分類モデルを用いた日本語学習者の格助詞誤り訂正

小川 耀一朗　　山本 和英
長岡技術科学大学
{ogawa, yamamoto}@jnlp.org

1 はじめに

日本に住む外国人の人口が年々増加している中で，日本語非母国語話者のための作文支援・訂正ツールの必要性は高まっている．文法誤り訂正システムは，非母国語話者の作文に含まれるあらゆる文法的な誤りを自動で訂正するシステムであり，日本語教師の作文チェック作業の支援や，e ラーニングとしての学習者の言語習得支援などに活用することができる．NAIST 誤用コーパス [1] から調べた日本語学習者が誤る箇所の割合では助詞の誤りが最も多いことから，学習者にとって助詞の使い方が難しいことがわかる．特に格助詞は文の意味を理解する上で重要であるため，格助詞誤り訂正システムの需要は高い．これまでの格助詞誤り訂正手法は訂正単語の周辺の情報のみを用いていた．しかし，離れた係り受け関係や，他の機能語との組み合わせなど，文全体を考慮した訂正が必要である．そこで本研究では，文全体を考慮した格助詞誤り訂正手法のための分類モデルを提案する．格助詞「が・を・に・で」を対象に，文中に誤りが 1 箇所のみ，かつ誤り箇所が既知であるという問題設定において，提案手法はベースラインである言語モデルを用いた手法よりも正解率 (%) が 8.59 ポイント向上し，提案手法の有効性を示した．

2 関連研究

英語の文法誤り訂正はこれまで様々な手法が提案されており，大きく機械翻訳手法と分類手法の 2 つに分けることができる．機械翻訳手法は誤りを含む文を正しい文に “翻訳する” タスクとして扱う手法である． Junczys-Dowmunt らは統計的機械翻訳 (statistical machine trainslation; SMT) を文法誤り訂正に適用し，学習者作文と添削文が対となっている “対訳” コーパス (以下，学習者コーパスと称す) を学習させることで高い訂正性能を示した [2]．近年ではニューラルネットワークを用いた手法が成功を収めており，Zheng らはニューラル機械翻訳 (neural machine trainslation; NMT) を初めて文法誤り訂正に適用した [3]．Chollampatt らは多層の畳み込み層で構成された encoder-decoder モデルを提案し，SMT 手法の性能を上回った [4]．Roman らは SMT と NMT を組み合わせた手法を提案し，最も高い性能を示している [5]．

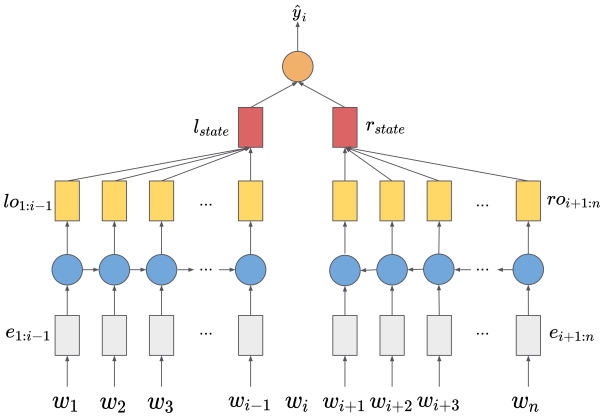

図 1: RNN モデル

分類手法は，誤っている単語に対して訂正候補の中から正しい単語を選択する分類問題として扱う手法である．Christopher らは大規模な単言語コーパスで訓練した N-gram 言語モデルを用いて，訂正候補の中から最もスコアが高くなる単語に訂正する手法の有効性を示した [6]．Wang らはリカレントニューラルネットワーク (RNN) を用いて入力文のコンテキスト情報から正しい単語を予測する分類手法を提案し [7]，さらにアテンション機構を加えることで機械翻訳手法に劣らない性能を示した [8]．分類手法には，機械翻訳手法とは対称的に，構築コストが非常に大きい学習者コーパスを必要としないというメリットがある．

日本語の文法誤り訂正は，あらゆる誤り種類を考慮した手法と，特定の誤り種類に限定した手法がある．水本らは Lang-8 から日本語学習者の作文を収集して構築した学習者コーパスで SMT を訓練させ，あらゆる誤りを訂正する実験を行なった [9]．笠原らは訂正対象を格助詞に限定し，Noisy channel model を用いて学習者の誤り傾向を考慮した訂正手法を提案した [10]．あらゆる誤りを訂正するには，まずは学習者の間違え

やすい格助詞の訂正性能を向上させる必要がある．格助詞の誤りを含む添削付きの学習者データは非常に少なく，機械翻訳手法では十分な精度は望めない．一方，分類手法は容易に手に入る大規模な日本語コーパスを用いることができる．そこで本研究では格助詞誤り訂正のための分類モデルを提案する．これまでの訂正対象単語の周辺情報のみを用いた手法と比べ，提案手法では文全体の情報を考慮することが可能となる．

