ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

A Survey on Text Summarization for News Articles Using Modified Neural Networks

Shankari V. Gajul

Assistant Professor, Department of Computer Engineering, A. G. Patil Institute of Technology, Solapur, India

ABSTRACT: A machine learning approach has been proposed that uses neural networks to produce summaries of arbitrary length news articles. A neural network is trained on a corpus of articles. The trained neural network is used to classify whether a word is summary worthy or not. A neural network is trained to learn the relevant features of sentences that should be included in the summary of the article. The neural network is then modified to generalize and combine the relevant features apparent in summary sentences. A sentence is ranked depending on the number of summary worthy word. A summary is produced by taking highly ranked sentences. Finally, the modified neural network is used as a filter to summarize news articles.

KEYWORDS: Adaptive clustering, feature fusion, neural networks, pruning, text summarization.

I. INTRODUCTION

Text summarization has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age. There is an abundance of text material available on the Internet; however, usually the Internet provides more information than is needed. Therefore, a dual problem is encountered: searching for relevant documents through a vast number of documents available, and absorbing a large quantity of relevant information. Summarization is a useful tool for selecting relevant texts, and for extracting the key points of each text. Some articles such as academic papers have accompanying abstracts, which present their key points. However, news articles have no such accompanying summaries, and their titles are often not sufficient to convey their important points. Therefore, a summarization tool for news articles would be particularly useful, since for given news topic or event, there are a large number of available articles from the various news agencies and newspapers. Because news articles have a highly structured document form, important ideas can be obtained from the text simply by selecting sentences based on their attributes and locations in the article.

We propose a machine learning approach that uses artificial neural networks to produce summaries of arbitrary length of news articles. A neural network is trained on a corpus of articles. The neural network is then modified, through feature fusion, to produce a summary of highly ranked sentences the article. Through feature fusion, the network discovers the importance (and unimportance) of various features used to determine the summary-worthiness of each sentence.

Paper is organized as follows. Section II describes related work, architecture of neural network with working of neural network. Problem statement is defined in Section III. After problem statement the methodology part which gives the detailed concepts with diagrams that is given in Section IV. Section V presents outcome. Finally, Section VI presents conclusion.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

II. RELATED WORK

A query based document summarizer based on similarity of sentences and word frequency is presented in [11].The summarizer uses Vector Space Model for finding similar sentences to the query and Sum Focus to find word frequency. In this paper they propose a query based summarizer which being based on grouping similar sentences and word frequency removes redundancy. In the proposed system first the query is processed and the summarizer collects required documents & finally produces summary.

Kamal Sarkar presented an approach to Sentence Clustering-based Summarization of Multiple Text Documents in [9] [10]. Here three important factors considered are: (1) clustering sentences (2) cluster ordering (3) selection of representative sentences from the clusters. For the sentence clustering the similarity histogram based incremental clustering method is used. This clustering approach is fully unsupervised & is an incremental dynamic method of building the sentence clusters. The importance of a cluster is measured based on the number of important words it contains. After ordering the clusters in decreasing order of their importance, top n clusters are selected. One representative sentence is selected from each cluster and included in to the summary. Selection of sentences is continued until a predefined summary size is reached.

The pervasion of huge amount of information through the World Wide Web (WWW) has created a growing need for the development of techniques for discovering, accessing, and sharing knowledge. The keyphrases help readers rapidly understand, organize, access, and share information of a document. Keyphrases are the phrases consisting of one or more significant words. Keyphrases can be incorporated in the search results as subject metadata to facilitate information search on the web [12]. A list of keyphrases associated with a document may serve as indicative summary or document metadata, which helps readers in searching relevant information.

Wang has proposed in a neural network based approach to keyphrase extraction, where keyphrase extraction has been viewed as a crisp binary classification task. They train a neural network to classify whether a phrase is keyphrase or not. This model is not suitable when the number of phrases classified by the classifier as positive is less than the desired number of keyphrases, K. To overcome this problem, we think that keyphrase extraction is a ranking problem rather than a classification problem. One good solution to this problem is to train a neural network to rank the candidate phrases. Designing such a neural network requires the keyphrases in the training data to be ranked manually. A Neural Network based approach to keyphrase extraction has been presented in that exploits traditional term frequency, inverted document frequency and position (binary) features. The neural network has been trained to classify a candidate phrase as keyphrase or not [4] [5].

