Team UKNLP at TREC 2017 Precision Medicine Track: A Knowledge-Based IR System with Tuned Query-Time Boosting

Jiho Noh[1] and Ramakanth Kavuluru[2]

***Abstract*—This paper describes the system architecture of the University of Kentucky Natural Language Processing (UKNLP) team's entry for the TREC 2017 Precision Medicine Track. The goal of the challenge is to retrieve useful precision medicine-related information (abstracts, clinical trials) for the given synthetic cancer patient cases, each of which consists of a neoplastic condition, genetic variants, demographic details, and any additional information (e.g., comorbidities). We explored query expansion techniques using well-known broad knowledge sources such as the Unified Medical Language System (UMLS) and the Medical Subject Headings (MeSH) for each abstract, and additional specialized sources such as the Catalogue Of Somatic Mutations In Cancer (COSMIC) database, which allowed us to construct boosted queries. We conducted several experiments with model averaging techniques and our final system architecture placed 6th (in terms of infNDCG and R-prec) among 29 teams that submitted runs to the scientific abstract retrieval task.**

I. Introduction

The 2017 TREC Precision Medicine (PM) Track is a successor of the Clinical Decision Support (CDS) Track from the previous three years. Previously the CDS track focused on finding biomedical articles relevant for answering clinical questions involving medical records. In 2017, the participating teams are asked to provide useful precision medicine-related information to the clinicians who treat cancer patients based on specific patient case information. Each case is described by the disease name (a specific type of cancer this year), the relevant genetic variants (gene and mutation names) for the patient, and other pertinent factors. Table I shows two examples of such cases, termed "topics" for this track. Teams need to construct an automated IR system that can (1) retrieve the most pertinent scientific articles to the topic and (2) recommend clinical studies that are suitable to the patient's case. Two corresponding document collections are provided:

1) Scientific abstracts: We were given the January 2017 snapshot of Medline abstracts for the scientific abstracts scenario. This collection consists of roughly 26 million articles that are stored in a structured XML format. Along with the textual content of the abstracts, detailed metadata of the each article is provided as well including *journal information*, *author list*, and *citation records*. Additionally, the TREC task organizers provided abstracts obtained from the most recent AACR (American Association for Cancer Research) and ASCO (American Society of Clinical Oncology) proceedings. Documents in this collection do not contain any metadata, which necessitates additional care in data pre-processing and querying procedures.
2) Clinical trials: For the clinical trial retrieval scenario, the April 2017 snapshot of ClinicalTrials.gov database was provided in a structured XML format.

[1]Jiho Noh is with the Department of Computer Science, University of Kentucky, Lexington, KY, USA. jiho.noh@uky.edu

[2]R. Kavuluru is the ***corresponding author*** and is with the Division of Biomedical Informatics (Department of Internal Medicine) and the Department of Computer Science, University of Kentucky, Lexington, KY, USA. rvkavu2@uky.edu

II. Methods

Our system follows a traditional information retrieval (IR) architecture with the following components.

1) *Indexing*: Document contents and attributes are the main elements compared between documents and topics for the retrieval task; Often such elements are quantified through term frequencies, the positions of terms in a document, and document lengths.
2) *Topic analysis/expansion*: The same term may be used with a different meaning in a different domain. Due to the ambiguity issues in natural language, additional work, such as query transformation and expansion, is often unavoidable.
3) *Retrieval models*: A variety of scoring techniques are typically explored in the document retrieval phase.
4) *Query-time boosting*: Based on empirically determined or expert defined weights, query-time boosting of terms/phrases can enhance the performance of an IR system.

A. Indexing

We used Apache Lucene (version 6.5.1)[1] to index our document collections with the indexing process slightly differing based on the collection type.

1) Fields of Interest: For each article, based on available metadata, we have chosen a set of fields that we think is highly useful in the precision medicine perspective. For indexing the PubMed abstracts, we used the following fields.

