International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Survey Paper on Extractive and Abstractive Techniques in Automatic Text Summarization

Vinit Agham[a*], Dr.V.K.Shandilya[b]

[a]*Research Scholar,Department of Computer Sci. & Engg.,Sipna College of Engineering,Amravati 444604,India*
[b]*Head of the Department,Department of Computer Sci. & Engg , Sipna College of Engineering, Amravati 444604,India*

A B S T R A C T

The technique of Automatic summarization is to reduce size of the text document with a computer algorithm or program with the intention of generation of summary that preserves the most significant points of the original document. As the amount of data has greater than before, so has interest in automatic summarization. It is very much difficult for human beings to do summarization of large documents text manually. Text Automatic Summarization methods can be broadly categorized into extractive and abstractive summarization. Methods with Extractive approach proceed by choosing a subset of existing words, phrases, or sentences in the original text to form the summary. Abstractive Text summarization fully understands the contents of document and generates a new document which is smaller than original text keeping the meaning of the original text unchanged. The abstractive summarization approach requires natural language generation techniques. An excessive research has been carried out into extraction-based algorithms, but very few works exist in the context of abstraction-based summarization. This paper gives the comparison of various text summarization models and also discusses the types of summarization based on categories and different approaches of abstractive as well as extractive text summarization.

Keywords:Abstractive Summarization, Automatic Text Summarization, Extractive Summarization, Neural Network Based Summarization

1. Introduction

Nowadays, people search their queries, retrieves information using search engines or any other Information Retrieval (IR) tool. However, with the rapid growth of information on the internet, information abstraction or summary of the retrieved results has become necessary for users. In the day to day life we find the notions of information overload, text summarization has become an important and timely tool for user to quickly understand the large volume of information. Now the question may arise that why we are interested to summarize the text? There are several valid reasons in favor of the automatic summarization. Here are just a few,

- Summaries reduce reading time.
- While doing research, researchers has to interpret the various scientific papers, summaries make the selection process easier.
- The algorithms in Automatic Text Summarization are less biased than human summarizers.

The literature provides various definitions of text summarization. Radev and et al. introduced the concept of multi-document summarization and the length of the summary in their definition (Radev et. al. ,2002):

A summary is a text that is produced from one or more texts, that convey important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that.

Automatic Text Summarization process represents the original information is in the shortened and conserved form. This representation not only save processing time, but also save storage space. Automatic Text Summarization is the method of automatically generating summaries from an input document while retaining the important points. There are two types of summarization i.e. Extractive and Abstractive. The approach of Extractive summarization systems form summaries by selecting parts of the source text through some measure of importance and then combine those part/sentences together to render a summary. Significance and interpretation of sentence is based on linguistic and statistical features. The Abstractive summarization

* *Corresponding author*
E-mail address: vinitagham@gmail.com

systems generate new phrases, possibly rephrasing or using words that were not in the original text. Abstractive approaches are harder. For achieving the faultless and perfect abstractive summary, main thing is that the model has to correctly comprehend the document and then express that understanding in short feasibly using new words and phrases. Majority of the research work has conventionally focused towards extractive due to the easiness of defining hard-coded rules for selecting important sentences than generating new ones. Also, it promises grammatically correct and coherent summary. Very less investigation has been done in abstractive summarization.

2. Types of Summarization Techniques

There are so many approaches on which we can categories the automatic text summarization methods. Different types text summarization techniques can be categorized based on below types.

2.1. Based on approaches

- Extractive methods select sections from the original text documents those sections can be phrases, sentences, words etc. and join or collect them to generate a summary without changing the original text.
- Abstractive method generates new words or phrases which are not in the source text for creating summary.

2.2. Based on details

- Indicative summary is used for quick view of a lengthy document and it provides only the main idea of source text that encourages a user to read the document.
- Informative summary serve as a substitution to the original document. It draws the brief information of the input document to the user.

2.3. Based on content

- In Generic summarization the generated summary is general in sense which can be used by any type of the user.
- In Query-based summarization Question-Answer system is there where the summary is generated as per user's query.

