語彙習得段階を考慮した 英文法多肢選択問題の誤答選択肢自動生成

戸崎　友輔† 　宮森　恒†

† 京都産業大学 情報理工学部　〒 603–8555 京都府京都市北区上賀茂本山
E-mail: †{g2054650,miya}@cc.kyoto-su.ac.jp

あらまし　本稿では，語彙習得段階を考慮した英文法多肢選択問題における誤答選択肢の自動生成手法を提案する．英語の多肢選択問題における選択肢は，与えられた問題文にふさわしく，既習の語彙で構成する必要がある．一方，問題作成者にとって，これら条件を満たした選択肢となっているかを逐一確認する負担は大きく，誤答選択肢を自動生成する研究が行われてきた．しかし，従来手法では，誤答選択肢が正解となってしまう可能性を必ずしも排除できていないという課題があった．本稿では，英文法多肢選択問題を取り上げ，語彙習得段階を考慮しつつ，誤答選択肢が正解とならないような枠組みを導入した誤答選択肢の生成手法を提案する．具体的には，事前学習済みモデルを用いて誤答選択肢候補を生成し，文章の自然さを表すスコアを利用して誤答選択肢候補が正解とならないと判断された候補のみに対して，語彙習得段階を考慮したスコアを用いたランキングにより誤答選択肢を選択する．実験では，問題文と生成した誤答選択肢の適切さ，語彙の習得段階としての適切さを評価し，人間によって作成された選択肢との大きな違いがないことを確認する．

キーワード　自動問題生成，多肢選択問題，事前学習済みモデル，Edtech，教育

1　はじめに

昨今，日本の教育現場では教員の業務量の増大が問題となっている [1]．その業務の中には，試験問題の作成が含まれる．試験問題の作成の課題点として，人手作業では時間や労力がかかってしまうことがあげられる．そこで最近では，自動問題生成（Automatic Question Generation，AQG）に関する研究が取り組まれてきている [2]- [5]．AQG は，自然言語処理や機械学習などの技術を活用し，試験問題を自動生成する技術全般を指す．AQG の主な手法として，ルールベースでの自動生成 [6]，機械学習での自動生成 [7] などが挙げられる．AQG により，人手作業による問題作成のコスト削減，多様な問題の大量かつ素早い作成，解答者の学習段階に合わせた問題作成を実現することが期待できる．

試験問題の形式としてよく用いられているものに，多肢選択問題（Multiple Choice Question，MCQ）がある．MCQ は複数の選択肢から正解の選択肢を選び，答える形式の問題である．MCQ は図 1 のように，問題文（Stem）と選択肢群（Options）で構成される．選択肢群は正解選択肢（Answer）と誤答選択肢（Distractor）から成る．MCQ は，1 つの文章で構成される「短文」に対して解答する形式の問題，複数の文章で構成される「長文」に対して解答する形式の問題がある．短文に対して解答する形式の問題では，言語の文法や語彙・語句について問うものが多い．一方で，長文に対して解答する形式の問題では，長文を読解した内容について問うものが多い．MCQ の形式は，機械による自動採点ができる，理解度の客観的評価ができる，問題作成の効率が良いなどといった点で優れている．しかし，誤答選択肢は解答者に適度に考えさせることができるように，問題文に適したものを採用する必要がある．これを満たすには，誤答が解答者にとって既習であるかどうかや，問題文や空欄，正答等の状況を踏まえてその誤答が選択肢として適切であるかなど，問題作成者が一つ一つ確認する必要があり，負担が大きい点が課題としてあげられる．

図 1　MCQ の構成

そこで，MCQ の誤答選択肢生成に関する研究も取り組まれてきている [10] [11]．しかし，従来の誤答選択肢生成手法では，本来誤答となる選択肢が，正解となってしまう可能性を必ずしも排除できていない問題がある．図 2 は誤答選択肢であるが正解となってしまう例である．問題文 “They ____ to go home.” に対して，実際の正解選択肢は “want” であるが，誤答選択肢の “have” を空欄に当てはめると文章の意味が成り立ち，正解となってしまう．

