International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 1 Jan 2022, pp: 20-25www.ijaem.net ISSN: 2395-5252

Text Summarization Using Fuzzy Logic and Sentence Ranking

Md. Rashed hasan[1], Prof. Dr. Md. Al Mamun[2]
Rajshahi University of Engineering & Technology

Submitted: 01-01-2022 Revised: 05-01-2022 Accepted: 10-01-2022

ABSTRACT – Due to an increasing information on the internet, it has become difficult to identify the most useful and necessary information in a text document and draw conclusions. Text summarization aims to concise larger text documents. It helps to precise summaries for long text documents without losing the main theme of text documents. Automatic text summarization has two approaches: i) abstractive text summarization and ii) extractive text summarization. In extractive method, important sentences or paragraphs are extracted and rejoin them to get the summary of the source content. In this paper, an extractive text summarization method is designed for large English text document. The proposed solution uses fuzzy logic for measuring the relevance of information and sentence ranking for choosing the most relevant sentences for summary generation. Experimental results show successful generation of the summary of a document having long text.

Keywords – text summarization, extractive text summarization, fuzzy logic, sentence ranking.

I. INTRODUCTION

In modern time, a large amount of information is generated everyday on the internet. Earlier, long text had been summarized by humans. But, in recent time, it is difficult for the human beings to cope up with the large amount of information. To alleviate this issue, meaningful information extraction from huge information repositories is much needed. Automatic text summarization is the way of solving the problem, which helps identify useful and meaningful information from a document or a set of related documents. It has reduced the burden of human works in summarization. It compresses the huge information and saves time for reading large document. It helps researchers for analyzing information with less effort. The baseline paper on automatic summarization published by Luhn of IBM in 1958 opened the curtain of research in this field [1].

Text summarization methods are divided into two categories: extractive and abstractive. Text summarization is accomplished using extractive techniques, which pick sentences from documents based on a set of parameters. By removing redundancies and clarifying the sentence contest, abstractive strategies aim to enhance the consistency of sentences. The most common method for extractive text summarization is sentence scoring. As a result, extractive summarization entails assigning a saliency metric to certain document unit (e.g. sentences, paragraphs) and extracting those with the highest scores to use in the description[2][3]. Single document and multiple documentscan be summarized using text summarization techniques.H. P. Luhn was the first to make an automatic text summarization system. It was grounded on the frequency of terms [4]. In 1958, he stated significance of words based of their frequency measures. According to him, words which have a medium frequency i.e. which neither occur much frequently nor less frequently are important. He also removed the stop-words from the text. In 1958, Baxendale proposed salient features based on the sentence position [5]. He stated that the beginning 7% of the document and the last sentences have most of the information. In 1969, Edmundson suggested new method of automatic text summarization which included two fresh features in addition to the position and term which were pragmatic words (cue words), title and heading words [6]. In 2005, Evans and Klavans proposed a new method for summarizing the clusters of documents on the same events based on text similarity [7]. English and Arabic language documents were used. In 2015, S. A. Babar and P. D. Patil proposed an approach to improve the performance of text summarization using Singular value Decomposition and Fuzzy inference system

[8]. . In 2017, an approach for auto text summarization was developed by H. A. Chopade and M. Narvekar using Deep network and Fuzzy logic which provided significant increase in the accuracy of the summary [9]. . In 2018, Nikhil S. Shirwandkar, Dr. Samidha Kulkarni proposed a method where RBM is used as an unsupervised learning algorithm along with fuzzy logic for improving the accuracy of the summary [10]. In 2019, J.N. Madhuri, Ganesh Kumar R proposed extractive based text summarization by using statistical novel approach based on the sentences ranking [11].

In this paper,we use fuzzy logic for evaluating the relevance of information and sentence ranking for choosing the most relevant sentences for summary preparation. We perform experiments on different documents to generate the summaries and claim that our summary outputs are competitive to the human summarizer.

