International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 6 (2014) pp. 626-634
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Suitability of Artificial Neural Network to Text Document Summarization in the Indian Language-Kannada

Jayashree R
Department of Computer Science
PES Institute of Technology
Bangalore, India
jayashree@pes.edu

Srikanta Murthy K
Department of Computer Science
PES Institute of Technology
Bangalore, India.
srikantamurthy@pes.edu

Basavaraj.S.Anami,
Department of ComputerScince ,
KLE Institute of Technology,
Hubli,India
anami_basu@hotmail.com

Abstract- The work explores the suitability of artificial neural network based summarizer to text document summarization in the Kannada language. A feed forward neural network, which is also called as back propagation network is trained on a corpus of text documents. The corpus is custom built for this purpose using Kannada web portals.

IndexTerms—feed forward neural network, training, thematic words, summarizer

I. INTRODUCTION

Information Retrieval (IR) is the process of finding material, usually document, of an unstructured type, that supplies the information needed within the given large collections of documents. Text summarization is the application of Information Retrieval. Text summarization based on machine learning is a widely used technology in the field of Information Retrieval (IR) and text mining, which have gained importance in the recent years. Further, text summarization helps to find the desired information quickly and efficiently. It is further noted that text document summarization is a predominant field of Natural Language Processing (NLP).

The important note is that, text document summarization is an accepted solution for the larger problem of content analysis. It is often seen as the task of passage extraction problem and the content delivered is the meaningful approximation of the original document. There are two variants of text document summarization: extractive summarization where the source text is transferred to constitute the summary text. Another approach is abstractive summarization which is intended to find individual manifestations of specified important notions regardless of status.

We have devised a summarizer based on artificial neural network approach. An artificial neural network (ANN), usually called as ANN or NN(Neural Network) , is a mathematical model or computational model inspired by the structural and/or functional aspects of a biological neural network. The figure1 depicts the structure of ANN.

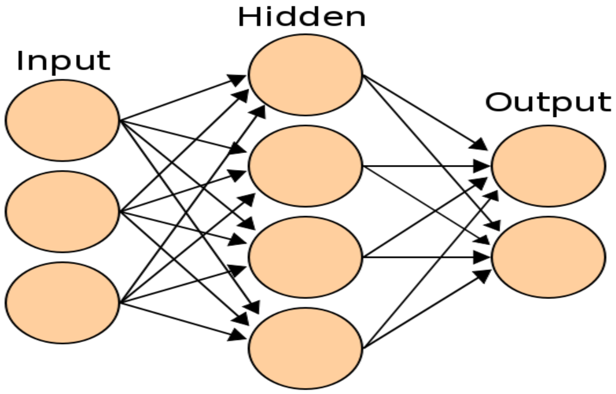


Figure 1: Artificial Neural Network

The first layer has input neurons, which send data via synapses to the second layer of neurons, and then via more synapses to the third layer of output neurons. More complex systems will have more layers of neurons with some having increased layers of input neurons and output neurons. The synapses store parameters called "weights" that manipulate the data in the calculations.

An ANN is typically defined by three types of parameters:

1. The interconnection pattern.

2. The learning process.
3. The activation function.

Mathematically, a neuron's network function $f(x)$ is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables.

The summarization process depicted here consists of training the artificial neural network. The parameters used are paragraph location, sentence length, whether paragraph follows title or not, thematic word ratio for each sentence, title word ratio, first sentence in paragraph, and sentence location.

The rest of the paper is organized as follows. Section-II highlights the literature survey; Section-III describes the methodology adopted in this work. Section-IV is about the Results and Discussion.

II LITERATURE SURVEY

(Mari-Sanna Paukkeri et.al, 2010) have proposed an approach, which selects words and phrases that best describe the meaning of the documents by comparing ranks of frequencies in the documents to the reference corpus.

(Gabor Berend et.al, 2010) have worked on the development of a frame work called SZETERGAK that treats the reproduction of reader assigned key words as a supervised learning task. In this work, a restricted set of token sequences are used as classification instances.

(Michael.J. Paul et.al, 2010) have proposed the work which uses an unsupervised probabilistic approach to model and extract multiple viewpoints in the given text. The authors have also used Lex rank, a novel random walk formulation to score sentences and pairs of sentences. The opposite view points based on both representations of the collections as well as their contrast with each other are considered. The word position information proves to play a significant role in document summarization.

(Letian Wang and Fang Li, 2010) have developed a methodology for key phrase extraction. This work is shown to be achieved using chunk based method. Keywords of a document are used to select key phrases from a candidate document.

(Su Nam Kim et al, 2010) have proposed the work on automatic production of key phrases for scientific papers. They have compiled a set of 284 scientific articles with key phrases carefully chosen by both their authors and readers.

(Fumiyo Fukumoto et.al, 2010) have presented a method for detecting key sentences from the documents that discuss the same event. To eliminate redundancy, they used spectral clustering. The classification of each sentence into groups comprises of semantically related sentences.

(Ahmet Aker Trevor Cohn, 2010) has developed techniques which have used A* algorithm to find the best extractive summary up to given length which is both optimal and efficient to run. Search is typically performed using greedy technique which selects each sentence in the decreasing order of model score until the desired length summary is reached.

(Xiaojun Wan et.al, 2010) have stated that cross language document summary is another upcoming research area in natural language processing, wherein the input document is in one language and the summary is in another language. The authors have worked with English document and summaries produced are in Chinese.

(Hitoshi Nishikawa et.al, 2010) have observed that a novel integer linear programming (ILP) formulation is used to generate summaries taking into consideration content coherence, sequence of sentences etc to get a better summary. They have also proved that size limitation becomes a bottleneck for content coherence which creates interest in improving sentence limited summarization techniques.

(You Ouyang et.al, 2010) have proved that the first occurrence of the word is important, and it decreases with the ordinal positions of appearances. The notion here is that the first sentence or a word in a paragraph is very important, which is not always true, because it depends on the writing style. Some prefer to give background first and keep conclusive sentences at the end.

(Feng Jin et.al, 2010) have presented a comparative study of two algorithms, the ranking problem and the selection problem. These are the two key algorithms existing for extractive summarization. From the labels and features of each sentence, they trained a model to infer the proper ranking of sentences in a document. They used a ranking algorithm based on neural networks.

(Vishal Guptha et.al, 2010) have presented a survey of text summarization by extractive techniques .There is a mention of neural network approach to text summarization. The neural network is trained using paragraphs for identifying sentences which are to be included in the summary and also to identify sentences to be excluded in the summary.

(Rajesh S. Prasad et.al, 2009) have worked on connectionist approach to generic text summarization. They have proposed an approach based on speech disambiguation using recurrent neural network. The approach has the ability to learn grammatical structure through experience. Text summarization using artificial neural network and fuzzy logic are suggested.

(Marina Litvak et.al, 2008) have observed that context factors such as input, purpose and output influence the process of summarization. They worked on graph-based keyword extraction for single-document summarization. In this approach they suggested two approaches, namely supervised and unsupervised for the cross-lingual keyword-extraction, which is used as the first step in extractive summarization of text document.

(Ronan Collobert et.al, 2007) have described the neural network architecture for solving the problem of semantic role labeling. The method directly maps from source sentence to semantic tags for a given predicate without the aid of a parser or a chunker.

