# Report on the CLEF-IP 2012 Experiments: Exploring Passage Retrieval with the PIPExtractor

Linda Andersson[1]
andersson@ifs.tuwien.ac.at

Parvaz Mahdabi[2]
parvaz.mahdabi@usi.ch

Allan Hanbury[1]
hanbury@ifs.tuwien.ac.at

Andreas Rauber[1]
rauber@ifs.tuwin.ac.at

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

**Abstract.** This technical report presents the work carried out for the Patent Passage Retrieval track of CLEF-IP 2012. Our aim was to create IR-Platform independent module for the Passage Retrieval process. For the Document retrieval Method - a Language Model based on IPC classes and for the Passage Retrieval a Passage Intellectual property Extractor (PIPExtractor) was implemented. Topics with the main language other than English were semi-manually translated by accessing the EPO Google Translation. We submitted five official runs one retrieving only document and four retrieving Passages.

**Keywords:** Passage Retrieval, Patent Search, Natural language Processing

## 1 Introduction

The CLEF-IP track started in 2009 with Prior Art Candidate Search track, since then several different tasks have been explored e.g. text classification, Image Retrieval etc [1]. This year Passage Retrieval was introduced as the text mining task. Previously, Passage Retrieval has been explored in NTCIR 4 and 5 [2].

For this Passage Retrieval track we set up two pre-conditions for our Passage Retrieval module i) it should be IR-platform independent i.e. it should not be incorporated in the indices; ii) we want to take the advantage of noun phrases which have shown to be effective in Patent Document Retrieval [3].

To use noun phrases as complement to bag-of-word method in IR is motivated by the fact that technical dictionaries, in majority, consist of terms with more than one word [20]. The technical multi-word phrases consist of noun phrases containing adjective, nouns and occasionally preposition (e.g. 'of').

However, using NLP-application in order to improve IR results has not been straight forward since sentence as "time is ripe to use Natural language processing for information retrieval" [4, p1] has generally been followed by "the impact of NLP on information retrieval task has largely been one of promise rather than substance" [5, p99]. Research involving Information Retrieval (IR) and Natural language processing

(NLP) shows the that shallow linguistic methods such as stop word, stemmer, etc. yield significant improvements, while deeper linguistic analyses such as Part-of – Speech tagging, parsing, word sense disambiguation etc. could even decrease accuracy [6]. Furthermore, as to use a NLP-application without any adaptations to the patent domain would affect the performance of the application considerable [7].

In experiment we use two-stage model costing of a Query model and a Passage Model. Only open word classes (adjective, verbs and nouns) and noun phrases are used in the similarity computation between paragraphs. The noun phrase extraction methods re-uses the lexico-syntactic patterns used in [3] with additionally pattern including noun phrases with preposition.

## 2 Patent Retrieval

The most used model in patent search is the Boolean retrieval model since it is transparent and the model will generate high recall, if the query constructed by the expert is well formed [8]. Here the search outcome lies in the hand of the searcher. The first search task of a patent examiner when given a patent application is to identify essential aspects and extract terms that can be used in the search query session. In real life this search is not limited to patent – since no prior publication shall exist in order to meet the uniqueness and novelty requirement.

Additionally, to the general linguistic IR problems presented in [9], previous studies in the patent genre have observed that patent writers intentionally try to use entirely different word combinations, not only synonyms, but also paraphrasing to recreate "concept" [10], [11].

In the patent genre the selection of alternative concepts and search keys have turned out to be more severe since a patent writer becomes his/her own lexicographist [10]. Furthermore, given the diachronical nature of the patent genre terms such as has "LP" and "water closet" could be regarded as instances of obsolescence [11]. 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.

The problem defined as referred and omitted search keys encompasses anaphoric and elliptical keys (e.g. pronoun, acronyms etc) [9]. In the patent genre both standard and non-standard acronyms is used [11]. The search key ambiguity addresses polysemy or homography in the patent genre phrases like "mouse trap" (a trap to catch mouse, a logic device) addressing wide variety of concept in different technological field, they are so called "shape shifter" [10]. Consequently, the patent collection accentuates several linguistic problems which still general IR systems are struggling with in other text genre such as term distribution diversity, vocabulary mismatch, dealing with paraphrases (including hyponymy relation, synonym etc)

Many Patent Retrieval studies have tried to address above IR-issues by applying variety of linguistic knowledge like lexico-syntactic pattern extraction, creation of domain semantic annotation and using ontologies. The studies have target different a

range of application from handle phrase indexing [12] and query reduction or expansion, semantic annotation [13] to sentence decomposition [14] and readability [7].

In Mahdabi et al a Part-of-Speech tagger was used for noun phrase extraction. The extracted were based upon manually observed lexico-syntactic patterns [3]. The noun phrases were then combined with different weight schema. Even though, errors were generated by the Part-of-Speech tagger when combining the extracted noun phrases with statistical methods the errors was minimized.

### 2.1 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 [15]. 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% [16].

A patent document consists of four main textual components (title, abstract, description, and claim) which intention is to fulfill different communication goals. The abstract section gives short and general summery, broad terms are generally used. The description section gives elaborative background information on the invention, here prior art in the field is mentioned. 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 [15].

