検索エンジンを用いた英文コロケーション誤りの検出と修正

谷本太郁由† 太田 学††

† 岡山大学工学部 〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 岡山大学大学院自然科学研究科 〒 700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †, ††{tanimotot,ohta}@de.cs.okayama-u.ac.jp

あらまし 英語を母語としない者の書く英文には不適切，あるいは不自然な表現がしばしば見られる．しかし，ネイティブユーザでなければ，その誤りに気づくことも，修正することも難しい．そこで本稿では，語の慣用的なつながりであるコロケーションの誤りの検出と修正を，検索エンジンを用いて行う手法を提案する．実験では，検索結果数を用いコロケーション誤りを検出した．また，コロケーション誤りが検出された場合，再検索を行い検索結果のサマリ中から修正候補を選別してユーザに提示し，その修正精度を評価した．
キーワード 検索エンジン，英作文，コロケーション，誤り検出，誤り修正

Detection and Correction of English Collocation Errors Using a Search Engine

Takayoshi TANIMOTO† and Manabu OHTA††

† Faculty of Engineering, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
†† Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama-shi, Okayama, 700–8530 Japan
E-mail: †, ††{tanimotot,ohta}@de.cs.okayama-u.ac.jp

Abstract Non-native English users often use inappropriate or unnatural expressions in English composition. However, it would be difficult for a non-native English user to notice and correct such errors. This paper, therefore, proposes a method to support detection and correction of English collocation errors using a search engine, where collocation means habitual co-occurrence of words. In experiment, the proposed method detected collocation errors using the number of search results. Moreover, in the case of detecting collocation errors, it suggested more appropriate word candidates found in summaries of the search results. We evaluated accuracy of detection and correction of the collocation errors.
Key words Search Engine, English Composition, Collocation, Error Detection, Error Correction

1. はじめに

英語を母語としない日本人が作成した英文に多く見られる誤りの中に，コロケーションの誤りがある．コロケーションとは文や句における，2 つ以上の単語の慣用的なつながりのことである．例えば,「夢を見る」の逐語訳は “see a dream” と思うかもしれないが，“have a dream” とするのが普通である．英文にこのようなコロケーションの誤りが含まれると不自然な表現となったり，意味が通じなかったりする．

このような問題に対処するにはコロケーション辞典を調べたり，母語話者に尋ねたりすることが挙げられる．しかし，コロケーション辞典を持っている人は少ないだろうし，母語話者が身近にいるとは限らない．

その他の対処法として，検索エンジンでフレーズ検索やワイルドカードを用いて検索して調べる方法がある．例えば，フレーズ検索によって自分が使いたい表現が実際に使われているか調べることができる．また，フレーズ中の自信のない単語をワイルドカードで置き換えて検索し，検索結果を調べて妥当な単語を選ぶ方法や，複数の候補があるなら各候補を含むフレーズで，候補を変えながらフレーズ検索を行い，ヒット件数からどの候補がより妥当であるか調べることもできる．しかし，適当な検索フレーズを作成するには手間がかかる．また，Web 上には誤った例文もあるので，フレーズ検索で同じフレーズが見つかったからといって，直ちに正しい表現であると判断するのは難しい場合がある．これは実際に，誤りや不自然な表現を含むフレーズでも検索結果数が 0 件になることは稀であるからである．また，膨大な量の検索結果を参照し，適当な例文を見つけることは大変な作業である．

そこで本稿では検索エンジンを利用して，英文コロケーション誤りの検出と修正を行う手法を提案する．提案手法は検討したい英文を入力すると，Web 検索によりその中に含まれるコロケーションが正しいかどうか判定し，誤りと判定すれば修正候補となる語を提示する．

2. 関 連 研 究

検索エンジンを用いた英文誤り検出や修正はさまざまな品詞に対して行われている．例えば，有冨ら [1] は英文中の前置詞誤りの自動修正を行うシステムを提案した．入力された英文の前置詞をワイルドカードに置き換え，その前後の単語から前置詞に関係が深いと思われる単語を動詞，名詞，形容詞，副詞の中から規則に従って選択し，検索フレーズを生成する．得られた検索結果のサマリからワイルドカード部分に相当する前置詞を抽出し，前置詞ごとの出現確率を求め，誤りの検出と修正を行っている．彼らは約半数の前置詞が誤っているデータに対して実験を行い，前置詞誤り検出の F 値 0.85，誤り修正精度 0.82 という性能を報告している．

大鹿ら [2] は検索エンジンを用いた英作文支援システムの一部として，英文の冠詞，前置詞，類義語などの誤り検出や妥当性の判断を支援するシステムを実装している．与えられたフレーズの検討したい箇所に対して，複数の候補を用意し，検討したい箇所をそれぞれの候補で置き換えながらフレーズ検索を行い，検索結果数を示すことによって，ユーザがフレーズの妥当性を判断することを支援する．候補を得るために，前置詞の場合はフレーズ中の前置詞をワイルドカードに置き換えて検索し，その結果から前置詞を抽出している．動詞，名詞，形容詞については辞書データベースを用いて類義語を取得している．

