小論文/エッセイの自動採点システム

—— 過去，現在，そして未来 ——

石岡 恒憲（大学入試センター）

評定者の影響を取り除くあるいは作文指導の観点から，近年コンピュータによる記述式テストにおける自動採点の研究が精力的に行なわれている。本稿では先ず英文における既存システムの開発の歴史を概説し，自動採点システムに対する従来からの批判を整理する。次に現在の代表的なシステムである e-rater, PEG, IEA, IntelliMetric, BETSYについて紹介し，相互の比較を試みる。また日本語エッセイを処理する，唯一の採点評価システムである Jessについて技術的詳細を紹介する。最後に残された課題や，日本語に固有な問題点について整理しておく。

1 はじめに

小論文/エッセイの自動採点および評価は，現在，教育測定における最もホットな話題の一つとなっている。その理由は，従来，知識工学的なアプローチの多かった自然言語処理に，膨大な言語集体(コーパス; corpus, *pl.* corpora)を利用した確率・統計的なアプローチが成功を収め，その有効性が多くの研究者や技術者に広く認知されてきたことによる。コーパスを用いた成功例のアプリケーションには，典型的なものだけでも，機械翻訳，音声認識，カナ漢変換，情報検索，文書要約などを挙げることができる。これより自然言語を必然的に取り扱うことになる小論文/エッセイテストの評価に，最近の自然言語処理での研究成果を取り込もうとする試みは，きわめて自然な流れであるということがいえる。

本稿では，このようなコーパスに基づく自動採点へのアプローチが開発されつつあった2000年以前の状況を過去，それ以降，現在までを現在と定義する。2節で過去，3節で現在における小論文/エッセイの自動採点システムを概説する。4節では未来において達成すべき課題や方向性を指摘しておく。

2 過去

2.1 先行研究の歴史

自動エッセイ評価の最初の研究は古く，1960年代のPage(1966)に始まるとされる。Pageの開発したシステムはProject Essay Grade, PEG と名付けられたが，その開発の目的は大規模テストにおけるエッセイ評価の教員の負担を減らすことにあった。教員は予め採点（グレード分け）してある学生のエッセイを用いて，テキスト特徴量に係る重回帰における重み係数を計算し，残りのエッセイスコアを予測する。PEGスコアと教員スコアとの相関係数は0.78で，教員同士の相関0.85に近いものであった。

当時，自動的に抽出される特徴量は表面的なもの，たとえば平均ワード長さ，エッセイの長さ（ワード数），コンマの数，前置詞の数，一般的でない(uncommon)ワードの数，といったものに限られていた。Pageはこれらの特徴量をproxiesと呼び，本来測定しようとする作文要素の代用とした。PEGのエッセイ評価予測はある程度の成功を収めたが，初期の段階では作文，および教育コミュニティに受け入れられるに留まっていた。それは作文スキルを直接的に測定していないことに起因すると考えられる。PEGに対しては，間接的な指標を用いているために，トリックを使って良いスコアを人工的に得ることができる，という批判がされた。他にもより本質的な批判

として，作文の重要な質であるところの，たとえば内容(contents)，組織化(organization)，文体(style)などを捉えておらず，このために学生への教育的なフィードバックを与えることができない，ということを指摘することができる。

1980年代の初期には，Writers Workbench (WWB)と呼ばれる作文ツールが開発された。これはスペリングや語法，可読性(readability)について，書き手に有用なヘルプを与えるものである。またWWBは可読性の指標を，文章に含まれるワード，文節，文の数に基づいて提示した。WWBはテキストの表面を粗くなぞっただけのプログラムということができるが，作文品質の自動評価を行うための1ステップと言うことができる。

わが国においてもこの時期，日本語版のWWBと呼ぶべき文書校正支援システムの原型が開発された。代表的なシステムにはNTTで開発されたREVISEを母体とし日経新聞社において利用されているVOICE-TWINや，COMETを母体とし講談社で用いられているSt.WORDS，産経新聞社で実稼働しているFleCSなどがある（池原ほか, 1993）。

