

Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results

Antonio Miranda-Escalada^a, Eulàlia Farré^a and Martin Krallinger^a

^aBarcelona Supercomputing Center, Spain

Abstract

Cancer still represents one of the leading causes of death worldwide, resulting in a considerable healthcare impact. Recent research efforts from the clinical and molecular oncology scientific communities were able to increase considerably life expectancy of patients for some cancer types. Most of the current cancer diagnoses are primarily determined by pathology laboratories, providing an essential source for information to guide the treatment of patients with cancer. Pathology observations essentially characterize the results of microscopic or macroscopic studies of cells or tissues following a biopsy or surgery. Clinicians and researchers alike, require systems that automatically detect, read and generate structured data representations from pathology examinations. The resulting structured or coded clinical information, normalized using controlled vocabularies like the ICD-O or SNOMED-CT is critical for large-scale analysis of specific tumor types or to determine response to specific treatments or prognosis. Text mining and NLP approaches are showing promising results to transform medical text into useful clinical information, bridging the gap between free-text and structured representation of clinical information. Nonetheless, in the case of cancer text mining approaches, most efforts were exclusively focused on medical records in English. Moreover, due to the lack of high quality manually labeled clinical texts annotated by oncology experts most previous efforts, even for English relied mainly on customized dictionaries of names or rules to recognize clinical concept mentions despite the promising results of advanced deep learning technologies. To address these issues we have organized the Cantemist (CANcer TExT Mining Shared Task) track at IberLEF 2020. It represents the first community effort to evaluate and promote the development of resources for named entity recognition, concept normalization and clinical coding specifically focusing on cancer data in Spanish. Evaluation of participating systems was done using the Cantemist corpus, a publicly accessible dataset (together with annotation consistency analysis and guidelines) of manually annotated mentions of tumor morphology entities and their mappings to the Spanish version of ICD-O. We received a total of 121 systems or runs from 25 teams for one of the three Cantemist sub-tasks, obtaining very competitive results. Most participants implemented sophisticated AI approaches; mainly deep learning algorithms based on Long-Short Term Memory Units and language models (BERT, BETO, RoBERTa, etc) with a classifier layer such as a Conditional Random Field. In addition to using pre-trained language models, word and character embeddings were also explored. Cantemist corpus: <https://doi.org/10.5281/zenodo.3773228>

Keywords

IberLEF, oncology, tumor histology, named entity recognition, deep learning, normalization, pathology, Gold Standard corpus, NLP, Plan TL, text mining, EHR

1. Introduction

Cancer is one of the leading causes of mortality worldwide, producing around one in six deaths in 2018. Lung, breast and colorectal cancers are amongst the most common types of cancer of the more

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

ORCID: 0000-0002-5654-001X (A. Miranda-Escalada); 0000-0002-2646-8782 (M. Krallinger)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

than 200 varieties [1], each with certain causes, symptoms, treatments as well as underlying tissue/cellular characteristics [2]. Despite recent advances in molecular oncology, precision medicine and the characterization of genetic cancer signatures, diagnosis of cancer does rely heavily on the study of macroscopic and microscopic samples of the tumor following a biopsy or surgery. The resulting observations are usually reported by pathologists and documented in pathology reports [3]. The observations described in pathology reports, are then used by clinicians to guide decision-making and to determine the appropriate treatment and prognosis of the tumor. Most pathology reports are still found in free text form, and the manner in which each pathologist describes a tumor sample often differs due to the idiosyncrasies of clinical language (including syntactic, morphological and orthographic variation). Pathology reports constitute only one of the many document sources related to the cancer domain. Additionally, there is a humongous body of cancer-related literature [4], as well as cancer clinical trial studies, patents, tumor biobank free text metadata and of course clinical records of various kinds, including hospital discharge summaries, oncology reports, imaging reports, or lab test results. Imaging or radiology reports offer specialist interpretations of patient clinical images and relate them to signs and symptoms that can aid in supporting a correct diagnosis. There is a pressing need to convert text into useful clinical information, being able to distill and normalize the semantic representation of the rapidly growing body of textual information found in narrative reports by means of automatic text processing tools. Text mining and natural language processing (NLP) systems are becoming a possible answer for bridging the gap between free-text clinical data and structured representation of cancer information [2, 4, 5]. To efficiently access current and past information on the pathology of tumors, we need NLP solutions that automatically detect, read and codify the description provided by pathologists. Such systems would not only facilitate the daily task of pathologists in busy, overburdened hospitals, but would allow large-scale analysis of relationships between the pathology of a specific tumor and response to specific treatments, prognosis and other.

Due to the highly specialized clinical languages and the need to standardize medical vocabularies, a range of different terminology systems has been constructed for oncology data. The availability of knowledge resources and medical terminologies is key for semantic interoperability and practical clinical coding. Schulz et al. provides an overview of controlled vocabularies relevant to oncology [6]. Terminological resources and cancer dictionaries were exploited by various text-mining approaches to process oncology data [7, 8, 9, 10, 11, 12, 13], while sophisticated advanced AI-solutions for oncology text processing do face the struggle of access to high quality manually labeled text corpora [14]. One of the key terminological resources in oncology is the International Classification of Diseases - Oncology (ICD-O), a domain-specific extension of ICD created originally in 1976, which can be regarded as the lingua franca of pathologists with an extensive use within tumor registries [15]. It is available also in multiple languages including Spanish [16]. ICD-O (or CIE-O in Spanish) has two main axes, one for the topography of malignant neoplasms and one for cancer morphology. It is being used by many hospitals for clinical coding purposes. The development of effective and practically useful clinical NLP tools is a complex task, which does require quality evaluation on properly annotated clinical data. Building such NLP tools and successfully exploiting the information stored in clinical narratives was not properly addressed in the oncological domain for data in Spanish. This was due to the lack of suitable resources in the form of manually annotated Gold Standard corpora prepared by clinical experts.

To this end, we have created the first Gold Standard text corpus of tumor morphology mentions, manually mapped to the latest Spanish version of ICD-O [17]. This gold standard corpus, named CAN-TEMIST corpus, constitutes the continuation of our previous efforts to generate publicly accessible high quality corpora annotated with relevant clinical entities in Spanish [18, 19, 20, 21]. It was built following detailed annotation guidelines and exhaustive manual text labeling by domain experts. The

creation of the CANTEMIST gold standard corpus can be subdivided into two discrete steps: 1) manual text annotation, where the annotator/s recognizes and tags tumor morphology mentions in text and 2) careful assignment a specific eCIE-O-3.1 code (the Spanish equivalent of ICD-O, version 3.1) to each mention. The complexity of the normalization or mapping step (ICD-O codification) resides in the considerable variability of expressions used by pathologists to describe the same histological finding, together with changes in terminologies and classifications over time due to scientific/clinical advances.

To increase the exploitation and impact of the CANTEMIST corpus, it was used for a shared task in the context of the IberLEF 2020 evaluation initiative. This paper provided an overview the results, data, methods, outcome and future outlook of the CANTEMIST shared task.

2. Task Description

2.1. Shared task goal

Cantemist explores the automatic detection of tumor morphology mentions in medical documents in the Spanish language, as well as the assignment of eCIE-O codes (*Morfología neoplasia*, in Spanish) to each mention. To the authors' knowledge, Cantemist is the first shared task specifically focusing on Named Entity Recognition of a critical type of concept related to cancer, namely tumor morphology, in Spanish. Previous community evaluation efforts within the cancer domain include the cancer genetics shared task using data in English [22].

2.2. Sub-tasks

The Cantemist task is structured into three independent sub-tasks, each taking into account a particularly important use case scenario:

- *Cantemist-NER track*. It requires finding tumor morphology mentions automatically in text. All tumor morphology mentions are defined by their corresponding character offsets (start character and end character) in UTF-8 plain text medical documents.
- *Cantemist-NORM track*. Clinical concept normalization task that requires to return all tumor morphology entity mentions together with their corresponding eCIE-O codes, i.e. finding and normalizing tumor morphology mentions.
- *Cantemist-CODING track*. It requires returning for each of the documents a ranked list of correct eCIE-O code assignments. This is essentially an indexing or multi-label classification task (oncology clinical coding).

2.3. Shared task setting

The Cantemist track was organized in three participation periods or phases:

1. *Training phase*. During the first participation period, 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 start building their systems.
2. *Development phase*. Then the development set was released (plain text documents and their annotations). This set was used to further fine-tune and evaluate the systems.

