The First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages

Piskorski, Jakub

Association for Computational Linguistics
2017

The First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages

Jakub Piskorski[1], Lidia Pivovarova[2], Jan Šnajder[3], Josef Steinberger[4], Roman Yangarber[2]
[1]Joint Research Centre, Ispra, Italy, first.last@jrc.ec.europa.eu
[2]University of Helsinki, Finland, first.last@cs.helsinki.fi
[3]University of Zagreb, Croatia, first.last@fer.hr
[4]University of West Bohemia, Czech Republic, jstein@kiv.zcu.cz

Abstract

This paper describes the outcomes of the First Multilingual Named Entity Challenge in Slavic Languages. The Challenge targets recognizing mentions of named entities in web documents, their normalization/lemmatization, and cross-lingual matching. The Challenge was organized in the context of the 6th Balto-Slavic Natural Language Processing Workshop, colocated with the EACL-2017 conference. Eleven teams registered for the evaluation, two of which submitted results on schedule, due to the complexity of the tasks and short time available for elaborating a solution. The reported evaluation figures reflect the relatively higher level of complexity of named entity tasks in the context of Slavic languages. Since the Challenge extends beyond the date of the publication of this paper, updates to the results of the participating systems can be found on the official web page of the Challenge.

1 Introduction

Due to the rich inflection, derivation, free word order, and other morphological and syntactic phenomena exhibited by Slavic languages, analysis of named entities (NEs) in these languages poses a challenging task (Przepiórkowski, 2007; Piskorski et al., 2009). Fostering research and development on detection and lemmatization of NEs—and the closely related problem of entity linking—is of paramount importance for enabling effective multilingual and cross-lingual information access in these languages.

This paper describes the outcomes of the first shared task on multilingual named entity recognition (NER) that aims at recognizing mentions of named entities in web documents in Slavic languages, their normalization/lemmatization, and cross-lingual matching. The task initially covers seven languages and four types of NEs: person, location, organization, and miscellaneous, where the last category covers all other types of named entities, e.g., event or product. The input text collection consists of documents in seven Slavic languages collected from the web, each collection revolving around a certain "focus" entity. The main rationale of such a setup is to foster development of "all-rounder" NER and cross-lingual entity matching solutions that are not tailored to specific, narrow domains. The shared task was organized in the context of the 6th Balto-Slavic Natural Language Processing Workshop co-located with the EACL 2017 conference.

Similar shared tasks have been organized previously. The first *non-English* monolingual NER evaluations—covering Chinese, Japanese, Spanish, and Arabic—were carried out in the context of the Message Understanding Conferences (MUCs) (Chinchor, 1998) and the ACE Programme (Doddington et al., 2004). The first shared task focusing on *multilingual* named entity recognition, which covered some European languages, including Spanish, German, and Dutch, was organized in the context of CoNLL conferences (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003). The NE types covered in these campaigns were similar to the NE types covered in our Challenge. Also related to our task is the Entity Discovery and Linking (EDL) track (Ji et al., 2014; Ji et al., 2015) of the NIST Text Analysis Conferences (TAC). EDL aimed to extract entity mentions from a collection of textual documents in multiple languages (English, Chinese, and Spanish), and to partition the entities into cross-document equivalence classes, by either linking mentions to a knowledge base or directly

clustering them. An important difference between EDL and our task is that we do not link entities to a knowledge base.

Related to cross-lingual NE recognition is NE transliteration, i.e., linking NEs across languages that use different scripts. A series of NE Transliteration Shared Tasks were organized as a part of NEWS—Named Entity Workshops—(Duan et al., 2016), focusing mostly on Indian and Asian languages. In 2010, the NEWS Workshop included a shared task on Transliteration Mining (Kumaran et al., 2010), i.e., mining of names from parallel corpora. This task included corpora in English, Chinese, Tamil, Russian, and Arabic.

