Evaluation of Stacked Embeddings for Bulgarian on the Downstream Tasks POS and NERC

Iva Marinova
LMaKP
IICT-BAS
Sofia, Bulgaria
iva.marinova@identrics.net

Abstract

This paper reports on experiments with different stacks of word embeddings and evaluation of their usefulness for Bulgarian downstream tasks such as Named Entity Recognition and Classification (NERC) and Part-of-speech (POS) Tagging. Word embeddings stay in the core of the development of NLP, with several key language models being created over the last two years like FastText (Bojanowski et al., 2017), ElMo (Peters et al., 2018), BERT (Devlin et al., 2018) and Flair (Akbik et al., 2018). Stacking or combining different word embeddings is another technique used in this paper and still not reported for Bulgarian NERC. Well-established architecture is used for the sequence tagging task such as BI-LSTM-CRF, and different pre-trained language models are combined in the embedding layer to decide which combination of them scores better.

1 Introduction

In this paper are reported the initial experiments for my PhD project which final goal is to build a system for extraction and classification of named entities, events and the relations between them from Bulgarian texts. The evaluation of the recent language models for Bulgarian is sufficient for my future work as it involves tasks such as NERC, Event Classification and Relation Extraction. All of them are considered downstream tasks and are often used to evaluate the language models and their usefulness. Currently, the tasks Event Classification and Relation Extraction are not addressed sufficiently. The data is available within the Bulgarian National Research Infrastructure for Language, Culture and History Resources and Tools — CLaDA-BG. In further experiments I will proceed with these data.

NERC and Event classification are considered both as sequence tagging tasks. Such tasks in the available manually annotated data from BulTreeBank (BTB) Project (Simov et al., 2002) are the part-of-speech tags and the named entities encoded in IOB (inside-outside-beginning) format.

Recent work on sequence tagging shows that BI-LSTM-CRF as proposed by (Huang et al., 2015) is the dominant solution applied to many different languages. This paper introduced the BIdirectional LSTM with CRF classfier (denoted as BI-LSTM-CRF) model to NLP sequence tagging tasks. The authors show that their model can efficiently use both past and future input features due to the bidirectional application of the LSTM component and use the sentence level tag information thanks to the CRF layer. This architecture reports state-of-the-art accuracy on POS, chunking and NERC tasks.

Here NERC and POS tagging are employed as fundamental tasks for the future experiments with Named Entity Recognition, Event Classification and Relation Extraction for Bulgarian texts. The next step will be to simultaneously solve these tasks together in a combined multitask model. These experiments should improve the interaction with linguistic information for Bulgarian.

The structure of the paper is as follows: the next section outlines the related work; description of the architecture and results of the experiments are available in Section 3; the last section concludes the paper and provides some ideas for future work.

2 Related Work

The identification of named entity (NE) mentions in texts is often implemented using a sequence tag-

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 48–54,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_008

ger, where each token is labeled with an IOB tag, indicating whether the token is beginning of a NE — (B), whether it is inside of a NE (I) or it is outside of a NE (O). This type of annotation is first proposed at CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003). The Bulgarian data is annotated with the same tags as the proposed in the above publication: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC, and O. In this way not only the structure of the NE is represented, but also its category. An example of an annotated sentence — Върна ли книгата на Петър Илиев? (Did you return the book to Peter Iliev?) — from the BulTreeBank is given here:

Върна	O
ли	O
книгата	O
на	O
Петър	B-PER
Илиев	I-PER
?	O

The NE *Peter Iliev* is annotated with the tags for PERSON marking the first token as a beginning of the NE and the second token as an internal token of the same NE. All other tokens are annotated by the tag O as outside tokens.

This dataset is used in some of the works on Bulgarian NERC, but in different splits and/or with some additions explained further.

(Georgiev et al., 2009) employ a rich set of features in their solution. At that time, CRFs was the dominant approach to NERC, but it required extensive and manual feature engineering, especially for morphological rich languages like Bulgarian. Their work was mostly devoted to construct a set of orthographic and domain-specific features. Using gazetteers, local/non-local morphology, feature induction and mutual information in the form of unlabeled texts they achieve **F1=89.40**. They used a development set during the training in order to improve the model and finally evaluated the model over the test set. The data split sizes are as follows: the training set contains 8,896 sentences; the development set contains 1,779 sentences; and the testing set contains 2,000 sentences.

