

Combining Semantics and Statistics for Patent Classification

Franck Derieux¹, Mihaela Bobeica¹, Delphine Pois², Jean-Pierre Raysz²

¹R&D, Jouve Lens, 30 Parc d'Activités du Gard, 62300 Lens, France

²R&D, Jouve Mayenne, 1 rue du Dr Sauvé, 53100 Mayenne, France

Abstract

For the patent classification task of the 2010 CLEF-IP evaluation we have used three different approaches combining semantics and statistics-driven techniques: first approach is based on an indexing-retrieval method using the Lemur system enhanced with a class calculation algorithm; the second approach combined a semantics-driven technique for class model building and the use of an advanced statistical classifier; the third approach combined the two previous methods, attempting to exploit their complementarity for results quality improvement. The results obtained for our system are encouraging: we ranked second in terms of precision on first candidate, which is, from an application point of view, the most pertinent score.

Introduction

The CLEF-IP track, a new track in the well-known CLEF evaluation campaign, was launched in 2009 to investigate IR techniques for patent retrieval. In 2010, CLEF-IP includes two types of tasks: a Prior Art Candidate Task and a Patent Classification Task. We have participated in the latter.

The patent classification task consists of classifying a patent document according to its IPC classes, namely the 3rd level classes (ie. subclasses).

EP Patents, Patent Classification and the IPC/ECLA Classifications

Patent applications classified at the European Patent Office have mainly two types of classes:

- IPC (International Patent Classification) classes;
- ECLA (European CLAssification) classes;

The ECLA and the IPC are hierarchical structures, divided and labelled in sections, classes, subclasses, groups and subgroups. At each sublevel of the hierarchy, the number of categories is multiplied by about 10.

The ECLA is an extension of the IPC and is the double of IPC in terms of number of classes: on the lowest level, the ECLA has about 135 000 classes, while the IPC contains 70 000 classes. Up to IPC subgroup level, the ECLA and IPC classification symbols are in most cases identical. The ECLA is allegedly more precise, more homogenous and more systematic than the IPC.

Neither the ECLA nor the IPC is a natural semantics classification but rather a search-oriented classification, used for prior art search. Many classes have an artificial composition and contain many limitations, exceptions, priorities etc:

- overlapping classes:

eg. C07 ORGANIC CHEMISTRY;
C08 ORGANIC MACROMOLECULAR COMPOUNDS;

- open class titles:

eg. C09K MATERIALS FOR MISCELLANEOUS APPLICATIONS, NOT PROVIDED FOR ELSEWHERE;

- unclear class boundaries, exceptions, precedence, limitations of scope:

eg. A42B HATS; HEAD COVERINGS (headbands, head-scarves A41D)

There are also many placement rules, references, indexes that represent a separate classification of patents according to special aspects of the invention, such as the technique or technology employed. Therefore, some IPC classes and subclasses serve the double purpose of classification and indexing.

These class features represent real difficulties for modelling. Moreover, training corpora available are made of patent documents that were manually classified by patent examiners. The document classifications are therefore characterized by the subjectivity inherent to any human, manual task.

It should also be noted that patent examiners use patent law knowledge for classifying a patent application. The use of this kind of knowledge can render the class choices obscure for non-specialists. For example, even if, in most cases, an examiner's classification work is patent claim-based, in some cases, the classes associated to a patent application are based on the information provided in the embodiment(s) detailed in the description. The use of this kind of knowledge makes it often challenging for an automatic system to choose as best candidate the same class as the patent examiner.

These classification-tree rules and manual classification rules used by patent specialists, together with the specificities of the vocabulary, the patent family particularities and multi-class labelling of patents make the automatic classification of patents, by far, one of the most difficult document classification tasks.

This paper describes the methods that we used for the CLEF-IP patent classification task.

We first describe the corpus provided by the organizers and we pursue by describing the structuring and the pre-processing of the corpus.

In sections 4, 5 and 6 we describe the different approaches that we have used for the classification of the test documents. Finally, we discuss and conclude on the results obtained for the evaluation.

