# Overview of the TREC 2011 Chemical IR Track

Mihai Lupu[1]*, Harsha Gurulingappa[3], Igor Filippov[4], Zhao Jiashu[2],
Juliane Fluck[3], Marc Zimmermann[3], Jimmy Huang[2], John Tait[5]
[1]Vienna University of Technology, Vienna, Austria
[2]Information Retrieval Lab, York University, Toronto, Canada
[3]Fraunhofer SCAI, St. Augustin, Germany
[4]National Institutes of Health, USA
[5]johntait.net Ltd, UK

**Abstract**

The third year of the Chemical IR evaluation track benefitted from the support of many more people interested in the domain, as shown by the number of co-authors of this overview paper. We continued the two tasks we had before, and introduced a new task focused on chemical image recognition. The objective is to gradually move towards systems really useful to the practitioners, and in chemistry, this involves both text and images. The track had a total of 9 groups participating, submitting a total of 36 runs.

## 1 Introduction

The Chemical Track of TREC aims to provide professional searchers with an understanding of the limits of the available tools and, at the same time, to stimulate interest from the research community. In organizing this track, we realized that we were in fact addressing two rather distinct research communities, which, although both focus on the chemical domain, do not usually interact. The text mining and image understanding processes are very different, but both of them are essential for the end user, as chemistry heavily relies on non-textual information.

In this campaign, we started in 2009 with a test, to see how current approaches to IR perform when given a corpus of chemical patents and scientific articles [2, 1]. We observed the limits and in the following year we extended the collection to include more scientific articles and, essential, the accompanying images and structure files for many of them. A large collection, of almost 500GB of compressed text and image data was provided, but no specific task was created to require participants to deal with the image data.

[*] Mihai Lupu is partially supported by the PROMISE NoE, funded under FP7 Grant no. 258191 and the Austrian Research Promotion Agency (FFG) FIT-IT project IMPEx (No. 825846)

We realized that handling both types is impossible for any research group, and, at the Chemical IR workshop organized last year at TREC, we decided to have a specific task to handle images. This task would require participants to, given an image file, provide the chemical structure of the compound there depicted. It was the most successful tasks this year, with 5 groups sending in their results.

## 2 Prior Art Task (PA)

### 2.1 Topics

As in the previous year, the effort this year was to create a balanced selection of topics. Of the total 1000 topics (numbered PA-1001 through PA-2000), 334 were from the European Patent Office, 333 from the US Patents and Trademarks Office and 333 from the World IP Organization. The sources were distributed randomly among the 1000 topics. Consequently, of the small set (PA-1001 through PA-1100) there were 20 from the EPO, 37 from the USPTO and 43 from the WIPO.

### 2.2 Results

This year we received runs from two participants for the Prior-Art Task. Their results are described in Figure 1.