(Krysta M. Svore et.al, 2007) have presented a new approach to text document summarization based on neural network. They extracted a set of features from each sentence in a document that highlights its importance in the document.

(Khosrow, 2004) has proposed an artificial neural network approach to text summarization. Neural Network is trained and modified through feature fusion to produce summaries of articles of arbitrary lengths.

(L.Galavotti et.al, 2000) have proposed two approaches to document summarization, supervised and unsupervised methods. The work made use of supervised approach, wherein a model is trained to determine if a candidate phrase is a key phrase. These methods first build a word graph according to word co occurrences within the document and then use random walk techniques to measure the importance of a word.

(Simone Teufel et.al, 1999) have observed that argumentative classification of extracted sentences acts as a first step towards flexible abstracting. They claimed that the rhetorical structure of a document is useful for automatic abstraction. The units of source text such as problem statement, conclusions and results could be used in automatic extraction.

(Regina Barzilay et.al, 1997) have stated that lexical chains for text summarization are used to produce summary of a document without full semantic interpretation. They presented a robust approach for merging several knowledge sources.

(Eduard Hovy et.al, 1997) have developed a system called ‘SUMMARIST’ which is a robust automated text summarization system. The equation that best describes the work on summarization is equal to summation of topic identification, interpretation and generation.

(Julian Kupiec, 1995) has developed a trainable document summarizer, wherein the author has introduced a classification function to estimate the probability that a given sentence is included in an extract. New extracts are generated by ranking sentences according to this probability and depending on a user specified number. The top scoring sentences are taken into consideration.

(Tomek Strzalkowski ,1986) has exploited regularities of organization and style. He called this approach as Discourse Macro Structure (DMS). The appropriate passages are extracted from the original text.

From the literature survey, it is learnt that significant amount of work has been carried out in the area of text summarization. But the researchers have considered English documents for the purpose of research. The word and sentence level methods of summarization are being tried. Supervised and unsupervised approaches are used. Some work on cross summarization is seen in the literature. We need to look into Kannada language keeping in mind the need for summarization and cross summarization. However, no specific work on Kannada document is seen in the literature. Hence, the work on Kannada document summarization is undertaken.

IV . METHODOLOGY

Training a neural network model essentially means selecting one model from the set of allowed models that minimizes the cost criterion. In our application we used the following two algorithms or methods for the implementation of the Network.

- Multi-Layer Perceptron.
- Back-Propagation.

Back propagation is a common method of training artificial neural networks so as to minimize the objective function. It is a supervised learning method, and is a generalization of the delta rule. It requires a dataset of the desired output for many inputs, making up the training set. It is most useful for feed-forward networks (networks that have no feedback, or simply, that have no connections that loop). The term is an abbreviation for "backward propagation of errors". Back propagation requires that the activation function used by the artificial neurons (or "nodes") be differentiable.

A neural network is a machine that is designed to model the way in which a human brain performs a particular task or function of interest. Multi layer feed forward neural networks is an important class of neural networks. The neural network consists of three levels: a set of sensory nodes called as input layer, one or more computation nodes called as hidden layer and an output layer. The input signal propagates through the network in a forward fashion, on a layer by layer basis. These neural networks are also referred to as Multi Layer Perceptrons (MLPs). Multi layer perceptrons have been successfully applied to solve some of the difficult problems by training them in a supervised manner using back propagation algorithm. We have used this algorithm for text document summarization. The methodology could be best described using the block diagram given below.

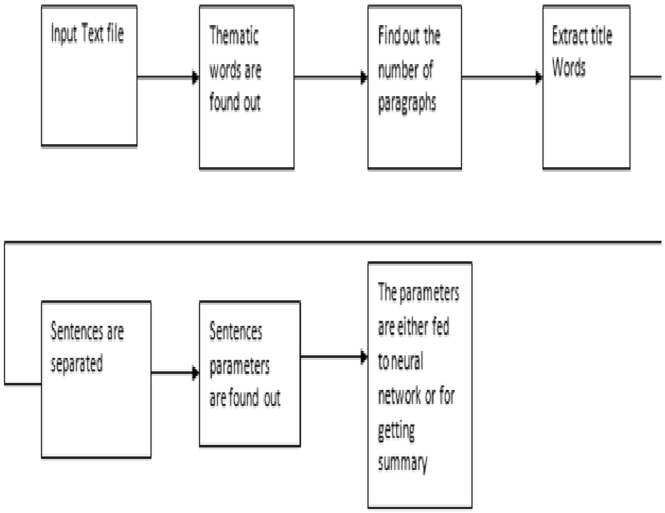

Input Text file
Thematic words are found out
Find out the number of paragraphs
Extract title Words
Sentences are separated
Sentences parameters are found out
The parameters are either fed to neural network or for getting summary

Figure 2: Block diagram of the summarizer

The first step is to consider a sample text file as shown in Box 1, its transliterated and translated versions are shown in Box 2 and Box 3. A sample document in sports category which is extracted from the web is considered here. The document pertains to cricketer Mr Rahul Dravid.

> ರಾಹುಲ್ ಶರದ್ ದ್ರಾವಿಡ್ (ಜನನ: ಜನವರಿ ೧೧, ೧೯೭೩) - ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ಆಟಗಾರರೊಲ್ಲಬ್ಬರು ಮತ್ತು ತಂಡದ ಮಾಜಿ ನಾಯಕ.ಮಧ್ಯಪ್ರದೇಶ ಮೂಲದವರಾದ ದ್ರಾವಿಡ್ ಪೂರ್ಣ ಕನ್ನಡಿಗರು. ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸುವುದರಲ್ಲಿ, ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸುನಿಲ್ ಗವಾಸ್ಕರ್ ನಂತರ ಮೂರನೇಯ ಭಾರತೀಯ. ಫೆಬ್ರುವರಿ ೧೪, ೨೦೦೭ ರಂದು ಅಂತರರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸಿದ ವಿಶ್ವದಲ್ಲಿ ೬ನೇ ಆಟಗಾರ,ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸೌರವ್ ಗಂಗೂಲಿ ನಂತರ ಮೂರನೇ ಭಾರತೀಯ. ಇವರು ಅಕ್ಟೋಬರ್ ೨೦೦೫ರಲ್ಲಿ ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ನಾಯಕನಾಗಿ, ಸೆಪ್ಟೆಂಬರ್ ೨೦೦೭ ರಲ್ಲಿ ತಂಡದ ನಾಯಕ ಸ್ಥಾನಕ್ಕೆ ರಾಜೀನಾಮೆ ಸಲ್ಲಿಸಿದರು.ರಾಹುಲ್ ಡ್ರಾವಿಡ್ ಭಾರತೀಯ ಪ್ರಿಮಿಯರ್ ಲೀಗ್ ನ, ರಾಯಲ್ ಚಾಲೆಂಜರ್ಸ್ ಬೆಂಗಳೂರು ತಂಡದಲ್ಲಿ ೨ ವರ್ಷ 'ಐಕಾನ್ ಆಟಗಾರ'ನಾಗಿ ಆಡಿ, ಈಗ ಜೈಪೂರದ ತಂಡವನ್ನು ಪ್ರತಿನಿಧಿಸುತ್ತಿದ್ದಾರೆ. ರಾಹುಲ್ ದ್ರಾವಿಡ್‌ಗೆ, ೨೦೦೦ರಲ್ಲಿ, "ವಿಜಡನ್ ಕ್ರಿಕೆಟರ್"ಅಂತ ಗೌರವಿಸಲಾಗಿದೆ. ದ್ರಾವಿಡ್‌ಗೆ, ೨೦೦೪ರಲ್ಲಿ, ವರ್ಷದ ಐಸಿಸಿ ಪ್ಲೆಯರ್ ಹಾಗೂ ವರ್ಷದ ಟೆಸ್ಟ್ ಆಟಗಾರನೆಂದೂ ಸನ್ಮಾನಿಸಲಾಗಿದೆ. ರಾಹುಲ್ ದ್ರಾವಿಡ್, ಟೆಸ್ಟ್ ಕ್ರಿಕೆಟ್ನಲ್ಲಿ ಅತಿ ಹೆಚ್ಚು ಕ್ಯಾಚ್ (೨೧೦) ಹಿಡಿದ ಆಟಗಾರರಾಗಿರುತ್ತಾರೆ.೭ ಅಗಸ್ಟ್ ೨೦೧೧ ರಂದು, ಒಂದು ದಿನದ ಹಾಗೂ ಟಿ೨೦ ಕ್ರಿಕೆಟ್‌ನಿಂದ ನಿವೃತ್ತಿ ಘೋಷಿಸಿದರು. ೨೦೧೨ ಮಾರ್ಚ್ ೯ ರಂದು ಟೆಸ್ಟ್ ಕ್ರಿಕೆಟ್‌ಗೆ ನಿವೃತ್ತಿ ಘೋಷಿಸಿದರು.

