Exploring Passage Retrieval with the PIPExtractor

Linda Andersson[1]
Parvaz Mahdabi[2]
Allan Hanbury[1]
Andreas Rauber[1]

[1] Vienna University of Technology, Austria
[2] University of Lugano, Switzerland

Passage Intellectual Property Extractor

Conditions:

i. It should be IR-platform independent
   - not incorporated in the indices

ii. Take the advantage of noun phrases which have shown to be partly effective in Patent Document Retrieval

(Mahdabi et al 2012)

(Passage) Patent Retrieval

- The most used model in patent search by patent searcher is the classification system (IPC) combined with Boolean retrieval model since it is transparent and the model will generate high recall, if the query constructed by the expert is well formed (Dulken 1999).
  - Here the search outcome lies in the hand of the searcher.

- There are several linguistic related issues when it comes to IR
  - *Selection of alternative concepts and search keys*
    - patent writer becomes his/her own lexicographist (Atkinson 2008).
    - diachronical nature of the patent genre terms such as has "LP" and "water closet" could be regarded as instances of obsolescence (Harris et al 2011).
    - the morphological variation of search keys in patent reflects in the high amount of chemical formulae and morphological variation of foreign spelling e.g 'sulfur-sulpur' and aluminum-aluminium.
  - A patent writers intentionally try to use entirely different word combinations, not only synonyms, but also paraphrasing to re-create "concept".

Vocabulary characteristics of the Patent genre

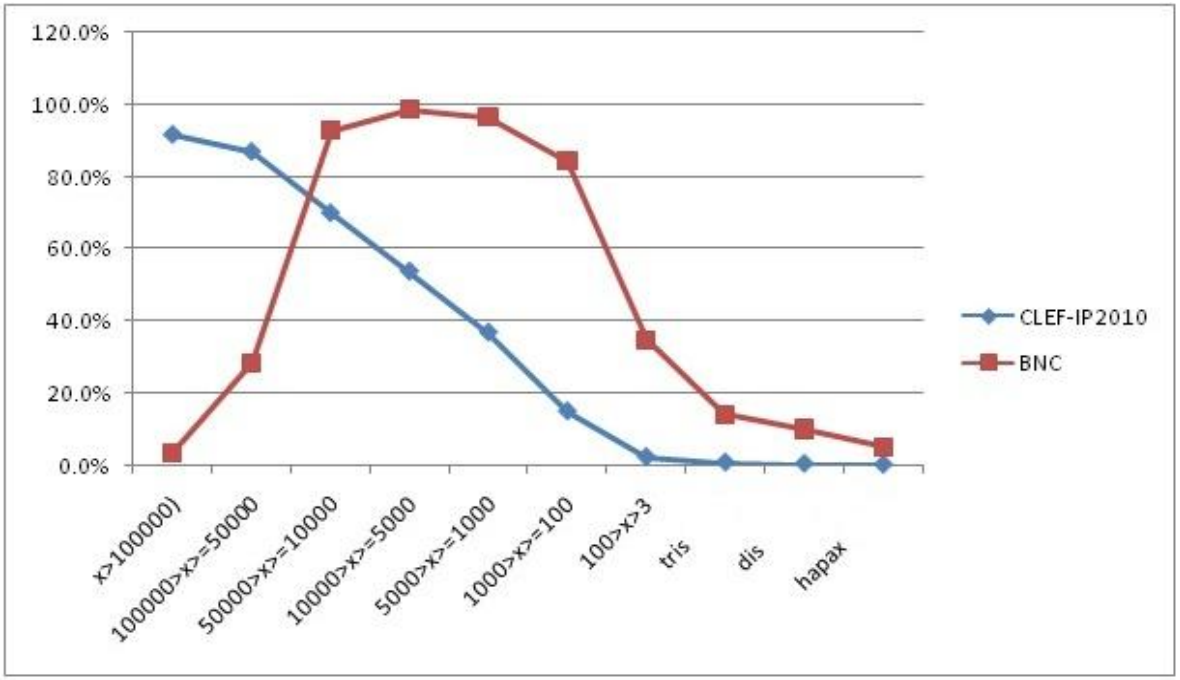
- Patent documents are associated with several interesting characteristics such as huge differences in length, strictly formalized document structure (both semantic and syntactic), acronyms and new terminology (Larkey 1999).
  - A patent document consists of four main textual components
    - the title, concise short description of invention
    - the abstract section gives short and general summery, broad terms are generally used.
    - the description section gives elaborative background information on the invention
    - the claim has its own very special conceptual, syntactic and stylistic/rhetorical structure and need to compose the essential component of the invention to make patent infringement difficult .
- When comparing general language resources (CLEX lexicon 160,568 English terms) with a patent corpus of 10,000 documents coverage on distinct word type (excluding chemical formulae and numerals) was 60% (D'hondt et al 2011).

Lexical coverage: WordNet

The graph shows the coverage of the WordNet for two different corpus CLEF-IP2010 (over 1.2 million patent document from EPO) and the British National Corpus (BNC) 100Million tokens – a balanced annotated corpus of the English language.

