Overview of the DIANN Task: Disability Annotation Task

Hermenegildo Fabregat[1 [0000−0001−9820−2150]], Juan Martinez-Romo[1,2 [0000000269057051]], and Lourdes Araujo[1,2 [0000−0002−7657−4794]]

[1] Universidad Nacional de Educación a Distancia (UNED), Department of Computer Science, Juan del Rosal 16, Madrid 28040, Spain http://nlp.uned.es/

[2] IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

Abstract. The DIANN task consists of the detection of disabilities in English and Spanish texts, as well as the detection of negated disabilities. The organizers have proposed a task with different elements concerning both, the language and the entities to be detected (disabilities, negation and acronyms). Two evaluation criteria have also been used: exact and partial. All these options have generated a large number of results and different classifications. This overview summarizes the participation of eight teams, all of them with results for both English and Spanish, totaling 37 runs (18 for English and 19 for Spanish).

Keywords: Biomedical Entity recognition · Biomedical corpus · Disability recognition · Negation detection

1 Introduction

Natural language processing techniques can be very useful for the biomedical domain, due to the large amount of unstructured information that it generates. There are many topics that have been addressed due to their great impact, for example the search of entities in medical texts, such as diseases, drugs and genes. A particular type of entity that has not been specifically considered is disabilities. There exist some tools for the annotation of medical concepts, especially in English, such as Metamap[3], and also some others that can be adapted for the annotation of some medical concepts in Spanish, such as Freeling-Med[19]. However, none of them consider terms such as disabilities as a distinctive concept. According to World Health Organization[18], the term disability refers to an umbrella term covering impairments, limitations of activities and restrictions on participation. The automatic processing of documents related to disabilities is an interesting research area if we take into account that, world health organization estimates that about 15% of the population suffers from some kind of disability. The task of detecting disabilities is a challenge that involves difficulties such as the freestyle used to write them. They can be mentioned using specific words, such as “blindness”, and also using descriptions such as “visual impairment”. Disabilities can also be mentioned in the presence of negation words, as

in "...had no expressive language". Given the relevance of this problem, as well as its difficulty, the goal of DIANN's task is to automate the mining process of research articles that mention disabilities in a multilingual scenario.
The remainder of this paper is organized as follows. Section 2 presents the task. Section 3 describes the datasets we released for training and test and the evaluation criteria. Section 4 summarizes the proposed approaches of the participants. Section 5 presents and discusses the results. Finally, conclusions are presented in Section 6.

2 Task Description

DIANN is a named entity recognition task which focuses on disability identification in biomedical research texts. As far as we know, the recognition of disabilities has not been addressed previously. So far, systems oriented to the detection of named entities in biomedical texts have not treated the concept of disability as an isolated entity, categorizing it in most cases as a disease or symptom (they do not make a clear distinction between a disability and a symptom or disease). This task aims to deal specifically with this kind of entities. We have compiled a collection of documents in English and Spanish, which has been annotated manually by three people. Due to the ambiguity present in the disability concept, the support of expert medical staff has been necessary during the annotation process. This corpus has been used to evaluate the performance of various named entity recognition systems in two different languages, Spanish and English. In addition to disabilities, negation has been annotated when it affects one or more disabilities. The rest of the negations presented in the corpus have not been annotated.
The corpus was divided into two parts, one for training and the other for test. To contextualize the problem, in addition to the training corpus, we provide a list of categories for the different disabilities identified in both Spanish and English languages. According to the scheduling specifications of the task, participants had one month to develop their systems since the publication of the training corpus. Then, we released the test set without annotations and participants had fifteen days to send their results to the task organizers. We indicated to each team of participants that they could present up to three different approaches per language. This document presents the evaluation of the different submissions in three categories (disability recognition, negated disability recognition and joint) through two different evaluation criteria, partial matching and exact matching.

3 Data and Evaluation

In this section we discuss the origin and characteristics of the dataset used in this task as well as the format in which it has been presented. We also discuss the methods or criteria used to evaluate the participant systems.

3.1 Data

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. The document search process has been restricted to documents with the abstract in both, English and Spanish languages, and at least contain a disability in both languages.

