Appendix: *PaintTeR*: Automatic Extraction of Text Spans for Generating Art-Centered Questions

Sujatha Das Gollapalli,[1] See-Kiong Ng,[1] Tham Ying Kiat,[1]
Chow Shan Shan,[2] Wong Jia Min[2], Kevin Lim[2]

[1]Institute of Data Science, National University of Singapore
[2]CoLab X-Innovation Lab, National Gallery Singapore
{idssdg,seekiong,tyk}@nus.edu.sg, {shanshan.chow,jiamin.wong,kevin.lim}@nationalgallery.sg

ArtQuest

We provide a snapshot of *ArtQuest* in Figure 1 for an example painting "Flowers and Insects" by Rachel Ruysch. Both the image and passage are in the public domain.[1] A demo is available at http://artquestapp.herokuapp.com/app.

A user engages with *ArtQuest* by first selecting an artwork presented as a list of thumbnails of painting images. After a choice is made, *ArtQuest* offers the options to either (1) let the user ask a question via the textbox ("Type your question or answer here") or (2) let the machine ask the user a question by clicking the "Ask me a question" button.

The backend Reading Comprehension module based on AllenNLP (Gardner et al. 2017) attempts to extract the answer from the passage in response to option (1) whereas a randomly-chosen art-centered question generated using the Question Generation model based on ProphetNet (Qi et al. 2020) is presented in case of option (2) using the text spans from *PaintTeR*. When the machine asks a question, the user has the choice to answer it or "pass". The number of (unused) art-centered questions available with the machine are shown alongside the "Ask me a question" button with the "score" signifying the number of questions correctly answered by the user listed at the top right corner of the screen.

The choice of QG and RC models used in *ArtQuest* was based on their state-of-the-art performance on SQuAD (Rajpurkar et al. 2016), the widely-used dataset for QG and RC research. Among other datasets available for QG such as HotPotQA (Yang et al. 2018), ComplexWebQA (Talmor and Berant 2018), and MS MARCO (Nguyen et al. 2016), the passages as well as questions in SQuAD closely resemble our setting. We also experimented with other QG models including a model incorporating passage-level context through Gated Self-Attention and Maxout pointer (Zhao et al. 2018) and a Transformer-based model (Lopez et al. 2020) as well as an unsupervised model (Lewis, Denoyer, and Riedel 2019). For RC, we also explored the DrQA model from Facebook AI Research (Chen et al. 2017).

[1]https://commons.wikimedia.org/wiki/File:Ruysch,_Rachel_-_Flower_Still-Life.jpg and https://www.britannica.com/list/25-famous-paintings-to-see-the-next-time-youre-in-florence

Collecting Crowd Annotations

We randomly sampled a subset of 10 articles from the *RAB* dataset and obtained human-annotations for the 310 questions generated from their candidate text spans. We set up our task through the crowdsourcing platform Amazon Mechanical Turk (AMT) following similar dataset collection efforts (Gao et al. 2019; Pan et al. 2020; Wang et al. 2020).

We required the crowdworkers to have greater than 95% HIT approval rate, a minimum of 10,000 HITs, and be located in the United States. The instructions were:
Given a passage related to a painting, a machine-generated question, and a machine-extracted answer for this question, evaluate the triplet with respect to the following four criteria.

1. *Fluent: Is the question grammatically correct, natural sounding and semantically valid?*
2. *Relevant: Does the question enable the viewer understand the subject/content/artistic elements of this painting better?*
3. *Answerable: Does the passage include the answer to the generated question?*
4. *Answer Correctness: Is the machine-extracted answer correct and complete?*

The crowdworker sees the passage and the one of the generated questions (for each assignment) and chooses one of the radio buttons from "Yes, No, or Acceptable/Partially Correct" for each of the four questions listed above. Several examples were provided illustrating the three options for each criterion. We ensure *ethical* and *quality* considerations for our collected data by using the AMT platform since it already ensures the *anonymity* and *privacy* of the crowdworkers. We offered a remuneration of $0.40 per assignment for each worker. Furthermore, the settings for the HIT approval rates, and location of the worker, described previously are set similar to previous QA/QG data collection efforts to ensure the English language skills of the data annotators. We also sample and manually examine subsets of the questions for their annotation quality to ensure the overall quality of the dataset. A total of 31 workers helped in creating our dataset, with about 41% of the workers labeling less than 5 questions each. We hope to make this dataset publicly-available through github since the passages were obtained from the public domain.

