February 13, 2024

Short paper: *For those who don't know (how) to ask: Building a dataset of technology questions for digital newcomers*

Dear AI4ED Chair and Reviewers,

Thank you all for the opportunity to respond to the reviews received.

Review ypKJ

Pros "The paper is well-motivated. The authors seem to rely on a network of resources and organizations that they would use while creating the dataset."

Our answer: Thank you for your kind remarks.

Cons "The paper only discusses what ought to be done but this is not yet tried, and hence there are no real results or evaluations."

Our answer: Thank you for pointing this out. We have just started this project and plan to have some very preliminary results to share in our poster at the workshop.

Other "Several typos need to be corrected."

Our answer: Thank you for pointing this out. The typos have been corrected.

Review GMnR

Pros "in my opinion, targeting learners that don't know what questions to ask seems like a valuable and worthwhile contribution. The authors will likely benefit from the discussion held at the conference, and I hope they receive insightful and encouraging feedback."

Our answer: Thank you for your kind regards, we look forward to getting feedback from others at the workshop.

Cons "while this is only a 2-page proposal, I think the dataset proposal is still very vague."

Our answer: Thank you for this comment, we expanded our discussion of the datasets to add clarity to our proposal:

> *We plan to develop three datasets. The first will come from mining existing online forums on computer-related topics; the second will tap into our extensive records of tutor-patron interaction from the BASIC program; and the third will be developed from interactions at our partners in the Superiorland Library Collective.*

We then proceed to describe the development of these three datasets.

Additional Comments "under 'Digital literacy tutor input', I found the sentence starting 'Volunteers interviewed would ...' to be quite vague; I would like to see this more refined, perhaps the initial investigation on mining the Yahoo dataset will help to focus the interviews."

Our answer: Thank you for this comment, we have adjusted the language to be

more clear. We have also added some discussion about how the different datasets will be used together.

Additional Comments "I would like to see the authors include a discussion about any expected limitations and any potential challenges involved in creating the dataset(s). "

> **Our answer:** Thank you for this suggestion, we have added the section below about limitations and expected challenges to address this note:
>
> *Much like the patron questions we are documenting, there is some uncertainty around the best means to accomplish our tasks. We anticipate challenges in ensuring diversity of the question topics and determining how best to utilize this dataset for generation.*
>
> *We have identified the potential for sampling bias in our proposed data sources. First, the interactions from Yahoo Answers involve people who are already online and therefore have a level of technological familiarity beyond some of our target audience.. Similarly, our in-person interactions come from a patron demographic with the means and resolve to show up and ask questions. Finally, our tutors are far removed from being in the position of not understanding core concepts necessary to ask questions. By combining these different sources and using them to prompt additional questions from the tutor group, we hope to mitigate some of this sampling bias. We also have identified the possibility for overfitting our question classification model to detect aspects of the data, such as specific language used by a patron or misspellings only present in the Yahoo Answers data.*
>
> *In order to train the model to learn tutors' dialogical flow, the training data must include full transcripts engineered into a chat format. As our tutoring notes and Yahoo Answers formats are not in dialog format, text data augmentation will be required. Care must be taken in this process to ensure that resulting dialog accurately represents authentic tutor-patron interactions.*

Additional Comments "in 'Question classification', I wonder if 'awareness' is better than 'understanding' within this context? Is understanding of a topic necessary in order to ask appropriate/on point questions?"

> **Our answer:** We have considered this comment and we think that "understanding" is closer to what we are trying to quantify than "awareness". Complete understanding of a topic is definitely not necessary to ask a good question, but just awareness of a topic could probably lead to questions with poor understanding.

Best regards, the authors
Evan Lucas
Kelly S. Steelman
Leo C. Ureel
Charles Wallace