3. *Test phase.* Finally, the test set was released. In this case, 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 the first two subtasks, Cantemist-NER and Cantemist-Norm, the main evaluation metric has been micro average f1-score. In addition, precision and recall have been computed.

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

For the Cantemist-Coding sub-track, the same metrics were computed. However, based on our experience with past clinical coding efforts, the primary evaluation metric used was Mean Average Precision (MAP). It is a ranking metric and therefore, participants has to submit their predictions ranked by confidence. Mean Average Precision (MAP) is a score extensively used in ranking problems:

$$AveP = \frac{\sum(P(k) * rel(k))}{\text{number of relevant documents}}$$

where $P(k)$ is the precision at the position k , and $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise. MAP has shown good discrimination and stability [23]. The evaluation library is available on GitHub ¹.

Additionally, as annotations with the code 8000/6 represent nearly 30% of all annotations, we also computed all metrics without taking into account this code for the Cantemist-Norm and Cantemist-Coding sub-tasks. The code 8000/6 corresponds to mentions of metastasis.

2.5. Baseline

We have compared every system to a baseline prediction, in this case a Levenshtein lexical lookup approach using a sliding window of varying length. The baseline systems essentially scans a novel input text looking for mentions previously found in the training or development annotations. For every test set document the baseline system performed the following steps:

1. Select one annotation from the training and development sets to generate a dictionary entry.
2. Scan the test set document with a sliding window the same size of the dictionary entry, plus 2 characters.
3. If the Levenshtein distance from the current text inside the window to the annotation is smaller than 1, annotate it and add to it the code that the original annotation had.
4. Select another annotation and repeat steps 2 and 3 with it.

Its results are found in Table 1.

¹<https://github.com/TeMU-BSC/cantemist-evaluation-library>

Table 1
Cantemist baseline results.

Cantemist-NER			Cantemist-Norm			Cantemist-Coding
P	R	F1	P	R	F1	MAP
0.181	0.737	0.291	0.18	0.73	0.288	0.584

Tratamiento
En espera de la realización de una nueva biopsia, se inició tratamiento de primera línea con cisplatino y gemcitabina, considerando que el tratamiento con gemcitabina podría ser efectivo en **MORFOLOGÍA NEOPLASIA** tumores de estirpe sarcomatoide.

Presentó tolerancia regular al tratamiento durante el primer ciclo, con mucositis grado 1, un episodio de fiebre sin focalidad y neutropenia grado 3, por lo que se pospuso el inicio del segundo ciclo.

Esta clínica obligó también a retrasar la biopsia quirúrgica que estaba prevista.

En la TC toracoabdominopélvica de revaloración tras dos ciclos, se observó disminución del tamaño de la masa en LSD (de 80 mm a 56 mm) y un **MORFOLOGÍA NEOPLASIA** aumento de tamaño de la lesión suprarrenal derecha (de 31 mm a 54 mm).

En octubre de 2016, el Equipo de Cirugía Plástica realizó nueva biopsia con exéresis de músculo piramidal, sin evidencia de **MORFOLOGÍA NEOPLASIA** malignidad, por lo que se realizó una TC/PET para planificar la biopsia con mayor rentabilidad.

Finalmente se realizó una biopsia guiada por TC de la lesión paramediastínica derecha, con el diagnóstico de **MORFOLOGÍA NEOPLASIA** carcinoma no célula pequeña **MORFOLOGÍA NEOPLASIA** (CPNCP) sugestivo de **MORFOLOGÍA NEOPLASIA** adenocarcinoma (TTF1 y p63 negativos), con estudio molecular negativo para ROS1, MET, ALK, KRAS y EGFR.

Figure 1: Annotated clinical case report visualized with Brat tool [24].

3. Corpus and Resources

3.1. Cantemist corpus

The Cantemist corpus is a collection of 1301 oncological clinical case reports written in Spanish. In addition to this official corpus post-workshop document collection will be released to sum a total of 1900 clinical cases. The training subset contains 501 documents, the development subsets 500, and the test subset 300. All documents of the corpus have been manually annotated by clinical experts with mentions of tumor morphology (in Spanish, “morfología de neoplasia”). Every tumor morphology mention is linked to an eCIE-O code (the Spanish equivalent of ICD-O). Figure 1 shows an example of an annotated document fragment. The Cantemist corpus is publicly available at Zenodo ².

To the best of our knowledge, there are no state-of-the-art, quality-controlled, publicly available NLP resources for oncologic histology in Spanish. The Cantemist corpus fills in this gap. It is a corpus of oncology clinical case reports written in Spanish, with tumor morphology mentions annotated and mapped to a controlled terminology, eCIE-O. It is publicly available, follows strict guidelines, and was annotated performing regular quality control analysis. It may be employed to develop new tools, to test the existing ones, and even to complete future oncology histology corpora, since the annotation guidelines are also published ³.

Spanish documents were chosen because it is the second world language with more native speakers, after Mandarin Chinese. In addition, it follows past efforts of the shared task organizers [18, 19, 20,

²<https://doi.org/10.5281/zenodo.3773228>

³<https://doi.org/10.5281/zenodo.3878178>

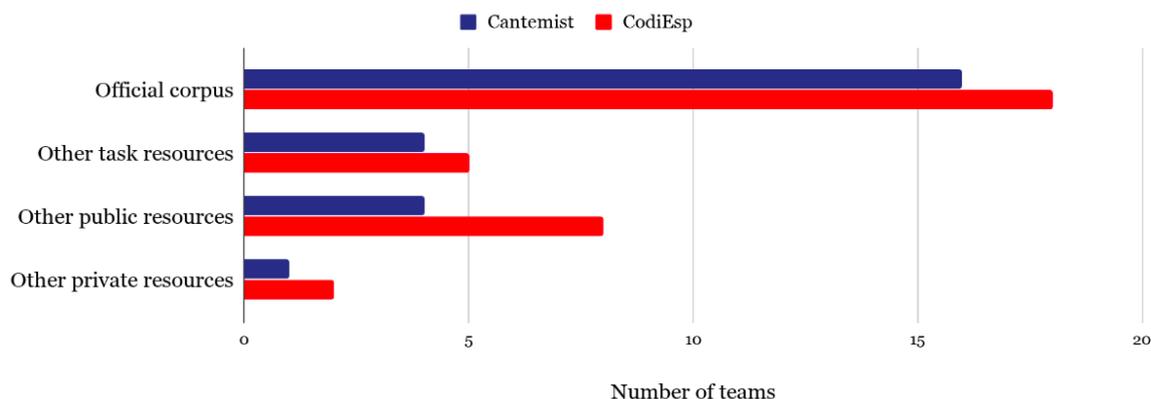


Figure 2: Comparison of the resources employed by participants between Cantemist and CodiEsp [25] shared tasks. Participants could find much more public resources for clinical coding than for cancer text mining in Spanish.

21, 25]. The selected type of text, clinical case reports, are fairly similar to hospital health records, and should facilitate a quick adaption of the systems to real-world scenarios.

To increase the usefulness and practical relevance of the Cantemist corpus, we selected clinical cases affecting all genders and that comprised most ages (from children to the elderly) and of various complexity levels (solid tumors, hemato-oncological malignancies, neuroendocrine cancer, etc.). We should emphasize that this corpus can be used toward generating further Gold Standard annotations, including temporal events and negations, or that focus on entities extracted from specific sections (for instance, imaging), of the individual clinical case records. The Cantemist cases include clinical signs and symptoms, personal and family history, current illness, physical examination, complementary tests (blood tests, imaging, pathology), diagnosis, treatment (including adverse effects of chemotherapy), evolution and outcome.

The lack of resources for oncology in Spanish was reflected in the training data employed by Cantemist participants. While in the past shared tasks we have organized participants make use of other public or private resources, this is rarely the case for Cantemist participants. See Figure 2 which compares the participants’ answers to the question “which datasets or corpus did you use?” in Cantemist and CodiEsp [25] shared tasks. The graph shows that, while CodiEsp participants found other public resources online that helped in building their systems, Cantemist participants did so less frequently.

Corpus annotation. The manual annotation of the Cantemist corpus was performed by clinical experts following the Cantemist guidelines. These guidelines contain rules for annotating morphology neoplasms in Spanish oncology clinical cases; as well as for mapping these annotations to eCIE-O.

