Vol.2011-NL-202 No.3
2011/7/15

多様な教育的観点を考慮した機械学習による日本語文章の評価と評価モデルの顕在化

藤田　彬[†]　藤田　央[†]　田村直良[††]

本稿では，文章に対する評点と教育的観点についての特徴量から，個々の評価者の文章評価モデルを学習する手法について述べる．また学習した文章評価モデルにおける観点毎の配分を顕在化する手法について述べる．評価モデルの学習には SVR を用いる．SVR の教師データには，「表層」「語」「文体」「係り受け」「文章のまとまり」「モダリティ」「内容」というカテゴリに分けられる様々な素性を用意する．これらには日本の国語科教育において扱われる作文の良悪基準に関わる素性が多く含まれる．なおかつ，全ての素性が評価対象文章に設定される論題の語彙的ドメインに依存しない汎用的なものである．本手法により，文章の総合的な自動評価，個々の評価者が着目する言語的要素の明示，さらに評点決定に寄与する各要素の重みの定量化が可能になる．

Automated Evaluation of Japanese Compositions based on Educational Points of View, and Manifestation of Individual Evaluation Model

Akira FUJITA[†]　Hiroshi FUJITA[†]
and Naoyoshi TAMURA[††]

We propose a method to learn an individual model which is to evaluate Japanese Compositions via Support Vector Regression, according to features based on educational points of view and scores marked by human in advance. Also, we propose a method to manifest breakdown of the model. Features in training data of SVR are categorized as 7 types according to what each features refers to. The features include some features regarding criterions of Japanese compositions in education. Besides, all the features do not depend on lexical domain of a composition's prompt. Our methods make it possible to score an integrated point of a composition automatically, and also to account elements considered by individual evaluator, to quantify weights of the each elements which contributes decision of scores.

1. はじめに

近年，学習者により書かれた文章を教育的な目的で自動評価する技術に対して，需要が高まっている．大学入試や就職試験等の大規模な学力試験において課される小論文試験の採点や，e-learning 等の電子的な学習システムにおいて学習者の能力を測るために出題される記述式テストの採点が例として挙げられる．このような，多数の文章を同一基準の下で迅速に評価する必要のあるタスクにおいて，対象となる全ての文章を人手で評価することは，多くの場合困難を伴う．

第一に，掛かる時間と労力が問題となる．短文を用いた記述式回答や小論文等の作文の評価は，選択式問題の回答の評価に比べて，評価者が捉えるべき情報と考慮すべき基準が多く，またそれらの情報や基準自体も複雑なものとなる．

第二に，評価基準の安定性が問題となる．文章の良悪を決定する基準は，評価者個々において完全に固定的なものではない．評価する順序による系列的効果や，ハロー効果[15]の影響も考えられる．このことから評価者間での基準の共有も難しいといえる．

図 1 に，複数人による同一文章の評価結果の例を示す．これは高校生により書かれた 584 編の小論文を，4 人の国語教育専門家が評価した際の，評点当たりの事例数とその分布や一致具合に関する諸々のデータである．評価は特定の観点に沿ったものではなく総合的なもので，10 段階の絶対評価により施される．評点は 10 点が最高点，1 点が最低点である．

	平均	標準偏差
評価者 A	5.420	1.081
評価者 B	5.438	1.037
評価者 C	6.639	2.092
評価者 D	5.856	1.436

	κ 係数		κ 係数
A-B	0.110	B-C	0.034
A-C	0.035	B-D	0.115
A-D	0.093	C-D	0.081

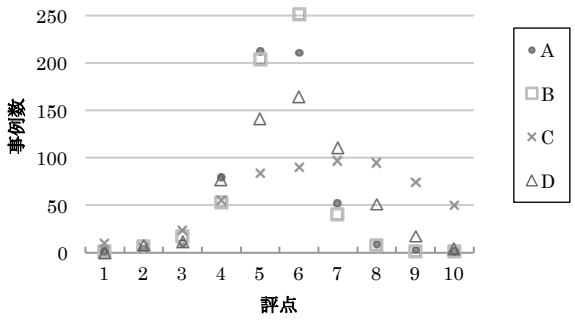

図 1 : (左上)各評価者の評点分布の平均と標準偏差，(左下)評価者間での評点の κ 係数
(右)各評価者の評点当たりの事例数

[†] 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences, Yokohama National University
[††] 横浜国立大学大学院環境情報研究院
Graduate School of Environment and Information Sciences, Yokohama National University

