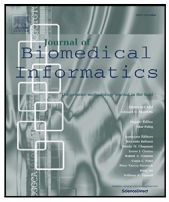
Journal of Biomedical Informatics 138 (2023) 104279

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Negation-based transfer learning for improving biomedical Named Entity Recognition and Relation Extraction

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

[a] Universidad Nacional de Educación a Distancia (UNED). ETS Ingeniería Informática, Juan del Rosal, 16, Madrid, 28040, Spain
[b] Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS), Spain
[c] Avature Machine Learning, Spain

ARTICLE INFO

Keywords:
Transfer learning
Named Entity Recognition
Negation detection
Relation Extraction

ABSTRACT

Background and Objectives: Named Entity Recognition (NER) and Relation Extraction (RE) are two of the most studied tasks in biomedical Natural Language Processing (NLP). The detection of specific terms and entities and the relationships between them are key aspects for the development of more complex automatic systems in the biomedical field. In this work, we explore transfer learning techniques for incorporating information about negation into systems performing NER and RE. The main purpose of this research is to analyse to what extent the successful detection of negated entities in separate tasks helps in the detection of biomedical entities and their relationships.

Methods: Three neural architectures are proposed in this work, all of them mainly based on Bidirectional Long Short-Term Memory (Bi-LSTM) networks and Conditional Random Fields (CRFs). While the first architecture is devoted to detecting triggers and scopes of negated entities in any domain, two specific models are developed for performing isolated NER tasks and joint NER and RE tasks in the biomedical domain. Then, weights related to negation detection learned by the first architecture are incorporated into those last models. Two different languages, Spanish and English, are taken into account in the experiments.

Results: Performance of the biomedical models is analysed both when the weights of the neural networks are randomly initialized, and when weights from the negation detection model are incorporated into them. Improvements of around 3.5% of F-Measure in the English language and more than 7% in the Spanish language are achieved in the NER task, while the NER+RE task increases F-Measure scores by more than 13% for the NER submodel and around 2% for the RE submodel.

Conclusions: The obtained results allow us to conclude that negation-based transfer learning techniques are appropriate for performing biomedical NER and RE tasks. These results highlight the importance of detecting negation for improving the identification of biomedical entities and their relationships. The explored techniques show robustness by maintaining consistent results and improvements across different tasks and languages.

1. Introduction

Named Entity Recognition (NER) and Relation Extraction (RE) are two Natural Language Processing (NLP) tasks especially related to the biomedical domain. However, a somehow fixed collection of types of entities usually studied and considered within these tasks can be found in the literature. For instance, drugs and disorders (and their potential relationships as adverse drug effects), or genes and their influence in causing specific diseases. In this research, we aim to move away from those entities that are normally dealt with in biomedical NLP research, and focus on two specific types of entities and the relationships that can be found among them: disabilities and rare diseases. These entities represent very important aspects of the biomedical research: for instance, it is estimated that 15% of the world's population suffers from some form of disability.[1] On the other hand, an estimated 300–400 million people worldwide are living with a rare disease. Despite these high incidence rates in disabilities and rare diseases, the detection of these kind of entities is not usually considered within classical NER or RE tasks in biomedical research. In order to cover the automatic detection of disabilities in scientific–medical texts and the identification of relationships between disabilities and rare diseases, two different

* Corresponding author at: Universidad Nacional de Educación a Distancia (UNED). ETS Ingeniería Informática, Juan del Rosal, 16, Madrid, 28040, Spain.
E-mail addresses: gildo.fabregat@lsi.uned.es (H. Fabregat), aduque@lsi.uned.es (A. Duque), juaner@lsi.uned.es (J. Martinez-Romo), lurdes@lsi.uned.es (L. Araujo).

[1] World Health Organization report on disability: https://www.who.int/disabilities/world_report/2011/report.pdf - Last visited: March 1, 2022

https://doi.org/10.1016/j.jbi.2022.104279
Received 10 May 2022; Received in revised form 3 December 2022; Accepted 30 December 2022
Available online 4 January 2023

corpora were created in previous works: RDD (Rare Diseases and Disabilities) and DIANN (DIsability ANNotation on documents from the biomedical domain) corpus. Each collection tries to cover specific aspects of information extraction in biomedical documents, and will be used in this research. As it will be detailed in subsequent sections of this work, while the RDD corpus is entirely written in English, the DIANN corpus presents documents in both English and Spanish. This allows us to expand our research for also analysing the differences that can be found between both languages when it comes to NER and RE tasks.

In the context of detecting disabilities and rare diseases, and relationships between them, the main objective of this work is studying the possible effects that negation detection might have in these biomedical tasks. Negation is an aspect of great interest in several domains and NLP tasks due to its implications on the speech. This linguistic element performs important functions of discourse polarization, being this aspect especially relevant in sentiment analysis and relationship extraction [1,2]. In this work, we explore techniques for performing a more targeted negation detection, and the effects this previous step may have in the above-mentioned tasks: NER and RE.

For this purpose, a deep learning architecture is built and trained for automatic detection of negation triggers and scopes, and subsequently, transfer learning techniques are explored for integrating this previously acquired knowledge into a different deep learning architecture devoted to performing NER and RE. Through this pipeline, we are able to extract useful information regarding the influence and impact of incorporating this negation-related knowledge on the recognition of named entities and the extraction of relationships between them. In particular, negation detection is an NLP task usually divided into two subtasks: detection of negation cues or triggers, i.e. words indicating the presence of negation within a sentence, and detection of the scope or part of the text affected by the negation cues. The main intuition behind the use of negation for improving NER and RE tasks is that a successful detection of the scopes involved in the negation detection task can be reflected in how the final system deals with the scopes of the entities. This can lead to the detection of longer and more complex entities, as well as to a better management of longer sentences for finding relationships between entities that are far apart from each other in the text.

The main contributions of this work are the following:

- We present specific neural architectures for addressing NER and RE tasks in the biomedical domain, as well as a separate neural model for detecting negation triggers and scopes.
- Different techniques are explored for transferring the knowledge acquired by training negation models into models performing biomedical NER and RE, and the results of this transfer learning are analysed.
- Experiments are developed for two different languages (English and Spanish).
- Two biomedical corpora devoted to the detection of disabilities and rare diseases and the relationships between those entities are employed for the development of this work.
- We show how successful negation detection leads to significant improvements on important tasks in biomedical NLP such as Named Entity Recognition and Relation Extraction.

The rest of the paper is structured as follows: Section 2 summarizes previous research and general background regarding negation detection, named entity recognition and relation extraction in the biomedical domain. Information regarding the different corpora used in this work, their relationships with the addressed tasks, and the main architectures developed is presented in Section 3. Results are shown in Section 4, separated according to the considered corpora and tasks. Finally, the main conclusions and future lines of work are depicted in Section 5.

2. Related work

Due to their importance within biomedical NLP, many different works can be found in the literature addressing named entity recognition and relation extraction in this specific domain. Both unsupervised and supervised techniques are usually developed for both tasks, depending on the availability of annotated data, which is one of the main challenges for developing successful systems in biomedical NLP. The generation of annotated document collections implies the consumption of large amounts of human and material resources. Considering biomedical NER, the existence of few annotated collections is mainly due to the need of relying on experts to perform this annotation, covering all the considered entities for a task and attending to their high degree of specificity. However, automatic systems can benefit from resources such as SNOMED CT [3] or the UMLS (Unified Medical Language System) metathesaurus [4], and associated tools for the semantic annotation of biomedical texts such as MetaMap [5], cTakes [6] or NCBO [7]. However, most of those resources are developed for the English language. Resources for other languages such as Spanish are much less common, although efforts are being made for their development [8], with tools such as FreelingMed [9]. In addition, quite a large collection of shared tasks and competitions related to biomedical NER in languages other than English are being carried out in the last few years, such as the CLEF eHealth [10] or the IberLEF eHealth-KD [11] initiatives.

Among supervised techniques for performing biomedical NER, deep learning methods such as Long Short-Term Memory (LSTMs) networks [12] and variants of these networks are normally considered to be state-of-the-art in the field, since these algorithms usually improve other techniques due to the sequential nature of the task. Vectorial representations of the input features are normally used in these works [13], and different types of embedding methods and models are analysed and compared [14]. In general, works combining Bi-LSTM [15] and Conditional Random Fields (CRF) layers are normally located among the best participating systems in sequence tagging tasks in general [16], and in biomedical NER competitions in particular [17–19].

On the other hand, relationship extraction is also a challenging task in the biomedical domain, involving both domain-specific knowledge and language-related aspects. As we mentioned for the NER task, resources such as corpora are easily found for the English language. The ADE corpus [20] for extracting relations between drugs and adverse events, or the DDI corpus [21] with annotations on drug-drug interactions are well known examples. In a similar way to NER tasks, deep learning methods achieve the state-of-the-art in relation extraction. Word embeddings on convolutional architectures of different sizes [22] and LSTM-based architectures [23] are normally used for modelling dependencies between entities or terms. Moreover, approaches based on CRF with complex representations and incorporating information about linguistic phenomena such as negation also offer interesting results [24]. This indicates that the use of Bi-LSTM and CRF layers may also lead to achieving competitive results in the RE task, as in the case of NER tasks. Methods of jointly modelling NER and RE have been proposed for exploring both tasks at the same time [23]. In this neural joint model, character embeddings and auxiliary representations for analysing out-of-vocabulary words are explored within a LSTM-based approach.

Combining deep learning stacks with CRF can be also found among state-of-the-art systems when it comes to negation triggers and scopes recognition. Bi-LSTM based systems usually achieve good performance in comparison to other simpler feed-forward neural networks [25,26], especially considering the exact matching, and in different datasets such as the SFU Review corpus [27] or the ConanDoyle-NEG corpus [28]. Moreover, the study is extended [29] to other domains and languages (Chinese), presenting, among others, results for the BioScope corpus [30]. Other types of architectures and deep learning stacks have been also explored. For instance, the use of Convolutional Neural

Networks is studied for detecting negation and speculation within the BioScope corpus, obtaining very competitive results, and highlighting the difficulties presented by long-distance syntactic dependencies [31].

Regarding the biomedical domain, some previous works using transfer learning techniques can be mentioned. In the context of biomedical NER, instance-based transfer learning can be performed by automatically developing big silver standard corpora and training models on those corpora in order to improve results achieved by only training the models on available gold standard corpora [32]. The use of pre-trained language representations, and more particularly contextual language models such as BERT [33] or ELMo [34] is often considered to be a network-based transfer learning technique. Biomedical versions of these language models, like BioBERT [35] or BioELMo [36] can then be used for input representation in many different biomedical NLP tasks [37]. State-of-the-art results are achieved using BERT or some of its variants for the NLP tasks addressed in this work: named entity recognition [38–40], relationship extraction [41] and negation [42]. When it comes to NER, researchers employ different biomedical corpora for these works: Sun and Yang [38] use the PharmaCoNER corpus [43] with substances, compounds and proteins and entities, while Chai et al. [39] make use of datasets oriented to the detection of chemicals, diseases, genes and proteins and species. Finally, Agrawal et al. [40] employ both biomedical datasets with cell-related entities (DNA, RNA, protein, cell-line and cell-type) and a non-biomedical dataset [44]. Works such as [41] address biomedical relation extraction using datasets containing relations between chemicals and diseases (BioCreative V CDR dataset [45]) or between different chemicals (CHR dataset [46]). Nevertheless, we have not found previous works offering results on the particular corpora used in this work, related to more specific and not usually considered entities such as disabilities and rare diseases.

Regarding negation detection, Gubelmann and Handschuh [42] study whether the functioning of the Transformer-based pre-trained language models is driven by simple shallow heuristics or by any real understanding of the languages that they are processing for the case of negation. They construct specific datasets using highly controlled, synthetic, and relatively simple sentences to guide the models, and find that most of the tested models are clearly sensitive to negation for the considered datasets. However, as far as we know, there are no previous works on studying direct transfer learning methods incorporating network information (more particularly, information about negation) from previously trained systems into deep learning models for performing biomedical NER and RE.

3. Materials and methods

The different corpora employed in this research are presented in this section and their main characteristics are described, in order to highlight their relation with the proposed experiments. Neural architectures developed for each of the experiments are also shown and detailed.

