Hybrid Citation Extraction from Patents

Olivier Galibert[*,1], **Sophie Rosset**[*], **Xavier Tannier**[*,†], **Fanny Grandry**[*]

[*]LIMSI-CNRS
B.P. 133 - F-91403 ORSAY Cedex
FRANCE
firstname.surname@limsi.fr

[†]University Paris-Sud 11
91405 Orsay Cedex
FRANCE

Abstract

The Quaero project organized a set of evaluations of Named Entity recognition systems in 2009, including reference extraction in patent text. The LIMSI participated in this evaluation. The task and its metrics is presented, followed by a complete system description and the evaluation results. The system obtained a (tied) first place in the evaluation.

1. Introduction

Quaero[2] is a program promoting research and industrial innovation on technologies for automatic analysis and classification of multimedia and multilingual documents.
Among the many research areas concerned by *Quaero*, a yearly evaluation campaign on named entity recognition (NER) is organized (Galibert et al., 2010). A sub-task of this campaign consists in extracting citations from patents, *i.e.* references to other documents, either other patents or general literature.
We present in this paper the results obtained by LIMSI's system for NER 2009 evaluation concerning citation extraction from patents.

2. Task and corpus

The objective of this sub-task is to detect citations of other documents in English-language patents. Additionally, these references must be classified between references to patents (*patcit*) and references to other literature (*nplcit*, non-patent literature citation). The training corpus contained 15185 annotated patents (215K nplcits and 198K patcits), the development corpus 1000 patents (14,3K nplcits and 8,5K patcits) and the test data 760 patents.
Patent citations refer to other patents. They often include a country code and a patent number, but sometimes take more complex forms:

(1) <**patcit**> *US 5828 849* <**/patcit**> *describes a method to derive edge extensions for wavelet transforms and inverse wavelet transforms in which a received input signal is filtered using wavelet filtering.*

(2) *Highly useful quasi-prepolymers are disclosed in* <**patcit**> *U.S. Patent No. 4,791,148* <**/patcit**> *and* <**patcit**> *U.S. application Serial No. 07/342,508, filed April 24, 1989* <**/patcit**>.

Non-patent citations refer essentially to articles, but can also concern other external entities such as databases (even if that has proven extremely rare):

(3) *Moreover, the calibration does not necessitate further standardization which is conversely required, for example, by the TaqMan technique (see* <**nplcit**> *Chatellard P. et al., J. Virol. Methods, 71:137-146, 1998* <**/nplcit**>*).*

As we can see, patent and non-patent citations have a very different nature. Patent citations tend to be short and strongly internally structured with little or no context. Non-patent citations tend to be longer, less strictly structured, and with more emphasis on contexts words such as *see*. Because of that, we decided to handle these two citation types with two independent, very different systems and then merge their results.
Some related works concern citation matching, *i.e.* systems normalizing different ways of expressing a reference to bibliography in scientific papers (Lawrence et al., 1999), or metadata extraction from already tokenized citations (authors, dates, etc.) (Councill et al., 2006; Cortez et al., 2007; Daya et al., 2007).
(Grover et al., 2000) propose a tokenization tool including named entity recognition with a specific focus on rule-based citation and references extraction in general scientific literature. Patents are not concerned.
Sections 3.1. and 3.2. respectively detail the techniques used for extracting patent and non-patent citations. Section 4. presents the official Quaero results as well as some additional experiments.

3. System description

3.1. Patent citations

We call *patent citations*, or *patcits*, references to other patents in the text. Automatic extraction of very regular and structured units of texts generally benefits from hand-created expert knowledge on the form of regular expressions. These patent citations cleary belong to that category.

[1]This author is now affiliated at LNE - Laboratoire National de Métrologie et d'Essais, 78197 Trappes Cedex, France.

[2]`http://www.quaero.org`

We defined some expressions representing structures and, when appropriate, contexts.
Patcits always contain a numeric or alphanumeric identifier that can take different forms: "*2003-294990*", "*W003/096095*", "*6,504,053*", "*11/059,282*", "*EP-A-0262894*", etc. These types of identifiers are important evidence, but are also quite ambiguous.
For example, in the following sentence, "AB024035" is not a *patcit* identifier, although it corresponds to a *patcit* identifier regular expression:

(4) *Location in database (accession number, location on bac) : AB024035, Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone: MHM17, complete sequence.*

Identifier candidates must then come together with strong clues or appropriate contexts, and must not occur within a "negative context".

- **Strong clues** are for instance:
  - country codes ("*EN*", "*FR*", "*DE*"...)
  - explicit triggers ("*pat.*", "*patent*", "*application*", "*no.*")
- **Appropriate contexts** can be specific expressions as "*herein by reference*", or dates, author names or another *patcit* on left context.
- **Negative context** is an expression introducing a non-patent citation. Such an expression will discard the candidate:
  - page numbers
  - a reference to a journal name or reference ("IEEE", "ISSN", ...)

