Exploiting SNOMED CT Concepts & Relationships for Clinical Information Retrieval:

Australian e-Health Research Centre and Queensland University of Technology at the TREC 2012 Medical Track

Bevan Koopman[1,2]*, Guido Zuccon[1], Anthony Nguyen[1],
Deanne Vickers[1], Luke Butt[1], Peter Bruza[2]

[1]Australian e-Health Research Centre, CSIRO
[2]School of Information Systems, Queensland University of Technology

Brisbane, Australia

Abstract

The Australian e-Health Research Centre and Queensland University of Technology recently participated in the TREC 2012 Medical Records Track. This paper reports on our methods, results and experience using an approach that exploits the concept and inter-concept relationships defined in the SNOMED CT medical ontology.

Our concept-based approach is intended to overcome specific challenges in searching medical records, namely vocabulary mismatch and granularity mismatch. Queries and documents are transformed from their term-based originals into medical concepts as defined by the SNOMED CT ontology, this is done to tackle vocabulary mismatch. In addition, we make use of the SNOMED CT parent-child 'is-a' relationships between concepts to weight documents that contained concept *subsumed* by the query concepts; this is done to tackle the problem of granularity mismatch. Finally, we experiment with other SNOMED CT relationships besides the is-a relationship to weight concepts related to query concepts.

Results show our concept-based approach performed significantly above the median in all four performance metrics. Further improvements are achieved by the incorporation of weighting subsumed concepts, overall leading to improvement above the median of 28% infAP, 10% infNDCG, 12% R-prec and 7% Prec@10. The incorporation of other relations besides is-a demonstrated mixed results, more research is required to determined which SNOMED CT relationships are best employed when weighting related concepts.

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

*Correspondence to bevan.koopman@csiro.au

combining statistical approaches in information retrieval and natural language processing with the formal semantics of the SNOMED CT medical ontology.

Our system used for the TREC Medical Records Track uses concept-based information retrieval models and medical domain knowledge provided by the SNOMED CT ontology; the system builds up from our experience at the 2011 Medical Records Track [2]. 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; this is done to address the problem of vocabulary mismatch. Queries and documents are transformed from their original terms to SNOMED CT concepts, retrieval is then performed by matching concepts. Concept-based approaches have previously demonstrated excellent results — Zhou et al. [10] concept-based system (using concept from UMLS ontology and MeSH headings) was the top performing at the TREC Geonomics Track.

We further investigate the use of relationship between concepts explicitly defined in the SNOMED CT ontology. Firstly, we utilise the parent-child ('is-a') relationship to weight concepts in a document that are children or *subsumed* by the query concepts. This is done to address the problem of granularity mismatch — where the concepts found in the documents are very detailed and specific, while those in the query are more general and high-level. The incorporation of subsumed concepts into the retrieval model had previously shown improvement in retrieval performance, as shown by Zuccon et al. [11]. Finally, besides the is-a relationship, we consider all types of relationships in SNOMED CT to weight concepts that are related to the query concept.

Results show that the concept-based system alone performed well above the median (+26% infAP). The inclusion of subsumed concepts via SNOMED CT is-a relationships demonstrated some additional improvements in performance. The inclusion of other SNOMED CT relationships demonstrated mixed results and require further investigation.

2 Methods

2.1 Concept-based Information Retrieval

In our system all queries and documents are converted from the original term-based representation into medical concepts. For this purpose we used MetaMap, which has been widely adopted in medical NLP [6, 7] and medical IR [3, 1, 5]. 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 same UMLS concept. Conversion to concepts aims to overcome some of the vocabulary mismatch that exists in medical text.

We represented both documents and queries not as bag-of-words but as bag-of-concepts. The overall process to translate from terms to concepts is as follows:

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

2. The UMLS concepts are then mapped to their SNOMED CT equivalents. This mapping is provided as part of the UMLS Metathesaurus. Queries and documents are now represented as a list of SNOMED CT concept ids.

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

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

5. A ranked list of document results is returned.

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

	#Docs	**Queries length**	**Documents length**	**#Vocab.**
Original terms	17,198*	8.4 terms / query	2338 terms / doc	218,574
SNOMED concepts	17,198*	12.9 concepts / query	6066 concepts / doc	54,143

*100,866 original reports collapsed to 17,198 patient *visit* documents.

Table 1: Collection statistics for the TREC MedTrack'11 corpus of clinical records. Statistics are provided for the original term corpus and subsequent corpus after conversion to SNOMED CT concepts.

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 1391 concepts / document, considerably smaller than the term-based 2338 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.