English data	Docs	Sents	Toks
Training	400	4782	70919
Test	100	1309	18406

Spanish data	Docs	Sents	Toks
Training	400	4639	78381
Test	100	1284	20567

Table 1: Number of articles (docs), sentences (sents) and tokens (toks) in each dataset

English data	Dis	Neg	Neg-Dis
Training	1413	40	42
Test	243	23	24

Spanish data	Dis	Neg	Neg-Dis
Training	1326	40	41
Test	229	22	23

Table 2: Number of disabilities (dis), negations (neg) and negated disabilities (neg-dis) in each dataset

The DIANN corpus was divided into two disjointed parts: training set (80%) and test set (20%). Table 1 and table 2 summarizes for both languages the size of the training and test sets and the data contained in them.

3.2 Format and Distribution

The dataset is structured in directories. Each folder corresponds to a specific language and contains the documents named with the associated PUBMED identifier. Each document is presented following an XML annotation format. For the disability annotations, the tag `<dis>` has been used:

> Fragile-X syndrome is an inherited form of `<dis>`mental retardation`</dis>` with a connective tissue component involving mitral valve prolapse.

The negation trigger and its scope, has been annotated using the tags `<neg>` and `<scp>`:

> In the patients `<scp><neg>`without`</neg>` `<dis>`dementia`</dis></scp>`, significant differences were obtained in terms of functional and cognitive status (Barthel index of 52.3438 and Pfeiffer test with an average score of 1.48 $\pm$ 3.2 (P<.001)).

The corpus is available in the following url: `https://github.com/gildofabregat/DIANN-IBEREVAL-2018/tree/master/DIANN_CORPUS`

3.3 Evaluation

In addition to the exact matching and due to the freedom with which a disability can be expressed, we have used a second evaluation criteria to compare the different systems. This second evaluation criteria, called partial matching, is based on the concept of core-term match introduced in [9]. To use this evaluation approach, as you can see below (annotation → annotation core), we have manually generated a file with the core of each annotation of the corpus.

irreversible visual loss → visual loss

moderate to severe dementia → dementia

severe mental disorder → mental disorder

For each evaluation criteria, the performance is measured with $F_{\beta=1}$ rate:

$$F_\beta = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \quad (1)$$

where precision is the percentage of named entities found by the system that are correct or partially correct and recall is the percentage of named entities present in the corpus that are found or partially found by the system.

4 Overview of the Submitted Approaches

A total of eight teams have participated, adding up to nineteen runs for English and twenty runs for Spanish. Although each document was presented in English and Spanish, none of the participating teams has exploited bilingualism. It may be because the abstracts of both languages are not parallel. They are written by the authors (they are not automatic translations) and sometimes contain different numbers of sentences and different numbers of disabilities.
In this section we explain the different approaches tested by each of the teams.

- The SINAI_18[12] team proposed different approaches for each language considered in the task. On the one hand, for English language, they have used Metamap and NegEx[6] to annotate concepts and to analyze negation; on the other hand, for Spanish language, they have used their own UMLS-based entity recognition system and, in the case of negation, they have used a method

based on bags of words. This second approach makes use of NLTK[4] to standardize text and CoreNLP[13] to perform syntactic analysis. Finally, in both approaches, after the recognition of the concepts, they perform both a filtering process based on semantic information of the identified concepts and the calculation of the similarity of each UMLS concept filtered with the term "disability" using word2vec[15].

- IxaMed[10] team presented a pipeline composed of a combination of multiple systems. First, they make use of a neural network-based architecture system for disability detection consisting of a Bidirectional Long Short Term Memory network (Bi-LSTM) and a Conditional Random Field (CRF) at the top. For English, this system uses Brown clusters[5] and word embeddings extracted from the MIMIC-III corpus[11]. For Spanish, they calculated the word-embeddings from Electronic Health Records and they did not include any Brown cluster. After disability detection, they have used a rule-based system for the detection of triggers associated with disabilities making use of a generic list of negation triggers. In addition and using a similar rule-based approach, they have designed a third module for the detection of disability-related abbreviations. Finally, and taking into account the aforementioned processes, they have designed a system based on neural networks for the identification of the negation scope.