そこで本稿では，語彙習得段階を考慮しつつ，誤答選択肢が正解となる可能性を排除する枠組みを導入した誤答選択肢の生

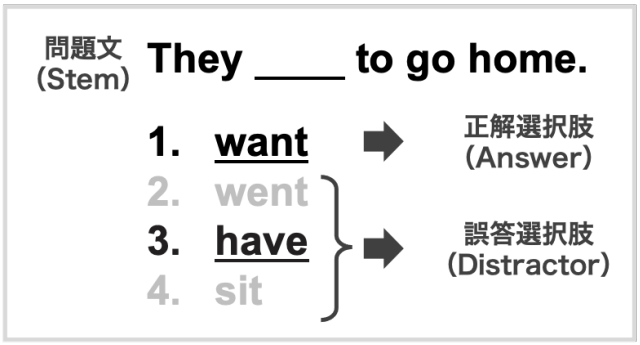

図 2　誤答選択肢が正解となってしまう例

成手法を提案する．具体的には，英文法の MCQ を取り上げ，文章の自然さを表すスコアを利用して誤答選択肢候補が正解とならないことを担保し，語彙習得段階を考慮したスコアを用いたランキングにより誤答選択肢を選択する．

誤答選択肢の生成は，大きく分けて 3 つの処理で構成する．まず，事前学習済みの大規模言語モデルを用いて誤答選択肢の候補集合を生成する．その候補集合から，問題文の空欄に当てはめたときに正解の選択肢候補とならないような妥当性を担保できた候補選択肢のみを絞り込み，抽出する．妥当性を担保できた誤答選択肢集合に対して，それぞれの誤答選択肢の単語分散表現に習得段階の情報を組み込んだ特徴など，複数の特徴を考慮することでスコアを算出・ランキングし，ランキングの上位 3 つの選択肢を誤答選択肢として採用する．

本稿で提案した英文法 MCQ の誤答選択肢自動生成手法の有用性を確かめるために，人手評価実験を実施した．

評価実験の結果,「選択肢の内容の適切さ」と「習得段階を考慮した難易度の適切さ」の評価項目において従来手法よりも高い評価結果が得られた．また，提案手法による選択肢は，従来手法によるそれよりも，人手によって作成された選択肢との違いが区別しにくいことが確認できた．

本論文での貢献は以下のとおりである．

（1）英文法の MCQ において，語彙習得段階を考慮しつつ，誤答選択肢が正解となる可能性を排除する枠組みを導入した誤答選択肢自動生成手法を提案した．

（2）自動生成した選択肢が，誤答選択肢として自然・適切か，語彙の習得段階を考慮できているか，人手で作成された選択肢と同等の品質であるかといった点で評価実験を行い，本手法の有用性について確かめた．

本論文の構成は次の通りである．2 節では，AQG，問題文生成，誤答選択肢生成のそれぞれについて関連研究を紹介する．3 節では，提案手法について，その構成要素である，誤答候補集合生成部，選択肢妥当性検証部，誤答候補集合選択部に分けて詳細に説明する．4 節では，提案手法を構築するために用いたデータセットについて述べ，5 節では，評価実験とその結果をまとめ，分析・考察する．6 節では，まとめと今後の課題について述べる．

2　関 連 研 究

2.1　AQG に関する研究

自動問題生成（Automatic Question Generation，AQG）の手法は，ルールベースによる生成，機械学習による生成の 2 種類に大別できる．

津森らは，理解度に適応した MCQ の生成をルールベースに基づく手法で提案している [6]．問題生成のもととなるデータを知識ベースに格納し，語彙情報をテンプレートに当てはめることによって問題生成を実現している．また，語彙の概念距離を考慮することで理解度に応じた適切な選択肢の組み合わせを探索している．

岩田らは，英語の MCQ の生成を機械学習に基づく手法で提案している [7]．問題文のもととなるパッセージを入力し，人手で作成された問題から統計的ルールを学習したモデルで穴埋め箇所を決めている．また，品詞の種類に応じてタイプを選定し，統計的な割合に従って辞書を用いて誤答選択肢を生成している．

