International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Multi Document Summarization: A Cross Document Structure Theory Approach

Ms Yogita K. Desai[1], Prof Prakash P. Rokade[2], Prof. Swati V. Sinha[3]
[1]*Department of Computer Engg., SND COE & RC, yogitadesai9@gmail.com*
[2]*Department of Computer Engg., SND COE & RC, prakashrokade2005@gmail.com,*
[3]*Department of Information Technology, SNJB's LS KBJ COE, swati_sinha_it@yahoo.co.in*

Abstract— the graph of use of internet is increasing day by day. All the information over the world can be retrieved on one click. But the information retrieved is so huge to read and sort. So to summarize this information, multi document summarization can be used. With this summarization user can read short and precise description of whole information. Proposed system uses CST (cross document structure theory) relations to identify relevance in information. Further dictionary based approach is proposed to identify these CST relations. Score of each sentence is calculated with the help of scoring model. Depending on the score of sentence, the system decides which sentence should be included in summary. Results are shown for number of sentences versus number of filtered sentences in input documents

Keywords- Multi document summarization, CST relations, scoring model, filtered sentences, and dictionary based approach.

I. INTRODUCTION

Numerous methods are described in last few decades for text summarization [1]. Because of use of internet, large amount of information is generated every day. This all information is maintained and provided to users by using different database techniques. But, when user sends a query to server (related to some information retrieval), much information is displayed and reading all this information is time consuming process. This leads to need of text summarization. The aim of text summarization is to give user a precise and compact representation of whole information by conserving its meaning. Fundamentally, text summarization can be divided into four approaches: informative, indicative, abstractive and extractive. In informative text summarization, summary describes about what topic the whole document is. If the summary generated according to user's query, then it is called indicative summary. Abstractive summarization is one in which, sentences chosen from novel document are further processed to reorganize them into the final summary. This process includes deep knowledge of natural language and sentence compression. Extractive approach for summarization identifies important sentences from document and those sentences are included in summary. In addition to this, extractive summarization is a process of two steps: first step is pre processing and second step is processing. In pre processing, document is represented systematically by the steps a) sentence segmentation: which divides the document in number of sentences by considering dot as a delimiter character. b) Common word removal: removes the words which are not representing relevant information or which actually do not contribute to relevant information c) stemming: refers to the process of obtaining stem of word, which highlights its semantics. Second step, processing refers to extracting features from pre processed document, allocating weights to these features. Then calculating weight of each sentence by using feature weight and finally high scored sentences are included in summary.

Further, summarization can be categorized as single document summarization involving summary for single document and multi document summarization indicating summary from several

documents, which are from same topic. A fact of showing related information as well as some multi document concepts like contradiction, ordering of information, redundancy should be handled by multi document summarization. Proposed work focuses on multi document summarization using extractive summarization approach. Documents from the same topic contains related information, by taking into account this fact, proposed system is based on CST (Cross document structure theory) relations. Cross document structure theory is introduced by Radev. Radev [2] proposed that CST based analysis of relevant documents can support multi document summarization. Relevance of information in multiple documents is denoted by four levels: word level (W), phrase level (P), sentence/paragraph level (PR) and document level (DOC).Fig 1 shows different levels for summarization.

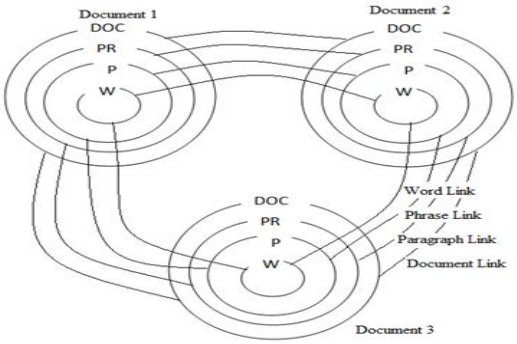

Fig. 1 MDS at different levels

In the proposed system, first step is to apply pre processing to the input documents, which includes segmentation i.e. to divide the document into number of sentences, common word removal and stemming means to obtain radix of given words. Second module is feature extraction and it is used to obtain features. Feature extraction is done with the help of six different features. By using these features CST relations are identified. Proposed system uses four CST relations identity, subsumption, overlap and description as it covers most of other relations. Upon identification of CST relations; sentences are scored using scoring model.

The proposed system represents automatic summarization using CST relations and dictionary approach. Section-II gives overview of multi document summarization approaches. Details of proposed system are given in section-III. Section-IV represents experiments and results. Proposed system is concluded in section-V.

