Automatic Text Summarization With Statistical, Linguistic and Cohesion Features

Nilesh R. Patil
*SSBT College of Engineering & Technology, Bambhori, Jalgaon-425001, Maharashtra, India

Girish Kumar Patnaik
SSBT College of Engineering & Technology, 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 importance sentences are added in summary 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 summary 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 articulate 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 adverte 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 adverte 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 exatract can be done by Morphological Analysis, Noun Phrase (NP)

2. **Sentence Position** - The weight of the book can as well 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. Linguistics 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.

\[
S = \text{No. Proper nouns in S}/(\text{Sentence Length (S)})
\]

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 summary. Cohesion devices that creates coherence in text are as Reference, Substitution, Ellipse, 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 proposed work of the automatic text summarization 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 overcome 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 summary. The disadvantage of this approach is use of citations with higher weight to sentence so unimportant information is added in summary.

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 summary. Cohesion devices that creates coherence in text are as Reference, Substitution, Ellipse, 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

Shah et al., in [5], The proposed arrangement alone focuses on extractive based summarization method. We will altercate two new approaches for summarizing the text. The aboriginal adjustment is to acclimatize the abundance of the words based on the basis anatomy of the word, and as well the abundance of its synonyms present in the text. The additional adjustment is to analyze sentences absolute citations or references and accord them a college weight. As apparent in Figure 1, the summarization arrangement takes ascribe as Wikipedia articles, processes it and gives the arbitrary sentences. Ascribe book abide of raw abstracts to be candy by the system.

Figure1: text summarization architecture [1].

The evaluated system with documents containing minimum 200 to 300 words. The summary generation is carried out as following steps.

A. Preprocessing the document - In the pre-processing stage the document mainly includes three steps
   a. Identification of sentences (or) word boundary - Generally in English language the sentence boundary is identified when stop word occur, similarly the words are separated by using space.
   b. Stop word Elimination - Common words with no semantics and which do not provide important information for the final summary are eliminated.
   c. Stemming - The purpose of stemming is to obtain the stem or radix (positive integer value) of each word, which helps to improve its frequency.

B. Calculating weight of words using term frequency - Term frequency depends on the occurrence of the word the text document using formula: \( \log(n) \) where \( n \) denotes no of times a word occurs. The extracted statistical, linguistic and cohesion features weights are involved in the assignment of
word weight. The sum of the term frequency and default feature weights will be the final weight of a particular word weight.

C. Calculating the weight of the sentence - Once all the term weights are calculated then the sentence weight can be attained by calculating the mean (or) avg of all the term weight of the particular sentence. Now these sentences along with their calculated weight are to be arranged as pair.

D. Generating the summary using iterative threshold - The value of threshold would initially be set by calculating the average weight of all the sentences. Then we select the sentences that satisfy the initial threshold, then again we calculate the threshold value by averaging the weight of selected sentences that satisfy the initial threshold. This procedure is repeated until number of sentences in summary would become equal to the number of paragraphs of a document.

E. 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.

F. Algorithm
Algorithm 1 shows the abstract procedure.
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
4.3: Calculate the total score of a sentence
5: Set the threshold value for selecting important sentences by computing the average weight of all sentences initially.
6: Select the sentences that meet the threshold value only.
7: To achieve both compression and retention
7.1: Check whether there are any sentences that are repeated. Remove the repeated sentences.
7.2: Count the total number of sentences n left, if n is equal to the number of paragraphs of a document, and then these are the sentences finally included in the summary.
7.3: Else, Re-Compute the threshold value by averaging the score of all the sentences, which met the initial threshold value.
8: Repeat the step 7 until our n value will be equal to the number of paragraphs of a given document.

IV. CONCLUSION
We proposed a summarization 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 summary compared with commercial online summarizer or Microsoft summarizer tool therefore by adding cohesion feature, the text overloading problem is solved and also increased effectiveness of the summary. In future, the effectiveness of summary can be increased by adding some advanced features such as biased words.

REFERENCES