TLR at BSNLP2019: A Multilingual Named Entity Recognition System

Jose G. Moreno **Elvys Linhares Pontes**[1,2] **Mickaël Coustaty**[1] **Antoine Doucet**[1]
University of Toulouse
IRIT, UMR 5505 CNRS
jose.moreno@irit.fr
1 - L3i laboratory, University of La Rochelle, La Rochelle, France
{firstname.lastname}@univ-lr.fr
2 - University of Avignon, Avignon, France

Abstract

This paper presents our participation at the shared task on multilingual named entity recognition at BSNLP2019. Our strategy is based on a standard neural architecture for sequence labeling. In particular, we use a mixed model which combines multilingual-contextual and language-specific embeddings. Our only submitted run is based on a voting schema using multiple models, one for each of the four languages of the task (Bulgarian, Czech, Polish, and Russian) and another for English. Results for named entity recognition are encouraging for all languages, varying from 60% to 83% in terms of Strict and Relaxed metrics, respectively.

1 Introduction

Correctly detecting mentions of entities in text documents in multiple languages is a challenging task (Ji et al., 2014, 2015; Ji and Nothman, 2016; Ji et al., 2017). This is especially true when documents relate to news because of the huge range of topics covered by newspapers. In this context, the shared task on multilingual named entity recognition (NER) proposes to participants to test their system under a multilingual setup. Four languages are addressed in BSNLP2019: Bulgarian (bg), Czech (cz), Polish (pl), and Russian (ru). Similarly to the first edition of this task in 2017 (Piskorski et al., 2017), participants are required to recognize, normalize, and link entities from raw texts written in multiple languages. Our participation is focused on the sole recognition of entities while other steps will be covered in our future work.

In order to build a unique NER system for multiple languages, we decided to contribute a solution based on an end-to-end system without (or almost without) language specific pre-processing. We explored an existing neural architecture, the LSTM-CNNs-CRF (Ma and Hovy, 2016), initially proposed for NER in English. This neural model is based on word embeddings to represent each token in a sentence. In order to have a unique embedding space, we propose to use a transformer-based (Vaswani et al., 2017) contextual embedding called BERT (Devlin et al., 2019). This pre-trained model includes multilingual representations that are context-aware. However, as noted by Reimers and Gurevych (2019), contextual embeddings provide multiple layers that are challenging to combine together. To overcome this problem, we used the weighted average strategy they successfully tested using (Peters et al., 2018).

The results of our participation are quite encouraging. Regarding the *Relaxed Partial* metric, our run achieves 80.26% in average for the four languages and the two topics that compose the test collection. In order to present comparative results against the state of the art, we run experiments using two extra datasets under the standard CoNLL evaluation setup. The remainder of this paper is organized as follows: Section 2 introduces the related work while Section 3 presents the proposed multi-lingual model. Section 4 presents the results while conclusions are drawn in Section 5.

2 Related Work

Named entity recognition has been largely studied through the organization of shared tasks in the last two decades (Nadeau and Sekine, 2007; Yadav and Bethard, 2018). The large variety of models can be grouped into three types: rule-based (Chiticariu et al., 2010), gazetteers-based (Sundheim, 1995), and statistically-based models (Florian et al., 2003). The latter type is a current hot topic in research, in particular with the return of neural based models[1]. Two main contributions

[1] In all their flavors, including attention.

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 83–88,
Florence, Italy, 2 August 2019.

have recently redrawn the landscape of models for sequence labelling such as NER: the proposal of new architectures (Ma and Hovy, 2016; Lample et al., 2016), the use of contextualized embeddings (Peters et al., 2018; Reimers and Gurevych, 2019), or even, the use of both of them (Devlin et al., 2019). The use of contextualized embeddings is a clear advantage for several kinds of neural-based NER systems, however as pointed out by Reimers and Gurevych (2019) the combination of multiples vectors proposed by these models is computationally expensive.

3 TLR System: A Neural-based Multilingual NER Tagger

This section describes our model which is based on a standard end-to-end architecture for sequence labeling, namely LSTM-CNNs-CRF (Ma and Hovy, 2016). We have combined this architecture with contextual embeddings using a weighted average strategy (Reimers and Gurevych, 2019) applied to a pre-trained model for multiple languages (Devlin et al., 2019) (including all languages of the task). We trained a NER model for each of the four languages and predict labels based on a classical voting strategy. As an example, the overall architecture of our model for Polish using the sentence *"Wielka Brytania z zadowoleniem przyjęła porozumienie z Unia Europejska"* (or "United Kingdom welcomes agreement with the European Union" in English) is depicted in Figure 1.

