Hulat-TaskAB at eHealth-KD Challenge 2019

Knowledge Recognition from Health Documents by BiLSTM-CRF

Cristóbal Colón-Ruiz[0000−0002−9167−809X] and Isabel Segura-Bedmar[0000−0002−7810−2360]

Universidad Carlos III de Madrid, Leganés, Spain
{ccolon,isegura}@inf.uc3m.es
http://hulat.inf.uc3m.es/

Abstract. Currently, the number of electronic health documents is increasing exponentially. Due to this, there is a growing interest in developing automatic systems to extract interesting information from these texts. In this paper, we describe a deep learning architecture for the identification and classification of named entities of interest in health documents. The architecture consists of two bidirectional Long Short-Term Memory layers and a final layer based on Conditional Random Fields. Our system (Hulat-TaskAB) participated in the ehealthkd-2019 sub-task A and obtained a micro-F1 of 76.63%.

Keywords: Information Extraction· Named Entity Recognition· Deep Learning · Long Short-Term Memory · Conditional Random Fields.

1 Introduction

Electronic health records (EHRs) are one of the richest resources for epidemiological studies in order to plan and evaluate strategies to identify and prevent diseases, among others. However, researchers have to manually review a large amount of information, which is a very costly and time-consuming task. Therefore, providing systems capable of automatically identifying and classifying entities or key phrases of interest in these records conforms a vital task.

The identification of specific entities of interest inside medical documents can be addressed as a Named Entity Recognition (NER) problem. This problem has been widely studied and the approaches normally used can be divided into several main categories: dictionaries and rule-based systems, machine learning, deep learning, and hybrid systems.

Dictionary-based methods are limited by the size of the dictionaries themselves, in addition to the constant growth of vocabulary and spelling errors. Rule-based approaches usually provide high precision, however, they do not usually

contemplate all existing cases as a result of the complexity of the language. For example, in [10], they used a technique of direct matching with fuzzy matching and stemmed matching. This approach was tested on the i2b2 "Heart Disease Risk Factors Challenge" dataset and achieved an F1 of 60.1% trying to maximize their recall with a limited impact on precision.

Furthermore, rule-based and machine learning methods require a previous generation of syntactic and semantic features, as well as domain-specific information. For example, in [7], they used an ensemble of support vector machines with different text annotation schemes, orthographic features, ngrams, path of speech (PoS) tags, among others. Their system was tested in the i2b2 2010 clinical data set, obtaining an F1 of 77.63.

Approaches based on deep learning methods automatically learn relevant patterns, allowing a certain grade of independence of language and domain. Moreover, these approaches have been shown to achieve better results than the best hybrid systems in i2b2 tasks. The system described in [3], which was based on Long-short Term Memory (LSTM) layers combined with Conditional Random Field (CRF) layers, scored 97.87% of F1 surpassing the winning ib2b 2014 system [13] with 96.11%, which was based on a hybrid model combining conditional random fields with keyword and rule-based approaches.

Considering the above, in this paper, we propose the use of an adaptation of the NeuroNer tool [2] for the subtask A of eHealthkd-2019 task [9] on health records in Spanish. This tool uses the combination of two bidirectional Long-short Term Memory (BiLSTM) layers with a final layer based on Conditional Random Fields.

The rest of the paper is organized as follows. In Section 2 we briefly describe the datasets provided for the eHealthkd-2019 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 data set consists of the annotated corpus of Spanish electronic health documents proposed in the ehealthkd-2019 sub-task A[1]. For this task, all the documents are provided in BRAT format[2], a standoff format where the different annotations are stored separately from the original text in a similar way to the BioNLP Shared Task standoff format[3].

As we can observe in Table 1, the training set is composed of 600 sentences annotated manually and contain a total of 4350 key phrases distributed among four classes of different categories. The development set is composed by 100 sentences annotated manually with a total of 4350 key phrases. The four categories are listed below:

[1] (https://github.com/knowledge-learning/ehealthkd-2019/tree/master/data)
[2] http://brat.nlplab.org/standoff.html
[3] http://2011.bionlp-st.org/home/file-formats

Table 1. Number of sentences manually annotated and key phrases.

	Number of sentences	Number of key phrases
Training set	600	4350
Development set	100	686

- **Concept**: Key phrase that indicates a relevant term or idea in the sentence.
- **Action**: Indicates a process or modification of concepts.
- **Predicate**: Represents a function or filter in a set of elements.
- **Reference**: Element that refers to a not explicit concept.

In addition, as we can see in Table 2, the different categories are represented unbalanced in a similar proportion for both sets.

