IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Extractive Technique for Text Summarization based on Ranking Scheme

A.A.Shrivastava[1*], A.S.Bagora[2], R.Damdoo[3]

[1]Dept. of CSE, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India
[2]Dept. of CSE, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India
[3]Dept. of CSE, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

Corresponding Author:shrivastavaa@rknec.edu, Tel.: +91-9834341678

Available online at: www.ijcseonline.org

Received: 20/Mar/2018, Revised: 28/Mar/2018, Accepted: 19/Apr/2018, Published: 30/Apr/2018

Abstract- Text Summarization is the process of creating a condensed form of text document which maintains significant information and general meaning of source text. Automatic text summarization becomes an important way of finding relevant information, precisely in large text, in a short span of time. In this paper, the proposed method uses sentence ranking of a topic-specific document to generate automatic summary. The method is based on the concept of extractive summary, in which the summary of a document is obtained by scoring, ranking and selecting the highest ranked sentences of the document. Initially, the text is pre-processed by tagging the document and selecting adjectives, nouns and verbs, and then the text is analysed and sentences most similar to all is ranked and selected for generation of summary. Experiments on these methods were conducted to compare the results on sentence ranking. The algorithm proposed was tested on different documents and has given accuracy of about 80% when compared to summarization tools available online.

Keywords: *Text Summarization, data mining, extraction-based summarization, sentence ranking.*

I. INTRODUCTION

As the amount of information available is increasing rapidly day by day in different formats such as text, video, images, etc. it becomes difficult for an individual to find relevant information related to a topic. To find appropriate information, a user needs to search through all the documents available, which leads to information overload and wastage of time and efforts.To deal with this problem, automatic text summarization plays a vital role. Automatic summarization condenses a source document into meaningful content without altering information, helping user to grab the gist of the document within a short time span (which reflects the gist of the document). If the user gets an effective summary, it helps to understand document at a glance without checking it entirely.

Summarization is mainly of two types: Abstractive summary and Extractive summary. Abstractive summarization generates a generalized summary by constructing new sentences which are short & concise, and the summary may contain new phrases and sentences that are not available in the source text. For generating abstractive summary, language generation and compression techniques are necessary[5].Extractive summarization method selects informative sentences from the document as they exactly appear in the source document, based on particular criteria and features, to form a summary. The main challenge before extractive summarization is to decide which sentences from the input document are significant and to be included in the summary. The extractive summarization process uses the meaning of the sentences for scoring and is usually doneby the traditional concept of lexical chaining. Fig1depicts a common method used for text extraction[4].

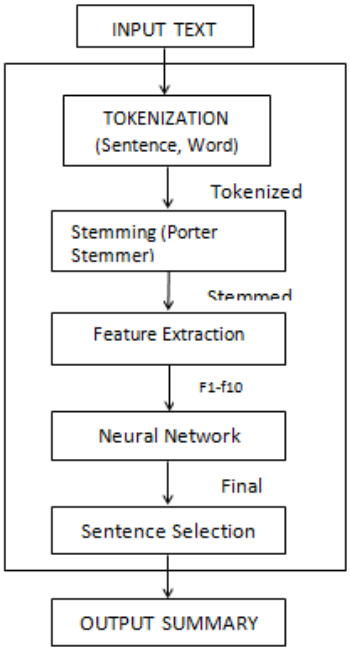

Fig 1: Approach for text extraction

In this paper, an extractive summarization methodology based on sentence ranking is proposed for generating automatic summary of a topic-specific document. Section I contains the introduction of summarization and its types, Section II contains the related work on extractive summarization techniques and ranking methods, Section III explains the methodology of the proposed algorithm with flow chart. The proposed method is broken down into the following steps: extraction of nouns, words and adjectives and stemming of the words, scoring and ranking of sentences by using similarity matrix, selection of the highest ranked sentence and generation of the summary. The results obtained from analyzing the proposed method on different documents and comparing it against standard available tools are discussed in the results section. Section IV describes results obtained when the algorithm was compared with available extractive summarization tools and discuses reasons for the inaccuracies found. Finally Section V concludes research work with future directions.

II. RELATED WORKS

A general procedure for extractive methods that are usually performed in three steps is discussed below [2]:

Step 1: Some pre-processing such as tokenization, stop word removal, noise removal, stemming, sentence splitting, frequency computation etc. is applied.

Step 2: In this step, sentence scoring is performed. There are two major solutions to measure the importance of sentences in sentence ranking.[3] One is supervised-based, which uses annotated text to train a model to predict the weight of each sentence. The features can be TF-IDF value or sentence position etc. Another is unsupervised-based, which does not need any training text. Unsupervised uses heuristic information like sentence position, cue phrases etc.

