Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230251

Text summarization quality detection based on GPT-3

Sihan Qi[1,3,†], Honghao Zhang [2,†]

[1]Department of computer science, Harbin University of Science and Technology, 150080, China
[2]Software engineering, Nanchang Hangkong University, 330063, China

[3]qisihan2021@gmail.com
[†]These authors contributed equally.

Abstract. To summarize lengthy text in a concise and accurate manner is called text summarization. With the increasing amount of information available, text summarization has become increasingly important in information retrieval, knowledge management, and sentiment analysis. The research on text summarization dates to the 1960s, and the methods have evolved from traditional template-based generation to statistical and neural network-based methods. Modern language model GPT-3 has demonstrated outstanding linguistic ability, including the ability to produce text that is cohesive and grammatically correct. However, evaluating the caliber of text produced by GPT-3 is challenging and requires careful evaluation criteria. This study evaluated the text summarization ability of GPT-3 using multiple evaluation models (ROUGE, BLEU, and CIDER), and found that the generated summaries exhibited high quality and accuracy.

Keywords: text summarization, GPT-3, evaluation matrix, natural language processing, quality detection.

1. Introduction

People continue to produce a lot of text material as the digital age progresses, including news articles, product reviews, medical literature, etc. Text summarization has impacted information retrieval, knowledge management, and many other significant academic domains due to the issue of having too much information [1]. Text summarization relies on the accuracy of the information to automatically synthesize brief language that captures the essence of the original text [2]. People need a massive amount of information to find what they need quickly, so text summarization is extremely important.

The quality and accuracy of the generated summaries directly determine its (text summarization) efficiency in the corresponding field. The popularity of smart gadgets and the burgeoning Internet age have led to a boom in the number of channels available for getting text information. To get enough text in a short amount of time, whether in the field of news or in popular sectors like knowledge graphs and sentiment analysis, the summary of text information is especially crucial. Text summarization plays an irreplaceable role. The quality and accuracy of the summary directly affect the user's control over the directionality and knowledge of the information in it. Suppose the quality and accuracy of the generated summary are low. In that case, it will easily lead to misreading by users. On the contrary, it can greatly save time and improve the user's reading efficiency of the text.

DOI: 10.54254/2755-2721/8/20230251

The study of text summarization dates back to the 1960s. Text summary technology is advancing along with the ongoing advancements in computer technology. Most early text summarization methods used template-based generation methods, first proposed by Karen Spärck Jones in the 1970s [3]. This method takes text templates as input and uses the format of text templates to generate summaries. In light of the developments in machine learning and NLP techniques, the text summarization method is gradually changing from the conventional template approach to the generation approach based on statistics and neural networks [4].

GPT-3 represents a significant advancement in natural language processing due to its vast size and impressive linguistic ability. Its massive training dataset allows it to generate high-quality text in various fields, including summarization [5]. One of the most significant advantages of GPT-3 is its ability to learn from a vast range of sources, making it highly adaptable to new contexts and domains.

The ability of GPT-3 to produce content that is both grammatically and coherently correct and nearly identical to writing produced by humans is another essential characteristic of the program. GPT-3 is capable of doing challenging tasks with this degree of linguistic proficiency, including creative writing, question-answering, and machine translation.

However, it is challenging to assess the quality of text produced by GPT-3, and the evaluation criteria must be carefully designed to ensure accurate results [6]. The accuracy and coherence of the generated text can be evaluated using human evaluators, although this process might take time and be subjective.GPT-3 nevertheless has the power to completely alter the NLP field. As such, the continued development and evaluation of GPT-3 and other language models are essential to ensure their continued usefulness and effectiveness.

This research used two datasets, one being the English Multi-News dataset and the other being the Chinese LCSTS dataset. The main aim of the study was to evaluate the text summarization performance of GPT-3 utilizing multidimensional evaluation measures including ROUGE, BLEU, and CIDER to grade the precision and quality of the dataset generated for text summarization.

