1st International Workshop on Advances in Patent Information Retrieval (AsPIRe'10)

Allan Hanbury
Information Retrieval Facility
Operngasse 20b, 1040 Vienna, Austria
a.hanbury@ir-facility.org

Veronika Zenz and Helmut Berger
Matrixware Information Services
Operngasse 20b, 1040 Vienna, Austria
{*v.zenz, h.berger*}*@matrixware.com*

1 Introduction

Patent Retrieval specialists in the 21st century face many challenges. They must search very large numbers of documents in multiple languages, expressing complex technological concepts through sophisticated legal clauses. Despite a great deal of theoretical development in Information Retrieval techniques and machine translation approaches, advanced search tools for patent professionals are still in their infancy.

Patent Information Retrieval is a cross cutting research area as it contains domains such as multilingual information retrieval; language processing; image processing and retrieval; and text categorisation, clustering and mining. The main goal of the workshop was to gather scientists from these areas together to foster interdisciplinary collaboration and spark discussions on open topics related to search in the Intellectual Property domain.

A few months before the paper submission deadline for the workshop, the IRF, supported by Matrixware, made available a collection of 400,000 patent documents in XML format for download – the AsPIRe'10 dataset. Groups submitting papers were encouraged to use this dataset for the experiments presented in the workshop papers. The AsPIRE'10 dataset was conceived to fill the gap in size between the 19 million patents in MAREC, the *MAtrixware REsearch Collection*, and the 20,000 patents in the *One Week of MAREC* collection. The AsPIRe'10 dataset continues to be available for download[1].

Six papers were submitted to the workshop, of which five were accepted for publication after thorough review by the members of the programme committee. Three of these papers make use of the AsPIRe'10 dataset.

[1]Information on obtaining these datasets is available here: `http://www.ir-facility.org/research/data`

The AsPIRE Workshop was held on Sunday, March 28th in conjunction with ECIR 2010 in Milton Keynes, UK. Although we did not have submissions from as many disciplines as hoped, the patent related research presented ranged from Machine Translation over NLP to Evaluation Measures and Retrieval Experiments. The talks stirred vivid discussion among the audience, and for almost every talk there were more questions than there was time for answering them. Barrou Diallo, head of the Research and Development Department of the European Patent Office (EPO), rounded up the workshop with a synthesis of the papers and topics discussed at the workshop. He began with an introduction to his work at the EPO and then summarised each of the five talks, highlighting the points that he judged as being most important for patent search practitioners.

2 Papers

The following summarises the five papers presented at the workshop:

Part-of-speech Language Model for N-Best List Re-ranking in Experimental Chinese-English SMT by Tao Jiang, Benjamin K. Tsou and Bin Lu. Tao Jiang and Bin Lu presented an extension by a list-reranking unit based on a part-of-speech model to a Chinese-to-English SMT system (Moses) that outputs several translation candidates. An advantage of part-of-speech (POS) language models is their small size compared to word surface from n-gram models, as the number of POS tags is very limited. It is therefore fast and easier to train language models. On the other hand POS models discard too much information associated with word surface form to be used independently. Thus the idea to use them in list-reranking. Depending on the model size (2-gram, 3-gram, etc.) the increase in the BLEU scores ranges between 0.001 and 0.0041 (for 9-gram). The question was raised whether these increases were significant: according to Benjamin Tsou it is very difficult to increase the BLEU scores of around 0.28 obtained, so even minor increases are worthwhile following. Barrou Diallo pointed out that this approach is particularly promising for adding new features to existing statistical machine translation systems.

Identifying Retrievability-Improving Model Features to Enhance Boolean Search for Patent Retrieval by Richard Bache and Leif Azzopardi. Richard Bache presented a study on the retrievability (or findability) of patent documents comparing retrievability in Boolean retrieval to ranked retrieval. He also introduced the Gini-Coefficient normally applied in social studies to measure retrievability. For a cutoff at 200 the best retrievability (i.e. least documents never retrieved) was attained by standard BM25 where only 0.05% of the documents were never retrieved. The worst results were shown by Boolean OR (57.28%). The Boolean result list was sorted chronologically, queries of length 2 were used (ORed) and the result list that was thus probably always very long was cut off at 200. Richard Bache pointed out that retrievability without precision is worthless. You could for example design a retrieval system that achieves maximum retrievability by returning a totally random result set every time without taking into account the query at all. He also pointed out that in retrievability studies the number of queries must be in the range of documents in the collection in order to give each document at least the theoretical chance of being retrieved. He concluded that a hybrid model which uses Boolean retrieval to filter the result set which is then ranked using e.g. BM 25 offers the best of both worlds for patent retrieval: it still gives the crisp cut-off of Boolean models (relevant/irrelevant) together with a better retrievability.

