Deriving a Test Collection for Clinical Information Retrieval from Systematic Reviews

Florian Boudin
DIRO, Université de Montréal
CP. 6128, succ. Centre-ville
H3C 3J7, Montréal, Canada
boudinfl@iro.umontreal.ca

Jian-Yun Nie
DIRO, Université de Montréal
CP. 6128, succ. Centre-ville
H3C 3J7, Montréal, Canada
nie@iro.umontreal.ca

Martin Dawes
Dept. of Family Medicine
McGill University, 515 Pine Av.
H2W 1S4, Montréal, Canada
martin.dawes@mcgill.ca

ABSTRACT

In this paper, we describe the construction of a test collection for evaluating clinical information retrieval. The purpose of this test collection is to provide a basis for researchers to experiment with PECO-structured queries. Systematic reviews are used as a starting point for generating queries and relevance judgments. We give some details on the difficulties encountered in building this resource and report the results achieved by current state-of-the-art approaches.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Search process

General Terms

Measurement, Standardization

1. INTRODUCTION

Information Retrieval (IR) systems are evaluated against test collections, which contain a collection of documents, a test suite of information needs (expressible as queries), and judgments (called qrels) as to which documents are relevant to which queries [8]. Retrieval systems return a ranked list of documents (or run) for each query. Retrieved documents are then marked for relevance using the qrels, and evaluation metrics calculated to measure the effectiveness of the run.

The development of test collections for ad hoc IR plays a major role in improving state-of-the-art retrieval methods. As an illustration of that, the retrieval system effectiveness has approximately doubled in the first six years of the Text REtrieval Conference[1] (TREC) [12]. Test collections can be used in repeated experiments to assess and compare retrieval results as well as to optimize system performance. But building a test collection is a long and costly process. A set of topics to run against the collection of documents have to be generated and relevance judgments performed by human assessors. These are very expensive to collect.

In this paper, we describe the construction of a test collection for clinical IR. We propose to use systematic reviews to semi-automatically produce relevance judgments. Clinical studies from which the synthesized results of the review were extracted, are selected as relevant documents. The corresponding clinical queries are then generated by human annotators in respect to the PECO framework [10]. The main purpose of this test collection is to provide a basis for researchers to experiment with PECO-structured queries which recently have gained much attention [3, 1].

Very few test collections have been developed for IR in the clinical domain. Originally created for the TREC topic detection track, OHSUMED [6] is probably the first medical test collection. It consists of a set of 348,566 MEDLINE references, 106 topics and 16,140 query-document pairs that have been judged for relevance using a three point scale: definitely, possibly or not relevant. Recently, Friberg [4] described a swedish medical test collection built from various types of documents (e.g. scientific articles, teaching material, guidelines, patient FAQs, ect.). It consists of a set of 42,255 documents, 62 topics and 7,044 relevance judgments made on a four graded scale. In addition to topical relevance, assessors judged each document for a specific target group (doctors or patients).

The purpose of this paper is threefold: to describe the methodology for building a test collection from systematic reviews; to report the results of state-of-the-art retrieval methods on the test collection; and to encourage the community to use this data for improving clinical IR.

This paper is organized as follows. We first describe the methodology employed for building the test collection. Next, we report baseline results on this data and conclude with a discussion.

2. METHOD

Systematic reviews try to identify, appraise, select and synthesize all high quality research evidence relevant to a single question. The best-known source of systematic reviews in the healthcare domain is the Cochrane collaboration[2]. It consists of a group of over 15,000 specialists who systematically identify and review randomized trials of the effects of treatments. Cochrane Reviews are internationally recognised as the highest standard in evidence-based health care. By providing a reliable synthesis of the available evidence on a given topic, systematic reviews adhere to the

[1] `http://trec.nist.gov`

[2] `www.cochrane.org`

DTMBIO'10, October 26, 2010, Toronto, Ontario, Canada.

principle that science is cumulative and facilitate decisions considering all the evidence on the effect of a treatment [5]. Cochrane reviews are published in the Cochrane Database of Systematic Reviews (CDSR)[3], which in June 2010 contained more than 4,200 reviews. Systematic reviews are not limited to medicine and are quite common in other sciences such as psychology or educational research.

From a general point of view, a systematic review is a summary of the best evidence contained in a set of clinical studies focused on one precise question. More specifically, a Cochrane review contains a reference section, listing all the articles used to answer the clinical question, and hence that can be considered as relevant. One can easily see how useful these reviews are for building a test collection. Indeed, we can use a review to generate a query from the adressed question and create a set of relevant judgments from the cited references. These two tasks are described in details in the following subsections.