Experimental results show that GPT-3 performs well in multidimensional evaluation of text summarization quality, generating highly accurate and high-quality text summaries in both Chinese and English datasets according to ROUGE and BLEU quality detection models. However, due to the different evaluation criteria of the CIDER model, which places more emphasis on semantic relevance and diversity of summary sentences and assigns higher weights to more informative, rare vocabulary and phrases, GPT-3's performance on the English dataset is not very satisfactory. Nevertheless, GPT-3 still maintains a high level of quality in text summarization on the Chinese dataset. In summary, GPT-3 performs well in terms of text summarization quality and accuracy.

2. Methods

A deep learning-based language model called GPT3 can learn and produce native language text automatically. The model was produced by OpenAI, one of the most successful companies in the NLP field and widely considered as one of the most cutting-edge language models. The complete name of GPT-3, a deep learning model built on the Transformer architecture, is "Generative Pre-trained Transformer 3". The model employs a method known as "pre-training," which uses enormous amounts of text data to train the model extensively before handling the job. This allows the model to perform well on unknown tasks because it already deeply understands natural language.

The GPT-3 model can perform various natural language processing tasks, including automatic summarization, text classification, translation, dialogue generation, question answering, and more. These tasks are realized based on the capacity of the model to comprehend and produce natural language. The GPT-3 model can automatically complete basic tasks, such as automatic summarization Based on the model's capacity for comprehending and producing natural language, these duties are accomplished. and text classification [7]. For example, when given a news article, the model can automatically extract the key information and generate a summary. Similarly, when given a piece of text, the model can automatically classify it, identifying which type of text it belongs to, such as news, technology, entertainment, etc. In addition to basic tasks, GPT-3 can also generate natural language text. For example,

DOI: 10.54254/2755-2721/8/20230251

in the dialogue generation task, the model can generate answers based on the user's input, making the dialogue fluent and natural. In the question-answering task, the model can generate answers based on the user's questions to help the user solve the problem.

3. Evaluation matrix

By analyzing the co-occurrence data of n-grams in the abstract, Chin-Yew Lin presented the automatic abstract evaluation method known as **ROUGE.** The fundamental concept is to calculate the quantity of overlapping basic units by comparing the automatically generated summary produced by the system with the manually generated standard summary [8]. It seeks to increase the evaluation system's stability and robustness. Numerous assessment techniques are included in the ROUGE index, such as ROUGE-N (N=1, 2, 3, 4) (W/S/SU, etc.) However, its evaluation only relies on the superficial similarity between peers and model summaries. It cannot fairly evaluate abstracts, including vocabulary changes and paraphrasing [9]. Due to the minimum lexical overlap, surface-based evaluation indicators such as ROUGE cannot capture the similarity and are often used to generate abstracts for long texts composed of multiple sentences or paragraphs. ROUGE and BLEU calculate textual similarity based on n-grams, but the former focuses on recall rather than precision.

Different rouge indicators have different calculation methods and applicable scenarios. For jobs requiring a single document summary, ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S are better options. ROUGE-1, ROUGE-2, ROUGE-S4, and ROUGE-S9 all perform well in multi-text summarization tasks when stop words are removed from the comparison.

Different granularity has an impact on the accuracy of ROUGE. From the word granularity analysis, Rouge calculation can better assess the degree of fitting of the model to proper nouns. However, this evaluation metric is greatly affected by the word segmentation results. Word segmentation errors will evaluate errors. It also reduces the importance of long words, and the model tends to fit short words that are easy to predict. It is unsuitable for long text generation.

BLEU was initially used in machine translation to evaluate the similarity between the target and source texts in machine translation. It is obtained from the precision number for the weighted n-gram. Recall is not used in the procedure; just accuracy is. It is better suited for analyzing short text generation tasks rather than long text scenarios since it struggles to assess the correlation in context understanding when employed in text generation [10].

CIDER is an evaluation index that can be utilized for related text generation activities as well as image description tasks. It assigns various n-grams varying weights using TF-IDF. It seems sense that words that occur more frequently have a lower weight than those that occur less frequently [11]. Treat each sentence as a document to start before comparing them to find similarities. Next, find the TF-IDF vector's cosine angle. The last step is to average the similarity of n-grams of varied lengths. With various TF-IDFs, different n-grams have varying weights. However, the more prevalent n-grams in the entire corpus convey less information, therefore decentralized non-keyword actions are required.

