社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－181 (17)
2007／9／26

文節内の特徴を用いた日本語助詞誤りの自動検出・校正

南保 亮太† 乙武 北斗† 荒木 健治†

†北海道大学大学院情報科学研究科

本稿では日本語学習者が誤りやすい助詞の誤りを，文節の特徴から帰納的学習を用いて検出および自動校正する手法を提案する．助詞誤りを検出する従来手法として，格フレームを用いた手法などがあったが，単文にしか対応できない点，格助詞以外の助詞に対応できない点など，制約が多いという問題点があった．そこで，提案手法では，これらの制約を軽減したうえで，構文解析によって得られた文節内の特徴から特徴スロットを抽出し，これと助詞を組み合わせることでルールとする．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・校正を行う．性能評価実験の結果，格助詞以外の助詞に対しても誤り検出・校正ができることが示された．

Automatic Error Detection and Correction of Japanese Particles Using Features within Bunsetsu

Ryota NAMPO† Hokuto OTOTAKE† Kenji ARAKI†

†Graduate School of Information Science and Technology Hokkaido University

In this paper, we propose an error detection and correction method of Japanese particles which are mistakable for learners of Japanese, using Inductive Learning from features within *Bunsetsu*. Before now, there also were detection and correction methods. However, they were so constrained. Therefore we propose the method which reduce this constraint. In our method, we extract features within *Bunsetsu* from the result of parsing, and combine them with a particle. And we get rules for choice of particles by Inductive Learning. Through the evaluation experiment, we confirmed that errors in particles except for case particles are able to be detected and corrected by our proposed method.

1. はじめに

日本語学習者にとって，助詞の習得とは努力を要するものであり，日本語の中でも最も難しい文法項目の一つであると言われている [1][2]．助詞の習得が難しい理由としては，次の 3 点が挙げられる．

- 助詞単体ではほとんど意味を持たず，他の品詞につくことでその前後関係から意味をなす
- 一つの助詞が複数の意味を持つ
- 類似した機能を持つ助詞が複数存在する

以上の理由に加えて，母国語の影響などから，日本語学習者は助詞を書き誤ることが多い．このような場合，自力で誤りを発見し，訂正することは労力を要する．

そこで，こうした現状を解決するために，日本語の助詞誤り検出を自動化する手法 [3][4][5] が提案されている．

今枝らの手法 [3] では，人手によって書かれたルールによる判定に加え，格フレームの照合を用いる．入力文の格フレームと辞書から得られた格フレームを比較することによって誤りの検出および校正を行う．しかし，格フレームを用いて扱うことができるのは格助詞のみである．また，入力は単文しか扱わない点，誤りは一箇所だけとする点，などの制約がある．

また，Suzuki らの手法 [4] では，文節内の特徴から素性を作成し，最大エントロピーモデルを用いた機械学習によって予測を行う．扱う助詞の対象は，10 の格助詞と「は」に加えて，「には，からは，とは，では，へは，までは，よりは」のような格助詞と「は」から成る 7 つの組である．各文節を 18 の格助詞と格助詞なしの 19 クラスに分類する．しかしこの手法は，機械翻訳における日本語文生成の準備段階と位置づけられており，誤りの検出・校正は対象としていない．

新納らの手法 [5] では，平仮名 N-gram を利用することで，助詞誤りに限ることなく，誤り検出および校正を行う．しかし，助詞の利用という観点から見ると，漢字やカタカナなど，平仮名以外の文字列に挟まれた場合には，対応することができない．

これら従来研究の問題点を踏まえ，本稿では日本語助詞誤りの検出と校正について更なる性能向上を目指し，文節内の特徴を用い，帰納的学習 [6] を行うことで日本語助詞誤りの検出・校正を行うシステムを提案する．本手法では，電子化コーパス中の日本語文における文節内の特徴を抽出し，助詞と組み合わせてルールとする．本稿における特徴とは，対象の文節や係り先の文節内の語や品詞情報，近傍の助詞などを要素として持つ特徴スロットのことを指す．特徴スロットの詳細については **3.1** で述べる．このような処理によって，文内の文脈を考慮したルールを獲得することができる．また，帰納的学習を用いて，抽出されたルール同士から抽象化したルールを

