Multilingual Clinical NER for Diseases and Medications Recognition in Cardiology Texts using BERT Embeddings

Manuela Daniela Danu[1,2,*], George Marica[1], Constantin Suciu[1,2], Lucian Mihai Itu[1,2] and Oladimeji Farri[3]

[1]Advanta, Siemens SRL, 15 Noiembrie Bvd, 500097 Brasov, Romania
[2]Automation and Information Technology, Transilvania University of Brasov, 5 Mihai Viteazul Street, 500174 Brasov, Romania
[3]Digital Technology and Innovation, Siemens Healthineers, 755 College Rd E, 08540 Princeton, NJ, United States

Abstract

The rapidly increasing volume of electronic health record (EHR) data underscores a pressing need to unlock biomedical knowledge from unstructured clinical texts to support advancements in data-driven clinical systems, including patient diagnosis, disease progression monitoring, treatment effects assessment, prediction of future clinical events, etc. While contextualized language models have demonstrated impressive performance improvements for named entity recognition (NER) systems in English corpora, there remains a scarcity of research focused on clinical texts in low-resource languages. To bridge this gap, our study aims to develop multiple deep contextual embedding models to enhance clinical NER in the cardiology domain, as part of the BioASQ MultiCardioNER shared task. We explore the effectiveness of different monolingual and multilingual BERT-based models, trained on general domain text, for extracting disease and medication mentions from clinical case reports written in English, Spanish, and Italian. We achieved an F1-score of 77.88% on Spanish Diseases Recognition (SDR), 92.09% on Spanish Medications Recognition (SMR), 91.74% on English Medications Recognition (EMR), and 88.9% on Italian Medications Recognition (IMR). These results outperform the mean and median F1 scores in the test leaderboard across all subtasks, with the mean/median values being: 69.61%/75.66% for SDR, 81.22%/90.18% for SMR, 89.2%/88.96% for EMR, and 82.8%/87.76% for IMR.

Keywords

MultiCardioNER, BioASQ, Cardiology, Named Entity Recognition, NER, unstructured data, BERT, Multilingual, English, Spanish, Italian

1. Introduction

With the increasing amount of available electronic health record (EHR) data, clinical natural language processing (NLP) tasks have become significantly important for extracting valuable information from unstructured clinical texts [1]. Named Entity Recognition (NER) is a key NLP task used to identify meaningful entities within these texts, such as anatomical structures, diseases and disorders, signs and symptoms, procedures, and medications [1, 2]. Consequently, this facilitates various data analysis applications, ranging from predicting future clinical events [3] to summarization [4] and relation extraction between entities (e.g., drug-to-drug interactions [5], symptom-disease relationship [6], patient-procedure association [7], etc.)

Despite recent advances in deep learning methods for NER [8, 9], extracting structured information from the vast amounts of unstructured and often noisy clinical documents in EHR systems remains challenging due to the highly specialized medical language, which varies considerably across different medical specialties, as well as due to the prevalence of misspellings, abbreviations, and use of synonyms to express clinical concepts [1].

While contextualized language models have recently improved the performance of NER systems for English corpora [8, 10, 11], there is a notable lack of research focused on clinical texts in low-resource

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
Disclaimer: The concepts and information presented in this paper are based on research results that are not commercially available.
*Corresponding author.
manuela.voinea@siemens.com (M. D. Danu); george.marica@siemens.com (G. Marica); constantin.suciu@siemens.com (C. Suciu); lucian.itu@siemens.com (L. M. Itu); oladimeji.farri@siemens-healthineers.com (O. Farri)

languages. To address this gap, our study aims to develop multiple deep contextual embedding models for English, Spanish, and Italian to enhance clinical NER in the cardiology domain, as part of the MultiCardioNER shared task [12, 13, 14]. The MultiCardioNER task is part of the twelfth edition of the large-scale biomedical semantic indexing and question answering challenge (BioASQ) [13, 14], a long-standing initiative aiming to advance research by developing methods and tools that leverage the vast amount of online information to meet the needs of biomedical researchers and practitioners. This initiative seeks to provide efficient and rapid access to the continuously expanding resources and knowledge in the biomedical field.

MultiCardioNER [12, 13, 14] is a shared task that aims to automatically identify two key clinical concepts in medical documents pertaining to cardiology, namely diseases and medications. This task focuses on adapting clinical NER systems to effectively work across multiple languages - primarily Spanish, English, and Italian - for two different subtasks: (1) diseases recognition in Spanish cardiology texts, and (2) medications recognition in cardiology texts written in Spanish, English, and Italian. Both subtasks involve reading and analyzing clinical texts to identify the clinical entities mentioned in the text and using the BRAT format to mark the starting and ending positions of these entities.

In this paper, we created four different monolingual models: (1) Spanish Diseases Recognition (SDR), (2) Spanish Medications Recognition (SMR), (3) English Medications Recognition (EMR), and (4) Italian Medications Recognition (IMR). Additionally, we developed two multilingual models: one specialized for Spanish Diseases Recognition (Multi-SDR) and another for Medications Recognition across all three targeted languages (Multi-MMR). We applied transfer learning techniques by fine-tuning BERT-based [15] contextual embeddings, originally trained on general domain text in each of the three languages, for the biomedical domain to extract diseases and medications from clinical reports.

