Spanish-language Platform for Drug-Disease Evidence Search based on Scientific Articles

Elvira Amador-Domínguez[1,2,*], Carlos Badenes-Olmedo[1,2]

[1]Ontology Engineering Group, Departamento de Inteligencia Artificial, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain

[2]Departamento de Sistemas Informáticos, ETSI Sistemas Informáticos, Universidad Politécnica de Madrid, 28031 Madrid, Spain

Abstract

The COVID-19 pandemic propelled the development of several Natual Language Processing (NLP) resources. These resources, however, are mostly developed exclusively in English, and lack accessibility for non-expert users, such as practitioners. This paper presents a platform for the analysis of clinical documents in Spanish, especially those related to COVID-19. This platform takes text in Spanish as input and comprises three different NLP modules: a named entity recognition module capable of detecting diseases and chemicals, a term linking module that links the detected entities with existing medical terminologies, and an evidence search module that retrieves scientific articles evidencing the relationship between pairs of entities, both diseases and medications. A series of experiments using the SPACCC corpus were conducted to assess the efficiency of the platform and the quality of the results.

Keywords

Named entity recognition, term extraction, term linking, biomedical text processing

1. Introduction

The COVID-19 pandemic propelled the release of considerable amounts of medical documents that, up to that time, were largely inaccessible to researchers [1, 2]. This phenomenon noticeably impacted the area of Natural Language Processing (NLP), leading to an increase in the development of models and data sets that were specifically targeted in COVID-19 documents. In 2024, the search "COVID" in HuggingFace[1] returns over 100 dataset results and over 500 language models. Considering that HuggingFace has around 500k pre-trained models available, COVID-19 related models represent around 10% of the total. This value is quite remarkable, especially considering that there may be models related to COVID-19 that do not include the term "COVID" in their name and are therefore not retrieved by search.

However, most of the available resources present two main shortcomings. First, most resources, both datasets and models, are available only in English [3]. Most of the data collected in the available datasets are gathered from two main sources: research papers and Twitter posts. Research papers are mostly published in English, and therefore data extracted from this source will also be in that same language. This is also the case with Twitter posts, in which, while there is a wider variety of languages, most available datasets are exclusively in English. Subsequently, most COVID-19 related NLP models are in English, since they are trained on datasets in this language. Therefore, there is a scarcity of COVID-19 NLP models available in Spanish. The scarcity of models in Spanish can hinder the exploitation of textual resources in this language and, as a result, limits the generation of new models and datasets. Providing tools capable of processing scientific resources in Spanish may not only be useful for the NLP community, but also may increase the impact of those resources.

Second, while these models are intended for their use in clinical practice, their use is not intuitive to practitioners. For non-expert users, using pre-trained NLP models can be a difficult task. Doting NLP models with an intuitive user interface based on natural language queries may help to approach these models to practitioners, subsequently enhancing their use.

This paper addresses the aforementioned issues, presenting a biomedical NLP platform that facilitates the exploration of large collections of scientific articles with clinical information to extract evidence of drug-disease, drug-drug, and disease-disease interactions. This platform includes several NLP models, including a Named Entity Recognition (NER) model to automatically detect diseases and chemicals within the text, as well as a term-linking module to retrieve additional information on the detected terms. In addition, the platform includes an evidence module to retrieve existing research documents that mention or relate selected terms. Moreover, since the medical field is highly expert-oriented, the platform follows a human-in-the-loop approach, enabling users to select and find information on medical terms that may not be correctly detected by the NER module.

The remainder of this paper is structured as follows. Section 2 presents the related work, Section 3 introduces the developed platform, describing its composing modules. Section 4 presents the experiments conducted on the platform to address its efficiency and results, and finally, Section 5 draws conclusions and future research lines.

2. Related Works

In the context of NLP, we can distinguish two types of resources developed during the COVID-19 pandemic: language models and datasets. The work of Shuja et al. [3] presents a comprehensive review of the different data sets, classified by task, that were developed during the pandemic. This work analyzes not only textual datasets but also speech and medical image datasets. One of the first corpus available was CORD-19 [1], released in 2020 by the Allen Institute for Artificial Intelligence. This corpus contained research articles in English extracted from different open repositories,

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.

