LSI2_UNED at eHealth-KD Challenge 2019

A Few-shot Learning Model for Knowledge Discovery from eHealth Documents

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. In this work, we describe a Few-Shot Learning approach for Named Entity Recognition (NER) in eHealth documents to identify and classify key phrases in a document (subtask A in the IberLEF eHealth-KD 2019 competition [10]). The architecture 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. The system obtained a F-score of 73.15% (baseline is 54,66%), with a 78,17% of precision, according to the eHealth-KD evaluation procedure. This improvement is reached mainly because (a) the correct selection of the hybrid model for NER that obtains better results using a POS tagger and (2) the addition of Wikidata entities to extend the vocabulary that improves the precision by nearly 10%.

Keywords: NER · Knowledge Discovery · Bi-LSTM · CNN · wikipedia2vec

1 Introduction

Currently, the number of medical data is growing at an exponential rate. Literature in the medical domain, moreover, is often found as unstructured or semi-structured data. In these cases, it is necessary to find methods to automatically extract and categorize the data contained in them, using different techniques as, for example, Named Entity Recognition (NER). NER aim is to recognize, identify and categorize pieces of information that refers to different entities of interest, i.e. a disease, a treatment or a patient name. First NER systems relied heavily on heuristic, hand-crafted features and language-specific knowledge as in the work presented by Rau[11] to extract and recognize company names.

In any research domain approximations based on the integration of different approaches or the integration of external resources are commonly used in order to improve the outcome of the research goal ([2], [3]). This is the case of neural

networks that are especially successful in complex NLP tasks [17], as for example, G. Fabregat et al. [5] work that use a deep learning model for disabilities and diseases recognition using Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Also research work with word embedding based techniques is frequently used, for example to simplify drug package leaflets written in Spanish [13] or to define reproducible experiments and replication datasets [8].

The aim of Few-shot Learning is to extract complex statistics and learn high level features using a very small set of training data. This problem has been addressed in several domains, such as [6] with one-shot learning, or [15] using zero-shot learning. M. Hofer et al.[7] demonstrate the effect of five sequential improvements on the learning capabilities of a neural network when having very few annotated examples, using as baseline the state-of-the-art NER architecture [4].

In this paper, we propose a hybrid Bi-LSTM CNN model following the work presented at [7]. Specifically, we have extended the model by adding a Part-of-speech (POS) tagging layer and information about multi-word entities. Moreover, in this work, we use wikipedia2vec [16], a pre-trained word embedding model from Wikipedia, and we extend the vocabulary by adding wikidata entities such diseases, health problems, etc. The results obtained in the eHealth-KD evaluation, improves the baseline by 18,5%.

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 work.

2 System description

The system process is divided into two steps. First, it is necessary to pre-process the data and prepare it to be the input of the neural network and secondly after to process the data using the implemented neural network it is needed a post-proccess of the output to be evaluated in the tasks of the IberLEF eHealth-KD 2019 competition [10]. In the next sub-sections both descriptions are included.

2.1 Pre and Post processing of the data

All documents are pre-processed following the next steps. First, sentences are splitted and tokenized using the Stanford CoreNLP natural language processing toolkit [9], 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). After the proccessing of the input data, the output data has to be converted into the BRAT format [14]. The BRAT format allows to include some aspects of the data original file, because it store all the information together with the labels of each category and the positions of the tokens in the text. Given this difference between data formats the final step is to process the documents as shown in the Table 1: concept, POS tags and BIO-label.

Word	POS tag	BIO-label
No	ADV	O
existe	VERB	B-Action
un	NUM	O
tratamiento	NOUN	B-Concept
que	CONJ	O
restablezca	NOUN	B-Action
la	DET	O
funcion	NOUN	B-Concept
ovarica	ADJ	I-Concept
normal	ADJ	B-Concept

Table 1. Structure of processed data in this work

2.2 Network architecture

The network architecture used in this work is shown in Figure 1. It has four input layers, named as character level, word level, casing input and POS tag level, described in the following:

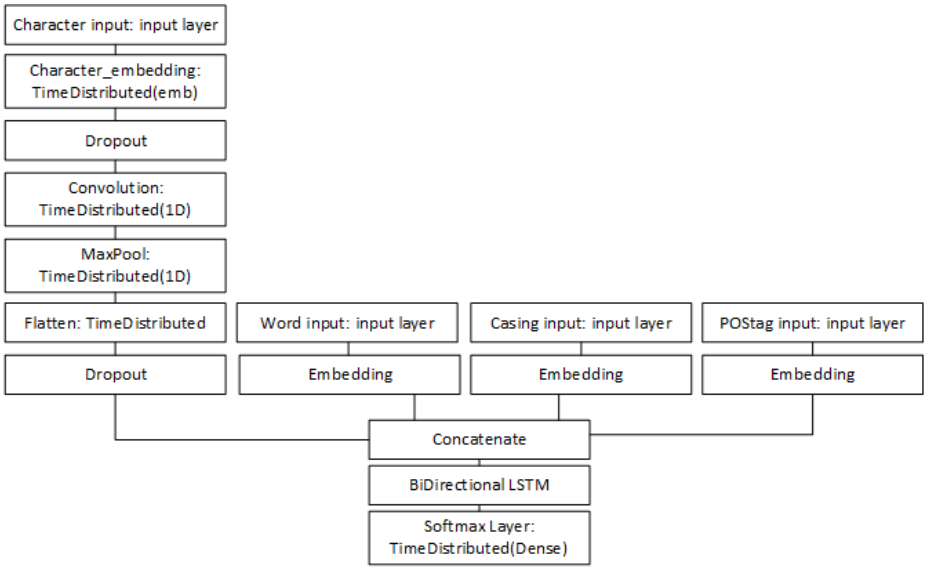

Fig. 1. Network architecture used in this work

– The first input layer corresponds to the character level. It 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.
- The second input layer uses the wikipedia2vec pretrained embeddings in Spanish language of 300 dimensions [3], 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 architecture starts processing these four inputs independently, to finally merge them into the last process. The bidirectional LSTM layer Bi-LSTM [12] transforms the input data into two vectors of 200 dimensions. In the last step, the softmax function is used to obtain a prediction for locating and classifying sequences of words in the input text.

