Identifying Cardiological Disorders in Spanish via Data Augmentation and Fine-Tuned Language Models

Antonio Romano[1,2], Giuseppe Riccio[1,2], Marco Postiglione[1] and Vincenzo Moscato[1,2]

[1] *University of Naples Federico II, Department of Electrical Engineering and Information Technology (DIETI), Via Claudio, 21 - 80125 - Naples, Italy*

[2] *Consorzio Interuniversitario Nazionale per l'Informatica (CINI) - ITEM National Lab, Complesso Universitario Monte S.Angelo, Naples, Italy*

Abstract

This study presents a novel approach to Biomedical Named Entity Recognition (BioNER), specifically tailored for the cardiology domain. The challenge of adapting models to specific fields is addressed through the integration of cross-domain transfer learning and data augmentation techniques. The process begins with the fine-tuning of a compact Biomedical Transformer model on the DisTEMIST corpus, enabling the capture of general biomedical concepts. This model is then further trained on the CardioCCC corpus, a cardiology-specific dataset, enhancing its ability to identify and interpret cardiological entities. A data augmentation strategy then is employed, leveraging Context Similarity and K-Nearest Neighbors (KNN) to generate augmented datasets. This enhances the model's ability to recognize medical entities. The final step involves a NER Fusion strategy, which combines outputs from multiple BioNER taggers to bolster robustness and accuracy in entity recognition. Experimental results from the MultiCardioNER challenge demonstrate the effectiveness of the proposed approach. Our framework surpasses the median F1 Score of 0.7566 by approximately 4%, achieving a score of 0.791, which is only 2% lower w.r.t. the top submission, despite being based on much smaller language models.

Keywords

Biomedical Named Entity Recognition, Data Augmentation, Language Models, EHRs

1. Introduction

In recent years, given the increasing volume of clinical data generated by medical personnel and the evolution of Artificial Intelligence (AI) models, it has become necessary to adopt techniques for the automatic extraction of medical concepts in order to support the development of personalized insights useful for patient health.

Specializing pre-trained BioNER models from general medical domains to specific fields like cardiology presents significant challenges due to limited specialized data availability, as highlighted by Nguyen et al. [1] and Chen et al. [2]. Transfer learning is a pivotal method for enhancing model performance in specific domains, as shown by Sasikumar and Mantri [3] and Zhou et al. [4], who adapted pre-trained biomedical models to specialized areas. Nevertheless, this approach is often insufficient due to the complexity of domain-specific texts.

Our approach attempts to address this problem by generating new data that increases the presence of less frequent medical entities by replacing them with similar medical entities. Therefore, for the identification of novel medical entities within the specified domain, it is necessary to establish a substitution strategy that, in contrast to other methodologies (Phan and Nguyen [5]; Ghosh et al. [6]), exploits the contextual similarity of the sentence in which the entity is to be augmented.

The proposed annotation methodology includes, in addition to data augmentation, a late fusion mechanism that leverages the use of various pre-trained models in the medical domain, similar to the

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

antonio.romano45@studenti.unina.it (A. Romano); giuseppe.riccio9@studenti.unina.it (G. Riccio); marco.postiglione@unina.it (M. Postiglione); vincenzo.moscato@unina.it (V. Moscato)

https://github.com/LaErre9 (A. Romano); https://github.com/giuseppericcio (G. Riccio); http://wpage.unina.it/vmoscato/ (V. Moscato)

0009-0000-5377-5051 (A. Romano); 0009-0002-8613-1126 (G. Riccio); 0000-0001-6092-940X (M. Postiglione); 0000-0002-0754-7696 (V. Moscato)

work proposed by Sun and Bhatia [7], and fine-tuned with cardiology data. This mechanism aims to improve the robustness and coverage of the generated annotations, as these models, trained on heterogeneous data, allow our system to recognize a greater number of medical entities through their combination.

We experimented our approach within the first track of the *MultiCardioNER*[1] [8] challenge, part of the *BioASQ 2024* [9] workshop. Specifically, we employed a diverse range of pre-trained models, each fine-tuned on combinations of the DisTEMIST dataset [10] as well as a new dataset (CardioCCC) of cardiology clinical cases annotated using the same guidelines. Our method surpassed the median *F1 Score* of 0.7566 by approximately 4% to achieve a score of 0.791. Interestingly, our score is close to the winning submission (only 2% lower) despite being based on much smaller transformer architectures.

The remainder of this paper is structured as follows: Section 2 discusses the existing literature related to our work; in Section 3 we outline the scope and objectives of our study; Section 4 presents the datasets used in our framework and their main characteristics. In Section 5, we detail our method, while experiments are presented in Section 6, discussing the results and their implications within the context of our research objectives. Finally, Section 7 summarizes the contributions of this paper and suggests avenues for future research.

2. Related Work

Adapting BioNER to specific medical domains, such as cardiology, presents significant challenges due to the complexity and variety of medical language.

Transfer learning has proven to be an essential method in the field of cross-domain BioNER for enhancing model resilience with respect to the medical concepts of specific domains. For example, Sasikumar and Mantri [3] leverage pre-trained models on biomedical corpora to adapt them to specific medical domains. Similarly, Zhou et al. [4] utilized transfer learning to leverage pre-trained features of general medicine models to improve the accuracy of specialized NER systems in clinical records.

However, despite the effectiveness of transfer learning, there are significant challenges. One of these is the adaptation of general clinical concept recognition systems to cardiology, a domain with unique complexity and specificity. Transfer learning alone may not be sufficient to address the challenges associated with domain-specific NER, due to the diversity and complexity of biomedical texts. To bridge this gap, our study proposes an approach that integrates transfer learning with Data Augmentation.

Therefore, extensive research has been conducted on strategies to increase text data in order to solve the issue of the lack of manually annotated data in specific medical domain (e.g. cardiological area). For example, Bartolini et al. [11] propose COSINER, which generates distinct increased data by using the contextual replacement of entities. Furthermore, Ghosh et al. [6] presented BioAug, which conditionally generates augmented data using the BART model to guarantee factual accuracy and diversity. Another approach for entity replacement proposed by Phan and Nguyen [5], creates new sentences by substituting entities with semantically equivalent ones using Gazetteer terms.

Following the cross-domain phase, to improve the robustness and coverage of the BioNER models, we utilize the merging of BioNER taggers. Sun and Bhatia [7] proposed merging the results to manage tag overlap and improve the complete concept extraction. Our approach, takes inspiration from the above-mentioned method, merging the results generated by different BioNER taggers to increase coverage, relying on these research results to ensure a more complete and accurate extraction of entities. In addition, merging taggers involves handling overlapping tags and conflicting results, ensuring that the final output is more precise and coherent.

Our contribution

In the proposed framework, we have adopted four pre-trained models. In particular, the basic models are those of Carrino et al. [12] and Carrino et al. [13], which have demonstrated excellent results in BioNER

[1] MultiCardioNER challenge website: https://temu.bsc.es/multicardioner/

in medical texts in Spanish. The other two models have been trained from the previous ones using wider medical datasets, such as those from Cohen et al. [14] and Llanos et al. [15], thus enriching knowledge of basic models and improving recognition of a greater number of medical entities. Subsequently, we implemented a phase of data augmentation on the CardioCCC cardiology-specific dataset in order to increase the number of medical concepts useful for cross-domain transfer learning, inspired by the entity replacement technique proposed by Phan and Nguyen [5]. Finally, our approach involves merging the predictions made by the different BioNER taggers, following an overlapping management strategy between the various annotations, inspired by the merging technique described by Sun and Bhatia [7].

