Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-210

Query Generation for Patent Retrieval with Keyword Extraction Based on Syntactic Features

Julien ROSSI [a,1,2], Evangelos KANOULAS [a]
[a] *Informatics Institute & Amsterdam Business School, University of Amsterdam, Netherlands*

Abstract. This paper describes a new method to extract relevant keywords from patent claims, as part of the task of retrieving other patents with similar claims (search for prior art). The method combines a qualitative analysis of the writing style of the claims with NLP methods to parse text, in order to represent a legal text as a specialization arborescence of terms. In this setting, the set of extracted keywords are yielding better search results than keywords extracted with traditional method such as tf-idf.

Keywords. Patent, Claims, Legal Text, NLP, Information Retrieval

1. Introduction

This work focuses on improving the effectiveness of the search for prior art. It is a high recall task, where the target is to retrieve from databases a list of documents that relate to the invention described in a patent and demonstrate whether it is indeed novel or not.

A number of methods have been proposed in the literature with the purpose of searching for prior art without the need to manually construct a Boolean Query, based on automated keyword-based query generation. Many of the previous works were based on keyword extraction out of the description or the claims, we refer to Konishi [3], Golestan et al. [1], Mase et al. [8], Lopez and Romary [4,5] and Verberne and dHondt [10].

In this work we also follow the methodology of keyword-based query generation, but different from all the previous work as it does not rely on term frequency to identify semantic salience (TF-IDF, BM25). Instead we depend on creating trees of words out of the claim section of the patent and based on these trees we attempt to identify novelty terms. Suzuki and Takatsuka [9] tackled the problem of identifying novelty-related keywords based on claims in an Information Extraction task. Our more generic method based on a constituency parser and chunking that relies on grammar to judge word salience. We show that this method can generate queries yielding to better search results, and we evaluate improved recall but also improved ranking.

[1]Corresponding Author: Julien ROSSI, e-mail: *j.rossi@uva.nl*
[2]This work is a part of a paid internship offered by the European Patent Office. Opinions expressed in this paper are authors only, and do not reflect the opinion of the European Patent Office.

To summarize, in this work we attempt to answer the following research question: *Can we improve the retrieval of relevant document by selecting keywords based on syntactical signals of semantic importance, rather than term-frequency?*

2. Methodology

In a nutshell our method works as follows: (1) Based on the complete claim set of a patent filing we generate a claim tree; (2) we then generate a specialization tree for each claim, and (3) score words based on their appearances in tree nodes; (4) we then submit the selected top-n keywords into a search engine.

2.1. Generating the Claim Tree

Each claim can establish itself as a refinement of one or more other claims. We used regular expressions to identify the dependencies and create the claim trees. [3]

Each claim is parsed by the Stanford Core NLP Constituency Parser [7], which provides the word tokenization, the POS tagging and the constituency parsing itself. Because of the unusual language of the claims, it is typical that words like "said" or "claim" are misclassified as verbs where they are instead used as relative adjectives. This incorrect POS tagging has repercussions in the chunking. We created a new annotator in the Stanford Core NLP Server to correct these tags, we refer to Hu et al. [2] for the correction.

2.2. Generating Specialization Trees

The constituency parsing tree is traversed depth-first, creating a string of tags, both POS and Chunks tags. Our set of regular expression identifies two types of patterns that can be expressed as head-to-child relation within a tree:

- Composition : a system comprising this and that, the constituency parsing allows for a lot of flexibility in the actual wording, as the chunk structure stays stable in that situation. Head node contains a system, attached to two child nodes this and that
- Specialization : a system made of this, which is on top of this and that, again the chunk structure is very stable and resistant to the diversity of the wording. Head node contains a system, attached to one child node this, itself attached to 1 child node this and that

We observe that the chunking produced by the Stanford Parser is very stable over the actual phrasing and choice of verbs, words and delimiters. It efficiently reduces lexical and morphological variations of the concepts of composition and specialization to a few chunk patterns. We leverage the chunking stability to then fold sentences into trees.

