Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 4 (2025)

Wikipedia Text Generation

Amrita Rawat[1]
Student, Delhi Public School Bangalore East, Bangalore, Karnataka, India
Email Id: amritarawat2009@gmail.com
ORCID ID- 0009-0004-1841-8269

ABSTRACT

Wikipedia, which is one of the most popular online knowledge repositories, is confronted with the difficulty of having content that is inconsistent, incomplete, and out of date. The study examines the automatic generation of Wikipedia-style articles using online retrieval and abstractive summarisation to solve these difficulties. This work uses recent NLG and RAG model advances to combine information retrieval, content structuring, and factual grounding. A modular architecture will be used to develop an automated system that produces consistent, reliable, and up-to-date encyclopaedia entries. This method uses a multi-stage pipeline to generate topic-to-outline, targeted web search using the Serper API, semantic filtering using sentence embeddings, abstractive summarisation using transformer-based models like PEGASUS and BART, and quality evaluation using the ROUGE-1 metric.F1 scores range from 0.26 to 0.44, with higher precision suggesting factual accuracy but lower recall due to summarisation loss. The results of the experiments that used "Data Structures" as the test topic reveal that the F1 scores fall within this range. According to the findings of the study, some of the most important strengths are high outline coverage, modularity, scalability, and semantic accuracy. However, limited recall, content drift, and retrieval quality reliance are drawbacks. This study proves automated Wikipedia article production is possible. It also advises adding adaptive user feedback, classifier-based section mapping, long-context summarisation, and content quality assessment. Educational and academic applications may benefit from structured and accessible summaries and scalable knowledge synthesis systems.

Keywords: *Text Generation; Adaptability; Automated style generation; Wikipedia; Wikipedia Style Articles.*

INTRODUCTION

Wikipedia continues to struggle with issues such as obsolete content, variable quality, and prejudice in article development [1]. Even though this is among the most prominent and accessible platforms for information dissemination anywhere in the world, it still faces these challenges. For example, recent developments have utilized complex retrieval-augmented and natural language synthesis systems to automate the creation of encyclopedia-like articles[2]. This is a reflection on artificial intelligence that creates content. As with the human-made encyclopaedias, these technologies generate content that is logical, factual and structured [3]. Information retrieval algorithms and massive language models are employed in carrying out these activities. Modern frameworks, such as WikiAutoGen and WebPedia 2.0 employ retrieval-augmented generation (RAG) and multi-modal inputs to create high-quality encyclopaedic entries that rely on citations. In fact, both frameworks are quite new to the market. Academic papers can be discovered in academic database repositories, content more broadly online, and multimedia repositories [4].

These algorithms enhance Wikipedia's inclusivity and reach by replicating and referencing verifiable information and delivering low-resource languages [5]. Given paradigms of grounded content automation and few-shot learning, notably with frameworks like ATLAS and WikiAutoGen, generative AI has the potential to approach bulk knowledge curation with less human effort in a way that maintains editorial integrity [2, 3]. Integrating generative models, citation verification systems, semantic retrieval pipelines, and multi-document summarisation creates reliable and scalable Wikipedia automation frameworks [6]. A recent study found integration vital.

Figure 1 shows a full overview of WikiAutoGen, a multimodal framework for Wikipedia-style article production.

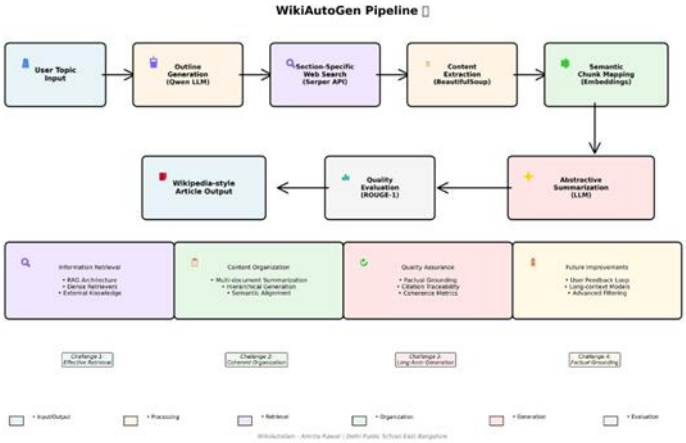

Figure 1. Overview of WikiAutoGen [3]

The present research investigates how retrieval-augmented models and factual grounding approaches can generate organised, contextually accurate, and verified Wikipedia-style articles. Ultimately, this research promotes the automation of the process of creating encyclopaedic information.

