CLUSTERING AND FEATURE SPECIFIC SENTENCE EXTRACTION BASED SUMMARIZATION OF MULTIPLE DOCUMENTS

A.Kogilavani[1] and Dr.P.Balasubramani[2]

[1]Department of CSE, Kongu Engineering College, Tamilnadu, India
vani_sowbar@yahoo.co.in
[2]Department of CSE, Kongu Engineering College, Tamilnadu, India
p_balu@kongu.ac.in

ABSTRACT

This paper presents an approach to cluster multiple documents by using document clustering approach and to produce cluster wise summary based on feature profile oriented sentence extraction strategy. Related documents are grouped into same cluster using document clustering algorithm. Feature profile is generated by considering word weight, sentence position, sentence length, sentence centrality, proper nouns in the sentence and numerical data in the sentence. Based on the feature profile sentence score is calculated for each sentence. According to different compression rates sentences are extracted from each cluster and ranked in order of importance based on sentence score. Extracted sentences are arranged in chronological order as in original documents and from this, cluster wise summary can be generated. Experimental results show that the proposed clustering algorithm is efficient and feature profile is used to extract most important sentences from multiple documents.

KEYWORDS

Feature Profile, Multi-Document Summarization, Sentence Extraction, Document Clustering.

1. INTRODUCTION

With the abiding development of online information, it has become progressively more essential to provide enhanced mechanisms to find and represent textual information effectively and efficiently. The vast amount of information available today has lead to information overload problem. Document summarization is one feasible key to handle this information overload problem. Document summarization is the process of taking a textual document, extracting content from it, and presenting the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs [1]. This process reduces the problem of information overload because only a summary needs to be read instead of reading the entire document. This can comprehensively help user to make out ideal documents within a short time by providing scraps of information.

Multi-document summarization is the process of producing a single summary of a set of related documents. In general two methods called extraction and abstraction are used to find out the summary from multiple documents. An extract summary consists of sentences extracted from document(s) while an abstract summary may contain words and phrases which do not exist in the original document(s) [2-3]. A document summary can also be either generic or query-dependent. A user-focused summary presents the information that is most relevant to the initial search query, while a generic summary gives an overall sense of the documents content. A generic summary should maintain a wide coverage of the document topics and keep low redundancy. In recent years several methods for automatic multi-document summarization have been proposed which deals with both approaches. The most common strategy used in these methods is the sentence extraction. The extraction process ranks and

extracts the representative sentences from multiple documents [4]. In sentence extraction strategy, clustering is used to avoid information redundancy resulted from the multiplicity of the source documents.

This paper discusses about feature profile oriented sentence extraction based summarization of multiple documents using document clustering approach. The clustering approach can be combined with the summarization technique in order to produce informative summaries [5]. This can be achieved by first performing a clustering of related documents and suppose if a user finds a particular cluster is interesting, then summarization is performed on that particular cluster only. The motivation behind automatic document clustering algorithm is to group similar documents into same cluster and new clusters are formed automatically or dynamically using threshold. To extract most important sentences the proposed approach generates feature profile by considering six sentence related features.

This paper is organized as follows. Section 2 discusses about related work on multi-document summarization. Section 3 presents an overview of the proposed approach. The experimental and evaluation measures are discussed in section 4. Finally section 5 concludes this paper and specifies future directions.

2. RELATED WORK

Document clustering has been widely applied to information retrieval systems for enhancing performance. Many clustering methods have been presented for browsing documents or organizing the retrieved results only for easy viewing [6]. Some researchers applied agglomerative clustering methods which start with all the documents as a separate cluster. At each step, the two most similar clusters are merged and this can be repeated until the desired number of clusters is obtained. But this method does not consider special properties of individual clusters so that it may make wrong merging decisions when noise is present.

Clustering is used to identify themes or subtopics of common information, because multiple documents relating to a particular topic are likely to contain redundant information in addition to information unique to each document [7]. Once themes have been known, a representative passage in each theme is selected and included in the summary. [8] Discusses about generation of an algorithm for information fusion, which merges similar sentences across documents to create new sentences based on language generation technologies. Although this approach can simulate, to some degree, portability is mere limitation of this approach.

