テキストカバー率からの読解成功判定法の一般化とその応用

江原 遥
静岡理工科大学情報学部/産業技術総合研究所人工知能研究センター
ehara.yo@sist.ac.jp

1 はじめに

語学学習者の理解を助けることを主目的とする語学学習支援においては，語学学習者が与えられたテキストを「読める」か（読解に成功するか）を判定するタスクは，様々な応用につながる基礎的なタスクである．例えば，学習者が読めないと判定されたテキストについてテキスト平易化を行い読みやすくする研究 [6, 14] や，コーパス中から学習者の学習にふさわしいテキストを収集してくる研究 [17] が広く行われている．

語学学習者が理解可能なテキストを簡便に判定する方法として，語学教育分野で，対象言語に寄らず広く利用されている方法が，テキストカバー率の閾値を用いた方法である．この方法では，テキスト中の延べ語数に対する，語学学習者の知っている語（既知語）の延べ語数の比率（テキストカバー率 (lexical text coverage) [1]）がある閾値を超えているかどうかで判定する [8, 11, 5, 12, 16]．英文のテキストの場合，テキストのドメインにも依存するが，テキスト中の 95%から 98%の語を知っていれば，十分な読解が得られる，とされている．直感的には，この結果は，テキストを十分理解するために，全ての語を知っている必要はなく，一部は文脈から推測可能であることを示している．

さて，このように広く使われているテキストカバー率による読解成功判別法であるが，従来法では，「学習者が知っている語」の判定が粗く，特に確率的な扱いができないという問題がある．現状，学習者が語を知っているかどうかは，学習者が知っている語彙サイズ（既知語数）を概算し，コーパスの頻度の降順に語を並べた時，語彙サイズまでは知っている，その既知語数より順位が大きい語は知らない，というごく単純な方法を取っている．実際には，既知語数の辺りでは，語を知っているか知らないかは正確に推定できないため，というのが正確であろうから，テキストカバー率の計算のときも，この効果を取り入れられた方が良いだろう．特に，短いテキストでは，判別を誤りやすい．節 1.1 で，従来法で問題となる図 1 の例を挙げ，詳述する．

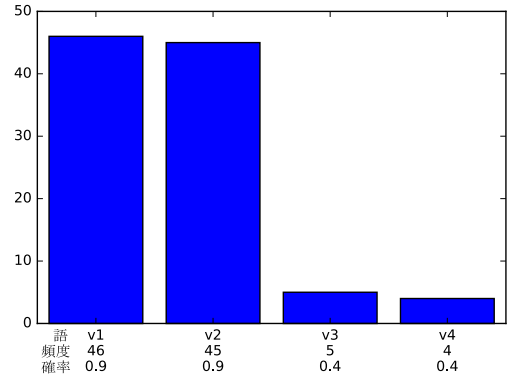

図 1: 従来法だと読解できないと判別されるが，厳密には読解に成功し得るテキストの例．異なり語数 4 語，延べ語数 100 語．図の縦軸はテキスト中の各語の頻度．横軸は，上の行から，語，このテキスト中の語の頻度，ある学習者が各語を知っている確率．

本稿では，こうした問題を解決するため，テキストカバー率を確率変数とみなし「テキストカバー率が閾値を超える事象の確率」を求める問題への一般化を示す．次に，この問題が NP 完全である事を示し，動的計画法による効率的な求解アルゴリズムを示す．そして，従来法が，この一般化において，取り得る確率値を 1 または 0 に限定した特殊ケースに含まれることを示す．

本研究の位置づけは，学習者のテキストカバー率と読解成功の関係に関する応用言語学の従来研究 [8, 11, 5, 12, 16] の流れと，個人化語学学習支援の目的で，「個々の語学学習者が語を知っているかを予測する」筆者の過去の語彙知識予測問題に関する研究 [3, 4, 2, 17, 10] の流れを理論的に結びつけるものといえる．

[1] この比率は，日本語では，文献によって，語彙カバー率，語彙被覆率，被覆率，既知語率など様々に訳されている．テキストカバー率は延べ語数同士の比率であるが，自然言語処理の文脈では，異なり語数同士の比率を被覆率と言い換えることがあること，また，既知語率は日本語教育分野 [18] 以外では使用例を見かけなかったため，本稿では，テキストカバー率とした．