## 3 Our Approach

**Data**, the corpus used in the Passage Retrieval track is consistence with the CLEF-IP 2011 collection both containing WO patent and EPO patent. The claim segments used as topics were extracted from 58 different patent application documents –generating 105 different topics. The claims segment used as topic was manually selected based on existing search reports. For the Passage the xpaths was used as Qrels.

**Method**, for the Document retrieval Method - a Language Model based on IPC classes was used (for detailed description see [17]). The Passage Intellectual Property Extractor (PIPExtractor) was implemented in Perl and consists of a two-stage method: a query model and a passage model. 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. The first dimension generated a cosine value based upon bag-of-word (only considering adjective, verbs and nouns) between sentences; the second dimension generated a cosine value based upon common noun phrases - a modification of technique used in [18].

In the second stage 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 values. 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) [19]. The noun phrase was extracted based upon 201 lexico-syntactic pattern include noun phrases with preposition 'of' and participle used as adjectives.

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.

Unfortunately, a late error in or script – generate incorrect format of xpath which explain the exceptionally low performance compared to the other participants runs. The script error was only discovered after the submission deadline. Therefore, we have chosen to present the corrected version of the official runs in this report.

In this report we also present three additional experiments using TF-IDF (inverse document frequency) and a stemmer (Porter) on both noun phrases and open word classes. The IDF was calculated within the retrieved documents generated of the Document Retrieval Method.

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.

## 4 Results

The official runs was also carried out on all 105 topics however since we only translated topic the overall performance on document was moderate with a MAP value of 0.0383 and PRES@100 of 0.1810 and Recall@100 of 0.1810.

The official runs for the Passage Retrieval track did not find any relevant passage for one runs and other runs had a MAP(D) value below 0.0002 and below 0.0006 for Precision (D) – this was caused by a error in the script generating the xpath. Therefore in the table 1 we present the official runs when the xpath has been corrected.

**Table 1.** Only English Topic at CLEF-IP2012 Passage Retrieval Track

| 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 |

Using only TF as weight technique decrease the ranking order considerable on the document level compared to the baseline. When using the ranked position value from the Document retrieval method the drop between baseline performances is reduced, significantly. When experiment with TF-IDF and Porter the performance improves on the document level but have negative effect on the Passage Retrieval Level. Combining all method (1.2.3.4) both the performance on document level as well as on Passage Retrieval level decreased.

## 5 Discussion and Conclusion

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. By using TF-IDF as weight schema and allow stemming we increase the discrimination power of words and phrases and reduce the affect of the morphological variation of search keys – generating an improvement on document level but it affected the Passage Retrieval performance negatively.

The paradox in the patent genre is that there is a large amount of data (e.g. higher frequency, lager document etc) but there is also issues related to data sparseness such as selection of alternative concepts and search keys, referred and omitted search keys and search key ambiguity.

The paradox is partly language depended, since in English combines common terms in order to create new terminology. At the same time the data sparseness occurs since each part of the new terminology can be substituted with synonyms or just have a different morphological suffix. When exploring Patent Retrieval consisting of English patent documents there is a need to identify the noun phrases boundaries as well handle the data sparseness in terms of stemming and expanding with synonyms. However, there are several pitfalls that can make this process perform lesser than good since we are depending on NLP-applications which are time consuming to adapt to the patent genre since it requires both linguistic knowledge as well as domain knowledge.

# 6 References


1. F. Prior and J Tait, CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. Workshop of the Cross-Language Evaluation Forum, LABS and Workshops, Notebook Papers. 2010
2. 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
3. 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, New York, NY, USA, 505-514. 2012
4. M. Lease, Natural Language Processing for Information Retrieval: the time is ripe (again). In Proceedings of the 1st Ph.D. Workshop at the ACM Conference on Information and Knowledge Management (PIKM). 2007
5. A. F. Smeaton, Using NLP or NLP resources for information retrieval task. In T. Strzalkowski, editor, Natural language information retrieval. Kluwer Academic Publisher, Dredrecth, NL, 99-111. 1999
6. T. Brants , Natural Language Processing in Information Retrieval. In Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands. 2003
7. 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.
8. S. van Dulken, S. Free patent databases on the Internet: a critical view, World Patent Information -Volume 21(4); p 253-257.1999
9. 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
10. 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.
11. 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
12. D'hondt E., Verberne S., Alink W. and Cornacchia R.(2011) Combining document representations for prior-art retrieval. Workshop of the CLEF-IP2011, LABS and Workshops, Notebook Papers.
13. L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, P. Gemma, I. Puhlmann, R. Gautam, M. Rotard, P. Schoester, L, Serafini and V. Zervaki, Towards content-oriented patent document processing World Patent Information -Volume 30(1), 21-30. 2008
14. P. Parapatics and M. Dittenbach, Patent Claim Decomposition for Improved Information Extraction. In W. B Croft., M. Lupu, K. Mayer, J. Tait and J.A. Trippe, (eds.) Current Challenges in patent Information Retrieval, Springer Berlin Heidelberg. 2011
15. 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
16. 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.

17. 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
18. K. Konishi, Invalidity patent search system of NTT DATA. In Working Notes of the Fourth NTCIR Workshop Meeting. 250--255. 2004
19. 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
20. 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