Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)
IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

Hybrid Auto Text Summarization Using Deep Neural Network And Fuzzy Logic System

Heena A. Chopade[1]
D.J.Sanghvi College of Engineering and Technology, Mumbai University,
Mumbai, India
heenachopade5@gmail.com

Dr.Meera Narvekar[2]
D.J.Sanghvi College of Engineering and Technology, Mumbai University,
Mumbai, India
narvekar.meera@gmail.com

Abstract- Amount of text for a particular topic is increasing at an exponential rate each day, causing availability of bulk text data, some of which will be relevant to our search while some data may be irrelevant. In such case we need to summarize the text document, so as to understand the key elements of the document. A Hybrid Automatic Summarizer using soft computing techniques namely, Fuzzy logic system and deep neural system is proposed. The summary is based on extractive summarization technique. Restricted boltzman Machine (RBM) is used in the deep neural network and fuzzy rule base on the sentences for feature extraction using a sentence matrix. Evaluation of hybridized systems gave better summary as compared to the individual systems.

Keywords- Fuzzy Logic System; Restricted Boltzman Machine(RBM); Deep neural network(DNN); fuzzy membership function

I. INTRODUCTION

The amount of information is increasing every day. Thus finding relevant data becomes hectic and time consuming, more over not all the data is relevant to the user's topic of interest. In order to find relevant data for user's search and to save time is it necessary to have a small summary of the documents. Summary made by humans is time consuming and tedious. Thus there is a need for automatically summarizing the text document to save time and to get quick results. Automatic Summarization can be defined as the art of condensing large text documents into few lines of summary, giving important information [8].

Text summarization can be classified on the basis of different criteria. [5] Summary based on construction, extractive summaries pick important sentences from the document based on certain conditions and display it to the user as they are[3] , abstractive summaries gives reconstructed summary which is not exactly the same as the original document. On the number of sources for the summary, single document summary which is produced from single document, multi-document summary which is obtained from multiple documents. Trigger based summaries can be of two types, generic based summary, present the summary in concise manner as the main topic of the data and query based summarization gives summary as an answer to the query given by the user. Also based on the important details of the summary it can be classified as indicative which gives information to the user whether the document should be read or not, and the informative summary provide all the relevant information to represent the original document.

The most challenging problem of auto summarization is to provide information that is relevant to user's topic of interest. To accomplish this a hybrid approach is proposed using fuzzy logic and Deep neural network (DNN) along with a provision to give seed word to the summarizer so that it summarizer the documents which include the seed word, so that the user gets all the relevant information related to the search thus saving a lot of time. The idea of using fuzzy logic is to determine the relevance of sentence with reference to the seed provided by the user. DNN provides the semantic space for the sentence so that the sentences with meaningful semantics can be extracted thereby reducing the data redundancy [7].

The remainder of this paper is organized as follows. In section II, the review of related work on text summarization is done. In section III, summarization using fuzzy logic and DNN is explained. In section IV, the model of hybrid auto text summarization is proposed. Finally we conclude with our work in section V.

II. RELATED WORK

In [4], important sentence extraction is done using fuzzy rules and fuzzy set for selecting sentences based on their features. Here gaussian membership function and IF-Then logic to summarize document. In [1], feature extraction for Wikipedia articles is done using ten different feature scores which is fed to the neural network and the neural network returns single value signifying the importance of the sentence in the summary. In [9], a hybrid approach for auto text summarization is proposed using genetic algorithm to optimize the rule sets of membership function of fuzzy system, which makes use of bell shaped membership function. After the features have been recognized it is fed to the fuzzy system that accurately finds the important sentences and lastly the genetic algorithm optimizing the important sentences. In [10], summary of multiple documents using query oriented, unsupervised deep learning. Here once the feature vectors are calculated, the restricted boltzman machine is used to filter out non relevant words and to discover the key words then the

reconstruction validation intends to reconstruct the data distribution by fine tuning the whole deep architecture globally and finally the dynamic programming is used to maximize the length of summary with length constraint.

III .SUMMARIZATION USING FUZZY SYSTEM AND DNN

A. Summarization using fuzzy system

[2] For facing the uncertainty the use of modeling processes like fuzzy logic is used to understand the semantic similarity between the words. Fuzzy logic system uses the fuzzy rules and fuzzy membership function for implicating the selection of sentences. The main role of the fuzzy logic system is to use proper fuzzy membership function and fuzzy rule set so as to overcome the uncertainty and properly correlate the syntactic and semantic relations between the sentences.