- The IXA_18_02[2] team presented one run for each language, both using the same entity recognition system. Known as ixa-pipe-nerc[1], this system aims to recognize named entities avoiding any linguistic motivated feature. The system makes use of typographic and morphological features of the text. Ixa-pipe-nerc makes use of the implementation of the Perceptron algorithm contained in the Apache OpenNLP project, incorporating the use of several features based on language representations such as Brown clusters taking the 4th, 8th, 12th and 20th node in the path; Clark clusters[7] and word2vec clusters.

- The work presented by the GPLSIUA[17] team consists of the use of its own general purpose automatic learning system called CARMEN[16] for disability annotation and a dictionary-based approach to negation detection. The annotation of disabilities has been divided into two modules, the first of which deals with the process of generating candidate expressions based on the extraction of noun phrases and the second one, based on the use of CARMEN, deals with the process of determining which of the candidate expressions can be considered as a disability. CARMEN consists of a machine learning system that makes use of Random Forest and is trained with syntactic and distributional features.

- UPC_018_3[14] team has presented two semi-supervised approaches for the task of recognition of the named entity. The first one is a conditional random

field model trained with syntactic features. The second one is a recurrent neural network using Bi-LSTM network and a CRF layer. As they explain, they have made use of a process to reduce possible over-fitting based on the addition of new unlabeled abstracts. Finally, to process the negation and its scope they have made use of a system called ABNER[20] based on CRF.

- The system presented by the UPC_018_2[21] team makes use of a CRF to annotate named entities. This system has been trained using both syntactic and some semantic features. For this purpose, they have also used a list of terms that appeared in the annotations that occurred in the training corpus, as a result of being an attribute associated with the inclusion or not in the list extracted from each term to be analyzed. Finally, they have used a NegEx-based system for negation detection, filtering the total of annotations to those where the negation trigger is less than 4 words away from the possible negated disability.

- The UC3M_018_1[22] team has submitted a proposal for each language based on the same architecture. The models presented use a two-phase architecture based on two layers of Bidirectional LSTM to capture the context information and a CRF to obtain the correlation of the information between the labels. Finally, they have jointly addressed entity detection and negation detection, making this approach a sequence to sequence (seq2seq) multi-proposal classification problem.

- Finally, the LSI_UNED[8] team presented an unsupervised approach to disability annotation that involves a process of generating variants and using lists of disabilities and body functions. The system extracts the noun phrases and creates their possible variants. To find the best candidate and taking into account the lists mentioned above, the total number of variants is filtered according to metrics such as centrality and variation. Finally, for both, the detection of negation in Spanish and for the detection of abbreviations in both languages, the system uses post-processing based on regular expressions. For detection of negation in English, the system uses NegEx.

5 Results and Discussion

In this section, we discuss the results for both languages based on three categories[3]: detection of disabilities, detection of only negated disabilities, and detection of both, negated and non-negated disabilities.

1. Disability recognition. These results correspond to the evaluation of the annotations of the participants without taking into account the negation. This means that all annotated disabilities included in the dataset are evaluated regardless of whether negations have been or not correctly annotated.

[3] The results of each of the following tables are sorted according to the F_β obtained.

2. Negated disability recognition. These results correspond to the annotation of negated disabilities. That is, only disabilities that are affected by negation are taken into account in this evaluation. In addition to the success of the disability annotation, both the correctness of the negation trigger annotation and the correctness of the negation scope are taken into account.

3. Global results. Finally, these evaluation results correspond to the joint evaluation of the annotations relating to negated disabilities and the annotations relating to non negated disabilities.

