Use of controlled vocabularies to improve biomedical information retrieval tasks

Emilie Pasche[a,b], Julien Gobeill[b], Dina Vishnyakova[a], Patrick Ruch[b], Christian Lovis[a]

[a] *Division of Medical Information Sciences, University Hospitals of Geneva and University of Geneva, Geneva, Switzerland*
[b] *Bibliomics and Text-Mining Group, University of Applied Sciences Western Switzerland, Geneva, Switzerland*

Abstract and Objective

The high heterogeneity of biomedical vocabulary is a major obstacle for information retrieval in large biomedical collections. Therefore, using biomedical controlled vocabularies is crucial for managing these contents. We investigate the impact of query expansion based on controlled vocabularies to improve the effectiveness of two search engines. Our strategy relies on the enrichment of users' queries with additional terms, directly derived from such vocabularies applied to infectious diseases and chemical patents. We observed that query expansion based on pathogen names resulted in improvements of the top-precision of our first search engine, while the normalization of diseases degraded the top-precision. The expansion of chemical entities, which was performed on the second search engine, positively affected the mean average precision. We have shown that query expansion of some types of biomedical entities has a great potential to improve search effectiveness; therefore a fine-tuning of query expansion strategies could help improving the performances of search engines.

Keywords:

Controlled vocabularies, Normalization, Information Retrieval

Methods

Our approach consists to expand users' queries with additional relevant terms in order to improve search effectiveness of two different search engines: a question-answering (QA) engine for clinical guidelines and a search engine for chemical patents. We evaluate the impact of query expansion by comparing the effectiveness of these two search engines with and without query expansion.

Query expansion

A dictionary-based approach is used to expand three different biomedical entities: diseases, pathogens and clinical conditions. First, we extract information from the Medical Subject Headings (MeSH) terminology and perform a slight cleaning of the vocabulary to exclude strongly ambiguous terms. Second, exact matching strategies are used to identify the biomedical terms present in the query and MeSH identifiers are assigned. Third, we extract synonyms available in the MeSH terminology for each assigned semantic identifier.

A hybrid approach is used to expand terms related to chemicals. First, chemical terms used in the query are identified using Oscar3, an open-source chemical entity recognition tool for text annotation. Oscar3 relies on a dictionary of chemical terms and a set of rules. Second, identified terms are automatically assigned an identifier using a MeSH categorizer. Third, we extract synonyms available in two different vocabularies: MeSH and PubChem.

Evaluation

The query expansion using the dictionary-based approach is evaluated in the context of a QA task. A set of 23 clinical questions related to antibiotherapy is asked to a QA search engine. The system answers these questions with a ranked list of antibiotics. We perform synonym expansion on the different biomedical terms (i.e. diseases, pathogens and clinical conditions) of the queries and evaluate how the synonym terms obtained for each of the biomedical entity types impact the top-precision (P0) of the QA system.

The query expansion using the hybrid approach is evaluated in the context of the Text REtrieval Conferences (TREC) Chemistry Track of 2010. A set of 30 topics manually created by chemical experts is used to query the search engine. The system answers by returning a set of relevant patents that fulfil the information need. We perform synonym expansion on the chemical terms of the topics and evaluate how the synonym terms obtained impact the mean average precision (MAP) of the search engine.

Results

The query expansion based on the pathogens' names resulted in an improvement of the P0 (+4.5%, $p<0.01$). In opposite, the query expansion based on the diseases' terms negatively impacted the P0 (-10.5%, $p<0.01$). Similarly, the query expansion of both the disease and pathogen entities resulted in a strong decrease of the P0 (-9.8%, $p<0.01$). The query expansion based on the clinical conditions failed to find synonyms for any of our queries.

The results obtained for the patent search engine show that query expansion based on all chemical terms identified by Oscar3 and their synonyms resulted in a strong increase of the MAP (+100%, $p<0.01$).

Conclusion

Such automatic large synonym expansion boosts the performances for some queries, but also strongly degrades them for other queries. We can therefore assume that synonym expansion has a great potential to improve search effectiveness, but some of the synonyms would need to be manually validated because they degrade the performance of the search engine. Moreover, the tuning of our systems was very sparse. We can thus expect that a fine-tuning of the terminology-driven expansion could help improve the performances.

Moreover, the approaches for query expansion described here are quite simplistic. Indeed, the recognition of entities does not allow insertions or deletions within a term. Therefore, we consider that using more sophisticated systems will be more powerful for such tasks.