Procesamiento del Lenguaje Natural, Revista nº 76, marzo de 2026, pp. 123-134 recibido 05-12-2025 revisado 20-01-2026 aceptado 13-02-2026

Enhancing Location Entity Recognition in Spanish Medical Texts by Leveraging Domain Language Models and Data Augmentation

Mejora del reconocimiento de entidades de localización en textos médicos en español mediante modelos de lenguaje de dominio y aumento de datos

Irati Garitano, Raquel Martínez
Dept. of Computer Languages and Systems, UNED, Spain
{igaritano, raquel}@lsi.uned.es

Abstract: This work focuses on the automatic recognition of location entities in Spanish clinical reports, using the MEDDOPLACE challenge (IberLEF 2023) as the experimental framework. We evaluated both general-domain pre-trained models and biomedical-specific models. Furthermore, we explored data augmentation techniques via back-translation and LLM-based paraphrase generation. Our results outperform previous state-of-the-art approaches, demonstrating the effectiveness of combining these data augmentation strategies with pre-trained clinical domain models.
Keywords: location entity recognition, data augmentation, medical domain, domain-specific language models.

Resumen: Este trabajo se centra en el reconocimiento automático de entidades de lugar en informes clínicos en español, utilizando el desafío MEDDOPLACE (IberLEF 2023) como marco experimental. Evaluamos tanto modelos preentrenados de dominio general como modelos específicos del ámbito biomédico. Además, exploramos técnicas de aumento de datos mediante traducción inversa y generación de paráfrasis basada en LLM. Nuestros resultados superan los enfoques previos del estado del arte, demostrando la eficacia de combinar estas estrategias de aumento de datos con modelos preentrenados del dominio clínico.
Palabras clave: reconocimiento de entidades de lugar, aumento de datos, dominio médico, modelos de lenguaje de dominio específico.

1 Introduction

Clinical texts contain unstructured information about patients' health, medical history, treatments, and geographical or institutional data, where mentions of locations can be relevant for clinical assessment, disease surveillance, or outbreak analysis.

For infectious diseases such as malaria, a patient's recent geographical movements or residences can aid accurate diagnosis. Geographical information has been shown to improve outbreak detection, treatment, and containment (Wang et al., 2018). The COVID-19 pandemic emphasized the need for tools capable of efficiently extracting such information, as case tracking, transmission route identification, and resource allocation relied on rapid processing of large volumes of unstructured clinical data (Chen et al., 2021). During this period, recognizing location entities, normalizing terms, and linking clinical concepts with locations was essential.

For non-communicable diseases, including genetic conditions, mental health issues (e.g., trauma from violence or war), and emerging diseases, identifying geographical areas and patient movements between institutions, facilities and departments remains important. Previous studies have explored integrating geographical components into clinical decision support systems ((Velupillai et al., 2018), (Culotta, 2014)).

The MEDical DOcuments PLACE (MEDDOPLACE) challenge[1], presented at the Iberian Languages Evaluation Forum (IberLEF) 2023[2], aimed to detect geographical, institutional, and functional entities in Span-

[1] https://temu.bsc.es/meddoplace/
[2] https://sites.google.com/view/iberlef-2023/home

ISSN 1135-5948 DOI 10.26342/2026-76-9

ish clinical texts, offering an experimental framework for evaluating systems. This initiative was the first of its kind in this context, and the collected, publicly released corpus stands as the first and only Spanish gold standard for this type of task.

This work focuses on the automatic detection of ten types of location entities in Spanish clinical reports, corresponding to Subtask 1 of the MEDDOPLACE challenge. The task is approached as a sequence labeling problem, and given the strong class imbalance in the corpus, with some categories being scarcely represented, we propose using data augmentation and evaluating both general-domain and biomedical-specific pre-trained models.

The main contributions of this work are threefold. First, we demonstrate that LLM-based paraphrase generation constitutes the most effective data augmentation strategy for this task, outperforming back-translation (BT). Second, we identify medical domain pre-trained models that surpass previously published results. Finally, we show that targeted data augmentation can effectively enhance performance without the need for extensive dataset expansion.

The remainder of the paper is organized as follows: Section 2 provides the background and context for the work. Section 3 details the corpus, the data augmentation strategies, the pre-trained models, and the experimental setup. Section 4 presents the obtained results, which are analyzed in depth in Section 5. Finally, Section 6 outlines the conclusions and future work.

2 Related Work

Named Entity Recognition (NER) is a core subtask of Information Extraction (IE), focused on identifying and classifying entities into predefined categories. In the context of this study, the extraction of specific locations, such as hospital names, medical units, or cities, has been addressed through various NER approaches tailored to the domain and the type of entity (Wang et al., 2018).

In recent years, pre-trained Transformer-based language models, such as BERT (Devlin et al., 2018), which provide a bidirectional contextual representation of language, have significantly enhanced the performance of NER systems in the biomedical domain, owing to their capacity to adapt to specific tasks.

