Don't use big words with me: An evaluation of English-Thai Statistical-based Machine Translation

Sanooch Nathalang
National Electronics and Computer Technology Center (NECTEC), Thailand

With the availability of many machine translation systems come the question of how effective they really are. The main purpose of this study is to evaluate the English-Thai statistical-based machine translation (SMT) developed by Human Language Technology Laboratory, National Electronics and Computer Technology Center (NECTEC), Thailand, by a human evaluator. We look in particular for potential areas of difficulty that may cause problems to the SMT system. The corpus from which the data for the current study were extracted consists of 200,000 English-Thai aligned sentences (around 1.3 million words) originally taken from bilingual sources that are mainly educational in nature such as dictionaries and phrase books. We consider the English sentences as the source, and the Thai sentences as the target. In the past few years, there were a few attempts to evaluate the translations generated by our SMT system (e.g. Porkeaw, Supnithi, Wutiwiwatchai, 2007), using the well-known BLEU metric. The results yielded were relatively low BLEU scores of 13-15. This immediately calls for an in-depth analysis of the difficulties that challenge the system. In this presentation, we based our linguistic analysis on the linguistic approach proposed by Baker (1992), paying particular attention to establishing equivalences between English and Thai at word level. Our investigation showed that simple words in English that can find their equivalences in Thai do not pose major problems in translation. However, problematic cases tend to occur where an English word corresponds to more than one word in Thai. Explanations to these problems can be drawn from a number of approaches, ranging from language typology, morphology and syntax, to lexicography. The results of the investigation lead us to conclude that the linguistic differences between the source language and the target language still play a significant role in developing and improving the SMT.