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

2007－DBS－143（24）
2007／7／2

検索エンジンを用いた英文冠詞誤りの検出

平野 孝佳† 平手 勇宇†,†† 山名 早人†††,††††

† 早稲田大学大学院基幹理工学研究科 〒169-8555 東京都新宿区大久保3-4-1
†† 早稲田大学メディアネットワークセンター 〒169-8050 東京都新宿区戸塚町1-104
††† 早稲田大学理工学術院 〒169-8555 東京都新宿区大久保3-4-1
†††† 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋2-1-2
E-mail: {hirano,hirate,yamana}@yama.info.waseda.ac.jp

あらまし 近年，英語の必要性はますます高くなっており，英作文を書く機会も増えてきている．本稿では，日本人の英作文によく見られる冠詞誤りを，検索エンジンを用いて検出する手法を提案する．検索エンジンを用いる手法は，検索エンジンがインデックス化している膨大なウェブページのテキストデータを利用することができるため，従来のコーパスを用いた手法では検出できなかった誤りを検出することが可能である．検索エンジンを用いた従来手法として，単純なフレーズを用いてフレーズ検索するLapataらの手法があるが，希出なパターンには対応できないという欠点があった．本稿では，冠詞前後の複数の単語について活用形を考慮したり、冠詞に影響を与えないと考えられる単語を除去するといった類似フレーズを用いた拡張を行い，パターンについても判定できるよう改善した．実験の結果，提案手法はLapataらの手法より，一般的な文章で0.04ポイント，技術的な文章で0.19ポイント高い性能(F-measure)で誤りを検出できることを確認した．

キーワード 冠詞誤り，英作文，検索エンジン

Detecting Article Errors in English using Search Engines

Takayoshi HIRANO†, Yu HIRATE†,††, and Hayato YAMANA†††,††††

† Graduate School of Fundamental Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan
†† Media Network Center, Waseda University 1-104 Totsuka-cho, Shinjuku-ku, Tokyo, 169-8050, Japan
††† Science and Engineering, Waseda University 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan
†††† National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
E-mail: {hirano,hirate,yamana}@yama.info.waseda.ac.jp

Abstract Recently, both the necessity for English and opportunities to write English have became higher and higher among non-native English speakers. But most of Japanese people tend to made many errors in English article usage when they write English. In this paper, we propose a method for detecting article errors in English by using search engines. Since search engines index great amounts of text data on web pages, search engine based methods are able to detect undetectable errors which conventional corpus based method cannot detect. Lapata et al. proposed a method for detecting article errors based on search engines. But Lapata's method using simple phase cannot detect low frequency words. Compared to the Lapata's conventional method, our proposing method generates similar phases to improve F-measure value especially in technical texts. Our experimental results show that our method is able to improve F-measure value 0.04 point in general texts and 0.19 point in technical texts.

Key words article errors, English composition, Search engines

1. はじめに

近年，企業や学校の英語教育の促進などの点から，英語に触れる機会や英作文を書く機会が増えている．これまで，英作文の誤り校正は，専門知識を持った人によって行われてきた．しかし，人手による校正は多くの時間と労力を必要とするため，文法誤りを自動検出するシステムの需要が増えている．文法誤りの中でも日本人の書く英作文には，冠詞の文法の誤りを多く

含むという傾向があると報告されている [1] [2].

