

EINLADUNG

Zeit: 29. Februar 2008, 10:00 Uhr
Ort: Raum 5056, Ahornstr. 55
Referent: Dipl.-Informatiker Richard Zens
Titel: Phrase-based Statistical Machine Translation:
Models, Search, Training

Abstract:

Machine translation is the task of automatically translating a text from one natural language into another. In this work, we describe and analyze the phrase-based approach to statistical machine translation. In any statistical approach to machine translation, we have to address three problems: the modeling problem, i.e. how to structure the dependencies of source and target language sentences; the search problem, i.e. how to find the best translation candidate among all possible target language sentences; the training problem, i.e. how to estimate the free parameters of the model from the training data.

We will present improved alignment and translation models. We will present alignment models which improve the alignment quality significantly. We describe several phrase translation models and analyze their contribution to the overall translation quality.

We formulate the search problem for phrase-based statistical machine translation and present different search algorithm in detail. We analyze the search and show that it is important to focus on alternative reorderings, whereas on the other hand, already a small number of lexical alternatives are sufficient to achieve good translation quality. The reordering problem in machine translation is difficult for two reasons: first, it is computationally expensive to explore all possible permutations; second, it is hard to select a good permutation. We compare different reordering constraints to solve this problem efficiently and introduce a lexicalized reordering model to find better reorderings.

We investigate alternative training criteria for phrase-based statistical machine translation. In this context, we generalize the known word posterior probabilities to n-gram posterior probabilities.

The resulting machine translation system achieves state-of-the-art performance on the large scale Chinese-English NIST task. Furthermore, the system was ranked first in the official TC-Star evaluations in 2005, 2006 and 2007 for the Chinese-English broadcast news speech translation task.

Es laden ein: Die Dozenten der Informatik