HapLap at eHealth-KD Challenge 2020

Sergio Santana[a], Alicia Pérez[a] and Arantza Casillas[a]

[a]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, 20080 Donostia, Spain

Abstract

We present the work carried out by the HapLap group in the in the subtask B of the eHealth-KD 2020 competition. Relation extraction was addressed with a pipeline system that makes use of a Joint AB-LSTM neuronal network together with a pre-process and a post-process phase. We obtained a result of 0.316 in Scenario 3.

Keywords

Entity recognition, Relation extraction, Joint AB-LSTM neuronal network.

1. Introduction

We present the work carried out by the HapLap group in the eHealth-KD 2020 task [1]. In this third edition the purpose of the task is to automatically extract knowledge, represented by means of thirteen semantic relations, from Spanish electronic health documents. We have taken part in the optional subtask B: the input is a plain text with entity annotations in a BRAT file and the output is the previous BRAT file with both the entities and relations. To address this, we have implemented a pipeline system that makes use of a Joint AB-LSTM neuronal network together with a pre-process and a post-process phase.

2. Related Work

In the last years various competitions related to relation extraction have been emerging such as: Semeval 2018 task 7 [2] to extract relations from scientific texts; eHealthKD 2018 [3], eHealthKD 2019 [4] or BioNLP [5] to extract and classify clinical relations from clinical texts. So the relation extraction problem is arousing interest in different areas and also in the clinical documentation area. Since the resurgence of neural networks, different approaches have been implemented for extracting clinical relations. DET-BLSTM system [6] makes use of a Bi-LSTM network. In [7] the authors presented a combination of two different networks gated recurrent unit (GRU) and convolutional neural network (CNN) to detect clinical relations. In [8] a convolutiona neural network is also used to classify relations. In [9] an Joint AB-LSTM neuronal network is used to extract adverse drug reaction relations. In this paper we present a Joint AB-LSTM neuronal, a

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: ssantana005@ikasle.ehu.eus (S. Santana); alicia.perez@ehu.eus (A. Pérez); arantza.casillas@ehu.eus (A. Casillas)

ORCID: 0000-0003-2638-9598 (A. Pérez); 0000-0003-4248-8182 (A. Casillas)

modification of the work presented in [10] network for the extraction of clinical relations in the context of eHealthKD 2020 competition.

3. Materials and Methods

For this work we have divided the system into three phases: First the pre-process, where we adapt the data format to use with the Joint AB-LSTM. After that we have the training phase, where we train and evaluate the neural network and we get the prediction. And after getting the predictions we have the post-process, where we convert those predictions into the data format that is used in the competition.

3.1. Pre-process

In the preprocess we do the following operations:

- Convert the input from the Brat standoff format to the format used in the eHeathKD 2019 challenge.
- Convert the data in the eHealthKD 2019 format into the format used by the Joint AB-LSTM.
- Create the NO_RELATION relations.

In the first part of the system we have pre-processed the input relations. We have converted the Brat Standoff input relation-format (also referred to as `ann`) to the format used in the previous eHealthKD 2019 competition by means of the `ann2txt` scripts (https://github.com/knowledge-learning/ehealthkd-2019/blob/master/scripts/ann2txt.py) provided there. Next, we needed to adapt it to what the Joint AB-LSTM requires. Three programs have been implemented for the pre-processing and their code has been posted on GitHub (https://github.com/Porobu/HAPLAP-MAL). These three programs load the instances that are in the eHealthKD 2019 data format and they join them into a single file.

In an attempt to enable the neural network learn to discriminate between positive and negative relations (absence of relation), both types of instances should be provided in the inference stage. To this end, in the pre-processing an auxiliary relation class, NO_RELATION, was also created. A critical point, hence, is how to choose instances that contain pairs of entities that could be related and, thus, are candidate relations and label them as negative instances. Both the selection and the proportions might be crucial. We have used a simple way of choosing them, that only creates negative (NO_RELATION) relations between entity pairs that have at least one positive relation instance in the data set. To further reduce the negative relations, we have only created these between entity pairs in the same sentence.

At this stage we have a set of data with the candidates marked as either related or not-related. At this point a multi-class approach enables us to predict whether a candidate pair is related with some of the relation-classes available (including NO_RELATION). This was, indeed, our **approach-1**: a pair of entities that could be related (are a relation-candidate) are directly classified by means of the Joint AB-LSTM.

Needless to say, in the aforementioned sample negative instances substantially exceed the positive ones leading to skewed class distribution. In table 1 we can see the number of positive

Table 1
Number of positive, negative and the total relations in the training and development data sets

Data set	Positive Relations	Negative Relations	Total
Training	8597	50812	59409
Development	1204	7144	8348

and negative relations in our training and development data sets. We have to remember that in our multi class classification approach (approach 1) the positive relation number contains all the thirteen classes, further skewing the data. Inference tends to be biased towards majority class. To cope with this we proposed to tackle the classification in two stages (our **approach-2**):

- In the first phase we have created the binary data set, and all the positive relations (target, causes...) have been grouped in the *RELATION* class. In this phase we filter all the negative relations, to reduce the imbalance.
- In the second phase we have now only the data set with the positive relations (arg, target, subject...), and we train the system to predict the relation.

