Abstract 92

A Study on the Utility of Hierarchical Phrase-Based Model for Low Resource Languages

Yashothara Shanmugarasa[1], Uthayasankar Thayasivam[2]

[1]University of Moratuwa
yashoshan@gmail.com

[2]University of Moratuwa
rtuthaya@cse.mrt.ac.lk

With the rebellion of internet, people got more opportunities to go global. There is the issue of communication, which is made more challenging due to difference in languages. English is the generally spoken language and there is no assurance that everyone is proficient in it. Therefore, translation plays a major role. Currently, South Asian languages are dominantly translated using traditional statistical and neural machine translation approaches. South Asian languages lack necessary natural language resources and tools, hence are classified as low resourced languages. This limits the effectiveness achievable in machine translation of those languages. Compared to English language, South Asian languages are morphologically rich and are commonly used in different sentence structures. For example, the structure of a sentence is subject-verb-object in English while it is subject-object-verb in most South Asian languages. As official languages of Sri Lanka are low resourced, when it is used to translate using traditional statistical machine translation, it is impossible to produce sentences with acceptable sentence structure because of sub-phrases which can only be reordered using distortion reordering model, are independent of their context. In addition, using phrases longer than three words barely improves the translation because such phrases are infrequent in the corpora due to data sparsity. To overcome this problem hierarchical phrase model translation, which uses grammar rules formed by the Synchronous Context Free Grammar, can be used. Moses is selected to build the baseline system. In the experiments, the system used 50000 parallel sentences for Tamil and English. Using BLEU as a metric, the hierarchical phrase-based model achieves 3.42 for Tamil to English translation and 1.73 for vice-versa. This score improves 0.72 from traditional approach. For Sinhala to Tamil, it achieves 11.18 and 10.73 for vice-versa. Moreover, the system could further be improved by establishing certain rules.

Key words: Hierarchical Model, Synchronous Context-Free Grammar (SCFG), BLEU
and Publication Unit, University of Kelaniya, Sri Lanka