また，Yi ら [3] は Web 検索を用いて，冠詞，動詞＋名詞のコロケーション，形容詞＋名詞のコロケーションの修正を行っている．動詞＋名詞からなるコロケーションの修正の場合，まず，品詞タグ付けとチャンクの解析によってコロケーションを含む部分を特定する．チャンクとは意味的にまとまったいくつかの単語のことである．検索は文，チャンク，語の 3 つの異なる粒度でクエリを生成して行う．文レベルでは調べたい動詞の前後で文を分割し，3 つのフレーズを AND 検索する．チャンクレベルでは文をチャンクに分けた後のフレーズを用い，語レベルでは文中の内容語のみで AND 検索を行う．検索結果のサマリ中で注目する動詞が，コロケーションの名詞と繋がりをもっているか調べ，その頻度が閾値より小さければ動詞を除いたクエリで再度検索を行い，修正候補を求めるという動作を各レベルに対して行う．彼らは実験で，英語学習者の書いた様々な誤りを含む英文に対して，冠詞と動詞＋名詞コロケーションの誤り修正を行っている．冠詞については精度 62.5%，再現率 49.7%で修正が可能であり，動詞＋名詞コロケーションでは精度 37.3%，再現率 30.7%で修正が可能であると報告している．

3. 提 案 手 法

3.1 提案手法の概要

提案手法の簡単な処理の流れを説明する．まず，ユーザが検討したい英文を入力する．次に，入力された英文を構文解析(注1)する．その構文解析の結果を用いて，検索クエリを生成し，フレーズ検索を行う．得られた検索結果数により，注目しているコロケーションが誤りであるか判定する．誤りを検出した場合は，修正候補取得のための検索クエリを生成し，再度検索を行って，検索結果のサマリを取得する．得られたサマリより修正候補を取得し提示する．

表 1　検索クエリ生成に用いる主要な構文要素

節	タグ	句	タグ	品詞	タグ
平叙節	S	動詞句	VP	動詞	VB, VBD, VBG,
		名詞句	NP		VBN, VBP, VBZ
		形容詞句	ADJP	名詞	NN, NNS, NNP,
		副詞句	ADVP		NNPS
		前置詞句	PP	形容詞	JJ, JJR, JJS

3.2 コロケーション誤りの検出

3.2.1 検索クエリ生成

英文中に含まれるコロケーションが正しいかどうか判定するためにフレーズ検索を行う．本稿では（1）動詞＋名詞,（2）名詞＋動詞,（3）名詞＋名詞,（4）形容詞＋名詞,（5）動詞＋副詞の 5 種類のコロケーションを誤り検出の対象とし，ここではそれぞれに対する検索クエリ生成法について述べる．

検索クエリは，入力された英文の構文解析結果の中から，検討したい英文のコロケーションを見つけ，3 つ生成する．動詞＋名詞の場合を例に挙げると，この 3 つの検索クエリは動詞クエリ，名詞クエリ，それらを連結したフレーズである共起クエリからなる．これら 3 つのクエリの検索結果数より共起の強さを計り，コロケーションの正誤を判断する．

構文解析では文の節，句，語にタグ付けされた構文木が得られる．検索クエリ生成には，平叙節 S，動詞句 VP，名詞句 NP などを用いる．検索クエリ生成に用いる主要な構文要素を表 1 にまとめる．なお，タグは Penn Treebank Ⅱ tags に準拠している．また，“I had a dream last night.” という例文の構文木を図 1 に示す．

本稿で扱う各コロケーション誤り検出のための検索クエリ生成法について以下で述べる．

（1） 動詞＋名詞コロケーション

動詞＋名詞コロケーションでは，動詞とその目的語の関係から誤りを検出することを目的とする．ここでは，VP，NP，PP の 3 種類のノードに着目し，構文木から動詞＋名詞コロケーションを探索する．まず，構文木から VP ノードを探索する．発見したノードの子ノードに VB，VBD，VBG などの動詞を示すノードと，NP ノードか PP ノードがあれば動詞＋名詞コロケーションと判断し，クエリを生成をする．生成する検索クエリは，動詞クエリ，名詞クエリ，それらを連結したフレーズからなる共起クエリの 3 つである．

動詞クエリは動詞と，その動詞に続く前置詞，不変化詞の 3 つの品詞からなる．動詞は VP ノードの子ノードが持つ動詞を

(注1)：Stanford Parser [4] の version 1.6.5 を使用した．
http://nlp.stanford.edu/software/lex-parser.shtml

