Vol.2022-HCI-200 No.13
Vol.2022-UBI-76 No.13
2022/11/8

個別学習支援システムのための自己認識による知識からの実際の知識の予測による測定負担の軽減手法

江原 遥[1,a)]

概要：個別学習支援システムでは、学習者が何を知らないかを予測して学習を支援する必要がある。しかし、学習者の知識を正確に測定するためには、通常、学習者に長時間に及ぶ試験をうけてもらう、あるいは長期間システムを利用してもらうなど、学習者の負担が大きい。一方、学習者が「自分が知っている」と信じている知識である「自己認知よる知識」は、学習者にその知識を持っているかどうかを尋ねるだけで得られるため、実際の知識よりもかなり簡単に測定することができる。したがって、学習者の「自己認識による知識」から実際の知識を高い精度で予測することができれば、学習者の負担を少なくして個別学習支援を行う事が期待できる。 本研究では、外国語語彙学習の分野において、自己認識による知識から実際の知識を予測するための信頼性の高いデータセットを作成した。そして、自己認識による知識から実際の知識を予測する事により、測定負担を軽減する手法を提案する。

1. はじめに

第二言語の学習には、多くの単語を習得することが必要である。学習者が知っている単語を判断するために、従来は主に 2 つの方法が提案されてきた。1 つは、学習者に単語を提示し、「この単語を知っているか」と直接質問する方法で、これは「自己申告式テスト」(self-report testing) と呼ばれている。[1], [2]。学習者がその単語を知っていると答えた場合、その単語は学習者が知っていると信じているため、「自己認識知識」の一部であると言われている。

もう 1 つは、ある単語について学習者の知識を正確に測定するために、多肢選択式 (*multiple-choice questions*) のように学習者に回答してもらう方法である。**表 1** に例示するような問題に回答してもらう [3], [4] 。このような語彙問題で学習者が正解する単語は、学習者が知っている可能性が高い単語である。しかし、多肢選択問題は、特に多くの単語を問う必要がある場合、テスト実施者が作成し、参加者が回答するのはかなり負担が大きい。例えば、優秀な第二言語学習者であれば 1 万語以上の単語を知っていると想定され [5], [6]、1 万語の質問に回答してもらうのは過大な労力であり非現実的である。

必要な労力を軽減するために、自己申告型テストを用いた別のアプローチを考えることができる。自己申告テストの結果から、機械学習を用いて多肢選択式問題の結果を高い精度で予測することができ、必要な労力を軽減することができる。

このとき、本稿では、次に挙げるようなリサーチクエスチョンに答えたい。このアプローチの方が信憑性が高いか？予測精度はどの程度か？最近のニューラル機械学習の技術を用いて精度を向上させることができるのか？予測精度を測定するためには、同じ学習者セットに対して、同じ単語セットに対する自己申告と多肢選択式テストの両方の結果を含むデータセットが必要である。我々の知る限り、そのようなデータセットは現在存在せず、本研究はこれらの困難な研究課題に答えるために望ましい特性を持つデータセットを初めて提供する。さらに、本研究では、このデータセットに対する詳細な分析・予測手法を提案する。

2. 関連研究

本研究のように、多肢選択式のテストの結果を自己申告式のテストの結果で置き換えた場合に予測精度がどのようになるかを測定する研究は少ない。

しかし、単語テストにおいて、多肢選択式のテストの有

[1] 東京学芸大学
Tokyo Gakugei University. Koganei-shi, Tokyo 184–8501, Japan.
[a)] ehara@u-gakugei.ac.jp

表 1 多肢選択式の試験問題の例。被験者は文中の下線部の単語に最も近い意味を持つ選択肢を選ぶよう指示される。

It was a difficult period.

a) question b) time c) thing to do d) book

効性、自己申告式のテストの有効性については、詳しく調べられている。[5] では、本研究でも用いている Vocabulary Size Test を提案している。これは、英語の受容語彙量を測る事を目的とした多肢選択式のテストである。テキストを読む際に理解できる語彙量（受容語彙量）は、一般に、テキストを作文する時に用いる事の出来る生産語彙量より多い事が知られている [7]。このほか、[8] では、改良された試験なども提案されいる。

