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

D2.1: Datasets, benchmarks and evaluation metrics for advanced cross-lingual NLP technology (T2.4)

Executive summary

The present report describes datasets and other resources collected for the benchmarking and the evaluation of the tasks of WP2 focused on 'Advanced NLP Technologies for Less-Resourced Languages'. For each task, the report describes the available resources. It further describes existing state-of-the-art tools and benchmarks, stating which solutions are intended to be used for the evaluation of WP2 tasks. A second and final version of this report will be delivered at M30 as D2.7.

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:	**D2.1**
Deliverable full title:	**Datasets, benchmarks and evaluation metrics for advanced cross-lingual NLP technology**
Deliverable short title:	**Datasets and evaluation for NLP technology**
Document identifier:	**EMBEDDIA-D21-DatasetsAndEvaluationForNLPTechnology-T24-submitted**
Lead partner short name:	**ULR**
Report version:	**submitted**
Report submission date:	**30/09/2019**
Dissemination level:	**PU**
Nature:	**R = Report**
Lead author(s):	**Jose G Moreno (ULR)**
Co-author(s):	**Elvys Linhares Pontes (ULR), Antoine Doucet (ULR), Leo Leppänen (UH), Andraž Repar (JSI), Senja Pollak (JSI)**
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
10/07/2019	v0.1	Jose G Moreno (ULR)	Initial version.
16/07/2019	v0.2	Elvys Linhares Pontes (ULR)	Datasets and benchmarks for NEL.
20/08/2019	v0.3	Antoine Doucet (ULR)	Proofread, restructured and added a deliverable introduction.
21/08/2019	v0.4	Antoine Doucet (ULR)	Restructured to easily accommodate for T2.2 and T2.3, updated the document consequently, in particular its introduction.
22/08/2019	v0.5	Andraž Repar (JSI), Senja Pollak (JSI)	Keyword and term extraction datasets
23/08/2019	v0.6	Leo Leppänen(UH)	Added Section 4 on Task T2.3
29/08/2019	v0.7	Jose G Moreno (ULR), Elvys Linhares Pontes (ULR)	Proofread and evaluation merging for NER and NEL.
30/08/2019	v1.0	Jose G Moreno (ULR)	Final changes before internal review.
10/09/2019	v1.1	Matthew Purver (QMUL)	Internal review
23/09/2019	v1.2	Saturnino Luz (UEDIN)	Internal review
24/09/2019	v1.3	Andraž Repar (JSI), Senja Pollak (JSI), Jose G Moreno (ULR), Antoine Doucet (ULR), Elvys Linhares Pontes (ULR), Leo Leppänen (UH)	Changes implementing comments of internal reviewers
25/09/2019	v1.4	Nada Lavrač (JSI)	Quality control
27/09/2019	v1.5	Jose G Moreno (ULR), Elvys Linhares Pontes (ULR), Antoine Doucet (ULR), Nicolas Sidère (ULR)	Final revision based on comments resulting from quality control
28/09/2019	v1.6	Andraž Repar (JSI), Senja Pollak (JSI)	Final revision based on comments resulting from quality control keyword sections
30/09/2019	submitted	Tina Anžič (JSI)	Submitted

Table of Contents

List of abbreviations

CLNEL	Cross-Lingual Named Entity Linking
ED	Entity Disambiguation
IPTC	International Press Telecommunications Council
KB	Knowledge Base
NE	Named Entity
NEL	Named Entity Linking
NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
SER	Slot Error Rate
STT	Finnish News Agency, *Suomen Tietotoimisto*
T	Task
WP	Work Package

1 Introduction

This report constitutes the first deliverable of EMBEDDIA's work package 2 (WP2) entitled 'Advanced NLP Technologies for Less-Resourced Languages'.

Within this work package, an essential task is Task T2.4 named 'public resource gathering, benchmarking and evaluation', where all technologies developed in WP2 are to be evaluated. These technologies correspond to 3 different tasks:

- Task T2.1 'Cross-lingual semantic enrichment' focused on named entity recognition and linking and event detection,
- Task T2.2 'Multilingual keyword extraction and matching' focused on monolingual and multilingual methods to extract topical terms and keywords, and
- Task T2.3 'Multilingual natural language generation', concerned with natural language generation.

The output of Task T2.4 is delivered in 2 batches; The present report D2.1, delivered at M9, is a first version of the final report to be delivered as D2.7 at M30.

The remainder of this report is organized in 3 sections corresponding to the different tasks of WP2. Section 2 presents the datasets and resources concerning named entity recognition and linking and event detection (Task T2.1). Section 3 describes monolingual and multilingual methods, datasets and metrics to perform and evaluate the extraction of topical terms and keywords (Task 2.2). Section 4 summarizes the work on resource gathering, benchmarking and evaluation for natural language generation (Task T2.3). Finally, conclusions are set out in Section 5.

2 Task T2.1: Cross-lingual semantic enrichment

In Subsection 2.1, we provide a short description of named entities and entity linking. Subsection 2.2 describes the available resources for named entity recognition and linking. In Subsection 2.3 we explain the metrics to evaluate NER and NEL systems.

2.1 Background

Named Entity Recognition (NER) is a traditional Natural Language Processing (NLP) task used in multiple text applications. The purpose of a Named Entity Recognition (NER) system is to extract and type mentions of named entities from raw texts, e.g. identify continuous[1]sequences of words –the mention of an entity– that refer to a unique entity and assign each one an entity type. Entities are usually typed as person (PER), organization (ORG), location (LOC), and miscellaneous (MIS). Entity types have been extended to include many other types including date, currency, geo-political, event, etc. NER systems are mainly studied in NLP, but used as input in other fields including biology, bio-medicine, web semantics, information sciences, etc.

NER emerged as a research topic in the middle of the 90s (Grishman & Sundheim, 1996). The early systems relied on rule-based approaches. Such techniques require costly efforts and time since the rules are defined by humans and based on dictionaries, trigger words and linguistic descriptors. Since then, efforts have been focused on machine learning techniques. These techniques include sequential tagging methods such as Hidden Markov Models (Bikel et al., 1998) and Conditional Random Fields (CRFs) (Filannino et al., 2013) as well as Support Vector Machines (SVM) (Asahara & Matsumoto,

[1] We are aware of languages (such as Croatian) where entity mentions are not continuous but these are rare situations in the NER corpus of Croatian (it happens only 4 times in 25258 sentences) presented in Section 2.2.1.