Xu et al. (2025) proposed a NER method for Chinese Electronic Medical Records (EMRs) that integrates ClinicalBERT (Alsentzer et al., 2019)[3], a language model pre-trained on clinical corpora, with structured knowledge from a medical knowledge graph. The proposed model targets entities such as diseases and diagnoses, surgical procedures, anatomical locations, and combines multiple character-level features—including positional labels, contextual category cues, and semantic embeddings—to improve boundary detection. Input texts are annotated using the BIOES (Begin, Inside, Outside, End, Single) tagging scheme and subsequently encoded by ClinicalBERT. The resulting representations are then passed through a bidirectional long short-term memory (BiLSTM) network and a conditional random field (CRF) layer for final label prediction, achieving an F1-score of 89.44% on publicly available datasets.

Beltagy, Lo, and Cohan (2019) introduced SciBERT[4], a pre-trained language model based on BERT and trained on scientific texts from Semantic Scholar. SciBERT was evaluated across a range of tasks and datasets in scientific domains and outperformed other BERT-Base models, such as BioBERT (Lee et al., 2020), which was trained on PubMed abstracts and PMC full-text articles. It achieved new state-of-the-art results on scientific NER tasks in datasets such as BC5CDR (Li et al., 2016). Other models, such as BlueBERT (Peng, Yan, and Lu, 2019)[5], a pre-trained model combining data from PubMed and MIMIC-III to cover both scientific literature and clinical language, have been evaluated on clinical NER tasks, outperforming the general-purpose BERT model across several datasets.

Collectively, these studies highlight the impact of domain-specific pre-training and the integration of structured knowledge for improving NER performance in both clinical and scientific text corpora.

Several pre-trained models exist for the Spanish medical domain. However, one of the main challenges for advancing NER in

[3] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[4] https://huggingface.co/allenai/scibert_scivocab_uncased

[5] https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-12_H-768_A-12

Spanish clinical texts is the scarcity of annotated corpora. Publicly available datasets of annotated medical texts in Spanish for NER-related tasks typically range from a few hundred to a few thousand documents. The MEDDOPLACE dataset[6] represents the only Spanish corpus annotated for location NER in medical texts.

The MEDDOPLACE challenge (Krallinger et al., 2023) was organized as part of the IberLEF 2023 evaluation campaign, held within the XXXIX International SEPLN Conference[7]. Its primary objective was to detect, normalize, and classify location-related mentions in Spanish medical texts. Four subtasks were proposed; however, in this work we focus on the first one: Location Entity Recognition. Two teams submitted proposals for this subtask.

The SINAI group (Chizhikova et al., 2023), which achieved the best results, presented five systems ranging from a pre-trained Spanish biomedical–clinical RoBERTa model (Carrino et al., 2022) fine-tuned on the training dataset, to recurrent BiLSTM+CRF classifiers built on top of contextual embeddings extracted from RoBERTa. They also proposed an ensemble of two recurrent classifiers with the same architecture: one trained on general labels and another on domain-specific clinical labels, applied at both the sentence and document levels. The best performance was obtained with the sentence-level ensemble approach, achieving an F1-score of 0.8512, followed by the BiLSTM+CRF sentence-level classifiers with an F1-score of 0.8387.

The URJC team (Roldán-Álvarez et al., 2023) presented a hybrid strategy based on the combination of a RoBERTa model and a recurrent architecture (BiLSTM). Finally, the outputs of both models were merged into a single file, removing duplicate predictions. The best performance achieved an F1-score of 0.49.

The results of the challenge left several open research directions that we aim to address in this work: text fragmentation into overlapping sentence segments (striding) to improve contextual coverage, the use of models capable of processing longer sequences, and data augmentation techniques to strengthen minority classes.

[6] https://temu.bsc.es/meddoplace/data/

[7] http://sepln2023.sepln.org/

3 Resources and methods

3.1 Corpus

The MEDDOPLACE Gold Standard corpus is a collection of 1,000 clinical case reports in Spanish, covering a wide range of medical specialties such as psychiatry, neurology, travel medicine, infectious diseases, cardiology, occupational medicine, and oncology. The corpus has been annotated, normalized, and classified by a linguist with the support of a clinical expert, ensuring both linguistic and domain accuracy. It is divided into two subsets: a training set comprising 750 clinical texts and a test set containing 250 instances.

Following the Automatic Content Extraction (ACE) framework (Doddington et al., 2004), three main types of locations are distinguished: geopolitical entities (GPE), facilities and buildings (FAC), and geographical locations and features (GEO). Inspired by the BBN Pronoun Coreference and Entity Type Corpus (Weischedel and Brunstein, 2005), each of these types is further classified into proper-name mentions (NOM) and generic mentions (GEN). In addition, four additional location categories are included: DEPARTMENT: encompasses hospital services, specialties, or areas such as "emergency room", which may refer either to physical locations or to organizational units; COMMUNITY: covers sociodemographic information related to the patient's origin or residence, including demonyms, religions, or ethnic groups; TRANSPORT: includes references to patient movements and means of transport, with terms such as "travel", "ambulance", or "transfer"; LANGUAGE: refers both to the languages spoken by the patient and to language barriers that may affect medical care.

