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

石岡 恒憲* ・ 亀田 雅之**

要 旨

アメリカで実施される適性試験のひとつである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Jessを試作した. Jess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.

1. はじめに

大学入試センターの平成10年度の調査(鈴木,1999)によれば, 国立大学の全学科・コース(以下, 学科と記す)のうち85.7%が小論文を実施し, また平成12年度の私立大学における郵送調査(柳井・鈴木,2002)においても, 回答の得られた325大学723学科のうち70.0%(495学科)が小論文を実施している. いまや小論文試験は学科試験や面接試験と並んで, ごく普通に行なわれている試験のひとつになっている.

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(1984)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(渡部 他,1988).

- 文字の巧拙(文字の上手さ, 綴りの正確性), Chase(1968, 1979), Marshall & Powers(1969)など
- 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), Hughes *et*

* 大学入試センター 研究開発部 試験作成支援研究部門, 〒153-8501 東京都目黒区駒場 2-19-23, E-mail: tunenori@rd.dnc.ac.jp
** 株式会社リコー ソフトウェア研究開発本部, 〒112-0002 東京都文京区小石川 1-1-17, E-mail: masayuki.kameda@nts.ricoh.co.jp

al.(1983)など

- 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), Meyer(1939)など
- その他種々の誤差要因(書き手の性別, 人種など), Chase(1986)など

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(Burstein *et al.*,1998; Foltz *et al.*,1999; Page *et al.*,1997; Powers *et al.*,2000; Rudner & Liang, 2002). 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 (Burstein *et al.*,1998; 石岡,2001) 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. 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. 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点の場合はモードである4点に近い方の値が選ばれる.)

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. 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.

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, 石岡・亀田(1999)などに詳しく, 従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. (ちなみにe-raterの満点は6点である.) 渡部 他(1988)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. この結果を踏まえて, ユーザが指定しないときの配点を5,2,3と定めた.

次章以降では, Jessは採点基準の詳細について説明する. 2章には修辞, 3章には論理構成, 4章には内容について述べる. 5章にはe-raterとの仕様の差異にまとめておく. 6章には実施例を取り上げ, そのときの実行時間について記す. 7章はまとめである.

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. (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, 細分化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(木下,1981). また日本語の文章作成に関する多くの本は, 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. 従って, 文の長さの中央値と最大値を指標のひとつとした. 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(安本,1994).

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ひとつの要因は読点(、—,)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ひとつとした.

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, これを評価指標のひとつとした. 漢字/カナの割合は, 一般には文体のひとつだと考えられている.

   ただ著者らが, この指標の適正な範囲を求めるのに, 新聞の社説やコラムを利用したのは必ずしも適当ではなかったかもしれない. なぜなら新聞では限られた字数で最大限の情報を盛り込もうとするため, 多くの漢字が使われる傾向があるためである.

5. 連体修飾(埋め込み文)の用言の数

   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. 従って, 連用形や接続助詞の句の並びの「最大値」のみを指標とした.

2.2 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,

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

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.

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. 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.

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(木下,1981). 従って, 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. 例えば, 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.

蛇足ながら, 我々が用いている茶筌では, 「れる」は助動詞ではなく, 動詞の語尾として扱われている. これだと本来の語尾であるラ行下一段動詞(「流れる」など)の場合と区別できないと懸念されるむきもあろうが, 茶筌では上一段活用, および下一段活用は「動詞-自立 一般」という品詞に判別されるので, その心配は不要である.

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(野矢,1997).

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 *et al.*(1985)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. 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, $\Sigma = \{a:$順接$, b:$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. 例えば, $P(a|\sqcup\sqcup)$は初期状態で最初に$a:$順接が出現する確率をいう.

これより, 論文中の$\{a:$順接$\}$と$\{b:$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. 例えば, $\{a,b,a,a\}$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$\{a:$順接$\}$の出現する確率は0.47で, $\{b:$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.

表 1: $\{a:$順接$, b:$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(例えば柳井・竹内,1983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, $T_0'T_0 = T_0T_0' = I_t$ 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,

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

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 *et al.*(1990)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, Berry(1992)の SVDPACK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・評価については石岡・亀田(1999)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Duff *et al.*(1989)にあるHarwell-Boeing sparse matrix format 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.

