N-019

学習者トークン埋め込みの導入と能力特性の抽出による解釈しやすい学習者反応予測手法

Interpretable Prediction Methods of Learners' Performance via Learner Token Embeddings

江原 遥[1)]
Yo Ehara

1 はじめに

学習支援システムにおいて，学習者が項目に回答できるかどうかを予測する事は，学習者に合った水準の項目（設問）の提示など，適応的学習支援を行うための基本的なタスクである．学習者が項目に回答した履歴のデータがあれば，教育心理学などで能力や難しさのモデル化に多用される項目反応理論 (Item Response Theory，以下 IRT) [1] を用いることで，学習者の能力と項目の難しさを推定し，学習者の反応予測を行う事ができる．

IRT に基づくモデルは通常，学習者の回答パターンにのみ依存し，項目（設問）が自然文で書かれていても文意を理解しない．自然言語処理においては，近年，Transformer モデルに代表される深層言語モデルが自然文理解で高い性能を示している [2]．従って，設問文の理解に，これらの深層言語モデルを用いたい．しかし，これらの言語モデルは，通常，言語のみをモデル化するため，学習者ごとに異なった判定を行うことができず，学習者反応の予測に用いることが難しい．これは，深層言語モデルを用いた設問文の文意を考慮した学習者適応がそのままでは行えないことを示している．

本研究では，深層言語モデルを設問文を考慮した学習者反応の予測問題に適用する簡便な方法を提案する．提案手法は，自然文で記述されている設問に対して，複数の学習者が正答/誤答が明瞭にわかる形式（多肢選択式など）で回答する試験結果データであれば，幅広く適用することが可能である．しかし，評価のためには，特定の問題に限定して，提案手法が設問文の文意を考慮した学習者反応の予測がどの程度行えているかを計測する必要がある．

そこで，本研究では，設問文の文意を考慮した判定が行えているかを評価するため，外国語学習の語彙学習支援における多義語の各意味を知っているかを問う語彙テストデータセットを作成し，これを用いて提案手法を評価した．以後，理解を容易にするため，この問題に限定した用語を用いて提案手法の解説や評価を行うが，技術的には，前述のように，幅広い問題に適用可能である．まず，学習支援システムのために，典型的な語義の知識状態から，非典型的な（意外な）語義の知識状態を予測する課題についての評価用データセットを作成する．具体的には，1 つの語について，典型的な語義で使われている文と意外な語義で使われている文を用意・作問し，クラウドソーシング上でデータ収集を行った（表 1，表 2）．設問は，複数の英語母語話者の確認の取れたものを用いた．IRT を用いて典型的/意外での設問の困難度等の分析を行い，学習者反応データ上でも，意外な語義の方が典型的な語義より難しい事を示す．作成したデータセット上で，典型的な語義のテスト反応から意外な語義への反応をどの程度予測できるか評価する（6 節）．

学習者反応の予測では，大別して 2 種類の手法を比較した．まず，教育心理学などで能力や難しさのモデル化に多用される，設問文の文脈を考慮しない，前述の IRT [1] を用いた手法である．次に，大規模な母語話者コーパスを事前学習に用いることで設問文の文脈を考慮する事ができる Transformer モデルの手法（Bidirectional Encoder Representations from Transformers, BERT [2] など）に基づく提案手法である．前述のように，Transformer モデルは，能力の考慮など，学習者によって異なる結果を予測する仕組みを通常持たない．本研究では，Transformer モデルを学習者反応予測問題に適用する手法をあわせて提案し，その予測性能が IRT による手法より高いことを意味する．また，IRT の利点は学習者の能力値等を合わせて推定できる解釈性にあるが，Transformer モデルから IRT で推定した能力値とよく相関する値を抽出する手法も提案する．本研究で作成したデータセットは，今後[1)]で公開する予定である．

本研究の内容は，教育データマイニングのトップ国際会議である Educational Data Mining 2022 の short paper に採択された（査読付き）[3].

2 関連研究

2.1 外国語学習支援の学習者反応データセット

本研究では，設問文を考慮した学習者反応予測を行いたい．そのためには，設問文の文意を考慮することが重要であるような設定で試験を行い，その結果を記録したデータセットが必要となる．3 節で詳述するこのデータセットの類似データセットについて，関連研究を述べる．1 つは，語学学習アプリ Duolingo 上の設問に対する回答データを用いた SLAM データセット [4] である．もう 1 つは，多数の語学学習者に対して，文中のわからない語をアノテーションさせた複雑単語推定（Complex Word Identification, CWI）のデータセット [5] である．

これらのデータセットと本研究で提示するデータセットの違いとして，各学習者は多くある設問のうちのごく一部にしか回答していないという点が挙げられる．言い換えると，学習者を行，設問を列とし，学習者の設問に

1） 東京学芸大学，Tokyo Gakugei University.

