Abstract

The Statistical Machine Translation (SMT) systems are developed using sentence aligned parallel corpus. The difficulty is that there is no parallel corpus at the required measure for many language pairs. The preparation of large scale parallel corpus takes time and demands the linguistics skill. In the present work, the various issues of a quality parallel corpus and a technique that extracts parallel corpus between Manipuri, a morphologically rich and resource constrained Indian language and English has been developed from a web based comparable news corpora. We explore the crux of the parallel corpora towards improving the translation quality through linguistics factors for the language pair.

References

- Gale, W. A., Church, K. W., 1991. A program for aligning sentences in bilingual corpora, In proceedings of 29th Annual meeting of ACL, Pages 177-184, Berkeley, California
Building Parallel Corpora for SMT System: A Case Study of English-Manipuri

111-119, Chiang Mai, Thailand, 2011.

- Singh, T. D., Bandyopadhyay, S. 2010d. Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM, Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), the 23rd International
Conference on Computational Linguistics (COLING), Pages 35–42, Beijing.

Index Terms

Computer Science  
Artificial Intelligence

Keywords

Sentence alignment  
Precision  
Recall  
English-Manipuri  
Agglutinative Morphology