本稿では，この冠詞誤りに注目し，冠詞誤りを検出する手法を提案する．これまでの多くの研究は，コーパスを用いる手法 [3] [4] [5] [6] と検索エンジンを用いる手法 [7] に分けられる．コーパスを用いる手法は，直接テキストデータを扱えるため，様々なアプローチが考えられるが，コーパスの文章量に限度があるためスパースネスの問題が起こりうる．一方，検索エンジンを用いる手法は，ウェブ上の 100 億を超える膨大なウェブページのテキストデータをインデックス化しているため文章量は現在最大の英文コーパス BNC コーパス [8] の数百倍の文章を対象とした解析ができる．これにより，辞書に載っていないような新語，流行語なども，検出可能である [9]．その一方で，解析には API を用いるため，用いる API の仕様によるため解析手法には大きな制約がある．

本稿で提案する手法は，検索エンジンを用いる Lapata らの手法 [7] を拡張させたものである．提案手法では文章から切り出した単純な N 単語からなるフレーズだけではなく類似の N 単語フレーズも用いる．従来手法より検索フレーズの条件を緩くすることで，より広い範囲の N 単語フレーズを用いた解析ができるようになる．これにより提案手法では，文章に含まれる冠詞を決定する特徴をより活かすことができるようになると考えられる．

本論文は，以下の構成を取る．第 2 節では，本手法に関する関連研究について述べる．第 3 節では，提案手法である検索エンジンを用いた手法について述べる．第 4 節では，第 3 章で提案した手法と既存手法との比較実験を行い，考察する．最後に第 5 節で，まとめを述べる．

2. 関連研究

本節では，冠詞誤り検出に関連する研究について述べる．冠詞誤りを検出する研究は，昔から行われている．古くの研究 [1] では，人手で作られたルールに基づいて冠詞誤りを検出していた．しかしルール作成には専門知識が必要であり，すべての冠詞の用法を人手で作成することは，非常に困難であるとされている．そこで，大規模なコーパスから誤りを判定するためのルールを自動で抽出する手法が研究されている [3] [4] [5] [6]．最近では，検索エンジンを利用して冠詞誤り検出を行う研究も行われている [7]．以下，本研究に関する冠詞誤りの関連研究について述べ，関連研究の特徴をまとめる．

2.1 コーパスベースの冠詞誤り検出 [3] [4] [5] [6]

コーパスベースによる手法では，英字新聞などから作られた大規模英文コーパスを用いる．そのため文章の書き手の信頼度が高いという利点がある．また，コーパス上の全ての文章を直接参照することができるため，様々な解析を行うことができる．一方で，コーパスベースの手法における問題点として，ルールの不足が上げられている．ルールが足りない原因は，コーパスに出現する単語の文章量が足りないためであり，スパースネスの問題といわれている．コーパスベースの冠詞誤り検出は，現在の言語資源では文章量の限界の問題があるため，複雑な言語モデルのルールを網羅することは難しい．コーパスベースの問題点であるスパースネスの問題を解決するためには，コーパスを拡大し文章量を増やすことが必要である．

2.2 検索エンジンのコーパスとしての利用可能性 [7] [10]

検索エンジンを用いて，自然言語処理に利用する際の考えられている問題点をまとめる．

検索エンジンを扱う上での一つの問題点として，意図しない用例の単語が含まれるということがある．例えば，"health care"という単語を検索した場合に，名詞句としての用法の頻度を得ることを望んでいても，care が動詞として使われている用法を含むということが起こってしまう．これと関連した問題として，多くの検索エンジンでは，"."（ピリオド）や","（カンマ），"()"（括弧），"/"（スラッシュ）などを無視している．そのため 2 単語の名詞句を調べても，文をまたいでいたり，括弧をはさんでいたりといったようなことが起こる．

次に，ヒット数の信頼性についても述べられている．例えば，同一の web ページ上に，w_1w_4, w_2w_4, w_3w_4 の 3 つの 2 単語の名詞句が存在しているとする．この場合，w_4 の実際の頻度は 3 であるのに，ページカウントは 1 となる．このようなことが頻繁に起こりうるため，実際のウェブ上での単語出現頻度よりも小さい値が検索エンジンからヒット数として返されると考えられる．またリンクやファイル名などテキストデータ以外の情報に単語が存在して，ページカウントされる場合もある．

