electronics

MDPI

Article

Improving Medical Entity Recognition in Spanish by Means of Biomedical Language Models

Aitana Villaplana [1], Raquel Martínez [2,*] and Soto Montalvo [3]

[1] VÓCALI Sistemas Inteligentes S.L., Parque Científico de Murcia, Carretera de Madrid km 388, Complejo de Espinardo, 30100 Murcia, Spain; aitana.villaplana@vocali.net
[2] Dept. of Lenguajes y Sistemas Informáticos, Escuela Técnica Superior de Ingeniería Informática, Universidad Nacional de Educación a Distancia, Juan del Rosal 16, 28040 Madrid, Spain
[3] Dept. Informática y Estadística, Escuela Técnica Superior de Ingeniería Informática, Universidad Rey Juan Carlos, C/Tulipán s/n, 28933 Móstoles, Spain; soto.montalvo@urjc.es
* Correspondence: raquel@lsi.uned.es

Abstract: Named Entity Recognition (NER) is an important task used to extract relevant information from biomedical texts. Recently, pre-trained language models have made great progress in this task, particularly in English language. However, the performance of pre-trained models in the Spanish biomedical domain has not been evaluated in an experimentation framework designed specifically for the task. We present an approach for named entity recognition in Spanish medical texts that makes use of pre-trained models from the Spanish biomedical domain. We also use data augmentation techniques to improve the identification of less frequent entities in the dataset. The domain-specific models have improved the recognition of name entities in the domain, beating all the systems that were evaluated in the eHealth-KD challenge 2021. Language models from the biomedical domain seem to be more effective in characterizing the specific terminology involved in this task of named entity recognition, where most entities correspond to the "concept" type involving a great number of medical concepts. Regarding data augmentation, only back translation has slightly improved the results. Clearly, the most frequent types of entities in the dataset are better identified. Although the domain-specific language models have outperformed most of the other models, the multilingual generalist model mBERT obtained competitive results.

Keywords: biomedical natural language processing; Spanish biomedical entity recognition; pre-trained language models; data augmentation

Citation: Villaplana, A.; Martínez, R.; Montalvo, S. Improving Medical Entity Recognition in Spanish by Means of Biomedical Language Models. *Electronics* **2023**, *12*, 4872. https://doi.org/10.3390/electronics12234872

Academic Editors: Shangsong Liang, Zaiqiao Meng

Received: 28 October 2023
Revised: 27 November 2023
Accepted: 30 November 2023
Published: 2 December 2023

1. Introduction

Healthcare professionals generate huge amounts of medical literature and clinical data that are stored digitally, resulting in a large availability of medical texts, many of which are stored in an unstructured text format. Thus, one of the most important challenges in medical data processing is the transformation of unstructured information into well-defined data.

NER is a Natural Language Processing task to identify entities in text and classify them into predefined categories. In the medical domain, NER plays a crucial role by extracting medical terminology, i.e., meaningful text segments, such as diseases, symptoms, drugs, etc.

A recent survey of the state-of-the-art proposals for NER [1] focused on deep learning approaches for the recognition of generic Named Entities (NEs) (e.g., organization, person and location) in English language. It concluded that deep-learning-based NER benefits from the advances made in pre-trained embeddings in modeling languages without the need for complicated feature engineering. Focused on clinical texts, ref. [2] presents a survey on NER and Relationship Extraction, concluding that linguistic model-based approaches are likely to continue to increase in the coming years. In addition, the authors of [3] compared domain-specific models and generalist models for NER in clinical trials in English, and concluded that domain models performed better than generalist models.

NER in the medical domain, and in languages other than English, is hampered by the scarcity of corpora and other resources and tools. In the case of Spanish, some annotated biomedical corpora have recently been created on certain (sub)categories of entities, e.g., only entities of a specific family of diseases [4–9]. An imbalance between classes of entities is common, leading supervised systems to recognize mainly the most frequent entities. Fortunately, although most of the available language models have been trained on general texts, models from the biomedical domain are becoming available.

Regarding the work for Spanish, ref. [10] presented an extension of the Freeling Spanish analyzer [11], FreelingMed, by extending the resources of Freeling with medical dictionaries and SNOMED-CT. Others works, such as [12,13], have focused on cross-lingual approaches in order to take advantage of the English resources and make different projections into Spanish. However, both works have Oracle terms as their starting point. Reference [14] presented UMLSMapper, a lexically/knowledge-driven system that relies on several terminological resources from UMLS. In [15], UMLSMapper is combined with cross-lingual approaches obtaining very promising results. Proposals for Spanish NER based on Bidirectional Long Short-Term Memory (Bi-LSTM) networks and Conditional Random Fields (CRFs) are presented in [7,16]. In [17], a Bi-LSTM network is used to resolve the NER task of clinical notes in Spanish and Swedish, evaluating several types of embeddings, both generated from in-domain and out-of-domain text corpora; the authors concluded that, with in-domain embeddings, the NER task is improved compared to with shallow learning methods. In [18], a pre-trained BERT language model on Spanish biomedical literature, fine-tuned for detecting pharmacological substances, compounds, and proteins, is presented. In 2020, the Cantemist (https://temu.bsc.es/cantemist/ (accessed on 3 July 2023)) evaluation campaign was presented, with the aim of exploring the automatic detection of mentions of tumor morphology in medical documents in Spanish, as well as the assignment of eCIE-O (ICD-O is an acronym for International Classification of Diseases for Oncology. It is an extension of the International Statistical Classification of Diseases and Related Health Problems applied to the specific domain of tumor diseases, and is the standard coding for the diagnosis of neoplasms https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_o_3.html (accessed on 3 July 2023)). In this case, the NEs involved are very specific to tumor morphology. The two best participant teams were [19,20]. In [19], NER is regarded as a machine reading comprehension problem, whose task is to answer questions regarding different types of entities based on given passages. The authors used a BERT model which was further pretrained using the CANTEMIST corpus. [20] and used an end-to-end deep-learning-based system from pre-trained BERT models as the basis for the semantic representation of the texts.

