語彙知識予測問題の概説 ～ロジスティック回帰と項目反応理論の対応関係を中心に～

江原 遥　　石川 博
首都大学東京 システムデザイン研究科
{ehara,ishikawa-hiroshi}@tmu.ac.jp

1 はじめに

本稿では，筆者がこれまで取り組んできた，語彙知識予測問題のこれまでの成果を，できるだけ数式を用いずに概説する．その際に，ロジスティック回帰と自然言語処理の対応関係に着目し，これを中心に，述べる．

語彙知識予測問題とは，語学学習者が「個々の単語」を知っているかどうかを予測する問題である[1]．具体的には，まず，数百語程度の単語テストをあらかじめ，ある学習者に受けさせ，その結果を用いて，残りの約1万語程度について，その学習者が個々の単語を知っているかどうかを当てる．学習者ごと，単語ごとに判定するのが，この問題のポイントである．

この問題は，語彙数を計測する問題の拡張になっている．語彙数は，個々の語彙を知っているかどうかを判別する事ができれば，簡単に分かる．具体的には，全語彙に対して，学習者が語を知っているかどうかを判別して，その学習者が「知っている」と判定された語の数を，その学習者の語彙数とすればよい．

このように，学習者が個々の単語を知っているかどうかを網羅的に予測したい理由は，幅広い工学応用が存在するからである．具体的には，例えば読解支援への応用として，与えられた文書の中にある全単語について，学習者がそれらの知っているかどうか判定し，学習者が知らないと判定された単語についてはあらかじめ訳をつけておくという事が考えられる．この問題は，[5] において，学習者からのフィードバックを用いて予測を改善するところまで含めて，詳細に述べられている．また，その他，作文支援への応用も容易に考えられる．例えば，学習者が外国語で作文している最中に，これまでに入力した文章から，次に学習者が書きたい単語を予測し，予測された単語がもし学習者が知らなさそうな単語であれば，例文や使い方を学習者に提示する，といった事が考えられる．

本稿では，語彙知識予測問題に関するこれまでの進展と，今後の方向性について述べる．

1.1 公開中のデータセットやツールの公開について

このような語彙知識予測問題を評価するためには，学習者がどの単語をどの程度知っているのか，という正解データを収集する必要がある．筆者は，16 人の主に日本語母語の大学院生について，1 人約 12,000 語の英単語について，5 段階でどの程度各単語を知っているかを自己申告させる方式で収集したデータを作成し，Web 上で公開している[2]．また，将来的には，予測用のコード・ツールキットも yoehara.com 上で公開する予定である．さらに，同時に，判別用のモデルも公開できる可能性が高い．これまでの筆者の研究では，有料で入手しにくい大規模コーパスからの単語頻度などを素性（特徴量）に用いて判別しているが，Python の nltk[3] ライブラリなどから無料でダウンロードすることが可能なコーパスのみを用いても，かなりよい判別精度が得られることが分かっている．

2 項目反応理論とのつながり

語彙知識予測問題は，いかにして解けば良いのだろうか．まず，直感的に分かる事は，学習者ごと，単語ごとに，その学習者がその単語を知っているかどうかを判別する問題であるので，学習者の能力，単語の難易度，といった値を，各学習者，各単語に設定してやる必要がありそうな事である．そして，学習者の能力が単語の難易度を上回っていれば，学習者が単語を知っている，そうでなければ，知らない，と判別するのである．

実は，この直観的な考え方が，まさしく，(教育）心

[1] 本稿では，第二言語学習者に対して，学習対象の言語の語彙に関する測定を行うことを前提としている．「第二言語学習者」という用語は長いので，簡単のため，語学学習者，また，単に学習者と呼ぶことにする．また，単に「語彙」といった場合，学習対象となっている言語の語彙を指すものとする．

