Question Generation for Reading Comprehension Assessment by Modeling How and What to Ask

Anonymous ACL submission

Abstract

Reading is integral to everyday life, and yet learning to read is a struggle for many young learners. During lessons, teachers can use comprehension questions to increase engagement, test reading skills, and to improve retention. Historically such questions were written by skilled teachers, but recently language models have been used to generate comprehension questions. However, many existing Question Generation (QG) systems focus on generating extractive questions from the text, and have no way to control the type of the generated question. In this paper, we study QG for reading comprehension where inferential questions are critical and extractive techniques cannot be used. We propose a two-step model (HTA-WTA) that takes advantage of previous datasets, and can generate questions for a specific targeted comprehension skill. We propose a new reading comprehension dataset that contains questions annotated with story-based reading comprehension skills (SBRCS), allowing for a more complete reader assessment. Across several experiments, our results show that HTA-WTA outperforms multiple strong baselines on this new dataset. We show that the HTA-WTA model tests for strong SRS by asking deep inferential questions.

1 Introduction

Reading is an invaluable skill, and is core to communicating in our digital age. Reading also supports other forms of development; when children read, it sharpens their memory, and improves social skills (Halliday, 1973; Mason, 2017). Yet, statistics show that one out of five children in the U.S. face learning difficulties (Shaywitz, 2005), especially in reading (Cornoldi and Oakhill, 2013). The coronavirus pandemic beginning in 2020 had a huge impact on the early reading skills of many children, and threatens to leave a lasting impact on a whole generation of young readers (Gupta and Jawanda, 2020).

The pandemic forced many children to learn online, putting in sharp relief the need for effective online education platforms. In particular, reading games have become popular, and can help fill the gap when teachers cannot read in person with students. These platforms present students with short passages and associated comprehension questions. These questions are key to assessing a reader’s comprehension of a passage, and can also enhance learning (Chua et al., 2017). But, writing diverse and engaging comprehension questions is no trivial task.

Teachers need to generate new comprehension questions whenever they incorporate new text into a curriculum. New text helps to keep material fresh and topical, and can allow teachers to customize lessons to the interests of a particular student cohort. After finding such custom reading material, teachers must write new comprehension questions to evaluate several reading aspects of comprehension (e.g. understanding complex words, recalling events, etc.).

Thus, to improve the educational process, and lighten the load on teachers, we need tools to automate Question Generation (QG): the task of writing questions for a given passage. Generated questions can be either inferential or extractive questions. Extractive questions can be answered using only information stated in the text, whereas inferential questions require additional information or reasoning. Previous work focused on this aspect of the questions in reading comprehension and discarded the comprehension skills (e.g. close reading, predicting, figurative language, etc.).

We take inspiration from continual learning (Parisi et al., 2019), which orders a set of learning tasks to improve model performance. We begin by training a model on the general task of QG (How to ask: HTA), and follow with our task of interest: generating a targeted question of a particular type (What to ask: WTA).
This paper focuses on the generation of questions for story-based reading comprehension skills (SBRCS), which are varied and cover many aspects of reading comprehension. We create a QG dataset for SBRCS. Although our aim in creating this dataset is to enrich educational applications, this dataset can be considered as a source for general QG and question answering (QA) systems in NLP.

Our focus here is to build a question generator without answer supervision as the case in a real-life application, where a story only will be given as input. This is a challenging task, as many different questions can be generated from a story when there is no answer supervision. QG with answer supervision is another prevalent research line in the literature (Zhao et al., 2018; Ma et al., 2020; Wang et al., 2020; Chen and Xu, 2021).

The contributions in this work are as follows:

- We build a novel QG dataset for SBRCS. The dataset contains advanced reading comprehension skills extracted from stories.

- We propose a two-steps method to generate skill-related questions from a given story. The method takes advantage of previous datasets to improve generalizability, and then, teaches a model how to ask predefined styles of questions.

- We demonstrate the efficiency of the proposed method after extensive experiments, and we investigate its performance in a few-shot learning setting.

The rest of the paper is structured as follows. In the next section, we present an overview of the literature work. In Section 3, we describe how we built our dataset. Section 4 describes the proposed methodology. The experimental setting is presented in Section 5. The results and the analysis are presented in Section 6. Finally, we draw some conclusions and possible future work for this study.

