SINAI at DIANN - IberEval 2018. Annotating disabilities in multi-language systems with UMLS

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, María Teresa Martín-Valdivia, and Salud María Jiménez-Zafra

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda,mcdiaz,maite,sjzafra}@ujaen.es

Abstract. In this paper we present our first participation as SINAI research group from the Universidad de Jaén at DIANN (Disability annotation on documents from the biomedical domain) task in IberEval. Our research aim is to create a Named Entity Detection system based on Natural Language Processing techniques in Spanish and to compare it with existing systems in other languages. For this, we identify disabilities in English and Spanish texts using techniques such as syntactic analysis and word embeddings, including a negation detection module. The results obtained are higher in English than in Spanish because MetaMap contains a good negation detection system.

Keywords: Natural Language Processing, Named Entity Recognition, Text Classification, Disability, Negation Annotation, Negation Scope, UMLS, MetaMap

1 Introduction

The concept of disability is defined as the condition that prevents or limits people in their daily lives and usually permanently. According to the World Health Organization (WHO)[1], disability is a general term that encompasses impairments, activity limitations and participation restrictions, so for there to be a disability, there must be a deficiency. A key finding of the World Report is that 1 billion people, 15% of the global population, have some disability [11]. In Spain there are more than 3.8 millions, 8.5% of the population.

It is important to consider tasks such as the DIANN proposal because most of the related work focuses on identifying medical concepts, but few recognize disabilities in the two most widely spoken languages in the world. Most of the existing systems are in English, for example, *Unified Modeling Language MetaMap Transfer* (UMLS MMTx)[2] [3, 12] is a configurable tool commonly used by biomedical system developers. It was created by researchers at the National

[1] http://www.who.int (last visited: May 31, 2018)
[2] https://www.nlm.nih.gov/research/umls/ (last visited: May 31, 2018)

Library of Medicine (NLM) and it is able to identify biomedical concepts from unstructured text and map them to the UMLS Metathesaurus [14] concepts. In reference to the SNOMED-CT ontology and the English language, there are several related research [1, 13, 6] for biomedical text processing.

On the other hand, another objective of the proposed task is the negation detection and the scope. This task is essential for properly understanding clinical texts.

This paper is organized as follows: in the next section, we introduce the collection of documents provided by the organizers. Our approach is described in Section 3. In Section 4 we include the results obtained and in Section 5, we comment conclusions and future works.

2 Collection

The corpus provided is composed of 500 abstracts of Elsevier journal papers related to the medical domain [5]. From this collection of documents, abstracts have been selected in Spanish and English.

Each document of the collection has been annotated with the disabilities present in it. Moreover, if a disability is negated it has been added information about the negation cue and the scope. For more details of the annotation, we provide several examples in both languages taken from the training corpus:

English

```
In the group <scp><neg>without</neg> <dis>cognitive impairment
</dis></scp>, the diagnosis was known by 83%, and 30% knew the
prognosis.
```

Spanish

```
Del grupo <scp><neg>sin</neg> <dis>afectación cognitiva</dis>
</scp>, un 83% conocían el diagnóstico, un 30% el pronóstico.
```

3 Our approach

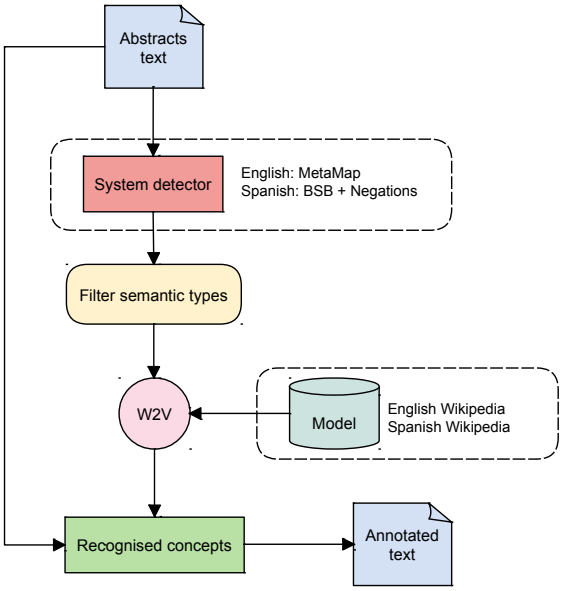
In the group SINAI, we have approached the two languages provided by the organizers, so we developed different systems and runs depending on the language. The figure 1 describes the general architecture developed for each language.

