LLMEval

Category: Research

1 Introduction

Text summarization is an extensively studied NLP task that has many downstream applications. The recent success of Large Language Models (LLMs) opens up even more application scenarios. With LLMs, a summarization can be easily generated with zero-shot or few-shot prompting. Prompt-based summarization provides opportunities for injecting user intent into the summarization process and is therefore highly customizable. Moreover, prompt-based summarization needs zero or few examples, further allowing users to craft examples tailored to their application contexts. Each prompt template can thus be seen as a summarization system that demonstrates better capability over previous neural models that do not respond to user intents and require a large amount of labeled data to train. The development of such prompt-based summarization systems has become a highly iterative process.

However, the evaluation of prompt-based summarization systems remains a challenge. The most common approach for evaluating summarization systems is to compare the average score of a computational metric such as ROGUE [6] on a labeled dataset, but such an approach has been criticized in many ways [1–3, 9, 10] even in previous neural-based summarization systems. Evaluating LLM summarization systems using computational metrics is even more questionable. First, computational metrics do not reliably quantify system improvements among state-of-the-art systems because they do not distinguish high-quality summaries very well [1, 2, 9]. Much of the improvements observed over such metrics between a newly proposed system and the baseline system are attributed to successfully distinguishing the 'easy' cases, i.e.cases where the baseline system fails very badly. LLMs have been demonstrated to be able to produce human-level summaries that the quality differences are too nuanced for computational metrics to capture [11]. Second, computational metrics rely on reference texts (ground-truths) to compute, which are often crafted by human writers. Such reference texts are expensive to craft and do not generalize well across domains. The usage of such reference texts also limits the customizability of LLMs. More recently, reference-free metrics such as Q^2 [5] or FEQA [3] have been proposed for evaluation without reference texts. However, they are reported to be capturing spurious correlations [4]. Another trend of reference-free evaluation is using LLM as evaluators, but they do not yet exhibit a consistent correlation with human judgment [11]. In addition, all these metrics have a high computation time that is not suitable for an iterative development process.

One commonality among the above approaches is that they are all automatic approaches that completely exclude humans in the evaluation process. We argue that the capability of LLMs in summarization has reached, if not exceeded, the human level. Prompt-based summarization can be applied to almost any application scenario, with a high level of customizability. The quality of a summarization system in real-world scenarios is arguably too complex to be quantified by computational metrics. Consequently, the evaluation of such summarization systems will likely be incomprehensive without a human in the loop.

In addition, embedding analysis has been shown to be effective in explaining LLM behaviors such as document relevancy ranking [7,8]. Following this direction, we propose to evaluate summarization systems by analyzing the embedding distribution of the input text and summarized text. We present experiments to show that the summarization system is essentially a transformation function that maps the input text embeddings to summarized text embeddings. Thus, the quality of a summarization system can be evaluated by analyzing the embedding distribution of the input text and summarized text.

Following the above discussion, we propose a visual analytic system that incorporates a human-in-the-loop approach to evaluate a summarization system. Our system provides a visualization of the embedding distribution of the input text and summarized text and interactions to analyze the distributions. In addition to the distribution visualization, we further incorporate existing computational metrics in the system as complementary signals. Our contributions are as follows:

- We show that embedding analysis is a promising approach for evaluating summarization systems.
- We propose to evaluate summarization systems in a human-in-the-loop approach and develop a visual analytic system to support embedding analysis for summarization system evaluation.
- We evaluate our system with quantitative experiments and qualitative expert reviews.

2 Related Works

2.1 Summarization Metrics and Meta Evaluation

Computational Metrics

Reference-free Metrics

LLM evaluator Metrics Outline 1. Existing problems with using automatic summarization metrics to guide prompt optimization/evaluate prompts: - LLM produces human-level summaries that quality differences are too nuanced for automatic metrics to capture [5] - Correlation with human judgment is questionable [6] - A trade-off between abstractive and faithfulness must be made for automatic metrics [10] - automatic metrics tend to correlate well with humans at the system level but have poor correlations at the instance level [2, 3] - No single metric can outperform (correlation) across datasets -¿ This suggests different datasets need to use different metrics [2] - Metrics can not reliably quantify improvements if the difference is too small – much success is attributed to ranking 'easy' cases [1, 2, 6]

2. Problems with LLM evaluators: outperforms automatic metrics, but do not exhibit consistent correlation with human judgment [5] 3. Problems with reference-free metrics: capturing spurious correlation (highly correlated with spurious metrics like text length, word overlap, and perplexity) [8] 4. LLM evaluators and reference-free metrics have higher computation time 5. New problem in summarization task when LLM is used: Robustness

Notes: 1. LLM benchmarks are neither informative nor actionable for application developers 2. evaluation of a new metric: system-level vs. instance level: which one would visual metric succeed at? 3. reference-based metrics are not able to capture factuality (faithfulness) errors [7] 4. Meaningful -¿ Grammatical/Readable/Formal -¿ Faithful -¿ Capturing gist

2.2 Latent Space Visualization

References

[1] M. Bhandari, P. Gour, A. Ashfaq, P. Liu, and G. Neubig. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*, 2020.

[2] D. Deutsch, R. Dror, and D. Roth. Re-examining system-level correlations of automatic summarization evaluation metrics. *arXiv preprint arXiv:2204.10216*, 2022.

[3] E. Durmus, H. He, and M. Diab. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754*, 2020.

[4] E. Durmus, F. Ladhak, and T. Hashimoto. Spurious correlations in reference-free evaluation of text generation. *arXiv preprint arXiv:2204.09890*, 2022.

[5] O. Honovich, L. Choshen, R. Aharoni, E. Neeman, I. Szpektor, and O. Abend. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*, 2021.

[6] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

[7] C. Lucchese, G. Minello, F. M. Nardini, S. Orlando, R. Perego, and A. Veneri. Can embeddings analysis explain large language model ranking? In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4150–4154, 2023.

[8] A. Mishra, U. Soni, A. Arunkumar, J. Huang, B. C. Kwon, and C. Bryan. Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models. *arXiv preprint arXiv:2304.01964*, 2023.

[9] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*, 2017.

[10] M. Peyrard, T. Botschen, and I. Gurevych. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 74–84, 2017.

[11] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.