ルールベースと BERT を用いた 日本語学習者向けの格助詞校正システムの提案

蔡 宇倫　　荻野 正樹
関西大学
{k540873,ogino}@kansai-u.ac.jp

概要

本研究は、格助詞の誤りを余剰(Unnecessary)・不足(Missing)・置換(Replacement)の 3 種類の誤りタイプに分類し、ルールベースと BERT 言語モデルを用いて、3 種類の誤りを検出・訂正する手法を提案する。日本語学習者による日本語作文データベースを用いて提案したシステムを評価した結果、従来のルールベースによる手法より誤り検出率が約 8.2%、誤り訂正率が約 14.3%上回っており、ルールベースと BERT を組み合わせた手法の文法誤り訂正における有効性を検証することができた。

1 はじめに

グローバル化が進んでいる現在では、英語や日本語をはじめ、“外国語”を活用してコミュニケーションを図ることは決して珍しいことではない。笠原ら[1]が NAIST 誤用コーパスから日本語学習者が誤る箇所を調べた調査では、助詞の誤りが 24%を占めており、日本語学習者に対する助詞の校正支援の重要が高いと考えられる。しかし、既存の日本語向けの自動修正ツールは誤字脱字の検出や不適切な表現の書き換えについての修正能力は高いが、日本語学習者によく見られる格助詞の誤りについては検出すらできないことが多い。

自然言語処理（NLP）において、入力したテキストの文法的誤りを自動で検出して訂正をする問題は「文法誤り訂正（Grammatical Error Correction: GEC ）」という[2]。文法誤り訂正の手法には、大きく分けてルールベース、分類器ベース、機械翻訳ベースの 3 つがある[3]。このうち、格フレームなどの情報を参照するルールベースの手法と、機械学習を用いた機械翻訳ベースの手法は格助詞誤り訂正の主流の手法である。

今枝ら[4]は入力文の格フレームと NTT 日本語語彙大系から得られる格フレームを比較することにより、格助詞誤りの検出および訂正をする手法を提案した。その結果、75.6%の誤り検出率と、62.5%の誤り訂正率を示した。南保ら[5]は構文解析によって得られた文節内の特徴から抽出した特徴スロットと助詞の組み合わせをルールとする。さらに帰納的学習を用いて得られた助詞選択ルール辞書を用いて誤りの検出および校正の手法を提案した。この手法は、今枝ら[4]と同等な結果を得られた。2011 年以降、ルールベースによる格助詞誤り訂正の手法はほんどん提案されなかった。

近年、文法的に誤りを含む文から文法的に正しい文への翻訳をする機械翻訳ベースが文法誤り訂正の主流の手法となっている。日本語の助詞は英語の前置詞に相当する。英語の文法誤り訂正の研究において、翻訳機を使って前置詞の誤りを検出・訂正する手法が多く提案されている。しかし、機械翻訳ベースでは膨大な学習データ（対訳データ）が必要であり、日本語では、研究に使う対訳データの不足が問題になっている。この問題に対し、今村ら[6]は日本語平文コーパスの利用と擬似誤りペア文による対訳データの拡張という二つの提案を行って、小規模誤りデータからの日本語学習者作文の助詞誤り訂正の手法を提案した。その結果、50.2%の適合率と 18.9%再現率を得ることができた。小川ら[7]は機械翻訳機にコピー機構を組み込んだモデルを用いて日本語学習者作文全般の誤りに対して訂正する手法を提案している。この手法では SMT を用いた手法より性能の向上が認められたが、英語に適用した場合と同等の性能は得られなかった。

本研究では対訳データを使わず、従来のルールベースの参照情報と BERT[8]を組み合わせる手法を提案し、格助詞の余剰(Unnecessary)・不足(Missing)・置換(Replacement)という 3 種類の誤りの検出および訂正をする。日本語学習者による日本語作文データベース[9]を用いて提案したシステムを評価する。

2 システムの概要

GEC の研究において、Errant[10]により文章の誤りは余剰(Unnecessary)・不足(Missing)・置換(Replacement)の 3 種類の誤りタイプに分類されている。本研究は格助詞の誤りを上記の 3 種類の誤りタイプに分類し、誤りタイプごとに校正手順を提案する。

本システムは図 1 で示すように、日本語学習者による作文を入力対象として、格助詞の誤りを含む文に対して、ルールベースと BERT を用いて、作文の中に格助詞が「余剰」・「不足」・「置換」の箇所を検出し、誤り箇所に対して、「削除」・「挿入」・「置換」の作業を行い、結果を出力する。

