Computerlinguistische Methoden für die Gesetzesredaktion

Sprachentag 2010 der Schweizerischen Bundeskanzlei

Stefan Höfler und Alexandra Bünzli

15:15-16:15

Einsatzmöglichkeiten...

... computerlinguistischer Methoden in der Gesetzesredaktion:

– maschinelle Rechtschreib- und Grammatikprüfung
– computerunterstützte Terminologie-Kontrolle
– «Translation Memories» und maschinelle Übersetzung
– maschinelle «Stil»-Kontrolle

Richtlinienverletzungen maschinell erkennen

Wie können Verletzungen von Richtlinien für die sprachliche Ausgestaltung von Gesetzen maschinell erkannt werden?

Ausgangspunkt: technische Redaktion

Firmen verfügen oft über **interne Stilregeln** für ihre technische Dokumentation (Benutzerhandbücher etc.), z.B.

- kein Passiv verwenden
- Titel nicht mit einem Artikel beginnen
- kein Futur verwenden
- Massangaben abkürzen
- ...

Maschinelle Stil-Kontrolle

Controlled Language Checkers

spezielle **Autorenwerkzeuge**, die einen Textentwurf maschinell auf die Einhaltung der Stilregeln überprüfen:

- markieren Textpassagen, die eine Stilregel verletzen;
- informieren den Benutzer darüber, welche Stilregel in der Passage (potentiell) verletzt wurde.

Funktionsweise

- maschinelle **Vorverarbeitung** des Textes
- **Suche** nach Richtlinienverletzung im vorverarbeiteten Text

Ein Text...

Art. 1 Oberste Recht sprechende Behörde

[1] Das Bundesgericht ist die oberste Recht sprechende Behörde des Bundes.

[2] Es übt die Aufsicht über die Geschäftsführung des Bundesstrafgerichts und des Bundesverwaltungsgerichts aus.

[3] Es besteht aus 35–45 ordentlichen Bundesrichtern und Bundesrichterinnen.

[4] Es besteht ausserdem aus nebenamtlichen Bundesrichtern und Bundesrichterinnen; deren Zahl beträgt höchstens zwei Drittel der Zahl der ordentlichen Richter und Richterinnen.[3]

[5] Die Bundesversammlung legt die Zahl der Richter und Richterinnen in einer Verordnung fest.

Art. 2 Unabhängigkeit

[1] Das Bundesgericht ist in seiner Recht sprechenden Tätigkeit unabhängig und nur dem Recht verpflichtet.

[2] Seine Entscheide können nur von ihm selbst nach Massgabe der gesetzlichen Bestimmungen aufgehoben oder geändert werden.

... ist intern einfach eine Zeichenkette...

```
Art. 1 Oberste Recht sprechende Behörde 1 Das Bundesgericht ist
die oberste Recht sprechende Behörde des Bundes. 2 Es übt die
Aufsicht über die Geschäftsführung des Bundesstrafgerichts und
des Bundesverwaltungsgerichts aus. 3 Es besteht aus 35–45
ordentlichen Bundesrichtern und Bundesrichterinnen. 4 Es besteht
ausserdem aus nebenamtlichen Bundesrichtern und
Bundesrichterinnen; deren Zahl beträgt höchstens zwei Drittel der
Zahl der ordentlichen Richter und Richterinnen.3 5 Die
Bundesversammlung legt die Zahl der Richter und Richterinnen in
einer Verordnung fest. Art. 2 Unabhängigkeit 1 Das Bundesgericht
ist in seiner Recht sprechenden Tätigkeit unabhängig und nur dem
Recht verpflichtet. 2 Seine Entscheide können nur von ihm selbst
nach Massgabe der gesetzlichen Bestimmungen aufgehoben oder
geändert werden. Art. 3 Verhältnis zur Bundesversammlung 1 Die
Bundesversammlung übt die Oberaufsicht über das Bundesgericht
aus.
```

