行動計量学 第 31 巻第 2 号 (通巻 61 号)
2004 年, 67 ~ 86

記述式テストにおける自動採点システムの最新動向

Latest Trends in Automated Essay-Scoring Systems

石岡 恒憲*
Tsunenori ISHIOKA

* 独立行政法人 大学入試センター 研究開発部
〒 153-8501 東京都目黒区駒場 2-19-23
(Research Division, the National Center for University Entrance Examinations;
2-19-23 Komaba, Meguro-ku, Tokyo 153-8501, Japan.
E-mail: tunenori@rd.dnc.ac.jp)

Abstract

With the aim of removing human errors and providing critical feedback and suggestions for improvement, considerable research has be done on computer-based automated essay-scoring systems. Examples of these include e-rater, PEG, IEA, IntelliMetric, and BETSY. This paper summarizes how these systems work in an attempt to comprehend their features. They are also compared. An automated Japanese essay-scoring system named Jess is introduced, including our analysis of its performance. Lastly, difficulties caused by its treatment of Japanese passages and related problems are discussed.

1 はじめに

1.1 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(池原・安田・島崎・高木, 1987) を母体とし日経新聞社において利用されている VOICE-TWIN(池原・小原・高木, 1993) や，COMET(福島・大竹・大山・首藤, 1986) を母体とし講談社で用いられている St.WORDS (福島・佐々木・赤石沢・竹元, 1992)，産経新聞社で実稼働している FleCS (奥村・脇田・金子, 1992) などがある．

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, 合成音声との対校方式を実装している．校正読みは, たとえば「今秋」を「コンアキ」と読ませるなどの, 同音/類義語の読み分けや, 句読点, 特殊記号を読むなどの点で, 自然読みとは異なっている. また校正のための知識やルールが容易に登録/更新できるよう工夫されている．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 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 NLP で用いている構文解析ツールを用いている．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．

GMAT の AWA テストでは，短いテキスト中で述べられている議論についての分析を問うものと，短いテキスト中で述べられている特定の問題について自ら意見を述べるものとの 2 題が出題されるが，e-rater 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 (Burstein, Kukich, Wolff, Lu, Chodorow, Braden-Harder, & Harris, 1998)．

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 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, Laham, & Landauer, 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．ちなみにこれらの値は総合点についての相関であるが，内容，文体，構成の単体に対しては，IEA と専門家との相関はそれぞれ 0.83, 0.68, 0.66 と少し小さくなる．

2000 年代に入り，ベイズ理論を採り入れた BETSY (Rudner & 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．

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が，これらの研究が盛んになってきている理由としては，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) 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が挙げられる．コーパス (corpus, *pl.* corpora) とは大規模な言語データベースのことであり，データの収集方法あるいは利用目的により様々な形態が存在する．電子化されたテキストを単に集めてきたもの (生コーパス, raw corpus) から品詞や構文情報などの各種の言語情報を付与したもの (タグ付きコーパス, tagged corpus; 解析済みコーパス, analyzed corpus) など

多様である．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よう．確率・統計的観点からの言語のモデル化と応用について解説したこの分野についての我が国最初の系統的成書として，北 (1999) を紹介しておく．N グラムモデル，隠れマルコフモデル，確率文法，最大エントロピーモデルなどについて詳しく知ることができる．自然言語処理全般についての良い教科書としては，長尾 (1996) の 600 ページからなる大著がある．

1.2 自動採点システムに望まれる要件

コンピュータによる自動採点についての初期の研究においては，主としてコンピュータの，人間の評定に対する信頼性に焦点が置かれていた．そのほとんどの研究において，コンピュータは人間とほぼ同等かそれ以上であることを示してきた．その後，エッセイの内容評価についての両者の比較 (Keith, 1998; Page, Lavoie, & Keith, 1996) とシステムの予測妥当性 (Myford & Cline, 2002) に関心が置かれた．その結果，これについても，コンピュータは人間とほぼ同等かそれ以上であることが示された．

Bennet & Bejar (1998) は，コンピュータの性能を評価する唯一の基準として人間の評定に必要以上に頼っていることを批判している．人間の評定は典型的には評価基準表 (rubric) に基づいているのであり，これはユーザが受容可能 (acceptable) と考える信頼性を確保するためのものだからである．作文の質の妥当性にしても，早急な結論を出す前に解決しなければならない以下のような概念的な幾つかの問題が存在すると指摘している．1 つ目の問題は，文脈にあるテーマを評定することが難しく，何が良い作文を構成しているかについての理論がない，ということである．2 つ目は，良い作文のためのルールは破られるためにある，ように思えることである．書き手が文法や使用法の一般的なルールを破るのは，通常の書き方だと書いたものに満足できないという読み手との合意がなされるときに限られるのであるが，その合意は一般には難しい．たとえ，書き手が良い訓練を受けていて，また良い評価基準表があったとしても，評定者間で高い信頼性を得ることは困難であろう．3 つ目は人間の評定者間で高い信頼性が得られたとしても，それはときには異なった理由によるものである，ということである．あるエッセイが良い (あるいは悪い) と判定されるときに，なぜそれを良い (あるいは悪い) 作文だと認識したかを表現できる評定者は決して多くない．

Bereiter(2003) も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．彼らによると $N = 20$ としたときに，ALEK(Assessment of Lexical Knowledge) と名付けられたこのテクニックがエラーと同定したうちの 79%が実際にエラーであった，と報告している．

日本語の場合でいうならば，助詞の誤り/脱落の例として「東京で行く」→「東京へ行く」，「計算機扱う」→「計算機を扱う」などを挙げることができる．また悪文の例として「犯罪を犯す」→「罪を犯す」，「改善する．対処します．」(不統一)，「~ しないと ~ しない．」(二重否定)，「背の高い社長の椅子」(曖昧な修飾関係) などが挙げられよう．これらは主として構文解析処理により誤りと断定することができるものである．

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 (Grosz, Joshi, & Weinstein, 1995) 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．

1.3 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．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．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．

本論文の構成

2 節に英文における代表的な既存システムである Electronic Essay Rater, e-rater (Burstein *et al.*, 1998), Project Essay Grade, PEG (Page, Poggio, & Keith, 1997), Intelligent Essay Assessor, IEA (Foltz *et al*, 1999), IntelliMetric (Elliot, 2003), Bayesian Essay Test Scoring sYstem, BETSY (Rudner *et al.*, 2002) について紹介し, 相互の比較を試みる. 各システムの紹介は，現在に近い状況を反映することは勿論のこと，可能な限りシステムの中身すなわち採点エンジンの仕組みが分かるように努めた．3 節には日本語エッセイ (以下, 小論文と呼ぶ) を処理する, わが国で最初で現時点で唯一の採点評価システム Jess (石岡 他, 2003b) について紹介する. 4 節には, 日本語小論文を評価する上で, 日本語に固有な問題点や解決すべき課題について整理しておく.

2 英文における既存システム

2.1 Electronic Essay Rater, e-rater

アメリカの経営大学院 (いわゆるビジネススクール) の入学試験である Graduate Management Admission Test, GMAT における小論文の採点に用いられており, 知名度という点で, おそらく最も有名な自動採点システムである. アメリカのテスト機関 Educational Testing Service, ETS の Burstein らの研究グループが開発し, 2000 年よりその補助機関である ETS Technologies に拡張開発, および運用が移管されている.

E-rater の GMAT における実際の採点においては, 採点の全てがコンピュータに委ねられてはいないことに注意する必要がある. ひとつの答案は人間とコンピュータが独立に採点し, その結果, 得点差が 6 点満点中 2 点以上あった場合に別の人間の評定者が最終的な得点を決定する. いわば採点の手間を文字どおり半減させる目的で利用している. 得点差が 1 点の場合はモードである 4 点に近い方の値が選ばれる. 専門家と e-rater による採点の一致率 (1 点差以内) は, Burstein & Wolska(2003) によれば 97%である. Burstein *et al.*(1998) では, その一致率は 89%であり, 運用開始から, かなり性能が向上していることがわかる.

E-rater は以下の 3 つの観点から小論文を評定する.

構造 **(Structure):** 文法の多様性, すなわちフレーズや文節, および文の配列が多様な構造で表現されているか.

エッセイ中の文はすべて MSNLP(2004) などの適当な構文解析プログラムによって構文解析され, 構成節, 従属節, 不定節, 関係節が判別される. それら各節の数や, 仮定法における助動詞 (would, could, should, might, may) の出現回数などの情報を得ることができる. これにより, 文ごとに構文構造タイプが決まり, それらの個数, あるいは出

現比率を調べることで, 1 つのエッセイにおける構文多様性の尺度を得ることができる.