3.1 English

We use different steps to achieve the goal. In the first step, we use MetaMap[3]. MetaMap is a software that discover UMLS Metathesaurus concepts in the text. We use MetaMap with NegEx for negation analysis [2, 4] and scope. The concepts detected belong to any category of UMLS, and not only refer to disabilities. Therefore, we need to remove all those concepts that are not disabilities.

[3] https://metamap.nlm.nih.gov/

Fig. 1. Architecture of the approach



For this reason, we filter by different semantic types of UMLS in the second step. Performing an analysis of existing disabilities in the training corpus, we have observed that most disabilities are within these three semantic types of UMLS:

- Mental or Behavioral Dysfunction
- Disease or Syndrome
- Finding

Finally, in the last step, we evaluate the similarity between each concept and the term "*disability*" using word embeddings [10]. Word embeddings are capable of capturing semantic and syntactic relations between the words. We use a model created from Wikipedia[4] with bag-of-words architecture (CBOW), size 400 and window size 5.

For each word in the UMLS concept we calculate the similarity between this word and the term "*disability*". Next, we select the maximum value of all similarities, and if it is greater than a given threshold this concept is selected. The threshold used for our experiments is 0.35, because empirically it has shown good results.

[4] https://en.wikipedia.org/ (last visited: May 31, 2018)

3.2 Spanish

In the case of Spanish, we use a biomedical entity recognizer created by the SINAI group. The entity detector normalizes the text changing all words to lowercase and removing special characters, both in the dictionaries used and in the text entered. The tool used in this case is the NLTK[5] library (*Natural Language Toolkit*) developed in the Python programming language. In addition, for greater precision in identifying terminology, the syntax analyzer included in the CoreNLP[6] tool developed by Stanford University for Spanish [9] is used. The system uses the UMLS concept dictionary in Spanish.

We made several improvements in our entity detector to get a higher hit score, for example, we added a list of disability abbreviations from various sources[7] and if the recognized term contained a quantifiable adjective in front or behind, it was added to take into account cases such as "*severe* functional impairment".

We also apply the filter for semantic types, as in the English language, using word embeddings with a model created with the Wikipedia[8] in Spanish with bag-of-words architecture (CBOW), size 300 and window size 5.

Finally, for the treatment of negation in Spanish we use the method developed in [8, 7] using a bag of words. The list of identified negation keys is the following:

Keywords negation

```
no (not), tampoco (neither), nadie (nobody), jamás (never),
ninguno (none), ni (nor), sin (without), nada (nothing),
nunca (never)
```

This method detects the negation and the scope, including all the negated words. In most cases, our scope contains more words than those observed in the training corpus. For this reason, we reduce our scope so that it begins with the first negation keyword and ends with the last word of the detected disability.

4 Results

Table 1 shows the 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 includes the results of the evaluation of the annotation of negated disabilities in English. For this evaluation, we have considered as negated disability the set of annotations (disability, negation trigger and scope of the negation).

[5] https://www.nltk.org/ (last visited: 31 May, 2018)
[6] https://stanfordnlp.github.io/CoreNLP/ (last visited: May 31, 2018)
[7] https://www.abbreviations.com/acronyms/DISABILITY
http://www.parentcenterhub.org/acronyms/
https://www.parentingspecialneeds.org/article/disability-acronyms-abbreviations
(last visited: May 31, 2018)
[8] https://es.wikipedia.org/ (last visited: May 31, 2018)

Table 1. Results in all English disabilities detection

	Exact			Partial		
Run	Precision	Recall	F1	Precision	Recall	F1
SINAI-Run1	0.016	0.593	0.032	0.019	0.704	0.038
SINAI-Run2	0.222	0.428	0.293	0.252	0.486	0.332
SINAI-Run3	0.625	0.370	**0.465**	0.688	0.407	**0.512**

Table 2. Results in English negated disabilities detection

	Exact			Partial		
Run	Precision	Recall	F1	Precision	Recall	F1
SINAI-Run1	0.250	0.391	0.305	0.556	0.87	0.678
SINAI-Run2	0.306	0.478	0.373	0.556	0.870	0.678
SINAI-Run3	0.526	0.435	**0.476**	1.000	0.826	**0.905**

Table 3 shows the results for English obtained evaluating jointly the annotation of disabilities and negation.

Table 3. Results in English non-negated disability + negated disability detection

	Exact			Partial		
Run	Precision	Recall	F1	Precision	Recall	F1
SINAI-Run1	0.015	0.543	0.029	0.019	0.691	0.037
SINAI-Run2	0.199	0.395	0.264	0.242	0.481	0.322
SINAI-Run3	0.573	0.337	**0.425**	0.685	0.403	**0.508**

Table 4 shows the 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 only includes the results of the evaluation of the annotation of negated disabilities in Spanish. For this evaluation, we have considered as negated disability the set of annotations (disability, negation trigger and scope of the negation).