- *Document ID*: PMID, the unique document identifier that is assigned by the National Library of Medicine (NLM)
- *Journal title*: the surface name of the journal title, which assumably defines the domain of the article.
- *Article title*: the title of the article
- *Abstract text*: the abstract of the article in free text
- *MeSH headings and corresponding UMLS identifiers*
- *Chemical names and corresponding UMLS identifiers*

[1]https://lucene.apache.org/core/

	patient 1	patient 2
Disease:	acute lymphoblastic leukemia	thyroid cancer
Variant:	ABL1, PTPN11	RET, BRAF
Demographic:	12-year-old male	63-year-old female
Other:	No relevant factors	Ecog grade of 2

TABLE I: An example of topics that describes the patient's case

We also indexed AACR/ASCO articles with the following matching fields.

- *Document ID*: the provided filename of the abstract is used as the identifier
- *Journal title*: the name of the meeting/conference
- *Article title*: the subject line of the article
- *Abstract text*: the abstract of the article in free text
- *MeSH headings and corresponding UMLS identifiers*
- *Chemical names and corresponding UMLS identifiers*

For clinical trials we extracted and used the following fields.

- *Document ID*: NCT ID, the unique clinical study identifier as it appears on ClinicalTrials.gov
- *Brief title*: a short title of the clinical study
- *Official title*: the full title of the clinical study
- *Brief summary*: a short description of the clinical study
- *Detailed descriptions*: extended description of the protocol
- *Overall status*: the recruitment status
- *Phase*: N/A, Early Phase 1, or Phases 1–4
- *Study type*: the nature of the investigation
- *Condition*: the names of disease or conditions being studied (corresponding MeSH terms or SNOMED CT terms)
- *Intervention*: the intervention(s) associated within the study
- *Eligibility criteria*: a limited list of criteria for selection of participants (inclusion/exclusion criteria)
- *Eligibility gender*: All, Female, or Male
- *Eligibility age*: minimum/maximum age of potential participants eligible for the study
- *Keywords*: keywords that best describe the protocol

2) Normalizing Age Groups: The age and gender information is given in a demographic field (e.g., "45-year-old female"). For a particular disease, this information is cruicial in filtering relevant scientific articles and clinical trials. For the scientific articles, age-related temporal phrases are transformed into corresponding MeSH terms as shown in Table II. The same patterns in the topics are transformed accordingly.

In trials data, the format used for the age information in the eligibility section is not inconsistent. The unit used in this field ranges from 'minute' to 'year'. We normalized this temporal information into a numeric value in 'days'. Given Lucene query parser allows us to use "range queries", the results can be restricted to the range of the numeric age values.

3) Data Pre-processing for the AACR/ASCO Collections: AACR/ASCO abstracts do not contain metadata such as MeSH terms or chemical components. In order to construct a consistent document index, we parsed them by using NLM's Medical Text Indexer (MTI) to obtain MeSH terms and corresponding UMLS concepts. After adding these MeSH terms of the AACR/ASCO documents, we have noticed 10–15% more occurrences of these documents in our top-ranked results.

B. Query/Topic Expansion

In several cases, topic representations are ambiguous leading to irrelevant results. To counter this, query expansion strategies are generally incorporated – Song et al. [4] used Google search results to expand the initial query and You et al. [5] obtained the MeSH terms from the citations of the retrieved documents. We have focused more on disease and gene variations by utilizing UMLS concepts and MeSH terms. The same query expansion techniques are used for searching both the scientific articles and the clinical trials.

1) UMLS Atoms for the Disease: We queried with the disease name mentioned in the topic to find the best matching UMLS concepts through NLM's UTS service. The "preferred" name of the identified UMLS concept often does not align with MeSH descriptions or the name mentioned in the topic. However, it provides a set of highly related key terms to the topic. We obtained up to five closest UMLS concepts and subsequently restricted the results to those arising from the National Cancer Institute (NCI) Thesaurus terminology. From the concept set, we further retrieved atomic phrases (from UMLS) of each concept. All such unique phrases are appended to the disease group when formulating the query. Listing 1 shows how a disease name can be expanded through an example.

