description is included here:

- **AIDA-CoNLL** data set (Hoffart et al., 2011) is based on CoNLL 2003 data that was used for NER task. This data set is divided into AIDA-train for training, AIDA-A for validation, and AIDA-B for testing. This data set contains 1393 Reuters news articles and 27817 linkable mentions.
- **AQUAINT** data set (Milne & Witten, 2008a; Guo & Barbosa, 2014) is composed of 50 short news documents (250-300 words) from the Xinhua News Service, the New York Times, and the Associated Press. This data set contains 727 mentions.
- **ACE2004** data set (Ratinov et al., 2011; Guo & Barbosa, 2014) is a subset of the ACE2004 coreference documents with 57 articles and 306 mentions, annotated through crowdsourcing.
- **MSNBC** data set (Cucerzan, 2007; Guo & Barbosa, 2014) is composed of 20 news articles from 10 different topics (two articles per topic: Business, U.S. Politics, Entertainment, Health, Sports, Tech & Science, Travel, TV News, U.S. News, and World News), having 656 linkable mentions in total.
- **Wikiann** (Pan et al., 2017) is a data set automatically built with name mentions extracted from Wikipedia. This data set is composed of 282 languages. More specifically, the Croatian, the

[15] The MUSE embeddings are available at: `https://github.com/facebookresearch/MUSE`

Estonian, the Finnish and the Slovenian data sets contain 76K, 69K, 341K and 67K mentions, respectively.

- Finally, **CWEB** (Gabrilovich et al., 2013) and **WIKI** (Ratinov et al., 2011) data sets are composed of 320 documents each.

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 EMBEDDIA language. Finally, we tested our model on the wikiann test data sets.

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.

3.4 Experimental Assessment

Entity embeddings performance. Table 4 shows the entity relatedness results using word2vec and MUSE embeddings for the English data set (Ceccarelli et al., 2013). 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 made the word2vec achieves the best results for all entity relatedness 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 the drop in performance, GH's approach using MUSE embeddings achieved better results than Yamada et al. (2016) and Milne & Witten (2008b) for all metrics.

Table 4: Entity relatedness quality for English.

Embeddings	NDCG1	NDCG5	NDCG10	MAP
Ganea and Hufmman (word2vec)	**0.632**	**0.609**	**0.641**	**0.578**
Ganea and Hufmman (MUSE)	0.613	0.568	0.592	0.536
Yamada et al. (2016)	0.59	0.56	0.59	0.52
Milne & Witten (2008b)	0.54	0.52	0.55	0.48

NEL analysis for mono- and multilingual embeddings. Advancing our analysis of the GH's system, we compared the F1 results for this system on English corpora using the word2vec and MUSE embeddings (Table 5). 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. Surprisingly, the GH's system using the MUSE embeddings achieved the best performance on the MSNBC data set.

Table 5: F1 results for the Ganea and Hofmann's approach on English corpora.

Word embeddings	AIDA	ACE2004	AQUAINT	CLUEWEB	MSNBC	WIKI
word2vec	**92.2**	**88.5**	**88.5**	**77.9**	93.7	**77.5**
MUSE	86.6	**88.5**	87.5	74.9	**94.4**	74.2

Cross-lingual NEL analysis. Table 6 presents the F1-measure results for the NEL on the wikiann corpora. We tested the NEL system using only the AIDA training data set to train the 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 6). The additional training process on the wikiann data set provided a minimal improvement on the performance of Ganea and Hofmman's for the wikiann test data sets.

Unfortunately, the wikiann data set is composed of short sentences with few context information. This characteristic makes the context analysis of the GH's system being less relevant and making the disambiguation process be decided mainly by the pairwise matching between mentions and entities on the $\log\ p(m|e)$. 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 6: F1 results for the Ganea and Hofmann's models on the test wikiann corpora.

