Supervised Machine Learning for Summarizing Legal Documents

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

[1]RALI-DIRO rali, Université de Montréal
{yousfim, lapalme}@iro.umontreal.ca

[2]*NLP Technologies Inc.*
NLP TECHNOLOGIES
farzindar@nlptechnologies.ca

May 31, 2010

Introduction

- Legal field: Thousands of decisions
  - 200 000/y in Canada
- Multilingual, multiple fields
- Access to specific decisions
  - Classification
- Access to relevant content
  - Summarization

DecisionExpress™: Classification and Summarization

Information	Mepham v. Canada (Citizenship and Immigration) (2009 FC 1188)
Subject: Skilled workers **Conclusion:** allowed **Judge:** Michel Beaudry **Tribunal:** Federal Court **Document Type:** Judgment View Summary	**Headnote:** This is an application for judicial review and Refugee Act, S.C. 2001, c. 27 , by L Officer at the Canadian High Commission Officer rejected her application for perm class. **Topics:** Selection criteria, Educational credentials **Location(s):** UNITED KINGDOM

Introduction	[1] This is an application for judicial review pursuant to subsection 72(1) of the Immigration Act), by Lynn Mepham (the Applicant), of a decision of an Immigration Officer (the Officer) a London, dated December 22, 2008, where the Officer rejected her application for permanent worker class.
Context	[4] The Applicant is a citizen of the United Kingdom and made an application for permanent skilled worker class. [5] The Applicant submitted information pertaining to her education.She listed that she had

Obective of our work

- Current system on *DecisionExpress^{TM}*:
  - ▶ **Symbolic** method
  - ▶ Manual linguistic rules

- Our goal:
  - ▶ Does a **statistical** method can get similar performances?
  - ▶ Will the results be **constant** over legal fields and languages?
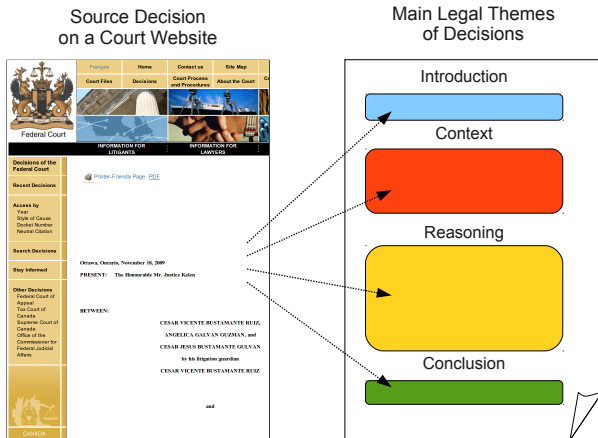
DecisionExpress™'s Architecture

