

# Transformer-Based Disease and Drug Named Entity Recognition in Multilingual Clinical Texts: MultiCardioNER challenge

Notebook for the BioASQ Lab at CLEF 2024

Anna Aksenova<sup>1,2</sup>, Aleksis Datseris<sup>2,3</sup>, Sylvia Vassileva<sup>3,\*</sup> and Svetla Boytcheva<sup>2,3</sup>

<sup>1</sup>Aalto University, Finland

<sup>2</sup>Ontotext, Sofia, Bulgaria

<sup>3</sup>Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

## Abstract

This paper presents a transformer-based approach for disease Named Entity Recognition (NER) in Spanish clinical texts using the DisTEMIST dataset and drug multi-lingual NER in Spanish, English and Italian clinical texts using the DrugTEMIST dataset. For the disease NER task, we use CLIN-X-ES, a BERT-based model pretrained on Spanish clinical texts and additional pretrained on a custom dataset, and fine-tuned on token classification, achieving F1 score 0.8049 on the test set. For the drug NER task, we experiment with language-specific clinical models as well as general domain multilingual models and achieved the best results with the language-specific models. For Spanish we fine-tuned the CLIN-X-ES model and our best model showed 0.9238 F1 score, for English we fine-tuned BioLinkBERT which scored F1 - 0.9223, and for Italian we pretrained the CLIN-X-ES model with a custom Italian dataset and achieved F1 score - 0.8838. Our system placed first on the English and Italian tracks in the drug subtask in MultiCardioNER challenge.

## Keywords

Named entity recognition (NER), Biomedical NLP, Medication extraction, Diagnosis extraction, Clinical NER, Multilingual NER

## 1. Introduction

Clinical narratives are a valuable source of healthcare information, but it requires design of special NLP models for effective data extraction. Automated identification of key terms such as diseases, medications, and procedures within clinical documents is known as clinical named entity recognition (NER). This process plays a significant role in clinical natural language processing (NLP) by facilitating the extraction of structured data from clinical narratives for subsequent analysis and interpretation by downstream healthcare applications. MultiCardioNER<sup>1</sup> [1] is a shared task part of CLEF BioASQ [2], which aims to detect diseases in Spanish clinical texts as well as drugs in a multilingual setting, including Spanish, English and Italian. The organizers have provided annotated datasets for training and evaluation of the systems - DisTEMIST for disease NER and DrugTEMIST for drug NER. The shared task consists of two subtask - Subtask 1 addressing disease recognition in Spanish, and Subtask 2 addressing drug recognition in multiple languages. In both subtasks, the challenge is to recognize terms in cardiology reports.

This paper describes our approach for disease and drug NER which we submitted for the MultiCardioNER challenge. Our code is available on GitHub<sup>2</sup>. The contributions of this paper are as follows:


---

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

\*Corresponding author.

✉ [anna.aksenova@ontotext.com](mailto:anna.aksenova@ontotext.com) (A. Aksenova); [aleksis.datseris@ontotext.com](mailto:aleksis.datseris@ontotext.com) (A. Datseris); [svassileva@fmi.uni-sofia.bg](mailto:svassileva@fmi.uni-sofia.bg) (S. Vassileva); [svetla@uni-sofia.bg](mailto:svetla@uni-sofia.bg) (S. Boytcheva)

ORCID [0000-0002-3489-874X](https://orcid.org/0000-0002-3489-874X) (A. Aksenova); [0000-0002-2257-0659](https://orcid.org/0000-0002-2257-0659) (S. Vassileva); [0000-0002-5542-9168](https://orcid.org/0000-0002-5542-9168) (S. Boytcheva)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup><https://temu.bsc.es/multicardioner/>

<sup>2</sup><https://github.com/svassileva/enigma-multicardioner>

- Developed a system for disease entity recognition in Spanish and performed different experiments with BERT-based models, achieving 0.8049 F1 score on the DisTEMIST dataset;
- Developed a system for drug entity recognition in Spanish, English, and Italian which achieved state-of-the-art (SOTA) results on English and Italian DrugTEMIST dataset and very competitive results on Spanish;
- We investigated and compared the performance of multilingual BERT-based models vs language-specific models for named entity recognition;
- We adapted a clinical Spanish RoBERTa model to the Italian language and showed the best result on the drug NER task for Italian;

## 2. Related Work

The state-of-the-art methods for biomedical named entity recognition are predominantly using deep learning based models [3], [4]. The most recent NER approaches for clinical documents also include some hybrid models like dictionary guided attention based model [5] and transfer learning [6],[7]. Besides classical approaches like machine learning [8], hidden Markov models (HMM) and conditional random fields (CRF) [9], an interesting application of Fourier Networks for NER and relation extraction were proposed in [10]. Another direction of research is using models based on knowledge graphs [11] as additional source for information. The limited availability of annotated data, imbalanced training datasets and lack of resources in low resource languages triggered another direction in the NER model development using data augmentation techniques to tackle these issues [12], [13].

