

# Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials.

Luis Gasco<sup>1</sup>, Anastasios Nentidis<sup>2,3</sup>, Anastasia Krithara<sup>2</sup>, Darryl Estrada-Zavala<sup>1</sup>, Renato Toshiyuki Murasaki<sup>4</sup>, Elena Primo-Peña<sup>5</sup>, Cristina Bojo Canales<sup>5</sup>, Georgios Paliouras<sup>2</sup> and Martin Krallinger<sup>1</sup>

<sup>1</sup>Barcelona Supercomputing Center. Barcelona, Spain

<sup>2</sup>National Center for Scientific Research "Demokritos", Athens, Greece.

<sup>3</sup>Aristotle University of Thessaloniki. Thessaloniki, Greece

<sup>4</sup>Centro Latinoamericano y del Caribe de Información en Ciencias de la Salud, Organización Panamericana de la Salud. São Paulo, Brasil.

<sup>5</sup>Instituto de Salud Carlos III. Biblioteca Nacional de Ciencias de la Salud. Madrid, Spain

## Abstract

There is a pressing need to exploit recent advances in natural language processing technologies, in particular language models and deep learning approaches, to enable improved retrieval, classification and ultimately access to information contained in multiple, heterogeneous types of documents. This is particularly true for the field of biomedicine and clinical research, where medical experts and scientists need to carry out complex search queries against a variety of document collections, including literature, patents, clinical trials or other kind of content like EHRs. Indexing documents with structured controlled vocabularies used for semantic search engines and query expansion purposes is a critical task for enabling sophisticated user queries and even cross-language retrieval. Due to the complexity of the medical domain and the use of very large hierarchical indexing terminologies, implementing efficient automatic systems to aid manual indexing is extremely difficult. This paper provides a summary of the MESINESP task results on medical semantic indexing in Spanish (BioASQ/ CLEF 2021 Challenge). MESINESP was carried out in direct collaboration with literature content databases and medical indexing experts using the DeCS vocabulary, a similar resource as MeSH terms. Seven participating teams used advanced technologies including extreme multilabel classification and deep language models to solve this challenge which can be viewed as a multi-label classification problem. MESINESP resources, we have released a Gold Standard collection of 243,000 documents with a total of 2179 manual annotations divided in train, development and test subsets covering literature, patents as well as clinical trial summaries, under a cross-genre training and data labeling scenario. Manual indexing of the evaluation subsets was carried out by three independent experts using a specially developed indexing interface called ASIT. Additionally, we have published a collection of large-scale automatic semantic annotations based on NER systems of these documents with mentions of drugs/medications (170,000), symptoms (137,000), diseases (840,000) and clinical procedures (415,000). In addition to a summary of the used technologies by the teams, this paper

---

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ lgasco@bsc.es (L. Gasco); tasosnent@iit.demokritos.gr (A. Nentidis); akrithara@iit.demokritos.gr (A. Krithara); darrylestrada97@gmail.com (D. Estrada-Zavala); murasaki@paho.org (R. T. Murasaki); eprimo@isciii.es (E. Primo-Peña); cbojo@isciii.es (C. B. Canales); paliourg@iit.demokritos.gr (G. Paliouras); martin.krallinger@bsc.es (M. Krallinger)

ORCID 0000-0002-4976-9879 (L. Gasco); 0000-0002-2646-8782 (M. Krallinger)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

shows that there was a clear improvement in terms of the best scoring systems when compared to previous efforts and there was also a clear time saving of up to 67% when using pre-indexing with these systems compared to manual indexing of documents. MESINESP corpus: <https://doi.org/10.5281/zenodo.4612274>

### **Keywords**

Hierarchical Text Classification, Semantic Indexing, Information Retrieval, Question Answering, Biomedical knowledge, Automatic Indexing, DeCS, Spanish

## **1. Introduction**

Since the beginning of the 21st century, we have been immersed in a digitalization process that, together with the advent of the information age, has facilitated the generation, dissemination, and access to digital content using computer-based tools. The production and accumulation of information have been particularly relevant in the field of health and biomedicine, in which difficulties have arisen in accessing relevant information on specific topics not only for practitioners but also for researchers, public healthcare decision-makers, and other health professionals.

The difficulties in finding the right information among so many documents become particularly evident in more demanding scenarios such as pandemics, when a high amount of new research is published every day. For example, more than 160,000 COVID-19 documents were published between January and October 2020, and estimates indicate that more than 700,000 articles may be published before the end of 2021 [1]. There is also a pressing need to enable and improve cross-lingual and multilingual search technologies, as a considerable number of publications, in particular those that present more clinically oriented results such as clinical case reports, do correspond to non-English content. Recent advances in machine translation approaches specifically adapted to the characteristics of medical language [2] as well as the use of multilingual controlled vocabularies exploited by tools like BabelMeSH<sup>1</sup> show that multilingual medical information retrieval will contribute to improve information access of healthcare professionals.

In an attempt to help health professionals access up-to-date and relevant information, initiatives have emerged to improve retrieval of COVID-19-related documents [3], new datasets have been generated to train better AI systems specialised in scientific literature [4] or even new platforms have appeared to highlight the most relevant research in the field [5, 6].

Despite all these innovative projects, the searching process is still centered on bibliographic databases, which incorporate functionalities to build complex semantic queries that allow to gather more specific results. Conversely, the results obtained rely on prior human indexing of the database records. Manual indexing consists of assigning a set of descriptors, which are part of a controlled vocabulary, to a manuscript to describe its content. The effectiveness of this process depends on the judgement, thoroughness and speed of the annotators, which makes it a slow, unsystematic and costly process.

To overcome these difficulties, previous initiatives such as the TREC genomics track (2003-2007) were launched with the aim of developing the state-of-the-art of automatic biomedical

---

<sup>1</sup><https://babelmesh.nlm.nih.gov>

semantic indexing [7]. More recently, from 2013, the BioASQ challenge focused on biomedical question answering and semantic indexing of scientific literature written in English [8].

However, it is a fact that there is a considerable amount of medically relevant content published in languages other than English. This is particularly significant for non-scientific literature records, such as clinical trials, EHRs, and patents, which are written entirely in the native language of each country, with some exceptions. Additionally, there is a great lack of interoperability in semantic search queries when looking for information in different data sources, a procedure mandatory if a health professional wants to get a complete vision about a specific topic. For example, if a practitioner wanted to know about adverse reactions to vaccines, he easily would be able to obtain research papers from scientific databases. When gathering more information on this topic, he should also search for information about the clinical trials conducted or about proprietary vaccine compounds in patents; nonetheless, search engines of these databases do not use the same controlled terminologies than scientific literature records, causing more efforts in creating queries with terminologies not known to the physician.

The MESINESP2 track, promoted by the Spanish Plan for the Advancement of Language Technology (Plan TL)<sup>2</sup> and organized by the Barcelona Supercomputing Center (BSC) in collaboration with BioASQ, aims to improve the state of the art of semantic indexing for documents written in Spanish, the second language with the most native speakers in the world<sup>3</sup>. We strongly believe that interoperability in semantic search queries is essential to improve the information retrieval procedure. For that reason, in this edition we propose to index not only scientific literature, but also clinical trials and patents to evaluate if well-known structured vocabularies can be used for other type of biomedical documents. The generation of automatic systems to index other documents would foster the interoperability of search queries and would be the first step towards unifying databases to include all available biomedical information by exploring cross-corpus training scenarios. Moreover, we briefly introduce new possible metrics based on semantic similarity to assess the quality of the generated systems as well as the usability of the results to improve the efficiency of manual indexing tasks using predictive models and manual indexing tools.

This paper presents the data and results of the MESINESP2 track, which was part of the CLEF-BioASQ 2021 challenge. In this document we firstly provide an overview of the task, the corpus and additional data resources we prepared for the participant teams. We also present and analyse the systems developed by the participants. We evaluate the performance of the systems using state-of-the-art evaluation measures, but we will also introduce new feasible metrics that may be used to evaluate the performance from a semantic perspective. Finally, we conclude with a discussion about the current quality of results and the current advantages of semantic indexing systems, as well as the future steps in the MESINESP track.

