学修番号 16890509

修士論文

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

金子　正弘

2018 年 3 月 30 日

首都大学東京大学院
システムデザイン研究科 情報通信システム学域

金子　正弘

審査委員：

小町 守 准教授　（主指導教員）

山口 亨 教授　（副指導教員）

高間 康史 教授　（副指導教員）

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*

金子　正弘

修論要旨

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がある．彼らの手法は，言語学習者コーパスがネイティブの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，先行研究で用いられている文法誤り検出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．これは，I would like to go **on/in** summer. のように前置詞誤りを含む文と正しい文が判別器に類似したベクトルの入力として扱われてしまう問題がある．

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ことでこの問題を解決する 3 つの手法を提案する．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

*首都大学東京大学院 システムデザイン研究科 情報通信システム学域 修士論文, 学修番号 16890509, 2018 年 3 月 30 日.

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 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．さらに，従来手法の C&W や word2vec では文法的妥当性が高いフレーズ対と低いフレーズ対の類似度が高くなるように学習してしまうが，提案手法である EWE, GWE と E&GWE は文法的妥当性が高いフレーズ対では類似度が高くなり，文法的妥当性が低いフレーズ対では類似度が低くなるように学習することを示した．このことから，EWE, GWE と E&GWE は文脈上の関連を維持しながら，文法誤りを含むフレーズ対と正しいフレーズ対の類似度が低くなるように学習することがわかる．

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) コーパスにおいて世界最高精度を達成した．
- FCE-public と NUS Corpus of Learner English (NUCLE) データに Lang-8 から抽出した誤りパターンを追加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．
- 誤りタイプごとの提案手法の有効性について分析を行った．

- 我々が提案した単語分散表現の学習方法は，正しい単語と誤ったフレーズ対を区別することができることを示した．
- 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可視化し分析した．
- 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．

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．次に，第 4 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 5 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評価を行い，最後に第 6 章でまとめる．

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

Masahiro Kaneko

Abstract

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 so that classifiers treat erroneous and correct words as similar inputs. We address the problem of contextual information by considering learner errors. Specifically, we propose two models: one model that employs grammatical error patterns and the other model that considers grammaticality of the target word. 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.

*Master's Thesis, Department of Information and Communication Systems, Graduate School of System Design, Tokyo Metropolitan University, Student ID 16890509, March 30, 2018.

目次

図目次

第 1 章　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 [1] 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 [1] 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．これは，下記の例文のように前置詞誤りを含む文と正しい文が判別器に類似したベクトルの入力（表 1 の word2vec と C&W の列）として扱われてしまう問題がある．

I would like to go ***on/in*** *summer.*

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ことでこの問題を解決する 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．

表 1.1: フレーズ対の cos 類似度

フレーズ対	word2vec	C&W	EWE	GWE	E&GWE
in summer & on summer	0.84	0.75	0.64	0.58	0.54
in summer & in spring	0.84	0.77	0.90	0.80	0.88
in summer & in English	0.40	0.46	0.36	0.25	0.30
on summer & on spring	0.85	0.71	0.82	0.76	0.80

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．

表 1.1 は，word2vec [2]，C&W [3] ，EWE，GWE と E&GWE それぞれのモデルのフレーズ対の cos 類似度を示している．フレーズ対の類似度はそれぞれの単語対の単語ベクトルの平均ベクトルの類似度によって計算した．in summer と on summer は前置詞誤りの関係であり，word2vec と C&W では類似度の高いベクトルとして学習されてしまっているが，EWE，GWE と E&GWE では類似度が低くなるように学習されている．そして，文法的妥当性が高いフレーズ対ではすべての提案モデルで類似度が高くなっている．一方で，文法的妥当性の低いフレーズ対では類似度が低くなっている．これらのことから，EWE，GWE と E&GWE は文脈上の関連を維持しながら，文法誤りを含むフレーズ対と正しいフレーズ対の類似度が低くなるように学習されていることが分かる．

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 [4] 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 [5] において世界最高精度を達成した．

