mBERT and Simple Post-Processing: A Baseline for Disease Mention Detection in Spanish

Antonio Tamayo[a], Diego A. Burgos[b] and Alexander Gelbukh[a]

[a] *Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Av. Juan de Dios Batiz, s/n, 07320, Mexico City, Mexico*
[b] *Wake Forest University, 1834 Wake Forest Road, Winston-Salem, NC 27109, Winston Salem, USA*

Abstract

Automatic disease mention extraction is a relevant task due to its various applications in the medical field. During the last decade, many related works have been published, which have accelerated the progress of this research area, but most of them have been carried out in English. In this work, we propose a deep-learning baseline for this task in Spanish. We report an approach based on transfer learning using multilingual BERT and a straightforward post-processing to tackle the problem. Our system does not use any external resources and rely only on efficient fine tuning, which makes it a fair baseline (Micro F1 = 0.5456) for disease mention identification in Spanish using transformer-based models.

Keywords

Disease mention detection, multilingual BERT, named entity recognition (NER)

1. Introduction

Named entity recognition (NER) has become a key component to more complex systems due to its contribution to document indexing and categorization, knowledge acquisition, etc. Disease mentions in clinical cases are considered a type of domain-specific named entity [1] and its recognition and extraction are of great help for higher-level tasks in the field of medicine. The way NER and other text mining tasks are tackled has rapidly changed over time, though. While text preprocessing (e.g., tokenization, pos-tagging, stopwords), feature engineering, and (limited) data were an essential input for classical probabilistic and machine learning models during years, modern deep-learning models are trained on huge data, but with much simpler feature representations [2].

In this paper, we propose a baseline for disease mention detection in Spanish as a contribution to the entities sub-track of DisTEMIST (Disease Text Mining Shared Task, 2022). This sub-track required the participants to automatically identify disease mentions in a dataset of clinical cases in Spanish [3]. The task is particularly challenging because the evaluation metric measures the system's ability to determine the exact starting and ending location of the disease mention in the document.

For this task, we followed a transfer learning approach, that is, we used and fine-tuned the multilingual BERT (mBERT) language model that was originally trained on a different, more heterogeneous data, i.e., not only medical documents, and that was also trained for a different task. Since we addressed the entities task as a sequence labeling problem, the training data provided by the organizers required substantial preprocessing in order to take it to the format required by the model. Likewise, we carried out simple post-processing on the model's output in order to concatenate subwords, clean up the output, and take the predictions to the DisTEMIST's format.

DisTEMIST is using a dictionary lookup baseline on Levenshtein distance, which looks for train and development annotations in the test set. The same approach for a similar task was used in [4], which yielded a Micro F1 baseline of 0.291. In the present work, we propose the Micro F1 that we reached for

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: atamayoh2019@cic.ipn.mx (A. 1); burgosda@wfu.edu (A. 2); gelbukh@gelbukh.com (A. 3)
ORCID: 0000-0002-5984-7463 (A. 1); 0000-0002-5784-3952 (A. 2); 0000-0001-7845-9039 (A. 3)


this task (0.5456) as a deep-learning baseline for disease mention detection in Spanish since we did not use any external resources such as terminological databases, POS tagging, and so on. As described in detail below, these results were achieved only by fine-tuning mBERT with the best hyperparameter combination found after several experiments and by using a BIO scheme [5] to annotate disease mentions in the training dataset. Moreover, while the literature for this task in English reports F1 scores as high as 89.7 [2], works in disease mention detection in Spanish before DisTEMIST 2022 are scarce, which makes a deep-learning baseline for this task in Spanish helpful for future work reference.

The paper is structured as follows. Section two presents relevant works related to NER and transfer learning in specific domains. Section three describes the dataset, preprocessing steps, the language model and fine-tuning used, as well as other methodological aspects. Section four reports and discusses the results, and Section five concludes.

2. Related Works

As many natural language processing problems, NER in the medical field has been addressed with some of the following three approaches: a) rule-based, b) machine learning, or c) deep learning techniques. As for the rule-based approach, [6] reports promising results in English using rules and regular expressions. On the other hand, the first works utilizing machine learning techniques to tackle NER in the medical field reached their best results using feature engineering, support vector machines (SVMs) and tree-based kernels [7].

