A Survey of Automatic Text Summarization

Niharika Verma, Prof. Ashish Tiwari
Computer Science & Engineering Department, Rajiv Gandhi Technical University
Airport Road, Gandhi Nagar Bhopal-462 033,India

Vindhya Institute of Technology and Science
Umrikheda, Khandwa Road, Indore (M.P.), India

***Abstract*- Text Summarization has now become a paramount research area. A technique in which a computer program summarizes a text is called Automatic Text Summarization. In today's world the web is an affluent source of information and data, therefore a substantial amount of data is scattered in the web domains. In addition of this there are number of articles which are publishing day by day and convivial blogs are additionally incrementing. Invigilation and monitoring of all the topics are not possible in a short time. Therefore automatic text summarization is auxiliary in order to reduce the texts with less time intricacy. In this paper a survey of automatic text summarization, its techniques and its applications are presented.**

Keywords- ***abstraction-predicated summary, automatic text summarization, extraction summary, feature extraction, text reduction.***

I. INTRODUCTION

The summary created by any summarization system must contain the most consequential points of the original text document. When a computer program reduces the text to obtain such a type of summary, this process is called as Automatic Text Summarization and this type of systems are called Automatic Summarization Systems. In this type of summarization systems generally the syntax is altered, but there will be no alteration in the semantics of the original text document [4]. The utilization of automatic summarization is incrementing day by day as the web information overload has become the incipient problem, and as the quantity of data has incremented so much. There is interest in computer program and technologies that can prepare an opportune summary by taking the variables such as length, indenting style and syntax. Now a day's search engines such as Google utilize this type of summarization system. Another example of summarization technology is Document Summarization.

The technique, where a computer program summarizes a document is kenned as Automatic Text Summarization [2], [4]. In the Automatic Summarization Systems a text is put into the computer and a mined text is obtained. Particularly there are two approaches of automatic summarization: extraction-predicated and abstraction-predicated. Extractive methods cull a subset of phrases, subsisting words, or sentences from the original text and then engender the summary[6]. On the other hand, abstractive methods use natural language generation techniques by building an internal semantic representation to compose a summary that is more proximate to the text what a human might additionally engender. This type of summary might contain words and phrases not explicitly present in the pristine text document [4]. The results obtained by abstractive methods are still quite impuissant, so extractive methods have got much interest and research has withal focussed on extractive-predicated approaches.

A. History of Automatic Text Summarization

The technique, Text Summarization has its roots in the early 1950's and has been developed during 30 years, but today with the incrementing utilization of Internet and the web, the technique has become more paramount [4]. Since 1997 Microsoft Word has a summarizer for documents of its utiliser. For Swedish SweSum is the first automatic text summarizer. Swedish news text in HTML/text format on the WWW is summarized by SweSum. During the summarization only 5-10 key words are taken and a mini summary is engendered. The highlighted (summarized) text with 84% precision is obtained. Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi (Persian) and German texts are withal summarized by SweSum.

Linguistically, heuristic and statistical are the methods in which automatic text summarization is predicated on, where the computer program calculates the paramount key words for the summarization (there are 700 000 possible Swedish ingresses that are possible in the Swedish system in which 40 000 Swedish base key words are pointed) and it becomes possible. These certain key words are termed as open class words. In the text which is obtained the calculation of frequency of the key words is done by the summarization system of the computer. The sentences of the documents and their location are then detected in the text. The text tagged with bold text tag is considered, first paragraph tag or numerical values are detected. Then the compilation of all this information is done and now this information is utilized to summarize the pristine text document. Due to the convivial blogs and gregarious networking sites, human conversational data in indicted forms are incrementing at a phenomenal rate. Also, organizations and individuals diligent in texting, email exchanges, blogging,, face-to-face meetings and other convivial media activities that are accumulating the data at a high rate. These "accumulated web data" is analysed and mined, with the advancement in natural language processing that provide more preponderant opportunities for engendering numerous incipient and valuable summarization systems.

B. Methods and Techniques of Automatic Text Summarization

1) *Extraction-predicated summarization:* In the literature extraction-predicated summarization are particular of two types that are often addressed. They are key phrase extraction and document summarization [6]. The key phrase extraction cull individual words or phrases to "tag" a document in, and the document summarization, culls whole sentences to engender a concise paragraph summary.

2) *Abstraction-predicated summarization:* Extraction techniques merely copy the information deemed most paramount by the system to the summary (for example, sentences, paragraphs or key clauses), while abstraction involves paraphrasing sections of the source document. In general, abstraction can condense a text more vigorously than extraction, but the programs that can do this are harder to develop as they require the utilization of natural language generation technology, which itself is a growing field[4]. While some work has been done in abstractive summarization (engendering an abstract synopsis like that of a human), the majority of summarization systems are extractive (culling a subset of sentences to place in a summary).