Step 3: In this step, sentences are selected based on the score for summary. Sentence ranking[3] is an important research issue in text analysis. Summarization using sentence ranking majorly includes extracting keywords from each sentence finding similarity in sentences, ranking sentences on the basis of the similarity and then picking up most significant sentences from the document.

1. EXTRACTIVE SUMMARIZATION METHODS

Extractive summarizers find out the most relevant sentences in the document. These also remove the redundant data. Extractive summarization is easier than abstractive summarization to bring out the summary. The common methods for extractive are TF/IDF method, cluster based method, graph based approach[6], machine learning approach, LSA (Latent Semantic Analysis) method[10], text summarization with neural networks[7], automatic text summarization based on fuzzy logic[9], query based extractive text summarization[8], concept-obtained text summarization, text summarization using regression for estimating feature weights, multilingual extractive text summarization, Hidden Markov SVM (HMSVM) and rhetorical structure of summarization[11], topic-driven summarization MMR (Maximal Marginal Relevance)[10] and centroid-based summarization, etc. [1]

2. RANKING METHODS

An iterative graph based ranking algorithm is designed on heterogeneous graph to rank sentences. J. Tian, M.Cao [3], designed three variants of the ranking algorithms which uses different combination of available structural information for ranking sentences. The results indicate that sentence information is more effective than paragraph information [3] and section information, and these can help improve the performance. These also compared the proposed algorithms with three other models: a TF-IDF based model, a graph-based model which uses sentences as vertices, and a graph-based model which uses word as vertices. All the results indicate that the structural information can improve the precision of sentence ranking.

III. METHODOLOGY

The proposed method first splits the document into sentences and then assigns indexes to each sentence. A file is created that contain nouns and a stemmed verb of every sentence. From the tagged sentences, nouns and verbs are picked and the verbs are stemmed. Each word of a sentence is tagged using POS Tagger. The algorithm then creates an intermediate representation of the original sentences that contains a list of extracted nouns and verbs from the original sentence. This intermediate representation of a sentence is called as a processed sentence. The similarity of the processed sentences Si with every processed sentence Sj is calculated and a matrix of similarity values is created which is then used for further ranking. After ranking, we get top-k processed sentences and use the indexes of the processed sentences to map it back to the original sentence in the document and generate the summary.

A. Assumptions:

This method works best on the documents with a strong central idea and theme, and not on documents dealing with abstract topic, as the proposed method focuses on ranking by comparing the common words in the sentences and not on the contextual meaning of the sentences.

B. Extraction of nouns, words and adjectives and stemming of the words

Before ranking the sentences, extraction of important words in summary, i.e. verbs, nouns and adjective, is done to later rank sentences based on these words. For extracting important words from the document, POS Tagger is used. A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads some text in a language and assigns parts of speech to each word, such as noun, verb, adjective, etc. For the proposed method, Stanford POS Tagger and english-left3words-distim tagger model is used as it is faster and is recommended for general use. The tagger takes a sentence as input and returns tagged sentence. We pick words tagged as verbs, nouns and adjectives as these are the main parts of the speech. After extracting verbs and nouns from the main text, we stem the extracted words, i.e. process of removing prefixes and suffixes from the words. This is done to later compare these root words and rank the sentences. As the algorithm uses POS tagger to extract the keywords or the words which gives sense to the sentence, it automatically removes all the stop