Previous Experiments and Useful Observations

Patent classification is a task that has been of interest to us for some time. Before the CLEF-IP evaluation campaign, we had already trained and tested on several patent data sets.

In the course of these previous experiments, we had made a certain number of observations related to the difficulties encountered in patent data, the challenges related to IPC classes, the classification rules used by patent examiners, the vocabulary specificities, the particular structure, content and sections found within patent documents.

Based on these observations, we developed methods for document selection, pre-processing and annotation, term extraction and filtering for document representation, threshold setting, feature weight calculations, parameter selection, and model building for automatic patent classification.

The methods used in the context of the CLEF-IP evaluation take advantage of these observations and methods. However, the detail of these observations and methods is out of the scope of this evaluation campaign and will not be presented hereafter. Where appropriate, references will be made to previous experiments used as background for the approaches taken for the CLEF-IP evaluation.

Corpus for the CLEF-IP Evaluation

Training Corpus

The training data is composed of all EP documents that have an application date before 2002, representing a total of about 2.7 millions documents corresponding to 1.3 million patents. The corpus is structured by "invention", that is, a folder contains the A and the B documents associated to an invention.

The corpus contains English, French and German documents, distributed as follows: 68% English documents, 24% German documents and 8% French documents.

The documents are in XML, in a format based on the international standard ST.36. The complete documents include bibliographic data (abstract included, for kind A documents), the description and claims sections.

The training corpus also included a certain number of empty documents (just bibliographic data present) or incomplete documents, containing only titles and abstracts, but no description and claims.

Test Corpus

Concomitantly with the training data, the IRF made available to the participants a small test corpus organized by topics: a topic is equivalent to a patent application (A1, A2 or A9). The test documents were complete, that is the title, abstract, description and claims are all present in the test files.

The test data comprises 2000 topics published after 2002, of which 1468 are in English, 409 in German and 123 in French.

Document Sections Used and Pre-processing

In our previous experiments, we have noted that methods based only on patent titles and abstracts score lower than approaches based on the full content of the patent documents. This is mainly related to the classification rules included in the classification trees and the very specific class association rules used by patent office examiners.

Therefore, in the context of this evaluation work, we have used the entire document content: title, abstract (where available), description and claims.

Therefore only full documents were used: title, description, claims, abstract, if available. For each invention, the latest version has been chosen, with a preference order going from kind B documents to A9 and A1/2.

After filtering as described above, about 670 thousands documents have been used for the English training; around 240 thousands documents for German and 75 thousands for French.

The linguistic pre-processing of the corpus included the classic steps such as phrase segmentation, tokenisation, POS-tagging, lemmatization.

Specific patent-oriented pre-processing included a “key-phrase” tagging step, namely the detection and tagging of those parts of the description that concisely describe the subject of a patent document.

During the pre-processing phase, language inconsistencies have been found: in some documents, the value of the language attribute found in the abstract/description/claims tag did not correspond to the actual language of the tagged text. Given that our pre-processing and model construction approaches are language-based, we have tried to estimate, using a language detector, the percentage of documents concerned by this type of errors. We have found less than 1% of documents containing inconsistencies. These documents have been ignored.

Similarity Method

The Lemur system has been used for indexing the whole training corpus. The documents used for building the index have not undergone any pre-processing. Language-specific stop-words lists were used

We have again used the entire document content, for both indexing and query: title, abstract (where available), description and claims.

Once the three indexes were built, one for each language, we have used the 2000 full test documents as queries. There was no pre-processing of the query documents. Based on our previous experiments, we chose to use the InQuery retrieval method available in Lemur that seemed to provide the best results.

The query results have subsequently been processed in order to calculate the candidate classes of the query document, based on the classes of the top most similar documents retrieved from the indexed document collection. The class calculation algorithm, built in the course of previous experiments, is based on the ranking order and on the similarity score obtained for the indexed documents retrieved.

Finally, we obtain a list of 20 candidate classes ranked by system confidence going from 10000 to 0.

Semantic and Statistic Method

Automatic patent classification is a supervised document classification task. We are required to classify documents in about 630 pre-defined classes. The number of classes, the possible coverage of these classes, their varied representativeness, are all factors disturbing the classification task.

