論文

正誤情報と文法誤りパターンを考慮した単語分散表現を用いた文法誤り検出

金子　正弘[†]・堺澤　勇也[†]・小町　　守[†]

本稿では，文法誤り検出のための正誤情報と文法誤りパターンを考慮した単語分散表現の学習手法を提案する．これまでの文法誤り検出で用いられている単語分散表現の学習では文脈だけをモデル化しており，言語学習者に特有の文法誤りを考慮していない．そこで我々は，正誤情報と文法誤りパターンを考慮することで文法誤り検出に特化した単語分散表現を学習する手法を提案する．正誤情報とは，n-gram 単語列内のターゲット単語が誤っているのか正しいのかというラベルである．これは単語単位の誤りラベルを元に決定される．誤りパターンとは，学習者が誤りやすい単語の組み合わせである．誤りパターンは大規模な学習者コーパスから単語分散表現の学習のために抽出することが可能である．この手法で学習した単語分散表現で初期化した Bidirectional Long Short-Term Memory を分類器として使うことで，First Certificate in English コーパスに対する文法誤り検出において世界最高精度を達成した．

キーワード：学習者支援，文法誤り検出，単語分散表現，深層学習

Grammatical Error Detection Using Error- and Grammaticality-Specific Word Embeddings

Masahiro Kaneko[†], Yuya Sakaizawa[†] and Mamoru Komachi[†]

In this study, we improve grammatical error detection by learning word embeddings that consider grammaticality and error patterns. Most existing algorithms for learning word embeddings usually model only the syntactic context of words and do not consider grammatical errors specific to language learners. Therefore, we propose methods to learn word embeddings specialized for grammatical errors by considering grammaticality and grammatical error patterns. We determine grammaticality of n-gram sequence from the annotated error tags and extract grammatical error patterns for word embeddings from large-scale learner corpora. Experimental results show that a bidirectional long-short term memory model initialized by our word embeddings achieved the state-of-the-art accuracy by a large margin in an English grammatical error detection task on the First Certificate in English dataset.

Key Words: *Educational Application, Grammatical Error Detection, Word Embedding, Deep Learning*

[†] 首都大学東京, Tokyo Metropolitan University

1 はじめに

作文中における誤りの存在や位置を示すことができる文法誤り検出は，第二言語学習者の自己学習と語学教師の自動採点支援において有用である．一般的に文法誤り検出は典型的な教師あり学習のアプローチによって解決可能な系列ラベリングのタスクとして定式化できる．例えば，Bidirectional Long Short-Term Memory (Bi-LSTM) を用いて英語の文法誤り検出の世界最高精度を達成している研究 (Rei and Yannakoudakis 2016) がある．彼らの手法は，言語学習者コーパスがネイティブが書いた生コーパスと比較してスパースである問題に対処するために，事前に単語分散表現を大規模なネイティブコーパスで学習している．

しかし，Rei と Yannakoudakis の研究を含む多くの文法誤り検出の研究において用いられている分散表現学習のアルゴリズムのほとんどは，ネイティブコーパスにおける単語の文脈をモデル化するだけであり，言語学習者に特有の文法誤りを考慮していない．一方で，単語分散表現に言語学習者に特有の文法誤りを考慮することは，より文法誤り検出に特化した単語分散表現を作成可能であり有用であると考えられる．そこで，我々は文法誤り検出における単語分散表現の学習に正誤情報と文法誤りパターンを考慮する 3 つの手法を示す．ただし，3 つ目の手法は最初に提案する 2 つの手法を組み合わせたものである．

1 つ目の手法は，学習者の誤りパターンを用いて単語分散表現を学習する **Error specific word embedding** (EWE) である．具体的には，単語列中のターゲット単語と学習者がターゲット単語に対して誤りやすい単語を入れ替え負例を作成することで，正しい表現と学習者の誤りやすい表現が区別されるように学習する．

2 つ目の手法は，正誤情報を考慮した単語分散表現を学習する **Grammaticality specific word embedding** (GWE) である．単語分散表現の学習の際に，n-gram の正誤ラベルの予測を行うことで，正文に含まれる単語と誤文に含まれる単語を区別するように学習する．この研究において，正誤情報とは周囲の文脈に照らしてターゲット単語が正しいまたは間違っているというラベルとする．

3 つ目の手法は，EWE と GWE を組み合わせた **Error & grammaticality specific word embedding** (E&GWE) である．E&GWE は正誤情報と誤りパターンの両方を考慮することが可能である．

本研究における実験では，英語学習者作文の文法誤り検出タスクにおいて，E&GWE で学習した単語分散表現で初期化した Bi-LSTM を用いた結果，世界最高精度を達成した．さらに，我々は大規模な英語学習者コーパスである Lang-8 (水本，小町，永田，松本 2013) を使った実験も行った．その結果，文法誤り検出においてノイズを含むコーパスからは誤りパターンを抽出して学習することが有効であることが示された．

本研究の主要な貢献は以下の通りである．

- 正誤情報と文法誤りパターンを考慮する提案手法で単語分散表現を初期化した Bi-LSTM を使い，First Certificate in English (FCE-public) コーパス (Yannakoudakis, Briscoe, and Medlock 2011) において世界最高精度を達成した．
- FCE-public と NUCLE データ (Dahlmeier, Ng, and Wu 2013) に Lang-8 から抽出した誤りパターンを追加し，単語分散表現を学習することで文法誤り検出の精度が大幅に向上することを示した．
- 実験で使用したコードと提案手法で学習された単語分散表現を公開した[1]．

本稿ではまず第 2 章で英語学習者作文における文法誤り検出に関する先行研究を紹介する．第 3 章では従来の単語分散表現の学習方法について述べる．次に，第 4 章では提案手法である正誤情報と誤りパターンを考慮した単語分散表現の学習モデルについて説明する．そして第 5 章では FCE-public と NUCLE の評価データである CoNLL データセットを使い提案手法を評価する．第 6 章では文法誤り検出モデルと学習された単語分散表現における分析を行い，最後に第 7 章でまとめる．

2 先行研究

文法誤り検出の研究の多くは前置詞の正誤 (Tetreault and Chodorow 2008)，冠詞の正誤 (Han, Chodorow, and Leacock 2006) や形容詞と名詞の対の正誤 (Kochmar and Briscoe 2014) のように特定のタイプの文法誤りに取り組むことに焦点が当てられている．一方で，特定のタイプの文法誤りではなく文法誤り全般に取り組んだ研究は少ない．Rei と Yannakoudakis (2016) は，word2vec を埋め込み層の初期値とした双方向の Bi-LSTM を提案し，FCE-public に対して全ての誤りを対象とする文法誤り検出タスクにおいて現在世界最高精度を達成している．我々も全ての文法誤り検出タスクの手法に取り組むが，正誤情報や学習者の誤りパターンを考慮した単語分散表現を使う．

