Lehrstuhl für Informatik 6
der RWTH Aachen
Prof. Dr.-Ing. H. Ney

Aachen, den 14. April 2011
M. Huck, M. Freitag

Software-Projektpraktikum Maschinelle Übersetzung 1. Übung

Bemerkungen:

Besondere Ankündigungen erfolgen normalerweise per E-Mail oder während der zweiwöchentlichen Vorbesprechungen. Ansonsten sind auf der Webseite
`http://www-i6.informatik.rwth-aachen.de/web/Teaching/LabCourses/SS11/Softwareprojektpraktikum/`
aktuelle Informationen abgelegt.

Thema:

Für die statistische Übersetzung arbeiten wir mit bilingualen Satzpaaren $f_1^J = f_1 \dots f_J$ (Quellsatz) und $e_1^I = e_1 \dots e_I$ (Zielsatz). Eine Zuordnung zwischen Wörtern der beiden Sätze wird als Alignment bezeichnet.
Auf den Webseiten des Instituts finden Sie die drei Dateien `f`, `e` und `Alignment`. Die Sätze in `f` und `e` entprechen sich jeweils zeilenweise. Das Alignment hat dabei folgendes Format:

```
          # (Indizes beginnen alle bei 0)
SENT: 2   # Informationen fuer das dritte Satzpaar

S 0 1     # dem ersten Wort im Quellsatz
          # ist das zweite Wort im Zielsatz zugeordnet
S 0 2     # und das dritte Wort auch
...

SENT: 3
...
```

Aufgabe:

1. Berechnen Sie folgende Korpus-Statistiken für Quell- und Zielseite der Trainingsdaten (Dateien `f` und `e`): Anzahl der laufenden Wörter, Größe des Vokabulars, durchschnittliche Satzlänge. Sie können dazu Unix-Tools wie `awk` und `sed` einsetzen.

2. Schreiben Sie ein Programm, welches über Kommandozeilenoptionen drei beliebige Dateien für jeweils die Quellsätze, die Zielsätze und das Alignment einlesen kann. Geben Sie satzweise alle einander zugeordneten Wörter aus.

3. Implementieren Sie die Klasse `Lexicon`. Die Klasse soll eine bidirektionale Zuweisung von Strings und Zahlenschlüsseln realisieren, um die Verarbeitung effizient zu halten.

4. Implementieren Sie die Klassen `Paircount` und `Singlecount`, die jeweils für ein Wort bzw. Wortpaar deren Häufigkeit mitzählen.

5. Berechnen Sie nun für alle gesehenen Wortpaare die relativen Häufigkeiten (relFreq), einmal gemessen an der Quellsprache, und dann an der Zielsprache. Geben Sie den errechneten Wert als negativen Logarithmus an.

   Die Ausgabe soll dabei folgendes Format haben:

   relFreq_f relFreq_e # sourceWord # targetWord

6. Schreiben Sie ein zugehöriges Makefile, das beim Aufruf `make all` ein Programm mit Namen `singleWordExtract` erzeugt. Der Aufruf von `make clean` soll alle automatisch erzeugten Dateien löschen.

Abnahmetermin: Donnerstag, 28. April, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (27. April, 18:00 Uhr) an

huck@i6.informatik.rwth-aachen.de.

Am Donnerstag erläutern Sie uns dann Ihre Lösungen und demonstrieren Ihre Programme.