Cross-Linguistic Disease and Drug Detection in Cardiology Clinical Texts: Methods and Outcomes

Notebook for the BioASQ Lab at CLEF 2024

Patrick Styll[1,*], Leonardo Campillos-Llanos[2], Wojciech Kusa[1] and Allan Hanbury[1]

[1]Data Science Research Unit (E194-04), Technische Universität Wien, Favoritenstraße 9-11, 1040 Vienna, Austria
[2]Institute of Language, Literature and Anthropology, Spanish National Research Council (CSIC), c/Albasanz 26, 28037 Madrid, Spain

Abstract

This paper presents our approach to the MultiCardioNER lab at CLEF2024, focusing on disease detection in Spanish texts and drug detection in Italian, Spanish, and English texts. We enhance model performance through several strategies: (1) fine-tuning on automatically translated TREC Clinical Trials admission notes using Masked Language Modeling (MLM); (2) data augmentation with translated MTSamples processed through a Spanish medical lexicon (MedLexSp) for accurate vocabulary matching; and (3) employing sliding windows with overlap to improve data capture. Additionally, we use transfer learning with a clinical trials corpus (CT-EMB-SP) to refine the outcomes. We further fine-tune several already established disease and drug extraction models to leverage their extensive vocabulary and compare their performance to models trained from scratch. Our methods and experiments demonstrate notable improvements in multilingual clinical NER, as evidenced by our track results.

Keywords

Clinical Named Entity Recognition, Transfer Learning, Data Augmentation, Cardiology

1. Introduction

The increasing volume of clinical text data presents both challenges and opportunities for the healthcare sector [1]. Extracting meaningful information from these texts, such as disease and drug mentions, is critical for applications such as patient care, clinical research and healthcare management [2]. In this context, the MultiCardioNER [3] task from the BioASQ [4] workshop at CLEF2024 provides an important platform for evaluating and advancing clinical named entity recognition (NER) technologies, both in monolingual and multilingual settings. MultiCardioNER is organized by the Barcelona Supercomputing Center's Natural Language Processing (NLP) for Biomedical Information Analysis group and is promoted by Spanish and European projects such as DataTools4Heart, AI4HF, BARITONE, and AI4ProfHealth. This shared task focuses on the multilingual adaptation of clinical NER systems to the cardiology domain. It includes two key tasks: disease detection in Spanish texts and drug detection across Italian, Spanish, and English texts. Our work addresses these tasks through innovative strategies designed to enhance model performance, which are detailed in this paper. Section 2 provides the background and an overview of the proposed techniques, baseline models and evaluation metrics. In Section 3, we take a look at the practical effect of the introduced methodology via preliminary experiments. Furthermore, we reflect on the results we obtained from the submitted runs via an extensive error analysis. Finally, in Section 4 we conclude our research and discuss and summarize our findings.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
patrick.styll@tuwien.ac.at (P. Styll); leonardo.campillos@csic.es (L. Campillos-Llanos); wojciech.kusa@tuwien.ac.at (W. Kusa); allan.hanbury@tuwien.ac.at (A. Hanbury)
https://github.com/Padraig20 (P. Styll); https://sites.google.com/view/lcampillos/index (L. Campillos-Llanos); https://wojciechkusa.github.io/ (W. Kusa); https://informatics.tuwien.ac.at/people/allan-hanbury (A. Hanbury)
0000-0003-3040-1756 (L. Campillos-Llanos); 0000-0003-4420-4147 (W. Kusa); 0000-0002-7149-5843 (A. Hanbury)

2. Methodology

In this Section, we give the background to our methodology. We describe our proposed techniques to further enhance results, the baseline models we used and discuss how results and outputs are effectively evaluated.

2.1. Proposed Techniques

- **Fine-tuning via Masked Language Modeling**
  We proposed fine-tuning on 240 automatically translated admission notes of the TREC Clinical Trials track via Masked Language Modeling [5]. This should help the model to produce a sense of what patient notes look like and enhance their understanding.
- **Data Augmentation**
  We used the NickyNicky/medical_mtsamples dataset from HuggingFace as a means of data augmentation. We extracted Cardiology Diseases and Drugs, automatically translated the texts to Spanish via the Google Translate API [6] and additionally processed the entities using a medical lexicon for Spanish (MedLexSp [7]) to ensure only correct medical vocabulary was used.
- **Sliding Windows with Overlap**
  We employed a sliding windows attention approach with overlap to handle long sequences of clinical text. This method has been effectively utilized in various Natural Language Processing (NLP) tasks to manage texts that exceed the input size limitations of standard models [5]. By breaking the text into smaller, overlapping segments, the model can better understand the context and connections between different Sections of the document.
- **Additional Fine-Tuning/Transfer Learning on general diseases/drugs**
  We fine-tuned several baseline models to detect diseases and drugs from the CT-EMB-SP corpus [8] with the goal of enhancing the model's vocabulary of specific medical data. This is a collection of 1200 texts about clinical trials in Spanish (500 journal abstracts and 700 trial announcements). It was annotated with entities for four semantic groups of the Unified Medical Language System [9]: ANAT, CHEM, DISO and PROC. This resource facilitates machine learning experiments for information extraction on evidence-based medicine. For information on the models and training process, see Table 3 and Figures 6a and 6b in Appendix A.

2.2. Baseline Models

This Section introduces all pre-trained models which we used in both experiments and submissions. It is explained how they were pre-trained, why the model is potentially useful and how we employed it in our research.

- **google-bert/bert-base-multilingual-cased**
  This is a multilingual version of the BERT model [5], which served as our baseline. It is a small model that we fine-tuned and used in the preliminary evaluation of track 1.
- **microsoft/mdeberta-v3-base** [10] [11]
  This large, multilingual general-domain model has recently gained recognition for its effectiveness in processing medical data. We fine-tuned it both from scratch and on the CT-EMB-SP corpus [8] for increased vocabulary. We used this model for every part of the track. Table 3 in Appendix A shows the parameters and performance of this model.
- **lcampillos/roberta-es-clinical-trials-ner** [8]
  This model is based on the RoBERTa architecture and is specifically fine-tuned for named entity recognition tasks in Spanish clinical trial texts. It is designed to effectively identify medical entities within the domain of clinical trials, enhancing the extraction of relevant information from these documents. On the evaluation set of its training data, it achieved a strong F1-score of

86.47%, demonstrating its effectiveness. We used it for every part of the track.
This model seemed very promising, since on preliminary testing on the MultiCardioNER data, it already achieved an F1-score of 45.52% for track 1 and 76.04% for track 2. This suggests that there is more difference between the general medical domain and the cardiology domain for diseases than for pharmaceuticals.

- **PlanTL-GOB-ES/bsc-bio-ehr-es** [12]
  This model is pre-trained on Spanish electronic health records (EHR), a large corpus of biomedical texts. It evidently outperformed other popular models on certain tasks, showcasing its performance. We used it for track 1 and the Spanish part of track 2.
- **IVN-RIN/bioBIT** [13]
  BioBIT (Biomedical Bert for Italian) is a model tailored for the biomedical domain, pre-trained on an Italian biomedical corpus derived from machine-translated PubMed abstracts. Built on the BERT architecture, BioBIT utilizes Masked Language Modeling and Next Sentence Prediction for pretraining. It excels in multiple tasks, including Named Entity Recognition (NER), achieving high accuracy across several biomedical datasets. We used it for evaluating the Italian part of track 2.
- **alvaroalon2/biobert_chemical_ner** [14]
  This BioBERT model is fine-tuned for named entity recognition (NER) tasks specifically targeting chemical entities. It has been trained on the BC5CDR-chemicals [15] and BC4CHEMD corpora [16], making it highly effective for identifying chemical mentions in biomedical texts. This model is a valuable tool for chemical NER in the biomedical domain, supporting advanced research and data extraction.

