Vol.2016-CE-133 No.19
2016/2/14

コンピュータ上で実施する記述式試験について

石岡恒憲[†1]

概要：自由形式で解答を記述するいわゆる記述解答試験において，エッセイタイプのものと短答式タイプのそれぞれについて，代表的なシステムにおける採点のロジックについて解説する．短答式解答については，この分野の研究が加速していることから，研究の方向性について説明する．また，全米学力調査(National Assessment of Educational Progress, NAEP)で実施された最新の記述試験のコンピュータによる出題形式について動画を交えて報告する．

キーワード：自動採点システム，エッセイ（小論文）採点，短答式テスト採点，記述試験，全米学力調査(NAEP)

Computer-based Writing Tests

TSUNENORI ISHIOKA[†1]

Abstract: Writing tests include essay and short-answer tests. We present both of their procedures and scoring systems. In particular, we explain the latest technologies using recognizing textual entailment in short-answer scorers. A writing test format that made full use of multimedia materials for the National Assessment of Educational Progress in 2011 is also presented.

Keywords: Automated Scoring System, Essay Scoring, Short-answer, Writing Test, National Assessment of Educational Progress

1. はじめに

オープンエンドといわれる，いわゆる自由形式で書くことのできる記述試験には大別して２つのタイプがある．ひとつはエッセイ（小論文）タイプの試験であり，もう一つは短答式の試験である．

前者（エッセイタイプの試験）では，基本的に正解がなく，日本語では 800 字程度から 1600 字程度を書く．評価の基準としては，修辞（文の上手さ，語法の使い方など），論理（論の進め方，論の掘り下げ，例示の使用など），内容（プロンプトと呼ばれる論題に十分に応えているか）などが用いられる．このタイプの試験については Educational Testing Service (ETS) の e-rater[1] や Vantage Learning 社の IntelliMetric[2]などハイステークスな試験に用いられている実用的なシステムがあり，人間（専門家）による評価に十分に近いことを示す多くの例証がある．他方，短答式試験は，通常，望まれる正解がある．したがって，システムは用意している正解と回答が同義であるか否かを判定する．正解は通常，１文か多くて２文である．この短答式記述の自動採点については，エッセイタイプの自動採点ほどには研究が進んでおらず，知られているシステムとしては Pacific Metrics 社の CRASE[3]や ETS の c-rater などわずかである．またこれらはおそらくまだハイステークスな試験には用いられていない（開発元の Web ページにはそのような情報の記述は認められない）．しかしながら，この「正解」と「解答」の同義判定やその含意関係を判定する技術は，「含意関係認識(Recognizing Textual Entailment)」と呼ばれ，いま自然言語処理の分野で最もホットな話題の一つになっている．このため，この短答式記述の自動採点の研究は急激に進歩することが期待されている．

本発表では，エッセイタイプの試験について，どのような自動採点システムが存在して，どのような観点から回答を評価するのかについて解説する．つぎに，短答式記述の試験について，現在行われている評価方法と，解決すべき課題について報告する．最後に，全米学力調査(National Assessment of Educational Progress, NAEP)[4]で実施された最新の記述試験のコンピュータによる出題形式について報告する．ただし NAEP の採点は人が行なう．

2. エッセイ（小論文）採点システム

代表的なエッセイ採点システム(Automated Essay-scoring System, AES)を表 1 に示す．Bill Burstein が開発し ETS が提供する e-rater[1]は 2004 年より Ver.2 が提供され，わずか 12 の変量によって採点を行うシステムに改変された．Ellis Page による AES の草分けである PEG[5]は 2002 年より Measurement Inc.がその権利を得て，現在も開発を続けている．Foltz と Landauer が開発した Intelligent Essay Assessor (IEA)[6]は当時，情報検索で主流になりつつあった潜在的意味解析(Latent Sematic Indexing, LSI)をいち早くエッセイ採点に取り入れたシステムであり，現在は Pearson Educational Technologies が提供する商用システムや試験に用いられている．IntelliMetric[2]はルール発見に基づく人工知能をベースとした AES である．Vantage

†1 (独)大学入試センター＆東京工業大学
The National Center for University Entrance Examinations, Tokyo Institute of Technology.

Learning 社が巨額の予算を投じて集中的に開発した．BETSY(Bayesian Essay Test Scoring sYstem)[7]は Lawrence Rudner によって開発されたベイズ・アプローチによる AES である．オープンソースとして公開されているが，これを用いた商用システムはまだない．現在，最も注目されている AES は Mitzel と Lottridge が開発し Pacific Metrics 社が提供する CRASE である．統計学と人工知能の両方のアプローチによって実装されている．このシステムは他の AES とは異なり，短答式記述採点エンジンをも含んでいる．日本語を処理する AES としては石岡と亀田が開発した Jess[8]がある．毎日新聞の社説やコラムから良い作文の作法を学習しており，統計学でいうところの異常値発見に基づいて採点を行う．

2012 年にはヒューレット財団がスポンサーとなり Automated Student Assessment Prize(ASAP)と呼ばれる Kaggle（企業や研究者による最適モデルのコンペ）が実施された．201 の応募者が参加し，8 つの論題に対して AES の性能を競った．9 つの AES ベンダーは招待された．調査の結果は AES が人間の評価者に比べ信頼できるというものであるが，それに対しては多くの反論がある．これらについても解説する．

表 1　エッセイ評価システムの比較

Table 1　The comparison of automated essay scoring system.