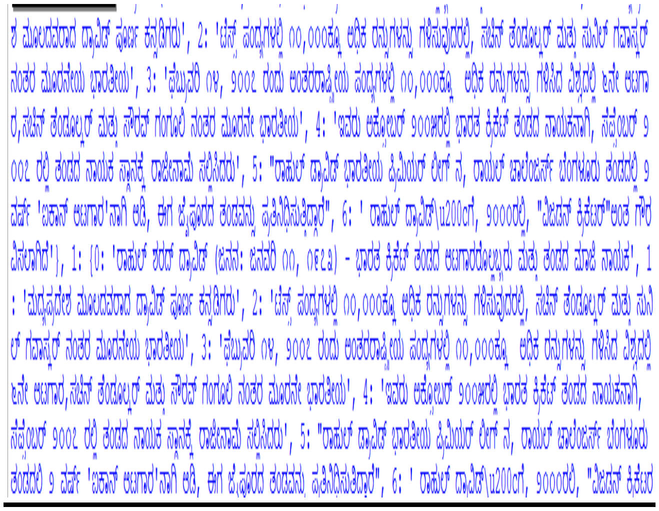
Box 1: A Sample Input Document

> Rāhul śarad drāviḍ (janana: Janavari 11, 1973) - bhārata kriket tanḍada āṭagārarollabbaru mattu tanḍada māji nāyaka.Madyapradēśa mūladavarāda drāviḍ pūrṇa kannaḍigaru. Ṭesṭ pandyagaḷalli 10,000kkū adhika rannugaḷannu gaḷisuvudaralli, sacin teṇḍūlkar mattu sunil gavāskar nantara mūranēya bhāratīya. Phebruvari 14, 2007 randu antararāṣṭrīya pandyagaḷalli 10,000kkū adhika rannugaḷannu gaḷisida viśvadalli 6nē āṭagāra,sacin teṇḍūlkar mattu saurav gaṅgūli nantara mūranē bhāratīya. Ivaru akṭōbar 2005ralli bhārata kriket tanḍada nāyakanāgi, sepṭembar 2007 ralli tanḍada nāyaka sthānakke rājīnāme sallisidaru.Rāhul ḍrāviḍ bhāratīya primiyar līg na, rāyal cāleñjars beṅgaḷūru tanḍadalli 2 varṣ'aikān āṭagāra'nāgi āḍi, īga jaipūrada tanḍavannu pratinidhisuttiddāre. Rāhul drāviḍge, 2000ralli, "vijaḍan kriketar"anta gauravisalāgide. Drāviḍge, 2004ralli, varṣada aisisi pleyar hāgū varṣada ṭesṭ āṭagāranendū sanmānisalāgide. Rāhul drāviḍ, ṭesṭ kriketnalli ati heccu kyāc (210) hiḍida āṭagārarāgiruttāre.7 Agasṭ 2011 randu, ondu dinada hāgū ṭi20 kriketninda nivr̥tti ghōṣisidaru. 2012 Mārci 9 randu ṭesṭ kriketge nivr̥tti ghōṣisidaru.

Box:2: Transliterated Version of Document in Box 1

> Rahul Sharad Dravid (born January 11, 1973) – one amongst them and former captain of India's cricket team. Dravid full my descendants were from Madhya pradesh. More than 10,000 run in Test matches, winning, third Indian after Sunil Gavaskar and Sachin Tendulkar. February 14, 2007, more than 10,000 runs in international matches and scored in the 6th player in the world, the third Indian after Sachin Tendulkar and Sourav Ganguly. Indian cricket team captain, who in October 2005, September 2007, served as team captain, Rahul dravid Indian Premier League, Royal Challengers Bangalore team 2 years 'icon player ' Nagy played, and now represents team Jaipur. Rahul dravidge, 2000, "vijadan Cricketer" I am honored. Dravidge, in 2004, the year the ICC Test Player of the Year was conferred on him. Rahul Dravid, Test cricket is the most catches (210) to will be players On 7 August 2011 , announced his retirement from one-day and Twenty20 cricket .Announced his retirement from Test cricket on March 9, 2012.