## 3 Technology Survey Task (TS)

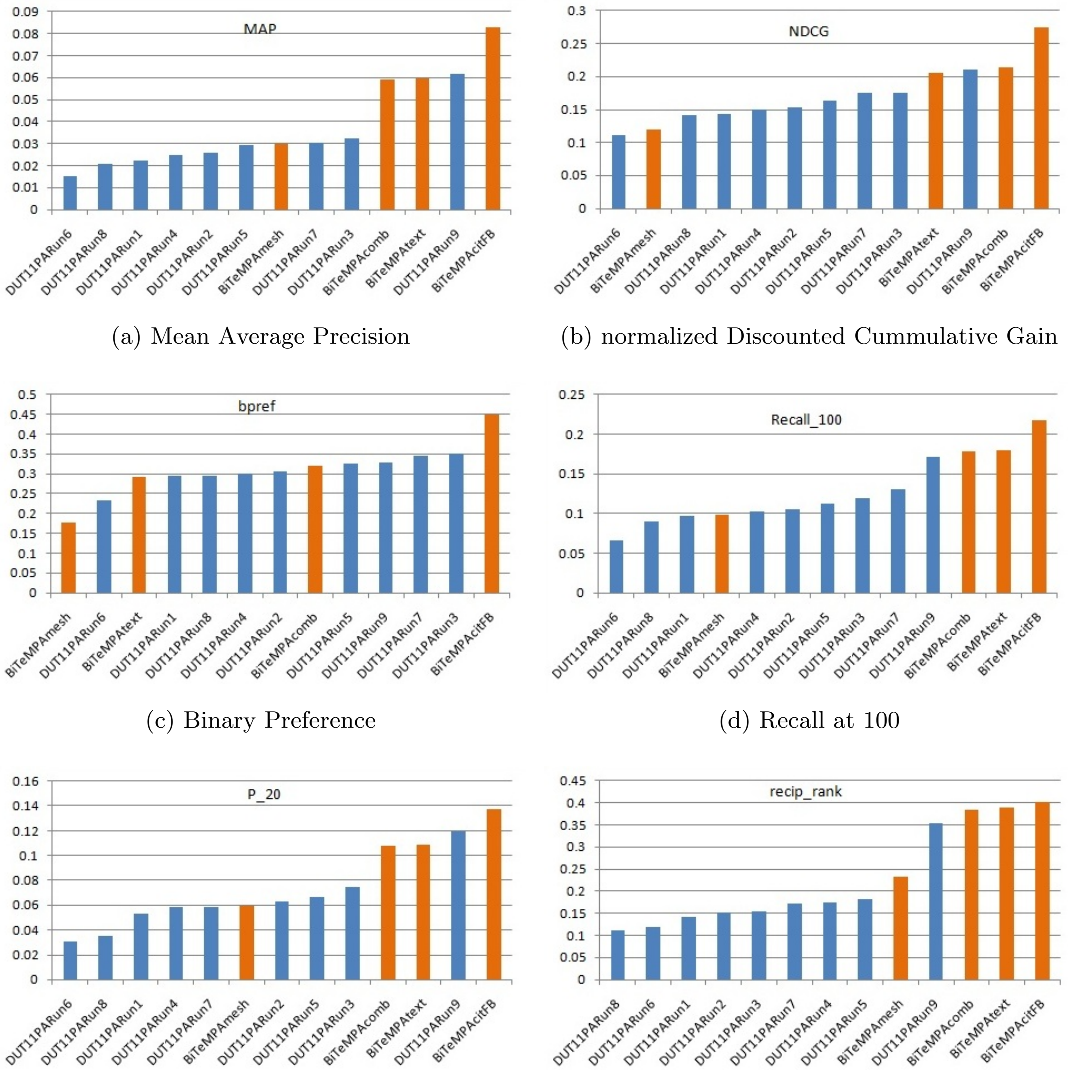
The TS task is a standard ad-hoc retrieval task where the challenge is to retrieve documents from the collection that can best answer the information need expressed in the topic. This year, the TS task focussed on biomedical and pharmaceutical topics, unlike previous years when it dealt with general chemistry topics. The primary motivation was to investigate the ability of state-of-the-art IR systems to deal with the textual diversity in biomedical patents and full-text articles by handling synonyms, abbreviations, naming variants, and many more. The second motivation is the availability of enormous terminological resources in the biomedical domain (such as MeSH[1], UMLS[2] , ATC[3] , and many more) and the necessity to investigate their adaptability to support information retrieval from patents and full-texts that has not been systematically tested in the past (according to author's knowledge). Although, the TREC Genomics track focussed on retrieval of biomedical full-text articles, there is a need for common platform-based evaluation of IR systems for patents.

In order to deal with this challenge, 6 questions concerning the information needs (also called as *TS-topics*) were provided by domain experts. Altogether, 4 teams sub-

[1] Medical Subject Headings, `http://www.nlm.nih.gov/mesh/MBrowser.html`

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

[3] Anatomical Therapeutic Chemical classification system, `http://www.whocc.no/atc_ddd_index/`