Models	**Croatian**	**Estonian**	**Finnish**	**Slovenian**
AIDA training data set (using MUSE)	60.97	57.82	62.51	69.78
pre-trained model on AIDA data set + wikiann training data set (using MUSE)	**61.53**	**58.47**	**63.04**	**70.31**

In order to improve the results of the transfer learning technique, we should create training data sets on the target languages that are composed of long sentences with rich context information to improve our NEL model.

4 Associated outputs

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

Description	URL	Availability
Crosslingual NER	`github.com/EMBEDDIA/bert-bilstm-cnn-crf-ner`	To become public*
Crosslingual NEL	`github.com/EMBEDDIA/cross-lingual_entity_linking`	To become public*

* Resources marked here as "To become public" are available only within the consortium while under development and/or associated with work yet to be published. They will be released publicly when the associated work is completed and published.

Parts of this work are also described in detail in the following publications, which are attached to this deliverable as appendices:

Citation	Status	Appendix
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. doi: 10.18653/v1/W19-3711	Published	Appendix A

5 Conclusions and further work

The present deliverable describes the initial version of named entity recognition and linking systems, where event detection was performed through a specific type of named entity. These systems provide the baselines for cross-lingual tasks in T2.1. We intend to combine state-of-the-art approaches and

transfer learning techniques in order to propose a new system capable of overcoming the limitations of less-resourced languages. The latest version of these systems will be delivered at M24 as D2.5.

Towards this final deliverable of Task T2.1, we will work towards the following targets. For NER, we are interested in a deeper exploration of contextualised embeddings. Indeed, recent works have shown that including the NER models within contextualised models will allow for better results (Devlin et al., 2018). It seems a promising direction for the project also because the results of EMBEDDIA embeddings may be improved by using our NER models and not only the other way around. An iterative improvement between embeddings and NER system is currently under discussion between the EMBEDDIA partners ULR and UL.

One key planned development is indeed the collaborative junction of the works presented in Deliverable D1.2 (where NER is used as a means for extrinsic evaluation of the quality of word embeddings) with the works of the present deliverable D2.2 (where cross-lingual algorithms for NER, given embeddings, are evaluated intrinsically).

When it comes to NEL, we intend to adapt and build resources to use Wikipedia in other languages than English. Then, we will provide monolingual NEL models for the remaining languages of the project (so far, our NEL system only links mentions to the English Wikipedia). We also intend to optimise our cross-lingual models with new training data sets that contain rich context information on the languages of the EMBEDDIA project (e.g., the AIDA data set for English). Finally, we will combine recent approaches of the state of the art (e.g., attention mechanisms (Ganea & Hofmann, 2017), latent relations between mentions (Le & Titov, 2018) and end-to-end models (Kolitsas et al., 2018)) to extract and analyse more context information of documents and improve the performance of cross-lingual NEL.

Finally, for event detection, we are developing specific tools for the detection of events, based on a rhetorical analysis of the news genre (building up on works of Lejeune et al. (2015)) and on the use of neural methods for event extraction at the sentence level as done by Boros (2018).

References

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.

Boros, E. (2018). *Neural methods for event extraction* (Unpublished doctoral dissertation). Université de Paris-Saclay.

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., & Trani, S. (2013). Learning relatedness measures for entity linking. In *Proceedings of the 22nd acm international conference on information & knowledge management* (pp. 139–148). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2505515.2505711 doi: 10.1145/2505515.2505711

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Cucerzan, S. (2007, June). 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). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D07-1074

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *2019 annual conference of the north american chapter of the association for computational linguistics (naacl-hlt)* (pp. 4171–4186).