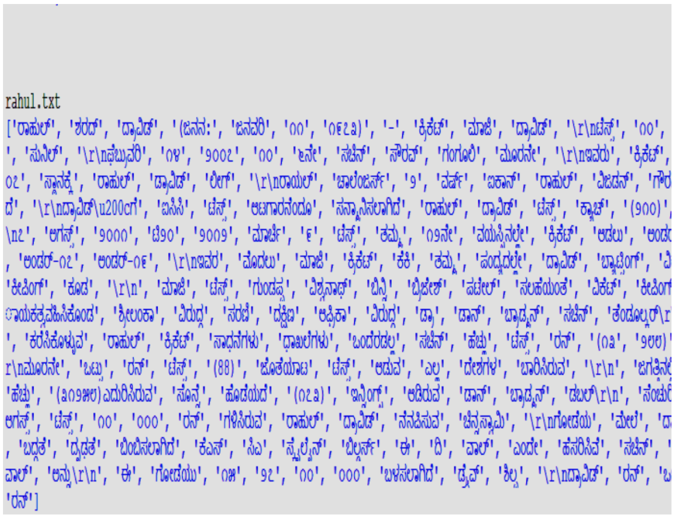
Box:3: English Meaning of Document in Box 1



```
ಶ ಮೂಲದವರಾದ ದ್ರಾವಿಡ್ ಪೂರ್ಣ ಕನ್ನಡಿಗರು', 2: 'ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸುವುದರಲ್ಲಿ, ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸುನಿಲ್ ಗವಾಸ್ಕರ್
ನಂತರ ಮೂರನೇಯ ಭಾರತೀಯ', 3: 'ಫೆಬ್ರುವರಿ ೧೪, ೨೦೦೭ ರಂದು ಅಂತರರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ  ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸಿದ ವಿಶ್ವದಲ್ಲಿ ೬ನೇ ಆಟಗಾ
ರ,ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸೌರವ್ ಗಂಗೂಲಿ ನಂತರ ಮೂರನೇ ಭಾರತೀಯ', 4: 'ಇವರು ಅಕ್ಟೋಬರ್ ೨೦೦೫ರಲ್ಲಿ ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ನಾಯಕನಾಗಿ, ಸೆಪ್ಟೆಂಬರ್ ೨
೦೦೭ ರಲ್ಲಿ ತಂಡದ ನಾಯಕ ಸ್ಥಾನಕ್ಕೆ ರಾಜೀನಾಮೆ ಸಲ್ಲಿಸಿದರು', 5: "ರಾಹುಲ್ ದ್ರಾವಿಡ್ ಭಾರತೀಯ ಪ್ರಿಮಿಯರ್ ಲೀಗ್ ನ, ರಾಯಲ್ ಚಾಲೆಂಜರ್ಸ್ ಬೆಂಗಳೂರು ತಂಡದಲ್ಲಿ ೨
ವರ್ಷ 'ಐಕಾನ್ ಆಟಗಾರ'ನಾಗಿ ಆಡಿ, ಈಗ ಜೈಪೂರದ ತಂಡವನ್ನು ಪ್ರತಿನಿಧಿಸುತ್ತಿದ್ದಾರೆ", 6: ' ರಾಹುಲ್ ದ್ರಾವಿಡ್\u200cಗೆ, ೨೦೦೦ರಲ್ಲಿ, "ವಿಜಡನ್ ಕ್ರಿಕೆಟರ್"ಅಂತ ಗೌರ
ವಿಸಲಾಗಿದೆ'}, 1: {0: 'ರಾಹುಲ್ ಶರದ್ ದ್ರಾವಿಡ್ (ಜನನ: ಜನವರಿ ೧೧, ೧೯೭೩) - ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ಆಟಗಾರರಲ್ಲೊಬ್ಬರು ಮತ್ತು ತಂಡದ ಮಾಜಿ ನಾಯಕ', 1
: 'ಮಧ್ಯಪ್ರದೇಶ ಮೂಲದವರಾದ ದ್ರಾವಿಡ್ ಪೂರ್ಣ ಕನ್ನಡಿಗರು', 2: 'ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸುವುದರಲ್ಲಿ, ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸುನಿ
ಲ್ ಗವಾಸ್ಕರ್ ನಂತರ ಮೂರನೇಯ ಭಾರತೀಯ', 3: 'ಫೆಬ್ರುವರಿ ೧೪, ೨೦೦೭ ರಂದು ಅಂತರರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ  ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸಿದ ವಿಶ್ವದಲ್ಲಿ
೬ನೇ ಆಟಗಾರ,ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸೌರವ್ ಗಂಗೂಲಿ ನಂತರ ಮೂರನೇ ಭಾರತೀಯ', 4: 'ಇವರು ಅಕ್ಟೋಬರ್ ೨೦೦೫ರಲ್ಲಿ ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ನಾಯಕನಾಗಿ,
ಸೆಪ್ಟೆಂಬರ್ ೨೦೦೭ ರಲ್ಲಿ ತಂಡದ ನಾಯಕ ಸ್ಥಾನಕ್ಕೆ ರಾಜೀನಾಮೆ ಸಲ್ಲಿಸಿದರು', 5: "ರಾಹುಲ್ ದ್ರಾವಿಡ್ ಭಾರತೀಯ ಪ್ರಿಮಿಯರ್ ಲೀಗ್ ನ, ರಾಯಲ್ ಚಾಲೆಂಜರ್ಸ್ ಬೆಂಗಳೂರು
ತಂಡದಲ್ಲಿ ೨ ವರ್ಷ 'ಐಕಾನ್ ಆಟಗಾರ'ನಾಗಿ ಆಡಿ, ಈಗ ಜೈಪೂರದ ತಂಡವನ್ನು ಪ್ರತಿನಿಧಿಸುತ್ತಿದ್ದಾರೆ", 6: ' ರಾಹುಲ್ ದ್ರಾವಿಡ್\u200cಗೆ, ೨೦೦೦ರಲ್ಲಿ, "ವಿಜಡನ್ ಕ್ರಿಕೆಟರ
```

