Generating Reflective Questions for Engaging Gallery Visitors in *ArtMuse*

Sujatha Das Gollapalli, Mingzhe Du, See-Kiong Ng

Institute of Data Science, National University of Singapore
{idssdg,mingzhe,seekiong}@nus.edu.sg

Abstract

Human guides in museums and galleries are professionally trained to stimulate informal learning in visitors by asking *low-risk, open-ended* reflective questions that enable them to focus on specific features of artifacts, relate to prior experiences, and elicit curiosity as well as further thought. We present *ArtMuse*, our AI-powered chatbot for asking reflective questions in context of paintings. Our reflective question generation model in *ArtMuse* was trained by applying a novel combination of existing models for extractive question answering and open-domain chitchat. User evaluation studies indicate that we are able to generate fluent and specific reflective questions for paintings that are highly-engaging.

Introduction

GLAMs (Galleries, Libraries, Archives, and Museums) present an opportunity for AI applications to bridge the gap between the highly engaging human docents and passive classical audio guides (Schaffer et al. 2018; van Strien et al. 2022). To this end, several recent works have focused on building quizzes, game-style interfaces, and question answering (QA)-based chatbots using expert-written passages and other metadata of the artefacts for enabling visitor interaction and enhancing their visit experiences (Boiano et al. 2018; Ueta et al. 2021; Gollapalli et al. 2022).

Museum experience studies have, however, consistently shown that users are often reluctant to ask questions even though they may be curious. Museum education researchers have thus recommended for human docents to proactively initiate visitor engagement through low-risk, open-ended questions that prompt them to *look closer and dig deeper* (Templeton 2011; Othman 2012). To this end, human docents often employ **reflective** questions that—rather than "test knowledge"—aim to stimulate informal learning in visitors by bringing their attention to specific aspects of the artifacts, evoking visitors' own prior experiences, and eliciting further thought and reflection (Burnham and Kai-Kee 2011; Diamond, Horn, and Uttal 2016).

Reflective questions are different from answer-seeking questions, which is the dominant focus of current question generation and conversational AI systems (Lu and Lu 2021; Gao, Galley, and Li 2019). As an example, consider a sentence from a passage[1] on the painting titled "Adoration of the Child" by the artist Filippino Lippi. Table 1 shows a question generated using a state-of-the-art Question Generation (QG) system ProphetNet (Qi et al. 2020), and a reflective question generated by *ArtMuse*, our prototype system.

Sentence: *The tenderness between mother and child* is moving; she is no longer depicted as a remote and inaccessible figure , and her face is reminiscent . . .
Question from ProphetNet: In adoration of the child, what is the mother no longer depicted as?
Question from *ArtMuse*: Do you think the painting conveys the tenderness between mother and child?

Table 1: Example for Illustration

ArtMuse engages with gallery visitors over a multi-turn session for each given painting. For every turn, relevant information on the painting is presented for viewer consumption along with a suitable reflective question as a follow-up prompt. In this manner, viewers not only learn more about the painting via the presented information (as in traditional learning), but are also able to indulge in further thought and analysis as prompted by the reflective question.[2]

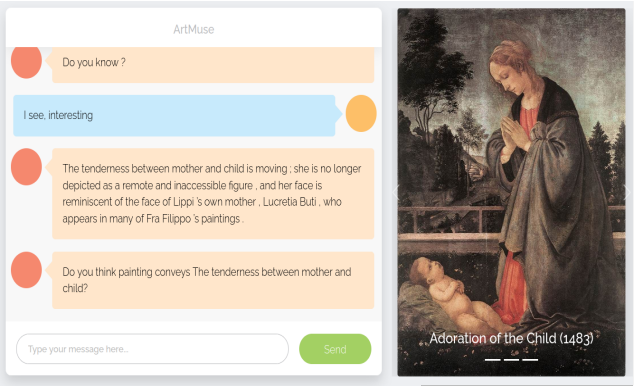

Figure 1: Reflective Questions in *ArtMuse*.

