Resource Creation for English-Maithili Machine Translation (EMMT) A Divergence Perspective

Ritu Nidhi and Tanya Singh

Amity Institute of Information Technology, Amity University, Noida, UP, India

Maithili is one of the 22 scheduled Indian languages with almost no language technology resource. Absence of basic tools in this language has affected resource creation. Since English is the dominant language, translation from it can help creating the required corpora for tools development in Maithili. The present work discusses efforts for Language Technology Resource (LTR) creation and divergence study for an EMMT system, which is a Statistical Machine Translation (SMT) system. Creating any SMT system requires sizeable parallel, aligned corpora for training and testing. Creating general-purpose source corpora for English language and creating translation equivalents with possible alignments requires time and effort. The paper focuses on the data collection methods, cleaning, the size and structure of the text corpora, alignment and parallelization strategies, training, testing and a study of divergence between the language pair.

Keywords: Divergence, Machine Translation, English-Maithili, Indian languages, MT Hub