

UTD HLTRI at TREC 2018: Precision Medicine Track

Stuart J. Taylor
Human Language Technology
Research Institute
Department of Computer Science
University of Texas at Dallas
Richardson, TX
stuart@hlt.utdallas.edu

Travis R. Goodwin
Lister Hill National Center for
Biomedical Communications
U.S. National Library of Medicine
U.S. National Institutes of Health
Bethesda, MD
travis.goodwin@nih.gov

Sanda M. Harabagiu
Human Language Technology
Research Institute
Department of Computer Science
University of Texas at Dallas
Richardson, TX
sanda@hlt.utdallas.edu

Abstract

In this paper, we describe the system designed for the TREC 2018 Precision Medicine track by the University of Texas at Dallas (UTD) Human Language Technology Research Institute (HLTRI). Our system extends the system submitted for the 2017 track which incorporates an aspect-based retrieval paradigm wherein each of the four structured components of the topic is cast as a separate aspect, along with two “hidden” aspects encoding the need that retrieved documents be within the domain of precision medicine and that retrieved documents have a focus on treatment. For the 2018 system, in addition to the aspect-based retrieval, we incorporated learning-to-rank (L2R). Our experiments show that our L2R approach leads to improved quality of retrieved clinical trials, but degrades performance for scientific articles.

1 Introduction

The 2018 Text REtrieval Conference (TREC) Precision Medicine Track aims to address the important clinical challenge of providing useful precision-medicine related information to clinicians treating patients with cancer. Specifically, participants are invited to devise automatic and/or manual systems capable of (1) obtaining pertinent scientific articles from the medical literature describing the precise treatment of tumors exhibiting specific genetic mutations or alterations; and (2) identifying clinical studies in the National Library of Medicine (NLM)’s ClinicalTrials.gov database that investigate drugs targeting the patient’s malignancy and for which the patient might be eligible.

In this paper, we present an extension of our system designed for the TREC 2017 Precision Medicine Track. As the track specifically emphasizes that retrieved scientific articles and clinical trials should have a focus on treatment, we cast both tasks as a hybrid question answering (Q&A) and information retrieval (IR) problem. Formally, we consider a topic (as exemplified in Figure 1) to be asking an implicit question, *What is the best treatment for the patient described by the topic?* To answer this question, we adapt techniques for Q&A from knowledge bases by generating and using a knowledge graph encoding relationships between drugs,

Topic 4
DISEASE: melanoma GENE: BRAF (V600E), NRAS (Q61R) DEMOGRAPHIC: 67-year-old male
Topic 14
DISEASE: melanoma GENE: KIT amplification DEMOGRAPHIC: 66-year-old female
Topic 22
DISEASE: melanoma GENE: no tumor infiltrating lymphocytes DEMOGRAPHIC: 74-year-old female

Figure 1. Example topics evaluated in the 2018 TREC Precision Medicine Track.

genes, and mutations. This allows us to retrieve articles discussing both the topic and its inferred answers by incorporating an aspect-based retrieval strategy based on rank fusion and learning-to-rank.

The remainder of this paper is organized as follows: Section 2 describes the data evaluated in the 2018 TREC Precision Medicine track, Section 3 details our approach, Section 4 describes each of the runs we submitted, Section 5 reports initial results, and Section 6 summarizes the conclusions.

2 The Data

Two separate document collections were used for the Precision Medicine track: scientific abstracts and clinical trials.

2.1 Scientific Abstracts

The scientific abstracts considered in Task A originated from two sources: (1) MEDLINE abstracts, and (2) Conference Proceedings.

MEDLINE Articles A January 2017 snapshot of MEDLINE abstracts was used for the scientific abstracts. The task organizers provided both a rich XML and simple textual representation for all abstracts. In this work, we considered the XML representation which provided, in addition to the textual

