Universität Zürich[UZH]

Institut für Computerlinguistik

Automatic detection of syntax-related style violations in law texts

Kolloquium in Computerlinguistik FS 2014

Kyoko Sugisaki

Background

SNF-Project

“Automated detection of styleguide violations in legislative drafts”

My PhD Project

- Scope: Detection of **syntax-related** style violations

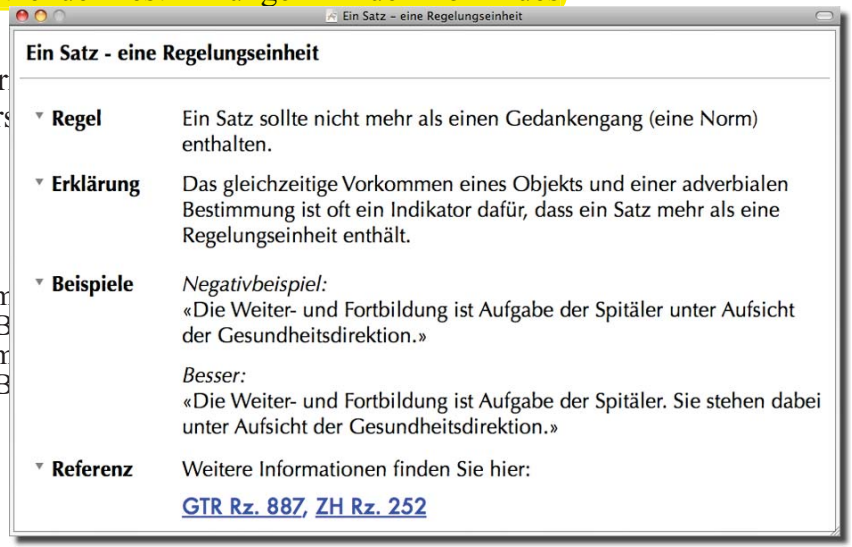
Art. 163 Form der Erlasse der Bundesversammlung

1 Die Bundesversammlung erlässt rechtsetzende Bestimmungen in der Form des Bundesgesetzes oder der Verordnung.

2 Die übrigen Erlasse ergehen in der For
beschluss, der dem Referendum nicht unters
bezeichnet.

83 Angenommen in der Volksabstimmung vom
BRB vom 4. Febr. 2002 – AS **2002** 241; BB
84 Angenommen in der Volksabstimmung vom
BRB vom 4. Febr. 2002 – AS **2002** 241; BB

50



Contents

Style guidelines: Syntax-related rules

“Avoid the modal verb of obligation, if the agent is an authority.”

Die Kantone müssen die Abrechnung dem BAG spätestens bis zum 30. Juni des folgenden Jahres einreichen .

Style guidelines: Syntax-related rules

“Avoid complex participle phrases”

„Die Leistungsstatistik der Spitäler muss in Abstimmung mit der nach dem Anhang zur Verordnung vom 30. Juni 1993 über die Durchführung von statistischen Erhebungen des Bundes erstellten Krankenhausstatistik und der Medizinischen Statistik der Krankenhäuser erstellt werden.”

Style checking tool: Pipeline

Step 1: Domain-specific preprocessing

- Text segmentation
- POS tagging (TreeTagger [Schmid 1995])
- Morphological analysis (Gertwol [Haapalainen and Majorin 1994])
- Rule-based tagger: Morphosyntactic disambiguation and grammar feature tagging (Constraint Grammar [Karlsson 1995])
- **Recognition of grammatical functions (Structured perceptron, conditional random fields)**

Step 2: Detection of syntax-related style violations

- **Rule-based approach (Constraint Grammar)**
- **Data-driven approach (Gaussian mixture model)**

Contents

Grammatical functions

- Grammatical functions: subject, indirect object, direct object, etc.
- German:
  - Indicators: Case, word order, animacy, etc.
    - Ambiguous case features

> Die Kantone können die Listen den regionalen Gegebenheiten anpassen.

