Query Modification through External Sources to Support Clinical Decisions

Raymond Wan[1], Jannifer Hiu-Kwan Man[2], and Ting-Fung Chan[1]

[1]School of Life Sciences and the State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong
({rwan,tf.chan}@cuhk.edu.hk)
[2]Department of Anaesthesia, Princess Margaret Hospital, Hospital Authority (Hong Kong), Laichikok, Hong Kong
(manhkj@ha.org.hk)

Abstract

For the Clinical Decision Support Track of TREC 2014, we looked into the effect of modifying queries by adding or removing terms. We considered both automatic and manual query modifications that use either external data sources or a domain expert. While each method gave slightly different results, we discovered that the manual method still performed slightly better among the methods we considered. This is despite the fact that the manual queries were formed through just term removal; no new terms were added.

1 Introduction

A team comprised of researchers from The Chinese University of Hong Kong and a clinician from Princess Margaret Hospital participated in the Clinical Decision Support Track of TREC 2014. While the problem of linking medical cases to biomedical literature offers many possible paths of investigation, our study focused on modifications to the query using external data sources. We submitted 5 runs in total using different variations to the queries – 4 of them were automated while one was manual. A couple of additional runs were also performed after the relevant judgements were released.

This report is structured as follows. Section 2 gives some background to the problem and the data. Our motivations for our approach is described in Section 3. Our workflow is detailed in Section 4. Results from all of our runs are discussed in Section 5 and our findings are summarized in Section 6.

2 Background

Both a document collection of biomedical texts and a set of topics were made available for participants of the track [Simpson et al., 2014].

The document collection consisted of articles in XML format using the National Library of Medicine's Journal Archiving and Interchange Tag Set. There were 733,328 articles in this NXML format[1], with each article uniquely identified by their PubMed Central Identifiers (PMCIDs). There were 30 topics in total, with

[1]Simpson et al. [2014] reports 733,138 articles – this slight difference might be due to our method used to pre-process the document collection.

Type	Description
Description	A 58-year-old African-American woman presents to the ER with episodic pressing/burning anterior chest pain that began two days earlier for the first time in her life. The pain started while she was walking, radiates to the back, and is accompanied by nausea, diaphoresis and mild dyspnea, but is not increased on inspiration. The latest episode of pain ended half an hour prior to her arrival. She is known to have hypertension and obesity. She denies smoking, diabetes, hypercholesterolemia, or a family history of heart disease. She currently takes no medications. Physical examination is normal. The EKG shows nonspecific changes.
Summary	58-year-old woman with hypertension and obesity presents with exercise-related episodic chest pain radiating to the back.
Manual	hypertension obesity exercise-related episodic chest pain radiating to the back

Figure 1: Description, Summary, and a manually simplified version of the Summary for topic #1.

10 topics for each of the following three pre-assigned categories (or *type*: diagnosis, test, and treatment). Each topic appears as a verbose Description and a shorter, more succinct Summary, all in XML format.

The task at hand for each query was to return a set of biomedical texts which correspond to the patient's diagnosis, additional test(s) required, or the prescribed treatment, depending on the topic's type.

3 Motivation

After examining the topics, we noticed that the Summaries have all been transformed uniformly from the associated Descriptions. We believe the experts that created the Summaries did this so as not to give TREC participants any "hints" for any particular topic. In some cases, we believe that the Summaries could be made more precise by *removing* additional terms manually.

As an example, the Description, Summary, and a Manual transformation for topic #1 is shown in Figure 1. This manual transformation by one of the authors who is a clinician (JHM) has the age and the gender of the patient removed, since both pieces of information were deemed unimportant to this case. Manual transformations such as this one by JHM formed the basis of our single manual run.

Thus, our aim was to see if retrieval effectiveness could be affected by straight-forward transformations to the query. Besides the aforementioned manual transformation where terms were removed, we also considered automatic transformations that both added and removed terms. We compared the effectiveness of our approaches against the baseline, which used the Summary queries as-is.

4 Workflow

Our investigation used version 4.0 of the Terrier retrieval system [Ounis et al., 2006]. The default stopword list, Porter stemming algorithm [Porter, 1980], and term frequency-inverse document frequency (TF-IDF) weighting were used throughout.

