日本人英語学習者の誤り例を用いた帰納的学習による英文冠詞誤りの自動検出

乙武北斗　　荒木健治
Hokuto OTOTAKE　　Kenji ARAKI
{hokuto,araki}@media.eng.hokudai.ac.jp

北海道大学　大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

1. はじめに

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．冠詞は文脈を明確にする働きがあるため，特に新聞記事や論文において正しい冠詞を用いることは重要である．しかしながら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，さらに専門知識も必要となる[1].

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が提案されている．Izumi らの手法[2]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（以下，The NICT JLE Corpus）[3]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する．また，我々が以前提案した手法[4]では，英字新聞等の冠詞の用法を実例として用いた帰納的学習により，自動的に冠詞誤りを検出・校正するルールを抽出する．

本稿では日本人英語学習者の誤り例を用いた帰納的学習による英文冠詞誤りの自動検出手法を提案する．提案手法では，英文中の単語出現状況における帰納的学習による冠詞誤り検出手法[4]をベースに，The NICT JLE Corpus[3]に含まれる冠詞の誤り例を誤り検出ルールに考慮することによる精度の向上を図る．

以下，**2.** で提案手法の概要を，**3.** では提案手法を用いた誤り検出の評価実験と考察について述べる．最後に **4.** でまとめを述べる．

2. システムの概要

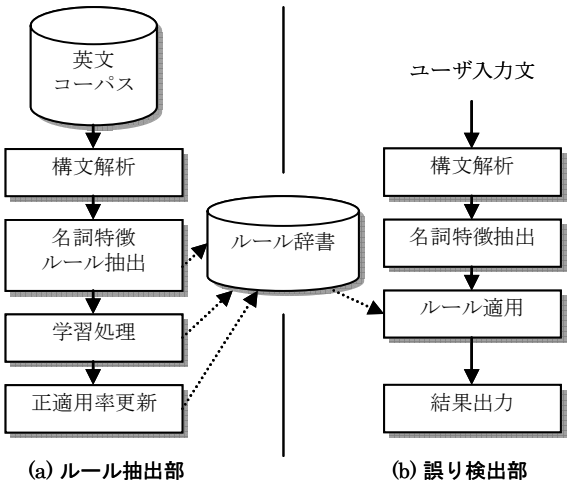

図 1　処理過程

本システムの処理の流れを図 1 に示す．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．ひとつはルール抽出部，もうひとつは誤り校正部である．

2.1 ルール抽出部

ルール抽出部では冠詞誤り検出のためのルールを The NICT JLE Corpus[3]から自動的に抽出する．以下，各処理の概要を述べる．

2.1.1 名詞特徴・ルール抽出

英語の冠詞についての知見[1][5]に基づき，冠詞の選択を決定する重要な要素をカテゴリとする特徴スロットを考える．なお，特徴スロットの詳細については文献[4]を参照されたい．

以下の The NICT JLE Corpus エラータグ付き英文

(a) This is **<at crr="the"></at>** only *book* which I bought yesterday.

対象		
	名詞	book
	主名詞	book
	属する句	NP
	前置詞	−
	名詞目的語動詞	be
	名詞主語動詞	−
	数	singular
	固有	no

前置修飾		
	修飾詞	only
	品詞	RB

後置修飾			
前置詞句	前置詞	−
	冠詞	−
	名詞	−
	主名詞	−
	修飾詞	−
不定詞句	動詞	−
	目的語冠詞	−
	目的語	−
	副詞	−
関係詞節	主語	I
	動詞	buy
	目的語冠詞	−
	目的語	−
	副詞	yesterday

図 2　特徴スロットの例

における対象名詞を *book* とした場合の特徴スロットを図 2 に示す.

book の前にあるエラータグにおいて, at は冠詞誤りであることを示し，crr 属性値が正解例を表わす．そして開始タグと終了タグの間に日本人英語学習者の誤り例が示される．英文(a)の例では正解例が the，誤り例は無冠詞となる.

誤り検出ルールは，特徴スロットと冠詞の正解例を組み合わせた正ルールと，冠詞の誤り例と組み合わせた負ルールの 2 種類がある．英文(a)の例において，正ルールは「ある名詞の特徴スロットが図 2 の特徴スロットと一致した場合，冠詞 the を付与」というものになる．一方で負ルールは「図 2 の特徴スロットと一致した場合，少なくとも無冠詞ではない」となる.

このようなルール抽出処理を学習用英文コーパスの全名詞に対して行う.

