Journal of Theoretical and Applied Information Technology
10th June 2015. Vol.76. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MALAY TEXT FEATURES FOR AUTOMATIC NEWS HEADLINE GENERATION

[1]MOHD SABRI HASAN, [2]SHAHRUL AZMAN MOHD NOAH, [3]NAZLENA MOHAMAD ALI

[1]Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43000 Malaysia

[2] Prof., Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia, 43000 Malaysia

[3]Assoc. Prof., Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43000 Malaysia

E-mail: [1]mohdsyberi@gmail.com , [2]shahrul@ukm.edu.my , [3]nma@ukm.edu.my

ABSTRACT

The diversity of natural language for meaning representation in documents is one of the causes of information overload in information retrieval. Headline generation is an automatic text summarization technique that can reduce or address such a problem. This research discusses an experimental study on the determination of Malay language characteristics from news genre documents. A Malay news corpus comprising 140 news documents was chosen from the BERNAMA news archive. The selection criteria were limited to hard news, a word count of 50 to 250 words, published between 2007 and 2012, and with news genres of economy, crime, education, or sports only. Three Malay linguistic experts were selected to produce a reference headline for each news document manually. Experiment results identify three characteristics. First, the first two sentences of a news document are suitable candidates for the most important sentences; second, sentences that contain an acronym definition also have the potential to become the most important sentences; and third, the ideal length of a headline is six words. Considering these characteristics will generate intelligent headlines for Malay news.

Keywords: *Headline Generation, Text Summarization, Malay News*

1. INTRODUCTION

Information and communications technology (ICT) nowadays has rapidly allowed information to be created, processed, copied, and disseminated in various languages. However, variations in languages cause critical information overload in information retrieval systems. This problem results from the differences between natural languages in delivering meanings. The volume of documents rapidly increases in various deliverable platforms, particularly the web. Huge document collections exacerbate the problem of information overload, in which the understanding of certain issues or decision-making processes becomes difficult because of the presence of too much information. Among the techniques proposed in the information retrieval field to address information overload problems is summarization. Within the context of textual documents, text summarization is the process of restating the main ideas in a document in as few or reasonable words as possible. Therefore, the automation of the summarization process involves creating a summary that retains the most important points of the original document. The output of summarization varies from snippets and titles to passages and headlines. This study focuses on automatic headline generation.

Headlines are the most important element in news. Headline generation produces a document version to reduce users' cognitive burden by including only important information in the simple version and by minimizing repetitive information. According to [1], headline generation is a potential tool to reduce information overload because this technique manages relevant document content in relation to users' needs. The technique also reduces cognitive burden during the scrutiny and selection of large volumes of documents.

Headlines are represented as a single sentence that summarizes the entire idea of a piece of writing, and as such, headlines are widely used for news genre. In practice, however, headline generation is usually adjusted to meet its intended purpose and focus and to cater to its expected audience, as determined by the news agency concerned [2].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A single document extraction method that generates headline consists of three processes, namely, selection of the significant words, selection of the most important sentences, and sentence realization. The complexity of this technique completely depends on the language system that represents the information in the document. The selection of the most important sentences aims to determine the main idea of the writing, whereas the selection of the most important words aims to identify the most important terms that should be found in the headline. Sentence realization is the final process that involves the reconciliation of morphology with the sentences chosen to form perfectly united sentences.

Most studies on headline generation have focused on proposing novel methods and techniques to improve the quality of headlines [3] [4] [5] [6] [7]. The features of the specific documents to be summarized, the language, and the genre need to be analysed and identified to automatically generate quality headlines. Ironically, this aspect has been largely ignored in the existing literature. To date, a significant work in this area is demonstrated in the study of [8], which focused on the characteristics of natural language in a headline generation technique for English news. This study suggests that the main idea of a news story is found on the first four sentences of the document, and the average length of a headline should be 10 words. The findings of this research have been used in most non-English news nowadays [9] [10]. However, different languages have different approaches to express the meanings of sentences or documents, so the quality of the generated headline for non-English text has yet to be improved.

