Supervised Machine Learning for Summarizing Legal Documents

Mehdi Yousfi-Monod[1], Atefeh Farzindar[2], and Guy Lapalme[1]

[1] Université de Montréal, Laboratoire RALI
RALI-DIRO Université de Montréal, C.P. 6128, succursale Centre-ville,
Montréal, Québec, Canada H3C 3J7
{yousfim, lapalme}@iro.umontreal.ca
[2] *NLP Technologies Inc.*
1255 University Street, suite 1212, Montréal, Québec, Canada, H3B 3W9
farzindar@nlptechnologies.ca

Abstract. This paper presents a supervised machine learning approach for summarizing legal documents. A commercial system for the analysis and summarization of legal documents provided us with a corpus of almost 4,000 text and extract pairs for our machine learning experiments. That corpus was pre-processed to identify the selected source sentences in extracts from which we generated legal structured data. We finally describe our sentence classification experiments relying on a Naive Bayes classifier using a set of surface, emphasis, and content features.

1 Introduction

Legal information is produced in large quantities and needs to be adequately classified in order to be reliably accessible. In Canada, federal and provincial courts produce around 200,000 decisions each year [1]. Classifying these documents is usually performed by legal experts and requires accuracy and speed. These legal experts often summarize decisions and look for information relevant to specific cases in these summaries. The high quality required for these summaries cannot be achieved by commonly available automatic summarization methods as was shown by Farzindar [2] who compared different summarization methods whose results were evaluated by legal experts. Using these results, *NLP Technologies Inc.* has developed a summarization system, named *DecisionExpress*^{TM}, based on a thematic segmentation of the text, specifically tailored to the legal field. Chieze *et al.* [3] detail the automatic summarization system as well as other legal information services offered by the company. As far as we now, there has been no other work dealing with the large scale and domain specific summarization of documents produced by Canadian federal courts.

DecisionExpress^{TM} relies on a symbolic approach based on a set of linguistic rules developed after a meticulous manual analysis of legal documents. The summaries are produced by extraction of whole sentences, often whole paragraphs, rather than by abstraction (rewriting). The reason is that an abstract may be less

accurate and less credible because it is not a direct citation of the decision; reformulation may lead to misinterpretation of the judge's intent. Extracts guarantee that the summary contains only original sentences that can be cited verbatim without having to refer to the original decision. This symbolic summarization approach was developed when no text and extract pairs corpus was available for supervised machine learning. Between June 2008 and June 2009, more than 4,000 decisions have been analyzed and summarized, providing us with a significant and valuable corpus. Unfortunately, the format of the extracts could not be used directly for supervised learning, so documents had to be pre-processed.

In the following section, the summarization process of *DecisionExpress*^{TM} is presented. In Section 3, we report our work on creating the corpus. Our experiments on a model for supervised learning, using the previously generated data, are described and discussed in Section 4. We conclude by introducing new perspectives.

2 Producing the Summary of a Legal Decision

Decisions are available on the Canada courts' websites in HTML.[3] which are analyzed in *DecisionExpress*^{TM} to produce an analytic sheet for each decision containing information extracted from the decision such as the decision's headline and conclusion, the judge's name, the court level and the addressed topics.

Most of the analysis relies on text content rather than the HTML document structure. Since HTML tags define the appearance of the decisions, rather than their structure, and since the presentation as well as its HTML definition is subject to change over time (and it has), we cannot rely on these tags alone to identify the structure of the decisions. Nevertheless, there are cases where text content is not enough and we rely on HTML emphasis to extract some structural elements, as explained below. Linguistic cues, text segments matched by a context-free grammar, are used to identify the decision structure as well as relevant factual information. The output of this analysis is saved using an XML data structure. The division into structural elements relies on a specific knowledge of the legal field [4, 5] and defines 4 decision sections or themes: **Introduction**, **Context**, **Reasoning** and **Conclusion**.

