CSIRO at 2017 TREC Precision Medicine Track

Vincent Nguyen
vincent.nguyen@csiro.au

Sarvnaz Karimi
sarvnaz.karimi@csiro.au

Sara Falamaki
sara.falamaki@csiro.au

Diego Molla-Aliod
diego.molla-aliod@mq.edu.au

Cecile Paris
cecile.paris@csiro.au

Stephen Wan
stephen.wan@csiro.au

CSIRO Data61
Marsfield, NSW, Australia

ABSTRACT

We report on our participation as the CSIROmed[1] team in the TREC 2017 Precision Medicine track. We submitted five runs for the scientific *abstracts* collection (MEDLINE and Cancer Proceedings), and five runs for the *clinical trials* collection. We experimented with a number of query expansion and search result re-ranking techniques. We used citation and MeSH-based re-ranking methods, as well as re-ranking based on a merging algorithm proposed for federated search. Our results show that boosting the gene variant in the query increases the relevance of the retrieved results. One of our five runs for clinical trials task was ranked in top 10 runs out of 133 runs submitted for this task.

1. INTRODUCTION

Precision Medicine (PM) is the development of treatment plans beyond observable signs and symptoms. It takes into account the patients' unique genetic markup, environmental influences, and lifestyle choices, as well as other biomarker information for an individual's prevention, diagnosis, and treatment strategies [2].

The TREC Precision Medicine track[2], a specialisation of the TREC Clinical Decision Support track, aims to tackle the challenge of including genetic information in designing treatment strategies. Concretely, it aims to provide, to the medical staff, clinical decision support for cancer patients with an emphasis on precise treatments based on the patient's genetic makeup. The task this year was to retrieve relevant biomedical literature and clinical trials for clinical decision support given a query with the patient's genetic mutations, past medical history, and demographic attributes.

In this report, we outline our approach, discuss the experimental setup and present our results.

2. DATASET

Three datasets were utilised for the Precision Medicine track (PMT). The first set of documents for the PMT'17 was taken from published medical literature on PubMed Central. It contained approximately 26.8 million journal abstracts from a January 2017 snapshot in XML format. The second set of documents were taken from an April 2017 snapshot of ClinicalTrials.gov. It contained 241,006 clinical trials in XML format. Finally, the last set of documents was taken from AACR (American Association for Cancer Research) and ASCO (American Society of Clinical Oncology) proceedings that were focused on cancer therapy. It contained 70,025 documents in plain text.

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

Figure 1: A TREC PM topic (Topic 1).

Topics for PMT'17 detailed four key pieces of information about the patient: disease, genetic variation, demographic attributes, and other relevant medical information. Figure 1 shows one of the topics which is related to a male patient in his late 30's with Liposarcoma with amplification of the CDK4 gene. The patient also has gastroesophageal reflux disease (GERD).

3. INDEXING

We created three separate indices (clinical trials, MEDLINE abstracts, and extra conference abstracts) using the default settings of the Apache Solr search engine[3]. We removed excess whitespace and newline characters in a preprocessing step on the MEDLINE abstracts. The Stop word removal and stemming was performed on the documents and queries automatically by Solr. Processing the abstracts was done at index time and only the following fields were retained in the index: pmid (Pubmed ID), pmcid (Pubmed Central ID), title, abstract, article type, MeSH headings, article keywords, and date published.

Likewise, only the following fields: title (brief and official), brief summary, detailed description, nct-id (clinical trial registry numbers), intervention type and intervention, inclusion and exclusion criteria, condition browse (a field containing MeSH keywords), and primary outcome were processed in the clinical trials. Inclusion and exclusion criteria were preprocessed from eligibility criteria by apply-

[1] CSIRO: Commonwealth Scientific and Industrial Research Organisation.

[2] http://trec-cds.appspot.com/2017.html (Accessed 16 Jan 2018)

[3] Solr version 6.6.0 http://lucene.apache.org/solr/ (Accessed 24 Oct 2017)

ing simple regular expressions and were used to restrict the query demographics for clinical trials. The age of patients was recorded in days in order to avoid floating point arithmetic. Any MEDLINE ids that were referenced by the clinical trial were also included in the index. Finally, the extra conference abstracts were processed with the included fields: id, meeting, title and abstract.

The indexing of these corpora was done on a single PC (8 physical cores), with parallel processing (8 processes), in approximately 8 hours. MEDLINE abstracts took the longest to index at seven hours.

