International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Short Survey of Automatic Text Summarization Techniques, Algorithms and Their Evaluation

Vipan

Submitted: 21/04/2020 **Accepted**: 08/06/2020

Abstract: Automatic text summarization has become an indispensable tool in the digital era, empowering users to efficiently extract and distill key information from vast troves of textual data. This comprehensive survey paper delves deep into the diverse techniques and algorithms employed in the field of automatic text summarization. It explores the core methods, including both extractive and abstractive approaches, as well as the underlying algorithms and techniques that power these summarization systems. The paper delves into the capabilities and limitations of these summarization methods, discussing their real-world applications in depth. It also examines the current state-of-the-art in automatic text summarization research, highlighting the notable advancements and innovations that are pushing the boundaries of this dynamic field. The survey provides a thorough and insightful analysis, equipping readers with a nuanced understanding of the key trends, challenges, and future directions in the realm of automatic text summarization.

Keywords: Automatic Text Summarization, Extractive Summarization, Abstractive Summarization, Text Mining, Natural Language Processing

A. Introduction

With the exponential growth and generation of digital content, both online and offline, there is an urgent need to produce shorter and more concise versions of lengthy texts. This can only be achieved through the use of automatic text summarization tools. These tools have the ability to save time and effort by reducing the amount of text that needs to be read, as people are generally more interested in brief versions compared to long, detailed documents.[1]

Automatic summarization tools play a crucial role in helping to save time and resources by condensing lengthy texts into more manageable and digestible formats. Moreover, the use of these tools can significantly improve the accessibility, quality, usability, and overall effectiveness of textual information by highlighting the most important points and eliminating redundancy.[2]

The process of text summarization involves distilling a lengthy piece of text into a shorter version without altering the core message or meaning of the original content. It is essential to maintain the original meaning and ensure that the summary text conveys the same level of information as the source material.[3]This is a crucial aspect of effective text summarization, as the summarized text should accurately reflect the intended message of the original document.

In the current data-driven landscape, automatic text summarization can be a valuable tool in dealing with the problem of information overload and topic modeling, where large volumes of data need to be processed and understood. By offering easy-to-understand summaries, these tools can make data more accessible to a wider range of users who may not have the time or expertise to delve into lengthy texts[4].

Text summarization systems have a wide variety of applications, including personalized news feeds that allow users to quickly stay up-to-date on the most relevant news, educational tools that can summarize complex research articles into concise and understandable content, and search engine results that can be summarized to help users quickly scan for the information they require.[5]

B. Types Of Summarization

There are two main types of text summarization methods.

Extractive summarization: it works by extracting or selecting the most relevant sentences from the

Department of Computer Applications, Sikh National College, Banga, Punjab, India.

input text for the final summary text. The identification of sentences for the selection can be based on various criterion like position of sentences, frequency of occurrence, relevance to the core message etc. [6]The selected sentences are directly taken from the source text for the final summary of text. The extracted sentences contain the most important information about the whole text. This technique is simple as compared to the abstractive method because the system needs not to understand the meaning of the text, it has to identify the most important and relevant sentences only.[6]

Abstract summarization: Obstructive Summarization method can generate short and concise summary by understanding the main idea or core message of the original text it is similar to the summaries generated by humans using sentences which may not exist in the source text.[3]

This method is more challenging because instead of selecting and extracting sentences from the original text, it generates a concise summary by understanding the content and rewriting it in a coherent form just like human generated summaries. It ensures that the generated summary is smooth and is linguistically correct.

With the advent of promising research in NLP, language models and deep learning techniques the quality of abstractive summarization system is improved, bringing it close to human generated summaries that effectively holds the essence of the original text.

C. Techniques And Algorithm Used In Automatic Summarization

Statistical Approaches

The automatic text summarization systems, which are based on statistical techniques, are focusing on various statistical features of the text, such as term frequency, sentence length, position of important keywords for the selection of most important sentences for the summary text.

Following are the some commonly used statistical methods for text summarization.