3 校正手順

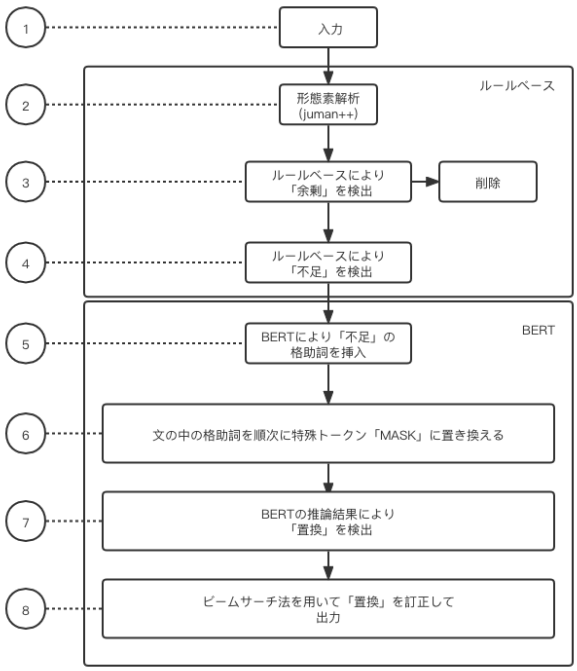

図 1 システムの校正手順

3.1 形態素解析

本研究では、Juman++[11]を用いて形態素解析を行う。Juman++は言語モデル Recurrent Neural Network Language Model (RNNLM) を用いた高性能の形態素解析システムである。単語の並びの意味的な自然さを考慮した解析が可能である。図 1 の①で入力した文節に対して、Juman++を用いて分かち書き処理をする。分かち書き処理後の単語を形態素とし、表 1 のように品詞情報を付け加える。

表 1　形態素解析の例

入力文	私は太郎です。
分かち書き	［私］［は］［太郎］［です］［。］
品詞情報	［私-名詞-普通名詞］［は-助詞-副助詞］［太郎-名詞-人名］［です-判定詞］［。-特殊-句点］

3.2 「剰余」の誤り検出・訂正

機械翻訳ベースによる手法では、翻訳機は「置換」のみに対して翻訳を行うのが一般的である。今村ら[6]は「不足」・「余剰」の誤りに対する「挿入」・「削除」の操作を空単語からある単語の置換とある単語から空単語の置換とみなせる手法を提案したが、「挿入」操作は全単語の間に挿入される可能性があり、非常に計算コストがかかる操作であるため、体言の後ろのみ挿入するように制約をかけた。

計算コストの制約とシステム構築の利便性を考えた上で、本システムは図 1 の②で得た分かち書き情報と詞品情報を用いて、文法的なルールに基づいて「不足」と「余剰」の誤りを検出する。

表 2　「余剰」の誤り一部例

例 1	用言（形容動詞を除く）と体言の間に「の」が挟む文 例文 1：日本語を使うの仕事を探します。 例文 2：美しいの花
例 2	同じ助詞が連続使用している文 例文：私は斎藤とと申します
例 3	隣接できない助詞を使用している文 例文 1：ラーメンをが食べたい 例文 2：私はそれしかを食べない
例 4	「を」が二つ以上含まれる文 例文：国を管理をする

「余剰」の誤りを検出および訂正する手法としては、表 2 で挙げる「余剰」の誤りに対して、分かち書き情報と詞品情報を用いて誤り箇所を検出する[4, 12]。そして誤り箇所に対して削除処理を行い、「余剰」の誤りを訂正する。

3.3 「不足」の誤り検出・訂正

「不足」の誤りは 3.2 節で述べる「余剰」の誤りを検出する手法と同様に、文法的なルールに基づいて誤りを検出する。

格助詞は「体言」と「用言」の間に使用するのが一般的である。そのほか、主語であることを示す場合や並立の関係を示す場合に、「体言」と「体言」の間に格助詞が使用される[4]。本システムでは「体言」と「用言」または「体言」と「体言」の間に格助詞が使用される状況を細分化し、図１の②で得た分かち書き情報と詞品情報を用いて、格助詞が「不足（脱落）」する箇所を検出する。

従来のルールベースでは辞書から得られる格フレームなどの情報を参照し、「不足（脱落）」の格助詞を挿入するのが一般的だが、格フレームを構築する作業量が膨大であり、全ての状況を考慮することが難しい。そこで、本システムではBERT言語モデルを用いて、「不足（脱落）」の格助詞を推論の手法で挿入する。

BERTは2018年にGoogleから発表されたニューラル言語モデルであり、文脈を深く考慮した分散表現を生成できるのが特徴である。今回提案したシステムはTransformersで提供されているクラスBERT For Masked LMを用いる。BERT For Masked LMは一部を特殊トークン「MASK」に置き換えた文章に対して、「MASK」に入る言葉を予測し、出力確率を出すことができる。

