Overview of MultiCardioNER Task at BioASQ 2024 on Medical Specialty and Language Adaptation of Clinical NER Systems for Spanish, English and Italian

Salvador Lima-López[1,*], Eulàlia Farré-Maduell[1], Jan Rodríguez-Miret[1], Miguel Rodríguez-Ortega[1], Livia Lilli[2,3], Jacopo Lenkowicz[2], Giovanna Ceroni[4,5], Jonathan Kossoff[5], Anoop Shah[4,5], Anastasios Nentidis[6,7], Anastasia Krithara[5], Georgios Katsimpras[5], Georgios Paliouras[5] and Martin Krallinger[1]

[1]*Barcelona Supercomputing Center, Plaça Eusebi Güell, 1-3, 08034 Barcelona, Spain*
[2]*Real World Data Facility, Gemelli Generator, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, 00168, Italy*
[3]*Catholic University of the Sacred Heart, Rome, 00168, Italy*
[4]*University College London, UK*
[5]*University College London Hospitals NHS Foundation Trust, UK*
[6]*National Center for Scientific Research "Demokritos", Athens, Greece*
[7]*Aristotle University of Thessaloniki, Thessaloniki, Greece*

Abstract

Transformers and large language models (LLMs) are increasingly used for clinical data analysis, mostly in English, but also in many other languages used within medical care systems. To comply with clinical standards, it is critical to evaluate the generated results by means of benchmarking efforts based on high-quality manually annotated corpora. To foster the adaptation of general clinical natural language processing (NLP) components to the characteristics of medical specialties, as well as exploring cross-language adaptation techniques, we propose the MultiCardioNER task at BioASQ 2024. MultiCardioNER focuses on the adaptation of named entity recognition (NER) systems trained on multispecialty clinical case reports to cardiology, since cardiovascular diseases are the leading cause of death globally. The MultiCardioNER task covered two entity types (diseases and medications) in case reports written in three languages (Spanish, English and Italian). To generate a comparable Gold Standard clinical NER corpus, we used neural translation, annotation projection and manual annotation correction by domain experts. Top scoring teams reached very competitive results for disease (F1-score 0.8199) and medication mentions (0.9277) in Spanish and also obtained very competitive scores for English (F1-score 0.9223) and Italian (F1-score 0.8842). These results suggest that adaptation of general clinical NLP components to a specific clinical specialty can improve the overall results and that cross-language adaptation of clinical NLP components using neural translation and expert-in-the-loop annotation might speed up the implementation of clinical entity extraction systems. The MultiCardioNER corpora, as well as a silver standard made up of predictions of participating systems over the background set, are available at: https://zenodo.org/records/11368861.

Keywords

named entity recognition, cardiology, subdomain adaptation, multilingual, clinical NLP

1. Introduction

Cardiovascular diseases (CVDs) represent a leading cause of death and morbidity worldwide and, therefore, are responsible for considerable disability costs every year. Cardiology is a medical field with

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
salvador.limalopez@bsc.es (S. Lima-López); eulalia.farre@bsc.es (E. Farré-Maduell); jan.rodriguez@bsc.es (J. Rodríguez-Miret); livia.lilli@policlinicogemelli.it (L. Lilli); jacopo.lenkowicz@policlinicogemelli.it (J. Lenkowicz); g.ceroni@ucl.ac.uk (G. Ceroni); j.kossoff@nhs.net (J. Kossoff); a.shah@ucl.ac.uk (A. Shah); tasosnent@iit.demokritos.gr (A. Nentidis); akrithara@iit.demokritos.gr (A. Krithara); gkatsibras@iit.demokritos.gr (G. Katsimpras); paliourg@iit.demokritos.gr (G. Paliouras); mkrallin@bsc.es (M. Krallinger)
0000-0002-7384-1877 (S. Lima-López); 0009-0000-0793-981X (J. Rodríguez-Miret); 0009-0000-0188-079X (M. Rodríguez-Ortega); 0009-0005-3319-7211 (L. Lilli); 0000-0002-8366-1474 (J. Lenkowicz); 0000-0002-8907-5724 (A. Shah); 0000-0002-3782-4412 (A. Nentidis); 0000-0003-0491-4507 (A. Krithara); 0000-0003-3697-941X (G. Katsimpras); 0000-0001-9629-2367 (G. Paliouras); 0000-0002-2646-8782 (M. Krallinger)

its own concepts and expressions of great relevance, as cardiovascular diseases are highly prevalent worldwide. Analysis of unstructured medical data, such as clinical notes or medical publications, may provide an opportunity to improve the characterization of cardiac pathologies. The extraction of clinical variables from medical content is key to enabling healthcare data analytics, improving patient care and advancing precision medicine. Information contained only as free text within Electronic Medical Records is currently mostly unused due to the difficulty of extracting relevant data from very diversely written data sources. Additionally, the distinctive language of each medical specialty calls for more specialized automatic semantic annotation resources in English and any language used in clinical services.

Due to its importance and the need to improve the extraction, use, and ultimately exploitation of patient data suffering from cardiovascular conditions, efforts have been made to implement natural language processing (NLP) solutions to classify or extract key variables from cardiology clinical content. Using results generated by NLP technologies might contribute to improving outcomes and understanding disease in cardiology. In order to account for the diversity and heterogeneity of NLP research applied to cardiology, several review articles have tried to systematically characterize the various NLP application scenarios adapted to handle cardiovascular disease medical documents [1]. These included applications related to heart failure [2], coronary artery disease, general cardiology or valvular heart disease [3, 4]. Efforts were also made to extract, by means of NLP tools, symptoms [5, 6], vital signs [7], heart function measurements [8], risk factors [4], or cardiovascular comorbidities [9] as well as social risk factors [10] or diagnostic codes of common cardiovascular diseases [11]. Some other attempts were also made to explore the use of NLP approaches to extract Framingham criteria [12] or New York Heart Association classifications from unstructured clinical notes [13, 14, 15].

General clinical domain pre-training does not necessarily transfer well to all medical sub-specialties or disciplines because of the use of highly specialized medical language, as encountered in cardiology clinical case reports or cardiology clinical notes. Domain adaptation strategies may have a great potential to improve NLP solutions for practical settings, real-world scenarios and industrial applications [16]. Also, adaptation of clinical NLP solutions across languages other than English is necessary and requires collaboration between researchers to accelerate progress in non-English clinical NLP [17].

In the case of Spanish, general clinical NLP datasets and resources, such as the DisTEMIST, SympTEMIST, PharmaCoNER, and MedProcNER corpora and systems, have been released. However, (a) the interplay and complementarity of multi-label entity extraction approaches were neither targeted nor evaluated, and (b) how such approaches could be adapted to handle multiple languages was not tested.

To address these issues and promote the development of comparable clinical NLP components adapted to a specific clinical domain across several languages, we have organized the MultiCardioNER shared task. This paper presents an overview of the data, methodologies and results of MultiCardioNER. It is structured as follows: Section 2 introduces the shared task, including its sub-tasks and evaluation methods. Next, Section 3 describes the different corpora used as part of MultiCardioNER, namely DisTEMIST, DrugTEMIST and CardioCCC, as well as other associated resources, while Section 4 presents the participation results and proposed methodologies. Finally, Section 5 concludes the paper with a discussion of some of the most interesting aspects, learned lessons, future work and more.

2. Task description

2.1. Shared task description

The MultiCardioNER task participants were asked to implement named entity recognition (NER) systems using a general clinical corpus annotated with disease and medication mentions. They were then required to adapt these NER systems to a particular medical specialty, namely cardiology. In addition, the MultiCardioNER task also explored the creation of clinical multilingual NER components or cross-language adaptation of these systems for three languages: Spanish, English and Italian.

The MultiCardioNER task relied on a previous resource exploited in a past shared task, called DisTEMIST [18]. The DisTEMIST corpus is a collection of 1,000 clinical case reports covering a wide range of specialties annotated for diseases by clinical experts. Furthermore, a previously unreleased corpus called DrugTEMIST was published as part of the task. The DrugTEMIST corpus provides drug or medication mention annotations for the same collection of clinical case reports as used for the DisTEMIST dataset.

For the adaptation to cardiology, we have constructed the CardioCCC corpus, a new dataset that consists of manually selected cardiology clinical case reports showing similar characteristics as cardiology discharge summaries. This resource was provided to participants to enable the exploration of different clinical subdomain/specialty adaptation strategies and to benchmark the resulting systems. To foster the generation of multilingual clinical NER corpora, DrugTEMIST and CardioCCC were automatically translated from Spanish into both English and Italian. The Gold Standard drug mention annotations were then mapped into both target languages and validated manually by clinical experts (native speakers of English and Italian). The three corpora and the underlying annotation projection process are described in more detail in Section 3.

The evaluation process relied on the comparison of participating team predictions against the manual annotations previously done by the clinical experts. Each team was allowed to submit up to 5 runs for each subtrack and language. The evaluation process and metrics are reported in Section 2.3.