In this work, we focus on the NER scenario proposed in the 2021 eHealth Knowledge Discovery (eHealth-KD) challenge (https://ehealthkd.github.io/2021 (accessed on 3 July 2023)) [21]. The goal was to identify four types of entities that are relevant terms representing semantically important elements in a sentence. The most successful systems in this NER challenge were based on the use of contextual language models. The winning team was PUCRJ-PUCPR-UFMG [22] with a transformer-based model, the multilingual version of BERT [23] (mBERT), with an end-to-end architecture, which not only addresses the NER task but jointly extracts relationships between entities and other eHealth-KD tasks. The second best team was Vicomtech [24] with the IXAmBERT transformer model [25], a multilingual model for English, Spanish, and Basque, and a classifier formed by a Neural Network that received the input tokens and jointly produced predictions for the NER and relation extraction tasks. The third best team was IXA [26] with a system designed as a pipeline for classifiers, each independently tuned for NER and relation extraction. For the NER task, texts were encoded using an XML-RoBERTa transformer model [27] and a Feedforward Neural Network (FNN) was used as the classifier. The top three teams in the NER task approached the task jointly with the relation extraction task, indicating that joint training improves entity identification.

To summarize, previous work shows that the use of Bert-type generalist large language models improves the results of other non-transformer-based approaches. On the other hand, the limited availability of Spanish open annotated corpora hampers the fair comparison of different approaches. In this sense, the eHealth-KD challenge provided a framework for experimentation that allows for comparing different techniques and models. In the latest 2021 campaign, participants did not use contextual language models from the biomedical domain, but instead used generalist contextual language models.

In this paper, we present a proposal for recognising named entities in Spanish medical texts using transfer learning and data augmentation techniques. Our goal is to improve the identification of medical entities in Spanish using recent public domain-specific language models, trying to overcome the limitation imposed by the imbalance between different types of entities. We hypothesize that the use of domain models can improve NER task performance in the experimental framework of the eHealth-KD Challenge 2021. As far as we know, no previous work has used this combination of techniques to improve the identification of biomedical entities in Spanish. The key contributions are as follows: (i) the use and fine-tuning of public transformer-based models previously trained on Spanish biomedical datasets to improve results in the NER task; and (ii) the selection of back translation as the data augmentation technique to mitigate data imbalance.

2. Materials and Methods

The eHealth-KD 2021 challenge proposed a framework to the automatic sentence-level annotation of multi-token entities and binary relations among them, attempting to capture a large part of factual semantics. Thus, two tasks were addressed: entity recognition and relation extraction. Here, however, we only focus on the entity recognition task.

2.1. NER Task

Four types of entities are considered [21]:

- Concept: identifies a relevant term, concept, or idea.
- Action: identifies a process or modification of other entities. It can be indicated by a verb or verbal construction, such as "afecta" (affects), but also by nouns, such as "exposición" (exposition), where it denotes the act of being exposed to the sun, and "daños" (damages), where it denotes the act of damaging the skin (see Figure 1).
- Predicate: identifies a function or filter of another set of elements, which has a semantic label in the text, such as "mayores" (older), and is applied to an entity, such as "personas" (people) with some additional arguments such as "60 años" (60 years) (see Figure 1).
- Reference: identifies a textual element that refers to an entity, of the same sentence or of different one, which can be indicated by textual clues such as "esta" (this), "aquel" (that one), etc.

Thus, not only biomedical terms are considered name entities in this challenge, but also some general language elements.

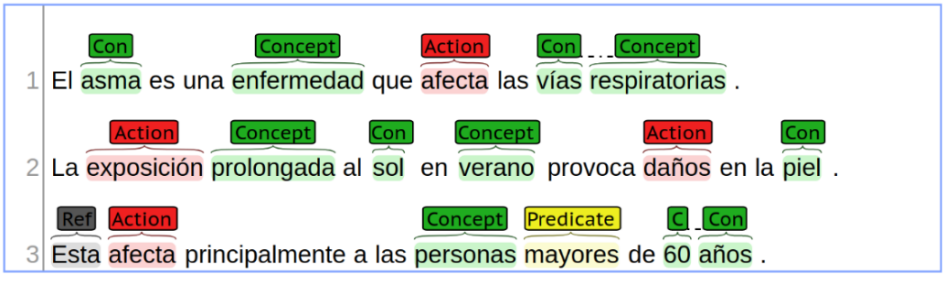


Figure 1. Examples of named entities, where text in green with labels "Con", "C" refers to "Concept" type; text in red to "Action" type; text in yellow to "Predicate" type; and text in grey with "Ref" label refers to "Reference" type [21].

