TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 83-88

Clasificación conjunta de frases clave y sus relaciones en documentos electrónicos de salud en español

Joint classification of Key-Phrases and Relations in Electronic Health Documents

Salvador Medina[1], **Jordi Turmo**[1]
[1]TALP Research Center - Universitat Politècnica de Catalunya
smedina@cs.upc.edu y turmo@cs.upc.edu

Abstract: This paper describes the approach presented by the *TALP* team for *Task 3* of *TASS-2018*: a convolutional neural network to jointly deal with classification of key-phrases and relationships in eHealth documents written in Spanish. The results obtained are promising as we ranked in first place in scenarios *2* and *3*.
Keywords: Relation extraction, joint classification, key-phrase classification, convolutional nerural networks.

Resumen: Este artículo describe el método presentado por el equipo *TALP* en la *Tarea 3* de *TASS-2018*: una red neuronal convolucional para tratar conjuntamente la clasificación de frases clave y sus relaciones en documentos de salud escritos en español. La propuesta quedó en primera posición en los escenarios *2* y *3*.
Palabras clave: Extracción de relaciones, clasificación conjunta, clasificación de frases clave, redes neuronales convolucionales.

1 Introduction

This article describes the model presented by the *TALP Team* for solving B and C sub-tasks of *Task 3* in the *Taller de Análisis Semántico en la SEPLN 2018* (TASS-2018) (Martínez-Cámara et al., 2018). TASS-2018's *Task 3* consists in recogniting and classifying key-phrases as well as identifying the relationships between them in Electronic Health Documents (i.e., eHealth documents) written in Spanish. *Task 3* is divided in sub-tasks *A*, *B* and *C*, which correspond to key-phrase boundary recognition, key-phrase classification and relation detection, respectively.

In this task, a key-phrase stands for any sub-phrase included in eHealth documents that is relevant from the clinical viewpoint and can be classified into *Concept* or *Action*. The relationships between them are classified into 6 types: 4 of them are between *Concepts* (*is-a, part-of, property-of* and *same-as*) while the rest are between an *Action* and another key-phrase (*subject* and *target*). The proposed task is similar to previous competitions such as *Semeval-2017 Task 10: ScienceIE*(Gonzalez-Hernandez et al., 2017), but uses a simpler categorization for key-phrases while considering a broader range of possible relationships.

Participants in the *Semeval-2017 Task 10: ScienceIE*(Gonzalez-Hernandez et al., 2017) shared task considered a large plethora of supervised learning models, ranging from *Convolutional* or *Recurrent Neural Networks* to *Support Vector Machines*, *Conditional Random Fields* and even rule-based systems, often applying radically different models for each one of the three sub-tasks. Note that some of the teams did not participate in all three sub-tasks, this was in fact the case for the winners of sub-tasks *BC* (MayoNLP (Liu et al., 2017)) and *C* (MIT (Lee, Dernoncourt, and Szolovits, 2017)).

1.1 Joint classification of key-phrases and relationships

In our implementation we tackle both the classification of key-phrases and the identification of the relationships between them, corresponding to scenarios 2 and 3 of TASS 2018's Task 3, as a single task. The intuition behind this decision is that the categories of key-phrases are influenced by the relationships they hold with other key-phrases. For instance, a verb is an *Action* key-phrase if and only if it relates to another *Action* or *Concept* by either being the *subject* or *target*, which means that sometimes phrases are

ISSN 1613-0073

not key-phrases by themselves but when they relate to other phrases.

2 Implementation

The architecture that we propose is represented in Figure 1 and consists of a two-layer Convolutional Neural Network (CNN) which takes a vectorial representation of the documents and the position of two key-phrases as input and applies several convolution filters for window sizes from 1 to 4 tokens. The outputs of these filters are then max-pooled and fed to a fully connected output layer, which has two outputs for the given key-phrase pairwise: the probabilities of either key-phases for being *Action* or *Concept*, and the probabilities of the pairwise for being each possible kind of relationship, including *"other"* for no relationship.

