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

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?

  Die PROMT Übersetzung weist die wenigsten syntaktischen und semantischen Fehler auf. PROMT ist ein kommerzielles System, so dass leider nicht viel über die angewendeten Verfahren bekannt ist.

- Was für Fehler finden sich und womit könnten diese zusammenhängen?

  Google Übersetzung:

  - Groß- und Kleinschreibung missachtet (z.B. Kontaktverhalten –> Contact behavior, lexikalischer Fehler)
  - Wortreihenfolge falsch – Frage nicht erkannt (Syntaxfehler)
  - Wortwörtliche Übersetzung –> unter die Lupe nehmen (lexikalisch/semantischer Fehler, Nichterkennung eines Idioms)

  Babelfish Übersetzung:

  - Münster mit cathedral übersetzt (Eigenname wurde nicht erkannt, semantischer Fehler)
  - Wortwörtliche Übersetzung –> unter die Lupe nehmen (lexikalisch/semantischer Fehler, Nichterkennen eines Idioms)
  - Which mit What verwechselt (Syntaxfehler)
  - Falsche Wortreihenfolge: use 80 per cent the network (Grammatik, Syntaxfehler)

  PROMT Übersetzung:

  - Wortwörtliche Übersetzung (unter die Lupe nehmen, lexikalischer Fehler)
  - university Münster –> Präposition weggelassen

- Geben Sie zwei weitere Anwendungsbeispiele und vorhandene Schwierigkeiten bei der automatischen Verarbeitung von natürlicher Sprache an.

  - Information Retrieval: Beim Ranken von natürlich sprachigen Dokumenten können u.a. die Worthäufigkeiten im Text eine Rolle spielen. Dabei müssen morphologische Varianten von Wörtern auf einen gemeinsamen Stamm heruntergebrochen werden. Dies kann dazu führen, dass Wörter mit unterschiedlichen Bedeutungen zusammengefasst werden. Zum Beispiel könnten *gallery* und *gall* beide auf den gemeinsamen Stamm *gall* reduziert werden.

– Bei der Textzusammenfassung (Text Summarization) soll mit automatischen Verfahren die relevanten Informationen eines Textes zusammengefasst werden. Neben einer statistischen Textanalyse (e.g. häufige Wörter, häufige Kollokationen etc.) kann auch die Semantik (Erkennung von Synonymen) eine Rolle spielen.

2 Part of Speech, Morphologie, Semantik

- Geben Sie fünf Beispiele für noun-noun oder verb-noun compounds.

  web server, car park, machine learning, washing machine, swimming pool

- Was ist der Bedeutungsunterschied in den folgenden beiden Sätzen?

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

  "her" ist ein Personalpronomen und bezieht sich auf ein eine aussenstehende Person. "herself" ist ein Reflexivpronomen und bezieht sich auf das nahestehende Subjekt, Mary.

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

  Komplemente und Adjunkte folgen auf den Kopf eines Satzes. Komplemente werden durch den Kopf festgelegt (obligatorische Ergänzungen), Adjunkte können beliebig hinzugefügt werden (oft Ort- und Zeitangaben, freie Angaben).

  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.*

  1 Adjunkt, 2 Komplement, 3 Adjunkt, 4 Komplement, 5 Komplement, 6 Komplement, 7 Adjunkt

- 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.
```

Our (PRP$) enemies (NNS) are (VBP) innovative (JJ) and (CC) resourceful (JJ),(,) and (CC) so (RB) are (VB) we (PRP). (.) They (PRP) never (RB) stop (VB) thinking (VBG) about (IN) new (JJ) ways (NNS) to (TO) harm (VB) our (PRP$) country (NN) and (CC) our (PRP$) people (NN), and (CC) neither (DT) do (VB) we (PRP) .(.)

- Können Sie Beispiele nennen, in denen eine solche Zuordnung nicht immer eindeutig ist?
  - Das Wort "book" kann sowohl ein Nomen (*hand me that book*) oder ein Verb (*book that flight*) sein
  - that kann entweder eine Determiner (*Does that flight serve dinner?*) oder eine Konjunktion (*I thought that your flight was earlier.*) sein.
- Welche Arten von lexikalischen Mehrdeutigkeiten können in einem Text vorliegen?

  Homonyme (siehe oben)

  Polyseme:
  - In *The house is at the foot of the mountains* bezieht sich "foot" auf den unteren Teil eines Berges, in *One of his shoes felt too tight for his foot* bezieht sich "foot" auf den unteren Teil eines Beines.
- 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

  Nicht kompositionale Ausdrücke unterscheiden sich von kompositionalen Ausdrücken dadurch, dass sie nicht aus der Bedeutung der Einzelteile errechnet werden können.
  1 yes, 2 no, 3 yes, 4 yes, 5 yes, 6 no

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.

S
-> NP VP
-> Det NN V NP
-> Det NN V Det NN

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

S
-> NP VP
-> Det NP V NP
-> Det NP V NP PP
-> Det NN V NP PP PP
-> Det NN V NP IN Det NN IN Det NN
```

- Welches Problem kann bei der Rekursivität wie sie im letzten Satz vorhanden ist, auftreten? Geben Sie ein Beispiel.