Figure 3: Thematic Words

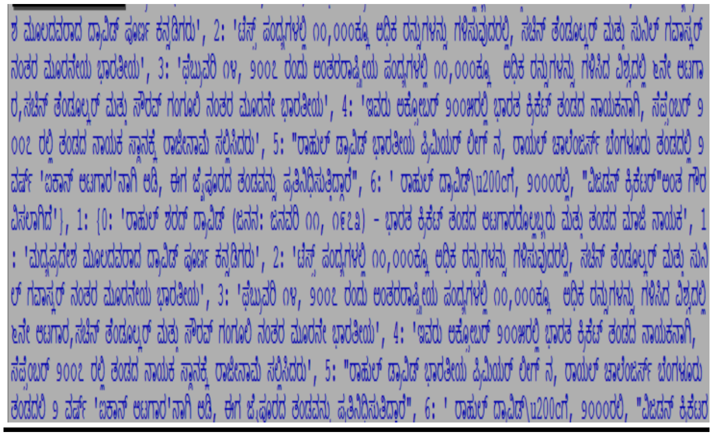
The thematic words contained in the document are found out and the number of paragraphs is selected in a given document, a sample such list is given below.



```
rahul.txt
['ರಾಹುಲ್', 'ಶರದ್', 'ದ್ರಾವಿಡ್', '(ಜನನ:', 'ಜನವರಿ', '೧೧', '೧೯೭೩)', '-', 'ಕ್ರಿಕೆಟ್', 'ಮಾಜಿ', 'ದ್ರಾವಿಡ್', '\r\nಟೆಸ್ಟ್', '೧೦',
', 'ಸುನಿಲ್', '\r\nಫೆಬ್ರುವರಿ', '೧೪', '೨೦೦೭', '೧೦', '೬ನೇ', 'ಸಚಿನ್', 'ಸೌರವ್', 'ಗಂಗೂಲಿ', 'ಮೂರನೇ', '\r\nಇವರು', 'ಕ್ರಿಕೆಟ್',
೦೭', 'ಸ್ಥಾನಕ್ಕೆ', 'ರಾಹುಲ್', 'ದ್ರಾವಿಡ್', 'ಲೀಗ್', '\r\nರಾಯಲ್', 'ಚಾಲೆಂಜರ್ಸ್', '೨', 'ವರ್ಷ', 'ಐಕಾನ್', 'ರಾಹುಲ್', 'ವಿಜಡನ್', 'ಗೌರ
ದೆ', '\r\nದ್ರಾವಿಡ್\u200cಗೆ', 'ಐಸಿಸಿ', 'ಟೆಸ್ಟ್', 'ಆಟಗಾರನೆಂದೂ', 'ಸನ್ಮಾನಿಸಲಾಗಿದೆ', 'ರಾಹುಲ್', 'ದ್ರಾವಿಡ್', 'ಟೆಸ್ಟ್', 'ಕ್ಯಾಚ್', '(೨೧೦)',
\n೭', 'ಆಗಸ್ಟ್', '೨೦೧೧', 'ಟಿ೨೦', '೨೦೧೨', 'ಮಾರ್ಚ್', '೯', 'ಟೆಸ್ಟ್', 'ತಮ್ಮ', '೧೨ನೇ', 'ವಯಸ್ಸಿನಲ್ಲೇ', 'ಕ್ರಿಕೆಟ್', 'ಆಡಲು', 'ಅಂಡ
, 'ಅಂಡರ್-೧೭', 'ಅಂಡರ್-೧೯', '\r\nಇವರ', 'ಮೊದಲು', 'ಮಾಜಿ', 'ಕ್ರಿಕೆಟ್', 'ಕೆಕಿ', 'ತಮ್ಮ', 'ಪಂದ್ಯದಲ್ಲೇ', 'ದ್ರಾವಿಡ್', 'ಬ್ಯಾಟಿಂಗ್', 'ವಿ
'ಕೀಪಿಂಗ್', 'ಕೂಡ', '\r\n', 'ಮಾಜಿ', 'ಟೆಸ್ಟ್', 'ಗುಂಡಪ್ಪ', 'ವಿಶ್ವನಾಥ್', 'ಬಿನ್ನಿ', 'ಬ್ರಿಜೇಶ್', 'ಪಟೇಲ್', 'ಸಲಹೆಯಂತೆ', 'ವಿಕೆಟ್', 'ಕೀಪಿಂಗ್
◌ಾಯಕತ್ವವಹಿಸಿಕೊಂಡ', 'ಶ್ರೀಲಂಕಾ', 'ವಿರುದ್ಧ', 'ಸರಣಿ', 'ದಕ್ಷಿಣ', 'ಆಫ್ರಿಕಾ', 'ವಿರುದ್ಧ', 'ಡ್ರಾ', 'ಡಾನ್', 'ಬ್ರಾಡ್ಮನ್', 'ಸಚಿನ್', 'ತೆಂಡೂಲ್ಕರ್\r
', 'ಕರೆಸಿಕೊಳ್ಳುವ', 'ರಾಹುಲ್', 'ಕ್ರಿಕೆಟ್', 'ಸಾಧನೆಗಳು', 'ದಾಖಲೆಗಳು', 'ಒಂದೆರಡಲ್ಲ', 'ಸಚಿನ್', 'ಹೆಚ್ಚು', 'ಟೆಸ್ಟ್', 'ರನ್', '(೧೩', '೨೮೮)'
r\nಮೂರನೇ', 'ಒಟ್ಟು', 'ರನ್', 'ಟೆಸ್ಟ್', '(88)', 'ಜೊತೆಯಾಟ', 'ಟೆಸ್ಟ್', 'ಆಡುವ', 'ಎಲ್ಲ', 'ದೇಶಗಳ', 'ಬಾರಿಸಿರುವ', '\r\n', 'ಜಗತ್ತಿನ
'ಹೆಚ್ಚು', '(೩೧೨೫೮)ಎದುರಿಸಿರುವ', 'ಸೊನ್ನೆ', 'ಹೊಡೆಯದೆ', '(೧೭೩)', 'ಇನ್ನಿಂಗ್ಸ್', 'ಆಡಿರುವ', 'ಡಾನ್', 'ಬ್ರಾಡ್ಮನ್', 'ಡಬಲ್\r\n', 'ಸೆಂಚುರಿ
ಆಗಸ್ಟ್', 'ಟೆಸ್ಟ್', '೧೦', '೦೦೦', 'ರನ್', 'ಗಳಿಸಿರುವ', 'ರಾಹುಲ್', 'ದ್ರಾವಿಡ್', 'ನೆನಪಿಸುವ', 'ಚಿನ್ನಸ್ವಾಮಿ', '\r\nಗೋಡೆಯ', 'ಮೇಲೆ', 'ದ
, 'ಬದ್ಧತೆ', 'ದೃಢತೆ', 'ಬಿಂಬಿಸಲಾಗಿದೆ', 'ಕೆಎಸ್', 'ಸಿಎ', 'ಸ್ಟೈಲ್ವನ್', 'ಬಿಲ್ಡರ್ಸ್', 'ಈ', 'ದಿ', 'ವಾಲ್', 'ಎಂದೇ', 'ಹೆಸರಿಸಿವೆ', 'ಸಚಿನ್', '
ವಾಲ್', 'ಅನ್ನು\r\n', 'ಈ', 'ಗೋಡೆಯು', '೧೫', '೨೭', '೧೦', '೦೦೦', 'ಬಳಸಲಾಗಿದೆ', 'ಡ್ರೈವ್', 'ಶಿಲ್ಪ', '\r\nದ್ರಾವಿಡ್', 'ರನ್', 'ಬ
'ರನ್']
```