... evtl. mit zusäztlicher Formatierungsinformation

```
<H5><B>Art. 1</B> Oberste Recht sprechende
Behörde</H5><P><SUP>1</SUP> Das Bundesgericht ist die oberste
Recht sprechende Behörde des Bundes.</P><P><SUP>2</SUP> Es übt
die Aufsicht über die Geschäftsführung des Bundesstrafgerichts
und des Bundesverwaltungsgerichts aus.</P><P><SUP>3</SUP> Es
besteht aus 35-45 ordentlichen Bundesrichtern und
Bundesrichterinnen.</P><P><SUP>4</SUP> Es besteht ausserdem aus
nebenamtlichen Bundesrichtern und Bundesrichterinnen; deren Zahl
beträgt höchstens zwei Drittel der Zahl der ordentlichen Richter
und Richterinnen.<SUP>1</SUP></P><P><SUP>5</SUP> Die
Bundesversammlung legt die Zahl der Richter und Richterinnen in
einer Verordnung fest.</P><H5><B>Art. 2</B>
Unabhängigkeit</H5><P><SUP>1</SUP> Das Bundesgericht ist in
seiner Recht sprechenden Tätigkeit unabhängig und nur dem Recht
verpflichtet.</P><P><SUP>2</SUP> Seine Entscheide können nur von
ihm selbst nach Massgabe der gesetzlichen Bestimmungen aufgehoben
oder geändert werden.</P>
```

Vorverarbeitung: mehr Information kennzeichnen

```
<Artikel><Ueberschrift><B>Art. 1</B> Oberste Recht sprechende
Behörde</Ueberschrift><Absatz><P><SUP>1</SUP> Das Bundesgericht
ist die oberste Recht sprechende Behörde des
Bundes.</P></Absatz><Absatz><P><SUP>2</SUP> Es übt die Aufsicht
über die Geschäftsführung des Bundesstrafgerichts und des
Bundesverwaltungsgerichts aus.</P></Absatz>
<Absatz><P><SUP>3</SUP> Es besteht aus 35-45 ordentlichen
Bundesrichtern und Bundesrichterinnen.</P></Absatz>
<Absatz><P><SUP>4</SUP> Es besteht ausserdem aus nebenamtlichen
Bundesrichtern und Bundesrichterinnen; deren Zahl beträgt
höchstens zwei Drittel der Zahl der ordentlichen Richter und
Richterinnen.<SUP>1</SUP></P></Absatz> <Absatz><P><SUP>5</SUP>
Die Bundesversammlung legt die Zahl der Richter und Richterinnen
in einer Verordnung
fest.</P></Absatz></Artikel><Artikel><Ueberschrift><B>Art. 2</B>
Unabhängigkeit</Ueberschrift><Absatz><P><SUP>1</SUP> Das
Bundesgericht ist in seiner Recht sprechenden Tätigkeit
unabhängig und nur dem Recht verpflichtet.</P></Absatz>
<Absatz><P><SUP>2</SUP> Seine Entscheide können nur von ihm
selbst nach Massgabe der gesetzlichen Bestimmungen aufgehoben
oder geändert werden.</P></Absatz></Artikel>
```

Vorverarbeitung (engl. *pre-processing*)

Der zu überprüfende Text wird maschinell analysiert und mit der so gewonnen **strukturellen und linguistischen Information** angereichert:

- **Tokenisierung**
  maschinelle Wortgrenzenerkennung
- **Textsegmentierung**
  maschinelles Erkennen von Kapitelgrenzen, Abschnittsgrenzen, Titeln, Fussnoten, Satzgrenzen, ...
- **Part-of-speech Tagging**
  maschinelle Wortartenerkennung
- **Morphologische Analyse**
  maschinelles Erkennen von Tempus, Genus, Modus, ...
- **Syntaktische Analyse (Parsing)**
  maschinelles Erkennen der Satzstruktur

(Demo im 2.Teil des Vortrags)

Suche nach Richtlinienverletzungen

Spezielle Suchregeln beschreiben die exakten Bedingungen, unter denen eine bestimmte Stilrichtlinie *verletzt* ist («**Fehlermodellierung**»).

Die Stilregel «Futur vermeiden» ist (unter anderem) verletzt, wenn...

```
Avoid_future

/* Example: ".. It will be necessary .." */

TRIGGER (80)==@will^1 [-@comma]* @verbInf^2

    ->($will, $verbInf)

    ->{mark : $will, $verbInf;}
```

... eine Wortform des Lemmas *will* von einem Infinitiv gefolgt wird und dazwischen null, ein oder mehrere Elemente (Token) sind, aber kein Komma.

In diesem Fall markiere *will* und den Infinitiv – und gib Fehler-meldung 80 aus.

