World Applied Sciences Journal 12 (Special Issue on Computer Applications & Knowledge Management): 39-46, 2011
ISSN 1818-4952

Summarizing Malay Text Documents

Norshuhani Zamin and Arina Ghani

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia

Abstract: Summarization is the art of generating the main points of a lengthy text document by removing redundant and less important information without losing the meaning of the original text. Summaries are significantly shorter than the original text and take a broad overview of the source material. With the increasing volume of digital information today, people find the manual process of summarization as hectic and time consuming. Having an automated text summarization system for electronic documents would very much help to encourage people to read, giving quick access to information thus helping them to a faster decision making process. Although many research and commercial text summarization tools are available, no research is officially reported for Malay language. Malay text summarizers are coming into demand when a lot of information in Malay language can now be accessed freely via the Internet. This paper presents a hybrid approach to an automated text summarization system for Malay language. The base system is built on SUMMARIST system and is expanded by combining with EstSum system. Experimental results show that expanding training data size significantly contributes to the performance. In general, our system produced acceptable results at the best case of 76% and the worst case of 31%.

Key words: Malay Text Summarization · Statistical Approach · Text Mining · Natural Language Processing

INTRODUCTION

Based on the survey of Malaysian reading profile in 1996 done by the Malaysia National Library, the average of reading activity for Malaysian is approximately two books in a year. Further survey made in 2005 on 60,441 Malaysian reported that the 'problem' has not shown significant improvement. Malaysia is far behind among most of well-develop countries in reading activities. One of the major causes to this problem is the reading habits develop very slowly in low income family as compared to higher income family. The mushrooming of the digital Malay texts in the Internet motivates us to develop an automated summarization system to encourage the reading habits of Malaysian by consuming less time to read lengthy documents.

The earliest research in text summarization was done in 1960s and the growth of the interest in this research continues in recent years. Most of the work found is for English text summarization but with the increasing demand of this tool for other languages, a number of research and development found for Estonian [1] Scandinavian [2], Thailand [3], Persian [4], Swedish [5] and the five in one system known as SUMMARIST [6] for summarizing Japanese, Arabic, Spanish, Indonesian and Korean text embedded with English translation feature.

A perfect summary is depending on users' requirement whether to have an indicative and informative summary. Indicative summary is a summary that highlights only the topic of the text while informative summary describes the central information in the text [7]. As this research is concerned with the synopsis of the text, the results shall produce informative summary. However, a development of text summarization tool for Malay language that has totally different grammatical structures than English is not only the challenge but also the accuracy of the results will be the important issue to be discussed. How good a summary is depending on the percentage of essence preserved in the summary and cohesion between one sentences to another. As it is difficult to measure the quality of the Malay summaries with no baseline research in existence, therefore, the outcome of this research will be compared with the analysis of human-made summaries.

Corresponding Author: Norshuhani Zamin, Department of Computer and Information Sciences Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia, E-mail: norshuhani@petronas.com.my.

Text mining refers to computational methods to discover previously unknown meaningful information from unstructured text. Text mining is closely related to data mining – finding interesting patterns and trends in a large dataset. The only different is text mining deals with natural language text while data mining requires structured databases or facts. The purpose of text mining is to link together the extracted information to form new facts or new hypotheses to be explored further [8].

In recent years, research in text mining covers diverse areas which include term association discovery, document clustering, text summarization and text categorization. Text mining consists of three basic steps: a) Text Preparation – the preprocessing of text to extract meaningful terms or features b) Text Processing – the use of computational methods to identify interesting patterns in preprocessed text c) Text Analysis – the evaluation of extracted output [9].

Text summarization is "the process of distilling the most important information from a text and to produce an abridged version for a particular task and user" [21]. Automatic text summarization refers to the use of computational methods to automatically derive the summary of a given text. Over the past half a century, text summarization research has been explored by the Natural Language Processing (NLP) community. The increasing availability of online information has necessitated intensive research in this area.

There are two main methods to automatic text summarization, abstraction and extraction [10]. Abstraction is a difficult technique yet promising where it generates new sentences from the original sentences through a process called paraphrasing. This technique involves syntactic and semantic study for the particular language and is useful for meaningful applications. While extraction method has been the current state of the art and commonly used by most of the existing tools. This method weighted each sentence in the original text with some specified characteristics and selects the original sentence and juxtapose in the summary. The basic of extraction-based summarization where each sentence is measured through special predefined properties, selecting the most relevant sentences based from the value of the properties and put them together in a summary. This research learnt the basic extraction based technique from [11].

