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

Bevan Koopman[1,2⋆], Peter Bruza[2], Lauriane Sitbon[2], and Michael Lawley[1]

[1] Australian e-Health Research Centre
CSIRO, Brisbane, Australia
{bevan.koopman,michael.lawley}@csiro.au
[2] Information Systems Discipline, Faculty of Science and Technology
Queensland University of Technology, Brisbane, Australia
{p.bruza,sitbon}@qut.edu.au

Abstract. This paper presents a novel approach to searching electronic medical records that is based on concept matching rather than keyword matching. The concept-based approach is intended to overcome specific challenges we identify in searching medical records. Queries and documents are 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 shows our concept-based approach outperforms a keyword baseline by 30% in Mean Average Precision. The concept-based approach provides a framework for further development of inference based search systems for dealing with medical data.

Keywords: Electronic medical records, Information retrieval, Semantic search and inference, Health informatics.

1 Introduction

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'[3]. 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 [8]. 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.

⋆ Corresponding author.
[3] Formal synonym for high blood pressure.

Our approach to the semantic gap problem is a concept-based information retrieval approach that uses medical domain knowledge from the SNOMED-CT ontology [11]. Queries and documents are transformed from their original terms to SNOMED-CT concepts, retrieval is 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 information retrieval model that addresses some of these problems while providing the foundation for further development; (3) empirical evaluation showing our concept-based system outperforms an equivalent keyword baseline; (4) analysis of how our system differs from a keyword baseline, specifically when dealing with hard queries.

2 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) medical domain knowledge, and specifically the SNOMED-CT ontology.

2.1 Concept-based IR

Broadly, concept-based information retrieval 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 [12] used general lexical thesauri such as WordNet[4] for the purposes of query expansion. Ravindran & Gauch [9] 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 [2] 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 [13]. Significant improvements using concept-based IR are in the area of genomic information retrieval. Zhou et al. [14] 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 information retrieval 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 [12, 9]. However, biomedical applications — which use domain specific ontologies — demonstrate the most improvements [14, 7]. For this reason we propose concept-based information retrieval for searching electronic medical records.

[4] WordNet is a large general English language ontology. Nouns, verbs adjectives and adverbs are grouped into cognitive synonyms each expressing a distinct concept [4].

2.2 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 biomedical in applications [13, 7, 14]. 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 [10]. Concepts are organised in an inheritance hierarchy and may be defined by relations to other concepts. An example is shown in Figure 1. The concept *Viral pneumonia* has a parent *Infectious pneumonia*. *Viral pneumonia* has a relationship *Causative agent* connecting it to the *Virus* concept.

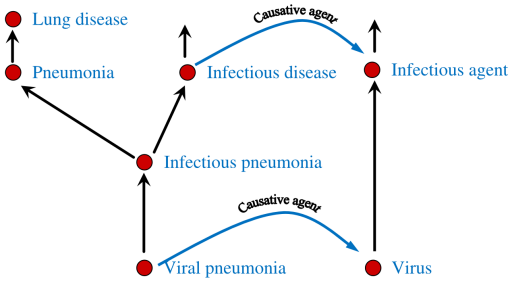


Fig. 1. SNOMED-CT ontology example showing concept hierarchy for *Viral pneumonia* with associated relationship to *Virus* [10].

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

3 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 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

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* $\approx$ *high blood pressure*	Associational
2. Specialisation / generalisation:		
Hyponyms/hypernyms, queries use general terms, medical records more specific	*Morphine* $\rightarrow$ *Opiate*	Deductive
3. Implied:		
Presence of certain term in medical records implies relevance to query	*Chemotherapy* $\rightarrow$ *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* [14]	Abductive

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.

