KlickLabs at TREC 2018 Precision Medicine track

Lediona Nishani, Maheedhar Kolla, Gaurav Baruah
Klick Labs, Klick Inc., Toronto, Canada.
{lnishani,mkolla,gbaruah}@klick.com

Abstract

Precision medicine aims to provide the most accurate course of treatment, personalized for each patient. The 2018 Precision Medicine (PM) track aims to build systems, that sift through biomedical articles to find relevant cancer treatments for patients, as well as search for clinical trials for which a patient might be eligible. Cancer, as a disease, can occur in various forms, be of different types, and affect multiple organs as well as bodily systems. Additionally, every patient–cancer combination presents with slight variations depending on genetics, age, sex, and other factors. As of 2017, 16% of PubMed articles related to cancer [5], and over 500,000 articles are added to PubMed each year, thereby, making quality information retrieval for PM an important problem to tackle, for the benefit of doctors and patients alike. In this work, we develop query-expansion methods, with an aim to compare their performance over the 2018 PM task. Our experiments indicate that medical ontology (NCIt) based expansion performs well for retrieving scientific abstracts from PubMed. Utilizing demographic information in queries improves the performance for clinical trial retrieval.

1 Introduction

The 2018 TREC Precision Medicine (PM) track aims to evaluate systems that identify the most relevant scientific articles and clinical trials relevant to patients suffering from cancer. Such shortlisting of pertinent documents for each patient could potentially aid the oncologist/physician to prescribe a course of treatment or medication, with high precision, for the particular patient.

The 2018 TREC PM track provides 50 synthetic cancer cases, developed by oncologists. Each patient case (topic) consists of three primary fields: Disease, Gene and Demographic. Disease field entails the specific type of sickness (form of cancer) that the patient is suffering from. Gene field comprises the genetic variants that are responsible for the disease, which can involve abnormal genes or particular variants of the gene. Demographics includes age and gender. The participants are tasked with retrieving the most relevant scientific abstracts from PubMed for each patient. Additionally, clinical trials, that are likely relevant to the patient, and for which the patient is eligible for, are also required to be retrieved from a snapshot of clinicaltrials.gov. Given that there are over 26 million scientific articles in PubMed, and there are over 240,000 clinical trials listed at clinicaltrials.gov, retrieving highly relevant documents for a patient is paramount for physicians to make informed decisions about patient care and treatment.

In order to generate high performance run for both tasks, we explored various query expansion strategies. We investigated expanding queries with a list of synonyms generated from UMLS and NCIT knowledge bases. We also tried to generate synonyms based on word embedding similarity, for query expansion. Finally, we tried a fusion based approach that fused results from differently expanded queries.

We submitted 4 runs for the first task and five runs for the second task. Based on our experiments, we find that using NCIt synonyms for query expansion improved results over our baseline queries for scientific abstract retrieval.

2 Task, Datasets, and Evaluation Metrics

The 2018 PM track specifies two sets of data, one for each respective task: scientific abstracts, and clinical trials. For both tasks, a common set of 50 topics are utilized as the query set.

2.1 Topic Structure

Each topic is a synthetic medical case developed by oncologists from the University of Texas MD Anderson Cancer Center. Each topic consists of the following fields:

- Disease: specifies the disease category (the type of cancer the patient is suffering from)
- Gene: specifies the gene responsible, or specifies several mutated genes and variations
- Demographics: specifies the age and the gender of the patient.

2.2 Document Corpora

For scientific abstracts, the set of documents consists of 26.6 million PubMed abstracts, along with supplementary abstracts from American Association for Cancer Research (AACR) proceedings. For clinical trials, the set of documents is a collection of 240,000 clinical trials from `clinicaltrials.gov`. Clinical trials are semi-structured articles, which have specific fields for covering information.

2.3 Evaluation Metrics

The primary metrics for the track are infNDCG [7], P@10 and R-precision. We also utilized MAP for comparing systems before submitting our runs.

3 Additional/External data used for developing systems

3.1 NCIt

NCI thesaurus (NCIt-part of the National Cancer of Institute) [6] encompasses a comprehensive classifier of cancer terminology and medical ontologies. It involves various human and machine literate concept of thousands of neoplasms in a systematic logic parentchild structure. This ontology provides mappings of concepts between NCI and all other medical resources, thus facilitating data analysis and exploration across medical dictionaries and ontologies.

