

個人に適した英語多肢選択問題の自動生成方法の提案

湯浅 成章^{1,a)} Andrew Vargo^{1,b)} 黄瀬 浩一^{1,c)}

概要：集中力やモチベーションは学習効率に大きく影響する。集中力やモチベーションを維持するためにはタスクの難易度を適切に調整することが重要である。従来の研究では問題の難易度を2, 3クラス程度の分類問題として扱っているため、適切な難易度の問題が存在しない学習者が現れる可能性がある。本研究では英語多肢選択問題を取り上げ、学習者個人に合わせて難易度を細やかに調整するシステムを提案する。問題の生成及び難易度の調整には Masked Language Model を用いる。実験参加者 15 名を対象として問題集での学習及び提案システムでの学習を実施し、学習の1日後、3日後にテストを行った結果、問題集での学習と比べて提案システムでの学習のほうが1日後で8パーセント、3日後で7パーセント正解率が向上し、その差が統計的に有意であることを確認した。また、提案システムが学習中の正答率に合わせて個別に難易度の調整を行っていることを確認した。

キーワード：学習増強, 知能増強, 英語学習, Masked Language Model

1. はじめに

英語など何かを学習する時に、知識の定着や実力の把握のために問題演習を行うのは一般的な学習方法の一つである。この学習方法の問題点として、既存の問題集などでは問題数に限りがあり、その中でも難易度などの要素を考えれば個人に適したものはさらに数が限られてしまうということがあげられる。簡単すぎる問題や難しすぎる問題ではやる気をなくしてしまい、パフォーマンスが低下してしまう [1]。それを避けるために、個人に適した問題を探すことに時間を浪費してしまうということも考えられる。これらの問題を解決する一つの方法は、個人に適した問題を自動生成することであろう。ここでは、個人に適した問題を、適切な正答率となる難易度の問題と定義する。

本研究では穴埋め形式の文法を問う英語多選択肢問題を扱う。この形式の問題は TOEIC などでも用いられており、記述式の問題と比べて解答にかかる時間が短く効率よく学習を進めることができる [2]。多選択肢問題の難易度は選択肢に大きく依存している [3]。そこで、本研究では学習者の知識や能力によって選択肢自体を変化させることで個人に適した難易度の問題を生成するシステムの作成を目指す。

個人に適した問題を生成する、あるいは生成した問題を

個人向けに変化させる研究はほとんど行われていない [4]。知識を扱う問題に関しては、自動生成した問題の難易度を制御する研究は存在する [5], [6]。ただし、これらの研究では難易度を2, 3クラス程度の分類問題として扱っているため、この方法では適した難易度の問題が存在しない学習者が出現する恐れがある。文法問題における誤答選択肢の生成を行う研究はこれまでも行われている [7], [8]。しかし、個人に向けて難易度の調整を行った研究ではない。

本研究では、問題の生成及び難易度の細やかな調整を行うために Masked Language Model (MLM) を利用する。MLM とは文章中の隠された単語を推定するモデルである。穴埋め問題の穴部分に入る単語を推定することで問題の生成を行う。MLM を用いることで、問題文の文脈を考慮した誤答選択肢の生成が可能となる。問題の難易度の調整には穴部分に入る単語の推定確率を利用する。推定確率を利用することで、問題文の文脈に応じた難易度の制御が細かな粒度で可能になると考えられる。本研究では学習者の正答率に合わせて選択肢を変化させることで個人に適した問題を生成できるシステムを作成することを目指す。15 名を対象とした実験の結果、問題集を用いた演習と比べて、提案システムを用いた演習が学力向上に有効であることがわかった。さらに、提案システムにおける難易度調整手法が実験参加者それぞれに対して個別の難易度調整をしていることがわかった。なお、本研究は大阪府立大学大学院工学研究科倫理委員会の承認を得ていることを付記しておく。

¹ 現在, 大阪府立大学大学院工学研究科

^{a)} yuasa@m.cs.osakafu-u.ac.jp

^{b)} awv@m.cs.osakafu-u.ac.jp

^{c)} kise@cs.osakafu-u.ac.jp

2. 関連研究

本研究の目的は、多肢選択問題の選択肢を自動生成することで、個人に適した学習を実現することである。そのためには、まず個人に適した学習とは何かを、問題の難易度の観点から考察した研究について述べる。その上で、問題の難易度を調査した研究や、その制御を試みた研究について述べる。

2.1 適切な正解率を調査した研究

Wilson らの研究 [9] は機械学習モデル及び知覚学習を模したモデルを用いてどのような正答率の時、最も学習効率が高くなるのかについて検証した。その結果、タスクや前提条件によるが、おおよそ 70 から 85 パーセントの正答率の時学習効果が最も高くなると述べている。さらに、この結果は人間にも適応できる可能性が高いと述べている。