Table 3 (exact matching) and the table 4 (partial matching) show the results obtained by the participants in the disability recognition task for both, Spanish (a) and English (b) languages. As can be seen, the IxaMed, UC3M_1 and UPC_3 teams have obtained the best results for the detection of disabilities in Spanish in both, partial and exact evaluation. These systems, based on a supervised approach (or semi-supervised, in the case of UPC_3), have in common the use of CRFs, being in the case of UPC_3 R1 and R2 systems based only on CRFs and in the rest of cases systems that use CRFs in the top layer of the proposed architecture (UPC_3 R3, IxaMed R1 and UC3M_1 R1 and R2). In the English scenario, the participants have not presented significant modifications respect to the approaches proposed for Spanish. The majority of these variations are modifications of the required resources, both in supervised and unsupervised approaches. Unsupervised approaches such as the one presented by LSI_UNED obtain notable improvements in the processing of documents in English, especially if we take into account the results of the partial evaluation.
The UPC_3 and IXA_2 teams have presented interesting solutions regarding the possible system over-fitting. The UPC_3 team has carried out a regularization process based on the incorporation of unannotated documents. If we take into account that the division into test and training has been generated trying to avoid the over adjustment of the systems due to the overlap between both sets, the consideration of an iterative learning scheme and the inclusion of new unannotated documents in the learning phase is a practice of great interest and that seems to have provided good results. The IXA_2 team, with a Perceptron-based model, has proposed the use of a set of shallow features to avoid any possible errors that might occur when processing the dataset with automatic text processing tools. This is of great interest if we consider that in the biomedical domain a specific terminology is used (disease names, abbreviations, drug names,...) which may not be included by these automatic processing tools and which may generate an accumulation of errors during the training phase. Regarding the annotation of acronyms, only the IxaMed, UPC_2 and LSI_UNED teams presented specific solutions for their annotation. Both IxaMed and LSI_UNED implemented solutions derived from the premise that an acronym is first presented at a maximum distance of X words from a disability. This way of dealing with acronym annotation is dependent on the accuracy of capturing the different disabilities. On the other hand, the UPC_2 team has used a boolean attribute to deal with whether a term or expression is part of a list of acronyms. In summary, all systems pro-

vide significant features to the task of entity detection. Regarding the results for the disability recognition task, the systems have performed better in general for English than for Spanish. In some cases, the difference between the results of the partial evaluation and the exact evaluation is very clear. However, in most cases the ranking for the systems is preserved.

With regard to the processing of negation, the approaches presented, as in the previous category, have been very diverse. While systems like UC3M_1 and IXA_2 deal with negation using the same entity detection system used in the entity annotation task (Neural Networks: IXA_2 - BiLSTM+CRF: UC3M_1 - CRF: UPC_3), others have used tools such as NegEx (English: SINAI, UPC_2 and LSI_UNED - Spanish: UPC_2), rule-based systems (Trigger detection for English and Spanish: IxaMed - Scope recognition for Spanish: LSI and GPLSI - Scope recognition for English: GPLSI) and lexicons, word bags, and so on (Trigger detection for English: GPLSI - Trigger detection for Spanish: GPLSI, SINAI, LSI_UNED). In most cases, the results obtained by the different systems show a strong relationship with the results of the disability detection task, with the GPLSIUA and SINAI teams in Spanish standing out. Although the systems have obtained very satisfactory results (table 5 and table 6), IxaMed, UPC_3 (R3, R1 and R2), IXA_2 (R1, R2 and R3) and UPC_2 stand out. Due to the size of the corpus and the criteria selected to consider a negation, few cases of negation have been included in the DIANN corpus, making it difficult to evaluate the significance of the negation detection in this task.

Finally, table 7 and table 8 show the results by jointly evaluating both the detection of disabilities and the recognition of negation. As you can see, these tables summarize the performance shown by the different systems. Due to the small number of negations, the results shown are strongly influenced by the results obtained in the detection of disabilities.

(a)

Spanish	P	R	F
IxaMed R1	0.757	0.817	0.786
UC3M_1 R2	0.818	0.646	0.722
UC3M_1 R1	0.801	0.651	0.718
UPC_3 R2	0.807	0.603	0.69
UPC_3 R1	0.814	0.594	0.687
UC3M_1 R3	0.801	0.563	0.662
IXA_2 R1	0.65	0.642	0.646
IXA_2 R3	0.636	0.655	0.645
UPC_3 R3	0.67	0.603	0.634
IXA_2 R2	0.641	0.616	0.628
UPC_2 R1	0.732	0.502	0.596
SINAI_1 R3	0.459	0.345	0.394
LSI_UNED R3	0.41	0.249	0.31
LSI_UNED R2	0.396	0.249	0.306
LSI_UNED R1	0.393	0.249	0.305
GPLSIUA_1 R1	0.813	0.17	0.282
GPLSIUA_1 R2	0.796	0.17	0.281
SINAI_1 R2	0.181	0.415	0.252
SINAI_1 R1	0.022	0.485	0.042

