論　文

項目反応理論に基づく難易度調整可能な読解問題自動生成手法

富川　雄斗[†a)]　　鈴木　彩香[†b)]　　宇都　雅輝[†c)]

Difficulty-Controllable Neural Question Generation for Reading Comprehension Based on Item Response Theory

Yuto TOMIKAWA[†a)], Ayaka SUZUKI[†b)], and Masaki UTO[†c)]

あらまし　読解力の育成方法の一つとして，学習者に多様な文章を読ませ，それに関連する読解問題に取り組ませるアプローチが知られている．しかし，多くの読解問題を人手で作成することは時間的・費用的コストが高い．この問題を解決する技術として，読解対象文からそれに関連する問題を自動で生成する読解問題自動生成技術が知られている．そのような問題生成技術を読解力育成のための学習支援として活用する場合，各学習者に適した難易度の問題を出題することが効果的と考えられる．そのため，近年では難易度調整可能な問題生成技術が提案されている．しかし，それらの既存手法には次の課題が残る．1) 読解対象文と答えを入力として問題を生成する手法として設計されているため，問題とそれに対応した答えの両方を生成することができない．2) 学習者の能力と問題の難易度の関係性を無視しているため，学習者の能力に合わせた難易度を指定して問題を生成することができない．これらの問題を解決するために，本研究では問題難易度の推定に項目反応理論を用いることで，学習者の能力にあった難易度を選択できるようにするとともに，項目反応理論の尺度で指定した難易度に合わせて問題文と答えのペアを生成できる深層学習手法を提案する．

キーワード　読解問題自動生成，項目反応理論，深層学習，適応的学習支援，適応型テスト，自然言語処理

1. まえがき

急速な情報化の進展に伴い，様々な情報の中から必要な情報を取捨選択し，内容を正確に理解する読解力がこれまで以上に求められている．読解力の育成方法の一つとして，学習者に多様な文章を読ませ，それに関連する読解問題に取り組ませるアプローチが知られている [1]．しかし，様々な読解対象文に対する多様な読解問題を人手で作成することは時間的・費用的コストが高いという問題がある．この問題を解決する方法の一つとして，読解問題自動生成技術が近年注目を集めている．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，教育分野においては読解力の育成・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 [2]〜[4]．

† 電気通信大学大学院情報理工学研究科，調布市
The University of Electro-Communications, Chofu-shi, 182–8585 Japan
a) E-mail: tomikawa@ai.lab.uec.ac.jp
b) E-mail: suzuki_ayaka@ai.lab.uec.ac.jp
c) E-mail: uto@ai.lab.uec.ac.jp
DOI:10.14923/transinfj.2023JDP7028

従来の読解問題自動生成手法では，人手で設計したルールやテンプレートを利用して原文を問題文に変換するアプローチが主流であったが，そのような変換を行う網羅的かつ適切なルールやテンプレートの作成はコストが高い (e.g. [5]〜[8])．この問題に対し，近年では，深層学習を用いた end-to-end の読解問題自動生成手法が多数提案されている (e.g. [3], [9]〜[13])．初期の手法は，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 [10] として実現されてきた．他方で近年では，事前学習済みの Transformer [14] をベースとする手法が多数開発され，読解対象文に対応したより流暢な問題の生成を実現している (e.g. [9], [15]〜[18])．

このような問題生成技術を読解力育成のための学習支援として活用する場合，各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と考えられる [19]．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幾つか提案されている (e.g. [4], [15], [20], [21])．

従来の難易度調整可能な問題生成として，Gao et.al. [15] は，問題を easy・hard の 2 値の難易度に分類し，その難易度ラベルを seq2seq モデルのエンコーダの入力に組み込んで学習することで，2 段階で難易度を調整できる問題生成手法を提案している．また，Cheng et.al. [20] は，答えにたどり着くために参照しなければならない文の数（推論ステップ数と呼ばれる）を問題難易度とみなし，推論ステップ数を指定して問題を生成できる手法を提案している．しかし，これらの既存手法には以下の問題点が残る．

（1） 読解対象文と答えを入力として問題を生成する手法として設計されているため，問題とそれに対応した答えの両方を生成することができない．

（2） 学習者の能力と問題の難易度の関係性を無視しているため，学習者の能力に合わせた難易度を指定して問題を生成することができない．

これらの問題を解決するために本研究では，テスト理論の一つである項目反応理論（Item response theory: IRT）[22] を用いて問題難易度を定量化し，その難易度値を指定して読解対象文から問題文と答えを自動生成できる手法を提案する．IRT は問題の難易度と学習者の能力の関係性をモデル化する理論であるため，IRT を活用することで学習者の能力にあった難易度値の選定が可能となる．本研究で提案する答えと問題の生成手法は，1) 読解対象文と IRT に基づく難易度を入力として，読解対象文から答えを抽出する難易度調整可能な答え抽出モデルと，2) 読解対象文と答え，IRT に基づく難易度を入力として，問題文を生成する難易度調整可能な問題生成モデルで構成される．難易度調整可能な答え抽出モデルでは基礎モデルに Bidirectional Encoder Representations from Transformers（BERT）[23] を，難易度調整可能な問題生成モデルでは基礎モデルに Text-to-Text Transfer Transformer（T5）[24] を利用する．また，提案手法を学習するためには，(読解対象文，問題文，答え，IRT に基づく難易度) の四つ組で構成されるデータセットが必要となるが，読解問題自動生成の従来研究で利用されるベンチマークデータセットは (読解対象文，問題文，答え) の三つ組で構成されている．そこで本研究では，従来のデータセット中の各問題を様々な質問応答（Question Answering: QA）システムに解かせることで，(読解対象文，問題文，答え，IRT に基づく難易度) の四つ組で構成されるデータセットを構築する手法の提案も行う．更に，学習者の能力に合った問題を生成するためには学習者の能力が既知である必要があるが，通常の学習場面ではこの前提は満たされない．そこで，本研究では，テストの出題方式の一つとして知られる適応型テスト（Computerized Adaptive Testing: CAT）の枠組みを利用することで，学習者の能力を効率的に推定しつつ能力に合った難易度の問題を逐次的に提示するアプローチも提案する．

本研究では，読解問題自動生成タスクにおいて広く用いられている SQuAD データセット [25] を用いた実験を通して，提案手法が所望の難易度を反映した問題と答えのペアを生成できることを示す．また，CAT の仕組みにより学習者の能力を効率良く推定でき，その結果，能力に応じた難易度を指定した問題生成が可能となることを示す．

2. タスク定義

本研究の目的は，読解対象文と IRT に基づく難易度を指定して，それに応じた問題文と答えを生成することである．

ここで，読解対象文を $\boldsymbol{w} = \{w_m \,|\, m \in \{1,\ldots,M\}\}$，それに関連する問題文を $\boldsymbol{q} = \{q_n \,|\, n \in \{1,\ldots,N\}\}$，及びその問題に対応する答えを $\boldsymbol{a} = \{a_o \,|\, o \in \{1,\ldots,O\}\}$ とする．ただし，w_m，q_n，a_o はそれぞれ $\boldsymbol{w}$，$\boldsymbol{q}$，$\boldsymbol{a}$ の m，n，o 番目の単語を表し，M，N，O はそれぞれ $\boldsymbol{w}$，$\boldsymbol{q}$，$\boldsymbol{a}$ の単語数を表す．ここで，答え $\boldsymbol{a}$ は読解対象文 $\boldsymbol{w}$ の単語列の部分文字列 $\boldsymbol{a} \subset \boldsymbol{w}$ とする．これは，問題生成や質問応答の研究で使用される代表的なデータセットである SQuAD のデータ構造に合わせた条件である．