3.1. Corpora

As mentioned in Section 1, the research presented in this work requires the use of two different types of corpora: first, corpora that allows us to train the initial neural network devoted to identifying negation triggers and scopes. In particular, the SFU Review SP-NEG corpus [47] will be used for the Spanish language, and the BioScope corpus [30] for the English language. Second, we need specific corpora for performing Named Entity Recognition and Relation Extraction tasks, in order to assess the influence of negation-based transfer learning on them. In this case, the DIANN corpus [48], will be employed for the NER task and the RDD corpus [49] for the RE task. The particular characteristics of each corpus will be depicted in the following subsections.

Table 1
Statistics of the negation corpora used in this research (extracted from [30] and [47]).

	Source	# Documents	# Sentences	# Negations
BioScope	Biomedical abstracts	1273	11872	1597
SFU Review	Product reviews	400	9455	3022

3.1.1. Corpora for detecting negation

The influence of negation detection in NER and RE tasks is studied in this work in two different languages: English and Spanish. For this reason, two different negation detection models are trained, each of them devoted to one of the considered languages. Hence, two different corpora are used for detecting negation: the SFU Review SP-NEG corpus [47] consists of 400 reviews related to 8 different domains (cars, hotels, washing machines, books, cell phones, music, computers and movies), written in Spanish. According to the information provided by the organizers, the corpus was randomly divided into training, development and test; ensuring 33 reviews per domain in training, 7 per domain in development and 10 per domain in test. On the other hand, the corpus employed for the negation detection phase in the English language is the BioScope corpus [30]. It consists of three parts: electronic health records (EHRs) presented in free text format, full biological articles, and biological paper abstracts. The discourse structure under each domain in the BioScope corpus is complex, with EHRs being the most different domain, due to the use of a free writing style. The subset of abstracts, which is the largest subset and contains more negations than the remaining domains (negations being marked all over the corpus, and not only around entities), stands out among the different domains. We use this subset given its careful formatting and its resemblance to the data we need to process.

Both corpora were annotated at token level with labels related to negation triggers and their linguistic scope. In addition, both collections used a similar annotation style and incorporated documents from different scenarios or domains. In short, these similarities allow us to build, from the same architecture, different models for the detection of different negation elements, covering different scenarios and domains, in Spanish and English documents. Table 1 shows some statistics of the two corpora devoted to negation detection employed in this work.

3.1.2. Corpora for performing NER and RE

In order to cover the automatic detection of disabilities and rare diseases in scientific–medical texts and the identification of relationships between these entities, two different collections of documents are used in this research: the DIANN (DIsability ANNotation on documents from the biomedical domain) corpus and the RDD (Rare Diseases and Disabilities) corpus. Each collection covers specific aspects of information extraction in biomedical documents.

The DIANN corpus [48] was developed under the umbrella of the IberEval (Evaluation of Human Language Technologies for Iberian Languages) 2018 conference [50]. The corpus was presented as a common evaluation framework of tools and approaches for the detection of disability mentions in documents written in Spanish and English. It was used as a benchmark for a homonymous task collocated in IberEval conference. The DIANN corpus is exclusively oriented to the study of named disabilities.

The DIANN corpus includes 1000 annotated documents, 500 published in English and 500 in Spanish. Only abstracts presenting at least a mention of a disability in both languages were gathered. Although the Spanish documents share the same contents as the English documents, they do not correspond to literal translations, i.e. the collected documents are abstracts of research articles distributed in both English and Spanish. As a consequence, it is possible that the number of annotations in the Spanish version of an abstract differs from the number of annotations in the English version. The DIANN corpus also presents annotated negations. In particular, a total of 1555 mentions to disabilities (564 unique mentions) were found for the Spanish language

Table 2
Statistics of the corpora used in this research for performing NER and RE tasks (extracted from [51] and [49]).

	Source	# Documents	# Entities	# Negations
DIANN (English)	Biomedical abstracts	500	1656 (disabilities)	63
DIANN (Spanish)	Biomedical abstracts	500	1555 (disabilities)	62
RDD (NER)	Biomedical abstracts	1000	3678 (disabilities)	90
RDD (RE)	Biomedical abstracts	1000	1957 (relationships)	N/A

and 1656 mentions (583 unique mentions) for the English language. Regarding negation, 63 instances were annotated for English and 62 for Spanish.

The RDD corpus [49] was developed with the aim of studying named entity recognition and relation extraction within scientific papers. It gathers a collection of abstracts of scientific articles concerning rare diseases. This corpus was annotated by three different annotators under the supervision of expert medical staff and presents annotations on different disabilities found in the text and hence potentially related to specific rare diseases. Given the importance of different linguistic phenomena in the area of information extraction, it also present annotated negations and speculations that affect one or more disabilities mentioned in each document. In addition, the corpus also contains a file with the relationships between rare diseases and disabilities stated within the documents, following a similar format to the ADE corpus [20] which illustrates relationships between drugs and adverse effects.

The RDD corpus is composed of 1000 abstracts in English, in contrast to the DIANN corpus, which collected documents in both English and Spanish, with an average of approximately 200 words per document and a total of 9657 sentences. The collected documents cover 578 rare diseases, and present 3678 annotations expressing a disability: 2792 are expressed as the impairment of a human function and 886 are stated using some disability term. Hearing, sight and motor skills are the physical functions most often affected by some kind of impairment, while the most frequently mentioned disability is ataxia, related to motor skills, followed by deafness and dementia. The corpus includes 90 negated disabilities and 194 speculation annotations affecting 264 disabilities.

Regarding relationships, a total of 1251 positive and 706 negative relationships between disabilities and rare diseases are identified within the corpus. A total of 362 different rare diseases are covered by the identified relationships. Finally, 186 disability tags were found to be expressed through an acronym.

Table 2 shows some statistics of the two corpora devoted to Named Entity Recognition and Relation Extraction employed in this work.

3.2. Neural models

In this section we present the different neural architectures developed for the different tasks involved in this research: detection of negation scopes and triggers, detection of disabilities and joint detection of named entities and relationships.

3.2.1. Negation detection

The deep learning model developed for the detection of negation scopes and triggers follows the approach proposed in [52] and can be seen in Fig. 1. This model is composed of a Bi-LSTM network which receives an input of different linguistic features in order to sequentially process all terms in a sentence. This first layer is followed by a Conditional Random Field (CRF) which labels each term according to its role as a negation trigger and part of the negation scope.

In order to simplify the study of a transfer learning process that makes use of knowledge obtained from the training of a negation detection system, the same configurations in terms of number of neurons per layer to be transferred have been explored for the different tasks (NER and joint NER and RE) and languages (English and Spanish). As main processing element, we explored the use of a Bi-LSTM with 150 neurons per layer in the negation detection architecture. In addition, each term was processed using an auxiliary character-level representation based on a Bi-LSTM of 50 neurons per layer. Finally, although corpora of different sizes and languages with different numbers of annotations have been studied, a uniform training has been performed using a batch size of 16 in all scenarios.

The input representation is conducted by using the following features: words, characters, Part-of-Speech (PoS) labels and casing information. Two different pre-trained Word Embedding models have been used for representing words, while a separate deep learning subarchitecture [53] is considered for character-level term processing. This way, 50-dimensional character embeddings are previously calculated using different sources of information from the biomedical domain. The use of character embeddings allows us to potentially deal with out-of-vocabulary words and hence improve the whole coverage of the system. This character processing technique has been exhaustively validated in previous works [54] and also successfully employed by some systems participating in the IberEval 2018 DIANN shared task [55] and in further experiments on the negation-based corpora employed in this work [56], as well as on the DIANN corpus [51].

On the other hand, both Part-of-Speech and casing information are represented through one-hot vectors. The use of one-hot vectors for representing casing and PoS information aims to reduce the number of trainable layers. PoS tags are extracted using the Python NLP package "Stanza" [57], which provides models for both the English and Spanish languages. In addition, the main intuition behind the use of the casing one-hot vector is to maintain some useful information that is usually lost when lowercasing all the words within a sentence, which is a common practice during the pre-processing step of these kind of tasks [58]. Through the use of the casing vector, we are able to format and model information that would be otherwise omitted, such as expressions starting with a capital letter or containing letters and numbers, which may be indicators of entities.

As previously mentioned, in this study we worked on the SFU Review SP-NEG corpus (Spanish) and the BioScope corpus (English). Then, two different models of Word Embeddings based on Word2Vec [59] were employed: for the Spanish language we used the Spanish Billion Word Corpus and Embeddings [60], built with 300-dimensional embeddings, while words for the English model were represented by biomedical 200-dimensional vectors [61].

We transformed the different datasets into the BILOU labelling scheme [62] (I: In - For tokens within the annotation; O: Out - For tokens outside the annotation; B: Begin - For the first token of each annotation; L: Last - For the last token of each annotation; U: Unique - Annotations composed of a single token). This annotation scheme allows the representation of partial overlapping and nested entities. We used this labelling scheme to represent both the scope and the negation triggers separately. Then, these two codifications were combined into a single one by means of concatenating the labels. For instance, if a token represents the beginning of a negation scope and the beginning of a negation trigger, its label in this scheme will be "BB". A total of 17 different labels were generated, given that there exist combinations of labels that cannot occur since a negated expression must always have an associated scope. As an example, Table 3 shows the instance "*No tendré jamás que aceptar un trabajo que no me gusta por el dinero.*", annotated following the BILOU format, where the first column contains the word and the second column contains the label after joining the scope label and the trigger associated label.

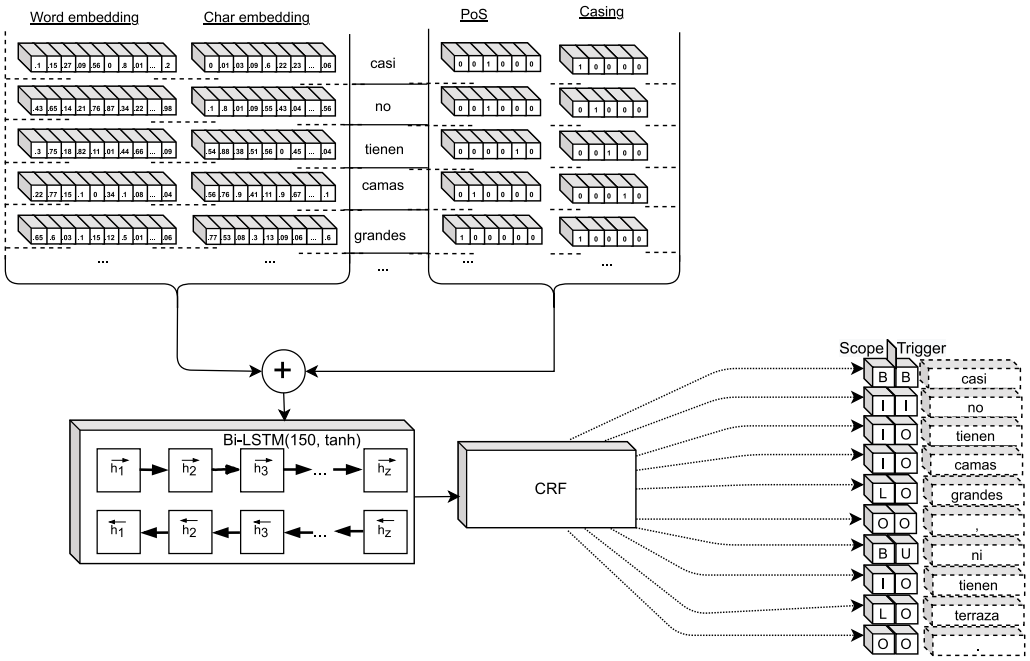


Fig. 1. Deep learning model for negation trigger and scope detection. Casing information and PoS-tagging are encoded using One-hot vectors. Bi-LSTM inputs are the concatenated features of each word. The model output covers the labelling of triggers and scopes simultaneously.

Table 3
Example of an instance (sequence of tokens) and its associated labels. Column "Word" contains each token and column "Label" presents the associated combination of labels for scope (first letter) and trigger (second letter) detection following the BILOU format.

Word	Label	Word	Label
no	BU	no	BU
tendré	IO	me	IO
jamás	IU	gusta	LO
que	IO	por	IO
aceptar	IO	el	IO
un	IO	dinero	LO
trabajo	IO	.	OO
que	IO		

This example shows the annotation of two negations: the first one spans from the first term "no" up to the term "dinero", while the second one is nested and spans from the second term "no" up to the term "gusta".

3.2.2. Named entity recognition

The first main task we want to assess the effect of negation detection on is the named entity recognition task that can be performed on the DIANN corpus. As it was stated before, both the Spanish and English languages are taken into account within this corpus. Fig. 2 shows the neural architecture proposed for this task.

