Vol.2010-DBS-151 No.36
2010/11/13

検索エンジンを用いた動詞名詞コロケーションに基づく英文動詞誤りの検出と修正

谷本　太郁由[†1]　　太田　学[†2]

英作文を行う際，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語の慣用的なつながりのことで，ネイティブスピーカーでなければ，誤りに気づくことも，修正することも難しい．本稿では，コロケーション誤りの中でも特に多い動詞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．具体的には，入力された英文より検索フレーズを生成し，検索結果数より動詞の誤りを検出する．誤りを検出した場合は，再度検索を行い，検索結果のサマリを収集し，そのサマリ中での動詞の出現回数に基づき修正候補の提示を行う．実験では，検出の F 値は 86.4% であり，修正精度は 38.6% であった．

Verb-Noun-Collocation-Based Detection and Correction of English Verb Errors Using a Search Engine

Takayoshi Tanimoto[†1] and Manabu Ohta[†2]

Japanese people often use miscollocations in English composition. Collocation means the habitual co-occurrence of words. Thus, it would be difficult for a non-native English speaker to notice and correct the errors. In particular, many miscollocations are verb-noun collocation errors. This paper, therefore, proposes a system for supporting detection and correction of verb errors based on verb-noun collocations by using a search engine. The proposed system automatically generates query phrases including a verb-noun collocation to be checked and detects verb errors using the number of the search results. In the case of detecting verb errors, our system i)searches again to collect summaries of the search result, ii)calculates occurrence probabilities of verbs, and iii)suggests correct verb candidates. The experimental results showed that F-measure was 0.864 in automatic error detection and the accuracy was 0.386 in automatic error correction.

1. はじめに

我々日本人が英語の文章を作成するとき，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,「スープを飲む」を逐語訳すると “drink soup” となるかもしれないが，実際は “eat soup” とするのが普通である．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．

対処法としてコロケーション辞典を調べたり，ネイティブスピーカーに質問したりすることがあげられる．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，ネイティブスピーカーが身近にいるとは限らない．

その他の対処法として，検索エンジンを用いてフレーズ検索やワイルドカードを用いた検索によっ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使われているか調べることができる．また，フレーズの中の自信のない単語をワイルドカードに置き換え検索結果を調べて妥当な単語を選ぶ方法や，複数の候補で迷っているならその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数の大きさでどの候補がより妥当であるか調べるなどの方法もある．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し，Web 上には誤った英文もあるので，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，正しい表現であると判断するのは難しいだろう．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．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の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，動詞と名詞からなるコロケーションによる動詞の誤り検出と修正を行うシステムを提案する．提案システムは検討したい英文を入力すると，Web 検索によりコロケーションが正しいかどうか判定し，誤りであれば修正候補となる動詞を提示する．

2. 関連研究

検索エンジンを用いた英文誤り検出は動詞以外の品詞に対しても行われている．例えば，

†1 岡山大学工学部
Faculty of Engineering, Okayama University
†2 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University

有冨ら[1] は英文中の前置詞誤りの修正を行っている．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．彼らは実験により，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，検索結果数を示すとこによって，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．検索で得たサマリ中で注目する動詞が，対応する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英文の動詞＋名詞コロケーションの誤り修正を行い，再現率 30.7% ，修正精度 37.3% で修正が可能であると報告している．

3. 提案システム

3.1 システムの概要

提案システムの簡単な処理の流れを説明する．まず，検討したい英文を入力する．次に入力された英文に対して OpenNLP[4] を用いて品詞タグ付けを行い，タグに基づき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より英文中に含まれる動詞+名詞のコロケーションが誤りであるか判別する．誤りとして検出された場合は，動詞の修正を行う．新たに動詞の修正のためのクエリを生成し，再度検索を行い，サマリを取得する．得られたサマリ中でコロケーションの中の名詞と共に用いられている動詞を抽出し，その出現回数を求め，それに基づきランク付けして動詞修正候補を提示する．

3.2 動詞＋名詞コロケーション誤りの検出

3.2.1 検索クエリ生成

英文中に含まれる動詞+名詞のコロケーションが正しいかどうか判別を行うためにフレーズ検索を行う．本節では提案するフレーズ検索のためのクエリ生成法について述べる．検索クエリは主に動詞から成る動詞クエリ，主に名詞から成る名詞クエリ，その 2 つを結合した共起クエリの 3 つを用いる．生成手順を以下の（1）から（5）に示す．

(1) 入力された英文に対して品詞のタグ付けを行う．
(2) 品詞タグに基づき，動詞を探す．
(3) 見つけた動詞の直後の品詞を調べ，前置詞か不変化詞が連続して出現している場合は，それらを含めて動詞クエリとする．
(4) （3）で求めた部分からそれ以降に出現する最初の名詞群までを名詞クエリとする．名詞クエリの終端は最初の名詞が出現して，その次に名詞以外の語が出現したところである．ただし，名詞が出現するまでに新しい動詞が現れた場合は，(2) で得られた動詞の修正は行わず，新しく出現した動詞に対し（3）以降の処理を行う．
(5) （3）, (4) で求めたフレーズを結合し共起クエリとする．

例えば，“It will go up the speed of the printing.” という誤りを含む文に対してクエリ生成を行うと，動詞クエリは “go up” となり，名詞クエリは “the speed” となり，共起クエリは “go up the speed” となる．

動詞直後の前置詞などは，句動詞の一部である場合が多く，そうでなくとも，その後に続く名詞よりも動詞との関連が強いので，動詞クエリに含める．

名詞の前にある冠詞などの違いによって共起する動詞が変化することもあるので，動詞クエリ直後から名詞直前までにある語は全て名詞クエリに含めている．複合名詞を扱えるように，名詞が連続して出現した場合も，全てクエリに含める．しかし，与えられた英文が第 4 文型（SVOO），第 5 文型（SVOC）であった場合，名詞クエリの妥当性が損なわれる可能性がある．例えば，目的語と補語両方に名詞が用いられている場合，後ろ側の名詞が冠詞や形容詞を伴わず単独で用いられていると複合名詞と判断し，名詞クエリに含むことになる．そのため，クエリが長くなったり，特殊なフレーズを含むことになり，検索結果数を大幅に減らすなどの悪影響を与える．また，第 4 文型の間接目的語に名詞をとる場合，人名などを

とることが多いので名詞と動詞との関係は弱いと考えられるため，動詞誤りの検出に用いるには妥当ではない．これらの問題に対しては構文解析を行ったり，検索結果数に応じてクエリに修正を施すなどの方法で対応していく必要がある．また，“tell me the story” のようなフレーズに対してクエリ生成を行うと，名詞クエリに代名詞 “me” も含まれる．しかしこのような場合，代名詞を含めることによって，検索結果数が極端に減るとは考えにくい．また，対象となる動詞が第 4 文型で用いられるのが適切かどうか判定する必要もある．よってこの場合，“me the story” を名詞クエリとする．

これらのクエリにスペルミスなど動詞誤り以外の誤りが含まれている場合，動詞誤り検出の精度が損なわれる可能性がある．そのため，事前にスペルチェックを行う必要がある．動詞直後の前置詞などは動詞クエリに含め修正対象の一部にした．名詞を修飾する形容詞に関しては，提案手法を拡張することによって対応できると考えている．冠詞は日本人が英作文を行う上で特に誤り易いので注意が必要だが，関連研究で述べたように多くの先行研究があるので本稿では扱わない．

3.2.2 検索結果数に基づく動詞誤り検出

生成した動詞クエリ，名詞クエリ，共起クエリそれぞれについてフレーズ検索を行い，検索結果数を求める．検索には Yahoo!デベロッパーネットワーク[5] 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動詞＋名詞コロケーションに対する MI スコア[6)7)] を求め，その値が閾値より小さい場合，誤りとして検出する．MI スコアとはコロケーション研究によく用いられている手法の一つであり，コロケーションの強度を測る指標である．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[7]．

