Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 241-253 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources

Detección, normalización y clasificación de especies, patógenos, humanos y alimentos en documentos clínicos: resumen de la tarea y los recursos LivingNER.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Darryl Estrada, Luis Gascó, Martin Krallinger
Barcelona Supercomputing Center, Spain
antoniomiresc@gmail.com

Abstract: There is a pressing need to generate tools for finding mentions of species, pathogens, or food from medical texts. To promote the development of such tools we organized the LivingNER task. LivingNER relied on a large Gold Standard corpus of 2000 carefully selected clinical cases in Spanish covering diverse specialties. It was manually annotated with species mentions that were also carefully mapped to their corresponding NCBI Taxonomy identifiers. Besides, we have generated Silver Standard versions of LivingNER for 7 languages: English, Portuguese, Galician, Catalan, Italian, French, and Romanian. LivingNER had three subtasks: LivingNERSpecies NER (species mention detection sub-task), LivingNER-Species Norm (species mention detection and normalization to NCBI taxonomy Ids), and LivingNERClinical IMPACT (a document classification task related to the detection of pets, animals-causing injuries, food, and nosocomial entities). We received and evaluated 62 systems from 20 teams from 11 countries worldwide, obtaining highly competitive results. Successful approaches typically modified pre-trained transformer-like language models (BERT, BETO, RoBERTa, etc.) and employed embedding distance metrics for entity linking. LivingNER corpus: doi.org/10.5281/zenodo.6376662

Keywords: named entity recognition, pathogens text mining, entity linking, NCBI Taxonomy.

Resumen: Existe la necesidad de generar herramientas para encontrar y normalizar menciones de especies, patógenos o alimentos en textos médicos. Para promover el desarrollo de tales herramientas hemos organizado la tarea LivingNER. La tarea LivingNER se basó en un corpus en español de 2000 casos clínicos cuidadosamente seleccionados, representando una diversidad de especialidades. El corpus fue anotado manualmente por expertos que también asignaron a las menciones sus correspondientes identificadores de la NCBI Taxonomy. Además, hemos generado versiones de LivingNER para otros 7 idiomas: inglés, portugués, gallego, catalán, italiano, francés y rumano. LivingNER se estructuró en tres subtareas: 1) LivingNER-Species NER (subtarea de detección de menciones de especies), 2) LivingNER-Species Norm (detección de especies y normalización a identificadores de NCBI Taxonomy) y 3) LivingNER-Clinical IMPACT (tarea de clasificación relacionada con la detección de mascotas, animales causantes de lesiones, alimentos y entidades nosocomiales). Recibimos y evaluamos 62 sistemas de 20 equipos de 11 países a nivel mundial, obteniendo resultados altamente competitivos. Generalmente, los enfoques más exitosos hicieron modificaciones a modelos de lenguaje basados en transformers (BERT, BETO, RoBERTa, etc.) y emplearon métricas de distancia de embeddings para la normalización de entidades. Corpus LivingNER: doi.org/10.5281/zenodo.6376662

Palabras clave: reconocimiento de entidades nombradas, minería de textos de patógenos, normalización de entidades, NCBI Taxonomy.

ISSN 1135-5948. DOI 10.26342/2022-69-21

1 Introduction

The semantic annotation of species or living organisms is critical to scientific disciplines like medicine, biology, ecology/biodiversity, nutrition, and agriculture. For instance, detecting species in clinical records underscores the burden of disease caused by pathogens in the case of infectious diseases; and identifying organisms and foods can reveal the cause of allergy-related conditions. Despite this undisputed relevance, organisms/species have relatively scarcely featured in NLP studies, particularly for non-English content.

Because of the significance of this task, hierarchical taxonomic relations have been developed over 250 years to determine rules and conventions to catalog species. And they have been recently transformed into computer-based terminological resources such as NCBI taxonomy (Schoch et al., 2020; Federhen, 2012), the Thompson scientific name list, the Catalogue of Life, the Global Names Index database, and the ITIS Catalogue. However, these efforts have not been adequately aligned with the development of automatic systems for semantic analysis of species mentions in text, especially when considering documents beyond English. Common challenges encountered are name changes (obsolete species names); homonymy with commonly used words (e.g., "spot" refers to the species *Leiostomus xanthurus* or "permit" to *Trachinotus falcatus*); abbreviations and acronyms (sometime highly ambiguous like EC, which can be used for the bacteria "*Escherichia coli*" and "*Enterobacter cloacae*," among others); misspelled names (*Escerichia coli* for *Escherichia coli*); coordinations and nested expressions ("human immunodeficiency viruses types 1 and 2"); vernacular forms (common names); and role names (e.g., athletes, responders).

To overcome these limitations, corpora and tools are already available for species identification in the English-language biomedical literature and their standardization to controlled vocabularies. For example, LINNAEUS (Gerner, Nenadic, and Bergman, 2010) and the SPECIES tool (Pafilis et al., 2013) are capable of detecting species mentions. Additionally, there have been shared tasks on information related to microorganisms/species, such as the Infectious Diseases (ID) task of BioNLP 2011 (Pyysalo et al., 2011). And the importance of detecting species mentions for gene mention entity linking to database records has been addressed using biomedical literature data in English (Krallinger, Leitner, and Valencia, 2010).

However, adapting these resources to languages other than English and document types different from biomedical literature is not trivial. This is aggravated by the lack of resources, common evaluation scenarios, and shared tasks in other languages.

The LivingNER task addressed these issues through (1) a challenge on Named Entity Recognition (NER) of species mentions, entity linking, and document classification and; (2) providing a manually, exhaustively annotated large corpus of Spanish clinical cases. All annotated organism/species mentions were manually mapped to the NCBI taxonomy and classified into four information axes related to relevant use cases.

The National Center for Biotechnology Information (NCBI) Taxonomy includes names of organisms classified primarily based on a phylogenetic hierarchy. The NCBI Taxonomy is a universal database, used by the International Nucleotide Sequence Database Collaboration (INSDC), which includes GenBank, the European Molecular Biology Laboratory (EMBL), and DNA Data Bank of Japan (DDBJ) as a single source of taxonomic classification to maintain consistency between databases. In NCBI, each unique code identifies a specific type of organism (e.g., Taxonomy ID: 5476 for *Candida Albicans*) or groups of organisms (Taxonomy ID: 40674 for mammals). NCBI Taxonomy was the controlled vocabulary chosen in the LINNAEUS corpus to standardize citations.

