サブワード非依存の辞書ベース BERT モデルを用いた日本語文法の誤り検出

山下稜心[1] 馬青[1]
[1]龍谷大学先端理工学研究科
y24m006@mail.ryukoku.ac.jp, qma@math.ryukoku.ac.jp

概要

本研究では，日本語文法誤り検出の実用性向上を目的とし，データセットとモデル構成の改善を行った.われわれの先行研究で用いたデータセットでは，未知語処理をサブワード分割に依存していたため，誤り文特有の未知語によりトークンが過度に細分化される問題があった．そこで，誤り文と訂正文の対応関係に基づいてトークンを再割当てし，データセットを再構築した．また，再構築したデータから生成した辞書に基づく独自トークナイザを設計した．さらに，誤り位置を直接推定する手法に加え，文レベルの正誤判定後に誤り位置推定を行う二段階モデルを提案し,構文情報の有効性についても検討した.

1 はじめに

近年,言語学習者向けの作文支援システムとして，文法誤りを自動で検出・訂正する手法や，作文を自動評価する手法が多数提案されており，学習者および教育者の負担軽減に寄与している [1]．なかでも文法誤り検出（Grammatical Error Detection: GED）は，誤り箇所を自動で特定することで即時フィードバックを可能にする技術であり，文法誤り訂正（Grammatical Error Correction: GEC）の前段階としても重要である.

GED に関する研究は主に英語を対象として発展しており, BERT [2] に代表される事前学習済み言語モデルを用いた系列ラベリング手法によって高い性能が報告されている [3]. 一方, 日本語を対象とした研究は限定的であり,われわれは先行研究において，BERT を用いた日本語文法誤り検出手法を提案した[4]．しかし，日本語学習者の誤り文には未知語や非標準的表現が多く含まれるため，既存のサブワード型トークナイザではトークンが過度に細分化され，誤り位置の解釈や学習者へのフィードバックの観点から実用性に課題があった.

そこで本研究では，誤り文と訂正文の対応関係に基づいてトークン分割を再設計し，サブワード分割に依存しない日本語文法誤り検出用のデータセットを再構築した．さらに，訂正文を伴わない入力文にも適用可能とするため，再構築したデータセットから生成した辞書に基づくトークナイゼーションを採用した.

また，実際の作文支援システムでは正しい文が入力される場合も多いことから，本研究では正しい文を含めた評価を行った．そのため，誤り位置を直接推定する手法に加え，文レベルで正誤判定を行った後，誤り文に対してのみ誤り位置を推定する二段階モデルを導入した.

さらに，日本語の文法誤りには，格助詞と用言の対応関係など構文的依存関係に基づく誤りが多く含まれる．このような誤りを捉えるため，依存構造解析器 Stanza[5] により得られた構文的特徴量をモデルに導入し，その有効性も検討した.

近年，大規模言語モデル（LLM）は高い言語生成能力を有するが，文法誤り検出では文の再生成よりも，誤り位置をトークン単位で正確に特定することが重要である．LLM は系列ラベリングによる定量評価との整合性を確保しにくい場合があるのに対し，BERT に基づく本提案手法は，各トークンに対する予測結果を明示的に出力でき，再現性の高い評価が可能である.

2 データセットの再構築

2.1 使用コーパス

Lang-8 コーパス [6]は学習者の作文と訂正文を文対で収集したコーパスで，本研究では日本語文対のみを抽出し,文法誤り検出用データとして利用した．誤り文には未知語や表記揺れが多く，サブワード型トークナイザでは過度な細分化が生じるため，本研究ではトークン分割の段階からデータセットを再設計した． 図 1 に，従来のトークン分割例を示す.