Prior work targeting NEs specifically for Slavic languages includes tools for NE recognition for Croatian (Karan et al., 2013; Ljubešić et al., 2013), a tool tailored for NE recognition in Croatian tweets (Baksa et al., 2017), a manually annotated NE corpus for Croatian (Agić and Ljubešić, 2014), tools for NE recognition in Slovene (Štajner et al., 2013; Ljubešić et al., 2013), a Czech corpus of 11,000 manually annotated NEs (Ševčíková et al., 2007), NER tools for Czech (Konkol and Konopík, 2013), tools and resources for fine-grained annotation of NEs in the National Corpus of Polish (Waszczuk et al., 2010; Savary and Piskorski, 2011) and a recent shared task on NE Recognition in Russian (Alexeeva et al., 2016).

To the best of our knowledge, the shared task described in this paper is the first attempt at multilingual name recognition, normalization, and cross-lingual entity matching that covers a large number of Slavic languages.

This paper is organized as follows. Section 2 describes the task; Section 3 describes the annotation of the dataset. The evaluation methodology is introduced in Section 4. Participant systems are described in Section 5 and the results obtained by these systems are presented in Section 6. Finally, lessons learnt and conclusions are discussed in Section 7.

2 Task Description

The data for the shared task consists of text documents in seven Slavic languages: Croatian, Czech, Polish, Russian, Slovak, Slovene, and Ukrainian. The documents focus around a certain entity—e.g., a person or an organization. The documents were obtained from the web, by posing a query to a search engine and parsing the HTML of the retrieved documents.

The task is to recognize, classify, and "normalize" all named-entity mentions in each of the documents, and to link across languages all named mentions referring to the same real-world entity.

Formally, the Multilingual Named Entity Recognition task includes three sub-tasks:

- **Named Entity Mention Detection and Classification.** Recognizing all unique named mentions of entities of four types: persons (PER), organizations (ORG), locations (LOC), miscellaneous (MISC), the last covering mentions of all other types of named entities, e.g., products, events, etc.
- **Name Normalization.** Mapping each named mention of an entity to its corresponding *base form*. By "base form" we generally mean the lemma ("dictionary form") of the inflected word-form. In some cases normalization should go beyond inflection and transform a derived word into a base word's lemma, e.g., in case of personal possessives (see below). Multi-word names should be normalized to the *canonical* multi-word expression, rather than a sequence of lemmas of the words making up the multi-word expression.
- **Entity Matching.** Assigning an identifier (ID) to each detected named mention of an entity, in such a way that mentions of entities referring to the same real-world entity should be assigned the same ID (referred to as the cross-lingual ID).

The task does not require positional information of the name entity mentions. Consequently, for all occurrences of the same form of a NE mention (e.g., inflected variant, acronym, or abbreviation) within the same document no more than one annotation should be returned.[1] Furthermore, distinguishing case information is not necessary since the evaluation is case-insensitive. In particular, if the text includes lowercase, uppercase or mixed-case variants of the same entity, the system should produce only one annotation for all of these mentions. For instance, for "*ISIS*", "*isis*", and "*Isis*" (provided that they refer to the same NE type), only one annotation should be returned. Note that the recognition of nominal or pronominal mentions of entities is not part of the task.

[1] Unless the different occurrences have different entity types (different readings) assigned to them, which is rare.

2.1 Named Entity Classes

The task defines the following four NE classes.

Person names (PER). Names of real persons (and fictional characters). Person names should not include titles, honorifics, and functions/positions. For example, in the text fragment "*... CEO Dr. Jan Kowalski...*", only "*Jan Kowalski*'" is recognized as a person name. Initials and pseudonyms are considered named mentions of persons and should be recognized. Similarly, named references to groups of people (that do not have a formal organization unifying them) should also be recognized, e.g., "*Ukrainians*." In this context, mentions of a single member belonging to such groups, e.g., "*Ukrainian*," should be assigned the same cross-lingual ID as plural mentions, i.e., "*Ukrainians*" and "*Ukrainian*" when referring to the nation should be assigned the same cross-lingual ID.

Personal possessives derived from a person name should be classified as a person, and the base form of the corresponding person name should be extracted. For instance, for "*Trumpov tweet*" (Croatian) it is expected to recognize "*Trumpov*" and classify it as PER, with the base form "*Trump*."

Locations (LOC). All toponyms and geopolitical entities (cities, counties, provinces, countries, regions, bodies of water, land formations, etc.), including named mentions of *facilities* (e.g., stadiums, parks, museums, theaters, hotels, hospitals, transportation hubs, churches, railroads, bridges, and similar facilities).

