統計数理(2024)
第 72 巻 第 1 号 23–41

特集「心理統計学の新展開」
[研究詳解]

項目反応理論に基づく教育のための自然言語処理のモデル

江原 遥†

(受付 2023 年 7 月 4 日；改訂 2024 年 2 月 7 日；採択 2 月 19 日)

要　　旨

教育応用において，学習者の能力を測定したり項目の困難度など項目の特性を計測することは，学習支援システム等に幅広い応用がある教育応用の基礎タスクである．単に学習者が所与の項目に正答するかを予測するモデルではなく，学習者の能力や項目の特性を人間の教員が解釈できれば人間の教員が教育するときにも活用できる．統計分野においては，古くから項目反応理論等を用いて試験の学習者の回答パターンから解釈可能なパラメタを推定するアプローチが取られてきた．一方，項目の大部分は自然言語で記述されている．自然言語を解析する自然言語処理分野では，項目のテキストから困難度等の項目特性を抽出する研究に関心が持たれてきた．特に，単語頻度などの技術的に抽出が容易な特徴量から，項目の困難度の多くを説明可能な値を抽出できる語学学習支援などへの応用ではテキストからの困難度推定などの研究が盛んであった．そこで本稿では，テキストからの困難度の推定が項目反応理論とどのように関わりを持つのかについて，外国語の語彙学習支援や読解支援・可読性判定を中心に，様々な分野の研究を引用しながら説明する．そして近年，テキストの意味を考慮した解析で高精度を達成している自己教師あり学習や Transformer 等の手法を取り上げて詳説する．

キーワード：項目反応理論，自然言語処理，語学学習支援．

1. はじめに

学習者の能力に合わせた教育を行うためには，学習者に適合する教材を提示することが重要と考えられている．この「適合する」は，多くの場合，学習者の能力に合った「困難度(難しさ，難易度，難度)」を持つ教材を提示することに相当する．学習者の能力や教材の困難度を何らかの形で数値化して，適切な能力を持つ学習者に適切な教材を割り当てたい．

学習者の能力や教材の困難度をどのように計測したら良いだろうか？教科ごとに独自の尺度を作成することは比較が難しい．また，試験を通じて学習者の能力を計測するにしても，試験の配点を作問者が自由に決めて良いとすれば，必ずしも適切に能力を計測できているとは限らない．難しい項目(テストの問題や設問)の配点を小さくしすぎているなど，試験結果のデータに沿っていない状況が考えられる．

項目反応理論(item response theory，IRT)(Baker and Kim, 2004)は，試験結果のデータから，学習者の能力値や項目の困難度の両方を推定する手法の一つである．これらのパラメタは項目の自然文による内容等は一切参照せず，学習者集合の各項目に対するパターンだけから計

† 東京学芸大学 教育学部：〒 184–8501 東京都小金井市貫井北町 4–1–1

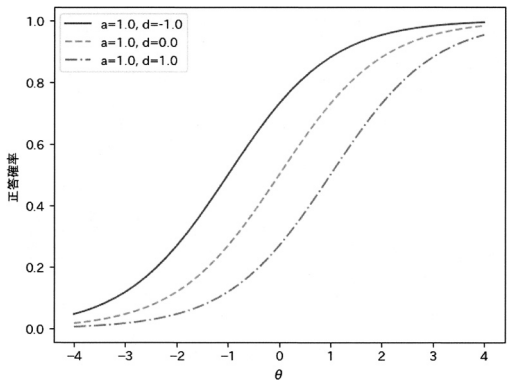

図 1. 困難度パラメタ d の異なる 3 項目に対する項目特性曲線.

測される．適切な項目集合を設定し，適切な等化(尺度調整)を行えば，異なる学習者集合間の能力を比較できることも知られている．

ここでは，IRT のモデルについて簡単に説明する．学習者の数を J，項目(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．

$$P(y_{ij} = 1|i, j) = \sigma(a_i(\theta_j - d_i)) \tag{1.1}$$

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．(1.1)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，項目の困難度を表すパラメタである．(1.1)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．

能力値パラメタ θ に対して，(1.1)で定義される正答確率をプロットした図を「項目特性曲線」という．項目特性曲線を用いると，項目パラメタが正答確率にどの様な影響を及ぼすかを視覚的に理解しやすい．図 1 に，困難度パラメタを $d = -1.0, d = 0.0, d = 1.0$ と変えた 3 つの項目特性曲線を図示した．θ の値が同じでも，困難度の値が大きくなるほど曲線が右に平行移動することにより，学習者の項目に対する正答確率が低くなることがわかる．これは直感的には，困難度パラメタの値が大きい項目には学習者は正答しにくいということを表す．

a_i が大きいと $\theta_j - d_i$ の絶対値が小さくとも，$\theta_j - d_i$ の正負により，正答確率が 1 か 0 のどちらかに偏るようになる．言い換えると，a_i が大きい項目は $\theta_j - d_i$ の正負により，「学習者 j が設問 i に正答するか否か」がはっきりと分かるようになる．これは，設問 i が，a_i の値が大きいほど，能力値が高い(能力値パラメタが d_i 以上の)学習者と，能力値が低い(能力値パラメタが d_i 未満の)学習者を正確に見分けられることを表しているため「識別力」と呼ばれる．識別力の値が大きいことは，基本的には項目が良い性質を持っていることを表し，出題に適した項目として評価される．

図 2 に，識別力パラメタを $a = 0.6, a = 1.2, a = 1.8$ と変えた 3 つの項目特性曲線を図示した．困難度パラメタは $d = 0.5$ で固定した．困難度の値が同じでも，識別力の値が大きくなる

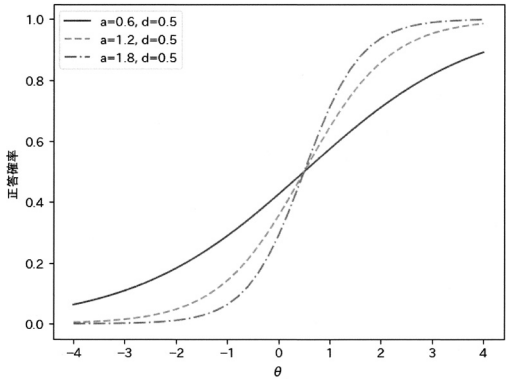

図 2. 識別力パラメタ a の異なる 3 項目に対する項目特性曲線.

