Literature Review on Optimizing Accuracy of Document Summarization

POONAM W. KOLHE[1], Prof. ASHISH KUMBHARE[2]

[1]Research Scholar, [2]Assistant Professor
Department of Computer Science & Engineering,
Shri. Balaji Institute of Technology & Management Betul, Madhya Pradesh, India

Abstract: **- In today's fast-growing information age we have an abundance of text, especially on the web. New information is continuously being generated. The growing accessibility of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community. Often due to time constraints we are not able to consume all the data available. It is therefore essential to be able to summarize the text so that it becomes easier to ingest, while maintaining the essence and understandability of the information. In this paper we aim to design an algorithm that can summarize ainput document by extracting action word and attempting to modify this extraction using a NLP tools. Our main goal is to reduce a given body of text to a fraction of its size, maintaining coherence and semantics of original text and it is Multi-lingual system.**

Index terms **- Automatic Summarization, Extraction, Abstraction, NLP.**

---

1. INTRODUCTION

Text summarization is a process to express the content of a document in a condensed form that meets the needs of the user. More and more electronic data is available on the Internet and it is not possible to read everything and hence some form of information condensation is needed. Summarization serves as a tool which helps the user to efficiently find useful information from immense amount of information.

Text summarization can be used by various applications; for instance researchers need a tool to generate summaries for deciding whether to read the entire document or not and for summarizing information searched by user on Internet. News groups can use multi document summarization to cluster the information from different media and summarize.

The subfield of summarization has been investigated by the Natural Language Processing community for nearly the last half century. Define a summary as a text that is produced from one or more texts, that convey important information in the original text, and that is no longer than half of the original text and usually significantly less than that". 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.

Natural language processing (NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human language. Natural language processing is a process of developing a system that can process and produce language as good as human can produce. The use of World Wide Web has increased and so the problem of information overload also has increased. Hence there is a need of a system that automatically retrieves, categorize and summarize the document as per users need. Document summarization is one possible solution to this problem.

2. TEXT SUMMARIZATION

2.1 Definition and types-

A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). According to, text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).

When this is done by means of a computer, i.e. automatically, we call this Automatic Text Summarization. Despite the fact that text summarization has traditionally been focused on text input, the input to the summarization process can also be multimedia information, such as images, video or audio, as well as on-line information or hypertexts. Furthermore, we can talk about summarizing only one document or multiple ones. In that case, this process is known as Multi-document Summarization (MDS) and the source documents in this case can be in a single-language or in different languages.

The output of a summary system may be an extract (i.e. when a selection of "significant" sentences of a document is performed) or abstract, when the summary can serve as a substitute to the original document. We can also distinguish between generic summaries and user-focused summaries (a.k.a query-driven). The first type of summaries can serve as surrogate of the original text as they may try to represent all relevant features of a source text. They are text-driven and follow a bottom-up approach using IR techniques. The user-focused summaries rely on a specification of a user information need, such a topic or query. They follow a top-down approach using IE techniques.

2.2 Automatic summarization-

It is the process of condensing textual content into a concise form for easy digestion by human, using a computer program. Summaries can be produced from a single document or multiple documents; they should be short and preserve important information. Summarization can be extractive or abstractive:-Extractive summarization aims to extract important and relevant parts of the given content and fuse them coherently. Abstractive summarization aims at paraphrasing the source document, similar to manual summarization. Automatic text summarization is a useful tool when there is a lot of textual information to be analyzed manually. Automatic summarization is used to condense the large amounts of textual data. This achieves the following benefits:

- Firstly, several redundancies can be removed. The user does not waste time reading repetitive data.
- Secondly, summarization allows you to remove data that is not essential to the understanding of the document.

There are many methods to proceed with automatic text summarization. In our model we use an extractive technique to obtain the summary from the given text. This summary is then improved further by replacing a few parts of it using an abstractive technique. The extraction of sentences from the document is done keeping coherence in mind and therefore the summary maintains the essence of the original document. The sentences are then ranked using a text- ranking algorithm, namely TextRank and the final cluster or summary is formed.

The important functions of the summarizer are:-

- Reducing a single document to a user-defined fraction of its original size while maintaining coherence.
- Choosing the most relevant and important sentences from the text.
- Improving the abstraction and/or length of the summary by using a thesaurus to replace semantically related units.

In effect, we aim to extractive summarize a single English document, not more than 300 sentences long, to a fraction of its original size, while maintaining cohesion, and then use a lexical database to abstract the generated summary.

The software uses the external tool WordNet to abstract the generated summary. WordNet is a lexical database that groups words by semantic relations. The Natural Language Toolkit (NLTK) for Python is used to access the database through the program. ROUGE is used for evaluating the summarization.