検索エンジンを自然言語処理に用いる場合には，このような問題を考慮しつつ，工夫して情報の信頼度の高い手法を考えなければならないという制約がある．

2.3 検索エンジンを利用した冠詞誤り判定に関する研究 [7]

検索エンジンを利用した冠詞誤り検出については，検索エンジンの API を利用するという限られた制限の中で行うため，工夫が必要である．

Lapata らによって，2005 年に提案された手法 [7] では，単純な N 単語からなるフレーズのフレーズ検索を行うことで高い精度で冠詞誤りを検出できるとしている．構文解析を用いて名詞句を抽出し名詞句の冠詞を {a/an, the , ϕ } に，変化させて 3 つのクエリを生成することによって，3 パターンのヒット数を取得し，比較している．ここでは，ϕ は無冠詞を表している．3 つのクエリの検索結果数のなかで，最も多かったものを答えとする手法である．この手法では名詞句と冠詞とその前の 2 つの単語のフレーズ検索から冠詞を求める手法が最適であるという結果になった．この手法は単純なものであるが，フレーズ検索によってもある程度冠詞誤りを検出できることを示している．しかし，これは単純な手法であり，名詞の単数・複数を考慮していないため全ての誤りを検出できないと考えられる．

2.4 冠詞誤り検出手法の先行研究のまとめ

本節では，冠詞誤りに関連する研究を述べた．コーパスベースの手法では，単語量の不足という欠点があるものの，全ての単語を様々な形で解析することが可能である．検索エンジンを用いる手法では，コーパスの数万倍の情報が扱える．しかし，Lapata らの手法は単純であり名詞の単数・複数を考慮していないため，全ての誤りを検出することが難しい．これらのことから Lapata らの手法を拡張した類似フレーズを用いる手法を

次節で提案する．

3. 提案手法

本節では，検索エンジンの検索クエリに類似フレーズを用いた冠詞誤り検出手法を提案する．以下，3.1 では提案システムの概要を述べ，3.2 で入力文章の解析と検索クエリの生成を述べる．次に 3.3 で検索の実行と誤り判定を述べ，3.4 でまとめを行う．

3.1 提案システムの概要

Lapata らの手法 [7] では，単純なフレーズ検索のみを検索クエリとして用いている．これは，よく使われるフレーズならば，有効である．しかし専門的な表現を多く含む文章では，単語のヒット数が少ないためフレーズでの検索ヒット数がさらに少なくなる．特に，論文など専門的な文章では複雑な名詞句がたびたび現れる．この欠点を補うため，本論文では，与えられた文章から，構文解析器を用いて必要の無い単語を除去することや，名詞や動詞の変化形を OR 演算子によりつなぐという類似フレーズを検索クエリとして追加することで，様々な文章により柔軟に対応できる手法を提案する．システムの主な流れは次のようになる．

（1） **入力文章の解析**：与えられた文章列から構文解析器を用いて，名詞句や動詞句などを抽出する．

（2） **検索クエリ生成**：構文解析の結果から，検索に必要なクエリを生成する．

（3） **ヒット数の取得**：生成されたクエリを用いてフレーズ検索を行う．判定に必要なヒット数が得られなかった場合は，結果の信頼性が低いと考え，(2) に戻り少し緩い条件の検索クエリを生成する．これにより広範な範囲のフレーズを用いて検索を行うことが可能となる．

（4） **冠詞誤り判定**：得られたヒット数から，冠詞誤りを判定する．

この流れを図式化すると図 1 のようになる．ヒット数取得後の判定には，閾値を用いている．

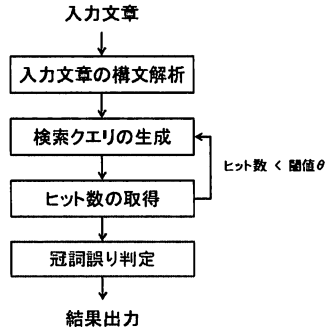

図 1 システムの概要

提案手法においては，文の先頭に出現する名詞については誤り判定を行わない．これは，文の先頭の単語では，フレーズ検索を用いた場合，無冠詞のヒット数が正確に取得できないからである．例えば，無冠詞単数形で検索すると，ヒット数の中には定冠詞単数形や不定冠詞単数形の結果も含まれてしまう．この数値を用いて判定を行うことはできない．この問題は，今後解決していく予定である．

