Experiments with citation mining and key-term extraction for Prior Art Search

Patrice Lopez and Laurent Romary
INRIA - Humboldt Universität zu Berlin - Institut für Deutsche Sprache und Linguistik
patrice_lopez@hotmail.com laurent.romary@inria.fr

Abstract

This technical note presents the system built for the IP track of CLEF 2010 based on PATATRAS (PATent and Article Tracking, Retrieval and AnalysiS), the modular search infrastructure initially realized for CLEF IP 2009. We largely reused the system of the previous CLEF IP but at a relatively smaller scale and with the improvement of three main components:

- A new citation mining tool based on Conditional Random Fields (CRF).
- A key-term extraction module developed for technical and scientific documents and adapted to patent document structures using a vast ranges of metrics, features and a bagged decision tree.
- An improvement of our multi-domain terminological database called GRISP.

We used the Okapi BM25 and the Indri retrieval models for the prior art task and a KNN model for the automatic classification task under the IPC subclasses. In both tasks, specific final re-ranking techniques were used, including multiple regression models based on SVM. Although the Prior Art task was more challenging and we used a more limited number of retrieval models, we maintained similar results as last year. We performed, however, miserably at the classification task, and we consider that an instance-based KNN algorithm is not competitive with standard classifiers based on preliminary large scale training.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Patent, Prior Art Search, Citation mining, Key-term extraction, Regression models, Re-ranking, Automatic classification

1 From CLEF IP 2009 to CLEF IP 2010

Our main motivations for participating to CLEF IP are to advance in the comprehension of scientific and technical information and documents at large, to develop new solutions for managing

the data deluge and the information overload in science, and to facilitate the exploitation and dissemination of patent information. CLEF IP is one of the rare evaluation event that permits to tackle these problems.

We focused our efforts this year on two main aspects: the quality of **citation mining** from the patent documents and the extraction of **key-terms** in order to capture human-understandable descriptions of the main concepts of a patent. In addition, we further extended and consolidated our multilingual terminological database (GRISP, General Research Insight in Scientific and technical Publications) by integrating more knowledge sources and by driving the merging of concepts from the different sources with machine learning techniques. Regarding the overall architecture, we reused the framework developed for CLEF IP 2009, called PATATRAS (PATent and Article Tracking, Retrieval and AccesS), with a more limited number of indexes. This presentation describes mainly the novel aspects of our work compared to the system of last year. For a detailed description of the system, the reader is invited to consult our technical note of CLEF IP 2009 [Lopez and Romary, 2009].

In the following description, the collection refers to the data collection of approx. 2,6 millions documents corresponding to 1,3 million European Patents. This collection represents the prior art. The *training set* refers this year to the 200 documents of *training topics* provided with judgements (the relevant patents to be retrieved). The *prior art (PA) patent topic* refers to the 2000 patents for which the prior art search is done and the *classification (CL) patent topics* are the 2000 patents to classify.

1.1 Prior Art Searches

Following the first CLEF IP in 2009, the prior art task this year has been reviewed to coincide more closely with the actual prior art performed by patent examiners. The PA patent topics are normal unexamined applications (i.e. A1 or A2 publications) in only one language and without amendments of the description. The description of the granted patent publications often includes acknowledgement of the most important document of the prior art which has been identified during the search phase. The topic documents are thus more challenging than last year because they offer less multilingual information and less document citations.

A fully automated prior art search based on the existing search reports produced by the patent offices has inherent limitations in relation to patent families, to the influences of procedural aspects, the impact of limited search tools of the patent examiners, and the absence of non patent literature [Lopez and Romary, 2009]. We could however note two issues that could be addressed for a future edition of the evaluation forum:

- The problem of missing patent application content for some PCT applications arriving to the European phase: The European Patent Office does not re-publish patent applications coming from the PCT phase, and thus it is more difficult to retrieve these documents than for a patent examiner who typically searches the full application documents from the WO patent publications.