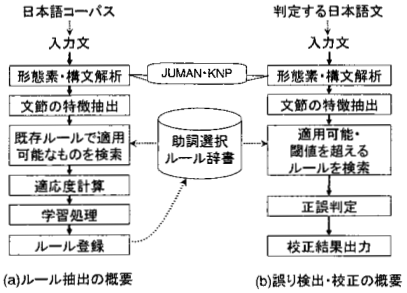

図 1 システム概要図

新たに自動生成する．次に獲得されたルールに基づいて助詞誤りの検出・校正を行う．

本手法の利点として，ルールを人手で作成する労力を必要としない点が挙げられる．また，ルールの抽象化を行うことで，助詞選択に関わる文脈要素を絞り込むことができる．更に，特徴スロットを用いた手法は扱う助詞の種類に関して制約を受けないため，本稿では，格助詞に加え，接続助詞，副助詞，係助詞，並立助詞を扱う．

なお，本手法で用いる特徴スロットの帰納的学習を用いた文法誤り検出・校正については，これまで，英語の冠詞誤りを対象として研究を行ってきた [7]．同様の枠組みを日本語の助詞に対して適用することで，前述の手法の汎用性を検証する．

以下，**2.** で本手法の概要を，**3.** でシステムの処理過程について述べる．**4.** で評価実験について述べ，**5.** では実験結果とその考察，**6.** では従来手法との比較を行い，**7.** でまとめを述べる．

2. システム概要

本システムは処理内容から大きく 2 つの処理部に分けられる．一つはルール抽出部，もう一つは誤り検出校正部である．本システムの基本的な機構は英文冠詞誤りの検出および校正システム [7] と同じであるが，日本語の助詞に対応させるために特徴スロットの要素や，適応度計算式などに対して修正を加えた．

2.1 ルール抽出部

ルール抽出の概要は図 1(a) に示す．入力文は日本語コーパスから自動的に取り出される．その入力文各々に対して形態素解析および構文解析を行い，構文構造などの情報を獲得する．形態素解析ツールとして JUMAN[8]，構文解析ツールとして KNP[9] を用いた．次に，構文解析された結果から各文節ごとにその文節と周辺の特徴を抽出する．この特徴は，助詞の選択を決定する要素をカテゴリとして持つ特徴スロットとして抽出される．これについては **3.** で詳細に述べる．また，この特徴スロットと対象の文節につく助詞を組み合わせたものをルールとする．得られた特徴スロットに対して，既存のルール中に適用可能なルールがあるかどうかを検索する．適用できるルールが存在した場合，そのルールに対して適応度の更新を行う．適応度とは得られたルールの確からしさを現す数値である．適応度については **3.2** で述べる．最後に，学習処理によって抽象化した新しいルールを生成する．学習処理の詳細については **3.** で述べる．

対象文節	*対象語*	みかん
	品詞	名詞
	品詞細分類	普通名詞
	活用形	名詞-*
	数量表現	なし
	日付表現	なし
	句読点	なし
	文節位置	文中
	係り受け	P

係り先文節	*最初の語*	りんご
	品詞	名詞
	品詞細分類	普通名詞
	句読点	なし
	文節位置	文中
	態	なし
	否定	なし
	係り受け	D
近傍文節	*前文節の助詞*	は
	次文節の助詞	も

図 2 特徴抽出の例

2.2 誤り校正部

誤り校正の概要は図 1(b) に示す．入力文は助詞の誤りの校正を行う文章を含むものとする．入力文に対して学習時と同じく，JUMAN および KNP を用いて形態素解析および構文解析を行う．次に，ルール抽出時と同様に文節の特徴抽出を行う．抽出された各々の文節の特徴に対してルール抽出で獲得された助詞選択ルール辞書の中から適用できるルールを検索し，助詞の校正を行う．最後に結果を出力する．