- FCE-public と NUCLE データ [6] に Lang-8 から抽出した誤りパターンを追加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．
- 我々が提案した単語分散表現の学習方法は，正しいフレーズと誤ったフレーズ対を区別することができる．
- 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*．

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．次に，第 4 章では FCE-public と NUCLE の評価セットである CoNLL-14 データセット [7] を使い提案手法を評価する．第 5 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評価を行い，最後に第 6 章でまとめる．

*https://github.com/kanekomasahiro/grammatical-error-detection

第 2 章　先行研究

文法誤り検出の研究の多くは前置詞の正誤 [8]，冠詞の正誤 [9] や形容詞と名詞の対の正誤 [10] 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，特定のタイプではなく文法誤り全般に取り組んだ文法誤り検出の研究は少ない．Rei と Yannakoudakis [1] 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 ら [11] 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 ら [12] 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選択誤りを自動訂正する手法がある．これらの研究とは，動詞選択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 ら [13] は，各単語の作文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．

文法誤り検出のための負例作成に関しては、Liu ら [14] の研究がある．ラベル付けされていないコーパスから負例を作成することで文法誤り検出を学習する．ただし，この研究は負例を用いた誤文作成が目的である。さらに，ルールベースを使い負例を作成している点も異なる．ルールベースは網羅性が欠点である．一方で，単語列に対して負例を作成している点が我々の研究と同じである．

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 [15] とニューラルネットワークを用いた同時解析モデル [16]

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．

第 3 章　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．これらのモデルは，既存の単語分散表現の学習アルゴリズムである C&W Embedding [3] 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，我々は初めに C&W の単語分散表現学習について説明する．そして，その次に提案手法である Bi-LSTM を使った文法誤り検出のための単語分散表現学習の具体的な方法について述べていく．

3.1　C&W Embedding

Collobert と Weston [3] の研究 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を提案した．具体的には，サイズ n の単語列 $S = (w_1, ..., w_t, ..., w_n)$ 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 $S' = (w_1, ..., w_c, ..., w_n | w_c \sim V)$ 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 は式 3.11 のように線形変換される．その後，隠れ層のベクトル i は式 3.12 のように線形変換され，出力 $f(x)$ を得る．

$$i = \sigma(W_{hi}x + b_h) \tag{3.11}$$

$$f(x) = W_{oh}i + b_o \tag{3.12}$$

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 3.13

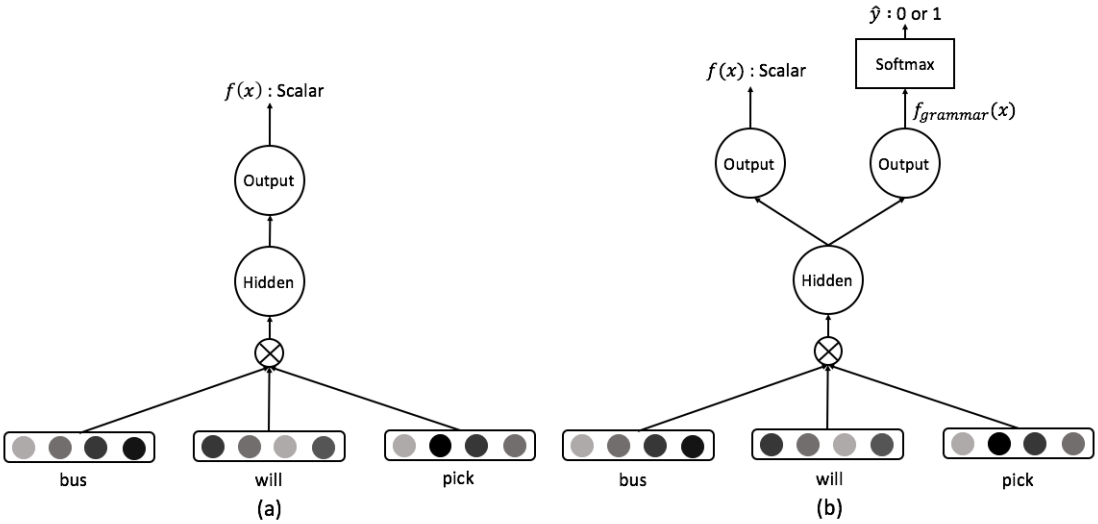

図 3.1: 単語分散表現を学習する提案手法 (a) EWE (b) GWE の構造．両方のモデルは window サイズの単語列の単語ベクトルを結合し隠れ層に入力している．その際，EWE の出力はスカラー値であり，GWE の出力はスカラー値と単語列の中央単語のラベルである．

