深層学習に基づく仮想受検者を用いた項目特性値推定手法の提案

Predicting Item Parameters Using Deep-Learning–Based Virtual Examinees

栗田 侑弥 *1，宇都雅輝 *1
Yuya Kurita*1, Masaki Uto*1
*1 電気通信大学
*1The University of Electro-Communications
Email: {kurita, uto}@ai.lab.uec.ac.jp

あらまし： 客観式テストにおいて広く利用される項目反応理論（IRT）に基づくテスト運用では，事前に難易度などの特性値が推定された項目プールが必要となる．項目特性値を推定するためには，項目を受検者集団に出題して反応データを収集する必要があるが，これにはコストがかかる．そこで本研究では，人間受検者の項目反応を模倣する仮想受検者モデルを深層学習を用いて開発し，これを用いて項目特性値を推定する手法を提案する．

キーワード： 項目特性値推定，深層学習，言語モデル，項目反応理論

1　はじめに

客観式テストにおいて受検者の能力を測定する手法の一つとして項目反応理論（IRT）が近年広く利用されている．IRT では，事前に難易度などの特性値が推定された項目を出題することで，個別の項目の特性を考慮した高精度な能力測定を実現する．したがって，IRT に基づく試験運用では，本番試験で出題する項目を事前に同質の受検者集団に出題して回答データを収集する手続きが必要となる．しかし，この工程はコストが高いことに加え，項目内容の漏洩につながるリスクも有する．これらの問題を解決するアプローチの一つとして，近年，自然言語処理技術を用いて問題文情報から項目特性値を予測する手法が提案されている [(1)]．具体的には，特性値が既知の項目データセットを用いて，項目のテキスト情報から特性値を予測する深層学習モデルを訓練し，新規項目の特性値を問題文から予測する．深層学習モデルには，BERT（Bidirectional Encoder Representations from Transformers）などの事前学習済み深層学習モデルが利用されている．しかしこのアプローチは比較的膨大な学習データセットで学習した予測モデルでも，実用に十分な精度には至っていない [(1)]．

一方でこれとは異なるアプローチとして，深層学習モデルに基づく質問応答（Question Answering:QA）システムに項目を解かせて正誤反応データを収集し，それを用いて項目特性値を推定するアプローチが提案されている [(2)]．この手法は，項目特性値予測モデルを用いた従来手法と比べて高い推定精度を達成している．他方で，この手法では，QA システムと人間の受検者が類似した傾向で回答することを仮定している．しかし，実際にはこの仮定が成立する保証はない．

この問題を解決するために本研究では，人間受検者の実際の項目反応を学習させた QA システム（以下，仮想受検者と呼ぶ）を構築し，それを用いた項目特性値推定手法を提案する．

2　提案手法

本研究で提案する仮想受検者は，受検者の能力値と項目の問題情報を入力として受け取って，その能力値を持つ受検者の正答確率を出力するモデルであり，人間受検者の実際の項目反応を学習させることで構築する．このシステムは，特定の能力の人間受検者の反応を模倣するため，「仮想受検者」とみなせる．なお，本研究では，項目の形式として多肢選択式を想定し，モデルへの入力を問題文と一連の選択肢に能力値を連結したテキストとして，指定した能力値の受検者の正答確率を出力する仮想受検者を設計する．具体的には，受検者の能力値を θ，問題文の単語系列を w，k 個の選択肢の単語系列をそれぞれ $q_1, q_2, \ldots, q_k$，そのうちの正答選択肢を q_1 とすると，モデルへの入力系列は次の通りとなる．

$$\text{[CLS]}\ w\ \text{[SEP]}\ \text{[COR]}\ \ q_1\ \text{[SEP]} \ldots\ q_k\ \text{[SEP]}\ \theta\ \text{[SEP]}$$

ここで [CLS] は入力の冒頭を表す特殊タグ，[SEP] は問題文と選択肢，能力値の区切りを表す特殊タグ，[COR] は正答選択肢を表す特殊タグである．

一般に BERT では，特殊タグ [CLS] に対応する出力を入力情報を集約した分散表現ベクトルとみなす．そのため本モデルでは，[CLS] タグに対応するモデルの出力ベクトル $\boldsymbol{h}$ に次式で定義される全結合層を適用することで，能力値 θ の受検者の正答確率 $\hat{P}$ を出力する．

$$\hat{P} = \boldsymbol{W}_{\hat{P}}\boldsymbol{h} + b_{\hat{P}} \tag{1}$$

ここで，$\boldsymbol{W}_{\hat{P}}$ と $b_{\hat{P}}$ は重みとバイアスを表すパラメータである．

モデル学習の際には，まず受検者の正誤反応データに IRT を適用して各受検者の能力値を推定しておき，各受検者の能力推定値と正誤反応データを利用して次式の二値交差エントロピー損失を誤差逆伝播法で最小化する．

$$-\frac{1}{N}\sum_{i=1}^{N}\Big(u_{ij}\log(\hat{P}_{ij}) + (1-u_{ij})\log(1-\hat{P}_{ij})\Big) \tag{2}$$