[2] http://yoehara.com/esl-vocabulary-dataset/

[3] http://www.nltk.org/

理学の分野で広く使われている項目反応理論 [17, 18] に対応する．ある学習者が単語を知っている/知らないを，学習者が単語の意味を答える設問に正答したか/誤答したかに，それぞれ対応させて考えよう．項目反応理論は，どの学習者がどの設問に正答/誤答したか，というテスト結果のデータを元に，各学習者の能力値と各単語の難易度を，データにフィットするように自動的に決定してくれる．さらに，同じ学習者が他のテストを受けた場合や，同じテストを他の学習者が受けた場合にも，等化と呼ばれる処理を挟むことで，能力値や難易度を，テスト問題やテスト受験者集団を超えて，比較することができるようなフレームワークが提供されている．項目反応理論では，伝統的に，ここでいう「設問」の事を項目 (item) という名前で読んでいるため，このような名称になっている．

2.1 項目反応理論の種類

項目反応理論は，実際には，様々なモデルの総称であり，目的によって次のように分かれている．

1. 各項目（設問）について，難易度以外のパラメタも考えるか
2. 正答/誤答だけでなく，例えば，「全く知らない」から「よく知っている」までの５段階，といったように，多値の値を考慮するか
3. 各項目の難易度に素性（特徴量）を入れた時を，どう考えるか

このうち，筆者のこれまでの研究では，1. と 2. については扱いがよく分かっているが，3. については，後述の「能力値や難易度の信頼性」との兼ね合いがよく分かっていない．

1. については，能力値の高い学習者とそうでない学習者で正答/誤答がきれいに分かれる程度を表す「識別力」や，学習者があてずっぽうで答えた時の正答のしやすさまで考慮する「当て推量パラメタ」を考慮したモデルが考えられる．難易度だけを考慮するモデルを，**1PL モデル**または **Rasch モデル**と言い，難易度と識別力を考慮するモデルを **2PL モデル**，難易度と識別力と当て推量パラメタを全て考慮するモデルを **3PL モデル**という [18]．

このうち，**1PL モデル**（Rasch モデル）については，実は，自然言語処理分野で，予測問題を解くときに多用されるロジスティック回帰の特殊な場合である．従って，LIBSVM[4] や LIBLINEAR[5] などの機械学習ライブラリで得られたパラメータを適切に変換してやることで，学習者の能力値や単語の難易度などを計測する事が可能である [4, 5]．また，このモデルについては，対数尤度関数が凸関数になるため，大域的最適解を得やすく，初期値依存の問題が起こりにくい．

2. については，大別して，Graded Response Model[14] と，Rating Scale Model[1] の２つの考え方がある．前者は，ある能力の学習者を考えた時，各設問の難易度がどこからどこまでの範囲であれば，学習者は「よく知っている」と答える，といったような，難易度の範囲をモデル化する．後者は，多値の各分類の背後にカテゴリがあると考え，カテゴリからの難易度のずれをモデル化する．

3. については，詳述したい．語彙知識予測問題については，自然言語処理分野からすれば，予測を正確にするために，例えばコーパスからの単語頻度を素性（特徴量）に加える，ということは，ごく基本的な発想である．学習者については，学習者の情報（例えば，TOEIC テストの点数など）が事前に分かることが少ないため，情報がないと仮定してもよいが，予測精度の向上のために，語についての素性を入れたい，と思うのは当然だろう．項目反応理論の世界では，Linear Logistic Test Model (LLTM)[7, 17] と呼ばれるモデルがこれに対応するが，考え方がかなり異なる．

「語の難易度に関する素性を入れる」ということを言い換えると，次のようになる．ある２つの語の難易度を考えた時に，２つの難易度がそれぞれバラバラに（独立に）決まるのではなく，背後には，例えばコーパス上での頻度といったような隠れた変数があり，その隠れた変数の値を通して，２つの語の難易度が決まっている，ということになる．実は，単純な 1PL, 2PL, 3PL モデルでは，ある学習者が各設問を解いた時の設問に対する反応には，独立性が仮定されている（局所独立性 [18]）．一方，語の難易度に素性を入れると，ある２つの語の背後は素性でつながっているため，この独立性は崩れる．項目反応理論の考え方では，「素性を入れる」ということは，局所独立性の仮定をゆるめたモデルを考えていることになる．