Exploiting text content allows the identification of most structural elements of decisions. Unfortunately, subsection titles seldom match regular lexical patterns, so the lexicon cannot be reliably used to identify such subtitles. It is important to locate these structural elements in order to improve the quality of automatic summaries and human reviews. Indeed, if subsections are not identified, their title and content are merged with the previous one, hence losing their legitimate and required salience. Fortunately, in our case there is a quite reliable clue: HTML emphasis. From a manual analysis of a sample HTML document, supported by legal advice, we defined a conditional rule covering most subtitles

[3] For example, `http://decisions.fct-cf.gc.ca/en/2009/2009fc1188/2009fc1188.html` presents a decision of the Federal Court of Canada, in English, on 19 November 2009.

emphasis: subtitles are generally sentences in either bold, underline or italic and not indented. We then identified a list of HTML elements and CSS attributes defining such emphases, and a specific XML attribute was added to matching sentences. The automatic summarization process relies on attributes added by the transducers and stylesheets to the XML tags of the sentences. The process uses a set of rules that match syntactic patterns relying on part-of-speech information and specific lexicon to define salient sentences for each decision's theme. The process consists in keeping a percentage of salient sentences for each theme.

Until May 2009, *DecisionExpress*^{TM} relied on a plain text reviewing system with which lawyers revised automatic extracts by cut-and-paste operations. The extracts were saved as plain text into the database. This is why we have to process such data before being able to use it for supervised learning, as described in the next section. The reviewing task now benefits from an interactive graphical Web interface, named RevSum, in which lawyers insert or remove whole sentences and paragraphs from the summary by simply clicking on them. This new interface saves the summary into the XML document, by adding attributes to selected sentence tags. Hence the XML structure is preserved during the whole process and these texts can be used directly in our learning algorithm experiments.

3 Building the corpus

DecisionExpress^{TM} relies on a symbolic approach based on linguistic rules according to a meticulous manual analysis of the legal documents, helped by legal experts (lawyers). Between June 2008 and May 2009, more than 4,000 decisions have been created and revised, providing us with a significant and valuable corpus. Unfortunately, the format of the extracts could not be used directly for machine learning. In order to train a categorization model, we need to know which sentences of the source documents were selected for the extracts. So the first step is to identify the source sentence of each extract sentence of the corpus.

Daniel Marcu [6] tackled a related problem in which summaries were abstracts, not extracts. His heuristic consists in removing clauses from the text until the resulting extract is similar enough to the abstract. Our case is simpler and different because sentences are expected to be at least similar so we decided to develop our own method.

Source documents being in XML, each sentence is delimited by an <S> tag. Plain text extract are split into four sections related to the legal themes described above. This difference gave rise to the following issues:

Sentence boundaries detection. The HTML <p> tags around many sentences in source documents that eased the parsing of sentences were not kept in the extracts; therefore, we had to rely on punctuation to deal with abbreviations (e.g. "Mr.") and sentences without end punctuation (e.g. bulleted lists). Fortunately, sometimes there are reliable markers showing the beginning of a sentence, namely paragraph numbers (e.g. "[25]").

Sentence alterations. When generating the summaries, legal experts may have modified sentences, even though, in principle, it was forbidden. The reason is that sometimes it is convenient to remove an unnecessary part of a sentence in a summary[4] or to merge two short sentences. So when matching sentences, we have to look for part of sentences and decide whether it is relevant to identify sentences that have been shortened. We also found cases in which sentences or parts of them have been rewritten. This may cause misspelled words, case changes and even translations. Thus, our similarity function has to be tolerant to slight modifications.

Sentence reordering. Within each of the four themes, we expected that sentences match with the order of the source sentences, however we have found several cases in which the legal experts had reordered the sentences. So we cannot always rely on the order of the source sentences for matching.

Sentence similarities. In the legal field, it is common to see repetition (phrases, clauses or even whole sentences) within a document. Therefore, when identifying sentences, selecting the first text and extract pair that matches may not be the best heuristic. We also have to exploit other clues like ordering.

