

PIDNA at BioASQ MESINESP: Hybrid Semantic Indexing for Biomedical Articles in Spanish

Yi Huang¹, Buse Giledereli^{2,3}, Abdullatif Koksall³, Arzucan Ozgur³ and Elif Ozkirimli⁴

¹Data and Analytics Chapter, Roche (China) Holding Ltd., Shanghai, China

²Data and Analytics Chapter, Roche Müstahzarları Sanayi Anonim Şirketi, Turkey

³Computer Engineering Department, Bogazici University, Turkey

⁴Data and Analytics Chapter, F. Hoffmann-La Roche AG, Switzerland

Abstract

Semantic indexing of biomedical articles is difficult due to the extensive use of domain-specific terminology. The task is even more difficult when the corpus is not in English and when there are only a limited number of training data points. In this paper, we describe a hybrid semantic indexing method for biomedical articles in Spanish with the data provided for the MESINESP task (subtrack 1) of the BioASQ challenge 2021. The method integrates transformer-based multi-label text classification and named entity recognition (NER). Our approach has outperformed the baseline methodology by a wide margin in microF1 and has ranked as the second team in the challenge.

Keywords

Biomedical semantic indexing, Text classification, Transformer-based framework, BioASQ challenge

1. Introduction

The growing body of scientific publications makes it very hard to keep track of recent advances. Indexing provides valuable article annotation for information retrieval, but automatic indexing of the articles remains a major bottleneck due to the long-tailed distribution of labels from a large set of controlled vocabulary. Indexing becomes even harder for text in non-English languages with limited training data.

BioASQ is a challenge in large-scale biomedical semantic indexing and question answering [1, 2]. The aim of the Medical Semantic Indexing in Spanish (MESINESP) task [3, 4, 5] is to provide a rich environment for studies in indexing large-scale medical and clinical clauses written in Spanish, which would help to keep track of different aspects of the literature. Many stakeholders such as pharmaceutical companies and researchers in clinical medicine would benefit from systematic labeling of this emerging number of biomedical articles. In MESINESP task, training and evaluation data are proposed for medical semantic indexing task in Spanish (detailed statistics in Table 1). The training data of biomedical articles in Spanish are labeled with DeCS (Health Sciences Descriptors). DeCS is a structured vocabulary developed based on

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

✉ yi.huang.yh4@roche.com (Y. Huang); buse.giledereli@roche.com (B. Giledereli); abdullatif.koksall@boun.edu.tr (A. Koksall); arzucan.ozgur@boun.edu.tr (A. Ozgur); elif.ozkirimli@roche.com (E. Ozkirimli)



© 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)

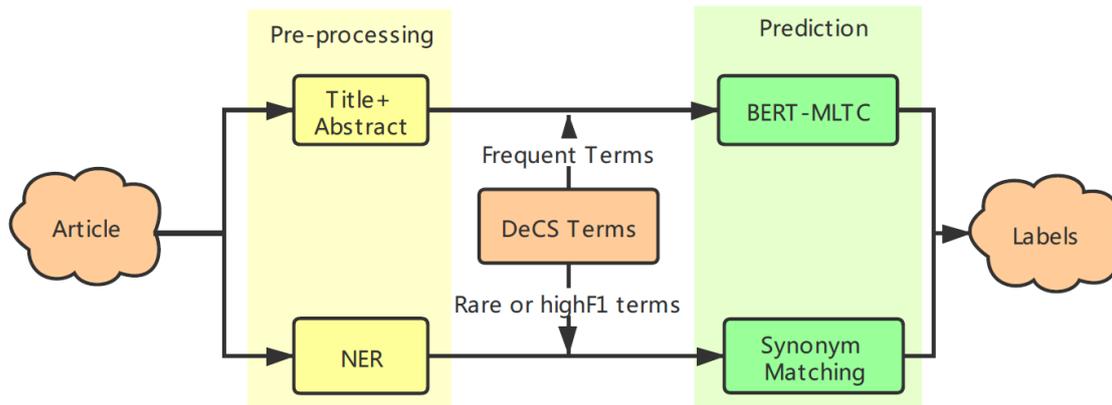


Figure 1: The procedure of hybrid semantic indexing. For frequent terms, transformer-based MLTC is applied. For rare terms or terms with high F1 in the validation step, a term is labeled if there is an entity identified with NER matching its synonyms. The results of BERT-MLTC and Synonym matching are pooled together as the final labels.

Medical Subject Headings (MeSH) terms and serves as a common terminology for consistent search in Spanish, Portuguese and English. For each DeCS heading, there is a list of descriptors and synonyms from both European and Latin Spanish DeCS 2019 data sets.