The specialization tree is the representation of one claim as a tree based on the relations of composition and specialization between chunks of the text. We identify words w as belonging to specialization tree nodes n_i: $\forall w, N(w) = \{n_i, where\, w \in n_i\}$, $\forall w, P(w) = \{(nd(n_i), nh(n_i), cd(n_i)), n_i \in N(w)\}$

[3] Strictly speaking, these are not trees, since each claim has an identified list of parent nodes.

nd, *nh* are the depth and height of the node within the specialization tree, *cd* is the depth of this claim within the claim tree of that patent.

2.3. Scoring keywords

Words are then grouped by stem, using the Porter Stemmer. For a specific stem, we record which one of all the words with the same stem had the highest number of occurrences: $\forall stem\, s,\, P(s) = \bigcup_{stem(w)=s} P(w)$

The scoring method has to favor words that are located deep within the specialization tree, as they relate to finer details of the invention, which is where we expect an invention to stand out of other similar inventions. We also want the scoring to favor words that are within claims that are deep into the claim tree, for the same reason as above, as a claim discloses finer details about the claims it depends on.

We devised two scoring methods: $CLST05(s) = \sum_{P(s)} e^{\alpha_{05} * \frac{nd}{nd+nh-1} + \beta_{05} * cd}$, and $CLST06(s) = \sum_{P(s)} e^{\alpha_{06} * \max(nd) + \beta_{06} * \max(cd)}$

The hyperparameters $\alpha_{05}, \alpha_{06}, \beta_{05}, \beta_{06}$ are determined by experimenting and keeping the values that generate the highest metrics.

The top-n stems with the highest scores are selected to construct a query, which is the concatenation of the words associated with these stems.

3. Experimental Setup

Dataset. We used the CLEF-IP 2011 Topic Collection as a basic dataset. This collection contains patent documents with qrels to identify the definitive list of relevant documents for each case. The search database is the complete historical worldwide repository of patents. The setting is to search for relevant documents based on the claims from the seed document, searching through the claims of the documents in the corpus.

Search Engine. We used an instance of Lucene search engine.

Baseline. The baseline is a system developed at the EPO, known under the MLT acronym, configured to act within the same parameters, using the claims as source for query, and the claims within the search database as corpus. We use it as it is representative of related work that uses TF-IDF term weighting to extract keywords.

System Configuration. Our methods are called CLST-05 and CLST-06, and each method has three variants: "As is", "BOOST" where the word scores are used as boost factors for the search engine, and "NO-RETAG" where the POS tagging is not corrected. The boost increases the scoring of a document that contains the boosted terms. We used the Lucene instance in place at the EPO. This search engine receives our query and returns a list of ranked search results.

Evaluation Metrics. We evaluate the performance of the different keyword-based query generation methods on the basis of Recall@100, and PRES@100, similar to the evaluation performed under CLEF-IP. PRES is a metric introduced by Magdy and Jones [6]. [4]

[4] We had to correct the formula presented in the original paper as it was producing results out of the range $[0,1]$. We used $\sum r_i = \sum_{i=1}^{nR} r_i + \sum_{i=nR+1}^{n} (N_{max} + n - (i - nR - 1))$

4. Results and Analysis

In the first place, we can compare the average PRES@100 and Recall@100 to the baseline. We developed 6 different systems that we can evaluate. For each system we can also select how many words we extract as keywords, from 50 to 100 by increment of 10. The performance of all those systems is given in Table 1.

Table 1. PRES@100 and Recall@100

	PRES@100					**Recall@100**					**Summary**	
	Number of Keywords					**Number of Keywords**					**R@100**	**PRES@100**
System Name	60	70	80	90	100	60	70	80	90	100		
CLST-05	0.18	0.19	0.19	0.19	0.19	0.24	0.24	0.25	0.25	0.25	**0.2479** (***)	**0.1923** (***)
CLST-05B	0.14	0.15	0.15	0.15	0.15	0.19	0.19	0.20	0.20	0.20	0.2019 (***)	0.1525 (***)
CLST-06	0.18	0.18	0.19	0.19	0.19	0.23	0.24	0.24	0.24	0.25	0.2463 (***)	0.1918 (***)
CLST-06B	0.12	0.12	0.12	0.13	0.13	0.16	0.17	0.17	0.17	0.18	0.1749	0.1275
CLST-06NR	0.18	0.18	0.19	0.19	0.19	0.23	0.24	0.25	0.25	0.25	0.2467 (***)	**0.1925** (***)
CLST-06NRB	0.12	0.12	0.13	0.13	0.13	0.16	0.17	0.17	0.18	0.18	0.1756	0.1294
MLT		0.13					0.17				0.1742	0.1325

The summary clarifies which results are a statistically significant improvements over the MLT baseline (randomization test, *** means $p < 0.001$)
The results show that BOOST is significantly decreasing performance. This can be interpreted as the scoring system having a good effect on selecting salient keywords over more general words, but not being able to catch the variations in relative importance of words in a way that is numerically in line with the boost factors of the search engine.

