ISSN:0975-9646

Automatic Text Summarization with Cohesion Features

Nilesh R. Patil
[#]*SSBT College of Engineering & Techonolgy,*
Bambhori, Jalgaon-425001, Maharashtra, India

Girish Kumar Patnaik
SSBT College of Engineering & Techonolgy,
Bambhori, Jalgaon-425001, Maharashtra, India

***Abstract*— In recent days, the large amount of information getting increased on internet; it is difficult for the user to go through all the information available on web. Automatic summarization system is used to reduce the user's time in reading the whole information available on web. Text summarization system is to identify the most important information from the given text and provided it to the end users without changing source text idea. The automatic text summarization with statistics and linguistics feature uses sentence scoring method for selecting important sentence according to their level of importance. The total number of importance sentences are equal to total number of paragraphs that time importance sentences are added in summery so meaningful information is not extracted effectively, it have text overloading problem. The automatic text summarization with cohesion features is grammatical and lexical linking within the text or sentences that hold together a sentence and provide meaningful sentences to end user without changing source text idea therefore it increase the effectiveness of summery and also solve text overloading problem.**

***Keywords*— Summarization, Linguistic Features, Statistical Features, Cohesion Features.**

I. INTRODUCTION

Text summarisation has been investigated mutually of analysis field by the tongue process community for nearly the half century. Text summarisation is one amongst the typical tasks of text mining. Data Mining is extraction of knowledgeable information from large amount of data. It is classified into two types i.e. information retrieval and information extraction. Automatic text summarization is part of information extraction. The main role of a text summarization system is to identify the most important information from the given text and provided it to the end users without changing the source text idea.

In recent days, the large amount of information getting increased on internet; it is difficult for the user to go through all the information available on web. Automatic summarization system is used to reduce the users time in reading the whole information available on web.

The automatic summarization is that the computer automatically extracts arbitrary from the aboriginal article, and in the ideal case, the arbitrary can call the capital content of commodity accurately and comprehensively, and the accent of the arbitrary is particular and smooth. Automatic text summarization can be classified into two categories: extraction and absorption. Extraction arbitrary is a alternative of sentences or phrases from the aboriginal argument with the highest score and put it calm to a new beneath argument without changing the antecedent text. Absorption arbitrary adjustment uses linguistic methods to appraise and adapt the text. Most of the accepted automatic argument summarization arrangement use extraction adjustment to after math summary. Automatic text summarization uses altered appearance for free the weights of the sentences and the appearance are broadly classified into statistical, Linguistic and cohesion features [1][5].

A. ***Statistics Feature*** - In statistical feature, the weight is assign for sentence according to their level importance for selecting important sentence. In this case, sentences statistical appearance are represented with some absence after amount for anniversary appearance advertence the position of sentences aural anniversary document, their breadth (in terms of words, they contain), their affinity with account to the certificate appellation and some bifold appearance advertence if sentences accommodate some cue-terms or acronyms begin to be relevant for the summarization task. These characteristics are then accumulated and the aboriginal p% of sentences accepting highest scores is alternate as the certificate summary [1].

1. *Key Word Feature* - Keywords are usually nouns and determined application tf × idf measure. Sentences having keywords are of greater affairs to be included in summary. these exatraction can be done by Morphological Analysis, Noun Phrase (NP) Extraction and Scoring, Noun Phrase (NP) Clustering and Scoring.
2. *Sentence Position* - The weight of the book can aswell be access by position of the book i.e area the book is present in certificate paragraph. This will accord acceptable after-effects in account articles.
3. *Term Frequency tf(w)* - Term abundance is calculated using both the unigram and bigram frequency. We considered alone nouns while accretion the bigram frequencies. A arrangement of two nouns occurring together denotes a bigram. The unigram/bigram frequency denotes the amount of times the unigram/bigram occurred in the document. Typically the bigrams action less number of times than the unigrams, so we acclimated a factor that catechumen the bigram abundance to unigram frequency as a chat akin feature. All the bigrams in which the chat occurs are taken, and normalized to unigram scale. Finally the best of the unigram and

normalized bigram abundance is taken as the term abundance of the word.

4. *Length of the Word l(w)* - Smaller words action more frequently than the beyond words, In adjustment to negate this after effect we advised the chat breadth as a feature.
5. *Parts of speech tag p(w)* - We acclimated Porter tagger to find the POS tag of the word. We ranked the tags and assigned weights, based on the advice that they contribute to the sentence.