(a) Mean Average Precision

(b) normalized Discounted Cummulative Gain

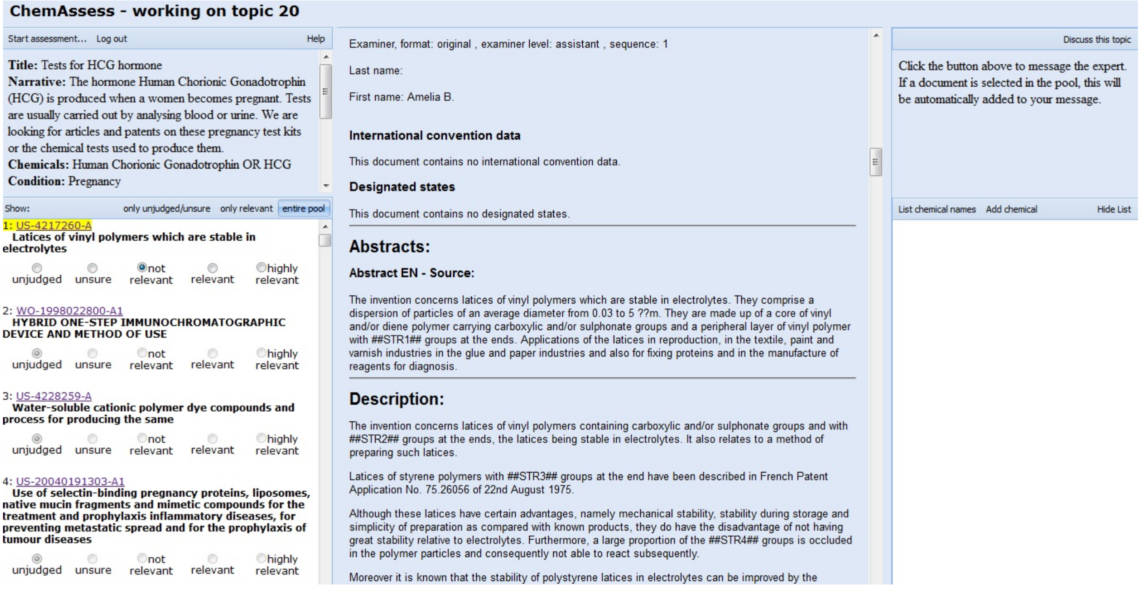
(c) Binary Preference

(d) Recall at 100

(e) Precision at 20

(f) Reciprocal Rank

Figure 1: Results for 6 measures for the Prior-Art task. The two participating groups are identified by different shades in the plots.



ChemAssess - working on topic 20

**Title:** Tests for HCG hormone
**Narrative:** The hormone Human Chorionic Gonadotrophin (HCG) is produced when a women becomes pregnant. Tests are usually carried out by analysing blood or urine. We are looking for articles and patents on these pregnancy test kits or the chemical tests used to produce them.
**Chemicals:** Human Chorionic Gonadotrophin OR HCG
**Condition:** Pregnancy

1: US-4217260-A
**Latices of vinyl polymers which are stable in electrolytes**
○ unjudged ○ unsure ● not relevant ○ relevant ○ highly relevant

2: WO-1998022800-A1
**HYBRID ONE-STEP IMMUNOCHROMATOGRAPHIC DEVICE AND METHOD OF USE**
● unjudged ○ unsure ○ not relevant ○ relevant ○ highly relevant

3: US-4228259-A
**Water-soluble cationic polymer dye compounds and process for producing the same**
● unjudged ○ unsure ○ not relevant ○ relevant ○ highly relevant

4: US-20040191303-A1
**Use of selectin-binding pregnancy proteins, liposomes, native mucin fragments and mimetic compounds for the treatment and prophylaxis inflammatory diseases, for preventing metastatic spread and for the prophylaxis of tumour diseases**
● unjudged ○ unsure ○ not relevant ○ relevant ○ highly relevant

Examiner, format: original , examiner level: assistant , sequence: 1

Last name:

First name: Amelia B.

**International convention data**

This document contains no international convention data.

**Designated states**

This document contains no designated states.

**Abstracts:**

**Abstract EN - Source:**

The invention concerns latices of vinyl polymers which are stable in electrolytes. They comprise a dispersion of particles of an average diameter from 0.03 to 5 ??m. They are made up of a core of vinyl and/or diene polymer carrying carboxylic and/or sulphonate groups and a peripheral layer of vinyl polymer with ##STR1## groups at the ends. Applications of the latices in reproduction, in the textile, paint and varnish industries in the glue and paper industries and also for fixing proteins and in the manufacture of reagents for diagnosis.