To deal with those issues, we went through several attempts to identify target sentences while we gradually discovered problematic cases. For each summary, we first determined the sentence boundaries within sections and then matched sentences from the abstract with the ones from the summary using the following three-steps procedure performed iteratively over the four themes:

1. As different sentences may have string similarities and may even include one another, we decided to reduce the risk of wrong matches by trying to first identify the longest sentences. We loop over target sentences, from the longest to the shortest, and for each one, we do a string comparison with each source sentence, also in a decreasing length order. We use the Levenstein edit distance algorithm to allow light modifications (set at 10% of character difference); this level of variation is also allowed in the next steps. We stop when we reach short sentences (less than 50 characters), which are processed in the next step.
2. The shorter the sentences, the greater the chances that they may be similar to others. Some may even be included into longer ones. To reduce risks of wrongful identification, we decided to partially rely on sentence order by trusting the matched sentences in the first step. Sentences are now sorted and processed according to their original order. Identification of the remaining sentences is done only within intervals defined by the previously identified sentences.
3. As some truncated or merged sentences may remain, we try to identify sentence inclusions. We match a summary sentence containing a source sentence

[4] For example, the reference after the colon in "This is a question of mixed fact and law, to be reviewed on a standard of reasonableness: Elezi v. Canada (Minister of Citizenship and Immigration), 2007 FC 240."

Domain	IMM	TAX	IP	Total
English	1 765	447	176	2 388
French	1 155	164	8	1 327
Total	2 820	611	184	3 715

Table 1. Corpus distribution over language and fields: immigration (IMM), tax (TAX) and intellectual property (IP).

if the former's length is within 50% and 150% of the latter. Outside this interval, we consider the sentence would bring noise to the summary, as its extra content was not selected by the experts.

Once all sentences of the corpus have been processed, we have a set of XML documents in which each sentence is either tagged as kept in one of the four sections of the summary or not tagged. For some summaries, a significant proportion of the sentences were not identified. This generally happens when sentences are rewritten by lawyers, usually translated. As such documents may bias the training process because unidentified sentences will be considered as negative examples, we decided to remove the documents in which less than 70% of sentences in the summary were matched. From 4,067 documents, we removed 352 and our final corpus is then composed of 3,715 documents, where 94% of the sentences and 93% of the words have been identified. Since we allowed a small editing distance, there are 1.9% of character insertions, deletions or substitutions among identified sentences. Table 1 presents the distribution of English and French decisions in three fields.

4 Machine Learning Experiments

4.1 Categories and Features

Our extract-based summarizer has to classify sentences as being in the summary or not, and our extracts are placed into four sections. We thus have a total of five categories: *not in summary (NIS)*, *Introduction*, *Context*, *Reasoning* and *Conclusion*. Table 2 presents the distribution of the instances of our corpus over all five categories. Depending on the legal field, documents have notable structure dissimilarities as well as differences in the summarization method used by legal experts. We therefore decided to train our models on a single field at a time. As some training features rely on vocabulary, we also decided to deal with one language at a time. In this paper we detail our experiments with a corpus of English decisions from the immigration field as it is the largest sub-corpus we have: 1,765 documents with 65,345 instances. We will describe some results in other fields and languages in Section 4.4.

For the learning process, we split the instances of the corpus into 2/3 for the training set and 1/3 for the test set. The classifier used is the Weka [7] implementation of the popular Naive Bayes, with the supervised discrimination

Classes	NIS	Introduction	Context	Reasoning	Conclusion	Total
# Instances	142 277	3 462	18 941	28 693	2 663	196 036
% Instances	72.6%	1.8%	9.7%	14.6%	1.4%	100%

Table 2. Distribution of instances over categories in the corpus.

option enabled. We also ran the classification with other bayesian-like and support vector machine classifiers as well as some based on tree decision algorithms but they did not yield better results. We then explored the relevance of several features for our categorization task:

Surface features Such common features exploit the decision structure of the source document: sentence position in a paragraph, paragraph position in a section, section position, sentence length of words and number of sentences in a paragraph.