4. QUERY PROCESSING

The topics were preprocessed before they were used as queries in the search engine. Aside from using topic terms as bag-of-words, where word order is discarded, we expanded the queries for the fields: *disease*, *gene* and *other*. The query expansion terms were weighted lower than the original query terms at 0.2 or 0.3 depending on the run. Every topic is always expanded with the term *neoplasm*. This is because all the topics are cancer-related. An example of query processings executed on the topic in Figure 1 is shown in Figure 2. These steps are explained in the following sections.

Gene Expansion

As the gene mentions in the query topics were often expressed in abbreviated forms, the abbreviations were expanded using the Human Gene Ontology[4]. All genes were expanded using the first result found in the ontology. For example, *CDK4* was expanded to *cyclin-dependent kinase 4*.

The expanded versions were already present in the documents in the corpus.

Disease Expansion

We expanded the disease names in the topics in two different ways: *Metamap filtering* and *semantic variation*.

In the *Metamap filtering* method, we use a combination concepts extracted by MetamapLite [1] and Wikipedia suggested terms, to expand mentions of diseases in the topics. That is, we queried each disease name on Wikipedia[5] through its *similar terms* API and retrieved the most similar words (T_W). We then ran Metamap over the retrieved terms T_W to find a new set of terms (T_M). We subsequently used the intersection of these two sets (T) as expanded terms: $T = T_W \cap T_M$. This helped limit the number of words added to the query to ensure that the retrieved documents were focused more on the query terms, rather than the expanded terms. This is important because the query itself is very short.

In the second method, we generate *semantic variations* of diseases mentioned in the queries using Wikipedia and MEDLINE word embeddings. These were trained using Word2Vec[6]. By using a high threshold, we limited the results to a maximum of three most similar words per query.

Demographic Attribute Expansion

Demographic attributes are precisely specified in clinical trials. By normalising the demographic attributes in the queries, we found exact matches in the clinical trial corpus. For example, a query containing the string *person at the age of 6* was replaced with the word *child*, and a query containing the term *female* was expanded with the words *woman women*. We normalised the ages by counting them in days in order to avoid date arithmetic. We then used Solr's boolean query operators to exclude all clinical trials that don't match the patient's demographic attributes. The corresponding boolean query is shown below.

```
fq:"−gender:male AND maximum_age:[0 TO 5110]"
```

This operator will exclude documents that are either for males or for individuals over the age of 15 (5,510 days). Conversely, this will restrict the results to only include patients that are female and under the age of 15.

5. NEGATION DETECTION

By performing negation detection (removing negated terms) we reduced false-positives in the retrieved results. For example, documents matching the query term 'CDK4' sometimes contained sentences like 'CDK4 amplification was negative'. Clearly, this document is irrelevant for the patient. We performed negation detection using the NegEx algorithm implemented in MetamapLite, which detected medically related negated terms.

Due to the long processing times of negation detection, we processed the top 300 documents of each query for each of the three indices and built a cache to speed up future queries.

6. RE-RANKING USING CITATIONS AND MESH TERMS

Clinical trials often reference MEDLINE articles. Since retrieved trials strictly match the patient's demographic attributes, we can boost the MEDLINE articles that they cite in their references, as they are more likely to be relevant. We do this by extracting the PMIDs from the retrieved clinical trial documents and boosting these by the reciprocal rank of the trial document. That is, a document in the retrieved MEDLINE set was boosted by the maximum amount if it was found in the most relevant clinical trial.

Another way that we approached re-ranking the search results using citations was for clinical trials. We used the ranking of the cited MEDLINE abstracts to change the ranking of the trial that cites it. That is, we ran searches separately on clinical trials and MEDLINE, and then re-ranked trials based on the reciprocal ranks of their cited articles. The intuition behind boosting the clinical trial is that if the clinical trial is ranked low, but it is linked to the most relevant MEDLINE abstract, then we assume the clinical trial is at least partially relevant and rank it higher. We note that, we did not use AACR and ASCO abstracts for citation boosting because we could not find any of the clinical trials in our set that had referenced those abstracts.

We also used MeSH terms to create a link between the MEDLINE and Clinical trial indices. In other words, we used MeSH terms in order to compute how similar two documents were: the more MeSH terms that matched between

[4] http://www.geneontology.org/ (Accessed 16 Jan 2018)

[5] https://pypi.python.org/pypi/wikipedia (Accessed 24 Oct 2017)

[6] https://radimrehurek.com/gensim/models/word2vec.html (Accessed 16 Jan 2018)

