Vol.2015-CH-105 No.8
2015/1/31

デジタル技術は人間の知性を再現できるか？ ―自動採点システムの現状と課題―

小林雄一郎[†1]

本稿の目的は，言語評価における自動採点の可能性を模索し，人間による採点と機械による採点の類似点と相違点を探ることである。そして，その目的を達成するために，非母語話者による英語（話し言葉）の自動採点に関する実験結果を示す。

Can Digital Technologies Duplicate Human Intelligence? Current Trends and Issues in Automated Scoring Systems

YUICHIRO KOBAYASHI[†1]

The purpose of the present paper is to explore the possibility of automated scoring in language assessment, and to investigate the similarity and difference between human scoring and automated scoring. This paper also shows the result of a pilot study on the automated scoring of non-native spoken English.

1. はじめに

言語教育の分野では，数多くの言語テストが存在し，中学，高校，大学など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ライティングやスピーキングを評価するという形式を取っている。しかしながら，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。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完全に一致するとは限らない (e.g., Bejar, Williamson, & Mislevy, 2006)。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に基づく自動採点の技術を開発することは，言語教育分野にとって非常に有用なことである。

しかしながら，人間の採点者と完全に代替可能な自動採点システムを開発するには，いくつかの大きな困難がともなう。それは主に，(1) 「デジタル技術が人間の知性を再現できるか」という技術的な問題と，(2) 「そもそも人間の知性とはどのようなものか」という哲学的な問題，の 2 つである。以下，本稿では，実際の自動採点研究の成果を踏まえつつ，これらの問題について議論していく。

2. 自動採点の方法論

自動採点は，学習者が産出した言語データから，対象となる学習者の習熟度が如実に反映されると思われる言語的特徴を抽出し，それらの頻度を統計的に解析するという手続きをとる。また，その目的は，言語学的あるいは教育学的な理論に基づき，あらかじめ定義されたレベル（CEFR における A2 や B1 など）やスコア（TOEIC や TOEFL の点数など）を予測することである。

何らかの言語的特徴を手がかりとして，分析対象とする言語データの所属グループを統計的に予測する方法論は，計量文献学，とりわけ著者推定 (authorship attribution) の分野で発展してきたものである。計量文献学の歴史は古く，その起源は，19 世紀の聖書研究に遡る (e.g., 村上, 1994)。その後，20 世紀後半になると，コンピュータ技術が飛躍的に進歩し，言語データから様々な言語的特徴を自動的に抽出する自然言語処理 (natural language processing) の技術が盛んに研究されるようになる (e.g., Manning & Schutze, 1999)。さらに近年，機械学習 (machine learning) やパターン認識 (pattern recognition) と呼ばれる分野で，統計的にスコアを予測したり，データを分類したりする技術が開発されている (e.g., Bishop, 2006)。自動採点システムでは，前述のような自然言語処理の技術を使って言語的特徴の頻度を算出し，機械学習の技術を使ってレベルやスコアを付与するという方法論が一般的なものとなっている (e.g., Larkey & Croft, 2003)。

3. 人間による採点と機械による採点

言語テスティングでは，(1) 妥当性（適切な言語項目が評価に用いられているか）と，(2) 信頼性（正確に評価できているか）の 2 つが重要となる。そして，一般的に，人間（熟練した評価者）は妥当性に優れているものの信頼性に難があり，機械（自動採点システム）は信頼性に優れているものの妥当性に難があると言われている (Williamson, 2013)。だが，実際は，それほど単純な二項対立ではない。

まず，妥当性に関しては，いかに熟練した評価者であ

†1 日本学術振興会

ったとしても，自分が評価に用いている項目を完全に理解している訳ではなく,「自分が評価に用いていると考えている項目」と「実際に評価に用いられた項目」が一致しないこともある (Kobayashi & Abe, 2014)。次に，信頼性に関しては，これまでの自動採点の研究において，機械と人間による評価の一致度は，複数の人間による評価の一致度と同程度であると報告されている (Shermis & Burstein, 2003)。

4. 自動採点の実際

小林・阿部 (2013) は，日本人英語学習者のスピーキングの自動採点に関する研究である。実験データは，NICT-JLE Corpus (和泉・内元・井佐原, 2004) を用いた。このコーパスは，ACTFL OPI に準拠した Speaking Standard Test (SST) を受験した日本人英語学習者 1,281 人の発話データを書き起したものである。SST の受験者は，1 枚の絵の描写，ロールプレイ，複数の絵を使った物語の作成といった，複数のタスクを 15 分間で行う。NICT-JLE Corpus は，SST を受けた学習者の発話から構築されているため，専門の評価官が判定した 9 段階の習熟度情報 (SST level) が全ての学習データに付与されているという大きな利点を持つ。小林・阿部 (2013) では，その 9 段階の SST level を自動採点の予測対象（目的変数）とした。

