An Automatic System for Summarization and Information Extraction of Legal Information

Emmanuel Chieze[1,⋆], Atefeh Farzindar[2], and Guy Lapalme[3]

[1] Département d'informatique
Université du Québec à Montréal
C.P. 8888, Succ. Centre-ville
Montréal, Québec, Canada, H3C 3P8
chieze.emmanuel@uqam.ca
[2] *NLP Technologies Inc*
3333, chemin Queen Mary, suite 543
Montréal, Québec, Canada H3V 1A2
farzindar@nlptechnologies.ca
[3] RALI-DIRO
C.P. 6128, Succ. Centre-ville
Université de Montréal
Montréal, Québec, Canada H3C 3J7
lapalme@iro.umontreal.ca

Abstract. This paper presents an information system for legal professionals that integrates natural language processing technologies such as text classification and summarization. We describe our experience in the use of a mix of linguistics aware transductor and XML technologies for bilingual information extraction from judgements 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. After having been developed on the immigration law domain, the system was easily ported to the intellectual property and tax law domains.

Keywords: Summarization, Natural Language Processing, Information Extraction.

1 Context of the Work

Legal information is produced in large quantities and it needs to be adequately classified in order to be reliably accessible. Indeed, legal experts perform relatively difficult legal clerical work that requires accuracy and speed. These legal experts often summarize legal documents, such as court judgements, and look for information relevant to specific cases in these summaries. These tasks involve

⋆ This work was performed while Emmanuel Chieze was at RALI-DIRO, Université de Montréal.

E. Francesconi et al. (Eds.): Semantic Processing of Legal Texts, LNAI 6036, pp. 216–234, 2010.

understanding, interpreting, explaining and researching a wide variety of legal documents.

To help in some of these tasks, *NLP Technologies*[1] has developed a series of advanced information technologies in the judicial domain. *NLP Technologies* is an automated language software company conducting the research, development and marketing of summarization and statistical machine translation software and related software tools and services. The company's services are available through the company's website and include access to four main tools: *DecisionExpress*, *SearchExpress*, *BiblioExpress* and *StatisticExpress* which are briefly described below.

The core technology underlying these tools is an automatic summarization system. Summaries help organize large volumes of documents so that finding relevant judgements for a specific case is both easy and efficient. That is why judgements 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 judgement and misclassifies it or produces an erroneous summary. Because of the high accuracy required in the classification and summarization of legal judgements, commonly available automatic classification and summarization methods are typically not suitable for this task. Based on the work of Farzindar [2], *NLP Technologies* has developed a summarization system specifically tailored for the legal domain based on a thematic segmentation of the text. Since 2005, the Federal Court of Canada has been a client of *NLP Technologies*'s automated legal analysis services for French and English documents. The summaries are available within 2 days of the publication date. Although this process must be adapted for new domains, the fundamentals stay the same and one of the goals of this work was to develop a methodology that allows an easy parameterization process through appropriate dictionaries and rules using advanced natural language processing tools such as transductors.

1.1 *DecisionExpress*

DecisionExpress is a weekly bulletin of recent decisions of Canadian federal courts and provincial tribunals. It processes judicial decisions automatically and makes the daily information used by jurists more accessible by presenting the summaries of the legal record of the proceedings of federal courts (such as Tax court, Federal court of appeal, etc.) and provincial tribunals in Canada.

Furthermore, it presents a factsheet for each decision that can save hours of reading by extracting the essential information and showing it in a user-friendly format for many cases of the same type.

Contrary to the traditional way of manually classifying and summarizing judgements to be saved in the database, *DecisionExpress* analyses and summarizes the judgements automatically. This brings numerous advantages both for those publishing legal information and the jurists using it:

[1] `http://www.nlptechnologies.ca`

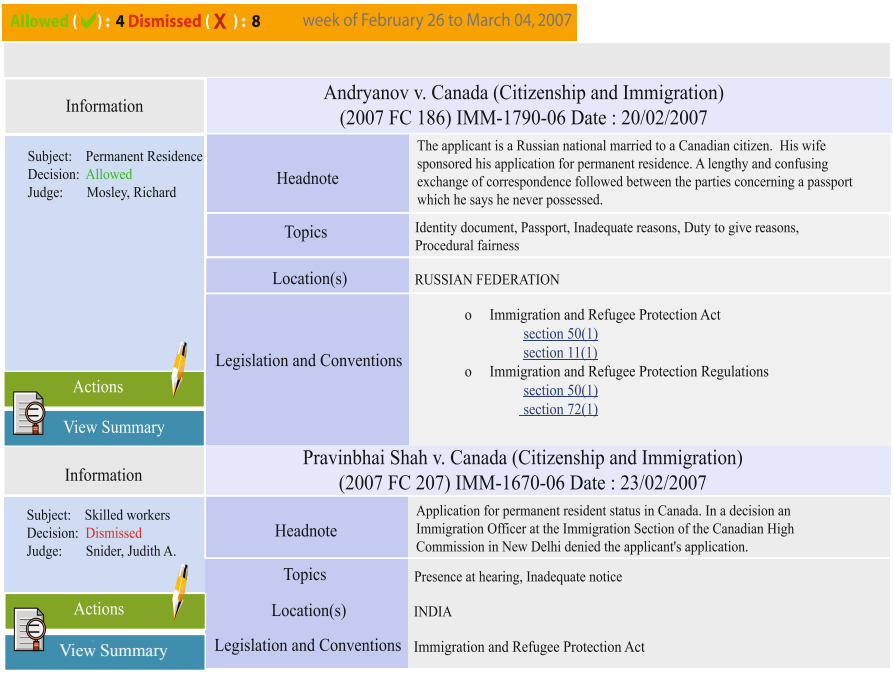

Fig. 1. Factsheet from *DecisionExpress* showing two cases from a week in which 4 immigration cases have been allowed and 8 dismissed. The left part gives the subject, the decision and the name of the judge while the right part gives a very short summary, the topics dealt with in this case, the country in which the applicant resided and the pertinent legislation that was cited in the case. Clicking on the appropriate button gives access to a longer summary (Figure 2) or the text of the original judgement.

- Significant cost reduction of the summary production process which can be passed back to those accessing the information as customers.
- Automatic summaries present sentences extracted from the judgement, whereas manual summaries consist of reformulations. A reformulation is less precise and less credible because it is not a direct source of law. In addition, an ambiguous reformulation can may lead to misinterpretations of what the judge meant and lead the user to erroneous beliefs.
- Automatic summaries provide greater consistency. The editors' abilities and concentration may vary, whereas the computer provides a stable level of performance. The machine is also better suited than a human for repetitive tedious tasks, such as the production of summaries of long articles.

DecisionExpress' other innovation is the production of a brief description of every decision analysed. This description allows a jurist to get the essential information of a decision in one glance. This way, he or she knows immediately if the decision is relevant enough to read the summary and eventually the whole judgement.

The thematic segmentation is based on specific knowledge of the legal field. According to our analysis, legal texts have a thematic structure independent of the category of the judgement [1]. Textual units dealing with the same subject form a thematic segment set. In this context, we distinguish four themes, which divide the legal decisions into thematic segments, based on the work of judge Mailhot[5]:

Introduction describes the situation before the court and answers these questions: who did what to whom?

Context explains the facts in chronological order: it describes the story including the facts and events related to the parties and it presents findings of credibility related to the disputed facts.

Reasoning describes the comments of the judge and the finding of facts, and the application of the law to the found facts. This section of the judgement is the most important part for legal experts because it presents the solution to the problem between the parties and leads the judgement to a conclusion.

Conclusion expresses the disposition, which is the final part of a decision containing the information about what is decided by the court.

The factsheet (see Figure 1) presents information such as the name of the judge who signed the judgement and the tribunal he or she belongs to, the domain of law and the subject of the decision (for example, immigration and application for permanent residence), a short description of the litigated point, the judge's conclusion (allowed or dismissed) and hyperlinks to the summary (Figure 2) and the original judgement.

These factsheets are highly appreciated by users because they present the essential information about a judgement more concisely than a summary. One glance is enough to determine if the decision is relevant. Moreover, the factsheets are automatically translated into French or English so that for every decision, the factsheet is available in both offical languages of Canada. This allows jurists to work in the language they are most comfortable with regardless of the language in which the decision was published.

1.2 *SearchExpress*

SearchExpress, integrated within *DecisionExpress* is a search engine that allows users to search the *NLP Technologies*' database rendered by Canadian federal courts and tribunals. In addition to the search functionality already offered by most providers of legal information such as QuickLaw[2] and Westlaw-Carswell[3], *SearchExpress* offers new possibilities. Search the factsheets generated by *DecisionExpress*. This way, the user can formulate the query based on the judge's name, his conclusion, the domain of law, the subject of the decision, the keywords, etc. In short, the query can be constructed using any information presented in the factsheets, which allows the user to refine his or her search.

