Clinical Named Entity Recognition and Linking using BERT in Combination with Spanish Medical Embeddings

Javier Reyes-Aguillón[1], Rodrigo del Moral[1], Orlando Ramos-Flores[2], Helena Gómez-Adorno[2] and Gemma Bel-Enguix[3]

[1]*Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México*

[2]*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México*

[3]*Instituto de Ingeniería, Universidad Nacional Autónoma de México*

Abstract

This paper presents our approach to the DisTEMIST 2022 shared task on finding disease mentions in clinical texts and linking them to a SNOMED-CT term. Our architecture combines a Spanish language BERT model with a neural network layer for sequence classification. We used a word embedding model generated from scientific documents related to the medical field for the entity linking problem. The obtained results confirm that a BERT-based system provides reasonable performance in the named entity recognition task. In the entity linking task, the use of word embeddings proved to be valuable for cross-database searching. Although the proposed solution is functional, there are still some areas for improvement in our system.

Keywords

DisTEMIST, Named Entity Recognition, Entity Linking, Machine Learning, BERT

1. Introduction

In recent years, the problem of named entity recognition within medical texts has received increasing attention from both scientific and clinical fields of expertise. Several shared tasks are being organized, such as CLEF eHealth 2021 [1] which included the creation of systems that recognize and classify entities in the Spanish language belonging to the area of radiology. IberLEF eHealth-KD 2019 [2], 2020 [3] and 2021 [4] have focused on the identification and classification of entities within articles extracted from the PubMED library, as well as from WikiNews and the CORD-19 corpus of COVID-19-related scientific resources. Entity recognition has also been applied in other fields that are health-related, such as CLEF 2020's CheMU shared task [5], which sought to identify entities related to chemical reactions.

This growing interest in the subject arises from the realization that many systems could

benefit from organizing information from all types of medical texts. It is essential to be aware, from the beginning, of the opportunities generated by linking concepts from disorganized texts with concepts inside several types of knowledge bases. Firstly, we would be able to delve deeper into these concepts and the information from other related entities. Other advantages of creating medical text analysis systems are the extraction of information in large volumes to obtain data that can train machine learning algorithms creating the possibility of knowledge generation in a semi-supervised or unsupervised form.

In this paper, we propose a system for identifying diseases mentions in clinical texts. Our system consists of a Spanish language BERT model combined with a neural network layer for token classification. In addition, our system links the identified mentions to the SNOMED-CT diseases database with the use of a word embedding model trained on medical texts.

The rest of this paper is structured as follows. Section 2 describes the data set used for the task and briefly presents the approach followed for its creation. Section 3 describes the methodology for the Named Entity Recognition and Entity Linking subtracks, as well as for pre-processing and post-processing the data. In section 4, the experiment configuration and results are reported. Section 5 discuses the obtained results and gives future work direction.

2. Data Set

The data set [6] consists of 1,000 clinical cases from more than a dozen medical specialties, all of them written in Spanish and manually annotated for disease mentions with the use of the *brat* tool [7]. The entity mentions also have their codes within the SNOMED-CT disease database (Systematized Nomenclature of Medicine - Clinical Terms). The annotation and standardization process was carried out by two specialist physicians and reviewed by a third physician over approximately two years. The entire corpus consists of 10,666 disease mentions and 10,318 unique standardized mentions.

The corpus was divided into two sets (training and testing) to leave data out to evaluate the shared task. The training set provided to the participants consisted of 750 annotated documents to train the models (584 normalized with their linked entities). The testing set consisted of 250 unannotated documents for the task evaluation. In addition, the organizers included 2,750 background documents in the test set. The models are required to make predictions on the 3,000 documents. However, these background documents were not considered for performance evaluations.

3. Methodology

The shared task is subdivided into two smaller tasks. For the first subtrack, we propose a Named Entity Recognition system to identify the diseases mentions in the documents provided in the DisTEMIST corpus. For the second subtrack, we designed and implemented an Entity Linking system that links each annotated mention to its ID codes inside the SNOMED-CT diseases database.

3.1. Subtrack 1: Named Entity Recognition

The system designed for the NER subtrack is comprised of three parts. First, a pre-processing stage converts the raw documents and provided annotations to a suitable format for the BERT model. Then, the model is trained using data from the DisTEMIST train set, and the test data set is processed through the model. Finally, a post-processing step prepares the submission file using the format required for the evaluation script. The specific steps for the entire process can be observed in Figure 1.

