UD_GU_BioTM at TREC 2017: Precision Medicine Track

A. S. M. Ashique Mahmood[1], Gang Li[1], Shruti Rao[2], Peter McGarvey[2, 3]
Cathy Wu[1, 3, 4], Subha Madhavan[2, 5], K. Vijay-Shanker[1]

1 Department of Computer and Information Science, University of Delaware, Newark, DE, USA.
2 Innovation Center for Biomedical Informatics, Georgetown University, Washington D.C, USA;
3 Protein Information Resource, Georgetown University Medical Center, Washington D.C, USA;
4 Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA;
5 Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington D.C, USA;

{ashique, ligang, wuc, vijay}@udel.edu
{sr879, pbm9, sm696}@georgetown.edu

Abstract— This paper describes the system developed for the TREC 2017 PM track. We employed a two-part system to generate the ranked list of clinical trials and scientific abstracts. The first part pertains to query expansion and document retrieval from document index. The second part pertains to generating the final ranked list by implementing a heuristic scoring method. The scoring for clinical trials involved grouping trials based on different trial fields and extraction of features based on occurrences of gene/disease and other terms in the trial. The scoring for scientific abstracts involved applying a NLP system to extract relations from text, as well as extraction of additional information relevant to precision medicine.

Keywords—TREC 2017, precision medicine, NLP

1. INTRODUCTION

Precision medicine (PM) focuses on finding personalized treatments based on patients' genetic profiles and medical history. The vast volume of ever-growing scientific literature for precision medicine makes it challenging for oncologists and clinicians to find the most appropriate treatment. The ability to quickly locate relevant information for a current patient using information retrieval (IR) has the potential to be an important tool for helping clinicians find the most up-to-date evidence-based treatment for their patients. The Precision Medicine Track of Text REtrieval Conference (TREC) 2017 aims to encourage data-driven approaches to identify the best treatment for a patient, by finding the clinical trials that best matches the patient condition, as well as finding evidence-based literature that suggests effective treatment.

```
<topic number="1">
    <disease>Liposarcoma</disease>
    <gene>CDK4 Amplification</gene>
    <demographic>38-year-old male</demographic>
    <other>GERD</other>
</topic>
```

Figure 1: A sample topic for the TREC 2017 Precision Medicine track. Each topic contains the disease name, the variant gene, patient demographic and optionally other conditions patient might have.

TREC 2017 PM track provided 30 topics, where each topic is a synthetic patient profile. Figure 1 shows one sample topic. There were two document collections for this track: clinical trials and scientific abstracts. The goal of retrieving clinical trials is to find trials for which the given patient would be eligible to enroll. The goal of retrieving abstracts is to identify documents that can suggest treatment for the given patient. Participants had to retrieve and rank documents from each collection separately for each of the 30 topics.

2. METHODS

In our participation of this track, we developed a two-part approach: (1) Indexing and retrieval of documents and (2) Ranking. For the first part, we indexed the given document sets. Then for each topic, we expanded the query by expanding diseases and genes with synonyms from external resources. In the second part, after the expanded query returned documents from index, we employed a scoring mechanism to rank the documents. The scoring mechanism was different for clinical trials and abstracts. We implemented the scoring based on our observations of example documents and suggestions from domain experts. The scoring for clinical trials used information such as study type and phase of the trial and occurrences of gene/disease in the trial. The scoring of scientific articles was mostly based on a NLP system. We extracted relations between genomic anomalies and outcome of cancer therapeutics from text, which strongly indicates an evidence-based treatment option for patients. Along with these relations, we extracted additional information that were combined together in a weighted fashion to rank the abstracts. Figure 2 presents a schematic diagram of our approach for this track.