Figure 1 shows the entities appearing in a set of sentences with their respective entity types. Note that some entities, such as "vías respiratorias" (airways) and "60 años" (60 years), are multi-word entities. When managing multi-word entities, IOB (Inside-Outside-Beginning) notation is often used to represent them. IOB provides labels to indicate the entity boundaries: B-entity (first word of the entity); I-entity (subsequent words); and O (non-entity words) [28]. For example, in Figure 1, the text segment "... las vías respiratorias" is represented as O for "las" (the), B-Concept for "vías", and I-Concept for "respiratorias".

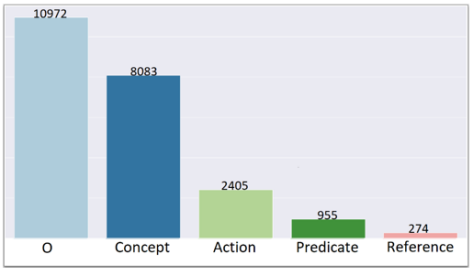
2.2. Dataset

The dataset provided by the organizers contains sentences extracted from MedlinePlus (https://medlineplus.gov/ (accessed on 1 September 2023)) (health information resource), Wikinews (https://www.wikinews.org/ (accessed on 1 September 2023)) (news), and the CORD-19 corpus [29] (scholarly articles about COVID-19), which are all related to health topics. All sentences are in Spanish, except those in the CORD corpus which are in English. The dataset is divided into three collections: training, development, and testing, as shown in Table 1 [21].

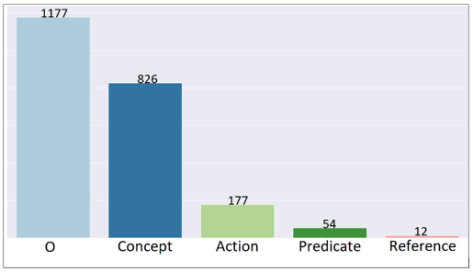
Table 1. Composition of the collections and their size in numbers of sentences.

Collection	Source	Language	Num. Sent.
Training	MedlinePlus	Spanish	1200
	Wikinews	Spanish	300
Development	MedlinePlus	Spanish	25
	Wikinews	Spanish	25
	CORD	English	50
Testing	MedlinePlus	Spanish	75
	Wikinews	Spanish	75
	CORD	English	150
Total			1800

Figure 2 shows the frequency of each type of entity and the total number of entities. Regarding the training dataset, it is unbalanced in terms of the number of entities of each type, the most frequent being "Concept" and the least frequent "Reference". The development dataset presents a similar imbalance, but it includes part of the CORD-19 corpus. We use both datasets for training the models.



(a) Frequencies in the training dataset.



(b) Frequencies in the development dataset.

Figure 2. Frequencies by type of entity in the dataset.

Note that there are 50 documents about COVID-19 in English (CORD-10 corpus) in the datasets used for training the models, but half of the test dataset contains that type of document. Transfer learning techniques may reduce the difficulty of correctly identifying and classifying COVID-19-related entities in the test dataset. On the other hand, the imbalance among the different types of entities can be a key factor in the performance of the models for the types that have very few examples. To overcome this limitation, we used data augmentation techniques.

2.3. Evaluation Metrics

For evaluation, we used Precision, Recall, and F1-Score metrics as defined by the eHealth-KD organizers, where "correct" (*C*), "partial" (*P*), "missing" (*M*), "incorrect" (*I*), and "spurious" (*S*) matches are based on the start and end of text spans and the corresponding entity type.

A "Correct" match is when the spans and entity type are equal; when the start and end values match, but not the type, this is an "incorrect" match; a "partial" match is when there is a partial match in the interval of [start, end] values; "missing" matches are those that appear in the goldstandard, but not in the output file; and "spurious" matches are those that appear in the output file but not in the goldstandard. Thus, *Precision* (P), *Recall* (R), and *F*1-Score metrics are defined as follows (Equations (1)–(3), respectively).

$$Precision = \frac{C + \frac{1}{2}P}{C + I + P + S} \tag{1}$$

$$Recall = \frac{C + \frac{1}{2}P}{C + I + P + M} \tag{2}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

2.4. Models and Techniques

In the following, we present the pre-trained models, the data augmentation techniques, and the proposed models.

2.4.1. Pre-Trained Language Models

Two generalist and three domain-specific models were selected. One of these models was trained on multilingual texts, and the other four on Spanish texts only.

Generalist models:

- mBERT: BERT multilingual base model [30] trained on 104 languages with data from Wikipedia to perform Masked Language Modeling (MLM).
- BETO: a Spanish version of BERT [31] trained with texts from Wikipedia, Wikinews, and Wikiquotes in Spanish, among other sources.