3.1.1. Pre-processing

The organizers provided annotations in the *brat standoff* format, with each of the entities referenced to the document where it belongs. However, to produce a suitable input for the entity recognition algorithm, the entities from each document in the data set had to be mapped to a BIO-tags representation. The process consisted of tokenizing all entities for the program to distinguish between beginning-tokens and inside-tokens (in the case of multi-word entities). Subsequently, the annotated entities were aligned with their positions in the original texts using the offsets provided by the *brat* format. Finally, the portions of the documents that did not yet have labels were tokenized, and all those tokens were assigned the label "O".

For tokenization, we decided to keep some of the non-alphanumeric characters since we noticed some entities contained inside hyphens, parentheses, and commas, among other symbols. Therefore, these symbols were treated as single tokens and annotated with the inside-token BIO-tag. In contrast, periods and line breaks were omitted from the tokenization since, in these instances, they are just for denoting the end of each sentence.

Finally, we perform sentence tokenization to train the model and predict the entities more efficiently.

3.1.2. BERT-based Classifier

Once the clinical texts were pre-processed and divided into sentences, we used a BERT pre-trained model for word classification to complete the subtrack on entity recognition.

The NER subtrack aims at identifying all diseases mentioned within a set of clinical documents. Thus we opted for a classifier system that could sort words into three different classes: "B-ENF" if the word is the first part of a disease mention (regardless of whether such mention consists of one or multiple words); "I-ENF" if the word is part of a disease mention, but is not the first word in that specific mention; and finally, the class "O" is used for terms that are not part of any disease mention.

Taking these considerations into account and being aware of the state of the art in NER problem solving, we decided to implement a system based on a Transformer pre-trained model; in this case, we used mBERT [8], and BETO [9]. In both cases, an additional layer was trained for token classification using CoNLL 2002 shared task data. Finally, we trained a top classification layer with data from the DisTEMIST corpus.
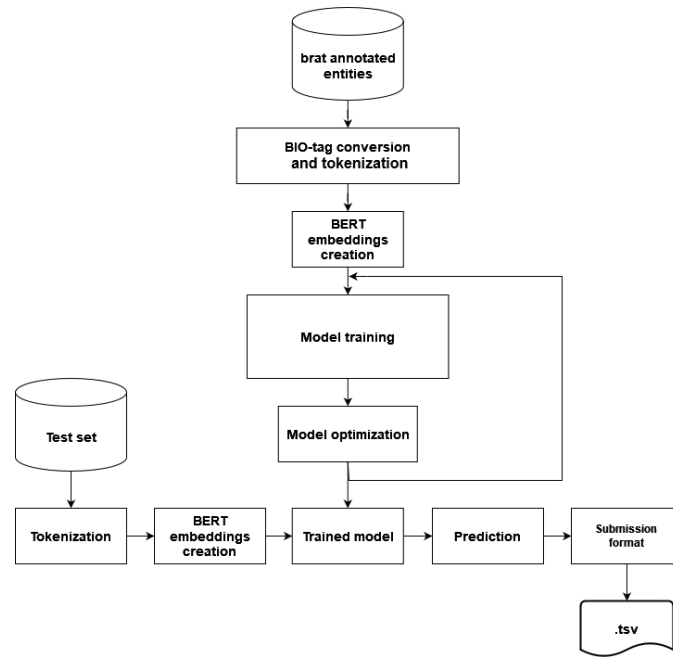
Figure 1: Subtrack 1 System Diagram

3.1.3. Post-processing

After the model's prediction, the output underwent a thorough post-processing stage to produce a definitive annotation document for its submission and evaluation, following the *brat standoff* format. The process started by joining all sub-embeddings generated by the BERT tokenizer, taking into account only the predicted label for the first sub-embedding and extending it through all the other fragments, ignoring their predicted labels. Then, we discarded all tokens predicted as the "O" label and tokens predicted as "I-ENF" despite not being immediately preceded by a "B-ENF". Other "I-ENF" labels were also discarded.

Further, we joined the multi-word entities and discarded their BIO-tags. With the set of identified entities, our algorithm discarded false positives that consisted only of punctuation marks or numbers. After that, we conducted an alignment of the identified entities with the original texts to obtain each entity's positions and append them to the submission form.

Finally, we included the recognized entities in a .tsv file for submission and evaluation.