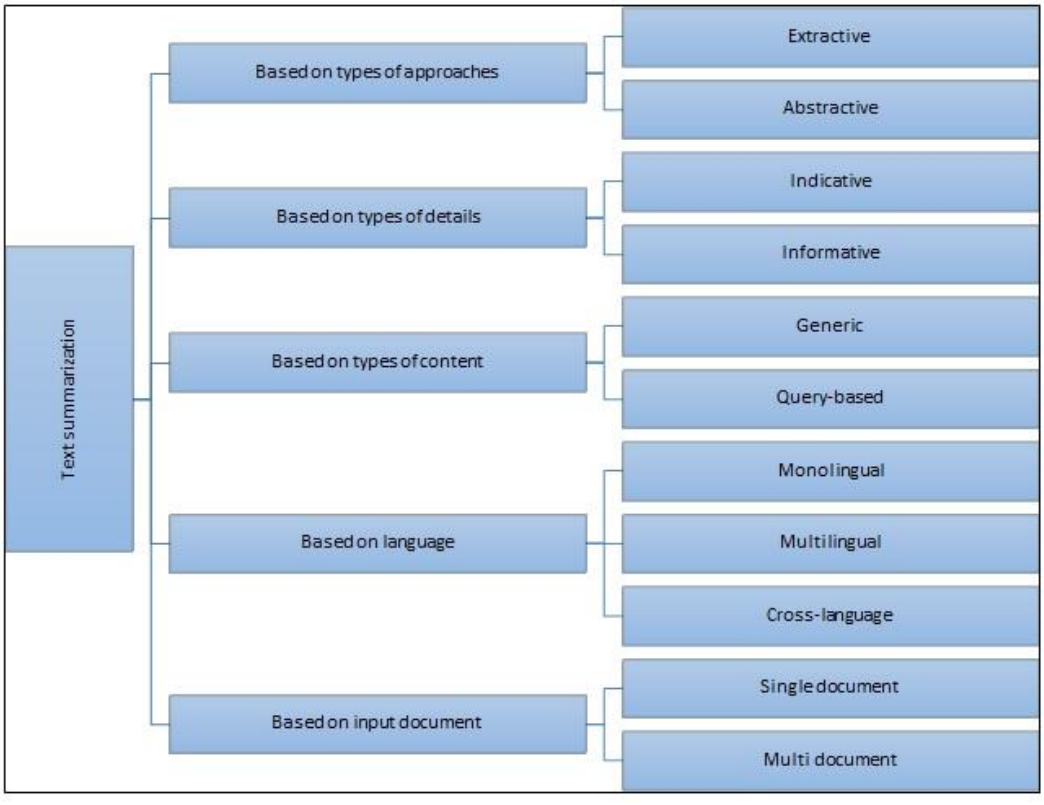

Fig. 1 - Different types of text summarization techniques

2.4. Based on language

- In Monolingual summarization input text language and output summary language is same but it designed to be performed on any single natural language.
- In Multilingual summarization input text language and output summary language is same but it designed to be performed on multiple natural languages.
- In Cross-lingual summarization the input text language and output summary language is different. For example: summarization of Marathi news to English.

2.5. Based on input document

- In Single document summarization there is only one input document to the summarizer.
- Multi-document summarization accepts more than one document as an input.

3. Difference between Extractive and Abstractive Text Summarization

Among various types of summarization most of the research is tend towards extractive and abstractive summarization. Extractive summarization is nothing but extracting or highlighting important parts of the text which is mostly enough to represent the original text. Abstractive text summarization summarizes which is similar how human write the summary and create more shortened summaries. These techniques i.e. abstractive are much tougher to implement than extractive summarization techniques in general. Following table shows the key differences between abstractive and extractive text summarization approaches.