The same data from BTB, with some additional data, is used by (Simeonova et al., 2019). The difference is that the supplement was annotated only on token level and the original data was annotated syntactically. In the current experiments this addition is not used.

(Simeonova et al., 2019) use LSTM-CRF on top of a word embedding layer too, but the authors employ morphosyntactic features in the data, using the position-aware morphosyntactic tags proposed by (Simov and Osenova, 2004). The word embeddings used in their experiments are Bulgarian FastText Vectors by (Bojanowski et al., 2017). They form the final vector representations of the word by combining FastText with character embeddings and further improve the test scores with POS and morphological representations. The best score achieved by their system is **F1=92.20**.

Since the data split used by (Georgiev et al., 2009) was not found and the new data used by (Simeonova et al., 2019) were not used in these experiments, the results from the experiments reported in this paper are not directly comparable with theirs.

Recently the Second Multilingual Named Entity Challenge in Slavic languages (Piskorski et al., 2019) explores the NERC task as part of a more complex solution including recognizing mentions of named entities in Web documents, their normalization, and cross-lingual linking. The challenge was performed on four languages including Bulgarian. The best achieved score for Bulgarian is **F1=87.5**. The data annotated within the shared task is in different format and is not used in my experiments.

There are many more works devoted to the POS tagging task for Bulgarian such as (Georgiev et al., 2012) and (Popov, 2016). (Georgiev et al., 2012) use guided learning, lexicon and rules and explore different tag sets achieving accuracy of 97.98, 98.85 and 99.30 with respectively 680, 49 and 13 tags.

Here their Table 5 is extended with results from the experiments done after the publishing of (Georgiev et al., 2012) including my own . Consult Table 1 for the complete overview. In the next section the experiment setup and the achieved results are described further.

3 Experiments

For the development of the models is used Flair[1], an NLP library implemented by Zalando Research on top of PyTorch[2]. Apart from their own pre-trained Flair contextualized string embeddings

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

[2] https://pytorch.org/

Tool/Authors	Method	Tags	Acc.
Tree Tagger	Decision Trees	680	89.2
ACOPOST	Memory-based learning	680	89.91
SVMTool	SVM	680	92.22
TnT	HMM	680	92.53
(Georgiev et al., 2009)	Guided learning	680	90.34
(Simov and Osenova, 2001)	RNN	160	92.87
(Georgiev et al., 2009)	Guided learning	95	94.4
(Savkov et al., 2011)	SVM + Lexicon + Rules	680	94.65
Tanev and Mitkov 2002	Rules	303	95.00
(Simov and Osenova, 2001)	RNN	15	95.17
Doychinova and Mihov 2004	Transform-based learning	40	95.50
Doychinova and Mihov 2004	Rules + Lexicon	40	98.40
(Georgiev et al., 2012)	Guided learning	680	95.72
(Georgiev et al., 2012)	Guided learning + Lexicon	680	**97.83**
(Georgiev et al., 2012)	Guided learning + Lexicon + Rules	680	97.98
(Georgiev et al., 2012)	Guided learning + Lexicon + Rules	49	98.85
(Georgiev et al., 2012)	Guided learning + Lexicon + Rules	**13**	**99.30**
(Popov, 2016)	BiLSTM Word Embeddings 100 (neurons)	153	91.45
(Popov, 2016)	BiLSTM Word Embeddings 125 (neurons)	153	91.13
(Popov, 2016)	BiLSTM Word + Suffix Embeddings 125 (neurons)	153	94.47
(Plank et al., 2016)	BiLSTM	16	97.97
(Yu et al., 2017)	CNN	16	98.23
(Yasunaga et al., 2017)	Adversarial training	16	98.53
experiment 1	BI-LSTM-CRF + Stacked embeddings (bg + flair-fast + char)	**16**	**98.90**
experiment 2	BI-LSTM-CRF + Stacked embeddings (bg + flair + char)	**16**	**99.10**

Table 1: Summary of all available POS systems for Bulgarian with different tag sets.