日本文の校正は，英語のスペルチェックに対応するものであるが，単に単語の辞書的照合を行うだけでなく，誤りの検出漏れを防ぐために，たとえばVOICE-TWINでは，音声出力機能を組合せ，合成音声との読み上げ校正方式を実装している。新聞，図書の出版分野においては，その語の使用法が厳密に定まっていることもあって，これらの文書校正支援システムは，現在でも校正の現場で実際に利用されている。

1990年代には自然言語処理(Natural Language Processing, NLP)や情報検索(Information Retrieval, IR)の急激な進歩と相まって，これらの技術を作文の品質測定に直接役立てる試みがなされた。アメリカの経営大学院への入学試験であるGraduate Management Admission Test, GMATの一部である作文テストAnalytical Writing Assessment, AWAにおけるエッセイ採点基準には，評価の観点として文法の多様性(syntax variety)，内容(topic content)，組織化(organization of idea)が挙げられている。Jill Burstein (1998)を中心とするETSのチームは，この3つの観点をより直接的に測定するためにNLPやIRの技術を用いて言語上の特徴量を抽出し，利用している。たとえば，彼らはエッセイ中に現れる文や句のタイプを計量するために構文解析ツールを用い，書かれている内容の妥当性については，当時，IRで主流であった単語の共起頻度に基づいたベクトル空間モデルを用いている。

彼らの開発したシステムはe-raterと名づけられたが，そのプロトタイプにおいては各々400以上のエッセイに対して人間の評定者と比較した結果，6点満点中2点以上異なった予測は全体の約10%であった。これは従来の2人の専門家による一致率とほぼ同等であり，これによりe-raterが専門家の一方に代替しうることの妥当性が検証された。

PEGにおいても作文品質をより直接的に測定できるように改良された(Page, 1994)。これによれば，“現在のプログラムは文章の繋がりやすさを測定するなど，より複雑で豊かな変数の採用と，その重み付けがなされている”としている。しかしながら，ほとんどの変数については未だに公開されていない。

またこの時期，Landauerらを中心とするグループが，作文品質をより直接的に測定する方法として，文書検索の最も著名な国際会議であるTREC(Text REtrieval Conference)などで盛んにその有用性が主張されてきた潜在的意味分析(Latent Semantic Analysis)を採り入れ，エッセイ中に含まれる語彙の文字列一致に基づかない，いわゆる意味的な内容の一致を測定するシステムIntelligent Essay Assessor, IEAを開発した (Foltz *et al.*, 1999)。IEAは，現在では改良がなされ，内容，文体，構成（メカニズム）の3つの観点から評価がされるが，15の話題について3,296編のエッセイについ

て評価したところ,2人の専門家による採点の相関が0.86であるのに対し,IEAと専門家による採点の相関はほぼ同じ0.85であった。

2.2 自動採点システムに対する批判

Shermis(2002)によれば，エッセイの自動採点には以下の3つの批判がされてきたという。1つ目は,コンピュータはテキストを正確に理解することができない,というものである(正確性）。適切なキーワードや同義語を用いて出題文に答えたとしても，これが必ずしも包括的に適切な答えになっているとは限らない。例えば以下のような文を考える。

「アメリカ女王は1492隻の船でサンタマリアへ航海した。彼女の夫，コロンブス王は，インディアンの探険家ニーナ・ピンタがイザベラ海岸に巨大な富を持っていることを知っていたが，フェルナンド大陸から香辛料を獲得することを我慢せざるを得なかった。」

勿論，この答案は荒唐無稽であるが，コロンブスの北米大陸発見に関連した多くの適切なキーワードが含まれているために，幾つかのシステムは，これに高スコアを与えるかもしれない。これ程の場合でなくても，望ましい答えに似た文章を書いた場合に，同じ問題が生じることは予想される。このために一部の研究者は，防護策として人間と機械との併用を推奨している。