1） http://yoehara.com/

表 1　実際の設問例．

It was a difficult period.
a) question
b) time
c) thing to do
d) book

対する回答内容を要素とする行列を考えた場合，これらのデータセットでは行列が疎になっている．項目反応理論は，学習者の設問に対する回答内容から，設問の難しさや学習者の能力値を推定を目標とするが，この推定のためには，各学習者がほぼ全ての設問に回答している形式のデータセットであることが望ましい．また，どちらのデータセットでも，文中の語に対する学習者の回答が記録されてはいるものの，設問について，今回のデータセットのような語の通常の用例と，意外と思われる用例といったようなアノテーションはされていない．さらに，[5] を含む CWI のデータセットでは，一般に，提示された文に対して，学習者が難しいと感じた語が記録されているだけであり，学習者が実際にその語の意味を適切に理解しているかテストを通じた確認はしていない．すなわち，意味は理解できたが難しいと感じてアノテーションした場合もあれば，単純に意味が分からなかった場合も含まれる．

2.2　提案手法の関連研究

近年にも，BERT を用いた教育応用が提案されているが [6, 7, 8]，これらの研究では個人化学習者支援については扱われていない．また，著者の知る限り，応用言語学の分野においても，外国語学習者を対象に，ある語の意外な意味／典型的な意味の 2 種類を同時に試験したデータセットを作成し，項目反応理論を用いて各意味の難しさを試験結果データから客観的に推定・分析した研究は見当たらない [9, 10, 11].

3　語彙テスト作成・データセット

語彙テスト作成・データセット作成は，著者が過去に語彙テスト結果データセット作成時の設定に準じて行った [12]. データセットはクラウドソーシングサービス Lancers[2)] から，2021 年 1 月に収集した．英語学習にある程度興味がある学習者を集めるため，過去に TOEIC を受験したことがある学習者のみ語彙テストを受けられると明記して，データを収集した．その結果，235 名の学習者（被験者）から回答があった．以後，用語の統一のため，被験者という語は用いず，学習者という語を用いる．Lancers の作業者は大部分日本語母語話者であるため，このデータセット中の学習者の母語は，大部分日本語を母語とするものと思われる．

まず，通常の語彙テストとしては，文献 [12] と同様に，Vocabulary Size Test（VST）[13] を用いた．ただし，VST は 100 問からなるのに対して，[12] では，低頻度語に関する設問では，Lancers 上のどの学習者もほとんどチャンスレートしか回答できていなかったことから，学習者の負担感を減らし的確な回答を集めやすくするため，低頻度語 30 問を削った．すなわち，残り 70 問を通常の語彙テストとして用いた．この設問例を表 1 に示す．文中の単語に下線が引かれてあり，学習者は，この単語と交換した際に元の文と意味が最も近くなる選択肢を選ぶように求められる．この際，文法的から選択肢を絞れてしまわないように，選択肢は下線部と文字通り置き換えても正文となるように作られている．例えば表 1 であれば，複数形の選択肢が内容に配慮されている．

2) https://lancers.co.jp/

表 2　意外な意味を問う設問例．

She had a missed ______.
a) time
b) period
c) hour
d) duration

一方，学習者にとって意外である思われる用例については，著者が作問し，英語母語話者を含む静岡理工科大学の教員複数名に問題として成立しているか確認を取る方法で，作成した．この際，表 1 と同様の形式にして，“period” という単語について 2 つの設問がある事が分かってしまうと，意外な語義については通常の語義以外の選択肢を選ぶことで，選択肢を絞り，意味を知らなくても回答できてしまう．そこで，本研究では，次の 2 つの工夫を行った．

1. 意外な語義を問う設問については，下線部の意味について問う形にはせず，空欄を埋める形式の問題とした．これにより，意外な語義については正答を知らなければ，どの語についての設問であるのかもわからないようにした．
2. 通常の語義についての設問を先に行ってしまうと，そこで出てきた単語と同じ語が正答であろう，という推測ができてしまう．そこで，意外な語義についての設問群を最初に行い，通常の語義についての設問群に移動したら，意外な語義についての設問群には戻れないようにした．

この 2 つの工夫を施した実際の設問例が表 2 である．“period” には通常の「期間」の他に「生理」という意味があり，これを問うている．学習者は，70 問の通常の用例の語彙テストの前に，表 2 のような設問を 13 問解くように求められる．ただし，先に解く表 2 の形式の選択肢が，表 1 の形式の問題に影響していないかどうかを後で確認できるよう，意外な語義ではあるが，通常の語義の設問群の側に対応する設問がない設問を 1 問設けた．これにより，対応する問題は 12 問となる．

4　項目反応理論

項目反応理論のモデルについて，説明する．学習者の数を J，設問（項目, item）の数を I とする．簡単のため，学習者の添字（index）と学習者，項目の添字と項目を同一視する．例えば，i 番目の項目を，単に i と書くことにする．y_{ij} は，学習者 j が項目 i に正答するとき 1，誤答であるとき 0 であるとする．試験結果データ $\{y_{ij}|i \in \{1,\ldots,I\}, j \in \{1,\ldots,J\}\}$ が与えられたとき，2 パ

ラメター—モデル (2PLM) では，学習者 j が項目 i に正答する確率を次の式でモデル化する．

$$P(y_{ij} = 1|i, j) = \sigma(a_i(\theta_j - d_i)) \quad (1)$$

ここで，σ は $\sigma(x) = \frac{1}{1+\exp(-x)}$ で定義されるロジスティックシグモイド関数である．σ は $(0, 1)$ を値域とする単調増加関数であり，$\sigma(0) = 0.5$ である．実数を $(0, 1)$ の範囲に射影し，確率として扱うために用いられている．式 1 において，θ_j は能力パラメタ（ability parameter）と呼ばれ，学習者の能力を表すパラメタである．d_i は困難度パラメタ（difficulty parameter）と呼ばれ，項目の難しさを表すパラメタである．式 1 より，θ_j が d_i を上回る時，学習者が正答する確率が誤答確率より高くなる．$a_i > 0$ は，通常，正の値を取り，識別力パラメタ（discrimination parameter）と呼ばれる．この値が大きいほど，$\theta_j - d_i$ が正答確率/誤答確率に大きく影響するようになる．$\theta_j - d_i$ を用いて，学習者 j が設問 i に正答するか否かが見分けやすくなる事を表しているため，「識別力」と呼ばれる．より直観的には設問 i が，能力値が高い学習者と低い学習者を正確に見分けられるという意味で良問であることを示している．