誤りパターンを考慮した研究としては，Sawai, Komachi, and Matsumoto (2013) の学習者誤りパターンを用いた動詞の訂正候補を提案する手法や，Liu, Han, Li, Stiller, and Zhou (2010) の類義語辞書および英中対訳辞書から作成した誤りパターンを元に中国人英語学習者作文の動詞置換誤りを自動訂正する手法がある．これらの研究とは，動詞置換誤りだけを検出対象としている点が異なり，Liu らの研究に関しては，我々が学習者コーパスから誤りパターンを作成している点が異なる．

正誤情報のような正解ラベルを考慮した単語分散表現を学習する研究としては，英語学習者作のスコア予測タスクにおいて Alikaniotis, Yannakoudakis, and Rei (2016) は，各単語の作文

[1] https://github.com/kanekomasahiro/grammatical-error-detection

スコアへの影響度を学習することによって単語分散表現を構築するモデルを提案した．具体的には，スコア予測により特定の単語の作文スコアに対する影響度を学習し，作成した負例とのランキングにより文脈を学習する．この研究では平均 2 乗誤差を用いて文書レベルのスコアから単語埋め込みを学習する．一方で，我々の研究ではヒンジ損失を用いて単語レベルの 2 値誤り情報から単語埋め込みを学習する．

大規模な言語学習者コーパスである Lang-8 を用いた文法誤り訂正の研究として，統計的機械翻訳手法 (Xie, Avati, Arivazhagan, Jurafsky, and Ng 2016) とニューラルネットワークを用いた同時解析モデル (Chollampatt, Taghipour, and Ng 2016b) などがある．我々の研究では上記の研究のように Lang-8 を直接学習データとして使うのではなく，Lang-8 から文法誤りパターンを抽出し単語分散表現の学習に使用した．Lang-8 を直接学習データとして使った LSTM ベースの分類器では期待するような結果は得られなかったが，誤りパターンとして有益な情報を抽出することで文法誤り検出の精度を向上させることが可能であることを示す．

3 単語分散表現学習の従来手法：C&W Embedding

本研究での提案手法は Collobert と Weston (2008) の研究を正誤情報と誤りパターンを考慮できるように拡張している．そのため，まず初めに C&W の単語分散表現学習について説明する．C&W は局所的な文脈を元にターゲット単語の分散表現を学習するための n-gram ベースのニューラルネットワーク手法である．具体的には，サイズ n の単語列 $S = (w_1, \ldots, w_t, \ldots, w_n)$ 中のターゲット単語 w_t の表現を同じ単語列に存在する他の単語 $(\forall w_i \in S | w_i \neq w_t)$ を元に学習する．分散表現を学習するために，モデルはターゲット単語 w_t を語彙 V からランダムに選択した単語と入れ替えることにより作成した負例 $S' = (w_1, \ldots, w_c, \ldots, w_n | w_c \sim V)$ と S を比較する．そして，負例 S' ともともとの単語列 S を区別するように学習する．単語列の単語を埋め込み層でベクトルに変換し，単語列 S と負例 S' をモデルに入力する．変換されたそれぞれのベクトルを連結し入力ベクトル $x \in \mathbb{R}^{n \times D}$ とする．D は各単語の埋め込み層の次元数である．そして，入力ベクトル x は線形変換式 (1) に渡される．その後，隠れ層のベクトル i は線形変換式 (2) に渡され，出力 $f(x)$ を得る．

$$i = \sigma(W_{hi}x + b_h) \tag{1}$$

$$f(x) = W_{oh}i + b_o \tag{2}$$

W_{hi} は入力ベクトルと隠れ層の間の重み行列，W_{oh} は隠れ層のベクトルと出力層の重み行列，b_o と b_h はそれぞれバイアス，σ は要素ごとの非線形関数 tanh である．このモデルは正しい単語列 S が単語を入れ替えたことによりノイズを含む負例 S' よりランキングが高くなるようにす

ることで分散表現を学習する．そして式 (3) によって正しい単語列とノイズを含む単語列の差が少なくとも 1 になるように最適化される．

$$loss_{context}(S,\ S') = \max(0, 1 - f(x) + f(x')) \tag{3}$$

x' は負例 S' の単語 w_c を埋め込み層で変換されたベクトルに変換することで得られた値である．$1 - f(x) + f(x')$ の結果と 0 を比較し，大きい方の値を誤差とする．

4　正誤情報と誤りパターンを考慮した単語分散表現

この章では提案手法である EWE, GWE と E&GWE における単語分散表現の学習方法について詳しく述べていく．

4.1　文法誤りパターンを考慮した表現学習 (EWE)

EWE は，C&W Embedding と同じモデルで単語分散表現を学習する．ただし，負例をランダムで作成するのではなく，学習者がターゲット単語 w_t に対して誤りやすい単語 w_c と入れ替えることで作成する．こうすることで，学習者の誤りパターンを考慮して負例を作成し，ターゲット単語の分散表現が誤りやすい単語と区別されるように学習される．学習の際，w_c は条件付き確率 $P(w_c|w_t)$ によりサンプリングされる．

$$P(w_c|w_t) = \frac{|w_c, w_t|}{\sum_{w_c\prime} |w_c\prime, w_t|} \tag{4}$$

ここで w_t はターゲット単語，w'_c は w_t と対応する w_c の集合である．学習者の誤りパターンとして，学習者コーパスから抽出した誤りの訂正前の単語に対して誤りの訂正後の単語を入れ替え候補とする．図 1(a) は EWE の表現学習におけるネットワーク構造を示している．

I think I can win this ***price/*prize****.*

上の文は FCE-public のテストデータに含まれている文である．この文では，price が誤りで prize が正しい単語である．この場合，w_t は price であり w_c が prize である．今回の実験では，1 対 1 の誤りパターンのみを使用する．

一方，入れ替え候補を学習者が誤りやすい単語にすることで，入れ替え候補がない単語や頻度の少ない単語で文脈を適切に学習できないという問題が生じる．この問題を word2vec を使い事前学習したベクトルを単語それぞれの初期値とすることで解決する．文脈が既に学習されたベクトルをファインチューニングすることで，入れ替え候補がない単語や少ない単語も文脈を学習することが可能になる．