提案手法は，事前学習済みの言語モデルを用いて誤答選択肢を生成するといった点で異なる．

2.2　問題文生成に関する研究

問題文生成は，データベースやコーパスから適切な問題文を抽出する手法と事前学習済みの言語モデルを用いて問題文を生成する手法に大別できる．

松森らは，問いたい語句が唯一の答えとなるような問題をコーパスから抽出する手法を提案している [9]．問題の空欄にマスクを適用し，マスクに問いたい語句が当てはまる推定確率が高いだけでは唯一性があるとは言えないと主張し，独自の Gap Score という指標を導入することにより，問いたい語句の唯一性を保証している．

鈴木らは，事前学習済みの言語モデルを用いて難易度を調節可能な読解問題自動生成手法を提案している [11]．具体的には，Decoder 型の事前学習済みモデルに読解問題文を入力し，その内容について問う問題を生成している．また，入力に項目反応理論を用いた難易度の指標を組み込むことで細やかな難易度別の問題生成を実現している．

これらは問題文生成を主な目的としているが，提案手法は誤答選択肢を生成することを目的としている点で異なる．

2.3　誤答選択肢生成に関する研究

誤答選択肢生成は，データベース・知識ベースから適切な選択肢を抽出する手法 [6] [7] と事前学習済みの言語モデルから生成する手法に大別できる．

湯浅らは，問題の空欄にマスクを適用し，マスクに入る単語の推定確率を考慮することで難易度調節可能な英語の多肢選択問題の生成を提案している [10]．Masked Language Model を問題集のデータを用いて追加学習することで，問題集で使用されるような誤答選択肢の特徴を反映させて生成している．問題集に応じた推定確率に着目し誤答選択肢を生成しているのに対し，提案手法では習得段階に着目して誤答選択肢を提案する点

で異なっている．

また，Shang らは，候補集合生成部（Candidate Set Generator，CSG）と誤答選択部（Distractor Selector，DS）で構成したシステムで誤答選択肢生成を提案している [12]．CSG では，事前学習済みモデルに問題文と正解選択肢を連結して入力し，正解選択肢の情報を加味した上で誤答選択肢を生成している．DS では，CSG で得られた候補集合をランキングするため，誤答選択肢がマスク適用文のマスク箇所に当てはまる推定確率，正解選択肢と誤答選択肢の非コサイン類似度，正解選択肢と誤答選択肢をマスク文のマスク箇所に当てはめた状態の文全体としての非コサイン類似度，正解選択肢と誤答選択肢の品詞の一致度といった特徴を Min-Max Normalization で重み付けした合計でスコアを算出している．誤答選択肢生成に関する研究では，現時点でこの研究で提案された手法が SOTA とされている．

本稿では，この手法に基づき，語彙習得段階の情報と誤答選択肢の妥当性を検証する仕組みを新たに導入することで，さらなる性能向上を目指す．

3 提 案 手 法

本稿で扱う誤答選択肢の生成は，問題文 S と正解選択肢 A が入力され，誤答選択肢 $D = \{d_1, d_2, d_3 ... d_n\}$ を出力する問題である．本稿では $n = 3$ と設定した．提案手法は，大きく分けて 3 つの処理で構成する．図 3 に提案手法の概要を示す．

3.1 候補集合生成部（Candidate Set Generator，CSG）

CSG では，事前学習済みモデルである BERT [13] を用いる．BERT に問題文 S と正解選択肢 A を入力したとき，[MASK] に当てはまるであろうと推論される誤答候補と推定確率を降順に得る．

$$CSG(S_{\otimes[MASK]}[SEP]A) \rightarrow d_i \tag{1}$$

ここで，$S_{\otimes[MASK]}$ は，S にマスクを適用した問題文，$S_{\otimes[MASK]}[SEP]A$ は，$S_{\otimes[MASK]}$ と A を $[SEP]$ で連結した文字列を表す．

