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

Aachen, den 10. Juni 2010
M. Huck, D. Stein

# Software-Projektpraktikum Maschinelle Übersetzung 4. Übung

## Thema:

Durch die Wort-für-Wort-Übersetzung entsteht das Problem, dass Kontext-Informationen verloren gehen. In der Praxis wird daher üblicherweise ein phrasenbasiertes Übersetzungsmodell angewandt. Für ein wortaligniertes Satzpaar sind alle Teilstrings gültige Phrasen, wenn auf Quellseite keine anderen Wörter der Zielphrase, und auf Zielseite keine anderen Wörter der Quellphrase zugeordnet sind, d.h. alle Zuordnungen sind innerhalb eines festen Blocks.
In dieser Aufgabe schreiben wir ein Programm, das die Extraktion auf Phrasenebene erweitert. Wir führen dann die Suche nach der besten Übersetzung mit Phrasen durch.
Die entstehenden Hypothesen werden danach mit einem Sprachmodell bewertet, um eventuell andere Hypothesen zu bevorzugen.

## Aufgabe:

1. Erweitern Sie Ihr Training, d.h. Extraktion und Berechnung der relativen Häufigkeiten, auf Phrasen. Benutzen Sie (selbstimplementierte) Präfixbäume für das Speichern der Daten.

2. Erweitern Sie Ihr Übersetzungsprogramm auf Phrasen.

3. Übersetzen und evaluieren Sie die Test-Sätze. Vergleichen Sie die Ausgabe mit der einzelwortbasierten Übersetzung. Um wieviel haben sich die Fehlermaße bzw. BLEU verbessert? Welche Fehler werden jetzt offensichtlich weniger gemacht? Geben Sie auch wieder die zehn besten Hypothesen aus.

4. Laden Sie das SRI Language Modeling Toolkit herunter und installieren Sie es (frei verfügbar für nicht gewerbliche Zwecke unter `http://www.speech.sri.com/projects/srilm/`). Erzeugen Sie damit ein Bigramm-Sprachmodell mit Kneser-Ney Smoothing auf allen Zielsätzen Ihres Trainingskorpus.

5. Führen Sie ein sogenanntes Rescoring mit Hilfe des Sprachmodells durch. Nutzen Sie dafür die vom SRI-Toolkit zur Verfügung gestellte Library.

   - Legen Sie eine Ngram und eine Vocab Klasse aus den gleichnamigen Klassen der SRI-Library an.
   - Leiten Sie Ihr Vokabular in das SRI-Vokabular um.
   - Laden Sie das Bigramm-Sprachmodell, das Sie in der vorherigen Aufgabe trainiert haben.

- Bewerten Sie die zehn besten Übersetzungshypothesen Ihres phrasenbasierten Decoders mit dem Sprachmodell. Addieren Sie das Ergebnis zu den Übersetzungskosten. (Achtung: Das SRI LM Toolkit erzeugt log-Scores, keine negativen log-Scores.)

Welche der Hypothesen hat nun die besten Scores? Vergleichen Sie die Ausgaben manuell und automatisch. Wiederholen Sie die Aufgabe mit den hundert besten Hypothesen.

## Abnahmetermin: Donnerstag, 24. Juni, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (23. Juni, 18:00 Uhr) an

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

Am Donnerstag erläutern Sie uns dann Ihre Lösungen und demonstrieren Ihre Programme im CIP-Pool.