RESEARCHERID
THOMSON REUTERS

[Tomar* *et al.*, 6(7): July, 2017]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

REVIEW OF AUTOMATIC SUMMARIZATION OF WEB DOCUMENTS

Shivani Tomar[*1] & Deepika Gupta[2]

[*1]M.tech student at M.I.E.T. College, Dr. A.P.J. Abdul Kalam Technical University (APJAKTU) Uttar Pradesh, India

[2]Assistant Professor at M.I.E.T, APJAKT University, Uttar Pradesh, India

DOI: 10.5281/zenodo.834544

ABSTRACT

A large number of information of all the domains are available online in the form of hyper text in web pages. Peoples are less interested to read whole of the documents, they keen to know that whether the document is of their interest or not, this is possible if the summary of the document is available various algorithms are present today for summarization of documents. In this paper various available summarization techniques are discussed along with their performance evaluation and applications.

KEYWORDS: summarization, summary, performance evaluation.

I. INTRODUCTION

The amount of data available today is big and increasing continuously. The Internet provides web pages, news articles, email, access to the databases around the world and much more. For individual users it is impossible to analyze and use the data effectively, leading to so called information overload. Document retrieval (DR) proved to be very useful in last decade reducing the burden on the users at various levels. DR is defined as, "given a set of documents and a query find a sub-set of documents most relevant to the query while retrieving as few irrelevant documents as possible". Many search engines based on this idea are in use; Google, Yahoo Search and AltaVista are some of them. Although the term document retrieval is more appropriate for this process. The term information retrieval has become well established. The difference between the two will be explained soon. The problem of information overload is not solved here. DR retrieves number of documents still beyond the capacity of human analysis, e.g. at the time of writing the query "information retrieval" in Google returned more than 30,100,000 results. Thus DR is not sufficient and we need a second level of abstraction to reduce this huge amount of data: the ability of summarization. This work tries to address this issue and proposes an automatic text summarization (TS) technique. Roughly summarization is the process of reducing a large volume of information to a summary or abstract preserving only the most essential items. The. A TS system [1] has to deal with natural language text and the complexities associated with natural language are inherited in the TS systems. Natural language text is unstructured and could be semantically ambiguous. TS is a very hard task as the computer must somehow understand what is important and what is not to be able to summarize. A TS system must interpret the contents of a text and preserve only most essential items. This involves extraction of syntactic and semantic information from the text and using this information to decide essentialness of the items. The following sub-section describes the need of TS systems with an example.

The process of summarization is described below-

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

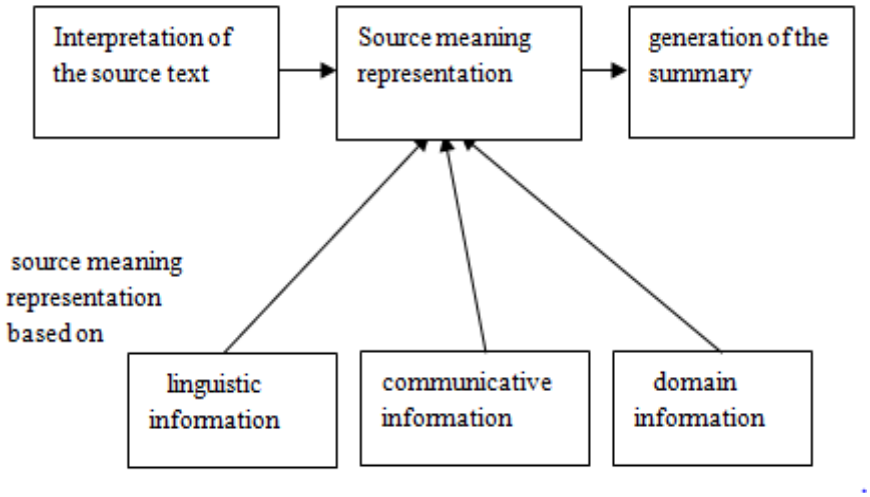

Figure 1: Summary Generation

II. AUTOMATIC TEXT SUMMARIZATION

Although there is no common definition of a summary, here a summary is defined as a condensed version of a text while preserving most of the information. Traditionally summarization [3] is performed by human professionals, like journalists and scientific writers. But due to so called information overload the amount of data available is well beyond the capacity of a single human and this call for the use of automatic means of summarization. Recently automatic summarization has gained widespread interest due to increasing amount of documents available in the electronic format which can be readily subjected to computer processing.

Text summarization or rather automatic text summarization [5] corresponds to the process in which a computer creates a shorter version of the original text (or a collection of texts) still preserving most of the information present in the original text. This process can be seen as compression and it necessarily suffers from information loss. Thus a TS system must identify important parts and preserve them. What is important can depend upon the user needs or the purpose of the summary. This chapter explains various aspects of text summarization followed by an overview of the past research.

