The Second Cross-Lingual Challenge on Recognition, Normalization, Classification, and Linking of Named Entities across Slavic Languages

Jakub Piskorski,[1] Laska Laskova,[5] Michał Marcińczuk,[4]
Lidia Pivovarova,[2] Pavel Přibáň,[3] Josef Steinberger,[3] Roman Yangarber[2]
[1]Joint Research Centre, Ispra, Italy first.last@ec.europa.eu
[2]University of Helsinki, Finland first.last@cs.helsinki.fi
[3]University of West Bohemia, Czech Republic {pribanp, jstein}@kiv.zcu.cz
[4]Wrocław University of Science and Technology, Poland michal.marcinczuk@pwr.edu.pl
[5]Bulgarian Academy of Sciences, Bulgaria laska@bultreebank.org

Abstract

We describe the Second Multilingual Named Entity Challenge in Slavic languages. The task is recognizing mentions of named entities in Web documents, their normalization, and cross-lingual linking. The Challenge was organized as part of the 7th Balto-Slavic Natural Language Processing Workshop, co-located with the ACL-2019 conference. Eight teams participated in the competition, which covered four languages and five entity types. Performance for the named entity recognition task reached 90% F-measure, much higher than reported in the first edition of the Challenge. Seven teams covered all four languages, and five teams participated in the cross-lingual entity linking task. Detailed evaluation information is available on the shared task web page.

1 Introduction

Due to rich inflection and 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 problem (Przepiórkowski, 2007; Piskorski et al., 2009). Fostering research on detection and normalization of NEs—and on the closely related problem of cross-lingual, cross-document *entity linking*—is of paramount importance for improving multilingual and cross-lingual information access in these languages.

This paper describes the Second Shared Task on multilingual NE recognition (NER), which aims at addressing these problems in a systematic way. The shared task was organized in the context of the 7th Balto-Slavic Natural Language Processing Workshop co-located with the ACL 2019 conference. The task covers four languages—Bulgarian, Czech, Polish and Russian—and five types of NE: person, location, organization, product, and event. The input text collection consists of documents collected from the Web, each collection centered on a certain "focal" entity. The rationale of such a setup is to foster the development of "all-round" NER and cross-lingual entity linking solutions, which are not tailored to specific, narrow domains. This paper also serves as an introduction and a guide for researchers wishing to explore these problems using the training and test data.[1]

This paper is organized as follows. Section 3 describes the task; Section 4 describes the annotation of the dataset. The evaluation methodology is introduced in Section 5. Participant systems are described in Section 6 and the results obtained by these systems are presented in Section 7. Conclusions and lessons learned are discussed in Section 8.

2 Prior Work

The work we describe here builds on the First Shared Task on Multilingual Named Entity Recognition, Normalization and cross-lingual Matching for Slavic Languages, (Piskorski et al., 2017), which, to the best of our knowledge, was the first attempt at such a shared task covering several Slavic languages.

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 several European languages, including Spanish, German, and Dutch, was organized in the context of CoNLL conferences (Tjong Kim Sang, 2002; Tjong Kim Sang

[1] bsnlp.cs.helsinki.fi/shared_task.html

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 63–74,
Florence, Italy, 2 August 2019.

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 Entity Discovery and Linking (EDL), (Ji et al., 2014, 2015), a track of the NIST Text Analysis Conferences (TAC). EDL aimed to extract entity mentions from a collection of 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).

3 Task Description

The data for the shared task consists of sets of documents in four Slavic languages: Czech, Polish, Russian, and Bulgarian. To accommodate entity linking, each set of documents is chosen to focus around one certain entity—e.g., a person, an organization or an event. The documents were obtained from the Web, by posing a keyword query to a search engine and extracting the textual content from the Web pages.

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 named mentions of entities of five types: persons (PER), organizations (ORG), locations (LOC), products (PRO), and events (EVT).

- **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 Linking.** Assigning a unique identifier (ID) to each detected named mention of an entity, in such a way that mentions 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. Thus, for all occurrences of the same form of a NE mention (e.g., an inflected variant, an acronym or abbreviation) within a given document, no more than one annotation should be produced.[2] Furthermore, distinguishing typographical case is not necessary since the evaluation is case-insensitive. 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 "*BREXIT*" and "*Brexit*" (provided that they refer to the same NE type), only one annotation should be produced. Note that recognition of common-noun or pronominal references to named entities is not part of the task.

3.1 Named Entity Classes

The task defines the following five NE classes.