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(曜日, 日など時間を表す副詞的な用法を持つ名詞, あるいは量や割合を表し副詞的に使うことのできる名詞, 「金曜」, 「一月」, 「少量」など), ナイ形容詞語幹(助動詞の「ない」の直前に現われて形容詞的な働きをする語, 「申し訳」, 「とんでも」など), 数, 非自立(連体詞や「の(格助詞)」, 活用語の基本形に接続するもの, 「嫌い」, 「以後」など), 特殊-助動詞語幹(他の品詞の連用形に接続して使われるもののうち, 学校文法で助動詞とされる「ようだ」の語幹部分, 「よう」, 「やう」など), 接尾(接尾語, 「君」, 「町」など), 接続詞的(単語と単語を接続する接続詞的な働きをするもの, 「(日本)対(アメリカ)」など), 動詞非自立的(〔助詞-接続助詞〕の「て」に接続するもので, 意味的には副詞的なもの, 「ご覧」など) は含まない.

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.

$$r(d_e, d_q) = \frac{(d_e, d_q)}{\|d_e\| \, \|d_q\|} \qquad (4.1)$$

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, (4.1)式はその相関係数と一致する. なおkの値は安全側にとって$k = 100$としてある. これはLSIの典型的な適用例である関連文書検索では, (4.1)式の計算は比較対象となる全ての文書に対して実施する必要があるが, 本稿での適用においては比較はわずか1回で済むためである.

われわれは, (4.1)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(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 *et al.*(1998)など)．e-raterではtf·idf法が用いられている．LSIとの比較については5章に後述する．

5. e-raterとの仕様の差異

本章ではJessのe-raterとの仕様の違いを整理しておく．

小論文を修辞，論理構成，内容の3つの側面から評価するという点において，両者は表記上の用語に違いはあるものの，実質的に同じである．しかしながら，これら評価尺度の重みづけをe-raterでは人間の採点に近づくよう線形回帰で求めるのに対し，Jessではユーザが指定できるようにした．これは採点者の意図を反映すべきだという立場による．しかしながらユーザが指定しない場合の既定値については，先行研究(渡部 他,1988)の結果を踏まえた値を用いている．またJessは毎日新聞の社説やコラムを学習しており，各種メトリクスの適正な範囲を定め，そこからの逸脱を検知することで減点するという仕組みになっている．

修辞の評価に用いたメトリクスには両者にかなりの共通がある．違いは主に対象言語の違いによるものであるが，e-raterのみに含まれるもので採点に少なからず影響を与える因子には「仮定を示す助動詞の数/割合」がある．一方，Jessのみに含まれる因子には，「漢字/カナの割合」がある．またJessでのみ考慮されている因子に，「ユール(Yule)のK特性値」，「ビッグワードの割合」がある．

論理構成の評価においても両者は似ている．e-raterがキューワードを用いているのに対し，Jessはその日本語に相当する接続詞により判断を行なう．ただしJessは接続詞の出現パターンも考慮している．また，議論を接続するために用いる指示代名詞の数，あるいはその割合にも着目している．

内容評価については，議論に相応しい単語が用いられているかどうかを判断するというコンセプトは同じであるが，実装方法は異なる．e-raterでは出現する単語に，単一文書中で出現する頻度(within-story term frequency)に応じて重みを与えるLuhn(1957)のtf法とその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(すなわちさまざまな文書に出現するありふれた単語の重みを低くする) Jones(1972)のidf法とを組み合わせたtf·idf法を用いているが，JessではLatent Semantic Indexing (LSI)を用いている．LSIの優位性はTRECなどで広く知られていることであるが，日本語の場合は，表記のゆれ(「インターフェース」と「インタフェイス」など) や異体字(「斉藤」，「斎藤」，「齋藤」など)，さらに翻訳語としての同義語(「コンピュータ」と「電子計算機」など)の問題があり，文字列一致によらないこの方法は，英語に比べさらに有利であると考えられる．

また，両システムとも得点だけでなく診断情報も併せて出力するが，e-raterが同じ評点に対し同じコメントを出すのに対し，Jessでは不適とされる項目について小論文に応じたコメントを表示する．

