学習者に特化した単語難易度の提案

江原遥 †，佐藤一誠 ‡，大岩秀和 †，中川裕志 ‡
† 東京大学大学院情報理工学系研究科，‡ 東京大学情報基盤センター

{ehara,sato,oiwa}@r.dl.itc.u-tokyo.ac.jp, n3@dl.itc.u-tokyo.ac.jp

1　導入

グローバル化に伴い，近年では，英語とはじめとする第二言語（外国語）の習得は益々必須となり，第二言語を用いたコミュニケーションの量は言語研究の上で決して無視出来ない程にまで増大している．自然言語処理分野においても，言語処理システムの利用者の第二言語を用いたコミュニケーションを支援する言語教育は，益々重要なテーマとして認知されつつある．

第二言語の学習においては，文法や発音以上に，語彙の習得も重要であることが分かってきている．第二言語学習者の語彙知識については，語彙数の推定 [21, 15, 18] や，学習者が習得すべき語彙の選別 [19] などが行われているが，学習者が実際にどのような語を知っているのかを具体的に予測する研究は，我々の知る限りない．

本研究では，学習者が知っている語を予測するタスクである，**語彙予測タスク**を扱う．このタスクは，理論上も応用上も重要である．理論上の意義としては，学習者が語彙を習得していく仮定上の傾向を把握し，第二言語の語彙習得論で論じられるような理想的な語彙習得を実際に行なっているかどうかの検証に役立てられる．応用上では，我々の既存研究 [4, 5] において，このタスクを Web 文書の読解支援に直接応用し，その有用性を実証している．

語彙予測タスクでは，モデルの予測性能に加えて，モデルの分析上の有用性も重要である．本研究では，語彙予測タスクにおいて重要となるモデルの性質を下記のように定めた．

解釈可能なパラメタ　モデルのパラメタを，一定の計算式を用いて**学習者の能力値**と**単語難易度**として解釈できること．例えば，Support Vector Machines (SVM) は確率値を出力しないため，この性質を満たさないと考えられる．

サンプル外設定　訓練データ中にない単語がテストデータに出現しても対応できること．

学習者に適応した単語難易度　語の難易度が全学習者にとって同じスケールであるという設定は非現実的である．例えば，能力値の低い学習者でも，音楽に強い興味を持っていれば，音楽に関する稀な語を知っていることはあり得る．従って，単語難易度も，学習者ごとに求まることが望ましい．

表 1: 提案モデルと既存モデルの比較．提案モデルのみ，全ての性質を満たす．

	解釈可能なパラメタ	サンプル外設定	学習者に適応した単語難易度
Rasch	✓	-	-
Ehara et al., 2010[4]	✓	✓	-
提案モデル	✓	✓	✓

図 1: (a) in-matrix 設定, (b) out-of-sample 設定.

表 1 に，提案モデルと既存モデルを比較した．提案モデルのみが全ての性質を満たす．

本論文の貢献を下記にまとめる．

- 語の難易度パラメータの一般形を導入した．
- 提案モデルは，このタスクのモデルに求められる性質を全て満たしている．
- 実験において，既存のモデルでは見ることが出来なかった第二言語学習者の特徴を捉えられた．

2　問題設定

U を学習者の集合，V を語彙の集合とし，学習者の数を $|U|$，語彙の数を $|V|$ と表す．1 つのデータは $y \in Y$，$u \in U$，$v \in V$ を用いて (y, u, v) の 3 つ組の形で表せる．ここで Y は語彙知識の程度を表す順序集合とし，本稿では簡単のため $Y = \{0, 1\}$ とする．訓練データを $\mathcal{D} = \{(y_1, u_1, v_1), \ldots, (y_N, u_N, v_N)\}$ で表す．

問題設定には，図 1(a),(b) にそれぞれ図示するように，*in-matrix*（サンプル内設定）と *out-of-sample*（サンプル外設定）の 2 種類がある．図 1 では，斜線部が訓練データを表し，空欄がテストデータを表す．テストデータ中の語に対する被験者の反応全てを予測しなければならないため，*out-of-sample* の方が *in-matrix* より予測が難しい設定である．

