生成AIに難しさ等を指示するための予測正答者数分布で設問の性質を表現できる設問難度推定法

江原 遥[1,a)]

概要：指定された指示文に従って設問等を生成できる生成 AI が大きな注目を集めている．生成 AI に難しさなどを指定して設問を生成する事を考えると，特定の分野によらない難しさの尺度の指定方法があれば便利である．分野非依存の尺度としては，項目反応理論の困難度や識別力といった尺度があるが，指示する側の人間の側が項目反応理論などの統計尺度に詳しい必要がある．本研究では，より人間にも生成 AI にも理解しやすい尺度として，正答者数の分布を用いる手法を提案する．具体的には，「ある 100 人の受験者集団では，次の設問は予測正答者数分布の平均が a 人，標準偏差 b と予想されています．予測正答者数分布の平均が c 人の全く新しい設問を生成してください」といった指示を行うことで，指示する側の人間が統計尺度に詳しくなくとも，生成 AI に，目的とする設問の性質を伝えられるようにする．本研究では，試験結果データと設問文を訓練データとして，設問文を考慮しながらマスク言語モデルを用いてテストデータ中の正答者数分布を推定する手法を提案する．この分布を，ポアソン二項分布として動的計画法を用いて推定する．英語の語彙試験データを用いた実験により，提案手法で，予測正答者数分布の平均・標準偏差を用いて，設問を ChatGPT に指示し，提案手法で，意図する設問の生成が行われることを定性的に示す．

Question Difficulty Estimation and Large Language Models

Keywords: Question Difficulty Estimation, Large Language Models

1. はじめに

2022 年 11 月 30 日に OpenAI によって発表された ChatGPT(`https://chat.openai.com/`) に代表される生成 AI が社会的に大きな注目を集めている．生成 AI では，プロンプトと呼ばれる指示文に従って，様々なテキストを生成できる．そこで，当然，生成 AI を用いて設問を生成することが考えられる．このとき，生成する設問の難しさを生成 AI に指示したい場合，どのような尺度を用いて生成 AI に指示すればいいだろうか？設問の難しさについては，例えば，その設問で問うている事項を教える学年などで指定する方法もあるが，学校教育で，どのような内容をどの学年で教えるかは，年代や学校の種類によって変わりうる．そのため，生成 AI が正しく指定された難しさを解釈できるとは限らないうえ，指示する側にとっても，学校教育の課程とは直接関係しない設問を生成しようとしている場合，指示することが難しい．学習者の能力については，個々の分野では，英語の TOEIC のように分野独自のテストの点数で表現される場合もある．しかし，学習者のテストの点数は，テスト中の設問の難しさを表すものではない．また，分野ごとに細かく尺度が分かれていると，生成 AI に指示する人間にも個々の分野の尺度に習熟する必要が生じてしまうので，個々の分野によらない設問の難しさ等の尺度を用いることが望ましい．こうした設問の難しさ等の性質は，教育心理学や統計では，項目反応理論 (Item Response Theory，以下 IRT) [1] という，設問の分野によらず，設問の難しさ等の尺度を数値で表現する手法が広く使われている．IRT では，こうした設問の特性を，学習者が項目に回答した履歴のデータから算出する．しかし，生成 AI での指示に IRT で用いられる「困難度」や「識別力」といった尺度を用いるとすると，指示する側の人間が，やはり，IRT のこうした尺度に習熟していなければならず，誰でも直感的に生成 AI への指示に利用できる尺度とは言

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

い難い．そこで，本研究では，より人間にも生成 AI にも理解しやすい尺度として，正答者数の分布を用いる手法を提案する．具体的には，「ある 100 人の受験者集団では，次の設問は予測正答者数分布の平均が a 人，標準偏差 b と予想されています．予測正答者数分布の平均が c 人の全く新しい設問を生成してください」といった指示を行うことで，指示する側の人間が IRT の尺度に詳しくなくとも，生成 AI に，目的とする設問の性質を伝えられるようにする．平均や標準偏差といった概念は高校の数学 I の学習内容であるので，指示する側の人間が数学 I を履修していれば生成 AI に指示が行えると期待される．