*Corresponding author.

elvira.amador@upm.es (E. Amador-Domínguez); carlos.badenes@upm.es (C. Badenes-Olmedo)

0000-0001-6838-1266 (E. Amador-Domínguez); 0000-0002-2753-9917 (C. Badenes-Olmedo)

[1]https://huggingface.co/

such as PubMed or bioRxiv. The corpus was first released in 2020, with a size of less than 1GB. By its last release, in 2022, it had a size of more than 18GB worth of COVID-19 related articles. Similarly, DisGeNet [2] provides an extensive dataset that labels diseases, genes, and variants within medical articles, providing evidence of their labeling. In addition to the benefits of academic sources, other works focused on the recollection of textual data from non-expert users. Lamsal et al. [4] present the COV19Tweets dataset, which contained more than 310 million tweets in English, intended for the sentiment analysis task. A geolocalized version of the dataset, named GeoCOV19Tweets, is also available [4]. Similarly, the COVIDSENTI dataset [5] provides more than 90,000 tweets collected in the early stages of the pandemic. Other works, such as Cheng et al. [6] and Depoux et al. [7] focused on recollection of rumors, fake news, and general misinformation for their further analysis.

Regarding the Spanish language, most of the available COVID-19 related datasets are extracted directly from social media sources. This is the case of the work by Yu et al. [8], which provides a dataset of tweets posted by two of the main Spanish newspapers during the pandemic. Similarly, Allés-Torrent et al. [9] present DHCovid, a curated Twitter corpus of digital conversations in the context of the pandemic.

Although there is no specific COVID-19 corpus available in Spanish, several biomedical corpus in this language have been released over the years. The most important corpus is SPACCC [10], which contains 1,000 examples of clinical case reports. This corpus covers a variety of medical areas, such as oncology, urology, or pneumology. This corpus has served as the base for more targeted datasets, such as PharmacoNER [11], a version of SPACCC specifically targeted at the identification of pharmacological substances in texts. Similarly, CANTEMIST [12] presents a set of oncological clinical reports focused on the identification of concepts related to tumor morphology. More recently, DISTEMIST [13] released a set of manually annotated clinical cases to identify clinical terms, including pharmacological substances, diseases, and symptoms.

These Spanish clinical data sets served as a scaffold for the development of several NLP models, each focused on different tasks. The most prominent work in this area is that conducted by the Barcelona Supercomputing Center [14], providing a set of pretrained models in Spanish for different tasks and datasets. More recently, the BioLORD [15] model was released, offering a multilingual solution for different NLP tasks, such as answering questions and recognizing named entities.

3. Description of the platform

This paper presents a platform designed for the extraction of evidence in Spanish from COVID-19 documents in English. The proposed platform [2] addresses the main shortcomings outlined in Section 1. First, it is devised for its use by non-expert uses, especially for practitioners, as it allows queries to be made in natural language. Therefore, it presents a centralized resource that comprises three NLP modules: a named entity recognition module, a term linking module, and an evidence retrieval module. The interaction between these modules is seamless to the user, who is only required to introduce an input text and make simple setting choices to visualize the output of the models without changing or interacting directly with the models. Secondly, the platform has been created in Spanish, both for the interface and for the language models. The platform extends previous work aimed at facilitating the exploration of large collections of clinical scientific articles, such as the EBOCA [16] ontology for evidence retrieval or the term linking approach presented in [17].

Figure 1 shows the workflow of the proposed platform. Once the user introduces the input text, the named entity recognition module finds and highlights all diseases and chemicals detected within the text. Diseases are highlighted in purple, while chemicals are highlighted in blue. As stated above, the user can manually select terms within the text and include them in the processing pipeline. In the sample, the user selects the term "respuesta febril" ("febrile response" in English), highlighted in gray. The platform then asks the user to denote whether the selected term represents a disease or a chemical and adds it to the list of detected entities. The term linking module then provides extended information on the detected entities, such as their identification codes on different medical taxonomies, such as MeSH or ATC.

A notable feature of the proposed platform is that it enables the inclusion of expert knowledge in real time. In this case, if there are entities related to diseases or chemicals that have not been detected by the NER model, they can be manually identified by the expert. Manually identified entities are added to the list of recognized terms and subsequently processed by the remaining modules.

The platform is devised for use both by non-expert users and by medical practitioners. In the second case, the platform can be used as an assistant, since it can easily retrieve external references of medical terms, which can then be queried on the corresponding medical taxonomies for further information. Moreover, thanks to the evidence module, the tool can be used to assess whether or not a certain chemical can be used to treat a certain disease. Given a pair of disease-drug, the evidence module returns scientific evidence relating both elements, which either supports or discourages the use of that chemical to treat the given disease.

3.1. Named-Entity Recognition Module

The first module comprises the Named Entity Recognition task. As stated in Section 2, there is a lack of NER models in Spanish trained specifically to detect COVID-19 related terms. More specifically, there is no benchmark NER model in Spanish dedicated specifically to the detection of diseases. In order to address this issue, a fine-tuned version of the BSC biomedical model is generated. This model is focused on disease detection and is trained using the DISTEMIST [13] corpus. The trained model is available in HuggingFace[3].