ファインチューニングを行うことにより，以下の損失関数 $\mathcal{L}$ を最小化するパラメータ θ を設定することを目指す．

$$\mathcal{L} = -\log(p(D \mid S, A; \theta)) \tag{2}$$

3.2 誤答妥当性検証部（Distractor Validator，DV）

DV では，CSG で得られた誤答選択肢候補集合の各要素に対して，正解の選択肢候補とならないような誤答選択肢であるかを検証し，誤答としての妥当性が確認された要素のみに絞り込む．CSG で得られた誤答選択肢を問題文の穴埋め箇所に当てはめ，文章の自然さを表すスコアである擬似対数尤度スコア（pseudo-log-likelihood score，PLL）[14] を算出する．以下は PLL を算出する定義式である．

$$PLL(W) = \sum_{t=1}^{|W|} \log P_{MLM}(w_t | W_{\setminus t}; \theta) \tag{3}$$

ここで，w_t は，[MASK] 箇所のトークンであり，$W_{\setminus t}$ を以下のように定義する．

$$W_{\setminus t} = (w_1, ...w_{t-1}, [MASK], w_{t+1}, ...w_{|W|}) \tag{4}$$

$P_{MLM}(w_t|W_t; \theta)$ は，W_t が与えられた際のトークン w_t を Masked Language Model で予測した確率を表す．各単語を [MASK] し，[MASK] 箇所を予測したときの推定確率の和をスコアとする．CSG で得られた各誤答選択肢に対してスコアを算出し，正解選択肢が穴埋め箇所に当てはめたときのスコアとの差が閾値以内であれば，その誤答選択肢が正解になりうる可能性が高いと判断し，誤答選択肢候補からは除外する．

3.3 誤答選択部（Distractor Selector，DS）

DS では，DV で妥当性を確認した誤答選択肢をランキングし，上位の誤答候補を誤答選択肢として採用する．ランキングのためのスコアは，従来手法で採用されていた 4 つのスコアに加え，語彙習得段階スコアを新たに追加することで計算する．

（1）信頼度スコア s_0

誤答選択肢 d_i が穴埋め箇所に入るであろうと推定される確率をスコア s_0 とする．

$$s_0 = p(d_i | S, A; \theta) \tag{5}$$

（2）単語埋め込みスコア s_1

正解選択肢 A と誤答選択肢 d_i のコサイン非類似度をスコア s_1 とする．

$$s_1 = 1 - \cos(\overrightarrow{A}, \overrightarrow{d_i}) \tag{6}$$

（3）文脈埋め込みスコア s_2

正解選択肢 A，誤答選択肢 d_i をそれぞれ穴埋め箇所に入れた文章のコサイン非類似度をスコア s_2 とする．

$$s_2 = 1 - \cos(\overrightarrow{S}_{\otimes A}, \overrightarrow{S}_{\otimes d_i}) \tag{7}$$

（4）品詞一致スコア s_3

正解選択肢 A と誤答選択肢 d_i の品詞が一致している場合に 1，不一致の場合 0 をスコア s_3 とする．

$$s_3 = \begin{cases} 1 & POS_A = POS_{d_i} \\ 0 & POS_A \neq POS_{d_i} \end{cases} \tag{8}$$

ここで，POS_A, POS_{d_i} は，それぞれ A, d_i の品詞を表す．

（5）語彙習得段階スコア s_4

正解選択肢の習得段階が誤答選択肢よりも上位の場合は，誤答選択肢の習得段階を正解選択肢の習得段階で除算した値，正解選択肢の習得段階が誤答選択肢よりも下位の場合は，0 とするスコアを s_4 とする．

$$s_4 = \begin{cases} \frac{Level_{d_i}}{Level_A} & Level_A - Level_{d_i} \geqq 0, \\ 0 & Level_A - Level_{d_i} < 0 \end{cases} \tag{9}$$

