TALP-UPC at eHealth-KD Challenge 2019

A Joint Model with Contextual Embeddings for Clinical Information Extraction

Salvador Medina and Jordi Turmo

Universitat Politècnica de Catalunya, Spain
Campus Nord, Carrer de Jordi Girona, 1, 3, 08034 Barcelona, Spain
Tel.: +123-45-678910
Fax: +123-45-678910
{smedina,turmo}@cs.upc.edu

Abstract. Most eHealth entity recognition and relation extraction models tackle the identification of entities and relations with independent specialized models. In this article, we show how a single combined model can exploit the correlation between these two tasks to improve the evaluation score of both, while reducing training and execution time. Our model uses both traditional part-of-speech tagging and dependency-parsing of the documents and state-of-the-art pre-trained Contextual Embeddings as input features. Furthermore, Long-Short Term Memory units are used to model close relationships between words while convolution filters are applied for farther dependencies. Our model was able to get the highest score in all three tasks of IberLEF2019's eHealth-KD competition[7]. This advantage was specially promising in the relation extraction tasks, in which it outperformed the second best model by a margin of 9.3% in F_1 Score.

Keywords: NERC · Relation Extraction · eHealth NLP · Contextual Embeddings

1 Introduction

This article describes the model presented by the *TALP* team for *IberLEF2019's eHealth-KD*[7] shared task, which includes the identification of relevant key-phrases and relations among them in Electronic Health (i.e., eHealth) documents written in Spanish. The task was divided in three scenarios: key-phrase identification and classification, relation extraction and full knowledge extraction. Our model outperformed the rest of competing models in all three scenarios.

IberLEF2019's eHealth-KD shared task supersedes and extends previous year's *Taller de Análisis Semántico en la SEPLN 2018's eHealth-KD (TASS-2018's eHealth-KD)*[4] shared task. There are, however, substantial differences between both tasks' classes[1] and evaluation metrics. Likewise, the task is inspired by previous competitions such as *Semeval-2017 Task 10: ScienceIE*[1].

The models presented for the aforementioned related task incorporate combinations of several techniques such as *Convolutional* or *Recurrent Neural Networks*, *Support Vector Machines*, *Conditional Random Fields* and even rule-based systems. Our team concurred to the key-phrase classification and relation extraction sub-tasks of *TASS-2018's eHealth-KD* with a joint *CNN*-based model[5], which ranked in first place for the relation extraction sub-task. The model did not support key-phrase recognition though, as it received pairs of key-phases as input. Our newly presented model overcomes this limitation by identifying key-phrases and all their related key-phrases at once.

Given the similarity to the aforementioned tasks, we decided to first try a model for *TASS-2018's eHealth-KD* data set and then used the model weights as an starting point. This idea of transferring some of the weights of a model trained for a different task (*transfer-learning*), has been extensively used in low-resource machine learning tasks such as image classification, text analysis, question answering and more[6][2].

2 Model

The model takes a document and a token's index and computes the boundaries and classes of the shortest key-phrase it belongs to, and the relations of every other entities' tokens to it. Hence, the model should be run for each token of the input document. This approach is inspired by attention-based translation models such as Transformer [8], in which the output is successively generated by running the model for one particular input token at a time.

The joint identification model's structure is visually described in Figure 1, and is composed of a set of shared layers and two independent output layers. Both output layers share the same structure, a fully connected layer and CRF; which respectively predict the target token's smallest entity sequence and each other token's relation to it. The core of the shared layer contains a recurrent layer composed by multiple bidirectional memory units (Either Gated Recurrent Units or Long-Short Term Memory units) followed by a convolution layer. The RNN and CNN's outputs are then fed to a fully connected layer with output dropout.

The recurrent and convolution layers allow for looking to both the local and global contexts of each input token. The local context is captured by the RNN Layer's output and the non-pooled convolution layer's output, which are concatenated for each time-step. The global context is captured by the max-pooled convolution layer's output. The global context information and the target token's local context information are added to all time-steps before being fed to the fully connected shared layer.