4 Concept-based information retrieval model

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

For concept extraction we utilise the natural language processing system MetaMap [1] 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 [5, 7]. Using MetaMap, queries and documents can be represented as sets of concepts rather than their original term-based representation. For example the text '`vascular dementia`' shall be translated to the UMLS concept `C0011269`. The translation process from terms to concepts is described in Figure 2 and consists of the following steps:

❶ MetaMap identifies the UMLS concepts in both medical records and queries.
❷ Documents and queries no longer contain their original terms, instead they are represented as UMLS concepts ids.

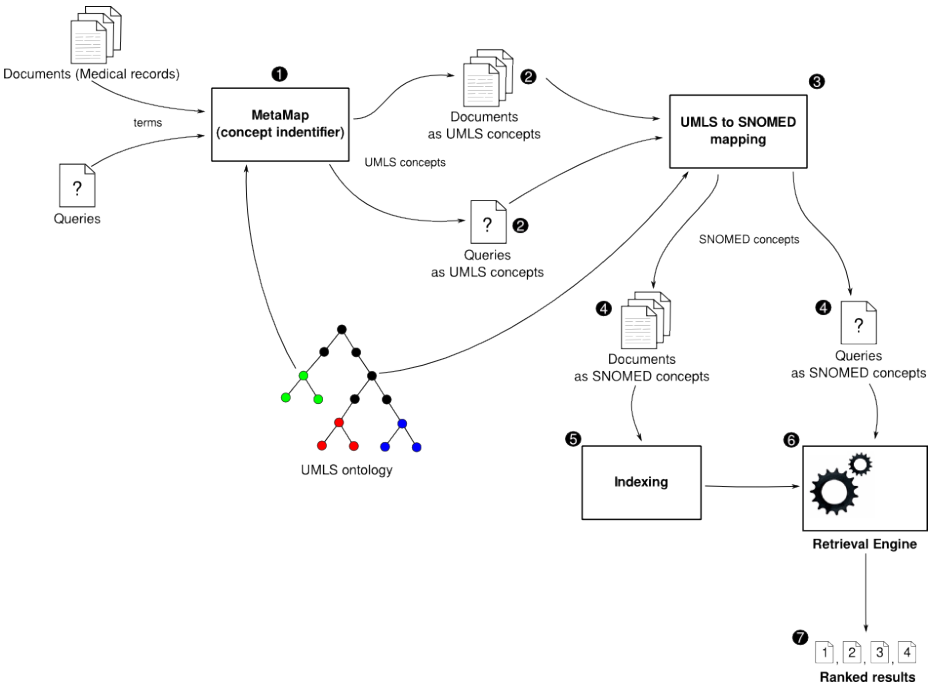


Fig. 2. Architecture of our concept-based medical information retrieval model.

❸ Using the UMLS Metathesaurus, UMLS concepts are 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 are included.
❹ Queries and documents are now represented as SNOMED-CT concept ids.
❺ Documents are indexed using a standard information retrieval engine and their new concept-based representation.
❻ The queries (represented as SNOMED-CT concept ids) are issued to the retrieval engine.
❼ A ranked list of document results is returned and can be compared to relevance judgements to determine retrieval performance.

5 Experimental design

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

A challenge for medical information retrieval is empirical evaluation. In previous work, we have developed a test collection specific for searching medical records [6]. The collection contains: (i) 81,617 de-identified clinical records from

multiple U.S. hospitals[5]; (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, ensuring that they had a significant number of relevance judgements, sufficient granularity and no inter query dependence [6]. 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[6], Porter stemmer and tf-idf weighting.

We evaluated the effectiveness of the retrieval systems using two widely adopted information retrieval performance metrics [3]: (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).

