

Automatic Summarization and Information Extraction from Canadian Immigration Decisions

Emmanuel Chieze^{*}, Atefeh Farzindar[#], Guy Lapalme^{*}

^{*}RALI-DIRO

C.P. 6128, Succ Centre-Ville
Université de Montréal
Montréal, Québec, Canada H3C 3J7

[#]NLP Technologies

3333, chemin Queen Mary, suite 543,
Montréal, Québec, Canada H3V 1A2

E-mail: chiezeem@iro.umontreal.ca, farzindar@nlptechnologies.ca, lapalme@iro.umontreal.ca

Abstract

This paper presents our experience in the use of a mix of linguistics aware transducer and XML technologies for bilingual information extraction from judgments in both French and English within a legal information and summarizing system. We present the context of the work, the main challenges and how they were tackled by clearly separating language and domain dependent terms and vocabularies. The use of Excel sheets for keeping dictionary information enables an easy to use customization approach for non linguists or non computer scientists.

1. Context of the work

One field in which information is produced in large quantities and needs to be adequately classified and be reliably accessible is the legal field. Indeed, legal experts perform relatively difficult legal clerical work that requires accuracy and speed. These legal experts often summarize legal documents, such as court judgments, and look for information relevant to specific cases in these summaries. These tasks involve understanding, interpreting, explaining and researching a wide variety of legal documents. The summary of a judgment is a compressed but accurate statement of the judgment's contents. Summaries help organize a large volume of documents so that finding relevant judgments for a specific case is easy and efficient.

That is why judgments are frequently manually summarized by legal experts. However, human time and expertise required to provide manual summaries for legal research make human-generated summaries relatively expensive. Also, there is always the risk that a legal expert misinterprets a judgment and misclassifies it or produces an erroneous summary.

Because of the high accuracy required in the classification and summarization of legal judgments, commonly available automatic classification and summarization methods are typically not suitable for this task.

NLP Technologies is an enterprise that develops solutions specifically for users of legal search tools. The company's services are available through the company's website¹ and include access to four main tools:

- *DecisionExpress* is the tool that processes judicial

decisions automatically and makes the information used by jurists daily more accessible by presenting the summaries of the legal record of the proceedings of federal courts in Canada as a table-style summary as shown in Figure 1. This service provides some form of continuing education for legal practitioners and saves hours of reading by extracting the essential information and showing it in a uniform format for many cases of the same type.

- *SearchExpress*, integrated within *DecisionExpress* is a search engine that allows users to search the NLP Technologies' database.
- *BiblioExpress* is a virtual law library providing access to legislations, regulations and international instruments.
- *StatisticExpress* is a specialized fact-finder providing fast and easy access to pertinent data and government statistics.

Since August 2006, the Federal Court of Canada has been using the services of NLP Technologies. The summaries are available within 2 days of the publication date.

FLEXICON (Smith & Deedman, 1987), SALOMON (Moens et al., 1999) and SUM projects (Grover et al., 2003) attest the importance of the exploration of legal knowledge for sentence categorisation and summarization (Moens 2007). NLP Technology's extraction of the most important units is based on the identification of the thematic structure in the document and the determination of argumentative themes of the textual units in the judgment (Farzindar & Lapalme, 2004; Farzindar, 2005).

¹ <http://www.NLPtechnologies.ca>





Allowed (✓): 4 Dismissed (X): 8 week of February 26 to March 04, 2007		
Information		
Andryanov v. Canada (Citizenship and Immigration) (2007 FC 186) IMM-1790-06 Date : 20/02/2007		
Subject: Permanent Residence Decision: Allowed Judge: Mosley, Richard  Actions  View Summary	Headnote	The applicant is a Russian national married to a Canadian citizen. His wife sponsored his application for permanent residence. A lengthy and confusing exchange of correspondence followed between the parties concerning a passport which he says he never possessed.
	Topics	Identity document, Passport, Inadequate reasons, Duty to give reasons, Procedural fairness
	Location(s)	RUSSIAN FEDERATION
	Legislation and Conventions	<ul style="list-style-type: none"> o Immigration and Refugee Protection Act section 50(1) section 11(1) o Immigration and Refugee Protection Regulations section 50(1) section 72(1)
Information		
Pravinbhai Shah v. Canada (Citizenship and Immigration) (2007 FC 207) IMM-1670-06 Date : 23/02/2007		
Subject: Skilled workers Decision: Dismissed Judge: Snider, Judith A.  Actions  View Summary	Headnote	Application for permanent resident status in Canada. In a decision an Immigration Officer at the Immigration Section of the Canadian High Commission in New Delhi denied the applicant's application.
	Topics	Presence at hearing, Inadequate notice
	Location(s)	INDIA
	Legislation and Conventions	Immigration and Refugee Protection Act