表 1 は，NICT-JLE Corpus におけるレベル別の学習者数と語数をまとめたものである。

表 1　SST レベル別の学習者データ

Level	Participants	Tokens
1	3 (0.23%)	428 (0.04%)
2	35 (2.73%)	7,701 (0.81%)
3	222 (17.33%)	95,169 (9.98%)
4	482 (37.63%)	308,177 (32.31%)
5	236 (18.42%)	203,759 (21.36%)
6	130 (10.15%)	130,492 (13.68%)
7	77 (6.01%)	85,309 (8.94%)
8	56 (4.37%)	68,470 (7.18%)
9	40 (3.12%)	54,341 (5.70%)
Total	1,281 (100.00%)	953,846 (100.00%)

（註: 学習者発話のみ，フィラーや繰り返しは削除）

自動採点の研究では，どのような言語的特徴に注目すれば，習熟度を正確に測定できるのか，ということが常に問題となる。学習者の言語を自動評価する場合，人間の評価者と同じ構成概念を用いることが理想ではあるが，人間は自分の評価基準に関する全てを理解している訳ではない (Attali, 2013)。それゆえ，自動評価プログラムを実装するにあたっては，学習者の習熟度と関連性があると思われる言語項目を可能な限り網羅的に考慮する必要がある。これは，予測に用いる説明変数を何にするか，という問題である。

小林・阿部 (2013) では， Biber (1988) が英語母語話者の話し言葉と書き言葉の分析に用いた 60 種類の言語的特徴，そして，総語数，異語数，平均文長の 3 項目を自動採点に用いた。このように様々な言語的特徴を推定に用いることで，学習者のパフォーマンスを多角的に評価することが可能になる。

そして，自動採点に用いるアルゴリズムは，ランダムフォレスト (Breiman, 2001) という機械学習の手法を用いた。この手法の長所としては，予測精度が高いこと，非常に多くの説明変数を効率的に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，などが挙げられる (e.g., Breiman & Cutler, n.d.; Hastie, Tibshirani, & Friedman, 2009)。

表 2 は，前述の 63 種類の言語的特徴を手がかり（説明変数）として，9 段階の習熟度（目的変数）を予測した結果である。

表 2　ランダムフォレストによる習熟度推定の結果

	L1	L2	L3	L4	L5	L6	L7	L8	L9	accuracy
L1	**0**	3	0	0	0	0	0	0	0	0.00%
L2	0	**27**	8	0	0	0	0	0	0	77.14%
L3	0	4	**146**	72	0	0	0	0	0	65.77%
L4	0	0	38	**398**	45	1	0	0	0	82.57%
L5	0	0	0	90	**124**	19	3	0	0	52.54%
L6	0	0	0	14	57	**41**	14	3	1	31.54%
L7	0	0	0	1	14	30	**23**	5	4	29.87%
L8	0	0	0	0	7	15	20	**7**	7	12.50%
L9	0	0	0	0	2	3	11	5	**19**	47.50%

この表を見ると，1,281 人のうち 785 人分の発話データの習熟度が正しく推定されており，全体の精度が 61.28%であることが分かる。

自動採点にランダムフォレストのような機械学習の手法を用いることの利点は，個々の説明変数（言語項目）が目的変数（習熟度）の予測に寄与する度合いを明らかにできることである (e.g., Crossley & McNamara, 2011)。前述のように，人間は自分の評価基準を必ずしも理解している訳ではないため，その評価基準を統計的に推定することの意義は大きい。