3.1 FastText Embedding

In this layer, we used pre-trained embeddings for each language trained on Common Crawl and Wikipedia using fastText (Bojanowski et al., 2017; Grave et al., 2018). These models were trained using the continuous bag-of-words (CBOW) strategy with position weights. A total of 300 dimensions were used with character n-grams of length 5, a window of size 5 and 10 negatives. The four languages of the task are included in this publicly available[2] pre-trained embedding (Grave et al., 2018). We have used the fastText library to ensure that every token (also in other alphabets) has a corresponding vector avoiding out of vocabulary tokens.

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

3.2 Case Encoding

This layer allows to encode each token based on the case information as proposed by (Reimers and Gurevych, 2017). We have used a one-hot encoding of the following seven classes: {'other', 'numeric', 'mainly_numeric', 'allLower', 'allUpper', 'initialUpper', 'contains_digit'}.

3.3 Multilingual BERT

We used the multilingual pre-trained embedding of BERT[3]. In particular, we used the model learned for 104 languages including the four of this task. This model is composed of 12 layers and 768 dimensions in each layer for a total of 110M parameters. Directly using the 12 layers can be hard to compute in a desktop computer. To cope with this problem, we used the weighted strategy proposed by Reimers and Gurevych (2019) and combined only the first two layers. When a token was composed of multiple BERT tokens, we averaged them to obtain a unique vector per token.

3.4 Char Representation

We used the char representation strategy proposed by Ma and Hovy (2016) where char embeddings are combined using a convolutional neural network (CNN). Thus, an embedding vector is learned for each character by iterating trough the entire collection. Note that the four languages include unique characters which make harder the sharing of patterns between languages. To deal with this problem, we transliterated each token to the Latin alphabet using the unidecode library[4] as a preprocessing step. This conversion is only applied at this layer and is not used elsewhere.

3.5 Language-Dependent and Independent Features

In Figure 1, we observe that the "char representation", "multilingual BERT", and "case encoding" layers are language-independent features[5] So, all the processing steps are applied without considering the language, including the transliteration to the Latin alphabet. It means that some tokens are translated even knowing that they are already in a

[3]https://github.com/google-research/bert/blob/master/multilingual.md

[4]https://pypi.org/project/Unidecode/

[5]We mean that as the four languages follow exactly the same process, those steps become completely independent in this specific context.