これらを要約した結果を表2に示す．

6. 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

表 2: e-raterとJessの仕様比較

	e-rater	Jess
評価尺度	構造, 組織, 内容	修辞, 論理構成, 内容
評価尺度への重み付け	専門家採点への線形回帰	ユーザ指定；既定値は先行研究に基づき設定
修辞の評価	仮定を示す助動詞の数/割合	漢字/カナの割合, ユールのK, ビッグワードの割合
論理構成の評価	キューワード	接続詞＋指示代名詞
内容の評価	tf-idf法	Latent Semantic Indexing
診断情報	同じ評点に同じコメント	小論文に応じたコメント
対象言語	英語	日本語

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．

我が国においては, オーソライズされた小論文の採点結果を得ることは, プライバシーの問題により事実上, 不可能である．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(その一例を付録Bに示す)．

採点結果を表3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 3: 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．参考までに, GMATにおける人間の採点の分布は正規分布ではなく, 6点満点中4点にモードがある左右非対称の分布である(Burstein *et al.*, 1998)．

表3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.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/jess/ で実行可能である.

7. おわりに

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. 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.

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教授(現, 東北大学大学院教育情報学研究部)にはe-rater見学のアレンジをしていただきました. 2名の匿名の査読者には, 論文の不備をご指摘いただき, 改良すべき事項についてご教示いただきました. 編集委員の南弘征先生, 編集委員長の水田正弘先生には審査ならびに編集上の多くのご尽力を賜わりました . ここに記して厚くお礼申しあげます.

付録

A 出力メッセージ一覧

Jessでは評点だけでなくその論文に相応しいコメントを出力する. e-raterでは同じ評点に対して同一のコメントしか出さないのであるが, Jessでは, 同じ評点でも異なったコメントを出力することができる. 以下は, Jessが出力するコメントのリストである.

A1 修辞に関する評価

“語彙の多様性が不足しています”
“語彙の多様性がやや不足しています”
“句(読点と読点の間, あるいは読点と句点の間)が総じて(平均的に)長いように見受けられます”
“句(読点と読点の間, あるいは読点と句点の間)が総じて(平均的に)やや長いように見受けられます”
“句(読点と読点の間, あるいは読点と句点の間)の長すぎる文があります”
“句(読点と読点の間, あるいは読点と句点の間)のやや長すぎる文があります”
“文が総じて(平均的に)長いです”
“文が総じて(平均的に)少し長いです”
“句の中の文節の数が総じて(平均的に)多いです”
“句の中の文節の数が総じて(平均的に)やや多いです”
“句の中の文節の数が多すぎる文があります”
“句の中の文節の数がやや多すぎる文があります”
“漢字の使用が少ないように見受けられます”
“漢字の使用がやや少ないように見受けられます”
“漢字が必要以上に使われているように見受けられます”
“漢字がやや必要以上に使われているように見受けられます”
“長くて難しい語が少ないように見受けられます”
“長くて難しい語がやや少ないように見受けられます”
“受動態の文が全体の分量に比べて多いように見受けられます”
“受動態の文が全体の分量に比べてやや多いように見受けられます”
“埋め込み文が全体の分量に比べて多いように見受けられます”

“埋め込み文が全体の分量に比べてやや多いように見受けられます”
“連用形や接続助詞の句の並びの多い文が, 幾つかあるように見受けられます”
“連用形や接続助詞の句の並びの多い文が, ややあるように見受けられます”

A2　論理構成に関する評価

“議論の接続が不十分であるように見受けられます”
“議論の接続がやや不十分であるように見受けられます”
“議論の掘り下げが不十分であるように見受けられます”
“議論の掘り下げがやや不十分であるように見受けられます”

A3　内容に関する評価

“問題文との関係が希薄であるように見受けられます”
“問題文との関係がやや希薄であるように見受けられます”

A4　分量に関する評価

“全体の分量が記号を除きxxx字以内で白紙答案とみなします”
“全体の分量が記号を除きyyy字以内とかなり少ないです”
“全体の分量が記号を除きzzz字以内とやや少ないです”

B　採点および出力メッセージ例

小論文A(表3)を採点した際の, 採点結果と出力メッセージは以下の通りである.