Duda presented a training algorithm for cluster prototypes while the LVQ approach modifies the position of the prototypes through a training procedure on all input data. Both these algorithms require the number of prototypes to be defined and good initialization of the prototype values before training [6].

Chen and Kaminsky and Strumillo describes the pruning methods, a large network is trained and then the least useful nodes or weights are removed. In crosswise propagation, linearly dependent hidden units are detected using an orthogonal projection method, and then removed. The contribution of the removed neuron is approximated by the remaining hidden units. In, the hidden units are iteratively removed and the remaining weights are adjusted to maintain the original input–output behavior. A weight adjustment is carried out by solving the system of linear equations using the conjugate gradient algorithm in the least squares sense. Unfortunately, if one large network is trained and pruned, the resulting error versus number of hidden units curve is not minimal for smaller networks. In other words, it is possible, though unlikely, for each hidden unit to be equally useful after training [13].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Phung and Bouzerdoum proposed the PyraNet, a neural network specifically developed for image recognition tasks. The PyraNet is based on CNN and on the pyramid images concepts (Gonzalez & Woods, 2010). The PyraNet learning algorithm tunes the nonlinear processing of each pyramid level to solve specific recognition problems. The PyraNet integrates the feature extraction and pattern classification steps in the same structure. Moreover, the PyraNet also maintains the spatial topology of the input image and presents a simple connection scheme with lower computational and memory costs than in other neural networks as demonstrated in Phung and Bouzerdoum (2007) [8].

Architecture of Neural network:

a) Concept of Neural Network:

The basic architecture consists of three types of layers: input, hidden, and output layers. In Neural networks, the signal flows from input to output units, strictly in fed-forward direction. There is a connection between each neuron of input layer to each input of hidden layer and each neuron of the hidden layer to each neuron of the output layer. A 'weight' is associated with each connection of the neurons.

The following fig 1 shows the three layers of neural network.

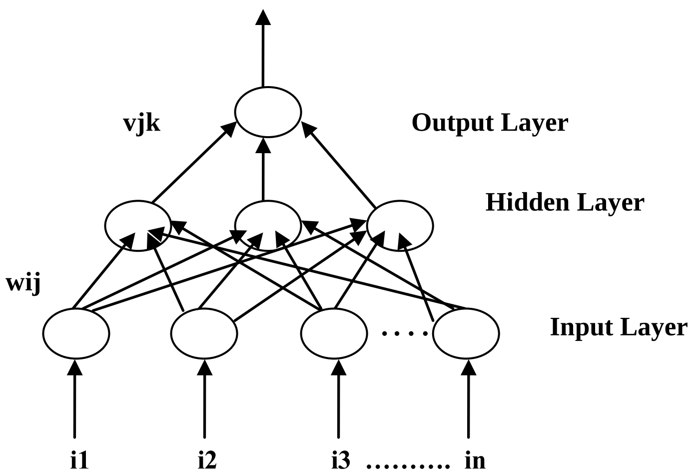

Figure 1: The Neural Network

A training sample, *X* = (i1, i2, i3…in) is fed to the input layer. Weighted connections exist between each layer, where, 'wij' denotes the input layer to hidden layer weight and 'vjk' denotes the hidden layer to output layer weight.

a) Working of Neural Network:

A vector of predictor variable values is presented to the input layer. In text summarization, this input vector is the feature vector, which is a vector of values of features characterizing the words of a text.

The value from each input neuron is multiplied by a weight and arrives at a neuron in the hidden layer, and the resulting weighted values are added together producing a combined value at a hidden node. The weighted sum is then fed into a transfer function (usually a sigmoid Input Layer Hidden Layer Output Layer function), which outputs a value.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The outputs from the hidden layer are distributed to the output layer. Arriving at a node (a neuron) in the output layer, the value from each hidden layer neuron is again multiplied by a weight, and the resulting weighted values are added together producing a combined value at an output node.