ここで，$Level_A$, $Level_{d_i}$ は，それぞれ A, d_i の習得段階に対応した正の実数を表す．

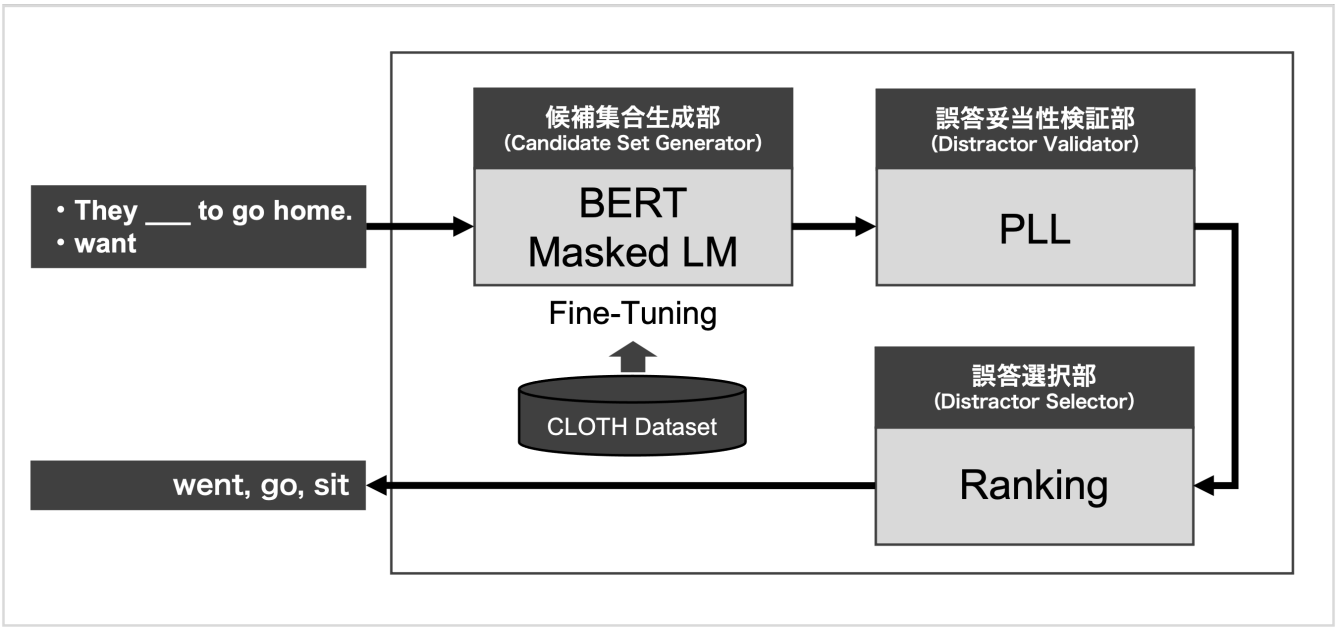

図 3　提案手法の概要図

最終的なスコア $score(d_i)$ は，各スコアを Min-Max Normalization した重み付け和として算出する．

$$score(d_i) = \sum_{i=0}^{4} w_i \cdot \text{MinMax-Norm}(s_i) \tag{10}$$

ここで，w_i は s_i の重みを表す．本稿では，$w_0 = 0.5, w_1 = 0.15, w_2 = 0.1, w_3 = 0.05, w_4 = 0.2$ と設定した．そして，$score(d_i)$ が上位 3 件に該当する d_i を最終的な誤答選択肢として採用した．

4　データセット

4.1　使用するデータセット

4.1.1　CEFR-J Wordlist

習得段階参照用データセットとして，CEFR-J Wordllist [15] を用いる．CEFR（Common European Framework of Reference for Language）は外国語の学習・教授・評価のためのヨーロッパ言語共通参照枠とされており，言語の枠を越えて，外国語の運用能力を同一の基準で測ることができる国際標準である．CEFR を日本の外国語教育に照らし合わせて整理したものが CEFR-J である．CEFR では A1〜C2 の 6 段階を基本としているが，CEFR-J Wordlist では日本の教育機関で習得する単語とされている A1〜B2 の 4 段階が公開されている．本稿では，A1〜B2 の段階をそれぞれ 1〜4 の正の整数値に置き換える．習得段階を各単語の分散表現の最後の要素に埋め込む．

