IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-9 E-ISSN: 2347-2693

Multi Document Summarization using Cross Document Relations

Yogita Desai[1*] and P. P. Rokade[2]

[1*,2]Department of Computer Engineering, University of Pune, India

www.ijcseonline.org

Received: Sep /06/2015 Revised: Sep/13/2015 Accepted: Sep/24/2015 Published: Sep/30/ 2015

Abstract— Multi-document summarization refers to the process of automatic extraction of text from multiple sources which belong to same topic. With the increase in usage of internet large amount of data has been generated day by day. It is quite difficult for anyone to distinguish and summarize this vast information gathered from various sources. Multi document text summarization has solution for this problem. Multi document summarization assembles information from different sources and summarizes the information up to necessary length. In this paper preprocessing is applied to unprocessed documents and different features are extracted. And then CST relations are identified from these extracted features document. Finally summary is generated depending on identified CST relations.

Keywords—*Multi Document Summarization, CST Realtions, Feature Extraction, Extractive Summarization.*

I. Introduction

For text summarization many research studies have been proposed in last few decades [1] [2]. The text summarization is informative describing precisely about entire contents of document, or can be considered as indicative when it is intimately linked with user's question [3]. In addition to this text summarization can be extractive or abstractive. Abstractive type of summarization collects original sentences from source documents process them and then the sentences are incorporated in absolute summary preserving the relevance of information. The study described by Gupta-lehal[1]and Kumar et.al.[2] takes into consideration extractive summarization in which key sentences are recognized and incorporated in summary. That means absolute summary is considered which comprises of sentences that are originally from the source documents [2]. Key sentences are determined by statistical as well as linguistic features of sentences. Word frequency measure is commonly computed by TF-IDF factor. For example, the input text document may contain the word 'CST' many times, so count the number of occurrences of the word 'CST' and that is considered to be word frequency and most frequently occurring words are ignored in case of TF-IDF. In news editorial if for any incidence time and date is specified then that can be considered as statistical information.

A further issue for summarization is the amount of information that is going to be processed. For example, in Ultimate Research Assistant text mining is carried out on internet search results to summarize, assist and categorize them and make it simple for the user to do online research [4]. Thus there is need for MDS (multi document summarization) for gathering multiple source text into a small, precise text. By considering the fact that if the documents are topically related then the documents have semantically associated information.

Based on this fact CST relations among the texts are identified. D. R. Radev proposed that multi document summarization can be smoothly progressed by analysis of relevant documents using CST model [5]. CST model can be represented as a cluster. Clusters of multi-documents are characterized by two data structures multi document cube and by multi document graphs. These data structures are defined at different levels such as word, phrase, and paragraph and document level. General process of multi document summarizations is described in fig.1

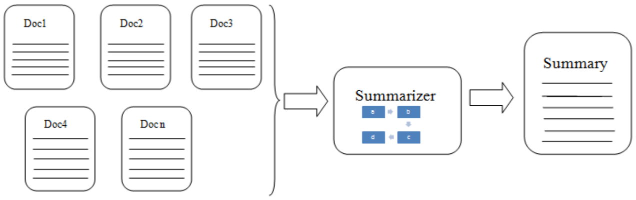

Fig 1: General Process of Summarization

In the proposed work, first step is to preprocess the document. Pre-processed document is then used as input to feature extraction. Feature extraction is the process of identifying keys in the document which is done with the help of six different features. In the next step, CST relations are identified. Then based on identified CST relations relevant sentences are included into final summary.

II. Motivation

The necessity of automatic text summarization has currently risen because of rise of information on the Internet. With the accessibility and internet speed, information search from online documents has been eased down to user's finger tips. However, it is not easy for users to manually summarize those large online documents. For example, when a user searches for information about earthquake which occurred in Sendai, Japan, the user will probably receive enormous articles related to that event. The user would definitely opt for a system that could summarize those articles. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning. Information overload has created an acute need for summarization. Typically, the same information is described by many different online documents. Hence, summaries that synthesize common information across documents and emphasize the differences would significantly help readers. Such a summary would be beneficial, for example, to a user who follows a single event through several news wires.