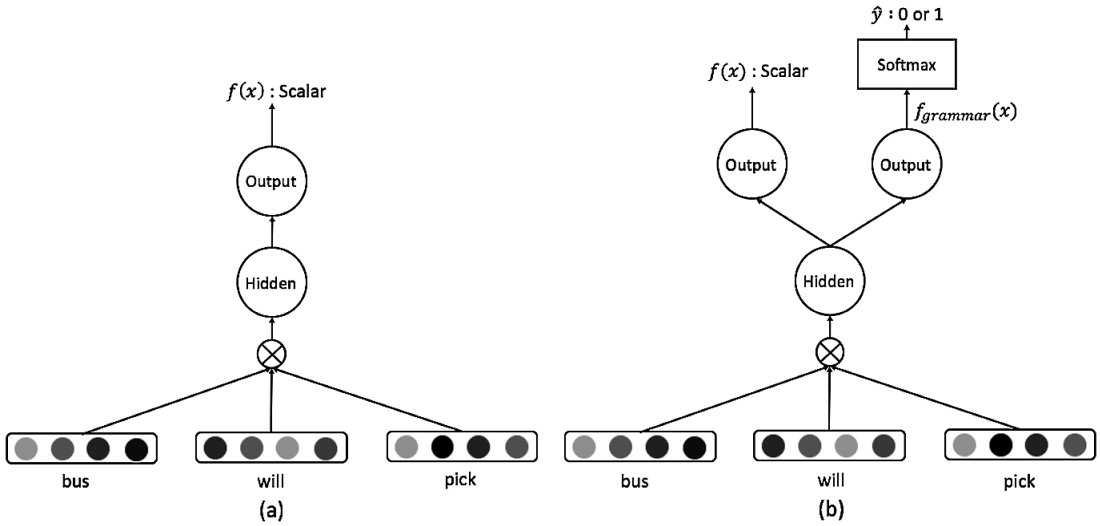


図 1　単語分散表現を学習する提案手法 (a) EWE (b) GWE の構造．両方のモデルは window サイズの単語列の単語ベクトルを結合し隠れ層に入力している．その際，EWE の出力はスカラー値であり，GWE の出力はスカラー値と単語列の中央単語のラベルである．

4.2　正誤情報を考慮した表現学習 (GWE)

Alikaniotis ら (2016) の作文スコア予測のように，C&W Embedding をそれぞれの単語の局所的な言語情報だけでなく，ターゲット単語がどれだけ単語列の正誤ラベルに貢献しているかを考慮して学習するように拡張する．図 1(b) は GWE の表現学習のネットワーク構造を示している．単語の正誤情報を分散表現に含めるために，我々は単語列の正誤ラベルを予測する出力層を追加し，式 (3) を 2 つの出力の誤差関数から構成されるように拡張する．

$$f_{grammar}(x) = W_{oh1} i + b_{o1} \tag{5}$$

$$y = softmax(f_{grammar}(x)) \tag{6}$$

$$loss_{predict}(S) = -\sum \hat{y} \cdot \log(y) \tag{7}$$

$$loss_{overall}(S, S') = \alpha \cdot loss_{context}(S,\ S') + (1-\alpha) \cdot loss_{predict}(S) \tag{8}$$

式 (5) の $f_{grammar}$ は，単語列 S のラベルの予測値である．式 (6) のように，$f_{grammar}$ に対してソフトマックス関数を用いて予測確率 y を計算する．式 (7) で交差エントロピー関数を用いて誤差 $loss_{predict}$ を計算する．ここで，$\hat{y}$ はターゲット単語の正解ラベルのベクトルである．そして，式 (8) のように 2 つの誤差を組み合わせて $loss_{overall}$ を計算する．ここで α は，2 つの誤差関数の重み付けを決定するハイパーパラメータである．

我々は，学習のための単語列の正誤情報として FCE-public にもともと付けられている正誤の

2 値ラベルを用いた．Lang-8 に関しては動的計画法を使いタグ付けを行った．GWE の負例は，C&W と同様にランダムに作成されている．

4.3　誤りパターンと正誤情報を考慮した表現学習 (E&GWE)

E&GWE は，EWE と GWE を組み合わせたモデルである．具体的には，E&GWE モデルは負例を EWE のように誤りパターンから作成し，GWE のようにスコアと正誤の予測を行う．

5　文法誤り検出の実験設定

我々は分類器と単語分散表現のための学習データとして，FCE-public 学習データ，NUCLE データと Lang-8 を用いる．そして，テストデータとして FCE-public テストデータと CoNLL-14 (Ng, Wu, Briscoe, Hadiwinoto, Susanto, and Bryant 2014) テストデータを用いる．表 1 はそれぞれのコーパスの統計情報を示している．我々の文法誤り検出では FCE-public，NUCLE，CoNLL と同様に，ある程度英文が書けるような上級レベルの英語学習者を対象にしている[2]．一方で，Lang-8 にはさまざまなレベルの英語学習者が含まれている．開発データはそれぞれ FCE-public 開発データと CoNLL-13 (Dahlmeier et al. 2013) 開発データとする．

単語の削除誤りに関しては，削除誤りの直後の単語に誤りタグを付けた．過学習を防ぐために，学習データ上で頻度が 1 の単語を未知語とした．

我々はまず単語分散表現の学習について，提案手法（EWE，GWE と E&GWE）と既存手法（word2vec と C&W）を比較する．そのために，従来手法と提案手法それぞれの単語分散表現で初期化された分類器 Bi-LSTM を FCE-public の学習データを使って学習し，文法誤り検出を行った．

***FCE-public* データセット**．FCE-public データセットは文法誤り訂正における最も有名な英語学習者コーパスの 1 つである．このコーパスには上級レベルで英語を評価する First Certificate

表 1　コーパスの統計情報

コーパス	学習データ	開発データ	テストデータ	ターゲット単語	入れ替え候補	
					トークン	タイプ
FCE-public	30,953	2,270	2,222	4,184	9,834	6,420
NUCLE & CoNLL	57,151	1,381	1,312	6,204	13,617	9,249
Lang-8	1,046,083	—	—	10,372	272,561	61,950