Krippendorff[11]の考察によれば，あるデータ間の κ 係数が 0.7 未満の場合，両データの関連を示すことは困難であることが多いとされる．図 1 において評点の κ 係数は全評価者間で 0.7 を下回っている．また，評点の分布にも異なりが認められる．このことから，本例において評価者の評価基準が共通のものであるとはいい難い．

このように評価者間で評価結果に差異が生じる要因の一つに，「個々の評価者が着目する言語的要素」と「評点決定に寄与する各要素の配分（重み）」の違いが挙げられる．これらの違いが定量的に示されることは，評価者の評価基準の統合の助けになる．

本稿では，評価対象文章に対する総合的な評点と，多種の観点についての特徴量から，個々の評価者の評価モデルを推定する手法について述べる．また，その評価モデルにおける観点毎の配分を顕在化する手法について述べる．

提案手法は，点数付けを順序付き多クラス分類として捉え，Support Vector Regression(SVR) [1]を用いて評価者が付けうる評点を予測する．SVR の教師データには，表層や使用語彙，構文，文章構造などの特徴に関する様々な素性を用意する．これらの素性には，日本の国語科教育において扱われる作文の良悪基準に関わる素性が多く含まれる．なおかつ，全ての素性が評価対象文章に設定される論題の語彙的ドメインに依存しない汎用的なものである．

ただし，文章の意味的な適切さの評価に関しては扱わない．ここでいう意味的な適切さとは，文章中の文が示す個々の内容の正しさを指す．例えば，「*月は西から昇る*」のような文が示す内容の不正確さについて，本研究では言及をしない．

2. 関連研究

自動的に文章の評価を行なうためのモデルを得る先行研究には，二通りのアプローチが存在する．一つは，評価者による文章のスコアをラベル，文章上の素性を事例として，教師付き学習により単一のスコア推定モデルを求めるものである[3)4)5)6)7)12)13)14)18)19)]．もう一方は，模範と考えられる文章上の素性値を基準として，その基準との距離を用いてスコア推定モデルを求めるものである[9)29)]．

e-rater[19)]は，12 の固定的な素性[a]を説明変数，評価者によるスコアを従属変数として重回帰分析を行い，得られた回帰式をスコア推定モデルとする．しかし，この手法では，多くの要素をまとめた素性を用いるため，得られるスコア推定モデルから詳細な言語的要素に関する評価基準の検討を行なうことができない． また，英文を対象にしており，日本語の文章を対象にする際は，扱う素性の調整を別途行なう必要がある．

Jess[9)]は，あらかじめ 3 種の観点（修辞，論理構成，内容）に沿って模範となる文章（新聞の社説やコラム）における種々の素性値の分布を獲得し，理想的な分布とする．評価対象文章の各素性値が模範文章における素性値分布の四分位数範囲の 1.5 倍を超える場合，外れ値とみなしてそれぞれについて点数を減ずる．しかしこの手法では，模範文章の選択の妥当性について，評価が行われる背景（試験の目的等）毎に検証が必要である．またその検証自体も，実用的に難しいと考えられる．

なおこれらの関連研究に関しては石岡によるサーベイ[30)]が詳しい．

[a] 変数選択を行うことがなく，常に 12 の素性を説明変数とする．

3. 評価モデルの学習

3.1 自動評価タスクの定式化と学習の手法

自動評価タスクにおいては，評点の各値がそれぞれ別のクラスであり，かつクラス間に順序関係をもつ．そこで本研究では，自動評価タスクを順序付き多クラス分類として定式化し，Support Vector Regression(SVR)を用いて評点の予測を行う．

順序付き多クラス分類問題を SVR を用いて解く先行研究には，岡野原ら[20)]の研究がある．岡野原らは，評判分類タスクを順序付き多クラス分類問題として定式化した上でレビューが評価対象に与える評価の度合を二極指標(実数値)で表す手法を提案している．この中で岡野原らは，(順序関係を考慮しない）多クラス分類問題を解く分類器である pairwise Support Vector Machine(pSVM)[16)]と SVR の間で，順序付き多クラス分類問題に対する適合性を比較している．その結果，SVR がより高い精度で分類を行うことが示されている．pSVM は，予測クラスを間違えた際のペナルティに全クラス間で差が無いため，SVR に比べて分類モデルに大きな誤差を含む可能性が高い．

3.2 用いる素性

評価者は，様々な観点から文章を捉えて評価を行なった上で，その判断結果を点数化する．本研究ではこの過程を教育的立場からモデル化する．

教育上，考慮されうる観点を羅列し，それぞれの観点を特徴付ける値を求める．得られる数値は，SVR の素性として評価モデルの訓練に用いられる．以下，提案手法において素性として用いる言語的要素をカテゴリ毎に列挙した上で，一部の素性についてその詳細を述べる．

