Assesment of Various Text summarization algorithms on the basis of ROUGE-L method

ROUGE-L method implementation using the Longest Comman Subsequence technique

Prashali Gupta
M.Tech Student
Department of Computer Science & Engineering
AIM & ACT, Banasthali University
Jaipur, Rajasthan
prashaligupta.15@gmail.com

Yogesh kumar Meena
Assistant Professor
Department of Computer Science & Engineering
Maliviya National Institute of Technology
Jaipur, Rajasthan
yogimnit@gmail.com

***Abstract*— Document summarization is the technique for understanding the main theme of any kind of document quickly. Document summarization depends on the user requirement, and classified in different ways. Here we study single document extractive text summarization and given different method for summarizing the text. This paper describes and performs a qualitative assessment of 10 algorithms for scoring the sentence. Sentences are scored on the basis of the feature of sentence and on the basis of the skeleton of the document. For extractive text summarization, sentences are extracted on the basis of the score and generate the required summary. These algorithms apply on the data set and for evaluation use ROUGE-L method.**

Keywords—ROUGE-L method; TF/ID; thematic;

I. INTRODUCTION

Text summarization is the process of automatically creating the shorter version of the text document or file. Text summarization provides users with summaries of text document, allowing them to quickly understand the main idea of documents. Summarization is a brief and accurate representation of input text such that the output covers the most important concepts of the source in a condensed manner.

Text summarization has become an important tool for analyzing and interpreting text documents in a fast growing information world. Different definition of text summarization is available first is - "Summary can be defined as a text that is produced from one or more texts, that contains a significant portion of the information in the original text, and that is no longer than half of the original text". Whereas second is - "text summarization as the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)".

Text Summarization helps users manage the vast amount of information available, by condensing documents' content and extracting the most relevant facts or topics included in them. Although Text summarization started in the late fifty [2], text summarization has experienced a great development in recent years, and a wide range of techniques and paradigms have been proposed to tackle this research field. However, to produce a summary automatically is very challenging issues such as redundancy, temporal dimension, co reference or sentence ordering, to name a few, have to be taken into consideration especially when summarizing a set of documents (multi-document summarization), thus making this field even more difficult.

II. RELATED WORK

Before Text summarization is depending on to the user requirement. There are three factors which are considered for text summarization give by [4].

- Input factors: language, type of text (xml, database etc.), genre, and etc.
- Purpose factor: depend on to the user requirement.
- Output factors: focus on the style and coverage, and normally derive from purpose factors.

Various taxonomies are available for text summarization like generic and query-focused summaries which are commonly known as user focused or topic focused summaries. Common methodology for text summarization is abstractive or extractive.

Most of the summarization system work is based on extraction of sentences from the original text. In the sentence based extraction technique sentences are score first then select on the basis of the score. Various scoring method is used to scoring the sentences. Some is based on the word scoring and some methods are based on the structure of the document like sentence position, sentence length etc. Word scoring based methods are word count [2], key phrases [3] etc.

FEATURE EXTRACTION METHODS

For generating best summary we need to select the sentences which cover more content related to the text. For selecting sentences scoring the sentence based on the different feature. Sentences have different features which are following:

1. Word Count

In the word scoring method count the frequency of each word in the document and for each sentence add the

frequency of all word. According to this method, more frequently the word occur, higher its score and important for the text. [2] 1st used word counts to generate the summary of documents to find the relevant sentences.

2. TF/IDF

TF / IDF [5] approach is based on the term frequency. The idea behind the TF/IDF is that frequent terms in a document are important only if they are not very frequent in the whole collection. Different methods are used for calculating the TF/IDF score for each word.

When a set of documents is given in advance, our system counts the term frequency (tf) and the document frequency (df) for each word w, then calculates the TF/IDF score as

$$\left(\frac{TF}{IDF}\right)(w) = \left(\frac{tf}{1+tf}\right)\log\left(\frac{DN}{df}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where DN is the number of given documents.