従って，LLTM では，いわゆる自然言語処理で多用されるロジスティック回帰における「素性」に対応する値は LLTM モデルに入れることができるが，あくまで，難易度は学習者の反応データだけから計測されることが前提になっていて，素性に対応する値には，どの変数とどの変数が独立か，といったような構造的な値を入れることが想定されている．

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvm/
[5] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

このように，モデルはほぼ同じでも解釈の仕方が異なるので，次節の「能力値や難易度の信頼性」を考えた場合に，LLTM でも同様に信頼性の議論ができるのかが，現時点での筆者にはよく分かっていない．この辺りを明確にすることが，今後の方向性であると筆者は考える．

LLTM の詳細については [17, 9, 8] が詳しい．実際の計算は，R では，1PL, 2PL, 3LP モデルなどは ltm パッケージで行える[6]．また，LLTM などは，eRm パッケージ [9, 8] で行える．

2.1.1 能力値や難易度の信頼性

さて，こうして求めた学習者の能力値や，項目の難易度といった情報は，どの程度信頼できるのだろうか．この「どの程度信頼できるか」を表す指標には，例えば，項目の難易度の値の分散といったように，様々な指標が考えられる．しかし，中でもテストを分析する上で重要性が高いのは，「ある項目に正答したかどうかで，学習者の能力がどの程度正確に測れるのか」という指標なのではないだろうか．

よく考えてみると，項目の難易度の信頼性は，学習者の能力に依存するはずである．例えば，英語の語彙知識予測問題であれば，“dog” や “cat” というような単語であれば，英語の初学者と全く英語を知らない学習者を見分けるには有効かもしれないが，大半の英語学習者にとっては簡単過ぎる設問であるため，中級英語学習者と上級英語学習者を見分ける設問としては不適切であるように思われる．このように，項目の難易度が，学習者の能力を見分ける上で，どの程度役立っているのかは，設問によって異なる．

項目反応理論は，この問題に対しても，項目の難易度の信頼性を学習者の能力ごとにはかる，**項目情報量**と**テスト情報量**というフレームワークを提供している．図 1 を用いて説明する．図 1 は，実際に，「はじめに」で説明した公開中のデータセットのうち，100 単語をランダムに選んで，R の “eRm” パッケージ[7] を用いて，項目情報量とテスト情報量を求めたものである．

図 1 上側の各曲線は，各項目の項目情報量を表している．横軸が学習者の能力であり，0 が平均の能力値で，大きければ大きいほど，能力値が高い，すなわち，優秀であることを示している．縦軸が，「学習者の能力をどの程度正確に測れるか」を表す値で，大きければ大きいほど正確に測れる．より正確にいえば，縦軸の値は，学習者の能力値の信頼区間の幅に関係していて，この数値が大きければ大きければ大きいほど，信頼区間の幅が狭くなる [18]．すなわち，ある能力値の学習者の能力を正確に計測することができるようになる．

