

Student Research Symposium 2003

Ying (Joy) Zhang

An Integrated Phrase Segmentation/Alignment Algorithm for Statistical Machine Translation

We present an integrated phrase segmentation/alignment algorithm (ISA) for Statistical Machine Translation. Without the need of building an initial word-to-word alignment or initially segmenting the monolingual text into phrases as other methods do, this algorithm segments the sentences into phrases and finds their alignments simultaneously. For each sentence pair, ISA builds a two-dimensional matrix to represent a sentence pair where the value of each cell corresponds to the Point-wise Mutual Information (MI) between the source and target words. Based on the similarities of MI values among cells, we identify the aligned phrase pairs. Once all the phrase pairs are found, we know both how to segment one sentence into phrases and also the alignments between the source and target sentences. We use monolingual bigram language models to estimate the joint probabilities of the identified phrase pairs. The joint probabilities are then normalized to conditional probabilities, which are used by the decoder. Despite its simplicity, this approach yields phrase-to-phrase translations with significant higher precisions than our baseline system where phrase translations are extracted from the HMM word alignment. When we combine the phrase-to-phrase translations generated by this algorithm with the baseline system, the improvement on translation quality is even larger.