Table 1 - Difference between Extractive approach & Abstractive approach of Text Summarization

Extractive Summarization	Abstractive Summarization
Select most important sentences to produce summary	Understand the whole document to produce summary
Produces a summary may not be grammatically correct.	Produces a summary which is grammatically correct.
Requires statistical, linguistic and heuristics procedures.	Requires NLG (Natural Language Generation) based procedures.

4. Extractive Summarization Methods

An extractive summarization method involves selecting important words or sentences or phrases or paragraphs etc. from the source document and represent them into proper sequence. The selection of entity (i.e. it may be word, phrase, sentence etc.) is based on how much it is important to the summary. This importance can be calculated using the features of the entity. Some of the parameters or features for sentence selection are:

Table 2- Extractive summarization methods.

Parameters/Features	Description
Word frequency	Word frequency of a word is defined as the ratio of the number of occurrence of each word in the entire text over document length.
Position of the Sentence	The position score of a sentence can be created as: the first sentence in a heading has a score value of 5 out of 5, the second sentence has a score 4 out of 5 and so on.
Title Similarity	It is the word overlap between the sentence and the document title. It is calculated as the ratio of number of overlapping title-sentence words and number of words in the title.
Named entities	Proper names such as names of celebrities, companies, groups of people and so forth are of high importance, especially if they are part of news texts. Therefore, in this section, the number of proper names for each sentence was calculated.
Term Frequency-Inverse Document Frequency (TF-IDF)	In TF-IDF, each word is given a weight based on its frequency in a document. This frequency shows the importance a word is in a document. While TF refers to word frequency in a document, IDF is used to calculate the final weight. The Inverse Document Frequency (IDF) was calculated by dividing the total number of documents by the number of documents in which the desired word appeared.

	Finally, by multiplying the two components of TF and IDF, the feature in question was obtained based on the weight of words.
Positive keywords	Positive keywords in the sentence are the keywords come many times in the summary.
Negative keywords	Negative keywords are the keywords that are unlikely to occur in the summary.
Sentence Centrality	Sentence centrality is the vocabulary overlap between this sentence and other sentences in the document.
Numerical data	The sentences having numerical information are most important one and it is most probably included in the document summary.
Presence of Brackets	After doing investigation it has been found that brackets do not contain important information and has lower probability to be included for the summary.
Presence of inverted Commas	Most often the information or text within inverted comma is important related to summary so such information or text has higher probability to be included for the summary.
Sentence Length	Sentences which are shorter in length may not show theme of a text document because of less number of words enclosed in it, though selecting longer length sentences are also not good for summary. So sentence length parameters are calculated in such a way that, shorter and longer sentences are assigned lower values. It is applied to avoid the selection of too short and too long sentences.

Figure 2 shows the general structure of automatic text summarization (Extractive) system. There are so many extractive text summarization developed till date. Here we try to cover some distinguished approaches towards extractive text summarization.

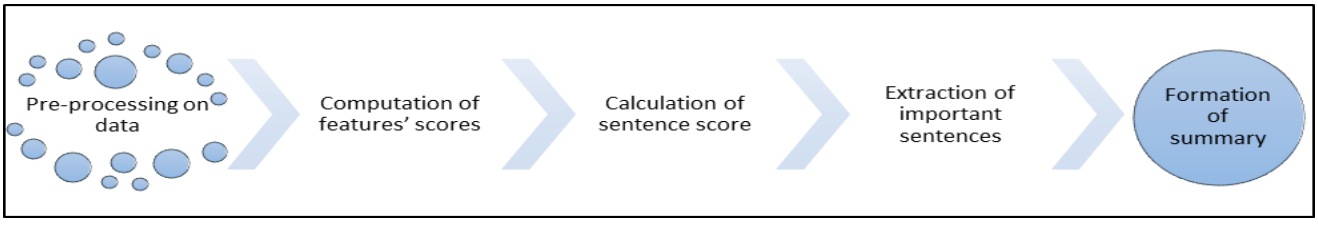

Fig. 2 - General Structure of Automatic Text Summarization (Extractive) system

4.1. CNN Based Extractive Text Summarization

Charitha and et al. developed the CNN based model which was capable of learning sentence features on its own. Feature extraction can be manual in most of the extractive systems but this system automatically extract features thus reduces overhead of extracting. Integer linear programming (ILP) is used to generate the summary based on sentence ranks. ILP is also helps to decrease the redundancy in the generated summary. This CNN model was trained so that it learns features of the sentences to rank them (Charitha, S et al., 2018).

