Prior Art Search in Chemistry Patents based on Semantic Concepts and Co-Citation Analysis

Harsha Gurulingappa[1,2], Bernd Müller[1,2], Roman Klinger[1], Heinz-Theodor Mevissen[1], Martin Hofmann-Apitius[1,2], Christoph M. Friedrich[1,3], and Juliane Fluck[1]

[1]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2]Bonn-Aachen International Centre for Information Technology (B-IT)
Dahlmannstraße 2, 53113 Bonn, Germany
[3]University of Applied Science and Arts Dortmund
Department of Computer Science, 44227 Dortmund, Germany

Abstract

Prior Art Search is a task of querying and retrieving the patents in order to uncover any knowledge existing prior to the inventor's question or invention at hand. For addressing this task, we present a contemporary approach that has been evaluated during TRECCHEM for its ability to adapt to text containing chemistry-based information. The core of the framework is an index of 1.3 million chemistry patents provided as a data set by TRECCHEM. For the prior art search task, the information of normalized noun phrases, biomedical and chemical entities are added to the full text index. Altogether, 7 runs were submitted for this task that were based on automatic querying with tokens, noun phrases and entities. In addition, the co-citation information was exploited in a systematic way to generate ranked citation sets from the retrieved documents. Querying with noun phrases and entities coupled with co-citation based post-processing performed considerably well with the best MAP score of 0.23.

1 Introduction

Automatic processing of chemistry literature is challenging due to the existence of different representations of chemical name mentions such as trivial names, IUPAC[1], brand names, InChI[2], and SMILES[3] [10]. For example, the drug name "Aspirin" is reported to have 25 synonyms and 95 brand names in DrugBank[4]. In order to address this challenge, TREC provides a workbench for large scale evaluation and comparison of different techniques for text retrieval in Chemistry. TREC-CHEM addresses this challenge in terms of a trier namely prior art search. Here, a test set of 1000 patents is provided and the task is to retrieve sets of documents from a large patent corpus that can invalidate each test patent.

Considering the ambiguity inherent to the chemistry-based literature, our approach focused on tagging the chemical and biomedical named entities in the documents. Tagging the entities and mapping them to standard database entries normalizes different forms of the same entity to one standard form. This helps to overcome the problems associated

[1]International Union of Pure and Applied Chemistry
[2]International Chemical Identifier
[3]Simplified Molecular Input Line Entry Specification
[4]`http://www.drugbank.ca/`

with multiple synonyms, acronyms and morphological variants in text. Moreover, document retrieval based on semantically tagged entities has demonstrated variable success in the past [12, 13]. A precondition for such an approach is the availability of named entity recognition techniques partly relying on comprehensive domain specific terminologies. Since the entities in chemistry patent space are not as well explored as in biomedical space, we propose to tag the noun phrases and normalize them to their canonical form before further assessments. From the querying and retrieval point of view, the performance of retrieval using tokens, noun phrases, and entities has been evaluated. Additionally, a strategy for post-processing of the citations of the retrieved documents is proposed and evaluated.

2 Prior Art Search Task

The data provided for the Prior Art (PA) search task contains approximately 1.3 million patents from the European Patent Office[5] (EPO), the US Patent and Trademark Office[6] (USPTO), the World Intellectual Property Organization[7] (WIPO) as well as 1000 test (query) patent applications. The task is to retrieve sets of documents from the patent corpus that can invalidate each test patent application. An example of such a task is *"PA-1: Find all patents in the collection that would potentially be able to invalidate US-6090800-A"*.

2.1 Data Preprocessing

The TREC corpus collection was provided in Extensible Markup Language (XML). As a preliminary measure, an analysis of different sections within the patents 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*, or *non-English abstracts*. The aim was to use only those fields that have high text-to-noise ratio and that encompass rich information content. Therefore, with a retrieval point of view, the following fields were chosen to be used for indexing and further assessments: UCID[8], publication date, priority date(s), patent citation(s), inventor(s), assignee(s), author(s), IPC[9] class, title, abstract, description, and claim(s).