At first glance, our architecture is similar to the one proposed by the *MIT* team for the *ScienceIE* task, which also consists of a CNN using word-embedding, relative position and *PoS-tags* as input features. However, it presents some noticeable differences. First of all, our architecture jointly tackles sub-tasks *B* and *C*. For this reason, it does not take the key-phrase category as an input and has two additional outputs which hold the source and destination key-phrases' classes. Moreover, we optimize all three outputs at the same time and consequently our loss function is designed to reflect this.

2.1 Layout of the network and parameter optimization

Artificial Neural Networks (ANN) and more specifically CNNs have proven to be capable of jointly identifying entities and relationships in various kinds of textual documents and relation extraction tasks, as it has been demonstrated in recent articles such as (Singh et al., 2013), (Shickel et al., 2017) and (Li et al., 2017). This joint identification takes advantage of the correlation that exists between linked entities aiming to provide better results for both named entity recognition classification and relation extraction tasks respect to a classical two-step system.

The loss function used by the parameter optimization algorithm is computed independently for the three outputs using soft-max cross-entropy, as classes are mutually exclusive for a single output, and is then combined by just adding the three losses. By adopting these three independent loss functions we can take profit of the fact that output classes for a single output are mutually exclusive and make their probabilities add up to one, independently of the other two outputs.

As for the optimization algorithm, we use *TensorFlow's Adam* optimizer with a learning rate of 0,005. The system was trained in batches of 128 sentences which were previously stripped to up to 50 tokens and padded. We also apply a *dropout* rate of 0,5 to the fully-connected output layer for regularization purposes. The parameter optimization process is stopped either when the average *loss* in the *development* corpus remains flat for 1000 iterations or when $1e5$ iterations have been run.

2.2 Input parameters and encoding

In order to come up with a manageable vectorial representation of the input sentences, they are previously tokenized using *FreeLing*'s with *multi-word* and *quantity* detection as well as *Named Entity Classification* (NEC) modules disabled, so that multiple tokens are never joined together. These tokens are then passed through a lookup table containing their pre-computed word-embeddings vectors, which are then joined one-hot encodings of the relative positions respect to the target source and destination key-phrases and their respective *Part-of-Speech* (PoS) tag determined by *FreeLing*'s *PoS-Tagger* module. A more detailed description of the input properties is listed below:

- **Word-Embedding**: 300-dimension vectorial representation of words in *word2vec* format. We used the pre-trained general-purpose vectors from SBWCE (Cardellino, 2016), trained from multiple sources.
- **Distance to source or destination key-phrase**: One-hot encoding of the distance respect to the key-phrases. We consider two types of distances: absolute distance in terms of the number of tokens between each token and key-phrase and number of arcs in the dependency tree between each token and key-phrase, not taking into account the dependency class. The latter option was finally selected as it yielded better results in the

