修得語彙量の分布の考慮による付随的語彙学習に適した読解用テキストの選択

江原遥
静岡理工科大学　情報学部
i@yoehara.com

1　はじめに

応用言語学分野において，外国語の語彙獲得における付随的学習 (incidental learning) とは，「語彙習得が主目的ではない活動の中で，偶発的に語彙が習得されること」([1]，p.21）である．例えば，テキスト中に出てきた分からない単語の意味を推測しながら外国語の語彙を学習する手法がこれにあたる．付随的学習に対して，「語彙習得を主目的とする活動の中で，語彙を学習すること」([1]，p.21）を「意図的学習」（intentional learning）と呼ぶ．例えば，いわゆる単語帳を利用する方法など，単語リストを記憶することで外国語の語彙を獲得する方法がこれにあたる．付随的学習は意図的学習に比べ，外国語の語彙量を増加させるには非効率である一方，文脈の中の語の使われ方など深い理解を促すとされている．

効率的な付随的学習のためには，学習者にあったテキストを用いることが重要である．既習語のみからなるテキストでは，新たな語彙修得は起こらない．一方，理解できない単語が多すぎれば，テキストの読解そのものに失敗するリスクが高く，付随的学習そのものが起こりづらい．[1] では [2] を参考に，「目安として、テキスト中の単語のうち、95%〜98%が既知語であれば、文脈から自然に語彙を学習することが可能になると言われています」と報告されている（p.22）．言い換えれば，**既知語比率が *95%〜98%* に入っていれば，テキスト中の既知語以外の語（未習語）を効率的に付随的学習できることが応用言語学分野における知見として示されている**．この結果は，自然言語処理分野で応用言語学分野の成果として引用されることの [3] ことの多い「既知語比率がこの範囲に入っていればテキストが読解できる」[4] という結果よりも，さらに踏み込んだ語彙修得に直結する結果である．

応用言語学分野でのこの結果は，**効率的な付随的学習のためのテキスト選択が投機的な性質を持つ問題である事を示している**．既知語比率の小さいテキストを選べば付随的学習による修得語彙量は増えるだろうが，既知語比率が閾値を下回ると，学習者がテキストを読み進めることができず，付随的学習が起こらず，修得語彙量が著しく低くなる．従って，戦略としては，**学習者が読解に失敗するリスクを一定に抑えながら，その中で，修得語彙量が最も多そうなテキストを選択する事が求められる**．

応用言語学分野では，この戦略を，学習者やテキストの全体的な傾向に基づいて実現している：具体的には，「学習者は均衡コーパスでの単語頻度の高い順に語を知っている」という仮定を置くことによって，学習者やテキストをレベル分けし，学習者に合ったテキストを選択している．こうした手法は全体的な傾向はつかめてはいるものの，個々の学習者やテキストの特性を考慮できない．また，修得語彙量がどの程度になるかの予測もできない．

一方，自然言語処理分野やデータマイニング分野では，こうしたテキスト選択の問題はあまり扱われていない．しかし，語学学習者の既知語判定の問題については，読解支援やテキスト簡単化の観点から，「レベル」という一次元的な尺度に頼らず，個々の学習者の特性を考慮した精緻な数理モデル化がなされている [5, 6, 7, 8, 9]．例えば，どの学習者がどの単語をどの程度の確率で知っているかを，個々の学習者の得意分野，単語の意味，人間の記憶の長さ，といった様々な点を考慮して，かなり正確に予測できるようになってきている [10, 11, 12, 13]．

本稿では，前述の戦略を機械学習の観点から数理的に定式化し，ある学習者があるテキストを読んだときの付随的学習による修得語彙量の確率分布を求める手法を提案する．これにより，学習者の特性やテキストの特性を考慮し得る，付随的学習のためのテキスト選択手法を可能にする．

2 問題の例

動機づけのため，例を挙げることで，できるだけ平易に解こうとしている問題を説明する．[a,b,c,d] の 4 種の単語からなるテキストを考える．頻度は，それぞれ [93, 3, 3, 1] とする．また，今，ある学習者に注目して，この学習者に対する過去の成績情報などから，[a,b,c,d] の各単語を知っている確率を機械学習などを使って予測し，入手できるものとし，それぞれ [0.9,0.6,0.5,0.2] とする．

