BERT による系列ラベリングを用いた文法誤り検出

岡本 昇也 [1]　南條 浩輝 [2]　馬 青 [1]
[1] 龍谷大学理工学研究科　[2] 滋賀大学データサイエンス学部
[1]t22m002@mail.ryukoku.ac.jp
[2]hiroaki-nanjo@biwako.shiga-u.ac.jp
[1]qma@math.ryukoku.ac.jp

概要

本研究では，文法誤り検出システムを実装し，先行研究と比較した．先行研究では文法誤り検出を系列ラベリングタスクとし，BiLSTM を用いた．我々は，BERT を用いて同じく系列ラベリングタスクとして解き，文法誤り検出システムを実装した．実験の結果，BERT を用いての文法誤り検出システムは，適合率，再現率，F(0.5) 値のすべてにおいて，先行研究の BiLSTM より高かった．また，文字レベルのモデル Flair を用いた実験も行い，適合率がさらに高くなったという結果が得られた．

1　はじめに

近年，言語学習者向けの作文支援システムは多く開発されている．その中には学習者の作文の文法誤りを自動で訂正するシステムや，学習者の作文を自動評価し点数をつけるシステムなどがある．このような作文支援システムにより，学習者及び教育者の負担を軽減することができる [1].

本研究では，作文支援システムのための文法誤り検出（Grammatical Error Detection: GED）に取り組む．これは，文法誤りを含む文を入力すると文法の誤っている位置をユーザに提示するシステムである．言語学習者が作文を書く際に，指導者なしで文法誤り位置を確認することを可能とする技術であり，作文支援システムの基本となる重要な技術である．文法誤り訂正（Grammatical Error Correction）は，文法誤り検出（GED）の結果に基づき，誤りを正しい表現に自動訂正するものであり，GED はこの前段階の技術としても重要である．

文法誤り検出（GED）に関しては，近年では深層学習を用いた研究が精度を上げ英語では BiLSTM を用いることで最高精度の誤り検出を行っている [2]. 英語では深層学習を用いた文法誤り検出の研究が多くなされている．これに対して，日本語では深層学習を用いての文法誤り検出の研究は先行研究 [1] などが行われているものの，十分であるとは言えない．このような背景に基づき，我々は日本語を対象とした深層学習に基づく文法誤り検出を研究し，その効果を確認する．

文法誤り検出を深層学習を用いて行うためには大規模のデータセットが必要となる．そこで，本研究では，大規模なデータセットである Lang-8 コーパス [3] を利用する．Lang-8 コーパスは，語学学習のための SNS Lang-8 の添削ログからクローリングして作られた，誤りを含む文とその訂正文からなるデータである．約 200 万文対からなり，文法誤り訂正用のデータとしては規模の大きいものである．

本研究では，文法誤り検出を系列ラベリングタスクとして解くことを目指す．系列ラベリングとは，系列の各要素もしくは部分系列に対してタグ付けを行う手法である．代表例として，品詞タグ付けやチャンキングなどがある．文法誤り検出は，作文を単語系列として扱い，各単語に正解または誤りのラベルを付与する系列ラベリング問題として表現できる．先行研究 [1] では BiLSTM を用いて，文法誤り検出を系列ラベリング問題として解いている．BiLSTM では，前後の文脈情報を考慮して正誤ラベルの判定を行えるが，遠く離れたデータの影響を扱うのが難しい問題が残る．日本語における文法誤り検出には，「全然〜ない」などの遠く離れた関係や，基本的に文末に置かれる用言と文中の格助詞の関係を考慮する必要がある．BiLSTM よりも，系列中の各単語同士の直接的に影響を考慮できる attention 機構をもつモデルのほうが望ましいと考える．そこで本研究では BERT を用いる．