Figure 1 illustrates the frequency of each entity type present in the training and test sets. In the training dataset, the most frequent classes are `FAC_GEN` (general facilities), `DEPARTMENT`, and `GPE_NOM` (named geopolitical entities), all exceeding 22% of the mentions. Conversely, the `LANGUAGE` and `GEO_NOM` (named geographical features) classes show a considerably lower frequency, with less than 1% of occurrences, which could potentially hinder their detection during the prediction phase. Overall, the class proportions remain highly similar between the two datasets. Nevertheless, slight differences are

observed; for instance, GPE_NOM and FAC_NOM exhibit a higher percentage in the training set, while FAC_GEN, DEPARTMENT, TRANSPORT, and COMMUNITY have a higher percentage in the test set.

Table 1 details document frequencies and percentages per entity category, as well as total entity counts per type, for each dataset. Consistent with the mention frequencies, the most prevalent categories—DEPARTMENT, FAC_GEN, and GPE_NOM—appear in more than 60% of the documents, while LANGUAGE and GEO_NOM are present in fewer than 5% of the documents.

3.2 Data augmentation

Given the limited availability of data for model training, we propose two data augmentation approaches: back-translation and paraphrase generation. Both methods aim to generate linguistic variations without altering the essential semantic content.

To address both model generalization and class imbalance, we propose two data augmentation strategies:

- **Global application:** The augmentation approaches are applied to the entire dataset to introduce semantic variability and improve the model's overall robustness against overfitting. However, since global augmentation preserves the original class ratios, it is insufficient for strictly mitigating the scarcity of underrepresented labels.
- **Targeted application:** Both approaches are applied exclusively to documents containing the minority classes (GEO_NOM and LANGUAGE). This specific oversampling ensures the model is exposed to more minority instances during training, addressing the imbalance shown in Table 1. As detailed, GEO_NOM appears in only 27 training documents (3.6% of the set; 0.5% of entities) and 10 test documents (4%; 0.4% of entities), while LANGUAGE is present in 26 training (3.5%; 0.5%) and 13 test documents (5.2%; 0.9%).

3.2.1 Back-translation

For the augmentation through back-translation we evaluated two schemes: a standard single-pivot approach (EN-SP) and a chained approach involving two intermediate languages (SP-EN-DE-SP):

- EN-SP Translation: We leveraged the challenge organizer's provision of the dataset in multiple languages, taking the English version (EN) and translating it into Spanish (SP).
- SP-DE-EN-SP Translation: A chained translation was applied to the original Spanish text, first to German (DE), then to English (EN), and finally back to Spanish (SP). The goal of this sequence was to increase the lexical and syntactic diversity.

The `GoogleTranslator` module from the `deep-translator` library[8] was employed for both schemes.

Regarding the realignment of entities after back-translation, a conservative string-matching approach was employed to map the original gold-standard entities onto the newly generated Spanish text. This method ensures that only entities maintaining their semantic and structural integrity are preserved, which explains why the total count of entities in the augmented datasets did not double despite doubling the number of documents.

3.2.2 Paraphrasing using LLMs

The second technique employed was the automatic generation of paraphrases, which involves reformulating the text while preserving its original meaning. To this end, several generative language models were evaluated, with Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) [9] offering the best results in terms of coherence and entity preservation. Paraphrasing was conducted at the document level to maintain clinical narrative integrity. Aligned with the Mistral-7B-Instruct-v0.1 context window (8,192 tokens), reports exceeding architectural limits were fragmented into segments to prevent information loss and ensure medical discourse preservation. While document-level paraphrasing ensures narrative integrity, according to (Li et al., 2022) sentence-level strategies could offer more entity integrity. Various prompts were tested, but the following was ultimately used for the experimentation, without any examples:

```
"[INST] Parafrasea el
siguiente texto
```

[8] https://pypi.org/project/deep-translator/

[9] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

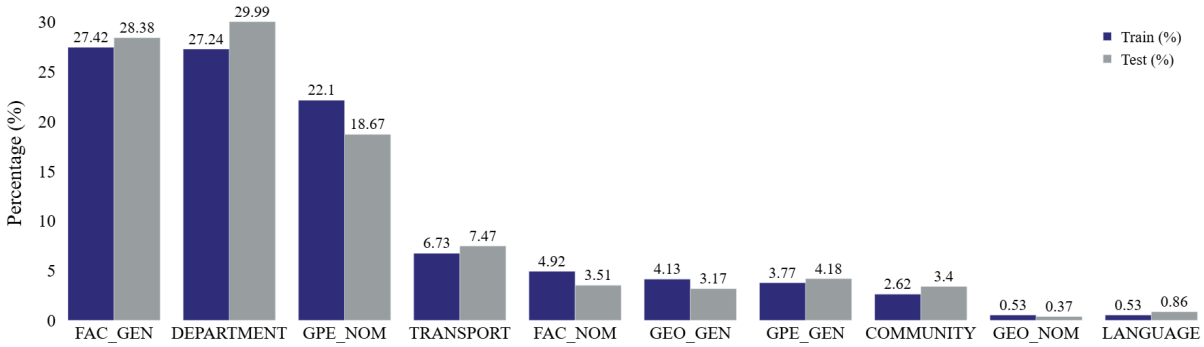

Figure 1: Percentage of occurrence by location entity category and dataset, sorted by frequency in the training set.