**Description:**

The invention concerns latices of vinyl polymers containing carboxylic and/or sulphonate groups and with ##STR2## groups at the ends, the latices being stable in electrolytes. It also relates to a method of preparing such latices.

Latices of styrene polymers with ##STR3## groups at the end have been described in French Patent Application No. 75.26056 of 22nd August 1975.

Although these latices have certain advantages, namely mechanical stability, stability during storage and simplicity of preparation as compared with known products, they do have the disadvantage of not having great stability relative to electrolytes. Furthermore, a large proportion of the ##STR4## groups is occluded in the polymer particles and consequently not able to react subsequently.

Moreover it is known that the stability of polystyrene latices in electrolytes can be improved by the

Click the button above to message the expert. If a document is selected in the pool, this will be automatically added to your message.

Figure 2: Illustration of the user interface of the ChemAssess tool.

mitted 14 runs for this year's task. Information about the topics can be found in the Appendix. Details of the evaluation and results can be found in the following subsections.

## 3.1 Sampling and Relevance Judgement

We employed the stratified sampling approach for generating a pool of documents for evaluation of each topic following the [3] method. We took all results returned in the top 10 by any run, 30% of those in the top 30 and 10% of the rest (down to rank 1000).

As last year, the TS topic evaluation was performed in parallel by junior and senior evaluators. A parallel evaluation is believed to enhance the level of interaction between the evaluators as well as reduce the overall time required for the evaluation process. Senior evaluators designate the patent experts who formulated the TS topics whereas the junior evaluators were graduate students who hold an academic major degree in biotechnology, bioinformatics, or medicine. One junior evaluator and one senior evaluator made judgments for each topic. Therefore, the judges team was composed of 6 junior and 4 senior evaluators each handling strictly one topic. The tool used for the evaluation was ChemAssess[4]. For a given pooled document set for each topic, a junior evaluator makes a judgement as the document is *unjudged*, *unsure*, *not relevant*, *relevant*, or *highly relevant*. The ChemAssess tool does not allow overlapping annotations and therefore each document ends up having only one judgement. Table 1 shows the number of documents contained in pooled document sets for each TS topic. An illustration of the user interface of the ChemAssess tool is shown in Figure 2.

[4] http://chemassess.ifs.tuwien.ac.at/

Table 1: Size of the pooled document sets for each TS topic.

| Topic | No. of documents |
|---|---:|
| TS-20 | 411 |
| TS-22 | 558 |
| TS-28 | 736 |
| TS-29 | 474 |
| TS-33 | 612 |
| TS-37 | 507 |
| Total | 3298 |

Table 2: Counts of comments or discussion passes over different topics by evaluators and track organizers.

| Topic | Junior Evaluator | Senior Evaluator | Organizer |
|---|---:|---:|---:|
| TS-20 | 3 | 2 | 0 |
| TS-22 | 7 | 4 | 0 |
| TS-28 | 6 | 7 | 1 |
| TS-29 | 1 | 1 | 0 |
| TS-33 | 5 | 5 | 0 |
| TS-37 | 0 | 0 | 0 |

For every document in the pool, the junior evaluator first decides the relevancy of a document based on his/her personal knowledge. Junior evaluators were facilitated with a set of biomedical terminological resources and databases in order to understand the synonyms and naming conventions of biomedical entities. For example, the junior evaluators referred to Swissprot for protein names, DrugBank for drug names, MeSH and UMLS for medical terms (such as diseases), and KEGG BRITE[5] for various biological entities such as pharmacological terms, pathways, and many more. In case of uncertainty in the decision, junior evaluators had an option to interact with the senior evaluators or track organizers within the ChemAssess environment. Based on the overall comments, the junior evaluators made the final judgement over a document's relevancy to the respective topic. Table 2 shows the number of comments or discussions passed over different TS topics by the evaluators. Table 3 shows the final results of judgement for every TS topic

## 3.2 Performance Measures

Performances of the submitted runs were evaluated using the evaluator's judgements as a golden standard. As evaluation metrics, inferred average precision (xinfAP) and the inferred normalized discounted cumulative gain (infNDCG) were applied.