TF-IDF (Term Frequency-Inverse Document Frequency): It is a very common measure that weighs words on their frequency of occurrence in the source text in relation with their occurrence in larger domain of the text. In other words, this method rely on the existence of words or phrases in the source document to determine their importance for the selection in the final summary. TF (Term frequency) means the no. of times a particular term or word appears in the source text. Whereas IDF (Inverse Document Frequency) means how common a particular term, in context of concerned domain.[7]

LSA (Latent Semantic Analysis): It is a mathematical approach that can be used to identify the latent or hidden relationships among the words and sentences in a corpus of text related to a particular domain. This technique applies SVD (Singular Value Decomposition) to matrix of term-document, which is then decomposed into three matrix namely Term matrix, Document matrix and Latent semantic matrix. The latent semantic factors that represents the underlying sentences of the document, can be used to identify sentences or topics that are present in the document. Finally, these semantic factor scores are used in selection of the sentences to represent the summary of the source text. This is a powerful[8] technique, which can be used in conjunction with other techniques like Text Rank to make summary text more informative and concise.

MMR (Maximal Marginal Relevance): This method focuses on the diversity of sentences along with the most important sentences for the final summary. This enhances the informativeness and diversity in gist of the source text. Initially this technique is used for the retrieval documents based on a query but later on, it is extended to text summarization.[9] MMR is an unsupervised method, which can reduce redundancy in final summary by promoting diversity of information, which covers different aspects of the topic. One limitation of MMR method is that it generated summaries, which are not well structured because it does not care about the semantic or syntactic relations.

Graph based methods: following are some most commonly used graph-based methods.

(i) Text Rank: This graph-based approach works by creating a graph for source text where nodes of graph represents a sentence and edges between the nodes of graph represents the relationships between the sentences. The core idea behind this approach is that the important sentences are more likely to connect with other important sentences or phrases. [10]This technique iteratively updates the score of sentences based on the scores of the neighboring

nodes/sentences. The scores are assigned by using a very popular page rank algorithms, which is used by Google to rank web pages. Finally, the summary is then generated by extracting the sentences having highest scores. This approach is very simple and effective for extractive summarization.

(ii) Lex Rank: This technique is actually an extension to the Page Rank method used for text summarization. Tf-idf and cosine similarity models are used to establish the relationship between sentences with the help of edges in between them. Lex Rank uses a measure, which focusses on the lexical similarity in sentences to generate summary.[11]

(iii) Event graph based sentence fusion: In sentence fusion method, the most important sentences of the input text are merged together to form a coherent summary. In this, an event graph captured effectively from the input text. Then these related events are organized in a structured way to guide sentence fusion.[12]

Machine Learning Approaches

Machine learning for text summarization can be divided into two main categories: supervised and unsupervised approaches.

Supervised Approaches: In supervised ML approaches, a model is trained on a large corpus of text-summary pairs, and this trained model is then used to predict the summary sentences for new, unseen text.

Some common supervised ML methods used for text summarization are:

(i) Sequence-to-Sequence Models: These powerful deep learning models use an encoder-decoder architecture to effectively generate abstractive summaries of the input text. The encoder component takes the full input text and encodes it into a compact, fixed-length representation that captures the essential meaning and semantics. The decoder component then uses this encoded representation to progressively generate a new, concise summary text, word by word, in a fluent and coherent manner. This end-to-end neural network approach allows the model to learn complex patterns and relationships in the text, enabling it to produce summary outputs that closely match human-written summaries, while preserving the key information and ideas from the original. By leveraging the strengths of sequence-to-sequence modeling, these techniques can go beyond simple extractive summarization to generate novel, abstractive summaries that are both informative and well-written.