この研究に基づき、適切な正解率となる難易度の問題をおおよそ 70 から 85 パーセントの正答率となる難易度の問題と定義する。

2.2 問題文の構成要素が難易度に与える影響を調査した研究

Susanti らの研究 [3] では、語彙を問う問題でその構成要素が難易度に与える影響を調査した。語彙を問う問題とは文中に存在する単語に対し、最も意味が近いものを選択するという形式の問題である。この研究では問題を文章自体の難易度、正答と誤答選択肢の類似度、誤答選択肢となる単語自体の難易度の 3 つに分解し、それぞれに対して簡単、難しいの 2 種類のものを作成し問題の正答率に与える影響を検証した。その結果、正答と誤答選択肢の類似度、誤答選択肢となる単語自体の難易度の 2 つが正答率に大きな影響を与えていることを示した。

文法問題においても選択肢の影響は大きいと考えられる。本研究では文法問題の誤答選択肢を調整することで個人に適した問題の生成を行う。

2.3 誤答選択肢の生成を行った研究

英語の文法問題における誤答選択肢を生成する研究として、岩田らの研究 [7] や Chen らの研究 [8] がある。これらの研究では正答となる単語から品詞などの情報を抽出し、類似する単語や正答となる単語の活用を変化させた単語を誤答選択肢として用いている。この種の手法は英語に限らず中国語などの文法問題においても現在も用いられている [10] [11]。

この種の方法では問題文の文脈の情報が考慮できていなかったり、生成される誤答選択肢がワンパターンになってしまっていたりするなどの問題がある可能性がある。本研究では文脈などの情報も含めてモデルの学習を行う。ま

た、誤答選択肢の難易度をコントロールできるシステムを作成する。

2.4 生成した選択肢の難易度を制御した研究

知識を問う問題における誤答選択肢の生成及び難易度の制御を行う研究として、Seyler らの研究 [12] や Faizan らの研究 [6] がある。これら研究では正答となる言葉と類似している言葉やオントロジー上で近い概念を表す言葉を誤答選択肢として使用し、その距離などに基づいて 2, 3 クラス程度の難易度に分類している。

知識を問う問題ではこの手法は有効であるが、文法問題を扱う場合では必ずしも有効であるとは限らない [5]。例えば、センター試験では“at most”という正答に対して“at least”という誤答選択肢が用いられていた問題がある。ここで most と least について Word2Vec [13] を用いてコサイン類似度を計算すると 0.35 となり類似しているともしていないとも言えない結果になった。このように、文法問題では類似度を誤答選択肢の指標として用いることは難しいと考えられる。さらに、正答が複数単語からなる場合、類似度などの情報を取得すること自体が難しくなってしまう。また、難易度が 2 クラス程度しかなく適した難易度がない学習者が現れる可能性がある。

本研究では文脈を考慮したうえで細かな粒度で難易度の調整が可能なシステムを構築する。

3. 提案システム

本章では、個人に適した問題を自動生成するシステムの詳細を説明する。本システムでは、問題を生成する、学習者が問題に解答する、機械学習モデルが難易度を調整するという操作を繰り返すことで学習者にとって適切な難易度の問題を生成できるようにする。

まず、本システムで用いられる 2 つの機械学習モデルについて説明する。1 つ目は大規模なデータを用いて事前学習された MLM である。ここではこのモデルのことを候補モデルと呼ぶ。2 つ目は大規模なデータで事前学習された MLM に対して、問題集のデータを用いて追加学習が行われたモデルである。ここではこのモデルのことを選定モデルと呼ぶ。この追加学習を行うことで、問題集で使用されるような誤答選択肢の特徴を学習することを期待している。

次に、提案システムにおける難易度の調整方法を説明する。難易度の調整手法を図 1 に示す。まず、問題文を候補モデルに入力し、問題文中の空白に入る単語を推定する。推定された確率が高いほどその単語は問題文によく合う単語である、すなわち、正答と区別しづらい難しい誤答選択肢になると考えられる。推定された確率が高い順に単語を並べ、上限と下限を設定し、その範囲内の単語を誤答選択肢の候補とする。

次に、選定モデルに問題文を入力する。その出力のうち、

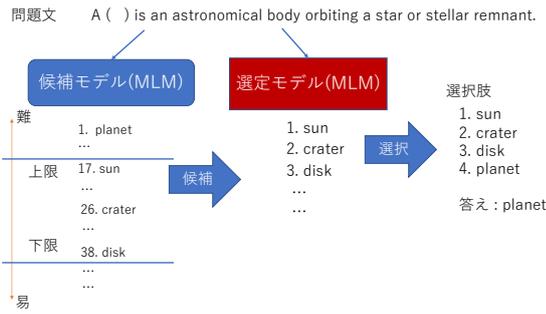


図 1 選択肢の生成と難易度調整方法

誤答選択肢の候補となっている単語のみを抜き出し推定確率の高いものから3つを誤答選択肢とする。

ここで、候補モデルの出力のみを用いないのは、前述したように、選定モデルが問題集の特徴を学習することを期待しているためである。例えば、センター試験では“successful”という単語が正答となる問題で、誤答選択肢として“success”, “succeeded”, “successive”を使用している。候補モデルのみでは品詞が正答と異なる単語は推定確率が低くなってしまいこれらのような誤答選択肢を生成することができない。すなわち、品詞に関する知識がなければ解けない問題がほとんど生成できなくなってしまう。

