日本語学習者の誤り傾向を反映した格助詞訂正

笠原 誠司 † 藤野 拓也 † 小町 守 † 永田 昌明 †† 松本 裕治 †

† 奈良先端科学技術大学院大学
情報科学研究科
{seiji-k, takuya-fu, komachi, matsu}@is.naist.jp

†† NTT
コミュニケーション科学基礎研究所
nagata.masaaki@lab.ntt.co.jp

1　はじめに

図 1 は 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を示したものであるが，助詞の誤りが 24% を占めており，学習にとって助詞が誤りやすい箇所であることが見て取れる．特に格助詞が間違っていると文の意味が理解しづらくなるため，学習支援での格助詞訂正の需要は高い．学習者がどの格助詞とどの格助詞を混同しやすいのかが分かれば精度の高い誤り訂正ができるはずだが，これまでの格助詞訂正のタスクで用いられてきた情報は，単語の頻度や品詞，係り受け関係などの，日本語母語話者が書いた文を元にして推測されるものであり，学習者の作文から得た誤り傾向を用いる試みは行われていなかった．近年の相互添削 SNS 利用者の増加により，大規模な学習者が書いた作文の添削情報が入手可能になり，信頼度の高い誤りの傾向を抽出できるようになった．そこで，本研究は日本語学習者の格助詞誤り傾向を反映した格助詞訂正手法を提案する．実験の結果，提案手法は言語モデルのみによって訂正するベースラインと比較して 7.6% 高い正解率を達成し，提案手法の有効性を示した．

2　関連研究

日本語学習者が誤りやすい箇所であるため，格助詞を訂正する試みは過去にも行われてきた．格助詞の自動訂正手法には，大きく分けて，格フレーム情報などの人手で作成された辞書を利用して訂正する手法と，機械学習を用いる手法の 2 つがある．

まず辞書を用いた手法について述べる．今枝らは NTT 日本語語彙大系を利用して誤りを検出し訂正する手法を紹介した [5]．南保らは文節内の特徴を用い，帰納的学習を行うことで，今枝らの手法と同等の結果が得られることを示した [7]．

次に機械学習を用いた手法について述べる．大山らは学習者の格助詞誤り検出を機械学習手法の SVM を用いて高い精度で行った [2]．大木らは，大山らの手法に，日本語文の誤り検出に有効な素性を新たに追加して再現率や適合率の改善を行った [6]．大山らや，大木らの手法は学習者の格助詞誤り訂正が目的という点で一致しているが，訂正対象となる助詞は 1 種類であり，提案手法ではすべての助詞を対象にしている点に違いがある．いずれの手法も学習者の誤りの傾向を反映したものではない．

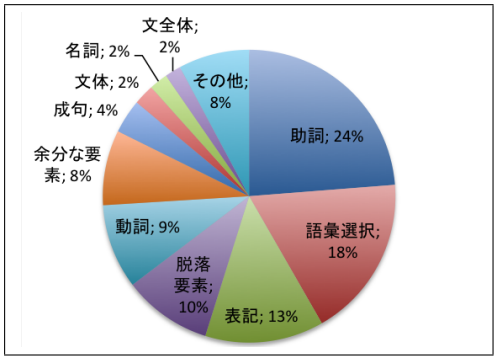

図 1　日本語学習者の作文誤りの傾向

鈴木らは機械翻訳の後処理として格助詞訂正を行った [4]．機械学習には最大エントロピー法を用いており，複数の格助詞を同時に高い精度で訂正しているが，係り受けや品詞などの情報を学習の素性として使用している．学習者の作文が対象のとき，これらの情報を正しく推定することは困難なので，学習者の作文の格助詞訂正に応用すると精度が低下する可能性がある．我々の手法は係り受けや品詞などの推定が困難な情報は使用していないので誤りを含む文でも頑健に対応できる．

Rozovskaya らは英語の非母語話者が書いた作文の前置詞訂正タスクで，学習者の母語に基づいた誤りの傾向を用いることにより精度が改善される事を示した [3]．学習者の誤り傾向を反映するという点で我々のタスクと共通しているが，対象としている言語が異なっているところに大きな違いがある．水本らは学習者が書いた作文のデータを用いて，統計的機械翻訳の手法を用いた日本語文誤り訂正を行った．日本語学習者の書いた作文を用いて誤り訂正を試みている点で類似しているが，訂正対象を助詞に限定しておらず，手法も異なっている [1]．