Both approaches (and both phases in the second approach) were implemented by means of the Joint AB-LSTM approach. Further details are given in the following section.

3.2. Joint AB-LSTM network

After pre-processing the instances we load them into the Joint AB-LSTM neural network. The Joint AB-LSTM neural network has been inplemented by using Tensorflow. The network also does its own pre-processing. First all tokens are lower-cased.

The network employed word-embeddings as the main feature. For this work we have used pre-trained embeddings from the clinical domain. The embeddings have been trained in corpora that consists of EHRs (electronic health records) that are not publicly available due to confidentiality issues. Other choices might have resulted more appropriate than ours since the amount and type of data employed has a big impact on the resulting embeddings. Apart from the word-embeddings, the network employs another powerful feature: the distance-embeddings. The distance is simply computed as the number of tokens between each annotated word in the sentence and the target word entity.

Having the relations completely pre-processed, the neural network is trained. This network combines two widely used neural networks in NLP: a Bi-LSTM with max pooling and an attentive Bi-LSTM. The Joint AB-LSTM is fed with the pre-processed sentences, their entities and relations between those, and the previously created distance embeddings.

We have optimised two hyper-parameters of the neural network, the dropout and learning rate to get the final model. We have trained the model with a mixture of the eHealthKD 2019 train+dev and the eHealthKD 2020 datasets, and we have used the eHealthKD 2020 dev dataset as validation. Note that this optimisation has been done over the so called *multiclass dataset* (approach 1), not over the *binary dataset* (approach 2). After doing the optimisation, we set 0.001 as the learning rate, and we used no dropout.

Table 2
Results on the eHealthKD 2020 dev dataset attained with Approach 1 (multi-class) and Approach 2 (working in two phases to filter binary relations).

	Precision	Recall	F1
Approach 1	0.336	0.298	0.316
Approach 2	0.328	0.306	0.316

3.3. Postprocess

After getting the predictions from the neural network, we postprocess them to get the output relations in the Brat Standoff format, respecting the IDs if the gold entities.

4. Results

As described in section 3.1, we provided two different approaches. The results achieved with each of them are given in table 2.

Approach 1 outperforms Approach 2 in terms of precision but with the recall occurs the opposite. Nevertheless, for both approaches the F1-measure has the same value.

5. Conclusions

Relation extraction was addressed with a neural approach, Joint AB-LSTM network. We applied two simple pre-processing approaches to get both positive and negative instances. This stage might result naive for the way in which the sampling was carried out and the proportions selected. We explored two pre-processing approaches: a straight one, approach 1, which just copes with multi-class problem; a filtered one (approach 2) that tried to get rid of negative candidates prior to the multi-class stage. None of them surpassed the other significantly. For future work, we should explore the embeddings provided to the network. Embeddings are the main source of knowledge in this stage with limited training sets and was proven significantly influential in related works.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science and Technology PAD-MED (PID2019-106942RB-C31) and by the Basque Government (IXA IT-1343-19 and a Grant for the student Sergio Santana published in the 12/03/2020 BOPV).

References

[1] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Diaz, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020, in: Proceedings of the Iberian Languages Evaluation Forum co-located with

36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, Spain, September, 2020., 2020.

[2] K. Gábor, D. Buscaldi, A.-K. Schumann, B. QasemiZadeh, H. Zargayouna, T. Charnois, Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers, in: Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 679–688.

[3] E. Martínez Cámara, Y. Almeida Cruz, M. C. Díaz Galiano, S. Estévez-Velarde, M. Á. García Cumbreras, M. García Vega, Y. Gutiérrez, A. Montejo Ráez, A. Montoyo, R. Muñoz, et al., Overview of tass 2018: Opinions, health and emotions (2018).

[4] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Munoz, A. Montoyo, 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.

[5] D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019. URL: https://www.aclweb.org/anthology/W19-5000.

[6] L. Li, J. Zheng, J. Wan, D. Huang, X. Lin, Biomedical event extraction via long short term memory networks along dynamic extended tree, in: Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, IEEE, 2016, pp. 739–742.

[7] B. He, Y. Guan, R. Dai, Convolutional gated recurrent units for medical relation classification, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018, pp. 646–650.

[8] S. Medina Herrera, J. Turmo Borras, Joint classification of key-phrases and relations in electronic health documents, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018) co-located with 34nd SEPLN Conference (SEPLN 2018): Sevilla, Spain, September 18th, 2018, CEUR-WS. org, 2018, pp. 83–88.

[9] S. Santiso, A. Perez, A. Casillas, Exploring joint ab-lstm with embedded lemmas for adverse drug reaction discovery, IEEE journal of biomedical and health informatics (2018).

[10] S. Santiso González, Adverse drug reaction extraction on electronic health records written in spanish (2019).