Information Retrieval Framework for Technology Survey in Biomedical and Chemistry Literature

Harsha Gurulingappa[1,2], Bernd Müller[1], Martin Hofmann-Apitius[1,2], and Juliane Fluck[1]

[1]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2]Bonn-Aachen International Center for Information Technology (B-IT)
Dahlmannstraße 2, 53113 Bonn, Germany
Corresponding Author: harsha.gurulingappa@scai-extern.fraunhofer.de

Abstract

The Technology survey task deals with the retrieval of information that can best answer a scientific question. This task is more challenging in biomedical and chemistry domains due to diverse conventions applied for naming the entities. In order to address this challenge, the work reported here presents an ad-hoc retrieval task that has been evaluated during the TRECCHEM-2011 for its ability to support retrieval from the biomedical and chemistry literature. The core of the framework contains nearly 1.3 million patents and full-text articles that were indexed with pre-selected biomedical concepts. Altogether, four runs were submitted based on different query formulation strategies and they exhibited competitive results.

1. Introduction

Information retrieval in biomedical and chemistry domains is challenging due to the presence of diverse denominations of concepts in the literature. A chemical name instance in text can be represented in terms of its trivial name, registry code (e.g. CAS number[1]), standardized nomenclature (e.g. IUPAC[2]), abbreviation, or brand name. For example, the drug 'Aspirin' is reported to have over 25 synonyms and 90 brand names according to the DrugBank[3] database. The naming diversification also applies to various classes of biomedical concepts such as gene or protein names, diseases, and many more.

In order to address the challenges associated with retrieval from patents and full-text articles, the TRECCHEM provides a platform for the development, evaluation, and comparison of systems for information retrieval in biomedical and chemistry domains. TRECCHEM defines two independent retrieval tasks namely the *Technology Survey* and the *Prior Art Search*. The first task provides a set of expert-defined natural language questions of information needs (also known as TS topics) for retrieving sets of documents from a predefined collection that can best answer those questions. The second task provides a set of test patents for retrieving sets of documents that can potentially invalidate the given test patents.

Considering the ambiguity inherent to biomedical and chemistry literature, tagging the chemical and biomedical concepts in documents followed by retrieval has demonstrated success in the past years [(Matos et al., 2010), (Gurulingappa et al., 2009b)]. Tagging the concepts and mapping them to standard database entries normalizes different forms of the same concept to one standard form. This helps to overcome the problems associated with synonyms, acronyms and morphological variants in text. However, an availability of comprehensive and domain specific dictionaries as well as consistent named entity recognition techniques are preconditions for such approaches.

The work presented here focusses on the technology survey task. Pre-selected biomedical concepts appearing in the documents were tagged using a dictionary-based named entity recognition technique. From the query and retrieval point of view, different query formulation strategies such as the manual query expansion and automatic query expansion (also referred as *semantic search*) have been systematically performed and evaluated.

2. Technology Survey Task

Data used for the Technology Survey (TS) task contains approximately 1.3 million patents from the European, US, and WIPO patent offices as well as nearly 130,000 full-text scientific articles from the PubMed Central (PMC). Six topics that were formulated by human experts as natural language narratives were provided. The task is to retrieve sets of documents from the collection that can best answer the topic questions. An example of a TS topic is:

Topic: *TS-29*
Title: *Inhibitors for acetylcholinesterase*
Narrative:*Acetylcholinesterase inhibitor is a potential target for Alzheimer's disease so identifying potent inhibitors of this human enzyme may lead to new treatments of this devastating disease.*
Chemicals: *Acetylcholinesterase inhibitors*
Conditions: *Alzheimer's disease*

Every TS topic contains a title, a narrative text of the information needed, and a separate indication of chemicals or conditions that the topic is looking for.

2.1. Data Preprocessing

The TREC collection was provided in the Extensible Markup Language (XML). As a preliminary step, an

[1]Chemical Abstract Service, `http://www.cas.org/expertise/cascontent/registry/regsys.html`
[2]International Union for Pure and Applied Chemistry, `http://www.iupac.org/`
[3]`http://drugbank.ca/`

```
Processing 00000000.tx.1: bacterial infection

Phrase: "bacterial infection"
>>>>> Phrase
bacterial infection
<<<<< Phrase
>>>>> Candidates
Meta Candidates (9): [Disease or Syndrome]
  1000 Bacterial Infection (Bacterial Infections)
```

Figure 1: Example of an arbitrary text mapped to UMLS concept by the MetaMap program.

analysis of different sections or zones within the patents and PMC articles was performed. Patent documents contain several fields that are presumably not necessary during retrieval and generate substantial noise while processing the documents. Examples of such fields are country, legal-status, non-English abstracts, etc. Similar examples within PMC articles are the affiliations, citations, editor, etc. The aim was to use only those fields that have high text/noise ratio and that encompass rich information content. Therefore, from a retrieval point of view, the following fields were chosen to be used for indexing and further assessments:

Patents: UCID, Publication date, Authors, Citations, IPC[4] class, Title, Abstract, Description, and Claims.