We used an Intel® Core™ i5-2400 CPU (3.10 GHz) with 6 GB cache and 16 GB RAM for all of our experiments, except for the **`icdqe`** run which also used a 2.33 GHz Intel® Core™ 2 Quad CPU Q8200 with 8 GB RAM. Both systems ran the Ubuntu 14.04 operating system.

Five runs were submitted for evaluation. After the relevance judgements were released, two additional runs were performed. For all of our runs, only the Summary queries were used; the longer Description queries were not considered.

```
<DOC>
<DOCNO>2630849</DOCNO>
<TITLE>
Immunomodulatory Effects of Domoic Acid Differ Between In vivo and In vitro Exposure in Mice ...
</TITLE>
<ABSTRACT>
The immunotoxic potential of domoic acid (DA), a well-characterized neurotoxin, has not been fully ...
</ABSTRACT>
<BODY>
1. Introduction Certain species of the marine diatom Pseudo-nitzschia produce the neurotoxin domoic ...
</BODY>
</DOC>
```

Figure 2: A sample document (PMCID 2630849), after pre-processing but prior to indexing by Terrier. The ellipses indicates lines have been truncated to facilitate easier display.

We described each of these runs in the context of the workflow used to pre-process the document collection and the topics before being input into Terrier.

4.1 Document Collection

Our first step in processing the documents was to removed the 137 articles from the list of duplicates[2]. Afterwards, each document was processed one-by-one as follows.

While the document collection consisted of well-formed XML files, we were interested in only certain fields. Even though Terrier could be configured to index parts of the articles based on certain XML tags, we decided to transform the documents ourselves so that the title, abstract, and body could be identified easily.

Both the title and abstract were taken from the document's <front> tag. Under this tag, the title was taken from the <article-title> tag. If that was not available, then the <subject> tag was used instead. The abstract was taken from the <abstract> tag.

As for the body, if sections were available as individual <sec> tags, they were concatenated together to form the body. If sections could not be found, then the entire article's body was taken.

Throughout this process, citations and external links (characterized by <xref> and <ext-link> tags, respectively) were also removed. An example document is shown in Figure 2, with only the first few words shown for each section.

Terrier was configured to index the title, abstract, and body of each article. In total, 733,190 documents were indexed by Terrier.

4.2 Topics

Pre-processing of the topics involved modifying the Summary topics prior to passing them to Terrier. Generally, two different approaches were considered, as shown in Figure 3. The first was query expansion – where additional terms were added to the query itself. The second was query reduction, where only terms that were found in an external source were retained. Thus, words not found in the external source were removed from the query.

Each run only used one of these two approaches even though both could have been employed simultaneously. Furthermore, for each run, we used at most one external source for query reduction.

[2]See http://www.trec-cds.org/duplicates-1.txt.

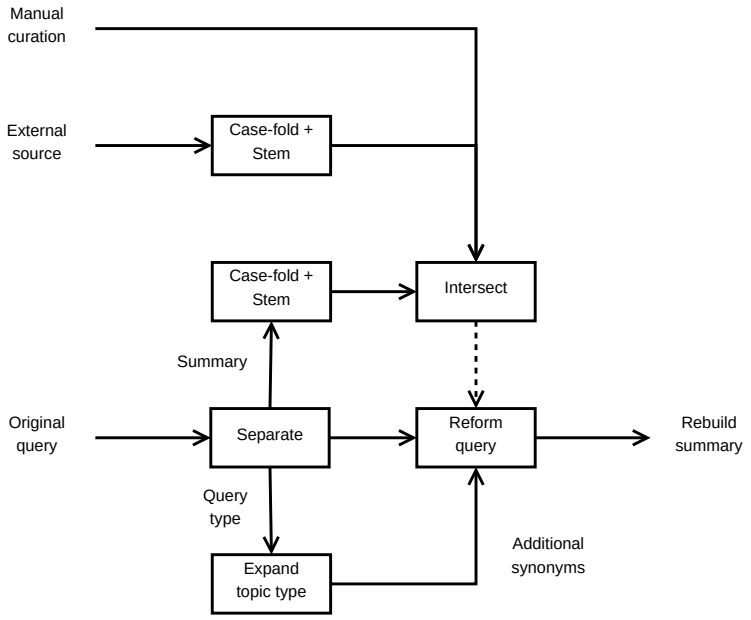

Figure 3: Overall workflow for transforming the queries that is applicable to all runs except for the **icdqe** run.