2. Related Work

In clinical and biomedical NER, recent studies have explored various methodologies to enhance performance. A key model in this domain is multilingual BERT (M-BERT) [15], trained on 104 Wikipedia languages, which excels in various tasks without explicit cross-lingual alignment [16], outperforming models based on cross-lingual embeddings [17].

[18] improved biomedical NER by incorporating syntactic information, enhancing recognition of complex entity relationships (ORCID). [19] focused on de-identifying Spanish medical texts via NER and entity randomization, achieving high recall rates on radiology reports and MEDDOCAN [20] challenge data. [9] developed BioELECTRA, a biomedical text encoder using discriminators, which outperformed several baselines on multiple biomedical NER benchmarks by leveraging ELECTRA's efficiency and accuracy in text encoding.

[21] developed a scalable NER system for large biomedical datasets, emphasizing real-time processing and high accuracy. [22] focused on pre-trained biomedical language models for clinical NLP in Spanish, addressing the need for multilingual capabilities in biomedical NER. [23] optimized a bi-encoder for NER using contrastive learning, introducing dynamic thresholding to improve accuracy, especially for nested entities, with significant gains on datasets like ACE [24] and GENIA [25]. [26] used a novel schema with distant supervision to enhance NER accuracy, showing that domain-specific schema can supplement limited annotated data effectively.

[27] used ChatGPT [28] for zero-shot clinical entity recognition with prompt engineering, showing it outperforms GPT-3 [29] but trails behind fine-tuned BioClinicalBERT [10] models. [30] leveraged transfer learning and asymmetric tri-training, combining labeled and pseudo-labeled data to boost NER performance across biomedical datasets.

To advance the development of medical NER systems, the BioASQ challenge proposed multiple clinical NER tasks to be solved over time, such as automatic detection and normalization of disease mentions from clinical texts (DisTEMIST) [31] or medical procedure detection and entity linking (MedProcNER) [32]. Most participating teams employed Transformer-based and large language models in their approaches.

3. Methods

3.1. Datasets

With a focus on adapting general medical NER systems for diseases and medications across multiple languages, the MultiCardioNER [12, 33] task leverages several datasets. Specifically, it utilizes a training collection of 1000 general clinical case reports in Spanish, covering various medical specialties such as oncology, urology, ophthalmology, dentistry, pediatrics, primary care, allergology, radiology, psychiatry, and more [33, 31]. These reports were annotated with diseases and medications, resulting in two distinct corpora, namely DisTEMIST [33, 31, 34] and DrugTEMIST [33]. The DrugTEMIST [33] corpus was also released in English and Italian. Since the original 1000 clinical case reports belong to the Spanish Clinical Case Corpus (SPACCC) [35], the multilingual DrugTEMIST [33] dataset was originally created in Spanish and then transferred into English and Italian using machine translation and lexical annotation projection. The result of this process was revised and validated by clinical experts who are native speakers of each language.

For the domain adaptation part of the task, MultiCardioNER [33] leverages a collection of 508 annotated cardiology clinical case reports (CardioCCC), divided into 258 for development and 250 for testing. The annotation process followed the same guidelines as the DisTEMIST [36] and DrugTEMIST [37] corpora, with the medication part also released in Spanish, English and Italian. In addition to the test set, an auxiliary collection of multilingual clinical case reports, referred to as the background set, is provided to facilitate the creation of a silver standard corpus and ensure the developed systems can effectively scale up to larger content collections.

All datasets were manually annotated by clinical experts using the BRAT annotation tool [38], following well-defined annotation guidelines [36, 37] defined after several cycles of quality control and annotation consistency analysis.

3.2. Experiments

In this work, we treated the automatic named entity recognition (NER) of diseases and medications in clinical case reports as a multi-label token classification task. To accomplish this, we employed pre-existing BERT models [15] for NER in the general domain for each of the three languages (Spanish, English, and Italian), as well as a multilingual model, and further fine-tuned them for the biomedical domain using the MultiCardioNER dataset [33].

We experimented with the following BERT-based models, specifically trained to perform NER:

- **bert-spanish-cased-finetuned-ner** [39]: a Spanish BERT cased model based on BETO [40]. Originally fine-tuned on the Spanish dataset of the CoNLL-2002 Shared Task [41], BETO was further fine-tuned on the Catalan and Basque subsets of the CoNLL-2007 dataset [42], resulting in the bert-spanish-cased-finetuned-ner model, which focuses on recognizing persons (PER), organizations (ORG), locations (LOC), and miscellaneous (MISC) entities within Spanish text documents.
- **bert-base-NER** [43]: a BERT cased model fine-tuned on the English version of the standard CoNLL-2003 dataset [44]. It was trained to recognize four types of entities, namely locations (LOC), organizations (ORG), persons (PER), and miscellaneous (MISC).
- **bert-italian-finetuned-ner** [45]: an Italian BERT cased model fine-tuned on the WikiANN dataset [46], which consists of Wikipedia articles annotated with LOC (location), PER (person), and ORG (organisation) tags.
- **bert-base-multilingual-cased-ner-hrl** [47]: a named entity recognition model for 10 high-resourced languages (Arabic, German, English, Spanish, French, Italian, Latvian, Dutch, Portuguese and Chinese) based on a fine-tuned multilingual cased BERT model. It has been trained to recognize three types of entities: locations (LOC), organizations (ORG), and persons (PER).