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

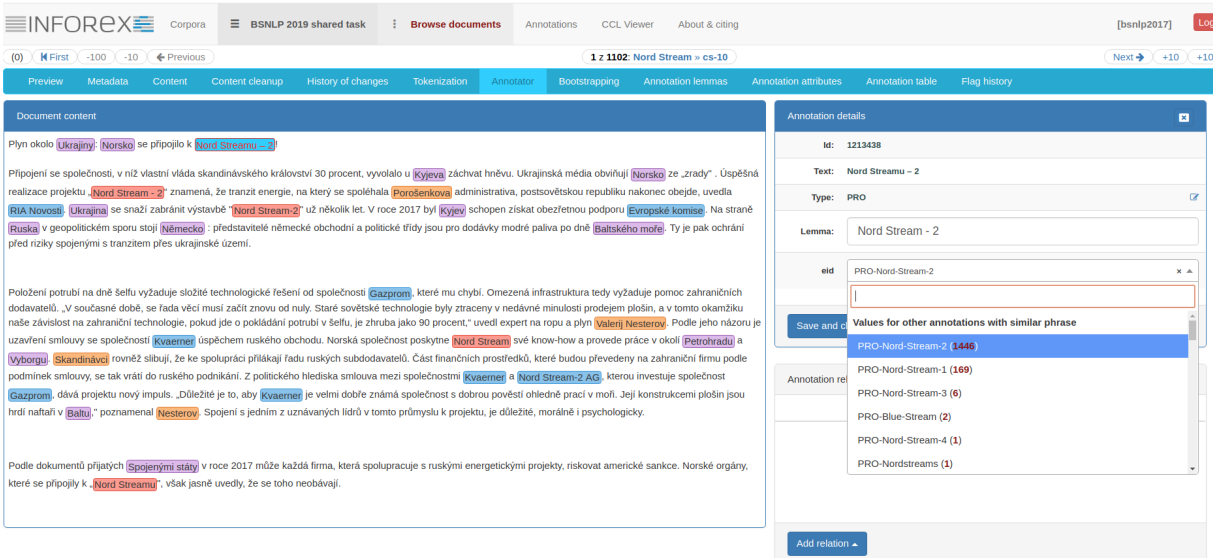

Figure 1: Screenshot of the Inforex Web interface, the tool used for data annotation

Person names (PER): Names of real (or fictional) persons). 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 receive the same cross-lingual ID.

Personal possessives derived from a person's name should be classified as a Person, and the base form of the corresponding name should be extracted. For instance, in "*Trumpov tweet*" (Croatian) one is expected to classify "*Trumpov*" 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 *also* refer to an organization, the LOC tag should be used. For example, from the text "*The Schipol Airport has acquired new electronic gates*" the mention "*The Schipol Airport*" should be classified as LOC.

Organizations (ORG): All organizations, including companies, public institutions, political parties, international organizations, 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 "*...Citi Handlowy w Poznaniu...*" (a bank in Poznań), the full phrase "*Citi Handlowy w Poznaniu*" should be extracted.

Products (PRO): All names of *products and services*, such as electronics ("*Motorola Moto Z Play*"), cars ("*Subaru Forester XT*"), newspapers ("*The New York Times*"), web-services ("*Twitter*").

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 PRO. However, when a company name refers to a service, expressing an opinion of the company, e.g., "*Fox News*", it should be tagged as ORG.

This category also includes legal documents and treaties, e.g., "*Traktat Lizboński*" (Polish: "Treaty of Lisbon").

Events (EVT): This category covers named mentions of events, including conferences, e.g. "24.

Japonci se ptají na czexit, říká Špicar ze Svazu průmyslu. "Odešli bychom z Česka," varovali ho Případné vystoupení České republiky z Evropské unie by bylo podle ekonomů, Hospodářské komory i Svazu průmyslu a dopravy ekonomickou sebevraždou. Odchod z EU by znamenal ztrátu stovek tisíc pracovních míst a česká ekonomika by se podle některých dostala na úroveň Běloruska. Praha 21:18 7. února 2018

cs-10

Japonci	Japonci	PER	GPE-Japan
czexit	czexit	EVT	EVT-Czexit
Špicar	Špicar	PER	PER-Radek-Spicar
Svazu průmyslu	Svaz průmyslu	ORG	ORG-Svaz-Prumyslu
Česka	Česko	LOC	GPE-Czech-Republic
České republiky	Česká republika	LOC	GPE-Czech-Republic
Evropské unie	Evropská unie	ORG	ORG-European-Union
Hospodářské komory	Hospodářská komora	ORG	ORG-Hospodarska-Komora
Svazu průmyslu a dopravy	Svaz průmyslu a dopravy	ORG	ORG-Svaz-Prumyslu
EU	EU	ORG	ORG-European-Union
Běloruska	Bělorusko	LOC	GPE-Belarus
Praha	Praha	LOC	GPE-Prague

Figure 2: Example input and output formats.

