Multi-Document and Multi-Lingual Summarization using Neural Networks

M.Karthi Keyan
PG Student
National Engineering College

K.G.Srinivasagan, Ph.D,
Profossor & Head
National Engineering College

ABSTRACT
This system proposes Multi-lingual (Tamil and English) Multi-document summarization by neural networks. The system involves three steps. In first step, the sentences of the documents are converted into vector form. In the second step weight values are assigned to vector form based on sentence features. Depend on sentence weight value, single document summarization is done. The output of single document summarization is used as an input for multi-document Summarization. Final step is a sentence selection, in which output summary is selected based on the similarity and dissimilarity measures. Sentence similarity and dissimilarity measures are used to compare the sentences. From that, resultant summary is produced. The proposed system can be able to summarize both Tamil and English online news papers.

Keywords
Neural Networks, Features, Summarization

1. INTRODUCTION
Information Retrieval (IR) is the science of searching for documents, information within documents, metadata about documents, relational databases and the World Wide Web. Summarization is a branch which deals with information retrieval. Automatic summarization is the process of creating a summary of one or more text documents. For instance, we may summarize a large amount of news from different sources .Many summarization techniques and their evaluation methods have been developed for this purpose. Such techniques are RANDOM, LEAD, MEAD and PYTHY etc. which are used to generate the summary. MEAD is the recent toolkit for summarization.

Automatic summarization has been tried in the last two decades vigorously. However it is hard to say how much greater interpretive sophistication, at sentence or text level, contributes to performance without use NLP, the generated summary may suffer from lack of cohesion and semantics. Summarization can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting vital sentence from the original document and concatenating them into shorter form. The importance of the sentences is decided based on its statistical and linguistic features. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words.

We propose a machine learning approach that uses artificial neural networks to produce summaries of arbitrary length of news articles. The system involves three steps. In first step, the sentences of the documents are converted into vector form. In the second step weight values are assigned to vector form based on sentence features. Depend on sentence weight value, single document summarization is done. The output of single document summarization is used as an input for multi-document Summarization. Final step is a sentence selection, in which output summary is selected based on the similarity and dissimilarity measures. Sentence similarity and dissimilarity measures are used to compare the sentences. From that, resultant summary is produced. The proposed system can be able to summarize both Tamil and English online news papers.

2. RELATED WORK
Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form [1]. The importance of sentences is decided based on statistical and linguistic features of sentences [2]. An Abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language [3]. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document [4].

The first step in summarization by extraction is the identification of important features. There are two distinct types of features: non-structured features and structured features. One group of researcher utilize only non-structured features; such features include ,Paragraph follows Title, Paragraph location, First sentence in paragraph[5], Positive keyword in the sentence, Negative keyword in the sentence[6], Sentence Centrality[7] ,Sentence relative length, Sentence resemblance to the title[8], Sentence inclusion of numerical data[9], Sentence inclusion of name entity[10] ,Sentence similarity with other sentence[11], Aggregate similarity measures[12] ,Term weight[13].This presents literature regarding summarization work based on grouping similar sentences and word frequency. Some Basic uses term frequency as an approach to identify important sentences reducing information redundancy [14]. Local Topic Identification [15] and word frequency are techniques used for Single Document Summarization [16]. Combination of other techniques with Similarity of first sentence for Multi Document Summarization [17] .The use of frequency has proven useful in literature. This is because authors state information in several ways. We calculate similarity of sentences using cosine similarity measure. Sum Focus is used to calculate word frequency [18]. After Pre-processing, producing the summary involves the following Steps.