3) *Aided summarization:* Automatic Summarization Systems have been prosperously adopted machine learning techniques from proximately cognate fields such as text mining or information retrieval These are the Plenarily or Fully Automated Summarizers (FAS), but the systems in which the task of summarization availed by the utiliser is MAHS(Machine Availed Human Summarization), it highlights the utiliser passages to be included in the summary, and the systems that depend on post-processing by a human are called HAMS(Human Availed Machine Summarization).

To identifying paramount content for automatic text summarization sundry approaches have been developed till now. Designator representation approaches do not aim at discovering topicality, in this type of approaches the text is represented by a diverse set of possible indicators of consequentiality. These designators are coalesced, by utilizing machine learning techniques, and scoring of the consequentiality of each sentence is done. Finally, by utilizing the avaricious approach the sentences are culled and a summary is engendered culling the sentences that will go in the summary piecemeal or globally optimizing the cull, culling the best set of sentences to compose a summary. On the other hand, in the Topic representation approaches an intermediate representation of the text is derived firstly that captures the topics discussed in the input. Sentences in the input document are scored for important predicated on these representations of topics [3]. In these subsisting approaches a broad overview with the particular distinctions is presented. Particular attention is on how representation, sentence scoring or summary cull strategies alter the overall performance of the summarizer. These are the different techniques and methods for text summarization.

II. RELATED WORK

This paper fixates on the automatic text summarization, its methods, techniques and applications. Therefore in this section overview of some text summarization predicated papers and their reviews are showed. If the utiliser wants to find the germane information in a summarized way then the Automatic Text Summarization techniques have been proved to be efficacious. To amend the efficiency and efficacy of a user’s performance in an information-seeking task he/she requires to only focus at a summary that includes the pertinent information presented in his/her preferred manner. In contrast, it might take more time to solve a target quandary by users or, it might additionally possible to make erroneous decisions, if the main conception is omitted or misrepresented from a summarized document. A personalized text summarization system must be designed to take into account both what a particular utiliser is currently fascinated with and how that utiliser perceives the information. The particular utiliser how receives the information is called as a user’s cognitive styles. Although there are sundry approaches subsists that takes into account a user’s fascinates and avail in the designing of a personalized text summarization system, In the personalised text summarization systems there has been inadequate fixate on cognitive styles and its exploration. When multi document [5] summaries are assessed, the impact of a user’s cognitive styles are studied in this paper presented by ***Hien Nguyen, Eugene Santos*** and ***Jacob Russell*** two dimensions of a user’s cognitive style—the wholist/analytic and imagery/verbal dimensions are culled by them and then a summary engendered from a set of documents is assessed to study the impacts of user’s cognitive styles [2]. Generally, they refer two types of a document set whether the set’s content is proximately cognate or loosely cognate. The reason behind utilizing a document set type is to explore if there are any differences in the summaries of the users’ assessments that they regenerated from sets of variants. The results of this paper show that different assessments have been given by different users in contrast with the information coverage and the way that information is presented in both proximately cognate document sets and loosely cognate document sets. They additionally found that there is significantly distinguishment between the wholist and analytic groups predicated on the coherency ratings that were given to summaries from the two types of document sets. Due to all these studies they investigate the impact of a user’s cognitive styles. Authors found that an utiliser’s ratings on the coherence of the summary of a multi document are affected by a user's cognitive styles and percentage of standalone concepts and graph entropy.

Makbule Gulein Ozsoy, Ilyas Cicekli and ***Ferda Nur Alpaslan*** in their research work present text summarization through LSA (Latent Semantic Analysis) that extract the consequential information from immensely colossal amount of text data [3]. Numerous methods have been developed in the literature that aims to identify paramount content for automatic text summarization and to engender out well-composed summaries. Latent Semantic Analysis (LSA) is one of them. Different LSA predicated summarization algorithms with two incipient LSA predicated summarization algorithms are proposed in this paper. One of algorithms performances the best on comparing utilizing their ROUGH-L scores and these algorithms works on Turkish documents.

Text summarization on the substratum of maximum coverage quandary and its variant are discussed by ***HiroyaTakamura*** and ***Manabu Okumura*** in their research work. Some decoding algorithms are explored, that one additionally never utilized for summarization formulation, including a randomized algorithm, a branch-and bound method and an avaricious algorithm with performance guarantee [1]. The augmentation of summarization model is done on the substratum of the results of comparative experiments. Through experiments, authors showed that the augmented model proves superior to the best-performing method of DUC'04 on ROUGE-1without any cessation word.

