Hulat-TaskA at eHealth-KD Challenge 2019

Sequence Key Phrases Recognition in the Spanish Clinical Narrative

Alejandro Ruiz-de-laCuadra[1], Jose Luis Lopez-Cuadrado[1], Israel Gonzalez-Carrasco[1], and Belen Ruiz-Mezcua[1]

Universidad Carlos III de Madrid, Computer Science Department, Av de la Universidad, 30, 28911, Leganes, Madrid, Spain
{aruizla,jllopez,igcarras,bruiz}@inf.uc3m.es

Abstract. Key phrases recognition is one of the open issues in Natural Language Processing. These entities are relevant to identify relations between phrases and allows extracting knowledge from unstructured text. This paper combines Recurrent Neural Networks and Conditional Random Fields to present a trending architecture to solve the Scenario 2 problem for identification and classification of key phrases at eHealth-KD Challenge 2019. With a performance measure F-score of 0.7903, this team HULAT-TaskA achieved the fourth position.

Keywords: Bi-LSTM · CRF · NER · Knowledge Discovery

1 Introduction

With the exponential growth of clinical documents, the task of structuring this data has become more complex and unfeasible. The automation of this process allows solving the scalability problem and continuing to offer knowledge mining for unstructured texts.

In the clinical domain, the extraction of key concepts and their relationships allow to have a better understanding of the diagnosis and to make a better follow-up of similar or related cases.

This language processing starts with the determinant task of named entity extraction (NER). Currently, the possible solutions to solve the NER problem are divided into methods based whether on dictionaries, rules or machine learning. Firstly, dictionaries are limited by the size and diversity of vocabulary, misspellings, the use of synonyms and abbreviations. Secondly, despite the fact that rule-based methods are at the peak of performance for this task, domain dependency to build effective rules makes NER a laborious and difficult to expand work. Finally, machine learning approaches have steadily progressing. The

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). IberLEF 2019, 24 September 2019, Bilbao, Spain.

CEUR-WS.org/Vol-2421/eHealth-KD_paper_2.pdf

competitive advantage of this last one lies in the simplicity of building and configuring the systems, the performance to extract characteristics or syntactic and semantic patterns, and its versatility in the domain and language.

While machine learning systems were already governed by Conditional Random Fields (CRF) methods, the emergence of hybrid systems combining deep learning and CRF was the necessary boost to be on a level with rule-based methods.

This paper describes the participation of the team HULAT-taskA in the IberLEF 2019 eHealth-KD challenge [4] . This challenge is oriented to the identification of key-phrases and their corresponding relationships in eHealth records in Spanish. The challenge is structured in two different subtasks: A and B. The goal of subtask A is to identify the key phrases per document and their classes. The goal of subtask B is to link the key phrases detected and labelled in each document. These two subtasks lead to three different scenarios. Scenario 1 covers the two subtasks as a pipeline, Scenario 2 evaluates the subtask A and Scenario 3 evaluates the subtask B. The team HULAT-TaskA takes part only in Scenario 2 (subtask A).

The core of the proposed tagger system is an adaptation of [3] bidirectional Long Short-Term Memory (bi-LSTM) - CRF model, successfully applied previously for temporal expression recognition. This tagger 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 in this task by the HULAT-TaskA team, F1 0.7903, show a performance similar to the obtained for the temporal expressions, applying a different approach at the character level. These results demonstrate the versatility of hybrid systems for the extraction of entities.

2 Dataset

The dataset is described in [4]. The provided corpus was divided in three parts: training, development and validation. The training set contained a 600 sentences manually annotated using brat standoff format and post-processed to match the input format. The development set was formed by 100 additional sentences for evaluating machine learning systems and tune their hyperparameters. For scenarios 2 and 3, only the 100 valid sentences are published. Participants could also freely use additional resources from other corpora to improve the systems. Although the team HULAT-TaskA used the previous year challenge dataset to extend the word and character vocabulary with more vectors and to start testing possible architectures, the team has not participated in the previous challenge.