2.3. Metrics

For evaluating the models, we used evaluation metrics based on entity-level [17]. Since we are working with a highly imbalanced dataset, this provides a more accurate assessment of Named Entity Recognition (NER) performance.

The International Workshop on Semantic Evaluation (SemEval'13) introduced four ways to evaluate Named Entity Recognition (NER) performance: *Strict*, *Exact*, *Partial*, and *Type*. These methods consider various aspects of matches between system predictions and ground truth annotations. The evaluation schemas assess correctness, incorrectness, partial matches, missed entities, and spurious entities differently, impacting the calculated precision, recall, and F1-scores.

Strict requires an exact boundary and type match, *Exact* requires just boundary match, *Partial* accepts partial boundaries, and *Type* requires some overlap. These metrics provide a comprehensive evaluation of Named Entity Recognition (NER) systems under different match criteria.

When assessing the performance of our models, we use the average of the four F1-scores of the evaluation metrics: *Strict*, *Exact*, *Partial*, and *Type*. This average F1-score ($F1_{avg}$) is calculated as follows:

$$F1_{avg} = \frac{Strict + Exact + Partial + Type}{4}$$

This method allows us to effectively determine the most performant model by considering a balanced view of different evaluation criteria.

3. Experiments and Results

3.1. Preliminary Experiments

The performance of the models is evaluated by cutting off excessive tokens from each patient note, where each model has an input size of 512 tokens. If models are not trained via sliding windows, excess tokens are simply cut off during the tokenization process. Note that, due to time constraints, we run the experiments just once for each model. However, a better methodology would be initializing each model with different seed values and reporting the average and standard deviation of all runs. This method would provide a more realistic overview of each model's performance.

Please note that these preliminary experiments do not yet include error analysis, as described in Section 3.3. The absolute performance of the models depicted here is not demonstrative, but the relative difference of the separate runs showcases success of the proposed techniques and some insights into how the models behave. For additional experiments where absolute performance of the models is depicted, please see Appendix C.

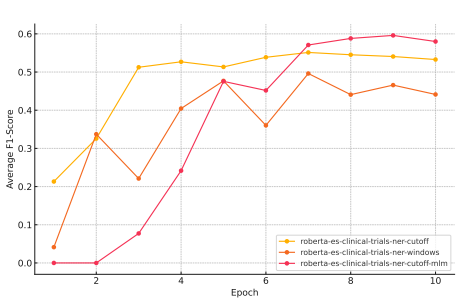
3.1.1. Experiments for Track 1

Baseline; Baseline values are given by the *multilingual-bert* and *bsc-bio-ehr-es*, which already achieved decent F1-scores on the development set, with *bsc-bio-ehr-es* being at 81.48% and *multilingual-bert* at 76.50%. This already suggests a clear benefit for fine-tuning on domain-specific data for specific tasks.

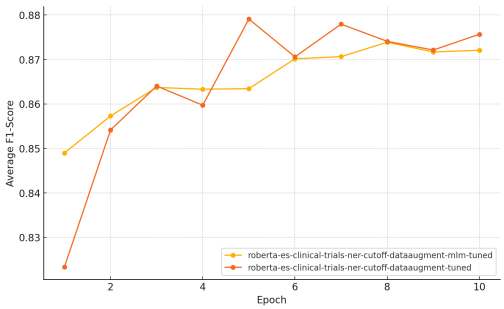
Domain-Specific Model; For this purpose, the *roberta-es-clinical-trials-ner* model (as introduced in Section 2.2) was used as a baseline. Since it was fine-tuned on general diseases in the Spanish language (not only cardiovascular conditions), it already started with a relatively high F1-score in the first epoch (see Figure 1a). We can see that the data-augmentation technique showed a promising influence in further fine-tuning the model to the cardiology domain. The sliding windows approach showed slightly worse results. However, the difference is not large enough for a conclusion.

After some hyperparameter-tuning, the model achieved 87.9% F1-score at its peak. As evident in Figure 1b, the Masked Language Modeling approach did not necessarily influence results. This might be due to the lack of data for this kind of fine-tuning, which may add unnecessary bias.

During the process of hyperparameter-tuning, we saw that a higher learning rate (i.e. $1e^{-4}$ instead of $2e^{-5}$) performed slightly better (approximately an increase of 4% in the evaluation metric). The same can be said for the batch size, where a higher size yielded better results (approximately 7% in F1-score). Unfortunately, experiments were rather limited here due to the lack of GPU RAM.



(a) Domain-Specific Model.



(b) Domain-Specific Model (tuned).

Figure 1: Domain Specific Model Experiments. *roberta-es-clinical-trials-ner-cutoff* used the *cutoff-strategy*, *roberta-es-clinical-trials-ner-windows* used the *sliding windows technique* and *roberta-es-clinical-trials-ner-cutoff-mlm* was first fine-tuned via Masked Language Modeling (MLM) on admission notes and then fine-tuned onto the cardiology domain via the *cutoff-strategy*. For the tuned models in Figure 1b, data-augmentation as well as the *cutoff-strategy* was performed for both runs, while *roberta-es-clinical-trials-ner-cutoff-dataaugment-mlm-tuned* was further fine-tuned via Masked Language Modeling (MLM) on admission notes.

Multilingual Model For this purpose, the *mdeberta-v3-base* model was used as a pre-trained model. We first fine-tuned the baseline model on the cardiology data provided by the shared task, which already showed promising results. It is also interesting to see that in the beginning the model's performance started much lower than those models that were already fine-tuned on general diseases (see Figure 2). Fine-tuning the model on general diseases using the CT-EMB-SP corpus showed promising changes in performance. Adding data-augmentation and Masked Language Modeling (MLM) as additional techniques only influenced the results slightly. In the end, it reached an F1-score of 87% at its peak.

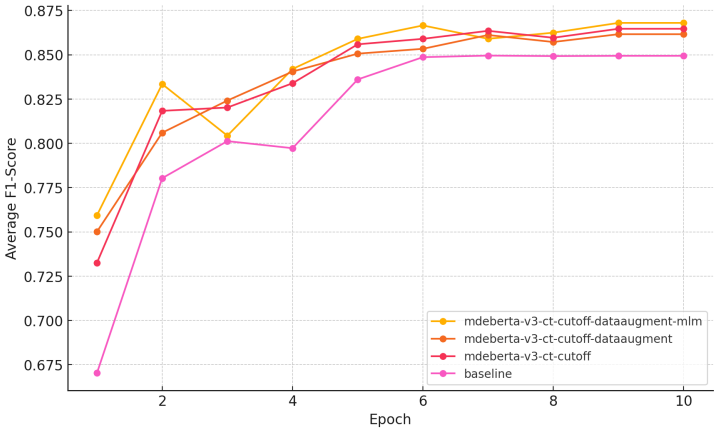
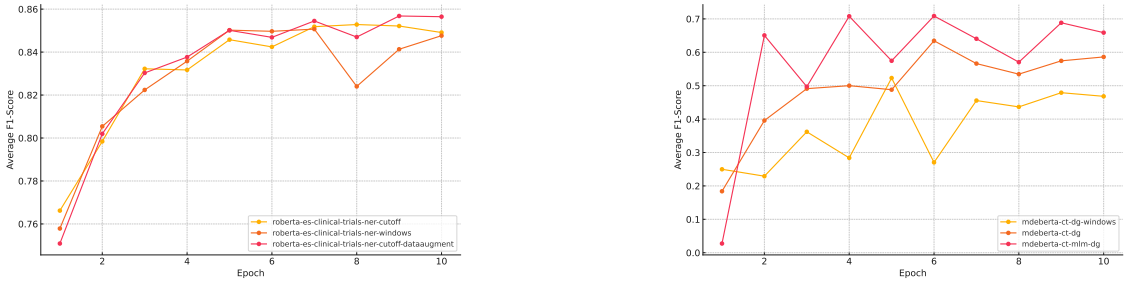


