I.J. Information Technology and Computer Science, **2014, 03, 38-46**
Published Online February 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.03.05

Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization

D.Y. Sakhare
Research Scholar, Bharati Veedyapeeth Deemed University, Pune, Maharashtra, India
E-mail: dysakhare@etx.maepune.ac.in

Dr. **Raj Kumar**
DRDO, Scientist 'D', DIAT, Khadakwasla, Pune, Maharashtra, India
E-mail: satya.may.jayate@gmail.com

Abstract— Recently, there has been a significant research in automatic text summarization using feature-based techniques in which most of them utilized any one of the soft computing techniques. But, making use of syntactic structure of the sentences for text summarization has not widely applied due to its difficulty of handling it in summarization process. On the other hand, feature-based technique available in the literature showed efficient results in most of the techniques. So, combining syntactic structure into the feature-based techniques is surely smooth the summarization process in a way that the efficiency can be achieved. With the intention of combining two different techniques, we have presented an approach of text summarization that combines feature and syntactic structure of the sentences. Here, two neural networks are trained based on the feature score and the syntactic structure of sentences. Finally, the two neural networks are combined with weighted average to find the sentence score of the sentences. The experimentation is carried out using DUC 2002 dataset for various compression ratios. The results showed that the proposed approach achieved F-measure of 80% for the compression ratio 50 % that proved the better results compared with the existing techniques.

Index Terms — Text Summarization, Dependency Grammar, Syntactic Structure, Feature Score, POS Tagger, DUC 2002

I. Introduction

Nowadays, enormous amount of digitally stored information is available on internet. In order to prevent sinking in it, filtering and extraction of information are necessary. A significant and opportune tool that assists and interprets huge quantities of text presented in documents is automatic text summarization (ATS).

The objective of ATS is to make a brief version of the original text with the most significant information at the same time retaining its main content and to enable the user to quickly comprehend huge quantities of information [1]. The summary should meet the major concepts of the original document set, should be redundant-less and ordered. These attributes are the basis of the generation process of the summary. The quality of summary is sensitive for those attributes relating to how the sentences are scored on the basis of the employed features. Consequently, the estimation of the efficacy of each attribute could result the mechanism to distinguish the attributes possessing high priority and low priority [1].

Single document summarization is the process of creating a summary from a single text document. Multi-document summarization shortens a collection of related documents; into single summary. User-focused summaries contain information most relevant to the initial search query; whereas generic summaries contain information about the overall perception of the document's content. Abstractive summary methods generate abstracts by examining and interpreting the text utilizing linguistic methods. Extractive summarization methods select the best-scoring sentences from the original document based on a set of extraction criteria and present them in the summary [2].

In this paper, we have presented a hybrid technique of text summarization with the combination of Dependency Grammar [3] and the sentence features.

Dependency Grammar is a type of linguistic theory, which constructs a syntactic structure based on the dependency relation between two words. The concept of dependency is that the syntactic structure of a sentence has a binary asymmetrical relation between the words of the sentence [3]. Many researchers have made an effort to find the linguistic text reduction techniques that maintain the meaning. These techniques differ drastically and some techniques are much tricky to implement than the others; however, all necessitate a fairly better syntactic analysis of the source text. These techniques are more robust to several prior systems because it involves a wide-coverage grammar, an efficient parser, and generation techniques. Syntactic trees have been constructed by the dependency grammar. These trees contain nodes that are related to the words of the sentence, and links between the nodes are labeled with grammatical relations (of the type “subject”, “direct object”, “subordinate clause”, “noun complement”, etc.). The grammar intends to do a complete syntactic analysis of the sentence. In case of failure that is, due to severe writer error or to restrictions of the grammar, it provides a series of incomplete analyses of fragments of the sentence [8]. However, determining the salient textual segments is only half of what a summarization system needs to do because, most of the time, the simple sequence of textual segments does not produce clear outputs [9]. Recently, many researchers have started to cope with the problem of creating articulate summaries by combining abstract and extraction based techniques.