組織化 **(Organization):** アイディアが理路整然と表現されていること. 例えば修辞的な表現, あるいは文や節の間の論理的な接続法が使われているか.

エッセイの議論を評価するために, まずエッセイを談話 (discourse) と呼ばれる意味的な議論の構成単位に分割する. この談話は, 形式上の段落とは必ずしも一致しないことに注意する.

談話単位に分割する方法には幾つかの方法があるが, e-rater では (一般的ではあるが最も古典的な) 手がかり語 (cue word) による方法が用いられている (Quirk, Greenbaum, Leech, & Svartvik, 1985). たとえば, "In summary" や "In conclusion" は要約を示す形容詞句であるとか, "perhaps" や "possibly" は議論を掘り下げるときに信念や考えを示す語である, といったものである. "this" や "these" は, 書き手が話題を変えずに関連をもたせるためにしばしば使われる. また新しい話題を始めることを示す不定詞句も同様に見つけることができる.

このようにして自動的に分割した談話単位に対し, 注釈プログラム (APA, Annotation Program) によって「議論の始まりを示している」あるいは「議論を掘り下げている」などのラベルを作成する. また「並列」と「対比」のような修辞的な関係を同定することも行う.

これにより, e-rater は, アイディアが理路整然と表現されているか, あるいは議論がよく掘り下げられているかを判定する.

内容 **(Contents):** トピックに関連した語彙が用いられているか.

良いエッセイは, 与えられたトピックに関連があって, 内容の乏しいエッセイに比べて専門的で正確な語彙が用いられる傾向がある. したがって, 良いエッセイは単語の選択において, 他の別の良いエッセイ (模範エッセイ) と似ていると考えられる. そこで e-rater は, 人間が評点してその結果, 評点 1 から 6 までとなった各カテゴリーに含まれるトレーニング用の複数のサンプル・エッセイに含まれる単語と, 採点するエッセイに含まれる単語とを比較することによって, 字句と内容の評価を行う. ここで使われている文書処理技術は, 「同一文書で何度も出現する単語の重みを大きくする TF(Term Frequency) 法」と, 「どの文書にも現われる (いわゆる一般的な) 単語の重みを小さくする IDF(Inverse Document Frequency) 法」を組み合わせた TF・IDF 法である. これを用いて, 採点エッセイの (評点 1 から 6 までの評点を有する) サンプル・エッセイとの類似度として, コサイン類似度を計算する.

このような方法は, 一般には「ベクトル空間モデル」による方法と呼ばれる. 基本となる考え方は, 互いに似たベクトルをもったもの同士は, 互いに近い関係にある, と判断するものである.

E-rater は最終的なエッセイの評点を, 人間の採点を目的変数とする線形の重回帰モデルにより算出する. 説明変数は, 前述の「構造」, 「組織化」, 「内容」から得られる 57 の説明変数である.

ただ全部の説明変数がいつも使われているわけではなく, 通常の場合は, このうち 8-12 変数が用いられている. エッセイの内容によって, 8-12 変数の組み合わせは異なるのであるが, 現在, (組み合わせの異なる) 75 のモデルが使われている. 75 のモデルのうち, 最も使われる変数は以下の通りである.

1. 単語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

2. 一般的な単語の重みを低くした単語の重みベクトルから得られるコサイン類似度スコア
3. 仮定を表す助動詞の数
4. 仮定を表す助動詞の数の全体の語数に対する割合
5. 議論を深めるための手がかり語の数
6. 議論の始まりに現われる代名詞の数
7. 議論の始まりに現われる補足句 (complement clause) の数
8. 議論の始まりに現われる要約語 (summary words) の数
9. 議論の始まりに現われる詳細語 (detail words) の数
10. 議論を深める修辞句 (rhetorical words) の数

これ以外の変数や, また偏回帰係数については, 公表されていない. 偏回帰係数は, 新しいテストエッセイを評点するたび更新される.

なお, e-rater の技術的な詳細に書かれた論文を `http://www.ets.org/research/erater.html` からダウンロードすることができる. システムそれ自体の説明は Burstein *et al.*(1998) が最も詳しく，最近の研究課題については Kukichi(2000) に詳しい．日本語で書かれた e-rater の紹介記事については石岡 (2001) がある. E-rater の予測妥当性については Powers, Burstein, Chodorow, Fowles, & Kukich (2000) に詳しい．ここには学部における 2 つの作文試験スコア，作文コースにおける GPA スコア，同僚/教官における作文評定スコアなど 9 つの指標と，(専門家および e-rater が採点する) GRE 作文スコアとの相関について報告している．標本サイズ $N = 721 \sim 890$ の調査において，多くの場合，e-rater の方が専門家よりやや劣るように見えるが，5%の危険率で有意となる程に両者に差はない．更にこれを詳しく調べると，スコアの両端で，すなわち GRE 作文スコアが 6 点満点中 5 ないし 6，あるいは 1 ないし 2 を得点する層において，e-rater は専門家より 9 つの指標との相関が小さくなることが示されている．

E-rater は現在, Critique という作文分析ツール (Critique Writing Analysis Tool) とともに Criterion — オンラインエッセイ評価サービス (Online Essay Evaluation Service) の機能の一部となっている. E-rater は全体的なスコアと簡単なコメント (同じ評定のエッセイに対しては全く同じ内容のコメント) を返すだけだが, Critique では文法, 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, 作文指導として利用されることを意図している. Critique の技術的な詳細に書かれた論文, たとえば Burstein(2003) の他, 幾つかの論文は, `http://www.ets.org/critique/research.html` から入手可能である.

また, ETS は Criterion の他に, c-rater と呼ばれる短答式 (short answer), および自由記述 (free answer) についての概念情報 (conceptional information) の解析を行うツールを用意している. これはライティングの質を評価するものではなく, 正解か不正解かを判定するものである.

2.2 Project Essay Grade, PEG

コンピュータによるエッセイ自動採点システムの草分けであり, デューク大学の Page を中心とするグループによって開発された (Page,1966). PEG 開発の背景は, SAT のような大規模テストにおける膨大なエッセイ評価の手間を軽減させることにあったようである. PEG は最初に人間による評価者によって評定された多くのサンプルが集められ, 様々な言語上の

特徴量を測定する. 次に重回帰分析により偏回帰係数を推定し, それぞれ学生のエッセイの評定を予測する.

Page はこの評価モデルを説明するために, 2 つの概念 (説明のための用語) を作っている. 一つは trins であり, これは流暢さ, 語法, 文法, 句読ほか多くの関心のある trinsic 変数である. これらの変数は直接的に測ることができず, そのために代用として proxes を考える. これらは trins の近似 (approximation) で, trins と強い相関がある. コンピュータによって計算される (エッセイについての) 変数は全て proxes である. たとえば, 「流暢さ」という trins は, 「語数」という proxes と強い相関があるといったものである (Page, 1994). trins を計算するために, 多くの proxes 変数を用いた多変量解析が行われるわけである. 多くの proxes 変数についてのデータを集めるために, 様々なソフトウェア製品, たとえば Microsoft WORD でバンドルされている Grammatik-5 や語や文を同定するプログラム, 電子辞書, 品詞タグ付け器, 構文解析ソフトなどが使われている.

なお, 予測変数 (proxes 変数) の大半は Page の著作に明示されていない. ただ変量の数は 1994 年時点で 26 であって, 最も影響の大きい変量は, ワード数の 4 乗根, 文の長さ, 句読を測定したものである. またこの 26 変量による PEG の予測スコアと人間の評価者による予測スコアの相関は 0.8 程度である.

なお, 最終的にユーザに示される評価の観点は, 以下の 5 つである.

1. 内容 Content
2. 組織化 Organization
3. 形式 Style
4. 技巧 Mechanics
5. 独創性 Creativity

PEG の最初のバージョン (Page,1966) では, 内容と形式の 2 つのみであった. 現在のバージョンは 1993 年に改訂されたものをベースにしている .

PEG に関する技術的詳細については, `http://134.68.49.185/pegdemo/ref.asp` より多くの論文をダウンロードできる.

2.3 Intelligent Essay Assessor, IEA

Landauer や Foltz を中心としたコロラド大学の研究グループ `http://lsa.colorado.edu/` が開発し, Knowledge Analysis Technologies (K-A-T) 社が販売するシステムである. 比較的最近まで, IEA のデモ・プログラムはコロラド大学のサイトで実行できていたが, 現在は K-A-T 社のサイトで動作する.