III. Literature Survey

An original notable effort in the area of automatic text summarization is by H. P. Luhn (1958). H. P. Luhn projected that number of occurrences of specific word in a source document is a helpful measure of relevance for text summarization on single document. Edmundson included two methods to generate summary [1969]. First method makes use of number of occurrences of word i.e. word frequency and second method deals with the heading of source document. The key sentences were scored by these features to incorporate them into summary. Jing presented a sentence diminution system for eliminating unrelated idiom like prepositional phrases, clauses from sentences [2000].

Hsun-Hui Huang proposed fuzzy-rough approach by examining features of sentence from conceptual space and then applying fuzzy-rough logic to identify significant sentences [6]. Depending upon this conceptual space various features of sentences are defined. These features are used to form feature space in which every sentence will be treated as an entity. Conceptual relationship is articulated by natural languages are intrinsically fuzzy, fuzzy approximation space is formed by rough theory and fuzzy set. Significant sentences are identified by computing sentence membership to the estimation of source texts. CPSL and LESM are two methods which are proposed by Md. Mohsin Ali. CPSL method is a mixture of MEAD and SimWithFirst methods [7]. MEAD is the extractive summarizer based on centroid and sentence scoring is done with the help of sentence level and inter-sentence level features. The features used in this method are centroid, position, length. In SimWithFirst method, every sentence is checked for similarity with first sentence. The second method, LESM comprised of CPSL and LEAD. LEAD allocates a score of 1/n to each sentence, where n is the number of sentence in specific document. Sentences with least value are not included in summary.

Other than very clear distinction in text input size, numerous other factors make the complication in MDS than single document summarizer. For example different source document comes from number of locations, from different authors and having different styles, even if they are relevant to the topic. Another fact that is to be considered as different source documents can be from different time frames or they may reflect the information which conflicting from each other. So multi document summarizer must deal with all these issues. Therefore summarizer must be designed in different way than the single document summarizer. Radev [2000] proposed the CST theory showing that CST can be a basis for cross document relations. And also CST relations are presented, based on RST (Rhetorical Structure Theory) RST is typically used for individual document. 24 CST relations are described having linking at different levels such as word (W), phrase (P), paragraph (PR) and document level (DOC). Fig 2 shows different levels of summarization. Out of the 24 CST relations identity, subsumption, overlap and description are considered in [2] as it covers most of other relations.

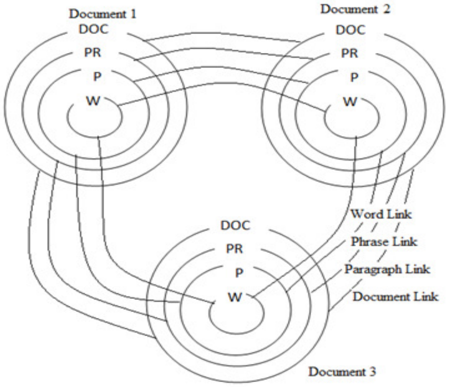

Fig 2: MDS graph at different levels

The quality of extractive summary is affected by CST relationships into consideration. Different kinds of CST relations have different effects on final summary [8]. As described in introduction, the cross document relations which are present amongst relevant documents are defined by CST model. Considering this fact, advantages of CST relations for summarization are addressed by numerous researchers. Zhang et al. stated that, the effect of enhancing CST is dependent on which CST relation is added into summary [8]. Jorge and Pardo scrutinized summarization based on CST. Methods based on content selection are proposed to generate inclination based and basic summary [9]. Major constraint of the mentioned researches are human experts are needed to manually identify the CST relations.

In the proposed work, this constraint is delighted by recognizing the relations amongst the sentences from the source texts. Z. Zhang et al [10] proposed boosting classification algorithm based on text in English, in which CST relations among sentences are recognized. But the classifier demonstrates the approximate average values of 46% precision, 33% recall and 36% F-measure showing poor performance in classification. As final result of the system is based on performance of classifier, the performance of classifier must be capable enough to see its effect in summarizer.

IV. Proposed System

A. *Work Breakdown Structure:*

Fig. 3 shows Work Breakdown Structure of the proposed system.