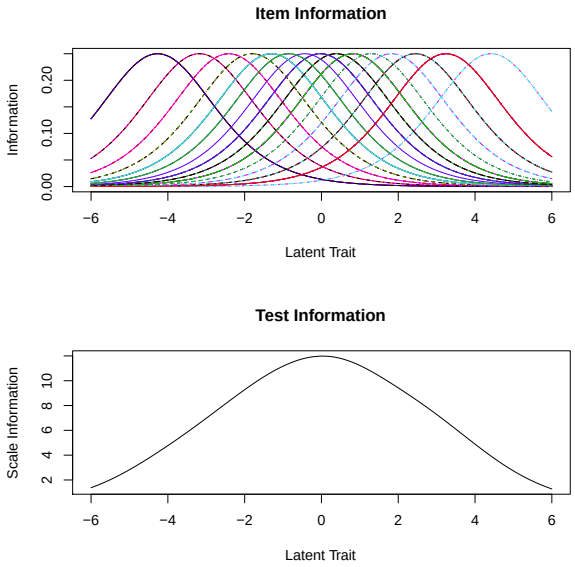

図 1: 1PL モデルによる，テスト情報量．上側の図が項目情報量，下側の図がそれらを足しこんだテスト情報量．横軸が学習者の能力．

例えば，図 1 上側の図で，横軸の能力値が 4 あたりのところでは，いくつかの曲線を除いて，ほとんどの曲線の縦軸の値が低い．これは，高すぎる能力値を持つ学習者の能力は，いくつかの難しい設問以外は正確に計測することができないということを表している．横軸の能力値の値が −4 あたりの所も，同様に，いくつかの曲線を除いて，ほとんどの曲線の縦軸の値が低くなっている．これは，低すぎる能力値を持つ学習者にとっては，いくつかの簡単な設問以外は難しすぎ，かんたんな設問でないとこれらの学習者の能力値を正確に計測することができないということを表している．

図 1 の上側の全ての曲線を足しこんだものが，図 1 の下側の図である．これは，**テスト情報量**と呼ばれ，テスト全体について，学習者の能力値をどの程度正確に計測できるかを表したものである．より専門的には，テスト情報量は，フィッシャー情報量と同じものであり，能力値の推定量の分散を，テスト情報量の逆数が下から抑える（Cramer-Rao の不等式）．

通常，テスト情報量は，図 1 のように，学習者の能力値の平均値を中心にして釣り鐘状のグラフになる．これは，簡単にいえば，やはり，能力値が極端な学習者については，うまく計測できない事を表す．図 1 は，公開中のデータセットを用いた例であり，テスト情報

[6] https://cran.r-project.org/web/packages/ltm/ltm.pdf
[7] https://cran.r-project.org/web/packages/eRm/eRm.pdf

量関数を見る限りでは，このデータセットは 1PL にうまくフィットしている事がわかる．

3 関連研究

3.1 自然言語処理分野においてパラメタの解釈

本稿は，端的にいえば，語彙知識予測問題を解く時のパラメタの解釈を，項目反応理論の知見を用いて，より，信頼性高く厳密なものにできないか考えよう，と提議をしている．

このような提議に至った背景には，自然言語処理分野において，予測モデルの重みパラメタの大小が，対応する素性の重要性として安易に解釈されすぎているのではないかという疑念がある．言語教育からは離れるが，例えば，[6] では，サポートベクトルマシン [16] の重みパラメタの値を，そのまま素性の重要度として分析し，知見を得ようとしている．これは，知見を得るための第一歩としてはあり得るが，重みパラメタの値は，予測モデルのパラメタの値は，使ったデータセットやデータ量によっても容易に変わり得るので，重みパラメタの値の大小がどの程度たしからしいかについては，より，厳密な議論が行える余地がある．

教育心理学の分野で発展してきた項目反応理論は，得られたパラメタを厳密に解釈するという点においては優れた枠組みであるが，自然言語処理分野のロジスティック回帰を使用する際に活用されているとは言いがたい．この点で，項目反応理論は自然言語処理に貢献できると筆者は考える．

3.2 語彙数計測の研究

学習者の知っている語彙「数」を計測する問題は，外国語教育学の研究分野において，長年研究されてきた [13, 10]．

テストの形式で大別して，２つの流れがある．まずは，Nation らによる，多肢選択式のテストが挙げられる．これは，広く知られている，複数の選択肢の中から，正しい意味を選び出すものである．次に，各単語について「知っている」「知らない」の２値を自己申告式でつける YES/NO テストがある [10]．この方法では，実際には単語ではない文字列をテスト中に混ぜておくことにより，信頼性を担保する．

