Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230977

A Question-Answering Approach to Evaluating Legal Summaries

Huihui Xu [a,b,1], Kevin Ashley [a,b,c]
[a] *Intelligent Systems Program, University of Pittsburgh*
[b] *Learning Research and Development Center, University of Pittsburgh*
[c] *School of Law, University of Pittsburgh*

Abstract. Traditional evaluation metrics like ROUGE compare lexical overlap between the reference and generated summaries without taking argumentative structure into account, which is important for legal summaries. In this paper, we propose a novel legal summarization evaluation framework that utilizes **GPT-4** to generate a set of question-answer pairs that cover main points and information in the reference summary. GPT-4 is then used to produce answers based on the generated summary for the questions from the reference summary. Finally, GPT-4 grades the answers from the reference summary and the generated summary. We examined the correlation between GPT-4 grading and human grading. The results suggest that this question-answering approach with GPT-4 can be a useful tool for gauging the quality of the summary.

Keywords. Summarization, natural language processing, question-answering, argument mining

1. Introduction

Readers need summaries to convey a rough idea of what a case is about and why it is important. This enables users to connect a case to their personal needs and to decide whether to read the case decision. As a result, the quality of legal summaries is important. Commonly used summary evaluation metrics such as ROUGE scores [1] focus primarily on surface-level aspects like word overlap and grammatical correctness. These metrics do not consider factors such as contextual understanding or the alignment of the summary with the reader's specific goals or preferences.

In this work, we propose a novel method to evaluate the quality of a legal summary by leveraging automated question-answering while incorporating legal argumentative structure. The argument structure comprises three elements: **Issue** – legal question that a court addressed in the case; **Reason** – elaboration of why the court reached the conclusion; and **Conclusion** – the court's final decision regarding the issue. Our method consists of three steps: (1) Given a reference legal summary, a question-answer generation model (GPT-4) produces a set of question-answer pairs based on the legal argumentative structure of the reference summary. (2) Then we use a question-answering model (GPT-4) to answer the questions from the reference summary based on the text of the generated summary.

[1] Huihui Xu is the corresponding author and can be contacted via email: huihui.xu@pitt.edu.

(3) Finally, GPT-4 compares the answers in step (1) from the reference summary with the answers in step (2) from the generated summary and assigns grades based on the similarity. Code is available at https://github.com/JoyceXu02/QA_evaluation.

2. Related Work

The Stanford Question Answering Dataset (SQuAD) introduced in [2] is useful for training and evaluating question-answering (QA) systems. It contains questions posed on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding passage. In this work, we are inspired by the idea of assessing the quality of summaries by asking questions about them: if a good summary retains all crucial information, then it should be able to answer questions about the original content accurately. Researchers in [3,4] proposed and concluded that human evaluators prefer QA-based metrics for evaluating abstractive summaries. Taking the QA-based evaluation research methods even further, [5] tackled the unfaithfulness of neural abstractive summarization by proposing QA-based automatic metrics, FEQA. Our work inherits the idea of using a QA-based method to evaluate the summary quality while taking legal argument into account.

Basing an evaluator on a large language model (LLM) has become increasingly popular lately. The authors of [6] proposed an evaluation framework, GPTScore, with generative pre-training models like GPT-3. They suggest that higher-quality text is more likely to be generated by following a given instruction in a given context. Researchers in [7,8] present a preliminary study of using LLMs as a Natural Language Generation (NLG) evaluator. Their LLM evaluation achieved a new state-of-the-art correlation in the summary evaluation task.

Few prior projects apply a question-answering approach to evaluation in a legal context. Our approach leverages recent progress in large language models while taking legal argumentative structure into account. Before the bloom of LLMs, an QA approach relied on corresponding curated datasets [9,10]. Our approach does not require a specific question-answering dataset and generates questions and answers automatically.

3. Methodology

3.1. Experimental Design

We utilize GPT-4 to generate question-answer pairs, incorporating the example prompt illustrated in Figure 1. This enhanced prompt enables us to generate not only question-answer pairs but also the corresponding question types. Subsequently, we utilized these questions as prompts to the model for predicting responses based on model-generated summaries. The prompt used for the prediction is shown in Figure 2. The prompt for GPT-4 to evaluate the answers is shown in Figure 3. We set the temperature to 0 for the GPT-4 generation part to get the most deterministic results. When human evaluators assess the quality of automatically generated summaries, they use previously generated question-answer pairs as a guide or reference. These pairs help the evaluators know what to look for and how to judge the summary's quality.