図 1 は，63 種類の言語的特徴に関して，習熟度推定における寄与度（ジニ係数の平均減分）の大きい順にプロットしたものである（上位 30 項目）。

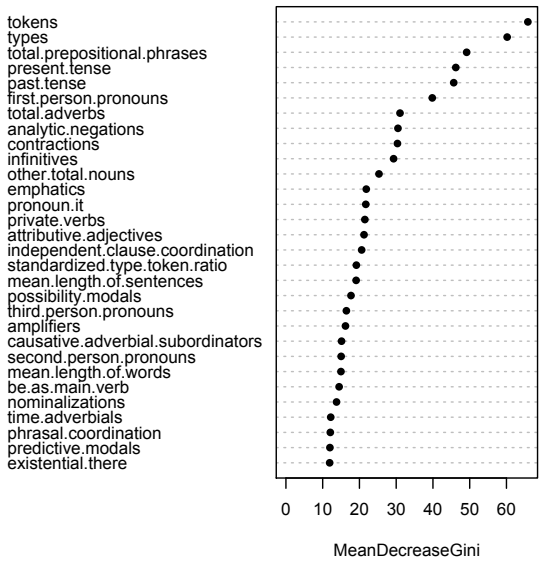

図 1　スピーキングの習熟度推定に寄与した言語的特徴（上位 30 項目）

この図を見ると，上位 10 項目は，総語数 (tokens), 異語数 (types), 名詞句 (total prepositional phrases), 現在形 (present tense), 過去形 (past tense), 1 人称代名詞 (first person pronouns), 副詞 (total adverbs), 否定 (analytic negations), 縮約 (contractions), 不定詞 (infinitives) である。その中でも，リアルタイムでの言語処理が要求されるスピーキングにおいて，限られた時間内にどれだけ多くの語を産出できるかという能力（総語数，異語数）が習熟度に反映されていることが分かる。このように寄与度上位の説明変数を吟味することで，スピーキングを評価する際にどのような言語的特徴に注目すればよいのか，という示唆が得られる。

5. おわりに

前述のように，自動採点システムを開発するには，(1)「デジタル技術が人間の知性を再現できるか」という技術的な問題と，(2)「そもそも人間の知性とはどのようなものか」という哲学的な問題，の 2 つの問題と向き合う必要がある。これらの問題に関して，前節で示した実験結果を見る限り，以下のように言うことができる。

まず，自動採点というタスクに関して，デジタル技術が人間の知性（採点結果）を再現できる割合は，6 割程度である。これは，複数の人間による採点が完全に一致する割合とほぼ同程度である (Page, 2003)。

次に，人間は自分の評価基準を完全に理解している訳ではない。しかしながら，機械学習に基づく自動採点を用いることによって，個々の評価項目が採点に寄与する度合いを明らかにすることができる。

言語的なパフォーマンスの評価は，人間にとっても，機械にとっても，簡単なタスクではない。しかしながら，人間の知性が飛躍的に向上することは期待できず，近い将来にデジタル技術が人間の知性を完全に再現できるかどうかも不明である。従って，現状では，人間と機械が互いの長所を生かし，短所を補うような仕組みを考えていく必要がある。具体的には，自動採点システムは，単に人間による評価を再現するだけでなく (e.g., Bennett, 2006; Bennett & Bejar, 1998)，人間による評価を補完するようなフィードバックができるようになることを目指すべきである。

註

本稿の一部は，小林 (2013) における議論に加筆修正を施したものである。

謝辞

本研究の成果の一部は，科学研究費補助金（特別研究員奨励費（PD 実験））「パターン認識と自然言語処理の技術を用いた習熟度判定」(代表：小林雄一郎) (2012-2014 年度)，科学研究費補助金（若手研究（B））「機械学習によるスピーキングの基準特性抽出と習熟度推定」(代表：小林雄一郎）(2014-2016 年度）によるものである。

参考文献

Attali, Y. (2013). Validity and reliability of automated essay scoring. In Shermis, M., & Burstein, J. (Eds.), *Handbook of automated essay evaluation* (pp. 181-198). New York: Routledge.

Bennett, R. E. (2006). Moving the field forward: Some thoughts on validity and automated scoring. In Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 403-412). Hillsdale: Lawrence Erlbaum Associates.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17(4),* 9-17.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49-81). Hillsdale: Lawrence Erlbaum Associates.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer-Verlag.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-23.

Breiman, L., & Cutler, A. (n.d.). Random forests,

http://www.stat.berkeley.edu/~breiman/RandomForests/ [Online].

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21(2-3)*, 170-191.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Second Edition. New York: Springer-Verlag.

和泉絵美・内元清貴・井佐原均 (編) (2004).『日本人 1200 人の英語スピーキングコーパス』 東京: アルク.

小林雄一郎 (2013). 「機械学習と自然言語処理の技術を用いた習熟度推定―現状と課題」『外国語教育メディア学会 (LET) 関西支部メソドロジー研究部会報告論集』*4,* 12-23.

小林雄一郎・阿部真理子 (2013). 「スピーキングの自動評価に向けた言語項目の策定」『電子情報通信学会技術研究報告』*113(253),* 1-6.

Kobayashi, Y., & Abe, M. (2014). The similarity and difference between human scoring and automated scoring. A paper given at the Applied Linguistic Association of Korea (ALAK) 2014.

Larkey, L. S., & Croft, W. B. (2003). A text categorization approach to automated essay grading. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 55-70). Hillsdale: Lawrence Erlbaum Associates.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing.* Cambridge: MIT Press.

村上征勝 (1994). 『真贋の科学―計量文献学入門』東京: 朝倉書店.

Page, E. B. (2003). Project Essay Grade: PEG. In Shermis, M., & Burstein, J. (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Hillsdale: Lawrence Erlbaum Associates.

Shermis, M. D., & Burstein, J. C. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective.* New York: Routledge.

Williamson, D. M. (2013). Developing warrants for automated scoring of essays. In Shermis, M., & Burstein, J. (Eds.), *Handbook of automated essay evaluation* (pp. 153-180). New York: Routledge.