A Survey On Hybrid Model Automatic Text Summarization

G. PadmaPriya
Assistant Professor
Department of CSE
K.S.R. College of Engineering
K.S.R. Kalvi Nagar
Tiruchengode
Tamilnadu, India

Dr. K. Duraiswamy
Dean / Academic
Department of CSE
K.S.Rangasamy College of Technology
K.S.R. Kalvi Nagar
Tiruchengode
Tamilnadu, India

T. Christi Jeyaseeli
Final Year ME/CSE
Department of CSE
K.S.R.College of Engineering
K.S.R. Kalvi Nagar,
Tiruchengode
Tamilnadu, India

Abstract

Automatic Text Summarization is a data compression process to exclude unnecessary details and include important information of a source text document in a shorter version. In this paper, we have presented the concept of four different techniques for automatic text summarization.

Keywords

MMI, PSO, Fuzzy Logic, Neural Networks, Text Summarization.

1. INTRODUCTION

The aim of text summarization is to generate high quality summary by extracting important sentences from the source document. Summary can be generated in two different forms: *(i) Abstractive* and *(ii) Extractive.*

An abstractive summary contains the main content of the original source document which are interpreted and presented in a shorter version.

An extractive summary is an extraction of important sentences from the source document and presented it in the same order as a summary.

Our hybrid model is an extractive based text summarization model. Four different techniques are involved in this model: MMI, PSO, Fuzzy Logic and Neural networks. MMI concentrates on filtering the similar sentences present in the document and selecting the most diverse sentences, PSO is used for scoring the sentence features and it identifies the most important sentence features and less important sentence features, Fuzzy Logic modifies the weight of the text features that are generated by PSO and the Neural network transforms the inputs into meaningful outputs. By integrating these four techniques, a good summary can be produced.

Rasim, Ramiz, Makrufa and Chingiz [1] described an MCMR text summarizer that produces a summary by deriving the significant sentences from the source document. It was done by directly determining the Key sentences in the source document and the vital information of the original document was covered in the summary. Also, it is suited for single and multi-document summarization.

Binwahlan, Salim and Ladda [2] says that the summary was generated by choosing obvious ideas “diversity” from the source document by means of three phases: Clustering, Binary tree and diversity based method. Clustering is used to group the similar sentences into different clusters, these clusters are represented in binary tree form and diversity based method selects the most important sentence from each cluster to be included in the summary.

According to Binwahlan, Salim and Ladda [3], summary was generated by combining two different techniques Swarm Intelligence and Fuzzy Logic. Swarm intelligence uses PSO algorithm to assign scores to sentences and Fuzzy Logic is used to modify the PSO generated score values.

Rasim, Ramiz, Makrufa and Chingiz [4] says that the summary was generated by deriving key sentences from the source document. It also concentrates on minimizing redundant information in the summary by using an Adaptive Differential Evolution algorithm.

According to Wei, Lim, Cheol and Ding [5], Fuzzy Evolutionary Optimization Modeling (FEOM) and its applications is used to categorization and

summarization. At first, FEOM is applied to clustering source documents and then applied to sentence clustering based summarization to choose the most vital sentence from each of the cluster to denote the comprehensive meaning of the document.

Sukanya, Petres and Tom [6] used an extended version of Tensor Term Importance (TTI) approach for summarizing text documents. The aim was to define the group of documents in a uniform form of term-sentence-document tensor and used a Higher-Order Singular Value Decomposition (HOSVD) to focus the important terms in each document. The main aim was to extract the important sentences and its related sentences to denote the comprehensive meaning of the document.

Jorge and Chalender [7] describes a hybrid sentence extraction summarization model to choose important sentences in source document by combining two methods. First method is a semantic analysis based, uses word senses and constructed by using lexical co-occurrences network. Second method is an integration of semantic and syntactical analysis to derive a set of relative sentences.

Rasim and Ramiz [8] describes a text summarizer that generates a summary based on sentence extraction method. The summarizer derives the sentences with highest sentence feature weight from the source document. Sentence feature weights are determined by genetic algorithms.

