Deep neural models for extracting entities and relationships in the new RDD corpus relating disabilities and rare diseases

Hermenegildo Fabregat[a], Lourdes Araujo[a,b,*], Juan Martinez-Romo[a,b]

[a] *Universidad Nacional de Educación a Distancia (UNED), Department of Computer Science, Juan del Rosal 16, Madrid 28040, Spain*
[b] *IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain*

Abstract

Background and Objective: There is a huge amount of rare diseases, many of which have associated important disabilities. It is paramount to know in advance the evolution of the disease in order to limit and prevent the appearance of disabilities and to prepare the patient to manage the future difficulties. Rare disease associations are making an effort to manually collect this information, but it is a long process. A lot of information about the consequences of rare diseases is published in scientific papers, and our goal is to automatically extract disabilities associated with diseases from them.

Methods: This work presents a new corpus of abstracts from scientific papers related to rare diseases, which has been manually annotated with disabilities. This corpus allows to train machine and deep learning systems that can automatically process other papers, thus extracting new information about the relations between rare diseases and disabilities. The corpus is also annotated with negation and speculation when they appear affecting disabilities. The corpus has been made publicly accessible.

Results: We have devised some experiments using deep learning techniques to show the usefulness of the developed corpus. Specifically, we have designed a long short-term memory based architecture for disabilities identification, as well as a con-

*Corresponding author: Lourdes Araujo
Email addresses: gildo.fabregat@lsi.uned.es (Hermenegildo Fabregat), lurdes@lsi.uned.es, Tel. 34-913797318 (Lourdes Araujo), juaner@lsi.uned.es (Juan Martinez-Romo)

volutional neural network for detecting their relationships to diseases. The systems designed do not need any preprocessing of the data, but only low dimensional vectors representing the words.

Conclusions: The developed corpus will allow to train systems to identify disabilities in biomedical documents, which the current annotation systems are not able to detect. The system could also be trained to detect relationships between them and diseases, as well as negation and speculation, that can change the meaning of the language. The deep learning models designed for identifying disabilities and their relationships to diseases in new documents show that the corpus allows obtaining an F-measure of around 81% for the disability recognition and 75% for relation extraction.

Keywords:

Biomedical corpora, Rare diseases, Disabilities, Deep neural networks, Entity recognition, Relationship classification

---

1. Introduction

According to the European legislation, a disease or disorder is defined as rare when it affects fewer than 5 in 2000 people. The number of rare diseases (RD) that are registered in official agencies is huge. Orphanet[1], the international database and portal on RDs and orphan drugs, has currently registered about 15,000 RDs. Many of these diseases involve different disabilities. Therefore, it is extremely important to know in advance the evolution of a disease in order to limit and prevent the appearance of disabilities and to prepare the patient's environment to manage the difficulties and needs of his daily life. Orphanet, is actually, collecting information [5] to improve the knowledge and visibility of disabilities associated with diseases, and to provide tools to help the affected people. In particular, they have indexed the functional consequences of each RD with the Orphanet Functioning Thesaurus[2], adapted from the "Activities and participation" and "Environmental factors" domains of the International Classification

[1] http://www.orpha.net

[2] http://www.orpha.net/orphacom/cahiers/docs/GB/Orphanet_Functioning_Thesaurus_EN.pdf

of Functioning, Disability and Health-Children & Youth version (ICF-CY [30]). It is done through a questionnaire sent to medical experts, disability specialists and patient organizations. According to the cited work, 857 RDs are already indexed and 540 more are in progress, with the contribution of hundreds of people and organizations from 43 countries. However, this manual process is highly time-consuming and expensive, and therefore it is desirable to perform it as automatically as possible.

Because of this, we propose to automate the process by mining from scientific papers the disabilities associated with diseases. Annotated corpora are required to develop such automatic extraction information systems. These corpora allow to train the systems and also to evaluate them. We have collected a corpus, RDD (Rare Disease-Disabilities), composed of scientific abstracts of articles related to some RD. The annotation includes disabilities, negation, speculation and also relationships between RDs and disabilities. We have used the Orphanet Functioning Thesaurus as the base of our annotation criteria, considering as disabilities the lack, delay or loss of motor skills, abilities for understanding, communicating with other, interpersonal relations, daily activities, social life, as well as limitation or delay in growth. Two kinds of disability expressions have been annotated in the corpus. On the one hand we have annotated expressions including words associated with disabilities. Examples are "progressive deafness", "congenital blindness", "cortical blindness" and "spinocerebellar ataxia". On the other hand, we have also annotated disabilities describing problems in individual functioning. Some examples are "hearing loss", "peripheral visual field loss", and "profound mental retardation". The list of disabilities resulting from the annotation process have been reviewed by two doctors.

We have used the developed corpus to evaluate two models for extracting both named entities and relationships respectively. These models use deep neural networks for learning. Deep learning (DL) with neural networks (NN) has become one of the most active areas of research in Natural Language Processing (NLP) [4]. It has been applied to many NLP tasks such as text classification, machine translation, etc. The great advantage of deep learning is that it requires much less feature engineering than traditional machine learning methods. DL networks use low dimensional vectors or embeddings [18] to represent the input information required to perform the task consid-

ered. Among the successful applications of DL to NLP are relationships classification and named entity recognition (NER).

Named entity recognition (NER) is paramount for extracting information from biomedical documents. This task, also known as concept identification, amounts to identifying terms of interest and mapping them to a set of pre-defined semantic categories. In our case these categories are diseases and disabilities. Most NER methods are supervised and their performance have improved a lot as more annotated corpora have become available [41]. Between the learning methods applied to NER, deep learning with neural networks [19] is leading to reach new levels of results [11, 23].

The classification of relationships between entities in documents is another key process in information extraction. It has been applied to different problems in the biomedical domain such as interactions between drugs and genes [33], between drugs and adverse affects [23], protein-protein interaction [1], etc. It has often been treated as a process that takes the identification of entities for granted. It has been applied to many different sets of entities, traditionally with machine learning methods for automatic classification. Deep learning techniques have also led to an explosion of work devoted to this problem [25, 43, 23].

Two main kinds of deep NNs (DNNs) have been considered for the tasks addressed in this work: recurrent neural networks (RNNs) [9] and convolutional neural networks (CNNs) [20]. RNNs are sequential architectures able to process arbitrary sequences of inputs, i.e. they have been designed to recognize patterns in sequences of data, such as text, handwriting, the spoken word, etc. Long short-term memory (LSTM) [13] is a particular case of RNN. These DNNs have the capability of "remembering" values over arbitrary time intervals, and therefore they are appropriate to process and predict time series given sequences of labels of unknown size. Their sequential nature have led to apply them to sequence modeling NLP problems which require to take the context into account.

CNNs, the other kind of DNN used in many NLP tasks [4], such as chunking, part-of-speech (POS) tagging or semantic role labeling, presents a hierarchical architecture. While traditional NN connect each input neuron to each output neuron, CNNs use convolutions over the input to compute the output. This way, CNNs establish local

connections linking regions of the input to an output neuron. They have often been applied to classification tasks, such as sentiment classification.

In this work we propose deep learning based methods to deal with the new problems of recognizing disabilities and their relationships to diseases. We have used a bi-directional (Bi) LSTM for disability and disease recognition, whereas we have designed a CNN for extracting relationships between disabilities and diseases. We have used Keras [3] for the implementation of both deep learning models. It is a high-level neural networks API, written in Python. Results show a high performance of both tasks, that is similar to the one achieved in other similar biomedical tasks, although the detection of disabilities presents more complex aspects due to the great freedom with which these concepts can be mentioned. Specifically, we achieve an F-measure greater than 81% for the disability recognition problem, and greater than 75% for the relation extraction problem using a CNN.

2. Related Work

Several annotated corpora have been developed in different biomedical domains. Most of them are focused on gene and protein annotations [15, 37]. Other corpora have been annotated for drugs and diseases. The BioText corpus [35] provides annotations for several relationships between disorders and treatments. It is composed of 100 titles and 40 abstracts from Medline 2001. The treatments include both, drugs and medical treatments. The annotations have been performed at the sentence level, including both, positive and negative relationships. The EU-ADR corpus [39] is composed of 300 Medline abstracts and annotated with drugs, diseases, targets, and their relationships. The corpus was pre-annotated automatically and missed or incorrect annotations were manually corrected by three annotators. The corpus provides drug-disease relations, indicating whether a particular drug may produce an adverse effect. The ADE corpus [10] is composed of 2972 Medline case reports manually annotated with relations between drugs and conditions representing adverse reactions, by three annotators and later harmonized. Oronoz et al. [32] have created the IxaMed-GS corpus composed of real electronic health records written in Spanish and manually annotated by experts

in pharmacology and pharmacovigilance. The DDI corpus [12] has been manually annotated by two experts in pharmacovigilance with four entity types: drugs, brands, groups of drugs and substances not approved for human use. Drug-Drug Interactions have been also annotated.

Some works have considered the annotation of some grammatical phenomena that can change the meaning of the language. Negation and speculation are paramount aspects to understand the language [27, 36]. Dealing with negation requires to know if a part of the text has the opposite meaning. Speculation refers to the degree of certainty about the facts being stated. Operators expressing negation and speculation have a scope [26], i.e. the words in the sentence that are affected by the negation or speculation expression. These operators may interact to each other in complex ways, as in the sentence "The drug may not have the expected effect", and thus they represent a challenge in processing text, in general, and in the biomedical domain in particular.

There are fewer works devoted to detect speculation. Some corpora have been annotated with negation and speculation in other areas [16], but they are focused on different problems, such as sentiment analysis [6]. This lack of resources in the considered domain, enhances the relevance of the RDD corpus.

Many works have appeared recently applying DNNs to extract different entities from different corpus. Collobert et al. [4] proposed a feed-forward neural network to classify word labels by using contexts within a window with fixed size. Later, more complex kinds of DNNs have been applied to the problem. Chiu and Nichols [2] applied Bidirectional LSTMs and a CNN for NER, obtaining competitive results on the CoNLL-2003 dataset. Ma and Hovy [24] proposed a DNN architecture combining bidirectional LSTMs, CNNs and Conditional Random Fields (CRF) for sequence labeling problems. They evaluate the system on two datasets, Penn Treebank WSJ corpus for POS tagging and CoNLL 2003 corpus for NER. For both datasets they obtained state-of-the-art performance. DNNs have also been applied to NER in the biomedical domain. For example, Zhao et al. [45] proposed a CNN for disease entity recognition, providing competitive results for two evaluation corpus, NCBI Disease corpus, and BioCreative V Chemical Disease Relation Task corpus.

For relation classification, both kinds of DNNs, CNN and LSTM, have been ap-

plied. Miwa and Bansal [25] use bidirectional LSTMs to extract relations from the ACE2005 and ACE2006 corpus. The authors found some improvement over a CNN based model on nominal relation classification for the data from the SemEval-2010 Task 8. Zeng et al. [44] use a CNN to learn sentence level features, providing information such as the positions of the related entities. There have also been several works specifically focused on the extraction of relationships in the biomedical domain. Huang et al. [14] tackled the problem of drug-drug interaction extraction using bidirectional LSTMs. Li et al. [23] extract adverse drug events between drug and disease entities from the ADE corpus, as well as resident relations between bacteria and location entities from the Bacterial Biomedical Task (BioNLP Shared Task Workshop 2016). They propose a neural joint model for entity and relation extraction.

In this work we are dealing with the new problem in the biomedical domain of extracting disabilities as well as their relationships to diseases. We have applied the most successful kinds of DNNs to extract these new kinds of entities and relationships.

3. Dataset

In order to develop systems able to automatically annotate disabilities in biomedical texts, we have compiled a corpus of scientific abstracts related to RDs. We have relied on information provided by Orphanet, for both, obtaining information related to RDs, and information concerning the functioning consequences that can be associated with RDs.

3.1. Methodology

The manual annotation of the RDD corpus was developed along one year roughly. The process followed the next steps:

- *Compilation of the set of documents*: To collect the corpus we have started from the list of RD given by Orphanet. For each RD we have downloaded a maximum of 100 abstracts (in order to include more variety of RDs in the corpus) and then selected those including at least a disability expression. We have selected and annotated a total number of 1000 abstracts. The total number of words in the

corpus is 181157, i.e. each abstract comprises around 200 words. Before the annotation process we have performed some text normalization. Specifically, we have removed all the abstract additional data, such as date, authors and urls, and divided the text into sentences. We have used the GENIASS software[3] for splitting the text.

Each file in the corpus is assigned a name of the form:

diseaseID–documentID–pubmedCod.txt

where diseaseID is the Orphanet number assigned to the RD in the Orphanet list of RD[4]. We have to take into account that a document can mention more than one RD. DocumentID is an internal identifier used to sort the documents with the same diseaseID. Finally, pubmedCod is the PubMed code corresponding to the scientific paper containing the abstract, and allows retrieving the original document at any time.

- *Definition of annotation criteria*: According to the experts, disability is the umbrella term for impairments, activity limitations and participation restrictions, referring to the negative aspects of the interaction between an individual (with a health condition) and that individuals contextual factors (environmental and personal factors) [21, 31]. Therefore, we have considered that disabilities are or tend to be permanent, and also that they are severe enough to disrupt the normal development of daily life. The details of the annotation criteria are explained below in the section 3.2.

- *Manual annotation of the documents*: We have used BRAT [38] as annotation tool. It is an online collaborative tool freely available for annotation visualization and editing. Three people (computer science scientists) have participated in the annotation of the set of documents. Two different people have annotated each

[3] http://www.nactem.ac.uk/y-matsu/geniass/

[4] `http://www.orpha.net/orphacom/cahiers/docs/GB/List_of_rare_diseases_in_alphabetical_order.pdf`.

document independently. Then, the annotations were compared, correcting them and the guidelines if needed.

- *Verification of terminology associated with disabilities*: After the annotation process two doctors have checked the lists of the obtained disabilities. Using their indications, we reviewed the annotations to produce the final corpus.

3.2. Annotation guidelines for Disabilities

We have annotated disabilities, as well as negation and speculation affecting them. We describe the details of the annotation criteria in what follows:

achromatopsia	dementia	paraparesis
aphasia	diplegia	paraplegia
apraxia	dysarthria	paresis
ataxia	dysautonomia	quadriparesis
autism	dyskinesia	quadriplegia
autistic	dystonia-dyskinesia	tetraparesis
blindness	hemiparesis	tetraplegia
deafness	hyperactivity	
deaf-mutism	paralysis	

Table 1: List of specific disability terms obtained after the annotation process.

Disabilities. To annotate disabilities we have considered that they are described either by a term directly related to a disability, or by a human function that is absent or limited. In both cases we have considered that disabilities imply a permanent condition and that they are severe enough to disrupt the normal development of daily life. After the annotation, we have obtained the list of specific disability terms appearing in Table 1. Notice that we have included in the table only one word representing the disability, and not all its possible derivative words. For example, we have included *blindness* in the list, but not *blind*, even if it appears in the corpus annotations.

Concerning the second kind of disability expressions, we have relied on the Orphanet Thesaurus of Functioning, which in turn relies on the International Classifica-

ability to recognise face	gait	processing speed
academic	growth	psychiatric
activities of daily living	hearing	psychological
attention	intellectual	receptive vocabulary

Table 2: Extract of the list of functions used to express a disability which have been identified during the annotation process.

tion of Functioning, Disability and Health (ICF, WHO 2001) [29]. An extract of the list of functions that we have found involved in some disability expressions appears in Table 2. The complete list is available in the corpus documentation. A function usually present in people which is absent or limited in a severe degree is considered a disability.

For annotating expressions that represent human functions whose limitation can entail a disability we have considered variants of the functions included in the Orphanet Functioning Thesaurus. For example we have found the expression *ability to recognise faces*, which can be considered related to *interacting with other people*, included in the Orphanet Thesaurus. There are two exceptions, *development* and *growth*, that are not included in the Orphanet Thesaurus, but are included in the International Classification of Functioning, Disability and Health-Children & Youth version (ICF-CY [30]), which is the source of the Orphanet Functioning Thesaurus.

Table 3 shows an extract of the list of words that indicate the absence or limitation of a common human function. The complete list is available in the corpus documentation. A function appearing negated can also represent a disability. Because of this, some negation terms are included in the list of impairment terms.

aberration	delay	limitation
abnormal	detachment	limited
absence	deterioration	loss

Table 3: Extract of the list of impairment words obtained after the annotation process.

The annotation of disabilities specifies the word indicating absence or limitation and the function affected by this. Each disability is annotated with the XML tag $<$

$dis\ >$. This element has an attribute which is an identifier to distinguish different disabilities within the same sentence:

$$< dis\ \ id = [0-9] + \ >< /dis >$$

The absent or limited functions that give rise to a disability are annotated with the XML tag $< fun >$:

$$< fun >< /fun >$$

The impaired words of a disability are annotated with the XML tag $< imp >$:

$$< imp >< /imp >$$

The negation of a function can also be a disability. In this case we also use the XML tag $< imp >$ to mark the negation word affecting a function.

Now, we are going to show some examples of annotated sentences.

- *We report a high incidence of <dis id="0">bilateral sensorineural deafness </dis> in transplanted patients, which highlights the systemic nature of the disease.*

  In this case the annotated disability includes the modifier *bilateral sensorineural*.

- *Inactivation of CBS results in CBS-deficient homocystinuria more commonly referred to as classical homocystinuria, which, if untreated, results in <dis id="0"> <fun>mental</fun> <imp>retardation</imp></dis>, thromboembolic complications, and a range of connective tissue disorders.*

  In this case, the disability is expressed as an impairment (*retardation*) of a human function (*mental*).

- *Furthermore, we evaluated the similarities of SMS adult food-related behaviors to those with <dis id="0"> <fun>intellectual</fun> <imp>disability</imp> </dis> and found that SMS adults had more a <dis id="1">severe <fun>behavioral</fun> <imp>problems</imp></dis>.*

  This sentence contains two disabilities, both of them being impaired functions, and having a different identifier.

- *He was <dis id="0"><imp>unable</imp> to <fun>walk</fun></dis> and had <dis id="1"> <imp>no</imp> expressive <fun>language</fun></dis>.*

  Here we can see two disabilities expressed as the negation of a function. We have annotated negation as any other impairment term.

Acronyms and Abbreviations. During the annotation process, we have found several disabilities that have been assigned an acronym. The list of acronyms and abbreviations found during the annotation process is available in the corpus documentation. These acronyms are annotated as disabilities, provided that the extended form has appeared in the same abstract.

Negation. We only annotate negation when it affects a disability, i.e. negative words affecting diseases, drugs, etc. are ignored. We have annotated negation specifying both, the expression indicating negation, and the scope of the negation. The XML tag for the negation expression is $< neg >$ and for the scope is $< scp >$. The tag for the scope has an identifier as attribute to identify the different disabilities negated within the same sentence.

$$< scp\ \ id = [0-9] + \ >< neg >< /neg >< /scp >$$

For the negation annotation we have followed similar criteria to those of BioScope [40]. Disabilities phrases including a negative keyword are not necessarily annotated for negation, for example because they may be in a speculative form. The scope of a negated disability depends on the syntax. The scope usually covers the biggest affected phrase. Generally, the scope of negative auxiliaries, adjectives and adverbs starts right with the negation word and finishes at the end of the phrase, as in the next example:

> *A case of homocystinuria with lenticular subluxation was misdiagnosed as Marfan syndrome since the patient had <scp id="0"><neg>no</neg> apparent <dis id="0"><fun>mental</fun> <imp>impairment</imp></dis> </scp> and had had a negative neonatal screen for homocystinuria.*

The list of negation terms affecting disabilities that have been found during the annotation process includes *absence, absent, doesn't, except, negative, no, no evidence, none, not, rather than, with the exception of*, and *without*.

Speculation. We only annotate speculation when it affects a disability. Similarly to the negation annotation, for speculation we have annotated the expression indicating speculation and the corresponding scope. The XML tag for the speculative expression is $< spe >$ and for the scope is $< ssc >$. The tag for the scope has an identifier as attribute to identify the different speculations within the same sentence.

$$< ssc\ \ id = [0-9] + >< spe >< /spe >< /ssc >$$

We consider the minimal unit expressing doubt for marking the speculative words. However, there are cases in which the doubt is expressed by several words together. Then, all of them are annotated as the speculative expression.

The list of speculation terms affecting disabilities that have been found during the annotation process includes *and/or, apparent, appear, suggest, etc.*

4. Registering Relations Between Disabilities and RDs

We have also extracted the relations between RDs and disabilities explicitly mentioned in the corpus. The scope of a relationship is restricted to the sentence level. Relationships have been annotated between the disabilities annotated in the corpus and RD included in the Orphanet list of RD. These relationships are annotated in a separated file, following a similar format to the one adopted by the ADE corpus for drugs and adverse effects [10]. The format is as follows:

Column-1: PubMed-ID
Column-2: Sentence
Column-3: Rare Disease
Column-4: Begin offset of RD at sentence level
Column-5: End offset of RD at sentence level
Column-6: Disability
Column-7: Begin offset of disability at sentence level
Column-8: End offset of RD at sentence level
Column-9: Indicative of speculation
Column-10: Orphanet ID of RD

The first column corresponds to the PubMed code of the article the abstract belongs to. The next field is the sentence. The third column contains the RD present in the sentence, followed by the start and end positions of the ER in the sentence, in columns 4 and 5. Column 6 is devoted to the disability, followed by its start and end positions within the sentence. Column 9 indicates if the relationship is in speculative form (1) or not (0). Finally, column 10 indicated the Orphanet ID of the RD.

We have collected both positive and negative relationships, which are recorded in different files. Let us consider an example of a positive relationship:

> *21838783 | Coffin-Lowry syndrome is a syndromic form of mental retardation caused by mutations of the Rps6ka3 gene encoding ribosomal s6 kinase (RSK)2. | Coffin-Lowry syndrome | 1 | 22 | mental retardation | 46 | 64 | 0 | 192*

This relation associates mental retardation with Coffin-Lowry syndrome. If there are more than one relationship in the same sentence, all of them are included in the corpus as different entries.

We have marked as speculative those relationships that express a possibility.

We have also collected a file of negative relationships. These relations include sentences mentioning a RD and a disability, but without stating a relation between them. Negative relations also include negated relations.

5. Statistics

We now provide some data related to the RDD corpus. It is composed of 1000 Medline abstracts of papers related to RDs, containing 9657 sentences. These abstracts contain information about 578 different rare diseases. Some of them are mentioned more than 50 times whereas other are mentioned only once, being 1.8, slightly less than 2, the average number of mentions. The corpus contains 3678 annotations of disabilities. From them, 2792 are expressed as the impairment of a human function, while 886 are stated using some disability term. In 186 cases, the disability corresponds to an acronym. The corpus includes 90 negated disabilities, corresponding to 83 negation

annotations, since a negation can include more than one disability. The corpus also includes 194 annotations of speculation, that affect to 264 disabilities.

The physical function most often found impaired is hearing. Sight and motor skills are often found impaired too. The second function most frequently found impaired affects cognitive capacities. The third one is related to development. The most frequently mentioned disability is *ataxia*, related to motor skills. *Deafness* appears in second place, whereas *dementia*, related to problems in cognitive functions, is the third one. *Autism* and *blindness* are also very frequent.

Concerning the relationships between RDs and disabilities, we have identified 1251 positive relationships and 706 negative. From them, 86 are speculative in the positive set and 8 in the negative set. The files include relationships for 362 different RDs.

Table 4 shows the agreement results for the different annotations considered. We do not include annotations that differ in one or two characters in the disagreement counts. Disagreements include omissions as well as inexact matches, in which the spams of the two annotations coincide in some word but not in all of them. The agreement in the annotation of disabilities and negation is high. The speculative annotations seem to be the more difficult ones, since words such as *indicate* are sometime used to express speculation, but sometimes it seems that are used as an assertion. The scope of the speculation is also difficult to establish in some cases. The disagreement in the annotation of the relationships only correspond to the positive relations, and it is affected by the disagreement in the annotation of the disabilities.

Annotation	Agreement	Disagreement	%Agreement
Disabilities	3222	456	87%
Negation	78	5	93%
Speculation	134	60	67%
Relationships	971	280	77%

Table 4: Agreement data from the annotation process.

6. Deep Neural Networks for extracting disabilities and their relationships to diseases

The annotated RDD corpus will allow the development of automatic annotation systems to identify disabilities and their relationship to diseases. In this section we have designed DNN models for performing these tasks using the RDD corpus for training and evaluation.

6.1. Deep learning model for disabilities and diseases recognition

As it is usual in the field of biomedical entity recognition [17, 23], the problem is addressed as an identification of label sequences. We follow the standard IOB labeling scheme [34]. We use two entity types which are Disabilities (I) and Diseases (U). For each kind of entity the first word is denoted by B (B-Disability or B-Disease) and the remaining words, if any with I (I-Disability or I-Disease). O indicates that the word does not correspond to any of the kind of entity considered. For example, the words in the next sentence will be assigned the labels appearing in brackets:
Many (O) neurodevelopmental (BI) disorders (II) exhibit (O) syndromic (BU) obesity (IU) including (O) SMS (BU)

BI stands for Beginning-Disability, II for Inside-Disability, BU for Beginning-Disease, IU for Inside-Disease and O for Others.

Figure 1 shows the architecture of the proposed model. The network is fed with two features represented by embeddings: lower case words, and a capital letter feature. A common practice [4] in NER tasks, that we also apply, is to reduce the number of entries in the dictionary by transforming all the words to lower case. However, capital letters can be an indicator of the position of RDs. The presence of these entities may be relevant information for capturing the presence of disabilities. Therefore, in order to keep some upper case information lost by the lower case transformation, we use the CASE embedding representation. Specifically, we use a CASE feature to indicate if a word is lowercase, is all uppercase, had first letter capital, or had at least one non-initial capital letter. Thus our model has two entries, the word embedding representing the words, and the CASE embedding representation. We create a matrix for the tokens, and for the casing of the words. We map each token to its index in the word

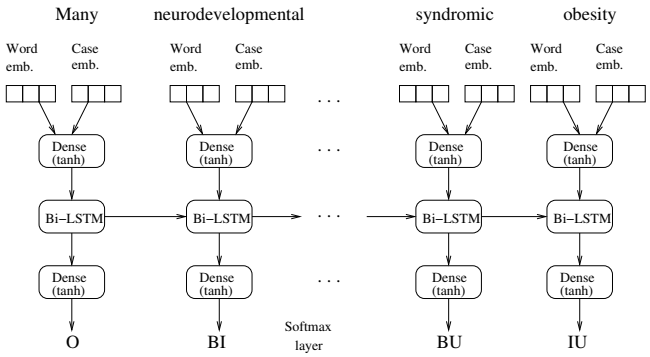

Figure 1: Deep learning model for disabilities and diseases recognition.

embeddings matrix and apply the same technique for word casing information, mapping the case information of each word to the index in the embedding lookup. We used pre-trained word embeddings [22] of size 300 to initialize our word embeddings. The *casing* embedding matrix is a hot-one encoding matrix of size 8. A *densely* connected hidden layer (Dense), with hyperbolic tangent (tanh) activation function, takes the concatenation of the embeddings as input. Then a bi-directional LSTM layer associates them a hidden state that captures the information of the current step, and also that in the previous steps. Finally, after another dense hidden layer, a softmax layer calculates the probabilities of all entity labels, what allows us to select the more appropriate label as output.

We have evaluated the model using the sentences extracted from the files of positive and negative relationships, because they contain annotations of both kinds on entities, disabilities and diseases. Similar to prior work [23] relations with nested gold annotations were removed (e.g., the RD "X-linked mental retardation" and the disability"mental retardation"). We evaluated our model using 10-fold cross-validation. The final results were displayed as macro-averaged scores. Some parameter values have been set following previous works [23]. We set the dimension of the bidirectional LSTM to 100. We have used AdaGrad as optimizer, setting the initial AdaGrad learning rate α and regularization parameter λ were set to 0.03 and 10^{-8}, respectively. The number of training iterations or epochs on each cross-validation fold has been set to 150. We have considered two different ways of evaluating. In the first one (evaluation

1) the tags 'B-' and I-' are only considered correct if they are in the correct sequence. The second one (evaluation 2) accounts for the matching of the different tags (excluding O) separately. We provide both, global results for both kinds of entities, and separate results for each kind of entity, RDs and disabilities.