Figure 2: Multilingual Model. *baseline* is the *mdeberta-v3-base* model without any special techniques. *mdeberta-v3-ct-cutoff* was fine-tuned on general diseases before being fine-tuned onto the cardiology domain via the *cutoff-strategy*. *mdeberta-v3-ct-cutoff-dataaugment* additionally used data-augmentation, and *mdeberta-v3-ct-cutoff-dataaugment-mlm* was additionally pre-trained via Masked Language Modeling (MLM) onto admission notes.

3.1.2. Experiments for Track 2

Domain-Specific Model (es) The *roberta-es-clinical-trials-ner* model was used as a baseline. Surprisingly, the scores were relatively low. This was unexpected, since we already measured much better performance by this model on the MultiCardioNER data. Eventually, we obtained an F1-score of 59.79%.

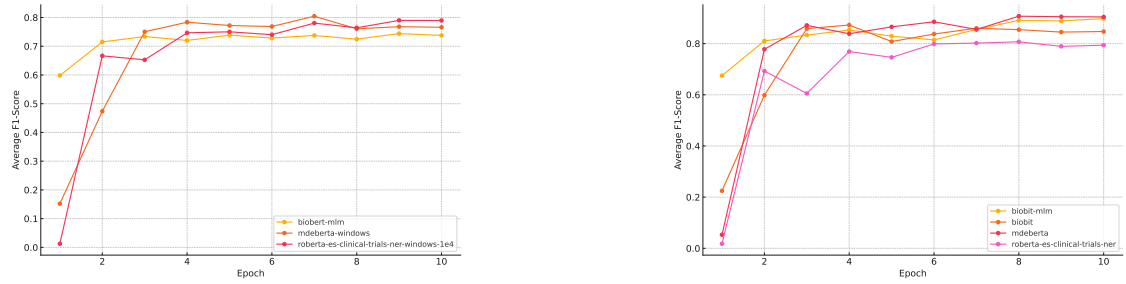
Multilingual Model (es) As in previous experiments, the *mdeberta-v3-base* model was used as a baseline and fine-tuned on drugs from the CT-EMB-SP corpus (which did not happen for the other languages where the base model was used). As can be seen in Figure 3b, the combination of Masked Language Modeling and data-augmentation actually brought much benefit. In the end, the model achieved an F1-score of 70.87% at its peak.



(a) Domain-Specific Model. (b) Multilingual Model.

Figure 3: Spanish track performance. *roberta-es-clinical-trials-ner-cutoff* used the *cutoff-strategy* when being fine-tuned onto the Cardiology domain, while *roberta-es-clinical-trials-ner-cutoff-dataaugment* additionally used data augmentation. On the other hand, *roberta-es-clinical-trials-ner-windows* used the *sliding windows approach*. *mdeberta-ct-dg* was first fine-tuned onto the general domain, before being fine-tuned onto cardiology via the *cutoff-strategy* and data-augmentation. *mdeberta-ct-mlm-dg* further used Masked Language Modeling (MLM) onto admission notes. On the other hand, *mdeberta-ct-dg-windows* used the *sliding windows approach* as well as data-augmentation.

Insights from other languages (en, it) Testing the *roberta-es-clinical-trials-ner* model on both the English and Italian track provided us with interesting results. Looking at Figure 4a of the English track, we can actually see that the Spanish model outperformed both the multilingual general domain model (i.e. *mdeberta-v3-base* and the domain-specific model (i.e. *BioBERT*). The same cannot be said for the Italian track, where it was outperformed by every other run (see Figure 4b). For the Italian track, the *mdeberta-v3-base* model won with an F1-score of 90.7%, while the *roberta-es-clinical-trials-ner* achieved 80.45% on the English track. These results suggest a *great multilingual overlap for pharmaceuticals*.



(a) English Track. (b) Italian Track.

Figure 4: English and Italian track performance. In Figure 4a, *biobert-mlm* used the *cutoff-strategy* as well as Masked Language Modeling (MLM) via admission notes. Both *mdeberta-windows* and *roberta-es-clinical-trials-ner-windows-1e4* used the *sliding windows approach*. In Figure 4b, all models used the *cutoff-strategy*, where *biobit-mlm* additionally used MLM via admission notes.

3.2. Official Submissions

The submission runs are described in Appendix D.

Table 1
Results of Submission Runs.

Track	Run name	P	R	F1
Track1	run1_mdeberta-ct-mlm-dg	59.28%	67.15%	62.97%
Track1	run2-mdeberta-ct	50.27%	68.84%	58.10%
Track1	run3_mdeberta-ct-dg	48.00%	67.73%	56.18%
Track1	run4-roberta-dg	65.65%	73.76%	69.47%
Track1	run5-roberta-dg-windows	65.46%	72.44%	68.77%
Track2_ES	run1_mdeberta-multilingual	39.14%	15.31%	22.01%
Track2_ES	run2_mdeberta-ct-multilingual	76.47%	35.56%	48.55%
Track2_ES	run3_roberta-ct-multilingual	87.05%	43.42%	57.94%
Track2_ES	run4_mdeberta_ct_mlm_dg	68.15%	38.36%	49.09%
Track2_ES	run5_roberta-ct-mlm	84.21%	39.12%	53.42%
Track2_EN	run1_mdeberta-multilingual	56.48%	24.81%	34.48%
Track2_EN	run2_mdeberta-ct-multilingual	84.53%	37.77%	52.21%
Track2_EN	run3_roberta-ct-multilingual	86.32%	43.64%	57.97%
Track2_EN	run4-mdeberta-windows	79.55%	43.17%	55.97%
Track2_EN	run5-biobert-mlm-windows	67.71%	44.10%	53.41%
Track2_IT	run1_mdeberta-multilingual	50.74%	20.94%	29.65%
Track2_IT	run2_mdeberta-ct-multilingual	74.33%	33.94%	46.61%
Track2_IT	run3_roberta-ct-multilingual	82.64%	42.06%	55.74%
Track2_IT	run4-mdeberta	74.81%	39.28%	51.51%
Track2_IT	run5-biobit-mlm	79.22%	35.17%	48.71%

It is important to note that these results do not reflect the absolute performance of the models (see Section 3.3 for further details and Appendix C for additional experiments with demonstrative performance), but the relative difference of the separate runs showcase the success of the proposed techniques and some insights into how the models behave.

