Aliababa DAMO Academy at TREC Precision Medicine 2020: State-of-the-art Evidence Retriever for Precision Medicine with Expert-in-the-loop Active Learning

Qiao Jin[1,2*], **Chuanqi Tan**[2], **Mosha Chen**[2], **Ming Yan**[2]
Songfang Huang[2], **Ningyu Zhang**[3] **Xiaozhong Liu**[4]
[1]Tsinghua University, [2]Alibaba Group
[3]Zhejiang University, [4]Indiana University Bloomington
jqa14@mails.tsinghua.edu.cn

Abstract

This paper describes the submissions of Alibaba DAMO Academy to the TREC Precision Medicine (PM) Track in 2020, which achieve state-of-the-art performance in the evidence quality assessment. The focus of the TREC PM Track is to retrieve academic papers that report critical clinical evidence for or against a given treatment in a population specified by its disease and gene mutation. We use a two-step approach that includes: 1) a baseline retriever using query expansion with Elasticsearch (ES) and 2) an automatic or expert-in-the-loop re-ranker: the automatic re-ranker uses features of the ES scores, pre-trained BioBERT scores, publication type scores and citation count scores; the expert-in-the-loop re-ranker uses expert annotations, fine-tuned BioBERT as well as features used in the automatic re-ranker. For the expert-in-the-loop re-ranker, we use a novel active learning annotation strategy that is sample-efficient: at each iteration of the annotation, 1) we fine-tune the BioBERT using all expert annotations of query-document relevance; 2) we let human experts annotate the actual relevance of the most relevant unannotated query-document pairs predicted by the fine-tuned BioBERT. Our submissions outperform the median topic-wise scores in the phase 1 assessment for general relevance and achieve state-of-the-art performance in the phase 2 assessment for evidence quality. Our analyses show that evidence quality is a distinct aspect than the general relevance, and thus additional modeling of it is necessary to assist IR for Evidence-based Precision Medicine

1 Introduction

Precision Medicine (PM) takes into account individual differences in people's genes, environments, and lifestyles when tailoring their treatment and prevention strategies (Initiative, 2016). PM is widely applied in oncology, where patients with different gene mutations receive different therapies though they have the same type of tumor. To facilitate data-driven approaches of PM, the Text REtrieval Conference (TREC) holds the PM Track annually since 2017. From 2017 to 2019, the TREC PM focuses on finding relevant academic papers or clinical trials of patient topics specified by their demographics, diseases and gene mutations (Roberts et al., 2017, 2018, 2019).

In 2020, the TREC PM focus has been changed to retrieve academic papers that report critical clinical evidence for or against a given treatment in a population specified by its disease and gene mutation. Clinical evidence denotes the patient-centered clinical research that studies the efficacy/safety of a treatment, diagnostic power of a test, prognostic value of a marker etc. In the scheme of Evidence-based Medicine (EBM, Sackett 1997), high-quality clinical evidence, e.g.: systematic reviews and randomized controlled trials, should be used to guide clinical practice, including PM.

Traditional Information Retrieval (IR) systems are based on BM25 or TF-IDF that only rank documents by their bag-of-word similarity to the input query. Such systems do not consider evidence quality of the documents, so they are unideal to assist Evidence-based Precision Medicine. This issue can be potentially solved by a re-ranker that models evidence quality based on supervised learning. However, training such re-rankers might require a large number of expert annotations, which can be prohibitively expensive. In this work, we propose a novel active learning scheme where biomedical experts iteratively label the most confident predictions of the re-ranker based on a pre-trained language model, BioBERT (Lee et al., 2020).

This paper describes the submissions of Alibaba DAMO Academy to the TREC PM 2020, which achieve state-of-the-art performance in the

[*]Work done during internship at Alibaba.

evidence quality assessment. Section 2 gives an overview of our two-step approach. Section 3 and Section 4 introduce the retriever and re-ranker in our two-step approach, respectively. Section 5 shows the competition results and analyses. We conclude this paper in Section 6.

2 Methodology Overview

We use a two-step approach to retrieve relevant PubMed documents for each given PM topic[1]: 1) A baseline retrieval strategy that is fast and can generate a relatively small number (e.g. thousands) of candidates out of millions of PubMed articles; 2) A re-ranker that finely re-ranks the retrieved documents based on their evidence quality. Ideally, the former step should guarantee high recall and the latter should have high precision.

Baseline retrieval strategy It is based on ElasticSearch[2] where the original queries are expanded by a list of weighted synonyms. The same baseline retrieval strategy is used by all submissions. We will introduce it in Section 3.