HTML Decision
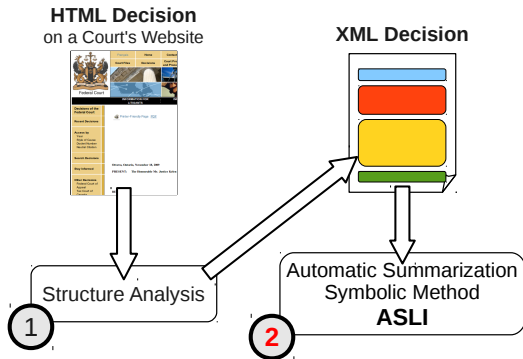on a Court's Website

Structure Analysis
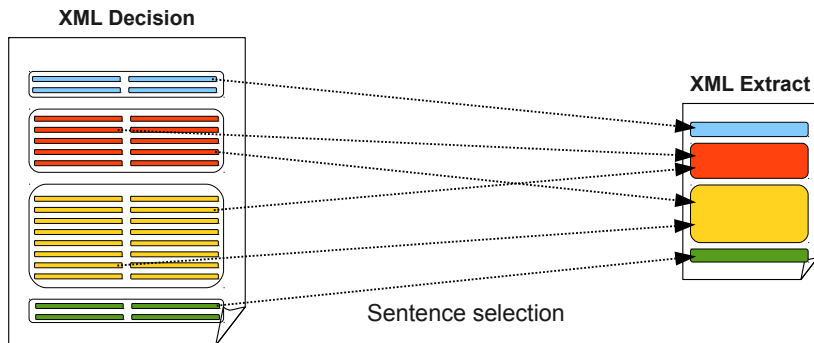
1

Automatic Structure Analysis

Source Decision on a Court Website

Main Legal Themes of Decisions

- Introduction
- Context
- Reasoning
- Conclusion

■ Structure elements: sections (legal themes), paragraphs, sentences

DecisionExpress™'s Architecture

Extracting Sentences
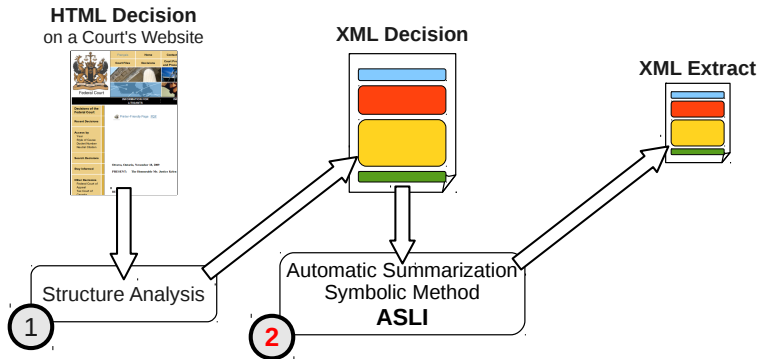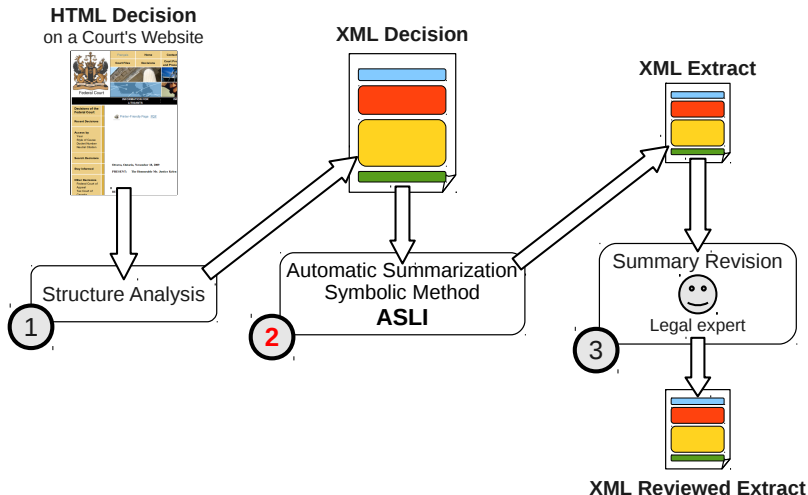

XML Decision

XML Extract

Sentence selection

- Select whole sentences
- Sentences may fall into another section

DecisionExpress™'s Architecture

HTML Decision
on a Court's Website

XML Decision

XML Extract

Structure Analysis
1

Automatic Summarization
Symbolic Method
ASLI
2

DecisionExpress™'s Architecture

HTML Decision
on a Court's Website

XML Decision

XML Extract

Structure Analysis
1

Automatic Summarization
Symbolic Method
ASLI
2

Summary Revision
☺
Legal expert
3

XML Reviewed Extract

rali

DecisionExpress™'s Architecture: Our Objective

HTML Decision
on a Court's Website

XML Decision

XML Extract

Structure Analysis

(1)

ProdSum

Classifier

(2)

(document, extract) pairs

Training Corpus

Summary Revision

Legal expert

(3)