(b)

English	P	R	F
IxaMed R1	0.786	0.86	0.821
UC3M_1 R1	0.778	0.72	0.748
UC3M_1 R2	0.759	0.663	0.708
UC3M_1 R3	0.775	0.65	0.707
UPC_3 R1	0.799	0.605	0.689
UPC_3 R2	0.795	0.605	0.687
UPC_2 R1	0.756	0.56	0.643
UPC_3 R3	0.655	0.617	0.636
LSI_UNED R3	0.671	0.597	0.632
LSI_UNED R2	0.639	0.597	0.617
LSI_UNED R1	0.633	0.597	0.614
IXA_2 R1	0.701	0.531	0.604
IXA_2 R2	0.706	0.494	0.581
SINAI_1 R3	0.625	0.37	0.465
GPLSIUA_1 R2	0.884	0.251	0.391
GPLSIUA_1 R1	0.881	0.243	0.381
SINAI_1 R2	0.222	0.428	0.293
SINAI_1 R1	0.016	0.593	0.032

Table 3: Disability recognition - (a) Spanish (b) English) - Exact matching. Precision (P), Recall (R) and F-measure (F).

(a)

Spanish	P	R	F
IxaMed R1	0.822	0.886	0.853
UC3M_1 R1	0.882	0.716	0.79
UC3M_1 R2	0.878	0.694	0.776
UPC_3 R2	0.889	0.664	0.76
UPC_3 R1	0.898	0.655	0.758
IXA_2 R3	0.712	0.734	0.723
UC3M_1 R3	0.876	0.616	0.723
IXA_2 R1	0.721	0.712	0.716
UPC_3 R3	0.743	0.668	0.703
IXA_2 R2	0.705	0.677	0.69
UPC_2 R1	0.828	0.568	0.674
LSI_UNED R2	0.847	0.533	0.654
LSI_UNED R1	0.841	0.533	0.652
LSI_UNED R3	0.842	0.511	0.636
SINAI_1 R3	0.512	0.384	0.439
GPLSIUA_1 R2	0.959	0.205	0.338
GPLSIUA_1 R1	0.958	0.201	0.332
SINAI_1 R2	0.204	0.467	0.284
SINAI_1 R1	0.026	0.568	0.05

(b)

English	P	R	F
IxaMed R1	0.842	0.922	0.88
LSI_UNED R3	0.856	0.761	0.806
UC3M_1 R1	0.822	0.761	0.791
LSI_UNED R2	0.815	0.761	0.787
LSI_UNED R1	0.808	0.761	0.784
UC3M_1 R2	0.835	0.728	0.778
UC3M_1 R3	0.828	0.695	0.756
UPC_3 R1	0.875	0.663	0.754
UPC_3 R2	0.865	0.658	0.748
UPC_3 R3	0.742	0.7	0.72
UPC_2 R1	0.822	0.609	0.7
IXA_2 R1	0.761	0.576	0.656
IXA_2 R2	0.788	0.551	0.649
SINAI_1 R3	0.688	0.407	0.512
GPLSIUA_1 R1	0.94	0.259	0.406
GPLSIUA_1 R2	0.913	0.259	0.404
SINAI_1 R2	0.252	0.486	0.332
SINAI_1 R1	0.019	0.704	0.038

Table 4: Disability recognition - (a) Spanish (b) English) - Partial matching. Precision (P), Recall (R) and F-measure (F).

(a)

Spanish	P	R	F
IxaMed R1	0.889	0.727	0.8
IXA_2 R1	1	0.545	0.706
IXA_2 R2	0.929	0.591	0.722
IXA_2 R3	0.923	0.545	0.686
UPC_2 R1	0.737	0.636	0.683
UPC_3 R3	0.688	0.5	0.579
UPC_3 R1	0.647	0.5	0.564
UPC_3 R2	0.647	0.5	0.564
SINAI_1 R3	0.667	0.091	0.16
SINAI_1 R2	0.333	0.045	0.08
GPLSIUA_1 R1	0	0	0
GPLSIUA_1 R2	0	0	0
LSI_UNED R1	0	0	0
LSI_UNED R2	0	0	0
LSI_UNED R3	0	0	0
SINAI_1 R1	0	0	0
UC3M_1 R1	0	0	0
UC3M_1 R2	0	0	0
UC3M_1 R3	0	0	0