Initially, the preprocessing steps are applied to the input document to extract the sentences and then, eight different feature score is computed for every sentence. Then the syntactic structure of every sentence is identified using the dependency grammar based POS tagger that is a necessary component of most text analysis systems, as it assigns a syntax class (e.g., noun, verb, adjective, and adverb) to every word in a sentence. Subsequently, two training matrices are generated with the help of feature score and the syntactic structure to train the neural network separately. For testing phase, every sentence is subjected to the feature extraction and structure extraction phase to generate the feature vector and the structure vector. Then, those two vectors are applied to the appropriate trained neural network to find the sentence score and sentence score obtained for every sentence from two neural networks are combined with the weighted average formula. Finally, sentences are selected as summary based on the sentence score obtained and the ordering process is carried out to make the summary in a meaningful manner.

The paper is organized as follows: Section 2 describes the review of recent works presented in the literature. Section 3 presents the contributions made in the proposed approach and section 4 presents the proposed approach for text summarization. Section 5 discusses the experimentation and discussion about the experimental results. Finally, conclusion is given in section 6.

II. Review of Related Works

Automated text summarization is an old eminent research area and dates back to the 1950s. As a result of the information overloading on the web there is large-scale interest in automatic text summarization during these days.

The early work on single-document summarization was done by Luhn [4]. He presented a method of automatic abstracting in the year 1958. This algorithm scans the original text document for the most important information. The features used here are word frequency and sentence scoring. Depending on a threshold value for important factors the featured sentences are extracted. The Weakness of this system is the summary produced lacks in quality. The system was restricted too few specific domains of literature. Baxendale [5] used sentence position as a feature to extract important parts of documents. Edmundson [6] proposed the concept of cue words. The strength of Edmondson’s approach was the introduction to features like sentence position in text, cue words and title and heading words.

Pollock [7] Used sentence rejection algorithm. The aim of the paper was to develop a system which outputs a summary conforming to the standards of the Chemical Abstracts Service (CAS).

The abstractive summary generation was pioneered by ADAM Summarizer [8]. Machine Learning frame work is used to generate summaries using sentence ranking. The strength of this approach was it’s potential to handle new domains in addition to redundancy elimination. K.R. Mc Keown in his thesis [8] generated the summary system using Natural Language Processing (NLP).The approach was based on a computational model of discourse analysis.

Truney and Frank[9], Mercer[10],Boguraev & Kennedy [11] all of them used key phrases extraction as a supervised learning task. For these systems a separate training document set with already assigned key phrases is required to function properly. This is again an open challenge for research community. [12] Presented Term Weighting and Sentence Weighting as important features to recognize the featured sentences. It has also addressed the problem of anaphora resolution.

Cut and Paste [13] is the first domain independent abstractive summarization tool. This was developed using sentence reduction and sentence combination techniques. Here a sentence extraction algorithm was implemented along with other features like lexical coherence, tf×idf score, cue phrases and sentence positions etc.

MEAD [14] was a multi document summarization toolkit it has used multiple position-based, TF×IDF, largest common subsequence, and keywords features. The methods for evaluating the quality of the summaries are both intrinsic (such as percent agreement, precision/recall, and relative utility) and extrinsic (document rank).A latest version of MEAD is based on centroid based multi document summarization. [15] Proposed a trainable summarizer based on feature selection and Support Vector Machine (SVM). [16] Has proposed keyword selection strategy. This is combined with the KFIDF measure to select the more meaningful sentences to be included in the summary. The Non-negative constraints used here are similar to the human cognition process. Evolutionary connectionist model for ATS is developed by [17] which is based on evolutionary, fuzzy and connectionist techniques. All the papers discussed above use various features for summary generation. The primary intention of this paper is to design and develop an efficient and hybrid approach for automatic text summarization. The proposed hybrid system combines the advantageous characteristics of both feature and syntactic structure-based methods to obtain better summary and at the same time, compactness of the sentences can also be preserved.

III. Proposed Syntactic and Sentence Feature-Based Hybrid Approach for Text Summarization

Initially, the input to the approach is a large document that has to be summarized. The document utilized for text summarization is prepared by a set of preprocessing steps namely, sentence segmentation, tokenization, stop words removal and word stemming. The preprocessed document is given to feature extraction, which involves the identification of significance features .Then, the syntactic structure of the extracted sentences is analyzed through the use of dependency grammar that converts the sentences into tokens/words and these tokens/words are connected using the dependency relations. These two techniques are effectively combined with neural network to obtain the final result that is a complete as well as a diminutive form of the input document. Fig.1 shows the block diagram of the proposed system.