III. PROBLEM STATEMENT

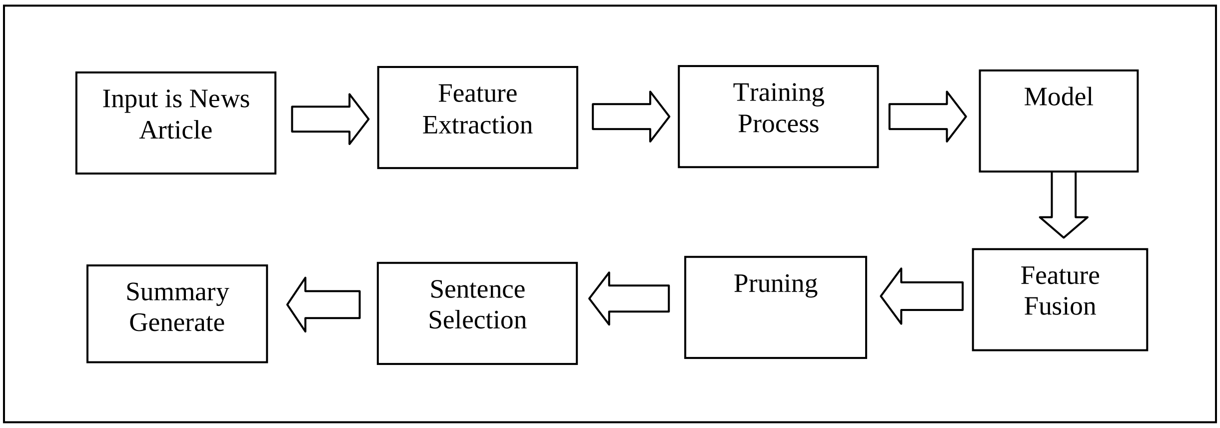

Figure 2: Architecture of neural network

The objective of the system is neural network is trained to learn the relevant characteristics of sentences that are included in summary of the article then neural network is modified to generalize and combine the relevant characteristics of sentences and finally the modified neural network is used as a filter to summarize news articles.

The above fig 2 show the flow of steps with news article is giving as input to the system. The detailed description of steps is discussed in further sections.

IV. METHODOLOGY

The following points are involved in the methodology:

1. *Input:*

The input for the system is a news articles. In this, we use pruning and clustering strategies to generate relevant characteristics of sentences that included in summary.

2. *Techniques:*

There are three phases in our process:

- **i.** *Neural network training*- The first step involves training a neural network to recognize the type of sentences that should be included in the summary.
- **ii.** *Feature fusion*- The second step, feature fusion, prunes the neural network and collapses the hidden layer unit activations into discrete values with frequencies.
- **iii.** *Sentence selection*- The third step, sentence selection, uses the modified neural network to filter the text and to select only the highly ranked sentences.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

i. Neural Network Training-

The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included [2] [10].

The fig 3 shows the neural network after training process which is taking the input as type of sentences that is consists of seven input-layer neurons, six hidden-layer neurons, and one output-layer neuron.

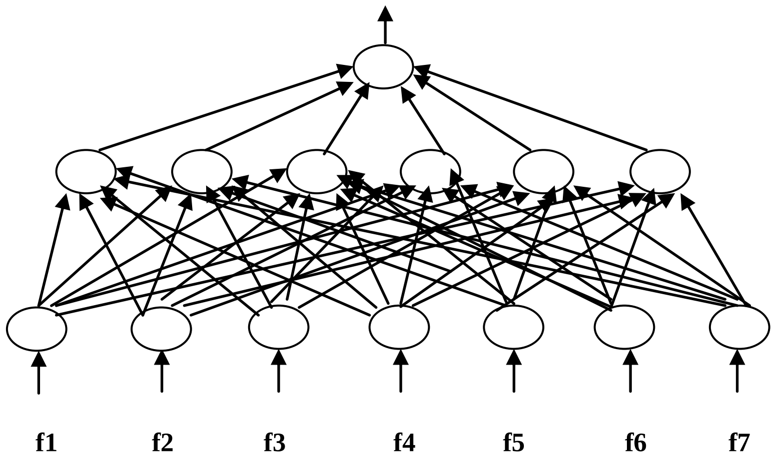

Figure 3: The Neural Network after Training

ii. Feature Fusion-

Once the network has learned the features that must exist in summary sentences, we need to determine the trends and relationships among the features that are inherent in the sentences. This is accomplished by the feature fusion phase, which consists of two steps:

1) Eliminating uncommon features - After the training phase, the connections having very small weights can be pruned for each input to hidden layer connection **,** and for each hidden to output layer connection; and

2) Collapsing the effects of common features - After pruning the network, the hidden layer activation values for each hidden layer neuron are clustered using adaptive clustering technique.