Domain-specific models:

- RoBERTa$_{Bio}$ [32]: based on the model RoBERTa [33], but trained with several Spanish biomedical corpora, such as Spanish Biomedical Crawled Corpus [34] and SciELO Spain (https://scielo.isciii.es/scielo.php (accessed on 13 July 2023)).
- RoBERTa$_{Clinical}$: trained with the same resources as RoBERTa$_{Bio}$, but, in addition, a corpus of clinical reports with more than 278,000 documents and clinical notes was used.
- RoBERTa$_{NER}$: fine-tuned for a NER task. This model is a refinement of RoBERTa$_{Bio}$, fine-tuned with the PharmaCoNER dataset (https://temu.bsc.es/pharmaconer (accessed on 13 July 2023)) and annotated with substance, protein, and compound entities.

2.4.2. Data Augmentation

Class imbalance is a common problem in this task. We propose the use of data augmentation techniques with the minority classes "Predicate" and "Reference" to mitigate it. We implemented two approaches to increase the number of samples: entity synonym generation and Back Translation (BT).

WordNet https://wordnet.princeton.edu/ (accessed on 17 July 2023) was used to generate synonyms. It is a lexical database that groups nouns, verbs, adjectives, and adverbs that are synonyms forming a synset, each of which expresses a different concept. In particular, we used the Open Multilingual Wordnet (https://github.com/globalwordnet/OMW (accessed on 17 July 2023)), which aims to facilitate the use of wordnets in multiple languages. For each entity of the classes we wanted to augment, we selected the most frequent synonym.

Back Translation [35] consists of translating the entities into another language, from Spanish to English in this case, and then translating them back into Spanish, assuming a high probability that some of the resulting entities are not exactly the same as the originals, but have the same meaning. For translating, we used the models provided by the Language Technology Research Group at the University of Helsinki, both for Spanish into English (https://huggingface.co/Helsinki-NLP/opus-mt-es-en (accessed on 17 July 2023)) and English into Spanish (https://huggingface.co/Helsinki-NLP/opus-mt-en-es (accessed on 17 July 2023)).

As a result of each type of augmentation process in training and development collections, the entity classes "Predicate" and "Reference" doubled in frequency: in the case of "Predicate" from 955 to 1910, and from 274 to 548 in the case of "Reference".

2.4.3. NER Models

We used the pre-trained models presented in Section 2.4.1 as base models, and fine-tuned them for the NER task using the training and development collections. These base models have been pre-trained using Masked Language Modeling (MLM), except the RoBERTa$_{NER}$ model. To fine-tune a base model, it receives as input a sequence of tokens, i.e., a sentence from the dataset, and returns a sequence of labels, where each label corresponds to each given input token. Thus, the architecture of the model is that of BERT, fine-tuned for the NER task. Figure 3 shows the pipeline of the proposed system, including a BT data augmentation step. The input words are a sample of the training data; the rest of the data presented are a simulation.

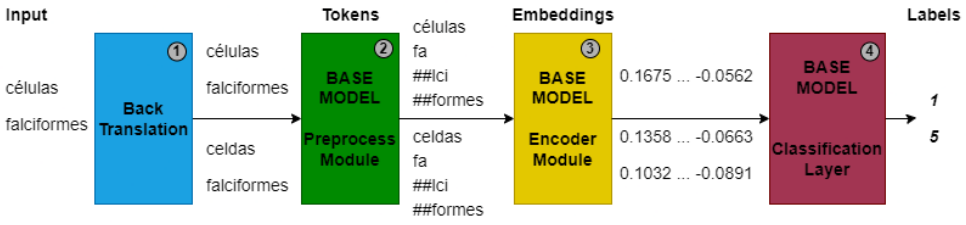


Figure 3. Pipeline of the proposed model.

First, the input data could be augmented by increasing the number of samples of the entities by BT (step 1 in Figure 3). Then, the input sentence is tokenized (step 2) using the WordPiece tokenization method [36]. This strategy tries to achieve a good balance between vocabulary size and out-of-vocabulary words. The algorithm segments the words into smaller parts and builds the vocabulary using the combination of these individual parts. Each of these parts or tokens will be converted to a 768-dimensional vector. Then, the input embedding is formed by a series of layers of embeddings (step 3), of which the output is the input of a classification layer (step 4). Specifically, the encoder module consists of 12 multi-head attention layers, where the self-attention mechanism is implemented, while the classification module is formed by a Feed Forward layer and a Softmax layer. Regarding

the hyperparameters, the models were trained with 25 and 40 epochs, a batch size of 64, and a learning rate of 2×10^{-5}. We used the Adam optimizer [37].

Text preprocessing techniques, such as stemming or lemmatization, as well as the removal of punctuation marks, were tried, but discarded because they did not improve the results. Finally, only non-alphanumeric characters and accents were removed. Adding an extra layer to process the POS tag of the entities was also tested, but the results did not improve.

Regarding computational resources, the training was performed on a private computer, with the following characteristics: CPU AMD Ryzen 9 5900HX, 16 GB RAM, ND 1T SSD M.2. No GPUs, cloud computing services or corporate servers were used. The training time with this configuration was between 4 and 5 h.

3. Results

Table 2 presents the results. The first column shows the pre-trained model and the number of epochs used in the fine-tuning. As, in the test collection, half of the documents are written in English, we show the overall results by language. The table is organized in three parts: the first corresponds to the original data sets, the second to the datasets augmented with WordNet, and the third part to the datasets augmented with Back Translation.