- **Fine-Tuning on General Diseases (Transfer-Learning)**
  The first three runs (multilingual runs) of track 2 show evidence that fine-tuning models on general diseases before focusing on the cardiology domain yielded great benefit. The first run, which was not fine-tuned on general diseases, showed worse performance than run 2 and 3, which were both fine-tuned on the CT-EMB-SP corpus.
- **Multilingual Overlap for Pharmaceuticals**
  Looking at run 2 and run 3 of track 2, we can see that the multilingual models performed similarly to the monolingual models. This suggests a big multilingual overlap for drugs in Spanish, English and Italian.
- **Possible Noise by Data Augmentation due to Machine Translation**
  Several runs (e.g. run 2 and run 3 of track 1) showed slightly worse performance when data augmentation was used. This suggests possible additional noise in the training data. For data augmentation, we used the MTSamples dataset, where we used the keywords as entities and translated all text via the Google Translate API. On first inspection, these keywords may refer to laboratory tests, procedures or anatomical entities. Therefore, we have processed the translated MTSamples with MedLexSp; we output only DISO and CHEM categories, and we re-named them to ENFERMEDAD and FARMACO, respectively. Nonetheless, it is unsure whether these data contain general diseases or only cardiology diseases. Furthermore, on closer inspection, there are some issues with the machine translation, which is also visible in the automatic translation of the TREC Admission Notes for Masked Language Modeling (MLM) fine-tuning. Either some words

are not translated or new words are created, possibly due to the sub-words of neural models. An example would be **leucitos en urino*, which should be *leucocitos en orina* ('white cells in urine').

3.3. Error Analysis

3.3.1. Data Reconstruction

Several problems arose during generating the runs; namely, reconstructing the output of the *mdeberta-v3-base* and *roberta-es-clinical-trials-ner* model.

mdeberta-v3-base This model posed several problems, particularly in generating the correct index of the span. Often, the start of the span would be one or two tokens off, leading to a decrease in the F1-score for the runs. Additionally, tokens might have leading spaces or newline characters at the beginning or end. These extraneous characters need to be removed to ensure the entity text is clean and accurate. This also includes adjusting the start and end offsets to reflect the new positions of the cleaned tokens.

The presence of punctuation at the end of tokens can create issues in entity recognition. Special rules are required to handle exceptions such as units (*mg.*) or cases where brackets are involved. Unnecessary punctuation needs to be removed, but care must be taken to preserve punctuation that is part of the entity. Furthermore, the model would sometimes add tabs instead of spaces into the extracted entity. When tokens are merged or cleaned, their character offsets in the text need to be recalculated. This ensures that the entities' positions in the text are accurately represented, which is crucial for tasks like text highlighting or linking entities back to their original context.

roberta-es-clinical-trials-ner This model exhibits significant issues with handling sub-words, often treating them as separate entities. Specifically, leading sub-words are represented as individual entities with a preceding space. This requires special considerations during the reconstruction process to ensure accurate entities.

General Remarks We analysed all errors made by the *roberta-es-clinical-trials-ner* model in run 4 of track 1; we used a Python script to count each type. There are several types of errors that the model made while generating the runs. Examples of this run may be found in Table 2.

1. **Scope Errors**
   a) **Incompletely predicted entities**
      These are entities where the model predicted only a part of the actual entity, missing some crucial parts.
   b) **Entities where too many words were predicted**
      Entities where the predicted span includes extra information not part of the actual entity.
   c) **Entities that would belong together**
      Entities where the predicted spans should be combined to form a single coherent entity.
2. **False Positives**
   Entities that were incorrectly identified by the model, but not labeled in the ground truth.
3. **False Negatives (Missed entities)**
   Entities that were missed by the model, but were labeled in the ground truth.

Looking into Figure 5, we can see that the high number of scope errors (correctly identified entities) indicates that the model generally has a strong baseline capability for recognizing entities correctly. Despite the high accuracy in identifying entities, the model still exhibits significant precision and recall

Table 2
Examples for incorrectly extracted entities from the test dataset. *T* is representative of the type of error that the model made while generating the runs.

Filename	T	Prediction	Test Set
multicardioner_test+bg_7336	1a	fumador	fumador activo
multicardioner_test+bg_7845	1a	taquicardia de QRS estrecho arrítmica	taquicardia de QRS estrecho
multicardioner_test+bg_7845	1a	vía accesoria lateral	vía accesoria lateral izquierda
multicardioner_test+bg_7560	1b	dilatación de VI con función severamente deprimida	dilatación de VI
multicardioner_test+bg_7845	1b	taquicardia ventricular originada en el músculo papilar posterior	taquicardia ventricular
multicardioner_test+bg_7845	1b	Cardiopatía hipertensiva con disfunción diastólica	Cardiopatía hipertensiva
multicardioner_test+bg_75	1c	IM jetivamente	IM subjetivamente
multicardioner_test+bg_7560	2	a IECA	
multicardioner_test+bg_7560	2	hipertrabeculación	
multicardioner_test+bg_543	3		IAM
multicardioner_test+bg_7336	3		carga trombótica

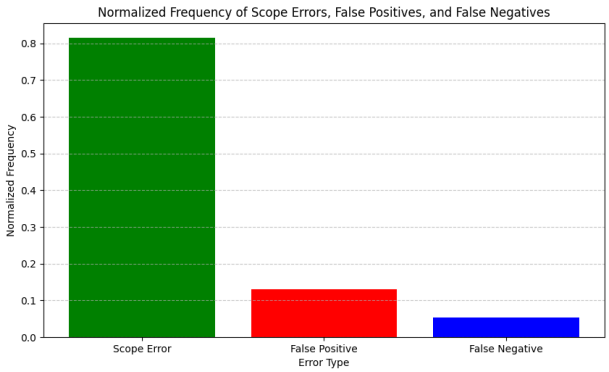


Figure 5: Normalized frequency of occurrences of different types of errors. All errors of run 4 of track 1 (see Table 1) has been chosen.

issues, as evidenced by the presence of false positives and false negatives. Furthermore, the high number of true positives amidst other errors implies that errors are not due to a fundamental flaw in the model but likely due to specific cases or contexts where the model's performance drops.

Nonetheless, score errors tend to cause less harm in the information extraction of clinical cases. Not detecting an entity (false negative, case 3 in Table 2) is more severe, whereas detecting *fumador* instead of *fumador activo* (case 1a in Table 2) is a mild error. This becomes even clearer when looking into errors where abbreviations were used. If written in the text is *diabetes mellitus (DM)*, the model would extract *diabetes mellitus (DM)*, while the test set would annotate both whole text and abbreviation separately. This happens very frequently, which makes the evaluation unsuitable for measuring the model's practical performance. In the end, a more relaxed evaluation metric could have been more appropriate and yielded higher results.

3.3.2. Data Capture

There are several factors that contributed to why the models underperformed in the submission runs (see Table 1). Further data analysis shows that the main issue was insufficient data capture during both training and evaluation. Plots may be found in Appendix B. In general, we refer to data capture to the

degree to which the model has difficulty with capturing all information from the patient notes. After all, in preliminary experiments and for generating submission runs, we simply cut off excess tokens that did not fit into the model.

General Remarks The relative distribution of entities over patient notes, as depicted in the density plots in Figure 8 (Appendix B), reveals several interesting insights.

For track 1, the training set displays a prominent spike at the beginning, followed by a relatively uniform distribution throughout the rest of the notes. In contrast, the development set and test set exhibit two significant spikes: one at the beginning and a larger one at the end, with a notably lower density in the middle. This pattern suggests that most diseases are mentioned either at the beginning or the end of the patient notes.

Turning to pharmaceuticals in track 2, we observe similar entity distributions across all plots. The training set again shows a more uniform distribution, whereas the development and test sets both feature two prominent spikes at the beginning and end, mirroring the pattern observed in track 1.

