

Text Summarization for Oil and Gas Drilling Topic

Y. Y. Chen, O. M. Foong, S. P. Yong, and Kurniawan Iwan

Abstract—Information sharing and gathering are important in the rapid advancement era of technology. The existence of WWW has caused rapid growth of information explosion. Readers are overloaded with too many lengthy text documents in which they are more interested in shorter versions. Oil and gas industry could not escape from this predicament. In this paper, we develop an Automated Text Summarization System known as AutoTextSumm to extract the salient points of oil and gas drilling articles by incorporating *statistical approach, keywords identification, synonym words and sentence's position*. In this study, we have conducted interviews with Petroleum Engineering experts and English Language experts to identify the list of most commonly used keywords in the oil and gas drilling domain. The system performance of AutoTextSumm is evaluated using the formulae of precision, recall and F-score. Based on the experimental results, AutoTextSumm has produced satisfactory performance with F-score of 0.81.

Keywords—Keyword's probability, synonym sets.

I. INTRODUCTION

COMPANIES realized that the data they are producing are ever increasing in numbers. The data do not always correspond to information since they need to be processed to generate knowledge or information [5]. As a result, information readers are overwhelmed with information. Information can be presented in a less congestive way. This can be done by providing readers with the gist of the document. Therefore, the need for a text summarization system is apparent.

Research on text summarization involves various methods to employ text categorization such as neural networks [1], regression models [4] and decision trees [2]. However, these methods have their own drawback which contributes to its poor development of classifiers due to performance variation using different types of data collection [3]. Therefore, numerous researches have been done to further enhance these methods in order to improve the performance of text categorization.

The objective of the work is to develop a running prototype that incorporates statistical approach, keywords identification, synonym words and sentence's position. The proposed system should produce a summary of any plain text document in English Language and within oil and gas drilling domain.

Authors are with Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia (e-mail: chenyokeyie@petronas.com.my, foongoimean@petronas.com.my, yongsuetpeng@petronas.com.my, i_onesutp@yahoo.com).

II. RELATED WORK

Automatic text summarization system would generate the summary of a given text document automatically. The summary generated by the system is dependent on the approach and end-objective of summarization of documents. For example, it could be indicative of what a certain topic is about, or can be informative about specific niceties of the same. It can differ in being a "generalized summary" of a document as opposed to "query-specific summary" [11]. It may be a set of sentences carefully chosen from the document or can be created by synthesizing new sentences on behalf of the information in the papers.

Assigning weights on the words in a document based on the frequency of its occurrence has become the key component of statistical analysis in text summarization [5]. This approach is less complex as compare to develop summaries through abstraction. Therefore, most of the researchers employ this method in their research on text summarization. For example, Neto, J.L *et al.* [8] research in text summarization algorithm is based on computing the value of *tf-isf* (Term Frequency – Inverse Sentence Frequency) measure for each word in the document. The use of *tf-isf* in developing a text summarization system is not new. Some of these can be found in [6] and [7]. The system has been evaluated with the real-world documents and the result is satisfactory.

Other promising approach include statistical analysis of term clustering, statistically based analysis of text structure, or discourse analysis and training algorithms that use human-generated summaries to determine probabilities that certain sentences from the source text should be included in the summary. [5]

The use of Bayesian model in text summarization system is popular due to its simplicity [9]. Julian Kupiec and his partners [5] employed an analysis technique which enables the learning progress of the application by using Bayesian statistics. However, their research found that based on the Bayesian algorithm alone does not provide satisfactory results. Other features that can improve the performance of the system are location, cue phrase and sentence length.

The statistical approach has its disadvantage in summarizing text. Those that have been identified were: the need for human intervention, ambiguous references, misapplied rhetoric, interpreting non-text objects and synonyms and other context-dependent terms [5]. Despite these problems, McCargar suggested that statistical approach is still an important strategy in developing text summarization system [5]. Recent research on text summarization has overcome some of the problems associated with statistical approach by combining other approaches. For example, Kraaij *et al.* [10] explored the use of

Naïve Bayesian with unigram model to perform multi document summarization. S.P. Yong *et al.* [6] worked on developing an automatic text summarization system combining both statistical approach and neural network to summarize documents. With this, incorporating statistical approach with keywords identification, and sentence's position can be a promising approach to develop a text summarization system that use in oil and gas industry.

