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

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

28. Mai 2008

1 Part-of-Speech Tagging

1.1 Grundlagen

Part-of-Speech Tagging stellt eine wichtige Zwischenstufe auf dem Weg dar, natürliche Sprache automatisch zu verarbeiten.

1. Erklären Sie mit Ihren eigenen Worten, worum es sich bei einem "Part-of-Speech"handelt und worum es beim Part-of-Speech Tagging geht.
2. Welchen Mehrwert kann ein korrekt getaggter Text im Vergleich zu einem nicht-getaggten Text haben? Zählen Sie einige Anwendunge auf, die diesen ausnutzen.
3. Welche Informationen können für das POS-Tagging verwendet werden? Welche Informationsquelle ist besonders nützlich und weshalb?

1.2 Visible Markov Model Taggers

Ein Ansatz zum POS-Tagging sind Visible Markov Models (VMM). Erklären Sie kurz, was in diesen Modellen den Parts-of-Speech entspricht.

1. Welche vereinfachenden Annahmen über die Struktur von Sprache liegen der Anwendung von Markov-Modellen zum POS-Taggen zugrunde?
2. Welche Phänomene natürlicher Sprache widersprechen diesen Annahmen?

Folgende Daten zur Abfolge bestimmter POS Tags und deren Häufigkeit für bestimmte Worte stammen aus dem Brown Corpus (siehe Manning/Schütze, S. 348f):

1. Beschreiben Sie in Ihren eigenen Worten, was passiert, wenn ein Visible Markov Model zum Taggen auf einen bestimmten Text trainiert wird. Was genau wird dabei "gelernt"?

Tabelle 1: Häufigkeiten der Abfolge einiger POS Tags aus dem Brown Corpus (idealisiert)

	zweites Tag					
Erstes Tag	AT	BEZ	IN	NN	VB	PERIOD
AT	0	0	0	48636	0	19
BEZ	1973	0	426	187	0	38
IN	43322	0	1325	17314	0	185
NN	1067	3720	42470	11773	614	21392
VB	6072	42	4758	1476	129	1522
PERIOD	8016	75	4656	1329	954	0

Tabelle 2: Häufigkeiten bestimmter Tags für einige Wörter aus dem Brown Corpus (idealisiert)

	AT	BEZ	IN	NN	VB	PERIOD
bear	0	0	0	10	43	0
is	0	10065	0	0	0	0
move	0	0	0	36	133	0
on	0	0	5484	0	0	0
president	0	0	0	382	0	0
progress	0	0	0	108	4	0
the	69016	0	0	0	0	0
.	0	0	0	0	0	48809
total (all words)	120991	10065	130534	134171	20976	49267

2. Berechnen Sie die Wahrscheinlichkeit der beiden Taggings `AT NN BEZ IN AT NN` sowie `AT NN BEZ IN AT VB` für den Satz *The bear is on the move*. Machen Sie dabei klar, an welcher Stelle das Markov Model trainiert wird.
3. Nachdem ein VMM trainiert wurde, kann es zum Taggen von unbekannten Corpora eingesetzt werden. Diskutieren Sie, inwiefern es sich dann noch um ein "visible" Markov Model handelt. Welche Verfahren kennen Sie, um mit einem fertig trainierten Modell zu taggen? Welches Verfahren ist effizient?
4. Verwenden Sie die Daten aus den Tabellen 1.2 und 1.2, um mittels des Viterbi-Algorithmus den Satz *The bear is on the move* zu taggen.
5. Ein grosses Problem beim POS-Tagging allgemein sind unbekannte Wörter. Skizzieren Sie kurz Lösungsansätze.

1.3 Hidden Markov Model Taggers

Auch Hidden Markov Models (HMM) werden für das POS-Tagging eingesetzt.

1. Unter welchen Voraussetzungen wird man ein Hidden Markov Model (HMM) zum POS-Tagging verwenden?
2. Das POS-Tagging mittels HMM teilt sich in drei Abschnitte auf: Initialisierung, Training, Tagging. Beschreiben Sie kurz, was in den drei Abschnitten passiert. Wo liegen die Unterschiede / Parallelen zu den Visible Markov Models?

1.4 Transformation-Based Taggers

Markov Models liefern gute Ergebnisse, sind aber für manche Eigenschaften natürlicher Sprache zu starr. Eine andere Methode, die ein breiteres Spektrum von lexikalischen und syntaktischen Regularitäten erfassen kann, sind transformations-basierte Tagging-Ansätze.

1. Welche Komponenten hat ein transformations-basierter Ansatz? Welche Art von Input is notwendig?
2. Bei der Initialisierung wird beim transformations-basierten Tagging im Traingsdatensatz zunächst jedem Wort ein initiales Tag zugewiesen. Im Buch ist die Strategie beschrieben, hierfür das häufigste Tag (laut einem Wörterbuch) zu verwenden. Eine andere Möglichkeit wäre, jedem Wort das gleiche Tag (z.B. *NN*) zuzuweisen, oder die Trainingsdaten von einem externen Tagger initial taggen zu lassen. Welche Vor- und Nachteile hat jede dieser Alternativen?

Tabelle 3: Vergleich verschiedener POS-Tagging Ansätze

	VMM	HMM	transformationsbasiert
nötiger Input			
Komplexität			
Annahmen			
Gefahr des Overtrainings?			
...			

1.5 Vergleich

Vervollständigen Sie die Tabelle 1.5, die zum Vergleich der vorgestellten POS-Tagging Methoden dient. Fallen Ihnen noch andere Vergleichs-Dimensionen ein?

2 Praxisübung - POS Tagging

Das NLTK-Toolkit sind einige POS-Tagger enthalten, unter anderem ein HiddenMarkov-ModelTagger. Dieser kann auch trainiert werden.

- Implementieren und trainieren Sie einen HMM Tagger für eine Teilmenge des Brown-Corpus. Bestimmen Sie die Grösse der Teilmenge so, dass das Training in vertretbarer Zeit abläuft. Das Training soll hierbei unüberwacht mittels des Baum-

Welch-Algorithmus stattfinden (dieser ist ebenfalls im NLTK-Toolkit verfügbar). Verwenden Sie dazu mindestens folgende Initialisierungen:

  - Ein Markov-Model, das mittels einem überwachten Lernverfahren gelernt wurde
  - Eine Initialisierung mittels Jelinek's Methode

- Wenden Sie Ihren Tagger auf folgende Testdaten an und vergleichen Sie die Accuracy des Taggings:
  - dieselbe Teilmenge, auf der der Tagger trainiert wurde
  - eine andere Teilmenge, ggf. den ganzen Korpus

  Wie verändert sich die Qualität Ihrer Ergebnisse in Abhängigkeit der Initialisierung? Wie verändert sich die Qualität Ihrer Ergebnisse in zwischen mehreren Testläufen mit derselben Initialisierung?

- *Zusatzaufgabe* (freiwillig): Vergleichen Sie die Ergebnisse des gelernten HMM-Taggers mit in NLTK verfügbaren Standard-Taggern (Brill, Bigramm, ...) oder eine Kombination aus diesen.

Hinweis: Im NLTK-Quellcode `nltk/tag/hmm.py` finden Sie bereits einige Beispiele, die Sie als Einstieg verwenden können.