ランダムフォレストを用いた英語習熟度の自動推定

小林 雄一郎　(大阪大学／日本学術振興会)
kobayashi0721@gmail.com

1. はじめに

英語教育の分野では，数多くの英語力テストが存在し，中学校や高校でのカリキュラムに組み込まれている場合もある。これらのテストの多くは，熟練した試験官や採点者が，学習者の英作文や発話を評価するという形式を取っている。しかし，熟練した試験官を育成するには，かなりの時間が必要とされるであろう。また，いかに熟練した試験官たちが厳密な基準に基づいて評価を下したとしても，複数の試験官の評価が一致しないこともある。そのような状況において，客観的な評価基準と統計モデルを用いて習熟度を推定する技術を開発することは，言語教育分野にとって非常に有意義なことである。

本研究では，習熟度の情報が付与された自由英作文を対象とし，20 種類の言語的特徴を説明変数の候補とするランダムフォレストと *k*-近傍法を用いて，書き手の習熟度を推定する。

2. 先行研究

英作文の評価は，テキスト全体に対する評定者の印象に基づく全体的評価 (holistic scoring)，作文を語彙や文法などの項目別に評定していく分析的評価 (analytical scoring)，そして，ある特定の要因（修辞的特徴など）がどの程度作文に反映されているかを基準とする特定要因の評価 (primary trait scoring) の 3 つに分けられる (Perkins, 1983)。このうち，比較的実用性が高く，多くのテストや評価で使用されているのは全体的評価であり (Weigle, 2002)，1960 年代から始まった，自由英作文の自動採点システムにおいても，表面的な特徴を用いて，自由英作文の全体的評価を行ってきた (水本, 2008)。

石岡 (2004, 2009) によれば，自動採点システムで使われることが多い説明変数は，単語の出現頻度ベクトル，文や句の長さ，代名詞や助動詞の数，議論を深めるための手がかり語や修辞句などである。

杉浦 (2008) は，日本人学習者による自由英作文における様々な言語的特徴を算出し，英語教育の専門家による評価スコアを予測する重回帰モデルを作成した。その結果，流暢さを表す総語数，統語的複雑さを表す文あたりの従属節数と平均文長，談話的特徴を表す文あたりの接続語句数が評価スコアに影響を与えていると報告している。また，水本 (2008) は，専門家による評価スコアを予測する重回帰モデルを作成した結果，総語数，Flesch-Kincaid Grade Level，平均単語長が予測に役立つと報告している。

3. 分析手法

3.1 説明変数

本研究で用いる説明変数は，以下の 20 変数である。

(1) 総語数 (Tokens)
(2) 異語数 (Types)
(3) 異語率 (TTR)
(4) Guirard Index (Guirard)
(5) 平均単語長 (MLW)
(6) 平均文長 (MLS)
(7) Flesch-Kincaid Grade Level (Readability)
(8) JACET8000 Index (JACET8000) [1]
(9) *I think* の文あたりの出現頻度 (I_think)
(10) 仮定を表す助動詞 (*could, may, might, should, would*) の文あたりの出現頻度 (Modal)
(11) 文頭接続詞 (*And, Because, But, So*) の文あたりの出現頻度 (Conjunction)

[1] JACET8000 (大学英語教育学会基本語改訂委員会, 2003) における 8 段階の語彙レベルに基づき，作文中でどれだけ難しい語彙が使われているかを測定する指標。例えば，レベル 1 での使用語彙が 49 語，レベル 2 が 3 語，レベル 3 が 1 語，レベル 4 が 1 語という場合には，(1×49)＋(2×3)＋(3×1)＋(4×1)＝62 という値が得られる (水本, 2008)。

(12) 関係節の文あたりの出現頻度 (Relative)
(13) 従属節の文あたりの出現頻度 (Subordinate)
(14) 受動態の文あたりの出現頻度 (Passive)
(15) 単語（頻度上位 100 タイプ）の出現頻度ベクトルから得られるコサイン類似度スコア (W1_dist) [2]
(16) 単語 2-gram（頻度上位 100 タイプ）の出現頻度ベクトルから得られるコサイン類似度スコア (W2_dist)
(17) 単語 3-gram（頻度上位 100 タイプ）の出現頻度ベクトルから得られるコサイン類似度スコア (W3_dist)
(18) 品詞（全て）の出現頻度ベクトルから得られるコサイン類似度スコア (P1_dist)
(19) 品詞 2-gram（頻度上位 100 タイプ）の出現頻度ベクトルから得られるコサイン類似度スコア (P2_dist)
(20) 品詞 3-gram（頻度上位 100 タイプ）の出現頻度ベクトルから得られるコサイン類似度スコア (P3_dist)

3.2 相関分析

相関係数とは，2 変数間の直線的関係の強さおよびその方向を表す測度である。本研究では，前節の 20 種類の説明変数に「習熟度 (Level)」という目的変数を加えた，全 21 変数の相関行列を作成し，各変数間の関係を探る。

