Multidomain Contextual Embeddings for Named Entity Recognition

Joaquim Santos[1 [0000−0002−0581−4092]], Juliano Terra[1 [0000−0003−3066−1531]], Bernardo Consoli[1 [0000−0003−0656−511X]], and Renata Vieira[1 [0000−0003−2449−5477]]

Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil
{joaquim.santos,bernardo.consoli,juliano.terra}@acad.pucrs.br,
renata.vieira@pucrs.br

Abstract. Neural Networks are widely used for Named Entity Recognition due to their capability of extracting features from texts automatically and integrating them with sequence taggers. Pretrained Language Models are also extensively used for NER, as their product, Word Embeddings, are key elements for improving the performance of NER systems. A novel type of embeddings, called Contextual Word Embeddings, can adapt according to the context it is inserted, something traditional word embeddings could not do. These contextual embeddings have proven to be superior to traditional embeddings for NER. In this work, we show the results of our network, which uses Neural Networks in conjunction with a contextual language model, on corpora composed of texts belonging to rarely tested textual genres, such as official police documents and clinical notes, as proposed by a task in IberLEF 2019.

Keywords: Neural Networks · Named Entity Recognition · Word Embeddings. · Flair Embeddings

1 Introduction

Named Entity Recognition (NER), a task in the field of Natural Language Processing (NLP), consists of finding proper nouns in a given text and to classify them on different predefined categories [14]. Modern approaches for the NER task utilize Neural Networks (NN) that automatically learn the features from raw text and making the use of manually constructed rules obsolete [4].

The use of vector representation of words (word embeddings) have helped to increase the quality of NER systems. These embeddings can be created through the training of a Language Model (LM) on a corpus of raw texts [4] - [10]. Modern LMs create *contextualized embeddings* in runtime, as opposed to retrieving information from static word-vector dictionaries as do traditional LMs. One of

these modern systems is *Flair Embeddings*, that achieved state-of-the-art results for German and English.

Recent State-of-the-Art results for sequential labeling tasks has used Deep Learning strategies. The Long-Short Term Memory (LSTM) networks has shown cutting-edge results on this type of task due to the fact that they allow to analyze the context in which a word is inserted in two directions of a given sentence: *forward* and *backward* [1], [3], [10].

On this work we present our system for Portuguese NER proposed by the Shared Task "Portuguese Named Entity Recognition and Relation Extraction Tasks (NerRelIberLEF2019) " on IberLEF 2019 [5]. Iberian Languages Evaluation Forum (IberLEF) is a forum that has that aims to develop the Natural Language Processing for Iberian American languages. On this year of 2019 the shared task NerRelIberlef presented three challenges (tasks) for researchers on the NLP area: a task about NER and other two about Relation Extraction. In what follows, this work describes our approach and resources used on the participation for the NER task, to which we will refer to *Task 1*. Our system makes use of a BiLSTM Neural Network fed by a composition of other two language models: Flair Embeddings and Word Embeddings. A final layer of a CRF classifier returns the token classification.

2 Related Work

Bidirectional Encoder Representations from Transformers (BERT) [6] achieved the State-of-the-Art using a complex network based around Transformer Neurons [23]. Recent works for NER in the English language showed excellent results with the use of *contextualized word embeddings*. BERT was recently surpassed by Flair Embeddings [1], which utilizes contextualized embeddings from a character-level Language Model.

Embeddings from Language Models (ELMo) is an approach that creates embeddings based on a deep BiLSTM network, this enables ELMo to analyze the context into which any given word is inserted. ELMo is also character based, enabling the model to create vector representations words it was not trained on [16].

In the context of Named Entity Recognition for Portuguese, there is a proposal for the use of a Deep Neural Network (DNN) called CharWNN [20]. CharWNN is a specific type of DNN utilized for sequence tagging, extracting features from word and character level.

Another approach also using Neural Networks was presented by [3]. It uses LSTM networks that has a Conditional Random Fields (CRF) classifier as its last layer, based on the network proposed by Lample et al. (2016) [10]. Table 1 shows the F1-measure of the cited works in this section.

