

Using Embeddings and Bi-LSTM+CRF Model to Detect Tumor Morphology Entities in Spanish Clinical Cases

Sergio Santamaria Carrasco^a, Paloma Martínez^a

^aUniversidad Carlos III de Madrid, Computer Science Department, Av de la Universidad, 30, 28911, Leganés, Madrid, Spain

Abstract

Extracting key concepts from electronic health records makes it easier to access and represent the information contained in them. The growing number of these documents and the complexity of clinical narrative makes it difficult to label them manually. This is why the development of systems capable of automatically extracting key information is of great interest. In this paper, we describe a deep learning architecture for the identification of named tumor morphology entities. The architecture consists of convolutional and bidirectional Long Short-Term Memory layers and a final layer based on Conditional Random Field. The proposed system (HULAT-UC3M) participated in CANTEMIST-NER and obtained a micro-F1 of 83.4%

Keywords

Named Entity Recognition, Deep Learning, Long Short-Term Memory, Word Embeddings

1. Introduction

With the exponential growth of clinical documents, natural language processing (NLP) techniques have become a crucial tool for unlocking critical information to make better clinical decisions. Understanding health-related problems requires the extraction of certain key entities such as diseases, treatments or symptoms and their attributes from textual data.

The identification of specific entities of interest inside medical documents can be addressed as a Named Entity Recognition (NER) problem. The different approaches to solving this problem range from dictionaries and rule-based systems, machine learning, deep learning, and hybrid systems. In similar shared tasks, such as eHealth 2019 [1], the most popular approaches were based on deep learning models, since they allow automatic learning of relevant patterns, allowing a certain degree of independence of language and domain.

This paper describes a participation of the team HULAT-UC3M in CANTEMIST 2020 challenge [2], CANTEMIST-NER subtrack. This challenge is oriented to named entity recognition


Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: sesantam@pa.uc3m.es (S.S. Carrasco); pmf@inf.uc3m.es (P. Martínez)

ORCID: 0000-0002-1923-7177 (S.S. Carrasco); 0000-0003-3013-3771 (P. Martínez)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of a critical type of concept related to cancer, namely tumor morphology that has been rarely adressed.

The core of the proposed system is an adaptation of [3] bidirectional Long Short-Term Memory (bi-LSTM) - CRF model, successfully applied previously for temporal expression recognition. This system combines several neural network architectures for the extraction of characteristics at a contextual level and a CRF for the decoding of labels. The results obtained by the system proposed by the HULAT-UC3M team reach a F1 score of 0.834. These results show the good performance of the approaches based on deep learning.

The rest of the paper is organized as follows. In Section 2 we briefly describe the datasets provided for the CANTEMIST task. In Section 3, we describe the architecture of our system. Section 4 presents the results obtained for our system. In Section 5, we provide the conclusions.

2. Dataset

The dataset [4] is described in [2]. The provided corpus was divided in three parts: training, development and validation. The training set contained 501 clinical cases manually annotated by clinical experts using BRAT standoff format¹ and following the rules for annotating morphology neoplasms in Spanish oncology clinical cases. The development set was formed by 500 additional clinical cases for evaluating machine learning systems and tune their hyperparameters. Participants could also freely use additional resources from other corpora to improve the systems.

Table 1 describes the statistics of the dataset relevant for the proposed system. A total of 22118 different words and 126 characters are in the training set, while 21031 words and 123 are in the validation set. In the union of both we reach 31329 words and 137 characters. The stop words have not been considered in the count.

Table 1
Vocabulary statistics

	Training set	Development set	Training and Development set
Words	22118	21031	31329
Characters	126	123	137

3. Methods and system description

3.1. Pre-processing

We pre-process the text of the clinical cases taking into account different steps. First, the texts are split into tokens and sentences using the Spacy², an open-source library that provides

¹<http://brat.nlplab.org/standoff.htm>

²<https://spacy.io/>

support for texts in several languages, including Spanish. We have decided to use Spacy, instead of other more specialized tools for processing clinical texts such as SciSpacy³, due to the fact that they do not support Spanish texts. Finally, the text and its annotations are transformed into the CoNLL-2003 format using the BIOES schema [5]. In this schema, tokens are annotated using the following tags:

- **B:** represents a token that conform the begining of an entity.
- **I:** indicate that the token belongs to an entity.
- **O:** represents that the token does not belong to an entity.
- **E:** marks a token as the end of a given entity.
- **S:** indicates that an entity is comprised of a single token.