The annotation guidelines were created by clinical experts in three phases:

1. Firstly, a preliminary initial version of the guidelines was created after the clinical experts revised neoplasm morphology annotations in the SPACCC corpus [18, 21, 25]. These original annotations in the SPACCC corpus followed the CodiEsp guidelines [25].
2. Later, a mature version of Cantemist guidelines was constructed after annotating iteratively sample sets of Cantemist corpus until the quality control (inter-annotator agreement) was satisfactory.
3. Finally, guidelines were further refined as manual annotation continued.

Similarly to what is required in the case of laboratory experiments, we should emphasize the need

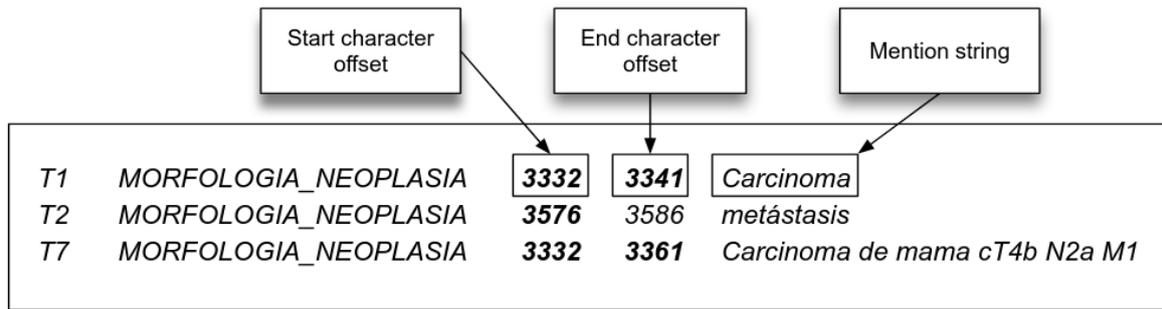


Figure 3: Example of annotation file for Cantemist-NER.

Table 2

Cantemist corpus summary.

	Documents	Annotations	Unique codes	Sentences	Tokens
Training	501	6396	493	25144	447903
Development	500	6001	520	23513	401994
Test	300	3633	386	14359	243604
Total	1301	16030	850	63016	1093501

for scientific rigor in the creation of NLP tools that assist in medical research. In our case, a medical doctor was regularly consulted by annotators (themselves scientists with PhDs on cancer-related subjects) for the most difficult pathology expressions. This same doctor periodically checked a random selection of annotated clinical records and these annotations were compared and discussed with the annotators. To normalize a selection of very complex cases, MD specialists in pathology from one of the largest university hospitals in Spain (Hospital Clínic, Barcelona) were consulted.

Corpus format. The Gold Standard Cantemist corpus is distributed in Brat standoff format [24]. Documents are released in plain text format with UTF-8 encoding. The annotations are included in a separate document (ANN file), with the same name as the plain text document name following the standards defined in Brat.

For Cantemist-NER, every line of the ANN file contains the mention string of the annotation, its start character offset, and its end character offset, which uniquely locate the mention in the text document. For Cantemist-Norm, the eCIE-O codes are included as comments in the annotations. See Figure 3 and 4 for examples of Cantemist-NER and Cantemist-Norm ANN files, respectively.

Finally, the Gold Standard Cantemist corpus was also distributed in the same format as previously used for the CodiEsp dataset [25]. In this case, documents are again distributed in plain text format. However, annotations are released in a tab-separated file. Every line of the tab-separated file contains the document name and an eCIE-O code. This format was employed in past clinical coding tasks, such as CodiEsp and the 2019 CLEF clinical coding shared task [26]. See Figure 5 for an example of the tab-separated file with the annotation information.

Corpus statistics. The Cantemist corpus contains 1,301 documents, with a total of 63,016 sentences and 1,093,501 tokens. All the 1,301 documents are annotated, i.e. tumor morphology mentions are found in them. There are 16,030 of such mentions, and each of them was manually mapped to an eCIE-O code. There are 850 unique codes. See Table 2 for a complete corpus overview.

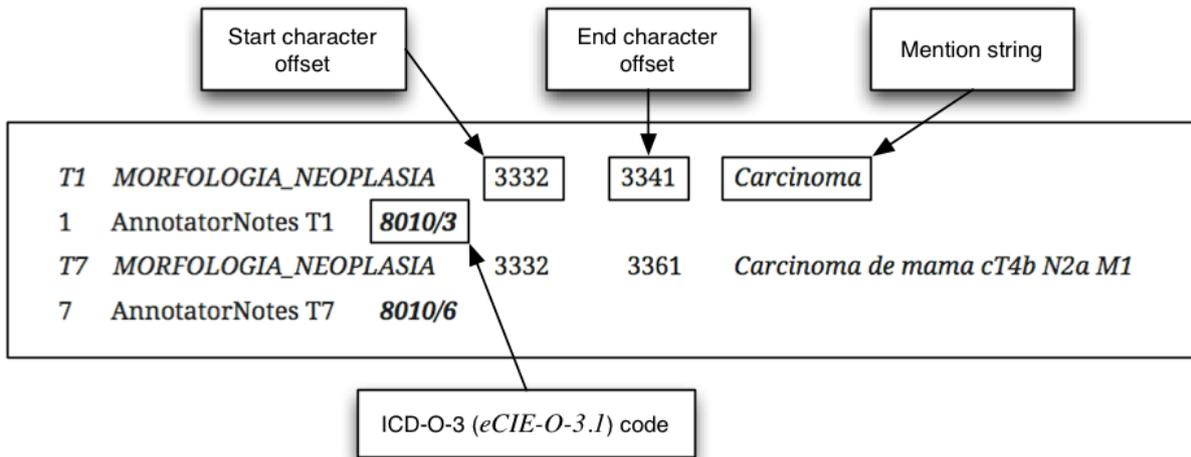


Figure 4: Example of annotation file for Cantemist-Norm.

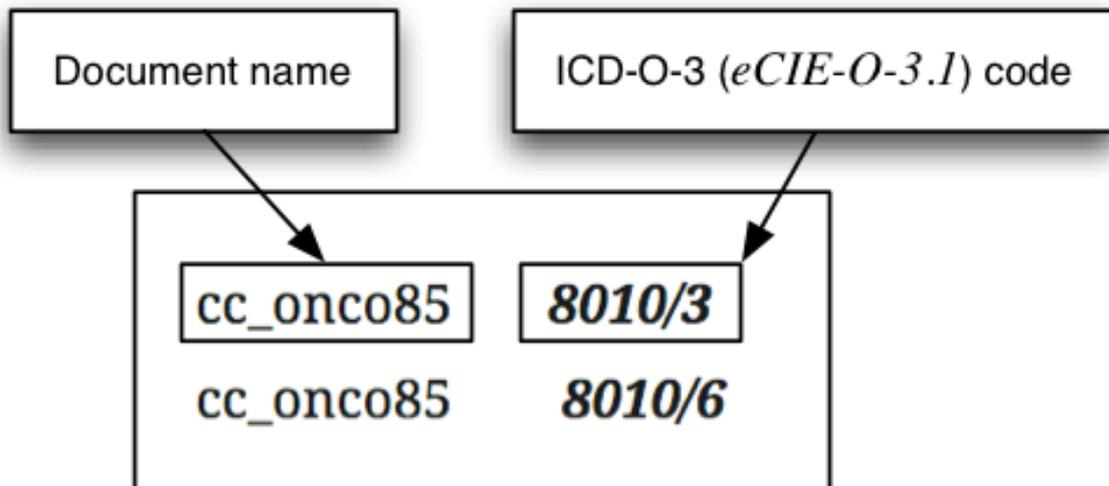


Figure 5: Example of annotation file for Cantemist-Coding.

3.2. Cantemist Silver Standard Corpus

The Cantemist test set was released together with an additional collection of 4,932 clinical case documents that belong to diverse medical disciplines. We call these 4,932 documents as the background set. Participants have generated automatic mention predictions for the test and also the background set, although they were only evaluated on the test set predictions. In that way, we examined whether systems were able to scale to larger data collections and prevented from manual annotation correction. Additionally, the code predictions for this background set constitute the Cantemist Silver Standard corpus, similar to the CALBC initiative [27] and to the CodiEsp initiative [25].