In case named mentions of facilities may also refer to an organization, the LOC tag should be used. For example, from the text phrase "*The Schipol Airport has acquired new electronic gates*" the mention "*The Schipol Airport*" should be extracted and classified as LOC.

Organizations (ORG). All kinds of organizations: political parties, public institutions, international organizations, companies, religious organizations, sport organizations, educational and research institutions, etc.

Organization designators and potential mentions of the seat of the organization are considered to be part of the organization name. For instance, from the text fragment "*Citi Handlowy w Poznaniu*" (a bank in Poznań), the full phrase "*Citi Handlowy w Poznaniu*" should be extracted.

When a company name is used to refer to a service (e.g., "*na Twiterze*" (Polish for "on Twitter"), the mention of "*Twitter*" is considered to refer to a service/product and should be tagged as MISC. However, when a company name is referring to a service which expresses the opinion of the company, e.g., "*Fox News*", it should be tagged as ORG.

Miscellaneous (MISC). All other named mentions of entities, e.g., product names—e.g., "*Motorola Moto X*"), events (conferences, concerts, natural disasters, holidays, e.g., "*Święta Bożego Narodzenia*" (Polish for "Christmas"), etc.

This category does not include temporal and numerical expressions, as well as identifiers such as email addresses, URLs, postal addresses, etc.

2.2 Complex and Ambiguous Entities

In case of complex named entities, consisting of nested named entities, only the *top-most* entity should be recognized. For example, from the text string "*George Washington University*" one should not extract "*George Washington*", but the entire string.

In case one word-form (e.g., "*Washington*") is used to refer to two different real-world entities in different contexts in the same document (e.g., a person and a location), the system should return two annotations, associated with different cross-lingual IDs.

2.3 System Input and Response

Input Document Format. Documents in the collection are represented in the following format. The first five lines contain meta-data; the core text to be processed begins from the 6th line and runs till the end of file.

```
<DOCUMENT-ID>
<LANGUAGE>
<CREATION-DATE>
<URL>
<TITLE>
<TEXT>
```

The <URL> field stores the origin from which the text document was retrieved. The values of

the meta-data fields were computed automatically (see Section 3 for details). In particular, the values of <CREATION-DATE> and <TITLE> were not provided for all documents, either due to unavailability of such data or due to errors in web page parsing during the creation process.

System Response. For each input document, the systems should return one file as follows. The first line should contain only the <DOCUMENT-ID> field that corresponds to the input file. Each subsequent line should contain the following, tab-separated fields:

<MENTION> TAB <BASE> TAB <CAT> TAB <ID>

The value of the <MENTION> field should be the NE mention as it appears in text. The value of the <BASE> field should be the base form of the entity. The <CAT> and <ID> fields store information on the category of the entity (ORG, PER, LOC, or MISC) and cross-lingual identifier, respectively. The cross-lingual identifiers may consist of an arbitrary sequence of alphanumeric characters. An example of a system response (for a document in Polish) is given below.

```
16
Podlascy Czeczeni    Podlascy Czeczeni    PER    1
ISIS    ISIS    ORG    2
Rosji    Rosja    LOC    3
Rosja    Rosja    LOC    3
Polsce    Polska    LOC    4
Warszawie    Warszawa    LOC    5
Magazynu Kuriera Porannego    Magazyn Kuriera\
        Porannego    ORG    6
```

3 Data

3.1 Trial Datasets

The registered participants were provided two trial datasets: (1) a dataset related to *Beata Szydło*, the current prime minister of Poland, and (2) a dataset related to *ISIS*, the so-called "Islamic State of Iraq and Syria" terrorist group. These datasets consisted of 187 and 186 documents, respectively, with equal distribution of documents across the seven languages of interest.

3.2 Test Datasets

Two datasets were prepared for evaluation, each consisting of documents extracted from the web and related to a given entity. One dataset contains documents related to *Donald Trump*, the recently elected President of United States (henceforth referred to as TRUMP), and the second dataset contains documents related to the *European Commission* (henceforth referred to as ECOMMISSION).