なお，項目反応理論には，多肢選択式の設問で分からなくても選択肢を無作為に選んで正答出来てしまう確率を考慮する 3 パラメータモデル（3PLM）が存在するものの，今回のデータセットの被験者数（学習者数）では，被験者数が少なすぎてパラメタ推定が不安定であるという報告 [14] があるため，よりパラメタ数の少ない 2PL モデルを用いた．

5　項目反応理論を用いた分析

前述のように，作成したデータセットでは，同じ語でも設問文の文意によって難しさ（困難度）が異なると思われる．本節では，まず，項目反応理論を用いて作成したデータセットを分析することで，実際に同じ語でも設問文の文意によって難しさが異なっていることを，困難度パラメタの値の違いを通じて確認する．

項目反応理論の困難度・識別力の各パラメタを求めるには，Python 言語のライブラリである pyirt（`https://github.com/17zuoye/pyirt`）を用いた．これは，周辺化最尤推定（Marginalized Maximum Likelihood Estimation）により項目反応理論を行うライブラリである．前述のデータセットに対して，2PL モデルを用いて困難度と識別力パラメタを求めた．表 1 と表 2 のように，設問のペアが 12 組ある．通常の用例，学習者にとって意外と思われる用例の困難度パラメタを，それぞれ横軸，縦軸に表し，横軸と縦軸の縮尺・範囲を同一にプロットし図 1 に示した．各点は語を表す．

困難度の比較　図 1 の左下から右上まで，点線で対角線を示した．図 1 の横軸・縦軸とも，困難度パラメタの値であり，この値が大きいほど難しいと判定される．そのため，この対角線より左上にある点は，通常の用例の困難度より，学習者にとって意外と思われる用例の困難度の方が，語彙テスト結果データからも学習者にとって

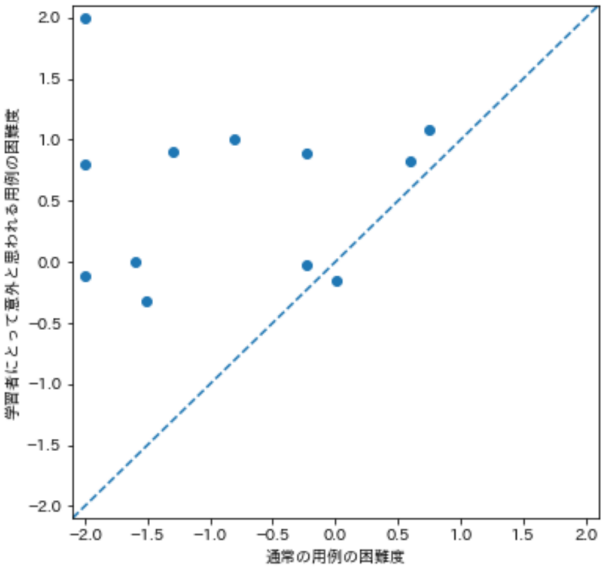

図 1　各語の，通常の用例の困難度（横軸）と学習者にとって意外と思われる用例の困難度（縦軸）のプロット．各点は各語を表す．

回答が難しいと判定された語ということになる．今回は設問数が少ないので，図 1 の結果が偶然得られた可能性がどの程度あるか検証するため，横軸の値の列と縦軸の値の列で統計的検定を行った．Wilcoxon 検定の結果，縦軸の値の列が統計的有意に横軸の値の列より大きかった（$p < 0.01$）．すなわち，縦軸の設問群の方が横軸の設問群より難しかった事が示唆される．

識別力の比較　識別力についても，図 1 と同様にプロットし，図 2 に示した．識別力は，直観的には，高いほど，その問題で（他の問題で推定される）能力値が高い学習者と低い学習者を分けることができるという意味で，良問である度合いを表す．学習者にとって意外と思われる用例は，能力値が高い学習者でも知らないことがあり，低い学習者でも知っていることがあるため，通常の用例よりも識別力が低いと予想される．全ての語について，通常の用例の方が，意外と思われる用例よりも識別力が高いと推定されている．この結果も，Wilcoxon 検定の結果，統計的に有意であった（$p < 0.01$）．識別力のプロットについては紙面の都合のため付録に記す．

困難度と識別力のプロット　識別力は，直観的には，他の設問で能力値が高いと推定された学習者が簡単な問題に誤答してしまう，また，他の設問で能力値が低いと推定された学習者が難しい問題に正答してしまう場合に低下する．今回の設定では，前者のケースはあまり見られないが，後者のケースで回答が分からない学習者がとりあえず選んだ選択肢に正答してしまう事はあるので，困難度の高い説問ほど，識別力が低く出ることが予想される．困難度の高い問題の識別力を向上させる 1 つの方法としては，選択肢に「わからない」や未回答を許すという方法が考えられる．しかし，クラウドソーシング上でこの方法を取ると，ほぼ全ての設問に対して「わからない」と回答するケースなどがあるため，今回はこの方

