# 小論文自動採点

Automated Essay Scoring

石岡恒憲

Abstract

本稿では代表的なエッセイの自動採点システムであるE-rater，PEG，IntelliMetric，IEA，BETSY，Jessについて取り上げ，最新の仕様を報告する．従来の研究ではこれら自動採点システムと人間との採点スコアには強い相関があり，スコア間には有意な差がないことが報告されてきたが，より詳細な最新の比較研究や，従来とは異なった結果を導いた研究について報告する．また自動採点システムに対する批判について整理する．

キーワード：エッセイ，公的共通テスト，自然言語処理

## 1. はじめに

アメリカにおいてはコンピュータによるエッセイの自動採点システムは既に実用段階にあり，共通テストとしての公的試験に実際に用いられている．アメリカの経営大学院（ビジネススクール）入学のための共通試験であるGMAT（Graduate Management Admission Test）における作文試験AWA（Analytical Writing Assessment）の採点にはETS（Educational Testing Service）で開発されたE-raterが1999年から使われている．2006年1月からGMATの開発及び運営はETSからPearson VUE & ACT Inc.に移り，これに伴いAWAエッセイの採点はVantage Learning社が開発したIntelliMetricが行うようになった．IntelliMetricはアメリカの医学大学院進学のための適性試験MCAT（Medical College Admission Test）の作文試験の採点にも2007年から用いられている．更にはカレッジ入学のためのインターネットベースのテストであるACCUPLACERプログラム中の作文テストであるWritePlacer Plusでも使われている．

また2007年3月14日のEducation Week誌の記事によると，2011年の初めには全米学力調査（NAEP：National Assessment of Educational Progress）で8年次と12年次の学生に対し従来の紙筆テストに替わってキーボード入力による作文テストを実施することが決定した（4年次の学生に対しては2019年度から）．この新しい作文評価では，1回の試験で30分のエッセイテストを2課題実施し，議論の掘り下げや説明力，経験の伝達力などを測定する．新しい作文テストの枠組みは，ACT Inc.が検討していることから，NAEPにもコンピュータによる自動採点（Automated Essay Scoring，以下AESと略記）が用いられることはほぼ間違いないと思われる．

AESが広く用いられる理由は幾つかが考えられるが，Rudner [1]によるとAESは，教師，テスト業者，研究者の3者にとって魅力的であるとしている．教師にとっては，多くのエッセイを読み採点するという重荷から開放され，問題作成やより深い理解に傾注することができるようになる．テスト業者にとってみれば，低コストで高品質の採点を可能にする．研究者にとっては，この魅力的な研究分野に自らの研究を統合させることができるであろう．

これ以外にもAESは，評定の系列的効果（小論文の評定が答案の中で何番目に行われたかにより評定が変わる），課題選択（異なる課題に基づいて書かれた小論文をどう一元的に評価するか，どのように等化をするか）などの問題を排除できるだけでなく，対話的な作文指導ができるといった点で，極めて有効であると考えられている．近年では説明責任といった点からも重要である[2]．

もちろんAESの妥当性は十分に検討されるべきである．AESの果たすべき最終目標は，詰まるところ，AESをいかに人間に近付けるかである．しかしこれは必ずしも人間が用いるのと同じテクニックをAESが用

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.92 No.12 pp.1036-1040 2009年12月

いることを意味しない．人間は典型的には，パッセージを読み，主題を探し，採点基準表に従ってパッセージに現れる内容や表現，及び言語スキルを評価する．一方，コンピュータは同じようにエッセイを評価することをしない．むしろ AES はコンピュータ特有の能力であるところのストップワードの同定や文と文のつながり具合など重要な特徴量の抽出で代表されるような，表層的な特徴量をカウントする．したがって，問題となるのは AES の採点結果が受容可能であるかどうかである．

もし最終目標が人間のスコアに十分近いかであるならば，答えは既にイエスである．Keith [3] によれば，人間同士のスコアの相関は .70 から .90 であり通常は .80 から .85 である．AES と人間のスコアの相関は人間同士の相関と差異がないという．また Page [4] によれば，PEG（Project Essay Grade）はチューリングテスト（Turing test：コンピュータの応答が人間のものと区別できるかどうかを判断するテスト）において，人間との差異を見つけられないとしている．これらの事柄は恐らくおおむね正しく，それゆえに現在，公的な試験に AES が採用されるようになっているのだと思われる．

現在，アメリカでは商用のシステムとして E-rater and Criterion [5]，PEG [4]，IntelliMetric and MY Access! [6]，IEA（Intelligent Essay Assessor）[7] がある．更にオープンソースとしての BETSY（Bayesian Essay Test Scoring System）[8] を加えた五つが比較的有名なシステムということができる．日本語小論文を取り扱うシステムとしては Jess（Japanese automated Essay Scoring System）[9],[10] がある．公的な試験の幾つかが E-rater や IntelliMetric により実施されるようになってから，10 年近くを経ているために，今ではこれらのシステムの比較やサーベイ論文も比較的，容易に得ることができるようになった（例えば文献(11)，(12)など）．日本語による初のサーベイ論文としては石岡 [13] があり，その後の進展については石岡 [2] にある．

そこで，本稿では，各エッセイシステムの現状の仕様について簡単に整理し（2.），最近の比較研究の中から興味深いと思われる結果について紹介しておく（3.）．4. には，作文教師を中心として今なお指摘されている AES についての批判や妥当性についての論議について整理しておく．

## 2. 代表的な自動採点システム

### 2.1 E-rater and Criterion

E-rater は世界最大のテスト機関である ETS の Burstein らの研究グループによって開発されたシステムであり，2004 年に新バージョン（Ver.2.0）が開発された [5]．E-rater では複数の言語上の特徴量に基づく重回帰によってスコアを計算するが，Ver.2 では用いられる変量の数が Ver.1 時代の 60 余りからわずか 12 に厳選され，論題によらずに固定となった．その 12 の変量とその変量に係る重みは表 1 に示すとおりである [5]．この重みは経験則によって定められている．

表 1　E-rater Ver.2 で用いる変量とその重み

| 変量 | 重み |
|---|---|
| (1)総ワード数に対する文法エラーの割合 | 0.05 |
| (2)総ワード数に対する語の使用法についてのエラーの割合 | 0.02 |
| (3)総ワード数に対する手順のエラーの割合 | 0.07 |
| (4)総ワード数に対するスタイルについてのエラーの割合 | 0.08 |
| (5)談話（discourse）ユニットの数 | 0.21 |
| (6)各ユニットにおける平均のワード数 | 0.12 |
| (7)当該エッセイの 6 点法によるコサイン類似度が最大となるスコア点 | 0.04 |
| (8)最高点（通常 6 点）を得たエッセイとのコサイン類似度 | 0.07 |
| (9)単語の繰り返しの程度を示す指標：全ワード数(token)に対する異なったワード種類(word type)の割合 | 0.08 |
| (10) Breland ら [14] の単語頻度指標に基づく語彙の困難度 | 0.03 |
| (11)平均の単語長さ | 0.03 |
| (12)単語の総数 | 0.20 |

Criterion は E-rater を採点エンジンとする Web アプリケーションである．400 を超える種々な論題が用意されており，これらを使うことができる．また ETS は Critique という E-rater の機能の一部を実装した作文分析ツール（Critique Writing Analysis Tool）を提供している．Critique は，文法，語の使用法，技巧，文体，組織化，展開などに対するリアルタイムのフィードバックを返すものであるが，現在では，ユーザの作文レベルに応じて返すコメントを変えるように改良されている．作文レベルには，小学生 (4-5 年生)，中学生 (6-8 年生)，高校生（9-12 年生），カレッジ（1-2 年生），上級 / 大学院受験 (GRE 相当)，非英語圏対象の英語 (TOEFL 相当) の各レベルがある．あるトピックにつき最低 465 編の採点 / 学習の後，提供されるとしている．

### 2.2 PEG

PEG はエッセイ評価のために開発された最初のシステムであり，Page によって最初のバージョンが 1966 年ごろに開発された．このバージョンでは，proxes と呼ばれる約 30 の特徴量が用いられ，これらを trins と呼ばれる本来測定しようとする作文能力を表す指標の代用とした．これら特徴量に係る重みを計算するために，E-rater 同様に重回帰モデルが用いられている．

PEG は 1993 年に改訂され [4]，最新の版 [15] では，PEG は総合点に加え，内容，組織化，スタイル，メカニクス，

独創性などの項目別のスコアを提供している．ほかのシステムにない革新的なこととしては，生徒の（作文上の）長所と短所について，より詳細なフィードバックを返すようにしている．

