Vol.2016-NL-227 No.7
2016/7/29

誤り傾向と文の容認性に着目した英作文のレベル判定

林 正頼[1]　笹野 遼平[1]　高村 大也[1]　奥村 学[1]

概要：英語教育において，学習者が書いた英作文が，どの程度のレベルであるかを把握することは，教育者，学習者双方にとって有用である．本研究では，英作文のレベル判定問題を順序回帰問題として定式化する．レベル判定の手がかりとして，語彙情報といった基本的な素性に加え，英作文に含まれる誤りの傾向や，文の容認性などを導入し，それらの有効性を検証する．

1. はじめに

語学教育において，学習者が書いた作文が，どの程度のレベルであるかを知ることは，教育者，学習者双方にとって有用である．たとえば，教育者側は，学習者のレベルに即した指導法を用意することで，より効率的な指導を行えることが期待でき，学習者側は，自分の現在の英作文レベルが明確になることで，学習意欲の向上が期待できる．

近年，語学レベルの指標として，ヨーロッパ言語共通参照枠 (Common European Framework of Reference for Languages : CEFR) [1] が欧州で普及しており，欧州の外国語教育者は，この CEFR レベルを参照して学習状況の評価や指導を行うことが増加している．本研究では，日本人英語学習者による英作文から構成される JEFLL (Japanese English as a Foreign Language Learner) コーパスを対象に，学習者が書いた英作文がどの CEFR レベルに該当するか，自動的に判定するシステムの提案を行う．英作文の CEFR レベルの自動判定ができれば，英作文の CEFR レベルを人手で判定していた教育者への負担が軽減され，また，システムが自動的に CEFR レベルを判定することで，学習者単独でも効率的な学習が行えるようになることが期待できる．

これまでにも，英作文のレベルを自動判定するシステムは複数提案されてきた．これらのシステムが英作文能力を判定する手がかりとして，学習者がどのような語彙が使えるか，どのような構造を持つ文が書けるかといった情報がよく使われている．しかし，学習者が書いた作文中には，さまざまな誤りが含まれる可能性が高く，どのような誤りをしているかという傾向も，作文のレベルを判定する上で重要な手がかりとなると考えられる．また，母語話者が学習者の書いた文を読むとき，文法的には正しい文であっても読みづらく感じることがあるが，このような母語話者が感じる文の容認性の度合いも作文レベルの判定の手がかりとなると考えられる．そこで，本研究では，学習者が書いた作文に含まれる誤り傾向と文の容認性に着目した英作文のレベル判定システムを提案する．

2. 関連研究

英作文のレベル判定と関連した研究として，英作文の自動採点に関する研究がある．英作文の自動採点に関する研究は，教師の負担を軽減する目的で古くから行われており，様々なアプローチが提案されている [2, 3]．代表的なアプローチとしては，Project Essay Grade (PEG) [2] や E-rater [3] は重回帰分析，Bayesian Essay Test Scoring sYstem (BETSY) [4] はナイーブベイズ分類器に基づいた方法を用いて，英作文の自動採点を行っている．また，自動採点を行うための手がかりの抽出法に関してもいくつかの手法が提案されており，Intelligent Essay Assessor (IEA) は潜在的意味解析 (LSA) [5]，IntelliMetric は決定木を用いている [6].

この中で，TOEIC や TOEFL で知られる Educational Testing Service (ETS) が提案した E-rater は，アメリカの Graduate Managemant Admissions Test (GMAT) の作文試験で使用されている．E-rater は，スコア予測のために，異なり語数や総単語数といった基本的な素性に加えて，本研究でも着目している誤りに関する素性も使用している．具体的には，英作文内に含まれる品詞バイグラムを，相互情報量と対数尤度比を基に誤り箇所を推定し，全体の作文中に含まれる誤りの出現割合を素性として利用している．これに対し本研究では，誤りの出現割合ではなく，誤り訂正モデルを導入することで，どのような誤りをしているかという誤り傾向を抽出し，分類の手掛かりとして使用する．

本研究でも用いる文法誤り訂正は，近年，文法誤り訂

[1] 東京工業大学 Tokyo Institute of Technology

正の精度を競う Shared Task も近年開催される [7, 8] など，盛んに研究されている分野である．誤りの傾向によって，ルールベースと機械学習手法を組み合わせを変える手法 [7]，さらに，機械翻訳のモデルとして Neural Machine Translation を用いた手法 [9] などが提案されている．本研究では，このうち，機械翻訳に基づく手法を文法誤り検出に使用する．