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

ここで，共起頻度，中心語頻度，共起語頻度はそれぞれ，共起クエリ，名詞クエリ，動詞クエリで得られた検索結果数に対応する．コーパス総語数は全 Web ページ中の総語数にあたるため，実験で求めた適当な定数を用いる．

この総語数はあるコロケーションの MI スコアの値がわかれば，共起語，中心語，それらの結合形をクエリとしてフレーズ検索を行い，得られた検索結果数を用いて見積もることができる．そこで石川[7] が，複数のコーパスを合わせたデータを用いて計算したコロケーションの MI スコアを用いる．**表 1** の左から 2 列目は，石川が形容詞 “large” を中心語として，

表 1 MI スコアと総語数の見積
Table 1 The Estimated Total Number of Words and MI scores

共起語	MI スコア	共起頻度（百万回）	共起語頻度（百万回）	総語数（$\log_2$）
quantities	9.0007	24.7	153	43.3
scale	8.929	146	845	43.1
amounts	8.3482	43.3	442	43.4
numbers	8.0419	52.7	1610	44.6
cities	7.1417	8.57	1930	46.6
extent	6.9715	12.9	408	43.6
number	6.8338	146	5710	43.8
amount	6.4699	47.5	1450	43.1
part	5.6847	56.5	5680	44.0
area	4.8488	15.1	4770	44.8

直後位置に共起した回数の大きかった上位 10 件の共起語に対して MI スコアを求めたものである．表 1 の他の列は，検索 API が返した共起語や共起フレーズの検索結果数と，それらを利用して逆算した総語数の対数である．なお，“large” の検索結果数，すなわちこの場合の中心語頻度は 34 億件であった．総語数の対数の平均は約 44 となった．また，文献 8) では，MI スコアが 3 より大きい語のペアは，興味深いものとなる傾向が観察された，と述べられている．よって，本稿では総語数を 2^{44} とし，閾値を 3 とした．

また，MI スコアはコロケーションの検出に用いられているものなので，誤り検出に対しても有効であるか予備実験を行った．動詞誤りを含んだコロケーションと誤りのないコロケーションに対して，3.2.1 節で示した手法により 3 種類のクエリを生成し，それぞれの検索結果数から MI スコアを求め，その度数分布を求めた（**図 1**）．図 1 からわかるように，おおよそ誤ったコロケーションの MI スコアは負，正しいコロケーションのそれは正に偏っているので，誤り検出にも十分用いることができると考えられる．MI スコアが特に 0 から 3 の範囲で重複している．正しいコロケーションのうち，MI スコアが低いものは，MI スコアがいくつ以上であればコロケーションである，というような厳密な定義がないのである程度はやむを得ない．また，誤ったコロケーションのうち，MI スコアが高いものは，意図とは別の意味で用いられているコロケーションであったり，MI スコアの性質によりコロケーションが過大に評価されたものである．これらについては 4.1.1 節で詳しく述べる．また，MI スコアが高いものの中にはクエリ生成法が適切でなかったためである例もあった．

