Universität Zürich[UZH]

Satzgrenzen-Erkennung in Gesetzestexten

(Programmierprojekt)

Cathrin Senn Baumgartner

21. Februar 2012

Übersicht

1. Ausgangslage
2. Programmierprojekt Satzgrenzen-Erkennung
   - Satz-Tagging
   - Evaluation
3. Erkenntnisse

Ausgangslage

Eugen-Huber-Regel

- Pro Artikel höchstens drei Absätze.
- **Pro Absatz ein Satz.**
- Pro Satz ein Gedanke.

Satzgrenzen-Erkennung bei Gesetzestexten

Satzgrenzen-Erkennung bei Gesetzestexten im Unterschied zu „normalen“ Texten:

- Spezielles Format der Texte
  - Artikel
  - Randtitel
  - Absätze
  - Aufzählungen
  - etc.

- Spezielle Arten von Sätzen
  - Halbsätze (durch Strichpunkt getrennte Teilsätze)
  - Sätze in Aufzählungen

Programmierprojekt Satzgrenzen-Erkennung

Erster Teil: Satz-Tagging

- Erkennen und Markierung von Sätzen und Halbsätzen

Zweiter Teil: Evaluation

- A. Auswertung der Eugen-Huber-Regel
- B. Auswertung der Satzgrenzen-Erkennung
- Anpassung des Codes

Programmierprojekt Satzgrenzen-Erkennung

Erster Teil: Satz-Tagging

- Erkennen und Markierung von Sätzen und Halbsätzen

Zweiter Teil: Evaluation

- A. Auswertung der Eugen-Huber-Regel
- B. Auswertung der Satzgrenzen-Erkennung
- Anpassung des Codes

Satz-Tagging: Vorgehen

Voranalyse

- Manuelle Auswertung der Punkte innerhalb von Absätzen von jedem fünften Erlass
- Sammlung der wichtigsten Beispiele, bei denen der Punkt keine Satzgrenze markiert
- Ergänzung von zufällig gefundenen, relativ häufigen Beispielen

Satz-Tagging: Vorgehen

PROGRAMMIERUNG SCHRITT 1: Marker einfügen

Markierung von Punkten ohne Satzgrenzen-Funktion

- Auslassungspunkte = ...
- Abkürzungen
  - Abkürzungen, die nie ein Satzende markieren wie *bzw.* oder *ca.*
  - Weitere Abkürzungen wie z.B. Art. in Tags und gefolgt von Zahlen:
    <ARTIKEL typ="Art." nr="160B">
    Art. 16
- Punkte in Fussnoten, Aufzählungs-Tags etc.
- Ordinalzahlen z.B. in Daten

Analog:

Markieren von Strichpunkten ohne Halbsatzgrenzen-Funktion

Satz-Tagging: Vorgehen

PROGRAMMIERUNG SCHRITT 2: Eigentliches Tagging

- Sätze
- Halbsätze
- Aufzählungen als Satz
- Sätze in Aufzählungen

Beispiel: Erlass SR 818.61

Verordnung über Transport und Beisetzung ansteckungsgefährlicher Leichen sowie Transport von Leichen vom und ins Ausland

Art. 1 Geltungsbereich

Den Bestimmungen dieser Verordnung unterliegen:

a. der Transport und die Beisetzung ansteckungsgefährlicher Leichen innerhalb der Schweiz;

b. der Transport aller Leichen vom Ausland in oder durch die Schweiz und ins Ausland.

Art. 2 Begriffsbestimmungen

[1] Unter Leiche im Sinne dieser Verordnung sind die Überreste einer verstorbenen Person zu verstehen; nicht unter diesen Begriff fällt die Leichenasche.

