International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:5, 2009

Text Summarization for Oil and Gas News Article

L. H. Chong, and Y. Y. Chen

***Abstract*—**Information is increasing in volumes; companies are overloaded with information that they may lose track in getting the intended information. It is a time consuming task to scan through each of the lengthy document. A shorter version of the document which contains only the gist information is more favourable for most information seekers. Therefore, in this paper, we implement a text summarization system to produce a summary that contains gist information of oil and gas news articles. The summarization is intended to provide important information for oil and gas companies to monitor their competitor's behaviour in enhancing them in formulating business strategies. The system integrated statistical approach with three underlying concepts: keyword occurrences, title of the news article and location of the sentence. The generated summaries were compared with human generated summaries from an oil and gas company. Precision and recall ratio are used to evaluate the accuracy of the generated summary. Based on the experimental results, the system is able to produce an effective summary with the average recall value of 83% at the compression rate of 25%.

***Keywords*—**Information retrieval, text summarization, statistical approach.

I. Introduction

MONITORING competitor's behavior is an important task in oil and gas companies to sustain the company's competitive advantage. Casual knowledge about the competitors is usually insufficient. Therefore, companies are always hunting for more valuable information which for them to formulate business strategies. One of the main sources of information is the news articles. The news articles contain all the information regarding the competitor's communication promotions, strategy plan as well as industry news. By tracking this information on daily basis, companies are able to monitor the behaviour and movement of the competitors.

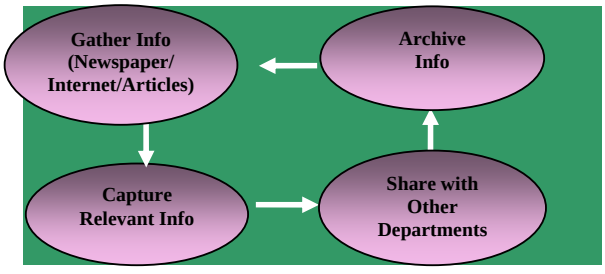

Fig. 1 Information Tracking Process

Fig. 1 shows a general information tracking process in an oil and gas company. The information gathering and sharing process are done manually. Human capability is limited in dealing with large textual database. In addition, it is time consuming in search and process important information from newspaper and Internet. The need for a text summarization system in a company is apparent. The objective of this paper is to develop a text summarization system which can produce a coherent summary by incorporating statistical approach with the three underlying concepts namely keyword occurrences, the title of the news article and the sentence location.

II. Related Work

The summary of an article can be an extract or abstract. An extract is produced by including some of the sentences of the original article, while an abstract is to provide the gist of the original article by rewriting it in a shorter length of text [1]. A text summarization system will generate summary for the users to obtain the gist of an article or document quickly. Most of the text summarization systems developed were mainly to produce an extract of the original article.

Neto, J.L *et al.* [2] text summarization system summarized a text based on the TFISF weight assigned to each word. TFISF is one of the statistical approaches that was widely used. This summarizing algorithm is different from TFIDF (term frequency inverted document frequency). TFIDF calculates the weight for each word in the entire documents. However, in TFISF (term frequency inverse sentences frequency), it extracts sentences having the high values that assigned to it. Evaluation has been conducted to measure the quality of the TFISF and the result proved that it is satisfactory.

Counting the word frequencies in an article is one of the techniques used to identify keywords in an article. Luhn, H.P. [3] claims that words that appear frequently appear in a document indicate the topic discussed. Words that appear the most in a text are most likely the keyword of the text excluding the stop words.

Despite the word frequencies, the position of a sentence is to be taken into consideration. Edmunson H.P. [4] stated that the important sentences always occur in specific positions: introduction and conclusion meanwhile the title of the text indicates its content.

Several other approaches have been used in text summarization system, for instance, neural network [5, 6], support vector machine [7], statistical approach [8] and Bayesian theorem [9]. However, each of these approaches has their own strength and weaknesses. In this paper, we proposed

to incorporate statistical approach with keyword occurrences, the paragraph position and the title of the document to summarize the text. This can be a promising approach which can produce a coherent summary that will be beneficial to oil and gas companies.

III. System Architecture

There are four main parts involved in the summarizing process which are preprocessing of words, weight assignment, sentence(s) extraction and, final filtering and assembling. The system architecture is illustrated as below.

