ID 645

Building a Sinhala-English Parallel Corpus for Neural Machine Translation Based on Exam Questions

MRM Rilfi[1#], UGYM Gunawansha[2], KAC Prasandika[2] and KGA Chandrani[2]

[1]Inoovalab Technologies, Sri Lanka
[2]University of Moratuwa, Sri Lanka

[#]rilfi@inoovalab.org

In any neural machine translation between two natural languages, parallel corpus is a compulsory part of the training process. The most crucial step in an MT system is to develop an effective method for gathering parallel corpus. The construction of a parallel corpus, on the other hand, necessitates substantial knowledge of both languages and is a time-consuming procedure. Due to these limits, digitizing documents becomes extremely challenging, lowering the quality of machine translation systems. This research offers a method for producing an English to Sinhala parallel corpus that is both faster and more efficient, while requiring less human intervention. This system generates a parallel corpus for language pair using the following steps: scanning the exam question papers using a special type of scanner, Image optimization for Optical Character Recognition (OCR), text extraction from images and converting unstructured text into structured form as parallel corpus.

Keywords: *parallel corpus, image optimization, text extraction, neural machine translation*