ACM Transactions in Speech and Language Processing, vol.2, no.2 (2005), pp.1-24

Marcello Federico and Nicola Bertoldi:

A word-to-phrase statistical translation model.

Abstract

This article addresses the development of statistical models for phrase-based machine translation (MT) which extend a popular word-alignment model proposed by IBM in the early 90s. A novel decoding algorithm is directly derived from the optimization criterion which defines the statistical MT approach. Efficiency in decoding is achieved by applying dynamic programming, pruning strategies, and word reordering constraints. It is known that translation performance can be boosted by exploiting phrase (or multiword) translation pairs automatically extracted from a parallel corpus. New phrase-based models are obtained by introducing extra multiwords in the target language vocabulary and by estimating the corresponding parameters from either: (i) a word-based model, (ii) phrase-based statistics computed on the parallel corpus, or (iii) the interpolation of the two previous estimates. Word-based and phrase-based MT models are evaluated on a traveling domain task in two translation directions: Chinese-English (12k-word vocabulary) and Italian-English (16k-word vocabulary). Phrase-based models show Bleu score improvements over the word-based model by 19% and 13% relative, respectively.