Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources

Antonio Miranda-Escalada[1], Luis Gascó[1], Salvador Lima-López[1], Eulàlia Farré-Maduell[1], Darryl Estrada[1], Anastasios Nentidis[2,3], Anastasia Krithara[2], Georgios Katsimpras[2], Georgios Paliouras[2] and Martin Krallinger[1]

[1] *Barcelona Supercomputing Center, Spain*

[2] *National Center for Scientific Research "Demokritos", Athens, Greece*

[3] *Aristotle University of Thessaloniki, Thessaloniki, Greece*

Abstract

There is a pressing need for advanced semantic annotation technologies of medical content, in particular medical publications, clinical trials and clinical records. Search engines and information retrieval systems require semantic annotation and indexing systems to support more advanced user search queries. Considering the relevance of disease concepts for clinical coding, automated processing of clinical trials and even patents, it is critical to provide access to high quality manually annotated documents labelled by clinicians for the development and benchmarking of disease mention recognition and grounding tools. This is particularly important for medical content beyond English, where even fewer annotated corpora have been released. To address these issues, we have organized the DisTEMIST (DISease TExt MIning Shared Task) track at BioASQ 2022. It represents the first community effort to evaluate and promote the development of resources for automatic detection and normalization of disease mentions from clinical case documents in Spanish. For this track we have released the DisTEMIST corpus, a collection of 1000 clinical case documents carefully selected by clinicians and annotated manually by a team of healthcare professionals following annotation guidelines and quality control analysis for consistency. Disease mentions were exhaustively mapped by these experts to their corresponding SNOMED CT concept identifiers. Moreover, we have created additional multilingual Silver Standard versions of the corpus for 7 languages (English, Portuguese, French, Italian, Romanian, Catalan and Galician), as well as mention normalization cross-mappings to 4 additional highly used terminologies. We received 38 systems or runs from 9 teams, obtaining very competitive results. Most participants implemented sophisticated AI approaches, mainly deep learning algorithms based on pre-trained transformer-like language models (BERT, BETO, RoBERTa, etc.), with a classifier layer for named entity recognition and embedding distance metrics for entity linking. Finally, some initial explorations of applicability and adaptation of disease taggers trained on the DisTEMIST corpus to different clinical records (discharge summaries, radiology reports and emergency records) were performed. DisTEMIST corpus: https://doi.org/10.5281/zenodo.6408476

Keywords

named entity recognition, disease text mining, entity linking, multilingual NLP, SNOMED CT, clinical NLP

1. Introduction

Systems able to detect and normalize disease mentions from medical content are crucial for a diversity of applications such as semantic indexing for improved retrieval/classification, clinical coding, drug repurposing, or relation extraction (disease-symptom, disease-drug/treatment, disease-gene/mutation) [1].

Focusing on semantic indexing, it was estimated that more than 20% of PubMed queries are related to diseases, disorders, and anomalies [2], stressing the importance for different users (researchers, clinicians, Pharma, biologists) to extract this key information. Indeed, this category has a significant presence in both scientific articles and clinical narratives [3] and it is also relevant to process other kinds of content like social media (e.g. SMM4H/COLING2022 track - SocialDisNER[1]). Consequently, the development of highly efficient systems capable of making these types of entities accessible by search systems has a great interest in the biomedical field. Detecting relevant disease entities can improve indexing systems by alleviating the need to consider entire documents.

Disease detection and normalization has been explored from various perspectives and in different languages. The earliest attempts consisted mostly of systems that tried to map the content of free text to biomedical knowledge sources like MeSH or the UMLS Metathesaurus [4] using lexical look-ups or rule-based methods. One example of such systems is MetaMap [5], a popular program developed by the National Institutes of Health (NIH) to extract UMLS concepts from text. MetaMap uses a fairly configurable algorithm and over 20 years after its release is still being updated [6]. It is also one of the main tools used to index biomedical literature by the National Library of Medicine (NLM)'s Medical Text Indexer [7].

Despite the usefulness of these systems, they have one main downside: concepts from knowledge sources oftentimes do not correspond to the expressions used in texts, especially when it comes to genres like Electronic Health Records, whose language can be noisy and more informal. Annotated corpora opened the door to detection methods with better recall that can tackle these problems, as well as more adequate system evaluation methods (which early mapping systems mostly lacked).

Some notable disease corpora in English are the 2010 i2b2 corpus [8] and NCBI-Disease corpus [9]. The former is a reference corpus of Clinical Records annotated for three tasks: medical problem concept extraction (which includes diseases, although not as a separate entity type), assertion detection (an extension of traditional negation and uncertainty extraction), and relation classification. Originally, no normalization was provided for the concept extraction task. However, a portion of the corpus was re-used as part of the n2c2 2019 shared task [10] and multiple clinical concepts, including diseases, were normalized to SNOMED CT and RxNorm. The NCBI-Disease corpus is a Gold Standard collection of PubMed abstracts annotated with disease concepts which were later normalized to MeSH and OMIM that served as the

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

antoniomiresc@gmail.com (A. Miranda-Escalada)
0000-0002-5654-001X (A. Miranda-Escalada)

[1] https://temu.bsc.es/socialdisner/

foundation of one of the first machine-learning based disease normalization systems: DNorm [11]. Similarly to DisTEMIST, the 2013 ShARe/CLEF eHealth evaluation lab data [12] also proposed participants the challenge of identificating diseases and normalizing them to SNOMED CT. The best performing systems in the strict evaluation setting obtained 0.75 and 0.58 F1-score in detection and normalization, respectively.

In Spanish, there are already some corpora that consider diseases, each of them with different characteristics. For instance, the IxaMed-GS corpus [13] is a collection of Electronic Health Records with annotations for diseases and drugs and relationships between both when they indicate adverse drug reaction events. DrugSemantics [14] is a corpus of summaries of product characteristics with multiple entity types, including diseases. The Chilean Waiting List Corpus [15] is another resource consisting of referrals from public Chilean hospitals.

Only two released corpora in Spanish include mappings to a knowledge source for its annotations. On the one hand, Campillos-Llanos et al. [16] present a corpus of clinical trials annotated with anatomical and chemical entities, disorders and procedures where a small fraction of the annotations are mapped to UMLS. On the other hand, the CodiEsp shared task [17] challenged its participants to associate clinical cases with their correct ICD-10 disease and procedure code (ICD-10 stands for International Classification of Diseases, 10th edition), and to find the specific mentions that supported the code choice. The CodiEsp corpus contains the same number of documents as the DisTEMIST corpus. However, the annotated text mentions in CodiEsp are targeted toward clinical coding with ICD-10 instead of SNOMED CT.

ICD-10 is an international standard for clinical coding designed to provide accurate statistics globally. The scope and granularity of ICD-10 are limited regarding representation of clinical language. In contrast, SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) is the most comprehensive and widely used multilingual clinical terminology [18]. The concepts, descriptions, and relationships in SNOMED CT are intended to accurately represent clinical information. The July 31, 2021 release of the SNOMED CT International Edition included more than 350,000 concepts.

To data and to the best of our knowledge, there are no resources in Spanish that annotate diseases and map them to SNOMED CT. To solve these limitations, we have created the first Gold Standard text corpus of disease mentions, manually mapped to the SNOMED CT terminology [19], the DisTEMIST Gold Standard corpus. DisTEMIST continues our efforts to generate publicly accessible high quality corpora annotated with relevant clinical entities [20, 21, 22, 23, 17, 24, 25] and for semantic indexing in Spanish [26, 3, 27]. It was built following detailed annotation guidelines and exhaustive manual text labeling by clinical experts. The DisTEMIST documents were also carefully selected to represent a wide range of diseases from multiple medical specialties (cardiology, ophthalmology, infectious diseases, urology, oncology, paediatrics, tropical diseases, internal medicine, dentistry and other), to facilitate knowledge transfer to different fields and textual sources. The creation of the DisTEMIST Gold Standard corpus was divided into two separate steps: 1) manual text annotation, where the annotator/s recognizes and tags disease mentions in the text and 2) assignment of specific SNOMED CT identifiers to each mention. The main challenges to the normalization to SNOMED CT reside in the considerable variability of expressions used by clinicians to describe the same disease, changes in medical terminology over time, and the richness of clinical entities and expressions, which not always found an optimal SNOMED CT identifier. To increase the exploitation and