IEA の最大の特徴は, 小論文の修辞上の側面を評価するのではなく, いかに適切な語彙が用いられているかという内容についての評価を行う点にある. このために IEA では, Deerwester, Dumais, Furnas, Landauer, & Harshman (1990) の Latent Semantic Analysis (文書検索の分野では Latent Semantic Indexing, LSI と呼ばれる) を用いている. この方法は, 大量の言語集体であるコーパス (彼らは "bag of words" と呼ぶ) を用いるものである. IEA では , コーパスとして , 百科事典と , 出題文の話題に応じた数冊の専門書が用いられている .

LSI はあらかじめ十分に多くの文書に出現する単語の頻度を表した $t \times d$ の行列 X (t は単語数, d は文書数) を特異値分解

$$X = T_0 S_0 D_0'$$

することから始まる. 一つの文書に複数の単語が同時に現れることを単語の共起 (coccurrence)

と呼び，共起する単語は互いに関係があると考えることができる．行列 X は単語-文書の共起マトリックスともいう．

T_0 および D_0 は, $T_0'T_0 = T_0T_0' = I_t$ および $D_0'D_0 = D_0D_0' = I_d$ を満たす直交行列である．ここで, I_t および I_d はそれぞれ t 次, d 次の単位行列である．また $0 \leq d \leq t$ とする．$'$ は転置を示し, S_0 の対角要素は大きい順とする．

ここで行列 S_0 の対角要素を k 番目までとり, これを新たな行列 S とする．それに応じて, T_0 および D_0 も k 列までを抜き出し, これを新たな行列 T および D とする．このとき,

$$\widehat{X} = TSD' \tag{1}$$

となり, $\widehat{X}$ は X の近似となる．ここで T は $t \times k$ 行列, S は $k \times k$ の正方対角行列, D' は $k \times d$ 行列である．

特異値分解は多変量解析の基本となるもので, $X'X$ の固有値問題は主成分分析に相当し, (1) 式において, TS は主成分得点, D' は主成分の係数を表す．また因子分析においては, T は共通性を 1 としたときの因子得点, SD' はその因子負荷行列に対応している．言語データ (たとえば Bellcore, ENCY で $56,530 \times 25,629$，NEWS で $35,796 \times 19,660$) の場合, Deerwester *et al.*(1990) によれば経験的に k は $50 \sim 100$ 程度にすればよいとしているが, IEA では $100 \sim 300$ の値を用いている．この違いは使用法の違いによるものと考えられる．つまり文書検索では，数万，あるいは数十万オーダ以上の文書の中から類似文書を見付ける必要があり，1 つの文書のベクトルサイズを小さくする必要があるが，一方，エッセイ採点では比較する文書は事前に人間が採点した学習データであるから，高々，数百の文書と比較すればよく，データサイズをそれほど縮約する必要がない．また文書検索では検索のゴミ (不適合文書) が混入することが許されるが，自動採点ではより正確に類似文書を検出することが要求されよう．

さて採点される小論文 e は, 形態素解析によりその小論文が含む t 次元の単語ベクトル x_e で表現することができ, これを用いて, 文書空間 D の行に対応する $1 \times k$ の文書ベクトル

$$d_e = x_e'TS^{-1}$$

を導くことができる．人間が予め採点してある小論文 q についても同様に k 次元ベクトル d_q を得ることができる．これより, 両文書の近似度 $r(d_e, d_q)$ は, 両文書ベクトルがなす角のコサインで与えることができる．

$$r(d_e, d_q) = \frac{(d_e, d_q)}{\|d_e\| \, \|d_q\|} \tag{2}$$

右辺分子の括弧は内積を, また $\|\cdot\|$ はユークリッド・ノルムを示す．(2) 式は相関係数の定義式であり，d_e と d_q が正規分布にしたがうとき, 両者の線形関係を示す妥当な指標となる．文書検索の分野では, (2) 式は一般にコサイン類似度 (cosine similarity) と呼ばれている．

なお, $r(d_e, d_q)$ の代わりに $r(x_e, x_q)$ を用いる方法は TF(term frequency) 法 (Luhn,1957) と呼ばれている．しかし TF 法が単独で用いられることはほとんどなく, 通常は単語が出現する文書数の逆数 (inverse document frequency) に応じて重みを与える Jones(1972) の IDF 法とを組み合わせた TF·IDF 法, もしくはその派生が用いられることが多い (これらの要約については Allan, Carbonell, Doddington, Yamron, & Yang, 1998 など)．他の多くのシステム (e-rater など) では TF·IDF 法が用いられている．

さて IEA は, 採点すべき小論文 e を人間が予め採点してある全ての小論文とのコサイン類似度を計算することで, 最も適切と考えられる評点を付与するものである．IEA では典型的

には類似度が大きい 10 件を取り出し, そこに含まれる評点の近さに応じた重み付けをして評点を与えている．またコサイン類似度だけでなく，ベクトルの大きさについても

$$\min(\|d_e\| - \|d_q\|)$$

となる類似文書 10 件を取り出し，重み付け評定を行っている．要するに潜在的意味空間における空間近さも評定の対象としている．

なお，Landauer らの研究グループが書いた 20 余りの関連論文は `http://lsa.colorado.edu/paper.html` から pdf 形式でダウンロード可能である．Foltz *et al.*(1999) によれば，IEA の評価基準は LSI による意味的評価のみであったが，現在 (Landauer, Laham, & Foltz, 2003) では, 以下の 3 つの観点

- Content: 内容
- Style: 文体
- Mechanics: 技巧

と Overall(総合点) を A-D, F で評価している．

Landauer *et al.*(2003) によれば，GMAT からの標準テストについて標本サイズ $N = 2,263$ について評価したところ，人間間の相関は 0.86 であるのに対し，IEA と人間 (single raters) との相関は 0.85 であったという．また心臓と循環系について書かれたコロラド大学での教室実験では，$N = 1,033$ に対し，人間間の相関は 0.75，IEA と人間との相関は 0.73 であった．しかしながら，複数の人間による調整点と IEA の相関はこれらより大きく，標準テスト ($N = 2,263$) で 0.88，教室実験テスト ($N = 1,033$) で 0.78 であった．これらを合わせる ($N = 3,396$) と，人間による調整点と IEA の相関は 0.85 となる．そのときの 3 つの観点に対する調整点と IEA の相関は，内容 0.83，文体 0.68, 技巧 0.66 であった．

エッセイの総合点に占める 3 つの観点の割合は，内容に対しては 70% ~ 80%，文体に対しては 10% ~ 20%，技巧に対しては 11%であり，内容の占める割合の多いことが報告されている．Chung & O'Neil (1997) には，IEA の妥当性について，さらに多くの実験結果をまとめている．典型的には人間とほぼ同等であるが，人間より良い場合もあれば，悪い場合もある．

K-A-T 社によると IEA の特徴として, 基準スコアの計算のために用意する学習データの数が少なくて済むという. 他のシステムでは事前に人間が採点したデータを 1 つの課題 (prompt) あたり 300 ~ 500 件が必要で, しかも各スコアあたり最低でも 20 ~ 30 件が必要であるのに対し, IEA では 1 つの課題あたり 100 件程度でよいとしている. その一方で, 応答時間は遅い. Foltz *et al.*(1999) によれば, 評価に要する時間は 20 秒とのことである. 使われているマシン，OS，クロック値等については不明であるが，同じ 1999 年に e-rater が Sun Workstation, Ultra-2, Solaris, 137MHz を用いて 2 秒程度で応答したことと比べれば明らかに遅い．LSI の手法は, この当時，既に幾つかの Web 検索エンジンとして使われてきており, この程度の大きさであれば瞬時に応答が返るべきである. 実際, 著者のグループが開発した Jess でもその機能の一部に LSI の手法を用いているが, この部分に要する時間は，Intel Pentium III, 800MHz, RedHat7.2 で，わずか 0.5 秒である. システムとしての完成度が低いのではないかと思われる.

蛇足ながら LSI を分かち書きされていない日本語文書に対して高速に文書検索する仕組みの実装については，著者らが 1999 年に特許を出願し, 2001 年に公開されている (石岡・亀田, 1999b). これを内容の評価に用いた採点システム (Jess) は，著者らが 2002 年に特許出願している (石岡・亀田, 2002)．

なお LSI は, 文書の内容の近さにパターンマッチを直接用いるのではなく, 潜在的な意味空間上での空間距離を測る方法といえる. ただ用いられている単語の出現の順番や論理展開について, この方法では評価していないことは認識すべきであろう.