A research in [12] categorized the approaches of the automatic text summarization into three: a) Shallow Approach – the simplest approach of all approaches where a summary is produced by extracting sentences from information source. However, the challenge in this approach is to preserve the original context when breaking the sentences b) Deeper Approach – this approach produces summary called 'abstract' where some of the text in summary may not be found in the original text. It finds the most specific generalization of concepts from texts and uses this for the summary and c) Hybrid Approach – a combination of two more existing computational methods and techniques. This research proposed a hybrid approach based from a multilingual text summarizer known as SUMMARIST [6] and an Estonian text summarizer known as EstSum [1]. Among the aims of text summarization are the single document and multiple documents summarization [23]. In the single document summarization, a summary that characterized the content of a single document is produced. Whilst, the multiple document summarization takes a group of documents as input and a condensation of the content of the entire group is produced as the summary. Multiple-document summarization has turned out to be much more complex than summarizing a single document. This research is focusing on the single document summarization.

Malay is not only a native language for Malaysia but also one of the languages used in Indonesia, Brunei, Singapore and southern Thailand. The Malay language is rich in colloquial, idiomatic expressions and literary allusions and like other languages, it possess its own unique structure and grammar. As the Malay language is used within the South East Asia region, it has become one of the less resourced languages in the world. Due to this, limited number of computational linguistic research was found related to Malay language. Although there are many studies related to Malay language, however, to the best of our knowledge, none has been officially reported on Malay text summarization. However, other computational linguistic studies for Malay language exist such as the Information Retrieval [13, 28], Essay Marking [14], Novelty Detection [15], Machine Translation [16], Corpus Analysis [29] and. Recently, an open source tool for Malay language corpus analysis was found in an ongoing research [17]. The tool provides access to Malay tokenisers, lemmatisers and part-of-speech taggers that are vital for the Malay linguistic research. A review on existing text summarizers in other languages is also conducted to investigate potential method for hybrid purpose.

SUMMONS [18] is the first example of multi-documents summarization system. It summarizes news articles about terrorism from different news agencies and produces a briefing merging relevant information about

each identified event. SUMMONS architecture consists of two major components: a) Content Planner – selects the important information from the combination of input templates with instantiated slots of predefined semantics and b) Linguistic Generator – selects the right words to express the information in grammatical and coherent text. An automated text summarization system called EstSum [1] is able to summarize newspaper articles in Estonian language. It constructs short summaries of text by selecting the key sentences that describes the document. The sentences are classified using a weighted combination of statistical, linguistic and typographic aspects such as the position, format and type of sentence and the frequency of each word appeared in the system. It achieves up to 60% accuracy on the evaluation done against the human-made summaries or newspaper articles. SUMMARIST [6] aims at generating both abstracts and extracts for arbitrary English and other languages texts. In this research extract is defined as "portions extracted verbatim of the original (they may be single words or whole passages)" and abstract as "novel phrasings describing content of the original (which might be paraphrases or fully synthesized text)". SUMMARIST combines statistical techniques and symbolic word knowledge derived from WordNet – a large lexical database of English. Its technique lies on the following 'equation':

$$\text{summary} = \text{topic identification} + \text{interpretation} + \text{generation} \quad (1)$$

The purpose of topic identification is to filter the input which retains only the most important central topics using various techniques such as stereotypical text structure, cue words, high frequency indicator phrases and discourse structure. Interpretation processes the topics, rephrases and compresses them. This process is vital to achieve further compaction and to remove redundancies, rephrase sentences and to merge related topics into more general one. Generation process aims to reformulate the interpreted data into new text. The SUMMARIST architecture is illustrated in Fig. 1.

SweSum [2] is a Swedish text summarizer. The sentences are extracted based on a combination of linguistic, statistical and heuristic methods. SweSum works in three different passes: a) Tokenization, Scoring and Keyword Extraction – the input text is split into sentences. Word boundaries are identified searching for periods, exclamation and question marks.