- The designation of the expected documents: The expected result this year were expressed as a list of patent publications (i.e. with a kind code) rather than simply a reference to a patent application. As a A publication is for instance always as relevant as the corresponding B publication (because the scope of B is always included in the one of the initial application document), the other publications for the same patent applications needed to be also considered as relevant. We view this way of building the expected results problematic because a patent with many publications will be repeated more often in the expected results as a patent with only one publication, and thus will have a stronger positive impact on the retrieving score. More generally this distinction between the publications appears artificial, because the final choice of citation of a particular publication by a patent examiner is very subjective.

We would like to thank the organizers for the progress toward a realistic prior art task which is remarkable and very beneficial for the participants. The developed systems could already be profitable to the actual search work of thousand of patent examiners and patent information specialists. This evaluation framework has also started to offer a sound basis for analyzing experimentally the impact of particular techniques on patent collections.

1.2 Automatic classification

CLEF IP this year introduced an automatic classification task. A set of 2 000 patent documents should be classified under one or several IPC subclasses (i.e. the four first characters of the IPC classification). The number of IPC sub-classes is approx. 600. This classification task corresponds to what is usually called the *pre-classification* [Krier and Zacc, 2002], where a patent application is routed to the appropriate a general level technical domain for being processed by the technically competent examiners. The *classification* is significantly more challenging as the complete IPC classification contains more than 60.000 subdivisions.

2 Advanced citation mining

2.1 The visible citation network

We observed last year a very strong impact of the interrelated cited patents on retrieval results. Citation relations between patents through time are manifestations of technological improvements and evolutions. These relations could be exploited for connecting a new patent application to a potentially relevant subset of the patent collection. The first kind of citations are the citations present in the search reports established by the patent examiners. This information are immediately exploitable because fully specified in the MAREC format (i.e. the XML format for the patent documents used in CLEF IP). Table 1 presents an overview of these citations available in the search reports.

Only the subset of the citations (EP) corresponds to documents present in the collection. It is possible from a citation to a non European patent to obtain the possible European version using patent family information. A patent family gathers all the different version of a patent application among the different geographical areas. The EPO proposes as web service (Open Patent Service, OPS) the access to the INPADOC database which permits to retrieve the possible European application of a given patent family given a non-European patent number. This service is however slow and limited by a fair use agreement. While it cannot be envisaged for a large number of patent references as present in the collection, we carried out a family look up for the patent topic set.

Source	Authority	#
Search Report	all	4 198 873
	EP	898 206
Description	all	6 257 511
	EP	890 754
Description + Family	EP	not processed

Table 1: Overview of citation relations in the patent collection.

2.2 Increasing the citation density

A scientific and technical work is often a contribution to previous existing works. Acknowledging and referring to previous realization and documents is therefore an inherent characteristic of any scientific and technical documents, including patent documents, which appears important to address. Following EPO's statistics, independently from the first kind of citation present in the

search report, the description body of patent application contains in average 9 citations from the initiative of the applicant, 7,5 references to other patents and 1,5 references to non patent literature. These citations correspond to the applicant's view of the state of the art and is a legal constraint (Rule 27(b) of the EPC, European Patent Convention). It is thus important for a patent examiner to evaluate these documents and possibly to cite some of these documents in the search report.

A patent document can contain several hundred of such references, while the number of citations in the search report is rarely more than ten. Extracting accurately these references can provide useful information for starting a search and understanding the key aspects of an application. The difficulty of this extraction task is a strong variability of contexts and patterns. Last year, we used a basic set of regular expressions for extracting patent citations in patent text bodies. The regular expressions were created based on a set of approx. 50 patterns of patent citations. Some analysis showed that we were missing at least 40% of the citations and that more advanced techniques were necessary.

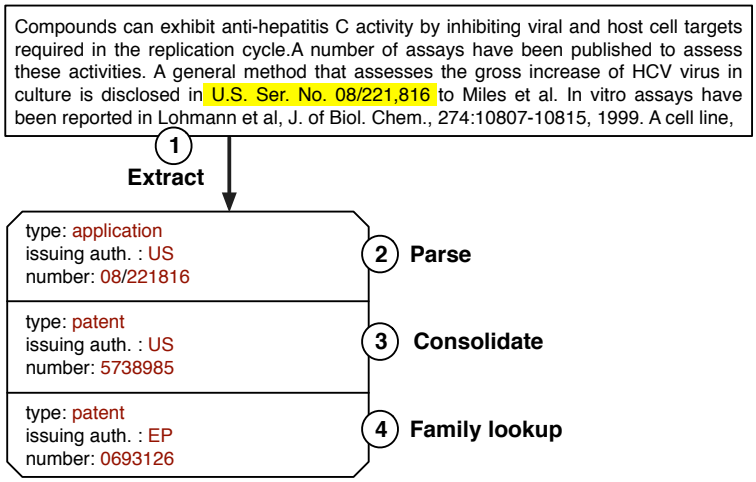

Figure 1: Example of patent reference extraction, parsing, consolidation and family lookup.