Table 2. Joint results by language of the fine-tuned transformer models: the first part corresponds to the original data sets, the second part to datasets augmented with WordNet, and the third part to those augmented with Back Translation, both augmentations being for the entity classes "Predicate" and "Reference". RoB$_{Cli}$ stands for RoBERTa$_{Clinical}$, RoB$_{Bio}$ stands for RoBERTa$_{Bio}$, and RoB$_{NER}$ stands for RoBERTa$_{NER}$. The best partial results are in bold and the best overall results are also underlined.

	ES + EN			ES			EN		
Model	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
RoB$_{Cli}$(25)	**0.736**	**<u>0.731</u>**	0.740	0.801	0.807	0.796	0.676	**<u>0.664</u>**	0.688
mBERT(40)	0.730	0.720	0.741	0.788	0.792	0.785	**<u>0.678</u>**	0.657	0.700
RoB$_{Cli}$(40)	0.728	0.727	0.728	0.798	0.806	0.791	0.663	0.657	0.670
mBERT(25)	0.727	0.706	**<u>0.750</u>**	0.792	0.785	**0.800**	0.670	0.640	**<u>0.702</u>**
RoB$_{Bio}$(40)	0.724	0.712	0.736	**0.803**	**0.808**	0.799	0.653	0.630	0.677
RoB$_{Bio}$(25)	0.723	0.710	0.736	0.783	0.789	0.778	0.669	0.641	0.698
RoB$_{NER}$(25)	0.718	0.716	0.719	0.792	0.798	0.786	0.650	0.643	0.658
RoB$_{NER}$(40)	0.716	0.711	0.721	0.786	0.790	0.783	0.653	0.642	0.664
BETO(25)	0.698	0.682	0.713	0.785	0.796	0.775	0.621	0.590	0.657
BETO(40)	0.693	0.674	0.712	0.784	0.790	0.780	0.612	0.580	0.650
	WordNet ES + EN			**WordNet ES**			**WordNet EN**		
Model	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
RoB$_{Bio}$(25)	**0.726**	**0.724**	**0.728**	0.790	0.805	0.775	**0.670**	**0.654**	0.684
RoB$_{Cli}$(25)	0.720	**0.724**	0.716	**0.795**	**0.815**	**0.776**	0.653	0.645	0.660
RoB$_{Cli}$(40)	0.720	0.716	0.724	0.783	0.807	0.760	0.665	0.641	**0.690**
mBERT(25)	0.718	0.717	0.718	0.787	0.810	0.765	0.657	0.640	0.674
mBERT(40)	0.711	0.714	0.707	0.773	0.803	0.746	0.656	0.641	0.671
RoB$_{Bio}$(40)	0.708	0.718	0.698	0.776	0.805	0.749	0.647	0.644	0.651
BETO(25)	0.696	0.692	0.699	0.778	0.802	0.756	0.623	0.602	0.645
BETO(40)	0.694	0.691	0.698	0.778	0.800	0.757	0.621	0.601	0.642
RoB$_{NER}$(40)	0.691	0.706	0.677	0.758	0.788	0.730	0.631	0.634	0.628
RoB$_{NER}$(25)	0.690	0.710	0.673	0.756	0.786	0.728	0.630	0.639	0.622

Table 2. *Cont.*

	Back Tr. ES + EN			Back Tr. ES			Back Tr. EN		
Model	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
RoB$_{Bio}$(40)	**0.739**	0.730	**0.750**	**0.812**	0.819	**0.804**	**0.674**	**0.653**	0.697
RoB$_{Bio}$(25)	0.738	**0.731**	0.747	**0.812**	**0.825**	0.799	0.673	0.651	0.697
mBERT(40)	0.723	0.717	0.730	0.779	0.795	0.763	**0.674**	0.651	**0.698**
RoB$_{Cli}$(40)	0.720	0.711	0.728	0.796	0.805	0.787	0.651	0.630	0.673
RoB$_{Cli}$(25)	0.717	0.726	0.708	0.791	0.814	0.769	0.650	0.648	0.651
mBERT(25)	0.715	0.713	0.716	0.797	0.815	0.780	0.641	0.626	0.656
RoB$_{NER}$(40)	0.707	0.712	0.701	0.796	0.805	0.787	0.651	0.630	0.673
BETO(40)	0.701	0.692	0.710	0.780	0.803	0.757	0.633	0.603	0.666
BETO(25)	0.697	0.681	0.713	0.780	0.805	0.758	0.626	0.587	0.670
RoB$_{NER}$(25)	0.696	0.706	0.687	0.765	0.795	0.738	0.634	0.630	0.638

Focusing on the first part, the results with the original corpus and regarding the overall results, RoBERTa$_{Clinical}$ obtains the best *F*1-Score, since it obtains good results in both Spanish and English. In all cases, the results in Spanish are better than the results in English. The RoBERTa$_{Bio}$ model stands out in Spanish, although the results with RoBERTa$_{Clinical}$ are very close. It is worth noting the good performance of BERT's multilingual model, mBERT, which obtains the best *F*1-Score in English and competitive results in Spanish. This may be due to the fact that, in the development and test collections, there are documents written in Spanish and English, since if we take into account only the results in Spanish, mBert's best performance is in the fifth position. Its better performance in English is what makes it rank second overall. It is remarkable that fine-tuning with only 50 English sentences from the development corpus allows it to recognize and classify English entities even starting from Spanish models.

