Biomedical Cross-Language Information Retrieval

Dolf Trieschnigg
HMI, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
trieschn@ewi.utwente.nl

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Theory, Experimentation, Measurement

Keywords

Biomedical IR, CLIR, Language Models

MEDLINE, the primary bibliographic database in life sciences, currently contains more than 17 million article citations and last year grew with 600,000 entries. Staying up-to-date, finding relevant information and even extracting new knowledge becomes increasingly difficult in this field [1].

The peculiarities of biomedical terminology make building an effective IR system a challenge [3]. Firstly, biomedical terms are highly synonymous and ambiguous. Secondly, multi-word terms such as 'Bovine Spongiform Encephalopathy' are commonly used, making a bag of words approach less effective. Thirdly, new terms and especially abbreviations are abundant. And finally there is the challenge of variation in terminology. Differences in spelling, use of hyphens and other special characters make it even more difficult to handle biomedical text.

The TREC Genomics benchmarks have demonstrated that overcoming these challenges is far from trivial. Different attempts have been pursued to map text to a notion of concepts. Explicitly, by mapping texts to entries in controlled vocabularies such as the Unified Medical Language System (UMLS). But also implicitly, by for example treating collocated words as 'concepts'.

The goal of my PhD project is to study how to optimize biomedical IR by including conceptual knowledge from biomedical ontologies, while maintaining a theoretical sound framework. To achieve this I propose to approach terminology issues in biomedical IR as a form of cross-lingual IR (CLIR). The two 'languages' distinguished are the textual representation of query and documents, and their conceptual representation in terms of concepts from a biomedical ontology. Traditional CLIR has shown the benefits of a tight integration of probabilistic translation into a retrieval framework of generative language models [2]. The interesting research questions for biomedical CLIR lie in the differences with traditional CLIR.

Firstly, in contrast to traditional CLIR, both the textual and conceptual representation can be available and thus be used for translation and matching. For the conceptual representation of the documents either a sparse but manually curated set of MeSH concepts can be chosen, or a completer but probably not flawless representation obtained from a biomedical concept recognizer. For a textual query a conceptual counterpart can be obtained by pseudo-relevance feedback from concept-tagged documents or again by using the concept recognizer.

Secondly, how to obtain translation resources. A rough alignment can be made between concepts and abstracts, or a finer alignment can be produced by using a biomedical concept recognizer to align phrases to concepts. These alignments can be used as parallel or comparable corpora to train translation models. Open questions are which textual unit should be used for translation (word, character- or word based n-grams) and what the impact is of using different parallel corpora and concept recognizers.

Finally, there is the choice of matching strategy. The matching can take place in textual, conceptual and combined representation. Moreover, the query and document representation can be a mix of the original representation and a translated representation.

During the doctoral consortium I would like to discuss the, hopefully, added value of a dual representation and a cross-lingual approach to this old issue. The conceptual representation seems particularly useful for user feedback. I would appreciate feedback how to investigate this.

Acknowledgements This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). I would like to thank my supervisors Franciska de Jong and Wessel Kraaij for their support in writing this proposal.

References

[1] L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, 7:119–129, feb 2006.

[2] W. Kraaij and F. de Jong. Transitive probabilistic CLIR models. In *Proceedings of RIAO 2004*, 2004.

[3] G. Nenadic, I. Spasic, and S. Ananiadou. Mining biomedical abstracts: What's in a term? *IJCNLP 2004*, pages 797–806.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.