A Syntactic-based Sentence Validation Technique for Malay Text Summarizer

ABSTRACT

In the automatic text summarization domain, a sentence compression technique is applied to the summary sentence to remove unnecessary words or phrases. The purpose of sentence compression is to preserve important information in a sentence and to remove unnecessary ones without sacrificing the sentence's grammar. The existing development of Malay natural language processing tools is still under study with limited open access. The issue is the lack of a benchmark dataset in the Malay language to evaluate the quality of the summaries and to validate the compressed sentence produced by the summarizer model. Therefore, this paper outlines a syntactic based sentence validation technique for Malay sentences by referring to the Malay grammar pattern. In this work, a new derivation set of syntactic rules based on the Malay main word class was proposed to validate Malay sentences that underwent the sentence compression procedure. This paper used the Malay dataset of 100 new articles covering the natural disaster and events domain to find the optimal compression rate and its effect on the summary content. An automatic evaluation using the benchmark ROUGE toolkit produced a result with an average F-measure of 0.5826 and an average recall value of 0.5925 with an optimum compression rate of 0.5 confidence conf value. Furthermore, a manual summary evaluation by a group of Malay experts on the grammaticality of the compressed summary sentence produced a good result of 4.11 and a readability score of 4.12 out of 5. This depicts the reliability of the proposed technique to validate Malay sentences with promising summary content and readability results.