2003) and Decision Trees (Sekine, 1998). More recently, neural networks have been shown to outperform other NER approaches (Collobert et al., 2011). They have reached very competitive results for NER in comparison to previous machine learning works (Dernoncourt et al., 2017; Peters et al., 2017, 2018). However, rule- and terminological-based approaches are still used and may achieve competitive performances in restricted domains when expert knowledge can be easily incorporated (Eftimov et al., 2017).

Named Entity Linking (NEL) extends the analysis of NER. In addition to identifying entities, NEL aims to retrieve the ground truth entities in a Knowledge Base (KB) referred to in a document by locating mentions, and for each mention accurately disambiguating the referent entity (Figure 1). However, NEL has to provide an unknown entry if a mention doesn't have a corresponding entry in the KB.

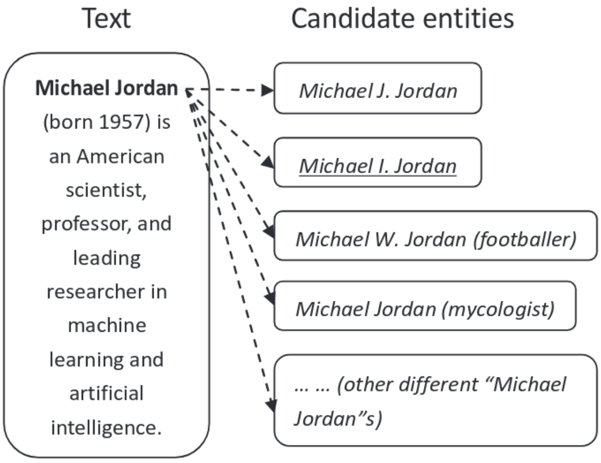
Text
Candidate entities
Michael Jordan (born 1957) is an American scientist, professor, and leading researcher in machine learning and artificial intelligence.
Michael J. Jordan
Michael I. Jordan
Michael W. Jordan (footballer)
Michael Jordan (mycologist)
… … (other different "Michael Jordan"s)

Figure 1: An illustration for the named entity linking task. The named entity mention detected from the text is in bold face; the correct mapping entity is underlined (Shen et al., 2015).

2.1.1 Named Entity

A Named Entity (NE) can be defined as a set of characters that, once assembled, can take different forms (word, compound word, group of words, acronym, date, etc.) and refers to a unique entity. We follow the standard definition of named entity given by Nadeau & Sekine (2007) which considers only entities to which one or many rigid designators stands for referent. This differentiates a set of characters forming a NE from another set of characters not forming a NE, and the possibility of placing it in one of the NE categories.

Indeed, a NE must be classifiable. In the context of a sentence, a NE must belong to only one category. This rule, as simple as it is, can lead us to rather complex situations. For example, the word "Washington" can both refer to a state, a city, or the name of a president. This type of problem is solved by an important activity of the NER systems, the disambiguation. This activity allows you to decide the meaning of a word in a sentence when it contains several words.

Various label sets have been used on the NER datasets (see Table 1). The labels *LOC* for location, *ORG* for organization, and *PER* for person are common in all of them. However, most works have used more refined labels:

- Person: A NE in this category represents a person, real or not, living or not. Anything that makes it possible to recognize a person (surname, first name) can belong to this category. For example, "Nicolas Sarkozy" represents a NE in the PERSON category just like "Nicolas" and "Sarkozy" individually.

- Location: A NE in this category represents a city, a country, a street, a place name that can be uniquely located. For example, “library” is not recognized as a location, but “national library of France” is well recognized as it.
- Organization: A NE in this category represents a name of a company, institution, government or organization. e.g. “UN” for United Nations.
- Miscellaneous: A NE in this category represents an entity that cannot be easily classified in the previous three classes. However, this category is not considered when more fine categories are added.
- Date/Time: A NE in this category represents a date, an hour, a month, or a relative mention.
- Event: A NE in this category represents an event. e.g. the “November 2015 Paris attacks” is a unique event.

2.1.2 Knowledge Base

A knowledge base is a centralized repository where information is stored, organized, and then shared. It stores complex structured and unstructured information. Knowledge bases (e.g. Wikipedia[2], DBpedia (Lehmann et al., 2015), YAGO (Suchanek et al., 2007), and Freebase (Bollacker et al., 2008)) contain rich information about the world’s entities, their semantic classes, and their mutual relationships. For instance, DBpedia knowledge base is composed of 111 different language editions of Wikipedia. Its largest knowledge base which is extracted from the English edition of Wikipedia consists of over 400 million facts that describe 3.7 million things (Lehmann et al., 2015).

2.1.3 Entity Linking

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 et al., 2015). NEL approaches can be divided into two classes:

- **Disambiguation approaches** only analyze gold standard named entities in a document and disambiguate them to the correct entry in a given KB (Ganea & Hofmann, 2017; Le & Titov, 2018; Raiman & Raiman, 2018).
- **‘End-to-End’ approaches** extract candidate entities from documents and then disambiguate them to the correct entries in a given KB (Kolitsas et al., 2018).

Recently, the study of less resourced languages has taken the main scene in NLP research. NEL has not escaped this trend and several approaches have been developed to analyze this problem in several languages (multilingual NEL) and cross-languages (cross-lingual NEL). Most precisely, Cross-lingual Named Entity Linking (CLNEL) aims to ground entity mentions written in any language to an English Knowledge Base (KB), such as Wikipedia (Sil et al., 2018; Upadhyay et al., 2018).

2.1.4 Event Detection

The aim of event detection is to markup potentially breaking events from large volumes of news stories. We first analyze event detection as a sub-problem of NER. In other words, we consider the event type as a valid type of a named entity. This definition has been previously used to evaluate the performance

[2] http://www.wikipedia.org/

of event detection (Piskorski et al., 2019, 2017). As an example, consider the sentence *"According to Maria Zakharova, the last straw was the story of the poisoning in Amesbury, where, unlike the incident with Russian citizens in Salisbury, a United Kingdom citizen died."* extracted from a training document of the BSNLP2019 shared task (Piskorski et al., 2019)[3]. In this task, it is expected that NER systems recognize and type *"poisoning in Amesbury"* as an event[4]. This definition of *Event Detection* extends the classical types used in NER to include events, but differs to the one known as slot filling. Later in the project, we might consider other definition such as the one proposed by Lejeune et al. (2013).

2.2 Available Resources

In this section, we list the existing resources available for the evaluation of the cross-lingual NLP applications targeted in WP2.