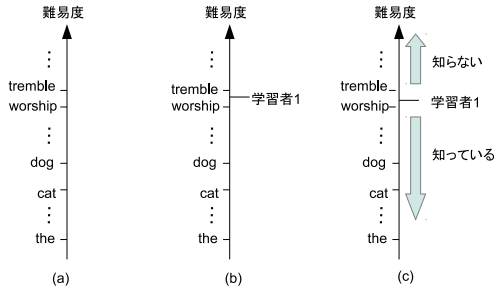

図 2: Rasch モデルによる予測の仕組み.

3 一般形

本節では，本稿で説明するモデルを全て特殊な場合として表現可能な一般形を導入する．この一般形は，尤度関数の一般形は式 (1) で与えられる．

$$P\left(y=1|u,v\right)=\sigma\left(a_u-f\left(u,v\right)\right). \tag{1}$$

$f(u,v)$ 語 v の学習者 u にとっての難易度を返す関数,

a_u 学習者 u の能力パラメタ.

Rasch モデルは，式 (1) で，$f(u,v)=d_v$ と置いた場合に相当する．ここで，式 (1) は Rasch モデルの尤度関数のみの一般形であることに注意されたい．

また，式 (1) において，$f(u,v)=\mathbf{w}^\top\phi(v)$ と置くことにより，難易度共通モデルが導入できる．ここで，$\phi: V\rightarrow\mathcal{R}^K$ である．難易度共通モデルは，[8] とほぼ同様のモデルであり，語彙判別タスクにおいては [4, 5] で用いられている．Rasch モデルは，難易度共通モデルで特殊な ϕ を設定したモデルと等価である．難易度共通モデルは，Rasch モデルと異なり，ϕ を適切に設定することにより *out-of-sample* 設定に適用可能である．新語がテストデータに現れたとしても，その新語の特徴量に対応する重みパラメタが訓練データ中で学習されていれば，$\mathbf{w}^\top\phi(v)$ によって単語難易度が計算できることによる．

Rasch モデルと難易度共通モデルの事前分布は表 2 にまとめた．

ところで，式 (1) は単語難易度を $f(u,v)$ の形に一般化した形であるが，学習者の能力パラメタ a_u を一般化して，学習者に関する様々な素性を入れたモデルも考えることが出来る．しかし，学習者については多くの情報が得られるとは限らないため，そのようなモデルの有用性は限定的と言えよう．

4 提案モデル

Rasch モデルと難易度共通モデルに共通する問題点として，全ての学習者が同じ単語難易度を共有するという点があげられる．例えば，図 3(a) においては，“tremble” という語は常に `` worship” という語より難しいと扱われる．即ち，全学習者が共通の難易度を用いているため，語の難易度が学習者に適応的でないと言える．実際には，能力の低い学習者が，様々な要因によって難しい単語を知っている事例は，ありふれたものである．例えば，音楽好きな学習者は，音楽分野の語彙については，語が難しくても知っている可能性が高い．このような，学習者毎の特性をモデルに組み入れることは，第二言語学習者に適応的な支援システムを設計する上でも必要不可欠である．

前述の**一般形**から，この問題点の根本的な原因は，Rasch モデルでも難易度共通モデルでも，単語難易度 $f(u,v)$ の語 v にのみ依存し，学習者 u に依存していないためであると分かる．従って，学習者に適応した単語難易度を導入するためには，$f(u,v)$ を u に依存させればよい．提案手法では，$f(u,v)=\mathbf{w}_u^\top\phi(v)$ とした．

$$P\left(y=1|u,v;\mathbf{w}_u\right)=\sigma\left(a_u-\mathbf{w}_u^\top\phi\left(v\right)\right). \tag{2}$$