Gabrilovich, E., Ringgaard, M., & Subramanya, A. (2013, June). *FACC1: Freebase annotation of ClueWeb corpora, Version 1 (Release date 2013-06-26, Format version 1, Correction level 0).*

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

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).*

Guo, Z., & Barbosa, D. (2014). Robust entity linking via random walks. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management* (pp. 499–508). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2661829.2661887 doi: 10.1145/2661829.2661887

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 782–792). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145521

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

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). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645530.655813

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 (volume 1: Long papers)* (pp. 1595–1604). Association for Computational Linguistics. Retrieved from http://aclweb.org/anthology/P18-1148

Lejeune, G., Brixtel, R., Doucet, A., & Lucas, N. (2015, October). Multilingual event extraction for epidemic detection. *Artificial Intelligence in Medicine*, *65*(2), 131–143. Retrieved from http://dx.doi.org/10.1016/j.artmed.2015.06.005 doi: 10.1016/j.artmed.2015.06.005

Ma, X., & Hovy, E. (2016, August). 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). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P16-1101 doi: 10.18653/v1/P16-1101

Milne, D., & Witten, I. H. (2008a). Learning to link with wikipedia. In *Proceedings of the 17th acm conference on information and knowledge management* (pp. 509–518). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1458082.1458150 doi: 10.1145/1458082.1458150

Milne, D., & Witten, I. H. (2008b). Learning to link with wikipedia. In *Cikm '08: Proceeding of the 17th acm conference on information and knowledge mining* (pp. 509–518). New York, NY, USA: ACM. Retrieved from http://www.cs.waikato.ac.nz/~dnk2/publications/CIKM08-LearningToLinkWithWikipedia.pdf doi: http://doi.acm.org/10.1145/1458082.1458150

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

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., & Ji, H. (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)* (pp. 1946–1958). Vancouver, Canada: Association for Computational Linguistics.

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Přibáň, P., Steinberger, J., & Yangarber, R. (2019). The Second Cross-Lingual Challenge on Recognition, Classification, Lemmatization, and Linking of Named Entities across Slavic Languages. In *Proceedings of the 7th workshop on Balto-Slavic natural language processing.* Florence, Italy: Association for Computational Linguistics.

Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., Yangarber, R., et al. (2017). The First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages. In *Proceedings of the 6th workshop on balto-slavic natural language processing.*

Ratinov, L., Roth, D., Downey, D., & Anderson, M. (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* (pp. 1375–1384). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002642

Reimers, N., & Gurevych, I. (2017, 09). Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 conference on empirical*

methods in natural language processing (emnlp) (pp. 338–348). Copenhagen, Denmark. Retrieved from http://aclweb.org/anthology/D17-1035

Reimers, N., & Gurevych, I. (2019). Alternative Weighting Schemes for ELMo Embeddings. *arXiv preprint arXiv:1904.02954*.

Sang, E. F. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Development* (Vol. 922, p. 1341).

Schuster, M., Paliwal, K. K., & General, A. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*.

Sniedovich, M. (2010). *Dynamic programming: Foundations and principles*. CRC Press.

Yamada, I., Shindo, H., Takeda, H., & Takefuji, Y. (2016, August). 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* (pp. 250–259). Berlin, Germany: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K16-1025 doi: 10.18653/v1/K16-1025

TLR at BSNLP2019: A Multilingual Named Entity Recognition System

Jose G. Moreno University of Toulouse IRIT, UMR 5505 CNRS jose.moreno@irit.fr

Elvys Linhares Pontes[1,2] **Mickaël Coustaty**[1] **Antoine Doucet**[1]
1 - L3i laboratory, University of La Rochelle, La Rochelle, France
{firstname.lastname}@univ-lr.fr
2 - University of Avignon, Avignon, France

Abstract

This paper presents our participation at the shared task on multilingual named entity recognition at BSNLP2019. Our strategy is based on a standard neural architecture for sequence labeling. In particular, we use a mixed model which combines multilingual-contextual and language-specific embeddings. Our only submitted run is based on a voting schema using multiple models, one for each of the four languages of the task (Bulgarian, Czech, Polish, and Russian) and another for English. Results for named entity recognition are encouraging for all languages, varying from 60% to 83% in terms of Strict and Relaxed metrics, respectively.

