Overview of MESINESP8, a Spanish Medical Semantic Indexing Task within BioASQ 2020

Carlos Rodriguez-Penagos[1], Anastasios Nentidis[2,3], Aitor Gonzalez-Agirre[1], Alejandro Asensio[1], Jordi Armengol-Estapé[1], Anastasia Krithara[2], Marta Villegas[1], Georgios Paliouras[2], and Martin Krallinger[1]

[1] Barcelona Supercomputing Center, Barcelona, Spain
{martin.krallinger, carlos.rodriguez1, marta.villegas}@bsc.es
[2] National Center for Scientific Research "Demokritos", Athens, Greece
{tasosnent, akrithara, bogas.ko, paliourg}@iit.demokritos.gr
Aristotle University of Thessaloniki, Thessaloniki, Greece
[3] National and Kapodistrian University of Athens, Athens, Greece

Abstract. In this paper, we present an overview of the novel MESINESP Task on medical semantic indexing in Spanish within the eighth edition of the BioASQ challenge, which ran as a lab in the Conference and Labs of the Evaluation Forum (CLEF) 2020. BioASQ is a series of challenges aiming at the promotion of systems and methodologies for large-scale biomedical semantic indexing and question answering. MESINESP represents the first attempt to generate resources for the development and evaluation semantic indexing strategies specialized on health-related content in Spanish. We have generate several publicly accessible Gold Standard collections of manually indexed content covering medical literature, clinical trials and health project descriptions associated to controlled terminologies in the form of the hierarchical DeCS vocabulary. Manual indexing of MESINESP documents was carried out by professional medical literature indexers. They used an indexing web interface particularly adapted for this task. The results obtained by participating teams was promising, showing that training data of semantically indexed medical literature can also serve to implement automatic indexing systems that assist manual indexing of other types of documents like clinical trials. MESINESP corpus: https://zenodo.org/record/3746596.Xo9WT$_IzaFA$

Keywords: Biomedical knowledge · Semantic Indexing · Question Answering

1 Introduction

There is a pressing need to facilitate more sophisticated search queries to retrieve relevant health-related content, in particular medical publications. This became

clear in case of the recent COVID-19 pandemic, where experts required finding medical articles describing certain aspects of this novel disease such as symptoms, co-morbidities or treatment related aspects [16, 14, 3]. Moreover, highly specialized information needs on complex subjects, for instance to select important articles to elaborate publications such as systematic reviews do require complex semantic search capabilities [8]. With the rapid accumulation of biomedical and clinical research publications, healthcare experts are increasingly relying on the results of so-called indexing initiatives to build more sophisticated semantic search queries that incorporate indexed terms from structured controlled vocabularies. Figure 1 provides a summary of the importance of semantic indexing and retrieval systems of medical literature content from the perspective of various stakeholders and end users.

