

EINLADUNG

Zeit: Freitag, 9. Dezember 2005, 15.00 Uhr
Ort: AH V, Ahornstraße 55 (ehem. PH)
Referent: Dipl.-Physiker Stephan Vogel
Thema: Statistical Machine Translation with Cascaded Probabilistic Transducers

Abstract:

In der statistischen maschinellen Übersetzung hat sich in den letzten Jahren ein Wandel von wortbasierten hin zu phrasenbasierten Ansätzen vollzogen. Hierdurch werden lokaler Kontext und lokale Wortumstellung besser erfasst, was zu deutlich besserer Übersetzungsqualität führt. Im Rahmen des Vortrags werden Trainings- und Dekodierungsverfahren vorgestellt, die eine Generalisierung des Phrasenalignments zu einem hierarchischen Alignment durch kaskadierte probabilistische Transducer ermöglichen.

Zunächst wird ein neues Phrasenalignment - Verfahren vorgestellt, das auf einer Erweiterung der Wortalignment-Algorithmen beruht. Gegenüber den derzeit verwendeten Standardverfahren besitzt es den Vorteil, weder die Maximum-Approximation noch Heuristiken zur Kombination der Viterbi Alignments zu verwenden.

Um ein hierarchisches Phrasenalignment zu trainieren, wird eine Erweiterung des sogenannten HMM Alignments zu einem Graphalignment vorgestellt. Durch die Anwendung einer Kaskade von Finite State Transducern auf ein bilinguales Corpus wird ein bilinguales partielles Parsing durchgeführt. Das Graphalignment findet dann die beste Zuordnung der beiden Parse-Strukturen und erlaubt die Schätzung der Wahrscheinlichkeiten der probabilistischen Transducer.

Anschließend wird ein Dekodierungsverfahren beschrieben, das in zwei Schritten arbeitet. Zuerst werden die Transducer angewendet, um Übersetzungen für einzelne Worte und Phrasen eines Satzes zu erzeugen. Das Resultat ist eine Graphstruktur. Im zweiten Schritt erfolgt eine single best path or n-best paths Suche durch diesen Graphen. Das Suchverfahren wird in einer Weise formuliert, dass Wortumstellungen zwischen Quell- und Zielsprache möglich sind. Rekombinierung von partiellen Hypothesen sowie eine flexible Pruningstrategie garantieren schnelle Dekodierung.

Es werden Ergebnisse auf verschiedenen Corpora vorgestellt: einerseits beschränkte Domänen mit kleinem Vokabular, andererseits weitgehend domänen-unbeschränkt mit sehr großen Vokabularen.

Es laden ein: Die Dozenten der Informatik