A Hybrid Bi-LSTM-CRF model to Recognition of Disabilities from biomedical texts

Renzo M. Rivera Zavala[1], Paloma Martinez[1], and Isabel Segura-Bedmar[1]

Computer Science Department, Universidad Carlos III de Madrid, Spain

Abstract. This paper describes the UC3M_018_1 participation. The model uses a two-phase model based on two layers of Bidirectional LSTM (Long Short-Term Memory) to capture context information and CRF (Conditional Random Fields) to obtain the correlation of the information between the labels. An interesting detail is that the authors jointly deal with entity detection and negation detection, which makes this approach a seq2seq multilabel classification problem. The system was presented in the task IberEval 2018 of Disabilities Annotation, achieving a satisfactory performance in the 2nd place in terms of F-Score (68.2 %) for the English task and (65.3%) for the Spanish task.

Keywords: NER · Bi-LSTM · CRF · Disability.

1 Introduction

The generation of Disability Annotations (DIANN) is the task of identifying disabilities in biomedical domain texts using computer methods. DIANN is a task of IberEval 2018 [1]. There are several approaches to address the problem of Named Entity Recognition (NER). The most popular methods are focused on the extraction of syntax, lexical, semantic, morphological and grammar patterns related to the structure of language. However, although the task of labeling many examples is not necessary, the ability to generalize and learn from new examples is poor or non-existent, as well as requiring linguistic knowledge and the specific domain. On the other hand, systems based on machine learning can automatically extract relevant patterns from the entities using annotated corpora, which allows the independence of specific domains or languages. Currently, many methods based on Deep Learning have been proposed to solve the NER problem achieving high performance with the definition of few features.

The NER task can be seen as a tag sequence problem. However, most machine learning models do not consider the sequence in which the input data was used by the training. Therefore, currently one of the most used models is the Recurrent Neural Network (RNN), specifically Long Short-Term Memory (LSTM) [3] due to its ability to keep in memory and relate parts of a sequence. There are multiple models for the NER problem that make use of RNN, being those that obtain better results hybrid models that implement more than one machine learning model, specifically models that combine long short-term memory

networks with conditional random fields (Bi -LSTM + CRF) [2]. In this work the NeuroNER model proposed by [4] is extended and applied to the task of DIANN. The NeuroNER model was enlarged in its initialization to capture context information regarding Part-of-Speech (POS) tags and for the recognition of overlapping or nested named entities. Likewise, in order to reduce training time and improve the results, learning transfer of pre-trained models of words embeddings GloVe [5] trained on the Wikipedia 2014 and GigaWord5 corpus and sense-disambiguation embedding Sense2Vec [6] trained on the Reddit corpus was performed. The study of the task shows that the proposed model achieves an exceptional performance for the task of DIANN.

2 Proposal

In this section we present the proposed model that consists of four processes: preprocessing, learning transfer, classification and post-processing. Both the training process and the validation process execute each of the after mentioned processes. The architecture and flow of the activities developed for our model is shown in Fig. 1.

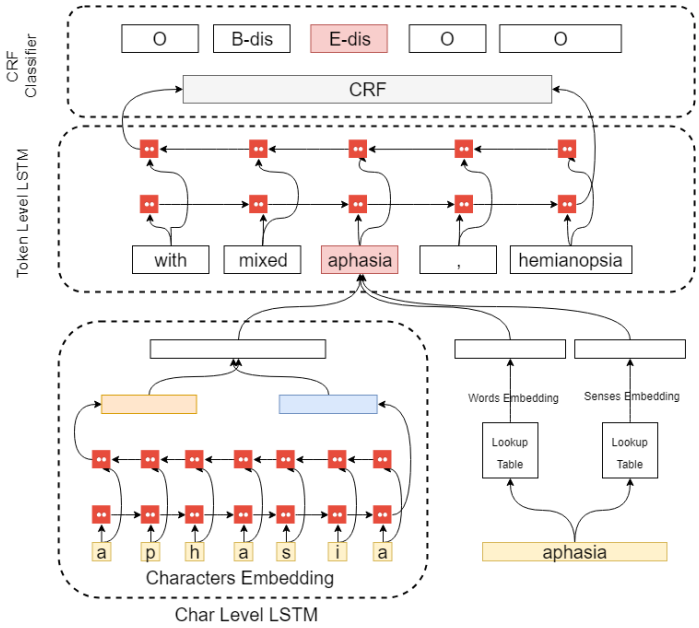

Fig. 1. Overview architecture of our hybrid LSTM-CRF model.

2.1 Datasets

The DIANN disability corpus is a gold standard corpus for the recognition of disabilities concepts. The corpus includes 500 abstracts of journals published in Elsevier related to the biomedical domain between the years 2017 and 2018. Only articles that had abstracts in English and Spanish were defined as inclusion criteria. Because the limits between disabilities and diseases are not clearly defined, a list of disability and function terms where the absence or limitation of the function is defined as a disability was considered. For the training and validation of the model the training dataset was divided into two parts: the training set consisting of 300 annotated abstracts and the validation set consisting of 100 annotated abstracts, the division of the training set was carried out randomly and subsequently stored for reproduction. Finally, the test set consisted of 100 abstracts without annotations.

2.2 Preprocessing

All abstracts in DIANN disabilities corpus were preprocessed in the following order: (1) sentence segmentation. Spacy [7] is an open source library for advanced natural language processing with support for 26 languages, it was used to divide the abstract text into sentences that are used by the model as input units. (2) Transformation of the XML label schema to the BRAT format [8]. BRAT is a web-based tool for the annotations representation following a specific annotation structure, specifically the annotation guidelines proposed in ACE 2005. (3) Tokenization, text segmentation in words and punctuation marks among others. (4) BMEWO-V extended tag encoding, to capture information about the structure and sequence of the tokens in the sentence. The BMEWO-V or BIOES-V encoding distinguishes B tags to indicate the start of an entity, the M or I tag to indicate the continuity of an entity, the E tag to indicate the end of an entity, the W or S tag for indicate a single entity, tag O to represent other elements that are not entities and finally label V added to this encoding to identify overlapping entities. This encoding format allows the detection of discontinuous entities, as well as overlapping or nested entities.

2.3 Learning transfer

Words Embedding Words embedding are distribution representations, expressing words or tokens as a vector, of low dimensionality and real value that can contain syntactic and semantic information for each of its dimensions. There are three models for the training of embedded words, the embedded word prediction model word2vec [9], the global aggregate model of word-word co-occurrence statistics [5] and the model morphological representation of fastText words [10]. In this work we used pre-trained GloVe models with 6 trillion words represented in 100 features vectors trained on 2014 Wikipedia and Gigaword 5 corpus for English disabilities annotations and GloVe with 1 billion words represented in 300 features vector trained on 2014 Wikipedia corpus for Spanish disabilities annotations.

Sense-Disambiguation Embedding embedding disambiguated senses are distribution representations, expressing the senses or meanings of words or tokens within a context, that is, a representation of a sentence as a vector of low dimensionality and real value that can contain syntactic and semantic information for each word meaning in its dimensions. Embedding disambiguated senses can increase the accuracy of syntactic dependency analysis in a variety of languages. There are several works that address this issue, however, sense2vec [6] outperforms the other proposals. In this work we used a pre-trained model generated with the sense2vec tool with 22 million words represented in 128 features vectors trained on the 2015 Reddit corpus for the recognition of disability annotations in English and Spanish.

2.4 Classifier

Character Embedding LSTM Although the words embedding allow to capture lexical information, morphological and orthographic characteristics are not considered immersed in the word itself. According to [11] it is identified that the use of character embedding improves learning for specific domains and is useful for morphologically rich languages. In the model, n-grams of the tokens generated in the preprocessing process converted to their vectors of corresponding embedded characters are generated, which are the input for the Bi-LSTM network, generating a feature vector that contains representations at the character level. Finally, a vector of 25 dimensions is generated representing character characteristics of the tokens.

Embedding Bi-LSTM Tokens The output of the Character Embedding LSTM process is concatenated with the feature vector of word embedding and sense-disambiguation embedding and used as input for a new Bi-LSTM network. At this point a sequence of scores is calculated which represent the probabilities of the labels for each word in the sentence. The labels used follow the encoding format BMEWO-V or BIOES-V used in the preprocessing process.

