How to cite this paper:

Bali Ranaivo-Malançon & Hazimah Iboi. (2017). Which extractive summarization method for Malay texts? in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference on Computing & Informatics (pp 577-582). Sintok: School of Computing.

WHICH EXTRACTIVE SUMMARIZATION METHOD FOR MALAY TEXTS?

Bali Ranaivo-Malançon and Hazimah Iboi

Universiti Malaysia,Sarawak, mbranaivo@unimas.my, ihazimah11@gmail.com

ABSTRACT. The number of texts written in Malay increases every day. When these texts are lengthy, interested readers tend to skim through them. Automatic text summarization may assist these readers to get access to the important parts of the texts without scanning from the beginning to the end. As of today, only few Malay text summarizers have been presented in the literature. Therefore, a comparative study of three extractive summarization methods (Luhn's method, Edmundson's method, and LexRank method) was undertaken and the results are reported in this paper. The aim of the study is to determine the adequate extractive method. Several experiments were conducted by comparing the results of three extractive methods with human extracts as well as human abstracts. It appears that the Luhn's method, one of the oldest automatic extractive summarization, shows a good performance while tested on 14 Malay abstract summaries and 20 Malay extractive summaries.

Keywords: extractive summarization, Luhn's method, Edmundson's method, LexRank method, Malay text

INTRODUCTION

As the world continues its progress, more and more Malay texts are created and many of them are available in digital form. However, the lengths of these texts are variable. Lengthy texts are usually skimmed through by readers. An automatic text summarization (ATS) can assist readers in getting access to the useful parts of any lengthy text. That is, only a subset sentences from the complete set of sentences will be presented to the readers. Unfortunately, there are only few works on Malay text summarization and thus no dedicated tool is immediately accessible to alleviate the reading of lengthy Malay texts. An ATS can be qualified as either extractive or abstractive. An extractive summary is the result of the selection of a few salient sentences from a full text. There are many extractive ATS approaches and they differ on the definition of "salient sentences". An abstract is a sketchy summary of the main points of a full text. Abstracting a text is not easy. Abstractive systems "are difficult to replicate, as they heavily rely on the adaptation of internal tools to perform information extraction and language generation" (Das & Martins, 2007). Therefore, the work reported in this paper focuses only on extractive methods. This study was undertaken to determine the adequate extractive summarization method for Malay texts. Three extractive summarization methods, that are Luhn's method (Luhn, 1958), Edmundson's method (Edmundson, 1969), and LexRank method (Erkan & Radev, 2004), were investigated and evaluated on extractive as well as abstractive Malay summaries (Figure 1). A full Malay text is summarized by human and by three automatic text summarizers. Thus, five kinds of summaries were obtained: a human