PMC: DOI[5], Publication date, Authors, Title, Abstract, Article body (front), and Article body (back).

2.2. Concept Identification in TS Topics

For the identification of concepts in TS topics, the MetaMap program was used. MetaMap is a publicly available tool that maps any arbitrary text to biomedical concepts in the UMLS[6] metathesaurus [(Aronson, 2001)]. Figure 1 shows an illustration of text-to-concept mapping performed by the MetaMap program. MetaMap was strictly applied to title, chemical, and condition sections of all TS topics.

Although the UMLS is a comprehensive terminological resource containing over 2 million concepts, it has been shown to lack several biomedical concepts [(Gurulingappa et al., 2009a)]. Elements in TS topics that could not be mapped to UMLS such as *DNA-based asymmetric catalysis*, *Asymmetric catalysis*, *hydrophobic amino acid*, and *endogenous phospholipid* were used as-is and stored for further processing. Constraints were applied on the MetaMap to restrict the semantic classes of mapped concepts to *chemicals and drugs*, *physiology*, and *disorders*. A threshold of 950 was applied for the confidence score of mapping in order to be accepted as a valid concept. During the concept mapping process, the MetaMap also indicates the source vocabularies from which concepts are derived from. Therefore, if a concept exists in the MeSH[7] hierarchy, its hyponym concepts (also known as *child concepts*) and their synonyms were extracted from MeSH. For example, the concept *Bacterial Infection* that appears in TS-28 co-exists in UMLS and MeSH. Since MeSH is hierarchically organized, it provides different hyponyms of *bacterial infections*. Figure 2 shows an illustration of hierarchical structure of MeSH from which the hyponym concepts were extracted.

Figure 2: An example of hyponyms of a concept *Bacterial Infection* in MeSH.

2.3. Concept Tagging in TREC Collection

Concepts obtained from TS topics and their hyponyms and synonyms were used to generate a dictionary of TS concepts. The dictionary contains 16 concepts obtained from 6 TS topics where 12 concepts were generated from automatic mapping and the remaining four concepts were extracted from topic annotations (e.g. the field *chemicals* of TS Topics). ProMiner [(Hanisch et al., 2005)], a dictionary-based named entity recognition system was applied for tagging the TS concepts in the patent and PMC collection. In patents, the *title*, *abstract*, *description*, and *claims* sections were tagged by ProMiner. In PMC, the *title*, *abstract*, and *body* sections of documents were tagged by ProMiner.

2.4. Document Indexing

SCAIView [(Hofmann-Apitius et al., 2008)] is a high performing and scalable Information Retrieval (IR) system based on Lucene [(Hatcher and Gospodnetic, 2004)]. It provides a framework for indexing several gigabytes of document data and to quickly perform complex searches over text as well as named entities. Free-text in the form of stemmed tokens appearing in *title*, *abstract*, *claims*, and *description* sections of patents were indexed. Meta data such as *publication date*, *assignee*, etc. were indexed as-is. Concepts occurring in *title*, *abstract*, *claims* of patents were merged and indexed as a separate field (referred to as CONCEPT-PAT-TAC). Concepts appearing in *description* section of patents were separately indexed (referred to as CONCEPT-PAT-DESC). Similarly, free-text in the form of stemmed tokens appearing in *title*, *abstract*, and *body* sections of PMC were indexed. The concepts occurring in *title* and *abstract* were merged and indexed (referred to as

[4]International Patent Classification, http://www.wipo.int/classifications/ipc/en/

[5]Digital Object Identifier, http://www.doi.org/

[6]Unified Medical Language System, http://www.nlm.nih.gov/research/umls/

[7]Medical Subject Headings, http://www.nlm.nih.gov/mesh/

Table 1: Counts of number of concepts occurring in patents and PMC, and number of documents containing at least one TS concept.

