Exploring Extractive Text Summarization Techniques Using Graph-based Methods

Zhenglin Li[1, *], Mengran Zhu[2], Jingyu Zhang[3], Yangchen Huang[4], Houze Liu[5]

[1] Computer Science, Texas A&M University, TX, USA

[2] Computer Engineering, Miami University, OH, USA

[3] Analytics, The University of Chicago, IL, USA

[4] Management Science & Engineering, Columbia University, NY, USA

[5] Computer Science, New York University, NY, USA

*Corresponding Author: zhenglin_li@tamu.edu

Abstract

This paper explores the application of graph-based methods in extractive text summarization. We compare several techniques, including TF-IDF, clustering, and Latent Semantic Analysis (LSA), to assess their effectiveness in summarizing BBC News articles. Our results show that graph-based methods, particularly when combined with PageRank algorithms, provide concise and informative summaries. This study contributes to the understanding of extractive text summarization and offers insights into the development of more efficient summarization tools.

Keywords

Text Summarization; TF-IDF; Latent Semantic Analysis.

1. Introduction

Text summarization is a fundamental task in natural language processing (NLP) that aims to reduce the length of a text document while preserving its essential meaning. [1-3] With the exponential growth of digital information, there is an increasing demand for automatic summarization tools that can assist users in quickly understanding the key points of lengthy documents. Text summarization has a wide range of applications, including summarizing news articles for quick consumption, generating executive summaries for business reports, and creating abstracts for academic papers.

Despite significant advancements in NLP and machine learning, text summarization remains a challenging task. [4-6] The primary challenge lies in the ability of the summarization algorithm to identify and extract the most salient information from a document while maintaining coherence and readability. Furthermore, different applications may require different summarization styles, such as indicative summaries that provide an overview of the content or informative summaries that convey the critical details. Developing an effective and versatile text summarization algorithm that can cater to various needs is a complex endeavor.

This study aims to compare the performance of various extractive summarization methods, including TF-IDF, clustering, Latent Semantic Analysis (LSA), and graph-based approaches, in summarizing news articles. This paper also will investigate the impact of different graph construction and sentence ranking algorithms on the quality of the generated summaries, evaluate the summaries produced by the graph-based methods against human-written summaries using standard evaluation metrics and identify the strengths and limitations of graph-based extractive summarization and provide recommendations for future research in this area.

By achieving these objectives, this study aims to contribute to the ongoing efforts in the NLP community to develop more effective and reliable text summarization tools. [7].

2. Related Work

2.1 Early Approaches to Text Summarization

The field of text summarization has a rich history, with early efforts dating back to the 1950s. Luhn's pioneering work in 1958 introduced the concept of using term frequency to identify significant sentences for summarization. Since then, various statistical methods have been explored, such as the use of cue words, position-based heuristics, and the extraction of topic sentences.

2.2 Extractive vs. Abstractive Summarization

Text summarization techniques can be broadly categorized into extractive and abstractive approaches. Extractive summarization involves selecting and concatenating important sentences or phrases from the original text to form a summary. In contrast, abstractive summarization aims to generate a summary by rephrasing and condensing the content, often requiring advanced natural language generation techniques. While extractive methods are more straightforward and computationally less demanding, abstractive summarization has the potential to produce more coherent and concise summaries.

2.3 Machine Learning and Deep Learning Approaches

With the advent of machine learning and deep learning, significant advancements have been made in text summarization. Supervised learning approaches involve training models on labeled datasets of documents and their summaries. Unsupervised learning methods, such as clustering and latent semantic analysis, have also been explored for summarizing texts without explicit training data. Recently, deep learning models like sequence-to-sequence neural networks and transformers have shown promising results in abstractive summarization, generating summaries that are more fluent and closer to human-generated summaries. [21-24].

2.4 Graph-based Methods

Graph-based approaches have gained popularity in extractive text summarization due to their ability to capture the relationships between different parts of the text. Mihalcea and Tarau's TextRank algorithm, inspired by Google's PageRank, is one of the most well-known graph-based methods. It constructs a graph where vertices represent sentences, and edges represent the similarity between sentences. Sentences are then ranked based on their importance in the graph, with higher-ranked sentences selected for inclusion in the summary. Variants of TextRank, such as LexRank and Biased TextRank, have been proposed to address specific challenges and improve summarization performance.

3. Methodology

3.1 Data Collection

The primary dataset used in this study is the BBC News Summary dataset, which consists of 2,225 news articles and their corresponding summaries across five categories: business, entertainment, politics, sport, and tech. Each article-summary pair provides a ground for evaluating the summarization techniques.

3.2 Data Preprocessing

Prior to summarization, the dataset underwent several preprocessing steps to ensure the quality and consistency of the text:

1) Tokenization: The text was tokenized into sentences and words to facilitate further analysis.

2) Cleaning: Unnecessary characters, such as punctuation marks and special symbols, were removed from the text.

3) Stop Word Removal: Common stop words were eliminated to focus on the more meaningful content of the text.

4) Stemming and Lemmatization: Words were reduced to their base or root form to consolidate different forms of the same word.