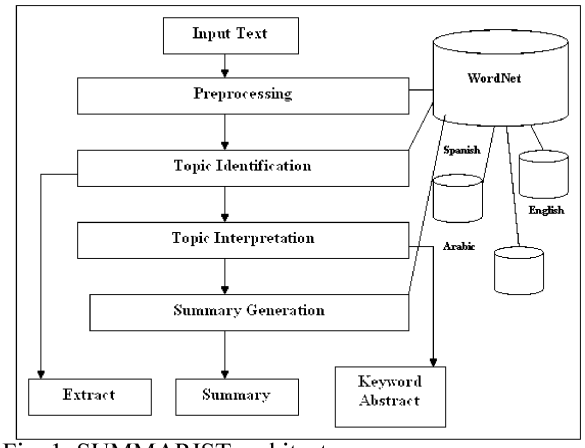

Fig. 1: SUMMARIST architecture

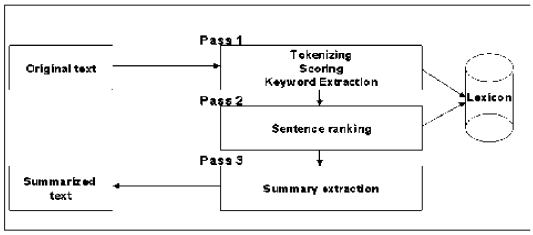

Fig. 2: SweSum architecture

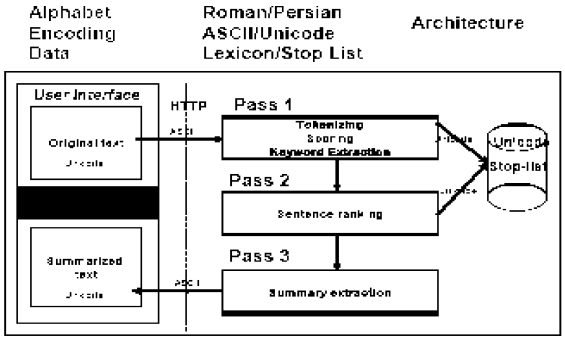

Fig. 3: FarsiSum architecture

The sentences are then scored by using statistical, linguistic and heuristic methods, b) Sentence Ranking - The score of each word in the sentence is calculated by a set of parameters, which can be adjusted by the user and total score is accumulated. Sentences containing common content words get higher scores and c) Summary Extraction – the final summary file is created in HTML format. These processes are schematically represented in Fig. 2. The lexicon is a database consists of key / value pairs where the key is the inflected word and the value is the stem / root of the word in Swedish.

FarsiSum [19] is a text summarizer for Persian built based on SweSum modules. The system is implemented as

a HTTP client/server application. The FarsiSum's tokenization module uses Persian's stop-list in Unicode format and a small set of heuristic rules. The stop-list is a file including the most common verbs, pronouns, adverbs, conjunctions, prepositions and articles in Persian. Fig 3 depicts the FarsiSum architecture with the each of the summarization steps numbered accordingly. The system is located on the server side and the client is a browser. The summarization steps are described as follows:

Step 1: The browser sends a summarization request to the Web server where FarsiSum is located. The URL of the document to be summarized is attached to the request where the original text is in Unicode format.

Step 2-5: The document is summarized in three phases similar to SweSum. However, words in the document are converted from ASCII to UTF-8.

Step 6: The summary is returned to the HTTP server. The browser then renders the summarized text to the screen.

An improved work on Persian summarization for single-document and multi-document using lexical chains and graphs using statistical heuristic methods to extract important sentences from the input texts is described in [30].

The objective of this research is to develop a Malay text summarizer giving the accuracy of at least 60% resemblances with the manual summaries. The development and evaluation considers summarization of various types of documents such as news articles, magazines, reports and story books in Malay language.

MATERIALS AND METHODS

In this research, we choose to work with extraction method. We believe that with further refinement to the formula, we can maximize the retention of the important information in the Malay text. We have considered adopting some techniques based from existing successful research. Basically we divide the whole summarization process into three phases: a) Preprocessing b) Text Extraction and c) Sentence Selection.

