

Exploring Deep Learning for Named Entity Recognition of Tumor Morphology Mentions

Gema de Vargas Romero, Isabel Segura-Bedmar

Computer Science Department, Universidad Carlos III de Madrid (UC3M), Leganés, 28911, Madrid, Spain

Abstract

This paper describes the development of a Named Entity Recognition (NER) system to automatically detect tumor morphology mentions in medical documents, known as ICD-O codes, International Classification of Diseases for Oncology. This study is developed as part of the Cantemist program of the Plan of Advancement of Language Technologies (Plan TL). This work tries to contribute to the existing NER technologies that focus on Spanish health-related documents. This is a necessary task given the amount of research, regarding the health sector, that is written in Spanish and the benefits it could bring to the medical environment. In fact, since most NER techniques are developed for English, this research cannot be completely exploited. In this research, we explore different machine learning techniques such as CRF, a Bidirectional Long short-term memory (Bi-LSTM) and a Bidirectional Encoder Representations from Transformers (BERT) to address the task of detecting tumor morphology mentions from clinical texts written in Spanish.

Keywords

Named Entity Recognition, BiLSTM, BERT

1. Introduction

Natural Language Processing (NLP) has become vital since the amount of information to which people have access nowadays cannot be easily managed. To solve this, NLP offers tools that vary based on the intention of the user, such as translating, summarizing or extracting information from a text. [1] Named Entity Recognition (NER) is a specific Natural Language Processing (NLP) task that focuses on information extraction.

Focusing on NER, there are currently many technologies that achieve state-of-the-art performance. However, there are two key aspects that force to keep developing in this field. On the one hand, the performance of NER systems is dependent on the domain. As a result, this makes it necessary to construct domain specific NER systems. In fact, this task becomes crucial in the medical domain given the amount of research in this field and the advantages it could bring to patient diagnosis, prognosis and further research. [2]

On the other hand, despite the generalization that some NER systems can achieve, the performance of the systems is also dependent on the language it was built for. Therefore, this makes it necessary to build systems specific for the different languages. For instance, focusing on

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: 100426253@alumnos.uc3m.es> (G.d.V. Romero); isegura@inf.uc3m.es> (I. Segura-Bedmar)

URL: <http://hulat.inf.uc3m.es/nosotros/miembros/isegura> (I. Segura-Bedmar)

ORCID: 0000-0002-0877-7063 (G.d.V. Romero); 0000-0002-7810-2360 (I. Segura-Bedmar)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Spanish, there is a huge amount of biomedical research written in this language. However, since most NER techniques are developed for English, this research cannot be completely exploited.

Disease recognition and normalization in medical texts is a challenging task given the variability and complexity of the disease mentions. Currently, the main technique followed in this context is the clinical coding, which consists in the collection of codes regarding the taxonomy of a disease, signs, symptoms and medical procedures. These are standard codes provided by the ICD10, International Classification of Diseases (CIE10 in Spanish). In particular, the ICD-O codes (CIE-O-3.1 in Spanish), International Classification of Diseases for Oncology, allows to code tumor morphology mentions in health-related documents written in Spanish. [2]

Under this context, this study proposes a NER system to automatically detect tumor morphology mentions in a medical document. It is developed as part of the Cantemist track [2], which is sponsored by Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) and is part of the IberLEF 2020 evaluation campaign. The Cantemist track is organized into 3 subtasks:

1. NER which consists on finding tumor morphology mentions;
2. NORM which involves finding tumor morphology mentions and assigning an ICD-O code;
3. CODING which consists on assigning an ICD-O code to documents.

We focus on the NER task. This way, it will aim at constructing a system to identify tumor morphology mentions in the corpus provided by the Cantemist organizers and contribute to the existing NER technologies and resources that focus on Spanish health-related documents. Our study explores three different machine learning approaches: 1) CRF classifier as baseline, 2) Bidirectional Long short-term memory (Bi-LSTM) and 3) Bidirectional Encoder Representations from Transformers (BERT) to address the task of detecting tumor morphology mentions from clinical texts written in Spanish.

This study will be focusing on domain specific NEs, more specifically, on a unique entity type: “Morfología Neoplasia”.

2. Related Work

NER aims to identify and categorize words or expressions inside a text that represent entities. NER systems must confront many challenges such as the construction of a generic entity tagger given the variability of the entity set among different domains, the diversity of languages and the ambiguity within each of them [3], the dependency on the quality of the annotations and the existence of infrequent entities.

Two key aspects in NER systems are to employ context information in the identification of entities and assume inter-dependency between the words in the text. [3]

Since entities can be words or phrases, they are usually labeled following the IOB format, where every token that conforms an entity is assigned a tag ‘I’, ‘O’ or ‘B’ based on its location, followed by the entity type. In fact, the IOB format stands for “Inside, Outside, Beginning”. Therefore, if a token is at the beginning of an entity, it will be labeled as “B”, as “I” if the token is inside an entity and as “O” if the token is not part of an entity. [4]

Along the years, several approaches have been employed to address the NER task. The main approaches include:

1. Rule-based;
2. Unsupervised learning;
3. Feature based supervised learning;
4. Deep learning.

Rule-based It is based on hand-crafted rules, that can be either semantic or syntactic. This approach does not require annotated data and is highly dependent on the domain and can rely on dictionaries. It usually provides high precision but with a low recall.[1]

Unsupervised learning It involves training unsupervised methods without using labeled data. The main unsupervised learning technique employed in NER is clustering. For this, the system clusters the data into groups based on “context similarity” and assigns them a named entity by means of “entity extraction”. [1]

Feature based supervised learning This approach involves prior feature engineering to construct features (usually vectors) that represent each instance (word or sentence). Features can be word level features or document and corpus features among others. This approach needs a dataset of annotations, which are used for train a model. In this scenario, the NER task is considered as a multi-class classification problem, where the variety of entity types conform the set of classes.

