Australasian Medical Journal [AMJ office use only]

Towards Semantic Search and Inference in Electronic Medical Records: an approach using Concept-based Information Retrieval

Bevan Koopman [1,2] Peter Bruza [2] Laurianne Sitbon [2] Michael Lawley [1]

1: Australian e-Health Research Centre, CSIRO, Brisbane, Australia

2: Science & Engineering Faculty, Queensland University of Technology, Brisbane, Australia

RESEARCH

Please cite this paper as: [Koopman, B., Bruza, P., Sitbon, L., Lawley, M. Towards Semantic Search and Inference in Electronic Medical Records: an approach using Concept-based Information Retrieval, other details to be added by AMJ]

Corresponding Author:
Bevan Koopman
Lvl 5, UQ Health Sciences Building 901/16
Royal Brisbane and Women's Hospital
Herston 4029
Queensland, AUSTRALIA
bevan.koopman@csiro.au

Abstract

Background
This paper presents a novel approach to searching electronic medical records that is based on concept matching rather than keyword matching.
Aims
The concept-based approach is intended to overcome specific challenges we identified in searching medical records.
Method
Queries and documents were transformed from their term-based originals into medical concepts as defined by the SNOMED-CT ontology.
Results
Evaluation on a real-world collection of medical records showed our concept-based approach outperformed a keyword baseline by 25% in Mean Average Precision.
Conclusion
The concept-based approach provides a framework for further development of inference based search systems for dealing with medical data.

Word count: 3091
Figures and Tables: 6

Key Words
Electronic medical records, Information retrieval, Semantic search and inference, Health informatics.

What this study adds:

1. Searching medical records presents some specific challenges that require tailored information retrieval (IR) systems.
2. It was found that a concept-based (rather than term-based) information retrieval system improved search accuracy.
3. The concept-based approach provides a framework for further development of inference based search systems for dealing with medical records.

Background

Searching medical records presents some specific challenges for information retrieval (IR) systems. Vocabulary mismatch – where relevant documents to a user's query may actually contain little or no shared terms – can hamper the performance of keyword-based retrieval. For example, a user searching for "high blood pressure" would want to retrieve documents mentioning "hypertension". Beyond vocabulary mismatch, certain queries require *inference* to determine relevant documents, for example the presence of a certain organism in a laboratory report denoting a certain disease, even though the disease it not stated explicitly [1]. Searching medical records requires an information retrieval system capable of overcoming the "semantic gap" – the mismatch between the terms found in documents and those in queries.

Our approach to the semantic gap problem is a concept-based approach that uses medical domain knowledge from the SNOMED-CT ontology [2]. Queries and documents were transformed from their original terms to SNOMED-CT concepts; retrieval was then done by matching concepts. The model is therefore less dependent on the specific terms used. The paper makes the following contributions: (1) an analysis of the types of semantic gap problem that exist when searching medical records, including the type of inference required to

handle each; (2) a concept-based IR model that addresses some of these problems while providing the foundation for further development; (3) empirical evaluation showing our concept-based system outperformed an equivalent keyword baseline; (4) analysis of how our system differs from a keyword baseline, specifically when dealing with hard queries.

Related Work

Related work is in two areas: (i) concept-based IR, that is representing queries and documents as concepts rather than terms; and (ii) domain knowledge, specifically the SNOMED-CT ontology.

Concept-based IR

Broadly, concept-based IR aims to make use of external knowledge sources (such as thesauri or ontologies) to provide additional background knowledge and context that may not be explicit in a document collection and user's queries. Early approaches by Voorhees [3] used general lexical thesauri such as WordNet for the purposes of query expansion. WordNet is a large general English language ontology. Nouns, verbs adjectives and adverbs are grouped into cognitive synonyms each expressing a distinct concept [4]. Ravindran & Gauch [5] used the Open Directory to create a concept index for query disambiguation.