II. PROPOSED METHODOLOGY

We provide our proposed text summarization approach in Fig. 1. First, the single document text is converted to lower case representation; on which we perform sentence segmentation and then tokenization. After that, we remove the stop words and punctuation marks and extract features. Then, we apply fuzzy logic on extracted features and rank the sentences. Finally, summary is generated. We discuss the insights of each block in details as follows:

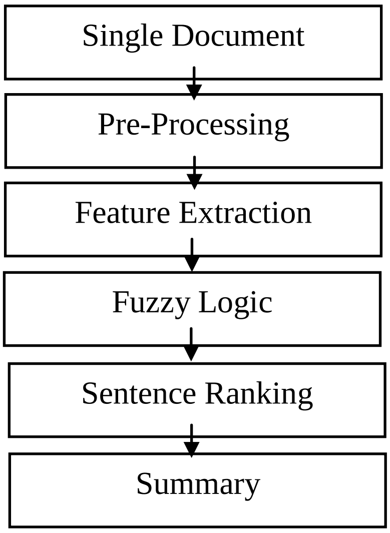


Fig. 1. Proposed method for text summarization

A. Single Document: A single text document is used as input in this process. It is the first stage of this process. Text documents are in English language.

B. Pre-processing: The imported text is passing through the pre-processing stage. In Convert Lower Case stage, the whole text is converted into lower case letter. In Sentence Segmentation, each sentence is extracted from the whole text and stored in an array according to the sentence positions. In Tokenization, words are extracted from each sentence. In Stop Word and Punctuation Removal, meaningless words such as a, an, the, and, or, punctuations etc. are removed.

C. Feature Extraction: After completing the pre-processing stage, features are extracted from the processed text. We use the following features in this work:

- Word Frequency: It is calculated using the following formulae:

$$\textbf{WF of } S_i = \frac{\mathbf{wf1+wf2+\cdots\cdots\cdots+wfn}}{\max(\mathbf{wf1,wf2,\cdots\cdots\cdots,wfn})} (1)$$

Where,**WF of S$_i$**is word frequency of **i**th sentence, **wf$_i$** is relative word frequency of word **i** in **S$_i$**, and **n** is the total number of words in **S$_i$** and **S$_i$**is any **i**th sentence of the document.

- Sentence Length: It is used to avoid too short and too long sentences. It is calculated as:

$$\mathbf{sentencelength}_i = \frac{\mathbf{wordsinsentence}_i}{\text{wordsinlargestsentence}} (2)$$

where, **sentence length$_i$** is the **i**th sentence length of the document.

- Sentence Position: In [5] author states that the first and last sentence of the document is always important and has maximum information. Position feature is calculated using:

$$\text{Sent. pos.} = \begin{cases} 1, & \text{if first or last sentence} \\ \frac{N \sim P}{P}, & \text{if others} \end{cases} \quad (3)$$

Where, Sent. pos. is the score of sentence positioning, N is sentence position, and P is total number of sentences.

- Numerical Token: It is the ratio between total number of numeric number and total words in a sentence.

$$\textbf{Numerical Token}_i = \frac{\textbf{totalnumericnumber}_i}{\textbf{totalwords}_i} (4)$$

Where, **Numerical Token$_i$** is numerical token score of **i**th sentence.

D. Fuzzy Logic: Four features are associated with each sentence. In each sentence, degrees of membership in the [0, 1] interval for four features are calculated using the triangular membership function.0 means unimportant and 1 means important. It will help remove too short and too long meaningless information from summarization.

Fuzzy triangular membership function helps finding the importance of features those are associated with each sentence. It follows:

$$\boldsymbol{\mu}_{\mathbf{Ai}}(x) = \begin{cases} 0, & \text{if} x < a \text{or} x > b \\ \frac{\mathbf{x-a}}{\mathbf{b-a}}, & \text{if others} \end{cases} \quad (5)$$

Where, $\boldsymbol{\mu}_{\mathbf{Ai}}: x \to [0,1]$is a membership function for feature A of ith sentence on the universe of discourse x,ais lower limit for removing too short meaningless information,b is upper limit for removing too long meaningless information. Here, a=0 and b=1.