2 Related Works

QG has progressed rapidly due to new datasets and model improvements. Many different QG models have been proposed, starting for simple vanilla Sequence to Sequence Neural Networks models (seq2seq) (Du et al., 2017; Zhou et al., 2017; Yuan et al., 2017) to the more recent transformer-based models (Dong et al., 2019; Chan and Fan, 2019; Varanasi et al., 2020; Narayan et al., 2020; Bao et al., 2020). Some QG systems use manual linguistic features in their models (Harrison and Walker, 2018; Khullar et al., 2018; Liu et al., 2019a; Dhole and Manning, 2020), some consider how to select question-worthy content (Du and Cardie, 2017; Li et al., 2019; Scialom et al., 2019; Liu et al., 2020), and some systems explicitly model question types (Duan et al., 2017; Sun et al., 2018; Kang et al., 2019; Zhou et al., 2019). The last group focused only on generating questions that start with specific interrogative words (what, how, etc.).

QG has been used to solve many real-life problems. For example, QG in conversational dialogue (Gu et al., 2021; Shen et al., 2021; Liu et al., 2021b) where models were taught to ask a series of coherent questions grounded in a QA style, QG based on visual input (Mostafazadeh et al., 2016; Shin et al., 2018; Shukla et al., 2019), and QG for deep questions such as mathematical, curiosity-driven, clinical, and examination-type questions (Liyanage and Ranathunga, 2019; Scialom and Staiano, 2020; Yue et al., 2020; Jia et al., 2021).

3 Data

Despite the recent efforts for building reading comprehension QA datasets, to the best of our knowledge, none of the available datasets explored SBRCS. Questions in previous datasets ask only either inferential or extractive questions from a given passage/story. Rogers et al. (2020), developed questions with general reasoning types based on text from news and blogs. We believe that those texts sources are not rich enough to examine reasoning skills. Advanced reasoning skills (e.g. Figurative Language) are usually used in stories to assess comprehension skills. In the following, we will show how we built our dataset. Table 5 gives an overview of the dataset. In Appendix A, Table A.1, we provide further dataset statistics.

3.1 Dataset Design

3.1.1 Stories Collection

Our stories (passages) are multi-genre, self-contained narratives. This content variety leads annotators towards asking non-localized questions...
that test for more advanced reading comprehension skills. The stories are generated using several resources: 1) acquired from free public domain content (Gutenberg Project\(^2\)), 2) partnerships with a publishing house (Blue Moon Publishers\(^3\)) and an educational curriculum development foundation (The Reimagined Classroom\(^4\)), and 3) authored by two professional writers, (the majority of the stories are from this last category). To provide good lexical coverage and diverse stories, we choose to write and collect stories that come from a varied set of genres (e.g. science, social studies, fantasy, fairy tale, historical fiction, horror, mystery, adventure, etc.). In total, we collect 726 multi-domain stories. The stories’ lengths range from a single sentence to 113 sentences.

### 3.1.2 Questions and Comprehension Skills

Previous comprehension question datasets focused on either inferential or extractive (literal) questions. Although these questions assess comprehension skills, they do not provide fine-grained evaluation of the reader comprehension. Thus, to build a more comprehensive list of question types, we started by reviewing curriculum documents available from Columbia University Teacher’s College Readers\(^5\) and Writers Workshop Program\(^6\). Then, we compiled a list of SBRCS, which we then expanded to include additional skills based on school teachers’ recommendations. Our final list contains the following skills:

1. **Basic Story Elements**: Can the reader identify the story’s main characters and setting?

2. **Character Traits**: Can the reader identify the traits attributable to certain characters in the story (e.g. character feelings, physical attributes)?

3. **Close Reading**: Can the reader extract the text span in a story where the author best describes or explains a key point?

4. **Figurative Language**: Is the reader able to recognize the implied meaning of a sentence?

5. **Inferring**: Can the reader infer what happened in between scenes if the time in-between is not explicitly described?

6. **Predicting**: Can the reader find textual clues and use them to guess what would happen next?

7. **Summarizing**: Is the reader able to recognize the main literary elements of the characters, the events, the problem, and the solutions?

8. **Visualizing**: Can the reader visualize scenes in her/his head to fully comprehend the story?

9. **Vocabulary**: Can the reader identify the right meaning of a word within a context when the word has multiple possible definitions?

Note that some of these SBRCS are prerequisites for others. For instance, the predicting skill may depend on the reader’s ability to identifying character attributes and to summarize story elements. In Section A.2, we present further details for each skill type.