3. 処理過程

本章では，**2.** の概要で述べた処理過程について詳細に述べる．

3.1 文節の特徴抽出

Suzuki らの研究 [4] では，最大エントロピー法を用いるための素性として，

- 格助詞の予測がなされている文節の素性
- その係り先の文節の素性
- その二つの文節間の関係に関する素性

の 3 種類の基本素性に加え，素性の組を 20 組使用している．これらの素性の抽出と評価は新聞記事データを用いて行われており，助詞の選択に有用な特徴であると考えられる．そこで本手法ではこれらの素性のうち，特に重要であると考えられるものを選択した．それに加え，隣接する文節の持つ助詞情報も利用した．以上から図 2 のように特徴スロットを定義する．図 2 の特徴スロットは以下の文，

(a) 私はみかんもりんごも好きです．

において，対象の文節を「みかんも」としたときに抽出される特徴スロットである．

図 2 に示すように，特徴スロットは三つのカテゴリー，1) 対象文節カテゴリー，2) 係り先文節カテゴリー，3) 近傍文節カテゴリーが存在する．

対象文節カテゴリーは，対象の文節に関する特徴を保持する．一番上のスロットにある対象語とは，助詞を除いて文節内に最後に来る語を指す．例文 (a) の場合，「みかんも」の助詞「も」を除いた「みかん」が対象語となる．その他の要素として，対象語の品詞や分類，活用形，数量・日付を表す表現や句読点の有無，文節の位置や係り受けの種類を保持する．

係り先文節カテゴリーは，構文解析結果で対象の文節が係る先の文節に関する特徴を保持する．一番上のスロットにある最初の語とは，文節内の最初に来る語を指す．例文 (a) の場合，りんごとなる．この他に，最初の語の品詞と分類，句読点の有無と文節位置，受身や使役を表す態，否定の有無，係り受けの種類を保持する．

近傍文節カテゴリーは，対象の文節から見て前後 1 文節ずつの助詞部分を保持する．例文 (a) の場合，前文節の助詞が「は」，次文節の助詞が「も」となる．対象の文節が文頭だった場合は前文節の助詞を，文末だった場合は次文節の助詞を，それぞれ「文節無し」とする．

一部のスロットについては空白化処理を行う．例えば，日付表現スロットは「なし」という値が入る割合が高いため，日付表現がないということは大きな特徴とはいえない．そのため，日付表現がある場合だけ「あり」という値を代入し，ない場合はスロットを空白にする．この処理を 7 つのスロットに対して行う．これによって無駄なルールが作成されるのを抑制し，効率的な学習が可能となる．

3.2 適応度の計算

日本語コーパスから新しくルールを抽出した際，既存のルールの中に適用可能なルールがあるかどうかを検査する．特徴スロットが一致した場合にルールを適用可能とする．ここで，適用回数と正適用回数を定義する．適用回数とは，新しく抽出されたルールの特徴スロットに対してルールが適用可能となった回数とする．正適用回数とは，適用回数のうち，助詞部分も一致した回数とする．適応度を式 (1) のように定義する．

$$\text{適応度} = \text{助詞頻度補正} \cdot \text{抽象度} \cdot \log_2(\text{正適用回数}) \cdot \text{正適用率} \tag{1}$$