一方、選定モデルのみを用いないのは、追加学習に用いることのできるデータ数が少なく単語の推定結果が安定しないためである。例えば、問題文によっては“inishing”のように存在しない単語が高い推定確率で出力されてしまうことがある。これは“in”と“ishing”が組み合わせられたものである。この“ishing”は“astounding”などの一部として扱われているため、前の単語に連続して出力される。これは問題集の誤答選択肢で文法的に誤ったものを学習したため、このようなありえない組み合わせの推定確率が高くなると考えられる。

本システムでは、誤答選択肢を取り出す際の上限と下限を変化させることにより、問題の難易度の調整する。具体的には、上限と下限をより上位の方向に変化させることにより難化、下位の方向に変化させることにより易化を行う。

4. 実験

本節では、提案システムが学習に有効であるか否かを検証するために行った実験について述べる。提案システムを用いた演習の前後でテストを行い、その差を取ることで提案システムの学習効果を検証した。また、比較対象として、同様の条件で問題集の学習効果も検証した。実験には日本人大学生、大学院生 15 名が参加した。このうち男性は 14 名であり女性は 1 名であった。平均年齢 22.1 歳、年齢の標準偏差は 1.58 であった。実験全体では 12 時間かかると想定し、謝礼は 12000 円分の金券とした。以下に詳細を述べる。

4.1 実験に用いたモデル

誤答選択肢を生成する MLM 及び候補モデルとなる MLM には RoBERTa [14] を用いた。RoBERTa は様々な自然言語処理タスクにおいて好成績を残している。その特徴として種々のタスクへのファインチューニングを行う前の事前学習が全て同じであるということが挙げられる。これは、事前学習により自然言語処理全般に有用な特徴が得られていることを意味する。誤答選択肢を生成するために必要な文脈などの情報を抽出できていると考え、学習済みの RoBERTa(HuggingFace [15] により作成された学習済みのモデル)を用いた。

4.2 選定モデル学習用データセット

追加学習するためのデータセットには 1990 年から 2020 年までのセンター試験本試および追試、英語の学習塾で用いられている教材、TOEIC スコアアップの教科書 *1, Nisshy の英語問題集 10515 *2 から得られた合計 4920 問を用いた。選定モデルへの入力の問題文の空白部分をマスクに置き換えたもの、正解ラベルは空白部分に誤答選択肢を入れたものを用いた。

4.3 実験用の問題セット

実験に用いた演習用教材は、英検 1 級の問題から得られた 225 問、英検準 1 級の問題から得られた 75 問、TOEIC 対策の文法問題集から得られた 200 問である。これらの問題はかなり難しい設定となっている。英検 1 級から 80 問、英検準 1 級から 20 問、TOEIC 対策の文法問題集から 100 問をランダムに抜き出し、200 問の問題セットを作成した。この 200 問のセットを、提案システム用と問題集用の 2 つ用意した。ランダムに問題を選択することで教材間の難易度差を軽減した。なお、この 2 つのセットで問題の重複はない。また、すべての問題は 4 択問題であり、正答の選択肢 1 つ、誤答選択肢 3 つを持つ。

プレテスト及びポストテストにおける誤答選択肢の作成方法を図 2 に示す。これらの選択肢には、問題集の誤答選択肢を 1 つ、提案システムから得られた選択肢を 2 つ、正答を 1 つの計 4 つを用いた。このとき、プレテスト、ポストテスト 1、ポストテスト 2 の誤答選択肢に重複がないようにした。問題集の誤答選択肢と生成された誤答選択肢を混ぜることで、問題集での学習した時のテストと提案システムで学習したときのテストの誤答選択肢による有利不利を軽減した。