[1] https://www.britannica.com/list/25-famous-paintings-to-see-the-next-time-youre-in-florence

[2] Our demo available at https://nlp-platform.online/artmuse

Learning Reflective Question Generation

In current practice, QG model learning involves the use of large pre-trained language models that are fine-tuned on novel datasets (Li et al. 2021). However, while question generation has been a prominent subject of recent NLP research (Lu and Lu 2021), none of the available QG datasets contain reflective questions required to fine-tune models for our setting (Rajpurkar et al. 2016; Reddy, Chen, and Manning 2019; Ko et al. 2020; Svikhnushina et al. 2022). Instead, only general guidelines for reflective questions are available as part of art appreciation and docent education manuals.[3] *How can we employ currently available NLP systems for related problems to build a suitable dataset for learning Reflective Question Generation (RQG) in ArtMuse?* We address the above question with the following approach:

Candidate Text Span and Type Extraction Consider the example in Table 1. Using a Reading Comprehension (RC) question such as "What feelings are portrayed in this painting?" on the associated passage, currently available *extractive* QA systems are able to *extract* the span "tenderness between mother and child" as the answer. Considering this span comprises an answer to a question on emotion/feelings, we can create a reflective question such as "Do you feel *the tenderness between mother and child* in the painting? Based on this insight, we used the available art appreciation guides to curate a list of 33 questions that can be used with existing extractive QA systems. Our list of questions can be broadly categorized into five types referring to the *main theme*, *visual aspects*, *emotions*, *background context*, and the *stylistic aspects* of paintings. For each of these curated extractive questions, we also compiled exemplars and templates for generating reflective questions using rules.[4]

Candidate spans obtained with extractive QA when used in the rule-based model do not always result in well-formed sentences and depending on the accuracy of the extracted answers, they could be wrong. For example, our QA model incorrectly extracted "Virgin" as an answer to the question "What colors were used in the painting?" for one of the passages in our dataset resulting in a rule-based question such as "What did you think of the Virgin color used in the painting?". Such errors are mitigated in deep learning based text generation when pre-trained language models trained on large-scale corpora are incorporated. We therefore harness large pre-trained language models and fine-tune a deep learning model for our RQG task while using the rule-based questions in a data augmentation setting to generate the required training data.

Data Augmentation Since real chat sessions with human-generated reflective questions are unavailable, we adopt the following data augmentation strategy for generating synthetic visitor interaction sessions in *ArtMuse* (Feng et al. 2021). In a synthetic session, sentences from painting passages are presented in a sequence over multiple turns and in every turn, one of candidate text spans from the sentence is selected randomly along with a randomly-chosen reflective question from our rule-based model. Next, we mimic visitor response to the presented sentence and question by employing a chitchat dialog model (Roller et al. 2021). We used data from these synthetic sessions to train our RQG model. By incorporating viewer utterances and randomness in the choice of rule-based questions as part of the training input, our model is able to go beyond the rule-based model and learns to generate different reflective questions for different viewer utterances, where possible. A snapshot of *ArtMuse* in action is shown in Figure 1.[2]

Models and Datasets: We used the painting passages from previous work (Gollapalli et al. 2022) for generating synthetic sessions for training our RQG model. The chitchat dialog model used in the data augmentation step was trained on the PersonaChat dataset (Zhang et al. 2018). In both cases, we fine-tuned the Text-to-Text Transfer Transformer or the T5 model (Raffel et al. 2020).[5] To ensure correctness of the answers extracted by our QA system while generating synthetic sessions, we applied textual entailment using answer containing sentences and only consider answers that meet an entailment confidence threshold.[6] RoBERTa-based models (Liu et al. 2019) from AllenNLP[7] and HuggingFace[8] were used for textual entailment and QA, respectively.

User Evaluation Results: We performed a user evaluation study of the questions generated by our RQG model. Approximately twenty questions each for the five types (*main theme*, *visual aspects*, *emotions*, *background context*, and *stylistic aspects*) were rated by crowdworkers on the annotation platform, Amazon Mechanical Turk. Each question was independently examined by five workers, who rated the generated questions on a Likert scale (Amidei, Piwek, and Willis 2019) from 1 (very poor) to 5 (very high) for fluency, relevance, engagingness, and specificity.[9] The averaged ratings from the crowdworkers are around the medium range indicating that our model is able to generate reasonably fluent, relevant, and engaging questions using specific content from the painting passages. In particular, questions related to emotions and moods captured in the paintings were found to be the most engaging and those on the artist's background and context were found to be the least engaging. About 81% of the questions had medium to high specificity, and overall, questions with higher specificity also had higher fluency, relevance, and engagingness scores.

