Two-level Alignment by Words and Phrases Based on Syntactic Information

Seonho Kim[1], Juntae Yoon[2],and Dong-Yul Ra[3]

[1] Institute of Language and Information Studies,
Yonsei University, Seoul, Korea
shkim@lex.yonsei.ac.kr
[2] NLP Lab., Daumsoft, Seoul, Korea
jtyoon@daumsoft.com
[3] Dept. of Computer Science, Yonsei University, Korea
dyra@magics.yonsei.ac.kr

Abstract. As a part of work on alignment of the English and Korean parallel corpus, this paper presents a statistical translation model incorporating linguistic knowledge of syntactic and phrasal information for better translations. For this, we propose three models: First, we incorporate syntactic information such as part of speech into the word-based lexical alignment. Based on this model, we propose the second model which finds phrasal correspondence in the parallel corpus. Phrasal mapping through chunk-based shallow parsing enables to settle mismatch of meaningful units in the two languages. Lastly, we develop a two-level alignment model by combining these two models in order to construct both the word and phrase-based translation model. Model parameters are automatically estimated from a set of bilingual sentence pairs by applying the EM algorithm. Experiments show that the structural relationship helps construct a better translation model for structurally different languages like Korean and English.

1 Statistical Translation Model

Pairs of sentences aligned with words or phrases can be exploited as a valuable resource for work on bilingual systems such as machine translation, cross languages information retrieval, bilingual lexicography, and bilingual parsing. As large scale bilingual corpora are available, a lot of statistical approaches have been proposed for finding sets of corresponding word tokens [2, 5], phrases [3, 6, 11, 12, 14], and syntactic structures [10, 13, 15] from a bitext.

Many of those works have been initiated with statistical machine translation (SMT) model pioneered by Brown et al. (See [2]), which estimates, directly from bilingual corpora, parameters for a word-to-word alignment model. In this framework, Korean-to-English machine translation is assumed that each source language ($\mathbf{k}$) is transformed to its target language ($\mathbf{e}$) by means of a stochastic process. Typically, the stochastic process is represented by $P(\mathbf{e}|\mathbf{k}) = P(\mathbf{e})P(\mathbf{k}|\mathbf{e})$, where $P(\mathbf{e})$ is called a language model (LM), and $P(\mathbf{k}|\mathbf{e})$ is referred to as a