文法誤り検出 BERT のためのマルチタスク追加事前学習

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，文法誤りの検出性能の向上を目的として取り組んだ．我々は，追加の事前学習として対照学習と MLM のマルチタスク学習を提案した．実験の結果，追加の事前学習を行うモデルは追加の事前学習を行わないモデルに比べて性能が向上した．この追加の事前学習により，各単語の意味表現，文の意味表現を維持しつつ，文が誤りを含むか含まないか区別させる学習ができたと考えられる．

1　はじめに

我々は，日本語学習者の作文支援システムの構築を目的として，文法誤り検出 (Grammatical Error Detection: GED) の研究を行っている [1]．これは，ユーザが文法誤りを含む文を書くとその箇所が自動的にユーザに提示されるシステムであり，作文支援システムの基本となる重要な技術である．

文法誤り検出（GED）に関しては，近年では深層学習を用いた研究が盛んであり，英語では文献 [2] など，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は先行研究 [3] 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，具体的には，BERT に基づく系列ラベリングを用いた文法誤り検出を研究している [1]．本研究では，この枠組みにおける BERT の学習方法である，新たな追加事前学習方法を提案し，その効果を示す．具体的には，BERT に誤りを含む文と含まない文を区別させる追加事前学習を提案する．すなわち，BERT から得られる文 embedding が文法誤りを含むか含まないかを区別できるように事前学習させることで，文法誤り検出において誤りを含まない文の token に対して，誤りラベルを付与してしまうことの抑制を狙う．その際，NSP の代わりに対照学習を導入することを提案する．これと MLM を同時に追加事前学習する（マルチタスク追加事前学習する）ことで，その後の下流タスクである系列ラベリングに基づく文法誤り検出性能が向上することを示し，提案手法の有効性を示す．

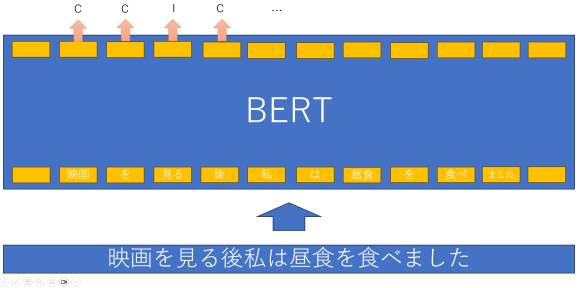

図 1　BERT を用いた系列ラベリングによる文法誤り検出

2　系列ラベリングによる文法誤り検出

2.1　BERT を用いた系列ラベリング

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．BERT の系列ラベリングを用いた文法誤り検出の様子を図 1 に示す．出力ラベルの I（誤り）と C（正解）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．

2.2　系列ラベリングを用いた文法誤り検出のためのデータセット

系列ラベリングを用いた文法誤り検出のため，我々は Lang-8 コーパス [4] を使用したデータセット

表 1 日本語誤用タグ付きデータセット（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

を構築している [1]．その概要を説明する．

2.2.1 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．ここから，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 万文対を取りだした．

2.2.2 データセットへの単語の正誤ラベル付与

誤用例文のどの単語が誤り単語であるかの情報が必要であ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，誤用例文の各単語にアライメントツール [5] を使用して，正誤ラベルを付与した．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．すなわち，学習者作文における置換誤りと挿入誤りの語に I のラベルを付与した．なお，本研究では，正用例の語に対応する誤用例の語がない（作文における削除誤り）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日本語誤用タグ付きデータセットとよぶことにする．これを学習データ，検証データ，テストデータに分割した（表 1）．

3 系列ラベリングによる文法誤り検出用 BERT の追加事前学習

日本語誤用タグ付きデータセットに対する BERT を用いた系列ラベリングによる文法誤り検出精度を向上させるため，BERT に対する追加の事前学習手法を研究する．具体的には，対照学習と MLM のマルチタスク学習を提案する．

3.1 対照学習

対照学習 [6] とは，あるデータの表現 (embedding) を同じクラスに属するデータの表現に近づけ，異なるクラスに属するデータの表現からは遠ざけるよう学習する手法である．自然言語処理においては，文の意味表現を得るために，ある文の embedding を，同じ意味をもつものの表層表現が異なる文の embedding に近づけ，意味が異なる文の embedding から遠ざけるというような使い方が一般的である．