3.2. Subtrack 2: Entity Linking

The proposed solution for the Entity Linking subtrack also consists of three parts. First, the pre-processing stage converts the annotations from the first subtrack into a list of entities to be used by the word embedding model. Then, the codes of the unlinked entities are predicted through a vector similarity search. Finally, a post-processing step prepares the submission file for the evaluation script with the predicted codes included. The specific steps for the entire process are shown in Figure 2.
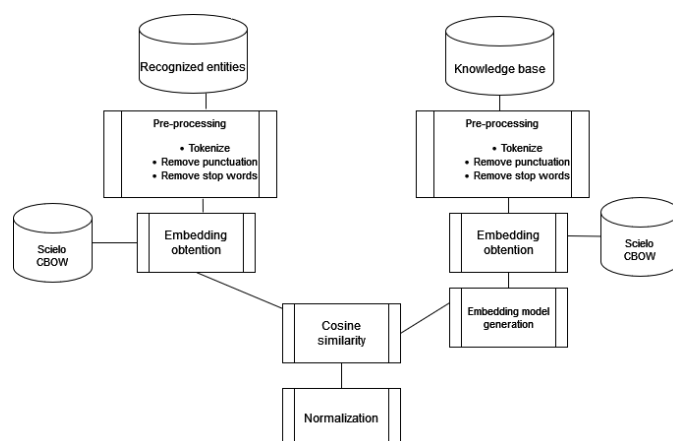
Figure 2: Subtrack 2 System Diagram

3.2.1. Pre-processing

We read the previously identified entities as strings and performed a tokenization process. We discarded punctuation marks, the capitalization of words, and stop words.

Subsequently, using embeddings from a pre-trained Spanish language model (SciELO-CBOW) with medical terms [10], the tokens of the named entities were mapped to their vector representations, obtaining an average vector for the case of multi-word entities.

We downloaded the most recent version of the SNOMED-CT database [1] and generated a plain text file from it. This text file contains the names of diseases mapped with their associated codes. Then, we mapped the diseases to their embeddings using the same model.

3.2.2. Entity Linking

To link the diseases to their identifiers in the SNOMED-CT database we compare the cosine similarity between the average vectors of each recognized entity and the SNOMED-CT disease term vectors obtained with the SciELO-CBOW model.

Vectors that showed a cosine similarity of 1, i.e., were equal, were assigned the label "EXACT" in the results. Vectors that did not have an exact match were given the ID value of the vector with the highest similarity and assigned the label "NARROW". When a vector matched more than one vector in its maximum cosine similarity, we assigned all the highest-ranked vector's IDs to that disease and added the "COMPOSITE" label.

3.2.3. Post-processing

The post-processing of the results from the entity linking model consisted of creating a new *.tsv file, to which "CODES" and "SEMANTIC RELATION" columns can be added to the columns already generated from subtrack 1.

---

[1]https://www.snomed.org/

4. Experiments and Results

We performed a series of experiments aimed at improving the model's performance. For the NER subtrack, we trained two different BERT-based models for both the validation and testing stages. Regarding the Entity Linking subtrack, we experimented with three different word embedding models before finding the optimal model for this specific corpus. In addition, we also present the evaluation results from the testing stage submitted to the DisTEMIST organizers.

4.1. Subtrack 1: Named Entity Recognition

We built our first system using a pre-trained mBERT model, with an extra layer for named entity recognition trained on the CoNLL 2022 shared task data. Such model had a final stage of training with 80% of the training data, i.e., 600 clinical texts, and validation with the remaining 150 documents. The optimization of the model was carried out using a stochastic gradient descent method with weighted decay. In Table 1, we show the parameters used for the final training of all models.

Table 1
NER Training Parameters

Parameter	Value
Sentence maximum length	256
Batch size	8
Epochs	5
Learning rate	constant

In the second system, we evaluated the BETO Spanish language model, with the same training of two additional layers as those placed for the first system.

Once the models were validated and the training parameters adjusted, we proceeded to the testing stage. For this stage, we re-trained the BERT based models using 100% of the training set. This was possible since the algorithm chosen for optimization does not need a validation set to tune the model. Using these final models, we predicted the annotations for the test data set and submitted them to the task organizers for evaluation.

Table 2 presents the evaluation results for both the experiments and the submitted predictions. The base model column indicates which base model was used to train the classification algorithm. The stage column denotes which data set was predicted with the NER model. The performance metrics provided by the evaluation script are micro-averaged precision, micro-averaged recall and micro-averaged F1 score. For the sake of comparison, the best performing results from the testing stage are highlighted.

