

Software-Projektpraktikum Maschinelle Übersetzung

5. Übung

Thema:

Unser statistisches maschinelles Übersetzungssystem kombiniert mehrere Modelle in einem log-linearen Framework:

$$\hat{e}_1^I = \arg \max_{e_1^I} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (1)$$

für Skalierungsfaktoren λ_m und Funktionen $h_m(e_1^I, f_1^J)$, $m = 1, \dots, M$. Nicht alle der Modelle $h_m(e_1^I, f_1^J)$ sind gleich zuverlässig bzw. nützlich zum Erreichen einer guten Übersetzungsqualität. Mittels der Skalierungsfaktoren λ_m sind wir in der Lage, die einzelnen Modelle unterschiedlich stark zu gewichten.

In dieser Aufgabe stellen wir die Skalierungsfaktoren der log-linearen Modellkombination mit geeigneten Algorithmen so ein, dass die Ausgabe des Decoders direkt bezüglich eines der automatischen Fehlermaße optimiert wird.

Aufgabe:

1. Implementieren Sie den Downhill Simplex Algorithmus.
2. Implementieren Sie MERT.
Wieviel Zeit kann durch effiziente Programmierung / Algorithmen gewonnen werden? Welche Quellcode-Parameter können verändert werden? Gibt es Verbesserungen?
3. Nutzen Sie beide Verfahren, um die Skalierungsfaktoren der im Dekoder verwendeten Modelle
 - Source-to-Target Phrasenübersetzungskosten,
 - Target-to-Source Phrasenübersetzungskosten,
 - Source-to-Target einzelwortbasierte Übersetzungskosten auf Phrasenebene,
 - Target-to-Source einzelwortbasierte Übersetzungskosten auf Phrasenebene,
 - Phrase Penalty,
 - Word Penalty,
 - Single Count Bit,
 - Source-Target Ratio und

- Unigramm-Sprachmodell

auf den auf der Webseite zur Verfügung gestellten Development-Daten bezüglich BLEU zu optimieren. Übersetzen Sie die Test-Daten mit den optimierten Skalierungsfaktoren. Vergleichen Sie die Resultate. Um wieviel hat sich das System gegenüber der unoptimierten Baseline verbessert?

Abnahmetermin: Donnerstag, 30. Juni, ab 14:00 Uhr

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte Ihre kommentierten Quelltexte bereits bis Mittwoch Abend (29. Juni, 18:00 Uhr) an

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

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