2.2 Incorporation of SNOMED CT relationships

In this section we describe our methods for extending the concept-based approach with the inclusion of SNOMED CT relationships. This approach includes in the weighting function not just the query concepts appearing in the document, but also the concept related to the query concepts that appear in the document. Documents are scored according to (1) the weight of query concepts in a document, and (2) the weight of concepts in a document that are related to a query concept. For each query concept c_i we obtain the list of related concepts $c_j \sqsubseteq c_i$ from the SNOMED CT ontology. These subsumed concepts are included in the retrieval function as follows:

$$RSV(d|q) = \sum_{c_i \in q} w(c_i, d) + \sum_{c_j \sqsubseteq c_i; c_i \in q} \delta(w(c_j, d)) \quad (1)$$

where $w(c_i, d)$ is the weight of concept c_i in document d, and $\delta(w(c_j, d))$ adjusts the weight of a related concept c_j. That is, the score of a document for a query q is the sum of the weights associated with the query concepts and the adjusted weights of the concepts that are related to the query concepts.

Equation 1 is a general method to integrate subsumed concepts into the retrieval function. A number of instantiations of both $w(c_i, d)$ and $\delta(w(c_j, d))$ are possible. In this instance we used the enhanced tf-idf described by Zhai [9] in which the Okapi formula is used for weighting term frequencies and where concepts are used instead of terms, i.e.:

$$w(c_i, d) = \frac{k_1 count(c_i, d)}{count(c_i, d) + k_1(1 - b + b\frac{l_d}{l_{avg}})} \cdot \log \frac{|D|}{|d(c_i)|} \tag{2}$$

where l_{avg} is the average document length, and k_1, b are the Okapi parameters.

2.2.1 Weighting Subsumed Concepts using IS-A Relationships

If a concept is subsumed by another concept in the SNOMED CT ontology it indicates it is a child or specialisation of the parent concept. We include subsumed concept weighting into the retrieval function in order to tackle the problem of granularity mismatch — a problem previously identified in medical IR.

From Equation 1 the related concepts, denoted by $c_j \sqsubseteq c_i$ is restricted to only is-a relationship. Thus, we consider how the weight of a concept should be adjusted if it was subsumed by the query. A straightforward approach would be to treat subsumed concepts in the same way as query concepts, i.e. $\delta(w(c_j, d)) = w(c_j, d)$. However, the presence of a subsumed concept in a document may offer a different indication of relevance than an actual query concept. A subsumed concept indicates a specialisation of the parent concept, and thus treated differently to an actual query concept. Intuitively a subsumed concept would be a weaker indication of relevance than a query concept. Thus, the weight for the subsumed concept c_j in the document is dampended according to the square root of the weight $w(c_j, d)$, i.e.:

$$\delta(w(c_j, d)) = \sqrt{w(c_j, d)} \tag{3}$$

In this case a subsumed concept contributes less evidence towards the score of a document than a query concept.

2.2.2 Inclusion of all SNOMED CT relationships

The previous section considered weighting of concepts related to the query concept via a is-a relationship, i.e., subsumed concepts. In this section we widen the scope to all types of SNOMED CT relationships. Thus, $c_j \sqsubseteq c_i$ from Equation 1 includes all relationships. For the implantation of the weighting adjustment function $\delta(w(c_j, d)) = w(c_j, d)$ we implement a similarity function *sim* that estimates the similarity between the two concepts c_i and c_j. Instead of dampening using the uniform $\sqrt{w(c_j, d)}$ applied to subsumed concepts we consider similarities between concepts to modulate weights. A number of different similarity measures are possible. Previous research by Pedersen et al. [8] found corpus-driven measures of similarity to be effective in the medical domain. Following the findings of Koopman et al. which evaluated a number of corpus-driven measures [4] we implement as our similarity measure the cosine angle between the two concepts c_i and c_j document vectors.

2.3 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.

3 Results and Analysis

Table 2 reports the results we obtained in comparison to median values obtained across all systems. Overall, our concept-based approaches demonstrate improvements over the median of the TREC systems.

	infAP (%Δ)	**infNDCG** (%Δ)	**R-prec** (%Δ)	**Prec@10** (%Δ)	**Recall**
Median	0.1689	0.4243	0.2960	0.4702	
AEHRC0	0.2130 (+26%)	0.4630 (+9%)	0.3251 (+10%)	0.4894 (+4%)	0.6406
AEHRCsub	0.2163 (+28%)	0.4682 (+10%)	0.3314 (+12%)	0.5043 (+7%)	0.6702
ARHRC1	0.2066 (+22%)	0.4614 (+8%)	0.3140 (+7%)	0.4596 (-2%)	0.6675
ARHRC2	0.2128 (+25%)	0.4608 (+8%)	0.3205 (+9%)	0.4553 (-3%)	0.6803

Table 2: Comparison of our concept-based approaches to the median result obtained across all TREC systems.