Table 2. Number of key phrases among classes.

	Training set	Development set
Action	977	168
Concept	2899	448
Predicate	343	46
Reference	131	24

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 Spacy[4], an open-source library that provides support for texts in several languages, including Spanish.

Finally, the text and its annotations are transformed into the CoNLL-2003 format[5] using the BIOES schema [11]. 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.

[4] https://spacy.io/
[5] https://www.clips.uantwerpen.be/conll2003/ner/

3.2 Network description

Bidirectional LSTMs are a type of recurrent neural network (RNN) where the context of words in the sentence is captured using information from previous words and information from subsequent words. In addition, to improve the accuracy of the predictions provided by the BiLSTM layer, the CRF layer uses information from neighboring (sentence level) tags to predict current tags. Due to the performance of this type of architectures in entity recognition tasks, we propose the use of the NeuroNer [2] tool for eHealthkd-2019 subtask A. This tool is composed of three main layers (see Figure 1):

1. Character-enhanced and token embedding layer.
2. BiLSTM prediction layer
3. CRF optimization layer

The first layer (Character-enhanced and token embedding layer) aims to generate vector representations of the tokens that conform the input sequences. The direct representation of token to vector (word embedding) can be pre-trained or can be learned in conjunction with the rest of the model by adjusting its weights. Pre-trained models can be obtained from a large amount of unlabeled data with methods such as word2vec, FastText or GloVe [6, 1, 8]. However, the different word embedding models do not contain representation for those tokens not included in their vocabularies. The first layer addresses this problem by incorporating a representation of tokens based on their characters (character embeddings). Each token character is represented by its own vector, allowing the network to learn morphological information even from tokens that are not included in the vocabulary of the word embedding model [5].

The character embedding sequence of each token is passed as input to a BiLSTM to obtain character-based word embedding as output. Finally, the representation of word embeddings and character-based word embedding are concatenated for each token, which will be the input for the second BiLSTM layer (BiLSTM prediction layer). This BiLSTM layer aims to obtain the sequence of probabilities for each token to pertain to a given label using the BIOES coding. The label for each token will be the one with the highest probability.

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 for the eHealthkd-2019 subtask A are listed below:

- **Word Embeddings**: randomly initialized and adjusted during training. The dimension of the vectors is 200.
- **Character Embeddings dimension**: randomly initialized and adjusted during training. The dimension of the vectors is 25.

– **First BiLSTM hidden state dimension**: 25 for the forward and backward layers
– **Second BiLSTM hidden state dimension**: 200 for the forward and backward layers
– **Optimizer**: ADAM optimizer [4], learning rate: 0.001
– **Dropout**: 0.5
– **Number of Epochs**: 100

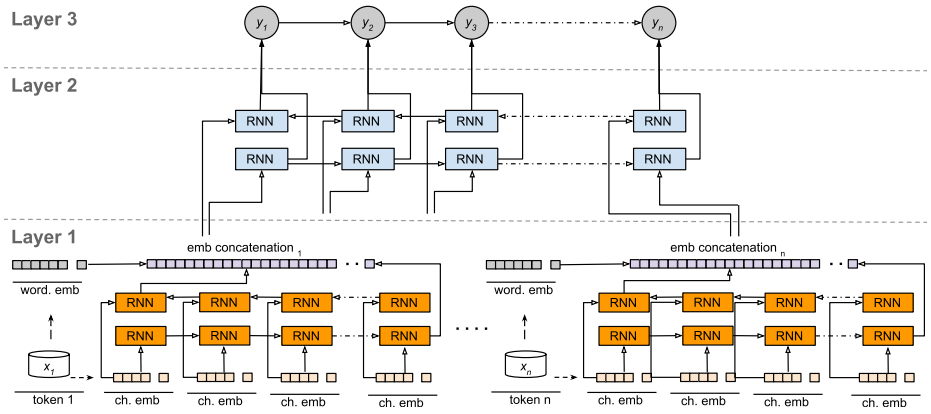

Fig. 1. Overview architecture of our system (Hulat-TaskAB)

4 Results

In our experiments, we use precision, recall, and F1 score to evaluate the performance of our system. In eHealthkd-2018 subtask A, several criteria are taken into account:

– **Correct matches**: When the spans of the predicted key phrase and its label match exactly with its entry in the gold file.
– **Partial matches**: The label matches between the predicted phrase and its entry in the gold file, but there is overlap in its spans.
– **Incorrect matches**: The spans of the predicted key phrase and its label match exactly, but not its labels.
– **Missing matches**: Those entries that appear in the gold file but not in the predicted file
– **Spurious matches**: Those entries that appear in the predicted file but not in the gold file

A more detailed description of the evaluation can be found on the website[6].