4. Experimental results and analysis

4.1. Dataset description

Multi-News dataset. The Multi-News dataset is an English dataset containing summaries of multiple news articles. Researchers at Carnegie Mellon University, Yale University, and Rice University jointly developed it. The dataset contains over 56,000 samples, each comprising 5-10 news articles and one or more abstracts. These articles and summaries come from reputable news sources like CNN, BBC, and The New York Times. The purpose of the Multi-News dataset is to provide a challenging benchmark for automatic summarization and multi-document summarization tasks. It can be used to evaluate the performance and effect of text summarization algorithms, and it can also be used to train and test automated summarization models. The Multi-News dataset is free to download and has been manually annotated and verified. Due to its size and quality, it has become one of the important datasets in automated text summarization.

DOI: 10.54254/2755-2721/8/20230251

LCSTS summary data set. The LCSTS summary data set is organized by the Harbin Institute of Technology. The data set is created based on news summaries published by news media on Weibo [13]. The data collection includes 2 million actual short Chinese text records as well as a summary provided by each text author. At the same time, the abstracts of 10666 texts were manually annotated by the collators. About 100 characters for each essay and about 20 characters for each abstract. The LCSTS summarization dataset was created to promote research on Chinese text summarization. It can train and evaluate automatic text summarization algorithms, such as those based on machine learning and deep learning. The feature of the LCSTS dataset is that it simplifies the text for training and evaluating summarization algorithms.

4.2. Results analysis

ROUGE Evaluation Results. This study bases its multidimensional evaluation of the generated text summary dataset on the GPT-3's text summary function, using ROUGE methodologies. Tables 1 and 2 display the evaluation findings of the data.

Table 1. English dataset.

Metrics	Recall	Precision	F Score
rouge-1	0.092	0.193	0.122
rouge-2	0.008	0.019	0.011
rouge-l	0.070	0.158	0.094

Table 2. Chinese dataset.

Metrics	Recall	Precision	F Score
rouge-1	0.482	0.144	0.217
rouge-2	0.193	0.047	0.074
rouge-l	0.426	0.108	0.169

The ROUGE-1 recall score for the English dataset was 0.092 with a precision score of 0.193, while the Chinese dataset had a higher recall score of 0.482 and precision score of 0.144, resulting in F-scores of 0.122 and 0.217 respectively. The higher ROUGE-1 recall score for the Chinese dataset may be attributed to the shorter length of the reference summaries, which contain only about 20 words.

This suggests that the performance of a text summarization model can be influenced by the complexity and length of the original texts and their corresponding reference summaries used in the datasets. Specifically, when the original texts and reference summaries are shorter, It may be easier for the model to capture key information from the text so that it can generate a summary that more closely matches the reference summary, resulting in a higher ROUGE-1 recall score. The performance of a text summarization model can be affected by several variables, including the extent of the dataset, the length of the text, and the complexity of the language. It is crucial to take these variables into account when interpreting the evaluation findings. The ROUGE evaluation results demonstrate the high quality and accuracy of the summaries generated by GPT-3 in both Chinese and English datasets.

BLEU Evaluation Results. The average number of terms in the generated summary that also appeared in the reference text for the English dataset was 13.61, or 13.61 percent. For the Chinese dataset, the BLEU score was 11.33. These scores indicate a moderate level of performance for a text summarization model. The reason for the slightly lower BLEU scores in the Chinese dataset compared to the English dataset is that spaces, like in English, do not separate Chinese words. Therefore, before the evaluation, the Chinese text was segmented using the third-party Python module jieba. However, the segmentation results may impact the evaluation of the BLEU metric. In summary, the segmentation of Chinese text before conducting the evaluation might affect the BLEU scores, which could be one of the

DOI: 10.54254/2755-2721/8/20230251

reasons for the lower BLEU scores in the Chinese dataset compared to the English dataset. This does not affect the demonstration of the high quality and accuracy of the summaries generated by GPT-3 in both Chinese and English datasets.