4.2. Subtrack 2: Entity Linking

We mapped the identified entities to the SciELO model for the entity linking system and calculated the cosine similarity for each vector obtained from the NERs and the vectors from the

Table 2
NER Results Evaluation

Base model	Stage	MiP	MiR	MiF
BERT	Validation	0.5600	0.3580	0.4368
BETO	Validation	0.7086	0.4401	0.5430
mBERT	Testing	0.4540	**0.4619**	0.4579
BETO	Testing	**0.6010**	0.4488	**0.5139**

SNOMED-CT knowledge base). With this process, we obtained a code for each entity according to the Top 10 list returned by the function.

Table 3 shows the results obtained from the entity linking model, using the same column structure as Table 2. We highlighted the best-performing results from the testing stage for comparison. It is worth mentioning that different metrics were obtained due to running different models with different parameters in the first subtrack.

Table 3
EL Results Evaluation

Base model	Stage	MiP	MiR	MiF
mBERT	Validation	0.2998	0.1460	0.1964
BETO	Validation	0.3556	0.1650	0.2254
mBERT	Testing	0.2267	**0.1494**	0.1801
BETO	Testing	**0.2754**	**0.1494**	**0.1937**

5. Conclusions

The results obtained by the named entity recognition system for the 250 test texts of the DisTEMIST corpus generally look congruent, suggesting that the use of Transformer networks may be a promising approach for this type of problem. We believe that the error produced by the system can be decreased by narrowing down the selection of medical specialties covered by the clinical texts, as well as increasing the size of the training data set. On the other hand, the results obtained with the entity linking system show that word vectors can help the search using similarity measures. The results from this task can benefit from using a dictionary of terms and codes to make direct inquiries in addition to the word embedding similarity search. However, there is still an excellent opportunity for improvement on both algorithms.

In the DisTEMIST shared task scoreboard, our team achieved 12th place out of 17 teams for the named entity recognition subtrack; and 13th place out of 15 in the entity linking subtrack. These results reassure us that we can continue improving the system for such problems.

Acknowledgments

References

[1] V. Cotik, L. Alonso Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L. Defrancesca, A. Dellanzo, M. Fernández Urquiza, Overview of CLEF eHealth Task 1-SpRadIE: A challenge on information extraction from Spanish radiology reports, in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. Bucharest, Romania, September 21st to 24th, 2021., 2021, pp. 732–750. URL: http://ceur-ws.org/Vol-2936/paper-61.pdf.

[2] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, 2019, pp. 1–16. URL: http://ceur-ws.org/Vol-2421/eHealth-KD_overview.pdf.

[3] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Diaz, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020). Málaga, Spain, September 23th, 2020, 2020, pp. 71–84. URL: http://ceur-ws.org/Vol-2664/eHealth-KD_overview.pdf.

[4] A. Piad-Morffis, S. Estevez-Velarde, Y. Gutierrez, Y. Almeida-Cruz, A. Montoyo, R. Muñoz, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021, Procesamiento del Lenguaje Natural 67 (2021) 233–242. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392.

[5] J. He, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa, A. Albahem, J. Wang, Y. Ren, Z. Zhang, Y. Zhang, M. H. Dao, P. Ruas, A. Lamurias, F. M. Couto, D. Lowe, J. Mayfield, A. Köksal, H. Dönmez, O. Arzucan, D. Mahendran, G. Gurdin, N. Lewinski, C. Tang, B. T. McInnes, P. R. Rao, S. L. Devi, L. Cavedon, T. Cohn, T. Baldwin, K. Verspoor, An extended overview of the clef 2020 chemu lab: Information extraction of chemical reactions from patents, Julien Knafou 10 (2020) 11.

[6] E. Farré-Maduell, L. Gascó, S. Lima, A. Miranda-Escalada, M. Krallinger, DisTEMIST Guidelines: detection and normalization of disease mentions in spanish clinical cases, 2022. URL: https://doi.org/10.5281/zenodo.6477407. doi:10.5281/zenodo.6477407, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[7] 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 Session at EACL 2012, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107.

[8] 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).

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020, pp. 1–10.

[10] F. Soares, M. Villegas, A. Gonzalez-Agirre, J. Armengol-Estapé, S. Barzegar, M. Krallinger, Fasttext spanish medical embeddings, 2020. URL: https://doi.org/10.5281/zenodo.3744326. doi:10.5281/zenodo.3744326, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) and the ICTUSnet project (https://ictusnet-sudoe.eu/en/).