In line with some works presented to the DIANN [48] competition [55,63], a model mainly based on Bi-LSTM and CRF is developed. As it can be seen in the figure, the proposed model presents a Bi-LSTM layer followed by a CRF performing the final classification. The input representation is similar to the one explained in the previous section, this is, it contains word and character embeddings, and one-hot vectors for representing Part-of-Speech and casing information. The word and character embedding models remain the same as for the negation detection task. As previously mentioned, the final size of the Bi-LSTM, configured jointly with the negation models developed for this task, has been 150 neurons for both the Spanish and English versions of the DIANN corpus.

Two different experiments are conducted for assessing the influence of negation detection on the Named Entity Recognition task: in the first one, all weights in the deep learning architecture are randomly initialized, while in the second experiment, the initial weights of the Bi-LSTM layer and the character processing sub-model are extracted from the corresponding negation detection model, trained with the SFU Review SP-NEG corpus for Spanish and with the BioScope corpus for English.

Considering training, and given the small size of the DIANN corpus, in both cases (with and without previous negation detection) the model is trained during 50 epochs using small batches of size 16. Regarding the learning rate, we explored an initial learning rate of 0.01 and a reduction of 10% of this rate after each epoch. An early stop criterion is applied which implies stopping the training if the F1 score of the validation set does not improve after three consecutive epochs. Finally, in order to avoid over-fitting, dropout layers of 0.25 were introduced between processing layers. The loss function employed for training the NER system is a Sparse Categorical Cross-Entropy function, considering that the system has to deal with multiple labels.

A final rule-based post-processing step is performed for improving the global coverage of the system. This step is devoted to the detection of abbreviations within the text, and is inspired by previous works that illustrate the usefulness of combining machine learning and deep learning systems with other modules more based on manual rules [63]. More specifically, the post-processing rule focuses on the detection of abbreviations in brackets, located within three words of a disability label in the text. All abbreviations detected following this rule are subsequently located in the text, and labelled as disabilities.

Regarding two different transfer learning classifications that can be found in the literature [64,65], the proposed transfer learning technique would be inductive transfer learning, since two domain-related although different tasks are involved (negation detection and NER), with available data for both. Moreover, this setting can also be considered as network-based deep transfer learning, since it is based on sharing complete parts of a previously trained neural network (in this case, the weights from the Bi-LSTM in the negation detection model).

3.2.3. Joint named entity recognition and relation extraction

A second task is proposed in this research for studying the impact of negation detection transfer learning on different biomedical tasks:

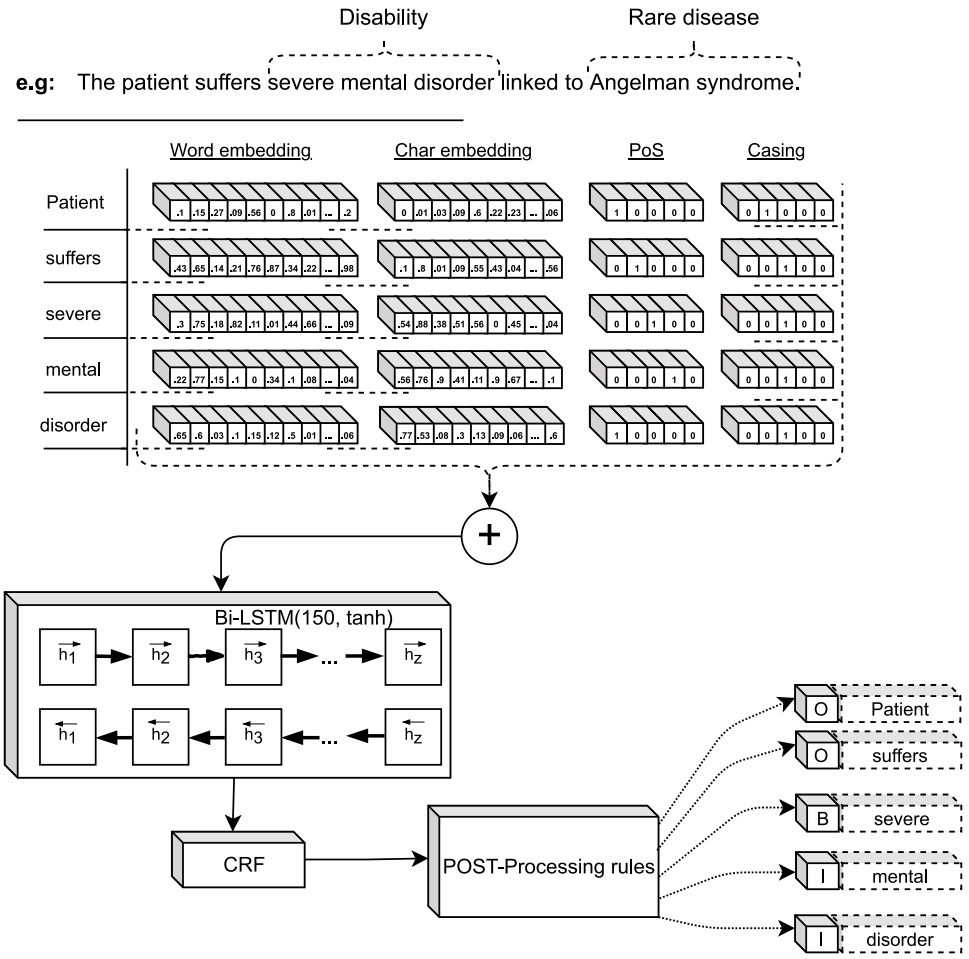


Fig. 2. DIANN Corpus: Deep learning architecture including its inputs and layers. Casing information and PoS-tagging are encoded using One-hot vectors. Bi-LSTM inputs are the concatenated features of each word. The post-processing rules are applied to the output of the deep learning model.

for this purpose, the RDD corpus allows us to explore a similar Named Entity Recognition task as the one explained in the previous section, combined with a Relation Extraction task. In particular, once the disabilities present in the documents of the RDD corpus have been detected, the extraction of relationships can be modelled as a binary classification task, by discriminating whether the tuple (Disability, Rare Disease) within a sentence is related or not, regardless of the relation labels.

Considering that we already studied a named entity recognition scenario in the previous section, and also considering the setting of the RDD corpus, which offers the possibility of studying NER and RE jointly, this particular task of named entity recognition and relationship extraction has been tackled from a multi-task approach. With this approach, we intend to explore the interconnection of both tasks using a weight-sharing technique between different models [66]. In this context, the concatenation of intermediate representation spaces obtained by a NER system is studied, which represent an enrichment of the features used to perform relationship detection. Fig. 3 illustrates this behaviour. The union of two explored subsystems is presented in the figure.

First, for the NER model (left) the architecture based on Bi-LSTM and CRF already explained in the previous section is employed. On the other hand, the architecture for the RE subtask (right) consists of a Bi-LSTM layer followed by a convolutional neural network (CNN), and finally a dense layer followed by a last dense network for performing the final binary classification. The input of the convolutional neural network is the concatenation of weights from both Bi-LSTMs, in the NER submodel and in the RE submodel. This way, we intend to share the latent space generated by the Bi-LSTM network of the NER model with the RE model.

Regarding the considered input features, the NER submodel presents the same collection of features already described in the previous section: word and character embeddings and one-hot vectors for representing PoS and casing. For the RE submodel, the input features are represented by the sentence level information provided by the position of the entities (disabilities and diseases) within the sentence. For each entity involved in the relationship to be classified, a feature representing the absolute distance of each term to the corresponding entity is generated. Fig. 4 represents these features. Then, this information is encoded using embeddings calculated during training.

Finally, an additional vector is added as a feature to the RE submodel, representing the average value of the word embeddings that can be found between the two studied entities. This vector is denoted "Context embedding". Regarding other types of lexical features such as those represented by word or character embeddings, PoS tagging or casing information, this input is incorporated to the joint model through the use of Bi-LSTM weights from the NER submodel. The concatenation of both types of features is processed using the aforementioned convolutional architecture, while the context embedding is added into the final dense network. A size of 50 neurons is used for both Bi-LSTMs, while the dense network of the NER submodel contains 25 neurons. On the other hand, the convolutional layer of the RE submodel contains a total of 16 filters with a kernel size of 3, and the final dense network presents 50 neurons. The learning rate of this model is set to

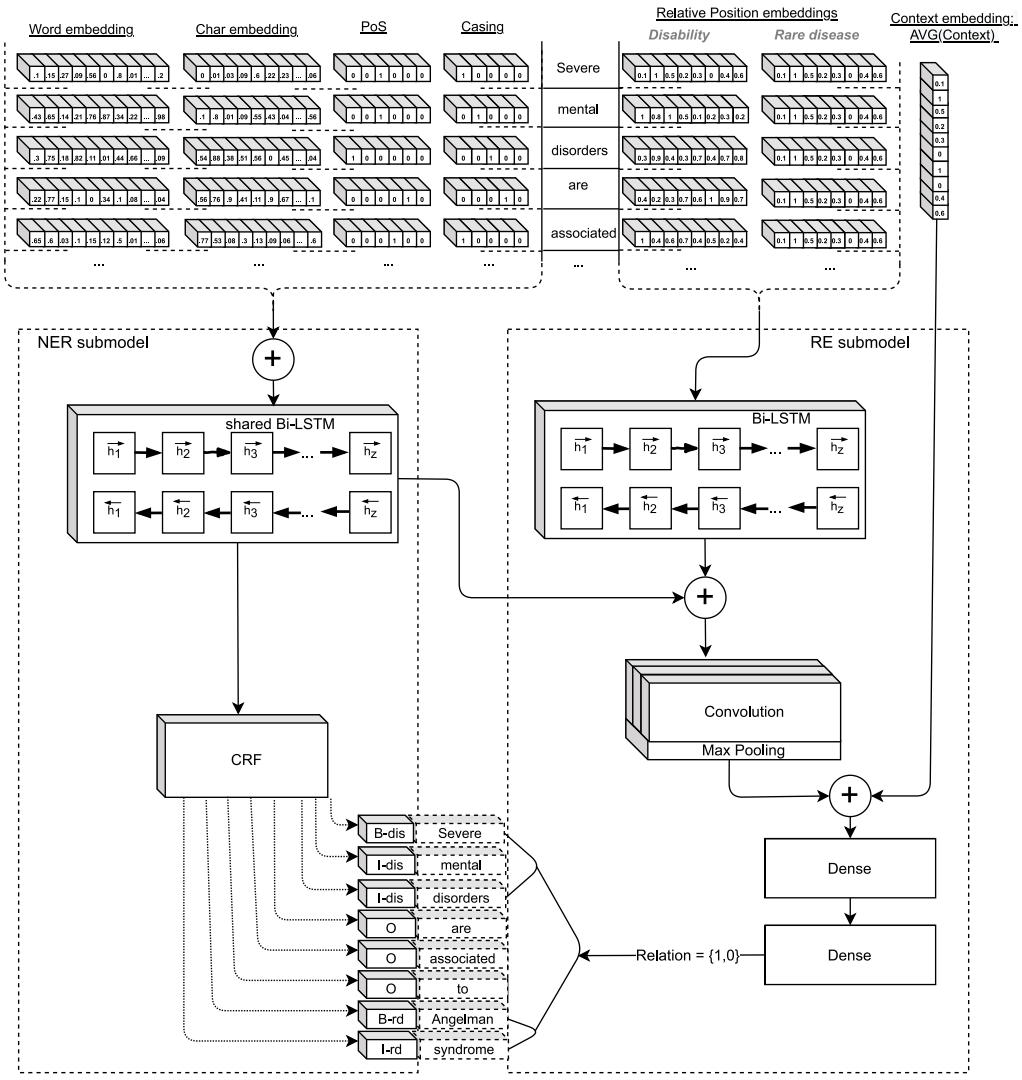


Fig. 3. RDD Corpus: Deep learning joint model for named entity recognition (NER submodel) and relationship extraction (RE submodel). NER submodel: a Bi-LSTM+CRF based stack to process the concatenation of the inputs (words and characters embeddings, and one-hot vectors representing PoS and casing information). RE submodel: (1) two Bi-LSTMs (one of them shared with NER submodel) and a convolutional network to process the inputs of more than two dimensions. (2) two densely connected networks (50 and 1 neuron/s) to process the concatenation of the pooling of (1) with the two-dimensional inputs (summarized context embeddings).