[5] http://www.genome.jp/kegg/brite.html

Table 3: Results of final judgements by evaluators for all TS topics. Number of documents (#doc) and percentage of documents (%doc) are provided.

| Topic | #doc | Highly Relevant | | Relevant | | Not Relevant | | Unsure | | Unjudged | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | #doc | %doc | #doc | %doc | #doc | %doc | #doc | %doc | #doc | %doc |
| TS-20 | 411 | 32 | 7.79% | 31 | 7.54% | 348 | 84.67% | 0 | 0.00% | 0 | 0.00% |
| TS-22 | 558 | 1 | 0.18% | 9 | 1.61% | 548 | 98.21% | 0 | 0.00% | 0 | 0.00% |
| TS-28 | 736 | 11 | 1.49% | 3 | 0.41% | 722 | 98.10% | 0 | 0.00% | 0 | 0.00% |
| TS-29 | 474 | 2 | 0.42% | 72 | 15.25% | 397 | 84.11% | 1 | 0.21% | 0 | 0.00% |
| TS-33 | 612 | 51 | 8.33% | 55 | 8.99% | 506 | 82.68% | 0 | 0.00% | 0 | 0.00% |
| TS-37 | 507 | 11 | 2.17% | 18 | 3.55% | 478 | 94.28% | 0 | 0.00% | 0 | 0.00% |

### 3.2.1 xinfAP

In [3], Yilmaz and colleagues extended the infAP measure by taking non-random samples from the pool of documents. We adopted this measure because it appears to estimate AP more accurately than infAP, given the same evaluation effort.

### 3.2.2 infNDCG

Also based on a stratified sampling approach, infNDCG extends nDCG, whose aim is to represent the common view that relevant documents returned higher in the ranked list are more important than similarly relevant documents returned lower in the list.

## 3.3 Results of Technology Survey Task

This year we received runs from 4 participants (for a total of 14 runs). Figure 3 describes the results of the evaluation. The figure shows a cummulative plot for each run, where each task is stacked on top of the others. This way, it is easier to observe how each run perfomed for each topic.

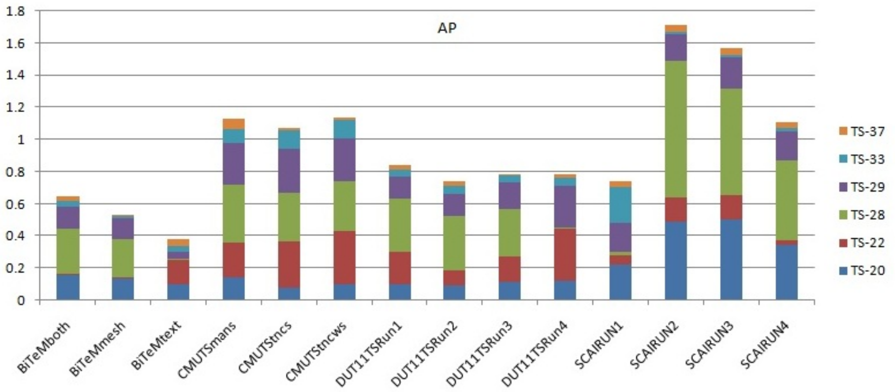
# 4 Image-to-Structure Task (I2S)

The main purposes of this task include evaluation of the state of the art in the chemical image recognition field and the applicability of image recognition for information retrieval goals.