この時，応用言語学のテキストカバー率の考え方によれば，テキスト中の 95% 以上の単語を知っていれば，テキストを読むことで知らない単語を学習することができる．具体的に，テキストカバー率が 95% を超える場合を列挙してみよう．この例では，単語 a が圧倒的に多いが，単語 a だけで 95% を超えることはできない．閾値を超えるのは，例えば，{a,b}を知っていて，{c,d}を知らない場合があげられる．知っている単語をだけに注目して書くと，閾値を超える場合は，{a,b},{a,c}，{a,b,d}, {a,c,d}, {a,b,c,d}の場合である．

さて，この学習者が知っている単語が{a,b}になる場合に注目しよう．{c,d}については知らないと考えていることに注意すると，{a,b}になる確率は，$0.9\times0.6\times(1-0.5)\times(1-0.2)$ と計算できる．この時，前述の応用言語学の知見 [1, 2] は，この学習者がこのテキストを読むことで，**知らない単語 $\{c,d\}$ を自然に修得できる**ということを示している．したがって，知らない側の単語に注目していけば，**このテキストを読むことで，この学習者が単語*{c,d}*を新たに修得できる確率が，$0.9\times0.6\times(1-0.5)\times(1-0.2)$ である**とも考えることができる．他の場合についても同様に考えられる．学習者が知っている単語が{a,c}である確率は，学習者がこのテキストを読むことによって新たに修得できる単語が{b,d}である確率とも読み替えられる．

こうして, すべての場合について考え, この学習者がこのテキストを読むことで [a,b,c,d] の各単語を修得できる確率を集計していくと，[0,0.18,0.27,0.576] になる．このうち，単語 a については修得できる確率が 0 になっているが，これは，単語 a を知らなければそもそもテキストが読めず付随的学習が起こらないので，付随的学習によって単語 a が修得できることはあり得ないことを示している．

さらに，この集計した確率値から，この学習者がこのテキストを読むことで修得できる語数の分布も求めることができる．例えば，3 単語を修得できる確率は，単語{b,c,d}をすべて修得する場合であるから，$0.18\times0.27\times0.576$ のように計算できる．こうして修得できる語数の分布が求められれば，そこから，修得できる語数のブレ（分散）も計算できる．修得できる語数のブレが大きいということは，確率的には修得できる語数が少なくなることがあり得るということである．すなわち，同程度の語数が修得できるテキストなら，より修得できる語数のブレが少ないもの（分散が小さいもの）を選ぶ方が，確実に語数を増やせる分，学習者にとっては得になる．すなわち，修得できる語数の分布における分散の値は，学習者にとっての「リスク」と考えることができる．このように，修得できる語数の分布が求まれば，修得できる語数の平均値のみならず，分散を考慮することで，確実に修得できる語数を増やせるように「リスク」を考慮したテキスト選択が可能となる．

3 提案する定式化

前節の内容を定式化する．今，I 種類の語彙 $\{v_1,\ldots,v_I\}$ を考え，注目しているテキスト中の v_i の個数を n_i とする．また，注目している学習者が v_i を知っている確率を p_i で表す．この時，テキストカバー率の閾値を τ とする（前節の例では $\tau=0.95$）．この時，テキストカバー率が閾値を超える確率は，次のように表せる．まず，テキスト中の総単語数は $N=\sum_{i=1}^{I}n_i$ と表せる．また，学習者が単語 v_i を知っている場合に 1，そうでなければ 0 になる次の確率変数を考える．ただし，$\{Z_1,\ldots,Z_I\}$ は互いに独立とする．

$$Z_i \sim \mathrm{Bernoulli}(p_i) \tag{1}$$

この時，学習者が知っている単語のテキスト中の出現数は $\sum_{i=1}^{I}Z_in_i$ であるから，テキストカバー率は $\frac{\sum_{i=1}^{I}Z_in_i}{N}$ と表せる．したがって，テキストカバー率が閾値を超える確率は $P(\sum_{i=1}^{I}Z_in_i\geq N\tau)$ と表せる．

学習者がテキストを読む付属的学習により語 v_i を新たに修得する確率は，次のように定式化できる．テキストを読む付属的学習が起こるためには，テキストが読める必要があるので，テキストカバー率が閾値を超えていなければならない．さらに，語 v_i が新たに修得されるためには，学習者は語 v_i を知らないことが必要である．したがって，この確率

