# PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track

**Aitor Gonzalez-Agirre**
Centro Nacional de Investigaciones Oncológicas (CNIO)
Barcelona Supercomputing Center (BSC)
aitor.gonzalez@bsc.es

**Montserrat Marimon**
Centro Nacional de Investigaciones Oncológicas (CNIO)
Barcelona Supercomputing Center (BSC)
montserrat.marimon@bsc.es

**Ander Intxaurrondo**
Centro Nacional de Investigaciones Oncológicas (CNIO)
Barcelona Supercomputing Center (BSC)
ander.intxaurrondo@bsc.es

**Obdulia Rabal**
Center for Applied Medical Research (CIMA)
University of Navarra
orabal@unav.es

**Marta Villegas**
Centro Nacional de Investigaciones Oncológicas (CNIO)
Barcelona Supercomputing Center (BSC)
marta.villegas@bsc.es

**Martin Krallinger**
Centro Nacional de Investigaciones Oncológicas (CNIO)
Barcelona Supercomputing Center (BSC)
martin.krallinger@bsc.es

## Abstract

One of the biomedical entity types of relevance for medicine or biosciences are chemical compounds and drugs. The correct detection these entities is critical for other text mining applications building on them, such as adverse drug-reaction detection, medication-related fake news or drug-target extraction. Although a significant effort was made to detect mentions of drugs/chemicals in English texts, so far only very limited attempts were made to recognize them in medical documents in other languages. Taking into account the growing amount of medical publications and clinical records written in Spanish, we have organized the first shared task on detecting drug and chemical entities in Spanish medical documents. Additionally, we included a clinical concept-indexing sub-track asking teams to return SNOMED-CT identifiers related to drugs/chemicals for a collection of documents. For this task, named PharmaCoNER, we generated annotation guidelines together with a corpus of 1,000 manually annotated clinical case studies. A total of 22 teams participated in the sub-track 1, (77 system runs), and 7 teams in the sub-track 2 (19 system runs). Top scoring teams used sophisticated deep learning approaches yielding very competitive results with F-measures above 0.91. These results indicate that there is a real interest in promoting biomedical text mining efforts beyond English. We foresee that the PharmaCoNER annotation guidelines, corpus and participant systems will foster the development of new resources for clinical and biomedical text mining systems of Spanish medical data.

## 1 Introduction

Efficient access to mentions of drugs, medications and chemical entities contained in clinical texts, scientific articles, patents or even the web is a pressing need shared by biomedical researchers and clinicians (Krallinger et al., 2017). Biomedical text mining is one of the most prolific application domains of natural language processing technologies (Zweigenbaum et al., 2007). The recognition of pharmaceutical drugs/chemical entities is a critical step required for the subsequent detection of relations with other biomedically relevant entities such as genes/proteins, diseases or adverse reactions (Vazquez et al., 2011). Text mining and information extraction systems were published that tried to find protein-drug relations (including ligand-protein interactions and pharmacogenomics information), medication-related al-



*Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10
Hong Kong, China, November 4, 2019. 

lergies, chemical metabolic reactions, drug-drug interactions (Herrero-Zazo et al., 2013), disease-drug relations, as well as drug safety-related issues. The correct identification of drug mentions is also needed for other complex relation types like drug dosage recognition, duration of medical treatments or drug repurposing.

The importance of chemical and drug name recognition motivated several-shared tasks in the past, such as the CHEMDNER tracks (Krallinger et al., 2015) or the i2b2 medication challenge (Uzuner et al., 2010b,a), with a considerable number of participants and impact (Doan et al., 2010; Yang, 2010).

Currently, most of the biomedical and clinical NLP research, is done on English documents, while only few tasks were carried out using non-English texts, or were multilingual. Nonetheless, it is important to highlight that there is a considerable amount of biomedically relevant content published in other languages than English, and particularly clinical texts are entirely written in the native language of each country.

Spanish is a language spoken by more than 572 million people in the world today, either as a native, second or foreign language. It is the second language in the world by number of native speakers with more than 477 million people. According to results derived from WHO statistics, just in Spain there are over 180 thousand practicing physicians, more than 247 thousand nursing and midwifery personnel or 55 thousand pharmaceutical personnel. These facts, and the extrapolation to other Spanish speaking countries explains why a considerable subset of the PubMed database records corresponds to Spanish medical articles. Moreover, PubMed does only contain a part of the medical literature originally published in Spanish, which is also stored in other resources such as MEDES, SciELO, IBECS or CUIDEN.

Following the outline of previous chemical/drug NER efforts, in particular the BioCreative CHEMDNER tracks, we have carried out the first task on chemical and drug mention recognition from Spanish medical texts, namely from a corpus of Spanish clinical case studies. Thus, this track addressed the automatic extraction of chemical, drug, gene/protein mentions from clinical case studies written in Spanish. The main aim was to promote the development of named entity recognition tools of practical relevance, that is, chemical and drug mentions in non-English content, determining the current-state-of-the art, identifying challenges and comparing the strategies and results to those published for English data.