3.2. Features

In this section we present the different attributes considered to be the input of the deep learning stack.

- **Words:** A representation based on pre-trained word embeddings has been used with FastText [6]. The word vectors presented in [7] have been selected due to their contribution of specific knowledge of the domain as they have been generated from Spanish medical corpora. These vectors have a total of 300 dimensions.
- **Part-of-speech:** This feature has been considered due to the significant amount of information it offers about the word and its neighbors. It can also help in word sense disambiguation. The PoS-Tagging model used was the one provided by the Spacy. An embedding representaton of this feature is learned during training, resulting in a 40-dimensional vector.
- **Characters:** We also add character-level embeddings of the words, learned during training and resulting in a 30-dimensional vector. These have proven to be useful for specific-domain tasks and morphologically-rich languages.
- **Syllables:** Syllable-level embeddings of the words, learned during training and resulting in a 75-dimensional vector is also added. Like character-level embeddings, they help to deal with words outside the vocabulary and contribute to capturing common prefixes and suffixes in the domain and correctly classifying words.
- **Meaning Cloud Named Entities:** MeaningCloud⁴ is a Software as a Service product that allows users to embed text analysis and semantic processing into any application or system. Its API, Topic's Extraction, allows extracting the entities present in a text, such as drugs or places. In addition, using a customized dictionary built from the different terms present in the National Cancer Institute's dictionary⁵, we also identify cancer treatments, types of cancer, medical tests, etc. This information is coded as a 15-dimensional one-hot vector.

