Vol.2012-DBS-156 No.2
2012/12/12

検索エンジンを用いた英文動詞誤り検出システム

谷本　太郁由[1,a)]　太田　学[1,b)]

概要：英語を母語としない日本人の書く英文には，動詞の誤りが多く含まれる．本稿では，英文中の動詞誤りを検出するシステムを提案する．本システムでは，主語の人称・単複と動詞の活用形の一致に関する誤り，動詞の時制に関する誤り，動詞の語彙選択に関する誤りの 3 種類の誤りを検出する．誤りであるかどうかの判定に検索エンジンから得られる検索結果数を用いる．実験では，日本人が書いた英文からこれら 3 種類の動詞に関する誤りを検出し，その検出性能を評価した．

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は様々な誤りが含まれる．その中でも，動詞の誤りは特に多い．阪上 [1] は，学習者コーパスである，The NICT JLE Corpus[2] と NICE[3] の一部のデータに対して誤用分析を行っている．この分析では，学習者を初級・中級・上級の三つのレベルに分類し，各レベルの誤用状況を観察している．ここで，書き言葉の英語学習者コーパスである NICE の誤用分析結果によると，日本人の英作文には習熟レベルに関係なく動詞と名詞に関わる誤用が多いことがわかる．

また，日本人英語学習者コーパスである Konan-JIEM Learner Corpus Third Edition（KJ コーパス）に付与してあるエラータグの統計でも，動詞に関する誤りが最も多かった．この KJ コーパスには，日本人大学生によって書かれた 233 の英作文が収録されており，それらの英作文中の文法的な誤りにエラータグが付与されている．KJ コーパスのエラータグ毎の出現回数の割合を**図 1** に示す．これらの結果をふまえ，本稿では動詞に関する誤り検出手法を提案する．

英文中の誤り検出には検索エンジンが利用できる．例えば，フレーズ検索の検索結果数からその英語表現が実際に使われているかどうかがわかる．また，複数の表現候補があるならば，それぞれの候補を含むフレーズで検索を行い，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．本稿では，このフレーズ検索を用いて，英文中の動詞誤りを検出するシステムを提案する．

[1] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama Uniersity
[a)] tanimotot@de.cs.okayama-u.ac.jp
[b)] ohta@de.cs.okayama-u.ac.jp

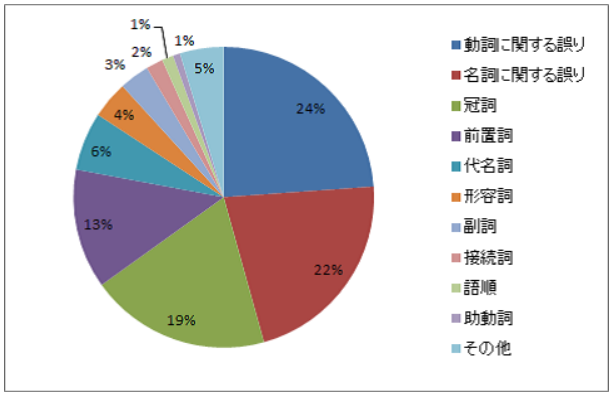

図 1　KJ コーパスのエラータグの出現割合

KJ コーパスの動詞に関する誤りは，主語-動詞の一致に関する誤り（一致誤り），時制に関する誤り（時制誤り），語彙選択の誤り（語彙誤り），その他の四つに分類されている．本システムではそのうち一致誤り，時制誤り，語彙誤りを扱う．一致誤りと時制誤りは複数の検索クエリを用いてフレーズ検索し，それらの検索結果数を比較することで誤りを検出する．また，語彙誤りは検索結果数を用いて前後の名詞との共起の強さを測り，その値が小さいとき語彙誤りとして検出する．実験では，KJ コーパスを用いてそれらの誤りの検出性能を評価する．

2. 関連研究

近年，多くの英文誤り検出，修正に関する研究が，特に冠詞や前置詞 [4][5] について多く行われている．また国内でも，2012 年には英文の「誤り検出・訂正ワークショップ（EDCW）」[*1]が初めて開催された．EDCW2012 では，動詞トラック，前置詞トラック，全ての誤りが対象のオープントラックの三つに分かれて，各トラックが対象とする誤り

[*1] https://sites.google.com/site/edcw2012/

の検出性能を参加チームが競いあった．動詞トラックは，動詞の一致誤りの検出が対象であり，我々の提案したルールと検索エンジンを用いたシステムが最良の F 値を達成した．