6 Results & 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 ∼30% 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	0.2055	0.3019
Concepts-based	0.2681 (+30.46%)‡	0.3667 (+21.46%)

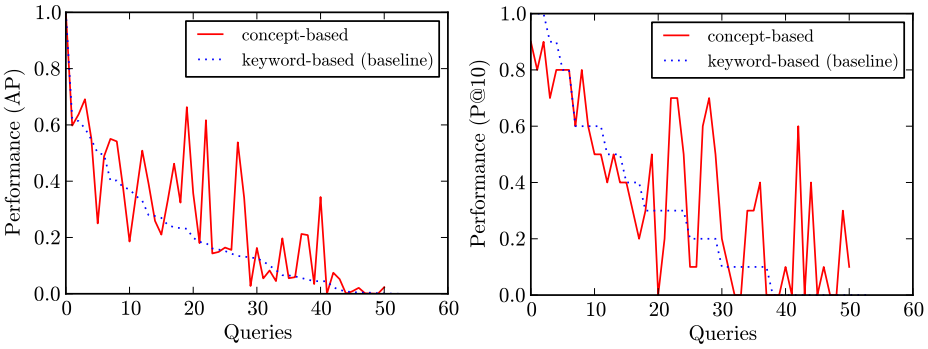
6.1 Per-query analysis

The above figures are a good overall comparison of the two systems but provide little understanding on how and why each system differs. We therefore conducted some per-query analysis to understand where each system is performing well. The plots in Figure 3 present the performance (y-axis) of each of the 54 queries

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

[6] The Lemur Project `http://lemurproject.org`

(x-axis), queries are ordered according to decreasing performance of the baseline system.



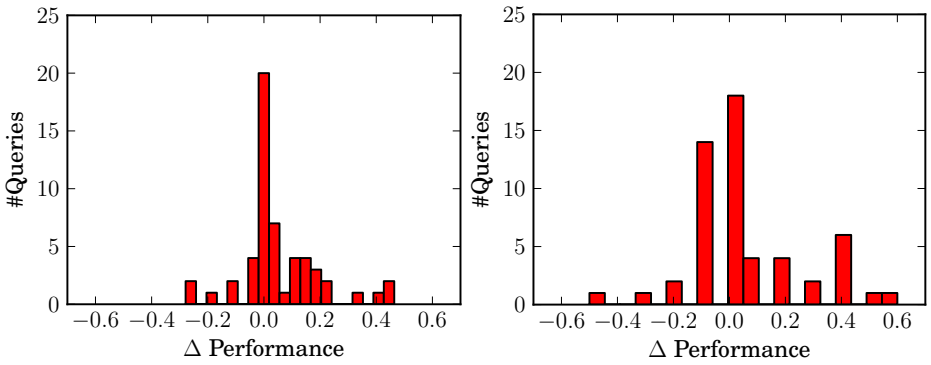
(a) Average precision (b) Precision @ 10

Fig. 3. Per-query comparison of concept-based and keyword baseline systems. Queries ordered in decreasing performance according to the baseline system. Results show some queries perform better using concept-based retrieval while others are suited to the keyword baseline.

We observe that certain queries perform better using our concept-based system while others are suited to a keyword-based system. It is important to understand whether performance gains are 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 few queries would demonstrate improved results. On the contrary, our system exhibits small gains across a large number of queries as shown by the histograms presented in Figure 4. 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 makes small improvements in a number of queries rather than large gains (or losses) on a few.

6.2 Hard vs. easy queries

The hypothesis that motivates our concept-based approach is it helps improve more challenging medical queries. We therefore provide some further analysis on how the concept-based system performs on hard queries (those showing poor performance in the baseline system) vs. easy queries. Our method is as follows, the 54 queries are sorted according to their performance in the keyword baseline system. They are divided into two subsets: 27 best performing queries and 27



(a) Average precision (b) Precision @ 10

Fig. 4. Histogram showing change in performance using concept-based system. We observe that the concept-based system makes small performance gains for a large number of queries. Significant changes in performance are only found for few queries.

worst performing queries. Table 3 presents the results for each hard and easy query subset on both the keyword baseline and concept-based system.

Table 3. Comparison of concept-based and keyword baseline systems for hard and easy queries. † Indiates statistical significance (pairwise t-test, $p < 0.05$).