By using a series of experiments, this study therefore aims to identify the characteristics of Malay language news that must be considered in headline generation. The experiments conducted are an adaptation of those done by [8]. The results are a set of proposed features and characteristics that emulate the human-expert approach and that are highly desirable to generate quality news headlines for the Malay language. This study consists of five sections. The first section contains the introduction, and the second one presents the research background and related works. The third section introduces the methodology of the experiment, and the fourth one discusses the results of the experiments. Finally, the fifth section concludes the characteristics of the Malay system that should be considered in the headline generation technique for Malay news genre.

2. BACKGROUD AND RELATED WORKS

The earliest automatic text summarization research was conducted by [11]. He used a statistical method to produce a short document version. However, this work did not receive attention in the field of computer science until the early 1990s due to the widespread use of the Internet and World Wide Web technologies, which resulted in the problem of information overload. The main factor involved in this challenge is the capability of technology to allow information to be created in many natural languages. In this regard, research has been oriented toward information retrieval, natural language processes, and artificial intelligence [1] [12].

Seminal work on headline generation techniques has been conducted for English news [3] [13] [14] [15]. Other languages, the simplest way to generating headlines is to employ the machine translation methods and exploit the headline generation techniques proposed for the English news, as illustrated in the works of [16] and [17]. In this approach, news is automatically translated into the English language, and the short document version is generated by the existing headline generation technique. Given the limitations of the natural language system, however, the results cannot be translated back into the original language.

The incorporation of machine translation method and existing headline generation techniques is made in the hope to develop an automatic text summarization technique for multiple natural languages without any human involvement. However, the uniqueness of the natural language factor becomes the main obstacle to the realization of this goal, which involves the following drawbacks: difficulty in determining the important sentences for translated documents [16], problem in accurately identifying the most important sentences based on the machine translation [17], and the challenge of ensuring cohesion in the translated sentence, which results in the failure to deliver the real meaning [18] [19].

Researchers have performed many studies on the headline generation technique based on the natural languages in which the document information is represented and without the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

involvement of machine translation. Shamfard et al. [19] and Alotaiby et al. [20] demonstrate that a text generation technique dealing with the Arabic language without the use of machine translation able to produces good results.

Our literature review shows that only [21] conducted research on automatic text summarization that is exclusive to Malay documents. However, the technique employed still incorporates machine translation methods without considering Malay language characteristics that represent the information in the document. Given the limitations of the natural language system, the results cannot be translated back into Malay.

According to [22], recent studies on headline generation have been oriented toward information extraction that incorporates the headline generation technique, natural language processing, and machine learning. Two information extraction methods have been used, namely, the supervised and unsupervised methods. Information extraction using the supervised method requires a set of training data to determine the most important sentences in particular texts. Models are constructed according to the characteristics that are identified to allow predication of the sentence selection made. The model must be capable of learning how to predict the most important sentences in texts from any set of training data. Through this model, the most important sentences are identified, and a short document version is generated. However, for languages such as Malay whereby the training data is rather limited, supervised methods usually produced unsatisfactory results.

Therefore, recent methods, especially those on headline generation, require a profound understanding of the natural language system characteristics that represent the information in documents. Understanding these characteristics involves identifying the morphology in a particular sentence structure and information on sentence syntax and syntax formulas that must be used to generate a perfect sentence. Comprehending the characteristics of the natural language stem, the natural language processing technique, and the machine learning technique allows the development of intelligent headline generation techniques. Subsequently, these techniques are expected to execute the generation task perfectly and produce results similar to those generated by humans.

3. METHODOLOGY

The Malay news corpus used in this experiment consists of 140 documents that were selected from the BERNAMA news corpus. The four criteria used in the selection were the news category, size, publication year, and genre. News documents are basically divided into three categories: hard news, soft news, and investigation news. Hard news generally refers to up-to-the-minute news and events that are reported immediately, whereas soft news generally refers to background information or human-interest stories. The hard news category was chosen because it is considered the main news, and 80% of this category is published in different forms. All the chosen hard news consisted of between 50 and 250 words and was published between 2007 and 2012. The four genres considered in this experiment were economy, crime, education, and sports.