In [9], researchers developed a multi-document summarizer, MEAD, which generates summaries using cluster centroid. It summarizes clusters of news articles automatically grouped by a topic detection system. MEAD uses Term Frequency-Inverse Document Frequency (*TF-IDF*) to calculate the weight for the word / term and three statistical features are used to select salient sentences. In automatic document summarization, the machine generated summary must be highly informative. To select informative sentences there is a need to include word sense along with word weight and more sentence specific features to calculate sentence score. To improve the accuracy of the word and the general importance of the word in the sentence, the proposed system adopts Term Synonym Frequency- Inverse Sentence Frequency (*TSF-ISF*) for calculating individual word weight. Sentence score is calculated for each sentence using six sentence specific features rather than three features in MEAD.

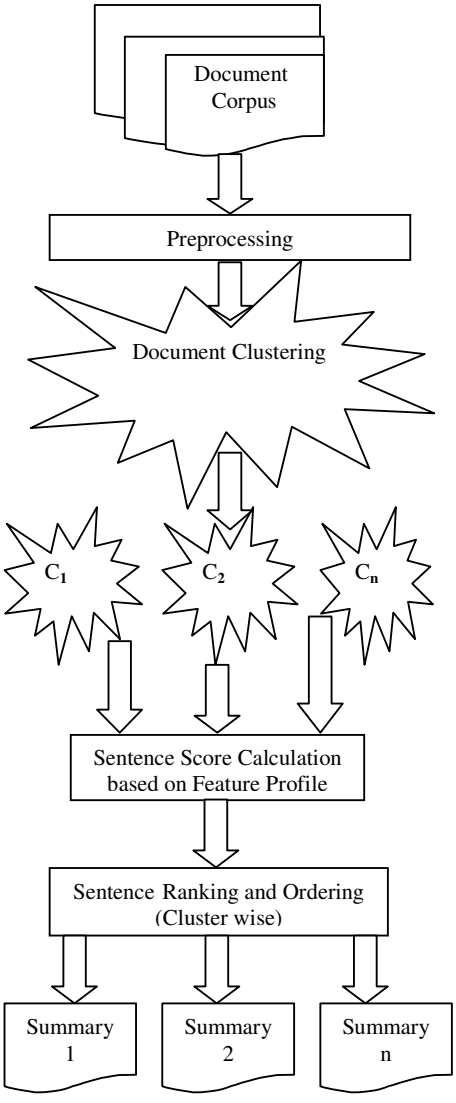
3. OVERVIEW OF THE PROPOSED APPROACH

Figure 1 illustrates an overview of the proposed approach for multi-document summarization system. The input to the system is a collection of documents. The output is a concise cluster-wise

summary providing the condensed information of the input documents. The proposed approach produces an extractive summary by selecting salient sentences from the documents cluster wise. All the relevant documents are grouped together into clusters by using threshold-based document clustering approach. Based on feature profile salient sentences from each cluster are identified and ranked according to their weights of importance. Based on the ranking of sentences, sentences are selected and ordered. The system then iteratively extracts one sentence at a time, until the required summary length is met for each cluster.

The proposed approach can be decomposed into five sub processes:

1. Preprocessing
2. Documents Representation and Clustering
3. Sentence Score Calculation based on Feature Profile
4. Cluster wise Sentence Ranking and Ordering
5. Summary Generation



3.1. Preprocessing

From the collection of documents, the boundaries of sentences are identified and the documents are split into sentences. Sentences are in turn split into words. Frequently

occurring insignificant words called functional words or stop words like “a”, “the”, “of” are removed because they do not contribute to the meaning of the sentence. Words are converted into their stems using enhanced Porter Stemmer algorithm.

3.2. Documents Representation and Clustering

After preprocessing documents are represented using vector space model. Let *D* be a collection of documents, *k* be the total number of documents in *D.* Each document has N number of sentences and collection of terms in each document is denoted as $d=\{t_1,t_2....t_m\}$. Each term in the document can be represented using a weighting scheme called *TSF-ISF.* The calculation of *TSF* involves synset extraction, comparison and term frequency calculation for each term. Synset extraction can be done with the help of WordNet which is a lexical database consisting of synonym sets for terms. The usage of *TSF* improves the quality of term weight. *ISF* is used to measure general importance of the term in the sentence. Term weight is calculated as