(ii) Classification-based methods: These supervised models use a binary classification approach to determine which sentences from the input text should be included in the summary. The model is trained on a large corpus of text-summary pairs to learn the features and patterns that distinguish important, summary-worthy sentences from less important ones.[13] During this training process, the model learns to recognize characteristics such as sentence position, length, presence of key terms, and semantic similarity to other important sentences that are indicative of a sentence's significance for the summary.[14] Once the model is trained, it can then be applied to new, unseen input text to score and identify the sentences that are most likely to be valuable for inclusion in the summary. The classifier assigns a probability score to each sentence, and the sentences with the highest scores are extracted to form the final summary. This approach allows the summarization model to adapt to the nuances of the input text and generate summaries that closely match human-written ones, by learning the complex patterns that define important, summary-worthy content[15].

(iii) Regression-Based Methods: Regression-based approaches to text summarization involve training a model to predict importance scores for sentences based on various features extracted from the input text. These features can include sentence-level characteristics such as length, position within the document, and the presence of key terms or phrases. The model is trained on a corpus of text-summary pairs, learning to recognize the relationship between these sentence-level features and the human-assigned importance scores for each sentence.[13] Once the regression model has been trained, it can then be applied to new, unseen input text. The model will score and rank each sentence according to its predicted importance, allowing the top-scoring sentences to be selected and extracted to form the final summary.[16] This data-driven, feature-based approach provides a flexible and adaptable way of identifying the most salient content for summarization, as the model can learn to recognize the complex patterns and nuances that define important, summary-worthy sentences within a given text. Regression-based summarization methods offer a powerful alternative to other

supervised and unsupervised techniques, leveraging machine learning to capture the semantic and contextual information that is crucial for generating high-quality, informative summaries.

Unsupervised Approaches: Unsupervised ML approaches for summarization do not require labeled text-summary pairs for training. These methods rely on statistical and graph-based techniques to identify the most important sentences in the input text.

Examples of unsupervised methods include:

(i) Clustering: Clustering techniques for summarization involve grouping together similar sentences or words based on measures of lexical, semantic, or structural similarity. By identifying clusters of related content, these methods can effectively capture the key themes and topics present in the input text. The clustering process typically involves computing pairwise similarity scores between sentences or words, and then partitioning the text into groups of related elements.[17]

Once the clusters have been formed, the most important or representative sentences within each cluster can be selected to create the summary. This approach helps to ensure that the final summary covers the breadth of topics and ideas present in the original text, rather than focusing narrowly on a few dominant themes. Clustering-based summarization techniques can be particularly effective for longer, more complex documents, where the input text discusses a variety of interconnected topics.[18][1] The specific algorithms and similarity measures used for clustering can vary, and may include techniques such as K-means, hierarchical clustering, and community detection in graph-based representations of the text. The choice of clustering method and similarity metric can have a significant impact on the quality and coverage of the resulting summaries. Careful tuning and evaluation of these components is often necessary to achieve optimal summarization performance.

(ii) Graph-Based Methods: In these approaches, the input text is represented as a graph data structure, where the individual sentences form the nodes, and the relationships or similarities between sentences are encoded as weighted edges connecting the nodes. This graph-based representation allows for the application of powerful network analysis algorithms to identify the most central and important sentences within the text.[10] One commonly used graph-based method for text summarization is the PageRank algorithm, which was originally developed for ranking web pages in search engine results. When applied to a sentence graph, the PageRank algorithm calculates a score for each sentence that reflects its relative importance or centrality within the overall text. Sentences with higher PageRank scores are then selected to form the summary, as they are deemed to be the most representative and salient content. The advantage of graph-based methods is their ability to capture the complex relationships and interdependencies between different parts of the text, going beyond simple surface-level features like sentence position or word frequency. By modeling the text as an interconnected network, these approaches can identify key sentences that may not necessarily be the longest or contain the most common words, but are nonetheless critical for conveying the overall meaning and themes of the document. [17][19]This holistic, data-driven perspective on text structure and importance can lead to more coherent and informative summaries compared to simpler extraction-based techniques.

Deep learning Approaches

Recent advancements in deep learning have led to the development of more sophisticated and powerful text summarization models. These approaches typically involve the use of neural networks, which can learn complex, non-linear relationships between the input text and the desired summary[20][15].