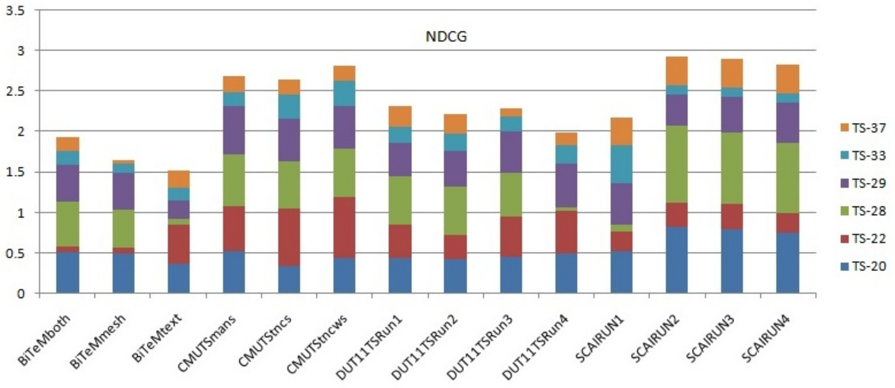
## 4.1 Topics

Two sets each containing 1000 images and the corresponding MOL[6] files have been selected to act as a training and an evaluation sets from the USPTO file collection. The following criteria were used in the selection process:

[6] http://en.wikipedia.org/wiki/Chemical_table_file



(a) Mean Average Precision



(b) normalized Discounted Cummulative Gain

Figure 3: A stack-like plot of the results of each participant in the TS task, for two measures: extended inferred AP and inferred nDCG

- No polymers (brackets in the molecule description), charges or isotopes - judged by the presence of text lines starting with “M” but not “M END” in the ctab block of the MOL file.

- No multi-fragment MOL files, only one molecule per file is allowed.

- Allow only ”organic” elements - C, N, O, S, F, Cl, Br, I, P, and H.

- Check that ctab records for all atoms correspond to the formal charge of 0 and that the isotope type is unspecified (default).

- Check that there are no stereobonds with stereo orientation specifically set to ”undefined”

- Check that the number of heavy (non-Hydrogen) atoms is greater than 6, and the molecular weight is lower that 1000 a.u.

- Check that InChI[7] can be created for the selected molecules.

These criteria allowed the organizers to focus on small organic molecules for which a reasonably widely accepted and well-defined chemoinformatic identity measure exists - namely InChI and InChI key. Those are also the types of molecules believed to be of the most interest to the chemical and pharmaceutical industry. Training set of images and MOL files and the evaluation set of images only have been made available to the participants.

## 4.2 Evaluation

Participatns of the Image2Structure task have been asked to submit the results of their runs in the form of SD files. SD file format is analogous to MOL with the exception that it allows for multiple molecules to be stored in a single file. This difference between the ground truth MOL format and the requested SD format was delebirate to allow for the possibility that recognition software may erroneously generate several output molecules for a single molecule input image. There are many free and commercial software utilities which allow interconversion between alternative formats, such as SMILES[8], and SDF. The runs were evaluated based on a recall measure by matching of the standard InChI keys computed from the original MOL file and the SD file representing recognition software output. Chemical identity is often a subject of ongoing debates among chemists about what constitutes a unique molecule. InChI - IUPAC International Chemical Identifier - is a text representation of a molecule which was designed to compute normalized, canonical text string from the original molecule representation. InChI takes into account certain forms of tautomerism, stereochemistry, etc. InChI key is a hashed version of InChI. Standard InChIs and InChI keys, while not completely free of their share of issues, are widely used as unique chemical identifiers by chemists worldwide. Therefore

[7] http://en.wikipedia.org/wiki/InChI

[8] http://en.wikipedia.org/wiki/Smiles

Table 4: Participants in the Image-to-Structure taks

| **Participating group** | **Run name(s) identifier** |
|---|---|
| University of Birmingham | UoB |
| SAIC-Frederik / NIH | OSRA |
| GGA Software | GGA |
| University of Michigan | ChemReader |
| Fraunhofer SCAI | ChemOCR |