4.1.2　CLOTH

候補集合生成部の学習用データセットとして，CLOTH を用いる [16]．CLOTH は中国内の中学校・高等学校の教師が学生の入試対策として作成したデータセットである．穴埋め文，選択肢，正解選択肢の 3 つ組のデータが，訓練データに 5513 件，評価データに 805 件，テストデータに 813 件含まれる．

5　評 価 実 験

5.1　実 験 設 定

本稿で提案した英文法 MCQ の誤答選択肢自動生成手法の有用性を確かめるために，人手評価実験を実施した．実験では，生成した誤答選択肢の内容の適切さ，習得段階を考慮した語彙の難易度の適切さ，人手によって作成された選択肢および従来手法による選択肢との品質の違いといった 3 つの観点から評価する．表 1 に示すように，18 問の英文法の MCQ のうち，中学卒業レベルと高校卒業レベルの MCQ をそれぞれ 9 問ずつ提示する．中学卒業レベル，高校卒業レベルのそれぞれ 9 問は，3 問が人手によって作成された選択肢，3 問が従来手法で生成された選択肢，3 問が提案手法で生成された選択肢である．評価者には，穴埋めの問題文とその習得段階（中学卒業レベル・高校卒業レベル），正解選択肢，誤答選択肢を提示する．穴埋め問題が，人手，従来手法，提案手法のいずれの方法で作成されたものであるかは知らせない．評価項目は以下の 2 点である．

（1）選択肢の内容の適切さ（5 段階評価，1 に近いほど不適切，5 に近いほど適切）
正解選択肢のほかに，正解となってしまう選択肢は存在していないか，解答者を適度に考えさせられる選択肢であるかどうかといった観点で評価する

（2）習得段階を考慮した難易度の適切さ（5 段階評価，1 に近いほど不適切・5 に近いほど適切）
正解選択肢よりも極端に難易度の高い・低い選択肢が含まれていないかといった観点で評価する

また，提案手法による誤答選択肢生成の有用性を確かめるため，提示された問題の選択肢が「人手によって作成された選択肢」か「自動生成による選択肢」のどちらであるかを予想してもらい，回答者の予想の平均正解率を確かめる．

平均正解率が 50 %に近いほど，人手による選択肢と提案手法による選択肢とで大きな違いがないことが示唆される．

なお，誤答選択肢生成に用いた問題文と正解選択肢について，中学卒業レベルは神奈川県公立高校 2011 年度〜2022 年度入試問題，高校卒業レベルは 2014 年度〜2020 年度のセンター試験問題からランダムに抽出した．また，評価実験参加者は 9 名で，いずれも塾講師経験済みの大学生である．

表 1　評価実験に用いた問題の内訳

	中学卒業レベル	高校卒業レベル
人手	3 問	3 問
従来手法	3 問	3 問
提案手法	3 問	3 問

5.2 実験結果

図 4 に評価項目「選択肢の内容の適切さ」の結果を示す．人手による選択肢は平均値 4.26，従来手法による選択肢は平均値 3.54，提案手法による選択肢は平均値 4.02 であり，人手による選択肢が最も高い値であった．また，提案手法による選択肢は従来手法による選択肢を上回った．

図 5 に評価項目「習得段階を考慮した難易度の適切さ」の結果を示す．人手による選択肢は平均値 4.22，従来手法による選択肢は平均値 3.52，提案手法による選択肢は平均値 4.06 であり，人手による選択肢が最も高い値であった．また，提案手法による選択肢は従来手法による選択肢を上回った．

表 2 に評価者による選択肢の人手・自動生成予想の平均正解率を示す．「従来手法＋人手」の正解率は 0.601，「提案手法＋人手」の正解率は 0.564 であり，「提案手法＋人手」の方がより人手によって作成された選択肢と提案手法によって生成された選択肢の区別ができていないことがわかる．また，対応のある t 検定を実施し，「従来手法＋人手」と「提案手法＋人手」のそれぞれの平均正解率には有意な差があることを確認できた（$t(107) = 1.98, p = .04$）．