3.1 Concept-based IR contribution

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, e.g. query 178 "Patients with metastatic breast cancer". Nevertheless, our concept-based baseline (AEHRC0) provides consistent improvements over the median of TREC systems.

3.2 Effect of Weighting Concept Relationships

We now consider the contribution of combining weights of query concepts and related concepts for scoring a document. Empirical results show that considering related concepts along with the original query concepts can improve retrieval effectiveness; which concept relationships to consider and how to weight these is however a challenging issue. Our AEHRC1 and AEHRC2 submission have proved that not all concept relationships lead to improvements of retrieval

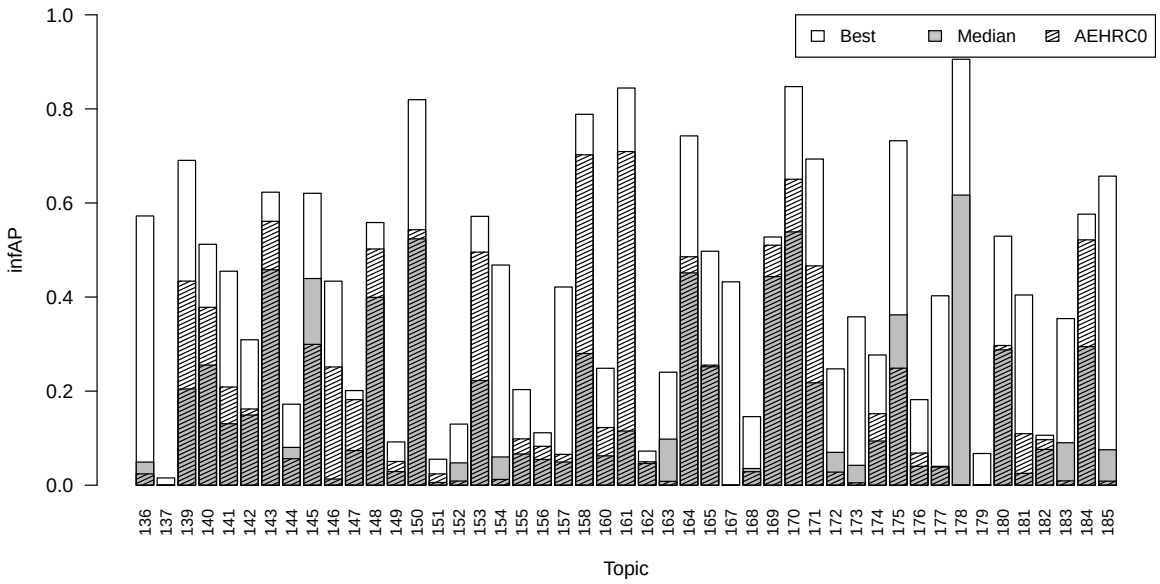

Figure 1: InfAP for lvl0 vs median

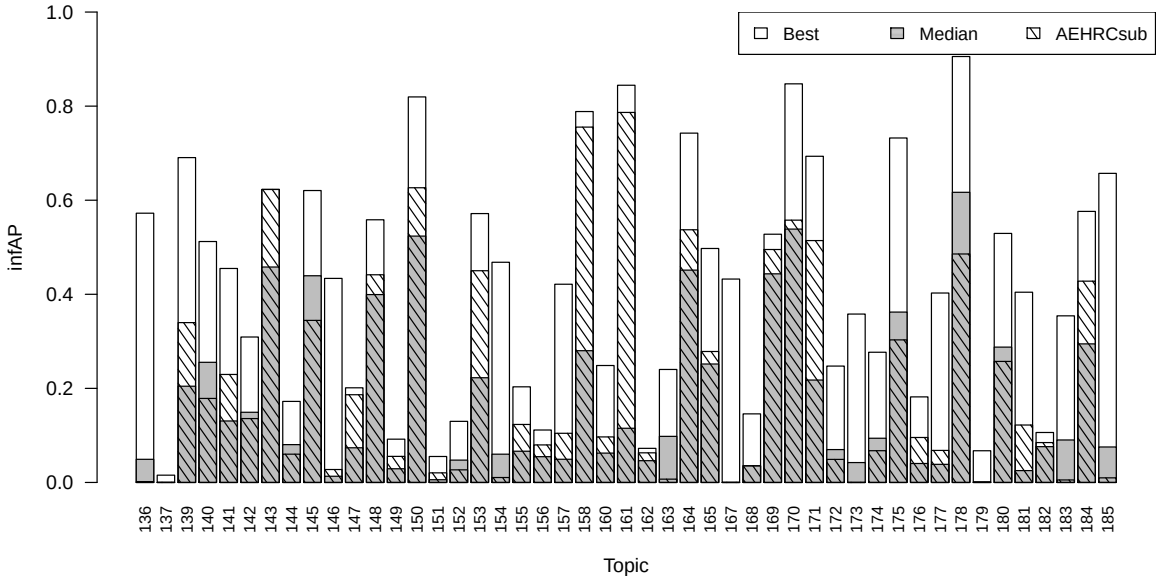

Figure 2: InfAP for sub vs median

effectiveness; while our AEHRCsub submission has shown that subsumption relationships do indeed provide relevant information that ultimately can lead to improvements. In addition, this submission demonstrate that improvements can be obtained by adjusting weights of subsumed concepts via an ad-hoc simple function. However, results obtained using this approach are not consistent throughout the whole query set; the approach itself provides small flexibility and interpretational power.

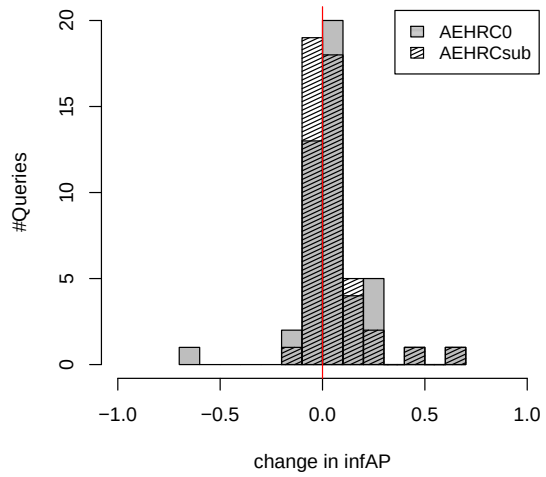

Figure 3: Change in infAP over the Median for AEHRC0 and AEHRCsub

4 Conclusion

We have presented approaches to searching electronic medical records 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. Relationships between concepts are also accounted for in our document scoring approaches.

Results show our approaches perform reasonably well, above the median value from all TREC systems. Our concept-based approaches that combine concept relationships in the retrieval function provide a platform for further development into inferencing based search systems for dealing with medical data.

References

