CSIRO at TREC Clinical Decision Support Track

Sarvnaz Karimi
sarvnaz.karimi@csiro.au

Sara Falamaki
sara.falamaki@csiro.au

Vincent Nguyen
vincent.nguyen@csiro.au

Data61, CSIRO
Marsfield, NSW, Australia

ABSTRACT

We report on the participation of the CSIRO[1] team, named as CSIROmed, in the TREC 2016 Clinical Decision Support Track. We submitted three automatic runs and and one manual run. Our best submitted run was the manual run using the summaries. We expanded the summaries with synonyms of diseases, metamap concepts, abbreviations as well as boosting phrases. We also report on experiments post TREC conference, where we analyse effectiveness of some of query processing methods.

1. INTRODUCTION

TREC Clinical Decision Support Track (CDS) is set to pave the way for investigating techniques for linking medical records to information relevant for patient care [5]. For the purpose of this track, the source of information was published medical literature in PubMed Central (PMC). For each medical case, the raw data (topic) provided is structured as a note, a description and a summary. These topics, written in natural language, capture both the past medical history of the patients, and the patient's current condition. The complexity of information need in searches by clinicians [1], a scenario replicated in the TREC track, means that query formulation is an important first step. Examples of complicated search for clinical questions can be found in search for biomedical systematic reviewing [3, 4] where multiple constraints must be satisfied to identify relevant articles. Particularly in the CDS track though, there was a language discrepency between provided information (clinical notes), and the documents to be searched over (published scientific articles), meaning a translation is a necessary first step.

In this report, we outline our approach to the query formulation, discuss the experimental setting, and present the results of using different sections of the provided queries.

2. DATASET

Documents for the CDS track were taken from published medical literature in PubMed Central. It contained 1.25 million journal articles published before 28 March 2016. These documents were in NXML format (XML format extended using National Library of Medicine (NLM) journal archiving and interchange tag library).

[1]Commonwealth Scientific and Industrial Research Organisation

Topics for the 2016 track was composed of three parts: note, description and summary. It was also labeled with a type: *diagnosis*, *test*, or *treatment*. For diagnosis topics, documents relevant to diagnosing the patient were saught. Test topic queries were to return articles guiding the physician in prescribing useful diagnostic tests, and treatment topics were to return articles about treating the patient's condition.

3. INDEXING

We indexed the documents using the Solr search engine. Similar to [2], we pre-processed the documents, replacing all numerals with a globally unique string. Each document was indexed with the following fields: title, abstracts, body, title Metamap concepts, and abstract Metamap concepts. An aggregate field containing all of the above data was also indexed, to aid searching.

Metamap was configured to identify the following concepts: therapeutic or preventive procedure, injury or poisoning, disease or syndrome, organ or tissue function, qualitative concept, body substance, pathologic function, pharmacologic substance, finding, and biologically active substance, and organic chemical. A script (written in groovy) was used to parse each xml file, run metamap on the relevant fields, and insert the data into our Solr search engine.

We indexed the documents using the CSIRO high performance computing systems. Each node processed the data in a single directory, meaning we could process the dataset concurrently. We ran two metamap servers to distribute the load. Our script had the ability to resume if a component raised an exception, which was critical to it being able to process the whole dataset. Each article took approximately 30 seconds to process, so on a single CPU, the dataset would take approximately 2 years to process. By running the processes concurrently, we completed the indexing in approximately 72 hours.

4. QUERY PROCESSING

The topics provided were preprocessed before being presented as queries to the search engine. For all the submissions, we used a set of heuristics to expand the medical shorthand in the topics (Table 1).

In all the runs we submitted, we configured Solr to boost particular elements in each document. This was so that the more concise parts of the document take precedence over the body of the text, and that the original parts of the document take precedence over our augmentations. We used

Shorthand	Expansion
M	Male
F	Female
hx	medical history
hotn	hotn hypotension
htn	htn hypertension
pt	patient
pmh	past medical history
pmhx	patient medical history
prn	when necessary
∼	approximately
h/o	history of
y/o	–
w/	with

Table 1: Heuristics for query processing.

the following boost factors: "titleˆ2 abstractˆ2 bodyˆ1.1 abstract metamapˆ1 title metamapˆ1". Search results were filtered to only retrieve documents that contained at least an abstract, ignoring title only publications. We also set proximity matching as "abstractˆ1.2 titleˆ2".

For our automatic runs, we used a script to extract metamap concepts from the notes, description and summary fields in the topic. We then tried different combinations of metamap data and original text. The metamap concepts used were the same as the ones used when indexing the articles.

The runs we submitted were as follows: CSIROmeta used notes both in the form of bag-of-words and metamap concepts. CSIROsumm used summaries as bag of words plus the metamap concepts in the summary. CSIROnote used notes only. CSIROdSum used both description and summaries. This run (mistakenly) did not comply with the instructions for submitting based on one of the fields only. We also submitted a manual run (CSIROmnul) using the summary fields. For the one manual run we submitted, we processed the queries further using the following steps:

- Used a *noun phrase extractor* in the NLTK toolkit to extract keyphrases in the topics;
- The extracted chunks were matched against Mayo Clinic or Wikipedia to expand names of diseases; and
- The extracted chunks were expanded against a dictionary of medical abbreviations, if dictionary words were found in the chunk.

5. RESULTS

Evaluation results from our four runs are listed in Table 2.