Konference Žárovného Zinkování" (Czech: "Hot Galvanizing Conference"), concerts, festivals, holidays, e.g., "*Vánoce*" (Polish: "Christmas"), wars, battles, disasters, e.g., "*Katastrofa Czernobylska*" (Polish: "the Chernobyl catastrophe"). Future, speculative, and fictive events—e.g., "Czexit" or "Polexit"—are considered as event mentions as well.

3.2 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 "*George Washington University*" one should not extract "*George Washington*", but only the top-level entity.

In case one word-form (e.g., "*Washington*") is used to refer to more than one 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.

In case of coordinated phrases, like "*European and British Parliament*," two names should be extracted (as ORG). The lemmas would be "*European*" and "*British Parliament*", and the IDs should refer to "*European Parliament*" and "*British Parliament*" respectively.

In rare cases, plural forms might have two annotations—e.g., in the phrase "*a border between Irelands*"—"*Irelands*" should be extracted twice with identical lemmas but different IDs.

3.3 System Input and Response

Input Document Format: Documents in the collection are represented in the following format. The first five lines contain meta-data:

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

The text to be processed begins from the sixth line and runs till the end of file. 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 4 for details). The values of <CREATION-DATE> and <TITLE> were not provided for all documents, due to unavailability of such data or due to errors in parsing during data collection.

System Response. For each input file, the system should return one output file as follows. The first line should contain only the <DOCUMENT-ID>, which corresponds to the input. Each subsequent line contains one annotation, as tab-separated fields:

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

The <MENTION> field should be the NE as it appears in text. The <BASE> field should be the base form of the entity. The <CAT> field stores the category of the entity (ORG, PER, LOC, PROD, or EVT) and <ID> is the cross-lingual identifier. The cross-lingual identifiers may consist of an arbitrary sequence of alphanumeric characters. An example document in Czech and the corresponding response is shown in Figure 2.

For detailed descriptions of the tasks and guidelines, please refer to the web page of the shared task.[3]

4 Data

The data consist of four sets of documents extracted from the Web, each related to a given *focus*

[3] bsnlp.cs.helsinki.fi/Guidelines_20190122.pdf

	BREXIT				ASIA BIBI				NORD STREAM				RYANAIR			
	PL	CS	RU	BG	PL	CS	RU	BG	PL	CS	RU	BG	PL	CS	RU	BG
Documents	500	284	153	600	88	89	118	99	151	153	137	130	146	149	149	87
PER	2 650	1 108	1 308	2 515	683	570	643	565	538	543	334	335	136	157	71	147
LOC	3 525	1 279	666	2 407	403	366	567	379	1 430	1 566	1 144	910	822	774	888	343
ORG	3 080	1 036	828	2 454	286	214	419	244	837	446	658	540	529	634	494	237
EVT	1 072	471	261	776	14	3	1	8	15	9	3	6	7	12	0	4
PRO	667	232	137	489	55	42	47	63	405	350	445	331	114	65	73	79
Total	10 994	4 126	3 200	8 641	1 441	1 195	1 677	1 259	3 225	2 914	2 584	2 122	1 608	1 642	1 526	810
Distinct																
Surface forms	2 813	1 110	771	1 200	507	303	406	403	843	769	850	500	514	475	394	322
Lemmas	2 133	839	568	1 092	412	248	317	359	634	549	568	448	420	400	327	314
Entity IDs	1 508	582	269	777	273	160	178	231	444	393	314	305	322	306	247	246

Table 1: Overview of the training and test datasets.

entity. We tried to choose entities related to current events covered in news in various languages. ASIA BIBI, which relates to a Pakistani woman involved in a blasphemy case, BREXIT, RYANAIR, which faced a massive strike, and NORD STREAM, a controversial Russian-European project.

Each dataset was created as follows. For the focus entity, we posed a search query to Google, in each of the target languages. The query returned documents in the target language. We removed duplicates, downloaded the HTML—mainly news articles—and converted them into plain text. This process was done semi-automatically using the tool described in (Crawley and Wagner, 2010). In particular, some of the meta-data—i.e., creation date, title, URL—were automatically extracted using this tool.

HTML parsing results may include not only the main text of a Web page, but also some additional text, e.g., labels from menus, user comments, etc., which may not constitute well-formed utterances in the target language.[4] The resulting set of partially "cleaned" documents were used to manually select documents for each language and topic, for the final datasets.

