UNED at DIANN 2018: Unsupervised System for Automatic Disabilities Labeling in Medical Scientific Documents

Hermenegildo Fabregat[1][0000-0001-9820-2150], Juan Martinez-Romo[1,2][0000-0002-6905-7051], and Lourdes Araujo[1,2][0000-0002-7657-4794]

[1] NLP & IR Group, Dpto. Lenguajes y Sistema Informáticos.
Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain
[2] IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid, 28019, Spain

Abstract. This paper introduces the LSI system participating in DIANN (Disability annotation on documents from the biomedical domain) task, framed in the IBEREVAL 2018 evaluation workshop. The identification of medical concepts in documents and, especially, the identification of disabilities, is a complex task mainly due to the variety of expressions that can make reference to the same problem. Our proposal implements an automatic annotation tool similar to UMLS MetaMap Transfer (MMTx) for extracting biomedical concepts. As MetaMap, our system generates different variants of the same disability aiming to improve coverage, and adapting them to the kind of entity considered. The first results of the system on a evaluation corpus of 500 scientific papers manually annotated indicate the potential of the proposal.

Keywords: Annotation · Disabilities · Biomedical domain · Negation.

1 Introduction

This paper presents a unsupervised approach to the Disability detection task DIANN proposed in the framework of IBEREVAL 2018 [6].

The main goal of DIANN task is the annotation of disabilities. These conditions affect to a large part of population. For example, they are present in many rare disease. Therefore, it is extremely important to collect information related to them. There are some tools for the annotation of medical concepts, especially in English, such as Metamap. However, they do not consider disabilities as a distinctive concept, but as any other sign. Thus, they do not allow to distinguish a disability, usually a permanent condition, from other signs associated to diseases. The task is evaluated in two sub-tasks, corresponding to the detection of entities in English and Spanish.

The dataset has been collected between 2017 and 2018. DIANN's corpus consists of a collection of 500 abstracts from Elsevier journal papers related to the biomedical domain. Every abstract has available both versions in Spanish and English.

Disabilities appearing in these abstracts have been annotated using the XML tag $< dis >$. Disabilities are usually expressed either with a specific word, such as "blindness", or as the limitation or lack of a human function, such as "lack of vision". A sample of annotatted disability is the following: *Fragile-X syndrome is an inherited form of* $< dis >$ *mental retardation* $< /dis >$ *with a connective tissue component involving mitral valve prolapse.* The boundaries among diseases, disabilities and signs are often unclear. Organizers provide a list of disability terms and a list of functions whose absence or limitation has been considered a disability.

Currently, there are very few annotators of entities adapted to the spanish medical domain. MetaMap Transfer (MMTx) [1] is an application that has two main features, on the one hand it can map medical texts to the UMLS thesaurus[3], and on the other hand it allows to discover thesaurus concepts in documents. This system applies a lexical/syntactic analysis to the input text that involves the following steps: tokenizer, lexical labelling and surface syntactic analysis and identification of the syntax core. The evaluation, which is carried out both in the correspondence of candidates and in the final proposals, is a linear combination of four linguistically inspired measures: centrality, variation, coverage and cohesiveness.

Due to this shortage of resources in Spanish, works such as [2] appeared in which they tried to adapt MetaMap to Spanish by translating the texts into English, and then applying the extraction of medical concepts using MetaMap. Later on, other works like MOSTAS arose that implemented a complete system. This morpho-semantic labeling system also performs text anonymization and spell checker functions with the aim of allowing the identification of clinical terms through the use of SNOMED CT. Castro et al. [3] presented a proposal for semantic annotation of clinical reports in Spanish. They implemented a tool similar to UMLS MetaMap Transfer (MMTx) for the identification of medical concepts on the Spanish ontology SNOMED CT. In another similar work [9, 8] an annotation tool has been developed that detects entities in the biomedical domain. Based on Freeling, the authors enrich their lexicon with biomedical terms from dictionaries and ontologies. The evaluation was conducted on drugs, substances and diseases. Vivaldi and Rodriguez [10] created a term extraction system that uses semantic information extracted from Wikipedia. The system was tested on a medical corpus, and according to the results, it could be considered a good resource for the extraction of medical terms. Conrado et al. [5] carried out an automatic extraction of medical terms, using nominal syntagmas previously recognized in medical texts in Spanish. The authors, using SNOMED CT, demonstrate that it is possible to extract medical terms using specific nominal syntagmas.