本研究では，まず，大規模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手法 [6] を説明する（5.3 節）．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．[6] は，この点を改良し，Bidirectional Encoder Representations from Transformers, BERT [3] などのマスク言語モデルを用いて，「どの学習者が，どの設問文からなる設問に，どの程度の確率で正答できるか」を出力する手法である．そして，この手法による個々の学習者の予測正答確率から，「ある学習者集団が個々の設問を解いた場合の，予測正答者数分布」を求める設問難度推定手法を提案する．これにより，前述の予測正答者分布の平均値や標準偏差といった値を算出することができる．そして，英語の語彙テストを題材に，予測正答者分布の平均値や標準偏差を用いて，生成 AI に設問の難しさ等を数値で指示することで生成する実験を行い，目的の生成が行えているか定性的に評価する．

2. 関連研究

2.1 教育学習支援の関連研究

近年にも，BERT を用いた教育応用が提案されているが [13], [14], [15]，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].

2.2 外国語学習支援の学習者反応データセット

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 [6] を用いて，提案手法を評価した．本稿では，理解を容易にするため，英語の語彙テストに限定した用語を用いて提案手法の解説や評価を行うことがあるが，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（**図 1**，**図 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（5 節）．

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

図 1 実際の設問例.

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 [4]. データセットはクラウドソーシングサービス Lancers[*1] 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 [4] と同様に，Vocabulary Size Test（VST）[2] を用いた．ただし，VST は 100 問からなるのに対して，[4] 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．この設問例を図 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

[*1] https://lancers.co.jp/

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

図 2 意外な意味を問う設問例.

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．例えば図 1 であれば，複数形の選択肢が内容に配慮されている．

実際の設問例が図 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 問の通常の語義の語彙テストの前に，図 2 のような設問を 13 問解くように求められる．意外な語義の 13 問のうち，12 問については同じ語の通常の語義の設問が，前者の 70 問の中に現れる．すなわち，12 問に対しては同じ語の意外な語義と通常の語義の両方の問題を解く．残り 1 問については，確認用に意外な語義だけに現れる語を用いた．

4. 項目反応理論

生成 AI への説明として有効な尺度の説明の前に，分野に依らない試験の設問の難しさの尺度を求めるために広く使われている項目反応理論について説明する．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．

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

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性質が良いことを示している．

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 [16] があるため，よりパラメタ数の少ない 2PL モデルを用いた．

図 3 学習者トークンの導入（[6]）.

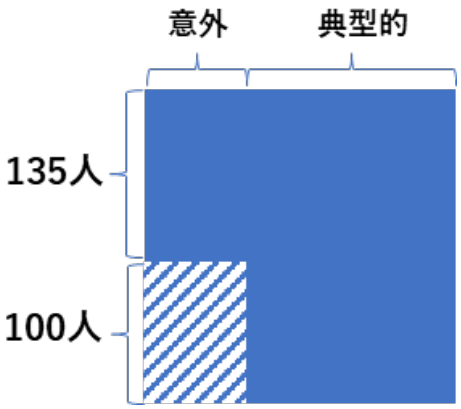

図 4 実験設定．青く塗られた部分がパラメタ推定に使われる訓練データ．斜線部が性能比較に用いられるテストデータ．今回も，学習者反応予測については，[6] と同一の設定を用いる．

表 1 図 4 斜線部の予測精度 (accuracy)（[6]）．設問文の文意をマスク言語モデルが考慮することで，設問文の文意を考慮しない手法より精度が向上することが分かる．

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
[6] の手法 (bert-large-cased)	0.674 (**)
[6] の手法 (bert-base-cased)	**0.688** (**)
[6] の手法 (bert-base-uncased)	0.655
[6] の手法 (roberta-base)	0.681 (**)
[6] の手法 (albert-base-cased)	0.671 (*)