According to Arman and Akbarzadeh [9], summary was generated by integrating the features of Genetic Algorithm and Genetic Programming to optimize the set of rules and membership function of fuzzy inference system. Structural part is handled by Genetic Programming and Membership Functions are handled by Genetic Algorithm, to extract the important sentences in the source document by minimizing the co-occurrences of text.

Yan-Xiang, De-Xi, Dong-Hong and Chong [10] introduced a multi-document summarizer which focuses the process of summarization as an optimization problem. Various sentence features are involved in the process. TFS was used to calculate the term frequency and genetic algorithm was used to select the optimal summary.

According to Khosrow [11], summary was created by using Neural Networks in three phases: neural network training, feature fusion, and sentence selection. Neural network training identifies certain sentences that should be inserted in the summary. Feature fusion is used to generalize the important features of the summary sentences by means of discovering sentence patterns and fusing the features. Sentence selection is a modified neural network that filters the text and selects the top ranked sentences by controlling the sentence selection in respect to their importance.

2. MAXIMAL MARGINAL IMPORTANCE (MMI)

Maximal Marginal Importance is based on the derivation of the sentences that are considered as most important. MMI considers various features that are grouped together in a linear fashion to focus the sentence importance.

3. SENTENCE FEATURES

Each sentence in the source document has a set of features. The features are:

S.No	*Sentence Feature*
3.1	Sentence location feature
3.2	Positive keyword feature
3.3	Negative keyword feature
3.4	Sentence centrality feature
3.5	Sentence length feature
3.6	Document title feature
3.7	First sentence feature
3.8	Term weight feature
3.9	Sentence similarity feature

Feature 3.1, Sentence location feature denotes the position of the sentence within the source document. Feature 3.2, Positive keywords in the sentence those are generally added in the summary and feature 3.3, Negative keywords in the sentence those are not probably included in the summary. Feature 3.4, Sentence centrality defines the term overlap between one sentence and other sentences in the source document. Feature 3.5, Sentence length feature is used to exclude the shortest sentences. Sentence length is calculated as the number of terms in the current sentence over number of terms in the lengthy

sentence in the document. Feature 3.6, Document title feature identifies all the sentences those matches with the title of the document. Feature 3.7, First sentence feature considers the similarities between the first sentence and other sentences of the document since the first sentence of any document is treated as an important sentence. Feature 3.8, Term weight feature calculates the number of occurrences of a term A term with highest frequency is used to determine the sentence importance. Feature 3.9, Sentence similarity feature determines the similarities between the current sentence and the other sentences in the source document.

4. PARTICLE SWARM OPTIMIZATION (PSO)

PSO is a robust stochastic optimization technique based on the movement and intelligence of swarms. PSO applies the concept of social interaction to problem solving. It was developed in 1995 by James Kennedy (social-psychologist) and Russell Eberhart (electrical engineer). It uses a number of agents (particles) that constitute a swarm moving around in the search space looking for the best solution. Each particle is treated as a point in a N-dimensional space which adjusts its "flying" according to its own flying experience as well as the flying experience of other particles.

Each particle keeps track of its coordinates in the solution space which are associated with the best solution (fitness) that has achieved so far by that particle. This value is called personal best, ***pbest***.

Another best value that is tracked by the PSO is the best value obtained so far by any particle in the neighborhood of that particle. This value is called ***gbest***.

The basic concept of PSO lies in accelerating each particle toward its pbest and the gbest locations, with a random weighted acceleration at each time step as shown below (fig (a)):

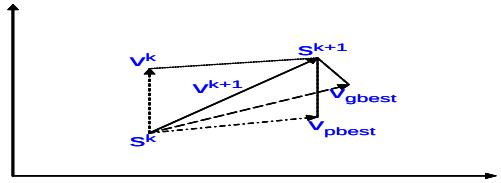

Fig (a): Concept of PSO algorithm towards selecting pbest and gbest values

s^k : current searching point.
s^{k+1}: modified searching point.
v^k : current velocity.
v^{k+1} : modified velocity.
v_{pbest} : velocity based on pbest.
v_{gbest} : velocity based on gbest