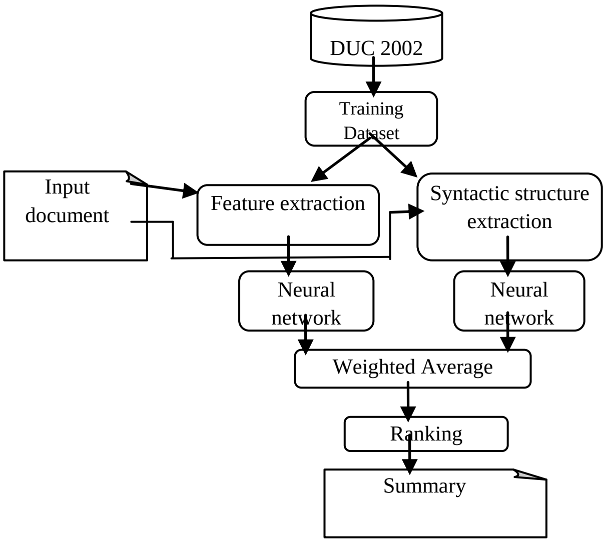

Fig. 1: The hybrid automatic text summarization system

3.1 Preprocessing

Preprocessing is the initial step involved in the system which is a three stage process consisting of sentence segmentation, removing stop words and, stemming. The output yielded after employment of preprocessing techniques is the individual sentences and their unique IDs which are obtained from the text document.

Segmentation results in separating the sentences in the document and will be useful for user to understand each individual sentence which is there in the document.

Stop words are removed from the document at the time of feature extraction step as they are considered as insignificant and include noise. Stop words are predefined words which are stored in an array and the array is made use of when the comparison with the words in the documents is carried out. The document consists of Individual words after the process in order to proceed with the word stemming process.

Word stemming transforms every individual word into its root base form. Word stemming basically removes the prefix and suffix of the concerned word to get the base form. This will in turn be used for comparison with other words.

3.2 Syntactic Structure-Based Neural Network

At first, the segmented sentence is given to the POS tagger for extracting the syntactic structure of the sentence. Every keywords of the sentence is tagged by the POS tagger that is used to extract the syntactic structure of the sentence. The dependency grammar constructs syntactic trees. These trees contain nodes which are corresponding to the words of the sentence, and links between nodes are labeled with grammatical relations (of the type “subject”, “direct object”, “subordinate clause”, “noun complement”, etc.). The grammar aims to do a complete syntactic analysis of the sentence. In case of failure that is, due to severe writer error or to limits of the grammar, it provides a series of incomplete analyses of fragments of the sentence. The theoretical tradition of dependency grammar is combined by the assumption that an important portion of the syntactic structure of sentences exist in binary asymmetrical relations holding between lexical elements. This dependency between the words in a sentence acts as the base for the compaction of the sentence and also maintains the sense of the original document. Also, the use of linguistic analysis for summarization purposes promises an increase in the efficacy.

3.3 Syntactic Structure Matrix for Training of Syntactic Structure-Based Neural Network

In our work, we have considered POS tagging defined in the dependency grammar for selecting the summary from a given document. Here, we have considered two different set of features for summary generation based on the perspective of considering dependency grammar. These two features consider the syntactic structure of particular sentence while finding the importance of the sentences.

3.3.1 Frequency of POS tags

Frequency of POS tags in a sentence is important for selecting summary since the frequency of tags is different for importance and unimportance sentence. When analyzing DUC 2002 summary, we can find the summary and unimportance sentence having the different frequency value for every POS tags. Here, we consider all the POS tags and its frequency for finding the importance of sentences. At first, we have taken the POS tags such as, noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. Then, we find the frequency of every POS tag in every sentence in the given document so that we can get the NxK matrix where, N is the number of sentence in a given document and 'K' is the number of tags.

3.3.2 Sequence of POS tags

Only considering the frequency of POS tags is not much significant for generating the summary. If we utilize the sequential placement of POS tags, the summary would be more precise and concise. By taking into consideration of the sequence, we have given number ID for every POS tag. Then, every sentence is converted to the number sequence by placing the number ID in the corresponding words.

The syntactic structure of the matrix is applied to the neural network to train the syntactic structure of the important sentences. Once the neural network trained with the important structure, the neural network can suggest the importance sentences based on the structures. Here, the syntactic structure matrix is represented as S with the size of NxK, in which N is the number of sentences in the document and $K-1$ is the number of unique POS tags identified by the POS tagger. Every element of the matrix is identified by finding the total number of corresponding POS tags presented in the sentence. But, the last vector column is the POS unique id sequence of the corresponding sentence. For finding the POS unique id sequence of a sentence, every POS tag is represented with the unique id and then the sequence of the unique id is generated for the POS sequence.