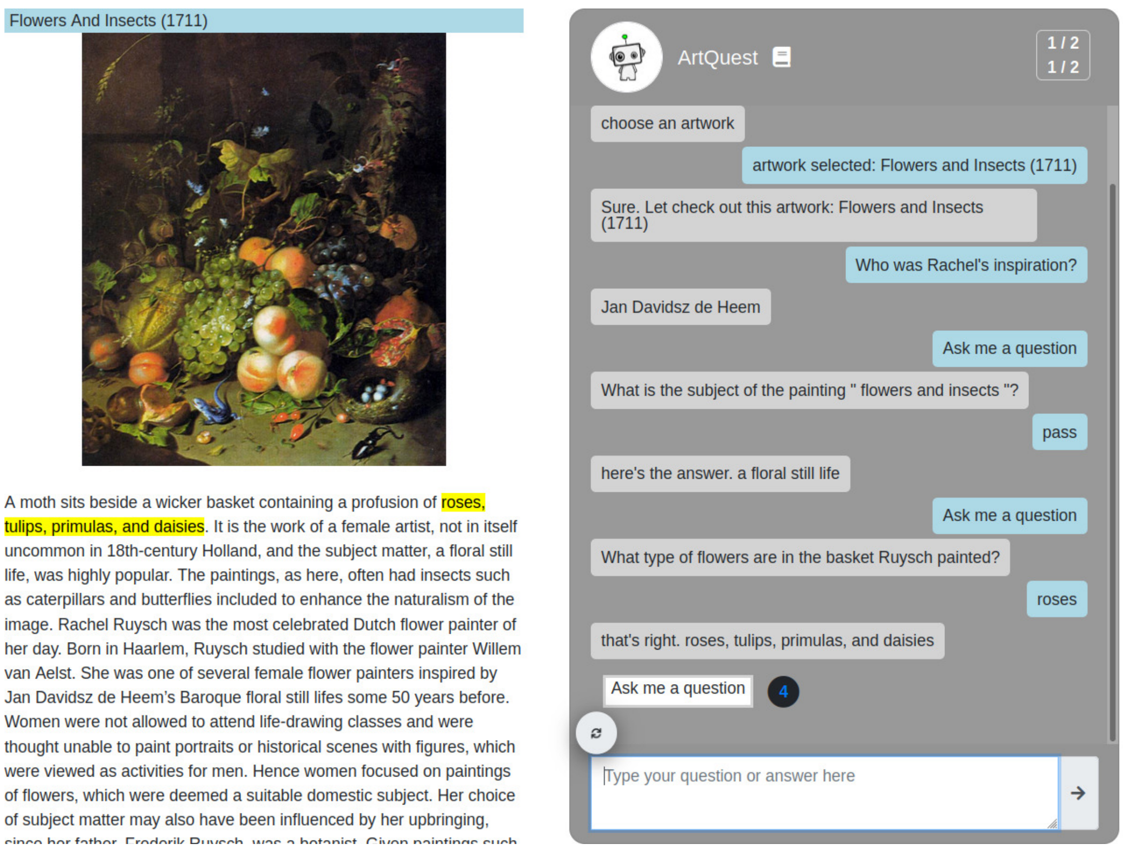

Figure 1: *ArtQuest* in use is shown for the painting “Flowers and Insects” by Rachel Ruysch.[10]

References

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *ACL*.

Gao, Y.; Bing, L.; Chen, W.; Lyu, M.; and King, I. 2019. Difficulty Controllable Generation of Reading Comprehension Questions. In *IJCAI*.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. S. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform.

Lewis, P.; Denoyer, L.; and Riedel, S. 2019. Unsupervised Question Answering by Cloze Translation. In *ACL*.

Lopez, L. E.; Cruz, D. K.; Cruz, J. C. B.; and Cheng, C. 2020. Transformer-based End-to-End Question Generation. *CoRR*, abs/2005.01107.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

Pan, L.; Xie, Y.; Feng, Y.; Chua, T.-S.; and Kan, M.-Y. 2020. Semantic Graphs for Generating Deep Questions. In *ACL*.

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*.

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

Talmor, A.; and Berant, J. 2018. The Web as a Knowledge-base for Answering Complex Questions. In *North American Association for Computational Linguistics (NAACL)*.

Wang, J.; Liu, J.; Bi, W.; Liu, X.; He, K.; Xu, R.; and Yang, M. 2020. Improving Knowledge-Aware Dialogue Generation via Knowledge Base Question Answering. *AAAI*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*.

Zhao, Y.; Ni, X.; Ding, Y.; and Ke, Q. 2018. Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks. In *EMNLP*.