Automated Summarization of Similar Articles Using LLM

Michal Chudoba

Seznam.cz, Radlická 3294/10, Prague, 150 00, Czech Republic
www.seznam.cz
michal.chudoba@firma.seznam.cz

Abstract. This paper presents an automated LLM-driven pipeline for summarizing clusters of related news articles, designed to enhance user engagement by providing concise, accessible summaries of similar content. By leveraging targeted prompting and re-prompting techniques, we refine LLM outputs, ensuring the quality and coherence of summaries. This pipeline offers a scalable solution for managing and summarizing extensive online news databases.

Keywords: summarization, clustering, LLM as a judge, automated evaluation

1 Introduction

This work introduces an LLM-powered solution for automatically summarizing clusters of similar news articles in our content database. The goal is to enhance user experience by providing concise and informative summaries of related content.
Our approach involves three key steps:

- **Article Clustering:** News articles are grouped together based on their semantic similarity, creating clusters of related content.
- **Cluster Summarization:** An LLM is employed to generate a summary of each cluster, capturing the main points from the article cluster.
- **Summary Evaluation and Re-prompting:** The quality of the generated summaries is assessed using a combination of metrics and an external LLM judge. Summaries that fall below a certain quality threshold are re-prompted with feedback to improve their accuracy and coherence.

This automated summarization pipeline enables the effective construction of new severable content that could be presented to the user with references to the original news articles.

2 Article Clustering

We use fastText [1] embeddings to compute semantic similarity, creating a graph structure where each article is a node. Edges represent similarity above a threshold, ensuring that clustered articles are meaningfully related. Maximal clique search is used to identify tightly connected subgroups, ensuring that each cluster contains articles with high mutual similarity, which is critical for generating coherent summaries.

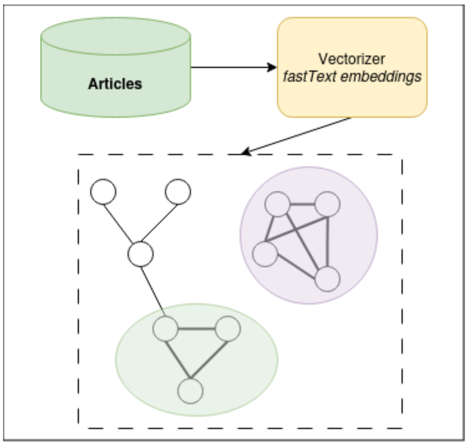

Fig. 1. Sketch of article clustering

Figure 1 visualizes the process with how the articles are clustered.

3 Cluster Summarization

Once articles have been clustered, the next step is to generate concise and informative summaries for each cluster. This is achieved through Cluster Summarization.
Our approach to cluster summarization involves the following steps:

- **LLM summary:** We utilize an LLM for summarization by providing it with the articles within the cluster. We constrain the model's output to given length and subsequently extract title and summary for the whole cluster
- **Summary Evaluation:** summarizations are evaluated using metrics of length, cosine similarity, clickbait score and grammatical correctness.
- **Re-prompting:** Underperforming summaries are re-prompted with explanation given to the LLM as on which metrics the summaries underpeformed.

The metrics were chosen based on the crucial parameters of being truthful and correct. Cosine similarity measures the semantic alignment of the summary with the cluster content and thus ensures truthfulness, while the clickbait score helps ensure that summaries are informative rather than sensational. Grammatical correctness was also measured as the LLM struggled when the original news articles contained errors themselves. This was done using the Small-e-Czech [2] developed internally at Seznam.cz.

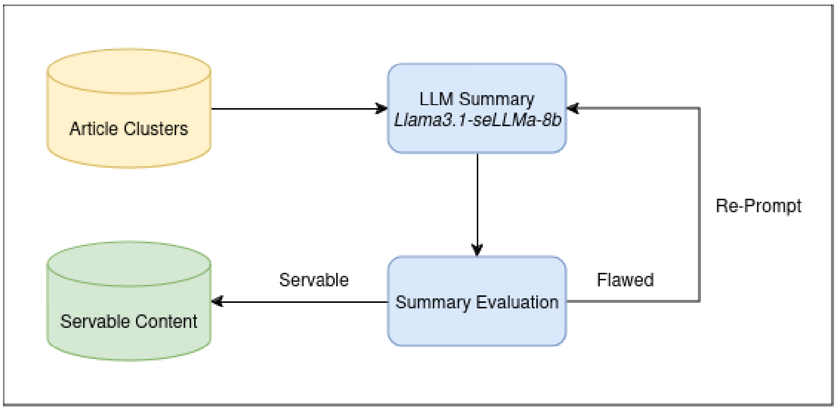

Fig. 2. Summarization pipeline with re-prompting.

4 LLM as a Judge

To ensure the quality of the generated summaries, we employ a LLM as a judge [4] as a final filter before the summaries are ruled as safe to serve to users. We have selected GPT-4o [3] as our judge model and followed the 1 to 4 grade scale with examples and explanations on what is considered a valid summary. By using the GPT-4o model as a final quality check, we aim to leverage its understanding of summarization and language comprehension to our advantage without needing to judge summaries that are already deemed faulty.

5 Results

We have utilized human annotators to validate the quality of the summarizations. For simplicity, the annotators were only tasked with determining if the given summary is severable based on the criteria of semantic accuracy, not being clickbait, grammatical accuracy, and length. This mirrors product-given thresholds for these metrics as employed in the summarization pipeline.

	Cosine Similarity	**Grammar**	**Clickbait**	**LLM Judge**	**Annotator**
Cosine Similarity	X	0.07	−0.41	0.45	0.47
Grammar	0.07	X	0.01	0.11	−0.02
Clickbait	−0.41	0.01	X	−0.32	−0.29
LLM Judge	0.45	0.11	−0.32	X	0.43
Annotator	0.47	−0.02	−0.29	0.43	X

Table 1. Measured spearman correlation from 152 cluster summaries, covering over 650 articles.

In Table 5 we showcase our positive correlation between Cosine Similarity and LLM Judge Score, while having a negative correlation between Clickbait Score and LLM Judge Score. Correlation with Grammar is low, as the summaries were almost always without grammatical errors. We were delighted to observe a non-trivial positive correlation between LLM Judge Score and Annotator Score, as it validates our approach.

6 Main Contributions

Our work demonstrates the feasibility of a fully automated summarization pipeline that not only generates accurate and readable summaries but also automatically detects and rectifies quality issues. The alignment of our metrics with human judgment offers a promising foundation for deploying scalable content summarization solutions.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
2. Kocián, M., Náplava, J., Štancl, D., Kadlec, V.: Siamese bert-based model for web search relevance ranking evaluated on a new czech dataset (2021), https://arxiv.org/abs/2112.01810
3. OpenAI: Hello gpt-4o (2024), https://openai.com/index/hello-gpt-4o/, accessed Oct 27 2024
4. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023), https://arxiv.org/abs/2306.05685