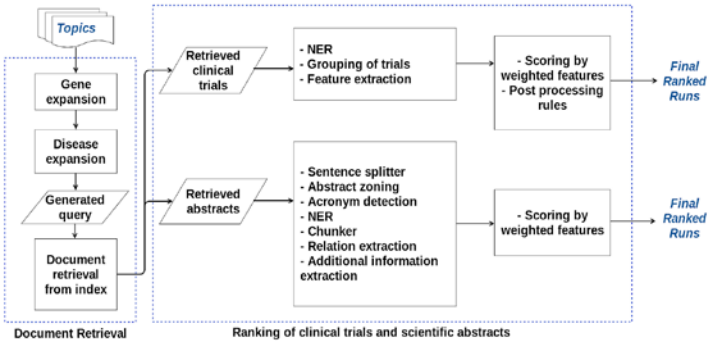

Figure 2: Flowchart of the system developed for TREC 2017 PM track.

2.1 Document Retrieval

A. Index

We used Lucene to index all the clinical trials and scientific abstracts. The standard analyzer, default scoring and empty set of stop words were used, as we only use Lucene to retrieve potential relevant trials and abstracts. We will score the trials and abstracts later using richer information extracted from them.

For clinical trials, the indexed fields are NCT_ID, brief_title, official_title, brief_summary, detailed_description, eligibility_criteria, gender, min_age, max_age, conditions, mesh_term, arm_description, primary_outcome, secondary_outcome, inclusion_text, exclusion_text. The fields are corresponding to the XML element tags in the XML files for the trials, except that inclusion_text and exclusion_text are extracted from the field eligibility_criteria. For scientific abstracts, the indexed fields are document ID (PMID), title and abstract texts.

B. Query Expansion

Gene Expansion

We use **Entrez** database to expand the gene names in the topic. Entrez database provides a list of synonyms for gene names. For each gene in the given topics, we retrieved the synonyms of gene names and included them as expansion in the query.

Disease Expansion

We use Disease Ontology [1] and MeSH terms to expand the disease names in the topic. **Disease Ontology** maintains a hierarchy of human disease terms. For each disease name in the topic, we first find the corresponding disease term in Disease Ontology, and obtain its grandparent, parent, child and grandchild terms. The names and synonyms of all the terms are collected to expand the disease name in the topic. We find the **MeSH** term for the disease name, and retrieve its name and synonyms as expansion of the disease name in the topic.

2.2 Ranking of Documents

Once we retrieved top documents from the index for each topic, we applied a scoring mechanism to score each document. The final submitted ranked list was based on this scoring. The scoring mechanism was different for the two tasks. The following sections describe the ranking methods for clinical trials and scientific abstracts, respectively.

A. Ranking of Clinical Trials

Grouping of Trials

To rank clinical trials for each topic, we first divided all the clinical trials into groups based on 5 fields in the trial: Study Type, Time Perspective, Intervention, Phase and Primary Purpose. We selected these 5 fields to group the trials as we believe that they are strongly associated with whether a trial focuses on treatment, prevention, or prognosis of cancer, one of the main judgement criteria in the relevance guideline. For example, if the Study Type is Interventional, Intervention uses Drug, and Primary Purpose is Treatment, then we know that the trial is about the treatment of cancer compared to a trial with Study Type of Observational.

Scoring of Trials

After we grouped the trials, we assigned scores to trials in each group. We searched for the expanded disease/gene names and other useful terms in the trial, and gave different scores to the mentions found according to the section type. The intuition is that some section types reflect more about what the trial is focused on. For example, if the disease name is found in the Title section, it is very likely that the trial is about the disease. On the contrast, if the disease name is only found in detailed description, then it is less likely to be the focus of the trial.

Beside the gene/disease names, we also searched for SNP mutation name in the topic and other useful terms for the gene name. For example, we used a list of general terms which may indicate mutation, amplification, expression, fusion, negativity of a gene and searched them in the context of gene name found in the trial. When such a term is found, it may be used to match up with the gene information given in the topic. For example, if the topic mentions gene amplification

and an amplification term is found around a gene name in the trial, we will rank this trial higher than those without an amplification term.

For each trial, we compute a combined score by first considering disease and gene scores, and then scores for other terms.

Post-processing Rules

Finally, we come up with a set of post-processing rules to rank down or discard some clinical trials. R2-8 discards a trial if it matches the rule.