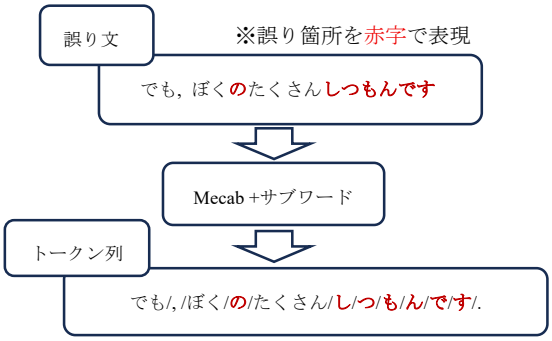

図 1 従来のトークン分割例

2.2 訂正文に基づくトークンの再割当

本研究では，訂正文との対応関係を利用して誤り文のトークンを再割り当てする手法を採用した．訂正文のトークン分割を基準に誤り文のトークンを決定することで，未知語や非標準的表現を含む誤り文でも安定したトークン割り当てが可能となる．分割例を図 2 で示す．

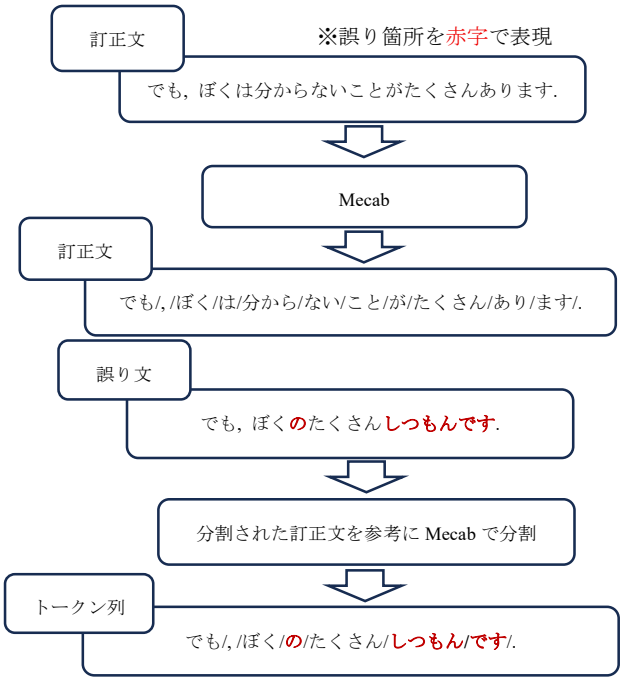

図 2 訂正文に基づくトークンの再割当例

3 手法

3.1 誤り検出の手法

本研究では，誤り位置を直接推定する手法と，文全体の正誤判定後に誤り文と判定された文のみを対象とする二段階モデルを採用した．直接推定手法はすべての文に対して一貫した処理が可能で誤り位置推定性能を直接評価できる一方，正しい文に対する不要な推定が含まれる可能性がある．二段階モデルでは，第 1 段階で文の正誤を判定し，第 2 段階で誤り文と判定された文にのみ誤り位置推定を行うことで，不要な誤り検出を抑制し，実運用に適した処理フローを実現できる．

3.2 構文情報の導入

日本語の文法誤りには語彙的誤りに加え，格助詞と用言の対応など構文的誤りも多く含まれる．本研究では，トークン列に加え，Stanza により得られた品詞情報および格情報を補助的特徴量として BERT 系列ラベリングモデルに統合した．統合方法としては，(1) 埋め込み表現として加算する方法，(2) one-hot 表現として連結する方法の二通りを検討し，性能への影響を比較した．

3.3 トークナイゼーション

文法誤り検出に適したトークン分割を実現するため，再構築したデータセットに基づく辞書を用いたハイブリッド型トークナイゼーション手法を採用した．本手法は貪欲法と最頻一致法を組み合わせ，両手法の欠点を補完する．例えば「いろんな行ったこともないの場所に行きました」という文では，貪欲法では「こ / ともない」と誤結合され，最頻一致法では「こ / と / も / ない」のように過度に細分化される．ハイブリッド型では，文字列長と語の出現頻度・文脈妥当性を併せて考慮することで，「こと / も / ない」と適切に分割されることが可能となる．このように，本手法は誤りを含む表現を適切な粒度で保持できることが期待される．

4 評価

データセット，トークナイゼーション，および文法誤り位置推定モデルについて評価を行った．

4.1 評価用データ

まず，評価全般に用いるデータを表 1 に示す．各評価に用いるデータについては，以下に述べる．

データセットにおけるトークン分割の妥当性を検証するために，誤り文のすべてから無作為に 100 文を抽出し，人手により正解トークンを付与した．

文法誤り検出タスクにおけるトークナイゼーションの評価にあたっては，辞書構築と評価データを厳密に分離した．具体的には，自作トークナイザに用いる辞書は，誤り文の学習データのみから作成し，

テストデータに含まれる文は辞書生成には一切使用していない．誤り文のテストデータに対して辞書に基づいたトークン分割を行い，その分割結果を評価対象とした．

文法誤り位置推定の評価には表 1 のデータを用いる．本評価では，結果のばらつきを抑えるため，実験データに対して 9:1 の比率で分割を行い，分割位置をローテーションさせた 10 パターンの学習・テストを実施し，その平均値を最終的な評価結果として採用した．なお，誤り文と正しい文の総数が異なるため，単純に両者を同一比率で分割すると，テストデータにおける誤り文・正しい文の数が同じでなくなる．そこで本研究では，まず誤り文のみを対象として 9:1 の分割を行った．分割された 1 割の誤り文と，それと同数の，（全正しい文から）無作為に抽出した正しい文をテストデータとして用いた．残りの誤り文と正しい文はすべて学習データとして使用した．

表 1 評価全般用データ

	誤り文	正しい文
学習データ	512,319	303,075
テストデータ	56,925	56,925
合計	569,244	360,000

4.2 モデル構成

4.2.1 直接推定モデル

先行研究と同様，誤り検出を系列ラベリングとして直接推定した．事前学習済み日本語 BERT [7]を用い，各トークン出力に対して分類層を追加することで正誤ラベルを付与した．ただし，トークンの獲得には，3.3 節で述べた自作トークナイザを用いた．

4.2.2 二段階モデル

第 1 段階：文の正誤判定

入力文が正しい文であるか，あるいは誤りを含む文であるかを判定する．本タスクは二値分類問題として定式化し，直接推定モデルと同じ日本語 BERT を用いた．この段階で「正しい文」と判定された文については，後続の誤り位置推定処理を行わないことで，不要な誤検出を抑制した．

第２段階：誤りの位置推定

第 1 段階で誤り文と判定された文に対し，直接推定モデルと同じ手法で誤り位置の推定をした．

4.2.3 構文情報の導入

構文情報として Stanza により付与された品詞情報（17 種類）および格情報（30 種類）を用いた．これらの情報は，BERT モデルに対して以下の二通りの方法で統合した．

(1) **埋め込み表現加算方式**：品詞および格情報を torch.nn.Embedding[8]で埋め込みベクトルに変換し，BERT のトークン埋め込みに加算する方式である．

(2) **One-hot 連結方式**：品詞・格情報およびトークンの係る head 情報を one-hot ベクトルとして BERT 埋め込みに連結する方式である．いずれの方式でも追加特徴量は学習時に更新される．

4.3 評価方法

データセットにおけるトークン分割の妥当性の検証においては，4.1 で述べた 100 文について，データセットに付与された分割結果と正解との比較を行った．評価指標には，トークン境界の一致に基づく適合率，再現率および F 値を用いた．

文法誤り検出タスクにおけるトークナイゼーションの妥当性の確認においては，貪欲法，最頻一致法，およびハイブリッド型の 3 種類のトークナイザを対象に評価を行った．評価では，テストデータセットに付与されたトークンとトークナイザの分割結果との一致率を算出した．

文法誤りの位置検出の性能評価はトークン単位で実施した．直接推定モデルでは誤り位置推定性能を，二段階モデルでは文レベルの正誤判定性能に加え，第 1 段階で誤りが含まれると判定された文を対象として誤り位置推定性能を評価した．また，学習支援において誤検出の抑制を重視する観点から，F 値は再現率よりも適合率を重視する F(0.5) を用いた．

4.4 評価結果

データセットに関する評価結果を表 2 に示す．従来のデータセットでは，トークン分割における適合率および再現率がともに低く，特に再現率が低いことから，人手で想定されるトークン境界が十分に再現されていないことが確認された．これは，誤り文に含まれる未知語や表記ゆれに対して，サブワード分割が過度に細分化を行う傾向にあるためと考えられる．一方，再構築したデータセットでは，適合率・再現率ともに大幅な改善が見られ，人手分割と整合的なトークン分割が実現されていることが示された．