HiroyaTakamura and ***Manabu Okumura*** propose generic summarization model which works in a multi-document and predicated on the budgeted median quandary. Sentence cull to engender a summary is efficaciously performed by this proposed model. In the document cluster assignment of every sentence is done and as much as possible each of them represented by a sentence in the summary [10]. This model is salutary in the way so that the document cluster entirely covers the pertinent information of the pristine document with the utilization of sentence assignment and textual entailment (asymmetric cognations between sentences) can incorporate facilely.

Now a day's probing on the Internet is most mundane. For this task the paramount role played by the field of information retrieval. The query processing systems predicated on keywords is efficaciously utilised by most of the information retrieval systems [4]. Words' matching with the lot of scattered data in the cyber world is a tough and consequential task. To retrieve the pertinent natural language is more challenging task in text document. In this paper, the concept of OpenNLP [7] implement is introduced by ***Harshal J. Jain et al*** that is salutary for natural language processing of text for matching of words [8]. Data mining document clustering algorithm are adopted for the extraction of query dependent and consequential information from an immensely colossal set of offline text documents. In addition to, clustering techniques and OpenNLP implements will utilize to analyse the performance of the summary and the best approach will be culled.

In this paper, presented by ***MücahidKutlu, Celal Cigir*** and ***Ilyas Cicekli*** a generic text summarization method is proposed. On the substratum of sentence ranking and their scores summaries of Turkish texts are engendered [4]. Surface level features are utilized for calculation of sentence scores. The sentences are ranked according to their scores and summary is engendered by taking into account highest ranked sentences from the pristine document. To cover the most pertinent information of the pristine text with negligible duplicity some features are taken into account such as centrality, designation kindred attribute, key phrase, sentence position and term frequency. A score function with its weights of the features and feature values computes the sentence rank. Furthermore, performance of the summary by comparing summarization outputs with manual summaries of two incipiently engendered Turkish data sets is analysed and it is well evaluated. The Turkish text summarization features are additionally included. The efficacy of the centrality feature in text summarization is showed and the surface level feature with the utilization of key phrase in text summarization is introduced in this paper, which was one of the first Turkish summarization systems.

III. APPLICATIONS

These summarization programs engender the summary of the pristine text, including query germane summaries or generic machine-engendered summaries. Depending on to which type of summary is focused by any summarization system there are variants of summaries. According to the utiliser needs the summarization systems engender both generic summaries and query germane text summaries [3]. Multimedia documents including images, audios, videos, pictures, movies etc can additionally be reduced or summarized. Summarization of multi-source documents with single-source documents is withal possible. Multiple source documents include a cluster of news stories on the same topic or it is additionally termed as document cluster. There are sundry applications of summarization expounded in this paper.

A. Text Compactor

Text Compactor was engendered by Keith Edyburn for Knowledge by Design, Incorporation. It is predicated on the Open Text Summarizer. Text Compactor is an online summarization implement which provides its accommodations to its utiliser free of cost. Huge amount of data and information is summarized in an opportune way. Struggling readers may take avail from this general approach and dispensed inundating amounts of data. Any diligent student, a professional or a teacher can utilise its facilities.

HOW IT WORKS

The web app of this website calculates the frequency of each word in the document text which is placed on the web page of Text Compactor. Based on the frequency count of the words, a score is computed for each sentence it contains. The sentence which receives most frequency count is considered to be the most paramount sentence. The best results are obtained conspicuously, when a document has only a few sentences.

Reference materials that are non-fiction and text-books provide more preponderant results additionally. Its circumscription is that Text Compactor does not work well with fiction such as stories about places, events and imaginary people).

B. Image Collection Summarization

Another application of automatic text summarization is Image Collection Summarization. Mainly it is utilized for summarization of pictures or images [9]. Some consequential sets of pictures are culled from an immensely colossal pool of pictures. The most representative pictures are showed and termed as image summary.

C. Tools 4 noobs

It also summarizes the text within seconds. Widgets and utilizable scripts are accumulated from astronomically immense amount of web data that anyone can utilize facilely. Tools 4 noobs keep some open source implements and all of its other implements are free of charge. It is gratuitous ***software.***

D. Anchor Text

Anchor text contains a hyperlink that is a readable text. The users reach to the contained link content when they click on the linked text for which they are probing. It contributes to the whole page as an overall probing data and its visibility. Usually it describes the hyperlink that contains the information of that particular word in detail. For example, a page about the “silver mountains” has a link containing the text "silver mountains" is found more facilely by search engines.