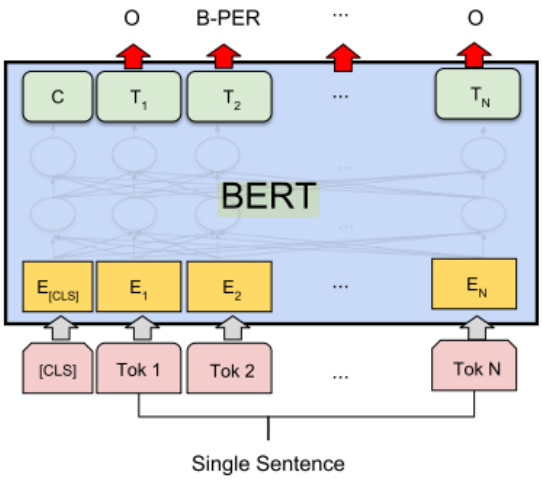

図 1　BERT の系列ラベリングの構造（文献 [6] の図 4 の (d) を引用）

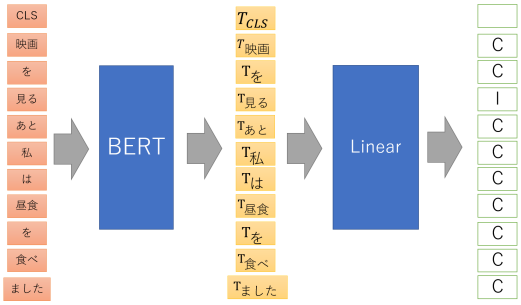

図 2　BERT での文法誤り検出

正用例	映画	を	見た	あと	私	は	昼食	を	食べ	ました
誤用例	映画	を	見る	あと	私	は	昼食	を	食べ	ました

図 3　Lang-8 データの一例

2　系列ラベリング

2.1　BERT による系列ラベリング

BERT（Bidirectional Encoder Representation from Transformer）は，attention 機構のみで構成される Encoder-Decoder モデルである Transfomer[4] の Encoder の部分を用いたモデルである．

BERT を用いた系列ラベリングは，BERT の出力に線形変換を適用したものとして実装される [5]．学習の際は，文を符号化したものを BERT に入力し，損失を計算する．損失関数はクロスエントロピーを用いた．BERT での一般的な系列ラベリングの様子を図 1 に示す．

2.2　系列ラベリングによる文法誤り検出

文法誤り検出は，各単語に正解または誤りラベルを付与するタスクに該当すると考えられる．そのため，本研究では，文法誤り検出を系列ラベリングタスクとして扱う．図 1 では，系列ラベリングの BIO 法で行っているが（B-PER は BIO の B に対応している），本研究では，BIO 法でなく IO 法で行った．BERT の系列ラベリングを用いた文法誤り検出の様子を図 2 に示す．出力ラベルの I と C は IO 法の I と O に対応している．

このような，モデルを用いて学習するために各単語に正解または誤りのラベルが付与されている必要がある．

3　データセットの作成

データセットの作成に Lang-8 コーパス [7] を使用した．

3.1　Lang-8

Lang-8 コーパスは学習者の作文であるエッセイとその添削結果からなるコーパスである．各エッセイには，エッセイ ID，ユーザー ID，学習言語タグ，母語タグが付与されている．すべての学習者の文には 1 つ以上の添削文が付与されている．本研究では，Lang-8 コーパスに含まれる，日本語学習者が書いた添削前の文である誤用例と日本語母語話者が添削した文である正用例のペア 72.2 万文対を用いる．その例を図 3 に示す．

3.2　データセットへの単語の正誤ラベル付与

Lang-8 コーパスは，学習者の作文（誤用例）とそれを訂正した作文（正用例）のペアからなるコーパスである．誤用例文のどの単語が誤り単語であるかは正用例文を見ることで確認できるが，元の誤用例文の各単語に直接的に誤りラベルが付与されているわけではない．したがって，系列ラベリング問題に使うためには，誤用例文の各単語に正誤ラベルを付与しておく必要がある．

そこで，アライメントツール [8] を使用して，各単語にラベル付けを行った．具体的には 72.2 万文対の日本語誤用例文と正用例文の対にそれぞれに対して，アライメントツールを用いて単語アライメン

正用例	映画	を	見た	あと	私	は	昼食	を	食べ	ました
誤用例	映画	を	見る	あと	私	は	昼食	を	食べ	ました
ラベル	C	C	I	C	C	C	C	C	C	C

図 4 Lang-8 コーパスのラベル付けの様子 (別の語に置換される場合)

正用例	映画	を		見た	あと	私	は	昼食	を	食べ	ました
誤用例	映画	を	を	見た	あと	私	は	昼食	を	食べ	ました
ラベル	C	C	I	C	C	C	C	C	C	C	C

図 5 Lang-8 コーパスのラベル付けの様子 (余計な語がある場合)