難易度共通モデルでは，全ての学習者が共通の単語難易度を用いているので，図 3(a) の例では，“tremble” が “worship” より難しいと仮定される．従って，“tremble” を知っているが “worship” は知らないような特異的な学習者が存在しても，これをモデル化することはできない．一方，図 3(b) に示す適応難易度モデルでは，学習者ごとに単語難易度を作るため，この様な特異的な学習者（ここでは仮に学習者 2 とする）に対しても,「“worship” が “tremble” より難しい，学習者 2 にとっての単語難易度」を作ることで，モデル化することが可能となる．

5 パラメタ推定

提案モデルのパラメタは，式 (3) の最小化で表される Maximum-a-posteriori (MAP) 推定によって求められる．

$$\begin{aligned} l\left(\mathbf{W},\mathbf{a},\mathbf{w_0}\right) &= \sum_{i=1}^{N} nll\left(y_i,u_i,v_i\right)+\frac{\lambda}{2}\sum_{u\in U}\left\|\mathbf{w}_u-\mathbf{w}_0\right\|^2 \\ &+ \frac{\eta_w}{2}\left\|\mathbf{w}_0\right\|^2+\frac{\eta_a}{2}\sum_{u\in U}a_u^2 \quad . \end{aligned} \tag{3}$$

ただし，$nll\left(y,u,v\right)\stackrel{\text{def}}{=}\log\left(1+\exp\left(-y\left(a_u-\mathbf{w}_u^\top\phi\left(v\right)\right)\right)\right)$ とし，$\mathbf{W}\stackrel{\text{def}}{=}\{\mathbf{w}_u|\forall u\in U\}$，$\mathbf{a}\stackrel{\text{def}}{=}\{a_u|\forall u\in U\}$ とする．式 (3) の関

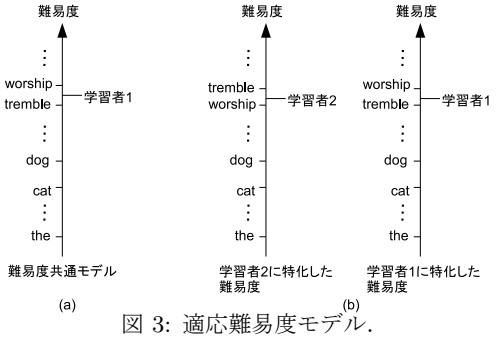

図 3: 適応難易度モデル.

表 2: 各モデルのまとめ．$\mathcal{N}$ は，適切な次元のガウス分布の確率分布関数．

名称	f の形	事前分布	備考
Rasch	$f(u,v)=d_v$	$P(a_u\|\eta_a)=\mathcal{N}(0,\eta_a^{-1})$ $P(d_v\|\eta_d)=\mathcal{N}(0,\eta_d^{-1})$	-
難易度共通モデル [4]	$f(u,v)=\mathbf{w}^\top\phi(v)$	$P(a_u\|\eta_a)=\mathcal{N}(0,\eta_a^{-1})$ $P(\mathbf{w}\|\eta_w)=\mathcal{N}(\mathbf{0},\eta_w^{-1}I)$	$\phi(v)$ が $\phi(v)=\mathbf{1}$ なる $\|V\|$ 次元ベクトルのとき，Rasch モデルに一致．
提案モデル	$f(u,v)=\mathbf{w}_u^\top\phi(v)$	$P(a_u\|\eta_a)=\mathcal{N}(0,\eta_a^{-1})$ $P(\mathbf{w}_0)=\mathcal{N}(\mathbf{0},\eta_w^{-1}I)$ $P(\mathbf{w}_u\|\mathbf{w}_0)=\mathcal{N}(\mathbf{w}_0,\lambda^{-1}I)$	$\mathbf{w}_u=\mathbf{w_0}$ $(\forall u\in U)$ の時，難易度共通モデルに一致．

数 l は，$\mathbf{W},\mathbf{a},\mathbf{w_0}$ の全ての変数について凸である [12] ため，式 (3) は大域的最適解を持つ．[12] に基づき，次の繰り返し操作を収束するまで，この大域的最適解を求める．また，表 2 に示す通り，Rasch モデルや難易度共通モデルも提案モデルで $\mathbf{w}_u=\mathbf{w_0}$ $(\forall u\in U)$ と置いた場合の特殊ケースである．従って，上記の前者の最適化のみで MAP 推定を行うことが可能である．

