2. Übung zur Vorlesung "NLP – Analyse des Wissensrohstoffes Text" im Sommersemester 2008
– mit Musterlösungen –

Dr. Andreas Hotho, Dipl.-Inform. Dominik Benz, Wi.-Inf. Beate Krause

30. April 2008

1 Grundlagen zur Arbeit mit Text

Die Grundlage für die automatische Verarbeitung natürlicher Sprache stellen elektronische Textsammlungen und -corpora dar. Diese können in verschiedenen Formen vorliegen bzw. erzeugt worden sein, z.B.:

1. manuell erstellte Transkriptionen z.B. von Gesprächen
2. mittels OCR digitalisierte Buch- oder Printmedien
3. primär digitale Medien, z.B. Webseiten, Emails, ...

Stellen Sie kurz dar, welche besonderen Herausforderungen jede dieser Korpus-Gattungen an eine automatische Verarbeitung auf niederer Ebene (z.B. Wortgrenzenerkennung, Satzgrenzenerkennung, ...) stellt. Fällt Ihnen ein "modernes" Textmedium ein, das eine besonders "harte Nuss" für solche Analysen sein könnte?

Lösung:

1. manuell erstellte Transkriptionen z.B. von Gesprächen
   - evtl. zusätzliche Meta-Informationen zur Gesprächssituation enthalten (z.B. *"(lacht)"*, *"(räuspert sich)"*, ...) → als Junk rausfiltern
2. mittels OCR digitalisierte Buch- oder Printmedien
   - Artefakte (Sonderzeichen, ...) durch OCR-Fehler
   - falsch erkannte Wörter
3. primär digitale Medien, z.B. Webseiten, Emails, ...
   - verschiedene Metainformationen (Dokumentenheader, HTML-Markup, ...), die nichts mit dem "eigentlichen" Text zu tun haben

In allen Fällen:

- Behandlung von Kopf-/Fuss-Zeilen, Tabellen, Abbildungen, ...
- Umgang mit Gross-/Kleinschreibung

Besonders schwierig zu analysieren, wenngleich auch leicht zu akquirieren, wäre wohl ein SMS-Korpus. Durch die Begrenzung von Nachrichtenlänge und die Umständlichkeit der Eingabe sind hier viele schwer handhabbare Phänomene wie *IchKommeSpäterBisDann*, Abkürzungen (*HDL, GLG, ...*), Zeichenkonventionen (Smileys...) und vieles mehr entstanden.

2 Wortgrenzenerkennung

Eine wichtige Verarbeitungseinheit natürlicher Sprache ist das *Wort*. Was genau allerdings als Wort betrachtet wird, ist selbst unter Linguisten umstritten. Eine mögliche pragmatische Definition ist die des *graphic word* von Kucera und Francis (1967), siehe Kapitel 4.2.2 Manning/Schütze.

- Geben Sie einen regulären Ausdruck an, der dieser Definition entspricht und somit auf jedes *graphic word* matcht.
- Geben Sie einige Beispiele an, in denen diese Definition nicht gültig ist. Fallen Ihnen Textmedien ein, in denen solche Fälle gehäuft vorkommen?

Lösung:

- Regex für graphic words: `\s([a-zA-Z0-9\'\"\_\-]+)\s`
- nicht gültige Beispiele: $22.50 (aus dem Buch), ...
- solche Beispiele kommen wohl gehäuft z.B. im Web vor (Emailadressen, ...)

Wie im Buch beschrieben, stellt die variable Notation von z.B. Telefonnummern eine grosse Herausforderung for Information Extraction Methoden dar. Finden sie drei andere Beispiele, wo dies der Fall ist. Fallen Ihnen Gegenbeispiele ein, in denen Information Extraction "leicht" ist? Woran liegt das?

