論　文

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

乙武　北斗[†a)]　　荒木　健治[†]

Correcting and Detecting Article Errors Method in English Using Inductive Learning from Conditions of Word Appearance

Hokuto OTOTAKE[†a)] and Kenji ARAKI[†]

あらまし　本論文では，日本人英語学習者が起こしやすい誤りの一つである英語の冠詞誤りを，単語出現状況から帰納的学習を用いて検出及び自動校正する手法を提案する．冠詞誤りを検出する従来手法として，電子化コーパスから獲得された統計量に基づくものがある．しかし，冠詞の用法には例外が多く，既存の手法では複雑な文脈を考慮するのは困難である．そこで，提案手法ではコーパス中の名詞及びその周辺の単語の出現状況をルールとして獲得することにより，文脈を考慮した冠詞誤りの検出，更に校正を行う．性能評価実験の結果，本手法の性能（F-measure=0.41）は関連手法の性能（F-measure=0.35）を上回り，優位性があることを確認した．

キーワード　冠詞誤り，帰納的学習，コーパス，自動校正，日本人英語学習者

1. まえがき

日本人英語学習者が起こしやすい誤りに，冠詞の誤用が挙げられる [1]〜[3]．文献 [4] では，冠詞そのものをもたない母語話者は，英語の冠詞を誤用する傾向があると報告している．

文献 [5] では，実験を行った結果，日本人の書いた英語には冠詞誤りが多いことが報告されている．また，文献 [6] では，冠詞誤りの中でも冠詞の脱落誤りが多いことが報告されている．そのため，英文を添削する際は多くの冠詞誤りを修正する必要がある．冠詞の用法には厳密な規則がない場合が多いため，辞書や用例から多くの事柄を調べる必要がある．このことから，冠詞誤りの添削には時間と労力，更に専門知識も必要となる．

こうした現状を解決するために，冠詞誤りの検出を自動化する手法 [1], [5], [7], [8] が提案されている．

永田らの手法 [1] では，英字新聞などの電子化コーパスから統計量を抽出して，それに基づいて冠詞誤りを検出している．また，永田らの手法 [7] では統計量を用いて名詞の可算・不可算を判定し，その結果を用いて冠詞誤りを検出している．和泉らの手法 [8] ではエラータグ付きの日本人英語学習者によるスピーキングコーパス [9] から統計量を抽出し，それに基づいて冠詞誤りを含む英文の誤りを検出している．このような統計量を用いた手法では，誤り検出用の辞書やルールを人手で作成する労力を必要としない利点がある．しかし，冠詞や名詞の統計量を用いていることから，複雑な文脈を考慮に入れることが困難である．また，冠詞の例外的な用法には対応できない．

河合らの手法 [5] では，構文解析などを用いて英文を解析し，人手で作成したルールに基づいて冠詞誤りを検出している．しかしながら，誤り検出に用いる辞書やルールを作成する際に多くの労力と専門知識を要する．また，すべての冠詞の用法を網羅したルールを作成することは非常に困難である．

以上の手法に加えて，欠落した冠詞を自動復元する手法 [4] が提案されている．この手法では，冠詞と名詞句の特徴との組合せの統計量を用いて，冠詞の復元を行う．ただ，欠落した冠詞の復元のみに対応しているため，例えば a を the と誤用したものなどについて

† 北海道大学大学院情報科学研究科，札幌市
Graduate School of Information Science and Technology, Hokkaido University, Kita 14 Nishi 9, Kita-ku, Sapporo-shi, 060–0814 Japan
a) E-mail: hokuto@media.eng.hokudai.ac.jp

は対応できない．

そこで，これら従来研究 [1], [4], [5], [7], [8] の問題点を解決するために，本論文では単語出現状況の特徴を用いた英文冠詞誤りの検出及び自動校正手法を提案する．本手法では，電子化コーパス中の英文における名詞句とその周辺の単語を特徴として抽出し，冠詞と組み合わせて冠詞選択ルールとする．本論文における特徴とは，対象名詞の単語や属する句の情報，修飾語句の単語や品詞を要素としてもつ特徴スロットのことを指す．特徴スロットの詳細は **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 [10], [11] 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新たに自動生成する．次に獲得されたルールに基づいて，冠詞誤りの検出・校正を行う．本手法の利点として，まず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，冠詞選択にかかわる文脈要素を絞り込むことができる．

以下，**2.** で本手法の概要を，**3.** ではシステムの処理過程について述べる．また，**4.** では評価実験を行い，**5.** で実験結果，及びその考察をする．**6.** で従来手法との比較を行う．最後に **7.** でまとめを述べる．

2. 概　　要

本システムは処理内容から大きく二つの処理部に分けられる．一つはルール抽出部，もう一つは誤り校正部である．

2.1 ルール抽出部

ルール抽出の概要を図 1 (a) に示す．

入力文は英文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構文解析を行い，構文構造を獲得する．構文解析ツールとして Apple Pie Parser [12] を用いた．

次に，構文解析された結果から名詞とその周辺の特徴を抽出する．この特徴は，冠詞の選択を決定する要素をカテゴリーとしてもつ特徴スロットとして抽出される．これについては **3.** で詳細を述べる．また，この特徴スロットと対象名詞に付属する冠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表す数値である．適応度のついては **3.2** で述べる．