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．

$$loss_{context}(S,\ S') = \max(0, 1 - f(x) + f(x')) \tag{3.13}$$

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．

3.2　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．

$$P(w_c|w_t) = \frac{|w_c, w_t|}{\sum_{w_{c'}} |w_{c'}, w_t|} \tag{3.21}$$

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 3.1(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．

The bus will pick you up right at your hotel ***entery/*entrance****.*

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, entery が誤りで entrance が正しい単語である．この場合，w_t は entrance であり w_c が entery 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．

3.3 正誤情報を考慮した表現学習 (GWE)

Alikaniotis ら [13] 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 3.1(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 3.13 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．

$$f_{grammar}(x) = W_{oh1} i + b_{o1} \tag{3.31}$$

$$y = softmax(f_{grammar}(x)) \tag{3.32}$$

$$loss_{predict}(S) = -\sum \hat{y} \cdot \log(y) \tag{3.33}$$

$$loss_{overall}(S, S') = \alpha \cdot loss_{context}(S,\ S') + (1 - \alpha) \cdot loss_{predict}(S) \quad (3.34)$$

式 3.31 の $f_{grammar}$ は，単語列 S のラベルの予測値である．式 3.32 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 3.33 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 3.34 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 と NUCLE 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．

3.4 文法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．

第 4 章　分類器：Bidirectional LSTM (Bi-LSTM)

我々は英語の文法誤り検出のすべての実験で分類器として Bi-LSTM [1] を用いる．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）を出した．LSTM は以下のように計算される：

$$i_t = \sigma(W_{ie}e_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \tag{4.01}$$
$$f_t = \sigma(W_{fe}e_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \tag{4.02}$$
$$c_t = i_t \odot g(W_{ce}e_t + W_{ch}h_{t-1} + b_c) + f_t \odot c_{t-1} \tag{4.03}$$
$$o_t = \sigma(W_{oe}e_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o) \tag{4.04}$$
$$h_t = o_t \odot h(c_t) \tag{4.05}$$

ここで，e_t は単語 w_t のベクトルであり，W_{ie}, W_{fe}, W_{ce} と W_{oe} は重み行列である．b_i, b_f, b_c と b_o はそれぞれバイアスである．LSTM は入力情報を制御するために入力ゲート i_t，メモリセル c_t，忘却ゲート f_t と出力ゲート o_t を持つ．g と h はシグモイド関数であり，α は tanh である．そして，$\odot$ はアダマール積である．

我々は，図 4.1 のように LSTM を両方向に拡張する．右方向と左方向の両方向から単語分散表現 e_i を LSTM に入力する．

$$o_t = W_{oh}(h_t^L \otimes h_t^R) + b_o \tag{4.06}$$

Bi-LSTM モデルは，それぞれの単語分散表現 w_t を隠れベクトル h_t^L と h_t^R にマッピングする．h_t^L と h_t^R はそれぞれ左から右方向 LSTM と右から左方向 LSTM の隠れベクトルを表してる．$\otimes$ は連結である．そして，W_{oh} は重み行列であり b_o はバイアスとする．先行研究と同様に，我々はさらに隠れ層と出力層の間に線形変換を行う追加の隠れ層を導入する．

出力 o_t に対してソフトマックス関数を適用することで予測ラベルの確率 y_t を得る．正解ラベルと予測ラベルの確率 y_t をもとに交差エントロピー関数を使い誤差を算出し学習する．