Looking at the words counts in the boxplots of Figure 7, we can clearly see that the training set exhibits significantly less words than both the development set and the test set. About 75% of patient notes from the test set have less than 550 words, which applies to less than 25% of patient notes from the development/test set.

The Venn diagrams in Figure 9 (Appendix B) are also worth mentioning. As we can see, track 1 seems to have only little overlap between the datasets, while track 2 has notably more overlap. This may imply that the model suffers from a few-shot learning problem, especially since results on track 2 are significantly better in terms of performance than those of track 1. Another factor may be the amount of unique entities in the datasets, which is way larger in track 1 than in track 2, further complicating the task for the model.

Implications for Training and Evaluation The previous training and evaluation strategies were significantly affected by these entity distribution patterns. During previous evaluation, a cutoff strategy was used, where all excess tokens were trimmed to fit the model's input layer, which was uniformly set at 512 tokens. This meant that only approximately 60% of the data was fully utilized during training. However, due to the high density of entities at the end of patient notes, this approach resulted in sub-optimal data capture. The situation was even worse during evaluation on the development set, where only less than 25% of the data fit into the models without token cutoff. This led to the model being evaluated on a non-representative portion of the dataset, inflating the performance metrics.

To improve data capture, we decided to split the patient notes into individual sentences using *spaCy* [18] for both training and evaluation. This change not only yielded better results, particularly for track 2, but also provided more reliable and representative metrics. Consequently, several experiments were re-conducted (refer to Appendix C). It is important to note that these new runs expand and confirm the trends observed in earlier experiments (see Section 3.1).

4. Conclusion

We can see some interesting trends in the data, allowing us to draw both conclusions about our proposed strategies as well as the provided data.

- **Fine-tuning via Masked Language Modeling:** This approach had very little influence on the model's results. This can be attributed to (i) the lack of sufficient data for this kind of fine-tuning, (ii) the fact that the patient notes are based on the general domain, and (iii) erroneous machine translation.

- **Data Augmentation:** The effects of data augmentation are still unclear. We have observed both positive and negative effects across different model architectures. More experiments with different models and types of data augmentation resources are necessary to draw definitive conclusions.
- **Sliding Windows with Overlap:** The impact of the sliding windows approach, as opposed to cutting off excess tokens, is also difficult to judge. Despite expecting better data capture, some experiments actually showed slightly worse results. This effect may be due to patient notes being split in random positions, resulting in incorrect grammar and split entities, which can disrupt the contextual information the model relies on. This issue becomes more evident when considering that processing patient notes at the sentence level improved results notably.
- **Additional Fine-Tuning/Transfer-Learning on general diseases/drugs:** This approach significantly improved the model's performance. Various experiments demonstrated that adapting a general model to a specific domain requires less effort and yields promising results with relatively little training.
- **Insufficient Data Capture:** Due to the high density of entities in the beginning and end of the patient notes, the cutoff strategy performed poorly due to missing entities at the end of the notes.
- **Overlap of Entities over Datasets:** There are significantly less overlapping entities between the training, development and testing dataset for track 1 than there are for track 2. This may explain the generally worse results for track 1, indicating that models may suffer from a few shot learning problem.
- **Multilingual Overlap for Pharmaceuticals:** We have shown that there is a big multilingual overlap concerning pharmaceuticals in Spanish, Italian and English. This can be largely attributed to the standardized pharmaceutical nomenclature, which suggests that a multilingual approach to drug entity extraction can leverage these similarities to enhance accuracy and consistency across different languages.

Acknowledgments

Leonardo Campillos-Llanos' work is conducted in the CLARA-MeD project (PID2020-116001RA-C33), funded by MICIU/AEI/10.13039/501100011033/, in call Proyectos I+D+i Retos Investigación.

References

[1] H. Dalianis, Clinical text mining: Secondary use of electronic patient records, Springer Nature, 2018.

[2] D. Demner-Fushman, N. Elhadad, C. Friedman, Natural language processing for health-related texts, in: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Springer, 2021, pp. 241–272.

[3] 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.), CLEF Working Notes, 2024.

[4] 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.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[6] Google Inc., Google translate api, https://cloud.google.com/translate, n.d. Accessed: 2024-05-21.

[7] L. Campillos-Llanos, MedLexSp – a medical lexicon for Spanish medical natural language processing, Journal of Biomedical Semantics 14 (2023) 2. URL: https://doi.org/10.1186/s13326-022-00281-5. doi:10.1186/s13326-022-00281-5.

[8] L. Campillos-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 and Decision Making 21 (2021) 69. URL: https://doi.org/10.1186/s12911-021-01395-z. doi:10.1186/s12911-021-01395-z.

[9] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[10] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.

[11] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[12] 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: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[13] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localizing in-domain adaptation of transformer-based biomedical language models, Journal of Biomedical Informatics 144 (2023) 104431. URL: https://www.sciencedirect.com/science/article/pii/S1532046423001521. doi:https://doi.org/10.1016/j.jbi.2023.104431.

[14] Á. Alonso Casero, Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature, 2021. URL: https://oa.upm.es/67933/, unpublished.

[15] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, Biocreative V CDR task corpus: a resource for chemical disease relation extraction, Database J. Biol. Databases Curation 2016 (2016). URL: https://doi.org/10.1093/database/baw068. doi:10.1093/database/baw068.

[16] M. Krallinger, O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, R. A. Sayle, R. T. Batista-Navarro, R. Rak, T. Huber, T. Rocktäschel, S. Matos, D. Campos, B. Tang, H. Xu, T. Munkhdalai, K. H. Ryu, S. V. Ramanan, S. Nathan, S. Žitnik, M. Bajec, L. Weber, M. Irmer, S. A. Akhondi, J. A. Kors, S. Xu, X. An, U. K. Sikdar, A. Ekbal, M. Yoshioka, T. M. Dieb, M. Choi, K. Verspoor, M. Khabsa, C. L. Giles, H. Liu, K. E. Ravikumar, A. Lamurias, F. M. Couto, H.-J. Dai, R. T.-H. Tsai, C. Ata, T. Can, A. Usié, R. Alves, I. Segura-Bedmar, P. Martínez, J. Oyarzabal, A. Valencia, The CHEMDNER corpus of chemicals and drugs and its annotation principles, Journal of Cheminformatics 7 (2015) S2. URL: https://doi.org/10.1186/1758-2946-7-S1-S2. doi:10.1186/1758-2946-7-S1-S2.

[17] D. S. Batista, Named-entity evaluation metrics based on entity-level, 2018. URL: https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/, accessed: 2024-05-21.

[18] Explosion-AI, spaCy: Industrial-strength Natural Language Processing in Python, https://spacy.io/usage/linguistic-features#sbd, 2023. URL: https://spacy.io/, version 3.0.

A. CT-EMB-SP Fine-Tuning

Table 3
Parameters and Performance for ENFERMEDAD and FARMACO using *mdeberta-v3-base*.