Table 6 shows the results for Spanish obtained evaluating jointly the annotation of disabilities and negation.

All the runs shown are composed as follows:

- **Run1**: The system annotated all the concepts detected and returned by the system (MetaMap or our Spanish detection system).
- **Run2**: Use the semantic type filter.
- **Run3**: Apply word embedding to get similarity.

Table 4. Results in all Spanish disabilities detection

	Exact			Partial		
Run	Precision	Recall	F1	Precision	Recall	F1
SINAI-Run1	0.022	0.485	0.042	0.026	0.568	0.05
SINAI-Run2	0.181	0.415	0.252	0.204	0.467	0.284
SINAI-Run3	0.459	0.345	**0.394**	0.512	0.384	**0.439**

Table 5. Results in Spanish negated disabilities detection

	Exact			Partial		
Run	Precision	Recall	F1	Precision	Recall	F1
SINAI-Run1	0	0	0	0.125	0.045	0.067
SINAI-Run2	0.333	0.045	0.08	0.667	0.091	0.16
SINAI-Run3	0.667	0.091	**0.16**	1	0.136	**0.24**

Table 6. Results in Spanish non-negated disability + negated disability detection

	Exact			Partial		
Run	Precision	Recall	F1	Precision	Recall	F1
SINAI-Run1	0.018	0.402	0.035	0.022	0.48	0.042
SINAI-Run2	0.157	0.349	0.217	0.18	0.402	0.249
SINAI-Run3	0.411	0.284	**0.336**	0.468	0.323	**0.382**

5 Conclusions and future work

We presented our approach for DIANN task of the IberEval workshop of the International Conference of the Spanish Society for Natural Language Processing (SEPLN) 2018 where the goal is to annotate disability content in papers related to the biomedical domain. The provided collection is made of two versions regarding the language: English and Spanish.

Our results obtained in English language are higher than Spanish. We notice that the improvement is significant in the annotation of the negations in English and we can see that MetaMap with NegExp does a good job and we must continue to research in Spanish to achieve more successful results taking into account the annotation guide provided by the organization to adapt our negation detection system for medical domain.

Perspectives for further work include, to improve our automatic detection system in Spanish, for example, using new models of words embedding according to the task and focused on disabilities. We will analyze the different medical ontologies to refine the identification of disabilities in a text.

Acknowledgments

This work has been partially supported by a grant from Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983) from the Spanish Government.

References

1. Allones, J.L., Martínez, D., Taboada, M.: Automated mapping of clinical terms into snomed-ct. an application to codify procedures in pathology. J. Medical Systems **38**(10), 134 (2014)
2. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association **17**(3), 229–236 (2010)
3. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
4. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics **34**(5), 301–310 (2001)
5. Fabregat, H., Martinez-Romo, J., Araujo, L.: Overview of the diann task: Disability annotation task at ibereval 2018. In: Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
6. Garla, V.N., Brandt, C.: Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC bioinformatics **13**(1), 261 (2012)
7. Jiménez Zafra, S.M., Martínez Cámara, E., Martín Valdivia, M.T., Molina González, M.D.: Tratamiento de la negación en el análisis de opiniones en espanol. Procesamiento del Lenguaje Natural **54**, 37–44 (2015)
8. Jimenez-Zafra, S.M., Valdivia, M.T.M., Camara, E.M., Urena-Lopez, L.A.: Studying the scope of negation for spanish sentiment analysis on twitter. IEEE Transactions on Affective Computing (2017)
9. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of ACL: system demonstrations. pp. 55–60 (2014)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Officer, A., Shakespeare, T.: The world report on disability and people with intellectual disabilities. Journal of Policy and Practice in Intellectual Disabilities **10**(2), 86–88 (2013)
12. Osborne, J.D., Lin, S., Zhu, L.J., Kibbe, W.A.: Mining biomedical data using metamap transfer (mmtx) and the unified medical language system (umls). Gene Function Analysis pp. 153–169 (2007)
13. Sánchez, D., Batet, M., Valls, A.: Web-based semantic similarity: an evaluation in the biomedical domain. International journal of software and informatics **4**(1), 39–52 (2010)
14. Wright, L.W., Nardini, H.K.G., Aronson, A.R., Rindflesch, T.C.: Hierarchical concept indexing of full-text documents in the unified medical language system information sources map. Journal of the Association for Information Science and Technology **50**(6), 514 (1999)