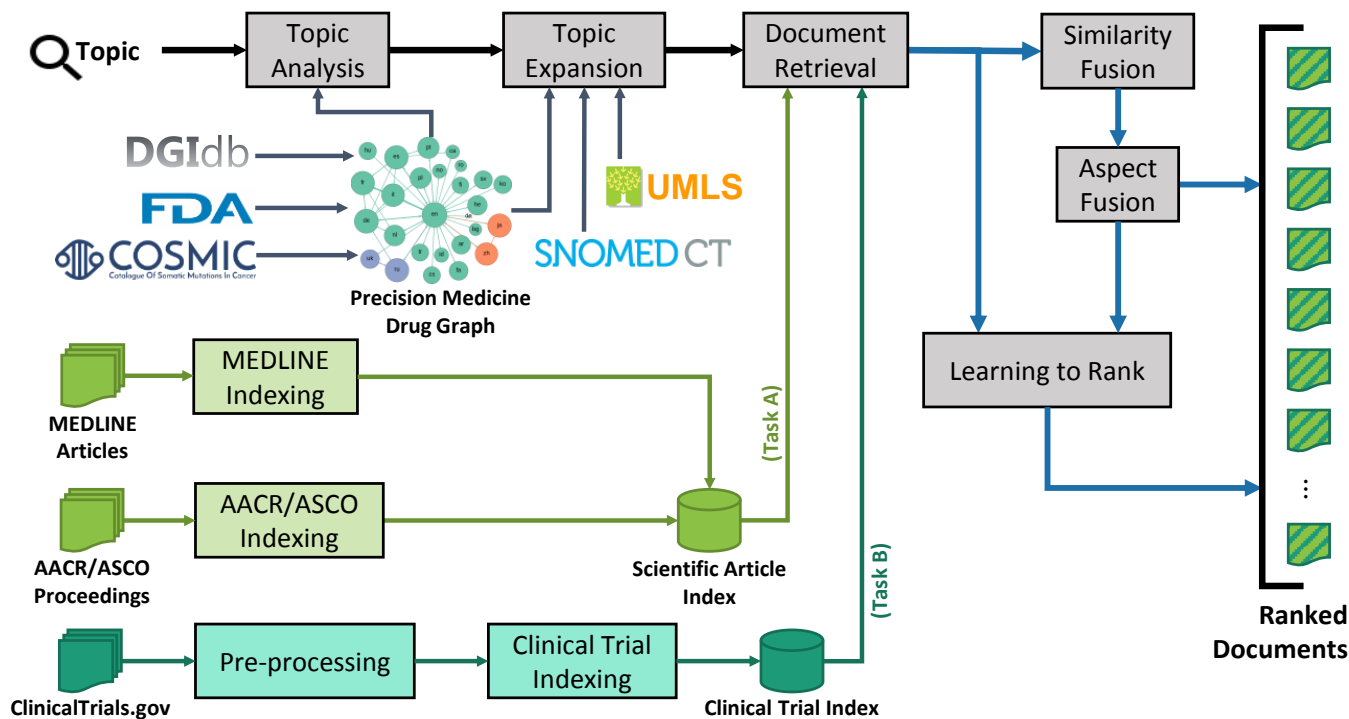


Figure 2. Architecture of Multi-task Retrieval System

content of the article, a variety of metadata including Medical Subject Headings (MeSH), keywords, and a controlled vocabulary of studied chemical compounds.

Conference Proceedings Because many of the topics evaluated in TREC PM were highly specific, the organizers also included abstracts of articles published in the proceedings of the annual meetings of the American Association for Cancer Research (AACR) and American Society of Clinical Oncology (ASCO). These articles were included with the intent to provide potentially relevant reports/articles describing precision medicine studies which are not included in MEDLINE. It should be noted that only a simple textual representation of articles presented at AACR/ASCO was available.

2.2 Registered Clinical Trials

For Task B, a snapshot of ClinicalTrials.gov obtained in April 2017 was considered. As with MEDLINE, both a rich XML and simple textual representation of each clinical trial was provided by the organizers. In this work, we considered the XML representation.

3 The Approach

Our system, illustrated in Figure 2, includes both “on-line” steps (i.e., steps applied for each topic) and “off-line” steps (steps applied before any topics are considered). In terms of on-line processing, the system has three mandatory steps:

1. *Topic Analysis*. The structured information of a given topic is analyzed to determine its main components or aspects;
2. *Topic Expansion*. The various aspects of the topic are expanded using external resources;
3. *Document Retrieval*. The aspects of the topic are represented as queries, and a set of ranked documents, i.e., scientific articles or clinical trials (depending on whether the system is configured for Task A or B), is retrieved;

as well as three “optional” steps designed to improve the quality of retrieved documents:

4. *Aspect Fusion (Optional)*. The ranked documents separately retrieved for each aspect of the topic are combined or *fused* together;
5. *Similarity Fusion (Optional)*. The ranked documents retrieved using multiple similarity measures (i.e., relevance models) are fused together.
6. *Learning to Rank (Optional)*. The ranked documents are re-ranked using a Random Forest classifier trained on the relevance judgments produced for the 2017 TREC-PM track.

In terms of off-line processing, the system relies on the availability of three data structures: (a) the Precision Medicine Drug Graph (PMDG), (b) an index of scientific articles, and (c) an index of clinical trials. In the remainder of this section

we describe the off-line steps used to create the above data structures, followed by the on-line steps of our system.