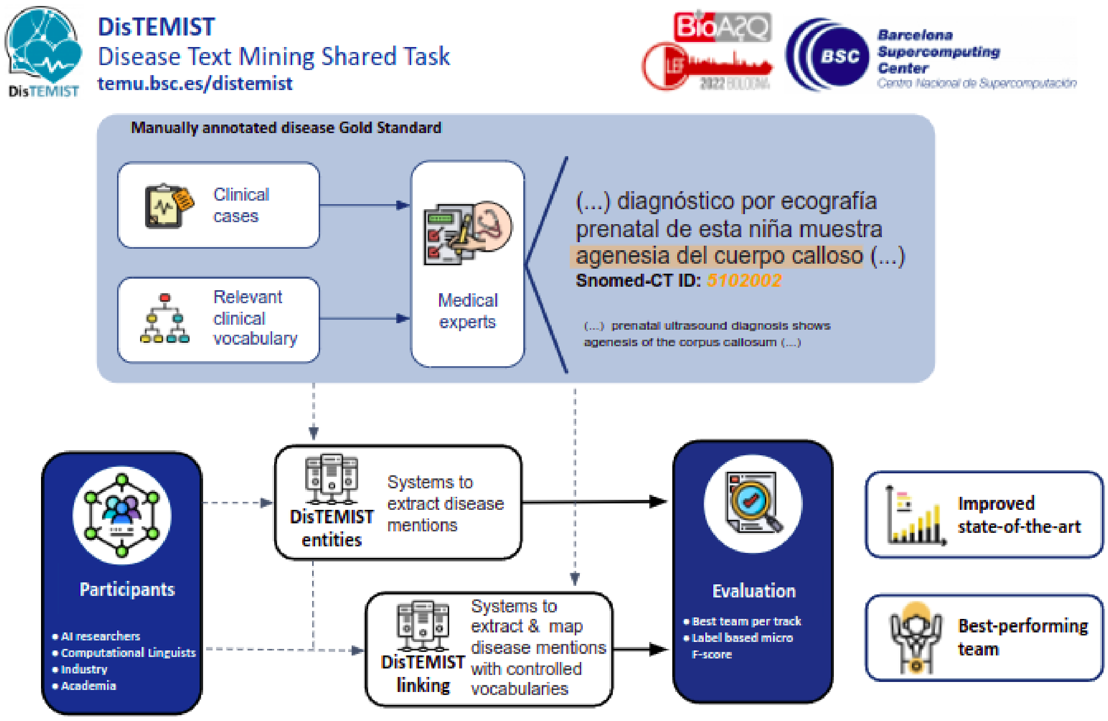

DisTEMIST
Disease Text Mining Shared Task
temu.bsc.es/distemist
DisTEMIST
BioASQ
CLEF 2022 BOLOGNA
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
Manually annotated disease Gold Standard
Clinical cases
Relevant clinical vocabulary
Medical experts
(...) diagnóstico por ecografía prenatal de esta niña muestra agenesia del cuerpo calloso (...)
Snomed-CT ID: 5102002
(...) prenatal ultrasound diagnosis shows agenesis of the corpus callosum (...)
Participants
• AI researchers
• Computational Linguists
• Industry
• Academia
DisTEMIST entities
Systems to extract disease mentions
DisTEMIST linking
Systems to extract & map disease mentions with controlled vocabularies
Evaluation
• Best team per track
• Label based micro F-score
Improved state-of-the-art
Best-performing team

Figure 1: Overview of the DisTEMIST Shared Task.

impact of the DisTEMIST corpus, it was used for a shared task in the context of the CLEF 2022 evaluation initiative. This paper provides an overview of the results, data, methods, outcome and future outlook of the DisTEMIST shared task.

To improve disease information extraction systems, the DisTEMIST Gold Standard and other resources described in this manuscript are released to the community through the DisTEMIST shared task -part of the CLEF and BioASQ 2022 initiative. DisTEMIST invites researchers, biomedical industry professionals, natural language processing and ontology experts to develop systems capable of indexing the content about diseases in biomedical texts, basing the choice on the existing evidence in the text by using Named Entity Recognition (NER), entity linking, and cross-ontology mapping techniques.

2. Task Description

2.1. Shared Task goal

The DisTEMIST shared task explores the automatic recognition of disease mentions in clinical documents in the Spanish language, as well as the assignment of SNOMED CT codes to each mention. DisTEMIST is the first community effort specifically focused on Named Entity Recognition and normalisation of diseases mentions to SNOMED CT in clinical cases written in

Spanish.

2.2. Sub-tasks

The DisTEMIST track is structured into two independent sub-tasks, each taking into account a particularly important use case scenario:

- *DisTEMIST-entities sub-task.* It requires automatically finding disease mentions in published clinical cases. All disease mentions are defined by their corresponding character offsets (start character and end character) in UTF-8 plain text clinical cases.
- *DisTEMIST-linking sub-task.* It requires automatically finding disease mentions in published clinical cases and assigning a SNOMED CT term to each mention.

2.3. Shared Task setting

The DisTEMIST track was organized in two participation periods or phases:

- *Training phase.* In this first phase, the training subset of the complete corpus was released, containing plain text documents and their annotations in the proper format (see section 3 for more details on corpus format). During this period, participants build their systems.
- *Test phase.* In this phase, the test set was released. Only the plain text documents were provided to the participants. They had to use their systems to predict the correct annotations for these documents. After the submission deadline, the organizers evaluated the participants' predictions against the manual annotations done by clinical experts. Each team was allowed to submit up to 5 runs.

2.4. Evaluation metrics

In both sub-tasks, DisTEMIST-entities and DisTEMIST-linking, the main evaluation metric has been micro-average f1-score. In addition, micro-average precision and recall have been computed.

$$\begin{aligned}\text{Precision (P)} &= \frac{\text{true positives}}{\text{true positives + false positives}}\\ \text{Recall (R)} &= \frac{\text{true positives}}{\text{true positives + false negatives}}\\ \text{F1 score (F1)} &= \frac{2 * (P * R)}{(P + R)}\end{aligned}$$

In the DisTEMIST-linking sub-task, only a set of SNOMED CT terms defined a priori were considered in the evaluation (see Section 3.4) to narrow down the SNOMED CT space that participants had to target. The DisTEMIST evaluation library is available on GitHub[2].

[2] https://github.com/TeMU-BSC/distemist_evaluation_library

2.5. Baseline

For the DisTEMIST-entities sub-task, we prepared two baseline models to compare with participants' systems.

- *DiseaseTagIt-VT*. This system uses a simple vocabulary transfer approach from training to test set. It follows a Levenshtein lexical lookup approach using a sliding window of varying length. For further reference, check out the CANTEMIST overview paper [17]. The code is available on GitHub[3].
- *DiseaseTagIt-Base*. This competitive baseline is a deep neural network system trained with the DisTEMIST training dataset. The network is a customization of the BiLSTM-CRF architecture and it employs word embeddings optimized for biomedical Spanish language [28]. For a more in-depth description of the system, check the PharmaCoNER tagger paper [29]. The code is available on GitHub[4]. A web demo of the DiseaseTagIt-Base system has also been made public[5].

In the DisTEMIST-linking sub-task, a baseline model has been developed for comparison with the participating systems. *TEMUNormalizer-Fuzzy* model applies a lexical similarity strategy between mentions and each of the gazetteer entries to make predictions. For each mention-lexical entry pair, the normalized Levenshtein distance is calculated, assigning the lexical entry code when the value is greater than a threshold. In the case of the sub-task, we selected 0.9. The code is available on GitHub [6].

3. Corpus and Resources

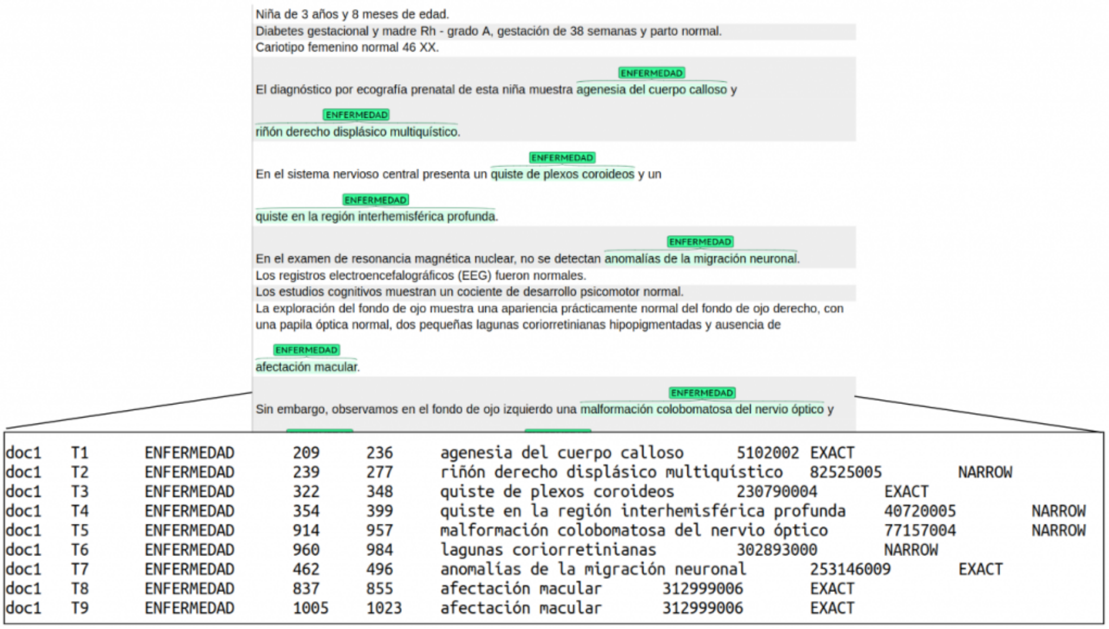
3.1. DisTEMIST Gold Standard Corpus

The DisTEMIST corpus is a collection of 1000 clinical case reports written in Spanish from miscellaneous medical specialties. Clinical experts have manually annotated all corpus documents with mentions of diseases (in Spanish, "enfermedad"). Every disease mention is manually linked to a SNOMED CT term. Figure 3 shows an example of an annotated document fragment. The DisTEMIST corpus is publicly available at Zenodo[7].

Novelty. This is the first Gold Standard manually annotated corpus of diseases in Spanish clinical documents with the mentions mapped to SNOMED CT. Additionally, the DisTEMIST guidelines are part of a pioneer effort to facilitate the creation and usability of annotated resources for clinical NLP in Spanish. The effort already includes other annotation guidelines such as the PharmaCoNER [21], CANTEMIST [23] and MEDDOCAN guidelines [22].

