絵本を題材とするクイズの生成と評価

水上 雅博 † 藤田 早苗 † 小林 哲生 †
†NTT コミュニケーション科学基礎研究所
{masahiro.mizukami, sanae.fujita, tessei.kobayashi}@ntt.com

概要

特定のドメインにおける知識量や理解度の確認を楽しく行うために，しばしばクイズが用いられる．クイズには高い需要がある一方で，クイズの作成は，クイズの題材となるドメインの理解だけでなく，クイズの目的や場面に応じた様々な明示的・暗示的条件を考慮する必要があり，作業のコストが高いという問題がある．本研究では絵本を題材とした子ども向けクイズの自動生成に向けて，クイズデータの人手作成，複数条件でのインストラクションの生成，LLM の追加学習を行い，インストラクションごとのクイズ生成の性能を評価することで，どのような情報がクイズ生成に有益であるか調査した．

1 はじめに

絵本を読むことは，子どもの言語発達と情操教育の両面で効果が期待できる [1]．NTT ではこれまで，子どもの興味や発達にあった絵本の推薦に取り組んできたが，絵本を読んだ後にも，内容に関するクイズを出題したり感想を話し合えるシステムを提供することで，理解度を測ると同時に読書への意欲を高めることを目指している．

特にクイズは，理解度を確認する有効性が示されており [2]，楽しみながら理解度の確認ができるという点で有用である．しかし，クイズを人手で作成するためには，実際に絵本を読み，クイズとして適切な問題文と回答を作成する必要があるため，高い人的コストを要する．そのため，クイズの自動生成（作問）技術が望まれている．

クイズの自動作問サービスは既に存在するものの，クイズの作問では「場面にあった主題と様々な明示的・暗示的な条件を満たす必要がある」と言われている [3]．本稿が目指す「子どもの絵本に対する理解度を確認するためのクイズ」においては，クイズの主題が絵本であるだけでなく，クイズで問う内容は絵本の内容に忠実かつ限定されている必要がある．また，クイズの問題文や回答が，その絵本を読み終わったばかりの子どもに対して適した難易度である必要がある．例えば，日本語の QA データセット JAQKET[4] には，絵本『ぐりとぐら』[5] に関するクイズが 4 問あるが，『ぐりとぐら』を読み終わった子どもへのクイズとしては，“おりょうりすることとたべることが好きな 2 匹の野ねずみの生活を描いた、中川李枝子（なかがわ・りえこ）の絵本シリーズは何でしょう？” よりも，“中川李枝子と山脇百合子による絵本『ぐりとぐら』で、ぐりとぐらが卵から作ったお菓子は何？” の方が適切だと考えられる．さらに言えば，『ぐりとぐら』に関するクイズだとわかっている場合には，下線部だけを出題すれば十分である．

これを踏まえ，本稿ではまず，子どもが絵本を読んだ後に回答することを想定したクイズデータを人手で作成し，インストラクションチューニング用のデータに変換する．その際，絵本のタイトル, 本文や想定読者の年齢情報を含めるかどうかなど，条件を変えてインストラクションデータを作成し，LLM の追加学習を行う．追加学習で得られたモデルを用いて，評価用の未知の絵本に対するクイズの生成を行い，クイズの品質を評価することで，インストラクションに含まれる情報がクイズの生成に与える影響を調査する．

2 関連研究

クイズの自動作問の研究では，AI 王[1] における問題作成部門や，橋本らによる早押しクイズのパラレル問題の自動生成 [6]，折原らによる時事問題を題材とした作問作業支援 [3] がある．また，QuizGenerator[2] や Questiogen[3] といったクイズ生成サービスも提供されている．クイズの自動作問に関連する研究としては，Q&A の自動生成の研究があ

1） https://sites.google.com/view/project-aio
2） https://quizgenerator.net/
3） https://www.questgen.ai/

り，製品マニュアルからQAを生成する取り組み [7] などがなされている．

具体的な生成手法については，ルールベースで質問を生成する方法 [8] やテンプレートを用いる方法 [7]，Encoder-Decoder や T5 などの NN を用いる方法 [9, 10, 11, 12] など様々な手法が提案されている．

先行研究の多くは，出題の対象者として大人を想定している．本研究は絵本を題材とする子ども向けのクイズの生成を目指しており，想定読者の年齢を考慮した生成を行うなどの点が異なる．