ルールが正適用された場合に適応度が上がり，適用されたが助詞が異なる場合に適応度が下がることとする．そのため，まず式 (2) に示す正適用率を用いる．

$$\text{正適用率} = \frac{\text{正適用回数}}{\text{適用回数}} \tag{2}$$

また，正適用率に正適用回数の対数を乗じることによって，適用回数が少ないにも関わらず正適用率が高くなることを防ぐ．これに加え，抽象度と助詞頻度補正を用いた．抽象度は式 (3) に示す通り，ルール内の特徴スロットのうちどれだけの割合で要素に値が入っているかを表す数値である．これを乗算することによって，スロットに空白の多い抽象的なルールの正適用回数が増加しても余計な適用を防ぐことができる．また，式 (4) に示す助詞頻度補正は学習コーパス中での頻度が高い助詞だけが早く学習され，優先的に適用されることを防ぐためのものである．

$$\text{抽象度} = \frac{\text{値が入っている特徴スロットの要素数}}{\text{特徴スロットの全要素数}} \tag{3}$$

$$\text{助詞頻度補正} = \log \frac{\text{コーパス内の全ての助詞数}}{\text{コーパス内の対象の助詞頻度}} \tag{4}$$

3.3 学習処理

3.3.1 帰納的学習

本論文における帰納的学習とは，「実例からそこに内在している規則を獲得すること」と定義している [6]．本手法での実例とは学習コーパス内の文章から抽出される特徴スロットである．この特徴スロット同士を比較し，各要素について共通部分と差異部分を再帰的に抽出することにより，抽象化したルールを次々に生成していく．このような学習処理を進めることによって，助詞選択の決定に必要な文脈要素のみを持つルールを生成することを目指す．

3.3.2 学習処理過程

日本語コーパスから新しいルールを抽出した際に，既存のルールの特徴スロットと比較し，**3.3.1** で述べた帰納的学習によって新たなルールを生成する．2 つのルールが持つ特徴スロットの各要素について，内容が一致した要素を共通部分とし，それ以外を差異部分とする．新しく生成されるルールの特徴スロットには共通部分が要素として残り，差異部分は空白化することにより抽象化される．学習処理によって新たなルールが生成される例を図 3 に示す．図 3 に示す上段のルールは，それぞれ以下の文 (1)(2) から獲得されたルールである．

(1) 冬は東京にいますが，夏にはハワイに行きます．
(2) 鈴木さんは，フランスに観光で行っていらしたんですか？

この例では，二つのルールの特徴スロットの比較により，対象文節の品詞，細分類，活用と，係り先文節の最初の語，品詞，細分類，そして助詞が共通部分で，その他が差異部分となることが分かる．これら二つのルールの帰納的学習の結果，共通部分が残り，差異部分が抽象化された新たなルールが生成される．図の“＊”部分はワイルドカードで，任意の要素の代入を許す．