III. SYSTEM ARCHITECTURE

The system architecture of *AutoTextSumm* can be divided into five main parts: preprocessing, word weight calculation, sentence weight calculation, sentence selection and final filtering as shown in Fig. 1.

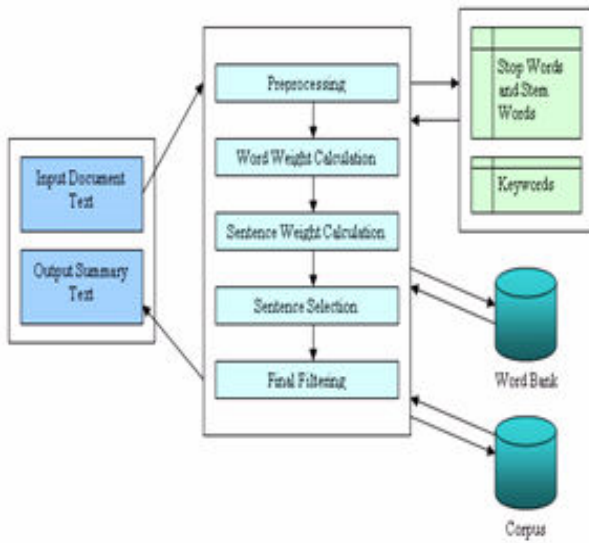


Fig. 1 System Architecture of *AutoTextSumm*

Words in the word bank are distinctive. Synonym of the words was considered as a single word could be represented by different words. For example, *offshore* could be replaced by *marine-based* or *sea-based*. These kinds of words should be taken in the same way since they all represent a single same meaning. The corpus (article database) focuses specifically on oil and gas drilling topic. The corpus was used in determining the weight for keyword. The system would process the corpus to find out how likely a keyword appears in the article. The keywords in the keyword list are provided by the experts from Petroleum Engineering field.

A. Preprocessing

There are three main activities performed in this stage: Tokenization, Stem Word Process and Stop Word Process. Tokenization is the process of separating the input article into individual words. The distinct words retrieved from the input article will be processed by removing prefixes and suffixes of each word. Any repetitive word found after the stemming process will be removed to avoid having huge amount of repetitive word in word bank. Stop words are the words which appear frequently in an article but contribute less meaning in identifying the important content of the article. Each stop

word found in the article will be given a smaller weight in order to rate the word as less important words in contributing the meaning to generate the summary.

B. Word Weight Calculation

After tokenizing the input article, the weight of each word in the article will be assigned according to the following formula [5]:

$$W(w) = tf * itf \quad (1)$$

where tf is the frequency of a specific word appears in the article, and itf is represented by the following formula [8]:

$$itf = \log\left(\frac{N}{n}\right) \quad (2)$$

where N = total number of articles in the corpus (articles database).

n = number of articles in which the word exists in the corpus.

C. Sentence Weight Calculation

The weight of the sentence in the article is affected by 3 factors, namely sentence's word weight, sentence's position weight and sentence's keyword weight. Therefore, the sentence total weight (St) for each of the sentences in the article is calculated based on the following formula:

$$St = Sw + Sp + Sk \quad (3)$$

where Sw = sentence's word weight

Sp = sentence's position weight

Sk = sentence's keyword weight.