$$Term\_Weight(t_i) = \frac{TSF(t_i).ISF(t_i)}{n}$$

where *i=1,2…m. TSF* of each term is calculated as

$$TSF(t_i) = \sum_{t_i \in \{\{t\} \cup synonym(t)\}} \alpha.TF(t_i)$$

In *TSF* calculation to incorporate term synonym into account the *TF* of each term and it’s synonym is multiplied by α where α = 1 for the term and α = 0.5 for synonym of the term. *TF* is calculated as

$$TF(t_i) = \frac{n_j}{\sum_k n_k}$$

where n_j is the number of occurrences of the term j in document collection and the denominator is the number of occurrences of all terms in the document collection.
ISF is calculated as

$$ISF(t_i) = \log \frac{N}{n_i}$$

where n_i is number of sentences that contain term *i.* After calculating *TSF-ISF*, term-document matrix is constructed.

3.2.1. Document Clustering Algorithm

Input:
Document Collection *D*, Term-Document Matrix
Output:
Clusters with related documents
Steps:
(1) The first document is assigned to the first cluster and that cluster centroid is calculated by adding *TSF-ISF* values of all the terms in the document.
(2) The remaining documents are clustered using the following steps
(a) Similarity between each cluster centroid and one of the remaining documents is calculated using cosine similarity measure.
(b) If the similarity value is greater than the given threshold range for any cluster, then the document is placed in that cluster and the centroid of that cluster is updated by taking the mean value of *TSF-ISF* values of all the terms in the cluster.
(c) If not, the document is placed in a new cluster and *TSF-ISF* values of the terms in the document is added and the result is assigned as new centroid of that cluster.
(3) Repeat step (2) until all the documents are clustered.

3.3. Sentence Score Calculation Based on Feature Profile

Feature profile is generated to capture the values of sentence-specific features of all sentences. The proposed work combines a feature called term feature [10] with five features in [11] like sentence position, sentence length, sentence centrality, number of proper nouns in the sentence and number of numerical data in the sentence to generate feature profile.

3.3.1. Term Feature

Term Feature (*T_F*) is defined as

$$T\_F(s_{i,k}) = \sum Term\_Weight(t).f(t, s_{i,k})$$

where $f(t, s_{i,k})$ is the frequency of each term t in sentence $s_{i,k}$.

3.3.2. Position Feature

Always the first sentence of the document is most important. The position feature is defined by considering maximum positions of 3. For example, the first sentence in a document has a score value of 3/3, the second sentence has a score 2/3 and third sentence has a score value of 1/3. Position Feature (*P_F*) is defined as

$$P\_F(s_{i,k}) = \frac{Position(s_{i.k})}{3}$$

3.3.3. Sentence Length Feature

The Length Feature (*L_F*) is defined as

$$L\_F(s_{i,k}) = \frac{N * length(s_{i,k})}{length(d_k)}$$

3.3.4. Sentence Centrality Feature

The Sentence Centrality Feature (*C_F*) is defined as

$$C\_F(s_{i,k}) = \frac{words(s_{i,k}) \cap words(others)}{words(s_{i,k}) \cup words(others)}$$

3.3.5. Sentence with Proper Noun Feature

In general the sentence that contains more proper nouns is an important one and it is most probably included in the document summary. The following formula is used to calculate the inclusion of proper nouns (*PN_F*) in the sentence.

$$PN\_F(s_{i,k}) = \frac{PN\_Count(s_{i,k})}{Length(s_{i,k})}$$

3.3.6. Sentence with Numerical Data Feature

Naturally the sentence that contains numerical data is an important one and it is necessary to be included in the summary. The following formula is used to calculate the inclusion of numerical data (*ND_F*) in the sentence.