Y-ax shows the coverage in percentage

X-ax shows the collection frequency for each token. Tris means that a token occurs only three times in the entire collection dis only 2 times and hapax only 1.



Benefits and disadvantages of using NLP in Patent Retrieval

- Noun phrases is motivated by the fact that technical dictionaries, in majority, consist of terms with more than one word (Justeson & Katz 1995).

- IR and NLP has shown to be successful when it comes to shallow linguistic method while deeper linguistic analyses could even decrease accuracy (Brants 2003).

- To use a NLP-application without any adaptations to the patent domain would affect the performance of the application considerable (Sheremetyeva 2003).

Known limitation of NLP

- Broad coverage?
  - Mostly trained on news text
  - Ambiguity in handling of non-letter strings
  - Unseen even (lexical, syntactic)
  - In order to obtain high accuracy the trained data need to be similar to the intended data
    - Source vs Target data
    - 97% accurcy can only be obtained in ideal scenarios i.e. the tagger is trained on a highly standard text, with low rate of errors and unusual patterns; and the test data is very similar to the training set (Gisbrecht and Evert 2009).
  - Limitation based upon linguistic theory behind the system, lexicon based, data-driven etc
- Domain Adaptation
  - Time consuming
  - Resource consuming
  - Not always generating a increase in target application performance

Previous use of NLP application tools in Patent Retrieval

- Many studies within patent retrieval have made use of variety of linguistic knowledge such as lexical-syntactic pattern, generic NLP tools and domain semantic annotation based on Ontologies.

- But very few of the previous studies addressed domain adaptation, since it is time consuming and require access to Natural Language Resources, domain knowledge as well as "deep" linguistic knowledge
  - The PHASAR system has been domain adapted towards the patent domain by increase of lexicon coverage.
  - It uses a special grammar based on AEGIR (an extension of Context Free grammar formalism). The rule-based grammar comes up with several interpretation suggestions but the statistical data promote the best interpretation. The statistical data is stored within a manually maintained database
    - Make the system both sensitive towards coverage and phrase weighting

  - In Sheremetyeva (2003) a supper tagger (lexical driven) was re-trained on domain knowledge. Terms from 5 million US patent documents were collected and additional domain dependencies relations rules was incorporated in the NLP processes. The grammatical dependency rules was derived from patent texts
    - Make the system sensitive to lexical as well as syntactic (grammar rules) coverage

Our adaption of NLP application

- The Stanford Part-of-Speech tagger (using the english-left3words-distsim.tagger model)

- The noun phrase was extracted based upon 201 lexico-syntactic pattern
  - Original candidate structure $((A|N)^+|((A|N)^*(NP)^?(A|N)^*)N$ (Justeson & Katz 1995)
  - Expansion of the pattern
    - noun phrases with preposition 'of'
    - adjective on final position in NP - if being chemical compound
- Benefits,
  - simplifies the pre-process, in terms of time and resources
  - partly language independent

- Disadvantage
  - syntactic coverage
  - require linguistic knowledge as well as domain knowledge
  - PoS-errors
    - Add rules to handle common errors of the application e.g. particples used as adjectives "the unwanted microbial growth" DT(A)*[VBG|VBN](A|N)*N

Our Approach

- Document retrieval Method
  - a Language Model based on IPC classes was used (Mahdabi et al 2011).

- The PIPExtractor consist of a two-stage method:
  - The query model consisted of two-dimensioned-matrix computing cosine similarity values pair wise for each sentence in the topic document in order increase query terms.
  - In the passage model a four-dimension-matrix was used generating cosine values for word and noun phrases in the original topic claim sentence and word and noun phrases used as query expansion keys. The computation across document boundaries was conducted per sentence; paragraph containing several sentences received a summation value. The term frequency was used as weight technique.

- Topics with the main language other than English were semi-manually translated by accessing the EPO Google Translation.
- All documents used as topics were Part-of-Speech tagged with the Stanford Part-of-Speech tagger (using the english-left3words-distsim.tagger model) (Toutanova et al 2003).
- In the official run only the TF of noun phrase and open word classes were used both in the query model and in the passage model. For each retrieved passage four different cosine values were generated; and then summed up in order to establish one value per retrieved passage.

Results

The baseline is generated by the Document Retrieval Model only listing retrieved document. Four different combinations were deployed at the passage level:

1. TF-Sum
2. The TF-Sum value was divided by the position rank value given by the Document retrieval model
3. Additional weight (0.2) for the noun phrases was given in calculation
4. TF-IDF and a Porter stemmer on word and noun phrases were deployed.