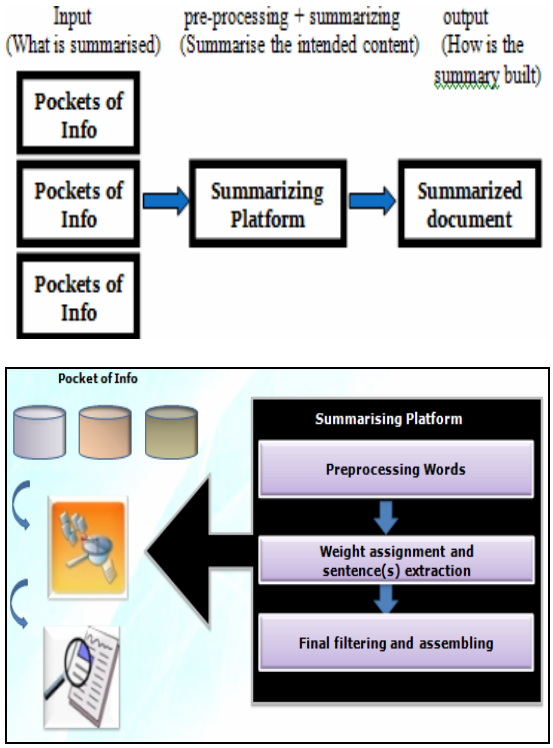

Fig. 2 System Architecture

A. Preprocessing

Preprocessing words in a news article involves two steps i.e. removing stop words and stemming. It is important to note that the proposed approach will take into consideration on the title of the article. Therefore, the title of the article will also need to go through the preprocessing phase. The words in a news article are stored in an array which is used to compare with a list of stop words that has been predefined and stored in a text file. If the word in the array matched the stop word in the text file, the particular word will be removed from the array. Following this, the array of text will split into individual words to go through the stemming process. This step is the process to derive the words to their root form. (E.g. going, gone >> go). This process is necessary to eliminate the suffixes and prefixes by applying Porter algorithm.

B. Weight Assignment and Sentence(s) Selection

Every sentence in an article is separated by "." (full stop) which is defined as a delimiter. Each word will be assigned with a basic weight (*wi*) using the following formula:

$$w_i = TFISF \tag{1}$$

Where *TF* is the frequency of a particular word appears in the article and *ISF* are represented by the following formula:

$$ISF(w) = log(S/SF(w)) \tag{2}$$

where w = word in a sample news article
S = number of sentences in a news article
$SF(w)$ = number of sentences that w occurred

After assigning weight to each word, w_i, all word's weight in a sentence are added up to represent the basic weight of the sentence, $s(w)$.

$$s(w) = \sum_{i=1}^{n} w_i \tag{3}$$

where $s(w)$ = sum of word weight in a sentence
n = number of words in the sentence
wi = the word weight of the ith word

$s(k)$, is calculated by comparing each word with a list of keywords that has been identified by the staff of the oil and gas company. Keywords that appeared in a sentence were given additional weight by using the following formula:

$$Sk = 2j \tag{4}$$

where Sk = total weight for keywords that appeared in a sentence
j = number of keywords exist in a sentence

The same weight assignment concept is applied to the title weighting method. Words that appear as the title of the text will be given additional weight by using the following formula:

$$St = 3z \tag{5}$$

where St = total weight for title words that appeared in a sentence
z = number of words that are same as title

As for the position of the paragraph, the first and the last sentences of each paragraph are assigned with extra weight.

$$Sl = 5 \tag{6}$$

where Sl = weight of the first/last sentence

Finally, the total weight of the sentence is calculated based on the following formula:

$$\text{Total weight} = Sk + St + Sl + s(w) \tag{7}$$

C. Final Filtering and Assembling

Sentence weight will represent its importance to the article. Each sentence will be ranked according to the weight of the sentence. In order to allow the system to produce a summary with good flow of information, each sentence was tagged with a reference number which is the order of appearance of this sentence in the original article. The higher the weight of the sentence, the more important is the particular sentence to appear in the summary. A summary should have a clearly arranged structure and written in a logical manner. Therefore, sentences that were selected to be included in the summary will need to be displayed based on the order of the reference number.

IV. Results and Discussion

The quality of the summary needs to be assessed in order to ensure the proposed approach is effective in summarizing news articles. Hence, an intrinsic method [10] has been used to evaluate the system generated summary by comparing it with human generated summary. The human generated summaries were obtained from the oil and gas company for evaluation purpose. The news articles were summarized at 4 different compression rates: 25%, 50%, 75% and 100%. This is to evaluate at which rate that the system is able to produce effective summary. The purpose of summary is to enable readers to capture the gist of the original article in a short time. Therefore, an effective summary is able to provide the gist of the article being summarized with small number of irrelevant information included.

