TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 89-95

TASS2018: Medical knowledge discovery by combining terminology extraction techniques with machine learning classification

TASS2018: Obtención de conocimiento médico mediante combinación de técnicas de extracción de terminologías y clasificación basada en aprendizaje automático

Jorge Vivaldi Palatresi[1], **Horacio Rodríguez Hontoria**[2]
[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Universidad Politécnica de Catalunya, Barcelona, Spain
jorge.vivaldi@upf.edu horacio@lsi.upc.edu

Resumen: En este artículo presentamos la aproximación seguida por el equipo UPF-UPC en la tarea TASS 2018 Task 3 challenge. Nuestra aproximación puede calificarse, de acuerdo a los códigos propuestos por la organización, como H-KB-S, ya que utiliza métodos basados en conocimiento y aprendizaje supervisado. El pipeline utilizado incluye: i) Un pre-proceso standard de los documentos usando *Freeling* (etiquetado morfosintáctico y análisis de dependencias); ii) El uso de una herramienta de etiquetado sequencial basada en CRF para completar las subtareas A (identificación de frases) y B (clasificación de frases), y iii) El abordaje de la subtarea C (extracción de relaciones semánticas) usando una aproximación híbrida que integra dos classificadores basados en *Regresión Logística*, y dos extractores léxicos para pares entity/entity y relaciones *is-a* y *same-as*.
Palabras clave: obtención de conocimiento médico, terminología médica, identificación de relations semánticas

Abstract: In this paper we present the procedure followed to complete the run submitted by the UPF-UPC team to the TASS 2018 Task 3 challenge. Such procedure may be classified, according the organization's codes, as H-KB-S as it takes profit from a knowledge based methodology as well as some supervised methods. Our pipeline includes: i) A standard pre-process of the documents using *Freeling* tool suite (POS tagging and dependency parsing); ii) Use of a CRF sequence labelling tool for completing both subtasks A (key phrase identification) and B (key phrase classification), and iii) Facing the subtask C (setting semantic relationships) by using a hybrid approach that uses two *Logistic Regression* classifiers, followed by lexical shallow relation extractors for entity/entity pairs related by *is-a* and *same-as* relations.
Keywords: health knowledge discovery, terminology extraction, identification of semantic relations

1 Introduction

Text mining and natural language processing (NLP) techniques have been applied to the biomedical domain for a long time. Automatic identification of relevant terms in medical texts (research and educational material as well as medical reports) and how they relate each other represent a major improvement for indexing and for search tools. Its results are useful for research as well as clinical and educational purposes.

In this paper we present two tools for facing the tasks proposed in the TASS-2018-Task 3 challenge: eHealth Knowledge Discovery. The first one is a term extraction tool for finding terminologically relevant substrings (key phrases) and classify them in one of the two classes proposed by the organization: Concept or Action. The second one is dedicated to recognize those semantic relationships chosen by the Organization between the recognized entities.

ISSN 1613-0073

2 TASS 2018 Task 3

2.1 Description of TASS 2018 Task 3

The corpus for this competition was compiled by sampling XML files produced by the National Library of Medicine, the world's largest medical library. It brings information about diseases, conditions, and wellness issues in understandable language. The full collection is available at https://medlineplus.gov/xml.html.

Given a collection of eHealth documents written in Spanish, TASS 2018 Task 3 has been conceived as a three task for three different scenarios. Each task may be described as follows:

A) To identify all the key phrases per document;

B) To assign a label (*Concept* or *Action*) to each of the key phrases;

C) to link the entities detected and labelled in each document through the following semantic relationships:

   (a) Concept-Concept: *is-a*, *part-of*, *property-of* and *same-as*;

   (b) Action-Concept: *subject* and *target*.

The output of each task is the input of the next one. Proceeding in this way the organization has considered the following three evaluation scenarios:

1. Only plain text is given (Subtasks A, B and C must be completed);
2. Plain text and manually annotated key phrase boundaries are given (Subtasks B and C must be completed);
3. Plain text with manually annotated key phrases and their types are given (only Subtask C must be completed).

More details about the tasks and scenarios may be obtained through the web site of the TASS-2018 competition[1] and the overview paper (Martínez-Cámara et al., 2018).