B. Summarization using Deep Neural Network (DNN)

Theoretically the complex structures are difficult to be extracted by the shallow architectures. [11] Deep architectures consist of multiple layers, where each layer trains data on distinctive set of features based on previous layer's output. DNN can perform automatic feature extraction unlike most traditional learning algorithms which can be done using Restricted Boltzman machine (RBM).

C. Proposed hybrid auto text summarization

This paper proposes hybridization of fuzzy logic system and deep neural network, thereby trying to discover the intrinsic features of sentences and increase the degree of importance and correlation, so as to identify the important sentences to create summarization. Figure (1) shows block diagram of proposed hybrid auto text summarization.

The framework for hybrid auto text summarization is as given below:

- The text document to be summarized is given to the summarizer. The summarizer first does feature extraction of sentences.
- Four features are used which gives extraction of sentences which are important for summary retrieval. The four features used are:
    1. Title similarity- Sentence is considered to be important if it matches the title of the text document.

$$F1 = \frac{s \cap t}{t} \quad (1)$$

Where,
S = set of words of sentences.
t = set of words of title.

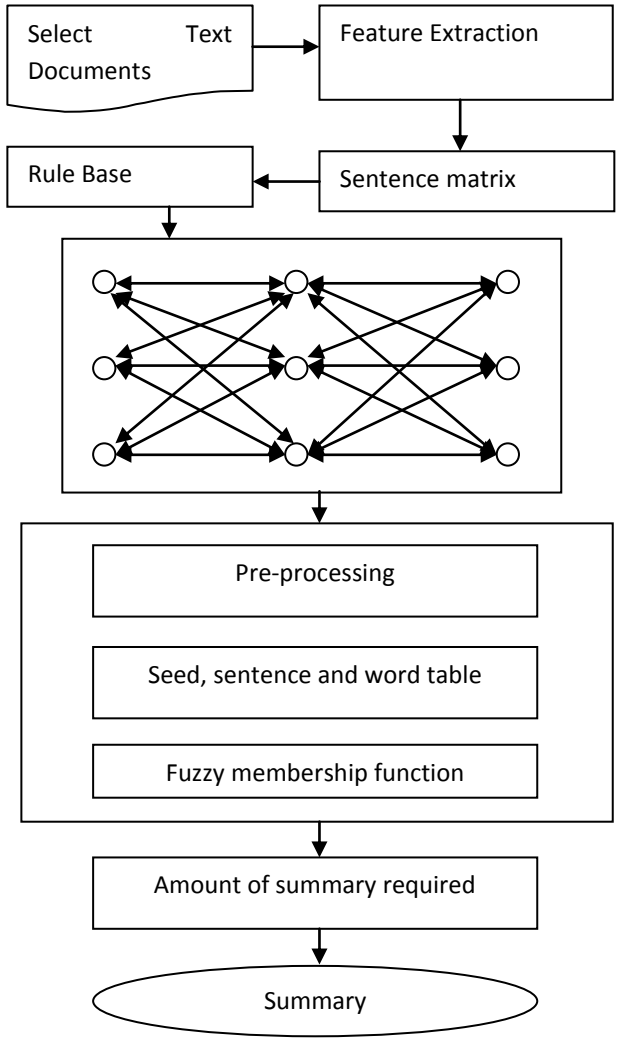

Figure 1: Proposed model of Hybrid Auto Text Summarizer

$s \cap t$ = common words in the sentence and title of the document.

2. Term weight- Sentence is important if the word it contains appears frequently in the text document.

$$F2 = tf * idf \quad (2)$$

Where,
tf = term frequency of the word in the document.
idf = inverse document frequency, refers to the word being rare or repeated in the sentence[6].

3. Named entities- Sentence is important if it is a proper noun, since named entities usually contain key information.

F3 = number of named entities (proper noun) in the sentences.