ほど，曲線の傾きが大きくなることがわかる．これは，能力値パラメタが $d = 0.5$ 以上か否かで正答確率がそれぞれ 1 か 0 に偏ることを表す．すなわち，この例の中では識別力が最も大きい $a = 1.8$ のとき，学習者の能力が 0.5 以上であるかどうかが，この項目に学習者が正答するか否かではっきりと見分けられる．

(1.1)では，各項目パラメタに対して d_i と a_i の 2 つの項目パラメタがあることが分かる．そのため，これは 2PL モデルや 2 パラメタモデルなどと呼ばれる．全ての項目に対して $a_i = 1$ であるときを，特別に 1PL モデルや Rasch モデルと呼ぶ．

なお，IRT には多肢選択式の項目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 パラメタモデル(3PLM)が存在する．3PLM では，2PLM に比べて項目パラメタの高精度な推定に必要とされるサンプル数(学習者数)が一般に多い．

(1.1)のパラメタ推定の方法について述べる．IRT のパラメタ推定には，周辺化最尤推定(Marginalized Maximum Likelihood Estimation，MMLE)がよく用いられる．この方法では，能力値パラメタ θ_j について，(1.1)に基づいて定義される尤度を(数値的に)周辺化し周辺化尤度関数を最大にする項目パラメタを探索する．この MMLE は R 言語では “ltm” パッケージにガウス・エルミート求積法を使う方法が実装されている．その他，Hanson (2000)を実装した **pyirt** (https://github.com/17zuoye/pyirt)も Python 言語では用いられている．Hanson (2000)も，結局は能力値パラメタを周辺化していることに相当することが示されている (Hsu et al., 1999)．

本稿の構成について述べる．まず語彙学習支援において，テキストからの難度推定について説明する．次に，近年の教育データマイニングなどの手法を項目文(設問文，問題文)からの困難度パラメタ推定に着目して整理する．最後に直近の方法である Ehara (2022b)の「学習者トークン法」について説明する．

2. テキストからの難度推定

前節では IRT について簡単に説明し，特に学習者の正答/誤答の反応パターンから項目の困難度をどのように計算するかについて説明した．その中で項目の内容(項目のテキスト表現)などは一切参照されないことも説明した．しかし実際には，項目の困難度は項目文と密接な関係がある．IRT では，どの様な項目の困難度も，結局は一回学習者を集めて出題してみないとわからない．これは大変に手間であるので，項目文から直接難易度パラメタなどを推定できれば学習者を集めなくとも学習者に個別適合した適切な出題が可能かもしれない．より平たく言え

ば，項目文の中身を見て困難度を推定する課題を解きたい訳である．本節では，こうしたアプローチの先行研究を語学学習支援を中心に説明する．

2.1 応用言語学分野の研究

応用言語学(Applied Linguistics)は，平たく言えば第二言語獲得のための言語学である．第二言語獲得においては，特に，語彙獲得に膨大な労力が必要とされる．英語の場合，大学の授業程度の内容を英語で理解するためには，10,000 語程度かそれ以上の語を学習する必要があると言われている（Nation, 2006）．他の教科と異なり，量が膨大であるために，これだけの知識を学習者が持っているかどうかを網羅的に試験することは難しく，どのように試験すればよいかが研究されてきた．

多くの研究があるが簡単に述べる．まず，学習者の語彙量を計測するテストが開発された．単純に British National Corpus(BNC)（BNC Consortium, 2007)等の収録ジャンルの偏りがなるべく小さくなるように設計された一般コーパス(General Corpus)の単語頻度をまず求める．この単語頻度の降順に，単語頻度順位を用いて語の困難度を表す指標を作る．具体的には，1,000 語の困難度が同程度であること，学習者が単語頻度の多い順に語を記憶していることなど，強い仮定を置いた上で，例えば単語頻度の降順に 1,100 位の単語であれば語の困難度は 1,001 位〜2,000 位の水準であるので，レベル 2，などと考える．こうしてレベルが等しいと仮定した 1,000 語の中から 5 語程度を選定して，多肢選択式のテストを作成する．こうして作成されたテストは Beglar and Nation（2007)などで公開されている．学習者の語彙量を実際の教育と結びつける研究としては既知語率(lexical threshold)と呼ばれる指標が使われ，テキスト中の延べ語数(トークン数)換算で 95% 以上の語を知らなければ，読解力試験の点数が大きく下がることが知られている（Laufer and Ravenhorst-Kalovski, 2010）．

このように応用言語学分野では，独自の基準で語の困難度(一般コーパス中の単語頻度の降順の 1,000 語区切り)を計測する枠組みが作られた．こうして作られた語の困難度の枠組みを，IRT で検証する研究が後から行われた．特に Beglar（2010)は重要であり，試験で計測された語彙量の数値は多肢選択式の試験結果データを Rasch モデル(1PL モデル)にかけて算出される能力値パラメタとよく相関すること，$-\log$（単語頻度）の値が，語彙テストの困難度パラメタの値とよく相関することが報告された．これは英語の語彙テストという限られた設定ではあるものの，問うている項目の一般コーパスの $-\log$（単語頻度）の値を線形変換すれば，実際に語彙テストを実施しなくとも(学習者の反応パターン情報がなくとも)困難度パラメタの値を推定することが可能であることを示した点で重要である．

2.2 自然言語処理分野の語彙学習支援の研究

困難度パラメタの値を項目文のテキストから推定する研究は，Beglar（2010)とは独立に自然言語処理分野でも行われてきた．Ehara（2018)はその 1 つであり，単語テストの困難度パラメタの値を Support Vector Regression を用いて単語頻度から具体的に予測しその値について報告している．この様な，語彙を中心にした簡単な語彙テストの困難度パラメタ推定を，テキストから行う研究を実応用したのが，語学学習アプリとして有名な Duolingo(当時)のグループが行った Settles et al.（2020)の研究である．この研究では，Duolingo の大人数の反応データを用いて，困難度パラメタをテキスト中の様々な特徴量から予測することで，学習者がいないか極めて少ない項目に対してその困難度パラメタを推定し Computer Adaptive Testing(CAT)を行う手法を提案している．CAT は，IRT の困難度パラメタがわかっている場合に個々の学習者の能力パラメタを逐次的に推定する手法の総称であり，様々な手法が提案されている．

2.3 Deep Knowledge Tracing

自然言語処理とは独立に，教育 AI 分野で特に近年注目されている手法が Knowledge Tracing（知識追跡）と呼ばれる手法である．この手法は，IRT とは基本的にタスク設定が異なり，スマートフォン上の教育アプリや Massive Open Online Courses（MOOCs）のように，大人数の学習者が多数の項目に答えたログを用いた分析や予測を目的とする．基本的に，スパースな時系列データであり，学習者の正答/誤答に時間情報が紐づいている点が違いの 1 つである．学習者の能力値の向上をみることや，学習者の正答/誤答の予測精度などに大きな関心がある．ただし，こうしたアプリ上の項目は，通常，作問者などがどの様な知識を問うための項目であるのかをわかりやすくするため，各作問がどの様な能力（スキル）を問うているのか等の情報が与えられる設定が多い．

この Knowledge Tracing の設定で深層学習を用いて，従来手法より高精度を達成した手法が，Deep Knowledge Tracing（Piech et al., 2015）である．これは Long-Short Term Memory（LSTM）（Schmidhuber et al., 1997）を用いる手法である．Piech et al.（2015）は学習者を表すパラメタを陽に持たないモデルであり，学習者個別の予測をどのように行っているのかが直感的に分かりにくい．前述のように教育アプリなどでは，学習者の学習履歴（どの項目にいつ，正答/誤答したか）が残っているため，この回答した項目と正答/誤答のペアの時系列情報を，そのまま，個々の学習者の表現に用いていることが，Piech et al.（2015）のポイントの 1 つである．予測する際は，学習者がある時点までの反応パターンを与えた上で，その次の項目に対する反応を予測する形を取る．従って，例えば，全く同じ項目集合に対して全く同じ正答/誤答のパターンを示した学習者が 2 名いる場合には，Piech et al.（2015）の手法は両者を区別できない．しかし，反応パターンまで完全に一致することは殆ど起こらないので，実質的には個々の学習者に対して個別適応した予測が可能となる．

Piech et al.（2015）では予測精度の向上が中心であり，Knowledge Tracing のような時系列データを対象に IRT の能力値パラメタや困難度パラメタを推定する手法については提案されていなかった．Knowledge Tracing の設定のデータを対象に，学習者の能力値や項目の難易度といった解釈性を持ち，なおかつ高い精度を兼ね備えて持つ手法については，博士論文 Tsutsumi（2023）とその一連の著作が詳しい．

一方 Knowledge Tracing の設定では，項目の中のテキスト表現からテキストの困難度を抽出しようといった試みはあまりなされてこなかった．この理由は，おそらく教育アプリなどの設定では，項目の困難度や項目がどの様な能力を試験しようとしているのか（スキル，と表現されることが多い）といった情報は，作問者がある程度タグ付けした情報が入手できることが多いこと，数学やプログラミングなど論理的な推論を必要とする項目も多いためであろう．すなわち，スキルのタグ情報を有効活用した方が，項目文の本体から困難度パラメタの値を推定するより容易であったためであると推定される．

ただし，Knowledge Tracing の設定でも項目の中の自然文を活用しようという手法はいくつかある．1 つは，Relation-aware Knowledge Tracing（Pandey and Srivastava, 2020）という手法であり，項目の間の意味的近さの計測に自然文を活用している．直近ではプログラミング課題の練習システム（例えば，日本では Aizu Online Judge などが有名である）のログを対象に，回答が正答/誤答の様な 2 値ではなく，プログラムのコードやテキストなどオープンドメインな回答される設定で，Knowledge Tracing を行う手法が提案されている（Liu et al., 2022）．ここでも，結局，項目の項目を表現する自然文だけから困難度を算出するのではなく，項目文と学習者が書いたプログラムコードの情報を合わせて利用することで，項目の困難度情報がモデル中に暗に表現されているものと思われる．

その他 Knowledge Tracing については，近年では Abdelrahman et al.（2023）という良質な

サーベイ論文が発表されている．

2.4 外国語学習支援の学習者反応データセット

語彙テスト結果に関して，学習者反応を記録したデータセットにはどのようなものがあるのかを，ここで整理する．項目文を考慮した学習者反応予測を行いたい．そのためには項目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．本節では，こうした語彙テスト結果データセットの先行研究を紹介する．ほぼ全ての公開されている英語学習者の語彙テスト結果データセットは，典型的な意味の語義についてのデータセットである．1 つは語学学習アプリ Duolingo 上の項目に対する回答データを用いた SLAM データセット（Settles, 2018）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（Yimam et al., 2018）である．