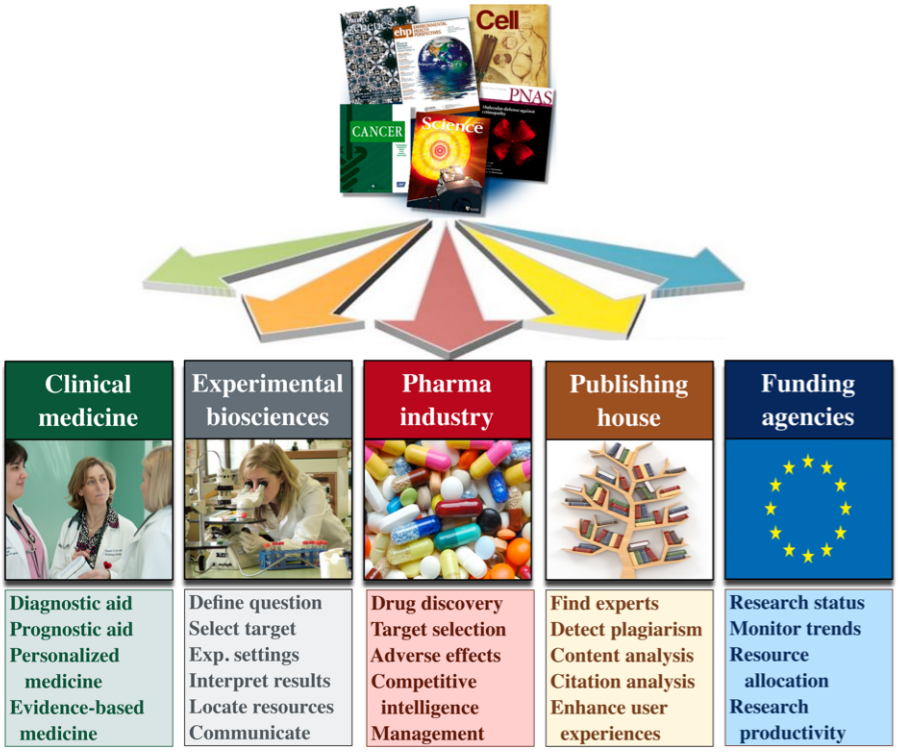

Fig. 1. Importance of semantic indexing of medical literature.

This paper aims at presenting the used data, settings and results of the MESINESP shared task, which was part of the CLEF-BioASQ 2020 challenge. Towards this direction we provide an overview of the MESINESP shared task and the corresponding corpus and additional data resources prepared for this track. We present a brief overview of the systems developed by the participating

teams for the different tasks. Detailed descriptions for some of the systems are available in the proceedings of the lab. We focus on evaluating the performance of semantic indexing strategies participating in this track systems using state-of-the-art evaluation measures. Finally we sum up the conclusion and future outlook of the MESINESP effort.

This year, the eighth version of the BioASQ challenge comprised three tasks: (1) a large-scale biomedical semantic indexing task (task 8a), (2) a biomedical question answering task (task 8b), both considering documents in English, and (3) a new task on medical semantic indexing in Spanish (task MESINESP). A detailed overview of these tasks and the general structure of BioASQ are available in [19]. In this paper, we describe the new MESINESP task on semantic indexing of medical content written in Spanish (medical literature abstracts, clinical trial summaries and health-related project descriptions), which was introduced this year [11], providing statistics about the dataset developed for it.

2 Data and Resources

There is a pressing need to improve the access to information comprised in health and biomedicine related documents, not only by professional medical users buy also by researches, public healthcare decision makers, pharma industry and particularly by patients. Currently, most of the Biomedical NLP and IR research is being done on content in English, despite the fact that a large volume of medical documents is published in other languages including Spanish. Key resources like PubMed focus primarily on data in English, but it provides outlinks also to articles originally published in Spanish. For English, Task 8a's aim was to classify articles from the PubMed/MedLine[4] digital library into concepts of the MeSH hierarchy. In particular, new PubMed articles that are not yet annotated by the indexers in NLM are gathered to form the test sets for the evaluation of the participating systems. The performance of the participating systems was calculated using standard flat information retrieval measures, as well as, hierarchical ones, when the annotations from the NLM indexers become available. Task 8a provided for training a dataset of 14,913,939 articles with 12.68 labels per article.

The main aim of MESINESP is to promote the development of semantic indexing tools of practical relevance of non-English content, determining the current-state-of-the art, identifying challenges and comparing the strategies and results to those published for English data.

[4] https://pubmed.ncbi.nlm.nih.gov/

MESINESP is focused on healthcare content in Spanish: IBECS[5], LILACS[6], REEC [7] and FIS-ISCIII [8]. In this task, the participants were asked to classify new IBECS and LILACS documents in Spanish. The classes come from the DeCS vocabulary [9] which was originally developed from the MeSH hierarchy. At present, this annotation is done manually, being costly and labor-intensive. Thus manual semantic indexing of Spanish medical literature would greatly benefit from a more systematic indexing strategy or the availability of manual indexing assistance software. Due to the burden of manual indexing, there is also a considerable delay from the date a record is published until is manually indexed, specially when compared to indexing speed of other databases like PubMed. The MESINESP task was promoted within the efforts of the Spanish Government's Plan for Promoting Language Technologies (Plan TL), that aims to promote the development of natural language processing, machine translation and conversational systems in Spanish and co-official languages in Spain.

