Re-ranking Biomedical Literature for Precision Medicine with Pre-trained Neural Models

Jiazhao Li*†, Adharsh Murali†, Qiaozhu Mei*, and V.G.Vinod Vydiswaran†*
*School of Information, †Department of Learning Health Sciences, Medical School
University of Michigan, Ann Arbor, Michigan, USA
Email: {jiazhaol, adharsh, qmei, vgvinodv}@umich.edu

2020 IEEE International Conference on Healthcare Informatics (ICHI) | 978-1-7281-5382-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICHI48887.2020.9374401

***Abstract*—We propose a biomedical literature retrieval approach that incorporates a domain-specific BERT model as an auxiliary re-ranker. Experiments on TREC Precision Medicine dataset show its effectiveness in improving retrieval performance by 6.2% in inferred NDCG and 6.8% in R-precision over the best-published results. The contribution of this study is to provide evidence of incorporating BERT in a biomedical literature retrieval system, which serves the overall goal to improve the information retrieval for precision medicine.**

***Index Terms*—biomedical retrieval, pretrained models, BERT**

I. Introduction

Clinicians are expected to follow up-to-date guidelines to treat patients through evidence-based interventions. However, many diseases are complex, and a more precise assessment and diagnosis for a patient's condition and an individualized plan may be needed. Such considerations give rise to the paradigm of precision medicine, where variability in genetic, environmental, and lifestyle-based factors could be considered while personalizing treatment and prevention strategies. Improved biomedical literature retrieval approaches could assist clinicians in seeking evidence from peer-reviewed articles linking genetic factors to specific diseases. We present an approach to improving state-of-the-art ad hoc retrieval performance on scientific abstracts by incorporating a neural language model to re-rank biomedical articles.

Previous approaches on this task implemented query expansion using external knowledge sources, such as building knowledge graph [1], mapping MeSH terms [2], or using iterative query construction [3]. Other researchers used annotated data to train topic-specific classifiers [4], [5] or neural network models [6] as re-rankers to boost relevant documents.

Pre-trained neural language models such as BERT [7] have been adapted to re-rank passages in ad hoc retrieval tasks [8], [9]. However, it is not known if BERT could be as effective for retrieval tasks with health-specific focus, such as in precision medicine. We propose an approach to retrieve scientific abstracts by combining BERT's output with an iterative retrieval-based pre-ranker. We examine the efficacy of both the original BERT model trained on Wikipedia and BookCorpus [7], and BioBERT [10] also trained on the biomedical text. The contribution of this study is to provide evidence of incorporating BERT to improve the performance of biomedical passage retrieval for precision medicine.

II. Re-ranking using neural models

The proposed architecture follows a two-step ranking approach with an initial ranker and a BERT re-ranker.

1) Initial Retrieval: We adopt an iterative approach [3] to construct queries that start with the strict matching of the input terms, such as diseases or genes, and progressively lead to the more lenient matching of terms. Each query is built as a boolean combination of whether a term should match a given query field. In addition to the original keywords in the structured input, common words such as 'patient' and 'treatment' are added as optional terms to give higher scores to clinical documents. With BM25 or query likelihood as the base models, different sets of documents are retrieved until the count of non-duplicate documents exceeds 500.

2) BERT as Re-ranker: Proposed by Delvin et al. [7], BERT has been shown to be effective in many NLP tasks with state-of-the-art performance. It is based on the bidirectional encoder layer from the Transformer model, where the self-attention mechanism is applied to learn contextual information from sequence data. Two available versions (base and large) differ in their number of encoder layer and overall parameters. BERT adapts WordPiece embeddings with a 30,000-token vocabulary learned from pre-trained corpus to maximize the probability of generating the whole text [11]. With WordPiece tokenization, unseen words are broken into subwords to alleviate the out-of-vocabulary problem.

When implementing BERT in information retrieval to re-rank passages, we followed the approach described by Nogueira and Cho [8] to frame this task as binary classification on whether the $\langle$query, passage$\rangle$ pair belong together, i.e. the passage is relevant to the query, or not. Query-passage pairs are then used as training data to fine-tune the pre-trained BERT with cross-entropy loss. The classifier's result is fed into a softmax function to produce the probability of a passage being relevant to the query. The score assigned to each passage is then used to create the ranked list.