R1: If there is a conflict between the gene information in the topic and terms found around the gene, rank down the trial. For example, if a topic does not contain any amplification term, but an amplification term is found around the gene in the trial, then there is a conflict.

R2: Gene/disease appears in exclusion criteria but not in inclusion criteria or title

R3: Terms in "other" field in the topic appears in exclusion criteria

R4: Age not matched

R5: Gender not matched

R6: Gene not found or only appears in detailed description

R7: Disease not found or only appears in detailed description

We submitted 5 runs for clinical trials retrieval. The detailed configuration of each run is in Table 1.

Run #	Description
1	Trials in groups with highest confidence, e.g., the study type is interventional and drug is used for intervention. R2 only checks inclusion criteria but not title.
2	Trials in Run 1 and trials in groups with lower confidence, e.g., the study type is observational and time perspective is prospective. R2 only checks inclusion criteria but not title.
3	Trials in all groups, R2 only checks inclusion criteria but not title
4	Trials in groups with highest confidence, e.g., the study type is interventional and drug is used for intervention. R2 checks both inclusion criteria and title
5	Trials in Run 4 and trials in groups with lower confidence, e.g., the study type is observational and time perspective is prospective. R2 checks both inclusion criteria and title

Table 1: Five Runs for Clinical Trials

B. Ranking of Scientific Abstracts

We used a combination of information to score each document. We looked into where in the abstracts the information mentioned in the topics appeared, as well as extracted information from text that solidifies the relevance of the abstracts to precision medicine. We employed NLP techniques to extract specific relations between genomic anomalies and response/outcome of cancer treatments. An abstract is deemed relevant for a topic if such relations are found in the abstract and the genomic anomaly matches the patient's variant in the given topic.

For each retrieved document, we applied the following procedure. The abstracts were first split into individual sentences and sectioned into five rhetorical zones (Title, Introduction, Method, Result, Conclusion) adapting the approach in [2]. We used Pubtator [3] NER annotations to tag disease, drug and gene entities. We applied an in-house developed mutation detector [2] for variant detection. It not only detects specific mentions (such as Tyr113His) but also generic mentions such as "BRCA1 mutation", "EGFR variant" etc. We used an acronym detector [4] to detect abbreviations, and accordingly extended entity recognition based on abbreviations. We then applied BioNex [5], a chunker to identify the shallow chunks from text, namely base noun phrases (NP) and verb groups. We also obtained larger NPs by taking into account prepositional phrase and relative clause attachments. We detected entities that represent the outcome of a treatment by looking for NPs headed by words/phrases that indicate response or outcome, such as "survival", "prognosis", "outcome", "response", and "efficacy" etc. Some examples of such occurrences are "progression free survival", "PFS", "overall survival", 'OR", "objective response rate" etc. Once the entities are tagged and phrases are detected, we employed the relation extraction module.

The relation extraction identifies associations between genomic anomalies and response to certain cancer treatments. We adopted the approaches mentioned in [6]. For instance, we would identify the following sentence in Example-1 as having an association between KRAS variants and treatment outcome. Additionally, from the context of the abstract, we identify the disease in concern, which is "ampullary adenocarcinoma" in this case.

<u>Example-1:</u> However, there was a significant correlation between KRAS mutation and worse RFS (HR = 2.74 , 95% CI : 1.52-4.92 , P = 0.0008). (PMID:27517148)

Additionally, we identify sentences that mentions genetic variants being involved in therapies. The following Example-2 is one such example. Again, from the context, we identify the disease (Breast cancer in this case).

Example-2: In addition, we detected a HER2 S855I mutation in two patients who had persistent benefits from anti-HER2 therapy. (PMID:28229982)

We extracted another type of sentences that suggest a potential therapy for patients with a specific genomic profile. Example-3 is an example of such sentences. This type of sentences were detected using simple patterns and looking for target phrases.