	Evaluation 1			Evaluation 2		
	Prec.	Recall	F1	Prec.	Recall	F1
SVM(RD+DI)	72.36 (3.14)	52.49 (1.34)	60.82 (1.79)	75.08 (1.89)	65.94 (2.30)	70.20 (1.94)
SVM(RD)	62.07 (4.76)	52.83 (4.57)	57.03 (4.30)	71.78 (4.88)	65.76 (4.36)	68.62 (4.53)
SVM(DI)	62.57 (2.74)	52.33 (3.44)	56.95 (2.85)	72.05 (2.36)	69.68 (2.80)	70.83 (2.39)
LSTM-W(RD+DI)	71.17 (3.41)	64.71 (3.15)	67.75 (2.80)	76.91 (3.13)	73.74 (2.33)	75.27 (2.44)
LSTM-W(RD)	58.39 (3.33)	58.85 (4.72)	58.58 (3.73)	64.95 (3.15)	66.70 (4.07)	65.80 (3.47)
LSTM-W(DI)	72.03 (3.90)	69.61 (3.19)	70.78 (3.35)	78.86 (3.51)	79.43 (3.08)	79.13 (3.08)
LSTM-W+C(RD+DI)	76.75 (2.42)	68.44 (3.26)	72.33 (2.63)	79.64 (2.91)	75.79 (2.90)	77.65 (2.62)
LSTM-W+C(RD)	63.24 (2.81)	61.90 (4.32)	62.52 (3.23)	69.69 (4.22)	70.03 (3.70)	69.81 (3.47)
LSTM-W+C(DI)	**76.31** (3.90)	**74.96** (4.71)	**75.58** (3.86)	**80.85** (2.31)	**81.44** (3.46)	**81.11** (2.41)

Table 5: Results of Precision (prec.), Recall and F-measure (F1) obtained identifying entities in the RDD corpus with a classic machine learning method, SVM (first frame), and with a LSTM neural network architecture, using only word embeddings (-W) (second frame) and using word and case (-W+C) embeddings (last frame). Standard deviation appears under each value. RD stands for rare disease, DI for disabilities and RD+DI for aggregate data. Evaluation 1 checks complete sequences of tags corresponding to each entity. Evaluation 2 checks the tags separately. Best results for detecting disabilities with both types of evaluation appear in boldface.

Table 5 shows the results obtained with both, a classic classifier, SVM, and the LSTM network. For the SVM classifier, inspired by other work [7], we have used

as features the POS tag of the current word and a vector representation of the word itself and the two following adjacent terms. We have used the SVM implementation provided by Weka [42] and default parameters.

6.2. Deep learning model for relationship identification

Once the involved entities have been identified in the sentences, we want to check in which cases there is an actual relation between them. Classic machine learning methods have been frequently applied to the problem of relationship classification. However, as in the case of entity recognition, they require a study of the most suitable set of features to characterize the problem. Because of this, we have resorted again to a DNN. In this case we have tested two different alternative networks. In addition to a network including a LSTM layer, we have used a CNN. CNNs have been successfully applied to different NLP tasks [4], and relationship classification [44, 28] in particular. Inspired by these proposals we have designed a CNN for extracting relations in our corpus. As it is usual in the deep learning approaches, our model does not require defining and extracting complex features from the text. Our model combines lexical information provided by embedding vectors corresponding to the sentence words and sentence level information provided by the position of the entities (disabilities and diseases) withing the sentence. As it has been done in previous works [44, 8] we use position embeddings to provide information about the entities positions. It is a relevant information since the presence of a relation between two target entities is usually determined from words which are close to the target entities. Figure 2 shows a scheme of the model. The word tokens and the positions of the first word in each entity are transformed into embeddings vectors. They are concatenated in a feature vector. Then, we use two convolutional layers, each of them followed by a pooling layer, which reduces the output dimensionality while keeping the most relevant information by performing a sampling operation. Finally, a softmax layer (not appearing in the Figure for simplicity) transforms the output into the probability of having found or not a relationship in the sentence.

Table 6 compares the results of the classic classifier, SVM (first row), and those obtained with the CNN proposed in this section, as well as an alternative network

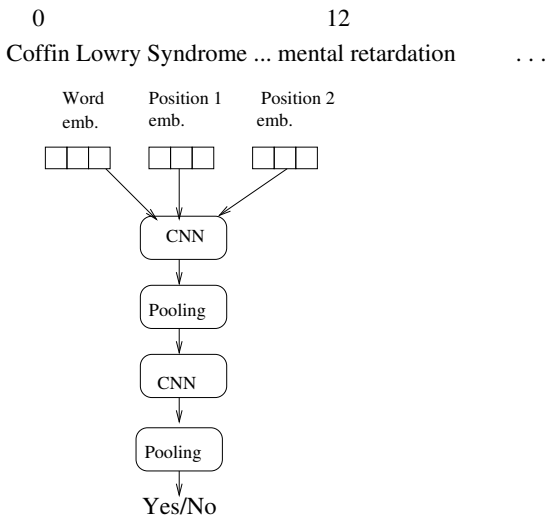

Figure 2: CNN for extracting relations between disabilities and diseases.