2) MeSH Headings for the Disease: MeSH is a frequently used biomedical terminology specifically designed to index biomedical articles to aid in subsequent retrieval by users through the PubMed search engine. Each article is typically assigned 10-15 MeSH headings. MeSH is hierarchical in nature which aids in better understanding a subject at several levels of specificity. We note from Section II-A that we already index MeSH terms for scientific literature abstracts. To be able to do exact matching in the MeSH heading field, we need to

Age Groups	MeSH Terms	Normalized in Days
Birth – 1 month	Infant, newborn	0 – 30
1 month – 24 month	Infant	30 – 720
2 years – 6 years	Child, preschool	730 – 2,190
6 years – 13 years	Child	2,190 – 4,745
13 years – 19 years	Adolescent	4,745 – 6,935
19 years – 45 years	Adult	6,935 – 16,425
45 years – 65 years	Middle aged	16,425 – 23,725
65 years – 80 years	Aged	23,725 – 29,200
80 years – 200 years	Aged, 80 and over	29,200 – 73,000

TABLE II: Normalized age groups for search strategies

Listing 1: The expansion details of the disease name "Pancreatic ductal adenocarcinoma"

```
{
    "disease name": "Pancreatic ductal adenocarcinoma",
        "UMLS": [{
            "code": "C1335302",
            "concept name": "Pancreatic Ductal Adenocarcinoma",
            "atoms": [
                "ductal adenocarcinoma of the pancreas",
                "pancreatic infiltrating duct carcinoma",
                "pancreatic tubular adenocarcinoma"
            ]
        }]
}
```

identify the MeSH term from the disease name in the input topic. For this, we ran NLM's concept mapping tool MetaMap on the disease name to identify UMLS concepts restricted to the MeSH source vocabulary to obtain the corresponding code. Furthermore, we used the synonymous UMLS atoms for the corresponding UMLS concept as part of the query expansion for the disease (in addition to atoms from NCI Thesaurus concepts from Section II-B1).

3) The Genetic Variations: The gene names are often abbreviated (JAK2 for 'Janus Kinase 2') and the variant names are more involved due to mutational information including the amino acid symbols and the loci. For the gene and mutation names, we have separately used the same expansion methods that leverage MetaMap in order to obtain UMLS concepts and associated MeSH terms. For the gene part, the retrieved set of the atoms are strongly relevant to the gene name from the topic and hence there are no concerns in using these phrases as synonyms. However, the UMLS atoms of the mutation part do not always present the mutation at a level that can lend itself to retrieval tasks. For example, V617F corresponds to the following set of atoms: "JAK2 Val617Phe", "Janus Kinase 2 V617F", "NP 004963.1:p.V617F" . Although the last atom gives us additional information (NCBI mutation accession number), searching Medline articles with the accession number does not improve the retrieval quality. Alternatively, adding '*Val617Phe*' to the query string does widen the range of relevant results. We used the contiguous numeric part of the mutation (which is mostly the reference sequence position number) to identify potentially useful tokens, such as '*Val617Phe*' from 617. Consequently, our method also covers the one/three-letter amino acid codes as well.

C. Query Term Weight Optimization

In the scoring procedure for IR systems, each component (term or phrase) in a query can be emphasized to the desired relevance level based on the corresponding perceived level of importance. Popularly called "query-time boosting", this approach can aid in assigning different weights (or boosts) to different sections of each query. In our query template, we introduce multiple boost factors on the constituent groups for each patient case. The Lucene disjunctive query used in our system is

$$\begin{aligned}&+[(D_1\,D_2\ldots)^{w_1}\,(\texttt{mt}:D_1'D_2'\ldots)^{w_2}]\\&+[(G_1\,G_2\ldots)^{w_3}\,(M_1\,M_2\ldots)^{w_4})]\\&(\texttt{mt}:G_1'\ldots)^{w_5}\,(\texttt{mt}:M_1'\ldots)^{w_6}\,(\texttt{mt}:F_1'\ldots)^{w_7},\end{aligned}$$