    - Relatively free word order

> Diese Listen können die Kantone den regionalen Gegebenheiten anpassen.

Grammatical functions

"Die Kantone können die Listen den regionalen Gegebenheiten anpassen."

Token	"Kantone"	"Listen"	"Gegebenheiten"
Case	nominative or accusative	nominative or accusative	dative
Topological fields	vorfeld	mittelfeld	mittelfeld
Animacy	organization	thing	thing
Definiteness	definite	definite	definite
Type	full noun	full noun	full noun

Task: Recognition of grammatical functions

- **Task**: label dependency grammar tags → Supertagging [Bangalore 1999]
- **Methods**: machine learning (ML) techniques for sequences
  - Structured perceptron [Collins 2002]
  - Conditional random fields (CRF) [Lafferty et. al 2001]
- **Targets**: 17 dependency classes (tag sets) restricted to nouns and prepositions
    - Subject
    - Accusative object
    - Dative object
    - Genitive modifier
    - …

Features: Recognition of grammatical functions

"Die Kantone können die Listen den regionalen Gegebenheiten anpassen."

Features	"Kantone"	"Listen"	"Gegebenheiten"
Case	nom. or acc.	nom. or acc.	dat.
Morphology	die	die	den
Topological fields	vorfeld	mittelfeld	nachfeld
Animacy	organization	none	none
Brown clustering	111	100	000
Verb	anpassen	anpassen	anpassen
word	Kantone	Listen	Gegebenheiten
Previous word	none	Kantone	Listen
Next word	Listen	Gegebenheiten	none
…			

Training: Recognition of grammatical functions

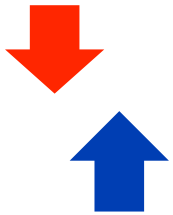
- Two MLs
  - Structured perceptron
  - CRF (wapiti)
- Training for MLs:
  - Data for supervised MLs:
    - TüBa DZ → 60% (training:700,888 tokens), 20% (cross validation: 231,277 tokens), 20% (evaluation 232,561 tokens)
    - Law texts → 300 sentences (XX tokens) for training, 200 sentences (XXX tokens) for evaluation
  - Preprocessing: sequence chunking and feature extraction
  - Experiments:
    - Scope: The tagger optimized for law texts

Experiments: Recognition of grammatical functions

Label accuracy = $\frac{\text{\# of correctly tagged tokens}}{\text{total \# of tokens}}$

- **Baseline**: Training on the training set of TüBa
  - TüBa: CRF → 89.32% (label accuracy), Perceptron → 80.83%
  - Law texts (CRF): 88.60%

- **Experiment 1:** Remove some domain-specific noisy data containing PAR and ROOT from TüBa
  - TüBa (CRF): 86.74%
  - Law texts (CRF): 89.71%

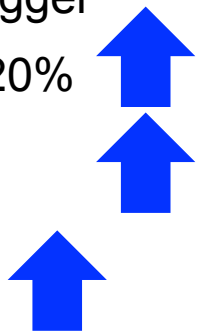


Experiments: Recognition of grammatical functions

Baseline: Law texts (CRF): 88.60%

- **Experiment 2**: Add 300 training sentences from the domain
  - Law texts (CRF): + 300 sentences → 91.06 %

- **Experiment 3:** Add the outputs from the domain-specific rule-based tagger
  - A: Law texts (CRF): + Case features (~56% = disambiguated): 91.20%
  - B: Law texts (CRF): + Dependency label (~53% = deterministic, 94.73% label accuracy): 92.07%
  - C: Law texts (CRF): + Case features & dependency label : 92.69%



Evaluation: Recognition of grammatical functions

- Reference parser: dependency parser developed by Bohnet [2010]
  - Best German parser for dependency labeling in CONLL-2009 shared task
  - Trained on the same training set (TüBa)
- Test data
  - Law texts → 200 sentences for evaluation
- F1 score for relevant dependency labels (e.g. grammatical functions)