Queries	**System**	**MAP (%Δ)**	**Prec@10 (%Δ)**
Hard	Keyword-based baseline	0.0489	0.1037
	Concept-based	0.1080 (+120.86%)†	0.1800 (+73.58%)
Easy	Keyword-based baseline	0.3535	0.4889
	Concept-based	0.4221 (+19.41%)	0.5462 (+11.72%)

The results support the hypothesis that concept-based information retrieval generally performs better on more difficult queries, with a 120% improvement over the baseline. Importantly, this is not at the expense of easy queries.

7 Discussion

Overall, the concept-based approach exhibits an improvement over a keyword baseline. Results are heavily dependent on the quality of concept extraction provided by the MetaMap system. MetaMap only identifies UMLS concepts, which are then mapped to SNOMED-CT concepts. 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 in turn 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* is unlikely to return documents that contain *oesophagitis*[7]. However, in the concept-based approach *oesophagitis* is represented in the query as part of concept `196600005`. The average precision for this query improved from 0.1285 to 0.3414. Another example is query `042` (*human immunodeficiency virus*) — relevent documents contained *HIV* or *AIDS* but did not mention *human immunodeficiency virus* (average precision increased from 0.2332 to 0.4622 for this query).

7.1 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 — 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 these relationships is the next step towards a system that supports the type of inferencing capabilities required to deal with the complex medical queries we have already outlined.

8 Conclusion

We have presented an approach to searching electronic medical records that is based on concept matching rather than keyword matching. Queries and documents are transformed from their term-based originals into medical concepts as defined by the SNOMED-CT ontology. Evaluation on a real-world collection of

[7] Inflammation of the esophagus caused by reflux.

medical records shows our concept-based approach outperforms a keyword baseline by 30% in MAP. In addition, the concept-based approach made significant improvement 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 inferencing based search systems for dealing with medical data.

References

[1] Aronson, A. R., and Lang, F.-M. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association 17*, 3 (2010), 229–236.

[2] Aronson, A. R., and Rindflesch, T. C. Query expansion using the UMLS Metathesaurus. *Proceedings of American Medical Informatics Association* (Jan. 1997), 485–9.

[3] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern information retrieval*. ACM Press, New York, 1999.

[4] Fellbaum, C. *WordNet: An electronic lexical database*. The MIT press, 1998.

[5] Hersh, W. *Information retrieval: a health and biomedical perspective*, 3rd ed. Springer Verlag, New York, 2009.

[6] Koopman, B., Bruza, P., Sitbon, L., and Lawley, M. Evaluating medical information retrieval. In *Proceedings of the 34st annual international ACM SIGIR conference on research and development in information retrieval* (Beijing, China, July 2011), ACM, pp. 1139–1140.

[7] Liu, Z., and Chu, W. W. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval 10*, 2 (Jan. 2007), 173–202.

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

[9] Ravindran, D., and Gauch, S. Exploiting hierarchical relationships in conceptual search. In *Proceedings of the 13th annual international ACM CIKM conference on information and knowledge management* (2004), ACM, ACM, pp. 238–239.

[10] Spackman, K. SNOMED Clinical Terms Basics. IHTSDO presentation, Aug. 2008.

[11] Spackman, K. A., and Campbell, K. E. Compositional concept representation using SNOMED: towards further convergence of clinical terminologies. In *Proceedings of the AMIA Symposium* (Orlando, FL, 1998), American Medical Informatics Association.

[12] Voorhees, E. M. 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, 1994), ACM, pp. 61–69.

[13] Zheng, H.-T., Borchert, C., and Jiang, Y. A knowledge-driven approach to biomedical document conceptualization. *Artificial Intelligence in Medicine 49*, 2 (2010), 67–78.

[14] Zhou, W., Yu, C., Smalheiser, N., Torvik, V., and 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, 2007), ACM, pp. 655–662.