Training phase: The same multi-layer perceptrons feed forward neural network is also utilized here as learning mechanism, in which the back-propagation algorithm can be utilized to train neural networks. The back-propagation algorithm can be utilized successfully to train neural networks. Here, the input layer is an individual (syntactic structure vector) obtained from the step and the target output is zero or one that signifies whether its importance or not. Testing phase: In testing phase, the input text document is preprocessed and the syntactic matrix of input document is computed. The computed syntactic score is applied to the trained network that returns the sentence score of every sentence presented in the input text document.

IV. Feature-based Neural Network

After preprocessing, the input document is subjected to feature extraction by which each sentence in the text document obtains a feature score based on its importance. The important text features used in the proposed system are: (1) Format based score (2) Numerical data (3) Term weight (4) Title feature (5) Co-relation among sentence (6) Co-relation among paragraph, (7) Concept-based feature and (8) Position data.

4.1 Feature Computation

Format based score: Expressing the text in diverse format E.g. Italics, Bold, underlined, big font size and more in many documents shows the importance of the sentences. This feature never depends on the whole document instead to some exact single sentence. Score can assigned to the sentence considering the format of the words in the text. The ratio of the number of words available in the sentence with special format to the total number of words in the sentence offers one to form the format which is dependent relative on the score of the sentence.

Numerical data: The importance stats concerning the vital purpose of the document are usually shown by the numerical data within the sentence and this has its own contributions on the basic thought of the document that usually make way to summary selection. The ratio of the number of numerical data that happens in sentence over the sentence length is thus used to calculate the score for this feature.

Term weight: Term weight (1) is a feature value which is employed to look into the prominent sentences for summarizing the text documents. The term weight of a sentence is calculated as the ratio of the sentence weight (2) to the maximum sentence weight in the given text document. The sentence weight is the summation of the weight factor of all the words in a sentence. The weight factor (3) is the product of word frequency and the inverse of the sentence frequency (4).

$$TW = \frac{S_w}{\underset{i \in D}{Max}\left(S_w(i)\right)} \tag{1}$$

$$S_w = \sum_{j=1}^{n} W_j \tag{2}$$

$$W_i = TF \times ISF \tag{3}$$

$$ISF(t) = l\ og\ (\ N / N(\mathrm{T})) \tag{4}$$

Where, S_w → Sentence weight

W_j → Weight factor of the word in a sentence

n → Number of words in a sentence

TF → The number of occurrences of the term or word in a text document

ISF → Inverse Sentence Frequency

N → Total number of sentences in a document

$N(\mathrm{T})$ → Total number of sentences that contain the term (T)

Title features: A sentence is given a good score only when the given sentence has the title words. The intention of the document is shown via the word belonging to the title if available in that sentence. The ratio of the number of words in the sentence that occur in title to the total number of words in the title helps to calculate the score of a sentence for this feature.

Co-relation among sentence: At first, the correlation matrix C is generated in a size of NxM , in which N is the number of sentence and the M is the number of unique keywords in the document. Every element of the matrix is filled with zero or one, based on whether the corresponding keyword is presented or not. Then, the correlation of every vector with other vector (sentence with other sentence) is computed for all combinations so that the matrix of NxN is generated where every element is the correlation of two vector (two sentences). Then, every element of the row vector is added to get the sentence score.

Co-relation among paragraph: Here, the correlation is computed for every paragraph instead of sentences. for that, the correlation matrix C is generated in a size of PxM , in which P is the number of paragraph and the M is the number of unique keywords in the document. Every element of the matrix is filled with zero or one, based on whether the corresponding keyword is presented or not in the paragraph. Then, the correlation of every vector with other vector (paragraph with other paragraph) is computed for all combinations so that the matrix of PxP is generated where every element is the correlation of two vector (two paragraph). Then, every element of the row vector is added to get the score of every paragraphs and the score of every will obtain the same score of what its relevant paragraph obtained.