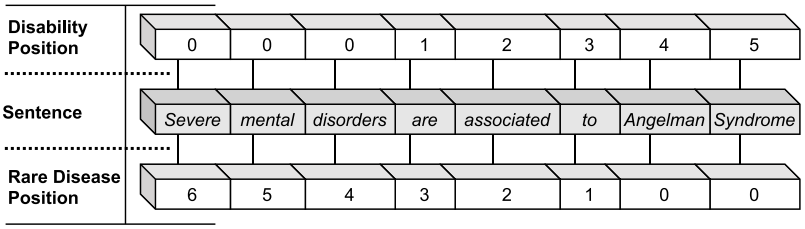


Fig. 4. Generation of vectors capturing the information related to the position of each entity being part of a relationship. In the example, Disability: *"Severe mental disorders"* and Rare Disease: *"Angelman Syndrome"*.

0.01, with a batch size of 16 and 50 training epochs. The early stopping criteria is the same used in the previous experiment: the training stops if there exists no improvement on the F1 score of the validation set during at least 3 consecutive epochs.

For this joint model, we propose a two-stage training pipeline: in a first phase, only the NER submodel is trained, in a similar way to that described in the previous section. In a second step, the whole model is trained jointly, this is, for every sentence to be processed, its lexical features are processed by the NER model for their subsequent use by the RE model through weight sharing, and the sentence level information is processed by the RE model. The initial "only NER" training allows us to study the differences introduced in the system that can be attributed to the use of the joint model, when compared to the results obtained by the NER model on its own. The loss function employed in the NER subsystem is also a Sparse Categorical Cross-Entropy, while the RE subsystem employs a Binary Cross-Entropy loss function, given that it is designed as a binary classification task. The update of the loss function is straightforward when it comes to the first training stage (only NER

Table 4
Results for the two subtasks of the negation detection system: scope recognition and trigger detection for both the English language (BioScope corpus) and the Spanish language (SFU corpus).

	Scope recognition			Trigger detection		
	PCS	F1s	F1t	Precision	Recall	F-measure
English (BioScope)	0.8852	0.8854	0.8005	0.9740	0.9428	0.9575
Spanish (SFU)	0.7429	0.8525	0.7200	0.9969	0.9185	0.9560

training). Regarding weight updates along the network in the second step (joint training), both losses are combined in an unweighted manner for performing this stage.

Finally, in a similar way to the experiment described in Section 3.2.2, the influence of previously acquired negation-based knowledge on the joint NER and RE task is studied by introducing this knowledge in the model. Hence, we compare the performance of the joint NER and RE model in two different scenarios: First, all the weights in the deep learning stack are randomly initialized. In the second scenario, the initial weights of the Bi-LSTM layer of the NER submodel are initialized using the trained weights from the negation detection model, in this case trained only with the BioScope corpus, since this task is only conducted in the English language. As previously mentioned, this is a type of inductive, network-based deep transfer learning.

4. Results

In this section we gather the main results that correspond to the experiments described in previous sections, particularly those tasks for which the influence of negation-based transfer learning is being analysed. Results regarding the negation detection system described in Section 3.2.1 are also shown in this section as a reference for better understanding, however, the main purpose of this research is to study their impact on the two proposed downstream tasks: Named Entity Recognition and Relation Extraction. Hence, the detailed behaviour of the systems designed for performing named entity recognition, on the one hand, and joint named entity recognition and relation extraction, on the other hand, will be shown thereafter. For all the performed experiments in the downstream tasks, we consider the usual metrics in the field: precision, recall and F-Measure as the harmonic mean of precision and recall. The aspects related to the evaluation of each task will be described in their specific subsections.

4.1. Negation detection

Table 4 shows results obtained by the architecture proposed in Section 3.2.1 for the two subtasks considered in the negation detection problem: scope recognition and negation trigger identification. Evaluation of trigger detection is conducted using standard precision, recall and F1 metrics. On the other hand, and due to the nature of the subtask itself, scope recognition is evaluated through the Percentage of Correctly identified Scopes (PCS), as well as the F1 measure at scope level (F1s) and the F1 measure at token level (F1t). A 10-fold cross-validation strategy is followed for model evaluation.

As previously mentioned, the analysis of the results achieved by the negation detection model is out of the scope of this research. However, we can observe how the proposed model is able to obtain satisfactory results, especially in terms of trigger detection. Regarding the different languages considered, the system performs better on the BioScope corpus (English language) when it comes to scope recognition, while results on the trigger detection subtask are similar for both languages. In general, these results lead us to hypothesize that applying transfer learning techniques between this negation detection model and the considered downstream tasks could be beneficial for the latter.

4.2. Named entity recognition: DIANN corpus

The first named entity recognition task is performed on the DIANN corpus, using the system described in Section 3.2.2. As we already mentioned, the influence of the negation detection system, trained for both the English and Spanish language, is analysed by initializing the weights in the Bi-LSTM layer with values obtained from this training of the negation detection system. Two different evaluation schemes are considered for this task: exact and partial matches. The exact matching criterion checks if every proposed annotation matches exactly with the ground truth. On the other hand, due to the freedom with which an entity (in this case, a disability) can be expressed, a second evaluation criterion called partial matched is also employed. This criterion is based on the concept of core-term match [67], and considers a label to be correct if at least the minimum unit or core contained in the ground truth is identified. This collection of core annotations is also provided within the DIANN corpus.

Tables 5 and 6 summarize the main results obtained with the proposed system for the NER task performed on the DIANN corpus. The different languages, evaluation criteria and models considered are shown in the tables. In order to determine whether the differences between the various configurations of our system are statistically significant, each experiment is repeated 10 times using different seeds for the initialization values of the employed networks. Tables show both the average and standard deviation for each proposed metric, in each experiment. Detailed results for each of the 10 repetitions of each experiment are provided in A.

Results clearly show how, both in English and Spanish, introducing information about negation using the described transfer learning technique leads to an overall improvement of global results, considering the achieved F-measures. When it comes to English, the main improvement can be seen in recall, while precision is the metric that presents the highest enhancement in the Spanish language. This improvements result, in both cases, in a higher F-measure when considering the influence of negation, over the base model. Regarding exact and partial matching criteria, logically the best results are always achieved by the less restrictive partial evaluation, however, the addition of information on negation is always beneficial. An interesting result is that the influence of the previous negation detection system is much more noticeable in the Spanish language: English results are quite higher than Spanish results using the base model (with rules), presenting around 4 percentage points more in the exact metric and around 8 points in the partial metric. However, after the boosting produced by negation information, and always considering the use of post-processing rules, the differences almost disappear: the English model is still better by around 1.5 points in the partial metric, but the Spanish model is able to present similar results to the English model when it comes to the exact metric. This is, negation-based transfer learning improves around 3.5 points the exact metric and barely 0.84 points the partial metric for the English language, but the improvement rises up to more than 7 points for the exact metric and around 7 points for the partial metric in the Spanish language. This could be due to the nature of the corpus employed for training the negation detection model in Spanish, which might present more similarities with the Spanish version of the DIANN corpus, and hence provide better assistance for refining entity detection. The density of negations within the different corpora used for performing the negation detection step could also influence these final results: the SFU corpus presents around 1 negation per 3.1 sentences, while the ratio in the BioScope corpus is 1 negation per 7.4 sentences. Moreover, the room for improvement is higher in Spanish case, since the performance of the base model is quite lower than that of the English base model. Finally, it becomes clear that the use of post-processing rules related to the detection of abbreviations is always beneficial for improving the F-measure of the systems, normally due to a significant improvement on recall scores. The table also shows the results of the Wilcoxon signed rank test for statistical significance

Table 5
DIANN Corpus: Results obtained by the proposed model for named entity recognition. "Base Model" shows results for the model with randomly initialized weights and "Negation" shows results when the Bi-LSTM initial weights come from the negation detection model. English experiments are shown, with negation model trained with the BioScope corpus. Rows "No rules" indicate results when post-processing rules are not taking into account, while rows "Rules" take those rules into consideration. Exact and partial results are also shown. Bold indicates the best results for each metric among the four studied configurations (base/negation model, rules/no rules), for each language and type of metric (exact or partial). Star (*) indicates statistically significant differences with respect to the base model, for the same experiment, metric and language (Wilcoxon Signed Rank Test).

			English (BioScope)		
			Precision	Recall	F-measure
Base model	No rules	Exact	0.7366 ±0.0207	0.6353 ±0.0193	0.6820 ±0.0154
		Partial	0.8345 ±0.0216	0.7197 ±0.0210	0.7726 ±0.0158
	Rules	Exact	0.7280 ±0.0176	0.7450 ±0.0142	0.7362 ±0.0108
		Partial	0.8434 ±0.0200	0.8630 ±0.0146	0.8529 ±0.0111
Negation	No rules	Exact	**0.7989 ±0.0409**(*)	0.6887 ±0.0198(*)	0.7385 ±0.0093(*)
		Partial	**0.8674 ±0.0453**(*)	0.7479 ±0.0234	0.8020 ±0.0074(*)
	Rules	Exact	0.7676 ±0.0440(*)	**0.7723 ±0.0163**(*)	**0.7689 ±0.0154**(*)
		Partial	0.8597 ±0.0469	**0.8651 ±0.0202**	**0.8613 ±0.0155**

Table 6
DIANN Corpus: Results obtained by the proposed model for named entity recognition. "Base Model" shows results for the model with randomly initialized weights and "Negation" shows results when the Bi-LSTM initial weights come from the negation detection model. Spanish experiments are shown, with negation models trained with the SFU Review SP-NEG corpus. Rows "No rules" indicate results when post-processing rules are not taking into account, while rows "Rules" take those rules into consideration. Exact and partial results are also shown. Bold indicates the best results for each metric among the four studied configurations (base/negation model, rules/no rules), for each language and type of metric (exact or partial). Star (*) indicates statistically significant differences with respect to the base model, for the same experiment, metric and language (Wilcoxon Signed Rank Test).

			Spanish (SFU)		
			Precision	Recall	F-measure
Base model	No rules	Exact	0.7640 ±0.0218	0.5085 ±0.0354	0.6100 ±0.0285
		Partial	0.8382 ±0.0196	0.5576 ±0.0335	0.6691 ±0.0245
	Rules	Exact	0.7465 ±0.0174	0.6455 ±0.0305	0.6919 ±0.0191
		Partial	0.8382 ±0.0184	0.7246 ±0.0267	0.7768 ±0.0125
Negation	No rules	Exact	**0.8079 ±0.0267**(*)	0.6839 ±0.0134(*)	0.7404 ±0.0064(*)
		Partial	**0.8710 ±0.0256**(*)	0.7375 ±0.0193(*)	0.7983 ±0.0094(*)
	Rules	Exact	0.7721 ±0.0242(*)	**0.7589 ±0.0233**(*)	**0.7649 ±0.0101**(*)
		Partial	0.8537 ±0.0249	**0.8393 ±0.0282**(*)	**0.8458 ±0.0124**(*)

Table 7
Ablation test on the negation detection model. Results obtained by the proposed model for named entity recognition (DIANN corpus), training the negation detection model only for trigger detection (upper rows) or for scope detection (lower rows). Rows "No rules" indicate results when post-processing rules are not taking into account, while rows "Rules" take those rules into consideration. Exact and partial results are also shown.

			English (BioScope)			Spanish (SFU)		
			Precision	Recall	F-measure	Precision	Recall	F-measure
Only triggers	No rules	Exact	0.8163	0.6723	0.7373	0.7990	0.6920	0.7416
		Partial	0.8980	0.7395	0.8111	0.8763	0.7589	0.8134
	Rules	Exact	0.7991	0.7689	0.7837	0.7566	0.7634	0.7600
		Partial	0.9039	0.8697	0.8865	0.8496	0.8571	0.8533
Only scopes	No rules	Exact	0.8254	0.6555	0.7307	0.8162	0.6741	0.7383
		Partial	0.8942	0.7101	0.7916	0.8703	0.7188	0.7873
	Rules	Exact	0.8000	0.7563	0.7775	0.7650	0.7411	0.7528
		Partial	0.8933	0.8445	0.8683	0.8387	0.8125	0.8254

between the base and negation models for each of the considered languages and metrics. As it can be observed, the negation-based transfer learning model achieves better results in terms of statistical significance with respect to the base model for most of the considered experiments.

With the aim of providing further information about the behaviour of the developed system, additional experiments have been carried out in order to run an ablation test on the negation detection model. In this test we separate the two features detected by the model (negation triggers and scopes), and analyse the results obtained on the DIANN corpus when the negation model is trained only for negation trigger detection, and when it is trained only for negation scope detection. Table 7 shows the main results of this ablation test.