All these BERT-based models utilize the standard Beginning-Inside-Outside (BIO) format [48] for tagging entities. This format is crucial as it allows NER to be approached as a multi-label classification

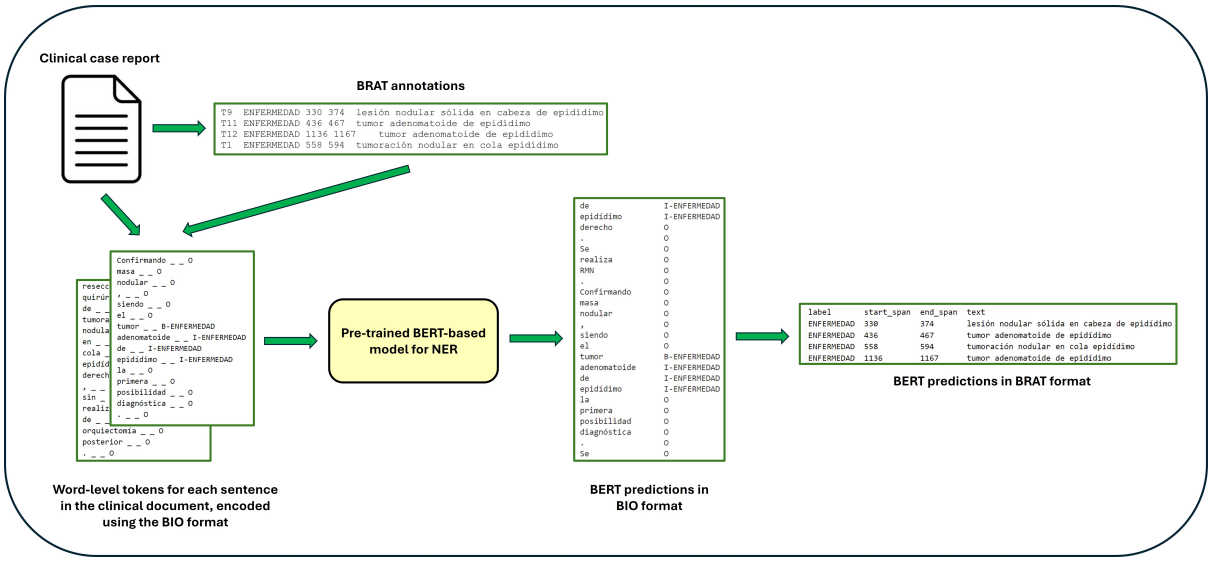

Figure 1: Overview of the prediction pipeline. As a pre-processing step, the clinical case reports are split into sentences, and further segmented into word-level tokens. By leveraging the available BRAT annotations, the word-level tokens are encoded using the BIO format and used to fine-tune BERT-based models on the MultiCardioNER dataset. The output from the BERT models, also in BIO format, is then post-processed to comply with BRAT format.

task, where words are labelled B if they represent the beginning of an entity, I if they are inside an entity, and O if they are outside any entity. This labeling method effectively distinguishes between the beginning and continuation of an entity, thereby simplifying the task of identifying entity boundaries.

Before performing the clinical domain adaptation of the general domain BERT-based models, the medical reports undergo a pre-processing step which involves splitting them into sentences to ensure a sequence length of less than or equal to 256. These sentences are then further segmented into word-level tokens while preserving their start and end offsets with respect to the original report. The word-level tokens are encoded in BIO format and used to fine-tune BERT-based models on the MultiCardioNER dataset. The output from the BERT models is then post-processed to comply with BRAT format. Figure 1 provides an overview of the prediction pipeline.

Details regarding the label lists used for each subtask, as well as the hyper-parameters configuration employed in the experiments, are provided in sections 3.2.1 and 3.2.2, respectively.

3.2.1. Subtask 1: Diseases Recognition in Spanish Cardiology Texts

For the subtask aiming to address the recognition of diseases in Spanish cardiology texts, we leveraged the pre-trained *bert-spanish-cased-finetuned-ner* and *bert-base-multilingual-cased-ner-hrl* models and further fine-tuned them on the MultiCardioNER dataset. Specifically, we employed the DisTEMIST corpora as the training set for the general clinical domain adaptation part of the task and used the disease-annotated version of the Spanish CardioCCC clinical cases as the development set to identify the best-performing models in the cardiology domain, resulting in the Clinical-SDR and MultiClinical-SDR models. We additionally experimented with fine-tuning these models on the CardioCCC development set, leading to the creation of the cardiology-specialized Cardio-SDR and MultiCardio-SDR models.

Following the standard BIO format [48], we defined our label list as follows: B-ENFERMEDAD, I-ENFERMEDAD, O, [CLS], and [SEP]. B-ENFERMEDAD and I-ENFERMEDAD denote the beginning and continuation of disease mentions within text sequences, whereas the label O corresponds to word-level tokens outside any recognized entity. Additionally, the [CLS] token indicates the commencement of a sentence, while the [SEP] token marks its termination. Notably, the [CLS] token also serves as a placeholder for the [PAD] token within the text sequences.