³<https://allenai.github.io/scispacy/>

⁴<https://www.meaningcloud.com/es>

⁵<https://www.cancer.gov/espanol/publicaciones/diccionario>

3.3. Deep Learning Model

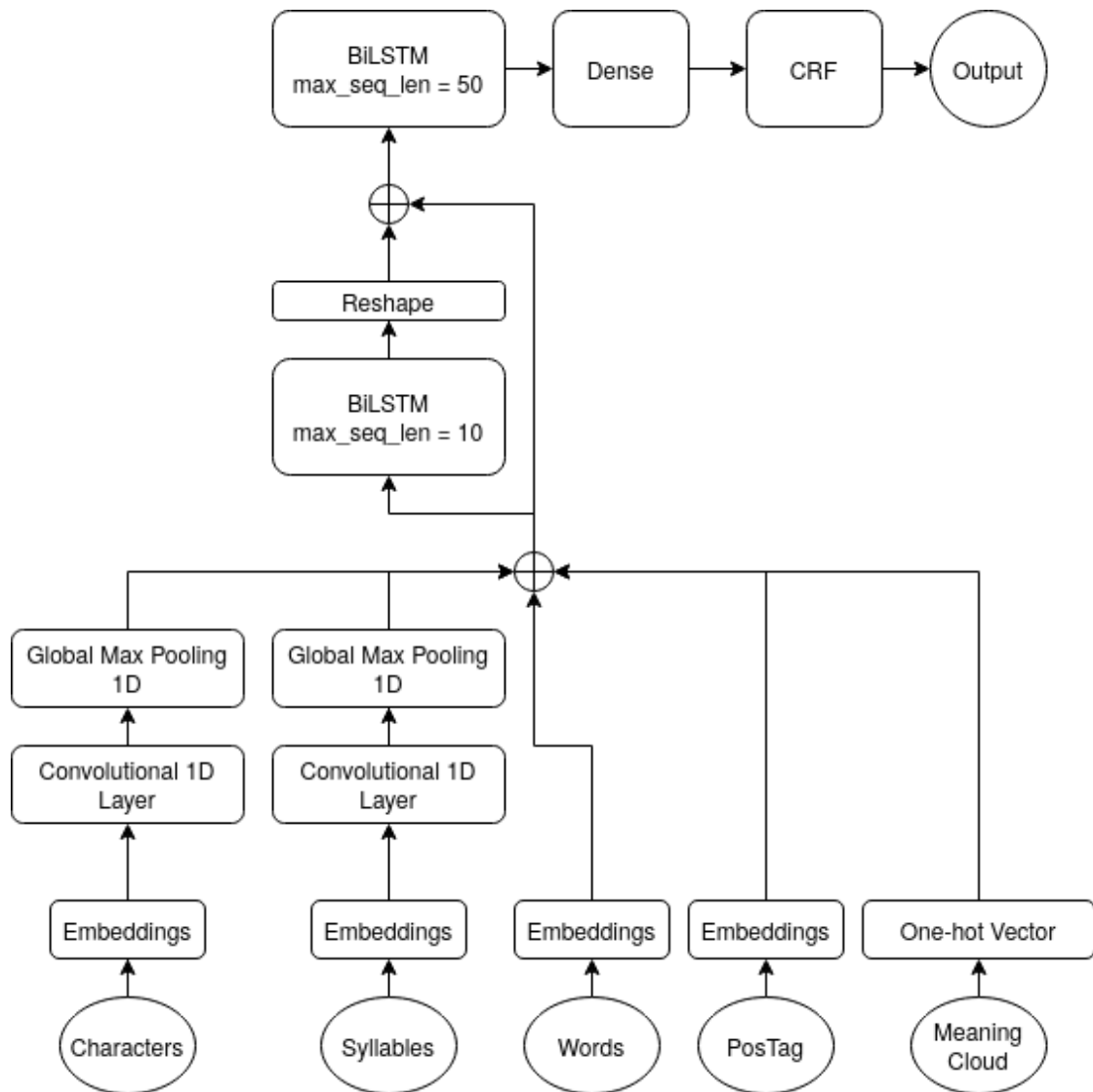


Figure 1: Architecture of the proposed model for named entity recognition.

The model implemented, as shown in Figure 1, works on two levels of maximum sequence length. First, a bidirectional LSTM layer receives the input features, the sequence of characters and syllables being previously processed by a convolutional and global max pooling block, for a maximum sequence length of 10. The reshaped output of this layer is concatenated with the input features, but this time for a maximum sequence length size of 50, and serves as the input for a new BiLSTM. This layer connects directly to a fully connected dense layer with *tanh* activation function.

The last layer (CRF optimization layer) consists of a conditional random fields layer. This layer receives as input the sequence of probabilities of the previous layer in order to improve predictions. This is due to the ability of the layer to take into account the dependencies between the different labels. The output of this layer provides the most probable sequence of labels.

The hyperparameters of our model used are listed below:

- **Maximum character sequence length:** 30
- **Maximum syllable sequence length:** 10
- **Character convolutional filters:** 50 filters of size 3
- **Syllables convolutional filters:** 50 filters of size 3
- **First BiLSTM hidden state dimension:** 300 for the forward and backward layers
- **Second BiLSTM hidden state dimension:** 300 for the forward and backward layers
- **Dense layer units:** 200
- **Dropout:** 0.4
- **Optimizer:** ADAM optimizer [8], learning rate: 0.001
- **Number of epochs:** 5

The system has been developed in python 3 [9] with Keras 2.2.4 [10] and Tensorflow 1.14.0 [11].

4. Results

In our experiments, we apply the standard measures precision, recall, and micro-averaged F1-score to evaluate the performance of our model. These metrics are also adopted as the evaluation metrics during CANTEMIST task.

We utilized the training set for training the model and exploited the development set to hyperparameter fine tuning. In the prediction stage, we combined the training and development sets to training the model. The detailed hyper-parameter settings are illustrated in Table 2 ‘Opt.’ denotes optimal.

Table 2

Detailed hyper-parameter settings in the CANTEMIST task.