Evaluation: Recognition of grammatical functions

	CRF (ex4C)	Bohnet (TreeTagger)
Training data	TüBa + Law texts	TüBa
Test data	Law texts	Law texts
Overall label accuracy	**92.60%** (1929, 154)	86.46% (1801, 282)
F1 score SUBJ	0.94 (0.94, 0.94)	0.90 (0.91, 0.89)
OBJA	0.86 (0.88, 0.84)	0.78 (0.77, 0.8)
OBJD	**0.85** (0.91, 0.80)	0.57 (0.71, 0.48)
GMOD	0.95 (0.96, 0.94)	0.92 (0.92, 0.92)
APP	**0.82** (0.73, 0.93)	0.67 (0.80, 0.57)

Contents

Rule-based Approach: Detection of syntax-related rules

“Avoid the modal verb of obligation, if the agent is an authority.”

Die Kantone müssen die Abrechnung dem BAG spätestens bis zum 30. Juni des folgenden Jahres einreichen .

“Assign the tag “&Style_Violation” to the token “müssen”, if the token is the subject and an authority and is not in passive voice“

ADD (&Style_Violation) TARGET (“müssen”)
IF (0 (SUBJ)) (0 (AUTHORITY)) (NOT 0* (&passive));

Rule-based Approach: Detection of syntax-related rules

“Avoid complex participle phrases”

„Die Leistungsstatistik der Spitäler muss in Abstimmung mit der nach dem Anhang zur Verordnung vom 30. Juni 1993 über die Durchführung von statistischen Erhebungen des Bundes erstellten Krankenhausstatistik und der Medizinischen Statistik der Krankenhäuser erstellt werden.”

“Assign the tag “&Style_Violation” to participle phrases”

ADD (&Style_Violation) TARGET (&participle_phrase);

Challenge: Too many participle NPs are retrieved

- "Sie haben dem Kind, insbesondere auch dem körperlich oder geistig gebrechlichen, eine angemessene, seinen Fähigkeiten und Neigungen soweit möglich entsprechende allgemeine und berufliche Ausbildung zu verschaffen.

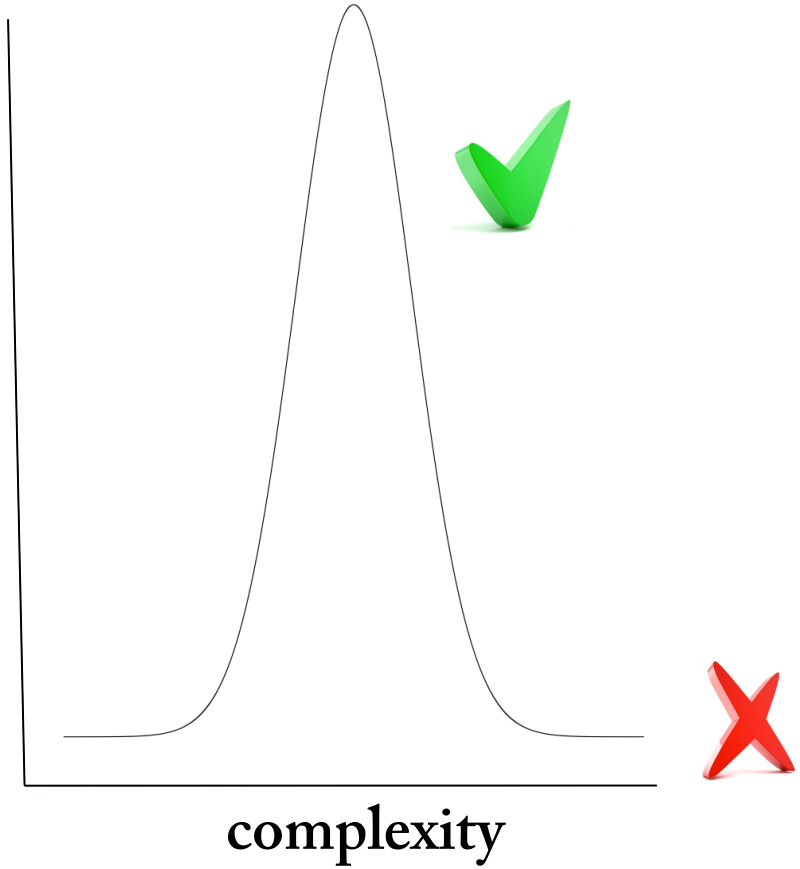
- "Die von der Eidgenössischen Bankenkommission oder vom Bundesamt für Privatversicherungen vor Inkrafttreten des FINMAG anerkannten Prüfgesellschaften und leitenden Prüferinnen und Prüfer gelten als zugelassen."