The re-ranker We use two types of re-rankers for the challenge: 1) The automatic re-ranker which unsupervisedly ranks the candidates by features of a pre-trained BioBERT and several article-level features; 2) The expert-in-the-loop re-ranker that predicts the relevance by active learning from interactive annotations of a biomedical expert. Each submission uses either the automatic re-ranker or the expert-in-the-loop re-ranker. We will discuss the re-rankers in Section 4.

3 Baseline Retrieval Strategy

We index the titles and abstracts of all PubMed articles using ElasticSearch.

For each topic, we denote its <disease> as d, the gene <variant> as g and the <treatment> as t. We find the synonyms of the d and g via the National Library of Medicine's web API: `https://ghr.nlm.nih.gov/condition/{d}?report=json` and `https://ghr.nlm.nih.gov/gene/{g}?report=json`, respectively. We denote the retrieved synonyms of d and g as $\{d_1, d_2, ..., d_m\}$ and $\{g_1, g_2, ..., g_n\}$, where $d_1 = d$ and $g_1 = g$. We do not expand the <treatment> because the provided term either has no synonym or is used in almost all articles.

For each synonym d_i and g_j, we count their document frequency $df(d_i)$ and $df(g_j)$, and calculate the weights of each synonym:

$$w(d_i) = df(d_i)/Z_d$$

$$w(g_i) = df(g_i)/Z_g$$

where

$$Z_d = \sum_i df(d_i)$$

$$Z_g = \sum_i df(g_i)$$

The weights are used in the ElasticSearch queries to lower the ranks of rare synonyms.

At the highest level, we query the ElasticSearch PubMed indices using a boolean query which *must match* the disease and treatment query, and *should match* the gene query and a list of keywords, where: 1) The disease, treatment and gene queries are all `dis_max` queries composed of their synonyms with the weights as boost factors. The `tie_breaker` is set to 0.8 and the title field has a 3.0 boost factor while that of the abstract field is 1.0; 2) The keyword list includes words like "trial" to serve as a weak classifier for evidence-based precision medicine papers.

We set the maximum number of retrieved documents of each topic as $10,000$. On average, we retrieve $1,589$ documents for each topic.

4 The Re-ranker

We submitted 5 systems, namely damoespb1, damoespb2, damoespcbh1, damoespcbh2, damoespcbh3. They use different re-rankers to rank the same set of documents retrieved by the baseline retrieval strategy described in Section 3. Damoespb1 and damoespb2 use the automatic re-ranker (Section 4.1) and are 'automatic' runs that don't rely on expert involvement. Damoespcbh1, damoespcbh2 and damoespcbh3 use the expert-in-the-loop re-ranker (Section 4.2). They are 'manual' runs that utilize expert annotations.

4.1 Automatic Re-ranker

The automatic re-ranker uses features of $\{a, b, c, d\}$, corresponding to:

a) the scores returned by the ElasticSearch;

[1] In TREC PM, each topic is a patient query that includes his/her disease, gene mutation and a potential treatment.

[2] `https://www.elastic.co/`

Publication Type	Score
Published Erratum	-2
Retraction of Publication	-2
Editorial	-1
Comment	-1
Journal Article	0
Review	0
English Abstract	0
Letter	0
Observational Study	1
Case Reports	1
Clinical Trial	2
Systematic Review	2
Meta-Analysis	2

Table 1: Mappings between publication types and clinical evidence quality scores.

b) the relevance scores predicted by a pre-trained BioBERT. BioBERT (Lee et al., 2020) is a pre-trained biomedical language model that can be fine-tuned to perform downstream tasks. We use the annotations from the previous TREC PM challenges to fine-tune the BioBERT: Specifically, we collect 54.5k topic-document relevance annotations from the qrel files of TREC PM 2017-2019. We only use the <disease> and <variant> fields of the topics as input and fine-tune the BioBERT to predict their normalized relevance in the annotations.

c) the publication type score. PubMed also indexes each article with a publication type, such as journal article, review, clinical trials, etc. We manually rate the score of each publication type based on the judgments of evidence quality. Our publication type and score mapping is shown in Table 1.

d) the citation count score. We rank the citation count of all PubMed articles and use the quantile of a specific article's citation count as a feature. Similar to but simpler than PageRank, this feature is designed to reflect the community-level importance of each article.

Finally, the re-ranking score of document i is calculated by:

$$r_i = w_a * \frac{a_i}{a_{\max}} + w_b * \frac{b_i}{b_{\max}} + w_b * \frac{c_i}{c_{\max}} + w_d * \frac{d_i}{d_{\max}}$$

where a_i, b_i, c_i, d_i are the above features of document i, $a_{\max}$, $b_{\max}$, $c_{\max}$, $d_{\max}$ are the maximum features among all documents, w_a, w_b, w_c and w_d are the weights associated with different features and are determined empirically (shown in Table 2).

