Educational Question Generation of Children Storybooks via Question Type Distribution Learning and Event-Centric Summarization

Zhenjie Zhao[♠ ♡] **Yufang Hou**[♣] **Dakuo Wang**[◇] **Mo Yu**[★ *] **Chengzhong Liu**[♦] **Xiaojuan Ma**[♦]
[♠] Nanjing University of Information Science and Technology [♡] Nankai University
[♣] IBM Research Europe [◇] IBM Research [★] WeChat AI
[♦] The Hong Kong University of Science and Technology
zzhaoao@nuist.edu.cn, yhou@ie.ibm.com, dakuo.wang@ibm.com
moyumyu@tencent.com, chengzhong.liu@connect.ust.hk, mxj@cse.ust.hk

Abstract

Generating educational questions of fairytales or storybooks is vital for improving children's literacy ability. However, it is challenging to generate questions that capture the interesting aspects of a fairytale story with educational meaningfulness. In this paper, we propose a novel question generation method that first learns the question type distribution of an input story paragraph, and then summarizes salient events which can be used to generate high-cognitive-demand questions. To train the event-centric summarizer, we finetune a pre-trained transformer-based sequence-to-sequence model using silver samples composed by educational question-answer pairs. On a newly proposed educational question-answering dataset *FairytaleQA*, we show good performance of our method on both automatic and human evaluation metrics. Our work indicates the necessity of decomposing question type distribution learning and event-centric summary generation for educational question generation.

1 Introduction

Listening to and understanding fairy tales or storybooks are very crucial for children's early intellectual and literacy development (Sim and Berthelsen, 2014). During the storybook reading process, prompting suitable questions with educational purposes can help children understand the content and inspire their interests (Zevenbergen and Whitehurst, 2003; Ganotice et al., 2017).

There is evidence that **high-cognitive-demand** (HCD) questions usually relate to good learning achievement (Winne, 1979). HCD questions usually correspond to application, analysis, synthesis, and evaluation questions in Bloom's taxonomy of cognitive process (Winne, 1979; Anderson et al., 2000), which are salient events merged from different elements across a session (Greatorex and Dhawan, 2016). However, it is challenging even for humans to ask educationally meaningful questions to engage children in storybook reading, which could be due to adults lacking the skills or time to integrate such interactive opportunities (Golinkoff et al., 2019). Recent research shows that AI-powered conversational agents can play the role of language partners to read fairy tales to children and ask them educational questions (Xu et al., 2021). This motivates us to investigate techniques to generate HCD educational questions for children's storybooks automatically. Automating the generation of such questions can have great value in supporting children's language development through guided conversation.

During storybook reading, HCD questions require children to make inferences and predictions. In contrast to low-cognitive-demand (LCD) questions describing facts in stories (*e.g.*, *Who is Snow White's mother?*), HCD questions are often related to events and their relations (*e.g.*, *Why did the queen want to kill Snow White?* or *What happened after the huntsman raised his dagger in the forest?*).

Most previous work on question generation (QG) focuses on generating questions based on predefined answer spans (Krishna and Iyyer, 2019; Pyatkin et al., 2021; Cho et al., 2021). Such systems that use "keywords" or specific events often generate LCD questions that are factual questions based on local context, but cannot work well on HCD cases, where we need to capture the salient events and understand the relations across multiple elements/events. Recently, Yao et al. (2021) released a fairytale question answering dataset **FairytaleQA** containing around 10.5k question-answer pairs annotated by education experts. Each question is assigned to a specific type, and some types, such as "*action*", "*causal relationship*", are high-cognitive-demanding. This makes it possible to investigate generating educational questions to support children's interactive storybook reading.

[*] This work was done while Mo was at IBM Research.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5073 - 5085
May 22-27, 2022 ©2022 Association for Computational Linguistics

In this paper, we propose a novel framework combining question type prediction and event-centric summarization to generate educational questions for storybooks. In the first stage, we learn to predict the question type distribution for a given input and add pseudo-label so that after prediction, we can know both the types of questions and how many questions of each type. In the second stage, conditioned on question types and the order of the question under the current question type, we extract salient events that are most likely for educators to design questions on and then generate an event-centric summarization of the original input. Finally, in the third stage, we use the output of the second stage to generate questions. Each summarization is used to generate one question. Note that it is difficult to obtain gold annotations for event-centric summarization. Instead, we rewrite annotated questions, and their corresponding hypothesized answers into question-answer statements (Demszky et al., 2018) as silver training samples. We hypothesize that HCD questions are around main plots in narratives and can guide our summarization model to focus on salient events. We evaluate our system on the FairytaleQA dataset and show the superiority of the proposed method on both automatic and human evaluation metrics compared to strong baselines.

2 Related Work

2.1 Question Generation

Question answering based on context has achieved remarkable results (Rajpurkar et al., 2016; Zhang et al., 2020b). The reverse problem, namely, question generation (Duan et al., 2017; Chan and Fan, 2019), usually relies on pre-selecting spans from an input text as answers and a single sentence as the context. However, to generate questions across a long paragraph in which the key information may come from multiple different sentences in fairy tales (Yao et al., 2021), these existing models relying on one text segment usually do not work well.

A few studies are focusing on generating questions that are based on multi-sentence or multi-document information fusion (Pan et al., 2020; Xie et al., 2020; Tuan et al., 2020). NarrativeQA (Kočiský et al., 2018) is an effort that tries to integrate key information across multiple locations of a paragraph for question answering/generation. Similarly, MS MARCO (Nguyen et al., 2016) is a dataset that integrates multiple locations of answers for users' queries in search engines. In Cho et al. (2021), a contrastive method is proposed that first trains a supervised model to generate questions based on a single document and then uses a reinforcement learning agent to align multiple questions from multiple documents. In Lyu et al. (2021), the authors use a rule-based method to generate questions with summaries and report to achieve good performance.