## 2. Track description

MESINESP2 proposed to participating teams the challenge of training AI models capable of assigning DeCS codes (*Descriptores en Ciencias de la Salud*, a terminology derived and extended

---

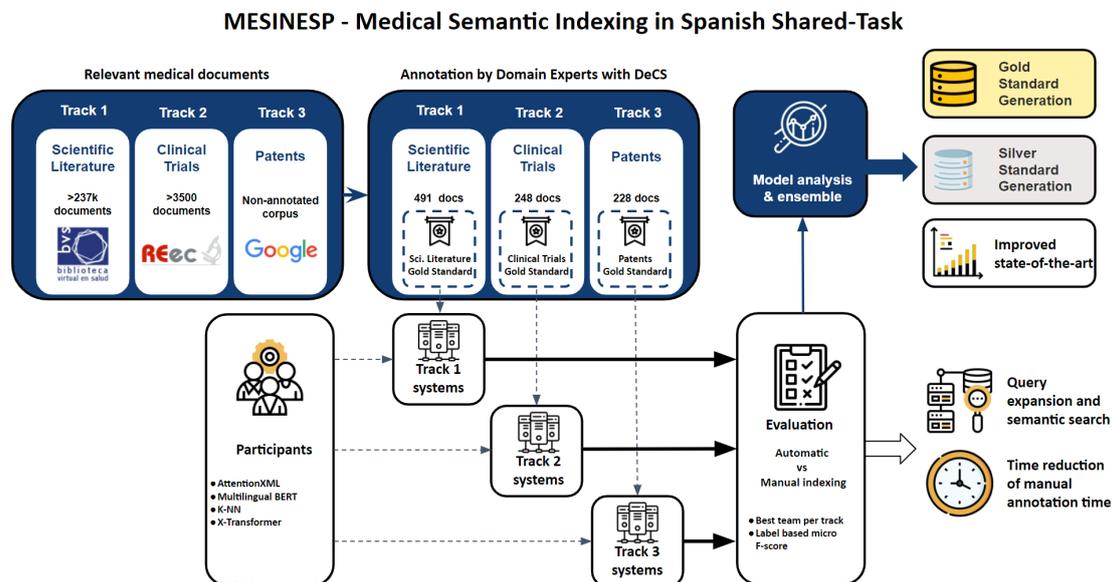
<sup>2</sup><https://plantl.mineco.gob.es>

<sup>3</sup><https://www.ethnologue.com/guides/ethnologue200>

from MeSH terms) to biomedical documents. The predictions were evaluated using a Gold Standard manually annotated by expert human indexers. We structured MESINESP2 into three independent subtracks launched consecutively, as shown in Figure 1, each focused on designing semantic indexing systems for different types of texts.

- **MESINESP-L - Scientific Literature (Subtrack 1):** This track required automatic indexing with DeCS terms of abstracts from scientific articles using two highly used databases in Spanish: IBECs<sup>4</sup> and LILACS<sup>5</sup>.
- **MESINESP-T - Clinical Trials (Subtrack 2):** This track asked to predict DeCS codes automatically for clinical trials from the REEC database<sup>6</sup>
- **MESINESP-P - Patents (Subtrack 3):** This track required automatic indexing with DeCS terms the content of patents in Spanish extracted from Google Patents<sup>7</sup>.

The rest of the section briefly describes the controlled terminology used in the shared task, the annotation process employed and a description of the corpora generated for the participants.



**Figure 1:** MESINESP2 workflow diagram showing the resources generated, as well as the areas it will improve.

<sup>4</sup>IBECs includes bibliographic references from scientific articles in health sciences published in Spanish journals. <http://ibecs.isciii.es>

<sup>5</sup>LILACS is the most important and comprehensive index of scientific and technical literature of Latin America and the Caribbean. It includes 26 countries, 882 journals and 878,285 records, 464,451 of which are full texts <https://lilacs.bvsalud.org>

<sup>6</sup>Registro Español de Estudios Clínicos, a database containing summaries of clinical trials <https://reec.aemps.es/reec/public/web.html>

<sup>7</sup>Google Patents is a public database that aggregates patents from many IP agencies including the OEPM (Oficina Española de Patentes y Marcas) <https://support.google.com/faqs/answer/7049585>

## 2.1. DeCS terminology used for semantic indexing

DeCS (*Descriptores Descriptores en Ciencias de la Salud*, Health Science Descriptors) is a structured controlled vocabulary created by BIREME to index scientific publications on BvSalud (*Biblioteca Virtual en Salud*, Virtual Health Library), the largest database of scientific documents in Spanish, which hosts records from the databases LILACS, MEDLINE, IBECs, among others.

The aim of this vocabulary is to facilitate the retrieval of records written in Spanish, Portuguese and English contained in BvSalud. This trilingual vocabulary is based on the MeSH (Medical subject Headings) terminology of the US National Library of Medicine (NLM), but also covers additional vocabularies to describe in more detail areas such as Public Health, Homeopathy, Health and Science and Health Surveillance. Similarly to MeSH, the DeCS vocabulary is organized basically in the form of a hierarchical tree structure. This enables improving search strategies by directly exploiting both more general or more specific term-relations through the hierarchical vocabulary structure.

For the MESINESP2 track we have used the 2020 version of DeCS, as at the time of releasing the datasets the new 2021 edition was not yet published. The 2020 version comprises a collection of 34,041 unique descriptors, 60,670 alternative terms in Spanish and a total of 77 qualifiers.

In order to improve the interoperability of the generated models, and because it was planned to use the generated systems for pre-annotations of BvHealth documents, in which there are a large number of COVID19-related documents, we have included an extension of COVID19-related codes that will be incorporated in the 2021 version.

## 2.2. Annotation process

Documents from BvSalud are being manually indexed by experts. Complementing the evaluation scenario of the BioASQ's English-language semantic indexing task, which relies on progressive indexing of PubMed articles, in case of the MESINESP track, evaluation is done using a subset of records indexed by professionals specifically for this evaluation initiative allowing us to control the quality of the annotations in the Gold Standard. The MESINESP Gold Standard evaluation data consists of a carefully selected subset of records (titles and abstracts) which are manually annotated by experts with DeCS codes following the BvSalud indexing guidelines, but focusing only on the title and abstract content<sup>8</sup>. We evaluated the quality of the models generated by registered teams by comparing automatic predictions of DeCS codes against a manually indexed Gold Standard.

After a thorough Inter-Annotator Agreement analysis of the seven indexers from last year's edition, this year we selected the three experts with the best agreement (around 0.9) to annotate the documents. These annotators were asked to index 1000 documents, so that each document was manually indexed at least twice. To perform the annotation, we used a new indexing tool called ASIT (Advanced Semantic Indexing Tool) that features performance improvements and allows recording annotation performance metrics such as the time a human indexer takes to annotate a document. In addition, ASIT provides a user-friendly and interactive interface to

---

<sup>8</sup>For a detailed description of the document selection criteria and the creation of the corpus, see section 2.3 of this document.

perform semantic indexing of documents, such as interactive descriptor search; and includes predictive systems to suggest which codes to use based on the content of the text.

The screenshot displays the ASIT User Interface. At the top, there is a header bar with the text 'Indizador del Plan TL' and a search icon. Below this is a table with the following columns: 'Identificador', 'Titulo', 'Fuente', 'Tipo', and 'Validado'. The table contains one row with the following data: 'biblio-1126907', 'Aspectos clínicos relacionados con el Síndrome Respiratorio Agudo Severo (SARS-CoV-2)', 'LILACS', 'article', and 'Sí'. Below the table, there is a pagination control showing 'Items per page: 10' and '0 of 0'. Below the table is a section titled 'Aspectos clínicos relacionados con el Síndrome Respiratorio Agudo Severo (SARS-CoV-2) (biblio-1126907)'. This section contains a 'RESUMEN' (Summary) in Spanish, followed by a search box labeled 'Términos' with several suggested terms: 'Betacoronavirus', 'Masculino', 'Neumonía Viral', 'Síndrome de Dificultad Respiratoria del Adulto', 'Coronavirus', 'Infecciones por Coronavirus', 'COVID-19', and 'vacunas contra la COVID-19'. Below the search box is a 'Completado' button. At the bottom of the interface, there is a footer bar with the 'Plan TL' logo, the text 'Developed by the Text Mining Unit at BSC © 2019. Code licensed under a MIT License.', and social media links for LinkedIn, GitHub, and Twitter.

**Figure 2:** ASIT User Interface is divided into a document selection panel and an annotation panel. The annotation panel displays the text content, as well as an interactive descriptor search box that can incorporate DeCS codes suggestions.