The new patent reference extraction module performs the following processing steps, as illustrated by Figure 1:

1. **Identification of reference strings:** The text body is first extracted from the patent document. The patent reference blocks are first indentified in the text body by a specific Linear-Chain CRF model.

2. **Parsing and normalization of the extracted reference strings:** The reference text is then parsed and normalized in order to obtain a set of bibliographical attributes. References to patent are parsed and normalized in one step by a Finite State Transducer (FST) which will identify (i) if the patent is referred to as a patent application or a patent publication, (ii) a country code, (iii) a number and (iv) a kind code.

3. **Consolidation with online bibliographical services:** Different online bibliograhical services are then accessed to validate and to enrich the identified reference. For patent references, we use OPS (Open Patent Service[1]), a web service provided by the EPO for accessing the Espacenet patent databases. This step permits for instance to retrieve the patent numbers from a reference to a patent application number.

[1] http://ops.espacenet.com

4. **Family lookup:** For the citations extracted from the patent topics, in case the citation is a non-European patent, we access OPS for patent family information and try to identify the corresponding European patent.

The CRF model has been trained based on 200 patent documents corresponding to approximatively 2 000 patent citations. In [Lopez, 2010], we evaluated the f-score of the extraction of patent reference blocks at 0.9540 based on a manually annotated corpus of patents from different sources, while the previous state of the art was around 0.75. In 97.2%, we were then able to parse correctly the citation block and identify the correct patent attributes.

The tool is also able to extract non-patent literature references with a specific CRF model, to parse the extracted reference for identifying a set of 12 bibliographical attributes (author, title, journal, date, etc.) and to consolidate the result with an access to Crossref. Although potentially very relevant to the Prior Art task, in particular in certain technical domains such as bio-technologies, computer and chemistry, this functionality has, however, not been used in the present work because of time and processing power constraints.

The result of these extraction are presented on Table 1 for the collection and on Table 2 for the set of topic patens.

Source	Authority	#
Search Report	all	0
	EP	0
Description	all	18 876
	EP	2 946
Description + Family	EP	7 706

Table 2: Overview of citation relations in the set of topic patents.

3 Key-term extraction of patent documents

3.1 Approach

Key terms (or keyphrases or keywords) provide general information about the content of a document. Key-terms constitute good topic descriptions of documents which can be used in particular for information retrieval, automatic document clustering and classification. Among the extracted terms for a given scientific document in a given collection, which key terms best characterize this document?

Our work is based on the system realized for Semeval 2010, task 5 *Automatic Keyphrase Extraction from Scientific Articles* [Lopez and Romary, 2010b]. Candidate phrases up to 5-grams are extracted from the textual content of the document. Phrases beginning or ending by a stopword are discarded. The ability of a candidate phrase to be considered as a key-term is estimated in a supervised manner by a bagged decision tree based on the key-terms selected by the authors and the readers of the training documents. The advantage of using examples annotated by the authors and the readers for selecting the key-terms is that the resulting extracted topic description will still be comprehensible for a human. The machine learning algorithm use three set of features:

- a first set of *structural features* characterizing the position of a term with respect to the document structure for each candidate: present in the *title*, in the *abstract*, in the *introduction*, in at least one *section titles*, in the *conclusion*, etc. the relative position of the candidate phrase in the document is also used,

- a second set of *content features* which tries to captures distributional properties of a term relatively to the overall textual content of the document where the term appears or the collection. For this we use a set of metrics: Generalized Dice Coeficient (GDC) as introduced

by [Park et al., 2002], TF-IDF and the frequency of the candidate phrase to be selected as key-term in the global corpus.

