Automatic Text Summarization

Roshna Chettri
M.tech, Computer Science and Technology
Sikkim Manipal Institute of Technology

Udit Kr. Chakraborty
Associate Professor,
Computer Science and Technology
Sikkim Manipal Institute of Technology

ABSTRACT
Summarization is the art of abstracting key content from one or more information sources [6]. Summarization includes text summarization, image summarization, and video summarization. Text summarization is one of application of natural language processing and is becoming more popular for information condensation [1].Information is accessible in great quantity for every topic on internet assembly the key information in the form of summary would benefit a number of users. Automatic text summarization system generates a summary, i.e. it contains short length text which comprises all the key information of the document. Summary can be generated through extractive as well as abstractive methods.

Keywords
Extractive, Abstractive, natural language processing

1. INTRODUCTION
Summarization is the way of abstracting important information from one or more sources [6]. It increases the likelihood of finding the points of texts, so the user will spend less time on reading whole documents. Text summarization is one among the typical tasks of text mining [6].The World Wide Web provide a huge information available to users and users are overloaded with lengthy text document where smaller version would do. Some people make decisions on the basis of reviews they have seen and with summaries they can make effective decision in less time. With increasing volume of information summarization play a very important role in terms of time saving.

Text summarization is a difficult task which preferably involves deep natural language processing capacities [5] and in order to simplify the issue current research is focused on extractive summary generation. Summary can be generated through either extractive or Abstractive summarization technique. Sentence based extractive summarization techniques are commonly used in automatic text generation.

Summarization task can be either supervised or unsupervised.in supervised learning training data is needed for selecting main content from the documents. Large amount of annotated or labeled data is needed for learning techniques. These systems are addressed at sentence level as two-class classification problem in which sentences belonging to the summary are termed as positive samples and sentences not present in the summary are named as negative samples [5]. Some of the classification methods used in machine learning is Support Vector Machine (SVM) [5] and neural networks [5].Unsupervised systems do not need any training data. They generate the summary by retrieving only the target documents. Therefore, they are appropriate for newly observed data without any advanced modifications.

1.1 Types of summaries
1.1.1 Extractive summaries (extracts):
This type of summary is generated by selecting few sentence(s) form the document and scores are assigned to important sentences in the documents and then highly scored sentences are chosen to generate the summary. It is performed by concatenating several sentences taken exactly as they appear in the input being summarized. Summary's length depends on the compression rate [5].

Classification of extractive approaches for summary generation

i. *Statistical based approaches*: statistical based approach deals with statistical features like, positive keyword (based on frequency count), negative keyword (based on frequency count), centrality of sentence (i.e. similarity with other sentences), position of sentence ,resemblance of sentence to the title, presence of numerical data in the sentence, relative length of the sentence etc. [5]. This approach is language independent [5]. Here weights of the sentences are identified considering features and based on the weight, score are assigned to the sentences. Sentence with the high score are chosen to generate the summary.

ii. *Topic based approaches*: Topic is defined by topic themes that are represented by events which occur frequently in the collection of documents [4].Topic is represented in five different ways [5]:

Topic signatures.

Enhanced topic signatures.

Thematic signatures.

Modeling the documents' content structure.

Templates.

iii.*Graph based approaches*: In Graph based approaches sentence or word are represented by nodes and edges which connect the related text elements (semantically related) together.in this approach similarity among two sentences is find and if similarity lies above a given limit, then connection between sentence is considered.After the connection is made, random walk on the graph is carried out and important sentences are selected.

iv *Approaches based on machine learning:* Machine learning approach can be supervised, unsupervised or semi-supervised. In supervised approach, there is a collection of documents and their corresponding human-generated summaries such that useful features of sentences can be learnt from them. It is supported by training data categories into "summary data" and "non-summary data". Unsupervised system does not contain training data and it generates the summary by retrieving only the target documents. This type of learning is appropriate for any anew observed data. Semi-supervised learning techniques require both labeled and unlabeled data to produce an appropriate function or classifier.

1.1.2 Abstractive summaries (abstracts)

An abstractive summary does not include the words or phrases from the original document instead it re-interpreted ideas or concepts taken from the original document and shown in a different form. It is written to covey the main information in the input and may reuse phrases or clauses from it, but the summaries are overall expressed in the words of the summary author. It needs extensive natural language processing [5]. Therefore, it is much more complex than extractive summarization [5].