また提案手法は，Lapata らの手法と同様，語順を保った検索を実現するためにフレーズ検索を用いている．「"」(ダブルコーテーション) で目的のフレーズを囲んで検索を行うと，その順序の並びのものだけを検索することができる．またダブルコーテーションでくくられた中で，OR 演算子を用いて類似フレーズの検索を実現している．

以下，システムの各部分について詳しく述べる．

3.2 入力文章の解析

入力された文章を解析するのには，英文構文解析器 Apple Pie Parser [12] を用いた．この構文解析器は名詞句の抽出に優れているという特徴を持つため，本手法に適していると考えられる．以下に英単語帳 DUO [11] の文を用いた解析結果の例を示す．

表 1 解 析 例

[例文] Medical breakthroughs have brought about great benefits for humanity as a whole.
[解析結果 1] (S (NPL Medical breakthroughs) (VP have (VP brought (NPL about great benefits) (PP for (NP (NPL humanity) (PP as (NPL a whole)))))) -PERIOD-)
[解析結果 2] Medical/NNP breakthroughs/NNS have/VBP brought/VBN about/RB great/JJ benefits/NNS for/IN humanity/NN as/IN a/DT whole/NN -PERIOD-/.

表 1 の解析結果 1 は，係り受け構造の解析結果を表し，解析結果 2 は，品詞情報の解析結果を表す．この解析結果を利用して 3.3 に示す手順でクエリを生成する．

提案手法では，前置詞情報を用いた手法 [3] と同様に冠詞クラス (表 2) 毎の利用頻度を基に冠詞誤りの判定を行う．そのため冠詞誤りを判定するターゲット名詞から 6 パターンの検索クエリを生成することが必要となる．

表 2 冠詞クラス [3]

	a/an	the	ϕ
単数	L_1	L_2	L_3
複数	L_4	L_5	L_6

3.3 検索クエリの生成

検索クエリ生成手順を以下に示す．

（1） 冠詞誤り判定の対象となる文から名詞句及び当該名詞句を含む動詞句を抽出する．

（2） 名詞句を単数形・複数形へと変化させ，各々3 種類の冠詞を付与することで 6 パターンの冠詞付き名詞を生成する．

（3） (1) の動詞句内の名詞句を (2) で示したバリエーションにより変化させたフレーズ群を検索クエリとする．

以下，詳細に説明する．まず，冠詞誤り判定の対象となる文に対し Apple Pie Parser を適用させ，表 1 の解析結果 1 に示すように，名詞句 (NPL) を抽出する．これを基本名詞句とする．また，冠詞誤り判定の対象となる名詞句を含む長い句である動詞句 (VP) を抽出する．この動詞句を基本フレーズとする．次に，基本フレーズを用いて検索エンジンにより検索を行う．冠詞誤り判定に必要なヒット数が得られなかった場合は，基本フレーズ内の名詞句を (2) で述べた方法により変更する．基本フレーズ内に複数の名詞句が含まれる場合は，各名詞句に対して (2) で述べた 6 パターンを生成する．すなわち，2 つの名詞句が基本フレーズ中に出現した場合は，6×6 の 36 パターンでの検索を行う．実際のクエリ生成にあたっては，表 3 に示すように OR 演算子を用いて全てのパターンを網羅するようなクエリを生成する．表 3 は，表 1 の解析結果の例で単語 whole の無冠詞単数形に対する検索クエリ群を示している．