E. Open Text Summarizer

An open source implement for summarizing texts is the Open Text Summarizer. The sentences cull is done by the program that reads a text and decides the consequentiality of sentences. The sentences which are more consequential are culled and the sentences which are less paramount are neglected. It works for Fedora, Linux and Ubuntu. At least 25+ languages are fortified by Open Text Summarizer. XML configure these languages. It is exalted and appreciated by sundry academic publications and they have benchmarked it. It is both a command line and a library implement. The command line implement enables us to summarize text on the console while the library implements include word processors such as KWord and AbiWord that can link to the library and reduce the documents. There are two ways of printing the summarized text-simply as text and the other one is as HTML. In the latter one the paramount text are highlighted. It works with UTF-8 encoding system and fortifies multi-lingual. The Open Text Summarizer summarizes texts in German, Esperanto, Spanish, English, Hebrew, Russian and other languages texts are summarized through OTS i.e. Open Text Summarizer. If XML file of rules is edited then it can withal tweak subsisting languages and can fortify many more languages.

F. The Text Tool

The Text Tool is a standard text implement utilized for paint applications. It is an implement that takes up the entire canvas area as the input, and returns the compressed output. The summarized text can be organized with key-board spaces and line breaks that follow the typewriter-style. The text is reduced by taking into account size, colour and weight depends upon the utiliser input.

AUTOMATIC TEXT SUMMARIZATION CAN BE USED:

1. To present summarized information of the search results in search engines such as Google.
2. To search in foreign languages.
3. To obtain an automatically translated summary of the automatically summarized text.
4. To summarize news to WAP-format or SMS or PDA and mobile phones.
5. To present the keyword directed subscription of news which is pushed to the user.
6. To read the summarized text by a computer because original text can be insipid to heedfully aurally perceive and too long.

IV. REVIEW

Table I summarizes the variants of methods utilized in different papers.

The workshop on Multilingual Multi-document Summarization is organized by ***George Giannakopoulos, NCSR “Demokritos” (Greece)*** and ***Georgios Petasis, NCSR “Demokritos” (Greece)*** in ***2013***. In this workshop they focus on summarization of multiple documents with multiple documents [13].

The research work which is done by ***Ning Zhong, Yuefeng Li*** and ***Sheng-Tang Wu*** in ***2012*** .In this research work, the quandary of text mining [11] is solved by an efficacious pattern evolution and revelation technique which works on the low-frequency and misinterpretation quandary for text mining.

The paper published by ***Hien Nguyen, Eugene Santos*** and ***Jacob Russell*** in ***2011*** focuses on multi-document summaries [2]. While assessing the multi-document summary they study the impact of a user's cognitive styles.

Makbule Gulein Ozsoy, Ilyas Cicekli and ***Ferda Nur Alpaslan*** in ***2010*** published an article on 'Text Summarization of Turkish Texts utilizing Latent Semantic Analysis' [3]. In this paper, two incipient LSA predicated summarization algorithms are proposed and different LSA predicated summarization algorithms are expounded withal.

In the paper published by ***Mucahid Kutlu, Celal Cigir*** and ***Ilyas Cicekli*** in ***2010*** to engender the summary of Turkish text a generic text summarization method is proposed .Summaries are engendered by extracting the highest ranked sentences from the pristine documents [4]. Scores are computed to find highest ranked sentences.

The review of different papers is explained with the help of a table in a summarized way.

TABLE I
REVIEW OF VARIOUS PAPERS

S no.	**Year**	**Published By**	**Method Used**	**Advantages**	**Limitations**
01.	2013	**The Association for Computational Linguistics**	Corpus Collection and Generation of Summaries	**1.** Summarizes the multiple languages withal multiple documents.	**1.** There are no translators for different languages. **2.** More attention is needed when translating into Arabic.
02.	2012	**Ning Zhong.Yuefeng Li** and **Sheng-Tang Wu.**	Pattern Taxonomy and Pattern Deploying Method(PTM & PDM)	**1.** Effectively performs term predicated methods, state-of-the-art method including SVM and BM25 and pattern mining predicated methods. **2.** The misinterpretation quandary and low frequency quandary is reduced.	**1.** Lack of accuracy and transparency.
03.	2011	**Hien Nguyen.Eugene Santos** and **Jacob Russell**	Assessment of User's Cognitive Styles	**1.** Find that different users have different coherence ratings and information ratings. **2.** Help to design an utiliser a text summarization system with centred features.	**1.** Does not expand DG's to define the appropriate graph theoretic measures from multiple documents. **2.** There is no trade-off between performance and adaptively.
04.	2010	**Makbule Gulein Ozsoy, Ilyas Cicekli** and **Ferda Nur Alpaslan.**	Latent Semantic Analysis by Cross and Topic Methods	**1.** Cross method observes noise in matrices of LSA and Topic method distinguishes main topics and subtopics. **2.** It is impervious to different input matrix engenderment methods.	**1.**Turkish text is evaluated only, cannot evaluate the English text
05.	2010	**Mucahid Kutlu, Celal Cigir** and **Ilyas Cicekli**	Generic Text Summarization Method using Sentence Extraction	**1.** Introduces the utilization of key phrase as a surface level feature. **2.** Shows the efficacy of the centrality feature in text summarization.	**1.** Does not include phrases conjunctions and answers of 5W1H questions.