以上の設定のもとで，本研究の目的は，$\boldsymbol{w}$ と **4.** で詳述する IRT によって尺度化された難易度 b から $\boldsymbol{q}$ と $\boldsymbol{a}$ を生成することである．

3. 既存手法

本研究では，近年高精度を達成している深層学習を用いた読解問題自動生成手法 (e.g., [9], [10], [26], [27]) を基礎技術として利用する．本章では，従来の深層学習ベースの読解問題生成手法と，それを基に開発された従来の難易度調整可能な問題生成手法を紹介する．

3.1 深層学習ベース手法の読解問題自動生成

初期の手法として，Du et.al. [10] は RNN を用いた seq2seq モデルに基づく読解問題自動生成手法を提案している．この手法では，読解対象文を RNN エンコーダに入力し，出力された特徴ベクトルを元に RNN デコーダが問題文を生成する．また，Zhou et.al. [26] は

答えと単語の品詞を表す Part-of-speech（POS）タグを考慮した RNN モデルを問題生成に利用することを提案している．

一方，近年では，様々な自然言語処理タスクにおいて，事前学習済みの Transformer ベースモデルが RNN に基づく seq2seq モデルより優れた性能を達成している (e.g. [14], [24], [28]). そのため，読解問題自動生成手法としても，近年では，Transformer ベースの手法が多数提案されている (e.g. [3], [9]〜[13]). 例えば，Chan and Fan [9] は，代表的な事前学習済み Transformer モデルの一つである BERT を用いて読解対象文と答えから問題文を生成する手法を提案している．また，Lee and Lee [16] は BERT と同様に事前学習済み Transformer モデルの一つである T5 [24] を用いた読解問題自動生成手法を提案している．本研究でも BERT と T5 を使用するため，付録 **1.** に各モデルの概略を示す．

3.2 難易度調整可能な読解問題自動生成

このような問題生成技術を読解力育成のための学習支援として活用する場合，各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と考えられる [19]. そこで近年では，難易度調整可能な読解問題自動生成手法が幾つか提案されている [15], [20].

例えば，Gao et.al. [15] は難易度を easy・hard の 2 種類で指定可能なモデルを提案している．まず，訓練データの各問題を二つの QA システムに解かせ，二つの QA システムが共に正解した場合に easy，共に誤答したときに hard という難易度ラベルを各問題に付与する．なお，一つの QA システムのみが正答した問題はデータから除去する．次に，これによって得られた難易度ラベルを含んだ訓練データセット（読解対象文，問題文，難易度ラベルで構成される）を用いて RNN を用いた seq2seq 型の問題生成モデルを訓練する．このモデルは，難易度ラベルが入力の一部として与えられる設計になっており，それにより難易度にあった問題生成が実現される．

また，Cheng et.al. [20] は解答に必要な推論ステップ数を難易度とする手法を提案している．この手法では，まず読解対象文から答えを根とする知識グラフを構築する．次に，読解対象文と答え，知識グラフから解答に必要な推論ステップ数が 1 となるような問題文を生成する．そして生成した問題文に対して，知識グラフを利用して推論ステップ数が増加するように問題文を修正することを繰り返す．

しかし，これらの難易度調整可能な読解問題自動生成手法には次の問題が残る．

（1） 読解対象文と答えから問題文を生成することはできるが，読解対象文から問題文と答えを生成することができない．

（2） 学習者の能力値と問題難易度の対応関係が考慮されておらず，能力値に合った難易度を選択できない．

上記の問題を解決するために，本研究では，IRT を用いて難易度を定量化し，その難易度値を指定して読解対象文から問題文と答えを自動生成することを目指す．

4. 項目反応理論（IRT）

IRT は，コンピュータ・テスティングの普及とともに近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つであり，特にハイステークスな試験で広く活用されている．IRT では，問題への学習者の正答確率を，問題の難易度などの特性を表すパラメータと，学習者の能力値を表すパラメータの関数として定式化する．このような数理モデル（IRT モデルと呼ばれる）を用いることで，問題への学習者の反応データの集合から，学習者の能力と問題の特性を区別して推定できる．IRT の利点として，以下のような点が挙げられる [29]〜[32].

（1） 問題の難易度などの特性を考慮して学習者の能力を推定できる．

（2） 学習者の能力値と問題の特性の関係性を解釈できる．

（3） 欠測データから容易にパラメータを推定できる．

一般に IRT は，正誤判定問題や選択式問題のように正誤を一意に判定できる客観式テストへの適用を想定し，正誤で表現される 2 値型の反応データを扱う．最も基礎的な 2 値型 IRT モデルとしてはラッシュモデル [33] が知られている．本研究でも，ラッシュモデルを IRT モデルとして採用する．

4.1 ラッシュモデル

ラッシュモデルは，学習者 j が問題 i に正答する確率 P_{ij} を次式で定義する．

$$P_{ij}(u_{ij}=1|\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1+\exp(\theta_j - b_i)} \tag{1}$$

ここで，u_{ij} は学習者 j が問題 i に正答するとき $u_{ij}=1$，誤答するとき $u_{ij}=0$ となる変数，b_i は問題 i の難易

度を表すパラメータ，θ_j は学習者 j の能力値を表すパラメータである．b_i は値が大きいほど問題の難易度が高いことを表す．また，θ_j は値が大きいほど学習者の能力が高いことを表す．

ここで，難易度が異なる三つの問題について，式 (1) に基づく項目特性曲線を図 1 に示す．図の横軸は，能力値 θ，縦軸は正答確率 P_{ij}，三つの曲線は難易度が異なる三つの問題に対応する項目特性曲線である．この図より，θ が等しい場合は難易度 b が小さいほど正答確率が高くなることがわかり，b が難易度を反映していることがわかる．また，図より，能力値と難易度の値が等しいとき，すなわち $\theta = b$ のとき，正答確率が 0.5 になることも確認できる．IRT では，このように能力値と項目特性の関係性を解釈できる．

4.2 適応型テスト

上述のように，IRT では学習者の能力と問題特性の関係性が表現されるため，学習者の能力に応じた問題を適応的に出題する CAT を効果的に実現できる．本研究でも，CAT の枠組みを利用するため，ここでは CAT について説明する．

CAT は学習者の回答履歴から逐次的に能力値を推定しつつ，能力値に適した特性の問題を適応的に出題するテスト実施方式である．具体的には，能力値の最ゆう推定量の分散がフィッシャー情報量の逆数に収束する性質を利用し，能力推定値のフィッシャー情報量が最大となる問題を出題する方式が一般的である [32], [34]．CAT では，問題への正誤反応を得るたびに能力値を更新し，その能力値に基づいて次の問題を選択するプロセスを繰り返すことで，少ない問題出題数で高精度に能力を推定できる [32], [34]．なお，ラッシュモデルでは，$b = \theta$ となるときにフィッシャー情報量が最大になるため，それまでの回答履歴から能力値を推定し，その推定値に近い難易度の問題を出題すれば良い．

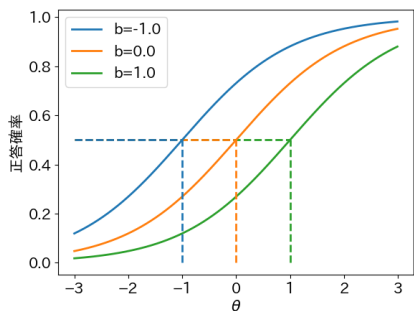

図 1　異なる難易度（$b = -1$，$b = 0$，$b = 1$）の問題に対するラッシュモデルの項目特性曲線

5. 提案手法

本研究では，IRT によって推定した難易度を基に問題を生成する手法を提案する．具体的には，IRT に基づいて推定される問題難易度を含んだデータセットを作成し，難易度に応じた答え抽出モデルと問題生成モデルを訓練する．加えて，CAT の枠組みを利用して，学習者の能力を逐次推定し，適切な難易度の問題を適応的に生成する手法も提案する．

5.1 難易度を含んだデータセットの作成

本章では提案手法で用いる，難易度を含んだデータセットの作成方法について説明する．

本研究では，質問応答・問題生成タスクで広く利用される SQuAD データセットを用いる．SQuAD は読解対象文とそれに対して作成された約 100,000 個の問題文，答えの組からなるデータセットである．SQuAD の読解対象文は Wikipedia から収集しており，問題文と答えはクラウドワーカーにより作成されている．

SQuAD データセットは $\boldsymbol{C} = \{\boldsymbol{w}_i, \boldsymbol{q}_i, \boldsymbol{a}_i \,|\, i \in \{1, \ldots, I\}\}$ で表せる．ここで，$\boldsymbol{w}_i = \{w_{im} \,|\, m \in \{1, \ldots, M_i\}\}$ は i 番目の読解対象文，$\boldsymbol{q}_i = \{q_{in} \,|\, n \in \{1, \ldots, N_i\}\}$ はそれに関連する問題文，$\boldsymbol{a}_i = \{a_{io} \,|\, o \in \{1, \ldots, O_i\}\}$ はその読解対象文と問題文に対応する答えである．また，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．なお，SQuAD では答え $\boldsymbol{a}_i$ は読解対象文 $\boldsymbol{w}_i$ の単語列の部分文字列 $\boldsymbol{a}_i \subset \boldsymbol{w}_i$ となる．

本研究では，後述する提案モデルを訓練するために，各問題に対して IRT に基づく難易度 b が付与されたデータセットが必要となるが，以上の定義のとおり，SQuAD では各問題に対する難易度は付与されていない．そこで，ここでは，難易度を含めた新たなデータセットの作成を以下の手順で行う．

（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 $\boldsymbol{q}_i$ に対する正誤反応データを収集する．ただし，人間の正誤反応データを収集することは費用的・時間的コストがかかることから，本研究では人間の解答者を複数の QA システムで代用する．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**：手順 (1) で収集し

た複数の QA システムからの正誤反応データを用いて，式 (1) のラッシュモデルに基づき各問題の難易度を推定する．推定には周辺最ゆう推定 [35] を用いる．

（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を加えた新しいデータセット $\boldsymbol{C}'$ を作成する．データセット $\boldsymbol{C}'$ は，読解対象文 $\boldsymbol{w}_i$，問題文 $\boldsymbol{q}_i$，答え $\boldsymbol{a}_i$，難易度 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から答えを抽出するモデルと，2) 読解対象文と難易度，答えから問題文を生成するモデルを訓練する．以降で各モデルの詳細を説明する．

5.2 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は，読解対象文 $\boldsymbol{w}$ と難易度 b を入力として受け取り，読解対象文中における答え $\boldsymbol{a}$ の開始位置と終了位置を予測することで答え文字列を抽出するモデルとして設計する．このような文章からの要素抽出タスクには BERT が広く利用されている (e.g. [23], [36]) ことから本研究でも答え抽出に BERT を利用する．

提案手法では BERT への入力として，読解対象文 $\boldsymbol{w}_i$ と難易度 b_i を区切りを表す特殊トークン [SEP] で連結した以下のデータを用いる．

$$b_i \,[\,\mathrm{SEP}\,]\, \boldsymbol{w}_i \tag{3}$$

そして，図 2 に示すように，読解対象文中における答えの開始位置と終了位置を予測する出力層を BERT に付与する．具体的には，読解対象文中の各単語に対して，その単語が答えの開始位置になる確率 $P_{im}^{(s)}$ と終了位置になる確率 $P_{im}^{(e)}$ を次式で求める出力層を追加する．

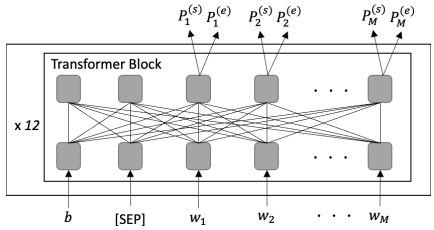

図 2　難易度調整可能な答え抽出モデルの概念図

$$P_{im}^{(s)} = softmax(\boldsymbol{S} \cdot \boldsymbol{T}_{im}) = \frac{\exp(\boldsymbol{S} \cdot \boldsymbol{T}_{im})}{\sum_{m'=1}^{M_i} \exp(\boldsymbol{S} \cdot \boldsymbol{T}_{im'})} \tag{4}$$

$$P_{im}^{(e)} = softmax(\boldsymbol{E} \cdot \boldsymbol{T}_{im}) = \frac{\exp(\boldsymbol{E} \cdot \boldsymbol{T}_{im})}{\sum_{m'=1}^{M_i} \exp(\boldsymbol{E} \cdot \boldsymbol{T}_{im'})} \tag{5}$$

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．

ファインチューニングの損失関数は以下で定義する．

$$-\sum_{i=1}^{I} \sum_{m=1}^{M_i} \{Z_{im}^{(s)} \log P_{im}^{(s)} + Z_{im}^{(e)} \log P_{im}^{(e)}\} \tag{6}$$

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．

提案手法を用いて読解対象文から答えを抽出する際には，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ればよい．

$$\hat{s} = \arg\max_m P_{im}^{(s)}, \quad \hat{e} = \arg\max_m P_{im}^{(e)}, \ (\hat{s} \le \hat{e}) \tag{7}$$

5.3 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は，読解対象文 $\boldsymbol{w}$ と難易度 b，答え $\boldsymbol{a}$ を入力とすることで，難易度に応じた問題 $\boldsymbol{q}$ を生成するモデルである．近年では，T5 を用いた読解問題自動生成が高精度を達成している (e.g. [2], [16]) ことから，提案手法は T5 を用いたモデルとして設計する．

提案手法では，入力として，読解対象文 $\boldsymbol{w}_i$ と答え $\boldsymbol{a}_i$，難易度 b_i を特殊トークンで連結した

$$b_i \,[\,\mathrm{QU}\,]\, \boldsymbol{W}_i \,[\,\mathrm{AN}\,]\, \boldsymbol{a}_i \,[\,\mathrm{AN}\,]\, \boldsymbol{W}_i' \tag{8}$$

を与え，次の形式で問題文 $\boldsymbol{q}_i$ が得られるように T5 を設計する．

$$[\,\mathrm{BOS}\,]\, \boldsymbol{q}_i \,[\,\mathrm{EOS}\,]. \tag{9}$$

ここで，$\boldsymbol{W}_i$ 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，[QU] は読解対象文の開始を表す特殊トークン，[AN] は答えの開始と終了を表す特殊トークン，[BOS] は問題文の先頭を表すトークン，[EOS] は問題文の終了を表す特殊トークンである．

本モデル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．

$$-\sum_{i=1}^{I}\sum_{n=1}^{N_i}\log\{P(q_{in}\mid q_{i1},\ldots,q_{i(n-1)},\boldsymbol{w}_i,\boldsymbol{a}_i,b_i)\} \tag{10}$$

ここで，

$$\begin{aligned}&P(q_{in}\mid q_{i1},\ldots,q_{i(n-1)},\boldsymbol{w}_i,\boldsymbol{a}_i,b_i)\\&\quad=softmax(\boldsymbol{G}\cdot\boldsymbol{T}^{q_{in}}_{pre_{i(n-1)}})\\&\quad=\frac{\exp(\boldsymbol{G}\cdot\boldsymbol{T}^{q_{in}}_{pre_{i(n-1)}})}{\sum_{v'=1}^{V'}\exp(\boldsymbol{G}\cdot\boldsymbol{T}^{q_{iv'}}_{pre_{i(n-1)}})}\end{aligned} \tag{11}$$

である．V' は T5 が扱う語彙の総数，$\boldsymbol{T}^{q_{in}}_{pre_{i(n-1)}}$ は単語列 $pre_{i(n-1)}=(\boldsymbol{w}_i,\boldsymbol{a}_i,b_i,q_{i1},\ldots,q_{i(n-1)})$ を所与とした場合の単語 q_{in} に対応する T5 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，式 (8) 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．

$$\hat{q}_{in}=\arg\max_v P(v\mid\hat{q}_{i1},\ldots,\hat{q}_{i(n-1)},\boldsymbol{w}_i,\boldsymbol{a}_i,b_i) \tag{12}$$

$\hat{q}_{in}$ は $\boldsymbol{w}_i,\boldsymbol{a}_i,b_i$ と，$n-1$ 文字目までの出力 $\hat{q}_{i1},\ldots,\hat{q}_{i(n-1)}$ を所与として，生成される確率が最も高い単語である．$\hat{q}_{in}=[\mathrm{EOS}]$ となった時点で生成を終了する．難易度調整可能な問題生成モデルの概念図は図 3 に示す．

5.4 CAT の枠組みによる能力推定と適応的問題生成

上記の提案手法により，任意の難易度を指定して問題と答えを生成することができる．適応的学習支援の研究では，学習者が 50% の確率で正答できる問題を出題することが学習の観点で効果的であることが示されている [19]．**4.1** で述べたように，ラッシュモデルでは，能力値と難易度が等しいときに正答確率が 0.5 になるため，学習者の能力値が分かれば提案手法によって，学習者の能力にあった難易度を指定して問題生成を行うことができる．しかし，通常の学習場面では学習者の能力は事前には未知であることが多い．そこで，本研究では，CAT の枠組みを利用することで，学習者の能力を効率的に推定しつつ，能力に合った難易度の問題を生成・出題する手法を提案する．

4.2 で述べたように，CAT では学習者の能力値 θ_j を効率良く測定するために，フィッシャー情報量が最大になる問題を逐次出題し，その回答を得て能力を推定するという手続きを繰り返す．本研究でも，この枠組みを採用し，次のように能力の推定と問題生成を行うことを提案する．

（1） ランダムに問題を数問生成して学習者に出題し，正誤反応データを得る．

（2） 得られた正誤反応データとラッシュモデルを用いて，学習者の能力値を推定/更新する．

（3） 推定した能力値を難易度として提案手法を用いて問題を生成・出題し，正誤反応データを得る．

（4） 手順 (2) と (3) を繰り返す．

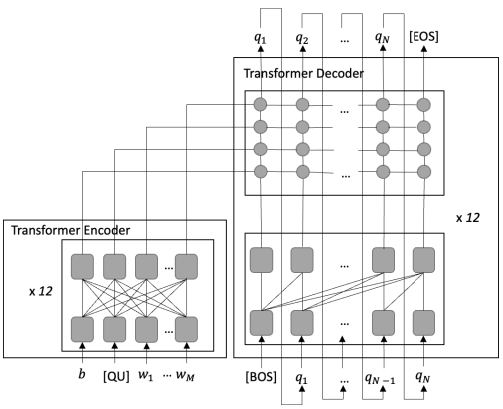

図 3 難易度調整可能な問題生成モデルの概念図

6. 提案手法の有効性評価実験

本章では提案手法の有効性評価について述べる．

6.1 実験手順

本実験では，**5.1** で説明した SQuAD データセットを用いて提案手法の性能を評価する．SQuAD データセットはあらかじめ訓練データ $\mathcal{D}^{(train)}$ (90%) とテストデータ $\mathcal{D}^{(eval)}$ (10%) に分割されている．このデータを用いた実験手順は以下のとおりである．

（1） $\mathcal{D}^{(train)}$ を用いて，精度の異なる 60 個の QA システムを構築した．具体的には huggingface (注1) で公開されている 12 個の QA モデル（BERT-base，BERT-large, RoBERTa-base, RoBERTa-large, DeBERTa-base, DeBERTa-large, DeBERTa-v3-base, DeBERTa-v3-large, ALBERT-base-v1，ALBERT-base-v2，ALBERT-large-v2，DistilBERT-base）を，$\mathcal{D}^{(train)}$ からランダムに抽出した 600，1200，1800，2400，3000 個のデータ

(注1)：https://huggingface.co/

を用いてそれぞれ訓練した．ここでは，異なる性質・性能をもつ QA モデルを，異なるサイズのデータサブセットで訓練したモデルを多数用意することで，異なる能力をもつ学習者集団の代替とみなしている．

（2） 60 個の QA システムに $\mathcal{D}^{(eval)}$ 中の各問題を解答させ，QA システムの解答と答え $\boldsymbol{a}_i$ の完全一致により正誤反応データを収集した．

（3） 収集した正誤反応データを用いて，ラッシュモデルで $\mathcal{D}^{(eval)}$ 中の各問題の難易度を推定した．

（4） 推定された難易度と $\mathcal{D}^{(eval)}$ を統合して，難易度を加えたデータセット $\mathcal{D}_b$ を作成した．なお，BERT や T5 の入力にはテキストが想定されているため，数値を扱いやすいように実数値で推定された難易度の小数第 1 位までを用いた．

（5） 作成したデータセット $\mathcal{D}_b$ を更に 90% と 10% に分割し，それぞれ $\mathcal{D}_b^{(train)}$，$\mathcal{D}_b^{(eval)}$ とした．以降では $\mathcal{D}_b^{(train)}$ を難易度調整可能な問題生成用の訓練データ，$\mathcal{D}_b^{(eval)}$ をテストデータとして扱う．

（6） まず，SQuAD の元々の訓練データ $\mathcal{D}^{(train)}$ を用いて，答え抽出モデルと問題生成モデルを難易度を考慮せずにファインチューニングした．この手順は必須ではないが，事前に難易度を考慮しないモデルを大量のデータで学習しておくことで，答え抽出と問題生成の基本性能が向上すると期待できる．

（7） 続いて，手順 (5) で作成した $\mathcal{D}_b^{(train)}$ で難易度を考慮した答え抽出モデルと問題生成モデルをファインチューニングした．このファインチューニングは，手順（6）で推定されたモデルパラメータを初期値として実施した．なお，手順（6）と（7）の問題生成モデルの訓練時には，答え抽出モデルの予測ではなく，訓練データ中の真の答えを入力としている．

（8） 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，$\mathcal{D}_b^{(eval)}$ 中のそれぞれの読解対象文に対し，−3.0 から 3.0 まで 0.1 刻みで難易度を指定して，難易度を考慮した答え抽出モデルによって答えを抽出した．続いて，指定した難易度と抽出した答えを，難易度を考慮した問題生成モデルに入力することで問題を生成した．以上の手順で生成された問題群と答え群を用いて，機械による評価と人間による評価を実施した．

6.2 機械による評価

機械による評価は，上記の実験手順 (8) で生成された問題群と答え群を以下の三つの観点で評価することで行った．

- 生成された問題群に対する QA システムの難易度別平均正答率
- 抽出された答え群の難易度別平均単語数
- 生成された問題群における先頭単語の疑問詞の難易度別出現割合

まず，生成された問題群に対する QA システムの難易度別平均正答率を図 4 に示す．横軸は難易度，縦軸は正答率を表している．なお，正答率は実験手順 (1) で訓練した 60 個の QA システムの平均正答率である．図から，指定する難易度が高いほど生成された問題群に対する QA システムの平均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．

ここで，IRT 尺度の難易度値に応じた問題が生成できたかを確認するために，$D_b^{(eval)}$ 中のそれぞれの読解対象文に対して $b = 1.0$ と指定して生成された問題群に対する 60 個の QA システムの正答率を図 5 に示した．図の横軸は能力値 θ，縦軸は正答率，橙線はロジスティック回帰モデルを最小 2 乗法によってフィッ

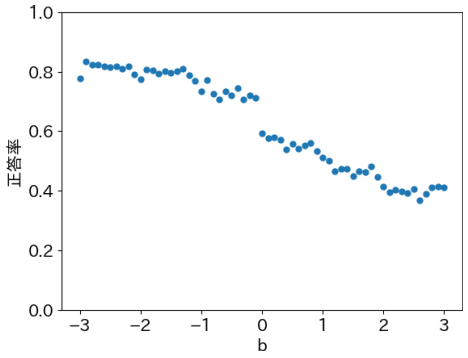

図 4　各難易度の問題群に対する平均正答率

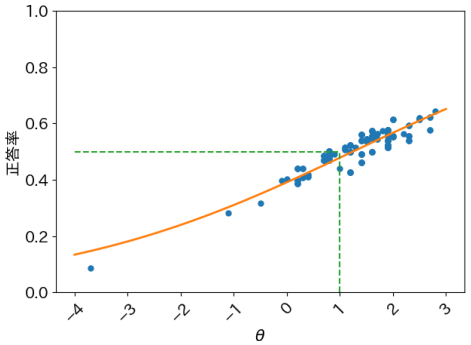

図 5　$b = 1.0$ の問題群に対する QA システムの平均正答率

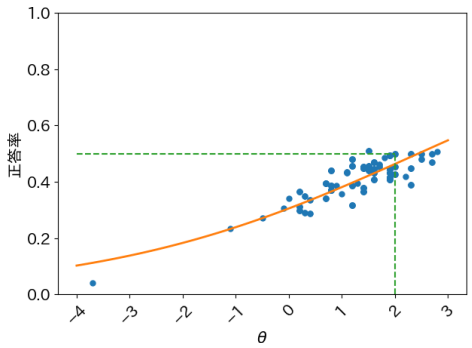

図 6　$b = 2.0$ の問題群に対する QA システムの平均正答率

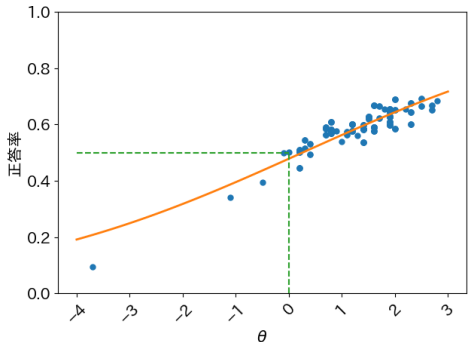

図 7　$b = 0.0$ の問題群に対する QA システムの平均正答率

ティングした関数を表す．ラッシュモデルでは $\theta_j = b_i$ のときに正答確率が 0.5 となることを踏まえ，緑破線は $\theta = 1.0$，正答率 = 0.5 を表す．図より，指定した難易度 $b = 1.0$ に対応する $\theta = 1.0$ 付近の QA システムが，正答率 0.5 程度を示しており，おおむね期待通りの問題生成ができていることが読み取れる．他の難易度でもおおむね同様の傾向が確認できた．例として，$D_b^{(eval)}$ 中のそれぞれの読解対象文に対して $b = 2.0$ と $b = 0.0$ と指定して生成された問題群に対する QA システムの平均正答率をそれぞれ図 6 と図 7 に示す．ここで，$b = 2.0$ の高難易度の問題群において，θ が 2.0 以上の QA システムの平均正答率が回帰曲線（ロジスティック回帰モデルをフィッティングした曲線）を下回る傾向が確認できる．この要因として，高難易度の問題群では，能力値全域において，同水準の能力をもつ QA システムの正答率にばらつきが生じやすくなっていることが挙げられる．このことを確認するために，表 1 に各 QA システムの平均正答率と回帰曲線との残差絶対値の平均を難易度帯別に示した．この表から，生成時に指定する難易度が高いほど QA システムの平均正答率が回帰曲線からばらつきやすくなっていることが確認できる．このことを踏まえると，$b = 2.0$ の問題群において，θ が 2.0 以上の QA システムの平均正答率が回帰曲線を下回って見えるのは，それらの QA システムの本質的な性能低下などが主要因ではなく，θ が 2.0 未満の一部の QA システムの平均正答率が誤差のレベルで若干上振れしているためと考えられる．

表 1　回帰曲線との残差絶対値の難易度帯別平均

難易度 b	残差絶対値の平均
$-3.0 \leq b < -2.0$	0.020
$-2.0 \leq b < -1.0$	0.022
$-1.0 \leq b < 0.0$	0.025
$0.0 \leq b < 1.0$	0.028
$1.0 \leq b < 2.0$	0.028
$2.0 \leq b \leq 3.0$	0.034

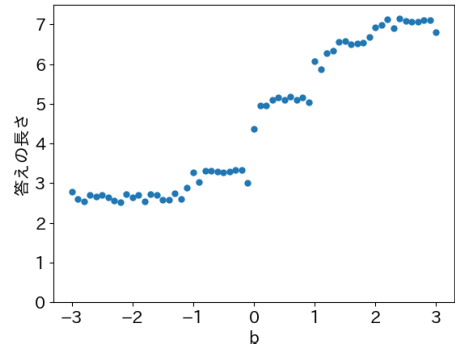

図 8　難易度別の平均単語数

次に，答え群の難易度別平均単語数を図 8 に示す．横軸は難易度，縦軸は答えの長さの平均を表す．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，提案手法によって難易度を反映した答え抽出が実現できたことが示唆される．また，出力された問題と答えの例を表 2 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．

次に，難易度調整が生成される問題文の特徴にどのように影響するかを調べるために，生成された問題文の先頭単語として What, Who, When, Why, Where, Which, How の 7 種の疑問詞が出現した割合を難易度別に求めた．結果を図 9 に示す．図の横軸は指定した難易度，縦軸は難易度別に生成された問題群において

表 2　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	−3.0
問題	How long do provisional deacons serve as?
答え	2–3 years
難易度	3.0
問題	What role are deacons assigned by the church?
答え	servant leadership within the church

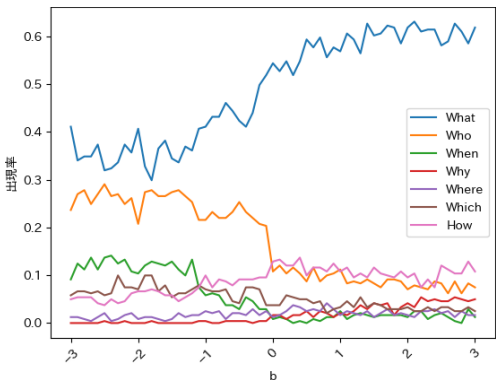

図 9　問題群の先頭単語における各疑問詞の難易度別出現割合

各疑問詞が先頭単語として出現した割合を表す．この図から，難易度を低く指定した場合は，特定の要素を聞く When や Who，Which から始まる問題文が，難易度を高く指定した場合に比べて生成されやすいことがわかる．反対に，難易度を高く指定した場合には，理由や方法を読み取って回答する必要がある Why や How から始まる問題文が，難易度を低く指定した場合に比べて生成されやすいことが確認できる．一般に，特定の要素を聞く問題よりも，理由や方法を読み取る問題の方が難しいと考えられるため，この結果は，難易度調整の効果として適当であると考えられる．なお，What については問題の難易度を高く指定した場合に出現率が高くなる傾向がある．これは，What には特定の要素を聞く用法だけでなく，How や Why のように理由や方法などを問う用法も多く存在するためと考えられる．

6.3　人間による評価

人間による評価では，上記の実験手順 (8) で生成された問題群の質と難易度を評価した．具体的には，生成された問題文と答えを難易度がバラつくようにランダムに 20 ペア選び，(読解対象文，問題文，答え) の三つ組で構成される問題セットを作成した．同様の方法で問題セットを計 5 セット作成し，それぞれのセットに対して 2 人の評定者を割り当てて，以下の評価観点で評価をさせた．

- 流暢性：文法的な正しさと流暢さの評価．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文の正しい答えとなっているかの評価．適当，不適当，不足，過剰の 4 段階で評価した．「不足」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．
- 実用可能性：問題文または答えを少し修正すれば問題として成立するかの評価．成立済み，成立する，成立しないの 3 段階で評価した．
- 問題難易度：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 が最も簡単であることを意味し，5 が最も難しいことを意味する．

なお，評定作業は，TOEIC900 点相当以上の英語スキルをもつクラウドワーカーに依頼した．

まず，流暢性，内容関連性，解答可能性，実用可能性の結果を表 3 に示す．表の数値は，評価観点ごとに，全 100 問を通して 2 名の評定者が各評価をつけた総数と割合（カッコ内）を表す．表から，9 割以上の問題が流暢または許容範囲な流暢性で生成されており，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．更に，6 割以上のケースで解答可能な

表 3　人間による評価の結果

流暢性	適当	許容範囲	不適当	
	128 (64.0%)	61 (30.5%)	11 (5.5%)	
内容関連性	適当	不適当		
	181 (90.5%)	19 (9.5%)		
解答可能性	適当	不適当	不足	過剰
	130 (65.0%)	18 (10.5%)	31 (15.5%)	21 (9.0%)
実用可能性	成立済み	成立する	成立しない	
	128 (64.0%)	52 (26.0%)	20 (10.0%)	

問題と答えのペアが生成できており，過剰や不足を含めると約 9 割の問題が部分的には適切とみなせることがわかる．このことは，実用可能性の評価において，少しの修正で成立する問題を合わせると約 9 割が肯定的な評価であったことからも確認できる．他方で，問題生成時に指定した難易度と評定者による難易度評価の Spearman の順位相関係数を確認したところ，0.15 と低い値となった．なお，相関係数の計算は，問題の難易度と各評定者の主観評価値をペアにしたデータに基づいて行った．具体的には，100 問に対する 2 名の評定者の評価値を，200 対の（問題難易度，人間評定者の評価値）のデータとして相関係数を計算した．

ここで，評定者間の評価の一致度を確認するために，2 人の評定者が付与した評価の一致率を各評価観点について確認したところ，流暢性が 0.63，内容関連性が 0.85，解答可能性が 0.53，実用可能性が 0.60，難易度が 0.22 であった．この結果より，流暢性，内容関連性，解答可能性，実用可能性については，比較的高い一致率を示している一方で，問題難易度は著しく一致率が低いことが確認できる．また，2 人とも解答可能と判定した問題のみを対象に評定者間の難易度評価値の Spearman の順位相関係数を算出したところ 0.16 と低かった．このことから，問題難易度の主観評価は評定者間でのバラつきが大きく，妥当な結果を得ることが難しいと判断できる．

そこで，より信頼性の高い難易度評価を行うために，生成された問題群を被験者に出題してその正答率に基づいて難易度を分析する実験を行った．具体的には，次の手順で実験を行った．1）まず，上記の実験で 2 人の評定者が共に解答可能と判定した問題の中から，生成時に指定した難易度がバラつくようにランダムに 30 問を選別した．2）先の実験とは異なる 10 人の被験者に，選定された 30 問からそれぞれに異なる 20 問を出題して回答を得た．本実験では多様な能力の被験者を必要とするため，TOEIC600 点以上相当の英語スキルをもつクラウドワーカーに依頼した．このとき各被験者への問題の割り当ては，各問題の出題回数が均等になるように行った．3）得られた回答を著者らが人手で採点した．人間による解答は QA システムの解答に比べて，冠詞や前置詞の有無など非本質的な揺れが生じやすいためである．今回選定した問題では，冠詞，敬称，前置詞の有無は本質的な正誤に影響しなかったため，これらの有無の差は許容して採点を行った．なお，これまでの実験で行っていた QA システムの解答評価でも，同様の揺れが正誤に影響している可能性があるが，このような揺れを考慮した解答自動評価は今後の課題とする．

各問題に対する人間被験者の正答率と生成時に指定した問題難易度について Spearman の順位相関係数を確認したところ −0.67 であった．また，無相関検定の結果，p 値が 0.01 を下回ったことから，1% 水準で有意な相関であることが確認できた．このことから，生成時に指定する難易度と実際の問題の難易度には一定の相関があり，人間が感じる難易度に即した問題が生成できていることがわかる．

6.4 適応的問題生成に基づく能力推定の性能評価

ここでは，**5.4** で述べた CAT の枠組みに基づく能力推定と適応的問題生成の有効性を評価する．ここでは，60 個の QA システムの θ の最小値が −3.658，平均値が 1.232，最大値が 2.766 であったことを踏まえ，θ が最小と最大の QA システム，及び，平均値に最も近い $\theta = 1.244$ の QA システムをそれぞれ学習者と見立てて，次の二つの方法で問題の生成を行い，能力推定値の変化を分析した．

適応的生成：最初にランダムな難易度の問題を 10 問出題し，それ以降は推定した θ に一番近い難易度の問題を 40 問生成・出題する．

ランダム生成：ランダムな難易度の問題を 50 問生成・出題する．

これらの方法で推定した θ の変化をそれぞれ図 10，11，12 に示す．図の横軸は出題数，縦軸は θ の値を表す．青の線は真値を表し，橙色と緑線は適応的出題とランダム出題による能力推定値の遷移を表す．なお，図では上記の実験を 500 回行い，各出題数ごとに θ の推定値の平均を取っている．

これらの結果より，ランダムな難易度の問題を出題するよりも，推定した能力値に応じた難易度の問題を出題する方が，より少ない出題数で真の能力値に効率良く近づいていることがわかる．このことは，CAT の枠組みによる出題を採用することで，能力にあった難易度の問題出題を提案手法でより効率的に行えることを意味している．

また，全ての QA システムを対象とした性能評価も行った．具体的には，上記の実験を 60 個の全てのモデルに対して 500 回行い，真の能力値と 50 問出題後の能力推定値の平均絶対誤差を次式で計算した．

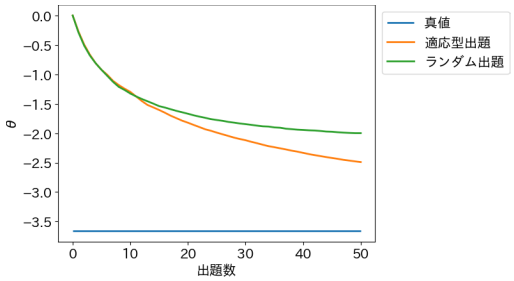

図 10　能力真値が −3.658 (最小の θ) の場合の能力推定値の変化

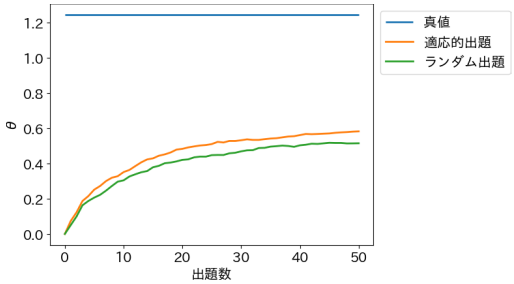

図 11　能力真値が 1.244 (平均値に最も近い θ) の場合の能力推定値の変化

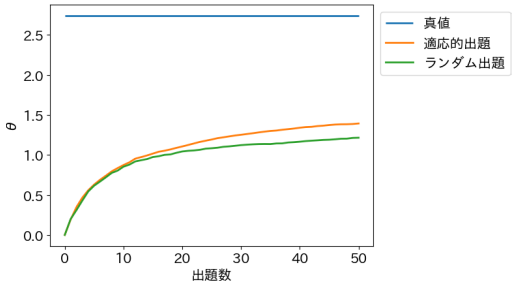

図 12　能力真値が 2.766 (最大の θ) の能力推定値の変化

$$MAE = \frac{1}{60 \times 500} \sum_{d=1}^{60} \sum_{k=1}^{500} |\hat{\theta}_{dk} - \theta_d| \qquad (13)$$

ここで，θ_d は $d \in \{1,\ldots,60\}$ 番目のモデルの真の能力値，$\hat{\theta}_{dk}$ は $k \in \{1,\ldots,500\}$ 回目の実験で得られた d 番目のモデルの能力推定値を表す．実験の結果，平均絶対誤差は，提案手法が 0.707，ランダム生成が 0.744 となった．このことから，全ての QA システムを対象とした性能評価でも，提案手法はランダム手法よりも高い精度で能力推定が行えていることが確認できた．

7. むすび

本研究では，IRT に基づく難易度を指定して問題と答えのペアを自動で生成する手法と，それを活用した適応的な問題生成手法を提案した．SQuAD データセットを用いた機械による評価実験により，提案手法は難易度を適切に調整した問題生成が可能であることを示した．また，人間による評価実験により，流暢性や内容関連性などの観点で適当な問題生成が行えていることを示すとともに，人間の能力に沿った難易度の問題生成が行えていることを示した．加えて，学習者の能力値 θ が未知という一般的な状況でも，適応的生成によってランダムに問題を出題する場合よりも効率良く θ を推定でき，能力にあった問題生成を効果的に行えることを示した．

また，**6.2** では，提案手法で生成された問題文の先頭の疑問詞が，指定する難易度によって異なる出現傾向を示すことを述べた．このことは，提案手法が，難易度に関連する問題や答えの特徴を難易度制御と連動して操作できる可能性を示唆している．提案手法では，問題の問い方や文法，話題選定，答えの語彙選択といった要因を明示的には制御していないが，人間による問題と答えの作成傾向を難易度と紐づけて学習するようにモデルが設計されているため，難易度に関連する一部の要因も難易度調整と連動して制御できたと推測される．他方で，このような多様な要因も直接的に制御することができれば，教育効果や技術活用の柔軟性を高める上でより効果的と考えられる．そのため，今後は難易度に加えてこれらの要因も制御可能な問題生成手法の開発を検討したい．なお，**6.2** では，問題文の先頭単語だけに着目して，生成される問題の特徴分析を行った．今後は構文解析などの言語処理技術を活用してより詳細に問題の問い方を類型化して，難易度との関係を調査したい．

提案手法は，訓練データ中の問題や答えの傾向を模倣するように訓練されるため，訓練データが読解力育成に有効な問題と答えで構成されていれば，それを模倣しようとする提案手法も学習に効果的な問題や答えを生成できると考えられる．他方で，本研究で用いた SQuAD データセットは一般的な質問応答タスク用に構築されたデータセットであるため，読解力育成に必ずしも効果的な問題と答えで構成されていない可能性がある．教育的な視点から作成されたデータセットを用いて提案モデルを構築することで，学習効果をより高める問題生成が行えると期待できるため，今後はそのようなデータセットの調査や構築を検討するとともに，データセットの特徴が教育効果に及ぼす影響につ

いても分析をしていきたい．

また，**6.2** では，問題の難易度が高くなるほど，生成された問題群に対する QA システムの平均正答率が，類似した能力値の QA システム間でばらつきやすなることも述べた．この理由としては，難易度が上がるほど，問題の問い方が複雑になり，答えも長くなる傾向があるため，QA システムの性能に揺れが生じやすくなるためと推察する．この理由についても，今後より詳細に分析したい．

更に，今後は，QA システムにより収集していた正誤反応データを人間の解答者から収集し，その正誤反応データを用いて QA システムで解答者を代替することの妥当性をより詳細に評価していきたい．更に，本研究では生成する問題の難易度を調整することを目指しているが，実際には読解対象文の難易度が問題の難易度にも影響することが考えられる．学習場面での利用を想定した場合，問題の難易度を学習者の能力に合わせるだけでなく，適切な難易度の読解対象文を選定することも重要であると考えられるため，今後は，能力に応じた読解対象文の選択手法や読解対象文の難易度も考慮した問題生成手法についても検討したい．

謝辞　本研究は JSPS 科研費 19H05663，20K20817，21H00898 の助成を受けたものです．

文　献

[1] 徳本浩子，“授業時間外のオンライン課題導入実践と英語読解力向上の相関性について，” ICT 利用による教育改善研究発表会，2011.

[2] B. Ghanem, L.L. Coleman, J.R. Dexter, S.M. von derOhe, and A. Fyshe, “Question generation for reading comprehension assessment by modeling how and what to ask,” Findings of the Association for Computational Linguistics, pp.2131–2146, 2022.

[3] R. Zhang, J. Guo, L. Chen, Y. Fan, and X. Cheng, “A review on question generation from natural language text,” ACM Trans. Information Systems, vol.40, no.1, pp.1–43, 2021.

[4] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. AI-Emari, “A systematic review of automatic question generation for educational purposes,” International Journal of Artificial Intelligence in Education, vol.30, no.1, pp.121–204, 2019.

[5] R. Mitkov and L.A. Ha, “Computer-aided generation of multiple-choice tests,” Proc. Building educational applications using natural language processing, pp.17–22, 2003.

[6] D. Lindberg, F. Popowich, J. Nesbit, and P. Winne, “Generating natural language questions to support learning on-line,” Proc. Natural Language Generation, pp.105–114, 2013.

[7] I. Labutov, S. Basu, and L. Vanderwende, “Deep questions without deep understanding,” Proc. the Association for Computational Linguistics and Natural Language Processing, pp.889–898, 2015.

[8] M. Heilman and N.A. Smith, “Good question! statistical ranking for question generation,” Proc. North American Chapter of the Association for Computational Linguistics, pp.609–617, 2010.

[9] Y.-H. Chan and Y.-C. Fan, “A recurrent BERT-based model for question generation,” Proc. Workshop on Machine Reading for Question Answering, pp.154–162, 2019.

[10] X. Du, J. Shao, and C. Cardie, “Learning to ask: Neural question generation for reading comprehension,” Proc. Annual Meeting of the Association for Computational Linguistics, pp.1342–1352, 2017.

[11] S. Subramanian, T. Wang, X. Yuan, S. Zhang, Y. Bengio, and A. Trischler, “Neural models for key phrase extraction and question generation,” Proc. Machine Reading for Question Answering,, pp.78–88, 2018.

[12] Y. Kim, H. Lee, J. Shin, and K. Jung, “Improving neural question generation using answer separation,” Proc. AAAI Conference on Artificial Intelligence, pp.6602–6609, 2019.

[13] X. Sun, J. Liu, Y. Lyu, W. He, Y. Ma, and S. Wang, “Answer-focused and Position-aware neural question generation,” Proc. Empirical Methods in Natural Language Processing, pp.3930–3939, 2018.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. International Conference on Neural Information Processing Systems, pp.5998–6008, 2017.

[15] Y. Gao, L. Bing, W. Chen, M. Lyu, and I. King, “Difficulty controllable generation of reading comprehension questions,” Proc. International Joint Conference on Artificial Intelligence, pp.4968–4974, 2019.

[16] S. Lee and M. Lee, “Type-dependent Prompt CycleQAG: Cycle consistency for multi-hop question generation,” Proc. International Conference on Computational Linguistics, pp.6301–6314, 2022.

[17] M. Rathod, T. Tu, and K. Stasaski, “Educational multi-question generation for reading comprehension,” Proc. Workshop on Innovative Use of NLP for Building Educational Applications, pp.216–223, 2022.

[18] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, “Generative language models for paragraph-level question generation,” Proc. Conference on Empirical Methods in Natural Language Processing, pp.670–688, 2022.

[19] M. Ueno and Y. Miyazawa, “IRT-based adaptive hints to scaffold learning in programming,” IEEE Trans. Learning Technologies, vol.11, no.4, pp.415–428, 2018.

[20] Y. Cheng, S. Li, B. Liu, R. Zhao, S. Li, C. Lin, and Y. Zheng, “Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting,” Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing, pp.5968–5978, 2021.

[21] T. Alsubait, B. Parsia, and U. Sattler, “A similarity-based theory of controlling MCQ difficulty,” Proc. Int. Conf. E-Learning and E-Technologies in Education, pp.283–288, 2013.

[22] F.M. Lord, Applications of item response theory to practical testing problems, Routledge, 2012.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

pp.4171–4186, 2019.
[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” J. Machine Learning Research, vol.21, no.1, pp.5485–5551, 2020.
[25] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ questions for machine comprehension of text,” Proc. Empirical Methods in Natural Language Processing, pp.2383–2392, 2016.
[26] Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou, “Neural question generation from text: A preliminary study,” National China Computer Federation Conference on Natural Language Processing and Chinese Computing, pp.662–671, 2017.
[27] L.E. Lopez, D.K. Cruz, J.C.B. Cruz, and C. Cheng, “Simplifying paragraph-level question generation via Transformer language models,” Pacific Rim International Conference on Artificial Intelligence, pp.323–334, 2021.
[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., “Language models are unsupervised multitask learners,” OpenAI, 2019.
[29] M. Uto and M. Ueno, “Empirical comparison of item response theory models with rater’s parameters,” Heliyon, vol.4, no.5, p.e00622, 2018.
[30] M. Uto, “A bayesian many-facet rasch model with markov modeling for rater severity drift,” Behavior Research Methods, pp.1–19, 2022.
[31] M. Uto, “A multidimensional generalized many-facet Rasch model for rubric-based performance assessment,” Behaviormetrika, vol.48, no.2, pp.425–457, 2021.
[32] 加藤健太郎，山田剛史，川端一光，R による項目反応理論，オーム社，2014.
[33] G. Rasch, Probabilistic models for some intelligence and attainment tests., ERIC, 1993.
[34] 宮澤芳光，宇都雅輝，石井隆稔，植野真臣，“測定精度の偏り軽減のための等質適応型テストの提案，” 電子情報通信学会論文誌 D, vol.101, no.6, pp.909–920, 2018.
[35] R.D. Bock and M. Aitkin, “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm,” Psychometrika, vol.46, no.4, pp.443–459, 1981.
[36] Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders,” Proc. Empirical Methods in Natural Language Proc. International Joint Conference on Natural Language Processing, pp.3730–3740, Nov. 2019.
[37] K. Clark, M.-T. Luong, Q.V. Le, and C.D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” Proc. Int. Conf. Learning Representations, pp.1–18, 2020.
[38] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” Proc. Annual Meeting of the Association for Computational Linguistics, pp.7871–7880, 2020.

付　　録

1. 事前学習済み深層学習モデル

ここでは，本研究で利用している事前学習済みの Transformer ベース深層学習自然言語処理モデルについて紹介する．事前学習済みモデルとは，大規模な言語コーパスを使用した自己教師あり学習によって，汎用的な言語構造を事前学習させた深層学習モデルである．近年では，様々なモデル [14], [23], [24], [28], [37], [38] が提案されており，幅広いタスクに対して高精度を達成している．以下では本研究で使用する事前学習済み深層学習モデルの BERT と T5 について説明する．

1.1 BERT

BERT は，Transformer のエンコーダ部分を利用したモデルである．1 億以上のパラメータをもち，33 億語以上の単語を含む文章データセットで事前学習されている．事前学習は，文章中の単語をランダムに欠損（Mask）させた入力に対して，欠損前の単語を予測する Masked Language Model と，与えられた 2 文が連続する 2 文であるか識別する Next Sentence Prediction と呼ばれる二つの自己教師あり学習で実現されている．事前学習によって汎用的な自然言語構造を獲得しているため，タスクに応じた出力層を追加してファインチューニングすることで，文書の分類や要素抽出などの様々なタスクで高い精度の予測を行うことができる．

1.2 T5

T5 は，Transformer のエンコーダとデコーダ部分を利用した自然言語生成モデルである．入力文を BERT 同様の双方向型のアテンションをもつ Transformer エンコーダで処理し，その出力を単方向型のアテンションをもつ Transformer デコーダに与えることで文章を生成する．T5 は，110 億以上のパラメータをもち，750GB の文書データで事前学習されている．事前学習には BERT の Masked Language Model と類似した Denoising Objective が採用されている．また，T5 は，様々な文章生成タスクを解くことができるように，入力にタスク指示を含めたマルチタスク事前学習がなされている点も特徴である．T5 は，翻訳や文章要約などの様々な文章生成タスクで高性能を達成している．

（2023 年 5 月 12 日受付，9 月 23 日再受付，
10 月 26 日早期公開）

富川　雄斗

2023 電気通信大学情報理工学域卒．同年，電気通信大学大学院情報理工学研究科情報・ネットワーク工学専攻入学，現在に至る．問題自動生成の研究・開発に従事．

鈴木　彩香

2022 電気通信大学情報理工学域卒．同年，電気通信大学大学院情報理工学研究科情報・ネットワーク工学専攻入学，現在に至る．問題自動生成の研究・開発に従事．

宇都　雅輝　（正員）

2013 電気通信大学大学院情報システム学研究科博士後期課程了．博士 (工学)．長岡技術科学大学特任助教を経て，2015 電気通信大学助教に着任，2020 に同大学准教授となり，現在に至る．e テスティング，e ラーニング，人工知能，ベイズ統計，自然言語処理などの研究に従事．