Specifically for the task of NER for medication extraction the SOTA methods are based on BiLSTM+CRF [14] reporting F1-score 0.93 for the best performing system for NER tasks over MIMIC-III<sup>3</sup> dataset. Another approach is based on Question-Answering (QA) for medication event extraction [15] translating the NER task to span identification task in QA, reporting NER F1-score 0.98. The classical models like SVM, CRF and rule-based models [16],[17] show comparable results. In the n2c2 shared task on medication event extraction in clinical notes [18] the top score model for NER task scores strict F1 0.97 using transformer based pretrained LLM, using a BERT-based model (RoBERTa-large-PM-M3-Voc) with classification layer and BIOES tags.

For the task of NER for diagnoses, the SOTA methods are also based on transformers [19], [20] of various architectures. Most widely used are domain and task adaptations of the transformer architecture, such as BERT, RoBERTa and ELECTRA. Additional enhancement using knowledge bases in deep learning models [21] can help in dataset annotation and expansion.

The organizers of MultiCardioNER have organized multiple challenges in the area of information extraction from Spanish clinical texts for different entity types - diseases, procedures, symptoms, etc. Transformer-based approaches are very commonly used for NER tasks in different languages. In previous challenges for diseases using the DisTEMIST dataset, the top teams have used BERT-based models trained for token classification like Spanish RoBERTa (PlanTL-GOB-ES/roberta-base-biomedical-clinical-es [22]) ([23], [24], [25]), mBERT and the Spanish BETO [26]. In a similar challenge for named entity recognition of medical procedures in Spanish text, competitors used an architecture including the same Spanish RoBERTa or XLM-RoBERTa model and adding BiLSTM and CRF layers on top ([27], [28]). In the task of symptoms NER, ensemble of transformer models for Spanish clinical text achieved the best result - F1 0.74 (strict) [29].

## 3. Data

### 3.1. Subtask 1

MultiCardioNER corpus for disease named entity recognition task consists of two data subsets of different nature: DisTEMIST and CardioCCC. The DisTEMIST dataset consists of 1000 clinical cases of

<sup>3</sup><https://mimic.mit.edu/>

different medical specialities (incl. oncology, otorhinolaryngology, dentistry, pediatrics, primary care, allergology, radiology, psychiatry, ophthalmology and more). CardioCCC is a collection of 508 cardiology clinical case reports, which are longer on average than the DisTEMIST reports. The CardioCCC dataset contains 508 documents, split in 258 for development and 250 for testing. Following the suggestions by the organizers, we used DisTEMIST as a train set, leaving CardioCCC documents for validation. Additionally, a custom dataset was used for clinical domain adaptation (see Section 3.2). Each team had to provide their model predictions on a dataset containing the CardioCCC test data and a large unlabelled background dataset. At the time of submission, the organizers had not released which examples were part of the test set. Therefore the number of documents in the prediction set is a lot higher than the train or validation sets, however only a small portion was used for evaluating the model performance by the organizers.

**Table 1**

Number of sentences per dataset after pre-processing

Dataset	No. sentences
DisTEMIST	15 885
CardioCCC	19 405
CardioCCC test + Background	197 430

### 3.2. Subtask 2

Similar to Subtask1, the MultiCardioNER corpus for the drug prediction task consists of two data subsets of different nature: DrugTEMIST and CardioCCC. The difference between the datasets is quite substantial. DrugTEMIST is a task-specific adaptation of DisTEMIST dataset that consists of 1000 clinical cases of different medical specialities. CardioCCC is a collection of 508 cardiology clinical case reports, meaning that reports are longer on average. Similarly to subtask 1, we used DrugTEMIST as a train set, leaving CardioCCC documents for validation.

One of the important aspects of the dataset is the sparsity of annotations. For DrugTEMIST only 9% sentences contain drugs, while for CardioCCC the share is even lower - 6%.

### 3.3. Custom Medical Text Dataset

To adapt foundation models for the clinical domain, we collected language-specific datasets with raw texts. The data statistics can be found in Table 2.

