Team LIA/LS2N at BioCreative VII LitCovid Track: Multi-label Document Classification for COVID-19 Literature using Keyword Based Enhancement and Few-Shot Learning

Yanis Labrak [1], Richard Dufour [2]

[1] LIA, Avignon University, Avignon, France

[2] LS2N, Nantes University, Nantes, France

Abstract — **Multi-label text classification consists in attributing, for each textual document, one or more labels. Due to its nature, the task is often considered to be more challenging than other types of classification problems since the number of labels to assign is unknown. In text documents, this difficulty is generally the result of a blurry border between lexical fields of the labels or an underrepresentation of some of them. In this paper, we seek to automatically associate categories to scientific articles related to the COVID-19. We propose to address this multi-label classification problem by integrating an original keyword enhancement method to the TARS transformer-based approach designed to perform few-shot learning. Experiments conducted during the BioCreative challenge on the multi-label classification task show that our approach outperforms the baseline (ML-Net), no matter the metric considered.**

Keywords* — *Multi-label Classification; Transformers; Biomedical; COVID-19; PubMed; BioCreative VII; LitCovid; BERT; TARS; Flair

I. Introduction

Since the beginning of the global pandemic, the scientific community has pooled its effort to fight or even understand how SARS-COV-2 works. This has resulted in the formation of a large amount of scientific articles related to the 2019 novel Coronavirus (COVID 19), with an estimate of 200,000 articles published for the year 2020 alone (Else, H. 2020). Information access is then a very important concern due to the rapid and continuous arrival of new scientific publications. As a result, information retrieval has become more important than ever and the ability to quickly retrieve relevant scientific literature to support researchers in the healthcare industry appears essential.

In this context, the National Institutes of Health (NIH) has implemented the LitCovid platform (Chen, Q. et al., 2020), making it possible to aggregate the scientific literature linked to the COVID 19. This hub is updated daily with articles from PubMed, which hosts millions of scientific articles related to the fields of biology and medicine.

To make the document's retrieval easier to scientists, policy makers, healthcare professionals or even the general public, LitCovid organizes articles into categories describing their research topics (*General*, *Mechanism*, *Transmission*, *Diagnosis*…). Note that the same document may be here assigned with several topics.

Such a process could not be done manually on this fast-growing large volume of textual documents. This led the computational biology branch of the National Center for Biotechnology Information (NCBI) and National Library of Medicine (NLM) to develop a attention based multi-label classifier, called ML-Net (Du, J. et al., 2019), to automatically annotate articles with their related topics.

This task then consists of a multi-label classification problem applied to text documents. Main difficulty with this kind of task is to predict an unknown number of labels for a document where the borders between topics are sometimes very blurry due to terms in common.

In this work, we introduce an original approach to automatically associate topics to COVID 19 related scientific articles based on the TARS (Task Aware Representation of Sentences) architecture (Akbik et al. 2020), enhanced with keyword tags extracted with a TF-IDF approach. This multi-label document classification problem takes place in the context of the track 5 of the BioCreative 7 (BC7) evaluation campaign *LitCovid track Multi-label topic classification for COVID-19 literature annotation*[1].

The paper is organized as follows. Section II briefly summarizes some of the current main approaches related to text document representation. Then, Section III presents the proposed approaches while Section IV describes the experimental protocol. Experiments are presented in Section V, before concluding and giving some perspectives in Section VI.

II. Related Work

Word and document representations are a historical problem in automatic language processing. It is notably the entry point of downstream tasks, and the quality of these

[1] *https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/*

representations generally depends on their performance. From more classical approaches allowing to obtain a fixed vector representation for each word, we have now arrived at representations of words that vary according to their surrounding context thanks to the use of complex architectures relying on massive quantities of textual data.

Non-contextual word representations. Word representations can be obtained from various ways. Historically, bag-of-words (BOW) approaches have been proposed, such as Term Frequency - Inverse Document Frequency (TF-IDF), to obtain a vector representation of the words (and by extension, the documents). The main drawback of these approaches relies on the low (or inexistent) consideration of the word context. This is particularly damaging for most natural language processing (NLP) tasks that generally require taking into account the context to disambiguate the words, e.g.:

I broke the pot.

