Automatic Query Expansion for Patent Passage Retrieval using Paradigmatic and Syntagmatic Information

Linda Andersson **Navid Rekabsaz** **Allan Hanbury**

TU Wien, Vienna, Austria
{surname}@ifs.tuwien.ac.at

Abstract

Patent text is a mixture of legal terms and domain specific terms. Patent writers tend to paraphrase standard terminology with hypernym, hyponym and synonym substitutions in order to avoid narrowing the scope of the patent invention. The practice of paraphrasing affects the exact match retrieval function negatively. There have been many success stories addressing vocabulary mismatching using pseudo-relevance feedback and distributional semantics. However, in the patent text genre these techniques have not yielded the same level of performance as in other text genres. In this paper we propose a combination of automatic query expansion methods to identify strong domain specific lexical-semantic relations. With our method we avoid a decrease in performance and we report an improvement in recall-oriented evaluation measurements for the CLEP-IP 2013 test collection.

1 Introduction

Many different retrieval approaches have been deployed during the CLEF-IP task challenges, but there has not been a real visible breakthrough until recently (Andersson et al., 2016; Tannebaum and Rauber, 2015; Mahdabi, 2014; Dhondt, 2014). The lack of breakthroughs for patent text mining applications compared to other text genres is due to a three-folded complexity:

- The *task complexity* i.e. the information seeking process is more challenging in the patent domain than other domains (Jürgens et al., 2014; Hansen, 2011),
- The *complexity of the languages*, especially word formation, increases vocabulary mismatch (Dhondt, 2014; Ferraro, 2012)
- The *complexity of the text genre*, i.e. the frequent re-use of common words in different multi word terms (MWT) (Oostdijk et al., 2010; Temnikova et al., 2013).

The search strategy in patent search consists of many complex queries targeting main topic and sub topics (i.e. different aspects) of an invention. The search outcome depends on searchers' ability to balance recall and precision in the search sessions (Hansen, 2011). In a search session, bibliographic data is combined with phrases and words in several iterations in order to narrow the scope of the search (Tannebaum and Rauber, 2015; Jürgens et al., 2014). Each aspect of an invention can be divided into pairs of terms consisting of a general term and a specific term. If an invention has three aspects A, B and C each of these three aspect term pairs needs to be combined in the search process (Adams, 2011). The complexity of the patent search task motivates usage of automatic query expansion (AQE) techniques and terminology identification. We propose an AQE method, which incorporate syntagmatic (i.e. MTW relations) and paradigmatic (i.e. lexical-semantic relations e.g. hyponymy relations) information on vocabulary present in the patent text domain, by merging research results from two previous publications (Andersson et al., 2016, 2014). We examine three different filters deployed on an ontology, automatically populated with domain specific lexical-semantic relations. We apply pointwise mutual information (PMI) as a pure syntagmatic filter, distributional semantics as a combined syntagmatic and paradigmatic filter, and the International Patent Classification (IPC) schema as a taxonomy filter. Our main contribution shows that a combination of paradigmatic and syntagmatic information will better recognize strong domain lexical-semantic relations compared to the IPC taxonomy.

2 Related work

In non-patent genres there have been many success stories regarding different retrieval methods, especially AQE techniques such as pseudo relevance feedback (PRF), exploring distributional semantic and external resources (Manning et al., 2008). In the patent text genre the success stories with PRF (Ganguly et al., 2011; Kishida, 2003), using Wikipedia (Al-Shboul and Myaeng, 2014) or random indexing (Lupu, 2014) have not shown the same enhancement, some methods have even decreased under baseline. A plausible explanation for this lack of improvement for PRF could be that the overall poor quality of the top *K* retrieved document are more non-relevant than relevant documents (Takeuchi et al., 2005; Magdy and Jones, 2011). Another explanation is the composite nature of the text genre (Temnikova et al., 2013), i.e. the majority of domain specific technical concepts are MWT composed of common English words e.g. *bus slot card* (SanJuan et al., 2005). The patent search terms are a mixture of words and MWTs composed of broad and general concepts (Adams, 2011). For instance, a typical hyponymy relationship in patent texts would be *thrips* is a hypernym to *bulb fly larvae*). Consequently, the bag-of-word methods will be limited due to the linguistic composite of the text genre. Knowledge based (KB) AQE methods would fit the linguistic composite of the text genre better, since they address paradigmatic relations composed of explicit lexical-semantic relations of different term lengths (Mandala et al., 2000). However, KB AQE methods have not shown the same robust enhancement for document retrieval (Voorhees, 1994), in comparison to automatically constructed thesauris using syntagmatic relations (Schutze and Pedersen, 1997).