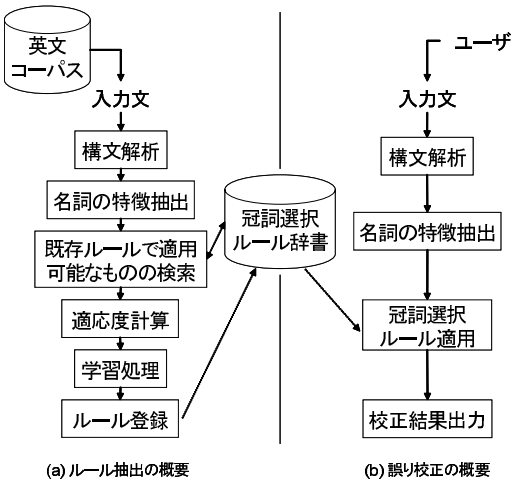

図 1　処 理 過 程
Fig. 1　Process flow.

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．

2.2 誤り校正部

誤り校正の概要を図 1 (b) に示す．

入力文は冠詞の誤りの校正を行う文章を含むものとする．入力文は構文解析されるが，用いる構文解析ツールはルール抽出時と同じく Apple Pie Parser とした．次に，ルール抽出時と同様に名詞の特徴抽出を行う．抽出された各々の名詞の特徴に対してルール抽出部で獲得された冠詞選択ルール辞書，及び人手で作成した明確な冠詞の用法を記述した解析的ルール辞書の中から適用できるルールを検索し，冠詞の校正を行う．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 名詞の特徴抽出

文献 [2], [3], [13], [14] によれば，冠詞の選択を決める重要な要素として以下のものが挙げられている．

- 対象名詞が可算名詞か不可算名詞か
- 対象名詞を修飾する修飾語句は何か
- 対象名詞が存在する文より前の文に出現する名詞との関連性
- 対象名詞が属する句は何か
- 対象名詞が受ける動詞は何か

- 対象名詞は繰り返し出現しているか

新聞記事や説明書，論文など，文脈が比較的明確な文章に関しては，以上の要素を考慮に入れたルールを適用することによって冠詞誤りを検出・校正することができると考えられる．

そこで，これら要素のうち文中文脈に関するもののみを考慮に入れたルールを構築するために，図 2 に示すような特徴スロット(注1)を定義する．図 2 の特徴スロットは以下の英文

(a) This is the only *book* which I bought yesterday.

において，対象名詞を *book* とした特徴スロットである．上の例文 (a) から抽出されるルールは，図 2 の特徴スロットと定冠詞 the の組合せとなる．

図 2 に示すように，特徴スロットには三つのカテゴリー，1) 対象カテゴリー，2) 前置修飾カテゴリー，3) 後置修飾カテゴリーが存在する．対象カテゴリーは，対象名詞に関する文法要素を保持する．例文 (a) の場合，対象名詞は *book* であるので “名詞”，“主名詞” 要素に *book* が入る．対象名詞が *tennis player* のように複数語で構成される場合だと，“名詞” 要素にそれら複数語が入り，“主名詞” 要素に複数語の中で最後尾の語である *player* が入る．その他の要素として，対象名詞が属する句，前置詞句に属する場合の前置詞，対象名詞の単複，対象名詞が固有名詞かどうかといったものがある．“名詞目的語動詞” 要素は，対象名詞を目的語とする動詞の原形が入る．例文 (a) の場合，*book* は *is* の目的語であると判断し，その原形である *be* が入る．また，“名詞主語動詞” 要素には対象名詞を主語とする動詞が入る．

前置修飾カテゴリーは，対象名詞の前置修飾語句に関する情報をもつ．例文 (a) の場合，対象名詞 *book* は *only* によって修飾を受けているので，*only* とその品詞情報が入る．この “品詞” 要素に入る品詞記号を表 1 に示す．

後置修飾カテゴリーは，対象名詞の後置修飾語句が存在した場合に，その情報を保持するカテゴリーである．後置修飾カテゴリーは 3 種類の後置修飾（前置詞句，不定詞句，関係詞節）によって更に細分化される．例文 (a) の場合，対象名詞 *book* は *which* 以下の関係詞節によって修飾を受けていると判断されるため，図 2 に示すように後置修飾カテゴリーの関係詞節のみ要素が入る．要素の内容としては，関係詞節内部の主語，動詞，その目的語，目的語に付属する冠詞，副詞

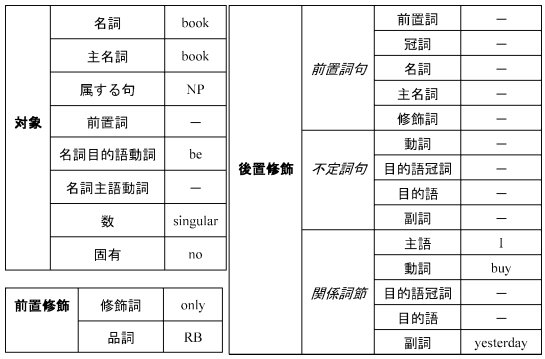

図 2　特徴スロットの例
Fig. 2　An example of a feature slot.

表 1　前置修飾カテゴリーで用いる品詞記号
Table 1　POS tags for the premodifier category.

記号	品詞
JJ	形容詞
JJR	形容詞　比較級
JJS	形容詞　最上級
RB	副詞
RBR	副詞　比較級
RBS	副詞　最上級
CD	基数

がある．

本手法では，一つの対象名詞から特徴抽出する際に

- 対象カテゴリーのみ
- 対象と前置修飾カテゴリー
- すべてのカテゴリー

のカテゴリーをもった三つの特徴スロットを抽出する．こうすることで異なる特徴範囲をもったルールを複数抽出し，ルール抽出の効率化を目指す．

3.2　適応度の計算

英文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索する．ルールが適用可能となる条件は，対象とルールの特徴スロットの一致とする．

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，冠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．

$$\text{適応度} = \log(\text{正適用回数})\frac{\text{正適用回数}}{\text{適用回数}} \qquad (1)$$

(注1)：図 2 で特徴スロットの要素における “—” は，該当する要素が存在しないことを示す．

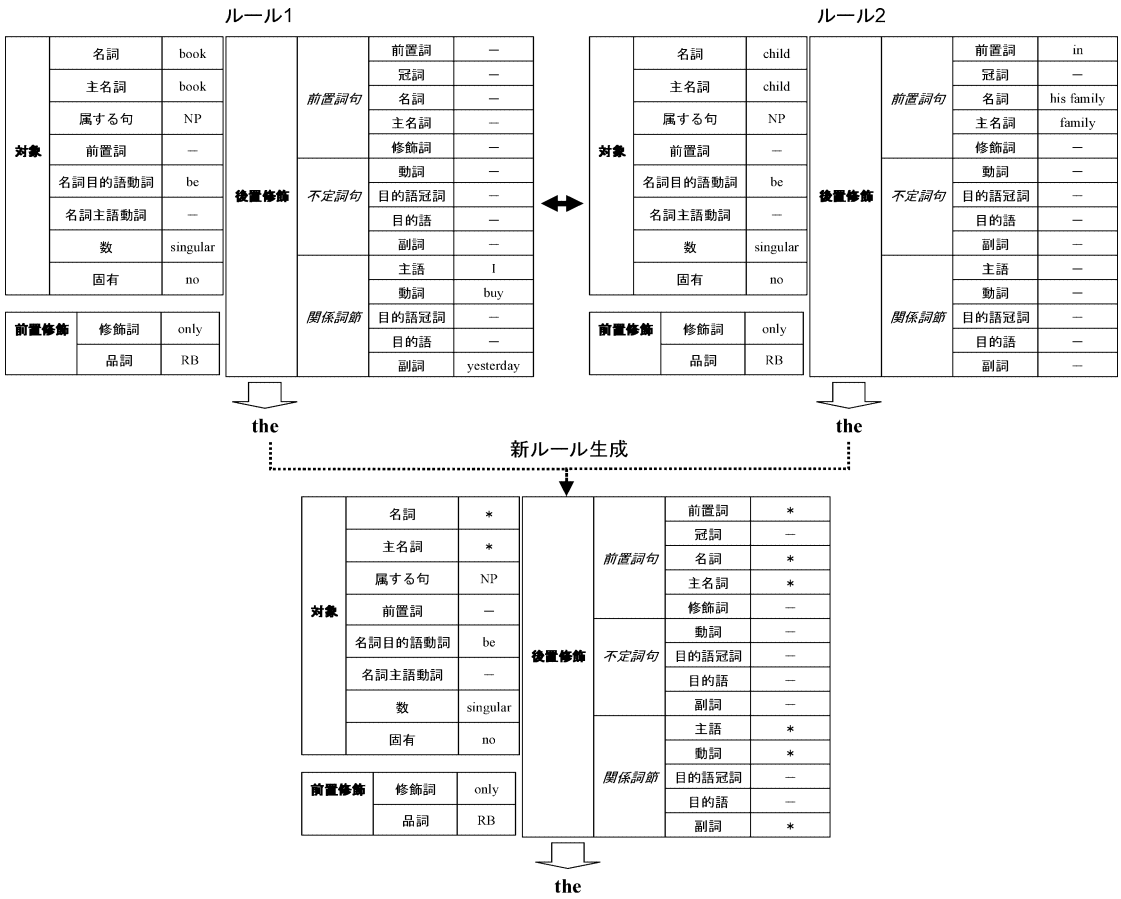

図 3　学習処理の例
Fig. 3　The learning process.

正適用回数の対数を掛けることによって，適用回数が少ないにもかかわらず正適用率が高くなることを防いでいる．

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 [11]. 本手法での実例とは学習用コーパスから抽出される英文に含ま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，冠詞選択の決定に必要な文脈要素のみをもつルールを生成することを目指す.

3.3.2 学習処理過程

英文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．二つのルールがも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，差異部分は変数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 に示す上段の二つのルール（ルール 1，ルール 2）は，それぞれ以下の英文（1），(2）から対象名詞 *book*，*child* として獲得されたルールである.

（1） This is the only *book* which I bought yesterday.

（2） Bobby is the only *child* in his family.