Three Malay linguistic experts were involved in analysing the news and determining the most important sentences for each news document available in the corpus. They were then required to produce the headline for each document. The original headlines for each document were not made available to the experts to avoid any biases. While producing the headline, the experts must comply with three conditions: headline generation must be conducted with the use of the extraction method, the word selection technique should be based on the select-word-in-order technique, and word morphological changes are allowed, as stated in the experiment of [8]. The generated headlines are considered ideal headlines that contain the main ideas for each piece of writing in the corpus.

The results of the experiment were analysed and compared with those obtained by [8] and were further discussed in Section 4 of this paper. The findings are regarded as the characteristics of the Malay language that should be considered in automatic methods for news headline generation.

4. RESULT AND DISCUSSION

Overall, 94.6% of the headlines produced by the experts follow the three guidelines mentioned in the previous section, with the exception of five headlines that did not comply with the select-word-in-order technique guideline. Therefore, the word orders in the resulting headlines were not parallel with those in the important sentences retrieved, as illustrated in Table 1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1. Headline non-compliance with the experimental condition

Document	Important Sentence	Headline
BM077	*Transformasi[1] kepada[2] sistem[3] pendidikan[4] negara[5] perlu[6] dilaksanakan ...*	*Sistem[2] pendidikan[3] negara[4] perlu[5] Transformasi[1]*
BM099	*... sukan SEA ke-26 yang memenangi[1] pingat[2] emas[3] selepas menewaskan Indonesian kelmarin, menerima ganjaran[4] wang tunai RM30,000[5].*	*Ganjaran[4] RM30,000[5] menang[1] Pingat[2] emas[3]*
BM119	*Bagi melahirkan[1] lebih ramai[2] usahawan[3] tempatan berdaya saing, kolej[4] komuniti[5] kini ...*	*Kolej[4] komuniti[5] lahir[1] ramai[2] usahawan[3]*
BM131	*Kementerian[1] Pelajaran[2] menyambut[3] baik[4] hasrat Universiti Kebangsaan Malaysia (UKM)[5] dengan[6] kerjasama[7] Permata[8] Negara[9] membangunkan*	*Kementerian[1] Pelajaran[2] sambut[3] baik[4] kerjasama[7] UKM[5] dengan[6] permata[8] pintar[9]*
BM135	*India[1] mengakui mereka semakin menghormati Malaysia[2] dan lebih berwaspada[3] apabila berhadapan[4] tuan rumah ...*	*India[1] berwaspada[3] berhadapan[4] Malaysia[2]*

The analysis shows three important characteristics of the Malay news document genre: the most important candidate sentences, the definition of acronyms in sentences, and the headline length.

4.1 The Most Important Sentences

The analysis of the 140 headlines by the experts indicates that only the first two sentences of the documents are selected as the most important sentences that represent the main idea. Figure 1 shows that 97.1% of the most important sentences are usually the first sentence, whereas 2.9% are the second sentence. The results of the experiment of [8] show that the main idea in English news is presented in the first sentence or the first 50 words of each news document.

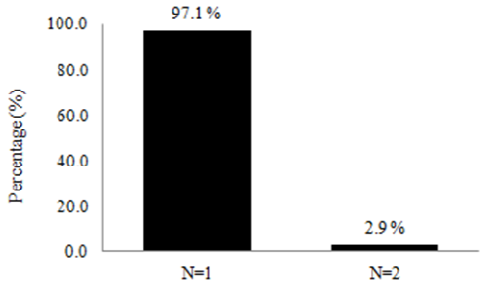

Figure 1. Number of important sentences.