2.2 Named Entity Recognition

A preliminary analysis of the IPC classes showed that a large portion of the corpus belongs to A61 (Medical and Veternary Science) and C07 (Organic Chemistry). The hypothesis is that named entity recognition of chemicals and biomedical terms helps to overcome the problems associated with synonyms by automatic query expansion. ProMiner was used for the task of named entity recognition in the title, abstract, claims and description sections of all the patents. The following classes of entities were used for tagging:

Chemical Names Chemical names including synonyms, formulae, IUPAC, and brand names of chemical compounds as extracted from DrugBank, KEGG[10] Drug and KEGG Compound databases. Additionally, a machine learning-based system[9] was applied for tagging the **IUPAC-like** names. It performs an internal normalization to map different variants to one base form.

Genes/Proteins Human genes and protein names as well as their synonyms that are extracted from EntrezGene[11] and UniProt[12] [6].

Diseases Disease names and their synonyms that are extracted from the Medical Subject Headings[13] (MeSH).

Pharma Terms Pharmacological terms that are extracted from the Anatomical Therapeutic

[5] http://www.epo.org/
[6] http://www.uspto.gov/
[7] http://www.wipo.int/portal/index.html.en
[8] User Reference Identifier
[9] International Patent Classification
[10] http://www.genome.jp/kegg/
[11] http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
[12] http://www.uniprot.org/
[13] http://www.nlm.nih.gov/mesh/

Chemical (ATC)[14] drug classification system. Since the ATC does not contain synonyms and term variants, this information was gathered from UMLS with the help of the MetaMap program [11].

Noun Phrases The OpenNLP-based NP chunker[15] was applied for tagging the noun phrases. Non-informative noun phrases were filtered off in a systematic way [5]. Examples of informative and non-informative noun phrases can be found in table 1. The remaining noun phrases were normalized using the LVG Norm program [1] provided within the Specialist NLP package by National Library of Medicine (NLM).

2.3 Indexing

Following the data preprocessing and name entity recognition, the document texts as well as the biomedical entities, chemical entities, and noun phrases occurring within them were indexed with SCAIView [7]. Figure 1 shows an overview of the workflow implemented for the PA task. Unlike a conventional index that contains only tokens, the used index additionally contains noun phrases, chemicals and biomedical entities. Table 2 shows the frequency of different entities occurring in the entire corpus as well as the number of documents that contain at least one entity of interest.

2.4 Querying and Retrieval

Altogether, 7 runs were submitted for the prior art search task. The queries were performed using different entity types occurring in the query documents. Based on the experiences from the previous TREC task, only the complete document searches were performed and the 4-digit IPC information was utilized. The documents were retrieved and ranked based on the Lucene-BM25 function[8] with the default parameters. Different objects that were used for querying are:

[14] http://www.genome.jp/kegg/brite.html
[15] http://opennlp.sourceforge.net/projects.html