4. Numerical data- Sentence is important if it contains numerical values, statistics etc.

F4 (S) = 1: if S has numerical value
0: otherwise

- The features are applied to the sentences based on the priority of the features considering the document type.
  [1] If the document contains more numerical data then high priority is given to the numerical data feature.
  [2] If the whole document contains only words then priority is given to title similarity, term weight and named entities and so on.
- Now a feature sentence matrix is formed. The sentence matrix gives us feature vector matrix score of the sentences.
- Once the feature vector matrix score is extracted, a fuzzy rule base is applied to the sentences. Three rules are applied to the sentences.
  1. (If the scores are high) then (the sentences are important.)
  2. (If the scores are medium) then (the sentences are average.)
  3. (If the scores are low) then (the sentences are unimportant.)
- These sentences are then given to the Restricted Boltzman machine(RBM) where the sentences are trained based on the training data. Here we use one input layer, two hidden layers and one output layer. The sentences from the matrix are fed to the input layer where the sentence score with a bias value are given to the first hidden layer. The same procedure is repeated for the second hidden layer with different bias value. The final output is given to the output layer, where, we get more refined extraction of sentences.
- After the training of sentences, the sentences are pre-processed so that user defined query can be given for extraction of sentences based on particular topic. Pre-processing includes stop word filtering, stemming and POS tagging.
  1. Stop word filtering, here the words such as “a”, “an”, “the”, “and”, “by”, full stop, comma, semicolon, question mark etc are removed as stop words as they are not as important as other words in the sentence.
  2. Stemming, the basic idea here is to bring the words to their root words by using singular form of words, removing “ing”, “ed” etc from the words.
  3. Parts of speech tagging is used to classify the words of the data into their parts of speech category they belong (noun, verb, adverb, adjective).
- Once the preprocessing is done we enter the seed word for the summarization i.e query based processing. A seed table is created where the seed word is entered with its random priority value. The priority value is used for fuzzy membership calculation of the seed query with the sentences.
- For all the sentences a sentence table is created which contains all the words of the sentences, the number of words in the sentence and its associated rank which is filled while the fuzzy membership score is calculated.
- A separate table for words of the sentence is created for calculating the frequency of words in the entire document.
- Once all the above tables are created, the fuzzy membership value is calculated with the help of following formula-

$$\text{Rank}(S_i) = \sum_{i=1}^{n} frequency(w_{ij}) + \sum_{i=1}^{n} membership\_with\_seed(w_{ij}) \quad (3)$$

Where,
n = total number of sentences in the sentence table.
S_i= i^{th} sentence.
W_{ij}= j^{th} word of i^{th} sentence.
Frequency ( ) finds the frequency of its argument.
Membership _with_seed ( ) returns the membership value of the argument with the seed defined by the user.

- Lastly ranking of sentences is done to obtain the final summary of document and the amount of sentences to be extracted is given by the user.
- Figure (1) shows the block diagram of proposed hybrid auto text summarizer.

IV. EXPERIMENTAL RESULTS

The objective of the system is to provide a summary of any given text document. A hybrid system is proposed to build the summarizer using soft computing approaches. Hybridization of deep neural network with fuzzy logic system was done. The hybrid auto text summarizer was tested on a ROUGE toolkit with the help of 25 text documents. The testing was compared based on the individual neural system and fuzzy system summary results v/s hybrid system summary results of recall, precision and f-measure of the system. The precision, recall and f-

measure of the summary are calculated using the ROUGE toolkit.

$$\text{Recall} = \frac{tp}{tp+fn} \quad (4)$$

$$\text{Precision} = \frac{tp}{tp+fp} \quad (5)$$

$$\text{F-measure} = \frac{2*precision*recall}{precision+recall} \quad (6)$$

True positive (tp) which means the result obtained and actual result is true. False positive (fp) means the result obtained is true, but is actually false. False negative (fn) means the result obtained is false but is actually true. Recall gives the amount of summary that is retrieved correctly. Precision gives the fraction of summary that is correct. F-m easure gives the accuracy of the system. The testing results of hybrid system gave an increased score of 29% to 34% as compared to individual systems. The accuracy of the hybrid system was increased up to 31%. Thus giving an accuracy of 84.73%.

Neural			Fuzzy			Hybrid		
Recall	Precision	F-Measure	Recall	Precision	F-Measure	Recall	Precision	F-Measure
0.08	1	0.16	0.22	0.088	0.035	1	0.91	0.95
0.08	1	0.15	0.18	1	0.31	1	0.88	0.93
0.18	0.87	0.3	0.14	1	0.25	1	0.6	0.75
0.03	0.91	0.07	0.14	0.82	0.25	1	0.91	0.95
0.11	1	0.21	0.11	1	0.21	1	0.6	0.75

TABLE 1: Scores of individual v/s hybrid system

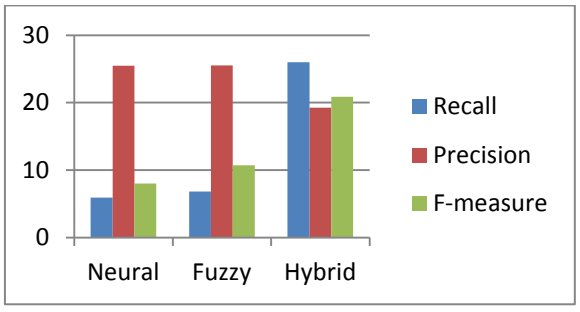

Figure 2: Performance comparison of Neural, Fuzzy and Hybrid systems.