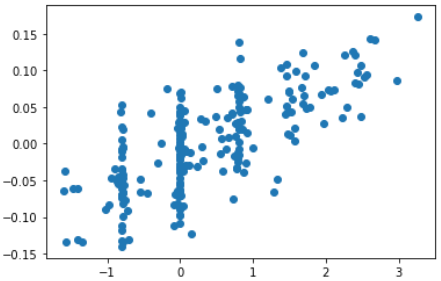

図 5 IRT の能力パラメタ（横軸，pyirt によって算出）と，学習者トークンの単語埋め込みベクトルの第一主成分得点（縦軸）([6]）．マスク言語モデルから学習者の能力値を抽出できることが分かる．

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

生成 AI に対する難しさの尺度の説明として IRT を説明したが，IRT は設問文のテキスト情報を一切使わない特徴がある．設問文のテキスト情報の利用の精度への貢献の測定を，語の意外/典型的な語義の設問に対する回答予測のデータセットで調査する．まず，235 人の学習者を 135 人と 100 人にランダムに分ける（図 4）．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 1 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．

5.2 マスク言語モデルと 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？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，Bidirectional Encoder Representations from Transformers (BERT)[3] に代表されるマスク言語モデルと IRT の予測性能を比較した．

マスク言語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，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，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 マスク言語モデル上の個人化判別

ここではマスク言語モデルを個人化判別に対応させる [6] の手法を説明する．個人化判別のために学習者トークンを加える手法は，著者の知る限り [6] 以前はない．ただし，マスク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．[6] 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ている（**図 3**）．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，図 1 のどの単語に下線が引かれているかや，図 1 や図 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，図 1 と図 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．

マスク言語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**，微調整の訓練には Adam 法 [8] を用いバッチサイズは 32 とした．

マスク言語モデルを用いた結果を，**図 1** 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．図 1 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，図 1 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は，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) に適していたためであると考えられる．

6. 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，マスク言語モデルでは，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，マスク言語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．

投稿時点ではより広く使われていることから，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 5**）．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 5 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．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** パッケージも用いて推定した．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 [12]．相関係数は 0.72 で，やはり統計的有意性を示した ($p < 0.01$)．

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．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．

7. 設問の難しさや識別力の抽出法

ここまでは微調整済の BERT モデルから学習者の能力値を抽出する方法であったが，さらに，設問の難しさや識別力に相当する値を抽出する方法を提案する．方法の概略を示す．BERT は被験者が設問文が指定されれば，その被験者がその設問に正答できるかどうかだけでなく，その確率値も予測として出力できる．ある設問に着目し，全被験者がその設問を解いた時の正答できる確率を BERT に出力させ，ここからその設問の正答者数の確率分布を計算する．被験者間の独立性を仮定すると，数学的には，成功確率が互いに異なる独立なベルヌーイ分布の和の分布であるポアソン 2 項分布を計算する事に相当する．この時，その設問の正答者数の確率分布の平均を設問の難易度，分散を識別力のような設問の良さと解釈する事が可能になる．

ここでは，被験者数を N 人とし，学習者の添字を n とする（厳密には，被験者の中から特定の被験者を選び N と J が異なる設定もあり得るので，違う文字でおいた）．項目数を I 個とし，項目の添字を i とする．学習データ上で予測器を微調整した後，予測器は学習者 n が項目 i に正しく回答する確率を出力することができる．この確率を $BERTProb(n, i)$ と表記する．簡単のために，ここからは設問 i に焦点を当てる．$BERTProb(n, i)$ を使って，N 人のうち，質問 i に正答する者の確率分布を求めたい．そこで $BERTProb(n, i)$ の確率で 1，そうでなければ 0 となるベルヌーイ分布に従う確率変数 A_n を $A_n \sim Bernoulli(BERTProb(n, i))$ と定義する．簡単のため，確率変数 $\{A_1, \ldots, A_n\}$ は互いに独立と仮定する．学習者について和をとり，項目 i の全 N

人の中での正答者数の確率分布は次のように書ける．

$$A_i = \sum_{n=1}^{N} A_n \tag{2}$$

式 2 は互いに独立なベルヌーイ分布の和であり，ポアソン 2 項分布と呼ばれる[*2]．この分布の計算は，動的計画法を用いて計算可能である．[5], [7] ではポアソン 2 項分布の計算を全く違うタスクに対して行う中で詳述しているので，計算アルゴリズムの詳細はこちらを参照されたい．