### 2.3. MESINESP2 corpora

The aim of MESINESP2 was to explore semantic indexing technologies applied to a variety of heterogeneous health related content in Spanish. On the one hand, for scientific literature, we considered the IBECs and LILACS records. For clinical trials, we considered the studies present in the Spanish Registry of Clinical Trials (REEC); and finally, we considered the patents in Spanish present in Google Patents. Content selection criteria included practical importance of the selected databases, size and number of records, practical impact of the resulting semantic indexing systems as well as access and redistribution of the data collections. We have prepared a corpus for every subtrack of the MESINESP task. For each of them, a process of data collection/harvesting, cleaning, data harmonization, subset selection and annotation of the records by experts was carried out.

### 2.3.1. MESINESP-L corpus

First, we crawled the BvSalud platform with a specific framework<sup>9</sup> to obtain records from IBECs and LILACS on 01/29/2021. This means that the data is a snapshot of that moment and that may change over time since LILACS and IBECs usually add or modify records/indexes after the first inclusion in the database. We obtained the title, abstract, language, journal and date of publication of more than 1.14 million records. The initial corpus contained documents in Spanish, English, Portuguese and French, and many of them were stored in the database without having been manually indexed with DeCS.

To generate a very large training dataset, we selected documents that were already indexed at this time point by literature databases. We then filtered records with empty abstracts and those that were not written in Spanish, as well as other articles that were not journal article publication types. We published a corpus of 237,574 scientific journal articles manually annotated by LILACS and IBECs experts with DeCS codes. This year we removed the DeCS qualifier information from those annotations, and assigned string descriptors to their corresponding DeCS identifiers. This prevented inconsistencies in the process of mapping descriptors to their identifiers by teams and ensured that they all had exactly the same set of labels.

For the development set, in order to resemble the characteristics of the test set used for evaluation purposes, we provided a set of records manually indexed by expert annotators. This dataset included 1065 articles (i.e. titles and abstracts) annotated with DeCS by the three indexers who obtained the best IAA (over 0.9) in the last MESINESP. Among all items in the development set, 285 were annotated by more than one annotator (we selected the union between annotations), and 852 articles were annotated by only one of those selected indexers.

To generate the evaluation dataset, we pre-selected all non-indexed Spanish scientific articles in the database that were published since 2020. As a novelty, and to align better medical indexing of literature content with indexing of medical records and EHRs, more than 8,000 pre-selected records were semantically compared with a corpus of clinical records provided by major hospitals in Spain (discharge summaries, clinical course and radiology reports). The 500 publications most similar to the medical records were selected for annotation by the three experts and were included in the final test set provided to the participants. A background set of 9676 Spanish-language clinical practice guidelines documents was also included in the test set distributed to participants, in order to evaluate the performance of automatic indexing systems on this type of biomedical documents in the future by manually validating team predictions by human indexers

Table 1 contains an overview of the MESINESP corpus statistics. The manually indexed corpora contained an average of 10 DeCS codes and about 190 tokens in length. Overall, the number of codes assigned to each document did not have a significant association with the length of its abstract as can be seen in Figure 3.

### 2.3.2. MESINESP-T corpus

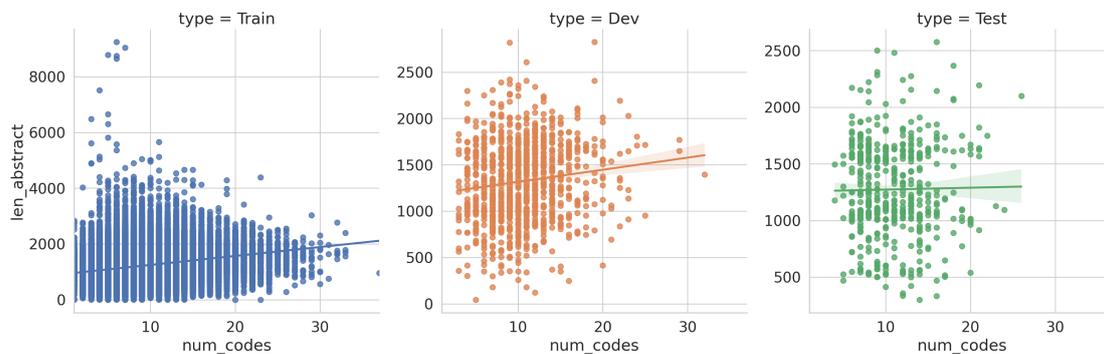
Clinical trials from the REEC database are currently not indexed on a regular basis with any controlled terminology. This makes efficient searches within clinical trials in this database quite

---

<sup>9</sup><https://github.com/ankush13r/BvSalud>

**Table 1**  
MESINESP-L (Scientific Literature) corpus statistics

<i>MESINESP-L</i>	Docs	DeCS	Unique DeCS	Tokens	Avg.DeCS/doc	Avg.token/doc
<b>Training</b>	237,574	~1.98M	22434	~43.1M	8.37(3.5)	181.45(72.3)
<b>Development</b>	1065	11283	3750	211,420	10.59(4.1)	198.52(64.2)
<b>Test</b>	491	5398	2124	93645	10.99(3.9)	190.72(63.6)



**Figure 3:** Correlation plots between the number of DeCS codes and abstract length for each of the three corpus subsets.

challenging. To account for the lack of a larger training set for this sub-track, we have used the silver standard (automatic predictions) generated by the MESINESP 2020 participants to create the subtrack training set surrogate [9]. Given that the quality of the task results were diverse, we only released the predictions made by the winning team of the last year’s task, which achieved a micro-averaged F-measure of 0.4254. For the development set, we released 147 records annotated manually by expert indexers in the last MESINESP edition.

For the test set, we downloaded the whole REEC database using the *reecapi* python library [10], a non-official library to access the REEC databases developed for this task. Since the number of indexed documents was substantially smaller than in MESINESP-L, we calculated the semantic similarity between the subtrack 1 training corpus and the 416 clinical trials published since 2020. Then, we selected the top 250 most similar articles, which included many COVID-19 clinical trials, being these records annotated by our indexers. In addition to the manually annotated data, which were used to evaluate the participating systems, we also included a background set of 8,669 documents from drug product data sheets to be automatically annotated by the participating systems.

In terms of corpus statistics (Table 2), clinical trials were longer documents dealing with more topics than scientific articles, which translates into a higher number of DeCS codes and a longer length in word tokens. However, the diversity of codes were much more limited than in subtrack 1, due to the health-focused nature of the documents.

**Table 2**  
MESINESP-T (Clinical Trials) corpus statistics

<i>MESINESP-T</i>	Docs	DeCS	Unique DeCS	Tokens	Avg.DeCS/doc	Avg.token/doc
<b>Training</b>	3560	52257	3940	~4.13M	14.68(1.19)	1161.0(553.5)
<b>Development</b>	147	2038	771	146,791	13.86(5.53)	998.58(637.5)
<b>Test</b>	248	3271	905	267,031	13.19(4.49)	1076.74(553.68)

### 2.3.3. MESINESP-P corpus

Similar to clinical trials, patents are not indexed using DeCS, but with *International Patent Classification (IPC)* codes. There are some studies that propose mapping between MeSH and IPC descriptors, but unfortunately at the time of the competition there were no public resources available to map these terminologies [11]. Because of this lack of resources, we decided to propose this track as a cross-corpus training challenge, in which participants should transfer previous models to the patent domain with a very low amount of annotated data. Improving patent search is of key practical relevance for competitive intelligence and intellectual property purposes.

We downloaded from Google Big Query<sup>10</sup> all the patents written in Spanish with the IPC codes “A61P” and “A61K31”<sup>11</sup>, some well-known codes for some of the task organisers [12]. After data acquisition, we obtained 65,513 patents, from which we chose the 250 most lexically similar to the MESINESP-L training set. This subset of patents was annotated by the expert indexers, 115 were used as the development set and the rest as the test set. In this subtrack, we also include a background set with a larger number of patents to be annotated, trying to increase the size of the dataset by improving the annotation process in future editions of MESINESP.

As can be seen in Table 3, the diversity of DeCS codes is the lowest of all subtracks because we decided to work with a subset of patents associated with only two IPC codes. The number of associated descriptors was similar to those found in the MESINESP-L annotated data, and the length of the documents was very diverse, with a mix of short and long documents.

