Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019)
IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8

ANALYSING FUZZY BASED APPROACH FOR EXTRACTIVE TEXT SUMMARIZATION

Aakanksha Sharaff, Amit Siddharth Khaire, Dimple Sharma
asharaff.cs@nitrr.ac.in[1], amitkhaire017@gmail.com[2], dimpleryp@gmail.com[3]
National Institute of Technology, Raipur
Raipur, India

***Abstract*— In today's era of information, there is gigantic amount of data available from various sources. Not only does the enormous volumes pose problems, searching of required information becomes a very difficult task. It is the need of the hour to have smaller but significant representation of large, bulky pieces of information in order to obtain the desired details. Text summarization is the process of condensing text documents into shorter and accurate representation conveying the meaning of the text precisely. It has found applications in numerous fields. In this paper, a process for extractive text summarization using fuzzy logic has been discussed meticulously. It takes various properties into account for identifying the most significant sentences for the formation of summary from a given text. The model proposed in this paper has been tested on BBC News Summary dataset and the results have been compared using the ROUGE measures. The obtained results indicate that the proposed model shows enhanced performance with improved f-measure values.**

***Keywords*—Fuzzy logic, extractive text summarization, membership function.**

I. INTRODUCTION

Text summarization has been gaining immense importance ever since information explosion. The amount of information is increasing rapidly every day. Not only is it difficult to manage these large volumes of data but also very tedious to look for the desired information. Today abundant information is available for each and every subject. It is not feasible to go through these gigantic magnitude of information. This is why many data mining techniques have come into play, text summarization being a significant one.

Text summarization refers to the process of shortening or reducing long textual documents. This is done in order to create summaries that bring out the crucial points delivering the meaning of the documents correctly. The summaries reduce reading time significantly as selection of the documents having the desired information becomes much easier. Text summarization is widely used in patent research, legal contract analysis, social media marketing, academic research, newsletters, financial research, etc.

Text summarization can be broadly classified into two types, namely extractive summarization and abstractive summarization. In extractive summarization, the most meaningful sentences are extracted from a given document and arranged. No modifications are made to the extracted sentences. In abstractive summarization, the most important sentences from the document are paraphrased. In the context of the proposed model, extractive text summarization has been performed. The most significant sentences are selected on the basis of the features of text (position, frequency of words in sentences) and application of fuzzy logic in the form of triangular membership function. Triangular membership function has been used for the conversion of input space point to a membership value lying between 0 and 1. The model proposed in the paper has been used over the BBC News Summary dataset. Results have been compared using ROUGE measures and tabulated.

Having covered the introduction in this section, paper is structured to discuss the past approaches for text summarization along with fuzzy logic in the next section. This is followed by the third section in which details of the proposed model have been elaborated. In the fourth section, performance of the model on various parameters have been analyzed with tabulated results. The last section concludes the work and enlists points for future work.

II. BACKGROUND

A. Fuzzy Logic

The term fuzzy means vague or unclear. Situations where state of being true or false cannot be determined, fuzzy logic comes to rescue. Fuzzy logic is an approach resembling human reasoning. In other words, fuzzy logic is a way of normalizing human ability of unspecific reasoning. Rather than producing binary true or false values (1 or 0), it is involved in computation of real numbers lying between 0 and 1 which indicate "degrees of truth". These are known as truth values of variables. The value 1.0 is for representing absolute truth while 0.0 is for

absolute false state. However, in fuzzy system, any intermediate value represents partial true or partial false states, i.e., all truths are basically partial [1]. Fuzzy logic has found a number of applications in various fields including automotive, business, defense, finance, marine, medical, psychology, etc.