A_i は確率分布なので，平均と分散を計算できる．A_i は，全 N 人のうち，項目 i の正答者数である事に注意すると，A_i の平均は，問題 i の難易度を表していると解釈できる．また，A_i の分散は，問題 i の正解者数の予測値のエラーバーと解釈できる．同じような難しさの設問の中では，分散が最も小さい，つまり，正答者数の予測がつきやすい問題が性質が良い．A_i の分散は，項目反応理論における「識別力」に似た性質を持つ指標である．項目反応理論の識別力は，項目が能力の高い被験者と低い被験者を識別する力を表す．直感的には，能力が本当は高い被験者が間違えてしまうような確率の少ない性質の良い問題である度合いを表す．A_i の分散も，項目反応理論の識別力のように性質の良い問題である度合を表すが，項目反応理論はモデルが固定されているのに対し，A_i の分散は予測器 $BERTProb(n, i)$ の確率値さえわかればどのような予測器を用いても計算でき，深層転移学習等，複雑な手法を用いた場合でも計算できる．

横軸に A_i の分散，縦軸に A_i の平均値をとることでリスク・リターンプロットを作成できる．まず，どの程度の難しさの設問を選びたいかを決めて縦軸の値に注目し，次に同程度の難しさの問題の中で横軸の値が最も小さいもの（最も左にあるもの）を選ぶことで，特定の難易度の性質の良い問題を選択可能である．この最も左にある点を結んだ線を「効率的フロンティア」という [7].

図 6 と**図 7** に，ある項目（設問）i について，受験者数がそれぞれ 5 人，100 人である場合の分布を描いた．(5 人については，ランダムに受験者を選んだ）図 7 には，難しい項目と簡単な項目の 2 つを選んで正答者数の分布を描いて重ねたグラフを示した．実際の正答者数は，図 7 の左側が 31 人，右側が 56 人であった．図 6 から，受験者数が少くとも，非対称な分布の形が計算できている事が分かる．

今回，提案手法は予測される正答者数の分布の平均を設問の難しさとして，標準偏差を設問の難しさ推定のしやすさとして出力できる．こうした値に類似した概念は，IRT においても，それぞれ，困難度，識別力という名前で知られている．**図 8**, **図 9** に，提案手法による値と，テストデータ中の値を与えたうえで IRT が推定した困難度・識別力の値の（簡単のため）負値を図示する．相関係数は，図 8 では 0.78 ($p < 0.01$)，図 9 では 0.62 ($p < 0.05$) であり，どちらも統計的有意な相関がみられた．また，**図 10** にリスク-リターンプロットを描いた．各点は前述の 12 問の設問であり，破線は効率的フロンティアである．各点は設問を表し，各点の番号は設問番号である．縦軸が同程度の値であれば，分散が小さい設問（図中左側の設問）が性質の良い設問である可能性が高いとされた．全 12 問のうち，効率的フロンティア上の設問を選ぶことで，3 問の性質の良い設問を選び出せている事が分かる．