The methods mentioned above usually do not consider the educational dimension and may not work well on fairy tales. Considering our research focus of fairytales, it is vital to generate questions that have educational purposes. In FairytaleQA (Yao et al., 2021), experts usually write different types of questions for separate paragraphs. We hypothesize that context plays a significant role in deciding the type of questions that should be asked during the interactive storybook reading with children. Therefore it is necessary to investigate not only how to summarize salient events but also how to learn the question type distribution.

2.2 Text Summarization

Summarization methods can be classified into extractive summarization and abstractive summarization. Extractive methods select sentences from the source documents to compose a summary; abstractive methods applies neural generative models to generate the summary token-by-token.

Extractive summarization methods, such as TextRank (Mihalcea and Tarau, 2004), feature-based methods (Jagadeesh et al., 2005; Luhn, 1958; Nallapati et al., 2017), and topic-based methods (Ozsoy et al., 2010), do not work to generate HCD questions on the fairytale scenario because such questions often are based on multiple sentences.

Abstractive methods based on encoder-decoder architectures usually encode an input document token-by-token sequentially (Rush et al., 2015) and cannot capture the fine-grained hierarchical relations in a document, such as actions, causal relationships. Graph neural network (GNN) models are recently used in summarization research (Wu et al., 2021; Wang et al., 2020; Xu et al., 2020; Li et al., 2021), thanks to their ability to model the complex relations in a document. For example, in Xu et al. (2020), researchers used a discourse-level dependency graph to encode a document and then decoded discourse-level embeddings to select sentences extractively. Similarly, in Wang et al. (2020), researchers have used a heterogeneous graph to encode both token-level and sentence-level relations in a document and then used it to extract sentences. Still, in the education domain, summarizing salient events of one paragraph that can be used to generate educational questions is an open problem. In this paper, we develop an event-centric summarization method based on BART (Lewis et al., 2020). To obtain the training data, we compose educational question-answer pairs through a rule-based method and use them as silver ground-truth samples.

3 Method

The overview of our educational question generation system for storybooks is shown in Figure 1, which contains three modules: question type distribution learning, event-centric summary generation, and educational question generation.

Given an input paragraph d, we first predict the type distribution of output questions $\boldsymbol{p} = (p_1, p_2, \ldots, p_T)$, where p_i denotes the probability of question type i, T is the total number of question types. We then transform the distribution into the number of questions under each question type $\boldsymbol{l} = (l_1, l_2, \ldots, l_T)$. Afterwards, we first generate l_i summaries of type i with the input paragraph d, and then generate l_i questions of type i with the corresponding summaries.

3.1 Question Type Distribution Learning

We fine-tuned a BERT model (Devlin et al., 2019), and adapt the output m dimensional class token $\boldsymbol{h}_c \in \mathbb{R}^m$ to learn the question type distribution. Specifically, the predicted distribution is obtained by $p_i = \frac{e^{(\boldsymbol{W}\boldsymbol{h}_c+\boldsymbol{b})_i}}{\sum_{i=1}^{T} e^{(\boldsymbol{W}\boldsymbol{h}_c+\boldsymbol{b})_i}}$, where $\boldsymbol{W} \in \mathbb{R}^{T \times m}$, $\boldsymbol{b} \in \mathbb{R}^T$ are learnable parameters, $(\cdot)_i$ denotes the operator of selecting the i-th element of a vector.

Assuming there are N training samples, we minimize the K-L divergence loss $\mathcal{L}_{K-L} = \sum_{j=1}^{N} \frac{1}{N} \sum_{i=1}^{T} p_i^{(j)} \log \frac{p_i^{(j)}}{\hat{p}_i^{(j)}}$, where $p_i^{(j)}$ denotes the probability of question type i for the j-th sample, and $\hat{p}_i^{(j)}$ is our predicted value.

To improve the prediction performance, similar to Zhang et al. (2018), we also conduct a multi-label classification task, where we use the question type with the maximal probability as the class of the output. In particular, we add a cross entropy loss $\mathcal{L}_{CE} = -\sum_{j=1}^{N} \frac{1}{N} \sum_{i=1}^{T} \mathbb{1}(y_i^{(j)}) \log \hat{y}_i^{(j)}$, where $\mathbb{1}(y_i^{(j)})$ equals to 1 if i is the question type with the maximal probability for the sample j.

In summary, we conduct a multi-task learning for question type distribution prediction, and the final training loss is a weighted sum of the K-L loss and the cross entropy loss: $\mathcal{L} = \gamma \mathcal{L}_{K-L} + (1-\gamma)\mathcal{L}_{CE}$, where γ is a weight factor.

To predict the number of questions for each question type during training, we add a pseudo label 1 to the original label $\boldsymbol{l} = (l_1, l_2, \ldots, l_n)$, *i.e.*, $\boldsymbol{l} = (l_1, l_2, \ldots, l_n, 1)$. We can then normalize it to get the ground-truth probability distribution $\boldsymbol{l} = (\frac{l_1}{\sum_{k=1}^{n} l_k+1}, \ldots, \frac{l_n}{\sum_{k=1}^{n} l_k+1}, \frac{1}{\sum_{k=1}^{n} l_k+1})$. During testing, assuming we get the predicted distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_n, p_{pseudo})$, we can obtain the number of each type of questions by diving the probability of this pseudo label p_{pseudo} as: $n_i = \lfloor \frac{p_i}{p_{pseudo}} + 0.5 \rfloor$.

3.2 Event-centric Summary Generation

In FairytaleQA, one paragraph usually has multiple questions with different question types, and information in one educational question may scatter across multiple parts. As mentioned before, we assume that context plays a big role to decide the type and the number of questions to be asked during the interactive storybook reading, and HCD questions are around salient events and the relations. With the output from the previous component, we can use the predicted question type distribution as a control signal, and select corresponding events for one particular question type.