The corpus also distinguishes between antibiotic-resistant pathogens and hospital-acquired (nosocomial) infections, an increasing cause of morbidity and mortality when existing drugs become ineffective in eliminating some bacteria. It also references and standardizes mentions of the different floras of the human organism in preparation for the literature and clinical cases related to the human microbiome. For clinical and microbiological use, all forms of parasitic cycles are also noted. Some of the most relevant applications are associated with extracting information about highly prevalent sexually transmitted diseases, animals causing injuries, and animal-transmitted diseases (zoonoses) originating from pets and animal husbandry. The correct extraction of species and infectious diseases facilitates the classification of bacteria in the context of antibiotic resistance and nosocomial pathogens and diseases. Another potential application relates to food (infection, intoxication, healthy and unhealthy diets, etc.), allergy triggers, and epidemiologically relevant mentions such as close contacts, people living in the same household, and relatives.

LivingNER is the first track on comprehensive species mention recognition and grounding of non-English content with a clear potential for multilingual adaptation, particularly for pathogens, to generate high-quality living being mention recognition components. The LivingNER annotation guidelines and corpus are indispensable resources for detecting and classifying species and infectious diseases in Spanish-language literature and medical reports.

2 *Task Description*

2.1 Shared Task goal

The LivingNER shared task explores the automatic recognition of species mentions in clinical documents in the Spanish language, the assignment of NCBI Tax IDs, and the classification of each mention into four categories. Notably, LivingNER incorporates a subtask in which participants solve four real-world health use cases.

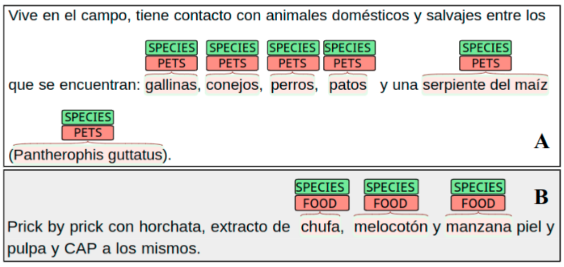
2.2 Sub-tasks

The LivingNER track contains three independent subtasks that are built one on top of the other:

LivingNER-Species NER track (Species mention entity recognition): given a plain text clinical case report document collection, participants must return the exact character offsets of all species mentions, both human and non-human.

LivingNER-Species Norm track (Species mention normalization): given a plain text clinical case report document collection, participating systems have to return all species mentions, together with their corresponding NCBI taxonomy concept identifiers.

LivingNER-Clinical IMPACT track given a collection of plain text documents, systems must (1) Perform a document classification according to information relevant to high-impact, real-world clinical use cases. The classification is multi-label, meaning that a single document may belong to several categories. And (2) Retrieve the list of NCBI Tax IDs that support the binary classification. Systems have to categorize the documents into the following information axes:


Vive en el campo, tiene contacto con animales domésticos y salvajes entre los que se encuentran: gallinas, conejos, perros, patos y una serpiente del maíz (Pantherophis guttatus).

Prick by prick con horchata, extracto de chufa, melocotón y manzana piel y pulpa y CAP a los mismos.

Figure 1: LivingNER example sentences annotated with Clinical Impact entities. (A) for Pets and farm animals, (B) for food species.

- *Pets and farm animals* in close contact with the patient (important for detecting animal-transmitted diseases such as toxoplasmosis, salmonellosis, cat-scratch disease, etc.).
- *Animals causing injuries.* Parasites are NOT included.
- *Food species.* It includes ingested aliments and any other food mentioned in the document. It excludes ingested items that are not food.
- *Nosocomial entities*: mentions corresponding to nosocomial/healthcare-associated infections.

2.3 Evaluation metrics

The micro-average f1-score has been the primary evaluation metric in the three subtasks. Additionally, micro-average precision and recall have been computed. The LivingNER evaluation library is available on GitHub (github.com/tonifuc3m/livingner-evaluation-library).

2.4 Baseline

For the LivingNER-Species NER subtask, we have employed the PathoTagIt-Base system. This competitive baseline is a deep neural network system trained with the LivingNER training dataset. The network is a customization of the BiLSTM-CRF architecture, and it employs word embeddings optimized for biomedical Spanish language (Soares et al., 2019). For a more in-depth description of the system, check the PharmaCoNER tagger paper (Armengol-Estapé et al., 2019). The code is available on GitHub (github.com/TeMU-BSC/PharmaCoNER-Tagger). There is also a web demo of the PathoTagIt-Base system (see temu.bsc.es/livingner/).

Finally, we have followed an indirect approach to create the document classifier of the LivingNER-Clinical Impact subtask. We have trained four different NER systems. The first NER system recognizes pet and farm animal mentions; the second mentions animals causing injuries; the third, food mentions; and the last, nosocomial entities. The four NER systems were run on the test set documents. The document containing it is automatically classified into the mention category if a mention is detected. For instance, in Figure 1 B, as soon as the NER system of food mentions recognizes ”chufa”, ”melocotón”, or ”manzana”, the document would be classified as a ”food document”.

3 Corpus and Resources

3.1 LivingNER Gold Standard Corpus

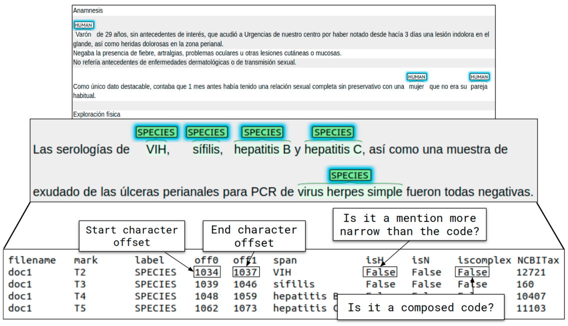
The LivingNER corpus is a collection of 2,000 clinical cases in Spanish from 20 medical specialties: infectious diseases (including Covid-19 cases), cardiology, neurology, oncology, ENT, dentistry, pediatrics, endocrinology, primary care, allergology, radiology, psychiatry, ophthalmology, psychiatry, urology, internal medicine, emergency and intensive care medicine, radiology, tropical medicine, and dermatology annotated with species [SPECIES] (including living organisms and microorganisms) and infectious diseases [ENFERMEDAD] mentions. Each mention in the corpus has been standardized to NCBI Taxonomy terminology. Finally, the species mentions have been classified into four classes of clinical interest to improve their usability (companion animals, animals causing injuries, food, and nosocomial entities).

