Effect of Feature Selection on Small and Large Document Summarization

D.Y. Sakhare* Rajkumar**
*Research scholar Bharati Veedyapeeth deemed university, Pune, Address Maharashtra, India
**DRDO Scientist 'D', DIAT, Pune Maharashtra, India

Article Info

Article history:

Received May 22, 2014
Revised Aug 8, 2014
Accepted Aug 20, 2014

Keyword:

Abstraction
DUC 2002
Extraction
Features
Summarizers

ABSTRACT

As the amount of textual Information increases, we experience a need for Automatic Text Summarizers. In Automatic summarization a text document or a larger corpus of multiple documents are reduced to a short set of words or paragraph that conveys the main meaning of the text Summarization can be classified into two approaches: extraction and abstraction. This paper focuses on extraction approach.The goal of text summarization based on extraction approach is sentences selection. The first step in summarization by extraction is the identification of important features. In our approach short stories and biographies are used as test documents. Each document is prepared by pre-processing process: sentence segmentation, tokenization, stop word removal, case folding, lemmatization, and stemming. Then, using important features, sentence filtering, data compression and finally calculating score for each sentence is done. In this paper we proposed various features of Summary Extraction and also analyzed features that are to be applied depending upon the size of the Document. The experimentation is performed with the DUC 2002 dataset. The comparative results of the proposed approach and that of MS-Word are also presented here. The concept based features are given more weightage. From these results we propose that use of the concept based features helps in improving the quality of the summary in case of large documents.

Corresponding Author:

D.Y. Sakhare,
Research scholar Bharati Veedyapeeth Deemed University,
Pune, Maharashtra, India.
Email : diptiysakhare@gmail.com

1. 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].

Automatic text summarization is utilized in a variety of applications, including search engine hit summarization (summarizing the information in a hit list fetched by certain search engine); physicians' aids (to summarize and compare the prescribed treatments for a patient); creating the brief of a book and so on [3]. Best performance of automatic text summarization is achieved if the document is well-structured, for example news, reports, articles and scientific papers [4]. Normally, automatic document summarization accepts one or more source documents as input and provides an elegant summary as output to the user by extracting the gist of the source(s). The process consists of three phases, namely, analysis, transformation and synthesis. In the analysis phase, a small number of significant features are chosen by analyzing the input document. In the transformation phase a summary corresponding to the user's need is generated by transforming the output of the analysis phase. Features selected are significant factors that influence the overall quality of the summary. In this proposed work the effect of feature selection on summarization is evaluated.

The rest of the paper is organized as follows: Section 2 describes the review of recent works presented in the literature. Section 3 describes the pre-processing step. Section 4 presents the mathematical modelling for feature selection. Section 5 presents the results and discussion. Section 6 concludes the paper.

2. LITERATURE SURVEY

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 [3]. 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 [4] useded sentence position as a feature to extract important parts of documents. Edmundson [5] proposed the concept of cue words. The strength of Edmundson's approach was the introduction to features like sentence position in text, cue words and title and heading words [5].

Pollock [6] 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 [7]. 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 [7] generated the summary system using Natural Language Processing (NLP).The approach was based on a computational model of discourse analysis.

[11] Presented Term Weighting and Sentence Weighting as important features to recognize the featured sentences. It has also addressed the problem of anaphora resolution. Boguraev & Kennedy [10], Mercer [9] in 1997, Truney and Frank [8] in 1999, 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.

Cut and Paste [12] 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 [13] 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] 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. [14] Proposed a trainable summarizer based on feature selection and Support Vector Machine (SVM).Evolutionary connectionist model for ATS is developed by [16] which is based on evolutionary, fuzzy and connectionist techniques. All the papers discussed above use various features for summary generatin.Our aim in this paper is to perform the comparative study on the use of various features used for document summarization depending upon the size and type of the document. The following section describes the various steps in the proposed study.

3. PRE PROCESSING

The proposed automatic text summarization system consists of the following components:

1. Preprocessing
2. Feature extraction
3. Model building
4. Sentence selection and assembly

This section deals with the pre-processing phase. The input document can be of any document format (doc, txt, pdf, html, rtf), hence the system first applies document converters to extract the text from the input document.

3.1. Text Prologuing

Pre-processing the text before incepting to summarization and categorization is Text Prologuing. It consists of six phases which are listed in the following subsections.