[2] `http://www.lexisnexis.ca`

[3] `http://www.carswell.com`

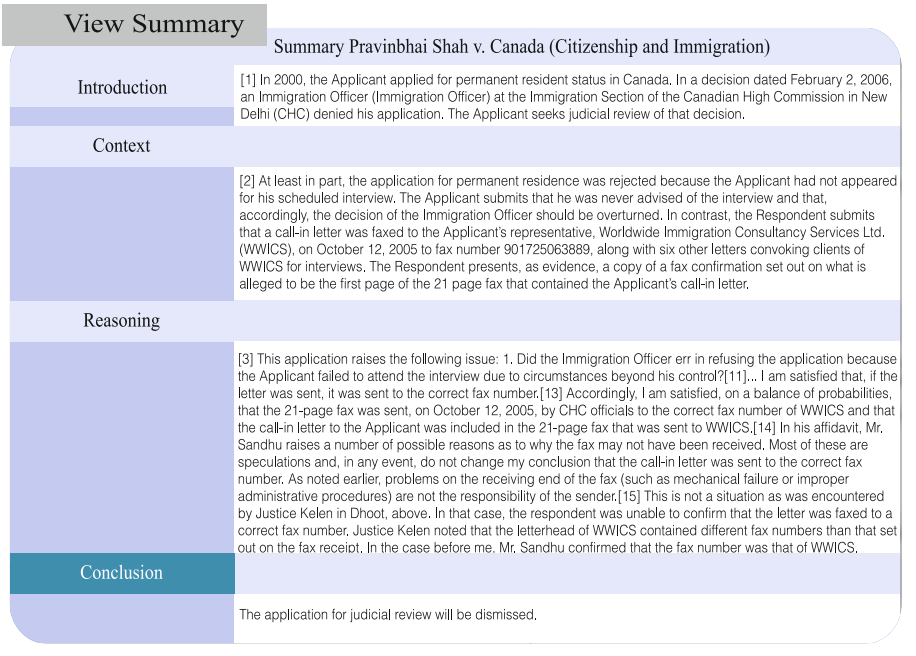

Fig. 2. Automatically generated, and manually reviewed, summary returned after clicking on the View Summary button at the bottom left of Figure 1. All sentences of the summary being taken verbatim from the original decision, they can thus be used more easily in legal argument. The sentences are classified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

Regardless of the type of search used, the results page presents, for every decision found, the factsheet of the decision as well as a hyperlink to the original text. This manner of presenting the results permits the user to save time in the preliminary sort of retrieved results. Instead of reading every decision retrieved to see if it is relevant, he or she can simply reject the decisions whose factsheets show clearly that they will not be useful. The overview presented in the factsheets also allows telling quickly if the query should be refined or otherwise modified.

Searching is done both in the full text of the judgement and in the factsheets generated by *DecisionExpress*. Consider a lawyer preparing a file for a client contesting in the Federal Court the refusal of his application for residence based on humanitarian considerations. On the other providers' websites, this lawyer could do a search in the decisions of the Federal Court by typing keywords such as immigration or application for refugee status (very broad) or humanitarian considerations (more precise but not always related specifically to the humanitarian application

process per se). With *SearchExpress*, the lawyer can search only those cases which have been correctly identified as judicial reviews of humanitarian applications per se among the Federal Court by limiting the results of the query to judgements labelled by *DecisionExpress* as immigration for the domain, and "humanitarian and compassionate application" for the subject. He or she could also choose only to retrieve decisions where the judge has granted the application for judicial review, or else limit the search to decisions with respect to applicants from the same country as his or her client. Cross-lingual (English-French) search allows the user to submit a query in one language and retrieve documents containing the terms of the query as well as their equivalents in the other language. The user can thus use a single query to retrieve all documents relevant to his or her case regardless of the language the judgement was made in. *SearchExpress* makes search easier by incorporating unique and useful search criteria (Figure 3) such as category (e.g., immigration and tax), court, name of judge, subject (e.g., investors, pre-removal risk assessment, and humanitarian considerations), conclusion (allowed or dismissed) and other relevant criteria.

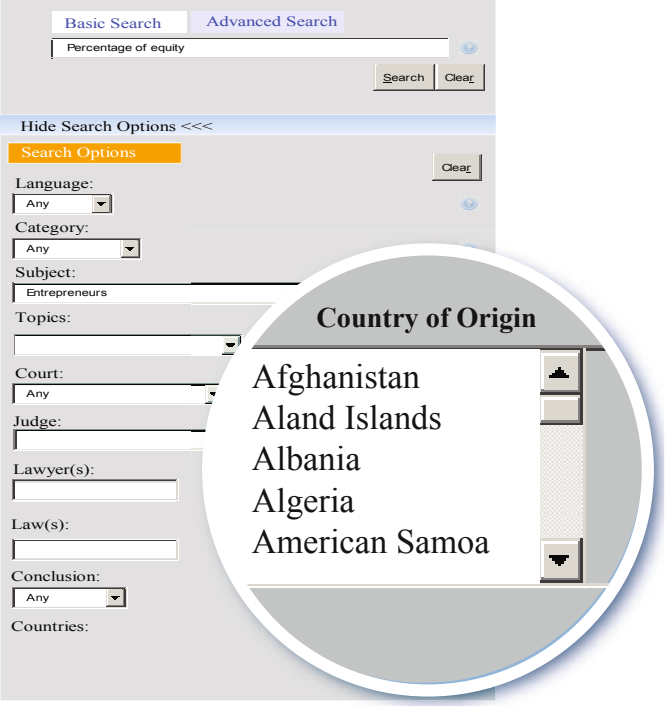

Fig. 3. With *SearchExpress*, it is possible to have access to relevant decisions for a research in progress by specifying the criteria. For example, a lawyer can carry out research on a precise case such as an entrepreneur who comes from a certain country with particular conditions to see how such situations have been treated historically in order to calculate the chance of success in court.

