2018 International Conference on Computing, Power and Communication Technologies (GUCON)
Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018

Review on Abstractive Text Summarization Techniques (ATST) for single and multi documents

Ms. Shivangi Modi
PG Scholar
Computer Engineering Department
SCET, Surat, India
smodi0123@gmail.com

Prof. Rachana Oza
Assistant Professor
Computer Engineering Department
SCET, Surat, India
oza.rachana7@gmail.com

***Abstract*— In recent times, there is an enormous amount of data available on the internet. It is laborious for users to encapsulate large amount of data manually. Automatic text summarization can solve this problem by generating summary automatically. It can be categorized into extractive and abstractive text summarization techniques. Existing techniques of extractive text summarization extract important sentences from original document and generate summary without any modification of actual data. This technique may not present conflicting information properly. Abstractive text summarization can solve this problem by representing the extracted sentences into another understandable semantic form. This paper discusses abstractive text summarization techniques and highlights the parametric evaluation of these techniques.**

Keywords — abstractive summary, text summarization, single document, multi documents, summary

I. INTRODUCTION

Now a days, amount of data available on internet increases day by day [1]. Due to the exponential growth in data, the need of information abstraction or summarization arises. Text summarization is the way of extracting important information and to represent that information in the structure of summary [1-7]. It is laborious for users to search and to choose which information is important from the massive volume of data [1][4]. To solve the difficulty of manual summarization, automatic text summarization is used. Automatic Text Summarization is the technique which is significant for users to automatically generate summary from text [3][4]. Main aim of automatic text summarization is to remove redundant data and to extract important information from source document and then represented in the more concise way possible [1-3]. The automatic summarization is the core subtle part of natural language processing [1][3]. Automatic text summarization used in numerous domain regions, for example, news-paper story summary, e-mail outline, information summary for students, short news on mobile, search engines etc [8].

Automatic text summarization can be categorized into extractive text summarization and abstractive text summarization method [1...5]. An extractive summarization method comprises of choosing important sentences from the original text document and connecting them into shorter form without changing or altering the main text. In Abstractive summary method, the original text gets changed over into another more understandable semantic form to get a shorter summary of original text document [1][2][3].

In extractive text summarization extracted sentences could become longer than the average [2][3]. Due to this some of the portion which are not important for summary that also gets included. Moreover the conflicting information may not be presented properly [2]. Abstractive text summarization can solve this problem by representing the extracted sentences into another more understandable semantic form [2]. In this paper we are studying different techniques of abstractive text summarization.

Summary can be generated from either single document or multi document [4]. Single document text summarization purpose is to extract most important information from single source document and create a short summary that can satisfy user's need [1][5]. Multi-document text summarization purpose is to extract important text from more than one documents and then collectively used for producing summary [4][5].

This paper aims to make survey of existing abstractive text summarization techniques along with parametric evaluation of these techniques.

This paper is organized as follows: Section 2 discusses various different text summarization techniques. The parametric evaluation of abstractive text summarization techniques is presented in Section 3. Finally, Section 4 concludes with a discussion of future research directions in this area.

II. EXISTING TECHNIQUES

This section gives a detailed description of various abstractive text summarization techniques. Depending on the input, text summarization approach can be classified according to single document text summarization and multi document text summarization.

A. Single document text summarization

Single document text summarization is to build summary from single source document. This type of text summarization technique accepts only one document as input, then uses different techniques to extract important sentences from source document and then after from extracted sentences summary to be generated. Generation of summary is in more understandable, syntactically or semantically correct and most important in reduced form. Various techniques of single document text summarization are discussed in the following section.

1) Semantic Graph Reduction Approach

IF Moawad et al. [1] approach performs summarization of input document by generating semantic graph. In this approach generation of semantic graph is identified as rich

semantic graph (RSG). Then that semantic graph can be further reduced and generates final abstractive summary from reduced semantic graph. In this approach system will take source document in English language. This approach consists of three phases. The first phase is RSG creation. The main aim of the RSG creation is to represent the input document semantically. In that tokens of source document represented as graph vertex and edges between nodes are represented as semantic relationship between them. In this way for each sentences of source document sub graph generated. At the end the sub graphs are merged together to represent entire document. The second phase is RSG reduction. In this phase generated semantic graph is reduced by applying certain set of rules like merging and deleting the graph nodes. Third phase is abstractive summary generation from reduced RSG. This approach succeeds to reduce the source document up to half of the original document. Limitation of this approach is that it will not take multiple documents as input to generate abstractive summary.

