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

江原 遥
Yo EHARA
静岡理工科大学情報学部
Department of Informatics, Shizuoka Institute of Technology
Email: ehara.yo@sist.ac.jp

あらまし： 応用言語学分野ではプレイスメントテストなどの目的で、学習者が所与の文書を十分に読解可能か判定する個人化読解判定が注目されており、文書中で学習者が知っている語の比率「テキストカバー率」が決定的な特徴量と示されている。この率の計算には学習者が単語を知っているかの識別器を用いる。従来は識別器が返す識別の信頼度を利用できなかった。本稿では、この信頼度を利用して当該判定を高精度に行う枠組みを提案する。

1 はじめに

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，語学教育や応用言語学分野で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) がある閾値を超えているかどうかで判定する [3, 5, 1, 6, 8]．英文のテキストの場合，テキストの分野にも依存するが，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．

このように，テキストカバー率による読解判定法では，テキスト中の各語に対して，学習者がその後を知っているか（既知語か）否かを識別することが求められる．この識別には，数十分程度で回答可能な語彙テストが用いられる．個々の学習者に対し，テキスト中に現れる全ての語についてテストすることは現実的ではないので，語彙テストは 100 語程度の一部の語に対してのみ行われる．したがって，語彙テストに出てきていない語については，学習者にとって既知語かどうかは不確かである．従来法では，この不確かさを読解判定法の際に考慮することができなかった．

本稿では，学習者にとっての既知語の識別における不確かさを，読解判定法の際にも考慮する手法を提案する．提案手法はテキストカバー率を確率変数とみなす自然な拡張になっているため，数多くの既存研究で実証されてきた読解に必要なテキストカバー率の閾値をそのまま用いることが可能なことが，提案手法の利点である．実際に読解力テスト結果データを用いた予測実験において，提案手法の精度の優位性を確認した．

2 定式化

テキスト $\mathcal{T}$ を考えよう．テキスト中に I 種類の語があり，その語彙集合を $\{v_1, \ldots, v_I\}$ とする．また，学習者は J 人いるとし，学習者の集合を $\{l_1, \ldots, l_J\}$ とする．また，テキスト $\mathcal{T}$ 中の語 v_i の頻度を $n(v_i)$ で表す事にする．すると，テキスト $\mathcal{T}$ の延べ語数は，$|\mathcal{T}| = \sum_{i=1}^{I} n(v_i)$ 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，$\sum_{i=1}^{I} y_{ij} n(v_i)$ 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．

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

従来法では，テキストカバー率の閾値 τ に対し，$TC_{\mathcal{T},j} \geq \tau$ の時，学習者 l_j は $\mathcal{T}$ を十分に読解できる，と判定する．τ は 0.95 から 0.98 の値をとる．

さて，従来法では，$y_{ij} \in \{0, 1\}$ は，単なる二値変数であり，学習者 l_j が語 v_i を知っているかの識別が不確かさを考慮することができなかった．

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．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 $\frac{n(v_i)}{|\mathcal{T}|}$ からなっていることがわかるので，テキストカバー率も確率変数となる．「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．直感的には，この確率は「テキストカバー率が閾値を超える」こと自体がどれぐらい不確かな現象であるかを示している．

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

式 2 による定式化は，従来法を特殊ケースとして含む自然な拡張となっている．実際，$\forall i$ について $P(y_{ij} = 1) \in \{0, 1\}$ と限定した場合が，識別の不確かさを考慮しない従来法に対応している．

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．

式 2 の確率を効率的に計算するアルゴリズムを Algorithm 1 に提案する．簡単のため，$n_i = n(v_i)$ とし，集合 $\{n_1, \ldots, n_I\}$ の部分和を単に「部分和」と呼ぶ．提案アルゴリズムは，ProbTCsurpass と SubsetSumP からなり，前者は閾値 τ に対してテキストカバー率 $\geq \tau$ となる確率を返す．前者の中で，部分和が閾値以上という条件を，「部分和が N と等しい」という条件に分解し，各条件を満たす確率を返す関数 $SubsetSumP(i, N)$ を繰り返し呼び出している．部分和が，ある整数 $N \geq 0$ と等しい確率を求める問題は**部分和問題**と呼ばれる問題と類似している．部分和問題は **NP 完全問題**であることがわかっているが，動的計画法による実用的なアルゴリズムが知られており [2]，Algorithm 1 でもこの考え方を用いている．

SubsetSumP の計算量は $O(I|\mathcal{T}|)$ 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．

