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 abstract (Abstract-H), a human extract (Extract-H), an extract from Luhn's method, an extract from Edmundson's method, and an extract from LexRank method. The automatic extracts are then compared against the each of the human (extract and abstract) summary using ROUGE metric (Lin, 2004).

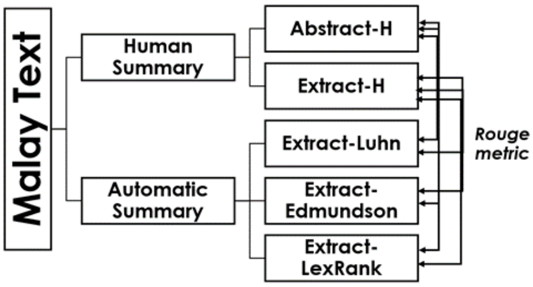

Figure 1. Overall Framework of the Comparative Study.

EXISTING MALAY TEXT SUMMARIZATION WORKS

Whereas the first work on automatic English text summarization was reported to be in the late 1950's, the first published works on Malay summarization was in 2010. Zamin and Ghani (Zamin & Ghani, 2010) proposed a hybrid Malay extractive summarizer. They combined the pre-processing step of SUMMARIST (Hovy & Lin, 2000) and the text extraction step of EstSum (Müürisep & Mutso, 2005), which is based on Edmundson's work. From the original SUMMARIST pre-processing component, Zamin and Ghani kept only the modules that split the input text into sentences, split sentences into tokens, and count the number of tokens. The EstSum text extraction aims to extract salient sentences based on three features: location, format, and keyword. For the dataset evaluation, they used 10 Malay news articles (general news, business news, and sports news). Four Malay language experts were asked to create extracts with no more than 30% of the original article's length. For the evaluation metrics, they used the common information retrieval metrics, that are, recall, precision, and F1-score. The summarization of the general news gave the best F1-score, 0.76 with 0.38. For the sports news, the F1-score is 0.422 and the recall is 0.31. The summarization of the business news seemed not easy, and thus the Malay hybrid summarizer could only reach 0.31 F1-score and 0.223 recall.

Jusoh and colleagues (Jusoh et al., 2011) proposed a sentence refinement approach to improve the result of a sentence extraction technique. The summarizer is called SRAETS. Three kinds of sentences are extracted: (1) the first and second sentence of a paragraph, (2) sentences containing all nouns in the title as well as keywords determined from the topic of the text, and (3) the last sentence of the paragraph. The extracted sentences are then refined by eliminating what the authors called "unnecessary words or phrases" and quoted texts. The proposed approach is quite difficult to replicate for two main reasons. First, the determination of the list of keywords require to have an understanding of the domain of the original text. Second, the refinement process seems to be ad-hoc. SRAETS was evaluated on 40 Malay texts and 100 English texts using extrinsic and intrinsic methods. The source of the Malay texts is not mentioned in the paper. The result of the intrinsic method is based on one single evaluator's opinion stating whether the summary is satisfactory or not. The authors wrote that the overall rating was satisfactory. The Compression Ratio metric (CR) was used for the extrinsic evaluation. CR is defined as "(the length of summary text) / (the length of full text) * 100". The CR values vary between 20.66% and 87.50% for the 40 evaluated Malay texts, giving an average of 54.08%. The authors interpreted the results as satisfactory since

SRAETS showed lower CR rate while compared with the results of Copernic text summarizer on the same Malay texts.

SELECTED EXTRACTIVE SUMMARIZATION METHODS

Luhn's method (Luhn, 1958) is one the earliest research on extractive summarization even though the author presented it as a method for getting abstracts. Luhn proposed two features to determine the importance of a sentence: the presence of significant words and the distance between these significant words. The significance of a word is based on its overall frequency in the original text. The distance is measured by the number of non-significant words between two significant words. The weight of a sentence is calculated by dividing the square number of significant words with the number of words occurring in the considered window. A window is a sub-sentence in which significant words occur with no more than four or five non-significant words.

Edmundson (Edmundson, 1969) calculated the weight of a sentence by adding the weights obtained from four features: Cue, Key, Title, and Location. The Cue method requires a predefined Cue dictionary containing selected words from a training corpus. The words are tagged as positive (stored in the Cue sub-dictionary called Bonus), negative (stored in the Cue sub-dictionary called Stigma), and irrelevant (stored in the Cue sub-dictionary called Null). In his experiment, Edmundson compiled the dictionary from 100 English documents. The Key method makes use of a Key glossary that contains high frequent words from the Cue dictionary. The Title method considers as positive all words that are non-Null and occurring in the title of the document as well as any other headings. The Location method uses a pre-stored Heading dictionary containing words that usually appear in headings such as "Introduction", "Purpose", "Conclusions", etc. Sentences are given positive weights if heading words occur in the sentences. In addition, sentences in the first and last parts of the first and last paragraphs are also given positive weights.