	Patent	PMC
No. of concepts	1,338,348	75,366
No. of documents	181,136	19,138

CONCEPT-PMC-TA) whereas, the concepts in body sections were indexed separately (referred to as CONCEPT-PMC-DESC). Counts of concept occurrences and documents containing at least one TS concept is shown in Table 1.

2.5. Query and Retrieval

For the 2011 TS task, altogether 4 runs were submitted based on different query formulation strategies. Lucene BM25F[8] was applied as a scoring function to measure the similarity between documents and the query. IPCCAT[9], a publicly available tool for the prediction of IPC classes for any input arbitrary text was applied to determine the potential IPC classes of TS topics. *Title* and *Narrative* sections of TS topics were used for the IPC prediction. Nevertheless, the information about IPC helps only in the retrieval of patents. Descriptions of runs submitted and the underlying query formulation strategies are discussed in the following subsections.

2.5.1. SCAIRUN1

SCAIRUN1 serves as a baseline run where queries were formed by manual extraction of key terms from the topic questions. Queries were formulated in a way to reflect knowledge-based human queries. This run provides a rationale for the comparison of performances of semantic and ontological searches against knowledge-based human searches. IPC information obtained from IPCCAT was supplied along with the queries.

2.5.2. SCAIRUN2

SCAIRUN2 applies semantic search strategy to search within the concept space of document index. Concepts extracted from the TS topics were searched against concepts indexed in CONCEPT-PAT-TAC, CONCEPT-PAT-DESC, CONCEPT-PMC-TA, and CONCEPT-PMC-BODY sections of the index. IPC information obtained from the IPCCAT was applied during querying.

2.5.3. SCAIRUN3

Search strategy applied during SCAIRUN3 was same as SCAIRUN2. Based on experiences from TRECCHEM 2010, exploiting the information about patent citations drastically improved the retrieval outcome in the PA task (Gurulingappa et al., 2010). Therefore, the outcome of search was systematically enriched with the patent citations information based on the co-citation ranking scheme defined by (Gurulingappa et al., 2010).

[8] http://nlp.uned.es/~jperezi/Lucene-BM25/
[9] https://www3.wipo.int/ipccat/

Table 2: Results of runs submitted to the TS task.

	AP	nDCG
SCAIRUN1	0.1241	0.3639
SCAIRUN2	0.2851	0.4888
SCAIRUN3	0.2611	0.4843
SCAIRUN4	0.1851	0.4713
SCAIRUN5 (*unofficial*)	0.2529	0.5535

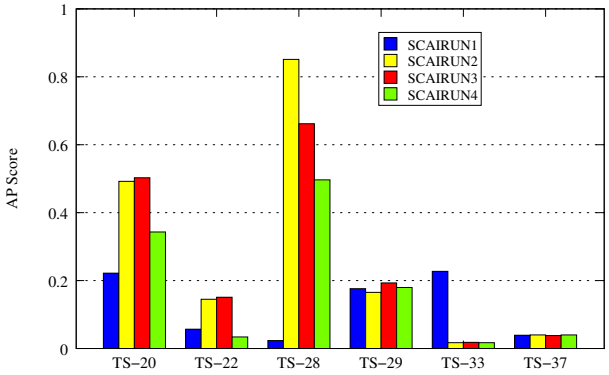

Figure 3: Results of runs for different TS topics.

2.5.4. SCAIRUN4

SCAIRUN4 is a combination of retrieval results of SCAIRUN1 and SCAIRUN2. Results were merged based on the BM25 scores. In case of overlapping documents i.e. if a document occurs in both SCAIRUN1 and SCAIRUN2, its maximum score amongst both runs was used. SCAIRUN4 roughly designed a run where concepts and keywords are applied for querying.

2.5.5. SCAIRUN5

This is not an officially submitted run. SCAIRUN5 is a combination of retrieved documents from runs SCAIRUN1, and SCAIRUN2. If a $Document$ occurs in more than one run, its final score was computed using $\sum \frac{BM25F(Document_i)}{Rank(Document_i)}$ where i indicates the run. Using this equation, a consensus ranked list of documents was generated in terms of decreasing scores.

3. Results and Discussion

3.1. Performance Evaluation

In information retrieval, along with the relevance of the retrieved documents, the order in which they are presented is important. For example, a system that returns maximum relevant documents within top N documents is worthier than the system that returns maximum relevant documents within middle N documents. Therefore, performances of the submitted runs were evaluated using the Average Precision (AP) and Normalized Discounted Cumulative Gain (nDCG) [(Voorhees, 2000)].