3.3 ランダムフォレスト

本研究で用いる回帰手法は，Breiman (2001) によって提案されたランダムフォレストである。端的に言えば，ランダムフォレストとは，回帰木のアンサンブル学習である。回帰木とは，非線形回帰分析の 1 つとして位置付けられ，説明変数の値を何らかの基準で分岐させ，予測モデルを構築する。分岐の過程は，木構造で図示することができ，IF-THEN のような簡単なルールで表すこともできる。また，アンサンブル学習とは，必ずしも精度の高くない複数の予測器の結果を組み合わせ，精度を向上させるパターン認識の手法である。

ランダムフォレストでは，まず，与えられたデータセットから，N 組のブートストラップサンプルを作成する。次に，各々のブートストラップサンプルデータを用いて，未剪定の最大の回帰木を生成する（但し，分岐のノードは，ランダムサンプリングされた説明変数のうち最善のものを使用する）。そして，全ての結果を統合し（回帰問題では平均），新しい予測器を構築する。

ランダムフォレストの長所としては，精度が高いこと，非常に多くの説明変数を扱うことができること，それぞれの説明変数が予測に寄与する度合いが分かること，欠損値を持つデータの正確さの維持に有効であること，などが挙げられる (金, 2007)。

3.4 *k*-近傍法

ランダムフォレストによる回帰の結果として得られる予測値は，「レベル 1」や「レベル 2」といったカテゴリーではなく，1.67 や 2.43 のような数値であるため，何らかの方法で予測値をカテゴリーに変換する必要がある。その 1 つの方法としては，四捨五入が考えられる (e.g. 劉・内田, 2012)。しかしながら，例えば「レベル 1」や「レベル 2」の最適な境界が 1.5 であるとは限らないため，本研究では，*k*-近傍法を用いて，それぞれのテキストが属するレベルを判定する。[3]

k-近傍法とは，判別すべき個体に関して，その周辺で最も近い個体を k 個見つけ，その k 個の多数決により，どのグループに属するかを判別する手法である。距離の測度としては，一般的にユークリッド距離が用いられる (金, 2007)。

4. 結果と考察

4.1 実験データ

本研究の実験データは，アジア圏英語学習者コーパスネットワークである CEEAUS (Corpus of English Essays Written by Asian University Students) の日本人英語学習者モジュール CEEJUS (Corpus of English Essays Written by Japanese University Students) の一部である。CEEJUS の作文は，「大学生のアルバイトの是非」と「レストランにおける禁煙の是非」という 2 つのテーマに関して，辞書使用不可，時間制限（20〜40 分），語数制限（200〜300 語），

[2] 変数 (15)〜(20) における出現頻度のベクトルには，TF-IDF による重みづけを行っている。また，単語は表記形，品詞は TreeTagger (Marcus *et al.*, 1993) による情報付与の結果を用いている。

[3] ランダムフォレストではなく，多項ロジットモデルや多項プロビットモデルのような順序回帰の手法を用いれば，*k*-近傍法を行う必要はない (e.g. 木村ほか, 2009; 田中ほか, 2007)。しかしながら，多くの場合，それらの順序回帰手法は，ランダムフォレストによる回帰よりも予測精度が低い。

PC 上での作成，スペルチェックの使用強制という統一的な枠組みで収集されたものである。また，それぞれの作文には，TOEIC(R)テスト型の模擬試験 (MT) に基づく 4 段階の習熟度が付与されている。具体的には，模擬試験スコア (Mock Test) を TOEIC(R)推定スコアに変換し，700 点以上を Upper，600 点以上を Semi-Upper，500 点以上を Middle，500 点未満を Lower としている(Ishikawa, 2009)。

表 1 は，本研究における実験データの概要である。なお，作文のテーマは，全て「レストランにおける禁煙の是非」である。

表 1　実験データの概要

Level	*N*	Tokens	Types
1 (Lower)	20	4172	773
2 (Middle)	20	4326	749
3 (Semi-Upper)	20	4197	746
4 (Upper)	19	4567	877

表 1 を見ると，語数制限の影響もあってか，総語数に大きな差は見られず，異語数に関しても，レベル 1〜3 には明確な差が見られない。また，一般的に学習データは 300 本以上，そして，各レベルに 20 本以上が必要であると言われているが (e.g. Elliot, 2003)，今回のデータセットには 79 本の英作文しか含まれていない。

4.2 各変数の関係

表 1 にある 79 本の英作文を対象に，Pearson の相関係数を用いて，説明変数に目的変数 (Level) を加えた 21 変数の相関行列を作成した。

目的変数と説明変数との相関係数に注目すると，Readability ($r = 0.294$) が最も高く，Types ($r = 0.285$), Tokens ($r = 0.271$)，MLS ($r = 0.266$)，Subordinate ($r = 0.241$), JACET8000 ($r = 0.238$)，Guirard ($r = 0.206$) と続く。

また，説明変数間の相関係数を見ると，Types と Guirard ($r = 0.889$) が最も高く，Types と JACET8000 ($r = 0.882$), Guirard と JACET8000 ($r = 0.829$)，MLS と Readability ($r = 0.816$)，Tokens と Types ($r = 0.760$)，P2_dist と P3_dist ($r = 0.723$)，TTR と Guirard ($r = 0.649$)，I_think と Conjunction ($r = 0.613$)，Tokens と JACET8000 ($r = 0.612$) と続く。