これらのデータセットの特徴として，各学習者は多くある項目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，項目を列とし，学習者の項目に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．IRT は，学習者の項目に対する回答内容から項目の困難度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なるべく多くの項目に回答している形式のデータセットが望ましい．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，項目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，Yimam et al.（2018）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．

IRT が適用しやすい多肢選択式の語彙テストを学習者に受験させたデータセットとしては，Beglar and Nation（2007）によるテストをクラウドソーシングで 100 人に受験させたデータセットが公開されている（Ehara, 2018）．語彙テスト結果の重要なデータセットとして，他に，個々の学習者に対して網羅的に語を知っているかどうかを自己申告式で回答してもらったデータセットがある．例えば，約 12,000 語を 15 人に対して自己申告式で回答させたデータセットが公開されている（Ehara et al., 2013）．また，最近，非公開ではあるが，NTT のグループが同種のデータセットを作成している例がある（藤田 他, 2023）．

2.5 可読性推定の研究

ここまで，項目文からの困難度推定の研究を語彙学習支援に関連する研究を中心に見てきたが，実は読みやすさ（可読性，リーダビリティ）をテキストから推定する研究も自然言語処理分野でほぼ独立に行われている．自然言語処理は人工知能の一分野という立ち位置もあり，人工知能は人間の知能を模倣するというのが典型的な研究の方向性であるため，「語学教師の読解教材に対する困難度の判断」を模倣させることで可読性を予測しようという一連の研究がある．つまり，こうした研究における「困難度」とは，語学教師などの人間の判断を記録したものであって，IRT の様な学習者の反応から求められるものではないという立ち位置である．このため，IRT を用いた語彙テストの研究と自然言語処理における可読性推定の研究は，ほぼ独立に発展してきたといってよい．このように，「困難度」という類似の概念を用いていても考え方に大きな違いがあるので，注意が必要である．

外国語学習における可読性推定の最近の研究としては，Vajjala and Lučić（2018）が複数の語

学教師に各テキストを実際に読解させて 3 段階でラベル付けした信頼性の高い可読性コーパスの研究がある．また，Martinc et al.（2021）の研究や Ehara（2022a）の研究によって，本稿でも用いている Bidirectional Encoder Representations from Transformers（BERT）（Devlin et al., 2019）等の近年の深層学習手法を用いた可読性推定の研究も行われている．

3. 学習者トークン法（Ehara, 2022b）の紹介

学習者トークン法は Ehara（2022b）が提案した方法である．項目文の文意を考慮して深層学習により学習者反応を予測し，更に，能力値や項目の困難度の抽出などもある程度可能な最近の手法であるので，以後，これについて詳解する．深層学習のアプローチで計算に多くの時間がかかる事前学習済モデルをそのまま用いて，微調整（fine-tuning）を行えるようにすることで，事前学習済モデルを有効活用できる点などに特徴がある．

前述のように，IRT（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（項目）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，自己教師あり学習の枠組みを用いた大規模言語モデルが自然文理解で高い性能を示している（Devlin et al., 2019）．従って，項目文の理解に，これらの大規模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，大規模言語モデルを用いた項目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．

Ehara（2022b）は，大規模言語モデルを項目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論文である．項目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．Ehara（2022b）の手法は，大規模な母語話者コーパスを事前学習に用いることで項目文の文脈を考慮することができる Transformer モデルの手法（BERT（Devlin et al., 2019）など）に基づく手法である．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．Ehara（2022b）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示している．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．この研究で作成したデータセットは，今後著者の Web サイト（http://yoehara.com/）で公開を予定している．

4. 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（Ehara, 2018）．データセットはクラウドソーシングサービス Lancers（https://lancers.co.jp/）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 名の学習者から回答があった．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であると思われる．

まず，通常の語彙テストとしては Ehara（2018）と同様に，Vocabulary Size Test（VST）（Beglar

表 1. 実際の項目例.

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

表 2. 意外な意味を問う項目例.

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

and Nation, 2007)を用いた．ただし VST は 100 問からなるのに対して，Ehara (2018)では，低頻度語に関する項目では 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．この項目例を表 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，文法的情報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．

一方，学習者にとって意外であると思われる用例については，英語非母語話者 1 名が作問し，英語母語話者を含む静岡理工科大学の教員複数名に項目が英語の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 つの項目があること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，Ehara (2022b) では，次の 2 つの工夫を行った．

(1) 意外な語義を問う項目については，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項目とした．これにより，意外な語義については正答を知らなければ，どの語についての項目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項目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項目群を最初に行い，通常の語義についての項目群に移動したら，意外な語義についての項目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項目例が表 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 のような項目を 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 の形式の項目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，通常の語義の項目群の側に対応する項目がない項目を 1 問設けた．これにより，対応する項目は 12 問となる．

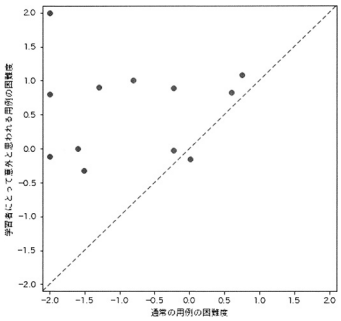

図 3. 各語の，通常の用例の困難度(横軸)と学習者にとって意外と思われる用例の困難度(縦軸)のプロット．各点は各語を表す．

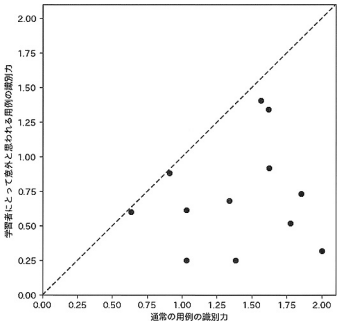

図 4. 各語の，通常の用例の識別力(横軸)と学習者にとって意外と思われる用例の識別力(縦軸)のプロット．各点は各語を表す．

4.1 作成したデータの IRT を用いた分析

前述のように，作成したデータセットでは，同じ語でも項目文の文意によって困難度が異なると思われる．本節では，まず IRT を用いて作成したデータセットを分析することで，実際に同じ語でも項目文の文意によって困難度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．