1.1 従来法で問題が起きる例

さて，従来法で問題が起きる場合を，簡単な例（図 1）で示そう．簡単のため，テキスト中には v_1, v_2, v_3, v_4 の 4 種類の語しかないものとし，それぞれの頻度が図 1 のようであるとする．また，学習者が各々の語を知っている確率も図 1 のようであるとする．この時，現状では，v_1, v_2 の 2 語のみ知っているとみなすので，テキストカバー率は 91% となり，閾値 95% を下回るので，学習者はこのテキストを読めないと判定される．また，確率値を扱う方法として，単純に各々の語の頻度と確率値をかけあわせて，テキスト中で知っている語の平均語数を求める方法が容易に思いつくが，これは，85.5 語/100 語となり，閾値 95% を下回るので，やはり，学習者はこのテキストを読めないと判定される．

さて，図 1 は簡単な例なので，学習者が知っている語がどのような組み合わせであれば，テキストカバー率が閾値を超えるのかを列挙して考えられる．具体的には，$\{v_1, v_2, v_3\}$ は知っているが v_4 を知らない場合，$\{v_1, v_2, v_4\}$ は知っているが v_3 は知らない場合，及び，全ての語を知っている場合の 3 通りで，テキストカバー率は 95% を超える．この確率を図 1 について計算すると，$0.9 * 0.9 * 0.4 * 0.6 + 0.9 * 0.9 * 0.6 * 0.4 + 0.9 * 0.9 * 0.4 * 0.4 = 0.518$ となり，実際には，学習者は半分程度の確率でこのテキストを読むことができることがわかる．

2 定式化

本節では，テキストカバー率を用いた読解成功判別法の定式化を行う．まず，従来のテキストカバー率がどのように表記できるか考えよう．

2.1 従来法の定式化

I 種類の語の集合 $\{v_1, \ldots, v_I\}$ と，J 人の学習者の集合 $\{l_1, \ldots, l_J\}$ を考えよう．また，あるテキスト $\mathcal{T}$ 中の語 v_i の頻度を $n(\mathcal{T}, v_i)$ で表す事にする．すると，テキスト $\mathcal{T}$ の延べ語数は，$|\mathcal{T}| = \sum_{i=1}^{I} n(\mathcal{T}, v_i)$ と書ける．また，y_{ij} を，語 v_i を学習者 l_j が知っている時は 1，知らない時は 0 を取るとしよう．すると，「テキスト $\mathcal{T}$ 中で学習者 j が知っている語」の延べ語数は，$\sum_{i=1}^{I} y_{ij} n(\mathcal{T}, v_i)$ と書ける．すると，テキスト $\mathcal{T}$ の学習者 l_j のテキストカバー率 $TC_{\mathcal{T},j}$ は，次のように書くことができる．

$$TC_{\mathcal{T},j} = \frac{\sum_{i=1}^{I} y_{ij} n(\mathcal{T}, v_i)}{|\mathcal{T}|} = \sum_{i=1}^{I} y_{ij} \frac{n(\mathcal{T}, v_i)}{|\mathcal{T}|} \quad (1)$$

ここで，従来，95%〜98% 程度とされてきた閾値を τ とおこう．$TC_{\mathcal{T},j} \geq \tau$ の時，学習者 l_j は $\mathcal{T}$ を十分に読解できる，と判定するのが，従来法である．

さて，従来法では，$y_{ij} \in \{0, 1\}$ は，単なる変数であり，「学習者 l_j が語 v_i を知っているか否か」を，粗くでもよいのでとにかく決めてしまうことによって値が決められていた．例えば，Vocabulary Size Test 法 [13] では，学習者 l_j の語彙サイズ $VS(l_j)$ を測り，大きなコーパス[2]上の頻度の降順で，語 v_i の順位 $\mathrm{rank}(v_i)$ をみて，$\mathrm{rank}(v_i) \leq VS(l_j)$ の時，$y_{ij} = 1$，そうでなければ $y_{ij} = 0$ としていた．

2.2 確率変数の導入

ここで，y_{ij} を単なる変数ではなく，確率変数としてみてみよう．ただし，$\{y_{ij} | i \in \{1, \ldots, I\}, j \in \{1, \ldots, J\}\}$ は互いに独立と仮定する[3]．また，$y_{ij} = 1$ となる確率を，$P(y_{ij} = 1)$ で表す．