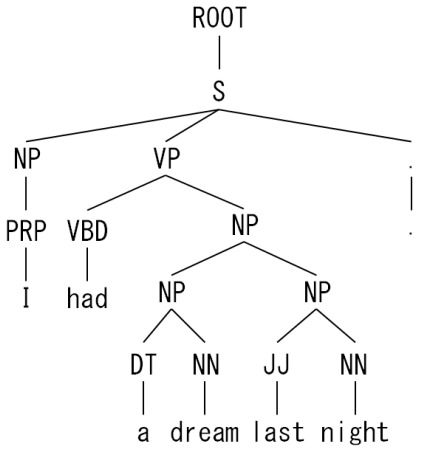

図 1　“I had a dream last night.” の構文木

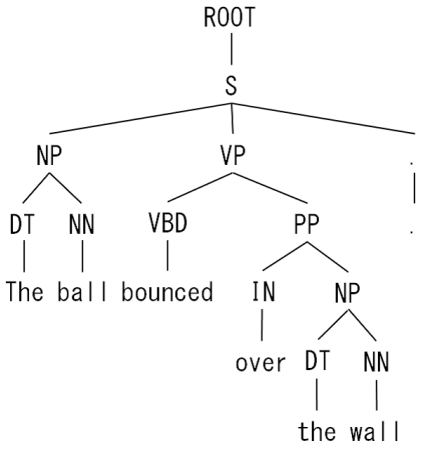

図 2　“The ball bounced over the wall.” の構文木

用いる．図 1 の例では “had” が動詞クエリとなる．このとき，VP ノードの子ノードに不変化詞を示すノードがあれば，それも動詞クエリに含める．さらに，子ノードに PP ノードがある場合は，そのノードの最も左側の子が前置詞を持っているので，それもあわせて動詞クエリとする．例えば，図 2 の “The ball bounced over the wall.” という例文では，PP ノードの最左の子ノードは “over” を持つので，動詞クエリは “bounced over” になる．

名詞クエリは，NP ノード，あるいは，PP ノードの子ノードの NP ノードの語句とする．名詞クエリは，NP ノード下にある語なら全ての品詞が含まれる可能性があるが，必ず 1 語以上の内容語を含むものとする．まず，子ノードに VB, JJ, NN などの品詞を示すノードを持つ NP ノードを探す．NP ノードの子に品詞ノードがなく，NP ノードがあるようであれば，NP ノードを深さ優先探索し，品詞ノードを持つ NP ノードを見つけ，その下の全ての語を名詞クエリとする．ただし，名詞クエリを得ようとする NP ノードの子ノードに，NP ノードと ADJP ノード以外の節や句を表すノードがある場合は，その節，句ノードより下の単語の獲得は行わない．図 1 の例では，“a dream”，図 2 の例では，“the wall” が名詞クエリとなる．

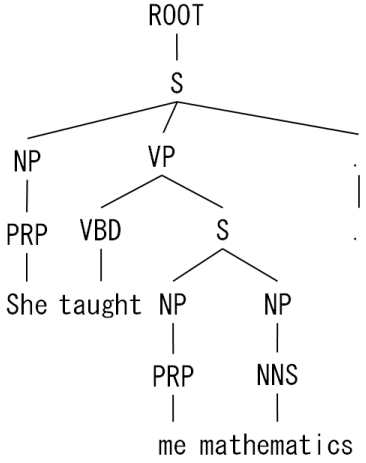

図 3　“She taught me mathematics.” の構文木

また，与えられた文が第 4 文型（SVOO 型）のとき，VP ノードの子ノードが S ノードである場合がある．例えば，“She taught me mathematics.” という文があるが，これを構文解析すると図 3 のようになる．このような場合は S ノードの子ノードまで調べ，子ノードが全て NP であれば深さ優先探索により，子に品詞ノードを持つ NP ノードを探索し，クエリを生成する．この例から生成される動詞クエリは “taught”，名詞クエリは “mathematics” となる．ここで，代名詞の “me” が例文中に出現しているが，代名詞のみからなるクエリは生成しない．

（2） 名詞＋動詞コロケーション

名詞＋動詞コロケーションでは，主語と動詞の関係から誤りを検出することを目的とする．まず，構文木より S ノードを探す．その S ノードが NP と VP を子ノードとして持っており，なおかつ，NP ノードが VP ノードより左側にある場合，名詞＋動詞コロケーションと判断する．NP ノードの下から得られる名詞クエリは，動詞＋名詞コロケーションの名詞クエリと同じ方法で取得する．図 2 の例では “The ball” が名詞クエリとなる．

一方，動詞クエリは動詞，前置詞，不変化詞に加え，助動詞，副詞も含む．動詞＋名詞コロケーションの場合と同様に動詞は必須である．これら 5 つの品詞が VP ノードの子にある場合は，それらを全て動詞クエリに含める．また，対象としている VP ノードの子に VP ノードが現れる場合はそのノードからも語句を獲得し，動詞クエリに含める．そのため，検討する英文が受動態の場合は，be 動詞に加えその直後の一般動詞もクエリに含める．助動詞などが含まれる場合の，VP を含む構文木の例を図 4 に示す．図 4 の例では，名詞クエリが代名詞のみになるので名詞＋動詞コロケーションとしての検索クエリを生成しないが，ここから，名詞＋動詞コロケーションの動詞クエリを生成する場合は “couldn’t believe” となる．

（3） 名詞＋名詞コロケーション

まず，NP ノードを構文木より探索する．NP ノードの子ノードに複数の名詞が連続して出現している場合，それらを，検索クエリとする．名詞 1 つがそれぞれクエリ 1 つに対応し，その共起クエリは隣接する名詞 2 語からなる．

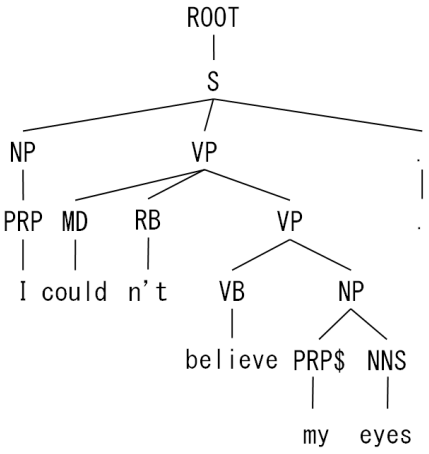

図 4　“I couldn’t believe my eyes” の構文木

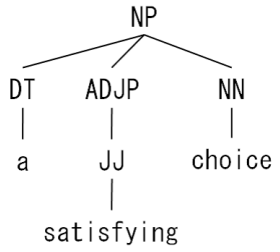

図 5　“a satisfying choice” の部分木

ここで，3 語以上の名詞が連続して出現している場合は，連続する名詞の先頭から順に 2 語ずつに対して，検索クエリを生成する．例えば，名詞 3 語からなる複合名詞 “Data Base System” について考えると，先頭の名詞と真ん中の名詞からなるコロケーション “Data Base” と，真ん中の名詞と最後の名詞からなるコロケーション “Base System” の 2 つの名詞＋名詞コロケーションを抽出し，それぞれに検索クエリを生成する．なお，3 語全体で 1 つのコロケーションとみなすことも考えられるが，本稿では扱わない．

（4） 形容詞＋名詞コロケーション

まず，名詞＋名詞コロケーションと同様に NP ノードを探す．形容詞＋名詞コロケーションは，NP ノードの子ノードに，名詞ノードの他に，形容詞を表すノードが出現している場合と，形容詞句ノード ADJP が出現している場合がある．形容詞ノードの場合，そのノードが持つ形容詞を形容詞クエリとし，NP ノードの子ノードのうち形容詞ノードより後ろの名詞ノードの名詞を名詞クエリとする．図 1 の例では “last night” が形容詞＋名詞のコロケーションとなる．また，NP ノードの下に形容詞ではなく動詞ノードがある場合がある．これは，動詞の分詞が形容詞的に用いられている場合である．構文要素としては形容詞を表す JJ の代わりに動詞の進行形を表す VBG などとされることがあるが，この場合の動詞は形容詞クエリとして生成する．ADJP ノードの場合は，そのノードの下にある全ての語を形容詞クエリとし，名詞クエリは ADJP ノードより後ろの名詞とする．図 5 に NP ノードが ADJP ノードを持つ例を示す．この例では “satisfying” が形容詞クエリ，“choice” が名詞クエリ，“satisfying choice” が共起クエリとなる．

（5） 動詞＋副詞コロケーション

まず，VP ノードを探す．VP ノードの子ノードに動詞ノードと，ADVP ノードがあれば，動詞＋副詞コロケーションと判断し，クエリを生成する．VP ノードの下の動詞を動詞クエリ，ADVP ノードの下の副詞を副詞クエリ，それらを並べたフレーズを共起クエリとする．

3.2.2　検索結果数に基づくコロケーション誤りの検出

3.2.1 項の方法で各コロケーションに対して生成したクエリを用いてフレーズ検索を行い，それぞれの検索結果数を求める．Web 検索には Yahoo!デベロッパーネットワーク(注2) で提供されている Yahoo!検索 API を利用する．得られた検索結果数より文中に含まれるそれぞれのコロケーションに対して，MI スコアと T スコアを求め，それらの値が閾値未満のとき誤りとして検出する．MI スコアと T スコアはコロケーション研究によく用いられている統計指標である [5].

MI スコアは相互情報量の考えに基づいており，ある語が共起相手の語の情報をどの程度持っているかを示す．実測値を期待値で割り対数を取ることで求められ，以下の式で定義される．

$$MI = \log_2 \frac{\text{共起頻度} \times \text{コーパス総語数}}{\text{中心語頻度} \times \text{共起語頻度}} \quad (1)$$

T スコアは，2 つの語の共起関係の統計的有意性を計る指標であり，以下の式で定義される．

$$T = \frac{\left(\text{共起頻度} - \frac{\text{中心語頻度} \times \text{共起語頻度}}{\text{コーパス総語数}}\right)}{\sqrt{\text{共起頻度}}} \quad (2)$$

ここで，共起頻度とは各コロケーションの共起クエリによる検索結果数とし，中心語頻度と共起語頻度は残りの 2 つのクエリ，例えば動詞＋名詞コロケーションでは動詞クエリ，名詞クエリによる検索結果数とする．なお，コロケーション誤りの検出では中心語と共起語を区別する必要はない．また，コーパス総語数は全 Web ページ中の総語数になるため，実験により適当な値を求めた．また，MI スコア，T スコアの閾値は，誤りを含むコロケーションと誤りの含まれないコロケーションを 50 件ずつ用い，適当な値を定めた．

3.3　コロケーション誤りの修正

コロケーション誤りが検出された場合，再検索を行い，検索結果のサマリより修正候補となる語を抽出する．抽出した語を出現回数によってランク付けし，その結果を提示する．

Web 上には膨大な英文があるので，再検索の際，適切なクエリを用いれば正しい表現を含む適切な英文を得ることが期待できる．よって，サマリ中における出現回数により語をランク付けし，適切な修正候補を提示する．

実験では，動詞，名詞，形容詞，副詞の修正を試みる．コロケーション誤りが検出された場合，どちらの語が誤りであるか判断する必要があるが，本稿では人が，どちらの語を修正対象とするか決定する．

(注2)：http://developer.yahoo.co.jp/

3.3.1 修正候補取得のための検索クエリ生成

まず，修正候補となる語を取得するための検索クエリを生成する．この検索クエリには，誤り検出に用いた共起クエリを除く 2 つのクエリのうち誤りが含まれていない方を用いる．例えば，形容詞＋名詞コロケーションから誤りを検出し，形容詞を修正する場合，名詞クエリを用いる．この誤りが含まれていないクエリを中心語として，これとコロケーションをなす語を獲得する．

これだけでは，中心語として定めたフレーズがどのような文脈で用いられているか特定できないので，文中で出現する他の語を用いて検索結果を絞り込む．絞り込みに用いる語は文中に出現する動詞，名詞，形容詞のうち，誤りとして検出されたコロケーションに含まれていないものとする．

例えば，“I went to a classic music concert.” という誤り文について考える．これは “classic” を “classical” に変更すれば正しい文となる．この文から抽出される形容詞＋名詞コロケーションは “classic music concert” であり，名詞クエリは “music concert” である．また，名詞クエリ，形容詞クエリに含まれていない動詞，名詞，形容詞は “went” となる．“I” は代名詞のため本研究では含めない．よって検索クエリは [“music concert” AND went] となる．

3.3.2 修正候補の取得

検索によって得られたサマリを解析し，修正候補となる語を取得する．まず，サマリを “.”，“!”，“?” で分割し，文を得る．この文に中心語としたフレーズが含まれていれば，構文解析する．

得られた構文木を探索し，修正候補を獲得する．構文木の探索は 3.2.1 項の誤り検出のための検索クエリ生成法と同様にして行う．例えば，形容詞＋名詞コロケーションより形容詞を修正する場合，構文木より NP ノードを探索する．見つけた NP ノードが子ノードに形容詞，動詞，形容詞句を持っていれば，その NP ノードの子ノードに名詞クエリのフレーズがあるか調べる．名詞クエリのフレーズが確認された場合は，この形容詞，動詞，形容詞句を修正候補として抽出する．

これを検索結果の全てのサマリに対して行い，抽出した語の出現回数をカウントし，出現回数順にランク付けし，修正候補として提示する．4.2 節の実験では修正対象のコロケーション毎に 400 件のサマリより修正候補となる語を抽出した．

4. 評価実験

実験では,「アメリカの子どものように英会話を覚える本 [6]」で挙げられている，コロケーション誤りを含む文 153 文とその正解例 153 文を用いる．誤りを含む文には 1 文につき一か所の誤りが含まれている．この誤りの内訳を表 2 に示す．

実験ではコロケーション誤りを含む 153 文に対して，誤り検出が行えるか，誤りのないコロケーションに対して誤検出がどのくらいあるか評価した．また，検出に成功したコロケーション誤りに対して修正候補を提示し，修正可能かどうか調べた．この際，英文校正を支援する Web サービスである NativeChecker と修正精度の比較を行った．

表 2　誤りの内訳

品詞	誤りの数
動詞	79
名詞	22
形容詞	45
副詞	7
合計	153

表 3　コロケーションの内訳

種類	正	誤	合計
動詞＋名詞	137	130	267
名詞＋動詞	20	19	39
名詞＋名詞	18	16	34
形容詞＋名詞	59	51	110
動詞＋副詞	9	10	19
合計	243	226	469

表 4　コロケーション誤りの詳細

種類	誤っている品詞	合計
動詞＋名詞	動詞: 73，名詞: 16，形容詞: 39，前置詞: 2	130
名詞＋動詞	動詞: 16，名詞: 2，形容詞 1	19
名詞＋名詞	前側の名詞: 11，後ろ側の名詞: 5	16
形容詞＋名詞	名詞: 3，形容詞: 48	51
動詞＋副詞	動詞: 3，副詞: 7	10

4.1 コロケーション誤りの検出実験

MI スコアと T スコアによるコロケーション誤りの検出では，それぞれに閾値を定め，コロケーションのスコアが閾値未満のとき誤りとする．ここで，MI スコアの閾値は 0，T スコアの閾値は 2 とした．また，式（1）,（2）のコーパス総語数は 2^{38} とした．

3.2.1 項で述べた方法によってコロケーションを抽出したところ，全 306 文から 469 のコロケーションが得られた．得られたコロケーションの内訳を表 3 にまとめる．また，そのうち誤りを含むコロケーションについてはどの品詞が誤っているか，詳細を表 4 にまとめる．

表 3 で，誤り総数が 226 となっており，表 2 の誤り総数 153 より大きくなっている．これは次のような例があるからである．“I went to a classic music concert.” という誤った文からは，動詞＋名詞コロケーションの “went to a classic music concert”，形容詞＋名詞コロケーションの “classic music concert”，名詞＋名詞コロケーションの “music concert” の 3 つのコロケーションが抽出される．この文は “classic” を “classical” に変更すれば正しい英文になるので，誤りが含まれているのは動詞＋名詞コロケーションと形容詞＋名詞コロケーションになる．このような例は表 2 では，形容詞誤り 1 件に相当するが，表 3 では動詞＋名詞と形容詞＋名詞コロケーションの両方で誤りと数えられる．また，このため表 4 の動詞＋名詞コロケーションと，名詞＋動詞コロケーションの誤っている品詞の内訳に形容詞が含まれている．

また，表 4 の動詞＋名詞コロケーションに前置詞の誤りが含まれているが，これは品詞タグ付けに失敗したためである．

表 5　コロケーション誤りの検出結果

	MI スコア			T スコア		
	検出	誤検出	非検出	検出	誤検出	非検出
動詞＋名詞	82	27	48	93	32	37
名詞＋動詞	16	7	3	16	7	3
名詞＋名詞	16	13	0	16	13	0
形容詞＋名詞	45	29	6	45	29	6
動詞＋副詞	10	6	0	10	6	0

表 6　コロケーション誤りの検出性能

	MI スコア			T スコア		
	再現率	適合率	F 値	再現率	適合率	F 値
動詞＋名詞	0.631	0.752	0.686	0.715	0.744	0.729
名詞＋動詞	0.842	0.696	0.762	0.842	0.696	0.762
名詞＋名詞	1	0.552	0.711	1	0.552	0.711
形容詞＋名詞	0.882	0.608	0.72	0.882	0.608	0.72
動詞＋副詞	1	0.625	0.769	1	0.625	0.769

よって表 3，表 4 の誤ったコロケーションの内訳には，正しいコロケーションの分類とは異なる分類がされているコロケーション誤りが含まれる．

これらに対し，MI スコア，T スコアをそれぞれ求めてコロケーションに含まれる誤りの検出を行った結果を表 5 に示す．また，その結果から再現率，適合率，F 値を求め，表 6 にまとめる．

表 6 の結果をみると動詞＋名詞コロケーション以外では MI スコアでも T スコアでも同じ検出性能となった．動詞＋名詞コロケーションで MI スコアの F 値が低い理由は，中心語か共起語の出現頻度が低いと，共起頻度が過大に評価され，式（1）のスコアが大きくなる性質があるためである．中心語と共起語の出現頻度は，動詞クエリと名詞クエリの検索結果数に対応する．ここで，名詞クエリは “classic music concert” のように形容詞＋名詞コロケーションや，名詞＋名詞コロケーションを包含する場合がある．名詞クエリ内のコロケーションが誤っている場合，名詞クエリ自体の検索結果数が少なくなり，結果として MI スコアが大きくなり，検出できないものが増えた．

本稿では行っていないが，誤りを修正するためにはコロケーション誤りを検出した後に，その中のどの語が誤っているか特定する必要がある．その一つの方法として，共通の語を含む複数のコロケーションが誤りとされた場合，その共通する語が誤りであると特定することが考えられる．例えば，動詞＋名詞コロケーションと名詞＋動詞コロケーションで誤りが検出されて，なおかつ，その動詞が同じ語であれば，動詞が誤りである可能性が高くなる．実際に，動詞が誤りである場合，両方のコロケーションで誤りを検出する例や，誤りとしなくてもスコアの値が低くなる例がいくつか見られた．しかし，動詞が誤っているとき，その動詞＋名詞コロケーションが誤っていてもその動詞を含む名詞＋動詞コロケーションが誤りであるとは限らないし，その逆もいえる．実験データには，このような共通の動詞を含む動詞＋名詞，名詞＋動詞コロケーションの例が少なかったため，データを増やすなどして確認する必要がある．また，代名詞など本稿では考慮しなかった品詞のコロケーションや，その他の組み合わせからなるコロケーションについても実験により，誤っている語の特定を試みることを考えている．

表 7　動詞誤りの修正結果

	修正件数	修正精度
1	15	23.1 %
2	18	27.7 %
3	22	33.8 %

表 8　名詞誤りの修正結果

	修正件数	修正精度
1	1	6.3 %
2	2	12.5 %
3	6	37.5 %

表 9　形容詞誤りの修正結果

	修正件数	修正精度
1	7	18.4 %
2	9	23.7 %
3	10	26.3 %

表 10　副詞誤りの修正結果

	修正件数	修正精度
1	1	14.3 %
2	2	28.6 %
3	2	28.6 %

4.2　コロケーション誤りの修正実験

ここでは，誤りを含んだ 153 文の英文から，コロケーションの誤りが検出できたものに対して修正を試みる．具体的には，コロケーションの中の誤っている動詞，名詞，形容詞，副詞の修正候補を検索によって取得して，提示する．

なお実験では，修正候補となる名詞は名詞＋名詞コロケーションのみに基づき取得する．名詞の修正候補取得に，動詞＋名詞など名詞＋名詞以外のコロケーションを用いることも考えられるが，本稿では扱わない．他の品詞については，動詞は動詞＋名詞コロケーションと名詞＋動詞コロケーション，形容詞は形容詞＋名詞コロケーション，副詞は動詞＋副詞コロケーションに基づき修正候補を取得する．