- "Für im Grundbuch eingetragene Berechtigte einer Dienstbarkeit gelten die Bestimmungen über die richterlichen Massnahmen bei Unauffindbarkeit des Eigentümers oder bei Fehlen der vorgeschriebenen Organe einer juristischen Person oder anderen Rechtsträgerin sinngemäss."

- "Ein solches von einem Kanton ausgesprochenes Verbot ist in der ganzen Schweiz gültig."

Hypothesis

- Not all participle NPs are stylistically inadequate.
- Stylistically adequate vs. stylistically inadequate
  - Stylistically adequate = domain-specific writing conventions = more frequent
  - Stylistically inadequate = less frequent
- Factors:
  - complexity of the structures
  - complexity of the contexts



Features

- Complexity of the participle phrase
  - Nouns and pronouns → 6
  - Prepositions → 2
  - Coordinating conjunctions → 1
  - Modifiers (adjectives and adverbs) → 1
  - Articles → 1
- Complexity of the context (field)
  - Nouns and pronouns → 1
  - Prepositions → 1
  - Coordinating conjunctions → 0
  - Modifiers (adjectives and adverbs) → 2
  - Articles → 1

„Es führt ein nicht öffentliches Verzeichnis über die von den kantonalen Vollzugsbehörden nach Artikel 11 Absatz 1 oder Artikel 8 Absatz 5 ChemRRV verfügten Massnahmen.

Training

- **Experiments:** Clustering with Gaussian mixture model (GMM) [scikit-learn & weka]
  - GMM = a weighted mixture of multivariate Gaussians
- **Hypothesis**:
  - Less probable to belong to a cluster = outliers = more likely to be style violations
  - 3 % of the weak members of clusters are good candidate for style violations
- **Data:** Swiss Legislation Corpus (1959 documents) [Höfler & Piotrowski 2011]
  - 1500 documents (~75%) for training

    → 18676 instances
  - 459 documents (~25%) for evaluation
- **Preprocessing:** extract participle phrases and features

Evaluation

- Swiss Legislation Corpus [Höfler & Piotrowski 2011]
  - 459 documents for testing
    - 100 instances for precision
    - manually annotated: Do you want to reformulate the participle NPs?
  - Results:

Annotation	Yes!	Rather yes	Rather No	No!	?	Bug
count	43	13	12	4	9	19

→ Precision: 62.22

Summary

- Recognition of grammatical functions (supervised MLs for sequences)
  - CRF > Structured perceptron
  - improving the performance:
    - Removal of domain-specific noises
    - Adding a small amount of in-domain data for training
    - Combining with a rule-based domain-specific tagger
- Detection of syntax-related style violations (example of participle phrases)
  - Simple GMM-based clustering: quite promising results
  - Difficulties in discriminating the preprocessing bugs and style violations

Future work

- Detection of syntax-related style violations
  - More work in preprocessing
  - Other syntax-related style violations
  - Poisson Mixture Model