GPLSIUA Team at the DIAAN 2018 task

Isabel Moreno[1][0000−0002−3052−7669], M.T. Romá-Ferri[2][0000−0002−3484−5621], and Paloma Moreda[1][0000−0002−7193−1561]

[1] Department of Software and Computing systems, University of Alicante, Spain
{imoreno,moreda}@dlsi.ua.es
[2] Department of Nursing, University of Alicante, Spain
mtrferri@ua.es

Abstract. This paper describes our participation in DIANN 2018 Task: DIsability ANNotation in English and Spanish documents. Our proposal detects disabilities as well as recognizes negated disabilities. To that end, our entity typing system is applied without tuning and it does not require any external knowledge. It consists of a Random Forest machine learning classifier whose feature set includes local entity information and profiles, generated unsupervisedly. Two experiments are presented in order to investigate performance of two types of profiles. Both proposals are able to reach promising and reasonable results, obtaining a partial-matching precision greater than 87% for disabilities and negated disabilities regardless of the language. Thus demonstrating the portability and adequateness of our approach regardless of type of profile.

Keywords: Disability · Negation · Named entity · Profiles · Machine learning · Random Forest · Language independent

1 Introduction

Disability is defined as "any condition of the body or mind (impairment) that makes it more difficult for the person with the condition to do certain activities (activity limitation) and interact with the world around them (participation restrictions)" [15]. The 2011 World Report on Disability [19] evince that more than one billion people of the world's population have some form of disability. Thus making the information gathering about disabilities of vital importance.

There are some efforts to annotate medical concepts for languages such as English [8, 18] or Spanish [11, 16], but the focus is on sign or disease. In other words, they do not delve into these two concepts in order to distinguish disabilities. This is why the goal of DIANN task is the DIsability ANNotation on scientific abstracts from the biomedical domain in English and Spanish [1, 5].

The request was not only to annotate disabilities (dis) but also to annotate the negation (neg) modifiers affecting at least one disability as well as its scope (scp), as illustrated by Example 1.

Example 1. In November 2000 several informative meetings were held for 41 residents of our center's Nursing Home who were selected as they presented <scp><neg>no</neg> potentially fatal disease or <dis>cognitive impairment</dis></scp>.

For the first edition of the DIANN task we were particularly interested in evaluating CARMEN [12–14], our general purpose Named Entity Typing (NET) system. Such NET system decides whether a possible chunk in a text corresponds or not to a disability. It does not use any external resource (e.g. UMLS Metathesaurus [2]) that contains any physical or intellectual conditions that when impaired give rise to a disability. The idea was to establish how well can we perform in this task without specific domain or language resources and without feature or parameter tuning. Despite the fact that our interest lies on classifying entities, a simple approach to detect negation and scope of negated disabilities is also proposed.

The rest of the paper is structured as follows. Next, our approach is defined in Section 2. The experiments are presented in Section 3. Results are discussed in Section 4. Last, conclusions and future work are outlined in Section 5.

2 Methods

Our approach, as can be seen in Figure 1, consists in five main steps: (i) preprocessing: this stage takes as input an annotated corpus to perform a linguistic analysis (see Section 2.1); (ii) disability annotation: this process implies the extraction of possible candidates in order to decide which ones should be typed as disabilities (see Section 2.2); (iii) negation annotation: this task finds all negation triggers in a given text (see Section 2.3); (iv) scope annotation: this phase determines which negation triggers, previously detected, affect a disability (i.e. must be kept) and its span (see Section 2.4); and last, (v) post-processing: this stage converts the resulting corpus to the competition format (see Section 2.5). In the following sections this work-flow is explained in detail.

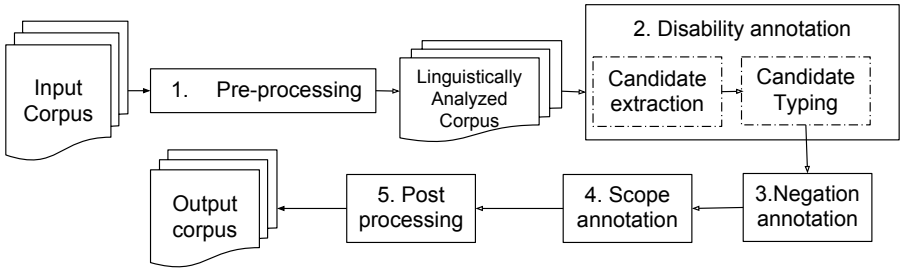

Fig. 1. GPLSIUA approach overview

2.1 Pre-processing

The system takes as input an annotated document following DIANN format. The DIANN format is plain text that can include XML tags to determine the anchor of the three elements that must be annotated (disabilities, negation and scope), as can be seen in Example 1. Such format is converted to a well-formed XML that can be processed by Gate [4]. To that end, special XML characters are escaped and a root tag is created. Besides, its raw text is tokenized, sentence-splitted, lemmatized, PoS-tagged and shallow parsed using Freeling [17]. Last, both outputs (i.e. each linguistically annotated text and its corresponding well-formed XML document) are merged into one document in GATE standoff format.