Beispiel: Erlass SR 818.61 in XML ohne Satz-Tagging

```
<?xml version="1.0" encoding="UTF-8"?>
- <ERLASS NR="818_61">
  - <ARTIKEL nr="1" typ="Art.">
      <ART_TITEL>Geltungsbereich</ART_TITEL>
    - <ABSATZ nr="1" quiet="true">
      - <ABS_TEXT>
          Den Bestimmungen dieser Verordnung unterliegen:
        - <AUFZAEHLUNG>
          - <AUFZ_ELEMENT nr="a.">
              <ELEM_TEXT>der Transport und die Beisetzung ansteckungsgefährlicher Leichen innerhalb der
                Schweiz;</ELEM_TEXT>
            </AUFZ_ELEMENT>
          - <AUFZ_ELEMENT nr="b.">
              <ELEM_TEXT>der Transport aller Leichen vom Ausland in oder durch die Schweiz und ins
                Ausland.</ELEM_TEXT>
            </AUFZ_ELEMENT>
          </AUFZAEHLUNG>
        </ABS_TEXT>
      </ABSATZ>
    </ARTIKEL>
  - <ARTIKEL nr="2" typ="Art.">
      <ART_TITEL>Begriffsbestimmungen</ART_TITEL>
    - <ABSATZ nr="1">
        <ABS_TEXT>Unter Leiche im Sinne dieser Verordnung sind die Überreste einer verstorbenen Person zu
          verstehen; nicht unter diesen Begriff fällt die Leichenasche.</ABS_TEXT>
      </ABSATZ>
```

Beispiel: Erlass SR 818.61 mit Satz-Tagging

```
<?xml version="1.0" encoding="UTF-8"?>
- <ERLASS NR="818_61">
  - <ARTIKEL nr="1" typ="Art.">
      <ART_TITEL>Geltungsbereich</ART_TITEL>
    - <ABSATZ nr="1" quiet="true">
      - <ABS_TEXT>
        - <SATZ nr="1">
            Den Bestimmungen dieser Verordnung unterliegen:
          - <AUFZAEHLUNG>
            - <AUFZ_ELEMENT nr="a.">
                <ELEM_TEXT>der Transport und die Beisetzung ansteckungsgefährlicher Leichen innerhalb der
                  Schweiz;</ELEM_TEXT>
              </AUFZ_ELEMENT>
            - <AUFZ_ELEMENT nr="b.">
                <ELEM_TEXT>der Transport aller Leichen vom Ausland in oder durch die Schweiz und ins
                  Ausland.</ELEM_TEXT>
              </AUFZ_ELEMENT>
            </AUFZAEHLUNG>
          </SATZ>
        </ABS_TEXT>
      </ABSATZ>
    </ARTIKEL>
  - <ARTIKEL nr="2" typ="Art.">
      <ART_TITEL>Begriffsbestimmungen</ART_TITEL>
    - <ABSATZ nr="1">
      - <ABS_TEXT>
        - <SATZ nr="1">
            <SATZ typ="Halbsatz">Unter Leiche im Sinne dieser Verordnung sind die Überreste einer verstorbenen
              Person zu verstehen</SATZ>
            ;
            <SATZ typ="Halbsatz">nicht unter diesen Begriff fällt die Leichenasche.</SATZ>
          </SATZ>
        </ABS_TEXT>
      </ABSATZ>
```

Beispiel Erlass 363.1: Satz-Tagging in Aufzählungen

Verordnung über die Verwendung von DNA-Profilen im Strafverfahren und zur Identifizierung von unbekannten oder vermissten Personen

Art. 7 Koordinationsstelle

[1] Das Departement bestimmt eines der anerkannten Labors als Koordinationsstelle.

[2] Die Koordinationsstelle hat folgende Aufgaben:

a. Sie überprüft die von den Labors erstellten Profile auf die Erfüllung der Qualitätskriterien und weiterer Vorgaben des fedpol.

b. Sie gibt die Profile in das DNA-Profil-Informationssystem (Informationssystem) ein und prüft sie auf Übereinstimmung mit den im Informationssystem vorhandenen Profilen (Profilvergleich). Das Ergebnis leitet sie an den Dienst für das Automatisierte Fingerabdruck-Identifizierungssystem (AFIS DNA Services) des fedpol weiter.

c. Sie arbeitet bei internationalen Ersuchen mit dem fedpol zusammen.

d. Sie vertritt die Interessen der anerkannten Labors gegenüber dem Bund.