3. JEFLL コーパス

JEFLL (Japanese English as a Foreign Language Learner) コーパス [*1] は，中学・高校の日本人英語学習者，約 1 万人分の自由英作文データをコーパス化したものである．表 1 に示した論題について書かれた英作文 (原文) と，原文を英語教育者が訂正したもの (正解文) のペアから構成される．原文と正解文は，一文の原文に対して一文の正解文のみ存在し，一対多の関係となるペアは存在しない．ただし，単語単位での対応付けはされていない．

表 1 JEFLL コーパスの論題

論題	概要
Urashima	浦島太郎のその後を想像
Rice or Bread?	朝食はどちらが良いか，理由も説明
Festival	自身の学校祭についての説明
Earthquake	地震に巻き込まれた時に何をするか，理由も説明
Otoshidama	お年玉をもらったら何を買いたいか
Bad Dream	これまで見た悪夢で一番の悪夢を説明

なお，以下の条件で原文は作成された．

- 授業中に作成
- 制限時間は 20 分
- 辞書不使用
- 事前準備なし
- 英語で書けない場合はローマ字等でも可

原文と正解文にはそれぞれ各単語の原形と品詞情報が追加されている．さらに，英語教育者によって各英作文毎に CEFR レベルが付与されている．CEFR レベルは A1，A2，B1，B2，C1，C2 の 6 段階から構成されており，C2 が最も高いレベルとなっている．なお，今回の JEFLL コーパスでは，B2 レベルが最高であった．

本研究では，英作文の平文からレベル判定を行うため，付与された情報は原則的に使用していない．ただし，後述する文法項目の抽出には，原文に付与された原形，品詞情報を，予備実験の評価の際には，正解文に付与された原形，品詞情報をそれぞれ使用した．それ以外の実験では，JEFLL コーパスの原文，および事例ごとに付与された CEFR レベルのみを使用した．

[*1] http://jefll.corpuscobo.net/index.htm

4. 提案手法

本研究では，英作文のレベル判定タスクを順序回帰問題として捉え，JEFLL コーパスから抽出した素性集合に基づく CEFR レベルの判定モデルを構築する．

4.1 分類方法

英作文のレベルを判定するということは，各事例が英作文レベルのどのクラスに属するかを判定する分類問題と考えられる．しかし，英作文のレベル判定を行う場合は，各クラスが完全に独立したものではないことから，広く使われる多クラス分類の手法を用いることは適切ではないと考えられる．

そこで，本研究では，順序回帰を用いて英作文のレベル判定を行う．順序回帰 (ordinal regression) は，順序付きのカテゴリを従属変数，素性集合を独立変数とする回帰問題であり，各事例間の順序を用いて判定を行う手法である．つまり，各事例に付与されたラベルが英作文のレベルであることを踏まえた上で判定する手法であり，今回の目的に対して有効な手法であると考えられる．

4.2 素性集合

本研究では，ベースラインとなる素性に加えて，英作文内に含まれる誤りに関する素性，英作文内で使用された文法に関する素性，文の容認性に着目した素性を使用する．以下では，それぞれのタイプの素性についての詳細を述べる．

4.2.1 ベースライン素性

ベースライン素性として，単語情報，文書情報，および単語の難易度情報を使用する．これらは自動採点に関する研究で広く使用されている素性である．品詞情報を表す素性としては，英作文中に出現する品詞タグのユニグラム，バイグラム，トライグラムの頻度を用いる．なお，品詞情報の付与には Stanford POS Tagger[*2] を使用した．単語の表層形は，品詞情報と同様に，単語ユニグラム, バイグラム，トライグラムの頻度を用いる．また，文書情報としては，一文あたりの平均単語数，一単語あたりの平均文字数，文章中の異なり語数を総単語数で割った値を使用する．

単語の難易度の情報は，投野によって分類されたリスト [10] を使用する．このリストでは単語は英語教育者によって 4 段階にレベル分けされており, 事例中に含まれる各レベルの割合を離散化したものを素性とする．