I ***don't*** *break the pot.*

I ***don't*** *know why he broke the pot.*

Such representations also fail to capture polysemy. The representation of a specific token would always be the same regardless if it appears in very different contexts, e.g.:.

A world ***record****. ≠ A* ***record*** *of the conversation.*

Efficient Estimation of Word Representations in Vector Space. In recent years, one of the most impactful word representations that use contextual elements was introduced with Word2Vec (Mikolov, T. et al. 2013). Unlike previous methods, Word2Vec uses a large amount of unlabeled data to compute a static word vector based on the surrounding context. This vector is called a word embedding and provides a pretty generic representation of the word context observed during the training phase. It allowed the community to do a huge leap forward to solve the word representation issues but suffers from two main concerns. The first one is the lack of consideration of word order for the surrounding context. The second issue is the inability to handle polysemous words and the change of context due to having a fixed embedding obtained during the training phase. Finally, although they have allowed major advances in terms of performance and quality of word representation, such approaches still suffer, to a lesser extent, from the problems we have observed with BOW approaches.

Bidirectional Encoder Representations from Transformers or BERT. Nowadays, word representations tend to use more and more surrounding context of the words to provide a dynamic representation and capture their different semantic meanings to address the issue of polysemy and the context-dependent nature of words. BERT is based on a deep neural network architecture (Jacob Devlin et al. 2019), called *transformers*, which follows that trend and uses a bi-directional approach coupled with a self-attention mechanism (Vaswani, A. et al., 2017) to learn the context of words in a more robust way. Rather than only trying to represent words based on their context, BERT masks 15% of the words in a sentence and forces the model to learn how to use information from the entire sentence to deduce which words are missing. Compared to other similar architectures, BERT has already been trained on a very large corpora of unlabelled text extracted from Wikipedia and books, which consolidate its vocabulary and word embeddings. However, this method is still being constrained by the size of the vocabulary, which can be an issue in some specific domains such as the healthcare industry, where the vast majority of the documents can then be composed of out-of-vocabulary tokens (*i.e.* absent from the BERT vocabulary).

Task Aware Representation of Sentences for Generic Text Classification. Regarding other approaches, such as transfer learning, TARS (Halder, K et al., 2020) has two main advantages. Firstly, the model is not constrained by the class number that changes from one targeted classification task to another, by training the system to detect the presence, or not, of each feeded class (binary classification). Secondly, the approach tries to integrate the semantic information of the targeted class name in the training process by linking it to the data, while most of the systems only focus on the training examples. As claimed by the authors, it allows us to reach higher performance when few data are available since it preserves information contained in the already trained linear layer and does not train it from scratch.

III. Methods

In this section, we describe the methods that we propose for our multi-label topic classification problem.

Baseline approach: SVM classifier with a TF-IDF 1-2-3 gram. For our baseline, we decided to use a Support Vector Machines (SVM) in conjunction with a TF-IDF for document representation.

During the computation of the TF-IDF, we represent documents at the token, bigram (two-word sequence) and trigram (three-word sequence) level to reduce ambiguities.

SVM classifier, TF-IDF vector representation and n-grams were obtained from the scikit-learn library (Pedregosa F. et al., 2011).

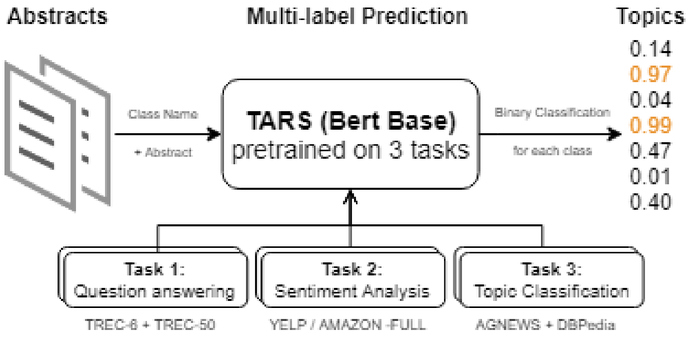

Figure 1. TARS Pipeline

Proposed approach: Task Aware Representation of Sentences for Generic Text Classification and keyword-based enhancement. For our best performing model, we are using a few-shot and zero-shot adaptation of the base version of BERT, called TARS (see Figure 1), which is optimized for text classification on a very small training corpus. The implementation of the model used during our experiments is part of the FlairNLP (Akbik, A. et al., 2019) library. For this first model, we simply trained the TARS model during 50 epochs to distinguish article topics based on their abstracts.