2つ目の批判は,各出題文に対するモデルをセットアップするために多大な労力を必要とするということである。自動採点システムの多くは重回帰モデルを用いており，採点をおこなうためには事前に多くの変量に係る重みを設定しておく必要がある。このために，実際にこれらのモデルが使われるのは，事前にデータを集めることが妥当となるような大規模テストの利用に限られている。

最後の批判は，書かれている内容の意味的妥当性を評価する内容重視の採点システムは,解答に正解が書かれているかについても十分な評価を行うべきである，というものである（正解性）。しかしながらこの指摘は適切ではない。多くの作文教師は，コミュニケーションの過程では修辞の側面，たとえば自分の意志を伝えるのに論理的な接続表現が用いられているか,あるいは話の筋が通っているか,などといった点を重視するという。実際，一部の出題では正しい答えのない場合がある。つまり作文スキルとして議論の展開の仕方だけに注目しているのである。もし答えが正しいことが重要なら，テストの様式はより効果的な別の形であろうし，その方が結果の妥当性もより上がるであろう。

3 現在

2000年代に入り，ベイズ理論を採り入れたBETSY (Rudner and Liang, 2002)や，ルール発見アルゴリズムに基づくIntelliMetric (Elliot, 2003), また日本語エッセイを処理する唯一のシステムであるJess （石岡・亀田, 2003b） なども新たに登場した。コンピュータによるエッセイの自動採点および評価は，評定の系列的効果（ある小論文の評定が答案の中で何番目に行な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，採点の手間を大幅に低減し，また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。

3.1 英文における自動採点システム

(1) IntelliMetric

このシステムの技術的な最大の特徴は，開発元であるVantage Learning社自身が「先進的な人工知能を有した」と称しているように，知識工学的なアプローチである「ルール発見」を採点に用いていることにある。すなわち，まず最初に予め採点が終っているスコアが出ている模範解答を「学習」し，各採点ポイントのデータを蓄積する。次にシステムはこれらのデータを用いて，人間の採点者の採点ル

ールの判断を推定する。Vantage Learning社が独自に開発したコグニサーチ(CogniSearch), クォンタムリーズニング(Quantum Reasoning), そしてインテリメトリック(IntelliMetric)は, 各採点ポイントにおける解答の特徴を学習し, その知識を採点に活用する。このアプローチは, 全体の採点を行う場合も同様である。

(2) Beyesian Essay Test Scoring sYstem, 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取られていることに最大の特徴がある(Rudner *et al.*, 2002)。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。

一般的には, 分類方法として 2つのベイジアンモデルが用いられる。一つは多変量Bernoulliモデルで, エッセイd_i が分類スコアc_i を受け取る確率を求めるものである。もう一つのモデルはmultinomialモデルで, 与えられたエッセイに対する各スコアの確率をエッセイに含まれる特徴の現れる確率の積で計算するものである。BETSYではエッセイの最初のパラグラフでどのような分野について書かれているかを判定すると言われている。

3.2 Jess

著者らが開発した日本語を処理する唯一の採点システムである。システムとしての最大の特徴は, 他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)の書いた文章を手本にしているところにある。

Jessでは, 模範と考えられる小論文/エッセイとしてある全国誌の新聞における社説とコラム（余録）を学習し, 理想とする文章の書き方についてのメトリクスの分布を予め獲得しておく。これらメトリクスの分布のほとんどは左右非対象の歪んだ分布となるが, この分布を理想とする小論文についての分布とみなす。採点の結果, 得られた統計量がこの理想とする分布において外れ値となった場合に, そのメトリクスにおいて「適当でない」と判断し, 割り当てられた配点を減じ, またその旨をコメントとして出力する。外れ値は四分範囲の1.5倍を越えるデータとする。