[*2] https://en.wikipedia.org/wiki/Poisson_binomial_distribution

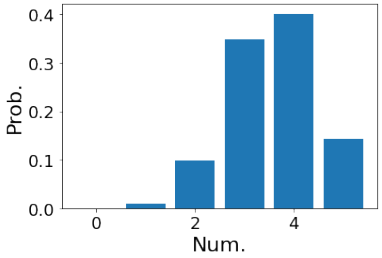

図 6 ある項目（設問）で，受験者数が 5 人のときに予測される正答者数の分布．

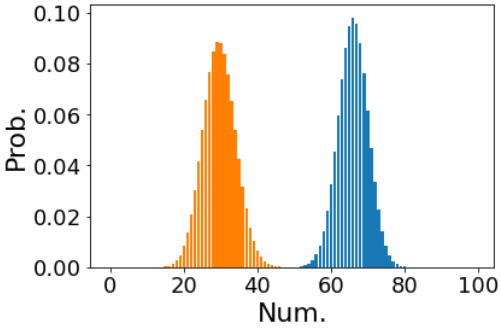

図 7 受験者数が 100 人であるときに予測される正答者数の分布．

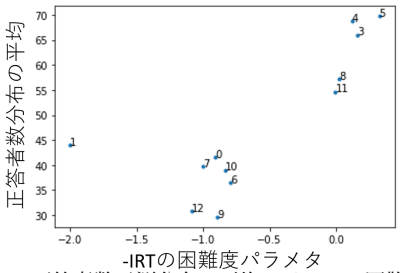

図 8 正答者数予測分布の平均と-(IRT の困難度).

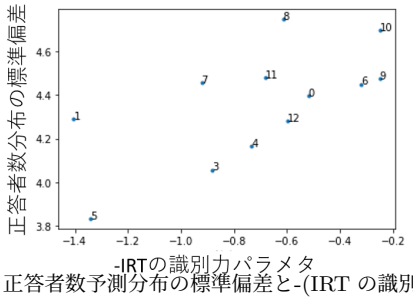

図 9 正答者数予測分布の標準偏差と-(IRT の識別力).

8. 生成 AI への指示による設問の生成

図 11 に，実際に [6] の「意外な意味」の語彙テストデータセットに含まれる語のテストの例を示す．図 4 の設定

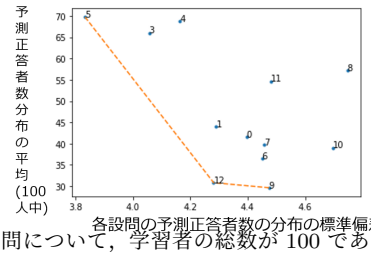

図 10 各設問について，学習者の総数が 100 である場合のリスク（横軸，各設問の予測される正解者数の分布の分散）とリターン（縦軸，各設問の予測される正解者数の分布の平均）．

で，最も性能の高い “bert-base-cased” を用いて予測したときの予測正答者数分布の平均は 69 人，標準偏差は 4.16 人であった．また，項目反応理論の 2PL モデルを用いてデータが全て利用できる場合（つまり，図 4 の斜線部もわかっている状態）での困難度と識別力を算出したところ，困難度は-1.2 であり，識別力は 0.738 であった．

この設問に対して，生成 AI として有名である ChatGPT の GPT-4 (ChatGPT May 24 Version) を用いて，新しい英語テストを生成する実験を行った．具体的には，次のような指示文のあとに，**図 12** のようなプロンプトで，具体的にどのような設問を生成してほしいのか日本語で指定した．指示文には日本語を用いた．これは，[6] のデータセットは，日本のクラウドソーシングサービスである Lancers で作成しているため，回答者の大部分は日本語を母語とする英語学習者であると推測されるためである．ただし，回答者のプライバシーなどの観点から，[6] では回答者の母語については入力等をしてもらったり，明示的に日本語を母語とする学習者のみに回答させたりはしておらず，あくまで推察・暗示されるだけである．このため，図 12 においても，日本語を用いた指示文で英語の英単語テストの設問を生成する事により，日本語を母語とする英語学習者向けの試験問題を生成していることが暗示されるようにするにとどめ，「日本語を母語とする英語学習者向けの試験問題を生成せよ」といった明示的な指示は与えなかった．

GPT-3.5 (ChatGPT May24 Version) でも図 12 のようなプロンプトを与えたが，指示文の中で例示した図 11 の設問の引きずられたためか，全く新しい設問は生成されなかった．そのため，実験には一貫して GPT-4 を用いた．GPT-4 でも，「全く違う単語を使って」を入れないと，指示文中で利用している図 11 と同じ単語や選択肢を含む設問が生成されることがあったため，明示的に「全く違う単語を使って」を指示文に加えた．これにより，図 11 と同じ単語や選択肢を含む設問が生成されることはなくなった．図 12 の指示では，特に新しい問題を複数生成しても良いが，1 回の生成に対して新しい問題がかならず 1 問生成されるようになった．また，Chain of Thought といって，「1 つずつ順序立てて考えてください」という指示を入れると，目的とする生成が出力されることが多くなるという報告がある (`https://arxiv.org/abs/2201.11903`) の