[2] 文法誤り検出または訂正で用いられているコーパスは上級者を対象としているものが多い．その他の代表的なコーパスとしては JFLEG (Napoles, Sakaguchi, and Tetreault 2017) と AWES (Daudaravicius, Banchs, Volodina, and Napoles 2016) がある．

in English (FCE) 試験を受けた英語学習者によって書かれた作文が含まれている．そして，文法誤りの種類に基づいてタグ付けがされている．我々は公式に分割されたコーパスを使用した：学習データ 30,953 文，テストデータ 2,720 文と開発データ 2,222 文である．FCE-public では，誤りパターンのターゲット単語として 4,184 単語が含まれている．入れ替え候補としては 9,834 トークン，6,420 タイプが含まれている．

***NUCLE* と *CoNLL*.** 提案手法による誤り検出精度の向上を FCE-public だけではなく他のデータでも検証するために，CoNLL-13 (Dahlmeier et al. 2013)，CoNLL-14 (Ng et al. 2014) の共通タスクのデータと NUS Corpus of Learner English (NUCLE) (Dahlmeier et al. 2013) を用いる．NUCLE は英語学習者であるシンガポールの大学の学生によって書かれた 1,414 個の作文が含まれている．作文のテーマとしては環境汚染や健康問題などがある．含まれている文法誤りは，英語を母語とするプロの英語教師によって訂正とアノテーションがされている．

学習データとして NUCLE の 57,151 文，開発データとして CoNLL-13 の 1,381 文そしてテストデータとして CoNLL-14 の 1,312 文を用いる．誤りパターンのターゲット単語として 6,204 単語が含まれている．入れ替え候補としては 13,617 トークン，9,249 タイプが含まれている．誤った文に対して動的計画法により正誤のタグ付けを行った．

***Lang-8* コーパス.** さらに，我々は単語分散表現の学習のために大規模な英語学習者コーパス Lang-8 を FCE-public と NUCLE に追加し使う．その際，分類器 Bi-LSTM の学習には FCE-public と NUCLE だけをそれぞれの実験で使う．Lang-8 コーパスには，英語学習者によって書かれた英文を人手でタグ付けした 100 万文以上のデータがある．Lang-8 を単語分散表現の学習に使うのは，大規模データにおける提案手法の効果について調べるためである．

Lang-8 は大規模な学習者コーパスであるが，専門家がアノテートしているわけではないため訂正されていない箇所が正用例と判断された結果訂正されていないとは限らず，単にアノテーションされていない場合もあるというノイズが含まれている (水本 他 2013)．専門家にアノテートされているコーパスであれば教育的観点から意図的に誤りを訂正しなかった場合も考えられるが，Lang-8 では意図しているのか意図していないのか区別することができない．一方で，訂正された箇所は正しい可能性が高いという特徴がある．

そのため我々は，Lang-8 を直接学習データとして用いるより誤りパターンを抽出し単語分散表現を学習したほうが文法誤り検出の精度が向上するのではないかと考え，これについても調査を行った．Lang-8 は誤りパターンのターゲット単語として 10,372 タイプが含まれている．そして，入れ替え候補として 272,561 トークン，61,950 タイプが含まれている．

5.1 動的計画法を用いたタグ付け

FCE-public はもともとデータに正誤タグが付与されているが NUCLE，CoNLL と Lang-8 には付与されていない．そのため，誤文に対しては動的計画法を用いて原文と正解文の単語のア

ライメントを取り正誤タグの付与を行う．一致にはスコア 1 を与え，不一致と対応する単語がない場合にはスコア 0 を各単語に与える．各単語単位のスコアを足し合わせアライメントのスコアを計算し，Viterbi アルゴリズムでスコアが最大となるアライメントを求める．得られたアライメント結果を元に，対応する単語が不一致またはない場合は誤用タグを付与し，一致だった場合は正用タグを付与する．

5.2 評価尺度

先行研究 (Rei and Yannakoudakis 2016) のように，我々はメインの評価手法として $F_{0.5}$ を使う．

$$F_{0.5} = (1 + 0.5^2) \cdot \frac{precision \cdot recall}{0.5^2 \cdot precision + recall} \tag{9}$$

この評価尺度は，誤り訂正タスクの CoNLL-14 の共通タスクでも用いられている (Ng et al. 2014)．$F_{0.5}$ は precision と recall の両方の組み合わせであり，precision に 2 倍の重みを割り当てている．なぜなら，誤り検出においては正確なフィードバックがカバレッジより重要であるからである (Nagata and Nakatani 2010)．

5.3 単語分散表現

先行研究 (Rei and Yannakoudakis 2016) で用いられていた単語分散表現と揃え，C&W，EWE，GWE と E&GWE の埋め込み層の次元数は 300 とし，隠れ層の次元数は 200 とした．単語分散表現の事前学習で用いられる word2vec (Chelba, Mikolov, Schuster, Ge, Brants, Koehn, and Robinson 2013) として Google News[3]からクロールしたデータから学習したモデルを用いる．単語列の長さは 3，予備実験により単語列から作成する負例は 600，式 (8) の線形補間の α は 0.03，パラメータの初期学習率は 0.001 とし，ADAM アルゴリズム (Kingma and Ba 2015) によって最適化した．そして GWE の初期値はランダムとし，EWE は事前学習された word2vec を初期値にした．

Lang-8 から誤りパターンを抽出した実験では，単語分散表現の学習のために FCE-public と Lang-8 の学習データを組み合わせて誤りパターンとした．しかしながら，Lang-8 の誤りパターンの数が FCE-public と比較して非常に多いため，我々はそれぞれの頻度の比率が 1 対 1 となるよう正規化した．単語分散表現を学習するために Lang-8 から誤りパターンを抽出する負例作成の過程は以下の通りである：

(1) 動的計画法を使い正しい文と誤った文から単語のペアを抽出する．
(2) 抽出された単語のペアが FCE-public によって抽出された語彙に含まれていた場合誤りパターンとする．

[3] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

5.4 分類器

EWE，GWE と E&GWE を Bi-LSTM を用いた文法誤り分類器の単語分散表現の初期値として使用し，入力文中の単語の正誤の予測を行う．Bi-LSTM はこのタスクにおいて Conditional Random Field (CRF) や Convolutional Neural Networks (CNN) などの他のモデルと比較して高い精度（世界最高精度）を出しており，Rei と Yannakoudakis の研究でも用いられている．彼らの Bi-LSTM は隠れ層と出力層の間に線形変換を行う追加の隠れ層が導入されている．ネットワークおよびパラメータの設定は，word2vec を初期値にした Bi-LSTM を使った先行研究 (Rei and Yannakoudakis 2016) と同じ設定である．具体的には，埋め込み層の次元数は 300 とし，隠れ層の次元数は 200 とし，隠れ層と出力層の間の隠れ層の次元数は 50 とした．初期学習率を 0.001 とした．そして，ADAM アルゴリズム (Kingma and Ba 2015) で，バッチサイズを 64 文として最適化した．