Category	Train		Test	
	N. Doc	N. Ent.	N. Doc	N. Ent.
COMMUNITY	136 (18%)	183 (2,6%)	56 (22%)	91 (3,4%)
DEPARTMENT	550 (73%)	1905 (27%)	200 (80%)	803 (30%)
FAC_GEN	533 (71%)	1918 (27%)	184 (74%)	760 (28%)
FAC_NOM	185 (25%)	344 (4,9%)	60 (24%)	94 (3,5%)
GEO_GEN	192 (26%)	289 (4,1%)	63 (25%)	85 (3,2%)
GEO_NOM	27 (3,6%)	37 (0,5%)	10 (4%)	10 (0,4%)
GPE_GEN	179 (24%)	264 (3,8%)	68 (27%)	112 (4,2%)
GPE_NOM	472 (63%)	1546 (22%)	160 (64%)	500 (19%)
LANGUAGE	26 (3,5%)	37 (0,5%)	13 (5%)	23 (0,9%)
TRANSPORT	259 (35%)	471 (7%)	91 (36%)	200 (7%)

Table 1: Document counts, percentages, and total entities per dataset and entity type.

```
en español manteniendo su
significado médico y sin
cambiar las entidades
clínicas:\n"
```

The English Translation of the prompt: "Paraphrase the following text in Spanish, maintaining its medical meaning and without altering the clinical entities".

The following parameters were utilized: `temperature=0.7` and `max_new_tokens=100` to maintain a balance between lexical diversity and the preservation of the original meaning. To ensure the integrity of the location entities during the augmentation process, a dual-stage preservation strategy was implemented. First, the prompt included an explicit negative constraint, instructing the model to maintain the medical meaning without altering the clinical entities. Second, a post-generation quality filter was applied to verify the presence of these entities. This filter acted as a safeguard for entity integrity; any generated paraphrase that failed to retain at least 20% of the original entities was discarded. While this threshold represents a minimum inclusion criterion, it was instrumental in filtering out noisy instances where the LLM failed to adhere to the prompt's constraints. As a result of this filtering, the final paraphrase-augmented dataset was smaller in volume than the back-translation sets.

3.2.3 Comparison of Methods

Table 2 presents the number of texts obtained through each data augmentation approach. For the datasets generated via back-translation, the number of texts was doubled, while the paraphrase-augmented dataset did not reach this volume due to the filtering criteria applied.

In the case of targeted augmentation, only 52 documents containing minority classes were augmented. In the BT EN-SP dataset, all 52 documents were successfully included, resulting in a total of 802 texts. In contrast, for the other augmentation approaches, the minority entities (`LANGUAGE` and `GEO_NOM`)

were lost due to syntactic variations introduced during back-translation or paraphrasing. This effect is particularly pronounced in the Paraphrases dataset, where only 18 of the original 52 documents were retained.

Augmen.	**Dataset**	**N. Docs**
	Original	750
Global	BT EN-SP	1,500 (+750)
	BT SP-DE-EN-SP	1,500 (+750)
	Paraphrases	1,294 (+544)
Targeted	BT EN-SP	802 (+52)
	BT SP-DE-EN-SP	792 (+42)
	Paraphrases	768 (+18)
	SP-DE-EN-SP+Para.	810 (+60)

Table 2: Number of documents per dataset, with augmented data in parentheses.

Table 3 presents the frequency of named entity occurrences for each category in the original dataset and in each of the augmented datasets. Although the number of texts was doubled through back-translation, the entities themselves were not duplicated, as some may have been lost during the transformation process. A similar pattern is observed in the paraphrase-augmented dataset.

Due to the limited data yield from the targeted strategy on the two minority categories, we decided to combine the samples generated via back-translation and paraphrasing, resulting in a fourth evaluation scenario, as shown in Table 3.

3.3 Pre-trained models

Four pre-trained Transformer models were selected to assess the influence of diverse model attributes on the task, based on three main criteria: domain specialization, language focus, and architectural capacity for long-range dependencies.

- **PlanTL-GOB-ES/roberta-base-biomedical-clinical-es (RoBERTa)** (Carrino et al., 2021)[10]: Model based on the RoBERTa architecture, trained on biomedical and clinical Spanish texts, with a maximum sequence length of 512 tokens. It represents the current standard for Spanish clinical Natural Language Processing (NLP) and allows for a direct comparison with previous participants in the MEDDOPLACE challenge.

