Cross Lingual issues in patent retrieval

Leo Sarasúa
European Patent Office
Rijswijk, The Netherlands
lsarasua@epo.org

Technical prior art searching in patent publications has specific characteristics, which set it apart from typical Information Retrieval (IR) applications. The language used in patents is generally well structured and gramatically correct, but tends to use generic terms and vague expressions, not to narrow the scope of the inventions. This greatly complicates the task of prior art searching and reduces the precision of the results. Further, the extensive use of acronyms and new words poses a great challenge to any Information Retrieval (IR) System.
Prior art disclosures are valid irrespective of the language used; so patent searching must be cross lingual to be complete. When searching in multiple languages the expansion of the query has to take the above aspects into account. A Cross Lingual IR system (CLIR) should also consider the multilingual capabilities of the users, as is the case of the European Patent Office.
Considering all the distinctive features of natural languages, the challenge is to create a canonical representation for language in general which can accommodate the different aspects of most languages, while being adapted to the special needs of patent retrieval.
In this paper we shall present bSmart, a prototype of search engine developed in the framework of the European Patent Office, which uses a shallow Case Based Grammar to analyze languages such as English, German, Spanish, or Japanese.

The main task of a patent examiner is the search of prior art disclosures. Other tasks are the classification of documents into a hierarchical structure of technical fields (the International Patent Classification, IPC) and the examination of the retrieved documents to assess the patentability of a patent application. The main purpose of document classification is to maintain the system of retrieval based on keywords and other tags manually assigned to documents, which convert the problem of Information Retrieval into one of Database Management.
Present systems rely heavily on manual classification schemes for retrieval. An exploding volume of patent applications renders this practice costly and ultimately untenable, making essential the use of an adequate IR tool.
An optimistic view is that new developments in IR will achieve a level of performance and quality good enough to make unnecessary the use of these expensive classification schemes. Therefore, in this paper we will focus on technical prior art retrieval. Our goal in this study is to identify those features inherent in patents, which could be exploited to simplify this task.

The first major difference between patent retrieval and other applications is the detail of the search: normally we do not search for documents on “electronic memory testing”, but rather “testing an electronic memory with a counter which skips odd or even positions”, or even more detailed.

Secondly, patent searching is mainly carried out by experts in a technical field, who interactively assess whether a retrieved documents is indeed relevant to the query. This allows a certain margin of error in the precision of the results, and the major consequence is that patent searching becomes mainly a problem of document retrieval rather than information retrieval. It is not uncommon to find complex descriptions such as: *“A compound with the general formula: CC-O-FA, in which CC is a carbon chain with at least 10 Carbon atoms, O is an Oxygen atom, and FA is an acillic group”*. This clearly requires a deep analysis of the expression, maybe with regular expression techniques. However, in the vast majority of the cases, it suffices to retrieve the relevant documents (preferably the relevant passages too) to let the user analyze them.

This is not only an advantage, but also a constraint imposed by another aspect of patents: the vagueness of the language used. Using too specific words could limit the protection of a patent. The implications for IR

are that patent attorneys tend to use vague and general expressions, such as “means compressible along an axis” instead of “a spring”. The challenge for IR is obvious, particularly when these vague expressions need to be translated to other languages for searching. In this case, query expansion combined with translation can yield a combinatorial explosion of terms and results.

Unlike distributed data sources, such as Internet, the structure of a patent document is rather homogeneous, which simplifies the processing of the text. Despite the use of general expressions, the language is gramatically correct and avoids colloquialisms, thus allowing its efficient analysis.
Further, the syntactic constructions employed in a typical patent publication are just a subset of the constructions considered gramatically correct in any language. For example, some verb forms, like the imperative form, the first or second person are never used. Past and future tenses are uncommon as well. These limitations allow the construction of efficient parsers with a very high degree of correctness.