		Prec.	Recall	F1
SVM		70.6(3.88)	70.9(8.13)	68.0(3.38)
CNN	Words	56.22(6.25)	56.38(6.14)	55.75(6.44)
	Words+Pos.	75.76(6.27)	75.90(5.14)	75.57(5.73)
LSTM	Words	62.49(9.05)	62.46(13.02)	60.93(11.53)
	Words+Pos.	70.02(5.70)	72.63(5.03)	70.22(5.25)

Table 6: Results of Precision (prec.), Recall and F-measure (F1) obtained classifying relationships in the RDD corpus with a classic machine learning method, SVM, and with the two different neural networks architectures proposed: CNN and LSTM, without and with using the information about the entity positions. Standard deviation appears in brackets.

using an LSTM similar to the one used for the previous problem. We have taken the annotations of disabilities and diseases provided by relationship files in the corpus. In this way, we do not propagate the possible errors of the entity detection phase. For the SVM classifier we have selected a quite refined set of features. These features include the named entities, POS tags of the terms, distance between the named entities, lemmas of the entities and of the words in their contexts, presence of negation and speculation, presence of other entities, entity length, overlapping between entities, and presence of abbreviations. For the SVM classifier we have used the Weka software and default parameters. We evaluated our model using 10-fold cross-validation in all the experiments. Reported results correspond to the average of the 10 runs. For the implementation of the convolutional layers we have used a linear rectifier as activation function. For the LSTM version of the network, we have set the dimension of the bidirectional LSTM to 100. In both networks we have used Adam (Adaptive Moment Estimation) as optimizer. The number of training iterations or epochs on each cross-validation fold has been set to 20 in both neural networks.

7. Discussion

Based on the results for disabilities and diseases recognition shown in Table 5, we observe that the LSTM architecture improves the results of the SVM classifier. Although, the SVM results could be improved by using a more sophisticated set of features, we want to show the advantage of the deep learning proposals when we avoid complex handcrafted feature engineering. Results also show that detecting RDs is a bit harder than detecting disabilities. It may be perhaps because they have very varied names, from a single word to a long expression, sometimes including proper names and sometimes not.

We observe that the use of word embeddings for the task provides high performance for both, precision and recall, reaching an F-measure of 79.13% for disability recognition. It is slightly improved by including case information in the model, reaching then an F-measure of 81.11%. Although we can not compare our results directly to other systems, because we are using a new corpus, we can compare them with other

related problems, such as the entity detection of drugs and adverse effects. Li et al. [23] reach an F-measure 84.6 for the drug and adverse effect identification problem using the ADE corpus [10]. Therefore, our results are similar to those for other related problems.

Concerning relationship extraction, we can observe in Table 6 that including the positions of the entities improves a lot the results. In this case, both kinds of networks beat the results obtained with the classic SVM classifier, although it has been built on a set of complex features that have required to be computed before the classifier's training. We can see that, although LSTM network results are better when using only word information, the CNN results are better than those of the LSTM network when using words and position information. Thus the LSTM network does not perform significantly better than the CNN while it is more computationally expensive to train.

The results obtained are competitive with relation extraction systems applied to other problems in the biomedical domain. Li et al. [23] obtain a F-measure of 71.3 for the relations in the ADE corpus between drugs and adverse effects. Zeng et al. [44] obtained a F-measure of 82.7 for the SemEval-2010 Task 8 dataset, which does not correspond to the biomedical domain. In this work the authors include in their model information from external sources, such as Wordnet. Thus, we can see that our simple model is able to obtain high values of F-measure for the new problem presented in this work.

8. Conclusions and Future Work

We have developed a new annotated corpus designed to support the annotation of disabilities, as well as the extraction of relations between diseases and disabilities. The corpus aims to help to develop and validate automatic systems to perform these tasks. It is particularly important in the area of rare diseases. There are several contributions in this paper. We have presented a detailed procedure to create a gold standard for the annotation of disabilities. We have annotated negation and speculation when they appear affecting a disability. We have also developed a corpus of relationships between rare diseases and disabilities appearing at sentence level.

The annotated RDD corpus has been applied to train and evaluate a deep learning system based on LSTM for disabilities identification, as well as an LSTM network and a convolutional neural network designed for extracting relationships between disabilities and diseases. Despite the simplicity of the proposed models, we have obtained results similar to those obtained for other problems in the biomedical domain.

In the future we plan to develop more sophisticated deep learning models that include additional information such as the POS tags of the words, or stems. We also intend to develop systems for annotating negation and speculation adapted to the case of biomedical concepts, and disabilities in particular.

The corpus has been made publicly accessible (`http://nlp.uned.es/~lurdes/RDDcorpus.zip`) in order to facilitate the research in related areas, such as annotations of disabilities, discovering of relationships between biomedical concepts and the identification of negated or speculated concepts. We have also published the code for the entity detection model (`https://github.com/gildofabregat/RDD-Named-Entity-Recognition-2018/tree/master`) and for the CNN deep learning model for relationship extraction (`https://github.com/gildofabregat/RDD-Relation-Extraction-2018/tree/master`).

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the projects EXTRECM (TIN2013-46616-C2-2-R), PROSA-MED (TIN2016-77820-C3-2-R) and EXTRAE (IMIENS 2017). The authors want to thank Doctors Sonia Albertos Rubio and David Garcia Illescas for their review of the disabilities extracted from the annotated documents.

9. Conflict of interest statement

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

[1] Chapman, W. W., Cohen, K. B., 2009. Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics 42 (5), 757–759.

[2] Chiu, J. P. C., Nichols, E., 2016. Named entity recognition with bidirectional lstm-cnns. TACL 4, 357–370.

[3] Chollet, F., et al., 2015. Keras. `https://github.com/keras-team/keras`.

[4] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., Nov. 2011. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537.

[5] de Chalendar, M., Daniel, M., Olry, A., Rath, A., 2014. Rare diseases and disabilities: improving the information available with three orphanet projects. Orphanet Journal of Rare Diseases 9 (1), O31.