The area was _____ in timber and coal.
a) inexpensive b) cheap c) poor d) not well off

図 11 実際に語彙テストデータセットに含まれる設問の 1 例．予測された正答者数分布は平均 69 人，標準偏差は 4.16 であった．また，項目反応理論の 2PL で算出した場合の困難度は-1.2，識別力は 0.738 であった．

で，図 12 に加えた．また，句読点については，日本語は圧倒的に「、。」を使ったテキストが多く，言語モデルでは句読点も生成されることから，指示文では工学系の一部論文等で使われる「,.」は用いず，明確に「、。」を句読点に用いたテキストを使用した．

図 11 に対して，正答者数の平均が 95 人になるように指示した結果が**図 13**，平均が 29 人になるように指示した結果が**図 14** である．実際に，それぞれ，設問が容易／困難になっていることが分かる．一方，図 11 に対して，正答者数分布の標準偏差を操作することにより生成させた設問が，**図 15** と**図 16** である．正答者数分布の標準偏差の場合は，人間でも作成した設問の標準偏差を予測することが難しいことから，具体的な数値目標は指定せず，単純に「大きく」,「小さく」と指定するにとどめた．正答者数分布の標準偏差を大きくした場合は，複数の選択肢が正解と思われる設問が生成された．一方，正答者数分布の標準偏差を小さくした場合は，より，正解が明確と思われる設問が生成された．

特に，図 14 と図 15 の違いが重要である．正答者数を少なくさせるような設問を生成するように依頼した図 14 では，単に，専門的な語（通常，コーパスの頻度が低い語）を答えさせる設問を生成したのに対して，図 15 使用されている語は一般的な語であるが，どの選択肢も答えになりうる設問が生成されている．これは，実際に，「正答者数分布の平均」の大小が設問に及ぼす影響と，「正答者数分布の標準偏差」の大小が設問に及ぼす影響の違いを，GPT-4 が正しく理解していることを示す．以上のように，GPT-4 は，正答者数分布の平均も標準偏差も，設問のどのような性質に対応しているかを理解していると定性的に示された．

一方，同様の指示を，項目反応理論の「困難度」や「識別力」といった専門的な語で指示した場合は，GPT-4 は，どのような設問を生成すればいいか，正しく理解できるのだろうか？実際に，このように項目反応理論の用語を用いて指示して設問を生成したものが**図 17**，**図 19**，**図 18** である．図 17 では，やはり，低頻度語を用いた設問で困難度を上昇させているのに対し，図 18 では，低頻度語を特に用いずに，設問を曖昧にすることで，識別力を低下させている．従って，驚くべきことに，GPT-4 は困難度や識別力といった項目反応理論の概念も適切に理解し，定性的には指示に沿った設問生成が可能であるように見える．

9. おわりに

本研究では，項目反応理論の「困難度」や「識別力」と

次の英単語テスト問題は、ある 100 人の英語学習者からなる受験者集団において、予測される正答者数の分布が平均 69 人、標準偏差 4.16 人と予測されています。この受験者集団において、予測される正答者数の分布が同程度で、標準偏差がより小さい英単語テスト問題を、全く違う単語を使って生成してください。１つずつ順序立てて考えてください。
====
"The area was _____ in timber and coal."の下線部に入る語を inexpensive, cheap, poor, not well off の 4 つから選びなさい。

図 12 実際に用いたプロンプト.

He turned off the lights and went to _____.
a) bed b) kitchen c) car d) park

図 13 正答者数 95 人になるように指示して生成された設問.

The researcher analyzed the _____ between the two variables.
a) apex b) correlation c) paradigm d) zenith

図 14 予測正答者数分布の平均が 29 人になるように指示して生成された設問.

He is very _____.
a) kind b) friendly c) evil d) generous

図 15 予測正答者数分布の標準偏差を大きくするように指示して生成された設問.