2.2 Disability Annotation

Disability annotation consists of two modules:

The first one, candidate extraction in Figure 1, identifies possible candidates from text. This module extracts noun phrases detected in shallow parsing to be considered as the set of candidates for being a disability.

The second component, candidate typing in Figure 1, establishes which of the previous noun phrases should be finally typed as disabilities (i.e. binary classification). For this purpose we applied CARMEN, our general purpose NET system [12–14]. CARMEN is a machine learning based system which employs Random Forest (RF) algorithm [3] with the default parameters from Weka 3.8 [7], but the number of iterations has been set to 45, as in [14]. Thus, CARMEN consists of two phases: (i) an offline processing step whose main goal is to train a ML model to perform NET, and (ii) an online processing step whose aim is to decide which previously extracted candidates are a disability.

For each candidate, CARMEN generates a feature vector that includes context and local information of the entity.Table 1 contains all the features generated for the sentence in Example 1, which are further explained next.

Context of the entity is built through profiles [10], specifically CARMEN defines one profile for each entity type (being a disability or not) in an unsupervised manner.

In brief, profiles are generated as follows: For each identified candidate, first we extract lemmas of nouns, verbs, adjectives and adverbs, in a window of size W ($\frac{W}{2} = 5$ words after and before the candidate) and their frequency as the number of occurrences. Second, for each entity type, the training corpus is divided in positive instances (e.g. disability) and negative instances (e.g. no_disability). Third, each lemma found only accompanying positive instances computes a relevance index based on the *term frequency, disjoint corpora frequency* [10] - TFDCF; whereas *relevance common index [10] - RC* - is used for the ones present in all training instances (positive and negative). This step produces a profile for each entity type (e.g. disability) of P elements. Each item in a profile is a pair representing a lemma and its relevance index (TFDCF or RC). The length of the profile (P) is the number of lemmas applying TFDCF and RC indexes. P has been set to maximum 1000 lemmas for both indexes. Once the profiles are built,

the feature vector of CARMEN can be enhanced with either all profile items (i.e. all pairs lemma and its relevance) or only with items that compute TFDCF relevance index.

As previously stated, the feature vector of CARMEN also contains local information of the entity, such features are inspired by state-of-the-art NET modules. These comprise words of the entity, length of the entity, suffixes and prefixes [14]. Besides, character n-grams are included as a new feature.

Table 1. Description of features included in CARMEN comprising context and local information

	Feature	Description	Example
Context	profile	Relevance of lemmas of nouns, verbs, adjectives and adverbs that appear in a window anchored in the candidate. Relevance is obtain according to TFDCF and RC indexes	relevance(disease)= RC(disease)=1.43, relevance(potentially)= TFDCF(potentially)=5.1, ...
Local Information	NE	Words of the entity	cognitive impairment
	NElen	Entity length without stop-words	2
	affixes	Suffixes and prefixes with a length of 1, 2, 3 and 4 characters from the first and last words	c, co, cog, cogn, ment, ent, nt, t
	character n-gram	All lowercase character bigrams, trigrams, fourgrams and fivegrams from the words of the NE	co, og, gn, ni, it, ti, iv, ve, e_, _i, im, mp, pa, ai, ir, rm, me, en, nt, t_, cog, ...

Last, it should be noted that adapting CARMEN to this new scenario was straightforward, since there was no need to change the NET architecture or its parameters. It only required: (i) a linguistic analyzer that is able to deal with the two languages tackled in DIANN task (i.e. English and Spanish) in order to perform sentence detection, tokenization, lemmatization, PoS-tagging and shallow parsing; and (ii) the DIANN training corpus, which was previously annotated with the target entity.

2.3 Negation annotation

Negation is tackled using a dictionary-based approach, thus having two phases:

- An offline step whose main goal is to build a lexicon of negation triggers. For each language, a lexicon is created directly from DIANN training corpus.
- A real-time processing step whose purpose is to annotate all negation triggers of a text. Each lexicon is used to instance a Hash Gazetteer, included within ANNIE plugin from GATE [4]. It performs case insensitive exact matching for each entry in a lexicon within a document.