For drug detection, the BSC Pharmaconer model is used, as it provides accurate results. In the first NER strategy, which will be referred to as the default strategy, these two models work in parallel with the input Spanish text, providing a set of detected diseases and chemicals in Spanish as output. A second NER strategy is considered, following a translation-based approach. Since most COVID-19 related NER models are in English and they offer a remarkable performance, a second processing pipeline is devised in which the original Spanish text is translated into English

[2] https://drugs4covid.oeg.fi.upm.es/platform_es

[3] https://huggingface.co/oeg/BioNER-es

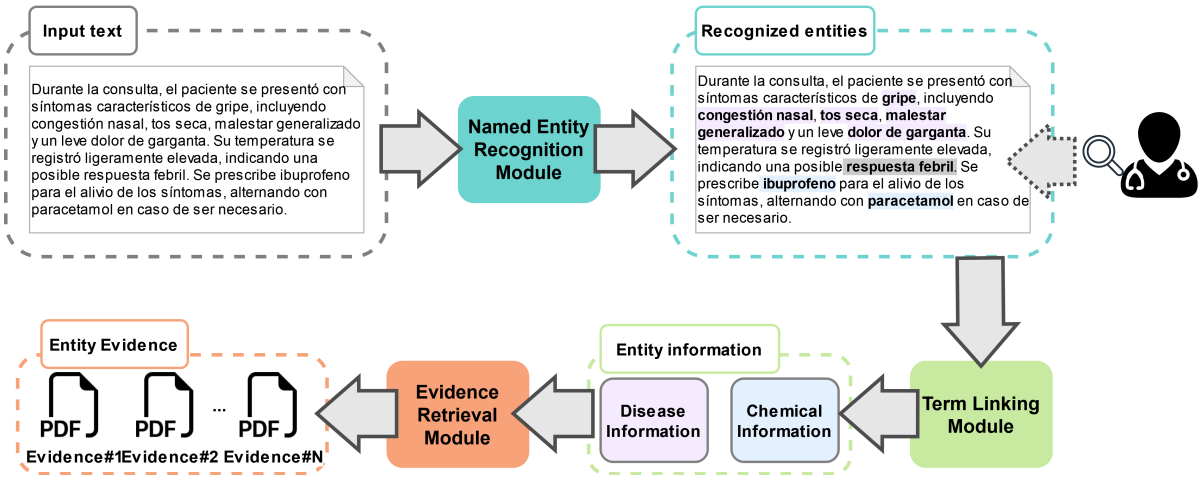


Figure 1: Overview on the workflow of the platform.

using the GoogleTranslator library, and processed by an existing pre-trained NER model. In this case, the selected model is the BC5CDR NER model. This model is contained within the Spacy library, and it accurately detects diseases and chemicals. The results provided by the NER model are then translated back into Spanish, making this process completely opaque to the user.

By default, the first strategy is followed, but the user can specify whether to use the default approach or the translation-based one.

3.2. Term Linking Module

Once biomedical entities are detected, the term linking module retrieves additional information related to each of them. The previous term linking resources developed within the DRUGS4COVID++ project were developed entirely in English. Therefore, a mechanism to precisely find the equivalent English term for a given entity in Spanish is required in order to recover its additional information from the already existing repositories. Some chemicals receive different commercial names according to their country of distribution. For example, the chemical "acetaminophen" is sold under the commercial name "paracetamol" in Spain and "tylenol" in the USA. Both terms refer to the same chemical, and therefore, they should be linked to the same term and return the same information.

The same translation strategy used for NER (Section 3.1) can be followed. However, literal translation may not be optimal in this case, as context is essential to correctly identify medical terms. An example of this is the term "callo", which refers to a hardening of the skin caused by friction. Without context, this term literally translates to "callous", which is not related to the medical domain.

Therefore, and to avoid mistranslated terms, the medical terminology SNOMED-CT is used as a pivotal element between the two languages. SNOMED-CT provides an official version in different languages, including English and Spanish. In order to maintain consistency between versions, each concept is assigned a unique identifier, which is identical for each version. This is exemplified in Figure 2, where the terms "flu" and its equivalent term in Spanish "gripe" are searched, and both relate to the same SNOMED-CT identifier.