The technique has its roots in the 60's and has been developed during 30 years, but today with the Internet and the WWW the technique has become more important.

Text summarization has number of applications; recently number of applications uses text summarization for the betterment of the text analysis and Knowledge representation.

The phenomenon of information overload has meant that access to coherent and correctly-developed summaries is vital. As access to data has increased so has interest in automatic summarization. An example of the use of summarization technology is search such as Google.

The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its in formativeness.

This simple definition captures three important aspects that characterize research on automatic summarization:

- Summaries may be produced from a single document or multiple documents.
- Summaries should preserve important information.
- Summaries should be short.

III. METHODOLOGY OF TEXT SUMMARIZATION

Text Summarization methods can be classified into extractive and abstractive summarization.

- **Extraction –** An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. In extractive summarization, the summary consists of verbatim textual units (e.g., phrases, sentences and paragraphs) from the original text. In this work the extraction unit is defined as a sentence. Sentences are well defined linguistic entities and have self contained meaning. So the aim of an extractive summarization system becomes, to identify the most important sentences in a text. The assumption behind such a system is that there exists a subset of sentences that present all the key points of the text. Extracting can be seen as selective copy and paste operation.
- **Abstraction –** An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

IV. APPLICATION OF TEXT SUMMARIZATION

Text Summarization has many potential application areas. Although the current state of the technology does not guarantee complete solution TS can assist by producing draft summaries. Some of the applications are listed here-

1. To summarize news to SMS or WAP-format for mobile phones/PDA.
2. To let a computer synthetically read the summarized text. Written text can be too long and boring to listen to.
3. In search engines to present compressed descriptions of the search results (see the Internet search engine like Google).
4. In keyword directed subscription of news which are summarized and pushed to the user (see) Nyhetsguident (In Swedish).
5. To search in foreign languages and obtain an automatically translated summary of the automatically summarized text.

V. EVALUATE A SUMMARY

1. Choose a granularity (clause; sentence; paragraph)
2. Create a similarity measure for that granularity (word overlap; multi-word overlap, perfect match)
3. Measure the similarity of each unit in the new to the most similar unit(s) in the gold standard, measure Recall and Precision.
4. Compression Ratio: *CR = (length S) / (length T)*
   Retention Ratio: *RR = (info in S) / (info in T)*

- **Summary Categorization Test**

Steps performed are as follows-

1. 1000 newspaper articles from each of 5 categories.
2. Systems summarize each text (generic summary).
3. Humans categorize summaries into 5 categories.
4. Testers measure *Recall* and *Precision*, combined into *F*: *How correctly are the summaries* classified, compared to the full texts?

- **Ad Hoc (Query-Based) Test**

1. 1000 newspaper articles from each of 5 categories.
2. Systems summarize each text (query based summary).
3. Humans decide if summary is relevant or not to query.
4. Testers measure *R* and *P*: *how relevant are the summaries to their queries?*

VI. CONCLUSION

The paper describes the automatic summarization of text documents, need of text summarization, applications of text summarization and performance evaluation of created summarization tool. This paper helps the researchers to develop summarization tools and also evaluate their tool performance.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

VII. REFERENCES

[1] James Allan, Jaime Carbonell, George Doddington,Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study: Final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.
[2] Chinatsu Aone, M. E. Okurowski, J. Gorlinsky, and B. Larsen. A scalable summarization system using robust NLP. In Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 66-73, Madrid, Spain, 1997.
[3] Breck Baldwin and Thomas S. Morton. Dynamic coreference-based summarization. In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada, Spain, June, 1998.
[4] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In Proceedings of the CL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pages 10-17, Madrid, Spain, 1997.
[5] A. Siddharthan, A. Nenkova, and K. McKeown. Syntactic simplification for improving content selection in multi-document summarization. In Proc. of COLING, 2004.
[6] L. Vanderwende, H. Suzuki, and C. Brockett. Microsoft Research at DUC2006: Taskfocused summarization with sentence simplification and lexical expansion. In Proc. of DUC, 2006.
[7] X. Wan and J. Yang. Improved affinity graph based multi-document summarization. In Proceedings of HLT-NAACL, Companion Volume: Short Papers, pages 181–184, 2006.
[8] D.Zajic, B. Dorr, andR. Schwartz.Automatic headline generation for newspaper stories. In Proc. of DUC, 2002.
[9] D. Zajic, B. Dorr, J. Lin, C. Monz, and R. Schwartz. A sentence-trimming approach to multidocument summarization. In Proc. of DUC, 2005.

CITE AN ARTICLE

Tomar, Shivani , and Deepika Gupta. "REVIEW OF AUTOMATIC SUMMARIZATION OF WEB DOCUMENTS." *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY* 6.7 (2017): 816-19. Web. 25 July 2017.