The test datasets were created as follows. For each "focus" entity, we posed a separate search query to Google, in each of the seven target languages. The query returned links to documents only in the language of interest. We extracted the first 100 links[2] returned by the search engine, removed duplicate links, downloaded the corresponding HTML pages—mainly news articles or fragments thereof—and converted them into plain text, using a hybrid HTML parser. This process was done semi-automatically using the tool described in (Crawley and Wagner, 2010). In particular, some of the meta-data fields—i.e., creation date, title, URL—were automatically computed using this tool.

HTML parsing resulted in texts that included not only the core text of a web page, but also some additional pieces of text, e.g., a list of labels from a menu, user comments, etc., which may not constitute well-formed utterances in the target language. This occurred in a small fraction of texts processed. Some of these texts were included in the test dataset in order to maintain the flavour of "real-data." However, obvious HTML parser failure (e.g., extraction of JavaScript code, extraction of empty texts, etc.) were removed from the data sets. Some of the downloaded documents were additionally polished by removing erroneously extracted boilerplate content. The resulting set of partially "cleaned" documents were used to select *circa* 20–25 documents for each language and topic, for the preparation of the final test datasets. Annotations for Croatian, Czech, Polish, Russian, and Slovene were made by native speakers; annotations for Slovak were made by native speakers of Czech, capable of understanding Slovak. Annotations for Ukrainian were made partly by native speakers and partly by near-native speakers of Ukrainian. Cross-lingual alignment of the entity identifiers was performed by two annotators.

Table 1 provides more quantitative details about the annotated datasets. Table 2 gives the breakdown of entity classes. It is noteworthy that a high proportion of the annotated mentions have a base form that differs from the form appearing in text. For instance, for the TRUMP dataset this figure is between 37.5% (Slovak) and 57.5% (Croatian).

[2]Or fewer, in case the search engine did not return 100 links.

	TRUMP		ECOMMISSION	
Language	# docs	# ment	# docs	# ment
Croatian	25	525	25	436
Czech	25	479	25	417
Polish	25	692	24	466
Russian	26	331	24	385
Slovak	24	453	25	374
Slovene	24	474	26	434
Ukrainian	28	337	54	1078
Total	177	3291	203	3588

Table 1: Quantitative data about the test datasets. *#docs* and *#ment* refer to the number of documents and NE mention annotations, respectively.

Entity type	TRUMP	ECOMMISSION
PER	48.4%	11.9%
LOC	26.9%	29.1%
ORG	18.3%	48.4%
MISC	6.4%	9.6%

Table 2: Breakdown of the annotations according to the entity type.

	Genitive	Nominative ("base")
hr	Europske komisije	Europska komisija
cz	Komisji Europejskiej	Komisja Europejska
pl	Evropskou komisí	Evropská komise
ru	Европейской комиссией	Европейская комиссия
sl	Evropske komisije	Evropska komisija
sk	Európskej komisie	Európska komisia
ua	Європейської Комісії	Європейська Комісія

Table 3: Inflected (genitive) forms of the name "*European Commission*" found in test data.

Table 3 provides examples of genitive forms of the name "*European Commission*" that occurred in the ECOMMISSION corpus frequently.

While normalization of the inflected forms in Table 3 could be achieved by lemmatization of each of the constituents of the noun phrase separately and then concatenating the corresponding base forms together, many entity mentions in the test dataset are complex noun phrases, whose lemmatization requires detection of inner syntactic structure. For instance, the inflected form of the Polish proper name *Europejskiego Funduszu Rozwoju Regionalnego* (*European*$_{GEN}$ *Fund*$_{GEN}$ *Development*$_{GEN}$ *Regional*$_{GEN}$) consists of two basic genitive noun phrases, of which only the first one ("European Fund") needs to be normalized, whereas the second ("Regional Development") should remain unchanged. The corresponding base form is "*Europejski Fundusz Rozwoju Regionalnego*". Since in some Slavic languages adjectives may precede or follow a noun in a noun phrase (like in the example above), detection of inner syntactic structure of complex proper names is not trivial (Radziszewski, 2013), and thus complicates the process of automated lemmatization. Complex person name declension paradigms (Piskorski et al., 2009) add another level of complexity.

