

## 設問の文意を考慮した解釈性の高い適応的回答予測

江原 遙<sup>†</sup>  
東京学芸大学<sup>†</sup>

## 1 はじめに

学習支援システムにおいて、学習者が項目（設問）に正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項目反応理論 (Item Response Theory, 以下IRT) を用いて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は、通常、被験者の回答パターンにのみ依存し、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、マスク言語モデル等の深層言語モデルが自然文理解で高い性能を示している。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等、学習者適応に用いることが難しい。

本研究では、学習支援システムのために、この問題に対処する簡便な方法を提案する。設問の文意の考慮が重要な設定にするため、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・使用した（節2）。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。典型的/意外での設問の困難度等の分析も行う。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（節3）。大別して2種類の手法を比較した。まず、前述のIRTであるが、これは、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[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[3]などを用いた手法を提案する。Transformerモデルは、能力の考慮など、被験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モデルを適応的回答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す。また、IRTの利点は被験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で推定した能力値や難しさとよく相関する値を抽出する手法も提案する。

## 2 データセット

設問文の文意の考慮が重要な設定で適応的回答予測を評価することが望ましい。そのため、外国語学習における、典型的な語義と意外な語義の語彙の語彙テストという設定で実験を行った。語彙テスト作成・データセット作成は、著者の過去の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<sup>\*1</sup>から、2021年1月に収集した。英語学習に一定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名の被験者から回答があった。Lancersの作業者は大部分

表1 典型的な語義を問う設問例

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

表2 意外な語義を問う設問例

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



図1 学習者トークンの導入

日本語母語話者であるため、学習者の母語は、大部分日本語であると思われる。典型的な語彙テストとしては、[4]と同様に、Vocabulary Size Test (VST) [2]を用いた。学習者にとって意外と思われる語義についての設問は、著者が作問後、複数名の英語母語話者を含む静岡理工科大学の教員に問題として成立している事を確認した。

IRTの困難度・識別力の各パラメタを求めるには、pyirt<sup>\*2</sup>を用いた。これは、周辺化最尤推定 (Marginalized Maximum Likelihood Estimation, 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、設問のペアが12組ある。結果、全てのペアで学習者にとって意外な語義を問う項目の困難度パラメタが、典型的な語義を問う項目のそれより大きく、難しいと判定された。すなわち、意外な語義の方が典型的な語義より難しいと示唆される。この結果は統計的有意であった (Wilcoxon 検定,  $p < 0.01$ )。

## 3 適応的回答予測の精度による評価

IRTを用いた手法は、学習者の回答のみに依存し、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被験者反応予測を行うと、被験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)[3]に代表されるTransformerモデルとIRTの予測性能を比較した。

**Transformerモデル上の適応的回答予測** Transformerモデルを適応的回答予測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、これを文頭に置くことによって判別を行う手法を提案する（図1）。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.” という文の設問に正答するか否かを予測する問

Prediction of Responses of Learners with Excellent Interpretability

<sup>†</sup> Yo Ehara, Tokyo Gakugei University

<sup>\*1</sup> <https://lancers.co.jp/>

<sup>\*2</sup> <https://github.com/17zuoye/pyirt>

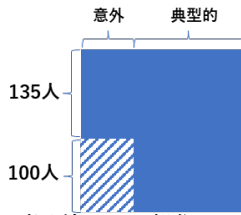


図2 実験設定. 青く塗られた部分がパラメタ推定に使われる訓練データ. 斜線部が性能比較に用いられるテストデータ.

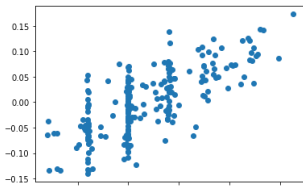


図3 IRTの能力パラメタ(横軸)と,学習者トークンの単語埋め込みベクトルの第一主成分得点(縦軸).

表3 図2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)	<b>0.688 (**)</b>
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 (*)

題とする. 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.今回は,入力文が短文であるため,学習者が1語でもわからなければ正答できない設問が多数であることから,文中のどの語に着目しているという情報はTransformerモデルでは与えない. 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. 微調整には Adam法を用い, バッチサイズは32とした.

Transformerモデルを用いた結果を,表3にまとめた.\*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. 大文字と小文字を区別するcasedなモデル(roberta-baseも含まれる)が,IRTと比較して統計的有意に予測性能が高い事が分かる. この実験結果は,設問文を考慮する事で,IRTより高精度な判別が行えることを示している.表3では,bert-base-casedが最も高い性能を示した.bert-large-casedよりも高い性能を示した理由として,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データに適合していた可能性が考えられる.

**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モデルは適応的の回答予測問題で高精度を達成しているため,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

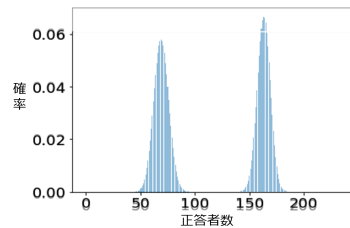


図4 正答者予測分布.

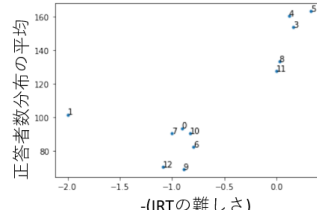


図5 提案手法とIRTの難しさパラメタの相関.

メタを比較した(図3).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.

**正答者数予測分布を通じた問題の難しさの抽出** さらに,提案手法からは,次のようにして,正答者数の予測分布の形で問題の難しさを抽出できる.それぞれの設問について,各被験者がどの程度の確率でその設問に正答するのかという確率値が取れる.各被験者が独立であると仮定すると,予測される正答者数の確率分布は,「正答確率がそれぞれ異なる,それぞれ独立な $n$ 人の試行を行ったときの正答者数の離散確率分布」となり,これは,ポアソン二項分布として知られており,動的計画法で確率分布を求められる[5].図4に難しい問題と易しい問題で,予測される正答者数の確率分布を図を提示する.易しい問題(正答者数が多い問題)は,わずかに,分散が小さくなっていることがわかる.図2における性能計測用の斜線部の回答まで埋まったデータを用いてIRTで問題の難しさを推定した.図5に,提案手法による正答者数と,IRTが全データから推定した困難度を示す(簡単のため難しさに-をつけて図示した).相関係数は0.78( $p < 0.01$ )で統計的有意に相関し,提案手法は高い解釈性を示した.

#### 4 結論

本研究では,外国語語彙学習を題材に,「学習者トークン」の導入により,BERT等の深層言語モデルを適応的学習支援に適用させられる簡便な方法を提案した.実際に,題材にした課題において,提案法は高精度な学習者の回答予測精度と,学習者の能力値や設問の難しさを,IRTが推定するものと統計的有意に推定できる高い解釈性を示した.(図5では,IRTは,提案手法が予測している図2斜線部の回答が見えている状態で推定している.)今後の課題としては,他のデータセットでも試すことがあげられる.

#### 謝辞

本研究は,JST ACT-X(JPMJAX2006)の支援を受けた.

#### 参考文献

- [1] Frank B. Baker. *Item Response Theory: Parameter Estimation Techniques, Second Edition*. CRC Press, July 2004.
- [2] David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, Vol. 31, No. 7, pp. 9–13, 2007.
- [3] 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.
- [4] Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.
- [5] Yo Ehara. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In *Proc. of EDM (poster)*, 2022.