(b)

English	P	R	F
UPC_3 R1	0.773	0.739	0.756
UPC_3 R2	0.773	0.739	0.756
UPC_3 R3	0.696	0.696	0.696
GPLSIUA_1 R1	0.647	0.478	0.55
UPC_2 R1	0.647	0.478	0.55
GPLSIUA_1 R2	0.611	0.478	0.537
IXA_2 R1	0.667	0.435	0.526
IXA_2 R2	0.75	0.391	0.514
SINAI_1 R3	0.526	0.435	0.476
IxaMed R1	0.476	0.435	0.455
SINAI_1 R2	0.306	0.478	0.373
SINAI_1 R1	0.25	0.391	0.305
LSI_UNED R2	0.188	0.13	0.154
LSI_UNED R3	0.188	0.13	0.154
LSI_UNED R1	0.176	0.13	0.15
UC3M_1 R1	0	0	0
UC3M_1 R2	0	0	0
UC3M_1 R3	0	0	0

Table 5: Negated disability recognition - (a) Spanish (b) English) - Exact matching. Precision (P), Recall (R) and F-measure (F).

(a)

Spanish	P	R	F
IxaMed R1	1	0.818	0.9
UPC_3 R3	1	0.727	0.842
UPC_2 R1	0.895	0.773	0.829
UPC_3 R1	0.941	0.727	0.821
UPC_3 R2	0.941	0.727	0.821
UC3M_1 R3	1	0.682	0.811
IXA_2 R3	1	0.591	0.743
IXA_2 R2	0.929	0.591	0.722
IXA_2 R1	1	0.545	0.706
UC3M_1 R2	0.909	0.455	0.606
SINAI_1 R3	1	0.136	0.24
UC3M_1 R1	1	0.136	0.24
LSI_UNED R1	0.75	0.136	0.231
LSI_UNED R2	0.75	0.136	0.231
LSI_UNED R3	0.75	0.136	0.231
SINAI_1 R2	0.667	0.091	0.16
GPLSIUA_1 R1	0.5	0.091	0.154
GPLSIUA_1 R2	0.4	0.091	0.148
SINAI_1 R1	0.125	0.045	0.067

(b)

English	P	R	F
IxaMed R1	1	0.913	0.955
UPC_3 R1	0.955	0.913	0.933
UPC_3 R2	0.955	0.913	0.933
UPC_3 R3	0.913	0.913	0.913
SINAI_1 R3	1	0.826	0.905
GPLSIUA_1 R1	0.941	0.696	0.8
UPC_2 R1	0.941	0.696	0.8
IXA_2 R1	1	0.652	0.789
GPLSIUA_1 R2	0.889	0.696	0.78
UC3M_1 R3	1	0.609	0.757
LSI_UNED R2	0.875	0.609	0.718
LSI_UNED R3	0.875	0.609	0.718
LSI_UNED R1	0.824	0.609	0.7
IXA_2 R2	1	0.522	0.686
SINAI_1 R1	0.556	0.87	0.678
SINAI_1 R2	0.556	0.87	0.678
UC3M_1 R2	0.875	0.304	0.452
UC3M_1 R1	1	0.043	0.083

Table 6: Negated disability recognition - (a) Spanish (b) English) - Partial matching. Precision (P), Recall (R) and F-measure (F).