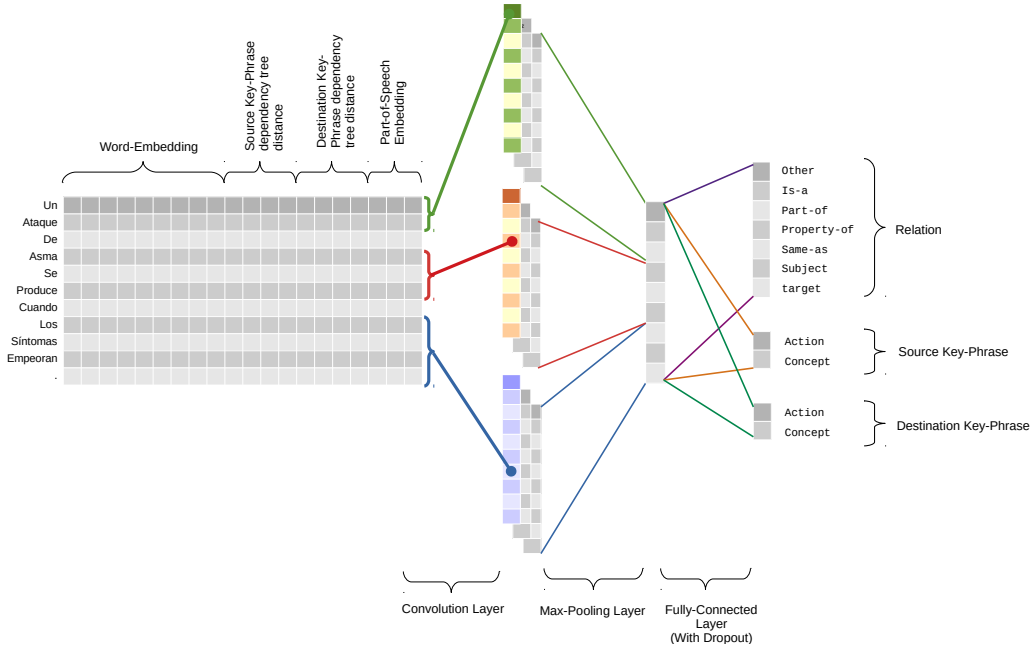

Figura 1: Layout of the proposed Convolutional Neural Network architecture

validation corpus.

- **Part-of-Speech tag**: One-hot encoding of the token's *PoS-tag* determined by *FreeLing*. For simplicity, we only contemplate the *category* and *type* positions of the *PoS-tag*, hence reducing the number of different tags to 33.

2.3 Data augmentation

Relation extraction is a difficult task and usually requires big amounts of training examples in order to be able to correctly generalize the relationship classes. This is specially so for ANN based models, which can be prone to over-fitting. The training corpus that was provided for the *TASS-2018* is very limited and classes are considerably unbalanced. To give an example, it only includes 30 instances of class *same-as* compared to the 911 examples provided for class *target*.

Because of this, we evaluated several data augmentation alternatives, which added slight modifications of the original training instances to the training set. These modifications included replacing some or all keyphrases by their class name or other keyphrases in the training corpus, or trimming the sentences removing some of their tokens. The alternative that worked best in the validation corpus and was used in the final model was to trim the context before and after the relationships to 1 and 3 tokens. For instance, in sentence "*Un ataque de asma se produce cuando los síntomas empeoran.*", the *target* relationship between *produce* and *ataque de asma*, adds "*Un ataque de asma se produce cuando*" and "*Un ataque de asma se produce cuando los síntomas*", as well as the full sentence.

3 Results

As it can be seen in Table 1, our model scored first in the evaluation scenarios 2 and 3, which evaluate sub-tasks *BC* and *C* respectively. As it was mentioned in Section 1, our system was designed for sub-tasks *B* and *C*, so no submission was sent for scenario 1, which also evaluates sub-task *A*. In terms of the individual sub-tasks, our system raked first for sub-task *C* but was outperformed by *rriveraz*'s model in sub-task *B*.

3.1 Analysis of errors

In this Subsection, we analyze the errors made by our model in *Scenarios 2* and *3*. Tables 2 and 3 show the confusion matrices for subtasks *B* and *C* in the evaluation of *Scenario 2*. Results for sub-task *C* in *Scenario 3* are analogous to *Scenario 2* and are not shown, as our model does not make use of the additional information given in *Scenario 2*.