1. Calculate similarity of sentences present in documents with sentence features.

2. After calculating similarity group sentences based on their similarity values.

3. Calculate sentence score using word frequency and sentence location feature.

4. Pick the best scored sentences from each group and put it in summary.

5. Reduce summary length to exact 100 words.

3. FEATURES

Each document is converted into a list of sentences. Each sentence is represented as a vector (f1, f2, f3, f4, f5, f6, f7, f8, f9, f9, f10, f11, f12, f13), composed of 13 features. In the below table .1, features f1 to f3 represent the location of the sentence within the document, or within its paragraph. It is expected that in structured documents such as news articles, these features would contribute to selecting summary sentences. Feature f4, positive keyword sentence, it is the keyword that frequently included in the summary. Feature f5, Negative keyword in the sentence is keywords that are unlikely to occur in the summary. Feature f6, sentence centrality is the vocabulary overlap between this sentence and other sentences in the document. Feature f7 is Sentence relative length. This feature is useful to filter out short sentences such as datelines and author name commonly found in news articles. The short sentences are not expected to belong in the summary. We use length of the sentences, which is the ratio of the number of word occurring in the sentence over number of word in the longest sentence in the document. overlap between this sentence and the document title .Feature f9 is Sentence inclusion of numerical data; Sentences that contain numerical data are more important than rest of sentences and are probably included in the summary. Feature f8 is Sentence resemblance to the title, is the vocabulary Feature f10 is the Sentence inclusion of name entity. Sentence that contains more proper nouns is an important one and it is most probably included in the summary. This feature measures the similarity between sentence and each other sentences. It measures how much vocabulary overlap between this sentence S and other sentences in the document. It is computed by cosine similarity measure with resulting between 0 and 1. The score of this feature for a sentence S is obtained by computing the ratio of similarity of sentence S with each other sentence over the maximum similarity between two sentences. Feature f12 is Aggregate similarity measures. Aggregate similarity measures the importance of a sentence. Instead of counting the number of links connecting a node (sentence) to other nodes (Bushy path), aggregate similarity sums the weights (similarities) on the links. Finally feature is f13 Term weight. The frequency of term occurrence within a document has often been used for calculating the importance of sentence. The score of sentence can be calculated as the sum of the score of word in the sentence. This feature is expected to be important because the salience of a sentence may be affected by the number of words in the sentence also appearing in the title. Next feature f11 is Sentence similarity with other sentence. These features may be changed or new features may be added. The selection of features plays an important role in determining the type of sentences that will be selected as part of the summary and, therefore, would influence the performance of the neural network.

Table 1: Features

Features	Description
F1	Paragraph follows Title
F2	Paragraph location
F3	First sentence in paragraph
F4	Positive keyword in the sentence
F5	Negative keyword in the sentence
F6	Sentence Centrality
F7	Sentence relative length
F8	Sentence resemblance to the title
F9	Sentence inclusion of numerical data
F10	Sentence inclusion of name entity
F11	Sentence similarity with other sentence
F12	Aggregate similarity measures
F13	Term weight

The importance of sentences is decided based on statistical and linguistic features of sentences

4. THE DOCUMENT SUMMARIZATION PROCESS

The proposed system is used to summarize the Multi document with Multilinguistics. The system will be used for summarizing both Tamil and English document. The corpus is collected for both Tamil and English document from the online news paper. The output summary is compared with the manual summarized output**.**

4.1 An Artificial Neural Network (ANN)

An Artificial Neural Network is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANN's, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. The figure .2 shows Artificial Neural Network.This system uses Mcculloch-Pitts model. It is the two stage artificial neural networks. This neural network is the linear threshold gate. It is a neuron of a

set of inputs $I_1, I_2, , , I_m$ and one output y . The linear threshold gate simply classifies the set of inputs into two different classes. Thus the output is binary. Such a function can be described mathematically using equations 1:

$$Sum = \sum_{i=1}^{N} I_i W_i \qquad (1)$$

$$Y = F(Sum)$$

$W_1, W_2, , , W_m$,Are weight values normalized in the range of either $(0,1)$ or $(1,-1)$ and associated with each input line, Sum is the weighted sum, and T is a threshold constant. The McCulloch-Pitts model of a neural network is simple yet has substantial computing potential. It also has a precise mathematical definition. However, this model is so simplistic that it only generates a binary output. In this network, input files are converted into vector form. The output of a neuron is a function of the weighted sum of the inputs plus a bias.

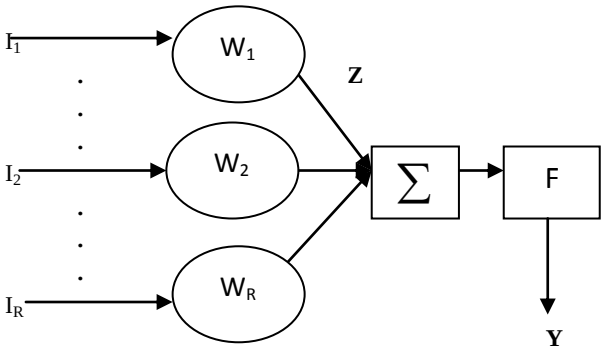

Fig 2: Artificial Neural Network

The Bias 'b' can treated as a weight whose input is always between 0 and 1.The weight values are assigned to these vectors based on specific features. Training is the act of presenting the network with some sample data and modifying the weights to better approximate the desired function. This process is done for document summarization. So these features are sentence features. The output of network considered as single document summarization

4.2 Sentence Selection

Sentence selection is the important step of this summarization system. In this phase, sentence select based on the similarity and dissimilarity of them. Similarity between the sentences identified by Cosine similarity .It is a measure of similarity between two vectors. The cosine of 0 is 1, and less than 1 for any other angle; the lowest value of the cosine is -1. The cosine of the angle between two vectors thus determines whether two vectors are pointing in roughly the same direction in the equation .2

$$Similarity = Cos(x) = \left(\frac{A.B}{\| A \| . \| B \|} \right) \qquad (2)$$

- Calculate similarity of sentences present in documents with sentence features.
- After calculating similarity group sentences based on their similarity values.
- Calculate sentence score using word frequency and sentence location feature.
- Pick the best scored sentences from each group and put it in summary.

