Text Summarizer Using Abstractive and Extractive Method

Ms. Anusha Pai
Lecturer in Comp. Engg
VCET Vasai Road (W)

***Abstract*— the challenge of how to make computer understand the document with any extension and how to make it generate the summary is the main motivation. Reducing the time and effort of the user of reading through entire document to know what the document is about is also the driving force behind this work. To summarize large documents of the text will be difficult for human beings. Extractive and abstractive summarization is two types of summarization. An extractive summarization method is concatenating important sentences or paragraphs without understanding the meaning of those sentences. An abstractive summarization method is generating the meaningful summary. The system uses is a culmination of both statistical and linguistic analysis of text document. Summary generated is better than mere statistical summarizers that generate summary based on word frequency calculation. Addition of plural resolution and abbreviation resolution adds more precision to summary. Concept of normalization introduced here makes sentences get their weights purely based on value of its content words and not on number of words it has. Therefore even a small but important sentence gets its place based on values of words it has. Adding linguistic features to the algorithm fine tunes the summary to higher level.**

Keywords—Automatic Summarization, extractive summary, abstractive summary.

I. INTRODUCTION

To reduce length, complexity, and retaining some of the essential qualities of the original document, will go for summarizer. Titles, key words, tables-of-contents and abstracts might all be considered as the forms of summary. In a full text document, abstract of that document plays role as a summary of that particular document. They are intermediates between document's titles and its full text that is useful for rapid relevance and quick assessment of the document. Auto-summarization is a technique generates a summary of any document, provides briefs of big documents, etc.

There is an abundance of text material available on the internet. However, usually the Internet provides more information than is needed. It is very difficult for human beings to manually summarize large documents of text. Therefore, a twofold problem is encountered. Searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. A good summary system should reflect the diverse topics of the document while keeping redundancy to a minimum.

Microsoft Word's AutoSummarize function is a simple example of text summarization. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization [1] method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An Abstractive summarization [2] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods [3] to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

II. REVIEW OF LITERATURE

Interest in automatic text summarization, arose as early as the fifties. An important paper of those days is the one in 1958, suggested to weight the sentences of a document as a function of high frequency words, disregarding the very high frequency common words. Automatic text summarization system in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights.

1. **Cue Method**: This is based on the hypothesis that the relevance of a sentence is computed by the presence or absence of certain cue words in the cue dictionary.

2. **Title Method**: Here, the sentence weight is computed as a sum of all the content words appearing in the title and (sub-) headings of a text.

3. **Location Method**: This method is based on the assumption that sentences occurring in initial position of both text and individual paragraphs have a higher probability of being relevant.

Summary Generation Based on Abstraction and Extraction

There are two very divergent methods to generate summaries automatically. Summarization based on Abstraction and Summarization based on Extraction.

Summarization by abstraction is concerned with issues of natural language processing, semantic representation and modification, text understanding and generation. This project focuses on extracting key sentences as a particular kind of computed document summary. There are several methods to do this. The first thing is to decide the important features. For instance the method discussed in [3] is based on sentence length and sentence position among other things. On the other hand researchers like [4] concentrate on the sentence length. These are typically called as the nonstructural features of the document. Another group of researchers rely on the linguistic characteristics such as the meaning of the words in the document and their relation to the understanding of the document.

Linguistic and Statistical summarizers

There are again two categories of summarizers based on the way summary is generated.

1) Linguistic summarizers.

2) Statistical summarizers.

Linguistic summarizers use knowledge about the language (syntax/semantics/usage etc) to summarize a document. Statistical summarizers operate by finding the important sentences using statistical methods (like frequency of a particular word etc).

Various techniques present involve finding the frequency of words, scoring the sentences, ranking the sentences etc. The summary is obtained by selecting a particular number of sentences (specified by the user) from the top of the list. It operates on a single document (but can be made to work on multiple documents by choosing proper algorithms for integration) and provides a summary of the document. Though there are several auto summarizers available, only a few try to combine the statistical and linguistic techniques. The linguistic summarizers are accurate but are time consuming, hence costly. On the other hand purely statistical summarizers may reduce time but are not accurate for summarizing text. Here is a system, which uses both the statistical and linguistic parameters simultaneously. This in turn will enable us to achieve accurate results while reducing the cost. Such a summarizing technique will be unique and provide a balanced and efficient system for summarizing documents.

The approach proposed [1]says the system consists of three main parts: preprocessing, analysis and selection.

Preprocessing: Tokenization, Stop-word removal, Conceptual.

Analysis and selection

-The number of main words, title words and query words

-The length of the sentence: Shorter sentences are more likely to appear in the summary.

The method suffers from shortcomings. First, it is clearly not considered abstract method. Second, Sentence score is calculated depend on number of words in the sentence and word frequency of words. Third, the system may be sensitive to the subjective user specified title and query. “Fuzzy logic” [2] is the method used for important sentence extraction using fuzzy rules and fuzzy set for selecting sentences based on their features. Each feature is given a value between ‘0’ and ‘1’.Following is the feature,-Sentence Centrality, Title feature: -Key word feature. Another approach [3], there are different approaches (statistical and linguistic) for selecting and scoring sentences. In statistical methods, sentence selection is done based on word frequency. There are several methods for determining the key sentences such as, The Title Method, The Location Method, The Aggregation Similarity Method, The Frequency Method, TF- Based Query Method, and Latent Semantic Analysis.

Fuzzy Genetic Semantic Based Text Summarization [4] :An automatic text summarization approach based on sentence extraction using fuzzy logic, genetic algorithm, semantic role labeling and their combinations to generate high quality summaries. Fuzzy IF-THEN rules were used to balance the weights between important and unimportant features. Information Retrieval by Text Summarization for an Indian Regional Language [5] The two methods of creating summaries are abstract and extract. The abstract method requires techniques like NLP, semantic parsing etc. The output is a collection of some important sentences of the text. The Local Salience Method [8] extracts phrases rather than sentences and paragraphs.

MEAD[7] this system produce both single and multi-document extractive summaries. It uses centroid-based two features, position and overlap with the first sentence. MEAD uses the CIDR Topic Detection and Tracking system to identify all the articles related to an emerging event. CIDR produces a set of clusters. From each cluster a centroid is built. Then, for each sentence, three values are computed: the centroid score, which measures how close the sentence to the centroid is; the position score indicates how far is the sentence with respect to the beginning of a document; and finally, the overlap with the first sentence or title of the document by calculating tf*idf between the given sentence and the first one. Then all these measures are normalized and sentences which are too similar to others are discarded and other sentences would be included in the summary.

WebInEssence [7] this system is a search engine to summarize clusters of related Web pages which provide more contextual and summary information. The overall architecture of the system can be decomposed into two main stages: the first one behaves as a Web-spider that collects URLs from the Internet and then it groups the URLs into clusters. The second main stage is to create a multi-document summary from each cluster.

NeATS [11]-Its architecture consists of three main components: content selection, content filtering and content presentation. The goal of content selection is to identify important concepts mentioned in a document collection. The techniques used at this stage are term frequency, topic signature or term clustering. For content filtering three different filters are used: sentence position, stigma words and redundancy filter.

NetSum [12]- The system produces fully automated single-document extracts of newswire articles based on neuronal nets. It uses machine learning techniques in this way: a train set is labeled so that the labels identify the best sentences. Then a set of features is extracted from each sentence in the train and test sets, and the train set is used to train the system. The system is then evaluated on the test set. The system learns from a train set the distribution of features for the best sentences and outputs a ranked list of sentences for each document.

GISTexter [12]-This system produces single and multi-document extracts and abstracts by template-driven IE. The system performs differently depending on working with single document or multi-document summarization. For single-documents, the most relevant sentences are extracted and compressed by rules learned from a corpus of human-written abstracts. In the final stage, reduction is performed to trim the whole summary to the length of 100 words. When multi-document summarization has to be done, the system, based on Information Ex- traction (IE) techniques, uses IE-style templates, either from a prior set (if the topic is well-known) or by ad-hoc generation (if it is unknown).

III. PROPOSED SYSTEM

A. Architecture

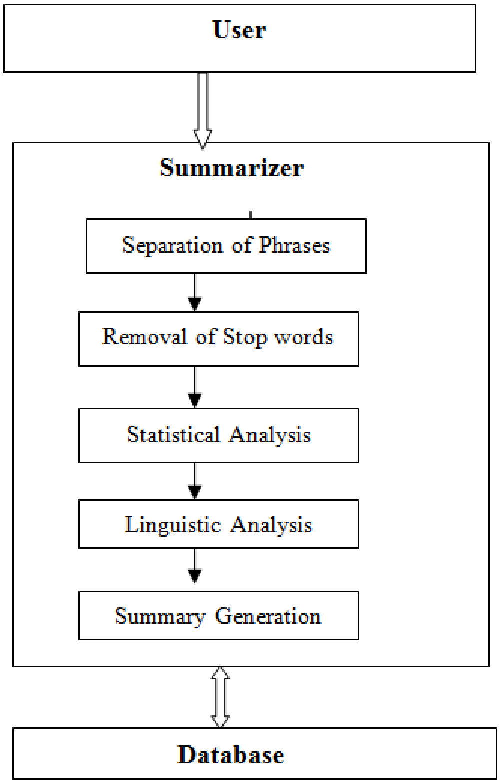

Fig.1 System Architecture

The system has three components.

1. **User**

First component is the user who uses the system. He/She enters the login details. If he/she is an authenticated user then he/she will be allowed to access the system.

2. **Summarizer**

The second component is summarizer, which generates the summary. If the user has given the keywords then system matches these words in the document and selects the sentences containing those words. Selected sentences will be displayed.

3. **Database**

Third component is database. In the database login details will be stored. There are separate tables for storing sentences, words, word frequency and sentence weight.

B. Steps used for summarization

Choice between summary generation and keywords search
Loading document, user keywords, percentage of summary expected and choosing between summary generation or keyword search.

All these inputs are taken from graphical user interface. A choice is given to the user at this point of time. The system gives two choices. The user can let system generate summary. Or he can go for keyword search. For example in a document about "data structures" if user only wants to know the uses of it, he/she can simply give key words like "use", "purpose" etc. to the system and can have only those phrases that have these words as output. A choice is made at this phase.

Various input parameters are fed into the system depending on the choice. One being percentage of summary the user wants if the choice is of summary generation. For example if the document has 100 lines. And if user wants top 40 sentences in the document he/she can give 40% as value to this parameter. If user goes for keyword search at the maximum of four to minimum of zero is expected from him/her.

Choice 1- summary generation

- **Separation of phrases**

To extract key phrases from a document it is imperative that, first separate them. The separation process is simple. Scan through the entire document and search for characters and punctuations that signify the end of a phrase, for ex. ".", "-", ";", ":", etc. this breaks the entire document into a list of phrases. It becomes easier to score phrases individually rather than the complete document.

- **Separation of words**

All words in the documents are separated and stored in the database. This is necessary for both statistical and linguistic analysis.

- **Removal of stop words**

Statistical analysis basically deals with word frequency calculation. This follows the concept that more important words occur more than once in the document. To effectively score the phrases, consider the words in the document which are contributing to the value of the phrase. All the unimportant words will have to be eliminated. Such words are called "stop words" in any language. Words such as "a", "an", "and", "as", "at", "by", "for", "from", "if", "in", "into", "on", "or", "of", "the", "to", "with" are examples of stop words for English. Thus, the next important phase will be to remove all the words from the document.

- **Word frequency calculation**

The real analysis of the document to be summarized begins at this stage. This is the statistical analysis of the document. In this stage go through the output generated in the previous stage and calculate the frequency of occurrence of each word. Thus, count the number of times each word is appearing in the document.

Consider an example that have a document on Linux operating systems. Such a document will most probably have the words "Linux" or "operating systems" or even "kernel" and "open source" occurring frequently. After going through a number of documents or articles, this trend is seen in more than 80 % of the documents irrespective of their subject or field. Thus, assume that the frequency of words is a good indicator of the content of the document itself. Now have a sequence of words w1, w2, w3,...,wn, with frequency counts of f1, f2, f3,..., fn respectively. These frequency counts are the primitives used for further scoring and will be called as word score hence forth.

- **Plural resolution**

Words "computer" and "computers" contribute to the same conceptual word computer. When refer to a particular child, it is child and a group of them as "children" but while summarizing both words child and children go together and they must contribute to one single conceptual word "child". But if it were to calculate frequency of words as the words are in the document, computation for computer and computers are done separately. Word child will have its own frequency and children its own. It requires for resolution here. In this phase such conflicts are resolved.

- **Abbreviation resolution**

Abbreviations occur frequently in the document. Standards specify that words abbreviated must be defined at its first occurrence. For example abbreviation "os" referring to "operating system" must be expanded in its first occurrence and defined as operating system (os).

When abbreviations are used in the document, they actually contribute to the words in their expansion. In this stage this problem is resolved.

- **Linguistic analysis**

In this phase language features are taken into account. The linguistic analysis focuses on the importance of words with the broad perspective of the language itself. In this part a list of important words in the English language itself. This file is maintained as a list of words and their respective multipliers ranging from 1 to 10. The multiplier is dependent on how important the word actually is. The multiplier can be considered as the priority of the word in the language. For example, words such as “firstly”, “secondly”, “therefore” are used to state new points or to conclude as is the case for “therefore”. It is obvious that phrases with such words have to be given highest consideration and so will be given a multiplier from 8 - 10. While other important words such as “state” or “consider” or “analysis” are important but not as much as those mentioned before. So the multiplier for these words will be around 4 to 7.

- **Sentence weight calculation and normalization**

In this phase weigh the sentences which had stored for analysis earlier, based on the values of words it contains. As specified before stop words are not part of analysis. Frequency of each word in the sentence is added to compute the weight of the sentence. The frequencies added will be the ones updated after plural resolution, abbreviation resolution and linguistic analysis.

More is the sentence length higher will be its weight. So there is a risk of losing a small but important sentence. Hence the concept of normalization has been introduced.

Normalized Weight = Total weight of the sentence/No. of words in the sentence.

These are weights of sentences/phrases in the document now. Stop words are not considered.

Display of summary

After sentence weights are calculated, depending on the percentage of summary requested by the user, those many sentences are highlighted in the document.

Choice 2-keyword search

If the user goes for simple keyword search of the document, the following phases occur.

- **Separation of phrases.**

This is same as the sentence separator of summary generation phase. All sentences are separated and stored in the database.

- **Display of filtered sentences**

Only those sentences are displayed on the screen that had the keywords user gave.

IV EXPERIMENTAL RESULTS AND PERF ORMANCE

Summary is generated of document, depending on the percentage of summary given as input by user. As per analysis, (i.e compared summaries of a few documents generated by both the summarizers). Found through logical deduction that summarizer chose better sentences for summary. For any type of document and any number of pages of the document, summary can be generated. Tested work by comparing with previous system and with the golden summary created by humans. To create a golden summary from the human summaries, gathered about 10 human summaries for each document. Finally chose the sentences with the most frequency in human summaries to be included in the golden summary.

For this task, examined the sentences occurring in the summary generated by the system to see how popular they have been in the summaries created by human and so how probable is their occurrence in the golden summary. A table for each document to compare system results with the golden summary. Table 1, 2, 3 shows result of single document containing 10 sentences (sentence 1 to sentence 10) and summary size is 40%. Comparison between System summary (ATS summary), Human summary, Microsoft word summary, SweSum system summary. Table 1 shows the situation for one of the good results obtained for summarizing a scientific article.

TABLE 1 : The golden summary with the results of Text Summarizer

No of chosen sentences(in the order of the most important)	% of sentence appearance in human summary	Presence of sentences in ATS summary
1	100	√
7	100	√
5	80	√
4	60	√
2	20	×
10	20	×
8	20	×

Compared Text Summarizer with Microsoft word 2007 Auto summarizer.

TABLE 2: Comparing Text Summarizer with Microsoft Word 2007 Auto summarizer

No of chosen sentences(in the order of the most important)	Presence of sentences in Microsoft word summary	Presence of sentences in ATS summary
1	√	√
7	×	√
5	×	√
4	√	√
2	×	×
10	√	×
8	×	×

Compared Text Summarizer SweSum Summrizer

TABLE 3: Comparing Text Summarizer with SweSum summarizer

No of chosen sentences(in the order of the most important)	Presence of sentences SweSum summary	Presence of sentences in ATS summary
1	×	√
7	√	√
5	×	√
4	√	√
2	×	×
10	√	×
8	√	×

A comparative analysis of 20% summary generated using the sample English article by machine with human is done. The article is given to five different individuals and asked them to produce the 20% summary. The average of those five summaries is generated by taking the common sentences of each summary and that average summary is used as human summary for comparison with machine generated summary.

The following results of both the summaries are compared.

• Total word count at 20%, 30% and 40% summary size.

• Common and uncommon sentences generated in both the summaries.

As shown in "Fig.1," it can be interpreted that in English summarization, the total number of word count in machine generated summary is more compared to the human generated summary and the percentage of word count has increased in both the summaries as the summary size is increased.

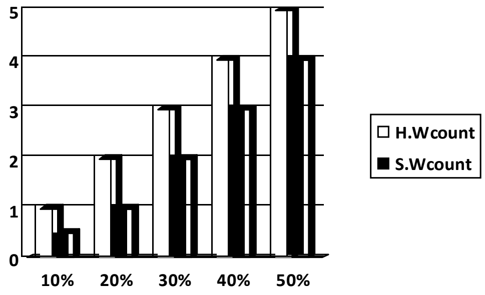

Fig.1. Total word count at different summary size

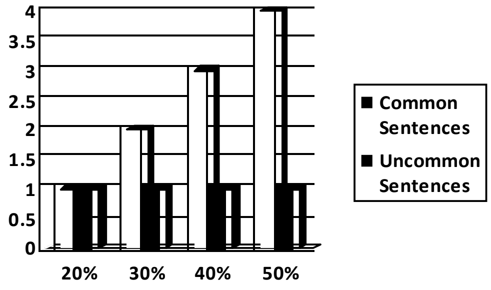

Fig.2. Common and Uncommon Sentences at different summary size

Based on the above Fig. 2, it can be interpreted that in 20% summary out of the two sentences one (50%) are common, in 30% summary out of the three sentences two (67%) are common and in 40% summary out of the four sentences three (75%) are common. It indicates that as the percentage of summary size increases the percentage of common sentences also increases.

CONCLUSION

To conclude that the system uses is a culmination of both statistical and linguistically analysis of text document. Summary generated is better than mere statistical summarizers that generate summary based on word frequency calculation. Addition of plural resolution and abbreviation resolution adds more precision to summary. Concept of normalization introduced here makes sentences get their weights purely based on value of its content words and not on number of words it has. Therefore even a small but important sentence gets its place based on values of words it has. Adding linguistic features to the algorithm fine tunes the summary to higher level. Thus in summary, system generates precise summary of documents saving the user of time and trouble of going through all the lines present in the document.

FUTURE SCOPE

NLP is one such enigmatic ocean which unfolds more and more mysteries the deeper go into it. All efforts aim towards making a dumb machine (computer) intelligent.

- **Synonyms resolution**

The word "use" and "purpose" are synonyms. Can consider it for word frequency calculation It will generate better summary. This can be done using software called "WorldNet".

- **Multiple documents summarization**

This concept can be extending for summarization of multiple documents. Two three documents on single domain and generate combined summary of it.

Making computer understand the documents by itself and generate precise summary which is finally the ultimate goal of all researchers on the planet.

REFERENCES

[1] Mehrnoush SHAMSFARD, Tara AKHAVAN, Mona ERFANI JOURABCHI." PARSUMIST: A Persian Text Summarizer". 978-1-4244-4538-2009 IEEE.

[2] Ladda Suanmali, Mohammed Salem Binwahlan and Naomie Salim Faculty "Sentence Features Fusion for Text Summarization Using Fuzzy Logic". 978-0-7695-3745-0/09 $25.00 © 2009 IEEE

[3] Saeedeh Gholamrezazadeh Mohsen Amini Salehi Bahareh Gholamzadeh," A Comprehensive Survey on Text Summarization Systems", 978-1-4244-4946-0/09/$25.00 ©2009 IEEE.

[4] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan," Fuzzy Genetic Semantic Based Text Summarization", 978-0-7695-4612-4/11 $26.00 © 2011 IEEE

[5] Jagadish S KALLIMANI, Srinivasa K G, Eswara REDDY B," Information Retrieval by Text Summarization for an Indian Regional Language", 978-1-4244-6899-7/10/$26.00 c2010 IEEE.

[6] 1 Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", *In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science*, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[7] Text Summarization : An Overview by Elena Lloret.

[8] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1- 12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[9] 2 Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", *In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science*, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.

[10] 3 Udo Hahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator", *Proceedings of the first International conference on Human language technology research*, *Association for Computational Linguistics* , ACM, Morristown, NJ, USA , 2001.

[11] TEXT SUMMARIZATION : AN OVERVIEW _ Elena Lloret Dept. Lenguajes y Sistemas Inform_aticos Universidad de Alicante Alicante, Spain elloret@dlsi.ua.es.

[12] A Survey on Automatic Text Summarization Dipanjan Das Andr_e F.T. Martins.