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

[2] Koopman, B., Bruza, P., Sitbon, L., and Lawley, M. AEHRC & QUT at TREC 2011 Medical Track : a concept-based information retrieval approach. In *Proceedings of 20th Text REtrieval Conference (TREC 2011)* (Gaithersburg, MD, USA, Nov. 2011), NIST, pp. 1–7.

[3] Koopman, B., Bruza, P., Sitbon, L., and Lawley, M. Towards Semantic Search and Inference in Electronic Medical Records: an approach using Concept-based Information Retrieval. *Australasian Medical Journal: Special Issue on Artificial Intelligence in Health 5*, 9 (2012), 482–488.

[4] Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., and Lawley, M. An Evaluation of Corpus-driven Measures of Medical Concept Similarity for Information Retrieval. In *21st ACM International Conference on Information and Knowledge Management (CIKM)* (Maui, USA, 2012).

[5] 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.

[6] Nguyen, A., Lawley, M., Hansen, D., Bowman, R., Clarke, B., Duhig, E., and Colquist, S. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *Journal of the American Medical Informatics Association 17*, 4 (2010), 440–445.

[7] Nguyen, A., Lawley, M., Hansen, D., and Colquist, S. A simple pipeline application for identifying and negating snomed clinical terminology in free text. In *HIC 2009: Proceedings; Frontiers of Health Informatics-Redefining Healthcare, National Convention Centre Canberra, 19-21 August 2009* (2009), Health Informatics Society of Australia (HISA), p. 188.

[8] Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., and Chute, C. G. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics 40*, 3 (2007), 288–299.

[9] Zhai, C. Notes on the Lemur TDIDF model. Tech. rep., School of Computer Science, Carnegie Mellon University, 2001.

[10] 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.

[11] Zuccon, G., Koopman, B., Nguyen, A., Vickers, D., and Butt, L. Exploiting Medical Hierarchies for Concept-based Information Retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium (submitted)* (2012).