さて，確率変数の定数倍や確率変数同士の和も確率変数である．ここで，テキストカバー率の定義である式 1 をよくみると，これは，まさに，確率変数 y_{ij} 同士の和と定数倍 $\frac{n(\mathcal{T}, v_i)}{|\mathcal{T}|}$ からなっていることがわかる．したがって，テキストカバー率も確率変数とみなすことができる．直感的には，これは，「テキストカバー率が閾値を超える」こと自体を確率的事象とみなすことができる事を意味する．具体的に，学習者 l_j を固定した時，「テキストカバー率が閾値 τ を超える確率」は，次のように書ける．

$$P(TC_{\mathcal{T},j} \geq \tau | l_j) = P\left(\sum_{i=1}^{I} y_{ij} n(\mathcal{T}, v_i) \geq |\mathcal{T}| \tau \middle| l_j \right) \quad (2)$$

式 2 による定式化は，従来のテキストカバー率を特殊ケースとして含む自然な拡張である．実際，$\forall i$ について $P(y_{ij} = 1) \in \{0, 1\}$ と限定した場合が，語を知っている/知らないに一旦 2 分してから計算する，従来のテキストカバー率に対応する．

[2] コーパスとしては，British National Corpus (BNC) や，Corpus of Contemporary American English (COCA) が用いられている．

[3] この仮定は，各語を設問とみなし，各学習者を受験者とみなした時のテスト理論（項目反応理論）を用いる際の一般的な仮定である．また，単語テストについては，従来法で計測した学習者 l_j の語彙サイズ $VS(l_j)$ と，最も単純な項目反応理論の内最も単純な Rasch を用いた際の各語の困難度パラメタがよく相関することが示されている [1] ことから，学習者の語彙サイズ計測の上では，このように仮定しても特段の問題がないことが示されていると言える．

では，具体的に「テキストカバー率が閾値 τ を超える確率」はどのように計算すればよいのだろうか？式 2 の右辺をみると，すぐに思いつく方法としては，I 個の 2 値確率変数の列 $\{y_{1j}, \ldots, y_{Ij}\}$ が取り得る全ての組み合わせを列挙し，式 2 の右側の括弧内に書かれた条件を満たす組み合わせの確率をすべて足し込む方法が考えられる．しかし，この方法は，2^I 通りの組み合わせを列挙することになるため，計算量は $O(2^I)$ であり，組み合わせ爆発を起こす問題があるため非現実的である．

3 提案手法：部分和問題への帰着と効率的なアルゴリズム

本節では，学習者 l_j がテキスト $\mathcal{T}$ を読んだ場合のテキストカバー率 $TC_{\mathcal{T},j}$ が，閾値 τ を超える確率 $P(TC_{\mathcal{T},j} \geq \tau | l_j)$ を効率的に計算するアルゴリズムを提案する．

式 2 の右側に注目しよう．今，簡単のため，$n_i = n(\mathcal{T}, v_i)$ とおく．すると，式 2 の右側部分は，$\{n_1, \ldots, n_I\}$ のうち，各 $y_{ij} \in \{0, 1\}$ が，各 n_i を使うかどうかを決めている，とみなすことができる．すなわち，集合 $\{n_1, \ldots, n_I\}$ の部分和が，「$|\mathcal{T}|\tau$ 以上」である組み合わせを網羅しようとしていることになる．本節では，以下，簡単のため，集合 $\{n_1, \ldots, n_I\}$ の部分和を，単に「部分和」と呼ぶ．

ここで，「部分和が $|\mathcal{T}|\tau$ 以上である場合」を具体的に書き下してみよう．実数 x 以上の最小の整数を返す天井関数 $\lceil x \rceil$ を用いると，これは，部分和が $\{\lceil |\mathcal{T}|\tau \rceil, \lceil |\mathcal{T}|\tau \rceil + 1, \ldots, |\mathcal{T}|\}$ のいずれかである場合と同値である．まとめると，手順は次の 2 つからなる．第 1 に，「部分和が $|\mathcal{T}|\tau$ 以上である場合」を，部分和が上記のいずれかの値になる場合にわける．第 2 に，部分和がある整数 N になるような $\{y_{1j}, \ldots, y_{Ij}\}$ の組み合わせを列挙し，この組み合わせが起こる確率を求めれば良い．