In the Preprocessing phase, we used the technique introduced in SUMMARIST [6]. The preprocessing algorithm considers only two modules in SUMMARIST: a) Tokenizer and b) Token Frequency Counter. There are two forms of Tokenizer which are word tokenizer and sentence tokenizer. Word tokenizer chunks each and every word in the input text to produce a set of tokenized text. The boundary of each word is determined based on the white space found between words. While sentence tokenizer chunks the text into sentences by taking the full stop (.) as the boundary between sentences. The tokenizer algorithm for Malay text is developed to recognize Malay words.

In the Token Frequency Counter module, the number of occurrence of each word appears in the original text is counted. The highest frequency will be considered as the keyword of the text. From this rank of word frequency, we select the first 10 words with the highest number of occurrence. Referring to these selected words, all sentences containing any of these words will be merged together as a preprocessed text. The original text is now simplified based on the frequency score of words. In the *Text Extraction* phase, we applied the Edmundson's statistical formula [20, 22] shown below and our main reference is the recent EstSum research [1]:

$$W(s) = \alpha P(s) + \beta F(s) + \gamma K(s) \quad (2)$$

Where,
$W(s)$ – weight function of sentences;
$P(s)$ – position-based score function;
$F(s)$ – format-based score function;
$K(s)$ – keyword-based score function;
α, β and γ are constants.

In this phase, each sentence in the preprocessed text is given three distinctive scores based on three properties identified by experts: a) Position – the score given to the location of the sentence. The regularities in text structures of many genres are helpful to rank the sentences based on their location in the text. For example, the first sentence appears in the text tend to contain important information and thus, will be given a higher score b) Format – the score given to the font style and format. For example, the word written in bold or italic will be given higher score as it addressed the importance of the word and c) Keyword – the score given to the frequency of word appeared in the text. Each of these scores is normalized using the following formula:

$$n = p.100 / t \quad (3)$$

Where,
n - normalized score;
p – assigned score for each distinctive features;
t - total assigned score for the text.

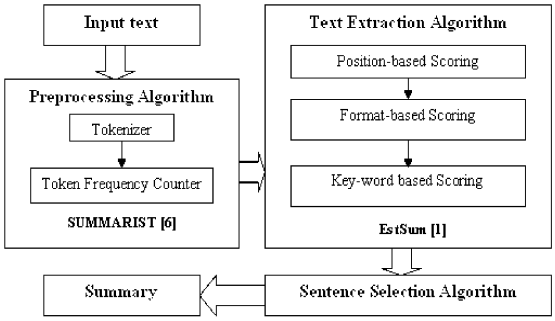

Fig. 4: Hybrid Malay Text Summarizer architecture

Then, each normalized score is multiplied with the constants - α, β, γ. These constants act as the tuning parameters and have been previously adjusted by hand using a manually created training corpus. A combination module combines the scores for each properties (Position score + Format score + Keyword score) and returns a single integrated score. This score gives a sentence a unique weight. Finally, in the Sentence Selection phase, by considering some threshold values, sentences with the higher score will be merged and taken as summary. The overall architecture of our proposed work is shown in Fig. 4.

Our experiment requires a training corpus. Due to the non existence of Malay corpus for text summarization, we create our own corpus consists of summaries compiled by four Malay language experts. A total of 10 original Malay news articles covering general, business and sports news were given to each of the four Malay language experts for manual summarization. Each expert submitted 10 summaries limited to 30% of the length of the source text giving 40 hand-created summaries. The process is illustrated in Fig. 5.

Scoring Mechanisms: The baseline of the scoring criteria is obtained from [1] with minor modifications made to suit the different structure of Malay text. As our training corpus of extracts is relatively small (only 40 summaries), we manually examined and compared each of the original text and its summaries. For the Position-based Scoring, we assigned the appropriate weight to a sentence by investigating the location of the sentence appeared in the summary using the following rules: a) The first 3 sentences of the original text b) The first sentence after each subtitle in the original text and c) The first 2 sentences of each paragraph the original text. An example of a Position-based Score for one of the summaries is shown in Table 1.