4.2.2 誤りに関する素性

学習者が書いた作文には，作文内に誤りが含まれる可能性が高く，どのような誤りをしているかという情報は，英作文のレベルを判定する際に有効な情報と考えられる．し

[*2] http://nlp.stanford.edu/software/tagger.shtml

かし，学習者が書いた作文の情報だけでは，どの箇所に誤りが含まれているか，どのような種類の誤りであるかは特定できない．本研究では，誤りを含む原文と，その誤りを訂正した正解文の対から，統計的機械翻訳の技術に基づく誤り訂正モデルを事前に構築することで，この問題に対処する．具体的な処理の流れは以下の通りである．

まず，誤りを含む原文を入力すると，その誤りを正した文 (システム出力文) を出力する誤り訂正システムを構築する．続いて，原文とシステム出力文の対応付けを行うことで，誤りの箇所，種類を特定する．この際，誤りの箇所を特定するためには，原文-システム出力文間の単語同士のアライメントを行うことが必要である．本研究では，望月ら [11] による編集距離を用いた動的計画法に基づく英文自動エラータグ付与ツールで，単語間のアライメントを行った．この特定した情報を学習者のレベル判定の手がかりとして使用する．システムが正しい訂正を行っている場合，アライメントの出力結果としては，以下の 4 通りが考えられる．

- **誤りなし**：原文とシステム出力文の単語が同一である
- **置換誤り**：システムが単語を書き換えた
- **脱落誤り**：システムが単語を追加した
- **余剰誤り**：システムが単語を削除した

本研究では，置換，脱落，余剰の誤りがあった単語の頻度を，その品詞ごとに集計し，素性として用いる．たとえば，システムが形容詞を追加しているような場合，「形容詞 (脱落)」という素性が用いられる．ただし，品詞が，前置詞や限定詞といった，機能語である品詞の場合は，単語の表層形をそのまま素性として使用する．すなわち，たとえば，誤り訂正システムが前置詞 'in' を 'at' に置き換えているような場合，「in (置換)」という素性が用いられる．

4.2.3 文法項目に関する素性

指示形容詞が使用されているかや，仮定法過去が使用されているかなどの，英作文内で使用されている文法についての情報は，英作文のレベル判定に有効な手がかりとなると考えられる．石井ら [12] はこのような情報を文法項目と呼び，表 2 に示すような合計 493 種類の文法項目のリストを整備している．

表 2 文法項目の例

ID	文法項目
1	人称代名詞主格 (I am)
3	人称代名詞主格 (he/she is)
11	指示形容詞 (this/that+名詞)
137	助動詞類 (should)
253	wish+仮定法過去

本研究では，石井らの文法項目リストを参照し，英作文に含まれる各文法項目の頻度を文法項目に関する素性として使用する．しかし，英作文中に使われている文法項目は，誤って使用されている可能性も考えられる．本研究では，誤って使用した文法項目の情報も素性として利用するために，前節で述べた，原文-システム出力文間の単語のアライメント結果を使用する．もし誤った使い方をしているならば，原文からは抽出されない文法項目が，システム出力文から抽出されることになる．したがって，システム出力文で文法項目が出現している該当箇所の単語に着目し，以下の基準で正用例と誤用例の 2 つに分類した上で，493 種類の文法項目それぞれに対し，正用例，誤用例の頻度を素性として利用する．

- **誤用例**：置換，脱落，余剰の誤りがある単語を含む
- **正用例**：上の誤りがある単語を含まない

なお，英作文から文法項目を抽出するためには，JEFLL コーパスに対応した正規表現パターンを使用しなければならない．本研究では，文法項目を抽出する時のみ，JEFLL コーパスの原文のデータを使用する．また，誤り訂正システムで出力された文は平文のままであり，このままでは文法項目を抽出できないため，JEFLL コーパスの作成方法をシステム出力文に適用し，XML 形式に変換を行ってから文法項目を抽出を行う．

4.2.4 容認度

母語話者が文を読んだ際，その文全体が自然に感じる（容認できる），もしくは不自然に感じる（容認出来ない）というように，文を容認できる度合いを容認度 (Acceptability) と呼ぶ．本研究では，英作文のレベルが高いほど容認度が高いとの仮定に基づき，容認度を英作文のレベル判定の素性として利用する．具体的には，Lau らの手法 [13] により文の容認度を算出する．Lau らの手法は，まず，確率言語モデルを大量の文書データから学習し，文の生成確率を出力する．その上で，表 3 に示す式により，文の生成確率を容認度に変換する式に適用する．なお，$P_m(\xi)$ は各言語モデルが出力する ξ の確率，$P_u(\xi)$ はユニグラム言語モデルによる ξ の確率，$|\xi|$ は単語数である．