Figure 1: *Factsheet* from *DecisionExpress* showing two cases from a week in which 4 immigration cases have been allowed and 8 dismissed. The left part give the subject, the decision and the name of the judge while the right part gives a very short summary, the topics dealt in this case, the country in which the applicant resided and the pertinent legislation that was cited in the case. By merely clicking on the appropriate button, it is possible to get a longer summary (shown in Figure 2) or even the text of the original judgment.

2. The Immigration and Refugee Law

We describe in more detail the process of dealing with decisions in the field of Immigration and Refugee Law. All Canadian immigration decisions are retrieved from the Federal courts web site when they become public, and are then processed in order to produce two valuable pieces of information (See Figure 1),: a *Factsheet*, which is a fixed set of structured information automatically extracted from the decisions (name of the judge, name of the case, docket number and neutral citation number, place of hearing ...), and an automatic summary of the decisions, a sequence of relevant sentences taken directly from the original decision and presented in a table (Figure 2). The factsheet clearly identifies such salient information as the subject matter, key words, presiding judge, result, legislation cited, etc., as well as an automatic summary composed of extracts from the decision and presented in a thematic table.

As the Court decisions in this domain are well structured, it is possible to identify three main parts and develop a specialized information extraction process for each:

1. **Prologue:** a list of semi-structured information such as the docket number, the place and date of hearing, the judge's, plaintiffs' and defendants' names. Each piece of information is usually introduced by a specific label but the concept extraction and the determination of the matter of the decision require a more detailed analysis.

2. **Decision:** a full-length text, structured in sections usually identified by titles or by specific sentences starting those sections. A typical decision is divided into six themes usually appearing in the following order: introduction, context, issues raised by the plaintiffs, reasoning, conclusion and the order. Some sections may be missing in some decisions, while additional sections may appear in other ones. The order in which sections appear may also vary.
3. **Epilogue:** another list of semi-structured information such as the lawyers' and solicitors' names.

The information from the prologue and epilogue are kept in a database and an automatic table style summary is produced for the decision. The result is then reviewed by a lawyer from NLP Technologies who can make some manual adjustments. The overall result is revised by an Editorial Board before the information becomes available to the company's subscribers on the Web. This mix of automatic processing and manual revision has been in operation for 2 years and has given very good results on Immigration decisions written in English.

We now describe a new version of that process, being gradually put in production, to extend the system to decisions in the same field written in French and to decisions in other fields covered by the Canadian Federal Law such as tax law and intellectual property law. Two core ideas have presided to this re-engineering: the use of linguistics aware technology and parameterization.

View Summary	
Summary Pravinbhai Shah v. Canada (Citizenship and Immigration)	
Introduction	[1] In 2000, the Applicant applied for permanent resident status in Canada. In a decision dated February 2, 2006, an Immigration Officer (Immigration Officer) at the Immigration Section of the Canadian High Commission in New Delhi (CHC) denied his application. The Applicant seeks judicial review of that decision.
Context	[2] At least in part, the application for permanent residence was rejected because the Applicant had not appeared for his scheduled interview. The Applicant submits that he was never advised of the interview and that, accordingly, the decision of the Immigration Officer should be overturned. In contrast, the Respondent submits that a call-in letter was faxed to the Applicant's representative, Worldwide Immigration Consultancy Services Ltd. (WWICS), on October 12, 2005 to fax number 901725063889, along with six other letters convoking clients of WWICS for interviews. The Respondent presents, as evidence, a copy of a fax confirmation set out on what is alleged to be the first page of the 21 page fax that contained the Applicant's call-in letter.
Reasoning	[3] This application raises the following issue: 1. Did the Immigration Officer err in refusing the application because the Applicant failed to attend the interview due to circumstances beyond his control?[11]... I am satisfied that, if the letter was sent, it was sent to the correct fax number.[13] Accordingly, I am satisfied, on a balance of probabilities, that the 21-page fax was sent, on October 12, 2005, by CHC officials to the correct fax number of WWICS and that the call-in letter to the Applicant was included in the 21-page fax that was sent to WWICS.[14] In his affidavit, Mr. Sandhu raises a number of possible reasons as to why the fax may not have been received. Most of these are speculations and, in any event, do not change my conclusion that the call-in letter was sent to the correct fax number. As noted earlier, problems on the receiving end of the fax (such as mechanical failure or improper administrative procedures) are not the responsibility of the sender.[15] This is not a situation as was encountered by Justice Kelen in Dhoot, above. In that case, the respondent was unable to confirm that the letter was faxed to a correct fax number. Justice Kelen noted that the letterhead of WWICS contained different fax numbers than that set out on the fax receipt. In the case before me, Mr. Sandhu confirmed that the fax number was that of WWICS.
Conclusion	The application for judicial review will be dismissed.