Within and across these two summaries there are two sub categories of summarization

Based on function and target reader

i. **Indicative summary**
This type of summary categorizes the topics of the document and characteristics such as length, writing style, etc. This sort of summary is required for writing an abstract for a less-structured document like an essay, editorial, or book. An indicative abstract is generally made up of three parts [8]:

a. Scope
b. Arguments Used
c. Conclusions

ii **Informative summary**
This summary is for writing an abstract for a strictly-structured document like an experiment, investigation, or survey etc. An informative abstract is made up of four parts [8]:

a. Purpose
b. Methodology
c. Results
d. Conclusions

iii. **Query focused summarization**:
It summarize only the information in the input document that is relevant to a specific user query.

Based on language:
i. **Mono-lingual summarization**:
This type of summarization include input document and the target document be in same language. Example: English to English.

ii.**Multi-lingual summarization**:
When source document is in a number of languages like English, Hindi, Punjabi and summary is also generated in these languages, then it is termed as a multi-lingual summarization system.

iii**. Cross-lingual summarization**:
This type of summary includes source document to be in one language and summary to be generated in some other language.

2. RELATED WORKS

There are lots of researches on Automatic text summarization and various techniques are being developed. Various researchers have proposed new techniques using multiple methodologies for automatic text summarization and some of them are mentioned below:

In 2003, Madhyastha, Harsha V.,et.al [1] proposed a method which makes use of the syntactic structure assigned to the input text by the link parser and its work lies in the working of the rules for prediction of subject, object and their modifiers. In the subject prediction scheme, the linkage of each sentence is considered one by one. If the subject is in some other sentence then it cannot be detected by this scheme.

Same year Johnson, Todd, S. Thede, and A. Vlahov, First Midstates [2] proposed method that different from the link parser. The method uses the mechanism similar to that used by Google search engine for ranking the most important ideas of the document. It is based on syntactic and semantic relationship between words and representation and is used within the program PARE. This method lacks the originality as sentence often appear mangled in the summary due to graph abstraction.

In 2004, Minqing Hu,et.al[3] proposed method which provide a summary of a customer reviews on online product. The method used is Feature-based opinion summarization. This method summarizes reviews in three steps 1. Mining features of the product that has been commented on by customers. 2. In each review identifying opinion sentences and decides either each opinion sentence is positive or negative.3. Result summarization.

Some of the researchers' were working on clustering and extraction method that provides summarization, therefore in 2009, Zhang pei-ying, et.al[4] developed a method which is based on the sentences clustering and extraction. First clustering the sentences in document is performed, and then on each cluster it calculates the accumulative sentence similarity based on the multi-features combination which then chooses the topic sentences by the rules.

Many were combining the techniques for better result, in 2010, Sonia Haiduc et.al[5] proposed a method where two different summarization techniques i.e. lead and VSM have been combined to generate summary of source code and this paper suggested that lead+VSM summaries are a good baseline for the automatic summarization of software artifacts.

In 2012, Surendranadha Reddy, et.al[6] proposed method for summarization of a single document which uses two sentence importance measures i.e. Frequency of the terms in the sentence and similarity to the other sentences. Ranking of sentence is done according to their individual scores and the sentences with top ranks are selected for summary. This method is best suited for fewer grams as with the increase in gram the performance is degraded.

Same year Kirti Bhatia, et.al[7] developed a statistical automatic text summarization approach, which uses K-mixture probabilistic model, to increase the quality of summaries. Sentences are extracted and ranked based on their semantic relationships significance values. Method includes parsing the input into a natural language and major part in the string is searched. From the abstracted symbol parse tree is constructed, and is analyzed based on the frequency of the abstracted symbols and prioritization. All the keywords and symbol is presented in a table and extract the sentence with those keywords and finally result is analyzed.

In 2013, Kamal Sarkar et.al [8] developed a single automatic summarization application in which one sentence that best possible elaborate the concept is selected and the best concept contribute to first line summary and second best line and so on. The proposed technique describes method in two phases. Phase one uses position information and document key phrases in an effective manner for summary sentence selection. Second phase is activated when phase 1 cannot produce summary of the desired length. It combines position

information and TFIDF. The proposed method depends on the extraction of the key content, therefore keywords plays important role in generation of summary.