は，$P(Z_i = 0, \sum_{i=1}^{I} Z_i n_i \geq N\tau)$ と表せる．この確率を q_i とおく．

さらに，このテキストからの付属的学習による修得語数の分布を求めたい．

$$A_i \sim \text{Bernoulli}(q_i) \tag{2}$$

とすると，修得語数の確率変数 A は，$A = \sum_{i=1}^{I} A_i$ と表せる．したがって，A の確率分布を求めれば，修得語数の分布も求まる．ただし，$\{A_1, \ldots, A_I\}$ は互いに独立とする．

4 解法

前節において，テキストカバー率が閾値を超える確率 $P(\sum_{i=1}^{I} Z_i n_i \geq N\tau)$ や，修得語数の分布 $A = \sum_{i=1}^{I} A_i$ を求めるためには，**異なる成功確率**を持った互いに独立な二項分布の和からなる確率変数の分布を求める必要がある．成功確率が等しければ二項分布の和は再生性を持つが，成功確率が異なるため，この和は二項分布にはならない．こうした分布は，**ポアソン二項分布**と呼ばれる．

$P(\sum_{i=1}^{I} Z_i n_i \geq N\tau)$ については，動的計画法を解くことで求める方法を過去に提案した [3, 14]．簡潔に言えば，$\sum_{i=1}^{I} Z_i n_i$ が整数であることを利用して，$N\tau$ 以上という条件を，「$\{n_1, \ldots, n_I\}$ の部分和が丁度いくつ」という部分和問題に帰着させる．この部分和問題を，「$\{n_1, \ldots, n_i\}$ までの数で丁度いくつになるものを作れる確率」からなる DP テーブルによる動的計画法で解ける．今回は，それに加えて，$P(Z_i = 0, \sum_{i=1}^{I} Z_i n_i \geq N\tau)$ を求める必要がある．こちらは，動的計画法の DP テーブルの各セルに，そのセル時点での $\{Z_1, \ldots, Z_I\}$ の確率値の集計値を記録する拡張を施すことで計算した．なお，ポアソン二項分布は，平均と分散を求めるだけであれば，$\sum_{i=1}^{I} p_i$ が平均，$\sum_{i=1}^{I} p_i(1 - p_i)$ が分散となる．後述の**図 2** を書く際には，この性質を用いて平均と分散を求めた．

5 実験

学習者が知っている単語については，著者がクラウドソーシングを用いて過去に公開したデータがある [11]．具体的には，クラウドソーシング上の学習者（TOEIC の受験経験があるものに限定）100 人に，100 問からなる語彙サイズ計測用の単語テスト Vocabulary Size Test(VST) (VST)[15] を受けてもらった結果のデータセットである．VST は多肢選択型のテストであり，**英文中に埋め込まれた**単語の言い換えとして適切な選択肢を 4 つの選択肢の中から選択するテストである．語形などから答えが分かってしまわないように，選択肢はすべて文中の文字列をそのまま選択肢に置き換えても文法的には問題がない選択肢になるように工夫されている．

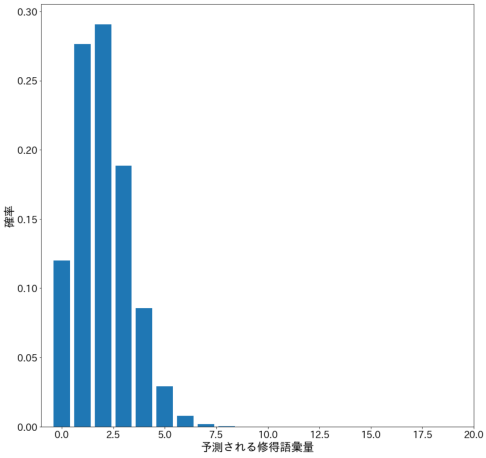

図 1 予測される修得語彙量の分布の例．

単語テストの結果を使って，各学習者が所与の単語を知っているかどうかを判別する確率的識別器を作成し，この確率値を**式 1** における p_i として用いた．二値判別問題であるため，ニューラルな識別器を用いることもできるが，今回は識別性能を向上させることが目的ではないので，単純なロジスティック回帰を用いて識別器を構成した．素性としては，COCA コーパスの頻度，British National Corpus (BNC) コーパスの頻度を用いた．ただし，頻度は $-\log($ 頻度 $)$ の形に直して素性として用いた．テキストとしては，Brown Corpus を用いた．テキスト長さが実験結果に影響しないように，Brown Corpus の 500 件の各テキストのうち，先頭から 300 語を切り出し，実験に用いた．この 500 件のテキストから，ある学習者の付属的学習に適したテキストを選択することが，我々の目的である．