Emphasis features As the analysis of the HTML source decisions preserves part of the emphasis in some sentences and as emphasis is closely related to salience, and thus to relevance, we decided to assess the usefulness of such information. Emphasis features are bold, underline, italic and indent, and take a boolean value.

Content features We tested 2 features relying on the vocabulary of the decisions. The first uses the sum of each word's $tf \cdot idf$ score, the result is normalized with the sentence length (in words). The second relies on the legal genre where there are specific words regularly used to express an opinion or declare a fact, which are in sentences generally relevant for the summary. Examples of such words are "apparently", "dismissed", "daughter" or "kill". Over the 1,765 documents of the training corpus, the word "dismissed" appears 1,623 times in all the extracts and 1,582 times in other sentences, while most words usually appear at least 2 to 3 times more in sentences not kept for the summaries. Other instances include the words "paragraphs", "relies" or "procedure". This led us to add such a score, based on the ratio of how many times a term appears in the extract sentences, to how many times it appears in other sentences. This score for a sentence S is the normalized sum of such a ratio of each word:

$$\frac{\Sigma_{w \in S}(\frac{tf_{se}(w)}{tf_{sne}(w)})^2}{length(S)} \quad (1)$$

$tf_{se}(w)$ represents the number of times the word w appears in the corpus in sentences selected for the extracts and $tf_{nse}(w)$ how many times it appears in sentences not selected for the extracts. The power applied to the ratio of frequencies helps discriminating words specific to extracts from others.

	Features	Intro.	Cont.	Reas.	Concl.	Summary
Precision	Sur	0.644	**0.523**	0.371	0.380	**0.466**
	Sur+Em	0.645	0.490	0.360	0.377	0.438
	Sur+Voc	**0.651**	0.505	**0.390**	**0.389**	0.458
	Sur+Em+Voc	0.649	0.492	0.388	0.387	0.448
Recall	Sur	0.789	0.499	0.190	0.621	0.360
	Sur+Em	0.795	0.595	0.291	0.617	0.447
	Sur+Voc	0.804	0.715	0.356	0.627	0.525
	Sur+Em+Voc	**0.809**	**0.741**	**0.432**	**0.630**	**0.574**
F_1-Measure	Sur	0.709	0.511	0.251	0.471	0.406
	Sur+Em	0.712	0.537	0.322	0.468	0.443
	Sur+Voc	0.719	**0.592**	0.372	**0.480**	0.489
	Sur+Em+Voc	**0.720**	**0.592**	**0.409**	**0.480**	**0.503**

Table 3. Classification Precision, Recall and F_1-Measure based on different feature groups, for each summary sections plus the whole summary of the English immigration corpus. Features are (Sur)face, (Em)phasis and (Con)tent.

4.2 Classification Results and Discussion

We tried different groups of features and the most relevant are shown in Table 3. All features have a positive impact on the classification when considering the F_1-Measure of the whole summary. Our best overall results are obtained by the use of all feature groups, we name this configuration ProdSum (PRobabilistic Decision SUMmarizer).

The emphasis features have no significant impact on introduction and conclusion categories because sentences in these sections are seldom emphasized. Surface cues greatly help for these two categories and are even enough to achieve our best results, which means that such sections are composed of sentences extracted from constant parts of the decisions. Introduction gets the highest score for both precision (0.649) and recall (0.809) because the most relevant text content of this legal theme is often made of sentences from the first paragraph of the decision. Surprisingly, regarding the context section, adding other features not only increases noise, but also increases recall. Context and conclusion sections reach an acceptable recall, respectively 0.741 and 0.630, mostly through surface features.

The reasoning section, which is usually the longest of the decision, needs other features than the surface ones in order to get an adequate recall; relevant sentences of this section do not solely depend on their position, so we need content and emphasis information to evaluate their relevance.