In particular, we add two control signals before an input paragraph: question type signal `<t>` and question order signal `<c>`, where `<t>` ∈ `T`, `<c>` ∈ `C`, `T` denotes the set of all question types, `C` denotes the set of order, *i.e.*, {`<first>, <second>, <third>, ...`}. We train a BART summarization model (Lewis et al., 2020) to conduct the event-centric summary generation task. The input of the BART model is: `<t> <c> d`, and the output of the BART model is a summary that collects related events for an educational question type, where `d` denotes the input paragraph.

Obtaining the golden summaries is difficult. However, a QA dataset, like FairytaleQA, provides both questions and their corresponding answers. We can therefore re-write the annotated questions and answers together to obtain question-answer statements, which are used as silver summaries to train our summarization model. We used the rule-based method in Demszky et al. (2018) which inserts answers into the semantic parsed questions

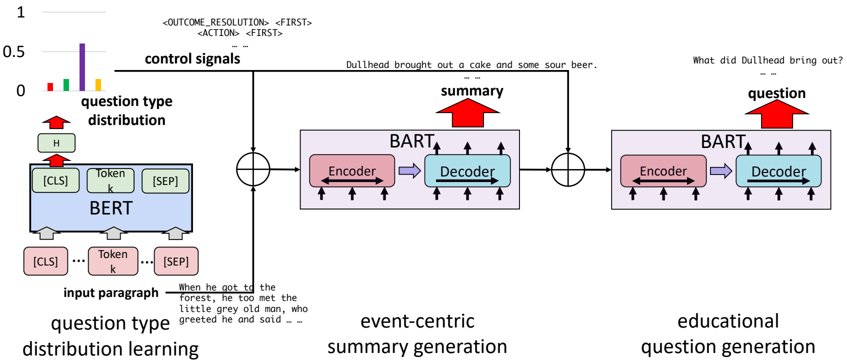
1
0.5
0
question type distribution
H
[CLS]
Token k
[SEP]
BERT
[CLS]
Token k
[SEP]
input paragraph
When he got to the forest, he too met the little grey old man, who greeted he and said … …
question type distribution learning
<OUTCOME_RESOLUTION> <FIRST>
<ACTION> <FIRST>
… …
control signals
Dullhead brought out a cake and some sour beer.
… …
summary
BART
Encoder
Decoder
event-centric summary generation
What did Dullhead bring out?
… …
question
BART
Encoder
Decoder
educational question generation

Figure 1: The overview of our educational question generation system of fairy tales.

and eliminates question words.

3.3 Educational Question Generation

With the summary generated in the second stage, generating an educational question is fairly straightforward. Because the summary has already contained all key events for the target educational question type, we can train a question generation model directly on top of it using the annotated questions. We fine-tune another BART model to generate questions, with the type and order control signals added before the input summary to control the generated results. Note that our question generation model does not reply on pre-selected answer spans.

4 Experimental Setup

To demonstrate the effectiveness of our proposed method, we conducted a set of experiments on the FairytaleQA dataset.

4.1 Dataset

The FairytaleQA dataset (Yao et al., 2021) contains annotations of 278 books, including 232 training books, 23 test books, and 23 validation books. Each book has multiple paragraphs, and for each paragraph of one book, there are several educational question-answer pairs annotated by education experts. The question type distribution is consistent among annotators. In total, there are seven types:

•**Character**: questions that contain the character of the story as the subject and ask for additional information about that character;

•**Setting**: questions that start with "Where/When";

•**Feeling**: questions that start with "How did/do/does X feel?";

•**Action**: questions that start with "What did/do/does X do?" or "How did/do/does X" *or* questions that contain a focal action and ask for additional information about that action;

•**Causal relationship**: questions that start with "Why" or "What made/make";

•**Outcome resolution**: questions ask about logic relations between two events, such as "What happened...after...";

•**Prediction**: questions that start with "What will/would happen...".

The first three are factual questions that are low-cognitive-demanding, and can be handled well by traditional span-based question generation methods (Yao et al., 2021). The remaining four types usually require people to make inferences from multiple elements (Paris and Paris, 2003), which correspond to high-level cognitive skills in Bloom's taxonomy (Anderson et al., 2000), and can be viewed as HCD questions. For the question type *prediction*, it usually asks for events that do not appear in storybooks, which is not our focus in this paper. We only consider *action, causal relationship*, and *outcome resolution*. There is a small portion (985 out of 10580) of questions that span multiple paragraphs. To control the cognitive-demand level for children, we also removed those questions. The statistics of the selected data is shown in section A of the appendix.

4.2 Baselines

We compared our system with two baselines: 1) the method proposed in Yao et al. (2021) (denoted as QAG), which is the only method that considers generating educational questions; 2) using FairytaleQA, we trained an end-to-end BART model.

QAG. The QAG model (Yao et al., 2021) use "keywords" (semantic role labeling) to identify entities and events and then generate questions, which contains four steps: 1) generate a set of answers based on semantic roles of verbs; 2) generate questions based on these answers; 3) generate answers based on the questions generated in the second step; 4) rank generated question-answer pairs and choose the top questions. We trained the question generation model in the second step and the answer generation model in the third step using the selected questions. We use the top 10/5/3/2/1 generated questions as baselines, denoted as QAG (top10), QAG (top5), QAG (top3), QAG (top2), and QAG (top1), respectively.

E2E. Using FairytaleQA with question types *action, causal relationship*, and *outcome resolution*, we trained one BART-large model to generate questions based one paragraph end-to-end. During testing, we used a maximal length 100 tokens (roughly 7 questions according to Table 11) and selected the first 2 questions as the output for evaluation. We denote this method as E2E.

