A review on building bilingual comparable corpora for resource-limited languages

ABSTRACT

Information retrieval tasks on certain Asian languages have the problem of limited knowledge resources such as the bilingual and multilingual dictionaries and corpora. Thus, there is a need to create multilingual resources for these languages. One of the ways is to automatically align document by identifying the chances that two documents are related to each other and these documents are not necessarily in one language. Multilingual corpora can then be automatically developed from these aligned documents. Numerous approaches for document alignment have been developed to date. In this paper, we gave an overview of recent progress made for bilingual and multilingual document alignments within the last 5 years. In addition, we also discussed the current progress made in developing bilingual comparable corpus especially on the Malay language, which is one of the resource-limited languages in Asia.

Keyword: Document alignment; Cross-lingual information retrieval; Comparable corpus; Malay language