B. Past approaches

In this section, some of the previous works with respect to text summarization have been discussed. Kiani-B and Akbarzadeh-T proposed a novel technique for text summarization using combination of Genetic Algorithm and Genetic Programming for optimizing rule sets and membership functions of fuzzy systems [2]. Patil and Kulkarni explored text summarization using fuzzy logic extraction approach [3]. Suanmali, Salim and Binwahlan focused on sentence extraction. Assigning numerical ranks to sentences helped in identifying the important sentences in document which were further used for making summary. Combination of fuzzy logic, genetic algorithm and semantic role labeling were the techniques used for the same. Semantic role labeling was used for capturing sentence contents for summarization [4]. Hannah et al. provided a sentence scoring and categorizing approach using fuzzy rules [5]. Dasari and Rao introduced a novel text summarization approach based on knowledge-corpus. It is capable of delivering significant performance irrespective of the size of content [6]. Text summarization is broadly divided into extraction and abstraction for which information retrieval comes into play. Gupta et al. came up with a new hybrid similarity measure to enhance information retrieval performance based on similarity between query and document. The hybrid similarity measure comprises of Jaccard and Cosine similarity measures which are combined by the means of fuzzy logic [7]. Alhabashneh et al. classified document relevance using fuzzy information retrieval system. It incorporated construction and integration of three relevance profiles followed by capture of relevance feedback and user queries [8]. Abbasi-ghalehtaki et al. presented a new text summarization model built on evolutionary algorithms, fuzzy logic system and cellular learning automata [9]. Goularte et al. utilized fuzzy rules for extracting features for automatic text assessment [10]. Megala et al. gave extraction based text summarization using ten feature extraction methods [11]. Kamya et al. produced a model for question paper generation with specified difficulty level from given database of questions using fuzzy logic [12]. Kyoomarsi et al. provided a new text summarization model using word-net and fuzzy logic. It extracts significant and relevant sentences by means of inference and fuzzy measures [13].

C. Previous Model

Goularte et al. have made use of fuzzy rules for text summarization for automatic text assessment. This assessment is to be utilized in the fields of Virtual Learning Environment and Computer Aided Assessment. For extraction of significant sentences, features including frequency and position of words in sentences have been used. The model is implemented in various stages. During the pre-processing phase, the tasks pertaining to tokenization, stop words removal and stemming are performed. This is followed by determining weightage of each and every sentence. Fuzzy analysis is executed with the help of bell membership function. The model was applied on Portuguese dataset and the obtained results were compared using measures of ROUGE. The model performed better than other methods in which fuzzy logic was not used.

III. PROPOSED MODEL

A. Corpus Selection

The dataset used in the model is the BBC News Summary dataset. It is one of the commonly used datasets for text summarization. It comprises of 417 news articles from BBC from 2004 to 2005 along with their respective standard summaries for the purpose of comparison. The news articles are on the subjects of business, entertainment, politics, sports and technology.

B. Pre-processing

Firstly, the input text is segmented, followed by stop-words removal. Next, indexing of words in the input text is performed which is utilized during computation of frequency and position parameters. The data-structure used for indexing is hash map due to its ease of use in storing and retrieving the information, be it word (string) or index number (integer). This hash map stores the words as keys with values assigned as numbers representing the index of the words in the text. The stop words are removed in the beginning, before applying indexing on the input text so that only useful words participate in the process of indexing. Figure 1 below shows the procedures involved in the proposed model.