which consists of seven components: (1) the disease names (D_i), (2) the disease MeSH codes (D_i'), (3) the gene names (G_i), (4) the mutation names (M_i), (5) the gene codes in MeSH (G_i'), (6) the mutation codes in MeSH (M_i'), and (7) the demographic codes in MeSH (F_i'). Each query component has a specific boost factor ($w_1, \ldots, w_7$) that needs to be tuned. We notice from the '+' components of the query that we at least require a disease and a gene/mutation variant to be part of the document with all other combinations being disjunctions.

Given the PM track is a new task introduced in 2017, we do not have a dataset with relevance judgments for tuning

	Average performance for Task A				
	infNDCG	R-prec	P@5	P@10	P@30
UKY_BASE	0.3800	0.2303	0.5267	0.4667	0.3756
UKY_CJT	0.3897	0.2333	0.5267	0.4800	0.3711
UKY_AGG	0.3852	0.2518	0.5533	0.4933	0.3944
UKY_COM	0.2572	0.1906	0.3933	0.3833	0.2933
UKY_MAN	0.3666	0.2354	0.5267	0.4867	0.3500

TABLE III: Performances of our runs for TREC PM Track (Task A)

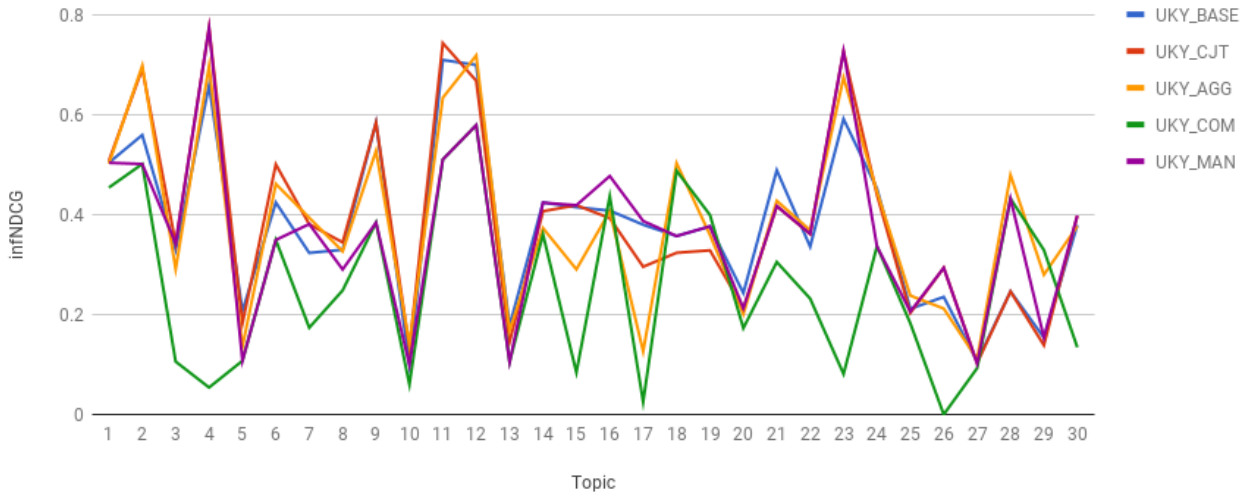

Fig. 1: infNDCG score distribution across all topics for our runs

the systems. To handle this, we used the genetic database of the Catalogue of Somatic Mutations in Cancer (COSMIC)[2] as a proxy for relevance judgements for this task. COSMIC provides a wide range of resources for somatic mutations in human cancer. Via its search engine, we were able to collate research articles with regards to cancer types and associated genes/mutations from our input topics. This allowed us to evaluate our system with ranking measures (infNDCG, infAP, p10, and bpref) and helped us tune the boost factors for our queries based on validation experiments.