Looking at the second part of Table 2, the results with data augmentation using Wordnet, it can be seen that this type of data augmentation does not improve the overall results of the task. Only one model, RoB$_{Bio}$(25), improved its *F*1-Score with respect to the first part of the table for Spanish entities.

Regarding the results after a data augmentation process by Back Translation, the third part of Table 2, the overall *F*1-Score slightly improves, with the best results being obtained with the RoBERTa$_{Bio}$ model, since this model seems to be the only one that takes advantage of the data augmentation in all cases and regardless of the number of epochs. The RoBERTa$_{Clinical}$ and RoBERTa$_{NER}$ models performs better without data augmentation, while mBERT only slightly improves the results of the Spanish documents and drops to third place overall.

Table 3 compares the results of our top three models with those of the top three participants in the eHealth-KD challenge. As can be seen, the best performing models are the domain-specific models RoBERTA$_{Bio}$ and RoBERTA$_{Clinical}$, especially those trained with data augmentation (with "-DA"). The number of epochs does not seem to be very significant. It is remarkable that all the models in the first part of Table 2, based on specific-domain models, improve the results of the challenge participants. Regarding the results with BT, just one specific-domain model, RoB$_{NER}$, and one generalist model, BETO, do not beat the best challenge participant.

Table 3. The results of our best models together with the best proposals submitted by eHealth participants.

Team/Model	***F*1-Score**	***Precision***	***Recall***
RoBERTa$_{Bio}$-BT40-BT	**0.739**	0.730	**0.750**
RoBERTa$_{Bio}$-BT25-BT	0.738	**0.731**	0.747
RoBERTa$_{Cli}$-25	0.736	**0.731**	0.740
PUCRJ-PUCPR-UFMG	0.706	0.715	0.697
Vicomtech	0.684	0.699	0.747
IXA	0.653	0.614	0.698

Error Analysis

The first part of Table 4 shows the evaluation metrics of the best model (RoBERTa$_{Bio}$-BT40-BT) corresponding to each type of entity shown with the IOB notation, while the second part shows the metrics for each class of entity, calculated by making a weighted average of the metrics for the IOB entity tags.

Table 4. Evaluation metrics for each entity shown with the IOB notation (first part) and calculated by the weighted average of the metrics for each class of entity (second part).

IOB Entity	*F*1-Score	*Precision*	*Recall*	Samples
B-Action	0.71	0.63	0.80	137
B-Concept	0.84	0.81	0.87	678
B-Predicate	0.44	0.65	0.34	98
B-Reference	0.35	0.43	0.30	10
I-Action	0.00	0.00	0.00	8
I-Concept	0.70	0.81	0.70	227
I-Predicate	0.00	0.00	0.00	20
Entity	***F*1-Score**	***Precision***	***Recall***	**Samples**
Action	0.67	0.59	0.75	137
Concept	0.80	0.81	0.83	678
Predicate	0.36	0.54	0.28	98
Reference	0.35	0.43	0.30	10
No entity	0.90	0.89	0.92	1110

Clearly, the model performs better in the case of components of the entities with a larger number of samples. On the other hand, in the cases with a smaller number of samples, such as I-Action and I-Predicate, no case is correct, but since the frequency is small, it hardly penalizes the overall results. The model is more accurate in recognizing the beginnings of the entities (B tag) than the rest of components (I tag), probably because there is a majority of one-word entities. In the case of class B-Reference, some predictions have been achieved despite the small sample size, and may be due to the fact that it is one of the classes to which Data Augmentation was applied. The I-Reference class does not appear in the test set.

4. Discussion

Most of our systems based on current and public domain-specific language models improve the results of the best eHealth challenge participants based on the use of generalist linguistic models. The best proposal so far was that in reference [22] based on mBERT, but with an end-to-end architecture that also extracts relationships between entities. Our way of using mBERT fine-tuned for the NER task obtains better results. The second [24] and third [26] best results so far were also based on generalist linguistic models but with a final classification layer using neural networks. These approaches have also been surpassed. All of our systems based on domain-specific models and fine-tuned with the original data sets outperform the best eHealth participant, indicating that they are a good choice even without performing any data augmentation.

Two of the domain-specific models, RoBERTa$_{Clinical}$ and RoBERTa$_{Bio}$ have beaten the other models in the two scenarios we evaluated: the as-is dataset, and the dataset enriched with data augmentation techniques.The main limitation of their use is that, although data augmentation by BT has improved the results of the NER task, the difficulty of recognizing certain entities that are very poorly represented in the datasets persists.

The only generalist model to compete with the domain-specific models was mBERT, the multilingual version of BERT, which achieved second place using the as-is corpus and third place using the augmented corpus, which may be due to its good performance in English. If we only consider the results with the Spanish sentences and the original datasets, the best configuration of mBERT moves from second to fifth place. mBERT has improved

the results of the Spanish generalist model BETO, even when only considering the Spanish part of the dataset. This indicates that a generalist model trained with huge amounts of text from different sources and languages can also obtain competitive results.

Interestingly, the domain-specific model, RoBERTa$_{NER}$, which was fine-tuned for a NER task, has shown the worst performance of the three. This may be due to the fact that the type of entities defined in the eHealth campaign framework do not fully correspond to those defined in other contexts. This is the only domain-specific model that performs worse than the generalist models when data augmentation is applied.

