解説

コンピュータ上で実施する記述式試験
――エッセイタイプ，短答式，マルチメディア利用について――

Computer-based Writing Tests

石岡恒憲

Abstract

自由形式で解答を記述するいわゆる記述解答試験において，エッセイタイプのものと短答式タイプのそれぞれについて，代表的なシステムにおける採点のロジックについて紹介する．短答式解答については，この分野の研究が加速していることから，研究の方向性について説明する．また，全米学力調査（NAEP：National Assessment of Educational Progress）で実施された最新の記述式試験のコンピュータによる出題形式についてスクリーンショットを交えて報告する．最後にセンター試験の後継としての記述テストについて私見を述べる．

キーワード：自動採点システム，エッセイ（小論文）採点，短答式記述採点，記述式試験，全米学力調査（NAEP）

1. ま え が き

オープンエンドクエスチョンといわれる，いわゆる自由形式で書くことのできる記述式試験には大別して二つのタイプがある．一つはエッセイ（小論文）タイプの試験であり，もう一つは短答式の試験である．前者（エッセイタイプの試験）では，基本的に正解がなく，日本語では800字程度から1,600字程度を書く．評価の基準としては，修辞（文のうまさ，語法の使い方など），論理（論の進め方，論の掘り下げ，例示の使用など），内容（プロンプトと呼ばれる論題に十分に応えているか）などが用いられる．このタイプの試験についてはEducational Testing Service（ETS）のe-rater[(1)]やVantage Learning社のIntelliMetric[(2)]などハイステークスな試験に用いられている実用的なシステムがあり，人間（専門家）による評価に十分に近いことを示す多くの例証がある．日本語を処理するシステムとしては筆者らが開発したJess[(3)]がある．他方，短答式試験は，通常，望まれる正解がある．したがって，システムは用意している正解と回答が同義であるか否かを判定する．正解は通常，1文か多くて2文である．この短答式記述の自動採点については，エッセイタイプの自動採点ほどには研究が進んでおらず，知られているシステムとしてはPacific Metrics社のCRASE[(4)]やETSのc-rater[(5)]など僅かである．またこれらは恐らくまだハイステークスな試験には用いられていない．（開発元のWebページにはそのような情報の記述は認められない．）しかしながら，この「正解」と「解答」の同義判定やその含意関係を判定する技術は，「含意関係認識（Recognizing Textual Entailment）」と呼ばれ，今自然言語処理の分野で最もホットな話題の一つになっている．このため，この短答式記述の自動採点の研究は急激に進歩することが期待されている．

本稿では数式を除くコンピュータによる記述式試験について現在の技術を解説する．2.では，エッセイタイプの試験を採点する代表的な自動採点システムを列挙し，それぞれがどのような観点から回答を評価するのかについて解説する．3.では，短答式記述の試験について，現在行われている評価方法と，解決すべき課題について報告する．4.では全米学力調査（NAEP：National Assessment of Educational Progress[(6)]）で実施された最新の記述式試験のコンピュータによる出題形式について報告する．ただしNAEPの採点は機械ではなく人が行う．5.ではまとめに代えて，センター試験の後継としての記述テストについて私見を述べる．

石岡恒憲　独立行政法人大学入試センター研究開発部
E-mail　tunenori@rd.dnc.ac.jp
Tsunenori ISHIOKA, Nonmember（Research Division, The National Center for University Entrance Examinations, Tokyo, 153-8501 Japan）.
電子情報通信学会誌　Vol.99 No.10 pp.1005-1011 2016年10月