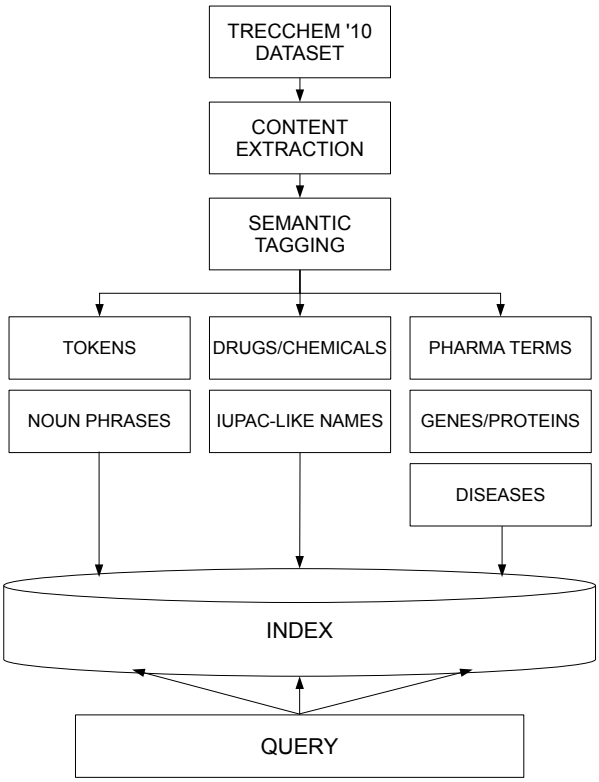

Figure 1: Overview of the workflow implemented for prior art search task.

Tokens: Search with all tokens that occur in a query patent

Noun Phrases: Search with all noun phrases that occur in a query patent

Entities: Search with all chemical entities (chemical names and IUPAC-like) and biomedical entities (pharma terms, genes/proteins and diseases) that occur in a query patent.

The retrieved documents were filtered based on the following criteria:

Priority date: The earliest priority date of the retrieved document must be older than the earliest priority date of the query document.

Informative Noun Phrases	Non-informative Noun Phrases
curable composition	1 2 3 1 2 m 4 R=H
methoxypropynyl group	the claims
biodegradable collagen	about 1800 mg/kg
self-adhesive CODAL tape	A)1>[M M]/(4 [M M] [M M])
tyrosine kinase inhibitor	such difficulties

Table 1: Examples of extracted noun phrases.

	No. of unique entities		No. of documents with one or more entities	
Entity Class	Large Corpus	Query Corpus	Large Corpus	Query Corpus
Chemical Names	12,296	2,467	1,151,477	999
IUPAC-like	2,656,128	18,374	283,677	484
Pharma Terms	479	232	725,325	915
Genes/Proteins	18,641	1,132	883,333	478
Diseases	4,222	833	565,763	336
Noun Phrases	10,158,177	167,851	1,276,229	1000

Table 2: Frequencies of dictionary entries occurring within the the large corpus as well as the query corpus and numbers of documents containing at least one entity of interest.

Family: The retrieved document and the query document must not belong to the same family.

Assignee: The retrieved document and the query document must not have the same assignee and title.

2.4.1 Co-Citaton Analysis

The experiences from 2009 TREC task showed that utilizing the citation information can boost the results [4]. Therefore, a new strategy was applied for the systematic utilization of citations of the retrieved documents. The applied strategy generates a ranked set of patents compiled from the citations of the retrieved documents for each query patent. Figure 2 shows the workflow adopted for co-citation analysis. From the set of retrieved document $\mathcal{D} = \{D_1, \ldots, D_n\}$ (where we use $n = 1000$ throughout this work) a set of citations $\mathcal{C}$ is generated. Let $c(D_j)$ denote the citation of D_j. Then the set of citations is $\mathcal{C} = \{c(D_1), \ldots, c(D_n)\}$. Analogously, $c^{-1}(D_j)$ denotes the set of all documents citing D_j. The function $\text{rank}(D_j)$ returns the rank of D_j with respect to BM25. Then, the co-citation score of a document D_i is

$$\text{co-citation score}(\mathrm{D_i}) = \sum_{D_k \in c^{-1}(D_i)} \frac{\mathrm{BM25(D_k)}}{\mathrm{rank(D_k)}}\,.$$

This is performed for all documents in $\mathcal{C}$. As an outcome of this post-processing strategy, the system returns sets of ranked patents complied from citations of the retrieved patents for a given query patent that can potentially invalidate it.

3 Results and Discussion

For the PA task, the reported results are based on the Binary Preference (bpref) and Mean Average Precision (MAP) scores [2]. Table 3 shows the results of retrieval using tokens, noun phrases and entities. The run with noun phrase queries outperformed the run with token queries with a boost in MAP score by 0.0379. Since the entities do not occur in all query

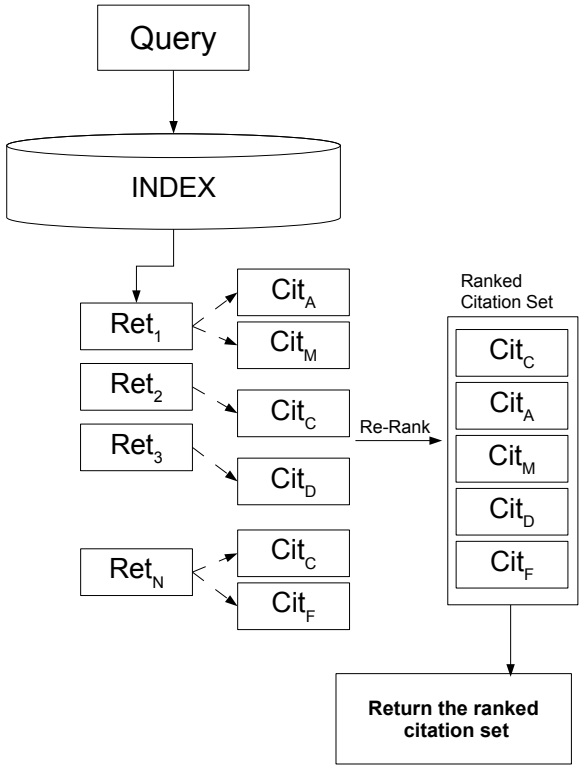

Figure 2: Workflow implemented for co-citation analysis. Cit_A, Cit_C, etc. are citations of the retrieved documents Ret_1, Ret_2, etc.

documents they were coupled with noun phrases and used for querying. A run with the combination of noun phrase and entity queries performed better than the run with the noun phrase queries alone with an improvement in the MAP score by 0.0114. In order to test the significance of using entities for querying, a paired t-test [3] was performed using the results of noun phrase queries and combined noun phrase and entity queries. A p-value lower than 0.0001 indicated that using the entities in combination with noun phrases can have a significant impact on the retrieval. Similary, a p-value lower than 0.0001 indicated that using the noun phrases for querying can significantly outperform token-based querying.

Table 4 shows the results of co-citation based document ranking with tokens, noun phrases and entities used as queries. In comparison to the baseline results, the performance of the system improved by nearly 4 folds. When the priority date filter was turned off, the co-citation based post-processing with noun phrase and entity queries yielded the MAP score of 0.4121 and bpref score of 0.7075 (run id: SCAI10CIENTP). Nevertheless, using the patents that have priority date later than the query patent makes the model unrealistic.

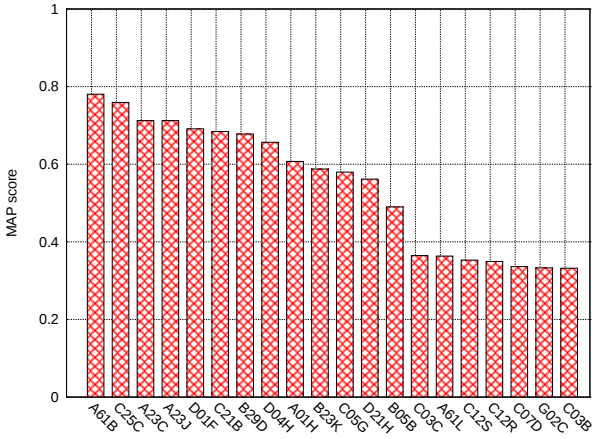

Figure 3: Average MAP scores achieved by the top 20 IPC classes of test patents.

In addition, the co-citation network based document re-ranking strategy proposed by Gobeill et al. [4] was tested. Querying with noun phrases and entities coupled with the post-processing as proposed by Gobeill et al. resulted in the MAP score of 0.1420 and bpref score of 0.5700. Therefore, the post-processing strategy implemented within this work resulted in a MAP score better than the proposed state-of-the-art strategy with a slight decrease in the bpref score.