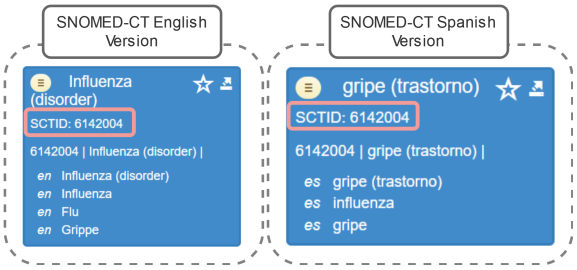


Figure 2: Example of SNOMED-CT identifier across versions.

Therefore, the Spanish term is queried into the Spanish version of SNOMED-CT, retrieving its SNOMED ID. This ID is then used to query the English version and retrieve the correct translation of the concept. The concept is then queried in our database [17], which includes the different identifiers for each medical concept in different terminologies, as well as the existing synonyms. In the case of diseases, additional information includes its CUI, MeSH ID, and ICD10, as well as its semantic type. For chemicals, its CID, MeSH ID, CID, ATC, and ATC levels are provided.

3.3. Evidence Retrieval Module

Once the terms have been identified and presented to the user, the platform provides the option of finding scientific evidence related to a term. The evidence is modeled in the form of a knowledge graph (KG), which uses the EBOCA [16] ontology as its basis. The EBOCA graph has two purposes. First, it describes the entities in the DISNET database and relates them with their corresponding scientific resources. Secondly, it also models metadata on evidences of associations between concepts. All evidence is extracted from curated sources, mostly research papers belonging to the CORD-19 corpus.

For each identified (or manually specified) entity, based on its disease/chemical classification, the system automatically enables a feature for user interaction. To access the evidence related to an entity, users initiate a process that explores the EBOCA KG, fetching the top ten results related to the term. These results are paragraphs retrieved from scientific resources in which the selected term is mentioned.