Results from the ablation test indicate that in most cases, training the negation model only for trigger detection incorporates enough information to the model to obtain similar results to those achieved by the system when the negation model has been fully trained. On the other hand, results on scope detection are slightly worse. However, differences between the different training regimes (only trigger detection, only scope detection, or trigger and scope detection) are not statistically significant. This might indicate that trigger and scope detection are quite similar subtasks, and no significant complementary information is extracted from any of the subtasks with respect to the other.

For better understanding the performance of the proposed system, Tables 8 and 9 show results achieved on the DIANN corpus by systems participating in the IberEval 2018 DIANN shared task, for the Spanish and English languages, respectively. Results obtained by the best configuration of the proposed system (negation-based transfer learning with post-processing rules) are also included at the bottom of both

Table 8
DIANN corpus: Comparative between systems participating in the IberEval 2018 DIANN shared task, Spanish language. Last row contains results obtained by the best configuration of the system presented in this work. Bold indicates the best result and the overall ranking is indicated in brackets, for each metric (Precision, Recall and F-Measure) and for each evaluation criteria (exact and partial matching).

	Spanish language					
	Exact matching			Partial matching		
	Precision	Recall	F-measure	Precision	Recall	F-measure
IxaMed	0.757 (5)	**0.817 (1)**	**0.786 (1)**	0.822 (6)	**0.886 (1)**	**0.853 (1)**
UC3M	**0.818 (1)**	0.646 (3)	0.722 (3)	0.882 (3)	0.716 (4)	0.79 (3)
UPC	0.807 (3)	0.603 (5)	0.69 (4)	0.889 (2)	0.664 (5)	0.76 (4)
IXA	0.65 (6)	0.642 (4)	0.646 (5)	0.712 (7)	0.734 (3)	0.723 (5)
SINAI	0.459 (7)	0.345 (6)	0.394 (6)	0.512 (8)	0.384 (7)	0.439 (6)
LSI_UNED	0.41 (8)	0.249 (7)	0.31 (7)	0.847 (5)	0.533 (6)	0.654 (7)
GPLSIUA	0.813 (2)	0.17 (8)	0.282 (8)	**0.959 (1)**	0.205 (8)	0.338 (8)
Ours	**0.772 (4)**	0.759 (2)	0.765 (2)	0.854 (4)	0.839 (2)	**0.846 (2)**

Table 9
DIANN corpus: Comparative between systems participating in the IberEval 2018 DIANN shared task, English language. Last row contains results obtained by the best configuration of the system presented in this work. Bold indicates the best result and the overall ranking is indicated in brackets, for each metric (Precision, Recall and F-Measure) and for each evaluation criteria (exact and partial matching).

	English language					
	Exact matching			Partial matching		
	Precision	Recall	F-measure	Precision	Recall	F-measure
IxaMed	0.786 (3)	**0.86 (1)**	**0.821 (1)**	0.842 (4)	**0.922 (1)**	**0.88 (1)**
UC3M	0.778 (4)	0.72 (3)	0.748 (3)	0.822 (5)	0.761 (3)	0.791 (3)
UPC	0.799 (2)	0.605 (4)	0.689 (4)	0.875 (2)	0.663 (5)	0.754 (5)
IXA	0.701 (6)	0.531 (6)	0.604 (6)	0.761 (7)	0.576 (6)	0.656 (6)
SINAI	0.625 (8)	0.37 (7)	0.465 (7)	0.688 (8)	0.407 (7)	0.512 (7)
LSI_UNED	0.671 (7)	0.597 (5)	0.632 (5)	0.815 (6)	0.761 (3)	0.787 (4)
GPLSIUA	**0.884 (1)**	0.251 (8)	0.391 (8)	**0.94 (1)**	0.259 (8)	0.406 (8)
Ours	0.768 (5)	0.772 (2)	0.769 (2)	0.860 (3)	0.865 (2)	0.861 (2)

tables for contextualizing them in the scope of the shared task. These results achieved by our system are extracted from Tables 5 and 6, in particular from the last two rows of the tables ("Negation" model with "Rules", "Exact" and "Partial" metrics), for both languages.

For the Spanish language, our system achieves the second best place in the overall ranking only overcome by the IxaMed system [63]. This team proposed a hybrid system based on deep learning architectures similar to those employed in this work, together with a couple of rule-based modules, one of them devoted to detecting disability-associated triggers, and a second module for identifying abbreviations. This team also presents slightly better results with respect to ours when it comes to the English language. Although the ranking of our system worsens in this language for some of the metrics such as precision of the exact matching criteria, we are able to obtain the second best results considering F-Measure for both exact and partial evaluation criteria. However, and on top of these interesting results, it is important to remark that the main purpose of the research presented in this work is not a thorough comparison with other state-of-the-art systems, but to test and validate the appropriateness of transfer learning techniques, and in particular negation-based transfer learning, in biomedical tasks such as those considered. Moreover, most of the hyperparameters involved in the developed neural architectures have not been exhaustively tuned, which indicates that the proposed system may still present room for improvement.

Regarding the approaches submitted by the other participating teams, different techniques for addressing the task were proposed. Both the UC3M team [55] and the UPC team [68] present architectures similar to ours, also based on LSTMs or Bi-LSTMs and CRF layers, although considering different input features. The IXA team makes use of an entity recognition system denoted "ixa-pipe-ner" [69], based on morphological and typographic features of the text [70]. The SINAI team presents two different approaches depending on the language: a technique based on the Metamap system [5] for the English language, and their own UMLS-based [71] entity recognition system for the Spanish language [72]. The system presented by the LSI_UNED team is an unsupervised approach that generates variants from an initial list of disabilities and body functions for detecting those variants within the text [73]. Finally, the GPLSIUA team presents their own general purpose automatic learning system, which generates candidate expressions and selects those that can be labelled as disabilities, through a Random Forest-based technique that employs syntactic and distributional features [74].

4.3. Joint named entity recognition and relation extraction: RDD corpus

As previously mentioned, the RDD corpus is employed for performing a more complex task: joint named entity recognition and relation extraction in the English Language. The two different steps of the proposed pipeline for this tasks are described in Section 3.2.3: in the first step, only the NER submodel is trained, while the second step involves the joint training of the NER submodel and the RE submodel. In this step, weights from the Bi-LSTM of the NER submodel are shared within the RE submodel. Table 10 contains the main results for this Named Entity Recognition and Relation Extraction Tasks. The employed metrics are Precision, Recall and F-Measure, and results are shown both after the first training step involving the NER submodel, and after the second step (NER and RE training). As in the previous experiment, two different configurations are compared: randomly initialized weights (base model), and negation-based transfer learning, in which the initial weights of the Bi-LSTM layer in the NER submodel are transferred from the negation detection task performed on the BioScope corpus for the English language.

The impact derived from transferring negation-based information into the proposed model is also quite clear in this case. Regarding the first step, involving the training of the NER subsystem, all the considered metrics improve, leading to an overall increase in F-Measure of around 13%. This is in line with the observations made in the previous experiment. Moreover, considering that we are working with the English language, the influence of negation knowledge appears to be more powerful in this experiment, which could indicate that the

Table 10
RDD Corpus: Results obtained by the proposed joint model for relationship extraction (RE) and entity recognition (NER). We report the results obtained using the base model with randomly initialized weights (row "Base Model") and the model using negation-based weights in the NER submodel (row "Negation"). The table shows the results obtained in both training phases: NER sub-model training (Step 1) and full system training (Step 2). Bold indicates the best results for each metric among the two studied configurations at the different steps of the joint model. The (+) symbol represents statistically significant differences with respect to the base model, while the (=) symbol represents lack of statistically significant differences (Wilcoxon Signed Rank Test).

		Step 1: NER training			Step 2: Joint training		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Base model	NER	0.6529	0.6048	0.6279	0.741	0.7587	0.7498
	RE				0.752	0.7282	0.7399
Negation	NER	**0.7815**(+)	**0.7365**(+)	**0.7583**(+)	**0.7596**(=)	**0.7661**(=)	**0.7628**(=)
	RE				**0.7637**(=)	**0.746**(=)	**0.7547**(=)

RDD corpus is more affected by negation than the English part of the DIANN corpus. When it comes to the second step, the influence of negation is quite less remarkable in the second training of the NER subsystem. However, both the NER and the RE model benefit from negation-based transfer learning in Step 2, with an improvement of around 1.5% in F-Measure for both models. This could mean that the previous improvement provided by negation information in the initial NER training (Step 1) is propagated along the joint system, eventually affecting the RE submodel. Finally, if we consider the first and second training steps involving the NER subsystem, results clearly show how F-Measure is always improved: around 7% in the system not using negation information, and around 0.4% in the system involving negation-based transfer learning. However, the main objective of this experiment is not evaluating the performance of the joint model, but the influence of negation-based transfer learning, which proves to be useful in the two studied tasks.

Nevertheless, there is no statistically significant differences between using the Base Model or the Negation Model when it comes to the joint training depicted in Step 2, while the differences in Step 1 are indeed statistically significant, as well as the differences between the NER subsystem in Step 1 and Step 2 of the Base Model. This might indicate that the boosting observed after the inclusion of negation-based transfer learning in Step 1 is already achieved in the NER subtask by considering the joint learning scenario, even when negation has not been considered (Base Model). Hence, negation-based transfer learning and joint learning appear to be two equivalent techniques in this case, although their combination does not seem to offer any additional improvement.

4.4. Case analysis

An analysis of examples in which entity and relationship detection is improved when negation-based transfer learning is applied has been carried out, in order to extract some clues on how this pre-training is helping to improve the final results. We have performed this analysis on instances from the RDD corpus, since this corpus presents the two tasks of interest in this research: Named Entity Recognition and Relation Extraction.

Below we show two examples for which the NER task has been improved through the use of negation-based transfer learning[2]:

- "We describe a girl with **motor and mental retardation**, macrocephaly, a 'coarse' face, choanal atresia, postnatal feeding difficulty, redundant skin with deep palmar and plantar crease, and histopathological evidence of altered elastic fibre, who died at the age of 11 months".
- "We report a patient with non-Down syndrome AML, also known as AMKL, with monosomy 7, who was also obese and had a **hearing impairment** and **mental retardation**".

In both examples, we indicate the entities of interest in boldface. In the first example, the base system detected the term "retardation" to be the complete entity, while the system improved with negation-based transfer learning is able to capture the whole entity: "motor and mental retardation". On the other hand, the base model captured the entity "hearing impairment and mental retardation" as a single entity in the second example, while the negation-improved model correctly separates those entities as "hearing impairment" and "mental retardation".

Regarding the nature of the DIANN and RDD corpus, it is quite difficult to find examples to directly analyse how the system deals with negated entities. However, we consider that the pre-training stage performed on negation as a linguistic phenomenon provides the final system with more linguistic knowledge, eventually allowing it to detect more complex entities and relationships. In the shown examples, we observe how long entities are better detected ("motor and mental retardation"), and how different entities can be detected separately ("hearing impairment" and "mental retardation"). In general, incorporating knowledge on negation to the final system tends to modify the general trend on long annotations: the number of false positives is reduced, hence increasing the overall precision of the system.

In the following two examples, we can also observe how the system behaves in terms of the Relation Extraction subtask on the RDD corpus:

- "**Costello syndrome** was delineated based on its distinctive phenotype including severe failure-to-thrive with macrocephaly, characteristic facial features, hypertrophic cardiomyopathy, papillomata, malignant tumours, and **cognitive impairment**".
- "**Leber congenital amaurosis** (LCA) is the most severe retinal dystrophy causing **blindness** or **severe visual impairment** before the age of 1 year".

In the first example, the relationship between the entities "Costello syndrome" (rare disease) and "cognitive impairment" (disability) is not detected by the base model but is detected by the negation-based model. In the second example, the negation-based transfer learning step allows the model to correctly detect the relationship between "Leber congenital amaurosis" (rare disease), and the two entities representing disabilities: "blindness" and "severe visual impairment", while the base model was not able to detect those relationships.

In a similar way to that mentioned when it comes to entity detection, the relation extraction model seems to be improved by negation-based transfer learning in terms of detecting distant relationships within the text, as well as correctly locating relationships between one rare disease and multiple disabilities.

In addition to this analysis, a manual study of the BioScope corpus indicates that those scopes associated to negation actually reflect expressions similar to the entities of interest in this case (e.g., disabilities), despite dealing with different topics.

[2] Although the term "mental retardation" may be considered offensive and has sometimes been replaced by terms such as "intellectual disability" in the biomedical literature, it has been maintained in the examples since they are directly extracted from abstracts of already published scientific papers, which belong to the corpora used in the experiments of this research.

