Structure-Aware Data Inpainting for Legal Dialogue Systems

Jisang Yu
Dept. Data Science

Daye Lee
Dept. Data Science

Yujin Jo
Dept. Data Science

Insu Han
Dept. Naval Architecture and Ocean Engineering

Sunyoo Kim
Dept. Data Science

Abstract

Dialog inpainting is a method of generating question-answering dialogues from documents. In this study, we present new data inpainting approaches to legal documents. We focus on three properties of legal documents including hierarchical structure, cross-referencing nature and semantical similarity. We make three data structures that represent each property, preprocess the legal documents reflecting each structure and turn the preprocessed data into question-context (QC) pair dialogue datasets via inpainting where the context is defined as each line of the original documents. Then, we restyle the QC datasets into question-answer (QA) datasets so that contexts are restyled into human-like natural answers. Finally, we use our inpainted data to fine-tune a Llama-2-7B chat model to generate answers to the questions of users and perform evaluation both by humans and GPT-3.5. Our results on accuracy, helpfulness, well-formedness, and human-likeness show improvements in the model's ability to engage in legally relevant and human-like dialogues.

1 Introduction

Motivation Although AI chatbots such as ChatGPT (OpenAI, 2023) based on pre-trained large language models (LLMs) are emerging as general problem solvers, they are still limited in specialized areas such as law. Collecting training datasets for interactive dialogue systems is both time-consuming and resource-intensive, especially for legal data. Although the previous dialog inpainting approach (Dai et al., 2022) partially solved this problem, it fails to capture the hierarchical structure and co-referencing relationship between passages in complex legal documents. Moreover, fine-tuning chatbots on raw law provisions also deprives them of the ability to generate human-like natural responses.

Problem definition To summarize, there are two problems to address. First, simply transforming legal documents into QC datasets does not capture the innate structures of legal documents. Secondly, the previous dialog inpainting method has limitations of resulting in artificial responses. In order to address these challenges, we develop three new dialog inpainting approaches to produce synthetic legal dialogue datasets. The resulting datasets can accurately reflect the complex nature of legal documents while being resource-efficient, and contain answers whose style resemble natural human conversation. As a real-world case study, we test our approach on the Seoul National University Laws.

2 Related work

Data augmentation addresses data scarcity by artificially generating new training examples. In natural language processing, methods such as word-level augmentation using synonyms (Niu and Bansal, 2018) or masked language models (Kobayashi, 2018) have been used, and backtranslation (Sennrich et al., 2016) at the sentence level. DADS (Liu et al., 2022) created synthetic training examples for low-resource areas such as the legal domain, but has a limited focus on dialogue summarization. Recently, some approaches such as AugGPT (Dai et al., 2023) directly used pretrained LLMs for data augmentation. This method outperforms traditional data augmentation methods but simply paraphrased the documents using ChatGPT without further consideration. In contrast, our method takes the structure of complex legal documents into account for data augmentation, resulting in differently structured preprocessed datasets.

Dialog inpainting (Dai et al., 2022) transforms documents into dialogues by taking a partial dialogue and generating a complete dialogue where the inpainter model fills in the unobserved turns. Recently, some studies use this approach to overcome the insufficient training data issue for dialogue systems. In the legal domain, one study (Yuan et al., 2023) aims to simplify complex legal documents by constructing a Legal Question Bank using GPT-3 for question generation and dialog inpainting. This method is effective in generating