Table 1. F1 measure of the presented related works

System	Training Corpus	Test Corpus	Language	F1
BiLSTM-CRF-Flair[1]	CoNLL-03 Train	CoNLL-03 Test	English	93.09%
$BERT_{LARGE}$[6]	CoNLL-03 Train	CoNLL-03 Test	English	92.80%
BiLSTM-CRF-ELMo[16]	CoNLL-03 Train	CoNLL-03 Test	English	92.22%
BiLSTM-CRF[10]	CoNLL-03 Train	CoNLL-03 Test	English	90.94%
BiLSTM-CRF[3]	I HAREM	MiniHAREM	Portuguese	76.27%
CharWNN[21]	I HAREM	MiniHAREM	Portuguese	71.23%
CRF+LG[18]	I HAREM	MniHAREM	Portuguese	60.36%

3 Background

In this section, we will present our approach's main concepts. These are LSTM recurrent neural networks, the CRF classifier and language models based on *Word Embeddings* and Flair Embeddings.

3.1 LSTM Networks

Recurrent Neural Networks (RNN) are currently considered to be the standard networks for NLP [15]. Of these networks, Long-Short Term Memory (LSTM) networks stand out in abundance of use. LSTM networks are a variation of the RNN with a refined architecture that shows better results for sequential tasks [22]. The LSTM architecture uses functions that determine whether previously added information should be kept, modified or discarded in order to better relate to newly added information. An important variation of LSTM networks are BiLSTM networks. These are composed of two LSTM working in parallel. One of the networks deals with a "forward" data sequence (*Forward* LSTM) and the other with a "backwards" data sequence (*Backwards* LSTM). This makes it so the network has a greater learning capacity [8].

3.2 CRF Classifier

Conditional Random Fields (CRF) is a classifier used for the construction of probabilistic models with the goal of segmenting and labeling sequential data. This type of classifier has been widely used on tasks of *Part-of-Speech Tagging* and NER [9]. Recent works about NER in the Portuguese language adopt CRF as a final component of the system in order to give the token it's final classification [3] [18] [2].

3.3 Word Embeddings

Neural Embeddings or *Word Embeddings* are ways to represent words in n-dimensional vector spaces. Recently, this type of representation has been widely

used in the field of NLP [11]. Among these, *Word2Vec* [13] stands out. *Word2Vec* is freely available and is based on RNN, being able to learn the representation of words in high-dimensional vector spaces [24].

Word2Vec is divided into two architectures, Continuous Bag of Words (CBOW) and Skip-Gram. CBOW uses a context as an input for the network and then the desired word gets returned, on the other hand, using Skip-Gram architecture, we use a word as an input and then the context in which the word is presented is returned [12].

For our system, we used the 300-dimensional *Word2Vec Skip-Gram (W2V-SKPG)* language model made available by the Interinstitutional Center for Computational Linguistics of São Paulo University (NILC). The embeddings are available in their website (http://nilc.icmc.usp.br/embeddings).

3.4 Flair Embeddings

Flair Embeddings is a recent language modelling architecture that works on both the character and word levels. It goes beyond traditional word embedding architectures like Skip-gram and CBOW, as it takes into account the character-level morphological features of words as well as the more traditional contextual information. Because of this, the authors consider Flair's embeddings to be *Contextual String Embeddings* [1].

The Flair Model used in our system, called FlairBBP, was trained with a corpus of over 4 billion tokens and is available for use in our GitHub page (https://github.com/jneto04/ner-pt).

4 Neural Network for NER

Our NER model is the product of one of our previous works, where we trained a Neural Network for this task. We used a BiLSTM-CRF Neural Network that has been previously used for NER in English and German [1]. The BiLSTM-CRF network was trained using a structure of embeddings concatenation called *Stacking Embeddings*. That is, each one of our tokens were represented by a compilation of two types of embeddings: Flair Embeddings (FlairBBP) and Word Embeddings (W2V-SKPG). The matrix $\mathbf{w}$ of the equation 4 shows our stacking of embeddings. The Neural Network used for this work is composed of two layers, a Character Language Model and a Sequence Labeling Model. First, all of the tokens are passed to the Character Language Model of the *Stacking Embeddings*, which then returns a vector r for each input token.

$$\mathbf{w} = \begin{bmatrix} w^{FlairBBP} \\ w^{W2V-SKPG} \end{bmatrix}$$

This vector r is then passed to the "Sequence Labeling Model" where a BiLSTM network receives the vectors r and pass it output to the CRF classifier that returns the token classification.

The training of the network was done with a corpus using the First HAREM (https://www.linguateca.pt/HAREM/), considering the categories Person, Place, Organization, Time and Value.

Table 2 shows the hyperparameters used on training.

Table 2. Our System's Hyperparameters

Hyperparameter	Value
Learning rate	$0.1 \sim 0.002$
Hidden Layers	256
Optimizer	GDW
Mini batch size	32
Epochs	150

5 Results

The system was evaluated using three different test corpora: a Police Dataset, a Clinical Dataset, and a General Dataset (Created from SIGARRA [17] and the second HAREM [7]). Table 3 presents the results provided by Task 1's coordination team using the CoNLL-2002 script [19].