The Spanish Diseases Recognition (SDR) models were fine-tuned on an NVIDIA GeForce RTX 3090

(24GB) GPU for 10 epochs. The utilized hyper-parameters configuration includes a maximum sequence length of 256, a batch size of 8, and a learning rate of $9e^{-6}$. Predictions were generated for both the test and background sets, but the evaluation exclusively considered the predictions achieved on the test set. In addition to the test set results, we also reported the development set results to identify any discrepancies between them and, hence, detect potential overfitting or any issues related to the data split distributions.

3.2.2. Subtask 2: Multilingual (Spanish, English and Italian) Medications Recognition in Cardiology Texts

For the second subtask, which focuses on the recognition of medications in cardiology texts written in Spanish, English, and Italian, we employed three monolingual pre-trained models (*bert-spanish-cased-finetuned-ner*, *bert-base-NER*, and *bert-italian-finetuned-ner*), each specialized for one of the three languages, as well as a multilingual model (*bert-base-multilingual-cased-ner-hrl*), and subsequently fine-tuned them on the MultiCardioNER dataset. Therefore, we leveraged the DrugTEMIST corpora in each of the three languages as the training sets for the general clinical domain adaptation part of the task and used the medication-annotated version of the CardioCCC clinical cases in Spanish, English, and Italian as the development sets to identify the best performing models in the cardiology domain, resulting in the Clinical-SMR, Clinical-EMR, Clinical-IMR, and MultiClinical-MMR models. We again conducted additional experiments by fine-tuning these models on the CardioCCC development sets, thereby achieving the cardiology-specialized Cardio-SMR, Cardio-EMR, Cardio-IMR, and MultiCardio-MMR models. It is worth noting that the multilingual model was trained on an aggregated dataset encompassing all three languages, but separately evaluated for each language to assess its performance across different linguistic contexts.

In accordance with the standard BIO format [48], we defined the label list for this subtask as B-FARMACO, I-FARMACO, O, [CLS], and [SEP]. The tags B-FARMACO and I-FARMACO denote the beginning and continuation of medication mentions within text sequences, while the label O marks the word-level tokens not associated with any recognized entity. Additionally, as in Subtask 1, the [CLS] token indicates the beginning of a sentence, while the [SEP] token marks its end. In this context, the [CLS] token also serves as a placeholder for the [PAD] token within the text sequences.

The Spanish Medications Recognition (SMR), English Medications Recognition (EMR), Italian Medications Recognition (IMR), and Multilingual Medications Recognition (MMR) models were independently fine-tuned on an NVIDIA GeForce RTX 3090 (24GB) GPU for 10 epochs. The utilized hyper-parameters configuration is identical to that employed for Subtask 1 and consists of a maximum sequence length of 256, a batch size of 8, and a learning rate of $9e^{-6}$. Predictions were generated for both the test and background sets. However, the evaluation exclusively considered the predictions obtained on the test set. In addition to the test set results, we also reported the development set results to identify any mismatch between them, which could indicate overfitting or issues related to the data split distributions.

4. Results

In this work, we evaluated the developed systems using a flat evaluation approach [49] by comparing the automatically generated results with those obtained by domain experts through manual annotation. The primary focus was on identifying and classifying clinical mentions of diseases and medications in cardiology reports. The performance metrics employed for flat evaluation include micro-averaged precision, recall, and F1-score (MiF). These metrics were computed based on the exact matches of the predicted entities and the annotated ground-truth. Table 1 summarises the evaluation results obtained on the development and test sets using the official-released evaluation library for the MultiCardioNER task. In the test set evaluation, we achieved the following F1-scores: 77.88% for Spanish Diseases Recognition (SDR), 92.09% for Spanish Medications Recognition (SMR), 91.74% for English Medications Recognition (EMR), and 88.9% for Italian Medications Recognition (IMR).

Table 1
Evaluation results on the development and test sets for the MultiCardioNER task. The best results on the test sets are highlighted in bold. The experiments marked with an (*) were conducted after the MultiCardioNER evaluation period and are not included in the official leaderboard.

