AEHRC & QUT at TREC 2011 Medical Track: a concept-based information retrieval approach

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

[1] Australian e-Health Research Centre, CSIRO

[2] Information Systems Disipline, Queensland University of Technology

Brisbane, Australia

Abstract

The Australian e-Health Research Centre and Queensland University of Technology recently participated in the TREC 2011 Medical Records Track. This paper reports on our methods, results and experience using a concept-based information retrieval approach. Our 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. Results show our concept-based approach performed above the median in all three performance metrics: bref (+12%), R-prec (+18%) and Prec@10 (+6%).

1 Introduction

The Australian e-Health Research Centre (AEHRC) is a multi-disciplinary research facility applying information and communication technology to improve health services and clinical treatment. The Health Data Semantic group aims to improve access for health data by combining statistical approaches in information retrieval and natural language processing with the formal semantics of the SNOMED CT medical ontology. Hybrid approaches using symbolic and statistical approaches are becoming increasing common [2].

Our system used for the TREC Medical Records Track is concept-based information retrieval using medical domain knowledge provided by the SNOMED CT ontology. In concept-based IR both documents and queries are represented using semantic concepts rather than keywords, retrieval is performed within this concept space. Using high-level concepts makes the retrieval model less dependent on the specific terms being used. Queries and documents are transformed from their original terms to SNOMED-CT concepts, retrieval is then done by matching concepts. Concept-based approaches have previously demonstrated excellent results — Zhou et al. [9] concept-based system (using concept from UMLS ontology and MeSH headings) was the top performing at the TREC Geonomics Track.

We received the track data only a few days before submission were due, as a result our system was used in its generic form as was not adapted to the specific corpus or topics for this years Medical Records Track. However we did develop the system using a subset of the BLULab NLP repository — the same collection from which the Medical Records Track data collection was taken.

2 Methods — concept-based information retrieval

This section describes our methods and the design of our concept-based system. We first report on our approach to treating patient visits as the unit of retrieval. We then describe the two main

*Correspondence to bevan.koopman@csiro.au

parts of our system: extracting SNOMED-CT concepts from free-text; and indexing and retrieval components.

2.1 Documents as visits

The guidelines for the Medical Records Track stated that the unit of retrieval should a single patient visit. A visit is a single admission for a single patient — if the same patient is admitted on two different occasions these will be viewed as two separate visits. Our approach was to treat individual reports as sub-documents and compile them together with all the other reports pertaining to a single patient admission into a single larger document. The unit of retrieval is then a 'patient visit' rather than individual medical reports. As all reports for a single visit are concatenated together we make no distinction as to the different reports type — radiology, discharge summary, etc.

2.2 Concept identification using MetaMap

In our system all queries and documents are converted from the original term-based representation into medical concepts. For this purpose we used MetaMap, a system developed by the U.S. National Library of Medicine [1]. It has been widely adopted in medical NLP [6] and medical IR [4, 8, 3, 5]. Comparisons with human subjects have shown that MetaMap is effective in concept identification tasks (84% precision, 70% recall) [7].

An example output of the MetaMap system using the input string '`heart attack`' is shown in Figure 1.

```
|: heart attack
|:
Established connection to Tagger Server on localhost.
Processing 00000000.tx.1: heart attack

Phrase: "heart attack"
Meta Candidates (8):  ❶
  1000 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
  861 C0018787:Heart [Body Part, Organ, or Organ Component]
  861 C0277793:Attack, NOS (Onset of illness) [Finding]
  861 C0699795:Attack (Attack device) [Medical Device]
  861 C1261512:attack (Attack behavior) [Social Behavior]
  861 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
  861 C1304680:Attack (Observation of attack) [Finding]
  827 C0004063:Attacked (Assault) [Injury or Poisoning]
Meta Mapping (1000):  ❷
  1000 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
```

Figure 1: MetaMap output for `heart attack`.

MetaMap first analyses the input string and produces a ranked list of possible matching candidate concepts (shown in Fig 1❶). From this list of candidates the system selects the highest ranking candidate as *Heart attack (Myocardial Infarction)* (`C0027051`) (shown in Fig 1❷). In our experiments we found including the candidate concepts from Fig 1❶ actually had a positive effect on retrieval.

The advantage of using concepts (rather than just terms) is that different terms with the same meaning are mapped to the same concept — for example the input text '`Myocardial Infarction`' and '`heart attack`' will both map to the UMLS concept `C0027051`. Conversion to concepts aims to overcome some of the vocabulary mismatch that exists in medical text.