2.2. Subtracks

MultiCardioNER was structured into two different subtracks:

- **Subtrack 1 (CardioDis)**. This track focuses on the adaptation of disease recognition systems to the cardiology specialty in Spanish. Participants could use the DisTEMIST corpus [18] as a base training set, together with a new collection of cardiology-specific clinical case reports annotated with diseases (CardioCCC) that could be used to fine-tune or adapt their systems to cardiology case reports.
- **Subtrack 2 (MultiDrug)**. This subtrack focuses on the multilingual or cross-language adaptation (Spanish, English and Italian) of medication recognition systems, specifically for cardiology clinical case reports. For this track, participants could use the DrugTEMIST dataset as NER training resource. This corpus can be seen as a complementary dataset to the previously-released DisTEMIST, ProcTEMIST and SympTEMIST corpora, as it incorporates annotations of medications for the same document collection. To enable adaptation to cardiology, the CardioCCC corpus was annotated with medication mentions and divided into development and test subsets. While the original versions of both datasets were created using Spanish texts, a machine-translated version in English and Italian was revised by hand and annotated by clinical experts.

2.3. Evaluation

The task was divided into distinct phases: training and test set prediction (evaluation). During the training phase, participants were provided with the DisTEMIST and DrugTEMIST datasets, as well as a subset of the CardioCCC corpus made up of 258 documents. The second batch of the CardioCCC collection was used as test set and released together with a larger background set to make sure that no manual post-editing was carried out by the teams and that the submitted systems could scale up to process larger data collections. These collections (test and background set) were released approximately one month after the start of the training phase. Participants were given two weeks to generate predictions for all documents. They were then evaluated using the Gold Standard annotations of the CardioCCC test set, reserving the predictions for the background set to create a participants' Silver Standard (discussed in Section 3.4). It was not mandatory to submit results for all three languages.

Both MultiCardioNER subtracks were evaluated using micro-averaged precision, recall and F1-score. These metrics are calculated as follows:

Table 1
Results of the baseline system (vocabulary transfer) for the two MultiCardioNER subtracks

Subtrack	Language	System Name	Precision	Recall	F1
CardioDis	Spanish	DisTEMIST vocabulary transfer	0.5178	0.3681	0.4303
MultiDrug	Spanish	DrugTEMIST vocabulary transfer	0.6366	0.7148	0.6734
MultiDrug	English	DrugTEMIST vocabulary transfer	0.3317	0.7269	0.4556
MultiDrug	Italian	DrugTEMIST vocabulary transfer	0.3320	0.6844	0.4471

$$\textit{Precision (P)} = \frac{\textit{True Positives}}{\textit{True Positives + False Positives}}$$
$$\textit{Recall (R)} = \frac{\textit{True Positives}}{\textit{True Positives + False Negatives}}$$
$$\textit{F1 score (F1)} = \frac{2 * (P * R)}{(P + R)}$$

As part of the task, an official MultiCardioNER evaluation library was released and is available on GitHub[1]. After the task results were released, the test set Gold Standard annotations were shared with participating teams to enable them to perform extra experiments and facilitate error analysis of their systems.

2.4. Baseline

To provide a baseline system for comparison, we used a simple vocabulary transfer approach that relied on generating a gazetteer of entities from the training sets (DisTEMIST/DrugTEMIST corpora), and carrying out dictionary look-up of these terms in the test set. Specifically, the system is a lexical lookup approach that tries to find the annotated strings in both corpora within the cardiology test set. The baseline results are shown in Table 1.

3. Corpus and resources

The MultiCardioNER task leverages an already-existing corpus, DisTEMIST, as well as two new releases, DrugTEMIST and CardioCCC. DisTEMIST and DrugTEMIST share the same document collection, which consists of clinical case reports from various clinical specialties such as oncology, infectious diseases, urology and psychiatry. This collection of texts has also been used for the procedures corpus MedProcNER/ProcTEMIST [19] and the signs and symptoms corpus SympTEMIST [20]. These corpora could be considered complementary since they have been annotated by the same clinical experts using the same methodology, which includes the creation of dedicated annotation guidelines. They were released as part of previous shared tasks in an effort to promote the development and accessibility of annotated resources for clinical information extraction in Spanish validated by clinical experts. Other resources resulting from this initiative include PharmaCoNER [21], LivingNER [22], MEDDOPROF [23] or MEDDOPLACE [24].

The CardioCCC corpus consists of cardiology-specific clinical case reports. It includes annotations for diseases and drugs created using the same guidelines as DisTEMIST and DrugTEMIST. Although all three corpora were created originally in Spanish, the texts and annotations related to drugs were translated into English and Italian and released for this task. Table 2 provides some statistics for the different datasets that make up MultiCardioNER, which are explained in detail in this section. All datasets described in this section are openly available on Zenodo[2].

[1] https://github.com/nlp4bia-bsc/multicardioner_evaluation_library
[2] https://zenodo.org/doi/10.5281/zenodo.10948354

Table 2
Statistics for the datasets provided for MultiCardioNER. "Annot." stands for "annotations", while "Chars" stands for "characters". Unique annotations refer to the number of distinct annotated strings after converting all annotations to lowercase. The number of tokens has been calculated using the following spaCy models: "es_core_news_sm", "en_core_web_sm" and "it_core_news_sm".

Dataset	Lang.	Entity	Docs	Tokens	Chars	Annot.	Unique Annot.	Mean Annot. Tokens	Mean Annot. Chars
DisTEMIST	ES	Diseases	1,000	406,137	2,335,968	10,664	6,739	3.20 ± 2.98	24.76 ± 18.89
DrugTEMIST	ES	Drugs	1,000	406,137	2,335,968	2,778	925	1.19 ± 0.56	11.34 ± 4.46
	EN	Drugs	1,000	404,194	2,230,631	2,814	875	1.25 ± 0.66	11.26 ± 0.52
	IT	Drugs	1,000	421,251	2,393,002	2,808	893	1.25 ± 0.69	11.49 ± 4.73
CardioCCC	ES	Diseases	508	568,297	3,215,774	18,232	7,692	3.32 ± 2.84	26.28 ± 19.06
	ES	Drugs	508	568,297	3,215,774	4,227	755	1.19 ± 0.71	11.60 ± 5.25
	EN	Drugs	508	576,772	3,114,833	4,231	734	1.21 ± 0.64	11.37 ± 4.74
	IT	Drugs	508	595,332	3,345,466	4,385	752	1.23 ± 0.72	11.85 ± 5.25

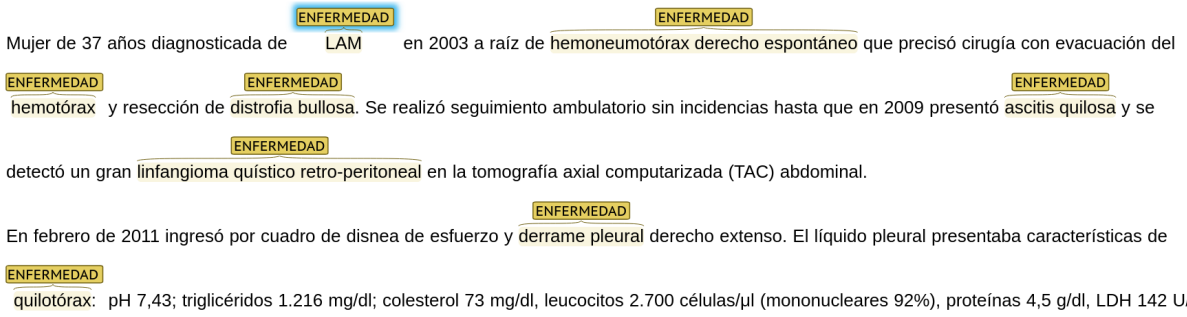


Figure 1: Excerpt from the DisTEMIST corpus with various annotated diseases. Translation with annotated entities in italics: "A 37-year-old woman diagnosed with *AML* (acute myeloblastic leukemia) in 2003 following a *spontaneous right hemopneumothorax* that required surgery with evacuation of the *hemothorax* and resection of *bullous dystrophy*. She was followed up on an outpatient basis without incident until 2009 when he presented with *chylous ascites* and a large *retroperitoneal cystic lymphangioma* was detected on an abdominal computed tomography (CT) scan. In February 2011 she was admitted for exertional dyspnea and extensive right pleural effusion. Pleural fluid showed characteristics of *chylothorax*: [...]".

3.1. DisTEMIST

DisTEMIST is a Gold Standard manually annotated corpus of disease mentions in Spanish clinical case documents normalized or mapped to SNOMED CT concept identifiers. It consists of 1,000 clinical case reports written in Spanish from miscellaneous medical specialties. Figure 1 shows an example of an annotated document.