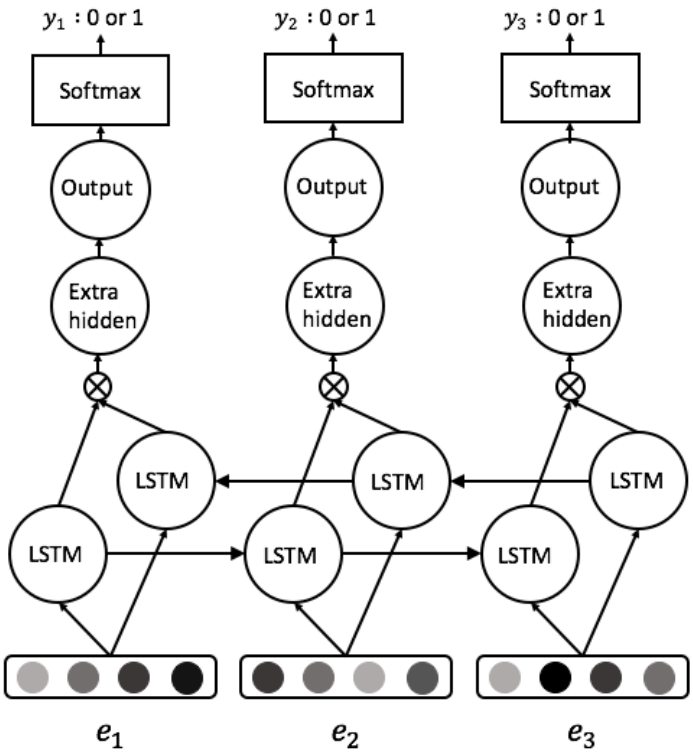

図 4.1: Bidirectional LSTM ネットワーク．単語ベクトル e_i が隠れ層に入力されそれぞれの単語のラベルを予測する．

第 5 章　英語学習者コーパスにおける文法誤り検出

5.1　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，評価データとして FCE-public テストデータと CoNLL-14 [7] テストデータを用いる．開発データはそれぞれ FCE-public 開発データと CoNLL-13 [6] 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 [6], CoNLL-14 [7] の共通タスクのデータと NUS Corpus of Learner English (NUCLE) [6] 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．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解のタ

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，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 [4]．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た．

この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 5.1b）; (2) word2vec で初期化された 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 5.1b）．

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学習データ（FCE-public か NUCLE）によって作成された語彙に含まれていた場合誤りパターンとする．

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．

FCE+EWE-L8 と FCE+E&GWE-L8 の実験では，単語分散表現の学習のため

に Lang-8 と FCE-public 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．

5.2 評価尺度

先行研究 [1] のように，我々はメインの評価手法として $F_{0.5}$ を使う．

$$F_{0.5} = (1 + 0.5^2) \cdot \frac{precision \cdot recall}{0.5^2 \cdot precision + recall} \tag{5.21}$$

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 [7]．$F_{0.5}$ 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 [17]．

5.3 単語分散表現

先行研究 [1] 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 [2] として Google News*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 [18] 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．

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．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 [1] と

*https://github.com/mmihaltz/word2vec-GoogleNews-vectors

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 [18] で，バッチサイズを 64 文として最適化した．

5.5 実験結果

表 5.1a 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を示している．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 [7] 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 [1] 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．表 5.1b は大規模なコーパスである Lang-8 を学習データに追加した文法誤り検出の結果を示している．そして，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 5.1a と 5.1b から，FCE-public と NUCLE における 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先行研究 [1] 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．さらに，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．

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±0.1⋆	28.0±0.1⋆	40.8±0.1⋆
FCE + GWE	46.5±0.1⋆	28.3±0.4⋆	41.2±0.2⋆
FCE + E&GWE	**46.7**±0.1⋆	**28.6**±0.1⋆	**41.4**±0.1⋆
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±0.4⋆	25.3±0.2⋆	25.6±0.3⋆
NUCLE + GWE	25.8±0.3⋆	25.6±0.1⋆	25.8±0.2⋆
NUCLE + E&GWE	**26.0**±0.1⋆	**26.0**±0.4⋆	**26.0**±0.1⋆

(a) 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&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⋆

(b) 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.

