Patent Terminlogy Analysis: Passage Retrieval Experiments for the Intellecutal Property Track at CLEF

Julia Jürgens, Sebastian Kastner, Christa Womser-Hacker, and Thomas Mandl

University of Hildesheim, Information Science, Marienburger Platz 22,
31141 Hildesheim, Germany
{juerge, kastners, womser, mandl}@uni-hildesheim.de

Abstract. In 2012, the University of Hildesheim participated in the CLEF-IP claims-to-passage task. 4 runs were submitted and different approaches tested. The tested approaches included a language independent trigram search approach, one approach formulating a query in the source language only and another approach with querys translated to English, German, French and Spanish. The results were not satisfactory and the task of passage retrieval as defined in CLEF-IP proved to be difficult for current technology.

General Terms

Experimentation

Keywords

Intellectual Property, Evaluation, Patent Retrieval System, Natural Language Processing

1 Introduction

In 2012, the Vienna University of Technology organized the CLEF-IP track and provided interested participants with several new tasks. Besides the claims-to-passage task, which is meant to represent the scenario of the patentability or novelty search, a flowchart recognition and a chemical structure recognition task were offered.

The data collection is a subset of the MAREC[1] corpus and consists of approximately 3.5 million XML documents from the European Patent Office (EPO) and the World Intellectual Property Office (WIPO) [1]. The training set was built up of 51 topics and the matching qrels with relevant passages from both EP and WO documents.

[1] http://www.ir-facility.org/prototypes/marec

The University of Hildesheim integrated different approaches which included a language independent trigram search, one approach formulating a query in the topics source language only and another approach with querys translated to English, German, French and Spanish.

This years claims-to-passage task is based on manual relevance assessments from patent specialists taken out of search reports. This provides the participants with a realistic data basis, which should be positively emphasized.

2 Pre-analyses

In order to get a better understanding of the task and the relevance assessments, the topics and the qrels from the trainings set were examined from different perspectives. The following paragraphs describe our approaches and results.

2.1 Overlap claims/passages

Since the task is to find relevant passages on the basis of given claims, it seemed natural to try to look for patterns and connections in the supplied training data. Given that the 51 topics were built from only 29 documents and relevant passages were often similar for two topics taken from the same document, we chose to examine ten topics (from distinct documents), which corresponds to about a third of the training data.

This first manual analysis, which was complemented by using an Online Text Analysis Tool[2], demonstrated that the absolute overlap of meaningful vocabulary (not counting stopwords nor regular patent terms) is minimal. In the ten documents, there were between two and seven matching terms in their base form comparing the query and the result set. In order to get to these comparable base forms, several methods like stemming, compound splitting and translation would have to be used. In three topics, two matching phrases consisting of two words (e.g. meat product) were also found. There was never a single overlap of longer phrases.

This analysis shows that it is extremely hard to find relevant passages just on the basis of linguistic methods/matching terms. This knowledge led us to examining other approaches, e.g. the overlap of classification codes.

2.2 Overlap classification codes

In real patent search scenarios, patent experts make immense use of the classification codes. Therefore, we analyzed the overlap of the classification codes from the topic documents and the documents with relevant passages. We automatically calculated the overlap of the IPCR-codes on the section, class and subclass level to be able to better decide if and where a cutoff could be made during retrieval.

[2] http://www.online-utility.org/text/analyzer.jsp

Table 1. Precision and Recall by IPCR level based on the training data

	Section	Section & Class	Section & Class & Subclass
Recall	99.82%	97.75%	97.75%
Precision	0.01%	0.07%	0.16%

For the calcuation of the overlap we calculated how many of the relevant passages were assigned to at least one of the topic document's IPCR codes section, section & class or section & class & subclass. The results (Table 1) showed that using this constraint a very high recall could be achieved. It also showed that the high recall is accomplished only at the expense of precision and therefore the usage of IPCR codes alone will not suffice to reach satisfactory results.

2.3 Language distribution

To get a better understanding on how to cope with multilingual challenges, the distribution of the languages in the data collection was analyzed. For the identification of the language we relied on the metadata supplied in the patent documents XML structure. The results were very interesting since they exposed that 40% of the documents neither contain abstract, nor claims nor description. The language distribution in the corpus can be seen in Table 2.

