AN AUTOMATIC DOCUMENT SUMMARIZATION SYSTEM USING A FUSION METHOD

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

[1]M Tech Student, Dept. of Computer science& Engineering, M.E.A Engineering College, Perinthalmanna, Kerala
[2]Assistant Professor, Dept. of Computer Science& Engineering, M.E.A Engineering College, Perinthalmanna, Kerala

---***---

Abstract - *Automatic summarization is the process of reducing a document by a computer program in order to create summary which covers the most important information from the original document. So document summarization is an emerging technique for understanding the main purpose of any kind of document. Here we propose a fusion approach for document summarization. This approach describes about Salton's vector method with N-gram language model. Salton's vector method producing a short, paragraph-length summary. Then a headline summary is created, which consists of a set of indicative words or phrases mentioned in the input. The headline summary helps to understand the main objective of the document using N-gram language model. The purpose of this project is to quick understand the main aim of the document and also helps to understand the overall idea about the document.*

***Key Words*: Document Summarization, Salton's vector method, N-gram Language Model, Tagging, Headline summary**

1. INTRODUCTION

Document Summarization is the most popular application in the Natural language processing. Document summarization is the process of generating a summary by reducing the size of input document and retaining important information of input document. There is arising a need to provide high quality summary in less time because at present, the growth of data is increasing tremendously on World Wide Web or on user's desktops so document summarization is the best tool for making summary in less time.

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 survey 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. Summarizer is a process to extract the important content from the documents. In general, the summaries are defined in two ways. They are single-document summarization and multi-document summarization. The summary which is extracted and created from single document is called as Single-document summarization whereas multi-document summarization is an automatic process for the extraction and creation of information from multiple text documents.

The main aim of summarization is to create summary which provides minimum redundancy, maximum relevancy and co-referent object of same topic of summary. In simple words, summary should cover all the major aspects of original document without irrelevancy while maintaining association between the sentences of summary. So, extractive summarization and abstractive summarization approaches are used. Extractive summarization works by selecting existing words, phrases or number of sentences from the original text to form summary. It picks the most relevant sentences or keywords from the documents while it also maintains the low redundancy in the summary. Abstractive summarization method which generates a summary that is closer to what a human might create. Basically this type of summary might contain words not explicitly present in the original document format. It provides abstraction of original document form in fewer words.

2. RELATED WORK

Different document summarization methods have been developed in recent years. Generally, those methods can be either extractive or abstractive ones. Extractive summarization create the summary from phrases or sentences in the input document, and the abstractive summary express the idea in the input document using different words. The abstractive summarization usually needs information fusion [1], sentence compression [2] or reformulation [3]. In this study, we focus on extractive summarization methods. This section discusses some of those existing summarization systems.

One of the main concept 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 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 [5]. Here the cluster quality heavily here depends on the sentence similarity measure used.

The graph based approach [6] 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 mentioned above use sentence similarity for either reducing redundancy or constructing a graph or both.

Another method as proposed by Theologos et al. [7] used N-gram technique to reorder a sentence if it was written incorrectly. However, we have tailored Theologos et al. model as per our requirement to generate all possible correct sentences. The filtering technique and ranking method used in our model is completely different from Theologos et al.

HOSVD [8] is the word-document-time tensor is used to extract concepts of words. For this propose, Firstly, Higher Order Singular Value Decomposition methods are applied to extract the main word's concept. These methods are able to extract important concepts from three-dimensional tensors. Then the sentences are arranged based on cosine similarity with the main concept of documents.

SRRank [9] is one of the important method for summarization. In this method the sentences in a document set are parsed with a SRL tool, and the resultant SRL 3-tuples are further decomposed into 2-tuples and nouns. Then we build a heterogeneous graph including these SRL 2-tuples and nouns together with sentences. Based on the heterogeneous graph, a heterogeneous ranking algorithm is applied to simultaneously rank sentences, SRL 2-tuples and nouns. Finally, each sentence is assigned with a saliency score, and we apply a greedy algorithm to remove redundancy and produce the summary.

Sentence clustering approach [10] is used to generate document summary. In this approach, single document summaries are combined using sentence clustering method to generate multi-document summary. Here each document is first pre-processed and then features are extracted based on which summary is created. The sentences appearing in the individual summaries are clustered. Sentences from each cluster are extracted to create a multi-document summary. The extracted sentences are arranged according to their position in the original document. For clustering, semantic and syntactic similarity between sentences is used. The semantic similarity between words is combined to get semantic similarity between sentences.

3. PROPOSED SYSTEM

In this section, we propose a simple and efficient document summarization system that generates a short summary and also generates a headline summary from input document. Our system can be divided into following three modules.

- Analysis
- Transformation
- Synthesis

Analysis or pre-processing includes stemming, stop word elimination, parsing, tagging 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.

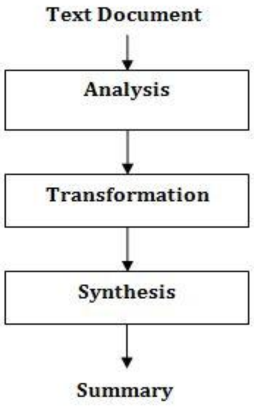

Fig-1: Module design for summarization

3.1 Salton's vector method

Salton's vector method provides a concise summary for a document. This method consist of mainly pre-processing, transformation and synthesis.