3 格助詞誤り訂正

誤りの種類は多岐に渡るが，それぞれの誤りに特化した訂正システムを繋ぎ合わせることで汎用的なシステムを構築することができる．本研究では，以下の問題設定において訂正性能の向上に努める．(1) 格助詞「が・を・に・で」を訂正対象および訂正候補とする．(2) 文中に対象誤りが 1 箇所のみ．(3) 誤り箇所には正解ラベルが付けられており，明示されている．

NAIST 誤用コーパス [1] から調べた日本語学習者が誤る箇所の割合では，助詞誤りが最も多く，全体の約 23%であった．笠原ら [10] は計 18 種類の助詞を訂正対象として実験を行なったが，本研究ではその中の，日本語作文において特に大きな役割を担っている「が・を・に・で」の 4 種類に限定して検証を行う．NAIST 誤用コーパスに付与されている誤用タグから，この 4 種類の格助詞が占める割合を集計したところ，助詞誤りタグの中では約 38%，全ての誤りタグの中では約 9%であった．

3.1 分類モデル

RNN モデルと CNN モデルの 2 つを構築した．RNN モデルは Wang ら [8] が提案した，アテンション付き RNN モデルを参考にして構築した．

図 1 に RNN 分類モデルの構成図を示す．まず，単語数 n の入力単語列 $\mathbf{w}$ 中の誤り単語 w_i を境に，文を左文脈と右文脈の 2 つに分割する．左エンコーダと右エンコーダそれぞれで，単語の分散表現 $e_{1:i-1}$ 及び $e_{i+1:n}$ を 2 層の LSTM ネットワークに入力し，得られた中間表現 $lo_{1:i-1}$ 及び $ro_{i+1:n}$ から，アテンションの状態 l_{state} 及び r_{state} を計算する．

$$score(lo_t) = lo_t^T W_a lo_{i-1} \quad (1)$$

$$a(t) = \frac{exp(score(lo_t))}{\sum_{j=1}^{i-1} exp(score(lo_j))} \quad (2)$$

$$l_{state} = (\sum_{t=1}^{i-1} a(t) lo_t) \oplus lo_{i-1} \quad (3)$$

r_{state} も同様に計算するが，LSTM は左エンコーダとは逆向きの右から左の方向に計算する．学習には multilayer perceptron(MLP) を用いる．左右のアテンション状態を結合し，ソフトマックスを通して各ラベルの確率を計算する．$L(x)$ は全結合層である．

$$L(x) = Wx + b \quad (4)$$

$$MLP(x) = softmax(L(ReLU(L(x)))) \quad (5)$$

最も確率の高いラベルをモデルの出力とし，正解ラベルとの交差エントロピーを最小化するための学習を行う．N は訓練データサイズ，$\hat{y}_i$ は予測ラベル，y_i は正解ラベルである．

$$loss = \frac{1}{N} \sum_{i=1}^{N} y_i \log \hat{y}_i \quad (6)$$

CNN モデルは，RNN モデルでのエンコーダ部分のネットワークを CNN に置き換えたモデルである．Le らはカーネルウィンドウ幅が 3,4,5 の 3 つの畳み込み層を用いた Shallow-and-wide CNN モデルを提案し，文章分類タスクにおいて高い性能を示した [11]．これを参考にし，エンコーダに Shallow-and-wide CNN を適用した CNN モデルを構築した．

共通して，入力の語彙数はコーパスの上位 4 万種類とし，それ以外は<UNK>ラベルに置き換えた．また，単語の分散表現の学習における初期値には NWJC2Vec[12] を利用した．これは国立国語研究所が「国語研日本語ウェブコーパス」に基づいて構築した，事前学習された単語の分散表現データである．

3.2 訓練データ

分類モデルの訓練には BCCWJ [1] を使用した．まず，このコーパスから対象格助詞が 1 つ以上含まれる計 3,313,451 文を抽出した．次にそれらのいずれか 1 つを正解ラベルとし，その左右の単語列をモデルの入力データとした．加えて，Lang-8 コーパス [9] も使用した．Lang-8 コーパスは言語学習 SNS サービス Lang-8[2]からデータを収集したものであり，学習者作文とその添削文の対となっている．このコーパスの添削文を正しい日本語文として考え，BCCWJ と同様のデータ前処理を行なった．ただし，“汚い” 文も多く含まれていたため，重複文の削除及び日本語と英語の混用文の削除を行ない，計 900,596 文となった．モデルの訓練にはなるべく大きいコーパスが望ましいため，この 2 つのコーパスを足し合わせて使用した．