[3] http://www.nlm.nih.gov/research/umls/

2 Description of the System

The system, which uses external resources to perform some language processing tasks, begins with the thesaurus (lists of disabilities and body functions) processing in which the variants of the disabilities it contains are generated. Then, given a document, it identifies the noun phrases and generates their variants. Variants of both the disabilities and the body functions are generated in the document using Wordnet [7]. It is therefore possible to configure the variant generation levels in both the document and the thesaurus. The disability annotation system is divided into several phases as can be seen in the Figure 1.

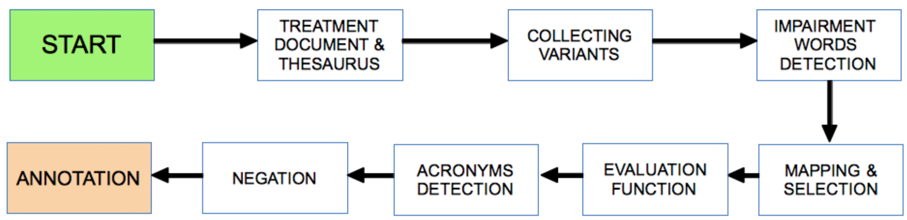

Fig. 1. Phases of the disabilities annotation system.

We have an initial phase that consists of processing the thesaurus by means of lists of disabilities, body functions, words of impairment and their variants. This list will provide us with basic terminology for identifying disability expressions in texts.

Then, for each document considered, the noun phrases (NP), the tokens of each NP and the variants of each token are obtained. It is also identified if the terms correspond to words of impairment, which may be an indication of disability. The correspondence between the noun phrases of the document (NPD) and the candidates of the thesaurus is established below. This ratio will be 1 to N.

Then, for the relationships (NPD-Candidates) obtained in the previous phase, an affinity calculation is performed, which allows the establishment of a ranking from which the best candidate is selected. This last phase employs an evaluation function which takes into account four measures: centrality, variation, coverage and cohesiveness:

- Centrality: The centrality value is simply 1 if the term in the thesaurus involves the head of the noun phrase and 0 otherwise.
- Variation: The variation value estimates how much the variants in the thesaurus term differ from the corresponding words in the noun phrase.
- Coverage: The coverage value indicates how much of the thesaurus term and the noun phrase are involved in the match.

- Cohesiveness: The cohesiveness value is similar to the coverage value but emphasizes the importance of connected components. A connected component is a maximal sequence of contiguous words participating in the match.

2.1 Acronym Detection

The following additions have been made to the implementation of disability detection by generating variants.

It is based on the premise that all abbreviations are presented in this format:

$$[text]ENTITY(ACRONYM\_ENTITY) \tag{1}$$

and they can be captured using a regular expression.

Variants such as the following are also considered:

$$ENTITY(ACRONYM\_ENTITY; [TEXT\_NOISE]) \tag{2}$$

where everything following the semicolon would not be included in the annotation. Once an acronym is recorded as present in the document, all of its appearances are annotated.

Once an acronym has been found, the document is re-processed from the line where the acronym was found by searching for words containing the initials of the acronym.

The main problem with this form of annotation is that it is totally dependent on the annotation of the entities because for the regular expression to work, the entity that precedes the acronym must be correctly annotated.

2.2 Negation

In the case of English, we have used pyConText tool [4]. Two restrictions are currently set up:

- A negation is a negation as long as it includes a disability. (As with acronyms, detection of negation is dependent on the disability annotation process).

About the scope, the following restrictions are described because they correspond to the annotation guidelines followed in the DIANN annotation:

- The scope of a negation does not start from where PyContext indicates (since it usually takes the entire phrase), but from the first annotated element that is included within the scope that it indicates, either the negation trigger or a disability.
- The scope of a negation ends at the last item annotated inside the scope annotated by PyContext.

In the case of Spanish we have used regular expressions. The following list has been used for trigger identification: [”no”,”ausencia de”,”sin”,”sin signos”,”sin signo”,”sin evidencias”,”sin evidencia”,”sin evidencias de”,”sin evidencia de”,”libre de”,”libres de”,”ausencia”].

