社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

コンピュータによる日本語小論文の自動採点システム

石岡　恒憲† 　亀田　雅之††

† 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
†† 株式会社リコー ソフトウェア研究開発本部 〒 112–0002 東京都文京区春日 1-1-17
E-mail: †tunenori@rd.dnc.ac.jp, ††masayuki.kameda@nts.ricoh.co.jp

あらまし　アメリカで実施される適性試験の一つである GMAT (Graduate management Admission Test) において，実際に小論文の採点に用いられている e-rater を参考にして，その日本語版ともいうべき jerater を試作した．jerater は，文章の形式的な側面，いわゆる文章作法を評価する「修辞」と，アイディアが理路整然と表現されていることを示す「論理構成」と，トピックに関連した語彙が用いられているかを示す「内容」の 3 つの観点から小論文を評価する．毎日新聞の社説およびコラム (「余録」) を学習し，これを模範とした場合に適切でないと判断される採点細目に対して減点することで採点を行なう．また，書かれた小論文の診断情報を提示する．システムは現在 UNIX 上で動作し，800–1,600 字の小論文を通常能力のパソコン (Plat'Home Standard System 801S, Intel Pentium III 800MHz, RedHat7.2) で 1 秒程度で処理する．

キーワード　イー・ティー・エス (ETS), イー・レーター (e-rater), 自然言語処理, 統計的アプローチ

Tsunenori ISHIOKA† and Masayuki KAMEDA††

† Research Division, National Center for University Entrance Examinations
†† Software Research Center, RICOH Co., Ltd.
E-mail: †tunenori@rd.dnc.ac.jp, ††masayuki.kameda@nts.ricoh.co.jp

Abstract　We have developed an automated Japanese essay scoring system named jerater. The system evaluates an essay from three features: (1) Rhetoric — syntactic variety, or the use of various structures in the arrangement of phases, clauses, and sentences, (2) Organization — characteristics associated with the orderly presentation of idea, such as rhetorical features and linguistic cues, (3) Contents — vocabulary related to the topic, such as relevant information and precise or specialized vocabulary. The final evaluated score is calculated by reducing an point assigned by learning editorial columns in MAINICHI daily news paper. The diagnosis for the essay is also given.

Key words　Educational Testing Service (ETS), e-rater, natural language processing, statistical approach

1. はじめに

小論文試験においては，実施者は受験者のある種の能力が答案に反映していることを期待しているわけだが，その得点結果には，様々な要因が複雑に関与している．Cooper [7] によれば，「小論文が Writing Ability を測定しているものと考えると，その得点に関して誤差要因として働くものには，書き手 (writer)，題目 (topic)，形式 (mode)，制限時間 (time-limit)，テスト状況 (examination situation)，そして評定者 (rater) がある」という．これらの大部分はいわゆる「試験」に共通している要因であるが，特に「評定者」の要因は小論文においては決定的なものである．

他にも小論文試験では，得点に影響を与える以下のような多くの要因が存在し，それらについての多くの研究がある [29]．

- 文字の巧拙 (文字の上手さ，綴りの正確性)[5], [6], [22]
- 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)[13]
- 課題選択 (異なる課題に基づいて書かれた小論文をどう評価するか)[23]
- その他種々の誤差要因 (書き手の性別，人種など)[4]

このような誤差要因を排除するため，あるいは公平性の立場から，近年，コンピュータによる小論文の自動採点の研究が精力的に行なわれている [3], [10] ~ [12], [26]．このうち最も有名なものは，アメリカのテスト機関 Educational Testing Service, ETS が開発し，現在は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 [3], [14] であろう．e-rater は現在，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いる．ただ採点の全てがコンピュータに委ねられているわけではない．一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．文字どおり，採点の手間を半減させる目的で利用している（得点差が 1 点の場合は人間の採点が優先する）．

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こと．

組織化 (Organization): アイディアが理路整然と表現されていること．たとえば修辞的な表現，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．

内容 (Contents): トピックに関連した語彙が用いられているか．

e-rater 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，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ことにより，得点のためのメトリクスにかかる回帰係数を定めている．翻って我が国の場合は，オーソライズされた得点の蓄積がなく，同じようなアプローチは事実上，不可能である．