時制誤りや語彙誤りに関しては田尻ら [6]，大鹿ら [7]，Yi ら [8] の研究が挙げられる．田尻らは，CRF（Conditional Random Fields）[9] を用いて時制誤りの検出と訂正を行っている．言語学習者向け SNS である Lang-8 から大規模な時制誤りコーパスを作成し，そのコーパスを用いて CRF を学習し，時制誤り検出と訂正に利用している．実験では，時制誤りコーパス中の 10 万エントリを学習に 1,000 エントリをテストに用いて検出・訂正性能を評価している．そのテストデータは 16,308 の動詞句を含み，そのうち 1,072 が時制誤りである．閾値を変えるなどしていくつか実験をしているが，時制誤り検出の最良の F 値は 0.336 であり，その時の再現率は 0.282，適合率は 0.416 であった．

大鹿ら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．この類義語の誤り検出には動詞が対象として含まれている．このシステムは，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，検討箇所をそれぞれの候補で置き換えながらフレーズ検索する．それらのフレーズの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 らは Web 検索を用いて，冠詞，動詞，形容詞を修正している．ただし，動詞と形容詞の修正には，名詞とコロケーションをなしていなければならないという制約がついている．なお，コロケーションとは語の慣用的なつながりのことである．誤り検出と修正には，冠詞の場合は，a or an/the/ϕ の三つの場合の検索結果数と検索クエリ長を用いて算出したスコアを用いる．動詞と形容詞の場合は，検索結果のサマリを用いる．品詞タグ付けと複数の単語の意味的なまとまりであるチャンクの解析を行い，それらの結果に基づき検索クエリを作成する．次に，得られた検索結果のサマリで，検討したい語がコロケーションを成す名詞とつながりを持っているか調べる．その頻度が閾値未満の時，誤りであるとし，その単語を除いたクエリで再検索し修正候補を得る．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．

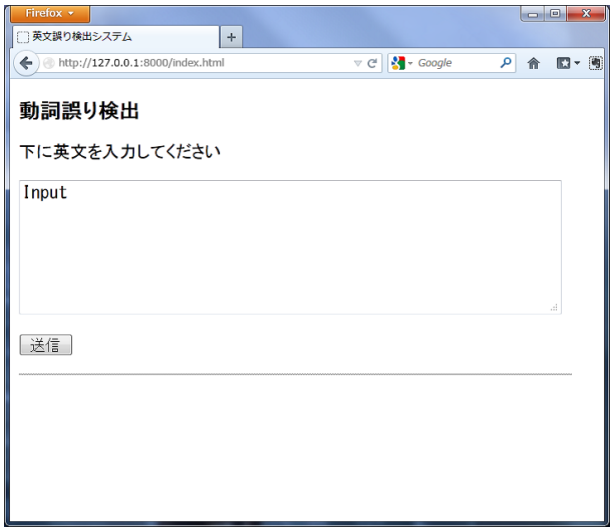

図 2　GUI 初期画面

図 3　GUI 実行画面

3. 提案システム

3.1 システム概要

まず，提案システムのインターフェースを**図 2** に示す．また，図 2 の入力欄に “Flowers is very beautiful.” という英文を与えたときの出力結果が**図 3** である．システムが誤りを検出した単語は赤色でハイライトされている．この英文は主語が “Flowers” であるが，動詞が “is” となっているため，動詞の一致誤りであると判定し，“is” が赤くハイライトされて表示されている．ここで，ハイライトされている部分にマウスオーバーすると，テキストボックスがポップアップされ一致，時制，語彙のうち，いずれの誤りを検出したかがわかる．

次に，動詞誤り検出の簡単な処理の流れを説明する．まず，与えられた英文の品詞タグとチャンク情報を取得する．この品詞タグとチャンク情報の取得は Stanford Parser

version 2.0.2[*2]を用いた．次に，一致，時制，語彙の誤りをそれぞれ検出する．最後に，それぞれの誤り検出結果をマージして表示する．一致誤り検出については 3.2 節，時制誤り検出については 3.3 節，語彙誤り検出については 3.4 節でそれぞれ詳しく説明する．

3.2 動詞の一致誤りの検出

主語-動詞の人称・数の一致に関する誤りは，まず品詞タグとチャンク情報を用いたルールベースの手法によって検出する．そして，品詞タグとチャンク情報から判定することが難しい英文に対しては，検索エンジンから得られる検索結果数を用いて，誤りを検出する．一致誤り検出の処理の流れを以下に示す．

(1) 主語-動詞の対応関係の取得
(2) 動詞の活用形の判定
(3) 主語の単複・人称の判定

ここで，(2) と (3) の対応が適切でない場合，動詞の活用形が誤っていると判定し，一致誤りとして検出する．

また，(3) では主語の単複・人称を判定することが難しい場合がある．例えば，主語に動名詞が含まれている場合，形容詞的用法か名詞的用法か判定しなければならない．このような場合，検索エンジンから得られる検索結果数を用いて誤りを検出する．検索エンジンを用いた一致誤り検出の流れを以下に示す．

(1) 主語と対応する動詞からなる検索クエリの生成
(2) (1) の動詞の活用形を変えた検索クエリの生成
(3) (1)，(2) を用いてフレーズ検索

(3) によって得られた検索結果数を比較し，(1) に比べて (2) の検索結果数の方が多い場合，動詞の活用形が誤っていると判定して検出する．

3.2.1 主語-動詞の対応関係の取得

まず，人称・数に関する誤りを含む可能性がある動詞と，それに対応する主語を探す．検討したい英文に以下のパターンでチャンクが出現した場合，それらのパターンに基づいて動詞とそれに対応する主語を取得する．

(1) NP + VP （名詞句 + 動詞句）
(2) NP + ADVP + VP （名詞句 + 副詞句 + 動詞句）
(3) WHNP + VP （関係代名詞句 + 動詞句）

ここで，各パターンの VP に含まれる動詞を誤り検出の対象とする．また多くの場合，これらのパターンに含まれている NP が主語である．しかし，そうでない場合もある．例えば，主語が前置詞句で修飾されている場合，NP + PP + NP + VP というパターンが出現する．ここで，PP は前置詞句を示す．このような場合，後ろの NP は主語の修飾なので，主語ではない．そのため，NP + PP + NP というパターンが出現した場合，PP の前の NP を主語とみなす．また，There is 構文の場合，(1) のパターンの NP は “there” になる．しかし，There is 構文は倒置なので，それに続く名詞句が主語となる．よって，There is 構文の場合，VP の直後のチャンクを調べ，そのチャンクが NP であった場合，その NP を主語とみなす．また，パターン (3) に適合した場合，WHNP の直前の NP を主語とみなす．このようにして，主語とそれに対応する動詞を取得する．

表 1 主語の人称・単複の判定に用いる基準

	三人称単数	三人称単数以外
イディオム	—	“a lot of”
代名詞	he, she, it, one	左記以外
限定詞	—	these, those
主語の最後の品詞	NN, NNP	NNS, NNPS

3.2.2 動詞の活用形の判定

次に，品詞タグに基づき動詞の活用形が三人称単数現在であるか，それ以外であるかを調べる．対象となる動詞は 3.2.1 項で見つけた VP の先頭の単語である．その単語の品詞タグが VBZ であった場合は三人称単数現在形であり，VBP であった場合はそれ以外である．なお，VP チャンクの先頭の単語が VBP，VBZ 以外の場合は，このチャンクの動詞に対する誤り検出を行わない．

ただし，VP の先頭の単語の品詞タグが RB（副詞）の場合は，次の単語の品詞タグより動詞の活用形を判定する．また，VP の先頭の単語が “was”，“were” の場合は，“was” を三人称単数の主語に対応する動詞，“were” をそれ以外と判定する．

3.2.3 主語の人称・単複の判定

次に，主語の人称・単複を判定する．ここでは，主語が三人称単数であるか，それ以外であるかに分類する．本システムで人称・単複の判定に用いている基準を**表 1** にまとめる．表 1 を上から順に適用して，主語が三人称単数かそうでないかを判定する．例えば，主語が “a lot of” というイディオムを含む場合，主語を複数とみなし三人称単数以外と判定する．多くの場合，主語の最後の品詞を調べることで人称・単複を判定できる．例えば，“database system” が主語ならば，“system” は名詞の単数形なので三人称単数に分類できる．

ただし，主語の最後の名詞の単複では，人称・単複を正しく判定できない場合がある．例えば，名詞がカンマや “and” を用いて列挙されている場合，一つ一つの名詞は単数形であっても，複数形として扱わなければならない．そのため，カンマと “and” で区切られて二つ以上の名詞が列挙されている場合は，三人称単数以外と判定する．この名詞の列挙による複数の判定は表 1 に示した判定基準より優先して行う．