Of the two data augmentation strategies we have studied, only Back translation improves the results, especially with the RoBERTa$_{Bio}$ model, regardless of the number of epochs, which is shown to be the most suitable model for this task with this experimental framework. The improvement is due to the Spanish part, as the data augmentation does not in any way improve the results in the English part. With respect to the competence of the best models with the different types of entities, the most frequent types of entities are clearly better identified.

5. Conclusions

In this paper, we have presented a system for NER in Spanish medical texts using transformers, transfer learning, and data augmentation techniques. The results allow us to conclude that these techniques are suitable for the task. Our hypothesis that the use of domain-specific models fine-tuned to the Spanish NER task can improve performance has been proven to be true. We have used the experimental framework proposed in the last eHealth challenge, outperforming the best systems that participated using generalist models. The Back Translation data augmentation technique has slightly improved the results, which invites further research along this line. A possible line of future work could be prompting large language models to generate new data. In this work, we have not focused on changing the properties of the models (such as the optimizer, number of layers, and learning rate) nor the optimization of the parameters, so a future line of work would be to explore other possible configurations of the system pipeline. Other future lines to explore would be the use of transfer learning with biomedical texts in other languages, such as English, which could be used in conjunction with multilingual models such as mBERT. Moreover, the dataset provided by the eHealth challenge is small; it combines specific entities of the biomedical domain with other more general ones, so another next step could be to evaluate our proposals in other Spanish corpora with different characteristics and annotation guidelines.

Author Contributions: Conceptualization: A.V.; methodology: R.M. and S.M.; software: A.V.; validation: A.V.; formal analysis: all authors; data curation: A.V.; writing—original Draft: All authors; writing—review and editing, All authors.; supervision: R.M.; project administration: S.M.; funding acquisition, R.M. and S.M. All authors have read and agreed to the published version of manuscript.

Funding: This work was partially supported by the projects DOTT-HEALTH (PID2019-106942RB-C32, MCI/AEI/FEDER, UE); ISCIII (PI20/00715, co-funded by ERDF/ESF, "A way to make Europe"/"Investing in your future"); the project M2297 from call 2022 for impulse projects funded by Rey Juan Carlos University; and GELP (TED2021-130398B-C21, MCI/AEI/10.13039/501100011033 and NextGenerationEU/PRTR).

Data Availability Statement: We used the dataset provided for the organizers of the eHealth-KD Challenge 2021. This dataset is available at https://ehealthkd.github.io/2021/resources (accessed on 3 July 2023). Regarding the language models, the RoBERTA$_{Bio}$ model is available at https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es (accessed on 13 July 2023); the RoBERTA$_{Clinical}$ model is available at https://huggingface.co/plncmm/roberta-clinical-wl-es (accessed on 13 July 2023); the RoBERTA$_{NER}$ model is available at https://huggingface.co/julian-schelb/roberta-ner-multilingual (accessed on 13 July 2023); the BETO model is available at https://github.com/dccuchile/beto (accessed on 13 July 2023); and the mBERT model is available at https://huggingface.co/bert-base-multilingual-cased (accessed on 13 July 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
BT	Back Translation
CRF	Conditional Random Field
FNN	Feedforward Neural Network
IOB	Inside-Outside-Beginning
NE	Named Entity
NER	Named Entity Recognition
SNOMED-CT	Systematized Nomenclature of Medicine—Clinical Terms
UMLS	Unified Medical Language System

References

1. Li, J.; Sun, A.; Han, J.; Li, C. A Survey on Deep Learning for Named Entity Recognition. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 50–70. [CrossRef]
2. Bose, P.; Srinivasan, S.; Sleeman, W.C.; Palta, J.; Kapoor, R.; Ghosh, P. A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Appl. Sci.* **2021**, *11*, 8319. [CrossRef]
3. Li, J.; Wei, Q.; Ghiasv, O.; Chen, M.; Lobanov, V.; Weng, C.; Xu, H. A comparative study of pre-trained language models for named entity recognition in clinical trial eligibility criteria from multiple corpora. *BMC Med. Inform. Decis. Mak.* **2022**, *22* (Suppl. S3), 235. [CrossRef] [PubMed]
4. Miranda-Escalada, M.; 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. In Proceedings of the Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, Bologna, Italy, 5–8 September 2022.
5. Gasco Sánchez, L.; Estrada Zavala, D.; Farré-Maduell, D.; Lima-López, S.; Miranda-Escalada, A.; Krallinger, M. The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: Methods, evaluation, guidelines and corpora. In Proceedings of the Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, Gyeongju, Republic of Korea, 12–17 October 2022.
6. Fabregat, H.; Martínez-Romo, J.; Araujo, L. Overview of the DIANN Task: Disability Annotation Task. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages, Sevilla, Spain, 18 September 2018.
7. 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. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, Virtual, 19 November 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020.
8. Sánchez González, L. Biomedical Entities and Relations on Spanish Clinical Case Corpus: BERSCCC. Zenodo. 2022. Available online: https://zenodo.org/records/7193681 (accessed on 4 September 2023).
9. 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. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, Malaga, Spain, 23 September 2020.
10. Oronoz, M.; Casillas, A.; Gojenola, K.; Pérez, A. Automatic annotation of medical records in Spanish with disease, drug and substance names. In Proceedings of the Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, Havana, Cuba, 13–20 November 2013; pp. 536–543.
11. Carreras, X.; Chao, I.; Padró, L.; Padró, M. FreeLing: An Open-Source Suite of Language Analyzers. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, 26–28 May 2004.
12. Roller, R.; Kittner, M.; Weissenborn, D.; Leser, U. Cross-lingual Candidate Search for Biomedical Concept Normalization. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
13. Yuan, Z.; Zhao, Z.; Sun, H.; Li, J.; Wang, F.; Yu, S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J. Biomed. Inform.* **2022**, *126*, 103983. [CrossRef] [PubMed]
14. Perez-Miguel, N.; Cuadros, M.; Rigau, G. Biomedical term normalization of EHRs with UMLS. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
15. Pérez, N.; Accuosto, P.; Bravo, A.; Cuadros, M.; Martínez-García, E.; Saggion, H.; Rigau, G. Cross-lingual semantic annotation of biomedical literature: Experiments in Spanish and English. *Bioinformatics* **2019**, *36*, 1872–1880. [CrossRef] [PubMed]
16. Fabregat, H.; Duque, A.; Martinez-Romo, J.; Araujo, L. Negation-based transfer learning for improving biomedical Named Entity Recognition and Relation Extraction. *J. Biomed. Inform.* **2023**, *138*, 104279. [CrossRef] [PubMed]