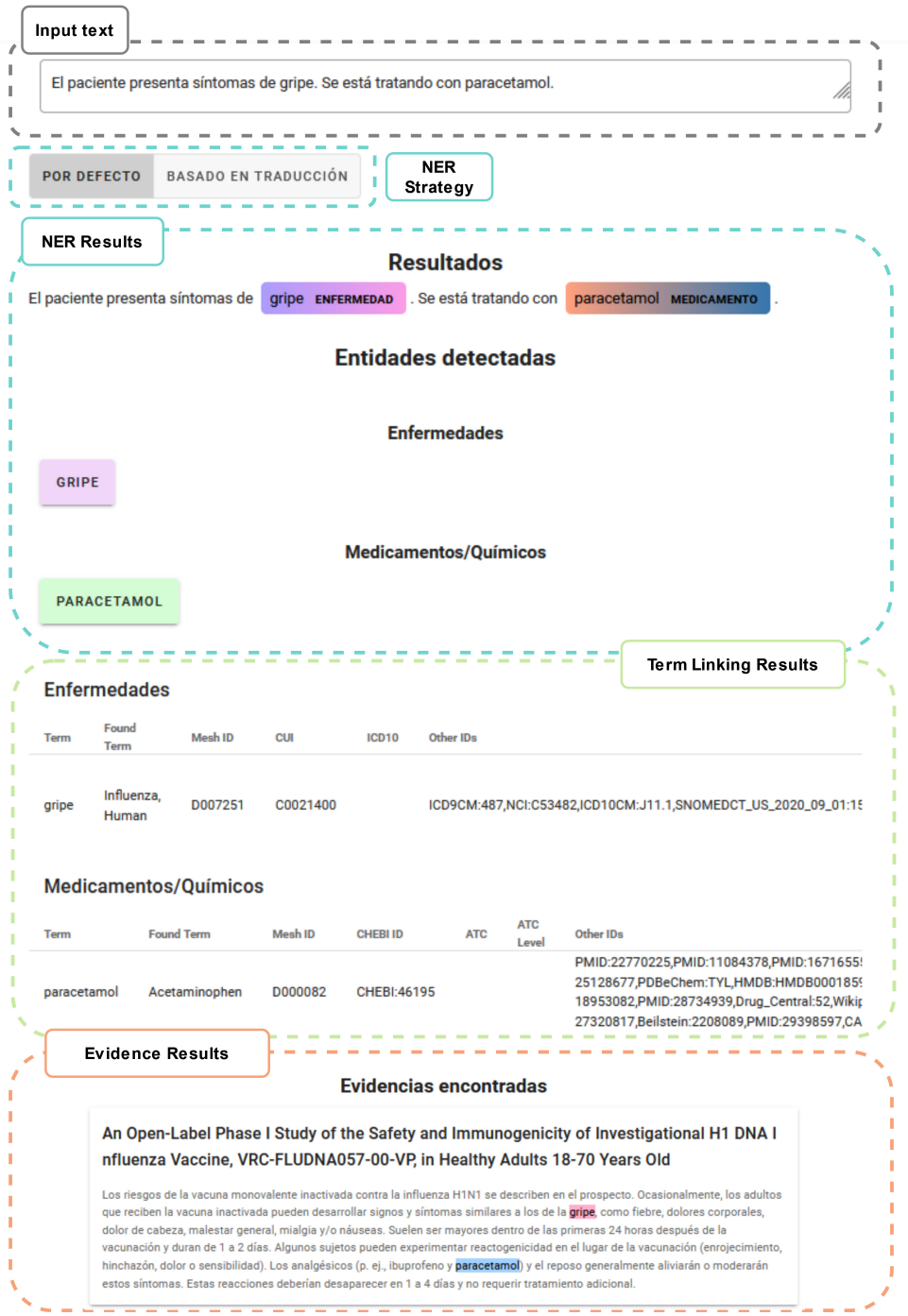


Figure 3: Overview of the platform.

The evidence content is presented in Spanish to improve comprehension, with occurrences of the relevant term emphasized to aid in reading.

In the EBOCA graph, the evidence involving compounds is organized in pairs, allowing only two terms to be chosen simultaneously to find related evidence. By selecting two terms, the system automatically retrieves the scientific literature that connects both entities from the graph by performing the appropriate SPARQL query. In these cases, the retrieved evidence is composed of paragraphs in which both entities are related.

3.4. Design of the platform

As previously stated, the platform is intended for use by individuals in the clinical domain who do not have experience with natural language processing (NLP). Therefore, it is essential to maintain simplicity, keeping the user oblivious to the internal working of the system while maintaining efficiency.

Figure 3 provides an overview of the visual aspect of the platform. Since the platform is self-contained, all modules are centralized on the same page, appearing dynamically following the workflow described in Figure 1. Therefore, the user first introduces a piece of text and selects the NER

Femenino de 42 años. Manifiesta 8 meses con herpes recurrente en boca. Ingresa por padecimiento de 2 meses con tos seca en accesos y disnea rápidamente progresiva. Además dolor torácico bilateral, fiebre hasta 39oC y pérdida de peso de 14kg. La recurrencia de la lesión herpética le ocasionaba disfagia y odinofagia. Al examen físico presentaba placas blanquecinas en orofaringe; la exploración del tórax, con disminución del ruido respiratorio y estertores finos. Al aire ambiente la saturación de oxígeno era del 86%. El reporte de la gasometría arterial con oxígeno suplementario al 70% fue: pH 7,30, pCO2 40,5mmHg, pO2 132mmHg, HCO3 19,5mmol/l, exceso de base -5,8mmol/l, saturación de oxígeno al 97,9%. Índice de oxigenación (IO) de 188. Los exámenes de laboratorio al ingreso destacan: linfopenia de 600células/mm3, Hb 11,8gr/dl, deshidrogenasa láctica de 971UI/l y albúmina 3,3gr/dl. La Rx de tórax presentaba opacidades bilaterales en parche con vidrio deslustrado y neumomediastino, por lo cual, en el diagnóstico diferencial se incluyó inmunosupresión asociada a VIH y neumonía por P. jirovecii (PJP). El análisis para VIH por ELISA fue POSITIVO, se confirmó por Western Blot. Se le realizó broncoscopia con biopsia transbronquial y lavado broncoalveolar (LBA). El estudio histopatológico se reporta en la figura 1. Recibió tratamiento con Trimetoprim/Sulfametoxasol y Prednisona en dosis de reducción por 21 días.
En el 7.o día de tratamiento presentó deterioro respiratorio y el IO desciende a 110, por lo cual ingresa a terapia intensiva en estado de choque y apoyo con ventilación mecánica invasiva. Al ingresar, los exámenes de laboratorio destacan leucocitos de 24,300células/mm3, Hb 10,8gr/dl, deshidrogenasa láctica 2033UI/l y albúmina de 2,26gr/dl. Se agrega Imipenem por sospecha de neumonía intrahospitalaria y luego de 12 días mejora la cifra leucocitaria a 5800células/mm3, Hb 8,7gr/dl, deshidrogenasa láctica 879UI/l, albúmina 2.41gr/dl e IO en 243.5, lográndose extubar. En las siguientes 24h, presenta hemoptisis masiva (volumen 250ml). Desciende la cifra de Hb a 6gr/dl, el IO disminuye a 106 y se apoya con ventilación mecánica invasiva. Por Swan-Ganz se mide una POAP de 10mmHg. La radiografía de tórax se muestra en la figura 2A. Se somete a LBA cuyo estudio de patología confirma la presencia de hemorragia alveolar reciente y activa. Además, por rt-PCR se documenta infección por CMV e inicia tratamiento con Ganciclovir 350mg/d durante 14 días.
Tiene buena evolución, mejora el IO hasta 277 retirándose de ventilación mecánica invasiva 9 días posteriores al evento. Luego de 37 días egresa a su domicilio. En el seguimiento, la cuenta de CD4 es de 109células/µl y la carga viral <40copias/µl.