## 2 Methods

### 2.1 Track Description

The PharmaCoNER track was one of the six tracks of the BioNLP-OST 2019 / EMNLP-IJCNLP workshop[1]. It was the first community challenge track devoted to the recognition of pharmaceutical drugs and chemical entities in medical texts in Spanish.

For this track, two scenarios or sub-tracks were proposed:

- *NER offset and entity classification*. The first sub-track focused on the recognition and classification of entities.

- *Concept indexing*. The second sub-track consisted of concept indexing, where, for each document, the participating teams had to generate the list of the unique SNOMED-CT concept identifiers, which were compared to the manually annotated concept IDs corresponding to the pharmaceutical drugs and chemical entities.

### 2.2 Track data

We prepared a manually classified collection of clinical case report sections derived from open access Spanish medical publications, named the Spanish Clinical Case Corpus (SPACCC)[2]. The corpus contained a total of 1,000 clinical cases / 396,988 words. It is noteworthy that this kind of narrative shows properties of both the biomedical and medical literature, as well as clinical records. Case reports are considered as the scientific paper of a single clinical observation. Moreover, the clinical cases were not restricted to a single medical discipline, covering a variety of medical disciplines, including oncology, urology, cardiology, pneumology or infectious diseases. This is key to cover a diverse set of chemicals and drugs.

The PharmaCoNER corpus had a total of 7,624 entity mentions, corresponding to four different mention types[3]. Figure 1 shows a screenshot of a

[1] https://2019.bionlp-ost.org/

[2] https://github.com/PlanTL-SANIDAD/SPACCC

[3] For a detailed description of the mentions types, see (Rabal et al., 2018).


Figure 1: PharmaCoNER annotation example.

clinical case annotated using the BRAT tool. The overall annotation statistics were:

- NORMALIZABLES (normalizable): 4,398 mentions of chemicals that could be manually normalized to a unique concept identifier (primarily SNOMED-CT).
- NO_NORMALIZABLES (not normalizable): 50 mentions of chemicals that could not be normalized manually to a unique concept identifier.
- PROTEINAS (proteins): 3,009 mentions of proteins and genes following an adaptation of the BioCreative GPRO track annotation guidelines. This class included also peptides, peptide hormones and antibodies.
- UNCLEAR: 167 cases of general substance class mentions of clinical or biomedical relevance, including certain pharmaceutical formulations, general treatments, chemotherapy programs, vaccines and a predefined set of general substances (e.g.: Estragn, Silimarina, Bromelana, Melanina, Vaselina, Lanolina, Alcohol, Tabaco, Marihuana, Cannabis, Opio and Gluten)[4].

The annotation process of the PharmaCoNER corpus was inspired by previous annotation schemes and corpora used for the BioCreative CHEMDNER (Krallinger et al., 2015) and GPRO tracks (Pérez-Pérez et al., 2017), translating the guidelines used for these tracks into Spanish and adapting them to the characteristics and needs of clinically oriented documents by modifying the annotation criteria and rules to cover medical information needs. This adaptation was carried out in collaboration with practicing physicians and medicinal chemistry experts. The adaptation, translation and refinement of the guidelines (Rabal et al., 2018) was done on a sample set of the SPACCC corpus and linked to an iterative process of annotation consistency analysis through inter-annotator agreement (IAA) studies until a high annotation quality in terms of IAA was reached. The final, IAA measure obtained for this corpus was calculated on a set of 50 records that were double annotated (blinded) by two different expert annotators, reaching a pairwise agreement of 93% on the exact entity mention comparison level and 76% agreement when also the entity concept normalization was taken into account. Entity normalization was carried out primarily against the SNOMED-CT knowledge base. Note that there is a SNOMED-CT version directly released by the Spanish Ministry of Health twice a year.

The PharmaCoNER corpus was randomly sampled into three subsets: the train set (500 clinical cases), and the development and test sets (250 clinical cases each). These clinical cases were

[4]Mentions of this class were not part of the entities evaluated by this track, but served as additional annotations of medical relevance.

manually annotated using a customized version of AnnotateIt. Then, the BRAT annotation toolkit (Stenetorp et al., 2012) was used to correct errors and add missing annotations. The statistics of the number of label for each datasets are shown in Table 1.

Table 1: Distribution of labels in the PharmaCoNER datasets.

| Label | Train | Dev | Test | Overall |
|---|---|---|---|---|
| NORMALIZABLES | 2,304 | 1,121 | 973 | 4,398 |
| NO_NORMALIZABLES | 24 | 16 | 10 | 50 |
| PROTEINAS | 1,405 | 745 | 859 | 3,009 |
| UNCLEAR | 89 | 44 | 34 | 167 |

Together with the test set, we released an additional collection of 3,501 documents (background set[5]) to make sure that participating teams were not able to do manual corrections and also to promote that these systems would potentially be able to scale to larger data collections.