3 Evaluation

The evaluation of the proposed model is carried out using the annotated corpus delivered in the 2019 competition that was extracted from the available MedlinePlus resources [4].

The IberLEF eHealth-KD 2019 corpus is divided in three sections: training, development and test. The training set contains a total of 600 sentences manually annotated in Brat and post-processed to match the input format. The development set has 100 annotated sentences, and the test data has 8800 non-annotated sentences for competition purposes.

Entity	Tags
Concept	B/I/O-Concept
Action	B/I/O-Action
Predicate	B/I/O-Predicate
Reference	B/I/O-Reference
Others	O

Table 2. Tokens labeled in this work

There are four categories or classes for key phrases:

1. **Concept**, a general category that indicates the key phrase is a relevant term, concept, idea, in the knowledge domain of the sentence.
2. **Action**, a concept that indicates a process or modification of other concepts.

[3] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/
[4] https://medlineplus.gov/

3. **Predicate**, used to represent a function or filter of another set of elements, which has a semantic label in the text
4. **Reference**, a textual element that refers to a concept of the same sentence or of different one, which can be indicated by textual clues.

In this work, tokens are annotated with the previous categories using the different labels (see Table 2) following the BIO encoding format.

Then the scores are computed (correct, partial, missing, incorrect and spurious matches). The expected and actual output files do not need to agree on the ID for each phrase, nor on their order. The detailed information of the evaluation is in the eHealth KD competition website [5].

3.1 Results

In this work has been carried out a series of experiments on the development corpus delivered by eHealth-KD 2019. The most interesting results are briefly described below, and they can be seen in Table 3.

Method	Recall	**Precision**	**F1**
wikipedia2vec (300) + wikidata entities + POStags	**0,6796**	**0,8429**	**0,7525**
wikipedia2vec (300) + wikidata entities	0,6887	0,8109	0,7449
wikipedia2vec (300 dim)	0,6788	0,8151	0,7407
wikipedia2vec (100 dim) + POStags	0.6515	0.7918	0.7148
wikipedia2vec (100 dim)	0,6432	0,7864	0,7077
fastext (300 dim)	0,6184	0,7638	0,6834
SBWC_glove	0,5828	0,6998	0,636
SBWC_fastext	0,5728	0,6906	0,6262
fastext (300 dim) + POStags	0.5646	0.6973	0.624
baseline	**0,6358**	**0,5416**	**0,5849**

Table 3. Results of experiments in this work

The experiments have been focused on the embeddings model used, and in the impact of the POS tagging in the neural network results. We used four embedding models Fastext [6], FastText and GloVe embeddings from SBWC [7] and wikipedia2vec [8]. The first experimental conclusions achieved are:

[5] https://knowledge-learning.github.io/ehealthkd-2019/evaluation
[6] https://github.com/facebookresearch/fastText/blob/master/docs/pretrained-vectors.md
[7] https://github.com/dccuchile/spanish-word-embeddings
[8] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/

1. The use of wikipedia2vec improves the performance and maintains the results from FasText in Spanish language.
2. Adding Wikidata entities improve the precision by approximate 10%.
3. POS tags do not improve results significantly in this task.
4. Adding fastext embeddings decreases system efficiency and does not improve results over wikipedia2vec.
5. Other embeddings in Spanish language are worse in terms of efficiency and accuracy.

4 Conclusions and Future Work

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. The vocabulary is improved using Wikidata entities such as diseases, health problems, treatments, etc. This entities are labeled as BIO-concepts and added in the corpus data as sentences. Our system can achieve satisfactory performance without requiring hand-crafted features. Our results demonstrated that in Spanish language, the wikipedia2vec pretrained embedding vectors has better performance in this task than other embeddings such as Fastext or Glove.

We plan to experiment with other BIO-based formats to detect discontinuous, overlapped or nested entities, such as BMEWO-V [18]. Moreover, we will extend the annotation using domain-specific formats and using external sources (such as Wikipedia with cui2vec format [1]).

Acknowledgements

Funding: This work was 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. Benavent, J., Benavent, X., de Ves, E., Granados, R., Garcia-Serrano, A.: Experiences at imageclef 2010 using cbir and tbir mixing information approaches. In: CEUR Proceedings. 2010 CLEF September, Padua, Italy. vol. 1176 (2010)
3. Castellanos, A., Cigarran, J., Garcia-Serrano, A.: Formal concept analysis for topic detection: A clustering quality experimental analysis. Information Systems (66), 24–42 (2017)
4. Chiu, J.P.C., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016)
5. 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)

6. 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)
7. 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)
8. Lastra-Diaz, 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)
9. 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)
10. Piad-Morffis, A., Gutiérrez, Y., Consuegra-Ayala, J.P., Estevez-Velarde, S., Almeida-Cruz, Y., Muñoz, R., Montoyo, A.: Overview of the ehealth knowledge discovery challenge at iberlef 2019. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS.org (2019)
11. Rau, L.F.: Extracting company names from text. In: [1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application. vol. 1, pp. 29–32. IEEE (1991)
12. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
13. Segura-Bedmar, I., Martinez, P.: Simplifying drug package leaflets written in spanish by using word embedding. Journal of Biomedical Semantics **8**(45) (2017)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)