In the area of biomedical information retrieval there have been a number of concept-based approaches. Aronson & Rindflesch [6] used the UMLS medical ontology for query expansion, while Liu & Chu [7] improve on standard query expansion with concept-based scenario-specific query expansion. More advanced approaches have gone beyond query expansion and use medical ontologies in both the indexing and retrieval process. For example Zheng et al. successfully used MeSH headings to build a concept-document matrix to facilitate biomedical document search [8]. Significant improvements using concept-based IR are achieved in genomic information retrieval. Zhou et al. [9] developed a concept matching algorithm that utilised both the UMLS ontology and MeSH headings; their system significantly outperformed keyword-based systems.

Performance in concept-based IR is highly dependent on the specific domain model or ontology used. General applications (those that utilise WordNet or Open Directory) struggle to outperform keyword-based systems [3, 5]. However, biomedical applications (which use domain specific ontologies) demonstrate the most improvements [9, 7]. For this reason we propose concept-based IR for searching electronic medical records.

Medical domain knowledge (SNOMED-CT)

The choice of domain model has been highlighted as an important consideration in concept-based IR. UMLS and MeSH are two domain models most often used in biomedical applications [8, 7, 9]. Recently there has been strong emphasis on the development of more formal, machine readable representations of medical knowledge, this has led to the develop of the SNOMED-CT ontology. SNOMED-CT is a medical terminology covering a large range of medical knowledge, including: disorder, procedures, organisms, body structure and pharmaceuticals [2]. Concepts are organised in an inheritance hierarchy and may be defined by relations to other concepts. For example the concept *Viral pneumonia* has a parent *Infectious pneumonia*. *Viral pneumonia* has a relationship *Causative agent* connecting it to the *Virus* concept.

SNOMED-CT contains approximately 390,000 concepts and 1.4 million relationships. SNOMED-CT's wide coverage and non-application specific focus was the reason it was chosen as the domain knowledge model for our concept-based IR system.

Requirements for semantic search and inference in medical records

We have introduced the "semantic gap" problem and stated that certain queries require *inference* rather than

Table 1: Classification of semantic gap queries found in medical records, including type of inference required to handle each.

Semantic Gap Query	**Example**	**Inference Required**
1. Keyword mismatch Synonyms, formal vs. colloquial terms:	*Hypertension ≈ high blood pressure*	Associational
2. Specialisation / generalisation: Hyponyms/hypernyms, queries use general terms, medical records more specific	*Morphine → Opiate*	Deductive
3. Implied: Presence of certain term in medical records implies relevance to query	*Chemotherapy → Cancer*	Deductive
4. Indirect relations: Causative and/or correlated	*Hepatitis B* causes liver damage, documents containing *Hepatitis B* sometimes mention the *HNF4* gene, therefore a query for "HNF4 liver function" should return the documents mentioning *Hepatitis B*	Abductive

keyword matching. To better understand the requirements for a semantic search system we have categorised the specific types of queries involved in searching medical records and the form of inference required to deal with each. These are provided in Table 1.

From these examples it is clear that bridging the semantic gap requires matching at the conceptual level and requires inference. At present our concept-based approach aims to deal with the first two types of query: keyword mismatch and specialisation / generalisation. However, it also provides a platform for further development on the more challenging inferencing problems highlighted. We now present details of our concept-based information retrieval model.

