短答式記述テストにおける自動採点

―その採点ロジックと課題について―

石岡 恒憲（大学入試センター）

エッセイの自動評価採点に比較し，短答式記述テストに対する自動評価システムの研究はさほど活発ではない。その評価システムとして，おそらく唯一の実用システムである c-rater の仕様を報告しながら，短答式テストをコンピュータで行なうことの必要性と現在の技術水準における採点ロジック，および残された課題について解説する。

1 はじめに

アメリカでは教育関係者のみならず一般においても主観が入るエッセイ試験においては，コンピュータによる評価採点の考えが受け入れられるようになってきた。実際，アメリカのビジネススクール入学のための共通テストである GMAT における作文（エッセイ）テストでは，1998 年より e-rater (Burstein et al., 1998) が，2006 年よりは IntelliMetric (Elliot, 1999)が採点を行なっている。他にも商用のシステムとして，PEG (Project Essay Grade; Page,1999) や IEA(Intelligent Essay Assessor; Foltz et al.,1999) があり利用に供されている。我が国においても著者のグループが Jess (Ishioka & Kameda, 2006) を開発し，その成果は国内外の多くのマスコミで紹介された。

しかし，今ここで議論しようとしている短答式記述(short-answer)テスト（以下，短答式テストと略記する）をコンピュータで採点することについては，その重要性は認められているものの技術的にさまざまな課題が未解決のままである。Vigilante(1999) は世界最大のテスト機関である ETS (Educational Testing Service) とニューヨーク大学とで共同研究を行い，この分野における最初の報告をした。Leacock & Chodorow (2003)は，ETS が開発した c-rater の最新の仕様について報告している。Pulman & Sukkarieh (2005)は，品詞決定に隠れマルコフモデルを適用した情報抽出技術を用い，これにより予め用意した正解文と同じ意味構造を持つ幾つかの文を自動生成する試みについて述べている。

このように，1998 年ごろから短答式テストにおける自動採点の研究が進んでいるものの，著者の知る限り実用システムとしては ETS が唯一 c-rater を試作しているに過ぎない。しかも，その性能であるところの専門家との一致率は，エッセイ自動評価採点システム (e-rater)のそれに比べ 10%以上も小さい。

本稿では，短答式テストのコンピュータで行なうことの必要性と現在の技術水準における採点ロジック，および残された課題について解説を行なう。

2 なぜ短答式テストなのか；なぜ自動評価なのか

2.1 短答式テストの利点

多肢選択テストの代わりに短答式テストが用いられる最大の理由は，短答式テストの方がより真正(authentic)で信頼できると考えられているからである。実際，現実世界における質問応答は，多肢選択ではなく，短答式テストに近い。

他の理由としては，経済性が挙げられる。高品質な多肢選択問題を作成することは，通常，コストと手間がかかる。日米の多くの共

通テストでは，4〜5択の問題がしばしば用いられるが，このうち2つの選択肢における選択率の合計がどのような学力層においても90%を超えるような，すなわち実質的には2択になっているような問題は，注意深く作られた問題であっても存在する。

最後の理由は，多肢選択テストはテスト戦略を立てやすく，学生の問題についての理解を正しく評価することが難しいことによる。また当て推量による効果も無視できない。たとえば5択の問題では，当て推量で平均20%の正解を得ることができるから，SN比を最大とする，すなわち誤差要因に対して学力要因を最も正しく測定することのできる場合の正解率は60%である。このことは，センター試験の正解率が60%を目指して作成されているという事実に符合するわけだが，逆にいえば，60%から大きく外れた場合の試験問題は，そのテストとしての識別力が一般には低いことを意味している。正解率の低い方，たとえば正解率が20%あるいはそれ以下の問題，すなわち難問は，多肢選択においては構造的に識別力の低い問題にならざるを得ない。

2.2 自動評価採点の利点

ではなぜ，短答式テストに対して自動評価採点なのであろうか？

教育テスト関係者におけるモチベーションとしては，経済性が挙げられる。もし2人の人間による評価採点のうち一方をコンピュータに置き換えることができれば，試験の評価に対するコストを文字通り半減することができる。このやり方は，アメリカのビジネススクールでの入学共通試験であるGMATにおける作文試験での適用の仕方と同じであり，最終判定点の決定が人間側に委ねられていれば，一般には受け入れ易いと考えられる。

2番目のモチベーションはより重要であるが，短答式の自動採点は，システムがよくできてさえいれば，即座のフィードバックを返すことができる。このフィードバックは，繰り返し行なうことができるから，対話的な学習訓練に向いており，個人教育的な側面を有しているといえる。