IRT 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，MMLE により IRT 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，項目のペアが 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 3 に示した．各点は語を表す．

困難度の比較．図 3 の左下から右上まで，点線で対角線を示した．図 3 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回答が難しいと判定された語ということになる．今回は項目数が少ないので，図 3 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，縦軸の項目群の方が横軸の項目群より難しかったことが示唆される．

識別力の比較．識別力についても，図 3 と同様にプロットし，図 4 に示した．識別力は，直

観的には，高いほど，その項目で(他の項目で推定される)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($p < 0.01$)．

5. 学習者反応予測による評価

5.1 IRT による学習者反応予測

語の意外と思われる語義の困難度を典型的な語義の困難度で代替してしまうと，学習者が項目に正答/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項目群(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，典型的な語義の項目群(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 $\times$12 問，計 1,200 件の回答データは用いていないことに注意されたい．(1.1)より，推定された学習者の能力値 θ_j，語義の困難度 d_i を用い，$\theta_j > d_i$ であれば学習者 j が項目 i に正答，そうでなければ誤答と判定できる．項目 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(accuracy)を比較した．予測精度の結果を表 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，語の意外な用例の困難度を語の典型的な用例の困難度で置き換えると，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．

5.2 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，項目文の意味などは全く考慮されていない．では，項目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？大規模言語モデルのうち，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，BERT (Devlin et al., 2019)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大規模言語モデルの代表的な手法であり，

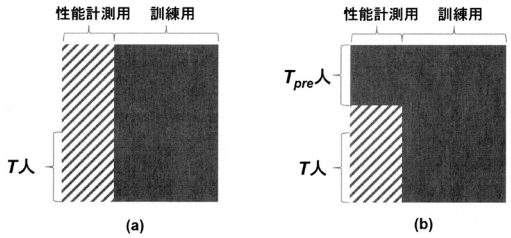

図 5. 実験設定．塗りつぶされた部分がパラメタ推定に使われる訓練データ．斜線部が性能計測に用いられるテストデータ．

表 3. 図 5 で，Ehara（2022b）のデータの「意外な語義」を性能計測用の項目セットに選び，$T_{pre} = 135$ の時の，斜線部のうち T 人の予測精度（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
Ehara（2022b）(bert-large-cased)	0.674 (**)
Ehara（2022b）(bert-base-cased)	**0.688** (**)
Ehara（2022b）(bert-base-uncased)	0.655
Ehara（2022b）(roberta-base)	0.681 (**)
Ehara（2022b）(albert-base-cased)	0.671 (*)

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．Ehara（2022b）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項目文と学習者の両方を入力し，微調整を行いたいが，通常大規模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．

5.3 Ehara（2022b）の手法

Ehara（2022b）の手法を説明する．適応的学習者反応予測のタスク設定を図 5 に示す．テスト結果データは，学習者を行，項目を列とし学習者の項目に対する反応を要素とする．図 5 では，塗りつぶされた部分が訓練データ，斜線部が性能計測用のテストデータとなる．図 5(a) の設定では，完全に新しい項目に対する性能計測用の T 人の反応（正答するか/誤答するか）を高精度に予測することが目的である．(a) の設定では，性能計測用データ中の項目は訓練データにない全く新しい項目なので，項目文の情報を一切用いない IRT 等の手法では項目のパラメタ推定が行えず学習者反応予測ができない．一般には，項目を少人数の学習者に事前に受験してもらい項目の特性を事前に確かめるケースもある．このケースを含み (a) を一般化した設定を図 5(b) に示した．「少人数の学習者」の人数を T_{pre} とした．(a) は，(b) で $T_{pre} = 0$ の場合とみなせる．

大規模言語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機械翻訳などの他タスクで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．Ehara（2022b）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した（図 6）．Ehara（2022b）では “It was a difficult period.” のような項目文中の語の語義を多肢選択式で問うデータセットを用いているため，図 6 もこの例にならっているが，語学学習

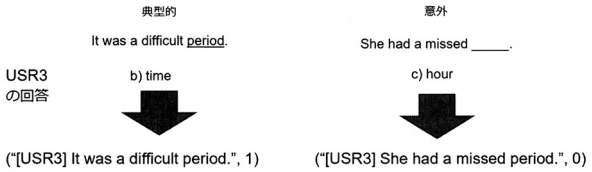

図 6. 学習者トークンの導入（Ehara, 2022b）.

支援に限らず，一般に項目文で示される項目を，指定した学習者が正答できるかを予測する識別器を構成している．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.” という文から成る項目に，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，入力文の直前に単純に学習者トークンを挿入する．導入するトークン数は学習者数と同数である．

重要な点として，Ehara（2022b）では，文中のどの語についての項目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，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．単純に項目文に学習者が正答できるか予測（識別）する予測器を構成していると解釈できる．これにより，Ehara（2022b）は，項目がテキストで構成されてさえいれば適用可能である．

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（Kingma and Ba, 2015）を用い，バッチサイズは 32 とした．

6. 実験結果

BERT 他，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，項目文の意味を考慮することで 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項目の項目文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，小さいモデルの方が微調整（fine-tuning）に適していた可能性がある．

また，Ehara（2022b）では，訓練データにない全く新しい項目に対する学習者反応予測を行う設定，すなわち $T_{pre} = 0$ の設定では実験していない．$T_{pre} = 0$ の設定で，本稿で新しく実験した結果を表 4，表 5 に示す．この設定では，学習者反応予測を行う「性能計測用」の項目につ