3　日本語学習者の格助詞訂正

格助詞誤りの訂正は誤り検出のタスクと誤り訂正のタスクにわけることができる．誤り検出のタスクとは，学習者の書いた作文中から誤りを含んでいる箇所を推測するタスクである．誤り訂正のタスクとは，誤りのある箇所がわかっていることを前提とし，各誤りがそれぞれ何に直されるべきなのかを推測するタスクである．実用時

[*1] NAIST 誤用コーパスについては 6.1.2 節で詳しく述べる

表 1　NAIST 誤用コーパスでの訂正先助詞の総数上位 10 個

訂正先の助詞	総数	割合
は	471	13.93%
に	419	12.40%
が	414	12.25%
を	375	11.09%
（空白）	371	10.98%
の	368	10.89%
で	236	6.98%
も	121	3.58%
では	80	2.37%
や	72	2.13%

には両方のタスクを段階的に，あるいは同時に解決する必要があり，それぞれのタスクが高精度で行われることが望ましい．本タスクは誤り訂正のタスクに焦点を当てる．従って，テストデータにおいて誤り箇所にはタグが付けられており，明示的に示されているものとする．

本稿では助詞の一部を予測訂正の対象とした．対象とする助詞は以下のような理由で選択したが，これらは鈴木らが使用したものを踏襲した．鈴木らの研究目的は我々のものと異なっているが，使用する助詞の選択は我々のタスクにおいても妥当であると判断した．対象とする助詞を選択するにあたり，学習者の誤り訂正にとって重要である助詞を調べるため，NAIST 誤用コーパスに付与されているタグから訂正先（添削者が付与した正解ラベル）の助詞の数を求めた．全助詞タグ数 3,380 に対する割合とともに表 1 に上位 10 個を示す．

■格助詞　補語が述語に対してどのような関係にあるかを表す助詞で，日本語文の生成で大きな役割を担っている．「が，を，の[*2]，に，から，と，で，へ，まで，より」の格助詞すべてを対象とした．

格助詞に加え係助詞の「は」も対象に含める．なぜなら，「は」は訂正先の 13.93% を占めており「が」や「を」に匹敵するほど頻度が高いためである．10 種類の格助詞と係助詞の「は」に加え，格助詞と「は」の組み合わせである，「には，からは，とは，では，へは，までは，よりは」も含めた，合計 18 の助詞を訂正タスクの対象とした[*3]．本稿で紹介する手法では，

$$
\begin{aligned}
&kakujoshiList = \\
&(\text{が}, \text{を}, \text{の}, \text{に}, \text{から}, \text{と}, \text{で}, \text{へ}, \text{まで}, \text{より}, \text{は}, \\
&\text{には}, \text{からは}, \text{とは}, \text{では}, \text{へは}, \text{までは}, \text{よりは}) \qquad (1)
\end{aligned}
$$

を使用する．

4　誤り傾向を反映した格助詞訂正手法

本稿では，言語モデルだけを使用した誤り訂正をベースラインとして実験し，Noisy channel model を用いて学習者の誤りの傾向を反映した場合の結果と比較した．また，誤りモデルの重みを調節することによる正解率の変化も調べた．

Noisy channel model を用いた手法について説明する．w_E を学習者の書いた単語，w_C を訂正された単語とすると，$P(w_C|w_E)$ は学習者の書いた単語がどの単語に直されたか，$P(w_E|w_C)$ はある単語をどの単語と間違えたか，$P(w_C)$ は正しい文中での単語の登場割合を，$P(w_E)$ は学習者の文中での単語の登場割合を表す確率である．本タスクでは w_E は kakujoshiList（式 1）の要素であるとする．訂正誤りを最小化するためには，事後確率 $P(w_C|w_E)$ を最大化する単語 $\hat{w}_C$ を求めればよい．

$$\hat{w}_C = \arg\max_{w_C} P(w_C|w_E) = \arg\max_{w_C} P(w_E|w_C)P(w_C)$$

この式では，求める確率を $P(w_E|w_C)$ と $P(w_C)$ のふたつの要素に分けて考えることができる．これらをそれぞれ誤りモデルと言語モデルと本稿で呼ぶ．これら誤りモデルと言語モデルの精度をそれぞれ改善することにより，誤り訂正の精度を改善することができる．

さらに，誤りモデルの重みを表すパラメータ λ $(0.0 \leq \lambda \leq \infty)$ を導入すると以下のような式となる．

$$\arg\max_{w_C}(P(w_E|w_C)^{\lambda}P(w_C))$$

λ を変化させて誤りモデルの影響を調節することにより，正解率が向上するかどうかを調べる実験も行った．適切な λ を求めるため開発データ[*4]を用いて予備実験を行った結果を図 2 に示す．