3.1.1. Text Segmentation

Text Segmentation is the process of decomposing the given text into its constituent sentences, calculating each sentence length and word count. This module divides the document into sentences. At first glance, it may appear that using end of sentence punctuation marks, such as periods, question marks, and exclamation points, is sufficient for marking the sentence boundaries.

3.1.2. Normalization

Normalization is the process of converting words into normalized form. The following are the processes that come under normalization techniques.

3.1.3. Tokenization

It is the process of splitting of the sentence into words

3.1.4. Stop word Removal

During the retrieval of relevant information we have to remove few words, numbers, and special symbols etc., which have less significance. A new approach is used for stop word removal. The stop words are classified as useful and useless stop word and the removed accordingly. This will help in faster operations at later stemming stage.

3.1.5. Case Folding

Converting entire words in the sentences into lower case so as to avoid repetition of same word in different cases like sentence case, capital case, title case, upper case etc.

3.1.6. Stemming

Mechanically removing or changing the suffixes of some nouns or verbs. Stemming improves the retrieval performance because they reduce variants of the same root word to a common concept. It also reduces the size of the indexing structure because the number of distinct index terms is reduced. The design of a stemmer is language specific, and requires some significant linguistic expertise in the language. Here we proposed an integrated stemming approach which involves both rule based approach and dictionary based approach. The proposed integrated model showed better impacting results with respect to words affected and computing time [17].

4. MATHEMATICAL MODELLING FOR FEATURE SELECTION

After pre-processing, 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. The concept based feature is used for the first time.

4.1. Feature computation

Once the features are decided, one needs to prepare the mathematical model for their computation. The following subsections describe the mathematical computation of these features.

4.1.1. Format based score:

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.

4.1.2. 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.

4.1.3. Term weight

Term weight 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 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 is the product of word frequency and the inverse of the sentence frequency.

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

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

$$W_i = TF \times ISF$$

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

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)

4.1.4. 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.

4.1.5. 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.

4.1.6. 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.

4.1.7. Concept-based feature

Initially, the concept is 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 following formulae are used to find the words that co-occurred together within each window.

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

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.

4.1.8. 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.

5. 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.

5.1. Training phase

Here**,** multi-layer perceptrons feed forward neural network is utilized for learning mechanism, in which the back-propagation algorithm is 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 pre processed 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

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.

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

Where, N_S → Total number of sentences in the document

C → Compression rate

6. 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) and the experimentation is carried out with i5 processor having 3GM RAM.

6.1. DUC 2002 dataset

For experimentation, we have used DUC 2002 dataset [18] that contains documents on different categories and extractive summary per document.

6.2. Experimental Results

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

Sentence ID	Feature score							
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
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

Table 2. Neural network score

Sentence ID	Neural network score
1.	0.1518
2.	0.1391
3.	0.1648
4.	0.0991
5.	0.0752
6.	0.0747
7.	0.1164
8.	0.1045

At first, the input document is given to the proposed 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. This matrix is given to the neural network 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.

6.3. Performance 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 and recall is called as F-measure.

$$Precision = \frac{|\{\text{Retrieved sentences}\} \cap \{\text{Relevant sentences}\}|}{|\{\text{Retrieved sentences}\}|}$$

$$Recall = \frac{|\{\text{Retrived sentences}\} \cap \{\text{Relevant sentences}\}|}{|\{\text{Relevant sentences}\}|}$$

Where, Relevant sentences → Sentences that are identified in the human generated summary
Retrieved sentences → Sentences that are retrieved by the system

$$\text{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6.4. Performance analysis

As per the application of above features, the analysis shows that different types of documents require different combinations of features to get precise Summary. The summary evaluation is done on different documents of Standard DUC Foundation. Documents are categorized as type 1 and type 2 documents.

a. Type 1 documents

Documents about a single short story not more than 15 sentences.

b. Type 2 documents

Documents about a biography of a person more than 15 sentences and less than 50sentences. Sentences.

We have compared MS Word Summary and our proposed approach using all eight features. The precision (Figure 1), recall (Figure 2) and f-measure (Figure 3) for the two type of documents are evaluated. The results show that our proposed approach (using all eight features) outperforms the MS word summaries.