(a)

Spanish	P	R	F
IxaMed R1	0.746	0.795	0.77
UC3M_1 R1	0.769	0.568	0.653
UPC_3 R2	0.772	0.563	0.652
UPC_3 R1	0.779	0.555	0.648
UC3M_1 R2	0.749	0.559	0.64
IXA_2 R1	0.644	0.616	0.629
IXA_2 R3	0.626	0.629	0.627
UC3M_1 R3	0.731	0.546	0.625
IXA_2 R2	0.633	0.594	0.613
UPC_3 R3	0.64	0.559	0.597
UPC_2 R1	0.71	0.48	0.573
SINAI_1 R3	0.411	0.284	0.336
LSI_UNED R3	0.424	0.245	0.31
LSI_UNED R2	0.409	0.245	0.306
LSI_UNED R1	0.406	0.245	0.305
SINAI_1 R2	0.157	0.349	0.217
GPLSIUA_1 R1	0.692	0.118	0.201
GPLSIUA_1 R2	0.659	0.118	0.2
SINAI_1 R1	0.018	0.402	0.035

(b)

English	P	R	F
IxaMed R1	0.746	0.811	0.777
UC3M_1 R1	0.749	0.626	0.682
UPC_3 R1	0.772	0.584	0.665
UPC_3 R2	0.768	0.584	0.664
UC3M_1 R3	0.712	0.609	0.656
UC3M_1 R2	0.706	0.572	0.632
LSI_UNED R3	0.657	0.568	0.609
UPC_3 R3	0.626	0.593	0.609
UPC_2 R1	0.724	0.519	0.604
LSI_UNED R2	0.624	0.568	0.595
LSI_UNED R1	0.616	0.568	0.591
IXA_2 R1	0.672	0.49	0.567
IXA_2 R2	0.685	0.457	0.548
SINAI_1 R3	0.573	0.337	0.425
GPLSIUA_1 R2	0.806	0.239	0.368
GPLSIUA_1 R1	0.812	0.23	0.359
SINAI_1 R2	0.199	0.395	0.264
SINAI_1 R1	0.015	0.543	0.029

Table 7: Negated and no negated disability recognition - (a) Spanish (b) English) - Exact matching. Precision (P), Recall (R) and F-measure (F).

(a)

Spanish	P	R	F
IxaMed R1	0.82	0.873	0.846
UC3M_1 R3	0.889	0.664	0.76
UPC_3 R2	0.88	0.642	0.742
UC3M_1 R2	0.865	0.646	0.74
UPC_3 R1	0.89	0.633	0.74
UC3M_1 R1	0.864	0.638	0.734
IXA_2 R3	0.7	0.703	0.702
IXA_2 R1	0.708	0.677	0.692
UPC_3 R3	0.735	0.642	0.685
IXA_2 R2	0.693	0.651	0.671
UPC_2 R1	0.819	0.555	0.661
LSI_UNED R2	0.803	0.48	0.601
LSI_UNED R1	0.797	0.48	0.599
LSI_UNED R3	0.803	0.463	0.587
SINAI_1 R3	0.468	0.323	0.382
GPLSIUA_1 R2	0.878	0.157	0.267
GPLSIUA_1 R1	0.897	0.153	0.261
SINAI_1 R2	0.18	0.402	0.249
SINAI_1 R1	0.022	0.48	0.042

(b)

English	P	R	F
IxaMed R1	0.841	0.914	0.876
LSI_UNED R3	0.843	0.728	0.781
UC3M_1 R3	0.832	0.712	0.767
LSI_UNED R2	0.801	0.728	0.763
LSI_UNED R1	0.79	0.728	0.758
UPC_3 R1	0.87	0.658	0.749
UPC_3 R2	0.859	0.654	0.743
UC3M_1 R2	0.817	0.663	0.732
UC3M_1 R1	0.803	0.671	0.731
UPC_3 R3	0.735	0.695	0.715
UPC_2 R1	0.822	0.588	0.686
IXA_2 R1	0.757	0.551	0.638
IXA_2 R2	0.784	0.523	0.627
SINAI_1 R3	0.685	0.403	0.508
GPLSIUA_1 R1	0.942	0.267	0.417
GPLSIUA_1 R2	0.903	0.267	0.413
SINAI_1 R2	0.242	0.481	0.322
SINAI_1 R1	0.019	0.691	0.037

Table 8: Negated and no negated disability recognition - (a) Spanish (b) English) - Partial matching. Precision (P), Recall (R) and F-measure (F).