近年では，他に説明責任が挙げられる。採点の論拠を示すことの重要性はますます増大してきている。

しかしながら，自動評価システムが成功するか否かは，システムの評価をいかに人間に近づけるか，あるいはそれが不可能なら，いかに一般に受け入れ易い採点の論拠を有しているかにかかっている。

3 短答式テスト

3.1 自動評価採点の仕組み

短答式テストをコンピュータに評価させるのに最も向いた出題形式の一つは，論説や物語において，読んだことの内容に関する質問に答えるものである。これは一般に読解テストと呼ばれている。

以下（図 1）は，読解テストのサンプル問題とその質問例である。

> ここに茶わんが一つあります。中には熱い湯がいっぱい入っております。ただそれだけでは何の面白みもなく不思議もないようですが，よく気をつけて見ると，だんだんにいろいろの微細なことが目につき，さまざまな疑問が起こってくるはずです。ただ一杯のこの湯でも，自然の現象を観察し研究することの好きな人には，なかなか面白い見物です。
> （中略）
> 湯気が上がるときにはいろいろの渦ができます。これがまたよく見ているとなかなか面白いものです。線香の煙でも何でも，煙の出るところからいくつかの高さまではまっすぐに上がりますが，それ以上は煙がゆらゆらして，いくつもの渦になり，それがだんだんに拡がり入り乱れて，しまいに見えなくなってしまいます。茶わんの湯気の場合だと，もう茶わんのすぐ上から大きな渦ができて，それ

がかなり早く回りながら上がっていきます。
茶わんの上で起こる渦のようなもので，もっと大仕掛けなものがあります。それは雷雨のときに空中で起こっている大きな渦です。陸地の上のどこかの一地方が日光のために特別に温められると，そこだけは地面から蒸発する水蒸気が特に多くなります。そういう地方の傍に，割合に冷たい空気におおわれた地方がありますと，前に言った地方の，温かい空気が上がっていくあとへ，入り代わりにまわりの冷たい空気が下から吹き込んできて，大きな渦ができます。そして雹がふったり雷が鳴ったりします。
（寺田寅彦「茶わんの湯」による。）

質問：――線部「そういう地方」とはどのような地方ですか？

図１：読解問題の例（第２回全国学力調査・中学・国語A（基礎），2008より）

もし我々が上記のような短答式問題への自動評価システムを構築しようとした場合に，最初に問題となるのは，なにをもって正解と判定するか，ということである。そのための最も安直な方法は，人間が正解か否かの判定を与えるというものである。そのための一つのアプローチは，いわゆる知能工学の分野では「教師学習」と呼ばれる，複数の正解あるいは不正解と判定される回答を学習し，トレーニングすることで分類の規則を構築することである。しかしながら不幸なことに，学習できるだけの信頼に足る「教師データ」を得ることは通常，難しい。現在，利用できるものとしては，1999年のTREC（Text REtrieval Conference：文書検索における著名な国際会議）コンテストで公開された200の質問について，応募した10－15のシステムの判定が利用できるに過ぎない。

そこで別のアプローチとしては，（これはETSのc-raterの開発グループがとっている方法であるが）回答をアンサーキーと比較するものである。アンサーキーは，もし文書がなんらかの正解を含んでいるとしたら，文書集合(back of the book)の中から正解を得ることができる。そうでないならば，人間の専門家（以下，専門家と略す）が適当な正解の集合を作る。アンサーキーは１つの質問に対して，１つ以上の正解からなる。例えば図２は図１からの質問とそのアンサーキーを示している。いったんアンサーキーを決めれば，入力された回答とアンサーキーの近さを，単語ベクトルの近さとして，たとえばTF-IDFモデルなどを利用して測定することができる。TFとは単語頻度(term frequency)のことで１つの文書に現れるある着目した単語の出現頻度をその重みとするものである。この重み付けの背景には「何度も繰り返し言及される概念は重要な概念である」という仮定がある。一方，ある単語がどの程度その文書に特徴的に現れるかという特徴性を示すために IDF (inverse document frequency)を用いる。IDFはある単語が全文書中でどの位の文書に出現するかの尺度であるが，実際は全文書数に対する出現文書数の逆数ではなく，その対数あるいはその対数に１を加えた値を取ることが多い。

