ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

A Text Summarization using Modern Features of Sentences

Birari Dhiraj Abhiman, Prof.P. P.Rokade

Dept of Computer Engineering, S.N.D College of Engineering & Research Center, Babhulgaon, Yeola, India

Dept of Information Technology. S.N.D College of Engineering & Research Center, Babhulgaon, Yeola, India

ABSTRACT: Automatic Text summarization evaluation is very important to the development of summarization systems which generates rational summaries that states the main goal of the given document. With the growth of amount of textual Information, automatic summarization of textual information is in urgent need for efficient processing of the gigantic information from huge, well-structured, coherent documents.
Automatic summarization is challenging problem in computational linguistics, since text summarization is an effective tool for processing large information resources in Computer world. Here we have proposed paper in which we have studied different features for text summary extraction from given large documents and studied its result in terms of number of features considered for extracting text summary. Extracted summary result naturally affected by size of documents if it is large then limited number of considered features may cause to the unexpected result like to generate poorly linked sentence or incoherent summary.

KEYWORDS: Text Summarizer, Extractive summary, POS tagging,Feature Extraction, Information Retrieval, Artificial Intelligence, Fuzzy Systems..

1. INTRODUCTION

Automatic Text summarization needed since last 1960 and last 30 years people are working to find out solution in better way. From 1990, WWW came to existence and rapidly data transaction and utilization increases. Due to different and huge size of documents, existing work not giving expected result in text summarization [6][7]. Here we are presenting one of the good approaches for finding out text summarization by using various features so it generate coherent sentence result in which user can see the linking between them. We are evaluating our results with existing technologies like MS Word, Manual Summary.

As per the limitations of the different summarizing systems due to selection of lack of features, it causes to generate irrelevant summary. Here WordNet is used to confirm the semantic correctness of the textual document generated at the syntactic analysis. It gives all hyponyms and synonymy for a selected noun to the user. We have used WordNet to find semantically related and similar meaning terms in text documents. It is used to find out words which are semantically related to each other. Also it is useful to calculate the words occurrences in documents and find out its frequency in the document.

In this paper, Preprocessing algorithm works in mainly three steps, first is sentence count in document, second is sentence segmentation and word steaming, third is sentence scoring. It uses score of sentences and ranks it. It focuses on frequencies, word occurrences, position of sentence in the document, indication words and phrases, and measuring lexical similarity. Here we have include few more features along with nine features for extraction of summary of text that are

i) Alpha Numeric Sentences
ii) Morphological Sentences
iii) Punctuations
iv) Capital letters
v) Adjectives

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

The above approach for summarization gives better performance as compared to other summarization tools. And we are confident about our result that top ranked sentences are most of the time extracted which are the most important ones.

II. RELATED WORK

Since many years ago for meeting text summarization, different evaluation methods and approaches have been developed like in 1998 Marcu developed such approach; in 2001 Chali Y. & Brunn M., also in 2001 Maybury and Mani was tried for text summarization; Mani 2001; Alonso and Castellon 2001.

In this classification, automatic text summarizers can be described as approaching the problem at the entity, surface, or discourse level. Since it observed that current summarizing systems having many limitations, constraints. And generated text summary contains poorly linked sentences and are not relevant to the subject[6][7].

Deerwester S. proposed a approach for text summarization by Indexing by latent semantic analysis [3] which is tried to overcome problem of retrieval techniques based on extraction result by using word queries and word of documents. But in latent semantic analysis there may be chances of selecteion of unimportant or irrelevant concepts from document. Because one word having many meaning and if we are failed to provide evidence for extracting text by using latent semantic techniques then users query may not find out expected output. Deerwester S. used Latent Semantic Indexing (LSI) for overcoming this unreliable output. It uses a Matrix technique which is based on Singular Value Decomposition method [4].

In "Summarizing text by ranking text units according to shallow linguistic features" [5], this approach identifying the most important sentences from given input text using shallow linguistic features. They have focused on degree of connectivity between sentences. It results into coherent and expected output which reduces non coherent sentences from resulting summary.

This is known as surface-level approach which considered mainly 6 points for ranking the sentences as well as sum of score of each word in each sentence in documents for extracting text summary are as follows;
i. Term Frequency of word ii. Location of word iii. Bias: meaning of word iv. Cue Word and Phrases v. Word co-occurrences: word and paragraph score is find out. vi. Lexical Similarity: Wordnet is used. For scoring word it uses vector space model, heuristics rules for coherent output.

Still it's having limitation of completeness because of extraction takes place at the sentence boundary only. This generate problem where highly compressed summary is required in that case it may left important data [5].
Rajesh S. Prasad, U. V. Kulkarni "Connectionist Approach to Generic Text Summarization," [6] also proposed a approach which aims for a large document's text summarization. It used POS tagging with repeated neural network concept [6].

Microsoft Office Word Summarizer tool [12] can be found in Microsoft Office Word 2003/2007. This tool produce summaries of few sentence like 10 to 20 or 100-500 sentences i.e. 10% up to 75% of words summary of the given input original document.

III. A MOTIVATING SCENARIO

- It uses modern featured base text summarization (MFBTS) algorithm for generating coherent and linked sentence summary.
- It uses stemming algorithm for removing affixes and suffixes of word.
- It uses WordNet [8] to find semantically similar terms, and for the gaining of synonyms. It is used to validate the semantic correctness of the sentences generated at the syntactic analysis.
- It also uses StopWord dictionary to restrict stop word to be included into summary.
- It uses modern features for extraction of summary like Alpha Numeric Sentences, Morphological Sentences, Punctuations, Capital letters, Adjectives.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

- We have used context-based text interpreter Algorithm (CFTI) which performs syntactical analysis and lexical semantic processing of sentences.
- It uses Vector Paragraph Model which allows ranking documents according to their relevance in word by finding out term frequency.
- We are using Fuzzy logic scoring for scoring sentences and paragraphs.
- We have also used Supervised Learning Model for processing the non-duplicate text, converts meaning text and calculate the Score of each text and calculate the summary of each text.

IV. IMPLEMENTATION

Here we are presenting how text summarization takes place effectively on given large document as an input.

We are performing number of functionality on given input documents such as Stemming algorithm, stop word dictionary, sentence counting and breaking sentence into segments, sentence scoring as well as paragraph scoring and finally generation of Summary.

Here we have proposed Modern Featured Based Summarization i.e. MFBS.

We illustrate the algorithm of this module by the following steps:

- **Step1:** Document Parser is done by using stemming algorithm. Stop words are removed by comparing input text with Stop word dictionary.
- **Step2:** By using Heuristic rules, input document is segmented into sentences and paragraphs. Also Sentence count is done.
- **Step 3: Feature extraction:** The document after preprocessing is subjected to feature extraction by which the properties of the sentences are extracted to score the sentence.
- **Step 4:** Vector paragraph model is used for ranking.
- **Step 5:** Indexing is done for respective word in document which bust up the performance of the system.
- **Step 6**: Sentence and paragraph scoring is done by using Fuzzy logic by considering cue phrases, word similarity in sentence as well as in paragraph, iterative query sore
- **Step 7:** Sentence with highest score is selected for summary by using supervised learning model.
- **Step 8:** Text Summary generation i.e. Synthesis.

V. SYSTEM ARCHITECTURE

Here we are presenting our system works which are manly depends on fourteen features for extraction of text summary with more accuracy. We have observed that with more features, we can get more precession and recall value as a performance parameter as compare with others.

In this implementation, we make clear the Summary Generated by the Word similarity among sentences, Word similarity among paragraphs, Iterative query score, Format based score, Numerical data, Cue-phrases, Term weight, Thematic features, Title features, Alpha Numeric Sentences, Morphological words, Punctuations, Capital Letters, Adjectives.

We have used the Stanford Part of Speech tagger to identify nouns and adjectives in the sentences which are present in document.

Following System Architectures shows functionality of our system.

5.1 Pre-Processing mainly three activity performed.

a. Tokenization is done by using parsing and POS tagger. Document is breaked into segmentation.

b. Stop word removal: Stop words are unimportant and these are already predefined in stop word dictionary. While comparing with input document, it is detached from extracted summary.

c. **Stemming**: it is used to remove suffixes & affixes. it contains few rules like;
   - If the word or concept is plural convert it into Singular form.
   - If the word or concept ends in 'ed', remove the 'ed'

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

- If the word or concept ends in 'ing', remove the 'ing'
- If the word or concept ends in 'ly', remove the 'ly'
- Different relationship between concepts words from "vocabulary-of-concepts" is recognized.

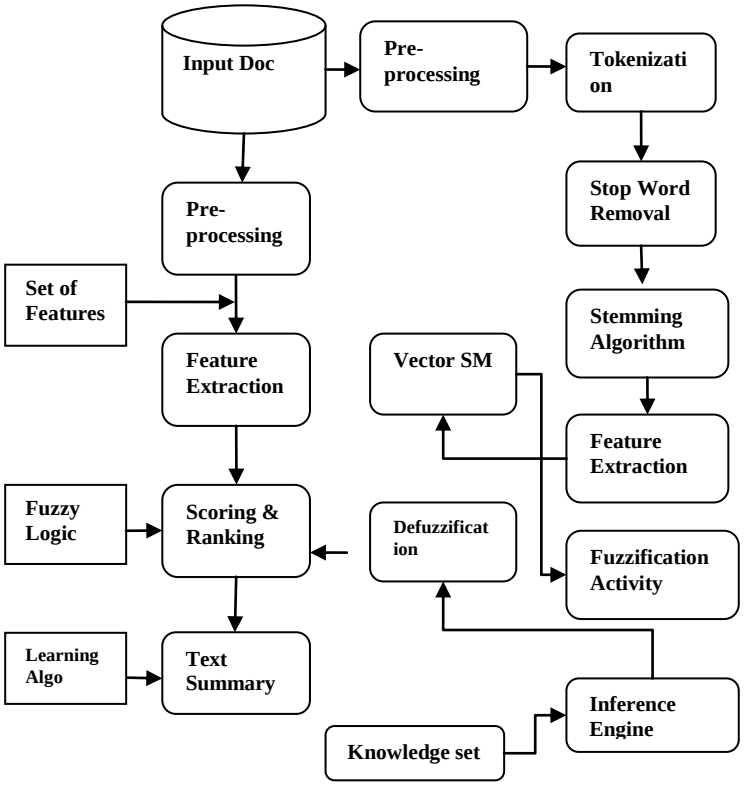

Fig A: Text Summarization system architecture.

5.2 Modern Features matching, extracting and word score

It uses mixture of fourteen features for extraction of text summary which is essential for large document. These features are useful for assigning score to the words, sum of word's score in sentences and also to the paragraph.

1) Numerical data:
2) Cue-phrases
3) Word similarity among sentences
4) Title features
5) Word similarity among paragraphs
6) Iterative query score
7) Format based score
8) Term weight
9) Thematic features
10) Alpha Numeric Sentences
11) Morphological words
12) Punctuations
13) Capital Letters
14) Adjectives

5.3 Fuzzy Logic

It is a mechanism for assigning score to sentences in paragraph. It is introduced in 1960 by Zadeh [9]. It assigns value between 0 to 1. It's having mainly 3 aspects;

i. Fuzzifier
ii. Inference Engine

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

iii. Defuzzifier

i. Fuzzifier

It converts input data into respective score values i.e. feature's score of each sentence in processing input document. This score value is presented into vary low, low, medium, high and very high which is in the form of linguistic value.

Fuzzy set is a class of objects. Let X be a space of point or objects.

Fuzzy Set = {x,f(x)} where x is extracted feature and $f_A(x)$ is membership function.

It is characterized by membership function.
I.e. Fuzzy Set A in X characterized by,

$f_A(x) = \{0,1\} = 0$ or 1

Ex. Suppose A is a set of integers from 0-1000 then

$f_A(0) = 0$; $f_A(17) = 0.1$; $f_A(500)=0.5$; $f_A(1000) = 1.0$.
$f_A(700) = 0.76$ etc….

ii. Inference Engine

It Compare generated set with knowledge base set and it assigns level of importance in terms of unimportant, average & important which are linguistic value.

iii. Defuzzifier

This mechanism converts linguistic value into crisp value (0 to 1).

Thus output of fuzzy logic i.e. crisp value is assigned to every sentence in document. Here different features plays important role for determining text summary.

5.4 Feature Extraction

We have defined fourteen different rules for finding out score of respective feature. Here we are also using **Vector Space Model (VSM)** for representing word in document. We can find out each word frequency speedily. Features like;

Numeric Data (ND) gives some important in paragraph and reduces noise. It gives preciseness of document. Therefore we are assigning score to numeric data as a ratio of,

$$\mathbf{ND(s)} = \frac{\text{Length of ND in sentences}}{\text{Sentence Length}} \qquad \text{---------- (a)}$$

Alpha Numeric (AN) Sentences are combination of alphabetic and numeric character. It may be keyword, password or any mathematical formula which plays important role for any conclusion.

$$\mathbf{AN(s)} = \frac{\text{No. of AN word in sentence}}{\text{No. of AN word in document}} \qquad \text{---------- (b)}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Morphological Word (MW) gives meaning and idea of word structure. How the word is related to the other word in given document. Words are made of morphemes at the basic level ex. Schoolyard = School+Yard. It may also stop word (SW) since that should be removed.

$$\textbf{MW(s)} = \frac{\text{No. of MW word in sentence - SW in sentence}}{\text{Sentence Length}} \quad \text{---- (c)}$$

Punctuations in documents also indicates importance of words, sentence as well as paragraph like hyphens uses in adjective or sentence connectivity, brackets, Quotations (“”), Question mark (?), exclamation mark (!) etc… For (?),(” ”), (!) We have assigned more score for considering in final summary.

Adjectives which describe and clarify noun. It defines properties of Noun. High score is given to the sentences which contains such adjectives.

$$\textbf{Adj(s)} = \frac{\text{No. of Adj word in sentence}}{\text{Total No. of Adj word in document}} \quad \text{------- (d)}$$

5.5 Ranking of sentence

As per the score assigned to the sentences in document, sorting of sentences done in descending order.

5.6 Text Summary

User predefines size of summary record and sentences are selected in final text summary as per the given size for summary.

5.7 System Mathematical Modeling

The proposed system S is defined as follows:

S = { I, O, F, U }

Where,

I: Input
O: Output
F: Functions
U: User

Where

I= {U, TS, FE , FL}

Where

U = User which having Text summarization
TS = Text Summary
FE = Different features extraction from given input text.
FL = Fuzzy Logic for assigning score to sentences.

O = {WS, SW, FE, SR, WI, TSG}

Where below are the output generated from system processing;

WS = Word steaming.
SW = Processed Text to remove unwanted stop words.
FE= Features Extraction by using fourteen keywords. SR = Sentence Ranking by using fuzzy logic mechanism.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

WI= Word Indexing by using fuzzy logic.
TSG = Finally Text Summary Generation.

U = {SV, OU, A}

Where
SV = System Visitor
OU = Online User
A= Administrator

F= {F1, F2, F3, F4, F5}

Where

Function F1 : Document Parser is done by using stemming algorithm. Stop words are removed by comparing input text with Stop word dictionary.

Function F2 : The document after preprocessing is subjected to feature extraction by which the properties of the sentences are extracted to score the sentence.

Function F3: Vector paragraph model is used for ranking sentences and Indexing of Words.

Function F4 : Sentence and paragraph scoring is done by using Fuzzy logic i.e fuzzification and defuzzification.
Fuzzy set is a class of objects. Let X be a space of point or objects.
Fuzzy Set = {x,f(x)} where x is extracted feature and $f_A(x)$ is membership function.

Function F5 : Sentence with highest score is selected for final summary by using supervised learning model.

VI. RESULT AND EVALUATION

The performance of the Text Summarization system can be assessed by determining the quality of text summary [12]. It is find out by precision and recall value. Precision denotes the ratio of preciseness of the sentences in the text summary and Recall value calculates the ratio of number of coherent sentences included within the summary.

Table 1: Result Analysis of different existing system.

Sr. No.	**Tools**	**Recall Value**	**Precision Value**
1	Copernic Summarizer(Feb 2003)	77.00%	80.00%
2	Intellexer	70.83%	82.50%
3	MS word	62.50%	59.16%
4	Fuzzy Logic (2009)	79.00%	79.00%
5	**Proposed System (MFBS)**	Above 90%	Above 90%

VII. CONCLUSION

Day by day, drastically increasing data load on server and finding out important summary or pattern from huge data is very crucial task to maintain accuracy in output text summary. Lots of work is done since MS-Word. It is

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

also providing summary but not giving accuracy. Current our research is focused namely on modifications of the existing approaches, or their combination.

We can observed from evaluation table 1 in which our proposed work will give more precision and recall value around 90% in terms of accuracy parameter and we are confident due to different combination of modern features that we are considered.

It proves that when large document is given as a input then it is must to consider all fourteen features for extraction of text summary with more accuracy. We can define here future work in our research that system should be able to find out necessary features while extraction of text summary so whenever document size is less, our system will be able to reduce number of features those are not required and improve time space complexity.

ACKNOWLEDGMENT

I am very thankful to the people those who have provided me continuous encouragement and support to all the stages and ideas visualize. I am very much grateful to the entire S.N.D College Of Engineering & Research Center for giving me all facilities and work environment which enable me to complete my task. I express my sincere thanks to Prof. P. P. Rokade, Prof. S. R. Durugkar, Head of the Computer Department, S.N.D College of Engineering & Research Center, Yeola who gave me their valuable and rich guidance and help in presentation of this research paper.

REFERENCES

[1] Using lexical chains, In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001, New Orleans, Louisiana, 2001.
[2] Brunn M., Chali Y., and Pinchak C. 2001, Text summarization using lexical chains, In Workshop on Text Summarization in conjunction with the ACM SIGIR Conference 2001, New Orleans, Louisiana, 2001.
[3] Deerwester S., "Indexing by latent semantic analysis", J. Ameri. Soci. Inf. Sci., Vol. 41, No. 6, pp. 391–407,1990.
[4] Landauer T.K., Foltz P.W. and LahamD. ,"Introduction to latent semantic analysis", Discourse Processes, Vol. 25,pp. 259–284,1998.
[5] Pankaj Gupta, Vijay Shankar Pendluri, Ishant Vats, "Summarizing text by ranking text units according to shallow linguistic features", Feb. 13~16, 2011 ICACT, 2011.
[6] Rajesh S. Prasad, U. V. Kulkarni, Jayashree R. Prasad, "Connectionist Approach to Generic Text Summarization,", World Academy of Science, Engineering and Technology 55, 2009.
[7] Uplavikar N.M., Wakhare S.S., Dr. R.S. Prasad "Feature Based Text Summarization" IJACIR, ISSN: 2277-4068, Volume 1– No.2, April 2012.
[8] WordNet (2.1)http://www.cogsci.princeton.eu/~wn/. Haruhiko Kaiya, Motoshi Saeki, 2005, "Ontology Based.
[9] Zadeh, L.A., 1965. Fuzzy sets. Inform. Control, 8: 338-353. DOI: 10.1016/j.fss.2004.03.027
[10] Copernic Summarizer in Feb 2003.
[11] Brunn M., Chali Y., and Pinchak C. 2001, Text summarisation using lexical chains, In Workshop on Text Summarisation in conjunction with the ACM SIGIR Conference 2001, New Orleans, Louisiana, 2001.
[12] René Arnulfo García-Hernández, Yulia Ledeneva, Griselda Matías Mendoza, Ángel Hernández Dominguez "Comparing Commercial Tools and State-of-the-Art Methods for Generating Text Summaries" 2009 Eighth Mexican International Conference on Artificial Intelligence by IEEE.