6 文法誤り検出の実験結果

6.1 FCE-public と NUCLE を用いた実験結果

表 2 は，Bi-LSTM を 2 つのベースラインで初期化したモデル（FCE + word2vec，FCE + C&W，NUCLE + word2vec と NUCLE + C&W）と提案手法を使ったモデル（FCE + EWE，FCE + GWE，FCE + E&GWE，NUCLE + EWE，NUCLE + GWE と NUCLE + E&GWE）の FCE-public と NUCLE を用いて学習した誤り検出の結果である．アスタリスクは Precision，Recall と $F_{0.5}$ のそれぞれが FCE + word2vec（(R&Y 2016) の再実装）または NUCLE + word2vec

表 2 上表は FCE-public だけ，下表は NUCLE だけで学習された Bi-LSTM と単語分散表現のそれぞれのテストデータにおける誤り検出精度

Bi-LSTM + embeddings	Precision	Recall	$F_{0.5}$
FCE + word2vec (R&Y, 2016)	46.1	28.5	41.1
FCE + word2vec （(R&Y, 2016) の再実装）	45.8 ± 0.1	27.8 ± 0.4	40.5 ± 0.3
FCE + C&W	45.1 ± 0.3	26.7 ± 0.4	39.6 ± 0.3
FCE + EWE	$46.1 \pm 0.1\star$	$28.0 \pm 0.1\star$	$40.8 \pm 0.1\star$
FCE + GWE	$46.5 \pm 0.1\star$	$28.3 \pm 0.4\star$	$41.2 \pm 0.2\star$
FCE + E&GWE	$\mathbf{46.7} \pm 0.1\star$	$\mathbf{28.6} \pm 0.1\star$	$\mathbf{41.4} \pm 0.1\star$
NUCLE + word2vec	25.1 ± 0.3	24.1 ± 1.1	24.9 ± 0.1
NUCLE + C&W	22.9 ± 0.2	22.6 ± 0.6	22.9 ± 0.1
NUCLE + EWE	$25.7 \pm 0.4\star$	$25.3 \pm 0.2\star$	$25.6 \pm 0.3\star$
NUCLE + GWE	$25.8 \pm 0.3\star$	$25.6 \pm 0.1\star$	$25.8 \pm 0.2\star$
NUCLE + E&GWE	$\mathbf{26.0} \pm 0.1\star$	$\mathbf{26.0} \pm 0.4\star$	$\mathbf{26.0} \pm 0.1\star$

に対して有意水準 0.05 で有意差があることを示す．FCE-public で学習したモデルは FCE-public のテストデータを使い，NUCLE で学習したモデルは CoNLL-14 (Ng et al. 2014) のテストデータを使い評価した．FCE + word2vec に関しては 2 つのモデルがある．FCE + word2vec (R&Y 2016) は先行研究 (Rei and Yannakoudakis 2016) で報告されている値である．FCE + word2vec（(R&Y 2016) の再実装）は先行研究の再実装の結果である．NUCLE + E&GWE と FCE + E&GWE は，それぞれのコーパスの EWE と GWE を組み合わせためモデルである．

まず，ベースラインと提案手法（FCE + EWE，FCE + GWE，NUCLE + EWE と NUCLE + GWE）を比較すると，一貫して精度が向上していることがわかる．このことから，誤りパターンと正誤情報を考慮する提案手法が文法誤り検出では有効であることがわかる．さらに，2 つの手法を組み合わせた FCE + E&GWE と NUCLE + E&GWE のほうが EWE と GWE それぞれを単体で使ったモデルより高い精度を出している．このことから，EWE と GWE を組み合わせることが有効であることがわかる．

6.2 Lang-8 を用いた実験結果

Lang-8 を直接学習データとして用いるより誤りパターンを抽出し単語分散表現を学習したほうが文法誤り検出の精度が向上につながることを検証するために，以下の 2 つの設定で比較する：(1) FCE-public と NUCLE それぞれの誤りパターンに Lang-8 から抽出した誤りパターンを追加する．そして，誤りパターンを用いて学習された単語分散表現によって初期化された Bi-LSTM を FCE-public と NUCLE のそれぞれだけを使い学習する（FCE + EWE-L8，FCE + E&GWE-L8，NUCLE + EWE-L8 と NUCLE + E&GWE-L8, 表 3）；(2) word2vec で初期化 Bi-LSTM の学習データとして FCE-public と NUCLE のそれぞれに直接 Lang-8 を追加する（FCE&L8 + W2V と NUCLE&L8 + W2V, 表 3）．表 3 は Lang-8 を学習データに追加した文法誤り検出の結果である．アスタリスクは Precision，Recall と $F_{0.5}$ のそれぞれが FCE + word2vec（(R&Y 2016) の再実装）または NUCLE + word2ve に対して有意水準 0.05 で有意差があることを示す．その際，我々はウィルコクソンの符号順位検定 ($p \leq 0.05$) を 5 回行った．

表 3　大規模な Lang-8 コーパスを追加で使い Bi-LSTM か単語分散表現のどちらかを学習した場合の FCE-public または NUCLE のテストデータにおける誤り検出精度

Bi-LSTM + embeddings	Precision	Recall	$F_{0.5}$
FCE&L8 + word2vec	12.3 ± 2.6	32.8 ± 2.2	14.0 ± 2.6
FCE + EWE-L8	50.5 ± 3.4⋆	**30.1** ± 1.2⋆	44.4 ± 2.7⋆
FCE + E&GWE-L8	**50.8** ± 3.6⋆	30.0 ± 1.2⋆	**44.6** ± 2.8⋆
NUCLE&L8 + word2vec	18.5 ± 0.1	18.6 ± 0.1	18.5 ± 0.1
NUCLE + EWE-L8	28.3 ± 0.2⋆	28.2 ± 0.2⋆	28.3 ± 0.1⋆
NUCLE + E&GWE-L8	**29.0** ± 0.1⋆	**28.8** ± 0.1⋆	**28.9** ± 0.1⋆