1 Introduction

Correctly detecting mentions of entities in text documents in multiple languages is a challenging task (Ji et al., 2014, 2015; Ji and Nothman, 2016; Ji et al., 2017). This is especially true when documents relate to news because of the huge range of topics covered by newspapers. In this context, the shared task on multilingual named entity recognition (NER) proposes to participants to test their system under a multilingual setup. Four languages are addressed in BSNLP2019: Bulgarian (bg), Czech (cz), Polish (pl), and Russian (ru). Similarly to the first edition of this task in 2017 (Piskorski et al., 2017), participants are required to recognize, normalize, and link entities from raw texts written in multiple languages. Our participation is focused on the sole recognition of entities while other steps will be covered in our future work.

In order to build a unique NER system for multiple languages, we decided to contribute a solution based on an end-to-end system without (or almost without) language specific pre-processing. We explored an existing neural architecture, the LSTM-CNNs-CRF (Ma and Hovy, 2016), initially proposed for NER in English. This neural model is based on word embeddings to represent each token in a sentence. In order to have a unique embedding space, we propose to use a transformer-based (Vaswani et al., 2017) contextual embedding called BERT (Devlin et al., 2019). This pre-trained model includes multilingual representations that are context-aware. However, as noted by Reimers and Gurevych (2019), contextual embeddings provide multiple layers that are challenging to combine together. To overcome this problem, we used the weighted average strategy they successfully tested using (Peters et al., 2018).

The results of our participation are quite encouraging. Regarding the *Relaxed Partial* metric, our run achieves 80.26% in average for the four languages and the two topics that compose the test collection. In order to present comparative results against the state of the art, we run experiments using two extra datasets under the standard CoNLL evaluation setup. The remainder of this paper is organized as follows: Section 2 introduces the related work while Section 3 presents the proposed multi-lingual model. Section 4 presents the results while conclusions are drawn in Section 5.

2 Related Work

Named entity recognition has been largely studied through the organization of shared tasks in the last two decades (Nadeau and Sekine, 2007; Yadav and Bethard, 2018). The large variety of models can be grouped into three types: rule-based (Chiticariu et al., 2010), gazetteers-based (Sundheim, 1995), and statistically-based models (Florian et al., 2003). The latter type is a current hot topic in research, in particular with the return of neural based models[1]. Two main contributions

[1] In all their flavors, including attention.

have recently redrawn the landscape of models for sequence labelling such as NER: the proposal of new architectures (Ma and Hovy, 2016; Lample et al., 2016), the use of contextualized embeddings (Peters et al., 2018; Reimers and Gurevych, 2019), or even, the use of both of them (Devlin et al., 2019). The use of contextualized embeddings is a clear advantage for several kinds of neural-based NER systems, however as pointed out by Reimers and Gurevych (2019) the combination of multiples vectors proposed by these models is computationally expensive.

3 TLR System: A Neural-based Multilingual NER Tagger

This section describes our model which is based on a standard end-to-end architecture for sequence labeling, namely LSTM-CNNs-CRF (Ma and Hovy, 2016). We have combined this architecture with contextual embeddings using a weighted average strategy (Reimers and Gurevych, 2019) applied to a pre-trained model for multiple languages (Devlin et al., 2019) (including all languages of the task). We trained a NER model for each of the four languages and predict labels based on a classical voting strategy. As an example, the overall architecture of our model for Polish using the sentence *"Wielka Brytania z zadowoleniem przyjęła porozumienie z Unia Europejska"* (or "United Kingdom welcomes agreement with the European Union" in English) is depicted in Figure 1.

3.1 FastText Embedding

In this layer, we used pre-trained embeddings for each language trained on Common Crawl and Wikipedia using fastText (Bojanowski et al., 2017; Grave et al., 2018). These models were trained using the continuous bag-of-words (CBOW) strategy with position weights. A total of 300 dimensions were used with character n-grams of length 5, a window of size 5 and 10 negatives. The four languages of the task are included in this publicly available[2] pre-trained embedding (Grave et al., 2018). We have used the fastText library to ensure that every token (also in other alphabets) has a corresponding vector avoiding out of vocabulary tokens.