2.1 Description of the datasets for MESINESP, and the annotation effort

First, we performed a web crawling against https://pesquisa.bvsalud.org/ (IBECS and LILACS) to obtain 1.1 million articles, extracting the title and the abstract (not the full text) among other article data such as journal and date of publication.

A training dataset [10] was released with 369,368 articles manually annotated with DeCS codes (*Descriptores en Ciencias de la Salud*, derived and extended from MeSH terms)[11]. Then, 1500 articles, published from 2018 onwards, were selected and annotated by 7 experts in the field of clinical text indexing with DeCS codes. Figure 2 shows a screen shot of the interface that was used for the first phase of the manual Gold Standard semantic indexing process.

Those articles have been distributed in a way that each article is annotated, at least, by two different annotators. The first phase consisted in adding DeCS codes to each document, and the second phase was about validating those DeCS codes viewing suggestions from codes added by other annotators on that same document (simple automatic DeCS term gazetteer look-up suggestions were also

[5] IBECS includes bibliographic references from scientific articles in health sciences published in Spanish journals. http://ibecs.isciii.es

[6] 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

[7] Registro Español de Estudios Clínicos, a database containing summaries of clinical trials https://reec.aemps.es/reec/public/web.html

[8] public healthcare project proposal summaries (Proyectos de Investigación en Salud, diseñado por el Instituto de Salud Carlos III, ISCIII) https://portalfis.isciii.es/es/Paginas/inicio.aspx

[9] http://decs.bvs.br/I/decsweb2019.htm

[10] https://zenodo.org/record/3826492

[11] 29,716 come directly from MeSH and 4,402 are exclusive to DeCS

Fig. 2. DeCS annotation tool developed at BSC for MESINESP: manual code assignment phase.

shown). This process results in the Gold Standard manual corpus comprising the development and test set records. Figure 3 shows a screen short of the semantic indexing validation interface used during the corpus construction phase.

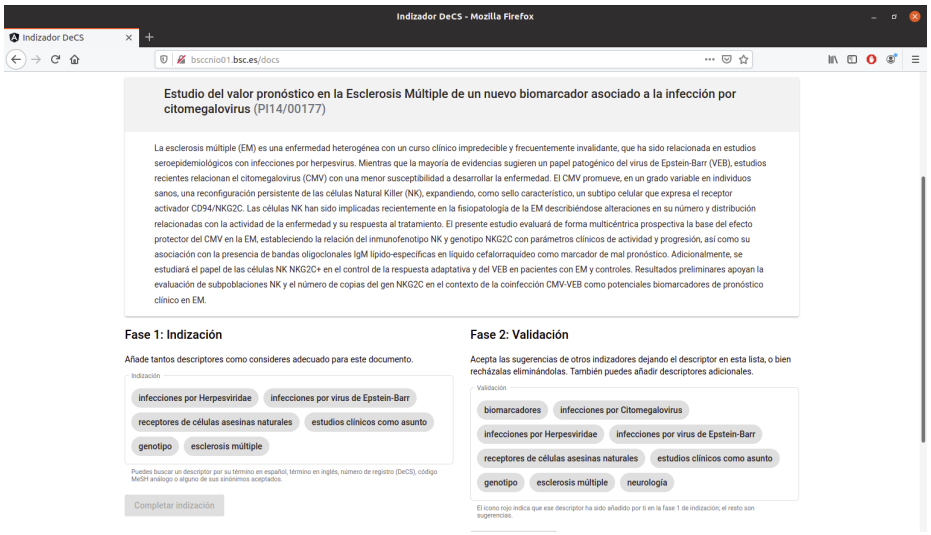

Fig. 3. Validation of DeCS annotation by two human annotators

A further additional background dataset was produced from diverse sources, including machine-translated text. This set did not have manual annotations but was distributed to teams to generate a silver standard corpus.

Consistently, the different collections averaged, per document, around 10 sentences, contained 13 DeCS codes, and 300 words, of which between 130 and 140 were unique ones.