Author name recognition only consists in extracting pairs or triplets of capitalized words, separated by "*and*" or ending by "*et al.*". Since we already suppose that we are in a citation block, these simple patterns are enough.
Boundaries of *patcits* are decided by collecting all internal clues (authors, identifiers, country codes, triggers) and extending the element to its maximum size. For example:

(5) *The present application is a Continuation in Part of <patcit> Application Serial No. 11/059,282 , filed February 16 , 2005 </patcit>, the entire contents of which are herein incorporated by reference .*

When several blocks are contiguous (several identifiers lead to several *patcits*), the following order of elements has been identified as being more frequent in texts: country code or name, explicit triggers, identifier. This order allows to separate distinct citations:

(6) *Heretofore , solutions to the above problems and demands have been proposed in <patcit> Japanese Laid-Open Patent Publication No. 2003-294990 </patcit>, <patcit> Japanese Laid-Open Patent Publication No. 2003-294992 </patcit>, <patcit> Japanese Laid-Open Patent Publication No. 2003-295000 </patcit>, <patcit> WO03/096095 </patcit>, <patcit> WO03/060584 </patcit>, and <patcit> WO03/098293 </patcit>.*

On the development data, we ended up with a correctness level of 77.6%. The errors consisted of 12.2% of insertions, 22.5% of deletions and the rest, 65.2%, of frontier errors. A large part of the frontier errors look like reference errors, but an interesting part of the remaining ones seem to be errors on whether to add additional information, often in parenthesis, to the citation. A statistical approach could probably help in that area.

3.2. Non-patent citations

The system for detecting non-patent citations, on the other hand, is purely stochastic. Given that the structure of these references is not so strict as for the references to patents, the number of different contexts is very high, we consider that a stochastic approach will better generalize such various informations.
The system first uses the TreeTagger (Helmut, 1994) to annotate the words of the patents with Part of Speech information, and then a CRF-based chunker (Lafferty et al., 2001) uses them to detect the citations. The open source toolkit *CRF++* (Kudoh, 2007) was used. The boundary of the citations in the training data is indicated using the standard *BIO* method. That means that all words are labelled with one of the following labels: B if the word begins a non patent citation, I if the word is inside a non patent citation, and O if the word does not belong to a non patent citation.
The feature generation rules used were the ones proposed by that toolkit for the CONLL chunking shared task. That means, we used as context the 2 precedings words, the current word and the 2 following words with their POS. The preceding class is also used as a clue (the one given by the model during the test or the one given by the manual annotation in the training data for the training phase).
The model was trained on about 5% of the available training data (10 million entries) because of memory usage issues. The labeling error rate on the 5% of the training data was of 0.006%, and 0.3% on the whole training data.
Turning these annotations into citations, we ended up with 56.2% correct citations on the development data. Among the errors, 7.8% were insertions, 13.5% deletions, leaving a whopping 78.6% of bad frontier errors. Studying these frontier errors shows that approximatively half of them are open to discussion (parenthesis and period inclusion issues in particular). More interestingly, a large number of errors come from missing elements like page numbers on the right side or part of the authors names on the left side. Adding regular expression-based structural clues to the feature set could help reducing errors of that kind.
In order to measure how much data we really need for such a task, we trained different models for the *nplcit* detection using different sizes of training data. The Figure 1 shows the main results of our various experiments. The results show an important decrease of the labelling error rate using 1% up to 15% of the available training data. Over 15% the decrease slows down. These results told us that the system

could benefit of using more training data but not that much. Further experiments are ongoing.

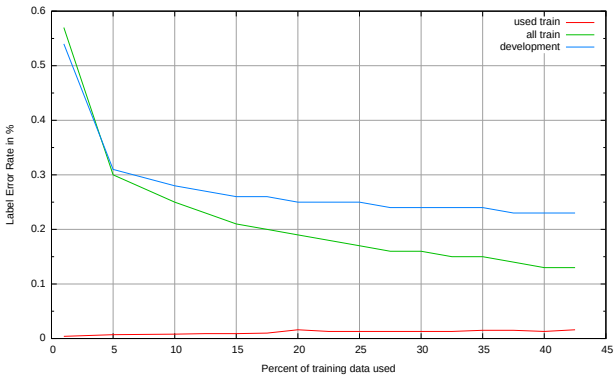

Figure 1: Non-patent citation detection: labelling error rate *vs.* training corpus size

3.3. Final system

Merging these two outputs was relatively simple since collisions were extremely rare (around 0.3% on the development data). In these rare cases the *patent* subsystem was systematically considered correct over the *non-patent* one.