表1　エッセイ評価システムの比較

評価システム	開発	評価基準	手法	特記事項
AutoScore	American Institutes for Research（AIR）	意味概念／段落間の意味的つながり／語の多様性／文法エラー	統計的手法	採点基準は論題依存
LightSIDE[8]	カーネギーメロン大学	内容／文体／構造／態	教師あり機械学習	オープンソース
Bookette	CTB/McGraw-Hill	構造／文法／意味／技巧	ニューラルネット	90の特徴量
E-rater[1]	ETS	構造／組織化／内容	重回帰モデル	12の評価指標
Lexile Writing Analyzer[9]	MetaMetrics	語彙使用の多様性／繰り返し使われる語彙の出現度合／文章としての流ちょうさの抑制	統計的手法	学年（grade），ジャンル，論題，句読法（punctuation）によらない
PEG[10]	Measurement Inc.	構造／組織化／形式／技巧／独創性	重回帰モデル	意味理解に着手中
Intelligent Essay Assessor, IEA[11]	Pearson Education	内容／文体／技巧	潜在的意味解析（LSI）	論理構成／語の出現順を評価しない
CRASE[12]	Pacific Metrics	アイデア／文章の流ちょうさ／組織化／態／語彙選択／慣習／プレゼンテーションのうまさ	機械学習＋統計（ベイズアプローチ）	Java言語で実装
IntelliMetric[2]	Vantage Learning	一貫性／内容／構成／文章の複雑さ／アメリカ英語への適応	ルール発見	論題ごとに大量のデータが必要
Jess[3]	Ishioka & Kameda	修辞／構成／内容	統計（異常値検出）	2016年からオープンソース

2. エッセイ（小論文）採点システム

2012年にヒューレット財団がスポンサーとなりAutomated Student Assessment Prize（ASAP）[7]と呼ばれるKaggle（企業や研究者による最適モデルのコンペ）が実施された．エッセイ採点については154のチーム（196の参加者）が参加し，八つの論題に対してエッセイ採点システム（AES：Automated Essay-scoring System）の性能を競った．八つの論題のうち四つは伝統的な作文ジャンル，すなわち意見文（persuasive），説明文（expository），叙述文（narrative）から出題された．残りの四つは素材に基づく（source-based）質問であり，素材文について誘発される質問について答えるものである．被験者は7，8，10年生（中1，中2，高1年生に対応）である．九つのAESベンダは招待された．うち一つはオープンソースであり商用ベンダは八つであるが，この八つで現在のアメリカにおける自動採点市場の97%を占めるとされている．このKaggleに招待された九つのAESを表1に示す．

なお，AESで処理をするために手書き文字をアスキー化する必要がある．英数字の認識率はある会社では98.12%であり，別の会社に再変換してもらったところ99.82%になったとしている．

これらの九つのシステムについて順に紹介する．

AutoScoreはAmerican Institutes for Research（AIR）が開発したスコアエンジンである．このシステムでは論題（prompt）ごとに採点基準を統計的尺度として作成する．この採点基準は既知の，かつ有効な採点スコア（機械学習でいうところの正確な教師データ）を基に作成される．採点基準には

- 高得点と低得点を識別する概念に基づいた意味的尺度
- 段落間の解念的なつながりを示す意味的尺度
- 語の使用の幅や文法的尺度

に加え明瞭性や主張性なども利用され，これらを統合して尺度（proxy）スコアを予測するとしている．

LightSIDE[8]はカーネギーメロン大学のTELEDIA研究室で開発されたオープンソースソフトウェアパッケージである．これはテキストマイニングの技術を非専門家でも様々な分野で簡単に利用できるよう指向したツールで，エッセイ評価も行うことができる．LightSIDEには様々なオプション，例えばデータ表示や機械学習，視覚化などの拡張機能のプラグインが用意されている．しかしながらASAPコンペにおいてモデルを訓練するための追加のプログラムはなく，標準の仕様でモデル構築を行った．

BookettはCTB/McGraw-Hillが提供する採点エンジンであり，モデル特性や総合点スコアにおいては，専門家（エキスパート）の人間の評価者と同等の信頼性を持つとしている．この採点エンジンはニューラルネットを用いた自然言語処理システムで，専門家の人間スコアをモデル化する．CTBは2005年からクラスルームにおいて，2009年から大規模試験において自動採点評価を使ってきた．採点は論題とジャンル依存で，採点のための各種パラメータを調整する．論題による調整によって人間採点をより良く再現するとしているが，ジャンルによる調整は，論題ほどには再現性が良くない．CTBのBookette エンジンは，90の構造的，文法的，意味的，技巧ベースの特徴量を取り扱うとしている．