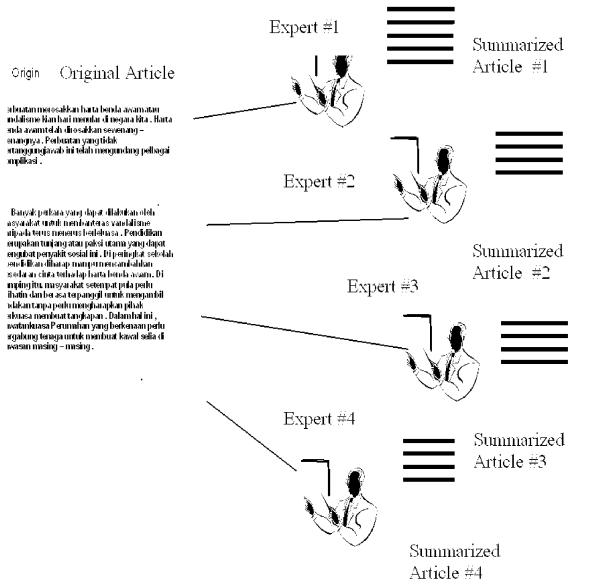

Fig. 5: Process of creating the training corpus

Table 1: An Example of a Position-based Score

Feature	% in extract	Given score (p)
1st sentence in article	100	10
2nd sentence in article	77	8
3rd sentence in article	45	5
1st sentence in paragraph	70	7
2nd sentence in paragraph	38	4
3rd sentence in paragraph	58	6

In addition to the Position-based Scoring, we defined a hypothesis that a paraphrased sentence found in an expert's summary will be treated as a multi-sentence during the manual investigation. Paraphrasing allows putting together multiple sentences in author's own word which is critical in natural language. This has been a challenge for text summarization research over a decade ago [22]. However, a research in [23] is found to be a promising start for automatically generating sentence paraphrases. At this development stage, we do not consider any paraphrasing in our system's generated summary.

For the Format-based Scoring, we considered the sentence based on the style of font (default, bold, italic) and punctuation marks (exclamation marks, question marks, double quotes). Unlike in [1], we excluded the score of the figure captions and text author as they were not present in our training data. Whilst for the *Keyword-based Scoring*, we used a general Malay word frequency table that is generated from all the 10 original texts by our *Token Frequency Counter* module. This helps to estimate

whether the word appears more frequently in original text than it normally does in the summarized text. This scoring takes the following rules: 1) The words belonging to the title (article headline) and subtitles are given higher scores and 2) All the other words are given similar lower weight.

Evaluation Metrics: The comparison of our proposed system's generated summaries is done against the human experts' summaries due to the non existence of commercially available Malay text summarizer. A survey in [24] describes and compares various human and automatic metrics to evaluate summaries. We employ the performance measures commonly used in the traditional Natural Language Processing task – Recall, Precision and F1 Score. These scores quantify how close the system's extract to human's. Precision shows the accuracy of the extracted sentence, Recall reflects how many good sentences the system has missed and F1 Score is a weighted average of the Precision and Recall [25]. Given an input text (original text), human-made summary and system generated summary, the following metrics are applied:

$$Precision\ (P) = correct / (correct + wrong) \quad (4)$$
$$Recall\ (R) = correct / (correct + missed) \quad (5)$$
$$F1\ Score = 2 \times (P \times R) / (P + R) \quad (6)$$

Where,
correct - the number of sentences extracted by the system and the human;
wrong - the number of sentences extracted by the system but NOT by the human;
missed - the number of sentences extracted by the human but NOT by the system.

The generated summary is judged correct if it contains sentences that were tagged in the human's summary or partially correct if the summary provides sufficient context for the passage. The generated summary is judged wrong if needed context was totally misleading or if the summary did not contain the expected passage at all. Finally, the generated summary is judged wrong if there is insufficient context for the passage. One standard marks a sentence as in the summary only when all four human experts agree.

RESULTS AND DISCUSSION

Table 2 shows the performance of our system. The second, third and fourth row of the table show the statistics of the three test collections. All the summaries agreed at a fixed-length compression rate of 30%.