Document selection. All clinical cases derived from various databases were gathered and preprocessed, and the actual clinical case section was extracted, removing embedded figure references or citations. These records were classified manually by a practicing oncologist and

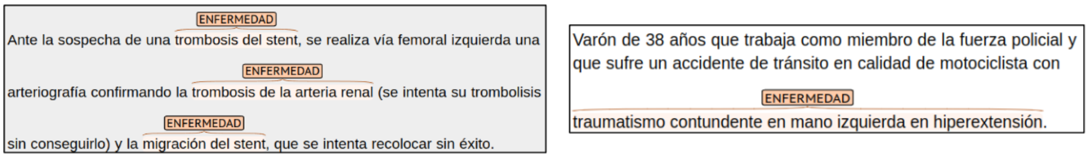
[3] https://github.com/tonifuc3m/character-lookup
[4] https://github.com/TeMU-BSC/PharmaCoNER-Tagger
[5] https://textmining.bsc.es/ner/04
[6] https://github.com/TeMU-BSC/TEMUNormalizer
[7] https://doi.org/10.5281/zenodo.6408476



Niña de 3 años y 8 meses de edad.
Diabetes gestacional y madre Rh - grado A, gestación de 38 semanas y parto normal.
Cariotipo femenino normal 46 XX.
El diagnóstico por ecografía prenatal de esta niña muestra agenesia del cuerpo calloso y riñón derecho displásico multiquístico.
En el sistema nervioso central presenta un quiste de plexos coroideos y un quiste en la región interhemisférica profunda.
En el examen de resonancia magnética nuclear, no se detectan anomalías de la migración neuronal.
Los registros electroencefalográficos (EEG) fueron normales.
Los estudios cognitivos muestran un cociente de desarrollo psicomotor normal.
La exploración del fondo de ojo muestra una apariencia prácticamente normal del fondo de ojo derecho, con una papila óptica normal, dos pequeñas lagunas coriorretinianas hipopigmentadas y ausencia de afectación macular.
Sin embargo, observamos en el fondo de ojo izquierdo una malformación colobomatosa del nervio óptico y

```
doc1	T1	ENFERMEDAD	209	236	agenesia del cuerpo calloso	5102002 EXACT
doc1	T2	ENFERMEDAD	239	277	riñón derecho displásico multiquístico	82525005		NARROW
doc1	T3	ENFERMEDAD	322	348	quiste de plexos coroideos	230790004	EXACT
doc1	T4	ENFERMEDAD	354	399	quiste en la región interhemisférica profunda	40720005		NARROW
doc1	T5	ENFERMEDAD	914	957	malformación colobomatosa del nervio óptico	77157004		NARROW
doc1	T6	ENFERMEDAD	960	984	lagunas coriorretinianas	302893000	NARROW
doc1	T7	ENFERMEDAD	462	496	anomalías de la migración neuronal	253146009	EXACT
doc1	T8	ENFERMEDAD	837	855	afectación macular	312999006	EXACT
doc1	T9	ENFERMEDAD	1005	1023	afectación macular	312999006	EXACT
```

Figure 2: Annotated clinical case visualized with Brat tool [30] and annotation tab-separated format.



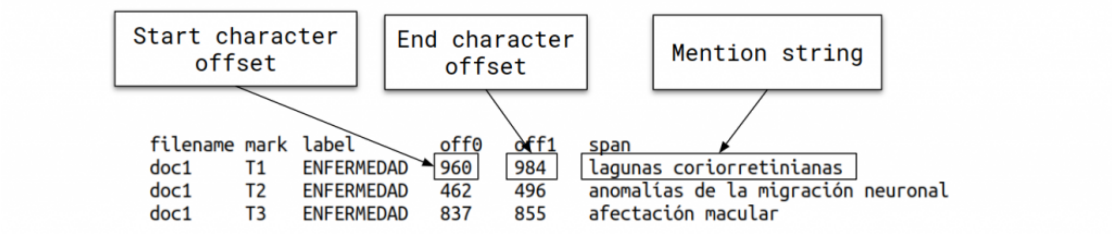
Ante la sospecha de una trombosis del stent, se realiza vía femoral izquierda una arteriografía confirmando la trombosis de la arteria renal (se intenta su trombolisis sin conseguirlo) y la migración del stent, que se intenta recolocar sin éxito.

Varón de 38 años que trabaja como miembro de la fuerza policial y que sufre un accidente de tránsito en calidad de motociclista con traumatismo contundente en mano izquierda en hiperextensión.

Figure 3: (A) Example sentence from DisTEMIST Gold Standard related to the complications caused by a stent (translation): "Suspecting stent thrombosis, an arteriography was performed via the left femoral artery confirming renal artery thrombosis (thrombolysis was attempted without success) and stent migration, which was unsuccessfully repositioned." (B) Example sentence from DisTEMIST Gold Standard related to a work accident (translation): "38-year-old male member of the police force who suffers a traffic accident as a motorcyclist with blunt trauma to the left hand in hyperextension."

revised by a clinical expert to ensure that they were related to the medical domain and their structure and content is relevant to process clinical content. The final collection of 1,000 clinical cases has 16,678 sentences, with an average of 16.7 sentences per clinical case.

Corpus annotation. The DisTEMIST corpus was annotated and standardized by two clinical experts from a Spanish tertiary hospital, with the support of a physician, who was also in charge of reviewing the mentions and their associated codes to reach a final version. The corpus annotation process took place between 2018 and 2020, with an approximate duration of 14 months, and the normalization process occurred in 2020 for approximately six months. The annotation and normalization review process lasted around two extra months each.



```
filename mark label      off0  off1  span
doc1     T1   ENFERMEDAD 960   984   lagunas coriorretinianas
doc1     T2   ENFERMEDAD 462   496   anomalías de la migración neuronal
doc1     T3   ENFERMEDAD 837   855   afectación macular
```

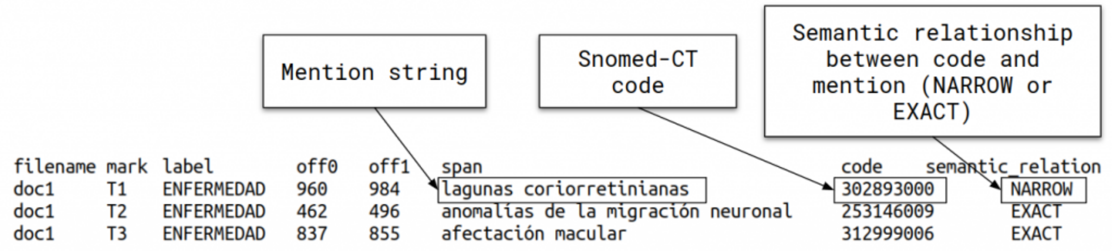
Figure 4: Example of DisTEMIST-entities annotation.

Before starting the annotation, the first draft of these guidelines was created based on previous works in the domain. The guidelines were refined through several rounds of inter-annotator agreement (IAA) consisting of parallel annotation of 10% of the corpus. 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.

In addition, during the DisTEMIST annotation process, an ongoing discussion took place on the content of the corpus, especially on complicated and ambiguous cases, to achieve the highest possible quality and refine these guidelines as much as possible. The DisTEMIST annotation guidelines are further discussed in Section 3.2.

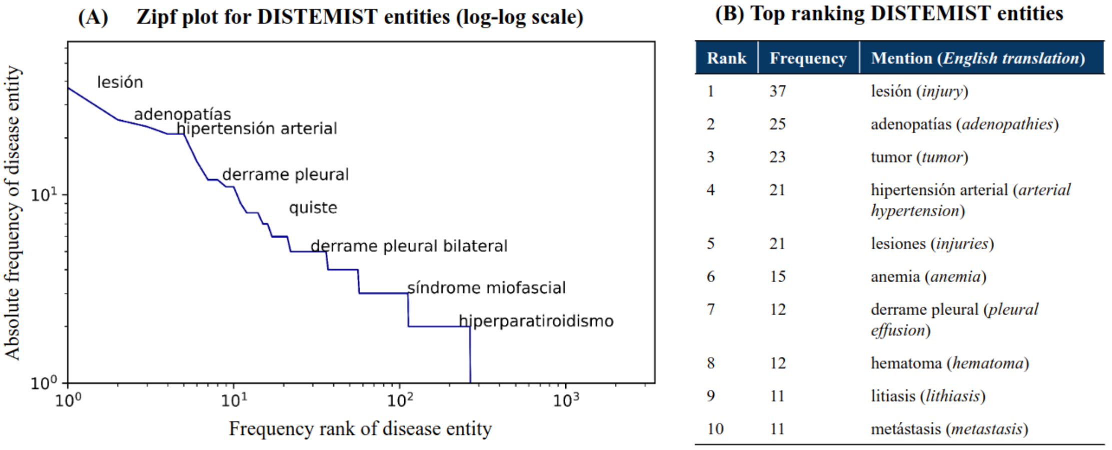
Corpus format. The DisTEMIST documents are released in plain text format with UTF-8 encoding. The annotations are included in a tab-separated document. For DisTEMIST-entities, the annotations file has the following columns: filename, mark (identifier mention id), label (ENFERMEDAD), off0 (starting position of the mention in the document), off1 (ending position of the mention in the document) and text span (see Figure 4). For DisTEMIST-linking, in addition to these six columns, the annotation file includes two more columns: codes (list of SNOMED CT concept codes linked to the mention; composite mentions with more than one associated code are concatenated by the symbol "+") and semantic relation (the relationship between the assigned code and the mention) (see Figure 4). There are two possible semantic relation values: EXACT –when the code corresponds precisely with the mention– and NARROW –when a mention corresponds to a narrower concept than the assigned SNOMED-CT code. For instance, the concept "Chorioretinal lacunae" does not exist in SNOMED-CT. Therefore, it was normalized to the SNOMED-CT ID 302893000: "Chorioretinal disorder" and a NARROW relation has been assigned to the mapping.

