Research Journal of Applied Sciences 7 (4): 229-233, 2012
ISSN: 1815-932X

Evaluation of Multi Document Summarization Techniques

[1]R. Nedunchelian, [2]R. Muthucumarasamy and [3]E. Saranathan
[1]Saveetha School of Engineering, Saveetha University, Chennai, India
[2]Sri Venkateswara College of Engineering, Pennalur, India
[3]Sastra University, 613402 Thanjavur, India

Abstract: Multi Document Summarization is carried out using MEAD extraction algorithm, Naive Bayesian classifier and genetic algorithm. The summary generated contains the selected sentences from each document and output them in the order prevalent in the original document, the order of the sentences in the summary may not be logical in occurrence. Hence to overcome this Timestamp concept is implemented. This gives the summary an ordered look, bringing out a coherent looking summary. Instead of taking up each sentence for comparison for summarization from all documents, it would be more than enough to summarize only the document (frequent document) which has been put to many numbers of readers. The Timestamp and Frequent document concepts are used to generate multi document summarization using MEAD extraction algorithm Naive Bayesian classifier and genetic algorithm and the results are compared and evaluated.

Key words: Timestamp, frequent document, compression rate, comprehensibility, readability, length of summary

INTRODUCTION

Automatic summarization (Goldstein *et al*., 2000) is the process of taking information source as the frequently used documents, 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 need. Based on the type of the input text the summary generators are classified as single or multi document summary generators. The two major differences between single and multiple document summarizations (Radev *et al*., 2000, 2001), first most approaches to single document summarization involve extracting sentences from the document, the multi document summarization involves methods that merge information stored in different documents and if possible, contrast their differences. Second, most single document summarization systems, to a certain extent, make use of the monolithic structure of the document. Multi document summarization systems usually rely less on the structures of the documents.

TIMESTAMP

The summary produced by summarization algorithms contains the selected sentences from each document and output them in the order prevalent in the original document. Sentences selected from the first document will appear before the sentences selected from the second document, similarly selected sentences from the second document will appear before the sentences selected from the third document and subsequently. The order of the sentences in the summary may not be logical in occurrence. Hence, to overcome this short coming the concept of Timestamp is implemented. The implementation of Timestamp is carried out by assigning a value to each sentence of the document depending on the chronological position in which it occurs in the document. Once the sentences are selected they are arranged in the ascending order depending on the Timestamp. This gives the summary an ordered look, bringing out a coherent looking summary (Ferreira, 2006).

FREQUENT DOCUMENTS SUMMARIZATION

Instead of taking up each sentence for comparison for summarization from all documents, it would be more than enough to summarize only the document which has been put to many numbers of readers. Since, researchers track for the document which is read frequently by many people, it is supposed to provide all the necessary information about the topic to the user so the user need not surf through other documents for information as the document in hand would be satisfactory. A frequently used multi document summarization using MEAD algorithm is developed and results are used to compare

Corresponding Author: R. Nedunchelian, Saveetha School of Engineering, Saveetha University, Chennai, India

the results of the summary from frequently used documents using Naive Bayesian classifier and genetic algorithms (Nedunchelian, 2008; Nedunchelian *et al*., 2009).

GENERATION OF SUMMARY USING MEAD

The final summary is generated based on the score. After processing the documents by using the MEAD (Radev *et al*., 2000) the sentences are arranged based on the score. High score sentence will appear first, next high score sentence and so on. The summary produced by MEAD contains the selected sentences from each document and output them in the order prevalent in the original document. Sentences selected from the first document will appear before the sentences selected from the second document, similarly selected sentences from the second document will appear before the sentences selected from the third document and subsequently. The order of the sentences in the summary may not be logical in occurrence. Hence, to overcome this shortcoming we have implemented a concept called Timestamp (Nedunchelian *et al*., 2009). The implementation of Timestamp is carried out by assigning a value to each sentence of the document depending on the chronological position in which it occurs in the document. Once the sentences are selected they are arranged in the ascending order depending on the Timestamp. This gives the summary an ordered look, bringing out a coherent looking summary.

SUMMARY GENERATED BY MEAD

The performance for summarization of the input documents using MEAD and Bayesian classifier has been analyzed and compared with frequent documents using MEAD and bayesian classifier. Totally there are 100 documents. Among them 10% of documents are selected as frequent documents for processing using MEAD. The score tables for frequent document summarization are shown in Table 1.

From the Table 1 it is understood that when MEAD is applied on the frequent documents the time taken to get the summary is 25 sec which is less than the time taken to summarize all the documents (Radev *et al*., 2000, 2001, 2004; Nedunchelian *et al*., 2011).