The automatic classification using semantics and statistics is defined mainly by two steps:

- representation of documents in an indexing perspective;
- training a statistical classifier using semantic information;

The document representation is a language dependent task. The approach described below was therefore applied, step by step, to each of the three corpora. The techniques presented took into account the specificities of each of the three languages.

Document Representation - Construction of Semantic Models

The training corpus structured according to the IPC subclasses has been used for the construction of the class-based semantic models.

In a first step, we have chosen to build, for each training class, a representative semantic model. The class representativeness is ensured by (1) terms extracted from documents, artificially-built term patterns and term-related concepts and by (2) a strong semantic relationship between these terms and the class identifier.

Information selection strategies have been applied in order to find a good compromise between class representativeness and processing time.

Our information selection approach is based on a combination of document term extraction, class title term extraction, semantic relation verification, filtering methods for selecting discriminatory terms and polysemy-based filtering methods.

The semantic relation establishment and the polysemy-based filters exploited WordNet as a lexical resource.

Different strategies were instituted, for example for fighting errors related to polysemy: low confidence in the semantic relations constructed for terms with high polysemy, concept reinforcement, etc.

Extraction of Terms

Many classification algorithms are based on the postulate that a document is a sequence of words, a “bag of words”.

For our experiments, we chose to extract terms rather than n-grams: the term length is not defined beforehand. These choices are motivated by the risk of losing information with a “bag of words” approach or an n-gram type method. The terms are obtained after breaking the text down into occurrences, labelling them, lemmatizing them, delimiting nominal groups and observing stability in corpus.

Semantic Relation with the Class Title

Each document class is defined by an identifier. The class representativeness is ensured by the patterns and terms extracted from the corpus and the class title. The relation linking the terms and term patterns to a class is a guarantee of representativeness of these terms for the class.

The relations between the identifier, the terms and term patterns constitute the semantic network of the class. This semantic network is supposed to be representative of this class.

Document Annotation and Training

In a second step, the training documents were annotated with the terms provided in the semantic models and in tight relation with their position within the patent document. An IPC subclass is thus described by the sum of the documents composing it. The documents composing a class are, in turn, described by the terms contained in the semantic models and their semantic relations, terms that are linked to their position in documents. The feature values are therefore calculated according to the feature position within documents: different weights are calculated depending on whether the term appears in the title, key phrases, claims, description section or in all together.

In a third step, an SVM classifier was trained for each one of the three languages. The classifier test output represented, for each tested document, the list of all the learned classes, ranked by probability.

Combined Method

In order to consolidate and maybe improve the results obtained with the methods described above, we have built a combined approach based on the first 3 best candidates obtained with the previous methods. We therefore attempt to exploit the complementarity of the two previous methods in order to improve system performance.

From previous experiments, we know that the performance on the top 3 candidates, for each of the two methods described above, is good enough to allow a performance gain of several points if combined.

With the two methods combined, we have a total of 3 to 6 candidate classes (subclass level) for each test document. SVM-based classifiers are built on the fly. We use a “one vs one” approach for each one of the 3 to 6 candidate classes. We therefore need to build from 3 (for the test documents having 3 candidates) to 15 classifiers (for documents with 6 candidates) for each test document.

The final score is the sum of the probability values obtained for each binary classifier.

Results and Discussion

Our system was ranked second in terms of precision for the first candidate. The IRF computed several other measures, in particular precision from 5 to 50 candidates, recall and F1 score for 25 and 50 candidates. Our system ranked first for the precision from 5 candidates upwards and for the F score on 25 and 50 candidates. For the recall scores, our system was disadvantaged by the low number of candidates sent (max 20), whilst this low number of candidates favoured our system in terms of computed precision and F scores.

However, from an application point of view, the most meaningful measure is the precision on the first candidate and, decreasingly less so, the precision on the top 5-6 candidates. Indeed, it is unlikely that a patent classification system user, be it an office examiner or an individual interested in patent classification, would actually go through the whole list of 25 to 50 classes in search of the most appropriate class. It is the first 3-4-5 candidates that likely seem the most pertinent and useful for a system user.

