Brown University at TREC Precision Medicine 2018

Prakrit Baruah, Riya Dulepet, Kyle Qian and Carsten Eickhoff
Brown University, USA
carsten@brown.edu

Abstract

This paper describes Brown University's submission to the TREC 2018 Precision Medicine (PM) track. This report details our efforts to query processing and retrieval modeling. As a key difference to last year's edition of the task, we are able to include supervised rankers based on existing patient-document relevance labels. We use this information in the form of cubic rank transformations between initial weak rankers and the final ranking. The empirical evaluation is based on 50 synthetic precision oncology patients and shows solid retrieval performance.

1 Introduction

Precision medicine is a modern field of study that aims to use genomic information in finding more effective treatments for patients. Due to the popularity of the new paradigm, the volume of annually published scholarly precision medicine articles has been growing rapidly in recent years. While this considerable amount of scientific research holds a rich and ever increasing well of knowledge, its sheer scale makes it intractable for manual inspection and mandates the development of dedicated automatic retrieval facilities.

In this paper, we present a supervised modular patient-centric information retrieval system for use in precision oncology settings. Based on patient demographics as well as information regarding the type of tumor and its genetic composition, we rank scholarly articles as well as clinical trials with respect to their relevance for a range of reference patients.

While the previous edition of this task exclusively relied on unsupervised approaches, this year, explicit relevance judgments are available and have opened up a range of supervised learning to rank techniques. Eventually, in an attempt to retain retrieval model generality, we decided against the use of highly parametric end-to-end trained rankers and instead propose an ensemble of light-weight retrieval models.

2 Methodology

We rely on literal as well as neural query expansion (Section 2.2) and ensembles of weak rankers whose initial rankings are transformed in a supervised manner using TREC PM 2017 data (Section 2.3).

2.1 Collection Indexing

The collections of scholarly abstracts and clinical trials, respectively, are indexed using Apache Lucene [2] and Elasticsearch [3]. We apply only standard processing and tokenization, stemming and stop word removal.

27th Text Retrieval Conference (TREC) Notebook 2018.

2.2 Query Processing

Before issuing queries to our retrieval system, all fields are projected to lower case characters and stemmed. Any punctuation characters and stop words are removed.

Next, we use two query expansion approaches to broaden the topical scope of the derived queries. The first approach involves gene name resolution using a thesaurus of HUGO Gene Nomenclature Committee (HGNC) data [1], and Medical Subject Headers (MeSH) for diseases. From HGNC, we extract the "approved names" and "synonyms" fields. In the case of MeSH, we extract the "entry terms" field. The thus obtained thesauri are used to append synonyms for the individual terms in the "genes" and "diseases" categories. The second approach involves using a word2vec [6] model trained on medical literature that maps individual words of the corpus to dense vectors. Word2vec was applied to a corpus of 10,876,004 English abstracts of biomedical articles from PubMed [7]. The resulting vectors are then used to predict the single most similar word for each term in the query.

The above methods allow for the creation of three query expansion variants: (1) Thesaurus, (2) Word embeddings (3) Both. Section 3 will discuss which expansion scheme is used in our official runs.

2.3 Retrieval Models

To retrieve relevant clinical trials and scientific papers, we perform three weak ranking passes using BM25 [8], a language model with Jelinek-Mercer smoothing and one with Dirichlet smoothing. These initial rankings can be refined by an optional re-ranking step in a cubic transformation whose parameters a_0, a_1, a_2, a_3 are found via linear regression on TREC PM 2017 relevance judgments [9].

$$rank' = a_0 + a_1 \ln rank + a_2 \ln rank^2 + a_3 \ln rank^3$$

The rankings induced by our three constituent retrieval models are then fused using either the CombSUM or CombMNZ schemes [4].

For scientific abstract searches, the contribution of the "title" field was boosted to twice the weight of the "abstract" field. For clinical trial searches, any trials not matching the patient's age and gender were removed. Here, the "official title" and "brief summary" fields were given twice the weight of any other fields.

3 Results

We submitted ten official TREC Precision Medicine 2018 runs for evaluation, five for scientific abstract retrieval and five for clinical trial retrieval. Table 1 lists the official abstract retrieval performance of each run in terms of inferred nDCG, precision at 10 retrieved documents and R-precision.

mnzAbs combines raw retrieval model scores in a CombMNZ scheme. Queries are not expanded.

sumAbs combines raw retrieval model scores in a CombSUM scheme. Queries are not expanded.

sumEW combines raw retrieval model scores in a CombSUM scheme. Queries are expanded using both word embeddings and thesauri.

cubicmnzAbs applies the cubic rank transformation and fuses the rankings using CombMNZ. Queries are not expanded.

cubicsumWAbs applies the cubic rank transformation and fuses the rankings using CombSUM. Queries are expanded via word embeddings.

We observe the scores of all five runs to lie close by each other with only mild differences between the rankings induced by individual runs. The relative ranking of runs is metric dependent, but there seems to be a weak signal in favor of query expansion to improve early precision. At the time of writing this report, no official evaluation results for the clinical trial task had been made available.

Table 1: Experimental results for scientific abstract retrieval.

Run Identifier	infnDCG	P@10	R-Prec
mnzAbs	0.391	0.480	0.234
sumAbs	0.391	0.480	0.235
sumEW	0.400	0.496	0.231
cubicmnzAbs	0.391	0.480	0.234
cubicsumWAbs	0.374	0.498	0.207

4 Conclusion

This report describes Brown University's entry to the TREC 2018 Precision Medicine Track, ranking scientific abstracts and clinical trial descriptions in response to structured precision oncology patient profiles. Our method relies on neural, as well as literal query expansion, cubic score transformations and retrieval model fusion. The scientific abstract retrieval results indicate only a mild precision-favoring effect when introducing query expansions. Future directions of inquiry include more aggressive expansion schemes [5] as well as weakly supervised variants of powerful neural retrieval models [10].

References

[1] Gene [internet]. bethesda (md): National library of medicine (us), national center for biotechnology information, 2004. [Cited 2017 Oct 15]. Available from: https://www.ncbi.nlm.nih.gov/gene/.

[2] Doug Cutting, M Busch, D Cohen, O Gospodnetic, E Hatcher, C Hostetter, G Ingersoll, M McCandless, B Messer, D Naber, et al. Apache lucene, 2008.

[3] BV Elasticsearch. Elastic search, 2015.

[4] Edward A Fox and Joseph A Shaw. Combination of multiple searches. *NIST special publication SP*, 243, 1994.

[5] Simon Greuter, Philip Junker, Lorenz Kuhn, Felix Mance, Virgile Mermet, Angela Rellstab, and Carsten Eickhoff. Eth zurich at trec 2016 clinical decision support. In *NIST Special Publication 500-321: The Twenty-Fifth Text REtrieval Conference Proceedings (TREC 2016)*. National Institute of Standards and Technology, 2016.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[7] Ioannis Pavlopoulos, Aris Kosmopoulos, and Ion Androutsopoulos. Continuous space word vectors obtained by applying word2vec to abstracts of biomedical articles, 2014.

[8] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[9] TREC. Precision medicine track, 2017. Available at: http://www.trec-cds.org/2017.html [Online; accessed 2017 Oct 18].

[10] Xing Wei and Carsten Eickhoff. Distant supervision in clinical information retrieval. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2018.