自然言語処理の教育応用において 学習者集団に非依存な難度の尺度は本当に必要か？

江原遥 [1]
[1] 東京学芸大学　教育学部
ehara@u-gakugei.ac.jp

概要

教育応用では，個々の学習者に合わせた個別最適な支援（自動作問など）が求められる．一方，作問された設問の「難度」は，設問を解く学習者に関わらず定まる，学習者集団に非依存な尺度であることが求められ，一般ユーザには解釈が難しい，アノテータ間の難度の基準の構築に労力が割かれてきた．しかし，自然言語処理の教育応用では，(学習者, 設問文) のペアから個々の学習者が設問に正答するかを予測する予測器さえ構築できれば応用上有用な場面が多く存在すると思われる．本稿では，こうした予測器が入手できる場面で，学習者集団中の予測正答者数の平均を一般ユーザにも解釈しやすい難度の代わりに使える尺度として使う手法を提案する．

1　はじめに

教育応用では，個々の学習者に合わせた個別最適な支援が求められる．例えば，学習者が過去に回答した設問の設問文や，その回答の正誤が分かっていれば，その設問文のテキスト情報と学習者情報の両方を考慮して，学習者が次に回答すると最も教育効果の高い設問を選んだり生成したりする支援が考えられる．個別最適な支援を行うためには，このように設問テキスト情報と，読み手となる学習者の両方を同時に考慮して，**個々の学習者にとっての**難度を考慮する事が望ましい．ところが，自然言語処理の教育応用の研究では，学習者とは独立に設問文だけから決まる「難度」の情報を教師やアノテータが付与し，これを予測する研究が大半であった．例えば，第二言語学習者向けにテキストのリーダビリティを推定する研究の大部分がこうしたタスク設定である．こうしたタスク設定では，アノテータ間で合意できる難度の尺度を作ることが難しいという問題がある．例えば，第二言語学習者にとっての英語のリーダビリティの代表的なデータセットである [1] では，複数のアノテータ間で合意しやすいように，難度の段階数を 3 段階まで絞っている．また，大規模言語モデルを用いた ChatGPT 等の生成 AI が社会に普及してきている現状では，こうした個々の難度の尺度を一般の利用者が理解して使用する事は現実的とは言い難い．

こうした状況の中，学習者の試験結果データを大規模言語モデルに組み入れる手法についての研究が進み，(学習者, 設問文, 正答/誤答) のデータがあれば，個々の学習者が所与の設問文の設問に正答する確率を予測する予測器を構成する事は容易になってきた．そこで，**自然言語処理の教育応用に，本当に設問文だけから決まる難度という概念が必要であるのか，こうした予測器さえ構築できれば，時間や労力をかけて，アノテータ間で合意できる難度の基準を作成する必要はない教育応用も多いのではないか？**，が本稿の主題である．

そこで，本研究では，こうした予測器が与えられている状況で，より人間にも生成 AI にも解釈しやすく難度のように使える尺度として，予測器から計算できる，学習者集団中の正答者数の分布を用いる手法を提案する．「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」といった指示を生成 AI に行い，正答者数分布を用いることで学習者集団に即した作問を行う手法を提案し，定性的に評価する．また，人間に対する解釈性の観点からも，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．学習者を表す特殊トークンを導入する事で，Bidirectional Encoder Representations from Transformers （BERT [2]）をはじめとする典型

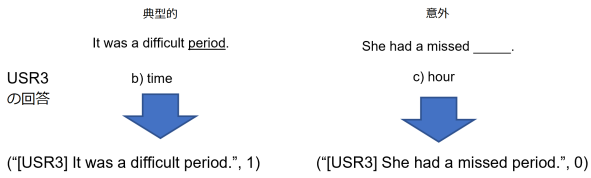

図 1 学習者トークンの導入（[3]）.

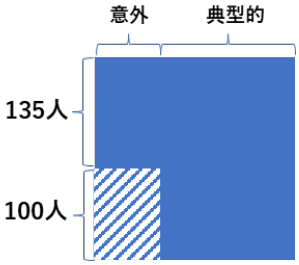

図 2 実験設定．青く塗られた部分がパラメタ推定に使われる訓練データ．斜線部が性能比較に用いられるテストデータ．今回も，学習者反応予測については，[3] と同一の設定を用いた．

的な大規模言語モデル・マスク言語モデルをモデルの改変なしにそのまま用いて，(学習者, 設問文, 正答/誤答) のデータから学習者の正答/誤答確率を出力する手法 [3] を紹介する．このように，特殊トークンを導入する手法は，Named Entity に相当するトークンを入れる [4] 等の手法にも用いられている．次に，予測器を用いて，個々の設問に対して，「訓練データ中の学習者集団に対する予測正答者数分布」を求める手法を説明する．