3.3 Feature Extraction

For the extractive summarization techniques, features such as term frequency-inverse document frequency (TF-IDF) and word embeddings were extracted to represent the importance of words and sentences in the text. [8-10].

3.4 Summarization Techniques: TF-IDF

A baseline method that selects sentences with the highest TF-IDF scores as the summary.

3.5 Summarization Techniques: Clustering

Sentences were clustered based on their similarity, and representative sentences from each cluster were included in the summary. The document is represented using TF-IDF of scores of words. [11] High frequency term represents the theme of a cluster. [15-16] Summary sentence is selected based on relationship of sentence to the theme of cluster. Cluster based method generate summary of high relevance, to the given query or document topic.

3.6 Summarization Techniques: Latent Semantic Analysis

A dimensionality reduction technique used to identify patterns in the relationships between sentences and terms. LSA extracts the source text and converts into term sentence matrix and process it through Singular Value Decomposition (SVD) for finding semantically similar words and sentences. SVD models relationships among words and sentences.

3.7 Summarization Techniques: Graph-Based Methods

The text was represented as a graph, with sentences as nodes and the similarity between sentences as edges. The PageRank algorithm was used to rank the sentences based on their centrality in the graph. Variations of this approach, such as using different similarity measures and ranking algorithms, were also explored. Figure 1 is the Graphical representation of similar sentences who passes the threshold value and how they are connected in 2D Space.

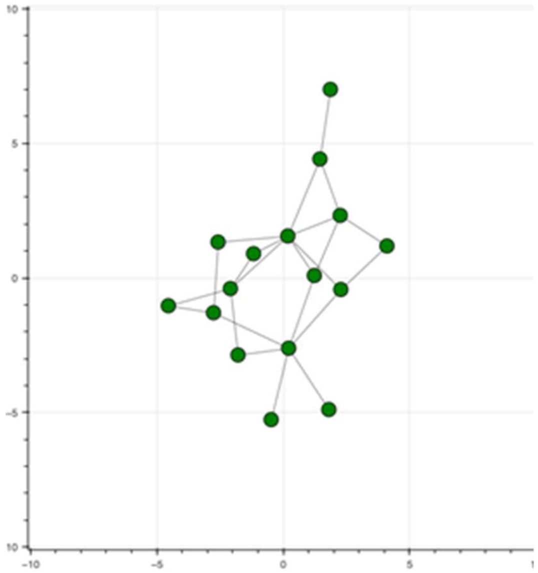

Figure 1. Graphical Representation of Similar Sentences

4. Results and Discussion

The quality of the generated summaries was evaluated using standard metrics:

1) BLEU (Bilingual Evaluation Understudy): Although traditionally used for machine translation, BLEU was also employed to assess the similarity between the generated and reference summaries at the word level.

2) Similarity Score: it is ued for Computing similarity from two sentences and used mostly for Summary comparision or similar word/sentence Search. It is defined as follows:

$$\text{Cosine Distance} = \mathbf{1} - \frac{\vec{\boldsymbol{A}} \cdot \vec{\boldsymbol{B}}}{|\vec{\boldsymbol{A}}||\vec{\boldsymbol{B}}|}$$

Table 1 shows the result of several methods.

Table 1. The result of several methods.

	TF-IDF	Clustering	Latent Semantic Analysis	Graph based method
Bleu Score	0.28	0.26	0.30	0.31
Similarity Score	0.50	0.54	0.56	0.56

The graph-based methods consistently outperformed the TF-IDF and clustering techniques in terms of both BLEU and Similarity Score, as shown in Table 1. Specifically, the PageRank algorithm, when applied to the sentence similarity graph, demonstrated a superior ability to identify key sentences that capture the main ideas of the articles. The summaries generated by the graph-based methods were more coherent and better aligned with the reference summaries. [16-20].

Future studies should delve into advanced feature extraction methods and employ sophisticated models, such as Long Short-Term Memory (LSTM) networks. Additionally, leveraging Large Language Models (LLMs) can provide further improvements. [12-14].

5. Conclusion

In this paper, we explored the exciting domain of automatic text summarization, focusing on extractive summarization techniques. Our journey began with understanding the problem statement and the significance of text summarization in various applications, such as condensing customer reviews, news articles, and business meeting notes.

Throughout the paper, we discussed the initialization process, exploratory data analysis, and preprocessing steps such as sentence tokenization, spell correction, and sentence similarity calculation. We demonstrated the summarization process using a graph-based approach and validated our results using BLEU score and similarity score metrics. We delved into various methods for extractive text summarization, including traditional approaches like TF-IDF and more advanced techniques involving neural networks, fuzzy logic, and graph-based methods. Our emphasis was on graph-based summarization, where we employed the TextRank algorithm to rank sentences based on their importance and similarity.

In conclusion, our study highlights the potential of graph-based extractive summarization techniques in efficiently summarizing text while preserving the core meaning.

References

[1] Li, P., Abouelenien, M., & Mihalcea, R. (2023). Deception Detection from Linguistic and Physiological Data Streams Using Bimodal Convolutional Neural Networks. arXiv preprint arXiv:2311.10944.