Documents were annotated using the Inforex[5] web-based system for annotation of text corpora (Marcinczuk et al., 2017). Inforex allows parallel access and resource sharing by multiple annotators. It let us share a common list of entities, and perform entity-linking semi-automatically: for a given entity, an annotator sees a list of entities of the same type inserted by all annotators and can select an entity ID from the list. A snapshot of the Inforex interface is in Figure 1.

In addition, Inforex keeps track of all lemmas and IDs inserted for each surface form, and inserts them automatically, so in many cases the annotator only confirms the proposed values, which speeds up the annotation process a great deal. All annotations were made by native speakers. After annotation, we performed automatic and manual consistency checks, to reduce annotation errors, especially in entity linking.

Using Inforex allowed us to annotate data much faster than in the first edition of the shared task. Thus we were able to annotated larger datasets and provide participants with training data. (In the first edition participants received only test data.) Data statistics are presented in Table 1.

Documents about ASIA BIBI and BREXIT were used for training and distributed to the participating teams with annotations. The testing datasets—RYANAIR and NORD STREAM—were released to the participants 2 days before the submission deadline. The participants did not know the topics in advance, and did not receive the annotations. Thus, we push participants to build a general solution for Slavic NER, rather than to optimize their models toward a particular set of names.

5 Evaluation Methodology

The NER task (exact case-insensitive matching) and Name Normalization (or "lemmatization") were evaluated in terms of precision, recall, and F1-measure. For NER, two types of evaluations were carried out:

- **Relaxed:** An entity mentioned in a given

[4]This occurred in a small fraction of texts processed. Some of these texts were included in the test dataset in order to maintain the flavor 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 documents were polished further by removing erroneously extracted boilerplate content.

[5]github.com/CLARIN-PL/Inforex

document is considered to be extracted correctly if the system response includes *at least one* annotation of a named mention of this entity (regardless of whether the extracted mention is in base form);

- **Strict:** 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., identifying all variants of an entity is required.

In relaxed evaluation we additionally distinguish between *exact* and *partial matching*: in the latter case, 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.

We evaluate systems at several levels of granularity: we measure performance for (a) all NE types and all languages, (b) each given NE type and all languages, (c) all NE types for each language, and (d) each given NE type per language.

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

$$Recall = \frac{N_{corrrect}}{N_{key}} \qquad Precision = \frac{N_{corrrect}}{N_{response}}$$

In evaluating document-level, single-language and cross-lingual entity linking we adopted 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 $K = \{k_1, k_2, \ldots, k_{|K|}\}$ denote the set of key entities and $R = \{r_1, r_2, \ldots, r_{|R|}\}$ the set of response entities, 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:

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

LEA brings several benefits. For example, LEA considers resolved co-reference relations instead of resolved mentions and has more discriminative power than other metrics for 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 lower-cased.

6 Participant Systems

Sixteen teams from eight countries registered for the shared task. Half of the registered teams submitted results by the deadline. Five teams submitted description of their systems in the form of a Workshop paper. The remaining teams submitted a short description of their systems.

We briefly review the systems; complete descriptions appear in the corresponding papers.

CogComp used multi-source BiLSTM-CRF models, using solely the BERT multilingual embeddings, (Devlin et al., 2019), which directly

allows the model to train on datasets in multiple languages. The team submitted several models trained on different combinations of input languages. They found that multi-source training with multilingual BERT outperforms single-source. Cross-lingual (even cross-script) training worked remarkably well. Multilingual BERT can handle train/test sets with mismatching tagsets in certain situations. The best performing models were trained on a combination of data in four languages, while adding English into training data worsen the overall performance, (Tsygankova et al., 2019).

CTC-NER is a baseline prototype of a NER component of an entity recognition system currently under development at the Cognitive Technologies Center, Russia. The system has a hybrid architecture, combining rule-based and ML techniques, where the ML-component is loosely related to (Antonova and Soloviev, 2013). As the system processes Russian, English and Ukrainian, the team submitted output only for Russian.

IIUWR.PL combines Flair[6], Polyglot[7] and BERT.[8] Additional training corpora were used: KPWr[9] for Polish, CNEC[10] for Czech, and data extracted using heuristics from Wikipedia. Lemmatization is partially trained on Wikipedia and PolEval corpora,[11] and partially rule-based. Entity linking is rule-based, and uses WikiData and FastText (Bojanowski et al., 2017).

JRC-TMA-CC is a hybrid system combining a rule-based approach and machine learning techniques. It is a corpus-driven system, lightweight and highly multilingual, exploiting both automatically created lexical resources, such as JRC-Names (Ehrmann et al., 2017), and external resources, such as BabelNet (Jacquet et al., 2019a). The main focus of the approach is on generating the possible inflected variants for known names (Jacquet et al., 2019b).