Table 3

Cantemist participation summary.

	Cantemist-NER	Cantemist-Norm	Cantemist-Coding	Total
Participant teams	23	10	9	25
Submitted runs	62	30	29	131

4. Results

Cantemist contained three independent subtasks: Cantemist-NER, Cantemist-Norm, and Cantemist-Coding. Participants could choose whether to submit results for one, two, or all subtasks. Participants could submit up to 5 runs for each subtask.

4.1. Participation overview

Cantemist has received considerable attention from the community. Indeed, 66 teams registered for this task, and 25 of them submitted predictions (one did not qualify due to a format error). From the 25 teams, 23 participated in the Cantemist-NER subtask, 10 in the Cantemist-Norm subtask, and 9 in the Cantemist-Coding subtask. Since five runs were allowed per subtask, the total number of systems participating in the shared task is considerably higher. Indeed, we received 62 prediction runs for Cantemist-NER, 30 for Cantemist-Norm, and 29 for Cantemist-Coding. In total, Cantemist shared task lead to the creation of 131 systems.

We should emphasize that 20% of the teams come from industry rather than academia. Additionally, as Table 4 shows, participants belonged to institutions (industry or academia) from a number of different countries including Spain, China, India, USA or Argentina.

4.2. System results

Table 5 shows the results of the best run obtained by each team. The top-scoring results for each subtask were:

- *Cantemist-NER*. HITSZ-ICRC team has obtained the highest f1-score, 0.87. Their system is highly balanced: the precision has been 0.871 and the recall 0.868. It is almost equivalent to the f1-score obtained by the Vicomtech team, 0.869.
- *Cantemist-Norm*. Again, the HITSZ-ICRC team has obtained the highest f1-score, 0.825. Their precision has been 0.824 and their recall 0.826. And again, Vicomtech has obtained a really close f1-score, 0.821.
- *Cantemist-Coding*. In this subtask, both Vicomtech and ICB-UMA have obtained equivalent scores for the main metric (MAP), 0.847. The Vicomtech team has developed a system with balanced precision (0.875), recall (0.836), and f1-score (0.855). Differently, the ICB-UMA team has maximized the MAP metric, since their system has a really low precision (0.007) and really high recall (0.928).

Table 4

Cantemist team overview. A/I stands for academic or industry institution. In the Tasks column, NE stands for Cantemist-NER, No for Cantemist-Norm and C for Cantemist-Coding.

Team Name	Affiliation	A/I	Tasks	Ref.	Tool URL
HITSZ-ICRC	Harbin Institute of Technology, China	A	NE,No	[28]	-
Vicomtech	Vicomtech Foundation, Spain	I	NE,No,C	[29]	-
SINAI	University of Jaén, Spain	A	NE,No	[30]	-
NLNDE	SSN College of engineering, India	A	NE,No,C	[31]	-
NCU-IISR	National Central University, Taiwan	A	NE	[32]	-
Recognai	Recognai, Spain	I	NE	[33]	-
mhjabreel	Hodeidah University, Yemen	A	NE,No,C	[34]	-
HULAT-UC3M	University Carlos III, Spain	A	NE	[35]	[36]
Fadi	Universitat Rovira i Virgili, Spain	A	NE,No,C	[37]	-
rrz-uc3m	University Carlos III, Spain	A	NE,No	[38]	-
baciero-fdez	-	-	NE	-	-
HULATUC3M-GI	University Carlos III, Spain	A	NE	[39]	-
IBS_Software	IBS Software Pvt. Ltd., India	I	NE	[40]	-
lasigeBioTM	Universidade de Lisboa, Portugal	A	NE,No,C	[41]	[42]
Tong Wang	Yunnan University, P.R.China	A	NE	[43]	[44]
DTIMAI	Siemens Healthineers, USA	I	NE	[45]	-
episource	Episource LLC, USA	I	NE,No,C	-	-
XIntao	Yunnan University, P.R. China	A	NE	[46]	-
UAB	Univ. of Alabama at Birmingham, USA	A	NE	[47]	-
Bigbyte	BigByteMX, Mexico	-	NE,No,C	-	-
PaccanaroLab	University of London, UK	A	NE	-	-
fernandez	Argentina	-	NE	-	-
ICB-UMA	University of Málaga, Spain	A	C	[48]	[49]
kathrync	DFKI, Germany	A	C	[50]	[51]

4.3. Error analysis

Missed annotations are more complex. The annotations that automatic systems fail to predict seem to be the more complex. We have extracted the annotations systematically missed by the top 5 participants, according to the f1-score: HITSZ-ICRC, Vicomtech, SINAI, NLNDE, and NCU-IISR. We have compared these annotations with the complete set of test annotations.

The missed annotations are longer. The median number of characters is 26 for the difficult annotations, while the median number of characters for all test set annotations is 14. In addition, there is a higher percentage of abbreviations from the Spanish Medical Abbreviation DataBase [52] in the missed annotations.

Missed codes are more specific and less frequent. The codes that automatic systems do not assign properly seem to be more specific. In the subset of missed annotations, 8% of the codes contain an “H”. This percentage is as low as 2% in the entire test set. Additionally, 13.2% of the missed annotations include the sixth differentiation digit in their code (the sixth digit in eCIE-O indicates the tumor differentiation). In contrast, this percentage is 5.6% in the entire test set. Besides, missed test codes are less frequent in the training and development sets. The median of appearances of the missed codes in the training and development set is 1, whereas for the test set codes is 3. Finally, 20.8% of the missed annotations have the metastasis code (8000/6), while this code accounts for 34.6% of the complete test set.

Table 5

Cantemist results. Best result per team. Best result bolded, second best underlined.

Team Name	NER			Norm			Coding			MAP
	P	R	F1	P	R	F1	P	R	F1	
HITSZ-ICRC	.871	<u>.868</u>	.87	.824	.826	.825	-	-	-	-
Vicomtech	<u>0.868</u>	0.871	<u>0.869</u>	<u>.822</u>	<u>.821</u>	<u>.821</u>	.875	.836	.855	.847
SINAI	.859	.851	.855	.763	.755	.759	-	-	-	-
NLNDE	.854	.852	.853	.767	.766	.767	.77	.771	.77	.749
NCU-IISR	.849	.851	.85	-	-	-	-	-	-	-
Recognai	.85	.84	.845	-	-	-	-	-	-	-
mhjabreel	.837	.84	.839	.775	.779	.777	.797	.812	.805	.737
HULAT-UC3M	.826	.843	.834	-	-	-	-	-	-	-
Fadi	.844	.818	.831	.798	.774	.786	<u>.826</u>	<u>.838</u>	<u>.832</u>	<u>.797</u>
rrz-uc3m	.823	.824	.823	.202	.14	.165	-	-	-	-
baciero-fdez	.808	.802	.805	-	-	-	-	-	-	-
HULATUC3M-GI	.828	.769	.797	-	-	-	-	-	-	-
IBS_Software	.765	.764	.764	-	-	-	-	-	-	-
lasigeBioTM	.787	.714	.749	.064	.058	.061	.211	.601	.312	.506
Tong Wang	.757	.736	.746	-	-	-	-	-	-	-
DTIMAI	.727	.741	.734	-	-	-	-	-	-	-
episource	.691	.758	.723	.557	.61	.582	.68	.681	.681	.575
XIntao	.716	.721	.719	-	-	-	-	-	-	-
UAB	.688	.744	.715	-	-	-	-	-	-	-
Bigbyte	.649	.469	.545	.645	.467	.542	.794	.73	.761	.68
PaccanaroLab	.159	.595	.251	-	-	-	-	-	-	-
fernandez	0	0	0	-	-	-	-	-	-	-
ICB-UMA	-	-	-	-	-	-	.007	.928	.013	.847
kathrync	-	-	-	-	-	-	.182	.51	.268	.394

4.4. Methodologies

Most successful teams employ deep learning for all subtasks (for NER, normalization, and clinical coding). Indeed, there are two types of architectures that most participant repeat:

- A transformer-based language model and a classification layer. For example, the DTIMAI [45] and Vicomtech [29] teams.
- A Bidirectional Recurrent Neural Network (in general, with LSTM memory units) and a Conditional Random Field (CRF). For example, the HULAT-UC3M [35] and Sinai [30] teams.