表 3　検索クエリの例

"(benefit OR benefits) for (humanity OR humanities) as whole"
"(benefit OR benefits) for (a OR the) (humanity OR humanities) as whole"

3.4 検索の実行

生成された検索クエリ群を用いて検索を実行し，各検索クエリに対するヒット数を比較することにより，冠詞誤りの判定を行う．検索の実行と冠詞誤り判定は，以下に示す (1) ～ (3) の手順で行う．

(1) 3.3 節で生成した検索クエリを用いて，検索を実行する．

(2) ヒット数が閾値を下回った場合は，条件を緩めた検索クエリを用い再検索する．

(3) (2) においても，ヒット数が閾値下回った場合は，動詞句内の単語を 1 つ削り，(1) から繰り返す．

表 2 で示した 6 つの冠詞クラスに対する検索ヒット数の総和が，少ない場合には統計としての情報の信頼性が落ちる．そのため閾値 θ を用いて判定を行う．ヒット数の総和が閾値を上回った場合は，検索を終了し，誤り判定を行う．一方，ヒット数の総和が閾値より少ない場合は，当該結果を用いない．以下，総ヒット数が閾値以下の場合の処理について説明する．

(a) 表 2 で示した 6 つの冠詞クラスに対する検索ヒット数の総和が閾値 θ を下回った場合は，基本フレーズ内の動詞の変化を認める．これは例えば，動詞が repeat の場合には，(repeat OR repeated OR repeats) のように単純規則により変更を行う．

(b) 基本フレーズ内の動詞の変化を認めた検索クエリを用いても，ヒット数が閾値 θ を下回った場合には，基本フレーズ内から副詞を除去する．

(c) 副詞を除去してもヒット数が閾値 θ を下回る場合は，基本フレーズ内の複合名詞の先頭から名詞を除去する．

(d) 名詞句を除去してもヒット数が閾値 θ を下回った場合は，基本フレーズ内の単語を 1 つ削る．削る単語は，フレーズのターゲット名詞の後ろに単語があれば，1 番最後の単語を削り，ターゲット名詞句がフレーズ中で一番後ろの単語であれば，一番前の単語を削る．

3.5 冠詞誤り判定

3.4 節の手順に従い検索を実行し，閾値 θ を超えるヒット数を得る．表 3 の 6 つの冠詞クラス毎に得られたヒット数を基に，冠詞誤り判定を行う．具体的には，もっとも検索結果の多かった冠詞クラスを正解とする．また，可算名詞/不可算名詞の判定も行う．不定冠詞単数形，定冠詞複数形，無冠詞複数形は可算名詞の場合にしか使用できない．無冠詞単数形は不可算名詞の場合にしか使用できない．よって冠詞クラス表 2 を用いて，$L_1 + L_5 + L_6$ と L_3 を比較することで判定ができる．$L_1 + L_5 + L_6$ が多ければ可算名詞と判定し，L_3 が多ければ不可算名詞と判定する．

4. 評価実験

本節では，提案した手法と従来手法との比較実験を行う．

4.1 評価尺度

冠詞誤り検出で用いられている評価尺度について述べる．冠詞誤りの研究では主に以下の 3 つの尺度が用いられている．

$$Recall = \frac{\text{正しく検出された誤りの数}}{\text{実際の誤りの数}} \quad (1)$$

$$Precision = \frac{\text{正しく検出された誤りの数}}{\text{検出された誤りの数}} \quad (2)$$

$$F - measure = \frac{2PR}{P+R} \quad (3)$$

検出率 (Recall) は，実験対象の全誤りのうち何割を検出できたかを評価する．精度 (Precision) は，誤りとして検出されたもののうち何割が実際に誤りであったかを評価する．第 3 の尺度，F 値 (F-measure) では，検出率を R とし，精度を P として両方を考慮して評価する．検出率と精度は互いにトレードオフの関係にあるため，両方を統合した尺度として用いられる．