II. OVERVIEW OF MULTI DOCUMENT SUMMARIZATION APPROACHES

In academia, number of research studies has been proposed for Multi document summarization. The attempt is to facilitate user by giving short description of whole content. This description preserves meaning as well as saves the time required to read all information. G. Erkan et. al.[3] proposed an approach for identifying sentence centrality. The sentence centrality is calculated using prestige scoring of sentences based on graph. Two different methods are introduced to calculate prestige in similarity graph first, degree and second LexPageRank. Xiaojun Wan and Jianwu Yang[4] proposed two models which makes use of theme clusters in document. Amongst the number of methods for multi document summarization, commonly used methods are feature based, graph based, cluster based and knowledge based. The methods are described below:

2.1 Method Based on Features of Sentence:

As discussed earlier, extractive summarizer selects important and relevant sentences and put them in final summary [5]. So for feature based extractive summarization, features that determine importance

and relevance of sentence are extracted. Some of the common features are word frequency, position of a sentence, title word, cue word etc.

2.2 Method Based on Graph:

Generally a graph is represented as a pair of V and E, where V is set of vertices and E is set of edges. In case of documents for summarization, sentence is represented by vertex and edge represents similarity among the sentences. Which is again can be measured by cosine similarity. A sentence is included in summary if it is strongly connected to other sentences. Erkan and Radev [6] proposed that similarity graph construction of sentences gives better identification of important sentences than compared to the centroid approach.

2.3 Method Based on Cluster:

In a cluster, similar objects are grouped into one class. When text summarization is considered, highly relevant sentences are grouped into one cluster, thus forming number of clusters. Relevance of sentences to form a cluster is identified by using cosine similarity. When summary is generated, it includes the process of selecting sentences from each cluster.

2.4 Method Based on Knowledge:

Generally the documents have its contents related to particular topic or heading. These documents from related topic belong to particular domain. And every domain has its knowledge structure. So researchers use this knowledge for summarization. Background knowledge is referred to as ontology. Nasir and Noor's study [7] facilitates converting unstructured information into the form that machine can understand.

III. PROPOSED SYSTEM

3.1 System Architecture:

Fig 2. Shows system architecture of proposed system.As shown in fig 2, the proposed system is divided into four parts

1. Preprocessing
2. Feature Extraction
3. Identification of CST relations
4. Sentence Scoring

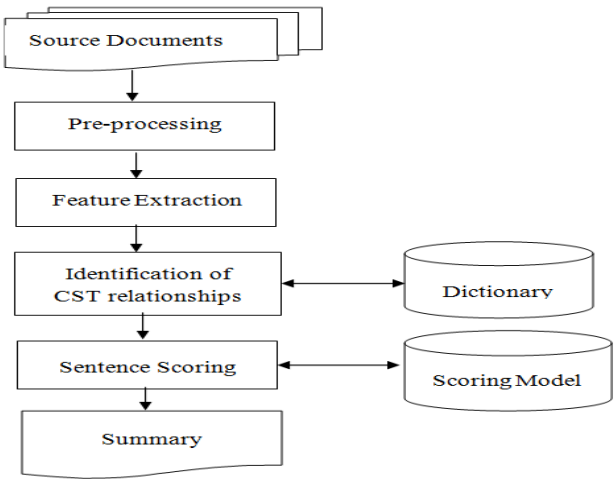

Fig.2 System Architecture of Proposed System

3.1.1 Pre Processing Module:

Performing pre processing on document makes the document useful for further processing. It takes into consideration four steps segmentation, common word removal, tokenization and stemming. Segmentation refers to the process of dividing the document into number of statements. For dividing the document into segments, system takes into consideration the delimiter character i.e. full stop. Step 3 is common word removal. Common words are the words which do not have any relevance to the document and also they do not contain any semantics. So this step removes such words. Tokenization performs decomposing sentence into the stream of texts which is known as tokens. These tokens are further processed to form radix or stem by using stemming.

3.1.2 Feature Extraction:

In this module, vector of features is formed to extract the number of features. These features are extracted from the document obtained by pre processing step. The process of transforming input information into a set of features is called as feature extraction. Six different features are considered for defining feature vector. Features used in proposed system are listed in table 1.

Feature 1 considers that first paragraph in the document which is followed by title contains relevant information of the document thus, first paragraph in document is important. F2 & F3 describes that location of paragraph and location of sentence in paragraph identifies important which described in F4. Long sentences contain usually more information than short sentences. This fact is represented in F5. If sentence contains number of thematic words or title word then sentence s should be included in summary which is described in F6.

Table 1 Features used in Proposed System

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

3.1.3 CST Relation Identification:

This module is used to identify CST relations using dictionary approach. As mentioned above four CST relation viz. Identity, subsumption, overlap and description are identified. Description of these relations is given in table 2.

Table 2CST Relations

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

Sentence pairs are used to find the type of relation between the sentences of documents. Sentence pairs are denoted by (S1, S2) where, S1 is first sentence and S2 is second sentence.