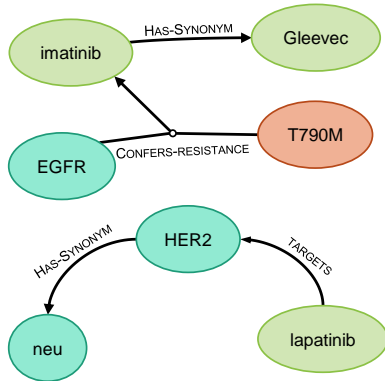


Figure 3. Examples of relations in the Precision Medicine Drug Graph (PMDG).

3.1 Building the Precision Medicine Drug Graph

Knowledge about the interactions between disease, genes, and drugs is available in a large variety of disconnected knowledge bases. Consequently, to unlock this knowledge, we create a unified knowledge graph which we refer to as the Precision Medicine Drug Graph (PMDG). The PMDG aggregates a subset of information from a variety of knowledge sources, including the Catalogue of Somatic Mutations in Cancer (COSMIC) [10], FDA Labels, as well as the 15 data sources incorporated within the Drug-Gene Interaction Database (DGIdb) [11].

Illustrated in Figure 3, the PMDG represents drugs, genes, and mutations as nodes. The PMDG encodes both binary relations such as $HER2 \xrightarrow{\text{HAS-SYNONYM}} neu$, as well as ternary relations such as $(EGFR, T790M) \xrightarrow{\text{CONFERS-RESISTANCE}} imatinib$. Specifically, the PMDG encodes four types of relations: (1) drug synonyms, (2) gene synonyms, (3) that a gene and locus confer resistance to a drug, and (4) that a drug can target (e.g. regulate) a specific gene and (possibly) a locus. In our system, the PMDG is used for both Topic Analysis and Topic Expansion.

3.2 Indexing the Data

We maintain two separate indices corresponding to the two tasks evaluated in the track.

3.2.1 Indexing for Task A

For task A, we maintain an tiered index including both MEDLINE articles and articles published in AACR/ASCO Proceedings. Because the MEDLINE articles are available in a rich XML format and the AACR/ASCO proceedings are only available as text, the indexing strategy varies for each dataset.

Indexing MEDLINE Articles The National Library of Medicine (NLM) provides an abundance of metadata about each article indexed in MEDLINE. In this work, we indexed a total of eight fields of metadata:

- *PubMed ID (PMID)*: the unique identifier assigned to each document in MEDLINE, used as the document ID for TREC PM submissions;
- *Journal Title*: the NLM version of the International Standards Organization (ISO) abbreviation of the title of the journal containing the article;
- *Publication Date*: the date the article was submitted to MEDLINE (specifically, the date when MEDLINE processing began for the article);
- *Article Title*: the (possibly-translated) title of the article;
- *Publication Type(s)*: the Medical Subject Headings (MeSH) [13] unique identifiers for all publication types associated with the article;
- *MeSH Term(s)*: the MeSH terms associated with the article;
- *Chemical Compound(s)*: the MeSH terms associated with all registered chemical compounds associated with the article; and
- *Abstract Text*: the full text of the abstract (note: structured abstracts are combined into a single passage).

Indexing AACR/ASCO Proceedings Because AACR/ASCO Proceedings were only available in a simple text format, we were only able to index a subset of the fields indexed for MEDLINE articles:

- *Article ID*: the filename of the abstract as provided by the organizers, used as the document ID for TREC PM submissions;
- *Journal Title*: the name of the conference;
- *Publication Date*: the year of the conference;
- *Article Title*: the title of the article; and
- *Abstract Text*: the unstructured text of the article;

3.2.2 Indexing for Task B

Task B relied on an index of clinical trials registered in ClinicalTrials.gov. While an abundance of metadata is available for each clinical trial, much of the metadata is represented as un-normalized free text, including date expressions, investigator names, descriptions of patient eligibility, etc. Producing the Clinical Trial index, consequently, entails two steps: (1) pre-processing each clinical trial, and (2) indexing each clinical trial.

Pre-processing Clinical Trials We observed three main inconsistencies in the metadata associated with each clinical trial on ClinicalTrials.gov:

1. *Processing Investigator Names*. Although the registry includes separate fields for the first, middle, and last name of each investigator, many investigators were provided with no first name, no middle name, and their

entire name as the last name. If an investigator had no provided first name, we relied on a series of rules to try to recover the first and last names of investigator from the provided names.