3.2.1 カテゴリ「表層」

文字数や文数，字種などの表層的特徴に関する素性を表 1 に列挙する．主に，文章の形式面の妥当性についての評価に役立つ．

FV2：文字数制限には，(Ⅰ)「～字以上」，(Ⅱ)「～字以内」，(Ⅲ)「～字程度」の 3 種類が存在する．制限（指定）文字数を r, 評価対象文章の文字数を n としたとき，下記のように達成度 d を算出することにする．ただし，ⅠⅡについては制限が守られなかった場合，達成度を 0 とする．

$$(\mathrm{I}): d = 1 \quad (\mathrm{II}): d = \frac{n}{r} \quad (\mathrm{III}): d = \frac{n}{r}\ (n \le r),\ d = \frac{2r-n}{r}\ (n > r)$$

FV7：文は長くなるほど，内部の係り受け関係に曖昧さを生じやすい [28].

3.2.2 カテゴリ「語」

単語（特に自立語）の用法や品詞，記法に関する素性を表 2 に列挙する．

FW2：自立語の異なり数(タイプ数)を延べ数(トークン数)で割った値とする．ただし用言については活用形を計数の考慮に入れず，異なり数，延べ数ともにその用言の原形を数える．

FW4：オノマトペは副詞に分類される．しかし，論説文におけるオノマトペの使用は特に着目されることが多いと考えられるため，副詞とは別途に扱う．

3.2.3 カテゴリ「文体」

文末の形式や文内で用いられる文体等に関する素性を表 3 に列挙する．

FF4：述語が“名詞句＋断定の助動詞(「だ」「である」「です」等)”で構成される文の出現数を，文の総数で割った値とする．

FF7：文中の助動詞に着目し，常体／敬体の混用が認められる場合は０，一貫している場合は１を素性値とする．文章中における常体／敬体の混用は，それを避けることが現行の小学校学習指導要領 [33]において指導事項として定められている．

FF8：式 $-\log\frac{n}{N}$ により算出する．ただし，nは文の最終文節の表記異なり数，Nは文の総数とする．

3.2.4 カテゴリ「係り受け」

同一表層格の多用や文節間の修飾関係の複雑さを捉える素性を表 4 に列挙する．これらの素性は，主に係り受けの適切さの評価に役立つ．

FD6：格助詞「が」を付属語に持つ文節が文中に出現する回数の最大値を素性値とする．文中で同じ助詞を繰り返し使用することで，文の表す内容が不明瞭になる場合がある [22].

例：*英語の早期教育がもたらす効果が測れないうちに実行に移すことが問題であることは，言うまでもない.*

FD14,FD15:一文中における名詞文節の出現回数を用言の出現回数で割った値を全文について平均した値，またその分散を素性値とする．

FD16：一文中で，用言の連用形や接続助詞等によって文が途中で中止される回数の最大値を素性値とする．中止法の多用は，係り受け関係を曖昧にする原因となりうる [22].

表 1：カテゴリ「表層」の素性群

ID	素性名	ID	素性名
FV1	文字数	FV9	カタカナ使用率
FV2	文字数制限の達成度	FV10	漢字使用率
FV3	文数	FV11	記号使用率
FV4	文の文字数の平均	FV12	英字使用率
FV5	文の文字数の分散	FV13	算用数字使用率
FV6	文の最小文字数	FV14	文当たりの読点数の平均
FV7	文の最大文字数	FV15	形式段落当たりの文数の平均
FV8	ひらがな使用率		

表 2：カテゴリ「語」の素性群

ID	素性名	ID	素性名
FW1	自立語の最大長	FW6	複合名詞の使用率[b]
FW2	自立語の多様性	FW7	カタカナ語の使用率[b]
FW3	副詞使用率[b]	FW8	自立語の文字数の平均
FW4	オノマトペ使用率[b]	FW9	自立語内のひらがな使用率
FW5	感動詞，間投詞の使用率[b]	FW10	数量表現における漢・算用数字混用の有無

表 3：カテゴリ「文体」の素性群

ID	素性名	ID	素性名
FF1	体言止めで終わる文の出現率	FF6	括弧（）による補足を含む文の出現率
FF2	体言が存在しない文の出現率	FF7	常体・敬体の混用の有無
FF3	用言が存在しない文の出現率	FF8	文末の単調度
FF4	名詞述語文の出現率	FF9	感嘆符が用いられる文の出現率
FF5	鉤括弧「」による引用を含む文の出現率	FF10	口語的表現の使用有無

3.2.5 カテゴリ「文章のまとまり」

文章のまとまりに関する性質として，**テキスト一貫性** [31] と**テキスト結束性** [8]が挙げられる．テキスト一貫性は，概念や事象の間の意味的なつながりの良さを指す．テキスト一貫性は，隣り合う二文間における一貫性を示す**局所的な一貫性**と，文章全体での話題遷移の一貫性を示す**大域的な一貫性**に区別できる．一方テキスト結束性

[b] これらの素性の素性値は，当該語の延べ数を全自立語の数で割った値とする．このとき複合名詞は一つの自立語として数える．

は，意味的なつながりではなく，文法的なつながりの良さを指す．Barzilay ら [2]は，局所的な一貫性のモデルとして「entity grid モデル」[c]を提案している．横野ら [20]は，結束性に寄与する要素[d]を entity grid モデルに組み込むことで，結束性と局所的一貫性を同時に捉えるモデルを提案している．

横野らは，結束性を考慮する目的で既存手法に下記の手法を組み込んでいる．

i. 語句要素の文間遷移確率の計算に文間の接続関係の考慮を加え，接続関係の種類別に遷移確率を計算する．
ii. 意味的な類似性に基づいて語句要素のクラスタリングを行なう．
iii. 参照表現が正しく機能している割合を素性として導入する．

また，Barzilay らの entity grid において扱われる構文役割は 4 種類(S:主語，O:目的語，X:その他，−:出現せず)であるが，横野らは構文役割の体系を日本語に特化する目的で，２種類の構文役割(H:主題，R:述部要素)を加えている．

本研究では，文章のまとまりについての特徴を捉える目的に，横野らのモデルに基づいた素性を用いる(表 5). ただし, FC122 のみ筆者らが独自に設定する素性である．接続関係の分類・同定方法，語句要素が持つ構文役割の同定方法，参照表現が機能している割合の導出方法は，それぞれ横野らの方法に従う．類似性に基づいた語句要素のクラスタリングについては, EDR 電子化辞書 [32]の日本語単語辞書内で同一の概念識別子を持つ語句要素を同じクラスタとして扱う．クラスタの持つ構文役割は，H>S>O>R>X という優先順位(cf.文献 [17])に基づいて，クラスタ毎に一つの構文役割を決定する．また，考慮する遷移確率は文 2-gram のみとする．

FC122：並列的に展開して述べる接続関係が文章中に存在する場合に 1，存在しない場合に 0 を素性値とする．並列的な接続を明示する特定の表現（人手で設定）の有無により決定する．

3.2.6 カテゴリ「モダリティ」

松吉ら [24]は，情報発信者の主観的な態度(モダリティ)に真偽判断や価値判断などの情報を統合した「拡張モダリティ」を提案し，体系化している．拡張モダリティは主に「態度表明者」,「相対時」,「仮想」,「態度」,「真偽判断」,「価値判断」の 6 項目から成る(文献 [26]を参考にした)．このうち態度[e]と真偽判断について文の特徴を捉え，素性として扱う(表 6).

松吉らは態度を 8 種類に, 真偽判断を 9 種類に分類している. この分類に基づいて，機能表現辞書つつじ [27]に助動詞型機能表現として収録される機能表現と，分類語彙表 [25]に「精神および行為／心」として収録される用言を手がかりにしたルールベース手法で，文の態度，真偽判断について分類を行なう．ただし，真偽判断については，肯

表 4：カテゴリ「係り受け」の素性群

ID	素性名	ID	素性名
FD1	格助詞「ノ」で終わる文節の最大連続数	FD9	文当たりの連体修飾回数の最大値
FD2	格助詞「ト」で終わる文節の最大連続数	FD10	文当たりの連体修飾回数の最小値
FD3	格助詞「ニ」で終わる文節の最大連続数	FD11	文当たりの連体修飾回数の平均
FD4	格助詞「ヲ」で終わる文節の最大連続数	FD12	文当たりの連体修飾回数の分散
FD5	係助詞「モ」で終わる文節の最大連続数	FD13	全文節中に占める名詞文節の割合
FD6	格助詞「ガ」で終わる文節の文内最大出現数	FD14	修飾の詳細さ（平均）
FD7	係助詞「ハ」で終わる文節の文内最大出現数	FD15	修飾の詳細さ（分散）
FD8	一文中での「ハ・モ・ガ」の最大出現数	FD16	一文中での中止法の最大使用回数

表 5：カテゴリ「文章のまとまり」の素性群

ID	素性名
FC1~FC36	論理的結合関係で接続する文間の構文役割遷移確率
FC37~FC72	多角的連続関係で接続する文間の構文役割遷移確率
FC73~FC108	拡充的合成関係で接続する文間の構文役割遷移確率
FC109~FC114	文章始めと第一文の間の構文役割遷移確率
FC115~FC120	最終文と文章終わりの間の構文役割遷移確率
FC121	参照表現が正しく機能している割合
FC122	並列接続の有無

表 6：カテゴリ「モダリティ」の素性群

ID	素性名	ID	素性名
FM1	態度「叙述」出現率	FM7	態度「許可」出現率
FM2	態度「意志」出現率	FM8	態度「問いかけ」出現率
FM3	態度「欲求」出現率	FM9	真偽判断の程度が断定的な文の出現率
FM4	態度「働きかけ-直接」出現率	FM10	真偽判断の程度が推量的な文の出現率
FM5	態度「働きかけ-間接」出現率	FM11	真偽判断の程度が不明な文の出現率
FM6	態度「働きかけ-勧誘」出現率	FM12	文末思考知覚感覚動詞使用率

[c] 文を行，文章中の語句要素を列，文における語句要素の構文役割を成分とする行列を用いて，語句要素の分布パターンを表現するモデル．行列から構文役割の遷移確率と構文役割の出現確率を成分とするベクトルを導出し，局所的一貫性の評価等に用いる．
[d] Halliday ら 8)は参照，接続，語彙的結束性，省略を挙げている．
[e] 言語学における「表現類型のモダリティ」23)に相当する．

否極性とアスペクトに関する区別をせずに判断の程度(強さ)のみを扱うこととし，断定，推量，判断程度不明の 3 種類を扱う．これら拡張モダリティ体系に準拠する素性(FM1～FM11)は全て，当該拡張モダリティカテゴリに分類される文が出現する回数を文の総数で割った値とする．

FM12：最終文節で思考動詞，知覚・感覚動詞が用いられる文の出現回数を文の総数で割った値を素性値とする．思考動詞，知覚・感覚動詞であるか否かの判断は，分類語彙表 [25]を典拠とする．

3.2.7 カテゴリ「内容」

文章の内容(筆者により書かれた行動，出来事，状態)について，その正しさを意味面で判断することは，本研究の目的としない．その代わりに，与えられた論題に対して適合した語彙が使用されていることを捉えるための素性を用意する．論題に含まれる名詞と文章中の名詞が，EDR 電子化辞書 [32]において同一の概念識別子を持つ，もしくは所属概念が直接の上位または下位関係にある場合，論題に適合する語彙として判断する．このように判断される語彙が文章中の全名詞中で占める割合を素性(FS1)とする．

4. 評価モデルの顕在化

SVR は，線形カーネルを使用して学習する場合，回帰係数wの成分値(以下，成分値)を参照することで，各素性がスコア推定モデルに寄与する度合を知ることができる．これにより，教師データにラベル(評点)をつけた評価者が，各素性に対して「どの程度の配分で評価するか」，また「加点要素とするか減点要素とするか」を定量化することができる．前者は成分値の絶対値，後者は成分値の正負に着目することでそれぞれ明らかになる．この方法により個々の評価者の評価モデルを顕在化する．

ただし，これらの特徴を素性間で比較する場合，成分値の大きさと素性値の大きさに従属性があるため，素性値の分布幅が正規化されている必要がある．しかし，提案する素性の素性値は，その分布幅が素性により異なる．そこで，成分値を素性間で比較するために，下記の 2 通りの方法で素性値を正規化する．

i. $$x_{regularized} = (x_{original} - min\,X)/(max\,X - min\,X)$$

ii. $$x_{regularized} = (x_{original} - Q_1)/(Q_3 - Q_1)$$

$x_{original}$は任意の文章データにおける任意の素性の素性値，Xは全教師データにおける任意の素性の素性値の集合，Q_1は集合X中の第 1 四分位値，Q_3は集合X中の第 3 四分位値である．

i は，全教師データ中の素性値が 0 から 1 の間に分布するように正規化する方法である．一方 ii は，全教師データの四分位数範囲に位置する素性値が 0 から 1 の間に分布するように正規化する方法である．i の方法を使って正規化した場合，ある素性に限って教師データ中に外れ値が含まれるとき，その素性は他の素性と比較して素性値分布の偏差が異なることになる．これに対して，外れ値を除外して正規化を行なうことを目的とする方法が ii である．

5. 実験・考察

5.1 設定

提案手法の評価のために実験を行なう．SVR の学習には 3 章に挙げた素性群と *SVM*^{light}[f]を用いる．また，素性の抽出には形態素解析器 MeCab[g]，係り受け解析器 CaboCha[h]を用いる．

教師データには第 1 章で用いた高校生による小論文を電子化したものを用いる．これらの小論文は，(論題 A)「小学校の授業における，英語の早期教育は必要であるか否かに対して意見を述べよ」，(論題 B)「グラフと説明文を読み，日本人の子育ての態度に関してどのような特色が読み取れるかに関して述べよ」という 2 種類の論題に沿って書かれている．また，400 字以内と 800 字以内の 2 種類の字数制限が存在する．事例は合計で 584 事例あり，論題 A を 400 字以内で記述するものが 153 事例，論題 B を 400 字以内で記述するものが 140 事例，論題 A を 800 字以内で記述するものが 147 事例，論題 B を 800 字以内で記述するものが 144 事例という内訳である．

これらの 584 事例に対して 4 人の評価者が総合的につけた 10 段階の評点を，各教師データのラベルとする．

5.2 実験

実験1. 評点推定の性能

教師データを用いて SVR を構築し，各評価者がつけた評点と SVR による評点推定結果の間の差について検討する．SVR による評点推定の評価指標には MAE(Mean Absolute Error)[i]を用いる．また素性値の正規化に，4 章で述べた二通り(i , ii)の方法を適用し，それぞれを教師データに用いて構築した SVR の MAE を比較する．SVR 構築には全て，線形カーネル，コストパラメータ C=10 を用いる．表 7 に 5 分割交差検定の結果を示す．なお以降の実験では素性値の正規化に ii を用いる．

[f] http://svmlight.joachims.org/
[g] http://mecab.sourceforge.net/
[h] http://sourceforge.net/projects/cabocha/
[i] $\sum_{i=1}^{n} \frac{|y_i - f_i|}{n}$ y_iを評価者による評点，f_iを SVR により推定される評点，nを事例数とする．

表 7：SVR の評点推定性能(MAE)
縦方向は正規化手法，横方向は教師データのラベル(評点)をつけた評価者名

	i	ii
評価者 A	0.746	0.732
評価者 B	0.814	0.743
評価者 C	1.677	1.598
評価者 D	1.070	1.050

表 8：評価者別の回帰係数wの成分値（絶対値が大きい上位 5 素性）
素性 ID と数値は対応する素性と回帰係数における成分値，{＋,−}は成分値の正負

	順位	評価者 A		評価者 B		評価者 C		評価者 D	
＋	1	FV2	0.111	FV2	0.084	FV2	0.437	FV2	0.297
	2	FV1	0.076	FV7	0.081	OX3	0.141	FW6	0.181
	3	FV10	0.075	FV15	0.076	FV14	0.115	NO1	0.136
	4	FD15	0.068	XN1	0.071	FW6	0.113	FV6	0.111
	5	FW6	0.068	FM9	0.065	FV1	0.107	FM9	0.105
−	1	FM12	-0.090	ON1	-0.079	FW9	-0.146	FW9	-0.120
	2	RN3	-0.078	FW4	-0.079	FW2	-0.119	FF8	-0.100
	3	FV8	-0.074	FS1	-0.070	FD8	-0.117	NH1	-0.100
	4	OE3	-0.070	FD10	-0.070	RN3	-0.101	SN1	-0.085
	5	BO3	-0.069	OE3	-0.069	FD1	-0.098	FM12	-0.084