- finally, a set of Lexical/Semantic features which are produced exploiting our multilingual terminological database GRISP and Wikipedia were introduced.

We further applied a post-ranking based on the statistics observed on HAL[2] research archive. HAL contains approx. 139,000 full texts articles described by a rich set of metadata, often including author's keywords. In Semeval 2010, we achieved a f-score of 27.5 for top the 15 key-terms. This level of performance must be considered knowing that the expected key-terms used for the evaluation were a relatively small and subjective selection by the authors and the readers.

The features have been adapted from this initial implementation for scientific articles to patent publications. The structure features were changed by using the available structural tag of the MAREC XML format. The TF-IDF were computed on the whole patent collection. Finally a set of 120 patent documents with annotated keywords have been used to retrain the bagged decision tree.

3.2 Extraction results

Table 3 give an example of the key-term extraction for the patent publication EP0381288A1. The score associated to a phrase evaluates to which extend the phrase can be viewed as a key-term for the document. We can see that this extraction corresponds at the same time to a topic modeling of the document, to a human-understandable summary of the key content of the document close to usual keywords attributed to a scientific and technical article, but also can be viewed as synthetic queries for which the documents itself is relevant.

word index	0.298805	structural ambiguity	0.0611575
syntactic interpretations	0.226326	word-specific ambiguity	0.0611575
governor	0.21987	interpretation index	0.060762
parsing process	0.193613	parsing analyses	0.0565435
word sequence	0.192174	representing the rank	0.0550731
choice point	0.173721	multiple syntactic interpretations	0.0525496
intermediate nodes	0.1588	natural language word	0.0520414
tree structures	0.152594	syntactic relation	0.0501529
multiple analyses	0.152235	tree	0.0494087
index	0.149352	interpretation	0.0492798
syntactic network	0.147376	word in the sequence	0.0492702
maximum phrases	0.143614	phrases	0.0488179
dependency grammar	0.142967	dependency	0.0470071
top node	0.131585	parsing algorithm	0.0467525
consistency check	0.129689	words of the sequence	0.0460244
word	0.126933	analyses	0.0428058
language word	0.125774	sentence	0.0417408
grammar	0.122467	programming language	0.0415976
parser	0.11987	natural language	0.0403096
syntactic	0.0943424	multiple syntactic	0.0394109
parsing	0.0787452	dependent	0.0384868
pointer node	0.0656035	choice	0.038086
unambiguously coding	0.0645593	ambiguity	0.0371675

Table 3: Example of key-term extraction for document EP0381288A1.

This processing scaled well the whole collection of documents since the extraction took on a low-end hardware in average 0.7 second per patent application.

[2]HAL (Hyper Article en Ligne) is the French Institutional repository for research publications: http://hal.archives-ouvertes.fr/index.php?langue=en

4 Extension of GRISP

GRISP (**G**eneral **R**esearch **I**nsight in **S**cientific and technical **P**ublications) is a multilingual terminological database based on the principles of ISO 16642 (TMF – Terminological Markup Framework) [Romary, 2001], a generic onomasiological (concept to word) model. This conceptual framework facilitates the combination of heterogeneous specialist resources and in different languages. [Lopez and Romary, 2010a] presents the overall framework, the different technical and scientific resources which have been combined and the usage of a machine learning approach for deciding when to merge two concepts coming from different resources in a single, enriched concept.

As compared to GRISP used in 2009, **ChEBI**[3] has been integrated. ChEBI is a freely available dictionary of molecular entities developed at the European Bioinformatics Institute [Degtyarenko and al., 2008]. ChEBI is a valuable source of chemical vocabulary with approx. 42.000 concepts, 97.000 terms, 28.000 semantic relations and multilingual terms in 5 languages. In addition, we update the partial Wikipedia resources with the latest 2010 XML dumps.

5 Overall Description of the Prior Art System

5.1 System architecture

Figure 2 gives an overview of the realized system. The system is close to the system realized for CLEF IP 2009, but with a limited number number of retrieval models and a redefined phrase retrieval model based on the extracted key-terms resulting from the preprocessing of the whole collection and the topic patents.