The best result obtained by the run SCAI10CITENT was analysed based on the different IPC classes. Figure 3 and Figure 4 show the average MAP and bpref scores achieved by the top 20 IPC classes of query patents respectively. Analysis of Figure 3 shows that the best MAP scores are achieved by the test patent that belong to the IPC class A61B

Query Type	Run ID	MAP	bpref
Tokens	SCAI10NRMTOK	0.0172	0.1536
NP	SCAI10NRMNP	0.0551	0.3702
NP + Ent	SCAI10NRMENT	0.0665	0.4171

Table 3: Results of baseline runs with tokens, noun phrases (NP) and entities (Ent) used as queries.

Query Type	Run ID	MAP	bpref
Tokens	SCAI10CITTOK	0.0947	0.2804
NP	SCAI10CITNP	0.2065	0.5110
NP + Ent	SCAI10CITENT	0.2336	0.5468

Table 4: Results of runs with tokens, noun phrases (NP) and entities (Ent) used as queries and co-citation based post-processing.

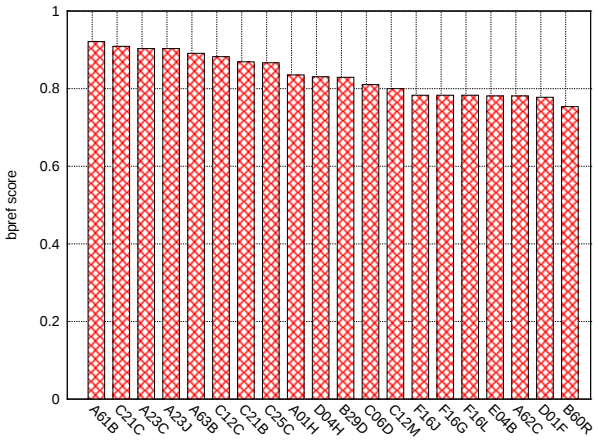

Figure 4: Average bpref scores achieved by the top 20 IPC classes of test patents.

(DIAGNOSIS; SURGERY; IDENTIFICATION) followed by C25C (PROCESSES FOR THE ELECTROLYTIC PRODUCTION, RECOVERY OR REFINING OF METALS; APPARATUS THEREOF) and A23C (DAIRY PRODUCTS, *e. g.* MILK, BUTTER, CHEESE; MILK OR CHEESE SUBSTITUTES; MAKING THEREOF). Figure 4 shows that the best bpref scores are acheieved by the test patent that belong to the IPC class A61B, C21C (processing of pig-iron, *e. g.* refining, manufacture of wrought-iron or steel) and A23C. Figure 5 shows the average MAP and bpref scores achieved by the test patents belonging to the different patent offices. Since the citations are used as a gold standard for evaluation and a major portion of TREC dataset is formed by the USPTO patents, this may be one potential reason for achieving the better performance with USPTO patents than EPO or WIPO patents.

Figure 6 shows the differences in MAP scores between noun phrase-based querying (run id: SCAI10NRMNP) and token-based querying (run id: SCAI10NRMTOK). It can be observed that over 60% of the test patents had an observable gain in the MAP score with noun phrase queries. For about 35% of the test patents, using the noun phrases did not show any effect. Whereas for nearly 5% of the test patents, using the noun phrases resulted in a decrease in MAP scores. The test patents that showed an improvement with using the noun phrase queries were analysed with respect to their IPC classes. It

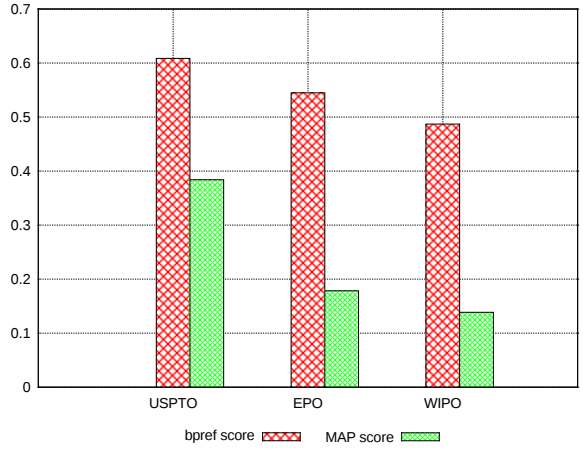

Figure 5: Average MAP and bpref scores achieved by the test patents from different patent offices.

