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

金子 正弘　　堺澤 勇也　　小町 守

首都大学東京

kaneko-masahiro@ed.tmu.ac.jp，sakaizawa-yuya@ed.tmu.ac.jp，komachi@tmu.ac.jp

1　はじめに

作文中における誤りの存在と位置を示すことができる文法誤り検出は，外国語学習者の自己学習と外国語教師の自動採点において有用である．現在，文法誤り検出で用いられているアルゴリズムのほとんどは，ネイティヴの書いた生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似た入力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文法誤りパターンと正誤情報を考慮することでこの問題を解決する 2 つの手法を示す．

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 (ESWE) 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 (GSWE) である．単語分散表現の学習の際に，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．

GSWE には，学習に必要となる負例がランダムに作成される問題がある．これを解決する手法が，ESWE と GSWE を組み合わせた Error & grammaticality specific word embedding (E&GSWE) である．

表 1 は，word2vec (W2V)，C&W [2] ，ESWE，GSWE と E&GSWE 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，W2V 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，ESWE，GSWE と E&GSWE では類似度が低くなるように学習されている．そして，文法誤りの関係と似ているフレーズ対ではすべてのモデルで類似度が高くなっており，似ていな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，ESWE，GSWE と E&GSWE 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．

表 1: フレーズ対の cos 類似度

フレーズ対	W2V	C&W	ESWE	GSWE	E&GSWE
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

英語学習者作文の文法誤り検出タスクにおいて，E&GSWE で学習した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．本研究の主要な貢献は以下の通りである．

- 学習者の誤りパターンを考慮した負例作成による単語表現学習が文法誤り訂正に効果があることを示した．
- 正誤情報を考慮した目的関数による単語表現学習が文法誤り訂正に効果があることを示した．
- FCE-public データセットにおける文法誤り検出において世界最高精度を達成した．

2　先行研究

誤り検出の研究の多くは前置詞の正誤 [10]，冠詞の正誤 [3] や形容詞と名詞の対の正誤 [5] 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 [8] 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，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 [9] 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 [6] ら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 Dimitrios ら [1] 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．

3 単語分散表現の学習

この節では，提案手法である ESWE，GSWE と E&GSWE の学習の詳細について示す．これらのモデルは，既存の単語埋め込み学習アルゴリズム C&W Embedding [2] を拡張し，文法誤りパターンと正誤情報を考慮した分散表現を学習する．

3.1 C&W Embedding

Collobert と Weston [2] は局所的な文脈を元にターゲット単語に対して分散表現を学習するニ ューラルネットワークのモデルを示した．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．

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

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

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．

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

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．

3.2 文法誤りパターンを考慮した表現学習

ESWE 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に対して誤りやすい単語と入れ替えることで作成する．その際，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．

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

Dimitrios ら [1] 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，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．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．

$$f_{grammar}(x) = W_{oh1} i + b_{o1} \quad (4)$$

$$f_{context}(x) = W_{oh2} i + b_{o2} \quad (5)$$

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

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

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

式 (4) の $f_{grammar}$ は，単語列 S のラベルの予測値である．式 (5) の $f_{context}$ は，C&W Embedding の式 (3) と同様に誤差 $loss_{context}$ を求めるために計算される．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．

4 実験

4.1 Bidirectional LSTM (Bi-LSTM)

ESWE，GSWE と E&GSWE をニューラルネットワークを用いた文法誤り検出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．そのために我々は，現在文法誤り検出で世界最高精度である Bi-LSTM を用いる．

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 [8] 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 [4] で，バッチサイズを 64 文として最適化した．

表 2: Bi-LSTM による文法誤り判定結果

初期値	Precison	Recall	$F_{0.5}$
word2vec [8]	46.1	28.5	41.1
word2vec（再実装）	45.8	27.8	40.5
C&W	45.1	26.7	39.6
ESWE	46.1	28.0	40.8
GSWE	46.5	28.3	41.2
E&GSWE	**46.7**	**28.6**	**41.4**

4.2 単語分散表現

先行研究 [8] で用いられていた単語分散表現と揃え，C&W，GSWE，ESWE と E&GSWE 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，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 [4] によって最適化した．そして，GSWE の初期値はランダムとした．

誤りパターンのターゲット単語数は 4,184 であり，入れ替え候補のトークン数は 9,834，タイプ数は 6,420 である．

4.3 実験設定

我々は，モデルの評価のために First Certificate in English dataset (FCE-public) [11] を使用する．このデータセットには，英語学習者によって書かれた作文が含まれている．2,720 文をテストデータとする．そして，30,953 文をトレーニングデータとし，2,222 文を開発データとした．

FCE-public データセットにおいて人手でラベル付けされた全ての単語を検出対象とした．単語の欠落誤りに対しては，単語が欠落している直後の単語に対して誤りラベルを付与する．実験の際，過学習を防ぐためにトレーニングデータにおいて出現回数が 1 回の単語に関しては未知語とした．

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

先行研究 [8] と同様に，誤り検出の評価として $F_{0.5}$ を使用する．これは，誤り検出において適合率が再現率よりも重要であることが多いためである [7].

4.4 実験結果

表 2 は，誤り判定のタスクにおいて word2vec，C&W，ESWE，GSWE，E&GSWE のそれぞれを初期値にしたモデルを比較した実験結果を示している．word2vec [8] は先行研究の実験結果であり，word2vec（再実装）は我々による [8] の再実装の実験結果である．E&GSWE，GSWE，ESWE，word2vec（再実装），C&W の順に Precison，Recall と $F_{0.5}$ のすべての評価において高い結果が示された．また，E&GSWE を用いた提案手法は，全ての評価尺度において最高精度である先行研究 [8] を上回った．

5 考察

表 3 は，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．まず，文法誤りパターンと正誤情報を考慮することによる特徴を調べるために，W2V と C&W の正解数と提案手法の正解数の差が最も大きかった動詞誤りと無冠詞を分析する．

*We have to **wear/dress** in **an** appropriate way.*

動詞誤りとは上記の例文の dress と wear のような誤りであり，無冠詞は an のような誤りである．動詞誤りは提案手法の方が，無冠詞では W2V と C&W の方が正解数が多い．無冠詞は提案手法では考慮されていない．このことから，学習可能な誤りだけを考慮することで，考慮していない他の誤りに対してはかえって文脈だけで学習された分散表現である W2V や C&W より精度が下がると考えられる．

次に，提案手法の文法誤りパターンと正誤情報の違いを調べる．そのために，ESWE と GSWE のそれぞれに対して正解数の差が最も大きかった接続詞誤りと前置詞誤りを分析する．when と while といった接続詞誤りでは GSWE の方が，on と in のような前置詞誤りでは ESWE の方が正解数が多い．学習データにおける接続詞誤りの平均タイプ数と平均トークン数は 18 と 38 であり，前置詞誤りは 20 と 202 である．このことから，タイプ数に対してトークン数が少ないと適切に誤りパターンを考慮することができないことがわかる．

6 おわりに

本論文では，正誤情報と学習者の誤りパターンを考慮する分散表現学習を提案した．学習された分散表現は，文の誤り判定を行う Bi-LSTM の埋め込み層の初期値として使うことで，英語学習者作文の文法誤り判定タスクにおいて，世界最高精度を達成することができた．

表 3: 誤りタイプごとの正解数

誤りタイプ	個数	W2V	C&W	ESWE	GSWE	E&GSWE
動詞誤り	131	56	53	60	62	**64**
無冠詞	112	**48**	46	37	43	40
接続詞誤り	21	14	9	6	**15**	12
前置詞誤り	126	58	52	66	60	**68**

今後は，無冠詞のような文脈の考慮では対応が難しい誤りに対して文法誤りパターンを考慮できるようにモデルを改良することが考えられる．

さらに，接続詞誤りのようなタイプ数に対してトークン数が少ないため，正解数が少なかった誤りも適切に誤りパターンを考慮して学習可能にする必要がある．これは，他の英語学習者の作文コーパスを用いてトークン数を増やし，誤りパターンをさらに学習することで可能になると思われる．

参考文献

[1] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *ACL*, pages 715–725, 2016.

[2] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.

[3] Na-Rae Han, Martin Chodorow, and Claudia Leacock. Detecting errors in English article usage by non-native speakers. *ACL*, 12(02):115–129, 2006.

[4] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[5] Ekaterina Kochmar and Ted Briscoe. Detecting learner errors in the choice of content words using compositional distributional semantics. In *COLING*, pages 1740–1751, 2014.

[6] Xiaohua Liu, Bo Han, Kuan Li, Stephan Hyeonjun Stiller, and Ming Zhou. SRL-based verb selection for ESL. In *EMNLP*, pages 1068–1076, 2010.

[7] Ryo Nagata and Kazuhide Nakatani. Evaluating performance of grammatical error detection to maximize learning effect. In *COLING*, pages 894–900, 2010.

[8] Marek Rei and Helen Yannakoudakis. Compositional sequence labeling models for error detection in learner writing. In *ACL*, pages 1181–1191, 2016.

[9] Yu Sawai, Mamoru Komachi, and Yuji Matsumoto. A learner corpus-based approach to verb suggestion for ESL. In *ACL*, pages 708–713, 2013.

[10] Joel R Tetreault and Martin Chodorow. The ups and downs of preposition error detection in ESL writing. In *COLING*, pages 865–872, 2008.

[11] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *ACL-HLT*, pages 180–189, 2011.