3. Problem formulation

Starting from a dataset of annotated sentences denoted as $ANN_D = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathbf{X} \times \mathbf{Y}\}$, where:

- $\mathbf{X}$ is the collection of all sentences.
- $i \in \{1, \dots, N\}$, N representing the total number of sentences in the dataset and the i representing the i-th sentence.
- Each sentence $\mathbf{x}_i$ is a sequence of tokens $x_j \in \mathbf{x}_i$ where $j \in \{1, \dots, H_i\}$ and H_i is the length of the sentence.
- $\mathbf{Y}$ is the set of possible labels. We use the IOB2 annotation scheme [16], thus $\mathbf{Y} = \{B, I, O\}$, where *B* marks the beginning of an entity, *I* marks the inside of an entity, and *O* marks tokens outside any entity.
- $\mathbf{y}_i$ assigns each token $x_j \in \mathbf{x}_i$ to its corresponding label y_j.

The objective of the BioNER model is to precisely assign the appropriate tag from $\mathbf{Y}$ to each token within a given input sentence.

4. Materials

In our study, we utilized data provided by the MultiCardioNER challenge, which encompasses various clinical domain corpora, including DisTEMIST and the smaller CardioCCC corpus. Specifically, the DisTEMIST corpus underwent manual annotation by clinical experts, following specific guidelines for annotating diseases in Spanish clinical cases, as outlined in the work of Miranda-Escalada et al. [10]. These guidelines were meticulously developed by clinical experts through multiple cycles of quality control and consistency analysis before the entire dataset received annotations.

The training set for recognizing designated entities within the DisTEMIST corpus comprises 1000 recorded clinical cases. Simultaneously, a similar procedure was applied to 508 documents related to cardiological clinical cases within the Corpus CardioCCC.

To enhance the annotated dataset, we leveraged the DisTEMIST gazetteer, which contains key terms and synonyms for clinical entities. This tool significantly improves coverage of terminological and semantic variations in cardiological clinical texts using similarity-based approaches, thereby enhancing the quality and accuracy of annotations.

5. Methodology

Figure 1 shows an overview of the methodological flow of our solution for the MultiCardioNER track. Starting with the DisteMIST ($\mathcal{D}$) corpus training set, annotated according to the guidelines explained previously, it has been adequately *preprocessed* to build a new dataset containing the document identifier ID, the tokens x_j representing it, and the tags y_j associated with the token, using the properly mapped BIO scheme. The need to tokenize clinical text sentences $\mathbf{x}_i$ into single tokens x_j arises from the limited number of input samples accepted by each model used. The same process is applied to prepare CardioCCC ($\mathcal{C}$) for the next fine-tuning phase.

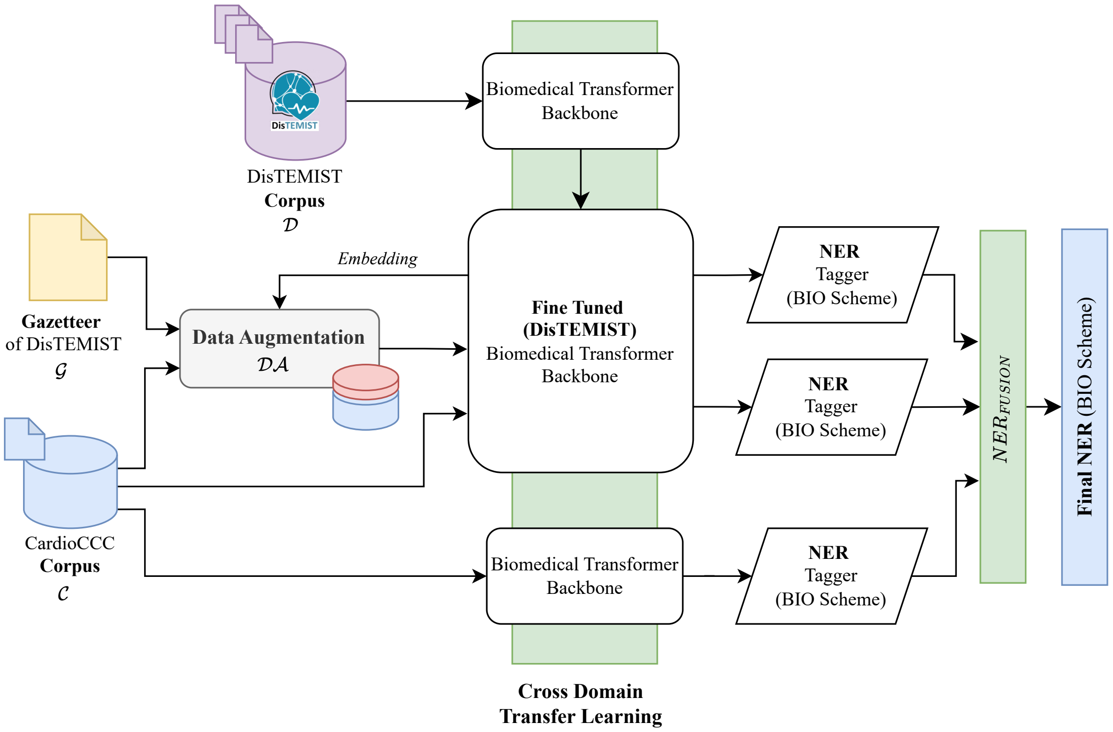


Figure 1: Overview of our NER solution for the MulticardioNER track. The proposed architecture employs *cross-domain transfer learning* to enhance disease recognition in cardiological domain. It begins with preprocessing the DisteMIST corpus data $\mathcal{D}$, followed by *fine-tuning a Biomedical Transformer Backbone* on various corpus. Subsequently, an innovative *Data Augmentation technique based on contextual similarity* CS is used to enrich the CardioCCC $\mathcal{C}$ corpus. The model undergoes further fine-tuning using augmented dataset $\mathcal{DA}$, ensuring improved accuracy and generalization capabilities in disease recognition. Finally, we merge the annotations generated by various models using a NER Fusion NER_{FUSION} strategy. This strategy effectively removes overlapping entities, prioritizing annotations based on a predefined scheme.

5.1. Cross-domain transfer learning

We propose an innovative cross-domain transfer learning solution to enhance disease recognition in cardiology. Our approach leverages a Biomedical Transformer Backbone, which is fine-tuned on various corpora provided by the challenge, to achieve superior predictive performance.

Initial Fine-Tuning on DisTEMIST We start by fine-tuning the Biomedical Transformer Backbone using the DisTEMIST corpus $\mathcal{D}$. This initial step tailors the model to understand the general biomedical language and disease entities present in this dataset. This corpus provides a broad foundation, enabling the model to capture essential biomedical concepts and terminologies.