Beispiel Erlass 363.1: Satz-Tagging in Aufzählungen

```
- <ARTIKEL nr="7" typ="Art.">
    <ART_TITEL>Koordinationsstelle</ART_TITEL>
  - <ABSATZ nr="1">
    - <ABS_TEXT>
        <SATZ nr="1">Das Departement bestimmt eines der anerkannten Labors als Koordinationsstelle.</SATZ>
      </ABS_TEXT>
    </ABSATZ>
  - <ABSATZ nr="2">
    - <ABS_TEXT>
      - <SATZ nr="1">
          <SATZ typ="Einleitung">Die Koordinationsstelle hat folgende Aufgaben:</SATZ>
        - <AUFZAEHLUNG>
          - <AUFZ_ELEMENT nr="a.">
            - <ELEM_TEXT>
                <SATZ typ="Aufzählung">Sie überprüft die von den Labors erstellten Profile auf die Erfüllung der Qualitätskriterien und
                  weiterer Vorgaben des fedpol.</SATZ>
              </ELEM_TEXT>
            </AUFZ_ELEMENT>
          - <AUFZ_ELEMENT nr="b.">
            - <ELEM_TEXT>
              - <SATZ typ="Aufzählung">
                  <SATZ nr="1" typ="MultiAufzählung">Sie gibt die Profile in das DNA-Profil-Informationssystem
                    (Informationssystem) ein und prüft sie auf Übereinstimmung mit den im Informationssystem vorhandenen
                    Profilen (Profilvergleich).</SATZ>
                  <SATZ nr="2" typ="MultiAufzählung">Das Ergebnis leitet sie an den Dienst für das Automatisierte Fingerabdruck-
                    Identifizierungssystem (AFIS DNA Services) des fedpol weiter.</SATZ>
                </SATZ>
              </ELEM_TEXT>
            </AUFZ_ELEMENT>
          - <AUFZ_ELEMENT nr="c.">
            - <ELEM_TEXT>
                <SATZ typ="Aufzählung">Sie arbeitet bei internationalen Ersuchen mit dem fedpol zusammen.</SATZ>
              </ELEM_TEXT>
            </AUFZ_ELEMENT>
          - <AUFZ_ELEMENT nr="d.">
            - <ELEM_TEXT>
                <SATZ typ="Aufzählung">Sie vertritt die Interessen der anerkannten Labors gegenüber dem Bund.</SATZ>
              </ELEM_TEXT>
            </AUFZ_ELEMENT>
          </AUFZAEHLUNG>
        </SATZ>
      </ABS_TEXT>
    </ABSATZ>
```

Satz-Tagging: Vorgehen

PROGRAMMIERUNG SCHRITT 3:

Entfernung der Marker

```
#============================================
# ENTFERNUNG DER MARKER
#============================================

# Auslassungspunkte wieder einfügen
$file =~ s| <auslassungs_punkte/> |\.\.\.|g;

# Abkürzungspunkte wieder einfügen
$file =~ s| <abkuerzungs_punkt/> |\.|g;

# Datenpunkte wieder einfügen
$file =~ s| <daten_punkt/> |\.|g;

# Ordinalzahlpunkte wieder einfügen
$file =~ s| <ord_zahl_punkt/> |\.|g;

# Randtitelpunkte wieder einfügen
$file =~ s| <randt_punkt/> |\.|g;

# Fussnotenpunkte wieder einfügen
$file =~ s| <fuss_punkt/> |\.|g;

# Fusstrichpunkte wieder einfügen
$file =~ s| <fuss_strichpunkt/> |;|g;

# Aufzählungselementpunkte wieder einfügen
$file =~ s| <aufz_elem_punkt/> |\.|g;
```

Programmierprojekt Satzgrenzen-Erkennung

Erster Teil: Satz-Tagging

- Erkennen und Markierung von Sätzen und Halbsätzen

Zweiter Teil: Evaluation

- A. Auswertung der Eugen-Huber-Regel
- B. Auswertung der Satzgrenzen-Erkennung
- Anpassung des Codes