Sw is calculated by adding up the weights of the words that form the sentence divided by the number of words in that sentence. Therefore, Sw can be expressed by the following formula:

$$Sw = \frac{\sum_{i=1}^n W_i}{n} \quad (4)$$

where W_i = the word weight of the i th word of the sentence

n = number of words in the sentence

The location or position of the sentence in the article was also taken into consideration in calculating the sentence weight. Sentences in the first two sentences of a paragraph are deemed to be important. The methods was also used in treating the last first or two sentences of the article, because they most likely to bring conclusion of what the article is about. Thus, if the sentence appears as the first two sentences or the last sentence of the paragraph, then the weight of the sentence should be higher and is calculated as:

$$Sp = Sw * 0.5 \quad (5. i)$$

Otherwise the sentence which appears in other parts of the paragraph should be given lower weight and is calculated by using the following formula:

$$Sp = Sw * 0.2 \quad (5. ii)$$

Sk is the sentence's weight based on keyword probability. The formula to calculate Sk is expressed by the following formula:

$$Sk = \prod_{i=1}^n P_i(k) \quad (6)$$

where $P_i(k)$ = probability of the i^{th} keyword in the sentence.
 n = number of words in the sentence

D. Sentence Selection and Final Filtering

After calculating the total weight for each of the sentences in the article, each sentence is ranked according to its sentence total weight. It leads us to the list of the sentences with their entrance reference (appear orderly in the article) and their weights. In short, the sentences are arranged in descending order according to its sentence weight. The higher the weight, the more relevant is the sentence to the content of the article. When displaying the summary, sentences which are selected to be included in the summary has to be ordered based on the entrance reference. In addition, the compression rate of the summary need to be defined by the user before the system is able to display the summary. The compression rate C is calculated using the formula below:

$$C = \frac{n_s}{N_f} \quad (7)$$

where n_s = number of sentences in generated summary
 N_f = number of sentences in the original full Text

The final filtering of the sentences would remove sentence which begins with quotes. Unimportant or trivial sentences are discarded based on the compression rate specified by the user.

IV. RESULTS AND DISCUSSION

In order to evaluate the effectiveness of the system, we have used an intrinsic method which aims to evaluate the quality of the summaries as compared to summaries produced by the system. We have used 5 different articles with the topic on oil and gas drilling in the field of petroleum engineering to evaluate the performance of the system. The limitation of the scope was aimed for the system to focus on an in-depth knowledge base. The reference summary for each article is obtained from the human experts to compare with the

summary produced by the system. The experts involved in generating and evaluating the summary would be from the area of Petroleum Engineering as well as English Language lecturer.

In calculating the overall performance of the *AutoTextSumm*, the summaries which generated based on different compression rate (from 10% to 90%) will be used for evaluation. In addition, the following information should be considered:

1. The reference summaries' selected sentences.
2. The sentences selected by *AutoTextSumm*.
3. The overlap between the *AutoTextSumm*'s summary and the reference summaries.

The performance measures used for the evaluation of the summary generated by the application are precision, recall, and F-score as shown in formula (8), formula (9) and formula (10) respectively. Precision measures the percentage of correctness for the total number of summaries judged by the summary assessor to be relevant. Precision also measures the usefulness of the summarizer while recall is a measure of the completeness of the summarizer.

Recall is a measure of how effective the system in including relevant sentences in the summary. It is 1.0 when all relevant sentences are retrieved. Precision is a measure of how effective the system in excluding irrelevant sentences from the summary. It is 1.0 when all documents returned to the system's users are relevant to the summary. Meanwhile, F-Score is a composite score that combines the precision and recall measures.

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

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

$$F - Score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (10)$$

To obtain the results of all performance measures, a reference output should be at hand. This section of evaluation uses a human-generated summary. The individuals involved in this process are the experts in Petroleum Engineering and the experts in English Language. The summary generated by experts would be used as a reference in obtaining the number of relevant sentences in a particular summary. Since there are 9 different summaries that will be generated for each article that put into testing, the summary which is most similar to the summaries generated by the experts will be used for evaluation.