2.4 IntelliMetric

アメリカ/ペンシルバニア州に本社を置く Vantage Learning 社が開発, 販売するシステムである. 開発の歴史としては 1991 年に設立された Vantage R & D 社において, 1995 年に教育, 心理測定, ホリスティック学の専門家が集まり, 記述試験問題のための採点ツールの開発に着手, 1997 年 7 月にペンシルバニア州の司法試験の採点を実施, 高い信頼性を証明した. 1997 年 11 月に IntelliMetric による論述形式の大学入学試験の採点を実施, 1998 年 2 月に世界で最初のインターネット上で論述式問題に対する自動採点を実施した. それと同時に, Vantage Learning 社を設立, そこでコンピュータ上で瞬時に採点することによって高等教育での学力判定を支援している. 実際のプログラムは `http://www.intellimetric.com/demosite/demo.html` にて閲覧/実施することができる. ちなみに開発までに 11 億円 (10 million dollars) 以上の経費をかけているとのことである.

このシステムの技術的な最大の特徴は, 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.

ルール発見のアルゴリズムには決定木 (decision tree; たとえば Berry & Linoff, 1997 など) が用いられているようである . 決定木を生成するアルゴリズムには CART (Classification and regression trees), CHAID(Chi-squared automatic interaction detection) の他 , エントロピーを利用した C4.5 や C5.0 などのアルゴリズムが知られているが , これらの派生を含めて , どれが用いられているかについては明らかにされていない .

IntelliMetric による評価の観点は, 文献により多少の違いがあるが, 概ね以下の 5 つである.

- Focus & Meaning: 主題に対してどの程度, 一貫性があるか.
- Development & Content: 内容の幅や発想の展開
- Organization: 論旨の展開など文章構成
- Language Use & Style: 文章の複雑さ, 多様性
- Mechanics & Conventions: アメリカ英語に対する適合度

それぞれの観点に対して, 通常 1 ~ 6 点のスコアが与えられ, それをもとに全体の評点が 6 点満点で与えられる. これらの観点は, ペンシルバニア州の教育者によって開発された基準に基づいているが, その基準では 4 段階らしく, そのため, それぞれの観点に対して, 1 ~ 4 点のスコアが与えられる満点が 4 点のバーションもあるようである (Sepos, 2000).

各観点に対するスコアは, 72 種類の素性 (features) により計算される. これらの素性は各観点に排他的に分類されるのではなく, 多少の違いはあれ, 全ての観点に重複して関与する.

Vantage Learning 社が主張する IntelliMetric の長所は, 以下の 2 つである.

1. 2 人の専門家同士の評点の相関よりも, IntelliMetric と各専門家との相関の方が, 相関が高く一致率も高い. 大学入学レベルの 1,202 件のデータを用いて 6 点法で採点した場合に, 2 人の専門家同士の評点の差が 1 点差以内の場合は 95%であるが, IntelliMetric と各専門家間では 99% である.
2. 論文の課題に応じて, 採点を個別に対応させることができる. このことは, 採点者に課題別の採点訓練を行うのと理論的には同じことである. IntelliMetric による採点と人間が行う採点との間に, 既存システムが示す以上の高い相関性と高い一致率が認められるのは, そこに理由がある.

Vantage Learning 社の主張が事実だとすれば, 1. は, IntelliMetric は人間より保守的な採点をしていると考えることで説明がつく. 人間は採点に際してある意味での思い入れ, たとえば「この論文は着眼がよい/切り口が斬新である」, あるいは「自分と共通体験がある」などの理由で, 他の採点者と比較し外れ値となるようなスコアを与えてしまいがちである. IntelliMetric はルールベースのシステムであるから, このような思い入れに類するようなデータは, 事例として相対的に少ないためにルールとして成立せず, したがって平均的な採点をするのだと考えられる. つまり真の Writing Ability を t, メソッド i による評価値を y_i は, 誤差項を e_i としたとき

$$y_i = t + e_i$$

で示されるわけだが, IntelliMetric ではこの誤差項 e_i のバラツキが人間によるバラツキに比べ小さいのだと考えられる. また, 既存システムが示す以上の高い相関性と高い一致率は, 2. で述べたように課題別に採点のルールを策定することによることは確かであろう.

しかしこの利点は諸刃の剣であって, 良い採点を行うために, 事前に良質の採点付き学習データを多数用意しておく必要があることは指摘しておいてよい. Elliot(2003) によれば,

- 学習データは 300 以上 (モデル決定には 50)
- 6 点法でスコア 1 及び, スコア 6 のデータが 20 以上
- 2 人以上の採点者

が必要であるという. 課題の数が限られていて, 多くの採点を行う場合には, 採点付き学習データを多数用意することがコスト的に割に合うが, 多種類少数の採点には割に合わないであろう.

また, 極めて注意深く書かれたいわゆる良いエッセイを正当に評価しない. たとえば 2001 年のポスト・ガセット誌には, その新聞の教育担当記者 (Eleanor Chute) が自分の書いたエッセイを IntelliMetric で評価したところ, 6 点満点中 4 点しかとることができず, 推敲を重ねてもそれ以上の得点を得ることができなかったことが報告されている (Chute, 2001). 実際, 主任責任者 (Chief Operating Officer) の Scott Elliott によれば, 3% から 7% の論文はルール適用が難しく, 類別することが通常, 困難 (too unusual to grade) であると言っている. また, 同じ評点を得た場合に同じコメントを返すようになっていることも不備な点として指摘してよい.

IntelliMetric の妥当性については Elliot(1999) に詳しい. これには 7 年生 (International, Age7), 11 年生, 14 年生における外部の作文テスト (External Measures of Writing) と IntelliMetric との相関について示している. 外部の作文テストには, 多肢選択テストと教師の評定との 2 つがあるが, IntelliMetric と多肢選択テストとの相関は平均で 0.60 であり, 教師の評定との相関は平均で 0.64 である. IntelliMetric の代わりに人間が採点した場合は, 相

関はそれぞれ 0.58 と 0.60 となる．Elliot によれば，学年によってその平均値に違いがある (学年が大きい程，相関の値が大きい) ことが報告されている．

2.5 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. ベイズ流のエッセイ採点を説明するために, 被験者の応答が, 適切 (Appropriate), 部分的に適切 (Partial), 不適切 (Inappropriate) の 3 つのいずれかに分類することを考える. 予め, エッセイの特徴量について以下の 3 つの確率を決定しておく. それらの確率は, 被験者の応答が適切/部分的に適切/不適切だと採点者が判断する場合にそのエッセイの中に着目する特徴量が含まれている確率である. それらを $P_i(u_i = 1|A),\ P_i(u_i = 1|R),\ P_i(u_i = 1|I)$ と表す. 添字 i は, 特徴量の識別子であり, u_i はエッセイがその特徴量を含んでいるか否かを示す. A, R, I は, それぞれ適切/部分的/不適切を示す. これらは専門家によって採点されたエッセイの集合から条件付き確率として与えられる. 例として, この条件付き確率を

$$\text{適切} : P_i(u_i = 1|A) = 0.7,\ \text{部分的} : P_i(u_i = 1|R) = 0.6,\ \text{不適切} : P_i(u_i = 1|I) = 0.1$$

とする. ここでの目的は, 被験者のエッセイがその特徴量に基づいて, 適切/ 部分的/ 不適切のいずれが最も尤もらしいか判定することである. 被験者の Ability について先験情報が与えられていないとき, それぞれの事前確率は等しい, すなわち $P(A) = P(R) = P(I) = 0.33$ を仮定する. それぞれの特徴量を検査し, $P(A), P(R), P(I)$ をその特徴量が被験者のエッセイに含まれているかに基づいて更新する. ベイズの定理から，被験者のエッセイがある特徴量を持っているときにそのエッセイが適切であるとする事後確率は

$$P(A|u_i = 1) = P(u_i = 1|A) * P(A)/P(u_i = 1)$$

である. このとき $P(A|u_i = 1) = 0.7 \times 0.33/P(u_i = 1) = 0.233/P(u_i = 1)$ となる. 同様に $P(R|u_i = 1) = 0.6 \times 0.33/P(u_i = 1) = 0.200/P(u_i = 1),\ P(R|u_i = 1) = 0.1 \times 0.33/P(u_i = 1) = 0.033/P(u_i = 1)$ となる. ここで各事後確率の分母の $P(u_i = 1)$ は, 分子の同時確率の総和であるから $P(A|u_i = 1) = 0.233/(0.233 + 0.200 + 0.033) = 0.5,\ P(R|u_i = 1) = 0.200/(0.233 + 0.200 + 0.033) = 0.429,\ P(I|u_i = 1) = 0.033/(0.233 + 0.200 + 0.033) = 0.071$ となる. この時点で, このエッセイが不適切 (I) であることは起こりそうにないことがわかる.

次にこれら事後確率を新しい事前確率として, 次の特徴量に対して $P(A),\ P(R),\ P(I)$ の推定値を再び更新する. このプロセスを全ての特徴量に対して繰り返す.

