Name: Bushra Rahman
Class: CS 4395.001
Assignment: ACL Paper Summary

Background Information
The title of the paper is "It is AI's Turn to Ask Humans a Question: Question-Answer Pair Generation for Children's Story Books". Its authors are Bingsheng Yao (Rensselaer Polytechnic Institute, 146 citations), Dakuo Wang (IBM Research, 2580 citations), Tongshuang Wu (University of Washington, 1508 citations), Zheng Zhang (University of Notre Dame, 270 citations), Toby Jia-Jun Li (University of Notre Dame, 811 citations), Mo Yu (WeChat AI, Tencent, 10167 citations), and Ying Xu (University of California Irvine, 615 citations). The paper has 5 citations on Google Scholar. It can be accessed online at https://aclanthology.org/2022.acl-long.54/.

Paper Summary
The researchers of this paper address the "underexplored" field of QAG (question-answer generation), for the purpose of RC (reading comprehension) in children (p. 1). Rather than a QA (question-answer) system, which is designed to use its knowledge base to answer questions, a QAG system is designed to use its knowledge base to determine potential answers to questions, and generate appropriate questions. Systems like this are applicable in the field of education because they can "support teachers to efficiently construct assessment questions (and its correct answer) for the students" (p. 1). A QAG system focused on reading comprehension could have a variety of real-world applications, including educational software for children (p. 8).

The researchers built their QAG system on an RC dataset called FairytaleQA, which contains 10,580 QA pairs from 278 narrative text passages of classic children's fairytales, targeting an age demographic from kindergarten to eighth grade. It was created to "complement the lack of a high-quality dataset resource for the education domain" (p. 3). While other RC datasets for QAG exist, they are designed to use shallow approaches like pattern-matching instead of focusing on the underlying narrative elements of the text. Conversely, FairytaleQA's QA set focuses on commonly-taught narrative elements like character, setting, feeling, action, prediction, etc. Each of these terms has a clear definition in Fairytale QA, and the questions generated center around those definitions — for example, a prediction-focused question for a story about a snowman may ask, "What will happen to the snowman when the weather changes?", to which the answer is "he will thaw", with both the question and the answer being extracted from the text (p. 12).

FairytaleQA is incorporated into the researchers' QAG system on a pipeline composed of three steps: rule-based answer extraction, NN-based question generation, and NN-based ranking (p. 5, Fig. 2). First, the text is scraped for potential answers. This is done through the use of a heuristic framework, which is designed to analyze the text for answers that fit the purpose of reading comprehension. For example, a line of text in the story about the snowman that talks about him thawing would be recognized by the framework as being a potential answer to a reading comprehension question. Next, a language model built on FairytaleQA is used to generate questions that are specifically intended to test for the narrative element present in the answer. So, for the example with the snowman, FairytaleQA would be used to recognize the narrative element of prediction in the answer, and would then generate an appropriate reading

comprehension question focused on testing prediction. Finally, a ranking system is used to rank the QA pairs by classifying which pairs most closely align with ground truth QA pairs created by education experts (p. 6). The entire pipeline altogether combines rule-based and neural-network (NN) based approaches in order to complement the limitations of each approach (p. 3).

The researchers evaluated their QAG system using a combination of metrics and human evaluation. Every QA pair was given a precision score, from which the average score was calculated and compared with the score of the ground truth (p. 7). These scores were compared against those of 2 existing state-of-the-art RC QAG systems: an automatically-generated baseline RC dataset called PAQ, and a 2-step system that was also built on FairytaleQA (p. 2). The results found that the researchers' QAG system outperformed both. The fact that the researchers' QAG system outperformed the 2-step system indicated that the ranking step incorporated into their own system was essential for quality control (p. 7). Human evaluation was also used to analyze the quality of the system. Five participants read a random selection of 7-10 literary passages and accompanying QA pairs generated by either the PAQ system, the researchers' system, or the ground truth as given by domain experts. They then rated the QA pairs based on a number of criteria pertaining to readability and relevancy. The results found that the ground truth QA pairs consistently outranked the other QA pairs in all of the criteria, but also that the researchers' model consistently outperformed the baseline model (p. 8).

Paper Importance and Contributions

The research in this paper is constructive because it focuses on the underexplored areas of both QAG models and machine learning applications for children's education. The researchers conclude their paper with a demo of a real-world software application they built using their system, even suggesting use case scenarios of how such software could be integrated into children's curriculum (p. 8). The application focuses on interactive engagement, thorough comprehension, and personalized performance evaluations. The research used to build this application illustrates how machine learning and natural language processing can be used to create constructive tools in fields that could greatly benefit from it, like education.