3.2. Results of the TS Task

For the TS task, the reported results are based on the AP and nDCG scores. Table 2 shows the results of official runs submitted for this task. Figure 3 shows results of runs over different TS topics.

Considering the overall results of TS task, SCAIRUN2 based on semantic search in the concept space outperformed rest of the runs with the best AP score of 0.2851. This shows the importance and advantages of searching at the concept level in comparison to the conventional text-based search. SCAIRUN3 that applied co-citation based document ranking performed competitively to other runs but its performance was lower than SCAIRUN2. This indicates that utilizing the citation information for technology survey search may not always be successful. A simple combination of concept-based search and manual text-based search (SCAIRUN4) reported poor performance although the results were better than searching with only manual text queries. A strategic combination of searches based on concepts and textual queries (SCAIRUN5) yielded superior results in terms of nDCG score.

Considering the topic-wise results from Figure 3, for topics TS-20, TS-22 and TS-28, search results with concepts showed superior results. Whereas for TS-29 and especially TS-33, the concept-based search strategy failed. One potential reason for the failure of SCAIRUN2 in TS-33 is because *respiratory tract disorder* in MeSH has 218 child concepts (hyponyms) and it could be fuzzy for human evaluators to accept certain hypernym/hyponym concepts for relevancy judgement.

4. Conclusions

This work reports on a semantic framework for information retrieval in biomedical-chemistry patents and full-text articles. Indexing with pre-selected concepts, their hyponyms and synonyms allows semantic search in the concept space. In addition, the performance of system with different search strategies has been systematically evaluated. Semantic search in the concept space indicated superior results in comparison to the conventional text-based queries. The result of run with concept search outperformed rest of the runs with best AP score of 0.28.

Currently, the system is indexed with pre-selected concepts that appear in TS topics. Indexing the biomedical and chemistry concepts that appear in complete MeSH or UMLS thesauri makes the system more applicable for general ad-hoc retrieval situations. However, this is not a trivial task since these thesauri contain substantial noise that may hinder the performance of retrieval. Currently, the performance of retrieval has been tested with 6 topics. In the future, it is necessary to evaluate the system using more questions. This minimizes the deviation of results from the standard average and gives a better estimation of actual system's performance. Document clustering is another topic that authors aim to explore during future experiments.

The system with comprehensively indexed biomedical and chemical terminologies is believed to substantially enhance the search performance. This helps domain experts and patent searchers to carry out prior art searches, invalidity searches, and technology survey searches more efficiently.

5. Acknowledgements

This work is partly funded by the B-IT Research School Fellowship Programme within the NRW state of Germany.

We thank our colleague Heinz-Theodor Mevissen for his valuable contribution to this work.

6. References

A. R. Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proc AMIA Symp*, pages 17–21.

Harsha Gurulingappa, Corinna Kolrik, Martin Hofmann-Apitius, and Juliane Fluck. 2009a. Concept-based semi-automatic classification of drugs. *J Chem Inf Model*, 49(8):1986–1992, Aug.

Harsha Gurulingappa, Bernd Mueller, Roman Klinger, Heinz-Theo Mevissen, Martin Hofmann-Apitius, Juliane Fluck, and Christoph M. Friedrich. 2009b. Patent retrieval in chemistry based on semantically tagged named entities. In *The Eighteenth Text RETrieval Conference (TREC 2009) Proceedings*.

H. Gurulingappa, B. Mueller, R. Klinger, H.-T. Mevissen, M. Hofmann-Apitius, J. Fluck, and C. M. Friedrich. 2010. Prior art search in chemistry patents based on semantic concepts and co-citation analysis. In *The Nineteenth Text RETrieval Conference (TREC 2010) Proceedings*.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. 2005. Prominer: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S14.

Erik Hatcher and Otis Gospodnetic. 2004. *Lucene in Action*. Manning Publications Co.

Martin Hofmann-Apitius, Juliane Fluck, Laura Furlong, Oriol Fornes, Corinna Kolrik, Susanne Hanser, Martin Boeker, Stefan Schulz, Ferran Sanz, Roman Klinger, Theo Mevissen, Tobias Gattermayer, Baldo Oliva, and Christoph M. Friedrich. 2008. Knowledge environments representing molecular entities for the virtual physiological human. *Philos Transact A Math Phys Eng Sci*, 366(1878):3091–3110, Sep.

Sérgio Matos, Joel P. Arrais, Joao Maia-Rodrigues, and Jose Luis Oliveira. 2010. Concept-based query expansion for retrieving gene related publications from medline. *BMC Bioinformatics*, 11:212.

Ellen Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716.