ここで，N はデータの総数，u_{ij} は受検者 j が項目 i に正答した時 $u_{ij} = 1$，誤答した時 $u_{ij} = 0$ となる変数を表し，$\hat{P}_{ij}$ は項目 i と能力値 θ_j を仮想受検者に入力して得られる正答確率である．

新規の項目に対して特性値を推定する際には，訓練時に利用したデータに存在するそれぞれの受検者の能力値を所与とした仮想受検者集団を用意し，対象の項目に対する各仮想受検者の正答確率データを収集した上で，その確率に基づいて各仮想受検者の正誤反応データを収集する．そのようにして得られた仮想受検者集団の正誤反応データと，各仮想受検者に設定された能力値を所与として，IRT を用いて対象項目の特性値を推定する．

3　評価実験

本研究では，実データ実験を通して提案手法の性能を評価した．本実験では，公開データセットである英語語彙テストを含む EVKD（The ESL Learners' Vocabulary Knowledge Dataset）[3] を使用した．このデータセットには，100 問のテスト項目が含まれ，それぞれに問題文，正解選択肢 1 つ，および 3 つの誤答選択肢，さらに 100 人の受検者からの正誤反応データが含まれている．実験手順は次のとおりである．

1. EVKD データセットを，90% の項目を含む訓練データと 10% の項目を含むテストデータに分割した．
2. 得られた訓練データを用いて仮想受検者モデルを構築し，テストデータ内の各項目に対する仮想受検者の正答確率データを収集した．なお，訓練データからの受検者の能力値推定には，2 パラメータロジスティックモデル（2PLM）を利用した．ここで，2PLM は，能力値 θ の受検者が難易度 β および識別力パラメータ α を持つ項目に正答する確率 p を式 (3) で表すモデルである．

$$p = \frac{1}{1 + exp(-\alpha(\theta - \beta))} \tag{3}$$

3. 得られた正答確率に対し，正答確率が 0.5 より大きい場合は正答，0.5 以下の場合は誤答とみなすことで収集した仮想受検者集団の正誤反応データを用いて，2PLM に基づいて各項目の難易度を推定した．
4. 得られた項目難易度と真値の一致度を評価した．難易度の真値は，訓練データから推定された各受検者の能力値を所与とし，テストデータに対する人間受検者の正誤反応データから推定された難易度値とした．また，一致度の指標には，相関係数を用いた．
5. 以上の実験を，訓練データとテストデータの分割方法をランダムに変えて 10 回行った．

比較のために，問題文から直接難易度を予測する従来手法についても同様の実験を行なった．

また，テスト項目の識別力を推定する実験についても行った．この場合，従来手法である特性値予測モデルは，問題文が入力されたとき，識別力を出力するように設計した．提案手法では，実験手順 3 において予測した正答確率 P に基づくランダムサンプリングにより正誤反応データの収集を行った．こちらの実験も，訓練データとテストデータの取り方をランダムに変えて 10 回行った．

なお，提案手法の仮想受検者と従来手法の予測モデルの基礎モデルとしては，DeBERTa-v3-large（Decoding-enhanced BERT with Disentangled Attention v3 Large）を利用し，実装は PyTorch で行なった．

表 1　提案手法と従来手法の特性値推定精度比較

	難易度	識別力
従来手法	0.440	0.374
提案手法	**0.510**	**0.416**

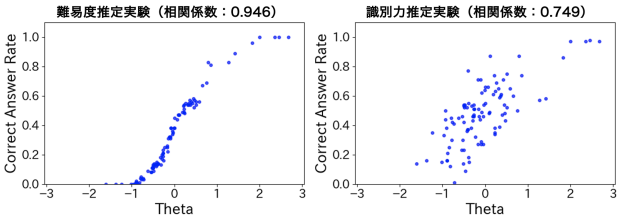

図 1　仮想受検者に指定した能力値と正答率の関係

4　実験結果

4.1　特性値推定精度評価

実験結果を表 1 に示す．表中の数値は 10 回の繰り返し実験の平均値を表す．値が大きいほど性能が良いことを意味し，各条件で性能が高い数値を太字で表記している．実験結果から，提案手法が従来手法より高い精度を達成できることが示された．

4.2　仮想受検者の能力評価

提案手法で構築した仮想受検者が，指定した能力値の人間受検者の反応傾向を適切に再現できているかを確認するために，指定した能力値と正答率の相関について分析を行った．それぞれの実験における実験結果を図 1 に示す．この結果から，指定した能力値と正答率に高い相関があることがわかり，提案手法が実際の受検者の能力値をある程度模倣できていることが示唆される．

5　まとめ

本研究では，人間受検者の反応を模倣する仮想受検者を用いた項目特性値推定手法を提案し，実験から提案手法の有効性を示した．

参考文献

(1) L. Benedetto. A quantitative study of nlp approaches to question difficulty estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education*, pp. 428–434, 2023.

(2) M. Uto, Y. Tomikawa, and A. Suzuki. Question difficulty prediction based on virtual test-takers and item response theory. In *Proceedings of the Workshop on Automatic Evaluation of Learning and Assessment Content*, pp. 1–11, 2024.

(3) Yo Ehara, Issei Sato, Hideki Oiwa, and Hiroshi Nakagawa. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of the International Conference on Computational Linguistics*, pp. 799–814, 2012.