しかしながら，現在は言語学研究の目的で日外アソシエーツより「毎日新聞」の 2001 年までの全記事 (http://www.nichigai.co.jp/newhp/cdeb/index4.html) を，また日経出版販売より「日本経済新聞」の 2000 年までの全記事 (http://www.nikkeish.co.jp/gengo/zenbun.htm) を入手することができる．社説，コラム（「余録」）等，模範と考えられえる小論文を電子媒体で獲得するのは容易である．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(http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html) 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(http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp.html) や奈良先端科学技術大学院大学の SAX, BUP(http://cactus.aist-nara.ac.jp/lab/nlt/{sax,bup}.html)，東京工業大学 田中・徳永研究室の MSLR パーザ (http://tanaka-www.cs.titech.ac.jp/pub/mslr/index-j.html) 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 [15] などに詳しく，したがって模範となるエッセイやコラムを学習するというアプローチを取ることで，e-rater と結果として同様のことを，すなわち日本語で書かれた小論文の自動採点システムを，技術的にはより優れた方法を用いて開発できる，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 jerater（ジェイ・イー・レーター）と名付けたが，jerater 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（ちなみに e-rater の満点は 6 点である；また e-rater の配点は専門家による採点への線形回帰により定められる）．ユーザが指定しないときの配点として，「修辞」の重みを「論理構成」，「内容」の重みより高くして 5,2,3 と定めるのは，渡部 [29] の結果に基づいている． この研究では，小論文における採点基準として (1) 誤字・脱字，(2) 用語力，(3) 文字，(4) 文法，(5) 文体，(6) 課題のとらえ方，(7) 発想，(8) 文の構成，(9) 表現力，(10) 知識，(11) 論理性・一貫性，(12) 思考力・判断力，(13) 一人よがり，(14) 読語感，(15) 親近感，の 15 の観点を取り上げ，観点ごとの評価値との相関係数を出しているが，それによると「修辞」に関係の深い (3) 文字の相関係数が 0.58 と最も大きく，(1) 誤字・脱字も 0.36 と比較的大きな値を示している．「論理構成」に関係の深い (8) 文の構成，(11) 論理性・一貫性の相関係数はそれぞれ 0.32, 0.26 と「修辞」ほど大きくなく，「内容」に関係が深いと思われる (6) 課題のとらえ方，(14) 読語感，はそれぞれ 0.27, 0.32 であった．

次節以降では，jerater は採点基準の詳細について説明する．2 節には修辞，3 節には論理構成，4 節には内容について述べる．5 節には実施例を取り上げ，そのときの動作時間について記す．6 節はまとめである．

2. 修　　辞

jerater では修辞を示すメトリクスとして [21], [24] 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ブレークダウンし，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（箱髭図においては 1.5 倍を越えない最大，あるいは最小のデータの位置まで髭が描かれる .) 採点において，ブレークダウンした各メトリクスの比重は同等とした．唯一の例外は「語彙の多様性」の尺度であり，これだけがその重みを 2 倍にしてある．これは，この項目が修辞だけでなく，内容にも関与する指標であると著者らが判断したことによる．

2.1 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

一般に文章を分かりやすくするためには，文の長さは短い方がよいとされる [18]．また日本語の文章作成に関する多くの本は，一文の最大長さを 40 ないし 50 字に納めるのが適当である，としている．したがって，文の長さの中央値と最大値を指標の一つとした．平均でなく中央値を用いるのは，多くの場合，文の長さの分布が歪んだ分布であることによる．中央値と最大値の評価における比重は同等（以下同じ）とした．

また文の長さは文体を知るのにかなりの効果があることが知られている [31]．

（2） 句の長さの中央値，最大値

句点（．）と並んで，読みやすさに影響を与えるもう一つの要因は読点（，）である．読点と読点の間をここでは句と呼び，句の字数についても評価指標の一つとした．

（3） 句中における文節数の中央値，最大値

人間は一時に多くのことを理解できない．人間の短期記憶の限界は一般に 7 だと言われており，それが句の長さを制限していると思われる．実際，著者らが毎日新聞の社説から句中の文節数を求めてみたところ，その中央値は 4 で，短期記憶の 7 と整合性が高いことが確認されている．

（4） 漢字/カナの割合

一般に文章を易しくしたり，読みやすくするために漢字を減らすということは意図的に行なわれる．小論文においても適当な漢字とカナの比率の範囲が存在すると考え，これを評価指標の一つとした．漢字/カナの割合は，一般には文体の一つだと考えられている．

（5） 連体修飾（埋め込み文）の数

連体修飾の用言は，いわゆる「埋め込み文」の存在を示しており，この多寡が文章の分かりやすさに影響を与えると考えられる．

ただ著者らが用いた形態素解析システムの茶筌やそのベースとなった JUMAN では『連体形』という活用形が存在しないことに注意されたい．茶筌では用言の活用形の名称は『未然形』，『連用形』『基本形』『仮定形』『命令』を基本的な活用形とし，例外的な形のものに対してのみ，IPA 品詞体系 (THiMC097) の活用形を使用している．形容動詞を除き，用言の助動詞の終止形と連体形は同形なので『基本形』と統一しているのだと考えられる．

