

EINLADUNG

Zeit: 10. September 2010, 14 Uhr

Ort: Raum 5052, Ahornstr. 55

Referent: Dipl.-Inform. Jia Xu
Lehrstuhl für Informatik 6

Titel: Sequence Segmentation for Statistical Machine Translation

Abstract:

In the last decade, while statistical machine translation has advanced significantly, there is still much room for further improvements relating to many natural language processing tasks such as word segmentation, word alignment and parsing.

Human language is composed of sequences of meaningful units. These sequences can be words, phrases, sentences or even articles serving as basic elements in communication and components for computational modeling. However, in monolingual text some sequences are not naturally separated by delimiters, and in bilingual text both sequence boundaries and their corresponding translations can be unlabeled. This work addresses solutions of sequence segmentation and alignment for statistical machine translation, including the following topics:

Chinese word segmentation: Different from the explicit word segmentation in trivial approaches, I introduce integrated Chinese word segmentation, where segmentation and alignment of words are trained jointly, and the decoding is performed on the lattice composed of alternative word segmentations. I show that direct translation on Chinese characters can achieve even better translation performance than translation on Chinese words;

Phrase training: Currently phrases are extracted in a heuristic way. I propose a mixture phrase pair model which is trained discriminatively allowing to combine multiple extraction processes and various resources, especially the underlying word alignment models discarded in the standard approach;

Parallel sentence exploitation: Training corpus acquisition is crucial for a data-driven translation system. I propose a maximum-entropy model where document pairs are partitioned recursively into sentence pairs using 'binary segmentation' without any requirement on sentence boundary markers;

Domain adaptation: A hierarchical clustering algorithm is applied to classify the training data into distinct domains. Domain specific language models and translation models are then combined to build a domain dependent system, and domain priors are estimated with a minimum error rate training.

Experimental results on state-of-the-art, large-scale Chinese-English tasks show that the training speed can be increased with a factor of four and each above mentioned method leads to an enhancement of the translation quality up to 6% relatively.

Es laden ein: Die Dozenten der Informatik