Subtask	Model	Fine-tuning	Dev Precision	Dev Recall	Dev F1-score	Test Precision	Test Recall	Test F1-score
Track1 (ES)	Clinical-SDR	No	0.6674	0.6243	0.6451	0.6758	0.6437	0.6593
Track1 (ES) *	Cardio-SDR	Yes	0.9713	0.9535	0.9623	**0.7739**	**0.7837**	**0.7788**
Track1 (ES) *	MultiClinical-SDR	No	0.6355	0.6118	0.6234	0.6387	0.6268	0.6327
Track1 (ES) *	MultiCardio-SDR	Yes	0.9406	0.9360	0.9383	0.7717	0.7788	0.7753
Track2 (ES)	Clinical-SMR	No	0.9019	0.8753	0.8884	0.8928	0.8778	0.8852
Track2 (ES) *	Cardio-SMR	Yes	0.9804	0.9562	0.9681	0.9289	0.9045	0.9165
Track2 (ES) *	MultiClinical-MMR	No	0.8783	0.8681	0.8732	0.8974	0.8807	0.8890
Track2 (ES) *	MultiCardio-MMR	Yes	0.9790	0.9482	0.9634	**0.9341**	**0.9080**	**0.9209**
Track2 (EN)	Clinical-EMR	No	0.8866	0.8625	0.8744	0.8685	0.8791	0.8738
Track2 (EN) *	Cardio-EMR	Yes	0.9575	0.9155	0.9360	**0.9277**	0.9018	0.9146
Track2 (EN) *	MultiClinical-MMR	No	0.8833	0.8594	0.8712	0.8920	0.8826	0.8873
Track2 (EN) *	MultiCardio-MMR	Yes	0.9681	0.9550	0.9615	0.9121	**0.9227**	**0.9174**
Track2 (IT)	Clinical-IMR	No	0.9122	0.8801	0.8958	0.8891	0.8689	0.8789
Track2 (IT) *	Cardio-IMR	Yes	0.9518	0.9250	0.9382	0.8994	**0.8789**	**0.8890**
Track2 (IT) *	MultiClinical-MMR	No	0.8868	0.8603	0.8734	0.8747	0.8378	0.8558
Track2 (IT) *	MultiCardio-MMR	Yes	0.9772	0.9455	0.9611	**0.9046**	0.8694	0.8867

These results surpass the mean and median F1 scores in the test leaderboard across all subtasks, with the mean/median values being: 69.61%/75.66% for SDR, 81.22%/90.18% for SMR, 89.2%/88.96% for EMR, and 82.8%/87.76% for IMR.

The experiments marked with an (*) in Table 1 were conducted after the MultiCardioNER evaluation period and are not included in the official leaderboard. However, these supplementary experiments provide further insights beyond the primary evaluation results. For instance, the fine-tuning process considerably enhances performance across all developed systems. Additionally, employing a multilingual model proves beneficial in certain substasks, such as Spanish Medications Recognition (SMR) and English Medications Recognition (EMR), resulting in an improved F1-score from 91.65% (achieved by the subsequent best performing model) to 92.09%, and from 91.46% to 91.74%, respectively.

By comparing the results from the development and test sets, we can assess potential discrepancies between these data splits and identify issues such as overfitting or distributional disparities. These insights are crucial for enhancing model robustness and generalization, which are essential for successfully utilizing the developed systems in real-world clinical scenarios. As illustrated in Table 1, non-fine-tuned models exhibit similar evaluation metrics on both the development and test sets. For these models, the development set was solely used to select the best-performing model across different checkpoints. This consistency confirms that the two data splits originate from the same distribution. In contrast, fine-tuned models – trained on the development set – demonstrate a performance gap between the two sets. While some degree of performance difference is expected due to the model's exposure to the development data during training, excessively large gaps suggest overfitting. This is the case of Spanish Diseases Recognition (SDR) models, where the performance gap between the development and test sets is 18.35% for Cardio-SDR and 16.3% for MultiCardio-SDR. For all other fine-tuned models, the F1-score on the development set is only slightly higher than that computed on the test set, with differences ranging from 1.95% to 7.44%. Although these differences may indicate some overfitting, they do not reach a severe extent. One plausible explanation for overfitting in these cases could be that the model is too complex for the limited diversity of cardiology-specific entities present in the development set. As a result, the model may capture specific patterns from the training data but struggle to generalize to new data.

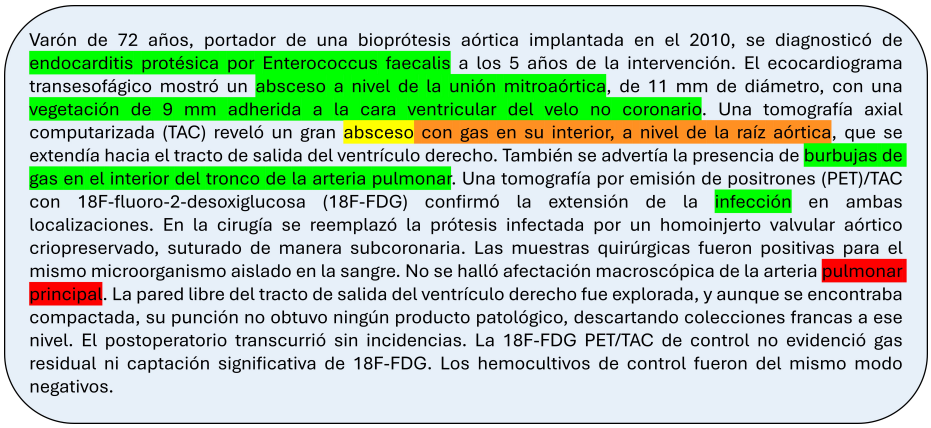

Figure 2: Prediction example for the Spanish Diseases Recognition (SDR) subtask, obtained using the best performing model in terms of F1-score. Green represents correctly identified mentions along with their spans. Red represents mentions that are not present in the ground-truth but predicted by the model. Yellow refers to mentions that are incompletely predicted by the model, while orange marks the full mention as present in the ground-truth.

