日本語小論文の自動採点システムJessについて

石岡 恒憲（大学入試センター）

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，日本語小論文のための自動採点システム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程度を示す「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。

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一つになっている。

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。それらの要因の大部分はいわゆる「試験」に共通している要因であるが，小論文試験に特有で，かつ決定的な要因として「評定者」が挙げられる。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 1988）。

- 文字の巧拙（文字の上手さ，綴りの正確性）
- 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）
- 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）
- その他種々の誤差要因（書き手の性別，人種など）

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 1997; Rudner 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)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。

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。

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(1999)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。）この配点は，渡部(1988)の結果に基づいている。

次節以降では，Jessにおける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。

2. 修辞

Jessでは修辞を示すメトリクスとして前川(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。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，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。
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）の用言の数
   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。
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。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。したがって，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 特性値と呼ばれる語彙の集中度を示す指標である。

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。

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。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。

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)。したがって，これも修辞に関する評価指標となる。

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。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。野矢 1997によると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。

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。

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。

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。接続表現の個数について，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。いま記号の集合として$\Sigma=\{a:$順接，$b:$逆接$\}$としたときに，トライグラムモデルでは$\{a:$順接$\}$および$\{b:$逆接$\}$の出現確率が，その2つ前までの出現状況に依存すると考える。これにより，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，予め得られた条件付き確率の積をとることで得ることができる。

一方，事前情報なしに$\{a:$順接$\}$および$\{b:$逆接$\}$の出現する確率が得ることができるから，たとえば順接が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を得ることができる。

トライグラムモデルにおける生起確率と，事前情報のない場合の生起確率とを比較し，後者の方が大きいならば，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

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（毎日新聞のCD-ROMに収められている全社説/コラム3年分を利用）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）を特異値分解することから始まる。得られた特異値ベクトルの特異値の大きい方からk番目までとり，これを対角要素とする対角行列をSとする。それに応じて，k列までを抜き出した左右の特異値分解行列をそれぞれT，Dとすれば，

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

となり，$\hat{X}$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。'は転置を示す。Deerwester(1990)によれば，言語データの場合，経験的にkは50～100程度にすればよい。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。出題文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$r(d_e, d_q)$を，両文書ベクトルがなす角の余弦で与えることができる。われわれは，ここで与えられる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。

5. 実施例

E-raterにおけるデモ(http://www.etctechnologies.com/html/eraterdemo.html)では，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。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Jessで採点した。出題文は以下で与えられる。

「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。」

4節との対応で述べると，出題文がq，採点すべき小論文がeとなる。結果を表1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。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。

表1：採点結果の比較

小論文	E-rater	Jess	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(3.5)	504	0.95

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。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，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。

表1の第5列に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。なおJessはhttp://zaza.rd.dnc.ac.jp/jess/で利用可能である。

6. 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，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，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．

参考文献

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, Montreal, Canada. Available online: http://www.ets.org/research/erater.html

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.

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

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. http://www.rd.dnc.ac.jp/~tunenori/doc/jjasSvd.{dvi,ps}

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

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.*

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

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』28: 143-164.

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.

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