〔2018 年度日本語教育学会秋季大会（プラサ ヴェルデ，2018.11.25）発表・ポスター発表⑦〕【予稿集】

「間違いが直す」格助詞誤り訂正システム

小川耀一朗（長岡技術科学大学大学院生）・山本和英（長岡技術科学大学）

1. はじめに

自然言語処理の分野では，非母国語話者の書く文章の中にある文法的な誤りを自動で訂正する研究が行われている。この文法誤り訂正タスクは，ある文章が与えられたときに，あらゆる文法的な誤りが訂正された文章を返すことが目標である。

自然言語処理の技術の活用が様々な領域で進む中で，言語教育の支援にも応用が期待されている。誤り訂正システムは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できると考えられる。

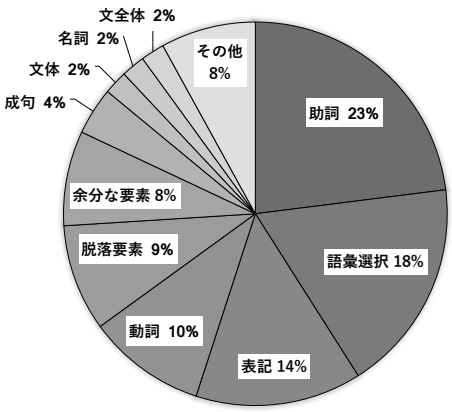

図 1 NAIST 誤用コーパスでの誤り傾向

日本語における文法的な誤りの種類は多岐に渡る。図 1 は NAIST 誤用コーパス（大山 2012）（大山 2016）における，誤り箇所の割合を示したものである。助詞の誤りが 23%を占めており，学習者にとって助詞が最も間違えやすいことがわかる。特に格助詞は文の意味を理解する上で重要である。そこで本研究では，日本語学習者の格助詞誤りを訂正するシステムを構築する。

2. 関連研究

文法誤り訂正タスクの研究では，誤りを含む文を正しい文に“翻訳する”タスクとして扱う機械翻訳手法が多く提案されている（水本 2013）（Yuan 2016）。機械翻訳手法は，学習者の書いた文とその訂正された文のペアで構成された学習者コーパスをモデルに学習させることで訂正を行う手法である。機械翻訳手法では大量の学習者コーパスを学習させる必要があるが，日本語で書かれた学習者コーパスの開発事例は少ない。Lang-8[(1)]から日本語の学習者データを収集し，規模の大きいコーパスで実験を行なった報告（水本 2013）では実用化できるまでの精度に至らなかった理由に学習者コーパスの不足を挙げている。

他の手法として，言語モデルを用いた訂正手法（Bryant 2018）や，訂正候補の中から正しい単語に分類する分類手法（Kaili 2018）がある。言語モデルを用いた手法では，対象言語のテキストコーパスのみで構築した言語モデルを用いるため，学習者コーパスを必

要としない。また，訂正対象を格助詞誤りに限定しているため，機械翻訳のようなあらゆる誤りを一度に“翻訳する”必要性は薄い。そこで本研究では，言語モデルを用いて格助詞誤りを訂正する。

3. 日本語学習者の誤り傾向

NAIST 誤用コーパス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」に誤用タグを付与したコーパスである。

図 1 に示した通り，学習者にとって助詞が最も間違えやすい。図 2 に NAIST 誤用コーパスでの助詞誤りの内訳を示す。助詞の不足による誤りの割合が 27%と最も多いことがわかる。また助詞の中でも「が・は・に・を・の・で」の誤用が多い。先行研究（笠原 2012）では全ての格助詞の誤用を対象に訂正を行ったが，助詞の不足に対しての補完は行ってない。しかし，日本語学習者は助詞を欠落してしまいやすいという傾向があるため，不足している助詞を補う処理は不可欠であると考える。本研究では，誤り頻度の高い格助詞「が・を・に・で」を対象に，誤用の訂正と不足箇所の補完を行う。

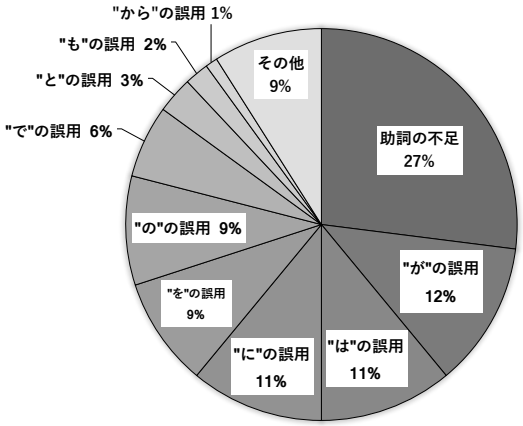

図 2 NAIST 誤用コーパスでの助詞誤りの内訳

4. 言語モデルに基づく格助詞誤り訂正手法

本研究では言語モデルを用いて誤り訂正を行う。言語モデルとは，文の自然さを表現するモデルであり，入力文sの自然さを言語モデル確率$P(s)$で表す。より自然な文には高い確率を与え，不自然な文には低い確率を与える。本稿では言語処理分野で広く利用されている，単語 n-gram モデルを使用する。