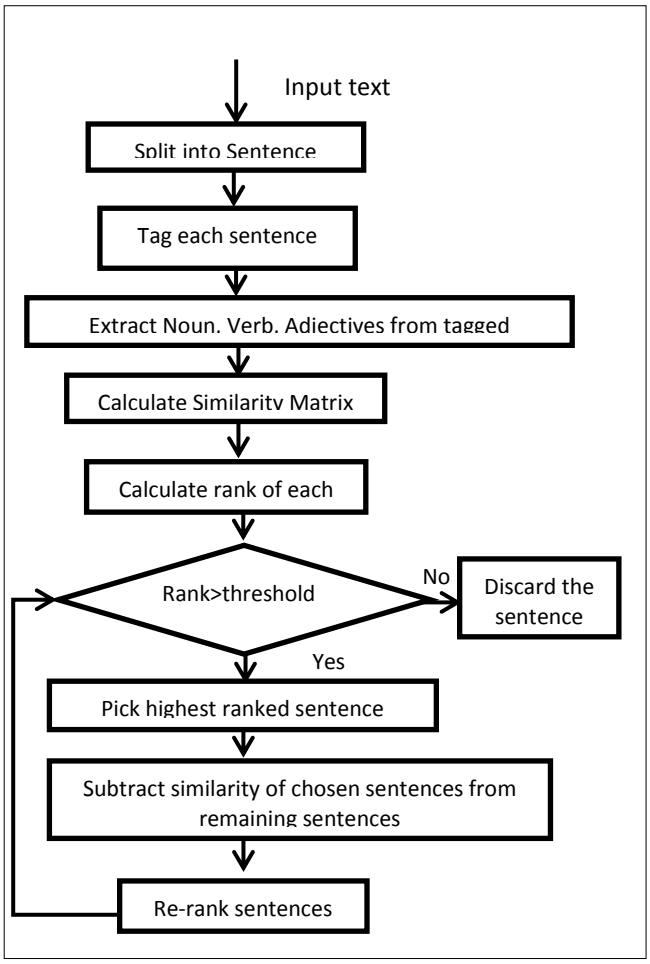

Fig 2: Flowchart for the proposed algorithm

words. So, words which have most weightage are selected rather than removing words that do not contribute to the sentence. Stop-words are necessary for constructing a syntactically and semantically correct sentence, but don't provide additional meaning to the sentences and hence, removed from the sentences before further processing.

C. Scoring of sentences and ranking of sentences

After extracting the words out of the sentences, sentences are ranked. Ranking of the sentences is done by assigning a 'score' to each sentence and the higher the score, higher the rank. The sentences are scored by comparing the 'similarity' between all the sentences. This similarity is based on the no of same words between the sentences. The similarity is calculated by the formula:

S(i, j) = 2*(no of common words between i and j)/ (no of words in i + no of words in j).(1)

where i, j are the sentences between which the similarity is to be calculated.

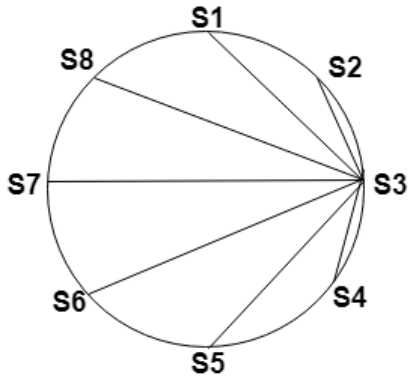

Fig3:Representation of association of one sentence-to-all sentences

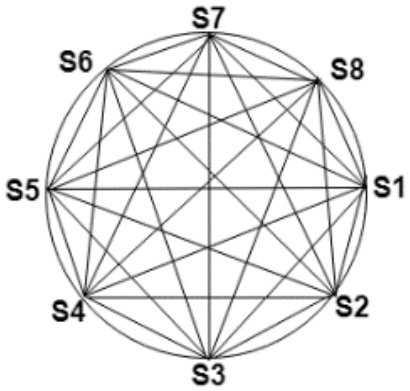

Fig4:Representation of association of all sentences-to-all sentences

Similarity of all the processed sentences is stored in a matrix, called 'similarity matrix', and each row element and column element represents one sentence. A sentence here is the list of extracted nouns and stemmed verbs in the original sentence. After calculating the similarity matrix, the ranking factor of each sentence is calculated and used to rank of each sentence. This is done by summing up the elements of the row, and each sum represents the ranking factor of that sentence. Sentences are rearranged in an ascending manner based on the ranking factor of the sentences and grouped to select the best possible sentence for the summary.

D. Selection of the highest ranked sentence

For grouping the sentences, ranking factor is used, with the assumption that closer the ranking factor values, more similar the sentences are in meaning. Rather than going for traditional methods of grouping, sentences are rearranged with some changes after one of the sentences is picked. So, when the highest ranked sentence is picked, we punish the rest of the sentences, i.e. subtract the similarity (between the selected sentence and all the remaining sentences) from the ranking factor and re-rank the sentences to pick up the highest ranked sentence. This method is based on the logic that since the most important sentence is already picked; it might cause redundancy if a sentence similar to already selected sentence is picked. Re-ranking and rearrangement of sentences is done to pick up the highest ranked sentence. The sentences that have fallen below the threshold are discarded.

E. Generation of summary

The sentences selected are added into the summary till we reach a minimum word limit or all the sentences have fallen below the threshold and no more sentences are left to consider.

IV. RESULTS& DISCUSSIONS

The proposed algorithm was tested on various documents and compared with the results of three online summarization tools i.e. Auto-summarizer, Summarizing.biz and Text Compactor. Table 1 shows percentage of matched sentences, between the proposed algorithm and the tools. It was found that proposed algorithm produced closest results with the tool summarizing.biz, with 87.6%.The algorithm has generated good results with each of the tools. All the major sentences were found to be picked by the proposed algorithm, as were picked by the online tools. Some sentences that were picked by our algorithm were missing from the summary generated by standard tools. Although, when it comes to topics which have a central theme, the algorithm was found to generate a very wholesome summary, but it generates weak results for documents which have very few similar sentences, or are not bound by a common theme. Further work could be done on identifying synonyms and merging them together in summary to generate a more accurate summary. This algorithm can be extended to find correlation between words and the crux of the document, or analyse and compare sentences as well as paragraphs, and to give proportionate scores to words according to its semantic correlation with the topic.