(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), BERT (Devlin et al., 2018).

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 as the developers claim that this method often gives best results and lately has become a common technique in sequence labeling models.

3.1 NE Recognition and Classification

The BTB dataset consist of 14,732 sentences from different genres like newspapers articles, legal documents — the Bulgarian Constitution, some user generated data, literature, etc. Data is split into training, development, and test sets. The sizes of the sets are as follows: the training set contains 10,979 sentences; the development set contains 1,487 sentences; and the testing set contains 2,266 sentences.

The hyperparameters used to train the BI-LSTM-CRF are as follows: the hidden vector size is 256; the learning rate is set to 0.1; the sequence length is 250; mini batch size is 32; and number of max epochs is 150. The model architecture as defined by (Huang et al., 2015) is depicted on Figure 1.

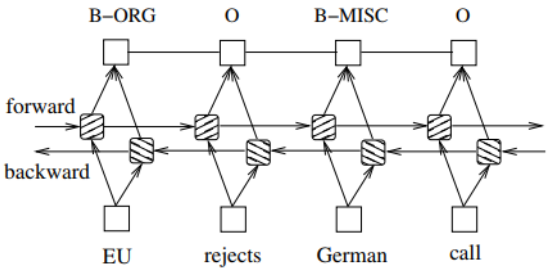

Figure 1: Bidirectional LSTM-CRF for Sequence Tagging (Huang et al., 2015)

The pre-trained language models used for the embedding layer are the following:

First, **BERT-base-multilingual-cased** model trained by (Devlin et al., 2018). This is their multilingual model. It is trained on 104 languages — the top languages with the largest Wikipedias. The model is implement as a 12-layer, 768-hidden, 12-heads, 110M parameters Bidirectional Transformer.

Second, **Bulgarian flair-forward and -backward** model trained by Stefan Schweter.[3] The author of the forward and backward Bulgarian

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

language models uses data from recent Wikipedia dump and corpora from OPUS. Training was done for one epoch over the full training corpus, which in 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 and combining them by concatenation contributes to the contextual vector representation of the words. There are two available -forward and -backward coupled models for Bulgarian:

bg-forward,**bg-backward**

and

bg-forward-fast,**bg-backward-fast**

The -fast models are CPU friendly and lightweight to train allowing for easy experimentation with a little damage to the result. The authors use vanilla SGD with no momentum, clipping gradients at 5 and employ a simple learning rate annealing method in which they halve the learning rate if training loss does not fall for 5 consecutive epochs (Akbik et al., 2018). The contextualization of the words is given by the utilization of the hidden states of the forward-backward recurrent neural network. From this forward-backward language model, they concatenate both the output hidden state after the last character in the word using the forward language model and semantic-syntactic information from the end of the sentence to this character with the backward language model. Both output hidden vector states are concatenated to form the final embedding and capture the semantic-syntactic information of the word itself as well as its surrounding context.

Another language model used in the experiments is **FastText**[4] obtained using CBOW ((Mikolov et al., 2013)) with position-weights, in dimension 300, character n-grams of length 5, a window of size 5 and 10 negatives as described in Learning Word Vectors for 157 Languages by (Grave et al., 2018).

The other embeddings used in the experiments are Character and OneHot embeddings obtained from the corpus. The Flair authors describe the use of stacked embedding in (Akbik et al., 2019a).

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

Results for NERC task

Table 2 shows the results of the experiments on the NERC task where the abbreviations in the left column represent the language models used from the following list:

1. **bg** = FastText wiki embeddings
2. **flair-fast** = bg-forward-fast + bg-backward-fast
3. **flair** = bg-forward + bg-backward
4. **char** = Character Embeddings
5. **onehot** = OneHot Embeddings
6. **bert** = bert-base-multilingual-cased

Model	Micro F1
bg + char	96.18
bg + flair-fast	95.75
bg + flair + char	**96.29**
bg + flair + onehot	96.21
bg + bert + char	86.08
bert + flair	83.37

Table 2: Evaluation of stacked embeddings for Bulgarian NERC.