4.4 実験手順

続いて、実験手順について述べる。提案システムの有効性を検証するために、提案システムでの演習環境と問題集

*1 <https://zitanstudy.com/>

*2 <http://www7a.biglobe.ne.jp/nisshy/>

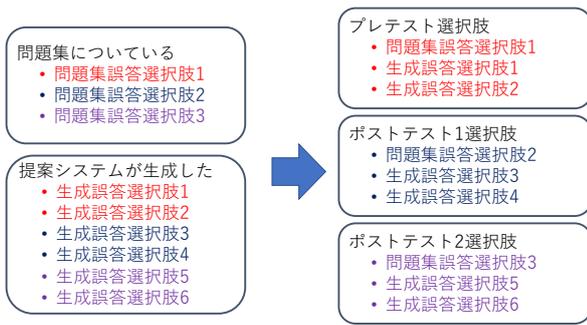


図 2 プレテスト及びポストテストにおける誤答選択肢の作成方法

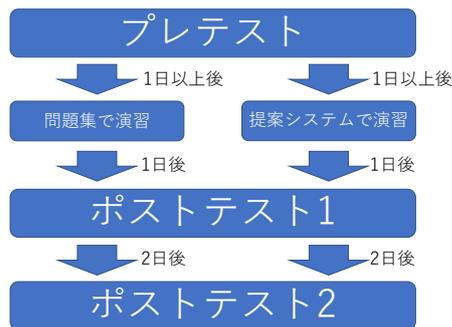


図 3 実験手順の概要

での演習環境を用意した。実験手順の概要を図 3 に示す。実験参加者はまず制限時間 50 分で 200 問のプレテストを受けた。その 1 日以上後、参加者は 160 分間演習環境で演習を行った。プレテストから 1 日以上空けるのは、プレテストで見覚えのある単語を選択することで正解できることを減らすためである。160 分間の演習では参加者が演習をさぼるのを防ぐために最低でも 2 周、すなわち 400 問解くように指示した。実際の演習に条件を近づけるため、2 周解いた後は 3 週目に入り、160 分の演習時間が終わるまで演習を行った。それぞれの周で問題の出現順序及び選択肢の順序をランダムに並べ替えた。ここで、同じ周で同じ問題が出題されることはない。問題への解答は図 4 に示す画面で行われた。画面には現在の進捗、問題文及び選択肢を表示した。すべての問題に解答していない場合は次に進めない処理になっている。演習中は問題の解答後、図 5 に示すように解答の正誤、正答となる選択肢、学習者が選択した選択肢を表示した。演習のちょうど 1 日後に 1 回目のポストテストを、ちょうど 3 日後に 2 回目のポストテストを行った。すべてのテストが終了したのちに、学習方法に関するアンケートを行った。

ランダムに選ばれた半数の実験参加者には提案システム、問題集の順で演習するよう指示し、残りの参加者には逆の順で演習するよう指示した。プレテスト、演習、ポストテスト 1、ポストテスト 2 の一部の日程がかぶることは許可したが、実験参加者の負担を考え同じ日に演習を 2 回することは禁止した。



図 4 問題の解答画面

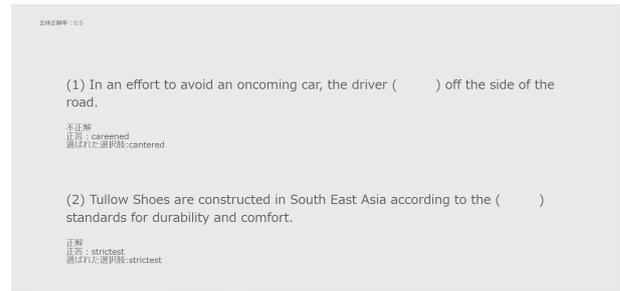


図 5 学習中の復習画面

4.5 誤答選択肢の生成に関する条件

提案システムでは難易度を適切に保つために正答率の履歴に基づいて候補モデルにおける上限と下限を調整する。履歴を用いるのは、直前に解答した問題のみを用いると問題文の難易度に影響が大きくなり、正答率がぶれやすくなると考えたからである。

演習を続けてある上限と下限に到達したときの正答率の履歴の平均と、その上限と下限で問題を解き続けた時の正答率の平均には若干の乖離があると考えられる。例えば、正答率が低い状態で演習を続け、正答率が、適切な正答率の下限である、0.7 になるような上限と下限に到達したとする。この時、正答率の履歴で平均を取ると 0.7 よりも低くなってしまふ。上限と下限を変化させない範囲と適切な正答率の定義が全く同じの場合、適切な上限と下限に到達しても難易度の調整がまだ行われるため、上限と下限を変化させない範囲が適切な正答率の定義より広くなるよう設定する。

本実験では以下のルールで上限と下限を調整した。なお、学習開始時の上限と下限はそれぞれ 10、20 とした。また、上限と下限の調整は 10 問解答するごとに行われた。提案システムで生成したすべての選択肢の単語数は、正答となる選択肢と同じ単語数で生成した。

- これまでの正答率が 0.9 以上の時、上限と下限をそれぞれ 10 ずつ推定確率が高い方向に変化させる。
- これまでの正答率が 0.6 以下の時、上限と下限をそれぞれ 10 ずつ推定確率が低い方向に変化させる。
- これまでの正答率が上記の範囲にない場合は上限と下限を変化させない。

4.6 実験結果

本節では提案システムを評価するために行った実験の結果について述べる。プレテスト及びポストテストの結果から提案システムを用いて演習した際の学習効果に与える影響を評価する。また、提案システムを用いて演習しているときの正答率から、提案システムの難易度調整手法が正答率に与えた影響を評価する。さらに、実験終了後のアンケートから提案システムが実験参加者の意欲やモチベーションに与えた影響について評価する。