	ENFERMEDAD	FARMACO
Learning Rate	2e-5	2e-5
Batch Size	16	16
Epochs	10	10
Input Size	512	512
Weight Decay	0.01	0.01
Optimizer	AdamW	AdamW
F1 Score	94.45%	93.89%

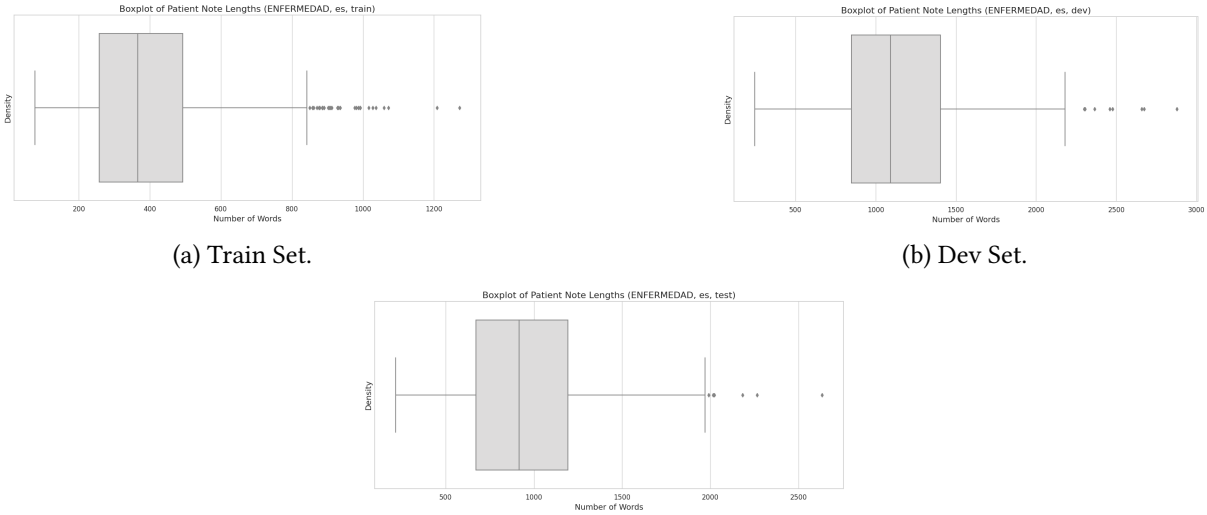
(a) Enfermedad.

(b) Farmaco.

Figure 6: Model Performance during Fine-Tuning.

B. Data Analysis

It is important to note that for track 2 (FARMACO), the density plots in Figure 8 and boxplots in Figure 7 look the same among the three different languages, despite translation. Trivially, the boxplots in Figure 7 look the same for both track 1 and track 2 since the same data was used.

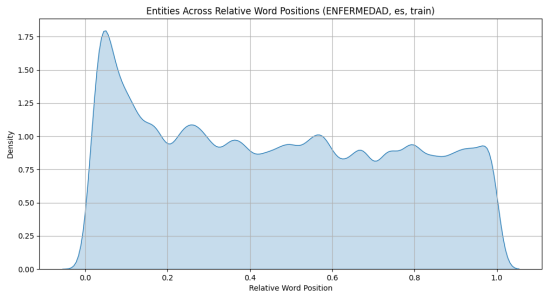


(a) Train Set.

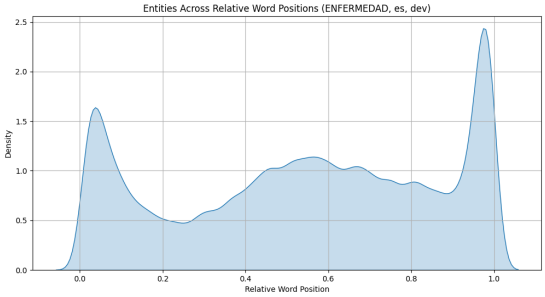
(b) Dev Set.

(c) Test Set.

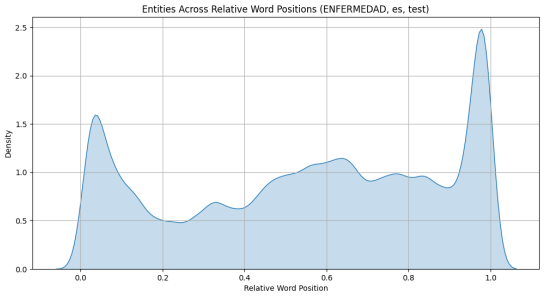
Figure 7: Amount of Words per Patient Note - Boxplot.



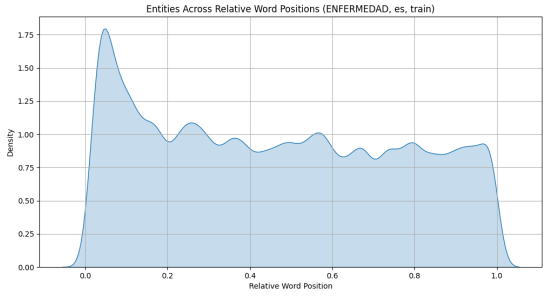
(a) Track 1 - Train Set.



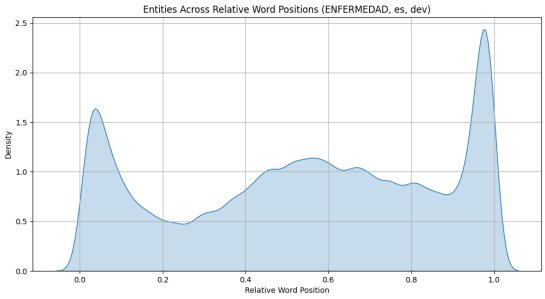
(b) Track 1 - Dev Set.



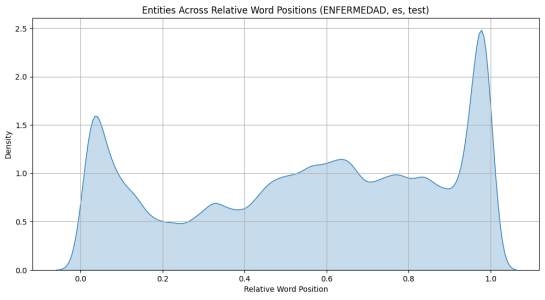
(c) Track 1 - Test Set.



(d) Track 2 - Train Set.

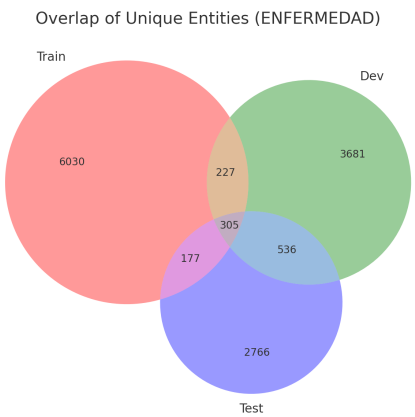


(e) Track 2 - Dev Set.

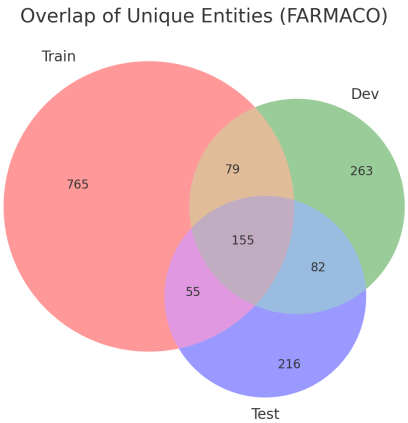


(f) Track 2 - Test Set.

Figure 8: Relative Entity Positions - Density Plot. Low values (close to 0) represent text positions at the beginning of the document; and values close to 1, positions at the end of the document.



(a) Track 1 - ENFERMEDAD.



(b) Track 2 - FARMACO.

Figure 9: Overlap of Unique Entities over train, development and test set.