なお，MI スコアでコロケーションの誤りを検出した場合，名詞が誤っている場合もある

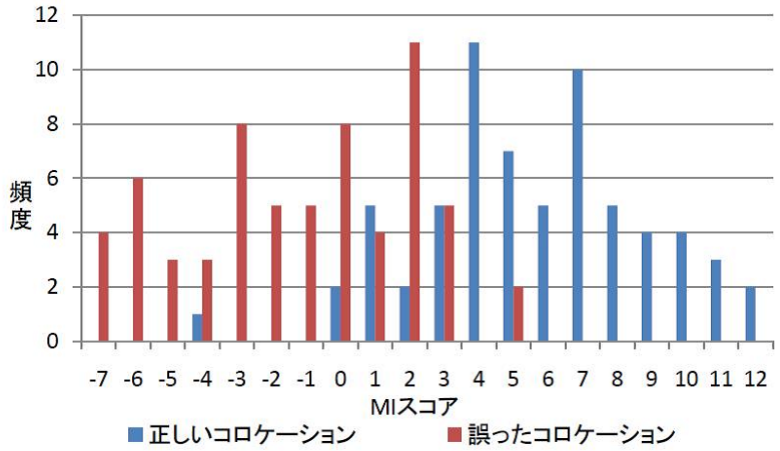

図 1 MI スコアの度数分布図
Fig. 1 Histogram of MI score

が，本稿では動詞のみを対象とする．

3.3 動詞修正候補の提示

3.3.1 動詞修正候補取得のための検索クエリ生成

動詞修正候補を得るための検索には，フレーズ検索と AND 検索を組み合わせて用いる．フレーズ検索で利用するのは検出の際に用いた名詞クエリであり，さらに文中に出現するそれ以外の名詞を AND 検索で追加する．例えば，“It will go up the speed of the printing.” に対し修正用のクエリを生成すると，“the speed” AND printing となる．

3.3.2 動詞修正候補の取得

検索によって得られたサマリを解析し修正候補となる動詞の取得を行う．動詞の抽出は以下の手順で行う．

(1) 得られたサマリをコンマとピリオドで分割する．
(2) 分割された文字列から名詞クエリのフレーズを探索する．
(3) フレーズが見つかればその文に対して Eric Brill の Monty Tagger[9] を使用して品詞タグ付けを行う．
(4) 文中でフレーズの直前に隣接して出現する動詞，あるいは動詞＋前置詞となっている部分を抽出する．

この処理を検索によって得た全てのサマリに対して行う．抽出した動詞は出現回数を数え，3.3.3 節で述べる方法でランク付けを行い提示する．なお，品詞タグ付けに用いるツールは動詞誤りの検出のものとは違うが，これは，OpenNLP は動詞タグと不変化詞タグのタグ付けに有利であり，Monty Tagger は実行速度に優れているためである．

本稿の実験では，Yahoo!検索 API で取得できる上限である 1000 件のサマリの解析を行っている．名詞クエリは主に名詞から構成されるためサマリ中に出現せず，タイトルや URL で用いられている場合がある．サマリ中で用いられている場合でも，キーワードとして使われていたり，文が途中で分割されていたりすることがあり，動詞を伴わないことがある．また，名詞クエリのフレーズが主語で用いられているなど期待する文型ではない場合もある．よって，必ずしも適切な動詞が得られるとは限らない．4.2.1 節の動詞修正の実験の際に抽出できた動詞の総数を確認したところ，最小 39 件，最大 1025 件で，平均すると 468 件だった．動詞修正候補の取得のために行う，検索や解析にかかるコストは大きいので，正しい動詞を十分に抽出し，ランク付けを行うために何件のサマリを解析するのが妥当かはさらに検討が必要である．

3.3.3 動詞修正候補のランク付け

3.3.2 節で得た動詞をランク付けをして，修正候補として提示する．ランク付けには動詞のサマリ中での出現回数と MI スコアを用いる．まず，出現回数が平均値以上のものと，それより小さいものに分ける．出現回数が平均値以上の動詞については，動詞クエリとして用い，名詞クエリと結合して共起クエリを作る．これらのクエリを用いて動詞誤り検出と同様に検索して MI スコアを計算する．ただし，ここで 3 種類のクエリを用いて検索を行う際に 3.3.1 節と同様に，文中に出現する名詞クエリ以外の名詞を用いた AND 検索で絞り込みを行い，検索結果数を得る．

こうして得られた修正候補となる動詞の出現回数と MI スコアを最大値が 1，最小値が 0 になるように以下の式で示すアフィン変換を行う．

$$x' = \frac{1}{\text{最大値} - \text{最小値}} x - \frac{\text{最小値}}{\text{最大値} - \text{最小値}} \tag{2}$$