The most viable approach for a CLIR system is to first translate the query into the target languages and then to carry out multiple monolingual searches in each of these languages [3]. Using a structured dictionary, which clusters words according to their different meanings, ensures that the translation and the expansion of the query will be more precise. This has the drawback, however, that maintenance of the dictionary becomes more complicated and error prone. Updating complex dictionaries with different relations such as homonyms, hyponyms, etc… (e.g. Wordnet) can represent a task too difficult for users without linguistic background. This difficulty is even more serious when using multilingual queries. Fortunately, when the system is destined for multilingual users (such as the European Patent Office), the task can be enormously simplified. In this case, the capture of new terms (a very real need in the world of patents, since new inventions demand new words) can be assisted by users. Also, query translation becomes independent of whether it is pre- or post-expansion, if the user can select both the synonyms and their translation.

Disclosing an invention to the public is the price that an inventor has to pay for the legal protection confered by a patent. It is logical that many inventors try to “disguise” their invention within large, unreadable documents. It is very common to find the relevant passages of an invention described in just a few lines in a document with, perhaps, hundreds of pages. This means that standard models, such as vector space, are not applicable in a search for technical prior art. We propose instead a “best match” metric, in which all sentences are considered independently (except for resolving ellipsis ambiguity across sentences) and in which the relevance of a document is the best match of any of its sentences.

Each language has its own unique characteristics with respect to IR. In English, one of the main stumbling stones is the Part Of Speech ambiguity of many words, mainly between nouns and verbs. For example, “a train” and “to train”. German is a more computer friendly language, once the stemming problem of compound words is solved (for instance, “Informationssucheergebnisse”). Romance languages such as French, Spanish or Italian, pose a challenge at the morphological analysis stage, due to the multiple suffixes and declensions of verbs and other words. They also present a problem with the construction of Noun Phrases by means of the prepositions “de”, “des”, “di”, etc… which slightly complicates syntax analysis.
A correct morphological analysis becomes essential in languages such as Japanese, in which word boundaries are not defined in written text, and where a suffix can change the whole meaning of a sentence (for instance “nomimasu”, “nomimasen”).
The question is then whether it is possible to find a canonical representation common to most languages, so that a single CLIR system can handle these particular constructions.
Despite all the differences, we find that phrases, particularly noun phrases, appear in all languages and therefore, can be used as the basic structure for indexing. Some authors have found similar concepts, like complex nouns [2], which could be used as well. Although we only have limited results, there is strong evidence suggesting that phrasal indexing could also yield higher precision than word indexing.

A patent retrieval system should be capable of handling the following common queries:

- numerical qualifiers: “a car with 5 wheels” could be inventive because of the number of wheels; however, numbers are usually considered stop words, even when written in words (“five”).
- negative queries: in standard systems, the query “a detergent without bleach” would likely return all documents *with* “bleach”.

- Nominative vs. Accusative disambiguation, or the problem of “who does what to whom”. We could mention the classical example of “a dog biting a man” vs. “a man biting a dog”. This is a problem extensible to all languages.
- Identification of ellipsis. When an invention is described in detail througout a document, it is natural to avoid repeating the subject. This problem occurs in all Western languages, and to a lesser degree in Japanese.

Also, the large size of the databases (around 20 Tbytes) imposes heavy constraints on the performance of the system.

All these factors suggested the architecture of our bSmart system, developed in the framework of the European Patent Office.
It combines phrasal indexing and a Case Based Grammar. The parser is probabilistic, with a simplified stemmer (removing normal suffixes “-ed”, “-ing”, etc... and prefixes) without lexicon look-up, for higher speed. Ellipsis across sentences is recognized however. The result is a shallow noun-phrase representation which preserves case.
Since indexing is a language dependant task, a modular system for each different language was the natural choice.

Weighting scheme

We found that the weight of the match between terms is the factor with most impact on the results. Many weighting approaches have been proposed [4]. In our case, we found the following as the most appropriate formula:

$W_i = pip_i \times idf_i$

where w_i is the weight of a given word,
pip_i is what we denominate “position in phrase” factor,
idf_i is the standard “inverse document frequency”.
The term pip is included to give more importance to the more relevant words of a phrase. For example, in the phrase “building block” the head nouns is “block”. However, in “block building”, “block” acts a qualifier of “building”, which is the head noun. In general, in English we find that the last word of a noun phrase is the head noun and, therefore, should be given more weight.