XML Reviewed Extract

Training Corpus
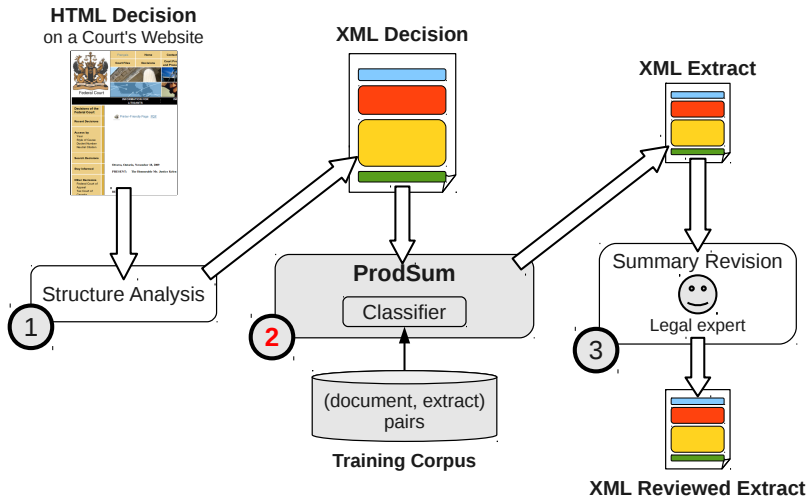
- 1 year of activity on *DecisionExpress^{TM}*
  - $\sim$ 4000 decision and extract pairs
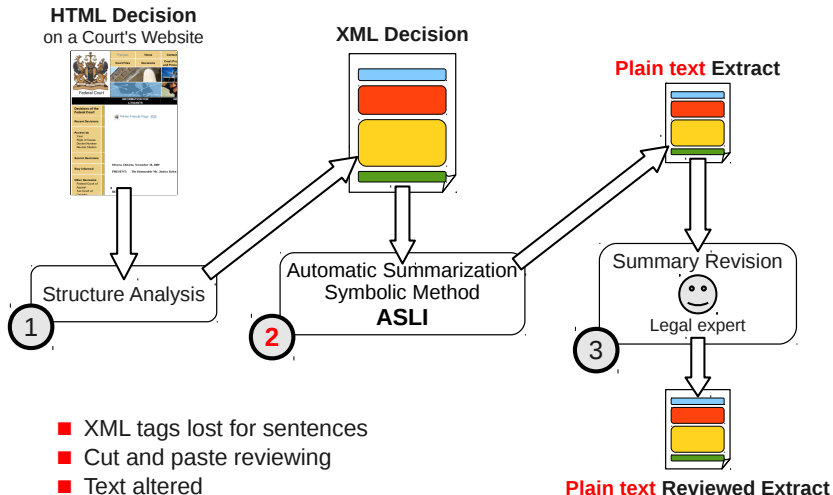
Field	IMM	TAX	IP
English	**1 765**	447	176
French	1 155	164	8

- Issue: Extract's sections were reviewed in plain text

DecisionExpress™'s Architecture
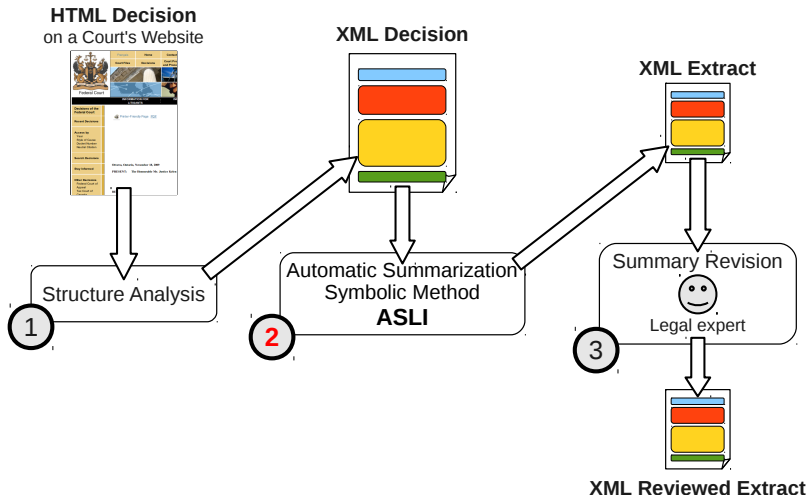
DecisionExpress™'s **Previous** Architecture