より一般的には，分類方法として 2 つのベイジアンモデルが用いられる (McCallum & Nigam, 1998)．一つは多変量 Bernoulli モデルで, エッセイ d_i が分類スコア c_j を受け取る確率が

$$P(d_i|c_j) = \prod_{t=1}^{V} [B_{it} P(w_t|c_j) + (1 - B_{it})(1 - P(w_t|c_j))]$$

で与えられる. ここで, V は特徴量の数, $B_{it} \in (0, 1)$ は特徴量 t がエッセイ i に含まれているか否かを示す識別子である. $P(w_t|c_j)$ は，特徴量 w_t がスコア c_j の文書中に少なくとも 1 回

現れる確率であり，予め採点された学習データから以下により計算することができる．

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{D_j} B_{it}}{J + D_j}$$

ここで D_j はスコア c_j のグループに含まれる学習データの数であり，J はスコアグループの数である．

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が以下のようにエッセイに含まれる特徴の現れる確率の積で計算されるものである.

$$P(d_i|c_j) = \prod_{t=1}^{V} \frac{P(w_t|c_j)^{N_{it}}}{N_{it}!}$$

ここで N_{it} は, 特徴量 w_t がエッセイ i に何回現れているかを示している. このモデルでは，$P(w_t|c_j)$ は，特徴量 w_t がスコア c_j であるエッセイに含まれる確率であり，学習データから以下のように計算される．

$$P(w_t|c_j) = \frac{1 + \sum_{i=1}^{D_j} N_{it}}{V + \sum_{i=1}^{D.} N_{it}}$$

ここで $D.$ はエッセイの総計である．音声認識の分野では，このモデルは “ユニグラム (unigram) 言語モデル” と呼ばれており，テキスト分類に Mitchell(1997) などが適用したものである．

BETSY の妥当性については，人間が予め 2 点法で採点した 462 の学習データを用いて，別の 80 編のエッセイ (各スコアに対して 40 編ずつ) を，特定の単語，フレーズ，論理展開の有無などの特徴量に基づき分類したところ，80 編中 64 編 (80%) が正しく判定された．なお, 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ようである.

2.6 エッセイ評価モデルの比較

全てのエッセイ評価システムは，複数の評価観点に基づき総合 (holistic) のスコアを与えるものであるが，そのスコアの付与の仕方は 2 つに分類することができる．一つは，採点スコアが最終的にはどこかに丸められるにせよ，本質的に連続量で与えられるものである．総合スコアが重回帰モデルによって計算される e-rater, PEG はその典型である．採点エッセイに最も近い 10 編を取り出し，その近さに応じた重み付けをして総合スコアを計算する IEA もこれに属する．

一方，別の分類方法は，採点スコアが離散値であることを前提とし，はじめから採点エッセイをスコア・カテゴリーに分類することを目的とするものである．IntelliMetric と BETSY が, これに属する．IntelliMetric は採点エッセイがどのカテゴリーに属するかを，予め学習により得られたルールに基づき判定するものである．すなわち BETSY はその採点エッセイが有する特徴量を基に，最も属することが自然な (尤もらしい) カテゴリーを見付けるものだということがいえる．この方法では，必ずしも採点カテゴリー間に自然な順序が保持されないことに注意する必要がある．たとえばスコアカテゴリーとして A,B,C,D,E,F の 6 段階

に分類することを考える．人間の判断では A が最も良く，F が最も悪い．ところがシステムの判断では，仮にいま採点エッセイがカテゴリー A に最も近いと判定したときに，次に近いカテゴリーは B であるということを必ずしも保証しない．C かもしれないし，それ以外かもしれない．もし，採点エッセイの空間的位置がカテゴリー A の中心と C の中心のほぼ中間にあるときには，A あるいは C に属すると判定するのではなく，カテゴリー間の自然な順序を考慮して B に属すると判定することもアルゴリズムとしては考えられよう．

各システム間の信頼性，妥当性については，相互の比較は決して容易ではない．Rudner & Gagne (2001) には，PEG, e-rater, IEA についてのそれぞれの妥当性について比較サーベイを行っているが，これによると IEA と e-rater は内容の評価において優れており，PEG は作文品質 (writing quality) を判定するのに優れているとしている．人間とシステムとの相関は典型的には 0.75 ~ 0.85 程度であると考えられるが，この数値は同一システムについての同じ著者による文献 (たとえば Page, 1996 と Page, 1997) でも異なる．Chung *et al.*(1997) では人間とシステム (IEA) との相関係数の比較による優劣さえも事例により逆転してしまう．これは学習データそのものが実験により異なるためであり，性能に依存する学習データの数も同じではない．システムそのものも改良により変化しているであろう．

Shermis(2002) によれば，総合点 (holistic score) で評価した方が，各観点ごとに評価するよりも人間との相関は高くなり，また人間との比較においても 1 人の人間 (single raters) との相関よりも複数の人間による評価 (たとえば平均値) との相関の方が高くなるという．また自動採点システムは一般に 500 ~ 1,000 ワードの短いエッセイにおいては，トピックの広い範囲についてのより記述的 (descriptive) なエッセイに向いているという．

唯一，Elliot(2003) が IntelliMetric と他の 2 つの有名なシステム (名前を伏せているが e-rater と PEG だと考えられる) との同一論題についての比較が数値を明示せずに文章で述べている．これによると IntelliMetric は，他の 2 つのシステムに比べ 6 点採点法で，人間の採点と全く同じ評点を与える割合は有意に大きいが，1 点差以内に収まる割合は小さいとしている．

本節の要約として, Wresch(1993) にならい, 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 については，従来より内容/概念的正当性を評価しないという批判がある．IEA は 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．

表 1: エッセイ評価システムの比較 (Wresch,1993 に準拠)

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.*(1996, 1997)
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)

3 日本語小論文の自動評価システム Jess

3.1 必要とする要素技術

欧米では専門家によって採点された膨大な数の小論文の蓄積があり, これを用いれば, たとえば, 専門家の得点とコンピュータによる得点とを線形回帰させる, あるいは統計的なクラスタリングの手法を用いて採点を行う, あるいは専門家の採点についてのルールを学習することでそのルールを適用した評点を与える, などの方法が可能となる.

一方, わが国の場合は, オーソライズされた利用可能な得点の蓄積がおそらくない. たとえば, 大学入試で用いられた採点データは, 入試選抜の目的でのみしか利用できない.

しかしながら模範と考えられる小論文, およびエッセイを電子媒体で入手することは現在ではさほど難しくない. たとえば, 「毎日新聞」の 2002 年までの全記事を, また日経出版販売より「日本経済新聞」の 2001 年までの全記事を入手することができる. これらの記事にはタグが付いていて, 社説あるいはコラム (「余録」) 等, 意図する記事を選択的に入手することができる. さらに著作権の切れた文学作品は青空文庫 `http://www.aozora.gr.jp/` から利用することもできる.

一方, 自然言語における日本語解析の最も基本となる形態素解析については, 京都大学 言語メディア研究室で開発された JUMAN や奈良先端科学技術大学院大学 松本研究室の茶筌 (ちゃせん, `http://chasen.aist-nara.ac.jp/`; 今回, 著者らが使用), 富士通研究所の Breakfast, NTT 基礎研究所の「すもも」などがフリーで利用でき, 構文解析についても京都大学の KNP や奈良先端科学技術大学院大学の SAX, BUP , 東京工業大学 田中・徳永研究室の MSLR パーザなどが同様にフリーで利用できる.

このように, 模範となるエッセイやコラムに加えて, それをコンピュータ処理すべきツールもいまや整いつつある. また小論文の採点においては内容の適切さ, すなわち書かれた内容が質問文に十分に応えた内容であるかの評価が不可欠となるが, これについてもインターネット・ウェブにおけるサーチ・エンジン等で用いられているパターン・マッチ (文字列一致) に拠らない意味的検索技術が利用できるようになった. その技術的な実装方法については, 石岡・亀田 (1999a,b) などに詳しく, 従って模範となるエッセイやコラムに如何に外れているかという判断のアプローチを取ることで, 日本語で書かれた小論文の自動採点システム

を開発できる, と著者らは考えた.

われわれは日本語で書かれた小論文の自動採点システムを Jess (ジェス) と名付けたが, 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ものとした. ユーザが特に指定しなければ, 配点は 5,2,3 とし, 合計を 10 点とした. この配点は渡部・平・井上 (1988) の研究成果を踏まえて, 著者らが定めたものである. ちなみに既存の多くのシステムの満点は 6 点である. これは評点の標準誤差より定まったものであるらしい. すなわち，もし個人が何度もこの試験を受けたとしたら，そのうちの 68%が与えられたスコアを得るようにスコア区間が定められている．0 点は，分量がきわめて少ないなどの特別の場合のために用意されている．

本節では, 以下に Jess は採点基準の詳細について説明する. 3.2 節に修辞, 3.3 節に論理構成, 3.4 節に内容について述べる. 3.5 節は実施例を取り上げ, そのときの実行時間について記す. 3.6 節はまとめである.