Moreover, we provided also the following resources: (1) Spanish medical text tokenizer, sentence splitter, lemmatizer and POS tagger; (2) Dictionary of chemicals, compounds and drugs in Spanish; (3) Sense inventory of Spanish medical abbreviation and their long forms; (4) Spanish drug naming file with prefixes and suffixes rules; and (5) a large background set of medical and health documents in Spanish.

### 2.3 Evaluation metrics

We released an evaluation script that supported the evaluation of the predictions of the participating teams. For both sub-tracks, the primary evaluation metrics used consisted of standard measures from the NLP community, namely micro-averaged precision, recall, and balanced F-score, the last one being the official evaluation measure:

$$\text{Precision: } P = \frac{TP}{TP+FP}$$

$$\text{Recall: } R = \frac{TP}{TP+FN}$$

$$\text{F-score: } F1 = 2 * \frac{(P*R)}{(P+R)}$$

where TP = true positives, FP = false positive and FN = false negative.

[5]The background set included the training, development and test sets, and an additional collection of 2,751 unlabeled clinical cases (total of 3,751 clinical cases).

Teams could submit up to five prediction files (or system runs) in a predefined prediction format: BRAT, for sub-track 1, and TSV files, for sub-track 2.

## 3 Participation and Results

### 3.1 Participation

To participate in the PharmaCoNER track it was necessary to register both on the official website[6] and in the CodaLab competition[7]. Training and development sets were made available for download on the official website[8], and the evaluation script was uploaded to GitHub[9], to ensure a transparent evaluation.

As we already said, submissions had to be provided in a predefined prediction format: BRAT, for sub-track 1, and TSV files, for sub-track 2. Additionally we plan to release the corpus also in the popular PubAnnotation format (Kim and Wang, 2012).

The participants had a period of almost two months to develop their system. In the middle of this period, the test and background sets were released with the 3,751 documents that the participants had to process and label, although the final evaluation was done only on the 250 documents corresponding to the test set. The intention was to use the background set to enable the construction of participant-generated Silver Standard corpus. As we have mentioned, the participants could submit a maximum of 5 system runs, and, once the submission deadline expired, we published the Gold Standard annotations of the test set, in order to ensure a transparent evaluation process and help participants to carry out a more detailed error analysis.

A total of 22 teams participated in the sub-track 1, submitting a total of 77 systems, and 7 teams in the sub-track 2, submitting a total of 19 runs. Teams from eleven different nationalities participated in the track: seven teams from Spain, three from China, and one team from each: Finland, France, India, Japan, Romania, Russia, United Kingdom and the United States. Three participants belong to a commercial institution. Table

[6]http://temu.bsc.es/pharmaconer/

[7]https://competitions.codalab.org/competitions/23159

[8]http://temu.bsc.es/pharmaconer/index.php/data/

[9]https://github.com/PlanTL-SANIDAD/PharmaCoNER-CODALAB-Evaluation-Script

Table 2: Overview of Team Participation in the PharmaCoNER track.

| Username | Organization/Institution/Company | Members | Country | Comm. |
|---|---|---|---|---|
| alily | Carlos III University of Madrid | 3 | Spain | No |
| ayan7246 | Unaffiliated | 1 | India | No |
| chaanim | University of Turku | 2 | Finland | No |
| CongSun | Dalian University of Technology | 3 | China | No |
| Edson | University of Côte d'Azur | 4 | France | No |
| foxlf823 | UMASS Lowell | 3 | United States | No |
| FSL | Unaffiliated | 2 | Spain | No |
| ghada.alfattni | University of Manchester | 3 | United Kingdom | No |
| ixamed | University of the Basque Country | 5 | Spain | No |
| JoyHan | - | - | - | - |
| lluisp | Universitat Politècnica de Catalunya | 1 | Spain | No |
| lukas.lange | Bosch Center for Artificial Intelligence | 3 | Germany | Yes |
| m-stoeckel | Goethe University Frankfurt | 2 | Germany | No |
| m.domrachev | Unaffiliated | 1 | Russia | No |
| naiven | JD | 1 | China | Yes |
| plubeda | Universidad de Jan | 4 | Spain | No |
| raduion | Research Institute for AI "Mihai Draganescu" | 3 | Romania | No |
| rriveraz | Carlos III University of Madrid | 3 | Spain | No |
| sohrab | National Institute of Advanced Industrial Science and Technology | 4 | Japan | No |
| tEarth | - | - | - | - |
| uyaseen | Siemens AG | 2 | Germany | Yes |
| VSP | Carlos III University of Madrid | 1 | Spain | No |
| xiongying | Harbin Institute of Technology | 4 | China | Yes |

2 summarizes the most relevant information about the participants (we lack the information from two of the teams, because they registered at CodaLab, but not at our website).