問題文: 人生において我々はしばしば自分がしたいことと, 自分がすべきだと感じることと, どちらを選んだらよいかに悩む場合がある. 自分がすべきだと感じることより自分がしたいことを優先させることがその人にとって良い場合があるとしたら, それはどのような場合だと思うか？あなた自身の経験, あるいはあなたが見聞したことから, その事例を挙げて, あなたの考えを述べなさい.

回答文: どのような状況においても, 人々はなすべき選択をもっている. ある者はときには自分がしたいことをするし, またあるときは自分がなすべだと感じたことをするであろう. 大抵の状況においては, 自分の望むことを押しやり, 他人がすべきであろうと感じることをすることがよいだろう.

大学に行くことは, あるものにとっては望むことでないかもしれないが, しなければいけないことである. これは真剣な選択である. 彼らがおこなう決定は, 物事の核心であり, 個々の人生に影響を与える. このような状況において, 忘れてはいけない最も重要なことは, いま自分が何を欲するかではなく, 人生の後になったときに何が欲しいかである. 大学に進むことによって, 彼らは自分の決定を後に延ばすことができる. 教育を受けて, よい仕事を得る能力をもっと得ることができるであろう. もし, 自分が望むからといって自分の教育を止めてしまったら, 数々の困難と共に道に倒れ伏してしまうかもしれない.

多くの10代の若者が直面する選択は, 自分の親のことを聞くかどうかである. 10代の若者には自分のしたいことが沢山ある. 問題なのは, 彼らが欲するということではなく, その要求が毎日変化するということである. 気のきいた者は, 自分にとっても優先度をつけ, 自分がなすべきだと感じることをするであろう. 多くの親達は, 自分の子供に正しい方向へ導こうとする. 大抵は彼らが自分の親のアドバイスを聞くことを知っているが, 彼らは

それを望んではいない. 将来になって, そのことは自分の要求を押さえ付けて, 自分のなすべきことを行なうのに役にたつであろう.

採点結果: 修辞3.0点(5点満点), 論理構成1.0点(2点満点), 内容2.9点(3点満点), 合計6.9点(10点満点),

出力メッセージ:

- 漢字が必要以上に使われているように見受けられます
- 埋め込み文が全体の分量に比べて多いように見受けられます
- 語彙の多様性が不足しています
- 議論の掘り下げが不十分であるように見受けられます

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`

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

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`

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

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.

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.

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

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.

Jones, K.S.(1972). A Statistical Interpretation

of Term Specificity and its Application in Retrieval, *Journal of Documentation*, **28** (1), 11–21.

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.

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

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

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.

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

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`

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.

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

石岡 恒憲・亀田 雅之(1999). 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.{dvi,ps}`

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.

木下 是雄(1981). 理科系の作文技術, 中公新書.

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.

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.

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.

鈴木規夫(1999). 3.1章 小論文総合問題に関する調査結果の概要, 平成8–12年度「大学の各専門分野への適性の評価を目的とする総合試験のあり方に関する共同研究」最終報告書, 大学入試センター研究開発部, 21–32.

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.

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.

柳井晴夫, 鈴木規夫(2002). 第3章 私立大学総合問題に関する調査結果, 「大学入学者選抜資料としての総合試験の開発的研究」平成11–

13年度 科研費補助金 基盤研究(B), 研究成果報告書.

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

JESS: AN AUTOMATED JAPANESE ESSAY SCORING SYSTEM

Tsunenori ISHIOKA* · Masayuki KAMEDA**

* Dept. of Applied Statistics and Measurement, Research Division, The National Center for University Entrance Examinations, Komaba 2-19-23, Meguro-ku, Tokyo 153-8501, Japan** Software Research Center, RICOH Co., Ltd., Koishikawa 1-1-17, Bunkyo-ku, Tokyo 112-0002, Japan

We have developed an automated Japanese essay scoring system named Jess. 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. Writing diagnosis will be also indicated.

Key words: Educational Testing Service (ETS), E-rater, Natural language processing, Statistical approach