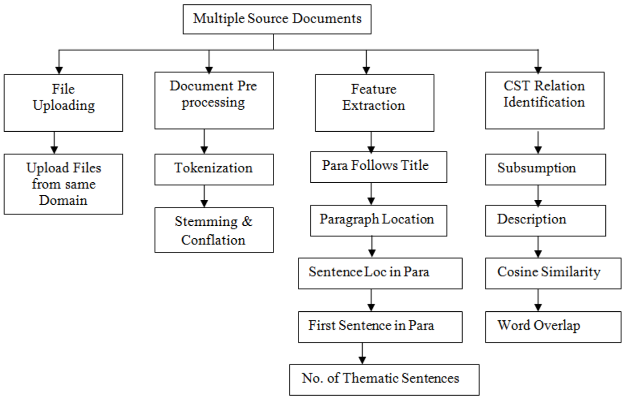

Fig 3: Work Breakdown Structure

B. *Document pre-processing:*

Pre-processing of the document means to produce structured information which is ready for processing. Segmentation is done to divide the input contents into number of sentences. For these sentences further common word removal and stemming algorithm is applied so that the tokens can be recognized from the input source documents.

C. *Feature Extraction:*

As the input information is too huge to be processed, the input information is transformed into a condensed representation of vector or the set of features. This process of converting input information into a set of features is named as feature extraction. For the feature extraction process, six different features are taken into consideration [11]. The vector [f1, f2.... f6] is considered for these six features. The feature selection plays a vital role in deciding the variety of sentences that will be chosen in final summary.

Table 1: Feature Vector

Feature	Description
F1	Paragraph Follows Title
F2	Location of paragraph in document
F3	Location of sentence in paragraph
F4	First sentence in paragraph
F5	Length of sentence
F6	Number of thematic words in sentence

.

D. *CST Relation Identification:*

Cross document relations are identified to include highly relevant sentences into summary. Four types of CST relations are considered viz. description, partial equivalence, subsumption and identity as these relations cover other relations in CST model. Cross document relation identification with the help of manually annotated text can require time period and resources. Inspired by this fact, a sentence pair is formed from all the input documents and from these pairs CST relations are identified. Table 2 describes CST relations used in proposed system.

Table2: CST Relations used in Proposed System

Relation Type	Level	Description
Description	P	First sentence describes an entity in second sentence
Partial Equivalence	P, DOC	First sentence provides some facts in second sentence (not all facts are provided in first sentence)
Subsumption	P, DOC	First sentence contains all information in second sentence including additional information which is not in second
Identity	Any	Same text appears in first and second sentence

V. Results And Discussion

To obtain the summary, common word removal, special words removal process should be done. And for every document unique words are identified. Different features or combination of features are applied to the input text documents. The system will generate the summary according to features selected. This summary is taken as an input for CST relation identification. And according to the selection of CST relations summary is generated. The result is checked for six features and four CST relations. Graph is shown against number of characters in input files and number of characters in summary for every relation. The result is shown for three different documents. Table 3 shows the values of number of characters in input files and number of characters in summary after selecting each CST relation individually and by selecting all CST relations.

Table 3: Summary for Different Documents

Files	Chars	All features	Identity	Subsumption	Overlap	Descritption	All CST
D1	224	518	303	35	215	180	216
D2	380						
D3	307						

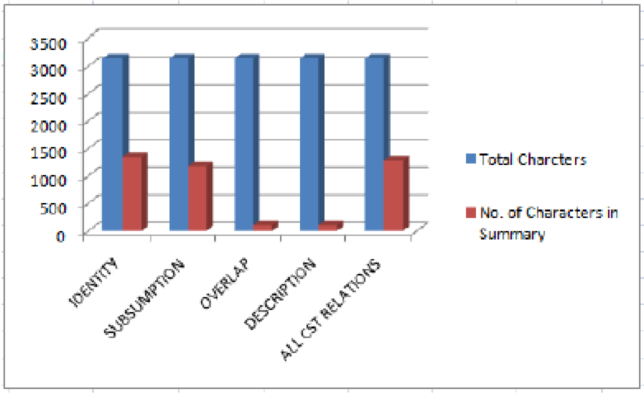

Fig 4: Summary Analysis for Data Set 1

VI. Conclusion and Future Scope

Thus, the system is implemented by uploading the files for summarization with the same domain. For any summarization system there are two steps 1. Pre-processing and 2. Processing step. So for these uploaded files pre-processing is done and then system implements feature extraction. CST relations are identified from the summary of feature extraction.