4.6.1 提案システムを用いて学習した際の学習効果に与える影響

まず、実験参加者全体の正答率を集計した結果を表 1 に示す。問題集での学習から 1 日後、3 日後のそれぞれについて、ポストテストの正答率とプレテストの正答率の差を求めて正答率の上昇幅を求めた。さらに問題集で学習した場合と提案システムで学習した場合の 2 つのグループにおいて、帰無仮説を「2 群間の正答率の上昇幅に差はない」とし、対立仮説を「2 群間の正答率の上昇幅に差はある」として対応のあるデータでの 2 標本 t 検定を行った。その結果、学習から 1 日後、3 日後の両方とも提案システムのほうが有意に正答率の上昇幅が大きかったことがわかった。

次に、実験参加者それぞれの正答率の上昇幅について述べる。図 6 に実験参加者それぞれの正答率の上昇幅を示す。横軸は実験参加者を表しており、縦軸は正答率の上昇幅の差を表している。この図より、提案システムが学習に有効だったものもあれば提案システムが逆効果だったものもいることがわかった。

この原因を、実験の終了後に行ったアンケート結果を用いて考察する。アンケートで尋ねた項目の 1 つに「選択肢が変わることについてどのような印象を覚えましたか?」というものがある。この質問への回答は自由記述であった。このうち学習方法に触れていたものを一部抜粋し表 2 に示す。この表より、提案システムのほうが 1 日後に学習効果が高かった p02, p09 は、正答となる選択肢に注目して演習していたと考えられる。一方、問題集のほうが 1 日後に学習効果が高かった p03, p11 は、正答以外の選択肢と比較して演習する傾向があることがわかった。本実験で行ったテストは暗記で解答できたため、この違いが提案システムと問題集の学習効果に影響を与えたと考えられる。また、p09 が回答しているように、提案システムが生成する選択肢を変更したことで間違えた選択肢の記憶が残りにくく、学習に良い影響を与えたと考えられる。この影響を排除しつつ提案システムを検証する場合は暗記で対応できないようなテストをデザインする必要がある。

4.6.2 提案システムの難易度調整手法が正答率に与えた影響

提案システムでは、実験参加者の正答率に従って、誤答選択肢の順位を変化させる。そのため、上限の推移は、演

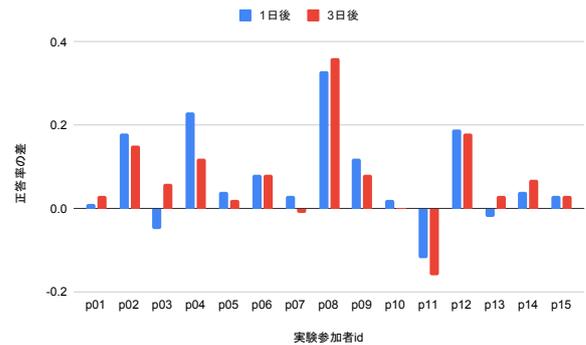


図 6 各実験参加者の正答率の上昇幅の差。提案システムで学習した時の正答率の上昇幅から問題集で学習した時の正答率の上昇幅を引いたものであり、グラフが上に伸びているほど提案システムで学習したほうが正答率の上昇幅が大きかったことを示している。

習の進展を表すものとなる。図 7 に推移を示す。ただし、ネットワークトラブルにより学習中のデータが正しく記録できなかった p09, p12 は含まれていない。本実験では上限と下限の差が 10 で固定されており、上限がわかれば下限がわかるため上限のみを示している。横軸は実験参加者が解いた問題数であり、縦軸は上限である。この図からわかるように、参加者全員に対して上限が高い部分では正答率が低いため次に解く問題の上限が下がっている。上限が下がってくると正答率が上がり、上限の下降が止まる。その後 p05 のようにさらに正答率が上がれば上限が上昇する。また、p08 のように実験中では上限の下降が止まらなかったものも存在する。この図より、提案システムの難易度調整手法によって実験参加者それぞれに対して個別の難易度調整がなされていることがわかった。

上限が変化しているときは学習者に適した難易度になっておらず、上限の変化が止まった時、学習者に適した難易度となっている。ここでは、この変化が止まった時の上限を適切な上限と呼ぶことにする。前述したように適切な上限は実験参加者によって異なる。これは英語能力の違いによると考えられる。そこで、適切な上限と TOEIC Listening & Reading^{*3} スコアの関係について考察する。TOEIC スコアは実験終了後のアンケートにて自由記述で尋ねた。記述のなかった p02, p06, p15 のデータはここでは省かれている。図 8 に TOEIC スコアと適切な上限の関係を示す。ただし、上限の変化が止まらなかった実験参加者は学習終了時の上限を適切な上限として用いた。図 8 より、適切な上限と TOEIC スコアには強い負の相関 (相関係数-0.747) があることがわかる。これは、TOEIC スコアを用いることで、提案システムで学習する際に最初から適切な難易度の設定で学習を始めることができる可能性があることを意味する。ただし、1 周 200 問のため、常に上限が変化する