4.3 Evaluation Metrics

We adopt both automatic and human evaluation to measure the performance of our method.

4.3.1 Automatic Evaluation

For automatic evaluation, similar to Yao et al. (2021), we use the Rouge-L score (Lin, 2004), and report the average precision, recall, and F1 values. Meanwhile, we also use BERTScore (Zhang et al., 2020a) to evaluate the semantic similarity of generated questions with the ground-truth questions, and report the average precision, recall, and F1 values. *In contrast to Yao et al. (2021), we mainly consider concatenating all generated questions into one sentence and comparing it with the concatenated ground-truth questions.* This is because for each paragraph, we need to evaluate the generated quality of not only each question but also the question type distribution for sub-skills required in education as a whole (Paris and Paris, 2003). Since the question order does not have much effects on Rouge-L, concatenating questions also partially takes individual question quality into consideration. Moreover, we also consider the same setup used in Yao et al. (2021) that takes the max score of each gold question against the generated questions, then averages the scores of all generated questions.

4.3.2 Human Evaluation

To evaluate the quality of our generated questions and their educational significance, we further conducted a human evaluation session. After regular group meetings, we concluded the following four dimensions, where children appropriateness is the main metric for our educational application:

1. **Question type**: whether the generated questions belong to any of the three event types.
2. **Validity**: whether the generated questions are valid questions according to the original paragraph.
3. **Readability**: whether the generated questions are coherent and grammatically correct.
4. **Children appropriateness**: to what extent would you like to ask this question when you read the story to a five year's old child?

4.4 Implementation Details

For re-writing silver summaries, there are 8 sentences that cannot be parsed successfully. In this case, we wrote the silver statements manually. We also corrected 5 low-quality statements manually.

The weight factor for question type distribution learning is set as 0.7 empirically. For question type distribution learning, we used a BERT cased large model. For summary generation, we used a BART cased base model. For question generation, we used a BART cased large model. The batch sizes of all training are set as 1. For the generation process, we only used a greedy decoding method. Automatic evaluation results were calculated with open sourced packages [1]. For all methods, we removed duplicated questions and questions that has less than 3 tokens. All experiments were conducted on a Ubuntu server with Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz, 32G Memory, Nvidia GPU 2080Ti, Ubuntu 16.04. Training our model took about three hours.

5 Results and Analysis

5.1 Automatic Evaluation Results

The results of automatic evaluation on both validation and test datasets are shown in Table 1. For Rouge-L, compared to E2E and QAG, our method can achieve the best results except for the recall values. In particular, our method outperforms E2E by about 20 points, and outperforms the best QAG model (top2) by about 10 points on the precision scores. For F1, our method outperforms E2E by about 10 points, and outperforms the best QAG model (top2) by about 5 points. These results show

[1] We used the package from `https://github.com/google-research/google-research/tree/master/rouge` to calculate Rouge-L, and the package from `https://github.com/Tiiiger/bert_score` to calculate BERTScore.

	Rouge-L			BERTScore		
Method	Pre (val/test)	Rec (val/test)	F1 (val/test)	Pre (val/test)	Rec (val/test)	F1 (val/test)
E2E	16.32/15.76	36.21/35.89	20.29/19.73	0.8855/0.8839	0.8425/0.8407	0.8632/0.8615
QAG (top1)	34.58/32.33	19.56/19.69	22.88/22.29	0.8599/0.8623	0.8776/0.8770	0.8684/0.8694
QAG (top2)	28.45/26.58	30.51/30.34	26.76/25.67	0.8830/0.8810	0.8745/0.8702	0.8786/0.8754
QAG (top3)	24.29/22.74	36.80/36.31	26.67/25.50	0.8866/0.8846	0.8663/0.8629	0.8761/0.8734
QAG (top5)	20.38/19.25	43.45/43.04	25.55/24.53	0.8883/0.8862	0.8571/0.8540	0.8722/0.8696
QAG (top10)	18.12/17.26	**46.57/47.04**	24.05/23.34	0.8873/0.8848	0.8503/0.8472	0.8681/0.8654
Ours	**33.49/37.50**	37.50/31.54	**31.81/30.58**	**0.8915/0.8862**	**0.8886/0.8930**	**0.8898/0.8893**

Table 1: The comparison results on Rouge-L and BERTScore by concatenating generated questions together.

Method	Pre(val/test)	Rec(val/test)	F1(val/test)
E2E	31.29/30.80	36.21/36.53	31.77/31.65
QAG (top2)	35.17/33.51	35.33/33.83	34.21/32.64
Ours	**48.30/44.05**	**39.55/36.68**	**41.78/38.29**

Table 2: The comparison results with the setup used by Yao et al. (2021).

that our method can match the ground-truth questions lexically better than other methods. However, the recall score of our method is not as good as E2E and QAG (top5 & 10). This is because for E2E and QAG (top5 & 10), they generally generate more questions than our method[2]. For BERTScore, our method achieves the best results on precision, recall, and F1. Although our method outperforms QAG (top2) by a small margin, it still outperforms other QAG models by at least 1 point. For the setup used by Yao et al. (2021), as shown in Table 2, our method also outperforms the best QAG model, *i.e.*, QAG (top 2), and E2E by a large margin in terms of Rouge-L. We believe that decomposing question types explicitly and using event-centric summaries to generate questions can capture the internal structure of educational question annotations and fit the data distribution in a more accurate way.

Some examples of the generated questions can be seen in Table 3. Our method usually can predict the correct question types, and cross multiple elements to generate HCD questions, with a limitation of factuality errors. More examples and comparison can be found in section C of the appendix.

Apart from the overall performance, we also investigated the performance of each module of our method. Because the performance values on both the validation and test data are similar, to simplify our experiment, in the following sections, we only conducted experiments on the test data.