3.1.1 Sub-Task B

Table 2 shows the confusion matrix for subtask *B* in *Scenario 2*. Our model achieves si-

Scenario	plubeda	rriveraz	upf_upc	VSP	baseline	Marcelo	TALP
1	0.71	**0.744**	0.681	0.297	0.566	0.181	N/A*
2	0.674	0.648	0.622	0.275	0.577	0.255	**0.722**
3	N/A*	N/A*	0.036	0.42	0.107	0.018	**0.448**
avg	0.461	**0.464**	0.446	0.331	0.417	0.151	0.39

Tabla 1: Micro-averaged $F1$ score for evaluation scenarios 1 to 3 and global average. *TALP* column shows our model's score. **N/A***: Not Available, counted as 0 in the average score.

true\pred.	Concept	Action	recall
Concept	**432**	7	0.984
Action	34	**120**	0.779
precision	0.927	0.945	$Acc = 0{,}931$

Tabla 2: Confusion matrix of our model's predictions for sub-task B in scenario 2.

milar *precision* for classes *Concept* and *Action*, but *recall* for the latter is 0.205 smaller. This is not only due to the fact that classes are unbalanced (439 and 154 instances of classes *Concept* and *Action* respectively), but also to other reasons listed below:

- The *Shared-Task's* description defines *Actions* as a particular kind of *Concept* that modifies another concept. Consequently, in some cases, the same phrase can either be an *Action* or a *Concept* depending on whether or not the modified *Concept* is explicitly mentioned. As an illustration, the noun *causa* (cause) is labeled as a *Concept* in sentence "*El tratamiento depende de la causa.*" (The threatment depends on the cause.). However, in sentence "*Es una causa común de sordera.*" (It is a common cause of deafness.), it is labeled as *Action*, as it is supposed to modify *sordera* (deafness).

- Errors which were in part due to incorrect dependency parsing or *PoS-tagging* by *FreeLing*, specially when verbs are identified as nouns.

  For example, the noun *oído* (ear) was identified as a verb in sentence "*Suele afectar sólo un oído.*" (It usually affects just one ear.) by *FreeLing*, which lead to confusion. Similarly, in sentence "*Esto causa una acumulación de sustancias grasosas en el bazo, hígado, pulmones, huesos y, a veces, en el cerebro.* (This causes an accumulation of fatty substances in the arm, liver, lungs, bones and, sometimes, the brain.), *causa* (causes) is incorrectly labeled as a noun.

- Other instances where it is difficult to determine the label assigned to the entity, even for us, as they do not seem to correspond to any of the criteria exposed in the description.

  For instance, in sentence "*Si usted ya tiene diabetes, el mejor momento para controlar su diabetes es antes de quedar embarazada.*" (If you alredy have diabetes, the best moment to control your diabetes is before getting pregnant.), the adverb *antes* (before) is labeled as *Action* and is related to *controlar* (control, keep) and *quedar* (get, become) as *subject* and *target* respectively.

  On the other hand, in sentence "*La exposición al arsénico puede causar muchos problemas de salud.*" (The exposition to arsenic can cause several health problems), the noun *exposición* (exposition) is labeled as *Concept*, while we understand it as the *Action* of being exposed to something. This is not coherent to other instances such as "*No se conoce la causa de la destrucción celular.*" (The cause of cell destruction is not known.), where *destruction* is labeled as *Action* - the *Action* of being destroyed.

3.1.2 Sub-Task C

Table 3 shows the confusion matrix for subtask C in *Scenario 2*. Class *other* is used for all pairs of entities that have no specified relationship in the training set, making it the most frequent class in the training set. The model seems to prioritize *precision* over *recall*, which vary from class to class. Recall and precision for *same-as*, although 0,000, are not significant, as just one instance is present in the test set. The list below describes multiple reasons for the most common errors produced by our model:

true\pred.	other	is-a	part-of	property-of	same-as	subject	target	recall
other	**0**	0	0	0	0	0	0	0.000
is-a	31	**58**	1	2	0	0	0	0.630
part-of	26	2	**5**	0	0	0	0	0.152
property-of	34	0	3	**18**	0	0	3	0.310
same-as	0	1	0	0	**0**	0	0	0.000
subject	65	0	0	2	0	**42**	8	0.359
target	84	0	1	7	0	12	**91**	0.467
precision	0.000	0.951	0.500	0.621	0.000	0.778	0.892	$F_1 = 0{,}431$

Tabla 3: Confusion matrix, precision and recall of our model's predictions for sub-task C in scenario 2. F_1 is micro-averaged for all classes.

