ACM Transactions on Speech and Language Processing 4 (3), 2007, art.6.

Christoph Tillmann and Tong Zhang:

A block bigram prediction model for statistical machine translation.

Abstract

In this article, we present a novel training method for a localized phrase-based prediction model for statistical machine translation (SMT). The model predicts block neighbors to carry out a phrase-based translation that explicitly handles local phrase reordering. We use a maximum likelihood criterion to train a log-linear block bigram model which uses real-valued features (e.g., a language model score) as well as binary features based on the block identities themselves (e.g., block bigram features). The model training relies on an efficient enumeration of local block neighbors in parallel training data. A novel stochastic gradient descent (SGD) training algorithm is presented that can easily handle millions of features. Moreover, when viewing SMT as a block generation process, it becomes quite similar to sequential natural language annotation problems such as part-of-speech tagging, phrase chunking, or shallow parsing. Our novel approach is successfully tested on a standard Arabic-English translation task using two different phrase reordering models: a block orientation model and a phrase-distortion model.