Question Type Distribution Learning. On the test set, the K-L divergence between the prediction results of our BERT-based model and ground-truth is 0.0089, which shows that the performance of our question type distribution learning module is relatively satisfactory. We also use the ground-truth question type distribution as an input and calculate the final Rouge-L score with our system. The results are shown in Table 4. Compared to the ground-truth question type distribution, our system still has lower precision and F1 scores. Having a more accurate question type distribution prediction is beneficial for improving the overall performance.

Event-centric Summary Generation To investigate the quality of the generated summaries, we compare the generated results with the silver summary ground-truth. Similar to the evaluation method of generated questions, we concatenated the generated summaries and calculated the Rouge-L score with the concatenated ground-truth summaries. The results are 15.41 precision, 30.60 recall, and 18.85 F1, which shows that there is still a lot of room to improve the summarization module.

Upper-bound Results with Silver Summary To see how the upper-bound performance is if we have perfect summaries, we input the silver summaries to our educational question generation model. The Rouge-L scores of generated questions are 92.71 precision, 85.65 recall, 87.67 F1, which shows the potential that once a good summary containing salient events is available, generating an educational question is relatively easy. The core challenge is to obtain good summaries, which we believe will be a valuable next step in future work.

5.2 Human Evaluation Results

We conducted a human evaluation with consent of our method against the best-performed baseline QAG (top2). We first randomly sampled 10 books from the test set. For each book, we randomly sam-

[2] On the test data, the mean of the generated questions by our method is 1.9 (std: 0.6), which is closer to the case of ground-truth (mean: 2.2, std: 1.5)

	Questions
QAG (top2)	**P1**: Once upon a time there was a farmer who had carted pears to market .? Why did the farmer want to cart pears?
	P2: What happened to the dwarf after he left? As for the silent earl and his irish sweetheart , they were married as soon?
Ours	**P1**: Why did the bonze want to get a good price for the pears? (causal relationship) What did the bonze ask for? (action)
	P2: What did the Islanders want to express when they were married? (action) Why did the Islanders hold to the belief that Snorro was spirited away? (causal relationship)
Gold	**P1**: Why did the farmer hope to get a good price for the pears? (causal relationship) What did the farmer do when he grew angry? (action)
	P2: What did Paul and Lady Morna do after Harold's funeral was over? (action) Why did Snorro lose all chance of finding the magic carbuncle? (causal relationship)

Table 3: Randomly selected examples of generated questions from two paragraphs (**P1** and **P2**).

Method	Pre	Rec	F1
Ours (gt)	**46.48**	**31.96**	**35.77**
Ours (tdl)	37.50	31.54	30.58

Table 4: The Rouge-L scores of our method with the ground-truth (denoted as gt) and predicted (denoted as tdl) on question type distribution learning.

pled 5 paragraphs. We then conducted experiments to evaluate the generated results on question type and quality. Participants are researchers or PhD students based in Europe, U.S., and China working on natural language processing and human-computer interaction in the education domain with at least 3 years of experience, and were recruited through word-of-mouth and paid $30. We had a training session to ensure the annotation among participants is consistent. This study is approved by IRB.

Question type. Three human participants annotated the types of all generated questions. The inter-coder reliability score (Krippendoff's alpha (Krippendorff, 2011)) among three participants is 0.86, indicating a relatively high consistency. The annotated results are shown in Table 5. Overall, our method demonstrates a much smaller K-L distance (**0.28**) to the ground-truth distribution, compared to QAG (**0.60**). We can see that our method has a better estimation of the distribution of question types, which is closer to the distribution of the ground-truth. QAG has a biased question type distribution and generates more outcome resolution questions.

	QAG(top2)	Ours	Ground-truth
Vague	17/17%	15/17%	0/0%
Action	21/21%	34/38%	47/48%
Causal	10/10%	36/40%	32/33%
Outcome	51/52%	5/6%	18/19%

Table 5: The human evaluation results on of question types (**vague** denotes question types that are hard to decide or questions that have grammar mistakes).

Question quality. We invited another five human participants and conducted a human evaluation to further evaluate the quality of the generated questions from our model against the ground-truth and QAG, including *validity*, *readability*, and *children appropriateness*. Among the three dimensions, the *children appropriateness* is most closely related to the educational purpose; the former two dimensions mainly measure the factual correctness and fluency respectively.

For the total 10×5 paragraphs, each participant is assigned 20 different paragraphs randomly, and each paragraph has annotation results from two participants. For each paragraph, participants need to read the paragraph and its corresponding questions and answers, and then rate the three dimensions on a five-point Likert-scale. The Krippendoff's alpha scores along the four dimensions are between 0.60 and 0.80 (validity: 0.80, readability: 0.69, children appropriateness: 0.60), indicating an acceptable consistency (Gretz et al., 2020).

We conducted an independent-samples t-test to compare the performance of each model. Our model is significantly better than QAG on the main evaluation dimension of *children appropriateness*: the mean score of our model and QAG are 2.56 and 2.22, with corresponding standard derivation 1.31 and 1.20 respectively. This gives a significant score with p-value=0.009, showing that the questions generated by our model can indeed better fit the education scenario. For reference, the ground-truth has a mean score and standard derivation of 3.96 and 1.02, indicating a still large space to improve.

On *validity* and *readability*, our model is on par with QAG. This is not surprising because both models are based on large pre-trained BART models that are good at generating natural and fluent sentences. For validity, our model (avg: 3.19, std: 1.53) is a bit lower than QAG (avg: 3.27, std: 1.62);

for readability, our model (avg: 4.19, std: 1.53) is a bit higher than QAG (avg: 4.12, std: 1.33). A further breakdown in Table 6 shows that QAG wins mainly on action questions, because it directly generates questions conditioned on verbs. For causal relationship and outcome resolution questions, our method generally outperforms QAG.