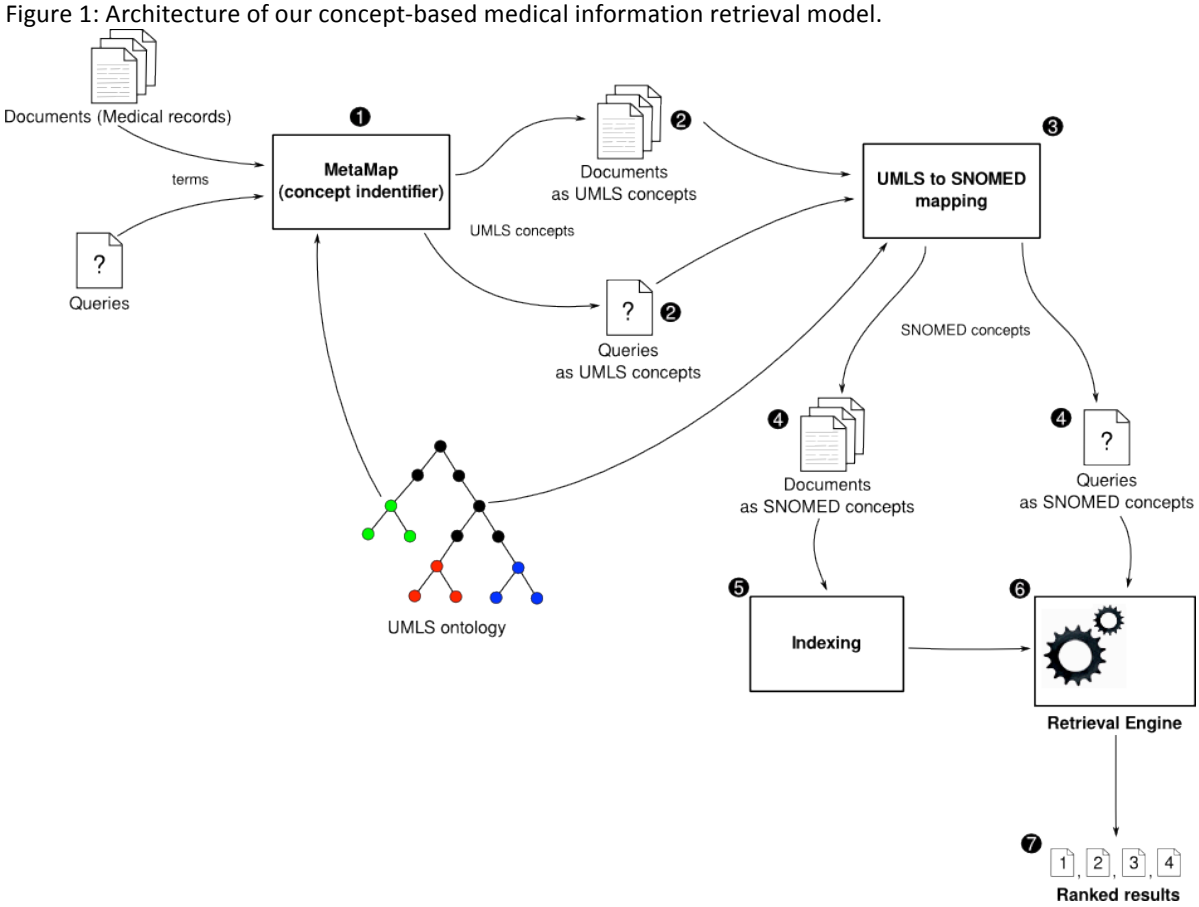
Method – Concept-based Information Retrieval

Our concept-based system has two main parts: a SNOMED-CT concept extractor from free-text; and the indexing and retrieval components.

For concept extraction we utilised MetaMap [10], the natural language processing system developed by the U.S. National Library of Medicine. MetaMap identifies UMLS concepts in biomedical text and is widely adopted in medical NLP and IR [11, 7]. Using MetaMap, queries and documents were represented as a bag-of-concepts rather than their original bag-of-words representation. For example the text "*vascular dementia*" can be translated to the UMLS concept "*C0011269*". The translation process from terms to concepts is described in Figure 1 and consists of the following steps:

1. MetaMap identified the UMLS concepts in both medical records and queries.[1]
2. Documents and queries no longer contain their original terms, instead they were represented as UMLS concepts ids.
3. Using the UMLS Metathesaurus, UMLS concepts were mapped to their SNOMED-CT equivalents. There is often a one-to-many mapping from UMLS to SNOMED-CT, in these cases all SNOMED CT concepts were included.
4. Queries and documents were then represented as SNOMED-CT concept ids.
5. Documents were indexed using a standard information retrieval engine and their new concept-based representation.
6. The queries (represented as SNOMED-CT concept ids) were issued to the retrieval engine.
7. A ranked list of document results was returned and compared to relevance judgements to determine retrieval performance.

Figure 1: Architecture of our concept-based medical information retrieval model.



Experimental Design

This section describes the experimental setup, including the test collection, associated queries and evaluation metrics.