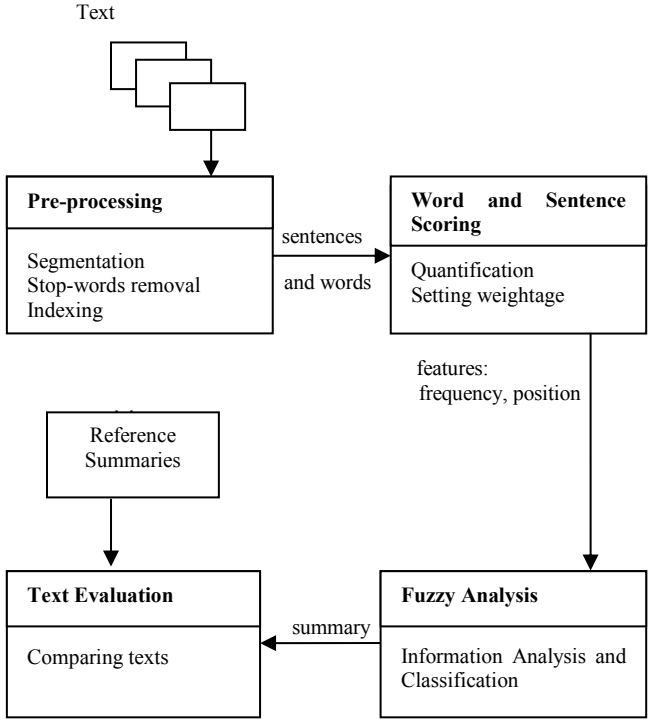

Fig. 1. Workflow diagram of Extractive Text Summarization using Fuzzy Logic

C. Word and Sentence Scoring

For calculation of word and sentence scoring, weightage of words and sentences are calculated. These values help in determining the importance of sentences, further enabling to make the decisions accordingly whether the sentence can be chosen as a part of summary or not. The below equations (Eq. 1, Eq. 2) are used to calculate the frequency and position parameters respectively for each sentence.

$$w_{fre}(j) = \frac{\sum_{i=0}^{m}\left(\frac{f_{ij}}{F_i}\right)}{x_j}, j = 0, \dots, n \quad (1)$$

$$w_{pos}(j) = \begin{cases} 1 - \frac{(p_j - 1)}{n}, if\ S_j \in B1 \\ \frac{p_j}{n}, if\ S_j \in B2 \end{cases}, j = 0, \dots, n \quad (2)$$

In the first equation (Eq. 1), $w_{fre}(j)$ represents weightage of the sentence j based on frequency. f_{ij} is frequency of the word i in sentence j. F_i denotes the frequency of word i in the text document. x_j represents the total number of words in sentence j. m is the number of words in the text document and n is the number of sentences in the text document. In the second equation (Eq. 2), $w_{pos}(j)$ represents weightage of sentence j based on position. p_j represents the position of sentence j. S_j represents sentence j. B1 and B2 represent the two sets where B1 has half set of sentences and B2 has another half of sentence. This equation (Eq. 2) has also been used in one of the previous works [10].

D. Fuzzy Analysis

In this step, information analysis and classification are done. The calculated values of the parameters obtained in last step undergo several operations. Mean values for all the sentences are calculated by adding $w_{fre}(j)$ and $w_{pos}(j)$, followed by dividing the obtained sum by 2. The mean values are then fed to the triangular membership function which is given in the third equation (Eq. 3). This membership function is used to map values from input space to a membership value between 0 and 1. The comparison of values of parameters a, b, c (a = 0.30, b = 0.60, c = 1.0) with the mean value (represented by x) indicates which formula is to be used for mapping. The mean values which represent weightage of sentences are used as input to the membership function on which fuzzy rule is applied. The final values obtained after applying the membership function are stored and then sorted in decreasing order. The top sentences with higher values are selected and included in the summary. Number of sentences to be selected is dependent on the requirement which may be specified in the form of percentage of total sentences in the text document or directly as the number of sentences required in the summary.

$$T(x, a, b, c) = \begin{cases} 0, x \le a \\ \frac{x - a}{b - a}, a < x \le b \\ \frac{c - x}{c - b}, b < x < c \\ 0, c \le x \end{cases} \quad (3)$$

E. Text Evaluation

During the text evaluation step, the generated summaries are compared against the given summaries from the dataset. The results compared using the measures of ROUGE have been tabulated below.

IV. RESULTS

The proposed model shows improvement in performance. The tables below are the comparison of results obtained by the proposed model with that of the previous model, discussed in section (II-C), on the basis of precision, recall and f-measure values respectively. Bell membership function has been used in the previous model whereas triangular membership function has been used in the proposed model.

$$Precision = \frac{number\ of\ words\ co-occur\ in\ system\ generated\ and\ reference\ summary}{number\ of\ words\ in\ reference\ summary}$$

$$Recall = \frac{number\ of\ words\ co-occur\ in\ system\ generated\ and\ reference\ summary}{number\ of\ words\ in\ system\ generated\ summary}$$

Table-I Performance comparison (20%)

Percentage of text summary	20%		
Parameter	Precision	Recall	F-measure
Bell Membership	0.417	0.398	0.406
Triangular Membership	0.410	**0.768**	**0.535**

Table-I gives performance comparison when the required summary is 20% of the total number of sentences in the given text document. Recall and f-measure values are indicating improvement in performance via proposed model. The precision value is not better than the previous model but is almost equal. As the recall value is higher for the proposed model, the f-measure value is also greater. Thus, there is an improvement in the overall performance. This reflects that better results can be obtained by using the proposed model even for generating small sized summary.