3.2 UMLS

UMLS-Unified Medical Language System [1] represents a repository of biomedical vocabularies designed from US National Library of Medicine. It comprises over 2 million names for more than 900,000 concepts for biomedical vocabularies, and additional 12 million associations across the underlying concepts. UMLS contains concepts (that could have many different names), and each concept has an associated Concept Unique Identifier (CUI). CUIs can be used to link those names with other medical ontologies and vocabularies (like NCIt). We identify CUIs from UMLS for query terms [9] in the provided topic. For instance, given the disease "melanoma", we find its CUIs from UMLS

$$\text{“}C3539018'' : \text{“}Melanoma'', \text{“}C0025202'' : \text{“}Melanoma'', \text{“}C0796561'' : \text{“}melanoma''$$

These CUIs help in finding synonyms from NCIt for query expansion.

4 Systems development and Runs

We primarily explore query expansion methods for the 2018 PM track tasks. We utilize ElasticSearch[1] as our underlying search engine to index and retrieve documents.

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

4.1 Indexing

For both tasks (scientific articles and clinical trials) the available documents in the corpus have been indexed by making use of ElasticSearch [3]. We created a separate index each for PubMed abstracts and clinical trials data. We indexed all fields of the clinical trials data. We extracted from each PubMed abstract the following fields for indexing:

- PMID: represents the unique identifier that serves as documents ID for submissions runs
- Abstract: specifies the full text of abstract, which summarizes the content of the paper
- Title: the title of the article
- Mesh-heading: the medical subject headings that are related to the articles
- Publication Type: the categories of publications and MeSH terms used to classify the articles categories
- Chemicals: Mesh terms related to all registered chemical compounds that are pertinent to an article

We have configured the ElasticSearch platform to return results using the Okapi BM25 ranking algorithm [2], with parameters k1 as 1.2 and b as 0.75. The queries to ElasticSearch consisted of the disease name and the gene variant from the topic, as well as expansion terms. The same query was used to retrieve abstracts and trial runs from respective indices. We ranked documents that contained the query terms within a documents title, abstract or mesh headings, and we returned the top 1000 documents, for each query/topic.

4.2 Baseline Run

In this run, we do not perform any query expansion. Our baseline query terms consist of the disease name and the gene variants.

4.3 Query expansion based on similarity between word embeddings

We obtained word embeddings [8] constructed over PubMed, PMC texts and Wikipedia; the word vectors were constructed using word2vec [4]. Word embeddings are essentially representation of words depending on their collocation context with surrounding words. Thus, words that appear in similar contexts may have word vectors that are close to each other. It is therefore possible to generate a list of likely synonyms for a given words word embedding vector.

We first compute an average word embedding for our baseline query. We find the 5 most similar word vectors to the average vector. We finally add the terms corresponding to the most similar word embeddings to the query.

4.4 Expansion with NCIt (synonyms) as knowledge base

To potentially increase recall, we utilize the NCIt thesaurus to find parent and child synonyms of the disease. Parent synonyms could be general terms for the specific cancer, and child synonyms could be more specific terms. For example, for the disease "liposarcoma", NCIt lists as parent synonyms, "Sarcoma" , "Malignant Lipomatous Neoplasm", and NCIt lists as child synonyms, "Myxoid Liposarcoma" , "Primary Liposarcoma" , "Adult Liposarcoma" , "Mixed Liposarcoma".

For this system, we first find UMLS CUIs for the query terms. In NCIt, each cancer related term is associated with a CUI. We use the query terms CUIs and look up the parent and child synonyms from the NCIt thesaurus. We add these synonyms to the baseline query for expansion.

4.5 Fusion runs (combine all runs)

For some Information Retrieval problems, fusion approach works very well, where results of multiple ranking algorithms are combined together. We combine the ranked list of documents from each of the previous query expansion methods and constructed a single ranked list of documents using Reciprocal Rank Fusion [2].