Transfer learning on CardioCCC The fine-tuned model is then trained on the CardioCCC ($\mathcal{C}$) corpus to generate the first set of predictions. This step allows the model to adapt its understanding specifically to cardiological contexts and terminologies. Focusing on this specialized corpus ensures that the model can recognize and interpret data relevant to cardiology. In fact, it generates a second set of predictions. This ensures that the model integrates the cardiology-specific knowledge more deeply.

5.2. Data Augmentation with Context Similarity

Frequency Study Prior to implementing data augmentation, a systematic *Frequency Study* was performed to identify underrepresented entities and contexts within the CardioCCC dataset ($\mathcal{C}$). This

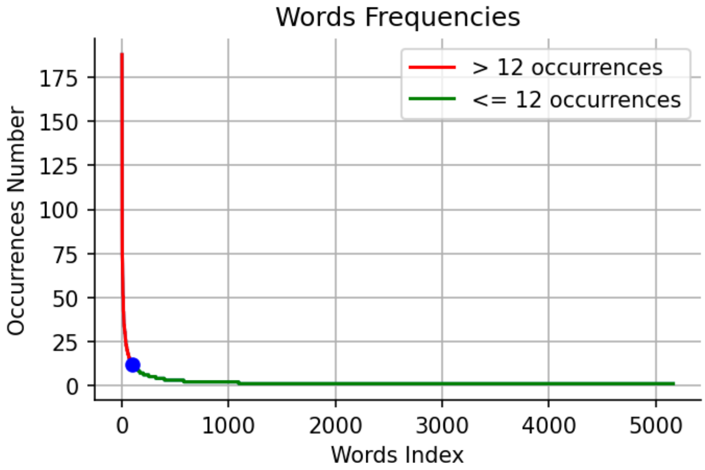


Figure 2: Frequency Study. In this analysis, we show the relative distribution of word occurrences. From this representation, it was possible to identify a threshold of 12, which roughly corresponds to the knee point of the curve. This allows for the exclusion of all words with occurrences greater than 12 (indicated by the red line), as they are sufficiently represented in the CardioCCC dataset ($\mathcal{C}$).

analysis involved examining the frequency of each medical mention in the CardioCCC dataset ($\mathcal{C}$). Following this analysis, a threshold was determined, corresponding approximately to the knee of the curve (Figure 2) that illustrates the distribution of word frequencies within the dataset. This threshold represents a balance point: words with frequencies above this threshold are sufficiently common and do not necessitate augmentation, while those with frequencies below are infrequent and can benefit from augmentation. This approach ensures that the augmentation process specifically targets these deficiencies, thereby optimizing the benefits of the additional data.

Through this analysis, we identified the entities that should be replaced to enhance the diversity and comprehensiveness of the CardioCCC dataset ($\mathcal{C}$). This enhancement improves the model's generalization capabilities, enabling it to recognize a broader spectrum of disease entities and ultimately achieve higher accuracy.

Entity Replacement using Context Similarity with KNN With the insights gained from the Frequency Study, we leverage the abundant information provided by the Gazetteer ($\mathcal{G}$) to fill gaps in the CardioCCC ($\mathcal{C}$) dataset, thereby enhancing its overall quality and utility for training the Biomedical Transformer Backbone fine-tuned using DisTEMIST ($\mathcal{D}$). To accomplish this objective, we employed a dataset augmentation technique utilizing Context Similarity with K-Nearest Neighbors (KNN) — K being set to $K = 1$ —, as illustrated in Figure 3.

This approach involves calculating the similarity between the embeddings of sentences in the CardioCCC dataset $Emb(\mathbf{x}_i)$ and the embeddings of entities in the Gazetteer $Emb(\mathbf{e}_i)$. By targeting sentences in CardioCCC ($\mathcal{C}$) annotated with B and I tags from the BIO Scheme, we identify the most contextually similar entities in the Gazetteer $\mathcal{G}$ and replace the original entities in the CardioCCC sentences $\mathbf{X}$ with these similar entities obtaining $\hat{\mathbf{X}}$.

To formalize Data Augmentation phase, we utilize a Gazetteer denoted as $\mathcal{G} = \{(\mathbf{e}_i, \hat{\mathbf{y}}) \in \mathbf{E} \times \hat{\mathbf{Y}}\}$, where $\mathbf{E}$ represents the collection of all entities in the Gazetteer, $\mathbf{e}_i$ is the i-th entity, with $i \in \{1, \ldots, N_E\}$, and $\hat{\mathbf{Y}} \in \{B, I\}$ represents the set of labels assigned to $\mathbf{e}_i$. Here, N_E denotes the total number of entities in the Gazetteer.

Subsequently, we employ Context Similarity (CS), computed using the K-Nearest Neighbors (KNN) function between the embeddings $Emb(\mathbf{x_i}) = \overline{\mathbf{x_i}}$ and $Emb(\mathbf{e_i}) = \overline{\mathbf{e_i}}$. The Context Similarity (CS) is defined as:

$$CS : KNN(\overline{\mathbf{x_i}}, \overline{\mathbf{e_i}}) \tag{1}$$

where CS represents the top-similar entities from $\mathcal{G}$ that are candidates for augment sentences. The

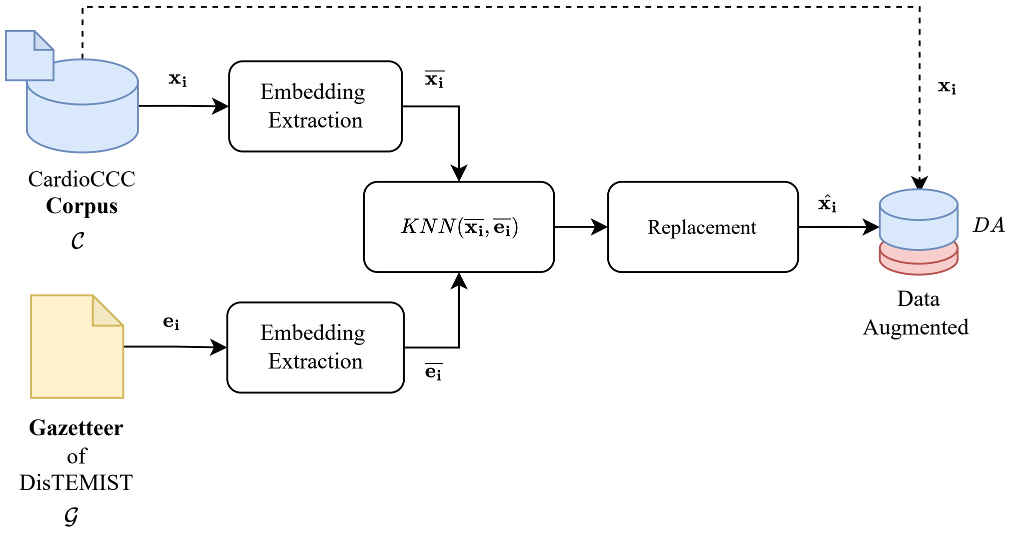


Figure 3: Overview of Data Augmentation. The embeddings for each sentence from CardioCCC ($Emb(\mathbf{x}_i)$) and the embeddings of entities in the Gazetteer of DisteMIST ($Emb(\mathbf{e}_i)$) are extracted, and the K-Nearest Neighbors (KNN) algorithm is used to calculate the similarities between these embeddings

augmented sentences $\hat{\mathbf{X}}$ are formulated as:

$$\hat{\mathbf{X}} : \{\hat{\mathbf{x}}_\mathbf{i} = CS(\mathbf{x_i}) \mid \forall \mathbf{x_i} \in \mathbf{X}\} \tag{2}$$

Consequently, the augmented dataset ($\mathcal{DA}$) is expressed as:

$$\mathcal{DA} : \{(\hat{\mathbf{x}}_\mathbf{i}, y_i) \cup (\mathbf{x_i}, y_i) \mid \forall \mathbf{x_i} \in \mathbf{X}, \forall \hat{\mathbf{x}}_\mathbf{i} \in \hat{\mathbf{X}}, \forall y_i \in \mathbf{Y}\} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ denote the original sets of sentences and their corresponding labels, respectively.