Our statistics (Table 1) show that, for instance, for our Run3 method, we have 97% chances of finding at least one correct class in the first 5 candidates, which is useful information for anyone using the system for classifying a patent.

The problem becomes more complex from 5 candidates upwards. As discussed above, intellectual patent classification follows complex rules that are described in the classification tree and in the particular patent office documents used by patent officers. Sometimes a class is very difficult to be found by a system based on document term analysis, statistics and similarity. There are many cases when classification is based on legal knowledge, on the examiner’s interpretation of the patent text or it is related to deficiencies in the classification scheme and difficult interdisciplinary or vague patent applications. In this case, automatic systems based on probability and similarity can only give deceiving results when considering the first top candidates. Although recall is less meaningful to a patent system user, it is an interesting measure that allows for the evaluation of systems’ accuracy for finding all or most of the classes relevant to a patent document.

Table 1. Results for documents having at least 1 relevant class in the top n candidates

Nb candidates	Similarity Method (1)	Semantic and statistic Method	Combined Method (3)	Delta between (1) and (3)
First candidate	77,5%	75,15%	82,1%	+4,6
Two candidates	86,6%	86,7%	92,05%	+5,45
3 candidates	91,15%	91,05%	95,35%	+4,2
4 candidates	93,5%	93,65%	96,6%	+3,1
5 candidates	94,8%	94,7%	97%	+2,2
6 candidates	95,8 %	95,25 %	97,05 %	+1,25
10 candidates	97,05%	97,25%		
20 candidates	98,4%	98,55%		

Our system results are different according to the language of the patent documents tested. Table 2 shows a very good performance for English documents (84.7%) while the results for the German documents are not less than 10 points below. This could be explained by the specificities of each language and the particular processing and resources used for building the models. The number of training documents has an important impact. As the CLEF-IP corpus shows, the English documents clearly dominate the other two languages in terms of document numbers.

Table 2. Results per language for Combined Method

Nb candidates	EN	DE	FR
First candidate	84,7%	74,1%	78,0%
Two candidates	93,9%	87,0%	87,0%
3 candidates	96,4%	92,4%	92,6%
4 candidates	97,3%	94,9%	93,5%
5 candidates	97,5%	95,8%	94,3%
6 candidates	97,6%	95,8%	94,3%
Nb patents	1468	409	123
% Nb patents	73,4%	20,45%	6,15%

References

- C. J. Fall, A. Töröcsvári, K. Benzineb, G. Karetka (2003), Automated Categorization in the International Patent Classification, SIGIR Forum 37 (1)
- Chong Huang, Yonghong Tian, Zhi Zhou, Charles X. Ling, Tiejun Huang (2006), "Keyphrase Extraction Using Semantic Networks Structure Analysis," icdm, pp.275-284, Sixth IEEE International Conference on Data Mining (ICDM'06).
- Cornelis H.A. Koster, Jean G. Beney (2009), Phrase-Based Document Categorization Revisited. Proceedings PAIR'09 Workshop.
- Cornelis H.A. Koster, Marc Seutter and Jean G. Beney (2003), "Multi-Classification of Patent Applications with Winnow", Proceedings PSI 2003, Springer LNCS 2890, pp 545-554
- IPC Guide. Available at http://www.wipo.int/classifications/ipc/en/guide/guide_ipc_2009.pdf
- Jae-Ho Kim , Key-Sun Choi (2007), Patent document categorization based on semantic structural information, Information Processing and Management: an International Journal, v.43 n.5, p.1200-1215
- Jean G.Beney and Cornelis H.A. Koster (2003), "Classification supervisée de brevets: d'un jeu d'essai au cas réel", Proceedings of the XXIeme congrès Inforsid, pp.50-59
- M. Krier and F. Zaccà (2002), Automatic categorization applications at the European patent office, World Patent Information 24, pp187--196
- R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008), LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research 9, pp 1871-1874. Software available at <http://www.csie.ntu.edu.tw/~cjlin/liblinear>
- WIPO Categorization Survey, C. J. Fall, K. Benzineb (2002), Available at http://www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/Documentation/wipo-categorizationsurvey.pdf