LexRank (Erkan & Radev, 2004) stands for "lexical PageRank". A text is modelled into a graph in which nodes are sentences and edges indicate the similarity relation between pairs of sentences. LexRank uses as a similarity function the cosine similarity to compute similarities between different sentences. Inspired by the concept of centrality or prestige in PageRank algorithm, a sentence is highly ranked if it is cited by many other highly ranked sentences. The summary is then formed by combining the top ranked sentences using a threshold or length cutoff.

EXPERIMENTS AND RESULTS

Datasets

14 Malay texts along with their corresponding abstracts (**Error! Reference source not found.**) were downloaded from several Malay blogs where primary and secondary school's teachers provide some hints to their students in preparing their national exams. The topics of the full text are economy, community, society, education, health, and sport. The average number of sentences of the 14 texts is 22 with the longest text having 29 sentences and the shortest text 17 sentences. Each abstract is made of three sections: an introduction, a main content, and a conclusion. For this study, only the main content was retained. The set corresponds to Abstract-H. The average number of sentences in the abstract is eight with the longest abstract having 10 sentences and the shortest one four sentences. The second set of summaries corresponds to 20 extractive summaries obtained from four Malay native speakers (Extract-H1). Their task was to summarize five texts that were randomly selected from the 14 Malay texts. These native speakers were asked to provide an extract of each text that must have at least five sentences selected from the original text. However, Morris and colleagues

(Morris et al., 1992) stated that "extracts containing 20% or 30% of original document are effective surrogates of original document". Therefore, the 20 extracts were reduced so that each of them will contain only 20-30% of the original text. This third set of extracts corresponds to Extract-H2. The reduction process is based on the CR formula as used by Jusoh and colleagues (Jusoh et al., 2011). The other sets of summaries were obtained by running the three selected extractive summarization methods on the 14 Malay texts. The results are named Extract-L, Extract-E, and Extract-LR from Luhn, Edmundson, LexRank methods respectively. To summarize, six kinds of summaries are used for the comparative study as shown in Table 1.

Table 1. Summary of the Dataset.

	Name	Description
(1)	Abstract-H	Human abstractive summary
(2)	Extract-H1	Human extractive summary (not compressed)
(3)	Extract-H2	Human extractive summary (compressed)
(4)	Extract-L	Luhn extractive summary
(5)	Extract-E	Edmundson extractive summary
(6)	Extract-LR	LexRank Extractive summary

Evaluation Metric ROUGE-N

The metric ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) has been selected as it is widely-used for measuring summary. "ROUGE is a recall-based metric for fixed-length summaries which is based on n-gram co-occurrence" (Erkan & Radev, 2004), and thus the name ROUGE-N, which is available in the ROUGE 2.0 evaluation toolkit. ROUGE-1 has been found to be a good measure for single document very short summary (Erkan & Radev, 2004), which is the case of the study reported in this paper. ROUGE-1 looks for unigram overlapping. To get a fair evaluation with ROUGE, the length of the summaries needs to be limited. ROUGE outputs three values: recall, precision, and F-score. The ROUGE-1 F-score for the summarization of clinical text notes using different summarization methods is in the range of [0.28-0.48] (Moen, et al., 2016). The ROUGE-1 F-score for different variations of TexRank summarization algorithm is in the range of [0.3903-0.4108] (Barrios et al., 2015). The ROUGE-1 F-score of LSA-based text summarization when applied on blog posts is 0.22 on negative posts and 0.21 on positive posts (Balahur et al., 2009).

Machine Extracts vs. Human Abstracts

In this evaluation, the Malay human abstracts, considered as the reference summaries, are compared to the Malay automatic extracts. The results of the evaluation are shown in Table 2.

Table 2. ROUGE-1 F-score: Abstract-H vs. Extract-L / Extract-E / Extract-LR

	Extract-L	Extract-E	Extract-LR
Max.	0.42	0.33	0.41
Min.	0.25	0.10	0.26

Avg.	0.36	0.22	0.35

Machine Extracts vs. Human Extracts

In this evaluation, the 20 extracts from the four Malay native speakers are compared against the extracts generated by the three selected extractive summarizers. The F-score results are shown in Table 3 when the human extracts are not compressed and Table 4 when the same human extracts are compressed to represent only 20-30% of the original text.

Table 3. ROUGE-1 F-score: Extract-H1 vs. Machine Extracts.

	Extract-L	Extract-E	Extract-LR
Max.	0.51	0.31	0.42
Min.	0.45	0.16	0.34
Avg.	0.43	0.24	0.38

Table 4. ROUGE-1 Result: Extract-H2 vs. Machine Extracts.