Lösung:

- Andere Beispiele von "variant coding":
  - ISBN-Nummern (mit / ohne Bindestrich)
  - Adressen allgemein (unterschiedliche Standards z.B. bezüglich der Reihenfolge Strasse, PLZ, ... in verschiedenen Ländern)
  - verchiedene Datums-Notationen
- Leicht ist information extraction überall da, wo ein festgelegter Standard über das Format der zu extrahierenden Information existiert - z.B. Emailadressen, URLs.

3 Satzgrenzenerkennung

Erklären Sie den Begriff *Haplologie* im Kontext der Satzgrenzenerkennung. Zu welchem "lexikalischen Problem" aus dem letzten Kapitel könnte man hier Parallelen ziehen?

Lösung: *Haplologie* (oder in diesem Fall wohl passender *Haplograhie*) beschreibt allgemein die Auslassung eines von zwei gleichgeschriebenen aufeinander folgenden Elementen in einem Satz, z.B. bei zwei Punkten an einem Satzende, wenn dieser mit einer Abkürzung endet:

```
The company is called Metal Inc.
```

Weit gedacht könnte man hier an eine Parallele zu Homographen sehen - ein Zeichen (in diesem Fall der Punkt) ist "mehrdeutig" in dem Sinne, dass es gleichzeitig zwei Funktionen erfüllt (nämlich Abkürzungszeichen und Satztrennzeichen).

4 Grammatikalisches Tagging

4.1 Markup

Vergleichen Sie HTML-Markup von Webseiten mit grammatikalischem Markup, wie es z.B. im Brown Korpus oder im Penn Treebank Korpus verwendet wird. Welche Gemeinsamkeiten gibt es, wo liegen die Unterschiede?

Lösung: Gemeinsamkeiten:

- in beiden Fällen wird bestehender Text um Metadaten erweitert
- je nach Korpus kann es Überlappungen geben, z.B. die Paragraph-Umgebung

Unterschiede:

- HTML beschreibt im Kern die logische Struktur eines Textes (Überschriften, Absätze, Tabellen, ...), während grammatikalisches Tagging meist auf Wortarten fokussiert.

5 Reguläre Ausdrücke

Betrachten Sie folgenden Text:

```
Das Fachgebiet Wissensverarbeitung im FB 16 - Elektrotechnik/Informatik wurde
im Zusammenhang mit der Einführung des Studiengangs Informatik in der
Forschungs- und Lehreinheit Informatik neu gegründet, und startete zum 1. April
2004 mit der Einrichtung einer Stiftungsprofessur der Gemeinnützigen
```

Hertie-Stiftung.

Wissensverarbeitung (Knowledge Engineering) beschäftigt sich mit der organisatorischen und technischen Unterstützung von Wissensprozessen. Wichtige Aktivitäten der Wissensverarbeitung sind das Entdecken und Strukturieren von Wissen, das Ableiten von neuem Wissen, und die Kommunikation des Wissens. Sofern es sich um anspruchsvolles Wissen handelt, spielt bei all diesen Aktivitäten der Mensch eine zentrale Rolle. Da das menschliche Denken begrifflich organisiert ist, entstand eine Reihe von Forschungsgebieten, die eine Semantik-basierte Unterstützung der Aktivitäten zum Thema haben: Knowledge Discovery, Ontologien/Metadaten, Semantic Web, Peer to Peer, Formale Begriffsanalyse, sowie Visualisierung und Interaktion. Das Fachgebiet beschäftigt sich schwerpunktmäßig damit, Methoden und Techniken auf den Schnittstellen dieser Forschungsgebiete zu entwickeln, um Synergien zu erreichen.

Das Fachgebiet Wissensverarbeitung ist Mitglied im Forschungszentrum L3S. In diesem Rahmen werden die im Fachgebiet erzielten Ergebnisse auf den Bereich des E-Learning übertragen.