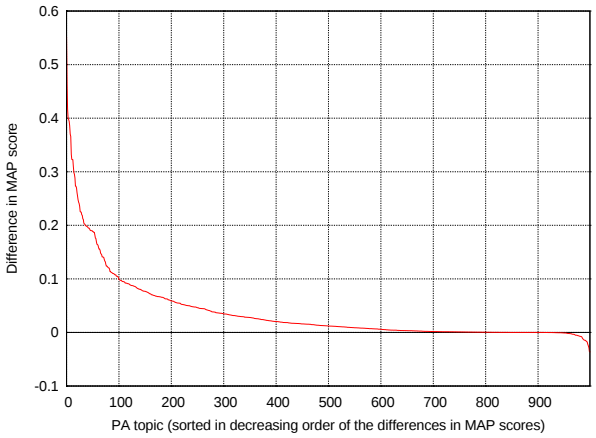

Figure 6: The difference in MAP scores between the runs SCAI10NRMNP and SCAI10NRMTOK. PA-topics are sorted in the decreasing order of the differences in MAP scores.

was observed that a large portion of test patents having an improvement in retrieval belongs to the following IPC classes: A61K (PREPARATIONS FOR MEDICAL, DENTAL, OR TOILET PURPOSES), C07D (HETEROCYCLIC COMPOUNDS) and A61P (SPECIFIC THERAPEUTIC ACTIVITY OF CHEMICAL COMPOUNDS OR MEDICINAL PREPARATIONS).

Figure 7 shows the differences in MAP scores between a combined entity-noun phrase querying (run id: SCAI10NRMENT) and noun phrase-based querying (run id: SCAI10NRMNP). It can be observed that nearly 50% of the test patents had an observable gain in the MAP score with a combined entity-noun phrase querying. Nearly 30% of the test patents had no impact with entities whereas nearly 20% of the test patents showed decrease in the performance. It was observed that a large portion of test patents having an improvement with entity-noun phrase querying belongs to the following IPC classes: A61K, A61P and C08B (POLYSACCHARIDES; DERIVATIVES THEREOF).

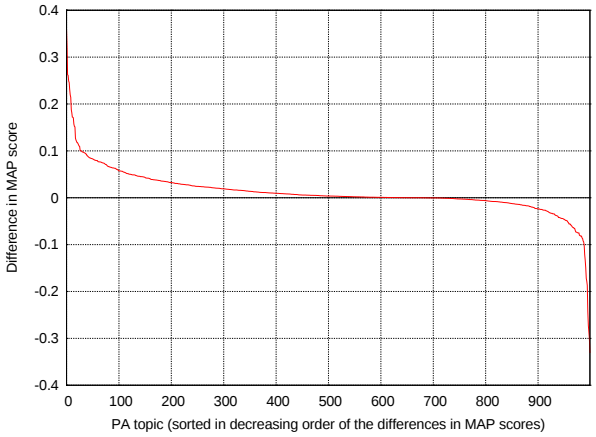

Figure 7: The difference in MAP scores between the runs SCAI10NRMENT and SCAI10NRMNP. PA-topics are sorted in the decreasing order of the differences in MAP scores.

4 Conclusions

For the Prior Art search task, the performance of retrieval using tokens, noun phrases and named enti-

ties as queries has been demonstrated. It was shown that using a combination of noun phrases and entities for querying can perform significantly better than using the tokens or noun phrases alone. The ability of co-citation based post-processing strategy for boosting the performance has been successfully shown. In comparison to state-of-the-art, the performance of adopted co-citation based post-processing has been shown to achieve relatively higher MAP score.