4.3 Comparison with a baseline and ASLI

In order to assess the performance of our classification we had to defined two baselines. Our first baseline, adapted from the *start-end* baseline of Farzindar [2], constructs an extract from the first N sentences of each section from the source

	System	Intro.	Cont.	Reas.	Concl.	Summary
Precision	Baseline	0.626	0.449	0.182	0.245	0.332
	ASLI	**0.699**	0.390	0.291	0.339	0.362
	ProdSum	0.649	**0.492**	**0.388**	**0.387**	**0.448**
Recall	Baseline	0.544	0.544	0.131	0.470	0.319
	ASLI	**0.878**	0.690	0.330	**0.666**	0.509
	ProdSum	0.809	**0.741**	**0.432**	0.630	**0.574**
F_1-Measure	Baseline	0.582	0.492	0.152	0.322	0.325
	ASLI	**0.778**	0.498	0.309	0.450	0.423
	ProdSum	0.720	**0.592**	**0.409**	**0.480**	**0.503**

Table 4. Comparison of ProdSum to a baseline and ASLI for the English immigration corpus.

document. While Farzindar retrieved 15% of the source document (12% from the start and 3% from the end) to create the baseline, we relied on the actual compression ratio of our training corpus, shown in Table 2, which amounts to an average of 27.5%. The extracts of our first baseline were composed of the extraction of the first 1.8% words of the decision's introduction, the first 9.7% of the context, and so on. The last sentence was added in full if it was to be cut by the percentage. It turned out that this compression rate worked better than the original 15%. Our second baseline is the current automatic summarization system of *DecisionExpress*^{TM}: ASLI. Scores of the latter baseline may be biased as sentences of our corpus have been first selected by ASLI's algorithm. However, the review process, achieved by legal experts, did alter that sentence selection and thus will reduce the bias.

Table 4 provides classification scores for the three systems tested: a baseline, ASLI and ProdSum.

The baseline, while not as efficient as other systems, still managed to get satisfactory scores for the introduction and context sections because most of the relevant information in these sections is found at the beginning. ASLI has a slight advantage over ProdSum when dealing with the introduction, but is generally outperformed for other sections, specifically with respect to the whole summary.

4.4 Results for other fields and language

Table 5 shows the results (F_1-Measure scores) of ProdSum compared to ASLI for two fields, immigration and tax, and two languages, English and French.

We do not provide results for the intellectual property field as there is not enough training data to yield relevant scores as of yet. The experiment on the English tax field resulted in a F_1-Measure score of 0.445, which is a bit lower than the 0.503 score of the immigration field but far greater than the corresponding score obtained by ASLI (0.190). ASLI obtained a low score because it selected too many sentences for the introduction, leading to a 0.058 precision score for

Language	Domain	# Documents	# Instances	ASLI	ProdSum
English	**IMM**	1765	65345	0.423	**0.503**
	TAX	447	21517	0.190	**0.445**
French	**IMM**	1155	40293	0.433	**0.483**
	TAX	164	8380	0.344	**0.368**

Table 5. F_1-Measure scores of ASLI and ProdSum for English and French languages and immigration and tax fields.

that section. It usually works well for the immigration field, but introductions in the tax field cover a large part of the decisions, so most of them should be removed to produce the extract. The underlying reason for that low score is the specialization of the symbolic approach to a specific field. Indeed, ASLI has been developed more specifically for immigration documents, which are more numerous than other fields, while ProdSum, as a statistic approach, adapts better to new fields. The French corpus also works well with ProdSum which yielded, for the immigration field, a F_1-Measure score of 0.483, proving that ProdSum is also suitable for French decisions. The small French tax corpus got a score of 0.368, which is comparable to the English version, but may not be relevant because of the low amount of available training data. Finally, ProdSum obtained the best overall results, notwithstanding the field and language.