[2]https://fasttext.cc/docs/en/crawl-vectors.html

3.2 Case Encoding

This layer allows to encode each token based on the case information as proposed by (Reimers and Gurevych, 2017). We have used a one-hot encoding of the following seven classes: {'other', 'numeric', 'mainly_numeric', 'allLower', 'allUpper', 'initialUpper', 'contains_digit'}.

3.3 Multilingual BERT

We used the multilingual pre-trained embedding of BERT[3]. In particular, we used the model learned for 104 languages including the four of this task. This model is composed of 12 layers and 768 dimensions in each layer for a total of 110M parameters. Directly using the 12 layers can be hard to compute in a desktop computer. To cope with this problem, we used the weighted strategy proposed by Reimers and Gurevych (2019) and combined only the first two layers. When a token was composed of multiple BERT tokens, we averaged them to obtain a unique vector per token.

3.4 Char Representation

We used the char representation strategy proposed by Ma and Hovy (2016) where char embeddings are combined using a convolutional neural network (CNN). Thus, an embedding vector is learned for each character by iterating trough the entire collection. Note that the four languages include unique characters which make harder the sharing of patterns between languages. To deal with this problem, we transliterated each token to the Latin alphabet using the unidecode library[4] as a preprocessing step. This conversion is only applied at this layer and is not used elsewhere.

3.5 Language-Dependent and Independent Features

In Figure 1, we observe that the "char representation", "multilingual BERT", and "case encoding" layers are language-independent features[5] So, all the processing steps are applied without considering the language, including the transliteration to the Latin alphabet. It means that some tokens are translated even knowing that they are already in a

[3]https://github.com/google-research/bert/blob/master/multilingual.md

[4]https://pypi.org/project/Unidecode/

[5]We mean that as the four languages follow exactly the same process, those steps become completely independent in this specific context.

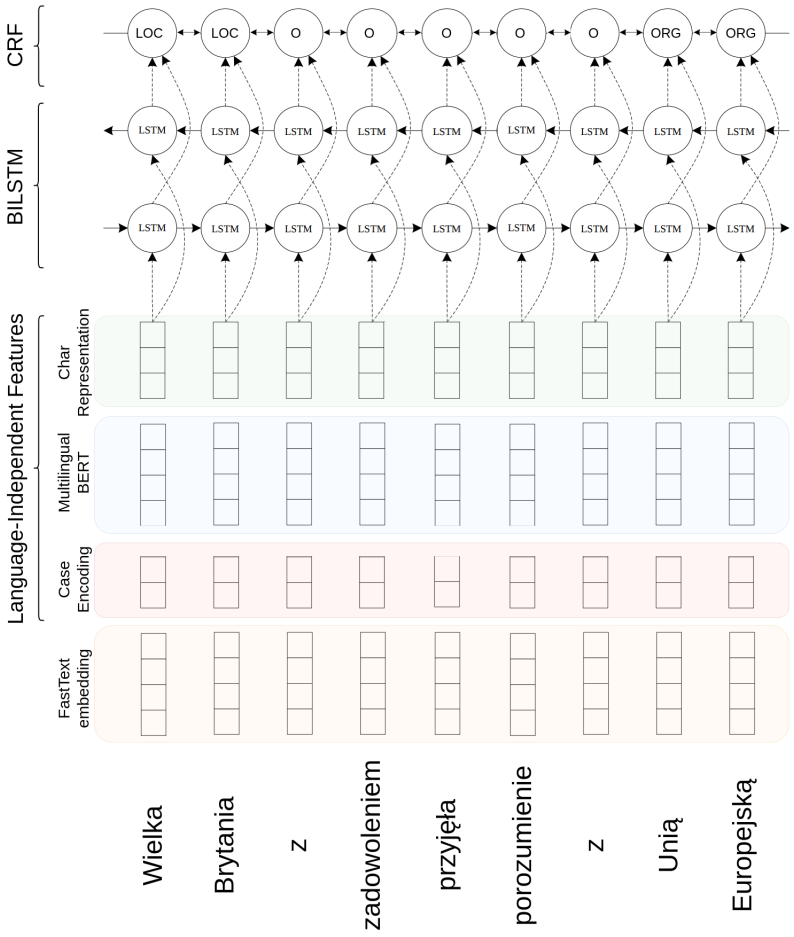

Figure 1: Architecture of a single-language model of our system. Note that for each token we provide a unique NER prediction.