The infectious diseases annotations are not used in the LivingNER shared task.

Document selection. The objective of document selection was to obtain a sufficient diversity of mentions representative of species in the clinical domain. We were mainly limited by the availability of relevant documents for certain specialties. For instance, obtaining clinical reports on tropical diseases was much easier than on pediatric allergies. The documents were also selected based on the richness of mentions, favoring the reports with a larger variety of species. Finally, we revised that certain diseases of great interest, notably COVID-19, but also zoonoses and parasite infections, AIDS, hepatitis C and others, were not excluded from our selection.

Corpus annotation. The LivingNER corpus has been annotated and standardized by a domain specialist with the support of a clinical specialist, who was also in charge of reviewing the mentions and their associated codes to arrive at a final version. The process of annotation and normalization of the corpus took place between 2020 and 2021, lasting approximately five months using the brat tool. Before starting the annotation, a first draft of these guides was created based on our previous annotation experiences MEDDOCAN (Marimon et al., 2019), CANTEMIST (Miranda-Escalada, Farré, and Krallinger, 2020) or MEDDOPROF (Lima-López et al., 2021) among others), and previous related work (Pafilis et al., 2013; Gerner, Nenadic, and Bergman, 2010). The annotation guidelines were refined by several rounds of interannotator agreement (IAA) consisting of parallel annotation of 5% of the corpus. After several rounds, a total IAA score of 0.942 for species and 0.885 for infectious diseases was reached. In addition, during the remainder of the LivingNER annotation, a random 10% of the papers were thoroughly reviewed to ensure that quality was maintained. There was also ongoing discussion about the content of the corpus, especially about difficult and ambiguous cases, with the aim of achieving the highest possible quality and refining these guidelines as much as possible.

The NCBI Taxonomy terminology was used to assign an identifier to each manual annotation, ensuring the usability of the corpus citations. The final version of the LivingNER corpus includes 30886 species mentions, of which 43.9% correspond to humans, 4580 are unique, and 29411 are normalized to NCBI Taxonomy. In addition, it contains 11841 infectious disease mentions, 4093 of which are unique, and 2283 are normalized to NCBI Taxonomy. The total is 42727 mentions. Finally, all species entries have been classified into four classes of clinical interest to improve their use (companion animals, animals causing injuries, food, nosocomial enti-



Anamnesis

HUMAN
Varón de 29 años, sin antecedentes de interés, que acudió a Urgencias de nuestro centro por haber notado desde hacía 3 días una lesión indolora en el glande, así como heridas dolorosas en la zona perianal.
Negaba la presencia de fiebre, artralgias, problemas oculares u otras lesiones cutáneas o mucosas.
No refería antecedentes de enfermedades dermatológicas o de transmisión sexual.

HUMAN HUMAN
Como único dato destacable, contaba que 1 mes antes había tenido una relación sexual completa sin preservativo con una mujer que no era su pareja habitual.

Exploración física

SPECIES SPECIES SPECIES SPECIES
Las serologías de VIH, sífilis, hepatitis B y hepatitis C, así como una muestra de

SPECIES
exudado de las úlceras perianales para PCR de virus herpes simple fueron todas negativas.

Start character offset

End character offset

Is it a mention more narrow than the code?

```
filename  mark  label    off0  off1  span         isH    isN    iscomplex NCBITax
doc1      T2    SPECIES  1034  1037  VIH          False  False  False     12721
doc1      T3    SPECIES  1039  1046  sífilis      False  False  False     160
doc1      T4    SPECIES  1048  1059  hepatitis B                          10407
doc1      T5    SPECIES  1062  1073  hepatitis C                          11103
```

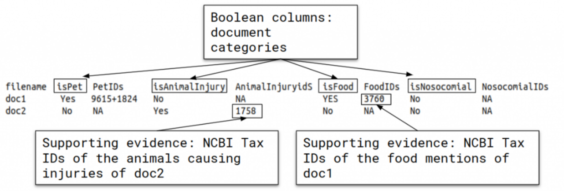
Is it a composed code?

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

ties, and antibiotic-resistant bacteria).

Corpus format. The LivingNER clinical case documents are released in plain text format with UTF-8 encoding. The annotations are included in a tab-separated document. In the LivingNER-SPECIES NER task, the annotations file has the following columns: filename, mark (identifier mention mark), label (SPECIES or HUMAN), off0 (starting position of the mention in the document), off1 (ending position of the mention in the document) and span. The LivingNER-Species Norm file, in addition to these columns, includes four more columns: isH (whether the span is narrower than the NCBITax assigned code), isN (whether the mention corresponds to a nosocomial infection), iscomplex (whether the span has assigned a combination of NCBITax codes) and NCBITax (mention code in the NCBI Taxonomy). Finally, the LivingNER-Clinical Impact annotation file has the following columns: filename, isPet, PetIDs (NCBITaxonomy codes of pet and farm animals present in document), isAnimalInjury, AnimalInjuryIDs (NCBITaxonomy codes of animals causing injuries present in document), IsFood, FoodIDs, (NCBITaxonomy codes of food mentions present in document), isNosocomial and NosocomialIDs (NCBITaxonomy codes of nosocomial species mentions present in document) (see Figure 3).

Corpus statistics. The LivingNER corpus contains 1,985 documents, which amounts to 65,373 sentences and 1,234,579 tokens. The corpus was randomly split into three subsets: training, validation, and test set. The test set is used for evaluation purposes of participating teams and consists of 485 records (15 extra records will be released shortly). Species and human mentions are found in all 1,985 documents. There are 30,604 such mentions (17158 species and 13446 human mentions) manually mapped to an NCBI Taxonomy ID. All human mentions have the 9606 NCBI Taxonomy ID, and there are 2,672 other unique codes. See Table 1 for the LivingNER corpus general statistics.