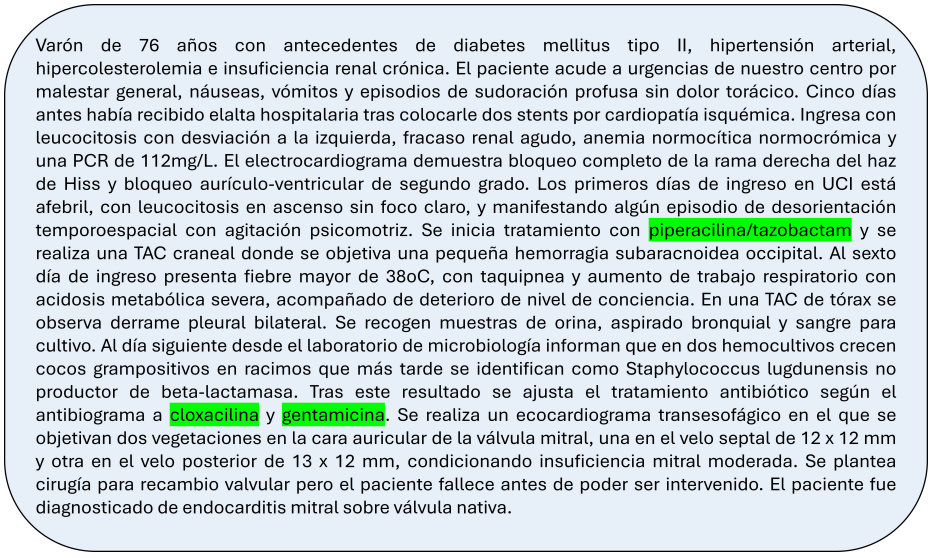

Figure 3: Prediction example for the Spanish Medications Recognition (SMR) subtask, obtained using the best performing model in terms of F1-score. Green represents correctly identified mentions along with their spans. In this particular example, there were no missed, incomplete, or incorrect predictions.

In addition to this performance analysis, we conducted a qualitative evaluation of the top-performing models across all subtasks. The qualitative analysis complements the quantitative metrics, providing a comprehensive assessment of the capabilities of the developed models in real-world clinical scenarios. The outcomes, as illustrated in Figure 2, Figure 3, Figure 4, and Figure 5, indicate that the models perform commendably in identifying medications within clinical texts across all three targeted languages. However, the Spanish Diseases Recognition (SDR) model exhibits room for improvement, as it occasionally produces incomplete or incorrect predictions.

5. Conclusions

In this paper, we investigated the utilization of BERT-based contextual embeddings, trained on general domain texts, for extracting mentions of diseases and medications from clinical case reports written in English, Spanish, and Italian. We developed four distinct monolingual models: (1) Spanish Diseases Recognition (SDR), (2) Spanish Medications Recognition (SMR), (3) English Medications Recognition

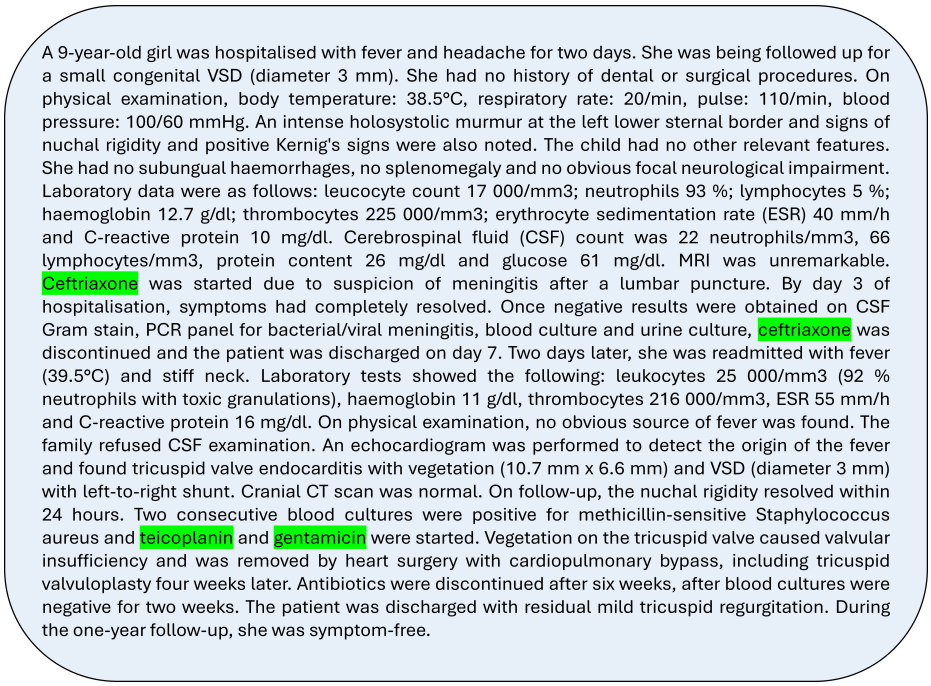

Figure 4: Prediction example for the English Medications Recognition (EMR) subtask, obtained using the best performing model in terms of F1-score. Green represents correctly identified mentions along with their spans. In this particular example, there were no missed, incomplete, or incorrect predictions.