2.1 Generating queries

The process of generating queries from systematic reviews is not straightforward. Although each Cochrane review is narrowly focused on one clinical question, it nevertheless covers various aspects of a topic and can hardly be summarized by only one query. To overcome this problem, we decided to generate a set of queries from each review. The goal is to produce multiple query variants (i.e. precise clinical questions) that capture the different aspects of the systematic review.

However, phrasing a precise clinical question that summarises what you want answered is a difficult task. Richardson *et al.* [10] identified the following four aspects as the key elements of a well-built clinical question:

- **Patient-problem**: what are the patient characteristics (e.g. age range, gender, etc.)? What is the primary condition or disease?
- **Exposure-intervention**: what is the main intervention (e.g. drug, treatment, duration, etc.)?
- **Comparison**: what is the exposure compared to (e.g. placebo, another drug, etc.)?
- **Outcome**: what are the clinical outcomes (e.g. healing, morbidity, side effects, etc.)?

These elements are known under the mnemonic PECO. Previous studies have validated the suitability of this structure as a knowledge representation for clinical questions [7]. This is the main query structure that we use in our test collection. This choice is motivated by the fact that physicians are educated to formulate their clinical questions in respect to this structure. Moreover, approaches trying to use the PECO framework in the retrieval process have recently received much attention [3, 1]. However, the almost total absence of PECO search interfaces forces clinicians to still use keywords queries as their main search method. Consequently and for comparison purposes, keywords queries are also generated for each systematic review.

We asked a group of annotators, one professor and four Master students in family medicine, to create queries for a given set of Cochrane reviews. Let us consider the Cochrane review about "*Lymphadenectomy for the management of endometrial cancer*"[4]. The first step was to generate a keywords query, which for the example is:

lymphadenectomy AND *endometrial cancer*

Keywords queries are composed of words or phrases separated by the conjuction and (e.g. *influenza vaccine* AND *asthma*). This decomposition is useful for phrase-based retrieval. Then, after reading the content of the review, annotators were asked to generate PECO structured queries by keeping in mind that these queries have to capture the main aspects of the review. For the previous example, the following queries were created:

1.	[*adult women with endometrial cancer*]	(P_1)
	[*pelvic lymphadenectomy*]	(E_1)
	[*no lymphadenectomy*]	(C_1)
	[*overall survival*]	(O_1)
2.	[*adult women with endometrial cancer*]	(P_1)
	[*pelvic lymphadenectomy*]	(E_1)
	[*pelvic lymph node sampling*]	(C_2)
	[*adverse event: lymphoedema or lymphocyst*]	(O_2)
3.	[*adult women with endometrial cancer*]	(P_1)
	[*pelvic lymphadenectomy*]	(E_1)
	[*no lymphadenectomy*]	(C_1)
	[*recurrence-free survival*]	(O_3)

We observe that the three above queries are quite similar. The Patient-problem (P_1) and Exposure-intervention (E_1) elements are the same among them. This is a normal phenomenon as a review is often focused on one couple of P-E elements and analyses the various possible clinical outcomes.

The various topics included in a Cochrane review, and by analogy the generated PECO queries, can be represented by a tree structure (Figure 1). This representation of the generated PECO queries synthesize the review' content. Decision trees were proposed by [2] to guide the extraction of critical information from randomized controlled trials. It is also possible to derive a partial semantic knowledge representation from the PECO tree. For our previous example, one can generate the following topical representations: In *women with endometrial cancer* (P_1), does *pelvic lymphadenectomy* (E_1) improves *overall survival* (O_1) or *recurrence-free survival* (O_2) compared to *no lymphadenectomy* (C_1)?; In *women with endometrial cancer* (P_1), does *pelvic lymphadenectomy* (E_1) or *no lymphadenectomy* (C_1) produces *adverse event: lymphoedema or lymphocyst* (O_3)?

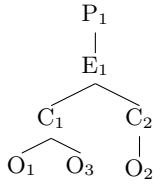

Figure 1: Tree structure representing the PECO queries about "*Lymphadenectomy for the management of endometrial cancer*".

[3] `www.thecochranelibrary.com`

[4] CDSR 2010, Issue 1. Art. No.: CD007585

2.2 Creating relevance judgments

A Cochrane review is a scientific investigation. It includes a comprehensive search of all potentially relevant studies and the use of explicit, reproducible criteria in the selection of studies for review. Reviewers search for relevant clinical studies in multiple medical databases such as MEDLINE[5], EMBASE[6] as well as specialized databases. The References section contains all the studies from which the interpreted and synthetized results of the review were extracted. Our idea is to use these clinical studies as ground truth for the generated queries. More specifically, the References section is composed of several subsections: "*References included in this review*", "*References excluded from this review*", "*additional references*" and "*References to other published versions of this review*". We asked the annotators to extract the citations from the *included in the review* subsection, as they are containing the scientific material needed to answer the clinical question. Although this structure may be used to define a graded scale of relevance, we choose not to do so as additional references contained in the other subsections are not all relevant to the question and should be manually filtered. We will leave this for our future work.