[1] Two additional key-phrase classes (*Predicate* and *Reference*) with their related relation classes (*in-time*, *in-place*, *in-context*, *domain* and *arg*) were added. Moreover, one relation class was removed (*property-of*) and others were added (*same-as*, *has-property*, *causes* and *entails*)

The final outputs are then generated by a Conditional Random Field (CRF) layer. Output CRF layers have proven to improve the capabilities of GRU and LSTM networks in low-resource sequence tagging tasks[3].

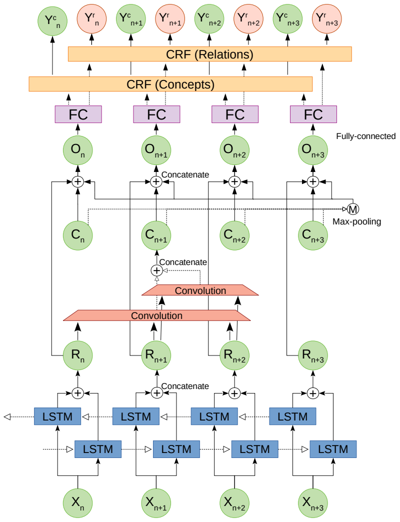
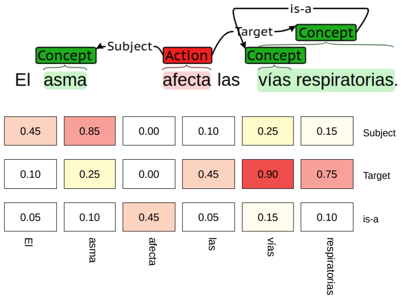


Fig. 1. Schematic architecture of the identification artificial neural network



	El	asma	afecta	las	vías	respiratorias
Subject	0.45	0.85	0.00	0.10	0.25	0.15
Target	0.10	0.25	0.00	0.45	0.90	0.75
is-a	0.05	0.10	0.45	0.05	0.15	0.10

Fig. 2. Visual representation of how relations are encoded by the network when *afecta* is the input token.

2.1 Input encoding and decoding

As described in Section 2, the model receives the sequence of tokens of a document and a token's index and outputs the bounds of the innermost key-phrase to which the token belongs. These bounds are encoded and decoded by assigning a Begin, Inside, Unitary and End tag to each token included in that key-phrase and Out to every other token (*BIOUE-tag*). Note that just one key-phrase is decoded for each token index. Consequently, in order to identify all key-phrases, the model has to be evaluated for every token.

We approach relation extraction at the level of tokens. Given a token, the list of relations' probabilities between the innermost entity to which the token belongs and each one of the tokens in the document is predicted. Note that for the source token, we only consider the innermost entity whereas for the target tokens we consider all parent entities. This restriction is imposed so that the encoded sequence is not ambiguous. Other alternatives such as source and target encoding were also considered but ultimately discarded as the increased decoding complexity did not yield improved results.

A visual representation of relations' probability predictions is shown in Figure 2. Relations are predicted from the target key-phrase if the aggregated score inside a key-phrase span surpasses a threshold. Only the key-phrase with the highest score is selected if multiple key-phrases overlap. A pseudo-code of the relation decoder is listed below.

Algorithm 1 Relation decoding algorithm

```
procedure DECODE($i_s, p^c, E, R$)    ▷ Relations from token $i_s$ with probabilities $p^c$ of class $c$
    $e_s \leftarrow$ innermost entity at index $i_s$ from $E$
    while not done do
        $\{e_t, p_t\} \leftarrow \{\phi, 0\}$
        for $e \leftarrow E$ do
            $p \leftarrow$ aggregate ${p^c}_i$ $\forall i \in$ bounds of $e$    ▷ Pre-defined probability aggregation function
            if $p > p_t \wedge p > p_{th}$ then    ▷ If probability is above the threshold
                $\{e_t, p_t\} \leftarrow \{e, p\}$    ▷ Probabilities in the span are set to 0
        if $e_t \neq \phi$ then
            $R_c \leftarrow R_c \bigcup \{\{e_s, e_t, c\}\}$    ▷ Add relation to the relation set
            ${p^c}_i \leftarrow 0$ $\forall i \in$ bounds of $e_t$
        else
            return $R$
```

2.2 Input features

In section 2 we describe how the model looks at sentences at the token level. The sentences are first tokenized, tagged and dependency-parsed by FreeLing.

