International Journal of Advance Research in Science and Engineering
Volume No.08, Issue No.06, June 2019
www.ijarse.com
IJARSE
ISSN: 2319-8354

Automatic Text Summarization Methods

Vikas Nirgude[1] ,Vaishali Nirgude[2]

[1]*Sanjivani College of Engineering, Kopargaon Ahmednagar, India*
[2]*Thakur college of Engineering and Technology,Mumbai-400001,India*

Abstract:

Test summarization is a process of extracting large and useful information from original documents and represent it in the summary/ abstract form. Nowadays huge amount of information is available on internet. There are many application where we need summary of the large documents such as news headlines, market review, email summary, short messages on mobile, research work, medical need such as patients' history for further treatment, business analysis, educational field, movie summary, minutes of meeting etc. Human being can give very accurate summary of the original document but it is very tedious job for human being to manually summarize large documents. Therefore to save time and efforts, software approach is used for text summarization. Automatic text summarization broadly classified into extractive and abstractive summarization. This paper mainly focus on comparative study of different extractive and abstractive text summarization techniques with algorithms, advantages and limitations of each technique. Study shows that most of the work has done in Extractive text summarization but still abstractive text summarization is a challenging area due to complexity of Natural Language Processing.

Keywords: LSA, ML, NLP, Ontology ,TF-IDF

I. INTRODUCTION

Nowadays large volume of data is available on the internet. Retrieving large volume of data is not a big problem today but due to time constraints, extracting required information from vast information is really a difficult task. Therefore, text summarization is very important and timely tool for human being to understand the large volume of data in summarize form within a specific time. Human summarization can be person-dependent, context-dependent, varies with human thought so for uniformity automatic text summarization is needed. In automatic text summarization process, extract or collect required information from original documents and represent the most important content to the user in condensed form. When text summarization is done through software it is called Automatic text summarization. In today's data growing age, need of abstract / summary has also increased such as business analysis, market survey, short messages on mobile, medical field, government offices, news headlines, research review, educational field for students and teachers, minutes of meeting etc. Many automatic text summarization

techniques are available to get successful summary, each technique has its own advantages and drawbacks. Initially researchers mainly focused on Single Document Summarization but as the need of automatic text summarization increases due to vast amount of information, researchers are focusing on multi-document summarization. Single document summarization produces summary of single input document. On the other hand, multiple document summarization produces summary of multiple input documents. Automatic text summarization approach include both machine learning and data mining. Automatic text summarization techniques are broadly classified into two categories, extractive summarization and abstractive summarization. Extractive summarization methods extract keywords, phrases, useful sentences etc from the input documents and generate the summary/ abstract. Whereas abstractive summarization methods include deep understanding of input text document and show semantic relation between sentences and then use natural language processing techniques to write new sentences and create a meaningful summary / abstract which is closer to human being.

II. LITERATURE REVIEW

AUTOMATIC TEXT SUMMARIZATION TECHNIQUES :

Automatic Text summarization techniques are discussed in following section:

1. EXTRACTIVE TEXTS UMMARIZATION TECHNIQUES

An extractive summarization method consists of ext racting important sentences, paragraphs, keywords, phrases etc. from

the original document(s) and concate them to produce summary.

1.1 ATTRIBUT ES / FEATUR ES FOR EXTRACTIVE TEXT SUMMARIZATION

One of the usual way to decide the importance of a sentence is to identify features or attributes. Following are few features

/ attributes which can be considered while including the important sentences into the final summary.

Position of Sentence: Usually first and last sentence of a text document are more important and are having greater

chances to be included in summary Score (S) = 1 / Position in the section.

Sentence length: Sentences which are short contain less information and long sentences are not appropriate to represent

summary. So, very large and very short sentences are not included in summary.

Proper Noun: Proper noun is name of a person, place and thing etc. Sentences containing proper nouns are carrying

important information, so can be included in summary.

Title words : The words in the Title and section or subsection heading words are generally included in summary.

Keyword: Sentences having keywords are of greater chances to be included in summary.

Font style: Sentences containing words appearing in upper case, bold, italics or Underlined fonts are usually more

important.

Pronouns: Pronouns cannot be included in summary.

Cue Phrases: Some words or phrases positively or negatively correlated to summary such as 'important', 'to conclude',

'exception' etc. Sentence containing any cue phrases are mostly added into the summary.

Term Frequency: Find most frequently used words in the original documents **.** The word frequency is calculated using TF-

IDF measure.

Presence of Irrelevant words: Some sentences contains non essential words such as 'because', 'due to', 'however' etc

EXTRACTIVE TEXT SUMMARIZATION STEPS :

Following are the steps to extract text from original documents and generate meaningful summary.