Figure 4: Sample SPACCC document.

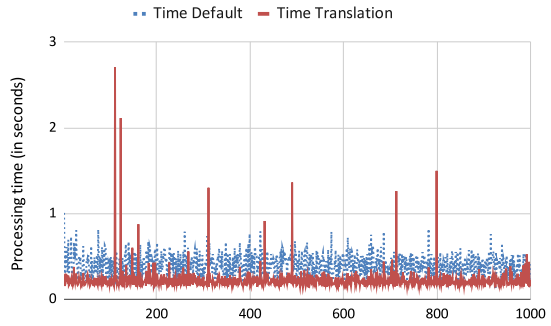


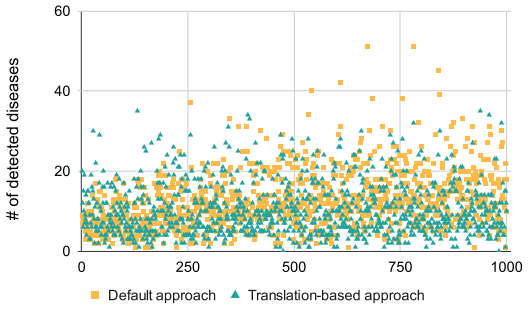
Figure 5: Processing time per strategy.

strategy to be followed. The system will then output the results of the NER module, providing a version of the text in which the detected entities are highlighted. The corresponding list of entities is also depicted, which can then be selected for evidence retrieval. Then, below the entities, the results of the term linking process are depicted, listing all codes and additional information related to the found entities. Finally, once the user selects one or more entities from the list, the corresponding evidence is presented.

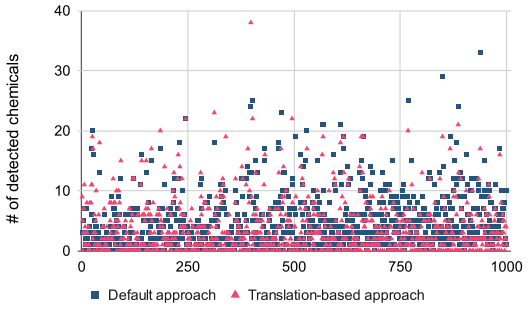
4. Experimentation

According to the design of the workflow, the NER module can be perceived as the cornerstone of the system. The process begins when entities are detected within the text, unchaining the execution of the subsequent modules. Therefore, ensuring that the NER module provides accurate results is essential for the correct functioning of the platform.

As presented in Section 3.1, two strategies are considered for the detection of named entities within the text: a default strategy based on Spanish NER models, and a translation-based strategy. Ideally, both approaches should yield the same results, identifying the same elements as diseases and chemicals, respectively. Experiments are conducted to assess the performance of both strategies. Since the Spanish NER model is fine-tuned on the DISTEMIST corpus, an unseen, valid corpus is required to evaluate such that the comparison is fair. For this purpose, the SPACCC corpus, described in Section 2, is used. Both performance and accuracy aspects are considered for the evaluation.



(a) Number of diseases detected.



(b) Number of chemicals detected.

Figure 6: (a) Number of diseases detected per NER strategy (b) number of chemicals detected per ner strategy.

From an efficiency perspective, the computation time per sample is computed for both strategies. The results are reported in Figure 5. As shown, on average, using a translation-based approach is slightly faster. Moreover, the results suggest that the default approach is more stable in

Table 1
NER results using both strategies on the sample SPACCC document depicted in Figure 4.