表 4. 図 5 で，Ehara (2022b) のデータの「意外な語義」を性能計測用の項目セットに選び，$T_{pre} = 0$ の時の，斜線部のうち T 人の予測精度 (accuracy)．すなわち，全く新しい，訓練データとも異質な項目に対する学習者反応を項目文を考慮することで予測する，最も過酷な設定での精度．

手法	精度
Ehara (2022b) (bert-base-cased)	0.601
Ehara (2022b) (bert-large-cased)	0.585
Ehara (2022b) (roberta-base)	**0.613**
Ehara (2022b) (roberta-large)	0.601

表 5. 図 5 で，Ehara (2022b) のデータの「典型的な語義」を性能計測用の項目セットに選び，$T_{pre} = 0$ の時の，斜線部のうち T 人の予測精度 (accuracy)．すなわち，全く新しいが，訓練データとは同質な項目に対する学習者反応を，項目文を考慮することで予測する，表 4 よりは過酷ではない設定での精度．

手法	精度
Ehara (2022b) (bert-base-cased)	0.632
Ehara (2022b) (roberta-base)	0.613
Ehara (2022b) (roberta-large)	**0.662**

いては回答した学習者がいないため，IRT では項目の項目パラメタを求められず性能を比較できないため，IRT を表から除外した．

表 4 は，表 3 とは $T_{pre} = 0$ であること以外は同一の設定である．ただし，Ehara (2022b) は，外国語の語彙テストを題材に，訓練データに典型的な語義の項目，性能計測用のテストデータに意外な語義の項目を置くというように，訓練データとテストデータの性質が違うものを計測している．

この設定は一般的とは言い難いので，より一般的に訓練データも性能計測用のテストデータの方も典型的な語義を問う同質の項目 (ただし，もちろん項目集合自体は完全に disjoint である) として，計測しなおしたものが表 5 である．

データが小規模であるためか，どのデータセットでも **bert-base-cased** で比較的高い精度が達成できていること，最も過酷な設定である表 4 では表 3 より全体的に精度が大きく低くなっていること，表 5 は，表 4 ほどには過酷ではない設定のため，全体的により高い精度が達成されていることがわかる．

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 7)．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．

7.1 IRT パラメタの推定手法に依らないことの確認

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 7 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(https://cran.r-project.org/web/packages/ltm/index.html)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（Paek and Cole, 2019）．その結果を図 8 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．

pyirt も **ltm** も，MMLE を用いていると記されている．ただし，同じ MMLE でも，周辺化する際の数値積分の方法が異なっており，**ltm** ではデフォルトでは 15 点からなるガウス=エルミート求積を使用していることが，ドキュメント(**ltm** の Reference manual である ltm.pdf)に記載されている．一方, **pyirt** の側では Hanson (2000)を使用していると記載されているだけで, 求積方法についてはドキュメント(https://github.com/17zuoye/pyirt/wiki/Model-Specification)に書かれていない．Hanson (2000)は，学習者の能力値パラメタを連続変数とみなしても実装

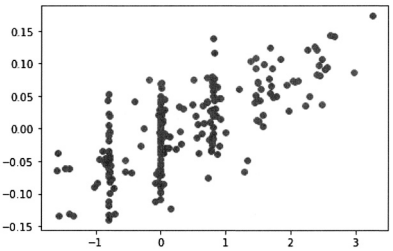

図 7. IRT の能力パラメタ(横軸，**pyirt** によって算出)と，学習者トークンの単語埋め込みベクトルの第一主成分得点(縦軸)．

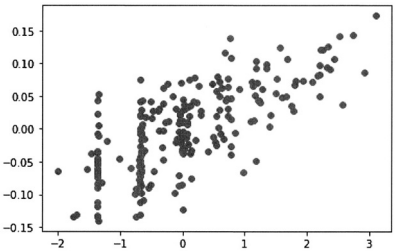

図 8. R 言語の **ltm** パッケージによって算出した IRT の能力パラメタ(横軸)と学習者トークン埋込ベクトルの第一主成分(縦軸)の関係．

上は有限の点で数値積分することになるため，学習者の能力値パラメタがはじめから決められた離散値しか取らないと仮定することで MMLE と EM アルゴリズムの関係性を整理し直した未出版のノートである．Hanson（2000）の手法は，記法が異なるだけで手法としては MMLE と同一であることが，その中で述べられている．Hanson（2000）の記述も数値積分の手法によらない記法となっているため，Hanson（2000）の手法を使用している事だけからでは **pyirt** がどのような数値積分を内部で行っているのかはわからない．そこで，筆者が **pyirt** のソースコードを読んで確認した限りでは，**pyirt** はデフォルトでは能力値パラメタについて $[-4, 4]$ の区間を 10 等分して離散値として扱い，単純な区分求積法を使っているようである．つまり，積分点は単純に区間の等分であって，多くの学習者が該当する能力値の区間でも細かく積分点が取られることはない．図 7 の縦の筋は，この荒い数値積分のためではないかと推察される．また **ltm** では，学習者の能力値パラメタの算出は Expected A Posteriori（EAP）で算出されている．**pyirt** ではドキュメンテーションに能力値パラメタ計算法についての明確な記載が無いが，著者がソースコードを読んだ限りでは，こちらも EAP で能力値を算出しているようである．

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，第二主成分得点についても能力値との相関係数を計算したが，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. 予測確率を用いた困難度の抽出