学習処理の対象となるルールの条件は，単に共通部分を持つルール同士である点だけではない．なぜ

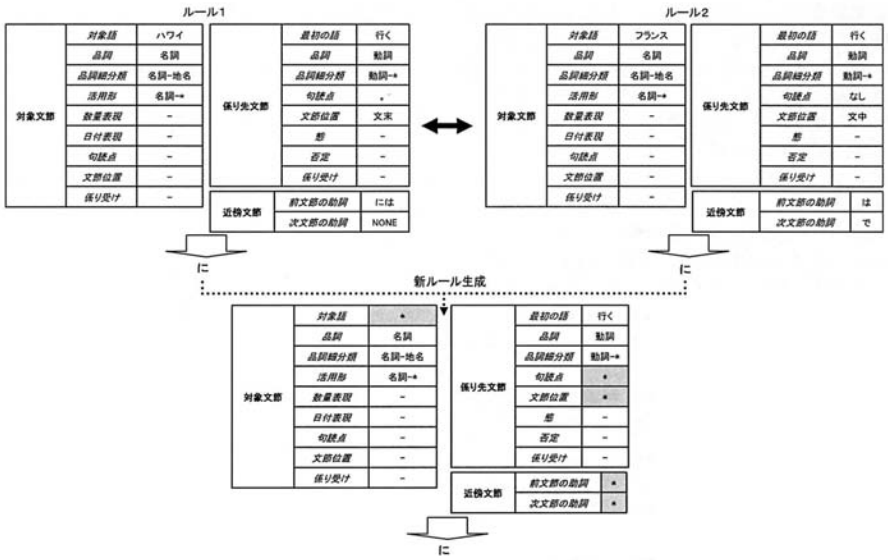

図 3 学習処理の例

なら，そうしてしまうと極度の抽象化によって，あらゆる特徴スロットに対して適用できるルールが作成されるためである．ゆえに，学習処理は何らかの制限のもとに行う必要がある．まず，学習対象となる二つのルールにおいて，助詞部分の一致は必須とした．また，対象文節カテゴリ・係り先文節カテゴリ内のスロットは最低一つ以上一致し，全体の 40%以上に要素が入っていることとした．これらの条件を満たすルールが存在する限り，学習処理は再帰的に行われる．

3.4 助詞選択ルール適用

誤り校正は，ルール抽出部において獲得されたルールを用いて行う．ここで，ある一定以上の適応度を持つルールだけを用いるためにしきい値 θ を設定し，それ以上の適応度を持つルールのみを検索し，用いることとする．

4. 性能評価実験

4.1 学習用日本語コーパス

学習用の日本語コーパスとしては，まず，日本語助詞について書かれた教科書 [1][2] 内の例文および章末問題から 1,441 文を用い，さらに 2000 年から 2001 年までの毎日新聞の新聞記事コーパスの中から，読者記事「みんな集合」の 829 記事 4,839 文を用い，計 6,280 文を学習データとした．

教科書の例文をコーパスとして利用したのは，多くの助詞を用いた例文が満遍なく掲載されているためである．また，新聞記事の中でも読者記事を選択したのは，常体（〜だ，〜である）と敬体（〜です，〜ます）がどちらも含まれており，日常的な話題が比較的多い記事となっているためである．

- 教科書巻末問題
  太田さんを迎え が 駅まで行ったわ.
- 格助詞誤り
  英語 を できる.

図 4 テストデータの例

これらのコーパスから獲得されたルール数は 131,426 個であった．そのうち，コーパスから直接獲得されたものは 32,696 個，学習処理によって生成されたものは 98,730 個となった．

4.2 実験対象

以下に示す 2 つの実験対象について実験を行った．どちらも，日本語学習者によって作られた助詞誤りを含む文章に対して実験を行った．

教科書巻末問題 教科書 [2] の巻末，General Quiz から，助詞選択問題 46 問を選び，3 名の日本語学習者が回答した文章をテストデータとした．設問部分の文節のみで，終助詞を除いた 142 箇所を対象とした．そのうち，誤りは 34 箇所存在した．

比較的単純な格助詞誤り 韓国の大学で日本語を受講し，それ以前に日本語学習経験のない韓国語母語話者への調査 [10] によって得られた誤り例 50 文と，その誤りを人手によって訂正し，

重複しているものを除いた 34 文の合計 84 文をテストデータとした．

4.3 実験手順

まず **2.1** で説明した手法に従って，助詞選択ルール辞書を作成した．次に，**2.2** で説明した方法を用いて，実験対象中の助詞誤りを検出し，校正を行った．しきい値 θ は 0.25 刻みで値を変化させて実験を行った．

4.4 評価方法

本手法の誤り検出を評価する尺度として，式 (5)(6) で定める Recall，Precision を用いる．

$$\text{Recall} = \frac{\text{正しく検出できた誤りの数}}{\text{助詞誤りの数}} \quad (5)$$

$$\text{Precision} = \frac{\text{正しく検出できた誤りの数}}{\text{システムが誤りと判断した数}} \quad (6)$$

また，誤り校正を評価する尺度として，式 (7)(8) で定める Recall，Precision を用いる．

$$\text{Recall} = \frac{\text{正しく校正できた誤りの数}}{\text{助詞誤りの数}} \quad (7)$$

$$\text{Precision} = \frac{\text{正しく校正できた誤りの数}}{\text{システムが誤りと判断した数}} \quad (8)$$