2.4 Scope annotation

This stage is carried out as a set of heuristics at sentence-level. In order to determine which negation triggers affect disabilities, the applied rule is: for each sentence, all negation triggers that do not co-occur with at least one disability are removed. To be able to define the scope of negated disabilities, scope is established as the anchor of negation trigger and disabilities. For instance, in Example 1, the scope starts at the negation position ("no") and ends with the disability ("cognitive impairment").

2.5 Post-processing

At this point, the results are stored in XML GATE standoff format. As a result, they need to be converted to DIANN format again. To that end, all documents are transformed to an inline XML format without a root tag using GATE Embedded [4].

3 Experiments

As mentioned before, our interest is focused on evaluating CARMEN, our entity typing system. Its feature set includes profiles of each entity type among other features. Profiles are composed of pairs lemma-relevance, but relevance is computed according to two different indexes (TFDCF and RC), as explained in Section 2.2. Therefore, two experiments were submitted aiming at studying the differences of using both relevance indexes or only one, namely:

R1 uses the full profile (both TFDCF and RC relevance indexes) in the feature vector of CARMEN.
R2 uses a reduced profile that only takes into account TFDCF relevance index in the feature vector of CARMEN.

4 Results and Discussion

Initially, the organization provided an annotated training set and an unannotated test for both languages. Next subsections present results for entity typing alone during training phase (Section 4.1) and official test results (Section 4.2) for our complete approach. Finally, the official results are analyzed (Section 4.3).

4.1 Entity Typing Training Results

Since no development set was provided, CARMEN was evaluated using 2 fold stratified cross-validation over the annotated training set. The purpose is to assess the entity typing task alone, assuming our candidate extraction module is "perfect" and all possible disabilities are included in the set of candidates. Hence, Table 2 summarizes results for being a disability reported by Weka.

Table 2. Entity typing training cross-validation results

Source	Language	Run	AUC	Exact		
				P (%)	R(%)	F1(%)
WEKA	English	R1	.974	97.2	76.7	83.9
WEKA	English	R2	.974	94.7	74.7	83.6
WEKA	Spanish	R1	.987	88.5	79.0	83.5
WEKA	Spanish	R2	.986	88.7	79.2	83.6

For each language and each run, values of Area Under the ROC Curve (AUC), Precision, Recall and F-score are reported. AUC is commonly used in biomedical informatics research to measure the performance of a classifier, thus allowing the comparison of several models under the same test [9].

The two first rows show the results of the two experiments for English. Similarly, the last two rows show the results of both experiments for Spanish

According to AUC, in the case of Spanish, it's better to use the full profile (R1 - 98.7) whereas there is no such difference for English (i.e. AUC is the same for the two experiments). Several scales for interpreting these AUC values exists, but there is a consensus that values greater than .96 indicate an excellent discriminatory ability [6]. Therefore, all runs have an excellent AUC regardless of the language. The best Precision, Recall and F-score is obtained for English using the full profile (R1). On the contrary, Spanish achieves the higher results thanks to the reduced profile (R2). In view of these results, combining local information and context (profiles) is appropriate for this task.

4.2 DIANN Test Results

The official results reported by DIANN task organizers over the test set can be found in Table 3. For each language and each run, values of Precision, Recall and F-score are reported for different types of disabilities. Two types of matching are used for the evaluation: partial and exact. First, performance for all disabilities, regardless being negated or not, are shown (type DIS). The first four rows show the results of the two experiments (R1 and R2) for English and Spanish, respectively. Similarly, the next four rows refer to negated disabilities (type NEGDIS). Finally, the last four rows concern non-negated disabilities as well as negated disabilities (type DIS + NEGDIS).

From Table 3, we can see that partial matching always benefits our results regardless the type of disability (DIS, NEGDIS or DIS+NEGDIS). Besides, English always gets the highest results. Concerning all disabilities (DIS), the best Precision, Recall and F-score is achieved for partial matching. As in the training phase, it should be noted that our Spanish system is more accurate using the reduced profile (see precision in Table 3), whereas English requires the complete profile. However, differences between training (see recall and F1 in Table 2) and test results suggest a problem in determining the boundaries of disabilities.