B. ***Lingustics Feature*** - In statistical feature, the weight is assign for sentence according to their level importance for selecting important sentence. In linguistic feature, Sentences containing such proper nouns and pronouns are having greater chances for including in summary; these chances are overcome through linguistic feature [1].

1. *Proper Noun feature* - Proper Noun may be name of person ,place etc. Sentences absolute such proper nouns are accepting greater affairs for including in summary.
   S= No. Proper nouns in S/(Sentence Length (S))
2. *Pronouns* - Pronouns such as "she, they, it" can be included in to arbitrary untill they are explained by their agnate nouns.

C. ***Cohesion Feature*** - In sentence to sentence cohesion feature, Cohesion is the grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning for increasing effectiveness of summery. Cohesion devices that creates coherence in text are as Reference, Substitution, Elipse, Lexical cohesion and Conjunction [3].

There are two types of cohesion.

1. *Grammatical Cohesion* – It referring structural content.
2. *Lexical Cohesion* - It referring language content of piece.

The organization of the report is given as - section 1 is introduction which consists information about domain, motivation. section 2 is a detailed related work. section 3 is praposed work of the automatic text summerization with statistics, linguistics and sentence to sentence cohesion features. Finally, the conclusion and future scope is discussed in last section.

II. RELATED WORK

Kumar et al., in[1], Sentences containing proper nouns and pronouns are having greater chances for including in summary, these chances are overcomes through statistical and linguistic approach. In Statistical and linguistic approach, the weight is assign for sentence according to their level importance for selecting important sentence. Advantage of both approach is work in any language. Disadvantage of approach is information diversity.

Freitas et al., in [2], The limitation of the approach is the problem of sentence ordering, as the system tries to find relevant sentences in groups of different topics. Sentence scoring method, sentence clustering and graph model are uses to overcome such a problem is to order the sentences using their relative position in the original documents and try to ''align'' the selected sentences. Advantage of this approach is dealing with information redundancy and diversity. Disadvantage of this approach is text overloading.

Blanka Frydrychova Klimova and Sarka Hubackova, in [3], The limitation of the approach is information overloading. Cohesion approach is the grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning to the end user without changing source text idea. The advantage of this approach is to overcome text overloading problem. The disadvantage of this approach is working in few languages.

Ferreira et al., in [4], A context based text summarization system is dealing with problem of Lack of fluency and coherence. The main contribution of this paper is finding the best combinations of sentence scoring methods for three kinds of documents: news, blogs, and articles. The advantage of this approach is improving the summarization results. The disadvantage of this approach is information overloading because summarization take place on context based document i.e. news, blogs and article.

Shah et al., in [5], Automatic text summarization of Wikipedia article is difficult to detect subtopic in document. There are two new approaches for summarizing the text. The first method is to adjust the frequency of the words based on the root form of the word, and also the frequency of its synonyms present in the text. The second method is to identify sentences containing citations or references and give them a higher weight. The advantage of this approach is effective sentence ranking in summery. The disadvantage of this approach is use of citations with higher weight to sentence so unimportant information is added in summery.

Jain et al., in [6], The limitation of the approach is dealing with problem of information redundancy, sentence ordering and fluency. Graph and Cluster Based approaches in Multi-document summarization and gives the idea to improve summary in less effort or even to construct new or hybrid procedure for next generation. The advantage of this approach is to generate smooth summaries as compared to ranking algorithm. The disadvantage of this approach is information loss during summarization.

Atefeh Ferdosipour, in [7], The effectiveness of sentence scoring method is depends upon length of document and the type of language used in document. Cohesion approach is grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning for increasing effectiveness of summery. Cohesion devices that creates coherence in text are as Reference, Substitution, Elipse, Lexical cohesion and Conjunction. Grammatical Cohesion is referring to structural content. Lexical Cohesion is referring to language content of piece. The advantage of this approach is to use of hybrid approach. The disadvantage of this approach is varies the document length.

Yogesh Kumar Meena and Dinesh Gopalani, in [8], The limitation of the approach is dealing with problem of no proper integration of document. Evolutionary algorithm approach gives a review of the growth and improvement in the techniques of Automatic Text Summarization on

implementing Evolutionary Algorithms techniques. The advantage of this approach is to use of genetic algorithm. The disadvantage of genetic algorithm is complex process.

Ferilli et al., in [9], The limitation of the approach is dealing with problem of stemming procedure. Sentence scoring approach uses stemming procedure. Stemming procedure is use to obtained the radix of each word. Radix is positive integer value which is use in sentence scoring method for selecting importance sentences. The advantage of stemming procedure works effectively. The disadvantage of stemming procedure is complex process for negative integer values.