ここで x は出現回数，または MI スコアを表し，x' は変換後のそれらの値を表す．変換後の出現回数と MI スコアの和を求め，値が大きい順にランク付けを行い，修正候補として提示する．

表 2 クエリごとの共起動詞
Table 2 Co-occurrence Verbs of Each Query

N-best	“speed”	“the speed”	“the speed” AND printing
1	use(11)	measure(25)	increase(66)
2	have(9)	travel at(18)	prove(36)
3	get(9)	break(16)	improve(32)
4	test(6)	travel(16)	offer(19)
5	provide(6)	exceed(13)	combine(16)
6	establish(5)	increase(10)	reduce(15)

3.3.4 動詞修正用の検索クエリに関する考察

名詞クエリは冠詞などを含むフレーズとなっており，冠詞の有無などによって共起する動詞に変化が見られる．例えば，“I’ll train my English.” という誤った例文があるが，これの名詞クエリは “my English” となる．この場合，3.3.2 節で述べた方法で共起する動詞を抽出すると，正解である “improve” が最もよく使われており 112 件抽出できる．しかし，単に “English” で検索すると最もよく用いられている動詞は 48 件抽出した “learn” となる．

“my English” のように名詞フレーズ単体で動詞を限定できればよいが，常にそのようなフレーズが生成されているとは限らないので，調べたい英文の特徴をより多く検索結果に反映させるため，文中の他の名詞も AND 検索で加える．例えばここで再び “It will go up the speed of the printing” という英文について考えてみる．**表 2** に “speed”，“the speed”，“the speed” AND printing のそれぞれのクエリから得た共起動詞の上位 6 件を示す．“the speed” で検索すると正解である “increase” は 6 番目の候補になるが，そこから “printing” で絞り込みを行うと 1 位になる．この例では，文中に含まれる語を多く含むクエリが正解動詞の獲得において有効に働いている．

4. 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[10]」で，挙げられている動詞誤りを含む例文 65 文とその正解例である 65 文を用いた．例文には，動詞＋名詞のコロケーションを含む部分がそれぞれ 66 か所あり，誤り文例はそのうちの 65 か所が誤っている．実験では動詞誤り検出と，修正を行った．動詞誤り検出実験では，動詞誤りを含む 65 のコロケーションに対して，誤り検出が行えるか実験すると共に，誤りのない 66 のコロケーションに対して誤検出を行わないか実験した．検出実験の結果については 4.1 節で述べる．動詞修正実験では，検出に成功した動詞誤りに対して修正候補を提示し，正解を提示できるかを調べた．この結果は 4.2 節で述べる．

4.1 動詞誤りの検出実験

4.1.1 動詞誤りの検出精度

3.2.2 節の結果を踏まえ，式 1 で定義した MI スコアが 3 以下となるコロケーションを誤りとして検出した．動詞誤りの自動検出結果を**表 3**，**表 4** に示す．検出の再現率は 87.7% ，適合率 85.0% ，F 値 86.4% という結果が得られた．誤りを含む例文に含まれている 65 の誤った動詞のうち，64 の動詞に対して検索フレーズを生成することができた．残りの 1 つは品詞タグ付けに失敗して，検索フレーズの生成が行えなかった．よって，64 の動詞について誤り検出を行ったところ，57 の動詞について誤り検出に成功し，7 の動詞については検出できなかった．うまく検出ができなかった理由の 1 つとして，意図した意味ではないが実際に使われるコロケーションに該当する事例があった．例えば，“She made a record in 1996.” という例文がある．これは，“彼女は 1996 年に記録を作った．” という意味で用いられており，よりふさわしい表現として “She set a record in 1996.” が挙げられている．“make a record” で “記録を作る” ということもあるが，“レコードを作る” という印象を与えるため推奨されていない．この場合は,“ レコードを作る ”という意味で多くの検索結果が得られ検出に失敗した．また，若干ではあるが MI スコアの性質によるものもあった．MI スコアはコロケーションに含まれるいずれかの単語の出現頻度が低い場合，そのコロケーションを過大に評価する性質がある．この場合，共起クエリの検索結果数が少ない場合であっても，動詞クエリと名詞クエリの検索結果数が小さいために，誤りが検出できないものがあった．