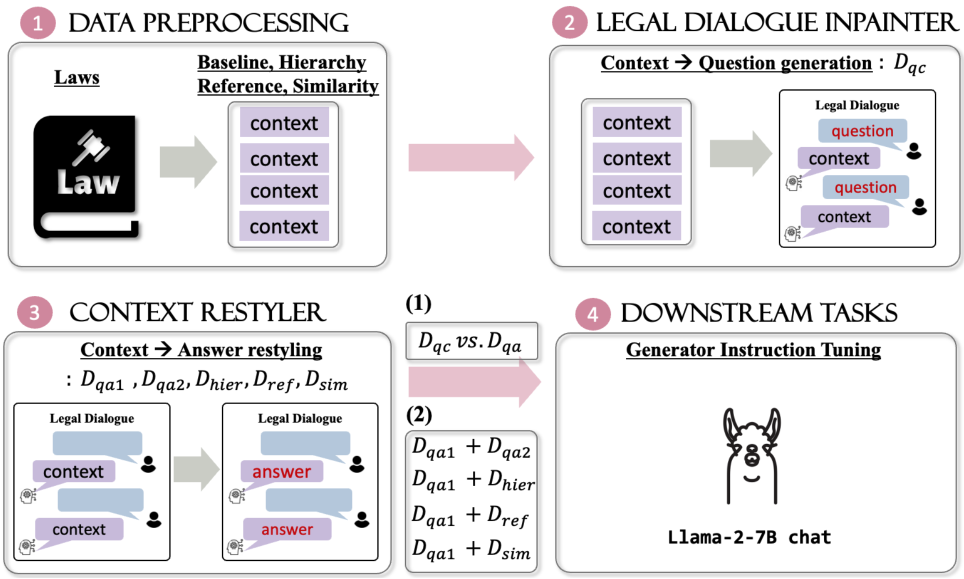

Figure 1: The Framework of our research. The "context" in the figure indicates the original legal paragraph in legal documents. On the other hand, the "answer" is the informal and human-like answer which is restyled from the "context"

diverse questions but the generated questions have poor accuracy. Dialog inpainting can enhance pretraining datasets for dialogue systems, but currently has limitations of resulting in artificial responses.

3 Methods

Our overall framework consists of four steps as shown in Figure 1. First, considering the characteristics of legal documents, we preprocess Seoul National University Laws to produce four preprocessed datasets where each dataset reflects each of the unique legal documents' properties (See Figure 2). Second, we reconstruct the datasets into question-context (QC) datasets using dialog inpainter based on ChatGPT 3.5, where contexts indicate that each line of the original law documents(Dai et al., 2022). After that, we apply context restyling to change each context into a human-like answer (question-answer (QA) datasets). Finally, we fine-tune a Llama-7b-chat model (Touvron et al., 2023) on our datasets to see how the model generate answers in the law-domain multi-turn dialog situation evaluating their effectiveness in an actual training setting.

3.1 Data preprocessing

3.1.1 Baseline

Legal documents generally consists of paragraphs that fall into the same [title-chapter-topic-article] structure as shown in Figure 2 which reflects the previous dialog inpainting approach (Dai et al., 2022). We named this straightforward structure where an article contains several paragraphs as the baseline structure. (see (1) in Figure 2)

3.1.2 Hierarchy, Reference, and Similarity

Legal documents are structured hierarchically, in a [title-chapter-topic-article-paragraph] format (see (2) in Figure 2, where each paragraph can contain or refer to each other (see (3) in Figure 2). Furthermore, for every paragraph, there can exist semantically similar paragraphs (see (4) in Figure 2. To preprocess the Seoul National University Laws so that those complex properties can be reflected, we convert the legal documents into a table format, organizing paragraphs by their titles, topics, chapters, and articles. At the same time, the table also include paragraphs' indices that the current paragraph are referencing to for taking the cross-referencing properties into account.

To conclude, we generated 4 tabular preprocessed datasets according to each structure - baseline, hierarchy, reference, similarity. The similarity is calculated with dot product.

3.2 Legal dialogue inpainter

We use the ChatGPT 3.5 model as our dialogue inpainter to turn our preprocessed four datasets into question-context (QC) datasets consisting of generated questions and contexts (paragraphs). To be clear, the context means the original content in the legal documents. For comparison with the

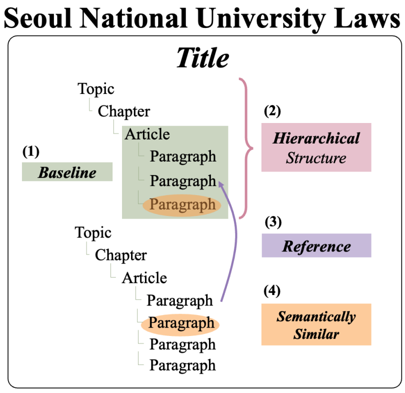

Figure 2: The Structure of Seoul National University Laws

dataset generated in the next step, we denote the baseline QC dataset as D_{qc}. Unlike the original dialog inpainting approach (Dai et al., 2022), we do not restrict the number of sentences due to the nature of legal documents.

3.3 Context restyler

Because our synthetic D_{qc} created in the previous step have direct quotes from the formal legal documents (contexts), their conversational style is formal and artificial. To mitigate this problem, we use a context restyler to transform our D_{qc} into question-answer (QA) datasets that more closely resembles human conversation. We use the ChatGPT 3.5 model as our context restyler, and use prompting to transform the datasets. We refer to the resulting four QA datasets as D_{qa}, D_{hier}, D_{ref}, and D_{sim}.

3.4 Downstream tasks

In order to evaluate our QA datasets, we adapt a fine-tuning approach. We hypothesize that if our dialog inpainting approaches are effective, training chatbots on our datasets would lead to better performance.

3.4.1 QC vs. QA

We fine-tune a Llama-2-7B chat (Touvron et al., 2023) model each on our D_{qc} and D_{qa}(see (1) in Figure 1) to evaluate the effectiveness of our conetxt restyling approach.

3.4.2 Legal domain-specific approaches

We want to see the effect of our domain-specific datasets, D_{hier}, D_{ref}, D_{sim}. Before fine-tuning, in order to evaluate each model fairly, we need to match the size of the datasets. Specifically, we create another D_{qa2} and sample all datasets except D_{ref} matching their size to the size of the smallest dataset (D_{ref}). Then, we fine-tune the Llama-2-7B chat model (Touvron et al., 2023) and create QA baseline (D_{qa1} + D_{qa2}), hierarchy (D_{qa1} + D_{hier}), reference (D_{qa1} + D_{ref}), and similarity model (D_{qa1} + D_{sim}) (see (2) in Figure 1). We use simple supervised fine-tuning (SFT) where the prompts and answers from the previous turn are accumulated as input to the next turn, and update our models with an autoregressive objective. We train each model for 5 epochs, and the resulting models are evaluated on a multi-turn generation task.

	# utterances			# tokens			
	avg	max	min	avg	25%	median	75%
D_{qa1}	4.51	22	2	94	52	86	125
D_{qa2}	4.43	20	2	94	52	87	125
D_{ref}	5.52	24	4	96	55	86	124
D_{hier}	5.62	6	4	171	144	200	200
D_{sim}	6.00	6	6	98	62	87	119

Table 1: Statistics of our datasets. The left table shows the number of utterances per one dialogue, and the right table shows the number of tokens per each utterance of one dialogue in our QA datasets.

4 Experiments

4.1 Dataset

As shown in Table 1, the number of turns per dialogue are set to approximately five, and the token number statistics for each utterance vary. During testing, we draw 10 conversations from the dataset that the model was fine-tuned on and use the first user question to initiate the dialogue.

4.2 Experimental settings

4.2.1 Metrics

A legal chatbot should be capable of delivering precise and accurate information based on legal documents. The chatbot should provide detailed and specific information, ensuring that it is helpful to users seeking legal information. Furthermore, the chatbot must be able to generate responses that are grammatically correct to maintain the reliability expected in the legal domain. Lastly, the chatbot should engage in natural and human-like conversations. This human-like interaction is essential for making users feel comfortable, when dealing with complex legal information. We evaluate legal chatbot's answer generation performance in a legal-domain multi-turn dialogue setting. **Accuracy** examines whether the model's responses were correct compared with the original legal contexts without any hallucination. **Helpfulness** measures if the model's responses were informative, understandable, and easily comprehensible. **Well-formedness** was to check the grammatical and structural integrity of responses. Finally, **Human-likeness** assesses the model's ability to mimic natural human conversation, fostering user comfort and engagement.

4.2.2 Human and GPT 3.5 evaluation

Human evaluation was conducted by 5 workers from our team. To complement possible bias by human evaluators, we additionally use the GPT 3.5 model for evaluation. We prompt GPT 3.5 with the ground truth law dataset, the goal of the evaluation task and metrics, the generated dialogues from our baseline model and from the other models, and ask it to compare which dialogue is better.

5 Results

5.1 QC vs. QA

We performed evaluation on our D_{qc} and D_{qa} in order to see the effect of context-restyling on how the fine-tuned chatbot model responds to the legal questions. Our baseline is the model fine-tuned on D_{qc} and compared it with the same model fine-tuned on D_{qa}.

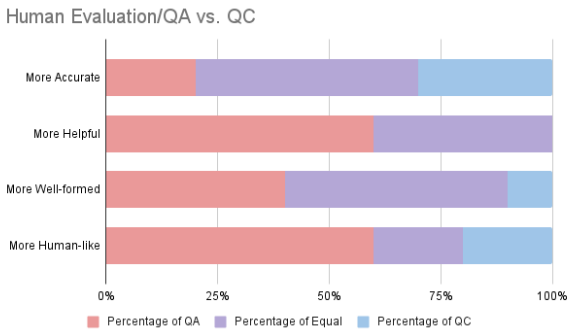

Figure 3: Evaluation results on QC vs QA

The QA model outperformed the QC model in terms of human-likeness in both Human Evaluation and GPT Evaluation. The QA model received higher ratings for being more helpful and for producing well-structured answers in human evaluations. Fine-tuning the model using QA datasets has the potential to create a chatbot that is more relatable to humans. However, the QC model scored higher in terms of accuracy. It appears that the introduction of noise during the transition from QC to QA had a negative impact on accuracy.

5.2 Legal domain-specific approaches

5.2.1 Hierarchy

We show the human evaluation and GPT-3.5 evaluation results of our domain-specific approaches in Table 2. When comparing the model fine-tuned on D_{hier} with the model fine-tuned on D_{qa}, the former exhibited higher accuracy, more helpfulness, and more human-likeness. Given the characteristics of the dataset that progressively refines questions from general to specific, the model appears to demonstrate a more profound understanding of the legal context.

5.2.2 Reference

The reference model outperformed the baseline model in terms of accuracy, helpfulness, well-formedness, and human-likeness. This effect can be attributed to learning the reference relationship between provisions unique to legal documents. While scalability is constrained by the limited number of provisions containing reference relationships, we hold the belief that it will complement other approaches.

5.2.3 Similarity

The performance of the similarity model was inferior to that of the baseline model in all aspects. This is likely due to the fact that the similarity in the embedding space based on the dot product may not accurately capture the actual semantic similarity.

6 Conclusion

Contributions The contributions of our research are threefold. First, we applied the dialog inpainting technique to the legal domain, where QA data collection for dialogue systems is highly expensive and more difficult than the general domain. Second, we introduced various inpainting methodologies reflecting legal document's cross-referencing nature, hierarchical structures, and semantic similarities between contexts. Third, we optimized the datasets for the multi-turn dialogue generation task by restyling the artificial answers into natural answers. This shows the novelty of our approach, considering chatbot models trained on QC datasets from previous dialogue inpainting methods (Dai et al., 2022)would lead to artificial and document-like answers.

Limitations & Future works The limitations of our study are that due to cost problems, fewer questions were used for fine-tuning compared to the

		Human evaluation				GPT-3.5 evaluation			
		Accurate	Helpful	Well-formed	Human-like	Accurate	Helpful	Well-formed	Human-like
QA	Ours	20	**60**	40	**60**	10	10	0	30
	Equal	**50**	40	**50**	20	30	30	**90**	**70**
	Baseline	30	0	10	20	**60**	**60**	10	0
Hierarchy	Ours	33	**40**	27	33	**47**	**47**	**40**	**47**
	Equal	**67**	27	**40**	**47**	33	20	33	13
	Baseline	0	33	33	20	20	33	27	40
Reference	Ours	**53**	33	20	**53**	**40**	**40**	7	40
	Equal	33	**47**	**80**	33	**40**	**40**	**93**	**47**
	Baseline	13	20	0	13	20	20	0	13
Similarity	Ours	13	13	13	33	20	20	13	20
	Equal	**80**	**47**	**87**	33	33	13	**53**	7
	Baseline	7	40	0	33	**47**	**67**	33	**73**

Table 2: Human and GPT 3.5 evaluation results

total possible number of dialogues. Also, the precise details of legal documents seems to be somewhat blurred going through the Korean-English translation-inpainting-restyling process. For future work, we propose a more delicate prompt design to form questions with legal details. Additionally, we hope to build Korean QA datasets since they would not have mistranslation issues and the blurring issues would be mitigated. Also, it is important to develop cost-free text data augmentation methodologies. Finally, finding a way to filter out similarity examples that are lexically but not semantically similar would reinforce our approach.

References

Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4558–4586. PMLR.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710, Seattle, United States. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Conference on Computational Natural Language Learning*.

OpenAI. 2023. Chatgpt. Accessed: [2023.12.11].

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Mingruo Yuan, Ben Kao, Tien-Hsuan Wu, Michael Cheung, Henry Chan, Anne Cheung, Felix Chan, and Yongxi Chen. 2023. Bringing legal knowledge to the public by constructing a legal question bank using large-scale pre-trained language model. *Artificial Intelligence and Law*, pages 1–37.