In the patent retrieval literature it has been reported that AQE KB methods, which incorporate citation graphs, classifications (e.g. IPC) and search reports increase the performance in comparison to standard PRF (Mahdabi et al., 2013; Feng et al., 2013; Tannebaum and Rauber, 2015). However, due to the tendency to avoid using standard terminology and the presence of neologism, the KB methods are limited in time and technical coverage (Nanba et al., 2009). Distributional semantic methods, which combine syntagmatic and paradigmatic information, have been successfully deployed in identification of medical concepts (Symonds et al., 2012). In (Chen et al., 2003) a patent document retrieval system incorporating syntagmatic and paradigmatic information was presented. However, due to computational complexity and the extensive pre-processing steps only the abstract was used for query formulation (QF) and they deployed a pre-defined list of MWTs, which limits the flexibility of their system.

3 Method

We experimented on the CLEF-IP 2013 test collection, which contains approximately 2.6M XML documents (representing 1.5 million patents). For the QF method and the retrieval architecture we re-used the solution presented in (Andersson et al., 2016). They compared several different features for QF such as phrases, words, bigrams, and MWTs on the English topics set (50) of the CLEF-IP 2013 passage retrieval task. Their best method **NLP** included a domain adapted NLP pipeline with additional machine learning for terminology extraction. In this experiment we examine if we can improve on the method **NLP** by adding lexical-semantic relations for AQE. The AQE is deployed on phrases since the majority of the automatically extracted hypernyms are composed of MWTs, or at least one entity of the hyponymy relation is a MWT e.g. *rape pollen beetles* and *thrips*. We re-used the seed ontology, presented in (Andersson et al., 2014), which was established by using lexico-syntactic patterns (Hearst, 1992). In the ontology, phrases such as *mechanical stress*, *remote communication*, *network lan* were extracted as candidate hyponyms for the term *communication link*. However, *remote communication* and *mechanical stress* have weaker termhood levels in comparison to *network lan* and *communication link*. Furthermore, only *network lan* and *communication link* have a hyponymy relation. In order to remove weaker candidates and noisy relations we deployed three filters:

- The taxonomy information filter explores explicit semantic categories and hierarchical structure of (**IPC**). Only terms belonging to the same sub technical field (i.e. IPC sub class) were to be used as expansion terms for a patent topic.

- The PMI filter explores the pure syntagmatic strength. The PMI was computed based upon document co-occurrence, Eq. 1, (Manning et al., 2008). For expanding over bigrams we used joint probability of all $P(w_i...w_n)$ over individual probabilities of all $P(w)_n$.
- The distributional semantic filter (**SEM**) reflects a filter composed of both paradigmatic and syntagmatic information. We computed the cosine similarity of word2vec representation for each term pair.

The word2vec model was trained on patent data and 300 dimensions were used, for further information see (Rekabsaz et al., 2017). However, word2vec methods are limited in their re-usability in the patent text domain, since they are modelled on unigrams or a fixed length of n-grams. In order to expand the existing cosine computation to include computation between arbitrary length of two MWTs, we sum the similarity values of each combination and in order to avoid bias towards longer MWTs we divided the sum by the number of token see Eq. 2. By summing up the cosine similarity of each member of the two MWTs we can be flexible regarding the number belonging to each set and thereby cover instances such as *rape pollen beetles* and *thrips*. For **SEM** and **PMI**, we decided to expand with a fixed set of terms (5, 10, 15) for each query since there were no clear cut threshold for either methods.

$$PMI = log\frac{P(x,y)}{P(x)P(y)} \quad (1)$$

- *P(x,y)* represents the probability of a given event. We defined a skip-gram constraint i.e. the words should co-occur within the range of three words additionally to the length of a given phrase sequence.
- *P(x)P(y)* represents the number of occurrence for each word independently of each other in the collection.

$$Joined_{similarity} = \frac{\sum \left(_{cos}(\overrightarrow{w_i}, \overrightarrow{w_n})..._{cos}(\overrightarrow{w_{i+1}}, \overrightarrow{w_n}) \right)}{N} \quad (2)$$

- $_{cos}(\overrightarrow{w_i}, \overrightarrow{w_n})$ represents each word vector pair cosine similarity of a given event; $_n$ the length of a given event defined by its number of members.
- *N* is the number of words for a given event.