TABLE I
EVALUATION ON OIL AND GAS DRILLING ARTICLES

Article No.	Precision	Recall	F-Score
1	0.47	0.89	0.61
2	0.50	1.00	0.67
3	0.88	0.84	0.86
4	1.00	0.92	0.96
5	0.94	0.93	0.93
Average			0.81

Based on the results shown in Table I, the average F-score for all articles is 0.81. This shows that identify important sentences for a summary from documents in a specific topic by using machine learning algorithm shows a similarity with the summaries generated by the expert (human-generated summaries). Therefore, the conclusion which have arisen from the results, suggest that this technique is suitable for a specific topic corpus.

Some reasons behind the acquisition of the findings (results) exist. First, the system being developed by the project used most of the articles with average length of 15-20 sentences. Difference in articles' length affects the analysis due to the difference on the overall articles' structures. Therefore, the project considered the location or position factor.

V. CONCLUSION

Automatic text summarization system's demand is increasing in nowadays high-technology environment. The advanced technology has caused more inventions found and more information shared. Therefore, information overloading has to be faced by users who are more interested in shorter version of lengthy documents. There exist some available text summarizers in the market; *Microsoft Word Auto Summarizer*, *NetSumm*, *Pertinence* and *Extractor*. However, rooms for further improvement need to be addressed in order to produce better summaries, which are similar to the human-generated summaries. The evaluation on the summarizer's effectiveness is still a huge area of research.

The reason why *AutoTextSumm* produced summaries nearer to the ideal standard of human-generated summary could be due to the topic specification. The developed system focuses on oil and gas drilling topic with the keywords and corpus as its knowledge base in predicting the likelihood of a sentence to be included in the summary. The summary generated by the expert is also done by considering the main theme of the article and then applies the experts' knowledge in generating the summary. For future work on text summarization system, other machine learning techniques such as the Support Vector Machine (SVM) and decision tree algorithm should be considered to improve the performance of the text summarization system.

REFERENCES

- [1] E. Qwiener, J.O. Pederson, and A.S.Weigned, "A neural network approach to topic spotting", in Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.
- [2] Joachims, T., "Text Categorization with SupportVector Machins: Learning with Many Relevant Features", in European Conference on Machine Learning (ECML), 1998.
- [3] Tsuruoka, Y., Kawaguchi-shi, Tsujii, J., "Journal of Biomedical Informatics archive", Vol.37(6), pp. 461-470, 2004.
- [4] Y.Yang and C.G.Chute, "An example-based mapping method for text categorization and retrieval", *ACM Transaction on Information Systems (TOIS)*, 12(3):252-277, 1994.
- [5] Victoria, M., "Statistical Approaches to Automatic Text Summarization", *Bulletin of the American Society for Information Science and Technology*, Vol3(4), April/May 2004.
- [6] S.P. Yong, Ahmad I.Z. Abidin and Y.Y. Chen, "A Neural Based Text Summarization System", in Proceedings of the 6th International Conference of DATA MINING, 2005.
- [7] Pardo, T.A.S., Rino, L.H.M. and Nunes, M.G.V., "GistSumm: A Summarization Tool Based on a New Extractive Method" in *Computational Processing of the Portuguese Language*. Vol. 2721/2003
- [8] Neto, J.L., Freitas, A.A. and Kaestner, C.A.A., "Automatic Text Summarization Using a Machine Learning Approach" in Proceedings of the 16th Brazilian Symposium on Artificial Intelligence: Advances in Artificial Intelligence, London, 2002.
- [9] Kim, S.B., Han, K.S., Rim, H.C. and Myaeng, S.H., "Some Effective Techniques for Naïve Bayes Text Classification" in *IEEE Transactions on Knowledge and Data Engineering*, 2006.
- [10] Kraaij, W., Spitters, M. and Heijden, M., "Combining a Mixture Language Model and Naïve Bayes for Multi-document Summarisation" <http://www-connex.lip6.fr/~amini/RelatedWorks/Kraaij01.pdf> [Accessed on 23th June 2008].
- [11] Albanese, M., "Extracting and Summarizing Information from Large Data Repositories" http://www.fedoa.unina.it/577/01/Tesi_MASSIMILIANO_ALBANESE.pdf [Accessed on 23th June 2008].