*3 <https://www.iibc-global.org/toEIC/test/lr/about.html>

表 1 正答率の上昇幅とその差

	上昇幅平均 (問題集)	上昇幅平均 (提案システム)	平均の差	p 値	有意差
1 日後	0.49	0.57	0.08	0.028	有意水準 0.05 で有
3 日後	0.51	0.58	0.07	0.031	有意水準 0.05 で有

表 2 誤答選択肢が変わったことに対するアンケート結果

参加者 id	1 日後の正答率の上昇幅の差	アンケートへの回答
p02	0.18	一度覚えた単語を再度選択することに集中しており、他の選択肢が変わったことにあまり気づかなかった。ただしスペルの 1 文字目が答えと同じものがあるとたまに混乱したので、選択肢が変わると改めてよく考え直した。
p09	0.12	正解の理由が分からずに選択肢の単語を覚えているので、正解よりも悩んだ結果間違えて選んだ選択肢の記憶の方が残っていることが多かった。次に同じ問題に出会ったときにその選択肢がなくなっていると、その次に記憶が濃く残ってる単語を選択すれば良いので、正解率は上がったと思う。同じ間違った答えを繰り返し覚えることもないので、正しい正解を暗記しやすくなった。
p03	-0.05	パニックになった。知らない単語が多いため、選択肢内の単語などを目印に覚えていた。しかし、選択肢が変わるとその目印が無くなるためほとんど覚えられなかった。
p11	-0.12	正解の選択肢を選びやすくなったような気がしたので、低い正答率が続いている時には意欲が続きやすいのかなという印象です。正答の選択肢以外の単語が簡単な単語になっていた気がしたので、消去法で選びやすかった気がしました。

としても変化回数は 20 回となり、1 回あたりの変化量が 10 であるため、1 周目の間に適切な上限が 200 を超えることはない。そのため、適切な上限が 200 を超えているものは 2 周目以降の段階で適切な上限に到達することになる。2 周目では 1 度解いたことのある問題を解いているため、初見の問題を解き続ける時と比べて正答率が高くなり、適切な上限に早く到達したと考えられる。適切な上限が 200 を超えた実験参加者らの本当の適切な上限は今回実験よりも高くなると考えられるため、この結果が提案システムの初期設定にそのまま使えるか否かについては検証が必要である。

次に、適切な上限となった後の正答率について述べる。4.5 節で述べたように、適切な上限に到達していても調整が続くことがあるため、問題が簡単になりすぎる可能性がある。そこで、適切な上限となった後の正答率を確認することで、問題が簡単になりすぎていないことを確認する。ただし、初見でない問題の正答率が高くなるのは自然なことなので、ここでは初見の問題を対象とする。よって、1 周目のうちに適切な上限へと到達した p01, p05, p07, p13 を用いる。図 9 に適切な上限に到達する前とした後の正答率の平均を示す。この図より、停止後の正答率の平均が適切な正答率の上限である 0.85 を超えていないことがわかる。よって、問題が簡単になりすぎていなかったと考えられる。ただ、p01, p05, p07 については適切な正答率の下限である 0.7 を下回っている。これは、適切な上限に到達した後に行われた調整が少なかったことを意味している。よって、上限と下限を変化させない範囲を見直す必要がある。

ると考えられる。

4.6.3 提案システムが実験参加者のモチベーションに与えた影響

実験後に行ったアンケートの結果について述べる。図 10 は提案システムでの学習と問題集での学習のどちらのほうが学習意欲が湧いたかという項目に対する回答結果である。選択肢は「選択肢が変化する」、「選択肢が変化しない」、「どちらも同じ」の 3 つである。提案システムの環境である、選択肢が変化する時のほうが問題集の環境より意欲が湧く参加者が多いという結果になった。提案システムのほうが意欲が湧いた参加者は選択肢が変わったことに対して「間違いやすい選択肢が変わったときは勉強になると思った」、「変わるほうが力がつくと思った」などと回答している。一方で、p03 と p10 の 2 名は問題集のほうが意欲が湧いたと回答している。選択肢が変わったことに対するアンケートでは p10 は「問題が難しく感じました」と回答している。p10 は上限の下降が止まらず、適切な難易度にたどり着けなかったと考えられる。これは、英語能力によって難易度を大きく調整したほうが良い可能性があることを示唆している。また、p03 は表 2 の通りほとんど覚えられなかった旨を回答している。p03 のように選択肢が変わること自体に否定的な場合は提案システムが逆効果となる可能性がある。

次に、「今後可能なら提案システムと問題集どちらを用いて学習したいか」と尋ねたアンケートの結果について述べる。問題集のほうが意欲が湧いた p03, p10 に加えて p05 が問題集を用いて学習したいと回答している。この参