Run ID	PRES@100	Recall@100	MAP	MAP(D)	Precision(D)
Baseline	0.2105	0.2653	0.0662	0.0000	0.0000
PIPExtractor-2.3	0.1552	0.2107	0.0421	0.0029	0.0315
PIPExtractor-1.2	0.1467	0.1869	0.0275	0.0011	0.0064
PIPExtractor-1.3	0.0274	0.0387	0.0035	0.0017	0.0227
PIPExtractor-1.3.4	0.0278	0.0384	0.0041	0.0023	0.0134
PIPExtractor-1	0.0228	0.0303	0.0033	0.0020	0.0292
PIPExtractor-1.4	0.0371	0.0655	0.0019	0.0008	0.0146
PIPExtractor-1.2.3.4	0.0809	0.1176	0.0227	0.0021	0.0128

Discussion & Future Work

Our aim for the Passage Retrieval task was to construct a module independent of IR-Platform and use the power of noun phrases to improve the performance. Although the position information from the IR system is very important in order to avoid a drop in performance.

- The paradox
  - large amount of data (e.g. higher frequency, lager document etc)
  - but also data sparseness (e.g.) selection of alternative concepts and search keys, referred and omitted search keys and search key ambiguity.
- This is partly language depended, since in
  - English combines common terms in order to create new terminology.
  - data sparseness occurs since each part of the new terminology can be substituted with synonyms or just have a different morphological suffix.

- Next step
  - Improve the NLP pre-process
  - Restrict the words and noun phrases based on additional constraint
    - Noun phrase >1
    - Different weight schema
  - Add expansion
    - Sloppy noun phrases (window of 5)
    - Accept synonyms for noun phrases
  - Extend method to include French and German

References

- K. H. Atkinson, Toward a more rational patent search paradigm. In Proceedings of the 1st ACM workshop on Patent information retrieval (PaIR '08). ACM, New York, NY, USA, 37-40. 2008.
- A Fujii., M. Iwayama and N. Kando,. Introduction to the special issue on patent processing. Inf. Process. Manage. 43, 5 (September 2007), 1149-1153.(2007
- J. Blitzer, R. McDonald, and F. Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (EMNLP '06). Association for Computational Linguistics, Stroudsburg, PA, USA, 120-128.
- T. Brants , Natural Language Processing in Information Retrieval. In Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands. 2003
- Coden R. A., Serguei V. S., , Ando K. R.,. Duffy H. P., and Chute Christopher G. C. (2005) Domain-specific language models and lexicons for tagging. J. of Biomedical Informatics 38, 6 (December 2005), 422-430. DOI=10.1016/j.jbi.2005.02.009 http://dx.doi.org/10.1016/j.jbi.2005.02.009
- E. D'hondt , S.Verberne , W. Alink and R. Cornacchia.(2011) Combining document representations for prior-art retrieval. Workshop of the CLEF-IP2011, LABS and Workshops, Notebook Papers.
- S. van Dulken, S. Free patent databases on the Internet: a critical view, World Patent Information -Volume 21(4); p 253-257.1999
- T. A. Hedlund,, A. Pirkola, and K. Järvelin, , Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval Information Processing and Management - Volume 37(1), 147-161. 2001
- C. G. Harris, R. Arens and P. Srinivasan P, Using Classification Code Hierarchies for Patent Prior Art Searches. Current Challenges in Patent Information Retrieval. The Information Retrieval Series, Vol. 29. Lupu, M.; Mayer, K.; Tait, J.; Trippe, A.J. (Eds.) 1st Edition, 2011, XIV, 402 p. 2011
- J. S. Justeson, and S. M. Katz, (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1):9–27.1995
- H-A C. Koster, J. Beney , S. Verberne, and M. Vogel (2011). Phrase-Based Documentation Categorization. *In* Croft W. B., Lupu M., Mayer K.,, Tait J., and Trippe J. A., (ed.) *Current Challenges in patent Information Retrieval*, Springer Berlin Heidelberg
- L. S Larkey, A patent search and classification system, In Proceedings of the 4th ACM conference on Digital libraries, (pp 179-187), (Berkeley, California, United States. 1999
- P. Mahdabi, L. Andersson, M. Keikha, and F.Crestani, Automatic refinement of patent queries using concept importance predictors. In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). ACM, Portland, USA, 505-514. 2012
- P. Mahdabi, L. Andersson., A. Hanbury and F.Crestani, Report on the CLEF-IP 2011 Experiments: Exploring Patent Summarization, Workshop of the Cross-Language Evaluation Forum, LABS and Workshops, Notebook Papers. 2011
- N. Oostdijk, H. van Halteren, E. D'hondt E, and S. Verberne , Genre and Domain in Patent Texts. In Proceedings of the The 3rd International Workshop on Patent Information Retrieval (PAIR) at CIKM 2010, pages 39-46, 2010.
- S. Sheremetyeva., Natural language analysis of patent claims. In Proceedings of the ACL-2003 workshop on Patent corpus processing - Volume 20 (PATENT '03), Vol. 20. Association for Computational Linguistics, Stroudsburg, PA, USA, 66-73. 2003.
- K. Toutanova, D. Klein, C. Manning, and Y. Singer, Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259. 2003
- Nivre, J. (2008) Sorting Out Dependency Parsing. In Proceedings of the 6th International Conference on Natural Language Processing (GoTAL), 16--27.