そこで

- 直後に名詞の類がくる「基本形」，あるいは
- 文末でもなく，終助詞に連ならない「基本形」

を「連体形」とみなした．ただし，形容動詞の場合は，活用語尾部の「体言接続」を連体形とみなした．

（6） 連用形や接続助詞の句の並びの最大値

連用形や接続助詞の句の並びが多いことも，文章の分かりやすさに影響を与えると考えられる．実際，マイクロソフト社の Word でも，接続助詞の句の並びはチェックしており，これが多すぎると赤字で警告を与えることは多くの人が経験しているであろう．

ただこの値は，平均的な大きさにはあまり意味がなく，係り受けの最大深さの方が，文章の分かり易さに影響を与える．係り受けの最大深さの代替として，連用形や接続助詞の句の並びの「最大値」を指標とした．

2.2 語彙の多様性

ユール [32] 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 特性値と呼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，

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

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．したがって 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．たとえば [28] などを参照されたい．

2.3 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．したが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．したがってヨミで 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．

なおヨミ以外に（日本語における）ビッグワードを，短単位の構成単語数で判断することも考えられる [17]．

2.4 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 [18]．したがって，これも修辞に関する評価指標となる．

受動態の文章は学校文法の品詞でいう助動詞の「れる」,「られる」で表記されることで能動態と区別される．もっとも「れる」,「られる」には，受け身とともに，尊敬，可能，自発の意味もある．このうち，能動であるにもかかわらず「れる」,「られる」が使われるのは尊敬の場合である．

しかしこの区別は形態素解析でも構文解析でもつかず，意味的なレベルでの解析が必要となる．たとえば，主語が「先生」や「ご主人」といった尊敬対象だった場合は尊敬の意味となるが，これは全くの意味の世界である．試験で用いられるような小論文には尊敬はないものとし，単純に「れる」,「られる」の有無だけで受動態とみなすこととした．

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．

ところが，日本語の文章においては一般に接続表現は敬遠されがちである．さらにいえば，曖昧な接続表現を好みさえする．そしてときには，曖昧に響きあう複数の叙述や問いかけが独特の効果を生み，名文ともなる [25]．

しかしながら試験で求められる小論文は名文ではない．意識的に接続表現を用いた論理的な文章である．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．実際，我々が参考とした e-rater においても，論文の「組織化 (Organization)」を測定するのに Quirk [27] にあるキュー・ワード (cue word, きっかけ語) による方法を用いている．これは “In summary” や “In conclusion” は要約を示す句であるとか，“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，といったことを判断するものである．

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 [25] 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」,「したがって」,「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「たとえば」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．

筆者らは，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．jerater 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 [19] を考えた．一般に N グラムモデルは確率有限オートマトンによって表現することができる．オートマトンの各状態は，トライグラムモデルにおいては，長さ 2 の記号列によりラベル付けされる．記号の集合は，$\Sigma = \{a : \text{順接}, b : \text{逆接}\}$ である．各状態遷移には表 1 に示す条件付き出力確率が割り与えられる．$\sqcup$ は何もないことを示す．初期状態は $\sqcup\sqcup$ である．たとえば，$P(a|\sqcup\sqcup)$ は初期状態で最初に a : 順接 が出現する確率をいう．

表 1　$\{a : \text{順接}, b : \text{逆接}\}$ の状態推移確率

$P(a\|\sqcup a) = 0.48$	$P(b\|\sqcup a) = 0.52$	$P(a\|\sqcup b) = 0.36$
$P(b\|\sqcup b) = 0.64$	$P(a\|aa) = 0.35$	$P(b\|aa) = 0.65$
$P(a\|ab) = 0.55$	$P(b\|ab) = 0.44$	$P(a\|ba) = 0.28$
$P(b\|ba) = 0.72$	$P(a\|bb) = 0.35$	$P(b\|bb) = 0.65$
$P(a\|\sqcup\sqcup) = 0.44$	$P(b\|\sqcup\sqcup) = 0.38$	

これより，論文中の $\{a : \text{順接}\}$ と $\{b : \text{逆接}\}$ の出現パターンに対する生起確率が，表 1 に示す条件付き確率の積をとることで得ることができる．たとえば，$\{a, b, a, a\}$ の出現パターンに対する生起確率 p は，$0.44 \times 0.52 \times 0.55 \times 0.28 = 0.035$ となる．

一方，事前情報なしに $\{a : \text{順接}\}$ の出現する確率は 0.47 で，$\{b : \text{逆接}\}$ の出現する確率は 0.53 であるから，順接が 3 回と

逆接が 1 回出現したときの，事前情報が与えられていないという条件のもとでの与えられた出現パターンの生起確率 q は $0.47^3 \times 0.53 = 0.055$ となる．

この例のように，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．

4. 内　　容

4.1 Latent Semantic Indexing

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 (以下 LSI と略す) を用いる．

LSI は予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 (たとえば [30] などを参照)

$$X = T_0 S_0 D_0'$$

することから始まる．T_0 および D_0 は，$T_0' T_0 = T_0 T_0' = I_t$ および $D_0' D_0 = D_0 D_0' = I_d$ 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 $0 \leqq d \leqq t$ 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，