**Table 2**

Domain adaptation dataset statistics in tokens

Data split	English	Spanish	Italian
Train	20 198 997	21 218 064	21 402 113
Dev	1 935 176	1 494 718	2 000 500
Test	1 387 661	2 660 891	2 271 500

The data was collected using the following sources:

1. **Wikidata concepts related to medicine:** We ran a SPARQL query over WikiData<sup>4</sup> extracting labels that are included in the following medical ontologies and classifications: ICD-11<sup>5</sup>; ICD-10,

<sup>4</sup><https://query.wikidata.org/>

<sup>5</sup><https://icd.who.int/es>

ICD-10 CM<sup>6</sup>, Symptom Ontology<sup>7</sup>, eMedicine<sup>8</sup>, DiseasesDB, MedlinePlus<sup>9</sup>, MONDO<sup>10</sup>, Human Disease Ontology<sup>11</sup>, SNOMED CT<sup>12</sup>, UMLS<sup>13</sup>.

2. **Wikipedia articles related to medical concepts:** Based on the extracted WikiData concepts, we went through the concept list and downloaded the texts of Wikipedia articles that were created for those concepts. For some concepts the articles did not exist. We used Mediawiki API<sup>14</sup> for text extraction.
3. **Abbreviation lists found online:** For each of the languages, we browsed for medical abbreviation lists available online. Some of them were extracted from the open-source articles, others were scraped from websites.
4. **Drug descriptions:** As drug descriptions usually contain quite useful information on symptoms, side effects and dosages, we leveraged multilingual drug description lists<sup>15</sup>.
5. **EMA medical documentation:** We leveraged a parallel corpus of the European Medical Agency Documentation<sup>16</sup>.
6. **Machine-translated data:** To enrich the amount of medical data, we decided to use machine translated datasets. For this purpose, we used medical abstracts corpus<sup>17</sup>.

### 3.4. Drug Gazetteer

As drug lists are dynamic and changing over time, specifically designed dictionaries were created based on various official drug sources to support model-based drug extraction. For the English drug dictionary, we used OMOP Standardized Vocabulary V5.0<sup>18</sup> (incl. ATC, RxNorm), DrugCentral<sup>19</sup> (FDA Approved Drugs, EMA Approved Drugs, PMDA Approved Drugs, PMDA+EMA+FDA Approved Drug), DrugBank<sup>20</sup>, DailyMed<sup>21</sup> (NIHS human drugs), Top250<sup>22</sup>, UnatedHealthcare<sup>23</sup>, and Drugs.com<sup>24</sup> (My Med List). For Spanish, we used Centro de información online de medicamentos de la AEMPS - (CIMA)<sup>25</sup> (incl. ATC Spanish version and Arbol Medicamentos DSCA Spanish). For Italian, we used ATC and Lists of Class A and Class H medicinal products of Italian Medicine Agency<sup>26</sup>. The drug names and synonyms are aggregated in three dictionaries for English, Italian and Spanish. The total number of generic and brand names of drugs included in the cleaned lists, after removing duplication are presented on Table 3. In addition to the drug dictionaries were used some procedures names for lab test from LOINC<sup>27</sup> for all languages in order to disambiguate some drug mentions from lab tests for measuring levels of some minerals and vitamins, like Vitamin D, Magnesium, Calcium, etc.

<sup>6</sup><https://www.eciemaps.sanidad.gob.es/browser/metabuscador>

<sup>7</sup><https://raw.githubusercontent.com/DiseaseOntology/SymptomOntology/main/src/ontology/symp.owl>

<sup>8</sup><https://emedicine.medscape.com/>

<sup>9</sup><https://medlineplus.gov/>

<sup>10</sup><https://obofoundry.org/ontology/mondo>

<sup>11</sup><https://www.disease-ontology.org/>

<sup>12</sup><https://www.snomed.org/>

<sup>13</sup><https://www.nlm.nih.gov/research/umls/index.html>

<sup>14</sup>[https://www.mediawiki.org/wiki/API:Main\\_page](https://www.mediawiki.org/wiki/API:Main_page)

<sup>15</sup><https://www.ema.europa.eu/en/medicines/human>

<sup>16</sup><https://live.european-language-grid.eu/catalogue/corpus/12729>

<sup>17</sup><https://github.com/sebischair/medical-abstracts-tc-corpus>

<sup>18</sup><https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary>

<sup>19</sup><https://drugcentral.org/>

<sup>20</sup><https://go.drugbank.com/>

<sup>21</sup><https://www.dailymed.nlm.nih.gov/dailymed/>

<sup>22</sup><https://clincalc.com/PronounceTop200Drugs/>

<sup>23</sup><https://www.uhc.com/member-resources/pharmacy-benefits/prescription-drug-lists>