4 Results

Table 1 shows the best performing AQE methods in comparison with the existing state-of-art (Andersson et al., 2016) (**NLP**) and the best official participant run of the CLEF-IP 2013 passage retrieval task (Luo and Yang, 2013) (**Georgetown**).

Table 1: AQE methods, state-of-the art (Andersson et al., 2016) (NLP) and best run in CLEF-IP 2013 (Georgetown) (Luo and Yang, 2013).

Run	**PRES**	**Recall**	**MAP**
NLP AQE SEM5	**0.558**†‡	**0.649**	0.269
NLP AQE PMI5	0.547†	0.631	0.270
NLP	0.544†	0.631	**0.285**†
NLP AQE IPC‡	0.477	0.568	0.244
Georgetown†	0.433	0.540	0.191

The AQE **IPC** method under performs in comparison to the pure **NLP** method (a decrease in all metrics). When applying either **PMI** or **SEM** filter, using 5 expansion terms we have slight increase in PRES (Magdy and Jones, 2010) and recall. However, the improvement is not statistically significant between **SEM** and **PMI** and the pure **NLP** method. For each metric, we performed an ANOVA to test the omnibus null hypothesis that all the runs are equal. This was rejected for MAP and PRES with ($p<0.05$), meaning that at least two runs are significantly different. The results indicate for each cell†‡ the runs to which it is statistically significantly different. As we can see, while results are visibly different, the relatively low number of topics in this track results in few clear cases of improvement.

5 Conclusion

In order to build a successful patent retrieval system we need to address:

- The *Language Complexity* in terms of linguistic characteristics. Especially word formation of new words are particular important for the patent text genre. By extracting candidate QE terms from the collection itself, we avoid the coverage issue which is the case when using external resources.
- The *Domain Complexity* in terms of diversity between general written text and the target text domain of a particular language. By recognizing the importance of MWTs in the patent text genre and adopting existing methods to handle MWTs we introduce a flexible AQE method.
- The *Task Complexity* needs to reflect the complexity of the target domain and the target language, as well as the information seeking process. Patent queries need to reflect different aspects of a patent invention. In this paper we explore AQE methods addressing both syntagmatic and paradigmatic information.

References

A. Adams. 2011. Personal correspondence. PatOlympics, 2011, Vienna.

Bashar Al-Shboul and Sung-Hyon Myaeng. 2014. Wikipedia-based query phrase expansion in patent class search. *Inf. Retr.* 17(5-6):430–451.

Linda Andersson, Mihai Lupu, Joao Palotti, Allan Hanbury, and Rauber Andreas. 2016. When is the time ripe for natural language processing for passage patent retrieval monitoring of vocabulary shifts over time. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM16.

Linda Andersson, Mihai Lupu, Joao Palotti, Florina Piroi, Allan Hanbury, and Andreas Rauber. 2014. Insight to hyponymy lexical relation extraction in the patent genre versus other text genres. In *Proc. of IPaMin KONVENS*.

Liang Chen, Naoyuki Tokuda, and Hisahiro Adachi. 2003. A patent document retrieval system addressing both semantic and syntactic properties. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, PATENT '03, pages 1–6.

Eva Dhondt. 2014. *Cracking the Patent using phrasal representations to aid patent classification*. Ph.D. thesis, Radboud University Nijmegen, Netherlands.

Wang Feng, Lin Lanfen, Yang Shuai, and Zhu Xiaowei. 2013. A semantic query expansion-based patent retrieval approach. In *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. pages 572–577.

Gabriela Ferraro. 2012. *Towards deep content extraction from specialized discourse: the case of verbal relations in patent claims*. Ph.D. thesis, Universitat Pompeu Fabra.