**Table 3**  
MESINESP-P (Patents) corpus statistics

<i>MESINESP-P</i>	Docs	DeCS	Unique DeCS	Tokens	Avg.DeCS/doc	Avg.token/doc
<b>Development</b>	109	1092	520	38564	10.02(3.11)	353.79(321.5)
<b>Test</b>	119	1176	629	9065	9.88(2.76)	76.17(27.36)

<sup>10</sup><https://cloud.google.com/blog/topics/public-datasets/google-patents-public-datasets-connecting-public-paid-and-private-patent-data>

<sup>11</sup>IPC Code reference: *A61P* (Specific therapeutic activity of chemical compounds or medicinal preparations, *A61K31* ( Medicinal preparations containing organic active ingredients)



*transformer based encodings using Deep Language Models* for the representation of text and labels using mostly Multilingual-BERT.

The **Fudan University team** built their *BertDeCS* system following their previously published *AttentionXML* architecture [16]. They made some modifications to make it perform better in Spanish domain. First, the encoding layer of their system uses *Multilingual BERT* which performs better on non-English documents. Once the word representation is obtained, the model uses label-level attention to get different representations for different labels. Finally, they used a fully-connected layer to get confidence scores for each label. They pretrained the model with English papers from the MEDLINE dataset, and then fine-tuned the model with Spanish articles from MESINESP2 corpus.

The **Vicomtech team** from the Basque Research and Technology Alliance implemented two systems based on transformers [17]. Specifically they used BERT pre-trained models to encode the text data and build a multi-label classification on top. The first system, CSS, relied on combining BERT-encoded tokens as input of a classification model. The second approach *LabelGLOSSES* was similar to Fudan's, and it includes an encoded representation of the DeCS descriptors built using a pre-trained BERT encoder, together with the layers of the first model, namely the text data encoder and the multilabel classifier.

The **Roche system** uses a hybrid semantic indexing method that integrates transformer-based multi-label text classification (MLTC), and named entity recognition (NER) provided by Barcelona Supercomputing Center [18]. The transformer-based solution is implemented with package "transformers" [19] and PyTorch with SentencePiece, the BERT pre-trained model [20] and auxiliary multiple binary classifiers. The system complements *rare classes through additional synonym matching* the entities in the articles to DeCS terms. The results are pooled together for each article to assign the final labels.

The **Iria systems**, a joint work of researchers from Universidade de Vigo and Universidade da Coruña, followed the approach described in [21]. They applied a k-NN approach using linguistically motivated index terms such as lemmas, syntactic dependencies, NP chunks, name entities and keywords. They also sent runs using a k-NN approach over indices storing dense representations of training documents obtained by means of sentence level embeddings using SentenceTransformers library [22].

The **Lasige-TEAM**, from Universidade de Lisboa, developed a prediction pipeline composed by two modules [23]. The first one was a graph-based entity linking model that used the Personalized PageRank algorithm in candidate disambiguation and a semantic similarity-based filter to select the most relevant entities in each documents. The second one was an adaptation of the X-Transformer algorithm for extreme multi-label classification [24] that had three components, being one of them the deep neural matcher, based on Multilingual BERT.

This year there has been teams that have decided to use traditional text representation methods to examine how they work in semantic indexing in Spanish. The team from Universidad de Chile used TF-IDF, word embeddings and cosine similarity measures to get the terms associated to each document.

The MESINESP2 baseline consisted of a simple textual lookup system. This approach used the descriptors and synonyms of each of the DeCS codes to search on the text and assign the code to the document if a match was detected. This approach obtained an MiF of 0.2876 for MESINESP-L, 0.1288 for MESINESP-L and 0.2992 for MESINESP-P.

### 3.2. System evaluation

The main evaluation metric used for the task was the Micro-averaged F-score. Based on this metric, the best results were obtained by the three MESINESP2 subtracks was the team from Fudan University. All their models were ranked first in each of the subtracks.

As happened in the last edition of MESINESP, the results obtained are lower than those of the English task when the same type of technologies are used [25]. In this edition we chose to generate a dataset exclusively with scientific articles published in journals in order to limit possible inconsistencies in documents of other types such as PhD dissertations. In addition, we opted to provide a list of mapped DeCS codes instead of textual descriptors to avoid inconsistencies among participants when assigning codes to records. The measures taken have resulted in a smaller but higher quality dataset which, together with the technological improvements implemented by the participants, has led to higher MiF values overall, with the winner obtaining a score of 0.4837.

The possible reasons for the drop in performance with respect to the English task may still be associated with a lower number of training documents and the fact that the documents obtained from BvSalud come mainly from two bibliographic databases that are indexed in slightly different ways [26]. On the other hand, since the update of deprecated DeCS codes is not performed simultaneously with the update of the controlled vocabulary, it may lead to some issues in the list of terms associated with the documents that may result in a loss of performance in the automatic indexing models developed by the participants.

Regarding the MESINESP-T task, there is no corresponding task in English dedicated to the indexing of clinical trials to compare the results. The TREC 2021 Clinical Trials Track had the aim to evaluate clinical trial retrieval systems, but using a different setting<sup>12</sup>. The achieved outcomes were significantly lower than those found in the scientific literature. Although this drop in performance could be linked to the fact of lacking a large set of Gold Standard training data (only a silver standard dataset was available), participants reported that they preferred to reuse the models trained with scientific literature, incorporating the development set, to make their final predictions. The average token length of the clinical trials was much longer than found in scientific literature (around 1000 vs. 200). Many of the participants opted to use BERT models, which have an input size limit of 512 tokens which may have caused some of the content not to be processed by the model, inherently losing indexing performance. Despite these drawbacks, the top scoring team was Fudan University with a MiF of 0.3640. However, in terms of accuracy this team was outperformed by a run of the Roche's models, achieving 0.4004.

The patent subtrack is one of the major novelties of MESINESP. Despite the existence of automatic patent indexing tasks using the IPC ontology, there were no public system that could be used for indexing patents with MeSH/DeCS terms [27]. MESINESP-P presented a great challenge for the participants due to the lack of training corpus, and a very small volume of development data. Since some of the statistics between scientific literature and patents corpora were similar, the participants opted to use the same models they had used in MESINESP-L to make the predictions of the test set. Despite being systems trained specifically for the scientific literature, the results obtained for the patent track are promising. The performance of some of the systems, especially those of the Fudan, Roche and Iria teams, remains at the same same

---

<sup>12</sup><http://www.trec-cds.org/2021.html>

level as for scientific literature.

Since the corpus was generated through the selection of two IPC code labels, the content of the articles might be more specific in some way. If the codes used to index patents were used extensively in the scientific literature corpus, the system may have learned to assign this type of descriptors more accurately. Moreover, patents were selected using a lexical similarity criteria with scientific articles, which would explain to some extent the similar results obtained. In MESINESP-P the best performing run was Fudan’s architecture, with a very balanced system between precision and recall. The system with the highest precision was the Roche’s one, which achieved a value of 0.525, the highest of all models and in all subtasks.

**Table 5**

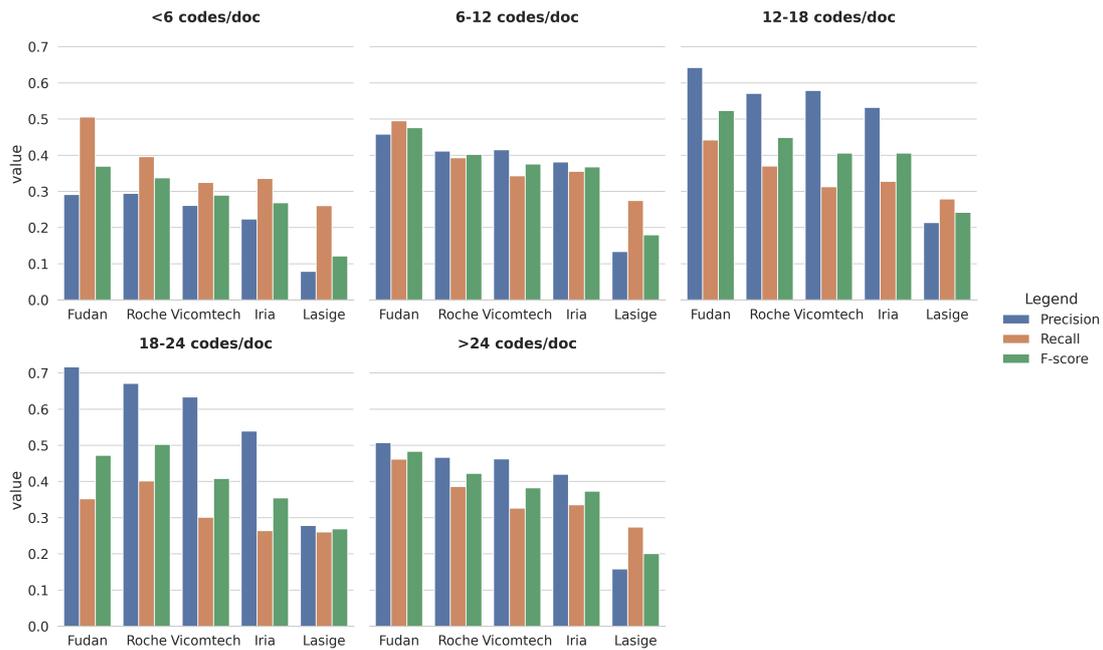
Performance of each of the models generated by participating teams in each of the subtracks.