本研究では、項目反応理論を用いているが、VST について、VST の方法で語彙量を予測した場合と、項目反応理論で困難度パラメタを推定した場合の相関については、[9] で大学生の日本人英語学習者を対象とした詳細な研究がある。これによれば、両者はよく相関するので、VST で計算される語彙量を教育上で用いることについては特段の問題がない事が示されている。また、語彙量も項目反応理論の能力パラメタも学習者に対して一次元の値を割り振る手法であるが、英語の一部の文章だけ非常に読解ができるような被験者がいるような設定では、能力値を一次元で表せない可能性もある。しかし、[9] では、この可能性は基本的に否定されていて、大学生の日本人英語学習者の能力は、基本的には一次元で表せるという結果になっている。

3. データセットの作成

データセットの構築には、VST(Vocabulary Size Test) を使用した [5]。VST は、20,000 語から抽出された単語を問う 100 問の質問から構成されている。その名前が示すように、VST は言語学習者の語彙数を測定するために設計されている。VST を使用して学習者の語彙数を計算するには、正解した問題の数に 20 を掛ける。VST には A 版と B 版があり、本稿では A 版を使用した。両者とも難易度は同じであり、受験者の語彙数は同じように計算できるように設計されている。

[10] では、学習者に語彙テストを受けさせるという語彙テストのデータセットを作成し公開した。比較のため、同様の設定を用いた。データセットの作成には、[10] と同じくクラウドソーシングサービス ”Lancers” (`https://lancers.co.jp/`) を利用した。

ランサーズは日本の企業であるため、参加者のほとんどは日本語を母国語とするであると想定される。しかし、全員が英語学習者というわけではない。そのため、無作為抽出の場合、英語を学習していない人が受験する可能性がある。そこで、英語検定協会が実施する「TOEIC」（ `https://www.ets.org/toeic` ）を受験したことがある学習者でなければ、受験できないようにした。日本人学習者が TOEIC を受験する場合、通常、日本円で 1 万円近い額が必要になる。そのため、過去に TOEIC を受験したことがある学習者であれば、所定の費用を支払ってでも自分の英語力を測りたいという意欲があると思われる。VST のバージョン A では、191 名の参加者（＝受験者）を得た。

VST は 20,000 語からサンプリングされた 100 語から構成されているため、各問題はその問題が問う単語の頻度で並べられている。したがって、ほとんどの学習者はテストのうち、最後の方の問題に正しく答えることができない事が過去の研究で分かっている [10]。本稿では、学習者が知っている語を調べる調査法に興味がある。そのため、学習者が知っている可能性が低いこれらの語は、単純に試験から除外し、本研究の対象とはしなかった。このようにした理由は、学習者が疲れてくると、ランダムに解答するようになる可能性があるため、この問題に対処するためでもある。そこで、最も難しい 35 問を削除し、問題数を 65 問に調整した。

受験者はまず、65 個の単語リストを提示され、自己申告式のテストで、その単語を知っているかどうかを回答するよう求められた。次のテストでは、65 個の単語について、VST の多肢選択式で、 表 1 のように回答してもらった。なお、受験者は多肢選択問題に移行した後、最初の自己申告テストパートに戻ることはできないと指示した。まとめると、各単語には自己報告問題と多肢選択問題の 2 つの問題があり、その両方に各受験者が回答しなければならないが、多肢選択式の問題を解いた後に、自己申告式の問題に戻って自己申告をやり直すことは禁じられている設定とした。

項目反応理論では、各設問に対して 1 つの難易度パラメータがある。1 語について 2 つ設問がある設定にしたので、各単語は項目反応理論（IRT）において 2 つの難易度パラメータを持つことになる。

4. 実験

4.1 IRT を用いた実験

IRT は，語彙テストの他、様々な人間を対象とする試験データの解析用いられる。具体的には、どの被験者がどの問題（通常、項目, item という言い方をされる）に正答/誤答したかという行列形式のデータが与えられたときに，学習者の能力や項目の難易度を推定する確率的モデルである。これは、教育心理学などの分野を含め、幅広く利用されている [11]。有名な前日の TOEIC などの試験でも内部

