検索エンジンを用いた英文冠詞誤りの検出

Detecting Article Errors in English using Search Engines

平野 孝佳* 平手 勇宇† 山名 早人‡
Takayoshi HIRANO Yu HIRATE
Hayato YAMANA

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，専門的な単語を含む稀なパターンは正確に検出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，稀な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，F-measure において一般的な文章で 0.04 ポイント，技術的な文章で 0.19 ポイント高い性能で誤りを検出できることを確認した．

Recently, both the necessity and opportunities to write English have become higher and higher among non-native English speakers. However, 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. Lapata et al. proposed a method for detecting article errors based on search engines. However, since Lapata's method uses simple phase, it cannot detect article errors correctly when we apply to phases that contain technical 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.

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実現が望まれている．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]．

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．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[3]を拡張させたものである．提案手法では，従来手法で用いていた単純な隣接する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，より多くの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，冠詞を決定する特徴を従来手法より多く用いることができると考えられる．

本論文は，以下の構成をとる．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冠詞誤りの検出手法について述べる．第4節では，従来手法との比較実験を行い，評価する．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．冠詞の用法は，それぞれの名詞によって異なるため，大規模なテキストデータからルールを自動生成する手法が広く用いられている．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[2]

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．この手法では，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，ルール数の不足が上げられている．これは，コーパス自体の規模を大きくすることが困難であるが故に，出現頻度が一般的に低い単語を含むルールを生成できないためである．したがって，コーパスベースの冠詞誤り検出では，複雑な言語モデルのルールを全て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一つの方法として，コーパスを拡大し文章量を増やすことが挙げられる．

2.2 検索エンジンを利用した冠詞誤り検出[3]

検索エンジンを利用した冠詞誤り検出は，API を用いてフレーズ検索のヒット数を元に判定を行う．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，検出手法には工夫が必要である．

Lapata らによって，2005 年に提案された手法[3]では，隣接するＮ単語からなる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，ϕ は無冠詞を表す．本手法は，3 つのクエリによる検索結果数を比較し，最も多かったものを正解とする．フレーズ生成にあたっては，名詞句と冠詞とその前の 2 つの単語を用いることにより最も高い F-measure 値が得られることを確認している．本手法は単純なものであるが，フレーズ検索によっても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．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．

* 学生会員 早稲田大学大学院基幹理工学研究科
hirano@yama.info.waseda.ac.jp
† 学生会員 早稲田大学大学院理工学研究科・早稲田大学メディアネットワークセンター
hirate@yama.info.waseda.ac.jp
‡ 正会員 早稲田大学理工学術院，国立情報学研究所
yamana@waseda.jp

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 以降で提案手法の各ステップについて詳しく述べる．

3.1 提案システムの概要

Lapata らの手法[3]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，Web 上に多く出現す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，フレーズでの検索結果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．ヒット数が少なくなると，例外的なフレーズでも正解となってしまうことが考えられる．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 演算子によりつないだ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) に戻り基本フレーズに近い類似フレーズを用いた検索クエリを生成する．この方法は，3.5 で詳細に説明する．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 θ を用いている．

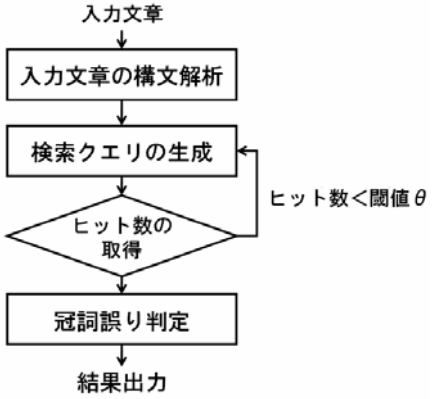

図 1 システムの概要
Fig.1 Overview of the system

なお，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の対象としない．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．

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 [4] を用いた．表 1 に英単語帳 DUO[5] の文を用いた解析結果の例を示す．

表 1 構文解析の例
Table 1 An Example of Syntactic Parsing

[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 中の解析結果 1 は，構文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，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
Table 2 Article class

	a/an	the	ϕ
単数	L1	L2	L3
複数	L4	L5	L6

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．

まず，冠詞誤り判定の対象となる文に対し構文解析を適用させ，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，基本フレーズ内の名詞句を冠詞クラスの 6 パターンに変更する．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 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．

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，3.3 で生成した全パターンのフレーズによる検索結果数の総和が閾値 θ を超えるまで類似フレーズを生成し検索を実行する．これは，検索結果数の総和が少ない場合には冠詞誤り検出の信頼性が落ちるからである．

検索の実行は，以下に示す（1）～（4）の手順で行う．

（1）3.3 節で生成した検索クエリを用いて，検索する．

（2）検索結果数の総和が閾値を下回った場合は，3.5 に示すように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，検索結果数の総和が閾値を下回った場合は，基本フレーズ内の単語を 1 つ削った上で，（1）からやり直す．削る単語は，フレーズ内の冠詞誤り判定の対象となる名詞句の後ろに単語があれば，フレーズ内の最後の単語を削る．冠詞誤り判定の対象となる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．

