Abstractive Multi-Document Summarization: An Overview

D.Y.Sakhare[1]; Dr.Rajkumar[2]
[1] Research scholar, Bharti Vidypeeth Deemed University, Pune, Maharashtra, India
Email: diptiysakhare@gmail.com
[2] Scientist 'D', DRDO, DIAT, Pune, Pune, Maharashtra, India

***Abstract* — In recent times, the necessity of generating single document summary has gained popularity among the researchers due to its extensive applicability. The text summarization can be categorized with different approaches like: extractive and abstractive from single document or multi document, goal of text summarization (intent, focus and coverage), characteristic of text summarization (frequency-based, knowledge-based and discourse-based), level of processing (surface level, entities level and discourse level) and kind of information (lexicon, structure information and deep understanding). Recently, the efforts in research are transferred from single document summarization to multi document summarization Multi-document summarization considerably differs from single in issues related to compression, speed, redundancy and passage selection are critical in the formation of useful summaries. In this paper, we review the techniques that have done for multi document summarization. Next, we describe evaluation method In conclusion; we propose our future work for multi document**

***Index Terms* — Multi-document, summary.**

I. Introduction

Text summarization is an important activity in the analysis of a high volume text documents. The text summarization has numerous applications; recently number of applications uses text summarization for the betterment of the text analysis and knowledge representation [14, 6]. A human performs following steps to do a summarization. Firstly, they understand the content of the document, then, he identifies the most vital pieces of information in the text, and finally, writes up this information. Automating the first and last part is beyond the state of the art for unconstrained documents. So, summarization which is based on semantics would be a future success, but for now, making summaries reduces to the task of Extraction [1]. Such summaries are very informative, but it is still hard for a user to understand why an aspect received a particular rating, forcing a user to read many, often highly redundant sentences about each aspect. To help users further digest the opinions in each aspect, it is thus desirable to generate a concise textual summary of such redundant opinions. Indeed, in many scenarios, we will face the problem of summarizing a large number of highly redundant opinions [3]. A challenging problem in implementing this approach in a particular domain is to devise a content selection strategy that identifies what key information should be presented. In general, content selection is a critical task at the core of both summarization and NLG and it represents a promising area for cross-fertilization [4].

The human understands a single or a cluster of text documents by consuming the main themes of the documents by applying some cognitive process. Multi-document summarization is a process, which produces a condensed representation of the contents of multiple related text documents collected from heterogeneous sources for human consumption. Thus, Multi-document summarization helps human to digest the main contents of multiple related text documents very rapidly [15]. Therefore, Multi-document summarization is an increasingly important task: as document collections grow larger, there is a greater need to summarize these documents to help users quickly find either the most important information overall (generic summarization) or the most relevant information to the user (topic-focused summarization) [10]. There are two methods of summarization called, extractive and abstractive. In extractive, sentences are extracted as summary based on the benchmark features. But, abstractive methods require a deeper analysis of the text and the ability to generate new sentences, which provide an obvious advantage in improving the focus of a summary, reducing its redundancy and keeping a good compression rate, there is an empirical limit intrinsic to pure extraction, as compared to abstraction [9] and researchers to reduce the time spent manually extracting the main ideas from text documents. For academics, it expedites, one way or another, the research progress by taking original text documents and generating shorter documents comprising of only the salient points [6].

For these reasons, as well as for the technical and theoretical challenges involved, we were motivated to come up with an abstractive multi document summarization model. Recent abstractive multi document summarization approaches have focused on rewriting techniques, without consideration for a complete model which would include a transition to an abstract representation for content selection. We believe that a "fully abstractive" approach requires a separate process for the analysis of the text that serves as an intermediate step before the generation of sentences. This way, content selection can be applied to an abstract representation rather than to original sentences or generated sentences [2].

II. Review Of Related Works

Elena Lloret and Manuel Palomar [5] have focused on abstractive text summarization for explore to what extent sentences generated employing a word graph-based method (which either compress or merge information) were suitable for producing abstracts. Moreover, in order to be decided which of the sentences should be included in the abstractive summary, and an approach based on extractive text summarization was developed (i.e., COMPENDIUM),

so that the most relevant abstractive sentences could be selected and extracted. As shown by the results obtained, these tasks were very challenging. But, the preliminary experiments carried out prove that the combination of extractive and abstractive information was a more suitable strategy to adopt towards the generation of abstracts.

Jackie Chi Kit Cheung and Gerald Penn [11] have investigated the abstractive summarization could advance past these paradigm towards robust abstraction by making greater use of the domain of the source text. They conducted a series of comparing human-written model summaries to system summaries at the semantic level of case frames. They showed that model summaries (1) were more abstractive and make use of more sentence aggregation, (2) do not contain as many topical case frames as system summaries, and (3) cannot be reconstructed solely from the source text, but could be if texts from in-domain documents were added. These results suggested that substantial improvements were unlikely to result from better optimizing centrality-based criteria, but rather more domain knowledge was needed.