C. Additional Experiments

C.1. Cardiology Domain Adaptation

This experiment serves to show how easily a general model, i.e. trained on general pharmaceuticals, can be adapted to a special medical domain. The *roberta-es-clinical-trials-ner* model was fine-tuned on general drugs using the CT-EMB-SP corpus in Spanish, and it was used as a base model and fine-tuned on the cardiology domain for pharmaceuticals. As previously mentioned in Section 2.2, it already achieved an F1-score of 76.04% before fine-tuning on cardiology data.

As seen in Figure 10, epoch 1 already shows an incredible performance on the development set. Evaluating this model of epoch 1 on the test set, we achieved a precision of 86.15%, a recall of 93.77% and an F1-score of 89.80%. Using a model trained on three epochs (which shows the peak in Figure 10) we obtained a precision of 90.25%, a recall of 94.30% and an F1-score of 92.23%.

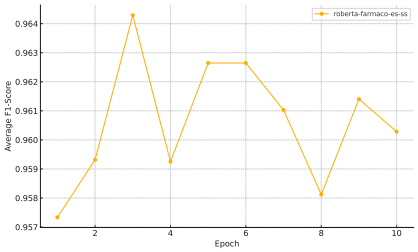


Figure 10: Fine-tuning *roberta-es-clinical-trials-ner* on the cardiology domain for Track 2 (FARMACO).

C.2. Effect of Data Augmentation

When looking into the effects of our proposed data augmentation, we trained *roberta-es-clinical-trials-ner* with and without data augmentation (same setup, i.e. same hyper-parameters). In Figure 11, we can see similar behaviour in training, but with less performance on the development set. When evaluating the model on the test set, we got a precision of 92.08%, a recall of 94.06% and an F1-score of 93.06%. Considering the model trained in Section C.1, data augmentation actually led to a slightly higher score.

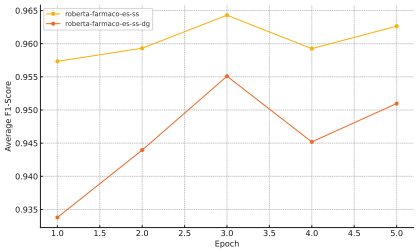


Figure 11: Fine-tuning *roberta-es-clinical-trials-ner* on the cardiology domain for the Spanish part for Track 2 (FARMACO) with data augmentation.

The same can be said when training the same model on track 1, where the plain model achieved a precision of 62.86%, a recall of 65.42% and an F1-score of 64.11%. With data augmentation, we achieved a significant improvement with a precision of 65.65%, a recall of 73.76% and an F1-score of 69.47%.

Nonetheless, when reflecting on the results in Section 3.2, we discussed possible negative effects due to incorrect machine translation. These effects are definitely visible when using e.g. *mdeberta-v3-base* as a baseline architecture (see also Table 1), which is why we are not entirely capable of judging the

effect of data augmentation. Although the outcomes of some models seem to support that it may help adapting a general model to a specific domain, we would need to experiment with more models and test more types of resources for data augmentation.

C.3. Effect of Fine-Tuning on General Domain - Transfer-Learning

In order to more precisely measure the benefit of fine-tuning a model on a general domain before fine-tuning it on a specialized domain, we conducted the following experiment on the Spanish part of track 2.

mdeberta-v3-base was used as a baseline. We compared the plain *mdeberta-v3-base* model with one fine-tuned on the general medical domain via the CT-EMB-SP corpus. Looking at the graph in Figure 12, the general model already showed greater performance than the base model in very early stages of training (validating once more what was seen with regard to adapting a general model to the medical domain, in Section C.1). The base model started with relatively low F1-score, but caught up in the last epochs.

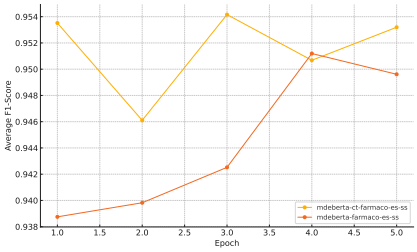


Figure 12: Fine-tuning *mdeberta-v3-base* on the cardiology domain for the Spanish part Track 2 (FARMACO). Model is either plain or originally fine-tuned on general pharmaceuticals in Spanish.

In Table 4, we can see the notable increase in performance not only on the development set, but also on the test set.

Table 4
Performance Metrics of the Plain and Fine-tuned Models Evaluated on the Ground Truth.

Model	Precision	Recall	F1 Score
Plain Model	87.56%	90.57%	89.04%
Transfer Learning	90.34%	93.60%	91.94%

C.4. Multilingual Capabilities

In order to check the assumption of a possible multilingual overlap for pharmaceuticals in Spanish, Italian and English, we trained and evaluated a multilingual general model (i.e. *mdeberta-v3-base*) on all three datasets simultaneously by having concatenated the data of all three languages.

As can be seen in Figure 13, the training exhibits a significant drop in performance to what can be seen in language specific models (see Figure 10 in Section C.1). The same can be said when looking into the results obtained when evaluating the model on the test set for each language separately (table 5). This can be explained via minor differences for special pharmaceutical words among the different languages, which may slightly add noise to the data.

The analysis of drug entity recognition across English, Spanish, and Italian (see Figure 14) demonstrates a significant overlap and similarity in the pharmaceutical terminology used in these languages. As seen

Table 5
Performance Metrics of the Multilingual Named Entity Recognition (NER) Model Evaluated on the Ground Truth for Each Language.

Language	Precision	Recall	F1 Score
Spanish	86.29%	89.00%	87.62%
English	85.37%	89.48%	87.38%
Italian	85.72%	86.10%	85.91%

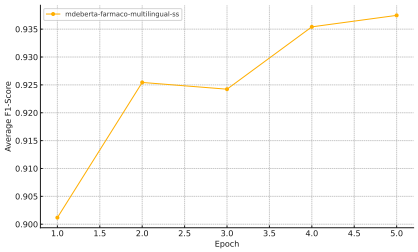


Figure 13: Fine-tuning *mdeberta-v3-base* on the cardiology domain for the all languages in Track 2 (FARMACO). The files for training and evaluation have simply been concatenated among all three languages.

in Table 6, many drug names exhibit minor variations that are primarily due to linguistic differences such as suffixes and spelling conventions. This overlap can be attributed to the standardized nature of pharmaceutical nomenclature and the widespread use of international nonproprietary names (INNs). These findings suggest that a multilingual approach to drug entity recognition can leverage these similarities to enhance accuracy and consistency across different languages.

Table 6
Examples of Similar Drug Names Across English, Spanish, and Italian.

Drug (Stemmed)	English Form	Spanish Form	Italian Form
lenalidomid	lenalidomide	lenalidomida	lenalidomide
caffein	caffeine	cafeína	caffeina
triamcinolone acetonid	triamcinolone acetonide	triamcinolona acetónido	triamcinolone acetonide
ampicillin	ampicillin	ampicilina	ampicillina
sulfacetamid	sulfacetamide	sulfacetamida	sulfacetamide

(a) No preprocessing.

(b) Stemmed and Lemmatized.

Figure 14: Overlap of pharmaceuticals in the training data for Spanish, Italian and English. To make the data more representative for similarity, we also provide entities that have been pre-processed via stemming and lemmatization. NLTK was used for Italian and English, while MedLexSp was used for Spanish.

D. Original Submission Runs

D.1. Track 1

run1-mdeberta-ct-mlm-dg The architecture of *mdeberta-v3-base* was used and fine-tuned on admission notes via Masked Language Modeling (MLM), with continual fine-tuning on general diseases. In order to further tune the model, we used data augmentation.