（4）検索結果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，(2)(3)を適用しても閾値に達しなかった場合は，当該結果を用いない．

3.5 類似フレーズの生成

3.4 の(2)で適用する類似フレーズの生成手法について詳細に説明する．

（a）表 2 で示した 6 つの冠詞クラスに対する検索結果数の総和が閾値 θ を下回った場合は，基本フレーズ内の動詞の活用変化を認める．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，検索結果数の総和が閾値 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．

3.6 冠詞誤り判定

表 2 の 6 つの冠詞クラス毎に得られた検索結果数を基に，冠詞誤り判定を行う．具体的には，最も検索結果数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，L1 +L5 +L6 と L3 を比較することで判定ができると考えられる．L1 + L5 + L6 が多ければ可算名詞と判定し，L3 が多ければ不可算名詞と判定する．

4. 評価実験

本節では，本提案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る評価尺度について述べる．以下の 3 つの尺度を用いて評価する．

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

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

$$F - measure = \frac{2PR}{P+R}$$

4.2 利用する検索エンジン API

今回の評価実験では，検索エンジンとして，Yahoo! 検索 Web サービス [6] を用いた．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比較を行うと，Yahoo! JAPAN WEB API が優れているからである．また，この実験ではヒット数の閾値として，θ = 100 を用いた．

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[2]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[7]．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，フレーズ検索のみで判定を行う Lapata らの手法[3] 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，英単語帳データと論文校正データを用いる．英単語帳データは，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[5] 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．また論文データは，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る．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エラーデータセットとする．エラーデータセットには 72 個の名詞が含まれ，その中の 23 個の名詞は冠詞誤りを含んでいる．提案手法と従来手法によって，エラーデータセットを入力文章とし，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 3 に示す．

表 3 コーパスベース手法との性能比較
Table 3 Comparison between Proposed Method and Corpus-based Method [2]

手法	正解数	正解率
永田らの手法[2]	166	0.84
提案手法	189	0.95

以下，不正解となった単語について考察する．従来手法で不正解となった 31 個の単語の内，ほとんどが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．ルールが存在しない単語は，fable，metropolis，curfew など出現頻度が低いと考えられる単語であった．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複数の意味を持つ単語であった．

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．Lapata らの手法と比較実験を行った結果を，表 4，表 5 に示す．表 5 に示されるように F-measure 値で 0.04 ポイントの向上が確認できた．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．

両手法の誤りの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，使用頻度も少ない．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，従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稀なヒット数で検出できたが，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．

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が原因と考えられる．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．

表 4 冠詞誤り検出の結果(英単語帳データ)
Table 4 Experimental Results of Detecting Article Errors (English Word Data)

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[3]	82	18	78	19
提案手法	86	14	80	17

表 5 冠詞誤り検出の性能比較(英単語帳データ)
Table 5 Comparison between Proposed Method and Web-based Method [3] (English Word Data)

手法	Recall	Precision	F-measure
Lapata らの手法[3]	0.82	0.70	0.76
提案手法	0.86	0.74	0.80

4.7 論文データを用いた実験

提案手法を用いて，実際の論文をテストデータとした実験も行った．実験結果を表 6, 表 7 に示す．表 7 に示すように，Lapata らの手法に比較し F-measure 値で 0.19 ポイントの向上が確認できた．

表 6 冠詞誤り検出の結果(論文校正データ)
Table 6 Experimental Results of Detecting Article Errors (Technical Text Data)

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[3]	14	9	31	18
提案手法	18	5	38	11

表 7 冠詞誤り検出の性能評価(論文校正データ)
Table 7 Comparison between Proposed Method and Web-based Method [3] (Technical Text Data)

手法	Recall	Precision	F-measure
Lapata らの手法[3]	0.60	0.43	0.50
提案手法	0.78	0.62	0.69

4.6 の実験と比較して，どちらの手法でも F-measure 値が低い．これは，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であると考えられる．

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 個であった．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検索エンジン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．

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．今後の課題としては，コーパスベースとの融合や英作文支援システムへの実装などが挙げられる．

[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] 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.
[3] 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.
[4] Apple Pie Parser, http://nlp.cs.nyu.edu/app/
[5] 鈴木陽一, “DUO3.0,” アイシーピー出版, 2000.
[6]Yahoo! デベロッパーネットワーク, http://developer.yahoo.co.jp/
[7] 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/

平野 孝佳 Takayoshi HIRANO
早稲田大学大学院基幹理工学研究科修士課程在学中．情報処理学会，日本データベース学会学生会員．

平手 勇宇 Yu HIRATE
2005 早稲田大学大学院理工学研究科修士課程修了．同大学同研究科博士課程在学中．2006 年より同大学メディアネットワークセンター助手．ACM，情報処理学会，日本データベース学会学生会員．

山名 早人 Hayato YAMANA
1993 早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．1989-1993 同大学情報科学研究教育センター助手．1993-2000 電子技術総合研究所．2000 早稲田大学理工学部助教授．2005 同大学理工学術院教授，現在に至る．IEEE, ACM，IEICE，IPSJ，日本データベース学会各会員．