3 クイズデータの作成

絵本 150 冊（うち 32 冊は児童書）に対し，人手でクイズを作成した．対象絵本は，図書館の推薦絵本，教科書掲載作品，課題図書[4] などから選定した．各絵本には，本文テキストから自動推定した対象年齢 (以下，age) が付与されている [13]．

クイズの作成者は著者 2 人を含む 3 人で，実際に絵本を読んだり内容を確認しながら，それぞれが独立にクイズを作成した．作成したクイズの例を表 1 に示す．作業の結果，914 問，1 冊平均 6.1 問のクイズが作成された．さらに，クイズの情報源となる絵本の本文がある場合，本文のテキスト (以下，ref) をページ単位で抽出した．なお，ref を抽出できなかったクイズには，絵をみて回答する必要があるものや，絵本全体を読まないと回答できないもの，絵本に出てこないものを回答させるもの等があった．

以降の実験では，ref を抽出できたクイズ (113 冊 535 問) を用いる．人手で作成したクイズは誤答 3 つを含む 4 択問題だったが，本稿では問題文 (以下，$question$) と正解の回答 (以下，$answer$) のみを用いる．以降，タイトル ($title$)，本文 (ref)，対象年齢 (age) の情報と合わせてクイズデータと呼ぶ．

4 インストラクションの生成

3 節で作成したクイズデータから，インストラクションチューニング用のデータを生成した．インストラクションでは，問題文と回答の両方を同時に生成するのではなく，与えられた任意の単語が回答となるような問題文を生成するようにした．これは，(1) 回答となる語を何らかの基準（単語親密度 [16] や学習指導要領など）で選択できるようにする，(2) 生成したクイズの人手評価およびクイズの答え合わせを簡略化する，という二つの目的がある．

本稿では，どのようなインストラクションが適切かを調査するため，クイズデータ $(title, ref, age)$ を網羅的に組み合わせてインストラクションを生成した．すなわち，利用可能な情報の全ての組み合わせ $S = \{(), (title), (ref), (age), (title, ref), \ldots, (title, ref, age)\}$ を用意し，ある組み合わせ $s \in S$ のインストラクション $inst$ を次のような手順で作成した．なお，手順中の $\{title\}$,$\{ref\}$,$\{age\}$,$\{question\}$,$\{answer\}$ は，それぞれクイズデータの中身である．

1. if $title$ in s: $inst$ += “絵本「$\{title\}$」について、”
2. if ref in s: $inst$ += “次に入力する本文を読んだ上で、”
3. $inst$ += “「$\{answer\}$」が答えになる”
4. if age in s: $inst$ += “$\{age\}$ 向けの”
5. $inst$ += “クイズ（問題文）を考え、応答として出力してください”
6. if ref in s: inst += “###入力: $\{ref\}$”
7. $inst$ += “###出力：”
8. $output$ = “$\{quiz\}$”

加えて，問題文の生成と同時に「回答の生成」も可能なマルチタスクモデルの学習に用いるため，与えられた問題文に回答するインストラクションも作成した．具体的には，手順 3 を $inst$ += “「$\{question\}$」という”，手順 5 を $inst$ += “クイズ（問題文）に対する答えを考え、応答として出力してください”，手順 8 を $output$ = “$\{answer\}$” に変更した．

最終的に，クイズデータの情報 $(title, ref, age)$ の組み合わせ $|S|$ が 7 種類，マルチタスク学習データ追加 (+MT) の有無で，合計 14 種類のインストラクションチューニング用のデータを作成した．表 2 に作成したデータの例を示す．

5 モデルの学習と生成

4 節で作成したインストラクションチューニング用データを用いて，LLM を Low-Rank Adaptation(LoRA)[17, 18] で追加学習した．ベースとなる LLM には，NTT が構築した日本語に強い LLM である tsuzumi[5] の 7B モデルを用い，LoRA での追加学習パラメータは表 3 の通り設定した．

インストラクションチューニング用のデータは，絵本ごとにまとめた上で，学習: 検証: 評価が 8:1:1 となるように分割した．その結果，学習 429 件，検証 54 件，評価 52 件となった．なお，マルチタスク学習（+MT）を行う場合は，学習データは倍の 858 件となる．学習中は 10step 毎に検証データに対する

4） https://www.dokusyokansoubun.jp/books.html

5） https://www.rd.ntt/research/LLM_tsuzumi.html

表 1 絵本を題材として人手で作成したクイズの例 (回答のみ記載)

タイトル	問題文	回答
はらぺこあおむし [14]	あおむしが最初に食べたものはなにかな？	りんご
ぐりとぐら [5]	ぐりとぐらが見つけた、大きなたまごで作ったものは？	カステラ
ぞうくんのさんぽ [15]	ぞうくんが最初に会ったのは、誰でしょう？	かばくん

表 2 インストラクションチューニング用のデータの例（本文の入力以下は省略）

利用した情報 s		
$(title, ref, age)$	*inst*	絵本「はらぺこあおむし」について、次に入力する本文を読んだ上で、「りんご」が答えになる 2 歳児向けのクイズ（問題文）を考え、応答として出力してください.
	out put	あおむしが月曜日に食べたものはなんでしょう？
$(title, ref, age)$+MT	*inst*	絵本「はらぺこあおむし」について、次に入力する本文を読んだ上で、「あおむしが日曜日に食べたものはなんでしょう？」という 2 歳児向けのクイズの答えを考え、応答として出力してください.
	out put	みどりのはっぱ

表 3 LoRA での追加学習パラメータ設定

Parameter	Value
r	16
alpha	32
Target Layer	Wqkv
epoch	50
Batch size	32
Optimizer	AdamW[19]
lr	3e-4
lr scheduler	cosine

Perplexity を評価し，最良のモデルを採用した．

採用したモデルと評価データを用いて問題文を生成し，beam 幅 4 の beam search を用いて最良の結果を 1 つ取得した．加えて，マルチタスク学習を行ったモデル (+MT) では，Q&A 生成で提案されている Roundtrip[20] が適用できるため，この手法を用いた生成結果 (+RT) も取得した[6]．表 4 に生成された問題文の例を示す．

6 評価

生成された問題文に対して，自動評価と人手評価を行った．自動評価としては，人手で作成した問題文 (GOLD) を比較対象として，ROUGE[22] と BLEU[23] のスコアを計算した．人手評価としては，(1) 妥当性評価：生成された問題文が指定の絵本に対して適切かどうかの 5 段階評価[7]，(2) 完全性評価：生成された問題文が指定された絵本の指定された単語が回答となる問題文として適切かどうかの 2 値評価[8]を行った．表 5 に評価結果を示す．また，各インストラクションで利用した情報 s が問題文の生成にどの程度影響したかを分析するため，ROUGE-1 と BLEU-1 のスコアに対して情報 s とマルチタスク学習 (+MT)，Roundtrip(+RT) の有無を説明変数とした重回帰分析を行った (表 6).

表 5 に示したように，自動，人手評価ともに $title, ref, age$ の全てを用いた場合の評価が最も高くなった．つまり，インストラクションが与える情報が増えるほど生成される問題文の評価が高くなっていた．これは，直感にも一致する結果と言えるだろう．重回帰分析 (表 6) では，ref を用いた場合の評価が有意に向上したことから，生成される問題文の評価に最も影響を与える情報は本文であることが示された．また，有意ではないものの $title$ を用いた場合も評価が向上する傾向が見られ，タイトルをインストラクションで与えることにより，LLM が持つ知識を活用できた可能性も示唆された．

妥当性評価における各点数の分布を確認したところ，指定された絵本の問題文として適切であると評価された 3-5 点の割合は 53.8%であり，半数は問題文としての体裁を保っていた．一方で，完全性評価において回答まで考慮して適切なクイズであると評価された割合は 30%であった．絵本に関連する単語を含み，流暢な問題文であっても，必ずしもインストラクションで与えられた指示に忠実な生成ができていないことがわかった (例えば表 4 (4)．問題文としては良いが，「ヤギ」が回答にはならない)．特に，本文を与えない場合は，別の絵本の知識を用いたり (表 4 (3))，それらしい内容を勝手に生成するハルシネーションが起こり，対象絵本のクイズとしては不適切な問題文が生成される現象が多く確認された．

マルチタスク学習を行ったモデル+MT と Roundtrip 推論を行う+RT では，評価は改善しな

6） ただし，計算時間の都合上，100 回施行して正しい回答が得られない場合は beam search の結果を用いた

7） 妥当性評価では，対象絵本のクイズとして成り立つ場合は 3 以上を付与するよう指示した

8） 評価作業の簡略化のため，完全性評価は妥当性評価で 3-5 と評価されたもののみを対象に実施し，妥当性評価で 1-2 と評価されたものは完全性評価は 1: 不適切とした

表 4　生成された問題文の例: 対象絵本『あらしのよるに』[21]

番号	モデル	回答	問題文の生成結果
(1)	GOLD	ヤギ	先に小屋にいたのは誰だったかな？
(2)	(*title*)		オオカミが，オオカミの友達になりたいと思ったのは，何だったかな?
(3)	()+MT+RT		チムと仲良しな動物は何でしょう?
(4)	(*age*)+MT		ヤギは，何を食べるでしょう?
(5)	(*ref*, *age*)+MT		くらやみの中で，誰かが入ってきたのは，誰の家?
(6)	(*title*, *ref*, *age*)+MT+RT		くらやみの 中にいたのは，誰でしょう?

表 5　生成された問題文に対する評価結果

	ROUGE-#		BLEU-#		人手評価	
モデル	1	2	1	2	妥当性	完全性
()	.28	.08	.02	.07	1.21	1.00
+MT	.30	.09	.25	.10	1.33	1.00
+MT+RT	.26	.09	.21	.09	1.02	1.00
(*title*)	.33	.11	.25	.11	2.04	1.04
+MT	.29	.09	.22	.09	2.15	1.06
+MT+RT	.29	.07	.24	.07	1.58	1.06
(*ref*)	.43	.20	.34	.21	2.90	1.29
+MT	.40	.15	.32	.18	2.67	1.08
+MT+RT	.36	.13	.28	.13	2.21	1.19
(*age*)	.29	.08	.23	.07	1.46	1.00
+MT	.26	.06	.21	.07	1.40	1.00
+MT+RT	.26	.05	.21	.07	1.06	1.00
(*title*,*ref*)	.43	.21	.36	.22	3.29	1.33
+MT	.44	.22	.34	.21	3.10	1.29
+MT+RT	.37	.11	.29	.13	2.04	1.08
(*title*,*age*)	.35	.12	.27	.13	2.31	1.12
+MT	.34	.11	.26	.13	2.06	1.10
+MT+RT	.31	.09	.25	.10	1.46	1.04
(*ref*,*age*)	.41	.18	.33	.20	3.02	1.29
+MT	.42	.18	.34	.18	2.92	1.25
+MT+RT	.36	.12	.28	.13	2.35	1.19
(*title*,*ref*,*age*)	**.45**	**.23**	**.37**	**.23**	**3.40**	**1.31**
+MT	.41	.17	.32	.18	2.75	1.21
+MT+RT	.35	.11	.28	.13	2.21	1.21

表 6　ROUGE-1 と BLEU-1 に対する重回帰分析

	ROUGE-1		BLEU-1	
説明変数	係数	P 値	係数	P 値
切片	0.30	9.6e-17	0.25	1.2e-15
title	0.028	.0033	0.017	.051
ref	0.11	1.7e-10	0.084	3.9e-9
age	0.0033	.70	0.00011	.99
+MT	-0.050	.00010	-0.044	.00029
+RT	-0.014	.19	-0.019	.068

表 7　生成された回答に対する評価結果

モデル	ROUGE-1	BLEU-1	人手評価
()	.025	.024	1.33
(*title*)	.032	.027	1.38
(*ref*)	.78	.77	4.38
(*age*)	.031	.030	1.25
(*title*,*ref*)	**.84**	**.83**	4.42
(*title*,*age*)	.052	.050	1.46
(*ref*,*age*)	.83	.83	4.40
(*title*,*ref*,*age*)	.82	.81	**4.54**

かった．学習そのものが成功しているかどうかを検証するため，問題文に対して回答を生成するインストラクションを用いて回答も生成し，自動評価 (BLEU-1, ROUGE-1) と，問題文に対して正しい回答になっているかの 5 段階の人手評価を行った．評価結果を表 7 に示す．