- **HiTZ/EriBERTa-base (EriBERTa)** (De la Iglesia et al., 2023)[11]: A RoBERTa variant pre-trained by the HiTZ Center[12] on scientific literature and clinical texts in both Spanish and English, with a maximum sequence length of 512 tokens.

- **dccuchile/bert-base-spanish-wwm-cased (BETO)** (Cañete et al., 2020)[13]: The BETO model is an adaptation of BERT for the Spanish language. Although it is not adapted to the clinical domain, it is included to analyze the impact of domain-specific pretraining compared to a robust, general-purpose model.

- **PlanTL-GOB-ES/longformer-base-4096-biomedical-clinical-es (Longformer)** [14]: Model based on the Longformer architecture, which extends RoBERTa to process sequences of up to 4,096 tokens. It has also been trained on biomedical and clinical corpora in Spanish.

As a baseline, a CRF model was selected, which has been widely used in NLP tasks, particularly in sequence labeling, before the rise of neural models (Lafferty, McCallum, and Pereira, 2001).

3.4 Training and evaluation

We fine-tuned each model on the training dataset, allowing it to capture task-specific patterns.

Given that only training and test datasets were provided, a 5-fold cross-validation was performed on the training set (using an 80%-20% split for training and validation in each fold) to optimize hyperparameters and ensure model stability. This procedure allowed for robust model selection before the final evaluation phase. Following the official MEDDOPLACE challenge protocol, the best-performing configurations were then evaluated on the independent test set. This single-run evaluation on the test set ensures a direct and fair comparison with the results

[10] https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es

[11] https://huggingface.co/HiTZ/EriBERTa-base

[12] https://www.hitz.eus/

[13] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[14] https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-biomedical-clinical-es

Category	Ori.	Global Aug.			Targeted Aug.			
		BT EN–SP	BT SP–DE	Para.	BT EN–SP	BT SP–DE	Para.	Para.+BT
COMMU.	183	+134	+120	+70	+40	+30	+4	+34
DEPART.	1905	+1446	+1090	+441	+150	+51	+14	+65
FAC_GEN	1918	+1745	+1084	+496	+126	+51	+16	+67
FAC_NOM	344	+243	+168	+97	+21	+11	+1	+12
GEO_GEN	289	+175	+164	+79	+46	+30	+3	+33
GEO_NOM	37	+35	+20	+13	+37	+26	+11	+37
GPE_GEN	264	+212	+167	+73	+38	+21	+1	+22
GPE_NOM	1546	+1402	+1163	+772	+223	+161	+27	+188
LANGUA.	37	+39	+28	+9	+37	+28	+12	+40
TRANSP.	471	+319	+247	+145	+69	+37	+2	+39

Table 3: Original frequency of entity types and increments introduced by each augmentation method.

obtained by the participating teams, SINAI and URJC.

Transformer hyperparameters were tuned using Optuna (Akiba et al., 2019) to maximize the average F1-score across folds, with 5 trials per configuration. Explored values included: learning_rate [1.5×10^{-5}, 5×10^{-5}] (logarithmic scale); weight_decay 0.0 or [10^{-5}, 5×10^{-2}] (logarithmic scale); warmup_ratio [0, 0.15]; and lr_scheduler_type {linear, cosine}. For the CRF models, the `lbfgs` algorithm was used, while for the Transformer models, early stopping and 10-epoch validation were incorporated. For experiments with the original dataset, default values were used, while experiments on augmented datasets were initialized from the best hyperparameters found on the original dataset.

To improve contextual coverage and reduce information loss, sentence-level segmentation with overlapping windows (striding) was applied before tokenization. The BIO (Beginning - Inside - Outside) annotation scheme was used.

Evaluation was conducted by comparing predictions and references using the official scoring script provided by the organizers[15], reporting micro-averaged Precision, Recall, and F1-score. Both strict and overlapping versions of these metrics were provided. The overlapping metrics consider a prediction correct if it partially intersects with the reference span, assuming the entity label matches. The strict metric F1-score was used to rank the participants.

[15] `https://github.com/TeMU-BSC/meddoplace_scoring_script`

4 Results

Table 4 presents the Precision, Recall, and F1-score obtained by the five models on the original test set, comparing the impact of training on the original data versus the augmented variants. Overall, the results show a substantial performance gap between the traditional CRF and Transformer-based models.

Among the Transformer-based models, EriBERTa, BETO, and Longformer achieve similarly competitive results, with F1-score values ranging from 0.85 to 0.87 across datasets.

In the Global application scenario, Longformer achieves the highest overall performance, reaching 0.8715 F1-score with paraphrased data. In contrast, augmentation via back-translation leads to performance drops, particularly in the BT EN-SP, BETO exhibits a similar pattern: paraphrasing yields the best results, while back-translation negatively impacts performance. In contrast, EriBERTa performs best on the original dataset. RoBERTa consistently lags behind the other Transformer-based models, though paraphrasing substantially improves its F1-score.