どちらの方式でも，単語難易度は，基本的には，BNC コーパス [15] のような大規模コーパスの単語頻度を元にしている．ただし，前者の場合，特に簡単な単語については人手で難易度の調整がされている [12]．

多肢選択式では，近年，単語難易度をコーパスからの単語頻度と，Rasch モデルによって計算された単語難易度を比較し，項目反応理論の枠組みで議論する研究が出てきている [2]．

また，計測された語彙数と，読解能力の関係については，[11] が詳しい．

3.3 これまでの研究

語彙知識予測問題について，筆者は，これまでに，大別して，３つの研究を行っている．

まず，語彙知識予測問題を読解支援に適用し，実際に，シミュレーション実験を通じて，読めない文書が有意に減少する事を確認した [5]．次に，単語の難易度は，学習者の専門などによっても違うので，項目反応理論の Rasch モデルを発展させ，学習者にとっての単語の難易度が計算できるモデルを提案した [4]．最後に，項目反応理論との繋がりは薄いが，単語テストに用いる単語を選び出すときに，単語間の意味的関係性をも考慮することにより，語彙知識予測問題の予測性能を向上させる手法を提案した [3]．

4 おわりに

本稿では，語彙知識予測問題についての，これまでの研究成果を概観し，今後の進展の方向性を示した．今後の課題としては，項目反応理論の観点からのテスト情報量が，自然言語処理で広く使われている「素性を使う」という設定で，どの程度信頼できるか，どのように解釈すればよいか，という問題が挙げられる．

謝辞

本研究は，JSPS 科研費 15K16059 の助成を受けた．

参考文献

[1] D. Andrich. A rating formulation for ordered response categories. *Psychometrika*, Vol. 43, No. 561–573, 1978.
[2] David Beglar. A rasch-based validation of the vocabulary size test. *Language Testing*, Vol. 27, No. 1, pp. 101–118, 2010.
[3] Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proc. of EMNLP*, pp. 1374–1384, October 2014.
[4] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proc. of COLING*, December 2012.
[5] Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, Vol. 4, No. 2, 2013.
[6] Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. In *Proc. of EMNLP*, pp. 1753–1764, October 2013.
[7] Scheiblechner H. Das lernen und losen komplexer denkaufgaben. [the learning and solving of complex reasoning items.]. *Zeitschrift fur Experimentelle und Angewandte Psychologie*, Vol. 3, pp. 456–506, 1972.
[8] P. Mair and R. Hatzinger. Cml based estimation of extended rasch models with the erm package in r. *Psychology Science*, Vol. 49, No. 26–43, 2007.
[9] P. Mair and R. Hatzinger. Extended rasch modeling: The erm package for the application of irt models in r. *Journal of Statistical Software*, Vol. 20, No. 9, pp. 1–20, 2007.
[10] Paul M. Meara. *EFL Vocabulary Tests (Second Edition)*. Lognostics (Center for Applied Language Studies, University of Wales), Swansea, 2010.
[11] Paul Nation. How large a vocabulary is needed for reading and listening? Vol. 63, No. 1, pp. 59–82, 2006.
[12] Paul Nation. The bnc/coca word family lists, September 2012.
[13] Paul Nation and Robert Waring. Vocabulary size, text coverage and word lists. In *Vocabulary: Description, acquisition and pedagogy*, pp. 6–19. Cambridge: Cambridge University Press, 1997.
[14] F. Samejima. Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs*, Vol. 34, No. (Suppl. 4), 1969.
[15] The BNC Consortium. The british national corpus, version 3 (bnc xml edition), 2007.
[16] 竹内一郎, 烏山昌幸. サポートベクトルマシン (機械学習プロフェッショナルシリーズ). 講談社, 2015.
[17] 豊田秀樹. 項目反応理論・理論編—テストの数理 (統計ライブラリー). 朝倉書店, 2005.
[18] 豊田秀樹. 項目反応理論 [入門編](第 2 版) (統計ライブラリー). 朝倉書店, 2012.