$$ND\_F(s_{i,k}) = \frac{ND\_Count(s_{i,k})}{Length(s_{i,k})}$$

The score of a sentence is the weighted sum of the scores for all terms in it. The following formula is used to calculate the score of the sentence

$$Sentence\_Score(s_{i,k}) = t_w.T\_F(s_{i,k}) + t_p.P\_F(s_{i,k}) + t_l.L\_F(s_{i,k}) + t_c.C\_F(s_{i,k}) + t_{pn}.PN\_F(s_{i,k}) + t_{nd}.ND\_F(s_{i,k})$$

where $t_w, t_p, t_l, t_c, t_{pn}, t_{nd}$ are weights of word, position, length, centrality, sentence with proper noun and numerical data features. These weights are given in order to normalize the values of sentence specific features such that $t_w + t_p + t_l + t_c + t_{pn} + t_{nd}$ must be 1. Here the values assigned for t_w is 0.3, t_p is 0.2, t_l is 0.2, t_c is 0.1, t_{nd} is 0.1, t_{pn} =0.1. A term t is keyword if

$$Term\_Weight(t) \geq 2 \times \frac{\sum_{t \in D} Term\_Weight(t)}{n_t(D)}$$

where $n_t(D)$ is number of terms in document cluster D.

3.4. Cluster wise Sentence Ranking and Ordering

Sentences are ranked according to their score values in descending order. After ranking the proposed system employs the following sentence ordering strategy according to their position and chronology of the original documents as specified in [9]. Suppose that $s_{i,1}$ and $s_{j,1}$ are the ith and jth sentence where $j>i$ in document d_1, $s_{r,2}$ and $s_{x,2}$ are the rth and xth sentence in document d_2. If d_1 is previous to d_2, these sentences should be ordered as $s_{i,1}$, $s_{j,1}$, $s_{r,2}$, $s_{x,2}$ and if d_2 is previous to d_1, should be ordered as $s_{r,2}$, $s_{x,2}$, $s_{i,1}$, $s_{j,1}$.

3.5. Summary Generation

After reordering all the sentences in each cluster, the summary is generated by extracting highly ranked sentences one at a time till the required summary length is met. In order to eliminate redundancy during summary generation, if the extracted sentence is already present in the summary then that sentence was eliminated and next highest ranking sentence is selected to form the summary. This process is repeated for each cluster and summary is generated depending upon various compression rates.

4. EXPERIMENTATION AND EVALUATION

The proposed system generates cluster-wise summary for the news genre. To form the experimental corpus 47 news articles are collected from web news portals. The proposed threshold-based document clustering algorithm is evaluated against centroid based clustering algorithm (MEAD). Experiment result shows that the proposed algorithm performs well. Human judges are asked to create summary for each cluster and it is evaluated against machine generated summary. The result shows that machine generated summary highly correlated with human summary.

The proposed system clusters 47 news articles into 13 clusters whereas MEAD clusters the articles into 14 clusters. Clustering quality is measured using precision rate which is calculated as

$$Precision Rate = \frac{A}{B}$$

where A is number of documents found by the clustering method and belonging to correct cluster, B is number of documents in the cluster.

Table 1 shows clustering precision rate of existing MEAD system and Table 2 represents precision rate of proposed document clustering algorithm. For the given query, MEAD system selects cluster 6 consists of 3 documents whereas the proposed system selects cluster 7 which consists of 4 documents all are related to given query. The result shows that even though both clusters have equal precision rate, the proposed system clusters all the related documents into same cluster whereas MEAD cluster the related documents into two different clusters**.**

Table 1. MEAD Clustering Precision Rate

Cluster No.	No. of Sentences	Precision Rate
1	5	60
2	6	66.67
3	5	40
4	1	100
5	2	50
6	3	100
7	10	20
8	1	100
9	1	100
10	4	50
11	1	100
12	3	100
13	1	100
14	4	75

Table 2. Proposed System Clustering Precision Rate

Cluster No.	No. of Sentences	Precision Rate
1	3	100
2	3	66.67
3	3	100
4	2	100
5	3	100
6	5	100
7	4	100
8	4	50
9	3	100
10	2	100
11	5	100
12	6	83.33
13	4	100

Table 3 shows sentence score calculation in MEAD system which uses only three features and sentence score calculation in proposed system which utilizes six sentence specific features. The values show that the proposed system gives high weight for sentences because of sentence specific features except for few sentences.

