Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4647–4652
Marseille, 11–16 May 2020

Reconstructing NER Corpora: a Case Study on Bulgarian

Iva Marinova, Laska Laskova, Petya Osenova, Kiril Simov, Alexander Popov
LMaKP, IICT-BAS
Sofia, Buglaria
iva.marinova@identrics.net
{laska, petya, kivs, alex.popov}@bultreebank.org

Abstract

The paper reports on the usage of deep learning methods for improving a Named Entity Recognition (NER) training corpus and for predicting and annotating new types in a test corpus. We show how the annotations in a type-based corpus of named entities (NE) were populated as occurrences within it, thus ensuring density of the training information. A deep learning model was adopted for discovering inconsistencies in the initial annotation and for learning new NE types. The evaluation results get improved after data curation, randomization and deduplication.

Keywords: NER corpora, Bulgarian, deep learning

1. Introduction

Named Entity Recognition (NER) and Named Entity Linking (NEL) have been active research areas for many years now, encompassing a wide range of topics: from rule-based systems to neural networks; from coarse-grained classification tasks (person, location, organization tagsets) to more fine-grained ones (including products, events, etc.); from local and global approaches to combined ones, etc. While these two related tasks are considered to be well covered in NLP for Germanic, Romance and other language groups, they are under-resourced for the Slavic languages, especially from a multilingual perspective.

Recently, a shared task was launched for Slavic languages on NER and linking to the same real-world entity in a cross-lingual context. The second edition in 2019 (Piskorski et al., 2019) included the following languages: Bulgarian, Czech, Polish, Russian. It focused on 5 types (or sorts and categories) of Named Entities (NEs): Person (PER), Location (LOC), Organization (ORG), Event (EVT), and Product (PRO).

In the work reported here we exploit this newly constructed Bulgarian dataset for NER. The second dataset was extracted from the Bulgarian HPSG-based treebank — BulTreeBank (Simov et al., 2002). The named entities in BulTreeBank were syntactically annotated and the category for each NE was assigned to the highest node in the syntactic annotation. For the aims of this work, the categories were inherited to the token level. In BulTreeBank, only form categories were annotated: Person (PER), Location (LOC), Organization (ORG), and Other (OTH). The category *Other* covers a wide range of entity types like Event (EVT), Product (PRO), and some of their sub-types — different kinds of products, for example.

The preparation of the Bulgarian corpus within a highly multilingual guidance environment was very useful for improvement on two fronts: the quality of the gold data, as well as the prediction power of the deep learning NER models trained on it. The former was achieved through the NE occurrences population and through discovering errors made by the deep model. The latter was achieved by the application of a deep learning model over the corpus extracted from BulTreeBank, in which the product type (PRO) was masked behind the more general type OTH, while the event type (EVT) had not been not annotated at all.

We focused only on Bulgarian, out of all the languages covered in the shared task, since we already had at our disposal the language-dependent processing tools. However, the approach we propose is applicable to other languages, given the availability of the appropriate resources and tools.

Our contributions are as follows: a) we show how type-centered information in a corpus can be token (occurrence) populated within the texts; b) with the help of a deep learning model, the annotations have been improved and new entities have been discovered within the data; c) the improved annotations, randomization and deduplication have led to better results on the NER task; d) the same model, trained on the improved dataset, has been used to populate the EVT and PRO categories within the golden NE datasets extracted from BulTreeBank. This step will further enrich the available language resources for Bulgarian with a contemporary and a much larger NER corpus. All the language resources and tools used for achieving the results reported here will be made available.

The structure of the paper is as follows: Section 2. discusses research relevant to our work; in Section 3., the data is described in more detail; Section 4. presents the methods that we have used, as well as the experiments that have been conducted. Section 5. concludes the paper.

2. Related Work

The Second Multilingual Named Entity Challenge in Slavic languages (Piskorski et al., 2019) explores the NER task as part of a more complex solution including the recognition of mentions of named entities in Web documents, their normalization, and cross-lingual linking. The challenge was performed on four languages: Bulgarian, Czech, Polish, Russian. The best score on the NER task for Bulgarian concerning the Ryanair corpus was **F1=87.5**, and for the Nord Stream corpus it was **F1=89.6**. Since the test data contains both topics, in table 5 we report the mean score on the whole testset **F1=88.55**.