Quelle
Lehmann, S. (2009) «Kontrollierte Sprachen und Sprach-technologie in der Industrie: das Autorenwerkzeug acrolinx.», UZH. https://cast.switch.ch/vod/clips/2pjsz3usia/link_box

Fehlermodellierung (engl. *error modelling*)

Die Methode der Fehlermodellierung anzuwenden, bedeutet also,

– anstatt die «Wohlgeformtheit» eines ganzen Textes zu überprüfen,

– zu antizipieren (bzw. «modellieren»), wie spezifische Richtlinien-verletzungen aussehen, und gezielt nach solchen zu suchen.

«***Fehler***modellierung» hat sich als technischer Begriff für diese Methode so eingebürgert.

Insbesondere in der Gesetzesredaktion sind sprachliche Richtlinien aber meist keine absoluten Stilregeln, sondern oft «nur» ***Empfehlungen*** oder ***Faustregeln***.

Gesucht wird also nicht nach eigentlichen «Fehlern», sondern nach ***potentiellen*** Richtlinienverletzungen.

Geht das auch für die Gesetzesredaktion?

[2] Über Rechtsfragen von grundsätzlicher Bedeutung oder auf Antrag eines Richter oder einer Richterin entscheiden sie in Fünferbesetzung. Ausgenommen sind Beschwerden gegen Entscheide der kantonalen Aufsichtsbehörden in Schuldbetreibungs- und Konkurssachen.

[3] In Fünferbesetzung entscheiden sie ferner über Beschwerden gegen referendumspflichtige kantonale Erlasse und gegen kantonale Entscheide über die Zulässigkeit einer Initiative oder das Erfordernis eines Referendums. Ausgenommen sind Beschwerden, die eine Angelegenheit einer Gemeinde oder einer anderen Körperschaft des kantonalen Rechts betreffen.

Art. 21 Abstimmung

[1] Das Gesamtgericht, die Präsidentenkonferenz, die Verwaltungskommission und die Abteilungen treffen die Entscheide, Beschlüsse und Wahlen, wenn das Gesetz nichts anderes bestimmt, mit der absoluten Mehrheit der Stimmen.

Geht das auch für die Gesetzesredaktion?

[2] Über Rechtsfragen von grundsätzlicher Bedeutung oder auf Antrag eines Richter oder einer Richterin entscheiden sie in Fünferbesetzung. Ausgenommen sind Beschwerden gegen Entscheide der kantonalen Aufsichtsbehörden in Schuldbetreibungs- und Konkurssachen.

[3] In Fünferbesetzung entscheiden sie ferner über Beschwerden gegen referendumspflichtige kantonale Erlasse und gegen kantonale Entscheide über die Zulässigkeit einer Initiative oder das Erfordernis eines Referendums. Ausgenommen sind Beschwerden, die eine Angelegenheit einer Gemeinde oder einer anderen Körperschaft des kantonalen Rechts betreffen.

Art. 21 Abstimmung

[1] Das Gesamtgericht, die Präsidentenkonferenz, die Verwaltungskommission und die Abteilungen treffen die Entscheide, Beschlüsse und Wahlen, wenn das Gesetz nichts anderes bestimmt, mit der absoluten Mehrheit der Stimmen.

Kommentar
Komplexe Koordinationen sind oft besser lesbar, wenn sie in eine Aufzählung (Bst. a, b, ...) aufgegliedert werden.

Weitere Informationen: Rz.187 ff.; ...

Beispiel: ...

Geht das auch für die Gesetzesredaktion?

Herausforderung für die Vorverarbeitung

Gesetzessprache

- komplex
- idiosynkratisch

➡ Methoden für die Vorverarbeitung müssen auf die Eigenheiten der Gesetzessprache angepasst werden.

Herausforderung für die Fehlermodellierung

sprachliche Richtlinien für Gesetzestexte

- domänenspezifisch
- abstrakt

➡ Es braucht eine spezielle Fehlermodellierung für Richtlinienverletzungen in Gesetzestexten.

Geht das auch für die Gesetzesredaktion?

Ziel des Vortrags

Sie haben eine Vorstellung davon, was computerlinguistische Methoden bei der maschinellen Stil-Kontrolle von Gesetzestexten erreichen könnten und was nicht.

- Was geht?
- Was geht nicht?
- Welche Vorverarbeitungsschritte werden benötigt?

«Pro Artikel höchstens drei Absätze; pro Absatz ein Satz»

Vorverarbeitung

Gliederungseinheiten erkennen und kennzeichnen (**Textsegmentierung**):

- **Artikel**, **Absätze**, ...
  (im CHLexML-Format des SVRI vorhanden)
- **Sätze**

Fehlermodellierung

Eine Richtlinienverletzung liegt vor,

- wenn ein Artikel > 3 Absätzen enthält.
- wenn ein Absatz > 1 Sätze enthält.