Table 3. Official results of the runs over the test set

Language	Type	Run	Exact			Partial		
			P (%)	R(%)	F1(%)	P (%)	R(%)	F1(%)
English	DIS	R1	88.1	24.3	38.1	**94.0**	**25.9**	**40.6**
		R2	88.4	25.1	39.1	91.3	25.9	40.4
Spanish	DIS	R1	81.3	17.0	28.2	95.8	20.1	33.2
		R2	79.6	17.0	28.1	**95.9**	**20.5**	**33.8**
English	NEGDIS	R1	64.7	47.8	55.0	**94.1**	**69.6**	**80.0**
		R2	61.1	47.8	53.7	88.9	69.9	78.0
Spanish	NEGDIS	R1	0	0	0	**50.0**	**9.1**	**15.4**
		R2	0	0	0	40.0	9.1	14.8
English	DIS + NEGDIS	R1	81.2	23.0	35.9	**94.2**	**26.7**	**41.7**
		R2	80.6	23.9	36.8	90.3	26.7	41.3
Spanish	DIS + NEGDIS	R1	69.2	11.8	20.1	**89.7**	**15.3**	**26.1**
		R2	65.9	11.8	20.0	87.8	15.7	26.7

DIS: all disabilities (included or not in a negation); NEGDIS: negated disabilities (considers disability, negation trigger and scope of the negation); DIS + NEGDIS: disabilities and negation (negated disability are considered correct if both negation and disability are correct).

Regarding negated disabilities (NEGDIS), there is an striking difference between English and Spanish performance and English is the highest by far. For both, the use of the complete profile (R1) seems more appropriate.

Last, concerning non-negated and negated disabilities (DIS+NEGDIS), our first experiment (R1) achieves the best results for English. Although Spanish obtains the best Precision with the complete profile, the highest Recall and F-score is accomplished for the second experiment (R2).

In general terms, our proposal performed reasonably well, particularly given that CARMEN system was applied without fine tuning parameters or features. Besides, no additional external knowledge has been used to find disabilities in this narrow domain.

4.3 Results Analysis

As previously stated, GPLSI team obtains high precision values, specially identifying disabilities regardless being negated or not, but recall values are a bit lower. These results are reasonably good, especially considering that (i) CARMEN system was applied off-the-shelf; and (ii) precise system are desired in a medical environment. In order to find reasons for the low recall, once the annotated test set was released, a 5% of the test set was examined carefully to find possible improvements to be implemented as future work.

Analyzing the results, regardless of the language, we found that most errors are related to the boundary of disabilities and negated disabilities. This is because the extraction module often includes extra tokens or misses some of them. Another problem found are disabilities represented by acronyms. Although the acronym definition usually appears in text, CARMEN is not able to classify it as

a disability. Another source of errors is concerned with detecting certain tokens as a disability in a document but not detecting them in another, so effects on recall are evident. This might be explained by two reasons. On the one hand, more local information and context may be needed in order to build a more robust representation of a disability. Examples of features to characterize local information of an entity mention could be its lemma, as well as its POS and shallow parsing tags. Additionally, such features could be also incorporated for our profile (i.e. lemmas in a window) jointly with word-embedding or brown clusters to enhance context. On the other hand, no additional knowledge has been use to determine disabilities, but it could avoid losing disabilities and has a direct effect on recall. Hence, as future work, we plan to implement new features to capture both.

The last source of errors concerns the negation and scope detection modules. On the one hand, our heuristics applied at sentence-level are too optimistic. Although a negation trigger appears in a sentence, it does not necessary affects all tokens in a sentence. This could be improved considering heuristics at a lower granularity, e.g. clause-level. On the other hand, the negation lexicon does not contain all possible triggers, thus some disabilities are not considered negated in testing. This could be improved gathering negation triggers from other corpora. Both issues are particularly problematic for Spanish due to its high flexibility and variance in comparison with English.

Finally, there were a few annotation issues which, in our opinion, could affect participant systems:

- Wrong tokenization in disability annotation: For example, "`<dis>`severe mental illness`</dis>`es" instead of "`<dis>`severe mental illnesses`</dis>`";
- Inconsistent disability annotation: Texts are the same for both languages, but the same disabilities are not present for both versions. For example, "nonagenarians with recent onset of `<dis>functional impairment</dis>` also benefit from rehabilitation in a medium-stay geriatric unit;[...]" but "los pacientes nonagenarios con incapacidad reciente también se benefician del tratamiento en una UME, [...]".

5 Conclusions

In this paper our proposal to detect both disabilities and negated disabilities is presented. On the one hand, negation and its scope is tacked with a set of simple heuristics and dictionaries. On the other hand, disabilities are extracted using our entity typing system, CARMEN, for this new task. It employs Random Forest, a supervised machine learning algorithm. Its feature set is based on profiles (context of the surrounding words) and information gathered from the NE itself. In this manner, the actual performance of CARMEN is studied when applied to a new genre and a new entity. Two experiments are presented in order to investigate performance of two types of profiles.

