Feature-based similarity method for aligning the Malay and English news documents

ABSTRACT

Corpus-based translation approach can be used to obtain reliable translation knowledge in addition to the use of dictionaries or machine translation. But the availability of such corpus is very limited especially for the low-resources languages. Many works have been reported for the alignments of multilingual documents especially among the European languages, but less focusing on the languages with less linguistics resources. One of the challenges is to align the available multilingual documents for the creation of comparable corpus for these kinds of languages. This article describes an alignment method that utilized the statistical features of the documents such as the documents' titles, texts of the contents, and also the named entities present in each document. This method will be focusing on the English and Malay news documents, in which in which the Malay language is considered as a low-resource language. Source and target documents were then compared in a pair. Accuracy, precision, and recall measurements were used in evaluating the results with the inclusion of three relevance scales; Same story, Shared aspect and Unrelated, to assess the alignment pairs. The results indicate that the method performed well in aligning the news documents with the accuracy of 96% and average precision of 81%.

Keyword: Document alignment; Feature-based method; Algorithm; Malay text processing; Corpus-based information retrieval