[1] MetaMap suggests a number of candidate concepts and finally a best fit concept. We included the best fit and all candidate concepts which produced better results than only including the best fit concepts.

A challenge for medical IR is empirical evaluation. To our knowledge no standardised test collection with associated queries and relevance judgements exists specific to medical records. Although there are test collections for medical journal articles (e.g. the OHSUMED collection of MEDLINE articles), these differ from medical records in that they focus specifically on well written journal articles. In previous work, we have developed a test collection specific for searching medical records [12]. The collection contains: (i) 81,617 de-identified clinical records from multiple U.S. hospitals[2]; (ii) 3249 clinical queries; (iii) relevance judgements indicating which documents are relevant to each clinical query.

For the purposes of this study we selected a subset of 54 queries. The rational for this was to obtain queries that contained (i) a significant number of relevance judgements; (ii) sufficient granularity, ranging from general queries to very specific queries; (iii) inter query dependence, an issue identified previously with some queries [12]; and (iv) examples of the semantic gap characteristics we outlined previously (Table 1). We ran the queries against two retrieval systems: a standard keyword based retrieval engine, this constitutes a baseline for comparison; and our concept-based retrieval system described in the previous section. Implementation of both the concept-based and keyword-based baseline systems was done using the Indri Lemur search engine[3], Porter stemmer and tf-idf weighting.

We evaluated the effectiveness of the retrieval systems using two widely adopted IR performance metrics [13]: (i) Mean average precision (MAP), which combines precision and recall while assigning higher importance to top ranked relevant documents; (ii) Precision at 10 (Prec@10), which measures the number of relevant documents in the top 10 results. Both measures range between 0.0 (worst, no relevant documents) and 1.0 (best, all relevant documents).

Results and Analysis

This section reports on the results of experiments evaluating our concept-based IR approach. Table 2 presents a comparison of our system against the keyword baseline. The concept-based approach outperforms the keyword baseline system by 25% in Mean Average Precision (MAP).

Table 2: Comparison of our concept-based system against the keyword baseline. ‡ Indicates statistical significance (pairwise t-test, $p < 0.01$).

System	MAP (%Δ)	Prec@10 (%Δ)
Keyword baseline	0.2012	0.2963
Concept-based	0.2532 (+25%) ‡	0.3462 (+17%)

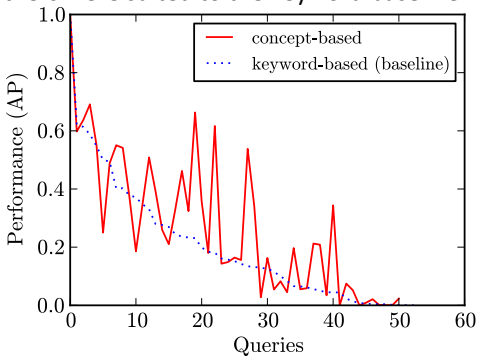
[2] The records are part of the BLULab NLP repository provided by the University of Pittsburgh at http://nlp.dbmi.pitt.edu/nlprepository.html.

[3] The Lemur Project http://lemurproject.org.

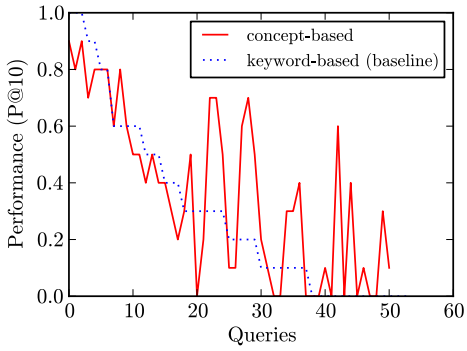
Per-query Analysis

The figures in Table 2 are a good overall comparison of the two systems but provide little understanding of how and why each system differs. We therefore conducted per-query analysis to understand where each system is performing well. The plots in Figure 2 present the performance (y-axis) of each of the 54 queries (x-axis), queries are ordered by decreasing performance of the baseline system.

Figure 2: Per-query comparison of concept-based and keyword-baseline systems. Queries ordered by decreasing performance of baseline system. Results show some queries performed better using concept-based retrieval while others were suited to the keyword baseline.