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

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．Deerwester [8] によれば，言語データの場合，経験的に k は $50 \sim 100$ 程度にすればよい．

行列 X は一般に巨大な疎行列 (sparse matrix) となるが，このような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・パッケージとして，SVDPACK [2] が知られる．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に適用した場合の比較・評価については [15] に詳しい．なお，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 Harwell-Boeing sparse matrix format [9] に変換する必要がある．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．

4.2 LSI による文書間の類似度

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

$$d_e = x_e' T S^{-1}$$

を導くことができる．問題文 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余弦で与えることができる．

$$r(d_e, d_q) = \frac{(d_e, d_q)}{\|d_e\| \, \|d_q\|} \tag{1}$$

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．d_e と d_q が標準正規分布にしたがうとき，(1) 式はその相関係数と一致する．

われわれは，(1) 式で与えられる r を「内容」に割り当てられた配点を乗ずることで「内容」に対する評点とすることとした．r は理論的には負の値を取りうるが，その下限を 0 にすることは妥当であろう．

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 [20] 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 idf 法 [16] 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 [1] など)．e-rater では tf·idf 法が用いられている．

5. 実　施　例

e-rater における デモ は http://www.etctechnologies.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．

著者らは上記の Web ページに示している小論文 A ~ G を和訳し，それらを jerater で採点した (表 2)．

表 2　採点結果の比較

小論文	e-rater	jerater	字数	CPU 時間 (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

2 列目が e-rater の得点，3 列目が jerater の得点であり，4 列目が各小論文の字数である．jerater 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 jerater 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 jerater との違いが現われているように思われる．たとえば小論文 C においては，e-rater は満点の 6 点を与えうるが，jerater 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．

表 2 の第 5 列に jerater 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．jerater 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.namagu.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/jerater/ で実行可能である．

6. おわりに

jerater 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．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，たと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．したがって，内容の分析においては，評価対象の小論文に応じて，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あった方がよい．

本稿では各観点を直接的に測る指標はないと立場から，e-rater と同様に各々間接的かつ測定可能な指標を用いた．たとえば，「修辞」の観点では，間接的指標を多数用意して，それらの組合せで多面的な代替評価（多くの証拠での評価）を行った．しかしながら「論理構成」や「内容」については，用いている指標の数が必ずしも十分でない，と考えている．特に「内容」については，わずか 1 つの指標しか用いていない．たとえば「内容」については tf·idf 指標も考慮するなど，今後，「論理構成」や「内容」の観点での複数の評価指標を用いた多面的な評価を検討したい．

謝辞　e-rater の調査に際しまして，当時 ETS におられました村木英治先生（現，東北大学大学院教育情報学研究部教授）には e-rater 見学のアレンジをしていただきました．ここに記して厚くお礼申しあげます．

文　献

[1] Allan, J., Carbonell, J., Doddington,G., Yamron,J. and Yang, Y.:Topic Detection and Tracking Pilot Study Final Report, *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, February 1998.

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

[3] Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M.D.: Automated Scoring Using A Hybrid Feature Identification Technique. In the Proceedings of *the Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, 1998. Available online: http://www.ets.org/research/erater.html

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

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

[6] Chase, C.I.: The impact of some obvious variables on essay test scores, *Journal of Educational Measurement* **5** (4), 315–318, 1968.

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

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

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

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

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

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

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

[14] 石岡 恒憲: 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, 2001.

[15] 石岡 恒憲・亀田 雅之: 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, 1999. http://www.rd.dnc.ac.jp/~tunenori/jjasSvd.{ dvi,ps}

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

[17] 亀田 雅之: 擬似キーワード相関法による重要キーワードと重要文の抽出, 言語処理学会第 2 回年次大会,A4-6, 1996

[18] 木下 是雄: 理科系の作文技術，中公新書, 1981.

[19] 北 研二: 確率的言語モデル，言語と計算 4，東大出版会, 1999.

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

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

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

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

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

[25] 野矢 茂樹: 論理トレーニング，哲学教科書シリーズ, 産業図書, 1997.

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

[27] Quirk, R., S. Greenbaum, G. Leech & J. Svartvik: *A Comprehensive Grammar of the English Language*, Longman, 1985.

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

[29] 渡部 洋，平 由実子，井上 俊哉: 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, 1988.

[30] 柳井晴夫，竹内啓 (1983). 射影行列・一般逆行列・特異値分解，UP 応用数学選書 10，東京大学出版会.

[31] 安本 美典 (1994). 文体を決める三つの因子，言語 **23** (2), 22–29.

[32] Yule, G.U.: *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Cambridge, 1944.