There are several ways to improve the performance of the retrieval. Currently, the breadth of knowledge sources that has been used is limited. For example, only the chemicals present in DrugBank and KEGG databases have been used. These databases are specialized to include the compounds that are of biomedical interest and does not focus on the chemicals present in ink formulations, cement or fertilizers. Considering the scope of IPC classes of the documents provided within the TREC data set, less than 50% of the documents belong to the biomedical domain. Therefore, indexing the entities using broader resources that cover terminologies beyond the biomedical domain has to be tested in future. Improving the recognition performance of the entity recognizers and the noun phrase chunker over patents can also contribute to the better retrieval. The optimization of functions for document retrieval and co-citation based post-processing has to be performed systematically. Therefore, our future work will focus on overcoming the limitations that have been mentioned previously and to optimize the retrieval system to better adapt to the chemistry-based patents.

5 Acknowledgements

This work is partly funded by Bonn-Aachen International Centre for Information Technology (B-IT) Research School within the NRW state. (`http://www.b-it-center.de`)

References

[1] Allen C. Browne, Guy Divita, Alan R. Aronson, and Alexa T. McCray. UMLS language and vocabulary tools. *Proceedings of AMIA Annual Symposium*, page 798, 2003.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *27th annual international ACM SIGIR conference on research and development in information retrieval*, pages 25–32, 2004.

[3] Bradley Efron. Student's t-test under symmetry conditions. *Journal of the American Statistical Association*, 64:1278–1302, 1969.

[4] Julien Gobeill, Douglas Teodoro, E. Patsche, and Patrick Ruch. Report on the TREC 2009 Experiments: Chemical IR Track. In *The Eighteenth Text RETrieval Conference (TREC 2009)*, 2009.

[5] Harsha Gurulingappa, Bernd Müller, Roman Klinger, Heinz-Theodor Mevissen, Martin Hofmann-Apitius, Juliane Fluck, and Christoph M. Friedrich. Patent Retrieval in Chemistry based on semantically tagged Named Entities. In *The Eighteenth Text RETrieval Conference (TREC 2009)*, 2009.

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

[7] Martin Hofmann-Apitius, Juliane Fluck, Laura Furlong, Oriol Fornes, Corinna Kolářik, Susanne Hanser, Martin Boeker, Stefan Schulz, Ferran Sanz, Roman Klinger, Theo Mevissen, Tobias Gattermayer, Baldo Oliva, and Christoph M Friedrich. Knowledge environments representing molecular entities for the virtual physiological human. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 366(1878):3091–3110, Sep 2008.

[8] Pérez-Iglesias Joaquín, Pérez-Agüera José, Fresno Víctor, and Feinstein Z. Yuval. Integrating the Probabilistic Models BM25/BM25F into Lucene. *Computing Research Repository (CoRR)*, 2009.

[9] Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*, 24(13):i268–i276, 2008. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB).

[10] Corinna Kolářik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. Chemical Names: Terminological Resources and Corpora Annotation. In *Workshop on Building and evaluating resources for biomedical text mining*, volume 6th edition of the Language Resources and Evaluation Conference, pages 51–58, Marrakech, Morocco, 2008.

[11] John D. Osborne, Simon Lin, Lihua Zhu, and Warren A. Kibbe. Mining biomedical data using MetaMap Transfer (MMTx) and the Unified Medical Language System (UMLS). *Methods in molecular biology*, 408:153–169, 2007.

[12] Dolf Trieschnigg, Wessel Kraaij, and Martijn Schuemie. Concept Based Document Retrieval for Genomics Literature. In *TREC Genomics Track*, pages 1–11, 2006.

[13] Jay Urbain, Nazli Goharian, and Ophir Frieder. IIT TREC-2007 Genomics Track: Using Concept-based Semantics in Context for Genomics Literature Passage Retrieval. In *TREC Genomics Track*, pages 1–4, 2007.