Table 1: Score table for frequent documents using MEAD

No. of documents	Time (sec)
4	5
7	8
9	16
10	25

GENERATION OF SUMMARY USING NAIVE BAYESIAN CLASSIFIER

Keywords are useful tools as they give the shortest summary of the document. Keywords are used rarely in the documents. Kupiec *et al*. (1995) proposes a method of training a Bayesian classifier to recognize sentences that should belong in a summary. The classifier estimates the probability that a sentence belongs in a summary given a vector of features that are computed over the sentence. Researchers propose a frequently used multi document summarization system with user interaction that would extract a summary from frequently used documents using Naive Bayesian classifier with supervised learning. Bayesian classifier works with an assumption that the feature values are independent. With this assumption, researchers can compute the probability that a word is a key given its TF*IDF score (T), the distance to the beginning of the paragraph (D), paragraph where the word is present (PT) and the sentence that it exists in (PS) by using Bayes Theorem (Nedunchelian *et al*., 2010):

$$P(\text{key}|T, D, PT, PS) = \frac{P(T|\text{key})\,P(D|\text{key})\,P(PT|\text{key})\,P(PS|\text{key})\,P(\text{key})}{P(T, D, PT, PS)}$$

$$P(T, D, PT, PS) = \sum \frac{P(T|\text{key})\,P(D|\text{key})\,P(PT|\text{key})\,P(PS|\text{key})}{P(\text{key})}$$

Where:

- $P(\text{key})$ = The prior probability that a word is a key
- $P(T|\text{key})$ = The probability of having TF*IDF score T given the word is a key
- $P(D|\text{key})$ = The probability of having distance D
- $P(PT|\text{key})$ = The probability of key with respect to the paragraph
- $P(PS|\text{key})$ = The probability of key with respect to the sentence
- $P(T, D, PT, PS)$ = The probability that a word having TF*IDF score T, neighbor distance D, position in the text PT and position in the sentence PS

Researchers have selected eight related documents and applied Timestamp based algorithm.

SUMMARY GENERATED BY NAIVE BAYESIAN CLASSIFIER

Totally there are 100 documents. Among them 10% of documents are selected as frequent documents for

Table 2: Score table for frequent documents using Naive Bayesian classifier

No. of documents	Time (sec)
4	4
7	10
9	14
10	18

Table 3: Score table for frequent documents using genetic algorithm

No. of documents	Time (sec)
4	3
7	8
9	12
10	15

processing using Naive Bayesian classifier. The score table for document summarization is shown in Table 2 (Nedunchelian *et al.*, 2010, 2011; Kupiec *et al.*, 1995).

From the Table 2 it is understood that when the Naive Bayesian classifier is applied on the frequent documents the time taken to get the summary is only 18 sec which is less than the time taken to summarize all the documents.

GENERATION OF SUMMARY USING GENETIC ALGORITHM

The modes of operation in genetic algorithm are training phase and testing phase. In training phase, features are extracted from original text and learning process is also started. A weighted score function is given in each sentence. In the testing mode, learned patterns are used to generate the summary. A set of highest score sentences are chronologically specified as a document summary based on the compression rate (Ferreira, 2006).

Totally there are 100 documents. Among them 10% of documents are selected as frequent documents for processing using genetic algorithms. The score table for document summarization is shown in Table 3.

From the Table 3, it is understood that when the genetic algorithm is applied on the frequent documents the time taken to get the summary is only 15 sec which is less than the time taken to summarize all the documents.

The following Fig. 1 shows the time taken by MEAD, Navie Baseyan classifier and genetic algorithm for summarizing the given multi documents. From the Fig. 1 it is understood that the time taken for summarization by genetic algorithm is lesser than others.

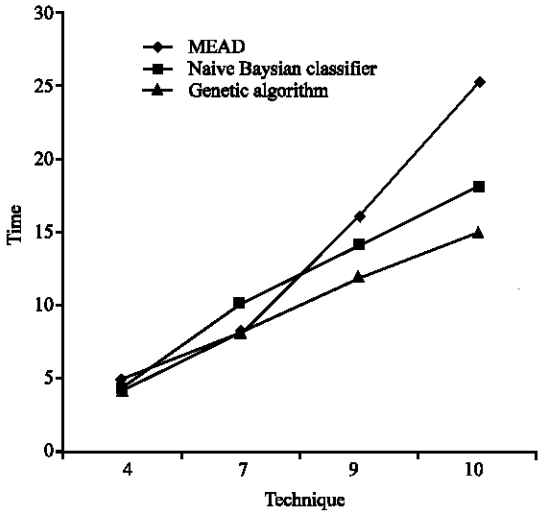

Fig. 1: Comparison of MEAD, Naive Bayesian classifier, and genetic algorithm

EVALUATION OF THE SUMMARY

Multi document summarization is carried out using MEAD algorithm, Naive Bayesian classifier and genetic algorithm for frequent documents with Timestamp. This is compared with human generated summary by human assessors consists of 5 professors, 9 lecturer and 6 students to find out whether the output is an informative summary which can be a substitute of the original documents. In order to satisfy this following points are considered important and assessed by human assessors:

Comprehensibility: The summary should include main content of target documents exhaustively.

Readability: The summary should be a self contained document and should be readable.