4.2 用いる API

今回の評価実験では，検索エンジンの API として Yahoo! JAPAN WEB API [13] を用いる．これは，検索の応答速度と 1 日のクエリの検索制限回数において他の API [14] [15] との比較を行うと，Yahoo! JAPAN WEB API が優れているからである．

4.3 比較対象

比較対象として，コーパスベース手法と検索エンジンベース手法の 2 つを用いる．コーパスベースによる手法の比較対象として，可算／不可算の情報を用いる永田らの手法 [4] を用いた．これは，ウェブ上で冠詞誤り検出システムとして公開されている [16]．このシステムでは，可算不可算のタグをつけることを目的とし，正解の冠詞と単数・複数を出力しない．そのため本手法の可算/不可算の判定部分とのみ比較を行う．

検索エンジンベースによる手法の比較対象として，フレーズ検索で判定を行う Lapata らの手法 [7] を用いる．Lapata らの手法は，冠詞を変化させた 3 パターンしか判定を行っていな

かったが，単数形・複数形も変化できるようにし6パターンに対応できるように変更した．

4.4 テストデータ

テストデータとして，ネイティブの英語の専門家によってチェックされたとしている英単語帳 DUO [11] の例文を用いる．DUO から，100 の例文を抜き出し，これを正解データセットとする．この文章の中の 197 個の名詞を判定対象とする．正解データセットの中に含まれる名詞のうち，100 個の名詞の冠詞と単数・複数をランダムに誤りに変化させたものをエラーデータセットとする．エラーデータセットを，実験の入力文章として用いる．提案手法と従来手法によって，エラーデータセットから冠詞誤りを検出し，3つの評価尺度によって評価を行う．

4.5 コーパスベース手法との比較結果

本実験では，正解テストデータを入力して可算名詞/不可算名詞であるか判定する実験を行う．対象とするのは，197 個の名詞である．永田らの手法と比較実験を行った結果を，表4に示す．以下，不正解となった単語について考察する．従来手法で不

表 4 可算/不可算タグ付与の結果

手法	正解数	正答率
永田らの手法	166	0.84
提案手法	189	0.95

正解となった 31 個の単語の内，13 個の単語は判定ルールが存在しなかった．ルールが存在しない単語は，fable，metropolis，curfew など出現頻度が稀であると考えられる単語であった．残りの 18 個においても，ルールにおいて文脈情報を利用できず，単語が可算と不可算どちらが多いかというデフォルトルールを用いたものが多く見られた．これもまた，コーパス内での出現回数の不足が原因であると考えられる．一方，提案手法では，文脈を利用しているため高い性能で判定できることが確認できた．8 個の誤りのうち 2 個は永田らの手法と共通した誤りであった．提案手法で判定できなかった 8 個の単語は，destiny や luxury など極めて曖昧性の高いどちらの意味でも使える単語であると考えられる．これらの可算・不可算を判定するのは極めて困難であると考えられる．

4.6 検索エンジンベース手法との比較結果

本実験では，冠詞誤りを含む文を入力して正しい冠詞を判定できるか実験を行う．対象文章中には，100 個のエラー単語と 97 個の正しい単語が含まれている．ヒット数の閾値として，$\theta = 100$ を用いた．エラー単語を正しい結果を検出したものを正解検出とし，エラー単語及び正しい単語を誤った結果を検出したものを誤検出とする．Lapata らの手法と比較実験を行った結果を，表5に示す．以下，提案手法で検出できなかった単語及び誤検出してしまった単語について考察する．

本実験では，冠詞誤りを含む文を入力して正しい冠詞を判定できるか実験を行う．対象文章中には，100 個の冠詞誤りと 97 個の正しい冠詞が含まれている．ヒット数の閾値として，$\theta = 100$ を用いた．冠詞誤りを正しく検出したものを正解検出とし，冠詞誤りを正しい冠詞として検出したものを誤検出とする．Lapata らの手法と比較実験を行った結果を，表5，表6に示す．また誤りを分類した結果のまとめを表7に示す．以下，提案手法で検出できなかった単語及び誤検出してしまった単語について考察する．