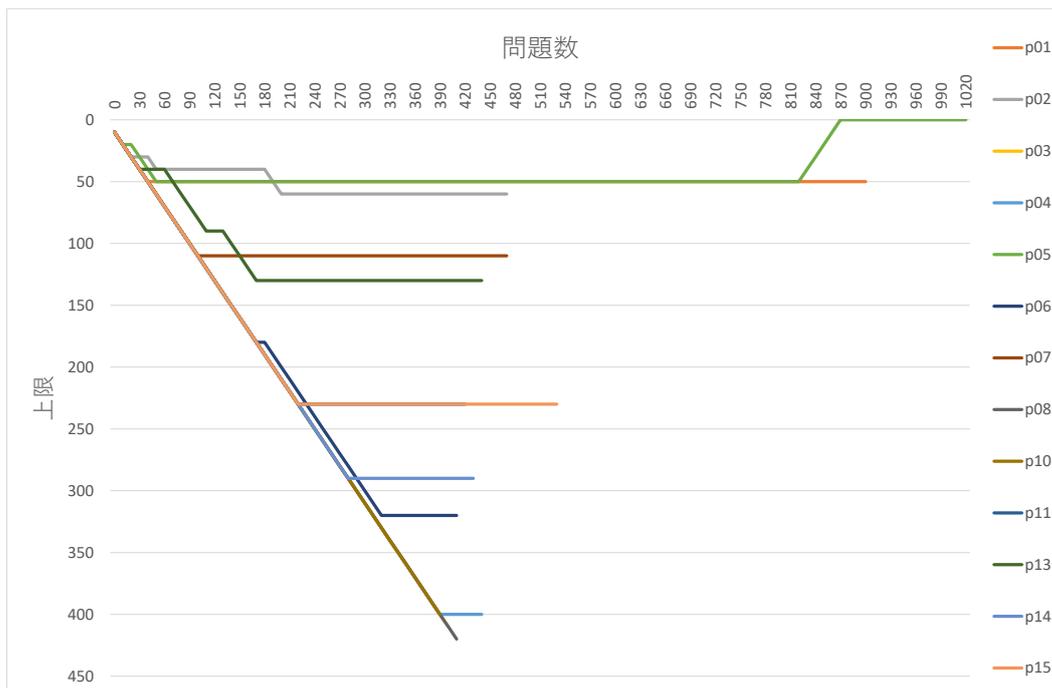


図 7 学習中における上限の推移

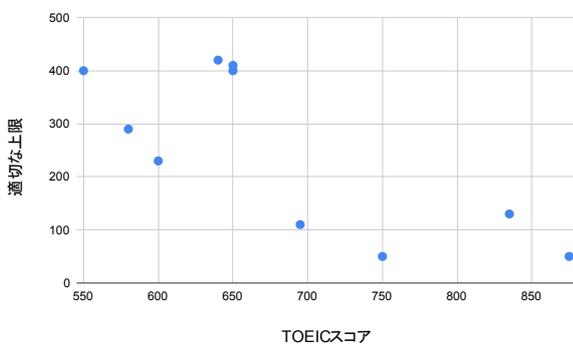


図 8 止まった時の上限と TOEIC スコアの関係

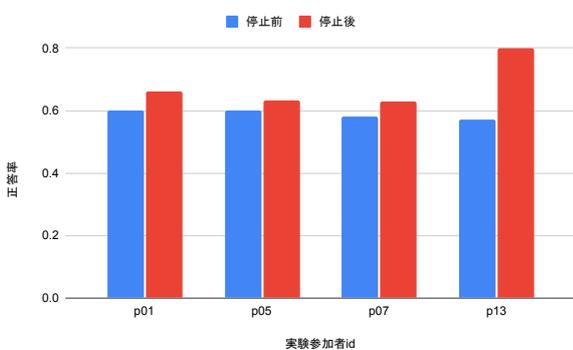


図 9 初見の問題に対する提案システムの上限と正答率の関係

加者は選択肢が変わったことについて「正解になりそうな選択肢が増えたことで、正解を覚えてないと解けない問題があった。」と回答している。p05 が解いた問題を確認すると、上限が 0、下限が 10 の段階で表 3 のような問題が生成されていた。この問題について英語を母国語とする人に意見を聞くと、おそらく“commune”を選択するが、同時に問題に苦情を言うだろうという回答を得た。その理由として、提示された選択肢の中で定義に財産や所有物を共有するということが含まれているのは“commune”のみだが、ほかの選択肢である“community”, “camp”, “center”も“commune”になりうるからだということ述べている。誤答選択肢を生成する際の上限が 0 に近い場合、このような選択肢が生成される可能性がある。このように誤答として生成されているにもかかわらず正解になりうる選択肢があることがわかった。図 7 より、p05 は上限が 0 の状態で多くの問題を解いているため、問題に解答する際の余計な負荷が大きくなり問題集での学習のほうがいと回答したと考えられる。問題集についている誤答選択肢を確認すると、すべて明らかに誤答となるものであった。提案システムには正答と比較して明らかに誤答であることを保証する機能が必要である。