まず，誤りを含んだ 153 文のうち，コロケーションの誤り検出に成功し，なおかつ，コロケーションが修正候補取得の条件と一致するものは 126 件あった．この条件とは，動詞は動詞＋名詞か名詞＋動詞コロケーション，名詞は名詞＋名詞コロケーション，形容詞は形容詞＋名詞コロケーション，副詞は動詞＋副詞コロケーションにそれぞれ基づいて修正候補を取得することである．この 126 件の内訳は動詞 65 件，名詞 16 件，形容詞 38 件，副詞 7 件であった．

この 126 件に対し修正候補を提示し，その上位 3 件までに正解の語が含まれている件数と，その際の修正精度を表 7 から表 10 に示す．

これらの表の修正精度をみると，上位 3 件までの修正候補に正解が含まれていればよいという条件でも精度はあまりよくな

い．この原因の 1 つは，検索クエリの不備である．検索クエリが元の文の意図を十分に反映できていなかったため，適切でない候補がランキングの上位に出現していることが多かった．また，サマリ中に期待する文型があまり出現しないこともある．例えば，形容詞＋名詞コロケーションを利用して形容詞を修正する場合，修正用の検索クエリに含めた名詞がサマリ中では形容詞を伴わず用いらることが多かった．また，名詞＋動詞コロケーションに基づき動詞を修正する際に，検索語の名詞が目的語などとして用いられている場合が多かった．名詞＋動詞コロケーションにおける動詞修正では，主語と動詞の関係から動詞修正候補を取得するので，その結果，獲得できる動詞の数が少なくなった．このような理由から修正候補の取得数が減ると，正解である語とノイズの差がほとんどなくなったり，そもそも正解である語が獲得できなくなったりして修正精度の悪化につながった．

また，誤りを含んだ文を構文解析するので，品詞タグ付けに失敗することがある．3.3 節で述べたコロケーション誤りの修正方法は，構文解析結果に依存するので，品詞タグ付けに失敗すると，意図したものと異なる品詞を収集する．

このような誤りに対しては，フレーズ検索とワイルドカード検索を組み合わせ，検索条件を厳しくすることが考えられる．例えば，“I went to a classic music concert” の誤り修正には，“went to a * music concert” という検索クエリを用い，ワイルドカードに相当する部分から修正候補となる語が得られるだろう．誤っている語を特定できれば，このように検索結果を絞り込んで，より適切な修正候補を獲得することも考えられる．

4.3 NativeChecker との比較

NativeChecker(注3) とは，英語のフレーズに対して，修正を支援する Web サービスである．検討したいフレーズを与え，修正したい単語と，その語をどのように修正するかを選択すると，その語を複数の修正候補で置き換えながらフレーズ検索する．それらの検索結果数の大きさを比較することで，より一般的な語を調べることができる．例えば，NativeChecker が提供する修正内容の一つである “類義語” を用いると，修正したい単語を類義語で置き換え，フレーズ検索をする．その検索結果数を比較することで，より妥当な表現を調べることができる．

本稿の実験では，それぞれの英文から人手により最適なフレーズを作成し，NativeChecker に与えた．また，修正内容は “類義語” を選択した．それによって提示された上位 3 件以内に正解となる語が含まれていれば修正できたとみなし，修正件数と修正精度を求めた．誤りを含む 153 文からフレーズを作成し，NativeChecker に与えたところ品詞タグ付けの失敗や，類義語の獲得ができなかったものを除くと，動詞 69 件，名詞 17 件，形容詞 38 件，副詞 7 件の誤りに対して検討が可能であった．これらに対し，修正件数と修正精度を求めたものを表 11 から表 14 にまとめる．

また，提案手法と NativeChecker の修正精度を比較したものを図 6 から図 9 に示す．これらの図をみると上位 3 件までに正解が含まれていればよいという条件では，提案手法が NativeChecker の修正精度を動詞，名詞，副詞で上回っている．また，形容詞については同じ修正精度となった．

形容詞の修正精度を比較している図 8 では，提示した修正候補の上位 3 件までに正解が含まれていればよいという条件では同じ修正精度となっているが，上位 1 件，または 2 件という条件では NativeChecker の方が高い修正精度を示している．名詞の修正精度を比較している図 7 についても同様の傾向がみられる．これは，形容詞や名詞を修正する場合，正しい語を元の英文の誤った語の類義語から得ることが容易な場合があるためである．例えば，動詞を誤っている “see a dream” というフレーズの “see” から正解である “have” を推測することは難しいが，形容詞を誤っている “classic music” では，“classic” から “classical” についてはシソーラスを用いれば容易に得ることができる．このことから，品詞によっては，修正候補のランキングに語の意味の類似性を考慮することにより，修正精度を向上させることが考えられる．

