[CICLING 2010: 11th International Conference on Intelligent Text Processing and Computational Linguistics, March 21-27, 2010, Iaşi, Romania]

Abstract

Thoudam Doren Singh and Sivaji Bandyopadhyay. Manipuri-English Example Based Machine Translation system

The paper reports our work on the development of a machine translation system for translating Manipuri to English using example based approach. Manipuri is a relatively free word-order language and makes use of a set of enclitics and morphological suffixes for correct meaning representation. The sentence level parallel Manipuri – English corpus is built from a comparable Manipuri and English news corpora. A Manipuri – English lexicon is being developed and is used during the alignment process. Preprocessing steps are applied on the sentence level parallel corpus in terms of POS tagging, morphological analysis, named entity recognition and chunking on both the source and the target sides. The developed sentence level parallel corpus is aligned at the phrase level. The translation process initially looks for an exact match in the parallel example base and returns the retrieved target output in case of a match. Otherwise, the maximal match sentence is identified for the input sentence. For word level mismatch, the unmatched words in the input are looked into the lexicon or transliterated if not found in the lexicon. Unmatched phrases are looked into the phrase level parallel example base, the target phrase translation identified and then recombined with the retrieved output from the maximal match parallel pair. In case of more than one maximal match pair, the most frequent pair in the example base is identified and the same translation process is carried out as for one maximal match. If there is no match (full or partial) with any of the parallel pairs then a phrasal EBMT method will be applied which is being developed. The EBMT system has been developed using 15319 Manipuri-English parallel sentences and evaluated three fold with a test set of 900 gold standard test sentences giving BLEU and NIST scores of 0.137 and 3.361 respectively. A baseline SMT system using MOSES with the same training and test data has been developed and subsequently evaluated for a BLEU and NIST scores of 0.128 and 3.195 respectively. Thus the proposed EBMT system has performed better than the baseline SMT system with the same training and test data.