

The Use of Thematic structure and Concept Identification for Legal Text Summarization

Atefeh Farzindar and Guy Lapalme

RALI, Département d'Informatique et recherche opérationnelle
Université de Montréal, Québec, Canada, H3C 3J7
{farzinda,lapalme}@iro.umontreal.ca

Abstract

LetSum is a summarization system developed for producing short summaries for legal decisions. LetSum is built with an approach based on the exploration of the document structure and thematic segmentation in order to produce a table-style summary for improving coherency and readability of the text. We present the components of the system and its implementation.

1 Introduction

As ever larger amounts of legal documents become available electronically, interest in automatic summarization has continued to grow in recent years. In this paper, we describe our method of text summarization which produces short summaries for legal documents. We have attempted to approximate the results of human expert reasoning by organizing and manipulating factual and heuristic knowledge.

We present our approach to summarize the legal record of the proceedings of federal courts in Canada and presenting it as a table-style summary for the needs of lawyers and experts in the legal domain. Our method investigates the extraction of the most important units based on the identification of thematic structure of the document and the determination of argumentative themes of the textual units in the judgment (Farzindar et al., 2004).

In Canada, the Canadian Legal Information Institute (CANLII) aims at gathering legislative and judicial texts in order to make a virtual library of Canadian law accessible for free on the Internet (<http://www.canlii.org>). The large volume of legal information in electronic form creates a need for the creation and production of powerful computational tools in order to extract relevant information in a condensed form.

The lawyers need to process the previous legal decisions to find a solution to a legal problem not directly indicated in the law, they look

for *precedents* of similar cases. Each decision contains the reasons which justify the solution for a legal problem. They constitute a *law jurisprudence precedent* from which it is possible to extract a legal rule that can be applied to similar cases.

One reason for the difficulty of the work in legal field is the complexity of the domain: specific terminology of the legal domain and legal interpretations of expressions produce many ambiguities. For example, the word *disposition* means nature, effort, mental attitude or property in general English but in legal terms it means the final part of a judgement indicating the nature of a decision: acceptance of an inquiry or dismissal. That is why we have constructed our conceptual dictionary containing 200 concepts of legal domain.

Thematic segments give the information which can be used to answer specific questions about the theme of the segment such as: what is theme? How a theme is used? How are the problem and the facts presented? How does a judge reason to reach a conclusion?

In this paper, we will describe how we dealt with the problem of the exploration of structure of document and content selection, according to the themes of a judgement.

1.1 Legal text summarization

Our approach to produce the summary is based on the identification of the thematic structure to find the argumentative theme of the judgement. This approach is a result of our corpus analysis in which we compared model summaries written by humans with the texts of the original judgments. The textual units considered as important by the professional abstractors were aligned manually with one or more elements of the source text. We look for a match between the information considered important in the professional abstract and the information in the source documents.

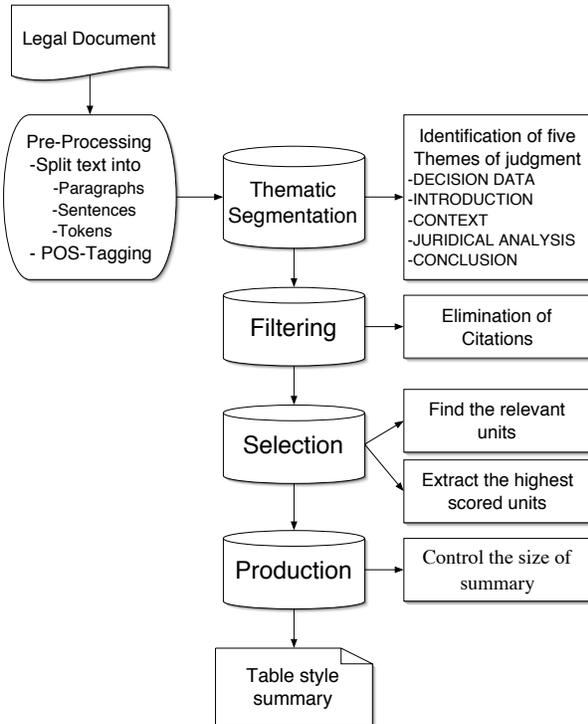


Figure 1: The procedural steps for generating of table-style summary

We extract the relevant sentences for each theme and present them as a table-style summary. Showing the information considered important could help the user read and navigate easily between the summary and the source judgment. If a sentence seems more important for a user and more information is needed about this topic, the complete thematic segment containing the selected sentence could be presented.