2.2 Our approach to TASS 2018 Task 3

After downloading the documents we have processed them using *Freeling* tool suite[2], (Padró and Stanilovsky, 2012). We have used basically tokenization, and POS tagging for subtasks A and B and EWN[3] tagging and dependency and constituency parsing for subtask C. In this section we present some details about the full system designed for these tasks. Figure 1 shows the overall scheme. It presents the main modules and its interconnection to complete the full task proposed in this competition.

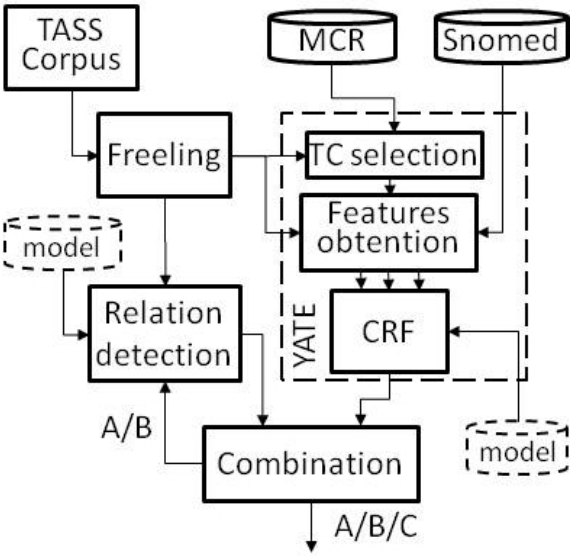

Figure 1: Full system architecture.

2.3 Subtasks A & B: key phrase identification and classification

For subtasks A and B we proceeded jointly using a single tool that is able to select the set of nominal term candidates, *TC*, included in the text under analysis. This approach is based in *YATE*, an in house term extractor that has been tuned for treating medical text. See (Vivaldi, 2001) and (Vivaldi and Rodríguez, 2010) for a full description.

Term extraction can be seen as semantic annotation task because it provides machine-readable information based on meaning. The way to attack the problem varies according the available resources for each language. Some languages (mainly English) disposes of lexical resources (like ontologies and/or term repositories) that can be used for reference while other languages have to identify term within text using other procedures that include linguistic/statistical strategies.

[1] http://www.sepln.org/workshops/tass/2018/task-3/

[2] http://nlp.lsi.upc.edu/freeling/

[3] Freeling can provide the set of possible synsets for each token when a WordNet is available. In the case of Spanish EuroWordNet is used.

YATE is a hybrid system whose keypoints are: i) the combination of heterogeneous detection strategies and ii) the use semantic knowledge in such strategies. Initially, it used the lexical ontology EuroWordNet (Vossen, 2004) (EWN) and since recently its evolution: MCR 3.0 (Gonzalez-Agirre, Laparra, and Rigau, 2012). In order to obtain domain terms, we mark on this resource some *domain borders*[4].

After applying standard linguistic analysis procedures, it starts by extracting a basic list of TCs. Such candidates are then analysed using a collection of heterogeneous methods.

A first method to evaluate the termhood of any candidate is its *domain coefficient*. It is calculated using the above mentioned *domain borders* and indicates in which degree a given *TC* belong to the domain of interest (medicine in this case).

Other methods included in *YATE* are based on statistical information, context information and the result to decompose a *TC* in its graeco-latin components. In the original tool all these informations were combined using a voting scheme or a boosting algorithm. For this competition, we decided to add information from Snomed-CT [5] and also, due to its specific requirements (analysing every mention of each nominal or verbal TC), we modify the original term extraction tool and chose to use a CRF classifier[6] for combining all the available information and to predict the BIO tag[7] for each token. In order to select which pieces of information are given as an input to the model a template has to be built for each *TC*. It allows to take into account both the features associated to the target token as well as the ones of its neighbours. The set of features chosen for each word in this task was the following:

- The lemmas of the target token and of the tokens appearing in a size-2 window around it;
- The reduced POS tag[8] of the target token and the tokens appearing in a size-2 window around it;
- The domain border to each of the tokens of the string detected as a *TC*;
- The main class for those *TC* included in Snomed-CT[9]. This information is applied to all the tokens of the string detected as a *TC*;
- The first/last three letters of each token.