2.1.2 学習処理

本稿では学習処理において「実例からそこに内在している規則を獲得すること」と定義される帰納的学習[6]を用いる．提案手法での実例とは学習用コーパスから抽出される特徴スロットである． 2 つの特徴スロットの各要素について，字面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．新たなルールの特徴スロットには共通部分の要素が残り，差異部分は抽象化される．このような処理を登録されているルールの全ての組み合わせに対して行い，その結果生成された抽象ルールに対しても再帰的に学習処理を行っていく[4].

2 つのルールが学習処理対象となる条件を以下に示す.

- 主名詞もしくは前置修飾詞が共通部．これら要素は冠詞選択において重要と考えられるからである.
- 正ルール同士もしくは負ルール同士の場合，冠詞の一致．結果生成されるルールにおいて，一致した冠詞が継承される.
- 正ルールと負ルールとの学習処理の場合，冠詞の不一致．結果生成されるルールは負ルールとなり，負ルール側の冠詞が継承される.

2.1.3 正適用率更新

これまで述べてきた処理によって得られたルールに対して学習用コーパスの英文と照らし合わせる処理を行うことにより，ルールの確からしさを算出する．その指標として，式(1)で定義する正適用率を用いる.

$$\text{正適用率} = \frac{\text{正適用回数}}{\text{適用回数}} \tag{1}$$

全てのルールはこの正適用率の降順で順位付けされる.

2.2 誤り検出部

誤り検出部では，まずルール抽出部と同様に構文解析された入力文から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構築されたルール辞書から適用できるルールを検索し，冠詞誤りの検出を行う．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．最後に結果を出力する.

ルール抽出部で獲得されたルールについては正適用率の閾値に応じて検索された上で用いられる．正適用率が設定された閾値 θ 以上の値を持つルールのみ用いられる.

3 性能評価実験

3.1 実験方法

本実験では学習データ, 評価データともに The NICT JLE Corpus[3]を用いた．同様のデータを用いた和泉らの手法[3]の評価実験に倣い，エラータグ付きデータ全 167 件から無作為に 151 件を学習用，16 件を評価用と

設定した. 学習データの総単語数は約 25 万語となった. 評価データの総単語数は約 3 万語，含まれる冠詞誤りの数は 431 個（脱落誤りが 292 個，置換／余剰誤りが 139 個）となった.

2.1 で説明した手法に従って誤り検出ルール辞書を構築した結果，獲得・生成されたルール数は 218,657 個であった．次に **2.2** で説明した手法を用いて，評価データ中の冠詞誤りの検出を行った.

本実験においてルールの正適用率における閾値 θ は仮に 0.8 と設定した．また，エラータグを考慮することによるルールの有効性を検証するために，エラータグから獲得される負ルールを用いる場合と用いない場合の両方で評価を行った．さらに，関連手法との比較のため，和泉らの手法[3]の評価実験との比較も行った.

3.2 評価方法

本実験では和泉らの手法[3]の評価実験に倣い，評価を 2 種類の冠詞誤りに分けて考える．一つは脱落誤りで，これは本来必要な冠詞が抜けてしまっている誤りを表わす．もう一つは置換／余剰誤りで，誤った冠詞を用いている，または本来冠詞が不要な部分に冠詞が挿入されている誤りを表わす.

また，本実験において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.

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

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

3.3 実験結果

初めに脱落誤りの検出結果を表 1 に示す．表 1 において提案手法のエラータグ考慮の有無を比較すると，エラータグを考慮することで僅かな Precision の低下は見られるものの，明確な Recall の向上が確認できる．また，和泉らの手法[3]と比較すると，Recall，Precision ともに提案手法の結果が上回った.

次に置換／余剰誤りの検出結果を表 2 に示す．表 2 において提案手法のエラータグ考慮の有無を比較すると，脱落誤り検出の場合と異なり，ほとんど性能に差が出ないという結果になった．和泉らの手法[3]と比較すると，Precision において同程度の性能を保ちながら Recall の改善を確認できた.

表 1　脱落誤りの実験結果

	Recall	Precision
提案手法	0.71	0.81
提案手法（エラータグなし）	0.58	0.84
和泉らの手法[3]	0.50	0.60

表 2　置換／余剰誤りの実験結果

	Recall	Precision
提案手法	0.47	0.30
提案手法（エラータグなし）	0.46	0.30
和泉らの手法[3]	0.15	0.30

表 3　冠詞誤りすべての実験結果

	Recall	Precision
提案手法	0.63	0.58
提案手法（エラータグなし）	0.54	0.56
和泉らの手法[3]	0.38	0.53

最後に脱落誤りと置換／余剰誤りとをまとめた検出結果を表 3 に示す．誤り全体としてみた場合，提案手法におけるエラータグ考慮の有無によって，特に Recall での違いが著しい．エラータグの考慮によって脱落誤りにおける Recall の向上の結果，誤り全体においてもその結果が反映されたと考えられる．また，和泉らの手法[3]と比較した場合でも，特に Recall において明確な性能向上が確認できた.