Pierre-Etienne Genest and Guy Lapalme [17] have showed that full abstraction could be accomplished in the context of guided summarization. They described a work in progress that relies on Extraction of Information, then statistical content selection and followed by Natural Language Generation. Early results already demonstrated the effectiveness. Extractive summarization was the strategy of concatenating to extracted taken from a corpus into a summary, while abstractive summarization involved paraphrasing the corpus used sentences.

Cheung [16] have defined a measure of corpus controversiality of opinions contained in evaluative text, and report the results of a user comparing extractive and NLG-based abstractive summarization at different levels of controversiality. The abstractive summarizer achieves better overall, the results suggested that the margin by which abstraction outperforms extraction was greater when controversiality was high, provided a context in which the need for generation based methods was especially great.

Fei Liu and Yang Liu [18] Have investigated could be applied sentence compression to extractive summaries to generate abstractive summaries. They used different compression algorithms, including integer linear programming with an additional step of filler phrase detection, a noisy channel used Markovization formulation of grammar rules, as well as human compressed sentences. Their experiments on the ICSI meeting corpus showed that when compared to the abstractive summaries, used sentence compression on the extractive summaries was improved their ROUGE scores; however, the best performance was still quite low and suggested the need of language generation for abstractive summarization.

Mohamad Ali Honarpisheh et al [8] have proposed multi-document multi-lingual text summarization techniques, based on the singular value decomposition & hierarchical clustering. They move toward relies on only two resources for any language: a word segmentation system and a dictionary of words along with its document frequency. The summarizer primarily starts with a collection of related documents, and converts them into a matrix; followed by singular value decomposition to the resultant matrix. Then by use of binary hierarchical clustering algorithm, the most important sentences from the most important clusters are used to form the summary. The appropriate place or sequence of each chosen sentence was determined by a technique. The system has been successfully tested on summarizing several Persian document collections.

Jade Goldstein et al [13] have discussed a text extraction approach to multi document summarization that builds on single document summarization methods used additional, available information about the document set as a whole and the relationships between the various documents. Multi-document summarization technique diverges from single in that the issues of compression, passage selection, speed and redundancy were critical in the formation of useful summaries. Their approach addresses these issues was used domain-independent techniques based mainly on speedy, statistical processing, a metric used for reducing redundancy and maximizing diversity in the selected passages and modular framework to allow easy parameterization for different genres, user requirements . Chong Long et al [9] have described for multi-document update summarization. The best summary was defined asone of which has the minimal information distance to the entire document set and the best update summary has the minimal conditional information distance to a document clustered given that a prior document cluster has already been read. They have two methods to approximate information distance between two documents, one by compression and the other by the coding theory. Experiments on the DUC 2007 dataset and the TAC 2008 dataset have proved that their method closely correlates with the human-written summaries and outperforms Lex Rank in many categories under the ROUGE evaluation criterion.

Heng Ji et al [7] have investigated to taking advantage of cross-document IE for multi-document summarization. They have multiple approaches to IE-based summarization and analyze their merits and demerits. One of them, re-ranking the output of a high performing summarization system with IE-informed metrics, leads to improvements in both manually-evaluated content quality and readability. Vikrant Gupta et al [12] have presented a statistical approach to automatic summarization based on the Kernel of the source text. The Kernel-based system, called Kernel Sum (KERNEL Summarizer), was used the Kernel as a guideline to identified and selected text segments to include in the final extract. the automatically created extracts were evaluated under the light of Kernel preservation and textuality

III. Evaluation Method

The task of automatic text summarization Evaluation is important and difficult. There are many criteria of summarization evaluation such as information coverage, grammatical and discursive coherence, readability etc. Evaluation of summarization can be intrinsic or extrinsic: Intrinsic methods measure a system's quality based on

analysis in terms of some set of norms. Extrinsic methods measure a system's performance task based on how it affects the completion of other application task. Evaluation can be evaluated by manual and/or automatic: Manual: Human judgment of the quality of a summary varies from person to person and chooses sentences from document to create manual-extract summaries. Human who evaluated is specialization in each topic. Automatic: Generated summaries and evaluated by computer such as ROUGE (Recall Oriented Understudy for Gist Evaluation) is the official scoring technique for Document Understanding Conference (DUC) 2004, TIPSTER [16]

IV. Proposed Methodology

In future we want to design and develop a system for automatic abstractive text summarization using artificial intelligence and NLP techniques. In the literature, lot of techniques has been presented for extractive summarization with the help of AI techniques. But, the contributions made in abstractive text summarization seems only a very few since it is very challenging task. By taking this challenging task, we have planned to do the research in the abstractive text summarization by effectively utilizing the fuzzy classifier and NLP techniques that use the subject object verb relationship.