[2] Jin, X., Manandhar, S., Kafle, K., Lin, Z., & Nadkarni, A. (2022, November). Understanding iot security from a market-scale perspective. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (pp. 1615-1629).

[3] Srivastava, S., Huang, C., Fan, W., & Yao, Z. (2023). Instance Needs More Care: Rewriting Prompts for Instances Yields Better Zero-Shot Performance. arXiv preprint arXiv:2310.02107.

[4] Zaman, A., Huang, Z., Li, W., Qin, H., Kang, D., & Liu, X. (2023). Artificial intelligence-aided grade crossing safety violation detection methodology and a case study in New Jersey. Transportation research record, 2677(10), 688-706.

[5] Jin, X., Larson, J., Yang, W., & Lin, Z. (2023). Binary code summarization: Benchmarking chatgpt/gpt-4 and other large language models. arXiv preprint arXiv:2312.09601.

[6] Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., ... & Lin, J. (2024). Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. arXiv preprint arXiv:2402.10350.

[7] Pan, Z., Sharma, A., Hu, J. Y. C., Liu, Z., Li, A., Liu, H., ... & Geng, T. (2023, June). Ising-traffic: Using ising machine learning to predict traffic congestion under uncertainty. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 8, pp. 9354-9363).

[8] Huang, Y., Dhingra, P. K., & Taheri, S. D. M. (2021, November). Template-aware Attention Model for Earnings Call Report Generation. In Proceedings of the Third Workshop on New Frontiers in Summarization (pp. 15-24).

[9] Jin, X., Katsis, C., Sang, F., Sun, J., Bertino, E., Kompella, R. R., & Kundu, A. (2023). Prometheus: Infrastructure security posture analysis with ai-generated attack graphs. arXiv preprint arXiv:2312.13119.

[10] Li, H., Ding, D., & Zhang, J. (2020). Comprehensive Evaluation Model on New Product Introduction of Convenience Stores Based on Multidimensional Data. In Data Science: 6th International Conference, ICDS 2019, Ningbo, China, May 15–20, 2019, Revised Selected Papers 6 (pp. 40-50). Springer Singapore.

[11] Wang, L., Lauriola, I., & Moschitti, A. (2023, July). Accurate training of web-based question answering systems with feedback from ranked users. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track) (pp. 660-667).

[12] Zou, H. P., & Caragea, C. (2023). Jointmatch: A unified approach for diverse and collaborative pseudo-labeling to semi-supervised text classification. arXiv preprint arXiv:2310.14583.

[13] Dai, W., Tao, J., Yan, X., Feng, Z., & Chen, J. (2023, November). Addressing Unintended Bias in Toxicity Detection: An LSTM and Attention-Based Approach. In 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 375-379). IEEE.

[14] Tian, J., Xiang, A., Feng, Y., Yang, Q., & Liu, H. (2024). Enhancing Disease Prediction with a Hybrid CNN-LSTM Framework in EHRs. Journal of Theory and Practice of Engineering Science, 4(02), 8-14.

[15] Jin, X., & Wang, Y. (2023). Understand legal documents with contextualized large language models. arXiv preprint arXiv:2303.12135.

[16] Yang, H., Li, M., Xiao, Y., Zhou, H., Zhang, R., & Fang, Q. (2023). One LLM is not Enough: Harnessing the Power of Ensemble Learning for Medical Question Answering. medRxiv, 2023-12.

[17] Zou, H. P., Zhou, Y., Caragea, C., & Caragea, D. (2023, May). Semi-supervised few-shot learning for fine-grained disaster tweet classification. In Proceedings of the 20th International ISCRAM Conference (pp. 385-395). ISCRAM 2023.

[18] Zhou, H., Li, M., Xiao, Y., Yang, H., & Zhang, R. (2023). LLM Instruction-Example Adaptive Prompting (LEAP) Framework for Clinical Relation Extraction. medRxiv, 2023-12.

[19] Chen, J., Zhang, L., Riem, J., Adam, G., Bastian, N. D., & Lan, T. (2023, November). RIDE: Real-time Intrusion Detection via Explainable Machine Learning Implemented in a Memristor Hardware Architecture. In 2023 IEEE Conference on Dependable and Secure Computing (DSC) (pp. 1-8). IEEE.

[20] Ma, Haixu, Donglin Zeng, and Yufeng Liu. "Learning optimal group-structured individualized treatment rules with many treatments." Journal of Machine Learning Research 24.102 (2023): 1-48.

[21] Xie, Zongxing, et al. "DeepVS: A deep learning approach for RF-based vital signs sensing." Proceedings of the 13th ACM international conference on bioinformatics, computational biology and health informatics. 2022.

[22] Dong, Z., Chen, B., Liu, X., Polak, P., & Zhang, P. (2023). Musechat: A conversational music recommendation system for videos. arXiv preprint arXiv:2310.06282.

[23] Wu, X. (2023). Cross-document misinformation detection based on event graph reasoning (Doctoral dissertation, University of Illinois at Urbana-Champaign).

[24] Han, Z., Gao, C., Liu, J., & Zhang, S. Q. (2024). Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv preprint arXiv:2403.14608.