Table 1 describes the statistics of the dataset relevant for the proposed system. A total amount of 2626 words were provided in the dataset, with 76 different characters and 8 labels for classifying the key-phrases. Figure 3 represent the relevant statistics related to the labels of the corpus, from the point of view of the IOB (Inside, Outside, Begin) format. As shown in the figure, some concepts are inside other ones. This is relevant for the configuration of the system, since we

decide not look for these elements in order to improve the results. We found that the percentage of F-measure lost due to the omission of these concepts with several words was lower than the percentage of F-Measure lost by the difficulties of the system to learn this type of concepts. So the system was only trained with one-word concepts.

Table 1. Vocabulary statistics

	Words	Tags	Characters
Vocabulary	2626	8	76

3 System Description

3.1 Pre-processing

Since the main architecture (Figure 1) has been designed based on the architecture of [3] for the recognition of multiple entities, it is necessary to include a pre-processing module for adapting the input data to the format expected by the system. This modular system allows adapting the proposed architecture to other problems.

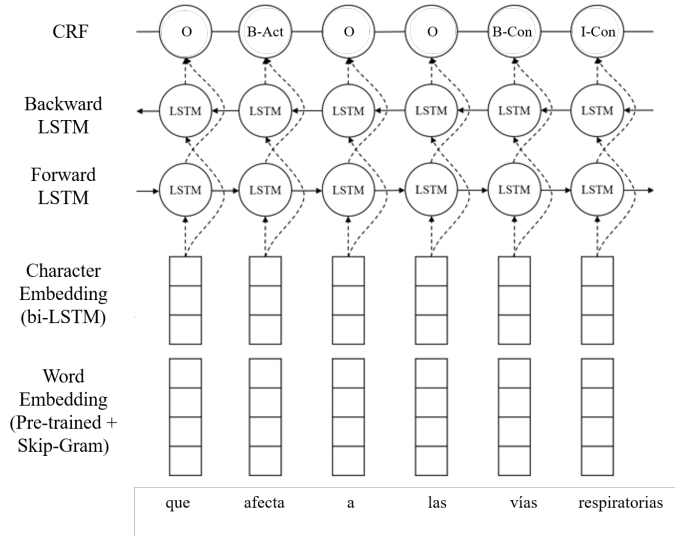

Fig. 1. Main system architecture

In the context of this workshop, data is provided in the format:

```
ID \tab START END ; START END \tab LABEL \tab TEXT
```

This module transform this format into a different token-oriented organization with its corresponding tags, as shown below:

```
docId \tab sentId \tab tokId \tab tokTxt \tab tag \tab tagId
\tab type \tab val
```

This new structure allows processing the tokens using the customized tagger module while minimizing the loss of information. At the same time, it saves the information required for reverting the process.

Finally, after adapting the data to the new format, the system can automatically generate the training and labelling files following the IOB format.

3.2 Model Architecture

Input The input of the model is composed by two levels: character and word.

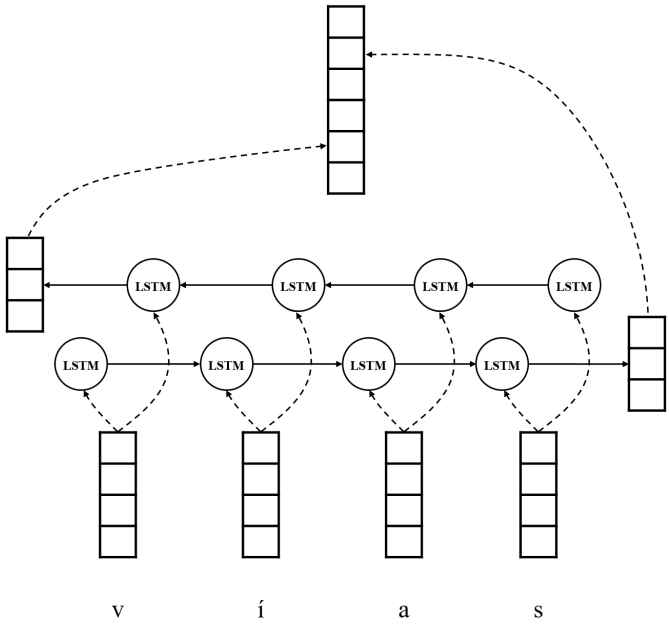

Fig. 2. Char representation using bi-LSTM

Character level representation provides additional lexical information relative to each word n in sentence s. Figure 2 depicts the design of the Recurrent Neural

Network (RNN) for obtaining the dimension 50 vector. A conversion table allows obtaining a numerical representation of each character n of the sentence s (c_n, s). This vector is the input of the RNN. By means of a many-to-one architecture the two LSTM layers in opposite directions represent more complex characteristics than the Convulutional Neural Network (CNN) (applied in previous research on temporal expressions). Finally, the outputs of the LSTM layers (dimension 25) are linked together and a dropout layer is included in order to avoid overfitting.