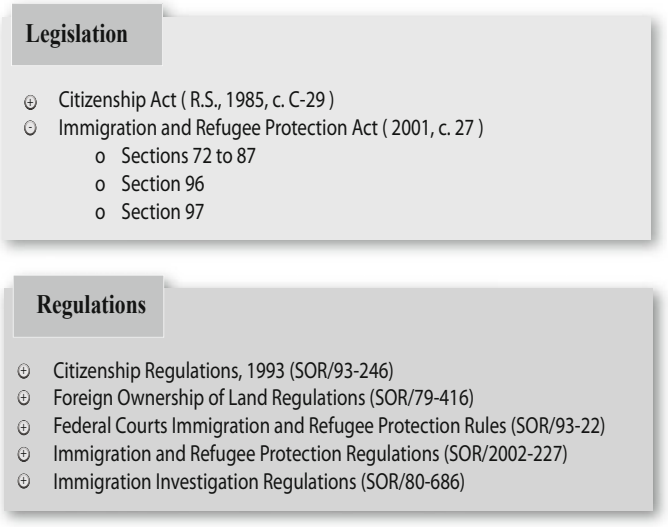

Fig. 4. The legal sources related to a subject are classified into different categories by *BiblioExpress*. For example, the links related to an immigration subject are centralized in one page.

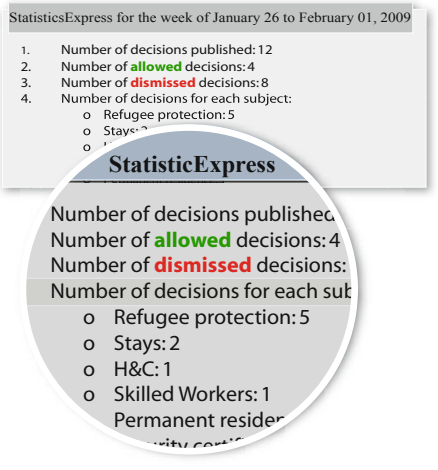

Fig. 5. Weekly statistics provided by *StatisticExpress*

1.3 *BiblioExpress*

BiblioExpress provides access to the text of federal legislations, rules, policy manuals and guidelines, as well as a range of inter-governmental agreements and international instruments in Canada. This service (see Figure 4) centralizes links to fundamental legal resources of three different categories: immigration, intellectual property, and tax. For instance, in the *Immigration of Canada* domain, *BiblioExpress* classifies the links to recourses into legislations, regulations, rules, conventions, guidelines, forms, agreements, etc.

1.4 *StatisticExpress*

StatisticExpress gives access to pertinent data and a variety of government statistics such as the annual and periodical reports of courts, tribunals and government agencies, international statistics, the performance reports of various government and international agencies and a specialized fact-finder providing statistics from *DecisionExpress*'s databases shown in Figure 5.

2 Research Background

The best source for an overview of legal text summarization is Moens [6] who presents an excellent survey of the area of summarization of court decisions. She describes the context in which court decisions are taken and published and the need for good quality summaries in this area which is comparable to the medical domain.

FLEXICON [9] is one of the first summarization system specialized for legal texts. It was based on the use of keywords found in a legal phrase dictionary. The summaries were not used as such but served for indexing a legal case text collection. SALOMON [7], developed for summarization of Belgian criminal cases, was the first to explicitly make use of the structure of a case. As such, the authors were more interested in identifying the structure than producing a complete summary. The SUM [4] project was developed to determine the rhetorical status of sentences of House of Lords judgements. This methodology could be used as a background technology for a complete summarization system.

These projects attest to the importance of the exploration of legal knowledge for sentence categorisation and summarization. *NLP Technologies*'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 judgement[1,2]. The system we describe in this paper is the only one that spans all steps from an original judgement to a complete summaries that can be used in the daily activity of legal professionals.

3 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 : the factsheet (See Figure 1) and an automatic summary of the decisions (Figure 2).

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:

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.

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.

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 reviewed by an editorial board before the information becomes available to the company's subscribers on the Web. This mix of automatic processing and manual review has been in operation for 4 years and has given very good results on Immigration decisions written in English. Using the parameterization process described below, we were able to extend, in the course of 2008, the system to decisions in the same field written in French and to decisions in tax and intellectual property laws. Two core ideas have presided to this re-engineering: the use of a linguistics aware technology and parameterization.