$\mathbf{W}$, $\mathbf{a}$ に関する最小化 $\mathbf{w}_0$ を固定し，$\mathbf{W}$ と $\mathbf{a}$ に関して $l(\mathbf{W},\mathbf{a},\mathbf{w}_0)$ を最小化する．[12] は Newton 法を用いたが，これは，$\mathbf{w}_0$ の次元数を K とすると，空間計算量が $O(K^2)$ であり，K が大きい時に問題となる．本稿では，$O(K)$ の空間計算量で済む L-BFGS[16] 法を，libLBFGS[20] を使用することによって用いた．

$\mathbf{w}_0$ に関する最小化 $\mathbf{W}$ と $\mathbf{a}$ を固定し，l を $\mathbf{w}_0$ に関して最小化する．この操作は，次のように解析的に実行できる．

$$\mathbf{w}_0=\frac{\lambda}{\eta_w+|U|\lambda}\sum_{u\in U}\mathbf{w}_u \quad . \tag{4}$$

6 評価

6.1 Dataset

[4] で公開したデータセットを用いた．このデータセットは 2009 年 1 月に日本で作成され，東大院生を中心とした 16 人が参加し，うち 15 人のデータを用いた．1 人 11,999 語について 5 段階で英単語を知っている度合いを付けてもらい，そのうち 5 のみを「知っている」場合とし，残りを「知らない」場合とした．

各コーパス[1] から，$-\log$(1-gram 確率) を素性として用いた．超パラメタは 5 交差検定で選択した．Rasch モデルと難易度共通モデルの η_d，η_a と η_w は，$\{0.01,2^{-3},2^{-2},2^{-1},1.0,2^1,2^2,2^4\}$ から，提案モデルの η_a，η_w と λ は，$\{2^{-2},2^{-1},1.0,2^1,2^2\}$ から選んだ．

6.2 単語難易度の学習者特化性の評価

提案モデルが既存モデルと最も異なる点は，学習者に特化した単語難易度が算出可能な点である．単語難易度がどの程度学習者に特化しているかを測るための指標として，学習者に対する単語難易度の分散である**学習者特化性**を次の $Var(v)$ で定義する．既存のモデルでは，全ての学習者で共通の単語難易度を用いているため，常に $Var(v)=0$ であることに注意されたい．ここで，$Mean(v)\stackrel{\text{def}}{=}\frac{1}{|U|}\sum_{u\in U}f(u,v)=\frac{1}{|U|}\sum_{u\in U}\hat{\mathbf{w}}_u^\top\phi(v)$ である．

$$Var(v)\stackrel{\text{def}}{=}\frac{1}{|U|}\sum_{u\in U}\left(\hat{\mathbf{w}}_u^\top\phi(v)-Mean(v)\right)^2. \tag{5}$$

表 3: 学習者特化性 $Var(v)$ の高い順から 30 語のリスト．

$Var(v)$	語	学習者特異性が高い理由
0.993	twitter	商品名
0.886	waltz	ドメイン依存: 音楽，母語で借用語
0.849	kindle	商品名
0.833	rink	母語で “link” と同音語
0.827	launder	母語で借用語
0.825	bass	ドメイン依存: 音楽
0.823	ultraviolet	ドメイン依存 : 化粧品
0.818	chime	ドメイン依存: 音楽
0.804	asphalt	母語で借用語
0.802	harry	母語で “hurry” と同音語
0.793	wooded	-
0.776	mantle	母語で借用語
0.767	trombone	母語で借用語
0.766	modulate	ドメイン依存 : 計算機
0.763	homeroom	母語で借用語
0.760	harness	-
0.760	bog	-
0.755	hearth	“health” と混乱
0.750	convent	-
0.748	hurdle	母語で借用語
0.733	parson	母語で “person” と同音語
0.732	vector	母語で借用語
0.731	haven	母語で “heaven” と同音語
0.719	gadget	母語で借用語
0.714	lizard	-
0.713	smelt	英語中で “smell” の過去分詞と同音語
0.709	shin	母語で “sin” と同音語
0.708	placebo	母語で借用語
0.707	lagoon	-
0.702	aha	-