NLP Cube[12] is an open-source NLP framework that handles sentence segmentation, POS Tagging and lemmatization. The low-level features obtained from the framework, such as part of speech tags, were used as input for an LSTM model. Each

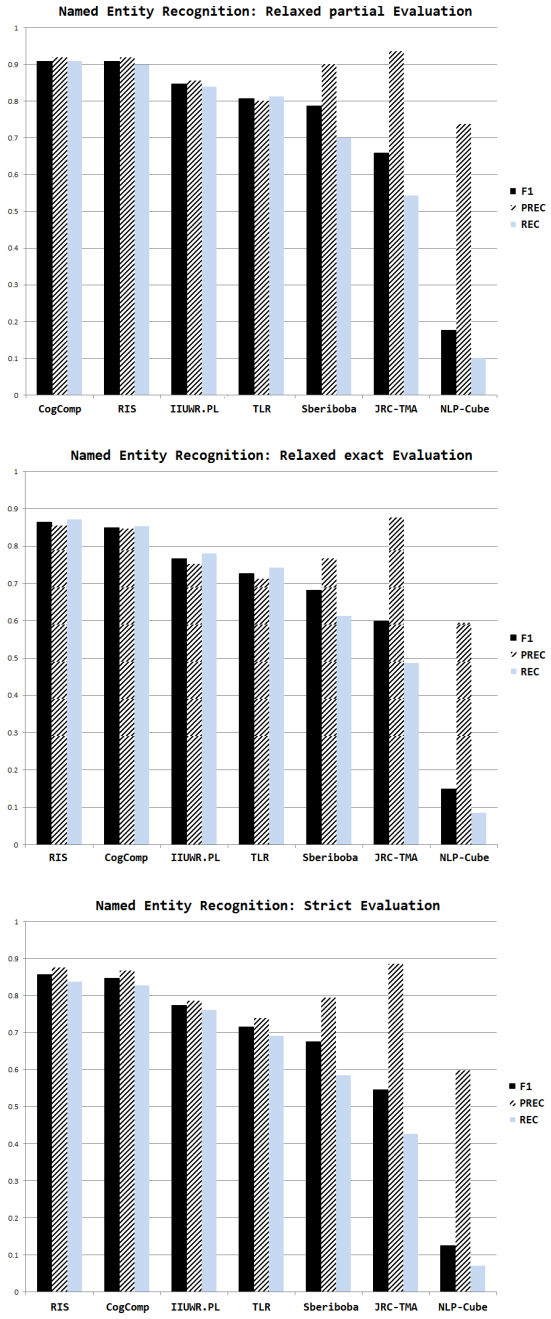

Figure 3: Average system performances on the test data

language was trained individually, producing four models. The models were trained using DyNet[13].

RIS is a modified BERT model, which uses CRF as the top-most layer (Arkhipov et al., 2019). The model was initialized with an existing BERT model trained on 100 languages.

Sberiboba uses multilingual BERT embeddings, summed with learned weights and followed by BiLSTM, attention layers and NCRF++ on the top (Emelianov and Artemova, 2019). Multilin-

[6] github.com/zalandoresearch/flair
[7] polyglot.readthedocs.io
[8] github.com/huggingface/pytorch-pretrained-BERT, github.com/sberbank-ai/ner-bert
[9] clarin-pl.eu/dspace/handle/11321/270
[10] ufal.mff.cuni.cz/cnec/cnec2.0
[11] poleval.pl/tasks/task2
[12] github.com/adobe/NLP-Cube
[13] dynet.io

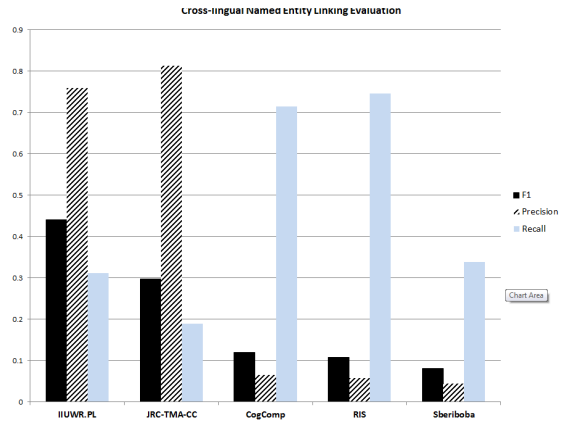

Figure 4: Evaluation results for closs-lingual entity linking. Averaged across two corpora.

gual BERT is used only for the embeddings, with no fine-tuning for the tasks.