The texts in the corpus were derived from SciELO (Scientific Electronic Library Online)[3], an electronic library that contains publications from scientific journals. The texts were manually selected by clinical experts so that their structure and content were clinically relevant and representative. The texts were then pre-processed to extract the appropriate sections of the clinical cases and to remove embedded figure references and citations to be as close as possible to real medical records. These texts were originally released under the name SpaCCC (Spanish Clinical Case Corpus). As shown in Table 2, the text collection includes a total of 406,137 tokens and 2,335,968 characters. In terms of annotations, the corpus includes a total of 10,664 entities, out of which 6,739 are unique after converting them into lowercase.

The DisTEMIST corpus was annotated and standardized by two clinical experts from a Spanish

[3]http://www.scielo.org

tertiary hospital. The annotated mentions and their normalization were post-processed and revised afterwards by a third physician. The annotations were created using the *brat* tool [25]. Annotation and normalization guidelines were created specifically for this task. The annotation involved discussions between physicians, particularly regarding complex mentions. This, together with multiple rounds of inter-annotator agreement (IAA) through parallel annotation of a section of the corpus (around 20%), resulted in an iterative refinement of the guidelines. After several rounds, a total IAA score of 82.3 (computed as the pairwise agreement between two independent annotators) for the disease mentions was achieved.

The result of this process is the DisTEMIST guidelines, openly available on Zenodo[4]. The document contains a total of 28 pages describing how to annotate diseases in clinical texts. There are a total of 52 rules divided into various types, such as general, positive or negative. There is also a set of rules specific to oncology mentions that was added as the language used in clinical cases related to this specialty proved to be more specific and harder to annotate. These are partially based on the CANTEMIST corpus [26]. The guidelines also include a discussion of the task's importance, a corpus characterization, basic information about the task and the annotation process, as well as indications and resources for the annotators. It is noteworthy that the DisTEMIST guidelines have been adapted to other domains, such as social media [27].

The DisTEMIST text documents are in plain text format with UTF-8 encoding. The annotations are presented in two different stand-off versions. The first version includes the original annotation files as outputted by *brat* [25]. These are .ann files, one for each text file, where each line represents an annotation, including its label, its start and end position and its associated text. The second version is a single tab-separated file (.tsv) which includes all annotations in the corpus. Similarly to the .ann files, this version includes one annotation per row with an additional field for the corresponding filename.

For MultiCardioNER, all 1,000 documents in the corpus are presented together. For anyone who wishes to use the original train/test split of the corpus (consisting of 750 and 250 documents, respectively), we advise downloading the original DisTEMIST Gold Standard[5] to retrieve the list of filenames belonging to each split. The original repository also includes the SNOMED CT mappings for the annotated mentions, as well as some additional data, such as a background set of related clinical documents and a Silver Standard of the corpus in 6 languages (English, Portuguese, Catalan, Italian, French and Romanian), created using annotation projection. The annotation projection methodology is described in Section 3.2, as well as in the original paper [18].

3.2. DrugTEMIST

DrugTEMIST is a collection of 1,000 clinical case reports from various clinical specialties annotated with mentions of medications. Figure 2 shows an excerpt of an annotated document from the corpus.

The corpus uses the same collection of texts as DisTEMIST, which is also shared by MedProcNER/ProcTEMIST [19] and SympTEMIST [20]. Unlike those corpora, DrugTEMIST hadn't been previously released and is one of the novelties of the MultiCardioNER task. Again, the corpus includes a total of 406,137 tokens and 2,335,968 characters, as well as 2,778 annotated entities (925 unique after converting them to lowercase).

Similarly to the DisTEMIST corpus, dedicated annotation guidelines were written to define what should be considered a medication and how to perform the annotations. These guidelines were created and refined using the same methodology used for DisTEMIST, including thorough discussions between physicians and the annotation of a sample of the corpus (around 20%). The final IAA of the corpus is 0.955. The DrugTEMIST annotation guidelines are also available in Zenodo[6]. They contain 17 pages and are quite similar to the DisTEMIST guidelines, with a total of 29 rules. The release format of the corpus is the same as that of DisTEMIST.

[4] https://zenodo.org/doi/10.5281/zenodo.6458078
[5] https://zenodo.org/doi/10.5281/zenodo.6408476
[6] https://zenodo.org/doi/10.5281/zenodo.11065432

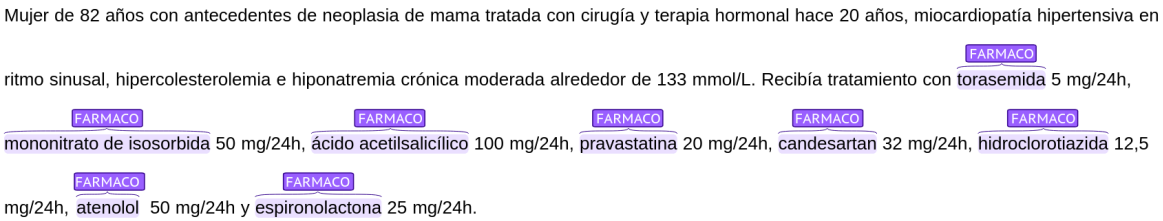


Figure 2: Excerpt from the DrugTEMIST corpus with various annotated medications. Translation with annotated entities in italics: "An 82-year-old woman with a history of breast neoplasia treated with surgery and hormone therapy 20 years ago, hypertensive cardiomyopathy in sinus rhythm, hypercholesterolemia and moderate chronic hyponatremia around 133 mmol/L. She was treated with *torasemide* 5 mg/24h, *isosorbide mononitrate* 50 mg/24h, *acetylsalicylic* acid 100 mg/24h, *pravastatin* 20 mg/24h, *candesartan* 32 mg/24h, *hydrochlorothiazide* 12.5 mg/24h, *atenolol* 50 mg/24h and *spironolactone* 25 mg/24h".

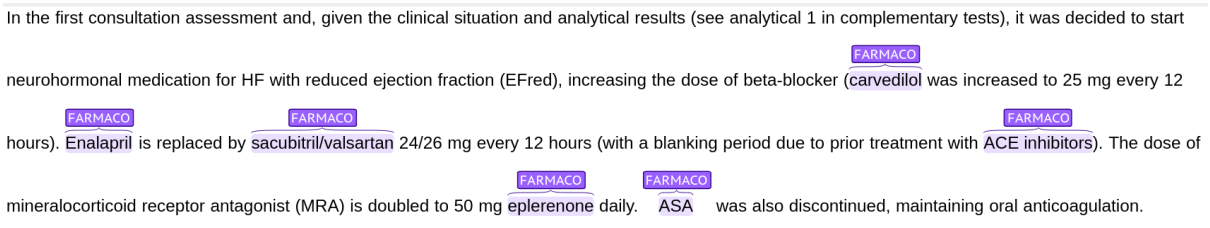
The original Gold Standard of the corpus was created in Spanish. For the multilingual part of the task, we created versions of the corpus in English and Italian using annotation projection techniques. These two languages were chosen due to their relevance for other related projects and the availability of clinical experts fluent in each language, who performed a manual revision of all documents to validate the annotation and the quality of the translation. Specifically, the annotation projection methodology consisted of the following steps:

1. An automatic translation of the Spanish documents was carried out (for the previous DisTEMIST task) using high-quality commercial machine translation systems. In a separate step, the Gold Standard annotations were translated without context (i.e. as a plain list of strings).
2. The translated annotations were next transferred into each document using a look-up system. For each document, only the annotations that existed in the original Gold Standard were looked up to prevent introducing false positives. The result of this step is an automatically annotated version of the corpus in each language, which could be considered a Silver Standard.
3. In order for the corpus to be used as a Gold Standard, a manual revision was performed. Experts compared the original Spanish version of the documents with the version in English and Italian using *brat*'s side-by-side comparison mode. They were tasked with correcting existing and adding new mentions if necessary to make the annotation as close as possible to the original. Additionally, these experts were asked to provide alternative translations to annotated entities that were incorrectly translated.
4. A post-processing step incorporated the alternative translations suggested by the annotators. These translations replaced the original annotated entity both in the text and in the annotation files.

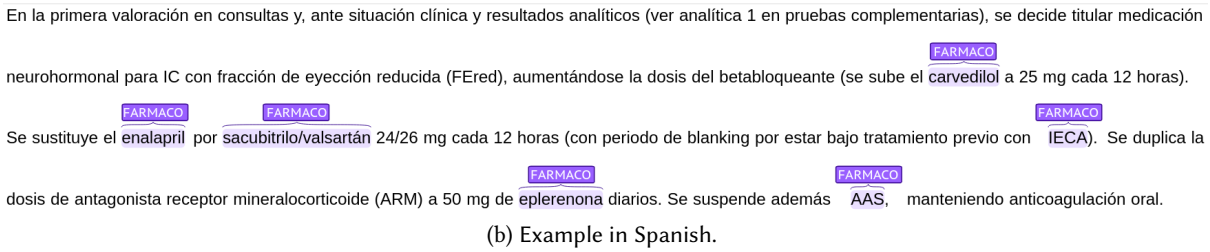
Statistics about the English and Italian versions of DrugTEMIST are also provided in Table 2. We should underscore that the different versions of the corpus do not contain the exact same number of annotations. This is mostly due to translation differences and errors introduced by the machine translation system.

