A Review on the Application of Automatic Text Summarization and Its Bias Analysis

Zehao Li

School of Computer Science and Engineering, Tianjin University of Technology, Tianjin, China

zsb@tjut.edu.cn

Abstract. In the era of information explosion, the Internet produces a large number of text documents daily, providing rich information but also posing a challenge: how to swiftly extract key information. Traditional manual reading is time-consuming and inadequate for handling vast data. Thus, automatic text summarization technology emerges as a crucial solution. This paper reviews the application and deviation analysis of this technology across various fields, focusing on addressing shortcomings of traditional methods, such as initial cluster center selection and redundancy. An automatic text summarization method based on an improved TextRank algorithm and K-Means clustering is introduced. Existing methods often struggle with inaccurate initial clustering center selection and high summary redundancy, especially with long texts, resulting in summaries that fail to reflect core content accurately. Furthermore, the widespread use of pre-trained language models introduces potential biases that can propagate to downstream tasks, affecting summary accuracy and impartiality. To address these issues, this paper proposes an innovative automatic text summarization method that optimizes initial clustering center selection and clustering refinement strategies to enhance summary accuracy and readability. Additionally, it discusses name-nationality bias in pre-trained language models and its propagation in text summary tasks, offering a theoretical foundation and practical guidance for developing a more just and reliable Natural Language Processing (NLP) system.

Keywords: Automatic text summarization, Deviation analysis, Pre-trained language models.

1. Introduction

As an important branch in the field of natural language processing, automatic text summarization technology is becoming more and more important under the background of information explosion. With the rapid development of Internet and digital technology, people are faced with a huge amount of text information every day. How to extract the key information quickly and accurately has become an urgent problem to be solved. Automatic text summarization technology is born to solve this problem. It aims to compress long text into a short summary containing the core content by means of computer automation, so as to help people acquire and understand information more efficiently.

The development of automatic text summarization technology can be traced back to the 1950s. As the figure shows, early research focused on simple methods based on keyword frequency and sentence placement. These methods, though simple, set the stage for subsequent research. By the '70s and' 80s, statistical methods began to play an important role in text summaries as computer technology advanced. Among them, techniques such as word frequence-inverse document frequency(TF-IDF) and text clustering have been widely used, which has significantly improved the quality of the abstract.

In the 1990s and early 2000s, machine learning methods began to make their mark in the field of automatic text summarization. Algorithms such as support, allowing the summarization system to better learn and utilize the features of the text. The research during this period laid an important foundation for the subsequent application of deep learning methods.

In the 2010s, the rise of deep learning revolutionized automatic text summarization technology. The introduction of sequence-to-sequence models and attention mechanisms allowed the summarization system to better capture the long-term dependencies and key information of the text. The application of these technologies has greatly improved the quality of the generated summary, making the automatically generated summary closer to the manual summary in terms of fluency and coherence.

In the 2020s, pre-trained models and multimodal techniques have become the new focus of automatic text summary research. With the application of Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-training Transformer (GPT) and other large-scale pre-trained language models, the abstract system can better understand the semantic and contextual information of text. At the same time, the development of multi-modal fusion technology also provides many possibilities for processing complex information types such as text and text mixing

The application of automatic text summarization technology is very wide, covering many important fields. Keyword extraction has been paid great attention in library science, information science, natural language processing and other fields. In the early stage, keyword extraction was completed by manual annotation with the help of human expert knowledge, which is a very heavy work. With the development of computer technology, automatic keyword extraction has attracted more and more attention. A large number of automatic keyword extraction techniques, frameworks and tools have emerged, and these methods have achieved certain achievements and good results. However, the performance of automatic keyword extraction is still low, and there is still a long way to go before the real solution of the task, so it is urgent to further improve the efficiency and quality of extraction or annotation [1]. In practical applications, the generative summary method has shown advantages in many fields. In terms of scientific and technological literature abstracts, the method based on deep learning can effectively extract the structural elements of scientific and technological literature [2]. This approach not only generates coherent abstracts, but also identifies and preserves key terms and research findings in the literature. In the aspect of scientific and technological literature summarization, automatic text summarization technology can help researchers quickly understand the core content of the paper and improve the efficiency of literature retrieval and reading. Shi Lei et al. 's research review pointed out that the sequence-based model based generative text abstract has shown a good application prospect in the field of scientific and technological literature [3]. News summary is another important application field, and automatic summary technology can help news organizations quickly generate news summaries to meet readers' needs for real-time information. Zhao Hong's research review discusses in detail the application of deep learning methods in generated automatic summarization, which provides important references for the development of news summarization technology [4].