表 3 言語モデルの生成確率から容認度への変換式

容認度	式
$LogProb$	$logP_m(\xi)$
$MeanLP$	$\frac{logP_m(\xi)}{\lvert\xi\rvert}$
$NormLP(Div)$	$-\frac{logP_m(\xi)}{logP_u(\xi)}$
$NormLP(Sub)$	$logP_m(\xi) - logP_u(\xi)$
$SLOR$	$\frac{logP_m(\xi) - logP_u(\xi)}{\lvert\xi\rvert}$

4.2.5 surprisal value

人間は自然言語を理解する際，文中に含まれる単語を逐次的に処理しているとされている．また，これまでの研究で，既知の情報を処理しているだけではなく，既知の情報を元に，次にどのような単語が来るかを予測していることが明らかになっている [14].

文を予測して読み進めるという仮定のもとに，文を読む時に生じる単語単位の処理負荷を計算する数理モデルが Hale によって提案されている [15]．Hale は予測を含んだ文処理モデルを Surprisal Theory と呼び，文中の $i-1$ 番目の単語までで予測される文の生成確率と，i 番目の単語を処理した後の文の生成確率の差が処理負荷と比例する，という説を提唱している．つまり，この説が正しいならば，i 番目の単語が，$i-1$ 番目まで読み進めた予測と大きく異なると処理負荷が大きくなる．この各単語に生じる処理負荷は surprisal value と定義され，i 番目の単語 w_i の surprisal value は以下の式で算出される：

$$Surprisal(w_i) = -log\frac{Pr(w_{1...i})}{Pr(w_{1...(i-1)})}.$$

なお，$Pr(w_{1...i})$ と，$Pr(w_{1...(i-1)})$ はそれぞれ，i 番目，$i-1$ 番目までの文の生成確率である．

また，Revy は人間は単語を予測しながら読み進めると同時に，文の構造も予測していると考え，文構造においてもこの surprisal theory が成り立つことを示した [16]. すなわち，i 番目の単語を見た時，$i-1$ 番目までで予測していた文の構造と大きく異なる場合に処理負荷が増大する．そこで本研究では，単語を予測した際に生じる surprisal value を lexical surprisal，文構造を予測した際に生じる surprisal value を syntactic surprisal とし，それぞれを素性として利用する．

5. 実験設定

5.1 使用データ

実験には，JEFLL コーパスを使用し，1 つの英作文を 1 事例として扱った．JEFLL コーパスのレベル別の事例数を表 3 に示す．なお，今回使用した JEFLL コーパスには B2 レベルと分類された英作文も存在するが，他のレベルと比較して事例数が非常に小さかったため，A1，A2，B1 の 3 つのレベルの事例のみを使用した．

表 4 JEFLL コーパスの事例数

レベル	事例数
A1	3362
A2	4900
B1	1520
B2	45
合計	9827

5.2 誤り訂正システム

5.2.1 設定

機械翻訳技術を用いた誤り訂正システムの構築のための学習データとして，Lang-8 Learner Corpora を使用した [17]．Lang-8 Learner Corpora は，投稿者が学習したい言語で文章 (投稿文) を投稿すると，その言語を母語とする添削者が，投稿文を添削して添削文を返信する，言語学習者向けの相互添削型 SNS である Lang-8[*3] のログデータをデータベース化したものである．この Lang-8 Learner Corpora から英語で書かれている作文のみを取り出し，投稿文，添削文のペアを学習データとし，誤り訂正モデルの構築を行った．水本ら [17] の前処理を参考に，最終的に 793,190 文対を学習データとして使用した．デコーダには，オープンソースの統計的機械翻訳システムである Moses Decoder を用いた [18].

