TASS 2018: Workshop on Semantic Analysis at SEPLN, septiembre 2018, págs. 65-70

# A Hybrid Bi-LSTM-CRF model for Knowledge Recognition from eHealth documents

## *Un Modelo Híbrido Bi-LSTM-CRF para el Reconocimiento de Conocimiento a partir de documentos electrónicos de eSalud*

**Renzo M. Rivera Zavala**[1], **Paloma Martínez**[1], **Isabel Segura-Bedmar**[1]
[1]Computer Science Department, University Carlos III of Madrid
100371920@alumnos.uc3m.es, pmf@inf.uc3m.es, isegura@inf.uc3m.es

**Abstract:** In this work, we describe a Deep Learning architecture for Named Entity Recognition (NER) in biomedical texts. The architecture has two bidirectional Long Short-Term Memory (LSTM) layers and a last layer based on Conditional Random Field (CRF). Our system obtained the first place in the subtask A (identification) of TASS-2018-Task 3 eHealth Knowledge Discovery, with an F1 of 87.2%.
**Keywords:** NER, Bi-LSTM, CRF, Information Extraction

**Resumen:** En este trabajo, describimos una arquitectura Deep Learning para el reconocimiento de entidades nombradas (NER) en textos biomédicos. La arquitectura se compone de dos capas bidireccionales LSTM (Long Short-Term Memory) y una última capa basada en Conditional Random Field (CRF). Nuestro sistema obtuvó el primer puesto en las subtareas A (identificación) y B (clasificación) de la competición TASS-2018-Task 3 eHealth Knowledge Discovery, con una F1 de 87.2%.
**Palabras clave:** NER, Bi-LSTM, CRF, Extracción de Información



### *1 Introduction*

Currently, the number of biomedical literature is growing at an exponential rate. The substantial number of research works makes it extremely difficult for researchers to keep up with the new development in their research areas. Therefore, the effective management of a large amount of information and the accuracy of knowledge is a vital task. Named Entities Recognition (NER) is one of the fundamental tasks of biomedical text mining, with the aim of identifying pieces of text that refer to specific entities of interest.

There are different scopes to address the NER problem. Among them, we can find methods based on dictionaries, which are limited by the size of the dictionary, spelling errors, the use of synonyms and the constant growth of vocabulary. Rule-based methods and Machine Learning methods usually require both syntactic and semantic features as well as characteristics of the language of the specific domain. One of the most effective method is Conditional Random Fields (CRF) (Lafferty, McCallum, and Pereira, 2001). Recently, Deep learning-based methods have also demonstrated state-of-the-art performance by automatically learning of relevant patterns from corpora, which allows the independence of a specific language or domain. However, until now, Deep Learning methods have not been able to provide better results than those obtained by classical traditional machine learning methods (Limsopatham and Collier, 2016).

In this paper, we propose a hybrid model combining two bidirectional Long Short Memory (Bi-LSTM) layers with a CRF layer. To do this, we adapt the NeuroNER model proposed in (Dernoncourt, Lee, and Szolovits, 2017) for the subtask A (identification) of TASS-2018-Task 3 eHealth Knowledge Discovery (Martínez-Cámara et al., 2018). Specifically, we have extended NeuroNER by adding context information, Part-of-Speech (PoS) tags and information about overlapping or nested entities. Moreover, in this work, we use two pre-trained word embedding models: i) a word2vec model (Spanish Billion Word Embeddings (Cardellino, 2016)), which was trained on the 2014 dump

ISSN 1613-0073

of Wikipedia and ii) a sense-disambiguation embedding model (Trask, Michalak, and Liu, 2015).

The rest of the paper is organized as follows. In Section 2, we describe the architecture of our system. Section 3 presents the results. In Section 4, we provide the conclusions.

## *2 System Description*

### 2.1 Pre-processing

All texts were preprocessed in four steps. First, sentences were split by using Spacy (Space.io, 2018), an open source library for advanced natural language processing with support for 26 languages. Second, sentences and their annotated entities were transformed to the BRAT format[1], a standoff format similar to BioNLP Shared Task standoff format. Then, sentences were tokenized. Finally, each token in a sentence was annotated using the BMEWO-V extended tag encoding, to capture information about the sequence of tokens in a given sentence. The BIOES label scheme introduced in the work of (Borthwick et al., 1998) arises in order to overcome the limitation of the BIO scheme for the representation of discontinuous entities. BIOES coding distinguishes the end of an entity through the E (End) tag and adds the S (Single) tag to denote entities composed of a single token. The BIOES-V or BMEWO-V encoding distinguishes the B tag to indicate the start of an entity, the M tag to indicate the continuity of an entity, the E tag to indicate the end of an entity, the W tag for indicate a single entity, and the O tag to represent other tokens that do not belong to any entity. The V tag allows to represent overlapping entities. This encoding scheme allows the representation of discontinuous entities and overlapping or nested entities.

### 2.2 Learning Transfer

In our work, we propose as input of our network two different embeddings: word embeddings and sense-disambiguation embeddings. Below we describe them in more detail.