Steps:

- Initially the summarizer separate the input document into sentences based on the separators.
- The next step is stop word removal, that is the unnecessary words are removed from the document.
- Next method is called stemming. The document after removing the stop words is revised again for the unique words. Unique words are the one which have the same meaning or might be redundant in the document. These are removed by a method called stemming.
- Then calculate the word count. By using the Stemming mechanism the occurrence of a word is computed and the results are displayed in the format of how many times they occur and the number of sentences they have occurred.
- Next step find the sentence weight. To find the weight of the each and every word that occurs in a sentence. The total weight of a sentence is the weight for each and every word.
- The final step is ranking. In this step the highest weight sentence is ranked the first position followed by the next consecutive sentences in the document.

3.2 N-gram Language model

N-gram language model provide a headline summary of the document. The headline summary is used to identify the main concept of the document. In this method we consider the input document as a salton's paragraph summary. The paragraph summary is first consider as a bag of words and generate bigram probability matrix for input bag of words. Bigram Probability matrix shows the probability of occurrence of two words together and also construct lattice of bigrams. It is a representation of Bigram Probability Matrix with directed edge and without any cycle. Then apply DFS search to generate only candidate sequences. Then tag in each generated candidate sequence and count the number of valid sequences. Then the highest score sentence will be taken in to account.

Tagging is the most important part in the headline generation. Part of speech tags represent what the role a word is playing in the sentence. To capture the syntax of language or to learn which word is preceding which another word we have used trigram model and have extracted the trigrams from the second annotated corpus of POS tags to generate a list of valid trigrams. We have applied Stanford Part-of- Speech Tagger [11] to the first text corpus. In the process we have obtained the annotated corpus of POS tags.

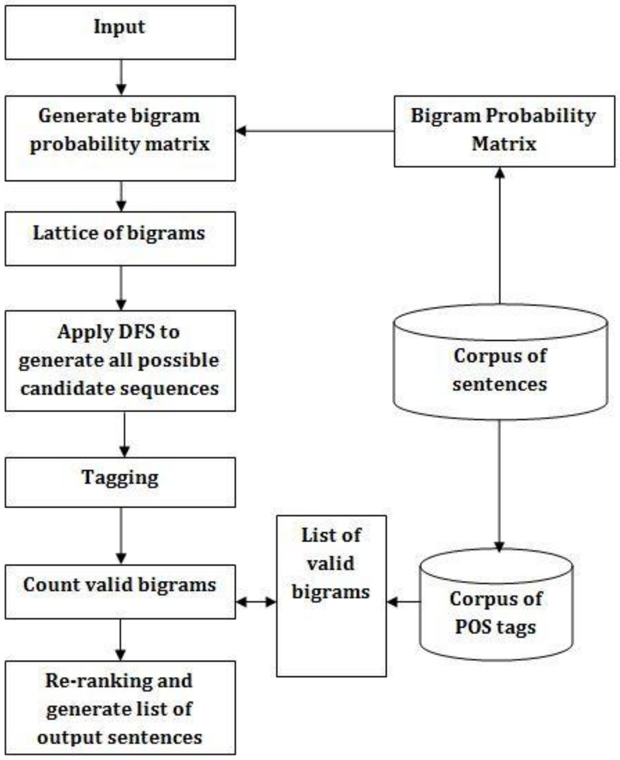
Input
Generate bigram probability matrix
Bigram Probability Matrix
Lattice of bigrams
Apply DFS to generate all possible candidate sequences
Corpus of sentences
Tagging
List of valid bigrams
Count valid bigrams
Corpus of POS tags
Re-ranking and generate list of output sentences

Fig -2: Block diagram of proposed work

4. EXPERIMENTS

4.1 Dataset:

Already available datasets will be used for the implementation of this project. To evaluate document summarization algorithms, the DUC data sets from Document Understanding Conference (DUC) will be used. Each dataset contains document clusters. Each cluster contains documents relevant to a query or topic description.

4.2 Data Analysis

For most of the cases the fluency of the summary can be easily identified with the help of linguistics used in newspaper, magazines etc. So, it is advised to collect data from these resources. The accuracy can be evaluated using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) toolkit (version 1.5.5) to evaluate the proposed methods, which is widely applied by DUC Conference for performance evaluation. It measures the quality of a summary by counting the unit overlaps between the candidate summary and a set of human written reference summaries. ROUGE can generate three types of scores: recall, precision and F-measure. F-measure is a balance of recall and precision results.

5. CONCLUSIONS

In this paper, we propose a fusion method for generating a fluent summary. The fusion method builds based on Salton's vector method with N-gram language model. The Salton's vector method gives a short summary about the document. Then applied N-gram language model to that summary and produce a single sentence called headline summary. In this case the user can easily understand the aim of large document and also provide a short description about the input document. That is the N-gram language model produces a headline summary for quick access of the document.

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] 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.

[6] 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

[7] 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.

[8] A. Biyabangard, "Word concept extraction using hosvd for automatic text summarization," proceedings of National Seminar on summarizationTechnology, vol. 6,no. 3, May 2015.

[9] 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.

[10] 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.

[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.