また，アライメントの手法によって誤り訂正の精度が変化すると考え，今回は 2 つのアライメントモデルを用意し比較を行った．1 つ目は GIZA++で，Moses decoder を用いる際にデフォルトで使われる，IBM モデルに基づくアライメント手法である [19]．2 つ目は交差するアライメントをとらないという制約がある Fast Align を使用した [20]. Fast Align を使用したのは，英語の誤りの傾向として，語順が入れ替わる誤りは，今回定義した置換，脱落，余剰誤りという単純な誤りに比べ少量であると予測し，この手法が今回のタスクにおいて有効であると考えためである．言語モデルとして IRST Language Modeling Toolkit (IRSTLM) の 5-gram 言語モデルを使用した [21].

本研究ではこのアライメントのためのパラメータはデフォルト値を使用した [*4].

5.2.2 誤り訂正の性能評価

機械翻訳手法を用いた誤り訂正がどの程度訂正されているかの評価を行うために，機械翻訳の性能評価で広く用いられる BLEU [23] と，単語単位の適合率, 再現率, 及び F 値で評価した．評価には，本実験でも用いる JEFLL コーパスの A1, A2, B1 レベルから無作為に 100 事例ずつ選び，合計 300 文書，2,998 文を使用した．

BLEU は，システムが出力した文と参照文の n-gram の適合率に基づいて評価する．今回は誤り訂正システムが出力した文と，正解文間で評価を行う．つまり，原文-正解文間の BLEU スコアに比べ，システム出力文-正解文間の BLEU スコアが向上しているならば誤り訂正が有効であると考えられる．結果は表 5 となり，両アライメント手法ともに，システム出力文-正解文間の BLEU スコアが向上していることを確認した．

表 5 BLEU スコア

レベル	原文	GIZA++	Fast Align
BLEU	44.66	46.96	46.31

また，適合率, 再現率, 及び F 値は図 1 のように，単語単位で原文，正解文，システム出力文を対応付けて評価

[*3] http://lang-8.com/
[*4] MERT (Minimum Error Rate Training) [22] を用いたパラメータチューニングを試したが，効果を確認できなかったためにデフォルト値を使用した．

原文:	He	eat	it	chocolate	.
正解文:	He	eats	a	chocolate	.
システム:	He	eat	a	chocolates	.
	TN	FN	TP	FP	TN

図 1 単語単位の誤り訂正の評価

を行った．TP (True Positive) は，システムが訂正を行いかつ正しい訂正だった箇所，FP (False Positive) は，システムが訂正を行ったが誤った訂正だった箇所，FN (False Negative) は，システムが訂正を行わずかつ訂正が必要であった箇所である．また，$N(x)$ は各箇所の総数を表す．

なお，適合率, 再現率, F 値は以下の式で算出する：

$$\text{適合率} = \frac{N(TP)}{N(TP) + N(FP)}$$

$$\text{再現率} = \frac{N(TP)}{N(TP) + N(TN)}$$

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{再現率} + \text{適合率}}.$$

結果は表 6 に示す．再現率に比べ，適合率は比較的高い値となった．適合率に関して注意すべき点として，原文に対する正解文が 1 つのみという実験設定で評価を行ったことが挙げられる．これは，ある誤りに対して，複数の訂正候補が考えられる場合もあり，人手による誤り訂正でも適合率が 1 とならない可能性が高い．また，2 つのアライメント手法に関しては，誤り訂正の精度に大きな差は見られず，また，英作文のレベル判定を行った結果にもほとんど違いが確認できなかった．このため，本稿では速度の面で優位なアライメント手法である Fast Align を使用した場合の結果を報告する．

表 6 単語単位の誤り訂正結果

アライメント	適合率	再現率	F 値
GIZA++	0.349	0.067	0.113
Fast Align	0.329	0.072	0.118

5.3 容認度

Lau らの報告 [13] によると，人手によって作成された容認度と最も相関が高かった組み合わせは，ニューラル言語モデルと，$SLOR$ の変換式であった．一方，英語学習者のエッセイから成るデータセットとの相関は，n-gram 言語モデルと，$NormLP(Div)$ の変換式が最良の結果となっている．このように，データにより最適なモデルは異なっていることから，本研究では，まず，ニューラル言語モデルと 4-gram 言語モデルの 2 つの言語モデルを用意し，先行研究で提案されている変換式すべてで容認度を算出し，本論文の実験に最適な組み合わせを検討した．

ニューラル言語モデルの学習には，Recurrent Neural Network Language Model (RNNLM) [24] を使用し，パラメータは隠れ層のユニット数が 600，学習率は 0.1 とした．n-gram 言語モデルの学習には，The SRI Language Modeling Toolkit (SRILM) [25] を使用し，4-gram で学習させ，Kneser-Ney スムージングを行ったものを使用した．また，言語モデルの学習データとしては，British National Corpus(BNC コーパス) を使用し，先行研究と同様に，1 文に含まれる単語が 8 単語未満の場合は学習データとして使用しなかった．この結果，総文数は 3,258,189 文で，語彙サイズは 424,621 単語となった．

用意した 2 つの言語モデルと，表 3 に示した各変換式で各文の容認度を計算し，CEFR レベルとの相関係数を算出した結果を表 7 に示す．ここで，容認度は 1 文ごとに算出されるため，各事例ごとの容認度の平均値と，付与されいてる CEFR レベルの相関係数を算出している．

表 7 CEFR レベルと容認度の相関係数

	RNNLM	4gram LM
$LogProb$	0.2635	0.2667
$MeanLP$	0.3449	0.3460
$NormLP(Div)$	0.2669	0.2839
$NormLP(Sub)$	0.3378	0.3372
$SLOR$	-0.0851	-0.1255

この結果から，先行研究では有効であった $SLOR$ は，どちらの言語モデルを使用した場合でも負の相関となり，本研究においては有効ではないことが確認された．なお，以降の実験ではニューラル言語モデルを使用し，$NormLP(Sub)$ の変換式によって得られた容認度を使用した．容認度を素性として利用する際は，各事例ごとの容認度の平均値を離散化し利用した．

5.4 構文解析器

surprisal value を計算する際に必要となる単語の生成確率を計算するための構文解析器として，Roark [26] らの Incremental top-down parser を使用した．Incremental top-down parser は句構造解析を行う構文解析器である．この構文解析器は，確率的文脈自由文法の一種で，一単語ずつ文頭から逐次的に解析することが特徴である．つまり，各単語位置までの途中状態の文の生成確率，および文構造の途中状態の生成確率を算出することができる．このため，途中状態の確率値を元に surprisal value を計算することが可能である．学習データには，Penn Treebank [27] 内にある Wall Street Journal の 2 章から 24 章までを利用した．

学習に使用する素性は，事例内の単語の中で最も高い値となる syntactic surprisal, lexical surprisal を使用し，それらの連続値を離散化したものを素性に加えた．

5.5 分類器

順序回帰の実装は mord: Ordinal Regression in Python [28] を使用した．mord では，損失関数が複数定義されているが，先行研究 [29] で最良の結果が報告されている All-threshold を使用した．また，順序回帰には正則化のパラメータである α がハイパーパラメータとして存在する．そこで，各素性集合において最適な α を用いて実験を行うために，学習データ内で 5 分割交差検定を行うことで α を決定した．このハイパーパラメータ α を決定する実験の評価方法として Mean Absolute Error (MAE) を用いた．MAE の式を以下に示す：

$$MAE(y, \hat{y}) = \frac{1}{N} \sum |y_i - \hat{y}_i|.$$

N は評価事例数，y_i は正解の値，$\hat{y}_i$ は予測値である．この際，α は 0.01 から 3 までの範囲で変動させ実験を行った．そして，各 α ごとに得られた MAE の平均値をとり，最も MAE の平均値が低くなる α を採用した．

なお本稿では，提案した各素性がそれぞれ有効であるか検討するために，以下の 6 通りで実験を行った．

- ベースライン素性
- ベースライン素性＋誤りに関する素性
- ベースライン素性＋文法項目に関する素性
- ベースライン素性＋容認度
- ベースライン素性＋ Surprisal Value
- ベースライン素性＋提案素性すべて

6. 実験結果

本研究では 5 分割交差検定を行うことで得られる F 値によって提案した素性の有効性を確認する．各レベル間，および全体の F 値を表 8 に示す．

表 8 実験結果 (F 値)

素性	A1	A2	B1	全体
ベースライン (Base)	.802	.879	.745	.834
Base + 誤り	.812	.883	.783	.844
Base + 文法項目	.812	.904	.788	.857
Base + 容認度	.801	.877	.754	.834
Base + Surprisal Value	.795	.880	.764	.835
Base + 提案素性すべて	.853	.889	.864	.873

