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

Aachen, den 24. Juni 2010
M. Huck, D. Stein

Software-Projektpraktikum Maschinelle Übersetzung 5. Übung

Thema:

Neben dem n-gram Sprachmodell und dem phrasenbasierten Übersetzungsmodell, die wir bis jetzt eingesetzt haben, sind viele weitere Wissensquellen denkbar, die zur Verbesserung der Übersetzungsqualität beitragen könnten.
In dieser Aufgabe integrieren wir weitere Modelle in unser statistisches maschinelles Übersetzungssystem. Da nicht alle Modelle gleich nützlich bzw. zuverlässig sind, sollen die Scores mit Skalierungsfaktoren unterschiedlich gewichtet werden. Dazu kombinieren wir die Modelle in einem log-linearen Framework:

$$\hat{e}_1^{\hat{I}} = \arg\max_{e_1^I} \left\{ p(e_1^I | f_1^J) \right\} \tag{1}$$

$$= \arg\max_{e_1^I} \left\{ \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{\tilde{e}_1^I} \exp\left(\sum_{m=1}^M \lambda_m h_m(\tilde{e}_1^I, f_1^J)\right)} \right\} \tag{2}$$

$$= \arg\max_{e_1^I} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\} \tag{3}$$

für Skalierungsfaktoren λ_m und Funktionen $h_m(e_1^I, f_1^J)$.
Die Skalierungsfaktoren stellen wir mit geeigneten Algorithmen so ein, dass die Ausgabe des Decoders direkt bezüglich einem der automatischen Fehlermaße optimiert wird.

Aufgabe:

1. Erweitern Sie Ihre Übersetzungstabelle um Scores mit

   - Phrase Penalty, $= 1$ für alle Einträge
   - Word Penalty, $= |e|$
   - einem Single Count Bit, $= 1$ falls $N(e)$ und $N(f) > 1$, sonst $= 0$
   - Source-Target Ratio, $= \frac{|f|}{|e|}$

   Passen Sie dazu Ihre Phrasenextraktion entsprechend an.

2. Setzen Sie die zusätzlichen Modelle in Ihrem Decoder ein. Erweitern Sie Ihren Suchalgorithmus um (als Kommandozeilenparameter übergebene) Skalierungsfaktoren, mit denen die verschiedenen Scores multipliziert werden.

3. Implementieren Sie den Downhill-Simplex Algorithmus und optimieren Sie Ihre Skalierungsfaktoren auf BLEU.

4. Implementieren Sie MERT und optimieren Sie Ihre Skalierungsfaktoren auf BLEU.

5. Beurteilen Sie die Software-Architektur einer anderen Gruppe dieses Praktikums. Achten Sie auf Dokumentation, Verständlichkeit, Modularität und allgemeine Sauberkeit des Codes und verfassen Sie dazu eine Zusammenfassung in der Größenordnung von einer maschinell geschriebenen DIN A4 Seite.

Abnahmetermin: Donnerstag, 8. Juli, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (7. Juli, 18:00 Uhr) an

stein@informatik.rwth-aachen.de.

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