This method augments the dataset by merging both the original sentences ($\mathbf{X}$) and contextually similar sentences ($\hat{\mathbf{X}}$), as can be seen from the flow of the data augmentation process shown in Figure 4.

5.3. Transfer learning on CardioCCC Augmented Corpus

The Biomedical Transformer Backbone, developed on DisTEMIST ($\mathcal{D}$), is further trained on the CardioCCC Augmented ($\mathcal{DA}$) dataset. This final training phase allows the model to generate a third set of predictions, benefiting from the increased diversity and richness of the data.

Through this strategy, we are able to enhance the model's ability to generalize and recognize diseases more accurately. By initially fine-tuning on the DisTEMIST ($\mathcal{D}$) corpus, the model gains a broad understanding of biomedical language, which is essential for accurate disease recognition. Training on the enriched CardioCCC ($\mathcal{C}$) corpus further refines the model to focus on cardiological data, ensuring its predictions are contextually relevant and precise.

5.4. BioNER Fusion

To enhance the coverage of entities extracted from clinical notes, we merge the annotations generated by different Biomedical Transformer Backbones. However, during the BioNER Fusion NER_{FUSION} phase, it is essential to define merging strategies to handle any overlapping annotations. To achieve this, we establish a priority level based on the predictive performance of the models, allowing us to correctly select the annotation in case of conflicts. Initially, we perform a fusion operation to remove duplicate extracted entities that are entirely overlapping.

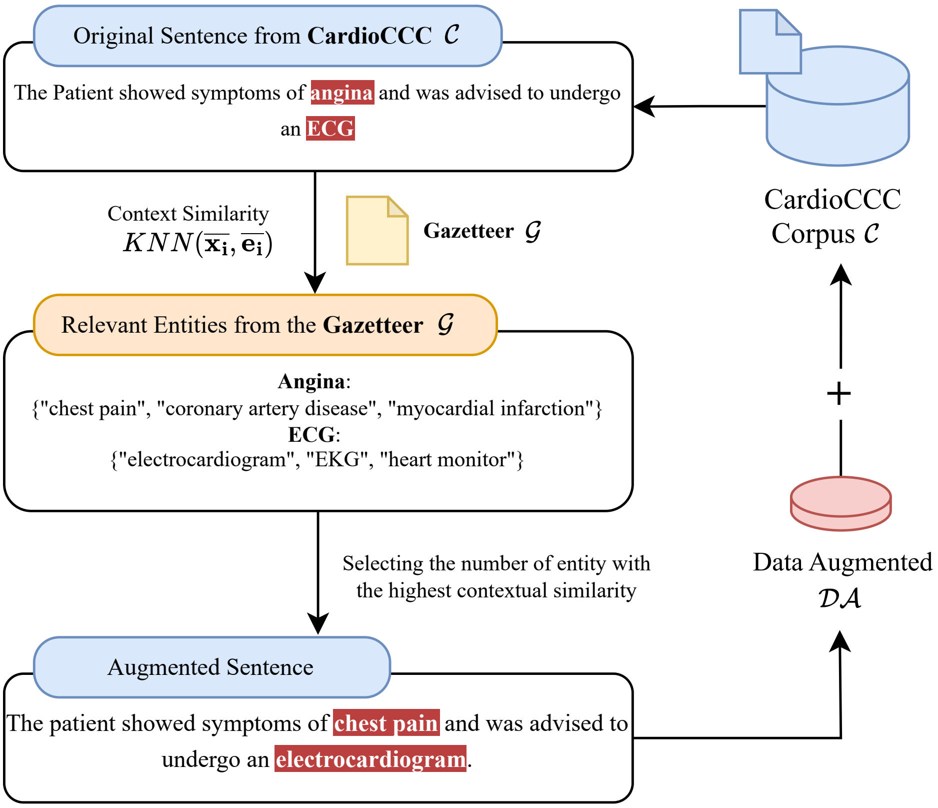

Original Sentence from CardioCCC $\mathcal{C}$
The Patient showed symptoms of angina and was advised to undergo an ECG
Context Similarity $KNN(\overline{\mathbf{x_i}}, \overline{\mathbf{e_i}})$
Gazetteer $\mathcal{G}$
Relevant Entities from the Gazetteer $\mathcal{G}$
Angina:
{"chest pain", "coronary artery disease", "myocardial infarction"}
ECG:
{"electrocardiogram", "EKG", "heart monitor"}
Selecting the number of entity with the highest contextual similarity
Augmented Sentence
The patient showed symptoms of chest pain and was advised to undergo an electrocardiogram.
CardioCCC Corpus $\mathcal{C}$
+
Data Augmented $\mathcal{DA}$

Figure 4: Example of Data Augmentation with Context Similarity CS. Starting from the *original sentence in CardioCCC* ($\mathbf{X}$), relevant entities from the Gazetteer are calculated to obtain the *augmented sentence* ($\hat{\mathbf{X}}$), which is then added to the initial dataset ($\mathcal{C}$).

For managing the Fusion of NER tagger generated by cross-domain transfer learning, we handle the overlapping with this function:

$$OL : \min(E_{x_i}, E_{x_{i'}}) - \max(S_{x_i}, S_{x_{i'}}) \geq 0 \tag{4}$$

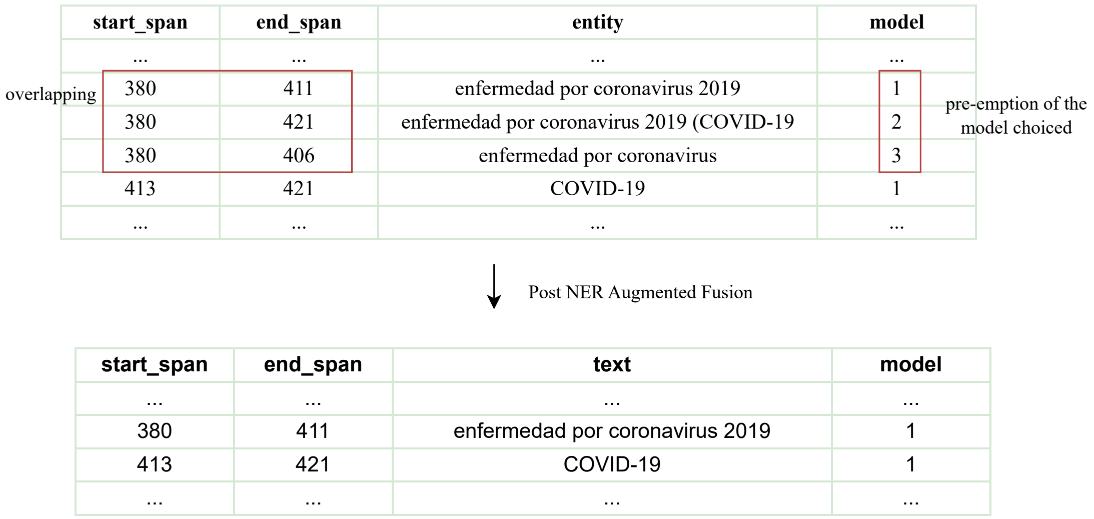
where E_{x_i} and S_{x_i} represent the End Span and Start Span of the x_i entity by the first model, while $E_{x_{i'}}$ and $S_{x_{i'}}$ refer to the second model.