It is worth mentioning that, for the sake of compliance with the NER guidelines in Section 2, documents that included hard-to-decide entity mention annotations were excluded from the test datasets for the present. A case in point is a document in Croatian that contained the phrase "*Zagrebačka, Sisačko-Moslavačka i Karlovačka županija*"—a contracted version of three named entities ("*Zagrebačka županija*", "*Sisačko-Moslavačka županija*", and "*Karlovačka županija*") expressed using a head noun with three coordinate modifiers.

4 Evaluation Methodology

The NER task (exact case-insensitive matching) and Name Normalization task (also called "lemmatization") were evaluated in terms of precision, recall, and F1-scores. In particular, for NER, two types of evaluations were carried out:

- **Relaxed evaluation**: An entity mentioned in a given document is considered to be extracted correctly if the system response includes at least one annotation of a named mention of this entity (regardless whether the extracted mention is in base form);
- **Strict evaluation**: The system response should include exactly one annotation for *each* unique form of a named mention of an entity in a given document, i.e., capturing and listing all variants of an entity is required.

In relaxed evaluation mode we additionally distinguish between *exact* and *partial matching*, i.e., in the case of the latter an entity mentioned in a given document is considered to be extracted correctly if the system response includes at least one partial match of a named mention of this entity.

In the evaluation we consider various levels of granularity, i.e., the performance for: (a) all NE types and all languages, (b) each particular NE

TRUMP		*Language*													
Phase	*Metric*	cz		hr		pl		ru		sk		sl		ua	
Recognition	Relaxed Partial	jhu	46.2	jhu	52.4	pw jhu	66.6 44.8	jhu	46.3	jhu	46.8	jhu	47.3	jhu	38.8
	Relaxed Exact	jhu	46.1	jhu	50.8	pw jhu	66.1 43.4	jhu	43.1	jhu	46.2	jhu	46.0	jhu	37.3
	Strict	jhu	46.1	jhu	50.4	pw jhu	66.6 41.0	jhu	41.8	jhu	47.0	jhu	46.2	jhu	33.2
Normalization						pw	60.5								
Entity matching	Document-level	jhu	5.4	jhu	7.3	jhu pw	6.3 10.8	jhu	11.2	jhu	10.1	jhu	9.5	jhu	0.0
	Single-language	jhu	19.3	jhu	17.6	jhu pw	18.2 4.9	jhu	18.9	jhu	22.6	jhu	28.7	jhu	10.7
	Cross-lingual	jhu	9.0												
ECOMMISSION		*Language*													
Phase	*Metric*	cz		hr		pl		ru		sk		sl		ua	
Recognition	Relaxed Partial	jhu	47.6	jhu	45.9	pw jhu	61.8 47.3	jhu	46.0	jhu	49.1	jhu	47.9	jhu	18.4
	Relaxed Exact	jhu	44.4	jhu	43.1	pw jhu	60.9 42.4	jhu	44.1	jhu	46.4	jhu	43.9	jhu	14.7
	Strict	jhu	47.2	jhu	46.2	pw jhu	61.1 44.8	jhu	46.5	jhu	46.1	jhu	47.8	jhu	10.8
Normalization						pw	48.3								
Entity Matching	Document-level	jhu	25.0	jhu	16.0	jhu pw	13.7 13.4	jhu	13.7	jhu	13.1	jhu	36.8	jhu	0.6
	Single-language	jhu	27.3	jhu	22.1	jhu pw	17.5 4.6	jhu	24.9	jhu	30.6	jhu	32.2	jhu	4.8
	Cross-lingual	jhu	2.6												

Table 4: Evaluation results across all scenarios and languages.

type and all languages, (c) all NE types for each language, and (d) each particular NE type per language.

In the name normalization sub-task, only correctly recognized entity mentions in the system response and only those that were normalized (on both the annotation and system's sides) are taken into account. Formally, let $correct_N$ denote the number of all correctly recognized entity mentions for which the system returned a correct base form. Let key_N denote the number of all normalized entity mentions in the gold-standard answer key and $response_N$ denote the number of all normalized entity mentions in the system's response. We define precision and recall for the name normalization task as:

$$Recall_N = \frac{corrrect_N}{key_N}$$

$$Precision_N = \frac{corrrect_N}{response_N}$$

In evaluating the document-level, single-language and cross-lingual entity matching task we have adapted the Link-Based Entity-Aware metric (LEA) (Moosavi and Strube, 2016) which considers how important the entity is and how well it is resolved. LEA is defined as follows. Let $K = \{k_1, k_2, \ldots, k_{|K|}\}$ and $R = \{r_1, r_2, \ldots, r_{|R|}\}$ denote the key entity set and the response entity set, respectively, i.e., $k_i \in K$ ($r_i \in R$) stand for set of mentions of the same entity in the key entity set (response entity set). LEA recall and precision are then defined as follows:

$$Recall_{LEA} = \frac{\sum_{k_i \in K}(imp(k_i) \times res(k_i))}{\sum_{k_z \in K} imp(k_z)}$$

$$Precision_{LEA} = \frac{\sum_{r_i \in R}(imp(r_i) \times res(r_i))}{\sum_{r_z \in R} imp(r_z)}$$

where imp and res denote the measure of importance and the resolution score for an entity, respectively. In our setting, we define $imp(e) = \log_2 |e|$ for an entity e (in K or R), $|e|$ is the number of mentions of e—i.e., the more mentions an entity has the more important it is. To avoid biasing the importance of the more frequent entities $\log$ is used. The resolution score of key entity k_i is computed as the fraction of correctly resolved co-reference links of k_i:

$$res(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)}$$

where $link(e) = (|e| \times (|e| - 1))/2$ is the number of unique co-reference links in e. For each k_i, LEA checks all response entities to check whether they are partial matches for k_i. Analogously, the resolution score of response entity r_i is computed as the fraction of co-reference links in r_i that are extracted correctly:

$$res(r_i) = \sum_{k_j \in K} \frac{link(r_i \cap k_j)}{link(r_i)}$$

Using LEA brings several benefits. For example, LEA considers resolved co-reference relations instead of resolved mentions and has more discriminative power than other metrics used for evaluation of co-reference resolution (Moosavi and Strube, 2016).

It is important to note at this stage that the evaluation was carried out in "case-insensitive" mode: all named mentions in system response and test corpora were lowercased.

5 Participant Systems

Eleven teams from seven countries—Czech Republic, Germany, India, Poland, Russia, Slovenia, and USA—registered for the evaluation task and received the trial datasets. Due to the complexity of the task and relatively short time available to create a working solution, only two teams submitted results within the deadline. A total of two unique runs were submitted.

JHU/APL team attempted the NER and Entity Matching sub-tasks. They employed a statistical tagger called SVMLattice (Mayfield et al., 2003), with NER labels inferred by projecting English tags across bitext. The Illinois tagger (Ratinov and Roth, 2009) was used for English. A rule-based entity clusterer called "kripke" was used for Entity Matching (McNamee et al., 2013). The team (code "*jhu*") attempted all languages available in the Challenge. More details can be found in (Mayfield et al., 2017).

The G4.19 Research Group adapted Liner2 (Marcińczuk et al., 2013)—a generic framework which can be used to solve various tasks based on sequence labeling, which is equipped with a set of modules (based on statistical models, dictionaries, rules and heuristics) which recognize and annotate certain types of phrases. The details of tuning Liner2 to tackle the shared task are described in (Marcińczuk et al., 2017). The team (code "*pw*") attempted only the Polish-language Challenge.

The above systems met the deadline to participate in the first run of the Challenge—Phase I. Since the Challenge aroused significant interest in the research community, it was extended into Phase II, with a new deadline for submitting system responses, beyond the time of publication of this paper. Please refer to the Challenge web site[3] for information on the current status, systems tested, and their performance.

6 Evaluation Results

The results of the runs submitted for Phase I are presented in Table 4. The figures provided for the recognition are micro-averaged F1-scores.

For normalization, we report F1-scores, using the $Recall_N$ and $Precision_N$ definitions from Section 4, computed for entity mentions for which the annotation or system response contains a different base form compared to the surface form. This evaluation includes only correctly recognized entity mentions to suppress the influence of entity recognition performance.

Lastly, for entity matching, the micro-averaged F1-scores are provided, computed using LEA precision and recall values (see Section 4).

System *pw* performed substantially better on Polish than system *jhu*.