The binary PSO is used to determine and optimize the relative weight of each feature. Each particle position is defined as bit string as shown below (fig (b)):

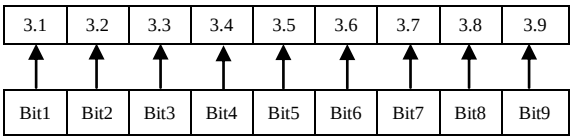

Fig (b) : Particle Position – Structure

Bit1 refers feature 3.1, bit2 refers feature 3.2, and so on. If the value of bit is 1 then the corresponding feature is selected and 0 refers the corresponding feature is not selected. By considering the velocity of the particle, each bit value is extracted from the sigmoid function. In each iteration, each particle selects a definite number of features and a summary is created based on the selected features. At the end of each iteration, nine evaluation values are present, since there are nine particles. Here, *pbest* contains the evaluation value of each summary and *gbest* contains the best evaluation value among the nine evaluation values. The value of *pbest* and *gbest* is updated in the successive iterations by comparing the current *pbest* value with its previous *pbest* value and the current g*best* value with its previous g*best* value. At the end of each run, the particle position with its gbest value will selected as vector for the best selected sentences of the source document.

4. FUZZY LOGIC

Fuzzy logic is used to calculate the score of each sentence by using fuzzy inference system.

The most commonly used fuzzy inference technique is the so-called Mamdani method. In 1975, Prof. **Ebrahim Mamdani** of London University built one of the first fuzzy systems. He applied a set of fuzzy rules supplied by experienced human operators.

4.1 Working Principle of Fuzzy Inference System

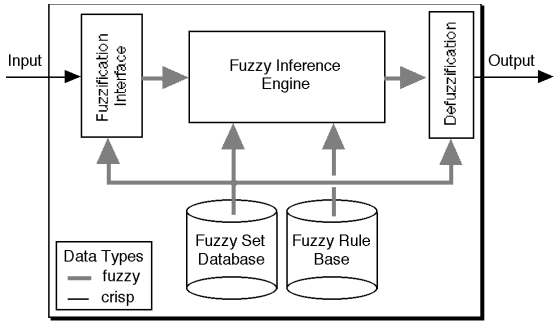

Fig (c): Fuzzy Inference System

Three phases involved in fuzzy inference system: Fuzzification, Fuzzy Inference Engine and Defuzzification. Fuzzification converts the crisp inputs (text feature values) to a linguistic variable using the membership functions stored in the fuzzy knowledge base.

Fuzzy Set Database (or dictionary) defining the MF (fuzzy values) used in the fuzzy rules and Fuzzy Rule Base contains a selection of fuzzy rules.

Fuzzy Inference Engine merges the facts from fuzzification with a set of IF..THEN rules to perform the fuzzy reasoning process.

Defuzzification transforms output fuzzy set into a single crisp value, which is the final score of the sentence.

5. ARTIFICIAL NEURAL NETWORKS

Warren McCulloch and Walter Pits were the two neurophysiologists who developed the first artificial neuron in 1943.

Neural network inspired by biological nervous systems, such as our brain. An artificial neural network is composed of many artificial neurons that are linked together according to specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs.

Neural networks are used for Classification (pattern recognition, feature extraction), Noise Reduction and Prediction.

5.1 Working Principle of Neural Networks

The output of a neuron is a function of the weighted sum of the inputs plus a bias. The function of the entire neuron network is simply the computation of the outputs of all the neurons.

$$y_{out}=f(\sum x_i w_i+Bias)$$

where, $x_1,x_2,\ldots x_i$ are inputs to neuron, $w_1,w_2,\ldots w_i$ are weights, 'Bias' is a function and y_{out} is the output.