From Table 2 it can be concluded that the best performing stack of embeddings is the concatenation of FastText, bg-forward, bg-backward, and Character embeddings. Table 3 shows the best results for the combinations of embeddings the per class.

Class	P	R	F1
LOC	95.54	96.62	96.08
ORG	95.28	93.74	94.50
MISC	97.14	82.93	89.47
PER	97.68	98.56	98.12

Table 3: Per class results from the best model. Precision (P), Recall (R) and F1

The combination of word embeddings, character embeddings, and the contextual string embeddings outperforms the other combinations, because in this way the words in the text are vectorized with respect to their contextual meaning and they are further represented as a bag of character n-grams. A vector representation is associated to

each character n-gram and thus words are represented as the sum of these representations. These models are fast and lightweight for training of the task. I am going to use them further in the experiments on the other tasks.

BERT (bert-base-multilingual-cased) did not improved the scores in these experiments, being most inaccurate in the classification of the MISC class, scoring particularly with Precision=64.71, Recall=40.24 and F1=49.62. Furthermore, the training of BI-LSTM-CRF with this language model is slow and needs a lot of computational resources.

Originally BERT (particularly bert-base-multilingual-cased) is tested on the XNLI dataset for machine translation on 6 out of 15 languages included in the data. The multilingual model scored 3% worse on English and Chinese than the single-language models for these languages. In my future work I envisage a training of a custom BERT embeddings for the Bulgarian in order to improve it's behaviour on the downstream tasks. Furthermore, the authors claim that the main idea behind BERT and the reason to propose it is to improve the fine-tuning based approaches, thus in the future experiments with Bulgarian NERC the idea should be tested. Fine-tuning is done by first initializing the language model with the pre-trained parameters, and all of the parameters are then fine-tuned using the labeled data from the custom corpus.

3.2 POS Tagging

The method used in the experiments with POS tagging is the same as the method used for NERC task presented above. POS tagging and NERC are both sequence tagging tasks so there is no need to change the proven architecture of the BI-LSTM-CRF tagger on top of the embedding layer. Moreover, the same stacks of language models are employed in the embedding layer and the Flair(forward+backward)+FastText+Character stack performed better than the other stacks again, showing that this combination of embeddings is very powerful for Bulgarian sequence tagging tasks.

Table 1 shows a summary of the previous systems with reported results for Bulgarian POS tagging, extending Table 5 from (Georgiev et al., 2009) with the experiments done after the publishing of the paper. Most of the systems before 2015 are concentrated on experiments for reducing the complexity of the morphosyntactic tagset which for Bulgarian consists of 680 tags.

In my experiments I am using the Universal Dependency version of BulTreeBank produced by (Osenova and Simov, 2015). Thus, I use the 16 tags of Universal POS tagset. The dataset can be downloaded at `https://github.com/UniversalDependencies/UD_Bulgarian-BTB`.

Tag	Num	Acc	Tag	Num	Acc.
NOUN	34,152	99.82	**ADV**	6,558	95.72
ADP	22,097	99.82	**CONJ**	4,860	99
PUNCT	22,058	100	**DET**	2,433	92.86
VERB	17,185	98.38	**NUM**	2,106	94.74
ADJ	13,591	96.57	**PART**	2,052	98.9
PRON	10,094	97.79	**SCONJ**	1,606	99.36
AUX	8,777	93	**INTJ**	143	98.64
PROPN	8,435	96.14	**X**	2	100

Table 4: Frequency of the universal tags in the Treebank. 156,149 tokens total. Column **Tag** contains the Universal POS tag, **Num** represent the number of occurrences, **Acc.** contains the per tag accuracy in %.

Table 4 represents the frequency of the tags within the data and the best accuracy for them achieved by my experiments. It is clear from the results that the main word categories expressing events and relations in text — verbs and nouns — are very well tagged — more than 98 % and 99 % respectively.