表 5 冠詞誤り検出の結果

手法	冠詞誤り		正しい冠詞	
	正解検出	誤検出	正解検出	誤検出
Lapata らの手法 [7]	82	18	78	19
提案手法	86	14	80	17

表 6 冠詞誤り検出の性能評価

手法	検出率	精度	F-measure
Lapata らの手法 [7]	0.82	0.70	0.76
提案手法	0.86	0.74	0.80

表 7 冠詞誤りの分類

	Lapata らの手法 [7]	提案手法
総誤検出数	37	31
共通する誤検出数	20	

提案手法では，冠詞誤りを正解検出できなかった 14 個と正しい冠詞を誤検出してしまった 17 個の計 31 個が正しく判定できなかった．Lapata らの手法では，冠詞誤りを冠詞誤りとして検出できなかった 18 個と正しい冠詞を誤検出してしまった 19 個の計 37 個が正しく判定できなかった．それぞれの単語を比較すると，両手法に共通する単語が 20 個，提案手法のみに含まれる単語が 11 個，Lapata らの手法のみに含まれる単語が 17 個であった．

共通する単語の特徴として，20 個のうちの 10 個が不定冠詞単数形を正しく判定できなかった．不定冠詞は限定的な用法であり，使用頻度も定冠詞と比較すると少ない．これに対しては，最終的にはユーザの意思によるところが大きいので，第2候補，第3候補を正解候補として提示することで解決できる可能性がある．残りの 10 個の単語に対しては，コーパスとの比較実験にも存在した曖昧性の高い単語が 6 個，定冠詞の不足・余剰が 4 個であった．

提案手法で判定できなかった 11 個の単語は，全て形容詞を取ったことに起因するものであった．テストデータは有名な英単語帳から作成したデータを用いているためウェブページ上にも少なからず該当フレーズが存在している．そのため，極めて特徴的なフレーズでも，来手法では当該フレーズがそのまま引用されているため検出できたが，一般化のため形容詞を取ったフレーズでは検出できなかったという場合があった．これは，普通の日本人によって英作文で書かれた文章については，起こりにくい誤りである．しかし，著名な文章が多く引用され，

ウェブ上に存在しているということも考慮する必要があると考えられる．

従来手法で判定できなかった冠詞誤りのほとんどは，変化しないフレーズによる特徴量の不足が原因と考えられるものであった．従来手法では，フレーズを変化させないためヒット数が少なく，数件程度又はヒットしないものも存在した．閾値を用いずに少ないヒット数で判定を行っているため，提案手法と比較すると例外的なものを正解としている例が確認できた．これらの誤りを提案手法では改善できたため，有効な手法であると考えられる．

4.7 論文を用いた実験

英単語帳だけではなく，実際の論文を用いても実験を行った．用いたデータは，ネイティブによる校正を受けた実際の投稿論文から抽出した文である，当該論文は，データマイニングに関連するものであり、テストデータ作成にあたっては，当該論文中からネイティブによる冠詞誤りを指摘された文を含む文を抽出した．ネイティブによる校正後のデータを正解データとし，校正前の日本人英語学習者が書いた文を実験を行うデータとする．テストデータには 72 個の単語が含まれ，その中の 23 個の単語は冠詞誤りを含んでいる．実験結果を表 8，表 9 に示す。

表 8　論文データを用いた結果

手法	冠詞誤り		正しい冠詞	
	正解検出	誤検出	正解検出	誤検出
Lapata らの手法 [7]	14	9	31	18
提案手法	18	5	38	11

表 9　論文データを用いた性能評価

手法	検出率	精度	F-measure
Lapata らの手法 [7]	0.60	0.43	0.50
提案手法	0.78	0.62	0.69

前の実験と比較して，どちらの手法でも F-measure が低い．これは，やはり論文に現れる単語やフレーズは，英単語帳と比較して使用頻度が低い単語が多いからであることが原因だと考えられる．