In addition, source code summarization and cross-language summarization are also emerging application fields of automatic text summarization. The research review of Song Xiaotao et al. discusses the automatic source code summarization technology based on neural networks, which provides new tools for software development and maintenance [5]. Zheng Bofei et al. 's research focuses on cross-language summarization methods, which provide important support for cross-language information processing and communication [6]. To solve this problem, this paper proposes an automatic text summary method based on an improved TextRank algorithm and K-Means clustering, which aims to improve the accuracy and readability of abstract by optimizing the initial clustering center selection and clustering refinement strategy.

2. Bias analysis in automatic text summarization

Figure 1 is the flow chart of the text automatic summarization method based on the improved TextRank algorithm and K-Means clustering [7]. In the figure, the method is divided into four main modules: vector generation, sentence computation, sentence clustering and sentence selection. Vector generation is based on the basic preprocessing of text, using Word2Vec to generate word vectors, and then using TFIDF to weight sentence vectors. The sentence calculation compares the similarity of each pair of sentence vectors using the improved BM25 model, and the TR score is obtained using the TextRank algorithm. Sentence clustering builds a similar sequence using TR scores and cosine similarity, followed by a similar sequence through TR scores and cosine similarity, an articulated approach that relies on clear criteria around similarity difference and maximum value, determining the initial center of the cluster, followed by a very detailed process of cluster subdivision and refinement. Sentence selection first obtains the cluster value based on the cluster size and cosine similarity of the sentence, then computes the sentence score based on three metrics, and finally selects the sum based on the above two scores.

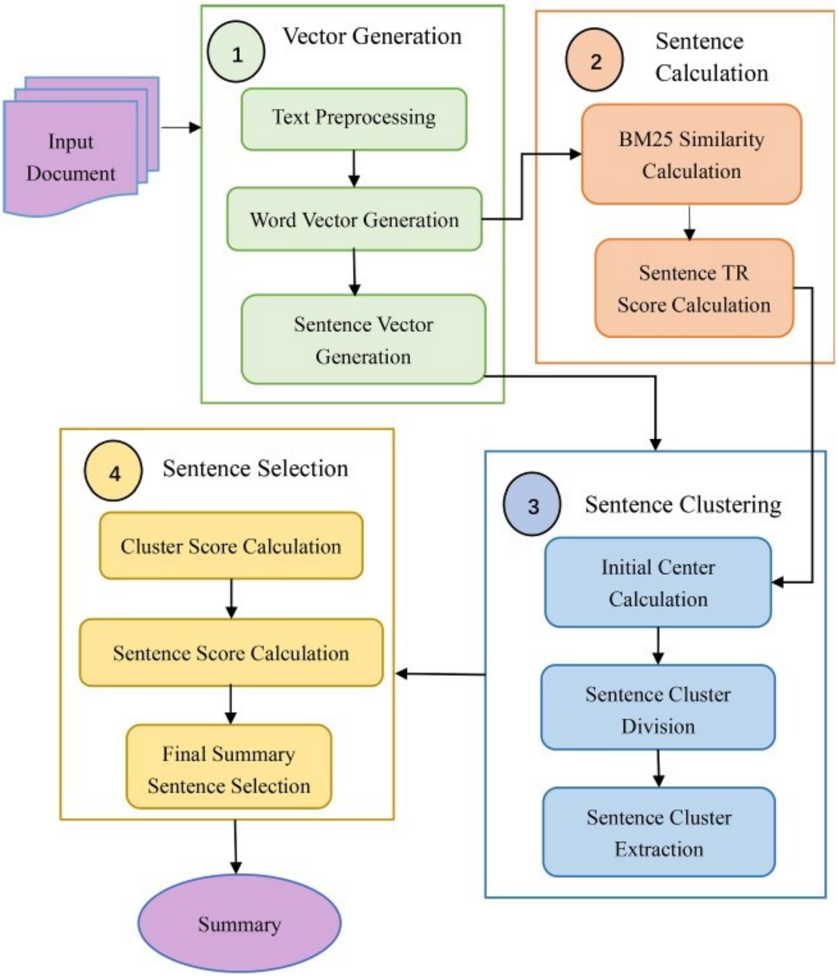
Input Document
1
Vector Generation
Text Preprocessing
Word Vector Generation
Sentence Vector Generation
2
Sentence Calculation
BM25 Similarity Calculation
Sentence TR Score Calculation
4
Sentence Selection
Cluster Score Calculation
Sentence Score Calculation
Final Summary Sentence Selection
3
Sentence Clustering
Initial Center Calculation
Sentence Cluster Division
Sentence Cluster Extraction
Summary