Table 2: System performance evaluation

Type of Test Collection	*General News*	*Business News*	*Sports News*
No. of Document	4	3	3
Average no. of sentences per document	110	85	94
Average no. of sentences per summary	33	32	25
Precision (%)	85.2	77.6	66.3
Recall (%)	38.0	22.3	31.0
F1 Score (%)	76.0	31.0	42.2

The average number of sentence per summary in General News is relatively high in comparison with the other test collections because the number of sentences in the body of documents is higher. Consequently the performance of General News is better on average than the other test collections. The reasons, based from the feedbacks provided by the human experts, why General News outperforms others are as following: 1) They are different from each other in genre 2) The Business News provides an extremely thorough analysis such as the stock market, foreign exchange market and mutual funds. The technical expressions that are regularly iterated reduced the average score and 3) most summaries generated by the system are hard to understand. 4) The number of the training data is relatively small for drawing any final statistical conclusions. However, further investigation need to be done in future in order to reveal clearer reasons.

Accuracy is any natural language application such as text summarization, machine translation, speech processing is always a big issue. As for text summarization, evaluation is an important aspect to ensure whether the system has reached the goal to resemble the human made summaries. Indeed, naturally that it is hard to find two similar human made summaries for the same language in the world. A study in [26] found that at best there was about 70% average agreement between two human made summaries. On average, our system produces summaries that are about 50% similar to the manually created summaries. Although the agreement between the human summaries is quite low but it can be a promising start for a Malay text summarization research.

CONCLUSION

This paper presents a Malay text summarization system using a hybrid approach – the preprocessing module introduced by the SUMMARIST and the statistical scoring methods described in the EstSum's text

extraction module. Experiment shows that using the combination of both techniques, the system is able to extract the most important sentence from Malay news articles. This is a cost-effective solution to reduce users' consuming time in document reading without losing the general issues for users' comprehension. Summary helps users to easily decide its relevancy to their interests and acquire desired documents with less mental loads.

Since the research in this area is still at its immature stage there are many things to be investigated in the future. One of the problems that should be given a highlight is the widespread use of disparate metrics. It is found that there is no standard human or automatic evaluation metric in text summarization to compare different systems and establish a baseline. Hence, in future, to increase the decision accuracy, we plan to conduct the following evaluation as proposed in [27]: 1) Quantitative Measures – involve the categorization of decision relevancy, summary time and summary length and 1) Qualitative Measures- involve user preferences and detailed feedback as to why the summary was or was not acceptable for a given task.

ACKNOWLEDGMENTS

We would like to thank the journal's anonymous reviewers for their valuable comments. We are grateful to Dr. Lai Weng Kin of MIMOS, Malaysia; Professor Dr. Alan Oxley and Dr. Mohd Nordin Zakaria of Universiti Teknologi PETRONAS, Malaysia and Professor Kamaruzaman Jusoff of Universiti Putra Malaysia for their suggestions which helped improve this work. We are indebted to the four Malay language experts for their efforts to manually summarize given documents in time.

REFERENCES

1. Kaili, M. and M. Pilleriin, 2005. EstSum - Estonian Newspaper Text Summarizer. 2nd Baltic Conference on Human Language Technologies, Talliin, Estonia, 4-5 April 2005, pp: 399-408.
2. Hercules, D., 2000. A Text Summarizer for Swedish. In Technical Report, ReportNo: TRITANAP0015.
3. Jaruskulchai, C. and C. Kruengkrai, 2003. A Practical Text Summarizer by Paragraph Extraction for Thai. 6th International Workshop on Information Retrieval with Asian Languages, Sappro, Japan, 7 July 2003, pp: 9-16.
4. Mazdak, M., 2004. FarsiSum – A Persian Text Summarize. Master Thesis, Department of Linguistics, Stockholm University.
5. Dalianis, H., 2000. SweSum – A Test Summarizer for Swedish. In Technical Report, Report No: TRITANAP0015.
6. Eduard, H. and L.C. Yew, 1997. Automated Text Summarization in SUMMARIST. Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, 11 July 1997, pp: 18-24.
7. Sagiion, H. and G. Lapalme, 2002. Generating Indicative-Informative Summaries with SumUM. J. Computational Linguistics, 28(4): 497-526.
8. Hearst, M.A., 1999. Untangling Text Data Mining. The Association Computational Linguistic, University of Maryland, College Park, Maryland, 20-26 June 1999, pp: 3-10.
9. Chang, T.M. and W.F Hsiao, 2008. A Hybrid Approach to Automatic Text Summarization. 8th IEEE International Conference on Computer and Information Technology, University of Technology, Sydney, Australia, December 2008, DOI: 10.1109/CIT.2008.4594651.
10. Mani, I., 2001. Automatic Summarization. John Benjamins.
11. Luhn, H.P., 1958. The Automatic Creation of Literature Abstract, IBM J. Research and Development, 2(2): 159-165.
12. Diola, A.M., J.T.T.O. Lopez, P.F. Torralba, S. So and A. Borra, 2004. Automatic Text Summarization. 2nd National Natural Language Processing Research Symposium, De La Salle University, Manila, Phillipines, 28-29 January 2004, pp: 39-42.
13. Tai, S.Y., C.S. Ong and N.A. Abdullah, 2000. On Designing an Automated Malaysian Stemmer for the Malay Language. 5th International Workshop Information Retrieval with Asian Language, Hong Kong,China, 30 September – 2 October 2000, DOI: 10.1145/355214.355247.
14. Aziz, M.J.A., 2008. Pola Grammar for Automated Marking of Malay Short Answer Essay-Type Examination. PhD Thesis, Universiti Putra Malaysia, Malaysia.
15. Kwee, A.T., F.S. Tsai and W. Tang, 2009. Sentence Level Novelty Detection in English and Malay. Lecture Notes in Artificial Intelligence, Springer-Berlin Heidelberg, 5476: 40-51.
16. Tong, L.C., 1986. English-Malay Translation System: A Laboratory Prototype. 11th International Conference on Computer Lingustics (COLING), University of Bonn, Germany, 25-29 August 1986, pp: 639-642.