4.2. Extractive Text Summarization based on RBM and Fuzzy Logic

N. S. Shirwandkar and S. Kulkarni proposed a method for Extractive text summarization that uses a combination of Restricted Boltzmann Machine and Fuzzy Logic to select important sentences from the text still keeping the summary meaningful and lossless. Two different summaries are generated using Restricted Boltzmann Machine as well as Fuzzy logic. Both summaries are then combined and get the final summary (Shirwandkar, N. S., &Kulkarni, S., 2018).

4.3. Heading-Wise Text Summarizer

P. Krishnaveni and S. R. Balasundaram introduced the model that summarizes the given input document using local scoring and local ranking. It provides heading wise summary. It makes ranking of the sentences heading wise and selects top n sentences from each heading. The ultimate summary formed by this method is a collection of summary of individual headings (P. Krishnaveni and S. R. Balasundaram, 2017).

4.4 Multi-document Text Summarizer

Rezaei and et al. introduced two multi-document extractive text Summarization systems in their model. This method uses auto-encoder neural network and deep belief network separately for scoring sentences in a document to compare their performances. Author also added some new features to score the sentences. Deep Neural Networks can improve the results by generating new features (Rezaei A et al., 2019).

5. Abstractive Summarization Methods

Abstractive Text summarization can be implemented using two approaches:

- Traditional or Generic Approach
- Neural Network Based Approach

5.1. Traditional approach for Abstractive Text Summarization

The methods or systems under this traditional or generic approach are compatible with rules. That is to implement this system first we have to design algorithm which follows some special logic or rule. All generic summarization systems tend to be unique in nature.

Structured Based Approach

Structure based approach takes the most important information through cognitive theories. It populates important sentences in a predefined structure without losing its meaning.

Table 3 - Structured Based Approaches and their significance

Structured Based Approach	Significance	Authors
Tree based method	This technique is based on a dependency tree for representing the text/contents of a document. The technique performs language generation.	R. Barzilay and K. R. Mckeown proposed a sentence fusion technique that uses treebased method (Barzilay&Mckeown, K. R., 2005).
Template based method	This technique uses a template to represent a full document. Linguistic patterns or extraction rules are compared to recognize text snippets that will be mapped into template slots.	S. H. Finley and S. M. Harabagiu proposed both single and multi-document summarization that uses a template based approach (Finley S. H. &Harabagiu S. M., 2002).
Ontology based method	The ontology i.e knowledge base can be used to boost the method of summarization. Maximum documents or information are domain connected on internet.	Lee and et al. proposed Chinese News Summarization based on fuzzy ontology. They built a model that process uncertain information and precisely define the domain knowledge (Lee et al., 2005).
Rule based method	In this technique, the input document is translated into classes and aspects. Data extraction rules causes system to response to the listed aspects. Content choice module is used to select the most effective candidate while generation patterns are used for making new sentences.	P. E. Genest and G. Lapalme used Rule based method in their abstractive summarization. They applied rules for extractions on the semantically related nouns and verbs (Genest, P. E. & Lapalme, G. , 2012)

Semantic based approach

It works on semantic representation of the document where this semantic information is feed to Natural Language Generation module to get the resulting desired summary. In this procedure, linguistics form of document(s) is generated to feed into natural language generation (NLG) system. This technique can identify noun phrases and verb phrases by processing linguistic data. Different methods using this approach are discussed here.