Kontakt:
Universität Kassel
Fachbereich Elektrotechnik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungslehrstuhl
Wilhelmshöher Allee 73
34121 Kassel
Tel.: ++49 561 804-6250
Fax: ++49 561 804-6259

Geben Sie (wenn möglich) reguläre Ausdrücke an, um folgende Daten zu extrahieren:

1. den Namen “Wissensverarbeitung”
2. alle Wörter
3. alle Sätze
4. das erste Wort jedes Absatzes
5. die Nummer des Fachbereichs (zu erkennen an einer zweistelligen Zahl nach dem Kennwort *FB*)
6. alle Telefon- und Faxnummern
7. Postleitzahl + Ort des Instituts
8. alle Relativsätze

9. alle Substantive
10. alle Wörter, die einen Umlaut enhalten
11. alle englischen Begriffe

Lassen sich Ihre Ausdrücke auch auf andere Texte anwenden?

Hinweise: Unter *http://www.fileformat.info/tool/regex.htm* können Sie Ihre Ausdrücke direkt testen. Ebenfalls praktisch zu Übungszwecken ist die graphische IDLE-Python-Umgebung in Kombination mit `nltk.re_show`. Hinweise zur Syntax der regulären Ausdrücke finden Sie unter *http://docs.python.org/lib/re-syntax.html*, und eine weitere kurze einführung unter *http://docs.python.org/dev/howto/regex.html*.

Lösung:

1. Namen "Wissensverarbeitung": `'Wissensverarbeitung'`
2. alle Wörter: `'([a-zA-Z0-9äöüÄÖÜ\-]+)'`
3. alle Sätze: `'(?:^|\.|\s)[A-Z][^\.]+\.'` (nicht perfekt, match auch z.B. auf 1. April)
4. das erste Wort jedes Absatzes: `'(?:\n\n|\A)([a-zA-Z0-9äöüÄÖÜ\-]+)'`
5. die Nummer des Fachbereichs: `'FB (\d{2})'`
6. Telefon-/Faxnummern: `'\+\+[0-9\s\-]+'`
7. PLZ+Ort des Instituts; `'^\d{5}\s.+'`
8. Relativsätze: `'\, (der|die|das)[\.]+'` (nicht perfekt, matcht auch "..., das Ableiten von neuem Wissen...")
9. alle Substantive: `'[A-Z][a-z0-9\-äöüÄÖÜ]+'` (nicht perfekt, matcht auch auf Satzanfang)
10. Wörter, die genau einen Umlaut enthalten: `'[a-zA-Z0-9\-]*[äöüÄÖÜ][a-zA-Z0-9\-]*'`
11. alle englischen Begriffe: nicht möglich (ausser durch Enumeration)

Alle Ausdrücke sind natürlich mehr oder weniger massgeschneidert für diesen (deutschen) Text, und lassen sich nur schwer auf andere Texte / Sprachen anwenden.

6 Praxisübung - Wort- und Satzgrenzenerkennung (Abgabe 13. Mai 2008)

Das Python NLTK beinhaltet ein vollständig geparstes Exzerpt (ca. 100.000 Wörter, 3000 Sätze) des Treebank Korpus. Auf der Übungswebseite finden Sie diesen Korpus in reiner Textform (`treebank_untokenized.txt`).

- Schreiben Sie einen Tokenizer, der alle Wörter bzw. Satz- und Trennzeichen aus

dem Text extrahiert.

- Schreiben Sie einen Tokenizer, der alle Sätze aus dem Text extrahiert.
- Vergleichen Sie Ihre Ergebnisse mit der korrekt geparsten Version des Korpus. Ermitteln Sie hierzu statistische Kennwerte wie die Gesamtanzahl der extrahierten Worte / Sätze und den Prozentsatz der richtig erkannten Worte / Sätze. Fallen Ihnen noch andere Metriken ein, um den Erfolg Ihrer Methode zu messen?
- Welche Worte / Sätze bereiten dabei die größten Probleme? Finden Sie hierzu eine Erklärung.

Hinweis: Bei der reinen Textversion des Korpus wurden alle Anführungszeichen (", ') umgewandelt in ''.