単語数がTの文 $s = s_1 \dots s_T$をs_1^Tと表記する。文sの言語モデル確率$P(s)$は以下のような式で表される。

$$P(s) = P(s_1^T) = \prod_{t=1}^{T+1} P(s_t | s_{t-n+1}^{t-1})$$

ここで$P(s_t | s_{t-n+1}^{t-1})$は条件付き確率であり，あらかじめ訓練データから全ての n-gram の頻度を算出して作成した辞書から計算する。

言語モデルを用いた誤り訂正は，誤りを含む文の言語モデル確率は誤りを含まない文と比べて低くなるだろうという考え方に基づいている。例えば「車が買う」の言語モデル確率は「車を買う」の言語モデル確率よりも低くなる。以下に訂正の手順を示す。

(1) 入力文の形態素解析を行い，各単語に対して以下の処理を繰り返し行う
(2) 置換
    ① 単語が格助詞「が・を・に・で」のいずれかの場合，他の各格助詞で

置換した訂正候補文を生成する

② 元の文および各訂正候補文の言語モデル確率を計算し，最も確率の高くなる文を選択する

(3) 補完

① 品詞が名詞または代名詞または接尾辞-名詞的の単語の後に，品詞が助詞または助動詞の単語が続いていなければ，格助詞不足と判断する

② 格助詞の不足と判断された箇所に，格助詞「が・を・に・で」をそれぞれ挿入した訂正候補文を生成する

③ 元の文および各訂正候補文の言語モデル確率を計算し，最も確率の高くなる文を選択する

各単語に対しての繰り返し処理を行う際は，文末の単語から文頭の単語まで順番に処理を行った。これについての考察は 7 章で述べる。

5. 格助詞誤り訂正実験

5-1. 言語モデルの構築

言語モデルには日本経済新聞記事コーパスを使用した。このコーパスは約 1800 万文の新聞データが収録されている。日本経済新聞記事コーパスの形態素解析には MeCab[(2)] と UniDic 辞書[(3)]を用いた。言語モデルは KenLM toolkit (Heafield 2011) を用いて単語 4-gram 言語モデルを構築した。

5-2. テストデータ

NAIST 誤用コーパスから格助詞「が・を・に・で」の誤用およびそれらの不足に関するタグが付与されている 879 文を抽出した。また文中の対象タグ以外の誤りは NAIST 誤用コーパスのタグに付与されている訂正先に置き換え，対象の誤りのみが含まれているデータセットを作成した。

5-3. 評価方法

適合率，再現率，F 値で評価する。適合率とは訂正を行った回数のうち，正しい訂正であった回数の割合である。再現率とは，誤っている箇所のうち，正しく訂正された箇所の割合である。F 値は適合率と再現率の調和平均で表される。

$$\text{適合率} = \frac{\text{正しい訂正の数}}{\text{訂正を行った回数}}$$

$$\text{再現率} = \frac{\text{正しい訂正の数}}{\text{誤り箇所数}}$$

$$F\text{値} = \frac{2\times\text{適合率}\times\text{再現率}}{\text{適合率} + \text{再現率}}$$

6. 結果

表 1 格助詞誤り訂正の実験結果

手法	適合率	再現率	F 値
ランダム	6.56%（174/2652）	44.1%（174/395）	11.4%
頻度	7.43%（206/2774）	49.2%（206/419）	12.9%
言語モデル	**50.9%**（493/969）	**60.3%**（493/817）	**55.2%**
言語モデル（置換）	49.3%（395/801）	69.3%（395/570）	57.6%
言語モデル（補完）	58.3%（98/168）	39.7%（ 98/247）	48.0%

表 1 に結果を示す。提案手法である言語モデルを用いた手法での格助詞訂正は，適合率が 50.9%，再現率が 60.3%，F 値が 55.2%となり，格助詞を選択するときにランダムに選択する方法（ランダム）及びテストコーパスでの訂正先となる頻度が最も高い格助詞を選択する方法（頻度）よりも高い性能を示した。

また，言語モデル手法での性能を置換と補完で切り分けて計測した結果も示す。置換での訂正では再現率が高く，適合率が低い。それに対して補完での訂正は適合率が高く，再現率が低い。置換では，誤り箇所を正しく訂正する割合が高い一方，本手法では文中の全ての対象格助詞を訂正の対象とするため，正しい格助詞を間違った格助詞に訂正してしまう場合が多く，適合率を下げる要因となっている。補完では，訂正すべき箇所数（247 箇所）に対して訂正を行った回数（168 回）が少ないため，再現率は適合率よりも低くなってしまう。

7. 議論

表 2 n-gram を変化させたとき結果

手法	適合率	再現率	F 値
3-gram 言語モデル	48.1%	59.2%	53.1%
4-gram 言語モデル	**50.9%**	**60.3%**	**55.2%**
5-gram 言語モデル	50.3%	59.8%	54.6%
6-gram 言語モデル	50.5%	59.7%	54.7%