2.2.1 Named Entity Recognition

The details of NER datasets are presented in Table 1. NER corpora can vary from each other on the annotation guidelines. Various guidelines have been proposed, most of them follow the IOB-scheme (Inside, Outside, Beginning) where every token is labeled as B-label if the token is the beginning of a named entity, I-label if it is inside a named entity but not the first token within the named entity, or O otherwise (Ramshaw & Marcus, 1999). Let us consider the following sentence as example:

> *The_O president_O of_O the_B-ORG Duma_I-ORG, Gennadi_B-PER Selevniov_I-PER, qualified_O from_O the_B-ORG Kremlin_I-ORG for_O Chechnya_B-LOC, Sergei_B-PER Yastrzhembski_I-PER.*

2.2.2 Named Entity Linking

In order to evaluate the performance of NEL approaches, we have collected datasets and benchmarks for NEL (Table 2). Some of these datasets were built automatically, semi-automatically or manually. Most available benchmarks are in English (AIDA, AQUAINT, ACE2004, WIKIPEDIA, MSNBC, and TAC2010). Statistics of English datasets are presented in Table 3.

McN-dataset was built from parallel document collections and crowdsourcing to generate ground truth in other languages (McNamee et al., 2011). The amount of queries and non-NIL[5] mentions on the McN-dataset are listed in Table 4. Finally, TH-dataset (Table 5) was built from documents of Wikipedia by taking the anchors (hyperlinked texts) as the query mentions and the corresponding English Wikipedia titles as the answers (Tsai & Roth, 2016).

Both the TAC2015 training data source corpus (444 documents) and evaluation data source corpus (500 documents) are comprised of approximately half newswire documents and half discussion forum threads in Chinese, English and Spanish (Ellis et al., 2015). See the statistics of this dataset in Table 6.

Wikipedia is a multi-lingual resource that currently hosts 294 languages and contains annotated markups and rich informational structures through crowd-sourcing. In this resource, name mentions are often labeled as anchor links to their corresponding referent pages (Pan et al., 2017). Taking advantage of this feature, Pan et al. (2017) developed an independent language framework to extract name mentions from Wikipedia articles in 282 languages and link them to the English Wikipedia (Wikiann dataset, Tables 2

[3] The sentence was originally written in Russian but automatically translated to English for readability purposes.

[4] Related to the events occurred on 30/06/2018. More details at `https://en.wikipedia.org/wiki/2018_Amesbury_poisonings`.

[5] NIL corresponds to the entity mention which entity record doesn't exist in the KB (Shen et al., 2015). In the contrary, non-NIL corresponds to any existing entity in the KB.

Table 1: The collected datasets for the NER task and their properties: acronym, name, year of publication, availability, languages, and link to the corpus location.

Acronym	Name	Year	Public	Language	Location
FIN-CLARIN	Finnish News Corpus for Named Entity Recognition (Ruokolainen et al., 2019)	2019	yes	fi	link
282NER	Cross-lingual name tagging and linking for 282 languages	2017	yes	bg, cs, de, en, et, fi, hr, lt, lv, pl, ru, sk, sl, sr, sv, uk	link
SlavicNER2017	1st shared task on multilingual named entity recognition	2017	yes	cs, hr, pl, ru, sk, sl, uk	link
SlavicNER2019	2nd shared task on multilingual named entity recognition	2019	yes	bg, cs, pl, ru	link
SETimes.HR+	The SETimes.HR+ Croatian dependency treebank	2013	yes	hr, sr	link link
GermEVAL2014	GermEval 2014 Named Entity Recognition Shared Task	2014	yes	de	link
KaggleNER	Annotated Corpus for Named Entity Recognition	2017	yes	en	link
EstNER	Estonian NER corpus	2013	yes	et	link
Finer-data	A Finnish News Corpus for Named Entity Recognition	2014	yes	fi	link
hr500k	Training corpus hr500k 1.0	2018	yes	hr	link link
TildeNER	accurat-toolkit/TildeNER	2012	yes	lt	link
LVTagger	PeterisP/LVTagger/NerTrainingData/	2013	yes	lv	link
factRuEval-2016	factRuEval-2016 dialog-21.ru	2016	yes	ru	link
ssj500k	Training corpus ssj500k 2.2	2019	yes	sl	link link
Slovene news	Slovene news - slavko.zitnik	2011	yes	sl	link
SwedishNER	Swedish manually annotated NER	2012	yes	sv	link

Table 2: The collected benchmarks for the NEL task.

Name	Year	Public	Language	Location
AIDA	2003	Yes	en	link
AQUAINT	2008	Yes	en	link
ACE2004	2011	Yes	en	link
CLUEWEB	2013	Yes	en	link
MSNBC	2007	Yes	en	link
WIKIPEDIA	2011	Yes	en	link
TAC2010	2010	No	en	link
McN-dataset	2011	Yes	ar, bg, cs, da, de, el, es, fi, fr, hr, it, mk, nl, pt, ro, sq, sr, sv, tr, ur, zh	link
TAC2015	2015	No	en, es, zh	link
TH-dataset	2016	Yes	ar, de, es, fr, he, it, ta, th, tl, tr, ur, zh	link
Wikiann	2017	Yes	282 languages	link

Table 3: Statistics of English datasets for the NEL task.

Dataset	#mentions	#docs	#mentions/doc
AIDA-train	18448	946	19.5
AIDA-A (valid)	4791	216	22.1
AIDA-B (test)	4485	231	19.4
AQUAINT	727	50	14.5
ACE2004	257	36	7.1
CLUEWEB	11154	320	34.8
MSNBC	656	20	32.8
WIKIPEDIA	6821	320	21.3

Table 4: Amount of queries and non-NIL on McN-dataset (McNamee et al., 2011) for the NEL task.

Language	Collection	#Queries	#Non-NIL
Albanian	SETimes	4190	2274
Arabic	LDC2004T18	2829	661
Bulgarian	SETimes	3737	2068
Chinese	LDC2005T10	1958	956
Croatian	SETimes	4139	2257
Czech	ProjSynd	1044	722
Danish	Europarl	2105	1096
Dutch	Europarl	2131	1087
Finnish	Europarl	2038	1049
French	ProjSynd	885	657
German	ProjSynd	1086	769
Greek	SETimes	3890	2129
Italian	Europarl	2135	1087
Macedonian	SETimes	3573	1956
Portuguese	Europarl	2119	1096
Romanian	SETimes	4355	2368
Serbian	SETimes	3943	2156
Spanish	ProjSynd	1028	743
Swedish	Europarl	2153	1107
Turkish	SETimes	3991	2169
Urdu	LDC2006E110	1828	1093

and 7). Despite being automatically built and does not contain all types of named entities, this dataset contains all languages of the EMBEDDIA project and can provide an analysis of the performance and limitations of cross-lingual NEL systems in less-resourced languages.