The identification of these structures separates the key ideas from the details of a judgment and improves readability and coherency in the summary. Therefore, in the presentation of a final summary, we propose to preserve this organization of the structure of the text in order to build a table-style summary.

Our corpus contains 3500 judgments of the Federal Court of Canada, which are available in HTML on <http://www.canlii.org/ca/cas/fct/>. We analyzed manually 50 judgments in English as well as their summaries written by professional legal abstractors. These judgments were suggested by the lawyers of CanLII project, as representing the *standard* judgements with *ideal* summaries. The rest of corpus is used for statistical computations.

2 Components of LetSum

To process an input decision to LetSum, the system first does some pre-processing. The summary is built in four phases (Figure 1): thematic segmentation, filtering of less important units such as citations of law articles, selection of relevant textual units and production of the summary within the size limit of the abstract (Farzindar and Lapalme, 2004).

Pre-Processing splits the input judgment into main units. First the body of the text of decision are identified. Some keywords like *Reasons for order*, *Reasons for judgment and order* separate the basic data (date, name of court, etc.), placed in the head of document, from the beginning of the judgment. The features used for the end of the decision are the date and place of hearing, name and signature of the judge. Then the document is divided into: section titles, paragraphs, sentences and tokens. To determine the Part-of-Speech tags, the tagger described by (Hepple, 2000) is used.

Thematic segmentation is based on the specific knowledge of the legal field. According to our analysis, the texts of jurisprudence have a thematic structure, independently of the category of judgment. 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 experimental work of judge Mailhot (Mailhot, 1998):

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

CONTEXT explains the facts in chronological order, or by description. It recomposes the story from the facts and events between the parties and findings of credibility on the disputed facts.

JURIDICAL ANALYSIS describes the comments of the judge and finding of facts, and the application of the law to the facts as found. For the legal expert this section of judgment is the most important part because it gives a solution to the problem of the parties and leads the judgment 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.

For thematic segmentation the following information are used: the presence of significant section titles, the positions of a segment, the identification of direct or narrative style (as the transition of CONTEXT and JURIDICAL ANALYSIS segments), certain linguistic markers. We present the heuristics and some examples of linguistic markers for each theme:

INTRODUCTION represents a short description on case with the cue phrases: *application for judicial review, application to review a decision, motion filed by, Statement of Claim.*

Position of this segment is at the beginning of the judgement after the title like: *Reasons for order, Reasons for judgment and order.*

Significant section titles are *Introduction* and *Summary.*

CONTEXT introduces the parties (eg. The Applicant is a 52 year old citizen of Sri Lanka).

It describes the application request using the markers: *advise, indicate, concern request* (eg. Motion concerns Air Canada's failure to provide ground services in the French language at the Halifax airport).

It explains the situation in the past tense and narration form.

Section titles are *Facts, Background, Factual background and Agreed statement of facts.*

JURIDICAL ANALYSIS, for this theme the judge gives his explanation on the subject thus the style of expression is direct using *I.*

Cue phrases are *In reviewing the sections No. of the Act, Pursuant to section No., As I have stated, In the present case, The case at bar is.*

Section titles are *Analysis* and *Decision of the court.*

CONCLUSION contains the final result of court decision using phrases such as: *The motion is dismissed, the application must be granted*

section titles are *Conclusion, Costs* and *Disposition.*

Filtering identifies parts of the text which can be eliminated, without losing relevant information for the summary. In a judgment, the

citation units (sentence or paragraph) occupy a large volume in the text, up to 30%, of the judgment, whereas their content is less important for the summary (according to our manual alignments between summaries and sources). This is why we remove citations inside blocks of thematic segments. We thus filter two categories of segments: submissions and arguments that report the points of view of the parties in the litigation and citations related for previous issues or references to applicable legislation. In the case of eliminating a citation of a legislation (eg. law articles), we save the reference of the citation in DECISION DATA in the field of authority and doctrine.