Boolean columns: document categories

```
filename  isPet  PetIDs     isAnimalInjury  AnimalInjuryidS  isFood  FoodIDs  isNosocomial  NosocomialIDs
doc1      Yes    9615+1824  No              NA               YES     3760     No            NA
doc2      No     NA         Yes             1758             No      NA       No            NA
```

Supporting evidence: NCBI Tax IDs of the animals causing injuries of doc2

Supporting evidence: NCBI Tax IDs of the food mentions of doc1

Figure 3: LivingNER Clinical Impact data format.

The 15 most common SPECIES mentions are shown in Figure 4B. It is noteworthy that seven out of the ten most common have the HUMAN label, despite there being fewer HUMAN annotations. This is because it is a more homogeneous entity type. Indeed, there are 707 different HUMAN mentions, while there are 3818 different SPECIES mentions.

In Figure 4.A, the 15 most common SPECIES NCBI Tax IDs are displayed. The main term of the code is shown instead of the numeric ID for clarity. While some terms are general (prokaryotes, viruses, eukaryotes), others are specific (HIV, *Enterobius vermicularis*, etc.) HIV appears very frequently partially because it is commonly mentioned in the context of patient serology results.

	Training	Validation	Test	Total
Documents	1000	500	485	1,850
SPECIES Annotations	9090	3817	4251	17158
HUMAN Annotations	7007	3289	3150	13446
Total Annotations	16097	7106	7401	30604
Unique codes	6738	2833	3101	12672
Sentences	34261	15107	16005	65373
Tokens	642813	296161	295605	1234579
Pets and farm animals	45	14	21	80
Animal causing injuries	107	12	22	141
Food species	255	107	163	525
Nosocomial entities	67	21	10	98

Table 1: DrugProt Gold Standard corpus statistics.

3.2 LivingNER Annotation Guidelines

The annotation guidelines posed many challenges, since many mentions of species in clinical documents are not identical to what

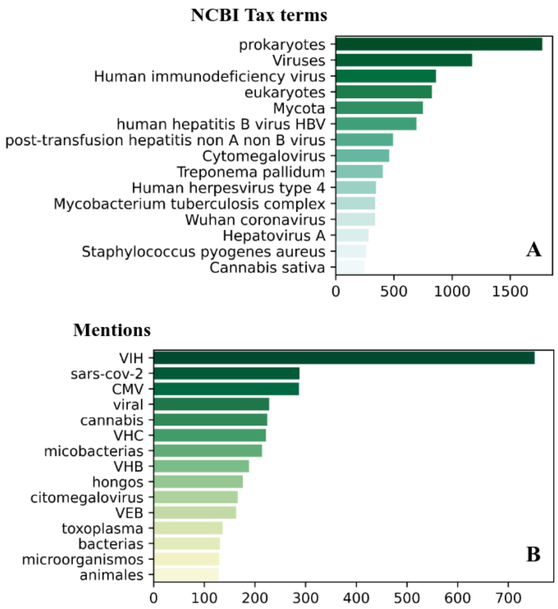


Figure 4: Number of appearances of (A) the main terms of the 15 most common codes, and (B) the 15 most common entities in the LivingNER Gold Standard.

we can find in the terminologies (for instance, hepatitis C virus is usually found as acronym, i.e., HCV, and *Staphilococcus aureus* as Staph A. and even its vernacular form, i.e., estafilococo). However, language was not the only challenge. We had to determine whether to include mentions undoubtedly related to infectious diseases and thus to pathogens, such as the term *vaccine*, if practically pathognomonic laboratory tests could be equated to the infection and thus the microorganism (i.e., VDRL to syphilis and thus to *Treponema pallidum*), and if we should distinguish between humans since their prevalence and responses to infection might greatly differ (for instance, neonates, children, males, females, the elderly). Naturally, these findings and decisions weighed heavily on the normalisation. We also wanted to include references to various parasite phases, since they are very important for microscopic diagnosis, and to the human microbiome, particularly the gut microbiome, since its study has imploded in the last decades, and the use of faecal transplant is already used to treat resistant *Clostridium difficile* infections and a large number of clinical trials to treat other conditions with human flora are under way. The current annotation guidelines are well adapted to capture pathogens and species, and also to expand with the advance of molecular microbiology and scientific knowledge.

3.3 LivingNER Multilingual Silver Standard

To foster the development of multilingual tools and generate systems not only for Spanish but also for content in English and various Romance languages, we have developed the annotated (and normalized to NCBI Taxonomy) LivingNER corpus in 7 languages: English, Portuguese, Galician, Catalan, Italian, French, and Romanian. The overview statistics of the Silver Standard are shown in Table 2. We refer to the DisTEMIST overview paper (Miranda-Escalada et al., 2022) for a complete description of the generation process since it is equivalent to that corpus. Find the Multilingual Silver Standard at Zenodo (doi.org/10.5281/zenodo.6376662)