Figure 1. text automatic summarization [7]

In the in-depth exploration of automatic text summarization technology, this experiment introduces and implements a series of detailed comparative experiments, aiming at comprehensively evaluating and comparing the performance and advantages of four automatic summarization algorithms, namely Lead-3, TextRank, MBM25EMB and IMTextRank-K-Means. The following is a detailed introduction of the methods:

2.1. The Lead-3 method: A basic and intuitive summary strategy

As a basic abstract strategy, the academic significance of the Lead-3 method is to extract information by selecting the first two sentences and the last sentence of an article as the content of the abstract, based on the general characteristics of the text structure -- that is, the author introduces the topic at the beginning of the article and summarizes it at the end. Despite the simplicity of the method, it performs well in text with a clear structure and clear topic, effectively capturing key points and providing a valuable reference baseline for subsequent automatic summarization algorithms.

2.2. TextRank method: An innovative breakthrough from the graph theory perspective

TextRank method is a major innovation in automatic summarization technology based on graph theory. This method regards text as a graph composed of sentence nodes, constructs a node connection graph by calculating the similarity between sentences, and assigns TextRank values to sentences with the help of an iterative algorithm. This process not only embodies the advantages of graph theory in dealing with complex relational networks, but also ensures the accuracy and representativeness of the content of the summary. The introduction of TextRank method significantly improves the performance of automatic summarization technology and provides important methodological support for subsequent research.

2.3. MBM25EMB method: a new method of unsupervised summarization based on fusion techniques

The MBM25EMB method further promotes the development of unsupervised automatic summarization technology. This method cleverly combines the advantages of word embedding technology with the traditional TF-IDF and BM25 models. The similarity between sentences is calculated by the improved BM25 model, and the summary is generated by TextRank's graph sorting method. This fusion strategy not only improves the accuracy and coherence of the abstracts, but also demonstrates the powerful potential of modern natural language processing techniques in automatic summarization tasks. The successful practice of MBM25EMB provides a new perspective and idea for the research of automatic summarization.

IMTextRank-K-Means: an innovative clustering optimization abstract strategy

The core contribution of this study lies in the proposed method of IMTextRank-K-Means. On the basis of inheriting the advantages of TextRank and BM25, this method innovatively introduces K-Means clustering algorithm. By optimizing the selection of initial cluster center and the process of cluster refinement, it effectively solves the problems of traditional methods in handling the selection of initial cluster center and redundancy control. IMTextRank-K-Means method not only improves the accuracy and diversity of abstract, but also significantly reduces the occurrence of redundant information, which sets a new milestone for the development of automatic text summary technology. The experimental results show that the method performs well on multiple evaluation indexes, providing a new method and tool for research in related fields.

In summary, the performance and characteristics of the four automatic summarization algorithms are fully demonstrated in this study through comparative experiments and in-depth analysis. Among them, IMTextRank-K-Means stands out for its innovation and excellent performance, which points out the direction for future research on automatic text summarization technology. At the same time, the paper are also aware of the challenges and opportunities existing methods have in dealing with sentence placement, multilingual support, and the introduction of advanced pre-training models, which will become important topics for future research.

In the experiment, in this study, the paper designed and implemented four automatic summarization techniques, namely Lead-3, TextRank, MBM25EMB and IMTextRank-K-Means. The experimental results show that IMTextRank-K-Means method is particularly outstanding in improving the performance of automatic summarization. This research is mainly divided into two parts: experimental design and experimental results. In the experimental design part, the method of automatic text summarization, the selection of initial data and the setting of evaluation indexes are described in detail. The experimental results mainly show the comparative data of ROUGE-1, ROUGE-2 and ROUGE-L.