Table 4 - Semantic Based Approaches and their significance

Semantic Based Approach	Significance	Authors
Multimodal semantic model	Context about is withdrawn by image captioning or any other procedure. The context is summarized. Text is also summarized. Next combine these two summaries to form final summary.	Greenbacker proposed a structure which generates summary that is abstractive in nature. Multimodal document contains both text and images (Greenbacker, C. F. , 2011)
Information item based method	Abstract representation of input document is created then Information is formed from this representation. The abstract representation plays the important role in summarization process instead of sentences in input document.	P. E. Genest and G. Lapalme established Information-Item based structure. This structure is consists of brief entities of coherent information in a text (Genest P. E. & Lapalme G., 2011).

Semantic Graph Model	Here linguistics graph is drawn by processing input document. This graph is also known as rich semantic graph (RSG).	Moawad and Aref prepared summary by creating Rich Semantic Graph. Then it decreases the generated semantic graph to get final abstractive summary (Moawad I. &Aref, M. , 2012)

5.2. Neural Network Based Abstractive Text Summarization Models

Neural network approaches do not follow manually compiled features and they are not bound to specific list of rules. The abstractive Text Summarization took next move when emergence of deep neural networks takes high acceptance. Most of the neural abstractive summarization system makes use of encoder-decoder architecture or sequence-To-Sequence architecture. The encoder captures the input or source data in sequence from which the decoder generates the target summary.

5.2.1 Attention-Based Summarization (ABS)

M. Rush and et al. recommended Attention-Based Summarization (ABS) system. This structure was based on Encoder-Decoder model. Three different encoders were used in this work like Bag-of-Words Encoder, Convolutional Encoder, and Attention-Based Encoder. Bag-of words encoder did not support word sequence in output. The decoder here is nothing but a language model based on Feed-Forward Neural Network (Neural Network Language Model).This model estimates the probability distribution that generates the word at each time step (Rush A. M. et al., 2005)

5.2.2Recurrent Attentive Summarizer (RAS)

Sumit Chopra and et. al made use of RNN and attention mechanism both in their system. Thus they called it as Recurrent Attentive Summarizer model. The model uses the encoder similar to the ABS system, but the weights were assigning in different way. The RAS summarizer used two decoder models which are based on the RNN and the LSTM (Chopra, S. et. al, 2016).

5.2.3 Sequence to Sequence Attentional Model (Seq2Seq)

A. See and et al. developed a model which was adopted from Nallapti's (Nallapati, R et .al , 2016)model (Sequence to Sequence RNN model).Authors used Nallapati's model as a baseline model. Context vector was created from input by the encoder. Here, encoder implemented using a single layer bidirectional LSTM while decoder implemented using a single-layer unidirectional LSTM (See, A. et al., 2017).

5.2.4 Multi-Modal Text-Image Summarization

Another concept based on Encoder-Decoder is Text-Image Summarization. J. Chen and H. Zhuge proposed multi-media summarization i.e. image plus text combine. They developed text-image summarization model based on abstractive approach .They call them as Abstractive text-image summarization. It was developed by using the attentional hierarchical Encoder Decoder model. This model summarizes text as well as its associated images simultaneously and then it makes alignment between the sentences and their associated images in summaries (Chen, J., &Zhuge, H., 2018).

5.2.5 Hybrid Model

Y. Zhang and et. al combined both the approaches i.e. extractive and abstractive. In the extractive part, they constructed a graph model and proposed sentence similarity measure. Then using this measure for ranking and extracting key sentences the model concatenates the important sentences into a smaller text as the input of the summary generator (Zhang, Y. et. al, 2018).

4. CONCLUSION

Various methods of Automatic Text Summarization (ATS) are discussed in this review. After implementing various summarization techniques authors concluded that abstractive summarization methods produce highly consistent, coherent and less redundant summary. Although the abstractive approach of summarization requires substantial computational models for summary generation. The purpose of this investigation is to provide broad survey of text summarization. This work also reveals the comparison of different techniques and approaches of abstractive summarization as well as extractive summarization. Positively, this study has been reformed in a way that new researchers to the area of text summarization can get a better understanding on different text summarization approaches.

REFERENCES

Juan, M.T. (2014). Automatic Text Summarization. Cognitive Science and Knowledge Management Series: Wiley Publications.