Piskorski et al. (2019) report a ***relaxed evaluation*** metric, which means that 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 of whether the extracted mention was in its base form). All the other systems reported in this paper, including ours, use the ***strict evaluation*** introduced at the SemEval'13 task.

Georgiev et al. (2009) employ a rich set of features in their solution. At that time, using conditional random fields (CRFs) was the dominant approach to NER, but it required extensive and manual feature engineering, especially for morphologically rich languages like Bulgarian. Their work was mostly devoted to constructing a set of orthographic and domain-specific features. Using gazetteers, local/nonlocal morphology, feature induction and mutual information in the form of unlabeled texts, they achieved **F1=89.40**.

Simeonova et al. (2019) employ an LSTM-CRF architecture on top of a word embedding layer. The authors explore a rich feature set for the language, using the position-aware morphosyntactic tags proposed by Simov and Osenova (2004). The word embeddings used in these experiments are Bulgarian FastText Vectors by Bojanowski et al. (2017). The final vector representations of the words are formed by combining FastText with character embeddings, thus further improving the test scores by using POS and morphological representations. The best score achieved by this system is **F1=92.20**.

The identification of named entity mentions in texts is often implemented using a sequence tagger, where each token is labeled with an IOB tag, indicating whether the token begins an NE — (B), whether it is inside of an NE (I), or whether it is outside of an NE (O). This type of annotation has been proposed for the first time at CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003). Here we follow this format of representation. Also, we focus only on the NER task, leaving NEL as a next step.

In our experiments we use the setup proposed in Marinova (2019), with **F1=96.29** on the BulTreeBank dataset. The model is explored further in Section 4.

3. Data Preparation

The corpus we have used for the work presented here was annotated by our team in accordance with the guidelines for the Second Multilingual Named Entity Challenge in Slavic languages organized for the 7th Balto-Slavic NLP Workshop 2019 (Piskorski et al., 2019). The shared task covered four Slavic languages – Bulgarian, Czech, Polish, and Russian – with Bulgarian being the only one that does not express the grammatical function of a noun phrase with an inflectional ending, and at the same time having adopted several types of endings as definite articles. We will address the effects that stem from these structural differences again in the following paragraphs. The multilingual NER task was divided into three parts (Piskorski et al., 2019): (1) NE mention detection and classification, (2) name normalization, and (3) entity linking. The first sub-task was to identify all the entities of **type** person (PER), organization (ORG), location (LOC), product (PRO), and event (EVT) named in a document. The NER system was not expected to output all the mentions of a given NE, but only the ones that exhibit some sort of grammatical idiosyncrasy.

Like in most NER tasks, NEs are considered to be non-recursive, non-overlapping, and whenever one NE is embedded in another NE, only the top-most entity is annotated (Yadav and Bethard, 2019). The evaluation was not case-sensitive, so the typographical variations were not taken into consideration; the same applied to common nouns and pronouns that were co-referential with NEs.

For the second sub-task, name normalization, each NE had to be paired with its so-called *canonical multi-word expression* or lemma: nominative case for the head and possibly the other components of a multi-word expression (Polish: Instytutu Jagiellońskiego, 'of the Jagiellonian Institute' ↔ Instytut Jagielloński, 'Jagiellonian Institute'), article-less form for Bulgarian (Европейския съюз, 'the European Union' ↔ Европейски съюз, 'European Union'), proper nouns from which possessive adjectives are derived (Czech: Trumpova (slova), 'Trump's (words)' ↔ Trump), corrected case form (Fox news ↔ Fox News).

The third sub-task was to assign a unique cross-lingual identifier to the NEs that referred to the same real-world entity.

The original Bulgarian corpus consists of 916 text files extracted from various news websites. The training dataset contains information on two topics – Brexit and the trial of the Pakistani Christian Asia Bibi, accused of blasphemy, while the main subjects for the test data are the Nord Stream 2 project and the recent developments in RyanAir's business history. Each document contains meta-data (document ID, language, creation date, URL, and title) represented at the beginning of the document and separated from the rest with an empty line (Table 1). The output or the annotation file contains all NE mentions followed by their lemma, type and cross-lingual ID. The list is organized in alphabetical order (Table 2). It is easy to see that the lemmas in the second column are identical to the mentions from the first column. As we have already indicated, in contrast to other Slavic languages, Bulgarian has no noun declension system, but it has a definite article. This results in a much poorer noun and adjective paradigm. For this reason, most Bulgarian NEs have only one grammatical form or; in other words, the mentions and the lemmas coincide.