### 3.2 Baseline system

We produced three baseline systems for the track: The first one is a very simple baseline based on vocabulary transfer, and the other two baseline systems are competitive baselines based on the PharmaCoNER Tagger (Armengol-Estapé et al., 2019), a deep learning-based tool for automatically finding chemicals and drugs in Spanish medical texts.

In the vocabulary transfer approach, each annotation from the train and development datasets was transferred to the test dataset using strict string matching. For those cases where the text was the same, but the entity type was different, we decided to annotate all entity types that matched that text.

In the two baselines based on the PharmaCoNER Tagger, we used the default parameters, a hidden layer of size 300, and early stop (best model at epoch 35). The models were trained using the GloVe embeddings (Pennington et al., 2014) from SBWC[10] (from now on *baseline-glove*) and the Medical Word Embeddings for Spanish (Soares et al., 2019) (from now on *baseline-med*). The corpus was tokenized using spaCy.

[10] https://github.com/dccuchile/spanish-word-embeddings

### 3.3 Results

Table 3 shows the results for sub-track 1 (*NER offset and entity type classification*), ordered by team performance (first column), then system performance (second column).

The top scoring system was submitted by *xiongying*, with an F-score of 0.91052, being relatively close to the next two participants: *FSL*, ranked 2nd with a F-score of 0.90968, and *m-stoeckel*, ranked 3rd with a F-score of 0.89888. Participant *Edson* submitted five systems that scored almost zero. Once he noticed the error, he submitted two fixed submissions. These submissions were made after the publication of the results but before the release of the test set with GS annotations. These late submissions of *Edson* are marked with an asterisks in the table, including the hypothetical ranking of his team/systems.

Note that all of the teams were well above the baseline based on vocabulary transfer, which would rank last if we ignored the submission with errors. The competitive baseline trained with the GloVe embeddings would rank 16, and the one trained with embeddings that are specific for clinical texts in Spanish would rank 13. It is remarkable that 12 teams out of 20 managed to beat a very competitive baseline based on a well known Deep Learning tool.

Table 4 shows the results for sub-track 2 (*Concept Indexing*), ordered by team performance (first

Table 3: Results for sub-track 1: *NER offset and entity type classification*.