2.4 Subtask C: Setting semantic relationships

For facing the subtask C we learned two multi-class classifiers, one for action/concept relations and the other for concept/concept relations. We used a simple *LR* Logistic Regression model[10] for both tasks with the same set of features, detailed below, but for learning two different classifiers. For defining the feature set we performed an initial learning process from the training documents consisting on the following steps:

- Collecting the whole set of correct entities. Computing the *tf*idf* and sorting the collection by descending *tf*idf* weight.
- Decomposing the multi-word terms in the previous collection into atomic components. Computing also their *tf*idf* weight and sorting them accordingly.
- Extracting the shapes of all the correct entities. We consider two types of shapes, long and short, the long shape simply maps the characters occurring in the term into a set of tags[11] while the short shape groups together sequences of identical tags. For instance, for the term "DM-2" the long shape is "AA*0" and the short one is "A*0". We also obtained an histogram of the length of entities in tokens, in order to constraint the length of the generated candidates.

[4] A domain border is defined as an EWN synset likely belonging itself and its descendants to such domain (eg. *disease*, *bodypart* and *medical_procedure* among many others).

[5] A comprehensive and well known clinical health terminology available for several languages including Spanish, https://www.snomed.org/

[6] CRF++, https://taku910.github.io/crfpp/

[7] BIO is a popular way of tagging tokens for detecting useful sequences, B stands for beginning of a sequence, I for inside it and O for out of it.

[8] The first character of the label.

[9] Snomed-ct is organized as a tangled taxonomy with 19 top classes. We have used these top classes as tagset.

[10] using Scikit-learn package.

[11] "A" stands for upper case letter, "a" for lower case, "0" for number, " " for the space, and "*" for other characters.

- Collecting all the labels occurring in the dependency trees of all the sentences in the training documents.
- Collecting histograms of the POS appearing as initial, middle, and ending tokens in valid multi-word terms, and those appearing in single-word terms. From these collections those POS occurring under a threshold were removed. For instance, in the initial set (resulting in 26 POS) the most frequent POS was "NCMS000"[12], that appears 90 times, in the middle set (18 POS) "DA0FS0"[13] occurred 20 times, in the single set (65 POS) "NCFS000" occurred 316 times, and in the ending set (10 POS), "NCFS000"[14] occurred 77 times.

It is worth noting that all these collections have been extracted independently for the two settings, Concept-Concept and Action-Concept. Once obtained these collections we performed a feature selection process using the development corpus, looking for different set of features (which we name "configurations"). The most reliable configuration was the following:

- A vector of the 1,000 most relevant lemmas (using *tf*idf* weigth) for both entities in the relation[15].
- A vector of the 200 most relevant atomic words (using *tf*idf* weigth) for both entities in the relation.
- A vector of the labels of the path between the two entities (is existing) in the dependency tree of the involved sentence. 32 labels occurred in the training set in the Concept-Concept case.
- A vector of the long word shapes for both entities. 51 different long word shapes were detected in the training set in the Concept-Concept case.
- A vector of the short word shapes for both entities. 13 different short word shapes were detected in the training set in the Concept-Concept case.
- The distance in tokens between the two entities.
- The distance in characters between the two entities.
- The length in tokens of both entities up to a maximum of 5 tokens.
- Whether the entities are simple or complex.

As samples for learning and testing we consider as positive examples all the pairs of entities from subtask B occurring in the same sentence and the text between them. For the first classifier one of the entities has to be an action and the other a concept, for the second one both entities have to be concepts. For negative examples we used pairs of entities so that only one element of the pair occurred as an entity in subtask B and the other should satisfy the POS constraints obtained above.

Besides these two classifiers we have applied two other simpler lexical relation extractors for some specific relations:

- We looked in each sentence for the co-occurrence of two entities being one of them an acronym and the other its long form. If it was the case the two entities are tagged as related by a *same_as* relation, (Montalvo et al., 2017).
- We looked for pairs of compatible entities where different degrees of compatibility were considered: equal word form, equal lemma, EWN synsets overlapping, approximate string matching, etc. In the case of compatibility *same_as* or *is_a* relations are set. The latter is set for the case of being one concept an hyponym of the other in EWN or when one concept is a prefix of the other.

3 Experimentation

In order to adjust the behaviour of our system to the specificities of this competition we use the package *develop* also provided by the organization. In this context we adjust some parameters to the characteristics of the text to be processed. In particular, for subtasks A and B, we reviewed the set of domain borders already defined adding three new ones.