Table 3. Sentence Score Calculation

Document Id	Sentence Id	MEAD Sentence_Score	Proposed System Sentence_Score
1	1	2.135	4.457
1	2	1.783	3.593
2	1	2.135	5.427
2	2	0.882	3.313
2	3	1.293	2.640
2	4	0.720	1.263
2	5	0.977	2.521
2	6	1.233	5.179
2	7	0.887	2.812
2	8	0.994	3.607
2	9	0.949	3.400
3	1	3.335	3.591
3	2	1.588	5.088
3	3	2.095	4.371
3	4	2.978	3.230
3	5	1.605	3.665
3	6	2.040	3.387
3	7	0.817	3.551
3	8	1.706	4.918
3	9	2.511	5.238
3	10	2.268	4.314
3	11	0.444	3.630
3	12	2.458	3.456
3	13	1.687	2.560
3	14	1.443	1.116
3	15	1.425	3.087
3	16	0.504	0.518
3	17	1.992	3.976
3	18	0.242	0.148
3	19	2.256	2.479
3	20	1.936	4.043
3	21	0.262	0.298
3	22	0.470	0.256
4	1	1.003	5.875
4	2	1.240	3.320
4	3	1.852	4.871
4	4	1.330	5.400
4	5	0.887	2.999
4	6	1.122	3.838

4.1 Evaluation of Precision, Recall, F-Measure Parameters

Precision and Recall can be calculated using terms and keywords in the summary. Precision is a measure of exactness and recall is a measure of completeness. F-measure is a weighted harmonic mean of precision and recall.

4.1.1. Using Terms - Precision

Precision (P) is defined as the ratio of number of common terms in both manual and machine summary to number of terms in machine summary.

$$P = \frac{N_o}{N_m}$$

where N_o is number of common terms in both manual and machine summary, N_m is number of terms in machine summary.

4.1.2. Using Terms - Recall

Recall(R) is defined as the ratio of number of common terms in both manual and machine summary to number of terms in manual summary.

$$R = \frac{N_o}{N_h}$$

where N_o is number of common terms in both manual and machine summary, N_h is number of terms in manual summary.

4.1.3. Using Keywords - Precision

Precision is defined as the ratio of number of common keywords in both manual and machine summary to number of keywords in machine summary.

$$P = \frac{K_o}{K_m}$$

where K_o is number of common keywords in both manual and machine summary, K_m is number of keywords in machine summary.

4.1.4. Using Keywords - Recall

Recall is defined as the ratio of number of common keywords in both manual and machine summary to number of keywords in manual summary.

$$R = \frac{K_o}{K_h}$$

where K_o is number of common keywords in both manual and machine summary K_h is number of keywords in manual summary.