Dataset	docs	DeCS	unique avg tokens	avg unique tokens	max.tokens	avg DeCS per doc	avg unique DeCS
Dev. (i)	750	6540	2756	295.2	141.1	3026	8.7	8.7
Dev. (u)	750	9847	3600	295.2	141.1	3026	13.1	13.1
TEST	938	12429	4061	273.8	135.3	2552	13.2	13.2
TRAIN	318658	2588925	23423	192.6	106.5	1725	8.1	7.2

Table 1. Core statistics for datasets used in MESINESP. (i) intersection, (u) union, Dev (development set)

In order to explore the diversity of content from this dataset, we generated clusters from the titles of semantically similar records from the background dataset distributed to the participants for automatic annotation. We used the lingo clustering algorithm from the Carrot Workbench project. [12]. The resulting 26 clusters are shown in ZZZ using the Foam visualization (see Figure 4).

Fig. 4. Content visualization of MESINESP evaluation dataset using the Lingo algorithm. Among subjects shown: Diabetis, AIDS, Suicide prevention,

[12] https://project.carrot2.org/

3 Results and participation

For the newly introduced MESINESP8 task, 6 teams from China, India, Portugal and Spain participated and results from 24 different systems were submitted.

The approaches were mostly the same as the ones used on the comparable English task (8a), and included KNN and Support Vector Machine classifiers, as well as deep learning frameworks like X-BERT and multilingual-BERT.

The LASIGE team from the University of Lisboa implemented a "X-BERT BioASQ" system that combines a solution based on Extreme Multi-Label Classification (XMLC) with a Named-Entity-Recognition (NER) tool. In particular, their system is based on X-BERT [5], an approach to scale BERT [7] to XMLC, combined with the use of the MER [6] tool to recognize MeSH terms in the abstracts of the articles. The system is structured into three steps. The first step is the semantic indexing of the labels into clusters using ELMo [13]; then a second step matches the indices using a Transformer architecture; and finally, the third step focuses on ranking the labels retrieved from the previous indices.

The Fudan University team also builds upon their previous "*AttentionXML*" [20] and "*DeepMeSH*" [12] systems as well their new "*BERTMeSH*" system, which are based on document to vector (d2v) and tf-idf feature embeddings, learning to rank (LTR) and DL-based extreme multi-label text classification, Attention Mechanisms and Probabilistic Label Trees (PLT) [9].

The Vigo and Grenada Universities "*Iria*" systems [15] implemented a multilabel k-NN classifier backed by an Apache Lucene indexing. In the official runs, only stemming and selected stem bigrams with high correlation were employed in citation representation and indexing. Finally, candidate subjects provided by the k-NN clasifier were enriched adding exact matches of subject labels taken from the abstract text using Apache UIMA ConceptMapper. For the MESINESP8 task runs, the k-NN approach remained the same. Several lingustically motivated text representations (content word lemmas, syntactic dependence triples, NP chunks) were tested using the Spanish models from the spaCy NLP toolkit to extract them from abstracts text.

System	Approach
Iria	bigrams, Luchene Index, k-NN, ensembles, UIMA ConceptMapper
Fudan University	AttentionXML with multilingual-BERT
Alara (UNED)	Frequency graph matching
Priberam	BERT based classifier, and SVM-rank ensemble
LASIGE	X-BERT, Transformers ELMo, MER

Table 2. Systems and approaches for Task MESINESP8. Systems for which no description was available at the time of writing are omitted.

A simple lookup system was provided as a baseline for the MESINESP task. This system extracts information from an annotated list. Then checks whether, in a set of text documents, the annotation are present. It basically gets the intersection between tokens in annotations and tokens in words. This simple approach obtains a MiF of 0.2695.

3.1 Evaluation metrics

Standard flat and hierarchical evaluation measures [2] were used for measuring the classification performance of the systems. In particular, the micro F-measure (MiF) and the Lowest Common Ancestor F-measure (LCA-F) were used to identify the winners for each batch [10].

The results in Task 8a show that in all test batches and for both flat and hierarchical measures, the best systems outperform the strong baselines. In particular, the "*dmiip_fdu*" systems from the Fudan University team achieve the best performance in all three batches of the task. More detailed results can be found in the online results page[13]. Comparing these results with the corresponding results from previous versions of the task, suggests that both the MTI baseline and the top performing systems keep improving through the years of the challenge.