The striking thing is that we do not use the traditional tf factor, particularly because of what was explained about the “disguising” of inventions within patent documents.
Further, the idf coefficient is calculated only per technical classes, so that the same term will have different idf’s in different fields. This greatly improves the results and also allows the splitting of the database into sub-parts, which can be indexed independently.

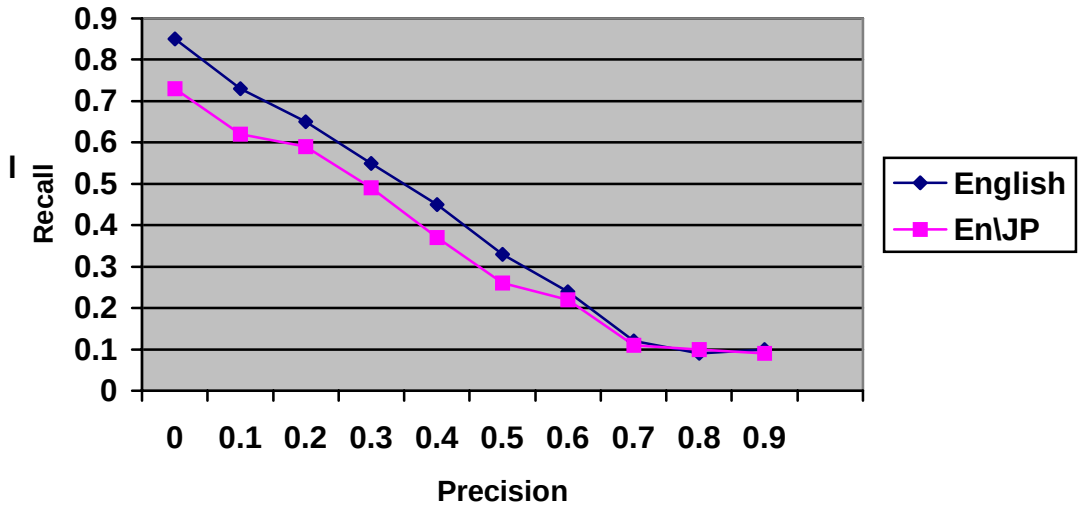
In Romance languages such as French or Spanish, the construction of a noun phrase does not follow the straightforward pattern of English and normally nouns have the inverse weight as in this language, i.e. the first word of an NP is the head noun, and the remaining words qualify it. In Japanese, we find a mixture of both cases, further complicated by the problem of unclear word boundaries. For a series of good examples, see Matsumura et al. [1].

Once the pip factors of each word are calculated, they are used to construct a normalized vector. The similarity between phrases is computed by the dot product of this vector and that of the sentences in the query.

Results

The experiment we performd consisted in 4 monolingual (English) searches in a collection of documents belonging to an IPC class, with 2280 documents, and 2 English searches in a collection of 623 Japanese

documents. The results were assessed by two experts in the technical field. The average length of the queries was 23 words.
The results we obtained can be summarized in the following recall/precision diagram:



The results are better than expected with a non patent database. The slight differences between the monolingual and the multilingual runs show that it is possible to achieve Japanese searches even if the user's knowledge of this language is very limited.

Conclusions

We have outlined the major differences between normal IR systems and patent retrieval. We have tested a prototype which exploits this differences to produce very promising results as a CLIR tool for patents.

References

[1] Atsushi Matsumura, Atsuhiro Takasu & Jun Adachi . The Effect of Information Retrieval Method Using Dependency Relationship Between Words. In *RIAO'2000 Proceedings.*
[2] Eduard Hoenkamp, Rob de Groot. Finding Relevant Passages using Noun-Noun Compounds: Coherence vs. Proximity. In *SIGIR '2000 Proceedings.*
[3] Ruth Sperer, Douglas W. Oard . Structured Translation for Cross-Lingual Information Retrieval. In *SIGIR '2000 Proceedings.*
[4] Gerard Salton, Christopher Buckley. *Term-Weighting approaches in automatic text retrieval.*