Sometimes, the overlap is not complete, but the “start span” and “end span” of one entity partially coincide with those of another entity (even if not identical), resulting in $OL > 0$.

In essence, the strategy of BioNER Fusion NER_{FUSION} is defined as:

$$NER_{FUSION} = \begin{cases} x_{IMP} & : OL \geq 0 \\ (x_i, x_{i'}) & : OL < 0 \end{cases} \tag{5}$$

In such scenarios, overlapping is resolved by prioritizing the entity with the higher priority level x_{IMP}. The priority scheme, assigned to the models, is fixed according to the performance of the models observed on the internal test set. For example, the model with priority level 1 has a higher priority than the model with priority level 2. This approach ensures that all extracted entities are retained in the final clinical note, thereby enhancing the overall entity extraction process.



start_span	end_span	entity	model
...	...	...	...
380	411	enfermedad por coronavirus 2019	1
380	421	enfermedad por coronavirus 2019 (COVID-19	2
380	406	enfermedad por coronavirus	3
413	421	COVID-19	1
...	...	...	...

start_span	end_span	text	model
...	...	...	...
380	411	enfermedad por coronavirus 2019	1
413	421	COVID-19	1
...	...	...	...

Figure 5: Example of NER Fusion NER_{FUSION}. In this example, there is an overlap between x_i, x'_i and, x''_i following our BioNER Fusion strategy, the higher priority model prediction (*model 1*) is chosen as the final annotation.

6. Experiments

The performance of the proposed approaches for BioNER was assessed by participating in the MultiCardioNER Shared Task[2] as part of the *BioASQ 2024* challenge. This section presents the results of our methodology on the final test set, along with preliminary experiments conducted on the training corpus provided by the challenge organizers.

6.1. Experimental Setup

6.1.1. Evaluation Metrics

The evaluation is performed by comparing the automatically generated results with those produced by the manual annotations of experts. The primary evaluation metrics for track 1 include **micro-averaged precision (MiP)**, **recall (MiR)**, and the **F1 score (MiF1)**. For the evaluation of the results, the library[3] realized by the organizers was used.

6.1.2. Configuration

The BioNER system was implemented using the `HuggingFace Transformers library` (v4.40.2) by exploiting the various Spanish Transformer biomedical networks in the repository. In the Table 1 are shown those selected for the experiment.

These models were chosen not only for their effectiveness but also for their relatively small size, making them executable even on less powerful hardware and thus suitable for low-resource environments. We fine-tuned our models in a Google Colab environment, which provided us with a Tesla T4 GPU.

In the phase prior to our submission, we studied the effects of various hyperparameters and the generalization error of our models by dividing the original corpus of clinical cases into three parts: (1) a

[2] MultiCardioNER challenge website: https://temu.bsc.es/multicardioner/
[3] MultiCardioNER Evaluation Library: https://github.com/nlp4bia-bsc/multicardioner_evaluation_library
[4] https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es
[5] https://huggingface.co/lcampillos/roberta-es-clinical-trials-ner
[6] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es
[7] https://huggingface.co/StivenLancheros/roberta-base-biomedical-clinical-es-finetuned-ner-CRAFT

Table 1
Spanish RoBERTa-based models for clinical biomedical language used for experimentation

Model Name	Size	Document types	Biomedical Corpora	Year
r-b-bio-clinical-es[4] [12]	504 MB	Clinical Docs and Reports	Medical Crawler [17], Clinical Cases, Misc Scielo [18], BARR2[19], Wikipedia Life Sciences, Patents, EMEA, mespen_Medline [20], Pubmed	2021
r-es-clinical-trials-ner[5] [15]	496 MB	Clinical Trials, Studies	CT-EBM-SP corpus [21]	2021
bsc-bio-ehr-es[6] [13]	499 MB	Clinical Docs and Reports	Medical Crawler [17], Clinical Cases, Misc Scielo [18], BARR2 [19], Wikipedia Life Sciences, Patents, EMEA, mespen_Medline [20], Pubmed	2022
r-b-bio-clinical-es-FT[7] [12]	502 MB	Clinical Docs and Reports	The Colorado Richly Annotated Full-Text (CRAFT) Corpus [14]	2022

training set (60% of the original corpus) used to train the model, (2) a validation set (20% of the original corpus) to evaluate the effects of the hyperparameters, and (3) a test set (20% of the original corpus) to assess the models' ability to generalize to unseen data (Internal Test Set Results).

6.2. Results

To conduct the experiments, we initially analyzed how variations in hyperparameters influenced our validation set. Firstly, we adjusted the batch size, determining that an optimal size was 4. Subsequently, we tuned the learning rate, discovering that 8e-5 yielded the best results. Finally, we modified the weight decay, concluding that a value of 0.2 was optimal. After identifying these hyperparameters, we increased the training epochs and implemented an early stopping criterion to halt training if performance on the validation set did not improve for five consecutive epochs. Further details on hyperparameters tuning are provided in the Appendix A.

6.2.1. Cross-domain BioNER Evaluation

To construct the optimal cross-domain transfer learning model, we conduct an *Internal Test* to select the best combination of pre-trained models used at various layers of the cross-domain process as shown in Table 2.

The best results are shown in bold and were selected in the *MultiCardioNER Test Results* released by the challenge organizer, as shown in Table 3.

The model *bsc-bio-ehr-es*, trained on CardioCCC ($\mathcal{C}$), achieved the best results in the external test with an F1 score of 0.7924, due to the specificity of the CardioCCC dataset, which includes terminology and concepts related to cardiology. In comparison, models pre-trained on DisTEMIST ($\mathcal{D}$) and both the combination of DisTEMIST and CardioCCC ($\mathcal{D} \cup \mathcal{C}$), as well as DisTEMIST and CardioCCC Augmented ($\mathcal{D} \cup \mathcal{DA}$) showed lower performance. This is attributed to the more general nature of the DisTEMIST dataset, which fails to effectively capture the specialized cardiology terms present in the test sets. Therefore, we are also considering the proposed model fusion approach.

Table 2
Internal Test Results. Results were obtained by splitting the CarcioCCC data into 80% training and 20% test datasets. The 'Dataset' column represents the dataset on which the pre-trained model was trained.