目的変数と説明変数の相関が高くなく，いくつかの説明変数の間で高い相関が見られることから，今回の実験データに対して，通常の重回帰分析を用いることが妥当ではないことが分かる。

4.3 習熟度の推定

ランダムフォレストによる予測にあたって，ランダムサンプリングする説明変数の数は，説明変数の数の正の平方根を取り，木の数は 100 とした。また，予測モデルの評価にあたっては，2-fold の交差検証を行った。

表 2 は，構築した予測モデルにおける説明変数の重要度をまとめたものである。なお，表中の重要度 (Importance) は，交差検証の結果として得られる値の平均値である。

表 2　説明変数の重要度

Rk.	Variable	Importance
1	TTR	4.697
2	MLS	3.493
3	Types	2.957
4	Tokens	2.816
5	Readability	2.728
6	Conjunction	2.685
7	P2_dist	2.579
8	Passive	2.492
9	W3_dist	2.351
10	I_think	2.256
11	JACET8000	2.172
12	W1_dist	2.116
13	MLW	2.111
14	P3_dist	1.995
15	Relative	1.469
16	Subordinate	1.428
17	P1_dist	1.426
18	W2_dist	1.387
19	Guirard	1.323
20	Modal	0.843

この表を見ると，ランダムフォレストによる回帰では，異語率や平均文長が予測に寄与していることが分かる。

そして，ランダムフォレストによる回帰の結果として得られる予測値を用いて，*k*-近傍法 ($k = 5$) を行った。表 3

は，79 本の英作文の書き手の習熟度を *k*-近傍法で推定し，交差検証を行った結果である。全体の精度は 58.23%であった。

表 3　推定結果の混同行列

	Level 1	Level 2	Level 3	Level 4	Accuracy
Level 1	12	4	3	1	0.600
Level 2	5	11	3	1	0.550
Level 3	2	5	9	4	0.450
Level 4	0	1	4	14	0.737

5. おわりに

本研究では，ランダムフォレストと *k*-近傍法を用いて，自由英作文における書き手の習熟度を推定した。予測精度は 58.23%で，主に異語率や平均文長などが予測に寄与していることが分かった。

今後の課題としては，データの数を増やすことが考えられる。それと同時に，Crossley *et al.* (2011) のように語彙の意味に関する指標を用いるなど，習熟度の推定に有効な説明変数を再検討する必要がある。

参考文献

Breiman, L. (2001). Random forests. *Machine Learning,* 24, pp. 123-140.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing,* 28, pp. 1-21.

大学英語教育学会基本語改訂委員会 (編) (2003). 『大学英語教育学会基本語リスト—JACET List of 8000 Basic Words』 東京: 大学英語教育学会.

Elliot, S. (2003). IntelliMetric: From here to validity. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Hillsdale: Lawrence Erlbaum Associates.

Ishikawa, S. (2009). Pharaseology overused and underused by Japanese learners of English: A contrastive interlanguage analysis. In Yagi, K., & Kanzaki, T. (Eds.), *Phraseology, corpus linguistics and lexicography: Papers from Phraseology 2009 in Japan* (pp. 87-98). Hyogo: Kwansei Gakuin University Press.

石岡恒憲 (2004). 「記述式テストにおける自動採点システムの最新動向」 『行動計量学』 31(2), pp. 67-86.

石岡恒憲 (2009). 「論述式項目の自動採点」 植野真臣・永岡慶三 (編) 『e テスティング』(pp. 95-120) 東京: 培風館.

木村恵・田中省作・八島等・依田みづき (2009). 「言語資源とその処理技術を活用した L2 語彙の習得レベル判定」 『英語コーパス研究』16, pp. 1-14.

金明哲 (2007). 『R によるデータサイエンス—データ解析の基礎から最新手法まで』 東京: 森北出版.

劉志宇・内田理 (2012). 「日本語を学習する外国人を対象とした日本語テキスト難易度推定手法」 『情報処理学会研究報告』 2012-NL-205, 6p.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19, pp. 313-330.

水本篤 (2008). 「自由英作文における語彙の統計指標と評定者の総合的評価の関係」 『学習者コーパスの解析に基づく客観的作文評価指標の検討』(統計数理研究所共同研究リポート 215) (pp. 15-28).

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly,* 17, pp. 651-671.

杉浦正利 (2008). 「英文ライティング能力の評価に寄与する言語的特徴について」 成田真澄 (編) 『学習者コーパスに基づく英語ライティング能力の評価法に関する研究』 平成 17 年度～平成 19 年度科学研究費補助金 (基盤研究(C)) 研究成果報告書 (pp. 33-58).

田中省作・木村恵・依田みづき・八島等 (2007). 「順序ロジットモデルに基づいた英単語の習得困難度の推定とその要因の分析」 『言語処理学会第 13 回年次大会発表論文集』(pp. 947-950).

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.