更に，Recall と Precision の両方を考慮して性能を評価するために，文書検索の分野などでシステム評価に一般的に用いられる F-measure を用いる．

$$\text{F - measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

5. 実験結果と考察

5.1 教科書巻末問題

図 5 に本手法の誤り検出・校正における Recall の結果，図 6 に Precision の結果を示す．横軸はしきい値 θ を示している．しきい値を 0 から上げていくにつれ，適用可能なルールが少なくなり，Recall の値が上がっていく．適用可能なルールが減りすぎると，誤り検出が行えなくなるために，Recall が下がり始める．また，Precision の値は，しきい値を 0 から上げていくと誤検出が発生するために下がり始めるが，しきい値をある程度まで上げ続けると，適応度が高く信頼できるルールだけが適用されるようになるため，再び Precision の値が上昇する．

5.2 比較的単純な格助詞誤り

5.1 と同様，図 7 に誤り検出・校正における Recall の結果，図 8 に Precision の結果を示す．グラフの軌道は教科書での結果と同様であるが，Recall，Precision ともに高い値を示した．θ =5.0 付近で Recall が急激に下がっているのは，テストデータ中の文体が画一的であるために，一度にほとんどの文節に対してルールの適用ができなくなるためである．

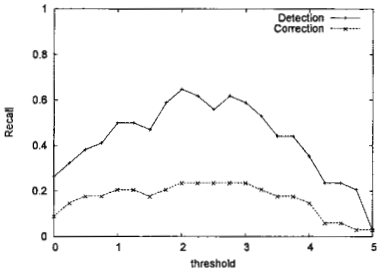

図 5 Recall 値の結果（教科書）

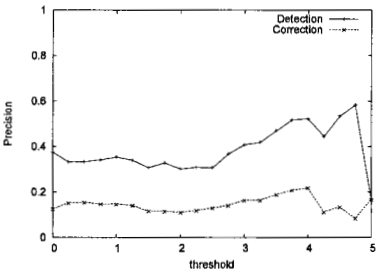

図 6 Precision 値の結果（教科書）

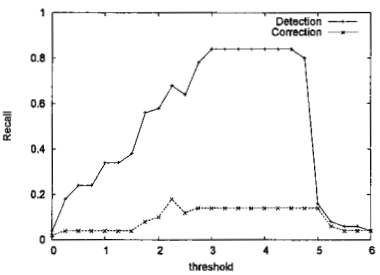

図 7 Recall 値の結果（格助詞）

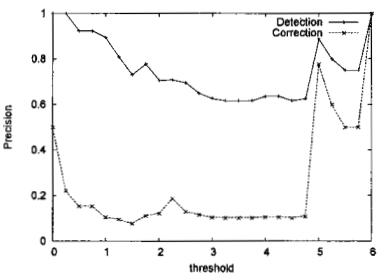

図 8 Precision 値の結果（格助詞）

表 1 関連手法との比較（検出）

	Recall	Precision	F-measure
教科書, $\theta = 2.0$	**0.65**	0.30	0.41
教科書, $\theta = 4.0$	0.35	0.52	0.42
格助詞, $\theta = 4.0$	**0.84**	0.64	**0.72**
格助詞, $\theta = 5.0$	0.16	**0.89**	0.27
今枝らの手法	0.59	0.89	0.71

表 2 関連手法との比較（校正）

	Recall	Precision	F-measure
教科書, $\theta = 2.0$	0.24	0.11	0.15
教科書, $\theta = 4.0$	0.15	0.22	0.18
格助詞, $\theta = 4.0$	0.14	0.11	0.12
格助詞, $\theta = 5.0$	0.14	**0.78**	0.24
今枝らの手法	0.49	0.33	0.39