The correction of the POS tagging, which was tried only on the system CLST-06, does not generate a statistically significant improvement over the vanilla version. We analyze that the distortions on the parsing occur at different places than those where a specialization or combination occurs, which makes the system oblivious to this correction to a certain extent. Nonetheless, well keep this correction in mind for future work, especially when additional features get extracted from the dependency parser.

The evaluation metrics keep increasing with the number of keywords, the difference being statistically non-significant between 80, 90 and 100 keywords. Our system overperforms the existing system, with a statistically significant improvement with 30 keywords. We keep the results based on 100 keywords.

The significant result is that both CLST-05 and CLST-06 largely outperform the TFIDF-based baseline. Results show significant improvement in this setting of Query Generation, although the setting mixes the performance of the keyword extraction and the tweaking of the underlying search engine. Nonetheless, the approach of going away from term frequency methods to identify salient words in presence of an enforced writing style is proven to make sense. The term frequency allows for identification in absence of other information on how the text is written, while we can leverage the additional information that authors are restricted to deliver information in a way that is reflected in grammar, thus enabling us to work at the semantic level by working at the grammatical level.

5. Conclusion

In this work we used the sentence morphological features to identify keywords within patent claims, and used these keywords as query terms to retrieve other relevant patents. In this setting we establish a significant improvement over the existing baseline, based on term-frequency weighting methods.

In the future we plan to apply and expand this work on other text sources with constrained writing style. We also see potential in adapting NLP tools that were designed or trained on conventional literature.

References

[1] Far, G., Mona, Sanner, Scott, Bouadjenek, Reda, M., Ferraro, Gabriela, David, H.: On term selection techniques for patent prior art search. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 803–806. SIGIR '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2766462.2767801, http://doi.acm.org/10.1145/2766462.2767801

[2] Hu, M., Cinciruk, D., Walsh, J.M.: Improving automated patent claim parsing: Dataset, system, and experiments. CoRR **abs/1605.01744** (2016), http://arxiv.org/abs/1605.01744

[3] Konishi, K.: Query terms extraction from patent document for invalidity search. In: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-5, National Center of Sciences, Tokyo, Japan, December 6-9, 2005 (2005)

[4] Lopez, P., Romary, L.: Multiple Retrieval Models and Regression Models for Prior Art Search. In: CLEF 2009 Workshop. p. 18p. Corfu, Greece (Sep 2009), https://hal.archives-ouvertes.fr/hal-00411835

[5] Lopez, P., Romary, L.: Experiments with citation mining and key-term extraction for Prior Art Search. In: CLEF 2010 - Conference on Multilingual and Multimodal Information Access Evaluation. Padua, Italy (Sep 2010), https://hal.inria.fr/inria-00510267

[6] Magdy, W., Jones, G.: Pres: A score metric for evaluating recall-oriented information retrieval applications. In: SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 611–618 (9 2010). https://doi.org/10.1145/1835449.1835551

[7] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 55–60. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/P14-5010, http://www.aclweb.org/anthology/P14-5010

[8] Mase, H., Matsubayashi, T., Ogawa, Y., Iwayama, M., Oshio, T.: Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing **4**(2), 190–206 (Jun 2005). https://doi.org/10.1145/1105696.1105702, http://doi.acm.org/10.1145/1105696.1105702

[9] Suzuki, S., Takatsuka, H.: Extraction of keywords of novelties from patent claims. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 1192–1200. The COLING 2016 Organizing Committee (2016), http://www.aclweb.org/anthology/C16-1113

[10] Verberne, S., d'Hondt, E.: Prior art retrieval using the claims section as a bag of words. In: et al., C.P. (ed.) CLEF 2009 Workshop, Part I, LNCS 6241. pp. 498–502. Springer-Verlag Berlin Heidelberg (2010)