	Default-approach	Translation-based
Diseases	'inmunosupresión asociada a vih', 'hemoptisis', 'neumonía intrahospitalaria', 'neumonía por p. jirovecii', 'herpes recurrente en boca', 'lesión herpética', 'disfagia'	'vih y p. neumonía de jirovecii', 'ey', 'deterioro respiratorio', 'choque', 'herpes', 'fiebre', 'encaja', 'hemoptisis', 'disnea', 'lesión herpética', 'neumonía adquirida en el hospital', 'dolor en el pecho', 'enfermedad', 'tos seca', 'odinofagia', 'ppp', 'infección por cmv', 'neumomediastino', 'disfagia', 'linfopenia'
Chemicals	'oxígeno', 'prednisona', 'trimetoprim/sulfametoxasol'	'prednisona', 'surgir', 'gotas', 'Apurarse', 'estoy abierto', 'oxígeno', 'ganciclovir', 'trimetoprima', 'Abucheo'

terms of efficiency, even though the average time is slightly longer. Although the difference is almost undetectable for single-document processing, it is definitely noticeable when processing large amounts of documents. The total amount of time required to process the 1,000 documents that comprise the SPACCC corpus is around 400 seconds when using the default approach and is reduced by almost half when using the translation-based approach ($\approx$ 220 seconds). This may be due to the fact that the processing time required by the NER model is significantly higher than the time required for document translation. The default approach comprises two separate NER models, while the translation approach employs only one NER model. Subsequently, the overall computation time, on average, is lower in the second approach.

Then both NER strategies are evaluated in the SPACCC corpus. The corpus is not devised for its evaluation on the NER task and therefore does not contain a list of the entities contained within the text. Therefore, a quantitative and manual approach is followed to assess the performance of both strategies. The SPACCC corpus comprises 1,000 medical reports samples. Figure 4 shows a sample document extracted from the SPACCC corpus. Each report is processed by both strategies, extracting the list of diseases and chemicals detected by each strategy. The results are then filtered to remove stop words or empty characters that may have been misdetected by the models. Then, the overlap between the results obtained by both strategies is measured. Ideally, both approaches should result in the same number of entities detected, and thus the overlap should be close to 100%. Figure 6 outlines the results achieved by both approaches for NER. The default approach identifies more entities than the translation-based approach. However, as can be observed, the difference is not quite significant. The number of elements detected by both approaches, while not identical, is fairly homogeneous. This is especially noticeable in chemical detection (Figure 6b), where while there are some files in which there is some disparity between the number of chemicals detected, the results are fairly homogeneous. In the case of disease detection (Figure 6a), the disparity between the results is quite noticeable. Therefore, an analysis of the entities detected per strategy is conducted to compare the results achieved.

Table 1 presents an example of the entities recognized by both strategies in a sample SPACCC documents. As can be observed for the case of chemicals, the default approach, while detecting a lower number of entities, correctly classifies all terms detected. In the translation-based strategy, all entities detected by the default approach are also correctly detected, but several incorrect terms are also detected. For example, the terms "gotas" and "surgir", which do not refer to any chemical. A similar phenomenon occurs in the case of diseases, where all elements detected by the default approach are detected by the translation-based approach. However, in this case, the translation-based approach detects a series of terms that are correct but undetected by the default approach, such as "linfopenia" of "fiebre". Nevertheless, disease detection is not as trivial as chemical detection, since there is no hard criteria on how to distinguish a disease from a symptom. For example, in the DISTEMIST corpus, which was used to train the NER model in the default approach, entities such as "dolor en el pecho" or "tos seca" are not labeled diseases, as they correspond to the symptoms of the labeled diseases. In the BC5CDR dataset, symptoms are also considered as diseases and thus labeled as such. This disparity in the labeling criteria of the datasets, paired with the potential mistakes introduced by translation (for example, the detection of the term "abucheo", which is an incorrect translation from the detected term "PO_2"), leads to results that, while different, both can be considered as correct up to a certain degree. This same pattern repeats in the remaining of the analyzed samples, thus leading to the

conclusion that while the default approach is more precise, the translation-based approach has a higher recall.

5. Conclusions and Future Work

This paper presents a self-contained platform that processes large-scale English biomedical documents to extract evidence through a natural language-based interface in Spanish, focusing especially on those related to COVID-19. The presented platform centralizes and adapts a series of resources developed within the DRUGS4COVID++ project, which were initially developed in English, for the Spanish language. The platform comprises three NLP resources: a named-entity recognition module, which detects diseases and chemicals within the text using two different strategies; a term-linking module that offers extended information on the detected entities; and an evidence retrieval module, which provides evidence in the form of scientific reports on the pairwise interaction between diseases and chemicals. Moreover, the platform considers the introduction of expert knowledge within the detection process, since NER modules can be flawed, such that users can manually incorporate diseases and entities in real time within the execution process. The NER module is then tested on the two offered strategies, a default strategy which uses two separate NER models in Spanish for the detection of diseases and chemicals, respectively, and a translation-based strategy that translates the text into English and uses existing pre-trained models for the detection of diseases and chemicals simultaneously. The experimentation shows a slight disparity in the results obtained by both approaches: while the results regarding chemicals are fairly similar between both approaches, the results regarding disease detection vary since the models employed in the default approach only consider diseases, while the translation-based approach labels symptoms as diseases as well, thus leading to a higher number of detected entities.