Concept-based feature: Initially, the concepts are extracted from the input document using the mutual information and windowing process. A windowing process is carried out through the document, in which a virtual window of size 'k' is moved from left to right until the end of the document. Then, the (5) is used to find the words that co-occurred together within each window.

$$MI(w_i, w_j) = \log 2 \frac{P(w_i, w_j)}{P(w_i) * P(w_j)} \qquad (5)$$

Where, $P(w_i, w_j) \rightarrow$ The joint probability that both keyword appeared together in a text window

$P(w_i) \rightarrow$ The probability that a keyword w_i appears in a text window.

The probability $P(w_i)$ is computed based on $\frac{|sw_t|}{|sw|}$, where sw_t is the number of sliding windows containing the keyword w_i and $|sw|$ is the total number of windows constructed from a text document. Similarly, $P(w_i, w_j)$ is the fraction of the number of windows containing both keywords out of the total number of windows. Then, for every concept extracted, the concept weight is computed based on the term weight procedure and the sentence score is also computed as per the procedure described in term weigh-based feature computation.

Position data: Position-based feature is computed with relevant to the sentence located in the document. With perspective of domain experts, initial sentence and the last sentence of the document is important than the other sentence. So, the maximum score is given for those sentences and the medium value is given to the sentence located in the starting and ending of every paragraph.

4.2 Feature Matrix for Training of Feature-Based Neural Network

This section describes the feature matrix used for training the feature-based neural network. The feature matrix is represented with the size of NxF , where N is the number of sentence and F is the number feature used in the proposed approach. (Here $F = 8$). Every element of the matrix is the feature score obtained for the corresponding sentence with the feature.

Training phase: Here, multi-layer perceptrons feed forward neural network is utilized for learning mechanism, in which the back-propagation algorithm can be effectively utilized to train neural networks. To train the neural network effectively, the input layer is an individual (feature vector) obtained from the feature computation steps and the target output is zero or one that signifies whether its importance or not. Testing phase: In testing phase, the input text document is preprocessed and the feature score of every sentence in the document is computed. The computed feature score is applied to the trained network that returns the sentence score of every sentence presented in the input text document.

V. Layered Neural Network

Here, layered neural network structure is designed by combining the feature-based neural network and syntactic structure-based neural network. The ultimate aim of combining two neural networks is to bring the feature as well as dependency grammar for text summarization. Without making use of syntactic structure of the sentences, the summary generated will not be concise and understandable manner. Furthermore, step used to combine the trained neural network is very important to generate summary by properly giving the importance to both feature as well as structure.

5.1 Combining Sentence Score

For the input text document, the sentence score obtained from both the neural network is combined with the following formulae. Here, weighted average formula (6) is used to combine the sentence score of both neural networks. The advantage of the weightage formula is that the different weights can be given to feature and syntactic structure with respect to its importance in text summarization.

$$S = \alpha * S^F + \beta * S^S \tag{6}$$

Where, S → Sentence score of the input sentence

S^F → Sentence score obtained by the feature-based neural network

S^S → Sentence score obtained by the syntactic structure-based neural network

α, β → Weighatge constants

5.2 Ranking of Sentence

Here, the ranking of sentence is carried out using the sentence score obtained from the previous step. Initially, sentences presented in the input text document are sorted in descending order according to the final sentence score. Then, the top-N sentences are selected for the summary based on the compression rate given by the input user. Finally, the selected top-N sentences are ordered in a sequential way based on the order of the reference number or unique ID to obtain the final summary.

$$N = \frac{C \times N_S}{100} \tag{7}$$

Where, N_S → Total number of sentences in the document

C → Compression rate

VI. Results and Discussion

This section describes the detailed the experimental results and it and analysis of the document summarization. The proposed syntactic and sentence feature-based hybrid approach is implemented in MATLAB (Matlab7.11)

6.1 DUC 2002 dataset

For experimentation, we have used DUC 2002 dataset [19] that contains documents on different categories and extractive summary per document.

6.2 Experimental Results

At first, the input document is given to the proposed hybrid approach for document summarization. Then, the feature score is computed for every sentence based on the features utilized in the proposed hybrid approach. The sample results obtained for the feature matrix is given in table 1. Subsequently, the syntactic feature is computed for the input text document those sample result is given in table 2. These two matrices are given to the two neural networks to obtain the sentence score. The final sentence score obtained from two neural networks are given in table 3. Here, the neural network is trained with the sentences available in the DUC 2002 and the corresponding target label is identified with the summary given in DUC 2002 dataset.