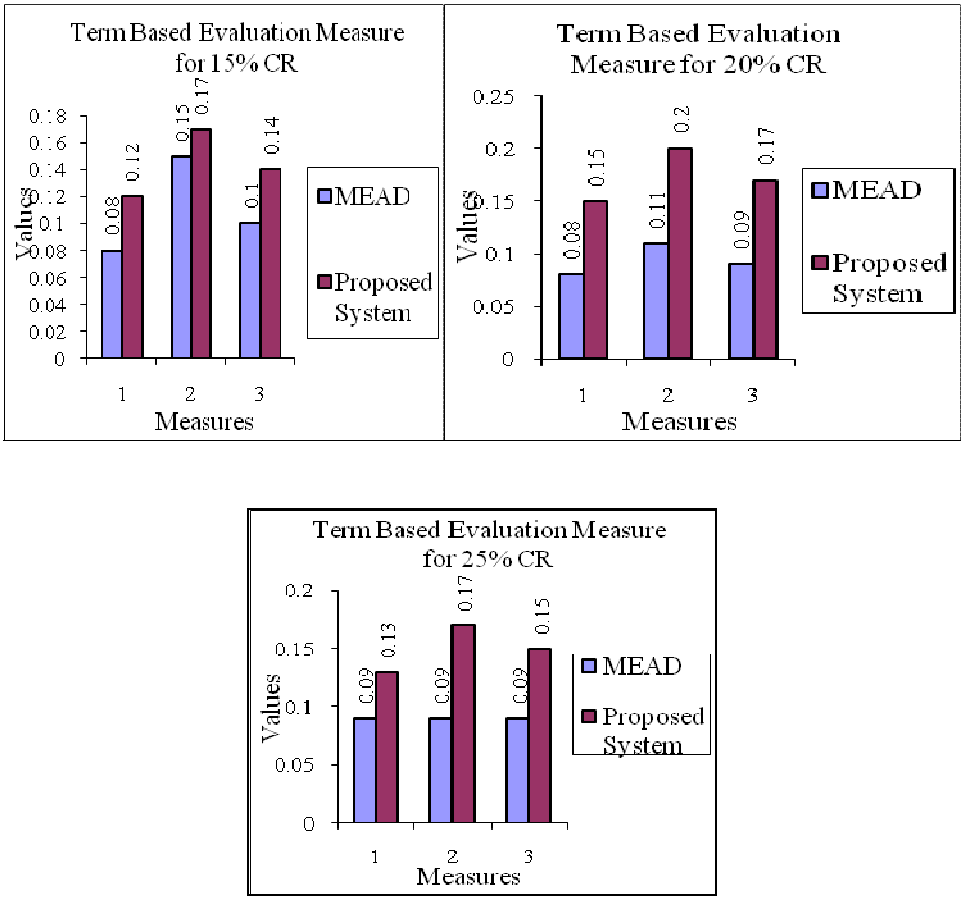
4.1.5. F-Measure

The formula for weighted harmonic mean of precision and recall is given by F_Measure(F_M) which is defined as

$$F\_M = \frac{2PR}{(P+R)}$$

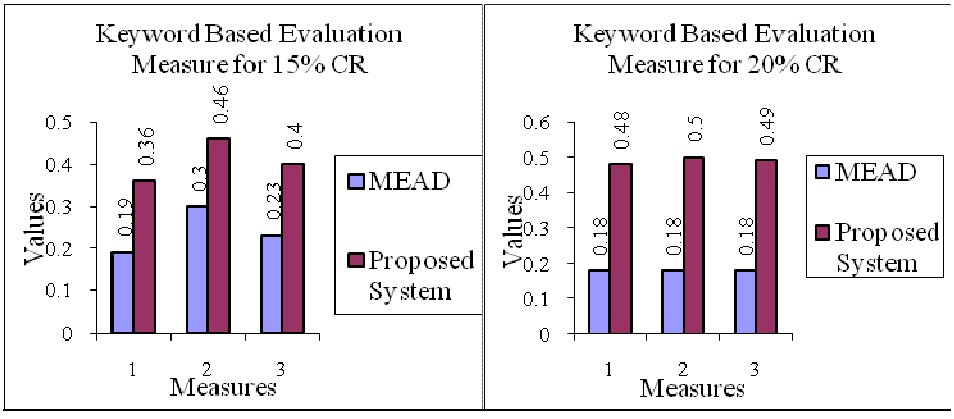
The following figure 2 denotes Term Based Performance Evaluation based on Precision, Recall, F-Measure for 15%, 20% and 25% Compression Rate (CR). For all the compression rates, the proposed system efficiency is high compared to existing MEAD system.

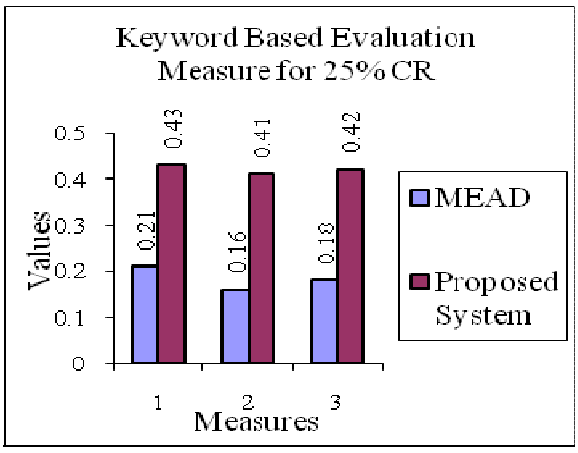
Figure 2. Term Based Performance Evaluation for 15%, 20% and 25% CR



The following figure 3 denotes Keyword Based Performance Evaluation based on Precision, Recall, F-Measure for 15%, 20% and 25% CR. For all the compression rates, the proposed system efficiency is high compared to existing MEAD system. At the same time the values are very high when the summary is evaluated based on keywords compared to terms.