この例では，二つのルールの特徴スロットの比較により，前置修飾カテゴリーの全要素と対象カテゴリー

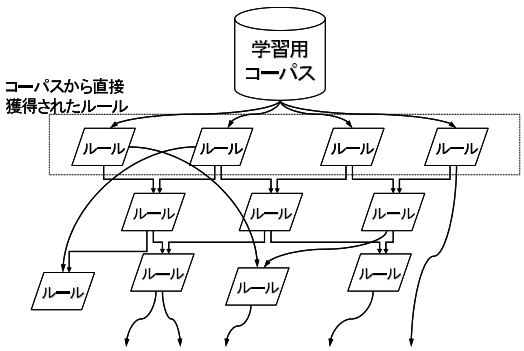

図 4　再帰的学習の例
Fig. 4　An example of recursive learning.

の一部要素，そして冠詞部分が共通部分で，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．対象カテゴリーの“*”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，単に共通部分をもつルール同士である点だけではない．なぜなら，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生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冠詞部分の一致は必須とした．更に，二つのルールの特徴スロットにおいて，少なくとも一つの共通部分をもつルール同士を学習処理の対象とした．これらの条件を満たすルールが存在する限り，学習処理は図 4 で示すように再帰的に行われる．

学習処理で生成されたルールに関しては，連続未使用回数がある一定値を超えた場合，ルールの淘汰処理の対象とする．これは，学習で生成されたルールがコーパスから直接獲得されたルールと比較して確実性が劣ると考えられるためである．本論文の **4.** で述べる性能評価実験においては，ルール淘汰処理を実行する連続未使用回数のしきい値を仮に 20,000 回と設定した．

3.4　冠詞選択ルール適用

誤り校正を行う際，ルール抽出部において獲得されたルールのほかに，人手で作成した解析的ルール辞書も用いる．文献 [2], [3], [15] によると，冠詞の用法として，名詞がある特定の単語に修飾を受けた場合，名詞の前に冠詞を付けない規則がある．このような明確な規則の場合，学習によって獲得したルールを用いて誤りを校正するよりも，人手で作成した解析的ルールを用いる方が確実である．現在，本手法において，名詞に冠詞を付けないという解析的ルールの適用条件は以下の三つである．

- 名詞が人称代名詞や名詞の所有格による修飾を受けた場合
- 名詞が限定詞による修飾を受けた場合
- 対象名詞が固有名詞で，かつ冠詞選択ルール辞書に適用できるルールがない場合

限定詞とは名詞を指定・限定する働きをもつ語句で，another や each 等がある．

ルール抽出部で獲得されたルールについては二つのしきい値に応じて検索された上で用いられる．一つ目のしきい値 θ は適応度のしきい値であり，設定された θ の値以上の適応度をもつルールのみ用いられる．二つ目のしきい値 n は，一つの対象名詞に対して適用できるルール数の上限値を表すしきい値であり，設定された n の値を超えるルールの同時適用は行わない．

4.　性能評価実験

4.1　学習用英文コーパス

本実験では，学習用の英文コーパスとして Reuters Corpus [16] に収録されている英文記事を使用した．記事数は 700，総単語数は 169,069 語であった．名詞の異なり単語数は 4,211 語であり，獲得・生成されたルール数は 90,131 個であった．このルールのうち，コーパスから直接獲得されたものは 64,178 個，学習処理によって生成されたものは 25,954 個となった．

4.2　実験対象

4.1 で述べた Reuters Corpus の英文記事の中で，学習には使用していない記事を無作為に九つ選んだものを用いた．それらの記事の冠詞部分を空欄に書き換え，日本人男子理系大学生 2 人が適切と考える冠詞を入力する．このような操作で得られた冠詞誤りが含まれる可能性のある英文記事を実験対象として用いる．記事数は 9，総単語数 1,586 語となり，含まれる冠詞誤りの数は 121 個であった．

4.3　実験手順

まず **2.1** で説明した手法に従って，冠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，実験対象中の冠詞誤りを検出し，校正を行った．

3.2 で述べた二つのしきい値 θ と n の変化による性能の違いを調査するため，θ は 4 種類の値を設定，n は 1 から 64 まで値を変化させて実験を行った．図 5

に，しきい値 θ とそれを満たすルールの個数の推移を示す．θ の設定は，すべてのルールを用いる ($\theta = 0.0$)，約 25,000 個のルールを用いる ($\theta = 0.5$)，約 18,000 個のルールを用いる ($\theta = 1.5$)，約 6,000 個のルールを用いる ($\theta = 3.0$) の 4 種類とした．

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 (2)，(3) で定める Recall，Precision を用いる．

$$Recall = \frac{\text{正しく検出できた誤りの数}}{\text{冠詞誤りの数}} \quad (2)$$

$$Precision = \frac{\text{正しく検出できた誤りの数}}{\text{検出した誤りの数}} \quad (3)$$

また，誤り校正を評価する尺度として，式 (4)，(5) で定める Recall，Precision を用いる．

$$Recall = \frac{\text{正しく校正できた誤りの数}}{\text{冠詞誤りの数}} \quad (4)$$

$$Precision = \frac{\text{正しく校正できた誤りの数}}{\text{検出した誤りの数}} \quad (5)$$

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 [17] を用いる．F-measure は式 (6) で定義される．

$$F - measure = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (6)$$

ここで β は，Recall に対して Precision を何倍重視して性能評価するかを示す．本実験では永田らの手法 [1] の評価実験に倣い，$\beta = 1$ とした．