Regarding augmentation strategies, paraphrasing emerges as the most effective technique for Transformer models, enhancing F1-score for RoBERTa, BETO, and Longformer. In contrast, back-translation methods are generally less beneficial, typically degrading performance, except in specific cases such as the CRF model or precision improvements in EriBERTa.

Regarding Targeted aplication scenario, in which only the documents containing the minority classes `LANGUAGE` and `GEO_NOM` (named geographical features) were augmented, we

Model	Metric	Origi.	Global Aug.			Targeted Aug.			
			BT EN-SP	BT SP-DE-EN-SP	Para.	BT EN-SP	BT SP-DE-EN-SP	Para.	Para. + BT SP-DE-EN-SP
CRF	Precision	0.6232	0.6513	0.7371*	0.6316	0.6208	0.6412	0.6252	0.6351
	Recall	0.3712	0.3118	0.2984	0.3727	0.3742	0.4663*	0.3719	0.3652
	F1	0.4652	0.4217	0.4248	0.4688*	0.4669	0.3663	0.4664	0.4637
RoBERTa	Precision	0.7932	0.8108	0.8142	0.8361	0.7867	0.7881	0.8385*	0.7881
	Recall	0.8514	0.8571	0.8665	0.8727*	0.8533	0.8537	0.8646	0.8423
	F1	0.8213	0.8333	0.8396	0.8540*	0.8186	0.8196	0.8515	0.8143
EriBERTa	Precision	0.8420	0.8494*	0.8455	0.8379	0.8408	0.8477	0.8461	0.8454
	Recall	0.8835*	0.8678	0.8704	0.8824	0.8738	0.8712	0.8786	0.8738
	F1	0.8622*	0.8585	0.8578	0.8596	0.8570	0.8593	0.8621	0.8593
BETO	Precision	0.8508	0.8481	0.8538	0.8522	0.8548	0.8479	0.8568*	0.8539
	Recall	0.8749	0.8611	0.8656	0.8764	0.8727	0.8596	0.8805*	0.8641
	F1	0.8627	0.8545	0.8596	0.8641	0.8636	0.8537	0.8685*	0.8589
Longformer	Precision	0.8503	0.8385	0.8490	0.8573	**0.8651**	0.8560	0.8528	0.8630
	Recall	0.8760	0.8648	0.8626	0.8861	0.8764	0.8768	0.8827	**0.8869**
	F1	0.8630	0.8515	0.8557	0.8715	0.8707	0.8663	0.8675	**0.8748**

Table 4: Comparison of Precision, Recall, and F1-score on the test set. The best results per dataset are underlined, while the best results per model are marked with an asterisk (*). The overall best results are highlighted in bold.

find that back-translation EN-SP achieves better results than Global application back-translation for BETO and Longformer, with Longformer improving the previous best F1-score (0.8707) and reaching the highest overall Precision (0.8651). In connection to back-translation, SP-DE-EN-SP further enhance performance for EriBERTa and Longformer compared to global augmentation. Paraphrasing improves BETO, achieving the best results for this model overall (0.8685), while EriBERTa reaches its best score among the augmented datasets, although it remains slightly below its performance on the original data.

Combining paraphrasing with back-translation yields the highest global Recall (0.8869) and F1-score (0.8748) with Longformer. Overall, Longformer benefits the most from the Targeted application, with BETO also exhibiting improvements, whereas RoBERTa is negatively impacted. BETO's good performance is particularly noteworthy given its nature as a general-domain model.

Interestingly, the targeted augmentation strategy outperformed the global approach; however, the improvement was not limited to the minority classes themselves. The most notable improvements occurred in the majority classes, such as `DEPARTMENT`, `FAC_GEN`, and `GPE_NOM`. This suggests that the targeted augmentation not only helps to slightly rebalance underrepresented classes but also provides additional syntactic variability that benefits the learning of robust decision boundaries across the dataset.

Table 5 presents the ranked list of our best systems alongside the challenge participants (SINAI and URJC). It can be observed that the Longformer, BETO, and EriBERTa models outperform the results obtained by the participants using RoBERTa; notably, EriBERTa surpasses them even without data augmentation.

5 Discussion

The experimental results indicate that the impact of data augmentation is not uniform across all evaluated architectures. While the ROBERTa model showed its best performance using global paraphrasing, achieving an F1-score of 0.8540, it was negatively impacted by the targeted augmentation approach. In contrast, the Longformer model achieved its highest overall results (0.8748 F1-score) precisely through the targeted application of paraphrasing and back-translation. These findings suggest that although targeted augmentation introduces beneficial syntactic variability for the learning of decision boundaries in some models, its effectiveness is highly dependent on the specific pairing of the augmentation strategy with the model's architecture. Consequently, the choice between global and targeted augmentation should be carefully considered based on the model being employed.