格助詞が「不足（脱落）」する箇所を特殊トークン「MASK」に置き換えて、BERTによる得た出力確率の最も高い格助詞を挿入するという貪欲法[8]が存在するが。文に複数の誤りが含まれている場合、文の後ろの誤りが前の特殊トークンの確率判定にマイナス効果を与えることがあるため、文頭から順に「不足」する箇所を出力確率の最も高い格助詞で置き換えても、最終的に文章全体の合計出力確率の最も高いものが出力される保証はない。今の段階では、まず貪欲法を使って、「不足（脱落）」の箇所に出力確率の最も格助詞を挿入する。3.4節でビームサーチ法[13]で改めて訂正を行う。

3.4 「置換」の誤り検出・訂正

「置換」の誤りはBERT言語モデルを用いて検出・訂正を行う。文の中の格助詞を特殊トークン「MASK」に置き換えて、BERTによって得たスコアリストの上位結果をチェックすることにより、「置換」の誤りを検出することが可能である。

前節にも論じたように、貪欲法を用いて順に「置換」の格助詞を出力確率の最も高い格助詞に置き換えても、最終的に合計確率が最も高いものが出力される保証はない。そこでビームサーチ法を使用する。

$$p_i = \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)} \quad (1)$$

$$p_{tot} = \sum_{k=1}^{N} p_k \quad (2)$$

$$p_{best} = MAX(p_{tot}) \quad (3)$$

BERT For Masked LMの出力層では、式(1)に示す通り、特殊トークン「MASK」の対する出力分類スコア $s=(s_1,s_2,\ldots,s_n)$ に対して、Softmax関数を適用し、出力確率 $p=(p_1,p_2,\ldots,p_n)$ を得ることができる。まず図１の②で得た詞品情報に基づいて、文節の中の格助詞を順次に特殊トークン「MASK」で置き換える。一個目の特殊トークンを出力確率 p が上位３位の格助詞に置き換えて、3つの文章を作る。次は得られた3つの文章に対して、次の特殊トークンに同じ処理を適用し、9つの文章を作る。そこでN個目の特殊トークンで得られた9つの文章ごとの合計確率 p_{tot} が式(2)に示す通り、各特殊トークンの出力確率 p の合計値である。9つの文章の中から P_{tot} が上位３位の文章を抽出し、それ以降の特殊トークンに対して以上の処理を繰り返す。最終的に、式(3)に示す通り、合計確率の最も高い文章 p_{best} を正解として出力する。

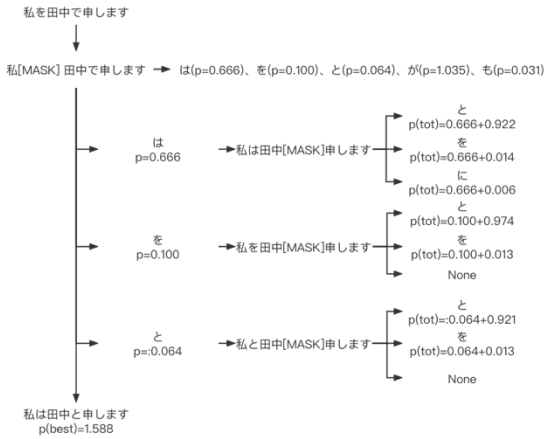

図 2 ビームサーチ法の例

4 実験

本章では、作文対訳データベース(対訳作文DB)[9]を用いて提案したシステムに対して評価を行う。対訳作文DB は

①日本語学習者による日本語作文

②作文執筆者本人による①の母語訳
③本語教師等による①の添削（ただし一部のみ）
④作文執筆者・添削者の言語的履歴に関する情報
という 4 種類のデータを大量に収集し，相互に参照することが可能な形で電子化したものである。

4.1 実験条件

今回の実験は対訳作文 DB の日本語学習者による日本語作文から、格助詞の誤りだけを含んで文法的に修正可能な文を 100 文抽出して、実験データとして用いる。日本語教師等による添削を正解データとする。抽出した文の中、正例が 179 箇所、誤用例が 124 箇所、合計 303 箇所の格助詞が使用されている。

4.2 実験結果

実験結果は表 3、表 4 で示すように、179 箇所の正例の中、152 箇所が正の結果と判断されて、訂正を行わず出力した。124 箇所の誤用例では、102 箇所が誤の結果と判断されて、その中、89 箇所が正しく訂正できて出力した。[4]と同様に検出・訂正の正答率を検出率と訂正率として求めると、

検出率…(152+102)/303=83.8%
訂正率…(152+89)/303=79.5%

という結果が得られる。

表 3　誤りの検出結果

		判定結果	
		正	誤
付与されたラベル	正	152	27
	誤	22	102

表 4　誤りの訂正結果

		訂正結果	
		未訂正	訂正済
付与されたラベル	正例	152	27
	誤例	35	89

誤文節で評価した結果が表 5 で示すように、100 誤文節の中、83 文の誤りが検出できて、71 文が修正できた。従って、文節に対する誤りの検出率が 83%、誤りの訂正率が 71%という結果が得られる。