形態素解析には形態素解析エンジン MeCab[3]及び UniDic 辞書[4]を使用した．

[1] https://pj.ninjal.ac.jp/corpus_center/bccwj/
[2] http://lang-8.com/
[3] http://taku910.github.io/mecab/
[4] https://unidic.ninjal.ac.jp/

表 1: 格助詞誤り訂正の実験結果

モデル	正解率 (%)
言語モデル (ベースライン)	74.68
CNN	81.23
CNN+emb	82.34
CNN+emb+kana	83.09
LSTM	78.07
AttnLSTM	82.53
AttnLSTM+emb	83.27
AttnLSTM+emb+kana	83.27

表 2: 出力例

テスト文	正解	言語モデル	提案モデル
この中には硫黄や硫化化合物<を>あります。	が	で	が
人々はこの規制<を>賛同の意志を表明しました。	に	が	に
それで貧困<を>解決できます。	が	を	を
たとえば、動物園や遊園地など<が>ある。	で	で	が

3.3 テストデータ

テストデータにはNAIST 誤用コーパスを使用した．これは国立国語研究所により収集された「日本語学習者による日本語作文と，その母語訳との対訳データベース」に誤用タグを付加したものである．NAIST 誤用コーパスの 6,730 文から，誤りが対象格助詞かつ訂正先も対象格助詞であるタグを含む 548 文を抽出し，テストデータとした．また対象格助詞以外の誤りはあらかじめ正しく訂正しておき，対象格助詞誤りが複数ある場合はランダムに 1 つを選択してそれ以外を正しい格助詞に置換しておいた．評価は誤用タグに記載されている添削とモデルの出力を比べたときの正解率である．

3.4 ベースライン

本実験のベースラインとして，言語モデルを用いた手法での格助詞誤り訂正実験を行なった．言語モデルは KenLM toolkit[13] を用いて単語 4-gram 言語モデルを構築した．3,4,5,6-gram で実験し，4-gram のとき最も性能がよかった．

4 実験結果と考察

表 1 に BCCWJ と Lang-8 を足し合わせたコーパスで訓練させた各モデルの実験結果を示す．“+emb” は事前学習された単語の分散表現を用いた場合である．また，“+kana” は訓練データ及びテストデータを全て平仮名に変換した場合である．ベースラインである言語モデルを用いた手法と比べて，どの提案手法も正解率が上回った．RNN 分類モデルでは，アテンション付き RNN(AttnLSTM) はアテンションなしのとき (LSTM) に比べてスコアが 4.46 ポイント向上した．CNN と RNN のどちらにおいても，事前学習された単語分散表現を用いることで正解率が向上した．平仮名に変換したときの結果は，RNN ではスコアが変わらなかったが，CNN では向上した．学習者は平仮名を多用し，訓練データに出現しないようなひらがな表記は性能を下げる．従って平仮名に正規化することの効果はあると考える．最終的には RNN モデルの方がわずかに高いスコアとなり，ベースラインよりも 8.59 ポイント向上した．

表 3: コーパスの漢字の割合及び平均文長

コーパス	漢字の割合 (%)	平均文長 (文字数)
NAIST 誤用コーパス	24.7	28.6
Lang-8	29.0	30.1
BCCWJ	35.7	32.8
日経新聞コーパス	54.0	47.7

表 4: コーパスの比較実験での正解率 (%)

	NIKKEI(3M)	BCCWJ(3M)
言語モデル	64.70	71.53
CNN+emb	73.61	82.16
AttnLSTM+emb	75.46	82.16

表 2 に，言語モデルでは正しく訂正できなかったが提案手法により改善された例 (1，2 番目)，および提案手法では訂正できなかった例 (3，4 番目) を挙げる．1 つ目の例は，訂正に必要なフレーズである「この中には」が 4-gram 言語モデルでは捉えられずに間違えている．2 つ目の例は，前後にある他の助詞との兼ね合いを考慮できていない．提案手法ではこれらを正しく訂正しているため，訂正対象単語の周辺だけでなく離れた単語との関係や他の機能語などを考慮することができている．3 つ目の例は，正解は “が” であるが，文法的には “を” でも正しい例である．学習者コーパスの中にはこのような正解が複数ある場合でも 1 つしか付与されていない事例がいくつか見られた．4 つ目の例も，正解は “で” だが，“が” でも間違ってはいない．この場合は前後の文や更に広い文脈情報が必要となる．

高い性能のためには，学習者コーパスに “近い” コーパスを訓練させる必要があると考える．表 3 に 4 つのコーパスでの漢字の割合及び平均文字数を集計した結果を示す．NAIST 誤用コーパスと Lang-8 はどちらも学習者コーパスであるからか，近い値となっている．