3.3. CardioCCC

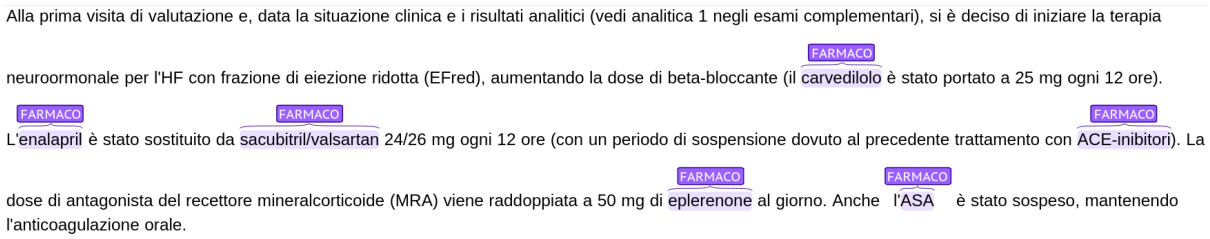
CardioCCC (which stands for Cardiology Clinical Case Corpus) is a collection of 508 cardiology clinical case reports. The documents were retrieved from open-access cardiology journals in Spanish. Within these journals, we tried to manually locate clinical case reports that would have a similar structure to real clinical health records. The candidates were then extracted and, in a similar fashion to the other two corpora presented so far, pre-processed to keep only the relevant article sections and to remove references to figures and tables. The cases were then revised by a clinical expert to confirm their validity. Figure 3 shows a parallel example of the drug annotations in Spanish, English and Italian.



(a) Example in English.



(b) Example in Spanish.



(c) Example in Italian.

Figure 3: Excerpt from the CardioCCC drug annotations in all three languages taken from the same document.

Table 3
Statistics for the two splits of the CardioCCC corpus. CardioCCC_DEV refers to the first batch of the corpus, which participants were allowed to use freely during the training phase. CardioCCC_TEST refers to the held-out test set used for evaluation. "Annot." stands for "annotations", while "Chars" stands for "characters". Unique annotations refer to the number of distinct annotated strings after converting all annotations to lowercase. The number of tokens has been calculated using the following spaCy models: "es_core_news_sm", "en_core_web_sm" and "it_core_news_sm".

Dataset	Lang.	Entity	Docs	Tokens	Chars	Annot.	Unique Annot.	Mean Annot. Tokens	Mean Annot. Chars
CardioCCC_DEV	ES	Diseases	258	315,047	1,777,279	10,348	4,749	3.34 ± 2.72	26.67 ± 18.47
	ES	Drugs	258	315,047	1,777,279	2,510	526	1.21 ± 0.74	11.67 ± 5.34
	EN	Drugs	258	319,289	1,718,585	2,510	513	1.22 ± 0.66	11.48 ± 4.85
	IT	Drugs	258	330,291	1,848,888	2,585	520	1.23 ± 0.73	11.82 ± 5.31
CardioCCC_TEST	ES	Diseases	250	253,250	1,438,495	7,884	3,784	3.29 ± 3.00	25.76 ± 19.80
	ES	Drugs	250	253,250	1,438,495	1,718	465	1.17 ± 0.66	11.51 ± 5.12
	EN	Drugs	250	257,483	1,396,248	1,721	460	1.20 ± 0.61	11.21 ± 4.58
	IT	Drugs	250	265,041	1,496,578	1,800	469	1.22 ± 0.71	11.89 ± 5.16

The corpus contains annotations for diseases and drugs, which were created following the same guidelines used for DisTEMIST and DrugTEMIST. The main annotator for CardioCCC was the same clinical expert who did the final annotation and revision step for the other two corpora, which was a big asset in accelerating the corpus annotation process. As with DrugTEMIST, the corpus's texts were translated from Spanish into English and Italian using machine translation. The Gold Standard drug annotations were also transferred into English and Italian via annotation projection and revised by clinical experts who are native speakers of each language.

As explained in Section 2.3, CardioCCC was released in two batches: one for training/development and another for evaluation. The statistics for these two parts are presented in Table 3, while Table 2 presents the statistics of the complete corpus. In terms of content, as shown by Table 2, the corpus contains 568,297 tokens and 3,215,774 characters. Despite having about half the documents as the SpaCCC corpus (i.e. the texts in DisTEMIST/DrugTEMIST), CardioCCC contains over 150,000 more tokens and one million more characters, meaning the documents are quite longer. This is also reflected in the number of annotations, with CardioCCC having around 8,000 more annotated diseases and 1,500 more drugs. Notably, despite the higher total number of annotations, CardioCCC contains fewer unique drug mentions (which is calculated by converting all annotations to lowercase). This might be due to the fact that in CardioCCC, drug mentions are usually more limited to cardiology-specific medications, while in DrugTEMIST, there is a wider variety of medications mentioned due to the varied clinical specialties it contains. As for the length of annotations themselves, all corpora seem to have a similar distribution in terms of character and token length. The high standard deviation with respect to the mean, especially for diseases, indicates that there's a number of long annotations in the datasets.

3.4. Background set

In addition to the three annotated corpora, an additional dataset was released as a background set. This dataset contains 7,625 text documents, both from the cardiology subdomain and other clinical specialties. While most documents were originally written in Spanish, some of them were also originally in English and Italian. All documents were translated to the other languages to have a comparable background set in all three languages. Together with the background set, we release a tab-separated values (.tsv) file that specifies the original language of each document and whether they belong to the cardiology domain or not.

As part of the task's evaluation period, participants were asked to create predictions for diseases and drugs using their systems. Their predictions were then used to create a Silver Standard, which we release in three different versions:

1. All mentions are kept, with the label name reflecting the team and run the prediction belongs to. This version inevitably includes many incorrect and redundant annotations.
2. Only predictions that have some overlap with the predictions of a different run are used. The overlapping annotations are then merged under a single annotation and a new label name. This version should have a reduced number of incorrect annotations, although some of the "correct" annotations might have extension problems, such as being too short or too long.
3. Only predictions that have a complete overlap with another prediction of a different run are used. This should, in theory, contain the highest number of correct annotations.

Table 4 shows some basic statistics about the text documents included within the Silver Standard. This new dataset can have multiple uses, such as bootstrapping manual annotations, system training using semi-supervised learning techniques or errors and data analysis, amongst others.

Table 4
Statistics of the documents in the background set.

Language	Documents	Tokens	Characters
Spanish	7,625	3,863,801	22,066,533
English	7,625	3,857,831	21,130,044
Italian	7,625	4,015,920	22,782,246

Table 5
Overview of the teams that participated in MultiCardioNER. In the Affiliation column, A/I stands for academic or industry institution. In the Tasks column, C stands for the CardioDis subtrack and M for the MultiDrug subtrack.

Team Name	Affiliation	Tasks	Ref.
BIT.UA	IEETA, University of Aveiro, Portugal [A]	C	[28]
DataScienceTUW	Technische Universität Wien, Austria & Spanish National Research Council (CSIC), Spain [A]	C/M	[29]
Enigma	OntoText, Bulgary & Sofia University, Bulgary [I/A]	C/M	[30]
ICUE	University of Edinburgh, UK & Imperial College London, UK [A]	M	[31]
NOVALINCS	NOVA School of Science And Technology, Portugal [A]	C/M	[32]
PICUSLab	Università degli Studi di Napoli Federico II, Italy [A]	C	[33]
Siemens	Siemens Advanta, Romania & Transilvania University of Brasov, Romania [I/A]	C/M	[34]

4. Results

4.1. Participation overview

A total of 31 teams registered for the MultiCardioNER task, out of which 7 teams submitted at least one run of their predictions. The participating teams originate from 8 different countries (some include collaborations between teams from different countries), and except for one group from the industry, the rest belong to academia. Table 5 shows the complete list of participating teams, along with their affiliation and the reference to their task paper.

As for the participation in each subtrack, 6 teams participated in the CardioDis subtrack, while 5 teams participated in the MultiDrug subtrack (with one of those teams participating only in the Spanish part). Overall, a total of 70 runs were submitted, with each team allowed up to 5 runs per subtrack and language: 20 for the CardioDis subtrack, 18 for the Spanish MultiDrug, 16 for the English MultiDrug and 16 for the Italian MultiDrug.

4.2. System results

All in all, the top scores for each subtrack were:

- **Subtrack CardioDis**. The team BIT.UA attained the top position with an ensemble of RoBERTa-based models (roberta-es-clinical-trials-ner) that also uses a multi-head-CRF approach [35]. Their runs integrated the provided datasets in different ways, with the highest scores achieved by the models that use both the DisTEMIST and CardioCCC data. Their best run achieved an F1-score of 0.8199 and a recall of 0.8243. The team with the next best F1-score (0.8049) is Enigma, which uses a CLIN-X-ES model also fine-tuned on the DisTEMIST and CardioCCC data. Interestingly, the team PICUSLab achieves the best precision (0.8886) by a wide margin by combining the predictions of multiple models trained on different parts of the data (including an augmented version of the CardioCCC corpus) and then using string matching techniques to enhance the final predictions.
- **Subtrack MultiDrug**. In Spanish, the best F1-score is achieved by the ICUE team (0.9277), who also achieved the best recall (0.9412). Meanwhile, in English and Italian, the winning team is Enigma, with an F1-score of 0.9223 and 0.8842, respectively.

