Key Phrases Annotation in Medical Documents: MEDDOCAN 2019 Anonymization Task*

Alicia Lara-Clares[1] and Ana Garcia-Serrano[2]

[1] Universidad Nacional de Educación a Distancia (UNED), Spain
alara@lsi.uned.es
[2] Universidad Nacional de Educación a Distancia (UNED), Spain
agarcia@lsi.uned.es

Abstract. There is a vast amount of digitized information about medical records, treatments and diseases, that used to be in an unstructured or semi-structured format. In order to take advantage of all the potential data that can be extracted from this information, it is necessary to deploy systems capable of converting it into annotated and structured information. In the context of the MEDDOCAN shared task of IberLEF2019, we use a Few-Shot Learning approach for Named Entity Recognition (NER) in medical documents to identify and classify key phrases in a document. The architecture of the system is an hybrid Bi-LSTM and CNN model with four input layers that can recognize multi-word entities using the BIO encoding format for the labels.

Keywords: NER · Bi-LSTM · CNN · wikipedia2vec

1 Introduction

Nowadays, there is a vast amount of digitized information about medical records, treatments and diseases, but is not completely annotated yet so there is unstructured or semi-structured information. In order to take advantage of all the potential data that can be extracted from this information, it is necessary to deploy systems capable of processing and converting it into structured information.

Recently, neural networks are shown to be especially successful in complex NLP tasks [14]. For example, G. Fabregat et al. [2] use a deep learning model for disabilities and diseases recognition using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Also the work with word embedding is one hot topic in this area, for example to simplify drug package leaflets written in Spanish [10] or to define reproducible experiments and replication datasets [6].

The MEDDOCAN task presented at the Iberian Languages Evaluation Forum (Iberlef) 2019 [8] has the objective of anonymize medical documents in Spanish. The task is structured in two sub-tasks: NER offset and entity type classification and sensitive token detection.

* Supported by UNED predoctoral grant started in April 2019 (BICI N7, November 19th, 2018)

Our work is based in the Few-shot Learning Model to learn high level features from datasets [3, 12]. We propose a hybrid Bi-LSTM CNN model by extending the model presented in [4] adding a Part-of-Speech (POS) tagging layer, that is, information about multi-word entities. Moreover, in this work, we use wikipedia2vec [13], a pre-trained word embedding model from Wikipedia. This approach to automatically extract and classify keywords is detailed at [5]. The code is available on Github [3].

The rest of the paper is organized as follows. In section 2, we describe the architecture of the system. Section 3 describes the evaluation process and results obtained. Finally, section 4 outlines the conclusions and future works.

2 System description

The system process is organized into (1) a pre-process of the data to be the input of the neural network, (2) its processing with the neural network and (3) the post-process of the output data format.

All documents are pre-processed following the next steps. First, sentences are splitted and tokenized using the Stanford CoreNLP natural language processing toolkit [7], ignoring all non-alphanumeric symbols. Then, each token is annotated using the BIO scheme, to preserve the multi-word entities. After that, we get the POS tag of each token (using the Stanford Core-NLP POS tagger).

The output of the system (annotated as shown in the Table 1: concept, POS tags and BIO-label) is converted into the BRAT format [11]. The BRAT format store all the information of the initial data together with the labels of each category and the positions of the tokens in the text.

Table 1. Structure of processed data in this work

Concept	**POS tag**	**BIO label**
Edad	PROPN	O
70	NUM	B-EDAD_SUJETO_ASISTENCIA
anyos	NOUN	I-EDAD_SUJETO_ASISTENCIA
Sexo	NOUN	O

The network architecture of this work is detailed also in [5]. It has four input layers, named as character level, word level, casing input and POS tag level, as can be seen in Figure 1.

– The character level starts with a character embedding that maps a vocabulary of 120 possible characters to an embedding initialized randomly. The maximum number of character per word is 52. It has a dropout layer (with drop rate 0.5) used to avoid the risk of overfitting. Finally, it has a convolutional layer to process the 1-dimension character layer.

[3] https://github.com/alicialara/lsi2_uned_at_MEDDOCAN2019

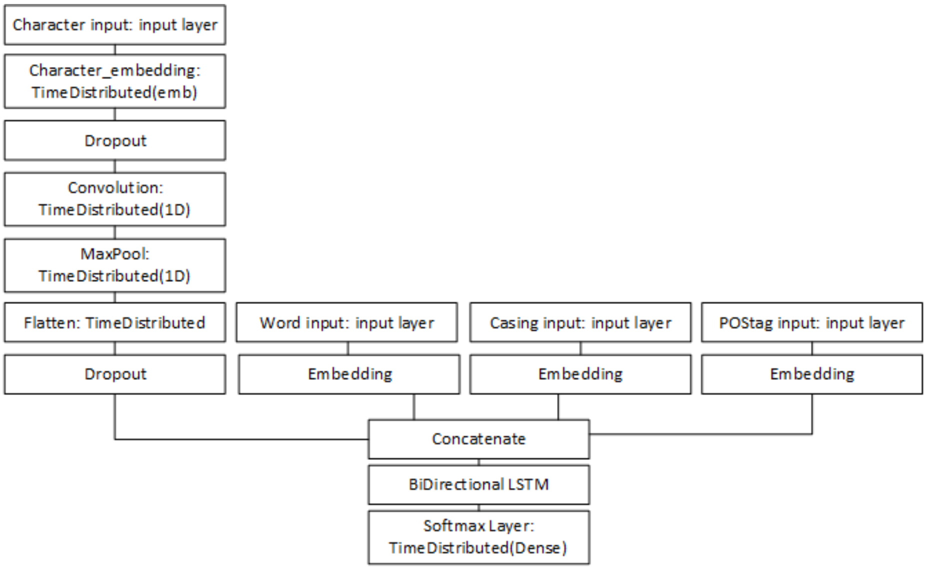
Character input: input layer
Character_embedding: TimeDistributed(emb)
Dropout
Convolution: TimeDistributed(1D)
MaxPool: TimeDistributed(1D)
Flatten: TimeDistributed
Word input: input layer
Casing input: input layer
POStag input: input layer
Dropout
Embedding
Embedding
Embedding
Concatenate
BiDirectional LSTM
Softmax Layer: TimeDistributed(Dense)