E-rater[1]は Bill Burstein が開発し ETS が提供する AES である．2004 年から Ver. 2 が提供され，僅か 12 の変量によって採点を行うシステムに改変された．1999 年に e-rater は公的な（＝ハイステークスな）試験である GMAT（ビジネススクール入学のための共通試験）において二つの採点者のうちの一方として使われ，その後，GRE（自然科学系の大学院入学のための共通試験）や TOEFL（英語を母語としない人を対象とした英語の試験）においても同様の使われ方をしている．

Lexile Writing Analyzer[9]は学年（grade），ジャンル，論題，句読法（punctuation）によらない採点エンジンである．開発元である MetaMetrics 社が独自に作ったレキシル作文尺度（Lexile Writer measures）に基づき採点を行い，人間の評価者による評価データを必ずしも必要としない．テキストの複雑度とテキストの特徴，及び作文能力の関係を調べたところ，導かれた有意な説明因子は多くはなく，語彙使用の多様性や，繰り返し使われる語彙の出現度合い，文章としての流ちょうさの抑制，といった非常に少ない数の結合でエッセイ分類のばらつきの 90% を予測できるとしている．

Project Essay Grade（PEG）[10]は，40 年以上前から Ellis Page により開発された AES の草分けであるが，2002 年から Measurement Inc. がその権利を得て，現在も拡張を続けている．近年，Measurement Inc. は，より深層的な意味レベルでテキストを高次元データを用いて評価できるツール群の開発を行った．これにより雑音を含んだことを前提とした高次元データによる予測アルゴリズムを考案し，統計的機械学習における進歩を目指すとしている．

Intelligent Essay Assessor（IEA）[11]は 2000 年当時，情報検索で主流になりつつあった潜在的意味解析（LSI : Latent Sematic Indexing）をいち早くエッセイ採点に取り入れたシステムである．開発は Foltz と Landauer であるが，現在は Pearson Educational Technologies が提供する商用システムや試験で用いられている．言語空間を構成するのに用いている語彙数は多く，1,200 万語である．IEA は現在，英語のみならず，スペイン語，アラビア語，ヒンディ語でのテキストをあらゆる分野において評価する．また English Language Arts（ELA）など高校生のカリキュラム教育のフィードバックシステムとして使われている．

CRASE[12]は Pacific Metrics 社による採点エンジンであり，大規模試験におけるそれぞれの質問に対し採点を行う．採点できる質問の種類には，（a）エッセイ論題のほか，（b）数学，ELA，理科についての短答記述，（c）数式解答，（d）コンピュータを用いた回答（ドラッグ＆ドロップ，項目と項目の結び付け＝グラフィング）などがある．このシステムの採点アルゴリズムとしての最大の特徴は，機械のみで採点するモデルのほか，人間のスコアと機械のスコアの両方を融合させて機械の採点を行うモデルの二つがあることである．Pacific Metrics 社は後者をハイブリッドモデルと呼んでいるが，機械学習の用語で言うところの半教師学習（semi-supervised learning）ということができる．スコアの分かっているいわゆる教師データに加えて，スコアの分からない（がデータの存在が分かっている）データを加えることで推定の精度を向上させる仕組みになっている．またモデル構築に際し，事前にスクリーニングを行い，モデルに組み入れるべきでないデータを除外する仕組みが組みこまれている．採点の観点としては，6+1 評価観点モデルと呼ばれるアイデア，文章の流ちょうさ，組織化，態，語彙選択，慣習といった六つの作文についての評価観点と，+1 のプレゼンテーションのうまさ（written presentation）が用いられる．採点は 1-4 点，あるいは 1-6 点に調整されるが，採点にはベイズアプローチが用いられている．CRASE は Java 言語で記述され Web サービスとして提供されている．

IntelliMetric[2]はルール発見に基づく人工知能をベースとした AES である．Vantage Learning 社が巨額の予算を投じて集中的に開発した．開発者によれば，IntelliMetric の採点の仕方は人間の採点のシナリオに基づいていると言う．すなわちエキスパートの採点者はそれぞれのスコアを得たエッセイを分析し，それによって採点の基準を知り，その後，それに基づいて採点をするとしている．