	QAG	Ours
Vague	2.06*/2.03*	**2.97***/**3.03***
Action	**3.69/4.76***	3.35/4.34*
Causal	**3.45**/4.45	3.10/**4.46**
Outcome	3.46/4.49	**3.50/4.80**

Table 6: The mean values of human evaluation on question qualities (validity/readability), where * denotes significant difference.

6 System Analysis

To further investigate the effectiveness of our method, we conducted a set of ablation studies.

6.1 Question Type Distribution Learning

To investigate the effects of our question type distribution learning, we conducted a comparison study. In particular, we removed the question type distribution learning module (denoted as w/o tdl), and directly trained the summarization and question generation models. In other words, during training, we concatenate all silver summaries as the output of the summarization model. During testing, we extract the first 2 sentences as the predicted summaries. The results are shown in Table 7. From the comparison, we can see that without knowing question types, the Rouge-L scores drop about 3 points overall, which implies the importance of our question type distribution learning module.

6.2 Event-centric Summary Generation

To investigate the effects of our event-centric summary generation module, we conducted a comparison with different summarization methods. The summarization methods include: **1) Lead3**. We select the first three sentences of a paragraph as the summary, and use them as input to the question generation model; **2) Last3**. We select the last three sentences of a paragraph as the summary, and use them as input to the question generation model. **3) Random3**. We select the random three sentences of a paragraph as the summary, and use them as input to the question generation model. **4) Total**. We use each sentence of a paragraph as the summary, and use them as input to the question generation model. **5) TextRank**. TextRank is a typical extractive summarization method. We use TextRank to extract a summary, and for each sentence in the summary, we input it to the question generation model.

Method	Pre	Rec	F1
Ours (w/o tdl)	32.62	29.89	27.42
Ours	**37.50**	**31.54**	**30.58**

Table 7: The Rouge-L scores of our method with and without question type distribution learning.

Method	Pre	Rec	F1
Lead3	25.20	30.76	24.73
Last3	24.35	29.97	24.05
Random3	23.75	28.88	23.07
Total	22.69	**34.34**	24.63
TextRank	30.72	21.74	21.94
Ours (w/o tdl)	**32.62**	29.89	**27.42**

Table 8: The comparison results (Rouge-L on question generation) of different summarization methods.

For other summarization methods, they cannot get the question type distribution like our method. For a fair comparison, we also remove the question type distribution learning module of our method, which is the same as the setting in section 6.1. The results are shown in Table 8, from which we can see that extracting sentences from the paragraph is not enough for covering salient events for educational question generation. Our event-centric summary generation method is an effective way for extracting educational events of fairy tales. Using all sentences (total) can have the highest recall score at the expense of accuracy, but the overall F1 score is still relatively low.

6.3 Multi-task Learning of Question Types

Currently, we use control signals to constrain generating questions of different types, which can be viewed as a multi-task learning framework for multi-type question generation. To investigate whether sharing parameters is a good way for our task, we trained individual summarization and question generation models using different question types. The results in Rouge-L are shown in Table 9. We can find that sharing parameters generally can achieve better performance because of the use of more training data. For only using one type of training data, owing to the error of question type distribution learning, the performance drops a lot, showing the importance of combining question

Method	Pre	Rec	F1
Action	35.97	20.68	24.29
Causal	13.70	11.23	11.54
Outcome	6.15	4.97	5.30
Ours (individual)	25.71	**33.08**	26.27
Ours (overall)	**37.50**	31.54	**30.58**

Table 9: The comparison results of training separate summarization and question generation models on each question type.

type distribution learning and multi-task learning with different types of training data.

7 Conclusion

In this paper, we propose a novel method for educational question generation for fairy tales, which can potentially be used in early childhood education. Our method contains three modules: question type distribution learning, event-centric summary generation, and educational question generation. Through question type distribution learning, we can decompose the challenges of educational question generation by extracting related events of one question type and generating educational questions with a short event-centric summary, which improves the performance significantly. On both automatic evaluation and human evaluation, we show the potential of our method. In the future, we plan to further investigate the event-centric summary generation module by considering discourse-level information to improve the summarization performance and improve the factuality error problem. We are also interested in deploying the system in real scenarios to benefit childcare-related domains.

Acknowledgments

The authors thank constructive suggestions from all anonymous reviewers, as well as all participants in our human evaluation session. This work is supported by the Hong Kong General Research Fund (GRF) with grant No. 16203421. Zhenjie Zhao is supported by the National Natural Science of China Foundation under Grant No. 62106109 and the Startup Foundation for Introducing Talent of NUIST.

References

Lorin W. Anderson, David R. Krathwohl, and Benjamin Samuel Bloom. 2000. A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives. *Longman*.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

Woon Sang Cho, Yizhe Zhang, Sudha Rao, Asli Celikyilmaz, Chenyan Xiong, Jianfeng Gao, Mengdi Wang, and Bill Dolan. 2021. Contrastive multi-document question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 12–30, Online. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv:1809.02922*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Fraide A. Ganotice, Kevin Downing, Teresa Ka Ming Mak, Barbara Chan, and Wai Yip Lee. 2017. Enhancing parent-child relationship through dialogic reading. *Educational Studies*, 43:51 – 66.

Roberta Michnick Golinkoff, Erika Hoff, Meredith L. Rowe, Catherine S. Tamis-LeMonda, and Kathy Hirsh-Pasek. 2019. Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child development*, 90 3:985–992.

Jackie Greatorex and Vikas Dhawan. 2016. Analysing the cognitive demand of reading, writing and listening tests. *ISEC Proceedings*.

Shai Gretz, Yonatan Bilu, Edo Cohen-Karlik, and Noam Slonim. 2020. The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 528–544, Online. Association for Computational Linguistics.