本実験で，顕著に従来手法と提案手法の差が開いているのは，論文によく見られる複合名詞で頻度の低い語のためだと考えられる．使用頻度の低い語と前の 2 単語を組み合わせた場合，検索結果が全て 0 になると考えられる．このような判定不能となったものが Lapata らの手法では 15 個存在したのに対し，提案手法では，4 個しかなかった．この点から，提案手法が頻度の少ない複合名詞に対して有効に検出することができることを確認できた．

4.8 実験のまとめと考察

本節では，提案手法と従来手法を同じテストデータを用いて比較する評価実験を行った．比較対象として，コーパスベース手法とウェブベース手法の従来手法を用いた．実験の結果，従来手法よりも検出率・精度ともに性能が高くなり，有効な手法であることが確認することができた．実際の論文データについてもテストして調べた結果，提案手法がより有効であることが確認できた．また，第 2 候補，第 3 候補の単語を正解候補に入れることでさらにシステムが改善される可能性があることが分かった．

5. おわりに

本論文では，検索エンジンの制約のある中で，検索クエリを変化させることで，類似フレーズを用いた冠詞誤り検出手法を提案した．提案手法と従来手法との比較実験を行った結果，多くの特徴を利用して判定を行うことができた．実験の結果，検出率・精度ともに向上した．誤り検出の性能を評価する F-measure が一般的な文章で 0.04 ポイント，技術的な文章で 0.19 ポイント向上することを確認した．以下，今後の課題について列挙する．

（1） **先頭名詞の判定**

提案手法では，先頭名詞を判定対象外としているため，先頭名詞の判定手法が必要である．

（2） **第 2 候補，第 3 候補の検討**

曖昧性の高い単語において，第 2 候補，第 3 候補を検討することで利便性の向上が可能である．

文　献

[1] 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,” 情処論, Vol.25, No.6, pp.1072-1079, 1984.

[2] 和泉恵美, 齋賀豊美, Thepchai Supnithi, 内元清貴, 井佐原均, “エラータグ付き日本人英語学習者発話コーパスを用いた学習者の冠詞習得傾向の分析,” 言語処理学会第 9 回年次大会発表論文集, pp.19-22, 2003.

[3] 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, 2005.

[4] 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, 2006.

[5] 乙武北斗, 荒木健治, “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,” 情報研報, 2006-NL-171, pp.25-30, 2006.

[6] J.Lee, “Automatic article restoration,” In Proc. of the human language Technology Conf. of the North American Chapter of the Association for Computational Lingstics, pp.31-36, 2004.

[7] M.Lapata, and F.Keller, “Web-based models for natural language processing,” ACM Trans. Speech and Language Processing, Vol.2, No.1, pp.1-31, Feb. 2005.

[8] British National Corpus, http://www.natcorp.ox.ac.uk/

[9] 安藤進, “Google に聞け!英語の疑問を瞬時に解決,” 丸善, 2004.

[10] P.Nakov, and M.Hearst, “A Study of Using Search Engine Page Hits as a Proxy for n-gram Frequencies,” In Proc. of the RANLP05, 2005.

[11] 鈴木陽一, “DUO3.0,” アイシーピー出版, 2000.

[12] Apple Pie Parser, http://nlp.cs.nyu.edu/app/

[13] Yahoo!デベロッパーネットワーク, http://developer.yahoo.co.jp/

[14] Google SOAP Search API, http://code.google.com/apis/soapsearch/

[15] Developer, http://search.msn.com/developer/default.aspx

[16] 可算／不可算の判定に基づいた英文誤り検出システム, http://www.ai.info.mie-u.ac.jp/ nagata/mc/