Dataset	Pre-trained Model	MiP	MiR	MiF1	Accuracy
DisTEMIST	bsc-bio-ehr-es	0.6871	**0.7313**	0.7085	**0.9628**
DisTEMIST	r-b-bio-clinical-es	0.6845	0.7203	0.7020	0.9598
DisTEMIST	r-b-bio-clinical-es-FT	0.6962	0.7186	0.7073	0.9610
DisTEMIST	**r-es-clinical-trials-ner**	**0.7023**	0.7284	**0.7152**	0.9615
CardioCCC	**bsc-bio-ehr-es**	0.8000	**0.8122**	**0.8061**	**0.9753**
CardioCCC	r-base-biomedical-clinical-es	0.7943	0.8089	0.8016	0.9731
CardioCCC	r-b-bio-clinical-es-FT	**0.8014**	0.8098	0.8056	0.9741
CardioCCC	r-es-clinical-trials-ner	0.7967	0.8052	0.8009	0.9752
DisTEMIST + CardioCCC	bsc-bio-ehr-es	**0.8082**	0.8069	0.8075	**0.9754**
DisTEMIST + CardioCCC	r-base-biomedical-clinical-es	0.7940	0.8152	0.8045	0.9743
DisTEMIST + CardioCCC	r-b-bio-clinical-es-FT	0.7959	0.8102	0.8030	0.9735
DisTEMIST + CardioCCC	**r-es-clinical-trials-ner**	0.7983	**0.8184**	**0.8082**	0.9751
DisTEMIST + CardioCCC Augmented	bsc-bio-ehr-es	0.8951	0.9047	0.8998	0.9886
DisTEMIST + CardioCCC Augmented	r-base-biomedical-clinical-es	0.8911	0.8930	0.8920	0.9875
DisTEMIST + CardioCCC Augmented	r-b-bio-clinical-es-FT	0.9012	**0.9080**	0.9046	0.9880
DisTEMIST + CardioCCC Augmented	**r-es-clinical-trials-ner**	**0.9125**	0.9038	**0.9081**	**0.9888**

Table 3
MultiCardioNER Test Results. Results obtained on the official test data. The 'Dataset' column represents the dataset on which the pre-trained model was trained and comparison with results of the Challenge

Combination	Dataset	Pre-trained Model	MiP	MiR	MiF1
$\phi_{\mathcal{D}}$	DisTEMIST	r-es-clinical-trials-ner	0.7184	0.7275	0.723
$\phi_{\mathcal{C}}$	**CardioCCC**	**bsc-bio-ehr-es**	**0.8046**	**0.7804**	**0.7924**
$\phi_{\mathcal{D}\cup\mathcal{C}}$	DisTEMIST + CardioCCC	r-es-clinical-trials-ner	0.7784	0.7744	0.7764
$\phi_{\mathcal{D}\cup\mathcal{D}\mathcal{A}}$	DisTEMIST + CardioCCC Augmented	r-es-clinical-trials-ner	0.7784	0.7744	0.7764
-	**Median of the Challenge**	-	-	-	0.7566
-	**Average of the Challenge**	-	-	-	0.6772
-	**Highest Precision**	-	0.8919	-	-
-	**Highest Recall**	-	-	0.8243	-
-	**Highest F1**	-	-		0.8199

6.2.2. BioNER Fusion Evaluation

We evaluate the impact of Fusion applied to the best combinations set previously. In Table 4, it is evident that the most promising results stem from the top submission presented during the competition.

Specifically, the BioNER Fusion (NER_{FUSION}) performed on the combination of CardioCCC and *bsc-bio-ehr-es* model ($\phi_{\mathcal{C}}$), DisTEMIST + CardioCCC with *r-es-clinical-trials-ner* ($\phi_{\mathcal{D}\cup\mathcal{C}}$), and DisTEMIST + CardioCCC Augmented with *r-es-clinical-trials-ner* ($\phi_{\mathcal{D}\cup\mathcal{D}\mathcal{A}}$) exhibited superior predictive performance. This integration, facilitated through our fusion strategy, yielded an enhancement in *Recall (MiR)*, showcasing the system's heightened ability to accurately identify relevant entities. This outcome implies that fusion enabled the system to offset individual model deficiencies, thereby contributing to an overall improvement in entity extraction effectiveness. Therefore, the fusion coupled with the application of *String Matching Cutter* exhibits *high Precision* but *low Recall*, indicating a conservative tendency of the system to recognize only highly probable entities. Conversely, the integration of *String Matching Adder* with the fusion is characterized by greater inclusivity, even if at the expense of *lower Precision.* In conclusion, examining the overall results of the challenge reveals that the leading models (e.g. *mdeberta, XLM-RoBERTa, CLIN-X-ES, ...*) utilized are at least 3-4 times larger than those employed in our approach. Despite this, the best result achieved by our system nearly matched the performance of the top models, with an F1 score of 0.791 compared to approximately 0.82. This demonstrates that our approach is not only effective but also more efficient in terms of computational resources, making it ideal for practical implementations with hardware constraints. Additionally, the fusion of models from various datasets (CardioCCC, DisTEMIST, and augmented datasets) has demonstrated

Table 4
BioNER Fusion results. "($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{DA}}$)" represents the combination of the models trained on CardioCCC ($\mathcal{C}$), DisTEMIST ($\mathcal{D}$) and DisTEMIST + CardioCCC Augmented ($\mathcal{D} \cup \mathcal{DA}$); "($\phi_{\mathcal{DA}} \times 4$)" represents the combination of the 4 models used during the experiments and then fine-tuned on Augmented Dataset ($\mathcal{DA}$); "+ *String Matching Cutter*" and "+ *String Matching Adder*" represent the application of String Matching using the Gazetteer. Specifically, *String Matching Cutter* removes entities not present in the Gazetteer, while *String Matching Adder* merges entities present in the Gazetteer that were not extracted by the model combination.

Fusion		MiP	MiR	MiF1
NER_{FUSION} ($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{DA}}$)		**0.7794**	**0.803**	**0.791**
NER_{FUSION} ($\phi_{\mathcal{DA}} \times 4$)		0.7346	0.7799	0.7566
NER_{FUSION} ($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{DA}}$)	+ String Matching Cutter	0.8919	0.4897	0.6323
NER_{FUSION} ($\phi_{\mathcal{DA}} \times 4$)	+ String Matching Cutter	0.8886	0.4744	0.6185
NER_{FUSION} ($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{DA}}$)	+ String Matching Adder	0.7079	0.7775	0.7411
NER_{FUSION} ($\phi_{\mathcal{DA}} \times 4$)	+ String Matching Adder	0.6568	0.6244	0.6402
Median of the Challenge		-	-	0.7566
Average of the Challenge		-	-	0.6772
Highest Precision		0.8919	-	-
Highest Recall		-	0.8243	-
Winner F1		-	-	0.8199

Table 5
Error Analysis. Error Analysis conducted on several BioNER Fusion strategy (*Note*: Total concepts in Ground Truth: 7884)

Fusion		N. Extr. Ent.	Correct	CFP	CFN	RLOS
NER_{FUSION} ($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{DA}}$)		8123	6331 (80,30%)	684	398	1122
NER_{FUSION} ($\phi_{\mathcal{DA}} \times 4$)		8370	6149 (77,99%)	934	294	1297
NER_{FUSION} ($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{DA}}$)	+ String Matching Cutter	4329	3861 (48,97%)	206	3745	271
NER_{FUSION} ($\phi_{\mathcal{DA}} \times 4$)	+ String Matching Cutter	4209	3740 (47,43%)	229	3886	244
NER_{FUSION} ($\phi_{\mathcal{C}}, \phi_{\mathcal{D}\cup\mathcal{C}}, \phi_{\mathcal{DA}}$)	+ String Matching Adder	8660	6350 (80,54%)	1187	367	1137
NER_{FUSION} ($\phi_{\mathcal{DA}} \times 4$)	+ String Matching Adder	7495	5146(65,27%)	1044	1245	1315

the system's capability to integrate and balance information from diverse sources, thereby enhancing overall performance and flexibility.