These experimental data strongly prove that, compared with other methods, IMTextRank-K-Means can generate more accurate summaries and show higher accuracy in the extraction of key sentences.

In the experiment, the performance of four automatic text summarization techniques was compared under the same conditions, and the data sets and initial data used were consistent. The evaluation criterion is ROUGE score, which directly reflects the degree of similarity between candidate abstracts and reference abstracts, and then reflects the advantages and disadvantages of automatic summarization. Among them, candidate abstracts are generated by automatic summarization, while reference abstracts are derived from the data set itself. The initial data for this study was derived from the DUC2004 dataset, which is provided by the Document Understanding Conference (DUC) and is one of the most authoritative evaluation resources in the field of text abstracts. DUC2004 dataset consists of five tasks, of which Task 2 is selected as the analysis object in this study. Task 2 consists of 50 English categories, each of which contains 10 documents, and each such category is treated as an independent document set in this study. For each set of documents, four reference summaries are provided for comparison. The performance of candidate abstracts and reference abstracts in the three dimensions of ROUGE-1, ROUGE-2 and ROUGE-L is emphasized to evaluate the performance of the proposed abstracts generation algorithm. Specifically, ROUGE-1 evaluates the accuracy of candidate abstracts at the word level by calculating the proportion of the number of the same words in the candidate abstracts and the reference abstracts. ROUGE-2 further takes into account the overlap of word pairs (i.e., two adjacent words), and measures the relevance of candidate abstracts in phrase combinations by comparing the overlap of word pairs in the candidate abstracts and the reference abstracts. ROUGE-L, on the other hand, calculates the length of word sequence shared between the two abstracts based on the concept of the longest common subsequence, which serves as the basis for evaluating the overall relevance of the candidate abstract and the reference abstract. In summary, these three indicators together constitute a comprehensive evaluation of the accuracy and relevance of candidate abstracts. This experiment was conducted on the DUC2004 dataset and compared with methods such as Lead-3, TextRank and MBM25EMB. The experimental results show that the IMTextRank-K-Means method proposed in this paper has excellent performance on the three evaluation indexes of ROUGE-1, ROUGE-2 and ROUGE-L, which is significantly superior to other comparison methods. This proves the effectiveness of the method in improving the accuracy of abstracts and reducing redundancies. The experimental performance of the data set is shown in the following three figures 2-4.

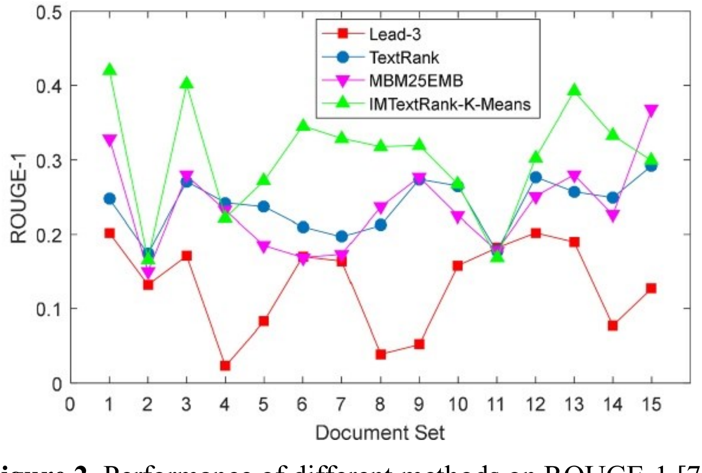

Figure 2. Performance of different methods on ROUGE-1 [7]

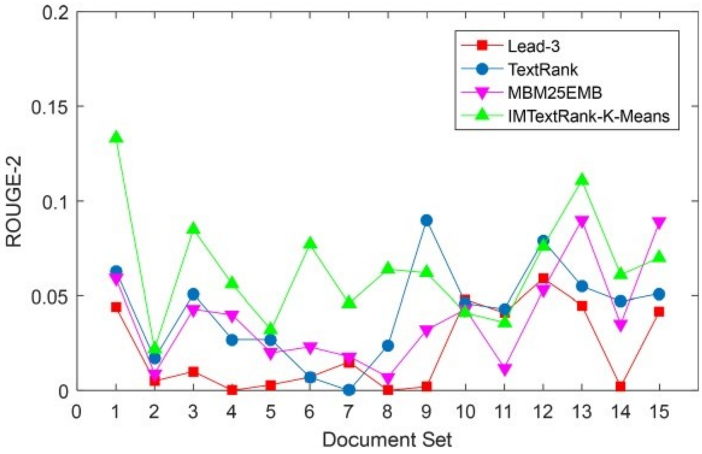

Figure 3. Performance of different methods on ROUGE-2 [7]