3 Results

Below are the results obtained by the system in the DIANN task. The results consist of six tables, the first three are for English and the next three for Spanish. For each language, the results of detecting only disabilities are shown first, then the annotation of negated disabilities, and finally the combinned annotation of disabilities and negation.

The DIANN task allows three runs per language to be sent. The system has been configured to filter the results obtained according to a threshold. This threshold corresponds to the evaluation function of the system and the three runs sent correspond to the values: 0.5, 0.6, and 0.7. Results with values below this threshold are not annotated.

3.1 English

Table 1 shows the results obtained evaluating the annotation of all disabilities in English. Runs 1, 2, and 3, correspond to a threshold of 0.5, 0.6, and 0.7 in the score of the evaluation function of the system. In this way, the 0.7 threshold gets the best score in F-measure and precision. The threshold used does not seem to have an significant impact on recall.

English Disability						
	Exact Evaluation			**Partial Evaluation**		
Run	**P**	**R**	**F-Measure**	**P**	**R**	**F-Measure**
LSI_1	0,633	0,597	0,614	0,808	0,761	0,784
LSI_2	0,639	0,597	0,617	0,815	0,761	0,787
LSI_3	0,671	0,597	0,632	0,856	0,761	0,806

Table 1. Results obtained evaluating the annotation of all disabilities in English (included or not in a negation). Both partial and exact evaluation results are included.

Table 2 shows the results of the evaluation of the annotation of negated disabilities in English. In this case, 0.6 and 0.7 thresholds get the best score in F-measure and precision. Exact evaluation gets poor results compared to partial evaluation, perhaps due to a faulty scope detection.

Table 3 show the results for English obtained evaluating jointly the annotation of disabilities and negation. In this evaluation, the 0.7 threshold gets the best score in F-measure and precision. The threshold used does not seem to have an significant impact on precision. The impact of the threshold on precision is significant with a difference of five points between 0.5 and 0.7 thresholds.

3.2 Spanish

Table 4 shows the results obtained evaluating the annotation of all disabilities in Spanish. As you can see in the table, the 0.6 threshold obtains the best

English Negated Disability						
	Exact Evaluation			Partial Evaluation		
Run	**P**	**R**	**F-Measure**	**P**	**R**	**F-Measure**
LSI_1	0,176	0,13	0,15	0,824	0,609	0,7
LSI_2	0,188	0,13	0,154	0,875	0,609	0,718
LSI_3	0,188	0,13	0,154	0,875	0,609	0,718

Table 2. Results of the evaluation of the annotation of negated disabilities in English. Both partial and exact evaluation data are shown.

English Non-negated Disability + Negated Disability						
	Exact Evaluation			Partial Evaluation		
Run	**P**	**R**	**F-Measure**	**P**	**R**	**F-Measure**
LSI_1	0.616	0.568	0.591	0.79	0.728	0.758
LSI_2	0.624	0.568	0.595	0.801	0.728	0.763
LSI_3	0.657	0.568	0.609	0.843	0.728	0.781

Table 3. Results for English obtained evaluating jointly the annotation of disabilities and negation (negated disability are considered correct if both negation and disabiity are correct). Both partial and exact evaluation results are included.

performance for both the F-measure and the precision. In the case of Spanish, we now see how the recall and F-measure fall as the threshold increases in the case of the partial evaluation.

Spanish Disability						
	Exact Evaluation			Partial Evaluation		
Run	**P**	**R**	**F-Measure**	**P**	**R**	**F-Measure**
LSI_1	0,393	0,249	0,305	0,841	0,533	0,652
LSI_2	0,396	0,249	0,306	0,847	0,533	0,654
LSI_3	0,41	0,249	0,31	0,842	0,511	0,636

Table 4. Results obtained evaluating the annotation of all disabilities in Spanish (included or not in a negation). Both partial and exact evaluation results are included.

Table 5 shows the results of the evaluation of the annotation of negated disabilities in Spanish. Before analyzing the results, it is necessary to say that the detection of noun phrases in the case of Spanish has been difficult and this has caused a general decrease in the results obtained. As a result of this, you can see the poor results obtained in the case of the exact evaluation. This is because the detection of the negation depends on the detection of disabilities and the latter has not been good enough.