Quantifying the Challenges in Parsing Patent Claims by Suzane Verberne, Eva D'hondt, Nelleke Oostdijk, and Cornelis H.A. Koster. The motivation for Suzane Verbene's talk was the

statement found in many other research papers on the complexity of the claims section in patents. The presented work aims to verify and quantify the challenges of patent claim processing. They concentrated on 3 challenges: (1) length of claim sentences, (2) novelty of terms and (3) complexity of patent claim structure. With a median length of 22 words and average length of 53 words they found patent claim sentences to be longer than sentences in the British National Corpus. To verify the second challenge (vocabulary) a lexical coverage test of single-word terms, frequency counts of ambiguous lexical terms and analysis of multi-word terms was performed. For dictionary coverage they used CELEX. They found that CELEX covers 60% of the dictionary and 99% of the word occurrences. Introductions of novel terms could be verified only with respect to multi-word terms. With respect to (3) the authors found that syntactic parsing of patent claims is a challenge, especially because the claims consist of sequences of noun phrases ("NP") instead of clauses, while syntactic parsers are designed for analyzing clauses.

A New Metric for Patent Retrieval Evaluation by Walid Magdy and Gareth J.F. Jones. Walid Magdy started his talk with the observation that although it is commonly claimed by patent IR scientists that patent retrieval is a recall oriented task, the most common measure used in its evaluation is still MAP. Based on four example ranked lists he showed the shortcomings of existing measures (MAP, Recall, F1). He introduced a new measure "PRES" derived from Normalized Recall (Rnorm). This measures the effectiveness in ranking documents relative to the best and worst ranking case. When the ranked list is depicted as a graph with rank on the x-axis and recall on the y-axis, PRES is the area between the actual and worst case divided by the area between the best and worst case. PRES, like Rnorm, has a parameter N_{max} that specifies the number of documents that the user is willing to check. The difference of PRES to Rnorm is the place where non-retrieved relevant documents are assumed to be found: for Rnorm this is at the end of the collection — because of this Rnorm converges to Recall for large collections with hundreds of thousands of documents. PRES in contrast assumes the non-retrieved relevant documents would have been found directly after the cut-off (N_{max}).

On the effects of indexing and retrieval models in patent search and the potential of result set merging by Veronika Zenz, Sebastian Wurzer, Michael Dittenbach, and Edgardo Ambrosi. Veronika Zenz presented a baseline study on search quality, which compares 15 runs using a variety of ranking models provided by the three most common open source IR systems: Lemur, Terrier and Solr/Lucene. After explaining some statistics about the lexicon, she introduced the test collection creation process and gave an overview of the IR systems the team used. Furthermore she presented the evaluation results and highlighted the potential of result set merging together with the first results achieved with the COMP* merging techniques. The analysis of the collection anatomy, especially the large amount of single-document frequency terms stirred a lot of discussion, e.g. on the reasons for this and consequences. The influence that the team assumed that OCR had in this respect was confirmed by Barrou Diallo, who stated that the early OCR systems introduced much noise. OCR is still used today by the EPO, but it is of much higher quality. Re-OCRing is unfortunately not possible as the original documents are not kept by the EPO. Information on which OCR technique was used during which time period is however available, so it is possible to infer the OCR quality from the patent date. Gareth Jones proposed to give different weights to the runs when merging, according to the retrieval results they achieved.

3 Acknowledgements

We would like to thank the ECIR for hosting this workshop. We would also like to thank the program committee for the time and effort invested in reviewing the papers. The programme committee consists of: Giambattista Amati, Bruce Croft, Hamish Cunningham, Barrou Diallo, Michael Dittenbach, Andreas Eisele, Karl Fröschl, Steffen Koch, Cornelis H. A. Koster, Josep L. Larriba-Pey, Birger Larsen, Josep Llados, Mihai Lupu, Thomas Mandl, R. Manmatha, Silvia Miksch, Andreas Pesenhofer, Andreas Rauber, Angus Roberts, Giovanna Roda, Patrick Ruch, Gerold Schneider, Oscar Täckström, John Tait, Karl Tombre, and Stefanos Vrochidis. Final thanks go to the paper authors and the participants for an excellent workshop.