<sup>24</sup><https://www.drugs.com/mednotes/>

<sup>25</sup><https://cima.aemps.es/cima/publico/nomenclator.html>

<sup>26</sup><https://www.aifa.gov.it/en/liste-farmaci-a-h/>

<sup>27</sup><https://loinc.org/>

**Table 3**

Drug dictionary after cleaning

Drug Dictionary	English	Spanish	Italian
Drug names	26 843	51151	39985

### 3.5. Data pre-processing

As clinical documents are quite lengthy, especially in the case of the CardioCCC corpus, we decided to split the documents into sentences using the following tools for different languages:

- English - MedSpaCy Sentence Splitting <sup>28</sup>
- Italian - Tint Sentence Splitting <sup>29</sup> [30]
- Spanish - SPACCC Sentence splitter <sup>30</sup>

Afterwards, we used the Brat tool <sup>31</sup> [31] for data transformation from BRAT to CONLL format. The dataset statistics after pre-processing are shown in Table 4.

**Table 4**

Number of sentences per dataset after pre-processing

Dataset	Spanish	English	Italian	Total
DrugTEMIST	15 885	16 342	15 913	48 140
CardioCCC	19 405	18 738	19 396	57 539
CardioCCC test + Background	197 430	195 053	198 084	590 567

## 4. Methods

### 4.1. Subtask 1

Our system was built using transformer-based models that were either multilingual or adapted to Spanish and that were preferably adapted to the biomedical domain as well. The models we used were:

- PlanTL-GOB-ES/roberta-base-biomedical-clinical-es [22]: Biomedical pretrained language model for Spanish. This model is a RoBERTa-based model trained on a biomedical-clinical corpus in Spanish collected from several sources.
- CLIN-X-ES [32]: This model is based on the multilingual XLM-R transformer (xlm-roberta-large) further pretrained on a Spanish clinical corpus.
- DeBERTa v3 [33]: A transformer-based model with disentangled attention trained using ELECTRA style pretraining. Both base and large versions were used and also a version of DeBERTa that was further pretrained on clinical data.
- mDeBERTa v3 <sup>32</sup>: A multilingual version of DeBERTa.

Additionally, some of the models were further pretrained on medical data from the Custom Medical Text Dataset described in Section 3.3. Approximately 0.08 of the tokens are annotated entities per sentence which corresponds to a very sparse annotation setting. Hence the majority of tokens to be evaluated by the model will be negative examples. Therefore, we tried to fine-tune our models with different class weight ratios (positive to negative samples) to try to make up for the class imbalance. The ratios chosen were inversely proportional to the prevalence of the class.

<sup>28</sup><https://github.com/medspacy/medspacy>

<sup>29</sup><https://github.com/dhfbk/tint>

<sup>30</sup>[https://github.com/PlanTL-GOB-ES/SPACCC\\_Sentence-Splitter](https://github.com/PlanTL-GOB-ES/SPACCC_Sentence-Splitter)

<sup>31</sup><http://brat.nlplab.org>

<sup>32</sup><https://huggingface.co/microsoft/mdeberta-v3-base>

## 4.2. Subtask 2

The drug name extraction task was reformulated as a drug Named Entity Recognition (NER) task. As the setting for this task was multilingual, we focused on two major approaches: building a single multilingual model capable of making predictions on all languages by leveraging knowledge transfer between languages during training, and training language-specific models which are independent from each other. In addition, as the drug name list is dynamic, we experimented with adding drug names from official drug registries such as DrugBank as a gazetteer. The gazetteer collection is described in Section 3.4.

In particular we experimented with the following methods:

- **Multilingual model**

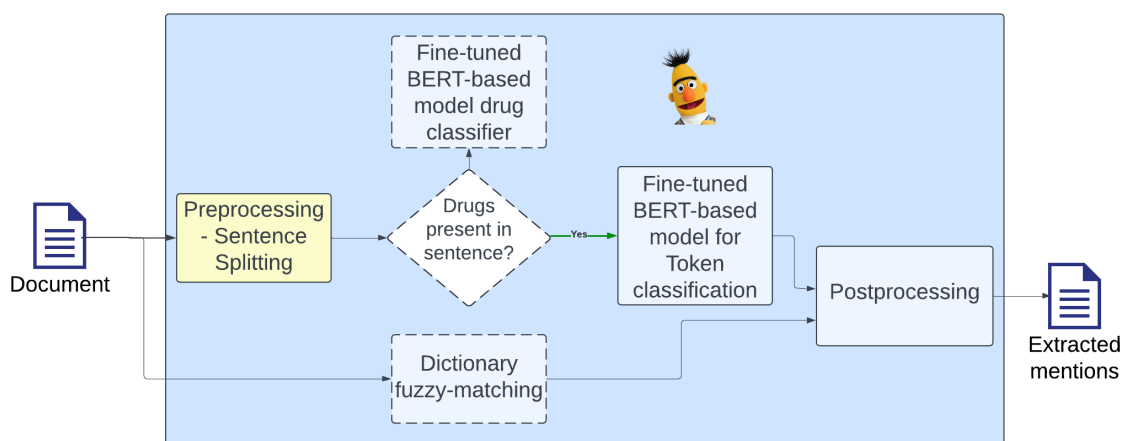
For the foundation multilingual model we used a FacebookAI/xlm-roberta-base<sup>33</sup> backbone. We experimented with training the original model and also performing domain adaptation. The adapted model was trained on English, Spanish and Italian medical datasets on the Masked Language Modeling objective using the dataset from Section 3.3. In addition, we experimented with a multilingual model pretrained on several NER datasets which has shown good results on similar tasks: numind/NuNER-multilingual-v0.1<sup>34</sup>

