1. Übung zur Vorlesung "NLP – Analyse des Wissensrohstoffes Text" im Sommersemester 2008

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

16. April 2008

1 Anforderungen an NLP anhand eines Beispieles

Eine bekannte Anwendung, in der viele der Schwierigkeiten beim Verstehen von Sprache eine Rolle spielen, ist die automatische Übersetzung von Text. Im folgenden finden Sie drei Beispiele für eine einfache Übersetzung.

Originaltext:

```
Haben Social Networks unser Kontaktverhalten verändert?
Kommunikationswissenschaftler der Uni Münster haben mit dieser
Fragestellung das Netz StudiVZ unter die Lupe genommen. Was sie
fanden: 80 Prozent nutzen das Netzwerk, um andere Profile
auszukundschaften - meistens heimlich.
```

Google Übersetzung:

```
Social Networks Have our Contact behavior changed? Communication
scientists from the University of Muenster have with this question
StudiVZ the power under the microscope. What they found: 80 percent
use the network to other profiles auszukundschaften - often secretly.
```

Babelfish Übersetzung:

```
Did Social networks change our contact behavior? Communication
scientists of the University of cathedral took the net StudiVZ with
this question under the magnifying glass. Which they found: use 80 per
cent the network, in order to explore other profiles - mostly
secretly.
```

PROMT Übersetzung:

```
Have Social Networks changed our contact behavior? Communication
scientists of the university Münster have taken the net StudiVZ under
the magnifying glass with this question. What they found: 80 percent
use the network to explore other profiles - mostly secretly.
```

- Welches System kommt einer korrekten Übersetzung am nächsten?
- Was für Fehler finden sich und womit könnten diese zusammenhängen?
- Geben Sie zwei weitere Anwendungsbeispiele und vorhandene Schwierigkeiten bei der automatischen Verarbeitung von natürlicher Sprache an.

2 Part of Speech, Morphologie, Semantik

- Geben Sie fünf Beispiele für noun-noun oder verb-noun compounds.
- Was ist der Bedeutungsunterschied in den folgenden beiden Sätzen?

  ```
  Mary defended her.
  Mary defended herself.
  ```

- Nennen Sie den Unterschied zwischen Adjunkten und Komplementen. Welchem Typ entsprechen die kursivgedruckten Satzteile?

  1. Peter washes his socks *in the bathroom.*
  2. Peter puts his socks *in the bathroom.*
  3. She goes to Church *on Sundays.*
  4. She went *to London.*
  5. Peter relies *on Mary* for help with his homework.
  6. The book is lying *on the table.*
  7. She watched him with *a telescope.*

- Identifizieren Sie mit den Tags aus dem Penn Treebank Korpus die folgenden Wortarten (eine Übersicht der Tags liegt der Übung bei).

  ```
  Our enemies are innovative and resourceful, and so are we. They never
  stop thinking about new ways to harm our country and our people, and
  neither do we.
  ```

- Können Sie Beispiele nennen, in denen eine solche Zuordnung nicht immer eindeutig ist?

- Welche Arten von lexikalischen Mehrdeutigkeiten können in einem Text vorliegen?
- Sind die folgenden Phrasen "nicht kompositional"?

  to beat around the bush, to eat an orange, help desk, not to do things by halves, big shot, have a good hand

3 Satzstruktur & Grammatiken

- Benutzen Sie die Ableitungsregeln und leiten Sie die folgenden Sätze ab.

```
S -> NP VP
NP -> Det NP
NP -> NP PP
NP -> NN
VP -> V NP
VP -> V
PP -> IN NP
PP -> IN Det NN
Det -> {der, die, das, den, einen}
NN -> {Bauer, Esel, Bauarbeiter, Bilder, Wohnungsinhaber, Flur}
V -> {brauchte, legten}
IN -> {in}
```

```
Der Bauer brauchte einen Esel.
```

```
Die Bauarbeiter legten die Bilder auf den Tisch in den Flur.
```

- Welches Problem kann bei der Rekursivität wie sie im letzten Satz vorhanden ist, auftreten? Geben Sie ein Beispiel.
- Welche Änderungen müssten Sie an der oben stehenenden Grammatik vornehmen, um deutsche "ungrammatische" Sätze wie "Das Bauarbeiter braucht den Esel in den Flur." auszuschließen?
- Entwickeln Sie eine kontextfreie Grammatik, welche für den Satz "Fed raises interest rates" mindestens drei linguistische Analysen liefert.

