特　集「テキストの自動評価」

小論文およびエッセイの自動評価採点における研究動向

Latest Trends in Automated Essay Scoring and Evaluation

石岡　恒憲　　独立行政法人 大学入試センター
Tsunenori Ishioka　The National Center for University Entrance Examinations
tunenori@rd.dnc.ac.jp, http://www.rd.dnc.ac.jp/~tunenori/

Keywords: nation-wide common test, natural language processing, visualization.

1. はじめに

小論文やエッセイの自動評価および採点についての研究は，教育測定の分野において最近 15 年の間で最も精力的に行われてきた研究の一つである．その背景としては，従来，知識工学的なアプローチの多かった自然言語処理の分野に 1990 年代頃からコーパスを用いた確率・統計的なアプローチが成功を収め，その有効性が多くの研究者や技術者に広く認知されてきたことがあげられる．成功例のアプリケーションとして代表的なものだけでも機械翻訳，音声認識，情報検索，自動要約などがある．これら研究のための道具立てもかなり整備され，自然言語であるところの小論文やエッセイの評価および採点に，最新の研究成果を活用しようとする試みは極めて自然な流れであるといえよう．

実際，少なくない数のシステムが 90 年代後半から開発され，実用化された．現在，商用とされるシステムにはアメリカ最大のテスト教育機関である Educational Testing Service：ETS が開発した e-rater[Burstein 98, Burstein 03]，パイオニアである Project Essay Grade：PEG[Page 66, Page 94]，意味的な内容の一致を測定する Intelligent Essay Assessor：IEA[Foltz 99, Landauer 00, Landauer 03], ルール発見アルゴリズムに基づく IntelliMetric[Elliot 99, Elliot 03a] の四つがある．2000 年頃から，これらシステムのエッセイ評価マシンとしての妥当性が証明されるにつれ，いくつかのシステムは，アメリカの入試における共通テストとしての公的試験に実際に用いられるようになった．E-rater は，1999 年からアメリカの経営大学院（いわゆるビジネススクール）入学のための共通試験である Graduate Management Admission Test：GMAT における作文試験 Analytical Writing Assessment：AWA の採点に使われた．2006 年 1 月より GMAT の開発および運営は ETS から Peason VUE & ACT, Inc. に移り，これに伴い AWA エッセイの採点は Vantage Learning 社が開発した IntelliMetric が行うようになった．IntelliMetric はまた，アメリカの医学大学院進学のための適性試験 Medical College Admission Test：MCAT の作文試験の採点に用いられることが 2007 年 1 月に決定した．これは，アメリカ医科大学協会（Association of American Medical Colleges：AAMC）がプロメトリック社と契約を締結し，従来の紙と鉛筆による試験からコンピュータを利用した試験（Computer-Based Testing：CBT）に変更することによる；これに伴い IntelliMetric 採点エンジンの導入（2007 年開始）が CBT の包括的なサービスを AAMC に提供する上で不可欠であるという判断がされた．

また商用のシステムではないが，メリーランド大学の Rudner らのグループによってベイズ理論を採り入れた BETSY [Rudner 02] が開発された．日本語を処理するシステムとしては，著者らのグループが最初でかつおそらく現時点で唯一のシステムである Jess[石岡 03, Ishioka 06] を開発した．

コンピュータによるエッセイの自動採点および評価は，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われたかにより評定が変わる），課題選択（異なる課題に基づいて書かれた小論文をどう一元的に評価するか；どのように等化をするか）などの問題を排除できるだけでなく，採点の手間を大幅に低減し，また対話的な作文指導ができるといった点で，極めて有効であると考えられている．近年では説明責任といった点からも重要である．

小論文，およびエッセイの自動採点システムのついての過去の歴史から現状を包括する日本語による初のサーベイ論文として，[石岡 04] がある．これは 21 ページからなる論文としてはかなり大部なもので，各システムの採点のロジック，いわば中身がわかるよう記述されている．また自動採点システムに今後求められる要件や，現状での問題点を含む，いわば近未来についても言及してある．

そこで本稿では，[石岡 04] 以降の近過去の話題をより詳しく入れて各種システムの紹介を行う．例えば，

e-raterは2004年にVer. 2.0が出され，以前のVer. 1.3に比べ，採点モデルを一新した．またいくつかのシステムでは対話的なフィードバックを返すよう，そのシステムの機能を拡張した．さらに本稿では，[石岡 04]で述べた自動採点システムに求められる要件が，どのように改良されたかを報告する．

2章では，アメリカで商用されている四つのシステム（e-rater, PEG, IEA, IntelliMetric）に加え，研究ベースのシステムとしてBETSYを取り上げ，計五つのシステムについて解説する．3章では，日本語を処理する現時点で唯一のシステムであるJessについて述べ，現在取り組んでいる課題について報告する．その課題とは，手掛かり語（cue words）に頼らずに論文全体の論理構造を把握するための工夫であり，その結果の一部を紹介する．4章はまとめであり，小論文の自動採点システムが今後取り組む方向性について改めて言及する．

2. 英文における自動採点システム

2・1 Electric Essay Rater：e-rater

E-raterは世界最大のテスト機関であるEducational Testing Service：ETSのBurstein らの研究グループによって開発されたシステムである[Burstein 98]．Ver. 1では約60の言語上の特徴量（feature）のスコアに基づいて，重回帰モデルによって採点を行う．ただしすべての変量（特徴量）が用いられるわけでなく，論題（prompt）によって8～12の変量が選ばれる．[Kukich 00]によると，変量の組合せの異なる75のモデルが用いられているという．

アメリカの公的試験では，エッセイのスコアは推定の標準誤差が1点となるように6点法で採点されることが多いが，専門家とe-raterによる採点の一致率（1点差以内）は，97%である[Burstein 03]．

2004年には新バージョン（Ver. 2.0）が開発された[Attali 05, Burstein 04]．この新バージョンでは，用いられる変数の数は12で，論題によらずに固定である．特徴量自体も，良い作文（good writing）を示す性質とより関係がつくように改良された．

これらの特徴量は以下の五つのカテゴリーに分類できる．（1）～（5）がカテゴリーで，1.～12.がそのカテゴリーに含まれる特徴量である．

（1）文法，語の使用法，手順（mechanics），スタイル
1. 総ワード数に対する文法エラーの割合
2. 総ワード数に対する語の使用法についてのエラーの割合
3. 総ワード数に対する手順のエラーの割合
4. 総ワード数に対するスタイルについてのエラーの割合

（2）組織化と掘下げ
5. 談話（discourse）ユニットの数

談話ユニットには，論点（thesis），中心となる話題（main point），それを指示する話題（supporting ideas），結論（conclusion）があるが，これらの談話ユニットのカウントには以下のような制限がある．

- supporting ideasユニットはmain pointユニットの後ろにあるときのみカウントされる．
- main pointユニットは一つのエッセイの中で多くとも三つまでである．

この考えは，Criterion（e-rater Ver. 1の時代からあったエッセイの作文指導のためのソフトウェア）で採用された良いエッセイを書くための「5パラグラフ戦略」に基づいている．この戦略によれば，初心者は典型的には，序文（introductory paragraph），三つの本文（main pointとそれを指示するsupporting ideasからなる），と結論から構成するとよい，とするものである．背景（background）の要素は，「5パラグラフ戦略」における序文には含まれないことに注意されたい．

したがって，e-rater 2.0では，このようにして得た談話ユニットの上限は8であり，実際の数から8を減じた数をスコアとして与える．すべての要素（thesis, conclusion, 三つのmain points, それに対応する三つのsupporting ideas）が備わったとき，スコアは0となる．

6. 各ユニットにおける平均のワード数

（3）トピック分析
7. 当該エッセイの6点法によるコサイン類似度が最大となるスコア点
8. 最高点（通常6点）を得たエッセイとのコサイン類似度；文書間の近さには[Salton 75]にあるtf・idf法が用いられる．

（4）単語の複雑度（complexity）
9. 単語の繰返しの程度を示す指標：全ワード数（token）に対する異なったワード種類（word type）の割合

例えば，“This essay is a long, long, long essay”．という文に対しては，五つのword type（this, essay, is, a, long）と八つのtoken（this, essay, is, a, long, long, long, essay）があり，その比率type/tokenは5/8 = 0.625となる．

10. [Breland 94]らの単語頻度指標に基づく語彙の困難度（difficulty）
11. 平均の単語長さ

（5）エッセイの長さ
12. 単語の総数

これら12変量に係る重み付けは経験則によって定められる．唯一の例外はエッセイの長さについてであり，この要因のスコアへの影響が過大に評価されないよう，これは0に設定される場合がある．TOEFLエッセイについては，それぞれの重みは順に0.05, 0.02, 0.07, 0.08, 0.21, 0.12, 0.04, 0.07, 0.08, 0.03, 0.03, 0.20であるとして

いる [Attali 05].

E-rater Ver. 2.0 のモデルは論題（prompt）によらずに自然でより単一の基準に基づこうとするものである．このモデルの最も重要な点は，特徴量の数が 12 と少ないことにある．この重みの値を前もって知らせることによって，ユーザは容易にスコアを推測することができるようになる．作文能力測定にかかわる理論的研究が進めば，これらの重みを今後，調整することも可能であり，このことも重要である．この改良は，e-rater（Ver.1）が従来“tricked”（トリックが使われている）という批判に対するものであり，スコアに対する説明責任を果たしているといえる．

2･2 Project Essay Grade：PEG

PEG はエッセイ評価のために開発された最初のシステムであり，Ellis Page によって最初のバージョンが 1966 年頃に開発された [Page 66]．このバージョンでは，proxes と呼ばれる約 30 の特徴量が用いられ，これらを trins と呼ばれる本来測定しようとする作文能力を表す指標の代用とした．

多くの特徴量は，数値化された作文上の表層的なもの，例えば平均の文の長さ，パラグラフの数，句の数などである．これら特徴量に係る重みを計算するために，重回帰モデルが用いられている．

PEG は 1993 年に改訂され，文法チェッカと談話分析器（part-of-speech tagger）などの自然言語処理ツールが使われるようになった [Page 94, Page 95, Page 03]. この版では trins をより正確に把握するために，“より豊かで複雑な特徴量の採用と重み付け”がされたとしている．典型的なスコアモデルには 30 〜 40 の特徴量が使われている．

最新版 [Shermis 02] では, PEG は総合点に加え, 内容, 組織化，スタイル，メカニクス，独創性などの項目別のスコアを提供している．ほかのシステムにない革新的なこととしては，生徒の（作文上の）長所と短所について，より詳細なフィードバックを返すようにしている．

しかしながら，PEG の項目別スコアや総合スコアを構成するために，どのような特徴量が用いられているかの詳細については公開されていない．

2･3 IntelliMetric

IntelliMetric は Vantage Learning 社によって，エッセイや自己完結型（open-ended）問題の採点のために開発された．IntelliMetric は，開発者サイドが自称するところの知能に基づいた（brain-based あるいは mind-based）モデルに基づいて情報処理理解を行っている [Elliot 03b]．技術的な背後にあるのは，人工知能，ニューラルネット，計算機言語学であるとしている．

与えられた論題（prompt）に対して，IntelliMetric は生徒の回答から 400 もの特徴量を抽出し，スコア推定に有効な特徴量を抽出し，スコアモデルに係る重みを推定する [Elliot 03b].

400 もの特徴量は，談話･修辞，内容･概念，意味･構造，メカニクスのクラスに分類することができるとしているが，それぞれのクラスに属する特徴量が何であるかについては公開されていない．

IntelliMetric による評価スコアの観点は，文献によって多少の違いがあり，また用いられているワーディングも一貫していないが，おおむね以下の五つである．

- Forms and unity: 目的や主題（メインアイディア）に対しての結束性や一貫性
- Development and elaboration: 内容の幅や発想の展開
- Organization and structure: 論旨の展開や文章構成
- Sentence structure: 文の完全性や多様性
- Mechanics and convention: 英語のルールへの適合

上記評価スコア観点への特徴量クラスへの対応は，すべての特徴量クラスがすべての評価スコア観点に影響する多対多の関係である [Elliot 03a].

2･4 Intelligent Essay Assessor：IEA

IEA はコロラド大学の Landauer や Foltz らの研究グループが開発し，現在，その開発は彼らが立ち上げたベンチャー企業である Knowledge Analysis Technologies（KAT）社に移管されている．KAT 社は IEA のコアとなっている KAT エンジンを，全米でも有数のテスト機関である Peason Education 社の傘下である Pearson Knowledge Technologies（PKT）社に提供し，PKT 社が IEA を販売，提供している．

IEA の特徴は主に論文の内容の評価に重きを置いているところにある．知識獲得と表現についてよく組織化された理論によってシステムがつくられている，としている．

そのアプローチは，Latent Semantic Analysis（LSA）に大きく依存している．LSA は，情報検索の分野からきた数学的な方法である．この方法で置いている仮定は，与えられた文書やテキストの潜在的な意味構造が，単語の共起を通して，コアの意味，すなわちテキストの内容を規定する代表的な行列（特異値行列）によって捉えることができるとするものである．

この方法では，内容の関係のあるテキスト（例えば特定目的の本）のさまざまな集合からつくられる情報が縮約され，単語と文書の関係を明示的に示した意味空間として定義される行列で表現される．この行列の単語 - 文書の関係は数値（重み）であり，因子分析における因子負荷量に似たものである．

エッセイ採点に用いた場合は，今，対象としているエッセイの内容が意味的空間において，ほかの採点されたエッセイにどの程度近いのかを知ることができるようになる．

IEA は総合スコアに加えて，通常三つの観点からユー

ザにスコアを提供する．

（1）内容：LSAから生成された二つの特徴量である文章品質とドメインとの関連性

（2）文体：首尾一貫性と文法

（3）技巧：句法，スペル

IEAの総合スコアは，各観点ごとの人間スコアとの回帰モデルによって計算される．IEAはもともと内容の評価を行うことを意図して設計されたが，今では作文スキルを評価することにも利用することができる．この点において，IEAのアプローチは，書かれている内容そのものと，その議論の展開の両方を正しく評価するものである [Landauer 00]．

2・5 Beyesian Essay Test Scoring sYstem：BETSY

BETSYはメリーランド大学のRudnerらのグループによって開発されたシステムで，エッセイ評価分類にベイジアンアプローチが取られていることに最大の特徴がある [Rudner 02]．エッセイの評点は，通常，4段階から6段階で評定されるので，これらの段階へのクラス分けとして考えることができる．

分類方法として二つのベイジアンモデルが用いられている [McCallum 98]．

一つは多変量Bernoulliモデルで，特徴量w_tがスコアc_jの文書中に少なくとも1回現れる確率を用いて，エッセイd_iが分類スコアc_jを受け取る確率を計算するものである．分類スコアc_jの確率の最も大きいものが，そのエッセイの得点として最も確からしいと判定する．

もう一つのモデルはmultinomialモデルで，特徴量w_tがエッセイiに何回現れているかを示す変数を用いて，特徴量w_tがスコアc_jであるエッセイに含まれる確率を学習データから獲得し，エッセイd_iが分類スコアc_jを受け取る確率を計算する．このモデルは“ユニグラム（unigram）言語モデル”と呼ばれており，テキスト分類に [Mitchell 97] などが適用したものである．

BETSYの適用できる範囲は限られているが，ソースコードは，研究目的であれば公式サイト [BETSY] からフリーでダウンロードできる．BETSYはVisual Basicに移植されており，間もなくソースコードが公開される予定である [Dikli 06]．

2・6 エッセイ自動採点システムに望まれる要件

著者は [石岡 04] において，今後のエッセイ自動採点システムに望まれる要件として

（1）システムの性能を評価する唯一の基準として人間の評定に必要以上に頼らないこと

（2）単なるスコアを返すだけでなく，対話的なフィードバックを返すための作文ツールとして精緻化させること

の2点を指摘した．本項では，[石岡 04] 以後，ここで紹介したシステムが上記をどの程度，反映したのか，について述べることにしたい．

（1）の問題を初めて指摘したのは [Bennet 98] である．欧米語に対するほとんどすべてのシステムは，基本的にプロによる人間の採点に機械の採点を限りなく近づけることを目的としている．しかしながら人間の評定は，典型的には評価基準表（rubric）に基づいているのであり，評価基準表の利用こそが，ユーザに受容可能（acceptable）と考える信頼性を確保するための手段となるべきはずのものである．したがって，評価基準表が論題（prompt）ごとに変わることは，（わずかな修正は許されるにせよ）通常はなく，論題ごとに重回帰モデルにおける重み係数が変わるようなモデルは本来奇妙なモデルということがいえる．その意味で，PEGやIntelliMetric, e-rater（Ver.1）は，必要以上に人間の評価者に頼りすぎているといってよい．

一方，改良されたe-rater Ver. 2.0は（そして後述するJessも），論題によらず評価モデルは一定であり，いわゆる評価基準表に従った採点を行っている．このような論題によらないモデルは，当然のことながら，論題ごとにモデルの重み付けや採点ルールの学習が不要であるから，モデルをセットアップするための準備が不要となり，即座の試験問題の採点に利用できるだけでなく，多種多様な論題に対して作文訓練を行うことができる．このことは（2）においても有利に働くことになる．

また，人間による採点には，いわゆるハロー効果のために，良い（あるいは悪い）印象がほかのすべての評価観点に良い（あるいは悪い）評価を与えることがある．事実，Fridmanが1980年代に行った研究によれば，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手によるエッセイを特別に高く評価することができなかった．人間の評価は，必ずしも絶対視すべきものではなく，e-raterの今回の改革は，著者の考え，および著者らのシステムの方向性の正しさを示したものといえよう．

（2）の観点，すなわち各システムが対話的なフィードバックを返す作文ツールとしての位置づけを高めることを意識していることは，もはや疑いがない．[石岡 04] の時点では，Critiqueというe-raterの機能の一部を実装した作文分析ツール（Critique Writing Analysis Tool）が唯一，提供されるに過ぎなかった．

Critiqueは，文法，語の使用法，技巧，文体，組織化，展開などに対するリアルタイムのフィードバックを返すものであるが，現在は，ユーザの作文レベルに応じて返すコメントを変えるように改良されている．作文レベルには，小学生（4～5年生），中学生（6～8年生），高校生（9～12年生），カレッジ（1～2年生），上級・大学院受験（GRE相当），非英語圏対象の英語（TOEFL相当）の各レベルがある．あるトピックにつき最低465編の採点・学習の後，提供されるとしている．

IEAでは，現在，各観点ごとの評点に加えて“Tools”という項目があり，コピー（copy; 剽窃を検出する），ス

ペル（spelling），冗長性（redundancy），文法（grammar）についてコメントを出すようになっている．IntelliMetric ではその機能の一部に基づいた My Access! という Web ベースの作文評価ツールが提供されるようになった．My Access! も e-rater 同様，レベルに応じたフィードバックを返すようになっており，高校最終年（K-12）レベルを標準（proficient）とし，初心者（developing），上級者（advanced）の合わせて三つのレベルがある．加えて，My Access! では多国語化を図っており，現在，英語，スペイン語，中国語を取り扱うことができる．ユーザには二つの選択肢があり，一つは英語，スペイン語，中国語のいずれかの言語でエッセイを書き，同じ言語でフォードバックを受け取る方法，もう一つは英語でエッセイを書き，英語あるいは母国語でフィードバックを受け取る方法である．

取り扱うエッセイにはさまざまなジャンルがあり，報知的なもの（informative），事実に基づく物語（narrative），文学，エッセイ（persuasive essay）など，現在 200 以上の論題が用意されている．

2･7　エッセイ自動採点の妥当性の研究

[Yang 02] によれば，エッセイ自動採点の妥当性に対する検討のアプローチは以下の三つに分類できるという．

（1）異なった評価者，例えば人間と機械によるスコアに焦点を当てたもの

（2）テストスコアと writing についてのほかの指標との関係に焦点を当てたもの

（3）評価プロセスに焦点を当てたもの

多くの欧米のシステムの設計思想は，人間のスコアにより近づけようとするものであり，したがって，機械と人間の間には高い一致が報告されている [Burstein 98, Elliot 99, Landauer 03, Page 95]．機械と人間の評価の一致は望ましく，かつ不可欠であるが，その品質基準（quality criterion）は必ずしも十分ではない [Bennet 06]．不幸なことに外的基準（[Yang 02] では second category と呼ばれているもの）については，ほとんど共通なものがない．利用可能な外的基準としては，多肢選択テスト，writing についての授業の成績，教師による生徒の writing スキルに対する評価，生徒による writing スキルに対する自己評価などがある．これらの分析のほとんどは，それが不完全なものとしても，その一致率はかなり低い（例えば [Elliot 01, Landauer 01] など）．

[Yang 02] による 3 番目のカテゴリー，すなわち評価プロセスについての妥当性の研究については極めて少ない．どの自動採点システムでも，独自のモデルに経験や理論に基づく実験を加えて意味のあるモデルを構築しているわけだが，Yang らは採点プロセスを評価するのに，記述的で質の高いアプローチを取ることが重要であると指摘している．具体的には人間と機械の不一致の同定と，そのパターンの分析が必要であり，そのことがどのような writing 特徴量を選択すべきかの決定と，またその重み付けの改良に役立つとしている．欠けている特徴量の同定も重要な課題であるとしている．[Attali 05] はこのカテゴリーに属する数少ない論文の一つである．

3. Jess

3･1　特　　徴

Jess のシステムとしての最大の特徴は，ほか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，このシステムは唯一，プロのライター（writer）の書いた文章を手本にしているところにある．このため採点モデルを論題ごとにセットアップする必要がなく，従来大規模な試験にのみしか実用的でなかったこの分野において，初めて小規模な試験での運用を可能とした．Jess では，模範と考えられる小論文やエッセイとしてある全国誌の新聞における社説とコラム（余録）を学習し，理想とする文章の書き方についてのメトリクスの分布をあらかじめ獲得しておく．これらメトリクスの分布のほとんどは左右非対象のゆがんだ分布となるが，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 倍を超えるデータとする．

Jess は採点基準については，アメリカの経営大学院（いわゆるビジネススクール）への入学試験である GMAT における AWA の採点基準 [GMA-Council 05] をほぼ踏襲しており，以下の三つの観点から評価を行う．

（1）修辞：文章としてよく書けているか．

（2）論理構成：アイディアが理路整然と表現されているか．議論が深められているか．

（3）内容：出題文に適切に応えているか．

このうち，（1）修辞については，文章の読みやすさ（文の長さ，句の長さ，句の数，埋込み文の存在など），語彙の多様性，ビッグワード（big word，長くて難しい語）の割合，受動態の文の割合など，良い文としての指標が比較的明確で，またそれについての一般的な合意が得られている．コンピュータで評価するのに最も適した観点ということがいえる．また（3）内容については，本質的には困難であるが，その代替として，Latent Semantic Indexing[Deerwester 90] などの手法を採り入れ，いわゆる意味的な内容の一致を測定することとしている．これはほかの多くのシステムでも同様であり，現時点での技術レベルでの限界を示しているといえよう．

残る（2）論理構成については，欧米語に対する多くのシステムでは，接続表現などの手掛かり語（cue words）に多く頼っている [Page 94, Burstein 98, Rudner 02].

このためJessでも

●順接／逆接を示す手掛かり語の出現の頻度と

●手掛かり語の出現パターンが特異か否か

について評価，および判断を行っている．出現パターンの判断については，順接／逆接の接続表現の出現にトライグラムモデルを用い，事前情報なしの場合の生起確率が，（新聞のコラムや社説であらかじめ獲得した）事前情報ありの生起確率に比べ大きいならば特異であると判断する．しかしながら，日本語の場合は，接続表現は意識的に避けられる傾向にある[野矢 97]．実際，この省略が独特のリズムを生み，美文または名文となるからである．事実，毎日新聞のコラム（余録）でも，1年365編のうち，平均で20編は，接続表現の全くない文章である．

もっとも，接続表現の出現個数は，文章全体の分量に大きく依存する．コラム（余録）では分量が700字であるために，接続表現なしに一つの話題で一気に最後までもっていくことが可能であるが，社説のように1200字ともなると，ある程度の論理構造が必要となり，接続表現なしに書き進めることはできなくなる．しかし不幸なことに，我が国の大学入試や大学院入試で用いられる小論文試験では，その分量はわずか600～800字程度である．このため，この程度の分量の評価を行うためには，接続表現だけに頼らない方法で，全体の論理構造の把握を行う方法を構築することが必要となる．次節以降では，そのための工夫について述べる．

3・2 手掛かり語に頼らない論理構造の把握

接続表現だけに頼らない論理構造の把握が必要とはいえ，やはり接続表現は論理構成の把握に極めて有効である．自動要約の分野でも，要約文の生成に，すなわち文章構造の把握のために手掛かり語に多く頼ってきた[Mani 01, Marcu 00]．ここでいう手掛かり語とは，主に順接や逆接の接続表現のことを指す．[野矢 97]によれば，順接の論理として以下の4通り存在するという．

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．ほかにも「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「例えば」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．

逆接の論理としても以下の4通りが存在しているとしている．

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のほう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のほう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：転換の一種と見ることもできるが，譲歩の場合は対話的構造が現れる．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．

前述の接続表現が用いられたときは，確かにそれに対応する論理を構成するであろう．しかし，上記の接続表現がない場合でも，接続の論理は起こり得る．それは明示的な接続表現が省略された場合であって，そのことを検出／判断する方法として以下の二つが考えられる．

一つ目は前の段落を受ける指示代名詞を検出することである．著者らは形態素解析として茶筌を用いているが，段落の最初の文において最初の句（文頭から最初の句点との間）に指示代名詞が存在したときに，前の段落あるいは前の段落中のなにかを示す指示代名詞であると判断する．

もう一つの方法は文末モダリティ分析である．モダリティ（modality）とは，時制や態などとは違って，文が指す内容に対する話し手の判断や心的態度をいう．「～すべきである」や「～と見られる」などがそれにあたる．

日本語の場合はモダリティが主に文末に位置するため，その識別は比較的容易である．本稿における実験では，各文の終端から文頭方向に10文字を切り取り，予め登録してあった文末パターンのデータベースと照合することでモダリティの識別を行っている．文末モダリティの分類体系の構築にあたっては，[日本語記述文法研究会 03]の体系を部分的に修正して採用することができる．

例えば順接を示す文末モダリティには以下があげられる．

付加：「～（は｜も）そうである」

解説：「～といえよう」，「～とまとめられる」，「～と要約できる」

論証：「～からである」，「（などの）理由による」，「～だと考えられる」

例示：「あげられる」，「列挙できる」

3・3 論理のつながりを示すポイントの付与

議論の流れをつかむとは，さまざまな主張のつながり具合を把握することにほかならないから，これら順接・逆接の論理がどのような順番でつながっているかを見れば，その論理展開を把握することができるであろう．

詳細については省略するが，順接の接続関係には，以下の強弱関係があるといえる．

付加＜解説＜例示＜論証

一方，逆接については論証の強弱関係は以下のように

なると考えられる．

制限＜対比＜（譲歩＝転換）

いま，順接を＋のポイント，逆接を－のポイントとし，前項で述べた順接／逆接の論証の強さに応じて，＋4 から－4 のポイントを付与することを考える．

便宜的であるが，順接については，付加（＋1），解説（＋2），例示（＋3），論証（＋4）とし，逆接については，制限（－2），対比（－3），譲歩（－4），転換（－4）を与える．

これにより，全体の論理のつながりを把握することができる．また，話題の転換の程度もわかるようになる．その際に，単に（＋4）～（－4）のポイントを付与するだけでなく，そのポイントに対応する段落の文字数（談話量）を考慮することが重要であろう．

このようにして得た論理構成を数値化すれば，例えば統計学で用いられる星座グラフ [Wakimoto 78] などで表現することができ，議論の展開を視覚化できる．ポイント化せずとも，接続の論理の種類はたかだか八つであるから，これらを直接的に扱い，異常とみなされる議論の展開を検出することも可能であろう．

著者らは 1999 ～ 2002 年，2006 年の毎日新聞 CD-ROM を用いて，そこに収められているコラムおよび社説に適用した．その結果，付加の論理により論証を進めていく文章が大半であり，それにいくつかの逆接の論証を交える書き方が多いことがわかっている．

4. おわりに

エッセイの自動採点システムが作文支援ツールとしての機能をより精緻化していくことが研究の方向性であることは，もはや疑いがない．現在，自動採点システムが指摘する項目は，単純な文法エラーか，構文解析で見つけることができるような誤りに限られている．ETS の研究グループは，中心化理論におけるラフ・シフトの検出や，「汚れ（pollution）」と呼ばれる語彙上の文法エラーを，ワード並びの N グラムモデルに基づいて発見しようという研究 [Kukich 00] を進めているようである．

「汚れ」の例としては前置詞の誤り／脱落や一般にいわれる悪文などがあげられる．今後は，当然の流れとして内容レベルでの誤りの指摘が求められるであろう．具体例としては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」），文意の誤りなどをあげることができる．これらは構文解析では解決することができず，文脈や一般常識を用いた解析により誤りと断定できるものである．

もう一つの方向性は，cheating と呼ばれるいわゆるズルをすることの対処である．剽窃や機械の裏をかくようなエッセイの検出が求められている．IntelliMetric では，論題と関係のない解答や，論題そのものの複写，および論題の半分の複写とオリジナルとの併用などの例については検出することを可能にしている．E-rater では，システムの評点が人間の評点より高くなる場合の論文について個別に分析を行い，良く書かれたパラグラフを何回も繰り返す方法や，接続表現の多用する方法で機械を欺くことができる場合のあることを報告している [Powers 01]．自動採点システムがテスト採点のツールとして利用される以上，cheating への対応についての研究は永久に続く課題なのかもしれない．

謝　辞

本研究については科学研究費補助金, 基盤研究（B）,「日本語小論文の自動評価における総合的研究」（課題番号: 19300292，研究代表：石岡恒憲 ）の援助を受けた．

◇ 参 考 文 献 ◇

[Attali 05] Attali, Y. and Burstein, J.: Automated essay scoring with e-rater v.2.0（ETS RR-04-45）, Princeton, NJ: Educational Testing Service（2005）

[BETSY] Beyesian Essay Test Scoring sYstem, BETSY, http://edres.org/betsy/

[Bennet 98] Bennet, R. E. and Bejar, I. I.: Validity and automated scoring: It's not only the scoring, *Educational Measurement: Issues and Practice*, Vol. 17, No. 4, pp. 9-17,（1998）

[Bennet 06] Bennett, R. E.: Moving the field forward: Some thoughts on validity and automated scoring, pp. 403-412, D. M. Williamson, R. J. Mislevy, I. I. Bejar（Eds.）*Automated Scoring of Complex Tasks in Computer-Based Testing*, Lawrence Erlbaum Associates（2006）

[Breland 94] Breland, M. H., Jones, J. R. and Jenkins, L. The college board vocabulary study, College Board Report no. 94-4（1994）

[Burstein 98] Burstein, J., Kukich, K., Wol., S., Lu, C., Chodorow, M., Braden-Harder, L. and Harris, M. D.: Automated scoring using a hybrid feature identification technique, *Proc. Annual Meeting of the Association of Computational Linguistics*, Montreal, Canada, http://www.ets.org/research/erater.html（1998）

[Burstein 03] Burstein, J. and Wolska, M.: Toward evaluation of writing style: Finding overly repetitive word use in student essays, *Proc. 11th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary（2003）

[Burstein 04] Burstein, J., Chodorow, M. and Leacock, C.: Automated essay evaluation: the Criterion online writing service, *AI Magazine*, Vol. 25, No. 3, pp. 27-36（2004）

[Deerwester 90] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 7, pp. 391-407（1990）

[Dikli 06] Dikli, S.: An overview of automated scoring of essays, *Journal of Technology, Learning and Assessment,* Vol. 5, No. 1（2006）

[Elliot 99] Elliot, S.: Construct validity of IntelliMetric with international assessment, Yardley, PA: Vantage Technologies（RB-323）（1999）

[Elliot 01] Elliot, S.: IntelliMetric: From here to validity, *Annual Meeting of the American Educational Research Association*（2001）

[Elliot 03a] Elliot, S.: IntelliMetric: From Here to Validity, Shermis, M. & Burstein, J. eds., *Automated Essay Scoring: A Crossdisciplinary Perspective*, pp. 71-86, Hillsdale, NJ:

Lawrence Erlbaum Associates（2003）
[Elliot 03b] Elliot, S.: How does IntelliMetric score essay responses?, *RB-929*, Newtown, PA: Vantage Learning（2003）
[Foltz 99] Foltz, P. W., Laham, D., and Landauer, T. K.: Automated essay scoring: applications to educational technology, *Proc. EdMedia'99*（1999）
[GMA-Council 05] Graduate Management Admission Council, The Official Guide for GMAT Review, 11th Edition（2005）
[石岡 03] 石岡恒憲，亀田雅之：コンピュータによる小論文の自動採点システムJessの試作，計算機統計学，Vol. 16, No. 1，pp. 3-18（2003）
[石岡 04] 石岡恒憲：記述式テストにおける自動採点システムの最新動向，行動計量学，Vol. 31, No. 2, pp. 67-87（2004）
[Ishioka 06] Ishioka, T. and Kameda, M.: Automated Japanese essay scoring system based on articles written by experts, *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 233-240, Available online: http://www.aclweb.org/anthology/P/P06/P06-1030（2006）
[Kukich 00] Kukich, K.: Beyond Automated Essay Scoring, The Debate on Automated Essay Grading, *IEEE Intelligent Systems*, Vol. 15, No. 5, pp. 22-27（2000）
[Landauer 00] Landauer, T. K., Laham, D. and Foltz, P. W.: The Intelligent essay assessor, the debate on automated essay grading, *IEEE Intelligent Systems*, Vol. 15, No. 5, pp. 27-31（2003）
[Landauer 01] Landauer, T. K., Laham, D. and Foltz, P. W.: The intelligent essay assessor: Putting knowledge to the test, *Emerging Technologies and Opportunities for Diverse Applications Conference*, Tucson, AZ（2001）
[Landauer 03] Landauer, T. K., Laham, D. and Foltz, P. W.: Automated scoring and annotation of essays with the intelligent essay assessor, Shermis, M. & Burstein, J. eds. *Automated Essay Scoring: A Crossdisciplinary Perspective*, pp. 87-112, Hillsdale, NJ: Lawrence Erlbaum Associates（2003）
[Mani 01] Mani, I.: *Automatic Summarization*, Natural Language Processing 3, John Benjamins Publishing Co.（June 2001）
[Marcu 00] Marcu, D.: *The Theory and Practice of Discourse Parsing and Summarization*, The MIT Press（Nov. 2000）
[McCallum 98] McCallum, A. and Nigam, K.: A comparison of event models for Naive Bayes text classification, *AAA-98 Workshop on "Learning for Text Categorization"*（1998）
[Mitchell 97] Mitchell, T.: *Machine Learning*, WCB/McGraw-Hill（1997）
[日本語記述文法研究会 03] 日本語記述文法研究会編：現代日本語文法4. 第8部モダリティー，くろしお出版（2003）
[野矢 97] 野矢茂樹：論理トレーニング，哲学教科書シリーズ，産業図書（1997）
[Page 66] Page, E. B.: The imminence of grading essays by computer, *Phi Delta Kappan*, pp. 238-243（1966）
[Page 94] Page, E. B.: New computer grading of student prose, using modern concepts and software, *Journal of Experimental Education*, Vol. 62, No. 2, pp. 127-142（1994）
[Page 95] Page, E. B. and Petersen, N. S. Computer assisted instruction; Computer simulation, *Phi Delta Kappan*, Vol. 76, No. 7, pp. 561-65（1995）
[Page 03] Page, E. B.: Project essay grade: PEG, In Mark D. Shermis & Jill C. Burstein,（Eds.）, *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pp. 43-54, NJ:Erlbaum, Mahwah（2003）
[Powers 01] Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. Stumping e-rater: Challenging the validity of automated essay scoring, GRE Board Professional Rep. No. 98-08bP, ETS RR-01-03（2001）
[Rudner 02] Rudner, L. M. and Liang, L.: Automated essay scoring using Bayes' theorem, *National Council on Measurement in Education*, New Orleans, LA. Available online: http://ericae.net/betsy/papers/n2002e.pdf,（2002）
[Salton 75] Salton, G., Wong, A. and Yang, C. S.: A vector space model for automatic indexing, *Commun. ACM*, Vol. 18, No. 11, pp. 613-620（1975）
[Shermis 02] Shermis, M. D., Koch, C. M., Page, E., Keith, T.Z., and Harrington, S.: Trait rating for automated essay grading, *Educational and Psychological Measurement*, Vol. 62, No. 1, pp. 5-18（2002）
[Wakimoto 78] Wakimoto, K. and Taguri, M.: Constellation graphical method for representing multi dimensional data, *Ann. Statist. Math.*, Vol. 30, Part A, pp. 77-84（1978）
[Yang 02] Yang, Y., Buckendahl, C. W.: A review of strategies for validating computer-automated scoring, *Applied Measurement in Education*, Vol. 15, No. 4, pp. 391-412（2002）

2007年11月9日 受理

著者紹介

石岡 恒憲

1959年生まれ．1983年東京理科大学大学院工学研究科経営工学専攻修士課程修了．同年株式会社リコー（ソフトウェア研究所）入社．1998年文部省大学入試センター研究開発部助教授．組織改編に伴い現在独立行政法人大学入試センター准教授．統計学，信頼性工学，情報数理に関する研究に従事．工学博士．IEEE Trans. on Reliability論文審査委員．日本信頼性学会（論文審査委員），応用統計学会（編集委員），言語処理学会，ACL各会員．Marquis Who's Who in the World, 25th Edition, 2008.