In future work, it would be interesting to evaluate the performance of the platform with respect to real expert users, to measure its utility, usability, and reliance. Moreover, it would be interesting to include evidence in Spanish, since the current evidence repository only considers scientific articles in English. Therefore, incorporating evidence in Spanish would enrich the evidence repository and reduce the need of use translation mechanisms, which can result in the loss of information.

Acknowledgments

The research work presented in this paper has been funded by the DRUGS4COVID++ project, with the financial support of BBVA.

References

[1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: K. Verspoor, K. B. Cohen, M. Dredze, E. Ferrara, J. May, R. Munro, C. Paris, B. Wallace (Eds.), Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL: https://aclanthology.org/2020.nlpcovid19-acl.1.

[2] Q. Chen, A. Allot, Z. Lu, Keep up with the latest coronavirus research, Nature 579 (2020) 193––193. doi:10.1038/d41586-020-00694-1.

[3] J. Shuja, E. Alanazi, W. Alasmary, A. Alashaikh, Covid-19 open source data sets: a comprehensive survey, Applied Intelligence 51 (2021) 1296–1325. doi:10.1007/s10489-020-01862-6.

[4] R. Lamsal, Design and analysis of a large-scale covid-19 tweets dataset, Applied Intelligence 51 (2020) 2790––2804. doi:10.1007/s10489-020-02029-z.

[5] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, J. Kim, Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis, IEEE Transactions on Computational Social Systems 8 (2021) 1003–1015. doi:10.1109/TCSS.2021.3051189.

[6] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, P. Bogdan, A covid-19 rumor dataset, Frontiers in Psychology 12 (2021). doi:10.3389/fpsyg.2021.644801.

[7] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, H. Larson, The pandemic of social media panic travels faster than the covid-19 outbreak, Journal of Travel Medicine 27 (2020). doi:10.1093/jtm/taaa031.

[8] J. Yu, Y. Lu, J. Muñoz-Justicia, Analyzing spanish news frames on twitter during covid-19—a network study of el país and el mundo, International Journal of Environmental Research and Public Health 17 (2020) 5414. URL: https://www.mdpi.com/1660-4601/17/15/5414. doi:10.3390/ijerph17155414.

[9] S. Allés-Torrent, G. del Rio Riande, J. Bonnell, D. Song, N. Hernández, Digital narratives of covid-19: A twitter dataset for text analysis in spanish, Journal of Open Humanities Data 7 (2021) 5. doi:10.5334/johd.28.

[10] A. Intxaurrondo, M. Krallinger, Spaccc, 2019. URL: https://doi.org/10.5281/zenodo.2560316. doi:10.5281/zenodo.2560316.

[11] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: K. Jin-Dong, N. Claire, B. Robert, D. Louise (Eds.), Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: https://aclanthology.org/D19-5701. doi:10.18653/v1/D19-5701.

[12] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020, pp. 303–323. URL: http://ceur-ws.org/Vol-2664/cantemist_overview.pdf.

[13] A. Miranda-Escalada, L. Gasco, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation

(CLEF) Forum. CEUR Workshop Proceedings, 2022. URL: https://ceur-ws.org/Vol-3180/paper-11.pdf.

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

[15] F. Remy, K. Demuynck, T. Demeester, BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights, Journal of the American Medical Informatics Association (2024) 38412333. URL: https://doi.org/10.1093/jamia/ocae029. doi:10.1093/jamia/ocae029.

[16] A. Á. Pérez, A. Iglesias-Molina, L. P. Santamaría, M. Poveda-Villalón, C. Badenes-Olmedo, A. Rodríguez-González, EBOCA: Evidences for BiOmedical Concepts Association Ontology, Springer International Publishing, 2022, pp. 152––166. URL: http://dx.doi.org/10.1007/978-3-031-17105-5_11. doi:10.1007/978-3-031-17105-5_11.

[17] C. Badenes-Olmedo, D. Chaves-Fraga, M. Poveda-Villalón, A. Iglesias-Molina, P. Calleja, S. Bernardos, P. Martín-Chozas, A. Fernández-Izquierdo, E. Amador-Domínguez, P. Espinoza-Arias, L. Pozo-Gilo, E. Ruckhaus, E. González-Guardia, R. Cedazo, B. López-Centeno, Ó. Corcho, Drugs4covid: Drug-driven knowledge exploitation based on scientific publications, CoRR abs/2012.01953 (2020). URL: https://arxiv.org/abs/2012.01953. arXiv:2012.01953.