Spanish Negated Disability						
	Exact Evaluation			Partial Evaluation		
Run	**P**	**R**	**F-Measure**	**P**	**R**	**F-Measure**
LSI_1	-	-	-	0,75	0,136	0,231
LSI_2	-	-	-	0,75	0,136	0,231
LSI_3	-	-	-	0,75	0,136	0,231

Table 5. Results of the evaluation of the annotation of negated disabilities in Spanish. Both partial and exact evaluation data are shown.

Finally, Table 6 shows the results for Spanish obtained evaluating jointly the annotation of disabilities and negation. In this case, depending on the type of evaluation that we analyze and the measure used, the performance will be different. If we look at the exact evaluation, the best threshold was 0.7, however in the case of the partial evaluation the threshold that obtained the best performance was 0.6.

Spanish Non-negated Disability + Negated Disability						
	Exact Evaluation			Partial Evaluation		
Run	**P**	**R**	**F-Measure**	**P**	**R**	**F-Measure**
LSI_1	0,406	0,245	0,305	0,797	0,48	0,599
LSI_2	0,409	0,245	0,306	0,803	0,48	0,601
LSI_3	0,424	0,245	0,31	0,803	0,463	0,587

Table 6. Results for Spanish obtained evaluating jointly the annotation of disabilities and negation (negated disability are considered correct if both negation and disabiity are correct). Both partial and exact evaluation results are included.

4 Conclusions

Our proposal to annotate disabilities in medical documents is based on the generation of variants from the terms and expressions considered. These variants are obtained both from the lists of disabilities provided by the task and from the noun phrases extracted from the text. The variants considered have been derivative and synonymous.

With respect to MetaMap, which is the reference tool in this type of task, we have included several improvements trying to adapt the identification of medical concepts to the specific problem of disabilities. In fact, the system detects negated disabilities and acronyms. However, these improvements can be applied to other types of medical concepts. The proposed system also allows to configure the level of generation of variants, both in analyzed documents and thesaurus.

The results obtained in the task indicate that the system is capable of achieving competitive levels of precision and recall, considering that it is an unsupervised system.

Acknowledgment

This work has been partially supported by the projects EXTRECM (TIN2013-46616-C2-2-R), PROSA-MED (TIN2016-77820-C3-2-R), and EXTRAE (IMIENS 2017).

References

1. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of AMIA, Annual Symposium. pp. 17–21 (2001)
2. Carrero, F.M., Cortizo, J.C., Gómez, J.M., de Buenaga, M.: In the development of a spanish metamap. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. pp. 1465–1466. CIKM '08, ACM, New York, NY, USA (2008)
3. Castro, E., Iglesias, A., Martínez, P., Castaño, L.: Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. In: Proceedings of the 1st ACM International Health Informatics Symposium. pp. 751–757. IHI '10, ACM, New York, NY, USA (2010)
4. Chapman, B.E., Lee, S., Kang, H.P., Chapman, W.W.: Document-level classification of ct pulmonary angiography reports based on an extension of the context algorithm. Journal of biomedical informatics **44**(5), 728–737 (2011)
5. Conrado, M.S., Koza, W., Dıaz-Labrador, J., Abaitua, J., Rezende, S.O., Pardo, T.A., Solana, Z.: Experiments on term extraction using noun phrase subclassifications. pp. 746–751 (2011)
6. Fabregat, H., Martinez-Romo, J., Araujo, L.: Overview of the diann task: Disability annotation task at ibereval 2018 (September 18, 2018)
7. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
8. Oronoz, M., Casillas, A., Gojenola, K., Perez, A.: Automatic annotation of medical records in spanish with disease, drug and substance names. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Lecture Notes in Computer Science, vol. 8259, pp. 536–543. Springer Berlin Heidelberg (2013)
9. Oronoz, M., de Ilarraza, A.D., Torices, O.: First steps in the manual and automatic annotation of clinical notes in spanish. Procesamiento del Lenguaje Natural **45**, 259–262 (2010), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/815
10. Vivaldi, J., Rodríguez, H.: Using wikipedia for term extraction in the biomedical domain: first experiences. Procesamiento del Lenguaje Natural **45**, 251–254 (2010)