The preeminence of deep learning against other approaches is particularly seen in the performance drop of teams that employed it in Cantemist-NER but chose other approaches in Cantemist-Norm (Figure 6, Table 5).

From the teams using deep learning, Recurrent Neural Networks, transformers (which are the core of the latest language models) and CRFs are the most utilized technologies. And within the language models, participants have chosen to use, mostly, BERT [53], BETO [54], XLM-RoBERTa [55] and one team included in their solution their own language model [39], as it is shown in Figure 6.

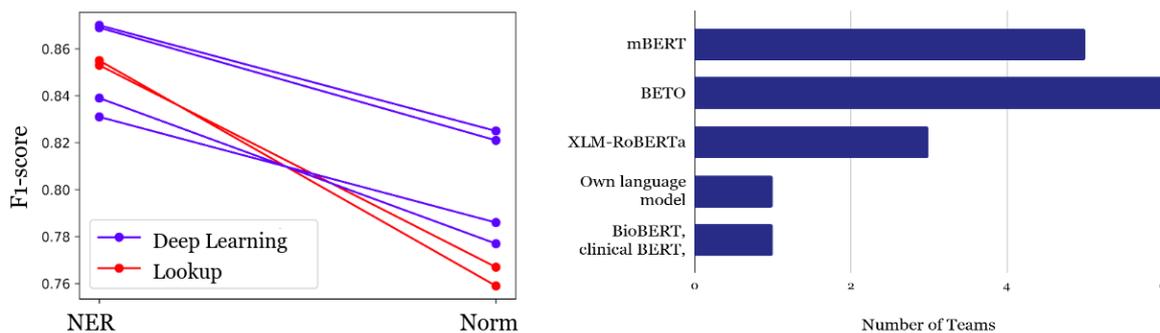


Figure 6: At the left, performance drop between Cantemist-NER and Cantemist-Norm colored by the method employed for normalizing the entities. At the right, the different types of language models employed by Cantemist participants.

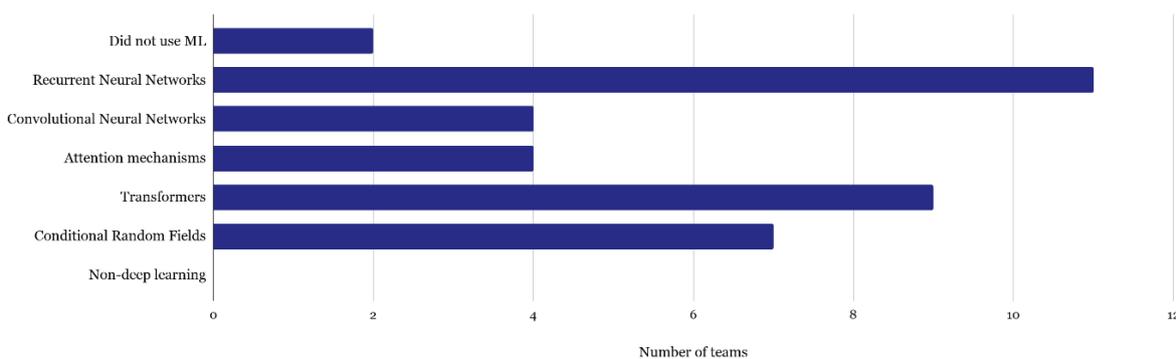


Figure 7: Machine Learning (ML) techniques employed by Cantemist participants.

5. Discussion

To the authors’ knowledge, Cantemist is the first NER shared task in Spanish focusing on cancer, specifically on tumor morphology. Notably, we use oncology clinical case reports, since their language and format are closer to clinical narrative texts from EHRs, but with no issues due to patient data privacy. Finally, in addition to NER, we introduce the normalization of tumor morphology concept mentions to a standard normative terminology, eCIE-O (the Spanish equivalent of the International Classification of Diseases for Oncology), which is globally used by pathologists for classification and statistical reporting purposes.

The Cantemist corpus is the first collection of clinical case reports in Spanish with annotated tumor morphology mapped to eCIE-O. We believe that the Cantemist Silver Standard may serve to further extend this corpus. Both corpora are publicly available at the Zenodo Medical NLP community⁴. We plan to also release a Cantemist post-workshop Gold Standard collection to further promote research and development of new tools beyond the shared task period.

Medical language is complex and varies considerably among different medical disciplines. Available corpora specific to a particular medical speciality or domain are essential for fine-tuning novel language models. We strongly encourage the community to replicate the Cantemist initiative for

⁴<https://zenodo.org/communities/medicalnlp/>

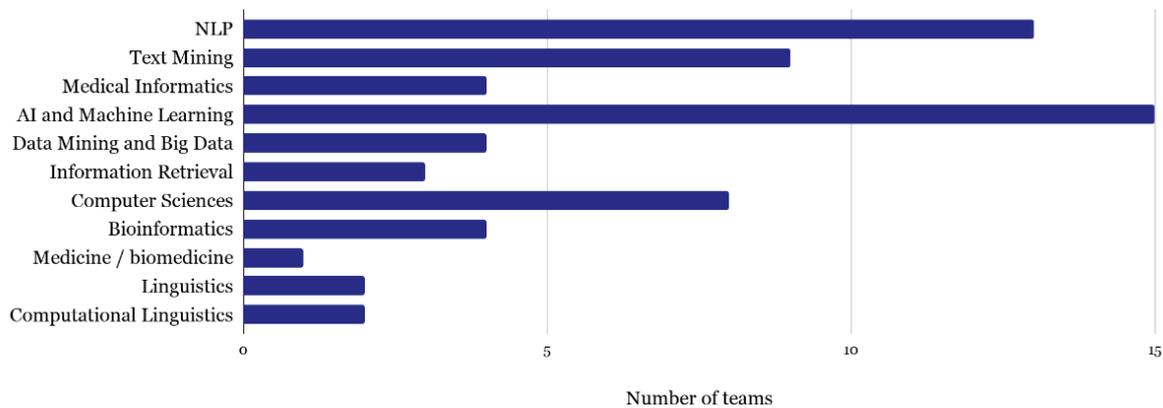


Figure 8: Background of expertise of Cantemist participants.

other medical specialties, such as cardiology or radiology. Notably, the current Cantemist corpus is limited to 1,301 clinical case reports with 16,030 annotations. As discussed in the Results section, codes that automatic systems fail to predict are usually the less frequent. Larger corpora will result in better NLP tools, and we are currently adding 600 case reports to our corpus. Finally, published clinical cases share many similarities with hospital records. However, we expect to be able to access real health records in the immediate future, since they would produce much better suited tools for application in real-world oncology reports.

Cantemist, a shared task on Spanish NLP, is included in the conference of the Spanish Society of Natural Language Processing (SEPLN). The clinical texts and the terminology are written in Spanish. Interestingly, teams from all around the world registered for this task, with 22 teams from Spain (22), followed in number of participant teams by China (9), India (8), and the USA (5). Other represented countries were Russia, Germany, Belarus and Taiwan.

Of note, participants originate from various locations and heterogeneous backgrounds (Figure 8). When asked about their team members' expertise, a high number of participants worked in artificial intelligence, machine learning and data mining. Others self-identified with NLP, information retrieval, and text mining fields of knowledge. Also, a small proportion of participants selected medicine or biomedicine as their background. There are also answers pointing to medical informatics or bioinformatics. Finally, 21.2% of registered teams work in commercial organizations, while the remaining teams work in academic institutions.

Named Entity Recognition is an established NLP task. Indeed, 90.9% of the participant teams that answered the question reported having worked on another NER task before (Figure 9). Moreover, the same percentage reported needing less than four weeks to complete the task (Figure 9). Technologies such as Recurrent Neural Networks and, more recently, transformer-based language models have taken over NER tasks and have become state of the art, which might explain the remarkable equality among top-scoring teams of Cantemist-NER: there are five teams with f1-scores between 0.85 and 0.871.