日本経済新聞コーパスは新聞記事データであり，漢字の割合及び平均文字数の値は大きく，学習者コーパスとは離れていると言える．一方，BCCWJ は日本経済新聞コーパスよりも学習者コーパスに近いと言える．そこで日本経済新聞コーパスと BCCWJ のそれぞれでモデルを訓練させたときの実験結果を表 4 に示す．コーパスサイズは比較のため 300 万文 (3M) に統一した．Lang-8 は文数が少なく十分な訓練ができなかったので割愛する．3 つのモデル全てにおいて，コーパスサイズが同じであるにも関わらず，BCCWJ で訓練させた方が高い正解率を示し，RNN モデルでは 6.70 ポイントの向上が見られた．このことから，誤り訂正にとって学習者コーパスに“近い”コーパスは重要であると言える．今回は漢字の割合と平均文字数を尺度としたが，他にも重要な尺度は存在すると考えられる．日本語の学習者コーパスは貴重であり，構築も難しい．しかし大規模な日本語コーパスから学習者作文に近い文を選定することができれば，性能を向上させる訓練データの拡充が容易になると考える．

5 おわりに

本研究では格助詞「が・を・に・で」を対象にした分類モデルによる訂正手法を提案した．CNN モデルと RNN モデルを構築し，RNN モデルでは言語モデル手法よりも正解率 (%) が 8.59 ポイント向上した．また，提案手法により改善された例を示し，訂正対象単語の周辺だけでなく文全体を考慮した訂正ができていることを確認した．2 つのドメインの異なるコーパスでの実験では，同じサイズでも学習者コーパスに近いコーパスで訓練させた方が高い性能を示したことから，性能の向上に貢献する訓練コーパスの選定が重要であることを確認した．

本研究では訂正対象が 4 種類の格助詞，誤りが文中に 1 つのみ，誤り箇所が既知という問題設定で実験を行なったが，実際には多くの種類の誤りがあり，また誤りは 1 文中に複数箇所存在する可能性が十分にあり，かつ誤り箇所も未知である．しかし，日本語文法誤り訂正に関する研究があまり進んでいない今，このような問題設定は今後の調査にとって現実的で有益であると考える．今回の知見を活かし，今後はさらに現実的な問題に取り組んでいきたい．なお，今回取り組んだ格助詞誤り訂正を Web で利用できるツールとして公開する予定だ．

謝辞

本研究は，平成 27〜31 年科学研究費助成金基盤 (B) 課題番号 15H03216，課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」の助成を受けています。

参考文献

[1] 大山浩美, 小町守, 松本裕治. 日本語学習者の作文における誤用タイプの階層的アノテーションに基づく機械学習による自動分類. 自然言語処理, Vol. 23, No. 2, pp. 195–225, 2016.

[2] Marcin Junczys-Dowmunt and Roman Grundkiewicz. Phrase-based Machine Translation is State-of-the-Art for Automatic Grammatical Error Correction. Association for Computational Linguistics, pp. 1546–1556, 2016.

[3] Yuan Zheng and Briscoe Ted. Grammatical error correction using neural machine translation. Association for Computational Linguistics, pp. 380–386, 2016.

[4] Chollampatt, Shamil and Ng, Hwee Tou. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[5] Grundkiewicz Roman and Junczys-Dowmunt Marcin. Near Human-Level Performance in Grammatical Error Correction with Hybrid Machine Translation. Association for Computational Linguistics, pp. 284–290, 2018.

[6] Bryant Christopher and Briscoe Ted. Language Model Based Grammatical Error Correction without Annotated Training Data. Association for Computational Linguistics, pp. 247–253, 2018.

[7] Chuan Wang, RuoBing Li and Hui Lin. Deep Context Model for Grammatical Error Correction. In Proc. 7th ISCA Workshop on Speech and Language Technology in Education, pages 167âĂŞ171, 2017.

[8] Zhu Kaili, Chuan Wang, Ruobing Li, Yang Liu, Tianlei Hu and Hui Lin. A Simple but Effective Classification Model for Grammatical Error Correction. arXiv preprint arXiv:1807.00488, 2018.

[9] 水本智也, 小町守, 永田昌明, 松本裕治. 日本語学習者の作文自動誤り訂正のための語学学習 SNS の添削ログからの知識獲得. 人工知能学会論文誌, Vol.28, No.4, pp.420–432, 2013.

[10] 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 回年次大会発表論文集, pp. 14–17, 2012.

[11] Hoa T. Le, Christophe Cerisara and Alexandre Denis. Do Convolutional Networks Need to Be Deep for Text Classification?. arXiv preprint arXiv:1707.04108, 2017.

[12] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木 稔. nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.

[13] Kenneth Heafield. KenLM: faster and smaller language model queries. In Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation, pp. 187–197, 2011.