Common supervised methods employed in NER are Hidden Markov Models, Decision Trees, Maximum Entropy Models, Support Vector Machines and Conditional Random Fields. [1]. This last method, CRF, considers context information and achieves to outperform the previous ones. [1] Therefore, we propose CRF as our baseline system.

Deep learning The use of deep learning in NER became more common in recent years. As an advantage to previous techniques, the extraction of features is learned automatically by the model, without complex feature engineering. This is done by stacking various layers in the neural network, where different abstractions of the data are obtained. Another advantage of deep learning is the use of non-linear activation functions that allow to learn more complex features. [1]

As previously mentioned, NER is a challenging task for various reasons such as the variability of the entity sets among different domains and the dependency on the quality of the annotations. However, when it comes to the biomedical domain, the NER task becomes even more challenging given the domain specific terminology, the use of acronyms or abbreviations and non-standard terms. Furthermore, Bio-NER systems are affected by the lack of comprehensive biomedical entities dictionaries. [5]

Many studies have shown that the combination of a Bidirectional Long short-term memory (BiLSTM) followed by a CRF achieves state-of-the-art performance in NER [6]. Zhai et al. [6] made a comparison of various systems combining deep learning models such as Bidirectional Long short-term memory (Bi-LSTM) and Convolutional Neural Network (CNN) to recognize chemical and diseases mentions. To initialize the deep learning networks, the author used a Word2Vec word embedding model trained over MEDLINE abstracts. The authors employed the

BioCreative V CDR corpus [7], which is a manually annotated dataset. 1,965 disease entities, 1,467 chemical entities and 1,038 chemical disease relations (CDRs) were employed in the training dataset. Moreover, 1,865 disease entities, 1,507 chemical entities and 1,012 CDRs were employed in the development dataset. Finally, 1,988 disease entities, 1,435 chemical entities and 1,066 CDRs were employed in the test set. [7] Focusing on the tasks of disease recognition, the best performance (F1=83.01%) was provided by a hybrid architecture combining a Bi-LSTM with a CRF classifier a CNN initialized with character embedding. [6]

Wei et al. [8] proposed various systems for disease named entity recognition combining Conditional Random Fields (CRF) and Bidirectional Recursive Neural Network (Bi-RNN). For this purpose, the authors also employed the BioCreative V CDR corpus [7]. They found that the best performance (F1=82.88%) was obtained when combining the results of a Bi-RNN with the CRF based model through a Support Vector Machine (SVM). On the one hand, the Bi-RNN was trained using pre-trained vectors on a PubMed corpus. On the other hand, the CRF-based model employed input features such as PoS tag, word shape, prefix and suffix among others [7].

Moreover, Choo and Lee [9] proposed a system, CLSTM (Contextual LSTM with CRF) that focuses on "capturing local context information based on n-gram characters" and employs word embeddings. The system was evaluated over three biomedical corpora: the National Center for Biotechnology Information (NCBI) [10], the BioCreative II Gene Mention (GM) corpus [11], and the BioCreative V [7] corpus. Their analysis showed that, focusing on a strict entity matching, the CLSTM system trained using word and character embeddings achieved the best performance (F1=86.44%) over the BioCreative V corpus. Regarding the results over the BioCreative V corpus, CLSTM with word level embedding and CLSTM with Character level embedding achieved F1 scores of 86.36 and 85.92 respectively, followed by GRAM-CNN (F1=85.79%), BERT [12] (F1=85.72%), BiLSTM-CRF (F1=85.50%) and BiLSTM (F1=81.88%) [9].

3. Methods

3.1. The Cantemist corpus

The Cantemist corpus [2] is formed by 3,000 clinical cases stored in different files. These are plain text with UTF-8 encoding. Each text file is associated to an annotation file (see Figure 1) in BRAT format which has been manually annotated by clinical experts. These annotations are tumor mentions, more specifically, ICD-O codes. The total corpus has been divided into 4 sets; train set, two development sets and test set. A detailed description of this dataset can be found in [2].

In the BRAT format, each line represents an entity mention. Each entity annotation includes an ID ('T1', 'T2'...) where 'T' stands for "text bound annotation", followed by the entity type ('MORFOLOGIA NEOPLASIA'), start-end offsets and the text of the annotation. The start-end offset indicates the position within the file of the first and last character in the entity. [4]

Entities can be formed by several tokens (words). In fact, the average number of words that conform each entity is 2. However, the BRAT format in which the annotations are given show continuous text-bound annotations, where only the start offset of the first word and the end offset of the last word that form the entity are given (see Figure 2). That is why, a pre-processing

```
[['T1', 'MORFOLOGIA_NEOPLASIA 2200 2214', 'adenocarcinoma\n'],
 ['T6', 'MORFOLOGIA_NEOPLASIA 1347 1357', 'metástasis\n'],
 ['T7', 'MORFOLOGIA_NEOPLASIA 1229 1234', 'tumor\n'],
 ['T9', 'MORFOLOGIA_NEOPLASIA 1112 1141', 'adenocarcinoma indiferenciado\n'],
 ['T10', 'MORFOLOGIA_NEOPLASIA 1504 1531', 'adecarcinoma indiferenciado\n'],
```