[6] https://knowledge-learning.github.io/ehealthkd-2019/evaluation

To evaluate the trained models, as well as their hyperparameters, we performed a set of experiments with the development dataset provided by the eHealthkd-2019 organizers. We used grid search to adjust the word embeddings dimension, the number of units in the BiLSTM hidden layer, the optimizer and the learning rate.

We can observe in Table 3 our best results on the development set. The run0 model was trained using the hyperparameters mentioned in section 3.2. The run1 model was trained using ADAM as optimizer, with a word embeddings dimension of 200 and 100 units in the second layer of the BiLSTM. The run2 model was trained using ADAM employing a word embeddings dimension of 100 and 100 units in the second layer of the BiLSTM.

Table 3. Results of Subtask A on development set. The top scores are bold.

	Precision	Recall	F1
run0	0,7886	0,7781	**0,7833**
run1	0,7696	0,7798	0,7747
run2	0,7846	0,7781	0,7814

Considering that the model run0 achieved the best results, this model was used to process the test set provided by eHealthkd-2019 in subtask A. The results obtained can be seen in Table 4.

Table 4. Results of subtask A on the test set.

	run0
F1	0,7663
Recall	0,7817
Precision	0,7515
Correct	476
Incorrect	57
Missing	55
Partial	58
Spurious	81

As we see in Table 4, the test set of subtask A contains 646 key phrases. In total, the spans and labels match correctly in 476 of them and partially in 58. One of the elements that most affect our result is the number of spurious phrases, resulting in decreased precision.

5 Conclusions

Electronic health records (EHRs) are sources for a wide range of studies to plan and evaluate strategies for identifying and preventing disease. However, due to the exponential increase in health records, manually reviewing that amount of information is a very expensive and time-consuming task. Therefore, providing systems capable of automatically identifying and classifying entities or key phrases of interest in these records is a vital task.

One of the biggest challenges of this shared task is that there are a large number of entities or key phrases of interest, but they are often unbalanced in the text. All this, together with the presence of nested, discontinuous or overlapping entities, results in difficulties in classifying them correctly.

In this document, we describe our system involved in the sub-task A proposed by eHealthkd-2019. It exploits the NeuroNer tool, a tool based on deep learning with bi-directional LSTM and CRF layers for the NER task. Considering the challenges described above, our system achieves a micro-F1 of 76.63% on the test equipment.

For future works, we plan to explore other deep learning architectures as well as exploiting pre-trained word embedding models, as well as other types of embeddings such as sense embeddings [12]. We also plan to explore the performance of our system by expanding the BIOES annotation scheme [11] to address the problem of overlapping, nested or discontinuous key phrases. This approach could reduce the number of partial matches and increase the number of exact matches. In addition, due to the unbalanced data, we also plan to explore how the weighting of different classes in training can affect the performance, as well as the use of different sampling methods.

Acknowledgements

Funding: 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. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
2. Dernoncourt, F., Lee, J.Y., Szolovits, P.: Neuroner: an easy-to-use program for named-entity recognition based on neural networks. arXiv preprint arXiv:1705.05487 (2017)
3. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association **24**(3), 596–606 (2017)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

5. Ling, W., Luís, T., Marujo, L., Astudillo, R.F., Amir, S., Dyer, C., Black, A.W., Trancoso, I.: Finding function in form: Compositional character models for open vocabulary word representation. arXiv preprint arXiv:1508.02096 (2015)
6. 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)
7. Nayel, H., Shashirekha, H.: Improving ner for clinical texts by ensemble approach using segment representations. In: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017). pp. 197–204 (2017)
8. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
9. 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 (2019)
10. Quimbaya, A.P., Múnera, A.S., Rivera, R.A.G., Rodríguez, J.C.D., Velandia, O.M.M., Peña, A.A.G., Labbé, C.: Named entity recognition over electronic health records through a combined dictionary-based approach. Procedia Computer Science **100**, 55–61 (2016)
11. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the thirteenth conference on computational natural language learning. pp. 147–155. Association for Computational Linguistics (2009)
12. Trask, A., Michalak, P., Liu, J.: sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. arXiv preprint arXiv:1511.06388 (2015)
13. Yang, H., Garibaldi, J.M.: Automatic detection of protected health information from clinic narratives. Journal of biomedical informatics **58**, S30–S38 (2015)