Challenges in Automating Style Checking for Legislative Texts

Stefan Höfler and Kyoko Sugisaki

Second Workshop on Computational Linguistics and Writing. Avignon, April 23, 2012.

Motivation

Current Situation:

① Legislative texts should meet linguistic quality requirements.

② Authorities have published style guidelines for legislative drafting.

③ Language experts review and edit the drafts.

Problem:

① Manual assessment is time-consuming.

② Authors and editors are prone to overlook some of the violations.

Aim:

To develop methods for an automated detection of
violations of existing style guidelines in legislative drafts.

Overview

- Introduction

- Types of rules
- Approach
- Challenges

- Conclusion

Rule types in style guidelines

Terminology:

① Abbreviations of titles should be less than 5 characters.

② Word “*beziehungsweise*” (respectively) should not be abbreviated as *bzw.*

Syntax:

① The main verb of a sentence should be introduced as early as possible.

② In general, avoid passive.

Discourse:

① Only include normative content; do not include explanations, justifications and descriptions.

② Put conditions before their consequences.

Approach

Step 1: Pre-processing

- Text segmentation: chapters, sections, articles, …
- Part-of-speech tagging (TreeTagger)
- Morphological analysis (Gertwol)
- Parsing (ParZu)
- Context recognition: legal definitions, statements of purpose, …

Step 2: Error detection

- Searching the preprocessed text for violations of style guidelines

Challenges

Pre-processing

- ➢ Peculiarities of legislative language
  - ➔ Domain-specific pre-processing required

Error detection

- ➢ Peculiarities of legislative style guidelines
  - ➔ Domain-specific error detection required

Challenges for error detection

① **Context-dependent rules**
The application of individual rules is dependent on their context.

② **Abstract rules**
Many rules are too abstract for an automatic detection of violations.

③ **Conflicting rules**
Rules do not constitute absolute constraints and may conflict with other rules.

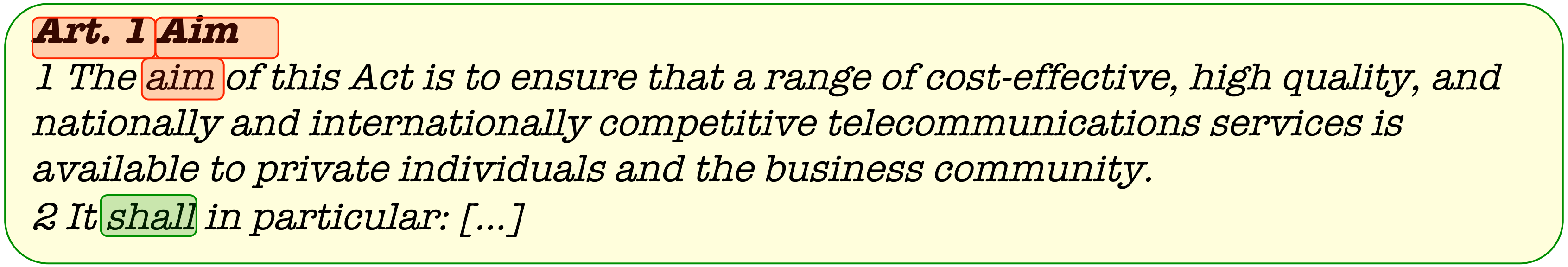
Challenge 1: Context-dependent rules

Don‘t use the modal verb „sollen“ (should/shall)
– unless it is in the statement of purpose.

Detection of violations of this rule:

① Detect statements of purpose (search for linguistic cues).

② Detect „sollen“ in provisions other than statements of purpose.



Art. 1 Aim

1 The aim of this Act is to ensure that a range of cost-effective, high quality, and nationally and internationally competitive telecommunications services is available to private individuals and the business community.

2 It shall in particular: […]

➔ Domain-specific pre-processing for context recognition needed.

Challenge 2: Abstract rules

Only include normative content;
do not include explanations, justifications and descriptions.

Detection of violations of this rule:

① Determine linguistic cues (e.g. discourse markers) for explanations, justification, descriptions, ...

② Search for these discourse markers.

Private household aids who give birth to a child during the processing of their permit may remain in Switzerland until their employment contract expires […]. Therefore, they have to leave the country after their contract has expired.

(Art. 16 Abs. 1 VPH, Version 12, 11 June 2010, emphasis added)

➔ **The domain-specific concretisation of the rules needed.**

Challenge 3: Conflicting rules

- Put conditions before their consequences. ➔ not violated
- The main verb of a sentence should be introduced as early as possible ➔ violated

Detection of violations of this rule:

- Determine linguistic cues for conditions and consequences and search for them.
- Determine the main verb of a sentence and search for it.

As far as the offender fails to pay the monetary penalty despite being granted an extended deadline for payment or a reduced daily penalty unit or fails to perform the community service despite being warned of the consequences,

the alternative custodial sentence is executed.

Judgment of these rules:

➔ Weighting of conflicting rules needed.

➔ In-depth corpus-based studies

Corpus for linguistic studies

The Swiss Legislation Corpus (SLC):

① comprises the whole body of contemporary legislative texts of the Swiss Confederation.

② is a parallel corpus (German, French and Italian).

- ➢ **1,915 texts** per language, currently.
- ➢ **800 to 1.3 million words** per text

③ is a corpus with inter- and intra-textual time depth.

- ➢ Inter-textual time depth: ca. **150 years**
- ➢ Inter-textual time depth: up to **122 years**

④ is an annotated corpus:

- ➢ **textual meta information** (text title, type of law…)
- ➢ **text segmentation** (article, paragraph, sentence boundaries)
- ➢ **date stamping** (date of origin of each individual text segment)
- ➢ **part-of-speech tagging** (TreeTagger)
- ➢ …

Conclusion

Aim of the project:

To develop methods for an automated detection of violations of guidelines for legislative drafting

Challenges for error detection:

① Context-dependent rules

➔ Domain-specific pre-processing for context recognition needed

② Abstract rules

➔ Domain-specific concretization of the rules needed

③ Conflicting rules

➔ Weighting of rules needed

In-depth **corpus-based research into legislative language** is a prerequisite for the development of automated style checking methods for the domain.

Acknowledgement

We thank

The Swiss National Foundation, Switzerland
Prof. Dr. Michael Hess, Institute of Computational Linguistics, University of Zurich
Prof. Dr. Felix Uhlmann, Institute of Law, University of Zurich
Dr. Rebekka Bratschi, Swiss Federal Chancellery

for their support of our project.

Homepage of our project:
http://www.cl.uzh.ch/research/maschinellestilpruefung/gesetzestextanalyse_en.html