The birds _____ south for the winter.
a) fly b) drive c) walk d) run

図 16 予測正答者数分布の標準偏差を小さくするように指示して生成された設問.

Despite the challenging conditions, the team remained _____.
a) resilient b) permeable c) solvable d) inflammable

図 17 項目反応理論の困難度を-1.2 から 0.89 に変更するように指示して生成された設問.

The sun _____ in the east.
a) rises b) falls c) sinks d) dives

図 18 項目反応理論の識別力を 0.738 から小さくするように指示して生成された設問.

He has a _____ to exaggerate things.
a) tendency b) progression c) direction d) development

図 19 項目反応理論の識別力を 0.738 から大きくするように指示して生成された設問.

いった専門知識を持たないユーザでも，生成 AI に目的とする設問の難しさ等の性質を伝えられるように，より直感的な予測正答者数分布の平均・標準偏差を用いて指示する手法を提案した．そのために，BERT などのマスク言語モデルを用いて，設問文の文意を考慮した個別学習者反応予測を行う [6] の手法を拡張し，テストデータの受験者集団に対する予測結果から，正答者数分布をポアソン二項分布に基づき算出する設問難度推定法を提案した．さらに，実際に，提案する設問難度推定法により，代表的な生成 AI である ChatGPT を用いて設問生成を行うことにより，指示を行う人間の意図に沿った設問の生成が行える事を定性的に確認した．本稿を通じて，簡単のため，英語の語彙テストのデータで実験したが，提案手法は，結果が正解／不正解に 2 値で与えられる形式であれば，プログラミング課題などの他の分野の設問にも適用可能である．

今後の課題として，予測正答者数分布の平均や標準偏差を数値指定して生成 AI で設問生成を行った場合に，実際にその設問を受験者集団に実施し，指定した数値に沿った設問生成を行えているか数値評価することが挙げられる．

謝辞

本研究は，科学技術振興機構 ACT-X 研究費 (JPMJAX2006) の支援を受けた．

参考文献

[1] Baker, F. B.: *Item Response Theory : Parameter Estimation Techniques, Second Edition*, CRC Press (2004).
[2] Beglar, D. and Nation, P.: A vocabulary size test, *The Language Teacher*, Vol. 31, No. 7, pp. 9–13 (2007).
[3] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. of NAACL* (2019).
[4] Ehara, Y.: Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing, *Proc. of LREC* (2018).
[5] Ehara, Y.: LURAT: a Lightweight Unsupervised Automatic Readability Assessment Toolkit for Second Language Learners, *Proc. of ICTAI*, pp. 806–814 (2021).
[6] Ehara, Y.: No Meaning Left Unlearned: Predicting Learners' Knowledge of Atypical Meanings of Words from Vocabulary Tests for Their Typical Meanings, *Proc. of Educational Data Mining (short paper)* (2022).
[7] Ehara, Y.: Selecting Reading Texts Suitable for Incidental Vocabulary Learning by Considering the Estimated Distribution of Acquired Vocabulary, *Proc. of Educational Data Mining (poster paper)* (2022).
[8] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *Proc. of ICLR* (2015).
[9] 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 (2010).
[10] Nation, I.: How Large a Vocabulary is Needed For Reading and Listening?, *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82 (2006).
[11] Nation, I. S. P. and Waring, R.: *Teaching Extensive Reading in Another Language*, Routledge (2019).
[12] Paek, I. and Cole, K.: *Using R for item response theory model applications*, Routledge (2019).
[13] Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D. and Chen, G.: Assessing Algorithmic Fairness in Automatic Classifiers of Educational Forum Posts, *Proc. of AIED*, pp. 381–394 (2021).
[14] Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S. and Lee, D.: Classifying Math Knowledge Components via Task-Adaptive Pre-Trained BERT, *Proc. of AIED*, pp. 408–419 (2021).
[15] Xu, S., Xu, G., Jia, P., Ding, W., Wu, Z. and Liu, Z.: Automatic Task Requirements Writing Evaluation via Machine Reading Comprehension, *Proc. of AIED*, Springer, pp. 446–458 (2021).
[16] 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).