Anmerkung
Der Entscheid, ob eine bestimmte Regel überprüft wird, muss dem Benutzer / der Benutzerin überlassen bleiben.

Die erste von Eugen Huber's Regeln wird man z.B. oft gar nicht anwenden wollen...

«Das Verb *sollen* ist zu vermeiden; ...»

Suche nach allen Wortformen (*soll*, *sollen*, *sollte*, ...) ist umständlich.

Vorverarbeitung

morphologische Analyse, insbesondere

Rückführung der Wortformen auf ihr Lemma (**Lemmatisierung**)

```
... <wordform lemma="sollen">sollte</wordform> ...
```

Fehlermodellierung

Eine Richtlinienverletzung liegt vor,

- wenn eine Form des Lemmas *sollen* im Text vorkommt.

«... in einer Zweckbestimmung ist die Verwendung von *sollen* zulässig.»

Vorverarbeitung

Kontext-Erkennung: Erkennen und Kennzeichnen von Zweckbestimmungen, Geltungsbereichsbestimmungen u.ä.

Zweckbestimmungen z.B. identifizieren anhand von Faktoren wie

- **Artikelüberschrift** (enthält Begriffe wie *Zweck*)
- **Position** (erster Artikel des Erlasses)

Fehlermodellierung

Ein Richtlinienverletzung liegt vor,

- wenn eine Form von *sollen* ausserhalb einer Zweckbestimmung im Text vorkommt.

Wie kann man Legaldefinitionen erkennen?

Für die Erkennung von Legaldefinitionen reichen Artikelüberschrift und Position im Text nicht aus: Sie kommen auch einzeln im Text vor.

Europäische Studien (*Legal Information Retrieval*)

– de Maat & Winkels (2010): maschinelle Identifizierung von **Legaldefinitionen in niederländischen Gesetzestexten**

– Walter & Pinkal (2009): maschinelle Identifizierung von **Definitionen in deutschen Gerichtsentscheiden**

Ansatz

– Legaldefinitionen anhand von **typischen Satzmustern** erkennen

Resultate

– 100% Präzision in Gesetzestexten; 70% Präzision in Gerichtsurteilen

---

Quellen

de Maat, E., Winkels, R. (2010) «Automated classification of norms in sources of law.» In: *Semantic Processing of Legal Texts*. Berlin, Springer.

Walter, S., Pinkal, M. (2009) «Definitions in court decisions: Automatic extraction and ontology acquisition.» In: *Law, Ontologies and the Semantic Web.* Amsterdam, IOS Press.

Satzmuster von Legaldefinitionen

Beispiele

Definiendum

Definiens

– **Satzmuster:** *Als* **NP (***im Sinne dieser Verordnung***)** *gilt/gelten* **NP**.

«Als Geflügel im Sinne dieser Verordnung gelten Hühnervögel (Galliformes), Schwimmvögel (Anseriformes) und Laufvögel (Struthioniformes).» (Bratschi 2009)

– **Satzmuster:** **NP**: **NP(**, *einschliesslich* **NP)**.

«Gehege: umgrenzter Bereich, in dem Tiere gehalten werden, einschliesslich Auslaufflächen, Käfigen, Volieren, Terrarien, Aquarien, Aufzuchtbecken und Fischteichen» (Bratschi 2009)

Vorverarbeitung

partielle syntaktische Analyse (**Chunking**),

z.B. Erkennen von Nominalphrasen (NP)

Quelle
Bratschi, R. (2009) «Frau im Sinne dieser Badeordnung ist auch der Bademeister: Legaldefinitionen aus redaktioneller Sicht.» *LeGes*, 2:191–213.

«Eine Legaldefinition ist nur nötig, wenn der definierte Begriff mindestens dreimal vorkommt.»

Fehlermodellierung

Eine Richtlinienverletzung liegt vor,

- wenn das Definiendum einer Legaldefinition im restlichen Text weniger als dreimal vorkommt.

«Der zu definierende Begriff darf normalerweise nicht mit sich selber erklärt werden.»

Fehlermodellierung

Eine Richtlinienverletzung liegt vor,

- wenn das Definiendum einer Legaldefinition (bzw. eine Komponente davon) auch im Definiens vorkommt.