これらのシステムに関する Kaggle の調査は Academic Advisory Board の委員長を務めた Sharmis 博士らによる報告[13]に詳しいが，結果は AES が人間の評価者に比べ信頼できるというものである．ただそれに対しては多くの反論がある．その典型は「コンピュータは文を正しく読めているわけではなく，有効な書き言葉による伝達の本質を測ることができない」というものである．MIT の研究者で AES の批評家として有名な Les Perelman 氏によれば，測定できない本質として，書かれている内容の確かさ（accuracy），論法（reasoning），証拠の適切性（adequacy of evidence），良識（good sense），倫理的スタンス（ethical stance），説得力（convincing argument），（文のまとまりとしての）意味のある組織化（meaningful organization），明瞭性（clarity），誠実さ（veracity）を挙げている．

日本語を処理する AES として筆者らによる Jess[3]がある．毎日新聞の社説やコラムから良い作文の作法を学習しており，統計学で言うところの異常値発見に基づいて採点を行う．人間の評価者による，いわゆる教師データを必要とせず，小規模な試験でも特段の準備を必要とせずに利用できる．これについても，AES 相互の比較のために表 1 に追記しておく．

3. 短答式記述採点システム

短答式記述採点を可能にする最も正統的なアプローチは含意関係認識であると思われる．この分野については現在，国立情報学研究所が主催する国際ワークショップ「NTCIR」の中で，含意関係認識に関するタスク「RITE」を2011年から実施している．そこに示されている例題[14]は，以下の二つの文章について，含意関係が成り立つかどうかを判定するものである．

t1　鎌倉幕府は1192年に始まったとされていたが，現在では実質的な成立は1185年とする説が支配的だ．
t2　12世紀に日本では鎌倉幕府が開かれた．

人間ならばt1が成り立つとき，t2も成り立つことを容易に判断することができる．しかしコンピュータがこれを判断するには，まず「鎌倉幕府（が）1185年（に）成立した」といった意味構造を正しく解析することに加えて，「1185年が12世紀である」という時間情報処理や「成立する≒開く」の含意関係知識が必要となる．含意関係認識は文章のレベルでコンピュータが人間の言葉の意味を理解することを目指しており，もしこれが実用レベルにまで達すれば，短答式記述の自動採点はほぼ実現されるといってよいだろう．ただ筆者の認識する限り，含意関係認識はまだ開発途中で，達成すべき課題や言語資源の整備が更に必要であると思われる．

含意関係認識が現在どのようなレベルにあるかを知るためには，受験ロボットによる話が分かりやすいだろう．同じ国立情報学研究所の「ロボットは東大に入れるか」プロジェクト[15]では，昨年（2015年）ベネッセのセンター試験模試である「進研マーク模試」をロボットが受験した．世界史Bの試験問題には国内外の七つの大学及び研究機関が参加し，その性能を競った．その最高点76点（100点満点，偏差値66.5）を取ったのは日本ユニシスのグループで，彼らは以下の三つの手法を組み合わせて解く方策を用いている．

① 質問応答
② 単語の相関
③ 構文木のマッチング

このうち①は設問選択肢の固有名詞を一つずつ隠し，隠した語を問う仮の問題を作った後，質問応答システムで隠した語彙を解答させ，それが元の隠した固有名詞と一致したならば，その設問選択肢は正しいと判断するものである．したがってこの手法は誤答（誤っている文）の検知に有効な方法である．②の方法は，単語の共起確率で設問選択肢文の正しさを評価するものである．①と後述する③の方法がピタリと適応した場合には絶対的な強みがある一方で適応しない場合もあり得て，その場合でもこの②の方法は適応できる．③の方法はいわゆるフィルモアの格文法によって二文（教科書等にある正しい文と設問選択肢の文）の一致性を判定するものである．格文法とは，文の意味の理解の中心に動詞を据えるもので，主格である「ガ格」や目的格「ヲ格」，場所格「ニ格」などの一致を判定する．もちろん，その場合にも同義語辞書などを利用する．