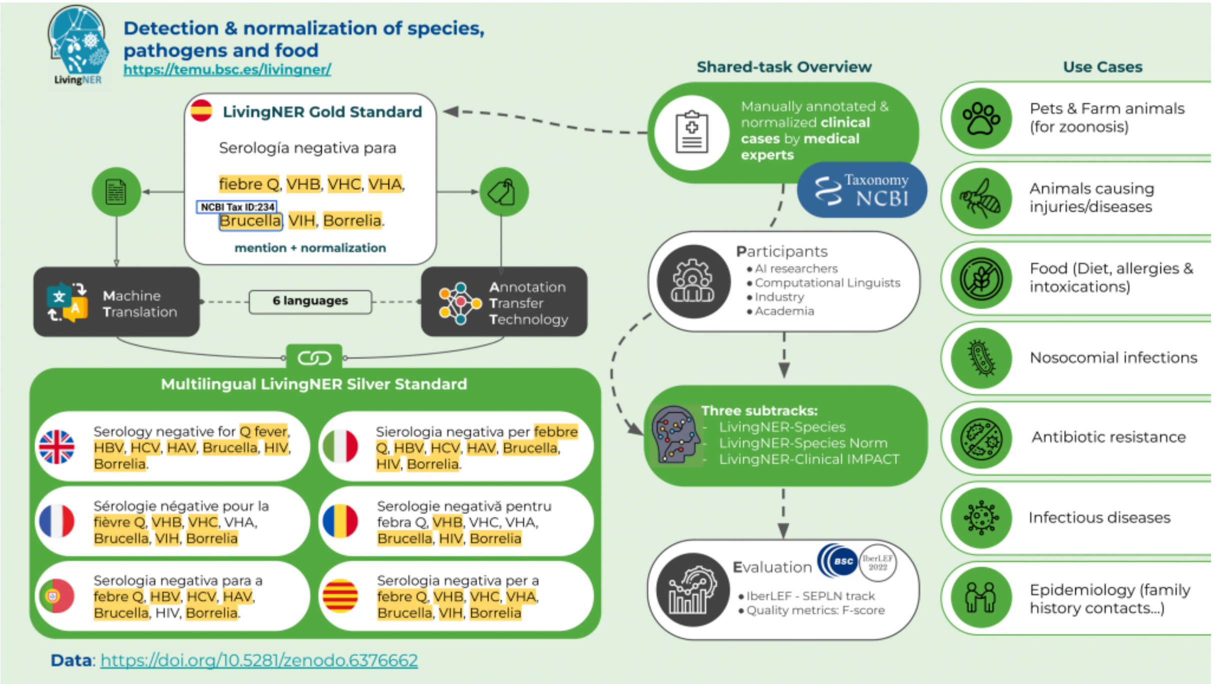
		Documents	Annotations	Unique NCBI Tax IDs	Sentences	Tokens
Catalan	**Training**	1000	14803	832	34173	642926
	Valid	500	6724	533	15073	297012
	Test	485	7709	548	15979	296124
English	**Training**	1000	13772	776	34430	624437
	Valid	500	6332	493	15180	287164
	Test	485	7225	513	16075	286419
French	**Training**	1000	12419	754	34552	697869
	Valid	500	5540	471	15225	322198
	Test	485	6428	498	16107	321766
Italian	**Training**	1000	12945	759	34373	649831
	Valid	500	5846	470	15130	299907
	Test	485	6703	499	16038	299470
Portuguese	**Training**	1000	12420	727	34330	641095
	Valid	500	5642	470	15143	295942
	Test	485	6738	490	16038	295587
Romanian	**Training**	1000	10522	699	34334	651595
	Valid	500	4799	427	15130	300773
	Test	485	5617	478	16029	300297
Galician	**Training**	1000	16633	875	34188	616216
	Valid	500	7319	555	15065	284023
	Test	485	7672	546	15983	283808

Table 2: LivingNER Multilingual Silver Standard corpus statistics.

3.4 LivingNER Terminology

It is the official NCBI Taxonomy FTP dump (ftp.ncbi.nlm.nih.gov/pub/taxonomy/) with the terms translated to Spanish by a Neural Machine Translator fine-tuned for the biomedical domain. It is a tab-separated file with the following columns: *tax_id* (the NCBI Taxonomy ID of node associated with this name), *name_txt* (the NCBI Taxonomy name), *unique name* (the unique variant of this name if the name is not unique), *name class* (synonym, common name, scientific name, ...), *Spanish name* (the NCBI Taxonomy name in Spanish).

Besides, we have added the following terms: 2560602 (Mumps orthorubulavirus), 2560526 (Human orthorubulavirus 4),


Detection & normalization of species, pathogens and food
https://temu.bsc.es/livingner/

LivingNER Gold Standard
Serología negativa para fiebre Q, VHB, VHC, VHA, Brucella, VIH, Borrelia.
NCBI Tax ID:234
mention + normalization

Machine Translation
6 languages
Annotation Transfer Technology

Multilingual LivingNER Silver Standard
Serology negative for Q fever, HBV, HCV, HAV, Brucella, HIV, Borrelia.
Sierologia negativa per febbre Q, HBV, HCV, HAV, Brucella, HIV, Borrelia.
Sérologie négative pour la fièvre Q, VHB, VHC, VHA, Brucella, VIH, Borrelia
Serologie negativă pentru febra Q, VHB, VHC, VHA, Brucella, HIV, Borrelia
Serologia negativa para a febre Q, HBV, HCV, HAV, Brucella, HIV, Borrelia.
Serologia negativa per a febre Q, VHB, VHC, VHA, Brucella, VIH, Borrelia

Data: https://doi.org/10.5281/zenodo.6376662

Shared-task Overview
Manually annotated & normalized clinical cases by medical experts
Taxonomy NCBI

Participants
- AI researchers
- Computational Linguists
- Industry
- Academia

Three subtracks:
- LivingNER-Species
- LivingNER-Species Norm
- LivingNER-Clinical IMPACT

Evaluation
- IberLEF - SEPLN track
- Quality metrics: F-score

Use Cases
Pets & Farm animals (for zoonosis)
Animals causing injuries/diseases
Food (Diet, allergies & intoxications)
Nosocomial infections
Antibiotic resistance
Infectious diseases
Epidemiology (family history contacts...)

Figure 5: LivingNER multilingual corpus overview.

2847144 (hepatitis C virus genotype 1a), _NOCODE_ (out of NCBI Taxonomy scope). The first three were added because they appear in the LivingNER corpus, and are present in the browser version of NCBI Taxonomy. The last one (_NOCODE_) is added to identify terms in the LivingNER corpus that are not present in the NCBI Taxonomy.

The terminology is available at Zenodo (doi.org/10.5281/zenodo.6390506).

4 Results

4.1 Participation Overview

The community has shown an active interest in LivingNER. There were 56 teams registered in LivingNER, and 20 successfully submitted their system results, totaling 62 submissions. 20 teams participated in LivingNER-SPECIES NER [41 runs], 8 also submitted their system predictions for LivingNER-SPECIES Norm [15 runs], and 5 did it for the LivingNER Clinical Impact track [6 runs]. Besides, as Table 3 shows, participants belonged to institutions (industry or academia) from different countries, including Spain, Romania, China and México.

