Review on Various Text Summarization Evaluation Methods

Narendra Sahu
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
sahun1000@gmail.com

Vrajesh Chawra
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
chawra05@gmail.com

Abstract

In modern days, the day-to-day hustle-bustle does not allow a human being to assign time for manually summarizing variety of lengthy documents. Hence it is of ultimate importance to devise for an application that will make possible the automated text summarization. To familiarize oneself with a subject area summaries play an important role. Text Summarization is a challenging problem these days. Summarization is very interesting and useful task that gives support to many other tasks as well as it takes advantage of techniques developed for related Natural Language Processing tasks. Evaluating summaries and automatic text summarization systems are not a straightforward process. This review paper discusses an overview of text summarization, various evaluation approaches on intrinsic and extrinsic techniques. In principle, text summarization is achieved because of the naturally occurring redundancy in text and because important (salient) information is spread irregularly in textual documents. Recognizing the redundancy is a challenge that hasn‘t been fully resolved yet.

Keywords— *Text summarization, Extraction-based summarization, Abstraction-based summarization, Automatic Text Summarization.*

I. INTRODUCTION

Nowadays, due to rapid growth of broadcast systems in the field of Internet, there is huge amount of information being available online. Search Engines obtain the process of constant indexing with respect to accumulate the rising information in World Wide Web. As soon as the user enters the search request, the required documents are retrieved. Here is a classic problem of Information Overload comes into play as the search engine fetches hundreds of documents as the search results.

The fetch time for the finding of document is very less, the user need to go through the whole documents in order to reach to the document he/she is searching for, because most of the in experiment users which are reluctant to make an unmanageable effort of going through each of the documents separately. With these massive amount of information available and their need for summarization not only for saving the search time but also for understanding the required information available. Summarization has been topic of study in the field of Computer Science for a very long time. Text Summarization [1] has become the study of hour. Although the attempts to generate automatic summaries began 50 years ago [2], in recent automatic Text Summarization has experienced an exponential growth [3], [4], [5] due to these new technologies.

Text summarization has involve the interest of so many scientists and researchers in last few years, meanwhile the textual data has become very useful for many real world applications and their problems. In the fast-moving world, it is very difficult to read all the text content. Therefore, the need for text summarization is being in the highlight. Automatic text summarization is a technique which concentrates the larger text to a shorter text which involves the essential information.

A. Types of Summarization

There are mainly two types of summaries: Extractive summaries and Abstractive summaries. Extractive summaries are generated by extracting the whole sentences from the source text document. To form the summary of extractive methods which work by choosing a subset of existing words, phrases, or sentences in the original form of text. Abstractive summaries are generated by the reformulating sentences of the source text document. Abstractive methods develop an internal semantic representation and natural language generating techniques is used to create a summary that is closer to what a human may generate. This type of summary could involve words not explicitly present in the original text.

Extraction-based Summarization

In this, without changing the objects themselves the automatic system extracts objects from the whole collection. Examples involve the key phrase extraction, in which the aim is to choose an individual words or phrases to "tag" a text document. The goal for document

summarization is to choose entire sentences (without modifying them) for making a short paragraph summary. In the same manner, from the collection system extracts images without changing itself, is image collection summarization.

Abstraction-based Summarization

Extraction techniques replica the information which are most important by the system to the summary (ex: key clauses, sentences or paragraphs), while abstraction involves paraphrasing sections of source document. In general, abstraction can shorten a text more strongly than extraction summaries can do, but the programs that can do is harder to generate as they need for the usage of natural language generation technology, which itself is a developing field. In abstractive summarization an abstract synopsis like that of a human is done, while majority of summarization systems are extractive where choice of subset of sentences to place in a summary.

B. Automatic Text Summarization

Automatic text summarization is a technique that gets a source in the form of text documents and presents the most related content in a condensed form as the user or task needs. Technologies that can make a relevant summary are taken such as writing style, length, and syntax. Traditionally, the process of automatic text summarization has been decomposed into three main stages [6], [2], [7]. The Spark Jones [7] approach, which is: the source text is interpreted to obtain a text representation, then transforming the text representation into a summary representation, and from summary representation summary of text is generated.

