Evaluating Medical Information Retrieval

Bevan Koopman*
Australian e-Health Research Centre, CSIRO
bevan.koopman@csiro.au

Peter Bruza
Queensland University of Technology
p.bruza@qut.edu.au

Laurianne Sitbon
Queensland University of Technology
sitbon@qut.edu.au

Michael Lawley
Australian e-Health Research Centre, CSIRO
michael.lawley@csiro.au

ABSTRACT

This paper presents a framework for evaluating information retrieval of medical records. We use the BLULab corpus, a large collection of real-world de-identified medical records. The collection has been hand coded by clinical terminologists using the ICD-9 medical classification system. The ICD codes are used to devise queries and relevance judgements for this collection. Results of initial test runs using a baseline IR system show that there is room for improvement in medical information retrieval. Queries and relevance judgements are made available at `http://aehrc.com/med_eval`.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

General Terms

Experimentation, Measurement

Keywords

Medical Information Retrieval, Evaluation

1. INTRODUCTION

With the advent of electronic medical records and the continued increase of medical documents online (e.g. MEDLINE) there is an increasing need for information retrieval systems tailored to searching medical free-text [1]. Searching medical records presents some specific challenges. For example the presence of an organism in a laboratory test might lead a human being to conclude a certain disease, even though this is not stated explicitly. Furthermore, a user formulating a query might specify the active ingredient of a pharmaceutical, whereas the patient record might only state the brand name of a drug containing that ingredient [2]. We require IR systems capable of bridging the 'semantic gap' – overcoming the mismatch between the terms found in documents and the terms used in queries.

Another challenge for medical IR is empirical evaluation. To our knowledge no standardised evaluation framework exists. There is no test collection with associated queries and relevance judgements specific to medical records. Although there are biomedical test collections (e.g. TREC Genomics Track), these differ from medical records in that they focus specifically on identifying genes and associated diseases.

Our contributions to medical information retrieval are: (i) development of an evaluation framework using real world (de-identified) medical records (ii) collection of queries extracted from the medical records using the standard ICD-9 medical code terminology (iii) relevance judgements devised from human classified ICD-9 medical records (iv) results of initial test runs on the collection using a baseline state-of-the-art IR system.

2. MEDICAL CORPUS & HUMAN TAGGING

Our test corpus is the BLULab NLP repository[1], a collection of 81,617 de-identified clinical records from multiple U.S. hospitals during 2007. The collection is available to the community for research purposes. A number of different medical record types are provided, including: History and Physical Exams, Progress Notes, Consultation Reports, Radiology Reports, Emergency Department Reports, Discharge Summaries, Operative Reports, Cardiology Reports.

Each record has been coded by professional clinical terminologists using the ICD-9 classification system. ICD (International Statistical Classification of Diseases and Related Health Problems) is a coding of diseases, symptoms, abnormal findings, complaints, social circumstances and external causes of injury, as classified by the World Health Organisation. These ICD codes are used to extract queries and relevance judgements.

3. EVALUATION ARCHITECTURE

The process for developing a test set of queries and relevance judgements is illustrated in Figure 1. The steps required are:

❶ For each medical record (document) we extract the ICD codes assigned to that record;
❷ Each ICD code is considered an individual query, the

*Also Queensland University of Technology.

Copyright is held by the author/owner(s).
SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

[1]BLULab provided by University of Pittsburgh, available online: `http://nlp.dbmi.pitt.edu/nlprepository.html`

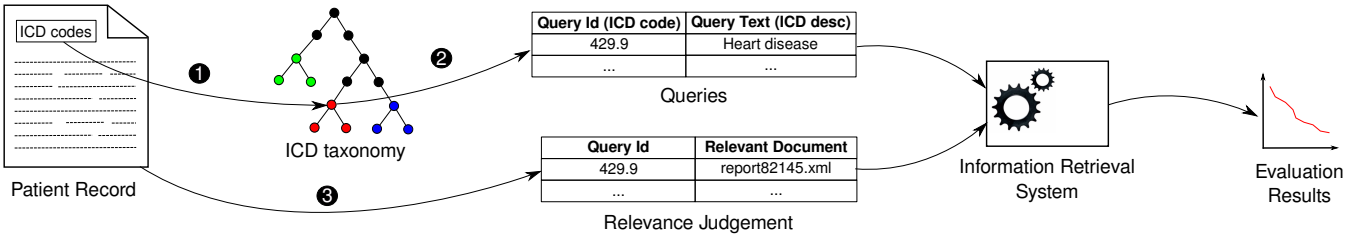

Figure 1: Evaluation architecture for BLULab collection.