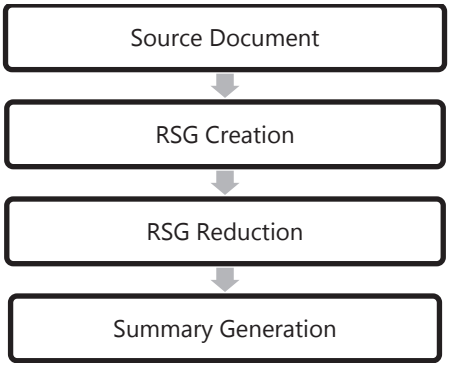

Fig.1. Architecture of semantic graph reduction approach [1]

2) Word Graph based Approach

Huong Thanh Le et al. [2] approach uses word graph to represent source document. This approach includes two phases. First phase is sentence reduction and second is sentence combination. The sentence reduction phase is depends on certain rules and syntactic constraints for removing redundant clauses and to complete the end of the reduced sentence. Word graph is used for sentence combinations and to represent word relations between texts. New sentences are generated from word graph. In word graph nodes are used to represent the information about words and their part of speech tagger and the adjacency relations between words pairs are represented on edges. This approach generate syntactically correct sentence but does not care about word meaning.

3) Sentiment Infusion Approach

R. Bhargava et al. [3] approach work on a graph based technique that makes summaries of redundant opinions and utilizes sentiment analysis to join the statements. This approach uses word graph for compressing and merging information and then summaries are generated from resultant sentences. The graph captures the excess in the document using words that happen more than once in the texts are mapped to the similar node. Moreover, the diagram creation does not require any domain learning. At the time of graph generation this approach will ensure the correctness of sentences. For getting abstractive summary, we give score to every one of the ways as well as the sentences have been combined. After that we rank the sentences in dropping request of their scores and expel duplicate sentences from summary using jacquard index technique for similarity measure. At that point the remaining top most sentences are chosen for the summary.

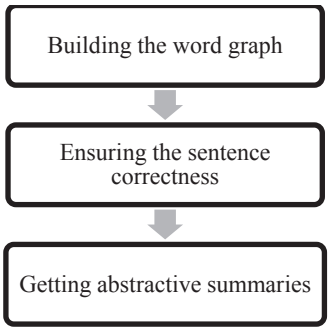

Fig.2. Architecture of Sentiment Infusion Approach [3]

B. Multi document text summarization

Multi document text summarization is to build summary from more than one source document. The purpose of multi-document text summarization is to extract meaningful information from each source documents and then generate a summary that can fulfill human's need. Various techniques using this type of approach are as follows.

1) Genetic Semantic Graph Based Approach

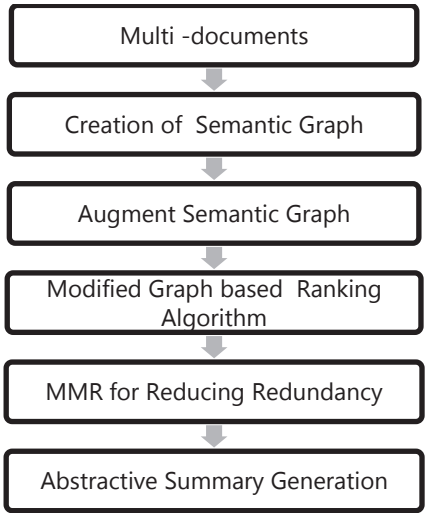

Fig. 3. Proposed Genetic Semantic Graph Based Approach [4]

Atif Khan et al. [4] approach work on a semantic graph based technique for multi document text summarization. Semantic graph is to be created for each document in such a way that the Predicate Argument Structure (PASs) is represented as graph vertices and the semantic similarity weight is to be represented on edges. For constructing PASs they use semantic role labeling. First this phase will extract predicates from each sentences of document then split predicate structure into meaningful tokens. After that they perform semantic similarity matrix operation. In that they make a framework for each pair of PAS. Semantic similarity measure is based on wordnet dictionary for calculating similarity between tokens. Once the semantic similarity matrix is built then they construct undirected weighted semantic graph. In this type of graph edges between two PASs is generated if similarity weight between them is

greater than one value. Then after in order to reflect the impact of document set on PASs, they augments semantic graph with PAS-to-document set relationship. The PASs is ranked using modified graph based ranking algorithm. To eliminate or to reduce redundancy they need to re-rank the PASs. For that purpose they used maximal marginal relevance (MMR) algorithm. After getting rank of each PASs, top ranked PASs are used to generate abstractive summary. This approach automatically merges similar information across the documents to reduce the overlapping information in summary.

2) Clustered Genetic Semantic Graph Based Approach

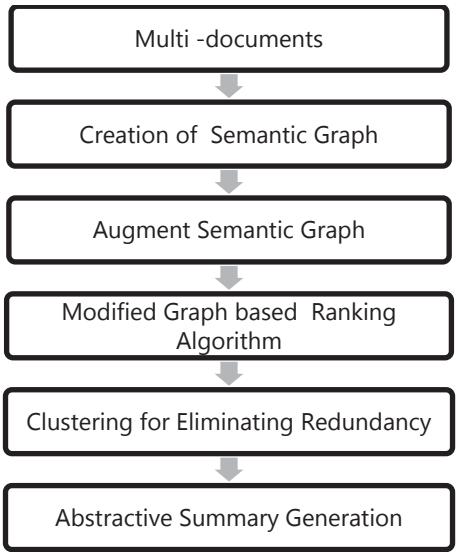

Fig. 4. Proposed Clustered Semantic Graph Based Approach [5]

Atif khan et al. [5] approach work on clustered semantic graph based approach for multi document abstractive text summarization. This approach is similar to genetic semantic graph based approach but here they used clustering algorithm to eliminate redundancy. In clustering algorithm PASs with highest similarity weight score from each cluster is chosen and apply to language generation. Then language generation rules are used to generate abstractive summary sentences. For making cluster they use Hierarchical Agglomerative Clustering (HAC) algorithm. HAC algorithm accepts the semantic similarity matrix as input. Algorithm merges most similar clusters and updates the semantic similarity matrix to represent similarity between the nearest cluster and the original cluster. End user will decide the compression rate of summary document. This process will be repeated until the user defined compression rate of summary is reached.

