ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Automatic Vector Space Based Document Summarization Using Bigrams

Rajeena Mol M. [1], Sabeeha K. P. [2]

P.G. Student, Department of Computer Science and Engineering, M.E.A Engineering College, Kerala, India[1]

Assistant Professor, Department of Computer Science and Engineering, M.E.A Engineering College, Kerala, India[2]

ABSTRACT: Automatic summarization is the process of reducing a text document by a computer program in order to create a summary which covers the most important information from the original document. Document summarization is an emerging technique for understanding the main purpose of any kind of the document. This paper proposes a vector space based document summarization method using bigrams. The method considers n-gram language model with vector space model of the document. The n-gram language model generates all possible sentences from a given bag of words in the input document. The method generates a summary based on bigram weight. Vector space method is used to calculate the bigram weight. In this proposed method, bigram weight is helps to rank the sentences in the input document. So we can propose a new vector space based document summary using bigrams. The main aim of this project is to understand the key points of the document and also helps to get information within less time and reduces the use of memory.

KEYWORDS: Document Summarization, N-gram Language Model, Weight Calculation, Probability Matrix.

I. INTRODUCTION

Document summarization is the most popular application in the Natural Language Processing. There is huge amount of data available in structured and unstructured form and it is difficult to read all data or information. The aim of this project is to get information within less time. Hence we need a system that automatically retrieves and summarize the documents as per user need in limited time. Document summarizer is one of the feasible solutions to this problem. Summarizer is a tool which serves as a useful and efficient way of getting information from large documents. It is a process to extract the important content from the documents.

Document summarization is the task of producing a concise and fluent summary to deliver the major information from an input document. Document summaries can be used for users to quickly browse and understanding the document. The document summaries can be helpful in information retrieval systems. In this study, single extractive summarization methods are focused. Extractive document summarization systems usually rank the sentences in a document according to some ranking strategy and then select a few highly ranked sentences into the summary.

In general, a summarization system can be divided into three modules. First module is analysis phase or pre-processing. Then transform of the document. The final module is synthesis phase. Analysis or pre-processing includes stemming, stop word elimination, parsing etc. Transformation means construct or transform the pre-processed data into some simple representation like graph, vector etc. Synthesis phase consist of score the sentences, ordering the sentences and select the sentences for creating summary.

Document summarization creates information reports that are both concise and comprehensive. The goal of a brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, comprehensive document summary should itself contain the required information.

Figure.1. shows the general module design for document summarization systems. The proposed system works based on this flow. First phase include sentence splitting, stop word elimination, stemming and tagging. Second phase consist of generation of bigrams and weight calculation. Synthesis phase consist of ranking and sentence selection.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

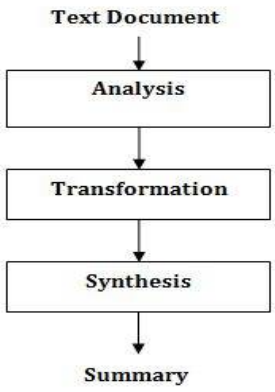

Fig. 1. Module design for Summarization

Paper is organized as follows. Section II describes related work, which reviews certain papers related to and referred for the project. Section III discuss about the proposed system. The flow diagram represents the step of the algorithm. Section IV presents experimental results showing the performance graph. Finally, Section V presents conclusion.

II. RELATED WORK

Automatic document summarization is a hot and important research area related with computer science and linguistics. There are different types of summarization approaches depending on what the summarization method focuses on to make the summary of the text. Different document summarization methods have been developed in recent years. Generally, those methods can be either extractive or abstractive ones. An extractive summarization method consists of selecting important sentences or phrases from the source document and concatenating this into shorter form. The importance of a sentence is based on statistical and linguistic features of the sentences. An abstractive summarization method consists of understanding the original text and retelling it in fewer words. Simply, express the idea in the source document using our words. The abstractive summarization usually needs information fusion [1], sentence compression [2] or reformulation [3].

One of the main concepts in the summarization process is the redundancy removal, so it is an important subtask. Some methods select the several top ranked sentences and reduce redundancy during summary generation using a popular measure called maximal marginal relevance (MMR) [4]

Another system is clustering based methods [5] are also used to ensure good coverage and avoid redundancy in summary. The clustering based approach divides the similar sentences into multiple groups to identify themes of common information and selects sentences one by one from the groups to create a summary [6]. Here the cluster quality heavily here depends on the sentence similarity measure used.