的に利用されている。

項目反応理論の実装としては、“pyirt” と呼ばれる Python 言語のライブラリを用いた[*1]。このライブラリは，周辺尤度最大化法（Marginalized Maximum Likelihood Estimation）を用いてパラメタ値の推定を行う。これは、項目反応理論のうち、被験者の能力パラメタとして正規分布などを想定し、尤度関数から能力パラメタを積分消去することにより、尤度関数を周辺化して、項目パラメタのみの周辺尤度関数を最大化するものである。これにより、能力パラメタと項目に関するパラメタを同時推定することにより、被験者数を無限大にした場合の一致性を確保する事ができるという理論上よい性質がある。

“pyirt” で用いられている周辺尤度最大化法は、統計で有名な R 言語の “ltm” パッケージなどの項目反応理論の他の実装でも用いられている信頼できる手法である。

項目反応理論のうち、パラメータを得るために 2PL モデルを使用した。2PL モデルでは難易度パラメータと識別パラメータの 2 つのパラメタが各項目から取得できる。“pyirt” ライブラリのデフォルトのクリッピング範囲をパラメタ推定に用いて，全パラメータを推定した.

4.2 項目反応理論による難しさの比較

自己申告問題と多肢選択問題の難易度パラメータをそれぞれ横軸と縦軸に示し、2 軸に同じ縮尺と範囲でプロットしたのが**図 1** である。各点は 1 つの単語を表している。

点線の対角線は 図 1 の左下から右上に引いたものである。

また、横軸と縦軸は難易度パラメータの値を示しており、値が高いほど難しいと判断されることを示している。したがって、対角線の右上にある点は、多肢選択問題の難易度が自己報告問題よりも低いことを表している。また、対角線は、両者の難易度が一致するケースを表す。

一般に多肢選択問題の方が自己申告式の問題より難しいので、この結果は、直感に反しているように見える。しかし、次のように考えれば、このような結果になる理由は容易に想像できるものである。

多肢選択問題は、たとえ難しい問題でも、答えをランダムに選ぶことで、偶然、正しく答えることができる場合がある。一方、自己申告式の試験では、難しい場合には、そもそも「知っている」と申告しないので、偶然正しく答えられる場合は少ない。この結果、多肢選択式問題の難しい問題については、答えをランダムに選択肢から解答する事によって、正答してしまった分、多肢選択式の難易度が低く見積もられているというように解釈する事ができる。このため，多肢選択式問題の難易度が低く見積もられているものと思われる．Wilcoxon 検定の結果、縦軸の値の列は横軸の値の列より統計的に有意に小さいことが示された（$p < 0.01$ ）。

縦軸の問題が横軸の問題よりも難しいかどうかを調べるために、次のような実験を行った（**図 2**）。まず、VST 試験のバージョン A の 191 名の受験者を 100 名と 91 名に分けた。実知識問題のパラメータは前者 100 人の回答のみから推定し、知覚問題については 91 人全員の回答から推定した。なお，後者の 91 人× 65 問，合計 5,915 問の回答データは使用していないことに注意して欲しい。この時、図 2 のうち、破線部分の回答の予測精度を評価対象とした.

IRT を用いて、図 2 の左下破線領域を予測する場合、主に 2 つの方法を用いた。IRT では、学習者ごとの能力値 $theta_j$ と問題ごとの難易度パラメータ d_i を算出することができる。**各語には多肢選択問題と自己申告問題があることに注意すると**、各語には 2 つの難易度パラメータがある。

どちらの難易度パラメータを予測に用いるかを変えることで、以下の 2 つの設定を比較した。1 つ目の方法は、100 人の学習者の回答のみから推定した、多肢選択問題の難易度を使用する方法である（ 図 2）. 2 つ目の方法は、全ての回答から推定した自己申告問題の難易度を、多肢選択問題の難易度として代用してしまう方法である。

実験結果によれば、破線部の予測精度は、第一の方法が 0.724 であるのに対し、第二の方法は 0.697 であった。これらの結果の差は Wilcoxon 検定により統計的に有意であった ($p < 0.01$). この結果は、多肢選択問題に対する回答を予測するためには、全受験者の回答から推定した自己申告の設問の難易度パラメタで代用する場合（第二の方法）は精度を低下させることを示すものである。すなわち、破線部以外の他の他の受験者のデータが少量でも入手できるのであれば、これを用いて、多肢選択式の設問の難易度を直接推定方が有効であることを示している。