V. Conclusion

In this paper we have reviewed the concept of text summarization which can characterize different approaches to text summarization we compare and discuss method that people used for multi document summarization. We have also described evaluation method for automatic text summarization. In the future, we will try to develop an algorithm or new model that supports multi document summarization area with fuzzy classifiers and natural language generation techniques.

References

[1] Vahed Qazvinian, Leila Sharif Hassanabadi and Ramin Halavati, "Summarizing text with a genetic algorithm-based sentence extraction", International Journal of Knowledge Management, vol. 2, no. 4, p p. 426-444, 2008.

[2] Pierre-Etienne Genest, Guy Lapalme, "Framework for Abstractive Summarization using Text-to-Text Generation", Workshop on Monolingual Text-To- The motivation of my research Text Generation, vol.23, no.9, p p. 64–73, June 2011.

[3] Kavita Ganesan, Cheng Xiang Zhai and Jiawei Han, "A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions", International Conference on Computational Linguistics, vol.44, no.12, p p. 847-864, 2010.

[4] Kathleen McKeown, "Query-focused Summarization Using Text-to-Text Generation: When Information Comes from Multilingual Sources", Language Generation and Summarisation, vol.6, no.3, p p.48-55,2009.

[5] Elena Lloret and Manuel Palomar, "Analyzing the Use of Word Graphs for Abstractive Text Summarization", Advances in Information Mining and Management, vol.1, no.5, p p.61-66, 2011.

[6] Oi Mean Foong, Alan Oxley and Suziah Sulaiman, "Challenges and Trends of Automatic Text Summarization", International Journal of Information and Telecommunication Technology, vol.1, no.1, p p.34-39, 2010.

[7] Heng Ji, Benoit Favre and Wen-Pin Lin, "Open-domain Multi-Document Summarization via Information Extraction: Challenges and Prospects", Theory and applications of natural language, vol.1, no.9, p p. 177-183,2013.

[8] Mohamad Ali Honarpisheh, Gholamreza Ghassem-Sani and Ghassem Mirroshandel, "A Multi-Document Multi-Lingual

[9] Automatic Summarization System", ACM SIGIR Conference on Research and Develop-ment in Information Retrieval, vol.2, no.4, 735-739, 2009.

[10] Chong Long, Minlie Huang, Xiaoyan Zhu and Ming Li, "Multi-Document Summarization by Information Distance", IEEE International Conference on Data Mining, vol.5, no.2, p p.866-871, 2009.

[11] Wen-tau Yih, Joshua Goodman, Lucy Vanderwende and Hisami Suzuki, " Multi-Document Summarization by Maximizing Informative Content-Words", International joint conference on Artificial intelligence, vol. 6, no.12, p p.1776-1782, 2007.

[12] Jackie Chi Kit Cheung and Gerald Penn, "Towards Robust Abstractive Multi-Document Summarization: A Caseframe Analysis of Centrality and Domain", Annual Meeting of the Association for Computational Linguistics, vol.4, no.9, p p.335-348, 2013.

[13] Vikrant Gupta, Priya Chauhan and Sohan Garg, " An Statistical Tool for Multi-Document Summarization", International Journal of Scientific and Research Publications, vol. 2, no. 5, p p. 1-5, May 2012.

[14] Jade Goldstein, Vibhu Mittal t, Jaime Carbonell and Mark Kantrowitzt, " Multi-Document Summarization by Sentence Extraction", ANLP/NAACL Workshop on Automatic Summarization, vol. 54, p p. 40-48, 2005.

[14] Naresh Kumar Nagwani and Shrish Verma, " A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm", International Journal of Computer Applications, vol. 17, no.2, p p. 36-40, March 2011.

[15] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, p p.325-335, July 2009.

[16] Giuseppe Carenini and Jackie Chi Kit Cheung, "Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality", International Natural Language Generation Conference, vol. 52, no.11, p p.33-41, 2008.

[17] Pierre-Etienne Genest and Guy Lapalme, "Fully Abstractive Approach to Guided Summarization", Meeting of the Association for Computational Linguistics, vol.2, p p. 354–358, 2012.

[18] Fei Liu and Yang Liu, "From Extractive to Abstractive Meeting Summaries: Can It Be Done by Sentence Compression", Association for Computational linguistics -IJCNLP, vol.3, no.1, p p.261-264, 2009.