Figure 1: Example of an ann file.

```
[ 'T10', 'MORFOLOGIA_NEOPLASIA 1504 1531', 'adecarcinoma indiferenciado\n']
```

Figure 2: Example of a continuous text-bound annotation in BRAT format.

```
[ 'T10', 'MORFOLOGIA_NEOPLASIA 1504 1515;1517 1531', 'adecarcinoma indiferenciado\n']
```

Figure 3: Example of a discontinuous text-bound annotation in BRAT format.

stage that involves converting such annotations into discontinuous text-bound annotations has been applied (see Figure 3). [4]

The IOB format is mostly used to tag each token. However, after analyzing the Cantemist corpus, the presence of nested entities, which are entities embedded in another entity, was noticed (see Figure 4). On the contrary, those entities that do not appear in various annotations are known as flat entities. To solve this problem, we have used an extension of the IOB format, the BIOES-V format. [6, 13] This way, if the entity is a single token entity it will be labeled as "S" and "V" if the token is part of a nested entity. To preprocess the texts and represent their tokens with the BIOES-V format, we have used Spacy, a popular library for NLP. Spacy allows us to perform several tasks such as sentence splitting, tokenization and PoS tagging.

Regarding the training dataset, over which the following methods will be trained, the distribution of tokens belonging to each BIOES-V tag is presented in Table 1. Table 1 also includes the distribution of tokens of the test set.

```

['T3', 'MORFOLOGIA_NEOPLASIA 2823 2847', carcinoma neuroendocrino'n'],
['T8', 'MORFOLOGIA_NEOPLASIA 2738 2749', tumor mixto'n'],
['T13',
'MORFOLOGIA NEOPLASIA 2751 2820',
adenocarcinoma pobremente diferenciado con células en anillo de sello'n'],
['T15',
'MORFOLOGIA NEOPLASIA 2738 2847',
tumor mixto: adenocarcinoma pobremente diferenciado con células en anillo de sello
y carcinoma neuroendocrino'n']]

```

Figure 4: Example of nested entities.

Table 1

BIOES-V statistic in the Cantemist corpus

Label	Training	Test
B	3,071	1,709
I	5,193	2,791
O	439,863	239,701
E	3,070	1,707
S	3,313	1,908
V	20	22
Total	454,530	247,838

Figure 5 shows the main preprocessing tasks involved in our study.

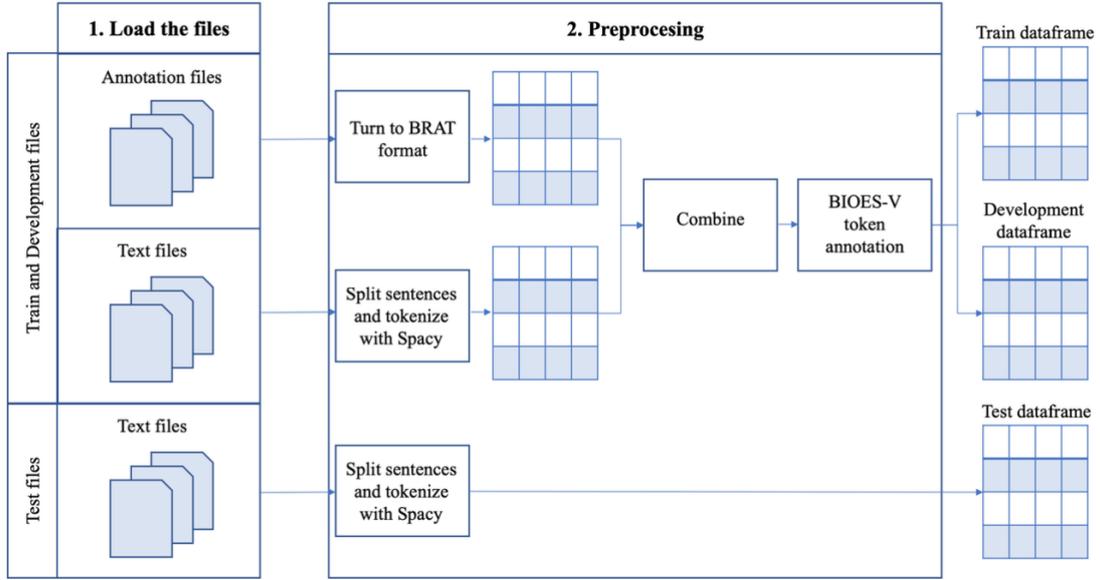


Figure 5: Preprocessing steps involved in the NER task.

3.2. Machine Learning models

As baseline, we propose the use of a CRF classifier, a type of probabilistic discriminative model which aims to predict an output variable while giving huge relevance to the sequence of predictions. The main difference with generative models is that these rely on strict dependency assumptions, whilst this discriminative model relaxed such assumptions. This way, it can employ features where it exists a dependency. Until the implementation of deep learning methods, CRF provided the best performance in NER tasks among machine learning methods.