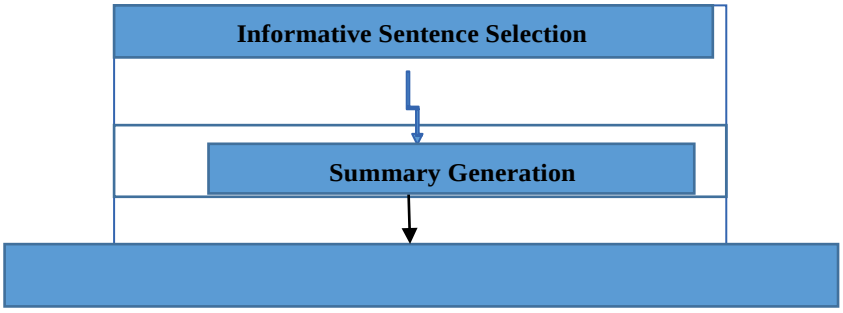

Fig. 1 Text Summarization Steps

a. Input Document: Source Text (Input):

- Source: single-document vs. multi-document
- Language: monolingual vs. multilingual
- Category: news vs. technical paper
- Specificity: domain-specific vs. general
- Length: short vs. long
- Media: text, graphics, audio, video, multi-media

b. Preprocessing:

Preprocessing is structured representation of the original document which includes:

- Sentence boundary identification: Sentence boundary is identified with presence of dot at the end of sentence.
- Stop word elimination: Common words with no semantic relation.
- Stemming : A word is reduced to common form.
- Tokenization: Source text divided into different tokens.

c. Feature Extraction: Identify features/ attributes such as keywords, location of sentences, title, phrases etc to find informative sentences.

d. Informative Sentence Selection: Select information rich senetences for summary using automatic te xt summarization techniques.

e. Summary Generation: Generated summary should be precise, information rich, semantically related , less redundant etc.

EXTRACTIVE TEXT SUMMARIZARION TECHNIQUES:

Most of the work has been done on extractive summarization. Extractive text summarization create the summary from phrases or sentences in the source documents. Information rich sentences are selected from original documents to form abstract / summary by using different ext ractive text summarization techniques.

1. Term Frequency-Inverse Document Frequency (TF- IDF)
2. Cluster Based Method
3. Graph Theoretic Approach
4. Machine Learning Approach
5. LSA Method
6. Text summarization With Neural Networks

7. Automatic Text Summarization based on fuzzy logic
8. Query Based Extractive Text Summarization

LIMITATION OF EXTRACTIVE TEXT SUMMARIZATION:

1. Extractive text summarization sometimes takes more time to form summary than average time.
2. Sometimes generate redundant summary.
3. Sometimes in final summary, sentences are not logically related to each other i.e . sematic relation is missing.
4. Multi-lingual text summarization is difficult.
5. Selected sentences for summary generally longer, so irrelevant part of sentences also get added in the summary.
6. In multi document summarization, there is possibility of contradiction between sources.

EVALUATION OF AUTOMATIC TEXT SUMMARIZATION:

There is no straightforward approach to evaluate summaries or automatic text summarization methods. There are two summary

evaluation methods: Intrinsic and Extrinsic. While evaluating summaries, following are two properties of the summary that

must be measured.

The Compression Ratio (how much shorter the summary is than the original):

CR= Length of summary / Length of input document

Retention Ratio (how much information is retained): RR= Information in summary/ information in input document

Table 1 Shows comparison of above Extractive Text summarization methods based on the different parameters such as algorithm, features, advantages and limitations etc.

Table 1. Comparison between Extractive Text Summarization Techniques

S.No	Method	Author	Explanation/ Features/Algorithm	Advantage	Disadvantage
1	Term Frequency-Inverse Document Frequency [10]	M.Fachrurrozi, Novi Yusliani, and Rizky Utami Yoanita, (2013)	Sentence / word / term -frequency is the number of sentences / words in the document that contain that term. Select those sentences which are similar to the query and the highest scoring sentences are picked to be part of the summary.	Query based sentences. The highest word frequency sentences are selected for summary generation	Redundancy in summary
2	Cluster Based Method [11]	Anjali R. Deshpande, Lobo L. M. R. (2013)	Similar documents are 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 Use of cosine similarity algorithm.	Clustering can be used to group similar sentences in different topics and generate a meaningful summary. Less repetition in summary.	Multidocument summarization
3	Graph Theoretic Approach [12]	Rada Mihalcea, Niraj Kumar, Kannan Srinathan and Vasudeva Varma (2013)	Every sentence is a node and there is edge between two sentences if they share common words. The nodes with high cardinality are the important sentences in the partition, and hence included in the summary.	Adjusted easily for visualization of inter and intra document similarity.	Less Semantics

4	Machine Learning Approach [13]	Kamal Sarkar, Mita Nasipuri,, Suranjan Ghose(2011)	The training dataset is used for reference. Sentences are classified as summary and non-summary sentences based on the features that they possess.	It provides a universal summary.	Machine learning techniques are computationally complex.
5	LSA ethod [14]	Hanane Froud, Abdelmonaime Lachkar and Said Alaoui Ouatik (2013)	Latent Semantic Analysis (LSA) in text processing. It gets this name LSA because SVD (Singular Value Decomposition (SVD) is a very powerful mathematical tool applied to document word matrices, groups documents that are semantically related to each other.	Automatically extract Semantically related sentences even though common words are not present , just like human brain	Complex calculations needed.