(a) Average precision

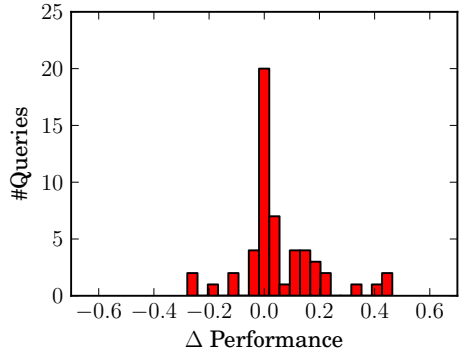


(b) Precision @ 10

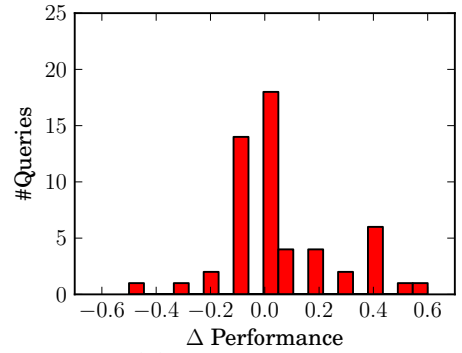
We observe that certain queries performed better using our concept-based system while others were suited to a keyword-based system. It is important to understand whether performance gains were a result of substantial improvements in a small set of queries or small gains across many queries. The former may provide good overall results but reduces the usability of the approach in practical terms as only a few queries would demonstrate improved results. On the contrary, our system exhibited small gains across a large number of queries as shown by the histograms presented in Figure 3. Both histograms report the change in performance (x-axis) compared to the baseline system, positive values reflect an improvement in performance, while negative values indicate cases where the baseline system performed better. The y-axis indicates the number of queries exhibiting that performance change. The histograms show that our concept-based system made small improvements

in a number of queries, rather than large gains (or losses) on a few.

Figure 3: Histogram showing change in performance using concept-based system. We observe that the concept-based system made small performance gains for a large number of queries. Significant changes in performance were only found for few queries.



(a) Average precision



(b) Precision @ 10

Hard vs. Easy Queries

The hypothesis that motivates our concept-based approach is it helped improve more challenging medical queries. We therefore provide some further analysis on how the concept-based system performed on hard queries (those showing poor performance in the baseline system) vs. easy queries. Our method was as follows, the 54 queries were sorted according to their performance in the keyword baseline system. They were then divided into two subsets: 27 best performing queries and 27 worst performing queries. Each query subset was evaluated on both the keyword and concept-based systems, results are presented in Table 3.

Table 3: Comparison of concept-based and keyword baseline systems for hard and easy queries. ‡ Indicates statistical significance (pairwise t-test, $p < 0.01$).

Queries	System	MAP (%Δ)	Prec@10 (%Δ)
Hard	Keyword baseline	0.0489	0.1037
	Concept-based	0.1000 (+104%) ‡	0.1667 (+60%)
Easy	Keyword baseline	0.3535	0.4889
	Concept-based	0.4064 (+15%)	0.5259 (+7%)

The results support the hypothesis that concept-based IR generally performed better on more difficult queries, with a 104% improvement over the baseline. Importantly, this was not at the expense of easy queries.

Discussion

Overall, the concept-based approach exhibited an improvement over a keyword baseline. Results were heavily dependent on the quality of concept extraction provided by the MetaMap system. MetaMap only identifies UMLS concepts, which were then mapped to SNOMED-CT concepts. The rational for converting to SNOMED-CT was its formal representation that provides scope for future inference techniques. Experiments using UMLS concepts showed comparable performance. However, mapping between terminologies may result in a loss in meaning from the original query or document. Certain UMLS concepts have no equivalent in SNOMED-CT. Such cases were found in the two worst performing queries in our experiments, these were query *454.9* (*asymptomatic varicose veins*) and *038.11*, (*methicillin susceptible staphylococcus aureus septicemia*). Advances in medical NLP, and the increasing popularity of SNOMED-CT, are likely to yield further improvements to tools such as MetaMap, for example direct SNOMED-CT concept identification that avoids the mapping via UMLS, this will avoid the mapping problem and, we conjecture, should improve our concept-based retrieval system.