4 Linguistics–Aware Information Extraction Process

Canadian immigration decisions are available on the Web as HTML documents either 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 thus analyze the text of the decision itself to discover the sentences of each section to appear in the summary.

Figure 6 shows a simplified view of the transformation pipeline combining different technologies to go from an original judgement as an HTML to an XML file that is saved in a data base from which the final summary, also in HTML, is

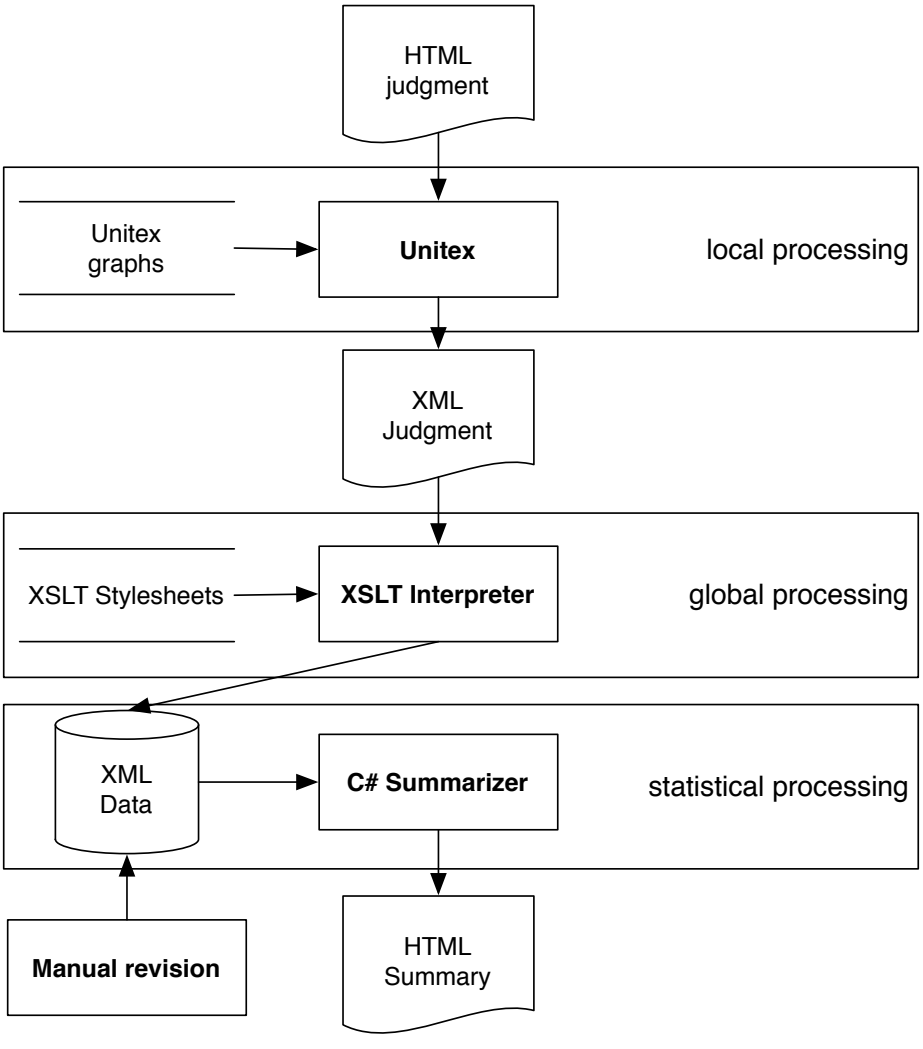

Fig. 6. System architecture going from the original to the summary. Unitex graphs are used for going from HTML to XML and for linguistic processing within a sentence or for short spans of text. XML Transformation Stylesheets (XSLT) are able to take into account long distance dependencies and the statistical computations for determining the most important sentences to appear in the summary are done by a C# program.

generated. *NLP Technologies* lawyers, through a specialized reviewer interface, can also change this XML file during the manual review process. 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 decided to use technologies that are appropriate for each step of the transformation. Transductors allow a great flexibility in sentence processing, XSLT stylesheets are an efficient means for selecting and transforming longer spans of texts and a procedural language is used for computing the final statistics to select the final sentences that appear in the summary.

The unit of work in all transformation steps is the whole sentence in order to guarantee that the summary contains only original sentences that can be cited verbatim without having to consult the judgement. Transductors and stylesheets add *hidden* information to sentences of the original text to provide hints to the final statistical summarization module that decides for each sentence whether it will appear in the summary or not, and if so, in which thematic segment it will be put. Even the manual reviewers work at the level of sentence and choose to either add or remove a whole sentence or not; they are not allowed to modify the wordings of sentences.