17. Baldwin, T. and S. Awab, 2006. Open Source Corpus Analysis Tools for Malay. 5th International Conference on Language Resource and Evaluation (LREC), Valetta, Malta, 22-28 May 2006, pp: 2212-2215.
18. McKeown, K.R. and D.R. Radev, 1995. Generating Summaries of Multiple News Articles. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 9-13 July 1995, DOI: 10.1145/215206.215334.
19. Hassel, M. and N. Mazdak, 2004. FarsiSum – A Persian Text Summarizer, Master Thesis. Department of Linguistic, Stockholm University.
20. Edmundson, H.P., 1969. New Method in Automatic Extracting. Journal of the Association for Computing Machinery, 16(2): 264-285.
21. Mani, I. and M.T. Maybury, 1999. Advances in Automatic Text Summarization, MIT Press.
22. Shigeru, M. and Y. Kazuhide, 2002. Some Research Topics and Future Prospects in Text Summarization. Statistical Method, Paraphrasing and More. J. Joho Shori, 43(12): 1310-1316.
23. Barzilay, R. and L. Lee, 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Los Angeles, California, USA, 27 May – 1 June 2003, pp: 16-23.
24. Lin, C.Y. and E. Hovy, 2002. Manual and Automatic Evaluation of Summaries. ACL-02 Workshop on Automatic Summarization, Philadelphia, Pennsylvania, USA, 11 -12 July 2002, pp: 45-51.
25. Hovy, E., 2005. Text Summarization, The Oxford Handbook of Computational Linguistic. Oxford University Press.
26. Hassel, M., 2003. Exploitation of Named Entities in Automatic Text Summarization for Swedish, Nordic Conference on Computational Linguistic, Reykjavik, Iceland, 30-31 May 2003, DOI: 10.1.1.105.7144.
27. Hand, T.F., 1997. A Proposal for Task-based Evaluation of Text Summarization Systems. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association of Computer Linguistics (ACL/EACL-97), Madrid, Spain, 11 July 1997, DOI: 10.1.1.11.3653.
28. Rais, N.H., M.T. Abdullah and R.A. Kadir, 2010. Query Translation Architecture for Malay-English Cross-Language Information Retrieval System, In Proceedings of the International Symposium on Information Technology, Kuala Lumpur, Malaysia, 15-17 June 2010, DOI: 10.1109/ITSIM.2010.5561589.
29. Zuraidah M.D., 2010. Processing Natural Malay Texts: A Data-Driven Approach. J. the Humanities and Social Sciences, DOI: 0.3176/tr.2010.1.06.
30. Shamsfard, M., T. Akhavan and J.M. Erfani, 2009. Persian Document Summarization by PARSUMIST, World Applied Science Journal, 7 (Special issue of Computer and IT): 199-205.