III. PARAMETRIC EVALUATION

This section shows comparison of previously discussed abstractive text summarization techniques. Table-1 shows comparison of techniques based on parameters are as follows. Type of text summarization parameter indicates that abstractive summary to be generated from single source document or multi documents. Source document representation parameter is to be constituted that the original text to be represented in which form. Content selection parameter represent that which techniques or algorithm used for extracting important information. Summary generation parameter describes that final abstractive summary generated in which form. Semantic summarization parameter and syntactically correct representation parameter indicates that generated summary is semantically and syntactically correct or not. This all techniques are based on mono-lingual language based techniques. There are other languages also available like multi lingual and cross lingual. When input and output both document's text languages are same then it will known as mono lingual language based technique. In multi-lingual language input document would be in more than one language and output will be in the user desired language and in cross-lingual language, input and output language is different from each other.

TABLE-I: PARAMETRIC EVALUATION OF ABSTRACTIVE TEXT SUMMARIZATION TECHNIQUES

Technique	**Type of Text Summarization**	**Original Text Representation**	**Content Selection**	**Summary Generation**	**Semantically Correct Summarization**	**Syntactically Correct Representation**	**Technique used for Eliminate Redundancy**
Semantic Graph Reduction Approach [1]	Single document	Rich semantic graph	Heuristic rules	Reduced semantic graph	Yes	No	-
Word Graph based Approach [2]	Single document	Word graph	Relation among words, Clauses	Word graph	No	Yes	-
Sentiment Infusion Approach [3]	Single document	Word graph	Sentiment analysis	Path Scoring, Sentence Fusion	Yes	Yes	-
Genetic Semantic Graph based Approach [4]	Multi document	Semantic graph	Semantic Role Labeling and Semantic Similarity Score	SimpleNLG and Simple Heuristic rule	Yes	Yes	Maximal Marginal Relevance (MMR) algorithm
Clustered Semantic Graph based Approach [5]	Multi document	Semantic graph	Semantic Role Labeling and Semantic similarity score	SimpleNLG and Simple Heuristic rule	Yes	Yes	Clustering algorithm

IV. Conclusion And Future Work

In this paper different abstractive text summarization techniques have been shown. These techniques are based on natural language processing, data mining and semantic similarity domains. These all techniques are used for to generate summary automatically from source document. We study different abstractive text summarization techniques based on single document and multi document. We study three techniques of single document text summarization and two techniques of multi document text summarization. These all techniques are mono lingual language based. In single document text summarization techniques, semantic graph based reduction approach produces concise, coherent and less redundant sentences. Sentiment infusion approach generates summary which is semantically and syntactically correct and in reduced form. Among multi document text summarization techniques, clustered genetic semantic graph based approach eliminate the overlapping semantic redundancy significantly.

Future work may include developing a more efficient technique with multi-lingual or cross-lingual structure based. One can also try to generate more concise and less redundant summary by designing new approach or by merging available techniques.

References

[1] IF Moawad, M Aref – "Semantic Graph Reduction Approach for Abstractive Text Summarization", Computer Engineering & Systems (ICCES), pp. 132-138, IEEE-2012.

[2] H.T Le, T.M Le –"An approach to abstractive text summarization", Soft Computing and Pattern Recognition (SoCPaR), IEEE-2013.

[3] R Bhargava, Y Sharma, G Sharma – "ATSSI: Abstractive Text Summarization using Sentiment Infusion", Procedia Computer Science, Elsevier-2016.

[4] A Khan, N Salim, YJ Kumar – "Genetic Semantic Graph Approach for Multidocument Abstractive Summarization", Digital Information Processing and Communications (ICDIPC), IEEE-2015.

[5] A Khan, N Salim, H Farman – "Clustered Genetic Semantic Graph Approach for Multi-document Abstractive Summarization", Intelligent Systems Engineering (ICISE), IEEE-2016.

[6] AR Pal, D Saha – "An approach to automatic text summarization using WordNet", Advance Computing Conference (IACC), IEEE-2014.

[7] KS Thakkar, RV Dharaska – "Graph-Based Algorithms for Text Summarization", Emerging Trends in Engineering and Technology (ICETET), IEEE-2010.

[8] J Zhan, HT Loh, Y Liu – "Gather customer concerns from online product reviews – A text summarization approach", Expert Systems with Applications, Elsevier-2009.