| Team Rank | System Rank | User | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 1 | 1 | xiongying | 0.91226 | 0.90879 | 0.91052 |
| | 2 | | 0.91589 | 0.90445 | 0.91013 |
| | 3 | | 0.91008 | 0.90662 | 0.90835 |
| | 4 | | 0.90751 | 0.90554 | 0.90652 |
| | 5 | | 0.90205 | 0.90988 | 0.90595 |
| 2 | 3 | FSL | 0.90625 | 0.91314 | 0.90968 |
| 3 | 7 | m-stoeckel | 0.90708 | 0.89082 | 0.89888 |
| | 8 | | 0.89297 | 0.89685 | 0.89491 |
| | 13 | | 0.88839 | 0.86369 | 0.87586 |
| 4 | 9 | CongSun | 0.90463 | 0.88056 | 0.89243 |
| | 10 | | 0.90704 | 0.87405 | 0.89024 |
| | 14 | | 0.89183 | 0.85939 | 0.87531 |
| | 17 | | 0.88732 | 0.85071 | 0.86863 |
| 5 | 11 | naiven | 0.90315 | 0.87079 | 0.88668 |
| 6 | 12 | lukas.lange | 0.88950 | 0.88274 | 0.88610 |
| | 27 | | 0.85162 | 0.87242 | 0.86189 |
| | 28 | | 0.86307 | 0.85885 | 0.86095 |
| | 31 | | 0.85078 | 0.86048 | 0.85560 |
| | 32 | | 0.85520 | 0.85288 | 0.85404 |
| 7 | 15 | chaanim | 0.87568 | 0.87188 | 0.87378 |
| 8 | 16 | foxlf823 | 0.88098 | 0.85993 | 0.87033 |
| | 22 | | 0.87218 | 0.85939 | 0.86574 |
| | 23 | | 0.87674 | 0.85342 | 0.86492 |
| 9 | 18 | ixamed | 0.90222 | 0.83659 | 0.86817 |
| | 21 | | 0.90088 | 0.83388 | 0.86608 |
| | 42 | | 0.82981 | 0.85233 | 0.84092 |
| | 49 | | 0.81914 | 0.80402 | 0.81151 |
| | 50 | | 0.81914 | 0.80402 | 0.81151 |
| 10 | 19 | sohrab | 0.86881 | 0.86645 | 0.86763 |
| | 26 | | 0.87079 | 0.85613 | 0.86340 |
| | 39 | | 0.85320 | 0.83931 | 0.84620 |
| | 41 | | 0.83665 | 0.84528 | 0.84094 |
| | 46 | | 0.88483 | 0.77579 | 0.82673 |
| 11 | 20 | uyaseen | 0.90581 | 0.83008 | 0.86629 |
| | 24 | | 0.90482 | 0.82573 | 0.86347 |
| | 25 | | 0.90482 | 0.82573 | 0.86347 |
| | 33 | | 0.84644 | 0.85885 | 0.85260 |
| | 37 | | 0.88941 | 0.81650 | 0.85140 |
| 12 | 29 | m.domrachev | 0.87073 | 0.84473 | 0.85754 |
| | 30 | | 0.87073 | 0.84473 | 0.85754 |
| - | - | *baseline-med* | *0.87020* | *0.83713* | *0.85335* |
| 13 | 34 | rriveraz | 0.88538 | 0.82193 | 0.85248 |
| | 35 | | 0.88538 | 0.82193 | 0.85248 |
| | 36 | | 0.88538 | 0.82193 | 0.85248 |
| 14 | 38 | raduion | 0.90189 | 0.80347 | 0.84984 |
| | 40 | | 0.90043 | 0.79533 | 0.84462 |
| | 47 | | 0.89327 | 0.76330 | 0.82319 |
| | 48 | | 0.78281 | 0.84528 | 0.81284 |
| | 52 | | 0.92530 | 0.71281 | 0.80527 |
| 15 | 43 | lluisp | 0.88882 | 0.78990 | 0.83645 |
| | 44 | | 0.89176 | 0.78719 | 0.83622 |
| | 45 | | 0.88991 | 0.78556 | 0.83449 |
| | 53 | | 0.81160 | 0.76710 | 0.78872 |
| | 61 | | 0.73211 | 0.73887 | 0.73548 |
| - | - | *baseline-glove* | *0.83259* | *0.80999* | *0.82113* |
| 16 | 51 | ghada.alfattni | 0.85039 | 0.77144 | 0.80900 |
| | 55 | | 0.82776 | 0.72530 | 0.77315 |
| 17 | 54 | plubeda | 0.88507 | 0.69815 | 0.78058 |
| | 56 | | 0.85992 | 0.69653 | 0.76965 |
| | 60 | | 0.92602 | 0.61835 | 0.74154 |
| | 62 | | 0.84404 | 0.64929 | 0.73397 |
| 18 | 57 | alily | 0.86034 | 0.68893 | 0.76515 |
| | 59 | | 0.86981 | 0.67101 | 0.75759 |
| 19 | 58 | VSP | 0.81621 | 0.71607 | 0.76287 |
| 20 | 63 | ayan7246 | 0.74668 | 0.61129 | 0.67224 |
| | 67 | | 0.43812 | 0.48046 | 0.45831 |
| | 68 | | 0.36910 | 0.47991 | 0.41728 |
| | 69 | | 0.33333 | 0.48046 | 0.39360 |
| | 70 | | 0.52283 | 0.19273 | 0.28163 |
| 21 | 64 | JoyHan | 0.88519 | 0.54098 | 0.67155 |
| | 65 | | 0.52523 | 0.52666 | 0.52594 |
| | 66 | | 0.88350 | 0.37193 | 0.52349 |
| - | - | *baseline-vt* | *0.67330* | *0.60641* | *0.63810* |
| 22 | 71 | Edson | 0.00280 | 0.00163 | 0.00206 |
| | 72 | | 0.00008 | 0.00163 | 0.00015 |
| | 73 | | 0.00007 | 0.00217 | 0.00014 |
| | 74 | | 0.00007 | 0.00217 | 0.00014 |
| | 75 | | 0.00002 | 0.00054 | 0.00004 |
| 20* | 60* | Edson | 0.80660 | 0.68920 | 0.74330 |
| | 70* | | 0.63350 | 0.14930 | 0.24160 |

Table 4: Results for sub-track 2: *Concept Indexing*.

| Team Rank | System Rank | User | Precision | Recall | F1 |
|---|---|---|---|---|---|
| 1 | 1 | FSL | 0.91108 | 0.92083 | 0.91593 |
| 2 | 2 | ixamed | 0.87964 | 0.82882 | 0.85347 |
| | 3 | | 0.87623 | 0.82810 | 0.85149 |
| | 9 | | 0.82374 | 0.83666 | 0.83015 |
| | 10 | | 0.81232 | 0.80884 | 0.81058 |
| 3 | 4 | xiongying | 0.82835 | 0.85021 | 0.83914 |
| | 5 | | 0.83809 | 0.83809 | 0.83809 |
| | 6 | | 0.82202 | 0.84665 | 0.83415 |
| | 7 | | 0.82032 | 0.84665 | 0.83327 |
| | 8 | | 0.81699 | 0.84379 | 0.83018 |
| 4 | 11 | sohrab | 0.87532 | 0.73609 | 0.79969 |
| | 12 | | 0.88003 | 0.73252 | 0.79953 |
| 5 | 13 | plubeda | 0.85207 | 0.63267 | 0.72616 |
| | 14 | | 0.82887 | 0.61840 | 0.70833 |
| | 15 | | 0.87879 | 0.55849 | 0.68295 |
| | 16 | | 0.83350 | 0.57846 | 0.68295 |
| 6 | 17 | VSP | 0.66502 | 0.55215 | 0.60335 |
| 7 | 18 | rriveraz | 0.50000 | 0.49287 | 0.49641 |
| | 19 | | 0.48641 | 0.49786 | 0.49207 |