Query processing step	**Orginial or expanded terms**
Demographic Attribute Expansion	
Age (map age to word): 38	adult
Gender (semantic variation of gender) : male	male men man
Gene Expansion	
Ontology expansion: CDK4 Amplification	amplification dependent cyclin kinase 4
Metamap: CDK4 Amplification	technique abnormality amplification gene
Metamap/Wikipedia set intersection for gene	amplification dependent cyclin kinase cks1b 6 9 p14arf inhibitor d1
Disease Expansion	
Metamap/Wikipedia set intersection for liposarcoma	sarcoma liposarcoma lipoblastomatosis myxoid lipoblastoma
Word embeddings (MEDLINE and Wikipedia)	sarcomas thymoma epithelioid osteosarcoma meningioma chondrosarcoma fibroma myxoid leiomyosarcoma hamartoma
Metamap/Wikipedia set intersection for GERD	gastroesophageal reflux disease
Word embeddings (MEDLINE and Wikipedia)	gord
Final query: (male 38 gerd) (male man men adult amplification dependent cyclin kinase cks1b 6 9 p14arf inhibitor d1 sarcoma liposarcoma lipoblastomatosis myxoid lipoblastoma gastroesophageal reflux disease sarcomas thymoma epithelioid osteosarcoma meningioma chondrosarcoma fibroma myxoid leiomyosarcoma hamartoma amplification dependent cyclin kinase cks1b 6 9 p14arf inhibitor d1 gord technique abnormality amplification gene neoplasm)^e (liposarcoma)^s (amplification cdk4)^g	

Figure 2: An illustration of the query processing steps on Topic 1. In the final query, e is the expansion weight, s is the disease boost, and g is the gene boost factor. Note that ^in solr query format is in fact multiplication.

two documents, the more similar the documents are. We applied a small boost that was unlikely to displace the overall ranking but to push up a relevant document by one to two ranks. The boost was applied to both the rankings of the MEDLINE documents and clinical trials and was guided by the cosine similarity between MeSH terms in the documents returned by both indices.

It is important to note that both citation and MeSH boosts were small, decreasing exponentially with the reciprocal ranking, R_d, of the corresponding document in the result sets:

$$S_d = S_d + b(R_d), \qquad (1)$$

where S_d is the score of a document, and b is a boosting function that uses the reciprocal ranking of a matched document in the second document set:

$$b(R_d) = \frac{1}{\exp(R_d)}. \qquad (2)$$

7. MERGING SEARCH RESULTS USING FEDERATED SEARCH

Our choice of having three different indices for the three document sets in this track led us to solve the problem of merging the search results for *abstract* runs. The two indices for the abstracts (MEDLINE and Cancer Proceedings) did not have directly comparable BM25 scores as the MEDLINE documents were much longer. As a result, we decided to use a merging algorithm proposed for federated search to normalise the scores. Specifically, we used the Generalised Document Scoring [3] (GDS) algorithm to achieve this.

The GDS algorithm calculates the generic document score by using the size of the intersection between the document and the query. This score is normalised to be between 0 (no overlap, lowest score) and $\sqrt{2}$ (complete overlap, highest possible score).

A drawback of using this method was that the query was treated as a bag-of-words (each term was given the same weight) even though the most important aspects of the query were the disease name and gene mutation. In order to mitigate this, after applying GDS, we applied a separate boosting factor to documents containing the disease name and gene mutation.

We note that in the clinical trials runs where we applied citation boost, we also used the GDS algorithm because it let us re-rank the retrieved documents based on important sections of the clinical trials (title, inclusion criteria and keywords).

8. SUBMITTED RUNS

We submitted 10 runs to TREC PM: five runs on clinical trials and five runs on medical literature. Details of these runs are described in Table 1. For expansion terms, we used a weight of $e = 0.2$ for the GDS runs, and 0.3 for the BM25 runs. The disease and gene expanded terms were boosted higher than the original query. Genes were boosted by $g = 1.3$ for GDS and 1.5 for BM25 for both clinical trials and scientific abstracts. Diseases were weighted $s = 1.15$ for GDS in scietific abstracts and 1.70 in clinical trials. In our additional post-TREC runs, disease boost was set to 1.5 in the solr query.