6.3. Error Analysis

Inspired by Moscato et al. [22], we conducted a detailed error analysis of the differences among the models by examining the number of correctly retrieved entity mentions (Correct). These errors can be categorized into three possible distinct types:

- **Complete False Positive (CFP)**: The model identifies an entity that is not annotated as a named entity.
- **Complete False Negative (CFN)**: The model fails to identify an entity that is annotated as a named entity.
- **Right Label Overlapping Span (RLOS)**: The model correctly identifies the presence of an annotated named entity, but the span of the entity is incorrect.

This categorization allowed us to better understand the strengths and weaknesses of our system. The results are shown in Table 5.

The error analysis corroborates previous evaluations, indicating that the fusion combined with string matching significantly reduces the number of false positives but drastically increases the number of false negatives, as it extracts only half of the relevant entities. The possible causes of these results may lie in the quality and size of the Gazetteer used. The best balance between precision and recall, which most effectively satisfies this analysis, is once again achieved by NER_{FUSION} ($\phi_{\mathcal{C}}$, $\phi_{\mathcal{D}\cup\mathcal{C}}$, $\phi_{\mathcal{D}\mathcal{A}}$).

7. Conclusion

In this study, we presented an innovative approach to address the challenge of BioNER Fusion in the biomedical domain, with a particular focus on cardiology. Our methodology integrates data augmentation techniques and data fusion mechanisms to enhance the robustness and coverage of the generated annotations. By utilizing pre-trained models on biomedical corpora and refining them with domain-specific cardiology data, we achieved significant results, overcoming the limitations related to the scarce availability of domain-specific data.

However, there are potential disadvantages to our approach. Data augmentation techniques, while increasing the diversity of the training data, might also introduce noise and potentially irrelevant information, which could hinder the model's performance. Additionally, the complexity of integrating multiple models through data fusion can increase computational requirements and may pose challenges in real-time applications.

The results obtained in the MultiCardioNER competition, part of the *BioASQ 2024* challenge, demonstrate the effectiveness of our approach. The key characteristics of our results include their efficacy, computational efficiency, domain adaptation, flexibility, balance between precision and recall, robustness, and innovativeness. These combined elements illustrate how our approach can be a valid and practical solution for entity extraction from biomedical texts, especially in contexts with limited computational resources. We exceeded the median F1 score by 4%, achieving a score of 0.791. This success highlights the potential of the proposed techniques in addressing BioNER challenges in specific biomedical contexts, paving the way for further improvements and applications in various clinical fields.

Acknowledgements

We acknowledge financial support from (1) the PNRR MUR project PE0000013-FAIR and (2) the Italian ministry of economic development, via the ICARUS (Intelligent Contract Automation for Rethinking User Services) project (CUP: B69J23000270005).

References

[1] N. D. Nguyen, L. Du, W. L. Buntine, C. Chen, R. Beare, Hardness-guided domain adaptation to recognise biomedical named entities under low-resource scenarios, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, 2022, pp. 4063–4071. URL: https://doi.org/10.18653/v1/2022.emnlp-main.271. doi:10.18653/V1/2022.EMNLP-MAIN.271.

[2] S. Chen, G. Aguilar, L. Neves, T. Solorio, Data augmentation for cross-domain named entity recognition, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 5346–5356. URL: https://doi.org/10.18653/v1/2021.emnlp-main.434. doi:10.18653/V1/2021.EMNLP-MAIN.434.

[3] N. Sasikumar, K. S. I. Mantri, Transfer learning for low-resource clinical named entity recognition, in: T. Naumann, A. B. Abacha, S. Bethard, K. Roberts, A. Rumshisky (Eds.), Proceedings of the 5th Clinical Natural Language Processing Workshop, ClinicalNLP@ACL 2023, Toronto, Canada, July 14, 2023, Association for Computational Linguistics, 2023, pp. 514–518. URL: https://doi.org/10.18653/v1/2023.clinicalnlp-1.53. doi:10.18653/V1/2023.CLINICALNLP-1.53.

[4] M. Zhou, J. Tan, S. Yang, H. Wang, L. Wang, Z. Xiao, Ensemble transfer learning on augmented domain resources for oncological named entity recognition in chinese clinical records, IEEE Access 11 (2023) 80416–80428. URL: https://doi.org/10.1109/ACCESS.2023.3299824. doi:10.1109/ACCESS.2023.3299824.

[5] U. Phan, N. Nguyen, Simple semantic-based data augmentation for named entity recognition in biomedical texts, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022, Dublin, Ireland, May 26, 2022, Association for Computational Linguistics, 2022, pp. 123–129. URL: https://doi.org/10.18653/v1/2022.bionlp-1.12. doi:10.18653/V1/2022.BIONLP-1.12.

[6] S. Ghosh, U. Tyagi, S. Kumar, D. Manocha, Bioaug: Conditional generation based data augmentation for low-resource biomedical NER, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 1853–1858. URL: https://doi.org/10.1145/3539618.3591957. doi:10.1145/3539618.3591957.

[7] Q. Sun, P. Bhatia, Neural entity recognition with gazetteer based fusion, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, Association for Computational Linguistics, 2021, pp. 3291–3295. URL: https://doi.org/10.18653/v1/2021.findings-acl.291. doi:10.18653/V1/2021.FINDINGS-ACL.291.

[8] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[9] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[10] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 179–203. URL: https://ceur-ws.org/Vol-3180/paper-11.pdf.

[11] I. Bartolini, V. Moscato, M. Postiglione, G. Sperlì, A. Vignali, COSINER: context similarity data augmentation for named entity recognition, in: T. Skopal, F. Falchi, J. Lokoc, M. L. Sapino, I. Bartolini, M. Patella (Eds.), Similarity Search and Applications - 15th International Conference, SISAP 2022, Bologna, Italy, October 5-7, 2022, Proceedings, volume 13590 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 11–24. URL: https://doi.org/10.1007/978-3-031-17849-8_2. doi:10.1007/978-3-031-17849-8_2.

[12] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, CoRR abs/2109.03570 (2021). URL: https://arxiv.org/abs/2109.03570. arXiv:2109.03570.

[13] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in spanish, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings

of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022, Dublin, Ireland, May 26, 2022, Association for Computational Linguistics, 2022, pp. 193–199. URL: https://doi.org/10.18653/v1/2022.bionlp-1.19. doi:10.18653/V1/2022.BIONLP-1.19.