4.2 System Results

Table 5 shows the best-run results by all teams for subtasks LivingNER-Species NER and LivingNER-Species Norm. In LivingNER-Species NER, the Vicomtech NLP team obtained the highest micro-average F1-score, 0.951. Team RACAI F1-score was almost tied with Vicomtech (0.9503), and it reached the highest precision (0.9622) and recall (0.9439) in different submissions.

In LivingNER-Species Norm, the highest F1-score (0.9304) and recall (0.9234) were obtained once again by the Vicomtech NLP team. The highest precision was obtained by the ClaC team (0.9641).

Table 4 contains the best-run results of the third subtask, LivingNER-Clinical Impact. In this case, participants had to classify the test set documents into four categories and include the NCBI Taxonomy codes justifying the classification. Results were computed for the document classification task and the document classification + code justification. The baseline system was available for the first task (document classification), and none of the participant teams outperformed it. We discuss this in the Discussion section. We must outline that only 4 test set documents were positive Nosocomial documents. Therefore, the results for this fourth classification axis are challenging to interpret.

Finally, the complete results of all runs, plus the disaggregated results

Team Name	Affiliation	Country	Tasks	Ref.	Tool URL
Vicomtech NLP	Vicomtech	Spain	NE/No/C	(Zotova et al., 2022)	–
racai	Research Institute for Artificial Intelligence ”Mihai Draganescu”	Romania	NE	(Avram, Mitrofan, and Pais, 2022)	–
READ-Biomed	RMIT University	Australia	NE	(Jimeno Yepes and Verspoor, 2022)	–
SINAI	Universidad de Jaén	Spain	NE/No/C	(Chizhikova et al., 2022)	–
plncmm	CMM, University of Chile	Chile	NE/No/C	(Rojas et al., 2022)	(plncmm, 2022)
Sumam Francis	KU Leuven	Belgium	NE	(Francis and Moens, 2022)	–
Clac	Concordia University	Canada	NE/No	(Bagherzadeh, Verma, and Bergler, 2022)	–
john_snow_labs	John Snow Labs	USA	NE	(Kocaman et al., 2022)	–
avacaondata	IIC (ADIC)	Spain	NE/No/C	(Vaca, 2022)	–
Pumas	Universidad Nacional Autónoma	México	NE/No/C	(del Moral et al., 2022)	–
IAM	University of Bordeaux	France	NE	(Cossin, Diallo, and Jouhet, 2022)	(IAM, 2022)
IGES	IGES Institut GmbH	Germany	NE/No	(Chapman, Schwarz, and Hässler, 2022)	–
NLP-CIC-WFU	Instituto Politécnico Nacional Wake Forest University	México & USA	NE/No	(Tamayo, Burgos, and Gelbukh, 2022)	(NLP-CIC-WFU, 2022)
Vitor	Universidade Federal do Rio de Janeiro	Brasil	NE	–	–
zzz	Yunnan University	China	NE	(Zhu and Wang, 2022)	(zzz, 2022)
Kformer-OEG	Universidad Politécnica de Madrid	Spain	NE	–	–
Mark	–	–	NE	(Hanjie and Xiaobing, 2022)	(Mark, 2022)
Han	Yunnan University	China	NE	(Han and Ding, 2022)	(tutorial, 2022)
Sapphire	–	–	NE	–	–
boun-ner	Bogazici University	Turkey	NE	–	–

Table 3: LivingNER team overview. In the Tasks column, NE stands for LivingNER-Species NER, No for LivingNER-Species Norma and C for LivingNER-Clinical Impact.

	Pets and farm animals			Animals causing injuries			Food species			Nosocomial entities		
Team Name	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**
LivingNER-Clinical Impact with codes												
Vicomtech	0	0	0	.0006	.125	.0012	.0088	.1154	.0164	0	0	0
SINAI	0	0	0	0	0	0	0	0	0	0	0	0
plncmm	.0317	.3636	.0584	0	0	0	.02	.3846	.038	0	0	0
avacaondata	0	0	0	0	0	0	0	0	0	0	0	0
Pumas	.024	.25	.0438	0	0	0	.0211	.2692	.0391	0	0	0
LivingNER-Clinical Impact												
Vicomtech	.0326	.25	.0577	.0058	.5	.0115	.0235	.3077	.0437	.0016	.75	.0032
SINAI	0	0	0	0	0	0	0	0	0	0	0	0
plncmm	.0397	.4167	.0725	.0282	.5	.0533	.0479	.9231	.0911	.006	.5	.0118
avacaondata	0	0	0	.0021	.125	.0041	0	0	0	0	0	0
Pumas	.024	.25	.0438	.0167	.25	.0312	.0211	.2692	.0391	0	0	0
PathoTagIt-Base	1	.1667	.2857	.032	1	.062	.8	.9231	.8571	.0513	.5	.093

Table 4: Results of LivingNER-Clinical Impact systems. MiP, MiR and MiF stands for micro-averaged precision, recall and F1-score.

by label (HUMAN and SPECIES), are published on a dedicated webpage (temu.bsc.es/livingner/results/).

4.3 Methodologies

Table 5 briefly describe the methodologies used by LivingNER participants, and for an in-depth description, we refer to their scientific articles, listed in Table 3. We have observed that the most successful approaches to LivingNER-Species NER included non-standard fine-tuning of pre-trained transformer-based language models. Typically, these language models are domain and language-specific, such as bsc-bio-es RoBERTa (Carrino et al., 2021), employed by teams READ-Biomed and SINAI, among others; or cross-lingual, such as XLM-RoBERTa (Conneau et al., 2019), chosen by team racai. The highest-scoring participant of LivingNER-Species NER, Vicomtech, has fine-tuned a transformer-based language model using a sliding windows technique that avoids hard, meaningless segmentation cuts that typically occur in these scenarios (Zotova et al., 2022).

In LivingNER-Species Norm, participants with the highest scores used a robust NER system to detect the species mentioned. And they were mapped to NCBI Taxonomy using traditional approaches such as string matching using Levenshtein distance and setting a heuristic cutoff. Additionally, other participants used, for instance, word embeddings similarity (Pumas) or TF-IDF matching (SINAI).

Finally, in LivingNER-Clinical Impact, the most successful approach has been the baseline: to train a simple NER system to recognize the entities of interest and label as positive any document with a detected named entity.