2. *Normalizing Temporal Expressions.* We found that minimum/maximum eligibility age was expressed in a variety of ways, e.g., by years, months, weeks, or even days. Moreover, the start date was provided in a variety of inconsistent formats. To account for this, we normalized temporal expressions using Natty¹.
3. *Recognizing Inclusion and Exclusion Criteria.* The structured data associated with each clinical trial includes a single unstructured field containing all eligibility criteria. To distinguish between inclusion and exclusion criteria, we used a simple rule-based strategy for distinguishing between sections of the eligibility criteria field that describe inclusion and exclusion criteria. Moreover, to account for the role of negation, we applied negation span detection using LingScope[1]. By detecting negation spans, we were able to parse negated inclusion criteria as exclusion criteria and negated exclusion criteria as inclusion criteria.

Indexing Clinical Trials When indexing clinical trials, we considered a tiered index encoding multiple fields of metadata:

- *NCT ID:* The unique identifier associated with each clinical trial in ClinicalTrials.gov, used as the document ID for TREC PM submissions.
- *Brief Title:* A brief summary of the clinical trial;
- *Official Title:* The official detailed title of the clinical trial;
- *Summary:* A summary of the role and purpose of the clinical trial;
- *Description:* A detailed description of the clinical trial including the goals, study design, and experimental setup;
- *Studied Condition(s):* The medical conditions studied in the trial;
- *Condition MeSH Term(s):* MeSH terms associated with the conditions studied in the trial (if any);
- *Studied Intervention(s):* The medical interventions (if any) evaluated in the trial;
- *Studied Intervention Type(s):* The type of medical interventions evaluated in the trial (e.g., genetic, drug, etc.);
- *Intervention MeSH Term(s):* MeSH terms associated with the interventions evaluated in the trial (if any);
- *Minimum Age:* The minimum age of eligible participants (if provided);
- *Maximum Age:* The maximum age of eligible participants (if provided);

- *Eligible Gender(s):* The eligible gender(s) of participants;
- *Inclusion Criteria:* Unstructured textual representation of all inclusion criteria parsed from the Eligibility Criteria of the trial; and
- *Exclusion Criteria:* Unstructured textual representation of all exclusion criteria parsed from the Eligibility Criteria of the trial.

3.3 Topic Analysis

Because the topics evaluated in the 2018 TREC PM task correspond to complex semi-structured medical cases, determining the relevance of a document (i.e. scientific article or clinical trial) to a topic requires accounting for many different *aspects*. In this, we consider a total of six aspects of the topic: four corresponding to the three semi-structured fields of the topic, and two additional implied or “hidden” aspects. Unlike the previous year, the “gene” field of the topic included both specific genetic variants (e.g., “BRAF (V600E)”) and general descriptions of tumors (e.g., “no tumor infiltrating lymphocytes”). Consequently, when representing the topic we distinguish between specific genetic variants, and other types of tumor descriptions. The four explicit aspects are represented as follows:

1. **DISEASE Aspect:** The disease (a type of cancer) is represented as a 2-tuple (pair) containing (1) the literal/surface form of the disease included in the topic, and (2) the set of zero or more concept unique identifiers (CUIs) in the Unified Medical Language System (UMLS) [3] corresponding to the disease. To identify the concepts in UMLS corresponding to each disease, we used a simple pattern-matching approach relying on the surface form of the disease and any matching *atoms* in UMLS;
2. **GENETIC Aspect:** Because a topic may indicate more than one genetic variant, we represent each genetic variant as a 3-tuple consisting of (1) the name of the gene, (2) the *type* of mutation, and (3) the locus of the mutation (if provided). The mutation type was determined using a small number of lexical patterns; the set of mutation types considered by our approach are: (1) AMPLIFICATION, (2) DUPLICATION, (3) TRANSLOCATION, (4) DELETION, (5) POINT-MUTATION, (6) INACTIVATION, and (7) UNSPECIFIED;
3. **DEMOGRAPHIC Aspect:** The demographic aspect of the topic was represented as a 2-tuple corresponding to two facets of demographic information, the patients age, and the patients gender; and
4. **OTHER Aspect:** The “other” aspect contains any genetic or tumor description from the “gene” field of the topic that is not a specific genetic variant.

¹<http://natty.joestelmach.com/>

To account for the requirement that retrieved documents have a focus on targeted treatment and precision medicine, we considered to additional “hidden” aspects of each topic.

5. **PRECISION MEDICINE Aspect:** This aspect indicates that retrieved documents must fall within the domain of precision medicine to be considered relevant; and
6. **TREATMENT Aspect:** This aspect indicates that retrieved documents must have a focus on treatment, and cannot be simple observational studies. To represent this aspect, we relied on the information encoded in the PMDG. Specifically, the treatment aspect is represented as a 2-tuple including (1) drugs targeting any gene/mutation in the topic, (2) any drugs that any gene/mutation in the topic confers resistance against.

Together, these six aspects are used to represent the diverse information needs expressed by each topic.