質問：――線部「そういう地方」とはどのような地方ですか？

正解文キー：
陸地の上のどこかの一地方が日光のために特別に温められると，そこだけは地面から蒸発する水蒸気が特に多くなります。

アンサーキー：
日光に特別に温められたために，地面から蒸発する水蒸気の量が特に多い地方。

回答文（回答例）：
地面から蒸発する水蒸気の量が特に多い地方。

回答中の単語　再現率／精度
キー：　[光，温められた，地面，蒸発，水蒸気，特に多い，地方]
回答：　[地面，蒸発，水蒸気，特に多い，地方]
再現率：　5/7=71%
精度：　5/5=100%

図2：アンサーキーと回答－単語 再現率

3.2 正解を判定するための指標

回答文の正しさを判定するために2つの指標，すなわち「文の正確性(sentence correctness)」と「回答中の単語の再現率(answer-word recall)／精度(precision)」を使う。文の正確性とは次のようなものである。まず専門家が正解文キーを作る。これはその質問に最もよく答えた段落の文から構成される。正確性は単純に正解文キーと回答文とを比較することである。あるいはもし文の番号付けがされていれば文番号と比較する。

すなわち図2にあるように，質問に対する正解文キーは「陸地の上のどこかの一地方が日光のために特別に温められると，そこだけは地面から蒸発する水蒸気が特に多くなります」である。もし回答文としてこの文を与えれば，システムはそれを正解として，また違う文を与えれば正解でないと判定することになる。

しかしながら，正解を与えるのに複数の文を必要とする場合が10%強は存在する。また正解文キーは，質問に答えるのに不要な文言をしばしば含んでいる。アンサーキーの長さ（35字）に比べ，正解文キーの長さ（55字）を比較されたい（図2）。これより明らかに，正解性の指標に加え，別の指標が必要であることがわかる。

彼らそして我々の目指すべきものは，テキストからの文を単に抜き出しただけのものを正解とするのではなく，適切な答えを正しいと判定できるシステムである。したがって，正確性より木目の細かい測定方法の開発が求められている。

そのための一つの案はアンサーキーと回答文との単語の重複に基づいたもの，再現率を使うことである。この再現率の指標は，回答文とアンサーキーとの単語の重複数をアンサーキー内の単語数で割ったものである。完全な(100%)の再現率は回答文中に全てのアンサーキーの単語が含まれていることである。0%の再現率は，アンサーキーの単語のどれ一つも回答文に含まれていないことを意味する。

また再現率に加えて，精度を用いる。これは，いわば回答におけるゴミの少なさを示す指標で，回答文とアンサーキーとの単語の重複数を回答文の単語数で割る。完全な(100%)の精度は回答文中に全ての単語がアンサーキーに含まれていることであり，0%の精度は，回答文の単語のどれ一つもアンサーキーに含まれていないことを意味する。

図2には，もし回答文として「地面から蒸発する水蒸気の量が特に多い地方」とした場合の例を示している。回答文には，幹となる内容を示す単語として「地面，蒸発，水蒸気，特に多い，地方」があり，アンサーキーには7つの内容を示す単語（光，温められた，地面，蒸発，水蒸気，特に多い，地方）があるから，再現率として5/7(71%)と精度5/5(100%)を与える。

ここで精度は「文の正確性」に代替しうるものであることに注意する。精度が低いならば回答文の長さは通常，長くなる。また回答文は異質の単語を含むようになるから，精度は必然的に小さな値をとる。

また当然のこととして再現率の高い回答であればある程，正解（アンサーキー）に近い。しかし一般に精度と再現率の間には，一方を高くすれば他方が低くなるというトレードオフの関係があるから，ある一定の再現率のもとで，ある閾値を超えた精度のものを正解と判定することになる。TRECデータをもとに，

正解を判定するための再現率ー精度曲線を求め，これを用いて正解／不正解を判定する。

なお，本項で頻出する再現率と精度は，情報検索の分野では，検索システムの性能を示す指標として，またその両方を併せ持ったF尺度などと共に最も広く用いられている。この統計学的な意味について再検討を行なった報告としてIshioka (2003)があるので，興味のある方は参考にされたい。

4 性能評価

Hirschman ら(2000)によれば，再現率が25%を超えれば，専門家が正しいと判定したもののうちシステムが正しいと判定をするヒット率は93.6%であり，専門家が不正解と判定するもののうちシステムが正しいと判定をする誤り率(alarm rate)は 6.6%であるという。一般に再現率と専門家が正解と判定する率との間にはよい相関があるとしている。