Figure 2 represents the inputs of the bi-LSTM layer. At the word level a conversion table has been applied, based on the numerical values calculated in the word embeddings Spanish Billion Words [2] which returns a 300 dimension vector (w_n, s).

LSTM Layer Weight training is focused on this layer of recurrent networks with a high number of nodes in order to address the complexity of this problem. The network takes the combination of both representations:

$X_{n,s} = C_{n,s} + W_{n,s}$

As in the representation of characters, the two layers are concatenated and a dropout layer is applied. As a result, a word representation is obtained in the sentence context (h1,..., hn):

$h_t = [\overrightarrow{ht}, \overleftarrow{ht}]$

This system allows us to capture multiple dependencies between the tags. Thanks to the use of two layers in both directions, the system allows exploiting the potential of the LSTM and capturing contextual information in both directions.

Conditional Random Fields To adjust the combined output of the LSTM layers, a CRF layer just after a dense layer has been include. This layer allows us to decode the output considering the neighbors against the Softmax function that makes the decision to tag independently.

As a function of loss, the function log-likelihood of tag sequences in a CRF has been used.

3.3 Post-processing

The post-processing module takes output from the tagger and sentence metadata from pre-processing in order to transform tagged data into brat format. Additionally, a set of rules have been applied to discard any incoherent IOB (Inside, Outside, Begin) labels. In fact, these rules have been created using data statics such as label distribution (Figure 3).

4 Results

Five different models were tested for the competition, but only the best result was sent. Table 2 summarises the different architectures tested. During the training phase, F-score, precision and recall measures were applied. TensorBoard [1]

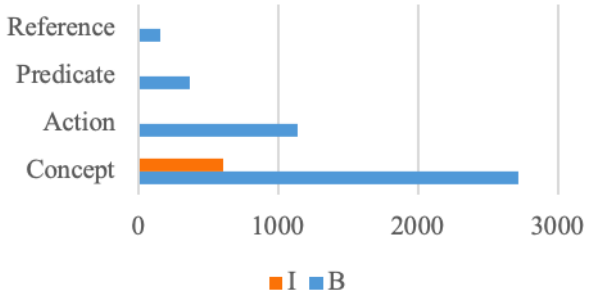

Fig. 3. Phrases statistics

Table 2. Models Configuration

Model	Char	Embedding	Batch	Dropout	Epochs
1	LSTM	LSTM	2	0.7	50
2	LSTM	LSTM	16	0.5	50
3	LSTM	LSTM	20	0.5	50
4	CNN	LSTM	20	0.5	50
5	LSTM	LSTM	16	0.2	50

was applied for analysing the non desired behaviors. For the final results, the proposed measure by the organization of the task was applied (https://knowledge-learning.github.io/ehealthkd-2019/evaluation).

Table 3 show the results obtained for each of the models proposed in Table 2. Results vary around 92% in the training set, and 78-79% for the rest of the models. This big difference makes relevant the selection of the dropout value in order to reduce the difference and avoid over fitting.

Models 1 and 2 (Table 3) obtain similar result (around 1% of difference). The batch size produces a higher cost in the training phase, but reduces the number of false positives.

The performance of the rest of the models do not improve the results of Model 1, so it was the selected for the competition.

Table 4 describes the results for scenario 2. A total of 9 teams improves the results of the baseline. The system proposed in this paper was ranked in the 4th position with a F-measure of 0.7903, 3 points below the winner TALP with an F-measure of 0.8203.

In future work, Model 1 can be improved by modifying the labeling format or adjusting the dropout of the different layers.