Table 2. Language distribution

	Abstract	**Claims**	**Description**	**Avg.**
en	44.83%	48.34%	41.40%	44.86%
unknown	50.88%	39.99%	39.98%	43.62%
de	11.22%	29.87%	14.04%	18.37%
fr	12.94%	23.81%	4.48%	13.74%
es	0.04%	0.05%	0.05%	0.05%
other	0.00%	0.05%	0.05%	0.03%

Table 2 shows that not one single language is predominant, but numerous patents are written in English, German and French. Therefore a translation of the documents was considered to be useful. The other languages included Russian, Japanese, Italian, Korean, Danish, Chinese, Swedish and Dutch.

3 System setup

The system used for both the training and the tests was based on Java and Apache Lucene and mostly on Lucene's built in language processing modules for language depending stemming, stopword removal etc. All translations were created with the help of the Microsoft Translator API.

3.1 Preprocessing and Indexing

Since the claim to passage task requires the matching of claims to passages in a patent, the passages in the patents were considered as documents themselves, and hence a passage based index was created. For each passage, the passage content was indexed and the patent document it appeared in was referenced by the patent document's file name.

For the training phase two indexes were created: a language independent trigram index (trigram index) and an index containing the passages original content as well as translations in clear text form (text index). Including the translations, the text index contained the passage content in the 4 most common languages in the CLEF 2012 corpus: English, German, French and Spanish.

For both the text index and the trigram index the passage content was processed before the actual indexing. The processing consisted of tokenization, lowercasing, stopword removal and stemming with language specific tools provided by the Lucene project. Hereby the language was detected by the data supplied by in the patent document files' XML structure.

The indexes had the following fields in common:

- **file name:** the name of the file the passage appeared in
- **IPCR** all the IPCRcodes assigned to the patent document the passage appeared in
- **language:** the language the passage was originally written in (according to the XML annotations in the patent document file)
- **type:** the type of the passage, i.e. claim, description, heading or abstract

Fields unique to the text index:

- **content_de:** German content
- **content_en:** English content
- **content_fr:** French content
- **content_es:** Spanish content

Instead of the language specific content fields, the trigram index contains only one language independent content field:

- **content:** the processed content of the passage split into trigrams

3.2 Search Process

Before searching, the contents of the topic passages were merged and subsequently processed. The text that was hereby created will from now on be referenced as topic text. The topic text was processed in the same way as the passage contents as described in Chap. 3.1.

Furthermore, extended stop word lists were created to reduce noise in queries. Since the topics are regular claims in the corpus, the 100 most common words in the corpus were manually analysed and searched for irrelevant words to find a broader list of stop words. The extended lists were created for all the topic

languages, i.e. English, German and French. The indexes remained untouched by the extended stop word lists.

3 different types of search were implemented:

1. **Language independent trigram based search (S1):**
   After processing the text was split into trigrams. The search itself was implemented language independent and there was no distinction between trigrams of different languages.
2. **Search with single language topic text (S2):**
   The topic text was not translated and it was only used in its original language. A boolean query was formed to query every content field in the clear text. A systematical outline of such a query:

   ```
   (content-de:TERM-1 OR content-en:TERM-1 OR content-fr:TERM-1)
   OR (content-de:TERM-2 OR content-en:TERM-2...)
   ... OR (content-de:TERM-X OR content-en:TERM-X...)
   ```

3. **Search with multilanguage topic text (S3):**
   The topic text was translated using the Microsoft Translator. Boolean queries were formed, querying the language specific field with the respective translation of the topic text only. A systematical outline of such a query:

   ```
   (content-de:DE-TERM-1 OR content-de:DE-TERM-2 OR
    content-de:DE-TERM-3 ... content-de:DE-TERM-X)
   OR
   (content-en:EN-TERM-1 OR content-en:EN-TERM-2 OR
    content-en:EN-TERM-3  ... OR content-en:EN-TERM-X)
   ...
   ```

Additionally to those types of search, a set of parameters meant to change and improve the search results, were defined. Those parameters were:

- **Boost Factor:**
  The Boost Factor, which is integrated into Lucene, was used to give a higher weight for reoccuring terms in the claims for search. For the calculation of the factor each term was boosted by the following formula was used:

$$1 + (BoostFactor * (TermFrequency - 1)) \tag{1}$$

- **Passage Limit:**
  The maximum amount of total passages to be retrieved
- **Extended Stopwords:**
  Denotes whether the extended stopwords should be used or not.
- **IPCRLevel:**
  Should the IPCRCodes be considered, i.e. should paragraphs only be returned, if at least one of the IPCRCodes of the document they were found in is consistent to at least one of the IPCRCodes of the topic document. If the IPCRCode was to be used, the IPCRlevel, i.e. the number of characters of the IPCRCode to be considered, could be defined as well.