4.3 BERT に基づく手法

図 2 の破線部分を予測するために、BERT（bidirectional encoder representations from transformers） [12] などの深層転移学習手法が高い予測性能を達成したと報告されているので、利用したいニーズがある。しかし、BERT は自然言語列のみを入力とすることができる。一方、我々のタスク設定では、異なる学習者に対する予測を行う際は、能力などの学習者の特性を考慮する必要があるので、学習者 ID などの情報を BERT が入力とできるように工夫しなければならない。

ここでは、本研究のタスク設定を、BERT が扱うことのできる系列分類タスク (Sequence Classification Task) に変換する簡便な手法を示す。

このために、2 種類の特殊なトークン（単語のように扱えるが言語学上単語ではない単位）を導入する [**USRn**] と [**SQ**] である。これらのトークンを入力文に加えることに

[*1] https://github.com/17zuoye/pyirt

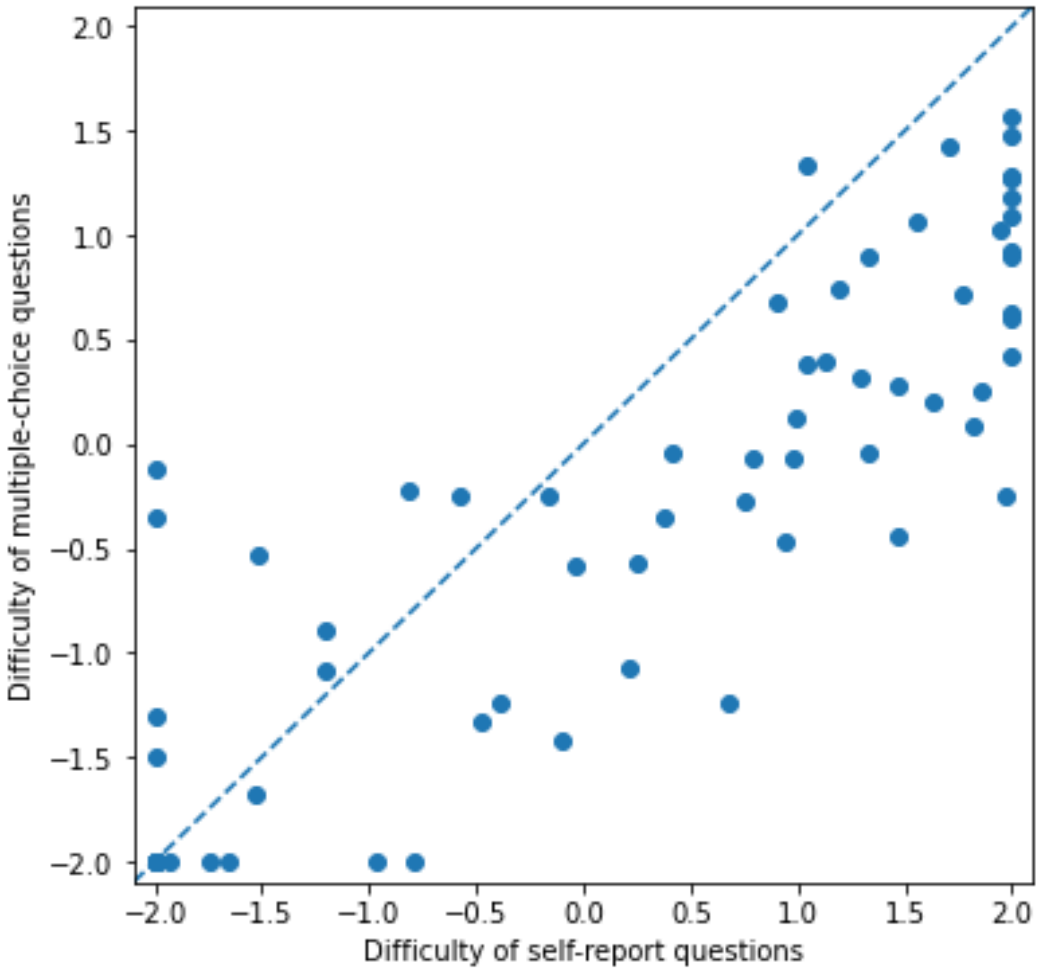

図 1 自己申告問題の難易度（横軸）と多肢選択問題の難易度をプロットしたもの（縦軸）。各点は単語を表す。

よって、学習者 ID などを含む情報を BERT が扱える配列分類タスクに変換することができる。

特殊トークン [**USRn**] について説明する。ここで、n は受験者 ID に置き換えられ、各受験者や学習者 (ユーザ) を表す。この特殊トークンは、入力文の先頭に置かれ、分類器はこのトークンで指定された受験者の応答を予測する必要があることを意味する。**図 3** の例では、“It is a difficult period.” という入力系列に対してユーザ **USR3** の反応 (正答/誤答) を予測する事が目的である事を、特殊トークンの付与によって、BERT に指示する。この場合、**USR3** は多肢選択問題を不正解であったため、多肢選択問題に対するラベルは 0 である。この正答/誤答の情報のみを微調整 (fine-tuning) に用いるデータとする。このとき、**USR3** が選択した選択肢自体は、BERT の fine-tuning に用いられないことに注意されたい。つまり 図 3 の例では、**USR3** が “question” を選択したことは無視される。

この挙動は、一見、情報を落としているように見えるが、IRT でも同様に情報を落としているので、IRT と公正な性能評価をするうえでは必要なことである。IRT を用いた手法では、受験者がどの誤答選択肢（distractor と呼ばれる）を選択したために不正解になったかはそもそも困難度パラメタなどのパラメタ推定時に用いられない。当然、受験者トークンは受験者数と同じ数だけ存在することになる。

また、今回の我々の設定では、BERT は自己報告式のテストを扱うことができる必要もある。そのために、自己報告式の質問を表す特別なトークン [**SQ**] を導入したので説明する。このトークンを、学習者を指定する [USRn] トークンの直後に置くことで、BERT 識別器に自己申告式の語彙問題を予測したいことを知らせる。これにより、BERT は [**SQ**] に続くトークンだけを考慮するように fine-tuning される。この受験者は、自己申告のテストで period という単語を知っていると回答したため、 図 3 の例ではラベルが 1 になっている。