Figure 2: Automatically generated, and manually revised, summary returned after clicking on the *View Summary* button at the bottom left of Figure 1. All sentences of the summary being taken verbatim from the original decision, they can thus be used as argumentation. The sentences are classified into meaningful sections: Introduction, Context, Reasoning and Conclusion. Note that sentences are not necessarily in the same order in the judgment and in the summary.

3. Overview of the linguistics aware Information Extraction Process

Canadian immigration decisions are available on the Web² as HTML documents and can be in English or French depending on the language used at the hearing. A decision may naturally be relevant for Canadian lawyers no matter in which language it is written. Since HTML tags define the presentation of those decisions, rather than their structure, and since the presentation as well as its HTML definition is liable to evolve over time (and it has...), we cannot rely on only these tags to identify the structure of the decisions. We will thus have to analyze the text of the decision itself to discover what parts of the text are part of each section that will appear in the summary.

Figure 3 shows a simplified view of the overall transformation pipeline combining different technologies to

go from an original judgment as an HTML file taken from the web site of the Canadian Federal Court to an XML file that is saved within a data base from which the final summary, also in HTML, is generated. This XML file can also be changed during the manual revision process by NLP Technologies lawyers that access it through a specialized revision interface.

This transformation process involves both local (within a sentence) processing, more global processing taking into account parts of the documents that can be farther apart and statistical processing for computing the salient sentences that will compose the final summary.

We have decided to use technologies that are appropriate for each step of the transformation. Transducers allow a great flexibility in sentence processing, XSLT stylesheets are an efficient mean for selecting and transforming longer spans of texts and a procedural language is used for computing the final statistics to select the final sentences to appear in the summary.