In case of the MESINESP task, it seems that is was more difficult when compared to results obtained for data in English (i.e. Task 8a), but overall we believe the results were pretty good taking into account that the provided data collection were considerably smaller. One problem with the medical semantic concept indexing in Spanish, at least for diagnosis or disease related terms, is the uneven distribution and high variability. [1], but the Task results show that this fact does not prevent good performance by advanced implementations. Compared to the setting for English, the overall training dataset was not only significantly smaller, but also the track evaluation test data set contained also clinical trial summaries and healthcare project summaries. Moreover, in case of the provided training data, two different indexing settings were used by the literature databases: IBECS has a more centralized manual indexing contracting system, while in case of LILACS a number of records were indexed in a sort of distributed community human indexer effort. The training set contained 23,423 unique codes, while the 911 articles in the evaluation set contained almost 4,000 correct DeCS codes. The best predictions, by Fudan University, scored a MIF (micro F-measure) of 0.4254 MiF using their AttentionXML with multilingual-BERT system, compared to the baseline score of 0.2695. Table 3 shows the results of the runs for this task. As a matter of fact, the five best scores were from Fudan. This team also outperformed all others in the comparable 8a indexing Task in English.

Although MiF represent the official competition metric, other metrics are provided for completeness [14].

[13] http://participants-area.bioasq.org/results/8a/

[14] It is noteworthy that another team (Anuj-ml, from India) that was not among the highest scoring on MiF, nevertheless scored considerably higher than other teams

System	MiF	EBP	EBR	EBF	MaP	MaR	MaF	MiP	MiR	Acc.
Model 4	**0.4254**	0.4382	**0.4343**	**0.4240**	0.3989	**0.3380**	**0.3194**	0.4374	**0.4140**	**0.2786**
Model 3	0.4227	0.4651	0.4146	0.4217	0.4201	0.3251	0.3122	0.4523	0.3966	0.2768
Model 1	0.4167	0.4596	0.4087	0.4160	0.4122	0.3153	0.3024	0.4466	0.3906	0.2715
Model 2	0.4165	0.4296	0.4247	0.4150	0.3918	0.3277	0.3082	0.4286	0.4051	0.2707
Model 5	0.4130	0.4538	0.4061	0.4122	0.4094	0.3162	0.3039	0.4416	0.3879	0.2690
PriberamTEnsemble	0.4093	0.5465	0.3452	0.4031	0.5944	0.2024	0.2115	0.5336	0.3320	0.2642
PriberamSVM	0.3976	0.4451	0.3904	0.3871	0.4602	0.2609	0.2543	0.4183	0.3789	0.2501
iria-mix	0.3892	0.5375	0.3207	0.3906	0.5539	0.2263	0.2318	0.5353	0.3057	0.2530
PriberamBert	0.3740	0.4477	0.3463	0.3678	0.4277	0.2002	0.2009	0.4293	0.3314	0.2361
iria-1	0.3630	0.5055	0.2980	0.3643	0.5257	0.1908	0.1957	0.5024	0.2842	0.2326
iria-3	0.3460	0.5432	0.2674	0.3467	0.5789	0.1617	0.1690	0.5375	0.2551	0.2193
iria-2	0.3423	0.4699	0.2837	0.3408	0.4996	0.1715	0.1719	0.4590	0.2729	0.2145
PriberamSearch	0.3395	0.4582	0.2824	0.3393	0.4971	0.1742	0.1776	0.4571	0.2700	0.2146
iria-4	0.2743	0.3070	0.2635	0.2760	0.2655	0.2925	0.2619	0.3068	0.2481	0.1662
BioASQ_Baseline	0.2695	0.2681	0.3239	0.2754	0.3733	0.3220	0.2816	0.2337	0.3182	0.1659
graph matching	0.2664	0.3708	0.2220	0.2642	0.4261	0.1424	0.1422	0.3501	0.2150	0.1594
exact matching	0.2589	0.2915	0.2395	0.2561	0.4356	0.0627	0.0575	0.2915	0.2328	0.1533
LasigeBioTM TXMC F1	0.2507	0.3641	0.1986	0.2380	0.3646	0.0799	0.0858	0.3559	0.1936	0.1440
Anuj_Ensemble	0.2163	0.2608	0.2082	0.2155	0.3641	0.1997	0.1746	0.2291	0.2049	0.1270
Anuj_NLP	0.2054	0.2499	0.1961	0.2044	0.3632	0.1996	0.1744	0.2196	0.1930	0.1198
NLPUnique	0.2054	0.2499	0.1961	0.2044	0.3632	0.1996	0.1744	0.2196	0.1930	0.1198
X-BERT BioASQ F1	0.1430	0.5057	0.0867	0.1397	0.3095	0.0186	0.0220	0.4577	0.0847	0.0787
LasigeBioTM TXMC P	0.1271	0.6609	0.0716	0.1261	0.6989	0.0081	0.0104	0.6864	0.0701	0.0708
Anuj_ml	0.1149	**0.7547**	0.0636	0.1164	**0.8020**	0.0005	0.0006	**0.7557**	0.0621	0.0636
X-BERT BioASQ	0.0909	0.5415	0.0508	0.0916	0.3422	0.0036	0.0045	0.5449	0.0496	0.0503