For the task C a summary of the features used by the Concept-Concept logistic classifier is shown in Table 1. Each candidate is represented as a vector of float in a 2,578

[12] Common noun, masculine, singular
[13] Determiner article, feminine, singular
[14] Common noun, feminine, singular
[15] All the feature vectors are used as Boolean indicators.

dimensional space. Note that some of the features, as the dependency labels or the distances are applied to the pair, while others to both entities in the pair.

In our setting we generated 1,700 examples, from which 994 positives. Figures for the Action-Concept classifier are similar.

Table 1: Features used for the logistic classifier Concept-Concept

Feature	**Quantity**
lemmas	2,000
atomic	400
labels DT	36
long shapes	102
short shapes	26
length	10
distance tk	1
distance ch	1
is simple	2
Total	**2,578**

We also performed some experimentation in obtaining some *is-a* relation for multiword terms. Consider for example the TC *síndrome de Marfan*, if the nucleus of this term (*síndrome*) has been validated for *YATE* it is reasonable to consider that the following semantic relation exists: *síndrome de Marfan* →*is-a* →*síndrome*. This procedure reach a precision of 66 %.

4 Results

The evaluation results obtained by the above described system and delivered to the organization for the evaluation are quantified in Table 2. This table shows both of our results as well as the baseline and best scores[16] [17]. As can be observed in Table 2 our results for sub-tasks A and B are at least acceptable as they are above the baseline. Overall we were ranged on the third position, with a global score of 0.446 (0.464 and 0.461 for the first and second teams). Unfortunately such results for sub-task C in two of the scenarios are bellow such baseline and this fact does not satisfy our expectation. For such reason we decided to revise in depth our procedures and algorithm for such task and we found a malfunction for this task. After solving such bug and performing some minor improvement for sub-tasks A and B, we run again all our system and evaluate the new results but using in this case the script provided by the Organization. The results obtained after correcting our procedures are shown in Table 3. Such table shows a clear improvement in the results obtained for Task C.

5 Discussion

In examining the text to be analysed we found some sentences whose inclusion in a health related corpus is not clear as they seems to be quite out of the domain. Example 1 shows a clear example of this kind of sentences.

(1) El CO se encuentra en el humo de la combustión, como lo es el expulsado por automóviles y camiones, candelabros, estufas, fogones de gas y sistemas de calefacción.

In this example, the annotator tags Concepts like "CO", "humo" and "fogones" among others. Also it was tagged an *is-a* relation involving "humo de la combustión" and "humo". It is not clear the reason that such Concepts and Relations are considered relevant in a Spanish health document. The consequence is that our term extractor identify such units but does not validated them as relevant in the domain. As a matter of fact, most of the evaluated as *missing* have similar characteristics (such as: "proveedor", "insecticidas", "pintura", ...).

Training phase has been completed using only those files provided by the organization. As *YATE* only obtain nominal terms. We rely in the training phase for obtaining Action terms.

In order to show the behaviour of the two main parts of system (*YATE* and semantic relationship detector) we present some additional information about such modules.

Table 4 shows the amount of different term candidates that has been validated and discarded for the text provided for scenario 1. It also shows the details for each term candidate analyser. It should be noted that a single candidate may be successfully evaluated by more than one analyser (ex. "botulismo", "diuréticos" and "psoriasis" among others).

[16] Note that the best score for each scenario correspond to the team with the best global result for each scenario. It means that it considers all subtasks for such scenario. This explains, for instance, that for the task C in scenario 1 the best scores are set to zero.

[17] Detailed results may be checked at the web site of the TASS-2018 competition. See footnote [1].

Table 2: Official evaluation results (ours plus baseline plus the best for each scenario)

Task	Metric	Scenario 1			Scenario 2			Scenario 3		
		Base.	**Best**	**Ours**	**Base.**	**Best**	**Ours**	**Base.**	**Best**	**Ours**
A	**Precision**	0.673	0.862	0.862	–	–	–	–	–	–
	Recall	0.536	0.882	0.755	–	–	–	–	–	–
	F1	0.597	0.872	0.806	–	–	–	–	–	–
B	**Accuracy**	0.932	0.959	0.945	0.774	0.931	0.954	–	–	–
C	**Precision**	0.262	0	0.18	0.676	0.487	0.431	0.714	0.506	0.263
	Recall	0.022	0	0.063	0.093	0.431	0.062	0.058	0.402	0.019
	F1	0.041	0	0.093	0.163	0.458	0.109	0.107	0.448	0.036