6 Conclusions

In this edition of Ibereval a new task of disabilities identification in biomedical research papers has been proposed. In spite of being the first edition of the task, we consider that it has been a success of participation with a total of 8 participants. The corpus that has been made available to the participants is a very interesting resource since it is a dataset of 1000 annotated abstracts, 500 in Spanish and 500 in English, all of them extracted from journal articles related to the biomedical area and each of them referring to at least one disability, both in its extended form or as an abbreviation. In addition to containing annotations of disabilities, the corpus contains annotations referring to negation when it affects at least one disability.

The participants used different approaches or resources that provided the task with different perspectives, all of which were very interesting. In summary, for each of the languages, the participants have not changed their models too significantly; in most cases, they have made use of alternative resources adapted to the language in which they work. In the case of Spanish, the systems with the best results have been supervised or semi-supervised systems, based on neural network models using a Bidirectional LSTM and CRF, and in the case of English, the use of neural networks has also been predominant among the best systems, although in this case there are unsupervised systems that have obtained a performance equal or higher than the previous ones. Regarding negation, many participants have adapted well known systems such as NegEx or ABNER, although there have also been some participants who have implemented their own negation detection systems based on rules or treating the problem like a classification problem.

In conclusion, the organizers have proposed a task with different elements concerning both, the language and the entities to be detected (disabilities, negation and acronyms). Two evaluation metrics have also been used: exact and partial. All these options have generated a large number of results and different classifications, highlighting the differences between the participating systems according to the aspect taken into account.

7 Acknowledgments

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

References

1. Agerri, R., Bermudez, J., Rigau, G.: Ixa pipeline: Efficient and ready to use multilingual nlp tools. In: LREC. vol. 2014, pp. 3823–3828 (2014)
2. Agerri, R., Rigau, G.: Simple language independent sequence labelling for the annotation of disabilities in medical texts. In: Proceedings of the Third Workshop

on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
3. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium. p. 17. American Medical Informatics Association (2001)
4. Bird, S., Loper, E.: Nltk: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. p. 31. Association for Computational Linguistics (2004)
5. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational linguistics **18**(4), 467–479 (1992)
6. 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)
7. Clark, A.: Combining distributional and morphological information for part of speech induction. In: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1. pp. 59–66. EACL '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003). https://doi.org/10.3115/1067807.1067817, https://doi.org/10.3115/1067807.1067817
8. Fabregat, H., Martinez-Romo, J., Araujo, L.: Uned at diann 2018: Unsupervised system for automatic disabilities labeling in medical scientific documents. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
9. Fukuda, K.i., Tsunoda, T., Tamura, A., Takagi, T., et al.: Toward information extraction: identifying protein names from biological papers. In: Pac symp biocomput. vol. 707, pp. 707–718 (1998)
10. Gonaega, I., Atutxa, A., Gojenola, K., Casillas, A., de Ilarraza, A.D., Ezeiza, N., Oronoz, M., Prez, A., de Viaspre, O.P.: A hybrid approach for automatic disability annotation. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
11. Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**, 160035 (2016)
12. Lpez-beda, P., Daz-Galiano, M.C., Martn-Valdivia, M.T., Jimnez-Zafra, S.: Sinai at diann - ibereval 2018. annotating disabilities in multi-language systems with umls. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
13. 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 the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
14. Medina, S., Turmo, J., Loharja, H., Padr, L.: Semi-supervised learning for disabilities detection on english and spanish biomedical text. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
16. Moreno, I., Rom-Ferri, M., Moreda, P.: Carmen: Sistema de entity typing basado en perfiles [carmen: Entity typing system based on profiles]. In: Congreso informtica para tod@s, IPT 2018 (2018)

17. Moreno, I., Rom-Ferri, M., Moreda, P.: Gplsiua team at the diann 2018 task. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
18. Organization, W.H., et al.: World report on disability: World health organization (2011)
19. Oronoz, M., Casillas, A., Gojenola, K., Perez, A.: Automatic annotation of medical records in spanish with disease, drug and substance names. In: Iberoamerican Congress on Pattern Recognition. pp. 536–543. Springer (2013)
20. Settles, B.: Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics **21**(14), 3191–3192 (2005)
21. Vecino, P.A., Padr, L.: Basic crf approach to diann 2018 shared task. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
22. Zavala, R.M.R., Martinez, P., Segura-Bedmar, I.: A hybrid bi-lstm-crf model to disabilities named entity recognition. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)