Figure 4: Selected sentences

Also, the sentence and paragraph locations are determined. Which are shown below:



```
ಶ ಮೂಲದವರಾದ ದ್ರಾವಿಡ್ ಪೂರ್ಣ ಕನ್ನಡಿಗರು', 2: 'ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸುವುದರಲ್ಲಿ, ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸುನಿಲ್ ಗವಾಸ್ಕರ್
ನಂತರ ಮೂರನೇಯ ಭಾರತೀಯ', 3: 'ಫೆಬ್ರುವರಿ ೧೪, ೨೦೦೭ ರಂದು ಅಂತರರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ  ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸಿದ ವಿಶ್ವದಲ್ಲಿ ೬ನೇ ಆಟಗಾ
ರ,ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸೌರವ್ ಗಂಗೂಲಿ ನಂತರ ಮೂರನೇ ಭಾರತೀಯ', 4: 'ಇವರು ಅಕ್ಟೋಬರ್ ೨೦೦೫ರಲ್ಲಿ ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ನಾಯಕನಾಗಿ, ಸೆಪ್ಟೆಂಬರ್ ೨
೦೦೭ ರಲ್ಲಿ ತಂಡದ ನಾಯಕ ಸ್ಥಾನಕ್ಕೆ ರಾಜೀನಾಮೆ ಸಲ್ಲಿಸಿದರು', 5: "ರಾಹುಲ್ ದ್ರಾವಿಡ್ ಭಾರತೀಯ ಪ್ರಿಮಿಯರ್ ಲೀಗ್ ನ, ರಾಯಲ್ ಚಾಲೆಂಜರ್ಸ್ ಬೆಂಗಳೂರು ತಂಡದಲ್ಲಿ ೨
ವರ್ಷ 'ಐಕಾನ್ ಆಟಗಾರ'ನಾಗಿ ಆಡಿ, ಈಗ ಜೈಪೂರದ ತಂಡವನ್ನು ಪ್ರತಿನಿಧಿಸುತ್ತಿದ್ದಾರೆ", 6: ' ರಾಹುಲ್ ದ್ರಾವಿಡ್\u200cಗೆ, ೨೦೦೦ರಲ್ಲಿ, "ವಿಜಡನ್ ಕ್ರಿಕೆಟರ್"ಅಂತ ಗೌರ
ವಿಸಲಾಗಿದೆ'}, 1: {0: 'ರಾಹುಲ್ ಶರದ್ ದ್ರಾವಿಡ್ (ಜನನ: ಜನವರಿ ೧೧, ೧೯೭೩) - ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ಆಟಗಾರರಲ್ಲೊಬ್ಬರು ಮತ್ತು ತಂಡದ ಮಾಜಿ ನಾಯಕ', 1
: 'ಮಧ್ಯಪ್ರದೇಶ ಮೂಲದವರಾದ ದ್ರಾವಿಡ್ ಪೂರ್ಣ ಕನ್ನಡಿಗರು', 2: 'ಟೆಸ್ಟ್ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸುವುದರಲ್ಲಿ, ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸುನಿ
ಲ್ ಗವಾಸ್ಕರ್ ನಂತರ ಮೂರನೇಯ ಭಾರತೀಯ', 3: 'ಫೆಬ್ರುವರಿ ೧೪, ೨೦೦೭ ರಂದು ಅಂತರರಾಷ್ಟ್ರೀಯ ಪಂದ್ಯಗಳಲ್ಲಿ ೧೦,೦೦೦ಕ್ಕೂ  ಅಧಿಕ ರನ್ನುಗಳನ್ನು ಗಳಿಸಿದ ವಿಶ್ವದಲ್ಲಿ
೬ನೇ ಆಟಗಾರ,ಸಚಿನ್ ತೆಂಡೂಲ್ಕರ್ ಮತ್ತು ಸೌರವ್ ಗಂಗೂಲಿ ನಂತರ ಮೂರನೇ ಭಾರತೀಯ', 4: 'ಇವರು ಅಕ್ಟೋಬರ್ ೨೦೦೫ರಲ್ಲಿ ಭಾರತ ಕ್ರಿಕೆಟ್ ತಂಡದ ನಾಯಕನಾಗಿ,
ಸೆಪ್ಟೆಂಬರ್ ೨೦೦೭ ರಲ್ಲಿ ತಂಡದ ನಾಯಕ ಸ್ಥಾನಕ್ಕೆ ರಾಜೀನಾಮೆ ಸಲ್ಲಿಸಿದರು', 5: "ರಾಹುಲ್ ದ್ರಾವಿಡ್ ಭಾರತೀಯ ಪ್ರಿಮಿಯರ್ ಲೀಗ್ ನ, ರಾಯಲ್ ಚಾಲೆಂಜರ್ಸ್ ಬೆಂಗಳೂರು
ತಂಡದಲ್ಲಿ ೨ ವರ್ಷ 'ಐಕಾನ್ ಆಟಗಾರ'ನಾಗಿ ಆಡಿ, ಈಗ ಜೈಪೂರದ ತಂಡವನ್ನು ಪ್ರತಿನಿಧಿಸುತ್ತಿದ್ದಾರೆ", 6: ' ರಾಹುಲ್ ದ್ರಾವಿಡ್\u200cಗೆ, ೨೦೦೦ರಲ್ಲಿ, "ವಿಜಡನ್ ಕ್ರಿಕೆಟರ
```

Figure 5: Sentences and Paragraph locations

The parameters thus found out are fed to the neural network for training. The method used for training is conjugate descent formulation of back propagation error in feed forward neural network. The common method for measuring the discrepancy between the expected output't' and the actual output 'y' using the standard error measure is given by the Equation (1).

$$E = (t - y)^2 \dots\dots\dots\dots (1),$$

Where E is the discrepancy error. The back propagation algorithm aims to find the set of weights that minimizes the error. The back propagation algorithm uses a gradient descent for minimizing the error. The derivative of the squared error function with respect to the weights of a network is given by the Equation (2)

$$E = \frac{1}{2}(t - y)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where, E is the squared error, t is the target output, y is the actual output of the output neuron. Error E depends on the output 'y'. The weighted sum of all its input is given by Equation (3),

$$y = \sum_{i=1}^{n} w_i \;\; x_i \dots\dots\dots\dots (3)$$

Where 'n' is the number of inputs to the neuron, w_i is the ith weight and x_i is the ith input value to the neuron.

The partial derivative of the error with respect to a weight w_i using the chain rule is given in the Equation (4);

$$\frac{\partial E}{\partial w_i} = \frac{dE}{dy}\,\frac{dy}{dnet}\,\frac{\partial net}{\partial w_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

$\frac{\partial E}{\partial w_i}$ gives how the error changes when the weights are changed. $\frac{dE}{dy}$ gives how the error changes when the output is changed. $\frac{dy}{dnet}$ gives how the output changes when the weighted sum changes. $\frac{\partial net}{\partial w_i}$ gives how the weighted sum changes as the weights change.

V . RESULTS

We have considered several categories of documents for producing the summary. The document categories include Literature, Entertainment, Politics, Social Sciences and Sports. From each of these categories, we have considered ten documents in some cases and 5 documents in some cases and summarized them with a compression factor of 20 and 10. The compression factor is the ratio of the number of sentences in the summarized document to the number of sentences in the actual document. We have compared the machine generated summary with that of summaries from two human experts. The results are shown below.