  Durch Rekursivität kann ein nicht terminales Symbol in viele kleinere Einheiten zerlegt werden. Dadurch könnten Worte, die syntaktisch verbunden sind, getrennt werden (Nicht-lokale Abhängigkeiten). Ein Beispiel für solche Abhängigkeiten ist zum Beispiel die Anpassung von Geschlecht und Zahl bei Subjekt Prädikat (Max who experimented with the phenomenon of flying flies was given an award.).

- Welche Änderungen müssten Sie an der oben stehenenden Grammatik vornehmen, um deutsche "ungrammatische" Sätze wie "Das Bauarbeiter braucht den Esel in den Flur." auszuschließen?

  - Statt einer einheitlichen Kategorie V für Verben wird eine Differenzierung transitive, nicht-transitive und Verben mit bestimmten Präpositionen notwendig.
  - Einführung von Genus-Informationen (männlich, weiblich), um nicht wohlgeformte Ausdrücke wie **die Esel* zu eliminieren. Dies muss sowohl für Artikel (Det) als auch für Nomina durchgeführt werden.

– Weiterhin braucht man Kasusinformationen, z.B. die Unterscheidung in Nominativ und Genitiv.

- Entwickeln Sie eine kontextfreie Grammatik, welche für den Satz "Fed raises interest rates" mindestens drei linguistische Analysen liefert.

```
S -> NP VP
NP -> N
NP -> N N
NP -> N N N
VP -> V NP
```

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	

Rang r	Wortform	Häufigkeit n	r * n	log r	log n	Steigung
10	sich	1.680.106	16.801.060	1	6.23	
100	immer	197.502	19.750.200	2	5.30	-0,93
500	Mio	36119	18.059.500	2.70	4.56	-1,06
1000	Medien	19041	19.041.000	3.00	4.28	-0,93
5000	Miete	3755	18.775.000	3.70	3.58	-1,00
10000	vorläufige	1664	16.640.000	4	3.22	-1,20

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

  Gesetz von Mandelbrot: $f = P(r + p)^{-B}$.

  Mit $B = 1$ und $p = 0$ gilt: $f = P(r + 0)^{-1} = P\frac{1}{r}$. Also gilt $fr = P$ mit $P = k$.

5 Kollokationen & Idiome

- Was ist eine Kollokation? Geben Sie drei deutsche Beispiele. Eine Kollokation bezeichnet das gehäufte benachbarte Auftreten von Wörtern. Beispiele sind "schwarzer Kaffee", "mittlerer Osten" oder "steife Brise".
- Was ist der Unterschied zwischen Idiomen und Kollokationen?

  Genau wie Idiome sind auch Kollokationen relativ festgelegte Ausdrücke. Ihre Bedeutung kann allerdings von den einzelnen Satzteilen abgeleitet werden. Z.B. ist "Tür - klopfen" eine Kollokation, aber kein Idiom. Es handelt sich um zusammengehörende Begriffe (man würde nicht sagen "an die Tür hauen"). Zusätzlich lässt sich der Sinn des Ausdrucks aus den einzelnen Begriffen ableiten.
- 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!

  1 Idiom, 2 Idiom, 3 Idiom, 4 Kollokation, 5 Teil-Idiom, 6 Idiom, 7 Idiom, 8 Teil-Idiom, 9 Kollokation, 10 Idiom, 11 Kollokation, 12 Kollokation, 12 Idiom
- Können Sie Beispiele für Kollokationen nennen, die ein deutscher Muttersprachler falsch in die englische Sprache übertragen würde?
  - einen Vortrag halten: to hold a talk, aber: to give a talk
  - ein Photo machen: to make a picture, aber: to take a picture
- Wie könnte man solche Kollokationen in einem Text ausfindig machen? Skizzieren Sie kurz Ihre Ideen.
  - Absolute Häufigkeiten aus den Texten extrahieren und schauen, welche Wörter mit welchen am Häufigsten vorkommen.
  - Aufbauend auf Frequenzanalyse: Syntaxmuster analysieren, z.B. "degrees of freedom" (NPN)
  - Anwendung eines probabilistischen Sprachmodells, bei dem Wortfolgen bestimmte Wahrscheinlichkeiten zugeordnet werden.

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).