In MESINESP 8 - 2020 [3], best performing teams made effective use of pretrained transformer models such as ELMO [6], multilingual BERT [7], and X-BERT [8] for text classification. There were also hybrid indexing approaches on top of pretrained transformer models. One of them from Priberam Labs used an ensemble of a Support Vector Machine model, a search engine and a BERT-based classifier [9] and proved the performance of multiple binary classifiers. Another team from the University of Lisboa leveraged an additional NER model to recognize MeSH terms in the abstracts [10, 11]. Following the same line of thought, in this work, we propose a hybrid approach where the BERT model trained on only Spanish and based on multiple binary classifiers is used with additional entity synonym matching in the articles to DeCS terms. Our results show that this hybrid method outperforms a solely transformers-based model in rare classes and prove that NER integration is helpful for the long-tailed distribution problem.

The paper is organized as follows. Section 2 describes the hybrid design of the system. Section 3 provides experimental details for our proposed approach. Section 4 discusses the results and the evaluation for the MESINESP dataset. Finally, Section 5 summarizes our conclusions and perspectives for future work.

2. Hybrid Semantic Indexing

Pretrained large language models such as BERT provide a powerful and high performance framework in many NLP tasks [7] for various languages, including Spanish [12]. They have been applied to the multi-label text classification (MLTC) task [13] to find the corresponding labels within a label set for semantic indexing. When the label set is large, the MLTC task

Table 1
Dataset Statistics of MESINESP subtrack 1

Dataset	Statistic
Number of articles (training)	237574
Number of articles (development)	1065
Number of articles (test)	10179 (500 with expert annotation)
Average number of labels per article (training)	8.37
Average number of articles per label (training)	88.64

often uses auxiliary multiple binary classifiers. However, for some of the labels, their binary classifiers are trained with few data points and the final multi-label model is often biased to high-frequency labels.

Besides MLTC, NER is another approach to extract key information from the text. Using NER, rare or unseen entities can be effectively recognized as long as there are enough training data of the same entity category. Therefore, in this work, we explore the possibility to integrate NER into the prediction pipeline as a pre-processing step, especially for those rare labels, to complement MLTC.

For this BioASQ challenge, the Text Mining Unit of the Barcelona Supercomputing Center has extracted entities related to medications, diseases, symptoms, and medical procedures for training, development and test sets. With the entities identified, there can be various ways to infer the labels. In this work, we evaluate a straight-forward approach by simply matching the entities to synonyms (provided by DeCS).

Figure 1 illustrates the procedure of the whole hybrid method. We take advantage of the transformer-based MLTC (BERT-MLTC) for frequent terms, and leverage the entities provided by the Barcelona Supercomputing Center for rare terms which occur fewer than 3 times in the training set, as well as the terms with high term-wise F1 score in the validation dataset. The semantic indexing result is the union of outputs of these two approaches.

3. Experiments

There are 237574 articles in the training set, 1065 in the development or validation set, and 10179 in the test set for the subtrack 1 of MESINESP (Table 1). 500 articles from the test data set are expert-annotated from LILACS and IBECs and used in the official evaluation. There are 34040 DeCS terms with synonyms, 22434 of which occur in the training set, and 17006 occur at least 3 times.

For BERT-MLTC, we use the *BertForSequenceClassification* backbone in the *transformers* library [14] with the BETO (bert-base-spanish-wwm-cased) pretrained model [12]. The BETO model has 110 million parameters. The training data are truncated with a maximal length of 512 and grouped with a batch size of 32. Only terms that occur at least 3 times in the training set (frequent terms) are included. We use AdamW with a weight decay of 0.01 as the optimizer, and determine the learning rate by hyperparameter search. The transformer-based framework is implemented in PyTorch.

Table 2

Evaluation results of hybrid semantic indexing, compared with BERT-MLTC only system as well as BERTDeCS version 4 (the best system) and MESINESP_baseline_t1 (the baseline system). *pi_dna* used hybrid semantic indexing, while *pi_dna_2* used BERT-MLTC

System	MiF	EBP	EBR	EBF	MaP	MaR	MaF	MiP	MiR	Acc.
BERTDeCS	0.4837	0.5077	0.4736	0.4763	0.5237	0.3990	0.3926	0.5077	0.4618	0.3261
<i>pi_dna</i>	0.4225	0.4919	0.3876	0.4120	0.4463	0.3149	0.3082	0.4667	0.3859	0.2722
<i>pi_dna_2</i>	0.3978	0.4836	0.3630	0.3915	0.4062	0.2717	0.2700	0.4520	0.3551	0.2546
baseline	0.2876	0.2449	0.3839	0.2841	0.3720	0.3787	0.3438	0.2335	0.3746	0.1710

For synonym matching, we perform an exact match of each full entity (recognized and provided by the Text Mining Unit of the Barcelona Supercomputing Center) to all DeCS synonyms in a case-insensitive manner. If the corresponding DeCS term occurs fewer than 3 times in the training set (rare term), or its term-wise F1 is larger than a threshold (0.01 in this work by hyperparameter search), the term will be set positive in the hybrid result.

The result of hybrid semantic indexing is named *pi_dna* in the official evaluation. We also keep the original result of BERT-MLTC, named *pi_dna_2*, to assess the improvement by integrating NER and synonym information.