現時点では，構文木の適切な解釈による正統的な含意関係認識よりも，アンサンブル学習的でこのようなアドホックな方式が優勢であるように思われる．

4. NAEPにおける記述テスト

2011年のNAEP[6]では，8年生と12年生に対して，作文テストが従来の紙筆テストに代わり初めてコンピュータによって実施された．そのテストは，単に従来の紙と鉛筆をコンピュータに置き換えただけのものではない．現在のディジタル技術が十分に活用できるよう，取り扱う作文のタイプには，テキストによる問い掛けのほか，写真を含むもの，音声によるもの，ビデオを見て問い掛けに答えるものの四つがある．このうち二つが釣り合い型不完備ブロック計画（BIB design : Balanced Incomplete Blocks design）に基づいて出題される．不完備というのは四つから二つを選ぶ組合せの一部のみが指定されることによる．被験者によって異なる組合せのブロックを与えることで，NAEPテストに割く時間を最小限に抑えながら，広範囲な学力を正確に測ることを目的としている．

以下，解答における操作を示す．図1は写真を含むテキストを見て，解答する問題であるが，解答ウィンドウの左には五つのアイコンが示される．図1ではこのアイコンを，画面左端に拡大して表示している．上から，「ボリューム調整」「読み上げ」「フォントサイズ調整」「マーカ」であり，左下にテストの残り時間を表示する「時計」の各ボタンが用意されている．文字の部分を選択（クリック）し，「読み上げ」ボタンを押すことで，選択部分を読み上げてくれる．音量は「ボリューム調整」ボタンのスライダを用いて調整する．

図2は写真を含むビデオを見て，解答する問題であるが，ビデオ画面中の矢印ボタンを押下することで試験は始まる．ビデオ音声についてはクローズドキャプションにより文字で表示することできる．ビデオ画面の下にある設問文については，テキスト問題と同様に，「読み上げ」や「マーカ」等が利用できる．図2の画面右側では，解答するワープロ画面を表示させ，ビデオを見ながら，解答することができる．ビデオは，適宜，止めたり，意図する箇所に移動したりすることができる．