In 2014, Rashmi Kurmi,et.al[9] developed method to reduce cost and time. The purposed method works on the principal of maximal marginal significance between word and sentence. To decide the maximal marginal significance unit step function is used. This method contains database where useless words or words which can't impact the meaning of document can be stored. The input document is traversed and words contain in the database is eliminated starting from the initial position of the sentence to the end.

In 2015, Luciano Cabral el at ,[10] proposed method for automatic summarization application which allows users to view summaries of news pages on Android-enabled mobile devices. The proposed method contain two approach first approach preprocesses web pages by reformatting or adapting them to a more appropriate way of viewing on small screens, without altering the original content Second approach selects the most salient and relevant content in a given page to the user, meeting their need for quickly grasping the fundamental information.

3. ANALYSIS OF RESEARCH GAP

Sl. No.	Researchers	Description	Research Gap
1	Johnson, Todd et.al 2003	Method used: similar to Google search engine. Sentences are kept in the form of graph that fits well with graph abstraction	Summary often appear mangled
2	Madhyastha et.al, 2003	Rules were defined for prediction of subject, object and their modifiers.	Subject presented in other sentence was not considered.
3.	Sonia Haiduc et. Al, 2010	Method used: Vector Space Model (VSM) and lead Proposed method: combined VSM and lead information retrieval techniques.	Structural information was not considered.
4	Y. surendranadha et. al, 2012	Sentence importance is measured based on frequency of the terms in the sentence and similarity to the other sentences.	Performance need to be enhanced with increase in grams.
5.	Kamal Sarkar et.al, 2013	Method is described in two phases. Phase one use position information and document key phrases in an effective manner for summary sentence selection. Second phase activate when phase 1 cannot produce summary of the desired length. It combines position information and TFIDF.	Improvement is required for key phrase extraction
6.	Luciano Cabral el at, 2015	Two approaches are used. First approach preprocesses web pages. Second approach select the most salient and relevant content in a given page to the user, meeting their need for quickly grasping the fundamental information.	It does not provide the possibility of combining this method to maximize result.

4. CONCLUSION

This survey focuses on different techniques and methodologies used by various researches for automatic text summarization. The aim of this paper is to make researchers aware of some important information related to the past of text summarization and current state-of-the-art. It is seen that extractive summarization is mostly used by researchers for text summarization but in future the features with better results of both abstractive and extractive summarization can be combined together to make better summarization of the text.

5. REFERENCES

[1]. Johnson, Todd, S. Thede, and A. Vlahov. "PARE: An Automatic Text Summarizer." First Midstates Conference for Undergraduate Research in Computer Science and Mathematics. 2003.

[2]. Madhyastha, Harsha V., N. Balakrishnan, and K. R. Ramakrishnan. "Event information extraction using link grammar." Research Issues in Data Engineering: Multi-lingual Information Management, 2003. RIDE-MLIM 2003. Proceedings. 13th International Workshop on. IEEE, 2003.

[3]. Zang pie-ying et al, "Automatic text summarization based on sentence clustering and extraction" IEEE, 2009.

[4]. Udo Hahn et al. "The Challenges of Automatic Summarization " IEEE, 2010.

[5]. Sonia Haiduc et al, "On the Use of Automated Text Summarization Techniques for Summarizing Source Code", 17th Working Conference on Reverse Engineering IEEE,2010.

[6]. Y. Surendranadha Reddy, "An Efficient Approach for Web document summarization by Sentence Ranking", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 7, July 2012.

[7]. Wang, Xuping, et al. "The application of automatic summarization technology in document management." Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on. IEEE, 2013.

[8]. Kamal Sarkar, el at "Automatic Single Document Text Summarization Using Key Concepts in Documents" , J Inf Process Syst, Vol.9, No.4, pp.602-620, December 2013.

[9]. E.Padma Lahari et,"Automatic Text Summarization with Statistical and Linguistic Feature using Successive Thresholds" IEEE(ICACCCT), 2014.

[10]. Luciano Cabral el at,"Automatic Summarization of News Articles for Mobile Devices". Fourteenth Mexican International Conference on Artificial Intelligence,2015.