4. Results

For each system, the MESINESP subtrack 1 evaluate performance with Micro F-Measure (MiF), Example Based Precision (EBP), Example Based Recall (EBR), Example Based F-Measure (EBF), Macro Precision (MaP), Macro Recall (MaR), Macro F-Measure (MaF), Micro Precision (MiP), Micro Recall (MiR) and Accuracy (Acc.). MiF is the official evaluation metric for this task. A summary of the results for our approaches, as well as the best and the baseline systems, are listed in Table 2.

Among the 26 systems of subtrack 1, our system of hybrid semantic indexing has ranked as the 6th (second as a team) in the challenge with a miF of 0.4225, whereas the plain BERT-MLTC achieved a miF of 0.3978. Compared to the baseline model MESINESP_baseline_t1, both models provide a significant improvement in MiF and MiP scores. This shows how strong the BERT-MLTC model is in frequent labels. The main improvement over the baseline is for the precision scores.

For macro scores, the baseline model already has high MaR (0.3787) and the performance in this metric improves only slightly for the highest scoring system (0.399). The hybrid model (*pi_dna*) provides an advantage in rare classes. It outperforms the BERT-MLTC based model (*pi_dna_2*) in every metric, with the biggest improvements being in MaP (+0.0401) and MaR (+0.0432) metrics. This improvement in macro scores shows the advantage of the NER based synonym matching in the rare classes, which is important in datasets with long-tailed distribution.

Noteworthy, our systems are efficient in both training and inference steps. With one GPU (A100), the training step takes less than 1 hour for one epoch, and the inference step takes 5 minutes for the whole test set (10179 articles).

This is our first time participating MESINESP, therefore, we also compare our results with previous systems with hybrid indexing approaches in past editions of MESINESP, with the baseline model system as a benchmark. Last year, the best system (on miF) from Priberam Labs, PriberamTEnsemble, achieved a miF of 0.4093 (+0.1398) and a maF of 0.2115 (-0.0701); the best system (on miF) from the University of Lisboa, LasigeBioTM TXMC F1, achieved a miF of 0.2507 (-0.0188) and a maF of 0.0858 (-0.1958). This year, our system (pi_dna) achieves a miF of 0.4225 (+0.1349) and a maF of 0.3082 (-0.0356). Our system works at a similar level as PriberamTEnsemble (with fewer base models) and better than LasigeBioTM TXMC F1.

5. Conclusions

In this study, we introduce a hybrid semantic indexing method for Spanish biomedical articles, and show its effectiveness and efficiency in the MESINESP task subtrack 1 of the BioASQ challenge 2021. We propose the integration of MLTC, NER and terminology as a promising approach for non-English biomedical text mining. The proposed hybrid approach can also be used to index document types in other domains with domain-specific language.

As future work, we are going to explore more options in the base models of the hybrid approaches. For example, synonym matching can be improved by taking DeCS synonyms into the NER step [11]. In addition, hyperparameters such as the rare term cutoff 3 can be further fine-tuned for higher performance in MaR.

References

- [1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, *BMC Bioinformatics* 16 (2015) 138. URL: <http://www.biomedcentral.com/content/pdf/s12859-015-0564-6.pdf>. doi:10.1186/s12859-015-0564-6.
- [2] A. Nentidis, G. Katsimpras, E. Vandorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. (2021).
- [3] C. Rodriguez-Penagos, A. Nentidis, A. Gonzalez-Agirre, A. Asensio, J. Armengol-Estapé, A. Krithara, M. Villegas, G. Paliouras, M. Krallinger, Overview of mesinesp8, a spanish medical semantic indexing task within bioasq 2020 (2020).
- [4] L. Gasco, A. Nentidis, A. Krithara, D. Estrada-Zavala, R.-T. Murasaki, E. Primo-Peña, C. Bojo-Canales, G. Paliouras, M. Krallinger, Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. (2021).
- [5] L. Gasco, M. Antonio, M. Krallinger, Mesinesp2 corpora: Annotated data for medical semantic indexing in spanish, 2021. doi:10.5281/zenodo.4707104, funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

- [6] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2227–2237.
- [7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [8] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, I. Dhillon, Taming pretrained transformers for extreme multi-label text classification, arXiv preprint arXiv:1905.02331 (2019).
- [9] R. Cardoso, Z. Marinho, A. Mendes, S. Miranda, Priberam at mesinesp multi-label classification of medical texts task, 2021. arXiv:2105.05614.
- [10] F. M. Couto, A. Lamurias, Mer: a shell script and annotation server for minimal named entity recognition and linking, *Journal of Cheminformatics* 10 (2018) 58. URL: <https://doi.org/10.1186/s13321-018-0312-9>. doi:10.1186/s13321-018-0312-9.
- [11] A. Neves, A. Lamurias, F. M. Couto, Extreme multi-label classification applied to the biomedical and multilingual panorama, 2020.
- [12] 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.
- [13] R. You, Y. Liu, H. Mamitsuka, S. Zhu, BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text, *Bioinformatics* 37 (2020) 684–692. URL: <https://doi.org/10.1093/bioinformatics/btaa837>. doi:10.1093/bioinformatics/btaa837.
- [14] 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>.