They are reported using four metrics: infAP, infNDCG, R-prec (precision at R where R is the number of known relevant documents), and P@10. P@10 values are exact since all top 10 documents retrieved were judged for each run. We averaged the TREC reported scores over all the runs in categories of notes and summaries, both for manual and automatic runs, and reported them in Table 2. *Avg. med. auto* and *Avg. best auto* are averaged over all topics in automatic runs for median and best reported results for the 26 groups that participated in this track. Similarly, we averaged the best and median scores for manual runs.

Run	infAP	infNDCG	R-prec	P@10
CSIROdSum	0.0077	0.1142	0.0628	0.1600
Summary				
CSIROsumm	0.0119	0.1358	0.0731	0.2167
Avg. med. auto.	0.0196	0.1859	0.1220	0.2633
Avg. best auto.	0.0868	0.4377	0.2554	0.6300
Note				
CSIROmeta	0.0078	0.0958	0.0401	0.2167
CSIRONote	0.0093	0.1052	0.0520	0.1600
Avg. med. auto.	0.0099	0.1228	0.0792	0.2000
Avg. best auto.	0.0599	0.3302	0.1994	0.5100
CSIROmnul	0.0168	0.1570	0.0898	0.2700
Avg. med. manual	0.0149	0.1593	0.0967	0.2433
Avg. best manual	0.0745	0.3805	0.1977	0.5800

Table 2: Results of CSIRO submitted runs compared to the average on best and median results of all submitted automatic and manual runs. Average taken over all 30 queries.

Run	infAP	infNDCG	R-prec	P@10
CSIROdSum	0.0073	0.1174	0.0854	0.1333
Summary				
CSIROsumm	0.0149	0.1626	0.1191	0.1867
Note				
CSIROmeta	0.0075	0.0869	0.0564	0.1000
CSIRONote	0.0098	0.1224	0.0783	0.1733
CSIROmnul	0.0200	0.1899	0.1306	0.2500

Table 3: Results of experiments post TREC conference. Boosting factors were removed.

The first row in Table 2 refers to our run with both summary and description. While this run was submitted in error, the results indicate that it performed worse than all our other runs that used only one source of information (note or summary).

Our automatic run with summaries and their Metamap concepts (CSIROsumm) was consistently below average on all four metrics. The same for CSIRONote which used notes only. When we added Metamap concepts to the notes, we improved P@10 over the averaged median results by approximately 0.22 (5%).

Our best submission was our manual run which resulted in higher than average infAP and P@10, at 0.0168 and 0.27 respectively. This means that adding synonyms for disease names and expanding abbreviations improved our results.

Our manual run was most successful in topics of type *test* and performed worst in topics of type *treatment*. However, one main drawback of our algorithm was ignoring the query types in the retrieval process. This is particularly important because topics were not in the form of questions and therefore the nature of the information required was ambiguous without referring to the query type.

After the release of the relevance judgements for the track, we ran complementary experiments to investigate the effect of our query processing steps. In particular, we were interested to measure the value of different boosting methods, such the article title boost. Results are shown in Table 3.

The results from the complementary experiments indi-

cated that the boosting factors that were used during query processing time of our runs worsened the recall and precision in some runs (CSIROnote, CSIROdSum, CSIROsumm, CSIROmnul) while having no effect on other runs (CSIROmeta). This is reflected in infNDCG, infAP and R-prec metrics. However, in terms of the P@10 (Precision at 10), boosting factors ensured more relevant documents, on average, were being placed in the top 10 query results. It should be noted that the overall changes observed in the results for the complementary runs are marginal and not statistically significant (paired t-test).

A common trend in the CSIROmnul, CSIROnote and CSIROmeta runs, without boosting factors, is that they performed better for topics in test (topics 11 - 20) and treatment (topics 21 - 30) categories than the previous runs that had the aforementioned boosting factors. This trend was not true for CSIROdSum and CSIROsumm however as these runs showed minor improvement in topics from diagnosis, treatment and test when boosting factors were removed.

6. CONCLUSIONS AND FUTURE WORK

According to the judgements received from TREC 2016, our best run was our manual run, augmented with disease synonyms and manual abbreviation expansion. In the future, we can build on this method and fully automate it.

In order to better understand the information needs in the query, we need to use the query type as part of our query expansion. This should help to achieve more relevant results.

The results we obtained are preliminary, and did not use any learning mechanism. As such they form a baseline of what is achievable. We will experiment with learning to rank methods and query expansion in our future query processing algorithms.

References

[1] Y. gang Caoa, J. J. Ciminob, J. Elyc, and H. Yua. Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics*, 43.

[2] S. Karimi, D. Martinez, S. Ghodke, L. Zhang, H. Suominen, and L. Cavedon. Search for medical records: NICTA at TREC 2011 medical track. In *Text REtrieval Conference*, 2011.

[3] S. Karimi, S. Pohl, F. Scholer, L. Cavedon, and J. Zobel. Boolean versus ranked querying for biomedical systematic reviews. *BMC Medical Informatics and Decision Making*, 10(58), 2010.

[4] S. Karimi and F. Scholer. Systematic reviews: A complex search episode for evidence based policy and practice. In *SIGIR 2011 Workshop on "entertain me": Supporting Complex Search Tasks*, page 7, 2011.

[5] K. Roberts, M. S. Simpson, E. Voorhees, and W. R. Hersh. Overview of the trec 2015 clinical decision support track. In *Text REtrieval Conference*, Gaithersburg, MD, 2015.