Table 5: Results by category for sub-track 1.

| | NORMALIZABLES | | | NO_NORMALIZABLES | | | PROTEINAS | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| Min | 0.31976 | 0.17986 | 0.28618 | 0.00000 | 0.00000 | 0.00000 | 0.32377 | 0.12224 | 0.19981 |
| Mean | 0.87217 | 0.81754 | 0.83880 | 0.19844 | 0.03276 | 0.04984 | 0.81654 | 0.76428 | 0.78494 |
| Median | 0.90977 | 0.86434 | 0.87922 | 0.00000 | 0.00000 | 0.00000 | 0.85626 | 0.82421 | 0.83445 |
| Maximum | 0.95924 | 0.94142 | 0.94253 | 1.00000 | 0.40000 | 0.38095 | 0.89831 | 0.89406 | 0.88709 |
| Std Dev | 0.12742 | 0.12967 | 0.13065 | 0.38104 | 0.06854 | 0.09247 | 0.11328 | 0.15302 | 0.13668 |
| Best team | raduion | FSL | xiongying | m-stoeckel<br>sohrab<br>xiongying | FSL | FSL | plubeda | xiongying | xiongying |

column), then system performance (second column). The top scoring system for sub-track 2 was submitted by *FSL*, with a F-score of 0.91593, showing a significantly better result when compared to the second best submission (more than 6 points) provided by *ixamed*, with a F-score of 0.85347. The third team was *xiongying*, the best participant in the sub-track 1, with a F-score of 0.83914.

Some statistics of the results are shown in Table 6. There was a high variability among the systems, with a difference of 6 point between the best system and the median for sub-track 1, and of 10 points for sub-track 2. The difference between the best system and the mean of all system was still higher. This proved that the task, was quite difficult.

As additional analysis, results by category, including the best teams for category and metric, are shown in Table 5. The performance of the systems was systematically better for the NORMALIZABLES category, 4-9 points better in respect with the PROTEINAS category. Surprisingly, the median for the NO_NORMALIZABLES category was 0, suggesting that at least half of the systems ignored this category.

Table 6: Statistics by track.

| Track | Measure | Precision | Recall | F1 |
|---|---|---|---|---|
| 1 | Minimum | 0.33333 | 0.19273 | 0.28163 |
| | Mean | 0.84211 | 0.77916 | 0.80493 |
| | Median | 0.88417 | 0.82791 | 0.85248 |
| | Maximum | 0.92602 | 0.91314 | 0.91052 |
| | Std Dev | 0.12071 | 0.13840 | 0.12819 |
| 2 | Minimum | 0.48641 | 0.49287 | 0.49207 |
| | Mean | 0.80152 | 0.72885 | 0.75936 |
| | Median | 0.82887 | 0.80884 | 0.81058 |
| | Maximum | 0.91108 | 0.92083 | 0.91593 |
| | Std Dev | 0.11975 | 0.14059 | 0.12057 |

### 3.4 Combination of systems

In this section, we present an experiment we performed to combine the systems submitted to the track to see if we could improve the results. We combined the systems using a voting scenario: we accepted as good the annotations that had been predicted by N systems.

The first system accepted all the annotations

predicted by, at least, one of the systems, while the last one accepted only the annotations that were predicted by, at least, N systems. The results of this experiment are shown in Table 7.

As expected, as the value of N increased (the number of required votes was increased), the recall got worse and the precision improved. Based on the maximum value of F-score for sub-track 1 on the train and development sets we selected 20 as the optimum value for combining systems (F-score of 0.98408). We used this value for N on the test set, obtaining an F-score of 0.92355, 1.3 points better than the best system. This score was lower than the best one that could be obtained for the test set (0.92426, with N = 18), but the difference was (in practice) negligible.

The combined systems did not improve the results for sub-track 2. The maximum value of F-score on the train and development sets was obtained combining 6-7 systems (F-score of 0.97352 in the Dev set for N = 6). This scored 0.87073 in the test set, 4.5 points below the best system. This was probably a consequence of amount of systems and the performance gap between the best systems and the others. For the future, we will combine the system using more sophisticated approaches.