	SPECIES NER				SPECIES Norm			
Team Name	MiP	MiR	MiF	Description	MiP	MiR	MiF	Description
Vicomtech NLP	.9583	.9438	**.951**	sophisticated fine-tune transformer model	0.9376	**0.9234**	**0.9304**	Semantic Text Search approaches
racai	.9569	**.9439**	.9503	fine-tune XLM-RoBERTa with lateral inibitory layer	-	-	-	-
READ-Biomed	.954	.9411	.9475	Fine-tune RoBERTa (bsc-bio-es)	-	-	-	-
SINAI	.9571	.9346	.9457	fine-tune RoBERTa (roberta-base-bne, bsc-bio-es & roberta-biomedical-clinical-es)	.8733	.8527	.8629	character-level TF-IDF matching and string matching w. Levenshtein distance
plncmm	.9455	.9373	.9414	Fine-tune RoBERTa (bsc-bio-es) w. FLERT	.9139	.906	.9099	string matching w. Levenshtein distance
Sumam Francis	.9443	.9307	.9375	Fine-tune BERT (BETO) pre-trained w. contrastive loss	-	-	-	-
Clac	.9385	.9256	.932	mi-RIM model	.9495	.891	.9193	string matching w. Levenshtein distance
john_snow_labs	.916	.9327	.9243	Bi-LSTM-CNN-Char & BertForTokenClassification	-	-	-	
avacaondata	.9228	.908	.9153	Domain adaptation of MarIA-Large	.512	.4799	.4954	-
Pumas	.9284	.8899	.9087	fine-tune RoBERTa (bsc-bio-es)	.9389	.8075	.8682	word embedding similarity
IAM	.9209	.8733	.8965	Complex dictionary lookup	-	-	-	-
IGES	.9112	.8638	.8869	SAPBert-XLMR + CRF	.8979	.8512	.874	FAISS indexes containing encoded synonyms
NLP-CIC-WFU	.8303	.8704	.8499	fine-tune mBERT & post-processing rules	.7768	.8143	.7951	dictionary lookup
Vitor	.9492	.5634	.7071	-	-	-	-	-
zzz	.8012	.6138	.6951	fine-tune BERT+BiLSTM	-	-	-	-
Kformer-OEG	.7306	.6057	.6623	-	-	-	-	-
Mark *pw	.8214	.6145	.703	BERT(BETO)+BiGRU+ CRF + adversarial learning	-	-	-	-
Han *pw	.5399	.1965	.2881	fine-tune BERT (BETO)	-	-	-	-
Sapphire	.6875	.0149	.0291	-	-	-	-	-
Boun-ner	0.126	0.078	0.0963	fine-tune BERT	-	-	-	-
PathoTagIt-Base	0.9461	0.8507	0.8958	Section 2.5	-	-	-	-

Table 5: Results of LivingNER systems, subtasks SPECIES NER and SPECIES Norm. **pw* means post-workshop submissions. MiP, MiR and MiF stands for micro-averaged precision, recall and F1-score.

4.4 LivingNER Spanish Silver Standard

The LivingNER test set was released together with a background set: an additional collection of 13,000 clinical case documents from various medical disciplines, all Spanish. The background set helps examine whether systems could scale to more extensive data collections and avoid manual annotation correction. Participants have generated automatic predictions for the test and the background set, although they were only evaluated on the test set predictions in the three subtasks.

Therefore, the background set predictions include automatic mention annotations (LivingNER-Species NER predictions), normalized to NCBI Taxonomy (LivingNER-Species Norm predictions) and document classifications with evidence (LivingNER-Clinical Impact predictions). The background set predictions from all participants will be harmonized and constitute the LivingNER Spanish Silver Standard corpus, similar to the CALBC initiative (Rebholz-Schuhmann et al., 2010), to the Cantemist (Miranda-Escalada, Farré, and Krallinger, 2020), CodiEsp (Miranda-Escalada et al., 2020), MESINESP2021 (Gasco et al., 2021), ProfNER (Miranda-Escalada et al., 2021), and PharmaCoNER (Gonzalez-Agirre et al., 2019) shared tasks.

Considering the large precision and recall of most LivingNER systems, the LivingNER Spanish Silver Standard will be a high-quality collection of annotated, normalized, and classified clinical documents in Spanish. Besides, it will serve to foster the development of species recognition and linking resources, as well as to generate more annotated data. The LivingNER Spanish Silver Standard will be released on the Zenodo Medical NLP community.

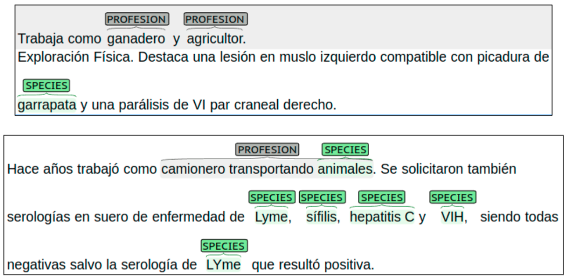
5 Discussion

There is a clear need to generate, extend and provide access to multilingual terminologies and glossaries for the biomedical domain. Providing access to bilingual medical glossaries such as MeSpEN, curated for species information and other clinical entities, might be helpful to foster exploitation for multilingual semantic annotation efforts (Villegas et al., 2018).

In this direction, the LivingNER initiative pioneers to structure the species information in clinical documents written in languages other than English. To foster the development of species NER and linking resources, we have released the LivingNER corpus: the first Gold Standard corpus of Spanish clinical documents with species mentions, manually mapped to the NCBI Taxonomy.

The LivingNER corpus was created following strict annotation guidelines that are made public to allow the corpus extension and adaptation to other languages or domains. It contains HUMAN annotations (a building block to collect relevant information from patient history, hereditary diseases, etc.) and SPECIES annotations. The latter is essential for diverse clinical applications, such as epidemiology.

To enhance the interoperability between different data sources, and taking into account (1) multilingual scenarios, (2) the multilingual potential of species mentions, and (3) the general lack of annotated data in other languages, we have released the LivingNER Multilingual Corpus. It contains the LivingNER corpus documents, translated to 7 languages (English, French, Italian, Portuguese, Catalan, Romanian, and Galician), and automatically generated species mention annotations mapped to NCBI Taxonomy.