As the purpose of the test collection is to be used by automatic retrieval systems, a unique document identifier have to be assigned to each clinical study. Citations were obtained by searching different sources but most of them are indexed in MEDLINE. Accordingly, we decided to keep only articles published in journals referenced in PubMed (e.g. conference proceedings are not considered). This choice is also motivated by the fact that MEDLINE is the most used medical database and that its citations are freely accessible. Relevant studies were manually mapped to PubMed unique IDentifiers (PMID). This is a very long process that was undertaken by two different annotators to minimize the number of errors.

One drawback of this methodology is that different queries generated from one review share the same set of relevant documents. It is clear that some citations are more relevant to a certain query variant than to the others. There is however no simple solution to this problem. Too much time would be required to analyse the relevance degree of each document in relation to the query variants.

A last point concerns the publication (or last assessed as up-to-date) date of the systematic reviews. Reviews cannot be comprehensive as the scientific literature increases continuously. We purposely included this time information in our test collection. IR systems are then able to prevent retrieving documents published after the review.

2.3 Test collection statistics

We selected in sequential order from the set of new systematic reviews and processed 156 Cochrane reviews. There was no restriction about the topics nor the number of included citations in the References section. The resulting test collection is composed of 423 queries and 8926 relevant citations (2596 different citations). This number reduces to 8138 citations once we remove the citations without any text in the abstract (i.e. certain citations, especially old ones, only contain a title). The average number of documents per query is 15.3 (min = 2, max = 108) while the average length of a document is 246 words. Keywords queries have on average 4.3 words while PECO queries have 18.7 words.

The Figure 2 shows the distribution of relevant documents among the systematic reviews we processed. Disregarding the average number of relevant documents by query, we observe that most of the reviews contain less than 20 relevant documents. This relatively small number of relevant documents reflects the large amounts of work that have been invested by reviewers in sorting and selecting relevant studies. It also gives us a glimpse of the task difficulty.

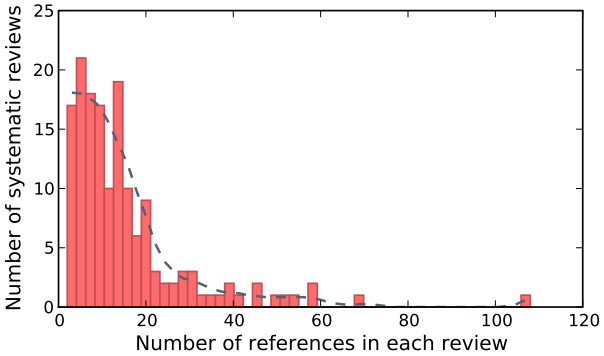

Figure 2: Histogram illustrating the number of relevant references in relation to the number of systematic reviews.

As a collection of documents, we gathered a large number of citations from PubMed using the following constraints: citations with an abstract, human subjects, and belonging to one of the following publication types: randomized control trials, reviews, clinical trials, letters, editorials and meta-analyses. As a result, 1,212,042 different citations were indexed. The goal was to extract a subset as representative as possible of MEDLINE but with a reasonable size.

3. EXPERIMENTS

In this section, we present the results obtained by several baselines on the test collection. We use the language modeling approach to information retrieval. This is one of the state-of-the-art approaches in current IR research. Retrieval tasks are performed using the Lemur toolkit[7] and queries expressed in Indri query language [11]. The number of retrieved documents is set to 1000 and the Dirichlet prior smoothing parameter to $\mu = 2000$. We use a standard list of stopwords (733 tokens) and evaluate the retrieval performance with the latest version of the `trec_eval`[8] tool.