With our list of SBRCS as a guide, we wrote question-answer pairs for each story. Given the difficulty of the task, we needed a large number of trained content writers to build the required questions. Each written question should fall into one of the mentioned skills, and obviously, should meet the educational goal. For that, a total of 25 professionals contributed to the writing process (18 teachers, 7 graduate students). We chose not to use crowdworkers (e.g. Amazon Mechanical Turk) to ensure high-quality and educationally-appropriate questions. To verify the quality of the generated content, a second team member reviews each question-answer pair before adding them to the dataset. In addition to annotating questions with a skills label, our content writers annotate each question as either *Literal* or *Inferential* question types. This information is important to measure the comprehension performance of the reader on each question type. Overall, we generate 4K question-answer pairs, with an average of 5.5 pairs per story.

### 3.2 Additional Data

In addition to the collected dataset, we use two well-known datasets, SQuAD and CosmosQA. We choose these two datasets because of their large size, and their focus on literal or inferential questions.

\(^2\)https://www.gutenberg.org/

\(^3\)https://bluemoonpublishers.com/

\(^4\)https://www.reimaginedclassroom.com/

\(^5\)https://www.tc.columbia.edu/curriculum-and-teaching/literacy-specialist/the-reading–writing-project/

\(^6\)https://readingandwritingproject.org/
Table 1: Collected dataset’s statistics. There are 726 stories, which can have questions from multiple skill types (described in Section 3.1).

<table>
<thead>
<tr>
<th># Stories</th>
<th>Character Tr.</th>
<th>Close Rec.</th>
<th>Figure Ex.</th>
<th>Infering</th>
<th>Predicting</th>
<th>Summarizing</th>
<th>Visualizing</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>269</td>
<td>280</td>
<td>448</td>
<td>219</td>
<td>449</td>
<td>152</td>
<td>360</td>
<td>153</td>
<td>403</td>
</tr>
<tr>
<td>390</td>
<td>415</td>
<td>719</td>
<td>292</td>
<td>695</td>
<td>162</td>
<td>560</td>
<td>163</td>
<td>604</td>
</tr>
<tr>
<td>274</td>
<td>120</td>
<td>606</td>
<td>108</td>
<td>16</td>
<td>111</td>
<td>464</td>
<td>36</td>
<td>168</td>
</tr>
<tr>
<td>115</td>
<td>295</td>
<td>113</td>
<td>148</td>
<td>679</td>
<td>151</td>
<td>96</td>
<td>127</td>
<td>436</td>
</tr>
</tbody>
</table>

**SQuAD** A reading comprehension dataset, consists of questions created by crowdworkers on a set of Wikipedia articles that cover a large set of topics (from musical celebrities to abstract concepts), where the answer to every question is a span from the corresponding reading passage (Rajpurkar et al., 2016). This dataset can be considered as an extractive QA dataset. It is one of the largest QA datasets in the literature. In this work, we use SQuAD 2.0 version with discarding the questions that have no answers. The size of the dataset is 100K paragraph/question/answer triplets.

**CosmosQA** It is another reading comprehension dataset consisting of 35.6K paragraph/question pairs that require commonsense-based reading comprehension. It is a collection of people’s everyday narratives, and it asks questions about the likely causes of events that require reasoning (Huang et al., 2019). We discard questions that have no answers in this dataset, resulting in 28K paragraph/question/answer triplets.

### 4 Methodology

Given the fact that including more data in a reading comprehension system is important for generalization (Chung et al., 2018; Talmor and Berant, 2019), and given that our created dataset has the SBRCS which are missing in previous datasets, we propose a two-steps method to generate skill-related questions from a given story: HTA followed by WTA. HTA teaches the model the typical format for comprehension questions using large previously released datasets. These previous datasets are not annotated with the question types outlined in Section 3.1, but the HTA phase allows us to take advantage of those datasets. WTA guides the model to generate questions to test the specific comprehension skills enumerated in Section 3.1. Thus, in HTA, we train (fine-tune) a model on large QG datasets, and then, we further train the model to teach the model what to ask (WTA). For the generation model, we use the pre-trained Text-to-Text Transfer Transformer T5 (Raffel et al., 2020), which closely follows the encoder-decoder architecture of the Transformer model (Vaswani et al., 2017). T5 is a SOTA model on multiple tasks, including QA.

#### 4.1 How to Ask (HTA)

Previous works showed that incorporating more data when training a reading comprehension model improves performance and generalizability (Chung et al., 2018; Talmor and Berant, 2019). However, we cannot incorporate previously released datasets with our new one, as they do not include compatible question skills information. However, they do contain many well-formed and topical questions. Thus, we train a T5 model on SQuAD and CosmosQA datasets to teach the model how to ask questions.