The identification of citations is based on two types of markers: direct and indirect. A direct marker is one of the linguistic indicators that we classified into three classes: verbs, concepts (noun, adverb, adjective) and complementary indications. Examples of verbs of citation are: *conclude, define, indicate, provide, read, reference, refer, say, state, summarize.* Examples of the concepts are: *following, section, subsection, page, paragraph, pursuant.* Complementary indications include numbers, certain preposition, relative clauses and typographic marks (colon, quotation marks). The indirect citations are the neighboring units of a quoted phrase. For example, citation segment in the phrase *paragraph 78(1), which reads as follows:* is identified using direct markers but it points to the textual units with no direct marker which are also quotations. We thus identify the enumerated sentences following a quoted sentence for determining a group of citations.

Selection builds a list of the best candidate units for each structural level of the summary. LetSum computes a score for each sentence in the judgment based on heuristic functions related to the following information: position of the paragraphs in the document, position of the paragraphs in the thematic segment, position of the sentences in the paragraph, distribution of the words in document and corpus ($tf \cdot idf$). Depending on the given information in each layered segment, we have identified some cue words and linguistic markers (Table 1). The thematic segment can change the value of linguistic indicators. For example, the phrase *application is dismissed* that can be considered as an important feature in the CONCLUSION might not have the same value in CONTEXT segment. At the end of this stage, the passages with the high-

Themes	Linguistic Markers
INTRODUCTION	Concepts : <i>decision, motion, application, determination, order, proceeding, against, rule, reason</i> Verbs : <i>raise, strike, determine, describ, declare, date</i> Expressions : <i>application for judicial review, application to review a decision, motion filed by, Statement of Claim</i>
CONTEXT	Concepts : party (<i>appellant, applicant, defendant, plaintiff, respondent</i>) Verbs : <i>advise, indicate, request, claim, allege, concern</i>
JURIDICAL ANALYSIS	Concepts: <i>opinion, conclusion, summary, because, cost, action, view, reason, I, my</i> Verbs : <i>note, accept, summarise, scrutinize, think, say, satisfy, discuss, conclude, find, believe, reach, indicate, persuade, agree, see, review, recommend, answer</i> Expressions : <i>in the case at bar, for all the above reasons, in my view, my review of, in view of the evidence, in the result, finally, thus, consequently</i>
CONCLUSION	Concepts : <i>motion, application, action, party</i> Verbs : <i>allow, deny, dismiss, grant, refuse, award</i> Expressions : <i>for (all) the above reasons, for all of the foregoing reasons</i>

Table 1: The linguistic markers used in selection module that signal the important sentences in a judgment.

est resulting scores are sorted to determine the most relevant ones.

Production of the final summary controls the size of summary and displays the selected sentences in tabular format. The final summary is about 10% of source document. The elimination of the unimportant sentences takes into account length statistics based on our observation from human abstracts. In the INTRODUCTION segment, units with the highest score are kept within 10% of the size of summary. In the CONTEXT segment, the selected units occupy 24% of the summary length. The contribution of the JURIDICAL ANALYSIS segment is 60% and the units with the role CONCLUSION occupy 6% of the summary.

Figure 2 shows an example of a table-style summary generated by LetSum. The summary is 15% of the source judgment.

3 Evaluating LetSum

We evaluate LetSum in two steps; first the evaluation of the modules of the system and second the global evaluation of produced summaries. Preliminary evaluations of components of LetSum are very promising; we obtained 0.90 F-measure for thematic segmentation and 0.97 F-measure for filtering stage (detection of 57 quoted segment correctly on 60). The final global evaluation will include the intrinsic and extrinsic tests (Spark-Jones and Galliers, 1995).

Intrinsic evaluations test the system in of itself and extrinsic evaluation test the system in relation to some other task.

Up to now, for intrinsic evaluations, we conducted a small pilot study on 5 long judgments (average 8 pages) with ROUGE (Lin, 2004). ROUGE is a package for automatic evaluation of summaries, which is a Recall-based measure to automatically determine the quality of a summary by comparing it to ideal summaries created by humans. We have compared the summaries of Microsoft Word, Pertinence (www.pertinence.net/), baselines and LetSum with the reference summaries written by professional abstractors. We have defined two baselines: Baseline-First and Baseline-Last. Baseline-First takes 15% words of the begin of the judgment. According our thematic segmentation, it takes the sentences from the themes INTRODUCTION and CONTEXT. Baseline-Last takes 15% words of the end of the judgment with themes JURIDICAL ANALYSIS and CONCLUSION. Table 2 shows the result of this evaluation. LetSum is the first with the best evaluation scores. We are completing the extrinsic evaluations with a series of the specific questions, which is based on legal expert judgement.