Jaya Jayashree Jagadeesh, Prasad Pingali, and Vasudeva Varma. 2005. Sentence extraction based single document summarization. *International Institute of Information Technology, Hyderabad, India*, 5.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *University of Pennsylvania*.

Kalpesh Krishna and Mohit Iyyer. 2019. Generating question-answer hierarchies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2321–2334, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Chenyang Lyu, Lifeng Shang, Yvette Graham, Jennifer Foster, Xin Jiang, and Qun Liu. 2021. Improving unsupervised question answering via summarization-informed question generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4134–4148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Makbule Ozsoy, Ilyas Cicekli, and Ferda Alpaslan. 2010. Text summarization of Turkish texts using latent semantic analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 869–876, Beijing, China. COLING 2010 Organizing Committee.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.

Alison H. Paris and Scott G. Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Valentina Pyatkin, Paul Roit, Julian Michael, Yoav Goldberg, Reut Tsarfaty, and Ido Dagan. 2021. Asking it all: Generating contextualized questions for any semantic role. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1429–1441, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Susan Sim and Donna Berthelsen. 2014. Shared book reading by parents with young children: Evidence-based practice. *Australasian Journal of Early Childhood*, 39(1):50–55.

Luu Anh Tuan, Darsh Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9065–9072.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219, Online. Association for Computational Linguistics.

Philip H. Winne. 1979. Experiments relating teachers' use of higher cognitive questions to student achievement. *Review of Educational Research*, 49(1):13–49.

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. 2021. Graph neural networks for natural language processing: A survey. *arXiv:2106.06090*.

Yuxi Xie, Liangming Pan, Dongzhe Wang, Min-Yen Kan, and Yansong Feng. 2020. Exploring question-specific rewards for generating deep questions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2534–2546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Penelope Collins, Hyelim Lee, and Mark Warschauer. 2021. Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers & Education*, 161:104059.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Tran Hoang, Branda Sun, Toby Jia-Jun Li, Mo Yu, and Ying Xu. 2021. It is AI's turn to ask human a question: Question and answer pair generation for children storybooks in FairytaleQA dataset. *arXiv:2109.03423*.

Andrea A. Zevenbergen and Grover J. Whitehurst. 2003. Dialogic reading: A shared picture book reading intervention for preschoolers. *Lawrence Erlbaum Associates Publishers*.

Tianyi Zhang, Varsha Kishore*, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text emotion distribution learning via multi-task convolutional neural network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4595–4601. International Joint Conferences on Artificial Intelligence Organization.

Zhuosheng Zhang, Hai Zhao, and Rui Wang. 2020b. Machine reading comprehension: The role of contextualized language models and beyond. *ArXiv*, abs/2005.06249.

Appendix

A Dataset Statistics

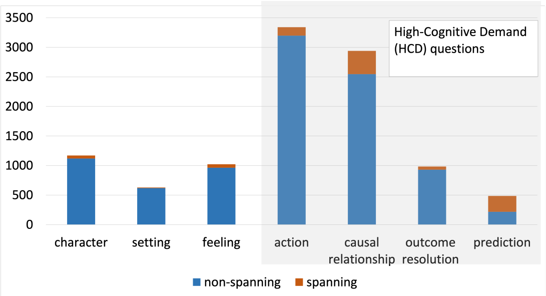

Figure 2: Question type distribution of FairytaleQA.

The question type distribution of FairytaleQA is shown in Figure 2. In particular, question types *action, causal relationship, outcome resolution, and prediction* are considered as HCD questions. For each question type, there are some questions that span multiple paragraphs (denoted as *spanning*, *character*: 53, *setting*: 11, *feeling*: 59, *action*: 143, *causal relationship*: 392, *outcome resolution*: 54, *prediction*: 266), which are discarded. We select question types *action, causal relationship, and outcome resolution* from FairytaleQA for conducting our experiments. In total, there are 2998 out of 4095 paragraphs used, including 2430 out of 3350 training paragraphs, 290 out of 380 validation paragraphs, and 278 out of 365 paragraphs. The number of QA pairs for each question type and the total number are shown in Table 10, and the token-level statistics of the selected training data can be found in Table 11.

	train	val	test	total
#action	2574	322	302	3198
#causal relationship	2057	246	244	2547
#outcome resolution	766	93	72	931
#selected	5397	661	618	6676
#total	8548	1025	1007	10580

Table 10: The numbers of QA pairs for question types *action*(#action), *causal relationship*(#causal relationship) and *outcome resolution*(#outcome resolution), the selected data (#selected), and all data of FairytaleQA (#total).

B Potential Risks

While High-Cognitive Demand (HCD) questions are considered in this paper, the cultivation of knowledge and ability is equally important for children. The experiment results show that our method is competitive to generate HCD questions, and therefore it is helpful to improve children's cognitive ability. However, because of the unexplainability of end-to-end training, we also find that sometimes our system may generate non-factual facts in terms of the original context, which has a potential risk on knowledge learning. Owing to the factuality error problem of our system, we suggest to further investigate constructing structured knowledge of fairy tales and knowledge-grounded question generation for real-world applications.

	mean	std
#question	2.2	1.5
#token (paragraph)	160.4	65.1
#token (summary)	17.8	7.2
#token (question)	10.1	3.1

Table 11: The mean and standard deviation (std) of the number of questions for each paragraph (#question) and the number of tokens (#token) in paragraphs, summaries, and questions in the training data.

C Examples of Generated Questions

Some randomly selected examples of the generated questions can be found in Table 12.

Paragraph: Once upon a time there was a farmer who had carted pears to market. Since they were very sweet and fragrant, he hoped to get a good price for them. A bonze with a torn cap and tattered robe stepped up to his cart and asked for one. The farmer repulsed him, but the bonze did not go. Then the farmer grew angry and began to call him names. The bonze said: "You have pears by the hundred in your cart. I only ask for one. Surely that does you no great injury. Why suddenly grow so angry about it?"