Two experiments were performed. In **experiment 1** are used the following embeddings: `bg-forward-fast + bg-backward-fast + Character embeddings + FastText`. The overall accuracy is 98.9 %. In **experiment 2** the embeddings used are: `bg-forward + bg-backward + Character embeddings + FastText` with overall accuracy 99.1 %.

4 Conclusions and Future Work

In these experiments are explored some combinations of the state-of-the-art embeddings on the NERC and POS tagging tasks for Bulgarian. The stack of Flair contextual string embeddings, FastText word embeddings, and Character embeddings outperformed all other combinations reported here.

The results are encouraging and the experiments will continue with training of custom contextual embeddings for Bulgarian and fine-tune

them on the different tasks. The idea of solving several tasks simultaneously in a combined model like (Simova et al., 2014) and (Zhikov et al., 2013) is interesting too. The authors of these articles suggest that several tasks can be solved by one model without much damage to the individual scores, and it is interesting to explore the idea further. Moreover, it combines tasks similar to the classification of NE, events and relations, which is the aim of my PhD.

5 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 DO01-164/28.08.2018.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 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.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 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.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. pages 1638–1649.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. 2009. Feature-rich named entity recognition for bulgarian using conditional random fields. In *Proceedings of the International Conference RANLP-2009*. pages 113–117.

Georgi Georgiev, Valentin Zhikov, Petya Osenova, Kiril Simov, and Preslav Nakov. 2012. Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, EACL '12, pages 492–502. http://dl.acm.org/citation.cfm?id=2380816.2380876.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Petya Osenova and Kiril Simov. 2015. Universalizing BulTreeBank: a linguistic tale about glocalization. In *The 5th Workshop on Balto-Slavic Natural Language Processing*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pages 81–89. https://www.aclweb.org/anthology/W15-5313.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 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*. Association for Computational Linguistics, Florence, Italy, pages 63–74. https://www.aclweb.org/anthology/W19-3709.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *arXiv preprint arXiv:1604.05529* .

Alexander Popov. 2016. Deep learning architecture for part-of-speech tagging with word and suffix embeddings. In Christo Dichev and Gennady Agre, editors, *Artificial Intelligence: Methodology, Systems,*

and Applications. Springer International Publishing, Cham, pages 68–77.

Aleksandar Savkov, Laska Laskova, Petya Osenova, Kiril Simov, and Stanislava Kancheva. 2011. A web-based morphological tagger for bulgarian. *Slovko* pages 126–137.

Lilia Simeonova, Kiril Simov, Petya Osenova, and Preslav Nakov. 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*. Association for Computational Linguistics, page in print.

Kiril Simov and Petya Osenova. 2001. A hybrid system for morphosyntactic disambiguation in bulgarian. In *Proceedings of the EuroConference on Recent Advances in Natural Language Processing*. Citeseer, pages 5–7.

Kiril Simov and Petya Osenova. 2004. BTB-TR04: BulTreeBank morphosyntactic annotation of Bulgarian texts. Technical report, Technical Report BTB-TR04, Bulgarian Academy of Sciences.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a linguistically interpreted corpus of bulgarian: the bultreebank. In *LREC*.

Iliana Simova, Dimitar Vasilev, Alexander Popov, Kiril Simov, and Petya Osenova. 2014. Joint Ensemble Model for POS Tagging and Dependency Parsing. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*. Dublin City University, pages 15–25.

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

Michihiro Yasunaga, Jungo Kasai, and Dragomir R. Radev. 2017. Robust multilingual part-of-speech tagging via adversarial training. *CoRR* abs/1711.04903. http://arxiv.org/abs/1711.04903.

Xiang Yu, Agnieszka Falenska, and Ngoc Thang Vu. 2017. A general-purpose tagger with convolutional neural networks. *CoRR* abs/1706.01723. http://arxiv.org/abs/1706.01723.

Valentin Zhikov, Georgi Georgiev, Kiril Simov, and Petya Osenova. 2013. Combining POS Tagging, Dependency Parsing and Coreferential Resolution for Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. pages 755–762.