3. PROCESS OF AUTOMATIC TEXT SUMMARIZATION-

Traditionally, summarization has been decomposed into three main stages. We will follow the Sparck Jones approach, which is:

- Interpretation of the source text to obtain a text representation,
- Transformation of the text representation into a summary representation, and
- Finally, generation of the summary text from the summary representation Effective summarizing requires an explicit and detailed analysis of context factors. Sparck Jones in distinguishes three classes of context factors: input, purpose and output factors.

Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text.

1.3.1 Frequency based approach

1] Term Frequency:

The term frequency is very important feature. TF (term frequency) represents how many time the term appears in the document (usually a compression function such as square root or logarithm is applied) to calculate the term frequency. The term identifying sentence boundaries in a document is based on punctuation such as (. (, “, [, {, etc.) and split into sentences. These sentences are nothing but tokens.

2] Keyword Frequency:

The keywords are the top high frequency words in term sentence frequency. After cleaning the document calculate the frequency of each world. And which words have the highest frequency these words are called keywords. The words score are chosen as keywords, based on this feature, any sentence in the document is scored by number of keywords it contains, where the sentence receives 0.1 score for each key word.

3] Stop word filtering:

In any document there will be many words that appear regularly but provide little or no extra meaning to the document. Words such as 'the', 'and', 'is' and 'on' are very frequent in the English language and most documents will contain many instances of them. These words are generally not very useful when searching; they are not normally what users are searching for when entering queries.

4. TYPES OF TEXT SUMMARIZATION TECHNIQUES

Different types of summary might be useful in various applications and summarization systems can be categorized based on these types. In addition to abstract and extract, there are various types of summaries. A full understanding of the major dimensions of variation, and the types of reasoning required to produce each of them, is still a matter of investigation. This makes the study of automated text summarization an exciting area in which to work. Various summarization methods can be compared based on the type of summary and application. Summarization system can be classified into the following categories, they are:

1) Based on approaches

There are two strategies for summarization those are summarization by extraction, which consists of extracting source sentences as it is and adding into a summary and summarization by abstraction, which involves generating novel sentences for the summary. The need for abstraction is especially high when opinions are diverse.

Summarization by extractive just extracts the sentences from the original document and adds them to summary. Extractive method is usually easy to implement and is based on statistical features not on semantic relation with sentences. Therefore the summary generated by this method tends to be inconsistent.

Summarization by abstraction needs understanding of the original text and then generating the summary which is semantically related. It provides more generalized summary but it is difficult to compute.

2) Based on type of details

Based on type of detail summary can be either informative or indicative. An indicative summary is used for quick view of a lengthy document and it provides only the main idea of the original text. These are usually small and it encourages a user to read the original document. For example while purchasing any novel a buyer reads the summary provided at back side of novel.

Informative summary serves as a substitution to the original document. It provides the concise information about the original document to the user.

3) Based on type of content

This classification is based on the type of content in the original document. Generic summarization is system which can be used by any type of the user and summary does not depend on the subject of the document. All the information is at same level of importance and which is not user specific.

Query-based summarization is question answer type where the summary is the result of query. It provides the users view and cannot be used by any type of user.

4) Based on limitation

Summary can be classified based on limitation of input text. Genre specific systems only accept special type of input like newspaper articles, stories, manuals etc. Limited to the type of input they can accept.

Domain independent system can accept different type of text. They are not dependent on the domain and can be used by any type of user. There are few systems that are domain dependent.

5) Based on number of input documents

Summarization can be classified based on whether a system accepts one or more documents as input. Single document summarization can accept only one document as input. They are usually easier to produce as it involves summarization of a single document.

Multi-document summarization accepts several documents of same topic as an input. It is more difficult to implement as there are multiple documents to summarize.

6) Based on language

Mono lingual system only accepts documents with specific language and output is based on that language only. Multi-lingual systems can accept documents in different languages and produce summary of different languages.

4. REVIEW OF RELATED WORK

This literature review describes different summarization techniques used for text summarization. It presents the previous work which had done on text summarization, including the analysis of various text summarization techniques.

[1] Dipanjan Das Andre F.T. Martins (November 21, 2007)

This survey emphasizes extractive approaches to summarization using statistical methods. A distinction has been made between single document and multi-document summarization.

[2] Archana AB, Sunitha. C (2013)

Archana AB, Sunitha. C describes comparative study on four different approaches to automatic text summarization. Text summarization approaches based on Neural Network, Graph Theoretic, Fuzzy and Cluster have, to an extent, succeeded in making an effective summary of a document.