Considering the entity types, performance was overall better for LOC and PER, and substantially lower for ORG and MISC, which is not unexpected. Table 5 and 6 provide the overall aver-

[3] http://bsnlp.cs.helsinki.fi/shared_task.html

Metric	Precision	Recall	F1
PER	74.8	65.9	69.8
LOC	73.0	75.4	74.2
ORG	47.1	22.1	30.0
MISC	7.9	14.4	10.2

Table 5: Breakdown of the recognition performance according to the entity type for TRUMP dataset.

Metric	Precision	Recall	F1
PER	68.2	59.4	62.9
LOC	73.1	57.8	64.5
ORG	45.0	49.0	46.6
MISC	18.7	12.0	14.2

Table 6: Breakdown of the recognition performance according to the entity type for ECOMMISSION dataset.

age precision, recall, and F1 figures for the relaxed evaluation with partial matching for TRUMP and ECOMMISSION scenario respectively.

Considering the tested languages and scenarios, system *jhu* achieved best performance on TRUMP in Croatian, its poorest performance was on ECOMMISSION in Ukrainian. System *pw* performed better on the TRUMP scenario than on ECOMMISSION. Overall, the TRUMP scenario appears to be easier, due to the mix of named entities that predominate in the texts. The ECOMMISSION documents discuss organizations with complex geo-political inter-relationships and affiliations.

Furthermore, cross-lingual co-reference seems to be a difficult task.

7 Conclusions

This paper reports on the First multilingual named entity Challenge that aims at recognizing mentions of named entities in web documents in Slavic languages, their normalization/lemmatization, and cross-lingual matching. Although the Challenge aroused substantial interest in the field, only two teams submitted results on time, most likely due to the complexity of the tasks and the short time available to finalize a solution. While drawing substantial conclusions from the evaluation of two systems is not yet possible, we can observe though that the overall performance of the two systems on hidden test sets revolving around a specific entity is significantly lower than in the case of processing less-morphologically complex languages.

To support research on NER-related tasks for Slavic languages, including cross-lingual entity matching, the Challenge was extended into Phase II, going beyond the date of the publication of this paper. For the current list of systems that has been evaluated on the different tasks and their performance figures please refer to the shared task web page.

The test datasets, the corresponding annotations and various scripts used for the evaluation purposes are made available on the shared task web page as well.

We plan to extend the Challenge through provision of additional test datasets in the future, involving new entities, in order to further boost research on developing "all-rounder" NER solutions for processing real-world texts in Slavic languages and carrying out cross-lingual entity matching. Furthermore, we plan to extend the set of the languages covered, depending on the availability of annotators. Finally, some work will focus on the refining the NE annotation guidelines in order to properly deal with particular phenomena, e.g., coordinated NEs and contracted versions of multiple NEs, which were excluded from the first test datasets.

Acknowledgments

We thank Katja Zupan (Department of Knowledge Technologies, Jožef Stefan Institute, Slovenia), Anastasia Stepanova (State University of New York, Buffalo), Domagoj Alagić (TakeLab, University of Zagreb), and Olga Kanishcheva, Kateryna Klymenkova, Ekaterina Yurieva (the National Technical University, Kharkiv Polytechnic Institute), who contributed to the preparation of the Slovenian, Russian, Croatian, and Ukrainian test data. We are also grateful to Tomaž Erjavec from the Department of Knowledge Technologies, Jožef Stefan Institute in Slovenia, who contributed various ideas. Work on Czech and Slovak was supported by Project MediaGist, EU's FP7 People Programme (Marie Curie Action), no. 630786.

The effort of organizing the shared task was supported by the Europe Media Monitoring (EMM) Project carried out by the Text and Data Mining Unit of the Joint Research Centre of the European Commission.

References

Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1724–1727, Reykjavík, Iceland.

S. Alexeeva, S.Y. Toldova, A.S. Starostin, V.V. Bocharov, A.A. Bodrova, A.S. Chuchunkov, S.S. Dzhumaev, I.V. Efimenko, D.V. Granovsky, V.F. Khoroshevsky, et al. 2016. FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. In *Computational Linguistics and Intellectual Technologies. Proceedings of the Annual International Conference "Dialogue"*, pages 688–705.

Krešimir Baksa, Dino Golović, Goran Glavaš, and Jan Šnajder. 2017. Tagging named entities in Croatian tweets. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 4(1):20–41.