本研究では，文法誤り検出を目的としており，分類したいクラスは文法誤りを含まない/含むの 2 値である．そこで，文法誤りを含まない文 (positive sample) の embedding を他の文法誤りを含まない文 (anchor sample) の embedding と近づけ，文法誤りを含む文 (negative sample) の embedding とは遠ざけることを行う．ここで，anchor sample は positive sample と表層的にも意味的にも類似している必要はない．一方，文法誤りを含む negative sample としては positive sample と表層的に似通っている（結果的に意味的に似通っている）文とする．このような sample は敵対的 sample ともいえる．意味や表層の違いに惑わされず，文法誤りを含まない文とは embedding が近づくように，文法誤りを含む文とは離れるように学習を行う．

3.2 MLM (Masked Language Model)

MLM (Masked Language Model) とは，文の穴埋めを行う言語モデルであり，このタスクは BERT の事前学習に用いられている．BERT の事前学習 MLM タスクでは，前後のコンテキストからマスクされた token を予測するよう学習がされており，これにより，各 token に対する BERT の出力は文脈を考慮した token の表現となるといえる．BERT を 3.1 節で述べた対照学習を行うと，文法誤りを含まない文同士は，意味が異なっても近い embedding になるように学習される．このことは，BERT がもつ各 token の意味表現を破壊しているとも考えられる．そこで，BERT に対して前述の対照学習を行う際に，MLM も同時に行うことを提案する．このようにすることで各 token の意味表現，文の意味的表現を維持しつつ，文法誤りを含まない文と文法誤りを含む文を分離できるようになることが期待できる．

4 追加事前学習の設定

4.1 実験データ

対照学習用データにも，Lang-8 のデータを使用する．

追加事前学習のためのデータセットとして，

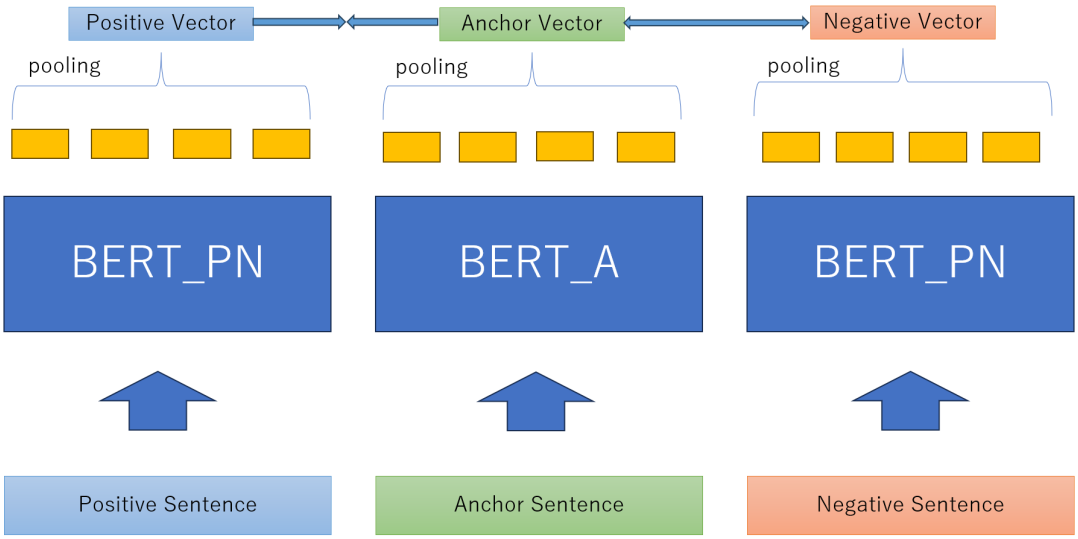

図 2　BERT での対照学習

表 2　対照学習のデータ

Anchor	Positive	Negative
360000	360000	360000

Anchor，Positive，Negative のデータをそれぞれ用意した．本研究では，日本語誤用タグ付きデータセットの学習データ 720,000 文から用意した．Negative は誤用例文，Positive は Negative のペアとなる正用例文，Anchor は Positive と異なる正用例文を用いた（表 2）．Anchor と Positive を別の正用例文とするため，720,000 のデータを 2 つにわけた．具体的には，誤用例文から 360,000 の negative sample を抽出し，それに対応する正用例文の 360,000 の positive sample を得た．残りの正用例文 360,000 を anchor sample とした．