- **Language-specific models**

As a set of language-specific models we focused on michiyasunaga/BioLinkBERT-base<sup>35</sup> for English, PlanTL-GOB-ES/roberta-base-biomedical-clinical-es<sup>36</sup> for Spanish and dbmdz/bert-base-italian-cased<sup>37</sup> for Italian. As the models for English and Spanish were originally pretrained on medical data, we expected them to perform well as is, whilst for the Italian model, we conducted additional domain adaptation. Furthermore, as Italian medical vocabulary is relatively close to Spanish, we conducted language adaptation of the Spanish model on the Italian medical dataset too.

- **Drug Gazetteer**

The gazetteer was applied before model predictions by finding exact match of drug names in the text. Dictionary statistics and description can be found in Section 3.4.



**Figure 1:** The architecture of the system for named entity recognition. Optional modules are displayed with a dashed border.

<sup>33</sup><https://huggingface.co/FacebookAI/xlm-roberta-base>

<sup>34</sup><https://huggingface.co/numind/NuNER-multilingual-v0.1>

<sup>35</sup><https://huggingface.co/michiyasunaga/BioLinkBERT-base>

<sup>36</sup><https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es>

<sup>37</sup><https://huggingface.co/dbmdz/bert-base-italian-cased>



As described in Section 3.2, the dataset is highly imbalanced and as a potential solution for boosting the precision of predictions, we experimented with training a binary classification model to identify sentences that contain drug annotations and sort out empty sentences. FacebookAI/xlm-roberta-base after domain adaptation was used to train the classifier model. Furthermore, as many drug names included punctuation marks, we added a post-processing step joining drug names divided by symbols / and +.

Figure 1 shows the overall architecture of the approach. Depending on the configuration we either include or do not include filtering and dictionary-based annotations.

## 5. Experiments & Results

### 5.1. Subtask 1

For the Named Entity Recognition task we employed a standard approach for token classification task. To simplify the setting and to avoid truncation due to limits in the input sequence length, we trained on the split sentences (see Section 3.5). Standard token classification pipeline from Huggingface Transformers was used. For the pretraining, we used a standard masked language modeling pretraining objective. In general, the models achieved better recall than precision. The models that achieved the best results were the models that had domain adaptation for both Spanish and clinical language domains. Despite that DeBERTa is generally a model that achieves better performance than XLM-R, the best model in our experiments was CLIN-X-ES, a XLM-R-based model pretrained on a Spanish clinical corpus. It is notable that DeBERTa-base achieved better performance than DeBERTa-large probably because the dataset was quite small. Using pretraining generally gave small improvements or no improvements, but pertained models converged much faster - on average it took them between 2-6 epochs less to converge.

**Table 5**

Models performance on dev dataset. Pretraining data 1 is proprietary biomedical data, and data 2 is the Custom Medical Text Dataset (see Section 3.2)

Model	Token Precision	Token Recall	Token F1
roberta-base-biomedical-clinical-es	0.744	0.760	0.752
CLIN-X-ES	<b>0.814</b>	0.823	<b>0.819</b>
CLIN-X-ES + weighted loss	0.793	<b>0.828</b>	0.810
Deberta-v3-base	0.786	0.784	0.785
mDeberta-v3-base	0.778	0.810	0.793
Deberta-v3-large	0.785	0.755	0.770
Deberta-v3-base + pretraining data 1	0.768	0.782	0.775
Deberta-v3-base + pretraining data 2	0.776	0.802	0.789
CLIN-X-ES + pretraining data 2	0.818	0.816	0.817

### 5.2. Subtask 2

For the second subtask the setup was quite similar to the first one. During the first set of experiments we compared the multilingual candidate models. The results are reported at Table 6. It could be observed that in-domain pretraining positively influences the overall performance of the model. For each of the languages, the F-score improves by circa 2%.

