Chanon Onman 2012: Thai – English Sentence Alignment using Weighted Cost Functions.
Master of Engineering (Computer Engineering),
Major Field: Computer Engineering, Department of Computer Engineering.
Thesis Advisor: Associate Professor Asanee Kawtrakul, D.Eng. 81 pages.

Machine translation plays an important role these days, since translating a large quantity of text in limited time is required. There are 2 approaches in machine translation which are rule based and corpus based approach. Rule based approach is to translate text by using rules and vocabularies provided by computational linguists. In the other hand, corpus based approach is to automatically extract translation knowledge from parallel corpus. Accordingly, corpus based approach is easier to extend translation capability and apply to new language pairs or domains than rule based approach.

However, manually creating parallel corpus is still laborious work, so developments of automatic sentence alignment tools are required. In previous works, sentence alignment tools utilized language information from initial parallel corpus, existing resources such as dictionary, Wordnet, or common linguistic features of specific language pairs such as punctuation markers, cognate. Previous sentence alignment tools have been proved to yield high accuracy with literal translated document. But in case of free translated document, there are some modifications in translation that subsequently cause insertion and deletion alignments and make the alignment task has more complexity and yields lower accuracy. Moreover, free translated documents tend to contain more loose translation examples than literal translated document and such examples are required for training machine translation in order to generate more naturally translation. The objective of this work is to study and develop a technique for aligning Thai – English parallel sentence in free translated documents. Since Thai – English language pair shares a few common linguistic features and have specific linguistic phenomena which causes more complexity in alignment such as using unknown words, word-description translation pairs, accordingly, a new alignment technique is needed. In this work, the proposed technique uses weighted cost functions which are analyzed from 3 features consisting of alignment type, sentence length, and translation probability. All of variables and weighting values are computed from an initial parallel corpus.

In this work, a 5-fold cross validation was performed with a set of free translated documents which consists of 115 paragraphs consisting of 941 sentence pairs. The proposed technique yields 0.728 precision, 0.752 recall, and 0.740 F-score. This technique is also used to evaluate with a literal document pair which is in legal domain. The experiment result yields 0.998 in precision and recall.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature