項目反応理論に基づく難易度調節可能な読解問題自動生成手法

鈴木彩香 [1]　宇都雅輝 [1]
[1] 電気通信大学大学院
{suzuki_ayaka,uto}@ai.lab.uec.ac.jp

概要

読解問題自動生成とは，読解対象文からそれに関連する問題を自動生成する技術である．近年では，深層学習を用いた手法により，柔軟で高品質な問題生成が実現されている．しかし，従来手法には，次の課題がある．1）問題に対応する答えを生成できない．2）学習者の能力に合わせた難易度の問題を生成できない．これらの問題を解決するために，本研究では，項目反応理論を用いて推定される難易度を考慮して，問題と答えのペアを生成する手法を提案する．提案手法は事前学習済み深層学習モデル（BERT と GPT-2）を拡張することで実現する．

1　はじめに

読解問題自動生成とは，読解対象文からそれに関連する問題を自動生成する技術であり，教育分野において読解力の育成・評価を支援する技術の一つとして活用が期待されている．

従来の読解問題自動生成手法は，人手で設計したルールやテンプレートを利用する手法が主流であったが，適切なルールやテンプレートの作成には大きなコストを要する [1, 2, 3]．この問題に対し，近年では，深層学習を用いた end-to-end の手法が多数提案されている [4, 5, 6, 7, 8, 9]．初期の手法としては，リカレントニューラルネットワーク（Recurrent Neural Networks：RNN）やアテンションに基づく sequence-to-sequence （seq2seq）モデル [6] が提案されてきた．一方で，近年では，事前学習済みの Transformer に基づく手法が多数提案され，読解対象文に対応した流暢な問題生成を実現している [5, 10, 11, 12, 13]．

一方で，問題生成技術を読解力を育成する学習支援として活用する場合，学習者の能力に合わせた適切な難易度の問題を出題することが効果的である．このような背景から，近年，難易度調整可能な問題生成技術がいくつか提案されている [10, 14, 15, 16]．しかし，既存手法には次の問題点がある．

1. 読解対象文と答えを与えて問題を生成するため，問題とそれに対応した答えの両方を生成することはできない．
2. 問題の難易度と学習者の能力の関係を無視しているため，学習者の能力にあった適切な難易度の問題生成を行うことができない．

これらの問題を解決するために，本研究では，項目反応理論（Item response theory：IRT）[17] を用いて定量化される難易度値を与えて，問題と答えのペアを生成する新たな読解問題自動生成手法を提案する．提案手法は，BERT と GPT-2 に基づく，2 つの事前学習済み深層学習モデルを用いて構成する．

2　提案手法

本研究では，読解対象文 $\boldsymbol{w}_i = \{w_{im} \mid m \in \{1,\ldots,M_i\}\}$ とそれに関連する問題文 $\boldsymbol{q}_i = \{q_{in} \mid n \in \{1,\ldots,N_i\}\}$，およびその問題に対応する答え $\boldsymbol{a}_i = \{a_{io} \mid o \in \{1,\ldots,O_i\}\}$ で構成されるデータセット $\boldsymbol{C} = \{\boldsymbol{w}_i, \boldsymbol{q}_i, \boldsymbol{a}_i \mid i \in \{1,\ldots,I\}\}$ が与えられている場合を考える．ここで，w_{im}，q_{in}，a_{io} はそれぞれ $\boldsymbol{w}_i$，$\boldsymbol{q}_i$，$\boldsymbol{a}_i$ の m，n，o 番目の単語を表し，M_i，N_i，O_i は $\boldsymbol{w}_i$，$\boldsymbol{q}_i$，$\boldsymbol{a}_i$ の単語数を表す．また，I はデータ数を表す．本研究では，このデータセットに各問題の難易度を加えたデータセットを次のように作成し，次節で説明する提案手法の訓練データとする．

1. **各問題に対する正誤反応データの収集**：データセット $\boldsymbol{C}$ に含まれる各問題 $\boldsymbol{q}_i$ に対する解答者の正誤反応データを収集する．ただし，本研究では人間の解答者を QA（Question Answering）システムで代用する．QA システムとは，読解対象文と問題文を入力して答えを予測するシステムであり，ここでは QA システムによる解答と答え $\boldsymbol{a}_i$ の完全一致により正誤判定を行う．
2. **IRT を用いた難易度推定**：本研究では問題の難易度推定に IRT を使用する．IRT は，数理モデルを用いたテスト理論の一つであり，

様々なハイステークス試験で活用されている [18, 19, 20]．IRT では，学習者の能力と問題の難易度を関連づけて推定でき，学習者の能力にあった適切な難易度値の選定が可能である．ここでは，最も単純な IRT モデルである次式のラッシュモデルを利用して，正誤反応データから各問題の難易度値を推定する．

$$p_{ij} = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \tag{1}$$

ここで，p_{ij} は i 番目の問題における j 番目の解答者の正答確率を表し，b_i は i 番目の問題の難易度値，θ_j は j 番目の解答者の能力値を表す．ラッシュモデルでは，$\theta_j = b_i$ のときに正答確率が 0.5 となる．適応的学習では，一般に正答確率が 0.5 となる問題を与えることが有効とされているため，$b_i = \theta_j$ を難易度値として指定して，問題を生成することが望ましいといえる．

3. **難易度を含んだデータセットの作成**：データセット $\boldsymbol{C}$ に IRT で推定された難易度値を加えた新しいデータセット $\boldsymbol{C}'$ を作成する．$\boldsymbol{C}'$ は，読解対象文 $\boldsymbol{w}_i$，問題文 $\boldsymbol{q}_i$，答え $\boldsymbol{a}_i$，難易度値 b_i の集合として，以下のように表記できる．

$$\boldsymbol{C}' = \{(\boldsymbol{w}_i, \boldsymbol{q}_i, \boldsymbol{a}_i, b_i) \,|\, i \in \{1, \ldots, I\}\} \tag{2}$$

このデータセット $\boldsymbol{C}'$ を用いて，提案手法では，1）読解対象文と指定した難易度値から答えを抽出するモデルと，2）抽出された答えと読解対象文，難易度値から問題文を生成するモデル，の 2 段階で問題生成を実現する．以降で各モデルの詳細を説明する．

2.1 難易度調整可能な答え抽出モデル

難易度調整可能な答え抽出モデルでは基礎モデルに BERT（Bidirectional Encoder Representations from Transformers）[21] を用いる．BERT は，1 億以上のパラメータを持つ Transformer ベースの深層学習モデルを，33 億語以上の単語を含む文章データセットで事前学習したモデルである．事前学習は，Masked Language Model と Next Sentence Prediction の二つの教師なし学習で実現されている．BERT は，文書の分類や回帰タスクをはじめとして，系列ラベリングや抽出型文章要約のような文章からの要素抽出タスクにも広く利用されている [22]．

本研究では，BERT を基礎モデルとして，読解対象文と指定した難易度値から答えを抽出するモデルを構築する．具体的には，読解対象文 $\boldsymbol{w}_i$ と難易度値 b_i を特殊トークンで連結したデータ [CLS] b_i [SEP] $\boldsymbol{w}_i$ を入力として受け取り，読解対象文における答えの開始位置と終了位置を出力するように BERT の出力層を設計してファインチューニング（追加学習）する．ここで，ファインチューニングの損失関数は以下で定義する．

$$-\sum_{i=1}^{I}\sum_{m=1}^{M_i}\{Z_{im}^{(s)} \log P_{im}^{(s)} + Z_{im}^{(e)} \log P_{im}^{(e)}\} \tag{3}$$

ここで，$Z_{im}^{(s)}$ と $Z_{im}^{(e)}$ は読解対象文 $\boldsymbol{w}_i$ 中の m 番目の単語が答えの開始位置と終了位置である場合にそれぞれ 1 を取るダミー変数である．また，$P_{im}^{(s)}$ と $P_{im}^{(e)}$ は次式で定義される．

$$P_{im}^{(s)} = softmax(\boldsymbol{S} \cdot \boldsymbol{T}_{im}) = \frac{\exp(\boldsymbol{S} \cdot \boldsymbol{T}_{im})}{\sum_{m'}^{M_i} \exp(\boldsymbol{S} \cdot \boldsymbol{T}_{im'})} \tag{4}$$

$$P_{im}^{(e)} = softmax(\boldsymbol{E} \cdot \boldsymbol{T}_{im}) = \frac{\exp(\boldsymbol{E} \cdot \boldsymbol{T}_{im})}{\sum_{m'}^{M_i} \exp(\boldsymbol{E} \cdot \boldsymbol{T}_{im'})} \tag{5}$$

ここで，$\boldsymbol{T}_{im}$ は読解対象文 $\boldsymbol{w}_i$ における m 番目の単語に対応する BERT の出力ベクトルを表し，$\boldsymbol{S}$ と $\boldsymbol{E}$ は学習される重みベクトルを表す．

このようにファインチューニングした BERT を用いた答えの抽出は，答えの開始位置 $\hat{s}$ と終了位置 $\hat{e}$ を次式で求め，その区間の単語列を読解対象文から抽出することで行う．

$$\hat{s} = \arg\max_m P_{im}^{(s)}, \quad \hat{e} = \arg\max_m P_{im}^{(e)} \tag{6}$$

2.2 難易度調整可能な問題生成モデル

難易度調整可能な問題生成モデルでは基礎モデルに GPT-2（Generative Pre-trained Transformer 2）[23] を用いる．GPT-2 は，15 億以上のパラメータを持つ Transformer ベースの深層学習モデルを，800 万以上の文書データで事前学習した言語モデルである．事前学習は，現時点までに入力された単語列から次に出現する単語を逐次的に予測させる Language Model と呼ばれる教師なし学習で実現されている．GPT-2 は，問題生成タスクを含む様々な文章生成タスクで広く利用されている．

本研究では，GPT-2 を基礎モデルとした問題自動生成手法 [24] を，問題の難易度を調整できるように拡張する．具体的には，読解対象文 $\boldsymbol{w}_i$ と答え $\boldsymbol{a}_i$，問題文 $\boldsymbol{q}_i$，難易度値 b_i を特殊トークンで連結した以下のデータを学習に用いる．

$$b_i \ <\text{QU}>\ \boldsymbol{W}_i \ <\text{AN}>\ \boldsymbol{a}_i \ <\text{AN}>\ \boldsymbol{W}_i' \ <\text{G}>\ \boldsymbol{q}_i \tag{7}$$

ただし，$\boldsymbol{W}_i$ と $\boldsymbol{W}_i'$ はそれぞれ読解対象文 $\boldsymbol{w}_i$ 中の答え $\boldsymbol{a}_i$ 以前と以降の単語列を表し，<AN>は答えの開始と終了を表す特殊トークンである．また，<QU>

と<G>は読解対象文と問題文の開始を表す特殊トークンである．このデータを用いた GPT-2 のファインチューニングは，以下の損失関数の最小化により行う．

$$-\sum_{i=1}^{I}\sum_{n=1}^{N_i}\log\{P(q_{in}\mid q_{i1},\ldots,q_{i(n-1)},\boldsymbol{w}_i,\boldsymbol{a}_i,b_i)\} \quad (8)$$

ここで，

$$\begin{aligned}P(q_{in}\mid q_{i1},\ldots,q_{i(n-1)},\boldsymbol{w}_i,\boldsymbol{a}_i,b_i)&=softmax(\boldsymbol{G}\cdot\boldsymbol{T}^{q_{in}}_{\boldsymbol{q}_{i(n-1)}})\\&=\frac{\exp(\boldsymbol{G}\cdot\boldsymbol{T}^{q_{in}}_{\boldsymbol{q}_{i(n-1)}})}{\sum_{v'}^{V'}\exp(\boldsymbol{G}\cdot\boldsymbol{T}^{q_{in}}_{\boldsymbol{q}_{i(n-1)}})}\end{aligned} \quad (9)$$

であり，V' は GPT-2 が扱う語彙の総数，$\boldsymbol{T}^{q_{in}}_{\boldsymbol{q}_{i(n-1)}}$ は単語列 $\boldsymbol{q}_{i(n-1)}=(q_{i1},\ldots,q_{i(n-1)})$ に続いて単語 q_{in} を入力した場合の GPT-2 の出力ベクトル，$\boldsymbol{G}$ は学習される重みベクトルである．

ファインチューニングされたモデルを用いた問題文の生成は，<G>までのデータを入力として与え，次式に従って一単語ずつ生成することで行う．

$$\begin{aligned}\hat{q}_{in}&=\arg\max_v P(v\mid\hat{q}_{i1},\ldots,\hat{q}_{i(n-1)},\boldsymbol{w}_i,\boldsymbol{a}_i,b_i)\\&=\arg\max_v\left(softmax(\boldsymbol{G}\cdot\boldsymbol{T}^{v}_{\hat{\boldsymbol{q}}_{i(n-1)}})\right)\end{aligned} \quad (10)$$

3 提案手法の有効性評価実験

提案手法の有効性を評価するために，質問応答・問題生成タスクで広く利用される SQuAD データセット [25] を用いて実験を行った．SQuAD とは，Wikipedia の様々な記事（読解対象文に対応）に基づいて作成された 98,169 個の問題とそれに対応する答えで構成されるデータセットである．このデータはあらかじめ，訓練データ（90%），テストデータ（10%）に分割されている．SQuAD データセットを用いた実験手順は以下の通りである．

1. SQuAD の訓練データを用いて，精度の異なる 5 つの QA システム [21, 26, 25, 27] を構築した．
2. 5 つの QA システムに SQuAD のテストデータ中の各問題を解答させ，正誤反応データを収集した．
3. 得られた正誤反応データを用いて，式 (1) のラッシュモデルで各問題の難易度値を推定した．得られた難易度値は，6 値（-3.96，-1.82，-0.26，0.88，2.00，3.60）であり，値が小さいほど簡単な問題であることを意味する．モデルが数値入力を理解しやすいように，実数値で推定した難易度値（-3.96，-1.82，-0.26，0.88，2.01，3.60）を正の整数値（1，29，49，64，79，100）に線形変換を行なった．
4. 得られた難易度値と SQuAD のテストデータを統合して，難易度値を加えたデータセットを作成した．このデータセットを 90%と 10%に分割し，90%を提案手法の訓練データ，10%を評価用データとした．
5. SQuAD の訓練データを用いて，難易度を考慮せずに答え抽出モデルと問題生成モデルを一度ファインチューニングしたのち，手順 4 で作成した提案手法のための訓練データで難易度を考慮したファインチューニングを行なった．
6. 所望の難易度に応じた出力が行えたかを評価するために，10%の評価用データ中の読解対象文に対して，6 パタンの難易度値をそれぞれ与えて生成した問題と答えを用いて，機械による評価と人間による評価を実施した．

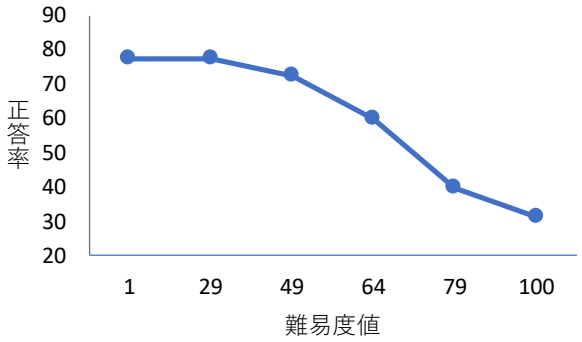

図 1 難易度別の正答率

3.1 機械による評価

上記の実験手順 6 における機械による評価は，以下の 2 つの観点で行なった．

- 生成された問題の難易度別正答率
- 抽出された答えの難易度別平均単語数

ただし，正答率の評価には 2 つの QA システム [21, 26] を使用し，先行研究 [10] と同様に 2 つの QA システムが正解した場合のみを正答として扱った．

まず，生成された問題の難易度別正答率を図 1 に示す．図から，難易度が高いほど生成された問題に対する QA システムの正答率が減少する傾向が確認できる．このことから，提案した問題生成手法が，指定した難易度を反映した問題生成を行っていることが示唆される．

次に，答えの難易度別単語数を図 2 に示す．図から，難易度が高いほど抽出された答えの平均単語数が増加する傾向が確認できる．一般に答えの単語数

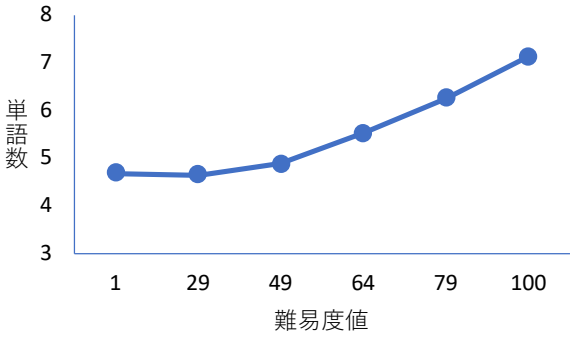

図 2　難易度別の単語数

表 1　出力された問題と答えの例

読解対象文	Deacons are called by God, affirmed by the church, and ordained by a bishop to servant leadership within the church. They are ordained to ministries of word, service, compassion, and justice. They may be appointed to ministry within the local church or to an extension ministry that supports the mission of the church. . . . Deacons serve supports the mission of the church. . . . Deacons serve a term of 2–3 years as provisional deacons prior to their ordination.
難易度値	1
問題	Who ordained deacons?
答え	bishop
難易度値	100
問題	How are deacons designated by the church?
答え	by God, affirmed by the church, and ordained by a bishop

が多くなるほど難しい問題であると予測できることから，提案手法が指定した難易度を反映した答え抽出を行っていることが示唆される．

また，出力された問題と答えの例を表 1 に示す．表から，難易度を低く指定すると，単一の用語を答えとする比較的簡単な問題が生成されたのに対し，難易度を高く指定すると，長めの文章を答えとする比較的難しい問題が生成されたことがわかる．

3.2　人間による評価

人間による評価では，評価用データからランダムに選択した 10 個の読解対象文について 6 段階の難易度別に生成された答えと問題（合計 60 ペア）を以下の 4 つの観点に基づいて，4 人の評定者で採点した．

- 流暢性：文法的な正しさや流暢さの評価．適当，不適当，許容範囲の 3 段階で評価した．
- 内容関連性：生成された問題が読解対象文の内容と関連しているかの評価．適当，不適当の 2 段階で評価した．
- 解答可能性：抽出された答えが生成された問題の正しい答えとなっているかの評価．適当，不適当，不十分，過剰の 4 段階で評価した．不十分は答えを部分的に含むが不足している場合を表し，過剰は抽出された答えに余分な部分が含まれていることを表す．
- 難易度：生成された問題の難易度の評価．1 から 5 の 5 段階で評価した．1 が最も簡単な問題を意味し，5 が最も難しい問題を意味する．

表 2　人間による評価

	適当	不適当	許容範囲
流暢性	219 (76.0%)	22 (7.6%)	47 (16.3%)

	適当	不適当
内容関連性	253 (87.8%)	35 (12.2%)

	適当	不適当		不十分	過剰
		内容関連	内容非関連		
解答可能性	194 (67.4%)	18 (6.3%)	26 (9.0%)	32 (11.1%)	18 (6.3%)

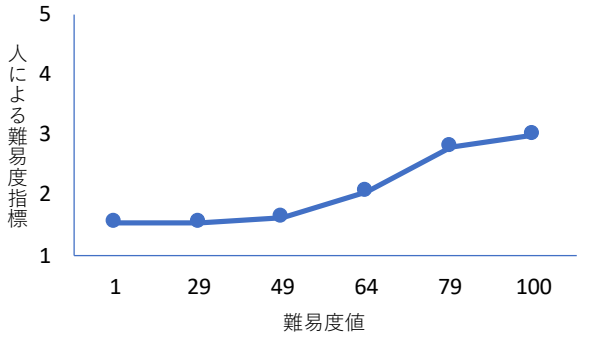

図 3　人間による難易度評価

流暢性，内容関連性，解答可能性の結果を表 2 に示す．なお，表中の「解答可能性」行に記載された「内容関連」と「内容非関連」の列は，解答可能性で不適当と判定された問題のうち，内容関連性が適当/不適当であったものの割合をそれぞれ表す．表から，7 割以上の問題が流暢な文法で生成されており，約 9 割の問題が適切に読解対象文の内容を反映していることがわかる．さらに，約 7 割のケースで解答可能な問題と答えのペアが生成できており，不十分/過剰を含めると 8 割以上のケースで少なくとも部分的には対応した問題と答えのペアが生成できたことがわかる．一方で，解答不適当と判定された残りの 1.5 割については，生成された問題と本文の内容が関連していないことが主要因であった．

次に，難易度の評価結果を図 3 に示す．図から，難易度を高く指定して生成した問題ほど人間評価者が難しいと判断したことがわかり，人間の主観的難易度にあった問題が生成できたことがわかる．

4　おわりに

本研究では，任意の難易度の問題と答えを自動生成する手法を提案し，実験から提案手法の有効性を示した．今後は，人間を対象にしたデータ収集を行ない，より厳密な評価実験を行う．

参考文献

[1] David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. Generating Natural Language Questions to Support Learning On-Line. In *Proc. Natural Language Generation*, pp. 105–114, 2013.

[2] Igor Labutov, Sumit Basu, and Lucy Vanderwende. Deep Questions without Deep Understanding. In *Proc. the Association for Computational Linguistics and Natural Language Processing*, pp. 889–898, 2015.

[3] Michael Heilman and Noah A. Smith. Good Question! Statistical Ranking for Question Generation. In *Proc. the North American Chapter of the Association for Computational Linguistics*, pp. 609–617, 2010.

[4] Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. A Review on Question Generation from Natural Language Text. *ACM Transactions on Information Systems*, 2021.

[5] Ying-Hong Chan and Yao-Chung Fan. A Recurrent BERT-based Model for Question Generation. In *Proc. Workshop on Machine Reading for Question Answering*, pp. 154–162, 2019.

[6] Xinya Du, Junru Shao, and Cardie Cardie. Learning to Ask: Neural Question Generation for Reading Comprehension. In *Proc. Annual Meeting of the Association for Computational Linguistics*, 2017.

[7] Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Yoshua Bengio, and Adam Trischler. Neural Models for Key Phrase Extraction and Question Generation. In *Proc. Machine Reading for Question Answering,*, pp. 78–88, 2018.

[8] Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. Improving Neural Question Generation Using Answer Separation. In *Proc. the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6602–6609, 2019.

[9] Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. Answer-focused and Position-aware Neural Question Generation. In *Proc. Empirical Methods in Natural Language Processing*, pp. 3930–3939, 2018.

[10] Yifan Gao, Lidong Bing, Wang Chen, Michael Lyu, and Irwin King. Difficulty Controllable Generation of Reading Comprehension Questions. In *Proc. International Joint Conference on Artificial Intelligence*, 2019.

[11] Seungyeon Lee and Minho Lee. Type-dependent Prompt CycleQAG : Cycle Consistency for Multi-hop Question Generation. In *Proc. International Conference on Computational Linguistics*, 2022.

[12] Manav Rathod, Tony Tu, and Katherine Stasaski. Educational Multi-Question Generation for Reading Comprehension. In *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*, 2022.

[13] Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. Generative Language Models for Paragraph-Level Question Generation. In *Proc. Conference on Empirical Methods in Natural Language Processing*, 2022.

[14] Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, 2021.

[15] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam AI-Emari. A Systematic Review of Automatic Question Generation for Educational Purposes. *International Journal of Artificial Intelligence in Education*, Vol. 30, pp. 121–204, 2019.

[16] Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. A similarity-based theory of controlling MCQ difficulty. In *Proc. International Conference on E-Learning and E-Technologies in Education*, pp. 283–288, 2013.

[17] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 2012.

[18] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon*, Vol. 4, No. 5, pp. 1–32, 2018.

[19] Masaki Uto. A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. *Behavior Research Methods*, 2022.

[20] Masaki Uto. A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. *Behaviormetrika*, Vol. 48, No. 2, pp. 425–457, 2021.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[22] Anirudh Srikanth, Ashwin Shankar Umasankar, Saravanan Thanu, and S. Jaya Nirmala. Extractive Text Summarization using Dynamic Clustering and Co-Reference on BERT. In *Proc. International Conference on Computing, Communication and Security*, pp. 1–5, 2020.

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI, 2019.

[24] Megha Srivastava and Noah Goodman. Question Generation for Adaptive Education. In *Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pp. 692–701, 2021.

[25] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[26] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proc. International Conference on Learning Representations*, 2020.

[27] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proc. Empirical Methods in Natural Language Processing*, pp. 193–203, 2013.