2.3 Evaluation measures

When designing NER and NEL systems, it is important to evaluate their effectiveness. Several measures have been defined to evaluate the performance of NER and NEL systems. These performances are usually measured in terms of recall, precision and F-measure or in terms of Slot Error Rate (SER) (Makhoul et al., 1999). Several weighted variants exist in the literature to calculate these measures.

Precision, recall and F1-measure analyze the performance of systems by analyzing the number of true positives, false positives, and false negatives. In the context of NER, these notions are defined as:

Table 5: Amount of training and test mentions of the TH benchmark (Tsai & Roth, 2016) for the NEL task.

Language	#Training	#Test
German	23124	9798
Spanish	30471	12153
French	37860	14358
Italian	34185	12775
Chinese	44246	11394
Hebrew	20223	16146
Thai	16819	11381
Arabic	22711	10646
Turkish	12942	13798
Tamil	21373	11346
Tagalog	4835	1074
Urdu	1413	1389

Table 6: The TAC2015 datasets (Ellis et al., 2015) for the NEL task.

Language	Train #docs (#news | #discussion)	Test #docs (#news | #discussion)
Chinese	147 (84 ‖ 63)	166 (84 ‖ 82)
English	168 (85 ‖ 83)	167 (82 ‖ 85)
Spanish	129 (82 ‖ 47)	167 (84 ‖ 83)

- **True positive (TP)**: represents the number of NEs correctly annotated.
- **False positive (FP)**: corresponds the incorrectly tagged entities.
- **False negatives (FN)**: are the ground truth entities that were not tagged.
- **Precision**: is the number of named entities correctly labeled compared to the number of tagged entities by a system.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

- **Recall**: is the number of named entities correctly labeled compared to the number of tagged entities in the reference.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

- **F-measure (F1)**: is defined as the harmonic mean of precision and recall.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

In the context of NEL, true positive corresponds to the correctly discovered and annotated mentions of NEs in a KB, false positive corresponds to incorrect mentions or entity annotation to a KB, and false negatives are ground truth mentions or entities that were either not annotated or annotated incorrectly. These measures can be calculated on a full corpus (micro-averaging) or averaged by document (macro-averaging). Since knowledge bases contain millions of entities, only mention-entity pairs where the ground-truth gives a known entity are analyzed.

On the other hand, the SER combines and weights the different types of errors on the NER task. Several metrics defined according to the types of errors have been taken into account. Weights are assigned to each type of error. We present here the metrics used to calculate the SER.

Table 7: Amount of name mentions on the Wikiann dataset (Pan et al., 2017) for the NEL task.

Lang.	#ment.	Lang.	#ment.	Lang.	#ment.	Lang.	#ment.	Lang.	#ment.	Lang.	#ment.
en	12M	tt	53K	nds	7.0K	kw	2.8K	kl	1.5K	chr	530
ja	1.9M	gl	52K	gd	6.7K	ilo	2.8K	sn	1.5K	ha	517
sv	1.8M	ka	49K	mrj	6.7K	mwl	2.7K	av	1.4K	ab	506
de	1.7M	vo	47K	so	6.5K	mai	2.7K	as	1.4K	got	506
fr	1.4M	lmo	39K	co	6.0K	nv	2.7K	stq	1.4K	bi	490
ru	1.4M	be	38K	pnb	6.0K	sd	2.7K	gv	1.3K	st	455
it	1.2M	mk	35K	pcd	5.8K	os	2.7K	wo	1.3K	chy	450
sh	1.1M	cy	32K	wa	5.8K	mzn	2.6K	xal	1.3K	iu	450
es	992K	bs	31K	frr	5.7K	azb	2.6K	nrm	1.3K	zu	449
pl	931K	ta	31K	scn	5.6K	bxr	2.6K	na	1.2K	pnt	445
nl	801K	hy	28K	fo	5.4K	vec	2.6K	ltg	1.2K	ik	436
zh	718K	bn	27K	ckb	5.3K	bo	2.6K	pam	1.2K	lrc	416
pt	576K	az	26K	li	5.2K	yi	2.6K	lad	1.2K	bm	386
uk	472K	sq	26K	nap	4.9K	frp	2.5K	tet	1.2K	za	382
cs	380K	ml	24K	crh	4.9K	myv	2.5K	sc	1.2K	mo	373
sr	365K	br	22K	gu	4.6K	se	2.5K	wuu	1.2K	ss	362
hu	357K	z-y	22K	km	4.6K	cdo	2.5K	ksh	1.2K	ee	297
fi	341K	af	21K	tg	4.5K	nso	2.5K	pfl	1.1K	dz	262
no	338K	b-x	20K	hsb	4.5K	gom	2.4K	haw	1.1K	ak	258
fa	294K	tl	19K	c-z	4.5K	ky	2.4K	am	1.1K	sg	245
ko	273K	oc	18K	jv	4.4K	n-n	2.3K	bcl	1.1K	ts	236
ca	265K	mr	18K	lez	4.4K	ne	2.3K	nah	1.1K	rn	185
tr	223K	bar	17K	hak	4.3K	sa	2.2K	udm	1.1K	ve	183
ro	197K	cv	15K	ang	4.2K	mt	2.2K	su	1.1K	ny	169
bg	186K	ba	15K	r-t	4.2K	my	2.2K	dsb	1.1K	ff	168
ar	185K	mg	14K	kn	4.1K	bh	2.2K	tpi	1.1K	ch	159
id	150K	hi	14K	csb	4.1K	vls	2.2K	lo	1.0K	xh	141
he	145K	an	14K	lij	4.1K	ug	2.1K	bpy	1.0K	fj	126
eu	137K	als	14K	nov	4.0K	si	2.1K	ki	1.0K	ks	124
da	133K	sco	14K	ace	4.0K	kaa	2.1K	ty	1.0K	ti	52
vi	125K	bug	13K	gn	4.0K	b-s	2.1K	hif	1.0K	cr	49
th	96K	lb	13K	koi	4.0K	krc	2.1K	ady	979	pi	41
sk	93K	fy	13K	mhr	3.9K	ie	2.1K	ig	968		
uz	92K	new	12K	io	3.8K	dv	2.0K	tyv	903		
eo	85K	ga	12K	min	3.8K	xmf	2.0K	tn	902		
la	81K	ht	12K	arz	3.8K	rue	1.9K	cu	898		
z-m	79K	war	12K	ext	3.7K	pa	1.8K	sm	888		
lt	79K	te	11K	yo	3.7K	eml	1.8K	to	866		
el	78K	is	11K	sah	3.6K	arc	1.8K	tum	831		
ce	77K	pms	10K	vep	3.5K	pdc	1.8K	r-r	750		
ur	77K	zea	10K	ku	3.3K	kbd	1.7K	om	709		
hr	76K	sw	9.3K	kab	3.3K	pap	1.7K	glk	688		
ms	75K	ia	8.9K	szl	3.0K	jbo	1.7K	lbe	651		
et	69K	qu	8.7K	tk	2.9K	diq	1.7K	bjn	640		
kk	68K	ast	8.3K	z-c	2.9K	pag	1.7K	srn	619		
ceb	68K	rm	8.0K	mn	2.9K	kg	1.6K	mdf	617		
sl	67K	ay	7.9K	kv	2.9K	m-b	1.6K	tw	572		
nn	65K	ps	7.7K	f-v	2.9K	rw	1.6K	pih	555		
sim	59K	mi	7.5K	gan	2.9K	or	1.6K	rmy	551		
lv	57K	gag	7.3K	fur	2.8K	ln	1.6K	lg	530		

