Einführung in die statistische Sprachverarbeitung

Detlef Prescher

May 26, 2007

Hausaufgabe 7

(i) Lade Dir die Datei `http://www.cl.uni-heidelberg.de/~prescher/tmp/data.tar.gz` herunter, und entpacke sie unter einem Linux/Unix-Betriebssystem. In der Datei `wsj.trainfragment` wirst Du ca. 2000 Bäume aus der Penn-Treebank finden (Format: ein Baum pro Zeile). Ferner wirst Du das GUI `viewtree` finden, welches Du wie folgt dazu benutzen kannst, um alle Bäume anzuschauen (eventuell noch TCL/TK installieren):

```
cat wsj.trainfragment | viewtree
```

Schaue Dir einige hundert Bäume an, um Dir einen Eindruck von der Penn-Treebank zu verschaffen. Besorge Dir weitere Informationen zur Penn-Treebank aus dem Internet. (Bedeutung der verwendeten POS-Tags, Nonterminals, usw.). Wieviele Bäume sind in dem Fragment exakt vorhanden?

(ii) Benutze eine geeignete Programmiersprache, um die Vorkommen aller POS-Tags in `wsj.trainfragment` zu extrahieren. Benutze die Linux/Unix-Befehle `sort` und `uniq`, um alle POS-Tags mit ihren Vorkommenshäufigkeiten auszugeben. Wie gross ist die durchschnittliche Satzlänge in dem gegebenen Fragment?

(iii) Gib für jedes der 12 häufigsten POS-Tags ein Satzfragment aus `wsj.trainfragment` an, welches die Bedeutung des jeweiligen POS-Tags besonders gut illustriert. Welches ist das seltenste POS-Tag? Gib auch hier ein Beispiel an.

(iv) Benutze eine geeignete Programmiersprache, um die Vorkommen aller Nonterminale in `wsj.trainfragment` zu extrahieren. Lösche anschliessend aus allen Nonterminalen die Funktionsinformation (d.h. anstelle von `NP-SBJ`, `ADVP-TMP`, `PP-LOC-PRD`, etc. sollen im weiteren Verlauf lediglich die syntaktischen Kategorien `NP`, `ADVP`, `PP` benutzt werden). Entferne aus Deiner Liste schliesslich auch den TOP-Knoten, sowie sämtliche POS-Tags. Verwende dann wie in Aufgabenteil (ii) die Linux/Unix-Befehle `sort` und `uniq`, um alle verbliebenen Nonterminale mit ihren Vorkommenshäufigkeiten auszugeben. Wieviele dieser Nonterminale hat ein Baum im Durchschnitt? Aus wievielen Knoten besteht ein Baum im Durchschnitt?

(v) Gib die Bedeutung der 12 häufigsten Nonterminale aus Aufgabenteil (iv) an. Gib für jedes dieser Nonterminale ein Baumfragment aus `wsj.trainfragment` an, welches die Bedeutung dieses Nonterminals besonders gut illustriert.

Abgabetermin: Dienstag, 4. Juni

Besprechung: Mittwoch, 6. Juni