MLM

- Epochs: 5
- Learning Rate: 5e-6
- Loss: 9.3413
- Perplexity: 11399.24

Cardiology Task

- Learning Rate: 1e-4
- Epochs: 5
- F1 Avg: 87.24%
- F1 Exact: 86.76%
- F1 Partial: 86.76%
- F1 Ent Type: 88.69%
- F1 Strict: 86.76%

run2-mdeberta-ct The architecture of *mdeberta-v3-base* was used and fine-tuned on general diseases.

- Learning Rate: 2e-5
- Epochs: 10
- F1 Avg: 86.66%
- F1 Exact: 86.61%
- F1 Partial: 86.61%
- F1 Ent Type: 88.31%
- F1 Strict: 86.10%

run3-mdeberta-ct-dg The architecture of *mdeberta-v3-base* was used and fine-tuned on general diseases, this time including data augmentation via MTsamples.

- Learning Rate: 2e-5
- Epochs: 10
- F1 Avg: 85.08%
- F1 Exact: 84.43%
- F1 Partial: 84.43%
- F1 Ent Type: 87.03%
- F1 Strict: 84.43%

run4-roberta-dg The architecture of *lcampillos/roberta-es-clinical-trials-ner* was used and fine-tuned on the task of diseases in cardiology. To further tune the model on identifying only cardiology diseases, we used data augmentation via MTsamples. The base model already has a solid understanding of diseases and reaches an F1-score of 45.52%.

- Learning Rate: 2e-4

- Epochs: 5
- F1 Avg: 87.91%
- F1 Exact: 87.52%
- F1 Partial: 87.52%
- F1 Ent Type: 89.08%
- F1 Strict: 87.52%

run5-roberta-dg-windows The architecture of *lcampillos/roberta-es-clinical-trials-ner* was used and trained on the task of diseases in cardiology. To further tune the model on identifying only cardiology diseases, we used data augmentation via MTsamples. To further data capture, we used the proposed sliding windows technique.

- Learning Rate: 2e-4
- Epochs: 3
- Window Overlap: 60 tokens
- F1 Avg: 86.07%
- F1 Exact: 85.48%
- F1 Partial: 85.48%
- F1 Ent Type: 87.78%
- F1 Strict: 85.94%

D.2. Track 2

D.2.1. Multilingual Models

We propose three types of multilingual models, where all data from all three languages are taken and concatenated for training and evaluation.

run1-mdeberta-multilingual The architecture of *mdeberta-v3-base* was used.

- Learning Rate: 2e-5
- Epochs: 5
- F1 Avg: 82.43%
- F1 Exact: 82.06%
- F1 Partial: 82.06%
- F1 Ent Type: 83.54%
- F1 Strict: 82.06%

run2-mdeberta-ct-multilingual The architecture of *mdeberta-v3-base* was used and fine-tuned on general drugs in Spanish.

- Learning Rate: 2e-5
- Epochs: 5
- F1 Avg: 83.22%
- F1 Exact: 82.92%
- F1 Partial: 82.92%
- F1 Ent Type: 84.14%
- F1 Strict: 82.92%

run3-roberta-multilingual The architecture of *lcampillos/roberta-es-clinical-trials-ner* was used and fine-tuned on the task of detecting cardiology drugs. We worked under the assumption that pharmaceuticals may have very similar or even the same names in Spanish, Italian, and English. The base model already has a solid understanding of general drugs and reaches an F1-score of 81.79% for exact matching and 76.04% for strict matching.

- Learning Rate: 8e-5
- Epochs: 10
- F1 Avg: 75.14%
- F1 Exact: 74.86%
- F1 Partial: 74.86%
- F1 Ent Type: 76.01%
- F1 Strict: 74.86%

D.2.2. Language Specific Models

Each language has 2 language specific runs. The purpose of these runs is to compare domain-specific models (i.e. models specially trained on the medical domain and use transfer learning to specialize the model on the cardiology domain) to large language-agnostic, base models (i.e. *mdeberta-v3-base*). Run 4 contains the base model, while run 5 contains the domain-specific model.

es

run4-mdeberta-ct-mlm-dg The architecture of *mdeberta-v3-base* was used and fine-tuned on general drugs in Spanish. Furthermore, it was fine-tuned on Spanish admission notes via Masked Language Modeling (MLM). Additional data via the automatically translated MTsamples dataset was used.

MLM

- Epochs: 5
- Learning Rate: 8e-6
- Loss: 8.7417
- Perplexity: 10589.27

Cardiology Task

- Learning Rate: 8e-5
- Epochs: 4
- F1 Avg: 70.03%
- F1 Exact: 69.23%
- F1 Partial: 69.23%
- F1 Ent Type: 72.41%
- F1 Strict: 69.23%

run5-roberta-ct-mlm The architecture of *lcampillos/roberta-es-clinical-trials-ner* was used and fine-tuned on Spanish admission notes via Masked Language Modeling (MLM).

MLM

- Epochs: 5
- Learning Rate: 1e-6
- Loss: 8.8788
- Perplexity: 7178.02

Cardiology Task

- Learning Rate: 1e-4
- Epochs: 10
- F1 Avg: 59.59%
- F1 Exact: 59.05%
- F1 Partial: 59.05%
- F1 Ent Type: 61.21%
- F1 Strict: 59.05%

en

run4-mdeberta-windows The architecture of *mdeberta-v3-base* was used, including the sliding windows approach to enhance data capture.

- Learning Rate: 1e-4
- Epochs: 10
- Window Overlap: 60 tokens
- F1 Avg: 80.45%
- F1 Exact: 80.22%
- F1 Partial: 80.22%
- F1 Ent Type: 81.15%
- F1 Strict: 80.22%

run5-biobert-mlm-windows The architecture of *alvaroalon2/biobert_chemical_ner* was used and fine-tuned on English (original) admission notes via Masked Language Modeling (MLM). Furthermore, we used the sliding windows approach to enhance data capture. It is worth mentioning that *lcampillos/roberta-es-clinical-trials-ner* (with the same specifications) actually achieved slightly better results than *alvaroalon2/biobert_chemical_ner*, i.e. an average F1-score of 79.00%.

MLM

- Epochs: 5
- Learning Rate: 1e-6
- Loss: 8.65492
- Perplexity: 5738.31

Cardiology Task

- Learning Rate: 1e-4
- Epochs: 5
- Window Overlap: 60 tokens
- F1 Avg: 75.50%
- F1 Exact: 74.94%
- F1 Partial: 74.94%
- F1 Ent Type: 75.54%
- F1 Strict: 77.18%

it

run4-mdeberta The architecture of *mdeberta-v3-base* was used, without any data enhancing techniques.

- Learning Rate: 1e-4
- Epochs: 10
- F1 Avg: 90.70%
- F1 Exact: 90.49%
- F1 Partial: 90.49%
- F1 Ent Type: 91.34%
- F1 Strict: 90.49%

run5-biobit-mlm The architecture of *IVN-RIN/bioBIT* was used and fine-tuned on Italian admission notes via Masked Language Modeling (MLM). It is worth mentioning that *lcampillos/roberta-es-clinical-trials-ner* (with the same specifications) achieved worse results than *IVN-RIN/bioBIT* this time, i.e. an average F1-score of 76.81%.

- Learning Rate: 1e-4
- Epochs: 10
- F1 Avg: 89.77%
- F1 Exact: 89.56%
- F1 Partial: 89.56%
- F1 Ent Type: 90.40%
- F1 Strict: 89.56%