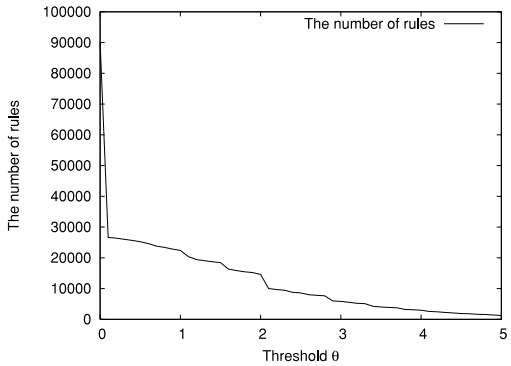

図 5　しきい値 θ を満たすルールの個数の推移
Fig. 5　θ and the number of rules.

5. 実験結果と考察

図 6 に本手法の誤り検出における Recall の結果，図 7 に Precision の結果を示す．これらの図は 4 種類のしきい値 θ で分けてあり，横軸はしきい値 n を示している．n の軸にある “inf” とは，しきい値 n が無限大，すなわちルールの同時適用数を制限しないことを表す．

図 6 より，しきい値 n を制限するほど Recall が高くなる結果となった．特に $\theta = 0.0$ の場合で $n = 1$ のときに Recall が最も高く，F-measure もしきい値の全組合せの中で最も良い結果 (F-measure=0.46) となった．

次に図 7 を見ると，しきい値 n を大きくするほど Precision が高くなることが分かる．特に $\theta = 0.5$，$n = \infty$ の組合せの場合に最も Precision がよい結果（Precision=0.83）となった．しかしながら同様のしきい値の組合せを図 6 で見ると，Recall=0.04 と非常に

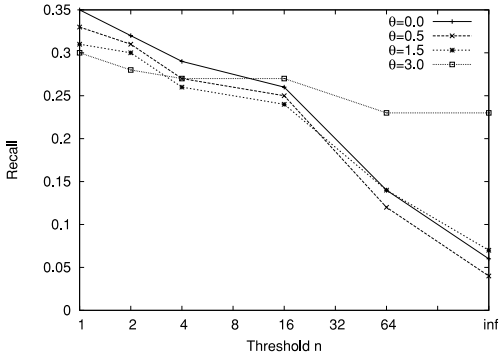

図 6　誤り検出における Recall 値の結果
Fig. 6　The recall ratio of the error detection.

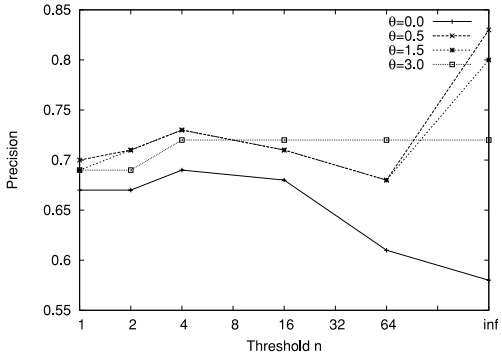

図 7　誤り検出における Precision 値の結果
Fig. 7　The precision ratio of the error detection.

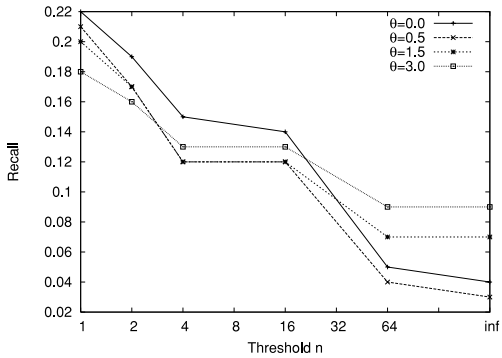

図 8　誤り校正における Recall 値の結果
Fig. 8　The recall ratio of the error correction.

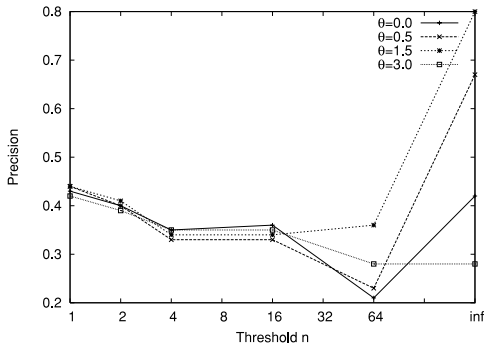

図 9　誤り校正における Precision 値の結果
Fig. 9　The precision ratio of the error correction.