表 3　図 5 斜線部の予測精度 (accuracy)．

手法	精度
IRT (能力 - 235 人から推定した典型的な語義の困難度)	0.544
IRT (能力 - 135 人から推定した意外な語義の困難度)	0.644
提案手法 (bert-large-cased)	0.674 (**)
提案手法 (bert-base-cased)	**0.688** (**)
提案手法 (bert-base-uncased)	0.655
提案手法 (roberta-base)	0.681 (**)
提案手法 (albert-base-cased)	0.671 (*)

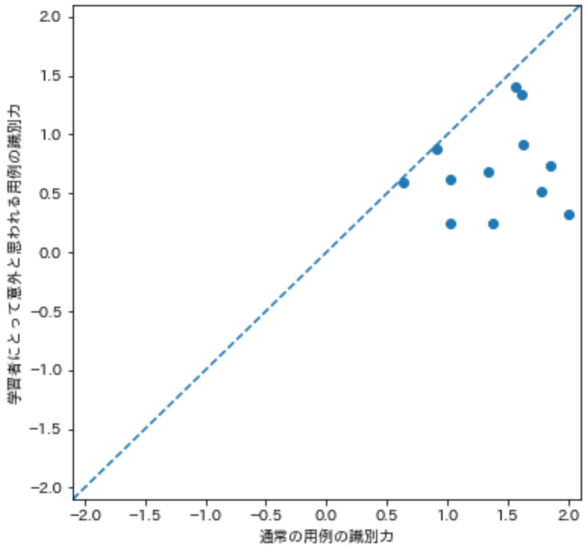

図 2　各語の，通常の用例の識別力（横軸）と学習者にとって意外と思われる用例の識別力（縦軸）のプロット．各点は各語を表す．

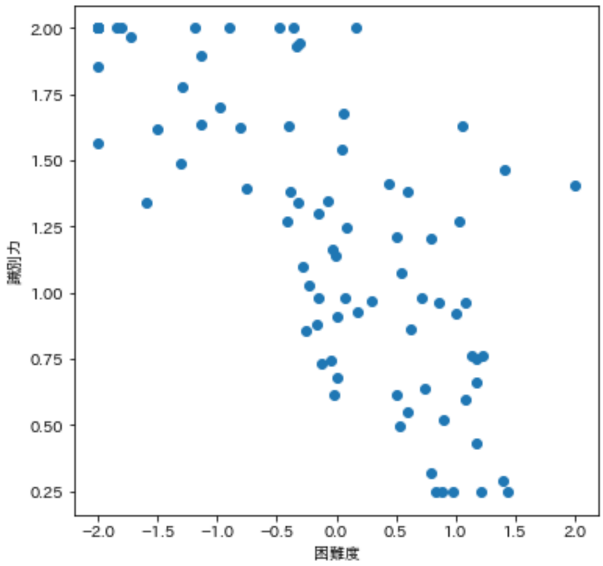

図 3　全語の困難度（横軸）と識別力（縦軸）の関係．

法は取らなかった．全ての語についての困難度パラメタと識別力パラメタのプロットを示した（図 3）．図 3 から，困難度が増加するにつれて，識別力が減少していく傾向が見て取れる．困難度パラメタと識別力の間の困難度パラメタと識別力パラメタの間のスピアマンの順位相関係数は-0.739（$p < 0.01$）で，「強い相関」が認められた．

6　学習者反応予測による評価

6.1　IRT による学習者反応予測

語の意外と思われる語義の難しさを典型的な語義の難しさで代替してしまうと，学習者が設問に正答/誤答するかを IRT で予測する際，どの程度の悪影響があるのだろうか？これを調べるために，次の実験を行った．まず，235 人の学習者を 135 人と 100 人に分ける（図 5）．意外と思われる語義の設問群（12 問）のパラメタについては前者の 135 人の学習者反応だけから，典型的な語義の設問群（70 問）のパラメタについては 235 人全員の学習者反応で推定する．この推定の際には，後者の 100 人× 12 問，計 1,200 件の回答データは用いていないことに注意されたい．式 1 より，推定された学習者の能力値 θ_j, 語義の困難度 d_i を用い，$\theta_j > d_i$ であれば学習者 j が設問 i に正答，そうでなければ誤答と判定できる．設問 i の困難度パラメタとして，意外と思われる語義の 12 問の困難度パラメタを直接用いた場合と，対応する語の典型的な語義の困難度パラメタで代替した場合で，この 1,200 件の回答データの予測精度を比較した．予測精度 (accuracy) の結果を表 3 に記す．その結果，直接用いた場合の予測精度は 64.4%，典型的な語義の困難度で代替した場合は 54.4% と，10 ポイントの差が出た．この差は，Wilcoxon 検定で $p < 0.01$ で有意であった．この結果から，学習者反応の予測における，語の語義ごとに困難度を推定することの重要性がわかる．より直接的に言い換えれば，この結果は，語の意外な用例の難しさを，語の典型的な用例の難しさで置き換えると，学習者反応予測の精度が著しく低下することを示唆している．

6.2　Transformer モデルと IRT の性能比較

IRT を用いた手法は，学習者反応のみに依存し，設問文の意味などは全く考慮されていない．では，設問文の意味をも考慮した学習者反応予測を行うと，学習者反応のみを用いた IRT の手法より高精度に予測できるのだろうか？深層言語モデルのうち，自然言語処理で文意を考慮した予測手法として近年多用される，Bidirectional Encoder Representations from Transformers (BERT)[2] に代表される Transformer モデルと IRT の予測性能を比較した．