Gold questions: Why did the farmer hope to get a good price for the pears? (causal relationship) What did the farmer do when he grew angry? (action)

Generated questions by our method: Why did the bonze want to get a good price for the pears? (causal relationship) What did the bonze ask for? (action)

Generated questions by QAG (top2): Once upon a time there was a farmer who had carted pears to market .? Why did the farmer want to cart pears?

Silver summaries: The farmer hoped to get a good price for the pears because they were very sweet and fragrant. (causal relationship) The farmer called the bonze names when he grew angry. (action)

Generated summaries by our method: The bonze wanted to get a good price for the pears because they were very sweet and fragrant. (causal relationship) The bonze asked for one. (action)

Paragraph: No one knew what had become of them; a few people were inclined to think that the Dwarf and his Raven had accompanied the Countess Fraukirk and the Countess Helga on their flight, but the greater part of the Islanders held to the belief, which I think was the true one, that the Powers of the Air spirited Snorro away, and shut him up in some unknown place as a punishment for his wickedness, and that his Raven accompanied him. At any rate, he was never seen again by any living person, and wherever he went, he lost all chance of finding the magic carbuncle. As for the Silent Earl and his Irish Sweetheart, they were married as soon as Earl Harold's funeral was over; and for hundreds of years afterwards, when the inhabitants of the Orkney Isles wanted to express great happiness, they said, "As happy as Earl Paul and the Countess Morna."

Gold questions: What did Paul and Lady Morna do after Harold's funeral was over? (action) Why did Snorro lose all chance of finding the magic carbuncle? (causal relationship)

Generated questions by our method: What did the Islanders want to express when they were married? (action) Why did the Islanders hold to the belief that Snorro was spirited away? (causal relationship)

Generated questions by QAG (top2): What happened to the dwarf after he left? As for the silent earl and his irish sweetheart , they were married as soon?

Silver summaries: Paul and Lady Morna they was married after Harold's funeral was over. (action) Snorro lost all chance of finding the magic carbuncle because he did not live on the island anymore. (causal relationship)

Generated summaries by our method: The Islanders wanted to express great happiness when they were married. (action) The Islanders held to the belief that Snorro was spirited away because he wickedness. (causal relationship)

Paragraph: The King danced again with the beautiful maiden, and thought she had never looked so beautiful. And while he was dancing, he put a gold ring on her finger without her seeing it, and he commanded that the dance should last longer than usual. When it was finished he wanted to keep her hands in his, but she broke from him, and sprang so quickly away among the people that she vanished from his sight. She ran as fast as she could to her little room under the stairs, but because she had stayed too long beyond the half-hour, she could not stop to take off the beautiful dress, but only threw the fur cloak over it, and in her haste she did not make herself quite black with the soot, one finger remaining white.

Gold questions: What did the king do to the maiden while they were dancing? (action) What happened because the maiden stayed too long beyond the half-hour? (outcome resolution) Why did the king want to keep the maiden's hand in his? (causal relationship)

Generated questions by our method: What did the maiden do because it had stayed too long beyond the half-hour? (action) What happened after the maiden danced again with her dress? (outcome resolution) Why did the maiden break from the King? (causal relationship)

Generated questions by QAG (top2): What happened after the king danced again with his beautiful maiden? What happened because the king danced again with his beautiful maiden?

Silver summaries: The king he put a gold ring on her finger without her seeing it to the maiden while they were dancing. (action) Because the maiden stayed too long beyond the half - hour she could not stop to take off the beautiful dress. (outcome resolution) The king wanted to keep the maiden's hand in his because he loved her. (causal relationship)

Generated summaries by our method: The maiden ran as fast as she could to her little room under the stairs because it had stayed too long beyond the half - hour. (action) After the maiden danced again with her dress she only threw it over it, and in haste did not make herself quite black. (outcome resolution) The maiden broke from the King because she had stayed too long beyond the half - hour. (causal relationship)

Paragraph: Art thou satisfied now?' said Matte to his wife. 'I should be quite satisfied,' said his wife, 'if only I had two servants to help, and if I had some finer clothes. Don't you know that I am addressed as Madam?' 'Well, well,' said her husband. So Maie got several servants and clothes fit for a great lady. 'Everything would now be perfect if only we had a little better dwelling for summer. You might build us a two-storey house, and fetch soil to make a garden. Then you might make a little arbour up there to let us have a sea-view; and we might have a fiddler to fiddle to us of an evening, and a little steamer to take us to church in stormy weather.' 'Anything more?' asked Matte; but he did everything that his wife wished. The rock Ahtola became so grand and Maie so grand that all the sea-urchins and herring were lost in wonderment. Even Prince was fed on beefsteaks and cream scones till at last he was as round as a butter jar. 'Are you satisfied now?' asked Matte. 'I should be quite satisfied,' said Maie, 'if only I had thirty cows. At least that number is required for such a household.' 'Go to the fairies,' said Matte.

Gold questions: What did Maie want Matte to build? (action) How many cows did Maie want? (action)

Generated questions by our method: What did Maie get? (action) What did the rock Ahtola become? (action)

Generated questions by QAG (top2): What did matte ask his wife to do? What did matte tell his wife to do?

Silver summaries: Maie wanted a two - storey house Matte to build. (action) Maie wanted thirty cows. (action)

Generated summaries by our method: Maie got several servants and clothes fit for a great lady. (action) The rock Ahtola became so grand and Maie was lost in wonderment. (action)

Table 12: Randomly selected examples of original paragraphs, their corresponding gold questions, questions generated by our method, questions generated by QAG (top2), silver summaries, and summaries generated by our method.