Throughout our research, we experimented with three models for generating summaries: Longformer Encoder-Decoder (LED) [11], BART [12], and GPT-4. LED and BART require fine-tuning in order to generate reasonable summaries while GPT-4 can generate summaries in a zero-shot setting.

3.2. Data

We developed a type system to annotate Canadian legal case summaries[13,14]. This type system includes three key components: Issue, Reason, and Conclusion. The dataset initially consisted of 1,049 annotated summaries along with their corresponding full-case decisions. We used the same dataset to support this work.

We used 90% of the data for fine-tuning LED and BART models. The remaining 10% of the data was for testing purposes. GPT-4 was used to generate summaries for this 10% subset of data without fine-tuning. Considering the cost associated with GPT-4 and human evaluation, however, we opted to leverage our QA approach to evaluate 10 summaries generated by each model.

```
Act like a legal professional and read the following legal text.
Use Issue, Reason, Conclusion sentences to generate question-answer pairs [...]

List those generated questions and answers in the following format:

Question: {{generated_qustion}}
Type:{{question_type}}
Answer:{{generated_answer}}
```

Figure 1. Prompt template for generating question-answer pairs based on annotated sentences.

4. Results and Discussion

There are 48 question-answer pairs for 10 cases. A human evaluator assessed whether the generated question-answer pairs adequately captured the necessary information and were addressed correctly. The evaluation options were limited to "YES" and "NO". Based on the results, 42 out of the 48 questions accurately captured the required information, while all 48 answers were correct and appropriately addressed the questions. Table 1 shows an example of GPT-4-generated QAs. This example shows that GPT-4 can generate coherent and contextually relevant answers to specific types of questions. Those QAs served as ground truth when comparing to the predicted grading.

The prompt in figure 2 is for predicting answers based on the previous generated questions and generated summaries. Figure 3 shows the prompt we used for grading the predicted answer against the ground truth. We converted GPT-4 grades (0-10 scale) into binary by setting a threshold. We recognize that the numerical meaning of a generated score might be different from human perception. To avoid misinterpretation, we show results at thresholds 5 and 6 for sensitivity analysis. The grades equal and above the threshold map to "YES" and grades below the threshold map to "NO". Furthermore, the human evaluator assessed whether the generated answer correctly addresses the given

```
Answer the {{type}} question based on the context
Context:{{context}}
Question:{{question}}
Answer:
```

Figure 2. Prompt template for predicting answers based on the model-generated summary.

```
You are a legal expert to judge the answers to questions.

You are judging the following question:
{{qustion}}
The real answer is:
{{real_answer}}
You are grading the following predicted answer:
{{predicted_answer}}

What grade do you give on a scale from 0 to 10, where 0 is the lowest (can’t find
the answer) and 10 is the highest (very close to the real answer).

Finally, give the explanation of the grade
Explanation: {{explanation}}
```

Figure 3. Prompt template to evaluate the predicted answer with the real answer.

Table 1. GPT-4 generated QA examples. The questions in red pertain to an issue; the one in blue is focused on the reason, and the question in teal color aims at drawing the conclusion.

Summary	Question	Answer
Warrant issued to search a dwelling house for weapons allegedly used in an attempted armed robbery. The affidavit in support referred to an unknown informant. Judge applied the test that the justice of the peace ‘must be satisfied on reasonable grounds.’ Substantial compliance found and warrant upheld.	What was the warrant issued for? What test did the judge apply to determine the validity of the warrant? What was the conclusion of the case?	The warrant was issued to search a dwelling house for weapons allegedly used in an attempted armed robbery. The judge applied the test that the justice of the peace’must be satisfied on reasonable grounds.’ Substantial compliance was found and the warrant was upheld.

question in relation to the model-generated summary. The evaluation options were also limited to “YES” and “NO”. During the assessment, the human evaluator found some of the answers were legally correct, but went well beyond the information provided in the generated summary. The evaluator also found that some answers contained hallucinations.

Table 2 shows the two types of correlation between GPT-4 evaluation grade and human evaluation at two set of thresholds (5 and 6). Pearson correlation measures the linear relationship between GPT-4 and human evaluations, while Spearman correlation measures the monotonic relationship between the two types of evaluations. Those metrics measure two aspects of the relationship between two variables. Here, we focus solely on the results at the threshold of 5 which closely align with those at threshold 6. Automatic evaluation of BART-generated summaries has highest correlation with human evaluation

Table 2. The correlation between the GPT-4 evaluation grade and the human evaluation. IRC is short for Issue, Reason and Conclusion. The IRC correlation is computed across all sentences related to issues, reasons, and conclusions without categorizing or grouping them by these types. Number in parentheses is the chosen threshold.

MODEL	TYPE	PEARSON(5)	SPEARMAN(5)	PEARSON(6)	SPEARMAN(6)
BART	ISSUE	0.67	0.72	0.67	0.72
	REASON	-0.07	-0.17	-0.07	-0.17
	CONCLUSION	0.29	0.29	0.29	0.29
LED	ISSUE	0.69	0.52	0.69	0.52
	REASON	0.43	0.31	0.45	0.36
	CONCLUSION	0.12	0.12	0.12	0.12
GPT-4	ISSUE	0.56	0.56	0.56	0.56
	REASON	-0.09	-0.20	0	-0.11
	CONCLUSION	0.57	0.57	0.57	0.57
BART	IRC	0.51	0.54	0.51	0.54
LED	IRC	0.87	0.84	0.88	0.85
GPT-4	IRC	0.50	0.48	0.52	0.48

on Issue types of answers with Spearman (0.72) correlations. LED-generated summaries have the highest correlation on Reasons (0.43 in Pearson and 0.31 in Spearman) and highest Pearson correlation on Issue (0.69). GPT-4 generated summaries have the highest correlation on Conclusions (0.57 in both Pearson and Spearman). In terms of Reason types of answers, we notice that the automatic evaluation has a negative correlation with the human evaluation on BART generated and GPT-4 generated summaries in both Pearson and Spearman correlation.

Overall, LED exhibits robust linear (0.87) and monotonic (0.84) relationships. The correlation results of BART suggest that the Spearman relationship is stronger (0.54). GPT-4 has a stronger Pearson correlation (0.50).

5. Conclusion and Future Work

In conclusion, our QA approach for evaluation tends to align with human judgements on Issue and Reason type answers. It could be useful assessing the quality of summarization systems. The strength of this correlation, however, can vary depending on the model and specific aspects of the evaluated summary. As a result, while the method offers valuable insights, it is best used along with other metrics for a comprehensive assessment.

While we show that GPT-4 achieves reasonable correlation with human evaluation of summaries, there are limitations that provide directions for future work: (1) Since GPT-4's performance as an evaluation metric is sensitive to the construction of prompts, we plan to explore various prompts to achieve better performance. (2) We need to scale up the experiment to show more robust comparison results. (3) Quality control of model generation is necessary, especially when the input context increases in length and structural complexity. We will further explore open-source models to calibrate the output and ensure consistency.

Acknowledgements

This work has been supported by grants from the Autonomy through Cyberjustice Technologies Research Partnership at the University of Montreal Cyberjustice Laboratory and the National Science Foundation, grant no. 2040490, FAI: Using AI to Increase Fairness by Improving Access to Justice. The Canadian Legal Information Institute provided the corpus of paired legal cases and summaries. This work was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-21176.

References

[1] Lin CY. Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out; 2004. p. 74-81.

[2] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016. p. 2383-92.

[3] Eyal M, Baumel T, Elhadad M. Question Answering as an Automatic Evaluation Metric for News Article Summarization. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019. p. 3938-48.

[4] Scialom T, Lamprier S, Piwowarski B, Staiano J. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 3246-56.

[5] Durmus E, He H, Diab M. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 5055-70.

[6] Fu J, Ng SK, Jiang Z, Liu P. Gptscore: Evaluate as you desire. arXiv preprint arXiv:230204166. 2023.

[7] Wang J, Liang Y, Meng F, Shi H, Li Z, Xu J, et al. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv e-prints. 2023:arXiv-2303.

[8] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, May 2023. arXiv preprint arXiv:230316634.

[9] Anantha R, Vakulenko S, Tu Z, Longpre S, Pulman S, Chappidi S. Open-domain question answering goes conversational via question rewriting. arXiv preprint arXiv:201004898. 2020.

[10] Adlakha V, Dhuliawala S, Suleman K, de Vries H, Reddy S. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. Transactions of the Association for Computational Linguistics. 2022;10:468-83.

[11] Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv preprint arXiv:200405150. 2020.

[12] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020. p. 7871-80.

[13] Xu H, Ashley KD. Multi-granularity Argument Mining in Legal Texts. In: International Conference on Legal Knowledge and Information Systems; 2022. .

[14] Xu H, Savelka J, Ashley KD. Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In: Legal Knowledge and Information Systems. IOS Press; 2021. p. 33-42.