Cosine Similarity: It is a measure to find out how much similar two sentences are. It is represented with TF – IDF (i) value [1].

$$\cos(S1, S2) = \frac{\sum S1, i * S2, i}{\sqrt{\sum (S1, i)^2} * \sqrt{\sum (S2, i)^2}} \quad (1)$$

Word Overlap: This measure is based on number of overlapping words in two sentences. The sequence of words in these sentences does not affect this measure [1].

$$overlap(S1, S2) = \frac{\#commonwords(S1, S2)}{\#words(S1) + \#words(S2)} \quad (2)$$

Length type: Type of first sentence with respect to length is calculated when first sentence is compared to second sentence [1].

$$Length\ type(S1) = 1\ if\ length(S1) > length(S2)$$
$$= -1\ if length(S1) < length(S2)$$
$$= 0\ if\ length(S1) = length(S2) \quad (3)$$

NP similarity: Noun Phrase (NP) Similarity among two sentences is represented by this measure[1]. Jaccard coefficient is used to find the NP similarity as given by equation 4[1].

$$NP(S1, S2) = \frac{NP(S1) \cap NP(S2)}{NP(S1) \cup NP(S2)} \quad (4)$$

VP similarity: Verb Phrase (NP) Similarity among two sentences is represented by this measure. Jaccard coefficient is used to find the VP similarity as given by equation 5 [1].

$$VP(S1, S2) = \frac{VP(S1) \cap VP(S2)}{VP(S1) \cup VP(S2)} \quad (5)$$

3.1.4 Sentence Scoring:

Scoring of each sentence is done with the help of scoring model. Based on the CST relations identified in previous module affinity matrix is calculated representing the CST relations. Affinity matrix is used to find initial score of sentence. Initial score is calculated by equation 6.

$$M\ i, j = Relsen(Si, Sj) \quad if\ i \neq j\ 0\ otherwise \quad (6)$$

Final score of sentence is calculated by making use of two rules. These rules are based on two CST relations, description and subsumption because these relations have 1-way directionality.

First Rule: (Rule based on Subsumption) Score of first sentence is updated, if first sentence (s1) contains all information in second sentence (s2) including additional information which is not in second sentence (s2).

Second Rule: (Rule based on Description) Score of second sentence is updated if first sentence (s1) describes an entity in second sentence (s2) as relevant information is in s2.

Final score is calculated as initial score plus the score obtained by the two rules. Depending on the scores of sentences reorganize the sentences. And finally, higher score sentences are included in summary.

IV. EXPERIMENTS AND RESULT

As described in previous section system is divided into four modules out of which two modules are implemented and other two modules implementation is in progress. Numbers of documents are taken as input and document pre processing is applied. For input documents, segmentation, common word removal and tokenization are applied. Stemming is performed by using affix stripping algorithm. These words are given as an input to feature extraction and six features are implemented. Out of which features are to be used is dependent on user. Graphical results are shown for number of sentences in input documents and number of filtered sentences for the same documents (fig.3)

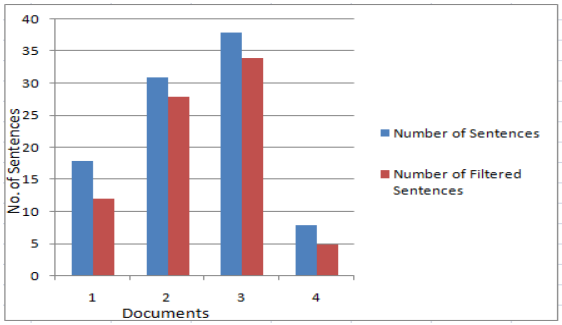

Fig. ***3 Results for Number of Filtered Sentences for Input Documents***

CONCLUSION AND FUTURE SCOPE

A method based on dictionary approach for CST relation identification is proposed. Dictionary is used to automatically identifying the CST relations which is the main limitation of other systems. At this point only two modules for the system are implemented viz. Document pre processing and feature extraction. This gives result of number of sentences in input documents to the number of filtered sentences. Currently, the proposed system takes into consideration the input documents from the same topic.

In future, the system can be implemented by considering input documents from different topics. And the result of the system can be checked by considering more CST relations.

REFERENCES

[1] 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

[2] 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

[3] Erkan G. and D.R. Radev, LexPageRank: Prestige in multi-document text summarization, University of Michigan, 2004

[4] Wan, X. and J. Yang, Multi-Document Summarization Using Cluster-Based Link Analysis. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, ACM, New York, USA., pp: 906. ISBN: 978-1-60558-164-4, 2008

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

[6] Erkan, G. and D.R. Radev, LexRank: Graph based lexical centrality as salience in text summarization. J. Artifi. Intelli. Res., 22: 457-479, 2004

[7] Nasir, S.A.M. and N.L.M. Noor, Automating the mapping process of traditional malay textile knowledge model with the core ontology. Am. J.Econ. Bus. Admin., 3: 191-196, 2011