The results for the CardioDis subtrack are shown in Table 6, while the results for the MultiDrug subtrack are presented in Table 7 for Spanish, Table 8 for English and Table 9 for Italian.

4.3. Methodologies

This section describes the methodologies used by each team, which are also summarized in Table 10.

Table 6
Results of the MultiCardioNER CardioDis subtrack, sorted by F1-score. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
BIT.UA	run1-all-full	0.8155	**0.8243**	**0.8199**
BIT.UA	run0-top5-full	0.811	<u>0.8181</u>	<u>0.8145</u>
Enigma	3-system-CLIN-X-ES-pretrained	0.8016	0.8082	0.8049
Enigma	2-system-CLIN-X-ES-14	0.8052	0.8007	0.803
PICUSLab	aug_fus_sub2	0.7794	0.803	0.791
BIT.UA	run4-all	0.7981	0.7827	0.7903
Enigma	1-system-CLIN-X-ES-12	0.7827	0.7938	0.7882
PICUSLab	aug_fus_sub1	0.7346	0.7799	0.7566
BIT.UA	run3-all-val	0.7544	0.7588	0.7566
BIT.UA	run2-best-val	0.748	0.7542	0.7511
DataScienceTUW	run4-roberta-dg	0.6565	0.7376	0.6947
DataScienceTUW	run5-roberta-dg-windows	0.6546	0.7244	0.6877
Siemens	run1_SDR	0.6758	0.6437	0.6593
PICUSLab	aug_fus_sm_sub2	0.8919	0.4897	0.6323
DataScienceTUW	run1_mdeberta-ct-mlm-dg	0.5928	0.6715	0.6297
PICUSLab	aug_fus_sm_sub1	**0.8886**	0.4744	0.6185
DataScienceTUW	run2-mdeberta-ct	0.5027	0.6884	0.581
DataScienceTUW	run3_mdeberta-ct-dg	0.48	0.6773	0.5618
NOVALINCS	1_bsc-bio-ehr-es_distemist_4	0.8018	0.3525	0.4897
NOVALINCS	2_bsc-bio-ehr-es_distemist_1	<u>0.8183</u>	0.3398	0.4802

Table 7
Results of the MultiCardioNER MultiDrug subtrack in Spanish, sorted by F1-score. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
ICUE	run2_single_pp	0.9146	**0.9412**	**0.9277**
ICUE	run4_GPT_translation	0.9146	0.9412	0.9277
ICUE	run5_GPT_translation_all	0.9146	0.9412	0.9277
Enigma	3-system-SpanishRoBERTa	0.913	<u>0.9348</u>	<u>0.9238</u>
Enigma	1-system-XLMR	0.904	0.9208	0.9123
Enigma	2-system-XLMR-filtering	<u>0.9148</u>	0.9005	0.9076
ICUE	run3_single	0.8777	0.9272	0.9018
Siemens	run1_SMR	0.8928	0.8778	0.8852
ICUE	run1_multilingual_pp	0.8287	0.9348	0.8786
Enigma	5-system-XLMR-filtering-dict2	0.7654	0.8871	0.8218
NOVALINCS	3_bsc-bio-ehr-es_drugtemist_4	**0.9242**	0.4965	0.646
NOVALINCS	4_bsc-bio-ehr-es_drugtemist_1	0.9076	0.4919	0.638
DataScienceTUW	run3_roberta-ct-multilingual	0.8705	0.4342	0.5794
Enigma	4-system-XLMR-filtering-dict1	0.4351	0.7899	0.5611
DataScienceTUW	run5_roberta-ct-mlm	0.8421	0.3912	0.5342
DataScienceTUW	run4_mdeberta_ct_mlm_dg	0.6815	0.3836	0.4909
DataScienceTUW	run2_mdeberta-ct-multilingual	0.7647	0.3556	0.4855
DataScienceTUW	run1_mdeberta-multilingual	0.3914	0.1531	0.2201

- **Team BIT.UA.**
  For subtrack CardioDis, this team builds on some of their previous work, namely the Multi-Head-CRF approach [35], which introduces a Multi-Head Conditional Random Field (CRF) classifier on top of a multi-class NER system. Starting from the "roberta-es-clinical-trials-ner" pre-trained

Table 8
Results of the MultiCardioNER MultiDrug subtrack in English, sorted by F1-score. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
Enigma	3-system-BioLinkBERT	0.8981	**0.9477**	**0.9223**
ICUE	run2_single_pp	**0.9086**	0.9128	0.9107
ICUE	run4_GPT_translation	0.9086	0.9128	0.9107
Enigma	1-system-XLMR	0.8823	0.9233	0.9023
Enigma	2-system-XLMR-filtering	0.9031	0.8989	0.901
Enigma	5-system-XLMR-filtering-dict2	0.8698	0.9047	0.8869
ICUE	run3_single	0.8734	0.8977	0.8854
ICUE	run1_multilingual_pp	0.8314	0.9343	0.8799
Siemens	run1_EMR	0.8685	0.8791	0.8738
Enigma	4-system-XLMR-filtering-dict1	0.8298	0.921	0.873
ICUE	run5_GPT_translation_all	0.8767	0.8635	0.87
DataScienceTUW	run3_roberta-ct-multilingual	0.8632	0.4364	0.5797
DataScienceTUW	run4-mdeberta-windows	0.7955	0.4317	0.5597
DataScienceTUW	run5-biobert-mlm-windows	0.6771	0.441	0.5341
DataScienceTUW	run2_mdeberta-ct-multilingual	0.8453	0.3777	0.5221
DataScienceTUW	run1_mdeberta-multilingual	0.5648	0.2481	0.3448

Table 9
Results of the MultiCardioNER MultiDrug subtrack in Italian, sorted by F1-score. The best result is bolded, and the second-best is underlined.

Team Name	Run name	Precision	Recall	F1
Enigma	1-system-XLMR	0.884	0.8844	**0.8842**
Enigma	3-system-Italian-Spanish-RoBERTa	0.8723	0.8956	0.8838
Enigma	2-system-XLMR-filtering	0.9016	0.8606	0.8806
Siemens	run1_IMR	0.8891	0.8689	0.8789
ICUE	run4_GPT_translation	**0.9114**	0.8461	0.8776
ICUE	run5_GPT_translation_all	0.9114	0.8461	0.8776
ICUE	run2_single_pp	0.8186	**0.9**	0.8574
ICUE	run1_multilingual_pp	0.8139	0.8867	0.8487
ICUE	run3_single	0.7879	0.8894	0.8356
Enigma	4-system-XLMR-filtering-dict1	0.5693	0.8578	0.6844
Enigma	5-system-XLMR-filtering-dict2	0.5707	0.845	0.6813
DataScienceTUW	run3_roberta-ct-multilingual	0.8264	0.4206	0.5574
DataScienceTUW	run4-mdeberta	0.7481	0.3928	0.5151
DataScienceTUW	run5-biobit-mlm	0.7922	0.3517	0.4871
DataScienceTUW	run2_mdeberta-ct-multilingual	0.7433	0.3394	0.4661
DataScienceTUW	run1_mdeberta-multilingual	0.5074	0.2094	0.2965

model[7], they present 5 runs of ensembled models, with some runs consisting on models fine-tuned only with the DisTEMIST dataset and others with DisTEMIST plus CardioCCC. Their best run is an ensemble of 17 systems trained on both corpora, which achieves the highest F1-score of the subtrack (0.8199).

- **Team Data Science TUW.**
  This team uses four main strategies throughout their experiments for both subtracks: pre-training via MLM (Masked Language Modelling), data augmentation, sliding windows with overlap and additional pre-training on general diseases and drugs using other corpora. The pre-trained models they use include the multilingual mDeBERTa [36, 37], the Spanish "roberta-es-clinical-trials-ner",

[7] https://huggingface.co/lcampillos/roberta-es-clinical-trials-ner

Table 10
General overview of the approaches presented by participants for the MultiCardioNER task. "*TEMIST corpora" refers to the joint version of the DisTEMIST, SympTEMIST, ProcTEMIST and DrugTEMIST corpora.