We represent each token by a vector, which results from the concatenation of the features listed below:

- One-Hot encoding of the *category* and *type* fields of the token's Part-of-Speech Tag from FreeLing's tag-set.
- Normalized vector encoding the dependencies found in the path between the token and the target token (the one that is being decoded). It is computed by adding the one-hot encoding representation of the dependency class for each hop in the dependency path and normalizing the resulting vector, not considering its direction. For instance, the representation of the token "I" in "I eat fish" when the target token is "fish" would be a vector with $\sqrt{2}$ in the positions corresponding to "subj" (subject) and "cd" (direct complement); whereas for "eat" it would be a vector with a single 1 in the "cd" position.
- One-Hot encoding of the distance between the token and the target token.
- Contextualized word embedding of the token, computed by extracting the weights of the last layer of a multi-language, general-purpose pre-trained[2]

[2] We used the **BERT-Base, Multilingual Cased** model (104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters) from the authors' repository (https://github.com/google-research/bert)

Bidirectional Encoder Representations from Transformer model (BERT)[2]. No fine-tuning of the BERT model is done.

2.3 Pre-Training

IberLEF2019's eHealth-KD' training data-set is arguably small considering the number of classes and the variability in the examples (3818 concepts and 3503 relations). In order to prevent over-fitting and consequently increasing recall, we opted for using previous years' *TASS-2018's eHealth-KD*[4] training and testing data-sets as a pre-training, transfer-learning step. Transfer-learning in this case is straightforward, as the two tasks are very similar and our model already shows positive results for *TASS-2018's eHealth-KD*'s tasks, as shown in Section 3. However, since output classes for both concept recognition and relation extraction do not match *IberLEF2019's eHeath-KD*, the output layers' weights are discarded after the pre-training phase.

3 Evaluation and results

The presented model was evaluated for tuning of hyper-parameters for using the provided development data-set. For this, we used the evaluation metrics and scripts provided by the challenge's organizers[3]. For more information about the data-set please refer to the tasks' overview paper.

In addition to this evaluation, we also evaluate the model learned in the pre-training step using *TASS-2018's eHealth-KD*'s testing data-set. For the sake of comparison, we also use the author's evaluation metrics and scripts and contrast the results to the challenge's contestants[4].

Table 1 shows the relevant evaluation metrics of *TASS-2018's eHealth-KD*'s best-performing models for the three scenarios, compared to the presented model (*Joint-RCNN*). The joint identification model clearly outperforms our previous model (*talp*) in all metrics, also beating the rest of the participants in sub-tasks B and C. In line with the original task's results, *rriveraz*'s model shows impressive results and surpasses our model in black-boxed sub-task A by a margin of 4.1% in F_1 score, which ranks in second position.

Table 2 shows the shared task's final evaluation results. Our model ranked in first position for all three evaluation scenarios. The largest advantage resides in the relation extraction task, as it outscores the second runner by a margin of 9.6% in F_1 score. After the competition was closed, we found that our model was outputting invalid combinations of key-phrase classes and relation classes. *fix-relations* shows the evaluation of the fixed model.

[3] *IberLEF2019's eHealth-KD* data-set and evaluation scripts were downloaded from https://github.com/knowledge-learning/ehealthkd-2019.

[4] *TASS-2018's eHealth-KD* data-set and evaluation scripts were downloaded from https://github.com/TASS18-Task3/data.