In a situation such as the one described in the previous task (a mature NLP task with an established state-of-the-art technology), we expect the emergence of stakeholders interested in transferring the knowledge from academia to industry. Indeed, more than 20% of Cantemist registered participants are from the industry (Figure 10). Additionally, most respondents to the participants' survey think that commercial firms and healthcare professionals could benefit from these tasks (Figure 11), and

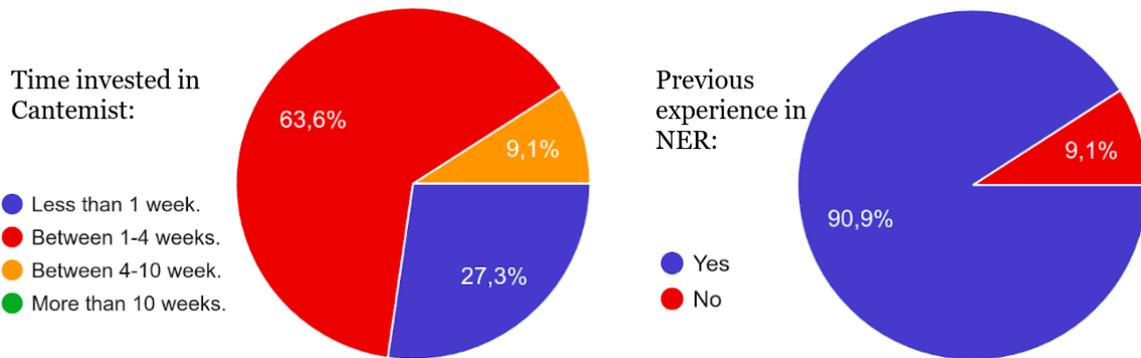


Figure 9: Time invested and previous experience of Cantemist participants.

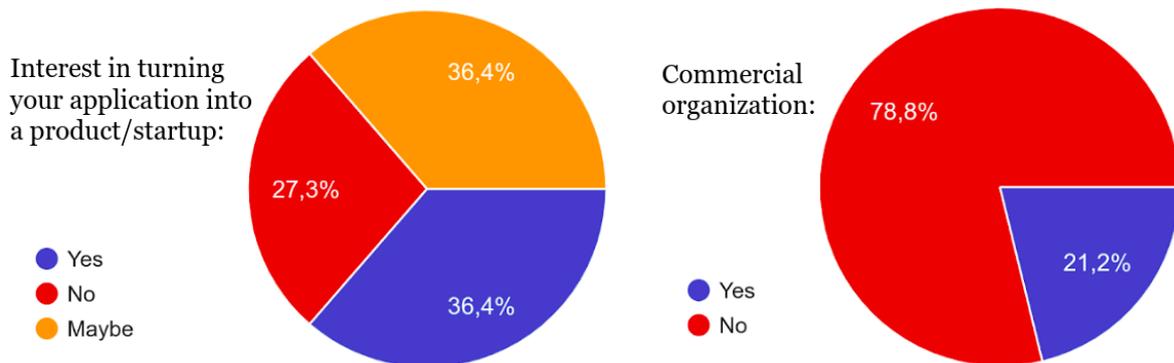


Figure 10: Interest in turning the developed application into a software product or startup and commercial origin of their current affiliation.

only 27% of them answered “no” to the question, “Would you be interested in support for turning your system into a software product/startup?” (Figure 10). Named Entity Recognition is a building block for downstream NLP tasks, and the resources originating from the Cantemist shared task have the potential to impact text mining applications for oncology.

The Cantemist shared task has been a participation success, considering the amount of participants and the interest of international teams in a Spanish text mining task for a Spanish conference. In addition, the Cantemist corpus is a pioneer work on the distribution of domain-specific medical NLP corpus, and in languages other than English. We would like to encourage similar initiatives in other medical specialties and other languages, including Basque, Catalan and Galician, which are of interest for the Spanish National Plan for the Advancement of Language Technology [56]. Finally, the developed systems may be ready to be implemented since the results, specially for Cantemist-Norm and Cantemist-Coding, are remarkable.

We propose that future evaluation efforts for oncology text mining in Spanish should also take into account the annotation of clinically important information such as negation or temporal expressions, as well as examining scalability and robustness related aspects [57] or the integration of the generated solution into interactive systems with experts clinicians in the loop, similar to what was done for biomedical text mining evaluation scenarios [58].

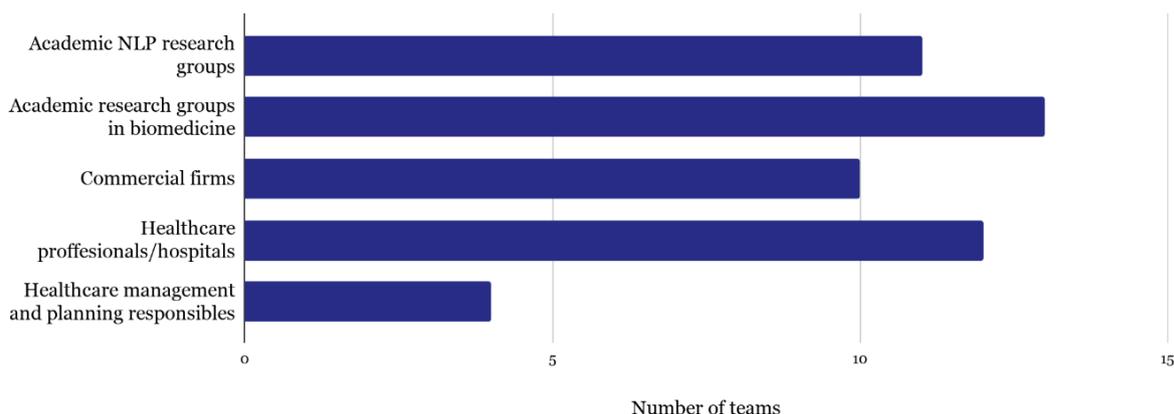


Figure 11: Potential Cantemist results beneficiaries, according to the participants.

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. Besides, we would also like to thank the organization of IberLEF. Finally, we would particularly like to thank the team of Bitac, Gloria González, and Toni Mas, who worked with us to create the dataset and are still working to make it grow. We do acknowledge positive support from Jose Antonio Lopez-Martin (Hospital 12 de Octubre) and the Sociedad Española de Oncología Médica (SEOM).

References

- [1] W. H. Organization, Cancer, https://www.who.int/health-topics/cancer#tab=tab_1, 2020. Accessed: 2020-08-31.
- [2] I. Spasić, J. Livsey, J. A. Keane, G. Nenadić, Text mining of cancer-related information: review of current status and future directions, *International journal of medical informatics* 83 (2014) 605–623.
- [3] N. C. Institute, How cancer is diagnosed - national cancer institute, <https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis#biopsy>, 2019. Accessed: 2020-08-31.
- [4] F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak, B. Shen, Biomedical text mining and its applications in cancer research, *Journal of biomedical informatics* 46 (2013) 200–211.
- [5] D. Dligach, S. Bethard, L. Becker, T. Miller, G. K. Savova, Discovering body site and severity modifiers in clinical texts, *Journal of the American Medical Informatics Association* 21 (2014) 448–454.
- [6] S. Schulz, P. Daumke, M. Romacker, P. López-García, Representing oncology in datasets: Standard or custom biomedical terminology?, *Informatics in Medicine Unlocked* 15 (2019) 100186.
- [7] C. D. Bajdik, B. Kuo, S. Rusaw, S. Jones, A. Brooks-Wilson, Cgmim: automated text-mining of online mendelian inheritance in man (omim) to identify genetically-associated cancers and candidate genes, *BMC bioinformatics* 6 (2005) 78.
- [8] B. Xie, Q. Ding, H. Han, D. Wu, mircancer: a microRNA–cancer association database constructed by text mining on literature, *Bioinformatics* 29 (2013) 638–644.