ここで求めた主語の人称・単複と 3.2.2 項で求めた動詞の活用形の対応が不適切であれば，動詞を一致誤りとして検出する．

[*2] http://nlp.stanford.edu/software/lex-parser.shtml

3.2.4 検索エンジンを用いた誤り検出

3.2.3 項で述べた方法では，主語の人称・単複を判定することが難しい場合がある．例えば，動名詞では形容詞的用法か名詞的用法か判定しにくい場合がある．また，入力が誤りを含む英文であるため，品詞タグ付け自体が誤ることも珍しくない．そのような場合を想定して，本システムでは検索エンジンを用い検索結果数を比較することで，誤りを検出する．

具体的には，3.2.1 項で主語と判定した NP に “ing” で終わる単語が含まれている場合と，主語とした NP の直前のチャンクが VP であり，なおかつ，その VP が品詞タグが VBG の単語 1 語からなる場合は検索エンジンを利用する．例えば，主語が “Reading books” である場合，その品詞タグとチャンクは [NP Reading/NN books/NNS] か，[VP Reading/VBG] [NP books/NNS] となることが多い．また，品詞タグ付けの誤りが疑われる場合として，以下の 3 パターンを定めた．

- 品詞タグが名詞単数形を示しているが，単語の語尾が “s” である場合
- 品詞タグが名詞複数形を示しているが，単語の語尾が “s” でない場合
- 品詞タグが動詞三人称単数現在形を示しているが，単語の語尾が “s” でない場合

このようなパターンで品詞タグとチャンクが出現した場合，検索結果数を用いて検出する．なお，“datum/data” のような場合は．単語の語尾だけでは品詞タグ付け結果が誤っているとは言い切れないので，今後，辞書を参照するなどして対応する予定である．

次に，検索クエリの生成方法について述べる．ここでは，二つの検索クエリを生成する．検索クエリは，3.2.1 項で主語と判定した NP とそれに対応する VP に含まれる単語を用いて生成する．この際，NP の直前の VBG のみからなる VP に含まれる単語も検索クエリに含める．これらのチャンクに含まれる単語を先頭から順に並べたものを一つ目の検索クエリとする．次に，主語に対応する動詞の活用形のみを変化させた検索クエリを生成する．主語に対応する動詞が原形であれば三人称単数現在形に，三人称単数現在形であれば原形に変化させる．ただし，動詞が be 動詞の場合は，原形の代わりに “are” を用いる．また，過去形の “was”，“were” にも対応している．例えば，“Reading books is 〜” という英文では，本システムは “Reading books is” と “Reading books are” という二つの検索クエリを作成する．

こうして二つの検索クエリを生成し，フレーズ検索により検索結果数を取得する．得られた検索結果数を比較し，動詞の活用形を変えた検索クエリによる検索結果数の方が多い場合，一致誤りであると判定する．なお，検索結果数の取得には Yahoo! デベロッパーネットワーク[*3]が提供する Web 検索 API を利用した．

3.3 動詞の時制誤りの検出

時制誤り検出は 3.2.4 項で述べた手法と同様に，複数の検索クエリを生成し，それらの検索結果数を比較することによって行う．ここで生成した検索クエリは検討したい動詞の時制が異なっている．動詞の時制を変化させた検索クエリを用いた検索結果数が，時制を変化させていない検索クエリを用いた検索結果数より大きい場合，その動詞は時制誤りであると判定する．

次に，検索クエリの生成方法について述べる．時制誤り検出にはフレーズ検索と AND 検索を組み合わせた検索クエリを用いる．フレーズ検索部分は検討する動詞から生成する．3.2.1 項で述べた手法によって VP チャンクを取得する．その VP に含まれる全ての単語を用いてフレーズ検索部分を生成する．AND 検索部分は，検討したい文から，その VP に含まれる単語を除いた，全ての単語を AND 結合したものである．フレーズ検索部分と AND 検索部分を，さらに AND 結合したものを検索クエリとして用いる．例えば，“I go to school everyday.” という英文からは，“[NP I/PRP] [VP go/VBP] [PP to/TO] [NP school/NN everyday/JJ] ./.” という品詞タグとチャンク情報が得られる．この英文から生成されるフレーズ検索部分は “go” であり，AND 検索部分は “I AND to AND school AND everyday” である．よって，検索クエリは “‘go’ AND I AND to AND school AND everyday” となる．ここで，シングルクォーテーションで囲まれた部分はフレーズ検索のクエリを示す．

比較に用いる検索クエリは，AND 検索部分は同じになるがフレーズ検索部分に違いがある．すなわち，フレーズ検索部分に含まれている動詞の時制が異なっている．例えば，“I go to school everyday.” の比較に用いるクエリのうち過去形のクエリを生成すると，“‘went’ AND I AND to AND school AND everyday” となる．