しかしこの相関をもって，短答式テストへの自動評価の妥当性を示したことにはならないであろう。まず明らかに，単語の重複での判定はあまりに単純である。正解のうち 6%は 0%の再現率である（すなわちアンサーキーと正解文に単語の重複がまったくない)。また不正解回答のうち 1.7%は逆に 100%の再現率である。

専門家と自動評価システム(コンピュータ)との不一致を理解するために，TREC にある評価データの中から990のレスポンスをランダムに選び，専門家とコンピュータとの評価の違う 72 件について調査してある(Hirschman et. al, 2000)。結果を表1に示す。これによると，うち7件は，そもそもTRECにおける人間の評定が誤っているものである；27 件は人間とコンピュータとでどちらが正しいのか不明瞭であるものである。つまり約半分（47%）が，自動評価システムの誤りとはいえないものである。残りの38件(53%)が単語の再現率の閾値の設定が悪いために明らかに自動採点システムの評定が間違っているものである。

表 1：専門家と自動採点との判定が異なった 72 件の分析（Hirschman et. al, 2000)

不一致の原因	件数	%
TREC 評価の誤り	7	10%
正解と関連があると思われる	27	37%
再現率閾値不適のための誤り	38	53%
計	72	100%

しかしながら，彼らによれば，これらの誤りのうちの半分(19/38)は，数学的表記の誤りに類するもので，例えば，回答が 3 でアンサーキーが“three”であるものや，Tuesday や“April 3”のようなものである。7/38 は回答の細かさ（粒度）や言い回しによるものである。たとえばアンサーキーが“George Washington”であるのに Washington と答えるものである。残りの 12 は別の問題に起因する。したがって，今後，開発が進めば，再現率は人間の判定に近づけることができ，これは多くの場合，最新の技術を実装することで実用的なレベルにまで引き上げることができるとしている。

Leacock & Chodorow (2003)では，2003年時点の c-rater の仕様が紹介されているが，これによれば同義語の辞書を完備し，専門家との一致率は 84%であるとしている。

5 おわりに

現在 c-rater で用いている回答中の単語の再現率は，非常に限定的な評価指標であり，回答を正しく評価するための以下のような重要な側面を無視しているといえる。

- 知性（単なる単語の集まりではなく，正解としての繋がりのよさ)
- 明瞭性（詳細さと余計なものを含んでいないこと；これについては精度がある程度は判定している)
- 弁明性（他の文からの論拠の提示)

・ 適切性（与えられた指示に適切に答えているか；よいフレーズを用いた最適な答えか）

精度のような詳細な指標が短答式の評価には求められるが，テキストからの単なる抜き出しをしたら，それとは相関が低くなるような指標もまた必要であろう。正しい答えであることとはどういうものであるかを判定するための新たな指標や，いままでの検討の次元を超えたような指標について研究・開発をする必要があろう。

本稿では，単なる技術紹介でなく我が国の大学入学者選抜に役立てるという観点から，英文における採点システムであるc-raterの仕様を，あえて日本語における例題を用いて説明することを試みた。我が国で将来このような採点システムを採用するか否かは，実際的な要請や社会的合意など多くの問題のために全くもって不明であるが，少なくとも技術的な課題や適用限界については知っておく必要がある。日本語で書かれた，また日本語を対象とする公刊されたこのような解説記事はかつてなく，本稿がこの分野に興味をもつ研究者や教育関係者の一助となれば幸いである。

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*, Montreal, Canada.

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

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

Hirschman, L., Breck, E., Light, M., Burger, J. D. and Ferro, L. (2000). Automated Grading of Short-Answer Tests. *IEEE Intelligent Systems, Trends and Controversies section*, 15(5): 22–37.

Ishioka, T. and Kameda, M. (2006). Automated Japanese Essay Scoring System based on Articles Written by Experts, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney Australia, July, 233–240.

Ishioka, T. (2003). Evaluation of Criteria for Information Retrieval, *IEEE/WIC International Conference on Web Intelligence (WI 2003),* Sponsored by IEEE Computer Society and Web Intelligence Consortium, Halifax, Canada, 425–431.

Leacock , C. and Chodorow , M. (2003). C-rater: Automated Scoring of Short-Answer Questions, Computers and the Humanities, Springer Netherlands, ISSN 0010-4817 (Print) 1572-8412 (Online), 37(4), 389–405.

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

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic Short Answer Marking. Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, *Association for Computational Linguistics.* 9–16.

Vigilante, R. (1999). Online Computer Scoring of Constructed-Response Questions, *Journal of Information Technology Impact*, 1(2), 57 –62.