Some variation might be expected, given that a multi-word NE, which typically refers to an organization, contains an adjective that – depending on its grammatical function – might assume a definite article. For example, a NE with a masculine noun as a head such as Европейски съюз, 'European Union' in subject position takes the form Европейският съюз, while in any other position it should be Европейския съюз. For comparison, the corresponding Russian NE Европейский союз has five different case forms: Европейский союз, Европейского союза, Европейскому союзу, Европейским союзом, Европейском союзе. Thus, name normalization proves to be a trivial task for Bulgarian.

The two datasets were pre-processed with the BTB-Pipeline (Savkov et al., 2011) in order to convert the BSNLP format to the well-established and more common

bg-108
bg
10/3/2018
https://business.dir.bg/kompanii/polagat-trabite-na-severen-potok-2-v-ruski-vodi-prez-noemvri
Полагат тръбите на “Северен поток” в руски води през ноември

Полагането на газопровода “Северен поток-2” в руски води ще започне в края на ноември, съобщи пред медии техническият директор на компанията оператор Норд стрийм 2 (Nord Stream 2) Сергей Сердюков, цитиран от ТАСС. Тръбопроводът трябва да мине по дъното на Балтийско море от руското крайбрежие до германския бряг. Капацитетът на двете му линии е 27,5 млрд. куб метра годишно и той ще удвои мощността на първия “Северен поток”, чийто маршрут като цяло повтаря. “Северен поток-2” ще заобиколи транзитните страни Украйна, Беларус и Полша и ...

Table 1: BSNLP input format.

bg-108			
Nord Stream 2	Nord Stream 2	ORG	ORG-Nord-Stream-2-AG
Балтийско море	Балтийско море	LOC	LOC-Baltic-Sea
Беларус	Беларус	LOC	LOC-Baltic-Sea
Норд стрийм 2	Норд стрийм 2	ORG	ORG-Nord-Stream-2-AG
Полша	Полша	LOC	GPE-Poland
Северен поток-2	Северен поток-2	PRO	PRO-Nord-Stream-1
Сергей Сердюков	Сергей Сердюков	PER	PER-Sergey-Serdyukov
ТАСС	ТАСС	ORG	ORG-TASS-Agency
Украйна	Украйна	LOC	GPE-Ukraine

Table 2: BSNLP annotation format.

BIO format for the purpose of experimentation.

All the metadata, except for the title, was stripped, the input files were segmented into sentences and tokens per line, and each token was combined with its corresponding POS tag and lemma. Then the text and the output files were merged in order to propagate the NE annotations within the text documents.

The POS and the lemma markup was used to annotate the text with the NEs listed in the original output file (see Table 2). For this phase of the pre-processing we used regular grammars to identify the NEs and assign to them their corresponding lemmas.

In cases where a NE is ambiguous and thus refers to more than one real-word entity, we assume that the lemma is a concatenation of the lemmas of the different words, type(s) and IDs. For example, there are mentions of two different people who share the same surname:

Браун → Lemma=Браун Type=“PER”
ID=“PER-Kerry-Brown; PER-Mark-Brown”

In the following example, not only the identifiers but also the types are different:

Азия → Lemma=Азия Type=“LOC;PER”
ID=“LOC-Asia; PER-Asia-Bibi”

Here, the NE might refer to the continent “Asia” or, as is often the case with the documents from the Asia Bibi dataset, to a person.

The disambiguation in such cases was performed in two steps: (1) automatic disambiguation with only one of these readings being used; and (2) manual selection for the cases that were not resolved after the first step. Only about 10 % of the cases were disambiguated manually.

The final BIO format is illustrated in Table 3., where the first token of a named entity is annotated with one of the tags: B-PER, B-LOC, B-ORG, B-PRO, and B-EVT, for the beginning of the NE. The following tokens of the named entities are annotated with I-PER, I-LOC, I-ORG, I-PRO, and I-EVT. The tokens that are not part of a named entity are annotated by the tag O.