- Annotated instances in both training and test sets are unbalanced. Relationship counts in the training set range from 991 for *target* and 693 for *subject* to 149 and 30 for *part-of* and *same-as* respectively. What is more, the auxiliary class *other* amounts to 16478 instances. More instances for the two less common classes seem to be required, as the model achieves much lower *recall* and *precision* than the most common ones.

- Relationships *subject* and *target* are prone to be mutually confused, specially for reflexive or passive verbs, and labeling is not always coherent. For example, in "*Algunos sarpullidos se desarrollan inmediatamente.*" (Some skin rashes are developed immediately.), *sarpullidos* (skin rashes) is *subject* of *se desarrollan* (are developed). However, in sentence "*Existen muchas razones para someterse a una cirugía.*" (There are several reasons to have surgery.), *razones* (reasons) is *target* of *existen* (there are).

- Multi-label relationships were not considered by our model, as we did not realize instances such as *Durante cada trimestre, el feto crece y se desarrolla.* (During each quarter, the fetus grows and develops.), where the relationships between *feto* (fetus) and *crece* (grows), and similarly between *feto* and *se desarrolla* (develops), are both *target* and *subject*.

- Errors due to incorrect parsing by *FreeLing*, which were already discussed in Section 3.1.1.

4 Conclusions and future work

In this paper, we have described the model presented by the *TALP* team for *Task 3* of *TASS-2018*. In addition we have presented some reasons for our model to wrongly classify key-phrases and relationships.

The results achieved by our model when compared to the rest of the challengers prove that a model that jointly classifies entities and relations can outperform traditional two-step systems in tasks where some entity classes are defined by the relationships they hold with others. There is however a big room for improvement, specially in the relation extraction task, mainly due to the increased complexity and the limited amount of examples available in the training set.

Our model was designed to solve the key-phrase classification and relation extraction tasks, leaving the key-phrase recognition as future work, as our focus was joint recognition and we did not have enough time to design and optimize a single model that could tackle all three tasks. We are committed to continue this line of investigation and extend the architecture so that it is also able to determine the key-phrases' boundaries.

Additionally, there are several improvements that could be applied to the current model, that we realized after analyzing the currently most common errors. To begin with, our model should allow for *multi-label* relation extraction, as mentioned in Section 3.1.2. Second, more syntactical features could be added, by for instance providing a complete and more appropriate encoding of the *PoS-tags* or by including not only the dependency tree distances but also the types.

Acknowledgments

This works has been partially funded by the Spanish Goverment and by the European Union through GRAPHMED project (TIN2016-77820-C3-3-R and AEI/FEDER,UE.)

References

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Gonzalez-Hernandez, G., A. Sarker, K. O'Connor, and G. Savova. 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of medical informatics*, 26(01):214–227.

Lee, J. Y., F. Dernoncourt, and P. Szolovits. 2017. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. *arXiv preprint arXiv:1704.01523*.

Li, F., M. Zhang, G. Fu, and D. Ji. 2017. A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics*, 18(1):198.

Liu, S., F. Shen, V. Chaudhary, and H. Liu. 2017. Mayonlp at semeval 2017 task 10: Word embedding distance pattern for keyphrase classification in scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 956–960.

Martínez-Cámara, E., Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, and J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. In E. Martínez-Cámara, Y. Almeida Cruz, M. C. Díaz-Galiano, S. Estévez Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez Vázquez, A. Montejo Ráez, A. Montoyo Guijarro, R. Muñoz Guillena, A. Piad Morffis, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Shickel, B., P. J. Tighe, A. Bihorac, and P. Rashidi. 2017. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*.

Singh, S., S. Riedel, B. Martin, J. Zheng, and A. McCallum. 2013. Joint inference of entities, relations, and coreference. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 1–6. ACM.