4.1 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 transductor technology, namely Unitex[4], a descendant of INTEX [8], to identify, mark and transform spans of texts by means of regular expressions which provides the following advantages:

- Regular expressions are represented with graphs (see Figure 7 for an example) instead of complex sequences of operators and their base unit is the word rather than the character. Language-dependent character equivalences are appropriately handled.
- 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.
- Graphs may be used as subgraphs of other more complex graphs, enabling graph reusability.

[4] `http://www-igm.univ-mlv.fr/~unitex/`

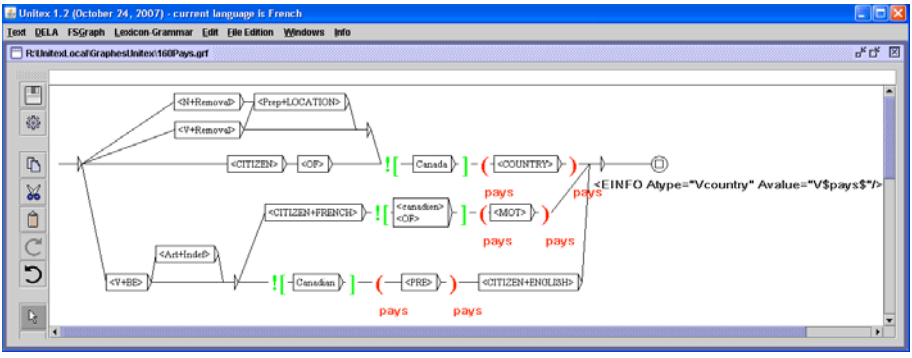

Fig. 7. A graph defines a set of paths matching words encountered in the text going from the entry node (the triangle on the far left) to the exit node (the circle containing a square) on the right. A node can match either be a single word (see `Canadian` above), or one word contained in a list defined in the dictionary (see `<COUNTRY>` above). When a path going from the entry to the 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.

– 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 judgements involve the use of 33 compiled graphs for transforming the HTML form of a judgement to a labeled XML file. An example of such a graph that detects the applicant's country of origin is displayed as Figure 7.

4.2 Global Processing

Although there is no theoretical limit on the span of input that can be processed by a Unitex transductor, in practice we have experienced many problems when the input is too long. Unitex is cumbersome for expressing long-range dependencies but there are however a few contextual or structural rules to implement, such as:

- A sentence that contains a pattern associated with salient phrases of a given section (introduction, context, citation, reasoning, conclusion). If a pattern, typical of a given section, is found in a sentence then the whole sentence is assigned to this 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.

4.3 Statistical Processing

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 per judgement.

5 Parameterization of the Information Extraction Process

As partly shown in Figure 7, 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 is 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 to assemble the various parameter files and to simplify the data definition. Excel macros are used for

validation and for cross-checks between those lists. Excel is also 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 judgement 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.

6 Application to the Intellectual Property and Tax Law Domains

Once the information extraction process was completed for the immigration domain, a natural step was to extend it to other law domains of interest to *NLP Technologies*, namely the intellectual property and the tax law domains. In both cases, federal decisions were the only ones taken into consideration.

The intellectual property domain was very easy to integrate: decisions from this domain emanate from the same courts as the immigration ones, and thus follow the same structure. The main differences between both domains lie in the subjects, topics, and laws associated with one domain or the other, as well as to some specific dictionary entries. Since these data are parameterized in an Excel file, it was very easy to add French and English data relevant to the intellectual property domain to the file. Only very minor reorganizations of this file were required, such as adding a domain field to the topic and the subject worksheets in order to facilitate their maintenance. The integration of the intellectual property domain took about three weeks.

Integrating the tax law domain was more challenging: decisions from this domain can originate from the Federal Court or the Federal Court of Appeal, as is the case for the previous domains, but also from the Tax Court of Canada. Decisions from the latter differ in structure from those issued by the Federal Court or the Federal Court of Appeal, especially in the order and way in which the prologue and epilogue information is presented, and thus required not only an update of the Excel file, but also some modifications of the local processing step. Since we wanted to keep one single processing unit for all decisions, we just added a parameter that states whether a decision is issued form the Federal Court of the Federal Court of Appeal, on the one side, or from the Federal Tax Court on the other side. It must be emphasized that these differences were big enough to justify adding two Unitex graphs for the Federal Tax Court decisions

Table 1. Number of topics and subjects associated with each domain. The *All domain* (not shown) indicates topics and subjects independent of a specific domain. They are associated with court practice questions that arise in all domains.