Length of summary to be generated: The length of summary generated by human and system is compared. Each participant generated summaries according to topic information given and submitted a set of summaries. Each topic corresponds to one IR result which consists of the following information:

- Topic ID
- List of keywords for query in IR
- Brief description of the information needs
- Set of documents IDs which are target documents of summarization. The number of documents varies from 3-20 according to the topic
- There are two lengths of summary, long and short. The length of long is twice of short

Table 4: Comparison of length of summary generated by human and system

	Length of summary generated by MEAD		Length of summary generated by Bayesian classifier		Length of summary generated by genetic algorithm	
Human assessor	Short	Long	Short	Long	Short	Long
Professor	3	2	4	1	5	-
Lecturers	6	3	7	2	8	1
Students	5	1	5	1	6	-

Table 5: Evaluating the comprehensibility of summary generated by system

	Comprehensibility by MEAD		Comprehensibility by Bayesian classifier		Comprehensibility by genetic algorithm	
Human assessor	Yes	No	Yes	No	Yes	No
Professor	4	1	4	1	5	Nil
Lecturers	8	1	9	Nil	9	Nil
Students	4	2	4	2	6	Nil

Table 6: Evaluating the readability of summary generated by system

	Readability by MEAD		Readability by Bayesian classifier		Readability by genetic algorithm	
Human assessor	Yes	No	Yes	No	Yes	No
Professor	5	Nil	5	Nil	5	Nil
Lecturers	9	Nil	9	Nil	9	Nil
Students	6	Nil	6	Nil	6	Nil

COMPARITIVE ANALYSIS OF SUMMARY GENERATED

About 20 human assessors are used to carry out the comparative study of summary generated by system and human in terms of comprehensibility, readability and the length of the summary.

The Table 4 shows information about comparison of length of summary by human assessors and system. From the Table 4 it is observed that the summary generated by MEAD, Bayesian classifier and genetic algorithm are shorter than summary produced by human as 70, 80 and 95% of the human assessors stated that the length of the summary generated by the system is short when produced by the three algorithms, respectively.

From the Table 5, it is observed that the summary generated by MEAD, Bayesian classifier and genetic algorithm are having all important contents as 80, 85 and 100% of the human assessors stated that the length of the summary generated by the system is comprehensible.

From the Table 6, it is observed that the summary generated by MEAD, Bayesian classifier and genetic algorithm are easy to read as 100% of the human assessors stated that the summary generated by the system is readable when produced by all the three algorithms.

From this analysis done using human assessors it is proved that the summary generated by the system is short and the quality of the summary generated is also goodbased on the two factors readability and comprehensibility.

CONCLUSION

Timestamp and Frequent document concept have been successfully implemented using MEAD, Bayesian classifier and genetic algorithm to generate the multi document summary. The results are evaluated, from this analysis done using human assessors it is proved that the summary generated by MEAD, Bayesian classifier and genetic algorithm is short and the quality of the summary generated is also good based on the two factors readability and comprehensibility.

REFERENCES

Ferreira, C., 2006. Genetic Programming: Mathematical Modeling by an Artificial Intelligence. 2nd Edn., Springer-Verlag, Germany.

Goldstein, J., V. Mittal, J. Carbonell and M. Kantrowitz, 2000. Multi-document summarization by sentence extraction. Proceedings of the NAACL-ANLP Workshop on Automatic Summarization, April 2000, Seattle, WA., USA., pp: 40-48.

Kupiec, J., J. Pedersen and F. Chen, 1995. A trainable document summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 9-13, 1995, New York, USA., pp: 68-73.

Nedunchelian, R., 2008. Centroid based Summarization of multiple documents Implemented using time stamps. Proceedings of the International Conference on Emerging Trends in Engineering and Technology, October 2008, Nagpur, India, pp: 480-485.

Nedunchelian, R., R. Muthucumarasamy and E. Saranathan, 2009. Multi document text summarization techniques. Int. J. Adv. Res. Comput. Eng., 3: 91-100.

Nedunchelian, R., R. Muthucumarasamy and E. Saranathan, 2010. An approach of the naive bayesian classifier for the summarization of frequently used documents implemented using timestamps. Int. J. Adv. Res. Comput. Eng., 4: 53-60.

Nedunchelian, R., R. Muthucumarasamy and E. Saranathan, 2011. Comparison of multi document summarization techniques. IJCSNS Int. J. Comput. Sci. Network Secur., 11: 155-160.

Radev, D.R., H. Jing and M. Budzikowska, 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility based evaluation and user studies. Proceedings of the Conference on ANLP/NAACL and Workshop on Summarization, April 12, 2000, Seattle, USA, pp: 21-30.

Radev, D.R., H. Jing, M. Stys and D. Tam, 2004. Centroid based summarization of multiple documents. J. Info. Processing Manage., 40: 919-938.

Radev, D.R., S. Blair-Goldensohn and Z. Zhang, 2001. Experiments in single and multi document summarization using MEAD. Proceedings of the10th EMNLP Conference on Empirical Methods in Natural Language Processing, October 9-11, 2010, New Orleans, LA, USA.