Encoder-Decoder Models: One of the most widely used deep learning architectures for summarization is the encoder-decoder model, which is commonly used in sequence-to-sequence (Seq2Seq) tasks such as machine translation. In this framework, the input text is first encoded into a compact, high-dimensional representation using an neural network encoder. This encoded representation is then passed to a decoder network, which generates the output summary one token at a time, using the encoded representation as a starting point. These encoder-decoder models can be further enhanced with attention mechanisms, which allow the decoder to selectively focus on different parts of the input text when generating each output token. This helps the model to better capture the relevant information from the source text and generate more coherent and informative summaries.

Pointer-Generator Networks: Another popular deep learning approach for summarization is the Pointer-

Generator Network, which combines the strengths of the encoder-decoder architecture with the ability to copy or "point" to specific words from the input text, rather than solely generating new words. This hybrid approach allows the model to seamlessly switch between generating new words and copying words directly from the source, which can be particularly useful for handling out-of-vocabulary terms and maintaining factual accuracy in the generated summaries.[21][22]

Attention Mechanisms: Attention mechanisms are a key component of many state-of-the-art deep learning models for text summarization. These mechanisms allow the model to selectively focus on the most relevant parts of the input text when generating each output token in the summary.[23] By dynamically attending to different parts of the input, the model can better capture the salient information and contextual relationships that are crucial for producing coherent and informative summaries. The attention scores computed by the model indicate the level of importance or relevance assigned to each input token, guiding the summary generation process.[24] Attention-based models have been shown to outperform traditional approaches that rely on fixed, predefined features or heuristics. The data-driven, adaptive nature of attention mechanisms enables the model to learn complex, non-linear relationships between the input text and the desired summary, leading to significant improvements in summarization quality. Incorporating attention into sequence-to-sequence architectures, such as encoder-decoder and pointer-generator networks, has been a key factor in the recent advancements of deep learning-based text summarization.[21]

Hybrid Approaches

While the deep learning techniques discussed above have shown impressive results, they are not without their limitations. To address these limitations, researchers have explored hybrid approaches that combine the strengths of different summarization methods. One such approach is the integration of graph-based techniques with deep learning models. These model uses an unsupervised graph-based method to capture the global context and long-distance relationships between sentences, which are then combined with a deep learning-based summary generation component. This hybrid approach aims to leverage the strengths of both graph-based and neural network-based methods, leading to improved summarization performance, especially for languages with limited resources. Another hybrid approach is the Hybrid MemNet model, which combines local and global sentence-level information to generate extractive summaries. The model jointly learns a unified representation of the document, capturing both the local and global sentential information, which is then used to identify the most salient sentences for the summary.[25]

These hybrid approaches demonstrate the potential of combining multiple summarization techniques, leveraging the complementary strengths of different methods to achieve more robust and effective text summarization, especially for challenging scenarios such as low-resource languages or specialized domains.[26]

D. Evaluation Matrics

The performance of text summarization systems is typically evaluated using a variety of metrics, which can be broadly classified into the following categories:

Human Evaluation

This involves having human assessors evaluate the quality of the generated summaries, typically based on criteria such as informativeness, coherence, fluency, and overall quality.

Intrinsic Evaluation

These are automatic metrics that compare the generated summary to one or more reference summaries, often written by humans. The most commonly used intrinsic metrics are:

ROUGE is a commonly used intrinsic evaluation metric for text summarization. It measures the overlap between the generated summary and the reference summaries, in terms of various lexical similarities such as n-grams and longest common subsequences. ROUGE has been widely adopted in the field of automatic text summarization and has become a standard way to assess the quality of summarization systems by comparing their outputs to human-written reference summaries. The use of ROUGE as an evaluation metric is supported by prior research in the area of text summarization, as evidenced by the citations from relevant sources [21] and [27].