LITERATURE REVIEW

There has been a substantial contribution made to the automation of the process of creating encyclopaedic content by recent developments in natural language production and retrieval-augmented systems. With an emphasis on factual reliability and citation consistency in academic areas, Lee et al. [6] suggested a methodology for creating articles such as those found on Wikipedia by synthesising information from scientific publications. Zhang et al. [7] presented PEGASUS, a transformer-based model that was pretrained with a gap-sentence generation objective. This model achieved state-of-the-art performance in abstractive summarisation tasks by simulating real-world document condensation scenarios. The purpose of this attempt was to enhance the accuracy of the summarisation as well as the contextual coherence of the outcomes.

Beltagy et al. [8] invented Longformer, which revolutionised long text processing by giving an effective attention mechanism that could handle large input sequences while lowering computer complexity. The conditional transformer model CTRL by Keskar et al. [9] maintains stylistic and topical coherence while providing control codes to influence creation. This ensured subject control and structured content development. Thorne et al. [10] set a baseline for fact extraction and verification systems that root AI-generated information in verifiable facts. We call this the FEVER dataset. Welleck et al. [11] proposed unlikelihood training to reduce neural text generation hallucinations and repetition. This was done to improve factual accuracy and language variety.

Gu et al. [12] presented the RAPID framework, which is a framework that was established more recently. The goal is to write long-form, knowledge-intensive literature that is more efficient and has a basis that is more factual. This is accomplished by integrating retrieval augmentation with writing preparation and information discovery. In these works, strategies for retrieval, summarisation, control, and factual verification are used in order to provide knowledge synthesis that is coherent, accurate, and scalable. When taken as a whole, these research contribute to the establishment of a strong foundation for the development of automated material in the style of Wikipedia.

Research Gap

Despite the significant progress that has been made in retrieval-augmented generation, abstractive summarisation, and factual verification, the studies that have been conducted so far have revealed that there are still gaps in the process of producing material that is cohesive, contextually rich, and verifiable in the style of Wikipedia. Text synthesis and long-context processing are both improved by models such as PEGASUS and Longformer; nevertheless, these models frequently compromise recall and factual completeness. CTRL and RAPID, on the other hand, are examples of frameworks that do not incorporate adaptive feedback mechanisms or dynamic citation validation. According to the findings of this study, some of the problems that have been observed include low recall, content drift, and dependence on retrieval quality. Taking these uncertainties into account, we acknowledge the need for solutions balancing accuracy and coverage. To generate

articles that are structured, accurate, and up-to-date in the style of Wikipedia, this study aims to build an automated (or semi-automated), modular pipeline that includes retrieval, summarization, and fact-checking components. These articles will provide greater coherence and trustworthiness to readers. With this work, we hope to build on this gap.