[6] Díaz, N. P. C., Taboada, M., Mitkov, R., 2016. A machine-learning approach to negation and speculation detection for sentiment analysis. JASIST 67, 2118–2136.

[7] do Amaral, D. O. F., Buffet, M., Vieira, R., 2015. Comparative analysis between notations to classify named entities using conditional random fields. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology, STIL 2015, Natal, Brazil, November 4-7, 2015. pp. 27–31.

[8] dos Santos, C. N., Xiang, B., Zhou, B., 2015. Classifying relations by ranking with convolutional neural networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. pp. 626–634.

[9] Goller, C., Kchler, A., 1996. Learning task-dependent distributed representations by backpropagation through structure. In: In Proc. of the IEEE International Conference on Neural Networks. Vol. 1. IEEE, pp. 347–352.

[10] Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L., 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics 45 (5), 885 – 892.

[11] Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., Leser, U., 2017. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33 (14), i37–i48.

[12] Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T., 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. Journal of Biomedical Informatics 46 (5), 914–920.

[13] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9 (8), 1735–1780.

[14] Huang, D., Jiang, Z., Zou, L., Li, L., 2017. Drug-drug interaction extraction from biomedical literature using support vector machine and long short term memory networks. Inf. Sci. 415, 100–109.

[15] Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J., 2003. Genia corpus-a semantically annotated corpus for bio-textmining. Bioinformatics 19 (suppl1), i180.

[16] Konstantinova, N., de Sousa, S. C., Cruz, N. P., Maa, M. J., Taboada, M., Mitkov, R., may 2012. A review corpus annotated for negation, speculation and their scope. In: Chair), N. C. C., Choukri, K., Declerck, T., Doan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, pp. 3190–3195.

[17] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C., 2016. Neural architectures for named entity recognition. In: NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. pp. 260–270.

[18] Le, Q. V., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. pp. 1188–1196.

[19] Lecun, Y., Bengio, Y., Hinton, G., 5 2015. Deep learning. Nature 521 (7553), 436–444.

[20] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., Jackel, L. D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Computation 1 (4), 541–551.

[21] Leonardi, M., Bickenbach, J., Ustun, T. B., Kostanjsek, N., Chatterji, S., 2006. The definition of disability: what is in a name? The Lancet 368 (9543), 1219–1221.

[22] Levy, O., Goldberg, Y., 2014. Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers. pp. 302–308.

[23] Li, F., Zhang, M., Fu, G., Ji, D., 2017. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinformatics 18 (1), 198:1–198:11.

[24] Ma, X., Hovy, E. H., 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. pp. 1064–1074.

[25] Miwa, M., Bansal, M., 2016. End-to-end relation extraction using lstms on sequences and tree structures. In: Proceedings of the 54th Annual Meeting of

the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. pp. 1105–1116.

[26] Morante, R., Liekens, A., Daelemans, W., 2008. Learning the scope of negation in biomedical texts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 715–724.

[27] Morante, R., Sporleder, C., 2012. Modality and negation: An introduction to the special issue. Comput. Linguist. 38 (2), 223–260.

[28] Nguyen, T. H., Grishman, R., 2015. Relation extraction: Perspective from convolutional neural networks. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA. pp. 39–48.

[29] Organization., W. H., 2001. ICF : International classification of functioning, disability and health / World Health Organization. World Health Organization Geneva.

[30] Organization, W. H., 2007. International classification of functioning, disability and health : children & youth version. Tech. rep., World Health Organization.

[31] Organization, W. H., Bank, T. W., 2011. World Report on Disability. World Health Organization, Geneva, Switzerland.

[32] Oronoz, M., Gojenola, K., Pérez, A., de Ilarraza, A. D., Casillas, A., 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. Journal of Biomedical Informatics 56, 318–332.

[33] Percha, B., Altman, R. B., 2015. Learning the structure of biomedical relationships from unstructured text. PLoS Computational Biology 11 (7), e1004216.

[34] Ramshaw, L. A., Marcus, M. P., 1995. Text chunking using transformation-based learning. CoRR cmp-lg/9505040.

[35] Rosario, B., Hearst, M. A., 2004. Classifying semantic relations in bioscience texts. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL '04. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 430–437.

[36] Saurí, R., Verhagen, M., Pustejovsky, J., 2006. Annotating and recognizing event modality in text. In: Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference, Melbourne Beach, Florida, USA, May 11-13, 2006. pp. 333–339.

[37] Smith, L., Tanabe, L. K., Ando, R. J. n., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C. A., Povinelli, R. J., Vlachos, A., Baumgartner, W. A., Hunter, L., Carpenter, B., Tsai, R. T.-H., Dai, H.-J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Maña-López, M., Mata, J., Wilbur, W. J., 2008. Overview of biocreative ii gene mention recognition. Genome Biology 9 (2), S2.

[38] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J., 2012. Brat: A web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. EACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 102–107.

[39] van Mulligen, E. M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J. A., Furlong, L. I., 2012. The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships. Journal of Biomedical Informatics 45 (5), 879 – 884, text Mining and Natural Language Processing in Pharmacogenomics.

[40] Vincze, V., Szarvas, G., Farkas, R., Mra, G., Csirik, J., 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9 (S-11).

[41] Wang, X., Yang, C., Guan, R., Sep 2015. A comparative study for biomedical named entity recognition. International Journal of Machine Learning and Cybernetics.

[42] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[43] Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., Jin, Z., 2015. Classifying relations via long short term memory networks along shortest dependency paths. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015. pp. 1785–1794.

[44] Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., 2014. Relation classification via convolutional deep neural network. In: COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland. pp. 2335–2344.

[45] Zhao, Z., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., Wang, J., Dec 2017. Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC Medical Genomics 10 (5), 73.