TABLE 1: COMPARISONOF PROPOSED ALGORITHM WITH OTHER AVAILABLE SUMMARIZATION TOOLS

	Comparison result of proposed algorithm (% matched sentences)		
	Auto – summarizer	**Summarizing.biz**	**Text Compactor**
Common sentences	75.55%	87.6%	81.1%
Common keywords	74.66	90.44%	80%

V. CONCLUSIONAND FUTURE SCOPE

In this paper, we explore the way to use the structural information to rank sentences based on graph ranking models. The structural information includes the word co-occurrence relation, the sentence-paragraph relation and the paragraph-section relation. By modeling those relationships into an iterative process over the relationship matrices, we can simultaneously obtain the rank of the words, sentences, paragraphs and sections. Proposed algorithm focuses on similarities between the sentences and how the sentences are ranked in the model. In order to have a better understanding of proposed model, we designed tasks to explore its characteristics and also compared proposed model with existing tools. Further work can be done on connecting words, recognizing semantic relationship network between words; as well as extend it to include outside sources such as Wikipedia and other literatures.

REFERENCES

[1] N. Andhale, L. Bewoor (Vishwakarma Institute of Information Technology, Pune), "*An Overview of Text Summarization Techniques*", International Journal of Scientific Research Engineering & Technology (IJSRET), Volume 6, Issue 3, 2017, pp. 146-150

[2] S. Akter et. al, "*An Extractive Text Summarization Technique for Bengali Document(s) using K-means Clustering*

Algorithm.", American Journal of Engineering Research (AJER) , Volume-6, Issue-1, pp-226-239

[3] J. Tian, M. Cao, J. Liu, X. Sun, Z. Hai, "*Ranking Sentences in Scientific Literatures*", In the proceedings of11th International Conference on Semantics, Knowledge and Grids, USA, pg . 275, 2015

[4] P. Gupta, R. Tiwari and N. Robert, "*Sentiment Analysis and Text Summarization of Online Reviews: A Survey*", In the proceedings of International Conference on Communication and Signal Processing, pg. 241-245, 2016, India.

[5] M Indu,, Kavitha K V, "*Review on text summarization evaluation methods*", In the proceedings ofInternational Conference on Research Advances in Integrated Navigation Systems, pg. 2016, India

[6] N.Moratanch, S.Chitrakala, "*A Survey on Extractive Text Summarization*", IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017) , India.

[7] K. Chen, S. Liu, B. Chen, H. Wang, E. Jan, W. Hsu, "*Extractive Broadcast News Summarization Leveraging Recurrent Neural Network Language Modeling Techniques*", IEEE / ACM Transactions On Audio, Speech, And Language Processing, Vol.23, No.8, pp. 1322-1334, 2015

[8] Y. Meena, P. Deolia, D. Gopalani, "*Optimal Features Set For Extractive Automatic Text Summarization*", Fifth International Conference on Advanced Computing & Communication Technologies, pp. 35, India, 2015

[9] Y. Zhang, M. Joo Er, M. Pratama, "*Extractive Document Summarization Based on Convolutional Neural Networks*", 42nd Annual Conference of the IEEE Industrial Electronics Society, USA, 2016

[10] J. Zhang, P. Fung, *"Learning Deep Rhetorical Structure For Extractive Speech Summarization*", IEEE International Conference on Acoustics, Speech and Signal Processing,, pp.5302-5305, USA, 2010

Author Profiels

Miss Ayushi Shrivastav is pursuing Bachelor of Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management, Nagpur affiliated by Rashtrasant Tukadoji Maharaj Nagpur University. Email: shrivastavaa@rknec.edu

Miss Aditi Bagora is pursuing Bachelor of Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management,Nagpur affiliated by Rashtrasant Tukadoji Maharaj Nagpur University.Email: bagoraas@rknec.edu

Prof. Rina Damdoo received the Masters in Technology in Computer Science and Engineering from Nagpur University in 2012. Currently she is serving as Assistant Professor at Shri Ramdeobaba College of Engineering and Management, Nagpur. She has a teaching experience of 15 years with expertise in subjects like Data Structures, Operating Systems, System Software, Design Patterns. Her research interest includes N-grams, Statistical Machine Learning and sentiment analysis in the domain of Template messaging. Email: damdoor@rknec.edu