² <http://decisions.fct-cf.gc.ca/fr/index.html>

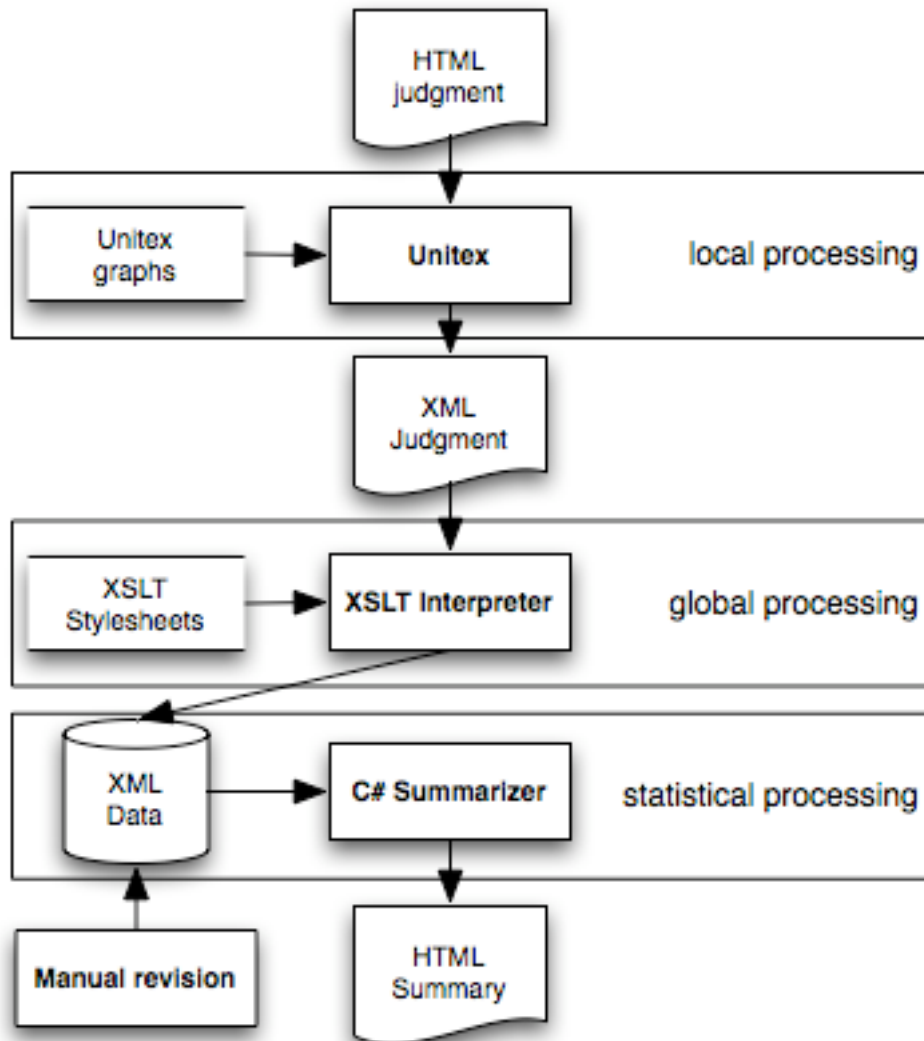


Figure 3: System architecture going from the original to the summary. Unitex graphs are used for going from HTML to XML and for linguistic processing within a sentence or for short spans of text. XML Transformation Stylesheets enable to take into account long distance dependencies and the statistical computations for determining the most important sentences to appear in the summary are done by a C# program.

3.1 Local processing

A first step is thus to convert HTML documents into text files and then use linguistic cues to identify the decision structure as well as the relevant factual information. Fortunately, decisions follow a rather stereotypical pattern and use recurrent information identifiers or section headings. Such identifiers have several variants, but there are usually a fixed set of them.

We decided to use XML tags to identify text structure and relevant factual information, since there are several general-purpose XML-based processing tools, such as structure validation or document transformation tools. So our process will first eliminate most HTML tags and transform others into paragraph markers.

Relevant information will then be identified through linguistic cues, which are phrases identifiable through context-free grammars. As we are aiming for power and flexibility we decided to make use of the transducer technology, namely Unitex³, a descendant of INTEX (Silberstein 1973), to identify, mark and transform spans of texts by means of regular expressions which provides the following advantages:

1. Regular expressions are represented with graphs (see Figure 4 for an example) instead of complex sequences of operators and their base unit is the word rather than the character. Language-dependant character equivalences are appropriately handled.