表 11　NativeChecker による動詞誤りの修正結果

	修正件数	修正精度
1	11	15.9 %
2	12	17.4 %
3	12	17.4 %

表 12　NativeChecker による名詞誤りの修正結果

	修正件数	修正精度
1	3	17.6 %
2	3	17.6 %
3	3	17.6 %

表 13　NativeChecker による形容詞誤りの修正結果

	修正件数	修正精度
1	9	13.7 %
2	10	26.3 %
3	10	26.3 %

表 14　NativeChecker による副詞誤りの修正結果

	修正件数	修正精度
1	1	14.3 %
2	1	14.3 %
3	1	14.3 %

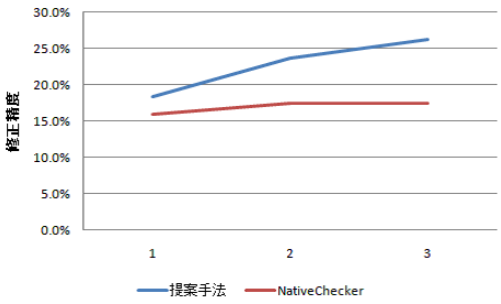

図 6　動詞修正精度の比較

(注3)：NativeChecker http://native-checker.com/

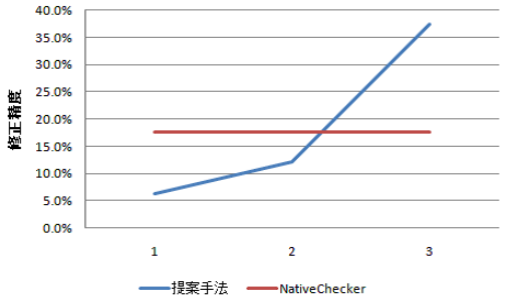

図 7　名詞修正精度の比較

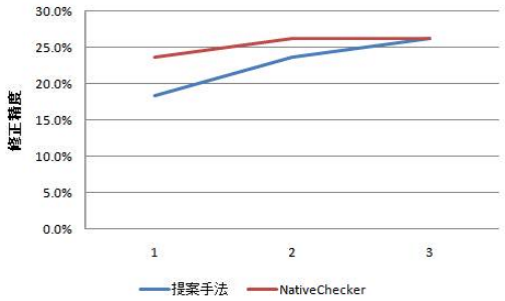

図 8　形容詞修正精度の比較

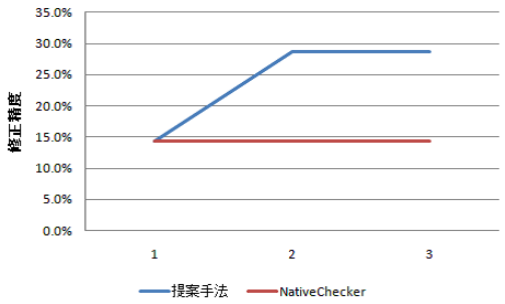

図 9　副詞修正精度の比較

5. ま と め

本稿では，検索エンジンを利用することで英語を母語としない者には判断が難しい英文コロケーションの誤りを検出し，誤りを検出した場合には，その語の修正候補を提示する手法について述べた．実験では，文中から 5 種類のコロケーションを抽出し，そのコロケーションが誤りかどうかを，MI スコア，T スコアの 2 つの統計指標を用いて検出した．また，検出したコロケーション誤りに対して，誤っている動詞，名詞，形容詞，副詞の修正候補を提示した．その結果，動詞＋名詞コロケーション誤り検出実験で得られた検出性能を示す F 値は，MI スコアで検出した場合 0.686 に対し，T スコアでは 0.729 となり，MI スコアより T スコアの方が高かった．しかし，その他のコロケーションでは MI スコア，T スコアのいずれを用いても同じ値となり，名詞＋動詞コロケーションは 0.762，名詞＋名詞コロケーションは 0.711，形容詞＋名詞コロケーションは 0.72，動詞＋副詞コロケーションは 0.769 となった．また，修正精度については提示した修正候補の上位 3 件までに正解が含まれていればよいという条件で，動詞 33.8%，名詞 37.5%，形容詞 26.3%，副詞 28.6%であった．

今後の課題として，複数のコロケーションを利用してコロケーション誤りを検出することで，誤っている語句を特定することが挙げられる．また，誤っている語句を特定することで修正候補の取得方法を改善したり，他の品詞誤りへ拡張したりすることを検討していきたい．

文　献

[1] 有冨　隼，太田　学，“検索エンジンを用いた英文前置詞誤り修正支援”，日本データベース学会論文誌，Vol. 9，No. 1，pp. 70–75，2010.
[2] 大鹿広憲，佐藤　学，安藤　進，山名早人，“Google を活用した英作文支援システムの構築”，DEWS2005，2005.
[3] Xing Yi，Jianfeng Gao，and William B. Dolan, “A Web-based English Proofing System for English as a Second Language Users”，the Proceeding of the third International Joint Conference on National Language Processing, 2008.
[4] Dan Klein, and Christopher D. Manning, “Accurate Unlexicalized Parsing”, Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430, 2003.
[5] 石川慎一郎，“言語コーパスからのコロケーション検出の手法–基礎的統計値について–”，統計数理研究所共同研究レポート，No. 190，pp. 1–14，2006.
[6] 足立恵子，“アメリカの子どものように英会話を覚える本”，中経出版，2007.