低い結果となっている．これは適応度のしきい値をあまり制限していない上にルールの同時適用数の制限もないので，冠詞の修正候補を絞りきれずに誤り検出ができないことが原因として考えられる．また，$\theta = 3.0$ の場合では Recall，Precision ともに n の変化の影響がほとんどないことが分かる．これは $\theta = 3.0$ という条件のみで，十分に冠詞の修正候補を一つに絞りきれていることによるものと考えられる．$\theta = 3.0$，$n = \infty$ のしきい値の組合せにおいて，複数の冠詞の修正候補を出力した割合は全体の 5 割であった．

図 8 に本手法の誤り校正における Recall の結果，図 9 に Precision の結果を示す．図 8，図 9 を見ると，誤り検出の Recall，Precision の推移とおおむね傾向が同じであることが分かる．しかしながら，誤り校正は誤り検出を正しく行えることができて初めて正しく校正できるので，Precision，Recall ともに誤り検出の場合よりも性能が低くなっている．正しく誤り検出を行えた場合でも，誤った修正候補や複数の修正候補を提示した場合に誤り校正の評価を下げることとなる．誤り校正の場合に最も性能が良好なしきい値の組合せは（$\theta = 0.0$，$n = 1$）で，F-measure が 0.29 という結果となった．また，Precision 重視のしきい値の組合せは（$\theta = 1.5$，$n = \infty$）で，Precision が 0.80 という結果となった．

次に，最も F-measure が高かったしきい値の組合せ（$\theta = 0.0$，$n = 1$）について詳細に考察する．冠詞誤りの検出に成功した際に用いられたルールは，学習処理にて生成された抽象ルールが全体の 79%，学習用英文コーパスから直接獲得されたルールが 19%，人手作成した解析的ルールが 2%となった．抽象ルールの利用率が約 8 割に達していることから，学習処理による汎用的なルールが有効であることが確認できたと考えられる．

次に，冠詞誤り検出に成功した例を以下に挙げる．

(i)　took a *part* in the vote.

(ii)　a *company* said.

上記の英文 (i)，(ii) は表 2 の (i)，(ii) に示す英文の一部分である．英文 (i) において，対象名詞 *part* に付与する冠詞は不定冠詞 a ではなく正しくは無冠詞であるが，この誤りを正しく検出することができた．この誤り検出に用いられたルールは，“前置詞 in に修飾を受け，名詞句に属する名詞 part ならば無冠詞” といったものであった．前置詞 in による前置詞句の特徴スロットは抽象化されており，前置詞 in を伴うならばどのような前置詞句でも適用できるルールであった．英文 (ii) においては，対象名詞 *company* に付与する冠詞は不定冠詞 a ではなく正しくは定冠詞 the である．英文 (ii) は文章中の一部であり，前文に出現した具体的な社名を受けて定冠詞をとる例となっている．本手法では文中文脈のみを考慮しているため，前文の影響を判断した冠詞誤り検出はできない．しかしながら，学習データ中に “company”，“said”，“定冠詞 the” の表現が頻出したため，英文 (ii) の誤りを検出することができた．

しきい値の組合せ（$\theta = 0.0$，$n = 1$）の誤り検出における Recall は 0.35 であったので，65%の冠詞誤りは検出することができなかった．検出できなかった誤りを更に分類すると，86%は適用できるルールがなかったために検出することができなかった．残りの 14%は，システムが正しい冠詞の用法であると判断してしまったために誤り検出できなかった．

適用できるルールがなかったために検出できなかっ

表 2　誤り検出例の全文
Table 2　The full text of the error detection examples.

No.	全文
(i)	About 9.5 million shares, or more than 80 percent of Waterhouse's outstanding shares, took a *part* in the vote.
(ii)	Waterhouse Investor Services Inc said on Tuesday that its stockholders overwhelmingly approved the merger with a newly formed subsidiary of Toronto-Dominion Bank. About 9.5 million shares, or more than 80 percent of Waterhouse's outstanding shares, took a part in the vote. Of those, more than 98 percent voted in a favor of the merger, a *company* said.
(iii)	In the promotion launched by the Coca-Cola Poland Services Sp. z.o.o., customers who submitted two bottle tops printed on inside with the matching dates and the venues for the *past Olympic Games*, plus the same sum of money, could win cash prizes.
(iv)	The bill will boost the wage, typically paid to the unskilled workers in the restaurants and the stores, by 50 cents to $4.75 on Oct. 1 and by 40 cents to $5.15 on Sept. 1, 1997. It will sweeten a pill for the mostly small businesses that will pay *higher labor costs* by handing them some $22 billion in tax breaks over 10 years, paying for this in part by reimposing 10 percent tax on the airline tickets.

た誤りの例を以下に示す.

(iii) for the *past Olympic Games*.

上記の英文 (iii) は表 2 の (iii) に示す英文の一部分である．英文 (iii) において，対象名詞 *past Olympic Games* の冠詞の正解は無冠詞であるが，適用できるルールは存在しなかったため，誤りを検出または校正することができなかった．

このような 86%の誤りに関しては，学習規模を拡大させることによって検出できる可能性がある．より多くの学習を行うことによって，ルールにおける名詞や文脈情報の網羅性を高めることで，Recall が改善できると考えられる．しかしながら，これらの検出できなかった誤りのうち，特定の会社名や地名などの固有名詞に起因するものが 26%含まれていた．これら固有名詞は様々なバリエーションが存在するため，学習データに複数回出現することがないような固有名詞も存在する．そのため，現在の手法のまま学習データの規模を拡大しても検出性能の改善の可能性は低いと考えられる．したがって，本手法で現在用いている特徴スロットによるルールの他に固有名詞専用のルールを用意することで，固有名詞に関する冠詞誤りの検出性能の改善ができると考えられる．

次に，システムが正しい用法であると判断してしまったために検出できなかった誤りの例を以下に示す.