Our training phase results for the entity typing task alone (AUC $>$ .96 and F1 almost 84%) show all runs have an excellent discriminatory ability regardless of

the language and profile type. Regarding official testing results, our recall values are a bit lower but partial-matching precision is greater than 87% for disabilities and negated disabilities for all languages and profile types. These results show our approach performs reasonably well when dealing with disabilities, specially considering the lack of external resources or parameter tuning.

Although the results are encouraging, there is still room for improvement. To that end, an analysis of the obtained output has been carried out to explain our results. Concerning the recall of the testing phase, it was found that (i) boundary detection needs to be more accurate and (ii) representation of entities may be enhanced with either more features or external knowledge. Thus, our participation in the DIANN task has given us an excellent opportunity to study which aspects should be considered to achieve a more versatile CARMEN.

Acknowledgments

This research is partially funded by the Spanish Government under the projects RESCATA (reference number TIN2015-65100-R) and REDES (reference number TIN2015-65136-C02-2-R).

References

1. Araujo Serna, L., Martínez Romo, J., Fabregat Marcos, H.: Diann: Disability annotation on documents from the biomedical domain (2018), http://nlp.uned.es/diann, (last accessed May 27, 2018)
2. Bethesda (MD): National Library of Medicine (US): Metathesaurus. In: UMLS® Reference Manual [Internet], chap. 2 (2009), https://www.ncbi.nlm.nih.gov/books/NBK9684/pdf/Bookshelf_NBK9684.pdf
3. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Dowman, M., Aswani, N.: Developing language processing components with GATE Version 8 (a user guide). Department of Computer Science (2016), https://gate.ac.uk/sale/tao/tao.pdf
5. Fabregat, H., Martínez-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. Fan, J., Upadhye, S., Worster, A.: Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine **8**(01), 19–20 (2006). https://doi.org/10.1017/S1481803500013336
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA data mining software: An update. SIGKDD Explorations Newsletter **11**(1), 10–18 (2009). https://doi.org/10.1145/1656274.1656278
8. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T.: The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. Journal of biomedical informatics **46**(5), 914–920 (oct 2013). https://doi.org/10.1016/j.jbi.2013.07.011

9. Lasko, T.A., Bhagwat, J.G., Zou, K.H., Ohno-Machado, L.: The use of receiver operating characteristic curves in biomedical informatics. Journal of Biomedical Informatics **38**(5), 404–415 (2005). https://doi.org/10.1016/j.jbi.2005.02.008
10. Lopes, L., Vieira, R.: Building and Applying Profiles Through Term Extraction. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology. pp. 91–100 (2015), http://aclweb.org/anthology/W15-5613
11. Moreno, I., Boldrini, E., Moreda, P., Romá-Ferri, M.T.: DrugSemantics: A corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics. Journal of Biomedical Informatics **72**, 8 – 22 (2017). https://doi.org/10.1016/j.jbi.2017.06.013
12. Moreno, I., Romá-Ferri, M.T., Moreda, P.: Combining profiles and local information for named entity classification: Adjustment of a domain and language independent approach. In: Proceedings of the Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, LTC 2017, Poznań, Poland, November 17-19, 2017. pp. 73–77 (2017), http://ltc.amu.edu.pl/book/papers/IRIE1-2.pdf
13. Moreno, I., Romá-Ferri, M.T., Moreda, P.: A domain and language independent named entity classification approach based on profiles and local information. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017. pp. 510–518 (2017). https://doi.org/10.26615/978-954-452-049-6_067
14. Moreno, I., Romá-Ferri, M.T., Moreda, P.: Carmen: Sistema de entity typing basado en perfiles [carmen: Entity typing system based on profiles]. In: Congreso informática para tod@s, IPT 2018, Madrid, Spain, April 19-20 (2018)
15. National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention: Disability overview (August 2017), https://www.cdc.gov/ncbddd/disabilityandhealth/disability.html, (last accessed May 27, 2018)
16. Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A.D., Casillas, A.: On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. Journal of Biomedical Informatics **56**, 318–332 (2015). https://doi.org/10.1016/j.jbi.2015.06.016
17. Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (may 2012)
18. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A.: Building a semantically annotated corpus of clinical texts. Journal of biomedical informatics **42**(5), 950–66 (oct 2009). https://doi.org/10.1016/j.jbi.2008.12.013
19. World Health Organization: World report on disability (2011), http://www.who.int/disabilities/world_report/2011/report/en/, (last accessed May 27, 2018)