Previous neural question generation models take the passage as input, along with the answer. However, encoders can pass all of the information in the input to the decoder, occasionally causing the generated question to contain the target answer. Since the majority of the questions in our created dataset are inferential questions, the answers are not explicitly given in the passages (unlike extractive datasets). Thus, we feed the stories to the encoder, but withhold the answers. Unlike previous systems, we then train the model to generate the questions and answers. We propose this setting to generate fewer extractive questions. During our experiments, we evaluated the effect of excluding the answers, and we found them useful to the system.

In Figure 1 we show the input-output format of the model. The encoder input is structured as `<STORY_TEXT> </s>`, where `<s>` is the end-of-sentence token. The decoder generates multiple question-answer pairs as `<QUESTION_TOKENS>1 ` `<as> ` `<ANSWER_TOKENS>1 ` `<sp> ` ... `<QUESTION_TOKENS>n`,
Figure 1: Input and output format of the How to Ask (HTA) model.

\(<\text{as}>\) <ANSWER_TOKENS>, \(</s>\), where \(<\text{as}>\) separates a question from its answer, and \(<\text{sp}>\) separates a question-answer pair from another. The model can generate more than one question-answer pair. We prepare the data to include all of a passage’s question-answer pairs in the decoder. Some passages include single question-answer pair, and some passages have up to fifteen pairs.

4.2 What to Ask (WTA)

QG models take a passage/story as input and generate a question. The type of generated question is not controlled and is left for the system to decide it. Thus, the generated question is usually undesired question. Thus, in order to control the style of the generated question, the system needs an indication about the skill that the system is expected to generate a question for. Liu et al. (2020) proposed a way to control the style of the generated questions (e.g. what, how, etc.). The authors built a rule-based information extractor to sample meaningful inputs from a given text, and then learn a joint distribution of <answer, clue, question style> before asking the GPT2 model (Radford et al., 2019) to generate questions. However, this distribution can only be learned using an extractive dataset (e.g. SQuAD); the model cannot learn to generate inferential questions.

To control the skill of the generated question, we use a specific prompt per skill, by defining a special token \(<\text{SKILL_NAME}>\) corresponding to the desired target skill. This helps us to control what to extract from the pretrained model. Thus, the encoder takes as input \(<\text{SKILL_NAME}>\) and \(<\text{STORY_TEXT}>\), where \(<\text{SKILL_NAME}>\) indicates to the model for which skill the question should be generated (see Figure 2). The data format in the decoder is similar to the one in the HTA step, but here the model generates a single question-answer pair. As a result, the encoding of the \(<\text{STORY_TEXT}>\) will be based on the given \(<\text{SKILL_NAME}>\). In this way, the model encodes the same story in a different representation when a different \(<\text{SKILL_NAME}>\) is given. A similar technique was used in the literature to include persona profiles in dialogue agents to produce more coherent and meaningful conversations (Scialom et al., 2020).

5 Experiments

5.1 Decoding Method

Decoding strategies are crucial and directly impact output quality. In general, Beam Search (Reddy, 1977) is the most common algorithm, in addition to some other sampling techniques such as Nucleus sampling (Top-p) (Holtzman et al., 2019). In Beam Search, the output of a model is found by maximizing the model probability. On the other hand, Nucleus sampling selects the smallest possible set of tokens whose cumulative probability exceeds the probability \(p\). Experimentally, we found that using the top-p (\(p=0.9\)) algorithm yields the best results in terms of the used scoring metrics, thus we use it in all of our experiments.

5.2 Evaluation Metrics

QG often uses standard evaluation metrics from text summarization and machine translation (BLEU (Papineni et al., 2002), ROUGE, METEOR, etc.). However, such metrics do not provide an accurate evaluation for QG task (Novikova et al., 2017), especially when the input passage is long (and many acceptable questions that differ from the gold question can be generated). Thus, to alleviate shortcomings associated with n-gram based similarity metrics, we use BLEURT (Sellam et al., 2020), which is state-of-the-art evaluation metric in WMT Metrics shared task. BLEURT is a BERT-based
model that uses multi-task learning to evaluate a generated text by giving it a value in between -2 to 1. In our experiments, we consider BLEURT as the main metric for the evaluation. We also report standard MT metric BLEU (1-4 ngrams), and perform an additional manual evaluation.