3.4 動詞の語彙誤りの検出

一致誤りと時制誤りの検出では複数の検索クエリを用いて得た検索結果数を比較し，その大小により誤りであるかどうか判定した．しかし，語彙誤りの場合，まず比較対象となる動詞の選出を行わないと，検索結果数の比較による誤り検出は困難である．すなわち，比較するためには検討したい動詞より適切な動詞を用意する必要があるが，そのような動詞を見つけることがそもそも困難である．そこで，検討したい動詞とその動詞の主語や目的語である名詞との MI スコア [10] を算出し，その値が閾値未満の場合，

[*3] http://developer.yahoo.co.jp/

誤りとして検出する手法を用いる [11].

語彙誤り検出の処理は以下のようになる.

(1) 検索クエリの生成と検索結果数の取得
(2) MI スコアの算出
(3) MI スコアが閾値未満の動詞を語彙誤りとして検出

3.4.1 MI スコア

MI スコアとは，ある二つの単語の共起の強さを測る指標の一つであり，以下の式で定義される.

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

ここで，共起頻度は二つの単語があるコーパスにおいて共起した回数であり，共起語頻度と中心語頻度はそれぞれの単語の出現回数である.

本システムでは，検索エンジンを用いるので，共起頻度，共起語頻度，中心語頻度は検索結果数とする．また，コーパス総語数は Web 上の総語数となるが，これを正確に求めることは困難である．また，式 (1) においてコーパス総語数は定数なので無視する．MI スコアは二つの単語の共起の強さを測る指標であるが，二つの単語の関係のみで誤りを検出するのは難しい場合がある．そこで，共起語頻度と中心語頻度は必要に応じて複数の単語からなるフレーズに拡張して用いる.

3.4.2 検索クエリの生成

MI スコアを求めるには，共起頻度，共起語頻度，中心語頻度それぞれに相当する検索結果数を得るために，三つの検索クエリが必要となる．共起頻度に相当する検索結果数を得るための検索クエリは残りの二つの検索クエリを並べたものである．共起語頻度，中心語頻度を得るための検索クエリは，それぞれ動詞句と，それに関わる名詞句から生成する.

まず，3.2.1 項で示した 3 種類のチャンクの並びによるパターンを用いて，対象となる VP チャンクを取得する．この VP チャンクを元に動詞クエリを生成する．まず，3.2.1 項で求めた VP チャンクに含まれる全ての単語を動詞クエリに含める．また，その VP チャンクに連続して VP, ADVP, PRT, PP チャンクが出現する場合，これらのチャンクに含まれる単語も全て動詞クエリに含める．例えば，“I live in Okayama.” という英文には，“[NP I/PRP] [VP live/VBP] [PP in/IN] [NP Okayama/NNP] ./.” という品詞タグとチャンク情報が付与される．この英文から動詞クエリを生成すると “live in” となる．これらの四つのチャンクには動詞の他に副詞や前置詞が含まれており，動詞を修飾したり，句動詞となったりするため動詞クエリに含めている．なお，句動詞が修飾されている場合など，VP, ADVP，PRT，PP が二つ以上連続して出現することもある．そのような場合は，VP, ADVP, PRT, PP 以外のチャンクに達するまで，全て動詞クエリに含める.

次に，名詞クエリを生成する．名詞クエリは 2 種類ある．主語の名詞からなるものと，目的語の名詞からなるものである．主語からなる名詞クエリは 3.2.1 項で主語と判定した NP チャンクに含まれる単語からなる．目的語からなる名詞クエリは，目的語と考えられる NP チャンクを探し，その NP チャンクに含まれる単語を用いて生成する．なお，目的語とする NP チャンクは動詞クエリの直後の NP チャンクである．ただし，その NP チャンクの直前に ADJP（形容詞句）が挿入されていてもよいものとする．NP チャンクの直前に ADJP チャンクがある場合，ADJP チャンクに含まれる単語は名詞クエリには含めない．例えば，“I found very interesting animals.” という英文には，“[NP I/PRP] [VP found/VBD] [ADJP very/RB interesting/JJ] [NP animals/NNS] ./.” という品詞タグとチャンク情報が付与される．この文から目的語からなる名詞クエリを生成すると “animals” となる.

共起頻度を求めるための検索クエリは動詞クエリと名詞クエリを並べたものである．ただし，名詞クエリを二つ生成したので共起頻度を求めるための検索クエリも二つ生成する．すなわち，主語の名詞クエリと動詞クエリを結合したものと，動詞クエリと目的語の名詞クエリを結合したものである．この際，主語の名詞クエリは動詞クエリの前に配置するが，目的語の名詞クエリは動詞クエリの後ろに配置する.