3.4 考察

まず提案手法におけるエラータグ考慮の効果について考察する．**3.3** で述べたように，提案手法においてエラータグから獲得される負ルールを用いることによって，冠詞の脱落誤り検出において Recall が改善されることが確認できた．また，冠詞誤り全体としてみた場合，Precision も若干の向上が見られており，エラータグによる冠詞誤り例の考慮は有効であると言える．しかしながら, 置換／余剰誤りに限定して注目した場合，エラータグの考慮による性能向上はほとんど見られない．この原因の一つとして，The NICT JLE Corpus[3] 中にエラータグが付与されている日本人英語学習者の冠詞誤りにおいて，置換／余剰誤りの数は脱落誤りの数と比較して半分以下しかないことが挙げられる．その結果，置換／余剰誤りに対応するルールの数は脱落

誤りに対応するものと比べて十分ではない数となってしまい，性能向上が見られなかったと考えられる．また，置換／余剰誤りの性能が脱落誤りの性能と比較して悪いことも，同様の原因が考えられる．置換／余剰誤りの検出性能を改善するためには，さらに数を増やした置換／余剰誤り例からルールを抽出する必要があると考えられる．

次に，提案手法と同様に The NICT JLE Corpus[3]をデータとして用いた和泉らの手法[3]との比較結果を考察する．**3.3** で述べたように，すべての冠詞誤りにおいて提案手法が優位性のある結果を確認できた．提案手法では特に Recall の性能が高いことが確認できた．これは，提案手法における，少ない学習データから汎用性の高い抽象的なルールを再帰的に生成できる帰納的学習の特徴によるものと考えられる．本実験で用いた誤り検出ルール辞書には 218,657 個のルールが含まれているが，そのうち 187,680 個は学習処理において生成された抽象ルールである．また，実際に誤り検出で用いられたルールの 9 割以上は抽象ルールである．したがって，提案手法の帰納的学習による汎用性の高さは和泉らの手法[3]と比較して優れている点であると考えられる．その一方で，帰納的学習は再帰的にルールを抽象化していく処理のため，学習データを増加させればさせるほど処理時間も著しく増加する．これは和泉らの手法[3]と比較して改善すべき点であると考えられる．

4 まとめ

本稿では、英文コーパス中の英文を実例とし，そこに内在する冠詞選択のルールを再帰的に自動獲得する冠詞誤り検出システムにおいて，The NICT JLE Corpus[3]に含まれるエラータグによる冠詞誤り例をルールに考慮する改善手法を提案した．性能評価実験の結果, 冠詞誤り全体として Recall が 0.63, Precision が 0.58 となった．エラータグを考慮しない場合と比較して約 1 割の Recall の性能向上が見られ，日本人英語学習者の誤り例をルールに考慮することは有効であることが確認できた．また，関連手法と比較したところ，提案手法が Recall の性能において優れていることが確認できた．

今後の課題としては，冠詞の置換／余剰誤りの性能向上が挙げられる．学習データを増加させた上で再実験を行い，性能向上に有効であるか調査する．また，The NICT JLE Corpus[3]には冠詞以外にも様々な文法誤りに対してエラータグが付与されている．提案手法は現在英文冠詞誤りを対象としているが，その他の日本人英語学習者が起こしやすい誤りに対しても対応し，エラータグによる誤り例の考慮が有効であるかどうかを確認したいと考えている．

謝辞

The NICT JLE Corpus のデータをご提供していただきました NICT の和泉絵美氏に感謝致します．

参考文献

[1] 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，東京，2000.

[2] E. Izumi，K. Uchimoto，T. Saiga，T. Supnithi and H.Isahara，“Automatic error detection in the Japaneselearners’ English spoken data”，The Companion Volumeto the Proceedings of 41st Annual Meeting of theAssociation for Computational Linguistics，pp.145-148，Sapporo，Japan，Jul.2003．

[3] 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，（株）アルク，東京，2004.

[4] 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.

[5] 石田秀雄，わかりやすい英語冠詞講義，大修館書店（株），東京，2002.

[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ー, 森北出版（株），東京，2004.