Team	Task	Approaches
BIT.UA	CardioDis	Ensemble of RoBERTa models with multi-head CRF and differences in the data used for training (only DisTEMIST or DisTEMIST + CardioCCC)
Data Science TUW	CardioDis	Transformer-based models with different pretraining settings, data augmentation and window sliding
	MultiDrug	Multilingual and language-specific Transformers with different pretraining settings, data augmentation and window sliding
Enigma	CardioDis	CLIN-X-ES model fine-tuned on the entire task data + custom clinical dataset
	MultiDrug	Multilingual and language-specific Transformers fine-tuned on the entire task data + custom drug dictionary
ICUE	MultiDrug	Multilingual and language-specific BERT models with re-training, post-processing rules + GPT 3.5
NOVALINCS	CardioDis	RoBERTa model fine-tuned on the standalone DisTEMIST corpus vs. joint *TEMIST corpora
	MultiDrug	RoBERTa model fine-tuned on the standalone DrugTEMIST corpus vs. joint *TEMIST corpora
PICUSLab	CardioDis	Ensemble of Transformer-based models trained on different datasets, including an augmented version of CardioCCC + post-processing via string matching
Siemens	CardioDis	Fine-tuned general domain BERT model
	MultiDrug	Fine-tuned language-specific general domain BERT models

the English "biobert_chemical_ner"[8] and the Italian BioBIT [38].
An important note about this team's results is that they had some problems with their submission that caused the overall low results. This is addressed in their system description, in which they re-evaluate their models with much better results, comparable to some of the task's best.

- **Team Enigma.**
For subtrack CardioDis, team Enigma fine-tuned a CLIN-X-ES model [39] on the DisTEMIST and CardioCCC corpora for a different number of epochs. One of their runs further pre-trains the model using Spanish Wikipedia pages and datasets from different challenges, achieving them a spot in the subtrack's top three F1-scores (0.8049).
For subtrack MultiDrug, the team uses a combination of different models, including a multilingual XLM-RoBERTa [40] and language-specific models such as a Spanish RoBERTa [41] (which they also use for Italian) and BioLinkBERT for English [42]. Their first run, which uses the multilingual XLM-RoBERTa, pre-trains the model on a custom multi-lingual dataset (including biomedical challenge data, European drug description data, Wikipedia) and then fine-tuned for token classification on all data for all languages. For Italian, this approach achieves them the highest F1-score of the Italian part of the subtrack (0.8842).
Their second run uses the same system but adds a classifier before it, which determines if there are any drugs in the sentence. For Spanish and English, their best run is the third one, which uses a language-specific model. This is not the case, however, for their third Italian run, which uses a Spanish model pre-trained on Italian data. Another interesting contribution by this team is the combination of neural systems and drug dictionaries obtained from resources such as DrugBank, ATC, DrugCentral or the NIHS. The two runs that use this approach achieve very good results, although not as good as their other ones.

[8] https://huggingface.co/alvaroalon2/biobert_chemical_ner

- **Team ICUE.**
  For the MultiDrug subtrack, this team compares the effectiveness of multilingual and monolingual BERT models. They also experiment with the inclusion of post-processing rules (specifically for composite drug mentions in Spanish), as well as with using Large Language Models (LLMs) such as GPT-3.5 [43] to translate predictions in Spanish to the other two languages. Their methodology achieves very good results, especially when they use monolingual models. In Spanish, they achieve the best F1-score (0.9277). It is noteworthy that some of their runs in the results table are repeated since they presented the same system with changes only for some languages.
  Team ICUE also includes some additional experiments in their system description paper, such as using GPT-3.5 and LLaMA [44] for entity recognition with competitive results.
- **Team NOVALINCS.**
  For CardioDis, this team fine-tunes the "bsc-bio-ehr-es" pre-trained RoBERTa[9] using the DisTEMIST corpus. They prepared two runs: one in which they only use the DisTEMIST annotations and another in which they also incorporate the other 3 entities from the complementary corpora (that is, procedures from MedProcNER/ProcTEMIST, symptoms from SympTEMIST and medications from DrugTEMIST). For MultiDrug, they only participated in the Spanish part using the same methodology, exchanging DisTEMIST with DrugTEMIST. Their overall results for both tasks are remarkable for their high precision and low recall, which may indicate the difficulty of the systems to adapt to the cardiology subdomain using only the general clinical domain data.
- **Team PICUSLab.**
  For the CardioDis subtrack, this team employs an ensemble transfer learning strategy. They train different models on DisTEMIST, CardioCCC and an augmented version of CardioCCC (created with the help of sentence similarity techniques and a gazetteer), and then fuse the predictions of the different models. To further improve their predictions, they use string matching to post-process them. Their best run earns them a spot in the subtrack's top 5 with an F1-score of 0.791.
- **Team Siemens.**
  This team participated in both CardioDis and MultiDrug with the same methodology. They use general domain BERT models ("bert-spanish-cased-finetuned-ner"[10], "bert-base-NER"[11] and "bert-italian-finetuned-ner"[12]) and fine-tune them for multi-label token classification using the different MultiCardioNER datasets. Despite not using clinical models, their results are quite good, especially for the MultiDrug subtrack (e.g. 0.8789 F1-score in the Italian part). In their overview paper, they also perform additional experiments that were not evaluated during the task's evaluation phase.

5. Discussion

Comparison with previous tasks.

MultiCardioNER is a novel task built upon the foundation of previous tasks and resources. In recent years, tasks such as DisTEMIST [18], PharmaCoNER [21] or MedProcNER [19] have provided the Spanish NLP community with a variety of corpora for the recognition (and normalization) of named entities in clinical texts. These corpora have progressively become reference corpora used to benchmark and model pre-training efforts [39, 45, 46, 47, 48, 49]. MultiCardioNER is different from these previous tasks in that it uses data from a single clinical specialty, rather than a general medical dataset. The CardioCCC corpus could become a reference for cardiology and subdomain adaptation in clinical NLP in Spanish. The corpus is expected to expand with the addition of case reports, more entity types (such as procedures and symptoms), and more languages.

[9] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es
[10] https://huggingface.co/mrm8488/bert-spanish-cased-finetuned-ner
[11] https://huggingface.co/dslim/bert-base-NER
[12] https://huggingface.co/nickprock/bert-italian-finetuned-ner

Subdomain adaptation is a major goal of MultiCardioNER. The task's results indicate the importance of using subdomain data to build systems with specific application fields. All top-performing systems incorporate the released 258 documents from the CardioCCC corpus. In contrast, participants that only use the DisTEMIST and DrugTEMIST corpora (consisting of clinical case reports from various specialties) achieve high precision but fail to recall, thus obtaining a comparatively lower F1-score. This suggests that, while these systems are able to retrieve many clinical entities correctly (i.e. high precision), they fail to recover concepts specific to the cardiology subdomain (i.e. low recall). Furthermore, comparing the results of the DisTEMIST shared task [18] with the CardioDis subtrack, the overall results are somewhat better in the latter task: DisTEMIST's winning team obtained an F1-score of 0.77, while the winning team of MultiCardioNER obtained an F1 of 0.81. This might point to the importance of using specialty-specific data, even within very similar clinical domains.

We should underline that compared with DisTEMIST, this task offers a higher volume of training data. While there seems to be a positive correlation with the use of subdomain-specific data, it remains a question whether these improvements can actually be attributed to subdomain adaptation, to differences in each of the tasks' test sets, or to simply having more data.

Similarity between the general domain and the cardiology corpora.

Given the task's focus on subdomain adaptation, and in order to further characterise the cardiology and SpaCCC datasets (i.e. the DisTEMIST and DrugTEMIST texts) of the shared task, a comparison analysis was conducted between these clinical case reports and documents belonging to other medical disciplines. These documents consist of a collection of clinical cases categorised into 22 different specialities with varying text structures and content, including oncology, COVID-specific reports, primary health care, neurology, etc. The data for the other specialties was extracted using the same methodology as for the CardioCCC (cardiology) corpus (explained in Section 3.3).

For the analysis, we tried to create a mathematical representation of the different document specialties and their subsequent visualisation in a two-dimensional space. To this purpose, the document embeddings were extracted using the pre-trained language model "roberta-base-biomedical-clinical-es" (RoBERTa-based and trained on a large Spanish biomedical corpus from different sources), resulting in tensors of $n \times m$ dimensions, where n is the number of sentences in the document and m is the size of the language model (768 for the RoBERTa model). Subsequently, a vector composition technique was employed to process the extracted document embeddings, as described in the work of Amigó et al. [50]. This involved utilising the proposed generalised composition function in Amigó et al. [50] and illustrated in Equation 1. In this expression, the first component determines the vector direction of the sum of two vectors ($\vec{v_1}$ and $\vec{v_2}$), while the second component represents its magnitude, which depends on the norm of single vectors and their inner product. By applying this function to pairs of successive sentences in a document and representing them as vectors, we are able to compute and represent each document as a single vector (embedding).