HTML Decision
on a Court's Website

XML Decision

Plain text Extract

Structure Analysis

1

Automatic Summarization
Symbolic Method
ASLI

2

Summary Revision

Legal expert

3

Plain text Reviewed Extract

- XML tags lost for sentences
- Cut and paste reviewing
- Text altered

Corpus' Extracts: Plain Text to XML

- **Goal**: Find extract sentences that match the source ones

XML Decision

Plain Text
Extract's Sections

plain text

plain text

plain text

plain text
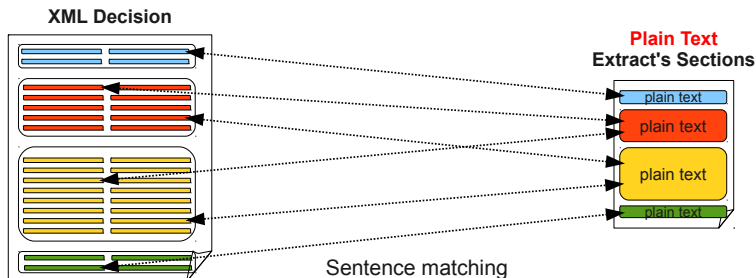
Sentence matching

- **Issues**: Sentence alterations, merging, splitting, reordering. . .

- **Algorithm**: String alignment allowing distance editing

- **Result**: 94% of matched sentences

Machine Learning

- Classifier: Naive Bayes
- Instances: Sentences
- Training set: 2/3rd
- 5 classes: Selected for a section or "not in summary"

XML Decision

Sentence classification

XML Extract

Not in summary
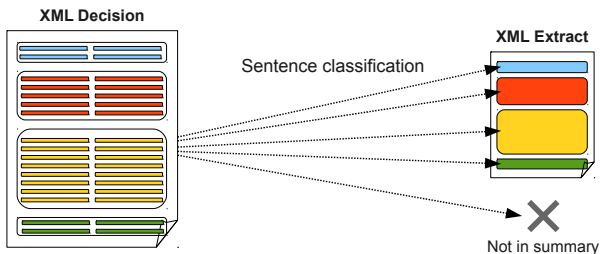
Classification Features

Surface

Position, length, count of sentences, paragraphs, sections.

Emphasis

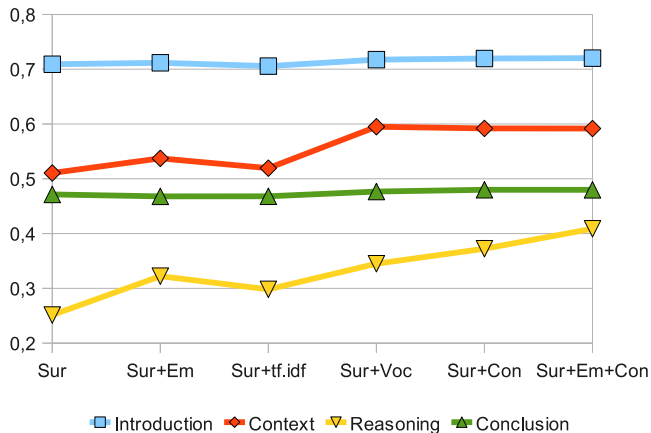HTML styles: bold, underline, italic, indent.