- Insertion [I]: the number of entities detected having no common word with an entity of the reference.
- Deletion [D]: the number of named entities in the reference fully undetected by the system.
- Type [T] errors: the number of entities detected with correct boundaries but an incorrect category.
- Border errors [F]: the number of entities detected with a correct category but incorrect boundaries.

- Type and border errors [TF]: the number of entities detected with incorrect category and boundaries.

$$SER = \frac{I + D + 0,5 \times T + 0,5 \times F + 0,8 \times TF}{number\ \ of\ \ NEs\ in\ the\ reference} \tag{4}$$

3 Task T2.2: Multilingual keyword extraction and matching

3.1 Background

In Task 2.2, we will use monolingual and multilingual methods to extract keywords and topical terms from the text. We identify datasets and benchmarks and evaluation measures related to keyword extraction, but as terminology extraction is a closely related task, we also present the corpora relevant for this task.

3.1.1 Keyword extraction

Keywords are terms (i.e. expressions) that best describe the subject of a document (Beliga et al., 2015). A good keyword effectively summarizes the content of the document and allows it to be efficiently retrieved when needed. Traditionally, keyword assignment was a manual task, but with the emergence of large amounts of textual data, automatic keyword extraction methods have become indispensable.

Despite a considerable effort from the research community, state-of-the-art keyword extraction algorithms leave much to be desired and their performance is still lower than on many other core NLP tasks (Hasan & Ng, 2014). The first keyword extraction methods mostly followed a supervised approach (Hulth, 2003; Nguyen & Luong, 2010; Witten et al., 2005): they first extract keyword features and then train a classifier on a gold standard dataset. For example, KEA (Witten et al., 2005), a state of the art supervised keyword extraction algorithm is based on the Naive Bayes machine learning algorithm. While these methods offer good performance, they rely on an annotated gold standard dataset and require a (relatively) long training process. In contrast, unsupervised approaches need no training and can be applied directly without relying on a gold standard document collection. They can be further divided into statistical and graph-based methods. The former, such as YAKE (Campos et al., 2018b,a), KP-MINER (El-Beltagy & Rafea, 2009) and RAKE (Rose et al., 2010), use statistical characteristics of the texts to capture keywords, while the latter, such as Topic Rank (Bougouin et al., 2013), TextRank (Mihalcea & Tarau, 2004), Topical PageRank (Sterckx et al., 2015) and Single Rank (Wan & Xiao, 2008), build graphs to rank words based on their position in the graph. From statistical approaches, a state-of-the-art keyword extraction algorithm is YAKE (Campos et al., 2018b), which is also one of the best performing keyword extraction algorithms overall; it defines a set of five features capturing keyword characteristics which are heuristically combined to assign a single score to every keyword. On the other hand, from graph-based approaches, Topic Rank (Bougouin et al., 2013) can be considered state-of-the-art; candidate keywords are clustered into topics and used as vertices in the final graph, used for keyword extraction. Next, a graph-based ranking model is applied to assign a significance score to each topic and keywords are generated by selecting a candidate from each of the top-ranked topics.

With the emergence of deep learning for NLP tasks, neural networks have also been applied to keyword extraction. The 2017 SemEval workshop (Augenstein et al., 2017) included a task on extracting keyphrases and relations from scientific publications featuring 54 teams in the development phase and

26 teams in the final competition proving that keyword extraction is a popular information extraction topic. The best results have been achieved with recurrent neural networks (RNN), e.g., TTI-COIN (Tsujimura et al., 2017) or s2_end2end (Ammar et al., 2017) which uses an additional CRF (conditional random fields) layer on top of the RNN network. Other methods were based on random forests (Wang & Li, 2017), support vector machines (Wang & Li, 2017), conditional random fields (Barik & Marsi, 2017; Berend, 2017; Prasad & Kan, 2017) etc.

Keyword matching, i.e. aligning extracted keywords across languages, has not yet been much addressed in this report, however we address the closely related task of terminology alignment (see below).

3.1.2 Terminology extraction and alignment

Terminology extraction refers to structuring terminological knowledge from unstructured text, and even if terminology extraction refers to specialized corpora, the keyword and terminology extraction tasks are very closely related, so we cover also the datasets for terminology extraction. In terms of input text, we can distinguish between monolingual terminology extraction, where terms are extracted from text in one language, and bilingual or multilingual terminology extraction, where the goal is to extract and align terms from text in two or more languages (Repar, 2019). At the highest level, bilingual terminology extraction can be divided into extraction from comparable and extraction from parallel corpora, where parallel corpora are composed of source texts and their translations in one or more different languages, while comparable corpora are composed of monolingual texts collected from different languages using similar sampling techniques (McEnery et al., 2006). Terminology matching (or alignment) can be considered a subfield of bilingual (i.e. multilingual) terminology extraction and is the process of aligning terms between two candidate term lists in two languages. Bilingual terminology alignment has a narrower focus than bilingual terminology extraction, but note that the two terms are often used interchangeably in various papers. We will explore machine learning approaches to term matching, such as Aker et al. (2013), and try to transfer them to the field of keyword matching.

3.2 Available resources

This section contains datasets for keyword extraction and terminology extraction.

3.2.1 Keyword extraction

Public datasets

We have identified several publicly available datasets for keyword extraction, including SemEval 2010 and SemEval 2017 shared tasks. Detailed dataset descriptions and statistics can be found in Table 8, while full statistics and files for download can be found online[6] (Campos et al., 2018b). Most datasets are from the domain of computer science or contain multiple domains. They are very diverse in terms of the number of documents—ranging from fao30 with 30 documents to KP20K with 570,809 documents, in terms of the average number of gold standard keywords per document—from 5.28 in KP20K to 48.92 in 500N-KPCrowd-v1.1—and in terms of the average length of the documents—from 75.97 in kdd to SemEval2017 with 8332.34. These datasets can be used for training supervised models, as well as evaluation of unsupervised methods.

