From Sparse to Dense
GPT-4 Summarization with Chain of Density Prompting

Tech Review by Hermann Rösch — hrosch2@illinois.edu CS410 Fall 2023

I. Introduction

In the paper titled "From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting," a deep dive is taken into the nuances and complexities of text summarization, an area where GPT-4, OpenAI's latest and most advanced language model, shines. The study's focal point is the concept of information density in summaries. It articulates a compelling argument: "An effective summary should inherently be denser, presenting a higher concentration of information from the source text, yet achieving this without sacrificing clarity or readability." (Adams et al., 2023). Here, the authors have unveiled an innovative technique termed "Chain of Density" (CoD) prompting. This technique stands out as it gradually weaves essential elements from the original text into a summary. Such an approach is not just about adding details; it's a delicate balancing act that enhances content richness while ensuring each summary remains digestible and maintains a consistent length.

This review will delve into the specifics of the Chain of Density method. We'll take a closer look at the evaluation methods employed by the researchers and discuss how their findings contribute to the broader landscape of machine-generated text summarization. The comparison of the CoD approach with traditional summarization techniques is integral to understanding the advancements this paper introduces and their potential influence on the future of natural language processing.

II. Chain of Density Prompting

The study introduces the Chain of Density (CoD) prompting method to generate a set of summaries with GPT-4 with varying levels of information density while controlling for length, which has proven to be a strong confounder when evaluating summaries (Fabbri et al., 2021; Liu et al., 2023b). The method begins with generating an entity-sparse summary, concentrating on just a few salient entities. Subsequently, it employs an iterative process where additional entities identified as missing from the prior summary are seamlessly integrated without increasing the overall word count.

To maintain the same length while increasing the number of entities covered, abstraction, fusion, and compression are explicitly encouraged, rather than dropping meaningful content from previous summaries. Figure 2 in the paper displays the prompt and an example output, illustrating how each iteration refines the summary.

Regarding data, the researchers randomly sampled 100 articles from the CNN/DailyMail summarization test set. This dataset choice allows for a robust comparison of the CoD-generated summaries. For a frame of reference, the paper compares CoD summary statistics to human-written bullet-point style reference summaries and summaries generated by GPT-4 with a vanilla prompt: *"Write a VERY short summary of the Article. Do not exceed 70 words."* The desired token length is set to match that of CoD summaries, as shown in Table 1 of the paper. This comparative analysis is crucial as it provides a benchmark to assess the CoD method's effectiveness against human-written summaries and those generated by standard GPT-4 prompts.

III. Statistics

Statistical analysis offers solid evidence of the effectiveness of the CoD approach, going beyond just theoretical concepts. The analysis includes direct and indirect statistical measures that shed light on the nuances of summarization performance. It shows that the summaries generated by CoD become more abstract and exhibit better fusion over time. Additionally, they move away from the lead bias present in standard GPT-4 prompt summaries.

As outlined in the study, direct statistics include a meticulous adherence to a fixed token budget and a marked increase in entity density; notably, even with their brief nature, CoD summaries outshine human and vanilla GPT-4 generated summaries regarding information density. These summaries start with an entity density below that of human and vanilla GPT-4 summaries (0.089), but after five densification steps, they reach an impressive 0.167, surpassing their counterparts.

Indirect statistics showcase that CoD summaries are getting more sophisticated. Measured by extractive density and concept fusion, abstractness is consistently increasing; with each iteration, summaries capture more content and do so more abstractly, indicating the method's efficacy in improving summary quality.

Moreover, the shift in content distribution in CoD summaries, from a predominant lead bias to a more holistic representation of the entire article, is particularly striking; this transition, underscored by comparative data, signals a shift towards more nuanced and comprehensive summaries. These findings underscore CoD's potential to redefine the benchmarks for machine-generated text summarization, pointing towards a future where AI can distill complex narratives with greater balance and depth.

IV. Results

The investigation's depth is further realized through a detailed examination of the CoD summaries' tradeoffs, as human and machine evaluators perceive. A preference-based human study involving the paper's first four authors reveals insightful patterns: a general preference for summaries generated in the middle steps of densification. Notably, this preference aligns with the entity density of human-written summaries. However, a low Fleiss' kappa score of 0.112 reflects the subjective nature of summary evaluation. Table 2 in the paper provides a breakdown of first-place votes, indicating a modal preference for step 2, with an aggregate expected preferred step around 3.06.