In this study we implemented two different composition functions derived from Equation 1, the summation (F_{sum}), obtained when the constants λ and μ are equal to 1 and -2 respectively, and F_{ind}, a particularization of Equation 1 when λ is equals to 1 and μ to 0.

$$F_{\lambda,\mu}(\vec{v_1}, \vec{v_2}) = \frac{\vec{v_1} + \vec{v_2}}{\|\vec{v_1} + \vec{v_2}\|} \cdot \sqrt{\lambda(\|\vec{v_1}^2\| + \|\vec{v_2}^2\|) - \mu\langle\vec{v_1}, \vec{v_2}\rangle} \quad (1)$$

Following the document vector representation and the composition function technique, we implemented a t-Distributed Stochastic Neighbour Embedding (t-SNE) algorithm with a perplexity of 30 and a maximum number of iterations of 800 to reduce the dimensionality of the document embeddings. This statistical method enables the visualisation of high-dimensional document embeddings in lower-dimensional spaces, in this case, two dimensions.

Figures 4 and 5 illustrate the scatter plots generated by the applied methodology, utilising the two composition functions previously mentioned, F_{sum} and F_{ind} respectively. Both figures reveal distinct clustering patterns depending on the specialty. Documents belonging to specific specialties form a well-defined cluster (see cardiology i.e. CardioCCC in black), highlighting the fact that each of them

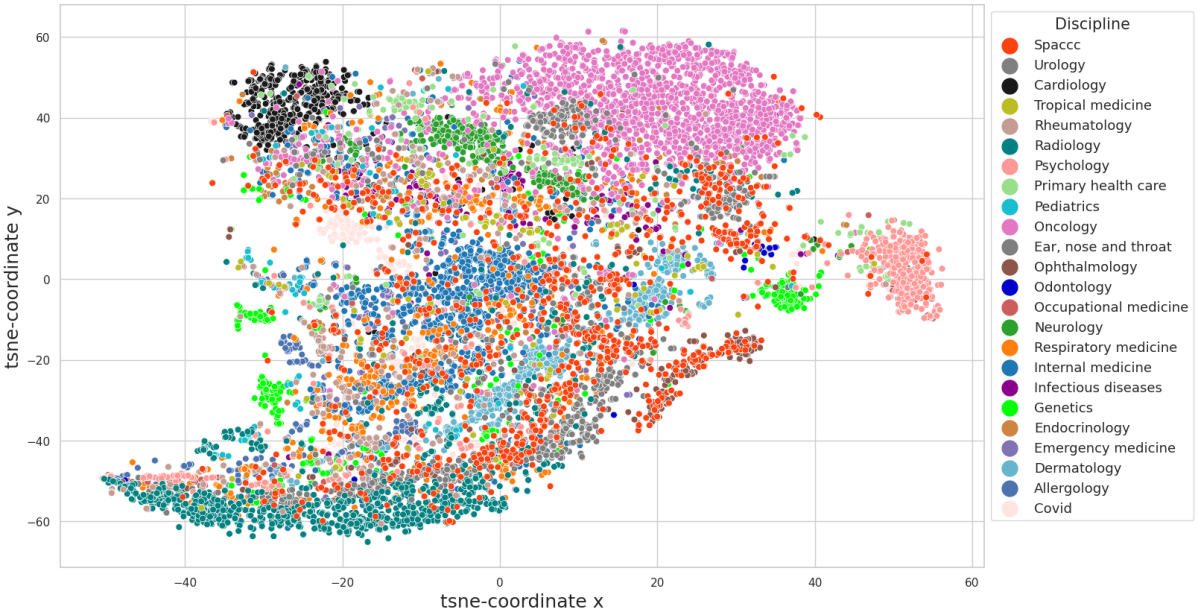


Figure 4: Document embeddings representation per each discipline after reduction of their dimensionality to 2-dimensions by applying the t-SNE algorithm and using F_{sum} as the composition function.

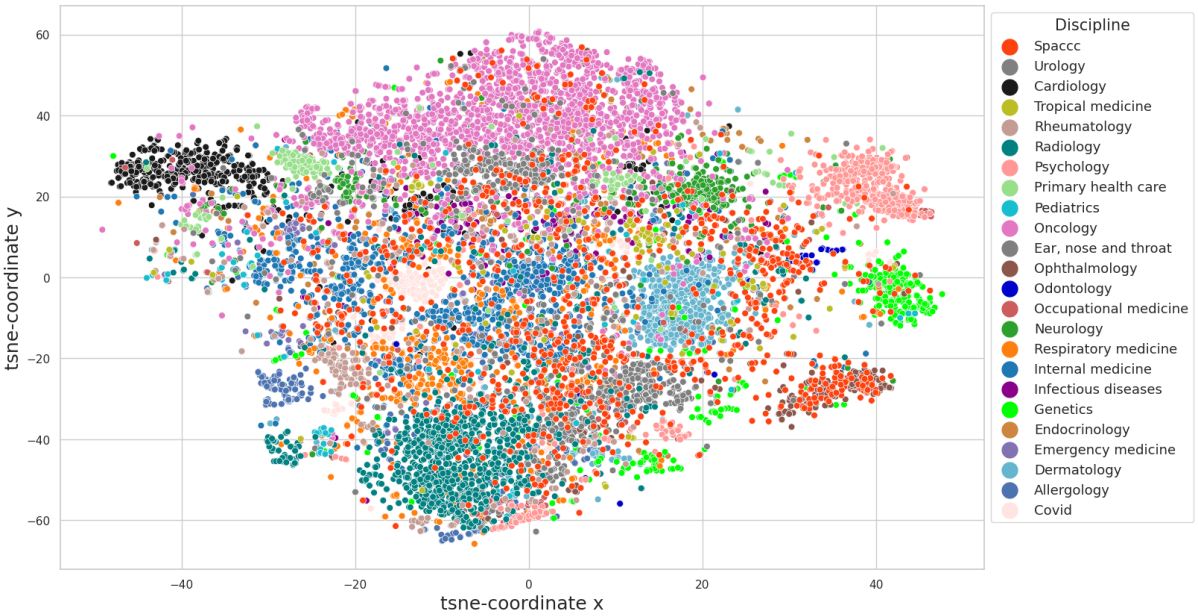


Figure 5: Document embeddings representation per each discipline after reduction of their dimensionality to 2-dimensions by applying the t-SNE algorithm and using F_{ind} as the composition function.

possesses unique features in terms of content and structure. In contrast, documents from the SpaCCC corpus (red points) are scattered across the plot, reflecting their diverse nature. This is due to the fact that they cover a wide range of medical disciplines, such as cardiology (CardioCCC), oncology, urology, pneumology or infectious diseases, among many others.

Future work and conclusions.

There is a pressing need to promote the development of annotated datasets to generate automatic clinical concept detection tools, not only for a single language but for several languages, following comparable annotation criteria and consistent results across multiple languages. Due to the complexity and considerable workload associated with the manual corpus construction process of clinical content,

the use of creative solutions such as neural translation and annotation projection strategies might provide an alternative solution to traditional corpus construction attempts. The results of the MultiCardioNER task indicate that it is feasible to create multilingual clinical corpora and use them to train and generate very competitive clinical NER systems with comparable results across several languages.

Moreover, an adaptation of clinical NLP components to specific medical specialties can improve the quality of the resulting systems for real-world scenarios. Typically clinical NLP application scenarios or use cases focus on content related to a particular medical discipline, disease or patient type. In this regard, the MultiCardioNER task also provides useful insights on how to adapt general-purpose clinical NLP systems to the characteristics of a medical specialty of interest.

We foresee that the results, resources, and strategies generated through the MultiCardioNER task (both by organizers and participants) might potentially promote also the creation of clinical NLP resources beyond the three chosen languages covered in this track. The MultiCardioNER silver standard corpus of predictions for Spanish, English and Italian could also constitute a valuable resource for data augmentation or corpus construction by manually validating the generated system predictions.

The presented annotation projection strategy obviously relies on the sufficient quality of the used medical translation systems. Therefore, systematic efforts to evaluate the quality of neural medical machine translation systems are critical. Initiatives like the Workshop on Machine Translation (WMT) Biomedical Translation shared task has provided insights on the quality and potential of neural translation technologies adapted to translate healthcare documents [51, 52].

Acknowledgments

The MultiCardioNER track was funded by Spanish and European projects such as DataTools4Heart (Grant Agreement No. 101057849), AI4HF (Grant Agreement No. 101080430), BARITONE (Proyectos de Transición Ecológica y Transición Digital 2021. Expediente Nº TED2021-129974B-C21) and AI4ProfHealth (PID2020-119266RA-I00 MICIU/AEI/10.13039/501100011033).

Google was a proud sponsor of the BioASQ Challenge in 2023. Ovid is also sponsoring this edition of BioASQ. The twelfth edition of BioASQ is also sponsored by Elsevier. Atypon Systems Inc. is also sponsoring this edition of BioASQ.

References

[1] A. Tariq, T. Santos, I. Banerjee, Natural language processing for cardiovascular applications, in: Artificial Intelligence in Cardiothoracic Imaging, Springer, 2022, pp. 231–243.

[2] T. Nagamine, B. Gillette, A. Pakhomov, J. Kahoun, H. Mayer, R. Burghaus, J. Lippert, M. Saxena, Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data, Scientific reports 10 (2020) 21340.