Table 1: Feature score for the text document (Cluster No. d071f and Document No. AP880310-0062)

Sentence ID	Feature score							
	F1	F2	F3	F4	F5	F6	F7	F8
1	0	0	0.2500	0.4002	0.0695	0.1850	0.2307	0.2500
2	0	0	0	0.5695	0.0044	0.1180	0.3283	0.2500
3	0.455	0	0	1.0000	0.3568	0.1640	0.5764	0.2500
4	0	0	0	0.3385	0.0141	0.0790	0.1951	0
5	0	0	0	0.2733	0.2838	0.0790	0.1575	0.2500
6	0	0	0	0.2470	0.6661	0.1386	0.1424	0
7	0.1000	0.1000	0	0.4426	0.0370	0.1386	0.2551	0.2500
8	0	0	0	0.5311	0.3792	0.4364	0.3062	0.2500

From the table 1, we can find that the feature score of sentences are varied between zeros to one. For the example in consideration, the fourth feature of the fourth sentence obtains the maximum value. Table 2 shows that F1 to F7 is the frequency-based POS tagging feature that provides the frequency of every POS tags for the particular sentence. F8, F9 and F10 are the sequence-based feature of the POS tags. When analyzing the table 3, some of sentences obtain approximately similar score for both the neural networks. But, we can find some difference especially, in fourth and fifth sentence. The combined feature score shows the improved importance level of every sentence in the given document [20].

Table 2: Syntactic structure-based feature score for the text document (Cluster No. d071f and Document No. AP880310-0062)

Sentence ID	**Syntactic structure-based feature score**									
	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**
1	0	8	0	0	3	0	4	616161114	114461	0
2	3	7	0	0	1	1	2	514061011	61011	0
3	5	15	0	0	2	0	5	611146114	0	1
4	1	6	1	1	2	0	1	246301411	111	0
5	0	2	3	1	4	1	2	661244224	3451	0
6	0	3	0	0	2	0	0	14411	0	0
7	1	5	1	0	5	1	1	244611411	44051	0
8	3	6	2	3	5	0	0	144201114	4	1

Table 3: Feature score of layered neural network

Sentence ID	**NN1 score**	**NN2 score**	**Combined score**
1	0.1437	0.1518	0.14775
2	0.1437	0.1391	0.1414
3	0.1437	0.1648	0.15425
4	0.1437	0.0991	0.1214
5	0.1437	0.0752	0.10945
6	0.0445	0.0747	0.0596
7	0.1437	0.1164	0.13005
8	0.1437	0.1045	0.1241

6.3 Performance Evaluation

6.3.1 Evaluation Measure

For performance evaluation, we have used the performance measure namely, precision, recall and F-measure. Precision measures the ratio of correctness for the sentences in the summary whereby recall is utilized to count the ratio of relevant sentences included in summary. For precision, the higher the values, the better the system is in excluding irrelevant sentences. On the other hand, the higher the recall values the more effective the system would be in retrieving the relevant sentences. The weighted harmonic mean of precision (7) and recall (8) is called as F-measure (9).

$$Precision = \frac{|\{\text{Retrieved sentences }\}\cap\{\text{Relevant sentences}\}|}{|\{\text{Retrieved sentences}\}|} \quad (7)$$

$$Recall = \frac{|\{\text{Retrived sentences}\}\cap\{\text{Relevant sentences}\}|}{|\{\text{Relevant sentences}\}|} \quad (8)$$

Where, Relevant sentences → Sentences that are identified in the human generated summary

Retrieved sentences → Sentences that are retrieved by the system

$$\text{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

6.3.2 Performance analysis

Here, the performance of the proposed hybrid approach is analyzed with the various compression ratios. The graphs (Fig.3, 4 and 5) show the precision, recall and F-measure for the proposed hybrid approach and the neural network 1. From the graphs, we can see that the proposed hybrid approach provides the greater F-measure compared with neural network. This concludes that the incorporation of dependency grammar-based syntactic structure into the feature-based text summarization provides the compact and brief summary compared with the previous feature-based summarization system.

Whenever the compression ratio is increased, the precision is also increased for three neural networks as per the graph shown in Fig.3. For the layered neural network, the precision value for the compression ratio (C=40) is 0.72 but, the precision value is 0.8 for the compression ratio, C=50. From the Fig.4, we can see that the recall is increased whenever the compression ratio is also increased for three neural networks. For the layered neural network, the recall value for the compression ratio (C=40) is 0.7 but, the recall value is 0.8 for the compression ratio, C=50. The F-measure value is 0.71 for the compression ratio (C=40) but, the F-measure value is 0.8 for the compression ratio, C=50.