Evaluation

A. Auswertung der Eugen-Huber-Regel «Pro Absatz ein Satz»?

Ausgangspunkt 1’861 Erlasse mit

- 117’749 Absätzen
- 2’149 Absätzen ohne Sätze
- 142’814 Sätzen
- 5’908 Halbsätzen

Durchschnittlich 1.2 Sätze pro Absatz
1.3 Sätze/Halbsätze pro Absatz

Maximum Durchschnittlich 2 Sätze pro Absatz in einem Erlass

Maximale Anzahl Sätze pro Absatz

Erlass 748.216.1 (Artikel zu Luftfahrzeugen)

```
- <ABSATZ nr="2">
  - <ABS_TEXT>
      <SATZ nr="1">Die Buchstaben und Zahlen müssen entweder auf beiden Seiten des Rumpfs zwischen der
        Flügelaustrittskante und der Eintrittskante des Heckleitwerks oder auf beiden Seiten des vertikalen
        Heckleitwerks angebracht werden.</SATZ>
      <SATZ nr="2">Bei mehreren vertikalen Heckleitwerkflächen sind sie nur auf der Aussenseite der äusseren
        Flächen anzubringen.</SATZ>
      <SATZ nr="3">Bei Flugzeugen und Helikoptern mit Rohr- oder Gitterrümpfen können sie an einer geeigneten
        Rumpffläche auf beiden Seiten angebracht werden.</SATZ>
      <SATZ nr="4">Die Darstellung auf der Aussenseite von Verschalungen am Rumpf angebrachter Triebwerke ist
        ebenfalls zulässig.</SATZ>
      <SATZ nr="5">Die Höhe der Buchstaben und Zahlen muss bei Flugzeugen und Helikoptern mit einer
        höchstzulässigen Abflugmasse von mehr als 5700 kg mindestens 30 cm, bei den übrigen Flugzeugen und
        Helikoptern sowie bei Motorseglern, Segelflugzeugen und Luftschiffen mindestens 20 cm
        betragen.</SATZ>
      <SATZ nr="6">Können infolge der Bauart des Luftfahrzeuges Buchstaben und Zahlen dieser Höhe nicht
        angebracht werden, beträgt die erforderliche Schrifthöhe mindestens3/5 des Rumpfdurchmessers bzw.
        der Rumpfbauhöhe, gemessen in der Mitte zwischen Flügelaustrittskante und Eintrittskante des
        Heckleitwerks.</SATZ>
      <SATZ nr="7">Bei Helikoptern entsprechender Bauart ist als Bezugsmass die Mitte des Leitwerk- oder
        Heckrotorträgers anzunehmen.</SATZ>
      <SATZ nr="8">Die Mindestschrifthöhe von 15 cm darf nicht unterschritten werden (Anhang, Fig.</SATZ>
      <SATZ nr="9">7).</SATZ>
    </ABS_TEXT>
  </ABSATZ>
```

Evaluation

B. Auswertung der Satzgrenzen-Erkennung

Ausgangslage

- Sortierung der Erlasse nach Anzahl Sätzen pro Absatz
- Manuelle Auswertung von jedem 20. Erlass der 1'861 Erlasse
- 94 Erlasse und 11'642 vom Code getaggte Sätze

Ergebnis

- Sätze: Precision 98%, Recall 97%
- Halbsätze: Precision 98%, Recall 99%
- Aufzählungen als Satz → siehe Sätze
- Sätze in Aufzählungen (verschiedene Tags): Precision ok, Recall schlecht → total 42 (sehr unterschiedliche) Fälle von denen 18 nicht erkannt wurden

Evaluation

Fehleranalyse der Satzgrenzen-Erkennung

Mögliche Fehlerquellen

- A. Erlasstext
- B. HTML-/XML-Formatierung
- C. Perl-Programm und Abkürzungsliste

False Positives

- «Falscher» Punkt (statt anderes, korrektes Satzzeichen) (A)
- Nicht erkannte Randtitel (B)
- Nicht definierte Abkürzungen (C)