We explain the procedure used for the 5 submitted runs, followed by the 2 additional runs performed.

4.2.1 Submitted Runs

Query expansion was applied to just the topic type. Recall that there were 30 topics in the query set with three different types: diagnosis, test, and treatment. These topic types were encoded within the <topic> tag as a value for the "type" attribute. Using WordNet 3.0, we expanded the query to include noun synonyms for these three terms. For example, "diagnosis" resulted in the addition of "designation" and "identification" to the search query. This was the approach used for our run called **origexp** and is indicated in Figure 3 by the box "Expand topic type".

On the other hand, query reduction treats each query as a set of independent words and intersected it with those from an external source. Only words that existed in both the query and the external source were kept.

As mentioned in Section 3, the first "external source" considered was manual curation by one of the authors (JHM). These efforts resulted in the only manual run (**manual**), as shown by the line[3] at the top-left of Figure 3. She identified which words were important but did not add or change any words. In hindsight, a more thorough modification might have yielded better results. For example, topic #2 mentions that the patient had just returned from a trip in Colorado. In this case, the exact place that the patient went to is not important but only that he had been away from home for a period of time. Capturing this idea is not easy for a manual run, and even more difficult for an automated run.

The two external sources which we considered were the International Classification of Diseases (ICD-10) [World Health Organization, 1992] and Medical Subject Headings[4] (MeSH).

The 2010 version of ICD-10 in Classification Mark-up Language (ClaML) format was obtained and all words that appear in the tag <Rubric> and then <Label> were identified. An example phrase in

[3]Though an example of an "external source", no case-folding or stemming was performed. This is unlike the other methods that we will discuss below.

[4]See http://www.nlm.nih.gov/mesh/filelist.html.

```
<DOC>
<DOCNO>1</DOCNO>
<NAME>
Certain infectious and parasitic diseases
</NAME>
<DESCRIPTION>
diseases generally recognized as communicable or transmissible
</DESCRIPTION>
</DOC>
```

Figure 4: A sample "document" from the re-parsing of the ICD-10 data.

the data is "Acute with both haemorrhage and perforation". All of the phrases were split into words with duplicates removed. The result was a list of 12,565 unique words.

Similarly, the 2014 MeSH descriptor records in XML format were downloaded and a word list was formed. Words that appeared in the tag <String> which was in turn under <DescriptorName>, <QualifierName>, <ConceptName>, or <Term> were combined to form a list of 58,720 unique words.

The words in either list were *intersected* with the words of the original queries to produce shorter sets of queries. For the purpose of the intersection of the word lists, stemming using Porter's stemming algorithm[5] and case-folding were used. This results in two separate runs: **icd** and **mesh**.

For our last submitted run (not shown in Figure 3), we again used the ICD-10 data. However, this time, our aim was to perform query expansion using terms from ICD-10 that are themselves associated with the original queries. In short, this meant employing Terrier twice. The second execution is as before. The first execution creates an index on the ICD-10 data. Each "preferred" <Rubric> was enclosed in a <NAME> tag. Each "inclusion" <Rubric> was transformed into a <DESCRIPTION> tag. An example of such a transformation is shown in Figure 4 shows an example of such a document. This parsing resulted in 4,148 documents.

We then used the original queries as input to this first instance of Terrier and obtained the top 3 documents for each query. All of the words from these documents were then *concatenated* on to the original query for the second Terrier run. This resulted in the **icdqe** run.

4.2.2 Additional Runs

Two additional runs were performed after the relevance judgements were released. One run is called **orig** and it represents submitting the queries to Terrier without any modification to serve as our baseline.

Also, after submitting **origexp**, we realized that the synonyms proposed by WordNet were added to the original query but not the original query types (i.e. "diagnosis", "test", and "treatment"). This mistake was corrected for a new run dubbed **origexp2**.

5 Discussion

Table 1 gives a summary of the 7 runs. Figure 5 presents the inferred normalized discounted cumulated gain (infNDCG) [Yilmaz et al., 2008] for our runs, using **orig** as the baseline (black lines). Along the horizontal axes is the topic IDs from 1 to 30. If a colored line is above the black line for a given query, then the run corresponding to that line performs better than the baseline for that query.

[5]The Perl implementation from http://tartarus.org/martin/PorterStemmer/perl.txt was used.