Table 11
Examples of negation triggers and scopes within the BioScope corpus.

<xcope><cue>without< /cue> antigenic stimulation</xcope> (without/IN antigenic/JJ stimulation/NN)
<xcope><cue>no</cue> detectable DPD activity</xcope> (no/DT detectable/JJ DPD/NNP activity/NN)
<xcope><cue>no</cue> Sp1 binding sites</xcope> (no/DT Sp1/NNP binding/NN sites/VBZ)
<xcope><cue>no</cue> enhancer activity of its own</xcope> (no/DT enhancer/NN activity/NN of/IN its/PRP$ own/JJ)

In the following examples, depicted in Table 11, the beginning and end of a scope is marked with the "xcope" and "/xcope" tags respectively, while the negations cues or triggers are represented between the "cue" and "/cue" tags. Part of Speech tags are also shown for each word in the sentence, using standard English PoS tags, as employed in the Penn Treebank POS tagset [75].

In these examples, POS patterns such as (JJ, NN), (JJ, NNP, NN), (NNP, NN, VBZ) or (NN, NN, IN, PRP, JJ) present in the detected scopes might represent an accurate source of information for detecting the scopes of the disabilities.

4.5. Contextual models

As mentioned in Section 2, pre-trained contextual models based on the Transformer architecture [76] have been used for performing similar tasks involving NER and RE in the last few years. More particularly, models such as BERT (Bidirectional Encoder Representations from Transformers) have shown to perform particularly well in many different NLP tasks, including those under study in this research. Therefore, additional experiments have been conducted in order to test whether pre-trained models are able to match or even overcome the results obtained by our proposed model.

In these experiments, we have employed BERT-based models for addressing Named Entity Recognition (NER) on the DIANN corpus and also on the RDD corpus. Although NER is just a subsystem of the whole system employed for our own experiments on the RDD corpus (Section 3.2.3), we consider that the results obtained by a BERT-based model on the NER subsystem might represent useful information for a potential comparison against our proposed model.

In particular, we have explored the use of the PubMedBERT model [77] for the English language. This biomedical model is pre-trained from scratch using abstracts from PubMed. For the Spanish language, the selected model is the RoBERTa-based biomedical model described in [78]. As the authors state in their paper, the pre-training of the Spanish model is performed using several biomedical corpora in the Spanish language. Default parameters of the selected models have been maintained, and a fine-tuning process has been performed on the selected corpora, also maintaining the main training parameters described in Sections 3.2.2 and 3.2.3, such as the early stopping scheme, maximum number of epochs or optimizers.

Tables 12–14 show the comparison between the best results regarding NER tasks obtained by the model proposed in this work, before and after applying negation-based transfer learning, and the results obtained by the BERT-based models on the same tasks. Tables 12 and 13 show results on the DIANN corpus (English and Spanish languages, respectively), while Table 14 refers to the NER subtask of the RDD corpus. Similarly to the experiments shown in Section 4.2, each experiment on the DIANN corpus is repeated 10 times using different seeds for the initialization values of the employed configurations for addressing the computation of statistical significance of the differences between them. Tables 12 and 13 show both the average and standard deviation for each proposed metric, in each experiment. Detailed results for each of the 10 repetitions of each experiment is provided in Appendix B.

Table 12
Comparison of results obtained by the proposed model, before and after the application of negation-based transfer learning (rows "Base Model" and "Negation", respectively), and by the BERT-based model (last row), on the DIANN dataset for the Exact metric in the English language. Bold indicates the best result for each metric. Star (*) indicates statistically significant differences with respect to the base model, for the same experiment, metric and language (Wilcoxon Signed Rank Test)..

	DIANN Corpus		
	English (BioScope)		
	Precision	Recall	F-measure
Base model	0.7280 ±0.0176	0.7450 ±0.0142	0.7362 ±0.0108
Negation	0.7676 ±0.0440(*)	**0.7723 ±0.0163**(*)	**0.7689 ±0.0154**(*)
BERT Model	0.7102 ±0.0213	0.7673 ±0.0258	0.7368 ±0.0090

Table 13
Comparison of results obtained by the proposed model, before and after the application of negation-based transfer learning (rows "Base Model" and "Negation", respectively), and by the BERT-based model (last row), on the DIANN dataset for the Exact metric in the Spanish language. Bold indicates the best result for each metric. Star (*) indicates statistically significant differences with respect to the base model, for the same experiment, metric and language (Wilcoxon Signed Rank Test)..

	DIANN Corpus		
	Spanish (SFU)		
	Precision	Recall	F-measure
Base model	0.7465 ±0.0174	0.6455 ±0.0305	0.6919 ±0.0191
Negation	0.7721 ±0.0242(*)	**0.7589 ±0.0233**(*)	**0.7649 ±0.0101**(*)
BERT Model	0.7186 ±0.0312	0.7356 ±0.0440	0.7254 ±0.0231

Table 14
Comparison of results obtained by the proposed model, before and after the application of negation-based transfer learning (rows "Base Model" and "Negation", respectively), and by the BERT-based model (last row), on the NER subtask of the RDD dataset. Bold indicates the best result for each metric.

	RDD Corpus (NER task)		
	Precision	Recall	F-measure
Base model	0.7410	0.7587	0.7498
Negation	**0.7596**	0.7661	0.7628
BERT model	0.7343	**0.7993**	**0.7654**

Results clearly show how, when it comes to the DIANN corpus, the employed BERT-based model is not able to overcome the results obtained by applying negation-based transfer learning on the proposed model. It can be seen how results offered by BERT are not too far from those reported in this work, which indicates that the transfer learning represented by the use of the pre-trained contextual model is also able to capture some of the information that is needed for performing accurate NER in this context. However, our model is still able to offer the best overall results both in the English and the Spanish language, which indicates that negation-based transfer learning represents a more adequate approach for the addressed task. The table also indicates that the results obtained by the negation-based model are significantly better according to the Wilcoxon signed rank test, with respect to both the base model and the BERT-based model.

Regarding the NER subtask of the RDD corpus, results show how the transfer learning techniques represented by the use of the BERT-based model are able to produce results comparable to those obtained by the proposed negation-based transfer learning technique. The differences between results achieved by the negation-based transfer learning technique and by the BERT-based model are actually not statistically significant (Wilcoxon signed rank test). However, considering the time and resource consumption derived from the use (pre-training and fine-tuning) of pre-trained contextual models, the proposed model could still be preferred in particular scenarios where these resources were not so readily available and where some negation-annotated corpus is available. As we have observed in the different experiments conducted, corpora used for performing negation detection need not necessarily

be specific to the domain of the data on which knowledge transfer is intended. In particular, in this work previously existing corpora have been employed, and, as in the case of SFU, some of those corpora belong to a different domain.

In addition, this study, by considering specific phenomena such as negation, separately from the tasks addressed (NER and RE), provides some intuitions about the type of information that could be handled by other models such as those based on Transformer architectures, which address the problems in a global way.

5. Conclusions and future work

In this paper we have presented a novel research on the influence that the acquisition of knowledge about negation detection may have in different tasks of Biomedical Natural Language Processing, in particular Named Entity Recognition and Relation Extraction. A specific deep learning architecture have been proposed and adapted in order to perform detection of negation triggers and scopes in two different languages: English and Spanish. This previously trained neural model has been then used for studying the impact of transferring this type of knowledge into other different neural network architectures, developed for named entity recognition and relation extraction tasks. In particular, the transfer learning is performed on the initial weights of the Bi-LSTM layers used for performing named entity recognition. The main purpose is to share the representation space generated during the training of the negation model with those models performing more complex tasks. Different corpora written in the two proposed languages have been employed: BioScope in English and SFU Review SP-NEG corpus in Spanish, both for training the negation detection model, the DIANN corpus in English and Spanish for named entity recognition, and the RDD corpus in English for joint named entity recognition and relation extraction. Through the use of different languages, corpora and tasks, we intend to cover a wide range of possibilities, in order to present the most robust and extrapolable possible results.

The achieved results are quite satisfactory and positive, since we have obtained significant improvements in the two main tasks that have been tackled within this research: Named Entity Recognition and Relation Extraction. In the first task, negation detection has an overall positive impact in all the studied cases, either due to improvements in precision or in recall, which unfailingly lead to enhance the final F-Measures. When it comes to relation extraction, which has been studied jointly with named entity recognition, the improvements introduced by negation-based transfer learning in NER rapidly spread along the joint model and have an impact on the results of the relation extraction subtask.

Future lines of work include the study of additional tasks related to biomedical and general domain natural language processing, which are normally influenced by negated entities and hence could potentially benefit from performing previous negation detection. This particular characteristic can be applied to many different NLP tasks and more complex pipelines which usually include named entity recognition, such as online content discovery (information extraction), efficient search algorithms, recommender systems or patient support. A different line of research would focus on modifying the neural models developed for this work, not only their main characteristics and parameters, but also the specific parts of the models in which the transfer learning could be performed. For instance, a named entity recognition model could present two different Bi-LSTM layers that could be combined later on. One of the layers could incorporate weights from previous negation detection tasks, while the other would be always randomly initialized. This way, negation information would be introduced in the model as an independent input feature represented in a complex way instead of within a shared representation space which may conflict with the representation of other input features. The representation schemes used in this work can also be effectively improved, for instance by incorporating different embedding models or studying the potential influence of contextual models for representing the input information. A systematic study on the impact that different linguistic characteristics such as parsing might have on the NER and RE tasks is also an intended future line of work. Finally, regarding the NER+RE joint model, another possible improvement would be the analysis of more sophisticated ways of combining the loss functions from both subsystems for their propagation through the network.

Table A.15
Statistical significance analysis for the differences between two pairs of experiments on the DIANN corpus: final results with the base model and with the negation-based transfer learning model, for both the English and Spanish language in the exact metric without using post-processing rules. Last row indicates whether the differences are statistically significant for all the considered metrics, according to the Wilcoxon Signed Rank Test.

		Exact match, no rules					
		English			Spanish		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Base model	Seed 0	0.7488	0.6639	0.7038	0.7895	0.5357	0.6383
	Seed 1	0.7028	0.6261	0.6622	0.7537	0.4509	0.5642
	Seed 2	0.7573	0.6555	0.7027	0.7682	0.5179	0.6187
	Seed 3	0.7286	0.6429	0.6830	0.7534	0.4911	0.5946
	Seed 4	0.7692	0.6303	0.6928	0.8000	0.5179	0.6287
	Seed 5	0.7423	0.6050	0.6667	0.7467	0.5000	0.5989
	Seed 6	0.7475	0.6218	0.6789	0.7310	0.4732	0.5745
	Seed 7	0.7327	0.6218	0.6727	0.7665	0.5714	0.6547
	Seed 8	0.7095	0.6261	0.6652	0.7469	0.5402	0.6269
	Seed 9	0.7269	0.6597	0.6916	0.7842	0.4866	0.6006
Negation	Seed 0	0.7746	0.6933	0.7317	0.8251	0.6741	0.7420
	Seed 1	0.7696	0.7017	0.7341	0.7772	0.7009	0.7371
	Seed 2	0.7951	0.6849	0.7359	0.7979	0.6875	0.7386
	Seed 3	0.7545	0.7101	0.7316	0.8483	0.6741	0.7512
	Seed 4	0.8785	0.6681	0.7589	0.7861	0.7054	0.7435
	Seed 5	0.7902	0.6807	0.7314	0.8352	0.6786	0.7488
	Seed 6	0.7425	0.7269	0.7346	0.8409	0.6607	0.7400
	Seed 7	0.7913	0.6849	0.7342	0.7778	0.6875	0.7299
	Seed 8	0.8634	0.6639	0.7506	0.8000	0.6786	0.7343
	Seed 9	0.8290	0.6723	0.7425	0.7908	0.6920	0.7381
p-value < 0.05		YES	YES	YES	YES	YES	YES

CRediT authorship contribution statement

Hermenegildo Fabregat: Conceptualization, Methodology, Software, Investigation, Validation, Writing – review & editing. **Andres Duque:** Conceptualization, Methodology, Investigation, Validation, Writing – original draft. **Juan Martinez-Romo:** Conceptualization, Methodology, Investigation, Validation, Writing – review & editing, Supervision. **Lourdes Araujo:** Conceptualization, Methodology, Investigation, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32 and OBSERMENH, Spain Project (MCIN/AEI/10.13039/501100011033 and NextGenerationEU/PRTR) under Grant TED2021-130398B-C21 as well as project RAICES (IMIENS 2022).