The queries that performed well using our concept-based approach were often characterised as having a number of possible variants in their keyword form. For example, the query *530.81* (*esophageal reflux*) which mapped to the SNOMED-CT concepts:

- *235595009* (*Gastroesophageal reflux disease*);
- *196600005* (*Acid reflux &/or oesophagitis*);
- *47268002* (*Reflux*); and
- *249496004* (*Esophageal reflux finding*).

In the keyword-based system a query for *esophageal reflux* was unlikely to return documents that contain *oesophagitis*[4]. However, in the concept-based approach *oesophagitis* was represented in the query as part of concept *196600005*. The average precision for this query improved from 0.1285 to 0.3414. Another example was query *042* (*human immunodeficiency virus*) – relevent documents contained the abbreviations *HIV* or *AIDS* but did not explicitly mention *human immunodeficiency virus* (average precision increased from 0.2332 to 0.4622 for this query).

Future work

Our current system represents queries and documents as SNOMED-CT concepts but does not make use of the additional information provided by the relationships between concepts. Some initial experimentation on using these relationships for query expansions proved difficult –

[4] Inflammation of the esophagus caused by reflux..

certain queries showed significant improvement, while others had significant degradation in performance. A more targeted approach that takes into account the semantic type (e.g. disease, treatment, symptom) of the specific query concept is required (this approach has been successful in other applications [7]). The use of inter-concept relationships is the next step towards a system that supports the type of inference capabilities required to deal with the complex medical queries we have already outlined.

Conclusion

We have presented an approach to searching electronic medical records that is based on concept matching rather than keyword matching. Queries and documents were transformed from their term-based originals into medical concepts as defined by the SNOMED-CT ontology. Evaluation on a real-world collection of medical records showed our concept-based approach outperformed a keyword baseline by 25% in MAP. In addition, the concept-based approach made significant improvements on hard queries. We have provided an analysis and classification of the type of queries used when searching medical records, emphasising that some require specific types of inference. Our concept-based approach provides a framework for further development into inference based search systems for dealing with medical data.

References

1. Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A, Kershenbaum A, et al. Matching patient records to clinical trials using ontologies. The Semantic Web. 2007;4825:816–829.
2. Spackman KA, Campbell KE. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. In: Proceedings of the AMIA Symposium. Orlando, FL; 1998. p. 201-211.
3. Voorhees EM. Query expansion using lexical-semantic relations. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval. Dublin, Ireland: ACM; 1994. p. 61–69.
4. Fellbaum C. WordNet: An electronic lexical database. The MIT press; 1998.
5. Ravindran D, Gauch S. Exploiting hierarchical relationships in conceptual search. In: Proceedings of the 13th annual international ACM CIKM conference on in-formation and knowledge management. ACM; 2004. p. 238–239.
6. Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. Proceedings of American Medical Informatics Association. 1997 Jan; p. 485–9.
7. Liu Z, Chu WW. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Information Retrieval. 2007 Jan;10(2):173–202.
8. Zheng HT, Borchert C, Jiang Y. A knowledge-driven approach to biomedical document conceptualization. Artificial Intelligence in Medicine. 2010;49(2):67–78.
9. Zhou W, Yu C, Smalheiser N, Torvik V, Hong J. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. New York, USA: ACM; 2007. p. 655–662.
10. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17(3):229–236.
11. Hersh W. Information retrieval: a health and biomedical perspective. 3rd ed. New York: Springer Verlag; 2009.
12. Koopman B, Bruza P, Sitbon L, Lawley M. Evaluating medical information retrieval. In: Proceedings of the 34th annual international ACM SIGIR conference on research and development in information retrieval. Beijing, China: ACM; 2011. p. 1139–1140.
13. Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. New York: ACM Press; 1999.

PEER REVIEW

Not commissioned. Externally peer reviewed

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

ETHICS COMMITTEE APPROVAL

BLULab data collection obtained with ethics approval from CSIRO Food and Nutritional Sciences Human Research Low Risk Review Panel – Proposal #LR13/2010.