Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 14 (2020) Pp-32-34

BBRC
Bioscience Biotechnology
Research Communications

Extractive Multi-Document Text Summarization by Using Binary Particle Swarm Optimization

Archana Potnurwar[1*], Anjusha Pimpalshende[2], Shailendra S. Aote[3] and Vrushali Bongirwar[4]

[1]Priyadarshini Institute of Engineering & &Technology, Nagpur, India
[3,4]Shri Ramdeobaba College of Engineering and Management, Nagpur, India
[2]CMR College of Engineering & Technology, Hyderabad, India.

ABSTRACT

The absence of a standard dataset and poor work for Hindi text summarization leads to develop a technique for better results. We have used a combination of Title feature, Sentence length, Sentence position, Numerical Data, Thematic word, Term frequency and Inverse Sentence Frequency for finding the results. Binary PSO is used for finding the optimal values of the features.

KEY WORDS: MULTI-DOCUMENT, TEXT SUMMARIZATION, SWARM INTELLIGENCE.

INTRODUCTION

Now a days, most of the information is searched through the Internet, as it is used as superior information retrieval tool like any search engine. Because of the large increase in the information on the internet and busy schedule of every individual, the summaries information is very important for the user. Summary of any text documents helps to easily understand the concepts and conclude something good out of it. A summary that is created by a human is called manual [Discussion] summarization, whereas, a summary that is created by the machine is called automatic text summarization [ATS].

ATS are typically divided into different approaches. In some techniques, input documents are used for text classification further used for the summary generation. The difference between single and multidocument summarization is that the first one uses only one text file for the summary generation, whereas second approach uses more than one text files, probably related to each other in some context. Two types of Summarization methods are mainly found in the literature: extraction and abstraction. An extractive summarization deals with selecting the most important sentences from the source documents and combining them into a summary. Abstractive summarization is a summary, at least some of whose material is not present in the input. Multi-document summarization is also one of the areas, which is used in large scale information retrieval.

Though many same techniques are shared between single and multi-document summarization, there are at least three ways by which they differ.

1. The degree of redundancy of the information present in topically similar documents is much higher than a single document. So the use of the anti-redundancy method is more preferred in multi-document summarization.
2. If the single document summarization demands 15% summary generation, still the multi-document summarization demands nearly the same number of sentences in the final summary. It means, for

ARTICLE INFORMATION

Received 15th Oct 2020 Accepted after revision 27th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)
A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.
Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.14/8

10 document summarization it generates 1.5% sentences in the summary. So the compression ratio plays a very important role in a multi-document summary generation.

3. The biggest problem is a coherence problem in multi-document summarization.

In general, finding a multi-document summary is difficult. Common steps used in multi-document summarization are preprocessing, feature extraction, single document summary generation followed by final summary generation.

literature review: A variety of either extractive or abstractive multi-document summarization (MDS) techniques has been developed in recent years. Extractive summarization is about finding scores of each sentences in the documents and selecting the sentences with highest scores. Abstractive summarization is complex w.r.t extractive summarization, because it involves sentence selection, it's fusion, compression and finally all these need to be reformulated. In this study, we focus on extractive summarization. This section aims to present an overview of the basics and types of multi-documents text summarization.

X Wan proposed a novel extractive approach based on the manifold ranking for topic-focused multi-document summarization. Yih and Suzuki tried to give a simple scheme where they first assign a score to each term in the document cluster, using only frequency and position information, and then find the set of sentences in the document cluster that maximizes the sum of these scores, subject to length constraints. Goldstein proposed a multi-document summarization based on a single document summarization by using relational similarity among documents. Most of the available extractive methods generates summary by considering one by one document. Because of which, structure patterns amongst the sentences are less redundant.

Ruifang He proposed a Group Sparse learning framework is proposed for the summarization, where learned group information is used for minimizing the error which in further reconstruct the original documents. A bottom-up approach is proposed for multi-document summarization to capture the association and order of two textual segments by Bolegala D, which is based on chronology, topical-closeness, precedence, and succession. Ordering sentences according to their publication date is also considered to be a superior method for the multi-document summarization. Chronological ordering improvement is proposed by Okazaki (2004).The unsupervised approach based on optimization is proposed by R. M. Alguliyev for automatic document summarization. M Xi, J Sun, W Xu proposed an improved quantum-behaved PSO algorithm by finding weighted mean of best positions. These positions are finding based on the fitness value of each particle.