Appendix A. Detailed results: NER experiments (DIANN corpus)

See Tables A.15–A.18.

Table A.16
Statistical significance analysis for the differences between two pairs of experiments on the DIANN corpus: final results with the base model and with the negation-based transfer learning model, for both the English and Spanish language in the partial metric without using post-processing rules. Last row indicates whether the differences are statistically significant for all the considered metrics, according to the Wilcoxon Signed Rank Test.

		Partial match, no rules					
		English			Spanish		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Base model	Seed 0	0.8483	0.7521	0.7973	0.8618	0.5848	0.6968
	Seed 1	0.7877	0.7017	0.7422	0.8358	0.5000	0.6257
	Seed 2	0.8447	0.7311	0.7838	0.8411	0.5670	0.6773
	Seed 3	0.8143	0.7185	0.7634	0.8356	0.5446	0.6595
	Seed 4	0.8615	0.7059	0.7760	0.8552	0.5536	0.6721
	Seed 5	0.8402	0.6849	0.7546	0.8200	0.5491	0.6578
	Seed 6	0.8535	0.7101	0.7752	0.8207	0.5313	0.6450
	Seed 7	0.8416	0.7143	0.7727	0.8323	0.6205	0.7110
	Seed 8	0.8286	0.7311	0.7768	0.8086	0.5848	0.6788
	Seed 9	0.8241	0.7479	0.7841	0.8705	0.5402	0.6667
Negation	Seed 0	0.8451	0.7563	0.7982	0.8907	0.7277	0.8010
	Seed 1	0.8387	0.7647	0.8000	0.8366	0.7545	0.7934
	Seed 2	0.8634	0.7437	0.7991	0.8756	0.7545	0.8106
	Seed 3	0.8170	0.7689	0.7922	0.9045	0.7188	0.8010
	Seed 4	0.9448	0.7185	0.8162	0.8458	0.7589	0.8000
	Seed 5	0.8732	0.7521	0.8081	0.9011	0.7321	0.8079
	Seed 6	0.8026	0.7857	0.7941	0.8864	0.6964	0.7800
	Seed 7	0.8689	0.7521	0.8063	0.8434	0.7455	0.7915
	Seed 8	0.9290	0.7143	0.8076	0.8789	0.7455	0.8068
	Seed 9	0.8912	0.7227	0.7981	0.8469	0.7411	0.7905
p-value < 0.05		YES	NO	YES	YES	YES	YES

Table A.17
Statistical significance analysis for the differences between two pairs of experiments on the DIANN corpus: final results with the base model and with the negation-based transfer learning model, for both the English and Spanish language in the exact metric using post-processing rules. Last row indicates whether the differences are statistically significant for all the considered metrics, according to the Wilcoxon Signed Rank Test.

		Exact match, with rules					
		English			Spanish		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Base model	Seed 0	0.7449	0.7605	0.7526	0.7653	0.6696	0.7143
	Seed 1	0.6965	0.7521	0.7232	0.7403	0.5982	0.6617
	Seed 2	0.7377	0.7563	0.7469	0.7525	0.6652	0.7062
	Seed 3	0.7317	0.7563	0.7438	0.7409	0.6384	0.6859
	Seed 4	0.7500	0.7437	0.7468	0.7807	0.6518	0.7105
	Seed 5	0.7414	0.7227	0.7319	0.7323	0.6473	0.6872
	Seed 6	0.7362	0.7269	0.7315	0.7302	0.6161	0.6683
	Seed 7	0.7202	0.7353	0.7277	0.7311	0.6920	0.7110
	Seed 8	0.7085	0.7353	0.7216	0.7317	0.6696	0.6993
	Seed 9	0.7126	0.7605	0.7358	0.7598	0.6071	0.6749
Negation	Seed 0	0.7379	0.7689	0.7531	0.7887	0.7500	0.7689
	Seed 1	0.7430	0.7773	0.7598	0.7510	0.8080	0.7785
	Seed 2	0.7811	0.7647	0.7728	0.7609	0.7813	0.7709
	Seed 3	0.7068	0.7899	0.7460	0.8049	0.7366	0.7692
	Seed 4	0.8333	0.7563	0.7930	0.7380	0.7545	0.7461
	Seed 5	0.7625	0.7689	0.7657	0.8029	0.7455	0.7731
	Seed 6	0.7138	0.8067	0.7574	0.7961	0.7321	0.7628
	Seed 7	0.7645	0.7773	0.7708	0.7489	0.7589	0.7539
	Seed 8	0.8219	0.7563	0.7877	0.7661	0.7455	0.7557
	Seed 9	0.8108	0.7563	0.7826	0.7632	0.7768	0.7699
p-value < 0.05		YES	YES	YES	YES	YES	YES

Appendix B. Detailed results: NER experiments (DIANN and BERT comparative)

See Table B.19.

Table A.18
Statistical significance analysis for the differences between two pairs of experiments on the DIANN corpus: final results with the base model and with the negation-based transfer learning model, for both the English and Spanish language in the partial metric using post-processing rules. Last row indicates whether the differences are statistically significant for all the considered metrics, according to the Wilcoxon Signed Rank Test.

		Partial match, with rules					
		English			Spanish		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Base model	Seed 0	0.8642	0.8824	0.8732	0.8520	0.7455	0.7952
	Seed 1	0.7977	0.8613	0.8283	0.8453	0.6830	0.7556
	Seed 2	0.8443	0.8655	0.8548	0.8384	0.7411	0.7867
	Seed 3	0.8374	0.8655	0.8512	0.8446	0.7277	0.7818
	Seed 4	0.8602	0.8529	0.8565	0.8556	0.7143	0.7786
	Seed 5	0.8578	0.8361	0.8468	0.8283	0.7321	0.7773
	Seed 6	0.8596	0.8487	0.8541	0.8307	0.7009	0.7603
	Seed 7	0.8436	0.8613	0.8524	0.8113	0.7679	0.7890
	Seed 8	0.8421	0.8739	0.8577	0.8098	0.7411	0.7739
	Seed 9	0.8268	0.8824	0.8537	0.8659	0.6920	0.7692
Negation	Seed 0	0.8306	0.8655	0.8477	0.8732	0.8304	0.8513
	Seed 1	0.8353	0.8739	0.8542	0.8257	0.8884	0.8559
	Seed 2	0.8755	0.8571	0.8662	0.8522	0.8750	0.8634
	Seed 3	0.7895	0.8824	0.8333	0.8829	0.8080	0.8438
	Seed 4	0.9259	0.8403	0.8811	0.8166	0.8348	0.8256
	Seed 5	0.8667	0.8739	0.8703	0.8894	0.8259	0.8565
	Seed 6	0.7955	0.8992	0.8442	0.8641	0.7946	0.8279
	Seed 7	0.8636	0.8782	0.8708	0.8326	0.8438	0.8381
	Seed 8	0.9132	0.8403	0.8753	0.8624	0.8393	0.8507
	Seed 9	0.9009	0.8403	0.8696	0.8377	0.8527	0.8451
p-value < 0.05		NO	NO	NO	NO	YES	YES

Table B.19
Statistical significance analysis for the differences between two pairs of experiments on the DIANN corpus: final results with negation-based transfer learning (with rules) and with the BERT-based model, for both the English and Spanish language in the exact metric. Last row indicates whether the differences are statistically significant for all the considered metrics, according to the Wilcoxon Signed Rank Test.

		English			Spanish		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Negation	Seed 0	0.7379	0.7689	0.7531	0.7887	0.7500	0.7689
	Seed 1	0.7430	0.7773	0.7598	0.7510	0.8080	0.7785
	Seed 2	0.7811	0.7647	0.7728	0.7609	0.7813	0.7709
	Seed 3	0.7068	0.7899	0.7460	0.8049	0.7366	0.7692
	Seed 4	0.8333	0.7563	0.7930	0.7380	0.7545	0.7461
	Seed 5	0.7625	0.7689	0.7657	0.8029	0.7455	0.7731
	Seed 6	0.7138	0.8067	0.7574	0.7961	0.7321	0.7628
	Seed 7	0.7645	0.7773	0.7708	0.7489	0.7589	0.7539
	Seed 8	0.8219	0.7563	0.7877	0.7661	0.7455	0.7557
	Seed 9	0.8108	0.7563	0.7826	0.7632	0.7768	0.7699
BERT Model	Seed 0	0.7270	0.7770	0.7484	0.7491	0.7148	0.7315
	Seed 1	0.7125	0.7677	0.7391	0.6903	0.7967	0.7397
	Seed 2	0.7422	0.7172	0.7295	0.7183	0.7607	0.7389
	Seed 3	0.7214	0.7845	0.7516	0.6601	0.7730	0.7121
	Seed 4	0.6886	0.7744	0.7290	0.7500	0.7574	0.7535
	Seed 5	0.6695	0.8047	0.7309	0.7096	0.7049	0.7032
	Seed 6	0.7051	0.7407	0.7225	0.7206	0.6426	0.6794
	Seed 7	0.7294	0.7441	0.7367	0.7043	0.7599	0.7310
	Seed 8	0.7052	0.7811	0.7412	0.7671	0.7344	0.7504
	Seed 9	0.7009	0.7811	0.7389	0.7162	0.7115	0.7138
p-value < 0.05		YES	YES	YES	YES	YES	YES

References

[1] I.G. Councill, R.T. McDonald, L. Velikovich, What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis, in: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, NeSp-NLP@ACL 2010, Uppsala - Sweden, University of Antwerp, 2010, pp. 51–59.

[2] M.F.M. Chowdhury, A. Lavelli, FBK-irst : A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information, in: Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta - USA, ACL, 2013, pp. 351–355.

[3] K.A. Spackman, K.E. Campbell, R.A. Côté, SNOMED RT: A reference terminology for health care, in: AMIA 1997, American Medical Informatics Association Annual Symposium, Nashville, TN, USA, AMIA, 1997, pp. 640–644.

[4] D.A. Lindberg, B.L. Humphreys, A.T. McCray, The unified medical language system, Methods Inf. Med. 32 (4) (1993) 281–291.

[5] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program, in: Proceedings. AMIA Symposium, 2001, pp. 17–21.

[6] M.G. Kersloot, F. Lau, A. Abu Hanna, D.L. Arts, R. Cornet, Automated SNOMED CT concept and attribute relationship detection through a web-based implementation of cTAKES, J. Biomed. Semant. 10 (1) (2019) 14.

[7] M. Dai, N.H. Shah, W. Xuan, M.A. Musen, S.J. Watson, B.D. Athey, F. Meng, et al., An efficient solution for mapping free text to ontology terms, AMIA Summit Transl. Bioinform. 21 (2008).

[8] N. Perez, P. Accuosto, À. Bravo, M. Cuadros, E. Martínez-Garcia, H. Saggion, G. Rigau, Cross-lingual semantic annotation of biomedical literature: Experiments in Spanish and English, Bioinformatics 36 (6) (2020) 1872–1880.

[9] M. Oronoz, A. Casillas, K. Gojenola, A. Perez, Automatic annotation of medical records in Spanish with disease, drug and substance names, in: Iberoamerican Congress on Pattern Recognition, Springer, 2013, pp. 536–543.

[10] V. Cotik, L.A. Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L.D. Francesca, A. Dellanzo, M.F. Urquiza, Overview of CLEF eHealth task 1 - SpRadIE: A challenge on information extraction from Spanish radiology reports, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021, pp. 732–750.

[11] A. Piad-Morfis, S. Estevez-Velarde, Y. Gutiérrez, Y. Almeida-Cruz, A. Montoyo, R. Muñoz, Overview of the ehealth knowledge discovery challenge at IberLEF 2021, Proces. Del Leng. Nat. 67 (2021) 233–242, URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392.

[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[13] H. Cho, H. Lee, Biomedical named entity recognition using deep neural networks with contextual information, BMC Bioinformatics 20 (1) (2019) 735.

[14] A. Casillas, N. Ezeiza, I. Goenaga, A. Pérez, X. Soto, Measuring the effect of different types of unsupervised word representations on medical named entity recognition, Int. J. Med. Inform. 129 (2019) 100–106, http://dx.doi.org/10.1016/j.ijmedinf.2019.05.022, URL https://www.sciencedirect.com/science/article/pii/S1386505618310311.

[15] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610, http://dx.doi.org/10.1016/j.neunet.2005.06.042.

[16] Z. Huang, W. Xu, K. Yu, Bidirectional LSTM-CRF models for sequence tagging, 2015, CoRR abs/1508.01991, arXiv:1508.01991.