- [9] Y.-C. Fang, P.-T. Lai, H.-J. Dai, W.-L. Hsu, MeinfoText 2.0: gene methylation and cancer relation extraction from biomedical literature, *BMC bioinformatics* 12 (2011) 471.
- [10] J. A. Strauss, C. R. Chao, M. L. Kwan, S. A. Ahmed, J. E. Schottinger, V. P. Quinn, Identifying primary and recurrent cancers using a sas-based natural language processing algorithm, *Journal of the American Medical Informatics Association* 20 (2013) 349–355.
- [11] M. A. Tanenblatt, A. Coden, I. L. Sominsky, The conceptmapper approach to named entity recognition., in: *LREC, Citeseer*, 2010, pp. 546–51.
- [12] S. S. Seda, F. d. P. P. León, J. M. Conde, M. C. G. Ruiz, J. M. Sánchez, G. Rodríguez, J. A. P. Simón, C. L. P. Calderón, Plataforma para la extracción automática y codificación de conceptos dentro del ámbito de la oncohematología (proyecto coco), *Procesamiento del Lenguaje Natural* 61 (2018) 65–71.
- [13] R. Kavuluru, I. Hands, E. B. Durbin, L. Witt, Automatic extraction of icd-o-3 primary sites from cancer pathology reports, *AMIA Summits on Translational Science Proceedings 2013* (2013) 112.
- [14] Y. Jin, R. T. McDonald, K. Lerman, M. A. Mandel, S. Carroll, M. Y. Liberman, F. C. Pereira, R. S. Winters, P. S. White, Automated recognition of malignancy mentions in biomedical literature, *BMC bioinformatics* 7 (2006) 492.
- [15] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, P. C. De Groen, Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model, *Journal of biomedical informatics* 42 (2009) 937–949.
- [16] A. Fritz, C. Percy, A. Jack, K. Shan, Clasificación internacional de enfermedades para oncología (cie-o), *Revista Española de Salud Pública* 77 (2003) 659–659.
- [17] W. H. Organization, International classification of diseases for oncology (ICD-O) – 3rd edition, 1st revision, 3rd ed. ed., 2013.
- [18] A. G. Agirre, M. Marimon, A. Intxaurreondo, 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.
- [19] M. Marimon, A. Gonzalez-Agirre, A. Intxaurreondo, 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.
- [20] A. Intxaurreondo, M. Pérez-Pérez, G. Pérez-Rodríguez, J. A. López-Martín, J. Santamaria, S. de la Pena, M. Villegas, S. A. Akhondi, A. Valencia, A. Lourenço, et al., The biomedical abbreviation recognition and resolution (barr) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to spanish biomedical abstracts (2017).
- [21] A. Intxaurreondo, M. Marimon, A. Gonzalez-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., in: *IberEval@ SEPLN*, 2018, pp. 280–289.
- [22] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, S. Ananiadou, Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013, *BMC bioinformatics* 16 (2015) S2.
- [23] C. D. Manning, H. Schütze, P. Raghavan, Introduction to information retrieval, Cambridge university press, 2008.
- [24] 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.
- [25] 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: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, 2020.
- [26] M. L. Neves, D. Butzke, A. Dörendahl, N. Leich, B. Hummel, G. Schönfelder, B. Grune, Overview of the clef ehealth 2019 multilingual information extraction., in: CLEF (Working Notes), 2019.
- [27] 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.
- [28] Y. Xiong, Y. Huang, Q. Chen, X. Wang, B. Tang, A joint model for medical named entity recognition and normalization, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [29] A. García-Pablos, N. Perez, M. Cuadros, Vicomtech at cantemist 2020, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [30] P. López-Úbeda, M. C. Díaz-Galiano, M. T. Martín-Valdivia, L. A. Urena-López, Extracting neoplasms morphology mentions in spanish clinical cases from word embeddings, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [31] L. Lange, X. Dai, H. Adel, J. Strotgen, Nlnde: Hybrid approaches for the extraction and normalization of tumor mentions from spanish clinical texts, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [32] J.-C. Han, R. T.-H. Tsai, Ncu-iisr: Pre-trained language model for cantemist named entity recognition, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [33] D. C. Fidalgo, D. Vila-Suero, F. A. Montes, Recognai’s working notes for cantemist-ner track, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [34] M. Jabreel, End-to-end neural coder for tumor named entity recognition, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [35] S. S. Carrasco, P. Martínez, Using embeddings and bi-lstm+crf model to detect tumor morphology entities in spanish clinical cases, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [36] Hulat-UC3M, Cantemist-participation, <https://github.com/ssantamaria94/CANTEMIST-Participation>, 2020.
- [37] F. Hassan, D. Sánchez, J. Domingo-Ferrer, Tumor entity recognition and coding for spanish electronic health records, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [38] R. Rivera-Zavala, P. Martinez, Deep neural model with contextualized-word embeddings for named entity recognition in spanish clinical text, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [39] G. de Vargas Romero, I. Segura-Bedmar, Exploring deep learning for named entity recognition of tumor morphology mentions, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [40] U. K. Sikdar, B. Gamback, M. K. Kumar, Tumor morphology mentions identification using deep learning and conditional random fields, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [41] P. Ruas, A. Neves, V. D. Andrade, F. M. Couto, Lasigebiotm at cantemist: Named entity recognition and normalization of tumour morphology entities and clinical coding of spanish health-related documents, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

- [42] LasigeBioTM, Cantemist-participation, <https://github.com/lasigeBioTM/CANTEMIST-Participation>, 2020.
- [43] T. Wang, Y. Zhang, Y. Li, A parallel-attention model for tumor named entity recognition in spanish, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [44] T. Wang, Cantemist, <https://github.com/18720936539/CANTEMIST/>, 2020.
- [45] R. Vunikili, S. H. N, G. V. Marica, O. Farri, Clinical ner using spanish bert embeddings, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [46] X. Tao, R. Liu, X. Zhou, A tumor named entity recognition model based on language model and multi-head attention, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [47] T. O’Leary, J. D. Monte, K. Sasse, W. H. Liang, J. D. Osborne, Identification of cancer entities in clinical text combining transformers with dictionary features, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [48] G. López-García, J. M. Jerez, F. J. Veredas, Icb-uma at cantemist 2020: Automatic icd-o coding in spanish with bert, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [49] ICB-UMA, Cantemist-2020, <https://github.com/guilopgar/CANTEMIST-2020>, 2020.
- [50] K. Chapman, G. Neumann, Cantemist 2020 coding task: Automatic icd code classification with label description attention mechanism, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.
- [51] Kathryn, Cantemist2020, <https://github.com/kathrynchapman/CANTEMIST2020>, 2020.
- [52] A. Intxaurrenondo, Abremes-db, 2018. URL: <https://doi.org/10.5281/zenodo.2207130>. doi:10.5281/zenodo.2207130, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
- [53] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
- [54] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: to appear in PML4DC at ICLR 2020, 2020.
- [55] 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, arXiv preprint arXiv:1911.02116 (2019).
- [56] M. Villegas, S. de la Peña, A. Intxaurrenondo, J. Santamaria, M. Krallinger, Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje, Procesamiento del Lenguaje Natural (2017) 141–144.
- [57] M. Pérez-Pérez, G. Pérez-Rodríguez, A. Blanco-Míguez, F. Fdez-Riverola, A. Valencia, M. Krallinger, A. Lourenço, Next generation community assessment of biomedical entity recognition web servers: metrics, performance, interoperability aspects of becalm, Journal of cheminformatics 11 (2019) 42.
- [58] C. N. Arighi, P. M. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, A. Chatr-Aryamontri, S. Clematide, P. Gaudet, M. G. Giglio, I. Harrow, et al., Biocreative iii interactive task: an overview, BMC bioinformatics 12 (2011) S4.

