Abstract

Word alignment in bilingual corpora has been an active research topic in the Machine Translation research groups. Corpus is the body of text collections, which are useful for Language Processing (NLP). Parallel text alignment is the identification of the corresponding sentences in the parallel text. Large collections of parallel level are prerequisite for many areas
of linguistic research. Parallel corpus helps in making statistical bilingual dictionary, in supporting statistical machine translation and in supporting as training data for word sense disambiguation and translation disambiguation. Nowadays, the world is a global network and everybody will be learned more than one language. So, multilingual corpora are more processing. Thus, the main purpose of this system is to construct word-aligned parallel corpus to be able in Myanmar-English machine translation. One useful concept is to identify correspondences between words in one language and in other language. The proposed approach is based on the first three IBM models and EM algorithm. It also shows that the approach can also be improved by using a list of cognates and morphological analysis.

Reference


Index Terms

Computer Science
Computational Linguistics

Key words

Word-aligned Parallel Corpus IBM Models EM Algorithm