The main focus of this system is identification of CST relations in text documents by implementing a new system which combines the result of feature extraction and CST relations. The implementation of feature extraction and identification of CST relations is done to reduce human efforts to summarize contents from huge information. Better performance is achieved through the system. Further it is on the user to opt for number of features and number of CST relations. The system implements the concepts of pre processing (removal of common words and obtain stem of word), feature extraction (features of language to obtain summary), CST relations (concepts use for multi document summarization).

In future the system can be implemented with more features and additional CST relations to obtain better summary. Also the system can be implemented online for search engines. When user searches for any information along with the links summary can be shown to user at one side. In future the system can also be implemented for other file formats.

VI. Acknowledgment

I am very much thankful to my respected project guide Prof. P. P. Rokade, H.O.D., Department of Information Technology, for his ideas and help, proved to be valuable and helpful during the creation of this paper and set me in the right path.I am also very much thankful to Head of Computer Department Prof. S. R. Durugkar and P. G. Coordinator Prof. I. R. Shaikh, Computer Engineering Department, for helping while selecting and preparing for dissertation work. I would also like to thank all the Faculties and Friends who have cleared all the major concepts that were involved in the understanding of techniques behind my work.

VII. References

[1] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques," Journal of Emerging Technologies in Web Intelligence, vol. 2, pp. 258-268, **2010**.

[2] Yogan Jaya Kumar, Naomie Salim, Albaraa Abuobieda, Ameer Tawfik, “Multi Document summarization based on cross-document relation using voting technique”, International conference on computing, electrical and electronic engineering (ICCEEE), **2013**.

[3] Y. J. Kumar and N. Salim, "Automatic multi document summarization approaches," Journal of Computer Science, vol. 8, pp. 133-140, **2011**.

[4] Ultimate Research Assistant, http://en.wikipedia.org/wiki/Ultimate_Research_Assistant, 27 Jan,**2015**.

[5] D. R. Radev, "A common theory of information fusion from multiple text sources step one: cross-document structure," presented at the Proceedings of the 1st SIGdial workshop on Discourse and dialogue – Volume 10, HongKong, **2000**

[6] D. R. Hsun-Hui Huang, Horng-Chang Yang, Yau-Hwang kuo, “A Fuzzy-Rough Hybrid Approach to Multi-document Extractive Summarization” , Ninth International Conference on Hybrid Intelligent Systems, **2009**

[7] Md. Mohsin Ali , Monotosh Kumar Ghosh, and Abdullah-Al-Mamun, “Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation”, International Conference on Future Computer and Communication, **2009**

[8] Z. Zhang, S. Blair-Goldensohn, and D. R. Radev, "Towards CST-enhanced summarization," presented atthe Eighteenth national conference on Artificial intelligence, Edmonton, Alberta, Canada, **2002**

[9] M. L. d. R. C. Jorge and T. A. S. Pardo, "Experiments with CST-based multidocument summarization," presented at the Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing, Uppsala, Sweden, **2010**

[10] Z. Zhang, J. Otterbacher, and D. Radev, "Learning crossdocument structural relationships using boosting," presented at the Proceedings of the twelfth international conference on Information and knowledge management, New Orleans, LA, USA, **2003**.

[11] Rajesh S.Prasad, Dr. U.V.Kulkarni, Jayashree R.Prasad, "A Novel Evolutionary Connectionist Text Summarizer (ECTS)", published in proceedingASID'09 Proceedings of the 3rd international conference on Anti-Counterfeiting, security, and identification in communication, IEEE Press Piscataway, NJ, USA, 20 Aug **2009**

AUTHORS PROFILE

Ms. Yogita K. Desai completed Bachelor Degree in Information Technology from University of pune, presently working with SNJB's KBJ COE, Chandwad, Nashik, Studying Masters in Computer Engineering from University of Pune at SND COE and RC, Nasik(MH), India. Her Research Interest is in Data Mining, Information Security, Computer Graphics.

Prof. P.P.Rokade working as Head of Department in Information Technology at SND COE and RC, Yeola. (MH), India. He has completed Masters in Computer Engg. from Bharti Vidyapeeth, Pune and Pursuing Ph.D in Data Mining. His Research area currently includes Text Mining, Web Mining, Information Security etc.