The graph based approach [7] [8] to text summarization represents the sentences in a document as a graph where a sentence is represented as a node of the graph and an edge between a pair of sentences is determined based on how much they are similar to each other. For measuring the importance of a sentence, a graph based method utilizes global information of the sentences in the graph, rather than depending only on local sentence specific information. The graph based methods also mainly uses the standard cosine similarity measure for building the similarity graph. Many existing extractive summarization systems [9] mentioned above use sentence similarity for either reducing redundancy or constructing a graph or both. The N-gram technique [10] is used to reorder a sentence if it was written incorrectly.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

III. PROPOSED SYSTEM

Figure 2 shows the proposed system architecture. First step is pre-processing. Pre-processing consist of sentence splitting, stop word elimination and stemming and also apply POS tagging method. POS tagging represent what the role of a word playing in the sentence. It helps to generate a valid bigram list for providing rank to the sentence at the latter stage. Next step is the generation of bigrams. Bigrams is a sequence of two adjacent elements from the sentence. Then generate a bigram probability matrix. The matrix shows the probability of occurrence of two words together. Then calculate the weight of bigrams based on vector method. Then rank the sentences based on bigram weight and probability. The final step is the sentence selection, that is highly ranked sentences are selected to the summary.

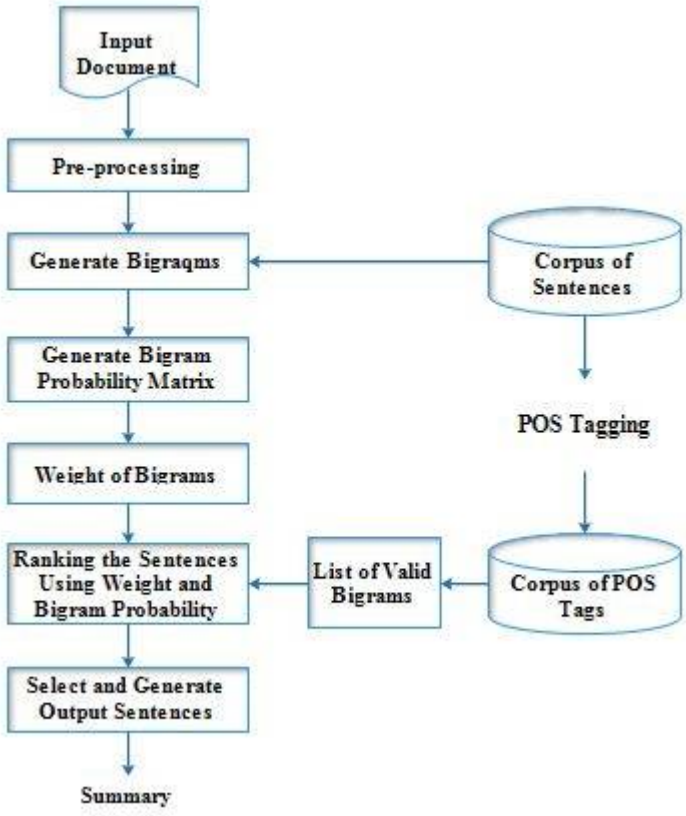

Fig. 2. System Architecture of vector space based document summarization method using bigrams

Figure 2 explains the overall working of the proposed method. The weight calculation of bigram is depending on the frequency of words. The frequency of the word helps to find the most important bigram. The bigram probability matrix shows the probability of occurrence of two words together. At this stage, a bigram was extracted from the bigram probability matrix that was trained on the textual corpus by N-gram language model.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

Words	<s>	W_1	W_2	.....	W_i	</s>
<s>	0	$P_{s,1}$	$P_{s,2}$	.....	$P_{s,i}$	$P_{s,e}$
W_1	0	0	$P_{1,2}$	.....	$P_{1,i}$	$P_{1,e}$
W_2	0	$P_{2,1}$	0	.....	$P_{2,i}$	$P_{2,e}$
.....	.....	.....	.....	.....	.....	.....
W_i	0	$P_{i,1}$	$P_{i,2}$	.....	0	$P_{i,e}$
</s>	0	0	0	0	0	0

Table 1. Bigram Probability Matrix

In the Table-1 W_i shows the i-th input words and the $P_{i,j}$ shows the probability of occurrence of next word(j-th) on a previous word(i-th).

IV. EXPERIMENTAL RESULTS

The experiments are performed on DUC 2004 dataset. The DUC dataset is helps to better evaluation. Figures show the evaluation results of the summarization methods. The proposed method for document summarization system has many advantages over the existing system. This method computes the weight of bigrams. It helps to improve the accuracy of the system. The system also helps to extract the information within limited time. So the extraction of information using this method is more efficient and easy.

The comparison chart is very helpful to understand the accuracy and time complexity in the two summarization systems. The proposed method is highly efficient compared to the existing methods and it also helps to get information within less time and reduces the use of memory.