付随的学習はテキストを読める必要があるので，好成績の学習者に起こりやすい．まずは，最も成績の良かった学習者（VST で 96 問正解）を対象に実験を行った．**図 1** にこの学習者があるテキストを読んだ時の修得語彙量の分布を 1 つ示す．**図 1** より，修得語彙量の分布には幅があり，単純に期待値が高いテキストを選べばよいわけではないことがわかる．

各テキストを読む際に予測される修得語彙量の期待値と分散を同時に考慮するため**図 2** に図示した．各点は Brown Corpus の各テキストである．修得語

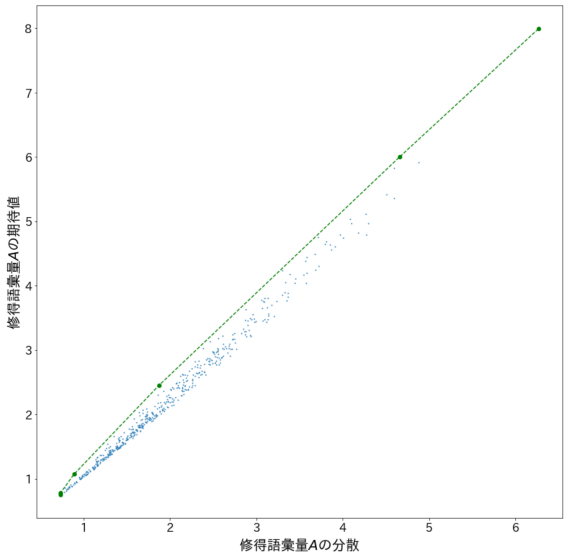

図 2　各テキストの修得語彙量分布の期待値と分散.

彙量を学習者にとっての利得と考えると，修得語彙量の期待値が同じであれば，できるだけ修得語彙量の分散が少ないテキストを選択する方が，確実に語彙を増やせるので，学習者にとっては得となる．すなわち，**図 2** の左上部分が，この学習者にとって最も効率的に付随的学習を行える文書群である．このように，**図 2** の縦軸は利得，横軸の修得語彙量の分散は**リスク**とみなせ，**図 2** は，経済分野で多用されるリスクとリターンの関係図とみなせる．

図 2 のようなリスクとリターンの関係図においては，左上部分が最も低リスクで利得を増やせる選択であり，この部分を**効率的フロンティア**という．**図 2** では，500 件あった選択肢の中から，凸包を用いて効率的フロンティアに属するテキスト 5 件が選択された．すなわち，学習者の付随的学習に適したテキストを 1/100 に絞ることができた．この 5 件の中でどのテキストを選択するかは，学習者がどの程度のリスクをとって語彙量を増やしたいかによって変わるので学習者に任せる方法が一案である．

効率的フロンティアに含まれるテキストは学習者によってどの程度変わるのだろうか？ 100 人のうち，成績の良い 30 人に対して，同様に各学習者が各テキストを読んだ場合に予測される修得語彙量の分布から効率的フロンティアを求めた．その結果，10 人以上で，効率的フロンティアに選ばれたテキストが 7 件あり，最も多いもので 14 人の効率的フロンティアに含まれたテキストがあった．この結果から，効率的フロンティアに含まれるテキストは比較的安定していることが見て取れる．

6　関連研究と展望

学習者が各単語を知っている確率 p_i については，本稿では単語テストを用いて推測する方法を用いたが，スマートフォンの語彙学習アプリなどからも推測することが可能である．例えば，語彙学習と関係が深い忘却曲線を考慮して，時間によって忘却する度合いをもモデル化する研究が近年複数なされている [12, 13]．こうした研究では，例えば最後に単語を学習してからの時間などを p_i のモデルに組み入れ，時間がたつと p_i が低くなるようにモデル化しており，実データとうまくフィットしたという報告もなされている．こうした手法を組み入れれば，語彙学習アプリのログから直接学習途中の各単語に対して，各単語を知っている確率 p_i が算出できる．