表 2　評価者による選択肢の人手・自動生成予想の平均正解率

	従来手法＋人手	提案手法＋人手
平均正解率	0.601	0.564

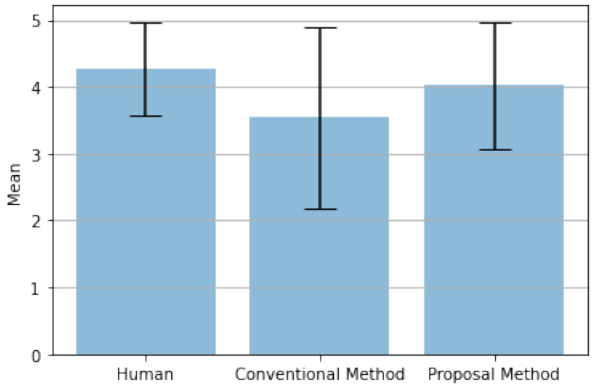

図 4　評価項目「選択肢の内容の適切さ」の結果

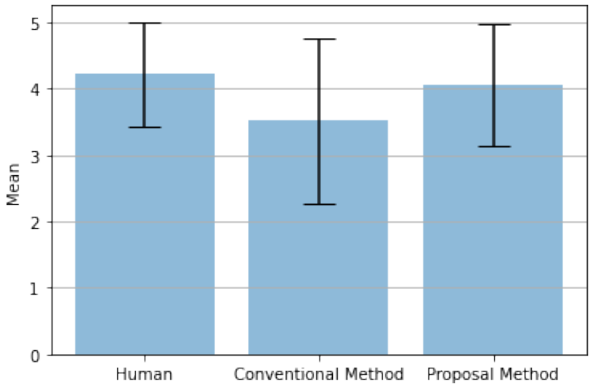

図 5　評価項目「習得段階を考慮した難易度の適切さ」の結果

5.3 考　　察

評価項目「選択肢の内容の適切さ」では，人手によって作成された選択肢が最も高い評価となった．人手によって作成された選択肢の評価が高くなったこととして，選択肢の品詞や使い方が同じであったことが考えられる．また，提案手法によって生成された選択肢の評価は従来手法によって生成された選択肢の評価を上回ったが，従来手法によって生成された選択肢に，誤答であるが正解選択肢となってしまうものが含まれており，従来手法の評価が低くなった原因であると考えられる．

評価項目「習得段階を考慮した難易度の適切さ」においても，人手によって作成された選択肢が最も高い評価となった．こちらは，評価項目「選択肢の内容の適切さ」で評価が低かった選択肢で同じく低い評価である場合が多く確認できた．これは，評価者が適切でない選択肢は難易度も適切でないと判断したことが考えられる．

評価者による選択肢の人手・自動生成予想の平均正解率は，「提案手法＋人手」の方が 50 %に近いことを確認できた．よって，人手によって作成された選択肢と提案手法によって生成した選択肢と比較しても選択肢の品質に大きな違いがないことが考えられる．

6 ま と め

本稿では，多肢選択問題の作成者が問題作成過程において誤答選択肢の妥当性を人手で確認する作業負荷を軽減するため，語彙習得段階を考慮しつつ，誤答選択肢が正解とならないような枠組みを導入した誤答選択肢の生成手法を提案した．事前学習済みモデルを用いて誤答選択肢候補を生成し，文章の自然さを表すスコアを利用して誤答選択肢候補が正解とならないと判断された候補のみに対して，語彙習得段階を考慮したスコアを用いたランキングにより上位の誤答選択肢を選択した．評価実験の結果，「選択肢の内容の適切さ」と「習得段階を考慮した難易度の適切さ」の評価項目において従来手法よりも高い評価結果が得られた．評価者による選択肢の人手・自動生成予想の平均正解率から，提案手法による選択肢は，従来手法によるそれと比べて，人手によって作成された選択肢との違いが区別しにくいことが確認できた．また，正解選択肢を含む文章と誤答選