CRF models the conditional distribution $p(y|x)$, which is the probability of label sequences given observation sequences and follows the formula below. [3]

$$p_{\theta}(y|x) = \frac{1}{Z_{\theta}(x)} \exp\left(\sum_{t=1}^T \sum_{k=1}^K \theta_k f_k(y_{t-1}, y_t, x_t)\right), \quad (1)$$

Here, θ_k represents the parameters of the distribution and f_k is the function that defines the transition from state y_{t-1} to state y_t , being y_{t-1} and y_t the sequences' labels. Z_{θ} is the normalization factor defined as follows. [3]

$$Z_{\theta}(x) = \sum_y \exp\left(\sum_{t=1}^T \sum_{k=1}^K \theta_k f_k(y_{t-1}, y_t, x_t)\right), \quad (2)$$

The training of the model involves working with sentences as the input sequence. In addition, for every token in a sequence, the model needs as parameters, the position of the current token,

its PoS (Part of Speech) tag, the BIOES-V label of the current token and the label of the previous token. This way, context information is also being considered. [14]

We also implement a BiLSTM (Long Short-Term Memory) network followed by a CRF classifier. On the one hand, the LSTM is a type of RNN (Recurrent Neural Network). The main aspect of these RNNs is that they employ context information, more specifically, information from past observations, to make predictions over the current observation. [15] In this study, the observations are tokens from the text that are used as sequential inputs to the network. This ability to “remember” information is crucial in NER since this is a sequence labelling problem where there is interdependency between the tokens in the text. However, RNNs suffer from vanishing gradient problem caused by the increasing amount of information to be “remembered”. To solve this, LSTMs differ from common RNNs by “forgetting” information and “making space” for more important one. [15] In addition, the implementation of a bidirectional LSTM becomes more appropriate when working with NER since it allows not only to remember information from past observations but also to consider future information. This is achieved by employing two independent LSTM processes working in opposite directions. As a result, the final layer (a CRF classifier) in the network receives two vectors, the outputs of the two LSTMs layers, as an input. Then, the CRF considers the label predicted for the previous observation in order to predict the label of the current one. This behavior is necessary in an NER task given the interdependency between tokens. We have implemented three different approaches based on BiLSTM-CRF.

Approach 1: BiLSTM-CRF with random initialization of vectors It consists on creating a word vocabulary from the words present in the training dataset and map each word to a numeric vector of 40 components. It must be mentioned that the length of each sentence is being set to a fixed size, 75. Then, it assigns a random vector for each token in the vocabulary. Unfortunately, it does not capture relationships or similarities between tokens. In addition, it may occur that certain tokens in the test dataset do not appear in the training dataset. Therefore, regardless of the similarity of these tokens to others in the vocabulary, these tokens will be mapped to the value that identifies the “Unknown” label. [15]

Approach 2: BiLSTM-CRF with a pre-trained word embedding in Spanish Word embedding is a technique that consists on representing words in a vocabulary as vectors of real numbers while capturing semantic and syntactic information. [6] However, word embeddings are not able to capture other features such as PoS tags. Among word embedding methods it can be found Word2Vec[16], which is a neural network-based model, and can be implemented using two different algorithms; Continuous Bag of Words model (CBOW) and Skip-gram model (SG). Other word embedding methods are Global Vector (GloVe)[17] and fastText[18]. The main disadvantage of using word embeddings is that there may be words in the test set that are not represented in the initial vocabulary. These are out-of-vocabulary words (OOV). For this study, the pre-trained word embedding employed is the Scielo + Wikipedia health cased skip-gram model developed by Plan TL [19]. This model was trained using FasText implementation over a corpus created as a combination of the SciELO database, which contains articles in Spanish, English and Portuguese, and a health-related subset of Wikipedia that includes Pharmacology,

Pharmacy, Medicine and Biology related documents. A very common pre-trained word embedding model employed in NLP tasks in Spanish is the Spanish Billion Words Corpus and Embedding, SBWCE. [20] However, given the specificity of the domain subject of this study, the Scielo + Wikipedia health cased skip-gram model [19] seems more appropriate since it contains specific medical terms that may be found in the Cantemist corpus. Thus, this approach involves representing each token in the training dataset as a 300-element numeric vector using the vocabulary from the pre-trained Scielo + Wikipedia health cased word embedding model. Also, this approach also involves fixing the parameter that defines the maximum sentence length to 75.

Approach 3: BiLSTM-CRF with character embedding This technique implements the pre-trained word embedding explained in approach 2. As a novelty from the previous approach, it represents each character in the words as a vector of a fixed dimension. For this, it performs character embedding as a previous layer to represent words as characters. This approach is useful when working with out of vocabulary words (OOV), which are words that do not appear in the vocabulary that the model was trained with. In sum, this approach combines two feature representations of the text as inputs to the system. On the one hand, every token in the text is encoded into a 300-element vector using a vocabulary from a pre-trained word embedding. On the other hand, every character that conforms a token is encoded using a character vocabulary also learned from the pre-trained word embedding. This way, this approach considers two fixed parameters; the maximum sequence length, fixed to 75, and the maximum word length, fixed to 10.

Finally, we also explore the use of Bidirectional Encoder Representations from Transformers (BERT) [12]. BERT is a more recent deep learning approach for NLP that has outperformed prior language models [12]. The main difference between BERT and other strategies is that BERT overcomes the limitation of standard language models based on a unidirectional constraint and allows to consider context information from both directions without the need to employ two independent LSTMs. [12]

Furthermore, the process of BERT consists of two stages. The first stage is the model pre-training over unlabeled data and the second stage is approached either by performing feature based or fine-tuning tasks. [12] In our work, we employ a pre-trained BERT model and then perform fine-tuning over it to find the optimal combination of parameters for this specific NER task. On the bright side, focusing on fine-tuning an already trained model is less time consuming and has a good generalization since very few parameters have to be learned.