Ehara（2022b）の手法では，マスク言語モデルである BERT 本体には，困難度を直接表現するパラメタは存在しない．また，パラメタも膨大であるため，fine-tune されたモデルから直接困難度パラメタを抽出することは難しい．しかし，このモデルは学習者個別の予測が可能である．そこで，同じ T 人の集合に対する各項目の正答確率を予測させた上で，その確率の平均値と T 人をかけあわせて，平均で何人正答者がいると予測されているかを計算することで，困難度を表現する値自体は抽出可能と考えられる．実際に図 9 に，これを行った結果を示した．縦軸は，前述の方法による予測正答者数である．IRT を用いて図 5 の全てのデータが見えている状態で推定した項目の困難度パラメタの値が横軸である．相関係数は，図 9 では 0.78（$p < 0.01$）であり，BERT モデルから IRT の困難度と統計的有意に相関する困難度パラメタの推定値を取り出せていることが分かる．

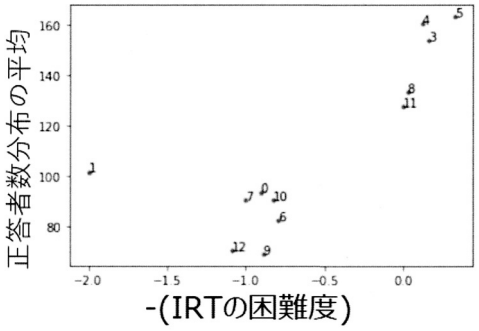

図 9. 予測正答者数と −(IRT の困難度).

9. 議論

表 1 と表 2 はどちらも多肢選択式であるが，表 1 が下線部の同義表現を選択肢から探させる項目，表 2 が空欄補充形式になっていることの影響について議論する．まず，「空欄補充」(fill-in-the-blank)といった場合，空欄を埋める文字列を学習者自身が思いつかないといけない形式の項目を意味する事も多くある．この意味での空欄補充であれば，「一般論としては空欄補充のほうが難易度が高い」といえる．

一方，今回の設問は，空欄に当てはまるものの選択肢が与えられているため，あくまで多肢選択式の出題方法の 1 つとして，項目文中の下線部の同義表現を探す形式(下線部形式)と空欄補充形式との間で困難度に差があるかどうかが議論の焦点になる．結論としては，著者の知る限り，多肢選択式の項目において項目文の下線部形式と空欄補充形式の差が，他の要因に比べて大きく困難度に影響することを示した既存研究はない．むしろ，下線部形式か空欄補充形式かの差よりも他の要因の方が困難度に大きく影響することが示されている．熊澤 (2010)では，言語テストにおける，多肢選択式の出題方法の 1 つとしての空欄補充形式と他の形式の困難度等の比較を，英語学習者 608 名を対象として行っている．熊澤 (2010)では，下線部形式を扱ってはいないものの，多肢選択式の空欄補充形式として，本稿のような 1 文中の空欄補充と複数文からなる短文会話の中の空欄を補充する「クローズ」という形式の 2 つが比較されている．多相ラッシュモデルを扱うソフトウェアである FACETS を用いて分析したところ，通常の空欄補充の困難度が −0.10 であり，クローズの空欄補充の困難度が 0.36 と報告されており，同じ空欄補充形式でも大きな違いが出ることが示されている．熊澤 (2010)の分類では，下線部形式と同様，「最初に提示される選択肢ではないテキストに空欄がなく」，「選択肢として英文が並べられ」，「1 つだけ正しい英文を選ぶ」という形式の項目は，提示された和文と同じ意味の英文を選択肢から選ぶ「和文英訳」しかないが，この形式の項目の困難度は −0.08 と，通常の空欄補充形式ともっとも近い値を取っている．

また，多肢選択式の分類では，下線部形式と空欄補充形式の様な項目文による分類の他に，選択肢による分類もあり得る．池上 (2015)の分類では，本稿の下線部形式も空欄補充形式も，どちらも「通常の択一式」に分類される．一方，「(他の選択肢に)正答なし」が選択肢に入る「「正答なし」を含む択一式」や，複数の選択肢が正答になり得る「複数選択式」があり，この順番に困難度が高くなることが報告されている．

以上のように，下線部形式か空欄補充形式か以外の要因の方が困難度に影響を及ぼすことが報告されており，本稿では，空欄の有無の違いはあるとはいえ，どちらも択一式の形式の項目を扱っていることから，空欄を用いる形式以外の要因が困難度に影響していると考えることが妥当であると思われる．すなわち，最初に設定した，通常の意味に関する項目か，意外な意味に関する項目であるかの差異が困難度により大きく影響していると推察される．

10. おわりに

本稿では，外国語の語彙学習支援を中心に，IRT に注目して，特に項目のテキスト表現から項目の困難度パラメタを推定する手法についてレビューを行った．そして，直近の BERT などのマスク言語モデルの事前学習済みモデルをそのまま活用して，能力値や困難度を抽出することの可能な手法を Ehara (2022b)に基づき，詳細に説明した．

今後のこの分野の研究の方向性について述べる．本稿では，語彙学習支援を中心に述べ，実験もこれに関して記した．しかし，もちろん，教育応用は語彙学習支援だけではない．大規模言語モデル(Large Language Models, LLMs)の性能が近年，急速に向上しており，特に ChatGPT (https://chat.openai.com/)等の大規模言語モデルを用いれば，項目文を解釈して，項目の困難

度を推定することはかなり可能であると思われる．実際，ChatGPT を用いて，外国語学習の項目で，困難度を数値指定できることについては，著者はある程度確かめている．外国語学習以外の応用に広げていることが今後の課題となろう．また，困難度が関心の中心であったが，大規模言語モデルから識別力の値を抽出する手法については，著者の知る限りほとんど報告がない．これも，著者はある程度予備実験を行っているが，今後のおもしろい研究課題になると考えられる．

参 考 文 献

Abdelrahman, G., Wang, Q. and Nunes, B. (2023). Knowledge tracing: A survey, *ACM Computing Surveys*, **55**(11), 1–37.