[14] K. B. Cohen, A. Lanfranchi, M. J. Choi, M. Bada, W. A. B. Jr., N. Panteleyeva, K. Verspoor, M. Palmer, L. E. Hunter, Coreference annotation and resolution in the colorado richly annotated full text (CRAFT) corpus of biomedical journal articles, BMC Bioinform. 18 (2017) 372:1–372:14. URL: https://doi.org/10.1186/s12859-017-1775-9. doi:10.1186/S12859-017-1775-9.

[15] L. C. Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, A. Moreno-Sandoval, A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine, BMC Medical Informatics Decis. Mak. 21 (2021) 69. URL: https://doi.org/10.1186/s12911-021-01395-z. doi:10.1186/S12911-021-01395-Z.

[16] L. A. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: D. Yarowsky, K. Church (Eds.), Third Workshop on Very Large Corpora, VLC@ACL 1995, Cambridge, Massachusetts, USA, June 30, 1995, 1995, pp. 82–94. URL: https://aclanthology.org/W95-0107/.

[17] C. P. Carrino, J. Silveira-Ocampo, A. Gonzalez-Agirre, A. Gutiérrez-Fandiño, M. Krallinger, M. Villegas, Spanish biomedical crawled corpus, 2022. URL: https://doi.org/10.5281/zenodo.5513237. doi:10.5281/zenodo.5513237.

[18] A. Intxaurrondo, Scielo-spain-crawler, 2019. URL: https://doi.org/10.5281/zenodo.2541681. doi:10.5281/zenodo.2541681.

[19] A. Intxaurrondo, M. Pérez-Pérez, G. P. Rodríguez, J. A. López-Martín, J. Santamaría, S. de la Peña, M. Villegas, S. A. Akhondi, A. Valencia, A. Lourenço, M. Krallinger, The biomedical abbreviation recognition and resolution (BARR) track: Benchmarking, evaluation and importance of abbreviation recognition systems applied to spanish biomedical abstracts, in: R. Martínez, J. Gonzalo, P. Rosso, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017, volume 1881 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017, pp. 230–246. URL: https://ceur-ws.org/Vol-1881/Overview1.pdf.

[20] M. Villegas, A. Intxaurrondo, A. Gonzalez-Agirre, M. Krallinger, Mespen_parallel-corpora, 2019. URL: https://doi.org/10.5281/zenodo.3562536. doi:10.5281/zenodo.3562536.

[21] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, A. Moreno-Sandoval, CT-EBM-SP - Corpus of Clinical Trials for Evidence-Based-Medicine in Spanish, 2022. URL: https://doi.org/10.1186/s12911-021-01395-z. doi:10.1186/s12911-021-01395-z.

[22] V. Moscato, M. Postiglione, C. Sansone, G. Sperlí, Taughtnet: Learning multi-task biomedical named entity recognition from single-task teachers, IEEE J. Biomed. Health Informatics 27 (2023) 2512–2523. URL: https://doi.org/10.1109/JBHI.2023.3244044. doi:10.1109/JBHI.2023.3244044.

A. Hyperparameters Tuning

We analyzed how variations in hyperparameters influence our validation set. Specifically, we have experimented each model's batch size, learning rate, and weight decay gradually to examine how well it performed in terms of precision, recall, and F1 score. We changed the batch size by first choosing from among potential candidates, and then we selected the value that corresponded to the best performance. We then experimented with various learning rates after fixing the batch size value; also in this case, we selected the value that yielded the highest scores. After setting the learning rate and batch size, we examined a small variation in the rate of weight decay and determined the ideal value based on earlier logic.

Batch size During training, we varied the *batch size*, initially set at 16, and then adjusted it to 8, 4, and 2. The results, as shown in the table 6 indicate that the optimal batch size is 4.

Table 6
NER hyperparameter selection - varying **batch size**

Models	Batch Size	MiP	MiR	MiF1
r-b-bio-clinical-es	16	0.7599	0.8003	0.7796
r-b-bio-clinical-es	8	0.7676	**0.8047**	0.7857
r-b-bio-clinical-es	4	**0.7899**	0.7963	**0.7931**
r-b-bio-clinical-es-FT	8	0.7690	0.7988	0.7836
r-b-bio-clinical-es-FT	16	0.7707	**0.7992**	0.7847
r-b-bio-clinical-es-FT	4	**0.7798**	0.7970	**0.7883**
bsc-bio-ehr-es	8	0.7534	0.7975	0.7749
bsc-bio-ehr-es	16	0.7514	0.7932	0.7717
bsc-bio-ehr-es	4	**0.7865**	**0.7979**	**0.7922**
r-es-clinical-trials-ner	8	0.7609	**0.7997**	0.7798
r-es-clinical-trials-ner	16	0.7602	0.7975	0.7784
r-es-clinical-trials-ner	4	**0.7877**	0.7979	**0.7927**

Learning rate After determining the optimal batch size, we varied the initial *learning rate* used by the AdamW optimizer, setting it between 2e-5 and 8e-5. The best results, as indicated in the as indicated in table 7, show that the optimal combination involves a learning rate of 8e-5.

Table 7
NER hyperparameter selection - varying **learning rate**

Models	Learning Rate	MiP	MiR	MiF1
r-b-bio-clinical-es	2E-05	0.6983	0.7768	0.7355
r-b-bio-clinical-es	8E-05	**0.7899**	**0.7963**	**0.7931**
r-b-bio-clinical-es-FT	2E-05	0.6924	0.7675	0.7280
r-b-bio-clinical-es-FT	8E-05	**0.7798**	**0.7970**	**0.7883**
bsc-bio-ehr-es	2E-05	0.6957	0.7733	0.7325
bsc-bio-ehr-es	8E-05	**0.7865**	**0.7979**	**0.7922**
r-es-clinical-trials-ner	2E-05	0.7018	0.7814	0.7395
r-es-clinical-trials-ner	8E-05	**0.7877**	**0.7979**	**0.7927**

Weight decay Finally, we adjusted the *weight decay* applied to all layers except the bias and LayerNorm weights in the AdamW optimizer, starting with a value of 0.1 and then increased it to 0.2, which proved to be the best solution, as reported in table 8.

Table 8
NER hyperparameter selection - varying **weight decay**

Models	Weight Decay	MiP	MiR	MiF1
r-b-bio-clinical-es	0.1	0.7384	0.7920	0.7643
r-b-bio-clinical-es	0.2	**0.7899**	**0.7963**	**0.7931**
r-b-bio-clinical-es-FT	0.1	0.7439	0.7872	0.7649
r-b-bio-clinical-es-FT	0.2	**0.7798**	**0.7970**	**0.7883**
bsc-bio-ehr-es	0.1	0.7367	0.7889	0.7619
bsc-bio-ehr-es	0.2	**0.7865**	**0.7979**	**0.7922**
r-es-clinical-trials-ner	0.1	0.7405	0.7923	0.7655
r-es-clinical-trials-ner	0.2	**0.7877**	**0.7979**	**0.7927**

As a result of these analyses, we determined the optimal hyperparameters as follows: a batch size of 4, a learning rate of 8e-5, and a weight decay of 0.2.

We selected the value ’5’ for the initial epochs based on preliminary studies indicating that the pattern tended to converge rapidly. Furthermore, we observed that after the fifth epoch, performance no longer improved significantly. Therefore, to avoid overtraining and optimize training time, we chose to stop at the 5 epoch.