	Extract-L	Extract-E	Extract-LR
Max.	0.67	0.32	0.65
Min.	0.41	0.19	0.34
Avg.	0.48	0.25	0.42

DISCUSSION AND CONCLUSIONS

In all experiments, Luhn and LexRank methods are nip and tuck with Luhn's method having the best performance. Moreover, the implementation of Luhn's method is simpler than the implementation of LexRank method. Edmundson's method performed very poorly for certainly one main reason. No lexicons (Cue, Key, and Heading wordlists) required by this method have been created specifically for Malay texts. Instead, a Malay stopword list has been used to replace all these lexicons. The topic of a text is not known in advanced, and thus, the classification of words as, for example, positive (Bonus words) and negative (Stigma words) is not straightforward. In addition, none of the Malay texts fed to the Edmundson's method has a title. Therefore, Edmundson's method is not applicable for Malay texts without pre-defined lexicons and original texts with some titles and headings.

A machine extract is gist summary of a full text, like Google Translate result is a gist translation of a source language text. As such, machine extracts should not be viewed as "imperfect" since they provide a quick glance of the content of a lengthy text. Therefore, if one needs a quick summary of a Malay text, Luhn's method is simple and can provide an "acceptable" gist summary. One reason that makes extractive summaries not appealing is that they present the information in disconnected and incoherent way. A possible improvement is to use some of the findings of Alias and colleagues (Alias et al., 2016) while studying a Malay text corpus made of news articles along with the corresponding human summaries.

Comparing human abstracts and automatic extracts is a little bit strange since an extract would never be better than an abstract, and the conception of an abstract is not the same as the conception of an extract. But, this kind of experiments can highlight the big gap as well as the similarity between an abstract and an extract. None of the values of ROUGE-1 F-score is equal to zero. It means that some fragments of the source texts have been copied by the school's teachers to be put in their abstracts. The highest average F1-score is 0.36 (from Luhn's method). This is not a bad score recalling that the ROUGE-1 F-score for a LSA-based text summarization when applied on blog posts is only 0.22 on negative posts and 0.21 on positive posts (Balahur et al., 2009). ROUGE is a recall-based metric. The highest recall value obtained by Zamin and Ghani (Zamin & Ghani, 2010) with their proposed Malay hybrid summarizer is 0.38 for the summarization of general news. The lowest recall value is 0.223 for business news.

By equalizing the length of the machine summaries and human summaries, Luhn's and LexRank methods show a substantial improvement (Table 3 and Table 4). This can be due to the fact that ROUGE metric works well when the evaluated summaries have equal length. This yields to a problem. By constraining the length of an automatic summary, we limit the capability of an ATS, and thus, some salient sentences may be missed.

REFERENCES

Alias, S., Mohammad, S. K., Gan, K. H., & Tan, T. P. (2016). A Malay Text Corpus Analysis for Sentence Compression Using Pattern-Growth Method. *Jurnal Teknologi, 78*(8), 197-206.

Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R., & Montoyo, A. (2009). Summarizing Opinions in Blog Threads. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, (pp. 606-613).

Barrios, F., López, F., Argerich, L., & Wachenchauzer, R. (2015). Variations of the Similarity Function of TextRank for Automated Summarization. *Proceedings of the 16th Argentine Symposium on Artificial Intelligence (ASAI)*, (pp. 65-72).

Das, D., & Martins, A. F. (2007). *A Survey on Automatic Text Summarization.* Technical Report. Retrieved from https://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the ACM (JACM), 16*(2), 264-285.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research, 22*, 457-479.

Hovy, E., & Lin, C.-Y. (2000). Automated Text Summarization in SUMMARIST. *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, (pp. 18-24).

Jusoh, S., Masoud, A. M., & Alfawareh, H. M. (2011). Automated Text Summarization: Sentence Refinement Approach. In V. Snase, J. Platos, & E. El-Qawasmeh (Eds.), *ICDIPC 2011, Part II, CCIS 189* (pp. 207-218).

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS).* Barcelona, Spain. Retrieved from http://anthology.aclweb.org/W/W04/W04-1013.pdf

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development, 2*(2), 159-165.

Moen, H., Peltonen, L.-M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T., & Salanterä, S. (2016). Comparison of Automatic Summarisation Methods for Clinical Freetext Notes. *Artificial Intelligence in Medicine, 67*, 25-37.

Morris, A. H., Kasper, G. M., & Adams, D. A. (1992). The Effects and Limitations of Automated Text Condensing on Reading Comprehension Performance. *Information Systems Research, 3*(1), 17-35.

Müürisep, K., & Mutso, P. (2005). ESTSUM - Estonian newspaper texts summarizer. *Proceedings of the Second Baltic Conference on Human Language Technologies*, (pp. 311-316). Tallinn.

Zamin, N., & Ghani, A. (2010). A Hybrid Approach for Malay Text Summarizer. *Proceedings of the 3rd International Multi-Conference on Engineering and Technological Innovation (IMETI).* Orlando, Florida, USA.