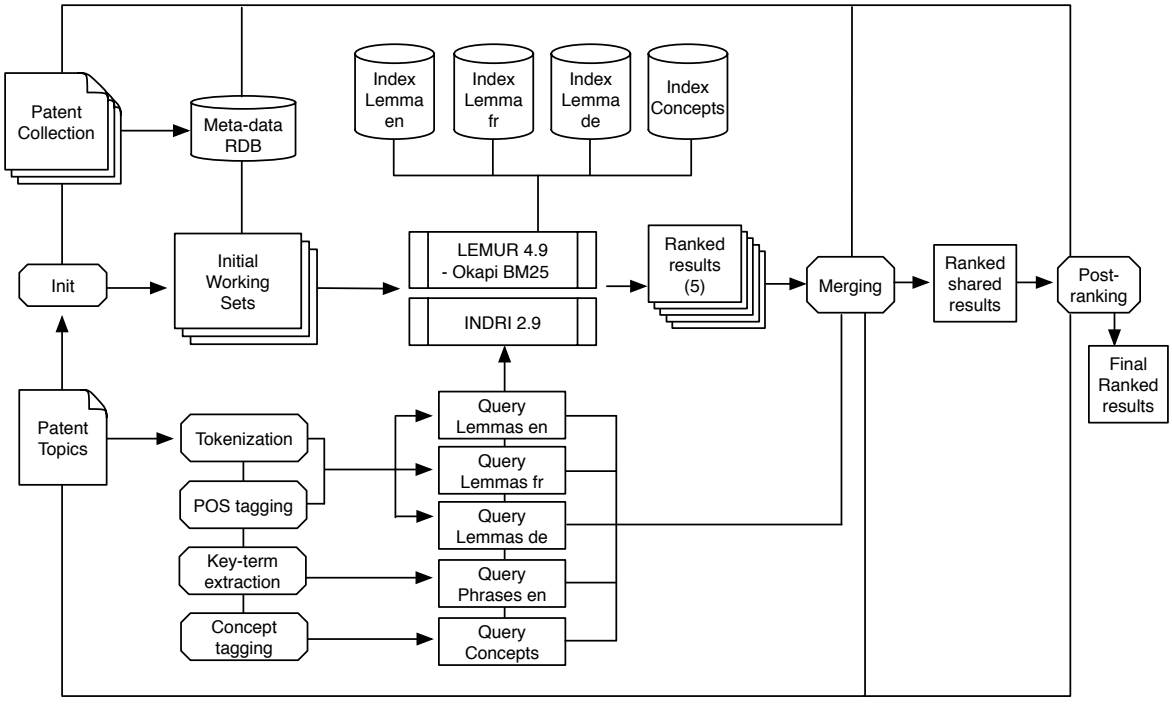

Figure 2: System architecture overview of PATATRAS millésime "CLEF IP 2010" for query processing. The arrow represents the main data flow from the patent topic to the final set of ranked results.

[3] http://www.ebi.ac.uk/chebi

5.2 Document preprocessing

The document preprocessing is similar as the previous year with two differences: the addition of the new citation mining processing and the extraction of key-terms as explained in section 2 and 3, and no systematic extraction of all phrases. The preprocessing result in particular in a database storing all metadata of the collection, including the new extracted citations and the key-terms. A few metadata fields were normalized: inventor and applicant names, similarly as last year, and a particular effort was made this year on cleaning and normalization of IPC and ECLA classes.

The concept tagging based on the controlled terminology of GRISP is similar as last year. The concept disambiguation was still realized on the basis of the ECLA classes (or by default the IPC classes) of the processed patent.

5.3 Indexes

The four following indexes were build using the Lemur toolkit [lem, 2001-2010] (version 4.9):

- For each of the three language (English, French, German), we built a full index at the lemma level.
- A crosslingual concept index was built using the list of concepts identified in the textual material for all three languages.

Similarly as last year, we do not index the collection document by document, but considered a "meta-document" corresponding to all the publications related to a patent application.