4. Results

The different metrics used for this evaluation were precision, recall, F-measure and slot error rate (SER). 4 teams participated to this task; our system obtained the best results of this evaluation campaign with a SER of 33.1%, a precision of 78.5%, a recall of 71.2% and a F-measure of 0.75. Second best is quite close (precision: 78.1%, recall: 69.4%) and the difference between the first two systems is poorly significant.
Looking specifically on the *patent* or *non-patent* citations detections, show a real and big difference (F-measure of 0.63 for *patent* and 0.47 for *non-patent* considering the strict metric, *i.e.* frontier error cost 1 point).
The Table 1 shows the complete and detailed results for this task.

# Tags	Rate (%)	Types
12231	50.1%	Correct
898	3.7%	Insert
1921	7.9%	Delete
2	0.0%	Bad type
10276	42.1%	Bad frontier
117	0.5%	Bad frontier and type
13214	54.2%	Total errors

Table 1: Detailed results for citations detection in patents

The Table 2 shows general scores and specific scores given the tag to be detected (*patcit* or *nplcit*).

5. Annotations error analysis

The patent corpus was initially annotated for human use, specifically for prior art review. As a result there were no precise annotation guides defined and no real attempt at consistency. Only the fact that the citations were annotated for later retrieval was important. As a result a large number of inconsistencies are present in the corpus, test included.
A first class of errors appears in the handling of punctuations. For instance the two following excerpts are issued from the same document:

Metrics	General	Patcit	Nplcit
Precision	55.3%	64.4%	49.9%
Recall	50.1%	61.0%	44.1%
F-measure	52.6%	62.6%	46.8%
SER	54.2%	-	-

Table 2: Results for all the tags or specific tags using strict scoring

(7) *The compound of the formula shown in Figure 8 has the following sequence of transitions [<**nplcit**>K.Praefcke, B.Kohne, B.Gündogan, D.Singer, D.Demus, S.Diele, G.Pelzl and U.Bakowsky, Mol. Cryst. Liq. Cryst., 198, 393-405 (1991) <**/nplcit**>]*

(8) *The compounds of the formulas shown in Figure 10 have the following sequence of transitions [<**nplcit**>T. J. Phillips, J.C.Jones and D.C.McDonnell, Liquid Crystals, 15, 203-215. (1993 <**/nplcit**>)]*

As can be seen, the closing parenthesis for the year is, in one case, considered inside the citation and, in the other case, outside. The same kind of variability can be observed with the period in *et al.* expressions.
Some cases show even more of this *just pointing to the citation is enough* effect. For instance in:

(9) *As to special frequency ratios, eg. as described in <**nplcit**> the publication IEEE Trans. Commun., Vol. 44, no. 6, pp. 742-748, June 1996 K. Murakami <**/nplcit**> 'Jitter in Synchronous Residual Time Stamp', a low frequency 'jitter' may be determined, i.e. a distortion due to synchronization errors, which is difficult to filter from the phase locked loop due to its low frequency.*

Not only the presence of *the publication* in the citation is debatable, but the publication title is left out, probably because of its unusual position in the end. The citation being annotated that would be no problem for the human patent reviewer, but it counts as an error for a system which annotates the whole citation.
These inconsistencies are not limited to non-patent literature citations. For instance a clear guideline is missing about whether the author names should be included in a patent citation:

(10) *Another technique for electroplating solder bumps is described in <nplcit> Patent Abstracts of Japan, vol. 017, no. 568 (E-1447), 14 October 1993 </nplcit> and unexamined <**patcit**> Japanese*

patent application No. 05-166815 </**patcit**> *by Matsumura, entitled Plating Bump Formation Method and Wafer Plating Jigs.*

(11) *The use of contact pins to withdraw lateral current from a UBM layer is more fully described in* <**patcit**> *U.S. Patent No. 5,342,495 to Tung et al.* </**patcit**> *entitled Structure for Holding Integrated Circuit Dies to be Electroplated.*

We can see that in two very similar cases the authors are included in one and excluded from the other. Possibly more annoying is a similar case with the patent prefixes:

(12) *In the expansion zone, the foam-forming mixture is allowed to expand freely without constraint by a leveling member such as the upper conveyor disclosed in U. S. Pat. No.* <**patcit**> *4,572,865* </**patcit**> .