Team	Country	Ref	System	MESINESP-L			MESINESP-T			MESINESP-P		
				MiF	MiP	MiR	MiF	MiP	MiR	MiF	MiP	MiR
Fudan University	China	-	BERTDeCS-CooMatInfer	0.4505	0.4791	0.4252	0.1095	0.1509	0.0859	0.4489	0.4462	0.4515
			BERTDeCS version 2	0.4798	0.5037	0.4581	<b>0.3640</b>	0.3666	0.3614	<b>0.4514</b>	0.4487	0.4541
			BERTDeCS version 3	0.4808	0.5047	0.4591	0.3630	0.3657	0.3604	0.4480	0.4454	0.4507
			BERTDeCS version 4	<b>0.4837</b>	0.5077	0.4618	0.3563	0.3589	0.3537	0.4514	0.4487	0.4541
			bertmesh-1	0.4808	0.5077	0.4591	0.3600	0.3626	0.3574	0.4489	0.4462	0.4515
Roche	Switzerland	[18]	bert_dna	0.3989	0.4662	0.3486	0.2710	0.3448	0.2232	0.2479	0.4143	0.1769
			pi_dna	0.4225	0.4667	0.3859	0.2781	0.3504	0.2305	0.3628	0.5250	0.2772
			pi_dna_2	0.3978	0.4520	0.3551	0.2680	0.4004	0.2015	-	-	-
			pi_dna_3	0.4027	0.4348	0.3750	-	-	-	-	-	-
			bert_dna_2	0.3962	0.4820	0.3364	0.2383	0.3754	0.1746	0.2479	0.4143	0.1769
Lasige-TEAM (Universidade de Lisboa)	Portugal	[23]	LASIGE_BioTM_1	0.2007	0.1584	0.2738	-	-	-	-	-	-
			LASIGE_BioTM_2	0.1886	0.1489	0.2573	-	-	-	-	-	-
			clinical_trials_1.0	-	-	-	0.0679	0.0575	0.0828	-	-	-
			clinical_trials_0.25	-	-	-	0.0686	0.0581	0.0838	-	-	-
			patents_1.0	-	-	-	-	-	-	0.0314	0.0239	0.0459
Vicomtech	Spain	[17]	Classifier	0.3825	0.4622	0.3262	0.2485	0.2721	0.2287	0.1968	0.2700	0.1548
			CSSClassifier025	0.3823	0.4509	0.3318	0.2819	0.2933	0.2715	0.2834	0.3188	0.2551
			CSSClassifier035	0.3801	0.4710	0.3186	0.2810	0.2888	0.2736	0.2651	0.2547	0.2764
			LabelGlosses01	0.3704	0.4526	0.3134	0.2807	0.2949	0.2678	0.2908	0.3596	0.2440
			LabelGlosses02	0.3746	0.4560	0.3179	-	-	-	0.2921	0.3890	0.2338
Iria (Uni Vigo, Uni. Coruña)	Spain		iria-1	0.3406	0.3641	0.3199	0.2454	0.2289	0.2644	0.1871	0.1926	0.1820
			iria-2	0.3389	0.3622	0.3185	-	-	-	0.3203	0.3657	0.2849
			iria-3	0.2537	0.2729	0.2369	0.1562	0.1419	0.1736	0.0793	0.0822	0.0765
			iria-4	0.3656	0.3909	0.3435	0.2003	0.1868	0.2158	0.2169	0.2232	0.2109
			iria-mix	0.3725	0.4193	0.3351	0.2003	0.1868	0.2158	0.2542	0.2750	0.2364
Universidad de Chile	Chile	-	tf-idf-model	0.1335	0.1405	0.1271	-	-	-	-	-	
YMCA University	India	-	AnujTagging	0.0631	0.0666	0.0600	-	-	-	-	-	-
			Anuj_ml	-	-	-	0.0019	0.0020	0.0018	-	-	-
			Anuj_NLP	0.0035	0.0053	0.0026	-	-	-	-	-	-
			Anuj_Ensemble	-	-	-	-	-	-	0.0389	0.0387	0.0391
<b>Baseline</b>				0.2876	0.2335	0.3746	0.1288	0.0781	0.3678	0.2992	0.4293	0.2296

### 3.3. Analysis

#### Does the performance of the models change depending on the number of document descriptors?

Gold standard documents have variability in the number of assigned codes. Some of the participating teams, such as Fudan University and Universidade de Lisboa, made the design decision of predicting a fixed number of codes for every document, which could affect performance when the actual number of descriptors is far from that design constant. Figure 5 shows the performance metrics of the best of the models for each team in MESINESP-L, splitting the test set into groups according to the number of codes manually assigned.



**Figure 5:** Performance of models for Gold Standard subsets generated using the number of associated codes.

The highest model performance is obtained for records containing between 12 and 24 codes. In general terms the most difficult records to classify are those with the lowest number of associated descriptors, although the sample of these documents is smaller than the rest of the groups. The accuracy of the models is substantially higher in the 18-24 code group, with precision values up to 0.7 in the case of the winning team. However, Fudan’s model, whose output was limited to 10 labels, drops substantially in performance for documents with less than 6 codes, although it maintains its clear lead in the other groups.

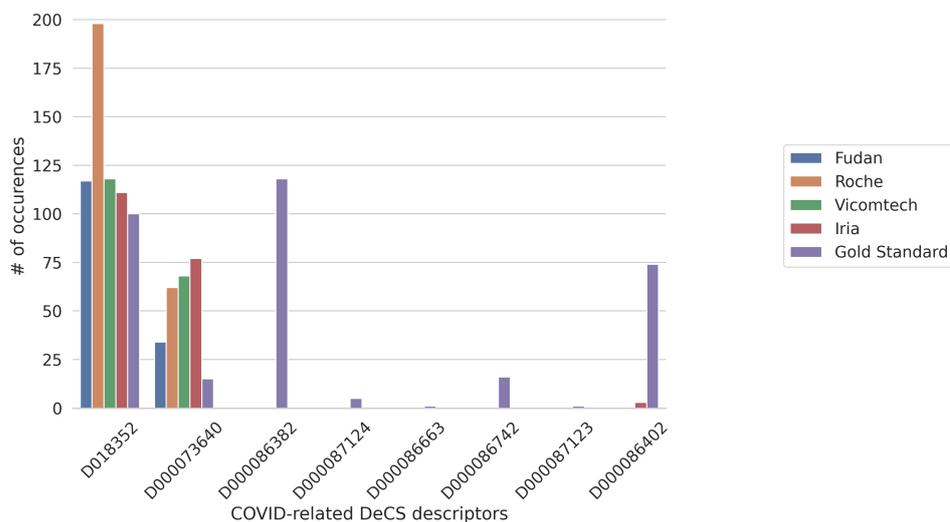
### How do systems behave with COVID-related descriptors?

Indexing tasks are critical in pandemic scenarios where it is necessary to assign descriptors to documents in order to retrieve them from databases. One of the analyses we have done of the participants’ predictions is focused on how the systems developed work with COVID-19 related records.

Although the training data were indexed with the COVID terms available in DeCS 2020, the annotation process of the test data was performed with the descriptors that will be incorporated in the 2021 version. This update of terms made it possible to test whether the systems developed by the participants were robust to new labels not present in the training resources, a topic of growing interest in the field [28, 29, 30, 31].