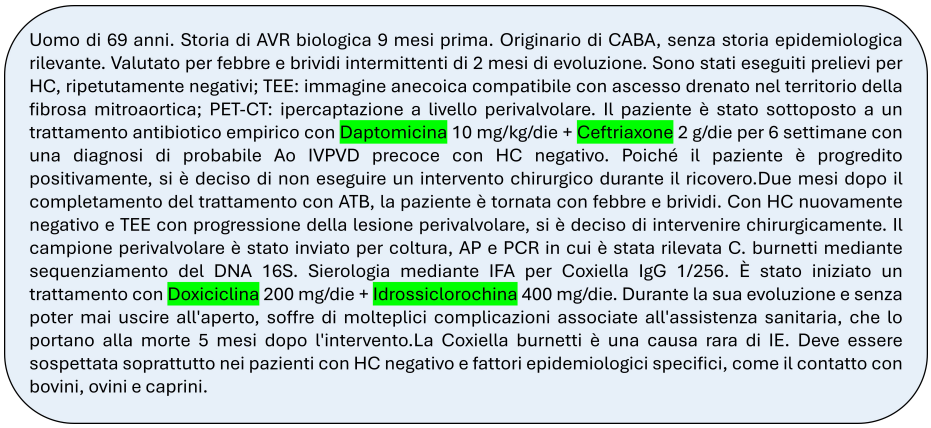

Figure 5: Prediction example for the Italian Medications Recognition (IMR) subtask, obtained using the best performing model in terms of F1-score. Green represents correctly identified mentions along with their spans. In this particular example, there were no missed, incomplete, or incorrect predictions .

(EMR), and (4) Italian Medications Recognition (IMR). Additionally, we created two multilingual models: one specialized for Spanish Diseases Recognition (Multi-SDR) and another for Medications Recognition across all three targeted languages (Multi-MMR). While the results show promising performance in identifying medications within clinical texts across all three languages, the models are not flawless. Some weaknesses arise in diseases recognition, where they occasionally produce incomplete or incorrect predictions. To address these issues, we aim to explore the capabilities of recent large language models (LLMs).

Acknowledgments

This work received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101057849 (DataTools4Heart project).

References

[1] E. T. Rubel Schneider, J. V. Andrioli de Souza, J. Knafou, L. E. Oliveira, Y. B. Gumiel, L. F. de Oliveira, D. Teodoro, E. C. Paraiso, C. Moro, et al., Biobertpt: a portuguese neural language model for clinical named entity recognition, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, 19 November 2020, 2020.

[2] S. R. Kundeti, J. Vijayananda, S. Mujjiga, M. Kalyan, Clinical named entity recognition: Challenges and opportunities, in: 2016 IEEE International Conference on Big Data (Big Data), IEEE, 2016, pp. 1937–1945.

[3] M. Jin, M. T. Bahadori, A. Colak, P. Bhatia, B. Celikkaya, R. Bhakta, S. Senthivel, M. Khalilia, D. Navarro, B. Zhang, et al., Improving hospital mortality prediction with medical named entities and multimodal learning, arXiv preprint arXiv:1811.12276 (2018).

[4] G. Riccio, A. Romano, A. Korsun, M. Cirillo, M. Postiglione, V. La Gatta, A. Ferraro, A. Galli, V. Moscato, Healthcare data summarization via medical entity recognition and generative ai (2023).

[5] D. Zaikis, I. Vlahavas, Drug-drug interaction classification using attention based neural networks, in: 11th Hellenic conference on artificial intelligence, 2020, pp. 34–40.

[6] M. Abulaish, M. A. Parwez, et al., Disease: A biomedical text analytics system for disease symptom extraction and characterization, Journal of Biomedical Informatics 100 (2019) 103324.

[7] B. Rink, S. Harabagiu, K. Roberts, Automatic extraction of relations between medical concepts in clinical texts, Journal of the American Medical Informatics Association 18 (2011) 594–600.

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

[9] K. R. Kanakarajan, B. Kundumani, M. Sankarasubbu, Bioelectra: pretrained biomedical text encoder using discriminators, in: Proceedings of the 20th workshop on biomedical language processing, 2021, pp. 143–154.

[10] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv preprint arXiv:1904.03323 (2019).

[11] F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, H. Yu, et al., Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: an empirical study, JMIR medical informatics 7 (2019) e14830.

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

[13] A. Nentidis, A. Krithara, G. Paliouras, M. Krallinger, L. G. Sanchez, S. Lima, E. Farre, N. Loukachevitch, V. Davydova, E. Tutubalina, Bioasq at clef2024: The twelfth edition of the large-scale biomedical semantic indexing and question answering challenge, in: European Conference on Information Retrieval, Springer, 2024, pp. 490–497.