(13) <**patcit**> *US Patent No. 5,329,585* </**patcit**> *discloses a subscriber line interface circuit for controlling ac and dc output impedance, in which two separate impedances must be set to germinate the ac and de feedback signals.*

Finally, a number of interesting cases show that automatic annotations have been used at one point, sometimes badly. In particular in the two following cases the *XP* reference is, as far as we can tell, an internal reference number within the document and not a patent reference:

(14) *The following document,* <**nplcit**> *Yushi Uno et al: "Complexity of the Optimum Join Order Problem in Relational Databases" IEICE Transactions, JP, Institute of Electronics Information and Comm. Eng. Tokyo, vol.E74, no.7, 1 July 1991 (1991-07-01), pages 2067-2075* </**nplcit**>, <**patcit**> *XP000263060* </**patcit**> *ISSN: 0917-1673, identifies the problem of finding the optimal join order (par. 2.6) including the representation of a query graph in mathematical terms as well as the computation of the cardinality of the multiple join query (p.2069, par. 2.5 and 2.6).*

(15) <**nplcit**> *SCHILIT B N ET AL: "TeleWeb: Loosely connected access to the World Wide Web" COMPUTER NETWORKS AND ISDN SYSTEMS, vol. 28, no. 11, May 1996 (1996-05), page 1431-1444* </**nplcit**>, <**patcit**> *XP004018240* </**patcit**> *describes a system in which costs are made visible to the user through annotated HTML; budget monitoring warns the user when operations exceed prespecified limits; actions may be postponed and later triggered when conditions are met; and user customisation and system configuration values may automatically adapt according to the changing conditions of use.*

The first case is particularly interesting in that the ISSN is lost in the process.

All these examples do not imply that the corpus is worthless, far from it. For a start, a more informative, application oriented result, could probably be obtained by ignoring frontier errors entirely. After all, the aim is to point to citations first, precise segmentation is obviously secondary. But also a large number of the encountered issues seems fixable automatically or semi-automatically, by a combination of rules and system error reports. The large number of patents annotated in that way (around 17,000) makes the effort probably worthwhile.

6. Conclusion and perspectives

These results confirm our first observations: *patent* citations are more structured and thus easier to detect than *non-patent* citations. The *patent* citations benefit from regular expressions even if their detection, specifically their frontiers, could be improved.

On the other side, the *non-patent* citations seem to be much more loosely structured; even a statistical approach does not lead to very good results, in particular where it comes to frontier detection. It seems possible that using more data for training could be useful, but probably not by a very large amount.

Observing the corpus, another problem appears: in a lot of cases, the frontiers are not well annotated in the reference. This corpus has not been created for such a machine learning project but for human use by prior art researchers. Obviously, a human reading an annotated document is perfectly able to deal with some annotation incoherences. Using that corpus as-is for machine learning approach shows its limits, in particular in the evaluation stage. Further progress will require cleaning it up, probably requiring a mix of correction rules and the use of statistical systems in a keep-one-out approach to point to problems.

In any case, this study has shown the interest of merging subsystems with very different approaches into a whole. An obvious next step will be to further hybridize the subsystems themselves, leveraging the benefits of both rule-based and statistical approaches.

7. Acknowledgement

This work has been partially financed by OSEO under the Quaero program.

8. References

Eli Cortez, Altigran S. da Silva, Marcos André Gonçalves, Filipe Mesquita, and Edleno S. de Moura. 2007. FLUX-CIM: flexible unsupervised extraction of citation metadata. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 215–224, New York, NY, USA. ACM Press, New York City, NY, USA.

Isaac G. Councill, Huajing Li, Ziming Zhuang, Sandip Debnath, Levent Bolelli, Wang Chien Lee, Anand Sivasubramaniam, and C. Lee Giles. 2006. Learning metadata from the evidence in an on-line citation matching scheme. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 276–285, New York, NY, USA. ACM Press, New York City, NY, USA.

Min-Yuh Daya, Richard Tzong-Han Tsaia, Cheng-Lung Sunga, Chiu-Chen Hsieha, Cheng-Wei Leea, Shih-Hung Wuc, Kun-Pin Wua, Chorng-Shyong Ongb, and Wen-Lian Hsu. 2007. Reference metadata extraction using

a hierarchical knowledge representation framework. *Decision Support Systems*, 43(1):152–167, FEB.

Olivier Galibert, Ludovic Quintard, Sophie Rosset, Pierre Zweigenbaum, Claire Nédellec, Sophie Aubin, Laurent Gillard, Jean-Pierre Raysz, Delphine Pois, Xavier Tannier, Louise Deléger, and Dominique Laurent. 2010. Named and specific entity detection in varied data: The Quæro Named Entity baseline evaluation. In *LREC'10*.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT - A Flexible Tokenisation Tool. In *Proceedings of Second International Conference on Language Resources and Evaluation*, pages 1147–1154.

Schmid Helmut. 1994. Part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Taku Kudoh. 2007. Crf++. http://crfpp.sourceforge.net/.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pages 282–289.

Steve Lawrence, C. Lee Giles, and Kurt Bollacker. 1999. Autonomous Citation Matching. In *Proceedings of the Third International Conference on Autonomous Agents*, Seattle, Washington, USA, may. ACM Press, New York City, NY, USA.