Jessは採点基準についてはe-raterの構造, 組織, 内容をほぼそのまま踏襲し, (1) 修辞, (2) 論理構成, (3) 内容の3つの観点から評価する。またそれら3つの観点に係る重み(配点)はユーザが指定できる。ユーザが特に指定しなければ, 配点は5,2,3で合計は10点である。この配点は渡部ほか(1988)の研究成果を踏まえて, 著者らが定めたものである。

(1) 修辞

Jessでは修辞の観点として, 文章の読みやすさ/語彙の多様性/ビッグ・ワード(big word, 長くて難しい語）の割合/受動態の文の割合を評価する。

(2) 論理構成

著者らは毎日新聞の社説に現れる接続関係を示す句を全て抜き出し, これを順接, 逆接各4通り, 計8通りに排他的に分類してある。Jessでは, 採点する小論文の談話(discourse, 議論のかたまり)に対して接続関係を示すラベルを付加し, これらの個数をカウントすることで議論がよく掘り下げられているかを判断する。個数についても, 修辞同様, 毎日新聞の社説で学習し, 模範とする分布において外れ値となった場合に配点を減ずる。

また, これら接続関係の出現パターンが, 社説のそれに比べて特異でないかを判断する。そのために順接と逆接の出現パターンについて, トライグラムモデル（北, 1999）が用いられている。Jessでは事前情報のない方がその生起確率が大きくなるとき, 順接と逆接の出現パターンは特異であると判断し, 議論の接続に割り当てられた配点を減ずる。

(3) 内容

書かれている小論文が問題文に対して適切な内容になっているかについては, TREC (Text REtrieval Conference)などでその有用性

が主張されているLatent Semantic Indexing, LSIが用いられている。このこと自体はIEAと同じであるが，その実装には，石岡・亀田(1999)の特許を利用した高速化のための工夫がしてある。

結果については 800字から1,600字程度の小論文に対しては，ある程度，妥当な結果を示すと考えられる。また入社試験の初期選抜における小論文試験での専門家との比較評価においても，専門家の評価と遜色のないことが確認されている（石岡ほか, 2003a）。

3.3 エッセイ評価モデルの比較

本節の要約として，各エッセイ評価システムの比較を表1にまとめる。第2列目はエッセイの評価基準で，第3列目は各評価システムが主として用いている手法を示す。第4列目の制限は，他の評価システムと比較した場合の弱点に類することが記載してある。第5列目は，人間との評定値との比較についての文献を示す。

評価基準は各システムともそれぞれ開発当初においては大きく異なっていたが，現在ではBETSYを除き，ほぼ同じような観点で評価がなされている。強いて違いを述べれば，e-raterでは評価指標が最も多く，どのようなタイプの論題についての適合できるようチューニングしてある。このためよくトリックが使われている，という批判がされる。また大量の学習データが必要である。

PEGとは逆に「内容」の占める割合が高めであるが，彼らのいう「内容」の中身それ自体に，例えば潜在的意味空間における文書ベクトル間の距離など，文書サイズにきわめて依存する，通常は表層的観点と考えられるような要因が含まれていることは知っておく必要があろう。また開発者が指摘しているように，論理構成や語の出現順を評価しないという問題点が残っている。IntelliMetricはルール発見のアルゴリズムに基づくが故に，論題毎に大量のデータが必要となる。BETSYはまだ開発中であり，利用できる分野が限られている。Jessは毎日新聞の社説やコラムで学習しているために，例えばコンピュータなどの科学技術分野については語の学習が十分でない。このため問題文に応えた内容の文章を書いているにもかかわらず，「内容」の評価が低くなる場合がある。

4 未来

4.1 自動採点システムに望まれる要件

Bereiter(2003)は専門家による採点の不完全さを指摘している。人間の採点には，良い（あるいは悪い）印象が他の全ての評価観点に良い（あるいは悪い）評価を与える，いわゆるハロー効果のあることが知られているからである。事実，Fridmanが1980年代に行った研究

