UTA and SICS at CLEF-IP

Antti Järvelin* Anni Järvelin* Preben Hansen**

*Department of Information Studies and Interactive Media
University of Tampere
Finland

**Swedish Institute of Computer Science

Outline

Introduction

- University of Tampere (UTA) and Swedish Institute of Computer Science (SICS) joined forces in CLEF-IP
- Our first try with patent retrieval – The goals were:
  - Getting a retrieval system up and running
  - Study the automatic query generation process
- For two topics, the extracted query words were compared to query keys selected by three human experts
- We participated in the main task with 8 XL runs

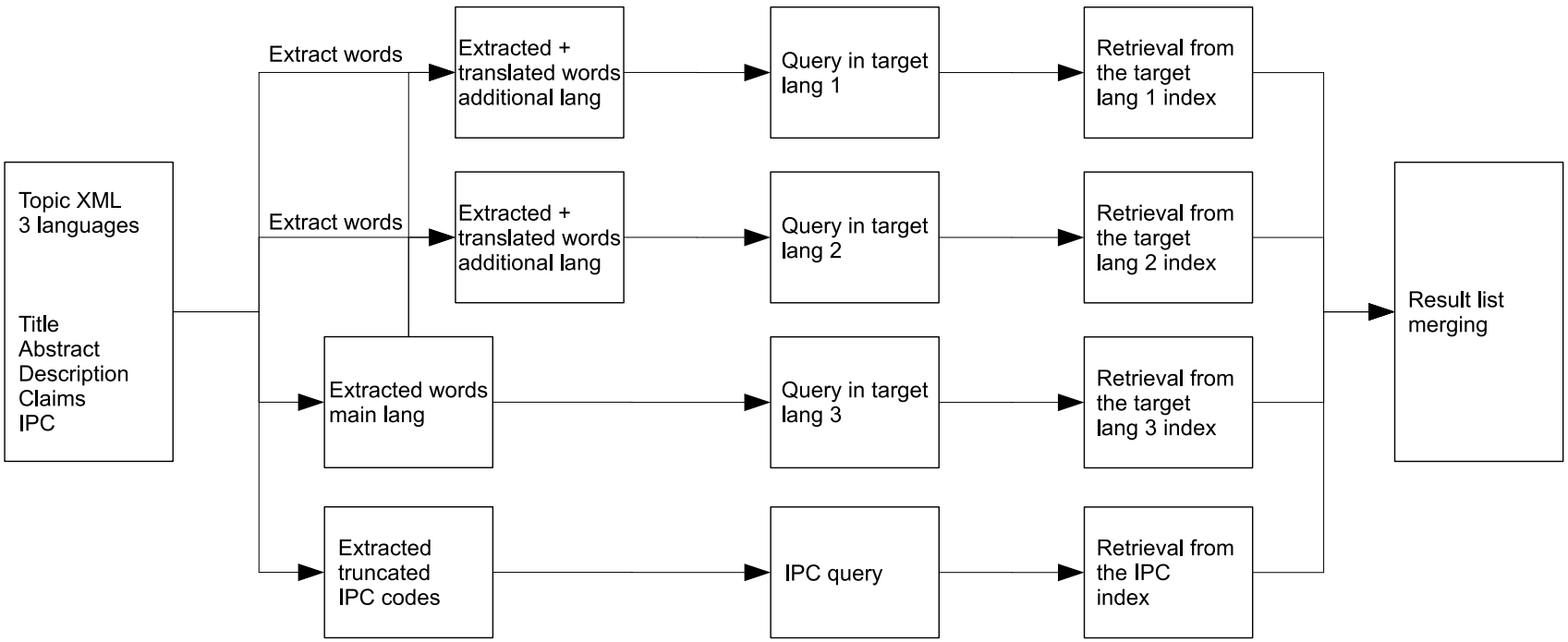
Automatic Query Generation

- Three approaches for picking query words from the topics
  - RATF-formula (e.g. Pirkola et al 2002) – does not account for the word frequencies in the topics
  - The "standard" $tf \cdot idf$ weighting of the topic words
  - Modified RATF-formula that accounts for the topic word frequencies
- Topic words were weighted using one of the previous formulas
- The top n words were then selected to form the query

Manual Queries by Patent Experts

- We had the opportunity to employ three patent engineers to analyse two of the topics (EP1186311 and EP1353525)
- We are aware that not much can be claimed based on an analysis of two topics only . . .
- . . . but we were hoping to get some first indications on how
  - patent examiners work
  - our automatic query generation procedure could be improved
- The manual queries were formed from the words that the patent engineers selected to be the top 10 representative words for the topic

System Details – Overview



Indexing

- "Virtual patent" based approach to indexing
  - Only the central fields (title, abstract, description and claims) were indexed and only the most recent version of each of the fields was indexed
- Separate monolingual index for each of the languages
  - Words were stemmed using the popular Snowball stemmer
- The IPC codes were indexed separately into a language independent index
  - The IPC codes were truncated after the fourth character

Retrieval

- Three monolingual queries and an IPC code query were run for each of the topics
- All the natural language queries in all runs were set to include 50 words, based on training results
- Missing patent fields were in some experiments translated from the main langugage's field using Google Translate
- The IPC queries included all the IPC codes present in a topic document
- Each index returned the top 2,000 best matches

Result List Merging

- The results from the four different queries were merged at query time using MAD (Mean Average Distance) merging model (Wilkins et al 2006)
  - Enables query based index weighting
- The scores of each index were min-max normalized before merging
- After merging top 1,000 docs were returned

Implementation

- A framework to study patent retrieval with the following properties:
  - Search engine independent – currently supported engines: Lemur, Lucene (experimental)
  - Environment for studying automatic query generation
  - Supports both query and document translation approaches for CLIR
- Implemented using Java-programming language
- Lemur-backend was used as a search-engine backend in our CLEF-IP runs

Results for the XL Runs

Run ID	**P10**	**MAP**	**nDCG**
UTASICS_abs-des-ratf	**0.0945**	**0.1237**	**0.4722**
humb_1	*0.1776*	*0.2802*	*0.5877*

Table: Our best run compared to the run by the Humboldt University, humb_1.

Results for the XL Runs

- The combination of the abstract and description fields seemed to be a better source of query keys than the other combinations
- Abstracts in general were the most promising source of query keys when no proper translation resources were available:
  - All topics contained the abstracts in all of the three target languages
- Using GT was not useful in general and seemed to perform especially badly on translation of the description fields
- $RATF_{\text{mod}}$ and $tf \cdot idf$ performed very similarly, and clearly better than the original RATF-formula

User Generated vs. Automatically Generated Queries

- The overlap between the user-generated and the automatically generated queries was usually four words
- The user generated queries performed worse than the automatically generated ones (based on MAPs):
  - 0.3333 vs. $<$ 0.01 for the topic EP1186311 and
  - 0.0004 vs. 0 for the topic EP1353525

Summary

- The modified version of the RATF-formula and the $tf \cdot idf$ weighting could be good candidates for intial query extraction in patent retrieval
- The combination of abstract and description fields was the best source for query words in our runs
- Our approach to using GT for translating the missing patent fields did not noticeably improve the results
- More user data would enable interesting evaluations of the system

Thank You

Questions?
Comments?
Suggestions?

Antti was partly funded by the Finnish Cultural Foundation
Anni was funded by Tampere Graduate School in Information Science and Engineering (TISE)

We wish to thank the Swedish patent bureau Uppdragshuset for their kind help and especially the three patent examiners for participating in the user assessments.