Beispiel

- *«Als Küchen- und Speisereste gelten Speisereste, die aus Einrichtungen stammen, in denen ....»*

(Legaldefinitionen können auch transitiv zirkulär sein!)

«Pro Satz eine Norm»

Problem

Diese Regel bezieht sich nicht mehr nur auf die **Form**, sondern auch auf die **Bedeutung** des Textes.

Bedeutung ist aber für den Computer nicht zugänglich.

Lösung

Wir benötigen **Indikatoren** («Symptome») in der **Form** des Textes, die auf das Vorhandensein einer Richtlinienverletzung hinweisen.

Forschungsfrage

Anhand welcher sprachlicher Indikatoren können Verletzungen solcher Richtlinien erkannt werden?

Indikatoren für die Präsenz von mehreren Normen

– Konjunktionen wie ***wobei***, ***dagegen***, ***jedoch (nicht)***, ***(nicht) aber***

«Das Bundesamt bestimmt die anzuwendende Methode, wobei die Flächenkostenpauschale die Regel bilden soll.»

– Präpositionen wie ***vorbehältlich***, ***mit Ausnahme (von)***, ***ausser***

«Das Vorverfahren ist grundsätzlich und vorbehältlich der Teilnahmerechte der Parteien geheim.»

– Objekt/Prädikativ und Adverbiale im gleichen Satz

«Die Weiter- und Fortbildung ist Aufgabe der Spitäler unter Aufsicht der Gesundheitsdirektion.»

– ...

Indikatoren sind nur «Symptome»

Wie medizinische Symptome greifen auch sprachliche Indikatoren für Richtlinienverletzungen oft zu weit oder zu kurz:

- **Genauigkeit** (engl. *precision*) **< 100%**
  Es werden auch Sätze gefunden, die in Ordnung sind.
- **Trefferquote** (engl. *recall*) **< 100%**
  Es werden nicht alle Sätze gefunden, die nicht in Ordnung sind.

Ziel

Indikatoren so verfeinern, dass Genauigkeit und Trefferquote möglichst hoch sind.

(Wir bitten um sachdienliche Hinweise!)

«Geltungsbereichsbestimmungen sollen keine materiellen Elemente enthalten.»

Problem

Materielle Elemente sind meist schwer erkennbar; es gibt Ausnahmen:

Beispiel

«Diese Verordnung gilt nicht für:

a. technisches Personal;
b. Hilfsassistierende; stellt ein Institut solche Personen ein, muss das Dekanat unterrichtet werden;
c. Mitarbeitende von Sekretariaten und Verwaltung.»

Indikatoren

für die Präsenz eines materiellen Elements:

– Modalverb ***muss***
– Abweichung vom typischen Satzmuster für Geltungsbereichs-bestimmungen

Pragmatik ist nicht maschinell überprüfbar

Beispiele

– «Ein Begriff sollte nur definiert werden, wenn er ohne Definition **missverständlich, unverständlich** oder **strittig** ist.»

– «Ein definierter Begriff ist innerhalb eines Erlasses immer **im definierten Sinn** zu verwenden.»

– «Die Definition soll **zweckmässig** sein im Hinblick auf den Regelungszweck.»

– «Die Definition soll **(in einem vernünftigen Mass) präzise** sein.»

– «Die Definition soll **adressatengerecht** formuliert sein und so weit wie möglich am allgemeinen Sprachgebrauch anknüpfen.»

Zusammenfassung

Ziel

Sie haben eine Vorstellung davon, was computerlinguistische Methoden bei der maschinellen Stil-Kontrolle von Gesetzestexten erreichen könnten und was nicht.

Fazit

Je nach Abstraktionsgrad einer Richtlinie ist eine maschinelle Überprüfung

- einfach zu realisieren
- aufwändig, aber grundsätzlich machbar
- machbar, aber mit eingeschränkter Genauigkeit und Trefferquote
- nicht möglich

Cui bono?

Gesetzesredaktoren / Gesetzesredaktorinnen

Vorteil

Auch wenn bei weitem nicht alle Richtlinien maschinell überprüft werden können, so wird doch ein Grundstock an potentiellen Richtlinienverletzungen eruiert, die nicht mehr mühsam «von Hand» gesucht werden müssen.

➡ spart Zeit und Aufwand

Autoren / Autorinnen von Erlassen

Vorteil

Man muss nicht alle Richtlinien auswendig kennen, sondern wird vom System von Fall zu Fall auf die entsprechende Richtlinie hingewiesen.

➡ Richtlinien würden evtl. konsequenter angewendet