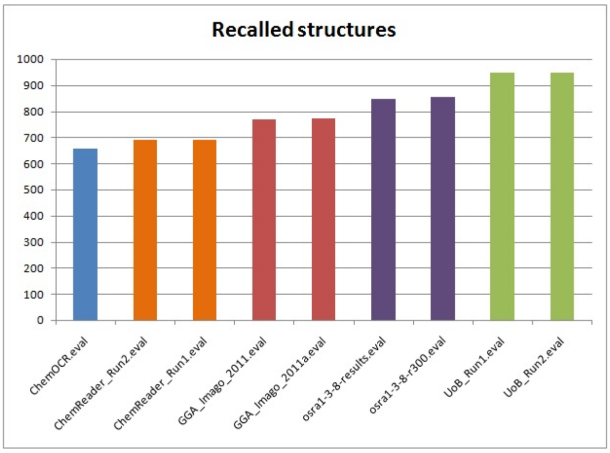


Figure 4: The objective of the Image-to-Structure task was to recognize structures. The plot shows how many correct structures each run identified.

standard InChI keys have been selected for Image2Structure evaluation as a relatively controversy-free chemical identity measure.

### 4.3 Results

This year, the Image-to-Structure task received 11 runs from 5 participants. Overall, results were very good, with all participants recognizing over 60% of the given structure images. Figure 4 shows the results for each run. Table 4 connects the names of the runs in Figure 4 with the participating groups.

## 5 Conclusion

This year it was clear that the Prior Art task had reached its final point. From it, we learned the extent to which we can, in one hit, get relevant documents to a patent application in the chemical domain. We were delighted to see engagement between students and senior evaluators in the Technology Survey task, and, in the following months, look to participants to provide a deep analysis of what worked and what did not work for the bio-chemical domains. We were also happy with the participation in the Image-to-Structure task. After all, there are only a handful of groups who do this professionally, and of all those contacted, only two could not send in results.

## Acknowledgements

We are particularly grateful to all the evaluators of the Technology Survey Task. The senior evaluators are Teresa Loughbrough, Carlos Faerman, Susan McKee and Guangyao Zhao. The student evaluators are Martins Oderinde, Yu Jing, Sameena Sultana Syeda, Mythri Bangalore, Anandhi Iyappa, Ramakanth Reddy Guntuka.

## References

[1] M. Lupu, J. Huang, J. Zhu, and J. Tait. TREC Chemical Information Retrieval - An Evaluation Effort for Chemical IR Systems. *WPI Journal*, 2011.

[2] M. Lupu, F. Piroi, J. Huang, J. Zhu, and J. Tait. Overview of the trec chemical ir track. In *Proc. of TREC*, 2009.

[3] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2008. ACM.

# 6 Appendix

Topics used in the Technology Survey Task of TREC-CHEM 2011.

| Topic | Title | Narrative |
|---|---|---|
| TS-20 | Tests for HCG hormone | The hormone Human Chorionic Gonadotrophin (HCG) is produced when a women becomes pregnant. Tests are usually carried out by analysing blood or urine. We are looking for articles and patents on these pregnancy test kits or the chemical tests used to produce them. |
| TS-22 | Uses of hormones in detection of menopause | The onset of menopause in women is complex and often difficult to detect accurately. We are looking for methods, devices, or kits using the fertility hormones Luteinizing hormone (LH), Follicle stimulating hormone (FSH) for detection of the menopause in women. |
| TS-28 | D-ala-D-ala ligase inhibitors | D-ala-D-ala ligase is a well known antibacterial target and the idea is to find potent inhibitors of this bacterial enzyme which have the ability to kill bacteria. Getting the additional MICs data (usually as tables in the body of a patent) is very important. |
| TS-29 | Inhibitors for acetylcholinesterase | Acetylcholinesterase inhibitors 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 |
| TS-33 | Respiratory tract disorders treatment using inhalation of porous particles containing hydrophobic amino acid and endogenous phospholipids | Find patents that claim use of porous inhalation particles that contain a hydrophobic amino acid (e.g. leucine, isoleucine, phenylalanine, etc.) and endogenous phospholipid(s) (e.g. phosphatidylcholines, phosphatidylethanolamines, etc.) for treatment of respiratory tract disorders. |
| TS-37 | DNA-based asymmetric catalysis | The users are looking for information about DNA-based asymmetric catalysis, especially sample reactions that have been done so far. The novel concept of DNA-based asymmetric catalysis was introduced only five years ago, thus there would not be too many results, most of which are enantioselective Diels-Alder reactions. |