Supervised Machine Learning for Summarizing Legal Documents

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

[1] Universite de Montreal, Laboratoire RALI
RALI-DIRO University de Montreal, C.P. 6128, succursale Centre-ville,
Montreal, Quebec, Canada H3C 3J7
{yousf im, lapalme}Qiro. umontreal. ca
[2] *NLP Technologies Inc.*
1255 University Street, suite 1212, Montreal, Quebec, Canada, H3B 3W9
farzindarOnlptechnologies.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 *DedsilnEtopnss*™, b i d 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*™ 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**