表 2 と表 3 から，Precision，Recall と $F_{0.5}$ に関してそれぞれの手法を以下のようにランク付けすることができる：(FCE, NUCLE) + E&GWE-L8 > (FCE, NUCLE) + EWE-L8 > (FCE, NUCLE) + E&GWE > (FCE, NUCLE) + GWE > (FCE, NUCLE) + EWE > (FCE, NUCLE) + word2vec > (FCE, NUCLE) + C&W．文法誤り検出において誤りパターンと正誤情報を考慮することで一貫して精度が向上している．このことから，提案手法が文法誤り検出では有効であることがわかる．そして，我々の提案手法は Lang-8 コーパスを使うことなく先行研究と比較して統計的有意差がある．我々の提案手法は FCE-public において全ての評価尺度において世界最高精度である Rei と Yannakoudakis の先行研究を上回った．そして，FCE&L8 + word2vec と FCE + EWE-L8 の結果から，直接分類器の学習データとして使うより誤りパターンとして抽出し使うほうが良いことがわかる．これは Lang-8 の正しい文にノイズが多く含まれているためと考えられる．Lang-8 のような専門家がアノテートしたわけではないノイズを多く含むコーパスを直接学習データとして使わず，大規模である利点を活かし多様な誤りパターンを抽出するために使う．そして，FCE-public や NUCLE のような小規模ではあるが専門家がアノテートしたコーパスを専門家のアノテーションの特徴を捉えた質の高い学習をするために使う．これにより，それぞれのコーパスの利点を活かした学習を行うことができていると言える．そして，上記の実験から GWE と組み合わせることでさらに精度が向上することがわかる．

7 考察

それぞれの手法ごとにどのような違いがあるかを調べるために，誤りタイプごとの正解数について見ていく．表 4 は，FCE-public のテストデータにおけるそれぞれのモデルの誤りタイプごとの正解数と正解率を示している．ここでは，比較対象のモデル同士でもっとも誤りタイプの正解数の差が大きかった 2 つの誤りタイプを上げている．従来手法と提案手法，Lang-8 ありの提案手法と Lang-8 なしの提案手法を比較した．誤りタイプは FCE-public にもともと付与さ

表 4 誤りタイプごとの正解数と正解率

	誤りタイプ	動詞置換誤り	限定詞欠損誤り	名詞置換誤り	名詞語形誤り
(a)	FCE + word2vec	56 (42.74%)	**48** (42.85%)	26 (33.76%)	9 (28.12%)
	FCE + C&W	53 (40.45%)	46 (41.07%)	24 (31.16%)	7 (21.87%)
(b)	FCE + EWE	60 (45.80%)	37 (33.03%)	29 (37.66%)	12 (37.50%)
	FCE + GWE	62 (47.32%)	43 (38.39%)	29 (37.66%)	11 (34.37%)
	FCE + E&GWE	64 (48.85%)	40 (35.71%)	31 (40.25%)	14 (43.75%)
(c)	FCE + EWE-L8	66 (50.38%)	36 (32.14%)	37 (48.05%)	**19** (59.37%)
	FCE + E&GWE-L8	**67** (51.14%)	40 (35.71%)	**39** (50.64%)	18 (56.25%)
	誤りの合計数	131	112	77	32

れていた正解ラベルを用いる[4]．

まず，従来手法と提案手法で最も正解数が異なる，動詞置換誤りと限定詞欠損誤りについて分析する（表 4 の (a) と (b)）．動詞置換誤りに関しては提案手法の正解数が多い．一方で，限定詞欠損誤りに関してはベースラインである FCE + word2vec と FCE + C&W のほうが正解数が多い．提案手法のほうが限定詞欠損誤りの正解数が少ないのは，誤りパターンが単語ペアを抽出し作成されており，単語が欠落している誤りが含まれていないためと考えられる．1-gram ベースの誤りパターンを用いた単語分散表現では入れ替え誤りに特化した学習を行うため，誤りパターンに含まれていないような他の誤りを文脈を手がかりに学習することは難しいと考えられる．

次に，我々は Lang-8 から抽出した誤りパターンを使うことによる影響について調べる（表 4 の (b) と (c)）．FCE + EWE と FCE + EWE-L8 は名詞置換誤りと名詞語形誤りにおいて最も正解数が異なる．名詞置換誤りとは suggestion と advice のような誤りであり，名詞語形誤りとは time と times のような誤りである．FCE + EWE-L8 は，名詞置換誤りと名詞語形誤りの両方で正解数が多い．理由としては，名詞置換誤りと名詞語形誤りともに Lang-8 に含まれている誤りパターンの数が FCE-public と比較して 10 倍ほど多いためと考えられる．

表 5 は従来手法である FCE + word2vec と最も精度の高い提案手法である FCE + E&GWE-L8 のテストデータに対する検出例を示している．表 5(a) は名詞置換誤りの検出例を示している[5]．FCE + word2vec は名詞置換誤りを検出できていないが，FCE + E&GWE-L8 は名詞置換誤りを検出することができている．名詞語形誤りに関しては表 5(b) で示されている．ここで，FCE + word2vec は誤りを 1 つも検出することができていない．一方で，FCE + E&GWE-L8 は名詞

表 5　FCE + word2vec と FCE + E&GWE-L8 を用いた誤り検出の例

	Bi-LSTM + embeddings	検出結果
(a)	Gold	The bus will pick you up right at your hotel *entrance*.
	FCE + word2vec	The bus will pick you up right at your hotel entery.
	FCE + E&GWE-L8	The bus will pick you up right at your hotel **entery**.
(b)	Gold	There are shops which *sell clothes*, *food*, and books …
	FCE + word2vec	There are shops which sales cloths, foods, and books …
	FCE + E&GWE-L8	There are shops which sales cloths, **foods**, and books …
(c)	Gold	All the buses and *the MTR* have air-condition.
	FCE + word2vec	All the buses and **MTR** have air-condition.
	FCE + E&GWE-L8	All the buses and MTR have air-condition.

