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

Bevan Koopman\textsuperscript{1,2}\textsuperscript{*}, Peter Bruza\textsuperscript{2}, Laurianne Sitbon\textsuperscript{2}, Michael Lawley\textsuperscript{1}

\textsuperscript{1}Australian e-Health Research Centre, CSIRO
\textsuperscript{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

\textsuperscript{*}Correspondence to bevan.koopman@csiro.au
2.1 Documents as visits

The guidelines for the Medical Records Track stated that the unit of retrieval should be 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.

![Figure 1: MetaMap output for heart attack.](image-url)

MetaMap first analyses the input string and produces a ranked list of possible matching candidate concepts (shown in Fig 1). From this list of candidates the system selects the highest ranking candidate as Heart attack (Myocardial Infarction) (C0027051) (shown in Fig 1). In our experiments we found including the candidate concepts from Fig 1 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:

1. 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.

2. 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.

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

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

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

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

7. A ranked list of document results is returned.

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

<table>
<thead>
<tr>
<th></th>
<th>Queries length</th>
<th>Documents length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original terms</td>
<td>9.9 terms / query</td>
<td>2053 words / document</td>
</tr>
<tr>
<td>UMLS concepts</td>
<td>8.2 concepts / query</td>
<td>7451 concepts / document</td>
</tr>
<tr>
<td>SNOMED concepts</td>
<td>12.5 concepts / query</td>
<td>8140 concepts / document</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>bpref (%Δ)</th>
<th>R-prec (%Δ)</th>
<th>Prec@10 (%Δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>0.4115</td>
<td>0.3087</td>
<td>0.4764</td>
</tr>
<tr>
<td>AEHRC1</td>
<td>0.4636 (+12.66%)</td>
<td>0.3640 (+17.91%)</td>
<td>0.5059 (+6.19%)</td>
</tr>
</tbody>
</table>

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 bpre 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
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


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.

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

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

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