5.3 考察

判定の失敗要因としては次のような事項が挙げられる．まず,「にも」,「なんかを」のようなコーパス中に現れる頻度が低い格助詞の組み合わせを誤りと判定することが多かった．また，手段や場所を表す「で」などは，動詞に対して必ずしも現れない任意格であるために，誤りと判定してしまうことがあった．また，助詞の直前に普通名詞が来た場合，適用されるルールに語そのものの情報がなく，名詞という品詞情報だけであったり，普通名詞という分類にしか値が入っていないという場合がある．こうした場合，ルールが抽象的であるために，誤検出が生じたり，誤った助詞を正しいと判定することが多かった．普通名詞の分類をもう少し細かく行うことによってこのような誤判定を減らすことができると考えられる．また，一部の助詞に関しては，判定部分が形態素解析で助詞と判断されないために校正が不可能となる場合も存在した．

なお，格助詞以外の助詞に対する性能を検証するため，対象の助詞を格助詞と格助詞以外に分け，しきい値 $\theta = 3.0$ でそれぞれ実験したところ，検出 F-measure はそれぞれ 0.43 と 0.38，校正 F-measure は 0.10 と 0.12 となり，それほど大きな差は生じなかったことから，格助詞以外の助詞に対してもこの手法が有効であることが示された．

6. 関連する手法との比較

同じ日本語助詞誤りの検出・校正システムとしては，今枝らの手法がある．文献 [3] での結果を基に同じ評価方法で比較した結果を，表 1，表 2 に示す．

今枝らの手法で扱われているのは格助詞のみである．また，用いられた実験データも，受身・使役・敬語表現を含まず，重文・複文も除外しているものであることから，条件的には，単純な格助詞誤りによる実験と同様であるといえる．誤り検出に関しては，ほぼ同等の性能であることが示された．

教科書での実験では，結果として低いものとなったが，格助詞以外の助詞を扱っているという点では本手法に優位性があるといえる．

また，しきい値をコントロールすることによって，Precision 重視，または Recall 重視というように，システムを調整することができるのは，本手法の利点といえる．

7. まとめ

本稿では，文節内の特徴を用いた日本語助詞誤りの検出および自動校正手法を提案した．実験の結果，格助詞以外の助詞に対しても誤りの検出・校正が行えることが示された．また F-measure による比較では，単純な助詞誤りによる評価実験で今枝らの手法とほぼ同等の結果を得ることができた．今後の課題としては Precision の向上が挙げられる．そのためには，まず頻度の低い助詞であっても正しく扱うために学習規模を大きくして実験する必要がある．また，特徴スロットの見直しを行うことによってさらなる性能の向上が見込めると考えられる．

参考文献

[1] 茅野直子．“助詞で変わるあなたの日本語 - All About Particles”，講談社インターナショナル，(2001)．

[2] 茅野直子．“比べて分かる日本語の助詞 - How to Tell the Difference between Japanese Particles”，講談社インターナショナル，(2005)．

[3] 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,2003-CE-68,pp.39-46, (2003)．

[4] Hisami Suzuki,Kristina Toutanova．“機械学習による日本語格助詞の予測”，言語処理学会第 12 回年次大会,pp.1119-1122, (2006)．

[5] 新納浩幸．“平仮名 N-gram による平仮名列の誤り検出とその修正”，情報処理学会論文誌 Vol.40, No.6 pp.2690-2698, (1999)．

[6] 荒木健治．“自然言語処理ことはじめ−言葉を覚え会話のできるコンピュータ”，森北出版，(2004)．

[7] 乙武北斗，荒木健治．“単語出現状況の帰納的学習による英文冠詞誤りの検出及び自動校正手法”，電子情報通信学会論文誌 D Vol.J90-D No.6 pp.1592-1601, (2007)．

[8] 黒橋禎夫，河原大輔．“日本語形態素解析システム JUMAN version 5.1 使用説明書”，(2005)．

[9] 黒橋禎夫，河原大輔．“日本語構文解析システム KNP version 2.0 使用説明書”，(2005)．

[10] 森山新．“日本語の格助詞習得はどのようになされるか - 韓国語母語話者に対する実験的研究”，東アジア日本語教育シンポジウム 論文集 pp.38-52, (2002)．