4.4 実装

本手法は、BERT [12] などの転移学習モデルを実装した標準ライブラリである **transfomers** ライブラリ [*2] の

[*2] https://github.com/huggingface/transformers

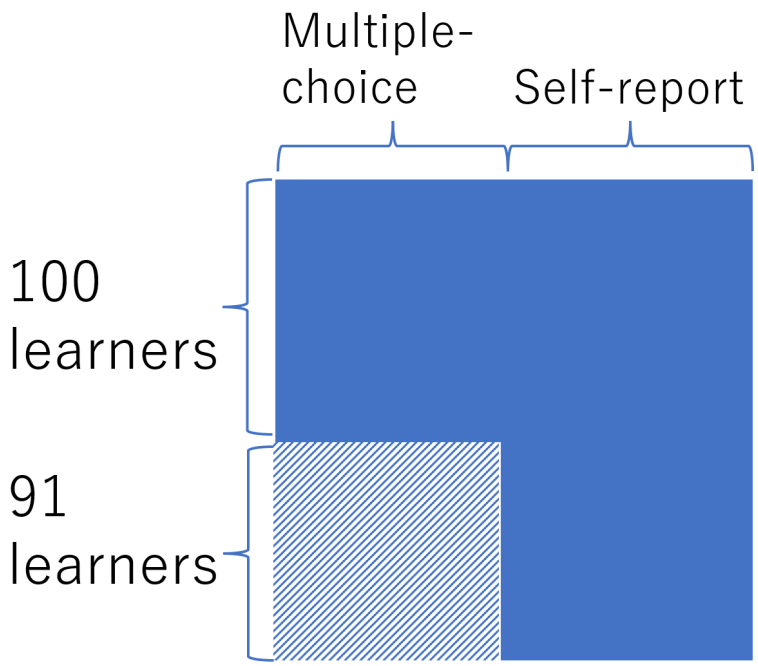

図 2 実験設定。塗りつぶした領域は、パラメータの推定に使用された訓練データを表す。塗りつぶした領域で識別器を訓練し、破線部の精度を評価する事になる。

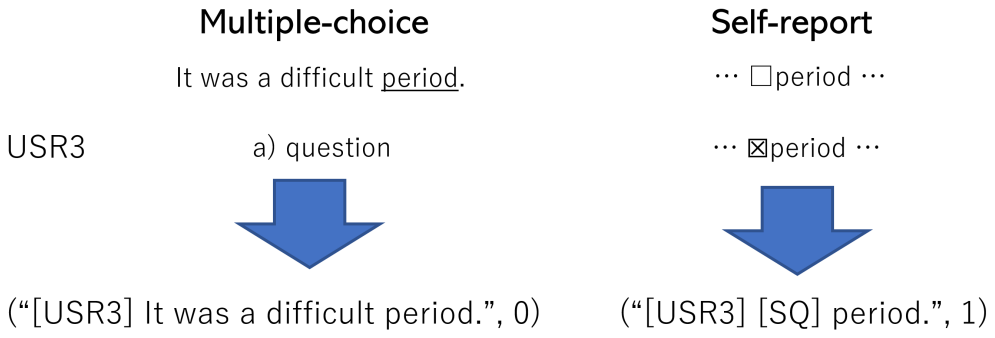

図 3 語彙テスト結果データセットを BERT 入力に変換する例。

BertForSequenceClassification 関数を用いて実装した。また、Transformers は、“pytorch” フレームワーク上で実装されているものを用いた[*3]。事前学習済みのモデルには、 **transformers** ライブラリからダウンロードできる **bert-base-uncased**、**bert-large-uncased**、**bert-base-cased**、**bert-large-cased** を使用した。

前述の特殊なトークンは **add_tokens** 関数で追加した。変換後の系列分類器を構築するための微調整には、単に **BertForSequenceClassification** を使った。Fine-tuning の最適化に用いる関数としては、標準的な Adam? 法を用いた。学習率は 10^{-5} とした。また、エポック数は全モデルで 9 回とした。

この BERT ベースの手法の精度は、9 番目のエポックで 0.729（ウィルコクソン検定、$p < 0.01$）と IRT ベースの手法の精度よりわずかだが統計的に有意に優れていた。この結果は、本タスクが語彙問題の文脈にあまり依存しないように見えるが、文脈を考慮することによって、その精度をわずかに向上させることができることを示している。最終的な結果は**表 2** にまとめられている。BERT（bert-large-cased） が、我々の実験において統計的に有意な最高の精度を達成したことが観察された。また、bert-base-cased と bert-base-uncased の精度値が低いことから、高い予測精度を実現するには、事前に学習したモデルの大きさが重要であることがわかる。