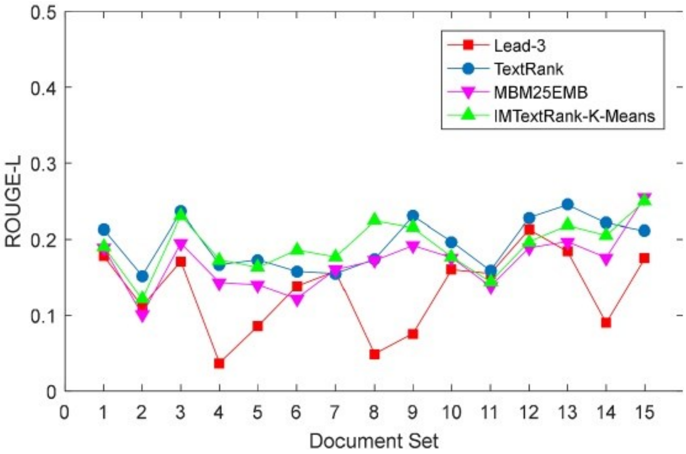

Figure 4. Performance of different methods on ROUGE-L [7]

In summary, the proposed automatic text summary method [8] based on the improved TextRank algorithm and K-Means clustering can effectively solve the problems of inaccurate and high redundancy of initial cluster center selection in traditional methods through innovative initial cluster center selection strategy and cluster refinement method. The experimental results show that the proposed method has significant advantages in improving the accuracy of abstracts and reducing redundancy. Future studies can further optimize sentence position processing, explore multi-language support, and introduce more advanced pre-training models to further improve summary performance

The method proposed in this experiment has achieved remarkable results in the field of automatic text summarization, but there is still some room for improvement. First, the existing method may sacrifice some sequential logic when dealing with the position of the sentence, resulting in a decrease in the readability of the summary. Future research may consider combining sentence position information for ordering optimization. Secondly, the proposed method is mainly based on English text, so its applicability in multilingual environments can be explored in the future. In addition, the use of more advanced pre-training models such as BERT for sentence vector modeling is expected to further improve the accuracy of sentence similarity calculation.

Research on the propagation of pretraining bias in text abstracts

With the rapid development of deep learning technology, large pre-trained language models (such as BERT, GPT, etc.) have made remarkable achievements in the field of natural language processing (NLP), and are widely used in multiple tasks such as text classification, machine translation, and text summarization. By pre-training on large-scale corpora, these models are able to capture rich linguistic knowledge and contextual information, thus performing well in a variety of downstream tasks. However, as the research progresses, scholars are gradually finding that these pre-trained models may contain sociocultural and other types of biases that may spread to downstream tasks under certain circumstances, affecting the performance and fairness of the models. Text summarization is an important task in the field of NLP, which aims to automatically extract key information from long texts to generate concise and clear summaries. However, existing text summary systems, especially those based on pre-trained models, can suffer from the phenomenon of "hallucination" when generating summaries, in which the generated information does not exist in the original text or contradicts the original information. This phenomenon of disloyalty not only affects the accuracy of the abstract, but may also raise ethical and legal issues. Therefore, studying how biases in pre-trained models spread to downstream tasks such as text summarization has important implications for developing more reliable and impartial NLP systems. In this experiment, the literature looked at one type of bias - name-nationality bias - and traced it from the pre-training stage to downstream summary tasks across multiple summary modeling choices. The paper show that these biases manifest as hallucinations in the summary, leading to factually incorrect summaries. Paper also find that the propagation of this bias is algorithmically dependent: more abstract models allow the bias to propagate more directly to downstream tasks as hallucinatory facts. Building on these observations, paper further analyze how changes in adaptive methods and fine-tuning datasets affect name nationality bias, and show that while they can reduce the overall incidence of hallucinations, they do not change the type of bias that does occur.