3.4 Topic Expansion

To account for the complexity of medical language in scientific articles and clinical trials, we incorporate query expansion techniques to expand (1) the medical problems within the disease and other aspects of the topic, (2) each genetic variant, and (3) any drugs in the treatment aspect of the topic.

Expanding the Disease and Other Aspects. To account for the role of synonymy in scientific articles and clinical trials, we incorporate two forms of query expansion: (a) we identify synonyms for each medical problem using UMLS and (2) we discover hyponyms using the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[7].

Expanding the Genetic Variations. Because genetic information can be described in a variety of ways, we expand each gene itself so as to include gene synonyms using the Precision Medicine Drug Graph (PMDG). We expand the mutation type to include basic synonyms. Finally, we expand the locus of each mutation by (1) expanding the amino acids to their three-letter abbreviations and (2) adding the protein identifier “p.” commonly used to indicate mutation loci in the literature.

Expanding Drugs. We expand drugs to include both brand names as well as generic names by using the PMDG. Specifically, we follow all SYNONYM relations originating from the nodes corresponding to any drug in the treatment aspect, and use those nodes as synonyms for the drug.

3.5 Document Retrieval

The role of the document retrieval step is to produce a ranked list of *documents* – scientific articles or clinical trials – relevant to a given topic. We considered two strategies for document retrieval:

- **Aspect Retrieval:** to prevent any single aspect from having too large of an impact on the score of a document, in the aspect retrieval strategy, we cast each aspect of the topic as a separate, independent query and obtain a separate ranked list of documents for each aspect; and
- **Joint Retrieval:** in the joint retrieval strategy, we cast each aspect of the topic as a clause in a single disjunctive Boolean query, obtaining a single ranked list of documents for the entire topic.

Both strategies rely on Apache Lucene² (version 6.6.0). It should be noted that the way in which each aspect was represented as a query (or clause) depends on whether the system is configured for Task A (scientific articles) or Task B (clinical trials).

prevention,	prophylaxis,	prognosis,
outcome,	survival,	treatment,
therapy,	personalized	

Table 1. Lexicon of words indicating treatment.

AMPLIFICATION	inhibitor antagonist
DUPLICATION	suppressor antisense
TRANSLOCATION	blocker
DELETION	agonist activator inducer
INACTIVATION	potentiator stimulator

Table 2. Lexicon of drug roles targeting each type of genetic mutation.

3.5.1 Retrieving Scientific Articles (Task A)

When retrieving scientific articles for a given topic, we encoded the DISEASE aspect as an additive disjunctive Boolean query with a clause representing the surface form of the disease as indicated in the topic as well as additional clauses representing each expansion. By contrast, the GENETIC aspect was represented as special type of disjunctive Boolean query in which the score of a document was determined as the *maximum* score obtained for any clause in the query (rather than the total). When encoding the GENETIC aspect, we considered up to three clauses corresponding to (1) the gene and its expansions, (2) the type of mutation and its expansions, and (3) the locus (if provided) and its expansions (if any). To ensure that scientific articles satisfy the TREATMENT aspect of the topic, we produced an additive disjunctive Boolean query in containing clauses encoding (a) the lexicon

²<https://lucene.apache.org/core/>

indicated in Table 1, (b) the patterns indicated in Table 2, (c) any drugs targeting the specific mutation and type of tumor indicated by the topic, and (d) any drugs known to be ineffective for the specific mutation. Finally, we address the PRECISION MEDICINE aspect by reducing the score of articles that focus on science rather than medicine and increasing the score of articles that appear that relate to clinical trials and studies involving humans. Specifically, we penalize the score of articles which (1) contain *cell*, *biochem*, *chem*, *molecular*, *cytogenetics*, *pathology*, or *★pathology* in the title of the journal, (2) mention the terms *cell*, *cell line*, or *cell cycle* in the abstract or title of the article, and (3) have a MeSH term of “Cell Line, Tumor”. To favor articles reporting the results of clinical studies, we increase the score of articles that have a MeSH term of “Human”, or include any of the patterns indicated in Table 3. We observed that, in general, scientific articles do not refer to the demographics or comorbidities of patients (an exception to this would be in the case of tumors occurring only in men or women- such as prostate or cervical cancer). Consequently, when searching scientific articles, we chose to ignore the DEMOGRAPHIC and OTHER aspects of the topic.

phase 1,	phase 2,	phase 3
phase I,	phase II,	phase III
trial,	randomized,	patient

Table 3. Lexicon of words indicating medical trials.