Since this task involves working with clinical cases in Spanish, a multilingual cased version of BERT is employed instead [21]. It has fixed model sizes: L=12, H=768, A=12, and Total Parameters=110M. Here, L denotes the number of layers (or transformer blocks), H is the hidden size and A is the number of self-attention heads. This way, the only parameters to focus on while performing fine-tuning are maximum length (the context or sequence length to consider), batch size, learning rate and number of epochs. [12]

In sum, we use an already pre-trained cased text BERT model. Therefore, no preprocessing steps has been performed over the corpus. Regarding the method's parameters, the maximum sequence length has been fixed to 75, as in previous approaches. When it comes to the training

of the model, a batch size of 32, a learning rate of $3e-5$ and 3 epochs have been employed. It must be mentioned that the number of epochs specify the number of times the model goes through all the data. Therefore, in order to avoid overfitting, this value must not be too large.

4. Results

The performance of the models has been evaluated using several metrics. The primary metrics employed are precision, recall and f1 score. What's more, it will also be employing micro-average scores instead of macro-average scores since the number of tokens belonging to each entity type is imbalanced. The main difference between micro-average and macro-average scores is that micro average scores are computed jointly for all the entity types whilst the macro average scores involve computing the metrics individually for each entity type and then combining them by performing the mean.

The following methods have been trained using the training dataset and fine-tuned employing the labeled development datasets. Then, once the optimal parameter combination had been achieved, the final model has been trained employing the training and development datasets.

First, we will present and discuss the performance of our approaches by considering their micro-average results calculated over the BIOES-V tags used to represent our tokens. Finally, we also present the general results provided by the Cantemist organizers.

First, we propose the CRF classifier, which has shown good performance in literature [3, 14]. This allows to establish a baseline. Focusing on the performance of the CRF classifier in identifying tokens that belong to an entity, this baseline method achieved a micro average F1 score 0.74 and 0.77 on the development datasets 1 and 2 respectively. Regarding the test set, this method has achieved a micro average F1 score of 0.78 as seen in Figure 6.

BiLSTM-CRF with random initialization achieves a micro average F1 score of 0.71 and 0.73 on the development datasets 1 and 2 respectively, when analyzing its performance in identifying tokens that belong to an entity. Regarding the test dataset, it has achieved a micro average F1 score of 0.78 as seen in Figure 7.

BiLSTM-CRF with a pre-trained word embedding in Spanish shows a micro average F1 score of 0.75 and 0.78 on the training and development datasets 1 and 2 respectively, when analyzing its performance in identifying tokens that belong to an entity. On the other hand, an f1 score of 0.81 has been obtained over the test dataset as seen in Figure 8.

BiLSTM-CRF with character embedding obtains a micro average F1 score of 0.75, 0.72 and 0.76 on the training and development datasets 1 and 2 respectively, when analyzing its performance in identifying tokens that belong to an entity. Regarding the test dataset, it has obtained an F1 score of 0.78 as seen in Figure 9.

Finally, BERT achieves a micro average F1 score of 0.78 and 0.80 on the development datasets 1 and 2 respectively, when analyzing its performance in identifying tokens that belong to an entity. Focusing on the test set, it achieved a micro average f1 score of 0.82.

In short, CRF established a baseline with a micro average f1 score of 0.78. Then, focusing on the deep learning methods, three strategies that implement a Bidirectional LSTM combined with CRF have been approached. The first strategy consisted on a random initialization of the vectors in the vocabulary and did not succeed in defeating the baseline method (see Figure

Table 2

Precision, Recall and F1 scores obtained over the test set.

Model	Precision	Recall	F1 Score
CRF	0.8	0.768	0.783
BiLSTM-CRF with random initialization	0.771	0.773	0.772
BiLSTM-CRF with pre-trained word embeddings	0.828	0.769	0.797
BiLSTM-CRF with word and character embeddings	0.784	0.759	0.771
BERT	0.756	0.775	0.765

11). The second BiLSTM-CRF strategy involved a pre-trained word embedding in Spanish for token representation. It achieved a better performance than the first BiLSTM-CRF approach and baseline approach, CRF, (see Figure 11). The third BiLSTM-CRF strategy consisted on incorporating a character embedding along with the token representation did not achieve a better performance than the second approach of BiLSTM-CRF nor the baseline approach (see Figure 11). Finally, the last deep learning approach to be assessed was BERT. This method achieved to defeat both the baseline method (CRF) and the previous Bi-LSTM approaches explored (see Figure 11). Therefore, we can conclude that BERT is the best option to recognize tumor morphology mentions in clinical texts.

However, the performance of the models must also be assessed based on the correctly identified entities in each clinical case. This way, this assessment does not consider if a token has been correctly identified as part of an entity, but if an entity, which can be formed by several tokens, has been identified in its complete form. The results of this evaluation have been captured in Table 2. It shows that the second approach of BiLSTM-CRF initialized with a pre-trained word embedding model has the best performance among the different methods explored.

5. Conclusion

This study has focused on developing a NER system to automatically identify tumor morphology entity mentions in health-related documents in Spanish. As previously mentioned, it has been developed under the Cantemist track, from which the corpus employed has been obtained.

For this purpose, we have explored different machine learning approaches such as CRF, Bi-LSTM and BERT. Although the approaches show very similar performance, we can conclude that Bi-LSTM with pre-trained word embeddings shows the top F1 (0.797). However, if we study the micro-average F1 calculated over the BIOES-V format to represent our tokens, we see that BERT provides better results (micro F1=0.82) than the other approaches.

As future work, we plan to extend our deep learning models by incorporating semantic information from biomedical dictionaries (such as entity embeddings). We will also explore other hybrid deep learning architecture.

6. Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R) and the Interdisciplinary Projects Program for Young Researchers at Universidad Carlos III of Madrid founded by the Community of Madrid (NLP4Rare-CM-UC3M).