[2] https://en.wikipedia.org/wiki/Okapi_BM25

4.6 Baseline with filtering (for clinical trials retrieval only)

In order to increase the likelihood of retrieving the relevant clinical trial for an eligible patient, we leveraged the demographic field of the topic . In this run we didnt employ any query expansion strategy. The formulated query involves information from the gender together with disease and gene of the patient. The baseline with filtering run appeared to outperform all the other clinical trials approaches over all metrics.

5 Results and Discussion

We developed the systems (listed below) based on different techniques of query expansion. Each system returned results for each of the 50 topics, for both, system abstracts and clinical trials.

- KLPM18T2Bl/KLPM18T1Bl: we construct a query that combines the terms from disease and gene for each topic.
- KL18absWV/KL18TrialWV: we expand the query by obtaining synonyms based on the similarity between the PubMed word embeddings.
- KL18absHY/KL18TriHY: we exploit the hierarchy from NCIT and extract parent synonym diseases and child synonym diseases to expand the query.
- KL18AbsFuse/KL18TriFuse: a system that fuses the retrieved results from the previous 3 systems using RRF.
- KL18TrialBF: baseline with demographic filtering. This was implemented for Clinical trial retrieval task only.

5.1 Scientific Abstracts results

runName	P10	infNDCG	R-prec	total	num rel retrieved
KLPM18T2Bl	0.50	0.4347	0.261	50000	3359
KL18absWV	0.284	0.2233	0.129	49950	2096
KL18absHY	0.534	0.4432	0.287	50000	3566
KL18AbsFuse	0.54	0.4367	0.274	50000	3422

Table 1: Abstract Runs results: Final runs for the 2018 topics for the scientific abstract task

In Table 1, we provide the PubMed Abstract results evaluated in P10, infNDCG, R-prec. Our first attempt, baseline run (KLPM18T2B1) performs reasonably well, with P10=0.482 and infNDCG=0.4347. Using NCIT synonym-based expansion (KL18AbsHY) increases performance over all metrics. Interestingly, the KL18absHY results in higher number of relevant articles retrieved than our other methods. Word embedding similarity based expansion (KL18absWV) seems to not improve the performance. The fusion run (KL18ABSFuse) slightly outperforms the NCIt run in terms of P10.

5.2 Clinical Trials results

In Table 2, we provide clinical trial results evaluated in P10, infNDCG, R-prec. The Baseline run of trials (KLPM18T1Bl) seems to perform well. The baseline with filtering(KL18TrialBF) outperformed all our runs over all metrics. Again, we notice that the word embedding similarity expansion KL18TrialWV performed poorly.

runName	P10	infNDCG	R-prec	total	num rel retrieved
KLPM18T1Bl	0.482	0.416	0.346	43522	1520
KL18TrialBF	0.494	0.437	0.358	42979	1538
KL18TrialWV	0.28	0.319	0.227	42662	1439
KL18TriHY	0.484	0.429	0.3524	43158	1518
KL18TriFuse	0.47	0.41	0.317	47097	1540

Table 2: Clinical Trial results: Final runs for the 2018 topics for Clinical Trials task

6 Conclusion

For the TREC 2018 Precision Medicine Track, we submitted 4 runs for scientific abstracts task: finding likely relevant scientific abstracts from PubMed, and submitted 5 runs for the clinical trial task: finding clinical trials tasks relevant for the patient to be qualified for. We observe a slight improvement in performance when using synonym-based query expansion.

Acknowledgments

We would like to thank the KlickLabs team at Klick Inc. for their feedback, and support, for this research.

References

[1] Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology, 2004.

[2] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759, 2009.

[3] Clinton Gormley and Zachary Tong. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Inc., 1st edition, 2015.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[5] Constantino Carlos Reyes-Aldasoro. The proportion of cancer-related entries in pubmed has increased considerably; is cancer truly the emperor of all maladies? *PloS one*, 12(3):e0173671, 2017.

[6] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.

[7] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610. ACM, 2008.

[8] Yongjun Zhu, Erjia Yan, and Fei Wang. Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Med. Inf. & Decision Making*, 17(1):95:1–95:8, 2017.

[9] Guido Zuccon and Bevan Koopman. Sigir 2018 tutorial on health search (hs2018): A full-day from consumers to clinicians. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1391–1394, New York, NY, USA, 2018. ACM.