E. Sentence Ranking: After completing the fuzzy logic operation, an average of the four features membership score for each sentence has been calculated for finding the sentence score. It follows:

$$\textbf{Sent. Sc}_i = \frac{\mu_{\mathbf{Ai}}(\mathbf{x}) + \mu_{\mathbf{Bi}}(\mathbf{x}) + \mu_{\mathbf{Ci}}(\mathbf{x}) + \mu_{\mathbf{Di}}(\mathbf{x})}{4} (6)$$

Where, **Sent. Sc$_i$** means **i**th sentence score,$\boldsymbol{\mu}_{\mathbf{Ai}}(x)$,$\boldsymbol{\mu}_{\mathbf{Bi}}(x)$,$\boldsymbol{\mu}_{\mathbf{Ci}}(x)$,$\boldsymbol{\mu}_{\mathbf{Di}}(x)$ are four features membership score for **i**th sentence.

According to Sentence Score, the whole sentences are sorted in descending order.

F. Summary: From sorted sentences, top 30% is taken for final summary from the source document.

Top ranked sentences are selected according to the proposed methodology and rejoin them. Finally, summarized text is generated.

III. EVALUATION AND RESULTS

We use Intel core i5 2.50GHz processor, 4GB RAM to perform the experiment. Python 3.6.10 with Jupyter Notebook as a platform, is used as a programming language.

In this experiment, 5 documents are tested where each document has successfully generated of the given document. TABLE I shows thesummary of the 1st source document using the proposed methodology.

Recall Oriented Understudy for Gisting Evaluation (ROUGE) is used for evaluating our method by comparing the number of overlaps between the automatically and manually generated summaries. ROUGE is an automatic summarization evaluation method proposed by Lin [12]. ROUGE scores are reported separately for each n-gram. The most commonly reported F1 scores are ROUGE-1, ROUGE-2 and ROUGE-L [13]. The parameters for performance evaluation are precision as p, recall as r and f measure. We compare system generated summary with human generated summary and smmry.com [14] generated summary using ROUGE metric. The result is shown in TABLE II.

According to the result shown in TABLE II, average precision, recall and f measure for system generated summary with respect to human generated summary are 0.52, 0.46 and 0.43 respectively, and smmry.com generated summary with respect to human generated summary are 0.31, 0.43 and 0.35 respectively. So it is obvious that the proposed method provides better results than smmry.com generated summary.

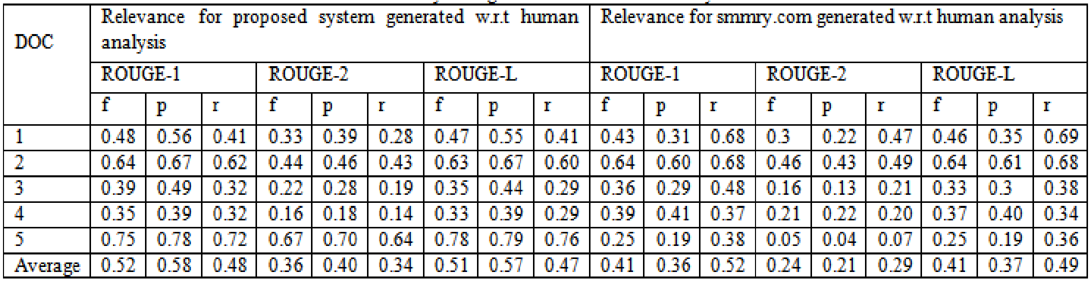
TABLE I. System generated summary along with input text.

Input Document			System Generated Summarized Document		
Text	No. of Sentences	No. of Word	Text	No. of Sentences	No. of Word
A good start at a Machine Learning	8	153	Basically, applications learn from previous	3	43

definition is that it is a core sub-area of Artificial Intelligence (AI). ML applications learn from experience (well data) like humans without direct programming. When exposed to new data, these applications learn, grow, change, and develop by themselves. In other words, with Machine Learning, computers find insightful information without being told where to look. Instead, they do this by leveraging algorithms that learn from data in an iterative process. While the concept of Machine Learning(ML) has been around for a long time (think of the WWII Enigma Machine), the ability to automate the application of complex mathematical calculations to Big Data has been gaining momentum over the last several years. At a high level, Machine Learning is the ability to adapt to new data independently and through iterations. Basically, applications learn from previous computations and transactions and use “pattern			computations and transactions and use “pattern recognition” to produce reliable and informed results.When exposed to new data, these applications learn, grow, change, and develop by themselves.ML applications learn from experience (well data) like humans without direct programming.		

recognition" to produce reliable and informed results.					

TABLE II. Performance comparison between results obtained by system generated w.r.t human analysis and smmry.com generated w.r.t human analysis.