表 1 学習データ，検証データ，テストデータの内訳

学習データ	検証データ	テストデータ
720000	1000	1000

トをとり，一致しているところを C（正解），一致しないところを I（誤り）とした．ラベル付けの様子を図 4，図 5 に示す．図 4 は，単語アライメントをとった際に対応する単語が誤っている場合のラベル付けの例である．図 5 は，単語アライメントをとった際に誤用例の方の単語に対応する単語がない場合のラベル付けの例である．本研究では，単語アライメントをとった際に正用例の方に対応する単語がない場合は扱っていない．このラベル付けされた誤用例データをデータセットとして実験で使用する．本論文では，このデータセットを誤用タグ付きデータセットとよぶことにする．

4 実験

4.1 実験データ

文法誤り検出の実験には，誤用タグ付きデータセットを使用する．72.2 万の誤用タグ付きデータセットを表 1 に示す通り，学習データ，検証データ，評価データに分割した．

4.2 評価方法

各単語に対して正誤ラベルを推定し，誤ラベルの一致度でその性能を評価する．評価尺度には，全ての誤ラベルのうち，どの程度を正しく検出できたかを表す再現率（Recall）と，誤ラベルと検出した中で正しく誤ラベルであった適合率（Precision）を用いる．再現率と適合率は通常トレードオフの関係にあるので，この両者の調和平均である F 値を用いる．F 値は式（1）で与えられる．ここで，$\beta(\geq 0)$ は適合率と再現率のどちらかをどの程度重要視するかのパラメータであり，$\beta = 1$ のときは両者を等しく扱い，$0 \leq \beta < 1$ のときは適合率を重視し，$1 < \beta$ のときは再現率を重視する．

$$F(\beta) = \frac{(\beta^2 + 1.0) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (1)$$

外国語学習のために誤り箇所を指摘する（フィードバックする）ことを考えたとき，文法誤りでないものを誤りとフィードバックすることは望ましくなく，正確なフィードバックの方がカバレッジの高い誤り検出よりも学習効果があるとされている [9]．つまり，多くの文法誤りをきちんと誤りとフィードバックする（再現率が高い）ことよりも，与えた誤りであるというフィードバックが正しい（適合率が高い）ことが学習にとって望ましい．そこで本研究では，適合率を重視する F(0.5) で評価した．

4.3 実験条件

本研究で使用した BERT は，東北大学が公開している日本語事前学習済みのモデルを用いた．BERT のハイパーパラメータは，最適化アルゴリズムを AdamW（学習率が 0.00001）に，バッチサイズを 32 文にして，検証データを用いて最もロスの少ないエポック数を決定した．

比較のために，先行研究の BiLSTM を用いたシステムを実装したものと，系列ラベリングタスクにおいてよい性能を示している Flair[10][11] を用いた実験を行った．

Flair では，文字エンベディングに基づく単語エンベディングのみを用いるモデルと，それに BERT エンベディングを加えた単語エンベディングを用いるモデルを用いた．このようにして得られた単語エンベディング系列を BiLSTM_CRF で系列ラベリングする．前者を Flair+BiLSTM_CRF，後者を Flair_BERT+BiLSTM_CRF と表記することとする．それぞれのモデル構造は図 6 と図 7 に示す．最適化アルゴリズムには SGD を使用した．ただし，学習率を 0.1 とし，5 エポック連続で学習損失が減少しない場合に学習率を半分にするアニーリングを行う [11]．バッチサイズは 32 とした．

4.4 実験結果

各モデルにおける文法誤り検出の評価結果を表 2 に示す．適合率（Precision）と再現率（Recall），F(0.5) 全てにおいて BiLSTM に比べて BERT の方が

表 2　日本語文法誤り検出の評価

	Precision	Recall	F(0.5)
BiLSTM（先行研究）	0.545	0.160	0.368
BERT	0.639	**0.226**	**0.468**
Flair+BiLSTM_CRF	**0.773**	0.107	0.343
Flair_BERT+BiLSTM_CRF	0.719	0.172	0.439

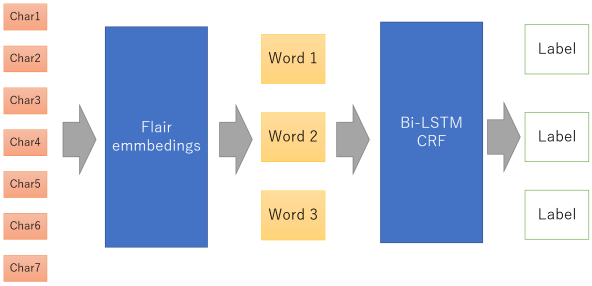

図 6　Flair のモデルの構造

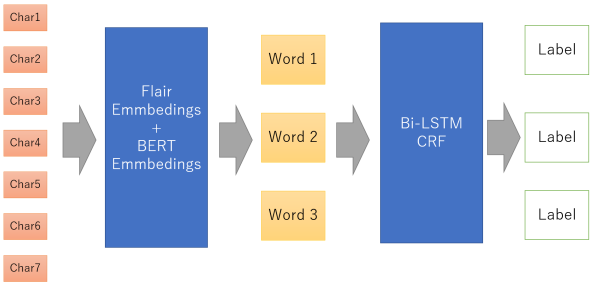

図 7　Flair_BERT のモデルの構造 [10][12]

上回っている．重視している適合率に注目すると Flair+BiLSTM_CRF が最も高い 0.773 となった．ただし，再現率は 0.107 と低く，比較的正確なフィードバックは行えるものの多くの文法誤りを見逃している結果となった．一方，BERT を用いたモデルは適合率が Flair モデルに比べるとやや低いものの 0.639 であり，再現率が 0.226 と Flair のモデルに比べて高かった．総合的な指標 F(0.5) では BERT が最も高かった．また，単語エンベディングに文字エンベディングと単語エンベディングを用いた Flair モデル（Flair_BERT+BiLSTM_CRF）については，文字エンベディングのみの Flair（Flair+BiLSTM_CRF）と BERT の中間的な結果となっており，Flair+BiLSTM_CRF の適合率と BERT の再現率のよさをそれぞれ活かせるハイブリッド型的なモデルになっている．文字レベルのモデルを用いると，正確なフィードバックが期待でき，BERT などの単語レベルのモデルを用いると文字レベル程の正確なフィードバックは期待できないが文字レベルのモデルに比べてカバレッジの高い文法誤り検出になると考える．

5　終わりに

本研究では BERT を用いて文法誤り検出システムを実装し，先行研究との比較を行った．実験の結果，BERT を用いての文法誤り検出システムは，適合率，再現率，F(0.5) 値のすべてにおいて，先行研究の BiLSTM より高かった．また，文字レベルのモデル Flair を導入することにより，適合率がさらに向上し，言語学習者へのより正確なフィードバックが期待できることがわかった．今後は Flair の高い適合率と BERT の高い再現率を生かしたハイブリッドモデルを実装し，文法誤り検出の性能向上を図りたい．

謝辞

Lang-8 のデータ使用に際して，快諾くださった株式会社 Lang-8 社長喜 洋洋氏に感謝申し上げます．なお，本研究は JSPS 科研費 19K12241 の助成を受けたものです．

参考文献

[1] 新井美桜, 金子正弘, 小町守. 日本語学習者向けの文法誤り検出機能付き作文用例検索システム. 人工知能学会論文誌, Vol. 35, No. 5, pp. A–K23_1–9, 2020.

[2] M. Rei and H. Yannakoudakis. Compositional sequence labeling models for error detection in learner writing. In **Proceedings of ACL**, pp. 1181—-1191, 2016.

[3] T. Mizumoto, T. Tajiri, T. Fujino, S. Kasahara, M. Komachi, M. Nagata, and Y. Matsumoto. Naist lang-8 learner corpora, 2012. https://sites.google.com/site/naistlang8corpora/home.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. In **Neural Information Processing Systems**, 2017.

[5] 近江崇宏, 金田健太郎, 森長誠, 江間見亜利. BERT による自然言語処理入門 − Transfomers を使った実践プログラミング −. オーム社, 2021.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. **arXiv**, 2018.

http://arxiv.org/abs/1810.04805.

[7] M. Tomoya, M. Komachi, and M. Nagata. Mining revision log of language learning sns for automated japanese error correction of second language learners. In **Proceedings of the 5th International Joint Conference on Natural Language Processing**, pp. 147–155, 2011.

[8] Z. Dou and G. Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In **Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, 2021.

[9] R. Nagata and K. Nakatani. Evaluating performance of grammatical error de tection to maximize learning effect. In **Proceedings of COLING**, pp. 894–900, 2010.

[10] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In **NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 54–59, 2019.

[11] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In **COLING 2018, 27th International Conference on Computational Linguistics**, pp. 1638–1649, 2018.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 260—-270, 2016.