BLEU is a metric originally developed for machine translation, but it has also been applied to text

summarization. It measures the precision of n-grams in the generated summary compared to the reference summaries. Specifically, BLEU calculates the geometric mean of the n-gram precisions, with a penalty for sentences that are too short. This provides a way to assess the quality of the generated summary by comparing it to one or more reference summaries, making it a useful intrinsic evaluation metric for text summarization systems[26].

Extrinsic Evaluation

These metrics assess the practical utility and real-world applicability of the generated summaries by evaluating their performance in specific downstream tasks. Unlike intrinsic metrics that focus on textual similarity to reference summaries, extrinsic evaluation measures the effectiveness of the summarization system in aiding higher-level applications such as question answering, text categorization, or information retrieval. [28]These extrinsic metrics provide a more holistic assessment of the summarization system's ability to produce summaries that are useful and beneficial for end-users, rather than just lexically similar to human-written references. By assessing the summaries' impact on the performance of downstream tasks, extrinsic evaluation offers insights into the practical value and relevance of the summarization system in real-world scenarios, which is crucial for the deployment and adoption of these technologies in various domains.[29]

E. Conclusion

The field of automatic text summarization has experienced substantial progress in recent years, driven by advancements in deep learning and natural language processing techniques. Hybrid approaches that combine the strengths of diverse summarization methods have also demonstrated promising results, particularly in addressing challenges related to low-resource languages and specialized domains. While current state-of-the-art models have achieved impressive performance, there remain opportunities for further improvement, especially in areas such as long-form summarization, handling of complex linguistic phenomena, and ensuring the factual accuracy and relevance of the generated summaries. As research in this domain continues to evolve, we can anticipate the emergence of additional advancements that will enhance the practical utility of automatic text summarization across a wide range of applications.

References

[1] K. Thakkar, R. V. Dharaskar, and M. Chandak, "Graph-Based Algorithms for Text Summarization," Nov. 01, 2010. doi: 10.1109/icetet.2010.104.

[2] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques," Aug. 20, 2010. doi: 10.4304/jetwi.2.3.258-268.

[3] A. M. Rush, S. Chopra, and J. Weston, "A Neural Attention Model for Abstractive Sentence Summarization," Jan. 01, 2015, Cornell University. doi: 10.48550/arxiv.1509.00685.

[4] Gevorg Poghosyan, gevorg.poghosyan@insight-centre.org, "Addressing Information Overload through Text Mining across News and Social Media Streams." Sep. 2019. [Available: https://dl.acm.org/doi/10.1145/3345645.3351105

[5] X. Wu, F. Xie, G. Wu, and W. Ding, "PNFS: PERSONALIZED WEB NEWS FILTERING AND SUMMARIZATION," Oct. 01, 2013, World Scientific. doi: 10.1142/s0218213013600075.

[6] S. Saiyed and S. S. Priti, "Literature Review on Extractive Text Summarization Approaches," Dec. 15, 2016. doi: 10.5120/ijca2016912574.

[7] E. Greussing and H. G. Boomgaarden, "Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis," Feb. 01, 2017, Taylor & Francis. doi: 10.1080/1369183x.2017.1282813.

[8] A. K. Singh, M. Gupta, and V. Varma, "Unity in Diversity: Learning Distributed Heterogeneous Sentence Representation for Extractive Summarization," Jan. 01, 2019, Cornell University. doi: 10.48550/arxiv.1912.11688.

[9] A. Singh, M. Gupta, and V. Varma, "Unity in Diversity: Learning Distributed Heterogeneous Sentence Representation for Extractive Summarization," Apr. 27, 2018, Association for the Advancement of Artificial Intelligence. doi: 10.1609/aaai.v32i1.11994.

[10] P. C. R. Raj, A. Bhandari, A. Singh, M. Puri, and S. Malik, "Comparison of Matrix Factorization and Graph-Based Models for Summary Extraction," Mar. 01, 2019. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=8991343

[11] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, "Graph-Based Text Summarization Using

Modified TextRank," in Advances in intelligent systems and computing, Springer Nature, 2018, p. 137. doi: 10.1007/978-981-13-0514-6_14.