6.3.3 Comparative analysis

This section presents the comparative analysis of the proposed system with the previous systems [17]. For compression ratio C=40, the proposed system achieved the precision of about 0.72 that is high compared with the precision of previous system. Similarly, recall and F-measure of the proposed system is also improved compared with the previous system. Table 5 gives the comparison in terms of precision, recall-measure for the compression ratio C= 50

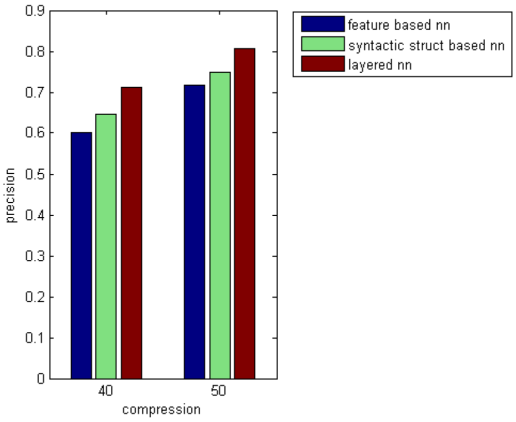

Fig. 3: Comparison graph for various ratio vs precision

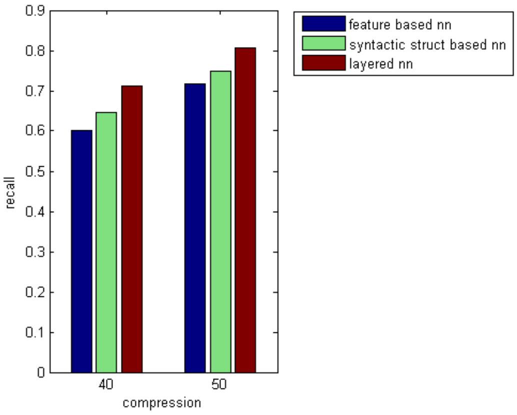

Fig. 4: Comparison graph for various ratio vs recall

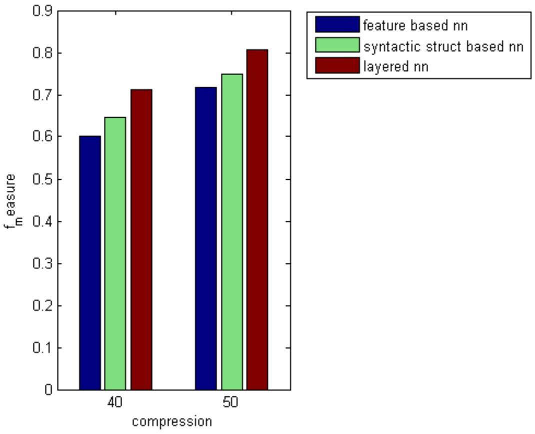

Fig. 5: Comparison graph for various ratios vs F-measure

Table 4: Comparison of compression ratio, C=40

Summarization system	Precision	Recall	F measure
Proposed system	0.72	0.7	0.71
Previous system [17]] (average values of from table 5-6)	0.865	0.45	0.59

Table 5: Comparison of compression ratio, C=50

Summarization system	Precision	Recall	F measure
Proposed system	0.8	0.8	0.8
Previous system [17] (average values of from table 5-6)	0.89	0.57	0.69

VII. Conclusion

We have presented a hybrid technique to text summarization with the combination of feature and syntactic structure. At first, neural network was trained based on the feature score obtained from the features including, (1) Format based score (2) Numerical data (3) Term weight (4) Title feature (5) Co-relation among sentence (6) Co-relation among paragraph, (7) Concept-based feature and (8) Position data, Then, the second neural network was trained with the syntactic structure of sentences. Finally, the two neural networks are combined with weighted average to find the sentence score of the sentences. The experimentation is carried out using DUC 2002 dataset for various compression ratios. The results showed that the proposed approach achieved F-measure of 80% for the compression ratio 50 %

References

[1] Automatic text summarization using sentence Features: a review, International J. of Engg. Research & Indu. Appls. (IJERIA). ISSN 0974-1518, Vol.4, No. IV, November 2011, pp. 31- 42.