To carry out this analysis, we selected records that addressed COVID topics. The training corpus has the items indexed with the COVID terms recommended in 2020, namely "D018352" and "D000073640". In contrast, the test set (Gold Standard) was annotated with the new, and

much more specific, descriptors. Figure 6<sup>13</sup> shows the occurrences of each COVID descriptor in the best-score participants' model with respect to the gold standard for MESINESP-L. Iria team was the only team able to detect these more specific terms, with 3 occurrences of the term "D000086402", but in general, none of the systems were able to detect these not-previously-seen terms, which opens the door to propose the implementation of automatic indexing systems that know how to predict previously unseen labels.



**Figure 6:** Number of occurrences of COVID descriptors in participants' predictions for the test set.

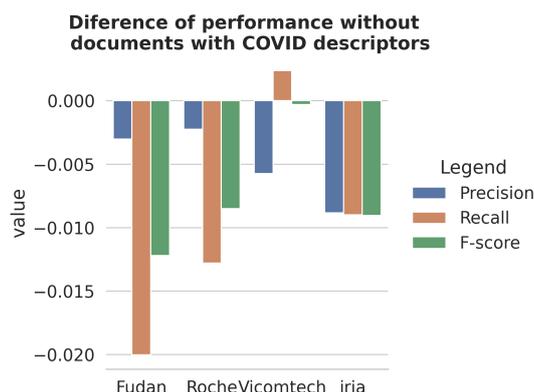
This ineffective assignment of COVID-related DeCS codes that were not present in the training corpus may significantly decrease model performance, given that a high proportion of the test set (119/491) had these types of descriptors. To assess the effect of these descriptors on overall performance, we recalculated model performance by considering only documents that did not have the new COVID codes.

The results are shown in Figure 7. When models are evaluated only with documents containing labels seen in the training set, the performance decreases minimally in general. This means that although there is a significant portion of documents in the test set whose COVID labels are not being correctly assigned, they are not being classified worse than the rest of the documents in which all the descriptors were present in the training set.

### Which descriptors do the systems predict best and worst?

MESINESP2 is an extreme multi-label labelling challenge. We provide a set of thousands of labels to be assigned to documents. There may be codes that are very easy for systems to identify and others that are difficult to assign correctly. Figure 8 presents the 4 descriptors that

<sup>13</sup>Descriptors reference: *D018352* (Coronavirus Infections), *D000073640* (Betacoronavirus), *D000086382* (COVID-19), *D000086663* (COVID-19 vaccines), *D000086742* (COVID-19 test), *D000087123* (nucleic acid test for COVID-19), *D000086402* (SARS-CoV-2)



**Figure 7:** Difference in model performance when removing COVID-related documents.

have been best identified by the top systems of the 4 most competitive teams in MESINESP-L, MESINESP-T and MESINEP-P. Conversely, Figure 9 shows the most poorly predicted labels by models, also showing the number of times they appeared in the test set.

For subtrack 1, the codes with the best success rate are repeated in all systems: "D006801", *Humanos*(Humans); "D018352", *Infecciones por Coronavirus* (Coronavirus infections); D008297, *Masculino* (Male); and D0052650, *Femenino* (Female). The human descriptor is one of the most frequent codes in the BvSalud database, so it was expected that systems would be able to detect this descriptor with ease and accuracy in the test set. Male and Female descriptors are special nodes within the DeCS ontology, since they are at the top of the hierarchical structure, without having any children. Finally, systems were able to detect with high precision the descriptor "Coronavirus infections", probably because the systems tended to over-assign this term, as seen in Figure 6. Regarding the worst predicted labels, as seen above, none of the systems was able to correctly predict the new COVID descriptors with which the test set was manually indexed. Additionally the code "D013812" (Therapeutic) and "D002363" (Case histories) were not assigned with adequate frequency, with the exception of the system of the Fudan team.

Clinical trials models show similar behaviour to those of scientific literature. Almost all of the descriptors "D006801" and "D018352" were correctly recognised. In addition, the descriptors "D011024" and "D016896", corresponding to the terms Viral Pneumonia and Treatment Outcome, showed very good accuracy. These codes have the peculiarity of being in the upper part of the structure, thus continuing the trend observed at MESINESP-L results. Concerning terms that were not correctly classified, once again we found that COVID descriptors have not been indexed correctly. This is an expected result after previous analyses and due to the lack of data labelled with these codes. We also have general terms, such as Therapeutics, but also the code "D016449" (Randomized Controlled Trial) which are very specific and models were not able to understand the context of the abstract to understand this concept.

For patents, the general trend observed so far in relation to well-detected terms changes. On the one hand, the Fudan and Roche models perform very well in detecting the Patent ("D020490") and Human ("D006801") descriptors, while the Vicomtech and Iria models seemed to have some

issues in correctly assigning these codes. Conversely, the Vicomtech and Iria models detect the descriptor Therapeutica ("D013812") very well, while the Fudan and Roche models did slightly worse. Although in general terms the systems are able to assign specific labels such as Oligonucleotides antisense ("D016376") or antineoplastic agents ("D010300"), they perform worse for more ontology-specific labels, although they are conceptually more generic (such as organic chemistry and Drug Compounding)

Track 1 - Best					Track 2 - Best					Track 3 - Best				
1	D006801 302/302	D006801 295/302	D006801 302/302	D006801 297/302	1	D006801 108/108	D006801 108/108	D006801 108/108	D006801 108/108	1	D020490 118/118	D020490 118/118	D013812 69/74	D013812 65/74
2	D018352 99/100	D018352 99/100	D018352 100/100	D018352 93/100	2	D018352 64/64	D018352 83/84	D018352 62/64	D016896 45/53	2	D006801 21/21	D006801 20/21	D000970 15/20	D010300 5/6
3	D008297 88/97	D008297 73/97	D008297 75/97	D008297 92/97	3	D011024 43/43	D011024 38/43	D011024 38/43	D000970 42/42	3	D000970 11/20	D000970 11/20	D019636 6/9	D019636 5/9
4	D005260 78/88	D005260 64/88	D005260 74/88	D005260 82/88	4	D016896 42/53	D010919 20/42	D016896 37/53	D018352 41/84	4	D016376 8/9	D010300 5/6	D000900 6/11	D000998 5/8
	Fudan	Roche	Vicomtech	Iria		Fudan	Roche	Vicomtech	Iria		Fudan	Roche	Vicomtech	Iria

**Figure 8:** Descriptors best predicted by the best model for each team in each subtrack. Each block contains the DeCS code, the number of times the model has predicted that code correctly and the total number of times the code appeared in the corpus.

Track 1 - Worst					Track 2 - Worst					Track 3 - Worst				
1	D000086382 0/118	D000086382 0/118	D000086382 0/118	D000086382 0/118	1	D013812 0/105	D013812 0/105	D013812 0/105	D013812 0/105	1	D013812 0/74	D013812 3/74	D020490 0/118	D020490 0/118
2	D000086402 0/74	D000086402 0/74	D000086402 0/74	D000086402 3/74	2	D016449 0/96	D016449 0/96	D016449 0/96	D016449 0/96	2	D004339 2/29	D004339 0/29	D004339 6/29	D004339 0/29
3	D013812 2/58	D013812 0/58	D002363 1/58	D002363 0/58	3	D000086382 0/89	D000086382 0/89	D000086382 0/89	D000086382 0/89	3	D002625 0/10	D002625 0/10	D006801 5/21	D000970 0/20
4	D002363 6/58	D002363 1/58	D013812 1/58	D013812 1/58	4	DDCS028570 0/87	DDCS028570 2/87	D000086402 0/80	D000086402 0/80	4	D006571 0/10	D006571 0/10	D002625 0/10	D006801 5/21
	Fudan	Roche	Vicomtech	Iria		Fudan	Roche	Vicomtech	Iria		Fudan	Roche	Vicomtech	Iria

**Figure 9:** Descriptors worst predicted by the best model for each team in each subtrack. Each block contains the DeCS code, the number of times the model has predicted that code correctly and the total number of times the code appeared in the corpus.

### 3.4. Silver standard generation

Within the test sets for each of the tasks, a set of records that were not used for evaluating the models was included in order to generate a silver standard.