#### 2.2.1 Words Embeddings

Word embedding is an approach to represent words as vectors of real numbers. There are different methods to obtain these vectors such as probabilistic models and neural networks. In last years, neural networks for training word embedding models have gained a lot of popularity among NLP community because they are able to capture syntactic and semantic information among words. The most popular methods are word2vec (Le and Mikolov, 2014), the global aggregate model of word-word co-occurrence statistics (Pennington, Socher, and Manning, 2014) and the morphological representation of fastText (Bojanowski et al., 2017).

In this work, we used the Spanish Billion Words(Cardellino, 2016), which is a pre-trained model of word embeddings trained on different text corpora written in Spanish (such Ancora Corpus (Taulé, Martí, and Recasens, 2008) 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

#### 2.2.2 Sense-Disambiguation Embedding

We also used the sense2vec (Trask, Michalak, and Liu, 2015) model, which provides multiple embeddings for each word based on the sense of the word. This is able to analyze the context of a word and then assign its more adequate vector. 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.

Reddit Vector is a pre-trained model of sense-disambiguation representation vectors presented by (Trask, Michalak, and Liu, 2015). This model was trained on a collection of comments published on Reddit (corresponding to the year 2015). The pre-trained Reddit vectors support the following "senses", whether partial or full PoS tags or entity tags. The details of the pre-trained model are the following:

- Corpus size: approximately 2 billions words
- Vocab size: 1 million
- Array size: 128
- Algorithm: Sense2Vec

[1] http://brat.nlplab.org/standoff.html

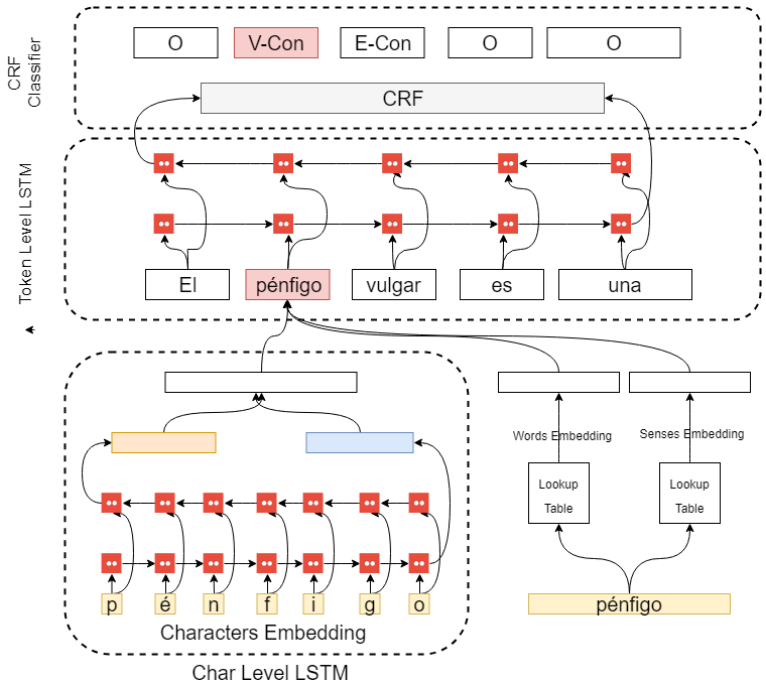


Figure 1: Overview architecture of our hybrid LSTM-CRF model.

## 2.3 The network

### 2.3.1 Character Embedding Bi-LSTM layer

Although the word embeddings are able to capture syntactic and semantic information, other linguistic information such as morphological information, orthographic transcription or PoS tags are not exploited. According to (Ling et al., 2015), the use of character embeddings improves learning for specific domains and is useful for morphologically rich languages. For this reason, we decided to consider the character embedding representation in our system. We used a vector of 25 dimensions to represent each character. The character alphabet includes all 121 unique characters in the TASS-2018-Task 3 eHealth Knowledge Discovery training, development and test datasets and the token PADDING. In this way, tokens in sentences are represented by their corresponding character embeddings, which are the input for the first Bi-LSTM network.

### 2.3.2 Word and Sense embedding Bi-LSTM layer

The output of the first layer is concatenated with the word embeddings and with the sense-disambiguation embeddings of the tokens in a given input sentence. This concatenation of features is the input for the second Bi-LSTM layer. The goal of this layer is to obtain a sequence of probabilities corresponding to each label of the BMEWO-V encoding format. In this way, for each input token, this layer returns six probabilities (one for each tag in BMEWO-V). The final tag should be that with highest probability.

The parameters of the sets and the hyper parameters of the models are the following:

- Words Embedding Dimension: 300
- 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

### 2.3.3 Conditional Random Fields (CRF) layer

To improve the accuracy of predictions, we also used a CRF model trained, which takes as input the output of the previous layer and obtains the most probable sequence of predicted labels.

### 2.4 Post-processing

Once tokens have been annotated with their corresponding labels in the BMEWO-V encoding format, the entity mentions must be transformed to the BRAT format. V tags, which identify nested or overlapping entities, are generated as new annotations within the scope of other mentions.