Compression factor = 20 lines of summary
Category –Sports

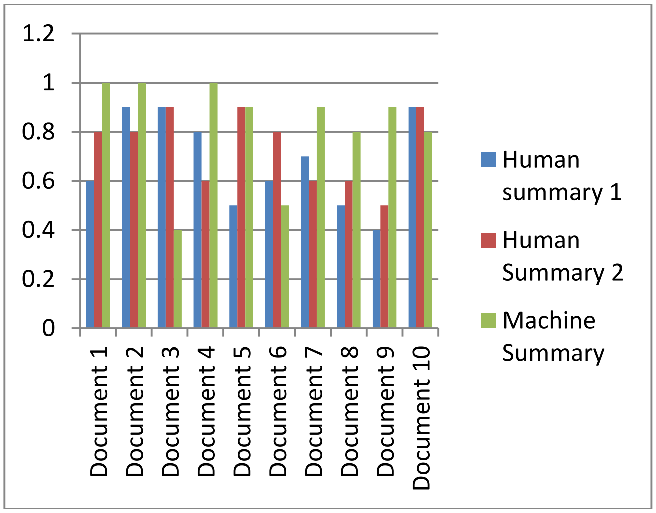


Figure 6: Effectiveness of Summary in Sports

Compression factor = 20 lines of summary
Category –Entertainment

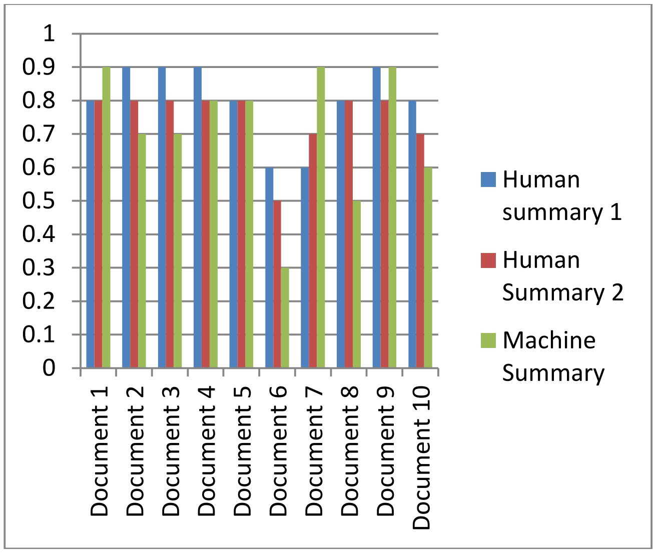


Figure 7: Effectiveness of Summary in Entertainment

Compression factor = 20 lines of summary
Category –Religion

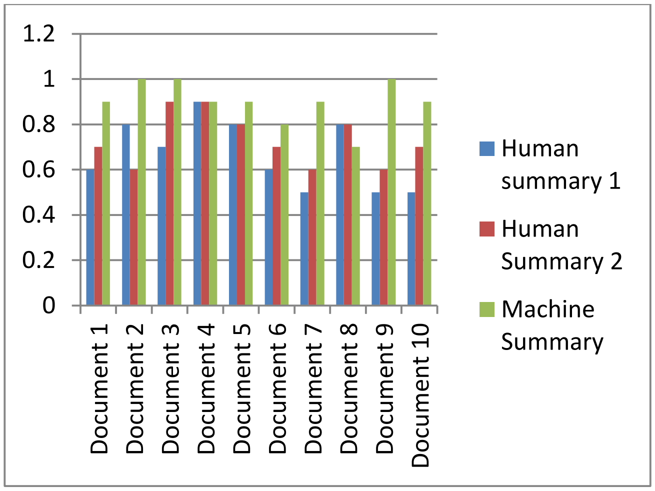


Figure 8: Effectiveness of Summary in Religion

Compression factor = 10 lines of summary
Category –Politics

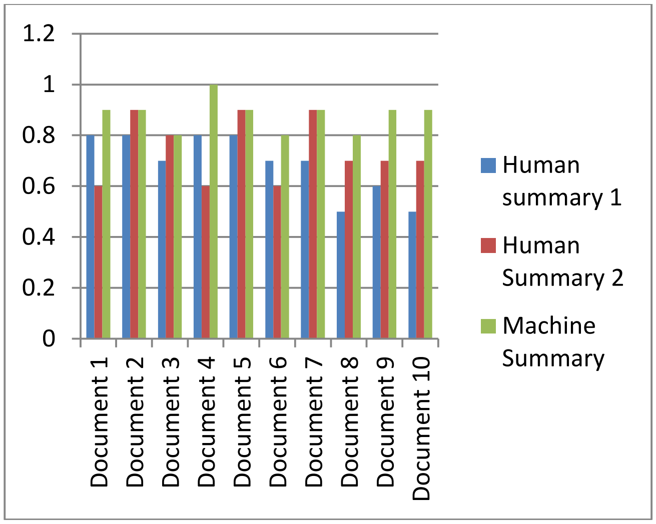


Figure 9: Effectiveness of Summary in Politics

Compression factor = 20 lines of summary
Category –Social Sciences

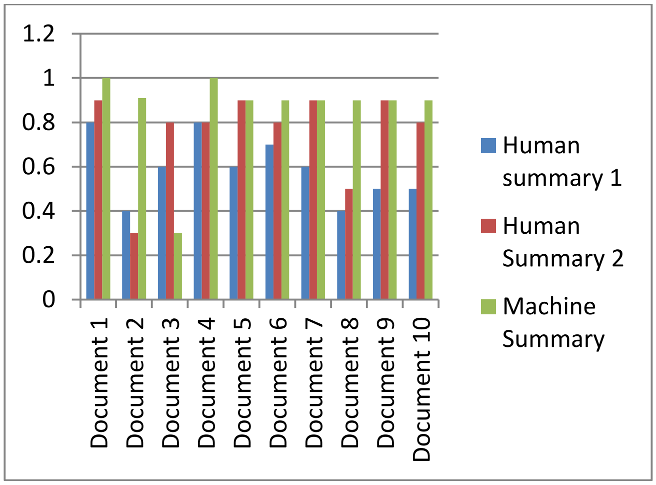


Figure 10: Effectiveness of Summary in Social Sciences

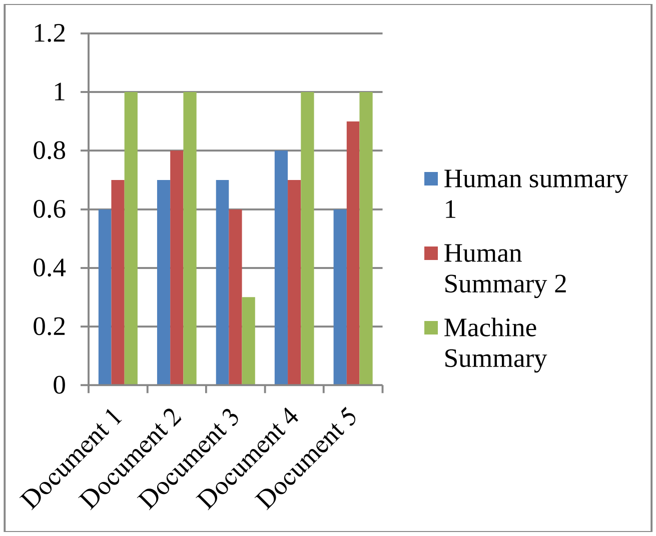


Figure 11: Results of Mahabharatha category

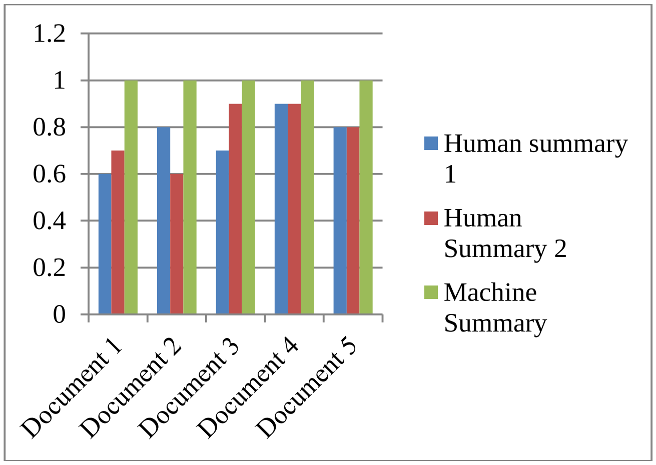


Figure 12: Results of Ramayana Category

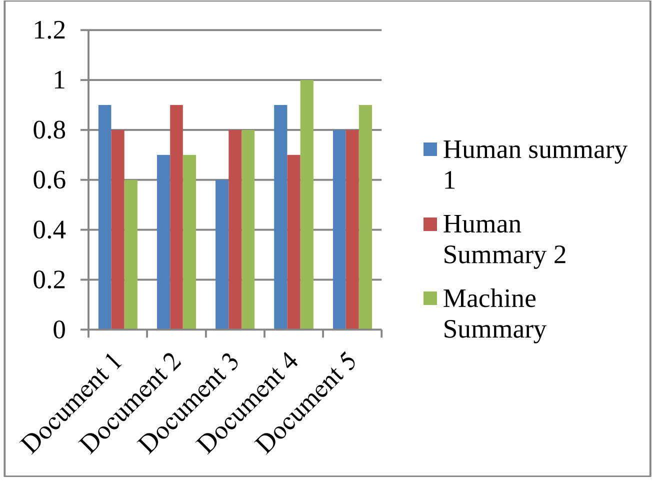


Figure 13: Results of Kannada Literature

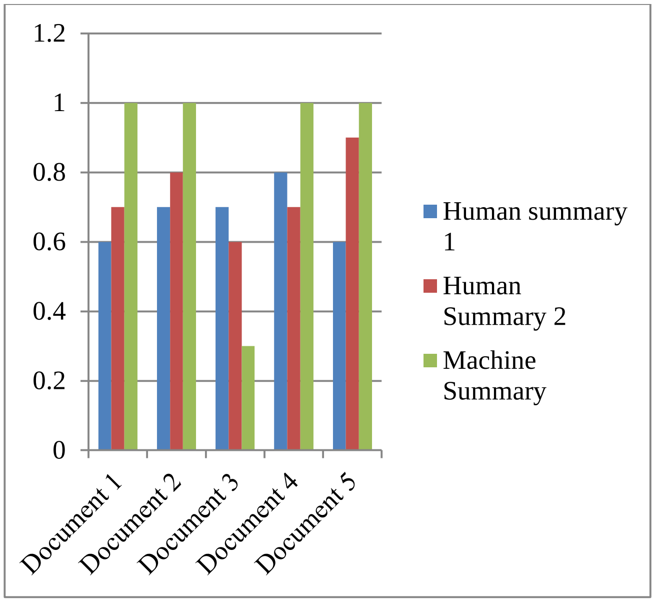


Figure 14: Results of Ramayana Category

Conclusion

In this paper, we have proposed a methodology for Kannada text document summarization. The ANN based method has given good results and also coherent summaries. The results indicate the fact that summarization is a challenging task and it is possible to render it effective if the methodologies incorporate coherence techniques. The production of a coherent summary matches the human summaries. Human summaries are found to be more effective. They will be much more effective if one has apriori knowledge of the domain. Therefore, there is scope for further research in this area.

The neural network could be trained according to the style of the human reader and to choose sentences which seem to be important in a paragraph. This, in fact, is an advantage our approach provides. Individual readers can train the neural network according to their own style. In addition, the selected features can be modified to reflect the user's needs.

The work at present focuses on only pre categorized data. But, the pre categorized data could influence human summarizer and hence when compared with machine summaries the results produced may not be effective.

RFERENCES:

[1] Ahmet Aker Trevor Cohn,'Multi-document summarization using A* search and discriminative training',Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 482–491, 2010.

[2] Eduard Dragut, Fang Fang, Prasad,Weiyi Meng, 'Stop Word and Related Problems in Web Interface Integration', Journal Proceedings of the VLDB Endowment,Volume2,Issue1, pages 349-360, 2009.

[3] Fumiyo Fukumoto Akina Sakai Yoshimi Suzuki,'Eliminating Redundancy by Spectral Relaxation for Multi-Document Summarization',Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, ACL 2010, pages 98–102, 2010.

[4] Feng Jin, Minlie Huang, Xiaoyan Zhu,'A Comparative Study on Ranking and Selection Strategies for Multi-Document Summarization', Coling 2010: Poster Volume, pages 525–533, 2010.

[5] Gabor Berend,Rich ́ard FarkasSZTERGAK : Feature Engineering for Keyphrase Extraction,Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 186–189, 2010.