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (tf-idf weights) cannot be negative.

5. RESULTS

We used articles from the Internet with various topics such as technology, sports, and world news to train the network. Each article consists of 19 to 56 sentences with an average of 34 sentences. The entire set consists of 2,'835 sentences. Every sentence, which is represented by a feature vector, is labeled as either a summary sentence or an unimportant sentence. A human reader performed the labeling of the sentences. A total of 163 sentences were labeled as summary sentence with an average of 12 sentences per article. The figure 3 shows that present a new extractive text summarization technique, for single documents based on the neural networks. Extractive text summarization works by selecting a subset of important sentences from the original document. The output of this single document summarization is used to input for the multi-document summarization. The System used text processing approaches as opposed to semantic approaches related to natural language. In order to assess the accuracy of all three neural networks, we selected 25 different news articles. The human reader and all three modified networks summarized the 25 news articles both Tamil and English document, independently.

Fig 3: Summarization

6. CONCLUSION

This system works on the main techniques to generate multi-document summarization, and describes the details of each step. The performance of the text summarization process depends predominantly on the style of the human reader. The selections of features as well as the selection of summary sentences by the human reader from the training paragraphs play an important role in the performance of the network. The neural network is trained according to the style of the human reader and to which sentences the human reader deems to be important in paragraph Individual readers can train the neural network according to their own styles. In addition, the selected features can be modified to reflect the reader's needs and requirements. To generate precise summarization, more in-depth understanding of the sentence (paragraph) is

required, so in future work we need to focus on the semantic and pragmatic information; it will be beneficial for the similarity calculation sentence extraction.

7. REFERENCES

[1] Hahn, U, "The challenges of Automatic Summarization", IEEE Transaction on Natural Language Process, vol.33, no.11, 2000.

[2] Pankaj Gupta, Vijay Shankar Pendluri, Ishant Vats, "Summarizing Text by Ranking Text Unit According to Shallow Linguistic Features", journal of Emerging Technologies in Web Intelligence ,vol.4,no.3,August 2011.

[3] Vishal Gupta,Gurpreet S.Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, vol.2,no.3,August 2010

[4] Yingxiong,Hongyan,LIU LeiLI, "Multi-Document Summarization Based on Improved Features and Clustering", IEEE conference on data mining and knowledge,vol.2,no.1,2010.

[5] Mohamed Abdel Fattah, Fuji Ren, "Probabilistic Neural Network Based Text Summarization" IEEE Conference on man and cybernetics, vol.6, 2008.

[6] Ramesh Vaishya,Surya Prakash Tripathi, "Strategic Approach for Automatic Text Summarization", journal of emerging technologies,vol.4,no.3,2011

[7] Khosrow Kaikhah "automatic text summarization with neural networks"IEEE conference on Intelligence System,vol.1,2004

[8] Yan-Xiang He, Hua Yang "Multi-document Summarization System by genetic algoritham", IEEE Conference on machine learning,2006

[9] Ryosuke Fujioka, Narazaki,"A neural network-based automatic summarization for local assemblies"on IEEE Conference on man and cybernetics , vol.4, 2006.

[10] A. P. Siva kumar,M.Ravi kumar,"Query-Based Summarizer Based on Similarity of Sentences and Word Frequency",IJCSI International Journal of knowledge engineering, Issues, vol. 8, Issue 3, no. 1, May 2011.

[11] Tao Li, Wei Peng, Charles Perng, Sheng Ma, and Haixun Wang, "An Integrated Data-Driven Framework for Computing System Management" IEEE Transaction on man,system,and cybernetics,vol.40, no. 1,2010.

[12] Tao Jiang and Ah-Hwee Tan, "Learning Image-Text Associations", IEEE Transaction on knowledge and data engineering, vol.18, no.6, 2008.

[13] Feifan Liu And Yang Liu, "Exploring correlation Between Rouge and Human Evaluation on Meeting Summarize", IEEE Transaction on audio, speech,language processing,vol.18,no.1,January 2010.

[14] Ohm Sornil "An Automatic text summarization approach using Content-Based and Graph-Based Characteristics"IEEE Conference on Intelligent system. vol.6, 2006.

[15] Tengfei Ma "Multi-Document Summarization Using Minimum Distortion", IEEE Conference on Dataniming, Vol.4, No.3, 2010.

[16] Lev Pevzner,Marti Hearst, "A Critique and Improvement of an

[17] Evaluation Metric for Text Segmentation", Association for Computational Linguistics, 1994.

[18] 17. A.V. Goldberg, R. Kennedy, "An Efficient cost scaling algorithm for the

Assignment problem", 1993