A. Appendix

Team Name	Run Name	P	R	F1
HITSZ-ICRC	1-RM_JOINT_NER_MERGED	.86	.871	.865
HITSZ-ICRC	2-JOINT_NER_MERGED	.866	.87	.868
HITSZ-ICRC	3-RM_SINGLE_NER_MERGED	.871	.868	.87
HITSZ-ICRC	4-JOINT_NER_LOSS_weight	.871	.868	.87
HITSZ-ICRC	5-JOINT_SPAN_NER	.866	.854	.86
Vicomtech	1-BETO	.862	.866	.864
Vicomtech	2-SciBERT	.854	.866	.86
Vicomtech	3-TwoExperts	.862	.861	.862
Vicomtech	4-OneRoundEnsemble	.869	.865	.867
Vicomtech	5-TwoRoundEnsemble	.868	.871	.869
SINAI	1-systemGlove	.809	.817	.813
SINAI	2-systemSME	.835	.855	.845
SINAI	3-systemFlair	.859	.828	.843
SINAI	4-systemSME+Flair	.859	.851	.855
SINAI	5-systemSME+Flair+Glove	.858	.847	.852
NLNDE	1-systemBiLSTM	.824	.83	.827
NLNDE	2-systemMeta	.815	.823	.819
NLNDE	3-systemBiaffine	.85	.835	.842
NLNDE	4-systemBiaffineDev	.854	.852	.853
NLNDE	5-systemEnsemble	.847	.808	.827
NCU-IISR	1-systemBETO	.849	.851	.85
NCU-IISR	2-systemBETO+MedlinePlus-TEI	.84	.85	.845
Recognai	1-xlmr-char-lstm	.85	.84	.845
Recognai	2-fasttext-char-lstm	.846	.844	.845
mhjabreel	1-run	.837	.84	.839
HULAT-UC3M	1-bilstm5	.826	.843	.834
HULAT-UC3M	2-bilstm7	.824	.832	.828
HULAT-UC3M	3-bilstm9	.843	.824	.833
HULAT-UC3M	4-bilstm15	.832	.817	.824
Fadi	1-BILSTM	.807	.83	.818
Fadi	2-BILSTM	.824	.824	.824
Fadi	CRF	.806	.776	.791
Fadi	CRF_BILSTM	.844	.818	.831
rrz-uc3m	1-BiLSTM_CRF	.823	.824	.823
baciero-fdez	1-Cantemist_baciero	.808	.802	.805
HULATUC3M-GI	1-systemCRF	.8	.768	.783
HULATUC3M-GI	2-systemBILSTM1	.771	.773	.772
HULATUC3M-GI	3-systemBILSTM2	.828	.769	.797
HULATUC3M-GI	4-systemBILSTM3	.784	.759	.771
HULATUC3M-GI	5-systemBERT	.756	.775	.765
IBS_Software	Run-1	.758	.746	.752
IBS_Software	Run-3	.756	.747	.751

IBS_Software	Run-2	.746	.745	.746
IBS_Software	Run-4	.697	.751	.723
IBS_Software	Run-5	.35	.175	.233
IBS_Software	Run-5-unofficial	.765	.764	.764
lasigeBioTM	1-run	.787	.714	.749
Tong Wang	1-systemDL	.737	.707	.722
Tong Wang	2-systemDL	.757	.736	.746
DTIMAI	1-systemBERT1	.727	.741	.734
DTIMAI	2-systemBERT2_Casing	.727	.741	.734
episource	process_1	.691	.758	.723
XIntao	1-run	.716	.721	.719
UAB	1-systemCasedBETO	.736	.609	.667
UAB	2-systemRegex	.688	.744	.715
UAB	3-systemMBERT	.673	.357	.467
BigByte	1-run	.649	.469	.545
PaccanaroLab	1-nn_merge	.159	.595	.251
fernandez	1-bilstm_merged	0	0	0
fernandez	2-bilstm_0	0	0	0
fernandez	3-bilstm_dropout	0	0	0
fernandez	4-bilstm_0_b	0	0	0

Table 6: Cantemist-NER results.

Team Name	Run Name	P	R	F1	No metastasis		
					P	R	F1
HITSZ-ICRC	1-JOINT_NORM_MERGED_new	.824	.826	.825	.848	.803	.825
HITSZ-ICRC	2-RM_JOINT_NORM_MERGED.	.803	.811	.807	.824	.78	.801
HITSZ-ICRC	3-JOINT_NORM_LOSS_weight.	.82	.808	.814	.828	.78	.803
HITSZ-ICRC	4-PIPELINE-SGM	.794	.791	.792	.799	.765	.782
HITSZ-ICRC	5-JOINT_SPAN_NORM_new	.677	.667	.672	.675	.603	.637
Vicomtech	1-BETO	.807	.808	.807	.806	.78	.792
Vicomtech	2-SciBERT	.801	.811	.806	.803	.782	.793
Vicomtech	3-TwoExperts	.798	.795	.797	.788	.759	.773
Vicomtech	4-OneRoundEnsemble	.822	.819	.82	.825	.79	.807
Vicomtech	5-TwoRoundEnsemble	.822	.821	.821	.823	.792	.807
SINAI	1-systemGlove	.728	.735	.732	.73	.707	.718
SINAI	2-systemSME	.747	.766	.756	.748	.732	.74
SINAI	3-systemFlair	.769	.741	.755	.751	.73	.74
SINAI	4-systemSME+Flair	.763	.755	.759	.749	.732	.74
SINAI	5-systemSME+Flair+Glove	.764	.754	.759	.753	.735	.744
NLNDE	1-systemBiLSTM	.743	.749	.746	.75	.709	.729
NLNDE	2-systemMeta	.735	.741	.738	.746	.709	.727
NLNDE	3-systemBiaffine	.767	.753	.76	.764	.714	.738
NLNDE	4-systemBiaffineDev	.767	.766	.767	.773	.726	.749
NLNDE	5-systemEnsemble	.767	.732	.749	.774	.702	.736
mhjabreel	1-run	.775	.779	.777	.782	.747	.764

Fadi	1-BILSTM	.765	.786	.776	.786	.769	.777
Fadi	2-BILSTM	.779	.78	.779	.787	.769	.778
Fadi	CRF	.775	.746	.76	.786	.735	.76
Fadi	CRF_BILSTM	.798	.774	.786	.801	.773	.787
rrz-uc3m	1-BiLSTM_CRF	.202	.14	.165	.202	.21	.206
lasigeBioTM	1-single_ont	.063	.057	.06	.059	.082	.069
lasigeBioTM	2-multi_ont	.064	.058	.061	.059	.08	.068
episource	process_1	.557	.61	.582	.504	.527	.515
Bigbyte	1-run	.645	.467	.542	.659	.436	.525

Table 7: Cantemist-Norm results.

Team Name	Run Name	MAP	P	R	F1	No metastasis			
						MAP	P	R	F1
Vicomtech	1-BETO	.829	.86	.824	.841	.807	.843	.792	.817
Vicomtech	2-SciBERT	.838	.858	.832	.845	.816	.843	.802	.822
Vicomtech	3-TwoExperts	.815	.85	.799	.824	.793	.83	.764	.796
Vicomtech	4-OneRoundEns.	.842	.875	.832	.853	.817	.862	.802	.831
Vicomtech	5-TwoRoundEns.	.847	.875	.836	.855	.822	.862	.807	.834
NLNDE	1-systemBiLSTM	.737	.755	.762	.759	.697	.727	.721	.724
NLNDE	2-systemMeta	.735	.748	.758	.753	.694	.719	.716	.718
NLNDE	3-systemBiaffine	.739	.759	.763	.761	.702	.73	.722	.726
NLNDE	4-systemBiaffineD.	.749	.77	.771	.77	.714	.743	.728	.736
NLNDE	5-systemEnsemble	.731	.772	.749	.76	.693	.746	.707	.726
mhjabreel	coding_processed	.737	.797	.812	.805	.721	.776	.78	.778
Fadi	1-BILSTM	.783	.813	.841	.827	.769	.795	.813	.804
Fadi	2-BILSTM	.797	.826	.838	.832	.785	.806	.813	.809
Fadi	CRF	.779	.834	.809	.821	.767	.818	.777	.797
Fadi	CRF_BILSTM	.787	.841	.826	.833	.773	.82	.799	.809
lasigeBioTM	1-X-Transformer.	.455	.151	.532	.235	.344	.113	.445	.18
lasigeBioTM	2-X-Transformer.	.449	.159	.517	.243	.333	.118	.427	.184
lasigeBioTM	3-X-Transformer.	.459	.197	.541	.289	.346	.151	.456	.226
lasigeBioTM	4-X-Transformer.	.463	.157	.549	.244	.35	.119	.466	.189
lasigeBioTM	5-X-Transformer.	.506	.211	.601	.312	.399	.167	.527	.254
episource	coding	.575	.68	.681	.681	.503	.637	.627	.632
Bigbyte	coding	.68	.794	.73	.761	.652	.771	.684	.725
ICB-UMA	1-Multi	.821	.007	.928	.013	.794	.006	.914	.011
ICB-UMA	2-Multi-Onco	.847	.007	.928	.013	.821	.006	.914	.011
ICB-UMA	3-Multi-Onco-M.	.837	.007	.928	.013	.813	.006	.914	.011
ICB-UMA	4-Scielo	.8	.007	.928	.013	.769	.006	.914	.011
ICB-UMA	5-Scielo-Onco	.812	.007	.928	.013	.784	.006	.914	.011
Kathrync	1-Cantemist-coding	.394	.182	.51	.268	.254	.135	.419	.205
Kathrync	2-Cantemist-coding	.381	.197	.472	.279	.237	.143	.375	.207

Table 8: Cantemist-Coding results.