The performance measures used to evaluate the quality of the summary is precision and recall ratio [11] which are shown accordingly in (8) and (9). Precision measures the ratio of correctness for the sentences in the summary whereby recall is used to count the ratio of relevant sentences included in summary. For precision, the higher the values, the better the system is in excluding irrelevant sentences. On the other hand, the higher the recall values the more effective the system would be in retrieving the relevant sentences. It is 1.0 when all relevant sentences are retrieved by the system.

$$Precision = \frac{|\{Relevant\ sentences\} \cap \{Retrieved\ sentences\}|}{|\{Retrieved\ sentences\}|} \quad (8)$$

$$Recall = \frac{|\{Relevant\ sentences\} \cap \{Retrieved\ sentences\}|}{|\{Relevant\ sentences\}|} \quad (9)$$

where *relevant sentences* = sentences that are identified in the human generated summary

retrieved sentences = sentences that are retrieved by the system

TABLE I
Evaluation on Summary of Oil & Gas News Articles

Compression Rate (%)	Precision	Recall
25	1.75	0.83
50	2.00	0.71
75	1.67	0.89
100	1.56	1.00

Based on Table I, average precision shows that the system can summarise the articles without including much of irrelevant sentences at the compression rate of less than 50%. While referring at the average recall, it shows that the recall value is increasing as the percentage of sentences length increases. This is because the more sentences included in the summary, the more relevant or important sentences are included. Therefore, summary that produced based on the compression rate of 75% and 100% will not be considered as an effective summary. However, we can observed that at the compression rate of 25%, the average recall value is higher as compare with the average recall value of 50% which is 0.71. This implies that at the compression rate of 25%, the system is able to retrieved most of the relevant sentences identified in human generated summary. In short, we can conclude that the system is able to produce a coherent summary at a shorter length of a text which is at the compression rate of 25%.

V. Conclusion

In this paper, we have developed a text summarization system which is able to produce a summary that has the similar result as the human generated summary. This is due to the topic and keyword specification from the oil and gas company. It may produce a less accurate result if the system is implemented in other fields without changing the specification. Thus, this would be the limitation of the system and there are still room for improvement. For future work, the system can be improved by integrating the system with intelligent process which enables it to learn such as incorporate the Support Vector Machine (SVM). With the intelligent process, the system is able to learn and produce better summaries in which the patterns of summarising are expected to be the same as human generated summaries. In addition, application of Bayesian theorem in keyword identification may overcome the limitation of this system.

References

[1] Hovey, E.E., Cross-lingual Information Extraction and Automated Text Summarization. Available from http://www.ics.mq.edu.au/~swan/summarization/glossary.htm [Accessed 12th November 2008].
[2] Neto, J.L., Freitas, A.A. and Kaestner, C.A.A. (2002), "Automatic Text Summarization Using a Machine Learning Approach" in Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advance in Artificial Intelligence, London, 2002.
[3] Luhn, H. P. (1999), "The Automatic Creation of Literature Abstracts", Advances in Automatic Text Summarization, MIT Press.
[4] Edmundson, H.P., "New Methods in Automatic Extracting", Journal of the ACM (JACM), Vol. 16, Issue. 2, p.264-285.

[5] S.P.Yong, Ahmad I.Z. Abidin and Y.Y. Chen (2005), "A Neural Based Text Summarization System", in Proceedings of the 6th International Conference of DATA MINING, Greece, 2005.
[6] Taeho, J., Marley, L. and Thomas M. G. (2006), "Keyword Extraction from Documents Using a Neural Network Model", in Proceedings of International Conference on Hybrid Information Technology, 2006.
[7] Joachims, T. (1998), "Text Categorization with SupportVector Machines: Learning with Many Relevant Features", in European Conference on Machine Learning (ECML), 1998.
[8] Y.Y. Chen, O.M. Foong, S.P. Yong, and Kurniawan Iwan (2008), "Text Summarization for Oil and Gas Drilling Topic", in Proceedings of World Academy of Scienc, Engineering and Technology, Singapore, 2008.
[9] Yamauchi, Y. and Mukaidono, M. (2000), "Probabilistic inference and Bayesian Theorem Rough Sets", Rough Sets and Current Trends in Computing, Springerlink.
[10] Mani, I, Klein, G., House, D. Hirschman, L., Firmin, T. and Sundheim, B. (2002), "SUMMAC: A Text Summarization Evaluation", Natural Language Processing, vol. 8, p. 43-68.
[11] Makhoul, J., Kubala, F., Schwartz, R. and Weischedel, R. (1999), "Performance Measures for Information Extraction", in Proceedings of DARPA Broadcast News Workshop, 1999.