	Scenario 1			Scenario 2			Scenario 3
Model	$F_1(A)$	$F_1(C)$	$\overline{F_1}(ABC)$	$Acc(B)$	$F_1(C)$	$\overline{F_1}(BC)$	$F_1(C)$
Joint-RCNN	82.3	**34.5**	72.8	94.3	**53.0**	**76.8**	**60.1**
talp	N/A	N/A	N/A	93.1	45.8	72.2	44.8
rriveraz	**87.2**	00.0	**75.7**	**95.9**	10.9	62.2	03.6
upf_upc	80.5	09.3	66.1	95.4	00.0	64.8	00.0

Table 1. Evaluation results comparing *TASS-2018's eHealth-KD* final results against the presented model for the 3 training scenarios and related test corpora.

	Scenario 1			Scenario 2			Scenario 3		
Model	P	R	F_1	P	R	F_1	P	R	F_1
Joint-RCNN	65.1	**62.9**	63.9	**80.7**	83.4	**82.0**	66.7	**59.2**	62.7
fix-relations	65.4	**62.9**	**64.1**	**80.7**	83.4	**82.0**	**67.3**	**59.2**	**63.0**
no-transfer	65.1	61.6	63.3	79.9	**83.9**	81.9	65.4	55.5	60.0
LASTUS-TALN(abravo)	**77.4**	46.6	58.2	80.0	83.4	81.7	17.1	35.2	23.0
IxaMed(iakesg)	69.0	37.6	48.7	65.7	71.1	68.3	52.0	37.5	43.6
NLP_UNED(lsi_uned)	65.6	47.0	54.7	**80.7**	70.8	75.4	62.4	46.7	53.4
coin_flipper(ncatala)	74.5	53.3	62.2	79.9	77.6	78.7	71.3	37.7	49.3
UH-MAJA-KD	56.4	48.0	51.9	80.0	83.2	81.6	43.1	43.7	43.4
VSP	45.5	40.6	42.9	51.3	58.5	54.7	58.9	42.4	49.3

Table 2. Final evaluation results of *IberLEF2019's eHealth-KD*. We also include the results when invalid relations are removed (*fix-relations*) and when no transfer-learning is used.

4 Conclusion

In this article we have presented a joint concept and relation identification and classification model that exploits the mutual information between the entities and their relations by using a single network that looks at both local and global textual features. This newly presented model significantly outperforms all other competing models in both *TASS-2018's eHealth-KD* and *IberLEF2019's eHealth-KD* shared tasks.

We hypothesize that the fact that the model shares both structure and weights allows the full model to more accurately capture the synergy between the two tasks and hence provide better precision and recall than traditional step-by-step models. In spite of this, the task is still not fully figured out, and we argue that further experimentation should be done in this line of research. We identify three noteworthy challenges:

- Entities and relations are formally very different. The first is usually encoded as a sequence and the second as a set of one-to-one labels. These two representations are difficult to combine, so more appropriate encoding is required to take full advantage of joint models.
- Optimization functions are also hard to define, as they have to balance both entity recognition and relation extraction, taking into account the different amounts and difficulty of instances of both tasks.

– Depending on the model's structure, the optimization of the model's hidden layer's parameters for different outputs may be mutually opposite. These hidden layers should be designed to promote reciprocal collaboration between both objective functions.

5 Acknowledgements

Funding: This contribution has been partially funded by the Spanish Ministry of Economy (MINECO) and the European Union (TIN2016-77820-C3-3-R and AEI/ FEDER,UE).

References

1. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. arXiv preprint arXiv:1704.02853 (2017)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
4. Martínez Cámara, E., Almeida Cruz, Y., Díaz Galiano, M.C., Estévez-Velarde, S., García Cumbreras, M.Á., García Vega, M., Gutiérrez, Y., Montejo Ráez, A., Montoyo, A., Muñoz, R., et al.: Overview of tass 2018: Opinions, health and emotions (2018)
5. Medina, S., Turmo, J.: Joint classification of key-phrases and relations in electronic health documents. Proceedings of TASS **2172** (2018)
6. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
7. 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)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)