Exploring Patent Passage Retrieval Using Nouns Phrases

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

[1] Vienna University of Technology, Austria
[2] University of Lugano, Switzerland
{andersson,hanbury,rauber}@ifs.tuwien.ac.at,
parvaz.mahdabi@usi.ch

Abstract. This paper presents experiments which initially were carried out for the Patent Passage Retrieval track of CLEF-IP 2012. The Passage Retrieval module was implemented independently of the Document Retrieval system. In the Passage Retrieval module we make use of Natural Language Processing applications (WordNet and Stanford Part-of-Speech tagger) for lemmatization and phrase (multi word units) retrieval. We show by applying simple rule-based modifications and only targeting specific language instances (noun phrases) the usage of general NLP tools for phrase retrieval will increase performance of a Patent Passage Information Extraction system.

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

1 Introduction

The CLEF-IP track started in 2009 with Prior Art Candidate Search track. In 2012, Passage Retrieval was introduced as the text mining task. The Boolean retrieval model is the most commonly used model in patent search due to its transparency as well as its high recall generation, given that the query constructed by the expert is well formed [1]. Here the search outcome depends on searcher's ability to distinguish between multiple meanings (senses) of a word – using phrases in several iteration steps to narrow the scope of the word's semantic field.

In the patent genre, issues addressing polysemy become more severe due to the ambiguous terminologies where terms represent a wide variety of concepts in different technological fields, the so called "shape shifters" (e.g. "bus")[1] [2]. We address the terminology ambiguity issue by extracting a set of noun phrases from each topic which according to the paper cited in [3] reflects technical concepts better than a single word and therefore distinguishes between different meanings of polysemous terminologies (e.g. bus: "bus card slots" versus "double-decker bus").

2 NLP and Patent Retrieval

Many Patent Retrieval studies have tried to address different search problems by applying linguistic knowledge. To use phrase retrieval and specially noun phrases as a

[1] i) motor vehicle, ii) an electronic subsystem transferring plurality of digit bits in group.

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 676–679, 2013.

complement to bag of word method in Information Retrieval (IR) is motivated by the fact that technical dictionaries, in majority, consist of terms with more than one word [3]. The technical multi-word phrases consist of noun phrases containing adjectives, nouns and occasionally prepositions (e.g. ‘of’). However, research involving IR and Natural Language Processing (NLP) shows that the shallow linguistic methods such as stop word removal, stemmer, etc. yield significant improvements, while deeper linguistic analyses such as Part-of-Speech (PoS) tagging, parsing, word sense disambiguation, etc. could even decrease accuracy [4]. In this paper, we demonstrate that deeper linguistic analyses can improve performance even in a text genre out of scope of the general PoS taggers.

3 Our Approach

Data, the CLEF-IP 2012 collection contains approximately 3.5 million XML document (representing 1,5 million patent documents). The claim segments used as topics were extracted from 58 different patent application documents – generating 105 different topics – in this experiment we only consider the English topics (35). The claim segment in a patent document was used as the topic and was manually selected based on existing search reports. Citations combined with XPaths were used as assessors (Qrel) in the Passage Retrieval track.

Method, for the Document Retrieval system - a Language Model based on IPC classes was used [5]. All topic documents were PoS tagged with the Stanford[2] [6] and all words (topic and retrieved document) were lemmatized (via WordNet[3]). In the noun phrase (NP) extraction process we re-used the lexico-syntactic patterns used in [7] with additional patterns including NPs with prepositions and participles used as adjectives. Approximately 2000 multi-word phrases were manually inspected in order to arrive at 201 linguistically accepted NP patterns. For the topic set the pre-established NP patterns produced 2,288 multi-word phrases (63 in average per topic).

The Passage Information Extraction (IE) module was implemented in Perl and is composed of a two-stage method: a Query Model and a Passage Model. The Query Model consists of a four dimensional matrix representing open word classes (1-dim) and NPs (2-dim) in the topic claim and associative open word classes (3-dim) and NPs (4-dim) extracted from the rest of the topic document. The claim sections are mostly composed of stylistic marked words rather than topic reflecting words [8]. In order to arrive at associative terms, cosine similarity values were computed pair wise between claims and other sentences in the topic document, similar to the technique used in [8].

In the Passage Model we expand the matrix to six dimensions by adding a three word window$_{\{NPlgth+3\}}$ for each NP dimension - claim NP (5-dim), associative NP (6-dim). We also gave extra weight .2 (Ew2) to the NPs when combining word and NP methods [5]. For each retrieved paragraph cosine similarity value was computed and

[2] English-left3words-distsim.tagger model (See).

[3] "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu

summed up; and then divided by the position rank value from the Document Retrieval system. For all cosine similarity computations only the term frequency (TF) was used, since TF requires no collection information. Five different cut off values were tested i.e. passage from the top10, top50, top100, top500 and top (1000) retrieved documents. In the result only the best cut off (10) level is presented.