4.5 Rouge evaluation

The Rouge metric is typically used to compare automatic extracts with human abstracts. While our reference summaries are extracts, misclassified sentences may contain relevant content which may be captured by Rouge measures. Rouge does not have a default configuration for the French language, and as we used the stemmer and stop word options, we only did runs for the English language, in the immigration and tax fields. Our main goal was to assess performances of our summarizer for each legal theme. Therefore, we evaluated each section separately. We were also curious to know if the full automatic summaries matched well with the references, so we did a run with the full summaries, i.e. extracts with the four sections. All runs used Rouge's configuration[5] of the DUC[6] 2007 conference, which yielded Rouge-{1, 2 and SU4} scores. Table 6 shows F_1-Measure Rouge scores of our experiment. We give F_1-Measure scores instead of the recall ones as we had almost no control on the size of our system and reference summaries, thus noise has to be taken into account. Scores are higher than those which traditional summarizers usually perform because we are comparing our extracts with other extracts made of sentences of the same

[5] Rouge version 1.5.5, with the following command line options: `-n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a` (use Porter stemmer on both models and peers, use 95% confidence interval, bootstrap resample 1000 times to estimate these 95%, compute F-measure with alpha = 0.5).

[6] `http://duc.nist.gov/`

Domain	Measure	System	Intro.	Cont.	Reas.	Concl.	Full
IMM	ROUGE-2	Baseline	0.692	0.496	0.173	0.414	0.435
		ASLI	**0.778**	0.499	0.270	**0.494**	0.577
		PRODSUM	0.768	**0.554**	**0.369**	0.479	**0.633**
	ROUGE-SU4	Baseline	0.690	0.503	0.192	0.402	0.452
		ASLI	**0.779**	0.503	0.279	**0.479**	0.590
		PRODSUM	0.766	**0.556**	**0.376**	0.461	**0.640**
TAX	ROUGE-2	Baseline	0.473	0.155	0.090	**0.414**	0.278
		ASLI	0.257	0.147	0.123	0.396	0.598
		PRODSUM	**0.507**	**0.403**	**0.445**	0.402	**0.661**
	ROUGE-SU4	Baseline	0.473	0.162	0.097	**0.408**	0.289
		ASLI	0.256	0.152	0.125	0.387	0.604
		PRODSUM	**0.507**	**0.414**	**0.453**	0.393	**0.667**

Table 6. F_1-Measure ROUGE scores of all systems for English immigration and tax documents.

source documents, not abstracts. ROUGE-1 scores are very similar to other ones so we do not display them.

PRODSUM gets the overall best results. When dealing with the smallest sections – introduction and conclusion – ASLI gets slightly better scores for the immigration field, due to better precision. The reason is that the symbolic method has rules to detect and exclude citations of the decision, which are not relevant to the summary, whereas our system does not have any feature dealing specifically with such cases. The low classification scores ASLI got for the context and reasoning sections of the tax field are confirmed by ROUGE scores. The baseline gets the best conclusion scores for the tax field because it selects few sentences, reducing noise thus increasing precision, which is favored by the $F_{alpha=0.5}$-Measure score. Finally, full summaries matched best with PRODSUM extracts, notwithstanding the field and measure. It is interesting to note the full summaries scores are globally higher than section scores, regardless of the system, thus indicating that some sentences were wrongly classified in a summary theme, while they actually belonged to another one.

The scores differences are not incidental according to significance tests we did on each legal theme and for full summaries. We used the standard paired t-test on PRODSUM and ASLI, for ROUGE-SU4 per evaluation score results, on both legal fields. While result differences of the small introduction and conclusion sections proved to be incidental ($0.3 < p - value < 0.9$), PRODSUM's results on context and reasoning sections, as well as full summaries, are significantly better ($p - value < 0.0001$) than ASLI's.

5 Related work

There have been a few other approaches dealing with automatic summarization of legal documents, and the best source for an overview of such works is certainly [8], where the author 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 field.