Domain	**Subjects**		**Topics**	
	examples	nb	examples	nb
Immigration	Appeal by Permanent Resident, Appeal by Protected Person, Citizenship application, Entrepreneurs, Family class application, Inadmissibility, Investors, Convention Refugee Abroad class, Refugee Protection, Enforcement of Removal orders, Skilled workers, Stay of removal orders, Study Permit, Visitors, Work permit, Source country class ...	33	Assurances against torture, Child custody order, Deserters, Failure to seek protection, Gangs, Habeas Corpus, HIV-positive, Identity, Irreparable harm, Obligation to avoid risk of persecution, Oral interview, Religious conversion, Removal from record, Review of detention, Risk assessment, Street gangs, Vengeance, Visa officer ...	1864
Intellectual Property	Patent, Trademark, Copyright, Industrial Design ...	4	adding parties, deadwood, elastomer, ex turpi causa, processability, prodrugs, Pseudonyms, recording medium, representation by non-lawyer, titles, work product ...	745
Tax	Income Tax, Income Tax Québec, Unemployment/Employment Insurance, Excise Tax, Goods and Services Tax GST, Canada Pension Plan, Old Age Security, Petroleum and Gas, Cultural Property Export and Import, Customs, War Veterans, Softwood Lumber, Tax Court Practice, Aboriginals ...	14	acupuncture, automobile allowance, bill of costs, business investment losses, constructive trust, contents of appeal book, foreign-based documents, incarceration, investment brokers, lawyers' disbursements, motion to reconsider, unjust enrichment, vehicle fees ...	1880
All		1		46
Total		52		4535

and making two versions of two existing graphs, but those modifications are still very minor and did not call for a major rethinking of the whole processing chain.

A few more technical adaptations of the process were required, such as splitting the execution of the topic graphs into five steps rather than as a whole for performance reasons. Once again, despite those minor updates, the integration of the tax law domain took only about six weeks. Whatever the domain associated with a decision, the latter is processed in the same way. It must be noted that the local

or global processing steps do not attempt to assign a domain to the decisions they process. Theoretically, an immigration topic or subject could thus be erroneously attributed to an intellectual property decision for instance. This is however very rare, and those mistakes can be corrected either by the statistical processing or the manual review that follow the local and global processing steps.

The integration of new domains (their number is indicated in Table 1) was thus almost effortless thanks to the parameterization approach described above to manage the immigration domain. This success in adding new domains shows that, although our methodology is primarily based on hand defined dictionaries and transductor graphs, these can be quickly adapted because we stay within the law domain for which the fundamentals stay constant.

7 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 were not processed correctly and involved some manual adjustment either by correcting the formatting of the input or by adding new words to 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 action 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 from current decisions.

Thus, we are confident that as time goes on, there will be increasingly less manual work to do by *NLP Technologies* legal staff, who will merely need to check that everything is all right for publication. This process is in production since the summer of 2008.

Farzindar [2] compared the approach underlying *DecisionExpress* and other state-of-the-art summarization systems, but we are not aware of any similar commercial legal summarization system.