2. ABSTRACTIVE TEXT SUMMARIZATION TECHNIQUES:

The abstractive summarization objectives to produce a generalized summary which is crisp, intelligent, information rich and semantically related. The abstractive summarization usually requires advanced language generation and compression techniques. Initial work only on single document summarization. Due to large amount of information on web , multi document summarization arose. Multi document summarization produces summaries from many source documents on the same topic.

Abstractive text summarization techniques are broadly classified into two categories [18] :

A. Structured based approach

B. Semantic based approach

A. Structured Based Approach:

Structured based approach improves the quality of summaries. All most all approaches produces abstract, relevant, information rich and less redundant summary. Only lead and body phrase method produces summary with redundant sentences. These methods most important in formation from the original documents through cognitive schemes such as templates, e xt raction etc. Five techniques under structured Based Approach are as follows:

1. Tree Based Method
2. Template Based Method
3. Ontology Based Method
4. Lead and Body Phrase method
5. Rule Based Method

B. Semantic Based Approach

Semantic based approach based on semantic representation of original documents. These methods produce concise, information rich, coherent, and less redundant summary. Three techniques under Semantic Based Approach are as follows:

1. Multimodal semantic model
2. Information Item Based Method
3. Semantic Graph Based method

Table 2 Compares all the Abstractive Text summarization methods based on the different parameters such as algorithm / features , advantages and drawbacks etc.

C. LIMITATION OF ABSTRACTIVE TEXT SUMMARIZATION:

1. Abstractive text summarization is difficult to implement compare to Extractive text summarization.
2. The quality of summaries are varying from system to systemor person toperson.
3. There is no generalized structure / template that humans can use for abstractive summarization.
4. Evaluating an abstractive summary is not a straightforward because there does not e xist an ideal summary for a given document.
5. Sometimes summary include grammatically incorrect sentences due to parsing errors.

III. DISCUSSION

Automatic Text summarization broadly classified into Extractive and abstractive. Extract ive summaries are created by reusing portions such as words, sentences etc of the input text. Search engines (Information Retrieval System) typically generate extractive summaries from WebPages. Most of the summarization work has been done today on extractive summarization. In abstractive summarization, information from the source text is re-phrased and deep analysis of input text is done to generate semantically meaningful summary. Human beings generally write abstractive summaries. Still Abstractive summarization is a challenging area because of the semantic representation, inference and complexity of natural language processing. In future, more focus should be done in the following direction in the field of automatic te xt summarization:

1. Text summarization for Indian languages such as Bengali, Telugu, Hindi, Marathi, Tamil etc
2. Multi-lingual text summarization.
3. More work should be done in multimedia summarization.

Table 2. Comparison between Abstractive Text Summarization Techniques

Sr. No	Name of the Method	Author	Explanation/ Features/Algorithm	Advantage	Disadvantage
1	Tree Based Method [1]	R. Barzilay and K. R. McKeown (1999,2005)	Use of dependency tree to represent the text/contents of a document. Theme intersection algorithm is used for content selection for summary. Finally sentence generation phase uses FUF/SURGE language generator.	Use of language generator significantly improved the quality of resultant summaries by reducing repetition	Context
2	Template Based Method [3]	S. M. Harabagiu and F. Lacatusu (2002)	1. Template is used to represent document Multi-document Summarization	This technique generate accurate summary because it depends on relevant information identified by Information Retrieval system.	This technique is works only when summary sentences are already present in the origi nal documents.
3	Ontology Based Method [4]	Lee and Jian (2005)	Use of fuzzy ontology with fuzzy concepts for text summarization. Chinese News summarization is done by news agent based on fuzzy ontology.	Handle Uncertain data	This technique is applicable only for Chinese News and not for English News.