The literature consists primarily of identifying and quantifying name-nationality bias in pre-trained models: Quantifying intrinsic bias in models by analyzing the accuracy of pre-trained models' predictions of different name-nationality associations. The second is to explore the transmission mechanism of name-nationality bias in the text summary: By constructing specific data sets and designing experiments, the paper analyzes how the bias in the pre-training model affects the output of the downstream text summary task, leading to the appearance of name-nationality illusion. And evaluate the impact of different modeling choices on bias propagation: including pre-trained models, fine-tuning datasets and adaptive methods, and explore how these factors change the mode and extent of bias propagation in downstream tasks. Finally, strategies to mitigate bias propagation in text summaries are proposed: Based on the research results, effective mitigation strategies are proposed to reduce the impact of pre-trained model bias on text summary tasks.

The research methods and processes mainly include data set construction. In order to study the transmission mechanism of name-nationality bias in text abstracts, a new evaluation dataset -- WIKINATIONALITY was constructed in this study. The dataset was constructed by collecting the entity list and scraping the biography page in the following steps: Scraping the corresponding Wikipedia biography page for each entity on the list. Next extract the intro paragraphs: Take the intro paragraphs of each biographical page as the input articles for the text summary model. Subsequently, by perturbing the entity names in the input articles and exchanging them with new names of different nationalities, this study was able to systematically evaluate the name-nationality illusion phenomenon in the model under different nationality entities

The literature on experimental design mainly focuses on the study and selection of two advanced text summary models, BART and PEGASUS, and fine-tuning on the XSUM dataset. To explore the effects of different modeling options on bias propagation, the following experiments were designed in this study: Pre-trained model selection: Compare BART and PEGASUS 'performance differences in generating summaries, especially the frequency of name-nationality hallucinations. Fine-tuning dataset selection: Fine-tuning the BART model on different datasets such as XSUM, CNN-DM, and NYT to observe the effects of fine-tuning datasets on bias propagation. Adaptive method selection: Compare the alleviating effects of standard fine-tuning, adapter fine-tuning, and fine-tuning of the last layer of the fine-tuning decoder on name-nationality illusion. Finally, the illusion is defined: The paper define the nationality illusion as a generated summary that references the original nationality of the inserted entity rather than the nationality in the input entry. The hallucination rate is simply the percentage of abstracts that contain the nationality hallucination. The paper measure hallucination rates at different levels of granularity -

each country, each continent, and each model, The comparison of data between the two groups is shown in Table 1.

Table 1. Compare the two sets of data

	ROUGE-L	Density	American	European	Asian	Afican
BART-XSUM	36.38	2.04	2.83	13.08	27.10	3.66
PEGASUS-XSUM	38.33	8.53	0.62	1.37	4.57	1.60

Density and hallucination rate for BART and PEGASUS. Hallucination rate refers to the percentage of abstracts that contain hallucinations related to nationality. The paper results suggest that PEGASUS is significantly more extractive than BART, and as a result, paper does not observe name-nationality hallucinations in PEGASUS as paper does in BART

Hallucination rates for BART fine-tuned on XSUM. Red corresponds to a higher hallucination rate, blue to a lower hallucination rate. Paper observed that Asian nationalities had higher hallucination rates This graph shows the hallucination rates for each pair of countries, i.e. when paper replace the entity from the original nationality with a new entity from the disturbed nationality. Paper observes that the hallucination rate is significantly higher for Asian nationalities. For example, the BART-XSum model directly produces Korean and Vietnamese nationalities in one-third of the generated summaries

Correlation of intrinsic bias in downstream summary tasks with extrinsic hallucination rates when the paper changed the pre-trained model and fine-tuned the data set. A strong positive correlation was present across all Settings

The adaptation method on XSum. Ovr is the overall hallucination rate for all countries. BART-adapter can achieve a lower hallucination rate while maintaining a similar ROUGE score, and with less extraction than bart-fine-tune

Table 2. observe the ovr data.