As for the monolingual model comparison reported at Table 7, we can observe that in general monolingual models exhibit slightly better performance for all the target languages except for Italian. Curiously, a Spanish model trained on medical data and fine-tuned on Italian medical data performs better compared to Italian foundation model adapted to the clinical domain.

Table 8 reports results of the submitted systems on the validation dataset. In general, language-specific models show better performance for this task. Filtering consistently increases precision but slightly

**Table 6**  
Multilingual models experiments

Model	F1-es	F1-it	F1-en	F1 overall
NuNER	0.848	0.849	0.852	0.844
XLMR	0.851	0.860	0.845	0.849
XLMR-med	<b>0.871</b>	<b>0.875</b>	<b>0.863</b>	<b>0.873</b>

**Table 7**  
Monolingual models experiments

Model	Language	F1
roberta-base-biomedical-clinical-es	es	<b>0.922</b>
BioLinkBERT	en	<b>0.878</b>
ClinicalBERT	en	0.863
italian-bert-base-cased	it	0.860
italian-bert-base-cased_med	it	0.863
roberta-base-biomedical-clinical-es	it	0.866
roberta-base-biomedical-clinical-es_it	it	<b>0.869</b>

reduces recall for medical XLM-R model pipelines for all the languages. The result is quite natural as by filtering out extra sentences we are adding more false negative examples to the final prediction. The most significant difference is observed for Italian.

**Table 8**  
Comparison of monolingual and multilingual pipeline performance on dev dataset.

Model	Lang	Token Precision	Token Recall	Token F1
XLMR_med	es	0.843	0.907	0.874
XLMR_med	en	0.825	0.895	0.859
XLMR_med	it	0.836	<b>0.892</b>	0.863
XLMR_med + filtering	es	0.853	0.902	0.877
XLMR_med + filtering	en	0.838	0.891	0.864
XLMR_med + filtering	it	0.855	0.889	<b>0.871</b>
roberta-base-biomedical-clinical-es	es	<b>0.915</b>	<b>0.928</b>	<b>0.922</b>
BioLinkBERT	en	<b>0.860</b>	0.897	<b>0.878</b>
roberta-base-biomedical-clinical-es_it	it	<b>0.864</b>	0.874	0.869
XLMR_med + filtering + dict1	es	0.610	0.909	0.731
XLMR_med + filtering + dict1	en	0.770	<b>0.911</b>	0.834
XLMR_med + filtering + dict1	it	0.569	0.863	0.686
XLMR_med + filtering + dict2	es	0.615	<b>0.911</b>	0.734
XLMR_med + filtering + dict2	en	0.794	<b>0.911</b>	0.848
XLMR_med + filtering + dict2	it	0.574	0.866	0.690

The final evaluation on the test set showed that medical XLM-R is a SOTA result for Italian and BioLinkBERT is the best approach for English.

As for the final submission, the best candidate models remained unchanged. Although on the validation set filtering out empty sentences proved to be a beneficial strategy, on the test set results of the pipeline without filtering are generally better.

### 5.3. Error Analysis

After the annotated test data was released, we conducted error analysis and found the following patterns. First, most of the errors related to recall are the cases of drugs that include numbers and special symbols.



**Table 9**

Test set F1 score on the entity level for all submitted systems

Model	Spanish	English	Italian
XLMR	0.912	0.902	<b>0.884</b>
XLMR_filtering	0.908	0.901	0.881
Monolingual	<b>0.924</b>	<b>0.922</b>	0.883
XLMR_filtering_dict1	0.561	0.873	0.684
XLMR_filtering_dict2	0.822	0.887	0.681

However, there were cases where the model predicted drugs names that were missing in the annotations (e.g. it was the case with Angiotensin-converting enzyme (ACE) inhibitors).

Usual treatment:

Sitagliptin + metformin 50/1000 mg, 1 tablet every 12 hours. Valsartan + amlodipine 40/10 mg, 1 tablet per day. Omeprazole 20 mg, 1 capsule per day. Acetylsalicylic acid (ASA) 100 mg, 1 tablet per day. Dapagliflozin + metformin 5/1000 mg, 1 tablet every 12 hours.

**Figure 2:** Inconsistent annotation example.

Based on the performance evaluation one can see that adding dictionary-based matching decreases precision of the models. This can be explained by several factors rooting from the nature of the dictionary we used. In the official drug registries there are quite a lot of drugs with ambiguous brand names (e.g. *vita* - life in Italian), which increases the number of false positive predictions. In addition, there are drug names that are homonymous with laboratory test measures, for instance sodium, vitamin K, etc. Rule-based matching does not rule out such cases. Lastly, official names of the medications include dosages and concentrations, while in the challenge data those were not included in the annotations (e.g. official name *lidocain 2%*, annotated name *lidocain*).