9. RESULTS

An overview of our results for the abstracts runs are shown in Table 2. The top row shows TREC Median results over 125 runs submitted by different teams as reported in the TREC Overview [4]. Two of the runs, aCSIROmedAll and aCSIROmedNEG, were higher or close to the TREC median for all three metrics of infNDCG, P@10 and R-Prec. Our best run in terms of infNDCG was aCSIROmedNEG which

Run	Technique						
	Ranking	Citation Boost	Gene Boost	Disease Boost	Negation	MeSH	Demographic Filtering
Abstracts							
aCSIROmedAll	BM25		✓				✓
aCSIROmedNEG	GDS		✓	✓	✓		✓
aCSIROmedPCB	GDS	✓	✓	✓		✓	✓
aCSIROmedMGB	GDS	✓				✓	✓
aCSIROmedMCB	GDS		✓	✓		✓	✓
(PostTREC) aAll+DB	BM25		✓	✓			✓
(PostTREC) aNEG-Neg	GDS		✓	✓		✓	✓
Clinical Trials							
cCSIROmedAll	BM25		✓				✓
cCSIROmedNEG	GDS	✓	✓	✓	✓	✓	✓
cCSIROmedMCM	GDS	✓	✓	✓		✓	
cCSIROmedHGB	GDS	✓	✓	✓		✓	✓
cCSIROmedMCB	GDS		✓	✓		✓	✓
(PostTREC) cAll+DB	BM25		✓	✓			✓
(PostTREC) cNEG-Neg	GDS		✓	✓		✓	✓

Table 1: Specification of the CSIROmed submitted runs as well as additional runs (PostTREC).

Run	**infNDCG**	**P@10**	**R-prec**
TREC Median	0.2766	0.3733	0.1761
aCSIROmedAll	0.2813	0.3933	0.1759
aCSIROmedNEG	0.3092	0.3733	0.2000
aCSIROmedPCB	0.2705	0.3176	0.1793
aCSIROmedMGB	0.1257	0.2200	0.0762
aCSIROmedMCB	0.2668	0.3233	0.1811
(PostTREC) aAll+DB	0.3023	0.4067	0.1885
(PostTREC) aNEG-Neg	0.2444	0.2733	0.1665

Table 2: CSIROmed results for search over abstracts.

Run	**P@5**	**P@10**	**P@15**
TREC Median	0.2896	0.2517	0.2253
cCSIROmedAll	0.4138	0.3586	0.3172
cCSIROmedNEG	0.2828	0.2552	0.2529
cCSIROmedHGB	0.3172	0.2897	0.2644
cCSIROmedMCB	0.2828	0.2655	0.2552
cCSIROmedMCM	0.2414	0.2552	0.2368
(PostTREC) cAll + DB	0.4345	0.3793	0.3241
(PostTREC) cNEG - Neg	0.2759	0.2621	0.2276

Table 3: CSIROmed results for search over clinical trials.

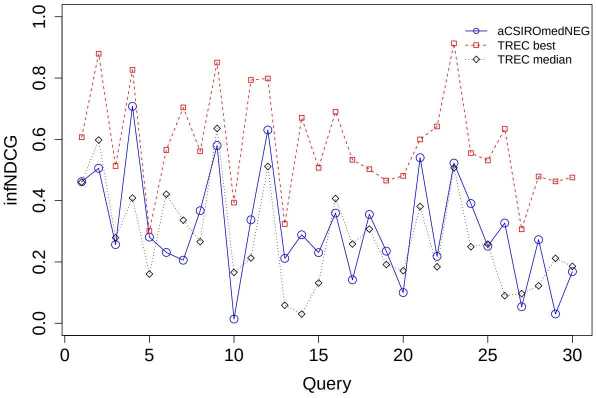

Figure 3: Per query comparison of our best run for abstracts versus the TREC best and median.

used negation detection, GDS ranking and demographic filtering. This could indicate the effectiveness of negation detection in this task. In one of our runs (aCSIROmedMGB) we had turned off gene boosting. Not surprisingly, that run achieved lowest scores for all three metrics which emphasises the importance of gene boosting for this task.

We analysed our best run for the abstracts, aCSIROmedNEG, per queries to see which queries were more successful and which were not (Figure 3). This run did substantially worse than average for two queries (Topic 10 and Topic 29). It substantially did better than average for Topic 13 (infNDCG of 0.2116 versus 0.0588), Topic 14 (0.2887 versus 0.0300 TREC median) and Topic 26 (0.3270 versus 0.0900 of TREC median). Note that the TREC best is the best submitted per query, and it does not correspond to one run.

Our results for the clinical trials runs are shown in Table 3. The top row shows TREC Median results over 133 runs submitted by all the participants. For these runs, evaluation metrics were precision at the cut-off of 5, 10, and 15. Our best-submitted run was cCSIROmedAll which made it to the top-10 submitted runs to this task as well. It used BM25 for ranking, without GDS re-ranking, plus boosting the trials using gene and disease names. It achieved P@5 of 0.4138, while the median of all the runs was 0.2896. Our second best run was cCSIROmedHGB which used every technique but negation detection. It used GDS ranking as well. This run led to above the median precision in all three cut-offs. We found all clinical trial runs that using document reranking performed considerably worse than the cCSIROmedAll, that used no re-ranking.