Manual evaluation is required in our collected dataset, because teachers wrote a single question per skill for a given story, where the model might generate other possible questions for the same skill.

5.3 Implementation Details

We fine-tune the T5 model (base) using the Adam optimizer with a batch size of 8 and a learning rate of $1e^{-4}$. We use a maximum sequence length of 512 for the encoder, and 128 for the decoder. We tested the T5-large model, but we did not notice any improvements considering BLEURT metric. We train all models for a maximum of ten epochs with an early stopping value of 1 (patience) based on the validation loss. We use a single NVIDIA TITAN RTX with 24G RAM.

For HTA, we validate on a combined version of the validation sets from both datasets (SQuAD and CosmosQA). Regarding the collected dataset validation set, we use stratified sampling: we took a random 10% of stories from each skill since the dataset is unbalanced. We apply the same strategy with the test set but with a value of 20%.

5.4 Baselines

To evaluate the performance of our model, we use a set of models that showed state-of-the-art results on several datasets. We obtain the results of those models by running their published GitHub code on our collected dataset. For all of the following baselines, we use SQuAD, CosmosQA, and the collected dataset for training and we test on the test part of the collected dataset:

- Vanilla Seq2seq (Sutskever et al., 2014): a basic encoder-decoder sequence learning system for machine translation. This model takes the story as input and generates a question.
- NQG-Seq (Du et al., 2017): another Seq2seq that implements an attention layer on top of a bidirectional-LSTM encoder. The authors use two encoders, one to encode the sentence that has the answer, and another to encode the whole document. The model then is trained to generate questions.
- NQG-Max (Zhao et al., 2018): a QG system with a maxout pointer mechanism and gated self-attention LSTM-based encoder to address the challenges of processing long text input. This model takes a passage and an answer as input and generate a question. The answer must be a sub span of the passage.
- CGC-QG (Liu et al., 2019a): a Clue Guided Copy network for Question Generation, which is a sequence-to-sequence generative model with a copying mechanism that takes a passage and an answer (as a span in the text) and generate the question. The text representation in the encoder (GRU network) is represented using a variety of features such as GloVe vectors, POS information, answer position, clue word, etc.
- AnswerQuest (Roemmele et al., 2021): a pipeline model that uses as a first step a previous model (Yang et al., 2019) to retrieve the relevant sentence that has the answer from a document. And then, the sentence is fed to a transformer-based sequence-to-sequence model that is enhanced with a copy mechanism.
- One-Step: a baseline that uses T5 model trained with all data in one step instead of having separate HTA and WTA steps. Because there is only a single step, the skill name is not included in the encoder’s input.
- T5-WTA: the WTA model trained using T5 model as a seed model. The HTA training step is not used here. We use this baseline to evaluate the effect of training WTA using HTA.

For all of the previous baselines that require the answer to be a sub-span in the passage, we use the semantic text similarity method that was proposed in (Ghanem et al., 2019) to retrieve the most similar span in the passage. The method extracts several ngrams features from a claim and text spans, and then compute cosine similarity to get the most similar span. In this work, we replace the ngrams features of a text with embeddings extracted from

---

\*We used the unofficial implementation in this GitHub repo: https://github.com/seanie12/neural-question-generation
RoBERTa model (Liu et al., 2019b). This process has been done on the inferential questions as their answers are not clearly given in the text.

6 Results and Analysis

Table 2 presents the results of the proposed HTA-WTA method with the baselines. We can see that out of the baselines, T5-WTA performs best in terms of BLEURT score (-1.17), followed by QG-Max with a value of -1.18. Given its high BLEURT score, it is surprising that T5-WTA model has low BLEURT-4. This implies that the generated questions use rich vocabulary, making them different from the gold in terms of overlapping ngrams, but semantically similar leading to higher BLEURT score.

As shown in the table, HTA-WTA’s BLEURT score outperforms all of the previous QG models by a noticeable margin, showing that including the skill name information plays an important role in generating the intended questions. Also, training on more QG datasets improves the performance.

Regarding the generated questions type, in Table 3 we show the performance of the T5-based models per question type (inferential and literal). Though One-Step and HTA-WTA models were trained on the same amount of data, the results show that HTA-WTA model clearly performs better than the One-Step model, especially on inferential questions. We see a similar scenario when comparing One-Step and T5-WTA models, yet, the gap is smaller. In general, we can notice that the performance gaps for the inferential questions are larger than the literal ones. Thus, we can conclude that HTA-WTA is generating more correct inferential questions, which is challenging. This experiment concludes that transformers-based models are capable of asking questions beyond the literal meaning of the text. This confirms what was shown by Liu et al. (2021a) regarding the skills that language models can acquire. Additionally, as some training questions directly quote text from the given story, the T5 model was able to learn how to quote the proper segment of the passage when generating questions.