正解をイタリック体とし検出結果を太字で表す．

[4] 誤りのタグ付け方法または種類については付録に掲載した．
[5] 実際の FCE-public では，最初に entery は entry に訂正され，その後 entrance に訂正されている．

語形誤りを検出することができている．これは，Lang-8 から抽出した誤りパターンに含まれていたためと考えられる．sale と cloths の検出は両方のモデルが失敗している．しかし，前者は構文的情報を必要とし，後者は常識を必要とするため誤り検出が難しいと考えられる．表 5(c) では，FCE + W2V は限定詞欠損誤りの検出に成功したが，FCE + E&GWE-L8 は検出に失敗した．この結果は限定詞欠損誤りと同様に誤りパターンの構造上，挿入誤りを適切に学習できていないことを示している．

図 2 は，学習データ内で高頻度な誤りの単語分散表現（FCE + word2vec と FCE + E&GWE-L8）を t-SNE を用いて可視化した図である．我々は誤りとして出現頻度が多い前置詞と動詞をいくつかプロットした．頻度を元に可視化したのは，高頻度で誤っているほど学習データに多く出現するため単語分散表現がよく学習され違いがわかりやすくなると考えたからである．ここでは誤りとして高頻度な単語を誤りやすい，低頻度な単語を誤りにくい単語としている．学習者が誤りにくい単語は FCE + E&GWE-L8 と FCE + word2vec で似たような位置として学習されている．一方で，学習者が誤りやすい単語に関しては誤りの出現頻度に比例して FCE + E&GWE-L8 と FCE + word2vec で離れた位置として学習されていることがわかる．例えば，under や walk のようにあまり誤りとして出現しない単語は FCE + word2vec の近くに位置している．一方で，was や at のようによく誤られる単語は FCE + E&GWE + L8 の点は FCE + word2vec と比較してより遠くに移動している．そして，この図中のほとんどすべての単語が上に移動しているので，上方向に移動する距離が誤りやすさに対応していると推測される．この

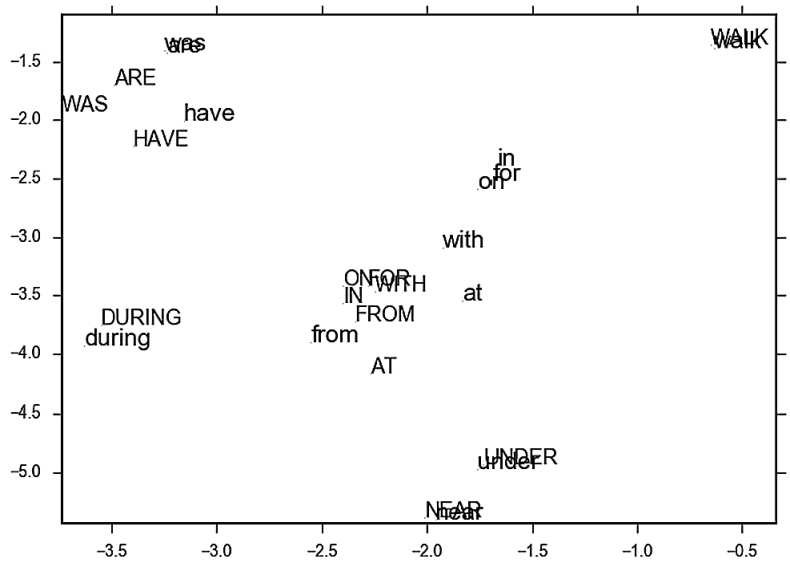


図 2 FCE + word2vec と FCE + E&GWE-L8 によって学習された単語分散表現の t-SNE による可視化．大文字が FCE + word2vec の単語であり，小文字が FCE + E&GWE-L8 の単語である．

可視化は学習者による誤りに対する分析に使うことができる．

8 まとめ

本稿で我々は，文法誤り検出のための正誤情報と文法誤りパターンを考慮した単語分散表現の学習手法を提案した．その結果，FCE-public と NUCLE の 2 つのコーパスにおいて文法誤り検出の精度向上を行うことができた．そして，提案手法で単語分散表現を初期化した Bi-LSTM モデルを使い FCE-public データセットにおいて世界最高精度を達成した．学習者コーパスによって学習された単語分散表現は正しいフレーズと誤ったフレーズを区別することが可能である．さらに我々は，Lang-8 コーパスを用いた追加の実験を行った．その結果，我々は誤りパターンを抽出して学習するほうが直接 Lang-8 コーパスを分類器の学習データに追加するより良いことがわかった．そして，いくつかの典型的な誤りに対して検出結果を分析し，学習された単語分散表現の特徴を明らかにした．

今回の提案手法では，Lang-8 の添削者を一律に誤りの見逃しなどのノイズを含む可能性があるとしている．一方で，Lang-8 の添削者中でも専門家のように質の高い添削を行っている添削者もおり，Lang-8 にも学習データとして直接使うことが可能な文が多く含まれていると考えられる．そのため，Lang-8 の添削者の評価や添削数などのメタ情報を活用した文法誤り検出などが考えられる．また，学習者の母語などのメタ情報を活用した文法誤り訂正の研究 (Chollampatt, Hoang, and Ng 2016a) が報告されている[6]．そこで，学習者の第二言語習得過程 (Settles, Brust, Gustafson, Hagiwara, and Madnani 2018) を考慮した文法誤り訂正にも取り組んでいきたい．

謝　辞

本研究は JSPS 科研費 JP16K16117 の助成を受けたものである．

参考文献

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). “Automatic Text Scoring Using Neural Networks.” In *ACL*, pp. 715–725.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). “One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling.” *arXiv preprint arXiv:1312.3005*.

[6] ネイティブ言語と学習言語の文法的な近さによって，学習時に文法的に近いため文法を混同する問題や文法の大きな違いによる問題などが発生する．

Chollampatt, S., Hoang, D. T., and Ng, H. T. (2016a). "Adapting Grammatical Error Correction Based on the Native Language of Writers with Neural Network Joint Models." In *EMNLP*, pp. 1901–1911.

Chollampatt, S., Taghipour, K., and Ng, H. T. (2016b). "Neural Network Translation Models for Grammatical Error Correction." In *IJCAI*, pp. 2768–2774.

Collobert, R. and Weston, J. (2008). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning." In *ICML*, pp. 160–167.

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). "Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English." In *BEA@NAACL-HLT*, pp. 22–31.

Daudaravicius, V., Banchs, R. E., Volodina, E., and Napoles, C. (2016). "A Report on the Automatic Evaluation of Scientific Writing Shared Task." In *BEA Shared Task*, pp. 53–62.

Han, N.-R., Chodorow, M., and Leacock, C. (2006). "Detecting Errors in English Article Usage by Non-native Speakers." *Natural Language Engineering*, pp. 115–129.

Kingma, D. and Ba, J. (2015). "Adam: A Method for Stochastic Optimization." *ICLR*.

Kochmar, E. and Briscoe, T. (2014). "Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics." In *COLING*, pp. 1740–1751.

Liu, X., Han, B., Li, K., Stiller, S. H., and Zhou, M. (2010). "SRL-based Verb Selection for ESL." In *EMNLP*, pp. 1068–1076.