ビデオ画面を閉じて，ワープロ画面のみを全画面に切

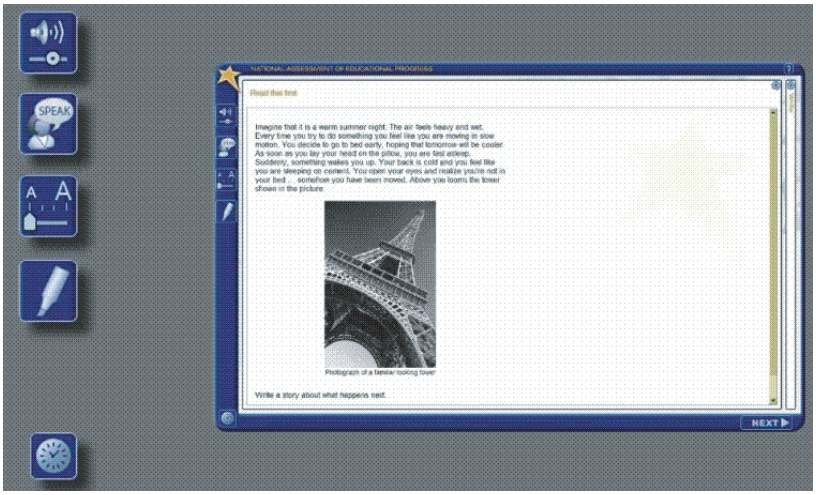

図 1　設問表示で使える各種ボタン

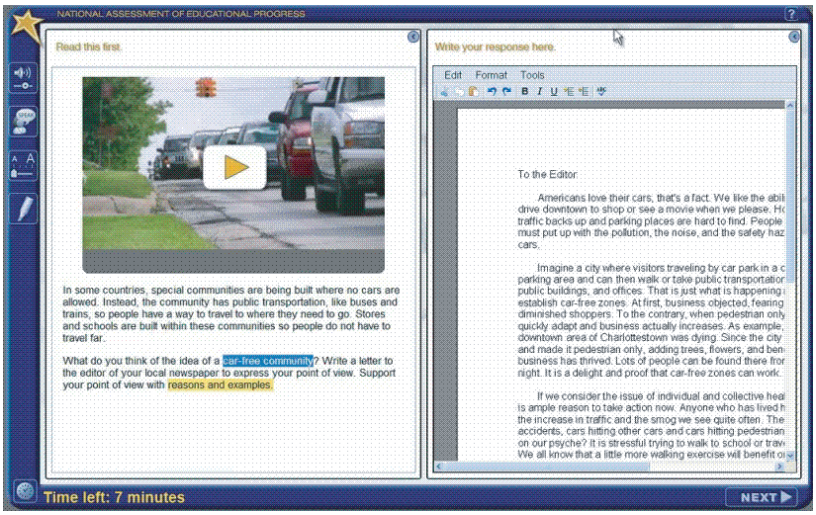

図 2　ビデオを見て解答する

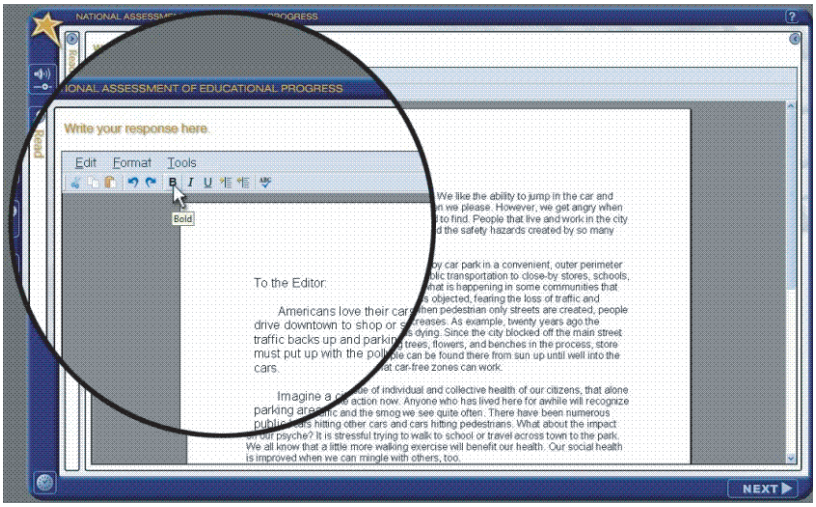

図 3　Word に似たワープロ画面

り替えることもできる（図 3）．このワープロ画面は，Microsoft Word に見た目が似ており，カット＆ペーストや，「繰返し」「元に戻す」ができる．また，太字やイタリックなどの簡単な文字修飾，インデント／アウトデントに加えて，スペルチェックのボタンが用意されている．このボタンによる操作は全て，プルダウンメニューによる代替操作が可能である．

この作文解答において特筆すべきは，障害者に対する配慮である．前項で見てきたように，出題文の音声読み上げや，「フォントサイズ調整」ボタンで 48 ポイントサイズにまで拡大したり，「マーカ」ボタンで地色を黄色にしたりすることができる．NAEP では障害者に対しても代替問題を使用せず，健常者と同じ問題を解答させることを基本としている．そのやり方が公平性を担保し，障害者の尊厳を損ねず，また作題や採点の手間を大幅に軽減するとしている．

5. センター試験の後継としての記述テスト

文部科学省・高大接続部会の最終答申において，新テストに記述式テストが導入されることが決定し，採点業務の効率・安定化のためにコンピュータの効果的な活用や人工知能の利用が盛り込まれた．本章では，センター試験の後継として用いられる記述テストについて，一研究者としての立場から私見を述べる．ただしこれは大学入試センターとしての見解ではない．