この第 2 の手順の中では，非負整数列 $\{n_1, \ldots, n_I\}$ の部分和のうち，ある整数 $N \geq 0$ を満たすものがあるかどうかを判定する問題を内包している．この問題は，**部分和問題**と呼ばれ，計算量理論の基礎的な問題である．部分和問題は **NP 完全問題**であることがわかっているが，動的計画法によるアルゴリズムが知られており，これを用いることによって実用的なサイズの問題を解くことができることが知られている [7]．

Algorithm 1 に，提案アルゴリズムを示す．これは ProbTCsurpass と SubsetSumP の 2 つの関数からなり，前者が，閾値 τ を入れた時，テキストカバー率 $\geq \tau$ となる確率を返す，目的の関数である．この中では SubsetSumP を呼び出している．$SubsetSumP(i, N)$ は，$n_1, \ldots, n_i$ の部分和が N になる確率を返す．

Algorithm 1 ProbTCSurpass:テキストカバー率が閾値を超える正確な確率を求めるアルゴリズム

Input: n_i:語 v_i のテキスト $\mathcal{T}$ における頻度，p_i:学習者 l_j が語 v_i を知っている確率値，τ:閾値，$|\mathcal{T}|$:テキストの全延べ語数，I:テキストの全異なり語数
Output: $p_{\mathrm{TCsurpass}}$:テキストカバー率が閾値を超える確率

```
function ProbTCSurpass(τ)
    p_TCsurpass ← 0
    for N = ⌈|T|τ⌉ to |T| do
        p_TCsurpass ← p_TCsurpass + SubsetSumP(I, N)
    end for
    return p_TCsurpass
end function
function SubsetSumP(i, N)
    if i ≤ 0 then
        if n = 0 then
            return 1
        else
            return 0
        end if
    end if
    if N ≥ n_{i−1} then
        return p_{i−1} * SubsetSumP(i − 1, N − n_{i−1})
                +(1.0 − p_{i−1}) * SubsetSumP(i − 1, N)
    else
        return (1.0 − p_i) * SubsetSumP(i − 1, N)
    end if
end function
```

本研究の目的では，所与の有限非負整数列の部分和で N を満たすものがあるかを単に判定するだけでなく，この列と同じ長さの確率値の列も与えられ，(独立性を仮定した上で) N を満たす確率も計算する必要がある．これは，既存の部分和問題のアルゴリズムを用いて解くことが可能であるが，従来法そのままではないため，Algorithm 1 において，P をつけ，関数の名前を SubsetSumP とした．簡単のため，Algorithm 1 では再帰を用いて記述した．実際に，メモ化 (memoization) を行い，過去の SubsetSumP の引数と返り値をキャッシュに入れておくことによって，SubsetSumP は十分高速に動作する．

部分和問題の動的計画法の計算量 [7] をもとに，提案法の計算量について述べる．SubsetSumP の計算量は，N と $|\mathcal{T}|$ の値が近いことを考えると，$O(I|\mathcal{T}|)$ となる．これを $(1.0-\tau)|\mathcal{T}|$ 回呼び出すので，Algorithm 1 の全体の計算量は $O(I|\mathcal{T}|^2(1.0-\tau))$ となる．

4 実験

実験のため，国内のクラウドソーシング Lancers 上で，59 名の被験者に，単語テストの後，読解問題に回

答させることで，データセットを作成した．単語テストとしては，英語の語彙サイズ計測の目的で標準的なテストである Vocabulary Size Test (VST) A[13] を用いた．VST では，BNC コーパス中の頻度の降順に各語を並べ，順位を 1,000 語単位の段階に区切り，同じ段階の難度を同程度とみなし，20,000 語までの 20 段階について，各段階から 5 語を試験する．

読解問題は，次の 2 問を用いた．まず，英語学習者の読解力を測定するための読解問題について精密に報告している文献 [9] の付録に収録されている問題で，もともとはイスラエルの大学の入学試験問題である．テキストは延べ 380 語で，4 択の選択問題 5 問が付随する．次に，機械読解の分野でベンチマークとして広く用いられており，SQuAD 2.0 データセット [15] から，“Normans” という設問を用いた．テキストは延べ 113 語で，記入式の問題 9 問が付随する．両問題とも，国内の英語の教科書/参考書で用いられており，作業者が過去に同じ問題に遭遇した可能性は低い．