次に，誤りのない例文を入力とした場合，66 の動詞に対して 11 の誤検出があった．誤検出の原因として考えられるのは検出率のために閾値を大きくしすぎたことと，検索結果数が想定以上に大きいものがあったことの 2 点である．検索結果数については共起クエリから得られる検索結果数が大きいにも関わらず，動詞クエリ，あるいは名詞クエリから得られた検索結果数がさらに大きい場合があった．例を挙げると，“She cut class.” という正解例文に対し，クエリを生成し検索結果数を求めてみると，共起クエリは 516,000 件，動詞クエリは 1,670,000,000 件，名詞クエリは 2,740,000,000 件となった．ここから，MI スコアを計算すると 0.99 となり，閾値に満たない．共起クエリのみを見ると十分多いように見えるが，動詞・名詞クエリの件数がさらに多いため誤検出してしまう．

4.1.2 動詞誤りの検出に関する考察

検出性能を向上させるには，クエリ生成法の見直し，他の指標を MI スコアと合わせて用

表 3 動詞誤りの自動検出結果
Table 3 Results of Automatic Detection of Verb Errors.

誤っている動詞 (65)		誤りのない動詞 (66)	
検出	非検出	非検出	誤検出
57	8	56	10

表 4 動詞誤りの自動検出性能
Table 4 Performance of Automatic Detection of Verb Errors.

検出率	検出精度	F 値
87.7%	85.0%	86.4%

いるなどの方法が考えられる．クエリ生成法の変更は，例えば，大鹿ら[2] が行ったようにクエリ検索の際に，動詞の活用形や名詞の複数形などを考慮し，フレーズ検索と OR 検索を組み合わせる方法がある．例えば，“cut class” なら，“（cut OR cuts OR cutting）（class OR classes）” をフレーズ検索する．この手法を用いると “cuts classes”，“cutting class” などのフレーズも検索結果に含まれる．このようにすることで，動詞の活用形や文の時制に左右されずに，実際に Web 上で用いられているコロケーションの数により近い値を得ることができ，正しい結果が得られる可能性がある．

コロケーション検出によく用いられている他の指標としては，t スコア，対数尤度比などが挙げられている[6][7]．これらの指標は MI スコアとは異なる性質を持つので組み合わせて用いれば，検出性能の向上に役立つ可能性がある．

4.2 動詞誤りの修正実験

4.2.1 動詞誤りの修正精度

4.1.1 節で示した動詞誤りの検出実験で誤りを検出できた動詞 57 個に対して自動修正を試みた．提示した上位 3 件までの候補の中に正解があれば修正できたものとみなし，その件数とその時の修正精度を，修正候補のランク付けの方法ごとに**表 5** にまとめる．表 5 に示すように，出現回数と MI スコアによるランク付けで 1 位となった動詞と正解の動詞が等しくなったものは 22 件で，その修正精度は 38.6% であった．また，上位 3 件までに正解があればいいという少しゆるい条件では，修正精度は 50.9% となる．

うまく修正できなかった主な理由は，検索クエリの不備が挙げられる．検索クエリが元の文の意味を十分に反映できていなっかたため，正解の動詞が上位に現れなかったと考えられる．これに対応するには文中の語をより多くクエリに加えたり，フレーズ検索を多用したりする方法が考えられるが，加減を間違えると検索結果数が 0 件になる可能性がある．

表 5 動詞誤りの修正結果
Table 5 Result of Correction of Verb Errors.