(iv)　*higher labor costs*

上記の英文 (iv) は表 2 の (iv) に示す英文の一部分である．英文 (iv) において，対象名詞 *higher labor costs* の冠詞の正解は定冠詞 the である．この英文は文章の一部分であり，前文で labor costs についての話題が触れられている．その前文の影響を受け，定冠詞をとる例となっているが，本手法で用いるルールでは前文を考慮していないため，文中文脈のみで判断してそのまま無冠詞で正しいという結果を出力した．

このような 14%の誤りに関しては，適用できるルールが存在するにもかかわらず誤りを検出することができなかった．これらの誤りには固有名詞に起因するものは全く含まれていなかったため，学習規模の拡大で改善できる可能性がある．また，これらの誤りのほとんどが定冠詞にかかわる誤りであったので，定冠詞と関連が強い前文の文脈情報を冠詞選択ルールにおいて考慮することで，検出できる可能性がある．現在本手法では **3.1** で述べた特徴スロットが示すように，冠詞選択ルールにおいて文中文脈のみ考慮している．

6.　関連する手法との比較

4.2 で述べた実験対象について，Web 上で公開されている永田らの手法 [1][(注2)], [7][(注3)] と本手法とで評価実験を行った．永田らの手法 [1] の精度/検出率パラメータは 1〜4 の 4 段階設定のうち，やや精度重視である 2 と設定した．永田らの手法 [1], [7] は冠詞誤り検出手法であるので，本手法の誤り検出の評価結果と比較を行うこととした．

永田らの手法 [1], [7] では，学習用の英文コーパスとして EDR 電子化辞書 [19] に収録されている英語コーパスと日英対訳辞書の英語語義文を約 300 万語の規模で用いている．可能な限り同条件で比較実験を行うために，本実験では EDR 電子化辞書の英語コーパスを学習データとした提案手法を比較対象として用いた．しかしながら，提案手法の場合，学習データから直接

(注2)：http://www.ai.info.mie-u.ac.jp/~nagata/error_detection/index.html
(注3)：http://www.ai.info.mie-u.ac.jp/~nagata/mc/system.html

表 3　関連手法との比較
Table 3　Comparison of our method and related work.

	Precision	Recall	F-measure
提案手法（$\theta = 0.0$，$n = 1$）	0.63	0.31	0.41
永田らの手法 [1]	0.63	0.22	0.33
永田らの手法 [7]	0.47	0.27	0.35

ルールを獲得するだけでなく，学習処理によって抽象化したルールを再帰的に生成する．そのため，統計モデルを用いる永田らの手法 [1], [7] と比較して，ルールの学習に時間を要すると考えられる．EDR 電子化辞書の英語コーパスの全データから学習を行うことは現実的でないので，提案手法では英語コーパスのデータをランダムに抽出してルールの学習を行う処理を，時間が許す限り繰り返した．その結果，実際に学習に用いたコーパスの規模は 95,988 語となった．

表 3 に本手法で最も性能が良好な結果となった（$\theta = 0.0$，$n = 1$）の場合と，永田らの手法 [1], [7] との性能の比較結果を示す．表 3 より，永田らの手法と比較して提案手法の学習データが約 30 分の 1 と非常に少ないながらも，Recall，若しくは Precision と Recall の両方において提案手法の結果が永田らの手法より良好なことを確認できた．

Precision については，永田らの手法 [7] を上回る結果となった．永田らの手法 [1] とは同程度の性能であったが，必要とする学習データの量は提案手法の方が少ないことが優れている点であると考えられる．ただ，永田らの手法 [1] において精度/検出率パラメータを Precision 重視である 1 にした場合，Precision は 0.67 程度まで向上する可能性がある．これは，永田らの手法 [1] のベースである文献 [20] におけるしきい値と冠詞誤り率の推移グラフから推定したものである．しかしながら，提案手法においても例えば $\theta = 1.5$，$n = 4$ を選択することで Precision が 0.70，Recall が 0.22 となり，Recall を極端に悪化させないまま Precision を向上させることができる．そのため，Precision における優位性は維持できると考えられる．一方で，永田らの手法 [1] において精度/検出率パラメータを 2 より上げて Recall 重視とした場合でも，Precision を同程度の性能に保つことはできないと推定されるため，Precision における提案手法の優位性はあると考えられる．

また，Recall については永田らの手法 [1], [7] の両方を上回る結果となった．永田らの手法 [1], [7] では未知名詞に対応できないという問題点がある．その点，提案手法ではたとえ未知名詞であっても，学習処理によって生成された抽象ルールを用いることによって，名詞以外の他の文脈要素から判断して誤りを検出することができる．実際に本実験において，提案手法で誤り検出を行う際に用いられたルールのうちの 90%は抽象化されたルールであった．提案手法の未知名詞に対応可能である点は，永田らの手法 [1], [7] と比較して優れている点であると考えられる．

7. むすび

本論文では，単語出現状況の特徴を用いた英文冠詞誤りの検出及び自動校正手法を提案した．実験の結果，誤り検出において最も性能が良好なもので，Precision が 0.67，Recall が 0.35 となり，F-measure が 0.46 となった．2 種類のしきい値によって Precision と Recall のバランスをある程度コントロール可能であることが確認できた．また，関連手法と同一実験対象で比較したところ，優位性のある結果を示し，本手法の有効性を確認することができた．