References

- [1] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, *IEEE Transactions on Knowledge and Data Engineering* (2020).
- [2] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, CEUR Workshop Proceedings, 2020.
- [3] T. Nguyen, D. Nguyen, P. Rao, Adaptive name entity recognition under highly unbalanced data, *arXiv preprint arXiv:2003.10296* (2020).
- [4] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 102–107.
- [5] V. Cotik, H. Rodríguez, J. Vivaldi, Spanish named entity recognition in the biomedical domain, in: *Annual International Symposium on Information Management and Big Data*, Springer, 2018, pp. 233–248.
- [6] Z. Zhai, D. Q. Nguyen, K. Verspoor, Comparing cnn and lstm character-level embeddings in bilstm-crf models for chemical and disease named entity recognition, *arXiv preprint arXiv:1808.08450* (2018).
- [7] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wieggers, Z. Lu, Biocreative v cdr task corpus: a resource for chemical disease relation extraction, *Database* 2016 (2016).
- [8] Q. Wei, T. Chen, R. Xu, Y. He, L. Gui, Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks, *Database* 2016 (2016).
- [9] H. Cho, H. Lee, Biomedical named entity recognition using deep neural networks with contextual information, *BMC bioinformatics* 20 (2019) 735.
- [10] R. I. Doğan, R. Leaman, Z. Lu, Ncbi disease corpus: a resource for disease name recognition and concept normalization, *Journal of biomedical informatics* 47 (2014) 1–10.
- [11] L. Smith, L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, et al., Overview of biocreative ii gene mention recognition, *Genome biology* 9 (2008) S2.
- [12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

- [13] R. M. R. Zavala, P. Martínez, I. Segura-Bedmar, A hybrid bi-lstm-crf model for knowledge recognition from ehealth documents., in: TASS@ SEPLN, 2018, pp. 65–70.
- [14] N. Patil, A. Patil, B. Pawar, Named entity recognition using conditional random fields, *Procedia Computer Science* 167 (2020) 1181–1188.
- [15] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional lstm-cnns-crf, *arXiv preprint arXiv:1603.01354* (2016).
- [16] Y. Goldberg, O. Levy, word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method, *arXiv preprint arXiv:1402.3722* (2014).
- [17] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [18] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics* 5 (2017) 135–146.
- [19] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, J. Armengol-Estapé, Medical word embeddings for spanish: Development and evaluation, in: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 124–133.
- [20] C. Cardellino, Spanish billion words corpus and embeddings (march 2016), URL <http://crscardellino.me/SBWCE> (2016).
- [21] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: *Practical ML for Developing Countries Workshop@ ICLR 2020*, 2020.

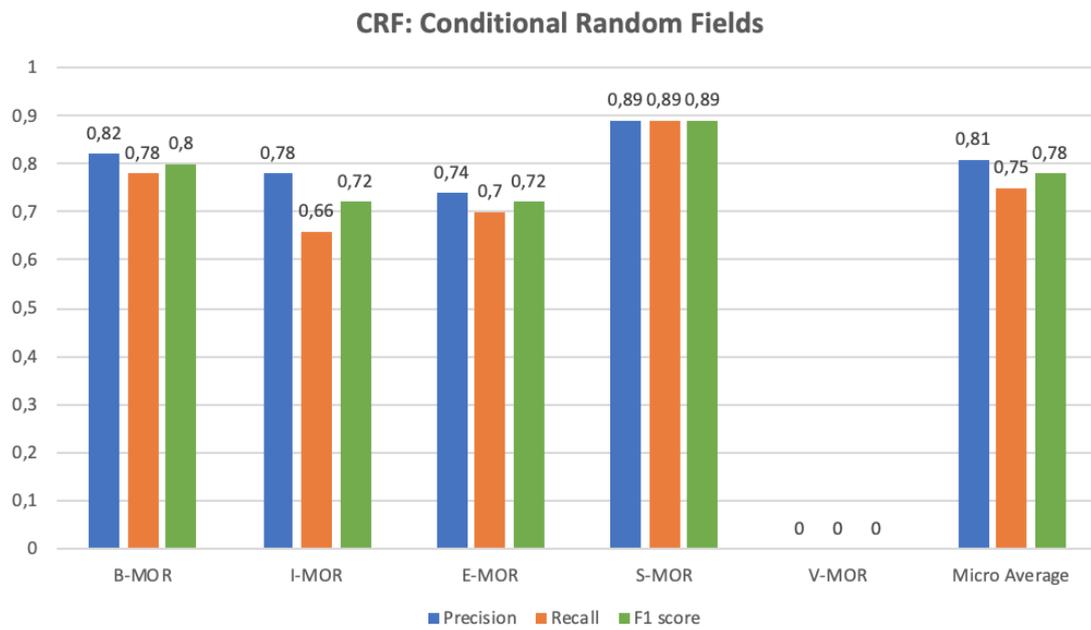


Figure 6: Evaluation of CRF over the test dataset.