4 Conclusion

In this paper, we have presented our approach to automatic summarization. Our method is

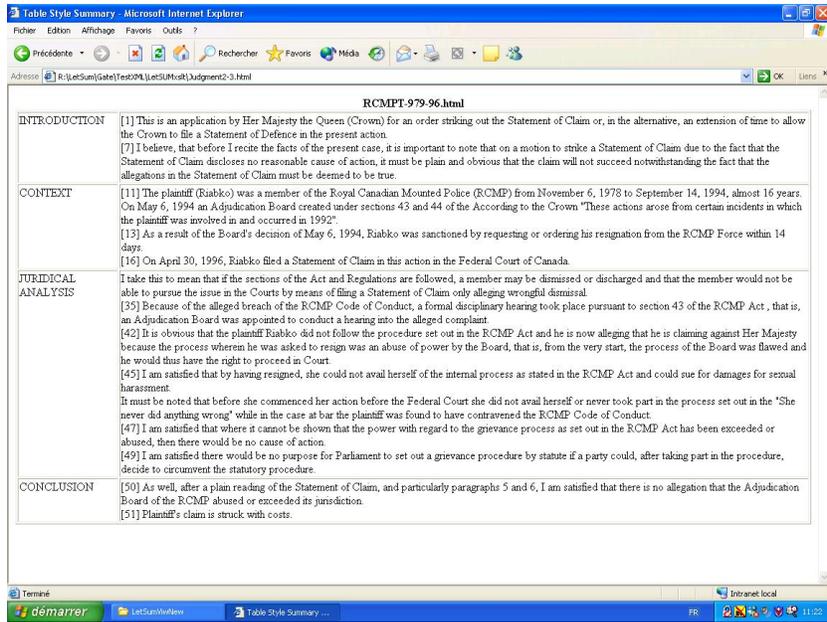


Figure 2: A table-style summary, example of the output of LetSum. The summary is 15% of the source judgment.

SYSID	ROUGE-1	Lower	Upper
LetSum	0.61911	0.45034	0.78788
Baseline-F	0.52660	0.45960	0.59360
MS Word	0.44745	0.39899	0.49591
Baseline-L	0.40052	0.30808	0.49296
Pertinence	0.35602	0.22727	0.48477

Table 2: Result of evaluation with ROUGE, LetSum is the first with the best evaluation scores.

based on the extraction of relevant units in the source judgment by identifying the document’s structure and determining the themes of segments in the decision. The generation of the summary is done in four steps: thematic segmentation to detect the legal document structure in four themes INTRODUCTION, CONTEXT, JURIDICAL ANALYSIS and CONCLUSION, filtering to eliminate unimportant quotations and noises, selection of the candidate units and production of table-style summary. The presentation of the summary is in a tabular form along the themes of the judgment.

References

Atefeh Farzindar and Guy Lapalme. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. In *Text Summarization Branches Out Work-*

shop held in conjunction with ACL’2004, pages 27–34, Barcelona, Spain, 25–26 July.

Atefeh Farzindar, Guy Lapalme, and Jean-Pierre Desclés. 2004. *Résumé de textes juridiques par identification de leur structure thématique. Traitement Automatique des Langues (TAL), Numéro spécial sur: Le résumé automatique de texte : solutions et perspectives*, 45(1):26 pages.

Mark Hepple. 2000. Independence and commitment: Assumptions for rapid training and execution of rule-based part-of-speech taggers. In *the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 278–285, October.

Chin Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out Workshop held in conjunction with ACL’2004*, pages 74–81, Barcelona, Spain, 25–26 July.

Louise Mailhot. 1998. *Decisions, Decisions: a handbook for judicial writing*. Editions Yvon Blais, Québec, Canada.

Karen Spark-Jones and Julia R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.