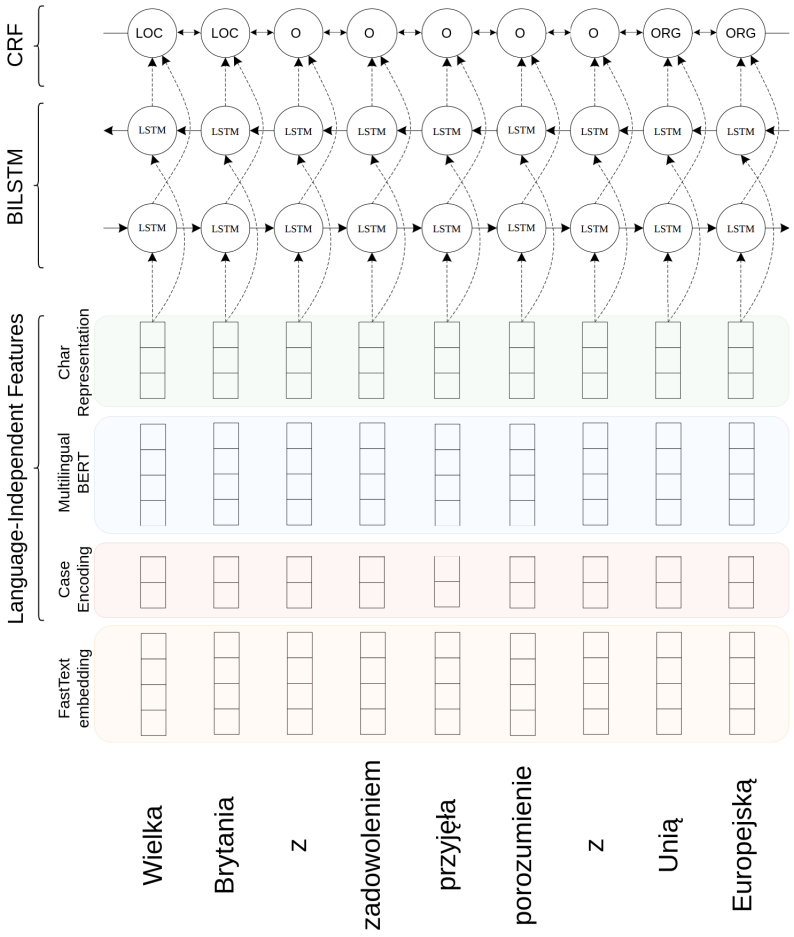

Figure 1: Architecture of a single-language model of our system. Note that for each token we provide a unique NER prediction.

Latin alphabet. On the other hand, "fastText embedding" is clearly a language-dependent feature. However, we intentionally reduce the language dependency by using the architecture in Figure 1 as many times as the number of languages involved in the task, e.g. four times. Each time we switched the "fastText embedding" model for the one corresponding to each language, this make a total of four different NER models. Our final prediction is obtained by applying a simple majority voting schema between these four NER models.

4 Experiments

4.1 Experimental Setup

We follow the configuration setup proposed by the task organizers. Two topics, "nord_stream" and "ryanair", were used to test our models. These topics include 1100 documents in the four languages. Further details can be found in the 2019 shared task overview paper (Piskorski et al., 2019). For training, we have used the documents provided for the task but also the ones in Czech, Polish, and Russian from the previous round of same task in 2017 (Piskorski et al., 2017). We additionally added the training example form the CoNLL2003 (Sang and De Meulder, 1837) collection in English (13879 train, 3235 dev, and 3422 test sentences). Used metrics include the officially proposed metrics and standard metrics for the CoNLL2003 dataset (F1 metric).

4.2 Official Results

The official results of our unique run are presented in Table 1 and identified as TLR-1. Note that only NER metrics are presented for the four languages. We have added the results for each language model using the partial annotations provided by the organizers[6]. Each result is identified with the language used for the "fastText embed-

[6]We were able to calculate "Recognition Strict" for these unofficial results.

NORD_STREAM		*Language*							
Phase	*Metric*	bg		cz		pl		ru	
Recognition	Relaxed Partial	TLR-1	83.384	TLR-1	82.124	TLR-1	80.665	TLR-1	73.145
	Relaxed Exact	TLR-1	76.114	TLR-1	74.106	TLR-1	71.423	TLR-1	62.168
	Strict	TLR-1	73.312	TLR-1	74.475	TLR-1	72.026	TLR-1	59.627
		bg	**72.873**	bg	67.841	bg	68.281	bg	54.922
		cz	68.821	cz	**78.225**	cz	71.509	cz	52.590
		pl	69.892	pl	73.636	pl	**75.820**	pl	53.939
		ru	72.661	ru	71.522	ru	70.356	ru	**58.399**
RYANAIR		*Language*							
Phase	*Metric*	bg		cz		pl		ru	
Recognition	Relaxed Partial	TLR-1	75.861	TLR-1	82.865	TLR-1	82.182	TLR-1	83.419
	Relaxed Exact	TLR-1	69.824	TLR-1	73.493	TLR-1	77.463	TLR-1	78.303
	Strict	TLR-1	68.377	TLR-1	72.509	TLR-1	75.118	TLR-1	78.028
		bg	**76.152**	bg	77.533	bg	79.168	bg	78.518
		cz	61.755	cz	**78.549**	cz	76.863	cz	75.280
		pl	67.876	pl	77.907	pl	**82.242**	pl	76.864
		ru	70.288	ru	74.805	ru	76.135	ru	**79.784**

Table 1: Evaluation results of our TLR submission. We have added extra results for the strict metric using each single model based on one of the four languages.

ding" layer in Figure 1. Based on strict recognition, most of the cases[7], the use of the correct language embedding improves the recognition of the respective language. However, the voting schema outperforms the individual models on average. This suggest that a system aware of the language of the input sentence could provide better results that our voting schema.