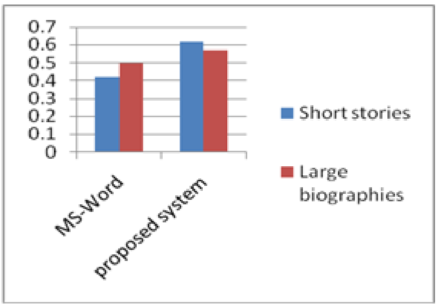

Figure 1. Effect on Precision

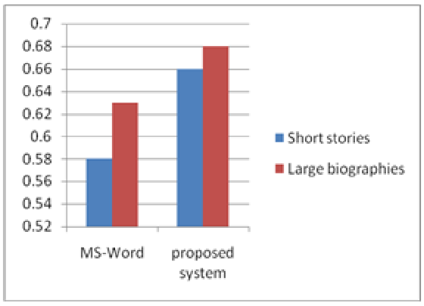

Figure 2. Effect on recall

Figure [1 -3] show that the proposed approach outperforms the MS word summaries. Figure 4 shows the effect of inclusion of concept based features on short and large documents.

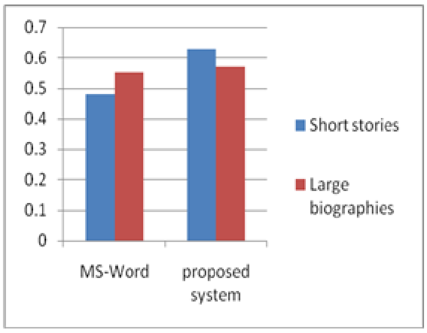

Figure 3. Effect on F measure

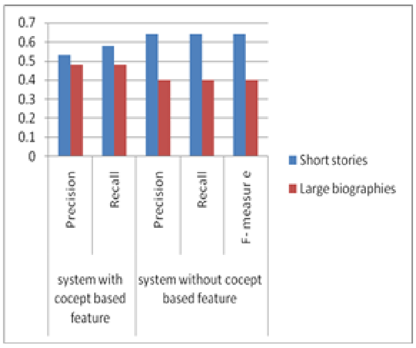

Figure 4. Comparison based on concept based feature

7. CONCLUSION

The results show that, summary generated using proposed module outperforms to that of MS-WORD module for all the performance parameters. The experiment is also carried out with consideration of concept based features and without concept based feature. The results, in figure 4, show that for large document summaries the concept based feature increases the quality considerably. So we can conclude that addition of the concept based features helps in improving the quality of the summary. These results achieved are a promising start toward further studies.

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] Luhn H.P, ‘The Automatic Creation of Literature Abstracts’, *IBM Journal* April 1958 pp. 159–165

[4] Baxendale, P. (1958), ‘Machine-made Index for Technical Literature’An Experiment‘, *IBM Journal of Research Development*, Vol. 2, No.4, pp. 354-361

[5] Edmundson H.P, ‘New Methods in Automatic Extracting’, *Journal of the Association for Computing Machinery*, Vol 16, No 2, April 1969, PP. 264-285

[6] J.J.Pollock and A. Zamora, ‘Automatic Abstracting Research at Chemical Abstracts Service’,*Journal of Chemical Information and Computer Sciences,* 15(4), 226-232(1975)

[7] Kathleen R. McKeown, ‘Discourse Strategies for Generating Natural Language Text’, Department ofComputer Science, Columbia University, New York, 1982

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

[9] Marcu, D. ‘*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

[10] B K Boguraev, C Kennedy, R Bellamy, ‘*Dynamic presentation of phrasal-based document abstractions’*, 32nd International Conference on System Sciences, 1999

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

[12] 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.

[13] 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

[14] 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.

[15] 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

[16] Rajesh Shardanand Prasad, Uday Kulkarni, Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization’ *Journal of Computer Science* 6(11):, 2010 ISSN 1549-3636, pp1366-1376.

[17] D.Y.Sakhare, Dr.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.

[18] duc.nist.gov/data.html

BIBLIOGRAPHY OF AUTHORS

D .Y.Sakhare is research scholar at Bharat Veedyapeeth,DeemedUniversity,Pune, Maharashtra,India. She is currently working as Assistant Professor in Department ofelectronics Engineering at MAE,Alandi,Pune. She has total Eight years teaching experience. Her teaching areas are: Digital systems,Information Retrival,VLSI Design

Raj Kumar was born on 14th May 1963 in Muzaffarnagar U.P., India. 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 the M. Tech. Programme in the Department of Electronics Engg. in 2010. He has written several technical paper in reputed International Journal and conferences.