Corpus statistics. The DisTEMIST corpus contains 1,000 documents, which include 16,678 sentences and 406,318 tokens. The corpus was randomly split into two subsets: training and test set. The test set is used for evaluation purposes of participating teams and consists of 250 records. All documents contain disease mentions. There are 10,665 disease mentions, and each of them was manually mapped to a SNOMED CT term. There are 7,303 unique codes. During the shared task, the entire training set was released annotated, but not all documents have the mentions normalized to SNOMED CT: only 585 documents have their mentions normalized to SNOMED CT, since we could not guarantee the quality of the normalization of the remaining 165 documents. The normalization of these 165 documents has been released after the shared task. See Table 1 for the DisTEMIST corpus general statistics.


Mention string
Snomed-CT code
Semantic relationship between code and mention (NARROW or EXACT)

```
filename mark label      off0  off1  span                                  code       semantic_relation
doc1     T1   ENFERMEDAD 960   984   lagunas coriorretinianas              302893000  NARROW
doc1     T2   ENFERMEDAD 462   496   anomalías de la migración neuronal   253146009  EXACT
doc1     T3   ENFERMEDAD 837   855   afectación macular                    312999006  EXACT
```

Figure 5: Example of DisTEMIST-linking annotation.



(B) Top ranking DISTEMIST entities

Rank	Frequency	Mention (*English translation*)
1	37	lesión (*injury*)
2	25	adenopatías (*adenopathies*)
3	23	tumor (*tumor*)
4	21	hipertensión arterial (*arterial hypertension*)
5	21	lesiones (*injuries*)
6	15	anemia (*anemia*)
7	12	derrame pleural (*pleural effusion*)
8	12	hematoma (*hematoma*)
9	11	litiasis (*lithiasis*)
10	11	metástasis (*metastasis*)

Figure 6: (A) Zipfs plot of all DisTEMIST ENFERMEDAD ["disease"] entities. (B) Most frequent ENFERMEDAD ["disease"] mentions of the DisTEMIST corpus

Figure 6 (A) shows the statistical profile of the disease entities present in the entire DisTEMIST corpus. It presents the relation between disease entity mention frequency and the corresponding entity rank when listing diseases according to absolute frequency. It matches the statistical corpus characteristics observed for token frequencies of other corpora. In Figure 6 (B), we can appreciate that the most common mentions are generally short and composed of one or two tokens (*injury, lymphadenopathy, tumor, hypertension*, etc). On the other hand, disease mentions with frequency equal to one are usually longer and include mentions such as "bloque anquilótico de la articulación témporomandibular" (*temporomandibular joint ankylosis*) or "atrofia del nervio óptico derecho" (*atrophy of the right optic nerve*).

Cross-mappings. Although SNOMED CT terminology is commonly employed in clinical scenarios, literature indexing applications are frequently based on Medical Subject Headings (MeSH)[8] or Descriptores en Ciencias de la Salud (DeCS)[9], while clinical coders might be more

[8] https://www.ncbi.nlm.nih.gov/mesh/

[9] https://decs.bvsalud.org/E/homepagee.htm

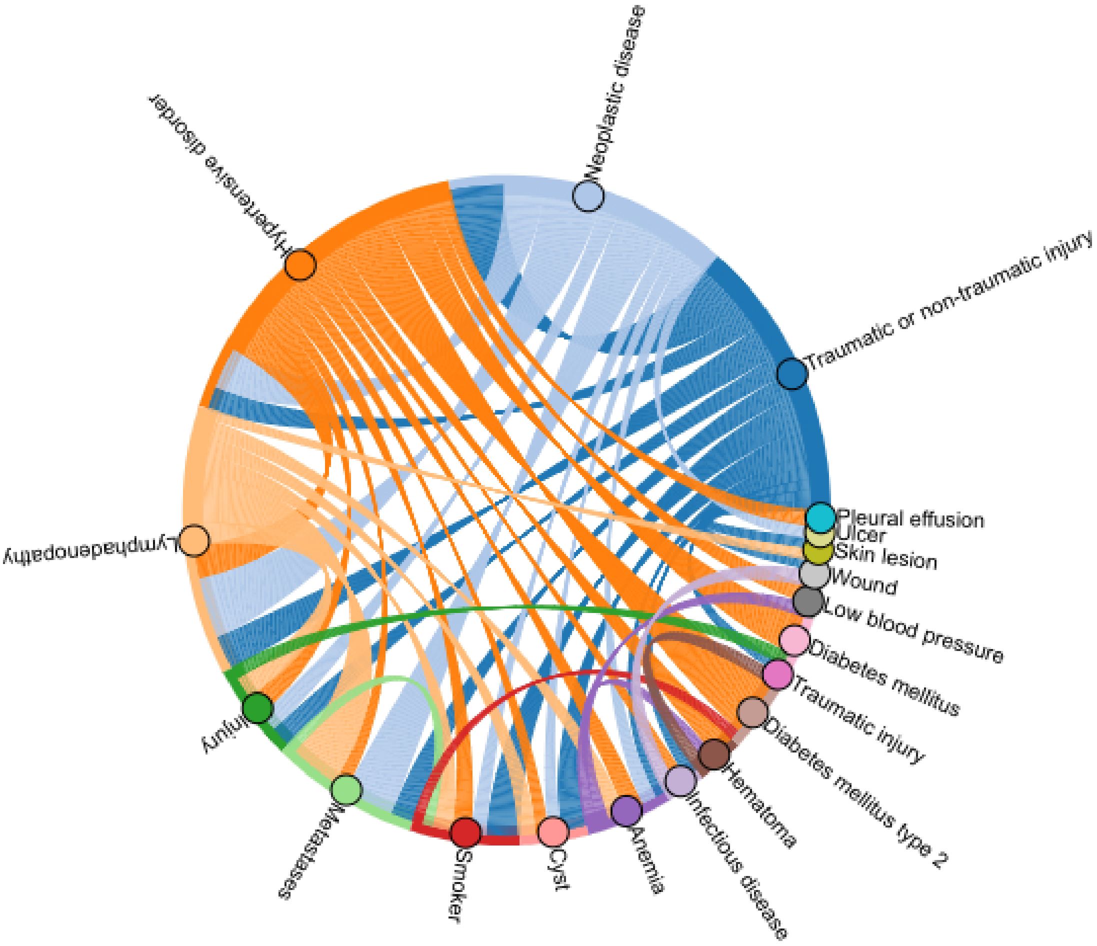


Figure 7: Chord diagram of the DisTEMIST corpus. Here we show the co-mentions of the top 20 SNOMED CT codes in the entire DisTEMIST corpus. Instead of the SNOMED CT IDs, we show the main terms to ease the understanding of the plot. Notably, the top 4 codes (417163006 or "Traumatic or non-traumatic injury", 55342001 or "Neoplastic disease", 38341003 or "Hypertensive disorder", and 30746006 "Lymphadenopathy") are frequently co-mentioned with many other codes.

familiar with the International Clasification of Diseases (ICD)[10].

To facilitate use of DisTEMIST identifiers, we have generated cross-mappings from the SNOMED CT code assignments of the DisTEMIST corpus to MeSH, ICD-10, HPO[11], and OMIM[12]. The cross-mappings were performed through the UMLS Metathesaurus and can be found at Zenodo[13].

[10] https://www.who.int/standards/classifications/classification-of-diseases

[11] https://hpo.jax.org/app/

[12] https://www.omim.org/

[13] https://doi.org/10.5281/zenodo.6408476

Table 1
DisTEMIST Gold Standard corpus statistics

	Documents	Annotations	Unique codes	Sentences	Tokens
Training	750	8,066	4,819	12,499	305,166
Test	250	2,599	2,484	4,179	101,152
Total	1,000	10,665	7,303	16,678	406,318

3.2. DisTEMIST Annotation Guidelines

The DisTEMIST corpus was manually annotated by clinical experts following the DisTEMIST guidelines. These guidelines contain rules for annotating diseases in Spanish clinical cases and for mapping these annotations to SNOMED CT.

Guidelines were created *de novo* by clinical experts and defined after several cycles of quality control and annotation consistency analysis. The DisTEMIST Annotation Guidelines are available at Zenodo[14].

The version 1 of the guidelines contains 28 pages. The annotation rules are distributed into general, positive, negative and special rules. General rules provide the basic do's and don'ts of the annotation process; positive and negative rules describe what must or mustn't be annotated as a disease, respectively; special rules are added for situations that are hard to generalize or exceptions to the other rules. Additionally, a set of rules specific to oncology mentions was added as the language used in clinical cases related to this specialty proved to be more specific and harder to annotate. The oncology rules are partially based in the CANTEMIST Guidelines [23].

All in all, there is a total of 52 rules: 16 general, 10 positive, 12 negative, 8 special and 6 for oncology. Other than the rules, the guidelines also include a short discussion of the task's importance, a short corpus characterization, basic information about the task and annotation process, as well as indications and resources for the annotators.