[*3] https://pytorch.org/

表 2 図 2 の塗りつぶされた部分を学習データとして、図 2 の破線部分に相当する、91 人の学習者の多肢選択問題の正解/不正解を予測する設定での各手法の予測精度。(**) は、太字部分の数値と下線部分の数値より統計的に有意な差がある事を示している ($p < 0.01$)。

手法	精度
IRT (91 人の各学習者の能力値 - 191 人の学習者の自己申告データから推定した難易度パラメタ) で判定	0.675
IRT (91 人の各学習者の能力値 - 100 人の学習者の多肢選択式データから推定した難易度パラメタ) で判定	<u>0.724</u>
BERT (bert-base-uncased)	0.718
BERT (bert-base-cased)	0.693
BERT (bert-large-uncased)	0.722
BERT (bert-large-cased)	**0.729** (**)

5. 考察：自己申告式の知識から実際の知識を予測する UI について

前節までの結果から、図 2 に示すように、自己申告式と多肢選択式を両方回答した被験者を最初に一定数用意すれば、時間のかかる多肢選択式の試験を一切行わなくとも、その結果は、0.729 程度の精度で予測する事が可能であることが示された。

これがどの程度実用的であるかは、具体的な応用と、多肢選択式と自己申告式の回答にかかる時間の差に依存すると思われる。今回用いた多肢選択式の VST では、一般に 100 語解くのに早い人で 30 分、実際に解いてみたところ、40 分程度はかかる?。これは、40 分で計算すると 1 語あたり、24 秒かかっている計算になる。VST では単語が簡単な方から難しい方に並んでいるため回答しやすいという効果があることを考えると、表 1 のような多肢選択式の問題を解くのに約 20 秒〜約 30 秒程度はかかると考えればよい。