表 3 訂正の順序を変化させたときの結果

手法	適合率	再現率	F 値
4-gram 言語モデル（文頭から訂正）	49.7%	57.1%	53.2%
4-gram 言語モデル（文末から訂正）	**50.9%**	**60.3%**	**55.2%**

表 2 に n-gram を変化させたときの言語モデル手法の格助詞誤り訂正の結果を示す。笠原（2012）は 3-gram 言語モデルを用いていたが，本実験では 4-gram が最も高い性能を示した。

表 3 に，本手法において文中の各単語に対して繰り返し処理を行う際に，文頭から順に行う方法と文末から順に行う方法のそれぞれの結果を示す。Bryant（2018）は文頭から順

に訂正を行っているが，本実験では文末から順に訂正を行う方が良い性能を示した。英語とは違い，日本語は主体の動作を表す述語を文の後ろの方に記述するため，より正しい格助詞を決定しやすい順番に訂正を行うことができるためと考えられる。

表 4 システムの出力例

	入力文	出力	正解	
1	それで学校に退学させられた，という話をよく聞きます。	を	を	正解
2	以上 *Φ* 私の意見です。	が	が	正解
3	人々はこの規制を賛同の意志を表明しました。	が	に	不正解
4	例えばテレビや洗濯機など *Φ* よく贈ります。	*Φ*	を	不正解
5	ときどき結婚式の時 *Φ* 音楽を流します。	に	*Φ*	不正解
6	この中には硫黄や硫化化合物をあります。	で	が	不正解
7	さいごで，次のような言い伝えも昔から伝わっている。	が	に	不正解

表 4 にシステムの出力例を示す。ただし *Φ* は何もないことを表している。1,2 は正解例，3〜7 は不正解例である。3 は置換訂正したが間違えた例である。前後の格助詞をより注視することでこのような間違いを防ぐことができるのではないかと考える。4 は補完するべき箇所を補完していない例である。「など」は助詞と解析されるため，補完操作が行われない。名詞,代名詞,接尾辞-名詞的の単語の後に，助詞,助動詞の単語が続いていなければ，格助詞の不足と判断したが，この条件では捉えられない不足誤りが存在するため，補完の条件を今後さらに検討したい。5 は補完しなくてよい箇所を補完してしまった例である。しかし，補完した方が自然な文であると思われる。NAIST 誤用コーパスにこのような訂正してもしなくても正しいような事例がいくつか見られた。6 は置換訂正したが間違えた例である。入力文を単語分割すると「この/中/に/は/硫黄/や/硫化/化合/物/を/あり/ます/。」となり，正しい訂正を行うために必要な「この中には」を，4-gram では捉えることができないため間違えてしまうと考えられる。文中の単語を広く捉えることのできる手法により精度が改善すると考える。7 は置換訂正が間違えた例である。「さいご」が平仮名で書かれており，訓練コーパス中に「さいご」から始まる文が含まれていないため，正しい判断ができなくなってしまうと考えられる。学習者の書く作文には漢字を使わず平仮名を多用する傾向があり，訓練コーパスに含まれない，もしくは正しい単語分割がされない事例が多く発生する。平仮名の多用への対策が今後の課題として挙げられる。

8. まとめ

本研究では学習者にとって最も間違えやすい格助詞を訂正対象とし，言語モデルを用いた格助詞誤り訂正システムを構築した。助詞誤りの中でも助詞の不足による誤りの割合が高いため，訂正手順に格助詞の補完操作を組み込むことで助詞不足の訂正に取り組んだ。NAIST 誤用コーパスでの訂正実験の結果，置換訂正に対して補完訂正の再現率が低いため，補完の訂正方法の改善が必要である。システムの出力例を見ると，対象格助詞の近傍だけ見ても正しい格助詞を選択することができない事例があった。格助詞の近傍の単語だけで

なく，機能語のような文法を構成する際に重要となる単語にも着目した仕組みを取り入れることで性能の向上につながるのではないかと考えている。また，日本語学習者の作文には平仮名が多用される傾向があるため，形態素解析誤りが生じてしまう。日本語における文法誤り訂正の性能向上のために，日本語学習者作文の平仮名に対する形態素解析の対策が必要である。

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。

注

(1) http://lang-8.com/

(2) http://taku910.github.io/mecab/

(3) http://unidic.ninjal.ac.jp/

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) 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，28 巻，5 号，pp. 420-432，2013.

(3) 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グつきコーパスの構築について-NAIST 誤用コーパスの開発-． 第一回テキストアノテーションワークショップ，2012.

(4) 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， 23 巻， 2 号， pp. 195-225，2016.

(5) Christopher Bryant, Ted Briscoe. Language Model Based Grammatical Error Correction without Annotated Training Data. Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 247-253, 2018.

(6) 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.

(7) Zheng Yuan, Ted Briscoe. Grammatical error correction using neural machine translation. Proceedings of NAACL-HLT 2016, pp. 380-386, 2016.

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