Algorithm 1 ProbTCSurpass: テキストカバー率が閾値 τ を超える確率を計算するアルゴリズム．

入力: n_i:語 v_i の $\mathcal{T}$ における頻度，p_i:学習者 l_j が v_i を知っている確率，τ:テキストカバー率の閾値, $|\mathcal{T}|$:テキスト $\mathcal{T}$ の延べ語数, I:語彙サイズ
出力: $p_{\mathrm{TCsurpass}}$:テキストカバー率が τ を超える確率

```
function ProbTCSurpass(τ)
    p ← 0
    for N = ⌈|T|τ⌉ to |T| do
        p ← p + SubsetSumP(I, N)
    end for
    return p
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
```

3　実験

実験のため，国内のクラウドソーシング Lancers 上で，100 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[7] を用いた．読解問題は，**Laufer** と **Short** を用いた．前者は，既存研究文献[4] の付録に収録されている問題で，イスラエルの大学の入学試験問題の公開部分をもとにしている．テキストは延べ 380 語で，4 択の選択問題 5 問が付随する．後者は，前者と同じ公開部分のうち，提示された短文の言い換えとしてふさわしい問題を選択する短文読解問題である．学習者が正答するためには，提示されたテキストと各設問の問題文の両方が読めなければならない．そこで，提示されたテキストの読解できる確率と各設問の問題文が読解できる確率の積を，各学習者が各設問に正答する確率として用いた．既存手法では読解できる/できないの 2 値が出力され，提案手法では読解できる確率（テキストカバー率が閾値を超える確率）が出力される．2 値出力と確率値出力の精度は，Mean Average Precision を用いて評価する事により比較可能であることが情報検索などの分野で知られており，本研究でもこの評価指標を用いた．この評価手法は，各手法を用いて，各学習者が読めそうな順に各設問を並べた時のランキングの精度を計測し，学習者全体について精度の平均をとったものである．ランキングの精度は，実際に学習者が読解に成功した（＝正答した）設問とランキングが完全に一致していれば 1.0 となる．1.0 となるランキングの逆順のランキングに対しては，ランキングの精度は 0.0 となる．

語彙テストの回答データは共通の試験[7] を用いた．この試験では，British National Corpus [?] 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 語を抽出し，4 択の選択式問題 100 問からなる．既知語かどうかの識別手法は **VST**, **LR**, **NN** の 3 種を比較した．**VST** は，既存手法[7] であり，正解数を単純に 20 倍して学習者の語彙量とし，学習者は BNC の頻度リスト順に語を知っているという仮定に基づき，BNC の頻度降順で語彙量までを既知語と識別する手法である．**LR**, **NN** は，それぞれ，ロジスティック回帰とニューラルネットを用いた確率的識別器である．**H-**を付けた読解成功判別手法は，確率的識別器を用いる際に，確率が 0.5 以上であれば，既知語，そうでなければ既知語ではない，というように二値化した後，テキストカバー率を計算する手法である．**A-**を付けた読解成功判別手法は，確率的識別器を用いる際に，単語頻度を確率値で重みづけした平均値を，テキストカバー率として用いる手法である．**H-**を付けた読解成功判別手法が提案手法であり，式 2 と Algorithm 1 を用いて，既知語の識別の不確かさを考慮しながら，直接テキストカバー率が閾値を超える確率を求める手法である．

表 1 に結果を示す．提案手法は，既存手法を 10 ポイント以上上回っていることが分かる．

表 1: 読解力試験に対する MAP スコア．$\tau = 0.98$.

	手法	Laufer	Short
既存手法	VST	0.4880	0.5437
	H-LR	0.5797	0.5304
	H-NN	0.5810	0.5393
	H-LR+GLOVE	0.5113	0.5613
	H-NN+GLOVE	0.5250	0.5631
平均法	A-LR	0.4880	0.4885
	A-NN	0.4880	0.4885
	A-LR+GLOVE	0.4880	0.4885
	A-NN+GLOVE	0.4880	0.4885
提案手法	UA-LR	**0.6314**	0.6533
	UA-NN	0.6172	0.6533
	UA-LR+GLOVE	0.6305	**0.6743**
	UA-NN+GLOVE	0.6159	0.6524

4　おわりに・謝辞

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，精度向上を確認した．この研究は，JST 戦略的創造研究推進事業（ACT-I, JPMJPR18U8）の支援を受けた．

参考文献

(1) 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.
(2) Jon Kleinberg and Eva Tardos. *Algorithm design*. Pearson Education India, 2006.
(3) Batia Laufer. What percentage of text-lexis is essential for comprehension. *Special language: From humans thinking to thinking machines*, Vol. 316323, , 1989.
(4) 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.
(5) Paul Nation. *Teaching and Learning Vocabulary*. Heinle and Heinle, Boston, MA, 1990.
(6) Paul Nation. How large a vocabulary is needed for reading and listening? Vol. 63, No. 1, pp. 59–82, 2006.
(7) Paul Nation and David Beglar. A vocabulary size test. Vol. 31, No. 7, pp. 9–13, 2007.
(8) 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.