This silver standard has been published in the task data repository, and contains two separate sections. On the one hand, the union of the labels of the best model of each participating team has been calculated, as long as this model had obtained at least an F-score of 0.2. On the other hand, the predictions of the best models of each participant have been included individually and anonymised.

The silver standard contains a set of 8642 scientific articles, 1537 text sections from Clinical Practice Guidelines, a set of 8458 text segments from Medication Data Sheets, 461 clinical trials from REEC and 5170 patents. The summary statistics of the silver standard generated are shown in Table 6.

**Table 6**

Summary statistics of the silver standard corpora generated in MESINESP.

Silver Standard Corpus	Docs	DeCS	Unique DeCS	Tokens	Avg. DeCS/doc	Avg.token/doc
Scientific papers	8642	257123	11557	~1.7M	29.75 (3.44)	198.31 (63.21)
Clinical Practice Guidelines	1537	46263	4018	132264	30.10 (3.51)	86.05 (87.66)
Medication data sheets	8458	190701	1609	~9.1M	22.55 (2.78)	1076.93 (423.71)
Clinical Trials	461	12277	4158	508793	26.63 (3.69)	1103.67 (558.31)
Patents	5170	101775	7319	641624	19.69 (6.05)	124.11 (145.77)

## 4. Discussion

### Real application of Spanish semantic indexing models

Despite the improvement of the F-score results by 0.06 with respect to last year, and considering the positive evolution of the results of BioASQ Task 9a, we believe that there is still wide scope for improvement in the semantic indexing of biomedical documentation in Spanish.

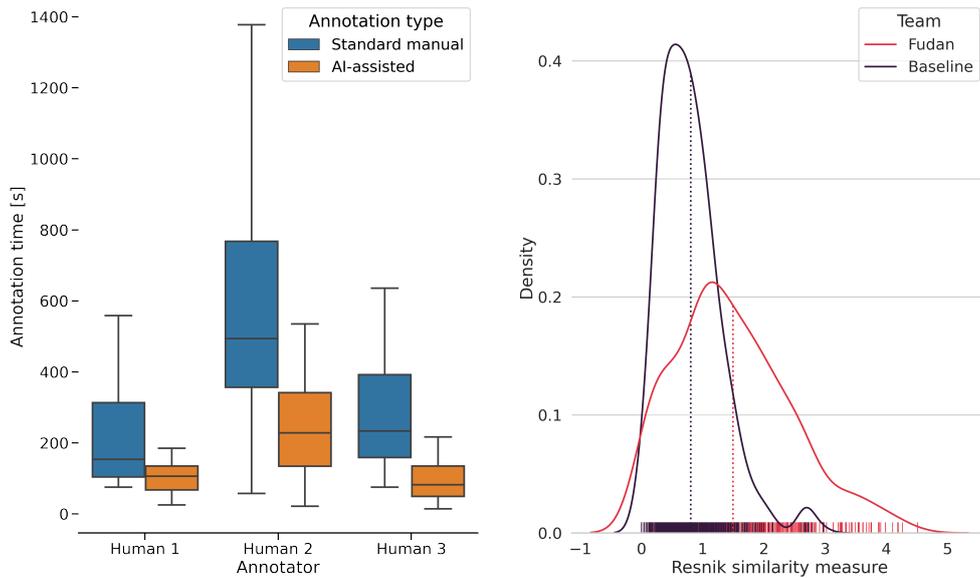
The applicability of semantic indexing models in Spanish is feasible and their use to assist manual indexing initiatives seems to be in reach, but with some limitations. The precision and recall values are not high enough to incorporate the predicted descriptors into databases without a validation process, as there are a significant number of codes that would not be added and would be poorly predicted. This validation process, which could be seen as an AI-assisted document annotation process, would speed up the process of including DeCS terms in documents and facilitate document retrieval. These automatic indexing systems could be incorporated into indexing assistance tools to pre-index documents for validation by expert annotators. Providing a ranked list of predicted codes would allow more flexibility in case of using totally automated predictions.

Using the ASIT tool, which allows annotation time logging, we measured the average time taken by three expert indexers to annotate 30 documents in a traditional manner and with the descriptors predicted by *BERTDeCS version4* system on a subset of the documents from the test set. Document pre-indexing has reduced annotation times by **more than half**. Figure 10a shows box plots of the indexing times of each of the annotators. When experts are confronted with a pre-indexed document their annotation times decrease substantially, with a maximum time reduction of one third of the total in the case of the A7 annotator.

### Cross-corpus training

The task has resulted in models with disparate quality for each type of document. The indexing models for scientific literature are the best in terms of performance, followed by those for patents and clinical trials, but how can be explained the differences in the quality of the systems if they have been trained with similar corpora?

When preparing MESINESP corpora, and in order to facilitate the cross-corpus training process, semantic similarity models were applied for the selection of documents to be annotated by experts. Despite using a document selection criterion, we believe that knowing the features that make a document well indexed would make it easier to know in advance in which systems



(a) Improvement of annotation time with an AI-assisted system compared to standard manual annotation. (b) Histogram showing the distribution of Resnik's similarity measure between the documents predicted by the baseline and Fudan with respect to the Gold Standard.

**Figure 10:**

semantic indexing would work better. This would make it possible to focus research on this type of document and even, if a sufficiently high indexing quality could be achieved, to train a model with literature data and index documents with similar properties without a new annotation, training and evaluation process.

For example, one of the documents that would be interesting to index would be the Electronic Health Records. We have launched automatic indexing systems on a subset of EHRs and obtained the metrics shown in Table 7. Since we do not know the common properties between scientific literature documents and EHRs, it would be very risky to use these models on this type of documents. But the annotation process needed to train indexing systems might discourage attempts to implement semantic indexing in this type of document.

**Table 7**

Statistics of DeCS descriptors found in different types of Electronic Health Records

EHR corpus	Docs	DeCS	Unique DeCS	Tokens	Avg. DeCS/doc	Avg.token/doc
<b>Clinical course</b>	1000	85562	4252	~3.7M	85.56 (90.37)	3780.47 (7023.16)
<b>Discharge summaries</b>	1000	7872	1242	83937	7.87 (11.43)	83.94 (142.46)
<b>Death reports</b>	42	312	182	2735	7.43 (14.18)	65.12 (158.73)
<b>Radiology reports</b>	461	3894	554	50420	1.30 (3.08)	16.81 (36.46)

## Quality in predictions due to the hierarchical structure

Semantic Indexing with DeCS is a complex task. As we have seen, DeCS is a constantly changing terminology, with very specific and some infrequently used terms, which makes model training challenging. The use of flat metrics, such as the F-score, to evaluate hierarchical models is too restrictive. The F-score penalises too much the failure of a prediction, as it does not take into account the distance between the predicted and the true code within the ontology. For example, in the predictions made by the participants we found many documents in which the systems, instead of matching the exact codes, had predicted the direct parent of the descriptors. These documents, despite not having a correct prediction, may have captured the content of the documents in some way, but they are not evaluated in the fairest way. Traditional ranking-based metrics, such as precision@k and nDCG [32], are also unable to capture this hierarchical structure, although they may be able to assess the presence of parent and child terms in the predictions. An alternative to this issue would be the use of other type of metrics such as Knowledge-based semantic similarity metrics, which measure the degree of common information between two documents by quantifying the distance among the concepts covered in the texts when mapped into an ontology [33]. Semantic similarity has been used for validating results from biomedical studies with specific ontologies such as Gene Ontology [34], and metrics such as Resnik's, Lin's, and Jiang [35] may be useful to evaluate the degree of similarity between gold standard and predictions obtained with a semantic indexing model.

To calculate Resnik's similarity at document level, we calculate the distances between the predicted and true labels for the same document. The higher the Resnik similarity value, the closer the predicted codes are to the true labels. This allows to consider systems that have predicted nearby codes within the hierarchical tree of the controlled vocabulary. Figure 10b shows the similarity distribution between predictions of the baseline and the winning team with respect to the MESINESP-L Gold Standard. As expected, Fudan's model obtains a considerably higher average similarity value than the baseline of the task. Despite some overlap between the two distributions, the Fudan model was able to assign codes to the documents.

## 5. Conclusions

This document has shown the overview of the MESINESP2 task within the 9th BioASQ Challenge (CLEF 2021). This time, the task has been focused on indexing documents concerning scientific literature, clinical cases and patents.

A brief analysis of the participant models and performance for each of the tasks has been introduced. Once again, the trend of using deep neural approaches is more evident, with a large percentage of the participants using Multilingual-BERT for text representation.