Effective summarizing requires an explicit and detailed analysis of context factors. In [7] three classes of context factors are distinguished: input, purpose and output factors. Whereas the Input factors describes the features of the text to be summarized vitally clarify the way of a summary can be obtained. This particular feature categorizes into three groups: text form; subject type and unit. Purpose factors are very important factor which fall under three categories: denotes the context within the summary to be used; the audience and use. Output factors group: implies with the material (i.e. content) format and style.

II. LITERATURE SURVEY

Karmakar Surajit [8] proposed review paper discusses a few of the extractive methods of text summarization. An extractive summary is a selection of important sentences from the original text that briefly describes the original text. Various methods of extractive approach have emerged in the past. But it is hard to say how much greater interpretive sophistication, at sentence or text level contributes to performance. Without the use of Natural Language Processing, the generated summaries may not be much accurate in terms of semantics. If the input documents cover multiple topics, it becomes difficult to generate a balanced summary. For this purpose, deciding proper weights of individual features is important as quality of final summary depends on it.

S. Ranwez [9] proposed a novel summarization framework (Opinosis) that uses textual graphs to generate abstractive summaries of highly redundant opinions. Evaluation results on a set of review documents show that Opinosis summaries have better agreement with human summaries compared to the baseline extractive method. The Opinosis summaries are concise, reasonably well formed and communicate essential information.

Reeve Lawrence H. [10] proposed the frequency of domain-specific concepts as a feature for identifying salient sentences in biomedical texts. We presented an evaluation of several existing summarization systems to determine a performance baseline. We then evaluated a stateof-the-art frequency algorithm using both terms and concepts as item units to show the use of the frequency of concepts is as effective, and sometimes an improvement over, the use of frequency of terms. We developed a new algorithm based on frequency distribution modeling and evaluate it using terms as well as concepts. In either case, our frequency distribution algorithm outperforms a current state-of-the-art frequency-based algorithm at the cost of higher computational complexity.

Sarda A.T. [11] proposed that people prefer to read summary of any document instead of reading whole document because the summary includes core part of the document. The selection of features & the selection of summary sentences to form better summary using neural network.

Ramezani Majid [12] proposed the paper in which author describe the existence of an automatic text summarization system will definitely facilitate it for one who deals with reading and results in time budgeting. In this paper with the aim of ontology-based automatic summarization of Persian documents, a system has been proposed which recognizes the semantic relations available in the text by using the FarsNet ontology and extracts the most important sentences for inclusion in the summary.

Table I. Shows comparisons of existing methods and its limitation

S. No.	Ref. No.	Method Used	Data Source	Approach	Strength	Limitation
1	8	Tree Based Method	Text data	It uses a dependency tree to represent the text of a document. -It uses either a language generator or an algorithm for generation of summary	It walks on units of the given document read and easy to summary.	It lacks a complete model which would include an abstract representation for content selection.
2	9	Semantic Graph Based Method	Review document	This method is used to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document	It produces concise, coherent and less redundant	This method is limited to single document abstractive summarization.
3	10	Multimodal semantic model	Biomedical text	A semantic model, which captures concepts and relationship among concepts, is built to represent the contents of multimodal documents.	An important advantage of this framework is that it produces abstract summary, whose coverage is excellent because it includes salient textual and graphical content from the entire document.	The limitation of this framework is that it is manually evaluated by humans.
4	11	Semantic Graph Based Method	articles	This method is used to summarize a document by creating a semantic graph called Rich Semantic Graph (RSG) for the original document, reducing the generated semantic graph.	It produces concise, coherent and less redundant and grammatically correct sentences.	This method is limited to single document abstractive summarization.
5	12	Ontology Based Method	Persian documents	-Use ontology (knowledge base) to improve the process of summarization. -It exploits fuzzy ontology to handle uncertain data that simple domain ontology cannot.	-Drawing relation or context is easy due to ontology - Handles uncertainty at reasonable amount	-This approach is limited to Chinese news only. - Creating Rule based system for handling uncertainty is a complex task.