Passages retrieved by the initial ranker for training queries are labeled using binary relevance judgments. These labeled query-passage pairs are then used to fine-tune the pre-trained BERT model to predict unseen pairs retrieved by the same initial ranker on test queries. The pairs are then re-ranked based on the probabilities generated by BERT.

3) Rank Fusion: The re-ranked list is then used to produce the final output. Existing re-ranking methods include ignoring

978-1-7281-5382-7/20/$31.00 ©2020 IEEE

the initial ranking and using the re-ranked BERT results directly [8], or interpolating scores from the two lists [9]. However, these are not suitable when the scores of the two ranked lists are not comparable. Instead, we use reciprocal rank fusion (RRF) [12] to combine the BERT re-ranked results with the initial rank. RRF sorts the passages using the formula:

$$RRF_{score}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} \qquad (1)$$

where D: set of documents d, R: set of rankings r, $r(d)$: ranking of d in r, and k: smoothing constant. The two ranked lists are combined using Eq.1 to produce the final ranking.

III. Experiments

1) Data set: We evaluated the approach on the scientific abstract dataset of the TREC Precision Medicine track. The corpus included a snapshot of over 26.7 million MEDLINE articles and articles from two cancer-related conference proceedings. The 80 synthetic case-based query topics (30 from 2017 and 50 from 2018) consisted of information about the type of cancer, a specific genetic variant of interest, and patient demographic information. Topics and relevance judgments from 2017 were used to train, and topics from 2018 to test.

2) Experimental setup: All topics were parsed into a series of queries for the initial retrieval step, and relevant passages were retrieved using BM25 and query likelihood. The two ranks were combined using RRF (Eq.1) to produce the initial ranked list. The top 500 ranked passages were output for each query.

The training data used to fine-tune BERT came from the 15K labeled query-passage pairs for the train topics using iterative retrieval strategy. These were fed to the BERT base model (cased, 12-layer, 768 hidden states, 110 million parameters) with a batch size of 32 and the total input length of 512 tokens. We fine-tuned the model with the learning rates of 2e-5 and 1e-5 for 32 epochs and selected the best-performing run to re-rank the retrieved passages on the test topics. We experimented with the original BERT model and two versions of BioBERT that added PubMed abstracts and then PubMed Central (PMC) articles to the training corpus [10]. Both of these were pre-trained with BERT base architecture over the same vocabulary created with WordPiece tokenization.

The results are presented using inferred NDCG, R-precision, and Precision@10. We compared eight systems: (a) **Iterative Retrieval Pattern (IRP)**: pre-ranker based on iterative retrieval on a set of parsed queries; (b) **BERT**: Use list re-ranked by BERT, as suggested by [8]; (c) **BioBERT (PubMed)**: pre-trained weights of the BERT model set to PubMed version of BioBERT; (d) **BioBERT (PubMed + PMC)**: same as (c), except pre-trained also on PubMed Central full-text articles; (e) **IRP + BERT**: combines initial ranker with BERT re-ranker, pre-trained on original corpus, based on reciprocal rank fusion (RRF) as described in Sec. II-3 with $k = 60$; (f) **IRP + BioBERT (PubMed)**: same as (e), except using (c) instead of BERT; (g) **IRP + BioBERT (PubMed + PMC)**: same as (e), except using (d) instead of BERT; (h) **IRP + BioBERT (tuned RRF)**: same as (g), except with $k = 120$. Table I shows the performance of these systems and compares against the state-of-the-art systems from TREC 2018 that have the highest single score for the three metrics [13]. The run with the highest R-precision is based on iterative retrieval, which is our initial ranker. Hence, we can directly evaluate the re-ranking strategy with BERT by comparing it against this run. The paired t-test on the R-precision showed a p-value lower than 0.01, a statistically significant improvement using the BERT re-ranker.