3.4.3 MI スコアを用いた動詞の語彙誤り検出

3.4.2 節で生成したクエリを用いてフレーズ検索し，得られた検索結果数を用いて MI スコアを二つ求める．すなわち，主語の名詞クエリと動詞クエリの MI スコアと，動詞クエリと目的語の名詞クエリの MI スコアである．いずれかの MI スコアが閾値未満である場合，動詞が語彙誤りを含むと判定する.

ただし，主語，目的語が代名詞のみからなる場合，その名詞クエリを用いた語彙誤り検出を行わない．例えば，主語が代名詞であった場合，主語側の名詞クエリを用いた語彙誤り検出をしないので，動詞クエリと目的語側の名詞クエリの MI スコアのみを用いて語彙誤りかどうか判定する.

また，動詞クエリが be 動詞 1 単語からなる場合は語彙誤りの検出をしない．これは，be 動詞のみだと主語や目的語がどのような単語であっても，多くの場合あまり不自然とならないためである．このような場合は，主語，動詞，目的語全てを用いた検索クエリを生成するなどの対応が考えられるが，それは今後の課題である.

4. 評価実験

実験では，KJ コーパスの Third Edition に新しく追加された EDCW2012 のフォーマルラン用のデータを用いて，誤り検出性能を評価した．KJ コーパスには，様々な誤りを含む英文が収録されている．それらの英文から，動詞の一致，時制，語彙それぞれに関する誤りを検出する.

表 2 主語-動詞の人称・数の一致に関する誤り検出結果

	検出	非検出	誤検出
提案システム	23	0	12
okayamaU	23	0	16

表 3 主語-動詞の人称・数の一致に関する誤り検出性能

	再現率	適合率	F 値
提案システム	1.000	0.657	0.793
okayamaU	1.000	0.590	0.742

表 4 KJ コーパスの全データに対する一致誤り検出結果

検出	非検出	誤検出
92	13	65

表 5 KJ コーパスの全データに対する一致誤り検出性能

再現率	適合率	F 値
0.876	0.586	0.702

ただし，KJ コーパスにおいて uk タグの付与されている部分に関しては，評価対象外とした．uk タグは，エラー分類が不可能である場合や，構成上の致命的なミスに対して付与されるエラータグである．また，エラータグが ch 属性を持つ場合も，評価対象外とする．元々，文法的に誤りではないが，他の誤りを修正することによって誤りとなる場合がある．KJ コーパスでは，このような場合にもエラータグが付与されるが，そのエラータグが ch 属性を持つ．また，語彙誤りには，必要な単語が不足している欠落誤りが含まれる．しかし，提案手法では欠落誤りの検出ができないので，これも評価対象外とした．なお，フォーマルラン用のデータには動詞の欠落誤りが 21 件あった．

4.1 動詞の一致誤りの検出実験

KJ コーパスに収録されている英文を与えて，一致誤りを正しく検出できるかどうか実験した．一致誤り検出結果を**表 2**，**表 3** にまとめる．比較のため KJ コーパスに付随している EDCW2012 の動詞トラックで我々のチームの結果（okayamaU）を載せる．

再現率は変わらないが，誤検出が減少しており，その結果として適合率と F 値が上昇している．これは，本稿の提案システムでは品詞タグの誤りを考慮したためである．

一致誤りについては，提案手法は閾値の設定や学習が不要なので，KJ コーパスに含まれている全英文に対しても誤り検出実験を行った．その結果を**表 4**，**表 5** にまとめる．

表 4 の検出もれの原因の一つには，文の構造が複雑であることが挙げられる．例えば，“and” を用いて複文を書くとき主語が省略されることがある．また，文中で主語の倒置が起こると 3.2.1 項で述べた手法では，正しい主語が得られない．このような文構造が原因で検出できなかった誤りが 4 件あった．また，品詞タグ付けが誤っているため，検出できなかった誤りが 4 件あった．

表 6 動詞の時制誤り検出結果

	検出	非検出	誤検出
提案システム	23	66	107
EDNII	20	69	155
ベースライン	63	26	582

正しい動詞を誤検出した原因は，英文中の他の誤りの影響が挙げられる．例えば，主語が複数形にも関わらず動詞が三人称単数現在形であった場合，本システムは動詞が誤りであると判定するが，実際には名詞の単複が誤っていることもある．また，一致誤り以外の動詞誤りを検出しなければならない際に，一致誤りを検出したものもある．このような他の誤りの影響を受けて誤検出したものが 37 件あった．また，品詞タグ付けの誤りによる誤検出が 12 件，正しい主語が獲得できなかったことによる誤検出が 5 件あった．