The One-Step model performs similarly to the baselines, although it has been trained using the T5 model and on all three datasets. This may be due to the fact that we did not include the skill name in the encoder, which guides the model to generate skill related questions. To better understand the differences between the outputs of One-Step and HTA-WTA models, we used human evaluation. This evaluation is to assess the quality of the generated question in terms of 1) Answerability (Ay), 2) Fluency (Fy), and 3) Grammaticality (Gy) categories, following Harrison and Walker (2018); Azevedo et al. (2020). We include these three criteria as questions may have high Fluency and Grammaticality scores, but not be answerable.

We select a sample of 110 story-question pairs from the test dataset, for both models. Then, we perform a human evaluation using crowdworkers on Amazon Mechanical Turk. We use a "master" qualification criteria to restrict the participation of workers in our evaluation study to those who have a high historical HIT accuracy, and workers are required to be located in an English speaking country. Workers received $0.41 USD for completing each HIT. Each HIT was answered by three workers. Each worker needs to read the story, and provides ratings (1-5, low to high) for the generated questions, and the three criteria. Table 4 shows the average rating assigned by the workers for the 3 criteria.

Originally, we hypothesized that adding the skill name to the input would force the model to formulate a specific SBRCS question, even if it is not applicable to the current passage. Omitting the skill name may allow the model score high values as it has been left to decide the question. The results show that both models are similar in terms of the given categories, except that HTA-WTA performs slightly better in all of the three categories. However, these results refute our claim and show that adding the skill information makes the model generates slightly better questions in terms of quality.

6.1 Impact of Skill Name Token

In order to quantify the impact of skill name in the input, we do another human manual evaluation to measure how accurate both models are when it comes to the generated question skill. Thus, we ask two professional persons who were involved in the annotation process to assign skill names to the generated questions of both One-Step and HTA-WTA models. We use the same question sample that was used in the previous human evaluation experiment. Few annotation conflicts were found and were solved after a discussion. We evaluate the results using accuracy (see Table 4). The result for One-Step model is 0.16, and 0.8 for HTA-WTA model. We can clearly see a large gap in accuracy between both models, and this becomes clear with
the skills that have a low number of instances in the dataset (e.g., Figurative Language, Precision, etc.). Table 6 in Appendix A.3 presents the F1 scores per skill name. We also notice that HTA-WTA model performed perfectly on the given sample of Predicting and Figurative Language (F1 is 1.0 for each skill). This is an interesting result given that the type of the questions for both skills is inferential, which is harder to generate compared to the extractive questions.

### 6.2 Few-Shot Generation

The process of manually writing questions to assess humans’ skills is difficult. In some stories, professional writers find obstacles in writing questions for some skills as those skills require high attention and advanced reasoning skills to be written. We can see that in our own dataset, as some skills have fewer questions (e.g., Predicting, Visualizing, etc.). Thus, in this experiment, we evaluate the performance of HTA-WTA model when we inject a low percentage of the skills’ instances into the training set. This experiment will simulate the case when training a model on a dataset that contains few skills’ instances. We use the stratified sampling technique when sampling fewer instances from the collected dataset. Figure 3 shows that injecting only 10% of the data led to a boost in performance of 0.22 (BLEURT). The result at 10% (-1.17) exceeds the results of most of the baselines and is similar to T5-WTA and QG-MAX models when trained on all the datasets (see Table 2). In Table A.4 in the appendix, we present the results considering other models and metrics. In most cases, the performance gradually improves as data grows. We notice a small drop when we move from 10% to 30%. This behaviour was previously reported by Stappen et al. (2020). Further research is needed to investigate the causes of this behaviour.

### 7 Conclusion and Future Work

In this paper, we presented a new reading comprehension dataset to assess reading skills using stories. Unlike previous datasets that focused on either inferential or extractive questions, our dataset has nine different SBRCs, each contains inferential and extractive questions. In addition to that, we proposed HTA-WTA model which uses two-steps fine-tuning processes to take advantage of previous datasets which have different question formats, and to learn how to ask skill-related questions. We evaluated the model on the collected dataset and compared it to several strong baselines. Our extensive experiments showed the effectiveness of the model. Additionally, HTA-WTA is able to generate high quality questions when only 10% of the dataset is used (~240 instances). In future work, we plan to extend our dataset with additional skills, and to investigate how our model can be integrated into online educational platforms.
References


Table 5: Collected dataset’s statistics. There are 726 stories, which can have questions from multiple skill types.