TABLE I
Results of our proposed approach (IRP + BERT) and best systems presented in TREC 2018 Precision Medicine track

	infNDCG	R-prec	P@10
Iterative Retrieval Pattern (IRP)	0.5481	0.3652	0.6140
BERT	0.5295	0.3240	0.5440
epochs: 20 BioBERT (PubMed)	0.5494	0.3416	0.5840
BioBERT (PubMed+PMC)	0.5614	0.3566	0.5920
IRP + BERT	0.5888	0.3803	0.6340
IRP + BioBERT (PubMed)	0.5921	0.3918	0.6400
IRP + BioBERT (PubMed+PMC)	**0.5975**	**0.3955**	0.6740
IRP + BioBERT (tuned RRF)	0.5973	0.3920	0.6840
Best run for infNDCG [5], [13]	*0.5626*	0.3214	0.6680
Best run for R-prec [3], [13]	0.5515	*0.3684*	0.6140
Best run for P@10 [4], [13]	0.5605	0.3656	***0.7060***

As shown in Table I, re-ranking with original BERT gives lower performance than the initial ranker, but BioBERT can improve over the initial ranker on inferred NDCG. Its poor performance on the other metrics could be due to two reasons. First, both BERT and BioBERT use the same vocabulary based on the WordPiece tokenization, which could impact BERT's effectiveness in the biomedical domain. Second, as WordPiece tokenization breaks words down into subword units, the efficacy to re-rank longer passages reduces.

For the combination of the two ranked lists, RRF is shown to be effective in using the re-ranked output by BERT by improving the results on inferred NDCG with 6.2% and R-precision with 6.8% (paired t-test, $p < 0.005$), compared with the state-of-the-art approaches. The performance on Precision@10 did not exceed the previous best run,

The last run in Table I shows that tuning k (set to 120) increases Precision@10 while maintaining the performance on the other two metrics.

IV. Conclusion

In this work, we proposed combining BERT as a re-ranker in the information retrieval model to search for biomedical literature effectively. Experiments with the scientific abstract dataset of TREC Precision Medicine show improvements on inferred NDCG by 6.2% and R-precision by 6.8%, compared with the state-of-the-art systems. The study offers insights into how pre-trained models can be effectively incorporated in a retrieval system that focuses on a biomedical literature search for precision medicine.

REFERENCES

[1] T. R. Goodwin, M. A. Skinner, and S. M. Harabagiu, “Utd hltri at trec 2017: Precision medicine track,” in *National Institute of Standards and Technology (NIST)*, 2017.

[2] B. Xu, H. Lin, and Y. Lin, “Learning to refine expansion terms for biomedical information retrieval using semantic resources,” *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[3] J. Liu, C. Kronk, W.-C. Su, D. T. Wu, and V. V. Vydiswaran, “Retrieving scientific abstracts iteratively: Medier at trec 2018 precision medicine track.” in *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC*, E. M. Voorhees and A. Ellis, Eds., Nov. 2018.

[4] M. Oleynik, E. Faessler, A. M. Sasso, A. Kappattanavar, B. Bergner, H. F. da Cruz, J.-P. Sachs, S. Datta, and E. Böttinger, “Hpi-dhc at trec 2018 precision medicine track,” in *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC*, E. M. Voorhees and A. Ellis, Eds., Nov. 2018.

[5] X. Zhou, X. Chen, J. Song, G. Zhao, and J. Wu, “Team cat-garfield at trec 2018 precision medicine track,” in *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC*, E. M. Voorhees and A. Ellis, Eds., Nov. 2018.

[6] L. Soldaini, A. Yates, and N. Goharian, “Denoising clinical notes for medical literature retrieval with convolutional neural model,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 2307–2310.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[8] R. Nogueira and K. Cho, “Passage re-ranking with bert,” *arXiv preprint arXiv:1901.04085*, 2019.

[9] W. Yang, H. Zhang, and J. Lin, “Simple applications of bert for ad hoc document retrieval,” *arXiv preprint arXiv:1903.10972*, 2019.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: pre-trained biomedical language representation model for biomedical text mining,” *arXiv preprint arXiv:1901.08746*, 2019.

[11] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.

[12] G. V. Cormack, C. L. Clarke, and S. Buettcher, “Reciprocal rank fusion outperforms condorcet and individual rank learning methods,” in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 758–759.

[13] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, and S. Pant, “Overview of the trec 2018 precision medicine track,” *NIST Special Publication*, pp. 500–331, 2018.