Transformer モデルは近年の深層転移学習による深層言

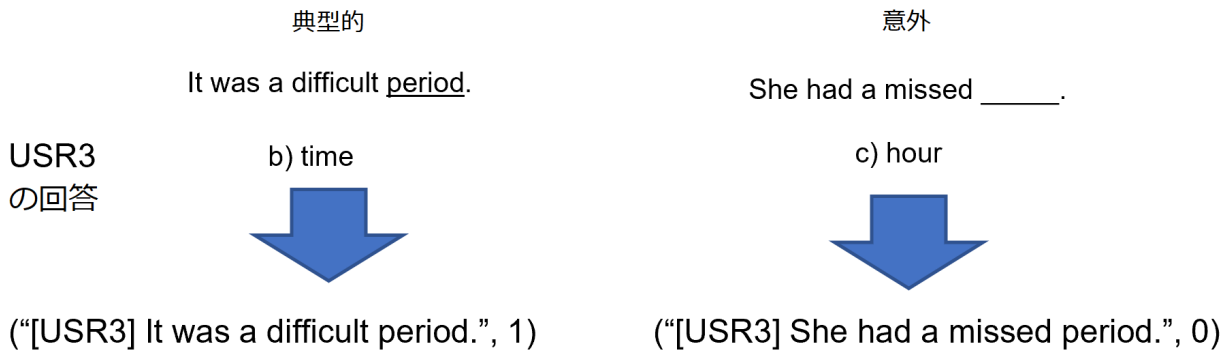

図 4　学習者トークンの導入．

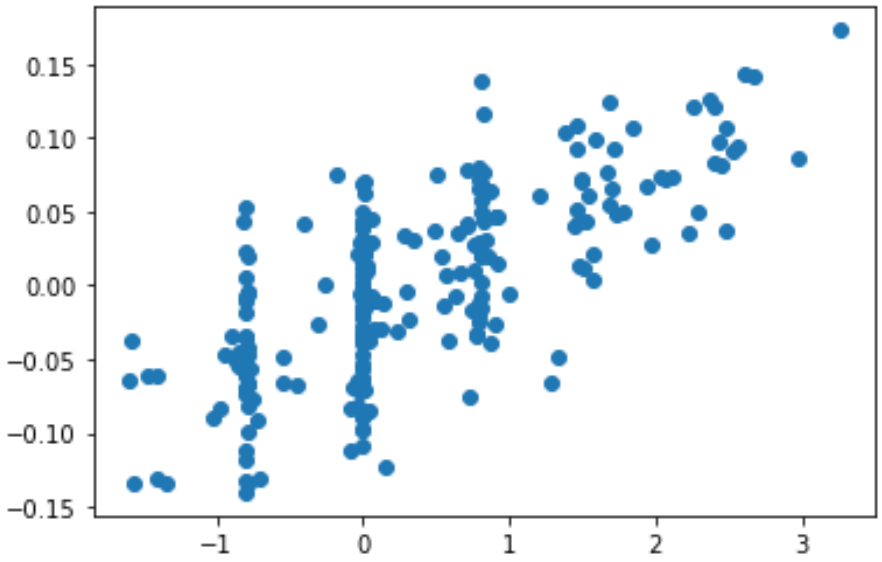

図 6　IRT の能力パラメタ（横軸，pyirt によって算出）と，学習者トークンの単語埋め込みベクトルの第一主成分得点（縦軸）．

語モデルの代表的な手法であり，大量のラベルなしデータからの事前学習 (pre-training) と，ラベル付きデータを用いた微調整 (fine-tuning) という 2 種類の学習からなる．事前学習では，大量のラベルなしコーパスを用いて，当該言語の基本的な構造を学習し，入力文の言語としての自然さを計算可能にする．この過程は計算量が非常に大きいが，様々なタスクに対して汎用的に用いることができる．そこで，通常，事前学習は，**bert-large-cased** 等の，英語版 Wikipedia 等を用いて訓練された **transformers**（`https://github.com/huggingface/transformers/`）の**事前学習済モデル**を用いる．事前学習済モデルの詳細情報，例えば事前学習に用いたコーパスなどの情報は `https://huggingface.co/models` に記載されている．多くのモデルは英語版 Wikipedia を使用している．

後段の微調整 (fine-tuning) では，実際に，目的とするタスクに合わせて，事前学習済モデルを追加訓練する．本研究のタスクにおいては，ラベルは，IRT 同様，学習者が正答する場合 1，誤答する場合を 0 とする 2 値判別問題である．事前学習済モデルに設問文と学習者の両方を入力し，微調整を行いたいが，通常，深層言語モデルの微調整では言語しか入力として扱えないため，学習者の情報を入力することができない．そこで，次節に述べる方法で，この問題を解決する．

6.3　提案：Transformer モデル上の個人化判別

Transformer モデルを個人化判別に対応させる手法は，自然言語処理の言語教育応用の目的では著者の知る限り知られていない．ただし，Transformer モデルに特殊なトークン（語）を加えて微調整を行い，様々な問題設定に対応させる手法は知られており，ライブラリ上で特殊なトークンを加える機能が用意されている．本研究では，この機能を利用することで，学習者に対応するトークン（**学習者トークン**）を作り，これを入力系列の最初に置くことによって判別を行う手法を提案する（図 4）．例えば，学習者 ID が 3 番の学習者を表すトークン “[USR3]” を導入し，“[USR3] It was a difficult period.” が入力であれば，3 番の学習者が “It was a difficult period.” という文から成る設問に正答するか否かを予測する問題に帰着させる．入力文はそのままで，入力文の前に，単純に学習者トークンが挿入されている点に注意されたい．導入するトークン数は学習者数と同数である．Transformer では各