The DisTEMIST guidelines have been successfully adapted to other data types, such as medical narratives and social media[15].

3.3. DisTEMIST Multilingual Silver Standard

To foster the development of multilingual tools and generate systems not only for Spanish but also for content in other languages, we have generated the DisTEMIST corpus in 6 languages (English, Portuguese, Catalan, Italian, French, and Romanian) by transferring the Gold Standard annotations to machine-translated versions of the corpus files. The resulting Silver Standards include not only these annotations but also the corresponding SNOMED CT mapping for each mention.

The Gold Standard transfer process was performed as follows:

1. The text files were translated from Spanish to the target languages with a neural machine translation system. Translations were done through combination of several machine

[14] https://doi.org/10.5281/zenodo.6458078
[15] https://temu.bsc.es/socialdisner/annotation-guidelines/

translation tools, with the exception of the Catalan translation which was obtained with the SoftCatala API[16]. These systems were chosen due to their perceived quality to ensure high quality translations. In addition, manual checks of the translation outputs were performed.

2. A list of all annotations, individually and without context, was translated with the same neural machine translation system.
3. The translated annotations were transferred to the translated text files using an annotation transfer technology. The transferred annotations carry as well the SNOMED CT normalisation. Therefore, the output multilingual corpus has disease mentions annotated and linked to SNOMED CT terms.

In more detail, the annotation transferring consisted of these steps:

1. For each language, a TSV file is created with the individual Gold Standard annotations, their translation and the translation's lemma. The lemmas were obtained using spaCy[17].
2. To add the GS normalization, an additional TSV file is created with the annotations in Spanish and their associated SNOMED-CT code.
3. Each document in the corpus is iterated through. The GS annotations are read and stored and the translated text file is retrieved.
4. For each document, a dictionary is created that contains all of its annotations and their corresponding translations, lemma and code.
5. A look-up system is used to find the translated text fragments in the translated text file and create new annotations. By using only the annotations present in each GS file, we aim to reduce the number of false positives and negatives. Three different options were used for each annotation: the translation, the lemmatized form and the original annotation in Spanish. The last two were only used if the previous one did not return any results.

The annotation created by the annotation transfer process constitute a Silver Standard as the resulting annotation is only approximate (Table 2). Still, these corpora might help fill a data gap in the chosen languages. Systems trained on the data should generate acceptable results and researchers who are native speakers of the language may correct the errors in the annotation without much difficulty by comparing it with the GS.

In order to manually assess the quality of the created annotations, we carried out a short, manual error analysis of the transferred files. It helped us understand the following error sources:

- **Synonyms.** Translating the full-text documents and the annotations separately results in inconsistencies between both, as the machine translation system might return different synonyms for concepts isolated or in context. For example, the system might translate 'niño' (*child*) in French as 'enfant' in one instance and as 'garçon' in the other.
- **Gendered words.** Another of these inconsistencies is related to current biases in Machine Translation systems. Without context, gender-neutral words in Spanish were often translated to their masculine counterpart in the target language. In contrast, the full-text translation took context into account and for the most part correctly gendered words.

[16] https://www.softcatala.org/traductor/

[17] https://spacy.io/

Table 2
DisTEMIST Multilingual Silver Standard corpus statistics

		Documents	Annotations	Unique Snomed IDs	Sentences	Tokens
Catalan	**Training**	750	8739	3365	12476	305004
	Test	250	1226	397	4169	100992
English	**Training**	750	6650	2388	12582	293771
	Test	250	1125	361	4212	97574
French	**Training**	750	6447	2290	12645	327480
	Test	250	1106	330	4219	108271
Italian	**Training**	750	6468	2424	12540	305250
	Test	250	1057	347	4197	100851
Portuguese	**Training**	750	6613	2199	12551	301231
	Test	250	1021	329	4197	99818
Romanian	**Training**	750	4338	1803	12531	305884
	Test	250	707	287	4189	101223

- **Word inflection.** Transferal to Romanian was probably the hardest due to the language's grammar. Being a more morphologically-complex language than Spanish in some aspects, many of the translated entities lacked grammatical information like noun cases. As a result, they do not match their translation in the full-text document, which used the grammatical information from the sentence context and correctly added case to many nouns.

There are various possible solutions to these issues, such as creating a complete list of the possible morphological variants of a translation, using a more intricate lookup system or creating a custom Machine Translation system that takes annotated entities into account. These options are outside the scope of our current work, but we plan to explore them in the future in order to expand this technology to more languages. As an initial solution, for some languages we manually enriched the entities list with some of the most common missing entities based on our error analysis.

An overview of the DisTEMIST Multilingual Silver Standard statistics is shown in Table 2.

The DisTEMIST Multilingual Silver Standard is available at Zenodo[18]. Besides, users can visualize the multilingual resources on a Brat server[19].

3.4. DisTEMIST Gazetteer

The July 31, 2021 release of the SNOMED CT International Edition comprised more than 350,000 clinically relevant concepts. Selection of the SNOMED CT subset with concepts relevant to DisTEMIST was thus necessary to facilitate normalization. The DisTEMIST gazetteer contains main terms and synonyms from the relevant branches of SNOMED CT for the grounding of disease mentions. Mentions belonging to SNOMED CT hierarchies such as disorder, finding, or

[18] https://doi.org/10.5281/zenodo.6408476

[19] https://temu.bsc.es/mDistemist/#/translations/, https://temu.bsc.es/mDistemist/diff.xhtml#/translations/en/train/S0004-06142005000900013-1?diff=/gold-standard/train/

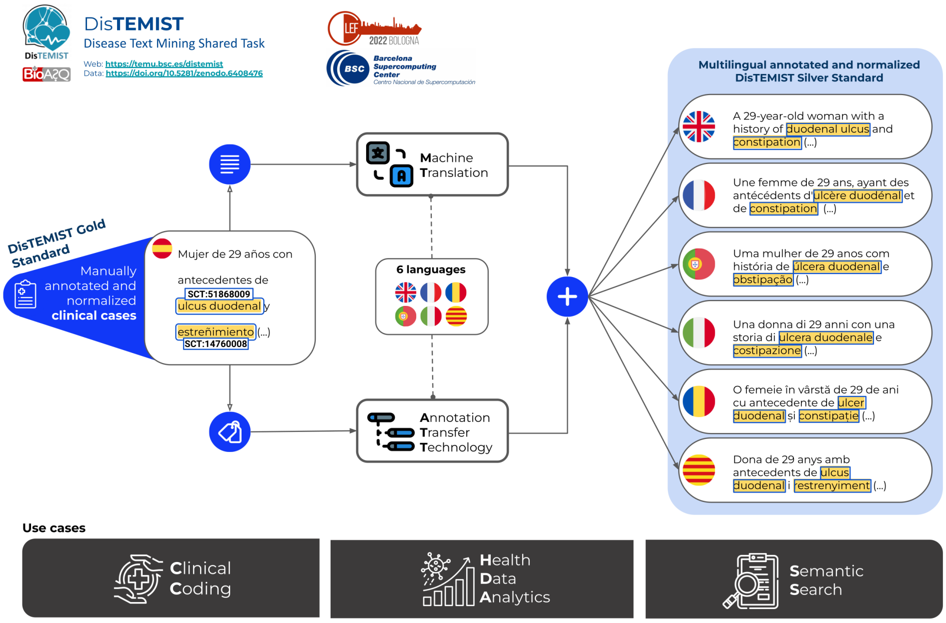

DisTEMIST
Disease Text Mining Shared Task
Web: https://temu.bsc.es/distemist
Data: https://doi.org/10.5281/zenodo.6408476
DisTEMIST
BioA2Q
LEF 2022 BOLOGNA
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación
DisTEMIST Gold Standard
Manually annotated and normalized clinical cases
Mujer de 29 años con antecedentes de
SCT:51868009
ulcus duodenal y
estreñimiento (...)
SCT:14760008
Machine Translation
6 languages
Annotation Transfer Technology
Multilingual annotated and normalized DisTEMIST Silver Standard
A 29-year-old woman with a history of duodenal ulcus and constipation (...)
Une femme de 29 ans, ayant des antécédents d'ulcère duodénal et de constipation (...)
Uma mulher de 29 anos com história de úlcera duodenal e obstipação (...)
Una donna di 29 anni con una storia di ulcera duodenale e costipazione (...)
O femeie în vârstă de 29 de ani cu antecedente de ulcer duodenal și constipație (...)
Dona de 29 anys amb antecedents de ulcus duodenal i restrenyiment (...)
Use cases
Clinical Coding
Health Data Analytics
Semantic Search

Figure 8: Overview of the DisTEMIST Multilingual Silver Standard corpus generation and use cases.

morphological abnormality were included. The test set mentions whose assigned SNOMED CT term is not included in the versions 1.0 and 2.0 of the DisTEMIST gazetteer were not considered for the shared task evaluation.

All concepts of the SNOMED CT disorder hierarchy were incorporated in the process of creating the dictionary. Additionally, sub-branches of other hierarchies have been incorporated after a manual validation carried out by the clinical experts who normalized the corpus. The dictionary has 147,280 entries, of which 111,177 are SNOMED CT main terms. There are 111,180 unique SNOMED CT codes from 18 SNOMED hierarchies, the most common one being disorder with 142,889 dictionary entries. The DisTEMIST Gazetteer is available in tab-separated format at Zenodo[20].