誤検出，非検出を減らすためには，主語の獲得方法を見直す必要がある．本システムでは，3.2.1 項で述べたようなシンプルなルールで主語と対応する動詞を獲得しているが，それを改善する余地がある．また，誤検出の原因で挙げた名詞の単複に関する誤り検出を検討することが挙げられる．名詞の単複に関する誤りは，名詞の可算/不可算や冠詞などから誤りを判断できるものがある．これを実装すれば，誤検出を減らし，適合率と F 値を向上させることができると考えている．

なお，一致誤りに関しては誤りを検出すれば，システムがその誤りを修正することは比較的容易であると考えられるので，今後実装したい．

4.2 動詞の時制誤りの検出実験

KJ コーパスでは，本来は過去時制とすべき動詞を現在時制で用いる誤りが極端に多かった．そのため，本実験では現在時制の動詞に対してのみ時制誤り検出を行った．また，比較する検索クエリは過去時制に変化させたもののみを用いた．

時制誤り検出実験の結果を**表 6**，**表 7** にまとめる．また，比較のため KJ コーパスに付随する ngan チームの結果（EDNII[12]）とベースライン手法の結果を載せる．EDNII は Temporal Centering[13] で示された理論を簡易化したものを実装し，時制誤りを検出している．ベースライン手法は 3.2.1 項で得た VP チャンクが現在形であれば，無条件で時制誤りと判定するものである．なお，提案システムは現在時制の動詞に対してのみ時制誤り検出を行っているが，表 6，表 7 の結果は全ての時制誤りに対しての検出結果を集計している．

表 6，表 7 を見ると提案システムが最も良い F 値を示しているが，改善の余地が随分ある．本実験で，提案システムは現在時制以外の動詞を無視したが，現在時制以外の時

表 7 動詞の時制誤り検出性能

	再現率	適合率	F 値
提案システム	0.258	0.177	0.210
EDNII	0.225	0.114	0.152
ベースライン	0.708	0.098	0.172

表 8 動詞語彙誤りの検出結果

検出	非検出	誤検出
20	51	167

表 9 動詞語彙誤りの検出性能

再現率	適合率	F 値
0.282	0.107	0.155

制誤りにも対応していく必要がある．

次に，比較用の検索クエリを増やすことが考えられる．本実験では，比較には過去時制のみを用いた．しかし，KJ コーパスではアスペクト（相）の誤りを時制の誤りに含めている．すなわち，現在進行形や現在完了形などの誤りも時制の誤りに含まれる．これを考慮すると単純には，現在形，過去形，未来形の 3 種類に，それぞれの完了形，進行形，完了進行形の 9 種類を加えた計 12 種類のクエリで比較することが考えられる．また，比較用の検索クエリを充実させ，最も検索結果数が多いものを正しい時制と判定すれば，時制誤りを修正できる可能性があるので，この点に関しても検討したいと考えている．

時制誤り検出に用いる検索クエリは，3.3 節で述べたようにフレーズ検索部分と AND 検索部分からなる．この，AND 検索部分は単純に文中の単語を羅列したものである．そのため，時制や時間に関係ありそうな単語を選定するなどすれば検出性能が向上すると考えられる．また，英文が短く，時制を判断できるような検索クエリを生成できない場合がある．そのような場合は，直前の文から検索フレーズに加えられそうな単語を探すことが考えられる．また，時制は前方の文の影響を受けることが多いので，前方の時制を併せて考慮することが考えられる．

4.3 動詞の語彙誤りの検出実験

3.4 節で説明した手法により，語彙誤りを検出できるかどうか実験した．なお，MI スコアの閾値は KJ コーパスに収録されている英文のうち，語彙誤り検出実験に使用しなかった英文を用いて，誤り検出の F 値が最も高くなる値を求めた．

フォーマルラン用の英文には動詞の語彙誤りは 71 件あり，そのうち 20 件を検出できた．またその際，167 件を誤検出した．語彙誤りの検出結果を**表 8**，**表 9** にまとめる．