The writing of Malay news in Malaysia is consistent with the results of the experiment. According to [2], the main idea in Malay news writing is in the news lead, and the headline is the summary of this news lead. The news lead is the first sentence of the first paragraph in any Malay news document [23]. Therefore, the first two sentences are considered the most important sentences in generating headlines for Malay news documents.

4.2 Acronym Definition In Sentences

An acronym is an abbreviation that combines the initial letters or syllables of words and is acceptable as one word [24]. In the Malay language system, an acronym is considered a single word that can be formed out of many words. In news writing, acronyms must be defined earlier before they can be used directly in a sentence. Figure 2 illustrates an example usage of an acronym. In this case, the phrase "Jabatan Perkhidmatan Awam" is represented as JPA.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AMPANG: Pelajar cemerlang yang mendapat 9A+ Sijil Pelajaran Malaysia (SPM) tahun lalu, tetapi tidak memperoleh biasiswa Jabatan Perkhidmatan Awam (JPA), diminta memaklumkannya kepada JPA untuk ditawarkan biasiswa itu.

Figure 2. Acronym definition in sentences.

Further analysis of the usage of acronyms in 140 news documents and headlines reveals that 45% of the news documents contain acronyms in the first sentence, the second one, or both; and that 37% of the headlines also contain acronyms. Ninety-eight percentages of the headlines were generated from the important sentences that contain acronyms. This result indicates that sentences containing acronyms have the potential to be considered the most important sentences in Malay news documents. However, this finding was not highlighted and discussed in the experiments of [8].

4.3 Headline Length

The Document Understanding Conference / Text Analysis Conference is one of the most active international communities that conduct research on automatic text summarization. This community set the ideal length of headlines for English news documents to 10 words [25]. This length was based on the work of [8]. However, the results of the experiments in the present study show that the average length of headlines produced by experts is six words, as illustrated in Figure 3. This length is similar to the standard one for writing headlines, as practiced by Malay news providers in Malaysia [2].

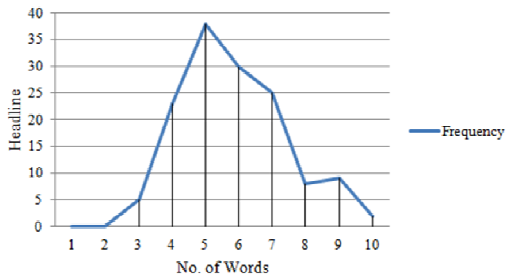

Figure 3. Number of Words.

5. CONCLUSION AND FUTURE WORKS

In this study, we have conducted an experiment to determine the characteristics of the Malay language system that must be considered in the generation of headlines for Malay news documents. These characteristics are as follows. First, the first two sentences of a news genre document are the best candidates for the headline generation technique. Second, sentences that contain an acronym definition also have a high potential to be selected as the most important sentences. Third, the appropriate length of headlines for Malay news is six words. The most important sentences are considered suitable for headline generation because they contain the main idea of the document. Considering these three characteristics is expected to produce consistent and accurate headlines. Our future works will focus on the development of methods and systems for the automatic headline generation of Malay documents on the basis of the three characteristics identified in this study.

REFRENCES:

[1] K. Kaikhah, "Automatic text summarization with neural networks", *Second International IEEE Conference on Intelligent System*, 2004, pp. 40 – 45.

[2] S. N. A. Rahman, "Kewartawan Malaysia: Praktis & cabaran dalam era revolusi digital", *Prentice Hall*, 2009.

[3] M. Banko, V. O. Mittal & M. J. Witbrock, "Headline generation based on statistical translation", *Proceedings of the 38th Annual Meeting on Association for Computational Linguistic (ACL-00)*, 2000, pp. 318 – 325.

[4] L. Zhou & E. Hovy, "Template-filtered headline summarization", *Proceedings of the Association for Computational Linguistics (ACL-04) Workshop on Text Summarization Branches Out*, 2004, pp. 56 – 60.