Parallel to human insights, GPT-4's automatic evaluation rating summaries on informativeness, quality, and coherence echo this preference for mid-range densification. Table 3 illustrates that while informativeness peaks at Step 4, qualities like quality and coherence decline after Steps 2 and 1, respectively; this nuanced evaluation by GPT-4 demonstrates the delicate balance between richness of detail and clarity, with intermediate steps often striking the most favorable balance.

Finally, the qualitative analysis underscores a tradeoff between coherence/readability and informativeness. Figure 4 exemplifies this with two CoD steps: one where adding details enhances the summary and another where it detracts from it; this balance is crucial in intermediate CoD steps, but the optimal equilibrium is yet to be determined and is left as an open area for future research.

V. Related work

Various studies have shed light on the evolving landscape of **GPT Summarization**. (Goyal et al. 2022) conducted a notable study benchmarking GPT-3 on news article summarization, uncovering a human preference for GPT-3's summaries over previously supervised baselines. Complementing this, (Zhang et al. 2023) discovered that zero-shot GPT-3 summaries could match the quality of human-written ones.

Regarding **Entity-Based Summarization**, (Narayan et al. 2021) introduced the concept of generating entity chains for the supervised fine-tuning of summarization models, offering a distinct contrast to the keyword-based methods detailed by (Li et al., 2020; Dou et al., 2021), as well as the purely extractive strategies highlighted by (Dou et al., 2021; Adams et al., 2023a).

Further enriching this field, entities have been increasingly utilized in various summarization strategies, ranging from their use as a control (Liu & Chen, 2021; He et al., 2022; Maddela et al., 2022) mechanism to enhancing faithfulness (Nan et al., 2021; Adams et al., 2022) and serving as a unit for evaluation (Cao et al., 2022; Adams et al., 2023b). This shift in methodology underscores the innovation of the CoD approach, particularly in its use of entity chains to modulate narrative density, marking a significant departure from more traditional summarization techniques.

VI. Conclusions and Limitations

The paper concludes by reflecting on the delicate interplay between densification and the intelligibility of summaries. It finds that while a degree of densification is preferred when summaries contain too many entities per token, readability and coherence become challenging to maintain (Adams et al., 2023); this conclusion underscores the core finding of the research: the value of moderation in densification to optimize information richness and readability. The authors have open-sourced an annotated test set as well as a larger unannotated training set, encouraging further research into fixed-length variable density summarization (Adams et al., 2023).

Even though the study was limited/focused on news summarization and the subjective nature of its annotations, the reliance on a closed-source model like GPT-4 creates transparency and replicability challenges. Nevertheless, the authors mitigate this limitation by sharing their evaluation data and encouraging the application of their methodology to open-source models, demonstrating a commitment to the collective advancement of the field (Adams et al., 2023).

VII. References

Adams, G., Fabbri, A., Ladhak, F., Lehman, E., & Elhadad, N. (2023). From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting. arXiv preprint arXiv:2309.04269. https://doi.org/10.48550/arXiv.2309.04269

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391-409. https://doi.org/10.1162/tacl_a_00373

Liu, Y., Fabbri, A., Liu, P., Zhao, Y., Nan, L., Han, R., Han, S., Joty, S., Wu, C., Xiong, C., & Radev, D. (2023b). Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4140–4170, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.228

Goyal, T., Li, J. J., & Durrett, G. (2022). News Summarization and Evaluation in the Era of GPT-3. arXiv preprint arXiv:2209.12356. https://doi.org/10.48550/arXiv.2209.12356

Narayan, S., Zhao, Y., Maynez, J., Simões, G., Nikolaev, V., McDonald, R. (2021). Planning with Learned Entity Prompts for Abstractive Summarization. Transactions of the Association for Computational Linguistics, 9, 1475–1492. https://doi.org/10.1162/tacl_a_00438

Dou, Z., Liu, P., Hayashi, H., Jiang, Z., & Neubig, G. (2021). GSum: A General Framework for Guided Neural Abstractive Summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4830–4842, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.384

Liu, Z., & Chen, N. (2021). Controllable Neural Dialogue Summarization with Personal Named Entity Planning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.8

Maddela, M., Kulkarni, M., & Preotiuc-Pietro, D. (2022). EntSUM: A Data Set for Entity-Centric Extractive Summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3355–3366, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.237

Cao, M., Dong, Y., & Cheung, J. (2022). Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.236