Four baselines are proposed. The first uses keyword queries with the traditional language modeling approach. This model considers each word in a query as an equal, independent source of information. For the query "*cyclosporine* AND *blood pressure*", this model uses the following Indri query:

```
#combine( cyclosporine blood pressure )
```

The second baseline considers multiword phrases. It is clear that finding the exact phrase "*blood pressure*" is a much stronger indicator of relevance than just finding "*blood*" and "*pressure*" scattered within a document. We use Metzler and

[5] `www.pubmed.com`
[6] `www.embase.com`
[7] `www.lemurproject.org`
[8] `http://trec.nist.gov/trec_eval/`

Croft's Markov Random Field model [9] to integrate that. In this model three features are considered: single term features (standard unigram language model features), exact phrase features (words appearing in sequence) and unordered window features (require words to be close together, but not necessarily in an exact sequence order). Features weights are set according to the authors's recommendation. For the query "*cyclosporine* AND *blood pressure*", baseline-2 uses the following Indri query:

```
#combine( cyclosporine
          #weight( 0.8 #combine(blood pressure)
                   0.1 #1(blood pressure)
                   0.1 #uw8(blood pressure) ) )
```

The third baseline uses PECO queries as bag-of-words with the traditional language modeling approach. This model allows us to compare keywords and PECO query search strategies. The idea is to evaluate if these longer queries are able to capture more aspects of the information need withour causing query drift problems. For the PECO query "$[cigarette\ smokers]^P$ $[reduction\ to\ quit]^E$ $[abrupt\ quitting]^C$ $[abstinence]^O$", baseline-3 uses the following Indri query:

```
#combine( cigarette smokers
          reduction to quit
          abrupt quitting
          abstinence )
```

The fourth and last baseline simply assigns a different weight to each PECO element in the query. Weights are set to the values found in Boudin *et al.* [1]. These were determined automatically by cross-validation. For the PECO query given in the previous example, baseline-4 uses the following Indri query:

```
#weight( 0.35 #combine(cigarette smokers)
         0.40 #combine(reduction to quit)
         0.15 #combine(abrupt quitting)
         0.10 #combine(abstinence) )
```

Results are presented in Table 1. As expected, baseline-2 is more precise than baseline-1 but returns less relevant documents. We observe that using PECO structured queries allows to retrieve more relevant documents and, in the case of baseline-4, improves significantly the retrieval effectiveness.

Model	MAP	P@5	P@10	#rel
Baseline-1	0.1288	0.1513	0.1513	5369
Baseline-2	0.1302	0.1735	0.1442	4650
Baseline-3	0.1255	0.1716	0.1359	5433
Baseline-4	0.1371	0.1853†	0.1579	5761

Table 1: Retrieval performance of the four baselines. †: significant to the 0.1% level ($\alpha = 0.001$) using Student' t test.

4. CONCLUSION

We presented the construction of a test collection for clinical IR. From a set of systematic reviews, we have generated 423 queries with relevance data. Relevance judgments were manually collected from the References section containing the citations from which the synthetized results of the review were extracted. We have sanity-checked the usability of our data by running the queries through a language modeling retrieval model and evaluating the results using standard software. We expect the collection to be useful for experimentating clinical IR using the PECO framework, for which there is currently no existing test collection.

The test collection introduced in this paper was named CLIREC (CLinical Information Retrieval Evaluation Collection). It will be available for download, along with the documentation given to the annotators, from:
`http://www-etud.iro.umontreal.ca/~boudinfl/pecodr/`

5. REFERENCES

[1] F. Boudin, J.-Y. Nie, and M. Dawes. Clinical Information Retrieval using Document and PICO Structure. In *Proceedings of the NAACL-HLT conference*, pages 822–830, Los Angeles, California, June 2010.

[2] G. Chung and E. Coiera. Are decision trees a feasible knowledge representation to guide extraction of critical information from randomized controlled trial reports? *BMC Medical Informatics and Decision Making*, 8(1):48, 2008.

[3] D. Demner-Fushman and J. Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[4] K. Friberg Heppin. MedEval - A Swedish Medical Test Collection with Doctors and Patients User Groups. In *Proceedings of the Second Louhi Workshop on Text and Data Mining of Health Documents*, pages 1–7, Los Angeles, California, USA, June 2010.

[5] S. Green, J. P. T. Higgins, P. Alderson, M. Clarke, C. D. Mulrow, and A. D. Oxman. *Cochrane Handbook for Systematic Reviews of Interventions*, chapter 1. The Cochrane Collaboration, 2008.

[6] W. Hersh, C. Buckley, T. Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the SIGIR conference*, pages 192–201, 1994.

[7] X. Huang, J. Lin, and D. Demner-Fushman. Evaluation of PICO as a Knowledge Representation for Clinical Questions. In *AMIA Annual Symposium Proceedings*, pages 359–363, 2006.

[8] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[9] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of the SIGIR conference*, pages 472–479, 2005.

[10] W. Richardson, M. Wilson, J. Nishikawa, and R. S. Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12, 1995.

[11] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[12] E. Voorhees. TREC: Continuing information retrieval's tradition of experimentation. *Communications of the ACM*, 50(11):51–54, 2007.