4.3 Unofficial Results

In order to compare our system to the state-of-the-art, we have evaluated our architecture using the CoNLL2003 dataset. Our results using two and six layers are presented in Table 2. Note that English is not part of our target languages. So, an underperformance of 2.5 is acceptable in our system[8]. It is also worth nothing that the use of more BERT layers increases our results. However, the amount of memory used is also increased manifold. We set the number of layers (hyperparameter) to two layers due to our computation constraints despite the downgrading in performances for English.

The number of epochs (hyperparameter) was set using the BSNLP2017 dataset (for ru, cs, and

Method		Metric		
	Set	P	R	F1
BRNN-CNN-CRF	Dev	94.8	94.6	94.7
(Ma and Hovy, 2016)	Test	91.3	91.0	91.2
BiLSTM + ElMo	Dev	95.1	95.7	95.4
(Reimers and Gurevych, 2019)	Test	90.9	92.1	91.5
BiLSTM + MultiBERT2L	Dev	92.3	93.0	92.7
(ours)	Test	88.2	89.7	89.0
BiLSTM + MultiBERT6L	Dev	93.2	93.8	93.5
(ours)	Test	89.3	90.3	89.8

Table 2: Evaluation results on the CoNLL 2003 dataset, an English only dataset.

[7] 6 out of 8, with differences smaller than 0.4 points.

[8] More experiments using BERT English-only model will be performed in our future work.

Language	BSNLP2017+CoNLL2003			
	P	R	F1	Epochs
en	78.9	82.8	80.8	10
bg	77.1	79.3	78.2	6
cz	78.7	82.2	80.4	24
pl	79.7	83.6	81.6	16
ru	79.1	83.4	81.2	21

Table 3: Evaluation results on the BSNLP2017 and CoNLL 2003 datasets, a multilingual dataset. Each row represents a model learned with a fastText language specific embedding.

pl) combined with CoNLL2003 as a validation set of our final models. Results for these combined datasets are presented in Table 3. Surprisingly, our results seem very similar independently of the fastText embedding. It suggests that our architecture is able to generalize the prediction for several target languages. Note that the worst results are obtained by the Bulgarian model, but no test examples were included for this language. In contrast, we believe that the examples provided in other languages were rich enough to help the predictions (also in English).

5 Conclusion

This paper presents the TLR participation at the shared task on multilingual named entity recognition at BSNLP2019. Our system is a combination of multiple representation including character information, multilingual embedding, and language specific embedding. However, we combine them in such a way that it can be seen as a generic multilingual NER system for a large number of languages (104 in total). Although top participants outperform our average score of 80.26% of "Relaxed Partial" (Piskorski et al., 2019), the strengths of the proposed strategy relays on the fact that it can be easily adapted to new languages and topics without extra effort.

Acknowledgements

This paper is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

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.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain Adaptation of Rule-based Annotators for Named-Entity Recognition Tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1002–1012. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named Entity Recognition through Classifier Combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

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)*.

Heng Ji and Joel Nothman. 2016. Overview of TAC-KBP2016 Tri-lingual EDL and its impact on end-to-end Cold-Start KBP. In *Proc. Text Analysis Conference (TAC2016)*.

Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. 2015. Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In *Proc. Text Analysis Conference (TAC2015)*.

Heng Ji, Joel Nothman, Ben Hachey, et al. 2014. Overview of TAC-KBP2014 Entity Discovery and Linking Tasks. In *Proc. Text Analysis Conference (TAC2014)*, pages 1333–1339.

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, Cash Costello, and Sydney Informatics Hub. 2017. Overview of TAC-KBP2017 13 Languages Entity Discovery and Linking. In *Proc. Text Analysis Conference (TAC2017)*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 260–270.

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), pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The Second Cross-Lingual Challenge on Recognition, Classification, Lemmatization, and Linking of Named Entities across Slavic Languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Florence, Italy. Association for Computational Linguistics.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, Roman Yangarber, et al. 2017. The First Cross-Lingual Challenge on Recognition, Normalization and Matching of Named Entities in Slavic Languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Nils Reimers and Iryna Gurevych. 2019. Alternative Weighting Schemes for ELMo Embeddings. *arXiv preprint arXiv:1904.02954*.

Erik F Tjong Kim Sang and Fien De Meulder. 1837. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Development*, volume 922, page 1341.

Beth M. Sundheim. 1995. Overview of Results of the MUC-6 Evaluation. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 13–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158.