4.2 Expert-in-the-loop Re-ranker

We show the expert-in-the-loop annotation strategy in Figure 1. A senior M.D. candidate (first author) is employed to annotate the evidence quality of a document for a given topic based on the criteria shown in Algorithm 1.

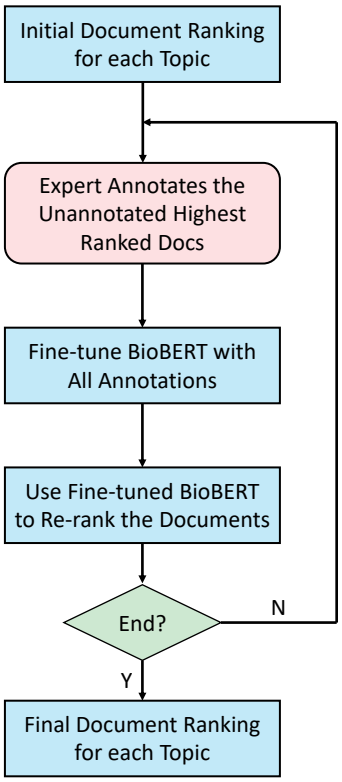

Figure 1: The architecture of our expert-in-the-loop annotation strategy.

We utilize the expert annotations to train a simple linear regression model using the above-mentioned features (a, b, c and d, in Section 4.1). The linear regressor then predicts the scores for all retrieved q-d pairs.

Besides, the manual re-ranker also uses features of: e) the relevance scores predicted by a fine-tuned BioBERT. The BioBERT is fine-tuned by $1,000$ query-document pairs we annotate.

For the annotated query-document pairs, we directly use the expert annotation. For the unannotated ones, the re-ranking score is a weighted sum of the relevance scores predicted by the fine-tuned BioBERT and the linear regression model:

$$r_i = w_{\text{lr}} * \text{LR}(a_i, b_i, c_i, d_i) + w_{\text{e}} * e_i$$

Features	ES score (w_a)	PT BioBERT (w_b)	Pub. Type (w_c)	Cit. Count (w_d)	Linear Reg. (w_{lr})	FT BioBERT (w_{e})	Expert
Weights in Submitted Systems							
damoespb1	1.0	0.5	1.5	0.0	–	–	–
damoespb2	1.0	0.5	1.0	0.0	–	–	–
damoespcbh1	–	–	–	–	1.0	1.0	*
damoespcbh2	–	–	–	–	1.0	2.0	*
damoespcbh3	–	–	–	–	1.0	5.0	*
Pearson r (%) with Qrels							
General Relevance (Phase 1)	38.92	**57.71**	-4.35	-6.21	13.41	37.33	21.57
Evidence Quality *ori* / *exp* (Phase 2)	7.52 4.74	6.21 3.38	6.96 8.06	25.64 27.72	27.28 28.16	**33.09** 28.47	29.37 **30.73**

Table 2: Feature weights of our 5 submissions and their correlations to the official scores. damoespb1 and damoespb2 are 'automatic' submissions, while damoespcbh1, damoespcbh2, damoespcbh3 are 'manual' submissions. *ori* / *exp*: the original and exponential scores of the phase 2. PT: pre-trained. FT: fine-tuned. Pub.: publication. Cit.: citation. Reg.: regressor. *Used when available.

Algorithm 1 `ExpertAnnotation`

Input: A document a (i.e. title and abstract of a PubMed article) and a topic q. q_d, q_g and q_t denote the topic's <disease>, <variant> and <treatment> fields, respectively.
Output: A relevance score $r \in [0, 1]$.
Let the expert annotate whether q_d, q_g and q_t are mentioned in a, getting $r_d \in \{0, 1\}$, $r_g \in \{0, 1\}$, $r_t \in \{0, 1\}$.
if $r_d = 1$ and $r_t = 1$ **then**
  Let the expert annotate whether treatment q_t for the disease q_d is the focus of document a, getting $f \in \{0, 1\}$
  if $f \neq 0$ **then**
    Let the expert annotate whether the document a discusses mono-treatment of q_t, getting $m \in \{0, 1\}$.
    Let the expert annotate the evidence quality e of document a for the topic q, $e \in [-1, 2]$.
    return $(r_d + r_g + r_t + f + m + e)/7$
  else
    return $(r_d + r_g + r_t)/7$
  end if
else
  return $(r_d + r_g + r_t)/7$
end if

5 Results

5.1 Main Results

The assessments of this year's TREC PM are divided into 2 phases, where phase 1 is "Relevance Assessment" and phase 2 is "Evidence Assessment". Phase 1 judgment follows the settings of previous years of the track, while phase 2 focuses on finding high-quality evidence (e.g.: systematic reviews and randomized controlled trials) for the given cancer treatment. The results are shown in Table 3.