.

Table-II Performance comparison (30%)

Percentage of text summary	30%		
Parameter	Precision	Recall	F-measure
Bell Membership	0.366	0.496	0.421
Triangular Membership	**0.755**	0.384	**0.452**

Table-II shows performance comparison when the required text summary is 30% of the given text document size. The results obtained by the proposed model show better precision and f-measure values. Therefore, the proposed model outperforms the previous model in generating moderate length summary. Recall is lesser than the previous model but is still comparable. Since precision is greater, it resulted in improved f-measure.

Table-III Performance comparison (40%)

Percentage of text Summary	40%		
Parameter	Precision	Recall	F-measure
Bell Membership	0.315	0.556	0.402
Triangular Membership	**0.777**	0.334	**0.468**

Table-III displays performance comparison when the percentage of the required text summary is 40% of the length of given text document. The proposed model achieves better precision and f-measure values, thus, showing improvement in performance. Hence, the proposed model surpasses the previous model and is capable of producing better summaries of greater lengths from given text documents.

The comparison of above obtained results can also be seen graphically for precision in figure 2, recall in figure 3 and f-measure in figure 4.

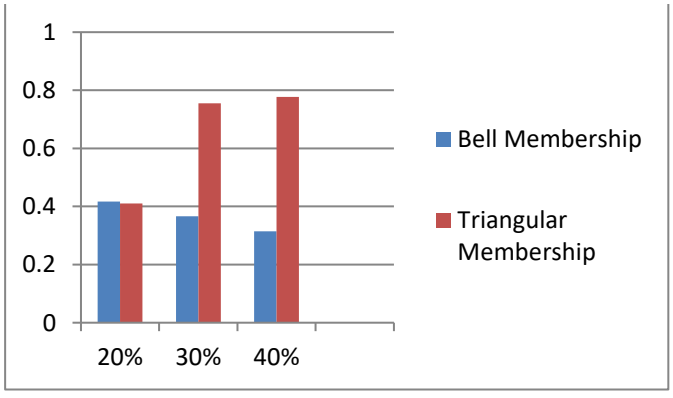

Fig. 2. Performance Comparison (Precision)

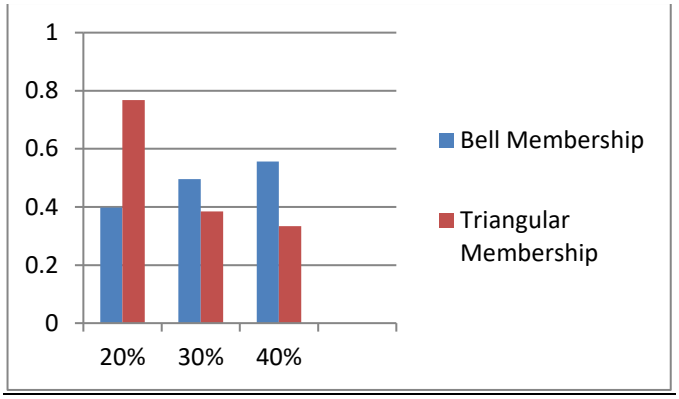

Fig. 3. Performance Comparison (Recall)