5. まとめ

本論文では学習者に合わせて英語多肢選択問題の難易度

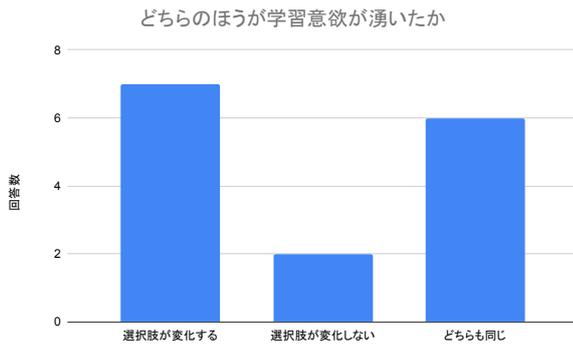


図 10 意欲に関するアンケートの結果

表 3 生成された選択肢が正答となりうる例

問題文	The religious sect established a () in a rural area where its followers could live together and share everything. No private property was allowed.
正答	commune
生成誤答選択肢 1	community
生成誤答選択肢 2	camp
生成誤答選択肢 3	center
問題集誤答選択肢 1	dirge
問題集誤答選択肢 2	prelude
問題集誤答選択肢 3	repository

を調整するシステムを提案した。さらに、提案システムの学習効果を検証する実験を行った。その結果、問題集を用いた学習と比べて提案システムのほうが有意に学習効果が高いことがわかった。この原因は提案システムが提示する選択肢が変わったことにより誤った選択肢を覚えることが減ったことだと考えられる。また、提案システムにおける難易度調整手法が正答率に与えた影響について検証した。その結果、実験参加者それぞれに対して個別の難易度調整がなされていることがわかった。また、適切な上限と TOEIC のスコアに強い負の相関があることがわかった。実験終了後のアンケートから提案システムは学習者の意欲に良い影響を与えることがあるが、逆効果になる人がいることがわかった。また、正答になりうる誤答選択肢が生成されると学習者に余計な負荷がかかるため、提案システムには生成された選択肢が誤答であることを保証する機能が必要である。

今後の課題は、実験から得られた結果を用いて難易度調整手法を改良すること、難易度の調整が学習効果に与える影響を調査すること、生成された選択肢が誤答であることを保証する機能を作成することなどがある。

謝辞 本研究の一部は JST CREST (JPMJCR16E1), JST Trilateral AI Research (JPMJCR20G3), JSPS 科研費基盤 (B) (20H04213), JSPS 国際共同研究強化 (B)

(20KK0235), 阪大 Society5.0 グランドチャレンジの補助による。

参考文献

- [1] Xu, J. and Metcalfe, J.: Studying in the region of proximal learning reduces mind wandering, *Memory & Cognition*, Vol. 44, No. 5, pp. 681–695 (2016).
- [2] 中田達也: 英単語の科学, 研究社 (2019).
- [3] Susanti, Y., Tokunaga, T., Nishikawa, H. and Obari, H.: Controlling item difficulty for automatic vocabulary question generation, *Research and practice in technology enhanced learning*, Vol. 12, No. 1, pp. 1–16 (2017).
- [4] Kurdi, G., Leo, J., Parsia, B., Sattler, U. and Al-Emari, S.: A systematic review of automatic question generation for educational purposes, *International Journal of Artificial Intelligence in Education*, Vol. 30, No. 1, pp. 121–204 (2020).
- [5] Alsubait, T., Parsia, B. and Sattler, U.: A similarity-based theory of controlling mcq difficulty, *2013 second international conference on e-learning and e-technologies in education (ICEEE)*, IEEE, pp. 283–288 (2013).
- [6] Faizan, A. and Lohmann, S.: Automatic generation of multiple choice questions from slide content using linked data, *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, pp. 1–8 (2018).
- [7] 岩田具治, 後藤拓也, 小尻智子, 渡邊豊英, 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*, Vol. 23, No. 7, pp. 16–19 (2011).
- [8] Chen, C.-Y., Liou, H.-C. and Chang, J. S.: Fast—an automatic generation system for grammar tests, *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 1–4 (2006).
- [9] Wilson, R. C., Shenhav, A., Straccia, M. and Cohen, J. D.: The eighty five percent rule for optimal learning, *Nature communications*, Vol. 10, No. 1, pp. 1–9 (2019).
- [10] Jiang, S. and Lee, J.: Distractor generation for chinese fill-in-the-blank items, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 143–148 (2017).
- [11] Murugan, S. and Balasundaram, S.: Affix-based Distractor Generation for Tamil Multiple Choice Questions using Neural Word Embedding., *Rupkatha Journal on Interdisciplinary Studies in Humanities*, Vol. 13, No. 2 (2021).
- [12] Seyler, D., Yahya, M. and Berberich, K.: Knowledge questions from knowledge graphs, *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 11–18 (2017).
- [13] Mikolov, T., Chen, K., Corrado, G. S. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781 (2013).
- [14] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [15] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Brew, J.: HuggingFace’s Transformers: State-of-the-art Natural Language Processing, *ArXiv*, Vol. abs/1910.03771 (2019).