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

Aachen, den 30. Juni 2011
M. Huck, M. Freitag

Software-Projektpraktikum Maschinelle Übersetzung 6. Übung

Thema:

In dieser Aufgabe führen wir ein sogenanntes n-best-Reranking durch. Wir trainieren ein Bigramm-Sprachmodell für Englisch und setzen dieses zusätzliche, nicht im Decoding verwendete Modell ein, um die n besten vom Decoder ausgegebenen Hypothesen erneut zu bewerten. Eventuell werden dann andere Hypothesen bevorzugt, und wir erhalten eine andere beste einzelne Übersetzung.

Aufgabe:

1. 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.

2. Führen Sie ein sogenanntes Reranking 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, 14. Juli, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (13. Juli, 18:00 Uhr) an

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

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