2 関連研究

本研究のように，学習者の特性を考慮しつつ，学習者集団とは独立に定義される難度の尺度も取り出そうとする試みには [5] が挙げられる．しかし，こうした研究では，BERT 等の一般的な自然言語処理のモデルではなく，内部に難度パラメタを持つ専用に設計されたモデルを使わなければならない．さらに，その難度の尺度は一般ユーザには解釈が難しい問題は残る．BERT などの言語モデルを用いた教育応用の既存研究はあるが [6, 7, 8]，これらの研究では難度の基準の難しさなどは扱われていない．

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．応用言語学の分野においても，語の語義ごとの難度を評価したデータセットは知る限りないため [9, 10, 11]，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 [3] を用いた．

3 設問の難度や識別力の抽出法

ここでは，BERT の予測器から，設問の難度や識別力に相当する値を抽出する方法を [12] に沿って説明する．また，項目反応理論（Item Response Theory, IRT）の困難度や識別力といった概念についても [12] を参照されたい．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．

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

式 1 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[1)]．この分布の計算は，動的計画法を用いて計算可能である．[13, 14] 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．同じような難度の設問の中では，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．

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度として，標準偏差を設問の難度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．[12] では，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・識別力の値が統計的有意に相関していることを報告している．また，[12] では，リスク-リターンプロットを用いて，正答者数分布を予測しやすいという意味で性質の良い問題を選び出せることなども詳細に報告している．

4 生成 AI への指示による設問の生成

表 3 に，実際に [3] 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 2 の設定で，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 2 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，表 4 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，[3] 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．

GPT-3.5 (ChatGPT May24 Version) でも表 4 のようなプロンプトを与えたが，指示文の中で例示した表 3 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表 3 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，表 3 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．表 4 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 [15] ので，表 4 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．

表 3 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表 5，平均が 29 人になるように指示した結果が表 6 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，表 3 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，表 7 と表 8 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

1) https://en.wikipedia.org/wiki/Poisson_binomial_distribution

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

図 3 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．

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 4 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 5 正答者数 95 人になるように指示して生成された設問．

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，表 6 と表 7 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表 6 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，表 7 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表 9，表 11，表 10 である．表 9 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，表 10 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．

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

図 6 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 7 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 8 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 9 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 10 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 11 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．

5 おわりに

本研究では，これまで学習者集団に依存しない設問文の難度について，難度の基準のすり合わせが難しい事を指摘した．そして，(学習者, 設問文, 正答/誤答) のデータから正答確率を予測する予測器を用いて，学習者集団に依存はするものの正答者数予測分布の平均や分散が，難度として利用できることを示した．提案手法は，平均や分散の概念さえ分かればよいため，項目反応理論などの難度の基準構築になじみのない一般ユーザでも，より直感的に生成 AI を用いて自動作問する際などに用いる事ができる．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．

本研究の今後の課題としては，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，JSPS 科研費 22K12287 の支援を受けた．

参考文献

[1] S. Vajjala and I. Lučić, “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, pp.297–304, 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proc. of NAACL, pp.••–••, 2019.

[3] Y. Ehara, “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), pp.••–••, 2022.

[4] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto, “LUKE: Deep contextualized entity representations with entity-aware self-attention,” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), eds. by B. Webber, T. Cohn, Y. He, and Y. Liu, pp.6442–6454, Association for Computational Linguistics, Online, Nov. 2020. https://aclanthology.org/2020.emnlp-main.523

[5] E. Tsutsumi, Y. Guo, R. Kinoshita, and M. Ueno, “Deep knowledge tracing incorporating a hypernetwork with independent student and item networks,” IEEE Transactions on Learning Technologies, pp.••–••, 2023.

[6] J.T. Shen, M. Yamashita, E. Prihar, N. Heffernan, X. Wu, S. McGrew, and D. Lee, “Classifying math knowledge components via task-adaptive pre-trained bert,” Proc. of AIED, pp.408–419, 2021.

[7] L. Sha, M. Rakovic, A. Whitelock-Wainwright, D. Carroll, V.M. Yew, D. Gasevic, and G. Chen, “Assessing algorithmic fairness in automatic classifiers of educational forum posts,” Proc. of AIED, pp.381–394, 2021.

[8] S. Xu, G. Xu, P. Jia, W. Ding, Z. Wu, and Z. Liu, “Automatic task requirements writing evaluation via machine reading comprehension,” Proc. of AIEDSpringer, pp.446–458 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, Oct. 2006.

[10] B. Laufer and G.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 R. Waring, Teaching Extensive Reading in Another Language, Routledge, Nov. 2019.

[12] 江原遥，“学習者回答予測モデルからの設問の正答者数予測分布推定，” 言現由佳 29 嫁実架肱塀，pp.••–••，2023.

[13] Y. Ehara, “Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners,” Proc. of ICTAI, pp.806–814, 2021.

[14] Y. Ehara, “Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary,” Proc. of Educational Data Mining (poster paper), pp.••–••, 2022.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., “Chain-of-thought prompting elicits reasoning in large language models,” Advances in Neural Information Processing Systems, vol.35, pp.24824–24837, 2022.

A　付録：実験設定の詳細

回答者のプライバシーなどの観点から，[3] 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，表 4 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．