[6] https://github.com/LIAAD/KeywordExtractor-Datasets. Last accessed: August 21, 2019

Table 8: Keyword extraction datasets. The labels in column *Doc. type* indicate the type of documents contained in the dataset, e.g. *Abstract* means paper abstracts, *Paper* means full papers.

Name	Lang	Doc. type	Desc	No. docs	Avg. keywords	Avg. doc. length
110-PT-BN-KP	PT	News	Broadcast news transcriptions	110	23.73	304
500N-KPCrowd-v1.1	EN	News	Broadcast news transcriptions	500	48.92	408.33
Inspec	EN	Abstract	Abstract Scientific journal papers from Computer Science collected between 1998 and 2002	2000	14.62	128.2
Krapivin2009	EN	Paper	Computer Science domain papers published by ACM	2304	6.34	8040.74
Nguyen2007	EN	Paper	Scientific conference papers	209	11.33	5201.09
PubMed	EN	Paper	Full-text papers collected from PubMed Central	500	15.24	3992.78
Schutz2008	EN	Paper	Full-text papers collected from PubMed Central	1231	44.69	3901.31
SemEval2010	EN	Paper	Scientific papers from the ACM Digital Library	243	16.47	8332.34
SemEval2017	EN	Paragraph	500 paragraphs selected from 500 ScienceDirect journal articles	493	18.19	178.22
WikiNews	FR	News	100 WikiNews articles	100	11.77	293.52
cacic	ES	Paper	Scientific articles published in the Argentine Congress of Computer Science	888	4.82	3985.84
citeulike180	EN	Paper	Full-text papers from the CiteULike.org	183	18.42	4796.08
fao30	EN	Paper	Agricultural documents from two datasets based on UN FAO	30	33.23	4777.7
fao780	EN	Paper	Agricultural documents from two datasets based on UN FAO	779	8.97	4971.79
kdd	EN	Paper	Abstracts from the ACM Conference on Knowledge Discovery and Data Mining (KDD)	755	5.07	75.97
pak2018	PL	Abstract	Abstracts of journals on technical topics collected from Measurement Automation and Monitoring	50	4.64	97.36
theses100	EN	Msc/Phd Thesis	Full master and Ph.D. theses from the University of Waikato	100	7.67	4728.86
wicc	ES	Paper	Scientific articles of the Workshop of Researchers in Computer Science	1640	4.57	1955.56
wiki20	EN	Research Report	Computer science technical research reports	20	36.50	6177.65
www	EN	Paper	Abstracts of WWW conference papers from 2004-2014	1330	5.80	84.08

Media partners datasets

The analysis of the publicly available resources revealed that there is a general lack of datasets involving project languages. However, project partners have keyword related datasets, that will be used to evaluate the keyword extraction methods.

Trikoder provides datasets from Styria Media Group, in particular the news from the Croatian media 24sata and one for Vecernji list. Article keywords at the 24sata and Vecernji list are manually chosen by the authors when writing the article text. The recommendation engine is used by 24sata that, based on the article text, suggests some keywords from the database of existing keywords. Journalists can pick recommended keywords, they can search from the database of existing keywords or they can type in new keywords. At Vecernji list, the recommendation engine is not used and their journalist can search their database of existing keywords or type in new keywords.

Ekspress Meedia tags are assigned manually to the news articles by the editors. Before adding the tags, editors get tags suggestion at the article admin system from which they choose more relevant ones. There are appr. 65 000 tags altogether at their admin system. There are plans that the cleaning and reducing of the tag set will be performed during the EMBEDDIA project. As Ekspress Meedia dataset contains articles in Estonian and Russian, it is a good candidate to be used also for evaluating keyword matching systems.

STT uses the global IPTC (International Press Telecommunications Council) system to arrange the metadata on subjects and topics. IPTC develops and promotes technical standards to improve the management and exchange of information between content providers, intermediaries and consumers. Its members include news agencies, publishers and industry vendors.

Every STT news article should have at least one (but preferably more) of these IPTC-keywords included. The journalists choose the keywords for their stories (there is no automation involved in this part of the work flow), but the keywords are limited to those existing in the IPTC-tree.

3.2.2 Terminology extraction and alignment

An ideal terminology extraction and alignment dataset would consist of a bilingual or multilingual (parallel or comparable) corpus along with reference (gold standard) term lists containing terms that can be found in the corpus. Such corpora are TTC[7] wind energy and TC mobile technology, which contain data for six languages (English, French, German, Spanish, Russian, Latvian, Chinese), or the Bitter corpus, which contains data for the EN-IT language pair. The first was used in Hazem & Morin (2016), while the second was used by Arčan et al. (2014). Since such datasets are scarce, researchers employ various methodologies for constructing their own datasets. One method, used by Aker et al. (2013), is to take one of the available multilingual translation memories containing EU documentation (such as Europarl (Koehn, 2005) or DGT (Steinberger et al., 2013)) as the corpus and a glossary (e.g., IATE (Johnson & Macphail, 2000)) or thesaurus (e.g., Eurovoc (Steinberger et al., 2002) as the terminology gold standard list. Another strategy, used by Hazem & Morin (2017), is to collect a parallel corpus manually (i.e. scientific articles in French and English from the Elsevier[8] website) and a domain specific terminological resource (i.e. ULMA[9]) as a reference termlist. Hazem & Morin (2017) also filter out those terms from the termlist that do not appear often enough in their corpus. In other cases (e.g., Haque et al. (2014)), the datasets are not available because the papers were written as part of industrial projects and the datasets are private. A somewhat different approach was taken by the authors of the KAS biterm dataset (Ljubešic et al., 2018) who annotated bilingual terms (EN-SL) in monolingual scientific publications in Slovene (by taking advantage of the patterns used to explain English terms to a Slovene audience).

In terms of strictly monolingual terminology extraction datasets, there are also a few available, such as the ACL RD-TEC 2.0 (QasemiZadeh & Schumann, 2016) for English, as well as CESART for French (El Hadi et al., 2004).

3.3 Evaluation measures

The performance of keyword and term extraction methods is usually evaluated against a gold standard set using standard measures of precision, recall and F1 measure, where a gold standard set of keywords means that each document in a dataset has a list of (manually) curated keywords against which the (automatically) extracted keywords are compared. Benchmark results gathered by authors of YAKE Campos et al. (2018b), can be accessed in their online repository[10]. For the media partner datasets,

[7] http://www.lina.univ-nantes.fr/?Reference-Term-Lists-of-TTC.html

[8] https://www.elsevier.com/

[9] https://www.nlm.nih.gov/research/umls/

[10] https://github.com/LIAAD/yake

Table 9: Terminology extraction and alignment datasets.