Figure 3. Keyword Based Performance Evaluation for 15%, 20% and 25% CR





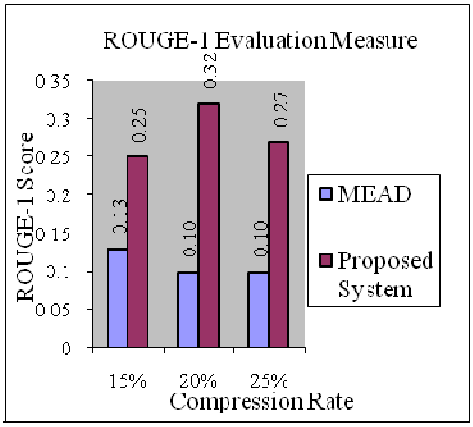
4.1.6. ROUGE-1 Score

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

$$ROUGE\_1\ Score = \frac{X}{Y}$$

where X is count of number of unigrams that occur in machine and manual summary and Y is total number of unigrams. The proposed method summary is compared against existing MEAD method by reevaluating summary generated by proposed system for different compression rates using *ROUGE-1 Score*. Human judge is asked to produce manual summary for *ROUGE-1 score* evaluation purpose. The following figure 4 compares *ROUGE-1 Score* of proposed system against MEAD. The result shows that by utilizing *TSF-ISF* and sentence specific features, the proposed system machine generated summary improves the accuracy of the summary.

Figure 4. ROUGE-1 Score for MEAD and proposed system



5. CONCLUSION AND FUTURE WORK

The proposed method discusses about grouping related documents using document clustering and cluster-wise summary generation using feature profile oriented sentence extraction strategy. Accuracy is improved by employing *TSF- ISF* measure. The summary generated using the proposed method is compared with human summary and its performance has been evaluated and the result shows that the machine generated summary coincides with the human intuition for the selected dataset of documents.

The future work includes implementation of the proposed system for more varied type of dataset with necessary changes to make it efficient. It has been planned to apply optimization techniques to produce optimal summary and also to implement the system in a grid like environment, to improve the speed of the system. This is necessary when the number of documents in the dataset is huge and processing time is very high. Thus it will improve the processing speed and efficiency of the proposed system.

REFERENCES

[1] Inderjeet Mani , (2001) "*Automatic Summarization"*, John Benjamins Publication.

[2] Inderjeet Mani, (1999) "*Advances in Automated Text Summarization*", MIT Press.

[3] Sun J, Shen D, (2005) "Web-page Summarization using Clickthrough data", *In proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[4] Fatma Kallel Jaoua, Maher Jaoua, (2008) "A Learning Technique to Determine Criteria for Multiple Document Summarization", *In proceedings of the WAS'00 Workshop on Automatic Summarization.*

[5] Manuel J, Mana Lopez, (2004) " Multi-document Summarization: An Added Value to Clustering in Interactive Retrieval", *ACM Transactions on Information Systems.*

[6] Lin C.Y., Hovy E, (2002) "NeATS in DUC 2002", *In Proceedings of the DUC 2002 workshop on Text Summarization.*

[7] Daniel N, Radev D, Allison T, (2003) "Sub-event based multi-document summarization", *In proceedings of the HLT-NAACL '03 workshop on Text Summarization.*

[8] Regina Barzilay, Michael Elhadad, Kathleen R.McKeown, (1999) "Information fusion in the context of multi-document summarization", *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics.*

[9] Dragomir R. Radev a, Hongyan Jing b and Małgorzata Stys, (2004) "Centroid-based summarization of multiple documents", *International Journal of Information Processing and Management.*

[10] Yan-Xiang He, De-Xi Liu,Dong-Hong Ji, Hua Yang, Chong Teng, (2006) "MSBGA: A Multi-Document Summarization System based on Genetic Algorithm", *In Proceedings of 5th International conference on machine learning and Cybernetics.*

[11] Mohamed Abdel Fattah, Fuji Ren, (2008) "Automatic Text Summarization". *In Proceedings of WASET.*

Authors

A.Kogilavani is currently working toward the Ph.D. degree in Document Summarization. Her research interests are Natural Language Processing, Knowledge Discovery, information Retrieval. She has authored 2 International Journals.

Dr.P.Balasubramanie received the Ph.D. degree in Theoretical Computer Science from Anna University in 1996. He has authored 6 books, 17 International Journals and 13 National Journals. He serves on the Editorial Board of the ACCST Research Journal. He was rewarded with CSIRJRF award.