With regards to deep learning, following a similar approach to the one we report in this paper, [8] used BERT [9] and ELMO [10] for NER in two corpora in English, namely, PubMed titles and abstracts and clinical notes. They used a BIO scheme and reached the highest F1 score (86.6) on disease mention detection with a BERT-base model on PubMed only using the configurations reported by [6]. [11] carried out transfer learning for biomedical NER in English also. They pre-trained a bidirectional language model (BiLM) on unlabeled data and used it to initialize weights instead of initializing them randomly. This strategy outperformed other experiments without pre-training or with unidirectional pre-training. Their F1 scores reached 87.34 on the NCBI-Disease corpus and 89.28 on the BC5CDR corpus. It is worth noting that besides disease mentions, BC5CDR includes chemical entities, which seem to respond better to bidirectional models. (cf. [8]). In [12, 13], the authors trained BERT from scratch on a huge biomedical corpus in English and then fine-tuned it for different tasks including NER; [12] accomplished an impressive improvement (0.51%) in biomedical NER. Likewise, [14] trained BERT but reports better results fine-tuning the model presented in [12] than with their own.

Regarding experiments with Spanish, [15] published a corpus with nested entities and addressed the NER problem using a biLSTM-CRF architecture and word embeddings trained over both clinical embeddings and Spanish Wikipedia. In [16], the authors proposed a novel join deep learning model to tackle NER and normalization in a corpus of cancer in Spanish achieving an F1 score of 0.87; an equivalent result was reported by [17] using ensemble pre-trained BERT models and post-processing. This latter work also followed a similar method to the one we present in this paper.

3. Methodology

The subsections below provide an outline of the dataset used as well as a description of our system.

3.1. Dataset

The dataset for this task consists of 1,000 clinical cases in Spanish in plain, unstructured text on one hand, and 750 structured text files on the other. Disease mentions in the original clinical cases were manually annotated by experts following thorough guidelines and these annotations were used to generate the structured files, which have the following fields (see also Table 1):

- *filename*: document name
- *mark*: identifier mention id

- *label*: mention type (ENFERMEDAD)
- *off0*: starting position of the mention in the document
- *off1*: ending position of the mention in the document
- *span*: text span

Table 1
Dataset sample

filename	mark	label	off0	off1	span
file_1	T1	ENFERMEDAD	89	111	síndrome postflebítico
file_1	T2	ENFERMEDAD	234	246	sepsis grave
file_1	T3	ENFERMEDAD	285	288	HTA

Out of the 1,000 clinical cases, the organization randomly chose 250 documents for the test dataset and the remaining 750 documents made up the training set. The test set was shuffled with 2,750 more clinical cases, for a total of 3,000 documents to be mined for disease mentions, but the organizers only evaluated NER on the 250 documents originally chosen.

3.2. System Description

Our approach to extract disease mentions from Spanish clinical cases is based on the transfer learning technique using BERT pretrained on a multilingual corpus plus simple post-processing. To implement the fine-tuning process with the mBERT model, the BIO (Begin, Inside, Outside) scheme was used. Since the dataset provided by DisTEMIST is formatted in a different way, a pre-processing was needed to take it to the BIO scheme. Below we describe our system in four subsections, namely, pre-processing, pretrained model description, transfer learning, and post-processing.

3.2.1. Pre-processing

Before converting the plain text dataset described in section 3.1 into the BIO scheme, we first tokenized it using SpaCy [18]. While tokenization is a straightforward task, some flaws in the plain text files tremendously hindered this step. A substantial normalization process was necessary in order to take full advantage of the tokenization stage. After this, each lexical unit in the documents was annotated with its corresponding label (B, I, or O) and the document's tokenization was reversed. For the annotation, we used the disease mentions in the structured dataset as a reference.

3.2.2. Pretrained model

We used the mBERT model which is a BERT version trained in 104 languages on the largest Wikipedias, including Spanish. It is a transformer-based model which was trained to predict masked words and the next sentence.

3.2.3. Transfer learning

To fine-tune the model, we modified its prediction head to tackle the extraction of disease mentions in clinical cases as a token classification problem using the transfer learning approach. In this work, we did not use sentence tokenization because the model's performance decreased. Instead, we fine-tuned the model by taking the whole clinical case text as a sample.

During the fine-tuning process, we conducted a hyperparameter tuning search to identify the best model's configuration using a grid search for the epochs (3, 5, 7) and the learning rate (5e-03, 5e-05, 5e-07). After that, the best results were found with the configuration shown in Table 2.

Table 2
Chosen hyperparameter configuration

Hyperparameter	Value
learning_rate	5e-5
train_batch_size	8
seed	42
optimizer	Adam betas=(0.9,0.999); epsilon=1e-08
lr_schedler_type	linear
num_epochs	7
eval_batch_size	8

For all the experiments, we used 562 texts for training and 188 for validation. The results that we show below correspond to the best performance achieved on the validation partition.