今後の課題としては，Recall の更なる改善が挙げられる．学習規模を拡大することで，特徴の網羅性を高めることができると考えられる．しかしながら，過学習による性能の低下も考えられるため，最適な学習量を推定することは重要である．また，本手法では前後の文の文脈は考慮していないが，冠詞が前後の文に影響されることは明らかであるので，今後考慮する必要がある．

現在本手法では学習データとして Reuters Corpus [16] を，英語母語話者によって書かれた正しい英文と仮定して用いている．それゆえ，本システムは冠詞の正しい使用例から学習を行い，冠詞誤りを検出・校正するものとなっている．しかしながら，誤りの検出や校正が目的の場合，正解例のみからの学習よりも，実際に日本人学習者が誤った例と正解例の組合せから学習する方が，より精度向上を期待できるルールを獲得できると考えられる．このような学習者による誤りと正解例の組合せを含んだコーパスの一つとして，The NICT JLE Corpus [9] がある．本コーパスは話し言葉であるため，今後は話し言葉における本手法の適用というテーマで別途研究を進める予定である．

文　献

[1] 永田　亮，井口達也，脇寺健太，桝井文人，河合敦夫，井須尚紀，“前置詞情報を利用した冠詞誤り検出，” 信学論（D-I），vol.J88-D-I, no.4, pp.873–881, April 2005.

[2] 石田秀雄，わかりやすい英語冠詞講義，大修館書店，2002.
[3] 原田豊太郎，例文詳解　技術英語の冠詞活用入門，日刊工業新聞社，2000.
[4] J. Lee, “Automatic article restoration,” Proc. HLT/NAACL Student Research Workshop, pp.195–200, Boston, USA, May 2004.
[5] 河合敦夫，杉原厚吉，杉江　昇，“英文の誤りを検出するシステム ASPEC-I,” 情処学論，vol.25, no.6, pp.1072–1079, Nov. 1984.
[6] 森田光宏，“日本人学習者の不定冠詞とゼロ冠詞の使用—抽象・具象名詞の視点から,” 日本学術振興会平成 13，14，15 年度科学研究費補助金基盤研究（C)(2）研究成果報告書，pp.47–60, 2003.
[7] 永田　亮，若菜崇宏，河合敦夫，森広浩一郎，桝井文人，井須尚紀，“可算/不可算の判定に基づいた英文の誤り検出,” 信学論（D)，vol.J89-D, no.8, pp.1777–1790, Aug. 2006.
[8] E. Izumi, K. Uchimoto, T. Saiga, T. Supnithi, and H. Isahara, “Automatic error detection in the Japanese learners’ English spoken data,” The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics, pp.145–148, Sapporo, Japan, July 2003.
[9] 和泉絵美，内元清貴，井佐原均，日本人 1200 人の英語スピーキングコーパス，アルク，2004.
[10] K. Araki and K. Tochinai, “Effectiveness of natural language processing method using inductive learning,” Proc. IASTED International Conference Artificial Intelligence and Soft Computing, pp.295–300, Mexico, 2001.
[11] 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.
[12] 関根　聡，“英語構文解析システム「Apple Pie Parser」,” 情処学論，vol.41, no.11, pp.1221–1226, Nov. 2000.
[13] 織田　稔，英語冠詞の世界，研究社，2002.
[14] 正保富三，英語冠詞がわかる本，研究社，1996.
[15] The Concise Oxford English Dictionary 10e, Oxford University Press, 2001.
[16] Reuters Corpus: http://www.reuters.com/researchandstandards/corpus
[17] C.J. Van Rijsbergen, Information Retrieval, 2nd ed., Butterworths, 1979.
[18] L. Burnard, ed., Users reference guide for the British National Corpus. version 1.0, Oxford University Computing Services, Oxford, 1995.
[19] Japan Electronic Dictionary Research Institute Ltd, EDR electronic dictionary specifications guide, Japan Electronic Dictionary Research Institute Ltd., Tokyo, 1993.
[20] 永田　亮，井口達也，脇寺健太，桝井文人，河合敦夫，“日本人英語学習者のための冠詞誤り検出,” 信学論（D-I)，vol.J87-D-I, no.1, pp.60–68, Jan. 2004.

（平成 18 年 8 月 24 日受付，19 年 1 月 8 日再受付）

乙武　北斗　（学生員）

2005 北大・工・情報工卒．現在，同大大学院博士後期課程在学中．文書の校正などの自然言語処理の研究に従事．

荒木　健治　（正員）

1982 北大・工・電子卒．1988 同大大学院博士課程了．工博．同年，北海学園大学工学部電子情報工学科助手．1989 同講師．1991 同助教授．1998 同教授．1998 北海道大学大学院工学研究科電子情報工学専攻助教授．2002 同教授．現在，北海道大学大学院情報科学研究科メディアネットワーク専攻教授．自然言語処理，特に，機械翻訳，音声対話処理などの自然言語処理の研究に従事．情報処理学会，言語処理学会，人工知能学会，認知科学会，ACL，IEEE，AAAI 各会員．