Syntax-Based Markov Models for Word Alignment

Roger Wechsler

January 11, 2017

Abstract

Word alignment is the task of aligning the words in one sentence with the corresponding words of another sentence that is a translation in another language. The resulting alignment links between the sentences are essential for machine translation systems as they allow the induction of a bilingual phrase dictionary, which serves as a lookup table for translation variants.

Most alignment methods attempt to align each word in one of the two sentences, the source sentence, to any word in the other sentence, the target sentence. The word alignment problem is then reduced to deciding for each source word to which target word it should be aligned. Such models perform particularly well if those alignment decisions are not made independently for each source word, but by considering to which target words, other source words are aligned. To keep the models tractable, the Markov assumption is made, which assumes that the alignment of a source word only depends on the alignment of the previous source word. The alignment decisions, thus, depend on each other in the form of a linear chain. Commonly used Markov models for word alignment are unsupervised generative Hidden Markov models (HMM) or supervised discriminative Conditional Random Fields (CRF).

From a linguistic perspective, the previous word in chain-structured models is often not directly related to the current word that is to be aligned. This weakness is more severe for languages with long distance dependencies in which words are separated by long phrases from their actual linguistic dependents. This thesis addresses this deficiency by incorporating a source-side syntactic dependency tree for word alignment models. We present and evaluate new HMM and CRF models, which use the syntactic information directly in their model or feature definitions.

Moreover, as opposed to the commonly used chain-structured models, we propose tree-structured models that use dependencies which mirror the structure of the source-side syntactic dependency tree. In a tree-structured model, a source word is aligned by considering the alignment decision that was made for its syntactic head. We explore how tree-structured alignment variables impact the models and we show which measures are required to render them competitive with chain-structured models.

All models are evaluated in word alignment experiments in terms of alignment error rate (AER). Moreover, we also measure their performance when they are directly applied in end-to-end machine translation, for which we report BLEU and METEOR scores. The experiments are conducted for the language pairs French-English and German-English and the results show that the alignment models benefit clearly from the information in the source-side syntactic tree. The tree-structured models are competitive with chain-structured models, particularly for the language pair German-English. Finally, our machine translation experiments show that the proposed models are competitive with the baseline models for all language pairs.