The issue with drug dosage and medication concentration is relevant for model predictions too. While the Italian models rarely included concentrations in the predictions, for English and Spanish this was often the case. Such inconsistencies originate from training data differences. The final models were trained on 2 different datasets (DrugTEMIST, CardioCCC) and those appeared to be annotated differently. While in DrugTEMIST dosages and concentrations were consistently included in the drug span, for CardioCCC this is not the case. Moreover, combined drug names were also annotated inconsistently being either split into 2 drugs, or combined into one. Figure 2 shows an example of drug annotations which do not include the respective dosages as part of the labelled span. The inconsistency of labelling between the train, dev, and test sets is a source of errors for the final trained model.

## 6. Conclusion

In this paper we presented transformer based models for drug and disease named entity recognition in multilingual clinical texts. The experiments for disease NER showed that the domain-adapted and language-specific model CLIN-X-ES scored 0.819 F1 and outperformed DeBERTa-based models. For the drug NER task, the best scores for English and Spanish were achieved using monolingual models BioLinkBERT and RoBERTa-base-biomedical-clinical-es respectively. For Italian, in contrast, the multilingual model XLMR\_med outperformed the monolingual one by a small margin. Although dictionaries showed contribution to the recall metric, the ambiguities in the drug names that contain some common vocabulary or the common confusion with lab test results cause a significant drop in the precision. We experimented with different approaches to attempt to tackle the label sparsity issue - adjusting class weights during training as well as adding a classification step which predicts whether

the sentence contains a drug name. The approach using a classifier as a filtering step showed improved performance on the validation set, however, did not work so well on the actual test set. As future work, using hybrid solution with the help of LLMs can improve the system performance and address the issues with the disambiguation of the term usage in context.

## Acknowledgements

This work was partially supported by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria [Grant Project No. BG-RRP-2.004-0008] and by Horizon Europe research and innovation programme project RES-Q plus [Grant Agreement No. 101057603], funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

## Limitations

The methods described in this paper were submitted as part of the MultiCardioNER challenge and were validated only on the challenge datasets. Further investigation on different datasets is needed to explore the generalizability of the approach. The specific labeling approach on the datasets impacts the model performance, as in the case of drug NER the inconsistencies of dosage labelling was a source of errors. In different settings, the labelling guidelines used might be different and therefore the presented approach may not perform as well.

## References

- [1] 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.
- [2] 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. Goeriot, 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.
- [3] H. Cho, H. Lee, Biomedical named entity recognition using deep neural networks with contextual information, *BMC bioinformatics* 20 (2019) 1–11.
- [4] L. Luo, C.-H. Wei, P.-T. Lai, R. Leaman, Q. Chen, Z. Lu, Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning, *Bioinformatics* 39 (2023) btad310.
- [5] Z. Zhu, J. Li, Q. Zhao, F. Akhtar, A dictionary-guided attention network for biomedical named entity recognition in chinese electronic medical records, *Expert Systems with Applications* 231 (2023) 120709.
- [6] H. Fabregat, A. Duque, J. Martinez-Romo, L. Araujo, Negation-based transfer learning for improving biomedical named entity recognition and relation extraction, *Journal of Biomedical Informatics* 138 (2023) 104279.
- [7] 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.