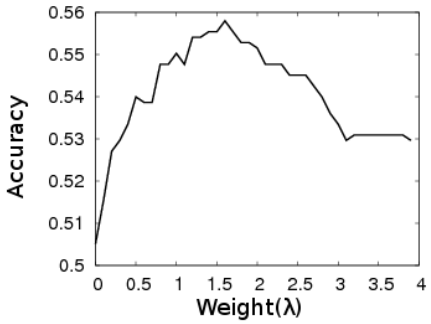

図 2　予備実験における Noisy Channel Model での誤りモデルの重み λ に対する正解率

$\lambda = 0$ の時が言語モデルのみの場合である．誤りモデルの重みを増加させるにつれ正解率が向上し，$\lambda = 1.6$ のときに最高値に達し，その後は重みを増加させるにつれ低下している．この結果より，言語モデルのみだけで誤り訂正を行った場合よりも，誤りモデルを考慮して訂正した方が正解率が向上するが，適切な重みで組み合わせることが重要であり，誤りモデルを重視しすぎることも正解率の低下をもたらすことがわかる．この予備実験では $\lambda = 1.6$ としたときが最も正解率が高かったため，この値を用いて実験を行う．

5　日本語学習者の誤りモデルの構築

本研究における誤りモデルは，学習者がどの格助詞とどの格助詞をどのくらいの確率で間違えるのか，という情報を得るために用いる．誤りモデルの構築には言語学習 SNS のひとつである Lang-8 を独自にクロールして収

[*2] 「の」は通常，接続助詞に分類されるが訂正先の 10.89% を占めているためタスクの対象に含めた．

[*3] 「より」，「からは」，「へは」，「よりは」はテストデータに登場しなかったため，実験の評価はこれらの助詞を除いて行った．

[*4] タスクの対象となる要素数は 1,486 であったが，約半数の 700 要素をテストデータとして使用し，残りの 786 要素を開発データとした．

表 2　Lang-8 から抽出した置換対の頻度上位 10 個

学習者のトークン	添削後のトークン	頻度
は	が	25,239
が	は	18,991
す	した	17,906
に	で	13,446
が	を	12,959
、	。	12,748
を	が	11,740
を	に	10,831
した	す	8,105
で	に	8,303

集したデータを用いた．

収集したデータは学習者の作文と添削文が対になっており，添削文は不要な文字が削除されていたり，必要な文字が追加されたりしている．Lang-8 では一般的には挿入した文字は青色にするなど，編集をした箇所がわかるようにマークアップが施されているが，添削のスタイルは添削者によってばらつきがあり明示的に何が何に変わったという情報が示されていないので，単純な方法では添削箇所の対応関係を取ることができない（参照 [1]）．

そこで，動的計画法によるマッチングを用いて置換対の抽出を行った．用いたデータは日本語学習者の作文，約 130 万文であり，HTML のタグの情報は利用せずにすべて取り除いた．添削前の文と添削後の文をそれぞれ文頭から 1 文字ずつ読み進めていき，添削前の文には存在するが添削後の文には存在しない文字列を挿入箇所として，添削後の文には存在するが添削前の文には存在しない文字列を削除箇所としてみなす．今回は挿入箇所と削除箇所が連続した部分を置換のペアとして抽出した．

表 2 に，この手法で獲得できた置換対の上位 10 個を示す．こうして抽出したデータから以下の式に基づき誤りモデルで用いる確率を求めた．

$$P(w'_E|w_C) = \frac{C(w'_E, w_C)}{\sum_{w_E} C(w_E, w_C)}$$

ここで，w_E, w_C はそれぞれ学習者のトークン，添削後のトークンであり，ともに kakujoshiList（式 1）に含まれる要素のみを使用した．

6　格助詞誤り訂正実験

6.1　使用データ

6.1.1　言語モデル

言語モデルには Web 日本語 N グラム第 1 版を使用した．このデータは 200 億文より得られた 2,550 億トークンを元に構築されており，総異なりトークン数は約 256 万トークンである．言語モデルで，それぞれの格助詞が現れる割合を調べた（図 3）．