V. CONCLUSIONS

In order to engender the summary of a long text there are sundry models which can find the most pertinent content from the data base. These are feature predicated models. But to amend the search methodology there are fewer efforts are found. In this survey different application, methods and techniques of automatic text summarization is expounded proposed which is able to handle the semantic gaps between the pristine text and summarized text. The issues of immensely colossal text summarization systems can be facilely understood. This study work is explicated in following major phase's first exordium of automatic text summarization with its methods and techniques, second some research works cognate with the topic, third its applications. Finally, review of some papers is presented in a tabular form.

ACKNOWLEDGMENT

This work has not been cognate with any type of industrial research work. I would relish expressing my very great appreciation to Prof. Ashish Tiwari for his valuable and constructive suggestions during the orchestrating and development of this research work. His inclination to give his time so benevolently has been very much appreciated. I additionally wish to acknowledge Lect.Mitesh Bargadiya and other to contribute in the preparation of this survey. This work would not be possible without their strenuous exertion.

REFERENCES

[1] HiroyaTakamura and Manabu Okumura, "Text Summarization Model based on Maximum Coverage Problem and its Variant" , Proceedings of the12th Conference of the European Chapter of the ACL, pages 781–789,Athens, Greece, 30 March – 3 April 2009. c 2009Association for Computational

[2] Hien Nguyen, Eugene Santos and Jacob Russell, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 41, NO. 6, NOVEMBER 2011

[3] MakbuleGulcinOzsoy,IlyasCicekli,FerdaNurAlpaslan, "Text Summarization of Turkish Texts using Latent Semantic Analysis", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–876,Beijing, August 2010

[4] MücahidKutlu, CelalCığır and Ilyas Cicekli, "Generic Text Summarization for Turkish", http://web.cs.hacettepe.edu.tr/~ilyas/PDF/2010_COMPUTERJOURNAL.pdf

[5] Marina Litvak, Natalia Vanetik, "Multilingual Multi-Document Summarization with POLY", Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 45–49, Sofia, Bulgaria, August 9 2013. 2013 Association for Computational Linguistics

[6] BertilCarlsson, Arne Jonson, "Using the pyramid method to create gold standards for evaluation of extraction based text summarization techniques", The Third Swedish Language Technology Conference (SLTC 2010) Linkoping October 27-29, 2010

[7] Md. MajharulHaque, SuraiyaPervin, and Zerina Begum Literature "Review of Automatic Single Document Text Summarization Using NLP", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 3 No. 3 July 2013, pp. 857-865 © 2013 Innovative Space of Scientific Research Journals.

[8] Harshal J. Jain, M. S. Bewoor, S. H. Patil, "Context Sensitive Text Summarization Using K Means Clustering Algorithm", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012 .

[9] Jorge E. Cameron and Fabio A. González., "A Multi-class Kernel Alignment Method for Image Collection Summarization", in Proceedings of the 14th Iberoamerican Conference on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP '09), Eduardo Bayro-Corrochano and Jan-OlofEklundh (Eds.). Springer-Verlag, Berlin, Heidelberg, 545-552.

[10] HiroyaTakamura and Manabu Okumura, "Text Summarization Model based on the Budgeted Median Problem", CIKM'09, November 2-6, 2009, Hong Kong, China.

[11] Ning Zhong, Yuefeng Li and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, Volume 24, No. 1,January 2012.

[12] Vishal Gupta and Gurpreet S.Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Volume 1, No. 1, August 2009.

[13] "MultiLing 2013: Multilingual Multi-document Summarization" , Proceedings of the Workshop August 9 2013, Sofia Bulgaria c 2013 The Association for Computational Linguistics(ACL).