CIDER Evaluation Results. CIDER score for the English dataset is very low, at only 0.00028937172126072683. This indicates a low correlation level between the generated summary and the reference text, and further optimization of the model's performance is needed. In contrast, the CIDER score for the Chinese dataset is much higher at 0.009093958439389904, indicating that the quality of summary generation is good. However, there is room for improvement in relevance to the reference text. The low CIDER score may be due to its different evaluation standards. CIDER also considers summary sentences' semantic relevance and diversity and assigns higher weights to more informative and rare vocabulary and phrases. This leads to significantly lower scores compared to other metrics.

5. Conclusion

GPT-3 has demonstrated impressive language processing abilities in evaluations, with its generated text displaying remarkable coherence and grammatical correctness. This study used the ROUGE, BLEU, and CIDER text summarization quality evaluation models to assess GPT-3's text summarization capabilities on two datasets, Multi-News and LCSTS. The results show that GPT-3 performs well in terms of quality and accuracy in text summarization, with good evaluation scores for ROUGE-1 recall, precision, F-score, and BLEU score in both Chinese and English datasets. Although GPT-3's CIDER score on the English dataset is not impressive, its performance on the Chinese dataset still has a high rating. Taking into account the possible differences in evaluation criteria, overall, GPT-3 has shown good performance in the quality and accuracy of text summarization. In the short term, this study confirms the excellence of GPT-3 in text summarization, alleviating concerns for potential users. In the long term, continuous development and evaluation of language models like GPT-3 are crucial in ensuring their usefulness and effectiveness in various natural language processing tasks and providing a valuable reference for related research.

References

[1] Soumya, S., Kumar, G.S., Naseem, R., Mohan, S. (2011). Automatic Text Summarization. In: Das, V.V., Thankachan, N. (eds) Computational Intelligence and Information Technology. CIIT 2011. Communications in Computer and Information Science, vol 250. Springer, Berlin, Heidelberg.

[2] Panchal, R., Pagarkar, A., Kurup, L. (2019).C. In: Kulkarni, A., Satapathy, S., Kang, T., Kashan, A. (eds) Proceedings of the 2nd International Conference on Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing, vol 828. Springer, Singapore.

[3] Mandal, S., Singh, G.K., Pal, A. (2019). PSO-Based Text Summarization Approach Using Sentiment Analysis. In: Iyer, B., Nalbalwar, S., Pathak, N. (eds) Computing, Communication and Signal Processing . Advances in Intelligent Systems and Computing, vol 810. Springer, Singapore.

[4] Steinberger, J., Ježek, K. (2009). Text Summarization: An Old Challenge and New Approaches. In: Abraham, A., Hassanien, AE., de Leon F. de Carvalho, A.P., Snášel, V. (eds) Foundations of Computational, IntelligenceVolume 6. Studies in Computational Intelligence, vol 206. Springer, Berlin, Heidelberg.

[5] Lucy, L., & Bamman, D. (2021, June). Gender and representation bias in GPT-3 generated stories. In Proceedings of the Third Workshop on Narrative Understanding (pp. 48-55).

[6] Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. Fundamental Research, 1(6), 831-833.

[7] Dale, R. (2021). GPT-3: What's it good for? Natural Language Engineering, 27(1), 113-118. doi:10.1017/S1351324920000601

[8] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text

DOI: 10.54254/2755-2721/8/20230251

summarization branches out (pp. 74-81).

[9] Song, X., Yang, C., Zhang, H., Zhao, X. (2018). The Algorithm of Automatic Text Summarization Based on Network Representation Learning. In: Zhang, M., Ng, V., Zhao, D., Li, S., Zan, H. (eds) Natural Language Processing and Chinese Computing. NLPCC 2018. Lecture Notes in Computer Science(), vol 11109. Springer, Cham.

[10] Lu Yuxuan, Sun Yueying. A method for automatic marking subjective mathematical questions based on BLEU [J]. Management Observation, 2019, No. 710 (03): 121-124

[11] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566-4575arXiv:1506.05865 [cs.CL]

[12] Li, Z., Peng, Z., Tang, S., Zhang, C., & Ma, H. (2020). Text summarization method based on double attention pointer network. IEEE Access, 8, 11279-11288.

[13] Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. The Stata Journal, 20(1), 131-148.