Clustering technique requires grouping together objects that are similar to each other. In adaptive clustering technique, each cluster is identified by its centroid and frequency. The centroid of each cluster represents the mean of the activation values in the cluster and can be used as the representative value of the cluster, while the frequency of each cluster represents the number of activation values in that cluster. By using the centroids of the clusters, each hidden layer neuron has a minimal set of activations. This helps with getting generalized output at the output layer.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The below fig 4 (a) (b) shows the neural network before and after pruning process which shows the collapses of hidden layer unit.

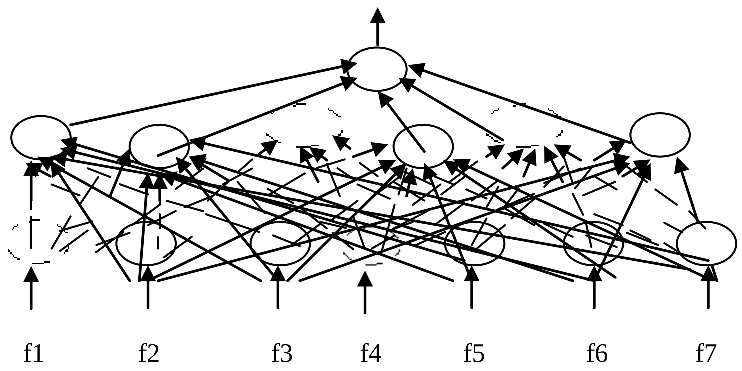

Figure 4 (a): The Neural Network before Pruning

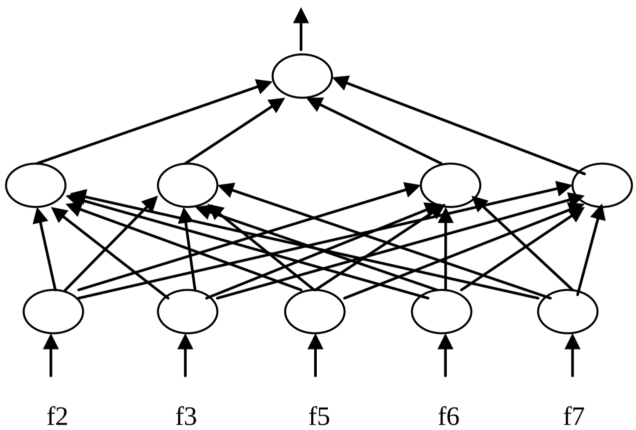

Figure 4 (b): The Neural Network after Pruning

iii. Sentence Selection-

In the sentence selection phase, the activation value of each hidden layer neuron is replaced by the centroid of the cluster, which the activation value belongs to. Basically an approach of clustering technique is dynamic; due to dynamic clustering the activation values are reclustered and the radius of new clusters is set to one-half of the original clusters [1] [2].

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The benefits of reclustering by a factor of two steps:

1) Because of reclustering the misclassified clusters are classified again according to activation values are formed into in appropriate cluster.
2) Due to reclustering eliminates any possible overlaps among clusters.

The combination of these two steps corresponds to generalizing the effects of sentence features. Once the network has been trained, pruned, and generalized, it can be used as a tool to determine whether or not each sentence should be included in the summary. This phase is achieved by providing control parameters for the desired radius and frequency of hidden layer activation clusters to select highly ranked sentences. The sentence ranking is directly proportional to cluster frequency and inversely proportional to cluster radius. Only sentences that satisfy the required cluster boundary and frequency of all hidden layer neurons are selected as high-ranking summary sentences. The selected sentences posses the common features inherent in the majority of summary sentences.

The beneath fig 5 which show the sentence selection procedure after feature fusion, which is uses the modified neural network to filter the text and to select only the highly ranked sentences.