FLEXICON [9] is one of the first summarization system specialized for legal texts, it was a symbolic approach 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 [10], developed for summarization of Belgian criminal cases, was the first to explicitly make use of the structure of a case. The system first identifies the discourse structure with text grammars a process similar to the one used in the first phase of *DecisionExpress*^{TM}. The next step produces the summary by selecting relevant paragraphs of each important document section. Paragraphs are represented as vectors of index terms and are grouped by a clustering algorithm. This process aims to removing redundant information and grouping paragraphs into thematically coherent units. SALOMON's approach uses shallow information as ProdSum do but does not rely on machine learning.

Hachey and Grover [11] present an approach closely related to ours. They exploit a corpus of 188 decisions of the House of Lords they have gathered and annotated. Sentences are tagged with rhetorical status, relevance and linguistic information. The authors performed sentence classification experiments with Naive Bayes and maximum entropy models, using shallow information and named entities as features. They only provide prediction scores for individual features, and the best one, F-Score of 31.2, goes to the "thematic words" feature for the Naive Bayes classifier. This feature is a basic $tf \cdot idf$ score, similar to ours. The authors have not performed any manual or automatic content-based evaluation.

6 Conclusion

In this paper we introduced an approach for selecting important sentences from legal documents using supervised machine learning. We first described a system for legal document analysis and summarization which is provided with a valuable and significant corpus of text and extract pairs. That corpus was processed to identify the source sentences contained in the plain text extracts. We also presented our work on generating an XML structured data, dealing with issues specific to the legal field. The machine learning step consisted in running a sentence classification algorithm, Naive Bayes, based on a set of surface, emphasis and content features. Our system, ProdSum, has been compared with a baseline system and with ASLI, the current automatic summarization system of *DecisionExpress*^{TM} (before revision). While ASLI may compete with ProdSum

on one or two of the smallest legal themes, our system obtained the best overall results.

While we only used rather standard features, it turned out to be enough to beat the symbolic method. To reach better classification scores, particularly for the context and reasoning legal themes, we plan to explore features based on events and factual information as it is the purpose of such sections to gather temporal and factual evidence in order to support the verdict.

Acknowledgment

We thank the Precarn for partially funding this work. We sincerely thank our lawyers Pia Zambelli and Diane Doray. The authors also thank Fabrizio Gotti and Farnaz Shariat for technical support, Vincenzo Mignacca for his meticulous proofreading of the final version of this paper, and reviewers for their comments and suggestions.

References

1. Plamondon, L., Lapalme, G., Pelletier, F.: Anonymisation de décisions de justice. XIe Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004) (May 2004) 367–376
2. Farzindar, A.: Résumé automatique de textes juridiques. PhD thesis, Université de Montréal et Université Paris IV-Sorbonne (Mar 2005)
3. Chieze, E., Farzindar, A., Lapalme, G.: An automatic system for summarization and information extraction of legal information. In: accepted in "Semantic Processing of Legal Texts". Springer (2009) 1–20
4. 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. IOS Press, Berlin (Dec 2004) 11–18
5. Farzindar, A., Lapalme, G.: Production automatique du résumé de textes juridiques: évaluation de qualité et d'acceptabilité. In: TALN 2005. Volume 1., Dourdan, France (Jun 2005) 183–192
6. Marcu, D.: The automatic construction of large-scale corpora for summarization research. In: University of California, Berkely. (1999) 137–144
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1 (2009)
8. Moens, M.F.: Summarizing court decisions. Inf. Process. Manage. **43**(6) (2007) 1748–1764
9. Smith, J., Deedman, C.: The application of expert systems technology to case-based law. In: Proceedings of the First International Conference on Artificial Intelligence and Law (Boston, Mass) The Center for Law and Computer Science, Northeastern University. (1987) 84–93
10. Moens, M.F., Uyttendaele, C., Dumortier, J.: Abstracting of legal cases: the potential of clustering based on the selection of representative objects. J. Am. Soc. Inf. Sci. **50**(2) (1999) 151–161
11. Hachey, B., Grover, C.: Extractive summarisation of legal texts. Artif. Intell. Law **14**(4) (2006) 305–345