表 5.1: Bi-LSTM を使った誤り検出の結果．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) に対して有意水準 0.05 で有意差があることを示す.

第 6 章　考察

表 6.1 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を示している．誤りタイプは FCE-public の正解ラベルを用いる．

まず，従来手法と提案手法で正解数が大きく異なる，動詞誤りと無冠詞の誤りについて分析する（表 6.1 の (a) と (b)）．動詞誤りに関しては提案手法の正解数が多い．一方で，無冠詞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．提案手法のほうが無冠詞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 6.1 の (b) と (c)）．FCE+EWE と FCE+EWE-L8 は名詞誤りと名詞曲用誤りにおいて正解数が大きく異なる．名詞誤りとは suggestion と advice のような誤りであり，名詞曲用誤りとは time と times のような誤りである．FCE+EWE-L8 は，名詞誤りと名詞曲用誤りの両方で正解数が多い．理由としては，名詞誤りと名詞曲用誤りともに 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 6.2 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 6.2(a) は名詞誤りの検出例を示している．FCE+word2vec は名詞誤りを検出できていないが，FCE+E&GWE-L8 は名詞誤りを検出することができている．名詞曲用誤りに関しては表 6.2(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 は名詞曲用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 6.2(c) では，FCE+W2V は冠詞誤りの検出に成功したが，FCE+E&GWE-L8 は検出に失敗した．この結果は無冠詞と同様に誤りパターンの構造上，挿入誤りを

	誤りタイプ	動詞誤り	無冠詞	名詞選択誤り	名詞曲用誤り
(a)	FCE + word2vec	56	**48**	26	9
	FCE + C&W	53	46	24	7
(b)	FCE + EWE	60	37	29	12
	FCE + GWE	62	43	29	11
	FCE + E&GWE	64	40	31	14
(c)	FCE + EWE-L8	66	36	37	**19**
	FCE + E&GWE-L8	**67**	40	**39**	18
	誤りの合計数	131	112	77	32

表 6.1: 誤りタイプごとの正解数

	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.

表 6.2: FCE+word2vec と FCE+E&GWE-L8 を用いた誤り検出の例．正解をイタリック体とし検出結果を**太字**で表す．

適切に学習できていないことを示している．

図 6.1 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．我々は典型的な前置詞と動詞をいくつかプロットした．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

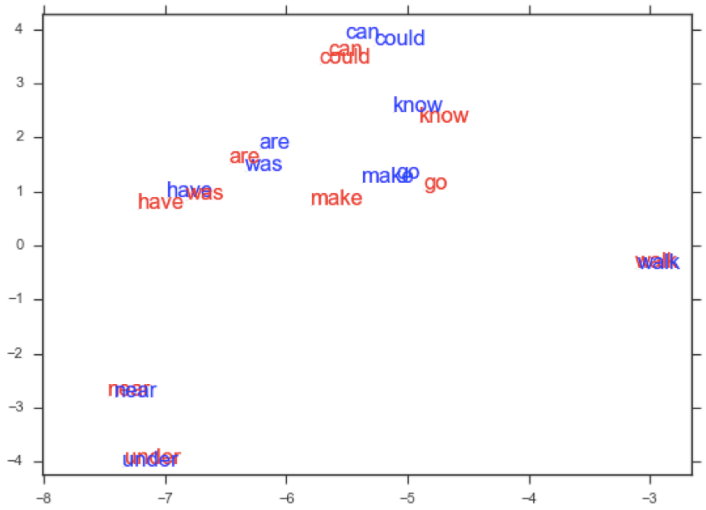

図 6.1: FCE+word2vec と FCE+E&GWE-L8 によって学習された単語分散表現の t-SNE による可視化．赤色が FCE+word2vec の単語であり，青色が FCE+E&GWE-L8 の単語である．

している．一方で，was や could 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．

第 7 章　おわりに

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．学習した単語分散表現は，NLP の応用先の 1 つである言語学習に役立つ一般的なものであることを願っている．

発表リスト

1. 金子正弘, 堺澤勇也, 小町守. **英語学習者の文法誤りパターンと正誤情報を考慮した単語分散表現学習.** 言語処理学会第 23 回年次大会, つくば, pp.729-732. March 15, 2017.
2. Kaneko Masahiro, Yuya Sakaizawa and Mamoru Komachi. **Grammatical Error Detection Using Error- and Grammaticality-Specific Word Embeddings.** In Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017), pp.40-48. Taipei, Taiwan. November 28, 2017.