[14] 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.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for

Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493. doi:10.18653/v1/P19-1493.
[17] S. Wu, M. Dredze, Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, arXiv preprint arXiv:1904.09077 (2019).
[18] Y. Tian, W. Shen, Y. Song, F. Xia, M. He, K. Li, Improving biomedical named entity recognition with syntactic information, BMC bioinformatics 21 (2020) 1–17.
[19] I. Pérez-Díez, R. Pérez-Moraga, A. López-Cerdán, J.-M. Salinas-Serrano, M. d. la Iglesia-Vayá, De-identifying spanish medical texts-named entity recognition applied to radiology reports, Journal of Biomedical Semantics 12 (2021) 1–13.
[20] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodríguez, J. A. Lopez Martin, M. Villegas, M. Krallinger, MEDDOCAN corpus: gold standard annotations for Medical Document Anonymization on Spanish clinical case reports, 2020. URL: https://doi.org/10.5281/zenodo.4279323. doi:10.5281/zenodo.4279323.
[21] V. Kocaman, D. Talby, Biomedical named entity recognition at scale, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I, Springer, 2021, pp. 635–646.
[22] 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, 2022, pp. 193–199.
[23] S. Zhang, H. Cheng, J. Gao, H. Poon, Optimizing bi-encoder for named entity recognition via contrastive learning, arXiv preprint arXiv:2208.14565 (2022).
[24] C. Walker, L. D. Consortium, ACE 2005 Multilingual Training Corpus, LDC corpora, Linguistic Data Consortium, 2005. URL: https://books.google.at/books?id=SbjjuQEACAAJ.
[25] T. Ohta, Y. Tateisi, J.-D. Kim, The genia corpus: an annotated research abstract corpus in molecular biology domain, in: Proceedings of the Second International Conference on Human Language Technology Research, HLT '02, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002, p. 82–86.
[26] A. Khandelwal, A. Kar, V. R. Chikka, K. Karlapalem, Biomedical ner using novel schema and distant supervision, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 155–160.
[27] Y. Hu, I. Ameer, X. Zuo, X. Peng, Y. Zhou, Z. Li, Y. Li, J. Li, X. Jiang, H. Xu, Zero-shot clinical entity recognition using chatgpt, arXiv preprint arXiv:2303.16416 (2023).
[28] OpenAI, Chatgpt, 2022. URL: https://chat.openai.com, accessed: 2024-06-10.
[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. arXiv:2005.14165.
[30] M. Bhattacharya, S. Bhat, S. Tripathy, A. Bansal, M. Choudhary, Improving biomedical named entity recognition through transfer learning and asymmetric tri-training, Procedia Computer Science 218 (2023) 2723–2733.
[31] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources., in: CLEF (Working Notes), 2022, pp. 179–203.
[32] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023., in: CLEF (Working Notes), 2023, pp. 1–18.
[33] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Krallinger, MultiCardioNER Corpus: Multilingual Adaptation of Clinical NER Systems to the Cardiology Domain, 2024. URL: https://doi.org/10.5281/zenodo.11368861. doi:10.5281/zenodo.11368861.

[34] A. Miranda-Escalada, E. Farré, L. Gasco, S. Lima, M. Krallinger, DisTEMIST corpus: detection and normalization of disease mentions in spanish clinical cases, 2023. URL: https://doi.org/10.5281/zenodo.7614764. doi:10.5281/zenodo.7614764.
[35] A. Intxaurrondo, M. Krallinger, Spaccc, 2019. URL: https://doi.org/10.5281/zenodo.2560316. doi:10.5281/zenodo.2560316.
[36] E. Farré-Maduell, L. Gascó, S. Lima, A. Miranda-Escalada, M. Krallinger, DisTEMIST Guidelines: detection and normalization of disease mentions in spanish clinical cases, 2022. URL: https://doi.org/10.5281/zenodo.6477407. doi:10.5281/zenodo.6477407.
[37] S. Lima-López, E. Farré-Maduell, M. Krallinger, DrugTEMIST Guidelines: Annotation of Medication in Medical Documents, 2024. URL: https://doi.org/10.5281/zenodo.11065433. doi:10.5281/zenodo.11065433.
[38] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.
[39] M. Romero, bert-spanish-cased-finetuned-ner, 2020. URL: https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner, accessed: 2024-06-07.
[40] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[41] E. F. Tjong Kim Sang, Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, in: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002), 2002. URL: https://aclanthology.org/W02-2024.
[42] J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, D. Yuret, The conll 2007 shared task on dependency parsing, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, pp. 915–932.
[43] D. S. Lim, bert-base-ner, 2020. URL: https://huggingface.co/dslim/bert-base-NER, accessed: 2024-06-07.
[44] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.
[45] N. Procopio, bert-italian-finetuned-ner, 2023. URL: https://huggingface.co/nickprock/bert-italian-finetuned-ner, accessed: 2024-06-07.
[46] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, H. Ji, Cross-lingual name tagging and linking for 282 languages, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1946–1958. URL: https://aclanthology.org/P17-1178. doi:10.18653/v1/P17-1178.
[47] D. Adelani, bert-base-multilingual-cased-ner-hrl, 2021. URL: https://huggingface.co/Davlan/bert-base-multilingual-cased-ner-hrl, accessed: 2024-06-10.
[48] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, in: Natural language processing using very large corpora, Springer, 1999, pp. 157–176.
[49] A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, I. Androutsopoulos, Evaluation measures for hierarchical classification: a unified view and novel approaches, Data Mining and Knowledge Discovery 29 (2015) 820–865.