DIR 2010: the tenth Dutch-Belgian Information Retrieval workshop

Wessel Kraaij[*][†] *w.kraaij@cs.ru.nl*
Suzan Verberne[‡] *s.verberne@let.ru.nl*
Max Hinne[*] *mhinne@sci.ru.nl*
Maarten van der Heijden[*] *m.vanderheijden@cs.ru.nl*
Theo van der Weide[*] *tvdw@cs.ru.nl*

Abstract

The Dutch-Belgian Information Retrieval workshop (DIR) is an annual event where the latest research results by Dutch and Belgian IR researchers are presented. The workshop is an important instrument to strengthen and maintain local contacts. This paper reports on the tenth DIR meeting, which took place in January 2010. The DIR 2010 programme included invited talks by Elizabeth D. Liddy and Cornelis (Kees) Koster.

1 Introduction

The 10th issue of the Dutch-Belgian Information Retrieval workshop took place on January 25, 2010 and was hosted by the Information foraging Lab (IFL) — a collaboration between researchers from the Institute of Computing and Information Sciences (iCIS) from the Faculty of Science and the Language and Speech Unit of the Faculty of Arts, both at Radboud University Nijmegen in The Netherlands. The primary aim of the DIR workshops is to provide an meeting place for Dutch and Belgian researchers from the domain of information retrieval and related disciplines, where they can exchange information and present new research developments.

The workshop is organised under the auspices of the Dutch Working Community on Information Sciences (WGI) and the Dutch research school for knowledge and information systems. This year, additional support came from Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO Exacte Wetenschappen), the Taalunie-programme STEVIN and Textkernel.

The first DIR workshop was organized by Jaap van den Herik in 2000 with Kal Järvelin as keynote speaker. Subsequently, DIR was located in Enschede (2001), Leuven (2002), Amsterdam (2003), Utrecht (2005), Delft (2006), Leuven (2007), Maastricht (2008) and

[*] Institute of Computing and Information Sciences, Radboud University Nijmegen, NL
[†] TNO, Delft, NL
[‡] Center for Language and Speech Technology, Radboud University Nijmegen, NL

Enschede (2009). The list of DIR keynote speakers includes distinguished researchers such as Stephen Robertson, Karen Spärck Jones, Charlie Clarke, Mounia Lalmas, Chengxiang Zhai, Maarten de Rijke, Thomas Hoffman, Keith van Rijsbergen, Hinrich Schütze and Gerhard von Weikum. The Information Retrieval community in the Netherlands and Belgium has grown significantly over time, judged by the number of accepted papers of Dutch and Belgian origin in top IR conferences and journals.

Information Retrieval has become a regular component of Dutch Computer Science and Information Science programmes. An important driver for this progress has been the involvement in benchmark conferences such as TREC, CLEF and INEX and the growth of the field as a whole due to the web revolution. The change in character and size of the DIR community has also had its impact on the format of DIR itself. Since researchers primarily target high impact conferences and journals, it became increasingly difficult to solicit high quality original work for the DIR workshop. Therefore, for DIR 2010, it was decided to encourage submissions of "compressed contributions" to present recent original work previously published in important venues by DIR members to the DIR community proper. Judging the number of submissions in this track, this was a success.

There were seventeen submissions of original work and eight submissions under the compressed contributions call. In creating the programme for DIR 2010, the programme committee has given priority to the first category of submissions: The oral sessions consisted of seven original papers and one compressed contribution. Seven original papers and seven compressed contributions were presented in the poster session.

2 Keynote talk: *NLP & IR — Wither We goest?*

Keynote speaker was Elizabeth D. Liddy, dean of the School of Information Studies at Syracuse University and ACM SIGIR chair for the 2007–2010 term. Her talk was entitled "NLP & IR — Wither We Goest?", discussing the mutual influence and cross-fertilization of the fields of Natural Language Processing and Information Retrieval. Both fields have a history of at least 50 years and it is instructive to understand why some levels of NLP were used and others were not used in IR applications at a particular point in time. New insights and more computing possibilities currently make other choices possible. Therefore, re-assessing the why and how of certain levels of NLP that have been put aside for IR systems could offer new avenues for research and a deeper understanding of the nature of IR. Liz gave several examples to illustrate the rich expressiveness of natural language in terms of structured queries as opposed to the 2.5 keyword paradigm. She also argued that NLP techniques such as paraphrasing are not so important for simple high precision web search. The raw size of the web softens the problem of the vocabulary gap. NLP is more likely to have an impact in vertical search applications where redundancy is less prominent and high recall is important. In the second half of the keynote talk Liz discussed the potential of NLP analysis on the level of pragmatics; what is the intended effect of the text, what is the sentiment evoked by a text, how could we enrich an application like CiteSeeer with polarity information; how can we assess the credibility and certainty of information. She finished her presentation by presenting several concrete research challenges involving credibility, certainty and complex high recall queries.

3 Oral sessions

Wouter Weerkamp from the University of Amsterdam opened the first paper session with a presentation based on the compressed contribution titled "A Generative Blog Post Retrieval Model that Uses Query Expansion based on External Collections", which was originally published in the proceedings of the Annual meeting of the ACL, 2009. Allan Hanbury from the Information Retrieval Facility presented work on patent categorization, specifically comparing the performance of Balanced Winnow and SVM. The last presentation in the first session was by Jan de Belder from the University of Leuven, reporting on sentence compression in Dutch and Flemish text.

Eva D'hondt presented work carried out at the Radboud University in the project 'Text Mining for Intellectual Property'. She explained the challenges of patent retrieval and presented a re-ranking approach based on syntactic dependencies on top of general bag-of-words retrieval. Marieke van Erp presented interesting work on her PhD research project, which is a collaboration between Tilburg University and the Dutch National Museum of Natural History. Her information retrieval system utilises domain knowledge from an ontology and external resources to aid retrieval of records from two natural history databases. Vera Hollink from the Centrum Wiskunde & Informatica (CWI) presented a paper on query modification analysis: how do users of search system modify their queries in order to get better results? She determined semantic relations between queries by first mapping them onto concepts in linked data sources and then identifying the relations between the concepts.

In the third session, Khairun Nisa Fachry from the University of Amsterdam presented her work on the influence of document summaries on user click-through behaviour. More specifically, she investigated what information in a summary triggers a (positive or negative) selection decision. Kien Tjin-Kam-Jet from the University of Twente discussed in his presentation the usage of SVM to learn functions to merge search results for Distributed Information Retrieval. His approach uses readily available document features, such as the title, the summary and the URL, so that bandwith usage can be minimized.

The day was concluded with an invited talk by Cornelis H.A. (Kees) Koster from the Radboud University Nijmegen, who recently retired as a professor. Kees' accomplishments for the Dutch computer science and computational linguistics communities were presented by Theo van de Weide. Kees talked about the 'Text mining for Intellectual Property' project which aims to improve access to patents by applying syntactic dependency analysis. Kees first highlighted several classical problems of natural language text that are even more prominent in technical documents such as morphological variation, synonymy, homonymy, polisemy and collocations. Kees then presented the PHASAR approach where text is represented by dependency triples instead of single keywords and the user has control over precision and recall in interaction with the system. He concluded his talk by explaining innovative elements in the AEGIR parser that are required to deal with syntactic ambiguity.

4 Conclusion

DIR 2010 was a very successful event. Over 60 participants attended the day, providing evidence that the local Dutch Belgian Information Retrieval community is thriving. The workshop proceedings, keynote slides and photos of the day can be found at the DIR website: `http://www.ru.nl/ds/ifl/dir_2010/`. DIR 2011 will be hosted by the University of Amsterdam.