Table 3. Task 1 System Evaluation Results

Corpus	Category	Prec	Rec	F1
Police Dataset	PER	94.21%	82.82%	88.15%
Clinical Dataset	PER	22.08%	41.46%	28.81%
General Dataset	Overall	75.28%	59.82%	66.66%
	ORG	65.13%	35.32%	45.80%
	PER	65.96%	54.33%	59.58%
	PLC	55.81%	61.40%	58.47%
	TME	94.43%	87.44%	90.80%
	VAL	88.68%	87.04%	87.85%

Out of all of the submitted systems that participated in the shared Task 1[5], our system achieved the best F1-measure for the General Dataset (Overall). We attribute that to the *Stacking Embeddings* and to the test corpus that was used for this part of the task. The General Dataset is composed of two corpora: SIGARRA and Second HAREM (Relation Version) that are relatively close structurally and linguistically to the HAREM collection with which our system was trained.

Our results for the Police Dataset were competitive, having achieved a 2.8% lower F1-measure score than the best system for this dataset. Our good performance with this test dataset is also attributed to the embeddings and to the fact that the texts used to build the Police Dataset are very well structured, like those used to build the HAREM collection [5].

In the case of the Clinical Dataset, the difference between the F1-measure of our system and the system with highest F1-measure of this dataset is 12.98%. We believe that this is due to the fact that the Clinical Dataset's unusual structure and language. It is composed of clinical notes containing abbreviations, medical terms and various other particularities found in texts from hospital environments. Texts like this differ greatly from the traditional style for the NER task in Portuguese, and these differences were not taken into account during the system's training.

6 Conclusion

This paper presented our proposed approach for "Task 1: Named Entity Recognition" in the NerRelIberLEF Shared Task in IberLEF 2019. Our approach involved the use of a BiLSTM-CRF that receives a compilation of highly representational embeddings: FlairBBP + W2V-SKPG. As such, we understand that our results come from the representational power of the Flair Embeddings architecture in representing a natural language.

As a future work, we plan to train our system with more corpora, as we believe this will yield even better metrics.

Acknowledgments

We thank CNPQ for their financial support.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
2. do Amaral, D.O.F., Vieira, R.: Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. Linguamática **6**(1), 41–49 (2014)
3. de Castro, P.V.Q., da Silva, N.F.F., da Silva Soares, A.: Portuguese named entity recognition using lstm-crf. In: International Conference on Computational Processing of the Portuguese Language. pp. 83–92. Springer (2018)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(Aug), 2493–2537 (2011)
5. Collovini, S., Santos, J., Consoli, B., Terra, J., Vieira, R., Quaresma, P., Souza, M., Claro, D.B., Glauber, R., Xavier, C.C.: Portuguese named entity recognition and relation extraction tasks at iberlef 2019 (2019)

6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., Santos, D.: Second harem: advancing the state of the art of named entity recognition in portuguese. In: quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association. European Language Resources Association (2010)
8. Graves, A., Jaitly, N., Mohamed, A.r.: Hybrid speech recognition with deep bidirectional lstm. In: 2013 IEEE workshop on automatic speech recognition and understanding. pp. 273–278. IEEE (2013)
9. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
11. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 302–308 (2014)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
14. Milidiú, R.L., Duarte, J.C., Cavalcante, R.: Machine learning algorithms for portuguese named entity recognition. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial **11**(36) (2007)
15. Murdoch, W.J., Szlam, A.: Automatic rule extraction from long short term memory networks. arXiv preprint arXiv:1702.02540 (2017)
16. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
17. Pires, A.R.O.: Named entity extraction from Portuguese web text. Master's thesis, Faculdade de Engenharia da Universidade de Porto, Porto, Portugal (2017)
18. Pirovani, J.P., de Oliveira, E.: Portuguese named entity recognition using conditional random fields and local grammars. In: LREC (2018)
19. Sang, T.K., Erik, F.: Introduction to the conll-2002 shared task: language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158 (2002)
20. Santos, C.D., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). pp. 1818–1826 (2014)
21. Santos, C.N.d., Guimaraes, V.: Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv:1505.05008 (2015)
22. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075 (2015)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

24. Zhang, D., Xu, H., Su, Z., Xu, Y.: Chinese comments sentiment classification based on word2vec and svmperf. Expert Systems with Applications **42**(4), 1857–1863 (2015)