Per query analysis of the results for clinical trials runs is shown in Figure 4. For two of the topics (3 and 10) our submission was (or one of) the TREC best. Out of 30 topics, our run was better than the TREC median for 17 topics (57%) and equal to the median for another five. Topic 28 (Figure 5) led to zero p@5 and p@10 for best and median for all the 133 submitted runs. Only for p@15, there was a ’best run’ achieving a low score of 0.0667. This query had two partially

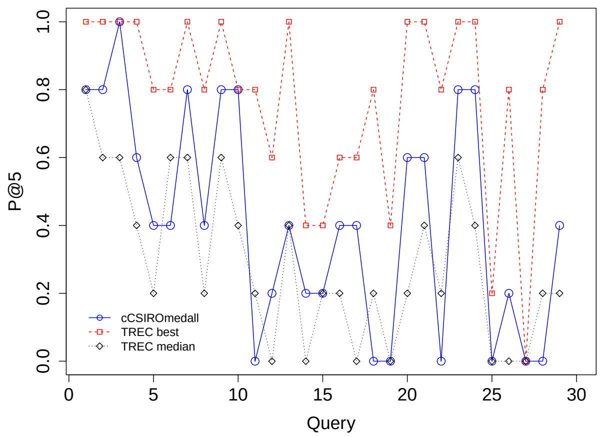

Figure 4: Per query comparison of our best run for clinical trials versus the TREC best and median.

```
<topic number="28">
  <disease>Pancreatic ductal adenocarcinoma</disease>
  <gene>ERBB3</gene>
  <demographic>73−year−old female</demographic>
  <other>hipple , FNA</other>
</topic>
```

Figure 5: Most difficult topic in TREC PM 2017 (Topic 28).

relevant clinical trials identified by the assessors.

Results produced from Post-TREC runs are included in Tables 2 and 3. These runs were produced to test the effectiveness of the boosting of diseases and the effectiveness of negation across clinical trials and abstracts. We found that there was a statistically significant improvement of 7.5% infNDCG in abstract retrieval when disease boosting was added (paired t-test, p-value= 0.0003). Adding disease boost for clinical trials (aCSIROmedall versus aAll+DB) led to improvements, but that was not significant (paired t-test, p-value= 0.0830). One way to see how much negation detection contributed to the submitted runs was to turn them off in post-TREC runs. For abstracts, turning off negation detection led to significant drop of infNDCG from 0.3092 to 0.2444, which translates to nearly 21% drop (aired t-test, p-value= 0.0002). Similarly we observed a drop in precision for clinical trials runs (cCSIROmedNEG versus CNEG-Neg); however, the drop was only statistically significant for p@15 (p-value= 0.0136). That means ignoring the negation detection step hurts abstract runs much more than clinical trial runs. This may be attributed to the facts that negation terms are much more common in the abstract indexes than the clinical trials, and that diseases and genes were commonly negated in the MEDLINE index.

10. CONCLUSIONS

In our submitted runs to the TREC Precision Medicine track, we experimented with a range of query expansion techniques as well as targeted boosting of disease and gene mentions, and negation detection and removal. When expanding the queries, we weighted the disease names and genetic variation(s) higher than the other terms. We limited the number of words added by query expansion to increase the number of relevant documents retrieved.

We submitted runs for both *abstracts* and *clinical trials* document sets. Our best run for the abstracts used negation detection, gene, disease and MeSH boosting. The best run for clinical trials, however, was a plain BM25 ranking where gene mentions were boosted. Our Post-TREC experiments showed varying importance of the disease name boosting and negation detection. In the future, we will extend our analysis of these results through query analysis in order to identify where these techniques improve and hurt retrieval effectiveness.

Acknowledgements. Funding for the team to participate in TREC was provided by the Decision Sciences Research Program of CSIRO Data61.

References

[1] D. Demner-Fushman, W. J. Rogers, and A. Aronson. MetaMap Lite: An evaluation of a new Java implementation of MetaMap. *J Am Med Inform Assoc.*, ocw177, 2017.

[2] I. Konig, O. Fuchs, G. Hansen, E. von Mutius, and M. Kopp. What is precision medicine? *European Respiratory Journal*, 50(4), 2017.

[3] P. Li, P. Thomas, and D. Hawking. Merging algorithms for enterprise search. In *Australasian Document Computing Symposium*, 2013.

[4] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. Hersh, S. Bedrick, A. J. Lazar, and S. Pant. Overview of the TREC 2017 precision medicine track. In *TREC*, Gaithersburg, MD, 2017.