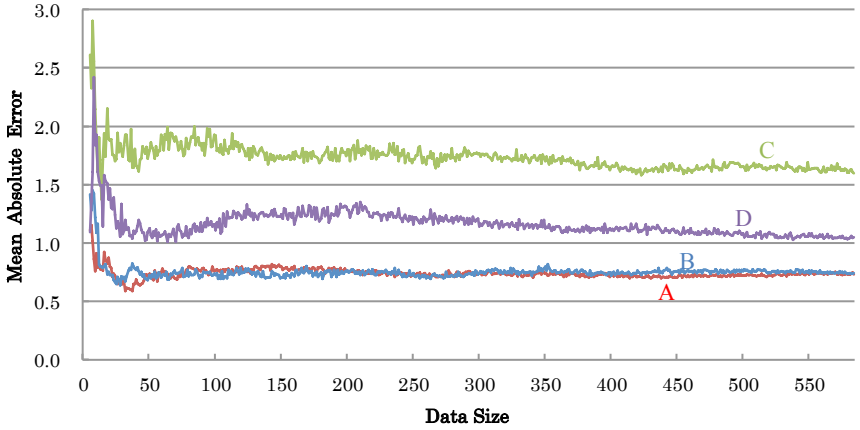

図 2：教師データ数の増加に伴う MAE の推移

実験2. 評価モデルの顕在化

教師データ 584 事例を全て訓練に用いて SVR を構築し，回帰係数の各成分値について検討する．表 8 に，成分値の絶対値が大きいもの上位 5 素性を正負別に示す．また，付録に全素性の成分値の分布を示す．ただし，FC1 から FC120 までの素性については，素性が示す意味自体を ID として表記する．これらの素性 ID はそれぞれ，1 文字目が前文，2 文字目が後文の構文役割を，3 文字目が 2 文間の接続関係を表す．構文役割”B”は文章始め，”E”は文章終わりのダミー要素に，”N”は前述の「-:出現せず」に対応する．また接続関係は，1 が論理的結合関係，2 が多角的連続関係，3 が拡充的合成関係に対応する．

実験3. 教師データ数と評点推定性能の関係

教師データの増加に伴う評点推定性能の推移について検討する．図 2 に 584 事例の教師データの部分集合(無作為抽出)を用いた 5 分割交差検定の結果を示す．

5.3 考察

表 7 より，正規化方法は ii を用いる方が良い結果が得られる傾向がある．教師データのラベル系列(評価者)間で MAE に差がみられるが，これは評価者による評点分布の違いに起因するものと考えられる．表 8 と付録より，「字数制限の達成度」，「複合名詞の使用率」はプラス評価，「自立語におけるひらがな使用率」，「文末の単調度」，「オノマトペの使用率」はマイナス評価であり，評価者によらない傾向がわかる．一方，ほとんど影響がでない素性もあることがわかる．カテゴリ「文章のまとまり」や「内容」に関する素性群には「OX3」や「RN3」など，成分値が大きくかつ評価者間で極性が共通する素性もあるが，大半の素性は成分値が 0 に近い．カテゴリ「モダリティ」に関する素性群の中では，「真偽判断の程度が断定的な文の出現率」が，評価者間で共通してプラス評価に働いている．一方，カテゴリ「内容」に関する素性は全評価者において成分値の絶対値が小さく，あまり評点に影響を与えていないことが分かる．また図 2 より，全体的にデータの増加に対する MAE の変化は収束傾向にあることがわかる．

6. おわりに

本稿では，日本語文章の教育的評価観点を「表層」「語」「文体」「係り受け」「文章のまとまり」「モダリティ」「内容」というカテゴリに分けられる素性群で表し，機械学習を用いて評価者の評価モデルを学習かつ顕在化する手法について述べた．この手法により，「個々の評価者が着目する言語的要素の明示」と「評点決定に寄与する要素の重みの定量化」が可能になる．しかし，本稿で言及した素性のうち「文章のまとまり」「内容」といったカテゴリの素性群は，ほかの素性群に比べ影響が小さいことが分

かった．原因として，高校生の小論文の「総合」評価を使って評価モデルの学習を試みたことが考えられる．我々が研究対象とした高校生の小論文データには，「総合」評価のほかに「語句」「表現」「語彙」「課題」「簡潔」「明確」「構成」「一貫」「説得」「独創」という観点からの評価も存在する．今後，「文章のまとまり」「内容」といったカテゴリの素性群の影響が大きく反映される観点(「課題」，「一貫」等)からの評価モデルの学習を試みる必要があると考える．

謝辞　本研究については，公益財団法人博報児童教育振興会の児童教育実践事業についての研究助成事業，「学習指導要領に立脚した児童作文自動点検システムの実現」（助成番号：11-B-081，　研究代表：藤田彬）の援助を受けた．