Debasis Ganguly, Johannes Leveling, Walid Magdy, and Gareth J.F. Jones. 2011. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM11, pages 1953–1956.

Preben Hansen. 2011. *Task-based Information Seeking and Retrieval in the Patent Domain: Processes and Relationships*. Ph.D. thesis, University of Tampere, FINLAND.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING92, pages 539–545.

Julia J. Jürgens, Christa Womser-Hacker, and Thomas Mandl. 2014. Modeling the interactive patent retrieval process: An adaptation of marchionini's information seeking model. In *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, New York, NY, USA, IIiX '14, pages 247–250.

Kazuaki Kishida. 2003. Pseudo relevance feedback method based on taylor expansion of retrieval function in ntcir-3 patent retrieval task. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing - Volume 20*. Association for Computational Linguistics, Stroudsburg, PA, USA, PATENT03, pages 33–40.

Jiyun Luo and Hui Yang. 2013. Query formulation for prior art search-georgetown university at clef-ip 2013. In *Proc. of CLEF*.

Mihai Lupu. 2014. On the usability of random indexing in patent retrieval. In *International Conference on Conceptual Structures*. Springer, pages 202–216.

Walid Magdy and Gareth J.F. Jones. 2010. Pres: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '10, pages 611–618.

Walid Magdy and Gareth J.F. Jones. 2011. A study on query expansion methods for patent retrieval. In *Proceedings of the 4th Workshop on Patent Information Retrieval*. ACM, New York, NY, USA, PaIR '11, pages 19–24.

Parvaz Mahdabi. 2014. *Query Refinement for Patent Prior Art Search*. Ph.D. thesis, University of Lugano, Switzerland.

Parvaz Mahdabi, Shima Gerani, Jimmy Xiangji Huang, and Fabio Crestani. 2013. Leveraging conceptual lexicon: Query disambiguation using proximity information for patent retrieval. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '13, pages 113–122.

Rila Mandala, Takenobu Tokunaga, and Hozumi Tanaka. 2000. Query expansion using heterogeneous thesauri. *Inf. Process. Manage.* 36(3):361–378.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Hidetsugu Nanba, Hideaki Kamaya, Toshiyuki Takezawa, Manabu Okumura, Akihiro Shinmori, and Hidekazu Tanigawa. 2009. Automatic translation of scholarly terms into patent terms. In *In Proc. of the 2nd Pair workshop*.

Nelleke Oostdijk, Eva D'hondt, Hans van Halteren, and Suzan Verberne. 2010. Genre and domain in patent texts. In *In: Proc. of the 3rd Pair workshop*.

Navid Rekabsaz, Mihai. Lupu, Artem Baklanov, Allan. Hanbury, Alexander Dür, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In *Proc. of ACL*. Association for Computational Linguistics, Stroudsburg, PA, USA.

Eric SanJuan, James Dowdall, Fidelia Ibekwe-SanJuan, and Fabio Rinaldi. 2005. A symbolic approach to automatic multiword term structuring. *Comput. Speech Lang.* .

Hinrich Schutze and Jan Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Inf. Process. Manage.* 33(3):307–318.

Michael Symonds, Guido Zuccon, Bevan Koopman, Peter D Bruza, and Anthony Nguyen. 2012. Semantic judgement of medical concepts: Combining syntagmatic and paradigmatic information with the tensor encoding model .

Hironori Takeuchi, Naohiko Uramoto, and Koichi Takeda. 2005. Experiments on patent retrieval at ntcir-5 workshop. NTCIR-5.

Wolfgang Tannebaum and Andreas Rauber. 2015. Learning keyword phrases from query logs of uspto patent examiners for automatic query scope limitation in patent searching. *World Patent Information* .

Irina P Temnikova, Negacy D Hailu, Galia Angelova, and K Bretonnel Cohen. 2013. Measuring closure properties of patent sublanguages. In *RANLP*. pages 659–666.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Springer-Verlag New York, Inc., New York, NY, USA, SIGIR94, pages 61–69.

6 Acknowledgement

This work has been partly supported by the Self-Optimizer project (FFG 852624) in the EUROSTARS programme, funded by EUREKA, the BMWFW and the European Union.