学習者特異性の高い順に 30 語を表 3 にまとめた．ただし，*in-matrix* 設定で，177,985 件を訓練，2,000 件をテストに用い，精度は 83.40%であった．

表 3 で最も注目すべき点は，“twitter” という商品名が表中の 1 位に来ていることである．これは，マイクロブログサービスである Twitter の影響であると思われる．“twitter” 自体は稀な英単語である．実際，BNC では “the” の頻度が 6,043,900 であるところ，“twitter” の頻度はわずか 17 である．BNC 中で同じ頻度の語としては，“abet”，“beguile”，“coddle” が挙げられる．これらの語も収集したデータセットに含まれているにも関わらず表中に現れていないことから，単語の頻度だけでは “twitter” が上位に来ていることを説明できない．考え得る妥当な説明としては，平均的には能力の低い学習者が，サービスとして Twitter を通じて，“twitter” という単語に対して「知っている」と答えたのであろう．データは，サービスとしての Twitter を知っている人間が今より少なかった 2009 年 1 月に日本で取られた．同様の

[1] British National Corpus (BNC)[22], American English (COCA)[3],Open American National Corpus (OANC) [11], Brown corpus [10], Google 1-gram [2].

例と考えられるのが “kindle” である．最初の Amazon Kindle は 2007 年に米国で発売されている．

このように，表 3 中の語について，表中で上位に来ている理由を推察し，下記にまとめた．もちろん，この分析は推測に過ぎないが，このような分析は検証が非常に難しいことにも留意されたい．通常，第二言語学習者は，いかなる理由でその語を知っているかまでについては記憶していないからである．下記の理由付けにおいては，日本語を母語とする英語学習者である著者自身の直観に依った．理由付け出来なかった語については “-” をつけた．

商品名 “twitter” や “kindle” がこの場合に相当する．語が英語としては難しくとも，有名な製品に使われていれば，低能力の学習者でも知っていることがあり得る．

母語で借用語 英単語が，日本語における借用語になっている場合，学習者の回答が分かれることが考えられる．日本語としての意味は知っているが，英語としての意味には確証を持てない学習者が，否定的な回答をするためであると考えられる．

母語で同音語 提示語より簡単で，学習者の母語の発音では同音になるような英単語が存在する場合には，低能力で注意深くない学習者が簡単な方の単語と混同するため，$Var(v)$ が大きくなると考えられる．例えば，表 3 の “rink” や “parson” は，それぞれ，日本人学習者が “link” や “person” といった簡単な語と混同した可能性が高い．

ドメイン依存 低能力の学習者でも，特に詳しいドメインがある場合，そのドメインに属する語は知っていることがあり得る．

英語中で同義語がある “smelt” は「精錬する」という動詞であるが，“smell” の過去分詞としての形の方が広く知られているため，後者と混同して肯定的に答えた学習者がいることが示唆される．

学習者が付けたラベルである y の分散からは，学習者の能力の高低は考慮されていないため，このような有意義な分析は不可能であることに注意されたい．例えば，今回の実験では，11,999 語中，1,408 語が 2 値化後の y の分散が最も高い語であった．また，表 3 とは反対に，最も小さい $Var(v)$ を持つ語は，全員が知っている語又は全員が知らない語であった．

最後に，サンプル外設定での精度を計測した．11,999 語中の 2,000 語をテスト，残りを訓練データとした．Rasch モデルは 66.32%，難易度共通モデルは 77.67%，提案モデルは 77.81% であった．

7 関連研究

提案モデルは，[6] や [12] と非常に近いが，目的が異なる．[6] はマルチタスク学習，[12] はクラウドソーシングを目的としており，当然ながら，Rasch モデルとの関係や難易度の一般形などについては述べられていない．

