Bilingual Extractive Text Summarization Model using Textual Pattern Constraints

ABSTRACT

In the era of digital information, an auto-generated summary can help readers to easily find important and relevant information. Most of the studies and benchmark data sets in the field of text summarization are in English. Hence, there is a need to study the potential of Malay language in this field. This study also highlights the problems in identifying and generating important information in extractive summaries. This is because existing text representation models such as BOW has weaknesses in inaccurate semantic representation, while the N-gram model has the issue of producing very high word vector dimensions. In this study, a bilingual text summarization model named MYTextSumBASIC has been developed to generate an extractive summary automatically in Malay and English. The MYTextSumBASIC summarizer model applies a text representation model known as FASP using three Textual Pattern Constraints, namely word item constraints, adjacent word constraints and sequence size constraints. There are three main phases in the framework of MYTextSumBASIC model, which are the development of the Malay language corpus, the development of MYTextSumBASIC model using FASP and the summary evaluation phase. In the summary evaluation phase, using the Malay language data sets of 100 news articles, the summaries produced by MYTextSumBASIC outperformed the summary generated by Baseline (Lead) and OTS summarizer with the highest average for retrieval (R) is 0.5849, precision (P) is 0.5736 and the F-score (Fm) is 0.5772. For manual evaluation by linguists, the MYTextSumBASIC method yielded a reading score of 4.1 and 3.87 for summary content generated using a random data set. Further experiments using the 2002 DUC English benchmark data set of 102 news articles have also shown that the MYTextSumBASIC model outperformed the best and lowest systems in the comparison with the mean retrieval values of ROUGE-1 (0.43896) and ROUGE-2 (0.19918). These findings conclude that the FASP text representation feature along with the textual pattern constraints used by our model can be used for bilingual text with competitive performance compared to other text summarization models.