しかしながら，PEGの項目別スコアや総合スコアを構成するために，どのような特徴量が用いられているかについては公開されていない．

### 2.3 IntelliMetric and MY Access!

IntelliMetricはVantage Learning社によって，エッセイもしくは自由回答形式（open-ended）の問題に対する採点のために開発された．IntelliMetricは，開発者サイドが自称するところの知能に基づいた（brainbasedあるいはmind-based）モデルに基づいて情報処理理解を行っている．技術的な背後にあるのは，人工知能，ニューラルネット，計算機言語学であるとしている．与えられた論題（prompt）に対して，IntelliMetricは生徒の回答から400もの特徴量を抽出し，スコア推定に有効な特徴量を抽出し，スコアモデルに係る重みを推定する[6]．

IntelliMetricによる評価スコアの観点は，文献によって多少の違いがあり，また用いられているワーディングも一貫していないが，おおむね以下の五つである．①目的や主題に対しての結束性や一貫性，②内容の幅や発想の展開，③論旨の展開や文章構成，④文の完全性や多様性，⑤英語のルールへの適合．

上記評価スコア観点への特徴量クラスへの対応は，すべての特徴量クラスがすべての評価スコア観点に影響する多対多の関係である[6]．

またIntelliMetricではその機能の一部に基づいたMY Access!というWebベースの作文評価ツールが提供される．MY Access!もE-rater同様，レベルに応じたフィードバックを返すようになっており，高校最終年（K-12）レベルを標準（proficient）とし，初心者（developing），上級者（advanced）と合わせて三つのレベルがある．加えて，MY Access!では多言語化を図っており，現在，英語，スペイン語，中国語を取り扱うことができる．取り扱うエッセイには様々なジャンルがあり，報知的なもの（informative），事実に基づく物語（narrative），文学，エッセイ（persuasive essay）など，現在200以上の論題が用意されている．

### 2.4 IEA

IEAはコロラド大学のLandauerやFoltzらの研究グループによって開発された．現在，その開発は彼らが立ち上げたベンチャー企業であるKnowledge Analysis Technologies（KAT）社に移管されている．KAT社はIEAのコアとなっているKATエンジンを，全米でも有数のテスト機関であるPearson Education社の傘下のPearson Knowledge Technologies（PKT）社に提供し，PKT社がIEAを販売，提供している．

IEAの特徴は主に論文の内容の評価に重きを置いているところにある[7]．知識獲得と表現についてよく組織化された理論によってシステムが作られている，としている．この方法で置いている仮定は，与えられた文書やテキストの潜在的な意味構造が，単語の共起を通して，コアの意味，すなわちテキストの内容を規定する代表的な行列（特異値行列）によってとらえることができるとするものである．このアプローチは，一般にLSA（Latent Semantic Analysis）と呼ばれている．

IEAは総合スコアに加えて，通常三つの観点からユーザにスコアを提供する．

（1） 内容：LSAから生成された二つの特徴量である文章品質とドメインとの関連性
（2） 文体：首尾一貫性と文法
（3） 技巧：句法，スペル

IEAの総合スコアは，採点者によるスコアとの回帰モデルによって計算され，各観点への重みが計算される．IEAは元々内容の評価を行うことを意図して設計されたが，今では作文スキルを評価することにも利用することができる．

IEAでは，現在，観点ごとの評点に加えて“Tools”という項目があり，コピー（copy：ひょう窃を検出する），スペル（spelling），冗長性（redundancy），文法（grammar）についてコメントを出すようになっている．

### 2.5 BETSY

BETSYはメリーランド大学のRudnerらのグループによって開発されたシステムで，エッセイ評価分類にベイジアンアプローチが採られていることに最大の特徴がある．エッセイの評点は，通常，4段階から6段階で評定されるので，これらの段階へのクラス分けとして考えることができる．分類方法として多変量Bernoulliモデルとmultinomialモデルと呼ばれる二つのベイジアンモデルが用いられている．

BETSYは700ワード以上のエッセイに対しては学習が十分でないほか，適用できる分野も限られているが，研究目的であれば公式サイト[8]からフリーでダウンロードでき，だれでも実行することができる．BETSYはPowerBasicで書かれており，Windows95，98，ME，and NTで動作する．