Evaluation

Fehleranalyse der Satzgrenzen-Erkennung

Beispiel False Positives (Erlass 832.313.13 Art. 2 Abs. 4)

4 Unter Bolzen sind alle Gegenstände (Stifte, Dübel. Schrauben usw.) zu verstehen, die mit Bolzensetzgeräten eingetrieben werden.

```
- <ABSATZ nr="4">
  - <ABS_TEXT>
      <SATZ nr="1">Unter Bolzen sind alle Gegenstände (Stifte, Dübel.</SATZ>
      <SATZ nr="2">Schrauben usw.</SATZ>
      ) zu verstehen, die mit Bolzensetzgeräten eingetrieben werden.
    </ABS_TEXT>
  </ABSATZ>
```

Evaluation

Fehleranalyse der Satzgrenzen-Erkennung

Mögliche Fehlerquellen

- A. Erlasstext
- B. HTML-/XML-Formatierung
- C. Perl-Programm und Abkürzungsliste

False Negatives

- Kein Punkt, wo einer sein sollte (A)
- Umbruch-Tag zwischen zwei Sätzen (B, C)
- Fussnote zwischen zwei Sätzen (C)
- Verschachtelte Aufzählungen (C)

Evaluation

Fehleranalyse der Satzgrenzen-Erkennung

Beispiel False Negatives (Erlass 321.0)

Art. 68

Unterdrückung einer Beschwerde

1. Wer eine von einem Untergebenen eingereichte Beschwerde oder eine Strafanzeige, in der Absicht, sie zu unterdrücken, zurückbehält oder ganz oder teilweise beseitigt,

wer über eine Beschwerde oder eine Strafanzeige wissentlich einen unwahren Bericht erstattet,

wird mit Freiheitsstrafe bis zu drei Jahren oder Geldstrafe bestraft.

2. In leichten Fällen erfolgt disziplinarische Bestrafung.

Evaluation

Fehleranalyse der Satzgrenzen-Erkennung

Beispiel False Negatives (Erlass 321.0)

```
- <ARTIKEL nr="68" typ="Art.">
    <RANDTITEL>Unterdrückung einer Beschwerde</RANDTITEL>
    <RANDTITEL>1. Wer eine von einem Untergebenen eingereichte Beschwerde oder eine Strafanzeige, in
      der Absicht, sie zu unterdrücken, zurückbehält oder ganz oder teilweise beseitigt,</RANDTITEL>
  - <ABSATZ nr="1" quiet="true">
    - <ABS_TEXT>
      - <SATZ nr="1">
          wer über eine Beschwerde oder eine Strafanzeige wissentlich einen unwahren Bericht
          erstattet,
          <UMBRUCH type="par"/>
          wird mit Freiheitsstrafe bis zu drei Jahren oder Geldstrafe bestraft.
        </SATZ>
        <UMBRUCH type="par"/>
        2. In leichten Fällen erfolgt disziplinarische Bestrafung.
      </ABS_TEXT>
    </ABSATZ>
```

Evaluation

Anpassung des Codes

- Sätze: Precision 98%, Recall 97%
  - Verbessertes Absatz-Splitting zur Erhöhung des Recalls
  - Weitere Abkürzungen (z.B., z. B., usw., Fig.) für Precision

  → nur marginale Verbesserung möglich

- Halbsätze: Precision 98%, Recall 99%
  - Hängt mit Satz-Tagging zusammen
  - Erkennung von 3 Halbsätzen nicht implementiert

- Sätze in Aufzählungen (verschiedene Tags):
  → Sehr selten, zu unterschiedliche Fälle, nicht implementiert

Erkenntnisse

Eugen Huber

- Motto «Pro Absatz ein Satz» wird durchschnittlich relativ gut eingehalten
- In fast der Hälfte der Erlasse (866) gibt es aber mindestens einen Absatz, in dem drei Sätze oder mehr vorkommen

Satzgrenzen-Erkennung in Gesetzestexten

- Domänenspezifisches Vorgehen nötig
- Fortpflanzung von Fehlern

Fragen?