5.4 Retrieval models

We used the two following well known retrieval models:

- Okapi weighting function BM25 ($K1 = 1.5, b = 1.5, K3 = 3$).
- Indri

Although KL-Divergence with Jelinek-Mercer smoothing ($\lambda = 0.4$) was the best performing retrieval model last year, it is also the most time and resource consuming retrieval algorithm. As our development timeframe was this year relatively limited, we did not submit runs including the result of this retrieval model.

The two models have been used with each of the previous four indexes, resulting in the production of 5 lists of retrieval results for each topic patent. Similarly as last year, the queries for lemma and concept representations were build based on **all** the available textual data of a topic patent.

There are many possibilities for exploiting a topic representation based on key-term extraction. For instance, in the context of language model information retrieval, [Zhou et al., 2007] uses a set of extracted keyphrases for building a topic signature language model used for a semantic smoothing method. We applied in this work a much simpler approach which can be viewed as a baseline. We used the Indri retrieval model applied to the English lemma index and built queries mixing phrases and single word terms. Due to the limit of the numbers of phrases in a query which could be processed in a reasonable time, we limit the number of multi-word key-term to a constant N, and then add the rest of phrases as individual words. For instance, for a list of n key-terms $(t_p, s_p)_p$ where s_p is the score associated to the term t_p, having a term formed by multiple words w, $t_p = (w_{pi})_i$, we build the query as follow:

$$\#\text{weight}(s_0\ \#1(t_0)\ s_1\ \#1(t_1)\ ...\ s_N\ \#1(t_N)\ ...\ s_p\ w_{pi}\ ...\ s_n\ w_{ni})$$

In our work, we limited the number N of phrases present in the Indri query to 4. Following this construction, an Indri query takes approximatively 15 second to be processed.

The baseline results of the different indexes and retrieval models are presented in Table 4, column (1). Given that this year the patent topic contains text content in only one language (the main language of application), the results presented in this table are restricted to the set of topics having text in this language, i.e. only 134 queries for French, 519 for German and 1 959 for English over the total of 2 000 patent topics. This restriction explains the high MAP results for French and German indexes.

Model	Index	Language	(1)	(2)
BM25	lemma	en	0.0842	0.1628
BM25	lemma	fr	0.124	0.2185
BM25	lemma	de	0.1081	0.1869
Indri	phrase	en	0.0758	0.1597
BM25	concept	all	0.0655	0.1529
KL	lemma	en	0.0911	0.171
KL	lemma	fr	0.1309	0.2244
KL	lemma	de	0.1085	0.1887

Table 4: MAP results of the retrieval models, Prior Art task. (1) Base MAP, Normal set (2 000 queries) (2) Map with initial working sets, normal set (2 000 queries). The results for a language dependent index are produced only for the patent topics having text in this language. The results for KL were not part of the submitted runs, they have been produced while preparing this technical note and are reported here for information.

The initial working sets have been created via an iterative process similarly as last year, exploiting cited documents and the whole range of available metadata. The process could take benefit this time from a larger number of citations extracted from the description to seed the sets. Using these working sets reduce the search space while containing approx. 75% of the expected documents. As one can see on Table 4, column (2), the initial working sets provide a significant improvements in term of retrieval precision which is superior to the one observed last year. The working sets remain, however, slow to build, are based on manual and intuitive rules and appear difficult to improve in term of recall. We plan to replace the current algorithm by a machine learning approach which could drive the process of selecting interesting patent documents in a monotonic process rather than iteratively.

5.5 Merging of results

The merging of the five result sets was realized as last year with a SVM model using a set of 4 631 training patents. We did not exploit the additional topic set of last year (10 000) and did not rebuild a specific model this year due to lack of time. As a result, the combination was not as effective as last year, but has still provided an improvement over the individual result sets.

5.6 Post-ranking and final results

We re-use the same final re-ranking model as build for CLEF IP 2009. This re-ranking permits in particular to boost the score of the patents initially cited in the description of the topic patent and the ECLA classes, resulting in a significant improvement. The regression model was trained using the set of 4 631 training patents which were compiled for CLEF IP 2009.

Measures	small	large
MAP	0.2731	0.2645
Prec. at 5	0.4244	0.4209
Prec. at 10	0.3625	0.3482

Table 5: Evaluation of official runs for the small (400 topic patents) and large (2 000 topic patents) topic sets.