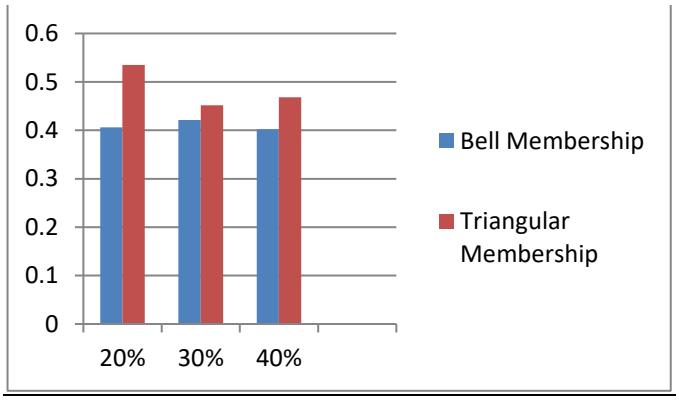

Fig. 4. Performance Comparison (F-measure)

V. Conclusion and Future Work

Text summarization has become one of the crucial data mining techniques since information explosion. The ever growing volumes of information is making it difficult to manage and moreover even more difficult to get the desired information. It is not practical to go through all such information sources. Text summarization produces small, shortened version of the entire textual data that underlines the main concepts covered in it. It provides the essence or gist of the entire document in a few words. This helps in determining the relevance of a text with desired information, saving enormous amounts of time and efforts. In this paper, the model proposed for extractive text summarization utilizes fuzzy logic using triangular membership function. This membership function performs mapping of input space points to membership values lying between 0 and 1. The results obtained using the proposed model show improvement in performance for various parameters including precision, recall and f-measure. Hence, the proposed model is capable of producing better summaries of any length be it small, medium or large. The sentences having the higher membership values are extracted to form summary of given text document. The features used for sentence extraction involve position and frequency of words in sentences. These features can be incorporated with some other features to enhance the process of sentence selection and yield improved summaries.

REFERENCES

[1] Ross, T. J. "Fuzzy logic with engineering applications" (Vol. 2). New York: Wiley, (2004).

[2] Kiani-B & Akbarzadeh-T, M. R. "Automatic text summarization using hybrid fuzzy ga-gp". IEEE International Conference on Fuzzy Systems (pp. 977-983), (2006).

[3] Patil, P. D., & Kulkarni, N. J. "Text summarization using fuzzy logic." International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1, (2014).

[4] Suanmali, L., Salim, N., & Binwahlan, M. S. "Fuzzy genetic semantic based text summarization." IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (pp. 1184-1191), (2011).

[5] Hannah, M. E., Geetha, T. V., & Mukherjee, S. "Automatic extractive text summarization based on fuzzy logic: a sentence oriented approach." International Conference on Swarm, Evolutionary, and Memetic Computing (pp. 530-538), (2011).

[6] Dasari, D. B., & Rao, K. V. G. "Single document text summarization by Knowledge-Corpus." IEEE International Conference in MOOC, Innovation and Technology in Education (MITE) (pp. 134-138), (2013)

[7] Gupta, Y., Saxena, A. K., Saini, A., & Sharan. "Development of hybrid similarity measure using fuzzy logic for performance improvement of information retrieval system." International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1-5), (2014).

[8] Alhabashneh, O.Iqbal, R.Doctor & Amin "Adaptive information retrieval system based on fuzzy profiling." IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8). IEEE, (2014).

[9] Abbasi-ghalehtaki, R., Khotanlou, H., & Esmaeilpour, M. "Fuzzy evolutionary cellular learning automata model for text summarization." Swarm and Evolutionary Computation, 30, 11-26, (2016).

[10] Goularte, F. B., Nassar, S. M., Fileto, R., & Saggion, H. "A text summarization method based on fuzzy rules and applicable to automated assessment.' Expert Systems with Applications. 115, 264-275, (2019).

[11] Megala, S. S., Kavitha, A., & Marimuthu, A. "Enriching text summarization using fuzzy logic." International Journal of Computer Science and Information Technologies, 5(1), 863-867, (2014).

[12] Kamya, S., Sachdeva, M., Dhaliwal, N., & Singh, S. "Fuzzy logic based intelligent question paper generator." IEEE International Advance Computing Conference (IACC) (pp. 1179-1183), (2014).

[13] Kyoomarsi, F., Khosravi, H., Eslami, E., & Davoudi, M. "Extraction-based text summarization using fuzzy analysis." Iranian Journal of Fuzzy Systems, 7(3), 15-32, (2010).