VSP at PharmaCoNER 2019: Recognition of Pharmacological Substances, Compounds and Proteins with Recurrent Neural Networks in Spanish Clinical Cases

Víctor Suárez-Paniagua
Computer Science Department,
Carlos III University of Madrid.
Leganés 28911, Madrid, Spain.
vspaniag@inf.uc3m.es

Abstract

This paper presents the participation of the VSP team for the PharmaCoNER Tracks from the BioNLP Open Shared Task 2019. The system consists of a neural model for the Named Entity Recognition of drugs, medications and chemical entities in Spanish and the use of the Spanish Edition of SNOMED CT term search engine for the concept normalization of the recognized mentions. The neural network is implemented with two bidirectional Recurrent Neural Networks with LSTM cells that creates a feature vector for each word of the sentences in order to classify the entities. The first layer uses the characters of each word and the resulting vector is aggregated to the second layer together with its word embedding in order to create the feature vector of the word. In addition, a Conditional Random Field layer classifies the vector representation of each word in one of the mention types. The system obtains a performance of 76.29%, and 60.34% in F1 for the classification of the Named Entity Recognition task and the Concept indexing task, respectively. This method presents good results with a basic approach without using pretrained word embeddings or any hand-crafted features.

1 Introduction

Nowadays, the task of finding the essential data about the patients in medical records is very difficult because of the highly increasing amount of unstructured documents generated by the doctors. Thus, the automatic extraction of the mentions related with drugs, medications and chemical entities in the clinical case studies can reduces the time of healthcare professionals expend reviewing these medical documents in order to retrieve the most relevant information.

Previously, some Natural Language Processing (NLP) shared tasks were organized in order to promote the develop of automatic systems given the importance of this task. The i2b2 shared task was the first NLP challenge for identifying Protected Health Information in the clinical narratives (Özlem Uzuner et al., 2007). The CHEMDNER task was focused on the Named Entity Recognition (NER) of chemical compounds and drug names in PubMed abstracts and chemistry journals (Krallinger et al., 2015).

The goal of the BioNLP Open Shared Task 2019 is to create NLP challenges for developing systems in order to extract information from biomedical corpora. Concretely, the PharmaCoNER Task is focusing on the recognition of pharmacological substance, compound and protein mentions from Spanish medical texts.

Currently, deep learning approaches overcome traditional machine learning systems on the majority of NLP tasks, such as text classification (Kim, 2014), language modeling (Mikolov et al., 2013) and machine translation (Cho et al., 2014). Moreover, these models have the advantage of automatically learn the most relevant features without defining rules by hand. Concretely, the LSTM-CRF Model proposed by (Lample et al., 2016) improves the performance of a CRF with hand-crafted features for different biomedical NER tasks (Habibi et al., 2017). The main idea of this system is to create a word vector representation using a bidirectional Recurrent Neural Network with LSTM cells (BiLSTM) with character information encoded in another BiLSTM layer in order to classify the tag of each word in the sentences with a CRF classifier. Following this approach, the system proposed in (Dernoncourt et al., 2016) uses a BiLSTM-

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 16–20
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

CRF Model with character and word levels for the de-identification of patient notes using the i2b2 dataset that overcomes the previous systems in this task.

This paper presents the participation of the author, as VSP team, at the tasks proposed by PharmaCoNER about the classification of pharmacological substances, compounds and proteins and the Concept Indexing of the recognized mentions from clinical cases in Spanish. The proposed system follows the same approaches of (Lample et al., 2016) and (Dernoncourt et al., 2016) for the NER task with some modifications for the Spanish language implemented with NeuroNER tool (Dernoncourt et al., 2017) because the architecture obtains good performance for the recognition of biomedical entities. In addition, a simple SNOMED CT term search engine is implemented for the concept normalization.

2 Dataset

The corpus of the PharmaCoNER task contains 1,000 clinical cases derived from the Spanish Clinical Case Corpus (SPACCC)[1] with manually annotated mentions such as pharmacological substances, compounds and proteins by clinical documentalists. The documents are randomly divided into the training, validation and test sets for creating, developing and ranking the different systems, respectively.

The corpus contains four different entity types:

- *NORMALIZABLES*: they are chemicals that can be normalized to a unique concept identifier.
- *NO_NORMALIZABLES*: they are chemicals that cannot be normalized. These mentions were used for training the system, but they were not taken into consideration for the results in the task of NER or Concept Indexing.
- *PROTEINAS*: this entity type refers to mentions of proteins and genes following the annotation schema of BioCreative GPRO (Pérez-Pérez et al., 2017).
- *UNCLEAR*: these mentions are cases of general substances, such as pharmaceutical formulations, general treatments, chemotherapy programs, vaccines and a predefined set of general substances.

Additionally, all mentions without the *NO_NORMALIZABLES* tag are annotated with its corresponding SNOMED CT normalization concept.