次に，各手法による読解成功判定の精度を計測するが，従来法がパラメタを持たない手法であるため，訓練･発展･テストに 3 分割する典型的な機械学習の精度計測法が必要ないことに注意が必要である．厳密には，テキストカバー率の閾値はハイパーパラメタたと言えるが，既存の報告から 95% か 98% のどちらかが決め打ちで用いられており，データを用いてその都度ハイパーパラメタを決定する機械学習の文脈とは異なる．

従来法がパラメタを持たない手法であるため，比較手法もパラメタを持たない手法とすることが適切であることに注意しつつ，各学習者が語 v_i を知っている確率を求める手法を次のように定めた．a) 従来法．VST から学習者の語彙量を求め，BNC 中の語の降順頻度順位 $\geq$ 学習者の語彙量であれば確率 1 で語を知っている，そうでなければ確率 1 で語を知らないとする方法である．前述のように，この方法は，従来法に一致する．b) 提案法．ここでは確率値の導入による効果をみるため，判定手法は a) に近づけ，VST 中の各段階の 5 問中，正解した問題数の比率を確率値とした．

4.1 結果

文献 [9] では,「十分に読解できる」スコアを，各読解問題に付随する設問の約 55% に回答可能であることと定義している．これに従い，1 問目では 3 問以上，2 問目では 6 問以上正答した場合，正解とした．提案法で確率値を評価する時,「テキストカバー率が閾値を超える確率」が 0.5 以上であれば読解成功，そうでない場合，読解失敗と判定した．表 1 より，提案法の方がわずかに従来法より優れていることがわかる．これとは別に，従来法と提案法で，判別に成功/失敗した問題数を調べ，2x4 の分割表で χ 2 乗検定を行ったところ $p < 0.01$ で統計的有意であった．

	問 1 正答	問 1 不正答	問 2 正答	問 2 不正答
従来法	34	25	34	25
提案法	31	28	29	23

表 1: 1 問目での正解数

5 おわりに

本稿では，テキストカバー率から学習者が読解に成功するかを判定する方法を一般化し，性能を確認した．紙面の都合で省いたが，提案手法はテキスト中の，テキストカバー率が高い範囲を検出する拡張が容易である他，確率を扱うことで多義語なども扱いやすいと考えられ，有望な性質を持っている．

謝辞

この研究は，JST 戦略的創造研究推進事業（ACT-I）の支援を受けた．

参考文献

[1] David Beglar. A rasch-based validation of the vocabulary size test. *Language Testing*, Vol. 27, No. 1, pp. 101–118, 2010.
[2] Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1374–1384, Doha, Qatar, October 2014. Association for Computational Linguistics.
[3] Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December 2012.
[4] 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.
[5] David Hirsh and Paul Nation. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a foreign language*, Vol. 8, pp. 689–689, 1992.
[6] Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pp. 59–73, 2013.
[7] Jon Kleinberg and Eva Tardos. *Algorithm design*. Pearson Education India, 2006.
[8] Batia Laufer. What percentage of text-lexis is essential for comprehension. *Special language: From humans thinking to thinking machines*, Vol. 316323, , 1989.
[9] Batia Laufer and Geke C Ravenhorst-Kalovski. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, Vol. 22, No. 1, pp. 15–30, 2010.
[10] 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.
[11] Paul Nation. *Teaching and Learning Vocabulary*. Heinle and Heinle, Boston, MA, 1990.
[12] Paul Nation. How large a vocabulary is needed for reading and listening? Vol. 63, No. 1, pp. 59–82, 2006.
[13] Paul Nation and David Beglar. A vocabulary size test. Vol. 31, No. 7, pp. 9–13, 2007.
[14] Gustavo Paetzold and Lucia Specia. Benchmarking lexical simplification systems. In *LREC*, 2016.
[15] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. Association for Computational Linguistics, 2018.
[16] Norbert Schmitt, Tom Cobb, Marlise Horst, and Diane Schmitt. How much vocabulary is needed to use english? replication of van zeeland schmitt (2012), nation (2006) and cobb (2007). *Language Teaching*, Vol. 50, No. 2, p. 212226, 2017.
[17] Chak Yan Yeung and John Lee. Personalized text retrieval for learners of chinese as a foreign language. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3448–3455, 2018.
[18] 松下達彦．オンライン日本語テキスト語彙分析器 j-lex．日本語教育方法研究会誌，Vol. 21, No. 1, pp. 8–9, 2014.