3.3 Experiments

As a baseline the search with single language topic text (S2) was chosen. As parameters, a Boost Factor of 0.5, a passage limit of 100, an IPCRlevel of 1 and no extended stopwords were used. With those settings, a precision of 0, a recall of 0.01 and an f-measure of 0 was achieved.

In order to improve those poor results, a batch tool was written, to allow the testing of several parameter configurations and combinations at once. Using the training set, precision, recall and f-measure were calculated for those combinations. Using the batch tool, a wide set of parameters was tested but no combination resulted in satisfying results. The best results with a precision of 0.01, a recall of 0.02 and an f-measure of 0.02 were achieved with a boost factor of 0, a passage limit of 200, an IPCRlevel of 6 and without extended stop words, using the single language approach (S2). Those results were not significantly better than many other combinations resulting in a precision of 0, a recall of 0.1 and an f-measure of 0.1.

3.4 Submitted Runs

After the experiments, the configurations that obtained the best results were chosen and handed in for the runs. This also meant that the trigram based approach was dropped and not handed in since it never yielded better results than the baseline. An overview of the runs and a description of their configurations can be seen in Table 3.

3.5 Results

The results (Table 4) show that the task of passage retrieval as defined in CLEF-IP is inherently difficult for current technology. The overall values are not satisfying. Only one group could achieve better results for the passage retrieval evaluation measures.

4 Outlook

We observed that the terminology between the legal parts (claims) and the technical parts (abstract and description) differs quite substantially. In the future. we intend to exploit these differences and knowledge about the distribution of terms in both parts. Optimized queries for the technical and the legal part of the patent will be created and sent to the proper index fields. In addition. a linguistic phrase parser [2] which has been used for the participation in 2011 [3] will also be adapted for the passage task.

Table 3. Explanation of submitted runs

Run	**Description**
multi-bf1-estrue-il1-pl100	– Single Language Topic Text (S2) – Boost Factor: 1 – IPCR Level: 1 – Extended stopwords: yes
multifield-bf0-il6-pl100	– Single Language Topic Text (S2) – Boost Factor: 0 – IPCR Level: 6 – Extended stopwords: no
translated-bf0-il6-pl100	– Multi Language Topic Text (S3) – Boost Factor 0 – IPCR Level: 6 – Extended stopwords: no
translated-bf1-estrue-il6-pl100	– Multi Language Topic Text (S3) – Boost Factor 0 – IPCR Level: 6 – Extended stopwords: yes

Table 4. Results of submitted runs

Run	**Passage Level**		**Document Level**		
	MAP(D)	**Prec.(D)**	**PRES@100**	**Recall@100**	**MAP**
multi-bf1-es-il1-pl100	0.0028	0.0087	0.0999	0.1173	0.0405
multifield-bf0-il6-pl100	0.0038	0.0170	0.1186	0.1544	0.0291
translated-bf0-il6-pl100	0.0027	0.0142	0.1117	0.1439	0.0340
translated-bf1-es-il6-pl100	0.0030	0.0192	0.1217	0.1653	0.0440

References

[1] Piroi, F., Lupu, M., Hanbury, A., Zenz, V.: Clef- ip 2011. In: CLEF (Notebook Papers/Labs/Workshop) 2011. (2011)

[2] Schulz, J.M., Becks, D., Womser-Hacker, C., Mandl, T.: A resource-light approach to phrase extraction for english and german documents from the patent domain and user generated content. In: Eighth International Conference on Language Resources and Evaluation (LREC) Istanbul. (2012)

[3] Becks, D., Eibl, M., Jürgens, J., Kürsten, J., Wilhelm, T., Womser-Hacker, C.: Does patent ir profit from linguistics or maximum query length? In: Working Notes 11th Workshop of the Cross-Language Evaluation Forum. CLEF 2011. (2011)