3.2 修辞

Jess では修辞を示すメトリクス (計量値/計数値) として前川 (1995), 長尾 (1996) に従い, (1) 文章の読みやすさ, (2) 語彙の多様性, (3) ビッグ・ワード (big word, 長くて難しい語) の割合, (4) 受動態の文の割合, を用いた. これらをさらに次項以下で述べるメトリクスに細分化し, それらの統計量の分布を, 毎日新聞の CD-ROM に納められている社説, あるいはコラムについて得た.

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.

(1) 文章の読みやすさ

文章の読みやすさを示す指標として以下を取り上げた.

1. 文の長さの中央値, 最大値
2. 句の長さの中央値, 最大値
3. 句中における文節数の中央値, 最大値
4. 漢字/カナの割合
5. 連体修飾の用言 (埋め込み文) の数
6. 連用形や接続助詞の句の並びの最大値

(2) 語彙の多様性

ユール (Yule,1944) は文体の解析に様々な統計量を使ったが, 最も有名なのが K 特性値とよばれる語彙の集中度を示す指標である.

K 特性値は, 文書中に n 回現われた語の個数を $f[n]$ で表すとき, 次式で与えられる:

$$K = \frac{T - S}{S^2} \times 10,000$$

ただし,

$$S = \sum_{n=1}^{n\text{ の最大}} (n \times f[n]), \quad T = \sum_{n=1}^{n\text{ の最大}} (n^2 \times f[n])$$

とする. S は語の出現回数の 1 次モーメントである. T は語の出現回数の 2 次モーメントであるが, n を 2 乗しているため, 出現回数の合計が同じであっても, 出現回数が偏っている程, T の値は大きくなる. 従って T の値そのものを語彙の集中度を示す指標としてもよいのだが, 全ての語が 1 回しか現われないときに K の値が 0 になるよう S を減じ, さらに長さに対して正規化する (文章が長くなると T も S も大きくなる) ために S^2 で割っている. これを 10,000 倍するのは人間にとって見やすくするためである.

K 特性値は, 語彙が集中しているほど大きくなり, 語彙が多様なほど小さくなる. 毎日新聞の社説では, K の値の中央値は 87.3 であり, コラムでは 101.3 であった. なお, 語彙の集中度を示す特性値には, ユールの K 以外にも多くが提案されている. 例えば Tweedie & Baayen(1998) などを参照されたい.

(3) ビッグ・ワードの割合

いわゆるビッグ・ワードをどの程度, 使っているかが, 読み手に与える印象は決して小さくないと思われる. さてビッグ・ワードを調べるに当たって, 日本語の場合は文節の長さだけではその判断を誤ってしまう危険がある. 英語の場合, ビッグ・ワードは大抵の場合長い語であるが, 日本語では漢字をカナで表せば長さは増え, 表記上は短い語もビッグ・ワードになる可能性がある. 従ってカナに変換したときの文字数, いわゆるヨミでもってビッグ・ワードを判断する必要がある.

毎日新聞の社説では, 用いられている名詞をカナで表記した場合の文字数を調べてみると, その中央値は 4 で, 第 3 四分位 (上位 25%) で 5 であった. 従ってヨミで 6 文字以上の名詞をとりあえずビッグ・ワードと仮定し, 改めてビッグ・ワードが文書中の名詞に含まれる割合を測定した. ヨミの字数は整数値であるために, この割合は必ずしも 25%にはならないが, それに近い値を平均とする分布が得られる.

(4) 受動態の文の割合

一般に文章はできるだけ能動態で書くべきで, 受動態の多い文章は悪文とされている. 従って, これも修辞に関する評価指標となる.

3.3 論理構成

議論の流れをつかむことは, さまざまな主張のつながり具合を把握することに他ならない. このため, 書き手はその理解を助けるために, 議論の接続を示す接続表現をしばしば用いることになる. そこで我々も論文中に現われる接続表現を検出することで, 文章の論理構造を把握することを試みた.

さて接続関係は, 大別して, 「順接」と「逆接」に区分できる. ここで「順接」という語はやや広い意味で用いており, 議論の流れが変わらない接続構造一般を指している. これに対して, 議論の流れを変えるような接続関係を「逆接」と呼ぶ. 「順接」と「逆接」の論理構造を主題的に分類すると以下のようになる. なお, この分類は野矢 (1997) による.

順接の接続構造には以下がある.

付加: 主張を加える接続関係である. 典型的には「そして」で表される. 他にも「しかも」や「むしろ」などがある. 省略されることも少なくない.

解説: 典型的には「すなわち」,「つまり」,「言い換えれば」,「要約すれば」といった接続表現で表される接続関係である. さらに細かく分類すると, 要約 (それまで述べていたことをまとめて述べる), 敷衍 (要約の逆で, まず大づかみなことを示しておき, それからその内容を詳述する), 換言 (内容的には同じことの繰り返しだが, 理解を助けるために, あるいはより印象的な表現を与えるために言い換えを行なう) がある.

論証: 理由と帰結の関係を示す. 理由を示す典型的な接続表現には,「なぜなら」,「その理由は」などがあり, 帰結を示すものとしては,「それゆえ」,「従って」,「だから」,「つまり」などがある. 接続助詞の「ので」や「からも」も理由–帰結を示す.

例示: 典型的には「例えば」で表される接続関係であり, 具体例による解説, ないし論証としての構造をもつ.

また逆接の接続構造には以下がある.

転換: ある主張 A に対して対立する主張 B が続けられるとき, B の方にいいたいことがくる接続関係をいう. 一般に「A だが B」,「A, しかし B」という表現をとる.

制限: 上記において, A の方にいいたいことがくる接続関係をいう. いわゆる「ただし書き」であり, 典型的には「ただし」や「もっとも」などがある.

譲歩: 転換の一種とみることもできるが, 譲歩の場合は対話的構造が現われる. 典型的には「たしかに」,「もちろん」などである.

対比: 典型的には「一方」,「他方」,「それに対して」といった接続表現で表される接続関係である.

我々は, 毎日新聞の社説に現われる接続関係を示す句を全て抜き出し, これを前述の順接, 逆接各 4 通り, 計 8 通りに排他的に分類した. 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した. 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こととした.

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した. そのために著者らは, 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) を考えた. 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考え, 議論の接続に割り当てられた配点を減ずることとした.

3.4 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 を用いる. このこと自体は IEA と同じであるが, その実装に以下のような高速化のための工夫がしてある.

一般に文書に現れる単語の出現頻度行列 X は一般に巨大な疎行列 (sparse matrix) となる. また, 我々の目的においては, 行列 X の特異値を全て計算する必要はなく, その特異値の大きいものだけ高々数百個を計算すればよい. そのことに着目すれば, 巨大な疎行列に対する特異値分解のためのソフトウェア・パッケージである Berry(1992) の SVDPACK を使うことが有効である. ここでは 8 通りのアルゴリズムが利用できるが, これらの日本語文書-単語の出現頻度行列に適用した場合の比較・評価を石岡 他 (1999a) らは既に行っていて, 最適な方法を見つけ出している. さらにこのパッケージを用いるためには行列 X のデータ格納形式として Duff, Grimes, & Lewis (1989) にある Harwell-Boeing sparse matrix format を用

いている. 疎行列に対してデータを効率よく格納できるので, ディスクの節約, ならびにデータ読み込み時間の大幅な低減をはかることができる. しかもデータはテキスト形式ではなく, バイナリ形式とし, 更なる速度向上を計っている.

参考までながら, ここでいう単語とは IPA 品詞体形 (THiMC097) でいう「名詞」のうち, 一般 (普通名詞), 固有名詞-一般 (一般的な固有名詞), 固有名詞-組織 (組織を表す名称, 「通産省」など), 固有名詞-地域-一般 (国名以外の地名), 固有名詞-地域-国 (国名), サ変接続 (格要素をとり, 後ろに「する」, 「できる」などが後接できるもの, 「悪化」, 「下取り」など), 形容動詞語幹 (いわゆる形容動詞の語幹で, 「な」の前に現われるもの, 「健康」, 「安易」など) とした. これ以外の名詞, 例えば代名詞, 副詞可能, ナイ形容詞語幹, 数, 非自立, 特殊-助動詞語幹, 接尾, 接続詞的, 動詞非自立的は含まない.