4 Eigenschaften von Text

- Zeigen Sie mit Hilfe eines Log-Log Plots, das für die Worthäufigkeiten aus der folgenden Tabelle annäherungsweise das Zipf Gesetz gilt.

Rang r	Wortform	Häufigkeit n	r * n
10	sich	1.680.106	
100	immer	197.502	
500	Mio	36119	
1000	Medien	19041	
5000	Miete	3755	
10000	vorläufige	1664	

- Zeigen Sie, dass das Gesetz von Mandelbrot die Vereinfachung von Zipf's Gesetz ist, wenn man $B = 1$ und $p = 0$ setzt.

5 Kollokationen & Idiome

- Was ist eine Kollokation? Geben Sie drei deutsche Beispiele.
- Was ist der Unterschied zwischen Idiomen und Kollokationen?
- Ordnen Sie die folgenden Ausdrücke nach den Kategorien Idiom, Teil-Idiom und Kollokation:

  jemandem den Fuß auf den Nacken setzen, jemandem sitzt die Angst im Nacken, reinen Tisch machen, den Tisch decken, ein rotes Tuch, mit der Wurst nach dem Schinken werfen, einen Frosch im Hals haben, etwas in den falschen Rachen bekommen, Geld abheben, in Geld schwimmen, Zeit investieren, die Zeit messen, die Zeit totschlagen!
- Können Sie Beispiele für Kollokationen nennen, die ein deutscher Muttersprachler falsch in die englische Sprache übertragen würde?
- Wie könnte man solche Kollokationen in einem Text ausfindig machen? Skizzieren Sie kurz Ihre Ideen.

6 Grundlegendes zu den Praxisübungen

1. Die Webseite zur Übung befindet sich unter `http://www.kde.cs.uni-kassel.de/lehre/ss2008/nlp/uebungen`. Dort liegt der Programmcode und ein Textkorpus `texte.zip`, der in dieser Übung zugrunde gelegt wird.
2. Machen Sie sich – soweit nicht schon geschehen – mit Phyton vertraut. Auf der Übungsseite stehen gute Referenzen für eine Einführung. In der nächsten Übung wird ein Einführungstutorial gegeben.

7 Praxisübung – Zipf's Law (Abgabe: 29.04.2008)

Schreiben Sie ein Python Programm, welches folgende Aufgaben erfüllt. Sie können dabei Funktionen aus dem Natural Language Toolkit (`http://www.nltk.org/`) zu Hilfe nehmen. Das zugehörige Buch auf dieser Webseite gibt in den ersten drei Kapiteln gute Tipps.

- Laden Sie den Textkorpus von der Übungsseite.
- Bearbeiten Sie den Korpus, so dass Sie Stopwörter entfernen und die einzelnen Token kleingeschrieben sind. Entfernen Sie die Zeichen am Satzende(".,;?!").
- Untersuchen Sie den Korpus. Implementieren Sie dabei jeweils eine Methoden, die
  - die 10 (oder allgemein n) häufigsten Wörter angibt
  - die 10 (oder allgemein n) häufigsten Wortpaare errechnet (ohne Stoppwörter)
  - die 10 (oder allgemein n) häufigsten Wortpaare relativ zur Gesamtanzahl der Vorkommen der einzelnen Wörter angibt.
- Plotten Sie die Worthäufigkeiten versus des Wortranks. (z.B. mit pylab.plot oder mit einem externen Grafikprogramm wie Gnuplot).