The team from Fudan University, winner in each of the subtracks, was able to improve the state of the art for indexing scientific literature in Spanish defined last year, and has been able to define it in the indexing of clinical trials and patents.

Despite the positive evolution in the performance of the MESINESP task, there is a drop in the performance of the model with respect to the English task. This reduction may be related to the volume of data available for model training, so we should evaluate whether following a multilingual approach, incorporating languages such as French, Russian or Chinese, could

favour the impact, interest and performance improvement of the systems by having a larger volume of labelled data.

We have experienced a rapid evolution in the quality of existing approaches to multi-label learning in recent years, for example from the first edition of BioASQ to the current edition we have experienced an increase in F-score of 0.12. However, when using such systems in real environments, the fact that the controlled vocabularies used as labels are updated periodically has been overlooked. In addition, the updating of database descriptors is not done instantaneously, which makes it difficult for the systems to be capable of new labels not previously considered.

Future evaluation setting for this or similar tasks could benefit from considering more interactive evaluation settings with the human (indexer) in the loop, similar to the BioCreative Interactive tracks [36] or the technical evaluation and integration and access of predictive systems through some meta-server settings [37].

## 6. Acknowledgements

The MESINESP task is sponsored by the Spanish Plan for advancement of Language Technologies (Plan TL). We thank Eulalia Farré, Antonio Miranda and Salvador Lima from Text Mining Unit at the Barcelona Supercomputing Center for giving advice and valuable opinions to organise MESINESP shared-task and input during the data preparation. Thanks to Alejandro Asensio (BSC) for his help in setting up the annotation tool, to BIREME indexers (Regina Chiquetto, Lucilena Bragion and Sueli Mitiko) for providing feedback about quality of manual annotations, and to Felipe Soares and Ankush Rana for helping in the data collection process.

## References

- [1] D. Torres-Salinas, N. Robinson-Garcia, F. van Schalkwyk, G. F. Nane, P. Castillo-Valdivieso, The growth of covid-19 scientific literature: A forecast analysis of different daily time series in specific settings, arXiv preprint arXiv:2101.12455 (2021).
- [2] R. Bawden, K. B. Cohen, C. Grozea, A. J. Yepes, M. Kittner, M. Krallinger, N. Mah, A. Neveol, M. Neves, F. Soares, et al., Findings of the wmt 2019 biomedical translation shared task: Evaluation for medline abstracts and biomedical terminologies, in: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), 2019, pp. 29–53.
- [3] S. MacAvaney, A. Cohan, N. Goharian, Sledge: A simple yet effective baseline for covid-19 scientific knowledge search, 2020. arXiv:2005.02365.
- [4] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al., Cord-19: The covid-19 open research dataset, ArXiv (2020).
- [5] E. Zhang, N. Gupta, R. Nogueira, K. Cho, J. Lin, Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned, 2020. arXiv:2004.05125.
- [6] Covid research: a year of scientific milestones, 2021. URL: <https://www.nature.com/articles/d41586-020-00502-w>.
- [7] W. Hersh, E. Voorhees, Trec genomics special issue overview, 2009.
- [8] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers,

- D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, *BMC bioinformatics* 16 (2015) 1–28.
- [9] M. Krallinger, A. Gonzalez-Agirre, A. Asensio, MESINESP: Medical Semantic Indexing in Spanish - Development dataset, 2020. URL: <https://doi.org/10.5281/zenodo.3746596>. doi:10.5281/zenodo.3746596, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
- [10] L. Gasco, REECapi: The unofficial library for accessing the REEC API, 2021. URL: <https://doi.org/10.5281/zenodo.4882820>. doi:10.5281/zenodo.4882820.
- [11] D. Eisinger, G. Tsatsaronis, M. Bundschuh, U. Wieneke, M. Schroeder, Automated patent categorization and guided patent search using ipc as inspired by mesh and pubmed, in: *Journal of biomedical semantics*, volume 4, Springer, 2013, pp. 1–23.
- [12] M. Krallinger, O. Rabal, A. Lourenço, M. P. Perez, G. P. Rodriguez, M. Vazquez, F. Leitner, J. Oyarzabal, A. Valencia, Overview of the chemdner patents task, in: *Proceedings of the fifth BioCreative challenge evaluation workshop*, 2015, pp. 63–75.
- [13] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, M. Krallinger, The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora, in: *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, 2021, pp. 13–20.
- [14] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020, in: *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, 2020.
- [15] 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.
- [16] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, S. Zhu, Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification, *arXiv preprint arXiv:1811.01727* (2018).
- [17] A. García-Pablos, N. Perez, M. Cuadros, Vicomtech at MESINESP2: BERT-based Multi-label Classification Models for Biomedical Text Indexing (2021).
- [18] Y. Huang, G. Buse, K. Abdullatif, A. Ozgur, E. Ozkirimli, Pidna at bioasq mesinesp: Hybrid semanticindexing for biomedical articles in spanish (2021).
- [19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [20] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: *PML4DC at ICLR 2020*, 2020.

- [21] F. J. Ribadas, L. M. De Campos, V. M. Darriba, A. E. Romero, Cole and utai at bioasq 2015: experiments with similarity based descriptor assignment, in: CEUR Workshop Proceedings, volume 1391, 2015.
- [22] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, arXiv preprint arXiv:2004.09813 (2020).
- [23] P. Ruas, V. D. T. Andrade, F. M. Couto, LASIGE-BioTM at MESINESP2: entity linking with semantic similarity and extreme multi-label classification on Spanish biomedical documents (2021).
- [24] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. S. Dhillon, Taming pretrained transformers for extreme multi-label text classification, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3163–3171.
- [25] A. Nentidis, E. Katsimpras, Georgios and Vandorou, A. Krithara, L. Gasco, G. . Krallinger, Martin and Paliouras, Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. (2021).
- [26] C. Rodriguez-Penagos, A. Nentidis, A. Gonzalez-Agirre, A. Asensio, J. Armengol-Estap e, A. Krithara, M. Villegas, G. Paliouras, M. Krallinger, Overview of MESINESP8, a Spanish Medical Semantic Indexing Task within BioASQ 2020 (2020).
- [27] D. Molla, D. Seneviratne, Overview of the 2018 alta shared task: Classifying patent applications, in: Proceedings of the Australasian Language Technology Association Workshop 2018, 2018, pp. 84–88.
- [28] A. Pham, R. Raich, X. Fern, J. P. Arriaga, Multi-instance multi-label learning in the presence of novel class instances, in: International Conference on Machine Learning, PMLR, 2015.
- [29] Y. Zhu, K. M. Ting, Z.-H. Zhou, Multi-label learning with emerging new labels, IEEE Transactions on Knowledge and Data Engineering 30 (2018) 1901–1914.
- [30] Y. Zhu, K. M. Ting, Z.-H. Zhou, Discover multiple novel labels in multi-instance multi-label learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 31, 2017.
- [31] J. Huang, L. Xu, K. Qian, J. Wang, K. Yamanishi, Multi-label learning with missing and completely unobserved labels, Data Mining and Knowledge Discovery 35 (2021) 1061–1086.
- [32] Y. Wang, L. Wang, Y. Li, D. He, W. Chen, T.-Y. Liu, A theoretical analysis of ndcg ranking measures, in: Proceedings of the 26th annual conference on learning theory (COLT 2013), volume 8, Citeseer, 2013, p. 6.
- [33] F. Couto, A. Lamurias, Semantic similarity definition, Encyclopedia of bioinformatics and computational biology 1 (2019).
- [34] G. O. Consortium, The gene ontology (go) database and informatics resource, Nucleic acids research 32 (2004) D258–D261.
- [35] X. Guo, R. Liu, C. D. Shriver, H. Hu, M. N. Liebman, Assessing semantic similarity measures for the characterization of human regulatory pathways, Bioinformatics 22 (2006) 967–973.
- [36] C. N. Arighi, P. M. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, A. Chatr-Aryamontri, S. Clematide, P. Gaudet, M. G. Giglio, I. Harrow, et al., Biocreative iii interactive task: an overview, BMC bioinformatics 12 (2011) 1–21.
- [37] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J. Hakenberg, C. Plake, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-C. Hung, W. W. Lau, et al., Introducing meta-services for biomedical information extraction, Genome biology 9 (2008) 1–11.