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

Aachen, den 28. April 2011
M. Huck, M. Freitag

Software-Projektpraktikum Maschinelle Übersetzung 2. Übung

Thema:

Für unsere statistische Übersetzung eines gegebenen Quellsatzes f_1^J wollen wir von allen theoretisch möglichen Zielsätzen e_1^I diejenige Hypothese $\hat{e}$ mit der maximalen Wahrscheinlichkeit finden:

$$\hat{e} = \text{argmax}_{e_1^I} Pr(e_1^I|f_1^J) \quad (1)$$
$$= \text{argmax}_{e_1^I} Pr(e_1^I) \cdot Pr(f_1^J|e_1^I) \quad (2)$$

Wir bezeichnen diesen Vorgang als *Suche*.
Um einen ersten Suchalgorithmus zu implementieren, wenden wir einige stark vereinfachende Restriktionen an:

- Der Algorithmus arbeitet wortweise.
- Die Übersetzung verläuft monoton von links nach rechts. Umordnungen in der Wortabfolge werden nicht vorgenommen.
- Der Zielsatz hat die gleiche Länge wie der Quellsatz, d.h. $I = J$.
- Die Sprachmodellwahrscheinlichkeit aus Gleichung 2 wird nicht berücksichtigt.

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, mit dem Sie die auf der Webseite zur Verfügung gestellten Test-Daten übersetzen.

   - (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 den nächsten 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.
   - (Pruning) Weil die Anzahl der möglichen Übersetzungen schnell zu groß werden kann, soll der Decoder in jedem Schritt nur die n PartialTranslations mit den besten (=niedrigsten) Kosten erzeugen. Enthält die Übersetzungstabelle mehr als n Übersetzungsoptionen für das aktuelle Quellwort, dann werden die schlechteren nicht betrachtet. Der Pruning-Parameter n sollte nicht hart kodiert, sondern frei einstellbar sein. Sinnvolle Größen für n liegen im Bereich 20 bis 100.

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

5. (Pruning der Übersetzungstabelle) Analog zum Pruning während der Suche kann der Suchraum eingeschränkt werden, indem schlechtere Übersetzungsoptionen bereits aus der Phrasentabelle entfernt werden. Manipulieren Sie Ihre Phrasentabelle so, dass für jedes Quellwort nur noch die n besten Übersetzungsoptionen enthalten sind, und testen Sie diese neue Tabelle mit ihrem Suchalgorithmus.

6. Implementieren Sie die A*-Suche und geben Sie für jeden zu übersetzenden Satz die zehn (oder wahlweise hundert) besten Hypothesen aus.

Abnahmetermin: Donnerstag, 12. Mai, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (11. Mai, 18:00 Uhr) an

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

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