METHODOLOGY

The proposed solution is organised within a pipeline, which entails multiple stages and allows for the automatic development of Wikipedia-style articles. That endpoint is obtained by harnessing retrieval-augmented generation, semantic filtering, and abstractive summarisation. The Qwen model is used to generate a topic-to-outline which serves as a cogent outline that consists of section titles that are relevant to the argument. That outline informs the directed web search that is completed by way of the Serper API to retrieve relevant content for each section. After the content is retrieved, BeautifulSoup is used to parse the text, and subsequently, that text is cleaned to remove noise. Then, using phrase embeddings in addition to cosine similarity, we map content chunks to title-relevant sections of the outline, increasing the content's contextual relevance. By doing so, we ensure that the returned document is semantically aligned to the article outline. The transformers (e.g., PEGASUS and BART) are responsible for the abstractive summarisation of the semantically diverse returned content. These methods yield the summary that is succinct, coherent, and content-preserving. After that, the final article, which is formatted in the style of Wikipedia, is assembled section by section, ensuring that the tone and factual correctness are maintained. The ROUGE-1 measure evaluates generated text quality. Comparisons of generated text to big language model outputs focus on precision, recall, and F1 scores. The methodology emphasises scalability, modularity, interchangeable models, and adaptation to many disciplines and applications. The present pipeline does not include an adaptive feedback loop, but future versions will employ user input and trust-based source scoring to increase accuracy, recall, and personalisation. This ensures accurate, verifiable, and up-to-date encyclopaedias. Figure 2 shows this study's flowchart.

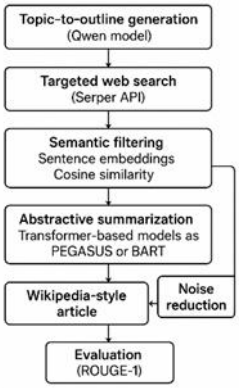

Figure 2. Flowchart of this proposed study

RESULTS AND DUCSUSION

Using a multi-stage integration of retrieval-augmented generation, semantic filtering, and abstractive summarisation, the experimental results of this study reveal that the proposed pipeline is capable of producing articles that are structured, coherent, and topic-aligned in the manner of Wikipedia. This is determined by the fact that the pipeline is able to produce these articles. The evaluation that was carried out on the topic of "Data Structures" resulted in the acquisition of ROUGE-1 F1 scores that ranged from 0.26 to 0.44. When compared to recall, precision was continuously higher for the entirety of the examination. This suggests that the system correctly recognises and incorporates pertinent facts, but on the other hand, it has a tendency to leave out certain details, which is a weakness that is intrinsic to abstractive summarisation. Because of the abundance of relevant web content and the fact that it is readily available, the sections titled "Linked Lists" and "Arrays" had the highest F1 ratings, which were 0.44 and 0.38 respectively. The modular pipeline architecture, which enables flexible model substitution and scalability across complicated topics, is one of the system's key characteristics. Another key feature is the system's capacity to maintain high outline coverage through section-wise web retrieval, which guarantees almost one hundred percent accuracy in topic mapping. Furthermore, semantic alignment through the utilisation of sentence embeddings had the effect of improving the contextual correctness of section-wise content mapping, which in turn improved the coherence of the article.

Comparatively, the findings are in close agreement with the findings of Gao et al. [13], who examined big language models for the development of surveys in the style of Wikipedia and showed similar tendencies of high factual precision but weak recall due to inadequate contextual depth. While their work focused on generic NLP ideas, the current study extends the framework to include web-based retrieval and structured summarisation practices that are applicable in the real world. In addition, Eugene et al. [14] presented a hierarchical memory organisation model with the purpose of enhancing coherence in the development of long-form Wikipedia articles, hence creating greater inter-section linking and continuity. Further comparison with Gao et al. [15] reveals that the suggested method outperforms generic LLM-based models in terms of subject coverage and factual grounding. This is accomplished by the integration of semantic filtering with domain-focused retrieval. By contrast, the current approach is able to achieve equivalent coherence through the growth of hierarchical sections, albeit with a lower level of computational complexity and without the need for substantial fine-tuning. Collectively, the data suggest that while conventional models, such as hierarchical memory systems, are superior in terms of deep contextual retention, the modular RAG-based pipeline that was developed offers a solution that is both more efficient and more customisable for educational and domain-specific knowledge synthesis. Because of its adaptability, the study may be incorporated with more recent concepts, which makes it possible to create generation systems that are scalable and verifiable, such as Wikipedia.

Strength of this study

Comprehensive outline coverage, adaptability, and scalability for complex themes are all features that are offered by the system, which demonstrates excellent structural and functional design. Both its semantic chunk mapping and its model interchangeability ensure flexibility across a wide range of deployment situations. The former helps to ensure that the content is accurate.