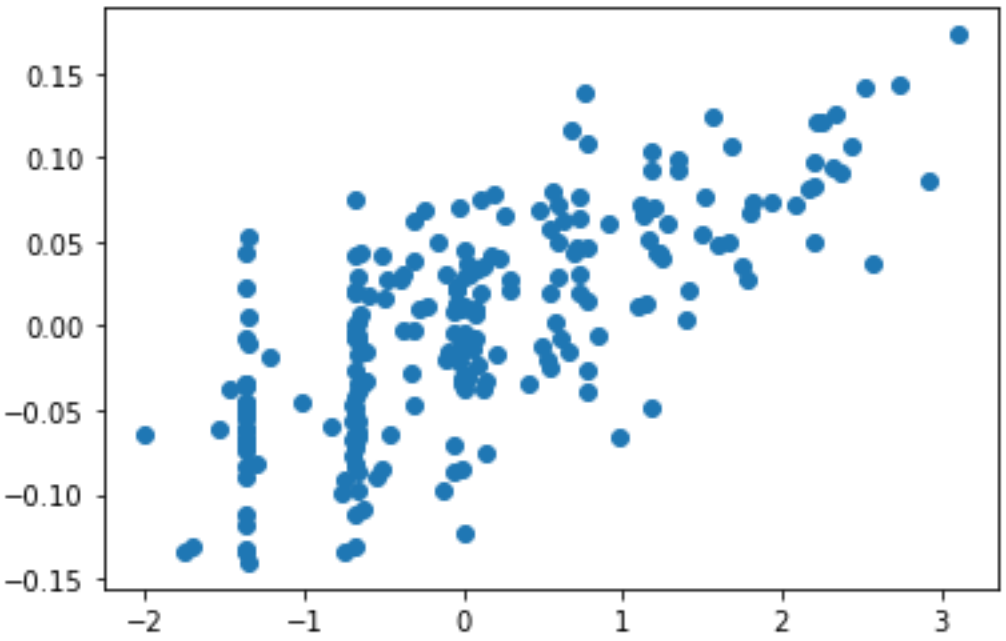

図 7　R 言語の ltm パッケージによって算出した IRT の能力パラメータ（横軸）と学習者トークン埋込ベクトルの第一主成分（縦軸）の関係．

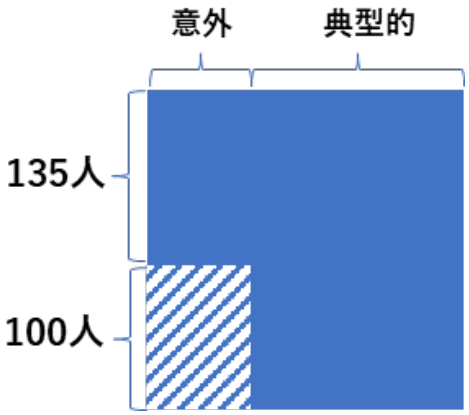

図 5　実験設定．青く塗られた部分がパラメタ推定に使われる訓練データ．斜線部が性能比較に用いられるテストデータ．

トークンに対して，その語としての機能を表現する**単語埋め込みベクトル**があるので，学習者トークンに対しても埋め込みベクトルが作られる．

重要な点として，提案手法では，文中のどの語についての設問であるかという情報や，誤答選択肢の情報は与えていない．すなわち，提案手法の判別器は，表 1 のどの単語に下線が引かれているかや，表 1 や表 2 の正答以外の選択肢の情報を用いない．提案手法は，単純に正解となる文を入力とし，これを学習者が理解できるか否かを判別する判別器を構成している，と解釈できる．これにより，提案手法は，表 1 と表 2 という仔細の異なる 2 種類の多肢選択式の問題に対応できる．このように，提案手法の適用範囲を広くとることができる．今回の設定では，入力文が短文であり，学習者が 1 語でもわからなければ正答できない設問で構成されているため，語義を知っている事と正解となる文を理解できるか否かは，同一視できる．

Transformer モデルのその他の実験設定については多用される設定とした．判別には，**transformers** ライブラリの **AutoModelForSequenceClassification** を用いた．微調整の訓練には Adam 法 [15] を用い，バッチサイズは 32 とした．

Transformer モデルを用いた結果を，表 3 に示す．*は IRT の最高性能と比較して Wilcoxon 検定で統計的有意であることを表し，**は $p < 0.01$，*は $p < 0.05$ を表す．また提案手法の（）内は用いた事前学習済モデル名である．表 3 では，まず，学習者トークンを導入した提案手法が，IRT を用いた従来手法より高い性能を達成していることが分かる．この実験結果は，設問文の意味を考慮する事で，IRT より高精度な判別が行えることを示している．

次に，“roberta-base” は cased（大文字・小文字を区別するモデル）であるのに対し、“albert-base-v2” は uncased（大文字・小文字を区別しないモデル）である．この結果から，良い精度を得るためには “cased”，すなわち，大文字と小文字を区別して扱うモデルでなければならないことが示唆される．この理由は，次のように推察される．この実験環境では，各質問は短い文から構成されているため，モデルは大文字で始まる文の開始を認識する必要があるためであろう．