[3] M. R. Turchioe, A. Volodarskiy, J. Pathak, D. N. Wright, J. E. Tcheng, D. Slotwiner, Systematic review of current natural language processing methods and applications in cardiology, Heart 108 (2022) 909–916.

[4] M. J. Boonstra, D. Weissenbacher, J. H. Moore, G. Gonzalez-Hernandez, F. W. Asselbergs, Artificial intelligence: revolutionizing cardiology with large language models, European Heart Journal 45 (2024) 332–345.

[5] R. Vijayakrishnan, S. R. Steinhubl, K. Ng, J. Sun, R. J. Byrd, Z. Daar, B. A. Williams, C. Defilippi, S. Ebadollahi, W. F. Stewart, Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record, Journal of cardiac failure 20 (2014) 459–464.

[6] S. Chae, J. Song, M. Ojo, M. Topaz, Identifying heart failure symptoms and poor self-management in home healthcare: a natural language processing study, in: Nurses and Midwives in the Digital Age, IOS Press, 2021, pp. 15–19.

[7] S. Khurshid, C. Reeder, L. X. Harrington, P. Singh, G. Sarma, S. F. Friedman, P. Di Achille, N. Diamant, J. W. Cunningham, A. C. Turner, et al., Cohort design and natural language processing to reduce bias in electronic health records research, Npj Digital Medicine 5 (2022) 47.

[8] O. V. Patterson, M. S. Freiberg, M. Skanderson, S. J. Fodeh, C. A. Brandt, S. L. DuVall, Unlocking echocardiogram measurements for heart disease research through natural language processing, BMC cardiovascular disorders 17 (2017) 1–11.

[9] A. N. Berman, D. W. Biery, C. Ginder, O. L. Hulme, D. Marcusa, O. Leiva, W. Y. Wu, N. Cardin, J. Hainer, D. L. Bhatt, et al., Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-canary comorbidity project, Clinical Cardiology 44 (2021) 1296–1304.

[10] J. R. Brown, I. M. Ricket, R. M. Reeves, R. U. Shah, C. A. Goodrich, G. Gobbel, M. E. Stabler, A. M. Perkins, F. Minter, K. C. Cox, et al., Information extraction from electronic health records to predict readmission following acute myocardial infarction: does natural language processing using clinical notes improve prediction of readmission?, Journal of the American Heart Association 11 (2022) e024198.

[11] X. Zhan, M. Humbert-Droz, P. Mukherjee, O. Gevaert, Structuring clinical text with ai: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases, Patterns 2 (2021).

[12] H. A. Alhakimi, T. E. Magzoub, Applications of natural language processing in cardiology using text clinical data: A systematic review, Advances in Clinical and Experimental Medicine 10 (2023).

[13] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, S. Speedie, Automatic methods to extract new york heart association classification from clinical notes, in: 2017 ieee international conference on bioinformatics and biomedicine (bibm), IEEE, 2017, pp. 1296–1299.

[14] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, S. Speedie, Discovering and identifying new york heart association classification from electronic health records, BMC medical informatics and decision making 18 (2018) 5–13.

[15] P. Adejumo, P. Thangaraj, L. S. Dhingra, A. Aminorroaya, X. Zhou, C. Brandt, H. Xu, H. M. Krumholz, R. Khera, A deep learning approach for automated extraction of functional status and new york heart association class for heart failure patients during clinical encounters, medRxiv (2024).

[16] P. Singhal, R. Walambe, S. Ramanna, K. Kotecha, Domain adaptation: challenges, methods, datasets, and applications, IEEE access 11 (2023) 6973–7020.

[17] E. Laparra, A. Mascio, S. Velupillai, T. Miller, A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records, Yearbook of medical informatics 30 (2021) 239–244.

[18] 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 (2022).

[19] 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: Working Notes of CLEF 2023, 2023.

[20] S. Lima-López, E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret, M. Krallinger, Overview of SympTEMIST at BioCreative VIII: Corpus, Guidelines and Evaluation of Systems for the Detection and Normalization of Symptoms, Signs and Findings from Text, in: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models, 2023.

[21] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 1–10.

[22] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of livingner shared task and resources, Procesamiento del Lenguaje Natural

(2022).
[23] S. Lima-López, E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, M. Krallinger, Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts, Procesamiento del Lenguaje Natural 67 (2021) 243–256. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6393.
[24] S. Lima-López, E. Farré-Maduell, V. Brivá-Escalada, L. Gascó, M. Krallinger, MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts, Procesamiento del Lenguaje Natural 71 (2023).
[25] 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.
[26] A. Miranda-Escalada, E. Farré-Maduell, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
[27] L. Gasco Sánchez, D. Estrada Zavala, E. Farré-Maduell, S. Lima-López, A. Miranda-Escalada, M. Krallinger, The SocialDisNER shared task on detection of disease mentions in health-relevant content from social media: methods, evaluation, guidelines and corpora, in: Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 182–189. URL: https://aclanthology.org/2022.smm4h-1.48.
[28] R. Jonker, T. Almeida, S. Matos, BIT.UA at MultiCardioNER: Adapting a Multi-head CRF for Cardiology, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[29] P. Styll, L. Campillos-Llanos, W. Kusa, A. Hanbury, Cross-Linguistic Disease and Drug Detection in Cardiology Clinical Texts: Methods and Outcomes, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[30] A. Aksenova, A. Datseris, S. Vassileva, S. Boytcheva, Transformer-Based Disease and Drug Named Entity Recognition in Multilingual Clinical Texts: MultiCardioNER challenge, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[31] C. Lee, T. I. Simpson, J. M. Posma, A. D. Lain, Comparative Analyses of Multilingual Drug Entity Recognition Systems for Clinical Case Reports In Cardiology, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[32] R. Gonçalves, A. Lamúrias, Team NOVA LINCS @ BIOASQ12 MultiCardioNER Track: Entity Recognition with Additional Entity Types, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[33] A. Romano, G. Riccio, M. Postiglione, V. Moscato, Identifying Cardiological Disorders in Spanish via Data Augmentation and Fine-Tuned Language Models, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[34] M. D. Danu, V. G. Marica, C. Suciu, L. M. Itu, O. Farri, Multilingual Clinical NER for Diseases and Medications Recognition in Cardiology Texts using BERT Embeddings, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.
[35] R. A. A. Jonker, T. Almeida, R. Antunes, J. R. Almeida, S. Matos, Multi-head CRF classifier for biomedical multi-class named entity recognition on Spanish clinical notes, Database (2024).
[36] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.
[37] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.
[38] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localizing in-domain adaptation of transformer-based biomedical language models, Journal of Biomedical Informatics 144 (2023)

104431.

[39] 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, Bioinformatics 38 (2022) 3267–3274.

[40] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[41] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained Biomedical Language Models for Clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[42] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: Pretraining Language Models with Document Links, in: Association for Computational Linguistics (ACL), 2022.

[43] OpenAI, Gpt-3.5 model, https://www.openai.com, 2023.

[44] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[45] G. García Subies, Á. Barbero Jiménez, P. Martínez Fernández, A comparative analysis of spanish clinical encoder-based models on ner and classification tasks, Journal of the American Medical Informatics Association (2024) ocae054.

[46] A. J. Tamayo Herrera, D. A. Burgos, A. Gelbukh, Clinical text mining in spanish enhanced by negationdetection and named entity recognition, Computación y Sistemas 27 (2023) 1169–1181.

[47] H. Verma, S. Bergler, N. Tahaei, Comparing and combining some popular ner approaches on biomedical tasks, arXiv preprint arXiv:2305.19120 (2023).

[48] A. V. Serrano, G. G. Subies, H. M. Zamorano, N. A. Garcia, D. Samy, D. B. Sanchez, A. M. Sandoval, M. G. Nieto, A. B. Jimenez, Rigoberta: a state-of-the-art language model for spanish, arXiv preprint arXiv:2205.10233 (2022).

[49] F. Gallego, G. López-García, L. Gasco-Sánchez, M. Krallinger, F. J. Veredas, Clinlinker: Medical entity linking of clinical concept mentions in spanish, in: International Conference on Computational Science, Springer, 2024, pp. 266–280.

[50] E. Amigó, A. Ariza-Casabona, V. Fresno, M. A. Martí, Information Theory–based Compositional Distributional Semantics, Computational Linguistics 48 (2022) 907–948. URL: https://doi.org/10.1162/coli_a_00454. doi:10.1162/coli_a_00454.

[51] R. Bawden, K. B. Cohen, C. Grozea, A. J. Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, et al., Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies, in: ACL 2019 Fourth Conference on Machine Translation, Association for Computational Linguistics, 2019, pp. 29–53.

[52] M. Neves, A. J. Yepes, A. Siu, R. Roller, P. Thomas, M. V. Navarro, L. Yeganova, D. Wiemann, G. M. Di Nunzio, F. Vezzani, et al., Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports, in: WMT22-Seventh Conference on Machine Translation, 2022, pp. 694–723.