### 2.6 Jess

Jessは筆者らのグループが開発した，我が国で最初の，そして恐らく現時点で唯一の日本語を処理する小論文の自動採点システムである[9],[10]．

このシステムの最大の特徴は，ほかの既存のシステム

がプロの評価者（rater）を手本にしているのに対し，このシステムは唯一，プロのライター（writer）の書いた大量の文章を手本にしているところにある．このため採点モデルを論題ごとにセットアップする必要がなく，従来大規模な試験にのみしか実用的でなかったこの分野において，初めて小規模な試験での運用を可能とした．

Jess は採点基準については，アメリカの経営大学院（ビジネススクール）への入学試験である GMAT における AWA の採点基準をほぼ踏襲しており，修辞，論理構成，内容の三つの観点から評価を行う．

## 3. 最近の比較研究

### 3.1 AWA テストの公平性

現在，GMAT の作文セッションにある AWA テストでは，IntelliMetric がその採点を行っている．AES スコアと採点者によるスコアに有意な差がないことは，Rudner らによって示されていたが，Guo [16] は，テストを受ける集団の違いによって AES スコア（6 点満点）の得点分布に違いが生じるかを調査した．その結果，アメリカ英語のネイティブとアメリカ英語以外の英語ネイティブとの間に違いがないほか，英語のネイティブと第 2 言語としての英語スピーカーとの間にも，男女間にも，またアメリカ市民とアジア人 / アフリカ系米国人 / ラテン系米国人との間にも差のないことが示された．

### 3.2 IntelliMetric の妥当性

同じ IntelliMetric を使った場合でも，3.1 とは異なった結果が導かれている例もある．Wang [12] は，カレッジ入学レベルの作文能力を測定する標準テストである WritePlacer Plus において，南テキサス州で実施された主にヒスパニック系の学生を対象に IntelliMetric と教員（faculty human rater）との評価スコアを 107 名で比較した．繰返しのある分散分析（ANOVA）の結果，IntelliMetric のスコアの方が教員のスコアに比べ有意に高いという結果を得ている．WritePlacer Plus のニューイングランド州で実施された過去の実験では，人間と IntelliMetric の間に有意な違いのないことが示されており，南テキサス州との違いは，ヒスパニック系の持つ言語上及び文化的背景に起因するものであろうと推測している．

### 3.3 トピックの違いによる影響

James [17] は，課題文の違いによる効果を調査している．WritePlacer Plus テストの三つの異なった課題に対して，IntelliMetric で採点を行った．その結果，個人間で，また男女間で，ネイティブとそれ以外とで，またコンピュータと人間との比較のいずれにおいても有意な差がなかった．唯一の例外は，女性向きのトピックを与えた場合に，平均スコアに有意な差が生じたとしている．

## 4. 自動採点に対する批判

AES の妥当性については今なお，多くの論議が存在する．代表的な批判は，AES はエッセイの表層的な面のみを過大評価しており，内容や独創性に対して無感覚である，というものである．また新しいタイプのズル（cheating）やテスト対策に対して無防備であるとしている [12]．

別の研究者は，AES による評価と人間による評価とが高い一致率を示すのは，エッセイの評価を構成する異なった要素に相関があることの可能性を指摘している．すなわち，よく組織化されたエッセイの書き手は，語彙も豊富で，注意深く修辞にも十分な推敲を重ねている．AES は良い作文の品質と相関する間接的な作文の特徴量をつかんでいるだけであり，AES がよく組織化されているが修辞が不十分であるエッセイや，語彙は豊富だがミススペルが多いようなエッセイを正しく評価できるかどうかは疑問であるとしている [18]．

AES への批判は，機械（コンピュータ）が評定すること自体に対しても向けられている．つまりしょせんコンピュータが採点するのだから，コンピュータが高得点を与えるアルゴリズムに合致するようにエッセイを書くことそれ自体に興味の焦点が置かれるようになるだろうとしている [19]．同じ関心は学生から（人間よりも）コンピュータ向きの作文を書くことを尋ねられる教師の側にも存在し，形式的な表現（formal display）を書くことに関心が置かれるようになる．それゆえ，作文指導は単なるデモンストレーションになり，他者にインパクトを与える言葉使いや，独創的な工夫といったものへの配慮はほとんどなされないであろう，と予想している．