Table 7: Combining systems using a voting scheme.

| Track | # | Train | Dev | Test |
|---|---|---|---|---|
| 1 | 1 | 0.77448 | 0.64485 | 0.36036 |
| | 2 | 0.89285 | 0.78679 | 0.71539 |
| | 3 | 0.94173 | 0.85545 | 0.78403 |
| | 4 | 0.95583 | 0.88711 | 0.82505 |
| | 5 | 0.96638 | 0.91222 | 0.85261 |
| | 6 | 0.97523 | 0.92725 | 0.87124 |
| | 7 | 0.98024 | 0.93859 | 0.88286 |
| | 8 | 0.98452 | 0.94902 | 0.89519 |
| | 9 | 0.98792 | 0.95772 | 0.90438 |
| | 10 | 0.98989 | 0.96386 | 0.90828 |
| | 11 | 0.99160 | 0.96906 | 0.91038 |
| | 12 | 0.99319 | 0.97280 | 0.91436 |
| | 13 | 0.99386 | 0.97431 | 0.91615 |
| | 14 | 0.99412 | 0.97760 | 0.91880 |
| | 15 | 0.99505 | 0.97808 | 0.92124 |
| | 16 | 0.99518 | 0.97856 | 0.92253 |
| | 17 | 0.99558 | 0.98162 | 0.92320 |
| | 18 | 0.99598 | 0.98368 | 0.92426 |
| | 19 | 0.99571 | 0.98362 | 0.92418 |
| | **20** | **0.99571** | **0.98408** | **0.92355** |
| | 21 | 0.99585 | 0.98244 | 0.92372 |
| | 22 | 0.99598 | 0.98182 | 0.92074 |
| | 23 | 0.99638 | 0.97740 | 0.91872 |
| | 24 | 0.99638 | 0.97569 | 0.91641 |
| | 25 | 0.99651 | 0.96952 | 0.91202 |
| | 26 | 0.99610 | 0.96665 | 0.90815 |
| | 27 | 0.99516 | 0.96294 | 0.90392 |
| | 28 | 0.99421 | 0.95715 | 0.90152 |
| | 29 | 0.99217 | 0.95135 | 0.89164 |
| | 30 | 0.99025 | 0.94474 | 0.88598 |
| | 31 | 0.98669 | 0.93575 | 0.88197 |
| | 32 | 0.98641 | 0.92965 | 0.87602 |
| | 33 | 0.98462 | 0.92197 | 0.86712 |
| | 34 | 0.98198 | 0.91708 | 0.85794 |
| | 35 | 0.98073 | 0.91311 | 0.85002 |
| | 36 | 0.97738 | 0.90290 | 0.84011 |
| | 37 | 0.97285 | 0.89545 | 0.82827 |
| | 38 | 0.97058 | 0.88458 | 0.81402 |
| | 39 | 0.96829 | 0.87545 | 0.80040 |
| | 40 | 0.96397 | 0.86574 | 0.78042 |
| | 41 | 0.95860 | 0.85169 | 0.75992 |
| | 42 | 0.95258 | 0.83427 | 0.73265 |
| | 43 | 0.94455 | 0.81401 | 0.69613 |
| | 44 | 0.93128 | 0.79135 | 0.66063 |
| | 45 | 0.91042 | 0.76815 | 0.57034 |
| | 46 | 0.88054 | 0.72273 | 0.50117 |
| | 47 | 0.84009 | 0.66146 | 0.46420 |
| | 48 | 0.75949 | 0.57517 | 0.41148 |
| | 49 | 0.34449 | 0.26516 | 0.20968 |
| 2 | 1 | 0.65485 | 0.63039 | 0.57822 |
| | 2 | 0.81233 | 0.80333 | 0.70237 |
| | 3 | 0.90781 | 0.91254 | 0.80749 |
| | 4 | 0.92741 | 0.93277 | 0.81967 |
| | 5 | 0.97599 | 0.96739 | 0.84716 |
| | **6** | **0.98009** | **0.97352** | **0.87073** |
| | 7 | 0.98298 | 0.97020 | 0.87106 |
| | 8 | 0.97983 | 0.97061 | 0.87719 |
| | 9 | 0.97207 | 0.94607 | 0.86444 |
| | 10 | 0.96992 | 0.93864 | 0.86435 |
| | 11 | 0.95120 | 0.90585 | 0.85479 |
| | 12 | 0.94721 | 0.88803 | 0.84603 |
| | 13 | 0.88231 | 0.80623 | 0.80337 |
| | 14 | 0.87566 | 0.78624 | 0.79056 |
| | 15 | 0.85368 | 0.73964 | 0.75661 |
| | 16 | 0.81724 | 0.68416 | 0.72358 |
| | 17 | 0.43734 | 0.38884 | 0.37916 |
| | 18 | 0.43231 | 0.39216 | 0.36889 |
| | 19 | 0.40541 | 0.36462 | 0.33152 |

## 4 Discussion and Conclusions

The results of the first chemical and drug named entity recognition track from clinical case reports in Spanish are very encouraging, both in terms of the number of participants, not only from Spanish-speaking countries, as well as in terms of the obtained system results, which are already reaching a level of performance that would make the resulting tools very valuable resources for processing the vast amount of medical data generated worldwide in Spanish.