Web 日本語 N グラムに格納されているデータは単語のユニグラムから 7 グラムの頻度であり，本稿では格助詞を含む単語 1-3 グラムを使用した．我々はこの頻度を用いて以下の式で表される確率を求めた．

$$P(w_1, w'_2, w_3) = \frac{C(w_1, w'_2, w_3)}{\sum_{w_2} C(w_1, w_2, w_3)}$$

我々が行った最も単純な手法のベースラインではこの確率のみを用いて訂正を行う．ただし w'_2 は kakujoshiList

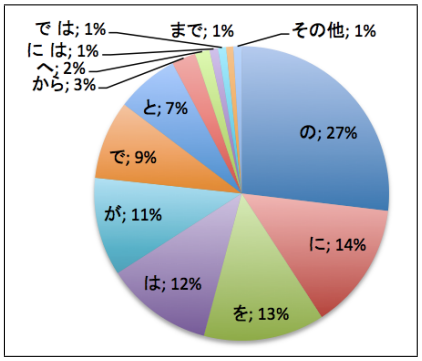

図 3　言語モデルにおける格助詞の割合

表 3　NAIST 誤用コーパスにおける助詞誤りの分類

	挿入	削除	置換
すべての助詞誤りタグ	368 (11%)	920 (27%)	2,093 (62%)
学習者の入力が kakujoshiList の要素	328 (10%)	920 (27%)	1,787 (53%)
学習者入力と訂正先が kakujoshiList の要素	328 (10%)	812 (24%)	1,485 (44%)

（式 1）の要素である．すべての格助詞に対してトライグラムの頻度が 0 だった場合，格助詞前後のバイグラム確率の積 $P(w_1, w'_2)P(w'_2, w_3)$ で推定し，バイグラムの確率でも最尤推定できなかった場合，ユニグラムの確率 $P(w'_2)$ で最尤指定を行った．

6.1.2　テストデータ

テストデータには NAIST 誤用コーパスを利用した．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（以下，「作文対訳 DB」）に誤用タグを付加したものである．作文対訳 DB の第一版はアジア 10 ヶ国の日本語学習者の作文が集められている．NAIST 誤用コーパスはこの第一版（アジア学生編）の作文の中で添削[*5]が施されている 313 編について誤用タグが付与されている．

本稿では NAIST 誤用コーパスから，訂正先の情報まで付与されている 6,685 文を取り出し，助詞誤りタグのついている箇所のうち，学習者の入力が kakujoshiList の要素である置換誤りについてテストを行った． 表 3 では，NAIST 誤用コーパス内で助詞誤りタグのついているものに対して，挿入・削除・置換のそれぞれの誤りの割合を示している．表から見て取れるように，置換の誤りがおよそ半分を占めている．また，全置換誤りのうち 70% は学習者の入力も訂正先も格助詞であり，提案手法によって訂正される可能性がある．

N グラム頻度を調べる際に入力文が形態素に分割されている必要があるため，前処理として IPADic-2.7.0[*6]で学習させた MeCab-0.98[*7]を用いて形態素解析を行った．[*8]また，助詞以外の誤りは 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た．

[*5] 添削の仕様は日本語教師によって定められた．
[*6] http://sourceforge.jp/projects/ipadic/
[*7] http://mecab.sourceforge.net/
[*8] 誤りを含む文に対する単語分割の精度は保証されていないが，本章の主目的は Noisy channel model を利用したときに精度がどれだけ向上するかを調べることなので，必ずしも文は正しく単語に分割されているという必要はないことに注意してほしい．

表 4　格助詞訂正の実験結果

手法	正解率	マイクロ平均	マクロ平均
言語モデル（ベースライン）	48.0%	58.3%	31.8%
Noisy Channel Model（提案手法 1）	54.4%	66.1%	31.8%
重み付き Noisy Channel Model（提案手法 2）	55.6%	67.5%	32.1%

表 5　助詞ごとの正解率の比較

	言語モデル		Noisy Channel Model		重み付き Noisy Channel Model		
	正解数	正解率	正解数	正解率	正解数	正解率	総数
が	75	60%	95	75%	99	**79%**	126
を	64	70%	65	**71%**	63	68%	92
の	38	**76%**	30	60%	28	56%	50
に	70	65%	82	76%	84	**78%**	108
から	0	0%	0	0%	0	0%	9
と	5	**63%**	3	38%	3	38%	8
で	33	52%	35	**55%**	35	**55%**	64
へ	0	0%	0	0%	0	0%	2
まで	0	0%	0	0%	0	0%	1
は	50	50%	71	71%	77	**77%**	100
には	0	0%	0	0%	0	0%	5
とは	0	0%	0	0%	0	0%	1
では	1	**11%**	0	0%	0	0%	9
までは	0	0%	0	0%	0	0%	1