評価システム	評価基準	手法	制限	開発
E-rater V.2	構造/組織化/内容	重回帰モデル	12の評価指標	ETS[1]
PEG	構造/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	Measurement Inc.[5]
IEA	内容/文体/技巧	LSI	論理構成/語の出現順を評価しない	Peason Education[6]
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	論題ごとに大量のデータが必要	Vantage Learning[2]
BETSY	表層	ベイズ的接近	分野が制限:開発中	L.Rudner[7]
CRASE	アイディア/組織化/態/語彙選択/文の流暢さ/慣習	統計＆ルール発見	短答式採点可	Pacific Metrics
Jess	修辞/論理構成/内容	外れ値検出&LSI	科学技術分野に弱い	石岡･亀田[8]

3. 短答式記述採点システム

短答式記述採点を可能にする最も正統的なアプローチは含意関係認識であると思われる．この分野については現在，国立情報学研究所が主催する国際ワークショップ「NTCIR」の中で，含意関係認識に関するタスク「RITE」を 2011 年から実施している．そこに示されている例題[9]は，以下の 2 つの文章について，含意関係が成り立つかどうかを判定するものである．

t1　鎌倉幕府は 1192 年に始まったとされていたが，現在では実質的な成立は 1185 年とする説が支配的だ．

t2　12 世紀に日本では鎌倉幕府が開かれた

人間ならば t1 が成り立つとき，t2 も成り立つことを容易に判断することができる．しかしコンピュータがこれを判断するには，まず「鎌倉幕府（が）1185 年（に）成立した」といった意味構造を正しく解析することに加えて，「1185 年が 12 世紀である」という時間情報処理や「成立する≒開く」の含意関係知識が必要となる．含意関係認識は文章のレベルでコンピュータが人間の言葉の意味を理解することをめざしており，もしこれが実用レベルにまで達すれば，短答式記述の自動採点はほぼ実現されるといってよいだろう．ただ著者の認識する限り，含意関係認識はまだ開発途中で，達成すべき課題や言語資源の整備がさらに必要であると思われる．

4. 全米学力調査における記述テスト

2011 年の全米学力調査(National Assessment of Educational Progress, NAEP)[3]では，8 年生と 12 年生に対して，作文テストが従来の紙筆テストに換わり，初めてコンピュータによって実施された．そのテストは，単に従来の紙と鉛筆をコンピュータに置き換えただけのものではない．現在のデジタル技術が十分に活用できるよう，取り扱う作文のタイプには，テキストによる問いかけのほか，写真を含むもの，音声によるもの，ビデオを見て問いかけに答えるものの 4 つがある．加えて，多くの学生が受験できるようユニバーサルデザインについての配慮がなされている．たとえば，問題文についての音声読み上げやフォントサイズの変更が可能である．また電子上のスペルチェックが利用できる．

発表では，この初めてのコンピュータ化された作文試験について，その仕様を動画を交え示すとともに，今後の方向性や果たすべき課題について論考する．

参考文献

1) Y. Attali and J. Burstein, Automated essay scoring with e-rater v.2.0 (ETS RR-04-45), Princeton, NJ：Educational Testing Service, (2005).
2) S. Elliot, IntelliMetric: From here to validity, Automated essay scoring: A cross disciplinary perspective, M. Shermis and J. Burstein, eds., pp.71-86, Hillsdale, NJ：Lawrence Erlbaum Associates (2003).
3) Pacific Metrix, CRASE, https://www.pacificmetrics.com/products-and-solutions/crase/
4) National Assessment of Educational Progress, NAEP (2011). http://nces.ed.gov/nationsreportcard/
5) E.B. Page, Project essay grade: PEG, Automated essay scoring: A cross disciplinary perspective, M. Shermis and J. Burstein, eds., pp.43-54, Hillsdale, NJ：Lawrence Erlbaum Associates, (2003).
6) K.T. Landauer, D. Laham, and W.P. Foltz, Automated scoring and annotation of essays with the intelligent essay assessor, Automated essay scoring: A cross disciplinary perspective, M. Shermis and J. Burstein, eds., pp.87-112, Hillsdale, NJ：Lawrence Erlbaum Associates, (2003).
7) Bayesian Essay Test Scoring sYstem, BETSY, http://edres.org/betsy/
8) T. Ishioka and M. Kameda, Automated Japanese essay scoring system based on articles written by experts, Proceedings of the 21st Intl Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics (Coling-ACL 2006), P06-1030, pp.233-240 (2006). http://www.aclweb.org/anthology/P/P06/P06-1030
9) 問われるのは意味を理解する力．暗記だけでは解けない社会科科目, [特集]人工頭脳プロジェクト「ロボットは東大に入れるか」NII Today 9, No.60 (2013). http://www.nii.ac.jp/userdata/results/pr_data/NII_Today/60/p8-9.pdf