Latin alphabet. On the other hand, "fastText embedding" is clearly a language-dependent feature. However, we intentionally reduce the language dependency by using the architecture in Figure 1 as many times as the number of languages involved in the task, e.g. four times. Each time we switched the "fastText embedding" model for the one corresponding to each language, this make a total of four different NER models. Our final prediction is obtained by applying a simple majority voting schema between these four NER models.

4 Experiments

4.1 Experimental Setup

We follow the configuration setup proposed by the task organizers. Two topics, "nord_stream" and "ryanair", were used to test our models. These topics include 1100 documents in the four languages. Further details can be found in the 2019 shared task overview paper (Piskorski et al., 2019). For training, we have used the documents provided for the task but also the ones in Czech, Polish, and Russian from the previous round of same task in 2017 (Piskorski et al., 2017). We additionally added the training example form the CoNLL2003 (Sang and De Meulder, 1837) collection in English (13879 train, 3235 dev, and 3422 test sentences). Used metrics include the officially proposed metrics and standard metrics for the CoNLL2003 dataset (F1 metric).

4.2 Official Results

The official results of our unique run are presented in Table 1 and identified as TLR-1. Note that only NER metrics are presented for the four languages. We have added the results for each language model using the partial annotations provided by the organizers[6]. Each result is identified with the language used for the "fastText embed-

[6]We were able to calculate "Recognition Strict" for these unofficial results.

NORD_STREAM		*Language*							
Phase	*Metric*	bg		cz		pl		ru	
Recognition	Relaxed Partial	TLR-1	83.384	TLR-1	82.124	TLR-1	80.665	TLR-1	73.145
	Relaxed Exact	TLR-1	76.114	TLR-1	74.106	TLR-1	71.423	TLR-1	62.168
	Strict	TLR-1	73.312	TLR-1	74.475	TLR-1	72.026	TLR-1	59.627
		bg	**72.873**	bg	67.841	bg	68.281	bg	54.922
		cz	68.821	cz	**78.225**	cz	71.509	cz	52.590
		pl	69.892	pl	73.636	pl	**75.820**	pl	53.939
		ru	72.661	ru	71.522	ru	70.356	ru	**58.399**
RYANAIR		*Language*							
Phase	*Metric*	bg		cz		pl		ru	
Recognition	Relaxed Partial	TLR-1	75.861	TLR-1	82.865	TLR-1	82.182	TLR-1	83.419
	Relaxed Exact	TLR-1	69.824	TLR-1	73.493	TLR-1	77.463	TLR-1	78.303
	Strict	TLR-1	68.377	TLR-1	72.509	TLR-1	75.118	TLR-1	78.028
		bg	**76.152**	bg	77.533	bg	79.168	bg	78.518
		cz	61.755	cz	**78.549**	cz	76.863	cz	75.280
		pl	67.876	pl	77.907	pl	**82.242**	pl	76.864
		ru	70.288	ru	74.805	ru	76.135	ru	**79.784**

Table 1: Evaluation results of our TLR submission. We have added extra results for the strict metric using each single model based on one of the four languages.

ding" layer in Figure 1. Based on strict recognition, most of the cases[7], the use of the correct language embedding improves the recognition of the respective language. However, the voting schema outperforms the individual models on average. This suggest that a system aware of the language of the input sentence could provide better results that our voting schema.