Acknowledgments

We thank Chow Shan Shan from National Gallery Singapore for inspiring our study on Reflective Questions and providing insights from the perspective of GLAMs.

[3] https://www.terraamericanart.org/tools-for-teachers/discussing-art-and-common-core-anchor-standards/

[4] All resources, details of experimental settings, and results are available at https://github.com/NUS-IDS/painter/tree/artmuse

[5] T5-large from https://huggingface.co/t5-large

[6] Our experiments with answer correctness models (Zhang, Yang, and Zhao 2021) are available at the github link.

[7] https://demo.allennlp.org/textual-entailment

[8] https://huggingface.co/deepset/roberta-base-squad2

[9] We used a 0/1/2 scale for No/Acceptable/Yes for Specificity

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Amidei, J.; Piwek, P.; and Willis, A. 2019. The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation*, 397–402.

Boiano, S.; Borda, A.; Gaia, G.; Rossi, S.; and Cuomo, P. 2018. Chatbots and New Audience Opportunities for Museums and Heritage Organisations. In *Proceedings of the Conference on Electronic Visualisation and the Arts*, EVA '18, 164–171.

Burnham, R.; and Kai-Kee, E. 2011. *Teaching in the Art Museum: Interpretation as Experience*. Getty Publications - Series. ISBN 9781606060582.

Diamond, J.; Horn, M.; and Uttal, D. 2016. *Practical Evaluation Guide: Tools for Museums and Other Informal Educational Settings*. American Association for State and Local History. Rowman & Littlefield Publishers.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988.

Gao, J.; Galley, M.; and Li, L. 2019. Neural Approaches to Conversational AI. *Foundations and Trends® in Information Retrieval*, 13(2-3): 127–298.

Gollapalli, S. D.; Ng, S.-K.; Tham, Y. K.; Chow, S. S.; Wong, J. M.; and Lim, K. 2022. PaintTeR: Automatic Extraction of Text Spans for Generating Art-Centered Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12503–12509.

Ko, W.-J.; Chen, T.-y.; Huang, Y.; Durrett, G.; and Li, J. J. 2020. Inquisitive Question Generation for High Level Text Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6544–6555.

Li, J.; Tang, T.; Zhao, W. X.; and Wen, J.-R. 2021. Pretrained Language Model for Text Generation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4492–4499.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Lu, C.-Y.; and Lu, S.-E. 2021. A Survey of Approaches to Automatic Question Generation:from 2019 to Early 2021. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, 151–162.

Othman, M. 2012. *Measuring visitors' experiences with mobile guide technology in cultural spaces*. Ph.D. thesis, University of York.

Qi, W.; Yan, Y.; Gong, Y.; Liu, D.; Duan, N.; Chen, J.; Zhang, R.; and Zhou, M. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *EMNLP Findings*, 2401–2410.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*, 2383–2392.

Reddy, S.; Chen, D.; and Manning, C. D. 2019. CoQA: A Conversational Question Answering Challenge. *TACL*, 7: 249–266.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In *EACL*, 300–325.

Schaffer, S.; Gustke, O.; Oldemeier, J.; and Reithinger, N. 2018. Towards chatbots in the museum. In *mobileCH@Mobile HCI*.

Svikhnushina, E.; Voinea, I.; Welivita, A.; and Pu, P. 2022. A Taxonomy of Empathetic Questions in Social Dialogs. In *ACL*, 2952–2973.

Templeton, C. A. 2011. *Museum Visitor Engagement Through Resonant, Rich and Interactive Experiences*. Ph.D. thesis, School of Design, Carnegie Mellon University.

Ueta, M.; Hashiguchi, T.; Pham, H.; Shoji, Y.; Kando, N.; Yamamoto, Y.; Yamamoto, T.; and Ohshima, H. 2021. Quiz Generation on the Electronic Guide Application for Improving Learning Experience in the Museum. In *BIRDS@SIGIR CEUR Workshop Proceedings*, volume 2863, 96–104.

van Strien, D.; Bell, M.; McGregor, N. R.; and Trizna, M. 2022. An Introduction to AI for GLAM. In *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop*, volume 170 of *Proceedings of Machine Learning Research*, 20–24.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *ACL*, 2204–2213.

Zhang, Z.; Yang, J.; and Zhao, H. 2021. Retrospective Reader for Machine Reading Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16): 14506–14514.