4. Results

DisTEMIST contained two independent sub-tasks: DisTEMIST-entities and DisTEMIST-linking. Participants could choose whether to submit results for one or both sub-tasks. Participants could submit up to 5 runs for each sub-task.

4.1. Participation Overview

DisTEMIST has received a large attention from the community. Indeed, 9 teams submitted their predictions and a total of 159 teams had registered for this task. All 9 teams participated in

[20] https://doi.org/10.5281/zenodo.6458114

Table 3
DisTEMIST participation summary.

	DisTEMIST-entitites	DisTEMIST-linking	Total
Participant teams	9	7	9
Submitted runs	19	13	32

Table 4
DisTEMIST team overview. A/I stands for academic or industry institution. In the Tasks column, E stands for DisTEMIST-entities, L for DisTEMIST-linking.

Team Name	Affiliation	Tasks	Ref.	Tool URL
PICUSLab	PICUS	E/I	[31]	–
HPI-DHC	University of Potsdam, Germany	E/I	[32]	[33]
SINAI	Universidad de Jaén, Spain	E/I	[34]	[35]
Better Innovations Lab Norwegian Centre for E-health Research	Better/NSE	E/I	[36]	–
NLP-CIC-WFU	Instituto Politécnico Nacional, Mexico	E	[37]	[38]
PU++	IIMAS UNAM, Mexico	E/I	[39]	-
Terminología	Hospital Italiano Buenos Aires, Argentina	E/I	[40]	-
iREL	IIIT Hyderabad, India	E	–	–
Unicage	Unicage, Portugal	E	[41]	–

DisTEMIST-entities, while 7 of them also submitted results for DisTEMIST-linking. Five runs were allowed per sub-task, so the total number of systems participating in the shared task is considerably higher: 19 for DisTEMIST-entities and 13 for DisTEMIST-linking. Additionally, as Table 4 shows, participants belonged to institutions (industry or academia) from different countries including Spain, India, Germany or Argentina.

4.2. System Results

Table 6 shows the complete results by all teams. The top-scoring results for each sub-task were:

- *DisTEMIST-entities.* PICUSLab team obtained the highest F1-score, 0.7770, as well as the highest precision (0.7915) and recall (0.7629). Teams Better/NSE, HPI-DHC and SINAI obtained as well F1-scores over 0.73.
- *DisTEMIST-linking.* The highest F1-score (0.5657), precision (0.6207) and recall (0.5196) were obtained by HPI-DHC.

4.3. Error Analysis

False Negatives. Missed annotations are longer. It is clear that annotations with low frequency in the training set are more difficult to predict in the test set. Indeed, there is a negative linear correlation of -0.21 between mention training set frequency and the number of False Negatives. However, analysing the participants' submissions, we found that a mention length is the strongest indicator of mention "difficulty". The linear correlation between mention

Table 5
DiseaseTagIt-Base performance comparison.

	Precision	Recall	F1-score
(1) DiseaseTagIt-Base w. DisTEMIST training	0.7225 ± 0.0226	0.6307 ± 0.0101	0.6733 ± 0.0113
(2) DiseaseTagIt-Base w. automatic predictions	0.7654 ± 0.0193	0.6964 ± 0.0182	0.7289 ± 0.0071
(3) DiseaseTagIt-Base w. augmented dataset	0.7342 ± 0.0089	0.7179 ± 0.0126	0.7259 ± 0.0101

length and number of False Negatives is 0.41. Examples of test set mentions not predicted by any of the systems are "tortuosidad de tronco celíaco y arteria hepática" (*tortuosity of the celiac trunk and hepatic artery*) or "retraso en la erupción de los incisivos inferiores y del canino" (*delayed eruption of lower incisors and canine tooth*). On the other hands, test set mentions predicted by all DisTEMIST systems are, for instance, some occurrences of "adenopatías" (*lymphadenopathy*), "cistitis"(*cystitis*) or "derrame pleural" (*pleural effusion*).

4.4. Methodologies

Participants have modelled the DisTEMIST-entities sub-task as a NER problem. Currently, large pre-trained transformer-like models are typically employed to solve this task. Following this trend, team PICUSLab has obtained the highest micro-average Precision, Recall and F1-score in the DisTEMIST-entities sub-task with a NER system based on a pre-trained biomedical Spanish transformer model. Similarly, HPI-DHC and SINAI team obtained the second and third highest Recall and F1-scores with the Spanish Clinical Roberta model.

Regarding DisTEMIST-linking, the preferred strategy has been to (1) detect the disease mentions with a NER system, (2) vectorize the disease mentions, as well as the target ontology terms/synonyms/descriptions, (3) find the best match by computing the vector similarity between the disease mention and the ontology terms, synonyms and descriptions. In this strategy, the quality of the NER system is extremely important to obtain a competitive linking system. HPI-DHC obtained the highest micro-average Precision, Recall and F1-score in the DisTEMIST-linking sub-task using an ensemble of a TF-IDF, character-n-gram based approaches and multilingual embeddings using SapBERT. On the other hand, team Better Innovations Lab / Norwegian Centre for E-health Research used FastText embeddings instead of the combination chosen by HPI-DHC, and their chosen distance metric was Approximate Nearest Neighbour similarity.

As shown by Figure 9, DisTEMIST participants have employed mostly spaCy and Python Deep Learning libraries (PyTorch and HuggingFace) to develop their systems. This is consistent with the architectures described in the previous paragraphs and with the system descriptions referenced in Table 4.

4.5. DisTEMIST Spanish Silver Standard

The DisTEMIST test set was released together with an additional collection of 2750 clinical case documents in Spanish from various medical disciplines, aka the background set. The background set is useful to examine whether systems were able to scale to larger data collections and to avoid

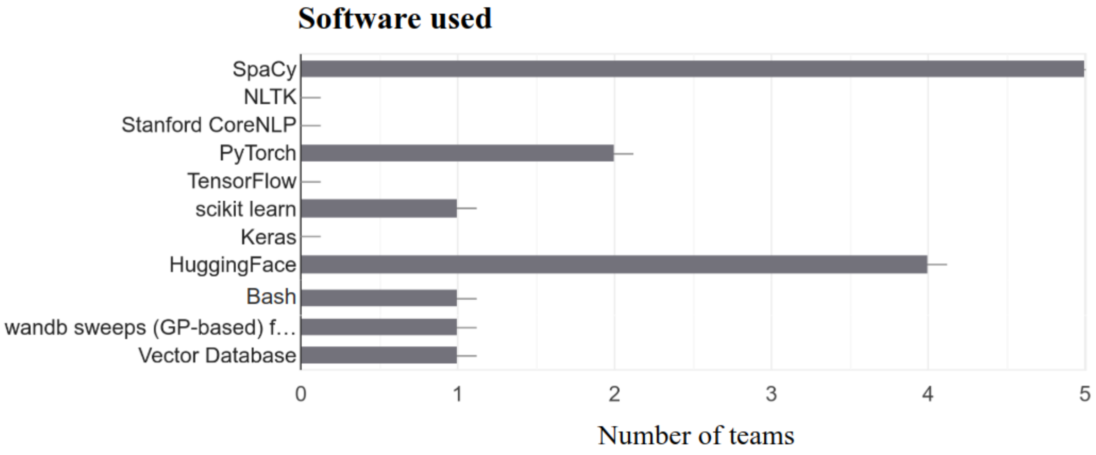


Figure 9: Overview of the software used by DisTEMIST participants.

manual annotation correction. Participants have generated automatic mention predictions for the test and the background set, although they were only evaluated on the test set predictions.

The predictions from all participants for this background set will be harmonized and constitute the DisTEMIST Spanish Silver Standard corpus, similar to the CALBC initiative [42], to the Cantemist [23], CodiEsp [17], MESINESP2021 [3], ProfNER [25], and PharmaCoNER [21] shared tasks.

The Spanish Silver Standard will be a high-quality collection of annotated clinical documents in Spanish and it will serve to foster the development of disease recognition and linking resources. To prove this claim, we have compared the DiseaseTagIt-Base baseline system (1) trained with the DisTEMIST training data (manually annotated data), (2) trained with the HPI-DHC background predictions (automatically annotated data) and (3) trained with the DisTEMIST training data + the HPI-DHC background predictions (augmented dataset) (Table 5). In the three cases, the network was trained with an Early Stop methodology based on the validation f1-score and patience equal to 3 epochs. The experiment was repeated three times.

When comparing the results obtained by the three systems, we can observe that the performance improves substantially when the automatically annotated data is added. The model trained with the augmented datasets shows an improvement of almost 0.5 F1-score with respect to the model that was trained using only manually annotated data.

The Spanish Silver Standard will be released on the Zenodo Medical NLP community.

5. Discussion

The DisTEMIST shared task has been a pioneer community effort on NER and normalization of diseases mentions to SNOMED CT in clinical documents written in Spanish. To foster the development of disease NER and linking resources, we have released the DisTEMIST corpus.