4	Lead and Body Phrase method [5]	Tanaka and Kinoshita (2009)	Insertion and substitution operations on phrases that have same syntactic head chunk in the lead and body sentences in order to rewrite the lead sentence.	It found semantically appropriate revisions for revising a lead sentence.	Due to parsing errors, reduces completeness sentences e.g. grammatically incorrect
5	Rule Based Method [6]	Genest and Lapalme (2012)	The documents to be summarized are represented in terms of categories and a list of aspects. Generation patterns designed for each abstraction scheme to	Generate a meaningful summary	All the generation patterns are manually written which is time consuming and tedious.
6	Multimodal sematic model [7]	C. F. Greenbacker (2011)	This technique focus on concepts and relationship among concepts. Multimodal document contains both images and text.	It generates abstract summary which contains all possible important information because it includes both textual and graphical content from the entire document.	It is manually evaluated by Human being.

IV. CONCLUSION

Text Summarization is one of the inspiring field of research because it has many applications where we need summary of the large documents such as news headlines, market review, short messages on mobile, medical, business analysis etc. This review paper focused on both the comparative study of extractive and abstractive text summarization techniques along with their advantages and limitations. An extractive summary is selection of important sentences from the original documents based on different attributes or features of sentences. Abstractive summary methods produces highly relevant, crisp, information rich and less redundant summary. A lot of work has been done in Extractive text summarization but Abstractive text summarization is a still challenging area because of the complexity of natural language processing.

V. REFERENCES

[1] R. Barzilay, *et al.*, "Information fusion in the context of multi-document summarization," in *Proceedings of the 37th annual meeting of the Association for Computational inguistics on Computational Linguistics*, 1999, pp. 550-557.

[2] R. Barzilay and K. R. McKeown,"Sentence fusion for multidocument news summarization," *Computational Linguistics*, vol. 31, pp. 297-328, 2005.

[3] S. M. Harabagiu and F. Lacatusu, "Generating single and multi-document summaries with gistexter," in *Document*

[4] C.-S. Lee, *et al.*, "A fuzzy ontology and its application to news summarization," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 35, pp. 859-880, 005.
Understanding Conferences, 2002

[5] H. Tanaka , *et al.*, "Synta x-driven sentence revision for broadcast news summarization," in *Proceedings of the*
2009 Workshop on Language Generation and Summarisation, 2009, pp. 39-47.

[6] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012, pp. 354-358.

[7] C. F. Greenbacker, "Towards a framework for abstractive summarization of multimodal documents," *ACL HLT 2011*, p. 75, 2011.

[8] P.E. Genest and G. Lapalme, "Framework for abstractive summarization using te xt-to-te xt generation," in Proceedings of the Workshop on Monolingual Te xt-To-Te xt Generation, 2011, pp. 64-73.

[9] I. F. Moawad and M. Are f, "Semantic graph reduction approach for abstractive Text Summarization," in Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, 2012, pp. 132-138.

[10] Fachrurrozi M., Yusliani Novi, and Yoanita Rizky Utami, "Frequent Term based Te xt Summarization for Bahasa Indonesia", International Conference on Innovations in Engineering and Technology Bangkok (Thailand), 2013.

[11] Deshpande Anjali R., Lobo L. M. R. J., "Te xt Summarization using Clustering Technique", International Journal of Engineering Trends and Technology (IJETT) , 2013, Vol. 4 Issue8.

[12] Kumar Niraj, Srinathan Kannan and Varma Vasudeva, "A Knowledge Induced Graph-Theoretica l Model for Extract and Abstract Single Document Summarization", Computational Linguistics and Intelligent Te xt Processing - 14th International Conference, 2013.

[13] Sarkar Kamal, Nasipuri Mita, Ghose Suranjan, "Using Machine Learning for Medical Document Summarization", International Journal of

Database Theory and Application, 2011.

[14] Froud Hanane, Lachkar Abdelmonaime and Ouatik Said Alaoui, "Arabic Te xt Summarization Based On Latent Semantic Analysis To Enhance Arabic Documents Clustering", International Journal of Data Mining & Knowledge Management Process (IJDKP), 2013, Vol.3, No.1. [15] Khosrow Kaikhan, "Te xt Summarization Using Neural Networks", Proceedings. 2004 Second IEEE International Conference on Intelligent Systems, 2004, Vol. 1.

[16] Ladda Suanmali, Salim Naomie, and Mohammed Salem Binwahlan, " Fuzzy Logic Based Method for Improving Te xt summarization" IJCSIS, 2009.

[17] Ibrahim Imam, Nihal Nounou, Alaa Hamouda, Hebat Allah Abdul Khalek, "Query Based Arabic Te xt Summarization", IJCST, 2013, Vol. 4, Issue Spl - 2.

[18] Atif Khan, Naomie Salim," A Review On Abstractive Summarization Methods", Journal of Theoretical and Applied Information Technology,2014, Vol. 59 No.1.

[19] Deepali K. Gaikwad1 and C. Namrata Mahender2,"

A Review Paper on Te xt Summarization", IJARCCE, Vo l. 5, Issue 3, March 2016.