³ <http://www-igm.univ-mlv.fr/~unitex/>

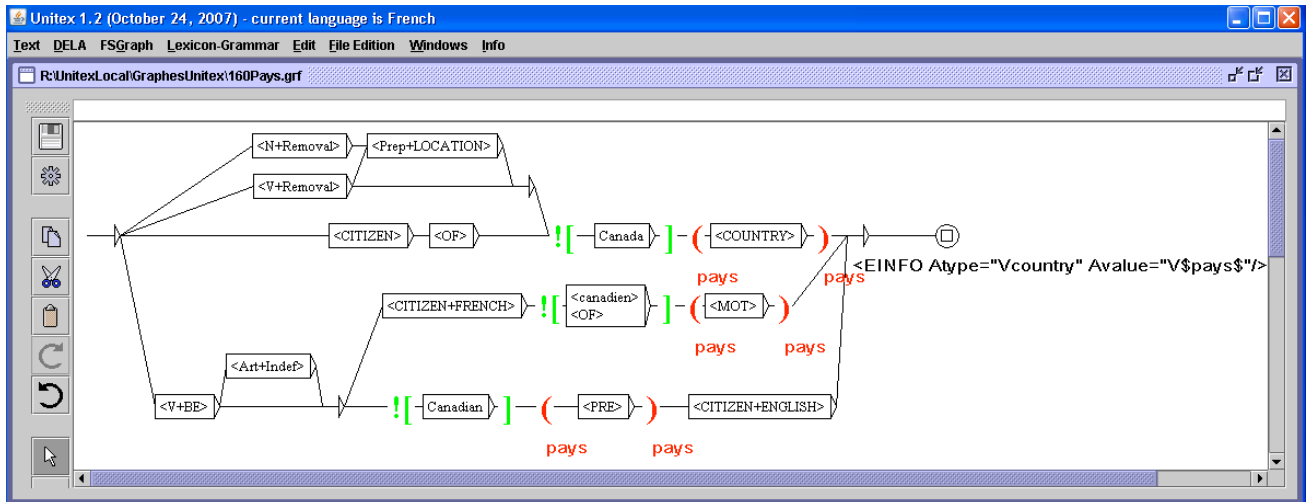


Figure 4: A graph defines a set of paths matching words encountered in the text going from the entry node (the triangle on the far left) to the exit node (the circle containing a square) on the right. A node can match either be a single word (see *Canadian* above), or one word contained in a list defined in the dictionary (see *<COUNTRY>* above). When a path going from the entry to exit has been found, information can be added (shown here in bold) to the original text. Here the occurrence detected is tagged with an XML tag named *EINFO* with attributes *ATYPE* having value *country* and *Avalue* having a value *pays* that was saved during matching this graph. This graph detects the country from which the applicant originates. The 4 paths out of the start state, from top to bottom, correspond respectively to: 1) a path that recognized phrases such as "his removal to Kenya", 2) a path that recognizes phrases such as "[is scheduled to be] removed to Kenya", 3) a path that recognizes phrases such as "[is a] citizen of Kenya", 4) a path that recognizes phrases such as "[is a] Kenyan citizen" or "[est] citoyen kenyan". Note that adjectives derived from country names, recognized by the last path, are not listed in the dictionary contrary to country names, which are listed.

2. It works with a user-defined dictionary in which words and phrases may be assigned various user-defined syntactic or semantic categories which may in turn be used in graphs. Flexional categories and morphological criteria can be almost freely combined with those syntactic and semantic categories, enabling the expression of complex search criteria without ever having to translate those criteria into character patterns.
3. Graphs may be used as subgraphs of other more complex graphs, enabling graph reusability.
4. Parameterized graphs (explained in the next section) add even more flexibility to our processing.

Unitex graphs have the power and efficiency of regular expressions, with the additional benefits of linguistic awareness and much improved user-friendliness. These grammars recognize word patterns most often limited to a single sentence. Unitex processing of the judgments involve the use of 33 compiled graphs for transforming the HTML form of a judgment to a labeled XML file. An example of such a graph that detects the applicant's country of origin is displayed as Figure 4.

3.2 Global processing

Although there is no theoretical bounds on the span of input that can be processed by a Unitex transducer, in practice we have experienced many problems when the input is too long. Unitex is cumbersome for expressing long-range dependencies but there are how-

ever a few contextual or structural rules to implement, such as:

- A sentence that contains a pattern associated with salient phrases of section X is a salient phrase of section X if and only if it appears in that section.
- All sentences of a paragraph following a sentence identified as a citation are also part of that citation.

We decided to express such structural rules with XSLT stylesheets applied to the resulting XML format of the documents.

Using XML provides the additional benefit of checking the conformity of the document structure to the XML schema associated with decisions. The XSLT processing uses 10 templates.

3.3 Statistical processing

The above processing has tagged the original text without modifying it but to identify the sentences to appear in the summary, some statistical computations are involved such as the computation of TF-IDF scores and other numerical values. This process is done with a C# program that parses the XML document produced by the previous two steps.

The HTML input files are about 30K characters long, corresponding to 16K words. On a stock desktop PC, the processing time for applying Unitex graphs, processing XSLT templates and computing statistics is about 40 seconds by judgment.

4. Parameterization of the Information Extraction Process

As shown in Figure 4, Unitex graphs can refer to words defined in a dictionary, a user-defined list of word forms associated with their root form as well as various syntactic and semantic categories and morphological features. It would be cumbersome to define all word forms by hand, especially in an inflected language like French in which semantic categories do not vary with the flexion, Unitex offers two types of dictionary definitions: the inflected dictionary, where it is possible to directly define word forms, and the non inflected dictionary, which will be inflected by Unitex using an inflexion graph provided by the user. Such graphs are language dependent but are application domain independent.

Unitex offers an additional mechanism called the parameterized graph, which combines a generic graph containing variables and a parameter file. The latter is a text file containing the values to be taken by the variables. More precisely, each line of the parameter file will generate a subgraph, and the whole family of subgraphs will be integrated as a single graph. Each subgraph thus represents an alternative and the main graph a disjunction of all those alternatives.

In order to maximize the parameterization of our system, we have made an extensive use of the dictionary as well as of parameterized graphs, so that many graph updates can simply be made through the update of those parameter files followed by a graph recompilation. We have used Microsoft Excel in order to gather the various parameter files in one single place, to make data definition user-friendlier, to validate it with an Excel macro, which includes cross-checks between those lists when applicable and to give the user a user-friendly way of consulting, sorting and filtering those parameter lists.

Some operators such as *X in-same-sentence-as Y* or *X near Y*, not available in Unitex have been developed with auxiliary graphs, and can be used in those lists to implement complex rules: there is a fixed list of them however, since we did not want to implement a general rule compiler.