誤検出が多いが，これは他の箇所の誤りの影響が大きいと考えている．語彙誤りの検出では，動詞クエリと名詞クエリを生成し，それらの MI スコアを求め，その値が閾値未満の場合，動詞が誤りであると判定する．しかし，名詞クエリに誤りが含まれていたり，名詞クエリと動詞クエリの間に欠落誤りがあっても MI スコアは低くなる．そのような誤りも動詞の語彙の誤りとして検出したため，誤検出が多くなったと考えられる．

一方，非検出の原因として，適切な閾値を用いて誤り検出ができなかったことが挙げられる．他の種類の誤りを動詞の語彙誤りとして誤検出するため，この誤検出を抑えるために MI スコアの閾値を低く設定した．そのため，本来なら誤りとすべき動詞で検出できないものが増えた．また，検索クエリの生成がうまくいかず検出できないこともあった．本システムは語彙誤り検出の対象の動詞を，3.2.1 項で得られた VP に含まれる動詞に限定している．そのため，動詞が不定詞や分詞などで修飾語句として用いられると検出対象とならないので，今後検討する必要がある．

5. まとめ

本稿では，検索エンジンを用いて，英文中の動詞誤りを検出するシステムを提案した．提案システムでは，主語-動詞の人称・単複の一致に関する誤り，動詞の時制に関する誤り，動詞の語彙選択に関する誤りの 3 種類の誤りを検出する．また，Konan-JIEM Learner Corpus Third Edition に含まれている英文を用いて，動詞誤り検出性能の評価実験を行った．

一致誤りの検出では，品詞とチャンク情報からルールに基づいて誤りを検出する．ルールに基づく誤り検出が難しい場合は，検索エンジンから得られる検索結果数を比較するという方法を提案した．一致誤り検出の F 値は，0.793 であり，EDCW2012 で最良の F 値を達成した我々のシステムの検出性能を上回った．時制誤りの検出は，動詞の時制が異なる複数の検索クエリを生成し，それらの検索結果数を比較することによって時制誤りであるか判定する．時制誤り検出の F 値は 0.21 であり，ベースラインの検出性能を上回った．語彙誤りの検出は，動詞と名詞句に基づく検索クエリを用いて得た検索結果数より，それらの MI スコアを求め，その値が閾値にみたないの場合，語彙誤りとして検出する．語彙誤り検出の F 値は 0.155 であった．

今後の課題として，本稿で提案した 3 種類の誤り検出の対象を広げることが挙げられる．例えば，本稿の時制誤り検出は検出対象を現在形に限定しており，語彙誤り検出は欠落誤りに対応していない．また，誤り検出ルールの適用順序や誤り修正についても検討したい．提案システムは，一致誤り，時制誤り，語彙誤りを並行して検出している．例えばこれを，一致誤りと時制誤りを修正した後に語彙誤りを検出するようにすれば，検出精度を向上させることができる可能性がある．最終的には，動詞以外の品詞に対しても誤り検出，修正機能を併せて実装したいと考えている．

参考文献

[1] 阪上辰也：The NICT JLE Corpus・NICE のレベル別誤用分析，自然言語処理技術を応用した英語学習者の誤用に関する包括的かつ体系的分析，名古屋大学 大学院国際開発研究科，pp. 137–146 (2008).
[2] 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，アルク (2004).
[3] 杉浦正利：英語学習者のコロケーション知識に関する基礎的研究，平成 17〜19 年度 科学研究費補助金 (基盤研究 (B)) 研究成果報告書 (2008).
[4] 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).
[5] 久保田朗，太田 学：検索エンジンを用いた英文前置詞誤りの自動検出を修正，情報処理学会研究報告，Vol. 2011-DBS-153, No. 2, pp. 1–8 (2011).
[6] 田尻俊宗，小町 守，松本裕治：大域的文脈情報を用いた英語時制誤りの検出と訂正，言語処理学会第 18 回発表論文集，pp. 357–360 (2012).
[7] 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，DEWS，4B-i8 (2005).
[8] 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).
[9] Laffery, J.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Proceeding of ICML, pp. 282–289 (2001).
[10] 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).
[11] 谷本太郁由，太田 学：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動詞誤りの検出と修正，情報処理学会研究報告，Vol. 2010-DBS-151, No. 36, pp. 1–7 (2010).
[12] Ngan L. T. Nguyen, Yusuke Miyao: ED-NII Error Detection Tool User Manual, `http://code.google.com/p/edcw2012-verb/source/browse/?name=ngan#git%2Fngan` (2012).
[13] Megumi Kameyama, Rebecca Passonneau, and Massimo Poesio: Temporal centering, Proceedings of the 31st annual meeting on Association for Computational Linguistics, Association for Computational Linguistics, pp. 70–77 (1993).