Fig. 1. Network architecture used in this work

- The second input layer uses the wikipedia2vec pretrained embeddings in Spanish language of 300 dimensions [4], mapping the existing vocabulary from the dataset.
- The third layer maps a vocabulary of eight casing types: numeric, allLower, allUpper, mainly_numeric, initialUpper, contains_digit, padding and other.
- The fourth layer maps into a one-hot embedding the POS tags existing in the vocabulary.

The system starts processing these four inputs independently, to finally merge them to be processed. The bidirectional LSTM layer Bi-LSTM [9] transforms the input data into two vectors of 200 units. Finally, the softmax function is used to obtain a prediction for locating and classifying sequences of words in the input text.

3 Evaluation and results

The evaluation of the proposed model was carried out using the MEDDOCAN corpus, that includes 1000 clinical cases, with around 495 thousand words, with an average of 494 words per clinical case. The corpus is annotated in both BRAT and i2b2 formats[5], and is divided in three sections: training, development and test. The training set comprises 500 clinical cases, and the development and

[4] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/
[5] https://www.i2b2.org/

test set 250 clinical cases each. The test set is an additional collection of 2000 documents previously non-annotated for competition purposes.

The detailed information of the evaluation is in the MEDDOCAN competition related paper [8]. There are 29 categories for key phrases and the evaluation is divided in two subtasks. The first task is an entity-based evaluation and the second one evaluates whether spans belonging to sensitive phrases are detected correctly.

In the first task, we have obtained a F-score of 90%. In the second one we have obtained a 91.5% of F-score. The documents are semi-structured, which facilitates the correct learning of certain entities. For example, patient names begin with "Name: ". The main difficulty was the detection of discontinuous, overlapped or nested entities. For example, the names in different lines are annotated discontinuously: the entity "T1 NOMBRE_SUJETO_ASISTENCIA 29 63 Pedro De Miguel Rivera" is annotated in this system as "T1 NOMBRE_SUJETO_ASISTENCIA 47 63 De Miguel Rivera" and "T2 NOMBRE_SUJETO_ASISTENCIA 29 34 Pedro". Other difficulties are the recognition of entities in the text, such as the recognition of numbers as years (ages) or dates.

4 Conclusions and Future Works

In this work, we propose a hybrid Bi-LSTM and CNN model with four input layers that can recognize multi-word entities using the BIO encoding format for the labels. Our system achieve a satisfactory performance without requiring hand-crafted features.

We plan to experiment with other BIO-based formats to detect discontinuous, overlapped or nested entities, such as BMEWO-V [15]. Moreover, we will extend the annotation using domain-specific formats and using external sources (such as Wikipedia with cui2vec format [1]).

Acknowledgements

This work was partially supported by the UNED predoctoral grant started in April 2019 (BICI N7, November 19th, 2018). The authors want to thank PhD Juan J. Lastra-Diaz for his support.

References

1. Beam, A.L., Kompa, B., Fried, I., Palmer, N.P., Shi, X., Cai, T., Kohane, I.S.: Clinical concept embeddings learned from massive sources of multimodal medical data. arXiv preprint arXiv:1804.01486 (2018)
2. Fabregat, H., Araujo, L., Martinez-Romo, J.: Deep neural models for extracting entities and relationships in the new rdd corpus relating disabilities and rare diseases. Computer Methods and Programs in Biomedicine **164**, 121 – 129 (2018). https://doi.org/https://doi.org/10.1016/j.cmpb.2018.07.007, http://www.sciencedirect.com/science/article/pii/S0169260718301330
3. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence **28**(4), 594–611 (2006)
4. Hofer, M., Kormilitzin, A., Goldberg, P., Nevado-Holgado, A.: Few-shot learning for named entity recognition in medical text. arXiv preprint arXiv:1811.05468 (2018)
5. Lara-Clares, A., Garcia-Serrano, A.: A few-shot learning model for knowledge discovery from ehealth documents (2019)
6. Lastra-Daz, J.J., Garcia-Serrano, A., Batet, M., Fernandez, M., Chirigati, F.: Hesml: a scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset. Information Systems **66**, 97–118 (2017)
7. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
8. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodrguez, H., Lopez Martin, J.A., Villegas, M., Krallinger, M.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). vol. TBA, p. TBA. CEUR Workshop Proceedings (CEUR-WS.org), Bilbao, Spain (Sep 2019), TBA

9. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
10. Segura-Bedmar, I., Martinez, P.: Simplifying drug package leaflets written in spanish by using word embedding. Journal of Biomedical Semantics **8**(45) (2017)
11. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics (2012)
12. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence (2018)
13. Yamada, I., Asai, A., Shindo, H., Takeda, H., Takefuji, Y.: Wikipedia2vec: An optimized implementation for learning embeddings from wikipedia. arXiv preprint arXiv:1812.06280 (2018)
14. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine **13**(3), 55–75 (2018)
15. Zavala, R.M.R., Martınez, P., Segura-Bedmar, I.: A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents. Proceedings of TASS **2172** (2018)