Trabaja como ganadero y agricultor.
Exploración Física. Destaca una lesión en muslo izquierdo compatible con picadura de garrapata y una parálisis de VI par craneal derecho.

Hace años trabajó como camionero transportando animales. Se solicitaron también serologías en suero de enfermedad de Lyme, sífilis, hepatitis C y VIH, siendo todas negativas salvo la serología de LYme que resultó positiva.

Figure 6: Actual examples of annotated species mentions and automatically recognized profession mentions.

The resources and the task have generated considerable interest in the community. Participant teams have developed 62 competitive systems based on pre-trained transformer language models evaluated against the LivingNER corpus manual annotations. Additionally, they have generated automatic predictions for nearly 13,000 documents that will be harmonized to create the LivingNER Spanish Silver Standard.

These resources can be used to obtain actionable information from clinical narratives. An example would be linking the species with the text's occupational information to fine-tune the work-related disease statistics. This linking is seen in Figure 6, in which Gold Standard SPECIES annotations are combined with an automatic system that recognizes profession mentions (trained with MEDDOPROF (Lima-López et al., 2021) corpus).

As future directions, we plan to generate more granular annotations for the HUMAN mentions that are needed for real-world applications. In addition, the third subtask on Clinical Impact applications lacked enough training and test data, and we plan to correct this issue in the future. Finally, the Multilingual Silver Standard will be manually reviewed to generate manually-generated parallel annotations in eight languages.

Acknowledgements

We acknowledge the Encargo of Plan TL (SEDIA) to BSC for funding and the scientific committee for their guidance and help. Due to the relevance of species for biomaterials and implants, this project is supported by the European Union's Horizon Europe Coordination & Support Action under Grant Agreement No 101058779. We acknowledge the support from the AI4PROFHEALTH project (PID2020-119266RA-I00). We thank the organization of IberLEF and SEPLN, and Bitac for collaboration during the corpus and guidelines construction.

References

Armengol-Estapé, J., F. Soares, M. Marimon, and M. Krallinger. 2019. Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics & informatics*, 17(2).

Avram, A.-M., M. Mitrofan, and V. Pais. 2022. Species entity recognition using a neural inhibitory mechanism.

Bagherzadeh, P., H. Verma, and S. Bergler. 2022. Multi-input rim for named-entity recognition in spanish clinical reports.

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.

Chapman, K., M. Schwarz, and B. Häussler. 2022. Multilingual medical entity recognition and cross-lingual zero-shot linking with faiss.

Chizhikova, M., J. Collado-Montañez, P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, and M. T. Martín-Valdivia. 2022. Sinai at livingner shared task 2022: Species mention recognition and normalization using transfer learning and string matching techniques.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cossin, S., G. Diallo, and V. Jouhet. 2022. Iam at iberlef 2022: Ner of species mentions.

del Moral, R., J. Reyes-Aguillón, O. Ramos-Flores, H. Gómez-Adorno, and G. Bel-Enguix. 2022. Species mention entity recognition, linking and classification using roberta in combination with spanish medical embeddings.

Federhen, S. 2012. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Francis, S. and M.-F. Moens. 2022. Task-aware contrastive pre-training for spanish named entity recognition in livingner challenge.

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

Gerner, M., G. Nenadic, and C. M. Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

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

Han, S. and H. Ding. 2022. Named entity recognition for livingner-species based on bert and span detection.

Hanjie, M. and Z. Xiaobing. 2022. Clinical text entity recognition based on pretrained model and bigru-crf.

IAM. 2022. Iamsystem. https://github.com/scossin/IAMsystem.

Jimeno Yepes, A. and K. Verspoor. 2022. The read-biomed team in livingner task 1 (2022): Adaptation of an english annotation system to spanish.

Kocaman, V., G. Pirge, B. Polat, and D. Talby. 2022. Biomedical named entity recognition in eight languages with zero code changes.

Krallinger, M., F. Leitner, and A. Valencia. 2010. Analysis of biological processes and diseases using text mining approaches. *Bioinformatics Methods in Clinical Research*, pages 341–382.

Lima-López, S., E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, and M. Krallinger. 2021. Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Procesamiento del Lenguaje Natural*, 67:243–256.

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

Mark. 2022. 33da. https://github.com/33Da/.

Miranda-Escalada, A., E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. *IberLEF@ SEPLN*, pages 303–323.

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

Miranda-Escalada, A., L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, and M. Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.

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

NLP-CIC-WFU. 2022. Nlp-cic-wfu- contribution-to-livingner-shared-task-2022. https://github.com/ajtamayoh/NLP-CIC-WFU- Contribution-to-LivingNER-shared-task-2022.

Pafilis, E., S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

plncmm. 2022. Livingner. https://github.com/maranedah/LivingNER.

Pyysalo, S., T. Ohta, R. Rak, D. Sullivan, C. Mao, C. Wang, B. Sobral, J. Tsujii, and S. Ananiadou. 2011. Overview of the infectious diseases (id) task of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 26–35.

Rebholz-Schuhmann, D., A. J. J. Yepes, E. M. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn. 2010. Calbc silver standard corpus. *Journal of bioinformatics and computational biology*, 8(01):163–179.

Rojas, M., J. Barros, M. Araneda, and J. Dunstan. 2022. Flert-matcher: A two-step approach for clinical named entity recognition and normalization.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. 2020. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020.

Soares, F., M. Villegas, A. Gonzalez-Agirre, M. Krallinger, and J. 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, June. Association for Computational Linguistics.

Tamayo, A., D. A. Burgos, and A. Gelbukh. 2022. Partner: Paragraph tuning for named entity recognition on clinical cases in spanish using mbert + rules.

tutorial. 2022. ner. https://github.com/songhan123123/ner.

Vaca, A. 2022. Named entity recognition for humans and species with domain-specific and domain-adapted transformer models.

Villegas, M., A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, and M. Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: multilingual biomedical text processing.*

Zhu, Z. and L. Wang. 2022. Bert-bilstm model for entity recognition in clinical text.

Zotova, E., A. García-Pablos, N. Perez, P. Turón, and M. Cuadros. 2022. Vicomtech at livingner 2022.

zzz. 2022. 2251821381. https://github.com/2251821381.