Table 3: Unofficial evaluation of our results for all scenarios and tasks

Task	Information	Scenario		
		1	**2**	**3**
A	Precision	0.91	–	–
	Recall	0.79	–	–
	F1	0.85	–	–
B	Precision	0.86	0.85	–
	Recall	0.76	0.76	–
	F1	0.81	0.80	–
C	Precision	0.42	0.43	0.39
	Recall	0.20	0.21	0.18
	F1	0.27	0.28	0.25

Table 4: Behaviour of YATE in Scenario1

Validated by	Quantity
... domain coefficient	55
... MCR database	137
... graeco-latins formants	19
... Snomed CT	181
Total accepted	212
Total discarded	59

Table 5 shows the amount of different relation candidates. We include the candidates generated by our *LR* classifiers (both the action-concept and the concept-concept ones) and the candidates generated by our lexical extractors. It is worth noting that no acronym/expansion pair was found in the test dataset, and, so, only one lexical extractor is included.

Table 5: Labels assigned by the relation extractor

Classifier	Label	Quant.
LR action-concept	subject	49
LR action-concept	target	97
LR concept-concept	same-as	1
LR concept-concept	is-a	24
LR concept-concept	part-of	11
LR concept-concept	property-of	22
lexical classifier	same-as	105
lexical classifier	is-a	181
Total LR	-	**204**
Total Lexical	-	**286**

6 Conclusions

We have presented the approach followed by the UPF-UPC team to the TASS 2018 Task 3 challenge. Our official results were not bad for the tasks A and B but were deceiving for task C. The results were consistent among the three scenarios with no clear improvement, as could be expected, from the first to the third. Anyway this problem seems to be general.

In our post-challenge analysis we detected a serious bug on the generation of our run for task C. Correcting it we improved the F1 for task C in the first scenario from 0.093 to 0.27. This is a clear improvement but the score continues to be small.

Using the *YATE* based approach seems to be a valuable way of facing tasks A and B.

The results of task C are clearly deceiving. Specially the recall score is very low, precision is closer to the winner.

Some issues detected merit to perform a more in depth analysis that we propose as

future work:

- Improve *YATE* accuracy when dealing with Actions.
- Improve the integration of SNOMED CT in YATE architecture
- Analyse why there is no clear improvement when moving from scenario 1 to 3.
- Analyse the contribution of the different features to the tasks A, B, and C.
- Analyse the poor recall of the two *LR* classifiers for task C
- In the case of task C we have used two linear multi-class classifiers *LR*. The set of features used was rather big and as the number of training samples was small the learning process was prone to overfitting. We plan to test other non linear classifiers and to reduce the number of features, specially the lexical ones, lemmas and atomic forms, probably using embeddings learned on the medical domain.
- Analyse the performance of the two *LR* classifiers separately. Only the overall results have been considered. The fact that basically the same feature set was used for learning the two classifiers was not the better solution.
- Taking into account that the number of classes in task C is rather small, moving from multi-class to binary classification, i.e learning one classifier per class, should be considered.

Acknowledgements

This work was partially supported by the projects TERMMED (FFI2017-88100-P, MINECO) and GRAPHMED (TIN2016-77820-C3-3R).

References

Gonzalez-Agirre, A., E. Laparra, and G. Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the Sixth International Global WordNet Conference (GWC'12).*, Matsue, Japan.

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.

Montalvo, S., M. Oronoz, H. Rodríguez, and R. Martínez. 2017. Biomedical abbreviation recognition and resolution by prosamed. In *Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages*, pages 247–254, Murcia, Spain, September. SEPLN.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Vivaldi, J. 2001. *Extracción de Candidatos a Término mediante combinación de estrategias heterogéneas*. Ph.D. thesis, Department of Computer and Information Science, Politechnical University of Catalonia, Barcelona, Spain, 06.

Vivaldi, J. and H. Rodríguez. 2010. Using Wikipedia for term extraction in the biomedical domain: first experience. In *Procesamiento del Lenguaje Natural*, volume 45, pages 251–254.

Vossen, P. 2004. EuroWordNet: a multilingual database of autonomous and language specific wordnets connected via an Inter-Lingual-Index. International. volume 17, pages 161–173.