Name	Lang	Gold standard terms	Domain
ACL RD-TEC 2.0	EN	6818	scientific articles
CESART Medicine	FR	22,861	medicine
CESART Education	FR	36,081	education
TTC wind energy	EN,FR,DE,ES,RU,LT,CH	277	wind energy
TTC mobile	EN,FR,DE,ES,RU,LT,CH	263	mobile technology
Bitter	EN,IT	237	information technology
Eurovoc+DGT	26 languages	around 7000	various
KAS biterm	EN,SL	3,732	scientific publications

the tags assigned to the articles will be considered as gold standard. As the document usually have a varying numbers of keywords assigned, a cutoff can distort the results. To tackle this problem Zesch & Gurevych (2009) propose to use the R-precision measure from information retrieval, which is translated to the keyword extraction task. R-p is the precision when the number of retrieved keyphrase matchings equals the number of gold standard keyphrases assigned to the document. An R-precision of 1.0 is equivalent to perfect keyphrase ranking and perfect recall. In R-p, the focus is on the precision on the first ranks.

Another option for the evaluation of term and keyword extraction is manual evaluation, where usually the focus is on precision.

The SemEval2017 task (Augenstein et al., 2017) introduced two additional evaluation settings alongside classic keyword identification described above: subtask B dealt with keyword classification (as process, task or material) and subtask C dealt with semantic relation extraction between keywords ("hyponym of" and "synonym of").

For keyword and term alignment, same measures can be used (precision, recall, F1 measure), if we have a gold standard list of aligned terms (for example Eurovoc is frequently used for term alignment). In manual evaluation, the focus is usually on precision. We also list the categories for manual evaluation used in Aker et al. (2014):

- **1 - Equivalence**: The terms are exact translations/transliterations of each other.
- **2 - Inclusion**: Not an exact translation/transliteration, but an exact translation/transliteration of one term is entirely contained within the term in the other language
- **3 - Overlap**: Not category 1 or 2, but the terms share at least one translated/transliterated word.
- **4 - Unrelated**: No word in either term is a translation/transliteration of a word in the other (e.g., *level*).

4 Task T2.3: Multilingual natural language generation

Task 2.3 develops basic technology for multilingual natural language generation (NLG) during M1–M18. WP5, 'Multilingual text generation', then builds on this work during M9–M33. The work on resource gathering, benchmarking and evaluation for NLG is joint for T2.4 and T5.4; T5.4 produces deliverable D5.1 on resource gathering, benchmarking and evaluation for NLG specifically. Given that WP5 carries the major responsibility for NLG and especially for its applications, we provide a full account of the datasets, evaluation methods and benchmarks for multilingual natural language generation only in Deliverable D5.1. We next provide the key conclusions from D5.1.

Natural language generation employs data in multiple roles. First, structured data is needed to act as the input of the system. Second, pairs of structured data and human-written texts based on that data can be used for both automated evaluation and for training an 'End-to-End' NLG system. Third, a wide variety of different datasets can be used to train machine learning components for use as sub-components of the NLG system. For use in this task, we have identified several datasets. A large dataset from Eurostat is to be used as system input, as it fulfills acceptably both scientific/technical and journalistic requirements. Second, in addition to word embedding models produced by the project partners, a corpus of news texts from STT is used as a starting point for research into incorporating machine learning components, and also as a source of qualitative examples.

No suitable dataset consisting of aligned input-output pairs has been located in the selected domain and producing such a dataset is prohibitively expensive. Due to this lack of an aligned corpus of inputs and outputs, the evaluation of the NLG methods is to be done using intrinsic human evaluations, where judges evaluate the output of case-study NLG systems on the 'Credibility', 'Liking', 'Quality', and 'Representativeness' axes. These axes, identified by Sundar (1999) are closest to a standard that research into automated news production has. At the same time, if suitable datasets for automated evaluation are produced by some third party, we intend to use them to measure system performance using standard evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) alongside these human evaluations. While other metrics such as METEOR (Lavie & Agarwal, 2007) and CIDEr (Vedantam et al., 2015) are often used in NLG research, they suffer from a series of flaws with regard to the multilingual setup of the work conducted in this task. Where human evaluations are limited by availability of online judges speaking the relevant languages, we will conduct quantitative evaluations where possible and complement those results with qualitative analyses.[11]

Notably, research into NLG, as applied to news production, is complicated and the setups not standardized. As such, we are not aware of any other works that would be directly applicable as quantitative benchmarks, i.e. by directly comparing numeric values obtained as results to determine which system is 'best'.

Please see Deliverable D5.1, 'Datasets, Benchmarks and evaluation metrics for multilingual natural language generation' for more details.

5 Conclusions and further work

We introduced resources collected to build and evaluate tasks T2.1 (named entity recognition and linking, and event detection), T2.2 (multilingual keyword extraction and matching), and T2.3 (multilingual natural language generation) in WP2 of the EMBEDDIA project. The collected resources contain large data in several languages to evaluate existing approaches. We intend that the list of datasets and methods will be extended during the project, resulting in the final deliverable of the present task, delivered at M30 (June 2021) as D2.5.

[11] Please refer to Deliverable D5.1 for more details.

References

Aker, A., Paramita, M., & Gaizauskas, R. (2013). Extracting bilingual terminologies from comparable corpora. In *Proceedings of acl* (Vol. 1, pp. 402–411).

Aker, A., Paramita, M. L., Pinnis, M., & Gaizauskas, R. (2014). Bilingual dictionaries for all eu languages. In *Lrec 2014 proceedings* (pp. 2839–2845).

Ammar, W., Peters, M., Bhagavatula, C., & Power, R. (2017). The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 592–596).

Arčan, M., Turchi, M., Tonelli, S., & Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a CAT environment.. doi: https://doi.org/10.13140/2.1.1019.8404

Asahara, M., & Matsumoto, Y. (2003). Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1* (pp. 8–15).

Augenstein, I., Das, M., Riedel, S., Vikraman, L., & McCallum, A. (2017). Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*.

Barik, B., & Marsi, E. (2017). Ntnu-2 at semeval-2017 task 10: Identifying synonym and hyponym relations among keyphrases in scientific documents. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 965–968).

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences*, *39*(1), 1–20.