Baker, F. B. and Kim, S.-H. (2004). *Item Response Theory*: *Parameter Estimation Techniques*, CRC Press, Boca Raton, Florida, USA.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test, *Language Testing*, **27**(1), 101–118, https://doi.org/10.1177/0265532209340194.

Beglar, D. and Nation, P. (2007). A vocabulary size test, *The Language Teacher*, **31**(7), 9–13.

BNC Consortium (2007). The British National Corpus, XML Edition, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554（最終アクセス日 2024 年 2 月 29 日）.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 4171–4186.

Ehara, Y. (2018). Building an English vocabulary knowledge dataset of Japanese English-as-a-second-language learners using crowdsourcing, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 484–488.

Ehara, Y. (2022a). Analyzing readability of scientific abstracts for ESL learners, *Companion Proceedings 12th International Conference on Learning Analytics & Knowledge* (*LAK22*), 86–88.

Ehara, Y. (2022b). No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings, *Proceedings of Educational Data Mining* (*short paper*), 50.

Ehara, Y., Shimizu, N., Ninomiya, T. and Nakagawa, H. (2013). Personalized reading support for second-language web documents, *ACM Transactions on Intelligent Systems and Technology*, **4**(2), 31:1–31:19, http://doi.acm.org/10.1145/2438653.2438666.

藤田早苗, 小林哲生, 服部正嗣 (2023). 日本語母語話者を対象とした英語の語彙数調査, Technical Report, No. 3, NTT コミュニケーション科学基礎研究所, 京都.

Hanson, B. (2000). IRT Parameter Estimation using the EM Algorithm, http://www.openirt.com/b-a-h/papers/note9801.pdf（最終アクセス日：2024 年 2 月 29 日）.

Hsu, Y., Ackerman, T. A. and Fan, M. (1999). The relationship between the Bock-Aitkin procedure and the EM algorithm for IRT model estimation, *ACT Research Report Series*, **99**(7), 1–35.

池上真人 (2015). 択一式と複数選択式の多肢選択文法問題の比較研究, 四国英語教育学会紀要, **35**, 15–24, https://doi.org/10.32276/seles.35.0_15.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization, *Proceedings of the 3rd International Conference on Learning Representations*, https://doi.org/10.48550/arXiv.1412.6980.

熊澤孝昭 (2010). 多肢選択式項目の項目形式が文法テストパフォーマンスに与える影響について, *JALT Journal*, **32**(2), 169–184.

Laufer, B. and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension, *Reading in a Foreign Language*, **22**(1), 15–30, https://eric.ed.gov/?id=EJ887873.

Liu, N., Wang, Z., Baraniuk, R. and Lan, A. (2022). Open-ended knowledge tracing for computer science education, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3849–3862, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, https://aclanthology.org/2022.emnlp-main.254.

Martinc, M., Pollak, S. and Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability, *Computational Linguistics*, **47**(1), 141–179.

Nation, I. (2006). How large a vocabulary is needed for reading and listening?, *Canadian Modern Language Review*, **63**(1), 59–82.

Paek, I. and Cole, K. (2019). *Using R for Item Response Theory Model Applications*, Routledge, Oxfordshire, UK.

Pandey, S. and Srivastava, J. (2020). RKT: relation-aware self-attention for knowledge tracing, *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1205–1214.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J. and Sohl-Dickstein, J. (2015). Deep knowledge tracing, *Advances in Neural Information Processing Systems*, **28**, 505–513.

Schmidhuber, J., Hochreiter, S. et al. (1997). Long short-term memory, *Neural Computation*, **9**(8), 1735–1780.

Settles, B. (2018). Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), Harvard Dataverse, V4, https://doi.org/10.7910/DVN/8SWHNO.

Settles, B., LaFlair, G. T. and Hagiwara, M. (2020). Machine learning–driven language assessment, *Transactions of the Association for Computational Linguistics*, **8**, 247–263, https://aclanthology.org/2020.tacl-1.17.

Tsutsumi, E. (2023). Item response theory based on deep learning with independent student and item networks, Ph.D. Thesis, Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo.

Vajjala, S. and Lučić, I. (2018). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 297–304, Association for Computational Linguistics, New Orleans, Louisiana, https://www.aclweb.org/anthology/W18-0535.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A. and Zampieri, M. (2018). A Report on the complex word identification shared task 2018, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78, Association for Computational Linguistics, New Orleans, Louisiana, https://doi.org/10.18653/v1/W18-0507.

Interpretable Natural Language Processing Models for Educational Applications

Yo Ehara

Faculty of Education, Tokyo Gakugei University

In education applications, measuring the ability of a learner or the characteristics of a question, such as the difficulty of a question, is a fundamental task that has broad utility in learning support systems and other applications. If human teachers can interpret the abilities of examinees and the characteristics of questions rather than simply using models to predict whether subjects will answer a given question correctly, the abilities and characteristics can also be used by human teachers when teaching. In statistics, item response theory (IRT) has been used to estimate interpretable parameters from the response patterns of test takers although IRT does not use the natural language text of each question. By contrast, in natural language processing, there has been interest in research to extract item characteristics such as difficulty from the text of questions. In particular, research on difficulty estimation from text has been active in applications such as language learning support, where values that can explain much of the difficulty of a question can be extracted from easy-to-extract features, such as word frequency. In the present paper, we explain how difficulty estimation from text is related to IRT, introducing research in various fields in addition to statistics, with particular focus on learning vocabulary and reading in second languages. We then discuss recent neural methods such as self-supervised learning and Transformers, which have achieved high accuracy in recent years in analyses that consider the meaning of text.

Key words: Item response theory, natural language processing, second language acquisition.