[17] H. Fabregat, A.D. Fernandez, J. Martinez-Romo, L. Araujo, NLP_UNED at eHealth-KD challenge 2019: Deep learning for named entity recognition and attentive relation extraction, in: IberLEF@ SEPLN, 2019, pp. 67–77.

[18] A. Bravo, P. Accuosto, H. Saggion, LastUS-TALN at IberLEF 2019 eHealth-KD challenge: Deep learning approaches to information extraction in biomedical texts, in: IberLEF@ SEPLN, 2019, pp. 51–59.

[19] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, Lsi_uned at CLEF ehealth2021: Exploring the effects of transfer learning in negation detection and entity recognition in clinical texts, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, in: CEUR Workshop Proceedings, vol. 2936, CEUR-WS.org, 2021, pp. 780–793, URL http://ceur-ws.org/Vol-2936/paper-65.pdf.

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

[21] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions, J. Biomed. Inform. 46 (5) (2013) 914–920, http://dx.doi.org/10.1016/j.jbi.2013.07.011, URL https://www.sciencedirect.com/science/article/pii/S1532046413001123.

[22] I.N. Dewi, S. Dong, J. Hu, Drug-drug interaction relation extraction with deep convolutional neural networks, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine, 2017, pp. 1795–1802.

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

[24] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H. Kriegel, Extraction of semantic biomedical relations from text using conditional random fields, BMC Bioinform 9 (2008).

[25] L. Lazib, Y. Zhao, B. Qin, T. Liu, Negation scope detection with recurrent neural networks models in review texts, in: Proceedings, Part I: Social Computing - Second International Conference of Young Computer Scientists, Engineers and Educators, Harbin - China, in: Communications in Computer and Information Science, vol. 623, Springer, 2016, pp. 494–508.

[26] F. Fancellu, A. Lopez, B.L. Webber, Neural networks for negation scope detection, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1, Berlin - Germany, ACL, 2016, pp. 495–504.

[27] N. Konstantinova, S.C.M. de Sousa, N.P.C. Díaz, M.J.M. López, M. Taboada, R. Mitkov, A review corpus annotated for negation, speculation and their scope, in: N. Calzolari, K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, European Language Resources Association (ELRA), 2012, pp. 3190–3195, URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/533.html.

[28] R. Morante, W. Daelemans, ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC'12, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 1563–1568, URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/221_Paper.pdf.

[29] F. Fancellu, A. Lopez, B.L. Webber, H. He, Detecting negation scope is easy, except when it isn't, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 2, EACL 2017, Valencia - Spain, ACL, 2017, pp. 58–63.

[30] V. Vincze, G. Szarvas, R. Farkas, G. Móra, J. Csirik, The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes, BMC Bioinform. 9 (S-11) (2008).

[31] Z. Qian, P. Li, Q. Zhu, G. Zhou, Z. Luo, W. Luo, Speculation and negation scope detection via convolutional neural networks, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, ACL, 2016, pp. 815–825.

[32] J.M. Giorgi, G.D. Bader, Transfer learning for biomedical named entity recognition with neural networks, Bioinformatics 34 (23) (2018) 4087–4094.

[33] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers), NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Association for Computational Linguistics, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/n19-1423.

[34] M.E. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-supervised sequence tagging with bidirectional language models, in: R. Barzilay, M. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers, ACL 2017, Vancouver, Canada, July 30 - August 4, Association for Computational Linguistics, 2017, pp. 1756–1765, http://dx.doi.org/10.18653/v1/P17-1161.

[35] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, Bioinform. 36 (4) (2020) 1234–1240, http://dx.doi.org/10.1093/bioinformatics/btz682.

[36] Q. Jin, B. Dhingra, W.W. Cohen, X. Lu, Probing biomedical embeddings from language models, 2019, arXiv preprint arXiv:1904.02181.

[37] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in: D. Demner-Fushman, K.B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019, Association for Computational Linguistics, 2019, pp. 58–65, http://dx.doi.org/10.18653/v1/w19-5006.

[38] C. Sun, Z. Yang, Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task, in: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 100–104, http://dx.doi.org/10.18653/v1/D19-5715, URL https://aclanthology.org/D19-5715.

[39] Z. Chai, H. Jin, S. Shi, S. Zhan, L. Zhuo, Y. Yang, Hierarchical shared transfer learning for biomedical named entity recognition, BMC Bioinformatics 23 (1) (2022) 1–14.

[40] A. Agrawal, S. Tripathi, M. Vardhan, V. Sihag, G. Choudhary, N. Dragoni, BERT-based transfer-learning approach for nested named-entity recognition using joint labeling, Appl. Sci. 12 (3) (2022) 976.

[41] J. Chen, B. Hu, W. Peng, Q. Chen, B. Tang, Biomedical relation extraction via knowledge-enhanced reading comprehension, BMC Bioinformatics 23 (1) (2022) 1–19.

[42] R. Gubelmann, S. Handschuh, Context matters: A pragmatic study of PLMs' negation understanding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4602–4621, http://dx.doi.org/10.18653/v1/2022.acl-long.315, URL https://aclanthology.org/2022.acl-long.315.

[43] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: J. Kim, C. Nédellec, R. Bossy, L. Deléger (Eds.), Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, BioNLP-OST@EMNLP-IJNCLP 2019, Hong Kong, China, November 4, 2019, Association for Computational Linguistics, 2019, pp. 1–10, http://dx.doi.org/10.18653/v1/D19-5701.

[44] D. Benikova, C. Biemann, M. Reznicek, NoSta-D named entity annotation for German: Guidelines and dataset, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources

and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014, European Language Resources Association (ELRA), 2014, pp. 2524–2531, URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/276.html.

[45] J. Li, Y. Sun, R.J. Johnson, D. Sciaky, C. Wei, R. Leaman, A.P. Davis, C.J. Mattingly, T.C. Wiegers, Z. Lu, BioCreative V CDR task corpus: A resource for chemical disease relation extraction, Database J. Biol. Databases Curation 2016 (2016) http://dx.doi.org/10.1093/database/baw068.

[46] S.K. Sahu, F. Christopoulou, M. Miwa, S. Ananiadou, Inter-sentence relation extraction with document-level graph convolutional neural network, in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, Vol. 1: Long Papers, ACL 2019, Florence, Italy, July 28- August 2, 2019, Association for Computational Linguistics, 2019, pp. 4309–4316, http://dx.doi.org/10.18653/v1/p19-1423.

[47] S.M.J. Zafra, M. Taulé, M.T. Martín-Valdivia, L.A.U. López, M.A. Martí, SFU ReviewSP-NEG: A Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns, Lang. Resour. Eval. 52 (2) (2018) 533–569.

[48] H. Fabregat, J. Martínez-Romo, L. Araujo, Overview of the DIANN task: Disability annotation task, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 1–14.

[49] H. Fabregat, L. Araujo, J. Martínez-Romo, Deep neural models for extracting entities and relationships in the new RDD corpus relating disabilities and rare diseases, Comput. Methods Programs Biomed. 164 (2018) 121–129.

[50] P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages co-located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain, in: CEUR Workshop Proceedings, vol. 2150, 2018.

[51] H. Fabregat, J. Martínez-Romo, L. Araujo, Understanding and improving disability identification in medical documents, IEEE Access 8 (2020) 155399–155408, http://dx.doi.org/10.1109/ACCESS.2020.3019178.

[52] H.F. y Lourdes Araujo y Juan Martínez-Romo, Deep learning approach for negation trigger and scope recognition, Procesamiento Del Lenguaje Nat. 62 (2019) 37–44, URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5950.

[53] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, H. Xu, Entity recognition from clinical texts via recurrent neural network, BMC Med. Inform. Decis. Mak. 17 (2) (2017) 53–61, http://dx.doi.org/10.1186/s12911-017-0468-7.

[54] W. Ling, C. Dyer, A.W. Black, I. Trancoso, R. Fermandez, S. Amir, L. Marujo, T. Luís, Finding function in form: Compositional character models for open vocabulary word representation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1520–1530, http://dx.doi.org/10.18653/v1/D15-1176, URL https://aclanthology.org/D15-1176.

[55] R.M.R. Zavala, P. Martínez, I. Segura-Bedmar, A hybrid bi-LSTM-CRF model to recognition of disabilities from biomedical texts, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 44–52.

[56] H. Fabregat, L. Araujo, J. Martínez-Romo, Deep learning approach for negation trigger and scope recognition, Proces. Del Leng. Natural 62 (2019) 37–44, URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5950.

[57] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C.D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: A. Celikyilmaz, T. Wen (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 101–108, http://dx.doi.org/10.18653/v1/2020.acl-demos.14.

[58] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P.P. Kuksa, Natural language processing (Almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537, http://dx.doi.org/10.5555/1953048.2078186, URL https://dl.acm.org/doi/10.5555/1953048.2078186.

[59] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013, URL http://arxiv.org/abs/1301.3781.

[60] C. Cardellino, Spanish billion words corpus and embeddings, 2019, URL https://crscardellino.github.io/SBWCE/.

[61] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, Distributional semantics resources for biomedical text processing, in: Proceedings of LBM 2013, 2013, pp. 39–44, URL http://lbm2013.biopathway.org/lbm2013proceedings.pdf.

[62] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, H. Isahara, Named entity extraction based on a maximum entropy model and transformation rules, in: 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong - China, ACL, 2000, pp. 326–335.

[63] I. Goenaga, A. Atutxa, K. Gojenola, A. Casillas, A.D. de Ilarraza, N. Ezeiza, M. Oronoz, A. Pérez, O. Perez-de-Viñasprre, A hybrid approach for automatic disability annotation, in: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, Sevilla - Spain, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 31–36.

[64] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.

[65] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 270–279.

[66] R. Collobert, J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, Helsinki - Finland, ICML '08, ACM, 2008, pp. 160–167.

[67] K.-i. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, et al., Toward information extraction: Identifying protein names from biological papers, in: Pac Symp Biocomput, Vol. 707, 1998, pp. 707–718.

[68] S. Medina, J. Turmo, H. Loharja, L. Padró, Semi-supervised learning for disabilities detection on English and Spanish biomedical text, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, SEPLN 2018, Sevilla, Spain, September 18th, 2018, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 66–73, URL http://ceur-ws.org/Vol-2150/DIANN_paper8.pdf.

[69] R. Agerri, J. Bermudez, G. Rigau, IXA pipeline: Efficient and ready to use multilingual NLP tools, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014, European Language Resources Association (ELRA), 2014, pp. 3823–3828, URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/775.html.

[70] R. Agerri, G. Rigau, Simple language independent sequence labelling for the annotation of disabilities in medical texts, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, SEPLN 2018, Sevilla, Spain, September 18th, 2018, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 25–30, URL http://ceur-ws.org/Vol-2150/DIANN_paper2.pdf.

[71] O. Bodenreider, The unified medical language system (UMLS): Integrating biomedical terminology, Nucleic Acids Res. 32 (Database-Issue) (2004) 267–270, http://dx.doi.org/10.1093/nar/gkh061.

[72] P. López-Úbeda, M.C. Díaz-Galiano, M.T. Martín-Valdivia, S.M.J. Zafra, SINAI at DIANN - IberEval 2018. Annotating disabilities in multi-language systems with UMLS, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, SEPLN 2018, Sevilla, Spain, September 18th, 2018, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 37–43, URL http://ceur-ws.org/Vol-2150/DIANN_paper4.pdf.

[73] H. Fabregat, J. Martínez-Romo, L. Araujo, UNED at DIANN 2018: Unsupervised system for automatic disabilities labeling in medical scientific documents, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, SEPLN 2018, Sevilla, Spain, September 18th, 2018, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 53–60, URL http://ceur-ws.org/Vol-2150/DIANN_paper6.pdf.

[74] I. Moreno, M.T. Romá-Ferri, P. Moreda, GPLSIUA team at the DIAAN 2018 task, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J.C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) Co-Located with 34th Conference of the Spanish Society for Natural Language Processing, SEPLN 2018, Sevilla, Spain, September 18th, 2018, in: CEUR Workshop Proceedings, vol. 2150, CEUR-WS.org, 2018, pp. 15–24, URL http://ceur-ws.org/Vol-2150/DIANN_paper1.pdf.

[75] B. Santorini, Part-Of-Speech Tagging Guidelines for the Penn Treebank Project, Tech. Rep. MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990, URL ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz.

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach,

CA, USA, 2017, pp. 5998–6008, URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[77] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2020, arXiv:arXiv:2007.15779.

[78] C.P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021, arXiv:2109.03570.