

## Abstract

The ever-increasing amount of parallel data opens a rich resource to multilingual natural language processing, enabling models to work on various translational aspects like detailed human annotations, syntax and semantics. With efficient statistical models, many cross-language applications have seen significant progresses in recent years, such as statistical machine translation, speech-to-speech translation, cross-lingual information retrieval and bilingual lexicography. However, the current state-of-the-art statistical translation models rely heavily on the word-level mixture models --- a bottleneck, which fails to represent the rich varieties and dependencies in translations. In contrast to word-based translations, phrase-based models are more robust in capturing various translation phenomena than the word-level (e.g., local word reordering), and less susceptible to the errors from preprocessing such as word segmentations and tokenizations. Leveraging phrase level knowledge in translation models is challenging yet rewarding: it also brings significant improvements on translation qualities. Above the phrase-level are the sentence- and document- levels of translational equivalences, from which \emph{topics} can be further abstracted as hidden concepts to govern the bilingual generative process of sentence-pair, phrase-pair or word-pair sequences. The modeling of hidden bilingual concepts also enables the learning to share parameters, and thus, endows the models with the abilities of \emph{learning to translate}.

Learning translational equivalence is the fundamental building block for machine translations. This thesis delves into learning statistical alignment models for translational equivalences at various levels: documents, sentences and words. Specific attention will be devoted to introducing hidden concepts in modeling translation. Models, such as \emph{Inner-Outer Bracket} models, are designed to model the dependency between phrases and the words inside of them; bilingual concepts are generalized to integrate topics for translation. In particular, \emph{Bilingual Topic-AdMixture} (BiTAM) models are proposed to formulate the semantic correlations among words and sentences. BiTAM is shown to be a general framework, which can generalize over different traditional alignment models with ease. In this thesis, IBM Model-1 and HMM are embedded BiTAM; BiTAM 1-3 and HM-BiTAM are designed with tractable learning and inference algorithms. Improvements of word alignment accuracies are observed, and also better machine translation qualities are obtained.

The models, proposed in this thesis, have also been applied successfully in the past a few statistical machine translation evaluations for the CMU-SMT team, especially for the scenarios of Chinese-to-English.