	出現回数順		MI スコア順		出現回数＋ MI	
N-best	修正件数	修正精度	修正件数	修正精度	修正件数	修正精度
1	21	36.8%	17	29.8%	22	38.6%
2	23	40.4%	22	38.6%	25	43.8%
3	28	49.1%	27	47.4%	29	50.9%

また，ランキングに MI スコアの値を十分に反映できなかったことも挙げられる．MI スコアによるランキングの 1 位が正解の動詞となったときに，出現回数順のとき 1 位でなかった動詞は 5 件あったが，出現回数と MI スコアの組み合わせでランクが 1 位となったものは 1 件しかなかった．これに対しては，ランキングの統合の際に重み付けを行ったり，MI スコアの代わりに他の指標でランク付けしたりすることが考えられる．

4.2.2 動詞誤りの修正に関する考察

実験で用いた例文の多くがそうであったように，1 文からそれほど多くの情報が得られるとは限らない．そこで，修正前の動詞の意味を考慮して，修正候補をランク付けする方法が考えられる．英作文をする際に，たとえ間違った動詞を選んでも，正解の動詞との間に多少なりとも意味の類似点があると考えられるので，それを考慮に入れれば修正精度の向上が期待される．例えば，“She had to grow a child.” という文から “grow” という動詞誤りを検出し，修正のためのクエリを生成すると “a child” となる．ここから得られる動詞は**表 6** に示すように，出現回数順と MI スコアでランキングすると 1 位は “sponsor” となり，正解である “raise” は 3 位となっている．“grow” と “raise” は植物などを栽培するという共通の意味を持っており，英語のシソーラスである WordNet[11] では，同じ同義語のグループに分類されている．WordNet のような，シソーラス上で，語間のパスの長さを測るなどの方法から意味の類似度を求め，それを考慮することによって順位をより妥当なものにできる可能性がある．

5. ま と め

本稿では，検索エンジンを利用することで一般の日本人には難しい動詞名詞コロケーションの妥当性を判定し，動詞の誤りが検出された場合には，その動詞の修正候補を提示するシステムを提案した．実験ではこのコロケーションに基づく動詞誤り検出の F 値は 86.4% で

表 6 “a child” の共起動詞
Table 6 Co-occurence verbs with “a child”

	修正候補動詞	出現回数	MI スコア
1	sponsor	71	8.34
2	adopt	38	8.94
3	raise	17	8.35
4	be	38	3.14
5	have	16	5.50

あり，検出した動詞誤りの修正精度は 38.6% であった．

本研究では，動詞名詞コロケーションを対象としたが，形容詞＋名詞のような他のコロケーションであっても，対象となる語が互いに隣接していれば応用が可能であると考えている．動詞名詞コロケーションの場合でも，語同士が隣接していない場合があるが，動詞と名詞の間に出現する語を動詞クエリか名詞クエリの一部として扱うことで対応している．しかし，動詞と名詞の間に出現する語数が多い場合は，コロケーション誤りの検出が不可能となる場合がある．そのため，検索クエリから余分な語を削除する方法を定め対応していく必要がある．また，提案手法では名詞は誤っていないと仮定して修正を行っている．しかし，実際は，名詞を誤ることもあるので，コロケーションを構成する語の中で，どの語が誤っているかわからない場合には，更なる検討が必要となる．

今後の課題は，検索フレーズの生成方法，MI スコアによる検出方法，修正候補となる動詞のランク付け方法の改良を通して検出と修正の精度を向上させることである．また，本稿で対象としなかった動詞以外の誤りへの対応についても検討したい．

参 考 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) Yi, X., Gao, J. and Dolan, W.B.: 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) OpenNLP: http://opennlp.sourceforge.net/.
5) Yahoo!デベロッパーネットワーク：http://developer.yahoo.co.jp/.
6) 齋藤俊夫，中村純作，赤野一郎：英語コーパス言語学 -基礎と実践- (改訂新版)，研究社出版 (2005).
7) 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).
8) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicpgraphy, *Comtutational Linguistics*, Vol.16, No.1, pp.22–29 (1990).
9) Monty Tagger: http://web.media.mit.edu/hugo/montytagger/.
10) 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).
11) WordNet: http://wordnet.princeton.edu/.