[12] K. Filippova and M. Strube, "Sentence fusion via dependency graph compression," Jan. 01, 2008. doi: 10.3115/1613715.1613741.

[13] J. Zhou and A. M. Rush, "Simple Unsupervised Summarization by Contextual Matching," Jan. 01, 2019. doi: 10.18653/v1/p19-1503.

[14] Abhishek Kumar Singh,Manish Gupta,Vasudeva Varma, "Unity in Diversity: Learning Distributed Heterogeneous Sentence Representation for Extractive Summarization." Dec. 2019. Available: https://arxiv.org/pdf/1912.11688v1.pdf

[15] A. M. Rush, H. Seas, S. Chopra, and J. Weston, "A Neural Attention Model for Sentence Summarization," Jan. 01, 2015.

[16] Z. Cao, F. Wei, S. Li, W. Li, M. Zhou, and H. Wang, "Learning Summary Prior Representation for Extractive Summarization," Jan. 01, 2015. doi: 10.3115/v1/p15-2136.

[17] K. V. Kumar, D. Yadav, and A. Sharma, "Graph Based Technique for Hindi Text Summarization," in Advances in intelligent systems and computing, Springer Nature, 2015, p. 301. doi: 10.1007/978-81-322-2250-7_29.

[18] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," Dec. 02, 2008, Elsevier BV. doi: 10.1016/j.eswa.2008.11.022.

[19] A. Sakhadeo and N. Srivastava, "Effective extractive summarization using frequency-filtered entity relationship graphs," Jan. 01, 2018, Cornell University. doi: 10.48550/arxiv.1810.10419.

[20] P. Ren, F. Wei, Z. Chen, J. Ma, and M. Zhou, "A Redundancy-Aware Sentence Regression Framework for Extractive Summarization," Dec. 01, 2016. Available: https://www.aclweb.org/anthology/C16-1004.pdf

[21] R. Paulus, C. Xiong, and R. Socher, "A Deep Reinforced Model for Abstractive Summarization," Jan. 01, 2017, Cornell University. doi: 10.48550/arxiv.1705.04304.

[22] S. Esmaeilzadeh, G. X. Peh, and A. Xu, "Neural Abstractive Text Summarization and Fake News Detection," Jan. 01, 2019, Cornell University. doi: 10.48550/arxiv.1904.00788.

[23] A. See, P. J. Liu, and C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," Jan. 01, 2017. doi: 10.18653/v1/p17-1099.

[24] D. Galanis, Γ. Λάμπουρας, and I. Androutsopoulos, "Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression," Dec. 01, 2012. Available: http://nlp.cs.aueb.gr/pubs/coling2012.pdf

[25] Abhishek Kumar Singh, abhishek.singh@research.iiit.ac.in, Manish Gupta, manish.gupta@iiit.ac.in, Vasudeva Varma, "Hybrid MemNet for Extractive Summarization." Nov. 2017. Available: https://dl.acm.org/doi/10.1145/3132847.3133127

[26] A. Bharadwaj, A. Srinivasan, A. Kasi, and B. Das, "Extending The Performance of Extractive Text Summarization By Ensemble Techniques," Dec. 01, 2019. doi: 10.1109/icoac48765.2019.246854.

[27] A. Singh, M. Gupta, and V. Varma, "Hybrid MemNet for Extractive Summarization," Nov. 06, 2017. doi: 10.1145/3132847.3133127.

[28] R. D. Gaudio, A. Burchardt, and A. Lommel, "Evaluating a Machine Translation System in a Technical Support Scenario," Jan. 01, 2015. Available: https://www.aclweb.org/anthology/W15-5705/

[29] G. Penn and X. Zhu, "A Critical Reassessment of Evaluation Baselines for Speech Summarization," Jun. 01, 2008. [Online]. Available: https://aclanthology.org/P08-1054/