In the phase 1 assessment, most of our submissions score higher than the topic-wise median submission, but the topic-wise best submission outperforms our submissions by large margins. Though utilizing expert annotations for learning, our manual runs (damoespcbh1-3) show no significant improvements over our automatic runs (damoespb1-2) in phase 1 assessment. We mainly model the evidence quality in our systems, which might give low scores to the highly related but of low evidence quality documents (e.g.: narrative reviews) and lead to the unideal performance in the phase 1 assessment.

In the phase 2 assessment, our submission damoespcbh3 and damoespcbh1&2 achieve the highest scores for NDCG@30 (0.4519) and NDCG@5 (0.4543), respectively.

5.2 Analyses

In Table 2, we show the Pearson correlations between the used features and the official relevance scores in both phases.

General Relevance (Phase 1): BioBERT that is further pre-trained by the annotations of previous TREC PM has the highest correlation with the

Submission	Evidence Quality (Phase 2)		General Relevance (Phase 1)		
	NDCG@30	NDCG@5	inferred NDCG	Precision@10	R-Precision
Best	**0.4519**	**0.4543**	**0.6818**	**0.7194**	**0.5603**
damoespcbh3	**0.4519**	0.4527	0.4424	0.4742	0.3472
damoespcbh1	0.4497	**0.4543**	0.4304	0.4742	0.3410
damoespcbh2	0.4495	**0.4543**	0.4384	0.4710	0.3414
damoespb1	0.4255	0.4357	0.4533	0.4742	0.3593
damoespb2	0.4254	0.4203	0.4112	0.4452	0.3237
Median	0.2857	0.2529	0.4316	0.4645	0.3259

Table 3: Topic-wise averaged performance of different submissions in the evaluation.

phase 1 scores, which is not surprising since such annotations are also about general relevance. The ES scores achieve the second highest correlation.

The expert annotations for the evidence quality, however, have only 0.2157 Pearson correlation with the general relevance scores. This indicates that relevant papers might not have high evidence quality. Surprisingly, the features of publication types and the citation counts, which are designed for the evidence quality ranking and are positively correlated with the evidence quality, are negatively correlated with the general relevance scores. It shows that the two assessment phases might have some opposite considerations.

Evidence Quality (Phase 2): we show the correlations between different features and the evidence quality scores. The trends are similar in both the original and exponential scores. BioBERT fine-tuned by the expert annotations achieves comparable performance to the expert annotations, and they are the most correlated features. Besides, the fine-tuned BioBERT outperforms the expert annotations by a large margin (37.33 v.s. 21.57) in the phase 1 assessment, indicating that it can re-rank the documents by evidence quality while remaining the original general relevance ranks.

The most correlated features of phase 1, i.e.: the pre-trained BioBERT and the ES score, have the lowest correlations with the phase 2 scores, which further confirms that the evidence quality assessment is distinct from the general relevance assessment. Interestingly, the simple citation count feature has high correlations with the evidence quality scores, probably because the highly-cited papers usually provide critical clinical evidence for precision medicine.

6 Conclusions

In this paper, we present the winning solution in the phase 2 of TREC Precision Medicine 2020. It uses 1) an ElasticSearch-based retriever with query expansion and keyword matching and 2) a re-ranker based on article features and the BioBERT fine-tuned by annotations from expert-in-the-loop active learning. Analyses show that the evidence quality is a distinct aspect than the general relevance, and thus additional modeling of it is necessary to assist IR for Evidence-based Precision Medicine.

References

The Precision Medicine Initiative. 2016. The precision medicine initiative.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, and Alexander J Lazar. 2018. Overview of the trec 2018 precision medicine track.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, and Shubham Pant. 2017. Overview of the trec 2017 precision medicine track. In *The... text REtrieval conference: TREC. Text REtrieval Conference*, volume 26. NIH Public Access.

Kirk Roberts, Dina Demner-Fushman, Ellen M Voorhees, William R Hersh, Steven Bedrick, Alexander J Lazar, Shubham Pant, and Funda Meric-Bernstam. 2019. Overview of the trec 2019 precision medicine track.

David L Sackett. 1997. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier.