教育テスト研究センターセミナー
2007年7月20日、東京オペラシティタワー

コンピュータによるエッセイ、小論文の自動採点について

石岡恒憲（大学入試センター）

小論文/エッセイの自動採点および評価

- 現在、教育測定における最もホットな話題の一つ
- 自然言語処理に膨大な言語集体(コーパス)を利用した確率・統計的なアプローチ
- 有効性が多くの研究者や技術者に広く認知されてきた

小論文/エッセイの自動採点および評価

- 成功例のアプリケーション:
  - 機械翻訳（日→英、英→日、アラビア→英）
  - 音声認識
  - カナ漢変換（IME）
  - 情報検索（Web検索）
  - 文書要約（重要文抽出→要約文生成）

隠れマルコフモデル（HMM）

Time flies like an arrow.

「光陰矢の如し」

Pi(名詞)Po(time|名詞)Pt(動詞|名詞)Po(flies|動詞)Pt(前置詞|動詞)Po(like|前置詞)Pt(冠詞|前置詞)Po(an|冠詞)Pt(名詞|冠詞)Po(arrow|名詞)

「時蝿は矢を好む」

Pi(名詞)Po(time|名詞)Pt(名詞|名詞)Po(flies|名詞)Pt(動詞|名詞)Po(like|動詞)Pt(冠詞|動詞)Po(an|冠詞)Pt(名詞|冠詞)Po(arrow|名詞)

- 品詞という状態がわからない→「隠れ」
- 前向きの遷移＜後向きの遷移
- ATOK → IME

小論文/エッセイの自動採点および評価

- 自然言語である小論文/エッセイのテストに最近の自然言語処理での研究成果を取り込む
- アメリカ国防省による潤沢な研究費
  - テロの予兆発見
  - 盗聴

自動採点および評価の利点

- 評定者による評点のバラツキ
  - ハロー（光背）効果
- 評定の系列的効果（何番目に評価したか）
- 課題選択（異なる課題に対してどう一元的に評価するか）
- 採点の手間を大幅に低減
- 対話的な作文指導
- 説明責任

発表の流れ

・コーパスに基づく自動採点システムの開発・実用化（2000年）以前→過去
- それ以降、現在まで→現在
- 構成
  - 過去
  - 現在
  - 未来
- Jess
  - デモ（Web版、Closed版）

過去

- 先行研究の歴史
- システム概説
- 自動採点システムに対する批判

先行研究の歴史

- Page(1966)に始まる
- Project Essay Grade, PEG
  - 大規模テストにおけるエッセイ評価の教員の負担低減
  - テキスト特徴量に係る重回帰における重み係数
  - PEGスコアと教員スコアとの相関係数は0.78
  - 教員同士の相関0.85に近い

Project Essay Grade, PEG (1966)

- 自動的に抽出される特徴量は表面的なもの
  - 平均ワード長さ、エッセイの長さ(ワード数)、コンマの数、前置詞の数、一般的でない(uncommon)ワードの数
  - 本来測定しようとする作文要素の代用
- 作文スキル（内容、組織化、文体）を直接的に測定していない
- 間接的な指標を用いているために、トリックを使って良いスコアを人工的に得ることができる

Writers Workbench (WWB)

- 1980年代の初期に開発された作文ツール
- スペリングや語法、可読性(readability)について書き手に有用なヘルプを与える
- 可読性の指標を、文章に含まれるワード、文節の数に基づいて提示
- テキストの表面を粗くなぞっただけのプログラム
- 作文品質の自動評価を行うための1ステップ

文書校正支援システム（1980s）

- WWBの日本語版
- NTTのREVISE
  - 日経新聞社のVOICE-TWIN
  - 音声読み上げ（自然読みと違う）
- COMET
  - 講談社のSt.WORDS
  - 産経新聞社のFleCS
- 現在でも校正の現場で実際に利用されている

1990年代

・自然言語処理や情報検索の急激な進歩
・作文の品質測定に直接役立てる試み

E-rater (1998)

- Educational Testing Service, ETS
- ETS Technologies, Jill C. Burstein
- GMATにおけるAWA
- 以下の3つの観点をより直接的に測定する

E-raterにおける採点の観点

- **構造(Structure):**
  - 文法の多様性
  - フレーズ/文節/文の配列が多様な構造で表現されているか
- **組織化(Organization):**
  - アイディアが理路整然と表現されているか
  - 修辞的な表現/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 (1998)

- 専門家によって採点された膨大な数の小論文の蓄積
- 専門家の得点とコンピュータによる得点とを線形回帰
- 得点のためのメトリクスにかかる回帰係数を決定
- プロトタイプにおいて400のエッセイ
  - 6点満点中2点以上の異なった予測は全体の10%
  - 従来の専門家による一致率とほぼ同じ
  - E-raterの専門家との代替の妥当性

PEG (1994)

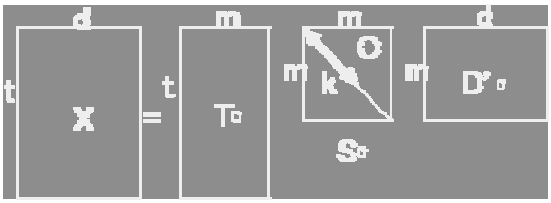
・作文品質をより直接的に測定できるよう改良
・“文章のつながり易さを測定するなど、より複雑で豊かな変数の採用とその重み付けがされている”
・変数については未公開

Intelligent Essay Assessor, IEA(1999)

・Latent Semantic Indexingによる意味的な内容の一致

Latent Semantic Indexing (LSI)

- TREC (Text REtrieval Conference)でその有用性が主張
- 特異値分解



$\hat{X} = TSD',\ k = 50 \sim 100$ → 次元低減

LSIによる文書間の類似度

- 採点される小論文e: t 次元の単語ベクトルx_eで表現できる
- 文書空間Dの行に対応する $1 \times k$ の文書ベクトル

$$d_e = x_e' TS^{-1}$$

- 出題文qについても同様のd_q
- 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コサイン

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

Intelligent Essay Assessor, IEA(1999)

- その後、改良
  - 内容、文体、構成の3つの観点から評価
- 15の話題について3,296編のエッセイを評価
  - 専門家同士の採点の相関0.86
  - IEAと専門家との相関0.85

自動採点システムに対する批判 (1/3)

- コンピュータはテキストを正確に理解することができない
- 適切なキーワードや同義語を用いて出題文に答えたとしても、これが必ずしも包括的に適切な答えになっているとは限らない
- 「アメリカ女王は1492隻の船でサンタマリアへ航海した。彼女の夫、コロンブス王は、インディアンの探険家ニーナ・ピンタがイザベラ海岸に巨大な富を持っていることを知っていたが、フェルナンド大陸から香辛料を獲得することを我慢せざるを得なかった。」→多くの適切なキーワード
- 望ましい答えに似た文章を書いた場合に同じ問題
- 防護策として人間と機械との併用

自動採点システムに対する批判 (2/3)

- 各出題文に対するモデルをセットアップするために多大な労力
- 自動採点システムの多くは重回帰モデルを使用
- 事前に多くの変量に係る重みを設定しておく必要
- 大規模テストの利用に限られている
- コーパスベースのシステムはこの問題を回避できる可能性

自動採点システムに対する批判 (3/3)

- 解答に正解が書かれているかについても十分な評価を行うべきである。→適切ではない（Shermis, 2002）
- 多くの作文教師は修辞の側面を重視
  - 論理的な接続表現が用いられているか
  - 話の筋が通っているか
- 答えが正しいことが重要ならテストの様式はより効果的な別の形

現在

- ルール発見アルゴリズムに基づくIntelliMetric(2003)
- ベイズ理論を取り入れたBETSY(2002)
- 日本語小論文を処理するJess(2003)
- エッセイ評価システムの比較

IntelliMetric

- Vantage Learning社が開発、販売
- 1997年7月:ペンシルバニア州の司法試験の採点を実施
- 1998年2月:世界で初めてインターネット上で論述式問題に対する自動採点を実施
- 開発までに11億円(10 million dollars)以上

IntelliMetricの技術的な特徴

- Vantage Learning社曰く「先進的な人工知能を有した」
- 「ルール発見」を採点に用いている
  - 最初に予め採点が終っている、スコアが出ている模範解答を「学習」
  - 各採点ポイントのデータを蓄積
  - 人間の採点者の採点ルールの判断を推定

IntelliMetricによる評価の観点

- 文献により多少の違いがあり
- Focus & Meaning: 主題に対してどの程度, 一貫性があるか.
- Development & Content: 内容の幅や発想の展開
- Organization: 論旨の展開など文章構成
- Language Use & Style: 文章の複雑さ, 多様性
- Mechanics & Conventions: アメリカ英語に対する適合度

IntelliMetricの評価法

- 各観点に対して通常1～6点のスコア
- それをもとに全体の評点(6点満点)
- 各観点に対して1～4点のスコア, 満点が4点のバーションもあり(ペンシルバニア州の基準に基づく)
- 各観点に対するスコアは 72種類の素性(Features)により計算される
- これらの素性は各観点に排他的に分類されるのではない

IntelliMetricの短所

- 良い採点を行うために, 事前に 良質の採点付き学習データを多数用意しておく必要
- 2000年の時点で49個の素性を決めるのに, 300個の人間による採点データが必要(フィラデルフィア・ビジネス・ジャーナル)
- 現在の版では素性の数は72と更に増えているから, より多くの採点データが必要
- 課題の数が限られていて, 多くの採点を行う場合には, 採点付き学習データを多数用意することがコスト的に割に合うが,
- 多種類少数の採点には割に合わない

IntelliMetricの短所(続き)

- 極めて注意深く書かれたいわゆる良いエッセイを正当に評価しない
- 2001年のポスト・ガゼット誌の例
  - 教育担当記者(Eleanor Chute)が 自分の書いたエッセイを IntelliMetricで評価
  - 6点満点中4点
  - 推敲を重ねても向上せず
- 主任責任者の Dr. Scott Elliottによれば, 3% から7% の論文は 類別することが 困難
- 同じ評点に 同じコメント

BETSY

- メリーランド大学のRudnerらによって開発
- エッセイ評価分類(4ないし6段階)にベイジアンアプローチ
- 性能
  - 2点法で採点した462の学習データ
  - 80編のエッセイ(各スコアに対して40編ずつ)
  - 特定の単語, フレーズ, 論理展開の有無などの特徴量に基づき 分類
  - 80編中64編(80%)が正しく判定
- 最初のパラグラフで 分野を判定

ベイズ流のエッセイ採点 (1/3)

- 適切(Appropriate), 部分的に適切(Partial), 不適切(Inappropriate) の3つのいずれかに 分類
- 着目する特徴量が含まれている確率

$$P_2(u_i = 1|A),\ P_2(u_i = 1|R),\ P_2(u_i = 1|I)$$

i:特徴量の識別子 , u_i :エッセイがその特徴量を含んでいるか否か

適切: $P_2(u_i = 1|A) = 0.7$, 部分的: $P_2(u_i = 1|R) = 0.6$, 不適切: $P_2(u_i = 1|I) = 0.1$

先験情報が与えられていないとき

$$P(A) = P(R) = P(I) = 0.33$$

ベイズ流のエッセイ採点(2/3)

u_i=1でそのエッセイが適切であるとする事後確率

$$P(A|u_i = 1) = P(u_i = 1|A) * P(A)/P(u_i = 1)$$

このとき

$$P(A|u_i = 1) = 0.7 \times 0.33/P(u_i = 1) = 0.233/P(u_i = 1)$$
$$P(R|u_i = 1) = 0.6 \times 0.33/P(u_i = 1) = 0.200/P(u_i = 1)$$
$$P(I|u_i = 1) = 0.1 \times 0.33/P(u_i = 1) = 0.033/P(u_i = 1)$$

$$P(A|u_i = 1) = 0.233/(0.233 + 0.200 + 0.033) = 0.5$$
$$P(R|u_i = 1) = 0.200/(0.233 + 0.200 + 0.033) = 0.429$$
$$P(I|u_i = 1) = 0.033/(0.233 + 0.200 + 0.033) = 0.071$$

ベイズ流のエッセイ採点(3/3)

- これら事後確率を新しい事前確率
- 次の特徴量に対して P(A), P(R), P(I)を更新
- 全ての特徴量に対して繰り返す
- より一般的には2つのベイジアンモデル
  - 多変量Bernoulliモデル→特徴量がエッセイに含まれているか否か
  - multinomialモデル→エッセイに含まれる特徴量が何回出現したか
- McCallum & Nigam, 1998

Jess

- 他の既存のシステムがプロの評価者(rater)を手本にしているのに対し、唯一、プロのライター(writer)の書いた文章を手本にしている
- 毎日新聞における社説とコラム(余録)を学習
- 理想とする文章の書き方についての特徴量の分布を予め獲得
- 得られた特徴量が理想とする分布において外れ値となった場合に減点

エッセイ評価システムの比較

評価システム	評価基準	手法	制限
E-rater	構造/組織化/内容	重回帰モデル	“tricked”の批判
PEG	内容/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
IEA	内容/文体/技巧	LSI	論理構成/語の出現順を評価しない
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
BETSY	表層	ベイズ的接近	分野が制限;開発中
Jess	修辞/論理構成/内容	外れ値検出&LSI	科学技術分野に弱い

未来

- 自動採点システムに望まれる要件
- 日本語固有の問題点

自動採点システムに望まれる要件(1/3)

- 人間の評定に頼りすぎない
- 人間の評価者は学生のエッセイの中に混入させたプロのエッセイを特別に高く評価できない(Friedman,1985)
- プロの評価者(rater)ではなくプロのライターを使う→Jessで実現

自動採点システムに望まれる要件(2/3)

- 対話的なフィードバックを返す作文ツール
- 単純な文法エラー検出はあたりまえ
  - “I concentrates”, “this conclusions”など
- 「汚れ(pollution)」と呼ばれる文法エラー検出
- 助詞の誤り/脱落の例
  - 「東京で行く」→「東京へ行く」,「計算機(を)扱う」
- 悪文の例
  - 「～しないと～しない．」(二重否定),「背の高い社長の椅子」(曖昧な修飾関係)

自動採点システムに望まれる要件(3/3)

- 内容レベルでの誤りの指摘
  - 実在しない固有名詞(「中僧根元首相」→「中曽根元首相」)
  - 矛盾する数値(「第五四半期」)
  - 文意の矛盾(「定率法と低額法」→「定額法」)
  - 文意の誤り

日本語固有の問題点(1/3)

- 分量の問題
  - アメリカの公的試験におけるエッセイ試験では字数制限がない
  - 日本では、600字あるいは800字の字数制限
  - 作文量についての指標が使えない

日本語固有の問題点(2/3)

・順接表現の省略
- 日本語では、順接表現は意識的に避けられる
- 手がかり語に頼らない文章の構成および展開の把握
- 文書要約の最新技術が利用できる？

日本語固有の問題点(3/3)

・機種依存文字の問題
- キーボード入力が可能となった場合であっても残る問題
- 利用者は必ずしも漢字コードに詳しくはない
- 機種依存文字（システム外字）を意識せずに使う可能性。例えば①②③
- ユーザは箇条書きで分かりやすく表現したつもりがシステムはこれを評価しない

Jess

- わが国における小論文採点の制限
- 要素技術
- 詳細
- 課題
- デモ

わが国における小論文採点の制限

- e-raterにおける採点の仕方
- 専門家によって採点された膨大な数の小論文の蓄積
- 専門家の得点とコンピュータによる得点とを線形回帰
- 得点のためのメトリクスにかかる回帰係数を決定
- わが国では同じようなアプローチは事実上，不可能

自然言語処理ツールの整備

- 形態素解析
  - 京都大学 言語メディア研究室の JUMAN
  - 奈良先端松本研の茶筌（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」
- 構文解析
  - 京都大学のKNP
  - 奈良先端のSAX, BUP，南瓜
  - 東工大 田中・徳永研究室の MSLRパーザ

模範と考えられる小論文の電子媒体での入手

- 「毎日新聞」の2006年までの全記事
- 「日本経済新聞」の2006年までの全記事
- 著作権の切れた文学作品（青空文庫）

内容の適切さの評価

- 書かれた内容が質問文に十分に応えた内容であるか
- パターン・マッチ（文字列一致）に拠らない
- Webにおけるサーチ・エンジン等で用いられている意味的検索（石岡・亀田,1999）

→高速化のための実装上の工夫

Jessの基本的なアイデア

- 模範となるエッセイやコラムの学習
- 外れ値検出
- 欧米の既存システムと同等のことを
- 技術的に，より優れた方法を用いて開発できる
- 少規模採点向き

採点基準

- e-raterの構造，組織，内容をほぼそのまま踏襲
  1. 修辞
  2. 論理構成
  3. 内容
- それぞれの観点に係る重み（配点）はユーザが指定
- ユーザが特に指定しなければ5,2,3（合計10点）、渡部(1988)

修辞を示すメトリクス

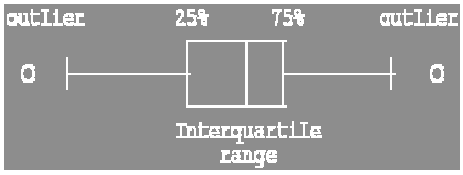
1. 文章の読みやすさ
   a. 文の長さの中央値，最大値
   b. 句の長さの中央値，最大値
   c. 句中における文節数の中央値，最大値
   d. 漢字/カナの割合
   e. 連体修飾（埋め込み文）の数
   f. 連用形や接続助詞の句の並びの最大値
2. 語彙の多様性; Yuleの*K*
3. ビッグ・ワード(big word, 長くて難しい語）の割合
4. 受動態の文の割合

各種メトリクスの統計量の分布

- 毎日新聞CD-ROM中の社説/コラムについて得た
- ほとんどは左右非対象の歪んだ分布
- この分布を理想とする小論文についての分布とみなす

外れ値検出アプローチ

- 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，割り当てられた配点を減じる
- 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



論理構成:議論の流れをつかむ

- さまざまな主張のつながり具合を把握すること
- 議論の接続を示す接続表現をしばしば使用する
- 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

「順接」を示す接続表現

- **付加:** 主張を加える接続関係
  - 「そして」,「しかも」,「むしろ」
- **解説:**
  - 「すなわち」,「つまり」,「言い換えれば」,「要約すれば」
- **論証:** 理由と帰結の関係を示す
  - 理由:「なぜなら」,「その理由は」
  - 帰結:「それゆえ」,「したがって」,「だから」,「つまり」
- **例示:** 具体例による解説/論証;「たとえば」

「逆接」を示す接続表現

- **転換:**
  - 「AだがB」,「A, しかしB」
- **制限:**
  - いわゆる「ただし書き」;「ただし」,「もっとも」
- **譲歩:**
  - 「たしかに」,「もちろん」
- **対比:**
  - 「一方」,「他方」,「それに対して」

接続表現の個数

- 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す
- 順接, 逆接各4通り, 計8通りに排他的に分類
- 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
- 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
- 「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

接続表現の出現パターン

- 社説に比べて特異でないかを判断
- 順接と逆接の出現パターンについてのトライグラムモデル(北,1999)
- 「順接」および「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(有限マルコフ過程)
- トライグラムモデルに従うときの ある出現パターン に対する生起確率が, 事前情報がないときの生起確率に比べ小さいならば、その出現パターンは特異であると判断

接続表現トライグラムの例

- パターン {*a, b, a, a*} の生起確率 p = 0.44 x 0.42 x 0.55 x 0.28 = 0.035.
- 事前情報無しの {*a*} の生起確率0.47; {*b*} の生起確率0.53
- 事前情報無しの順接3回、逆説1回の生起確率q = 0.47^3 x 0.53 = 0.055.
- この出現パターンは特異;$p < q$
- 議論の接続に割り当てられた配点を減ずる

内容

- Latent Semantic Indexing (LSI)
- SVDPACKC (Michel Berry)
- 行列Xの特異値問題は以下の対称行列の固有値問題と同じ

$$\begin{pmatrix} 0 & X \\ X' & 0 \end{pmatrix} \qquad X'X$$

- 部分空間法、トレース最小法、ランチョス法、ブロックランチョス法による比較（石岡・亀田1999)

実施例

http://www.etstechnologies/com/html/eraterdemo.html

採点結果の比較

Essay	E-rater	Jess	文字数	応答時間（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 (3.6)	504	0.95

- e-raterが良い得点を与える小論文には Jessも良い得点を与える
- 得点もかなり一致している

専門家との比較（480編）

Jessスコア、専門家による平均スコア、言語理解テストの相関

	Jess	専門家平均スコア
専門家平均スコア	0.57	
言語理解テスト	0.08	0.13

- 相関0.57は、専門家同士の相関0.48よりも大きい
- 言語理解テストとの相関はともに小さい
- 言語理解テストは別の学力を測っている？

大学生143 編のデータ結果

- 国立国語研究所で収集したデータ
  - 似た調査結果:
- 「喫煙」について
  - Jessと専門家との相関 0.83 > 専門家同士の相関(0.73)
- 「日本の祭り」について
  - Jessと専門家との相関 0.84 > 専門家同士の相関(0.73)

デモ

- Web版
- Windows closed版
  - 大量処理用

Jessの課題

- 分野による使用辞書の切り替え
- 手がかり語（接続表現）によらない接続関係の把握
- 日本語では接続表現は意識的に避けられる
  - 指示語に注目
  - 省略時は順接
  - 接続関係の図式表現
  - 分量に依存

マスコミ紹介

- 2005年2月朝日新聞夕刊1面トップ
- ニッポン放送
- アサヒパソコン、コンピュータ・ピープル欄
- 2006年Yahoo! Internet Guide 6月号「インターネットでできること300」
- 2006年6月商標登録
- 2007年2月韓国KBSテレビ

謝辞

- 亀田雅之（株リコー、共同研究者）
- 井上達紀（早稲田大学、Windows版移植）
- 生田和重（徳島文理大学、評価）
- 鷺坂由紀子（リクルートマネージメント、評価）
- 宇佐美洋（国立国語研究所、データ提供）
- 村木英治（東北大学、前ETS)
- 企画セッション講演者、等々

ご清聴ありがとうございました

http://coca.rd.dnc.ac.jp/jess/