Radev, D., Winkel , A. & Topper M. (July 2002). Multi document centroid-based text summarization. In 40th Meeting of the Association for Computational Linguistics (ACL '02), Demonstrations Session, Philadelphie, PA, ACL, 112–113.

Dilawari, A. & Khan M. U. G. (2019).ASoVS: Abstractive Summarization of Video Sequences. In IEEE Access ,7, 29253-29263.

Charitha, S., Chittaragi, N. B., &Koolagudi, S. G. (2018). Extractive Document Summarization Using a Supervised Learning Approach. IEEE Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER 2018). doi:10.1109/discover.2018.8674133.

Shirwandkar, N. S., &Kulkarni, S. (2018). Extractive Text Summarization Using Deep Learning.Fourth International Conference on Computing Communication Control and Automation (ICCUBEA-2018). doi:10.1109/iccubea.2018.8697465

Krishnaveni, P., &Balasundaram, S. R. (2017).Automatic text summarization by local scoring and ranking for improving coherence.International Conference on Computing Methodologies and Communication 2017. doi:10.1109/iccmc.2017.8282539

Rezaei A.,Dami S. and Daneshjoo P. (2019). Multi-Document Extractive Text Summarization via Deep Learning Approach. Conference on Knowledge Based Engineering and Innovation (KBEI), Tehran, Iran, 680-685. doi: 10.1109/KBEI.2019.8735084.

Afsharizadeh, M., Ebrahimpour-Komleh, H., &Bagheri, A. (2018).Query-oriented text summarization using sentence extraction technique.International Conference on Web Research (ICWR 2018). doi:10.1109/icwr.2018.8387248.

Barzilay, R., &Mckeown, K. R. (2005).Sentence Fusion for Multidocument News Summarization. Computational Linguistics, 31(3)

Finley S. H. &Harabagiu S. M. (2002). Generating single and multidocument summaries with gistexter.In U. Hahn & D. Harman (Eds.), Proceedings of the workshop on automatic summarization, 30–38.

Lee C. S., Jian Z.-W. & Huang L. K. (2005).A fuzzy ontology and its application to news summarization. In Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 35, 5,859–880.

Tanaka, H., Kinoshita, A., Kobayakawa, T., Kumano, T., & Kato, N. (2009).Syntax-driven sentence revision for broadcast news summarization.Proceedings of the 2009 Workshop on Language Generation and Summarisation - UCNLG Sum 09. doi:10.3115/1708155.1708163.

Genest, P. E. & Lapalme, G. (2012).Fully abstractive approach to guided summarization.Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA. 2. 354–358.

Greenbacker, C. F. (2011).Towards a framework for abstractive summarization of multimodal documents.ACL HLT 2011, 75.

Genest, P. E. & Lapalme, G. (2011).Framework for abstractive summarization using text-to-text generation.Proceedings of the Workshop on Monolingual Text-To-Text Generation, Association for Computational Linguistics, Stroudsburg, PA, USA.64–73.

Moawad, I. &Aref, M. (2012).Semantic graph reduction approach for abstractive text summarization.Seventh International Conference on Computer Engineering Systems (ICCES), 132– 138.

Rush, A. M., Chopra, S., & Weston, J. (2015).A neural attention model for abstractive sentence summarization.arXiv:1509.000685v2.

Chopra, S., Auli, M., & Rush, A. M. (2016).Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. doi:10.18653/v1/n16-1012

See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). doi:10.18653/v1/p17-1099

Nallapati, R., Zhou, B., Santos, C. D., Gulcehre, C., & Xiang, B. (2016).Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning. doi:10.18653/v1/k16-1028

Chen, J., &Zhuge, H. (2018).Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN.Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. doi:10.18653/v1/d18-1438

Zhang, Y., Chen, E., & Xiao, W. (2018).Extractive-abstractive summarization with pointer and coverage mechanism.Proceedings of 2018 International Conference on Big Data Technologies - ICBDT 18. doi:10.1145/3226116.3226126