文意を考慮する深層言語モデルを適応的学習支援に適応させる簡便な手法
Easy Adaptation of Deep Language Models to Adaptive Learning Systems

江原 遥
Yo EHARA
東京学芸大学　教育学部
Faculty of Education, Tokyo Gakugei University
Email: ehara@u-gakugei.ac.jp

あらまし： 学習反応予測のため，設問の文意を考慮し項目の難しさを考慮したい．BERT 等の深層言語モデルは文意の考慮に有用だが，能力等の学習者特性の考慮が難しい．本稿では，学習者を表す特殊な語，学習者トークンを導入し，深層言語モデルを適応的学習支援に適応させる簡便な手法を提案する．外国語語彙学習支援で評価し，提案法の高い予測性能と，提案モデルから抽出した能力値が項目反応理論の能力値と有意に相関する事を確認した．
キーワード：自然言語理解，適応的学習支援，学習者特性，学習支援システム

1　はじめに

学習支援システムにおいて，学習者が項目に回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（設問）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．

項目反応モデルは通常，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，語彙学習支援における多義語の学習支援を題材に，この問題に対処する簡便な方法を提案する．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（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，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 IRT [1] を用いた手法である．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 モデルの手法（[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 モデルを被験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 で推定した能力値とよく相関する値を抽出する手法も提案する．本研究で作成したデータセットは，今後[1]で公開する予定である．本稿の内容は EDM short paper に採択された [5]．

2　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[2]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 名の被験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，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．

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

IRT の困難度・識別力の各パラメタを求めるには，pyirt [3]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　被験者反応予測による評価

IRT を用いた手法は，被験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 モデル上の個人化判別　Transformer モデルを個人化判別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

[1] http://yoehara.com/
[2] https://lancers.co.jp/
[3] https://github.com/17zuoye/pyirt

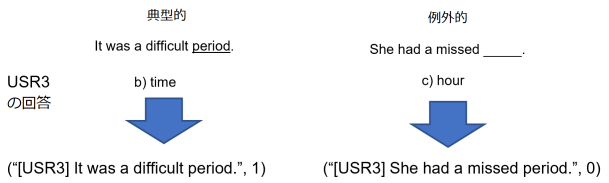

図 1: 学習者トークンの導入

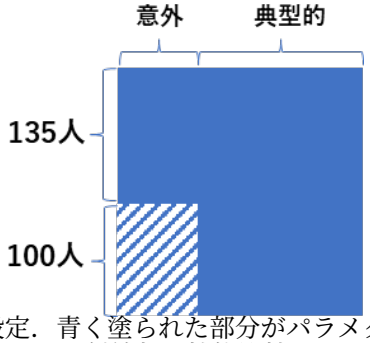

図 2: 実験設定．青く塗られた部分がパラメタ推定に使われる訓練データ．斜線部が性能比較に用いられるテストデータ．

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 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．微調整の訓練には 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．

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．

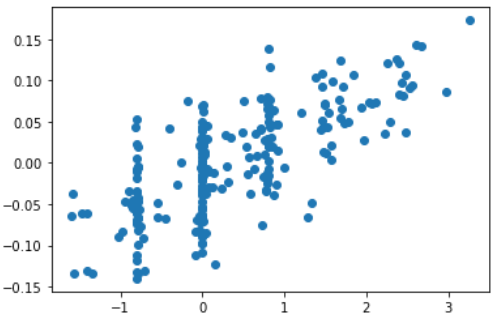

図 3: IRT の能力パラメタ（横軸）と，学習者トークンの単語埋め込みベクトルの第一主成分得点（縦軸）．

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)	**0.688** (**)
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 (*)

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．

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は，提案法は，高精度な被験者反応の予測精度と，学習者の能力値を抽出できる高い解釈性を示した．提案手法は，原理的には語彙学習支援以外にも設問の文意を考慮する必要がある一般の適応的学習支援に用いることが可能であり，今回の題材以外での実証実験が今後の課題として挙げられる．

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. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proc. of EDM (short)*, 2020.