Conditional Random Fields (CRF) To improve the accuracy of predictions, we use a CRF model trained from the outputs of the previous process, obtaining as a result the most probable sequence of predicted labels.

2.5 Post-processing

After the generation of the labels for the tokens within the sentence the mentions of entities are recognized with the proposed model. The labels obtained following the BMEWO-V encoding format are transformed to the BRAT format following the encoding guides of this format. V tags that identify nested or overlapping entities are generated as new annotations within the scope of these tags

3 Evaluation

3.1 Datasets

The evaluation of the proposed model was carried out using the corpus of disability annotations in English and Spanish proposed in the task of IberEval2018 (http://nlp.uned.es/diann/). In both the DIANN training corpus, an analysis was developed to transform all the documents in text format with XML tags, later the whole training set was collected as the training data and the whole test set as the test data. The training set is made up of 400 documents, 3500 disability annotations. The test set consists of 100 documents without annotations as can be seen in Table 2. The corpus is made up of three types of entities: disabilities, negations and scope, so that the labels for words within a sentence can be labeled as seen in Table 1 following the BMEWO-V encoding format. In our experiment, the processing performed can be seen in the preprocessing process in the proposal section.

Table 1. Tokens tag in sentence.

Entity	Tags
Disability	B/I/V/E/S-Disability
Negation	B/I/V/E/S-Negation
Scope	B/I/V/E/S-Scope
Others	O

Table 2. Tokens tag in sentence.

Dataset	Files	Disabilities	Negation	Scope
Train	40	3500	80	80
Train-Train Data	300	2789	70	70
Train-Validation Data	100	711	10	10
Test	100	0	0	0

3.2 Pre-trained Models: Words Embedding and Senses Embedding

GloVe.6B, SpanishMillionWords and Reddit Vectors are pre-trained models of words and sense-disambiguation used to initialize the training set for training the model. Glove.6B is a pre-trained model of word representation vectors presented by [5]. The corpus used for the generation of this model are Wikipedia articles (until January 2014) and the Fifth Edition of Gigaword in English which is a complete text news data file that has been acquired for several years by the LDC in the University of Pennsylvania (which contains news from previous versions and adds news from 2009 to 2010). Some details of the pre-trained model:

- Corpus size: 6 billions words
- Vocab size: 400k
- Array size: 100
- Algorithm: GloVe

Spanish Billion Words is a pre-trained model of embedded words presented by [12]. The corpus of this model is made up of texts and web resources (SenSem, Ancora Corpus, Tibidabo Treebank and IULA Spanish LSP Treebank, OPUS Project, Europarl and Wikipedia). The details of the pre-trained model are the following:

- Corpus size: approximately 1.5 billions words
- Vocab size: 1000653
- Array size: 300
- Algorithm: Skip-gram Bag of Words

Reddit Vector is a pre-trained model of sense-disambiguation representation vectors presented by [6]. The corpus used for the generation of this model are comments published on Reddit (corresponding to the year 2015). The pre-trained Reddit vectors support partial or full Part-Of-Speech tags and entity tags.

3.3 Evaluation Metrics

In our experiment, we use precision, recall and F1 score to evaluate the performance of our system, and we use the evaluation criteria provided by the IberEval 2018 task, the Partial matching (a tagged disability name is correct only if there is some overlap between it and a gold disability name) and Exact matching (a tagged disability name is correct only if its boundary exactly matches with a gold disability name). A detailed description of evaluation is in the web (http://nlp.uned.es/diann/#evaluation). Finally, we use evaluation script (https://github.com/diannibereval2018/evaluation) provided by the shared task organizers to evaluate our system.