Полагат	O	
тръбите	O	
на	O	
“	O	
Северен	B-PRO	B-PRO-Nord-Stream-2
поток	I-PRO	I-PRO-Nord-Stream-2
”	O	
в	O	
руски	O	
води	O	
през	O	
ноември	O	

Table 3: BIO representation of the title from the above example obtained after the BTB-Pipeline processing. We also keep the entity identifiers like B-PRO-Nord-Stream-2 and I-PRO-Nord-Stream-2.

The initial distribution of the documents in the training and the test datasets (i.e. the one following the design of the shared task) was found suboptimal after an analysis of the results from the first series of experiments. This led us to the conclusion that we have to improve the quality of the annotation of the two sets. First of all, there was a substantial number of documents with similar content. This is due to the fact that many websites copy the information provided by one news agency and publish it without

	Brexit	Asia Bibi	Nord Stream	RyanAir	Brexit	Asia Bibi	Nord Stream	RyanAir
	TRAINING				TEST			
	INITIAL SETUP							
1	600	99	0	0	0	0	130	87
2	596	**23**	0	0	0	0	130	78
	FINAL SETUP							
3	567	21	30	20	30	2	100	58

Table 4: Training and test datasets. Rows 2 and 3 show the number of documents after de-duplication.

any changes. Some of the topics, especially that of Asia Bibi, proved to have been of little interest to Bulgarian news outlets, which had reduced the number of unique collected documents even further.

Secondly, some types of NEs were distributed unevenly between the sets dedicates to various topics, and what is worse, in the training and the test sets (see further discussion in Section 4.) For that reason, we decided to randomize the initial partition.

Thirdly, we discovered some inconsistencies in the annotation that had to be corrected. Some NEs were left out, for example the service "My RyanAir", and more than 10 multi-word names, based on the function of the referred administrative bodies, such as Транспорт и околна среда, 'Transport&Environment' or Столична община, 'Sofia Municipality'. Another interesting case was presented by two capitalized common words, Острова, 'the Island' and Залива, 'the Gulf', the first referring to the island of Great Britain and, by metonymy, to the state (LOC in both cases), and the latter referring to the Persian Gulf. Furthermore, we decided to tag as NEs all capitalized partial mentions of various organizations such as Комисията for the European Commission (Европейска комисия) and Съюзът for the European Union (Европейски съюз). These cases represent an interesting trend in the contemporary writing style that becomes more and more commonplace in Bulgarian newswire articles. We plan to use this annotation in our work on co-reference resolution.

Table 4. presents the change in the number of documents, in both training and test sets, after annotation correction, deduplication, and randomization.

4. Methods and Experiments

In this section, we first present our approach for training the neural network models, and then — the results from the experiments.

4.1. Methods

For the experimental work, Flair[1] was used, an NLP library implemented by Zalando Research on top of PyTorch[2]. Apart from their own pre-trained Flair contextualized string embeddings (Akbik et al., 2019b), the library provides access to many other state-of-the-art language models, such as FastText (Grave et al., 2018), GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018) and the Transformers provided by HuggingFace[3]: BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet .

Stacking the embeddings is one of the most important features of the library and the functionality is used in the experiments to concatenate language models together. The developers of the library claim that this method often gives the best results and lately has become a common technique in sequence labeling models (Grave et al., 2018). The pre-trained language models used for the embedding layer are described in the following paragraphs.

The first one is the **Bulgarian flair-forward and -backward** model trained by Stefan Schweter.[4] The author of the forward and backward Bulgarian language models uses data from a recent Wikipedia dump and corpora from OPUS parallel corpora. Training was done for one epoch over the full training corpus, which for Bulgarian consists of 111,336,781 tokens. The hyperparameters used to train the contextual string embeddings are the following: the hidden vector size is 2048; the number of the hidden layers is 1; the sequence length is 250; and the mini batch size is 100. One model is trained in a forward direction and one backward; combining them by concatenation contributes to the contextual vector representation of the words.

The second language model used in the experiments is **FastText**[5], obtained using CBOW ((Mikolov et al., 2013)) with position-weights and dimensionality of 300, character n-grams of length 5, a window of size 5 and 10 negative samples as described in "Learning Word Vectors for 157 Languages" (Grave et al., 2018).

Additionally, we use character embeddings obtained from the corpus.