III. PROPOSED SYSTEM

The proposed system solely focuses on extractive based summarization method. We will discuss two new approaches for summarizing the text. The first method is to adjust the frequency of the words based on the root form of the word, and also the frequency of its synonyms present in the text. The second method is to identify sentences containing citations or references and give them a higher weight. As shown in Fig. 4.10, the summarization system takes input as Wikipedia articles, processes it and gives the summary sentences. Input file consist of raw data to be processed by the system. The sentences are selected according to their level of importance which are added in summery according to their score of rank.

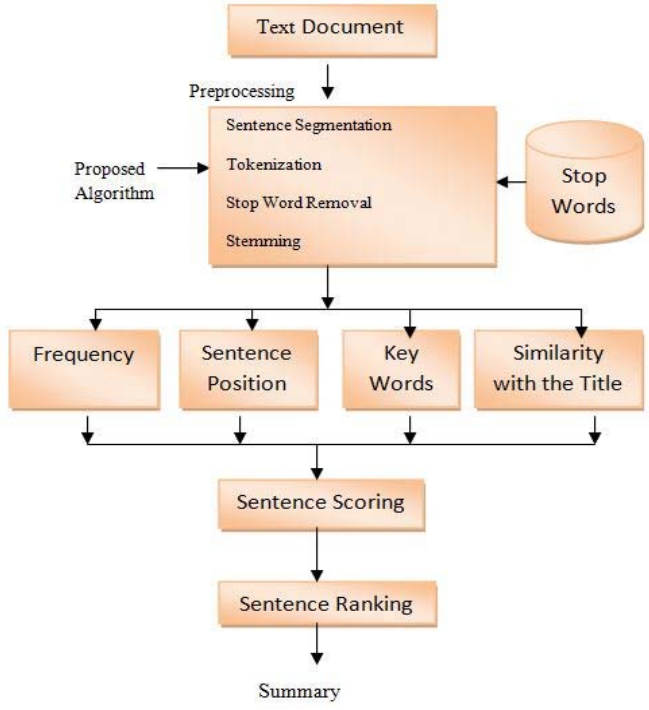

Figure1: text summarization architecture [1].

The Goal of extractive text summarization is selecting the most relevant sentences of the text. The Proposed method uses statistical and Linguistic approach to find most relevant sentence. Summarization system consists of 3 major steps, Preprocessing ,Extraction of feature terms and algorithm for ranking the sentence based on the optimized feature weights.

A. Preprocessing the document - This step involves Sentence segmentation, Sentence tokenization, Stop word Removal and Stemming.

a. *Sentence Segmentation* – It is the process of decomposing the given text document into its constituent sentences along with its word count. In English, sentence is segmented by identifying the boundary of sentence which ends with full stop (.) , question mark (?), exclamatory mark(!).

b. *Tokenisation* - It is the process of splitting the sentences into words by identifying the spaces, comma and special symbols between the words. So list of sentences and words are maintained for further processing.

c. *Stopword Removal* - Stop words are common words that carry less important meaning than keywords. This words should be eliminated otherwise sentence containing them can inuence summary generated.

d. *Stemming* - A word can be found in different forms in the same document. These words have to be converted to their original form for simplicity. The stemming algorithm is used to transform words to their canonical forms. In this work, Porters stemmer is used that splits a word into its root form using a predefined suffix list.

B. Feature Extraction – After an input document is tokenized and stemmed, it is split into a collection of sentences. The sentences are ranked based on four important features: Frequency, Sentence Position value, Cue words and Similarity with the Title.

a. Frequency – Frequency is the number of times a word occurs in a document. If a words frequency in a document is high, then it can be said that this word has a significant effect on the content of the document. The total frequency value of a sentence is calculated by sum up the frequency of every word in the document.

b. Sentence Position Value - Position of the sentence in the text, decides its importance. Sentences in the beginning defines the theme of the document whereas end sentences conclude or summarize the document. The positional value of a sentence is computed by assigning the highest score value to the first sentence and the last sentence of the document. Second higher score value to the second sentence from starting and second last sentence of the document. Remaining sentences are assigned score value zero.

c. Key words - Key words are the important words in the document. These key words are inputted from the user. If a sentence contains keywords then assign score value one to the sentence, otherwise score value is zero for the sentence.

d. Similarity with the Title- The similarity with the title consists of the words in titles and headers. These words are considered having some extra weights in sentence scoring for summarization. . If a sentence contains words in title and header then

assign score value one to that sentence, otherwise score value is zero for the sentence.