We had structured the track into two sub-tracks to cover different practical aspects of the resulting systems. The named entity recognition track of chemicals/drugs had the aim of serving as a building block task for future down-stream text mining of more complex information types, including the detection of medication duration, dosage, drug-drug-interactions, therapeutic target relations and drug/chemical induced adverse effects. The concept-indexing sub-track was more concerned with the development of sophisticated semantic retrieval engines and the exploitation of high impact normative terminologies such as SNOMED CT.

Surprisingly we had a considerably higher number of participants for the NER sub-track when compared to the concept-indexing sub-track. Future evaluation efforts should potentially consider also an entity grounding/normalization of chemical and drug mentions in clinical case reports.

Most of the participating systems were based on the use of sophisticated deep learning and neural net approaches, which are becoming the state of the art methods for named entity recognition tasks also in specialized domains such as biomedicine or for non-English data.

When analyzing the more difficult mention types for participating teams, it is still clear that very short abbreviations (1-2 letters) are cumbersome to recognize correctly, due to their high level of implicit ambiguity. Solving such cases would probably require larger manually annotated corpora or the generation of other complementary resources specifically suited for the recognition and resolution of short abbreviations. We did not observe any particular issues related to the clinical disciplines of the case reports, thus it seems that drug NER systems should work well across all medical specialties. It is important to place the very competitive results obtained for PharmaCoNER into its context, in terms data collections used. When compared to the biomedical literature or medicinal chemistry patents, clinical case reports show a lower degree of variability in terms of the chemicals and drug mentions used, as in the clinic only a limited number of medications and chemical entities are being used for treatment, biochemical testing or explored in clinical settings and analysis.

The construction of high quality Gold Standard manually annotated corpora can be considered one of the major bottlenecks for the development of biomedical named entity recognition systems. During this task, we have promoted the collaborative generation of a larger Silver Standard corpus generated through the predictions of all the participating teams. A more detailed examination of this resource and approaches on how to optimally merge/combine multiple annotations and in turn train new systems using this silver standard dataset might give new insights on how to speed up the creation of new NER tools/annotated datasets.

One of the difficulties we have also encountered during this task was due to the use of a very popular third party platform for organizing online shared tasks on data mining tasks, including text mining and NLP. The explored resource, Codalab, had a server crash, and no proper up to date backup system in place (including user registration info, as well as data collections). Thus, the use of resources with a more focused support for biomedical text mining datasets, corpora, services and shared task organization, such as PubAnnotation would have been a better choice for hosting all the relevant data and predictions for biomedical shared tasks.

## Acknowledgments

We acknowledge the Encargo of Plan TL (SEAD) to CNIO and BSC for funding, and the scientific committee for their valuable comments and guidance. We would also like to thank Siamak Barzegar for his help in setting up PharmaCoNER at CodaLab.

## References

Jordi Armengol-Estapé, Felipe Soares, Montserrat Marimon, and Martin Krallinger. 2019. Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics Inform*, 17(2):e15–.

Son Doan, Lisa Bastarache, Sergio Klimkowski, Joshua C Denny, and Hua Xu. 2010. Integrating existing natural language processing tools for medication extraction from discharge summaries. *Journal of the American Medical Informatics Association*, 17(5):528–531.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Jin-Dong Kim and Yue Wang. 2012. Pubannotation: A persistent and sharable corpus and annotation repository. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pages 202–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):S2.

Martin Krallinger, Obdulia Rabal, Analia Lourenco, Julen Oyarzabal, and Alfonso Valencia. 2017. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117(12):7673–7761.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Martin Pérez-Pérez, Obdulia Rabal, Gael Pérez-Rodríguez, Miguel Vazquez, Florentino Fdez-Riverola, Julen Oyarzabal, Alfonso Valencia, Anália Lourenço, and Martin Krallinger. 2017. Evaluation of chemical and gene/protein entity recognition systems at biocreative v. 5: the cemp and gpro patents tracks. In *BC V.5 - Workshop Proceedings*.

Obdulia Rabal, Ander Intxaurrondo, and Martin Krallinger. 2018. Guías de anotación y normalización de compuestos químicos.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. 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*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010a. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010b. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

Miguel Vazquez, Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2011. Text mining for drugs and chemical compounds: methods, tools and applications. *Molecular Informatics*, 30(6-7):506–519.

Hui Yang. 2010. Automatic extraction of medication information from medical discharge summaries. *Journal of the American Medical Informatics Association*, 17(5):545–548.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B Cohen. 2007. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5):358–375.