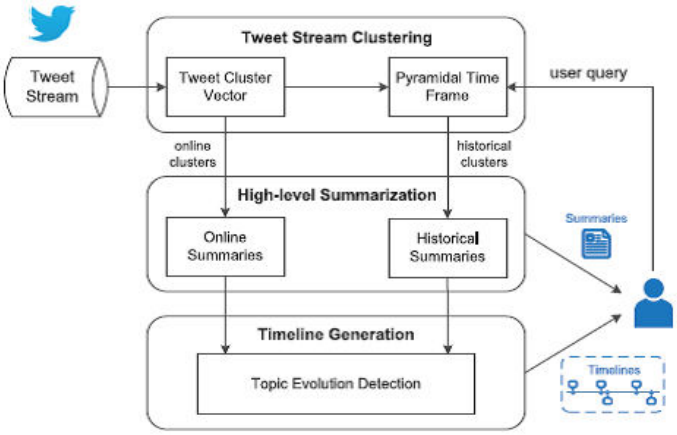

Fig.1. Workflow of Text Summarization Method

DKL is the Kullback-Leibler divergence (KLD) which defines the divergence of distribution M from S. The motivation of below Equation is analogous to that of maximal marginal relevance (MMR). In query-oriented summarization, MMR combines query relevance and information novelty.

$$D_{KL}(S||M) = \sum_{w \in V} p(w|S) log \frac{p(w|S)}{p(w|M)}.$$

III. TOOLS USED

There are many tools available for processing data and extracting user information from collection of datasets. Various tools can be used are described below.

a. Statistical Package of R: Used to train the datasets.
b. SOM-PAK package of Python: The SOM PAK program package contains all programs necessary for the correct application of the Self Organizing Map algorithm in the visualization of complex experimental data.
c. SciPy (pronounced “Sigh Pie”) is a Python-based ecosystem of open-source software for mathematics, science, and engineering.

IV. CONCLUSION

Due to the huge amount textual data generated daily, it is impossible to read all the data and from that it is impossible to extract the relevant information for a human being. For such type of problem, text summarization is become the one of the vital solution. Improvement in modification of data to its summary can be possible with efficient models, accuracy can be achieved up to greater extent in this area too.

In this paper, a brief summary of automatic text summarization techniques for various text document has been described. From the review, we can observe that a good work has been done for various types of document. Automatic summarization system for many other types of document is still lacking. We can also conclude that various combination of features works differently for different types of content. Hence, it is challenging to create a single summarizer for various types of content.

REFERENCES

[1] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[2] Mani, I., Maybury, M. T., Ed. Advances in Automatic Text Summarization. The MIT Press, 1999.

[3] Hovy, E., Lin, C. Y., Zhou, L., et al. Automated Summarization Evaluation with Basic Elements. In Proceedings of the 5th International Confer- ence on Language Resources and Evaluation (LREC). Genoa, Italy, 2006.

[4] Jackson, P., Moulinier, I. Natural language processing for online applications. John Benjamins Publishing Company, 2002. [5] Padro Cirera, L., Fuentes, M.J.,Alonso, L., et al. Approaches to Text Summarization: Questions and Answers. Revista Iberoamericana de Inteligencia Arti_cial, ISSN 11373601,(22):79{102, 2004.

[5] Hovy, E. H. Automated Text Summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics, chapter 32, pages 583{598.Oxford University Press, 2005.

[6] Spark Jones, K. Automatic summarizing: factors and directions. In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization. MIT Press, 1999.

[7] Karmakar Surajit, Lad Tanvi, Chothani Hiten, “A Review Paper on Extractive Techniques of Text

Summarization", International Research Journal of Computer Science (IRJCS), Issue 1, 2015, Vol. 2.

[8] S. Ranwez, B. Duthil, M. F. Sy, J. Montmain, P. Augereau, and V. Ranwez, "How ontology based information retrieval systems may benefit from lexical text analysis," in New Trends of Research in Ontologies and Lexical Resources, P. Vossen, L. Qin, and E. Hovy, Eds. Springer, 2013, pp. 209–231.

[9] Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C., Schwimmer Tamara A., Brooks Ari D., "Concept Frequency Distribution in Biomedical Text Summarization", ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006.

[10] Sarda A.T. and Kulkarni A.R., "Text Summarization using Neural Networks and Rhetorical Structure Theory", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, 2015.

[12] Ramezani Majid, Feizi-Derakhshi Mohammad-Reza, "Ontology Based Automatic Text Summarization Using Fars Net", ACSIJ Advances in Computer Science: an International Journal, 2015, Vol. 4, Issue 2, No.14.