水本智也，小町守，永田昌明，松本裕治 (2013). 日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得. 人工知能学会論文誌, **28** (5), pp. 420–432.

Nagata, R. and Nakatani, K. (2010). "Evaluating Performance of Grammatical Error Detection to Maximize Learning Effect." In *COLING*, pp. 894–900.

Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). "JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction." In *EACL*, pp. 229–234.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). "The CoNLL-2014 Shared Task on Grammatical Error Correction." In *CoNLL Shared Task*, pp. 1–14.

Nicholls, D. (2003). "The Cambridge Learner Corpus: Error Coding and Analysis for Lexicography and ELT." *Corpus Linguistics*, pp. 572–581.

Rei, M. and Yannakoudakis, H. (2016). "Compositional Sequence Labeling Models for Error Detection in Learner Writing." In *ACL*, pp. 1181–1191.

Sawai, Y., Komachi, M., and Matsumoto, Y. (2013). "A Learner Corpus-based Approach to Verb Suggestion for ESL." In *ACL*, pp. 708–713.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., and Madnani, N. (2018). "Second Language

Acquisition Modeling." In *BEA Shared Task*, pp. 56–65.

Tetreault, J. R. and Chodorow, M. (2008). "The Ups and Downs of Preposition Error Detection in ESL Writing." In *COLING*, pp. 865–872.

Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., and Ng, A. Y. (2016). "Neural Language Correction with Character-based Attention." *arXiv preprint arXiv:1603.09727*.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). "A New Dataset and Method for Automatically Grading ESOL Texts." In *ACL*, pp. 180–189.

付録

7 章の考察で取り上げた FCE-public で用いられている誤りタイプ (Nicholls 2003) について説明する．2 つのタグから誤りタイプは構成されている．1 つ目のタグは誤りの種類を表しており，2 つ目のタグは対象単語のクラスを表す．2 つのタグを組み合わせることで誤りタイプを表現する．例えば，動詞置換誤りであれば 1 つ目のタグが置換の R，2 つ目のタグは動詞の V，この 2 つを組み合わせた RV として表す．

一般的な誤り（1 つ目のタグ）

F	語形誤り (wrong Form used)
M	欠損 (something Missing)
R	置換 (word or phrase needs Replacing)
U	不必要 (word or phrase is Unnecessary)
D	派生誤り (word is wrongly Derived)

単語クラス（2 つ目のタグ）

A	照応 (Anaphoric)
C	接続詞 (Conjunction)
D	限定詞 (Determiner)
J	形容詞 (Adjective)
N	名詞 (Noun)
Q	数量詞 (Quantifier)
T	前置詞 (Preposition)
V	動詞 (Verb)
Y	副詞 (Adverb)

記号誤り（誤りの種類 + P）

MP	記号欠損 (punctuation Missing)
MP	記号置換 (punctuation needs Replacing)
UP	記号不必要 (Unnecessary punctuation)

一致誤り（AG + 単語クラス）

AGA 照応一致誤り (Anaphoric agreement error)
AGD 限定詞一致誤り (Determiner agreement error)
AGN 名詞一致誤り (Noun agreement error)
AGV 動詞一致誤り (Verb agreement error)

可算名詞誤り（C + 単語クラス）

CN 可算名詞誤り (countability of Noun error)
CQ 可算名詞による数量詞誤り (wrong Quantifier because of noun countability)
CD 可算名詞による限定詞誤り (wrong Determiner because of noun countability)

空似言葉 (False friend)（FF + 単語クラス）

全ての空似言葉は FF でタグ付けされる．必要な単語クラスは A，C，D，J，N，Q，T，V と Y のいずれかである．この誤りは空似言葉を扱っていることが確実な場合にのみ使用される．その他の場合は置換 R が使われる．

その他の誤り

AS 項構造誤り (incorrect Argument Structure)
CE 複合誤り (Compound Error)
CL コロケーション誤り (CoLlocation error)
ID 慣用句誤り (IDiom error)
IN 名詞複数形の形成誤り (Incorrect formation of Noun plural)
IV 動詞の不正な活用 (Incorrect Verb inflection)
L 不適切なレジスター (inappropriate register)
S スペリング誤り (Spelling error)
SA アメリカ英語 (American Spelling)
SX スペル混同誤り (Spelling confusion error)
TV 動詞の時制誤り (wrong Tense of Verb)
W 語順誤り (incorrect Word order)
X 否定形誤り (incorrect formation of negative)

CN は，学習者が意図された意味で利用できない名詞形を使用したことを表す．例えば，the country's natural beauties や two transports などである．一方で，可算または不可算に関わらず間違った形が使用された場合，その誤りは FN とする．例えば，vacation と vacations である．

AS（項構造誤り）は MT（前置詞の欠損，例えば he explained me）または UT（不必要な前置詞，例えば he told to me）では網羅できない誤りを対象とする．AS は，特に第 4 文型をとる動詞に対して使用される．例えば，it caused trouble to me は it caused me trouble と 1 つの誤りとして訂正する．

CE（複合誤り）は，意図した意味が推定できない複数の誤りや単語の集合をカバーする包括的な誤りである．この誤りを用いることで，学習者の誤りに関する有用な情報をほとんど得ら

れない箇所を除外することができる．

SX（スペル混同誤り）は，スペルの混同の可能性をカバーする．例えば to と too，their と there や weather と whether などである．

略歴

金子　正弘：2016 年北見工業大学工学部情報システム工学科卒業．同年，首都大学東京システムデザイン研究科博士前期課程に進学．2018 年博士前期課程修了．同年，首都大学東京システムデザイン研究科博士後期課程に進学．

堺澤　勇也：2015 年首都大学東京システムデザイン学部システムデザイン学科情報通信システムコース卒業．同年，同大学院システムデザイン研究科博士前期課程に進学．2017 年博士前期課程修了．現在株式会社ジャストシステム勤務．

小町　　守：2005 年東京大学教養学部基礎科学科科学史・科学哲学分科卒業．2007 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．2008 年より日本学術振興会特別研究員 (DC2) を経て，2010 年博士後期課程修了．博士（工学）．同年より同研究科助教を経て，2013 年より首都大学東京システムデザイン学部准教授．大規模なコーパスを用いた意味解析および統計的自然言語処理に関心がある．情報処理学会，人工知能学会，言語処理学会，ACL 各会員．

（2017 年 12 月 9 日　受付）
（2018 年 2 月 26 日　再受付）
（2018 年 4 月 23 日　再々受付）
（2018 年 5 月 15 日　採録）