The fine-tuning process was carried out using the Transformer library with the mBERT cased model available at Hugging Face (https://huggingface.co/bert-base-multilingual-cased) and, as infrastructure, we used Google Colab Pro with a GPU Tesla P100 with 27.3 gigabytes of available RAM. The clean version of the data we used for our training process together with the source code to replicate this work are available at a GitHub repository (https://github.com/ajtamayoh/NLP-CIC-WFU-Contribution-to-DisTEMIST-shared-task-2022).

3.2.4. Post-processing

Once we obtained the predictions from our model, simple post-processing was carried out. As mBERT works with subword tokenization, the first step was to concatenate all the subwords belonging to the same disease mention found by the model.

Secondly, we concatenate all the predictions of disease mentions which were found by the model one after the other. This means that if the model detects a disease mention whose final character position plus one concurs with the first position of the next disease mention detected, these two mentions are considered part of the same entity by our system. This was necessary because the model tagged some words as "B" but they were actually a continuation of a previous entity and their label should be fixed to "I".

We also applied simple post-processing based on some orthographic and grammatical rules, which were applied in the order that they appear in Table 3. Moreover, because our model was fine-tuned with a BIO scheme, its predictions come out in the same format. Therefore, the last step consisted in decoding the predictions from the BIO format to the structured format required by DisTEMIST, which was described in section 3.1.

Table 3
Post-processing rules

If the predicted disease mention…	… then apply this rule
Concurs with non-content words or punctuation marks	Leave out of the entities detected
Ends with non-content words or punctuation marks	Delete the match and fix the off1
Contains a mark of new line	Replace the match with a space
Contains a space before and/or after a hyphen or a parenthesis	Delete the space(s) and fix the off1

4. Results and discussion

The best configuration was found by intuitively tuning the hyperparameters instead of using brute force with a grid search. We started from a default configuration and carefully decreased the learning rate while increasing the number of epochs. The results achieved by our system for both training and test can be seen in Table 4. Micro-Precision (MiP), Micro-Recall (MiR), and Micro-F1 (MiF1) were used to measure the model performance. The evaluation on our local training set aimed at determining the best hyperparameter configuration. For this preliminary evaluation we used BIO annotations as gold standard rather than the starting and ending locations of disease mentions in DisTEMIST's data. This is the reason why a comparison with and without post-processing is not presented here as our post-processing rules for this shared task were designed to work on DisTEMIST's format rather than on the BIO format.

Table 4
Results for training and test

	Training			Testing		
Model	**MiP**	**MiR**	**MiF1**	**MiP**	**MiR**	**MiF1**
mBERT	0.5637	0.5801	0.5718	0.6095	0.4619	**0.5456**

While mBERT is inherently a complex model, our approach can be considered simple enough for a baseline, but in turn, compared to a dictionary lookup baseline (cf. [4]), our results (0.5456) are not negligible. A caveat must be added here with regards to the model's ability to identify disease mentions vs the way this task is evaluated. Metrics for disease mention extraction typically assess an exact match of the extracted mention with the starting and ending position of the mention in the gold standard. However, a qualitative scan of the mentions we extracted show our model's capacity to identify disease mentions effectively. The reason why the values for MiR and MiP in Table 4 are low is not necessarily that the model missed to detect the mentions in the gold standard or that it detected irrelevant text spans, but also that it may have truncated or extended a text span that contains a correct mention and that our post-processing was not robust enough. See Table 5 for some examples.

Table 5
Error analysis - Truncated (t), extended (e), and missed (-) disease mention examples

Gold Standard	**Model's prediction**
Amaurosis	amaurosis de uno o ambos ojos (e)
hemihipoestesia derecha	hemihipoestesia (t)
nido vascular epicraneal frontal izquierdo	-
complicaciones neurológicas	-
hematoma protuberancial posterior derecho	hematoma protuberancial posterior derecho de (e)

The examples in Table 5 suggest that, for a baseline, our model can be considered a decent disease mention extractor. The ability of the model to correctly predict some simple and complex disease mentions according to the gold standard is also remarkable, e.g., *hemorragia frontal de angioma cavernoso frontal derecho*, *angioma protuberancial izquierdo*. We acknowledge, though, that a reliable system must meet the text span match requirement to the greatest extent possible.

We intend this baseline to emulate an expert's identification of specialized knowledge in a text without any external resources. The task of accurately delimiting a term in context, or a disease mention in this case, is hard, not only for machines but also for human experts. [19], for example, proved that human experts are good at identifying approximate windows of specialized mentions in context, but they do not perform as well to determine the starting and ending position of the term. We would expect a language model properly fine-tuned in combination with external resources such as dictionaries, ontologies, grammars, taggers, etc. to dramatically outperform our baseline, just like an expert with similar resources would do.