[5] K. Muurisep & P. Mutso, "ESTSUM – Estonian newspaper texts summarizer", *Proceedings of the Second Baltic Conference on Human Language Technologies*, 2005, pp. 311 – 316.

[6] R. Soricut & D. Marcu, "Abstractive headline generation using WIDL-expressions", *Information Processing and Management*, Vol. 43, No 6, 2007, pp. 1536 – 1548.

[7] S. Xu, S. Yang & F. C. M Lau, "Keyword extraction and headline generation using novel word features", *Proceedings of the Twenty-Fourth AAAI*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Conference on Artificial Intelligence 2010, 2010, pp. 1461 – 1466.

[8] D. Zajic, B. Dorr & R. Schwartz, "Headline Generation For Written And Broadcast News", *Thechnical Report of the Language and Media Processing Laboratory*, Institute for Advanced Computer Studies, University of Maryland, Report No: UMIACS-TR-2005-07, 2005.

[9] F. A. Alotaiby, "Automatic headline generation using character cross-correlation", *Proceedings of the Association for Computational Linguistics – Human Language Technology (ACL-HLT 2011) Student Session*, 2011, pp. 117 – 121.

[10] J. A. Daniel, "Headline generation for Dutch newspaper articles through transformation-based learning", *Master Thesis, Information Science, University of Groningen*, 2008.

[11] H. P. Luhn, "The automatic creation of literature abstract", *IBM Journal Research and Development*, Vol. 2, No. 2, 1958, pp. 159-165.

[12] J. Steinberger & K. Jezek, "Evaluation measures for text summarization", *Computing and Informatics*, Vol 28, 2009, pp. 1001 – 1026.

[13] D. Zajic, B. Dorr & R. Schwartz, "Automatic headline generation for newpapers stories", *Proceedings of the Document Understanding Conference 2002*, 2002.

[14] L. Zhou & E. Hovy, "Template-filtered headline summarization", *Proceedings of the Association for Computational Linguistics (ACL-04) Workshop on Text Summarization Branches Out*, 2004, pp. 56 – 60.

[15] D. Marcu, "From discourse structures to text summaries", *Proceedings of the Association for Computational Linguistics (ACL-97) Workshop on Intelligent Scalable Text Summarization*, 1997, pp. 82-88.

[16] B. Dorr & D. Zajic, "Cross-language headline generation for Hindi", *ACM Transaction on Asian Language Information Processing*, Vol. 2, No. 3, 2003, pp. 270 – 289.

[17] F. S. Dauzidia & G. Lapalme, "Lakhas, an arabic summarization system", *Proceedings of the Document Understanding Conference 2004*, 2004.

[18] N. Mazdak, "A Persian Text Summarizer", *Master Thesis. Department of Linguistics, Stockholm University*, 2004.

[19] M. Shamfard, T. Akhavan & M. E. Joorabchi, "Persian document summarization by PARSUMIST", *World Applied Sciences Journal 7 (Special Issue of Computer & IT)*, 2009, pp. 199 – 205.

[20] F. Alotaiby, S. Foda & I. Alkharashi, "A new approach to HMM-based automatic headline generation", *Proceedings of the Computational Linguistics - Applications Conference 2011*, 2011, pp. 3 – 8.

[21] N. Zamin & A. Ghani, "Summarizing malay text documents", *World Applied Science Journal 12 (Special Issue on Computer Applications & Knowledge Management)*, 2011, pp. 39 – 46.

[22] D. Shen, "Encyclopedia of Database Systems", *Springer*, 2009.

[23] M. S. Kassim, "Kewartawan: Teori & praktis", *Dewan Bahasa dan Pustaka*, 1985.

[24] N. S. Karim, F. M. Onn, H. H. Mohammad & A. H. Mahmud, "Tatabahasa dewan edisi ketiga", *Dewan Bahasa dan Pustaka*, 2010.

[25] P. Over, H. Dang & D. Harman, "DUC in context", *Information Processing and Management*, Vol. 43, No. 6, 2007, pp. 1506 – 1520.