この結果から，誤りに関する素性，および文法項目に関する素性は特に有効であることが示された．一方，容認度，surprisal value といった容認性に関する素性は有効とは言えない結果となった．しかし，本研究で提案したすべての素性を組み合わせた場合に最も良い F 値となっており，素性の組み合わせ方によってはこれらの素性も有効である可能性も考えられる．このことから，素性の組み合わせについての検討が必要であり，今後の課題として挙げられる．

7. まとめ

本研究では，学習者が書いた英作文を対象に，英作文のレベルを順序回帰に基づき自動的に判定するシステムを提案した．提案システムでは，英作文の自動採点に使われる語彙情報といった基本的な素性に加え，文法誤り訂正システムを用いることで得られる誤りの傾向に関する素性や，文の容認性に関する尺度である容認度や surprisal value を素性として導入した．また，実験を通して，誤りの傾向に関する素性が英作文のレベル判定に有効であることを確認した．今後の課題としては，本研究で有効性が確認できなかった容認度や surprisal value の導入方法の検討や，文間の関係に着目できる談話構造の利用などが挙げられる．

参考文献

[1] Fred Davidson and Glenn Fulcher. The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching*, Vol. 40, No. 3, pp. 231–241, 2007.

[2] Ellis B. Page. The Imminence of Grading Essays by Computer. *The Phi Delta Kappa*, Vol. 47, No. 5, pp. 238–243, 1966.

[3] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, NAACL-HLT'11, pp. 180–189, 2011.

[4] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *In AAAI-98 workshop on FOR TEXT CATEGORIZATION*, pp. 41–48. AAAI Press, 1998.

[5] Peter. W. Foltz, Darell. Laham, and Thomas. K. Landauer. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Education Journal of Computer enhanced learning On-line journal*, Vol. 1, No. 2, 1999.

[6] Lawrence M. Rudner, Veronica Garcia, and Catherine Welch. An evaluation of the intellimetric[sm] essay scoring system. *Journal of Technology, Learning, and Assessment*, Vol. 4, No. 4, 2006.

[7] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14, 2014.

[8] Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–12, 2013.

[9] Zheng Yuan and Ted Briscoe. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 380–386, 2016.

[10] 投野由紀夫. CAN-DO リスト作成・活用英語到達度指標 CEFR-J ガイドブック, 6 2013.

[11] 投野由紀夫, 望月源. 編集距離を用いた英文自動エラータグ付与ツールの開発と評価. 『コーパスに基づく言語教育研究報告』, No. 9, 2012.
[12] 投野由紀夫, 石井康毅. 英語 cefr レベルを規定とする基準特性としての文法項目の抽出とその評価. 言語処理学会第21回年次大会, pp. 884–667.
[13] Jey Han Lau, Alexander Clark, and Shalom Lappin. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1618–1628, 2015.
[14] Katherine DeLong, Thomas Urbach, and Marta Kutas. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. In *Nature Neuroscience*, pp. 1117–1121, 2005.
[15] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL'01, pp. 1–8, 2001.
[16] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, Vol. 106, No. 3, pp. 1126–1177, 2008.
[17] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining revision log of language learning sns for automated Japanese error correction. *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 28, No. 5, pp. 420–432, 2013.
[18] Hieu Hoang and Philipp Koehn. Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP'08, pp. 58–65, 2008.
[19] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, 2003.
[20] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, 2013.
[21] Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. Irstlm: an open source toolkit for handling large scale language models. pp. 1618–1621, 2008.
[22] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL'03, pp. 160–167, 2003.
[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pp. 311–318, 2002.
[24] Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pp. 1045–1048, 2010.
[25] Andreas Stolcke. SRILM - an extensible language modeling toolkit. pp. 901–904, 2002.
[26] Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 324–333, 2009.
[27] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330, 1993.
[28] Fabian Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI, Feb 2015.
[29] Jason D. M. Rennie. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the International Joint Conference on Artificial Intelligence Multidisciplinary Workshop on Advances in Preference Handling*, pp. 180–186, 2005.
[30] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, Vol. 42, pp. 109–142, 1980.
[31] Gaurav Kharkwal and Smaranda Muresan. Surprisal as a predictor of essay quality. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 54–60, 2014.
[32] Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, Vol. 13, No. 4, pp. 519–549, 2008.
[33] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'02, pp. 133–142, 2002.