From the above figure 2, the testing results of proposed system gave the best results in case of Recall and F-measure. The accuracy of the hybrid system is increased drastically. The recall and precision results show that the true positives retrieved are more as compared to the false positives. As a result the recall value is increased and the precision value is decreased. The decrease in the precision values shows that the amount of falsely obtained result is reduced.

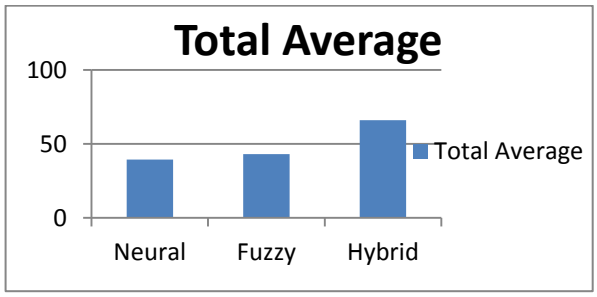

Figure 3: Performance comparison of overall average evaluation of individual v/s hybrid systems.

V. CONCLUSION

As the data is increasing at an exponential rate, Automatic Summary generator is essential in time bound situations and retrieval of accurate text document. A lot of research has been done on summarization techniques and most of them are extractive summarizers. A hybrid approach of soft computing techniques is proposed using deep neural network and fuzzy logic system. Here the training of sentences over a set of data and applying rules base over it which gives a human reasoning to the summary. Prioritizing the features is best for particular document types where numerical data is very important. We have also used user query based summary extraction, thereby mapping the user query with the sentences using membership function. The proposed system is found to give 84.73% accuracy in giving summary which is to an average 31% more than the individual summary generator.

The system scope can be increased by applying back propagation network to the deep neural network. Also a large set of rule base can be applied to the features.

REFERENCES

[1] Dharmendra Hingu, Deep Shah, Sandeep S Udmale, “Automatic Text Summarization of Wikipedia Articles”, IEEE, 2015

[2] Mehdi Jafari, Amir Shahab Shahabi, Jing Wang, Yongruri Qin, Xiaohui Tao, Mehdi Gheisari, “ Automatic Text Summarization using Fuzzy Inference”, IEEE, 2016

[3] Jyoti Yadav, Dr. Yogesh Kumar Meena, “ Use of Fuzzy Logic and wordNet for Improving Performance of Extractive Automatic Text Summarization”, IEEE, 2016

[4] Ladda Suanmali, Mohammed Salem Binwahlan, Naomie Salim, “Sentence Features Fusion for Text Summarization Using Fuzzy Logic”, International Conference on Hybrid Intelligent Systems, IEEE, 2009

[5] Saiyed Saziyabegum, Priti S. Sajja, “ Literature Review on Extractive Text Summarization Approaches”, International Journal of Computer Applications, 2016

[6] Shweta Karwa, Niladri Chatterjee, "Discrete Differential Evolution for Text Summarization", International Conference on Information Technology", 2014

[7] Chengwei Yao, Jianfen Shen, Gencai Chen, "Automatic Document Summarization via Deep Neural Networks", International Symposium on Computational Intelligence and Design, IEEE, 2015

[8] Nazreena Rahman, Bhogeshwar Borah, "A Survey on Existing Extractive Techniques for Query Based Text Summarization", International Symposium on Advanced Computing and Communication, IEEE, 2015

[9] Arman Kiani-B, M.R.Akbarzadeh-T, "Automatic Text Summarization Using: Hybrid Fuzzy GA-GP",IEEE, International Conference on Fuzzy Systems, 2006

[10] Yan Liu, Sheng-hua Zhong, Wenjie Li, "Query Oriented Multi-Document Summarization via Unsupervised Deep Learning", Proceedings of the Twenty-sixth AAAI Conference on Artificial Intelligence, 2012

[11] Jayashree R, Srikanta Murthy K, Basavaraj.S.Anami, "An Artificial Neural Netowrk Approach to Text Document Summarization in the Kannada Language", IEEE, 2013