Content

- $tf \cdot idf$ normalized sum
- "Extract's specific vocabulary" score
    - High score: "apparently", "dismissed", "daughter", "kill"...
    - Low score: "paragraphs", "relies", "procedure"...

Classification Results for English Immigration Decisions

F1-Measure

Sur = Surface features
Em = Emphasis features
Voc = Extract's Voc.
Con = Content features

Introduction   Context   Reasoning   Conclusion
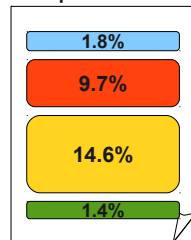
Comparison with a baseline and ASLI

- **PRODSUM**: Surface + Emphasis + Content

- **ASLI**: Current symbolic system

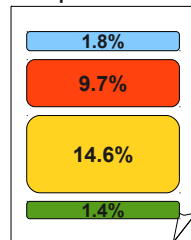- **Baseline**: First sentences of each section based on the average compression ratio

Average Extract's Compression Ratio

1.8%
9.7%
14.6%
1.4%

Comparison with a baseline and ASLI

- **PRODSUM**: Surface + Emphasis + Content
- **ASLI**: Current symbolic system
- **Baseline**: First sentences of each section based on the average compression ratio
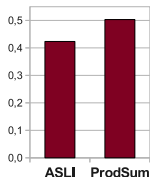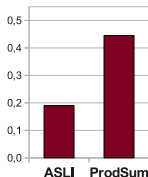
Average Extract's Compression Ratio

1.8%	
9.7%	
14.6%	
1.4%	

System classification scores

- Introduction
- Context
- Reasoning
- Conclusion
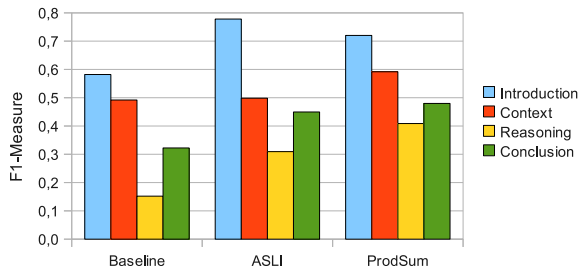
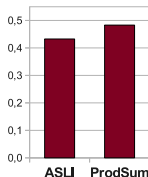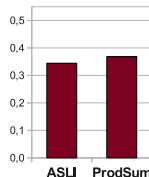# Comparison over legal fields and languages

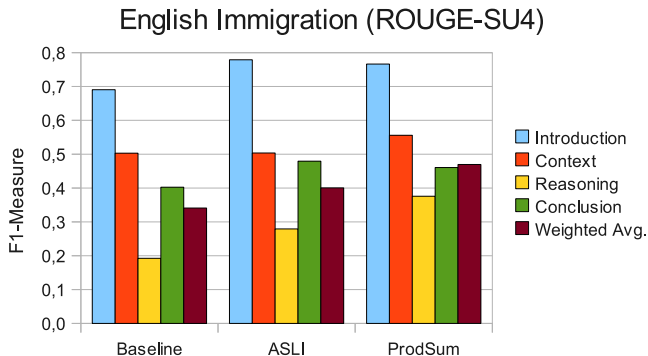**English**

Immigration   Tax

French

Immigration   Tax

Weigthed average F1-Measure scores of all sections

- ASLI's rules do not adapt well to the tax field
- PRODSUM is more reliable over fields

Content evaluation: ROUGE scores

English Immigration (ROUGE-SU4)

- No control over extract's sizes $\Rightarrow$ F1-Measure scores
- PRODSUM works best for longer sections
- Introduction and Conclusion are better handled with ASLI's rules

Conclusion

- Corpus of 4000 decisions suitable to ML
- Emphasis and "extract's vocabulary" features improve results
- Basic features generally beat the symbolic method
- PRODSUM adapts better to new legal fields

Perspectives

Features based on events, entities and factual information.

Merci

Thank you for your attention.

Questions, comments?