To interpret the results, a more detailed analysis was conducted on the two top-performing models from the evaluation set:

	Precision	Recall	F1
Longf. Targ. Aug. Para. + BT	0.8630	0.8869	0.8748
Longf. Glo. Aug. Para.	0.8573	0.8861	0.8715
BETO - Targ. Aug. Para.	0.8568	0.8805	0.8685
EriBERTa - Ori.	0.8420	0.8835	0.8622
RoBERTa - Para.	0.8361	0.8727	0.8540
SINAI*	0.8639	0.8391	0.8512
URJC*	0.43	0.57	0.49
CRF - Para.	0.6316	0.3727	0.4688

Table 5: Strict metrics of the best proposed models compared to participating systems (*) in the MEDDOPLACE challenge, sorted by F1-score.

(i) Longformer trained with paraphrase and BT using the Targeted approach, and (ii) Longformer trained with paraphrase augmentation using the Global approach. Figure 2 shows the confusion matrices for the two best models. The most frequent errors involve distinguishing entity tokens from non-entity tokens (labeled as O). Conversely, inter-class confusions are rare, demonstrating that Longformer has learned to effectively differentiate between entity categories.

Interestingly, data augmentation targeting the two minority categories boosts performance on well-represented categories (`DEPARTMENT`, `FAC_GEN`, `GPE_NOM`, `TRANSPORT`) and even on other less frequent ones (`COMMUNITY`, `FAC_NOM`, `GEO_GEN`, `GPE_GEN`). Yet, it only benefits one of the targeted minority classes, `LANGUAGE`, while the other, `GEO_NOM`, is negatively affected.

Among the entity errors, the most common is confusing `GEO_GEN` (general geographical features) with `FAC_GEN` (general facilities), due to the presence of terms related to outdoor or camping environments, such as "road", "camping", "cabin", "artificial grass field", or "pasture area", which may share semantic context.

Another frequent error, labeling a `FAC_NOM` (named facilities) as `DEPARTMENT`, occurred on all three occasions because of the entity "UME Galatzó". This entity refers to a psychiatric Medium-Stay Unit in Palma de Mallorca, which is a specific healthcare facility rather than a department. `COMMUNITY` and `LANGUAGE`, two entities that can be closely related, are also sometimes confused, as in errors such as "English", "American" and "Italian".

Table 6 compares the lengths of correct and incorrect predictions. Model errors tend to occur in longer sequences. Correct predictions have an average length of 12.42 tokens (max 77), whereas incorrect ones have a higher average of 18.99 tokens (max 103). These results indicate that the model struggles more with longer sequences due to their syntactic complexity.

The lower performance of back-translation could be linked to semantic drift and the limitations of recovering entities through string matching after the translation process. Regarding the latter, our analysis reveals frequent cases of truncated entities where the model only predicts a fragment of the full span, such as 'Hospital of', 'University of', or 'Foundation'. These truncations suggest that syntactic reordering and lexical variation during the translation chain hinder string matching of the original multi-word entity spans, often causing the loss of final tokens. Furthermore, semantic drift of the surrounding context is evidenced by the misclassification of entities like 'zona de potrero' (pasture area), which was correctly labeled as `FAC_GEN` but predicted as `GEO_GEN`. This indicates that the translation process can shift the linguistic context of a term from a functional facility to a natural feature, thereby introducing noise that affects the model's ability to distinguish between semantically related categories.

	freq.	mean	std	max
Correct	2744	12.42	8.78	77
Incorrect	304	18.99	14.85	103

Table 6: Frequency, average, standard deviation and maximum token lengths for correct and incorrect predictions.

6 Conclusions

The experiments clearly show that Transformer-based models outperform the CRF, maintaining F1-score above 0.82,

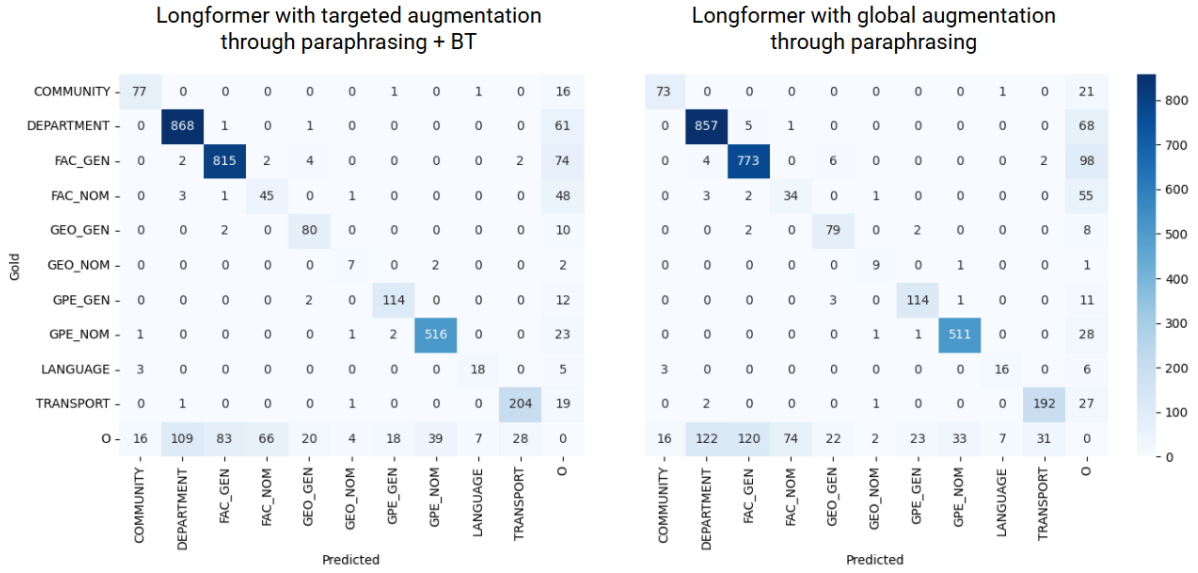

Figure 2: Confusion matrix of the best two models.