3.5.2 Retrieving Clinical Trials (Task B)

We observed that clinical trials often focused on general diseases rather than specific types of cancer. For example, when manually browsing ClinicalTrials.gov, we found that many clinical trials study any type of “solid tumor” rather than the specific type(s) provided in the topics evaluated in the Precision Medicine track. Consequently, when retrieving clinical trials for a given topic, we encoded the DISEASE aspect as a disjunctive Boolean query with clauses representing (1) the surface form of the disease as indicated in the topic, (2) each expansion of the disease, as well as the phrases (3) “solid tumor” and (4) “solid neoplasm”. When searching clinical trials, the GENETIC and TREATMENT aspects were represented in the same way as when searching scientific articles. The DEMOGRAPHIC aspect was represented as a conjunctive Boolean query with three clauses ensuring (1) the minimum eligible age for the trial was \leq the age of the patient in the topic, (2) the maximum eligible age for the trial was \geq the age of the patient in the topic, and (3) the gender of the patient in the topic matched the eligible genders of the trial. To encode the OTHER aspect of the topic, we constructed a disjunctive Boolean query containing each disease and its expansions, and penalized the score of clinical

trials which matched this query in their *Exclusion Criteria* field, and slightly increased the score of clinical trials which matched this query in their *Inclusion Criteria* field. Finally, the PRECISION MEDICINE aspect was encoded as a disjunctive Boolean query favoring clinical trials which have at least one intervention of type “DRUG” or “GENETIC”, and/or which include the term “Phase” in their *Brief Title* field.

3.6 Aspect Fusion

When documents are retrieved using the *Aspect Retrieval* strategy, it is necessary to combine the ranked list of documents obtained for each aspect to produce a single ranked list of documents relevant to the topic. This is accomplished using a technique known as rank fusion or answer fusion. Rank fusion is a process for combining the ranked list resulting from multiple searches to produce a single ranked list of results. While a number of different methods for rank fusion have been published, they predominantly rely on combining the different relevance scores associated with the same document in different ranked lists. In our system, the scores produced when searching for each aspect can vary by several orders of magnitude, making score-based rank fusion techniques difficult to apply. Thus, we rely on a method known as *Reciprocal Rank Fusion*[6] (RRF). Rather than combining the different scores associated with a single document, RRF combines the reciprocal rank of the document in each ranked list. Formally, given a set of D documents to be ranked, and a set of rankings R , we determine the new score of each document $d \in D$ as:

$$\text{RRF-Score}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} \quad (1)$$

where $r(d)$ is the rank of document d in ranking r , and k is a parameter intended to reduce the impact of low ranks on the score (in our experiments, we used $k = 60$ as recommended by the original authors). Thus, the role of the aspect fusion step is to combine the ranked list of documents obtained for each aspect of the topic using Equation 1 to produce a single ranked list of documents for the topic.

3.7 Similarity Fusion

It has been widely shown that for many machine learning evaluations, the top performing systems often combine a large variety of models. We were interested in learning whether this behavior was true for information retrieval problems as well. Thus, the role of the similarity fusion step entailed two steps: (1) the document retrieval process was repeated using a number of different similarity measures (or relevance models), and (2) the resultant ranked lists of documents were combined to produce a single ranked list. As with aspect fusion, we relied on reciprocal rank fusion (RRF) to combine the ranked list of documents retrieved for each similarity measure. In our experiments, we considered

Feature Description	Feature Description
BM25 <i>statistics</i> from each disease (without expansions) of t_i for f_i in d_i	\vdots
LMD <i>statistics</i> from each disease (without expansions) of t_i for f_i in d_i	F2Exp <i>statistics</i> from each gene (without expansions) of t_i for f_i in d_i
F2Exp <i>statistics</i> from each disease (without expansions) of t_i for f_i in d_i	DFI <i>statistics</i> from each gene (without expansions) of t_i for f_i in d_i
DFI <i>statistics</i> from each disease (without expansions) of t_i for f_i in d_i	TF-IDF <i>statistics</i> from each gene (without expansions) of t_i for f_i in d_i
TF-IDF <i>statistics</i> from each disease (without expansions) of t_i for f_i in d_i	
BM25 <i>statistics</i> from each disease (with expansions) of t_i for f_i in d_i	BM25 <i>statistics</i> from each gene (with expansions) of t_i for f_i in d_i
LMD <i>statistics</i> from each disease (with expansions) of t_i for f_i in d_i	LMD <i>statistics</i> from each gene (with expansions) of t_i for f_i in d_i
F2Exp <i>statistics</i> from each disease (with expansions) of t_i for f_i in d_i	F2Exp <i>statistics</i> from each gene (with expansions) of t_i for f_i in d_i
DFI <i>statistics</i> from each disease (with expansions) of t_i for f_i in d_i	DFI <i>statistics</i> from each gene (with expansions) of t_i for f_i in d_i
TF-IDF <i>statistics</i> from each disease (with expansions) of t_i for f_i in d_i	TF-IDF <i>statistics</i> from each gene (with expansions) of t_i for f_i in d_i
BM25 <i>statistics</i> from each gene (without expansions) of t_i for f_i in d_i	BM25 <i>statistics</i> from each medical problem of t_i for f_i in d_i
LMD <i>statistics</i> from each gene (without expansions) of t_i for f_i in d_i	LMD <i>statistics</i> from each medical problem of t_i for f_i in d_i
\vdots	F2Exp <i>statistics</i> from each medical problem of t_i for f_i in d_i
	DFI <i>statistics</i> from each medical problem of t_i for f_i in d_i
	TF-IDF <i>statistics</i> from each medical problem of t_i for f_i in d_i