This algorithm is much faster than other algorithms and better global convergence. A Fuzzy Inference System is proposed by S.Babar, P. Patil for extractive text summarization. This method selects the most relevant sentences and words for summarization. A Support vector-based regression model was proposed by Y. Ouyang, Qin Lu for the sentence ranking in query-focused multi-document summarization. Text summarization for the Nepali language is proposed by Sarkar S. based on the hybrid PSO and k-means clustering technique. The Nepali word net is used for summarization. Inter-cluster similarity and intra-cluster similarity are used as the measure for the performance evaluation of the algorithmA hybrid model of symmetric non-negative matrix factorization (SNMF) and sentence-level semantic analysis (SLSS) is proposed for multi-document summarization. SNMF divides the sentences into groups then SLSS findsrelationships between sentences.

Summerization Algorithm: Hindi is also widely used language in some part of the world, mostly in India. Hindi is normally spoken using a combination of 52 sounds - 12 vowels, 35 consonants, nasalization and a kind of aspiration. Preprocessing of the documents plays a very important role in data mining applications for better results. Preprocessing generally involves three tasks: Tokenization, Stop word removal and stemming. The next step in the summarization is feature extraction. Features are used to extract salient sentences from the text. In the literature, more than 10 features are suggested. Any number of combinations of those features is used in the summarization techniques.But based on the literature, only six features are found to be suitable for a summary generation. Therefore,in this study six features are selected to score each sentence in the document. These features are Title feature [TF], Sentence Length [SL], Sentence Position [SP], Numerical Data [ND], Thematic Word [TW], TermFrequencyInverse Sentence Frequency [TfIsf].

After calculating the value of each feature for each sentence, it is necessary to calculate the overall value of each sentence. The numbers of documents are taken as an input to the algorithm. For each document, initial feature values are calculated. These values are passed to the Binary PSO. At the end of maximum iterations, it sends the optimum feature values back to the algorithm. Add all six feature values for each sentence, to get the final score of each sentence. Sort the sentences in increasing order of the final score. The next step is to find the similarity of each sentence to other sentences to remove redundancy in the final summary. This similarity is simply the final value comparison. If the similarity is more than 70%, then remove the duplicate sentence from the final summary. It is also necessary to arrange the text in the summary coherently.

RESULTS

We have created political news data set, consisting of three documents for each news. We have 5 types of news in each category. Recall, precision and f-measure are used to test the performance automatic summarization.

Recall = (human-generated summary ∩ Automatic summary) / (human-generated summary)

Precision = (human-generated summary ∩ Automatic summary) / (Automatic summary)

F Measure = (2 X precision X recall) / (precision + recall)

Table 1. Shows the precision, recall and f-measure values for the datasets.

News No.	No of sets of Documents	Precision	Recall	F measure
1	2	0.8	0.73	0.76
	3	0.875	0.8	0.84
2	2	0.81	0.8	0.8
	3	0.78	0.76	0.77
3	2	0.825	0.795	0.81
	3	0.79	0.76	0.77
4	2	0.82	0.79	0.8
	3	0.8	0.785	0.79
5	2	0.785	0.775	0.78
	3	0.83	0.82	0.824

CONCLUSION

The multi-document summarization for Hindi documents is a major issue due to very poor work is performed in this direction. We have proposed an algorithm to solve the problem. Six features are used to find the weight value of each sentence. Binary PSO is used to find the optimum value of each feature.

REFERENCES

J Goldstein, V Mittal, 2000, Multi-Document Summarization By Sentence Extraction", pp 40-48,.

Shailendra S. Aote, Raghuwanshi M M, Malik L, "A New Particle Swarm Optimizer with Cooperative Coevolution for Large Scale Optimization", Springer- AISC, Vol.327, ISBN NO: 978-3-319-1933-5,14th -15th Nov 2014, pp 781-790.

Shailendra S. Aote, Raghuwanshi M M, Malik L 2016 Particle Swarm Optimization based on winner's strategy", LNCS Vol. 9873, pp. 201–213.

X. Wan, J. Yang and J. Xiao, 2007, Manifold-Ranking Based Topic-Focused Multi-Document Summarization", IJCAI-07, pp 2903-2908.

W. Yih 2007, "Multi-Document Summarization by Maximizing Informative Content-Words", IJCAI-07, pp 1776-1782.