C. *Sentence Scoring* – The final score is a Linear Combination of frequency, Sentence positional value, Key Words and Similarity with the title of the document.

D. *Sentence Ranking* - After scoring of each sentence, sentences are arranged in descending order of their score value i.e. the sentence whose score value is highest is in top position and the sentence whose score value is lowest is in bottom position.

E. *Summery Extraction* - After ranking the sentences based on their total score the summary is produced selecting X number of top ranked sentences where the value of X is provided by the user. For the readers convenience, the selected sentences in the summary are reordered according to their original positions in the document.

F. Sample Experiment - The experiment on different text documents and extracted the output and then analysed by using manual summaries. And also the summary generated by available summarizers such as Microsoft and online summarizers.

G. Algorithm

a. *Algorithm 1*- It shows the proposed method-1.

1: Count the total number of paragraphs in the given document.
2: Pre-Process the given document, segmenting the given document into sentences and then segment the each sentence into words.
3: Carry out an analysis of stop words, and then apply stop word removal and stemming procedure.
4: Assign a Score for each sentence in a document based on features (both linguistic and statistic) as follows
4.1: Assigning a weight for each word based on its level of importance.
4.2: Calculate the total weight of each word of a sentence
Wg=tf*mathlog(scnt/df)
4.3: Calculate the total score of a sentence
total weight of sentence=Wg1+Wg2+...+Wgn
5: rearrange all sentences according their score or ranking.
6: Generate Summery.

Algorithm 2 - It shows the proposed method-2.

1: Count the total number of paragraphs in the given document.
2: Pre-Process the given document, separating the given document into sentences through full stop and then separating sentences into words through space.
3: Repeated words are stored in wordmap variable in the form of key set and key value ie.key set is repeated word and key value is how many time that word is repeated.
4: compared each words of all sentences.
4.1: Those sentences have repeated word that sentences are added in summery.
4.2: Those sentences doesn't have repeated word that sentences are skipped from summery.
5: Generate Summery.

H. *Flow chart*

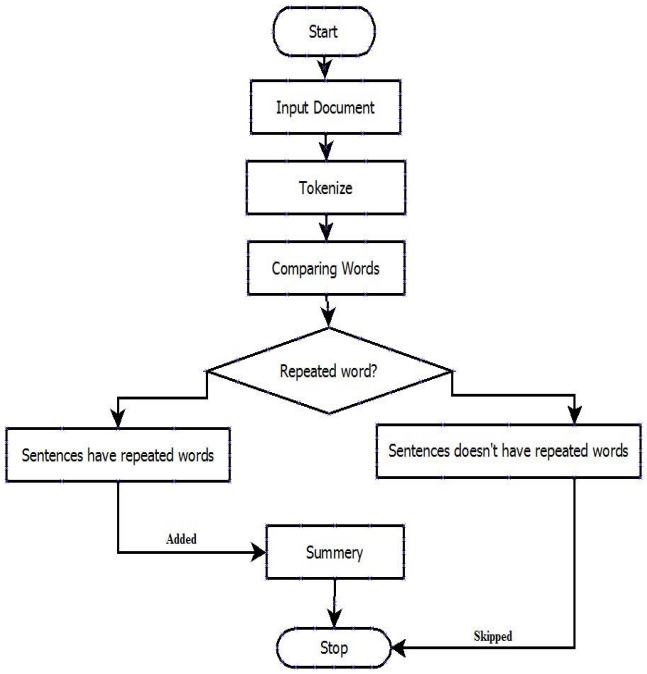

Figure2: Flow Chart of Proposed Method-2 .

IV. COMPARATIVE ANALYSIS AND RESULTS

Text overloading problem in summery exits in existing systems while proposed system does not have that problem. And irrelevant information in summary does not occur in proposed system while existing system have problem of irrelevant information. From the result point of view, the Sentence to sentence cohesion approach is more effective than the Statistics approach and linguistcs. Cohesion approach consist of Gramatical and Lexical Linking within the text and
sentence that hold the sentence which gives important information. Cohesion approach also solve the text overloading problem, therefore effectiveness of summery is increase through cohesion approach. The evaluated system with documents containing minimum 200 to 300 words. The summary generation is carried out as following steps. In the pre-processing stage the document mainly includes identification of sentence or boundary word, comparing word procedure. Identification of Sentences (or)Word Boundary is Pre-Process the given document, separating the given document into sentences through full stop and then separating sentences into words through space. Repeated words are stored in wordmap variable in the form of key set and key value ie. key set is repeated word and key value is how many time that word is repeated. Compared each words of all sentences. Those sentences have repeated word that sentences are added in summery. Those sentences doesn't have repeated word that sentences are skipped from summery. The final summery is generated shown in following figure2.