For the second submission, we used the same TARS model but fed with more relevant data such as abstracts, titles and keywords. Then, we applied what we called a keyword-based enhancement to the data. This enhancement consists of applying a first TF-IDF pass on the data to extract the specific terms of each topic with a score greater than 0.65. These terms are then framed by tags (Caubriere A. et al., 2020), the idea being to explicitly give more importance to these terms during their modeling by the TARS model. This original proposition is schematized in Figure 2. The model has been trained during 10 epochs.

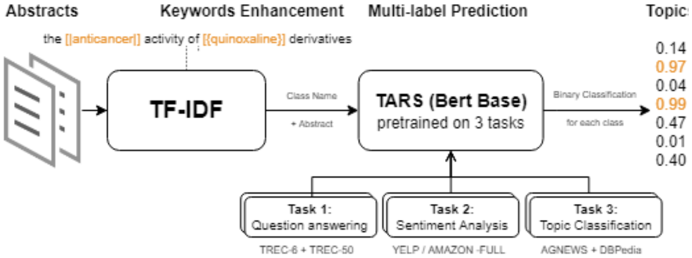

Figure 2. TARS + Keywords Enhancement Pipeline

IV. Experimental Protocol

A. Dataset & Labels

The LitCovid platform lists more than 170,000 articles published on PubMed since the beginning of the COVID-19 global pandemic, while approximately 10,000 new articles are added every month (Chen, Q. et al., 2021).

For the purpose of the Biocreative 7 (BC7) challenge, organizers then selected a subset from LitCovid of 33.7k articles, including their associated metadata (title, abstract, keywords...). Table I presents the statistics of the BC7 LitCovid dataset.

TABLE I. Statistics on BC7 LitCovid dataset.

	Corpora			
	Train	***Dev***	***Test***	***Total***
#docs	24,960	6,239	2,500	**33,699**
#tokens titles	314.8K	79.0K	32.2K	**426.0K**
#tokens abstracts	611.7K	151.1K	64.6K	**827.4K**
Total #tokens	**926.5K**	**230.1K**	**96.8K**	**1.25 M**

Articles of the dataset have been manually reviewed and annotated by the National Library of Medicine (NLM). Each article has been assigned with up to eight possible topics: *Treatment, Mechanism, Prevention, Case Report, Diagnosis, Transmission* and *Epidemic Forecasting*. As previously said, the same article may be associated with multiple topics. On average, we have 1.37 topics per article and at most 5 topics for one article.

Concerning *Treatment* and *Diagnosis*, their over representation in the corpora had forced the organizers to bottleneck them in the LitCovid curation pipeline. However, we can observe in Table II that the corpora is still significantly imbalanced between topics.

TABLE II. Statistics on BC7 LitCovid labels.

Label	***Train+Dev***	***Proportion***
Prevention	13,852	32.40%
Treatment	10,924	25.55%
Diagnosis	7,739	18.10%
Mechanism	5,511	12.89%
Case Report	2,545	5.95%
Transmission	1,344	3.14%
Epidemic Forecasting	837	1.95%
Total	**42752**	**100%**

B. Evaluation Metrics

In order to evaluate the results obtained by the different methods for the multi-label classification task, BC7 organizers have chosen to use the precision, recall and F1-score metrics.

Recall. The recall is the number of relevant topics found compared to the total number of topics proposed for a given document.

$$Recall \ = \ \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Precision. The precision is defined by the number of relevant topics found compared to the number of relevant topics that the corpora has.

$$Precision \ = \ \frac{True\ Positives}{True\ Positives + False\ Positives}$$

F1-Score. A measurement that combines precision, recall and their harmonic mean.

$$F1 - Score \ = \ \frac{2 * precision * recall}{precision + recall}$$

V. Results

For reproducibility reasons, both data and source code used during the challenge are available on our Github repository[2].

