Towards Exploiting Inference from Semantic Annotations for Medical Information Retrieval

Guido Zuccon[1], **Bevan Koopman**[2,1], **and Peter Bruza**[1]
[1]School of Information Systems, Queensland University of Technology, Australia
[2]Australian E-Health Research Centre, CSIRO, Australia
{g.zuccon, b.bruza}@qut.edu.au, bevan.koopman@csiro.au

Search is a core capability that can support clinicians and health practitioners in performing their activities. Advances in semantic information retrieval have the promise to radically enhance how clinicians search and access medical information, e.g. when searching electronic health reports for clinical trial cohort selection[1], or when searching medical literature to inform evidence-based medicine[2]. In this talk we review some of the efforts made by the information retrieval community to enhance search for specific tasks within the medical domain (including methods to support health consumers searching for health advice), and in particular within the umbrella of a number of research community shared-tasks and resources, e.g., [12, 3, 8].

While early work in medical information retrieval has mainly translated techniques developed outside of the medical domain [4], there has been a recent increase in research that exploits resources and factors that uniquely characterise the medical domain. The medical domain in fact offers a plethora of carefully curated, structured, semantic resources (such as the UMLS and SNOMED CT), along with well established entity extraction and linking tools (e.g. Metamap [1]). Recent studies [6] have furthermore suggested that medical information retrieval is characterised by problems that are inherently inferential and thus, integrating inference within the retrieval process may provide significant breakthroughs in the way information is searched in this domain. These characteristics are rarely found together in general information retrieval settings. In this talk we show however that current work in medical information retrieval has only scratched the surface of what can be done with semantic annotations to improve search on medical content. We argue that integrating semantic inference in information retrieval has the potential to uncover a large amount of information that otherwise would be inaccessible; but inference is also risky and, if not used cautiously, can harm retrieval.

Velupillai has discussed the potential benefit that semantic annotations have in the clinical domain, in particular for hypothesis generation, adverse event identification and patient similarity [11]. Koopman and Zuccon [6] exhibited some of the reasonings that users execute when assessing the relevance of clinical documents to queries aimed at identifying patient cohorts that satisfy a number of inclusion criteria for clinical trial. They showed that users apply inference mechanisms to determine temporality, the relations between query aspects, cause-effect relations, etc. Similarly, Koopman [5] performed a meta-analysis of the different factors that require semantic inference to augment retrieval in the medical domain; Koopman identified the following as the primary factors: (1) semantic similarity and (2) granularity, (3) conceptual implication, (4) temporality, (5) level of uncertainty, (6) negation and contextually, (7) dependencies between entities.

Rudimentary inference mechanisms were used in a heuristic-driven model that exploited semantic concepts and is-a (a.k.a. subsumption) relations [15], achieving a form of semantic query expansion that was driven by the inference that, for example, a sought concept was the parent of a more specific concept present in a potentially relevant document. Cohen et al. [2] has recently demonstrated the use of more sophisticated levels of semantic inference accomplished through analogical reasoning based on high-dimensional vector representations to drive pseudo-relevance feedback. Although this

[1] e.g.,http://trec.nist.gov/data/medical.html
[2] e.g., http://www.trec-cds.org/

method does encode higher level semantic inferences it does not provide substantial gains in terms of retrieval effectiveness and is indeed inferior to well engineered statistical approaches to search.

A full-fledged semantic inference model for medical information retrieval has been proposed by Koopman [5] (the Graph Inference Model, GIN). The model, partially inspired by early literature on formal logic-based models of information retrieval, performs an implicit query expansion at retrieval time by traversing a graph-based semantic representation of the corpora, where both queries and documents are annotated with semantic concepts (forming the nodes in the graph) and edges encode different semantic relations between concepts (e.g., is-a, finding site, active ingredient, etc.). In the GIN, concepts that are inferred from the query concepts (by exploiting semantic relations directly encoded in the underling knowledge base) allow to retrieve a higher number of relevant documents than what the original query concepts retrieved in first instance. Thorough empirical experimentation however demonstrated that semantic inference, as encoded in the GIN, is a risky mechanism that, if not used cautiously, can harm retrieval [5]. Similar conclusions were observed in the work of Cohen et al. [2]. An important observation springing from Koopman's work is that it is the quality of the structured knowledge resource that influences the quality of inferences used for retrieval: in particular, inferences that may be logically valid from a representational perspective may not provide valuable information when used for retrieval [5].

Despite the previously mentioned efforts (among others) to integrate inference mechanisms within retrieval models for medical search, the most successful approaches in this domain used semantic annotations, but only with weak inference mechanisms, by combining structured domain knowledge with document and corpus statistics. For example, using concepts, semantic types (higher level groupings of concepts, e.g., diseases, organisms) and corpus statistics, [13] were able to derive implicit relations between concepts, which could be used for query expansion; this was the best approach at the TREC Genomics Track [14]. Similarly, recent work by Limsopatham et al. [7] partially exploits semantic dependency information between concepts, but relying on a powerful statistical diversification approach for document ranking adapted from web retrieval [9] rather than more semantically-grounded inference mechanisms.

Thus, while current state-of-the-art medical information retrieval systems solicit semantic information, more investigation is required to effectively integrate (semantic) inference in the retrieval process. As other authors have stated in the past, e.g., [10], techniques that effectively understand the content of documents and the intent of queries and are able to exploit the inferential process for retrieval are expected to provide significant improvements in search effectiveness, even more so in the medical domain.

References

[1] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proc. of AMIA*, page 17. American Medical Informatics Association, 2001.

[2] T. Cohen, D. Widdows, and T. Rindflesch. Expansion-by-analogy: A vector symbolic approach to semantic search. In *Proc. of QI'14*, 2014 (to appear).

[3] L. Goeuriot, G. J. Jones, L. Kelly, J. Leveling, A. Hanbury, H. Müller, S. Salanterä, H. Suominen, and G. Zuccon. Share/clef ehealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports. In *Proc. of CLEF'13*, 2013.

[4] W. Hersh. *Information Retrieval: A Health and Biomedical Perspective: A Health and Biomedical Perspective*. Springer, 2008.

[5] B. Koopman. *Semantic Search as Inference: Applications in Health Informatics*. PhD thesis, School of Information Systems, Queensland University of Technology, 2014.

[6] B. Koopman and G. Zuccon. Why assessing relevance in medical ir is demanding. In *MedIR 2014*, 2014.

[7] N. Limsopatham, C. Macdonald, and I. Ounis. Modelling relevance towards multiple inclusion criteria when ranking patients. In *Proc. of CIKM'14*, 2014 (to appear).

[8] D. Molla and M. E. Santiago-Martinez. Development of a corpus for evidence based medicine summarisation. In *Proc. of ALTA'11*, pages 86–94, 2011.

[9] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proc. of WWW'10*, pages 881–890. ACM, 2010.

[10] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM TOIS*, 9(3):187–222, 1991.

[11] S. Velupillai. Semantic annotations in clinical documentation: Exploring potentials for future information retrieval. In *Proc. ESAIR'10*, pages 9–10. ACM, 2010.

[12] E. Voorhees and R. Tong. Overview of the trec 2011 medical records track. In *Proc. of TREC'11*, 2011.

[13] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proc. of SIGIR'07*, pages 655–662, Amsterdam, The Netherlands, 2007.

[14] W. Zhou, C. Yu, V. Torvik, and N. Smalheiser. A concept-based framework for passage retrieval in genomics. In *Proc. of TREC'06*, pages 14–17, 2006.

[15] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting medical hierarchies for concept-based information retrieval. In *Proc. of ADCS'12*, pages 111–114. ACM, 2012.