DOC	Relevance for proposed system generated w.r.t human analysis									Relevance for smmry.com generated w.r.t human analysis								
	ROUGE-1			ROUGE-2			ROUGE-L			ROUGE-1			ROUGE-2			ROUGE-L		
	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r	f	p	r
1	0.48	0.56	0.41	0.33	0.39	0.28	0.47	0.55	0.41	0.43	0.31	0.68	0.3	0.22	0.47	0.46	0.35	0.69
2	0.64	0.67	0.62	0.44	0.46	0.43	0.63	0.67	0.60	0.64	0.60	0.68	0.46	0.43	0.49	0.64	0.61	0.68
3	0.39	0.49	0.32	0.22	0.28	0.19	0.35	0.44	0.29	0.36	0.29	0.48	0.16	0.13	0.21	0.33	0.3	0.38
4	0.35	0.39	0.32	0.16	0.18	0.14	0.33	0.39	0.29	0.39	0.41	0.37	0.21	0.22	0.20	0.37	0.40	0.34
5	0.75	0.78	0.72	0.67	0.70	0.64	0.78	0.79	0.76	0.25	0.19	0.38	0.05	0.04	0.07	0.25	0.19	0.36
Average	0.52	0.58	0.48	0.36	0.40	0.34	0.51	0.57	0.47	0.41	0.36	0.52	0.24	0.21	0.29	0.41	0.37	0.49

IV. CONCLUSION AND FUTURE WORK

In extractive text summarization method, meaningful sentences are selected from given document. It is a complex task to find meaningful information from a large text. In this paper, we propose an extractive text summarization method using fuzzy logic and sentence rankingand compare with an online based summary generator. Finally, the proposed method provides better results than the online based summary generator.

In future, we plan to work for Bengali multiple document summarization. The proposed method can be extended for hybrid text summarization where extractive and abstractive text summarization are combined for better performance. Multi document and multiple languages can be summarized.

REFERENCES

[1] Jones, Karen Sparck. "Automatic summarising: The state of the art." Information Processing & Management 43.6 (2007): 1449-1481.

[2] S. Akter, A. S. Asa, M. P. Uddin, M. D. Hossain, S. K. Roy and M. I. Afjal, "An extractive text summarization technique for Bengali document(s) using K-means clustering algorithm," 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Dhaka, Bangladesh, 2017, pp. 1-6, doi: 10.1109/ICIVPR.2017.7890883.

[3] P. B. Tumpa, S. Yeasmin, A. M. Nitu, M. P. Uddin, M. I. Afjal and M. A. A. Mamun, "An Improved Extractive Summarization Technique for Bengali Text(s)," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2),Rajshahi, 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465609.

[4] H. P. Luhn, "The automatic creation of literature abstracts," IBM Journal of Research Development, vol. 2, no. 2, pp. 159-165, 1958.

[5] P. Baxendale, "Machine-made index for technical literature - an experiment." IBM Journal of Research Development, vol. 2, no. 4, pp. 354- 361, 1958.

[6] H. P. Edmundson, "New methods in automatic extracting," Journal of the ACM, vol. 16, no. 2, pp. 264-285, 1969.

[7] D. K. Evans, "Similarity-based multilingual multidocument summarization," Technical Report CUCS-014- 05, Columbia University, 2005.

[8] S. A. Babar, and P. D. Patil, "Improving performance of text summarization," Procedia Computer Science, vol. 46, pp. 354-363, 2015.

[9] H. A. Chopade and M. Narvekar, "Hybrid auto text summarization using deep neural network and fuzzy logic system," 2017 International Conference on Inventive Computing and Informatics (ICICI), COIMBATORE, India, pp. 52-56, 2017.

[10] N. S. Shirwandkar and S. Kulkarni, "Extractive Text Summarization Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697465.

[11] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 International

Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-3, doi: 10.1109/IconDSC.2019.8817040.

[12] Lin , C hin -Y e w . "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. (2004).

[13] S. Ren and K. Guo, "Text Summarization Model of Combining Global Gated Unit and Copy Mechanism," 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2019, pp. 390-393, doi: 10.1109/ICSESS47205.2019.9040794.

[14] Online summary generator. From https://smmry.com