TLR used a standard end-to-end architecture for sequence labeling, namely: LSTM-CNN-CRF, (Ma and Hovy, 2016). It was combined with contextual embeddings using a weighted average (Reimers and Gurevych, 2019) of a BERT model pre-trained for multiple languages (including all of the languages of the Task).

As seen from these descriptions, most of the teams use the BERT model, except NLP Cube, which uses another deep learning model (LSTM), and JRC, which uses rule-based processing of Slavic inflection.

7 Evaluation Results

Figure 3 shows system performance averaged across all languages and two test corpora. We present results for seven teams, since CTC-NER submitted results only for Russian. For each team, we present their best-performing model.[14]

As the plots show, the best performing model, CogComp, yields F-measure 91% according to the relaxed partial evaluation, and 85.6% according to the strict evaluation. Also, the only hybrid model, JRC-TMA-CC, reaches the highest precision—93.7% relaxed partial, and 88.6% strict—but lower recall—54.4% relaxed partial, 42.7% strict.

Five teams submitted results for *cross-lingual entity linking*. The best results for each team, averaged across two corpora, are presented in Figure 4, and in Table 2. The plots show that this task is much more difficult than entity extraction.

[14]Complete results available on the Workshop's Web page: bsnlp.cs.helsinki.fi/final_ranking.pdf

NORD STREAM		RYANAIR	
System	F1	System	F1
IIUWR.PL	41.5	IIUWR.PL	48.7
JRC-TMA	31.0	JRC-TMA	27.0
RIS	11.1	CogComp	13.0
CogComp	11.1	RIS	10.3
Sberiboba	05.6	Sberiboba	10.2

Table 2: Cross-lingual entity linking.

The best performing model, IIUWR.PL, yields F-measure 45%. As seen from the plot, for this task it is harder to balance recall and precision: the first two models obtain much higher precision, while the last three obtain much higher recall. The two best-performing models used rule-based entity linking.

Note that in our setting the performance on entity linking depends on performance on name recognition and normalization: a system had to link entities that it extracted from documents upstream, rather than link a correct set of entities.

Tables 3 and 4 present the F-measure for all tasks, split by language, for the RYANAIR and NORD STREAM datasets; Table 2 shows performance on the final phase—cross-lingual entity linking. We show one top-performing model for each team. For recognition, we present only the *relaxed* evaluation, since results obtained on the three evaluation schemes are correlated, as can be seen from Figure 3.

The tables indicate that the test corpora present approximately the same level of difficulty for the participating systems, since the values in both tables are similar. The only exception is *single-language* document linking, which seems to be much harder for the RYANAIR dataset, especially for Russian. This needs to be investigated further.

In Table 5 we present the results of the evaluation by entity type. As seen in the table, performance was higher overall for LOC and PER, and substantially lower for ORG and PRO, which corresponds with our findings from the First shared task, where ORG and MISC were the most problematic categories (Piskorski et al., 2017). The PRO category also exhibits higher variation across languages and corpora than other categories, which might point to some annotation artefacts. The results for the EVT category are less informative, since there are few examples of this category in the dataset, as seen in Table 1.

