Report on
Workshop on “Information Retrieval seminar series”

By Dr. Karin Friberg Heppin, University of Gothenburt

1. Summary

The spring of 2012 Språkbanken, University of Gothenburg (GU), has made an effort to make GUa platform for Swedish Information Retrieval Research. The main event was an Information Retrieval seminar series, where Professor Kalervo Järvelin held two of the seminars. Professor Järvelin also acted as scientific advisor for the MedEval project, enhancing the Swedish medical test collection MedEval, and making the collection, with user groups doctors and patients, available for researchers. Professor Järvelin was also available for personal discussions for students and researchers. He also was advisor in planning a course in Information Retrieval which will be given for the first time at GU this autumn.

2. Description of scientific content of and discussion at the event

The Swedish medical test collection MedEval is unique in that is contains documents assessed for target groups, that is if they have experts or lay persons as their main audience. The test collection gives the user the possibility to choose user group, doctors or patients, and in doing this a recall base that is adjusted to the user is selected. In the first version of MedEval the recall bases for the two user groups were created by downgrading the relevance score of the documents from the wrong user group. We had long and interesting discussions for the coming version of MedEval on how to create recall bases for the different user groups in a manner that would be more scientifically motivated. These discussions resulted in a decision not to only make assumptions on how valuable different documents are for different groups of users, but to make this part of the evaluation methodology. That is, we have since hired lay assessors which will judge the same document pools as the medically trained assessors.

Professor Järvelin helped formulate the syllabus for the IR course at GU, and we discussed how the plan of the course could be. We decided that the course book would be Croft, Metzler and Strohman, 2010.

Below are the abstracts for the two seminars of Professor Järvelin:

Light Statistical Morphology for IR

Traditional morphological tools seek to treat the morphological variation of a language comprehensively. While the results tend to be good, at least linguistically, the down side is complexity of construction, maintenance and use of such tools. During the past few years, several statistical methods for morphological processing have been proposed for use in Information Retrieval (IR). These methods may be characterized as unsupervised, semi-supervised, light-weight, and/or (partially) language-independent. Generally, they are easy to set up for a new language or collection and provide competitive results for treating morphological variation in IR. The talk introduces some generative and reductive light statistical methods for use in IR and discusses their effectiveness and limitations.

Managing Morphologically Complex Languages in Information Retrieval

Morphologically complex languages are rich in inflectional and derivational morphology and compound formation. In contrast to languages like English, Hindi or Chinese, complex languages may offer tens or even hundreds of inflectional forms for nouns, for instance. Such variation is a challenge to Information Retrieval (IR) methods that are based on matching keywords to text indexes. The talk discusses reductive (such as stemming and lemmatization) and generative (inflectional stem and full word form generation) methods for IR in several languages, covering both index construction and query processing. Emphasis is given to IR effectiveness and the contribution of morphological processing in this.

3. Assessment of the results and impact of the event on the future direction of the field

The seminars of Professor Järvelin were very appreciated and attended by many researchers and students. The advice he gave: personal, for the MedEval Project and for the Information Retrieval course was invaluable. We hope that this effort will establish Information Retreival as a research field at GU. The MedEval test collection is a valuable resource for this.

Professor Järvelin helped plan the IR course at GU and helped finalize the course syllabus. Our hope is that the seminar series and the IR course will have inspired some of the students at the Masters' programme of Language Technology at GU and that there will be some students which choose to write their master's thesis about Information Retrieval, preferably using the enhanced test collection, MedEval.

The discussions about the assessment of the MedEval documents and the creation of recall bases for different user groups resulted in a decision to hire assessors without medical training in addition to the previous assessors which all had around 3 years of medical studies behind them. This is to see how the judgements really differ, if they do, from the judgement of the medically trained.

The focus put on IR by GU this spring, with the seminar series and visit by Professor Järvelin as highlight, has paid off in an obvious way as GU will take over as the Swedish partner in the Promise Network of Excellence, July 1 2012.

4. Final programme of the meeting

The first days of the visit Professor Järvelin was available for discussions and meetings. Thursday May 10 and Friday May 11 he held his seminars which were part of the larger seminar series on Information Retreival. In more detail the programme looked like this:

May 7-8	Professor Kalervo Järvelin will support the planning of the Course in Information Retrieval and Give advise on future studies using MedEval, the Swedish medical test collection built at the Univesity of Gothenburg.

May 9	Professor Kalervo Järvelin will be available for individual discussions on IR.
May 10	CLT seminar – main audience senior researchers.
May 11	Seminar - main audience student at the Master Programme of Language Technology.

Information Retrieval Seminar Series 2012

Friday March 23. *Peter Ingwersen*: **Basic conceptual models for Information Retrieval research**

Friday April 13. *Anni Järvelin*. **Different approaches to translation in cross-language information retrieval: What to translate and how.**

Friday April 20. *Anni Järvelin and Karin Friberg Heppin*: **Information retrieval evaluation**

Thursday May 3 *Jussi Karlgren*; **Genres on the web – what is a genre anyway?**

Thursday May 10. *Kalervo Järvelin*: **Light Statistical Morphology for IR**

Friday May 11. *Kalervo Järvelin*: **Managing Morphologically Complex Languages in Information Retrieval**