3.5 実施例

e-rater における実施例は `http://www.etstechnologies.com/html/eraterdemo.html` で見ることができ, ここで 7 通りの回答パターン (7 つの小論文) に対する評価を見ることができる. 得点の内訳は, 6 点満点中, 6 点, 5 点, 4 点, 2 点のものが各 1 つで, 3 点のものが 3 つである. そこで上記の Web ページに示している小論文を著者が和訳し, それらを Jess で採点した. 採点結果を表 2 に示す. 2 列目が e-rater の得点, 3 列目が Jess の得点であり, 4 列目が各小論文の字数である.

表 2: 採点結果の比較

小論文	e-rater	Jess	字数	CPU(秒)
A	4	6.9(4.1)	687	1.00
B	3	5.1(3.0)	431	1.01
C	6	8.3(5.0)	1,884	1.35
D	2	3.1(1.9)	297	0.94
E	3	7.9(4.7)	726	0.99
F	5	8.4(5.0)	1,478	1.14
G	3	6.0(3.6)	504	0.95

Jess は標準では修辞 5 点, 論理構成 2 点, 内容 3 点の計 10 点で採点するが, e-rater の得点と比較するために, 6 点換算の得点を括弧書きで示した. これを見るに e-rater が良い得点を与える小論文には Jess も良い得点を与えており, 得点もかなり一致していることがわかる. だが e-rater は (そしておそらく人間は) 同じような形式で書かれた小論文であるならば, 分量の多いものにより多くの点を与える傾向があり, そこに減点法で採点する Jess との違いが現われているように思われる. 例えば小論文 C においては, e-rater は満点の 6 点を与えうるが, Jess では減点法なので, 論文の有する多少の悪い点を分量で補うということをせずに, 6 点満点換算で 5 点程度としてしまうと考えられる.

ちなみに人間が評価すると, 7 つの論文の評点の平均をどこに置くかによって個人差 (評点者差) が生じ, e-rater/Jess での判定とは必ずしも合致しない場合がある. しかし, 論文の順位 (論文間の優劣, あるいは同等か) の判定は, e-rater/Jess での判定とほぼ同等であることが確認されている.

表 2 の第 5 列目に Jess の処理時間 (CPU 時間) を示した. 使用マシンは Plat'Home Standard System 801S; Intel Pentium III 800MHz; RedHat7.2 である. Jess は C シェルスクリプト, jgawk, jsed, C で書かれており, 全部で 1 万行弱のプログラムである. 動作させるために, 形態

素解析システム茶筌の他に, 漢字/カナ変換プログラム kakasi(http://kakasi.namazu.org/) が必要である. 現在は UNIX 上でのみ動作する. Web 上では http://zaza.rd.dnc.ac.jp/jess/ で実行可能である. 現在，早稲田大学アジア太平洋教育センターの井上達紀先生らが Windows への移植作業を進めており，2004 年秋に Windows 版の提供を予定している．提供の環境が整い次第，上記アドレスにて連絡する．

3.6 課題

Jess は大学入試における小論文の採点システムに用いることを念頭において作成された. このため, 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・鷺坂・二村, 2003a). さらに他の全てのシステムが, 与えられた課題について事前にいくつかの小論文を人間が採点しておく必要があるのに対し, Jess はその必要が全くないことはその優位性として主張してよいであろう.

しかしながら, 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く, 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」の評価が低い事例のあることがわかっている. 従って, 内容の分析においては, 書かれている記事に応じて, 用いるべき単語-文書の共起マトリックスを自動選択できるような仕組みが必要となるかもしれない.

4 日本語小論文評価における問題点

日本語が分かち書きをしないいわゆる膠着言語であることが, 英米語に対するシステムの日本語への転用を阻害している, という考え方はナンセンスである. いまや高性能の形態素解析や構文解析ツールが整備され, これらを容易に使用することができるからである. 日本語も英米語も，どのようなエッセイをより好ましいと考えるかについての基本的な認識に違いはない．しかしながら試験文化やエッセイについての価値観は必ずしも同じではない．このため以下のような問題点が存在する．

4.1 分量の問題

英米語と比較した場合, 最も大きな問題点は, 日本語の字数制限である, と著者は考えている. 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. たとえば, アメリカの経営大学院の入学試験 GMAT における AWA では, 以下の 2 つのタイプの論題が出され, 各 30 分で解答する.

1. 論点 (issue) の分析 (analysis of an issue): 論点に対して自分の意見を述べる問題である. 効果的で説得力のあるエッセイであることが求められる.
2. 議論 (argument) の分析 (analysis of an argument): 議論に対する批判と, どうすれば議論が良くなるかを述べる問題である. 論理的批判能力と分析力が求められる.

平均的な受験者は 30 分で 300 ~ 400 ワード程度を書くようであるが, 中には 800 ワード近くまで書く者も決して少なくない. 通常, 翻訳業界では和文の 400 字を英文の 200 ワードに換算するから, 800 ワードは 1,600 字となる. これには改行による空白分は含まれていないから, 正味の 1,600 字は 400 字詰めの用紙で確実に 4 枚を越える. 起承転結の論理構造なし

にこの程度の量を書くことは実際上できないし，論理展開もそれなりになされると考えるのが自然であろう．そうすれば，誰が採点しても 6 点満点中 (モードが 4 点であるから)5 点ないし 6 点を得るであろう．

一方わが国においては，600 字ないし 800 字を書くのに十分な時間が与えられるから，高い作文能力 (writing ability) を有している人もそうでない人も，ほとんど同じ程度の分量を書くことになる．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．著者らは言語学の専門家を含めて，これら約 20 編の記事を全て調べてみたが，決して不自然な/表現的に悪い文章ではなかった．850 字程度の分量だと起承 (転) 結なしに一気にかけてしまうのだと考えられる．

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ことが予想される．実際，公開できないものも含めて我々の調査によると，事前に採点基準を定めた専門家による評価であっても，専門家同士の評点のピアソン相関係数はわずか 0.5 を少し越える程度である (石岡 他, 2003a など)．これは英米語の同様の調査 (Powers *et al.*, 2000) に比べ，明らかに小さい．

ちなみに毎日新聞の社説は 1,600 字であり，これだと接続表現が必ず出現することを確認している．現状より，人間の誤差の少ない評価が期待できることは明らかである．

4.2 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．実際，この省略が独特のリズムとなり，名文ともなる．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．階層的談話関係を示した Marcu (2000) は注目に値するだろう．

4.3 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．我々の開発したシステムは内部的には拡張 UNIX コード (EUC) を用いているが，入力文字コードとして EUC の他に，パソコンでよく用いられているシフト JIS コード，あるいは通信でよく用いられている新旧の JIS コードを許容し，Web インターフェースにおいては文字化けを起こさないよう工夫がされている．しかしながら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．

5 おわりに

エッセイの自動採点およびその評価は来るべき時代の学問であり，また実際的であるために社会的にも要求が高い．本稿では，自動採点システムの現状について妥当性を含めて説明するだけでなく，残された課題や今後の方向性についても一通り言及したつもりである．本稿がこの分野に関心のある研究者の一助となれば幸いである．

謝辞

2 人の匿名の査読者からは修正すべき多くの事項についてご指摘いただいた．自動採点システムの「最新動向」だけでなく，歴史的経緯やさまざまな議論，特に自動採点システムに望まれる要件や幾つかの妥当性についての議論を詳述することで，論文の内容を充実させることができたのは査読者のご教示によるものである．また論文の改訂にあたり，柳井晴夫教授，石井秀宗氏 (ともに大学入試センター) より御助言をいただいた．ここに記して深謝申しあげる．なお本稿は日本行動計量学会 第 7 回春のセミナー「知識社会のための情報・統計科学」での著者の原稿に加筆修正したものである．関係者各位に感謝したい．本研究については文部科学省科学研究費補助金基盤研究 (C)(研究代表者 石岡 恒憲，課題番号 16500628) の補助を受けた．

参 考 文 献

Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y.(1998):Topic Detection and Tracking Pilot Study Final Report, *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, 194–218. Available online: `http://ciir.cs.umass.edu/pubfiles/ir-137.pdf`

Bennet, 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.

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

Berry, M.W.(1992). Large scale singular value computations, *International Journal of Supercomputer Applications*, **6** (1), 13–49.

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

Beyesian Essay Test Scorint sYstem, BETSY, `http://edres.org/betsy/`

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & 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, 1998. Montreal, Canada. Available online: `http://www.ets.org/research/erater.html`

Burstein, J. & Wolska, M. (2003). Toward evaluation of writing style: Finding overly repetitive word use in student essays. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, Hungary.