表 2 データセットの性能

	適合率	再現率	F 値
従来のデータセット	0.55	0.41	0.47
再構築したデータセット	0.87	0.86	0.86

トークナイゼーション手法の評価結果を表 3 に示す．貪欲法および最頻一致法はいずれも一定の一致率を示したものの，分割精度には限界が見られた．一方，両者を組み合わせたハイブリッド型は，局所的な分割戦略と頻度情報を併用することで分割誤りを抑制し，3 手法の中で最も高い一致率を示した．

表 3 トークナイザの性能

	一致率
貪欲法	0.79
最頻一致法	0.77
ハイブリッド型	0.96

文の正誤判定に関する評価結果を表 4 に示す．適合率，再現率，F(0.5)はいずれも 0.77 であり，第 1 段階の文判定モデルとして一定の性能を有していることが確認された．誤り位置推定に関する評価結果を表 5 に示す．結果，二段階モデルが直接推定モデルを一貫して上回る性能を示した．これは，第 1 段階で誤り文を適切に選別することで，第 2 段階における不要な誤検出が抑制されたためである．

さらに，学習データに正しい文を含めるか否かの効果について，直接推定モデルでは正しい文を含めた学習が有効である一方，二段階モデルでは，第 2 段階に入力される文が誤り文に偏る特性から，誤り文のみを用いた学習が適していることが示された．

表 4 文の正誤判定の性能

適合率	再現率	F(0.5)
0.77	0.77	0.77

表 5 誤り位置推定の性能

	学習に正しい文を含むか否か	適合率	再現率	F(0.5)
直接推定	含まない	0.38	0.29	0.36
直接推定	含む	0.58	0.20	0.42
二段階	含まない	0.56	0.25	0.44
二段階	含む	0.60	0.19	0.42

構文情報を導入した評価結果を表 6 に示す．埋め込み表現加算方式および One-hot 連結方式のいずれにおいても明確な性能向上は確認されず，特に依存関係の head 情報を導入した場合には性能低下が見られた．以上より，本研究で扱う文法誤り検出タスクにおいては，構文情報を明示的に付与することは必ずしも有効ではないことが示唆された．

表 6 誤り位置推定に構文情報が導入された場合の性能

構文情報	付与方式	適合率	再現率	F(0.5)
品詞	加算	0.57	0.19	0.41
格	加算	0.53	0.19	0.39
品詞＋格	加算	0.55	0.18	0.38
品詞＋格	連結	0.47	0.16	0.33
品詞＋格＋head	連結	0.13	0.14	0.13

5 終わりに

本研究では，日本語文法誤り検出において，サブワード分割に起因するトークンの過度な細分化の問題に着目し，誤り文と訂正文の対応関係に基づいてトークン分割を再設計したデータセットを構築した．本データセットから構築した独自の辞書に基づくトークナイゼーション手法を BERT に導入した．評価の結果，人手分割と整合的なトークン分割が実現されていることが示された．

また，誤り位置を直接推定する従来手法に加え，文レベルの正誤判定を前段に設けた二段階モデルを提案し，その有効性を検証した．その結果，二段階モデルは正しい文に対する不要な誤検出を抑制しつつ，直接推定モデルを上回る検出性能を示した．さらに，学習データに正しい文を含める効果についても検討を行った．その結果，直接推定モデルでは正しい文を含めた学習が有効である一方，二段階モデルでは誤文のみを用いた学習が適していることが明らかとなった．

一方，構文情報の導入については明確な性能向上は確認されず，特に依存関係の head 情報は性能低下を招いた．今後は，誤り文に対する構文解析の信頼性の検証および向上や，有効な構文情報の選定を通じて，日本語文法誤り検出手法のさらなる性能向上を目指す．

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Representations. arXiv 2019. https://arxiv.org/abs/1810.04805

[3] Rahul Nihalani, Kushal Shah. Enhancing Grammatical Error Detection using BERT with Cleaned Lang-8 Dataset. Representations. arXiv 2024. https://arxiv.org/abs/2411.15523.

[4] 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. 言語処理学会第 29 回年次大会発表論文集. pp.1607-1611, 2023

[5] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In the Association for Computational Linguistics (ACL) System Demonstrations. 2020.

[6] 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.

[7] tohoku-nlp/bert-base-japanese-whole-word-masking, tohoku-nlp/bert-base-japanese-whole-word-masking · Hugging Face (accessed 2025-01-01).

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Representations. arXiv 2019. https://arxiv.org/abs/1912.01703