Table 3. Models results

Model 1	F1	Precision	Recall	Correct	Incorrect	Partial	Spurious	Missing
Training	0.9281	0.9319	0.9243	3491	108	76	112	143
Development	0.799	0.7951	0.803	461	46	48	55	49
Testing	0.7879	0.7688	0.7879	499	52	46	82	49
Model 2	F1	Precision	Recall	Correct	Incorrect	Partial	Spurious	Missing
Training	0.9209	0.921	0.9208	3471	117	89	140	141
Development	0.7824	0.7791	0.7856	451	55	47	56	51
Testing	0.7778	0.76	0.7964	489	52	51	85	54
Model 3	F1	Precision	Recall	Correct	Incorrect	Partial	Spurious	Missing
Training	0.9268	0.9321	0.9216	3482	117	73	103	146
Development	0.7784	0.7863	0.7707	442	51	47	52	64
Testing	0.7795	0.7661	0.7933	488	56	49	76	53
Model 4	F1	Precision	Recall	Correct	Incorrect	Partial	Spurious	Missing
Training	0.8383	0.8558	0.8215	3053	248	167	197	350
Development	0.7643	0.7863	0.7434	430	57	38	46	79
Testing	0.7483	0.7465	0.75	458	63	53	75	72
Model 5	F1	Precision	Recall	Correct	Incorrect	Partial	Spurious	Missing
Training	0.9637	0.9774	0.9505	3613	45	32	23	128
Development	0.7891	0.7978	0.7806	451	46	41	53	66
Testing	0.7674	0.7564	0.7786	476	57	54	78	59

Table 4. Results for Scenario 2

No.	Team	F-score	Precision	Recall	System
1	TALP	0.8203	0.8073	0.8336	Joint-BERT-RCNN
2	LASTUS-TALN (abravo)	0.8167	0.7997	0.8344	
3	UH-MAJA-KD	0.8156	0.7999	0.8320	MeDeepCal
4	Hulat-TaskA	0.7903	0.7706	0.8111	RNN-ICK
5	coin_flipper (ncatala)	0.7873	0.7986	0.7763	Voting LSTMs
6	Hulat-TaskAB	0.7758	0.7500	0.8034	
7	NLP_UNED	0.7543	0.8069	0.7082	DeepNER+ARE
8	lsi2_uned	0.7315	0.7817	0.6873	
9	IxaMed	0.6825	0.6567	0.7105	
10	baseline	0.5466	0.5129	0.5851	Baseline
10	VSP	0.5466	0.5129	0.5851	Baseline

5 Conclusions

This paper has presented RNN-ICK a bi-LSTM architecture for recognizing key phrases in the context of the Scenario 2 (subtask A) of the IberLEF 2019 eHealth-KD challenge. The proposed system combines RNN architecture with a CRF in the output. The proposed architecture was adapted from a previous one applied to the temporal expression recognition, modifying the input processing and adapting the labels to be processed. The proposed system reached the 4th position in the competition. The results obtained were similar to the temporal scenario. These results show the ability of hybrid systems for adapting to several NER scenarios.

Moreover, the results show the importance of generating more information at the character level, being the configurations with LSTM at all times superior to the configurations with CNN. This detail also clarifies that it is necessary to give more knowledge to the model, since the pre-trained embeddings by itself are not enough, either due to the Skip-Gram algorithm or to the specific domain. In the previous edition (2018), the winning system used a very similar architecture [5], although achieving a result 8 percentage points above this F-score. In future work, the use of BIOES-V labeling to differentiate the beginning and end of the entity, together with embeddings at sentence level, will add enough information to address the complexity of the problem. These modifications are feasible to add thanks to the modular design of the proposed system. Further research will also include the use of specific Word Embeddings based on clinical data in Spanish as well as the testing of other architectures.

Acknowledgments

Funding: This work was supported by the Research Program of the Ministry of Economy and Competitiveness Government of Spain (Project DeepEMR: Clinical information extraction using deep learning and big data techniques TIN2017-87548-C2-1-R).

We also thank the organization committee from eHealthKD Challenge 2019 for providing all resources.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, P., Chen, Z., Citro, C., Corrado, G.S., Davis, A., et al.: Tensorflow: Large-scale machine learning on heterogeneous systems, 2015 (2015), tensorflow.org
2. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (March 2016), https://crscardellino.github.io/SBWCE/
3. Genthial, G.: Tensorflow - named entity recognition (2018), https://github.com/guillaumegenthial/tf_ner
4. 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)

5. 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)