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

Aachen, den 6. Mai 2010
M. Huck, D. Stein

# Software-Projektpraktikum Maschinelle Übersetzung 2. Übung

## Thema:

Für unsere statistische Übersetzung eines gegebenen Quellsatzes $f_1^J$ suchen wir von allen theoretisch möglichen Zielsätzen $e_1^I$ diejenige Hypothese $\hat{e}$, die uns die Wahrscheinlichkeit

$$\hat{e} = \text{argmax}_{e_1^I} Pr(e_1^I | f_1^J) \tag{1}$$

maximiert. Wir bezeichnen diesen Vorgang als *Suche*.
Um einen ersten Suchalgorithmus zu implementieren, treffen wir stark vereinfachende Annahmen:

- Die Übersetzung verläuft monoton, das heißt, es sind keine Umordnungen des Quellsatzes von Nöten.
- Der Zielsatz hat die gleiche Wortlänge wie der Quellsatz.

## Aufgabe:

1. Schreiben Sie ein Programm, das zunächst die Übersetzungstabelle aus der vorherigen Aufgabe, und ein zu übersetzendes Dokument einlesen kann.
2. Legen Sie die Datenstrukturen `HypothesisNode` und `PartialTranslation` an. Die einzelnen Einträge sollen folgende Werte beinhalten:
   - HypothesisNode
     - die aktuell besten Kosten
     - einen Vektor von eingehenden PartialTranslations
   - PartialTranslation
     - Kosten der Teilübersetzung
     - "Übersetzung" der Teilübersetzung
     - ursprüngliche HypothesisNode
3. Schreiben Sie nun einen einfachen, monoton arbeitenden Suchalgorithmus.
   - (Initialisierung) Bei einem Satz der Länge $J$, legen Sie einen Vektor von $J+1$ HypothesisNodes an.
   - (Schleife) Durchlaufen Sie die Quellwörter, und legen Sie einen Übergang in die nächste Node mit allen möglichen Übersetzungen des aktuellen Wortes.

- (Update) Aktualisieren Sie dabei die besten Kosten. Achtung: Da Sie im negativ-logarithmischen Bereich arbeiten, müssen die Kosten addiert werden, nicht multipliziert, außerdem müssen Sie die Kosten minimieren.

4. Geben Sie für jeden zu übersetzenden Satz die Hypothese mit den niedrigsten Kosten aus.

5. Implementieren Sie die A*-Suche und geben Sie für jeden zu übersetzenden Satz die zehn besten Hypothesen aus.

## Abnahmetermin: Donnerstag, 20. Mai, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (19. Mai, 18:00 Uhr) an

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

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