A Appendix

A.1 Full Dataset Details

A.2 Further Details on Skills

1. Basic Story Elements: Determining what are the main story elements is one of the comprehension skills to assess the reader understanding. Using this skill, we can understand whether the reader is able to identify the main characters and environment settings of the stories.

2. Character Traits: Identifying permanent traits that can be assigned to characters or describe character development. For instance, knowing what most likely X character felt during the story, recognizing facts about X, identifying main adjectives that X has, etc.

3. Close Reading: Identifying the place in a story where the author best describes or explains a key point. Also, it includes questions to identify the purpose of a quote or a sentence. This skill requires advanced reading comprehension ability from the reader since its answers cannot be extracted directly from the story text, where inferential skills are needed.

4. Figurative Language: Figurative language is common in stories as it makes ideas and concepts easier to visualize by the reader. Also, it is an effective way of conveying an idea that is not easily understood. With this skill, we examine the reader ability of recognizing the implicated meaning of a sentence or a type of figurative language.

5. Inferring: Writers sometimes jump into the action or skip forward in their stories. Good readers must infer what happened in between scenes if the time in-between is not explicitly detailed. In addition, readers must infer their characters’ emotions if their characters do not share those aloud.

6. Predicting: Predicting involves guessing what will happen next. It is different from inferring; inferring is guessing what is happening now or what happened before. Good readers do not let books passively happen to them, they work to "solve" the story before it reaches its end by finding clues and using them to guess what will happen next or to guess how the conflict will be resolved.

7. Summarizing: Consolidating a text into a precise synopsis of only the most key information. Summarizing skill contains the main literary elements of the characters, the problem, and the solutions. Key events from the beginning, middle, and end are included in a summary.

8. Visualizing: This skill requires readers to visualize scenes in their heads to fully comprehend the story. It can assess readers ability of imagining specific events or elements in the stories. An example of visualizing questions is: What do you imagine when reading this sentence "quote"?

9. Vocabulary: Identifying the meaning of unfamiliar words in the text is a key skill for readers to fully comprehend the story. In this skill, the reader should identify the right meaning of a word within a context when the word has multiple possible definitions. Additionally,
the reader should be able to identify vocabulary based questions related to identifying synonyms, antonyms, homophones, compound words, and word types (e.g. noun, verb, etc.).

A.3 Manual Evaluation Results of Questions’ Skills

In Table 6, we show the fined-grained results per skill name after the manual labeling experiment for the generated questions from both One-Step and HTA-WTA models.

A.4 Few-Shot Question Generation Results

In Table 7, we show the few-shot experiment’s results considering both scoring metrics (BLEU, and BLUERT). We do not experiment with One-Step model as we need to sample SQuAD and CosmosQA datasets when we sample the collected data; it is hard to set up a fair comparison here as, for instance, sampling 10% of SQuAD dataset is larger than the whole collected dataset.

A.5 Samples of the Generated Questions

In this section, we list some random examples from HTA-WTA model for inferential questions:

- Story: “The Line 1 Toronto train was a subway like many others you’ve seen. He rocketed down Yonge Street, around the Union loop, and rattled off towards Vaughn. At Vaughn he’d let out a loud, hissing sigh and a clanking sort of grunt, then reverse and do the whole thing backwards all over again. He liked his transit union job well enough, but he couldn’t help thinking about the lights at the end of his tunnels. No matter how long he’d been running, or how much he wished for anything else, that little hopeful point of light always turned out to be just one more dirty subway platform.”

  Generated Figurative Language question: “Reread this sentence: "He rocketed down Yonge Street, around the Union loop, and rattled off towards Vaughn." Which figurative language technique is being used here?”

  Generated answer: “Alliteration”.

- Story: “"The map says left", said Bri. "But my heart says right!" cried Rob. "Is your heart full of hidden treasure?" asked Bri. "Yes." Rob replied. "At least, that’s what my mom says.""

  Generated Inferring question: “Why do you think Bri’s heart says "But my heart says right!"?”

  Generated answer: “Because she thinks she has found something”.