Table 3. Final scores for MESINESP task submissions, including the official MiF metric in addition to other useful metrics.

3.2 Dataset releases and creation of a Silver Standard

A *Silver Standard* that contains 5.851.870 entries was created from the submissions, that is automatically generated indexing results by participating teams for a collection of 23.873 documents [15]. Each entry in the MESINESP silver standard corpus contains:

- Submission/Run Name
- The document Id
- Our own MESINESP Id
- The source DB
- A DeCS code
- The Spanish Term or descriptor
- The MiF (Micro-F1) scored by this run

with Precision metrics such as EBP (Example Based Precision), MaP (Macro Precision) and MiP (Micro Precision). Unfortunately, at this time we have not received details on their system implementation.

[15] https://zenodo.org/record/3946558

– The MiR (Micro-Recall) scored by this run
– The MiP (Micro-Precision) scored by this run
– The Accuracy scored by this run
– A consensus across all runs (e.g. how many runs attributed this DeCS to this document)

The last five fields can help asses the reliability of the automatic annotation. Since some of the teams used various non-official sources to train their systems, there were some DeCS codes that were not included in the mapping file distributed/used or in the training dataset, and they were removed from the Silver Standard since no descriptor could be linked to it. 513 DeCS codes were thus removed, some appearing only once, but at least 4 of the appearing hundred of times.

In addition to the automatically annotated Silver Standard, a the full, manually-annotated dataset from the 7 human annotator will be releases, containing 66.271 datapoints with:

– Annotator ID
– DocumentId
– DeCS Code
– Annotation Timestamp
– IF validated or not by another annotator
– Spanish Descriptor
– MESINESP doc ID
– Document Source

We have also generated additional resources of relevance for this task, including a machine translated collection of PubMed abstracts generated [16] using a system adapted for medical text translation English-Spanish [17] that particpated in the medical machine translation track of WMT 2019 [4]. Moreover, participants had access to medical word embeddings[17] for Spanish [18].

4 Discussion and conclusions

This paper provides an overview of the MESINESP Task within the eighth BioASQ challenge (CLEF 2020). The new MESINESP task on semantic indexing of medical content in Spanish ran for the first time and showed strong results across the board and a good international participation. The addition of the new challenging task on medical semantic indexing in Spanish, revealed that in a context beyond the English language, there is even more room for improvement, highlighting the importance of the availability of adequate resources for the development and evaluation of systems to effectively help biomedical experts dealing with non-English resources.

[16] https://zenodo.org/record/3826554
[17] https://zenodo.org/record/3744326

The overall shift of participant systems towards deep neural approaches, already noticed in the previous years, is even more apparent this year. Most of the systems adopted on neural embedding approaches, notably based on BERT and BioBERT models, for all tasks of the challenge.

Overall, as in previous versions of the challenge, the top preforming systems were able to advance over the state of the art, outperforming the strong baselines on the challenging shared tasks offered by the organizers. In addition, a very valuable Silver Standard resource with 5.8 data points will enhance the semantic indexing resources for Spanish. Therefore, we consider that the challenge keeps meeting its goal to push the research frontier in biomedical semantic indexing and question answering. The future plans for the challenge include the extension of the benchmark data though a community-driven acquisition process.