また，高校生の小論文答案をお貸しいただき，研究利用を認めて下さった揚華氏，宇佐美慧氏，東京工業大学大学院社会理工学研究科の前川眞一教授に感謝の意を表す．

参考文献

1) A. Smola. and B. Sch.: A tutorial on Support Vector Regression, Technical report, NeuroCOLT2 Technical Report NC2-TR-1998-030(1998).
2) Barzilay, R. and Lapata, M.: Modeling Local Coherence: An Entity-based Approach, Computational Linguistics, 34(1), pp.1-34(2008).
3) Burstein, J. et al.: Automated Scoring Using A Hybrid Feature Identification Technique, ACL '98 Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1(1998).
4) Burstein, J. et al.: Toward Evaluation of Writing Style: Finding Overly Repetitive Word Use in Student Essays, EACL '03 Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics, Volume 1(2003).
5) Elliot, S.: How Does IntelliMetric Score Essay Responses?, RB-929, Newtown, PA: Vantage Learning(2003).
6) Ellis B.Page.: The Imminence of Grading Essays by Computer, The Phi Delta Kappan, Vol. 47, No. 5, Jan(1966).
7) Ellis B.Page.: New Computer Grading of Student Prose, Using Modern Concepts and Software, Journal of Experimental Education(1994).
8) Halliday, M. A. K. and Hasan, R.: Cohesion in English, Longman, London(1976).
9) Ishioka Tsunenori, Masayuki Kameda.: 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(2006).
10) J.A.Keats.: An Introduction to Quantitative Psychology, John Willey & Sons Australasia Pty Ltd(1971).
11) Klaus Krippendorff.: Content Analysis: An introduction to its methodology, Sage Publications(1980).
12) Landauer, T. K., Laham, D. and Foltz, P. W.: The Intelligent Essay Assesor, the Debate on Automated Essay Grading, IEEE Intelligent Systems, Vol.15, No.5, pp.27-31(2003).
13) Landauer, T. K., Laham, D. and Foltz, P. W.: Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor, Sherims, M. & Burstein, J. eds. Automated Essay Scoring: A Crossdisciplinary Perspective, pp.87-112, Hillsdale, NJ: Lawrence Erlbaum Associates(2003).
14) Peter W. Foltz, Darrell Laham, and Thomas K Landauer.: Automated Essay Scoring: Applications to Educational Technology, Proc. EdMedia'99(1999).
15) Richard E. Nisbett. and Timothy DeCamp Wilson.: The halo effect: Evidence for Unconscious Alternation of Judgements, Journal of Personality and Social Psychology, Vol.35, No.4, 250-256(1977).
16) U. Kresel.: Pairwise Classification and Support Vector Machines Methods, MIT Press(1999).
17) Marilyn Walker, Masayo Iida, Sharon Cote: Japanese Discourse and the Process of Centering, Computational Linguistics, 17(1), pp.21-48(1994).
18) Yigal Attali, and Don Powers.: A Developmental Writing Scale, ETS Research Report No. RR-08-19, Educational Testing Service(2008).
19) Yigal Attali. and Jill Burstein.: Automated Essay Scoring With E-rater v.2.0, Article in Journal of Technology, Learning, and Assessment, Vol. 4, No. 3(2006).
20) 横野光，奥村学.: テキスト結束性を考慮した entity grid に基づく局所的一貫性モデル，自然言語処理, Vol.17, No.1, pp.161-182(2010).
21) 岡野原大輔，辻井潤一.: レビューに対する評価指標の自動付与，自然言語処理, Vol.14, No.3, pp.273-295(2007).
22) 岩淵悦太郎.: 第三版 悪文，日本評論社(1979).
23) 益岡隆志.: モダリティの文法，くろしお出版(1991).
24) 松吉俊，江口萌，佐尾ちとせ，村上浩司，乾健太郎，松本裕治.: テキスト情報分析のための判断情報アノテーション，電子情報通信学会論文誌 D Vol.J93-D No.6(2010).
25) 国立国語研究所.: 分類語彙表増補改訂版データベース(2004).
26) 松吉俊，佐尾ちとせ，乾健太郎，松本裕治.: 拡張モダリティ付与コーパス作成の基準 version 0.8β, http://cl.naist.jp/nltools/modality/manual.pdf(2011).
27) 松吉俊，佐藤理史，宇津呂武仁.: 日本語機能表現辞書の編纂，自然言語処理, Vol.14, No.5, pp.123-146(2007).
28) 森岡健二.: 文章構成法 文章の診断と治療，至文堂(1963).
29) 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).
30) 石岡恒憲.: 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，人工知能学会誌, Vol.23, No.1, pp.17-24(2008).
31) 田窪行則，西山佑司，三藤博，亀山恵，片桐恭弘.: 談話と文脈，言語の科学 7，岩波書店(2004).
32) 日本電子化辞書研究所.: EDR 電子化辞書 V4.0，独立行政法人情報通信研究機構(2010).
33) 文部科学省.: 小学校学習指導要領国語編，東洋館出版社(2008).

Vol.2011-NL-202 No.3
2011/7/15

付録