2.3 Concept-based architecture

We represented both documents and queries not as term-based vectors but as concept-based vectors. The overall process to translate from terms to concepts is illustrated in Figure 2. The steps required are:

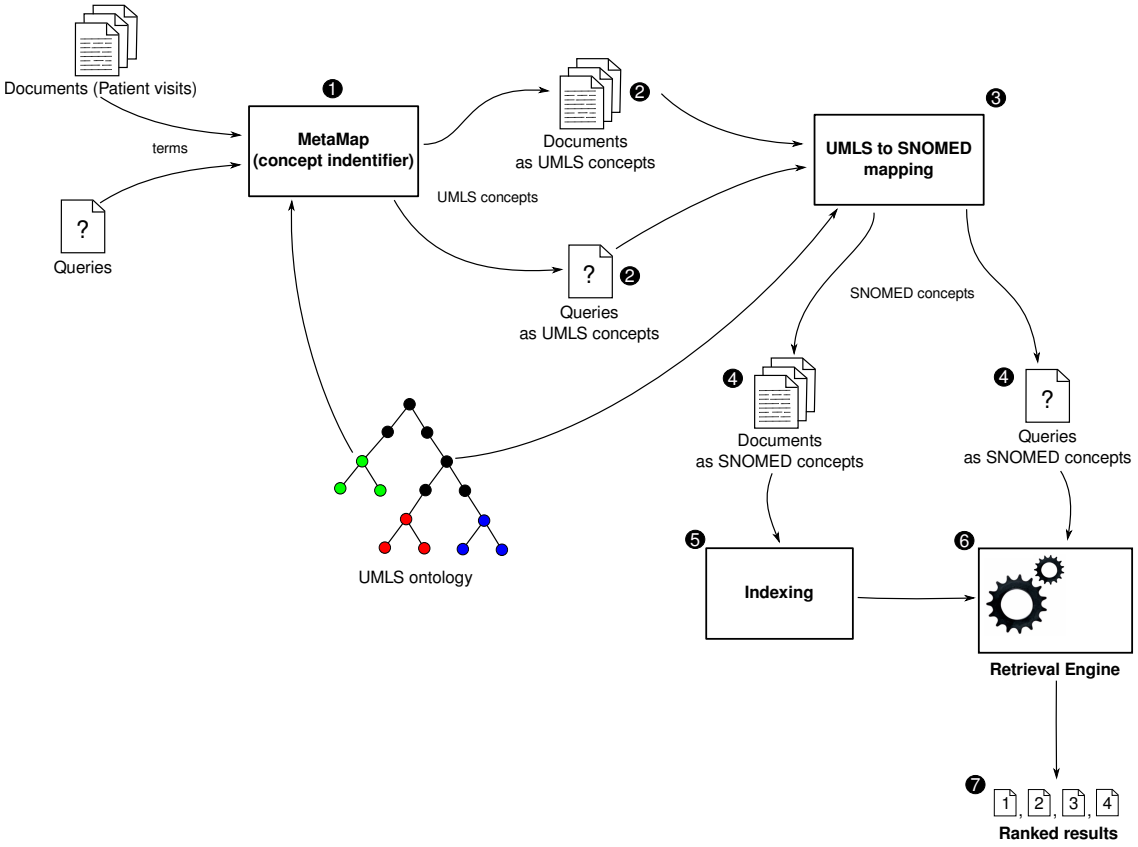


Figure 2: Architecture for concept-based medical information retrieval.

❶ Original queries and documents are fed to the MetaMap information extraction system. MetaMap identifies medical concepts using the UMLS ontology and returns their corresponding UMLS concept ids.

❷ Each document and query is now represented as a list of UMLS concept ids (e.g. `C0027051`) rather than the original terms (e.g. `heart attack`). Documents now only contain medical concepts.

❸ The UMLS concepts are then mapped to their SNOMED CT equivalents. This mapping is provided as part of the UMLS Metathesaurus.

❹ Queries and documents are now represented as a list of SNOMED CT concept ids.

❺ Documents are indexed using the Indri Lemur search engine. The system treats the documents as a bag-of-concepts.

❻ The queries (represented as SNOMED CT concept ids) are issued to the retrieval engine.

❼ A ranked list of document results is returned.

Appendix A provides an example of converting a single term document into SNOMED CT concepts.

Table 1 provides a comparison of the term and concept based representations. It shows average query and document (visit) length for term-based, UMLS and SNOMED CT based representations.