These language models are stacked at the embedding layer of the BiLSTM-CRF classifier. Flair authors describe the use of stacked embeddings in Akbik et al. (2019a).

4.2. Experiments

It is interesting to see how various modifications have influenced the experimental results. For example, the baseline showed very low results for EVT and PRO. This is because there are over 2200 EVT annotations in the training data as the Brexit topic is full of political events, but there are only 15 EVT annotated entities in the test data. The events in the test data are not semantically related to the events in the training data, but there are still some events that are rec-

[1] https://github.com/zalandoresearch/flair

[2] https://pytorch.org/

[3] https://github.com/huggingface

[4] https://github.com/stefan-it

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

Authors	Method	Data	Categories	F1
(Georgiev et al., 2009)	SVM	BTB NE	ORG | PER | LOC | OTH	89.40
(Simeonova et al., 2019)	LSTM-CRF	BTB NE	ORG|PER|LOC|OTH	92.20
(Marinova, 2019)	BiLSTM-CRF	BTB NE	ORG|PER|LOC|OTH	**96.29**
(Piskorski et al., 2019)	BILSTM-CRF	BSNLP	ORG|PER|LOC|PRO|EVT	***88.55***
1. Test on RyanAir and Nord Stream	BiLSTM-CRF	BSNLP†	ORG|PER|LOC|PRO|EVT	86.36
2. Test on all topics	BILSTM-CRF	BSNLP‡	ORG|PER|LOC|PRO|EVT	**95.77**
3. Test on BTB dataset	BILSTM-CRF	BSNLP‡ + BTB NE	ORG|PER|LOC|PRO|EVT|OTH	**76.50**

Table 5: Summary of all available NER systems for Bulgarian compared to our experiments. All results are Micro Averaged F1 strict evaluation except of original BSNLP evaluation which is relaxed F1. † = Semi-automatically annotated and curated ; ‡ = Semi-automatically annotated, curated, de-duplicated and randomized.

Experiment	ORG	PER	LOC	PRO	EVT
1. Test on RyanAir and Nord Stream	83.69	92.99	96.83	38.77	41.66
2. Test on all topics	92.28	96.81	98.63	91.42	97.56
3. Test on BTB dataset	69.9	84.83	87.54	TBA	TBA

Table 6: Results per Class.

ognized there, such as "Великден" (Easter), even if they were not present in the training data. The PRO category, on the other hand, is widely covered in the topics streams for RyanAir and Nord Stream in the test data, but has no equivalent in the train split where the only products are in the form of media content providers.

The randomization, deduplication of the dataset, together with the curation, substantially improved the annotation of EVT and PRO, as well as the annotation of the other types. This is due to the randomization of the topics. In this way we used documents from all the topics to train and test our model, which led to improved test scores. The deduplication ensured the sustainability and reliability of the results.

The main source of errors are the embedded named entities. For example, *Brexit* is classified under EVT type. However, when it is part of a product name like 'законопроектът за Брекзит' (the so-called *Brexit bill*) it is mistaken with EVT again. Another source of errors is the confusion between ORG and PRO, and between EVT and ORG. For example, the news agency Ведомости (Vedomosti) is annotated as ORG, but it is PRO. Also, the event Международен арктически форум (International Arctic Forum) has been annotated as ORG instead of EVT.

In the next experiment, the model was tested on the gold data from BulTreeBank. The idea was to re-annotate the type OTHER and ORG in the data as PRO where appropriate, as well as to annotate EVTs (such as the Second World War or the Russian-Turkish War, or the April Rebellion). We manually evaluated the newly found entities in the data. The model discovered and strictly annotated 208 entities of type PRO and 33 entities of type EVT.

Some noise was inevitably introduced via this process. For example, a residence quarter (кв. Младост, kv. Mladost) should be tagged as LOC, but was annotated as PRO. Some of the OTHER annotations were wrongly re-annotated as EVT; for instance: the food supplement Розкалипт (Rozkalipt) or the football team Берое (Beroe). Interestingly, a LOC in a non-standard form was annotated as EVT as well; an example of this type of error is Европата (the Europe).