3.4 Result

The purpose of the first experiment was to compare tagging accuracy of two implementations: Bi-LSTM + CRF or Baseline model and Embedding-Bi-LSTM-CRF-Extended or the proposed model. To do this, we evaluated these implementations on the DIANNs dataset.The parameters of the pre-trained models and the hyper parameters of the models are the following:

- Words Embedding Dimension: 100 for English and 300 for Spanish
- Sense-Disambiguation Embedding Dimension: 128
- Characters Embedding Dimension: 25
- Hidden Layers Dimension: 100 (for each LSTM: for the forward and backward layers)
- Learning method: SGD, learning ratio: 0.005

– Dropout: 0.5
– Epochs: 100

The results are shown in Table 3. As can be seen in the table, the extension of the tag encoding format improves the result of the predictions. On the other hand, the use of pre-trained models of words and embedded senses reduces training time and increases the accuracy of labeling. In the first experiment we train the model over the NeuroNER baseline using the after mentioned configuration, datasets and pre-trained models. Due the absence of model to capture entities between the scope we decide to extend the tag encoding format to BIOES-V in order to capture this type of entities. Also, sense-disambiguation embeddings was apply in order to deal with different meanings for the same word in sentence context.

We compare the proposed Hybrid Bi-LSTM + CRF model with other published results for the DIANN task. Results are presented in the Table 4. The Hybrid Bi-LSTM + CRF model significantly outperform other approaches on DIANNs dataset. However, the result on exact negated disabilities can't be identify by the proposed model.

Table 3. English Non-negated Disability + Negated Disability.

Model	Exact		
	Precision	Recall	F-Score
Baseline	0.706	0.572	0.632
Embedding-Bi-LSTM-CRF-Extended	0.749	0.626	0.682

Table 4. Comparison in result of our experiment with others work in IberEval 2018.

Model	Exact			Partial		
	P	R	F	P	R	F
Our Model	0.749	2789	0.682	0.803	0.671	0.731
LSI-018-1	0.657	0.568	0.609	0.843	0.728	0.781
UPC-018-3	0.772	0.584	0.665	0.87	0.658	0.749
IxaMed	0.746	0.811	0.777	0.841	0.914	0.876
UPC-018-2	0.724	0.519	0.604	0.822	0.588	0.686

4 Conclusions

Named Entity Recognition (NER) is a crucial tool in information extraction tasks. In this work, we discuss the recent advance in NER that are able to

achieve satisfactory performance without requiring specifically domain knowledge or hand-crafted features. It is also important to highlight the specific domain independence as well as the language independence that are key to multi-language tasks. Inspired in these new works, we propose a hybrid Bi-LSTM and CRF model adding sense-disambiguation embedding and an extended tag encoding format to detect discontinuous entities, as well as overlapping or nested entities. Our results demonstrated that the extended tag encoding format improves the result of the predictions and the use of pre-trained models of words and embedded senses reduce training time and increase the accuracy of labeling and achieve state-of-the-art for the Disabilities NER.

We propose as future work the inclusion of the three models for the training of embedded words, the embedded word prediction model for syntactic and semantic information, the global aggregate model of word-word co-occurrence statistics and the model morphological representation of fastText words. Also extended the tag encoding format to capture different types of overlapping or nested entities.

5 Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain (project DeepEMR: Clinical information extraction using deep learning and big data techniques-TIN2017-87548-C2-1-R)

References

1. H. Fabregat, J. Martinez-Romo, and L. Araujo, Overview of the DIANN Task: Disability Annotation Task at IberEval 2018, in Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), 2018.
2. G. Lample, M. Ballesteros, , S. Subramanian, K. Kawakami, and C. Dyer, Neural Architectures for Named Entity Recognition.
3. S. Hochreiter and J. Schmidhuber, Long short-term memory., Neural Comput., vol. 9, no. 8, pp. 173580, Nov. 1997.
4. F. Dernoncourt, J. Y. Lee, and P. Szolovits, NeuroNER: an easy-to-use program for named-entity recognition based on neural networks, 2017.
5. J. Pennington, R. Socher, and C. D. Manning, GloVe: Global Vectors for Word Representation.
6. A. Trask, P. Michalak, and J. Liu, sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings, Nov. 2015.
7. Explosion AI, spaCy - Industrial-strength Natural Language Processing in Python. [Online]. Available: https://spacy.io/.
8. Standoff format - brat rapid annotation tool. [Online]. Available: http://brat.nlplab.org/standoff.html.
9. Q. Le and T. Mikolov, Distributed Representations of Sentences and Documents.
10. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, Enriching Word Vectors with Subword Information, 2016.

11. W. Ling et al., Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation.
12. C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2016. [Online]. Available: http://crscardellino.me/SBWCE/.