In total, there are 10 worksheets in this Excel file: each of them parameterizes a specific aspect of the information extraction process. The dictionary itself contains 432 uninflected single words, 840 inflected single words or single words without any flexion and 812 phrases. Those figures combine both English and French entries. In a specialized information extraction setting like this one, we only have to deal with words that are used for segmenting the judgment or for identifying specific information like dates, names of parties. Most of the words encountered in the text are simply taken as is and will be given back verbatim if it happens that the sentence as a whole is chosen to appear in the summary.

5. Maintenance of the Information Extraction Process

The information extraction transductors were developed originally by the manual inspection of about 60 decisions in both English and French published in 2007. Only a few (about 5%) of current decision are not processed correctly and imply some manual adjustment either by correcting the formatting of the input or by adding new words in the dictionary.

We have also tested the transductors on 14 380 *historical* decisions published between 1997 and 2006. Only 15% of those decisions were incorrectly processed by the original information extraction process, i.e. the resulting XML document was not well-formed, usually because the beginning of a section was detected but not its end or vice-versa. This happens because these complementary elements are tagged independently.

Resolving the problems caused by 9 decisions helped resolve the problems encountered in 49 additional decisions (over 90 decisions tested). In other words, a single problem occurred on average on 6.5 decisions among the 90 decisions on which corrections were tested. Among those 9 problems, 3 implied adding entries in the dictionary, 5 implied modifying existing graphs in order to improve their flexibility. We decided not to take any step on the last one which was caused by a misspelling in the decision. It is yet unclear whether our parameterization effort has been sufficient, since only 3 problems out of 8 could be solved without modifying any graph. We are just at the beginning of the correction process however, and we hope that, as time goes on, a higher proportion of problems will be solved through dictionary update, as well as we can hope that one single correction will have a positive impact on more decisions. Moreover, we know that decisions have been presented in a considerably more homogeneous way since 2003, so that historical results are worse than those obtained on current decisions.

So we are confident that as the time goes on, there will be less and less manual work to do by NLP technology legal staff who will merely check that everything is all right for publication. It is too early to do a formal evaluation of the new process both in terms of the efficiency of extraction and in the reduction of manual corrections needed before the judgment is put on the NLP Technologies web site. But as the process is good enough to be gradually put in production, we are very pleased with the result.

6. Conclusion and Perspectives

DecisionExpress is the first service in the world based on an automatic summarization system developed specifically for legal documents. It is implemented in a real-life environment and currently produces summaries for large collections of judgments (between 25 and 50 each week) written in English in the immigration domain.

In this article, we have presented our recent work for extending the applicability of the system to French and to other domains such as financial field and intellectual property field. The main idea was to separate the linguistic cues used to achieve a precise information extraction in different domains. The output of the system is systematically reviewed by a lawyer but the goal is to have the system do as much work as possible.

To allow NLP Technologies client to work in the language they are most comfortable with, a project of automatic translation summary of judgments is under way. That would help users during the (up to nine) months it takes for the official translation to be published. As the summaries are obtained with extracts of the original judgment, the decision could be summarized both in English and in French, regardless of the original language of the judgment by taking the corresponding extracts from the automatic translation.

7. Acknowledgement

This work is partially funded by the PRECARN Alliance Program of Canada and The National Research Council Canada (NRC).

8. References

- Farzindar, A. (2005). *Automatic summarization of legal texts*, Ph.D. Thesis, University of Montreal and University of Paris IV-Sorbonne.
- Farzindar, A. and Lapalme, G. (2004). LetSUM, an automatic Legal Text Summarizing System. In Thomas F. Gordon (editors), *Legal Knowledge and Information Systems, Jurix 2004: the Seventeenth Annual Conference*, IOS Press, Berlin, pp. 11-18.
- Grover, C., Hachey, B. and Korycinski, C. (2003). Summarising legal texts: Sentential tense and argumentative roles. In Radev, D. and Teufel, S., editors, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, Edmonton, Alberta, Canada, pp. 33-40.
- Moens, M.F. (2007) Summarizing court decisions, *Information Processing and Management*, vol 43, pp. 1748-1764.
- Moens, M.F., Uyttendaele, C. and Dumortier, J. (1999). Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50(2), pp. 151-161.
- Silberztein, M.D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson, 234 p.
- Smith, J.C. and Deedman, C. (1987). The application of expert systems technology to case-based law. *ICAIL*, pp. 84-93.