表 5　誤文の検出及び訂正結果

	誤文	検出文	訂正文
文数	100	83	71

4.3 考察

本研究で提案したシステムの性能は，［4,5］で提案されている手法と比較すると、誤り検出率で約 8.2%、誤り訂正率で約 14.3%上回っていた。「不足」と「置換」の誤り訂正は BERT For Masked LM を用いることで、格フーレムの構築手順も約まる。

しかし、

① 一部の文法的に正しくない文節に対して、Juman++が正しく解析できないため、文法的なルールに基づいて誤りを検出する時には、誤りは検出できない。
② BERT For Masked LM の事前学習のデータは，ウィキペディアの記事であるため文章の固い表現が多い。このため BERT を用いて文節を訂正するときに「…がある」よりも「…である」の出力確率が高く、文節を正しく訂正できない場合がある。

上記の二つの問題が存在する。ルールベースの検出条件を細分化し、小説などの平文を用いて BERT For Masked LM に転移学習させることで、誤りの検出率及び訂正率の向上が期待できると考えられる。

5 おわりに

本研究は、日本語学習者が苦手とする格助詞の誤りを題材に、格助詞の余剰(Unnecessary)・不足(Missing)・置換(Replacement)の 3 種類の誤りタイプに対するそれぞれの検出・訂正手法を提案した。

「余剰」の誤りには、形態素解析の結果に基づいて、文法的なルールを適用することにより誤りを検出・訂正する。「不足」の誤りは「余剰」の誤りと同様に、文法的なルールを適用することにより誤りを検出する。訂正の段階では、BERT の推論結果を用いた。訂正率が上がる上で格フーレムの構築手順も約まった。「置換」の誤りはビームサーチ法による作文と比較することにより、誤りの検出・訂正ができた。

対訳作文 DB を用いて提案したシステムを評価した結果、83.8%の誤り検出率と 79.5%の誤り訂正率が得られた。従来のルールベースによる手法より、誤り検出・訂正率を向上させることができた。機械翻訳ベースによる手法と比べて、対訳データを使わずに、日本語の平文で構築できる利点がある。ルールベースと BERT を組み合わせた手法の文法誤り訂正における有効性を検証した。

参考文献

1. 笠原誠司, 藤野拓也, 小町守, 永田昌明, 松本裕治. 日本語学習者の誤り傾向を反映した格助詞訂正. 言語処理学会第 18 回年次大会, pp. 14-17, 2012.
2. Yu Wang,Yuelin Wang,Jie Liu,and Zhuo Liu. A compre- hensive survey of grammar error correction. arXiv preprint arXiv:2005.06600, 2020.
3. 水本智也. 自然言語処理による文法誤り訂正. 人工知能 2018 年 33 巻 6 号, p. 893-900, 2018.
4. 今枝恒治, 河合敦夫, 石川裕司, 永田亮, 桝井文. 日本語学習者の作文における格助詞の誤り検出と訂正. 情報処理学会研究報告. コンピュータと教育研究会報告, No. 13, pp. 39-46, 2003.
5. 南保亮太, 乙武北斗, 荒木健治. 文節内の特徴を用いた日本語助詞誤の自動検出・校正（語学学習支援・自動校正）. 情報処理学会研究報告. 自然言語処理研究会報告, No. 94, pp. 107-112, 2007.
6. 今村賢治, 齋藤邦子, 貞光九月, 西川仁. 小規模誤りデータからの日本語学習者作文の助詞誤り訂正. 自然言語処理 2012 年 19 巻 5 号, pp. 381-400, 2012.
7. 小川耀一朗, 山本 和英.日本語誤り訂正における誤り傾向を考慮した擬似誤り生成.言語処理学会第 26 回年次大会, pp. 505-508, 2020.
8. 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. オーム社, 2021.
9. 作文対訳データベース.（引用日： 2021 年 12 月 10 日.） https://mmsrv.ninjal.ac.jp/essay/
10. Christopher Bryant,Mariano Felice,Ted Briscoe.Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction.In Association for Computational Linguistics,pp793-805,2017.
11. 日本語形態素解析システム JUMAN.（引用日： 2021 年 12 月 10 日.）https://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
12. 内田茂. 留学生の日本語作文に見られる助詞の誤用について. 教育研究所紀要 34 号, pp. 45-50, 1998.
13. 甫立健悟, 金子正弘, 勝又智, 小町守. 文法誤り訂正における訂正度を考慮した多様な訂正文の生成. 自然言語処理 2021 年 28 巻 2 号, pp. 428-449, 2021.