3 Method

This section presents the Neural architecture for the classification of the entity types and the concept normalization method in Spanish clinical cases. Figure 1 presents the process of the NER task using two BiLSTMs for the character and token levels in order to create each word representation until its classification by a CRF.

3.1 Data preprocessing

The first step is a preprocessing of the sentences in the corpus, which prepares the inputs for the neural model. Firstly, the clinical cases are separated into sentences using a sentence splitter and the words of these sentences are extracted by a tokenizer, both were adapted for the Spanish language. For the experiments, the previous processes were performed by the spaCy tool in Python (Explosion AI, 2017). Once the sentences were divided into word, the BIOES tag schema encodes each token with an entity type (B tag is the beginning token, I tag is the inside token, E tag is the ending token, S tag is the single token and O tag is the outside token). In many previous NER tasks, using this codification is better than the BIO tag scheme (Ratinov and Roth, 2009), but the number of labels increases because there are two additional tags for each class. Thus, the number of possible classes are the 4 tags times the 4 entity types and the O tag for the PharmaCoNER corpus.

3.2 BiLSTM layers

RNNs are very effective in feature learning when the inputs are sequences. Concretely, the Long Short-Term Memory cell (LSTM) (Hochreiter and Schmidhuber, 1997) defines four gates for creating the representation of

[1] https://doi.org/10.5281/zenodo.2560316

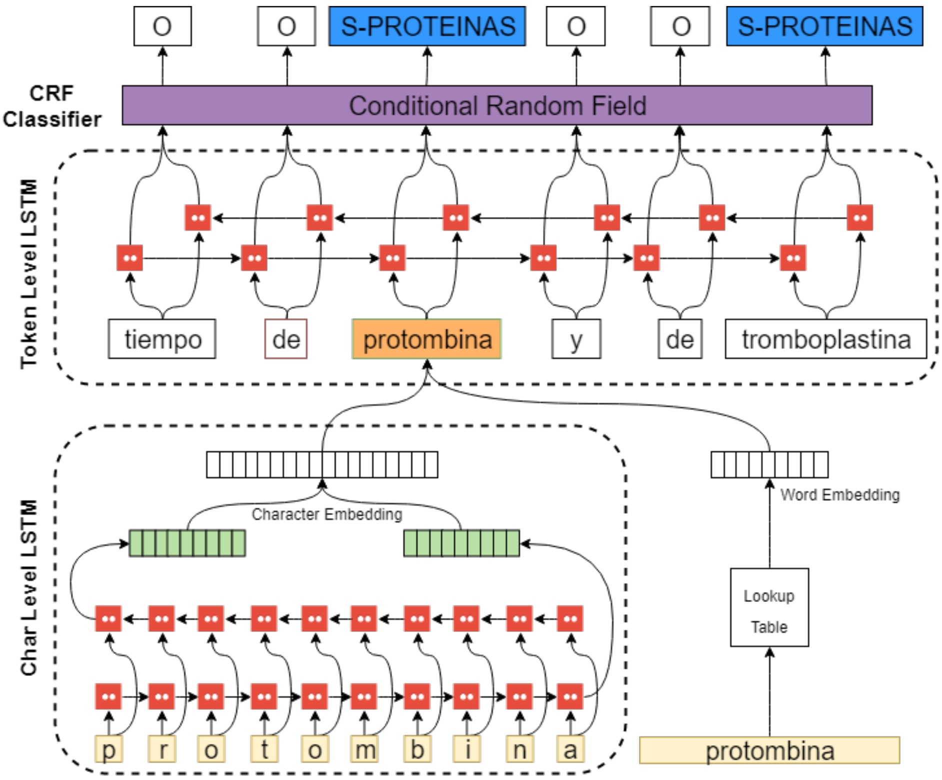

Figure 1: Neural model for the recognition of mentions in Spanish clinical cases using the PharmaCoNER task 2019 corpus.

each input taking the information of the current and previous cells. Thus, each output is a combination of the current and the previous cell states. Furthermore, another LSTM can be applied in the other direction from the end of the sequence to the start in order to extract the relevant features of each input in both directions.

3.2.1 Character level

The first layer takes each word of the sentences individually. These tokens are decomposed into characters that are the input of the BiLSTM. Once all the inputs are computed by the network, the last output vectors of both directions are concatenated in order to create the vector representation of the word according to its characters.

3.2.2 Token level

The second layer takes the embedding of each word in the sentence and concatenates them with the outputs of the first BiLSTM with the character representation. In addition, a Dropout layer is applied to the word representation in order to prevent overfitting in the training phase. In this case, the outputs of each direction in one token are concatenated for the classification layer.

3.3 Contional Random Field Classifier

CRF (Lafferty et al., 2001) is the sequential version of the Softmax that aggregates the label predicted in the previous output as part of the input. In NER tasks, CRF shows better results than Softmax because it adds a higher probability to the correct labelled sequence. For instance, the I tag cannot be before a B tag or after a E tag by definition. For the proposed system, the CRF classifies the output vector of the BiLSTM layer with the token information in one of the classes.