Berend, G. (2017). SZTE-NLP at SemEval-2017 task 10: A high precision sequence model for keyphrase extraction utilizing sparse coding for feature generation. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 990–994). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S17-2173 doi: 10.18653/v1/S17-2173

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1998). Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 acm sigmod international conference on management of data* (pp. 1247–1250). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1376616.1376746 doi: 10.1145/1376616.1376746

Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (ijcnlp)* (pp. 543–551).

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018a). A text feature based automatic keyword extraction method for single documents. In G. Pasi, B. Piwowarski,

L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 684–691). Cham: Springer International Publishing.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018b). Yake! collection-independent automatic keyword extractor. In *European conference on information retrieval* (pp. 806–810).

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, *12*(Aug), 2493–2537.

Dernoncourt, F., Lee, J. Y., & Szolovits, P. (2017). Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*.

Eftimov, T., Seljak, B. K., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, *12*(6), e0179488.

El-Beltagy, S. R., & Rafea, A. (2009). Kp-miner: A keyphrase extraction system for english and arabic documents. *Information Systems*, *34*(1), 132–144.

El Hadi, W. M., Timimi, I., & Dabbadie, M. (2004). Evalda-cesart project: Terminological resources acquisition tools evaluation campaign. In *Lrec.*

Ellis, J., Getman, J., Fore, D., Kuster, N., Song, Z., Bies, A., & Strassel, S. M. (2015). Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Proceedings of the 2015 text analysis conference, TAC 2015, gaithersburg, maryland, usa.*

Filannino, M., Brown, G., & Nenadic, G. (2013). Mantime: Temporal expression identification and normalization in the tempeval-3 challenge. *arXiv preprint arXiv:1304.7942*.

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

Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In *Coling 1996 volume 1: The 16th international conference on computational linguistics* (Vol. 1).

Haque, R., Penkale, S., & Way, A. (2014). Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In *Proceedings of the 4th international workshop on computational terminology (computerm)* (pp. 42–51).

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of acl* (pp. 1262–1273). Retrieved from `http://www.aclweb.org/anthology/P14-1119`

Hazem, A., & Morin, E. (2016). Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of coling* (pp. 3401–3411).

Hazem, A., & Morin, E. (2017). Bilingual word embeddings for bilingual terminology extraction from specialized comparable corpora. In *Proceedings of ijcnlp* (pp. 685–693).

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 216–223).

Johnson, I., & Macphail, A. (2000). IATE–Inter-Agency Terminology Exchange: Development of a single central terminology database for the institutions and agencies of the European Union. In *Proceedings of the workshop on terminology resources and computation, lrec 2000 conference.*

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th machine translation summit* (Vol. 5, pp. 79–86).

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`

Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 228–231).

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`

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., . . . Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, *6*(2), 167–195.

Lejeune, G., Brixtel, R., Lecluze, C., Doucet, A., & Lucas, N. (2013). Added-value of automatic multilingual text analysis for epidemic surveillance. In *Conference on artificial intelligence in medicine in europe* (pp. 284–294).

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).

Ljubešic, N., Erjavec, T., & Fišer, D. (2018). Kas-term and kas-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing.

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).

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

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. Retrieved from `https://www.aclweb.org/anthology/I11-1029`

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing.*

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Nguyen, T. D., & Luong, M.-T. (2010). Wingnus: Keyphrase extraction utilizing document logical structure. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 166–169).

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.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Peters, M., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1756–1765).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (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)* (pp. 2227–2237).

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Přibáň, P., Steinberger, J., & Yangarber, R. (2019). The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th workshop on balto-slavic natural language processing* (pp. 63–74). 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.* Association for Computational Linguistics.

Prasad, A., & Kan, M.-Y. (2017). Wing-nus at semeval-2017 task 10: Keyphrase extraction and classification as joint sequence labeling. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 973–977).

QasemiZadeh, B., & Schumann, A.-K. (2016). The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the tenth international conference on language resources and evaluation (lrec 2016)* (pp. 1862–1868).

Raiman, J., & Raiman, O. (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* (pp. 5406–5413).

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157–176). Springer.

Repar, M. M. P. S., Andraž. (2019). Replication, analysis and adaptation of a term alignment approach. *Article submitted for publication*.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1–20.

Ruokolainen, T., Kauppinen, P., Silfverberg, M., & Lindén, K. (2019). A finnish news corpus for named entity recognition. *Language Resources and Evaluation*. Retrieved from `https://doi.org/10.1007/s10579-019-09471-7` doi: 10.1007/s10579-019-09471-7

Sekine, S. (1998). Nyu: Description of the japanese ne system used for met-2. In *Proc. of the seventh message understanding conference (muc-7).*

Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, *27*(2), 443-460. doi: 10.1109/TKDE.2014.2327028

Sil, A., Kundu, G., Florian, R., & Hamza, W. (2018). Neural cross-lingual entity linking. In *Thirty-second aaai conference on artificial intelligence.*

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th international conference on language resources and evaluation (lrec'2012).*

Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. *Computational Linguistics and Intelligent Text Processing*, 101–121.

Sterckx, L., Demeester, T., Deleu, J., & Develder, C. (2015). Topical word importance for fast keyphrase extraction. In *Proceedings of the 24th international conference on world wide web* (pp. 121–122).

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1242572.1242667` doi: 10.1145/1242572.1242667

Sundar, S. S. (1999). Exploring receivers' criteria for perception of print and online news. *Journalism & Mass Communication Quarterly*, *76*(2), 373–386.

Tsai, C.-T., & Roth, D. (2016). Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 589–598). San Diego, California: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N16-1072 doi: 10.18653/v1/N16-1072

Tsujimura, T., Miwa, M., & Sasaki, Y. (2017). Tti-coin at semeval-2017 task 10: Investigating embeddings for end-to-end relation extraction from scientific papers. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 985–989).

Upadhyay, S., Gupta, N., & Roth, D. (2018). Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2486–2495). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D18-1270

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566–4575).

Wan, X., & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the 23rd national conference on artificial intelligence - volume 2* (pp. 855–860). AAAI Press.

Wang, L., & Li, S. (2017). Pku_icl at semeval-2017 task 10: keyphrase extraction with model ensemble and external knowledge. In *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* (pp. 934–937).

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (2005). Kea: Practical automated keyphrase extraction. In *Design and usability of digital libraries: Case studies in the asia pacific* (pp. 129–152). IGI Global.

Zesch, T., & Gurevych, I. (2009). Approximate Matching for Evaluating Keyphrase Extraction. In *Proceedings of the 7th international conference on recent advances in natural language processing* (pp. 484–489). Borovets, Bulgaria. Retrieved from http://www.aclweb.org/anthology/R09-1086