RYANAIR		*Language*							
Phase	*Metric*	bg		cz		pl		ru	
Recognition	Relaxed Partial	CogComp	87.5	CogComp	94.2	RIS	92.1	CogComp	94.3
		RIS	85.8	RIS	93.5	CogComp	91.4	RIS	92.5
		IIUWR.PL	75.9	IIUWR.PL	84.1	IIUWR.PL	84.1	CTC-NER	91.0
		TLR	75.9	TLR	82.2	TLR	82.2	TLR	83.4
		JRC-TMA	64.2	Sberiboba	80.5	Sberiboba	80.5	IIUWR.PL	78.9
		Sberiboba	64.6	JRC-TMA	53.6	JRC-TMA	53.6	JRC-TMA	63.7
		NLP Cube	14.7	NLP Cube	18.8	NLP Cube	18.0	Sberiboba	76.9
								NLP Cube	16.4
Normalization		CogComp	83.4	CogComp	88.7	RIS	87.4	RIS	91.3
		RIS	78.1	RIS	87.4	CogComp	86.3	CogComp	90.3
		TLR	68.3	IIUWR.PL	80.7	IIUWR.PL	78.9	CTC	85.9
		IIUWR.PL	68.0	Sberiboba	74.9	TLR	75.1	TLR	78.0
		JRC-TMA	61.3	TLR	72.5	Sberiboba	73.1	JRC-TMA	74.2
		Sberiboba	55.9	JRC-TMA	50.2	JRC-TMA	52.6	IIUWR.PL	73.5
		NLPCube	11.2	NLPCube	11.0	NLPCube	15.2	Sberiboba	66.9
								NLPCube	14.8
Entity linking	Document level	IIUWR.PL-5	35.5	IIUWR.PL	51.8	IIUWR.PL	58.6	IIUWR.PL	29.4
		JRC-TMA	15.8	JRC-TMA	51.7	JRC-TMA	54.6	CogComp	09.4
		CogComp	10.5	CogComp	16.7	CogComp	25.7	RIS	09.3
		RIS	07.1	Sberiboba	16.2	Sberiboba	23.2	CTC-NER	05.4
		Sberiboba	03.1	RIS	13.9	RIS	22.3	Sberiboba	05.4
								JRC-TMA	02.7
	Single language	IIUWR.PL	60.2	IIUWR.PL	70.0	IIUWR.PL	61.9	IIUWR.PL	55.9
		JRC-TMA	48.8	JRC-TMA	36.3	JRC-TMA	28.3	JRC-TMA	49.6
		CogComp	13.9	RIS	13.4	RIS	23.3	RIS	14.8
		RIS	07.4	Sberiboba	12.7	CogComp	23.1	CogComp	12.6
		Sberiboba	05.2	CogComp	11.3	Sberiboba	16.9	CTC-NER	12.4
		NLP Cube	02.0	NLP Cube	00.7	NLP Cube	02.0	Sberiboba	11.9
								NLP Cube	03.1

Table 3: F-measure results for the RYANAIR corpus

8 Conclusion

This paper reports on the Second Multilingual Named Entity Challenge, which focuses on recognizing mentions of NEs in Web documents in Slavic languages, normalization/lemmatization of NEs, and cross-lingual entity linking. The Challenge attracted much wider interest compared to the First Challenge in 2017, with 16 teams registering for the competition and eight teams submitting results from working systems, many with multiple systems variants. Many of the systems used state-of-the-art neural network models. Overall, the results of the best-performing systems are quite strong for extraction and normalization, while cross-lingual linking appears to be substantially more challenging.

We show summary results for the main aspects of the challenge and the best-performing model for each team. For detailed, in-depth evaluations of all submissions systems and their performance figures please consult the Shared Task's Web page.

To stimulate further research into NER for Slavic languages, including cross-lingual entity linking, our training and test datasets, the detailed annotations, and scripts used for evaluations are made available to the public on the Shared Task's Web page.[15] The annotation interface is released by the Inforex team, to support annotation of additional data for expanded future tests.

This challenge covered four Slavic languages. For future editions of the Challenge, we plan to expand the training and test datasets, covering a wider range of entity types, and supporting cross-lingual entity linking. We also plan to cover a wider set of languages, including *non-Slavic* ones, and recruit more annotators as the SIGSLAV community expands. We will also undertake further refinement of the underlying annotation guidelines—always a highly complex task in a real-world setting. More complex phenomena also need to addressed, e.g., coordinated NEs, contracted versions of multiple NEs, etc.

We hope that this work will stimulate research into robust, end-to-end NER solutions for processing real-world texts in Slavic languages.