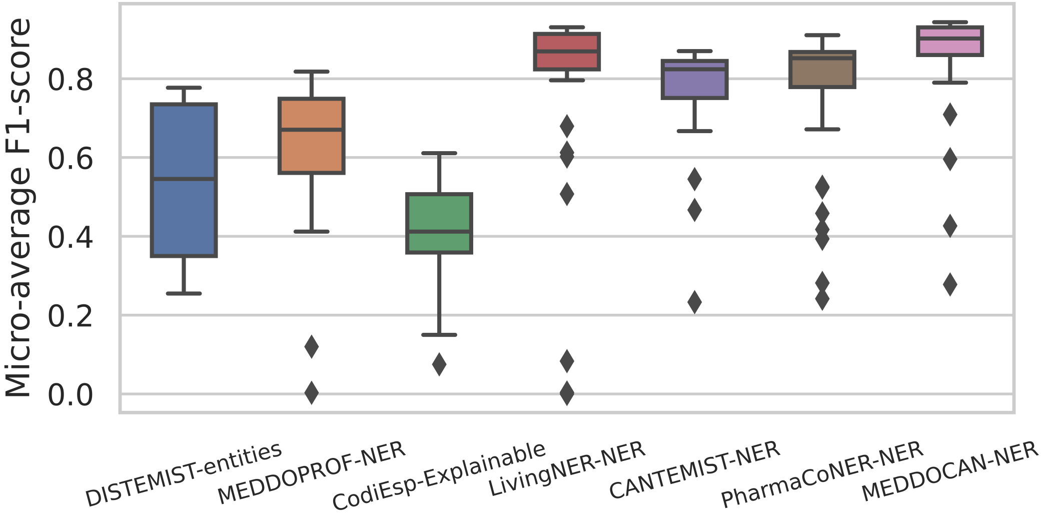


Figure 10: Comparison of micro-average f1-score obtained by participant teams of clinical NER shared tasks in Spanish.

It is the first Gold Standard text corpus of Spanish clinical documents with disease mentions, manually mapped to the SNOMED-CT terminology.

The DisTEMIST corpus was created following strict annotation guidelines that are made public to allow the corpus extension and adaptation to other languages or domains. Our group has applied the DisTEMIST annotation guidelines (and systems trained with the DisTEMIST corpus) to Electronic Health Records (in particular, to hospital discharge and radiology reports) and social media with promising results.

To enhance the interoperability between different data sources, we have generated two additional resources as part of DisTEMIST. On the one hand, taking into account multilingual scenarios and the general lack of annotated data in other languages, we have released the DisTEMIST Multilingual Corpus. It contains the DisTEMIST corpus documents, translated to 6 languages (English, French, Italian, Portuguese, Catalan, and Romanian) and with automatically generated disease mention annotations mapped to SNOMED CT. On the other hand, we have also released the DisTEMIST cross-mappings, which links the Gold Standard mappings to SNOMED CT to four different terminologies (MeSH, ICD-10, HPO and OMIM).

All these resources have attracted the attention of the community. Participant teams have developed 38 competitive systems, mainly based on pre-trained transformer language models, evaluated against the DisTEMIST corpus manual annotations. Additionally, they have generated automatic predictions for 2,750 documents that will be harmonized to create the DisTEMIST Spanish Silver Standard. We expect that this resource can help enhance current disease recognition and entity linking, as it was shown to improve the performance of a deep learning engine.

In the last years, disease mention detection systems have been implemented and used to process a diversity of content types (e.g. scientific publications, clinical records, clinical trials, patient fora or social media) resulting in a component integrated into a variety of practically relevant application types, such as health data analytics software and study of disease trajectories, disease outbreak monitoring and surveillance, as well as epidemiology tools, extraction of disease phenotype or co-morbidities, drug discovery, repurposing and off label indications, occupational health studies, pharmacogenomics or clinical coding of diagnosis.

With the DisTEMIST shared task and resources, we envision expanding these use cases to other languages, multilingual or code-switching scenarios, and to different data sources. Besides, current systems will also benefit from the generated resources since we have proven that their addition improves system performances.

Indeed, DisTEMIST is part of a continued effort carried out by the Plan de Tecnologías del Lenguaje and the Barcelona Supercomputing Center to generate publicly accessible high-quality corpora in clinical documents in the co-official languages of Spain. This project includes MEDDOCAN [22], on the extraction of entities relevant for clinical document anonymisation, PharmaCoNER [21], the pharmacological substances, compounds and proteins and NER track, Cantemist [23], that focused on the recognition of tumor morphology mentions, CodiEsp [17], related to the detection of ICD-10 entities, MEDDOPROF [24], on the recognition of occupations, and LivingNER [43], about the recognition of species and pathogens. The seven shared tasks aimed to recognize different medical entities of relevance in clinical documents in Spanish. They have provided the community with resources (annotated corpora, evaluation libraries, etc.) to solve the challenges. Table 7 compares the main characteristics of the seven shared tasks, and the Figure 10 shows side-by-side the micro-average f1-score distribution across all seven shared tasks.

In the future, we plan to expand the DisTEMIST Multilingual Silver Standard to other under-represented languages such as Galician and to generate a Gold Standard subset of each language to create high-quality benchmarks in the seven languages. Furthermore, we are currently preparing the same Gold Standard corpus for other medical entities (symptoms and medical procedures).

A. Additional tables

B. Acknowledgments

We acknowledge the Encargo of Plan TL (SEAD) to BSC for funding, and the scientific committee for their valuable comments, their guidance, and their help with the review of the proceedings. Due to the relevance of diseases for biomaterials and implants, this project has received funding from the European Union's Horizon Europe Coordination & Support Action under Grant Agreement No 101058779. Finally, we acknowledge as well the support from AI4PROFHEALTH project (PID2020-119266RA-I00). In addition, we would like to thank the organization of BioASQ and CLEF, as well as the team from Hospital 12 de Octubre and Veratech, who worked with us to create the dataset and annotation guidelines. We acknowledge positive support from Jose Antonio Lopez-Martin.

References

[1] M. Krallinger, F. Leitner, A. Valencia, Analysis of biological processes and diseases using text mining approaches, Bioinformatics Methods in Clinical Research (2010) 341–382.

[2] R. Islamaj Dogan, G. C. Murray, A. Névéol, Z. Lu, Understanding pubmed® user search behavior through log analysis, Database 2009 (2009).

[3] L. Gasco, A. Nentidis, A. Krithara, D. Estrada-Zavala, R. T. Murasaki, E. Primo-Peña, C. Bojo Canales, G. Paliouras, M. Krallinger, et al., Overview of bioasq 2021-mesinesp track. evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials, CEUR Workshop Proceedings, 2021.

[4] P. L. Schuyler, W. T. Hole, M. S. Tuttle, D. D. Sherertz, The umls metathesaurus: representing different views of biomedical concepts., Bulletin of the Medical Library Association 81 (1993) 217.

[5] A. R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program, Proc. AMIA Symp. (2001) 17–21.

[6] A. R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (2010) 229–236.

[7] J. G. Mork, A. Jimeno-Yepes, A. R. Aronson, et al., The nlm medical text indexer system for indexing biomedical literature., BioASQ@ CLEF 1 (2013).

[8] O. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, Journal of the American Medical Informatics Association 18 (2011) 552–556. URL: https://doi.org/10.1136/amiajnl-2011-000203. doi:10.1136/amiajnl-2011-000203. arXiv:https://academic.oup.com/jamia/article-pdf/18/5/552/33015279/18-5-552.pdf.

[9] R. Islamaj Doğan, R. Leaman, Z. Lu, Ncbi disease corpus: A resource for disease name recognition and concept normalization, Journal of Biomedical Informatics 47 (2014) 1–10. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001974. doi:https://doi.org/10.1016/j.jbi.2013.12.006.

[10] Y.-F. Luo, S. Henry, Y. Wang, F. Shen, O. Uzuner, A. Rumshisky, The 2019 n2c2/UMass Lowell shared task on clinical concept normalization, Journal of the American Medical Informatics Association 27 (2020) 1529–e1. URL: https://doi.org/10.1093/jamia/ocaa106. doi:10.1093/jamia/ocaa106. arXiv:https://academic.oup.com/jamia/article-pdf/27/10/1529/39739985/ocaa106.pdf.

[11] R. Leaman, R. Islamaj Doğan, Z. Lu, DNorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (2013) 2909–2917. URL: https://doi.org/10.1093/bioinformatics/btt474. doi:10.1093/bioinformatics/btt474. arXiv:https://academic.oup.com/bioinformatics/article-pdf/29/22/2909/888873/btt47

[12] H. Suominen, S. Salanterä, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, G. Zuccon, Overview of the ShARe/CLEF ehealth evaluation lab 2013, in: Lecture Notes in Computer Science, Lecture notes in computer science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 212–231.

[13] M. Oronoz, K. Gojenola, A. Pérez, A. Ilarraza, A. Casillas, On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions, Journal of Biomedical

Informatics 56 (2015). doi:10.1016/j.jbi.2015.06.016.

[14] I. Moreno, E. Boldrini, P. Moreda, M. T. Romá-Ferri, Drugsemantics: A corpus for named entity recognition in spanish summaries of product characteristics, Journal of biomedical informatics 72 (2017) 8–22.

[15] P. Baez Benavides, F. Villena, M. Rojas, M. Durán Fernández, J. Dunstan, The chilean waiting list corpus: a new resource for clinical named entity recognition in spanish, 2020, pp. 291–300. doi:10.18653/v1/2020.clinicalnlp-1.32.

[16] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, A. Moreno-Sandoval, A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine, BMC Med. Inform. Decis. Mak. 21 (2021) 69.

[17] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020., in: CLEF (Working Notes), 2020.

[18] T. Benson, Principles of health interoperability HL7 and SNOMED, Springer Science & Business Media, 2012.

[19] K. Donnelly, et al., Snomed-ct: The advanced terminology and coding system for ehealth, Studies in health technology and informatics 121 (2006) 279.