[3] Simran kaur1, wg.cdranil chopra2 (02 Mar 2016)

Simran kaur1, wg.cdranil chopra2presented approach towards 'k means clustering Automated Text Summarization'. This approach attempts to generate a text summary from the article of newspapers, while avoiding the repetition of identical or similar information and presenting the information in such a way that makes sense to the reader.

Advantages- Another major issue to be handled in this study is to generate a "user-friendly" summary at the end.

[4] Anjali R. Deshpande #1, Lobo L. M. R. J. *2 (August 2013)

This paper presented a new approach to multi-document summarization. It is the clustering based approach that groups first, the similar documents into clusters & then sentences from every document cluster are clustered into sentence clusters. And best

scoring sentences from sentence clusters are selected in to the final summary. We find similarity between each sentence & query. To find similarity "cosine similarity measure" is used
Advantages- This method ensures good coverage and avoids redundancy.

[5] Priya Ganguly1, Dr. Prachi Joshi2 (, January 2016)
This survey paper explains about various accounts of extractive summarization. An extractive summery is choice of main sentences from the corresponding documents. The importance of sentences is based on applied statistical and linguistic features of sentences.
Advantages-Itis usually easy to implement and is based on statistical features not on semantic relation with sentences.

[6] N. R. Kasture1, NehaYargal 2, NehaNityanand Singh3, Neha Kulkarni4 and Vijay Mathur5 (November 2014)
The proposed system includes understanding the main concepts and relevant information of the main text and then expressing that information in short and clear format. Abstractive summarization techniques can again be classified into two categories- structured based and semantic based methods. So, author follows the abstractive summarization methods.
Advantages- Abstractive summarization methods produce more coherent, less redundant and information rich summery. Generating abstract using abstractive summarization methods is a difficult task since it requires more semantic and linguistic analysis.

[7] Richa Sharma, Prachi Sharma (April 2016)
This survey paper gives the details on extractive text summarization features and its methods. The extractive text summarization is a process of selecting important sentences from the document and including those sentences as it is in the final summary of the document and the selection procedure of sentences is done on the basis of statistical and linguistic features of the sentences.

[8] Vishal Gupta and Gurpreet Singh Lehal, (August 2010.)
In this paper author describes the extractive summarization methods which comprises of two parts Pre Processing and Processing. In this paper, pre-processing step is further divide into other sub processes which are sentence segmentation, stop word removal and stemming. In processing step, the weights are given to the features used for extraction of summary from the large document respectively.

[9] Saranyamol C S and Sindhu L (2014)
In this paper the author describes about the various techniques used in automatic text summarization which are extractive text summarization and abstractive text summarization respectively.

[10] Rafael Ferreira, Luciano de Souza Cabrala, Rafael DueireLins , Gabriel Pereira Silva , Fred Freitas , George D.C. Cavalcanti , Rinaldo Lima a, Steven J. Simske , Luciano Favaro (2013)
This paper, gives the brief description of various features used to perform extractive summarization and it also describes the methods for summary evaluation.

[11] K. Vimal Kumar, DivakarYadav(2015)
This paper mainly laid emphasis most importantly on the Hindi text summarization. It also describes various features used for the Hindi summarization using extractive approach of text summarization. The author had proposed a system which can generate the summary with 85 % accuracy.

[12] Vishal Gupta (2013)
The author of this paper has proposed a hybrid algorithm for Hindi and Punjabi text summarization. The algorithm proposed by the author is the first algorithm which can summarize both Hindi as well as Punjabi text. It also suggests some new methods for Hindi and Punjabi text.

[13] **Dr. Annapurna P Patil ,ShivamDalmia, Syed Abu Ayub Ansari, TanayAul, VarunBhatnagar (2014)**
According to the autherextractive summary is advantageous for certain formats of documents. The abstraction is slight and marginally improves readability and length. However, the abstraction does not strictly generate an abstractive summary in the true sense, as natural language processing techniques are not used.
Advantages-The advantage of this method is that it operates completely algorithmically, and does not require sophisticated techniques. However, often the replacement is not sufficiently appropriate or ideal.

[14] Rafael Ferreira∗†, FredericoFreitas∗, Luciano de Souza Cabral∗, Rafael DueireLins∗, Rinaldo Lima∗, Gabriel Franc¸a∗, Steven J. Simske‡, and Luciano Favaro§(2014)
This paper is the thesis that the quality of the summary obtained with combinations of sentence scoring methods depend on text subject. Such hypothesis is evaluated using three different contexts: news, blogs and articles. The results obtained show the validity of the hypothesis formulated and point at which techniques are more effective in each of those contexts studied.
Advantages-This paper suggests and brings experimental evidence that the effectiveness of sentence scoring methods for automatic extractive text summarization algorithms depends on the kind of text one wants to summarize, the length of documents,

the kindof language used, and their structure. Different combinations of sentence scoring algorithms yield different results both in the quality of the summaries obtained and the time elapsed in generating them.