3. Upper Case

Upper Case [6] method assigns the higher score to the words that contain one or more upper case letters. These words may be a noun, a place and any other important thing. So these are the important part of the text and contain important matter. For calculating the score of each word, formula used:

$$CPWTW(j) = \frac{NCPW(j)}{NTW(j)} \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Here

CPWTW = Ratio of total number of capital word present in the sentence to the total number of word present in the sentence.

NCPW = Number of capital word present in the sentence

NTW = Total number of word present in the sentence

$$UCS = \frac{CPWTW(j)}{MAX(CPWTW(j))} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Here, UCS = Upper case score

4. Cue-Phrases

In general, the sentences started by "in conclusion", "our investigation" or "the aim of this paper" may be good indicators of relevant information [6]. Sentences which contain these words are assigned to the higher score.

5. Thematic Feature

Thematic feature is based on thematic words. Thematic words are the most frequently occurring words in the document. The top n frequent words are considered as thematic words. The score for this feature is calculated by the following formula:

$$Thematic\ Feature = \frac{number\ \ of\ thematic\ \ words\ \ in\ sentence}{maximum\ \ \ number\ \ of\ thematic\ \ words} \quad (4)$$

6. Numeric Value

Sentences are score on the basis of presence of numerical data [7]. Numeric values indicate mostly time, price, date, address etc. contain the important information of the document. So the sentences which contain numeric values have higher score then other. Formula for calculating the score of sentence based on numerical value:

$$NS = \frac{NNW(j)}{NTW(j)} \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Here

NS = Score based on numerical value

NNW = Number of numerical word in sentence

NTW = Total number of word in sentence

7. Word Co-occurrence

Word co-occurrence is measure the chance of two terms form a text alongside each other in a certain order. To implement this method using the n-gram, this is the contiguous sequence of n-term in the text. It gives higher score to the sentence there co-occurrence word appears.

8. Sentence Position

Generally the sentence which is occur in the starting or in the ending of the document is consider to be more important than the others. The sentence which is occurring in starting contains the theme of the document and the ending sentences contain the conclusion of document. According to [7], the 1st sentences of a paragraph are the most important. They rank a paragraph sentence according to its position in the paragraph and consider maximum positions of 5. For instance, the 1st sentence in a paragraph has a score value of 5/5, the second sentence has a score 4/5, and so on.

Here we use threshold value, define how many sentences are selected in beginning and at the end. The feature weight of these sentences is:

SP = 1

For remaining sentences, weight is computed as follow:

$$SP = Cos\left((CP - MinV) * \left(\frac{Max\theta - Min\theta}{MaxV - MinV}\right)\right) \ldots\ldots\ldots\ldots (6)$$

Where:

TRSH = Threshold Value

MinV = NS * TRSH (Minimum Value of Sentence)

MaxV = NS * (1 - TRSH) (Maximum Value of Sentence)

NS = Number of sentences in document

Min θ = Minimum Angle (Minimum Angle=0)

Max θ = Maximum Angle (Maximum Angle=180)

CP = Current Position of sentence

9. Sentence Length

Sentence length is also a feature for scoring the sentence. Sentences which are too sorted or too long [8] are not considering as optimal selection in the summary. Sentence length is the number of words in the sentence. On the basis of the sentence length is scoring the sentences. According to the [5], find the certain length (CL) of the sentence if sentence is sorter then certain length then panelize the sentence.

$$score_{len}(si) = Li \quad (if\ Li > CL) \ldots\ldots\ldots\ldots\ldots (7)$$
$$Li - CL \ \ (otherwise)$$

10. Sentence Similarity

Sentence similarity is the vocabulary overlap between the two sentences. If more words are similar between two sentences then two sentences are closed to each other. For finding the similarity score, the number of similar words between two sentences is divided by the longest sentence length.

Aggregate score

Aggregate score is the summation of the similarity from each sentence.

Bushy path

Bushy path is the number of connecting link to the sentence.

11. Sentence resemble to Title

Title contains set of words that represents gist of the document. In this feature find the similarity between the sentence and title of the document. Sentences which are more similar to the title are considered into the summary.

In the next section we study the different algorithms and on the basis of these algorithms we find the summary and calculate the precision, recall, and f-measure.

III. PROPOSED WORK

Implementation of algorithms

Algo1 (word count) :- there is three step 1st remove all stop word; 2nd count the number of each word from text; and 3rd for each sentence add up the word frequency score of each word for sentence;

Algo2 (tf/idf):- this algorithm is divide into three step: 1st remove all stop word; 2nd calculate the tf/idf score for each word from formula (1); 3rd for each sentence add up the word frequency score of each word for sentence;

Algo3 (Upper case):- it divide into: 1st remove all stop word; 2nd count the number of word with capital letter in each sentence; 3rd used the formula (2) and (3) for finding the upper cases core.

Algo4 (thematic feature):- it divides into: 1st count the frequency of each word in the document and select n top word as a thematic words; then use the above formula (4) for finding the score.

Algo5 (numerical value):- this algorithm use regular expressions to verify if some numerical data is present. Here 1st find how many numerical words are present in the sentence and divided by total number of word in the sentence.

Algo6 (word co-occurrence):- it divide into: 1st compute n-gram measure for n=2; 2nd for each sentence, add up the n-gram score of each word in a sentence;

Algo7 (sentence position):- for finding score based on the sentence position apply the formula given in above formula (6)

Algo8 (sentence length):- it divides into: 1st find the certain length (CL); 2nd calculate the score for each sentence and panelized the sentence which are lesser then the CL.

For the algo9 and algo10, 1st finding similarity score, the number of similar words between two sentences is divided by the longest sentence length. So threw this create a graph in this sentence is node, nodes are connected through the link and on the links weight is assign which is similarity score.

Algo9 (aggregate score):- In the aggregate score for each sentence add up the weight of all connected link to that sentence.

Algo10 (bushy path):- In the bushy path score for each sentence count all the link connected to that sentence.

IV. RESULTS

In table show the evaluation results after comparing the automated generate summary with the human generated summary. Precision, Recall, F-measure are used to measure the relevance of the summary. For evaluation various methods is used. Here I used ROUGE- L method.

TABLE I. EVALUATION RESULTS

#	Feature	Precision	Recall	F-measure
1	Word frequency	0.3830	0.7078	0.4944
2	TF/IDF	0.6536	0.5824	0.6120
3	Uppercase	0.4858	0.6943	0.5637

#	Feature	Precision	Recall	F-measure
4	Aggregate score	0.3539	0.6377	0.4532
5	Thematic feature	0.3839	0.7377	0.5023
6	Numerical value	0.5118	0.7327	0.5940
7	Word co-occurrence n=2	0.3932	0.7195	0.5059
8	Length	0.3828	0.7684	0.5084
9	Position	0.4849	0.6829	0.5623
10	Bushy path	0.3698	0.6369	0.4663

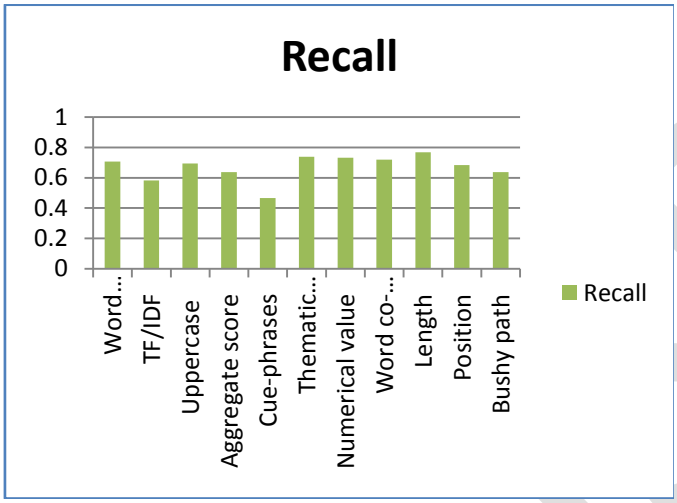

Fig:1 Recall chart

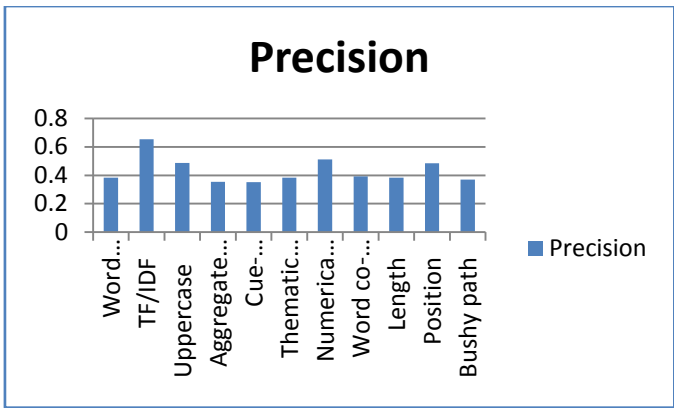

Fig:2 Precision chart

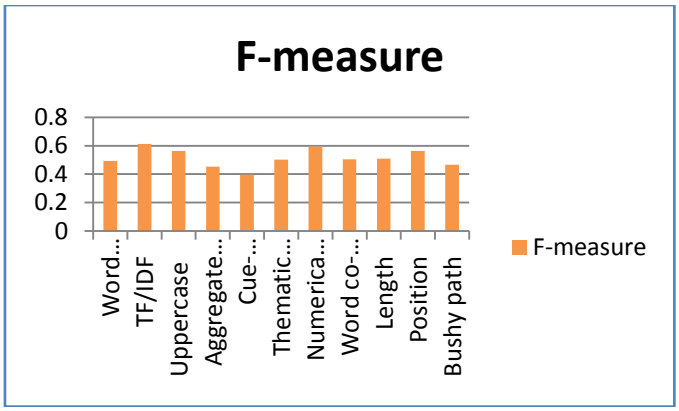

Fig:3 F-measure chart

CONCLUSION

In this paper we have study different type of summary generation methods. Methods depend on the sentence features and on the skeleton of the document. In some case for example if document don't have any title then can't generate the summary on the basis of title. So no single feature can generate the desired summary. For generating the best summary need to combine the feature and generate the summary. Algorithms used for summary generation is depend on to the field from which document belong. Performance enhanced using the basic search method to remove the redundant sentences from the input text. It will reduce the summary text size as well as the time required to generate the text summary if we apply it before the text to summary conversion algorithm (as per result shown).

REFERENCES

[1] Lin, Chin-Yew (2004). Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens (Ed.), Text summarization branches out: Proceedings of the ACL-04 workshop (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics.

[2] Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, 2(2), 159-165

[3] Edmundson, H.P. (1969). New method in automatic extracting. Journal ACM, 16(2), 264-285.

[4] Spärck Jones K (1999). Automatic summarizing: factors and directions. In: Advances in automatic text summarization. pp 1–14.

[5] Satoshi, Chikashi Nobata., Satoshi, Sekine., Murata, Masaki., Uchimoto, Kiyotaka., Utiyama, Masao. & Isahara, Hitoshi. (2001). Keihanna human info-communication. Sentence extraction system assembling multiple evidence. In Proceedings 2nd NTCIR workshop (pp. 319–324)

[6] Prasad, Rajesh Shardanand, Uplavikar, Nitish Milind, Wakhare, Sanket Shantilalsa, Jain, Vishal, Yedke & Tejas Avinash (2012). Feature based text summarization.International Journal of Advances in Computing and Information Researches, 1.

[7] Fattah, Mohamed Abdel, & Ren, Fuji (2009). Ga, mr, ffnn, pnn and gmm based models for automatic text summarization. Computer Speech and Language, 23 (1), 126–144.

[8] Abuobieda, A., Salim, N., Albaham, A. T., Osman, A. H., & Kumar, Y. J. (2012). Text summarization features selection method using pseudo genetic-based model.In International conference on information retrieval knowledge management (pp.193–197).