Table 4. Features extracted for field f_i in document d_i retrieved for topic t_i .

eight similarity measures: Base Model 25[14] (BM25), TF-IDF, Divergence from Randomness (DFR)[2], Information-based Similarity[5], Dirichlet-smoothed Language Model Similarity[16], Jelinek-Mercer-smoothed Language Model Similarity[16], Axiomatic Similarity[9], and Divergence from Independence[12].

3.8 Learning to Rank

We leveraged the relevance judgments produced for TREC-PM 2017 by training a Random Forest to re-rank the documents retrieved for both tasks. Learning-to-rank (L2R) is used to learn a precision-medicine-specific relevance model based on the relevance judgments produced for TREC-PM 2017. Specifically, we (1) we extract features characterizing the relevance between that an article/trial and the topic and represent each article/trial retrieved for each topic as a feature vector, and (2) train a relevance model to infer relevance of an article/trial to the topic based solely on the extracted features and relevance judgments produced during TREC-PM 2017.

Task	Field
Task A	Abstract Text
	Article Title
	MeSH Terms
Task B	Official Title
	Description
	MeSH Terms
	Keywords
	Inclusion Criteria
	Exclusion Criteria

Table 5. Fields used for each task.

3.8.1 Feature Extraction

Determining if an article or clinical trial is relevant to a topic requires considering information about the topic’s disease, genetic variants, and demographics. For this reason, ranking articles and clinical trials requires a rich set of features. When extracting features we consider (1) the different fields of the article or trial (e.g. title, main body of text, and MeSH terms), (2) the different aspects of the topic, (3) the expansions of each aspect (e.g. genetic variants), and (4) multiple relevance models.

As shown in Table 4, each feature measures the relevance between an aspect of the topic and a field of the document. We considered five measures of relevance: (1) Best Match 25[14] (BM25), (2) Dirichlet-Smoothed language model probability [16] (LMD), (3) Axiomatic relevance [9] (F2Exp), (4) Divergence from Independence [12] (DFI), and (5) Term Frequency–Inverse Document Frequency [15] (TF-IDF). As Task A and Task B use different document collections, we use different fields for each task which are shown in Table 5. We additionally extract features both with and without expanding each aspect. To account for the variance in the number of genes and diseases, we measure five *statistics* capturing the similarity between each disease and genetic variant for each topic, namely, the mean, minimum, maximum, variance, and sum.

3.8.2 Training the Relevance Model

We trained a Random Forest[4] to measure the relevance of each retrieved article/trial to a given topic by (1) extracting features from all articles/trials retrieved for each topic in TREC-PM 2017, and (2) maximizing the mean average precision (MAP) of the re-ranking produced by the model using the relevance judgments produced for TREC-PM 2017. To

System (Run)	Task A			Task B		
	infNDCG	P@10	R-Prec	infNDCG	P@10	R-Prec
Run 1: UTDHLTRI_NL	0.4797	0.6160	0.2870	0.4794	0.5380	0.3675
Run 2: UTDHLTRI_SF	0.3668	0.5360	0.1996	0.4139	0.4640	0.3102
Run 3: UTDHLTRI_SS	0.3554	0.4920	0.1936	0.4554	0.5360	0.3920
Run 4: UTDHLTRI_RF	0.3827	0.5300	0.2088	0.4587	0.4960	0.3444
Run 5: UTDHLTRI_RA	0.4113	0.5440	0.2291	0.4629	0.5380	0.3785

Table 6. Average performance across each topic for TREC PM.

