短答式記述テストにおける自動採点

石岡恒憲（大学入試センター）

Lynette Hirschman (The MITRE Corp.)

エッセイの自動評価

- アメリカでは受け入れられる共通認識
  - GMATにおける作文テスト
    1998ー2005： e-rater
    2006ー ： IntelliMetric
  - ーMCATにおける作文テスト（2007ー： IntelliMetric）
- 商用システム
  - PEG（Project Essay Grade）
  - IEA（Intelligent Essay Assessor）
- 日本語処理： Jess

短答式記述(short-answer)テスト の自動採点

- 多くの技術的未解決の問題
- ETS とニューヨーク大学との共同研究 (Vigilante, 1999)
- c-rater, 試作（Hirschman et al. 2000）
- c-rater, 同義語辞書の追加 (Leacock & Chodorow , 2003)
- 正解文との同義文を自動生成（Pulman & Sukkarieh, 2005)

短答式テスト自動採点の現状

- 1998年頃から研究開始
- 実用システムはc-raterのみ
- 性能（専門家との一致率）
  - エッセイ自動評価 (e-rater, 6点法)：96%
  - c-rater:84%

目 次

- 短答式テスト自動採点の必要性
- 現在の技術水準における採点ロジック
- 残された課題

なぜ短答式テストなのか

- 短答式テストの方がより真正(authentic)で信頼できる
  - 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
- 経済性
  - 高品質な多肢選択問題の作成は、コストと手間がかかる
- 多肢選択テストはテスト戦略を立てやすい
  - 理解力を正しく評価することが難しい
  - 当て推量による効果

なぜ自動評価なのか

- 教育テスト関係者からは「経済性」
  - 2人の人間のうち一方を機械に置き換え
  - 評価コストの半減
  - GMATにおける作文試験での適用と同じ
  - 最終決定が人間側に委ねられている
- 即座のフィードバック
  - 繰り返し;対話的な学習訓練;個人教育的な側面
- 説明責任
  - 採点の論拠を示すことの重要性

読解問題の例（素材文）

- （2000年1月18日）NASAの火星探査機からの信号を失って1か月以上経ち、ミッション・コントローラたちは、それを探す希望を捨てた。火星探査機の火星のおけるミッションは、その大気を調査し、水や、火星上に生命がいるか否かを科学者に判断させるための資料を探すことである。火星探査機は、90日間の任務のために12月3日に火星の南極の近くに着陸した。探査機からは降下し始める数分前から何も聞かれなくなった。3本足の探査機との最後の交信は、欲求不満を抱えたまま、月曜日午前8時に終了した。
  “我々は何も見ることはできなかった”、とNASAのジェット推進研究所のプロジェクトマネージャーであるリチャードクック氏は、言った。火星でのミッションは、米国政府に2億ドル以上の費用をかけて、結局は失敗に終わった。いまや宇宙船エージェンシーの科学者やエンジニアたちは、何が悪かったのかを突き止めようとしている。彼らは次のミッションで同じような過ちを起こしたくはないのである。コントローラたちは、次の着陸で起こりうるいくつもの困難なシナリオを想定し、何度もテストを繰り返している。（出典：CBCオンラインニュース、CBCラジオニュース、NASA）

読解問題の例（設問）

A) 火星探査機のプロジェクトマネージャーは誰ですか？

B) 火星探査機の使命は何ですか？

C) コントローラたちが探査機と交信する希望を捨てたのはいつですか？

D) 火星上のどこに宇宙船は着陸したのですか？

E) なぜNASAは火星探査機が水を探すことを望んでいるのですか？

評価システム構築 の仕方

- 教師学習
  - 複数の正解/不正解と判定される解答を学習
  - 分類の規則を構築
  - 信頼に足る「設問」と「教師データ」の不足
- 解答した答えをアンサーキーと比較
  - ETSのc-raterの開発グループ
  - 開発初期段階は正解文キーと比較
    →解答に不要なフレーズを含む

アンサーキー との比較

- 質問:
  B) 火星探査機の使命は何ですか？
- アンサーキー
  - 火星の大気を調査し水を探すため
  - 火星上に生命がかつて 存在していたか否かを科学者に判断させるための手助けをするため
- 解答文（解答例）
  - 大気を調査するため

解答中の単語　再現率／精度

- キー1:　[火星, 大気, 調査, 水, 探す]
- 解答:　[大気, 調査]
- 再現率:　2/5=40%　←必要なものが拾えているか
- 精度:　2/2=100%　←ゴミのなさ
- キー2:　[火星, 生命,かつて,存在,否か,科学者, 判断,手助け]
- キー1の方が大きな値を与える→採用

再現率と精度はトレードオフ

- 解答文を冗長に (長めに)
  - 再現率→大きくなる
  - 精度→小さくなる
- 解答文を簡潔に
  - 再現率→小さくなる
  - 精度→大きくなる
- ある再現率のもとで、ある閾値を超えた精度のものを正解と判定 (TRECデータで取得)

性能評価

- 再現率25%超のとき、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：93.6%
- 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)：6.6%

→自動評価の妥当性を示したことにはならない

- 単語の重複での判定はあまりに単純
  - 正解のうち6%は0%の再現率
  - 不正解解答のうち1.7%は逆に100%の再現率

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%

- TRECにある評価データの中から990のレスポンスから選択

システムの誤り38件の分析

- 半分(19/38)は数学的表記の誤り
  - アンサーキーが“three” で解答が3
  - Tuesdayや“April 3”
- 7/38は解答の細かさ (粒度) や言い回しによる
  - アンサーキーが“George Washington”で解答がWashington
- 残りの12は別の問題に起因
- 今後、開発が進めば、再現率 は向上

c-raterの現状(Leacock, 2004)

- 辞書的類似性(lexical similarity)の強化
  - 文字列の部分一致
  - 態に依らない
    - Walter wants money.
    - Money concerns Water.
    - Money is important to Walter.
- スコアの信頼度の提示

課題：無視している重要な側面

- 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)
  - 精度がある程度は判定
- 弁明性
  - 他の文からの論拠の提示 → 構文解析
- 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

おわりに

- 正しい答えであることとはどういうものであるか
- 判定するための新たな指標
- いままでの検討の次元を超えたような指標

→F_1尺度、フォールアウトの利用

ご清聴ありがとうございました