Automated detection and resolution of legal cross references

Approach and a study of Luxembourg’s legislation

Morayo Adedjouma\textsuperscript{1,2} Mehrdad Sabetzadeh\textsuperscript{1} Lionel Briand\textsuperscript{1}

\textsuperscript{1} University of Luxembourg, Luxembourg
\textsuperscript{2} McMaster University, Canada
Luxembourg

- Population of 530,000
- The only Grand Duchy in existence!
- Three official languages

- Language of administration and law is French
How did this work come about?

• Collaboration with Government of Luxembourg

• **CTIE**: Government’s IT Centre

• **ACD**: Tax Administration Department

• New tax system under development

• System needs to be **compliant** with the law
Art. 2 Individuals are considered resident taxpayers if they have either their fiscal or habitual residence in the Grand Duchy. Individuals are considered non-resident taxpayers if they neither have their fiscal nor their habitual residence in the Grand Duchy, but have a local income within the meaning of Art.156.

Resident taxpayers are subject to income tax over both local and foreign income.

Non-resident taxpayers are subject to income tax only over their local income within the meaning of Art.156.

R1: The system shall levy taxes on non-residents’ local income as per the annual tax scale.
R2: For non-residents, rental and lease income earned in the Grand Duchy shall be considered as local income.
Why should RE care about CRs?

• CRs important for requirements change analysis [Ghanavati et al., 2014]

• CRs entail (among other things) exceptions, constraints, and priorities [Maxwell et al., 2012]

• Ignoring or misunderstanding CRs leads to non-compliance in software [Maxwell et al., 2013]
CR detection and resolution

- Automatically identify and link CRs to targets
Related work

- CR detection and resolution: a well-trodden path

  - Dutch legislation [de Maat et al., 2006]

  - HIPAA Privacy Rule [Kiyavitskaya et al., 2008]

  - US regulations [Breaux, 2009]
What’s new?

• More thorough treatment of CR patterns

• Systematic resolution process

• Improved automation
CR patterns

- Grounded Theory study of Luxembourg’s Income Tax Law
  - first drafted in 1967;
  - 189 pages; 236 articles; 767 paragraphs
- All 1223 CRs analysed
What did we learn?

• CR patterns carry over across languages
  • thus, opportunities for reuse
• All previously-known patterns were observed
• New patterns were identified
• … among which some are ambiguous
Example of ambiguity

articles 109, 1st paragraph, numbers 1 to 3, 127 and 154ter
Approach

1. Define schema for structure of legal text
2. Transform into markup text
3. Resolve cross references
4. Visualization & Analysis

Text schema from recommendations
Tailored schema
Text with structure markup

CR patterns
Solution component discussed so far

Legal text (non-markup)
Text schema for the tax law

New or deviating from guidelines
Markup scripts derived from the text schema

**Rule: MarkArticleSegment**
```
(((ArticleHead)):start
 {{LawHead} | {PartHead} | {BookHead} | {TitleHead} | {ChapterHead} | {SectionHead} | {SubsectionHead} | {SubpartHead} | {ArticleHead} | {EOD}}:end)
```

```
:ref -->:ref.ArticleSegment= {} 
```

**Rule: MarkArticleHead**
```
(((Split))+ ({{Token.string=="Art"} | {Token.string=="Article"}}
 {{Token.string=="."}}? {Token.kind=="alphanum"}
 {Token.string=="."})
 :ref -->:ref.ArticleHead= {} 
```
Example markup

- Markup generated all the way down to sentences
CR Resolution

- Resolution algorithm discussed in the paper
CR visualization and analysis

• Legal text navigation

• Trace detection

• Circularity analysis
Completeness of CR patterns

- Studied 164 randomly-selected pages from a corpus of 13 legal texts (1640 pages)
  - Oldest text drafted in 1808; newest in 2011
- 1852 CRs in the selected pages
  - Correctly identified 1813 CRs
  - Detected 38 CRs only partially
  - Missed 1 CR completely
  - 5 false positives

Detection accuracy

**Precision:** 99.7%

**Recall:** 97.9%
Accuracy of resolution

• Resolution attempted for tax law’s 928 internal CRs

• 45 CRs (5%) marked for manual resolution

• 874 CRs (94%) automatically resolved (1736 links)

• 9 failure-to-resolve warnings
  • 8 due to anomalies in the text
  • 1 false positive due to limitations of the approach

Resolution accuracy over CR links

Precision: 99.9%
Recall: 97.5%
Scalability

• Results for the income tax law (189 pages):

<table>
<thead>
<tr>
<th>Task</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural markup generation + CR identification</td>
<td>34 sec</td>
</tr>
<tr>
<td>Resolution, generation of hyperlinked legal text, traceability matrix and diagnostics</td>
<td>290 sec</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Approx. 5 minutes</strong></td>
</tr>
</tbody>
</table>
Tool Support

Eclipse

GATE NLP Workbench

Crocopat

Analysis Rules

Analysis Results

Resolution Diagnostics

Hyperlinked Legal Text

Text Schema

Legal Text

Domain-Specific Lists

Logical Markup & CR Links

LST

JAPE

CR Detection and Resolution Scripts

Concept Markers List

Structure Markup Scripts
A beneficial byproduct

• Automated legal portal generation
Limitations

• All CR types detected but only internal ones are currently resolved

• Limited to legislative documents

  • Regulations, circular letters, directives, and parliamentary proceedings not covered
What’s next?

• Generation of semantic annotations for CRs
• Change impact analysis for legal requirements
Take away

• Example of a transition from research to a palatable solution to the industry

• Lesson learned from the transition: There is many a slip between the cup and the lip!
Acknowledgements

- **CTIE**: Ludwig Balmer, Marc Blau, Yves Boland, Michaël Masseroni, Gautier Barrère
- **ACD**: Thierry Prommenschchenkel, Guy Heintz, Vincent D'Angelo
- **SnT**: Domenico Bianculli