D. Retrieval Models

As mentioned above, the relevance judgment data was not available for this year. In order to compare the relevance models with respect to the topics and collections of interest, we examined the results over the previous years' datasets from TREC Clinical Decision Support tracks. For this experiment, we ran tests using the open source Terrier[3] system for the following methods:

- Vanilla TF-IDF without smoothing functions
- The DFR version of BM25 (DFR_BM25) [1]
- Poisson estimation for randomness (PL2) [1]
- Log-logistic DFR model (LGD) [2] [3]
- Inverse expected document frequency model for randomness (In_expB2 and In_expC2) [1]

The *In_expC2* was the best among the models in this experiment. Even though *BM25* was not the best model in this experiment, the difference was negligible compared to the results of the Lucene platform; thus we used Lucene's default *BM25* model without changing any of its settings.

III. Experiments

Five models were prepared for the submission in different aspects.

- **RUN1 (UKY_BASE)**: This is our baseline run where each query component (disease, gene, mutation, demographic info) is combined in a disjunctive query.
- **RUN2 (UKY_CJT)**: To our baseline, we add an additional conjunctive component in which the terms from the initial disjunctive query are ANDed. The idea is to have those documents that match all criteria rise to the top of the ranked results.

[2]http://cancer.sanger.ac.uk/cosmic
[3]http://terrier.org/

- **RUN3 (UKY_AGG)**: The results of RUN2 and RUN4 (see next) are aggregated via Borda count to impose a new ranking
- **RUN4 (UKY_COM)**: In this run, the query-time boost factors are optimized based on the COSMIC reference, and additional MeSH terms are added using the MeSH on Demand (MOD) service (https://meshb.nlm.nih.gov/MeSHonDemand) in addition to the MetaMap results. With some topics, we noticed that the MoD results had slightly more general concepts which may have a slight positive effect on the retrieval performance.
- **RUN5 (UKY_MAN)**: Each query is manually modified based on manual observations of the topic.

IV. Results

The scope of our experiments is focused on Task A (retrieval of scientific abstracts) alone although we submitted some baseline runs for the clinical trial dataset. Table III shows the average performances for infNDCG, R-prec, and P@5/10/30. The best infNDCG was achieved by UKY_CJT, which added a conjunctive query of the baseline query terms to the original disjunctive query. UKY_AGG returned the highest scores in all other measures except infNDCG. The worst score in the same category is from UKY_COM, in which the MoD terms are appended to the queries. This indicates that the topic expansion technique if not handled properly may result in weaker outcomes. Overall, the combination of query-time boosting and knowledge-based query expansion resulted in our best results.

V. Conclusion

This paper describes our IR system and its results based on runs submitted to the the TREC 2017 Precision Medicine track. We imbued both input topics and document collections with external knowledge (from UMLS and MeSH) to improve the recall and ranks of relevant documents. Overall, the topic expansion techniques using the UMLS concepts and MeSH terms improved the results (specifically for the AACR/ASCO collection). Our systems ranked 6th (in terms of infNDCG and R-prec) among 29 teams that submitted runs to the abstract retrieval task. In the future, we would like to experiment with well-known learning-to-rank and recent neural approaches to rerank the top few results to improve our performances.

References

[1] Gianni Amati, Cornelis Joost, and Van Rijsbergen. Probabilistic models for information retrieval based on divergence from randomness. 2003.

[2] Stéphane Clinchant and Eric Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for ir. In *Conference on the Theory of Information Retrieval*, pages 54–65. Springer, 2009.

[3] Stéphane Clinchant and Eric Gaussier. Information-based models for ad hoc ir. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 234–241. ACM, 2010.

[4] Y Song, Y He, Q Hu, and L He. Ecnu at 2015 cds track: Two re-ranking methods in medical information retrieval. In *Proceedings of the 2015 Text Retrieval ...*, jan 2015.

[5] R You, S Peng, S Zhu, and Y Zhou. FDUMedSearch at TREC 2015 Clinical Decision Support Track. *TREC*, jan 2015.