[20] A. Intxaurrondo, M. Marimón, A. González-Agirre, J. A. Lopez-Martin, H. Rodriguez, J. Santamaria, M. Villegas, M. Krallinger, Finding mentions of abbreviations and their definitions in spanish clinical cases: The barr2 shared task evaluation results., IberEval@ SEPLN 2150 (2018) 280–289.

[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] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results., in: IberLEF@ SEPLN, 2019, pp. 618–638.

[23] A. Miranda-Escalada, E. Farré, 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., IberLEF@ SEPLN (2020) 303–323.

[24] 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.

[25] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, M. Krallinger, The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora, in: Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task, 2021, pp. 13–20.

[26] C. Rodriguez-Penagos, A. Nentidis, A. Gonzalez-Agirre, A. Asensio, J. Armengol-Estapé, A. Krithara, M. Villegas, G. Paliouras, M. Krallinger, Overview of mesinesp8, a spanish medical semantic indexing task within bioasq 2020, Working Notes of CLEF (2020) 1–12.

[27] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 239–263.

[28] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, J. Armengol-Estapé, Medical word embeddings for Spanish: Development and evaluation, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 124–133. URL: https://aclanthology.org/W19-1916. doi:10.18653/v1/W19-1916.

[29] J. Armengol-Estapé, F. Soares, M. Marimon, M. Krallinger, Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts, Genomics & informatics 17 (2019).

[30] 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.

[31] V. Moscato, M. Postiglione, G. Sperl[í], Biomedical spanish language models for entity recognition and linking at bioasq distemist (2022).

[32] F. Borchert, M.-P. Schapranow, Hpi-dhc @ bioasq distemist: Spanish biomedical entity linking with cross-lingual candidate retrieval and rule-based reranking (2022).

[33] HPI-DHC, distemist_bioasq_2022, https://github.com/hpi-dhc/distemist_bioasq_2022, 2022.

[34] M. Chizhikova, J. Collado-Montañez, P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, M. T. Martín-Valdivia, Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions (2022).

[35] SINAI, Spanish_disease_finder, https://huggingface.co/chizhikchi/Spanish_disease_finder, 2022.

[36] M. Bernik, R. Tovornik, B. Fabjan, L. Marco-Ruiz, Diagñoza: a natural language processing tool for automatic annotation of clinical free text with snomed-ct (2022).

[37] A. Tamayo, D. A. Burgos, A. Gelbukh, mbert and simple post-processing: A baseline for disease mention detection in spanish (2022).

[38] NLP-CIC-WFU, Nlp-cic-wfu contribution to distemist 2022 sub-track 1 entities, https://colab.research.google.com/drive/1nRpOSYd5iaLbo6O-An1hIputoHG7SIJz?usp=sharing, 2022.

[39] 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 (2022).

[40] J. Castano, M. L. Gambarte, C. Otero, D. Luna, A simple terminology-based approach to clinical entity recognition (2022).

[41] A. Neves, Unicage at distemist - named entity recognition system using only bash and unicage tools (2022).

[42] D. Rebholz-Schuhmann, A. J. J. Yepes, E. M. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, U. Hahn, Calbc silver standard corpus, Journal of bioinformatics and computational biology 8 (2010) 163–179.

[43] 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).

[44] L. Lange, H. Adel, J. Strötgen, Nlnde: The neither-language-nor-domain-experts' way of spanish medical document de-identification, arXiv preprint arXiv:2007.01030 (2020).

[45] Y. Xiong, Y. Shen, Y. Huang, S. Chen, B. Tang, X. Wang, Q. Chen, J. Yan, Y. Zhou, A deep learning-based system for PharmaCoNER, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 33–37. URL: https://aclanthology.org/D19-5706. doi:10.18653/v1/D19-5706.

[46] N. García-Santa, K. Cetina, L. Cappellato, C. Eickhoff, N. Ferro, A. Nevéol, Fle at clef ehealth 2020: Text mining and semantic knowledge for automated clinical encoding., in: CLEF (Working Notes), 2020.

[47] S. Cossin, V. Jouhet, Iam at clef ehealth 2020: Concept annotation in spanish electronic health records., in: CLEF (Working Notes), 2020.

[48] Y. Xiong, Y. Huang, Q. Chen, X. Wang, Y. Nic, B. Tang, A joint model for medical named entity recognition and normalization, Proceedings http://ceur-ws. org ISSN 1613 (2020) 17.

[49] L. Lange, H. Adel, J. Strötgen, Boosting transformers for job expression extraction and classification in a low-resource setting, arXiv preprint arXiv:2109.08597 (2021).

Table 6

Results of DisTEMIST systems. MiP, MiR and MiF stands for micro-averaged Precision, Recall and F1-score. DisTEMIST-e stands for DisTEMIST-entities and DisTEMIST-l stands for DisTEMIST-linking. The best result is bolded, and the second-best is underlined.

		DisTEMIST-e			**DisTEMIST-l**		
Team Name	**Run Name**	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**
PICUSLab	NER_results	0.7915	**0.7629**	**0.777**	-	-	-
	EL_results	-	-	-	0.2814	0.2748	0.278
	EL_results (post-workshop)	-	-	-	0.274	0.2712	0.2726
HPI-DHC	1-r.c.e.-linear-lr	0.7302	0.7363	0.7332	-	-	-
	2-r.c.e.-constant-lr	0.7302	0.7259	0.728	-	-	-
	3-r.c.e.-linear-lr-post-process	0.7434	0.7483	0.7458	-	-	-
	4-r.c.e.-constant-lr-post-process	0.7417	0.7371	0.7394	-	-	-
	1-tf_idf_ngrams_distemist	-.	-	-	0.3576	0.3646	0.3611
	2-sap_umls_large_distemist	-	-	-	0.3641	0.3738	0.3689
	3-ensemble	-	-	-	0.4678	0.389	0.4248
	4-ensemble-reranking	-	-	-	0.5427	0.4513	0.4928
	5-ensemble-reranking-postprocess	-	-	-	**0.6207**	**0.5196**	**0.5657**
SINAI	run1-clinical_model	0.7519	0.7221	0.7367	-	-	-
	run2-biomedical_model	0.752	0.7259	0.7387	-	-	-
	run1-clinical_model	-	-	-	0.4163	0.4081	0.4122
	run2-biomedical_model	-	-	-	0.4134	0.4069	0.4101
Better Innovations Lab & Norwegian Centre for E-health Research	run1-ner	0.7724	0.6925	0.7303	-	-	-
	run2-ner-limited	**0.7926**	0.6574	0.7187	-	-	-
	run1-snomed	-	-	-	0.5478	0.4577	0.4987
	run2-snomed-limited	-	-	-	0.5497	0.4549	0.4978
NLP-CIC-WFU	System_mBERT	0.6095	0.4938	0.5456	-	-	-
PU++	run1_mbertD5	0.454	0.4619	0.4579	-	-	-
	run2_mbertM5	0.601	0.4488	0.5139	-	-	-
	run1-scieloBERT	-	-	-	0.2267	0.1494	0.1801
	run2-scieloBERT	-	-	-	0.2754	0.1494	0.1937
Terminología	distemist-subtrack1	0.5622	0.3772	0.4515	-	-	-
	distemist-subtrack2	-	-	-	0.4795	0.2292	0.3102
iREL	iREL	0.4984	0.3576	0.4164	-	-	-
Unicage (post-workshop)	STEM_XXL_LEX_3spc	0.2055	0.3464	0.258	-	-	-
	XL_LEX_3spc	0.2486	0.3303	0.2836	-	-	-
	XL_LEX_spc	0.205	0.338	0.2552	-	-	-
	XXL_LEX_3spc	0.2478	0.3306	0.2833	-	-	-
	XXL_LEX_spc	0.2045	0.338	0.2548	-	-	-
BSC baselines	DiseaseTagIt-VT	0.1568	0.4057	0.2262	0.1003	0.1621	0.124
	DiseaseTagIt-Base	0.7146	0.6736	0.6935	0.3041	0.2336	0.2642

Table 7
Comparison of Named Entity Recognition shared tasks in Spanish clinical documents.

Shared task	Entities	Participants	Submissions	Training instances	Test instances	IAA	Best methodology
MEDDOCAN [22]	miscellaneous	18	63	17134	5661	0.98	Recurrent Neural Network (BiLSTM-CRF) architecture with post-processing rules [44]
PharmaCoNER [21]	FARMACO	29	96	5748	4159	0.93	Transformer pre-trained language model (mBERT-CRF) architecture fine-tuned for NER [45]
CodiEsp [17]	DIAGNOSTICO & PROCEDIMIENTO	22	168	13658	4777	0.81	Transformer pre-trained language model (mBERT) fine-tuned for NER [46] and a dictionary lookup [47]
Cantemist [23]	MORFOLOGIA _NEOPLASIA	25	131	12397	3633	0.84	Transformer pre-trained language model (mBERT) fine-tuned jointly for NER and normalization [48]
MEDDOPROF [24]	miscellaneous	15	94	3658	1085	0.9	Transformer pre-trained language model (XLM-CRF) architecture fine-tuned for NER [49]
LivingNER [43]	SPECIES & HUMAN	19	62	23205	7402	0.942	Ensemble of transformer pre-trained language models (XLM-RoBERTa) fine-tuned for NER
DisTEMIST	ENFERMEDAD	9	38	8066	2599	0.82	Transformer pre-trained biomedical Spanish language model fine-tuned for NER [31]