Semantic search and inferencing in health informatics

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

1. Introduction

Consider a person searching electronic health records, a search for the term 'cracked skull' should return documents that contain the term 'cranium fracture'. A information retrieval systems is required that matches *concepts*, not just keywords. Further more, determining relevance of a query to a document requires *inference* – its not simply matching concepts. For example a document containing 'dialysis machine' should *align* with a query for 'kidney disease'. Collectively we describe this problem as the 'semantic gap' – the difference between the raw medical data and the way a human interprets it.

This paper presents an approach to semantic search of health records by combining two previous approaches: an ontological approach using the SNOMED CT medical ontology; and a distributional approach using semantic space vector space models. Our approach will be applied to a specific problem in health informatics: the matching of electronic patient records to clinical trials.

2. Description of the work

Identifying and recruiting eligible patients for participation in clinical trials is an important step in the success of a clinical trial. A clinical trial represents a *long-query* expressing an information need and patient records can be viewed as a collection of documents with varying relevance. The semantic gap between the clinical trial and patient record can be significant, for example a trial may contain the name of disease (e.g. appendicitis), whereas the patient record might only mention procedures/drugs used to treat it (e.g. appendectomy). This makes it a good candidate for a semantic search approach. Our approach draws on two main areas of previous work:

SNOMED medical ontology. Firstly we leverage the definitional knowledge contained within SNOMED CT, a machine readable medical terminology covering a large range of concepts, including: disorders, procedures, organisms, body structures and pharmaceuticals [1]. It contains approximately 283,000 concepts, 732,000 active terms and 923,000 active relationships. The Australian e-Health Research Centre has developed *snorocket* – a Java implementation for efficient classification of the SNOMED CT ontology [2].

Semantic spaces. Secondly, we make use of *semantic spaces* – a pragmatic high dimensional representation of words found in a free-text corpus. Semantic spaces capture the relationships between words by analysing how words co-occur with each other or within documents. Semantic spaces have a proven track record in semantic tests (such as synonym tests) designed for humans [3]. The meaning of a word within the semantic space is determined by its relation to other words around it. The SemanticVectors software package is a Java implementation for indexing, constructing and searching semantic spaces [4].

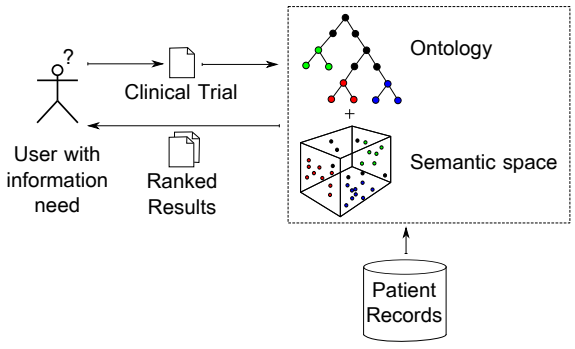

Figure 1: A combined ontological and semantic space approach to information retrieval.

[1]Information Systems Discipline, Faculty of Science & Technology, Queensland University of Technology
[2]Australian e-Health Research Centre, CSIRO

Additionally, clinical trials are an example of a verbose query. As current search engines typically do not perform well with long queries [5] we will make use of techniques for improving performance of verbose queries, these include summarising the query to a subset of key terms [6] or assigning relevant weights to query terms [5].

Our approach to semantic search, illustrated in Figure 1, is to integrate the SNOMED CT ontology and distribution approaches such as semantic spaces models. The ontology may be used for logical deductive tasks – for example inferring that a 'dialysis machine' may be used to treat 'renal failure'. The semantic space can then be used to infer more indirect relationships – for example 'renal disease' and 'liver disease' are semantically similar in that they are both treated using 'dialysis'. This combined approach will make use of both ontologies and semantic spaces to deal with the semantic gap problem. A proof-of-concept system will be implemented with the use of the SemanticVectors and *snorocket* software packages.

3. Conclusion

In this paper we have outlined the need for search systems that go beyond keyword matching of health data and are able to match concepts, thus overcoming the 'semantic gap' problem. Our approach is to make use of symbolic representations of medical knowledge such as the SNOMED CT ontology and semantic spaces – distribution methods of representing words in a free-text corpus. This combined approach offers the advantage of formal deductive reasoning afforded by the ontology and associational relationships in the semantic spaces build from free-text. Evaluation will be conducted by applying our approach to improving the matching of patient records to clinical trials, a good example of a semantic retrieval problem in health informatics.

References

[1] Kent Spackman, "SNOMED clinical terms basics", IHTSDO presentation, August 2008, International Health Terminology Standards Development Organisation (IHTSDO).

[2] Michael Lawley, "Exploiting fast classification of SNOMED CT for query and integration of health data", in *Proceedings of the Third International Conference on Knowledge Representation in Medicine (KR-MED '08)*, R. Cornet and K.A. Spackman, Eds., Phoenix, Arizona, May 2008.

[3] T. K. Landauer and S. T. Dumais, "Solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge", *Psychological Review*, vol. 104, no. 2, pp. 211 – 240, 1997.

[4] Dominic Widdows and Kathleen Ferraro, "Semantic vectors: a scalable open source package and online technology management application", in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008, European Language Resources Association (ELRA), http://code.google.com/p/semanticvectors.

[5] Michael Bendersky and W. Bruce Croft, "Discovering key concepts in verbose queries", in *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2008, pp. 491 – 498, ACM.

[6] Giridhar Kumaran and James Allan, "A case for shorter queries, and helping users create them", in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, New York, April 2007, pp. 220–227, ACM.

Photographs of authors


Bevan Koopman


Peter Bruza


Michael Lawley


Laurianne Sitbon