Run	Simple	Aspect	Similarity	Joint	L2R
UTDHLTRI_NL		✓	✓		
UTDHLTRI_SF			✓		✓
UTDHLTRI_SS	✓				✓
UTDHLTRI_RF		✓	✓		✓
UTDHLTRI_RA	✓	✓	✓	✓	✓

Table 7. System configuration for each run, where “Simple” indicates the run used a simple rule-based search strategy, “Aspect” indicates that the run used Aspect-fusion, “Similarity” indicates that the run used Similarity-fusion, “Joint” indicates the Joint Retrieval strategy was used, and “L2R” indicates that the run incorporated learning-to-rank to re-rank retrieved documents.

do this, we used RankLib[8], a Java library with implementations of many learning-to-rank algorithms.³

4 Overview of Runs

We submitted five runs relying on (a) different configurations of our system as well as (b) a simple rule-based strategy considering only the disease and genes of the topic. Table 7 presents the configurations of the system for each run; in the case of learning-to-rank (L2R) runs, the table indicates the configuration of the system when retrieving the initial set of article/trial which are re-ranked using L2R. Each of these runs is detailed below:

- **Run 1: UTDHLTRI_NL** This run relied on the Aspect Retrieval strategy and incorporated both Aspect Fusion and Similarity Fusion. This run did not use our L2R model;
- **Run 2: UTDHLTRI_SF** This run relied on Similarity Fusion to perform the initial retrieval of documents, and re-ranked the returned documents using our L2R model;
- **Run 3: UTDHLTRI_SS** This run used a very simple rule-based search strategy considering only the disease and genes of the topic for the initial retrieval, and re-ranked the returned documents using our L2R model;

- **Run 4: UTDHLTRI_RF** This run relied on Aspect Fusion and Similarity Fusion to perform the initial document retrieval, and re-ranked the returned documents using our L2R model (this is essentially Run 1 with L2R incorporated); and
- **Run 5: UTDHLTRI_RA** This run combined the results of the Simple, Joint, and Aspect-based strategies into a single list of results, and re-ranked the entire list using our L2R model.

5 Results

Table 6 presents the average performance of each of our runs for both tasks. For both tasks, we report the inferred NDCG (infNDCG), Precision at 10 (P@10), and R-Precision (R-Prec). In general, the best performance was obtained by Run 1 which incorporated Aspect Retrieval and Aspect Fusion. For Task B, Runs 3 and 5 obtained higher R-Prec than Run 1, but lower infNDCG and P@10, suggesting that, for some topics, the simple retrieval strategy was able to retrieve relevant documents that were not retrieved by the aspect nor joint retrieval strategies.

6 Conclusion

In this paper, we described our system designed for the TREC 2018 Precision Medicine track. For both tasks, we submitted five runs corresponding to alternative configurations of our system. Our system incorporates an aspect-based retrieval paradigm wherein each of the four structured components is considered as an aspect, along with two “hidden” aspects encoding the need for retrieved documents to be within the domain of precision medicine and to have a focus on treatment. To this end, we construct knowledge graph encoding the relationship between genes, mutations, and drugs. Additionally, for this year’s evaluation, we added a learning to rank component. Experimental results suggest that learning-to-rank provided no substantial improvement in performance.

References

- [1] Shashank Agarwal and Hong Yu. 2010. Biomedical negation scope detection with conditional random fields. *Journal of the American Medical Informatics Association: JAMIA* 17, 6 (2010), 696.

³ Random Forests were selected based on training set performance.

- [2] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 357–389.
- [3] Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, suppl 1 (2004), D267–D270.
- [4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [5] Stéphane Clinchant and Eric Gaussier. 2010. Information-based models for ad hoc IR. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 234–241.
- [6] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. 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, 758–759.
- [7] Roger A Côté, College of American Pathologists, et al. 1977. *Systematized nomenclature of medicine*. College of American Pathologists.
- [8] V. Dang. 2009. The Lemur Project-Wiki-RankLib. [Online] <http://sourceforge.net/p/lemur/wiki/RankLib>. (2009). Accessed: 2018-08-14.
- [9] Hui Fang and ChengXiang Zhai. 2005. An exploration of axiomatic approaches to information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 480–487.
- [10] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al. 2010. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* 39, suppl_1 (2010), D945–D950.
- [11] Malachi Griffith, Obi L Griffith, Adam C Coffman, James V Weible, Josh F McMichael, Nicholas C Spies, James Koval, Indrani Das, Matthew B Callaway, James M Eldred, et al. 2013. DGIdb: mining the druggable genome. *Nature methods* 10, 12 (2013), 1209–1210.
- [12] İlker Kocabaş, Bekir Taner Dinçer, and Bahar Karaođlan. 2014. A nonparametric term weighting method for information retrieval based on measuring the divergence from independence. *Information retrieval* 17, 2 (2014), 153–176.
- [13] Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.
- [14] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *NIST Special Publication (SP)* 500-225 (1995), 109–126.
- [15] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press.
- [16] Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* 22, 2 (2004), 179–214.