3.4 Concept Indexing

After the NER task, the concept indexing is applied to all recognized entities in the sentences for the term normalization. To this end, the Spanish Edition of the SNOMED CT International Browser[2] searches each mention and gives its normalization term. Moreover, The Spanish Medical Abbreviation DataBase

[2] https://prod-browser-exten.ihtsdotools.org/

(AbreMES-DB)[3] is used in order to disambiguate the acronyms and the resulting term is searched in the SNOMED CT International Browser. In the cases where there are more than one normalization concept for a term, a very naive approach is followed where the first node in the term list is chosen as the final output.

4 Results and Discussion

The architecture was trained over the training set during 100 epochs with shuffled mini-batches and choosing the best performance over the validation set via stopping criteria. The values of the two BiLSTM and CRF parameters for generating the prediction of the test set are presented in Table 1. Additionally, a gradient clipping keeps the weight of the network in a low range preventing the exploding gradient problem. The embeddings of the characters and words are randomly initialized and learned during the training of the network. The main goal of this work is to test the performance of the proposed neural model on this dataset without using pretrained word embeddings or any hand-crafted features. In future work, the impact of different pretrained word embeddings will be covered.

Table 1: The parameters of the neural model and their values used for the PharmaCoNER results.

Parameter	**Value**
Character embeddings dimension	25
Character-level LSTM hidden units	25
Word embeddings dimension	300
Word-level LSTM hidden units	256
Optimizer	SGD
Learning rate	0.001
Dropout rate	0.5
Gradient clipping	5

The results were measured with precision (P), recall (R) and F-measure (F1) using the True Positives (TP), False Positives (FP) and False Negatives (FN) for its calculation. Table 2 presents the results of the system over the test set of the PharmaCoNER tasks. The performance for the entity type classification and the performance for the Concept Indexing task are 76.29% and 60.34% in F1, respectively.

Table 2: Official results of the neural Model for the two tasks of the PharmaCoNER.

Task	**R**	**P**	**F1**
NER	71.61%	81.62%	76.29%
Concept Indexing	55.22%	66.5%	60.34%

Table 3 presents the results of the NER task for each entity type independently. It can be observed that the number of FN is higher than FP in all the classes giving better results in Precision than in Recall. The performance of the classes are directly proportional of the number of instances in the training set. In order to alleviate this problem, the use of oversampling techniques will be tackled in future works to increase the number of examples of the less representative classes and making this dataset more balanced.

5 Conclusions and Future work

This paper presents a model where a neural model classifies mentions from clinical texts in Spanish and the Concept Indexing uses the SNOMED CT search engine for their normalization. The neural architecture is based on RNNs in both direction of the sentences using LSTM for the computation of the outputs. Finally, a CRF classifier performs the classification for tagging the entity types. The results shows a performance of 76.29% in F1 for the classification of the pharmacological substances, compounds and proteins in the PharmaCoNER corpus and the normalization system reaches to 60.34% in F1. In spite of the basic approaches, the results are very promising in both tasks. As future work, it is proposed to pretrain the word embeddings with collections of biomedical documents and the aggregation of other embeddings such as Part-of-Speech tags, syntactic parse trees or semantic tags, that could increase the representation of each word in order to improve its classification. Moreover, fine-tuning the parameters of the model according to the PharmaCoNER corpus will be useful in order to increase the performance of the method. Furthermore, adding more layers to each BiLSTM is proposed to be included in the architecture. In addition, other complex concept indexing rules could be applied to chose the best nor-

[3] https://zenodo.org/record/2207130#.XHPEFYUo85k

Table 3: Performance of the neural model for each category in the Named Entity Recognition Task of the PharmaCoNER.

Label	**TP**	**FN**	**FP**	**R**	**P**	**F1**
NORMALIZABLES	707	266	94	72.66%	88.26%	79.71%
PROTEINAS	612	247	203	71.25%	75.09%	73.12%
UNCLEAR	20	14	6	58.82%	76.92%	66.67%

malization term in the cases that they are multiple possibilities.

References

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association : JAMIA*, 24.

Explosion AI. 2017. spaCy - Industrial-strength Natural Language Processing in Python.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Y. Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. Chemdner: The drugs and chemical names extraction challenge. *J. Cheminformatics*, 7(S-1):S1.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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.

Martin Pérez-Pérez, Obdulia Rabal, Gael Pérez-Rodríguez1, Miguel Vazquez, Florentino Fdez-Riverola, Julen Oyarzabal, Alfonso Valencia, Anália Lourenço, and Martin Krallinger. 2017. Evaluation of chemical and gene/protein entity recognition systems at biocreative v.5: the cemp and gpro patents tracks. In *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, page 11–18.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155. Association for Computational Linguistics.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550 – 563.