さらに，表 3 では，**bert-base-cased** が最も高い性能を示した．より大きな事前学習済モデルである **bert-large-cased** よりも **bert-base-cased** が高い性能を示した理由として次のことが考えられる．学習者特性を表す学習者トークンの単語埋め込みベクトルは，今回作成した比較的小さい訓練データで訓練しているため，小さいモデル

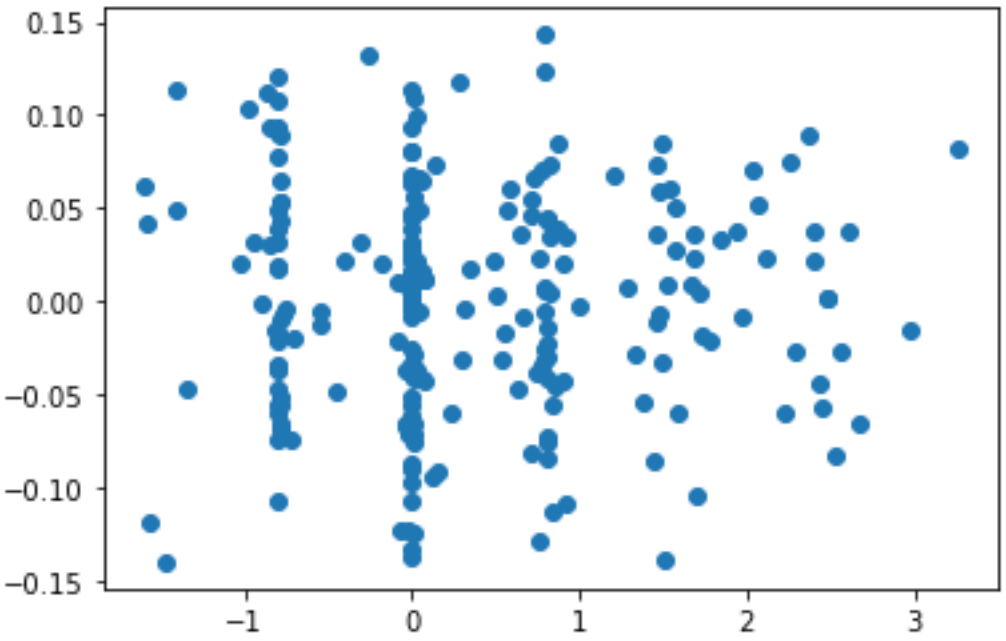

図 8　学習者能力パラメタ（横軸）と学習者トークン埋込ベクトルの第二主成分得点（縦軸）の散布図．

の方が微調整（fine-tuning）に適していた可能性がある．

7　解釈性―学習者トークンからの能力値抽出

IRT は，学習者の能力パラメタを持つことにより，学習者の特性について解釈しやすい．一方，Transformer モデルでは，学習者の特性は学習者トークンに対する単語埋め込みベクトルという多次元の形で表現されており，そのままでは直感的な解釈が難しい．しかし，Transformer モデルは個人化判別問題で高精度を達成しているので，学習者トークンの単語埋め込みベクトルの中に能力値の情報が含まれていると考えられる．

微調整後の **bert-large-cased** の場合の学習者トークンに対する単語埋め込みベクトルのみを集めた．すなわち，学習者の人数分の単語埋め込みベクトルの集合がある．このベクトル集合に対して主成分分析を行い，その第一主成分得点と IRT の能力値パラメタを比較した（図 6）．各点は学習者を表す．IRT の能力値パラメタの算出には，Python の **pyirt** ライブラリを用いた．両者は相関係数 0.72 という強い相関を示した（$p < 0.01$）．これにより，提案手法を用いた場合でも，能力値は学習者トークンの第一主成分得点として容易に抽出できることが分かった．これにより，提案手法は文意を考慮することにより IRT より高い精度を達成しながら，IRT と同様に「能力値を取り出せる」という高い解釈性を持つことが示された．

図 6 では，縦に筋が入っているように見える部分がある．これは，pyirt の内部で使われている IRT のパラメタ推定アルゴリズムの性質で，横軸の学習者の能力値パラメタの推定の際，能力に大きな差がない能力値パラメタは 1 つの値にまとめられる性質があるため，横軸が同じ値を取る学習者が存在するためである．

7.1　IRT パラメタの推定手法に依らない事の確認

IRT の能力値パラメタは，データが同じでも，どの推定用ソフトウェアを用いるかによって，多少の違いが生じることが知られている．前節では，図 6 では pyirt ライブラリを用いたが，この推定用ソフトウェアの特性によって統計的有意性が生じた可能性もある．

そこで，確認のため，同じデータを，pyirt とはプログラミング言語も異なる全く独立の実装である R 言語の **ltm** パッケージを用いて推定した．これは，教育心理学の標準的な教科書で使用されているソフトウェアである [16]．その結果を図 7 に示す．この場合も，目で見て取れる相関があり，実際に相関係数は 0.72 で，やはり統計的有意性を示した（$p < 0.01$）．従って，IRT の能力値パラメタを算出するソフトウェアによらず，学習者トークンから能力値が抽出できることが示された．

7.2　第二主成分分析との相関

学習者トークンの埋め込みに第一主成分得点には能力値が含まれていることが示されたが，第二以降の主成分得点には能力値との相関はあるのだろうか？この疑問について調べるために，第二主成分得点と能力値をプロットしたものが図 8 である．図 8 からは，目に見えて分かる通り，相関がないように見える．実際に，相関係数を計算しても，統計的に有意な相関は得られなかった．このことから，学習者の能力値は，各学習者の学習者トークン埋め込みベクトルの第一主成分得点にのみ保持されていることがわかる．