- Story: “Mary looked at it, not really knowing why the hole was there, and as she looked she saw something almost buried in the newly-turned soil. It was something like a ring of rusty iron or brass and when the robin flew up into a tree nearby she put out her hand and picked the ring up. It was more than a ring, however; it was an old key which looked as if it had been buried a long time. Mistress Mary stood up and looked at it with an almost frightened face as it hung from her finger. "Perhaps it has been buried for ten years," she said in a whisper. "Perhaps it is the key to the garden!"

  Generated Vocabulary question: “Reread this sentence: "Perhaps it has been buried for ten years" What is the correct definition of the word "frightened" as it is used here?"

  Generated answer: “Scared”.

Generated answer: “Because she thinks she has found something.”
## Table 6: F1 score results per skill name.

<table>
<thead>
<tr>
<th>Model</th>
<th>Basic Sp.</th>
<th>Character Tr.</th>
<th>Close Res.</th>
<th>Figurative Ls.</th>
<th>Inferring</th>
<th>Predicting</th>
<th>Summarizing</th>
<th>Visualizing</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>#instances</td>
<td>12</td>
<td>8</td>
<td>23</td>
<td>7</td>
<td>14</td>
<td>6</td>
<td>14</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>One-Step</td>
<td>0.13</td>
<td>0.00</td>
<td>0.31</td>
<td>0.00</td>
<td>0.19</td>
<td>0.00</td>
<td>0.07</td>
<td>0.00</td>
<td>0.18</td>
</tr>
<tr>
<td>HTA-WTA</td>
<td>0.88</td>
<td>0.93</td>
<td>0.68</td>
<td>1.00</td>
<td>0.69</td>
<td>1.00</td>
<td>0.81</td>
<td>0.18</td>
<td>1.00</td>
</tr>
</tbody>
</table>

## Table 7: Few-shot performance of the HTA-WTA and T5-WTA models over a percentage of added few-shot samples. 1 means single instance per skill (9 instances).

<table>
<thead>
<tr>
<th>Instances Ratio</th>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>BLEURT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T5-WTA</td>
<td>8.61</td>
<td>3.38</td>
<td>1.71</td>
<td>1.04</td>
<td>-1.41</td>
</tr>
<tr>
<td></td>
<td>HTA-WTA</td>
<td>10.2</td>
<td>4.74</td>
<td>2.85</td>
<td>1.96</td>
<td>-1.39</td>
</tr>
<tr>
<td>0.1</td>
<td>T5-WTA</td>
<td>14.8</td>
<td>6.68</td>
<td>3.63</td>
<td>2.22</td>
<td>-1.34</td>
</tr>
<tr>
<td></td>
<td>HTA-WTA</td>
<td>16.55</td>
<td>9.54</td>
<td>6.28</td>
<td>4.37</td>
<td>-1.17</td>
</tr>
<tr>
<td>0.3</td>
<td>T5-WTA</td>
<td>16.02</td>
<td>8.3</td>
<td>5.07</td>
<td>3.45</td>
<td>-1.5</td>
</tr>
<tr>
<td></td>
<td>HTA-WTA</td>
<td>16.14</td>
<td>9.7</td>
<td>6.64</td>
<td>4.82</td>
<td>-1.20</td>
</tr>
<tr>
<td>0.5</td>
<td>T5-WTA</td>
<td>16.32</td>
<td>8.25</td>
<td>4.77</td>
<td>3.00</td>
<td>-1.24</td>
</tr>
<tr>
<td></td>
<td>HTA-WTA</td>
<td>15.48</td>
<td>9.25</td>
<td>6.34</td>
<td>4.61</td>
<td>-1.21</td>
</tr>
<tr>
<td>0.75</td>
<td>T5-WTA</td>
<td>18.9</td>
<td>10.12</td>
<td>6.24</td>
<td>4.19</td>
<td>-1.17</td>
</tr>
<tr>
<td></td>
<td>HTA-WTA</td>
<td>18.69</td>
<td>11.53</td>
<td>7.97</td>
<td>5.74</td>
<td>-1.14</td>
</tr>
<tr>
<td>All</td>
<td>T5-WTA</td>
<td>18.53</td>
<td>9.99</td>
<td>6.07</td>
<td>3.93</td>
<td>-1.17</td>
</tr>
<tr>
<td></td>
<td>HTA-WTA</td>
<td>22.15</td>
<td>14.3</td>
<td>10.2</td>
<td>7.67</td>
<td>-1.10</td>
</tr>
</tbody>
</table>

Table 7: Few-shot performance of the HTA-WTA and T5-WTA models over a percentage of added few-shot samples. 1 means single instance per skill (9 instances).