A Evaluation results

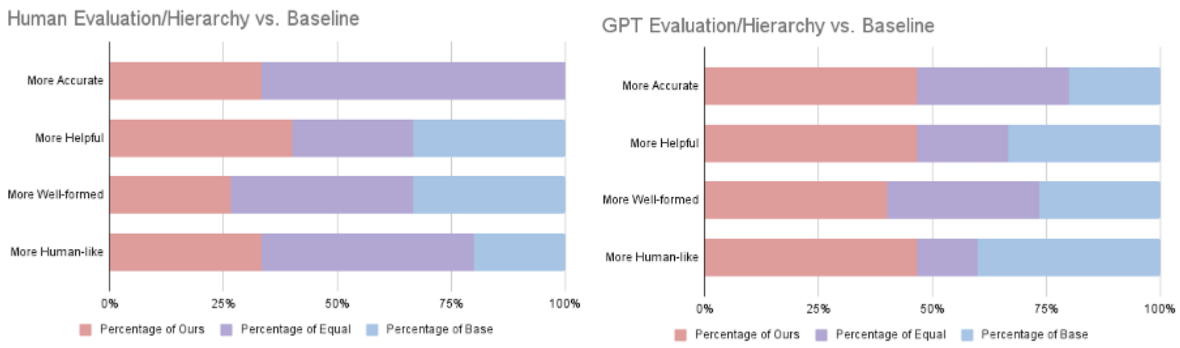

Figure 4: Evaluation results on D_{hier}

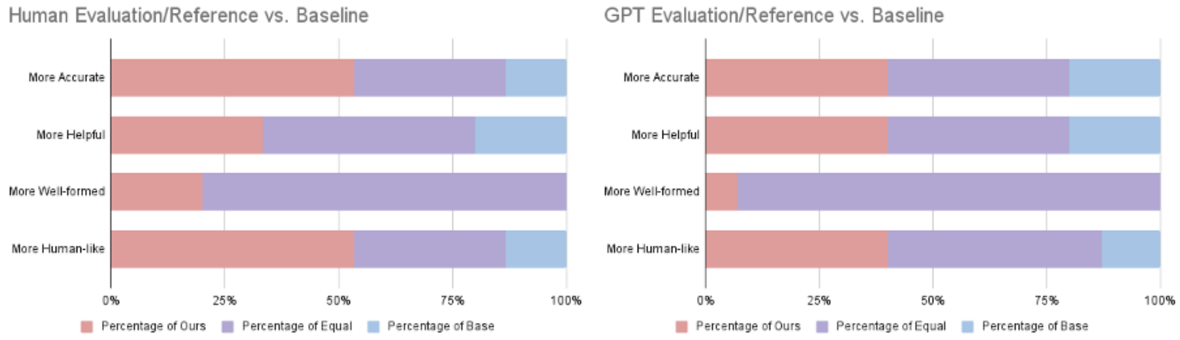

Figure 5: Evaluation results on D_{ref}

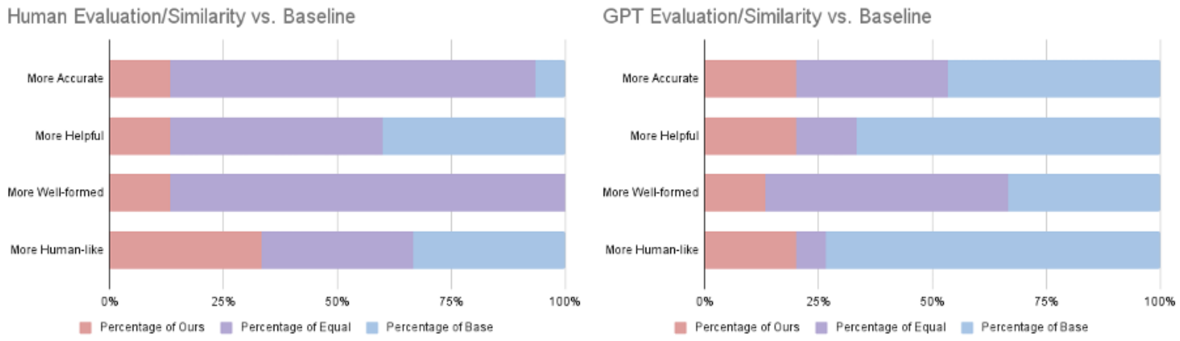

Figure 6: Evaluation results on D_{sim}