4 Results

Table 1 shows the performance of different methods on document level and passage level.

Table 1. Results (sorted by MAP(D) at passage level)

	Document level			Passage level	
Run ID	PRES@100	Recall@100	MAP	MAP(D)	Precision(D)
1.2.3.4-DimEW2	0.1955	0.1965	0.0605	**0.0354**	0.0221
5.6-Dim	0.1954	0.1965	0.0595	0.0332	0.0221
2.4-Dim	0.1955	0.1965	0.0614	0.0325	0.0222
1.2.3.4-Dim	0.1954	0.1965	0.0581	0.0307	0.0221
1.3-Dim	0.1954	0.1965	0.0532	0.0285	0.0222
Document Retrieval	0.2105	0.2653	0.0662	0	0

The method (1.2.3.4-DimEW2) combining words and NPs where extra weight is added to the NP dimension achieves the highest performance in terms of MAP(D). MAP(D) is a micro version the standard Mean Average Precision (MAP), computing average precision for each relevant XPaths retrieved for a single relevant document.

All methods using NP either in combination with words or as the solitary query method increase the MAP(D) value compared to the method using only words (1.3-Dim). Figure 1 shows the difference in performance, MAP(D), between solitary NP methods (5.6-Dim, 2.4-Dim) and the bag-of-word Method (1.3-Dim).

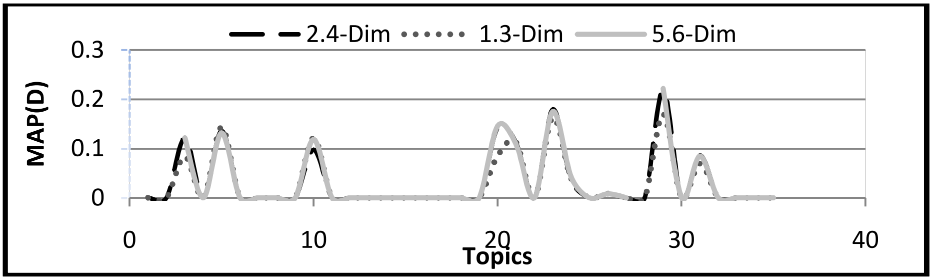

Fig. 1. MAP(D) per topic

Using 5.6-Dim (window$_{\{NPlgth+3\}}$) as the solitary query method increased the performance with an average of 11% per topic over 1.3-Dim (bag-of-word). But for four topics the 1.3-Dim was more effective, since the relevant paragraphs only had one word overlap with the topic. Although, the usage of NP either as complementary or as solitary query method increase performance for passages the loss on document level is still considerable. The Document Retrieval System retrieved relevant document for 23 out 35 topics and for the Passage IE module the number of topic retrieving relevant documents is 13. The loss in performance compared to the Document Retrieval system is partly caused by the simplicity of the weight method (only using TF) and partly due to the low number of overlapping terms between the topic terms and relevant paragraphs.

5 Conclusion

In this paper, we address the terminology ambiguity issue by comparing a query method using only words as opposed to a query using noun phrases or a combination. We set up a twofold hypothesis, first claiming that multi-word units better capture technical concepts since they reduce the polysemous terminologies used in the patent genre (e.g. bus: “bus card slots” vs. “double-decker bus”). In the second part, we claimed that applying simple rule-based modifications to a general PoS tagger and only targeting specific language instances will increase performance for a Patent Passage IE module compared to using only words. Although, our results support our claims, due to the small amount of topics used in the CLEF-IP Passage track we cannot state with higher level of certainty that this is in fact the case. Our results rather indicate, i) noun phrases are useful in order to improve performance in terms of precision, ii) a general PoS tagger can be used successfully in the patent genre if used in combination with observed syntactic pattern from the patent genre.

References

1. van Dulken, S.: Free patent databases on the Internet: a critical view. World Patent Information 21(4), 253–257 (1999)
2. Atkinson, K.H.: Toward a more rational patent search paradigm. In: PaIR 2008 Workshop, pp. 37–40. ACM, New York (2008)
3. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1), 9–27 (1995)
4. Brants, T.: Natural Language Processing in Information Retrieval. In: Proc. 14th CLIN (2003)
5. Mahdabi, P., Andersson, L., Keikha, M., Crestani, F.: Automatic refinement of patent queries using concept importance predictors. In: Proc. 35th SIGIR 2012. ACM, Portland USA (2012)
6. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proc. HLT-NAACL, pp. 252–259 (2003)
7. Andersson, L., Mahdabi, P., Rauber, A., Hanbury, A.: Report on the CLEF-IP 2012 Experiments: Exploring Passage Retrieval with the PIPExtractor. In: CLEF-IP (Notebook Papers/Labs/Workshop) (2012)
8. Konishi, K.: Invalidity patent search system of NTT DATA. In: 4th NTCIR (Notebook Papers/Labs/Workshop), pp. 250–255 (2004)