8　おわりに

本研究では，設問文を考慮しつつ解釈しやすい学習者反応予測を行うため，学習者を表すトークンを導入することで BERT 等の深層言語モデルを学習者反応予測問題に適用する，簡便な手法を提案した．提案手法が設問文

を考慮することによる効果の評価のため，語学学習者が設問文中の単語の意外な意味を知っているかどうかを予測する課題に取り組み，そのための評価データセットも作成した．提案手法の予測性能は，統計的に有意な差をもって IRT の予測性能より優れていることが分かった．また，提案手法では，学習者トークン埋込みの第一主成分得点を用いることで，学習者の能力値を容易に取得できることを示した．この結果は，提案手法の解釈性が高いことを示しており，提案手法が教育利用にも適していることを示している．

今後の課題，特に今後の研究上の発展性について述べる．まず，語彙学習支援の方向性での今後の課題として，語彙テスト結果データでの微調整後の語の単語埋め込みベクトル集合だけに注目することが考えられる．これらのベクトルの中に単語の難しさの情報が埋め込まれており，主成分分析などの方法によって取り出せる可能性はある．日本語母語話者にとっての英単語の難しさについては，代表的な British National Corpus [17] などのコーパス上での単語頻度や，CEFR-J 語彙プロファイル（https://github.com/openlanguageprofiles/olp-en-cefrj）などの様々な指標が提案されており，こうした指標と単語埋め込みベクトルから抽出した主成分得点などの統計量の相関等の調査も今後の課題の 1 つであろう．

より重要な今後の課題の 1 つは，今回対象とした語彙学習支援以外の応用について，「学習者トークン」を用いる提案手法の有用性を検証することである．本稿で用いた BERT は，高度な文と文の間の意味的処理を扱う含意関係認識等でも高い性能を示している（https://gluebenchmark.com/）．従って，本稿で用いたような BERT 等の手法を用いて，設問の難しさを設問文から意味的に推定可能と考えられる設問文については，本稿と同様の手法で，学習者トークンから能力値を抽出可能と予想される．

今回の提案手法では，どの部分に下線が引かれているかや，誤答となる選択肢は入力せず，単純に正解文を学習者が理解できるかを判定する判定器を構成した．これにより，提案手法は幅広い問題設定に適用可能であると考えられる．しかし，教育上は，下線部や誤答となる選択肢の情報をどうしても入力させたい場合もあると考えられるので，これらの情報をどのように深層言語モデルに与えるかは，今後の課題である．

また，教育上重要な今後の課題としては，学習者トークンの単語埋め込みから，学習者の能力パラメタより高度な情報を取り出すことが挙げられる．本研究で導入した学習者トークン埋め込みは多次元である．第一主成分のスコアが受験者の能力パラメータと有意に相関することを示したが，他の成分は学習者の他の種類の能力の情報を保持している可能性がある．IRT においても，学習者の能力を多次元ベクトルとしてモデル化する同様の考え方があり，多次元 IRT と呼ばれている [1]．したがって，学習者トークン埋込みの他の主成分と多次元 IRT を比較することが，この方向性での今後の発展的課題として考えられる．

謝辞

本研究は JST ACT-X 研究費 (JPMJAX2006) の助成を受けたものです．理研 miniRAIDEN，産総研 ABCI を使用しました．データ作成に協力頂いた谷口ジョイ先生他，静岡理工科大学の皆様に感謝します．

参考文献

[1] Frank B. Baker. *Item Response Theory : Parameter Estimation Techniques, Second Edition*. CRC Press, July 2004.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, 2019.

[3] Yo Ehara. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proc. of EDM*, 2022.

[4] Burr Settles. Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), 2018.

[5] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. A report on the complex word identification shared task 2018. In *Proc. of BEA*, June 2018.

[6] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil Heffernan, Xintao Wu, Sean McGrew, and Dongwon Lee. Classifying math knowledge components via task-adaptive pre-trained bert. In *Proc. of AIED*, pp. 408–419. Springer, 2021.

[7] Lele Sha, Mladen Rakovic, Alexander Whitelock-Wainwright, David Carroll, Victoria M Yew, Dragan Gasevic, and Guanliang Chen. Assessing algorithmic fairness in automatic classifiers of educational forum posts. In *Proc. of AIED*, pp. 381–394. Springer, 2021.

[8] Shiting Xu, Guowei Xu, Peilei Jia, Wenbiao Ding, Zhongqin Wu, and Zitao Liu. Automatic task requirements writing evaluation via machine reading comprehension. In *Proc. of AIED*, pp. 446–458. Springer, 2021.

[9] I. Nation. How Large a Vocabulary is Needed For Reading and Listening? *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82, October 2006.

[10] Batia Laufer and Geke C. Ravenhorst-Kalovski. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, Vol. 22, No. 1, pp. 15–30, April 2010.

[11] I. S. P. Nation and Rob Waring. *Teaching Extensive Reading in Another Language*. Routledge, November 2019.

[12] Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.

[13] David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, Vol. 31, No. 7, pp. 9–13, 2007.

[14] 荘島宏二郎, 豊田秀樹. テストが複数の出題形式を含むときの項目母数の推定. 教育心理学研究, Vol. 52, pp. 61–70, 2004.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[16] Insu Paek and Ki Cole. *Using R for item response theory model applications*. Routledge, 2019.

[17] BNC Consortium. *The British National Corpus*. 2007.