問題文の評価 (表 5) と同様，本文 (*ref*) を利用することで大幅に評価が向上し，問題文に対して正しい回答が生成できた．このことから，今回の結果ではマルチタスク学習自体は成功しているが，マルチタスク学習による回答の生成能力の向上が問題文の生成能力の向上に寄与しなかったことがわかった．

7　まとめ

本稿では子ども向けの絵本に関するクイズの自動作問を目指し，クイズデータの作成と，LLM の LoRA を用いたインストラクションチューニングを行い，生成された問題文の評価を行った．

評価結果から，インストラクションには絵本のタイトル，本文，対象年齢の情報を与えた場合に評価が最も高くなり，人手評価でも生成結果の 50%が問題文としては妥当，31%がインストラクションに忠実に従った問題文であることがわかった．一方で，問題文から回答を生成するタスクを含めたマルチタスク学習では，回答を生成することはできたものの，問題文を生成する性能は向上しなかった．

今後の課題としては，ハルシネーションに起因する不適切な問題文の生成を抑える方法を探るとともに，4 択問題などで提示する誤答の自動生成にも取り組みたい．さらに，生成されたクイズによって子どもの理解度を測ることが可能かどうかや，読書意欲の向上につながるかの検証を行いたい．

参考文献

[1] 藤田早苗, 服部正嗣, 小林哲生, 奥村優子, 青山一生. 絵本検索システム「ぴたりえ」～子どもにぴったりの絵本を見つけます～. 自然言語処理, Vol. 24, No. 1, pp. 49–73, 2017.

[2] Pedro Azevedo, Bernardo Leite, Henrique Lopes Cardoso, Daniel Castro Silva, and Luís Paulo Reis. Exploring nlp and information extraction to jointly address question generation and answering. 2020.

[3] 折原良平, 鶴崎修功, 森岡靖太, 島田克行, 狭間智恵, 市川尚志. クイズビジネスにおける作問作業支援. 言語処理学会 第 28 回年次大会 発表論文集, 2022.

[4] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語 QA データセットの構築. 言語処理学会 第 26 回年次大会 発表論文集, 2020.

[5] おおむらゆりこ, なかがわりえこ. ぐりとぐら. 福音館書店, 1963.

[6] 橋元佐知, 佐藤理史, 宮田玲, 小川浩平. 早押しクイズのパラレル問題の自動生成. 言語処理学会 第 28 回年次大会 発表論文集, 2022.

[7] 佐藤紗都, 伍井啓恭, 奥村学. 製品マニュアル文からの質問自動生成. 人工知能学会 第 32 回全国大会 論文集, 2018.

[8] 田村吉宏, 山内崇資, 林佑樹, 中野有紀子. Wikipedia 記事情報に基づく歴史学習問題の自動生成手法. 人工知能学会 第 28 回全国大会 論文集, 2014.

[9] 牧野拓哉, 野呂智哉, 吉川和, 岩倉友哉, 関根聡, 乾健太郎. 自動生成した質問に基づく質問応答学習手法の提案と評価. 言語処理学会 第 24 回年次大会 発表論文集, 2018.

[10] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Association for Computational Linguistics, 2017.

[11] Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, 2018.

[12] Ying-Hong Chan and Yao-Chung Fan. A recurrent BERT-based model for question generation. In **Proceedings of the 2nd Workshop on Machine Reading for Question Answering**, 2019.

[13] 藤田早苗, 小林哲生, 南泰浩, 杉山弘晃. 幼児を対象としたテキストの対象年齢推定方法. 認知科学, Vol. 22, No. 4, pp. 604–620, 2015.

[14] エリック＝カール, もりひさし. はらぺこあおむし. 偕成社, 1976.

[15] なかのひろたか, なかのまさたか. ぞうくんのさんぽ. 福音館書店, 1968.

[16] 藤田早苗, 小林哲生. 単語親密度の再調査と過去のデータとの比較. 言語処理学会 第 26 回年次大会 発表論文集, 2020.

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In **arXiv**, 2021.

[18] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft, 2022.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **arXiv**, 2019.

[20] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic qa corpora generation with roundtrip consistency. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, 2019.

[21] きむらゆういち, あべ弘士. あらしのよるに. 講談社, 2000.

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, 2004.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.