Table 1: Summary of our 7 runs. Only the first 5 were submitted for assessment.

Run ID	Description
manual	Manual removal of terms from the Summary query by a domain expert.
origexp	Additional synonyms of the topic type appended to the Summary query.
mesh	Intersection of the Summary query with the MeSH vocabulary.
icd	Intersection of the Summary query with the ICD-10 vocabulary.
icdqe	Indexed ICD-10 data and concatenated ICD-10 terms associated with the query on to the original query.
orig	Baseline run with no modifications done to the query.
origexp2	Correction to **origexp** which now includes the original query type.

Table 2: Averaged results across the #30 queries for all of the runs. R-precision, Precision @ 10, and infNDCG are shown.

	orig	manual	origexp	origexp2	mesh	icd	icdqe
R-prec	0.1554	0.1537	0.1266	0.1290	0.1507	0.1315	0.0877
P @ 10	0.2700	0.2933	0.2300	0.2300	0.2767	0.2300	0.1700
infNDCG	0.1775	0.1821	0.1466	0.1476	0.1724	0.1549	0.0947

Overall, the results show that most of the runs perform close to the baseline, since most of the time, the colored lines coincided with the dashed black lines. On a few queries, the **manual** run performs slightly better, except for query #30 (see Figure 5(a)). This is interesting because the **manual** run merely *removed* words from the query. It suggests that a manual run based on domain knowledge that permitted terms to be added or modified may prove useful.

On the other hand, automated runs generally performed worse than the baseline run. According to Figure 5(b), there was a negligible difference between **origexp** and **origexp2**. Thus, the mistake from our original submission had no impact on retrieval results.

The benefit of using MeSH terms and ICD-10 to filter terms in the query is unclear (see Figure 5(c)). But, it would appear that MeSH terms (in red) performed slightly better than ICD-10. Finally, the last panel clearly shows that using two retrieval systems, with one building an index on the ICD-10 data does not work. Besides the added complexity, it also performs much worse than the baseline.

Results for the 7 runs averaged across all of the queries are presented in Table 2. The results in this table reflects our findings from Figure 5. The **manual** run performs the best with **orig** and **mesh** coming second. **icdqe** is the worst out of all 7.

6 Summary

This report summarizes our participation in the 2014 Clinical Decision Support Track. Using the Terrier retrieval system, a bag-of-words approach was combined with TF-IDF weighting to index the Summary topics. We submitted a single manual run (**manual**) and three runs that made use of MeSH and ICD-10 vocabularies for transforming the queries. A fifth submitted run (**origexp**) used synonyms from WordNet to expand the queries. After the relevance judgements were released, two further runs were performed – one to serve as a baseline (**orig**) and another called **origexp2** which corrected a mistake made in generating **origexp**. This report discusses these 7 runs.

The non-baseline runs all modified the queries by simply adding or removing terms through methods such as: (1) manual term removal, (2) adding synonyms to the topic type, (3) intersecting query terms with external data, and (4) augmenting queries with ICD-10 terms associated with the original query, as a