	Queries length	**Documents length**
Original terms	9.9 terms / query	2053 words / document
UMLS concepts	8.2 concepts / query	7451 concepts / document
SNOMED concepts	12.5 concepts / query	8140 concepts / document

Table 1: Comparison of average query and document lengths for term and concept-based representations. Documents as patient visits.

The concept-based representations are considerably longer than the original term-based documents. This is a result of including all the candidate concepts suggested by the MetaMap program, not just those top-ranked concepts. Without candidate concepts the SNOMED CT average document length was 1168 concepts / document, considerably smaller than the term-based 2053 terms per document. Later experimental results show that retrieval performance is improved by including all candidate concepts rather than just choosing the top-ranked concepts suggested by MetaMap. Including candidates could be considered a type of basic query expansion.

3 Results and analysis

Table 2 showing the results we obtained in comparison to median values obtained across all systems.

	bpref (%Δ)	**R-prec (%Δ)**	**Prec@10 (%Δ)**
Median	0.4115	0.3087	0.4764
AEHRC1	0.4636 (+12.66%)	0.3640 (+17.91%)	0.5059 (+6.19%)

Table 2: Comparison of our concept-based approach (AEHRC1) to the median result obtained across all systems.

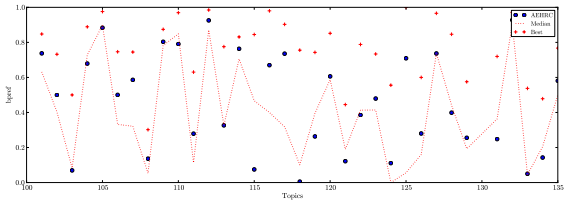
Overall, our concept-based approach demonstrates an improvement the median. Analyses for individual topics is provided in Figure 3 for the three metrics bpref 3(a), R-prec 3(b) and Precision @ 10 3(c).

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 worst performing queries.

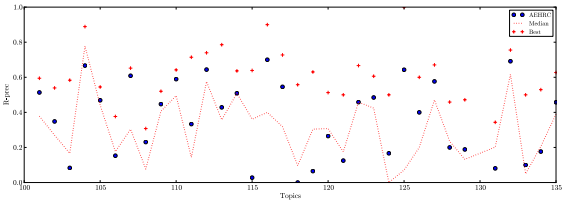
Performance on individual topics seems to correlate with that of the median, i.e. the concept-based approach does not differ drastically. Particular topics of interest that showed significant improvement included:

`107` *Patients with chronic back pain who receive an intraspinal pain medicine pump*:
In our system this query was converted to the single SNOMED CT concept `82711006`, described as *Infiltrating duct carcinoma*. Our system equalled the best score of 0.9 Prec@10.

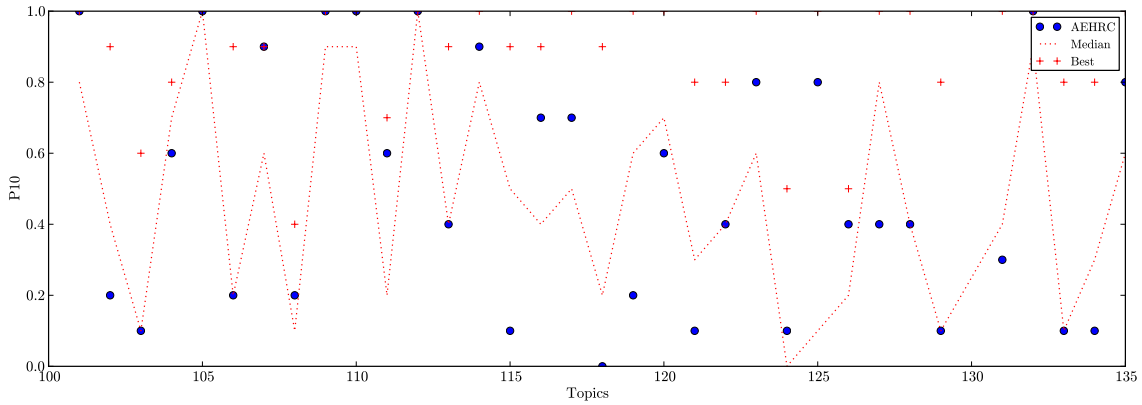
`125` *Patients co infected with Hepatitis C and HIV*:
HIV may have a number of different variants: HIV, AIDS or Human immunodeficiency virus, similarly Hepatitis C could be written as Hep-C or Hep C. Converting to SNOMED CT



(a) bpref



(b) R-prec



(c) Precision @ 10

Figure 3: Plots comparing concept-based, median and best systems for each topic.

concepts meant these different variants all mapped to the SNOMED CT codes `86406008` (*Human immunodeficiency virus infection*) and `62944002` (*Hepatitis C virus*). Prec@10 was 0.8, median 0.1.