択肢を含む文章の自然さを表すスコアの差を用いて妥当性を検証するだけでは，誤答選択肢が正解となる可能性を完全に排除できたとはいえないため，人手によって作成された選択肢を分析し，より高い妥当性をもって選択肢を生成する枠組みを考案する必要がある．

謝　　辞

本研究の一部は科研費 18K11557 の助成を受けたものである．ここに記して感謝の意を表します．

文　　献

[1] 大内裕和. “教員の過剰労働の現状と今後の課題”. 日本労働研究雑誌, Vol. 2021, No. 730, pp. 4–13, 2021.

[2] Dhawaleswar Rao CH, Sujan Kumar Saha. “Automatic Multiple Choice Question Generation From Text : A Survey”. IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 13, NO. 1, JANUARY-MARCH 2020

[3] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler and Salam Al-Emari. “A Systematic Review of Automatic Question Generation for Educational Purposes.”. Int J Artif Intell Educ 30, pp. 121–204, 2020.

[4] Chao-Yi Lu, Sin-En Lu. “A survey of Approaches to Automatic Question Generation : from 2019 to Early 2021”. The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021) Taoyuan, Taiwan, October 15-16, 2021

[5] Srivastava, Goodman. “Question Generation for Adaptive Education”. ACL-IJCNLP 2021.

[6] 津森伸一, 海尻賢二. “理解状況に適応した多肢選択式問題の自動生成に関する構想.”. 教育システム情報学会研究報告 21.4, pp. 3–8, 2006.

[7] 岩田具治, 後藤拓也, 小尻智子, 渡邉豊英, 山田武士. “機械学習に基づく英語穴埋め問題の自動生成”. NTT 技術ジャーナル, 9 巻, 10 号, pp. 16–19, 2011

[8] Sophia Chan, Swapna Somasundaran, Debanjan Ghosh, and Mengxuan Zhao. “AGReE: A system for generating Automated Grammar Reading Exercises.”. In Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing System Demonstrations, pp. 169–177, 2022.

[9] 松森 匠哉, 奥岡 耕平, 柴田 遼一, 井上 南, 吉野 哲平, 福地 庸介, 岩沢 透, 今井 倫太. “マスク言語モデルを利用した Open Cloze 問題の自動生成”. 人工知能学会全国大会論文集, JSAI2022 巻, No. 36, pp. 3N4GS1004–3N4GS1004, 2022.

[10] 湯浅成章, Andrew Vargo, 黄瀬浩一. “個人に適した英語多肢選択問題の自動生成方法の提案”. 情報処理学会研究報告（Web）, Vol. 2022, No. 5, pp. 1–8, 2022.

[11] 鈴木彩香, 宇都雅輝. “Transformer モデルを用いた難易度調節機能付き読解問題自動生成手法”. 日本行動計量学会大会録集, 50 巻, No. 50, pp.126–129, 2022.

[12] Shang-Hsuan Chiang, Ssu-Cheng Wang, Yao-Chung Fan. “CDGP: Automatic Cloze Distractor Generation based on Pre-trained Language Model”. Association for Computational Linguistics, EMNLP 2022, pp. 5835–5840, 2022.

[13] Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.”. arXiv preprint arXiv:1810.04805, 2018.

[14] Julian Salazar and Davis Liang and Toan Q. Nguyen and Katrin Kirchhoff. “Masked language model scoring”. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2699-2712, 2019.

[15] Negishi, Masashi, Tomoko Takada, and Yukio Tono. “A progress report on the development of the CEFR-J.”. Exploring language frameworks: Proceedings of the ALTE Krakow Conference, 2013.

[16] Qizhe Xie, Guokun Lai, Zihang Dai, Eduard Hovy. “Large-scale Cloze Test Dataset Created by Teachers”. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2344–2356, 2018.