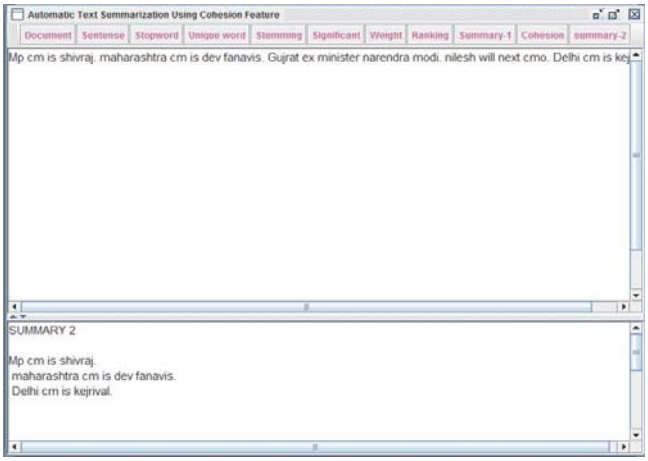

Figure2: Result Generated by Proposed Method-2.

TABLE 4.1
TABLE SHOWING COMPARISON OF VARIOUS APPROACHES OF TEXT SUMMARIZATION

Sr.No.	Text Summerization Features	Performance	Easy or Difficult
1	Statistics	Average	Easy
2	Lingustics	Good	Easy
3	Cohesion	Best	Difficult

TABLE 4.2
TABLE SHOWING COMPARISON BETWEEN EXISTING SYSTEM AND PROPOSED SYSTEM

Sr.No.	Problems	Existing System	Proposed System
1	Text Overloading in summery	Yes	No
2	Irrelevant Information in summery	Yes	No

Evaluating a summary result is a difficult task because there does not exist an ideal summary for a given document or set of documents. The absence of a standard human or automatic evaluation metric makes it very hard to compare different systems and establish a baseline. Besides this, manual evaluation is too expensive: as stated by Lin (2004), large scale manual evaluation of summaries as in the DUC conferences would require over 3000 hours of human efforts.

V. CONCLUSIONS

W Automatic summarization is technique which selects the important sentences based on statistical, linguistic and cohesion features. Cohesion feature is grammatical and lexical linking within sentences that hold the sentence and provide meaningful text to end user without changing source text idea. Cohesion Feature through generated summery compared with commercial online summarizer or Microsoft summarizer tool
therefore by adding cohesion feature, The text overloading problem is solved and the effectiveness of summery was increased.

In future, the effectiveness of summary can be increased by adding some advanced cohesive features such as synonymy, antonymy, collocation, enumeration, parallelism and transition feature.

REFERENCES

[1] D.V.N.Siva Kumar, E.Padma Lahari and S. Shiva Prasad, "Automatic text summarization with statistical and linguistic features using successive thresholds," IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), 2014.

[2] Frederico Freitas, Rafael Ferreira, Luciano de Souza Cabral and Rafael Dueire Lins, "A multi-document summarization system based on statistics and linguistic treatment," Expert Systems with Applications,2014.

[3] Blanka Frydrychova Klimova and Sarka Hubackova, "Grammatical cohesion in abstracts", IEEE Transactrions, 2013.

[4] Frederico Freitas, Rafael Ferreira and Luciano de Souza Cabral. "A context based text summarization System," 11th IAPR International Workshop on Document Analysis System, 2014.

[5] Deep Shah, Dharmendra Hingu and Sandeep S. Udmale, "Automatic text summarization of Wikipedia articles," 2015.

[6] Yogesh Kumar, Ashish Jain and Dinesh Gopalani, "Survey on graph and cluster based approaches in multi-document text summarization," International Journal of Computer Applications, 2014.

[7] Atefeh Ferdosipour, "The Effectiveness of cohesion of science text by means of the paragraph on attitude," 2nd GLOBAL CONFERENCE on PSYCHOLOGY RESEARCHES, 2014.

[8] Yogesh Kumar Meena and Dinesh Gopalani, "Evolutionary Algorithms for Extractive Automatic Text Summarization," International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014), 2015.

[9] Stefano Ferilli, Floriana Esposito, and Domenico Grieco, "Automatic Learning of Linguistic Resources for Stopword Removal and Stemming from Text," 10th Italian Research Conference on Digital Libraries (IRCDL), 2014.