5.1 記述テストでのみ思考力が測れるのか

始めに指摘しておきたいのは，記述テストが思考力，判断力，表現力を測る唯一の手段ではない，ということである．従来のセンター試験でも，すなわちマークシート式でも，思考力や判断力が測れるよう歴代の作題委員の創意工夫が重ねられてきている．記述テストは現実の質問応答を具現した形で出題できるために，より真正な学力測定ができると考えられている．筆者もそれに異を唱えるものではないが，記述テストでマークシートでは測れない思考力や判断力が測れるようになると考えるのは全くの幻想である．出題の形式（記述式かマークシート式か）が問題なのではない．むしろ問いの設定の仕方や，素材の工夫が，思考力や判断力を測定するか否かを決定する．

そもそもテストは，学力測定そのものである．もちろんテストにはどのような分野の問題を問うのかによって受験生に伝えたいメッセージというのがあるのかもしれない．ただテストを学力測定のツールとして考えるならば，テストの構成は測定誤差の少なくなるように工夫する必要があり，それにはある程度の設問数が必要である．記述式は一般に解答時間を多く要するために，記述式を導入することにより設問数は減らさなくてはならない．試験としての妥当性を減じてまでも記述式を導入すべきかは十分検討しなければならない．またマークシート方式が多くの設問数を設けることを可能にし，それゆえに高校指導要領に沿った必要十分な領域のテストを可能にしていることにも留意しなければならない．

5.2 短答式記述テストの自動採点は実用的か

短答式記述テストについては筆者らの研究グループも今懸命に研究を進めている．しかしながら，技術的にも自然言語の意味理解は難しく，採点はギミックの積み重ねである．本来，測定したい変量を測定する代わりに，それに関連する（多くは相関する）と思われる複数の変量を収集し，それを最終的な人による採点結果を教師データとして機械学習を行うことにより採点を実施する．したがってこれは万能ではなく，推定精度の点においてもいまだ不十分である．今人工知能と呼んでいるものの多くは，「何らかの前処理や工夫をした上での機械学習」という表現に置き換えてもよいと思われる．少なくとも教師あり機械学習は「ルール発見」や「ルール構築」と言ってよいだろう．ルールになるためには，そのように答えるある一定数が必要であり，例外には対応し切れない．実際，2012 年のヒューレット財団による ASAP でも，機械による判定で最大 2% までの「判定不能」が許されている．したがって，教師ありデータによる機械学習の方法を取る限り，自動採点は決して万能にはなり得ず，機械による採点ミスを許容できないなら，人間との併用は避けられないだろう．

事実，アメリカで公的な試験に用いられているエッセイの採点は人との併用である．人間と機械でそれぞれ独立に採点し，6 点満点中 1 点以内の違いならばその平均を最終点とする．2 点の違いがあったなら，第 3 者による人間が採点し，調整点を決定する．これより，我が国においても短答式記述テストの自動採点が行われた場合には機械と人間の併用が現実的であり，これこそが受け入れやすい形式と思われる．その併用の仕方には

① 人と機械が独立に採点
② 機械採点（評価指標の数値及び最終得点を提示）の人間によるチェック
③ 人間採点の機械によるチェック

の三つが考えられる．ただこのうちどれにするかは「記述の採点は本来どうあるべきか」という哲学的な判断を含めて，十分な討議がなされるべきだろう．

5.3 字数について

解答の字数については，高大接続部会の会議の中でも様々に揺れている．30 字程度から最大 300 字程度までの幅があるようである．しかしこの字数は「人工知能の

利用」という観点から議論されたものではない．ただ自然言語処理技術の立場からみれば，この字数の問題は極めて重要である．30字ならば模範解答におけるキーワードの抜き出しと若干の言い換えなどの表層的な字句のパターンマッチや格文法辺りで済む可能性はあるが，「より多い文字数の記述」では言い換え辞書，シソーラスの活用が必須となり，困難度が指数関数的に増大する．幾つかの議論が現状技術についての正しい理解を踏まえた上での討議になればよいと期待する．

付記　本稿は2016年2月13日（土）～14日（日）に東京農工大学小金井キャンパスで開催された情報処理学会133回コンピュータと教育研究会発表会における筆者による招待講演の予稿に加筆修正を行ったものである．講演の機会を設けて頂いた東京農工大学工学部情報コミュニケーション工学科の中川正樹教授に感謝申し上げる．