Nancy Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*.

Jonathan B. Crawley and Gerhard Wagner. 2010. Desktop Text Mining for Law Enforcement. In *Proceedings of IEEE International Conference on Intelligence and Security Informatics (ISI 2010)*, pages 23–26, Vancouver, BC, Canada.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) program—tasks, data, and evaluation. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 837–840, Lisbon, Portugal.

Xiangyu Duan, Rafael E. Banchs, Min Zhang, Haizhou Li, and A. Kumaran. 2016. Report of NEWS 2016 Machine Transliteration Shared Task. In *Proceedings of The Sixth Named Entities Workshop*, pages 58–72, Berlin, Germany.

Heng Ji, Joel Nothman, and Ben Hachey. 2014. Overview of TAC-KBP2014 entity discovery and linking tasks. In *Proceedings of Text Analysis Conference (TAC2014)*, pages 1333–1339.

Heng Ji, Joel Nothman, and Ben Hachey. 2015. Overview of TAC-KBP2015 tri-lingual entity discovery and linking. In *Proceedings of Text Analysis Conference (TAC2015)*.

Mladen Karan, Goran Glavaš, Frane Šarić, Jan Šnajder, Jure Mijić, Artur Šilić, and Bojana Dalbelo Bašić. 2013. CroNER: Recognizing named entities in Croatian using conditional random fields. *Informatica*, 37(2):165.

Michal Konkol and Miloslav Konopík. 2013. CRF-based Czech named entity recognizer and consolidation of Czech NER research. In *Text, Speech and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 153–160. Springer Berlin Heidelberg.

A Kumaran, Mitesh M. Khapra, and Haizhou Li. 2010. Report of NEWS 2010 transliteration mining shared task. In *Proceedings of the 2010 Named Entities Workshop*, pages 21–28, Uppsala, Sweden.

Nikola Ljubešić, Marija Stupar, Tereza Jurić, and Željko Agić. 2013. Combining available datasets for building named entity recognition models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):35–57.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki. 2013. Liner2—a customizable framework for proper names recognition for Polish. In Robert Bembenik, Lukasz Skonieczny, Henryk Rybinski, Marzena Kryszkiewicz, and Marek Niezgodka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 231–253. Springer.

Michał Marcińczuk, Jan Kocoń, and Marcin Oleksy. 2017. Liner2—a generic framework for named entity recognition. In *Proceedings of the Sixth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Valencia, Spain.

James Mayfield, Paul McNamee, Christine Piatko, and Claudia Pearce. 2003. Lattice-based tagging using support vector machines. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 303–308, New York, NY, USA. ACM.

James Mayfield, Paul McNamee, and Cash Costello. 2017. Language-independent named entity analysis using parallel projection and rule-based disambiguation. In *Proceedings of the Sixth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, Valencia, Spain.

Paul McNamee, James Mayfield, Tim Finin, and Dawn Lawrie. 2013. HLTCOE participation at TAC 2013. In *Proceedings of the Sixth Text Analysis Conference, (TAC 2013)*, Gaithersburg, Maryland, USA.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 632–642, Berlin, Germany.

Jakub Piskorski, Karol Wieloch, and Marcin Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information retrieval*, 12(3):275–299.

Adam Przepiórkowski. 2007. Slavonic information extraction and partial parsing. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Radziszewski. 2013. Learning to lemmatise polish noun phrases. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 701–709.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Agata Savary and Jakub Piskorski. 2011. Language Resources for Named Entity Annotation in the National Corpus of Polish. *Control and Cybernetics*, 40(2):361–391.

Magda Ševčíková, Zdeněk Žabokrtskỳ, and Oldřich Kruza. 2007. Named entities in Czech: annotating data and developing NE tagger. In *International Conference on Text, Speech and Dialogue*, pages 188–195. Springer.

Tadej Štajner, Tomaž Erjavec, and Simon Krek. 2013. Razpoznavanje imenskih entitet v slovenskem besedilu. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, 1(2):58–81.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jakub Waszczuk, Katarzyna Głowińska, Agata Savary, and Adam Przepiórkowski. 2010. Tools and methodologies for annotating syntax and named entities in the National Corpus of Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA'10)*, pages 531–539, Wisła, Poland. PTI.