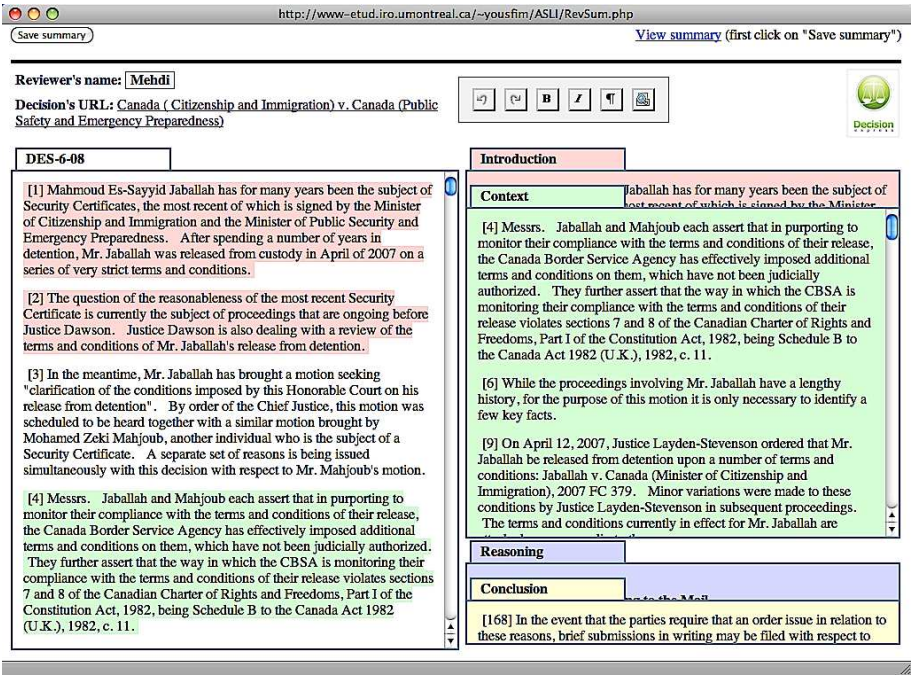

Fig. 8. Interface for the manual review (bottom left box of Figure 6) of the summaries produced by the automatic process within *DecisionExpress*. The left part shows all sentences of the judgement. Paragraphs selected by the automatic process are highlighted with a different color according to the theme they were assigned to. In this figure, the reviewer is working with the *Context* theme whose the tab is currently *opened*. To remove a paragraph from the summary, the reviewer double-clicks on a paragraph from the theme, to add a paragraph to this theme she double-clicks on a paragraph in the judgement. It is possible to add or remove single sentence but its content cannot be changed. This ensures that the original text is preserved in the summary. Simple sentence or paragraph highlighting (bold or italics) buttons are available at the top right. Once the reviewer is satisfied, the resulting summary can be saved in the database using the top left button.

Although we did not conduct any formal evaluation, the feedback given by the Federal Courts is that they find the results of the summaries produced and reviewed by *DecisionExpress* 100% precise and very useful. The electronic dissemination of the judicial decisions within the Federal Courts offices made possible by *DecisionExpress* also brought an interesting *environmental* benefit. The Federal Court used to print its weekly decisions for all of its judges about 1.5 million pages yearly. When the decisions were no longer used by the judges, they were picked up and stored. After the implementation of *DecisionExpress*, a poll taken amongst judges showed that a massive majority agreed that the Court should stop providing printed copies of the judgements.

Human reviewers find that about 70% of the sentences or paragraphs are identified correctly by the automatic system described in this paper. We are currently improving the system using statistical methods now that we have a corpus of reference summaries that *NLP Technologies* has produced over the years. Even though all summaries have to be validated by human reviewers, this process takes less than 15 minutes per decision. We expect the review process to be even faster now that we are implementing the specialized review process interface shown in Figure 8.

8 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 judgements (between 50 and 100 each week) written in English or French in the immigration domain.

In this article, we have presented our recent work with respect to extending the applicability of the system to French and to other domains such as tax and intellectual property law. The main idea was to elaborate on an information management platform to organize the linguistic cues and semantic rules to achieve a precise information extraction in different fields. 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.

In order for a client to work in the language they are most comfortable with, the RALI and *NLP Technologies* have developed a bidirectional French and English statistical machine translation (SMT) engine for judgements [3]. The SMT output sentences are reviewed before publication, similar to the process used by *NLP Technologies* for summaries.

As the summaries are extracts of the original judgement, we are also developing an interface to keep track of revisions (removal of selected sentences by the system or adding of new sentences) done on the summaries so that the corresponding translated sentences now form the summary in the other official language of Canada.

NLP Technologies is currently studying the possibility of extending the system to other courts and countries. US courts are particularly targeted because of the number of decisions and the proximity to Canada, but they are quite challenging because of different source formats and a different legal system.

Acknowledgments

We thank the CRIM-Precarn Alliance program and National Research Council Canada - Industrial Research Assistance Program (NRC-IRAP) for partially funding this work. We acknowledge the collaboration from the Federal Courts and their feedback. We sincerely thank our lawyers Pia Zambelli and Diane Doray. The authors also thank Fabrizio Gotti, Mehdi Yousfi-Monod, Farzaneh Kazemi and Jimmy Collin for technical support and Elliott Macklovitch for many fruitful discussions.

References

1. Farzindar, A., Lapalme, G.: LetSUM, an automatic Legal Text Summarizing System. In: Gordon, T.F. (ed.) Legal Knowledge and Information Systems. Jurix 2004: the Sevententh Annual Conference, pp. 11–18. IOS Press, Berlin (2004)
2. Farzindar, A.: Résumé automatique de textes juridiques. Ph.D. Thesis, Université de Montréal and Université de Paris IV-Sorbonne (2005)
3. Gotti, F., Farzindar, A., Lapalme, G., Macklovitch, E.: Automatic translation of court judgements. In: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai, pp. 370–379 (2008)
4. Grover, C., Hachey, B., Korycinski, C.: Summarising legal texts: Sentential tense and argumentative roles. In: Radev, D., Teufel, S. (eds.) HLT-NAACL 2003 Workshop: Text Summarization (DUC 2003), Edmonton, Alberta, Canada, pp. 33–40 (2003)
5. Mailhot, L.: Decisions: a handbook for judicial writing, Editions Yvon Blais, Québec, Canada (1998)
6. Moens, M.F.: Summarizing court decisions. Information Processing and Management 43, 1748–1764 (2007)
7. Moens, M.F., Uyttendaele, C., Dumortier, J.: 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), 151–161 (1999)
8. Silberztein, M.D.: Dictionnaires électroniques et analyse automatique de textes. Le système INTEX, Paris, Masson (1993)
9. Smith, J.C., Deedman, C.: The application of expert systems technology to case-based law. In: Proceedings of ICAIL 1987, pp. 84–93 (1987)