The final results are presented on Table 5, and shows comparable accuracy as last year. Given that the prior art task of this year was more challenging as the topic patents were real application documents, and given that we reduced the number of retrieval model and not updated our regression models for result merging and re-ranking, this result shows the positive impact of a high quality extraction of applicant's citations in the patent descriptions and the potential of key-term extraction.

6 Automatic Classification task

As we started to prepare the classification task very late, we could not experiment any algorithms requiring a training on the document collection. We thus opted for an instance-based approach, and more particularly for a KNN algorithm, simply re-using the existing system build for the prior art task. We use the existing prior art search system for providing a list of ranked results for a given patent topic to be classified and the KNN implementation of WEKA [Witten and Frank, 2005], with $N = 25$. Such algorithm could be developed and produced in just a few hours.

Run	Metric	Score
patatras	MAP	0.5083
	Prec. at 1	0.56
	Prec. at 5	0.252
ssft_CEC0_run7	MAP	0.7951
	Prec. at 1	0.835
	Prec. at 5	0.3662

Table 6: Evaluation of official runs for the classification task with the best system (Simple Shift).

Unfortunately, our system suffered from several implementation errors which make the interpretation of the results difficult. The final results are presented in Table 6 with a comparison with the best run. The difference between the two systems is very important. Even by correcting implementation errors, we consider that an instance-based KNN algorithm is not competitive with state of the art classifiers based on preliminary large scale training, and *a fortiori* with the advanced system realized by Simple Shift.

7 Future Work

We plan to focus our future efforts on the automatic recognition and the exploitation of the structures of patent documents. The main goal is to improve the formulation of the queries and to build more specialized indexing processes. The recognition of entities of special interest such as non patent references and numerical values is a second axis of future work which appears promising in certain technical domains such as biotechnology, chemistry and computer sciences.

References

[lem, 2001-2010] 2001-2010. *The Lemur Project*. University of Massachusetts and Carnegie Mellon University.

[Degtyarenko and al., 2008] Degtyarenko, K. and al., 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36:344–350.

[Krier and Zacc, 2002] Krier, M. and F. Zacc, 2002. Automatic categorisation applications at the european patent office. *World patent Information*, 24:187–196.

[Lopez, 2010] Lopez, P., 2010. Automatic Extraction and Resolution of Bibliographical References in Patent Documents. In H. Cunningham, A. Hanbury, and S. Rüger (ed.), *First Information Retrieval Facility Conference (IRFC)*. Vienna, Austria: Springer, Heidelberg.

[Lopez and Romary, 2009] Lopez, P. and L. Romary, 2009. Multiple retrieval models and regression models for prior art search. In *CLEF 2009 Workshop, Technical Notes*. Corfu, Greece. http://hal.archives-ouvertes.fr/hal-00411835.

[Lopez and Romary, 2010a] Lopez, Patrice and Laurent Romary, 2010a. GRISP: A Massive Multilingual Terminological Database for Scientic and Technical Domains. In *Seventh international conference on Language Resources and Evaluation (LREC) 2010*. La Valette, Malte. Available at http://hal.inria.fr/inria-00490312.

[Lopez and Romary, 2010b] Lopez, Patrice and Laurent Romary, 2010b. HUMB: Automatic Key Term Extraction from Scientic Articles in GROBID. In *SemEval 2010 Workshop*. Uppsala, Suède. Available at http://hal.archives-ouvertes.fr/inria-00493437.

[Park et al., 2002] Park, Y., R.J. Byrd, and B.K. Boguraev, 2002. Automatic glossary extraction: beyond terminology identification. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics.

[Romary, 2001] Romary, L., 2001. An abstract model for the representation of multilingual terminological data: Tmf - terminological markup framework. In *TAMA (Terminology in Advanced Microcomputer Applications)*. Antwerp, Belgium. Available at http://hal.inria.fr/inria-00100405.

[Witten and Frank, 2005] Witten, I.H. and E. Frank, 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2nd edition.

[Zhou et al., 2007] Zhou, X., X. Hu, and X. Zhang, 2007. Topic signature language models for ad hoc retrieval. *IEEE Transactions on Knowledge and Data Engineering*:1276–1287.