4.3 Unofficial Results

In order to compare our system to the state-of-the-art, we have evaluated our architecture using the CoNLL2003 dataset. Our results using two and six layers are presented in Table 2. Note that English is not part of our target languages. So, an underperformance of 2.5 is acceptable in our system[8]. It is also worth nothing that the use of more BERT layers increases our results. However, the amount of memory used is also increased manifold. We set the number of layers (hyperparameter) to two layers due to our computation constraints despite the downgrading in performances for English.

The number of epochs (hyperparameter) was set using the BSNLP2017 dataset (for ru, cs, and

Method		Metric		
	Set	P	R	F1
BRNN-CNN-CRF	Dev	94.8	94.6	94.7
(Ma and Hovy, 2016)	Test	91.3	91.0	91.2
BiLSTM + ElMo	Dev	95.1	95.7	95.4
(Reimers and Gurevych, 2019)	Test	90.9	92.1	91.5
BiLSTM + MultiBERT2L	Dev	92.3	93.0	92.7
(ours)	Test	88.2	89.7	89.0
BiLSTM + MultiBERT6L	Dev	93.2	93.8	93.5
(ours)	Test	89.3	90.3	89.8

Table 2: Evaluation results on the CoNLL 2003 dataset, an English only dataset.

[7] 6 out of 8, with differences smaller than 0.4 points.

[8] More experiments using BERT English-only model will be performed in our future work.

Language	BSNLP2017+CoNLL2003			
	P	R	F1	Epochs
en	78.9	82.8	80.8	10
bg	77.1	79.3	78.2	6
cz	78.7	82.2	80.4	24
pl	79.7	83.6	81.6	16
ru	79.1	83.4	81.2	21

Table 3: Evaluation results on the BSNLP2017 and CoNLL 2003 datasets, a multilingual dataset. Each row represents a model learned with a fastText language specific embedding.

pl) combined with CoNLL2003 as a validation set of our final models. Results for these combined datasets are presented in Table 3. Surprisingly, our results seem very similar independently of the fastText embedding. It suggests that our architecture is able to generalize the prediction for several target languages. Note that the worst results are obtained by the Bulgarian model, but no test examples were included for this language. In contrast, we believe that the examples provided in other languages were rich enough to help the predictions (also in English).

5 Conclusion

This paper presents the TLR participation at the shared task on multilingual named entity recognition at BSNLP2019. Our system is a combination of multiple representation including character information, multilingual embedding, and language specific embedding. However, we combine them in such a way that it can be seen as a generic multilingual NER system for a large number of languages (104 in total). Although top participants outperform our average score of 80.26% of "Relaxed Partial" (Piskorski et al., 2019), the strengths of the proposed strategy relays on the fact that it can be easily adapted to new languages and topics without extra effort.

Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain Adaptation of Rule-based Annotators for Named-Entity Recognition Tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1002–1012. 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 *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named Entity Recognition through Classifier Combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

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)*.

Heng Ji and Joel Nothman. 2016. Overview of TAC-KBP2016 Tri-lingual EDL and its impact on end-to-end Cold-Start KBP. In *Proc. Text Analysis Conference (TAC2016)*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In *Proc. Text Analysis Conference (TAC2015)*.

Heng Ji, Joel Nothman, Ben Hachey, et al. 2014. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proc. Text Analysis Conference (TAC2014)*, pages 1333–1339.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, Cash Costello, and Sydney Informatics Hub. 2017. Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking. In *Proc. Text Analysis Conference (TAC2017)*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 260–270.

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.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The Second Cross-Lingual Challenge on Recognition, Classification, Lemmatization, and Linking of Named Entities across Slavic Languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, Roman Yangarber, et al. 2017. The First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Nils Reimers and Iryna Gurevych. 2019. Alternative Weighting Schemes for ELMo Embeddings. *arXiv preprint arXiv:1904.02954*.

Erik F Tjong Kim Sang and Fien De Meulder. 1837. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Development*, volume 922, page 1341.

Beth M. Sundheim. 1995. Overview of Results of the MUC-6 Evaluation. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 13–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

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.