One final round of experiments were conducted in order to determine optimal regularization rates for training and evaluation on the curated and randomized BSNLP data. Training was done using an analogous BiLSTM-CRF architecture with the following values for the hyperparameters: batch size = 32; number of training epochs = 100; 1 hidden layer; size of the hidden layer = 256; optimizer = Adam. The models all have a stacked embedding layer combining FastText embeddings, Flair-backward and Flair-forward embeddings, as well as byte-pair and character-level embeddings. The only parameter that varies in the reported results is the amount of dropout applied to the output vectors (the concatenated hidden states from the BiLSTM). Table 7 shows the results obtained with different regularization levels. The optimal value for this architecture is found at 0.6 – compared to the result for no regularization, this model fares much better, demonstrating that the randomization allows for almost perfect learning of the topics in the corpus. Of course, it must be borne in mind that such a model would probably behave differently when presented with a corpus that is more heterogeneous and that contains a greater diversity of entities.

Dropout	F1
0.0	92.31
0.1	94.63
0.2	98.25
0.3	94.27
0.4	94.94
0.5	94.45
0.6	**98.60**
0.7	92.27
0.8	96.66

Table 7: F1-scores on the BSNLP data (semi-automatically annotated, curated and randomized, including all five types of entities), with varying amounts of regularization.

5. Conclusions and Future work

Our work has explored the following language resources: Piskorski et al. (2019) and Simov et al. (2002). We had several goals: (1) to annotate NEs in BulTreeBank with EVT and PRO categories, using a neural network model trained on the BSNLP data; (2) to analyze the compatibility of both datasets with respect to possible errors in the manual annotations; and (3) to achieve a new state of the art for the NER task for Bulgarian. In our future work we plan to explore the possibility of combining these datasets into a new BIO-formatted gold corpus for Named Entity Recognition in Bulgarian. Besides the categories of Person (PER), Location (LOC), Organization (ORG), Event (EVT), and Product (PRO), we plan to add identifiers for all recognized NEs, including links to the Bulgarian Wikipedia.

The data in the BTB dataset is rich in terms of the topics covered and of the diversity of the NE mentions, but it consists of data collected in the period 2000-2001. Its enrichment with the newly crawled data from the 2019 BSNLP task will improve the NER data for Bulgarian and will be a stable ground for further multilingual experiments and tasks devoted to NER. Our next steps also include: the curation of the predicted EVT and PRO categories in the BTB dataset; the alignment of the categories in both datasets. The data will be further randomized and split into train, development, and test parts.

The semi-automatically annotated and randomized datasets are available at `https://github.com/usmiva/bg-ner`.

Acknowledgements

This research was partially supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-272/16.12.2019. Alexander Popov's contribution has been supported by *the Bulgarian Ministry of Education and Science under the National Research Programme "Young scientists and postdoctoral students"*, approved by DCM # 577 / 17.08.2018.

6. Bibliographical References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019a). Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Akbik, A., Bergmann, T., and Vollgraf, R. (2019b). Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 724–728.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Georgiev, G., Nakov, P., Ganchev, K., Osenova, P., and Simov, K. (2009). Feature-rich named entity recognition for bulgarian using conditional random fields. In *Proceedings of the International Conference RANLP-2009*, pages 113–117.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

Marinova, I. (2019). Evaluation of stacked embeddings for bulgarian on the downstream tasks pos and nerc. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 48–54, Varna, Bulgaria, September. INCOMA Ltd.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proc. of NAACL.*

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Přibáň, P., Steinberger, J., and 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*, pages 63–74, Florence, Italy, August. Association for Computational Linguistics.

Savkov, A., Laskova, L., Osenova, P., Simov, K., and Kancheva, S. (2011). A web-based morphological tagger for bulgarian. *Slovko*, pages 126–137.

Simeonova, L., Simov, K., Osenova, P., and Nakov, P. (2019). A morpho-syntactically informed lstm-crf model for named entity recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2019*, page in print. Association for Computational Linguistics.

Simov, K. and Osenova, P. (2004). BTB-TR04: BulTreeBank morphosyntactic annotation of Bulgarian texts. Technical report, Technical Report BTB-TR04, Bulgarian Academy of Sciences.

Simov, K., Osenova, P., Slavcheva, M., Kolkovska, S., Balabanova, E., Doikoff, D., Ivanova, K., Simov, A., and Kouylekov, M. (2002). Building a linguistically interpreted corpus of bulgarian: the bultreebank. In *LREC.*

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050.*

Yadav, V. and Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models.