while the CRF reaches a maximum F1-score of only 0.4688.

The Longformer trained on the dataset augmented with Targeted strategy with paraphrase and BT achieves the best overall performance (0.8748). By providing a significantly larger context window than the 512-token limit of architectures like EriBERTa or RoBERTa, it seems that the Longformer is better positioned to manage the syntactic complexity of clinical reports and capture document-level dependencies that are often lost in shorter-context models. BETO performs competitively (0.8685) and excels in paraphrase-augmented datasets, while EriBERTa achieves the top F1-score (0.8622) on the original data. RoBERTa yields the lowest results among the Transformers, with its highest F1-score (0.8540) obtained on the paraphrase dataset.

Data augmentation demonstrates clear benefits. Paraphrasing improves F1-score across all models, whereas back-translation often reduces performance, likely due to semantic drift or entity misalignment. Notably, SP-DE-EN-SP back-translation outperforms the standard EN-SP approach for most models, suggesting that using three languages in the BT augmentation introduces linguistic variability effectively.

Targeted augmentation for minority classes (`LANGUAGE` and `GEO_NOM`) further enhances performance, with strategies combining paraphrasing and back-translation yielding the highest Recall (0.8869) and F1-score (0.8748) for Longformer. This demonstrates that augmenting the entire dataset is not necessary to achieve superior results. Overall, Longformer benefits most from these targeted approaches and BETO shows improvements, whereas RoBERTa is negatively affected.

Finally, error analysis reveals that the main limitations are incomplete or overextended entity spans, particularly for long entities, and confusion between semantically related classes.

Future research will include improving the recognition of long, nested, and discontinuous entities, evaluating sentence-level paraphrasing to optimize entity preservation and explore alternatives beyond the BIO format. We will also continue exploring class-balancing techniques through data augmentation.

Acknowledgements

This work was supported by the projects EDHER-MED (PID2022-136522OB-C21; MCIN/AEI/10.13039/501100011-033/FEDER, UE), funded by the Spanish State Research Agency and the Spanish Ministry of Science, Innovation and Universities.

References

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th*

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2623–2631.

Alsentzer, E., J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Beltagy, I., K. Lo, and A. Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Carrino, C., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, and M. Villegas. 2022. Pretrained biomedical language models for clinical nlp in spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199.

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

Cañete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Papers with Code*.

Chen, Q., R. Leaman, A. Allot, L. Luo, C.-H. Wei, S. Yan, and Z. Lu. 2021. Artificial intelligence (ai) in action: Addressing the covid-19 pandemic with natural language processing (nlp). *arXiv preprint arXiv:2010.16413*.

Chizhikova, M., J. Collado Montañez, M. C. Díaz-Galiano, L. A. Ureña-López, and M. T. Martín Valdivia. 2023. Sinai@meddoplace: Detecting, normalizing, and classifying places and related information in spanish medical texts. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with SEPLN 2023*, CEUR Workshop Proceedings, Jaén, Spain.

Culotta, A. 2014. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122.

De la Iglesia, I., A. Atutxa, K. Gojenola, and A. Barrena. 2023. EriBERTa: A Bilingual Pre-Trained Language Model for Clinical Natural Language Processing.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2023. Mistral 7b.

Krallinger, M., S. Lima Lopez, E. Farré Maduell, L. Gascó Sánchez, and V. Briva Iglesias. 2023. Meddoplace shared task on location entity recognition in spanish clinical narratives. IberLEF 2023.

Lafferty, J. D., A. McCallum, and F. C. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 282–289.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Li, J., Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05.

Peng, Y., S. Yan, and Z. Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Roldán-Álvarez, D., M. Ángel Rodríguez-García, S. Montalvo-Herranz, and R. Martínez-Unanue. 2023. Urjc-team at

meddoplace 2023: Bi-lstm and transformers for medical document place-related content extraction. In *Proceedings of IberLEF 2023*, Jaén, Spain. CEUR Workshop Proceedings. September 2023.

Velupillai, S., H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, and R. Dutta. 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics*, 88:11–19.

Wang, Y., L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu. 2018. Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, 77:34–49.

Xu, X., Z. Li, H. Zhang, and K. Ma. 2025. Named entity recognition for chinese electronic medical records by integrating knowledge graph and clinicalbert. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.