In addition to our approaches, we also include, for sake of comparison, a simple baseline provided by the organizers based on the ML-Net (Du, J. et al. 2019) architecture. Table III presents the results obtained with the various methods on the LitCovid multi-label topic classification. In addition to the baseline and the proposed approaches, we also include the *teams mean* and the *teams quartile 3* performance, which correspond to the mean performance obtained by considering all participant submissions in the task and to the upper quartile (Q3) of all submissions respectively.

Globally, *TARS + Keywords Enhancement* is our best performing system and outperforms the ML-Net baseline. Compared to the other participants, we are above the mean for all the F1-scores.

TABLE III. Statistics on BC7 LitCovid Results.

	Micro			Macro			Instance Based		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Baseline ML-Net	87.56	81.42	84.37	83.64	73.09	76.55	88.49	85.14	86.78
SVM 3g TF-IDF	89.51	82.80	86.02	88.14	77.23	81.74	87.87	86.10	86.98
TARS 50 epochs	87.60	86.59	87.09	84.98	81.38	82.31	89.81	89.42	89.61
TARS + Keywords 10 epochs	86.99	**89.66**	**88.30**	82.98	85.70	83.66	89.93	**91.98**	90.94
Teams Mean	89.67	86.24	87.78	86.70	80.12	81.91	89.85	88.87	89.31
Teams Quartile 3	**90.79**	85.55	86.70	**92.51**	**89.64**	**90.83**	**93.53**	91.92	**92.54**

VI. Conclusions and Perspectives

In this manuscript, we described our submission in the BioCreative 7 *LitCovid Multi-label topic classification for COVID-19 literature annotation* track. The results demonstrate that the proposed TARS transformer-based system coupled with our keyword based enhancement method can be more performant than a classic TF-IDF SVM approach as well as the organizer ML-Net approach.

Due to the lack of time, our best performing model was trained on only 10 epochs and does not have the time to converge. More time could improve the overall performance of the system and should allow us to be in the third quartile.

Recent works also demonstrate the importance of using domain-specific word embeddings to improve various NLP tasks, including multi-label classification. We could use biomedical-specific embeddings based on the BERT architecture such as PubMedBERT (Gu, Y. et al. 2020), BioBERT (Lee et al., 2019) or BlueBERT (Peng, Y. et al. 2019) to improve overall performances.

Switching to character-level word embedding such as Contextual String Embeddings for Sequence Labelling (Akbik, R. et al. 2018) could also be a significant improvement, even more so if they are trained on domain-specific data. Such modification should allow a better handle of out-of-vocabulary (oov) words, and especially rare and misspelled words, by capturing subword structures such as prefixes and endings. This kind of behavior is very interesting in domains such as healthcare where the vocabulary is too vast to be entirely defined and could improve overall performances.

Acknowledgment

This work was financially supported by Zenidoc.

References

1. Akbik, R. (2018). Contextual String Embeddings for Sequence Labeling. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1638–1649). Association for Computational Linguistics.
2. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In NAACL 2019 (pp. 54–59).
3. Caubriere, A., Rosset, S., Estève, Y., Laurent, A., & Morin, E. (2020, May). Where are we in Named Entity Recognition from Speech?. In Proceedings LREC (pp. 4514-4520).
4. Chen, Q., Allot, A. and Lu, Z., 2021. LitCovid: an open database of COVID-19 literature. Nucleic acids research, 49(D1), pp.D1534-D1540.
5. Chen, Q., Allot, A., & Lu, Z. (2020). Keep up with the latest coronavirus research. Nature, 579(7798), 193-194.
6. Chen Q., Allot A., Leaman R., Islamaj Doğan R., and Lu Z.. Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation. Proceedings of the seventh BioCreative challenge evaluation workshop. 2021.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
8. Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C. and Lu, Z., 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association, 26(11), pp.1279-1285.
9. Else, H. (2020). How a torrent of COVID science changed research publishing-in seven charts. Nature, 553-553.
10. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv preprint arXiv:2007.15779.

[2] https://github.com/ganastek/BioCreative-VII-Track-5

11. Halder, K., Akbik, A., Krapac, J., & Vollgraf, R. (2020). Task Aware Representation of Sentences for Generic Text Classification. In COLING 2020.

12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

13. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

15. Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In BioNLP 2019 (pp. 58–65).

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).