また，今回の文書選択では，「次に付随的学習に用いるテキスト」を選択したが，将来的な利得の総和を考慮するアプローチも考えられる．このアプローチは強化学習でうまく記述できる．今回のようなリスクを考慮した選択も，累積報酬の期待値だけではなく分布全体の形を考慮して強化学習を行う分布強化学習の理論面で研究されている [16, 17].

7　おわりに

本研究では，付随的学習による語彙修得に適したテキストを選択するため，応用言語学の知見に基づき，個々の学習者・個々のテキストに対して付随的学習によって修得される語彙量の推定値を算出する手法を提案した．修得語彙量を学習者にとっての利得と考えることで，金融工学などで使われる効率的フロンティアの考え方を語彙学習支援の研究に導入した．さらに，多くの学習者の効率的フロンティアに含まれる，付随的学習に適した「お得」なテキストが存在することを実験的に示した．

今後の課題としては，単語修得アプリのログなどを用いた経時的な評価実験，将来的な利得分布を考慮できる分布強化学習の導入，単語埋め込みベクトル空間上の領域の考慮した多義語対応，一度に複数のテキストを読む場合の複数テキストの選択を現代ポートフォリオ理論 [18] や整数計画問題を絡めて行うことなどが挙げられる．

謝辞

本研究は JST 戦略的創造研究推進事業 ACT-X（JPMJAX2006）の支援を受けた．

参考文献

[1] 中田達也. 英単語学習の科学. 研究社, 2019.

[2] Paul Nation. How much input do you need to learn the most frequent 9,000 words? 2014. Publisher: University of Hawaii National Foreign Language Resource Center.

[3] Yo Ehara. テキストカバー率の確率的拡張に基づく語彙テストのみからの個人化読解判定. 2019.

[4] I. Nation. How Large a Vocabulary is Needed For Reading and Listening? *Canadian Modern Language Review*, Vol. 63, No. 1, pp. 59–82, October 2006.

[5] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, Vol. 26, pp. 267–275, 2018.

[6] John Lee and Chak Yan Yeung. Personalizing Lexical Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 224–232, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[7] Chak Yan Yeung and John Lee. Personalized Text Retrieval for Learners of Chinese as a Foreign Language. In *Proc. of COLING*, pp. 3448–3455, August 2018.

[8] John Lee and Chak Yan Yeung. Personalized Substitution Ranking for Lexical Simplification. In *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 258–267, Tokyo, Japan, October 2019. Association for Computational Linguistics.

[9] John Lee and Chak Yan Yeung. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pp. 1–4, April 2018.

[10] 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, 2014.

[11] Yo Ehara. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*, May 2018.

[12] Burr Settles and Brendan Meeder. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1848–1858, Berlin, Germany, August 2016. Association for Computational Linguistics.

[13] Siddharth Reddy, Sergey Levine, and Anca Dragan. Accelerating human learning with deep reinforcement learning. In *NIPS workshop: teaching machines, robots, and humans*, 2017.

[14] Yo Ehara. Uncertainty-Aware Personalized Readability Assessments for Second Language Learners. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1909–1916, December 2019.

[15] David Beglar and Paul Nation. A vocabulary size test. *The Language Teacher*, Vol. 31, No. 7, pp. 9–13, 2007.

[16] 牧野貴樹, 澁谷長史, 白川真一, 浅田稔, 麻生英樹, 荒井幸代, 飯間等, 伊藤真, 大倉和博, 黒江康明, others. これからの強化学習. 2016. Publisher: 森北出版.

[17] 森村哲郎. 《第 11 回》リスク考慮型強化学習. 計測と制御, Vol. 52, No. 9, pp. 818–823, 2013. Publisher: 公益社団法人 計測自動制御学会.

[18] Harry M Markowitz. Foundations of portfolio theory. *The journal of finance*, Vol. 46, No. 2, pp. 469–477, 1991. Publisher: JSTOR.