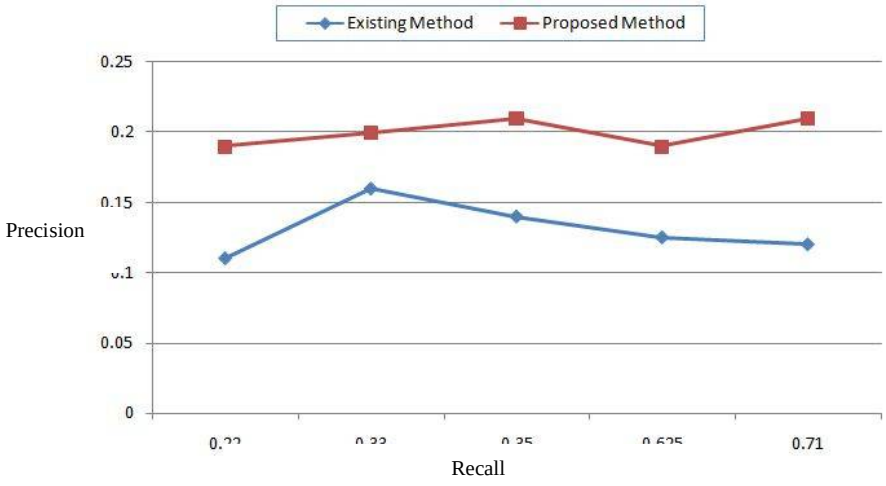

Fig. 3 Analysis based on precision and recall

Figure 3 is the analysis graph based on precision and recall. The proposed method achieves good precision. So the proposed method provides an accurate summary comparatively existing method. The bigram weight is helps to ensure an accurate summary

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 7, July 2016

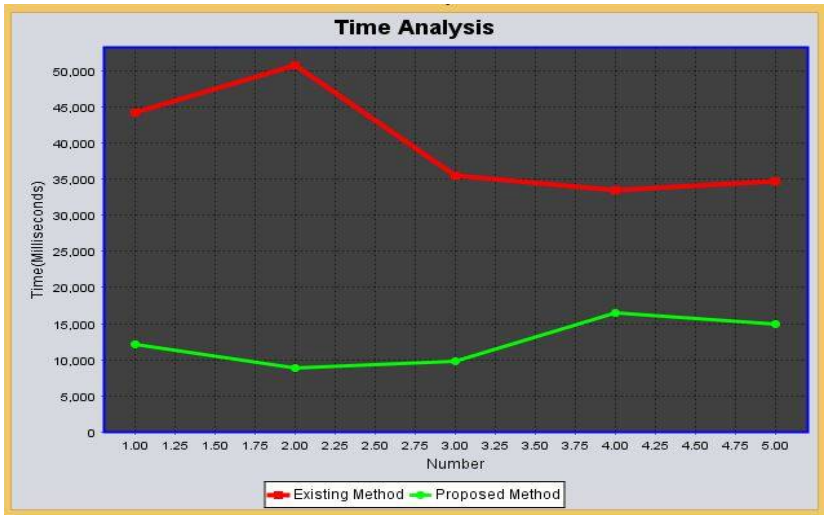

Fig. 4 Analysis based on time

Figure 4 is the analysis graph based on time. The graph shows proposed method generates a summary within limited time. So the proposed method ensures an accurate summary within limited time.

V. CONCLUSION

The proposed work has presented a vector space based document summarization method using bigrams. This method builds the system based on n-gram language model and vector space model of the document. It helps to summarize the document in limited time. So the proposed method ensures to get an accurate summary within less time and also reduce the use of memory. In this work, high frequency of correct sentences is observed. So the proposed model discussed here is able to provide better result by using the bigram method.

REFERENCES

[1] R. Barzilay, K. R.McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in Proc. ACL'99, 1999.
[2] K. Knight and D.Marcu., "Summarization beyond sentence extraction:A probabilistic approach to sentence compression," Artif. Intell., vol.139, no. 1, pp. 91–107, 2002.
[3] K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E.Eskin, "Towards multidocument summarization by reformulation:Progress and prospects," in Proc. AAAI, 1999.
[4] J. G. Carbonell, and J. Goldstein, "The use of MMR, diversity-based re-ranking for reordering documents and producing summaries," In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 335–336, 1998.
[5] G.Erkan and D.R.Radev, "Multi-document summarization using sentence clustering," 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, vol. 3, pp. 653-658, 2012.
[6] E. Boros, P. B. Kantor, and D. J. Neu, "A Clustering Based Approach Creating Multi-Document Summaries," In Proceedings of the 24th ACM SIGIR Conference, LA, 2001.
[7] G. Erkan, and D. R. Radev, "LexRank: graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, pp. 457-479, 2004
[8] S. Yan and X. Wan, "Srrank: Leveraging semantic roles for extractive multi-document summarization," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, December 2014.
[9] A. Biyabangard, "Word concept extraction using hosvd for automatic text summarization," proceedings of National Seminar on summarizationTechnology, vol. 6,no. 3, May 2015.
[10] Athanaselis Theologos, Mamouras Konstantinos, Bakamidis Stelios and Dologlou Ioannis, "A Corpus Based Technique for Repairing formed Sentences with Word Order Errors Using Co-Occurrences of n-Grams". International Journal on Artificial Intelligence Tools, pp. 401-424,2011.
[11] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network ". In Proceedings of HLT-NAACL , pp. 252-259,2003.
[12] Brill, E. "A Simple Rule-Based Part of Speech Tagger". In Proceedings of the Third Conference on Applied Natural Language Processing, 1992