表１：エッセイ評価システムの比較

評価システム	評価基準	手法	制限	人間との比較（文献）
e-rater	構造，組織化，内容	重回帰モデル	"tricked"の批判あり	Powers *et al*, (2000)
PEG	内容，組織化，形式，技巧，独創性	重回帰モデル	内容/概念的正当性を評価しない	Page *et al* (1994)
IEA	内容，文体，技巧	LSI	論理構成/語の出現順を評価しない	Landauer *et al* (2003)
IntelliMetric	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	論題毎に大量のデータが必要	Elliot (2003)
BETSY	表層	ベイズ的接近	分野が制限されている；開発中	Rudner *et al* (2002)
Jess	修辞，論理構成，内容	外れ値検出&LSI	科学技術分野に弱い	石岡・亀田 (2003b)

によれば，人間の評価者は学生のエッセイの中に混入させたプロの手によるエッセイを特別に高く評価することができなかった。このため，彼(Bereiter)は自動採点システムを改良する方法の一つとして，専門家の評定者(rater)を使うのではなく，専門家のライターを使うことを提案している。著者らのグループが開発したJessは，専門家のライターによる文章を評価基準とするという点で世界で最初のシステムということがいえる。

一方，自動採点システムは単にスコアを返すだけでなく，現在では対話的なフィードバックを返すための作文ツールと見なすこともできる。このような立場では，低いスコアを得た学生には，書いたエッセイのどの部分に問題があるかを適切に提示する必要があろう。このために，現在，e-raterの開発チームは以下の問題に取り組んでいるという(Kukich, 2000)。1つは単純な文法エラー（たとえば“I concentrates”，“this conclusions”など）でない，一般に「汚れ(pollution)」と呼ばれる語彙上の文法エラーを，ワード並びのNグラムモデル（たとえば北, 1999など）に基づいて発見しようというものである。「汚れ」の例としては前置詞の誤り/脱落や一般にいわれる悪文などが挙げられる。

2つ目の課題は，言語学で用いられる中心化理論(centering theory)におけるラフ・シフト(rough-shift)を検出しようとする試みである。中心化理論は，代名詞と先行名詞の照応関係を決定する手法であり，トランスレーションの自然な順に「接続(continue)」>「保持(retain)」>「スムース・シフト(smooth-shift)」>「ラフ・シフト(rough-shift)」 の関係がある。100件のAWAエッセイを調査したところ，ラフ・シフトの割合とエッセイスコアとには負の相関があることがわかっており，したがってラフ・シフトを含む文を修正を要するものとして指摘することが正当化される。日本語の場合は，係り受けの深さや埋め込み文の存在などがこれに相当するものと考えられる。

これら2つのことは，まさに今，達成されつつある課題であるが，当然の流れとして将来は内容レベルでの誤りの指摘が求められるであろう。具体例としては実在しない固有名詞（「中僧根元首相」→「中曽根元首相」），矛盾する数値(「第五四半期」)，文意の矛盾(「定率法と低額法」→「定額法」），文意の誤りなどを挙げることができる。これらは構文解析では解決することができず，文脈や一般常識を用いた解析により誤りと断定できるものである。対話的フィードバックの重要性についてはCalfee (2000)にも詳しく述べられているが，自動採点システムを作文支援ツールと考える場合は，従来のWWBの機能それ自体をより精緻化することの方向性が窺えよう。

4.2 日本語評価固有の問題点

(1) 分量の問題

少なくともアメリカの公的試験におけるエッセイ試験では字数制限がないのに対し，わが国の場合は，600字あるいは800字の字数制限が設けられている。その結果，わが国においては作文能力の高い者もそうでない者もほぼ同じ分量を書くことになり，一般に量についての評価は不適である。しかも600字ないし800字という分量は，論理構造を表現するのには少なすぎる分量である。実際，毎日新聞のコラム（余録）の字数は850字であるが，1年365編のコラムの中で約20編は接続表現の全くない記事である。

