Biomedical Question Answering using Extreme Multi-Label Classification and Ontologies in the Multilingual Panorama[*]

Andre Neves, Andre Lamurias, Francisco M. Couto

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. Deep learning models achieve state-of-the-art results in Natural Language Processing (NLP) tasks, such as Question Answering (QA), across different domains, mostly thanks to pre-trained language models such as BERT [1]. However, there is a lack of models designed for NLP tasks in the multilingual panorama, especially in specific domains such as the biomedical sciences, mostly due to the lack of datasets available in non-English languages.

In this short paper, we propose the development of a QA system using state-of-the-art deep learning models and combining it with a deep learning Extreme Multi-Label Classification (XMLC) solution along with ontologies, in order to improve the results achieved by the model. The proposed model shall be able to answer biomedical questions in English, Spanish and Portuguese.

Keywords: Deep Learning, Question Answering, Multilingual, Extreme Multi-Label Classification, Ontologies.

1 Introduction

QA aims at the automatic retrieval of the most relevant and informative answers to questions made by the user. For these answers to be precise, the system requires the capability to understand and process natural language, which can be achieved through deep learning techniques that automatically discover patterns in the text by giving raw text data as input [2].

The choice of datasets and corpora to use is essential when training the model. However, most deep learning models and corpora are designed exclusively for the English language, thus being difficult to apply them to QA and other NLP tasks in other languages. This can be even more significative when dealing with specific domains, like biomedical sciences, where structured domain specific corpora may not exist in non-English languages.

Although there are multilingual deep learning models such as BERT [1], and biomedical models such as BioBERT [3] and SciBERT [4], multilingual biomedical QA

[*] This work was supported by FCT through project DeST, ref. PTDC/CCIBIO/28685/2017, and the LASIGE Research Unit, ref. UIDB/00408/2020

still remains a challenge. However, there might be a possibility to increase the quality of the results in multilingual question answering by combining it with deep learning XMLC. This technique consists in assigning to the text multiple labels from several candidate labels from a dataset that can achieve thousands or even millions [5]. The choice of labels to use can be from any language and from any domain. Thus, with the right choice of labels, it can be applied to multilingual biomedical QA.

2 Objectives

We propose to develop a deep learning QA model designed for the biomedical and multilingual panorama, namely for the Spanish, Portuguese and English languages.

Since XMLC algorithms can classify documents with labels that are related with the contents of the document, and usually the first step of a QA system is to retrieve documents relevant to the question, the hypothesis is, that by using these labels, the QA system will be able to give more accurate answers, since it had a series of multilingual labels to identify the most related documents to the question. This label classification can be adapted to the multilingual biomedical panorama thanks the biomedicals terms chosen as labels, which can be controlled vocabularies that exist in multiple languages such as DeCS (Descriptores en Ciencias de la Salud) or ICD (International Classification of Diseases) terms, in addition to the dataset and pre-trained language model used to train the algorithm.

An additional objective consists in incorporating biomedical ontologies with the deep learning model, which is expected to improve the understanding of the context of the lexical terms in the text, since ontologies provide a structured representation of the knowledge about a given domain.

3 Methods

To achieve these objectives, an XMLC algorithm named X-BERT, which consists in a deep learning approach to scale the BERT model to XMLC [6], will be adapted to the biomedical multilingual panorama. Our approach will use the BERT multilingual pre-trained model in combination with three datasets of scientific articles, one in each language (English, Spanish and Portuguese), retrieved from PubMed and the Virtual Health Library databases.

Then, since the algorithm uses labels to classify the data, one can use as labels the DeCS terms, which consists in a hierarchy of terms that were developed from MeSH (Medical Subject Headings), to label biomedical articles in Portuguese, Spanish and English. Since almost every DeCS term has a corresponding MeSH term, it is possible to make a conversion between MeSH and DeCS terms and thus label multilingual data. For example, the MeSH term D006801, which corresponds to the label “Humans”, corresponds to the DeCS term 21034, which has the same label and a corresponding description in English, Portuguese and Spanish. Thanks to this conversion between MeSH and DeCS terms, it is also possible to access the relations between the terms and their ancestors, and even incorporate ontologies in this multilingual solution, as long as they have MeSH or DeCS terms associated.

Finally, the QA part will use a deep learning model trained with pre-trained language models for scientific text, such as BioBERT [3] and SciBERT [4], along with a QA dataset that gives scientific documents as answers to the questions. The adapted X-BERT model will also be used to previously label the answers of the dataset. This way, it is expected that the QA model retrieves more accurate results.

4 Discussion

Deep learning QA models achieved state-of-the-art results across different domains. However, there is a lack of models designed for the multilingual panorama, especially in specific domains such as the biomedical sciences. With this proposal, it is expected that a reliable QA model can be developed for the biomedical and multilingual panorama, by combining question answering and XMLC.

We have previously explored two types of English biomedical QA models. The first one consisted in classifying question pairs and question-answer pairs according to their similarity [7], while the other consisted in retrieving relevant documents to each question based on posts from a Q&A website [8]. We intend to adapt these two approaches to the multilingual panorama and, along with the XMLC approach herein proposed, consolidate them as a complete multilingual biomedical QA system.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved from http://arxiv.org/abs/1810.04805
2. Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature. https://doi.org/10.1038/nature14539
3. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz682
4. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Retrieved from http://arxiv.org/abs/1903.10676
5. Bhatia, K., Jain, H., Kar, P., Varma, M., & Jain, P. (2015). Sparse local embeddings for extreme multi-label classification. Advances in Neural Information Processing Systems.
6. Chang, W. C., Yu, H. F., Zhong, K., Yang, Y., & Dhillon, I. (2019). X-BERT: eXtreme Multi-label Text Classification with BERT. Retrieved from http://arxiv.org/abs/1905.02331
7. Lamurias, A., & Couto, F. M. (2019). LasigeBioTM at MEDIQA 2019: Biomedical Question Answering using Bidirectional Transformers and Named Entity Recognition. https://doi.org/10.18653/v1/w19-5057
8. Lamurias, A., Sousa, D., & Couto, F. M. (2020). Generating Scientific Question Answering Corpora from Q&A forums. Retrieved from https://arxiv.org/abs/2002.02375