謝辞

自然言語処理について何も知らない自分を外部から取り自然言語処理を研究するチャンスを下さった小町守先生に深く感謝します．そして、研究の指導だけでなく進路についても人生の先輩としてアドバイスしてくださりありがとうございました。今の自分があるのは先生のおかげです。指導してくださった堺澤さん，佐藤さんと相談に乗ってくれた同期のみなさんありがとうございます．梶原さんには，時間を惜しまず熱心に指導していただき深く感謝しています．そして，副査を引き受けてくださった山口先生と高間先生に感謝します．

参考文献

[1] M. Rei and H. Yannakoudakis, “Compositional Sequence Labeling Models for Error Detection in Learner Writing,” ACL, pp.1181–1191, 2016.

[2] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson, “One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling,” arXiv, 2013.

[3] R. Collobert and J. Weston, “A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” ICML, pp.160–167, 2008.

[4] T. Mizumoto, M. Komachi, M. Nagata, and Y. Matsumoto, “Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners,” IJCNLP, pp.147–155, 2011.

[5] H. Yannakoudakis, T. Briscoe, and B. Medlock, “A New Dataset and Method for Automatically Grading ESOL Texts,” ACL, pp.180–189, 2011.

[6] D. Dahlmeier, H.T. Ng, and S.M. Wu, “Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English,” BEA@ NAACL-HLT, pp.22–31, 2013.

[7] H.T. Ng, S.M. Wu, T. Briscoe, C. Hadiwinoto, R.H. Susanto, and C. Bryant, “The CoNLL-2014 Shared Task on Grammatical Error Correction,” CoNLL Shared Task, pp.1–14, 2014.

[8] J.R. Tetreault and M. Chodorow, “The Ups and Downs of Preposition Error Detection in ESL Writing,” COLING, pp.865–872, 2008.

[9] N.-R. Han, M. Chodorow, and C. Leacock, “Detecting Errors in English Article Usage by Non-native Speakers,” Natural Language Engineering, pp.115–129, 2006.

[10] E. Kochmar and T. Briscoe, “Detecting Learner Errors in the Choice of Content Words Using Compositional Distributional Semantics,” COLING, pp.1740–1751, 2014.

[11] Y. Sawai, M. Komachi, and Y. Matsumoto, “A Learner Corpus-based Approach to Verb Suggestion for ESL,” ACL, pp.708–713, 2013.

[12] X. Liu, B. Han, K. Li, S.H. Stiller, and M. Zhou, “SRL-based Verb Selection for ESL,” EMNLP, pp.1068–1076, 2010.

[13] D. Alikaniotis, H. Yannakoudakis, and M. Rei, “Automatic text scoring using neural networks,” ACL, pp.715–725, 2016.

[14] Z. Liu and Y. Liu, “Exploiting Unlabeled Data for Neural Grammatical Error Detection,” J. Comput. Sci. Technol., pp.758–767, 2017.

[15] Z. Xie, A. Avati, N. Arivazhagan, D. Jurafsky, and A.Y. Ng, "Neural Language Correction with Character-based Attention," arXiv, 2016.

[16] S. Chollampatt, K. Taghipour, and H.T. Ng, "Neural Network Translation Models for Grammatical Error Correction," IJCAI, pp.2768–2774, 2016.

[17] R. Nagata and K. Nakatani, "Evaluating Performance of Grammatical Error Detection to Maximize Learning Effect," COLING, pp.894–900, 2010.

[18] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," ICLR, 2015.

[19] D. Nicholls, "The cambridge learner corpus: Error coding and analysis for lexicography and elt," CL, 2003.

付録

FCE-public で用いられている誤りタイプ [19] 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．