4 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. Results show our approach performs reasonably well, above the median value from all system for all performance metrics. Our system was generic and due to time constraints was in no way tuned or adapted to the test corpus. Our concept-based approach provides a platform for further development into inferencing based search systems for dealing with medical data.

References

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

[2] M. Clark, Y. Kim, U. Kruschwitz, D. Song, D. Albakour, S. Dignum, U. C. Beresi, M. Fasli, and A. D. Roeck. Automatically structuring domain knowledge from text: An overview of current research. *Information Processing & Management*, In Press,, 2011.

[3] A. Gaudinat, P. Ruch, M. Joubert, P. Uziel, A. Strauss, M. Thonnet, R. Baud, S. Spahni, P. Weber, J. Bonal, C. Boyer, M. Fieschi, and A. Geissbuhler. Health search engine with e-document analysis for reliable search results. *International Journal of Medical Informatics*, 75(1):73–85, 2006.

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

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

[6] S. Meystre and P. J. Haug. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599, 2006.

[7] W. Pratt and M. Yetisgen-Yildiz. A study of biomedical concept identification: MetaMap vs. people. In *Proceedings of American Medical Informatics Association Symposium (AMIA)*, pages 529–533, Jan. 2003.

[8] L. W. Wright, H. K. G. Nardini, A. R. Aronson, and T. C. Rindflesch. Hierarchical Concept Indexing of Full-text Documents in the UMLS Information Sources Map. *Journal of the American Society for Information Science*, 50(6):514–523, 1999.

[9] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. 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*, pages 655–662, New York, USA, 2007. ACM Press.

A Converting terms to concepts

The sections provides an example of converting an original medical document into SNOMED CT concepts. Figure 4(a) shows the original term document. This document is converted to UMLS concepts (b) by the MetaMap system. UMLS concepts are then mapped to SNOMED CT concepts (c). The description for each of the SNOMED CT concepts is provided in Table 3.

(a) Original medical document

```
LEFT ANKLE:
**DATE[Jul 3 07] 8:59 PM
FINDINGS: There is moderate
soft tissue swelling.
There is no fracture or
dislocation.
The ankle mortise is
intact. IMPRESSION: NO
ACUTE FRACTURE. J4 END OF
IMPRESSION
```

(b) UMLS concepts

```
C1280015 C0230448
C0011008 C0243095
C0205081 C0037580
C0016658 C0012691
C0003086 C0003087
C1283839 C0039316
C0205266 C0564590
C0205178 C0016658
C0442779 C0444930
C0442779 C1522314
C0564590
```

(c) SNOMED CT concepts

```
241784008 51636004
118573002 246188002
6736007 298349001
72704001 157257005
344001 70258002
361292008 108371006
11163003 286781002
53737009 72704001
260253008 261782000
260253008 422117008
286781002
```

Figure 4: Example document from the BLULab corpus represented as original text, UMLS concepts and SNOMED CT concepts (Report Id: 20070703RAD-0JXYWK9UldBF-392-867771537).

Id	Preferred term	Id	Preferred term
241784008	Entire left ankle (body structure)	70258002	Ankle joint structure (body structure)
51636004	Structure of left ankle (body structure)	361292008	Entire ankle region (body structure)
118573002	Date (property) (qualifier value)	108371006	Bone structure of tarsus (body structure)
246188002	Finding (finding)	11163003	Intact (qualifier value)
6736007	Moderate (severity modifier) (qualifier value)	286781002	Character trait finding of level of suggestibility (finding)
298349001	Soft tissue swelling (finding)	53737009	Acute (qualifier value)
72704001	Fracture (morphologic abnormality)	157257005	[Dislocations &/or sprains &/or strains] or subluxations (disorder)
260253008	J4 (finding)	344001	Ankle region structure (body structure)
422117008	Stop (qualifier value)	261782000	End (qualifier value)
286781002	Character trait finding of level of suggestibility (finding)		

Table 3: Preferred term descriptions for SNOMED CT concepts taken from Figure 4(c).