Example-3: The KRAS (G12D) mutation identifies a subset of AA patients with poor prognoses and may be used to identify patients at risk of early recurrence and poorer survival who may benefit from adjuvant therapy. (PMID:25616942)

Once these relations are identified, we confirm that the genomic anomaly and the disease matches the topic description. We extracted additional information from the abstracts that were used for ranking. We determined whether the article talks about a treatment or prognosis for the given disease by looking for certain phrases. We checked whether the given gene and disease (and their synonyms and variations) appear in prime locations such as title, result or conclusion sections. We penalized the abstract if it contains other genes and/or diseases in title or conclusions. All these extracted information were used as features. We applied a simple formula to calculate the score for each abstract by summing a weighted list of the features. We submitted 5 different runs for scientific abstracts by slightly varying the weights of the features. A brief description of the 5 runs are shown in Table 2.

Run #	***Description***
1	Articles with target genes and diseases more frequent in title or conclusions, received higher ranking.
2	Articles with relationship between variant and drug responses received higher ranking.
3	Articles with sentences suggesting potential therapy (Example-3) received higher ranking.
4	Articles with exact or more specific mentions of disease received higher ranking.
5	Features from all previous four runs are evenly combined to rank articles.

Table 2: Brief description of the 5 runs for scientific abstracts.

3. RESULTS

Table 3 and 4 lists the values for the evaluation metrics for all runs for clinical trials and scientific abstracts, respectively. Top scores are marked by bold font. The last rows of Table 3 and 4 represent the average of the median values for all topics over 133 and 125 runs, respectively.

Run ID	**P @ 5**	**P @ 10**	**P @ 15**
UD_GU_CT_1	0.5172	0.4241	0.3701
UD_GU_CT_2	0.5379	0.4414	0.3816
UD_GU_CT_3	**0.5448**	**0.4448**	**0.3885**
UD_GU_CT_4	0.5214	0.4214	0.3690
UD_GU_CT_5	0.5429	0.4357	0.3786
Average_median	0.2929	0.2536	0.2262

Table 3: Evaluation scores for 5 automatic runs for clinical trials.

Run ID	**infNDCG**	**P @ 10**	**R-prec**
UD_GU_SA_1	0.3872	0.5933	0.2400
UD_GU_SA_2	0.4024	0.6233	0.2413
UD_GU_SA_3	0.3884	0.5833	0.2400
UD_GU_SA_4	0.4027	0.6200	0.2434
UD_GU_SA_5	**0.4135**	**0.6400**	**0.2477**
Average_median	0.2685	0.3586	0.1738

Table 4: Evaluation scores for 5 automatic runs for scientific abstracts.

4. CONCLUSION

In this paper we described the system developed for the TREC 2017 PM track. We employed a two-part system to generate the ranked list of clinical trials and scientific abstract. The first part pertains to query expansion and document retrieval from index. We expanded the gene and disease names by using The Entrez database and Disease Ontology/MeSH, respectively. The second part pertains to generating the final ranked list by implementing a heuristic scoring method. The scoring for clinical trials involved grouping trials based on different trial fields and extraction of features based on occurrences of gene/disease and other terms in the trial. The scoring for scientific abstracts involved applying a NLP system to extract relations from text, as well as extraction of additional information relevant to precision medicine. In total, we submitted 5 different runs for each of the document sets.

ACKNOWLEDGEMENT

This effort is supported by the NIH BD2K Data Wrangling Awards – Grant Number: 1 U01 HG008390-01. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the poster.

REFERENCES

[1] Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res. 2015;43: D1071–8.

[2] Mahmood ASMA, Wu T-J, Mazumder R, Vijay-Shanker K. DiMeX: A Text Mining System for Mutation-Disease Association Extraction. PLoS One. 2016;11: e0152725.

[3] Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41: W518–22.

[4] Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput. 2003; 451–462.

[5] Narayanaswamy M, Ravikumar KE, Vijay-Shanker K. A biological named entity recognizer. Pac Symp Biocomput. 2003; 427–438.

[6] Mahmood ASMA, Rao S, McGarvey P, Wu C, Madhavan S, Vijay-Shanker K (2017) eGARD: Extracting associations between genomic anomalies and drug responses from text. PLoS ONE 12(12): e0189663. https://doi.org/10.1371/journal.pone.0189663