[2] Oi Mean Foong, Alan Oxley and Suziah Sulaiman, 'Challenges and trends in automatic text summarization, IJITT, Vol. 1, Issue 1, 2010 pp 34-39'.

[3] Joakim Nivre, "Dependency Grammar and Dependency Parsing", In MSI report 05133, 2005.

[4] Luhn , 'The Automatic Creation of Literature Abstracts', IBM Journal April 1958 pp. 159–165.

[5] Baxendale, 'Machine-made Index for Technical Literature'An Experiment', IBM Journal of Research Development, Vol. 2, No.4, pp. 354-361, 1958.

[6] Edmundson, 'New Methods in Automatic Extracting', Journal of the Association for Computing Machinery, Vol 16, No 2, April 1969, PP. 264-285.

[7] Pollock and Zamora, 'Automatic Abstracting Research at Chemical Abstracts Service',Journal of Chemical Information and Computer Sciences, 15(4), 226-232,1975.

[8] Kathleen McKeown, 'Discourse Strategies for Generating Natural Language Text', Department ofComputer Science, Columbia University, New York, 1982

[9] Turney, Learning to extract keyphrases from text', technical report ERB-1057. (NRC#41622), National Research Council, Institute for Information Technology, 1999.

[10] Marcu, 'The automatic construction of large-scale corpora for summarization research'. In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, University of California, Berkeley, August 1999.

[11] Boguraev, Kennedy and Bellamy, 'Dynamic presentation of phrasal-based document abstractions', 32nd International Conference on System Sciences, 1999.

[12] Brandow, R., Mitze, K., Rau,' Automatic condensation of electronic publications by sentence selection'. Information Processing anagement,31(5):675-685, 1995.

[13] Radev, R., Blair-goldensohn, S, Zhang, Z., 'Experiments in Single and Multi-Docuemtn Summarization using MEAD'. In First Document Understanding Conference, New Orleans, LA, 2001.

[14] Jing, Hongyan and Kathleen McKeown., 'Cut and paste based text summarization'. In 1st Conference of the North American Chapter of the Association for Computational Linguistics , 2000

[15] Nadira Begum, Mohamed Abdel Fattah, Fuji Ren, 'Automatic text summarization using support vector machine', International Journal of Innovative Computing, Volume 5, pp 1987-1996, 2009.

[16] Rafeeq Al-Hashemi, 'Text Summarization Extraction System (TSES)Using Extracted Keywords', International Arab Journal of e-Technology, Vol. 1, No. 4, pp 164-168, 2010

[17] Rajesh Prasad, Uday Kulkarni, Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization' Journal of Computer Science, 2010 ISSN 1549-3636, pp1366-1376.

[18] Dipti..Sakhare, Rakjumar 'Syntactical Knowledge based Stemmer for Automatic Document Summarization', CIIT international journal of data mining knowledge engineering print: ISSN 0974 – 9683 & online: ISSN 0974 – 9578 Issue: march 2012 doi: dmke032012002.

[19] DUC.nist.gov/data.html

[20] Dipti..Sakhare, Rakjumar "Neural network based approach to study the effect of feature selection on document summarization" (IJET) ISSN: 0975-4024, Vol 5, Issue No 3, Jun-Jul 2013 pp 2585-2593

Authors' Profiles

Dipti Y Sakhare: Research scholar at Bharati Veedyapeeth, Deemed University, Pune, Maharashtra, India. Her areas of interest are: Digital systems, information retrival, VLSI Design.

Dr. Raj Kumar: He has completed his M. Sc.(Electronics) Degree in 1987 from University of Meerut, Meerut. He has been awarded M. Tech. and Ph. D degree in 1992 and 1997 respectively from University of Delhi, New Delhi. He worked at CEERI Pilani from 1993 to 1994 as a research associate. From May 1997 to June 1998, he worked as Assistant Professor in Department Electronics and Communications Engg, Vellore College of Engg. (Now VIT), Vellore. He worked in DLRL (DRDO), Hyderabad as Scientist from June 1998 to August 2002 and later on came in DIAT (DU) in Sept 2002. At present, he is Scientist 'E' in Department of Electronics Engg., DIAT (Deemed University), Pune. He established a Microwave and Millimeter Wave Antenna Laboratory in DIAT (DU), Pune and formulated thM. Tech. Programme in the Department of Electronics Engg. in 2010. He has written several technical papers in reputed International Journal and conferences.