Run	Documents used	Docs	Queries	MAP	Precision @5	Precision @10	Precision @20
A	All.	89268	3500	0.0485	0.1249	**0.1085**	**0.0918**
B	Discharge Summaries, History & Physical Exams, Emergency Department Reports. No laboratory based reports.	33671	3434	0.0817	0.1263	0.1072	0.0877
C	Discharge Summaries.	7837	2841	0.1322	0.1234	0.0983	0.0766
D	Discharge Summaries (excluding non-clinical queries).	7837	2420	**0.1445**	**0.1347**	0.1074	0.0834
E	Discharge Summaries, queries mapped to SNOMED CT.	7837	2489	0.1123	0.1075	0.0873	0.0697

Table 1: Experimental retrieval results for the BLULab NLP collection using ICD codes as gold standard.

query id is the ICD code id and the query text is the ICD code description as defined in the ICD taxonomy;

❸ The ICD code and record filename are then added to the relevance judgement file.

Queries length ranged from 1 to 98 words with an average of 18 words. Each query had on average 231 relevance judgements. Average document length was 365 words.

The ICD terminology is primarily used by hospitals for administrative and billing purposes. Consequently, there were initially a number of limitations as a result of both the characteristics of the ICD terminology and the manner in which terminologists assign codes to documents:

Granularity of coding Choice of ICD code is highly dependent on the diseases being described. For example general respiratory diagnoses codes are typically used rather than a more specific code, even though one exists. Conversely, certain high level codes such as "Kidney" (`198.0`) are hardly used as the terminologist would favour the use of a code specific to a disease. This affects the quality of relevance judgements.

ICD term hierarchy ICD codes in the form `XXX.XX` can only be understood in the context of their parent, e.g. "Septicemia" (`038`) has a child "Other" (`038.49`). "Other" obviously cannot be used as the query text for our evaluation so in these cases the query is formed by concatenation of the parent and child, thus forming "Septicemia Other".

4. INITIAL RESULTS

Using the evaluation framework described we provide the results of a number of baseline experiments on the BLULab corpus. Experiments were conducted on the entire corpus and on a number of subsets. The results presented in Table 1 were obtained with the Indri search engine, Porter stemmer and tf-idf term weighting, which performed better in comparison to a state-of-the-art BM25 system.

Run A, using all the documents, resulted in very low performance. This can be attributed in part to the many laboratory reports which contain little or no natural language. Excluding these documents in Run B improved results. Run C used only Discharge Summaries which represent a good overview of the patient encounter, using only these documents again improved performance. Run D excluded ICD codes of type "E" and "V" which are administrative rather than clinical in nature. Run E mapped ICD codes to SNOMED CT, a formal ontology for medical knowledge. This was done to determine whether SNOMED CT had better concept descriptions for use as query text.

In short, our framework provides a meaningful evaluation of medical IR. The low performance prompted some manual review which reaffirmed the challenges in medical IR, specifically the 'semantic gap' that exists between queries and documents. Bridging this gap involves more than matching keywords, it requires inference. The results also show plenty of scope for future work and strong motivation for a semantic search approach to medical information retrieval.

5. CONCLUSIONS

An evaluation framework for medical IR is provided using a human classified corpus of de-identified medical records. Queries and relevance judgements are devised from the human assigned ICD-9 codes. Initial results from test runs using a state-of-art baseline IR system show there is room for improvement and future work. Our queries and relevance judgements are available for other researches in medical information retrieval at `http://aehrc.com/med_eval`.

6. REFERENCES

[1] W. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Verlag, New York, 2009.

[2] C. Patel, J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivasclass. Matching patient records to clinical trials using ontologies. *The Semantic Web*, 4825:816–829, 2007.