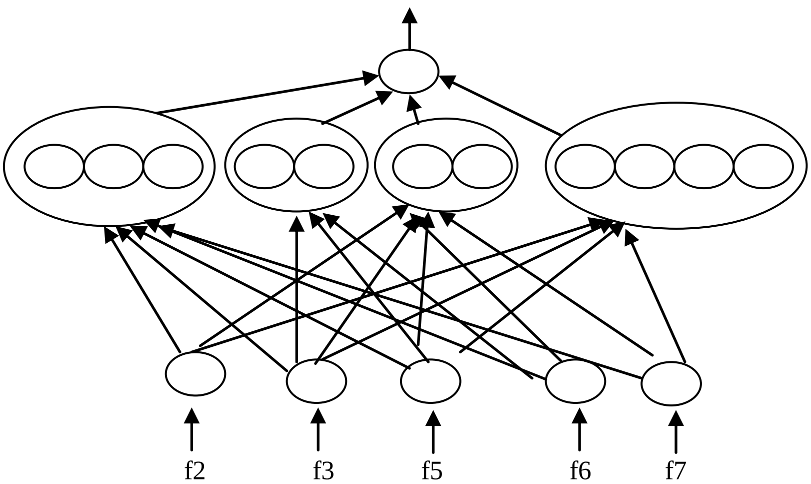

Figure 5: The Neural Network after Feature Fusion

V. OUTCOME

The text summarization using neural network system has module like Neural Network training, feature fusion, Sentence selection. During training, a set of news articles and associated summaries are given to the system. The feature fusion module process the articles and extracts sentences or words from that articles and then extracts features from the articles based on word or sentence. These features are given to the NN for learning the pattern. During feature fusion, prunes the neural network and collapses the hidden layer unit activations into discrete values with frequencies. This step generalizes the important features that must exist in the summary sentences by fusing the features.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Finally in sentence selection phase uses the modified neural network to filter the text and to select only the highly ranked sentences. During testing a set of news articles is given to the system and passed through the feature fusion module to get feature vector or test instances. This test instances are given to inputs of NN and the trained network

classify whether an input (generally a word) is summary worthy or not. Depending on summary worthy words a sentence is ranked and the system selects only top ranked sentence for summary.

VI. CONCLUSION

Our system performed well on the test paragraphs with accuracy of neural network summarization is nearer to 90 %. The performance of the text summarization process, selection of features, and selection of summary sentences depends on the style of the human reader. The network is trained by the human reader and to which sentences the human reader holds to be important in a paragraph. This, in fact, which is helpful to individual readers, can train the neural network as per their own styles. As well as the selected features can be modified with respect to the reader's needs and requirements.

REFERENCES

[1] Lawrence Wong M, "Text Summarization: An Overview", 2006.
[2] Khosrow Kaikhah, "Text Summarization Using Neural Networks", Ph.D, 2004.
[3] Vishal Gupta, "A Survey of Text Summarization Extractive Technique", Vol. 2, No. 3, pp. 258-268, 2010.
[4] Kamal Sarkar, Mita Nasipuri and Suranjan Ghose, "A New Approach to Keyphrase Extraction Using Neural Networks", Vol. 7, Issue 2, No 3, pp. 1694-0784, 2010.
[5] J. Wang, H. Peng, J.-S. Hu, "Automatic Keyphrases Extraction from Document Using Neural Network", 2005.
[6] Pramod L. Narasimhaa, Walter H. Delashmitb, Michael T. Manrya, Jiang Lic, Francisco Maldonadod, "An integrated growing-pruning method for feedforward network training", pp. 2831–2847, 2008.
[7] Bruno J.T. Fernandes a,b, George D.C. Cavalcanti a, Tsang I. Ren a, c, "AutoAssociative Pyramidal Neural Network for one class pattern classification with implicit feature extraction", pp. 7258–7266, 2013.
[8] Anjali R. Deshpande, Lobo L. M. R. J,"Text Summarization using Clustering Technique", Vol4 Issue8, pp. 3347-335, 2013.
[9] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", vol. 2, no.1, 2009.
[10] A. P. Siva kumar, Dr. P. Premchand and Dr. A. Govardhan, "Query-Based Summarizer Based on Similarity of Sentences and Word Frequency", International Journal of Data Mining & Knowledge Management Process, vol.1, no.3, 2011.
[11] Y. Matsuo, Y. Ohsawa, M. Ishizuka, "KeyWorld: Extracting Keywords from a Document as a Small World", Vol. 2226, pp. 271– 28, 2001.
[12] Y. B. Wu, Q. Li, "Document keyphrases as subject metadata: incorporating document key concepts in search results",Volume 11, Number 3, 229-249, 2008.
[13] Xudong Jiang, Alvin Harvey Kam Siew Wah, "Constructing and training feed-forward neural networks for pattern classification", pp. 853 – 867, 2003.