6.2　評価尺度

評価はタグに記載されている添削とシステムの出力を比べたときの正解率である．また，全体としての評価と，総数が少ない助詞も重視した評価の両方を行うためにマイクロ平均とマクロ平均での評価も行った．本実験のベースラインとして，誤りモデルを使用せず言語モデルだけを使用して誤り訂正をする実験を行った．

6.3　実験結果

表 4 に示すように，Noisy channel model ではベースラインよりも 6.4% 正解率が向上した．また，誤りモデルの重みを調節することによりさらに正解率の改善が確認できた．この結果より，学習者の誤りの傾向を反映することで，訂正精度が改善されることが示された．一方でマクロ平均では大きな差はみられなかった．これは，総数の少ない助詞も多い助詞と同様に重要だと考えると大きな改善はないことを示している．

表 5 に，各モデルの傾向をより詳しく比較するため，助詞ごとの正解数，正解率，総数を示す．言語モデルと比べて Noisy channel model を用いた場合，「が」「を」「は」などの総数の多い格助詞の正解率が改善している一方で，「の」「と」「で」「では」などの比較的総数の少ない格助詞では，言語モデルよりも正解率が下がっている．

7　議論

表 6 に「と」と「の」について，ベースラインでは正解できていたが提案手法により正解できなくなっていたものの例文を示す．考えられる理由としては，これらの助詞は体言にかかることが多く，他の助詞が主に用言にかかることと比較すると性質が違うことがあげられる．対象としている助詞が用言を修飾しているのか体言を修飾しているのかは，助詞前後の情報だけでなくより広い文脈を見なければ判断できない．そこで，係り受けなどの情報を用いることが有効だと考えている．またその際

表 6　提案手法により正解できなくなった文の例（下線部は誤り箇所を示す．）

学習者の文	正解	システム出力
これは人々がたばこの理解していること を 表明だと思います。	の	が
アイヂルビトリ に 前に一ヶ月間に 断食をする。	の	で
たばこを一種の嗜好品 に 考えられるように 努力しなければならなく	と	を
男性 を 女性はお互いの物を 買いそろえます 。	と	が

に，「の」や「が」以外の助詞では，一般に 1 つの述語に同じ種類の助詞が複数係るのは不自然である．そこで，1 つの述語に係る同じ助詞は 1 つのみといった制約を加えることで，文全体をみた訂正が可能になると考える．

マクロ平均では大きな改善がみられなかったが，学習者にとっては，少数の格助詞を正しく使いたいという事も想像できるので，マイクロ平均のみでなく，マクロ平均もともに改善できる手法の研究に取り組む必要がある．

今回言語モデルには Web 日本語 N グラムを用いたが，Web から集めてきたコーパスと日本語学習者の作文ではドメインにずれがあるため N グラムの分布に違いがある可能性がある．現代日本語書き言葉均衡コーパス[*9]は書籍や雑誌，新聞，ブログなど多様な日本語から構築されており，Web 日本語 N グラムよりも N グラム分布の偏りが少なく，学習者の作文に近いと考えられる．このようなコーパスを組み合わせて言語モデルを構築することにより正解率の改善が期待できる．また，機械学習を適用することで，関連研究で用いられていた，品詞や係り受けなどのよりリッチな情報を組み合わせる事ができる．誤りモデルに関しては，今回ユニグラム情報しか用いなかったが，バイグラムやトライグラムを用いればより広い文脈の情報を反映した訂正が行える．

今回は教師データの量が十分ではなかったため母語による分類を行わなかったが，学習者の母語によって誤りの傾向が異なると考えられるので，母語別の訂正モデルを構築すれば正解率が改善されるのではと考えている．

参考文献

[1] Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of IJCNLP*, 2011.

[2] Hiromi Oyama. Automatic Error Detection Method for Japanese Particles. *ICTATLL*, pp. 235–245, 2010.

[3] Alla Rozovskaya and Dan Roth. Generating Confusion Sets for Context-sensitive Error Correction. In *Proceedings of EMNLP*, pp. 961–970, 2010.

[4] Hisami Suzuki and Kristina Toutanova. Learning to Predict Case Markers in Japanese. In *Proceedings of ACL*, pp. 1049–1056, 2006.

[5] 今枝恒治, 河合敦夫, 石川裕司, 永田亮, 桝井文人. 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.

[6] 大木環美, 大山浩美, 北内啓, 末永高志, 松本裕治. 非日本語母国話者の作成するシステム開発文書を対象とした助詞の誤用判定. 言語処理学会第 17 回全国大会論文集, pp. 1047–1050, 2011.

[7] 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りの自動検出・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.

[*9] http://www.tokuteicorpus.jp/