5 Acknowledgments

The MESINESP task is sponsored by the Spanish Plan for advancement of Language Technologies (Plan TL) and the Secretaría de Estado para el Avance Digital (SEAD). BioASQ is also grateful to LILACS, SCIELO and Biblioteca virtual en salud and Instituto de salud Carlos III for providing data for the BioASQ MESINESP task.

References

1. Almagro, M., Unanue, R.M., Fresno, V., Montalvo, S.: Icd-10 coding of spanish electronic discharge summaries: An extreme classification problem. IEEE Access **8**, 100073–100083 (2020)
2. Balikas, G., Partalas, I., Kosmopoulos, A., Petridis, S., Malakasiotis, P., Pavlopoulos, I., Androutsopoulos, I., Baskiotis, N., Gaussier, E., Artieres, T., Gallinari, P.: Evaluation framework specifications. Project deliverable D4.1, UPMC (05/2013 2013)
3. Bansal, M.: Cardiovascular disease and covid-19. Diabetes & Metabolic Syndrome: Clinical Research & Reviews (2020)
4. Bawden, R., Cohen, K.B., Grozea, C., Yepes, A.J., Kittner, M., Krallinger, M., Mah, N., Neveol, A., Neves, M., Soares, F., 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). pp. 29–53 (2019)
5. Chang, W.C., Yu, H.F., Zhong, K., Yang, Y., Dhillon, I.: X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers. arXiv preprint arXiv:1905.02331 (2019)
6. Couto, F.M., Lamurias, A.: MER: a shell script and annotation server for minimal named entity recognition and linking. Journal of Cheminformatics **10**(1), 58 (dec 2018). https://doi.org/10.1186/s13321-018-0312-9
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**(Mlm), 4171–4186 (oct 2018), http://arxiv.org/abs/1810.04805

8. Giustini, D., Boulos, M.N.K.: Google scholar is not enough to be used alone for systematic reviews. Online journal of public health informatics **5**(2), 214 (2013)
9. Jain, H., Prabhu, Y., Varma, M.: Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. pp. 935–944. ACM Press, New York, New York, USA (2016). https://doi.org/10.1145/2939672.2939756
10. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Mining and Knowledge Discovery **29**(3), 820–865 (2015)
11. Krallinger, M., Krithara, A., Nentidis, A., Paliouras, G., Villegas, M.: Bioasq at clef2020: Large-scale biomedical semantic indexing and question answering. In: European Conference on Information Retrieval. pp. 550–556. Springer (2020)
12. Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., Zhu, S.: Deepmesh: deep semantic representation for improving large-scale mesh indexing. Bioinformatics **32**(12), i70–i79 (2016)
13. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. Proceedings of the Conference on Empirical Methods in Natural Language Processing pp. 31–40 (feb 2018), http://arxiv.org/abs/1802.05365
14. Rajkumar, R.P.: Covid-19 and mental health: A review of the existing literature. Asian journal of psychiatry p. 102066 (2020)
15. Ribadas, F.J., De Campos, L.M., Darriba, V.M., Romero, A.E.: CoLe and UTAI at BioASQ 2015: Experiments with similarity based descriptor assignment. CEUR Workshop Proceedings **1391** (2015)
16. Salehi, S., Abedi, A., Balakrishnan, S., Gholamrezanezhad, A.: Coronavirus disease 2019 (covid-19): a systematic review of imaging findings in 919 patients. American Journal of Roentgenology pp. 1–7 (2020)
17. Soares, F., Krallinger, M.: Bsc participation in the wmt translation of biomedical abstracts. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). pp. 175–178 (2019)
18. Soares, F., Villegas, M., Gonzalez-Agirre, A., Krallinger, M., Armengol-Estapé, J.: Medical word embeddings for spanish: Development and evaluation. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 124–133 (2019)
19. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics **16**, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6
20. You, R., Zhang, Z., Wang, Z., Dai, S., Mamitsuka, H., Zhu, S.: Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. arXiv preprint arXiv:1811.01727 (2018)