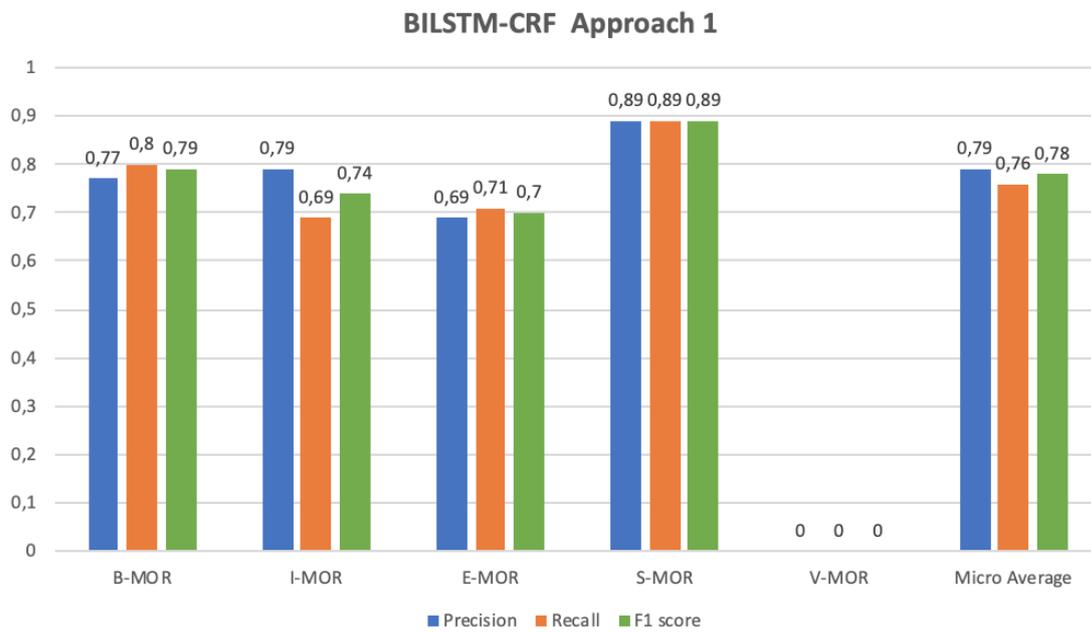


Figure 7: Evaluation of BiLSTM-CRF approach 1 over the test dataset.

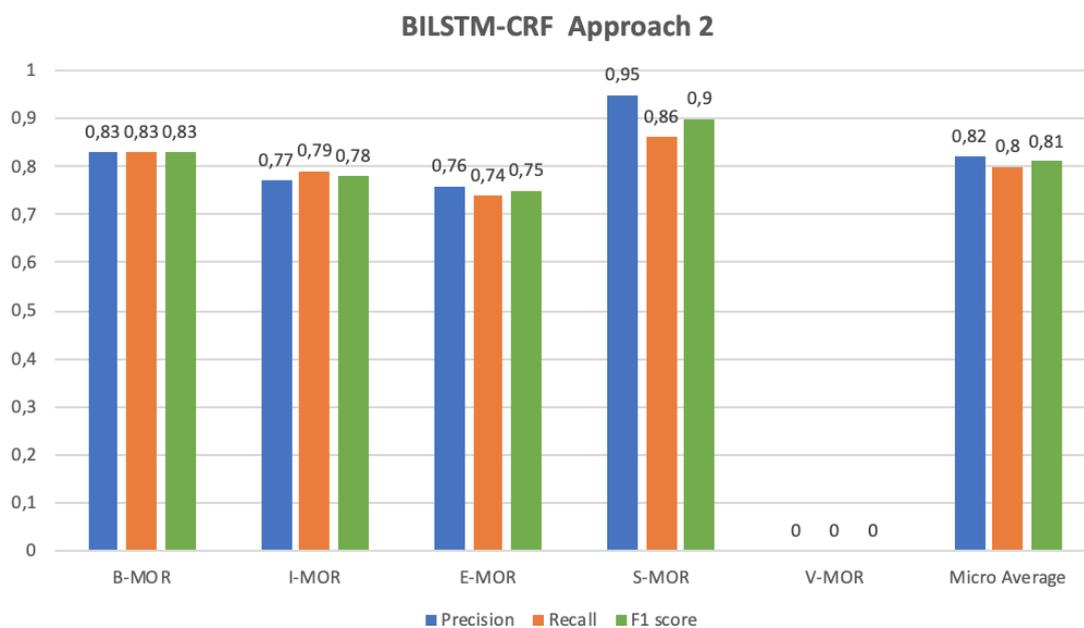


Figure 8: Evaluation of BiLSTM-CRF approach 2 over the test dataset.