4.2　学習設定

BERT の事前追加学習（対照学習）の方法を図 2 に示す．

Anchor の文ベクトルを出す BERT（BERT_A）と Positive，Negative の文 embedding を出す BERT（BERT_PN）の二つの BERT を用意し，Anchor，Positive，Negative の文を入力しそれぞれの文 embedding を求め，損失関数を Triplet Loss として学習する．なお，いずれの BERT も同じ事前学習モデルを用いる．anchor 文に対しては embedding が変わらないようにするため，BERT_A のパラメータは固定する．

表 3　対照学習ハイパーパラメータ

最適化アルゴリズム	AdamW
学習率	0.000001
バッチサイズ	64
エポック	15

文 embedding の求め方は，予備実験によって最も性能が良かった Average pooling とした．対照学習の学習を行う際の BERT のハイパーパラメータとしては表 3 のとおりである．

5　実験

5.1　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），これらの調和平均である F 値（式（1））を用いる．

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

言語学習のための誤り箇所の指摘においては，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 [7]．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

表 4 系列ラベリングモデル学習用ハイパーパラメータ

最適化アルゴリズム	AdamW
学習率	0.00001
バッチサイズ	32
エポック	検証データで決定

表 5 日本語文法誤り検出における追加の事前学習の効果

追加の事前学習	Precision	Recall	F(0.5)
なし	0.639	0.226	0.468
あり	0.725	0.244	0.520

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．本研究では，適合率を重視する F(0.5) で評価する．

5.2 実験条件

本研究で，東北大学が公開している日本語事前学習済みの BERT モデルを用いる[1)]．日本語誤用タグ付きデータセットの学習データを用いて系列ラベリングモデルを学習し，評価データで評価する．

事前学習済み BERT をそのまま系列ラベリングモデルとして学習したものと，事前学習済み BERT に対して対照学習と MLM で追加の事前学習を行った上で系列ラベリングモデルとして学習したものを比較する．

系列ラベリングモデルの学習のためのハイパーパラメータは表 4 のとおりである．

5.3 実験結果

文法誤り検出における追加の事前学習の効果を表 5 に示す．

追加事前学習を行ったモデルと追加事前学習を行っていない BERT とでは，適合率（Precision) と再現率（Recall），F(0.5) 全てにおいて追加事前学習を行わない BERT に比べて追加事前学習をした BERT の方が上回っている．特に Precision が大きく向上しており，言語学習の GED モデルに適しているといえる．この結果は，提案する追加事前学習が GED のための系列ラベリングモデル学習にとって有効であったことを示している．

5.4 Ablation Study

次に，追加の事前学習において対照学習のみの場合と MLM のみの場合を比較した．結果を表 6 に示す．追加事前学習を行わない BERT に比べて性能の向上は見られなかった．対照学習のみでは文法誤りを含むと含まない文の表現（各単語 embedding の平均）の区別は学習はできるものの同時に各単語 embedding を適切に修正できていなかったため，系列ラベリングモデル学習に効果が見られなかった可能性があると考えられる．また，MLM のみでは文法誤りを含むと含まない文の表現の区別は学習できないことに加え，もともとの BERT が持っていた単語 embedding を過剰に修正してしまったことが性能が向上しなかった原因の一つと考えられる．

表 6 追加事前学習 (Ablation Study)

追加事前学習タスク	Precision	Recall	F(0.5)
なし	0.639	0.226	0.468
contrast+mlm（提案法）	0.725	0.244	0.520
contrast のみ	0.638	0.225	0.467
mlm のみ	0.612	0.168	0.401

提案法では，各単語の意味表現，文の意味的表現を維持しつつ，文法誤りを含まない文と文法誤りを含む文を分離できるよう学習ができたと考えられる．

6 おわりに

BERT を用いた文法誤り検出システムの研究を行った．追加の事前学習として対照学習と MLM のマルチタスク学習を提案した．提案した追加事前学習を行ったモデルと行わなかったモデルとの比較を行い，追加の事前学習の効果を確認した．

各単語の意味表現，文の意味的表現を維持しつつ，文法誤りを含まない文と文法誤りを含む文を分離できるよう学習ができたと考えられる．

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] 岡本昇也, 南條浩輝, 馬青. Bert による系列ラベリングを用いた文法誤り検出. 言語処理年次大会, 2023.

[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] 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

1） https://huggingface.co/docs/transformers/model$_doc/bert$

会論文誌, Vol. 35, No. 5, pp. A–K23_1–9, 2020.

[4] 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.

[5] 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.

[6] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. **Technologies**, Vol. 9, No. 1, p. 2, 2020.

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