[15] bsnlp.cs.helsinki.fi/shared_task.html

NORD STREAM 2		*Language*							
Phase	*Metric*	bg		cz		pl		ru	
Recognition	Relaxed Partial	RIS	89.6	CogComp	94.4	RIS	93.7	CTC-NER	86.1
		CogComp	89.4	RIS	94.1	CogComp	93.2	CogComp	85.9
		IIUWR.PL	84.5	IIUWR.PL	88.1	IIUWR.PL	91.3	RIS	84.8
		TLR	83.3	Sberiboba	84.3	Sberiboba	84.4	IIUWR.PL	76.5
		JRC-TMA	77.9	TLR	82.1	TLR	80.6	Sberiboba	73.5
		Sberiboba	73.3	JRC-TMA	65.9	JRC-TMA	59.3	TLR	73.1
		NLP Cube	16.4	NLP Cube	23.8	NLP Cube	15.2	JRC-TMA	69.5
								NLP Cube	17.1
Normalization		RIS	84.9	CogComp	89.3	RIS	89.2	RIS	78.0
		CogComp	84.3	RIS	89.1	CogComp	86.4	CogComp	72.5
		TLR	73.3	IIUWR.PL	83.3	IIUWR.PL	85.9	CTC-NER	69.4
		IIUWR.PL	70.7	TLR	74.4	TLR	72.0	IIUWR.PL	65.0
		JRC-TMA	66.7	Sberiboba	71.1	Sberiboba	67.9	Sberiboba	60.3
		Sberiboba	63.3	JRC-TMA	50.3	JRC-TMA	42.4	TLR	59.6
		NLP Cube	13.5	NLP Cube	15.6	NLP Cube	09.0	JRC-TMA	53.0
								NLP Cube	10.5
Entity linking	Document level	IIUWR.PL	46.8	IIUWR.PL	71.9	IIUWR.PL	74.3	IIUWR.PL	52.8
		JRC-TMA	17.0	CogComp	20.1	JRC-TMA	18.8	RIS	18.2
		RIS	11.3	RIS	19.0	CogComp	15.4	Sberiboba	14.6
		CogComp	10.3	Sberiboba	14.2	RIS	14.4	CogComp	12.3
		Sberiboba	08.6	JRC-TMA	11.5	Sberiboba	12.2	JRC-TMA	11.3
								CTC-NER	06.7
	Single language	IIUWR.PL	58.9	IIUWR.PL	67.2	IIUWR.PL	68.6	IIUWR.PL	48.8
		JRC-TMA	54.8	JRC-TMA	35.3	JRC-TMA	31.5	JRC-TMA	38.0
		RIS	12.1	RIS	20.1	RIS	15.5	RIS	08.8
		CogComp	10.6	CogComp	18.6	CogComp	14.4	CTC-NER	06.8
		Sberiboba	07.8	Sberiboba	08.7	Sberiboba	06.0	Sberiboba	05.9
		NLP Cube	01.0	NLP Cube	01.0	NLP Cube	01.3	CogComp	05.6
								NLP Cube	00.7

Table 4: Evaluation results (F-measure) for the NORD STREAM 2 corpus

	NORD STREAM				RYANAIR			
	bg	cs	pl	ru	bg	cs	pl	ru
Per	93.9	95.7	93.0	93.3	97.8	96.3	97.7	97.4
Loc	94.8	98.3	95.5	98.7	98.3	97.1	97.6	96.6
Org	85.1	95.0	95.5	92.5	90.1	90.1	89.9	83.4
Pro	59.5	79.6	54.1	65.1	72.8	92.3	90.4	57.1
Evt	0.50	0.55	100.0	-	50.0	18.2	50.0	40.0

Table 5: Recognition F-measure (relaxed partial) by entity type—best-performing systems for each language.

Acknowledgments

We thank Tomek Bernaś, Anastasia Golovina, Natalia Novikova, Elena Shukshina, Yana Vorobieva, and Alina Zaharova and many others for contributing to data annotation. We also thank Petya Osenova and Kiril Simov for their assistance.

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

Work was supported in part by investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

Work was supported in part by ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no. CZ.02.1.01 / 0.0 / 0.0 / 17_048/0007267), and by Grant No. SGS-2019-018 "Processing of heterogeneous data and its specialized applications."

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.

AY Antonova and AN Soloviev. 2013. Conditional random field models for the processing of Russian. In *Computational Linguistics and Intellectual Technologies: Papers From the Annual Conference "Dialogue"(Bekasovo, 29 May–2 June 2013)*, volume 1, pages 27–44.

Mikhail Arkhipov, MAria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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.

Maud Ehrmann, Guillaume Jacquet, and Ralf Steinberger. 2017. Jrc-names: Multilingual entity name variants and titles as linked data. *Semantic Web*, 8(2):283–295.

Anton Emelianov and Ekaterina Artemova. 2019. Multilingual named entity recognition using pretrained embeddings, attention mechanism and NCRF. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

Guillaume Jacquet, Jakub Piskorski, and Sophie Chesney. 2019a. Out-of-context fine-grained multi-word entity classification: exploring token, character N-gram and NN-based models for multilingual entity classification. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1001–1010. ACM.

Guillaume Jacquet, Jakub Piskorski, Hristo Tanev, and Ralf Steinberger. 2019b. JRC TMA-CC: Slavic named entity recognition and linking. participation in the BSNLP-2019 shared task. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

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.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1064–1074.

Michal Marcinczuk, Marcin Oleksy, and Jan Kocon. 2017. Inforex - a collaborative system for text corpora annotation and analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2-8, 2017*, pages 473–482. INCOMA Ltd.

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, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 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.

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.

Nils Reimers and Iryna Gurevych. 2019. Alternative weighting schemes for elmo embeddings. *arXiv preprint arXiv:1904.02954*.

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. 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.

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.

Tatiana Tsygankova, Stephen Mayhew, and Dan Roth. 2019. BSNLP2019 shared task submission: Multisource neural NER transfer. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. 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.