Parameters	Tuned range	Opt.
Sequence length	[(100,10), (50,10), (100,20), (50,20)]	(50,10)
Train batch size	[8, 32, 64]	32
Dev batch size	32	32
Test batch size	32	32
Learning rate	[0.01, 0.001, 0.0001]	0.001
Epoch number	[5, 7, 10 ,15]	5
Dropout	[0.4, 0.5]	0.4

Considering the optimal hyper-parameters, we use these in our model to process the test set provided by CANTEMIST in subtask CANTEMIST-NER. Our proposal, as the Table 3 shows,

reaches a F1-score of 0.834, with an precision of 0.826 and a recall of 0.843. The results show that a proposal based on BiLSTM-CRF can be interesting.

Table 3

Results of CANTEMIST-NER on the test set

Precision	Recall	F1-score
0.826	0.843	0.834

Furthermore, we manually analyzed the errors generated by our system on the corpus set after the CANTEMIST-NER task. The main errors can be classified into three categories:

- **Incorrect boundaries:** This type of error is usually caused by including more information, especially related to the location of the tumor, than is necessary. An example of this found in the test set is where our system predicts "nódulo pulmonar contralateral" (contralateral pulmonary nodule) as an entity while the gold standard annotation shows "nódulo pulmonar" (pulmonary nodule). This error is also caused by not including important information such as the stage of the tumor.
- **Missing the tumor morphology mention (Missed):** This error occurs when the system does not recognize a tumor morphology mention in the clinical text. This type of error is often caused when the model confuses terminology that in other documents refers another medical condition, but in the current document refers to tumor morphology. An example of this is "lesión pulmonar" (lung lesion), which depending on the context may or may not refer to tumor morphology.
- **Incorrectly distinguishing the tumor morphology mention (Incorrectly distinguished):** This type of error is often caused when the model confuses terminology that in other documents refers to tumor morphology, but in the current document refers to another medical condition. An example of this is "lesión ósea" (bone lesion), which depending on the context may or may not refer to tumor morphology.

Table 4 details the proportion of errors made by our system in the predictions on the test set in the CANTEMIST-NER task.

Table 4

Proportion of errors made in CANTEMIST-NER task.

Incorrect boundaries	Missed	Incorrectly distinguished
39.68%	34.52%	25.80%

By analyzing these error examples, we infer the document-level information may be helpful for our system.

5. Conclusion

Due to the exponential increase in electronic health records, the need to develop systems capable of automatically identifying and classifying entities in order to facilitate the use of electronic information and resources has become stronger.

The shared task CANTEMIST is focused on the recognition of tumour morphology entities with the aim of contributing to oncological and clinical research and collecting potentially important clinical variables for cancer treatments hidden in medical texts, which are essential to promote health quality improvements and personalised medicine in a more systematised way for the advancement of cancer care.

In this document, we propose a system based on deep learning with bidirectional LSTM and CRF layers for the NER task. Our system achieves a micro-F1 of 83.4%.

In future work, our goal is to explore different ways to include document-level information, as well as examine to other deep learning architectures and other types of embeddings such as contextual embeddings.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R).

References

- [1] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the ehealth knowledge discovery challenge at iberlef 2019., in: IberLEF@ SEPLN, 2019, pp. 1–16.
- [2] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [3] G. Genthial, Tensorflow - named entity recognition, 2018. URL: https://github.com/guillaumegenthial/tf_ner.
- [4] A. Miranda-Escalada, E. Farré, M. Krallinger, Cantemist corpus: gold standard of oncology clinical cases annotated with CIE-O 3 terminology, 2020. URL: <https://doi.org/10.5281/zenodo.3978041>. doi:10.5281/zenodo.3978041, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
- [5] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), 2009, pp. 147–155.
- [6] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).

- [7] F. Soares, M. Villegas, A. Gonzalez-Agirre, J. Armengol-Estapé, S. Barzegar, M. Krallinger, Fasttext spanish medical embeddings, 2020. URL: <https://doi.org/10.5281/zenodo.3744326>, doi:10.5281/zenodo.3744326, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
- [8] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [9] G. Van Rossum, F. L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.
- [10] F. Chollet, et al., Keras, 2015. URL: <https://github.com/fchollet/keras>.
- [11] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

A. Online Resources

The sources for the HULAT-UC3M participation are available via

- GitHub <https://github.com/ssantamaria94/CANTEMIST-Participation>,