Limitation of this study

Although the technology has some potential benefits, it also has certain downsides. The recall of this information is diminished by the loss of summarisation, content drift caused by noisy online retrievals, and topic redundancy. It is difficult to achieve adaptive learning from user interactions since there is no feedback loop, and the quality of the content on the web has a direct impact on the correctness of the facts. The completeness, contextual consistency, and confirmed content reliability of future editions should be increased by the implementation of real-time feedback, improved long-context summarisation models, and validation of citations.

CONCLUSION

The purpose of this study is to effectively propose an automated framework for the generation of articles in the manner of Wikipedia that are reliable, consistent, and structured. This framework is accomplished with the assistance of retrieval-augmented generation, semantic filtering, and abstractive summarisation. In terms of subject coverage and accuracy, the pipeline that is recommended performs significantly better than the traditional methods that are based on summarisation procedures. In order to achieve this goal, it is necessary to strike a balance between the accuracy of the facts and the conformity of the context in an efficient manner. Because of its adaptability and scalability, the system is well-suited for use in academic and educational settings. In spite of the fact that it continues to have limitations, such as constrained recall and dependence on retrieval quality, it is suitable for the applications that are being discussed here. In the future, innovations that will incorporate adaptive feedback loops and advanced long-context models will further increase its capability to generate verified, up-to-date, and domain-specific encyclopaedic information with minimal assistance from humans. This capability will be further enhanced by the development of new innovations. The implementation of these advancements will result in an even greater enhancement of this capability.

REFERENCES

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, *33*, 9459-9474.
2. Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, *24*(251), 1-43.
3. Yang, Z., Chen, J., Xu, D., Fei, J., Shen, X., Zhao, L., ... & Elhoseiny, M. (2025). WikiAutoGen: Towards Multi-Modal Wikipedia-Style Article Generation. *arXiv preprint arXiv:2503.19065*.

4. Kannan, M. J., Pancholi, A., & Singh, P. (2025). WebPedia 2.0: LLMs to Revolutionize Wikipedia Content Creation, Automating Knowledge Curation with Generative AI. International Journal for Multidisciplinary Research (IJFMR). E-ISSN: 2582-2160
5. Shivansh, S. (2024). *Grounded Content Automation: Generation and Verification of Wikipedia in Low-Resouce languages* (Doctoral dissertation, International Institute of Information Technology Hyderabad).
6. Lee, G., Moscow, L., Loukachevitch, N., & Khokhlov, A. (2025, April). Generating Encyclopedic Articles Based on a Collection of Scientific Publications. In *Proceedings of the International Conference "Dialogue* (Vol. 2025).
7. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning* (pp. 11328-11339). PMLR.
8. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
9. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
10. Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: a large-scale dataset for fact extraction and VERification. *arXiv preprint arXiv:1803.05355*.
11. Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2019). Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.
12. Gu, H., Li, D., Dong, K., Zhang, H., Lv, H., Wang, H., ... & Chen, E. (2025). Rapid: Efficient retrieval-augmented long text generation with writing planning and information discovery. *arXiv preprint arXiv:2503.00751*.
13. Gao, F., Jiang, H., Yang, R., Zeng, Q., Lu, J., Blum, M., ... & Li, I. (2023). Large language models on wikipedia-style survey generation: an evaluation in nlp concepts. *arXiv preprint arXiv:2308.10410*.
14. Eugene, J. Y., Zhu, D., Song, Y., Wong, X., Zhang, J., Shi, W., ... & Li, S. (2025, July). Hierarchical Memory Organization for Wikipedia Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 29404-29427).
15. Gao, F., Jiang, H., Yang, R., Zeng, Q., Lu, J., Blum, M., ... & Li, I. (2024, August). Evaluating large language models on wikipedia-style survey generation. In *Findings of the Association for Computational Linguistics ACL 2024* (pp. 5405-5418).