文　　献

（1） Y. Attali and J. Burstein, Automated essay scoring with e-rater v. 2.0 (ETS RR-04-45), Educational Testing Service, Princeton, NJ, 2005.
（2） S. Elliot, “IntelliMetric from here to validity,” in Automated essay scoring : A cross disciplinary perspective, M. Shermis and J. Burstein, eds., pp. 71-86, Lawrence Erlbaum Associates, Hillsdale, NJ, 2003.
（3） T. Ishioka and M. Kameda, “Automated Japanese essay scoring system based on articles written by experts,” Proc. of the 21st Intl Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics (Coling-ACL 2006), no. P06-1030, pp. 233-240, 2006, http://www.aclweb.org/anthology/P/P06/P06-1030.pdf
（4） Pacific Metrix, CRASE, 2012, https://www.pacificmetrics.com/products-and-solutions/crase/
（5） C. Brew and C. Leacock, “Automated short answer scoring, principles and prospect,” in Handbook of Automated Essay Evaluation, M. Shermis and J. Burstein, eds., pp. 136-152, Routledge, New York and London, 2013.
（6） National Assessment of Educational Progress, NAEP, 2011, http://nces.ed.gov/nationsreportcard/E.B.
（7） The Hewlett Foundation, Automated Student Assessment Prize (ASAP), 2012, https://www.kaggle.com/c/asap-aes
（8） E. Mayfield and C.P. Rosé, “LightSIDE : Open source machine learning for text,” Handbook of Automated Essay Evaluation, M. Shermis and J. Burstein, eds., pp. 124-135, Routledge, New York and London, 2013.
（9） M.I. Smith, “The reading-writing connection,” Position papers, The Lexile Framework for Reading, MetaMetrics, 2009, https://d1jt5u2s0h3gkt.cloudfront.net/m/uploads/positionpapers/TheReading-WritingConnection.pdf
（10） E.B. Page, “Project essay grade : PEG,” Automated essay scoring : A cross disciplinary perspective, M. Shermis and J. Burstein, eds., pp. 43-54, Lawrence Erlbaum Associates, Hillsdale, NJ, 2003.
（11） K.T. Landauer, D. Laham, and W.P. Foltz, “Automated scoring and annotation of essays with the intelligent essay assessor,” Automated essay scoring : A cross disciplinary perspective, M. Shermis and J. Burstein, eds., pp. 87-112, Lawrence Erlbaum Associates, Hillsdale, NJ, 2003.
（12） S.M. Lotridge, E.M. Schulz, and H.C. Mitzel, “Using automated scoring to monitor reader performance and detect reader drift in essay scoring,” Handbook of Automated Essay Evaluation, M. Shermis and J. Burstein, eds., pp. 233-250, Routledge, New York and London, 2013.
（13） M.D. Sharmis and B. Hamner, Contrasting State-of-Art Automated Scoring of Essays : Analysis, Contrasting Essay Scoring, 2012.
（14） 国立情報学研究所，“問われるのは意味を理解する力．暗記だけでは解けない社会科科目，［特集］人工頭脳プロジェクト「ロボットは東大に入れるか」，” NII Today, no. 60, pp. 8-9, 2013, http://www.nii.ac.jp/userdata/results/pr_data/NII_Today/60/p8-9.pdf
（15） ロボットは東大に入れるか　Todai Robot Project, http://21robot.org/

（平成28年5月6日受付）

石岡（いしおか）　恒憲（つねのり）

昭58東京理科大・工・経営卒．昭60同大学院修士課程了．同年㈱リコー入社．ソフトウエア研究所．以来，UNIX上のDBMS応用システム，例えば図書館システムLimedioの設計開発等に従事．平10文部省・大学入試センター助教授．その後省庁再編を経て現在，独立行政法人大学入試センター教授．工博．平12文部省長期在外研究員（カーネギーメロン大Language Technologies Institute客員研究員）．平24から東工大大学院社会理工学研究科連携教授．