- [8] S. Archana, J. Prakash, An effective undersampling method for biomedical named entity recognition using machine learning, *Evolving Systems* (2024) 1–9.
- [9] X. Wang, C. Yang, R. Guan, A comparative study for biomedical named entity recognition, *International Journal of Machine Learning and Cybernetics* 9 (2018) 373–382.
- [10] A. Yazdani, D. Proios, H. Rouhizadeh, D. Teodoro, Efficient joint learning for clinical named entity recognition and relation extraction using fourier networks: A use case in adverse drug events, *arXiv preprint arXiv:2302.04185* (2023).
- [11] Y. Lou, X. Zhu, K. Tan, Dictionary-based matching graph network for biomedical named entity recognition, *Scientific Reports* 13 (2023) 21667.
- [12] S. Ghosh, U. Tyagi, S. Kumar, D. Manocha, Bioaug: Conditional generation based data augmentation for low-resource biomedical ner, in: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023*, pp. 1853–1858.
- [13] I. Bartolini, V. Moscato, M. Postiglione, G. Sperli, A. Vignali, Data augmentation via context similarity: An application to biomedical named entity recognition, *Information Systems* 119 (2023) 102291.
- [14] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, et al., A study of deep learning approaches for medication and adverse drug event extraction from clinical text, *Journal of the American Medical Informatics Association* 27 (2020) 13–21.
- [15] T. Tsujimura, K. Yamada, R. Ida, M. Miwa, Y. Sasaki, Contextualized medication event extraction with striding ner and multi-turn qa, *Journal of Biomedical Informatics* 144 (2023) 104416.
- [16] S. Doan, N. Collier, H. Xu, P. H. Duy, T. M. Phuong, Recognition of medication information from discharge summaries using ensembles of classifiers, *BMC medical informatics and decision making* 12 (2012) 1–10.
- [17] S. Doan, H. Xu, Recognizing medication related entities in hospital discharge summaries using support vector machine, in: *Proceedings of COLING. International conference on computational linguistics, volume 2010, NIH Public Access, 2010*, p. 259.
- [18] D. Mahajan, J. J. Liang, C.-H. Tsou, Ö. Uzuner, Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes, *Journal of Biomedical Informatics* 144 (2023) 104432.
- [19] X. Liu, Y. Zhou, Z. Wang, Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network, *Journal of Visual Communication and Image Representation* 60 (2019) 1–15.
- [20] M. Polignano, M. de Gemmis, G. Semeraro, et al., Comparing transformer-based ner approaches for analysing textual medical diagnoses., in: *CLEF (Working Notes), 2021*, pp. 818–833.
- [21] S. Silvestri, F. Gargiulo, M. Ciampi, Iterative annotation of biomedical ner corpora with deep neural networks and knowledge bases, *Applied sciences* 12 (2022) 5775.
- [22] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. *arXiv:2109.03570*.
- [23] V. Moscato, M. Postiglione, G. Sperli, Biomedical spanish language models for entity recognition and linking at biosq distemist, in: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022*, pp. 315–324.
- [24] F. Borchert, M.-P. Schapranow, Hpi-dhc @ biosq distemist: Spanish biomedical entity linking with pre-trained transformers and cross-lingual candidate retrieval, in: *Conference and Labs of the Evaluation Forum, 2022*. URL: <https://api.semanticscholar.org/CorpusID:251471783>.
- [25] M. Chizhikova, J. Collado-Montañez, P. López-Úbeda, M. C. Díaz-Galiano, L. A. U. López, M. T. M. Valdivia, Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions, in: *Conference and Labs of the Evaluation Forum, 2022*. URL: <https://api.semanticscholar.org/CorpusID:251471813>.
- [26] J. Reyes-Aguillón, R. del Moral, O. Ramos-Flores, H. Gómez-Adorno, G. Bel-Enguix, Clinical named entity recognition and linking using bert in combination with spanish medical embeddings, in: *Conference and Labs of the Evaluation Forum, 2022*. URL: <https://api.semanticscholar.org/>

CorpusID:251471813.

- [27] T. Almeida, R. A. A. Jonker, R. Poudel, J. M. Silva, S. Matos, Discovering medical procedures in spanish using transformer models with mcrf and augmentation, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
- [28] M. Chizhikova, J. Collado-Montañez, M. C. Díaz-Galiano, L. A. Ureña-López, M. T. Martín-Valdivia, Coming a long way with pre-trained transformers and string matching techniques: Clinical procedure mention recognition and normalization, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
- [29] F. Gallego, F. J. Veredas, ICB-UMA at BioCreative VIII @ AMIA 2023 Task 2 SYMPTEMIST (Symptom TExt Mining Shared Task), in: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models, Zenodo, 2023. URL: <https://doi.org/10.5281/zenodo.10104058>. doi:10.5281/zenodo.10104058.
- [30] A. Palmero Aprosio, G. Moretti, Italy goes to Stanford: a collection of CoreNLP modules for Italian, ArXiv e-prints (2016). arXiv:1609.06204.
- [31] 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 Session at EACL 2012, Association for Computational Linguistics, Avignon, France, 2012.
- [32] L. Lange, H. Adel, J. Strötgen, D. Klakow, Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain, 2021. URL: <https://arxiv.org/abs/2112.08754>. arXiv:2112.08754.
- [33] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.
- [34] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.

## Appendix A

### Hyperparameters

For fine-tuning the models we used the following hyperparameters settings:

- learning rate: We used AdamW[34] optimizer with between 1-e5 and 5-e5 learning rate.
- number of epochs: We experimented between 5-20 epochs depending on how long it took the model to converge.
- batch size: Initialized to 8 with gradient accumulation steps 1-2 giving an effective batch size of 8-16 due to GPU memory limitations).
- learning rate scheduler: linear

For pretraining we used the following hyperparameter settings:

- learning rate: We used AdamW optimizer with 5-e5 learning rate.
- number of epochs: 3 epochs due to resource limitations.
- weight decay: 0.01.
- batch size: Initialized to 8 due to GPU memory limitations.
- learning rate scheduler: linear