6. CONCLUSION:

In this paper, we have described a general overview of automatic text summarization. The status, and state, of automatic summarizing has radically changed through the years. It has specially benefit from work of other asks, e.g. information retrieval, information extraction or text categorization. Research on this field will continue due to the fact that text summarization task has not been finished yet and there is still much effort to do, to investigate and to improve. Definition, types, different approaches and evaluation methods have been exposed as well as summarization systems features and techniques already developed. In the future we plan to contribute to improve this field by means of improving the quality of summaries, and studying the influence of other neighbor tasks techniques on summarization.

ACKNOWLEDGMENT:

The author is thankful to Assistant Prof. AshishKumbhare, faculty of Computer science and Engineering, SBITM, Betul, RGPV University Bhopal, MP for providing necessary guidance to prepare this paper.

REFERENCES:

[1] Archana AB, Sunitha. C "An Overview on Document Summarization Techniques" International Journal on Advanced Computer Theory and Engineering (IJACTE) Volume-1, Issue-2, 2013

[2] Dipanjan Das Andre F.T. Martins "A Survey on Automatic Text Summarization" Language Technologies Institute Carnegie Mellon University, November 21, 2007

[3] Simran kaur1, wg.cdranilchopra2 "Document Summarization Techniques" International Journal of Computer Science Engineering (IJCSE) Vol. 5 No.02 Mar 2016

[4] Anjali R. Deshpande #1, Lobo L. M. R. J. "International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- August 2013"

[5] Liadh Kelly, Johannes Leveling, Shane McQuillan, SaschaKriewel, Lorraine Goeuriot, Gareth Jones " Report on summarization techniques" European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development. 28 February 2013

[6] AniNenkova, Kathleen McKeown "A SURVEY OF TEXT SUMMARAZATION TECHNIQUES"

[7] Priya Ganguly1, Dr. Prachi Joshi2 International Journal of Science and Research (IJSR) Volume 5 Issue 1, January 2016

[8] SherifElfayoumy, Jenny Thoppil "A Survey of Unstructured Text Summarization Techniques" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 4, 2014

[9] N. R. Kasture1, NehaYargal 2, NehaNityanand Singh3, Neha Kulkarni4 and Vijay Mathur5 "A Survey on Methods of Abstractive Text Summarization" INTERNATIONAL JOURNAL FOR RESEARCH IN EMERGING SCIENCE AND TECHNOLOGY, VOLUME-1, ISSUE-6, NOVEMBER-2014

[10] Richa Sharma, Prachi Sharma "International Journal of Advanced Research in Computer Science and Software Engineering" Volume 6, Issue 4, April 2016

[11] Vishal Gupta &Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010.

[12] Saranyamol C S and Sindhu L, "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies, Vol. 5(6), pp. 7889-7893, 2014.

[13] Vishal Gupta and G.S Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, pp. 60-76, August 2009.

[14] Neelima Bhatia and ArunimaJaiswal, "Trends in Extractive and Abstractive Techniques in Text Summarization", International Journal of Computer Application (0975-8887), Vol. 117- No. 6, May 2015.

[15] V. Gupta and G.S. Lehal, "A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages," Journal of Emerging Technologies in Web Intelligence, Vol. 5, pp. 157-161, 2013.

16] Elena Lloret "TEXT SUMMARIZATION : AN OVERVIEW " Dept. Lenguajes y SistemasInformaticos Universidad de Alicante, Spain.

[17] Vishal Gupta," Hybrid Algorithm for Multilingual Summarization of Hindi and Punjabi Documents" 2013 in Springer International publishing Switzerland 2013.

[18] K. Vimal Kumar, DivakarYadav "An Improvised Extractive Approach for Hindi Text Summarization" Springer India 2015, J.K. Mandal et al. (eds.), Information Systems Design and Intelligent Applications, Advances in Intelligent System and Computing 339, DOI 10.1007/978-81-322-2250-7_28.

[19] Dr. Annapurna P Patil ,ShivamDalmia, Syed Abu Ayub Ansari, TanayAul, VarunBhatnagar "Automatic Text Summarizer" 2014 International Conference on Advances in Computing,Communications and Informatics (ICACCI).

[20] Rafael Ferreira∗†, FredericoFreitas∗, Luciano de Souza Cabral∗, Rafael DueireLins∗, Rinaldo Lima∗, Gabriel Franc¸a∗, Steven J. Simske‡, and Luciano Favaro§ "A Context Based Text Summarization System" 2014 11th IAPR International Workshop on Document Analysis Systems.