EasyChair Preprint

№ 5377

A Text Summarization Using Multi Linguistic Features and Fuzzy Logic Technique of Sentences

Dhiraj Birari and Yogesh Palve

April 25, 2021

A Text Summarization Using Multi Linguistic Features and Fuzzy Logic Technique of Sentences

BIRARI DHIRAJ ABHIMAN[a,1] and PALVE YOGESH HIRAMAN[b]
[a]*Department of Information Technology*
MVPS's KBTCOE, Nashik-13
birari.dhiraj@kbtcoe.org

[b]*Department of Information Technology*
MVPS's KBTCOE, Nashik-13
palve.yogesh@kbtcoe.org

Abstract. Automated word specification assessment is very important to the progress of characterization systems which develop analytical analysis that case the main target of the given evidence. Along the improvement about bulk of text file material, automated characterization of text file material is now in necessary use for productive preparing of the enormous material against enormous, strong - framed, consistent material. Automated characterization is assert issue now in combinational syntax, during the time that word characterization act an efficient mechanism for prepare large material assets now in PC earth. present individually gain scheduled paper now in that individually include prepared other appearance as content survey eradication against given broad material also prepared owned conclusion now in details of total containing presence treated for derive content analysis. derive arbitrary conclusion commonly concerned by amount of material in case appeal is broad then narrow number of treated appearance can begin to the abrupt conclusion alike to achieve diseased related decision either irrational abstract.

Keywords. Word compile, removal synopsis, POS label, Feature derivation, data Improvement, Artificial Intelligence, flossy structure.

1. Introduction

Automated content characterization required as final 1960 and final 30 years community do running to search out result in superior approach. Against 1990, WWW came to alive and quickly input action and usage raise. Require to unlike along with enormous amount of material, current task not giving normal result in content characterization [6][7]. Now we are given one of the good access to conclusion away content characterization through applying individual attribute so it create consistent sentence conclusion through which user can look the joining between them. Individually are appraise our conclusion along with current automation alike MS Word, Manual Summary.

In the act of constraint of the different summarizing structure requires to selection of lack of characteristics, it source to generate irrelevant summary. Here WordNet is used to approve the semantic correctness of the textual document generated at the syntactic analysis. It gives all hyponyms and synonymy for a preferred noun to the user. We have used WordNet to find semantically related and similar meaning terms in word

material. It is used to find out words which are semantically associated to each other. Likewise, it is useful to calculate the words accident in documents and find out its frequency in the document.

In this paper, preprocessing algorithm works in primarily three steps, first is sentence calculation in document, second is sentence segmentation and word steaming, and third is sentence tally. 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 combined 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

The above way for characterization offer better performance as correlated to other characterization tools. And we are satisfied about our result that top ranked sentences are most of the time extracted which are the most important ones.

2. Related Work

Since many years ago for meeting content characterization, 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 word characterization can be described as approaching the problem at the entity, surface, or discourse level. Since it noticed that current characterizing systems having many limitations, constraints. And generated text summary contains poorly linked sentences and are not relevant to the subject [6][7].

Deerwester S. recommended a approach for word characterization by Indexing by latent semantic analysis [3] which is tested to overcome problem of retrieval techniques established on extraction result by using word queries and word of material. But in latent semantic analysis there may be chances of selection of unimportant or irrelevant concepts from document. as one word having many meanings 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 determine the most essential sentences from given input text using shallow linguistic features. They have focused on degree of connectivity among sentences. It results into coherent and expected output which reduces non coherent sentences from resulting summary.

This is known as surface-level approach which treated 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 add word it uses vector space model, heuristics rules for coherent output. Still it’s having restriction 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].Second paragraph.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 produces summaries of few sentences like 10 to 20 or 100-500 sentences i.e., 10% up to 75% of words summary of the given input original material.

3. A Motivating Scenario

- It uses modern featured base text summarization(MFBTS) algorithm for generating logical and linked sentence summary.
- It uses stemming algorithm for delete affixes and suffixes of word.
- It uses WordNet [8] to identify semantically similar condition, and for the gaining of synonyms. It is used to validate the semantic correctness of the sentences develop at the syntactic analysis.
- It also apply Stop Word dictionary to restrict stop word to be admitted into summary.
- It applies modern features for extraction of summary like Alpha Numeric Sentences, Morphological Sentences, Punctuations, Capital letters, Adjectives.
- We have used context-based text interpreter Algorithm(CFTI) which performs syntactical analysis and lexical semantic preparing of sentences.
- It applies Vector Paragraph Model which grant ranking documents according to their relevance in word by finding out term frequency.
- We are applying Fuzzy logic scoring for scoring sentences and paragraphs.
- We have also apply 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.

4. Implementation

Here we are producing how content characterization takes place effectively on given large material as an input.

We are performing number of functionalities 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 analysis. Here we have proposed Modern Featured Based characterization 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. Likewise 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 complete for respective word in document which bust up the performance of the system.
- **Step 6:** Sentence and paragraph scoring is done by applying Fuzzy logic by considering cue phrases, word similarity in sentence as well as in paragraph, iterative query score
- **Step 7:** Sentence with highest score is choose for summary by using supervised learning model.
- **Step 8:** Text Summary generation i.e., Synthesis.

5. System Architecture

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

In this execution, 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 fragment 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 delete 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'
    - 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.

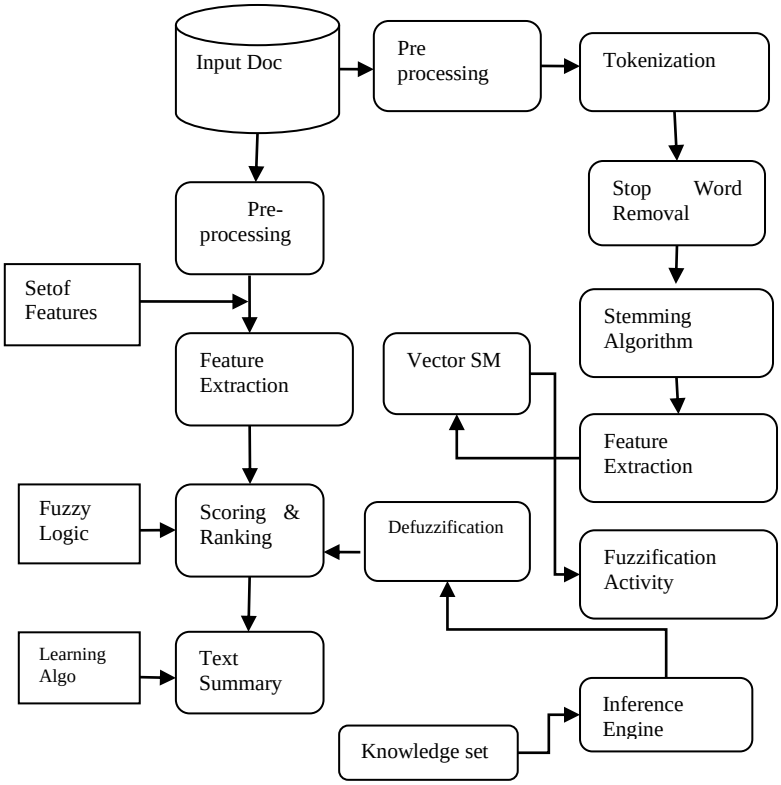

Figure 1. Content characterization system architecture.

5.2 Modern Features matching, extracting and word score

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

- Numerical data:
- Cue-phrases
- Word similarity among sentences
- Title features
- Word similarity among paragraphs
- Iterative query score
- Format based score
- Term weight
- Thematic features
- Alpha Numeric Sentences
- Morphological words
- Punctuations
- Capital Letters
- Adjectives

5.3 Fuzzy Logic

It is a process 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;

- Fuzzifier
- Inference Engine
- Defuzzifier

I. Fuzzifier

It transform input data into respective score values i.e., feature's score of each sentence in processing input document. This score value is intorduce 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 fA(x) is membership function.

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

FA(x) = {0,1} = 0 or 1

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

FA (0) = 0; fA (17) = 0.1; fA (500) =0.5; fA (1000) = 1.0.
FA (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.

II. Defuzzifier

This procedure 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 play main role for determining text summary.

5.4 Feature Extraction

We have interpret fourteen different rules for finding out score of respective features. Here we are also apply Vector Space Model (VSM) for representing word in document. We can observe out each word frequency speedily. Characteristic like;

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

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

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

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

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

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

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 describes properties of Noun. High score is given to the sentences which contains such adjectives.

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

5.5 Ranking of sentence

As by 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 choose 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.
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 cut 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 fA(x) is membership function.
Function F5: Sentence with highest score is selected for final summary by using supervised learning model.

6. Result and Evaluation

The performance of the content characterization system can be assessed by determining the character of text summary [12]. It is observe 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. Following figure shows fuzzification and defuzzification of input doc for generation of text summary.

Figure 2. Content characterization Using Modern Features of Sentences front screen

Graph 1: Comparison of Recall and Precision Value in existing and proposed system Modern featured base text summarization.

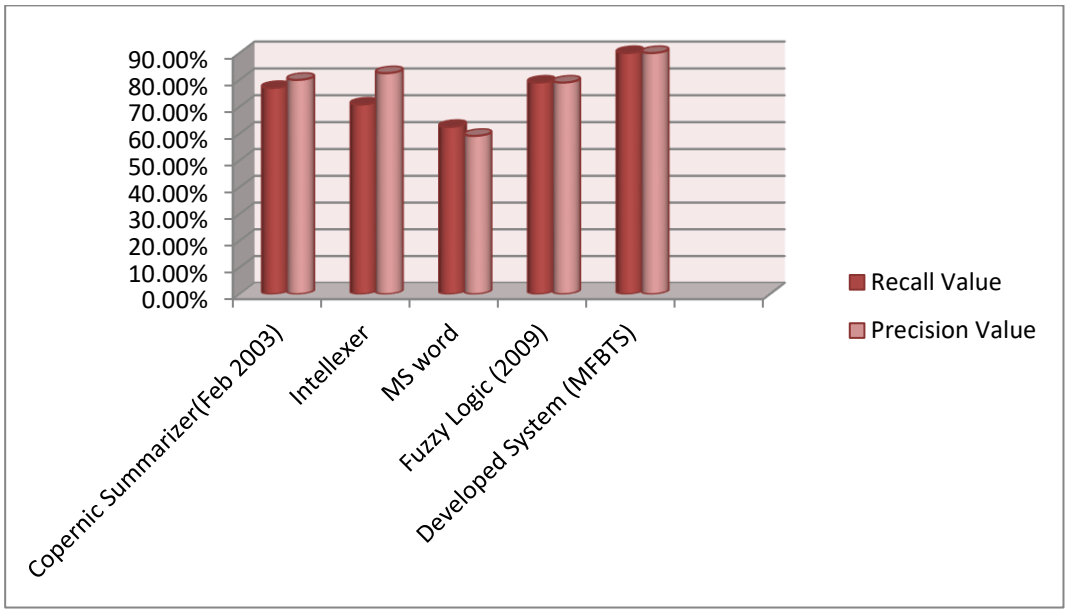

Figure 3. Performance evaluation graph.

Performance Analysis of Modern featured base text summarization (MFBTS) with existing Tools.

Table 1: Comparison between MFBTS and Existing Tools.

Features Extraction	Copernic Summarizer (Feb 2003)	Intellexer	MS word	Fuzzy Logic (2009)	Developed System (Considering 14 features for extraction)
Alpha Numeric Sentences	NO	NO	NO	NO	YES
Morphological Sentences	NO	NO	NO	NO	YES
Punctuations	NO	NO	NO	NO	YES
Capital letters	NO	NO	NO	NO	YES
Adjectives	NO	NO	USER	YES	YES

Tables shows that developed System, there are very few tools those can extract summary information such as non-repetitive and as brief as possible. A summary should be indicative. It should indicate the document's relevance to the reader.

Thus, the results of this initial performance evaluation are very encouraging and support developed approach here and the potential of this technology in general.

Conclusion:

Day by day, extremely improve data load on server and detection out important summary or pattern from huge data is very essential task to maintain efficiency in output text summary. Lots of work is done since MS-Word. It is also providing summary but not giving accuracy. Present our research is concentrated namely on modifications of the existing approaches, or their combination.

I distinguish from evaluation table 1 in which our proposed work will offer 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 show that when huge 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 structure 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 increase time space complexity. In future work we will try to extend out usefulness of a content characterization using modern features of sentences for supporting huge database as well as for multiple languages.

Acknowledgment

I am very thankful to the people those who have provided me continuous encouragement and support to all the stages and ideas visualization of this 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 summarization using lexical chains, In Workshop on Text Summarization 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.

[13]Uplavikar N.M., Wakhare S.S., Dr. R.S. Prasad "Feature Based Text Summarization" IJACIR, ISSN: 2277-4068, Volume 1– No.2, April 2012.

[14]Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artif. Intell., 139(1):91– 107, 2002.