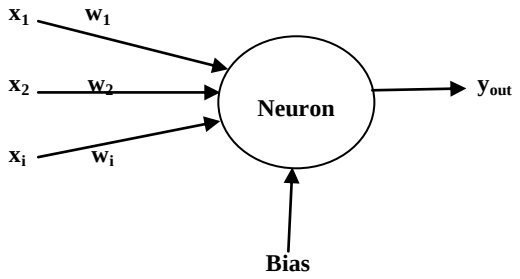

Fig (d): Neuron Network

Bias acts as a control function such that the output of a neuron in a neural network is between (0,1) or (-1,1).

5.2 An Artificial Neural Network

An artificial neural network is composed of many artificial neurons that are linked together according to specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs.

Fig (e) shows an Artificial Neural Network. Here, N_1, N_2 & N_3 are input neurons, N_5 & N_6 are hidden neurons and N_7 & N_8 are output neurons.

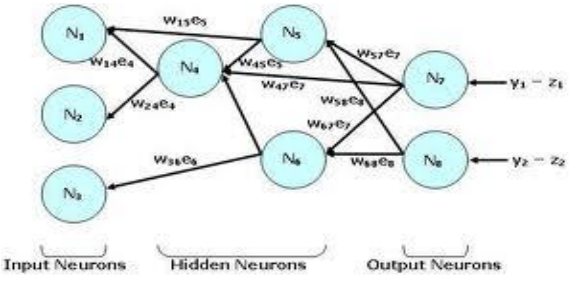

Fig (e): An Artificial Neural Network

6. CONCLUSION

In this paper we have presented the concept of four different text summarization techniques such as MMI, PSO, Fuzzy Logic and Neural Networks. By integrating these four techniques, we can get a information rich summary.

7. REFERENCES

1. Rasim M.Alguliev, Ramiz M.Aliguliyev, Makrufa S.Hajirahimova, Chingiz A.Mehdiyev, MCMR: Maximum Coverage and Minimum Redundant Text Summarization Model, *Expert Systems with Applications 38(2011) 14514-14522.*
2. Binwahlan.M.S., Salim.N., & Suanmali.L (2009a), MMI Diversity based text summarization, *IJCSS International Journal of Computer Science and Security, 3(1), 23-33.*
3. Binwahlan.M.S., Salim.N.,& Suanmali.L (2009d), Fuzzy Swarm based text summarization, *Journal of Computer Science, 5(5), 338-346.*
4. Rasim M.Alguliev, Ramiz M.Aliguliyev, Chingiz A.Mehdiyev, Sentence Selection for Generic Document Summarization using an Adaptive Differential Evolution Algorithm, *Swarm and Evolutionary Computation 1(2011) 213-222* .
5. Wei Song, Lim Cheon Choi, Soon Cheol Park, Xiao Feng Ding, Fuzzy Evolutionary Optimization Modeling and its applications to unsupervised categorization and extractive Summarization, *Expert Systems with Applications 38(2011).*
6. Sukanya Manna, Zoltan Petres and Tom Gedeon, Tensor Term Indexing: An application of HOSVD for Document Summarization, *ISCIII 2009, 4th International Symposium on Computational Intelligence and Intelligent Informatics, 21-25 October 2009 Egypt.*
7. Jorge Garcia Flores, Gael de Chalender, Syntactico-Semantic Analysis: a Hybrid Sentence Extraction Strategy for Automatic Summarization, *2008 Seventh Mexican International Conference on Artificial Intelligence.*
8. Rasim M.Alguliev, Ramiz M.Aliguliyev, Effective Summarization Method of Text Documents, *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence(WI'05).*
9. Arman Kiani, M.R.Akbarzadeh, Automatic Text Summarization Using Hybrid Fuzzy GA-GP, *2006 IEEE International Conference on Fuzzy Systems, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, July 16-21, 2006*
10. Yan-Xiang He, De-Xi Liu, Dong-Hong Ji, Chong Teng, MSBGA: A Multi-Document Summarization System Based on Genetic Algorithm, *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Daliban, 13-16 August 2006.*
11. Khosrow KaiKhah, Automatic Text Summarization with Neural Networks, *Second IEEE International Conference On Intelligent Systems, June 2004.*