The scarce literature on disease mention detection in Spanish also suggests that there is still much to do in this line. We chose mBERT as we think it fairly represents the Spanish language, but some languages, e.g., English, may benefit more from its subword tokenizer. For example, many disease mentions in English of the pattern noun + noun are translated into Spanish as noun + *de* + noun. It seems the model learns that *de* is a prototypical part of complex terms in Spanish and it may explain the extended text span in the last example in Table 5.

Lastly, in these results we also need to factor in the tokenization issues due to our need to handle the BIO scheme and to the way some acronyms and punctuation interacted with disease mentions. Of course, the combination of all the involved variables may have also impacted the model's performance, that is, the model itself and its configuration during training, which may have not built the best word representation for extracting disease mentions, the data, and the pre- and post-processing.

5. Conclusions

In this work, we propose a baseline for disease mention identification in a corpus of clinical cases in Spanish. It is our contribution to the sub-track 1 of DisTEMIST shared task. Our system is based on the transfer learning technique using the multilingual version of the well-known language model called BERT. Besides the simple post-processing that we carried out, the fine-tuned model reported here can be considered a strong baseline (Micro F1=5456) since it does not use any external resources.

Our qualitative analysis suggests that the model performs well to identify disease mentions, but that accuracy in text span delimitation needs improvement. An effective use of external resources, such as the ones made available by DisTEMIST organizers, should help fixing text span truncation and extension and improving recall and precision. However, the system, as described here, can be used as a starting point for an information extraction system in the medical field.

The optimization of subword tokenization for Spanish is an open line for future work. We think such an improvement would boost performance and would decrease the need for post-processing. Likewise, we plan to use the baseline proposed here and the state of the art from DisTEMIST and other sources to set up a competitive disease mention identification system. We plan to integrate domain-specific resources as well as to replicate and strengthen our experiments with other language models.

6. Acknowledgements

This work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

7. References

[1] Wang Y, Tong H, Zhu Z, Li Y. Nested Named Entity Recognition: A Survey. ACM Transactions on Knowledge Discovery from Data (TKDD). 2022.

[2] Hahn U, Oleynik M. Medical information extraction in the age of deep learning. Yearbook of medical informatics. 2020 Aug;29(01):208-20.

[3] Miranda-Escalada A, Gascó L, Lima-López S, Farré-Maduell E, Estrada D, Nentidis A, Krithara A, Katsimpras G, Paliouras G, Krallinger M. Overview of DISTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings. 2022.

[4] Miranda-Escalada A, Farré E, Krallinger M. Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. IberLEF@ SEPLN. 2020 Sep:303-23.

[5] Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. InNatural language processing using very large corpora 1999 (pp. 157-176). Springer, Dordrecht.
[6] Eftimov T, Koroušić Seljak B, Korošec P. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PloS one. 2017 Jun 23;12(6):e0179488.
[7] Patra R, Saha SK. A kernel-based approach for biomedical named entity recognition. The Scientific World Journal. 2013 Jan 1;2013.
[8] Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474. 2019 Jun 13.
[9] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
[10] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv 2018. arXiv preprint arXiv:1802.05365. 1802;12.
[11] Sachan DS, Xie P, Sachan M, Xing EP. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. InMachine learning for healthcare conference 2018 Nov 29 (pp. 383-402). PMLR.
[12] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.
[13] Akhtyamova L. Named entity recognition in Spanish biomedical literature: Short review and bert model. In2020 26th Conference of Open Innovations Association (FRUCT) 2020 Apr 20 (pp. 1-7). IEEE.
[14] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. 2019 Apr 6.
[15] Báez P, Villena F, Rojas M, Durán M, Dunstan J. The Chilean Waiting List Corpus: a new resource for clinical named entity recognition in Spanish. InProceedings of the 3rd clinical natural language processing workshop 2020 Nov (pp. 291-300).
[16] Xiong Y, Huang Y, Chen Q, Wang X, Nic Y, Tang B. A joint model for medical named entity recognition and normalization. Proceedings http://ceur-ws. org ISSN. 2020;1613:0073.
[17] García-Pablos A, Perez N, Cuadros M. Vicomtech at cantemist 2020. InProceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings 2020.
[18] Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017 Jul;7(1):411-20.
[19] Estopà, R, Martí, J, Burgos, D, Luna, J, Monserrat, S, Montané, A, Muñoz, P, Quispe, W, Rivadeneira, M, Rojas, E, Sabater, M, Salazar, H, Samara, A, Santis, R, Seghezzi, N, Fernández, S, Souto, M. La identificación de unidades terminológicas en contexto: de la teoría a la práctica. 2006:1000-30.