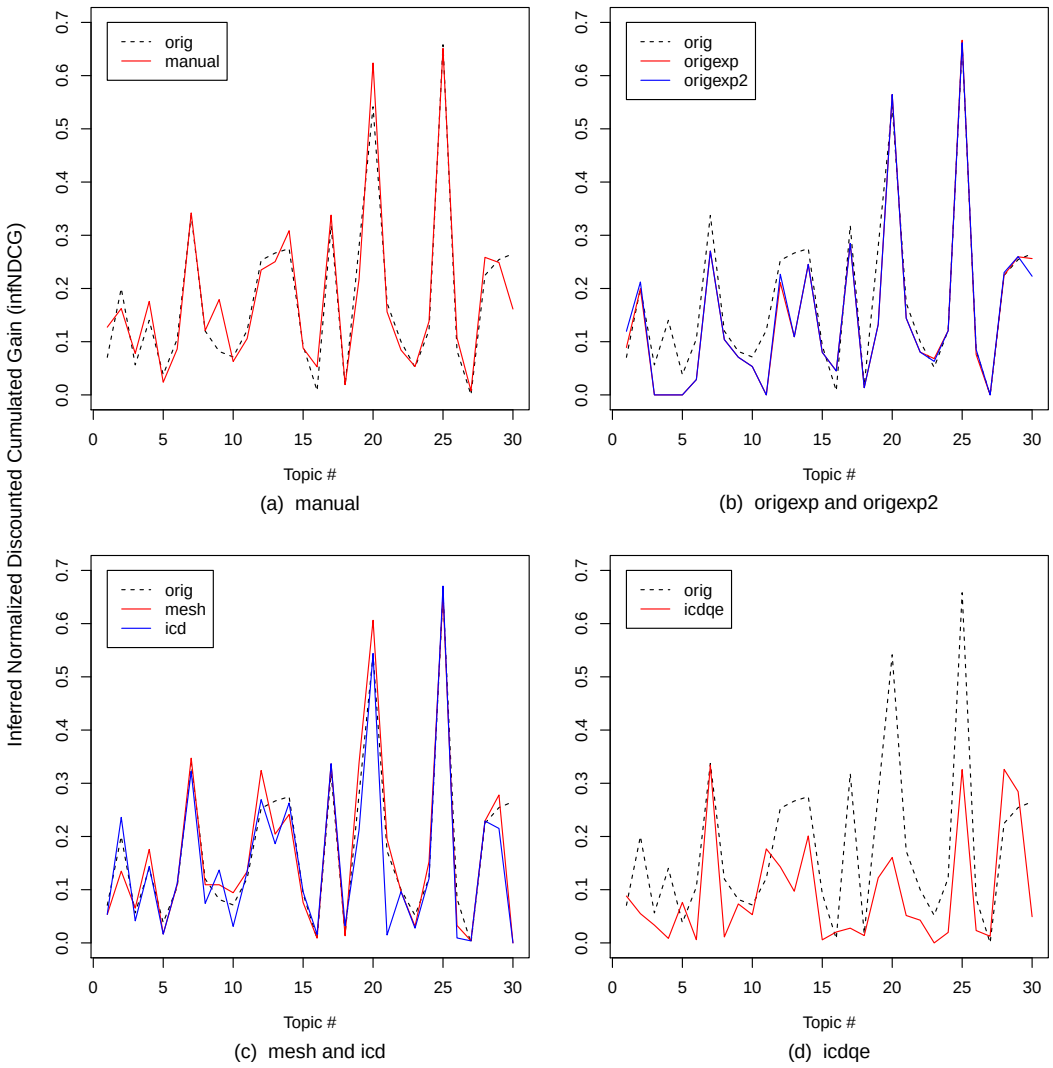

Figure 5: Inferred Normalized Discounted Cumulated Gain (infNDCG) of our runs, using **`orig`** as the baseline (in black).

variation to (3).

Though all of the methods gave similar results, small differences do indicate that simple changes to the queries can affect retrieval performance. While the use of MeSH terms and ICD-10 terms as filters did not necessarily give better results than **`manual`**, we note that all 3 runs simply removed terms. Yet, **`manual`** removed terms in such a way that it performed slightly better than the baseline. Thus, even a simple process such as shortening a query can improve performance.

We anticipate that runs that permit changes other than just query removal could perform better – if so, the questions are under what guidelines could this be done for a manual run and whether or not an automatic run could do the same. For example, as noted for the TREC Medical Records track [Voorhees, 2013], medical records consist of domain specific abbreviations and terms as well as negative language (i.e., "no discomfort") – this is easy to do for a manual run if specific guidelines were provided, but potentially difficult for an automatic run.

Acknowledgements

This study is supported by the Hong Kong RGC General Research Fund (GRF461712), Collaborative Research Fund (CUHK3/CRF/11G), the Lo Kwee-Seong Biomedical Research Fund, and the Lee Hysan Foundation granted to TFC. We thank Kirk Roberts (NLM) for helpful comments on a preliminary version of this manuscript.

References

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, et al. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, August 2006.

M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980. Reprinted in Readings in Information Retrieval, pages 313–316, 1997.

M. S. Simpson, E. Voorhees, and W. Hersh. Overview of the TREC 2014 Clinical Decision Support Track. In *Proc. 23rd Text Retrieval Conference (TREC 2014)*. National Institute of Standards and Technology (NIST), 2014.

E. M. Voorhees. The TREC Medical Records Track. In *Proc. International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 239–246, 2013.

World Health Organization. The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines, 1992.

E. Yilmaz, E. Kanoulas, and J. A. Aslam. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, 2008.