## *3 Evaluation*

### 3.1 Datasets

The evaluation of the proposed model was carried out using the annotated corpus proposed in the TASS-2018-Task 3 eHealth Knowledge Discovery (https://github.com/tass18-task3/data).

The training set is made up of 5 documents with 3276 entities annotations. The development set consists of 1 text document with 1958 entities annotations. The test set consists of 1 text document (see Table 2). There are two types of of entities: concepts and actions. For this reason, tokens can be annotated with different labels (see Table 1) following the BMEWO-V encoding format.

| **Entity** | **Tags** |
|---|---|
| Concept | B/M/E/W/V-Concept |
| Action | B/M/E/W/V-Action |
| Others | O |

Table 1: Tokens Tag in Sentence

| **Datasets** | **Files** | **Concept** | **Action** |
|---|---|---|---|
| Train | 5 | 2427 | 849 |
| Development | 1 | 1525 | 434 |
| Test | 1 | 0 | 0 |

Table 2: Dataset Statistics

In our experiments, we used precision, recall and F1 score to evaluate the performance of our system. The TASS-2018-Task 3 considers two different criteria: the partial matching (a tagged entity name is correct only if there is some overlap between it and a gold entity name) and exact matching (a tagged entity name is correct only if its boundary exactly match with a gold entity name). A detailed description of evaluation is in the web (`http://www.sepln.org/workshops/tass/2018/task-3/evaluation.html`). Moreover, we used evaluation script (`https://github.com/TASS18-Task3/data/blob/master/score_training.py`) provided by the shared task organizers to evaluate our system.

### 3.2 Results

As it was described above, our system is based on network with two Bi-LSTM layers and a last layer for CRF. In the first Bi-LSTM layer, we consider the character embeddings. In the second layer, we concatenate the output of the first layer with word embeddings and sense-disambiguate embeddings. Finally, the last layer uses a CRF to obtain the most suitable labels for each token.

Table 3 compares the results obtained using the NeuroNER system with our extended version using pre-trained embeddings models and the BMEWO-V encoding format. Our extended version of NeuroNER achieves a significant improvement of the results (more than 7.2% in F1).

| **System** | **P** | **R** | **F1** |
|---|---|---|---|
| NeuroNER | 0.824 | 0.785 | 0.804 |
| **ext. NeuroNER** | **0.862** | **0.882** | **0.872** |

Table 3: Comparison of NeuroNER and our extended version.

In the substask A (identification of key phrases), our system obtained the top micro F1 (0.872) (see Table4). It significantly outperform the rest of participating systems. We will wait to review the proposed systems in greater depth in order to establish comparisons and possible improvements to our implementation.

| **System** | **P** | **R** | **F** |
|---|---|---|---|
| **Extended NeuroNER** | **0.862** | **0.882** | **0.872** |
| plubeda | 0.77 | 0.81 | 0.79 |
| upf-upc | 0.86 | 0.75 | 0.80 |
| VSP | 0.31 | 0.32 | 0.32 |
| Marcelo | 0.11 | 0.32 | 0.17 |

Table 4: Results of the participating systems in the subtask A.

## 4 Conclusions

Named Entity Recognition (NER) is a crucial tool in text mining tasks. In this work, 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 system is able to achieve satisfactory performance without requiring specifically domain knowledge or hand-crafted features. It is also important to highlight the language independence, which is key to multi-language tasks. Our results demonstrated that the extended BMEWO-V encoding improves the result of the predictions. Moreover, the pre-trained models help to reduce training time and increase the accuracy of labeling, achieving the highest F1 for the subtask A.

We plan to try with other embeddings models such as the FastText model, which contains morphological information. Moreover, we will extend the encoding format to capture distinct types of overlapping or nested entities.

## Acknowledgement

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

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Borthwick, A., J. Sterling, E. Agichtein, and R. Grishman. 1998. Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition. Technical report.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings.

Dernoncourt, F., J. Y. Lee, and P. Szolovits. 2017. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102. Association for Computational Linguistics.

Lafferty, J., A. McCallum, and F. C. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, pages 282–289.

Le, Q. and T. Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Limsopatham, N. and N. Collier. 2016. Learning orthographic features in bi-directional lstm for biomedical named entity recognition. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, pages 10–19.

Ling, W., T. Luís, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black, and I. Trancoso. 2015. 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*, page 1520–1530.

Martínez-Cámara, E., Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, and J. Villena-Román. 2018. Overview of TASS 2018: Opinions, health and emotions. In E. Martínez-Cámara, Y. Almeida Cruz, M. C. Díaz-Galiano, S. Estévez Velarde, M. A. García-Cumbreras, M. García-Vega, Y. Gutiérrez Vázquez, A. Montejo Ráez, A. Montoyo Guijarro, R. Muñoz Guillena, A. Piad Morffis, and J. Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September. CEUR-WS.

Pennington, J., R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Space.io. 2018. spaCy · Industrial-strength Natural Language Processing in Python.

Taulé, M., M. A. Martí, and M. Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *LREC 2008*, pages 96–101.

Trask, A., P. Michalak, and J. Liu. 2015. sense2vec - a fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, abs/1511.06388.