自己申告式の場合、この時間と比べてどの程度早く回答できるかが問題になる。TOEIC の受験者が目指すべき読解速度は、1 分間に 150 語程度であるとされている（`https://eigo-kochi-training.com/sokudokunitsuite/`）。つまり、語を読むだけであれば、1 語につき、0.4 秒程度しかかからないことになる。実際には、これに加えて、その後を自分が知っているかどうか考えて反応する時間が加わるが、多肢選択式のように、1 語につき 20 秒も回答にかかる事は稀であると考えられ、長く見積もっても 1 語につき 4〜5 秒程度ではないだろうか。（今回の設定では、被験者数を確保するため、試験時間を計測する事はしていない。作業中に一時中断する被験者などもいる可能性があるためである。）

概算としては、実際の知識を多肢選択式の設問を解かせて回答する時間に比べて、5 6 倍程度は早く自己申告式テストに回答する事ができると考えられる。すなわち、100 語について、5 6 分程度で自己申告を行う事ができると考えられる。

これは、例えばわからない単語を学習者に合わせて個別指定して、読解支援システムなどで、最初に使い始めるまでの時間を短縮する対策（使い始めではユーザに関するデータがないコールドスタート問題に対する対策）としては、相当に有効であると考えられる。英語学習のような長期にわたるシステムでは、必要なデータは学習を進めるうえで時間があるときに取得する事も可能であると思われるので、まずは、自己申告式のテストで、短時間でデータを収集し、支援を行う仕組みが有効であると考えられる。

6. おわりに

本研究では、コストのかかる実際のテスト結果を予測するために、多肢選択式と自己申告式の質問を比較するデータセットを開発した。実験の結果、IRT に基づく最適な手法は、BERT に基づく手法よりもわずかに、しかし統計的に有意に低い精度を達成することが示された。この結果は、語彙問題の単語の意味的文脈を考慮することにより、語彙問題の長さは一般的に短く文脈を考慮する効果は薄いと思われるにもかかわらず、精度をわずかに向上させることができることを示唆している。

今後の課題として、実際に、自己申告式の試験から多肢選択式の回答を予測する事でコールドスタート問題にどの程度対処できるのかを検討する実験を行う事が挙げられる。

また、本稿のような多肢選択式テストを自己申告式テストで代用する方法は、英語学習以外に数学や他の教科でも、代表的な用語を知っているかどうかの確認テストの形で実施する事が可能であることが容易に推察される。そこで、英語学習以外の設定で、本研究のような自己申告式テストの結果から、多肢選択式テストなどの実際に「解かなければならない」テストの結果をどの程度予測する事ができるのか計測する事があげられる。

謝辞 本研究は、JST ACT-X 研究費（JPMJAX2006）の支援を受けた。

参考文献

[1] Wesche, M. and Paribakht, T. S.: Assessing second language vocabulary knowledge: Depth versus breadth, *Canadian Modern Language Review*, Vol. 53, No. 1, pp. 13–40 (1996).
[2] O'Dell, F., Read, J., McCarthy, M. et al.: *Assessing vocabulary*, Cambridge university press (2000).
[3] Nation, I. and Beglar, D.: A vocabulary size test. The

Language Teacher, 31 (7), 9–13 (2007).
[4] Beinborn, L., Zesch, T. and Gurevych, I.: Predicting the Difficulty of Language Proficiency Tests, *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 517–530 (online), DOI: 10.1162/tacl$_{a}$0200(2014).
[5] Nation, I.: How large a vocabulary is needed for reading and listening?, *Canadian modern language review*, Vol. 63, No. 1, pp. 59–82 (2006).
[6] Laufer, B. and Ravenhorst-Kalovski, G. C.: Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension, *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30 (online), available from ⟨https://eric.ed.gov/?id=EJ887873⟩ (2010).
[7] Nation, I.: How Large a Vocabulary is Needed For Reading and Listening?, *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82 (2006).
[8] Beglar, D. and Nation, P.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).
[9] Beglar, D.: A Rasch-based validation of the Vocabulary Size Test, *Language Testing*, Vol. 27, No. 1, pp. 101–118 (online), available from ⟨https://doi.org/10.1177/0265532209340194⟩ (2010).
[10] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).
[11] Baker, F. B.: *Item Response Theory : Parameter Estimation Techniques, Second Edition*, CRC Press (2004).
[12] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. of NAACL*, Minneapolis, Minnesota, pp. 4171–4186 (2019).