このような少ない分量だと，文章の論理構造，あるいは展開を採点者は正しく判定することが難しく，したがって採点者個人による違いの影響が相対的に大きくなってしまう。

(2) 順接表現の省略

日本語では，順接表現は意識的に避けられる傾向にある。このため日本語では特に手がかり語に頼らない文章の構成および展開の把握が必要となる。

エッセイをその内容に応じてブロックごとに分解し，その関係を分析する方法は，一般

に談話分析(discourse analysis)と呼ばれ，現在，多くの研究がなされている。重要文抽出あるいは文書要約の基本となるためである。しかしながら，エッセイの自動採点においては，談話の関係に階層構造を採り入れたものはまだない。

(3) 機種依存文字の問題

現在，わが国では小論文の試験は手書きで行われているが，今後キーボード入力が可能となった場合であっても機種依存文字の問題が生じ得る。利用者は必ずしも漢字コードに詳しくはなく，このためJISのコード表に定義されていない機種依存文字（システム外字とも呼ばれる）を意識せずに使用する可能性がある。たとえばWindows（シフトJIS）の①②③はそうである。

小論文では箇条書きを使用する可能性は少なくなく，この危険は十分に想定される。Jessでは機種依存文字は空白に置き換え，システム上，破綻することはないが，ユーザは箇条書きで分かりやすく表現したつもりがシステムはこれを評価しないことになる。

参考文献

Bereiter, C. 2003, Foreword. In Shermis, M. and Burstein, J. eds. *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berry, M. J. A. and Linoff, G. S. 1997, *Data Minig Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., and Harris, M.D. 1998, Automated Scoring Using A Hybrid Feature Identification Technique. In the Proceedings of *the Annual Meeting of the Association of Computational Linguistics*, August, Montreal, Canada.

Calfee, R. 2000, To Grade or Not To Grade, The Debate on Automated Essay Grading, *IEEE Intelligent Systems*, **15**(5), 35-37.

Elliot, S. 2003, IntelliMetric: From Here to Validity, 71–86. In Shermis, M. and Burstein, J. eds. *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P.W., Laham, D. and Landauer, T.K. 1999, Automated Essay Scoring: Applications to Educational Technology. In proceedings of *EdMedia'99*.

池原 悟・小原 永・高木 伸一郎 1993, 文書校正支援システムにおける自然言語処理, 情報処理, **34**(10), 1249-1258.

石岡 恒憲・亀田 雅之 1999, 特許：データベース作成装置および関連文書/関連語検索装置, データベース作成方法および関連文書/関連語検索方法ならびに 記憶媒体, 出願番号:出願平11-188613, 公開番号:公開2001-14341.

石岡 恒憲・鷲坂由紀子・二村英幸 2003a, Jess: 日本語小論文の自動採点システム —入社試験による作文データの評価—, 2003年度 統計関連学会連合大会, 講演報告集, 298-299.

石岡 恒憲・亀田 雅之 2003b, 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, **16**(1), 3-18.

北 研二 1999, 確率的言語モデル, 言語と計算4, 東京大学出版会.

Kukich, K. 2000, Beyond Automated Essay Scoring, The Debate on Automated Essay Grading, *IEEE Intelligent Systems*, **15**(5), 22–27.

Page, E.B. 1966, The imminence of Grading Essays by Computer, *Phi Delta Kappan*, 238–243.

Page, E.B. 1994, New Computer Grading of Student Prose, Using Modern Concepts and Software, *Journal of Experimental Education*, **62**(2), 127–142.

Rudner, L.M. and Liang, L. 2002, Automated essay scoring using Bayes' theorem, *National Council on Measurement in Education*, New Orleans, LA.

Shermis, M.D., Koch, C.M., Page, E., Keith, T.Z., and Harrington, S. 2002, Trait Rating for Automated Essay Grading, *Educational and Psychological Measurement*, **62**, 5–18.

渡部 洋, 平 由実子, 井上 俊哉 1988, 小論文評価データの解析, 東京大学教育学部紀要, 第28巻, 143-164.