	ROUGE-L	Density	American	European	Asian	African	Ovr
BART-fine-tune	36.38	2.05	2.83	13.08	27.10	3.66	12.87
BART-adapter	35.11	1.72	2.06	8.14	12.76	1.37	6.71
BART-last-layer	32.63	4.67	0.71	3.04	11.58	1.03	4.55

The experimental results are shown in Table 2. Through the analysis of experimental results, the research found that:

There is a strong correlation between the inherent bias in the pre-training model and the hallucination in the downstream task; Pre-trained models with stronger name-nationality associations were more prone to name-nationality hallucinations when generating summaries.

Different modeling choices have a significant effect on bias propagation: more abstract models (such as BART) allow bias to propagate more directly into downstream tasks; Models that are fine-tuned on cleaner datasets reduce hallucinations; Methods such as adaptor fine-tuning and fine-tuning the final layer of the decoder can also mitigate bias transmission to some extent.

3. Conclusion

This paper presents a thorough examination of automatic text summarization technology, highlighting its pivotal role within the realm of natural language processing and its extensive application across multiple fields. The origins of this technology can be traced back to the 1950s, when it began with rudimentary methods based on keyword frequency and sentence position. Over time, it has evolved significantly, embracing sophisticated algorithms that leverage deep learning and pre-trained models. The study underscores the importance of automatic text summarization in the current era of information overload. As the volume of data continues to grow exponentially, the ability to efficiently process and condense this information becomes increasingly crucial. The paper delves into specific techniques used in automatic summarization, with a particular focus on those that incorporate the improved TextRank algorithm and K-Means clustering. Four distinct automatic summarization algorithms are explored in detail: Lead-3, TextRank, MBM25EMB, and IMTextRank-K-Means. Lead-3 is a straightforward approach that selects the first two and last sentences of a text as the summary. TextRank, on the other hand, identifies important sentences by constructing a sentence connection graph and calculating TextRank values. MBM25EMB combines word embeddings, TF-IDF, and the BM25 model to create summaries. Lastly, IMTextRank-K-Means utilizes an enhanced BM25 model and TextRank to rank sentences, followed by K-Means clustering to generate diverse summaries. Experimental results demonstrate that IMTextRank-K-Means excels in terms of accuracy and reducing redundancy.

Furthermore, the paper delves into an important issue: name-nationality bias in pre-trained language models and its propagation in text summaries. By constructing the Wiki-nationality dataset and designing targeted experiments, the research analyzes the bias exhibited by models such as BART and PEGASUS during the summary generation process. The findings reveal that the biases inherent in pre-trained models are closely linked to the biases observed in the resulting text summaries. This underscores the need for ongoing research to address and mitigate these biases, ensuring that automatic text summarization technology remains fair and unbiased. In conclusion, this paper provides a comprehensive overview of the evolution, applications, and challenges facing automatic text summarization technology. It highlights the importance of this technology in managing the vast amounts of information available today and underscores the need for continued research to address biases and improve the accuracy and diversity of automatically generated summaries.

References

[1] Zhao, J. S., Zhu, Q. M., Zhou, G. D., Zhang, L. (2017). Review on Automatic Keyword Extraction. Journal of Software.

[2] Zhao, D., Mou, D.M, Bai, S. (2021). Deep Learning-based automatic extraction of structural elements from scientific and technological abstracts. Data Analysis and Knowledge Discovery.

[3] Shi, L., Ruan, X. M., Wei, R. B., Cheng, Y. (2019). A review of generative text Summarization based on sequence-to-sequence model. Information Science Journal.

[4] Zhao, H. (2020). A Review of Deep Learning methods for Generative Automatic Summarization. Journal of Information Science.

[5] Song, X. T., Sun, H. (2021). Based on neural network source code automatically in technology review. journal of software.

[6] Zheng, B. F., Yun, J., Liu, L. M., Jiao, L. (2023). A Review of Cross-language Summarization Methods, Journal of Computer.

[7] Liu, C., Sun, Y. Y., Yu, B., Wang, H., Peng, C., Hou, M. S., Guo, H., Wang, H. (2024). Automatic Text Summarization method based on improved TextRank algorithm and K-Means Clustering. Knowledge-Based Systems.

[8] Faisal, L., Esin, D., Mirac, S., Zhang, T., Dan, J., Kathleen, M., Tatsunori, H. (2023). When does pre-training bias propagate to downstream tasks? Case Studies in text Summaries.