17. Weegar, R.; Pérez, A.; Casillas, A.; Oronoz, M. Recent advances in Swedish and Spanish medical entity recognition in clinical texts using deep neural approaches. *BMC Med. Inform. Decis. Mak.* **2019**, *19* (Suppl. S7), 274. [CrossRef] [PubMed]
18. Akhtyamova, L. Named Entity Recognition in Spanish Biomedical Literature: Short Review and Bert Model. In Proceedings of the 26th Conference of Open Innovations Association FRUCT, Yaroslavl, Russia, 20–24 October 2020.
19. Xiong, Y.; Huang, Y.; Chen, Q.; Wang, X.; Nic, Y.; Tang, B. A Joint Model for Medical Named Entity Recognition and Normalization. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, Malaga, Spain, 23 September 2020; pp. 499–504.
20. García-Pablos, A.; Perez, N.; Cuadros, M. Vicomtech at CANTEMIST 2020. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, Malaga, Spain, 23 September 2020; pp. 489–498.
21. Piad-Morfis, A.; Estevez-Velarde, S.; Gutierrez, Y.; Almeida-Cruz, Y.; Montoyo, A.; Muñoz R. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021. *Proces. Del Leng. Nat.* **2021**, *67*, 233–242.
22. Pavanelli, L.; Terumi Rubel Schneider, E.; Bonescki Gumiel, Y.; Castro Ferreira, T.; Ferro Antunes de Oliveira, L.; Vitor Andrioli de Souza, J.; Paiva, G.P.M.; e Oliveira, L.E.S.; Moro, C.M.C.; Paraiso, E.C.; et al. PUCRJ-PUCPR-UFMG at eHealth-KD Challenge 2021: A Multilingual BERT-based System for Joint Entity Recognition and Relation Extraction. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), Malaga, Spain, 21 September 2021.
23. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Minneapolis, MI, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186.
24. García Pablos, A.; Pérez, N.; Cuadros M. Vicomtech at eHealth-KD Challenge 2021: Deep Learning Approaches to Model Health-related Text in Spanish. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), Malaga, Spain, 21 September 2021.
25. Otegi, A.; Agirre, A.; Campos, J.A.; Soroa, A.; Agirre, E. Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 436–442.
26. Andrés, E. IXA at eHealth-KD Challenge 2021: Generic Sequence Labelling as Relation Extraction Approach. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), Malaga, Spain, 21 September 2021.
27. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Virtual, 5–10 July 2020; pp. 8440–8451.
28. Ramshaw, L.A.; Marcus, M.P. Text chunking using transformation-based learning. In *Natural Language Processing Using Very Large Corpora*; Springer: Amsterdam, The Netherlands, 1999; pp. 157–176 .
29. Wang, L.L.; Lo, K.; Chandrasekhar, Y.; Reas, R.; Yang, J.; Burdick, D.; Eide, D.; Funk, K.; Katsis, Y.; Kinney, R.; et al. CORD-19: The COVID-19 Open Research Dataset. In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Virtual, 9–10 July 2020.
30. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
31. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.; Kang, H.; Pérez, J. Spanish pre-trained bert model and evaluation data. *arXiv* **2020**, arXiv:2308.02976.
32. Carrino, C.P.; Armengol-Estapé, J.; Gutiérrez-Fandiño, A.; Llop-Palao, J.; Pàmies, M.; Gonzalez-Agirre, A.; Villegas, M. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv* **2021**, arXiv:2109.03570.
33. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
34. Carrino, C.P.; Armengol-Estapé, J.; Bonet, O.; Gutiérrez-Fandiño, A.; Gonzalez-Agirre, A.; Krallinger, M.; Villegas, M. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models. *arXiv* **2021**, arXiv:2109.07765.
35. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding back-translation at scale. *arXiv* **2018**, arXiv:1808.09381.
36. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.