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

Chase, C.I.(1979). The impact of achievement expectations and handwriting quality on scoring essay tests, *Journal of Educational Measurement*, **16** (1), 293–297.

Chase, C.I.(1986). Essay test scoring : interaction of relevant variables, *Journal of Educational Measurement*, **23** (1), 33–41.

Chung, G. & O'Neil,Jr. H. F. (1997). Methodological Approaches to Online Scoring of Essays, *CSE Technical Report* 461, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards and Student Testing, Available online: `http://www.cse.ucla.edu/CRESST/Reports/TECH461.pdf`

Chute, E.(2001). PG writers take intelliMetric software for a test drive, PG news, post-gazette.com, Available online: `http://www.post-gazette.com/regiostate/20011216essaysidep9.asp`

Cooper, P.L.(1984). The assessment of writing ability: a review of research, *GRE Board Research Report*, GREB No.82-15R. Available online: `http://www.gre.org/reswrit.html#TheAssessmentofWriting`

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R.(1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41** (7), 391–407.

Duff, I.S., Grimes, R.G., & Lewis, J.G.(1989). Sparse matrix test problem, *ACM Trans. Math. Software*, **15**, 1–14.

Electronic Essay Rater, e-rater, `http://www.ets.org/erater/index.html`

Elliot, S.(1999). Construct validity of IntelliMetric with international assessment, Yardley, PA: Vantage Technologies (RB-323).

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

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

福島 俊一・大竹 暁子・大山　裕・首藤 友喜 (1986). 日本語文書校正支援システム COMET, 信学技報, OS 86-21, 15–22.

福島 俊一・佐々木 仲太郎・赤石沢 元博・竹元 義美 (1992). 日本語文書校正支援システム St.WORDS, 情報処理学会 第 45 回全国大会, 6C-1. 275–276.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203–225.

Huang, X.D., Ariki. Y., & M.A. Jack, M.A. (1990). *Hidden Marcov Models for Speech Recognition*, Edinburgh University Press, Edinburgh.

Hughes, D.C., Keeling, B., & Tuck, B.F. (1983). The effects of instructions to scorers intended to reduce context effects in essay scoring, *Educational and Psychological Measurement*, **43**, 1047–1050.

Intelligent Essay Assessor, IEA, `http://www.knowledge-technologies.com/`

池原 悟・安田 恒雄・島崎 勝美・高木 伸一郎 (1987). 日本文訂正支援システム (REVISE), 研究実用化報告 **36**(9), 1159–1167.

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.

IntelliMetric, http://www.intellimetric.com/

石岡 恒憲・亀田 雅之 (1999a). 単語の共起に基づく関連文書検索, 算法と検索事例, 応用統計学, **28** (2), 107–121. Available online: http://www.rd.dnc.ac.jp/~tunenori/doc/jjasSvd.{pdf,ps}

石岡 恒憲・亀田 雅之 (1999b). 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 (2001). コンピュータによるエッセイの自動採点システム e-rater について, 大学入試フォーラム, **24**, 71–76.

石岡 恒憲・亀田 雅之 (2002). 特許：文章評価採点装置, プログラム及び記憶媒体, 出願番号, 特願 2002-31300.

石岡 恒憲・鷺坂由紀子・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. Available online: http://www.rd.dnc.ac.jp/~tunenori/doc/jess_kt.{pdf,ps}

Ishioka,T. & Kameda,M. (2004). Automated Japanese Essay Scoring System : Jess, *DEXA 2004 (15th International Conference on Database and Expert Systems Applications)*, Zaragoza Spain, 4–8. Available online: http://www.rd.dnc.ac.jp/~tunenori/doc/Ishioka_T_Jess.ps

Jones, K.S.(1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval, *Journal of Documentation*, **28** (1), 11–21.

Keith, T. Z.(1998) Construct Validity of PEG, American Educatiotional Research Association, San Diego, CA.

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.

Landauer, T.K., Laham, D., & Foltz, P.W. (2000). The Intelligent Essay Assessor, The Debate on Automated Essay Grading, *IEEE Intelligent Systems*, **15**(5), 27–31.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor, 87–112. In Shermis, M. & Burstein, J. eds. *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luhn, H.P.(1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information, *IBM Journal of Research and Development*, **1** (4), 307–317.

前川 守 (1995). 文章を科学する, 1000 万人のコンピュータ科学 3, 岩波書店.

Marc, D.(2000). *The Theory and Practice of Discourse Parsing and Summarization*, MIT Press, Cambridge, Massachusetts.

Marshall, J.C. & Powers, J.M. (1969). Writing neatness, composition errors and essay grades, *Journal of Educational Measurement*, **6** (2), 97–101.

McCallum, A. & Nigam, K. (1998). A comparison of event models for Naive Bayes Text Classification. *AAA-98 Workshop on "Learning for Text Categorization."* Available online: `http://citeseer.nj.nec.com/mccallum98comparison.html`

Meyer, G. (1939). The choice of questions on essay examinations, *Journal of Educational Psychology*, **30** (3), 161–171.

Mitchell, T. (1997). *Machine Learning*, WCB/McGraw-Hill.

MSNLP (2004). `http://research.microsoft.com/nlp/`

Myford, C.M. & Cline, F. (2002) Looking for Patterns in Disagreements: A Facets Analysis of Human ... Rater's and e-rater Scores on Essay Written for the Graduate management Admission Test (GMAT), *Annual Meeting of the American Educational Research Association*, April 1-5, 2002, New Orleans, LA. Available online: `http://www.ets.org/research/dload/AERA2002-myf.pdf`

長尾 真 (編)(1996). 自然言語処理, 岩波講座ソフトウェア科学 15, 岩波書店.

野矢 茂樹 (1997). 論理トレーニング, 哲学教科書シリーズ, 産業図書.

奥村 薫・脇田 早紀子・金子 宏 (1992). 日本語校正支援システム FleCS : 新聞社における実用化報告情報処理学会第 45 回全国大会講演論文集, 151–152.

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.

Page, E.B., Lavoie, M.J., & Keith, T.Z.(1996). Computer Grading of Essay Traits in Student Writing, *Annual Meeting of the National Council on Measurement in Education*, New York.

Page, E.B., Poggio, J.P., & Keith, T.Z.(1997). Computer analysis of student essays: Finding trait differences in the student profile. *AERA/NCME Symposium on Grading Essays by Computer.*

Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M.E., & Kukich, K. (2000). *Comparing the validity of automated and human essay scoring* (GRE No. 98-08a). Princeton, NJ: Educational Testing Service.

Project Essay Grade, PEG, `http://134.68.49.185/pegdemo/`

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*, Longman.

Rudner, L. & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, **7**(26). Available online: `http://PAREonline.net/getvn.asp?v=7&n=26`

Rudner, L.M. & Liang, L. (2002). Automated essay scoring using Bayes' theorem, *National Council on Measurement in Education*, New Orleans, LA. Available online: `http://ericae.net/betsy/papers/n2002e.pdf`

Shermis, M.D., Koch, C.M., Page, E., Keith, T.Z., & Harrington, S. (2002). Trait Rating for Automated Essay Grading, *Educational and Psychological Measurement*, **62**, [1], 5–18.

Sepos, M.(2000). Grading essay tests is going online in PA., 2000-11-06, Philadelphia Business Journal, Available online: `http://philadelphia.bizjournals.com/philadelphia/stories/2000/11/06/fofus7.html`

Tweedie, F.J. & Baayen, R.H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective, *Computers and the Humanities*, **32**, 323–352.

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.

Williams, R. (2001). Automated Essay Grading: An evaluation of four conceptual models, *Teaching and Learning Forum 2001.* Available online: `http://lsn.curtin.edu.au/tlf/tlf2001/williams.html`

Wresch, W. (1993). The Imminence of Grading Essays by Computer - 25 Years Later. Computers and Composition, 10(2), 45-58. Available online: `http://corax.cwrl.utexas.edu/cac/archiveas/v10/10_2_html/10_2_5_Wresch.html`

Yule, G.U.(1944). *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge.

著者紹介

石岡 恒憲 (いしおか つねのり)
(1) 独立行政法人 大学入試センター 研究開発部 試験作成支援研究部門 助教授
(2) 東京理科大学 大学院 工学研究科 経営工学専攻 修士課程 修了 (1985), 工学博士 (1992)
(3)

- Evaluation of criteria for information retrieval, *Sytem and Computers in Japan*, **35** (1), 42–49, 2004. (Translated from Denshi Joho Tsushin Gakkai Ronbunshi, J86-D-I (5), 293–300, 2003)
- 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, 2003.
- Maximum likelihood estimation of Weibull parameters for two independent competing risks, *IEEE Trans. on Reliability*, R-40 (1), 71–74, 1991.