本稿では第二言語学習者にとっての単語難易度を拡張したが，一方で，単語難易度は第一言語の語彙発達の研究でも重要であり，この方面の拡張も計算言語学の分野で扱われている [14]．第一言語の語彙を網羅的に計測した研究として，[1] が挙げられる．また，計算言語学分野では，文書のリーダビリティを予測する研究が広く行われており [9, 7, 13]，本稿で扱った語彙知識とリーダビリティの関連は [19] で述べられている．第二言語習得論においては，第二言語学習者の持つ語彙の「サイズ」を測定する研究は広く行われている．主に，多肢選択式の [18] と，*Yes/No* 式の [17] に分けられる．

8 結論

本稿では，学習者に特化した単語難易度を用いるモデルを提案し，従来不可能であった学習者特化性の分析を可能とし，予測性能も既存モデルと同等以上であると確認した．今後の課題としては，順序変数への拡張や，超パラメタの自動最適化が挙げられる．

参考文献

[1] S. Amano and T. Kondo. Estimation of mental lexicon size with word familiarity database. In *Fifth International Conference on Spoken Language Processing*, 1998.

[2] T. Brants and A. Franz. Web 1T 5-gram Version 1, 2006. LDC2006T13.

[3] M. Davies. N-grams data from the corpus of contemporary american english (coca), 2011. Downloaded from `http://www.ngrams.info` on June 23, 2012.

[4] Y. Ehara, N. Shimizu, T. Ninomiya, and H. Nakagawa. Personalized reading support for second-language web documents by collective intelligence. In *Proceedings of the 15th international conference on Intelligent user interfaces (IUI 2010)*, pp. 51–60, Hong Kong, China, 2010. ACM.

[5] Y. Ehara, N. Shimizu, T. Ninomiya, and H. Nakagawa. Personalized reading support for second-language web documents. *ACM Transactions on Intelligent Systems and Technology*, 4(2), 2013.

[6] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2004)*, pp. 109–117. ACM, 2004.

[7] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010): Posters*, pp. 276–284, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[8] G. Fischer. Logistic latent trait models with linear constraints. *Psychometrika*, 48(1):3–26, 1983.

[9] T. François and C. Fairon. An “ai readability” formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pp. 466–477, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[10] W. N. Francis and H. Kucera. Brown corpus manual, 1979. Brown university, Rhodes island, third edition.

[11] N. Ide and K. Suderman. The open american national corpus (oanc), 2007. Corpus available at `http://www.AmericanNationalCorpus.org/OANC/` (Retrieved on October 24, 2012).

[12] H. Kajino, Y. Tsuboi, and H. Kashima. A convex formulation for learning from crowds. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI-2012)*, pp. 73–79, Tronto, Ontario, Canada, July 2012.

[13] R. Kate, X. Luo, S. Patwardhan, M. Franz, R. Florian, R. Mooney, S. Roukos, and C. Welty. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 546–554, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[14] K. Kireyev and T. K. Landauer. Word maturity: Computational modeling of word knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pp. 299–308, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[15] B. Laufer and P. Nation. A vocabulary-size test of controlled productive ability. *Language testing*, 16(1):33–51, 1999.

[16] D. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

[17] P. Meara and B. Buxton. An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2):142–154, 1987.

[18] I. S. P. Nation. *Teaching and Learning Vocabulary*. Heinle and Heinle, Boston, MA, 1990.

[19] I. S. P. Nation. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1):59–82, 2006.

[20] N. Okazaki. *libLBFGS: L-BFGS library written in C*, 2007. Software available at `http://www.chokkan.org/software/liblbfgs/` (Retrieved on October 24, 2012).

[21] N. Schmitt, D. Schmitt, and C. Clapham. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, 18(1):55–88, 2001.

[22] The BNC Consortium. The british national corpus, version 3 (bnc xml edition), 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: `http://www.natcorp.ox.ac.uk/` (Retrieved on October 26, 2012).