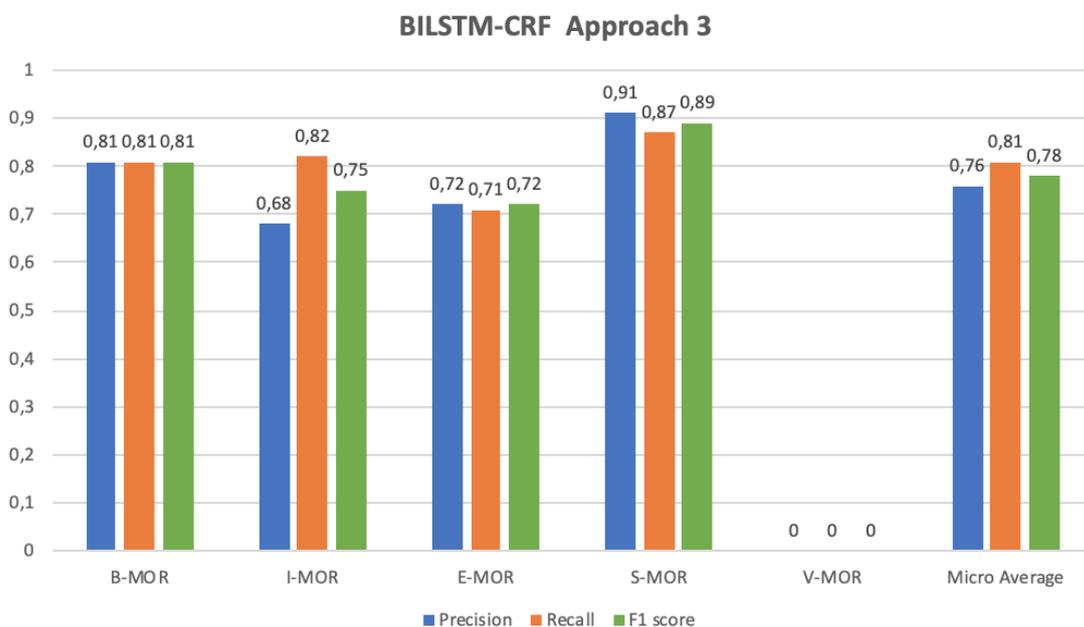


Figure 9: Evaluation of BiLSTM-CRF approach 3 over the test dataset.

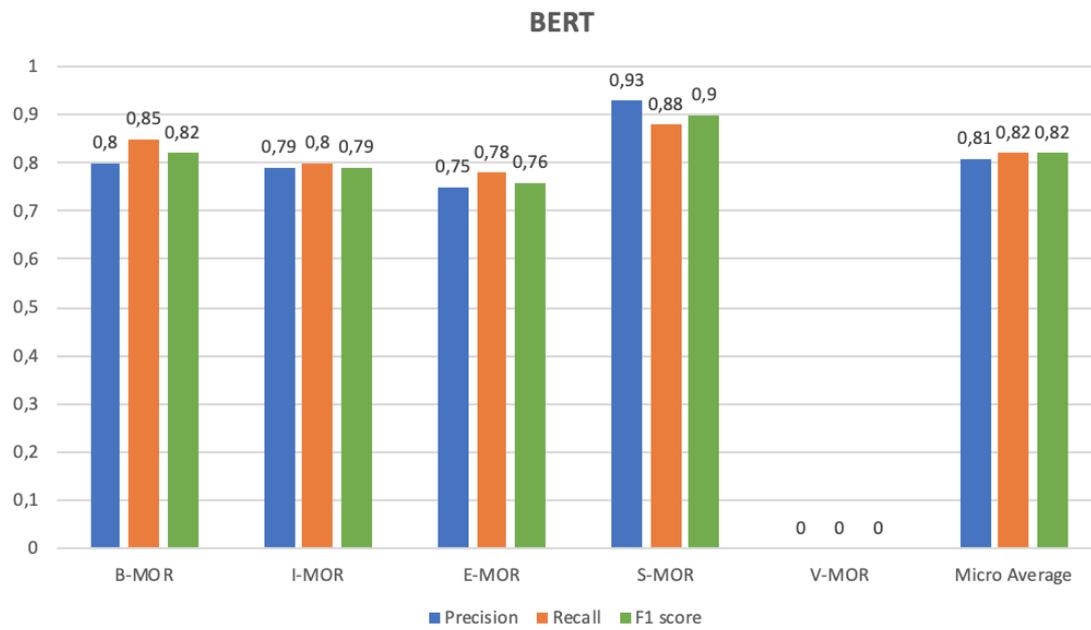


Figure 10: Evaluation of BERT over the test dataset.

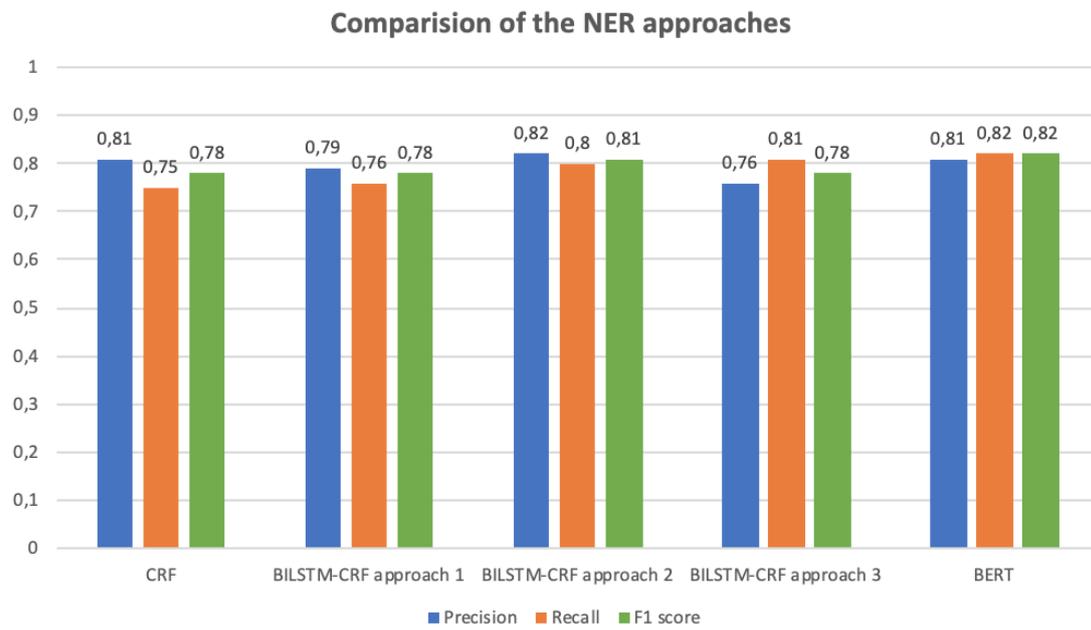


Figure 11: Comparison of the evaluation metrics of the NER approaches over the test dataset.