[6] Galavotti L, F. Sebastiani, and M. Simi,'Experiments on the use of feature selection and negative evidence in automated text categorization' Proc. 4th European Conf. Research and Advanced Technology for Digital Libraries, Springer-verlag, pages 59-68,2000.

[7] Hitoshi Nishikawa, Takaaki Hasegawa, Yoshihiro Matsuo and Genichiro Kikui,' Opinion Summarization with Integer Linear Programming Formulation for Sentence Extraction and Ordering', Coling 2010: Poster Volume, pages 910–918, 2010.

[8] Julian Kupiec, Jan O. Pedersen, and Francine Chen:' A Trainable Document Summarizer', SIGIR, pages 68-73, 1995.

[9] Khosrow Kaikhah,'Text Summarization using Neural Networks',Faculty Publications ComputerScience, https://digital.library.txstate.edu/handle/10877/3819.

[10] Krysta M. Svore, Lucy Vanderwende, Christopher J.C. Burges,' Enhancing Single-document Summarization by Combining RankNet and Third-party Sources', Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 448–457, 2007.

[11] Letian Wang, Fang Li, SJTULTLAB: Chunk Based Method for Keyphrase Extraction, Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010,pages 158– 161, 2010.

[12] Michael J. Paul,ChengXiang Zhai,Roxana Girju,'Summarizing Contrastive Viewpoints in Opinionated Text',Proceedings of the 2010 Conference on Empirical

Methods in Natural Language Processing, pages 66–76, 2010.

[13] Marina Litvak and Mark Last ‘Graph-Based Keyword Extraction for Single-Document Summarization’, Proceedings of the 2nd Workshop on Multi-source, Multilingual Information Extraction and Summarization, pages 17-24, 2008.

[14] Rajesh S. Prasad, U.V.Kulkarni, and Jayashree.R.Prasad ‘Connectionist Approach to Generic Text Summarization’, World Academy of Science, Engineering and Technology, pages 340 - 345, 2009.

[15] Ronan Collobert, JasonWeston,‘Fast Semantic Extraction Using a Novel Neural Network Architecture’, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 560–567, 2007.

[16] Regina Barzilay and Michael Elhadad,`Using Lexical Chains for Text Summarization' ,In Proceedings of the Intelligent Scalable Text Summarization Workshop(ISTS'97),ACL, Proceedings of the ACL workshop on intelligent scalable text summarization, Volume 17,Issue1,pages 10-17 1997

[17] Teufel S, Moens M, ‘Argumentative classification of extracted sentences as a first step towards flexible abstracting’ I. Mani, M. Maybury (eds.), Advances in automatic text summarization, MIT Press, Advances in automatic text summarization ,pages 155-175,1999.

[18] Su Nam Kim, Ä Olena Medelyan,~ Min-Yen Kan} and Timothy BaldwinÄ,'SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles',Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 21–26, 2010.

[19] Tomek Strzalkowski: An Approach to Non-Singular Terms in Discourse, pages 362-364, 1986.

[20] Vishal Gupta, Gurpreet Singh Lehal,’ A Survey of Text Summarization Extractive Techniques’,Journal of Emerging Technologies in Web Intelligence, vol. 2, no. 3, pages 258-268, 2010.

[21] Xiaojun Wan, Huiying Li and Jianguo Xiao,'Cross-Language Document Summarization Based on Machine Quality Prediction',Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 917–926, 2010.

[22] Xiaojun Wan, Huiying Li and Jianguo Xiao,'Cross-Language Document Summarization Based on Machine Quality Prediction',Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 917–926, 2010.