AES の提案者及び擁護者が AES の正当性や効率性を示す妥当性研究の結果を示してさえなお，AES への批判もまたその正統性を有している．コンピュータの採点ロジックが人間のそれと異なっている以上，AES への批判は決して消えることがない．この論争は，実用上問題がなく有用であれば良しとする実用主義（プラグマティズム）とエッセイの評価はこうあるべきだとする理想主義との相克であるといえよう．

### 文　献


（1） L.M. Rudner, “Automated essay scoring: A cross-disciplinary perspective,” Comput. Linguist., M.D. Shermis and J.C. Burstein, vol.30, no.2, pp.245-246, 2004. http://dx.doi.org/10.1162/coli.2004.30.2.245

（2） 石岡恒憲，“小論文およびエッセイの自動評価採点における研究動向，”人工知能誌，vol.23, no.1, pp.17-24, 2008.

（3） T.Z. Keith, “Validity and automated essay scoring systems,” Automated essay scoring: A cross-disciplinary perspective, M. Shermis and J. Burstein, eds., pp.147-167, Hillsdale, NJ：Lawrence Erlbaum Associates, 2003.

（4） E.B. Page, “Project essay grade: PEG,” Automated essay scoring:A cross-disciplinary perspective, M. Shermis and J. Burstein, eds., pp.43-54, Hillsdale, NJ：Lawrence Erlbaum Associates, 2003.
（5） Y. Attali and J. Burstein, Automated essay scoring with e-rater v.2.0 (ETS RR-04-45), Princeton, NJ：Educational Testing Service, 2005.
（6） S. Elliot, “IntelliMetric: From here to validity,” Automated essay scoring: A cross-disciplinary perspective, M. Shermis and J. Burstein, eds., pp.71-86, Hillsdale, NJ：Lawrence Erlbaum Associates, 2003.
（7） 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, Hillsdale, NJ：Lawrence Erlbaum Associates, 2003.
（8） Bayesian Essay Test Scoring sYstem, BETSY, http://edres.org/betsy/
（9） T. Ishioka and M. Kameda, “Automated japanese essay scoring system based on articles written by experts,” Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL 2006), P06-1030, pp.233-240, Association for Computational Linguistics, 2006. Available online：http://www.aclweb.org/anthology/P/P06/P06-1030
(10) 石岡恒憲，亀田雅之，“コンピュータによる小論文の自動採点システム Jess の試作，” 計算機統計学， vol.16, no.1, pp.3-18, 2003.
(11) S. Dikli, “An overview of automated scoring of essays,” The Journal of Technology, Learning and Assessment, vol.5, no.1, 2006.
(12) J. Wang and M.S. Brown, “Automated essay scoring versus human scoring: A comparative study,” The Journal of Technologies, Learning, and Assessment, vol.6, no.2, 2007.
(13) 石岡恒憲，“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， vol.31, no.2, pp.67-87, 2004.
(14) M.H. Breland, J.R. Jones, and L. Jenkins, The college board vocabulary study, College board report, no.94-4, 1994.
(15) M.D. Shermis, C.M. Koch, E. Page, T.Z. Keith, and S. Harrington, “Trait rating for automated essay grading,” Educational and Psychological Measurement, vol.62, no.1, pp.5-18, 2002.
(16) F. Guo, Fairness of Automates Essay Scoring of GMAT AWA, GMAC Research Reports, RR-09-01, Jan. 2009.
(17) C.L. James, “Electronic scoring of essays: Does topic matter?,” Assessing Writing, vol.13, no.2, pp.82-90, 2008.
(18) R. Calfee, “To grade or not to grade,” IEEE Intell. Syst., vol.15, no.5, pp.35-37, 2000.
(19) D. Baron, “The college board’s new essay reverses decades of progress toward literacy,” The Chronicle of Higher Education, vol.51, no.35, May 2005.

（平成 21 年 6 月 11 日受付　平成 21 年 7 月 21 日最終受付）

石岡（いしおか）　恒憲（つねのり）

1985 東京理科大大学院工学研究科経営工学専攻修士課程了．同年(株)リコー（ソフトウエア研究所）入社．1998 文部省大学入試センター研究開発部助教授．組織改変に伴い現在独立行政法人大学入試センター准教授，統計学，信頼性工学，情報数理に関する研究に従事．工博．IEEE Trans. on Reliability 論文審査員，日本信頼性学会（論文審査委員会委員），応用統計学会（編集委員），言語処理学会，ACL 各会員．Marquis Who’s Who in the World, 2008/2009/2010.