英語学習支援
―誤り自動校正手法とその応用―

基般

乙武 北斗
福岡大学工学部

冠詞が苦手な日本人

現在，数多くの日本人が英語を学んでいる．2011 年度から小学校 5・6 年生で英語学習が必修化され，近年の国際化と相まって，日本人英語学習者の数は今後も増加を続けることが予想される．

中学校・高校の 6 年間を経て，主に英語の文法や単語を学ぶ人は非常に多く，さらに高いレベルの学習に取り組む人も多い．そのような日本人英語学習者が共通して間違いやすい文法項目が，日本語には存在しない「冠詞」と「前置詞」である．

日本人大学生によって書かれた英文エッセイをネイティブチェックによる注釈とともにまとめた Konan-JIEM learner corpus[1)]（以下，KJ learner corpus）を使用し，どのような誤りが含まれるかを調べてみたところ，図 -1 のような結果となった．図 -1 は全 2,737 個 20 項目の誤りにおける割合を示したものであるが，上位 4 項目の誤りの数が全体の 5 割を超えている．冠詞と名詞の単数／複数形の選択は密接なつながりがあるため，これらをまとめて冠詞誤りと見れば，上位 4 項目の内訳は冠詞，前置詞，動詞の時制誤りとなる．

これら誤りの中でも特に「冠詞」と「前置詞」は，日本人以外の英語学習者にとっても難しい文法項目であり，誤用が目立つことがさまざまな研究から報告されている．そのため，この 2 項目の誤りを自動的に検出・校正する試みは盛んに研究されている．

以下，本稿では，はじめに英文の文法誤りを自動的に検出・校正を行う研究の歴史について述べた後，各文法項目の誤りをどのようにして校正するかについて説明する．その後，文法誤り自動校正手法の応用例としての英語学習支援について述べる．

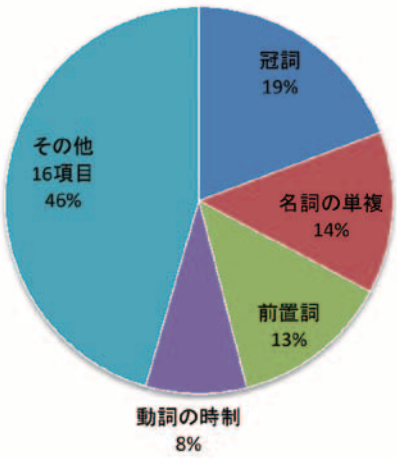

図 -1　KJ learner corpus に含まれる英文誤りの分類

文法誤り校正の自動化の歴史

図 -2 に英文の文法誤り自動校正を実現するために用いられてきた手法の概要を登場順に示す．

1980 年代に入るまでは，文法的な解析処理を含む手法よりも，文字列の一致を調べることによって誤りを検出する手法が主流であった．このような文字列マッチングの応用の 1 つとして，図 -2 の概要図でも示している単語のスペルチェックが挙げられる．また，不規則動詞の活用（たとえば *eat* の過去形が *ate*）を情報として持つことで，*eated* のような誤りを修正することも可能である．

1980 年代に入ると，文法的な解析処理を備えた手法が登場する．図 -2 の解析的手法で示されるように，入力文の構文構造を解析し，それに対して人手で作成されたルールを適用することで，文字列マ

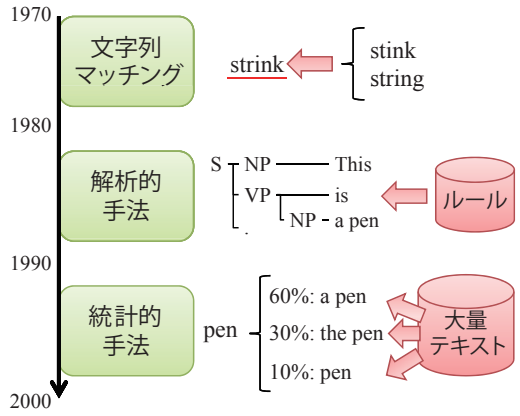

図 -2　誤り校正手法の移り変わり

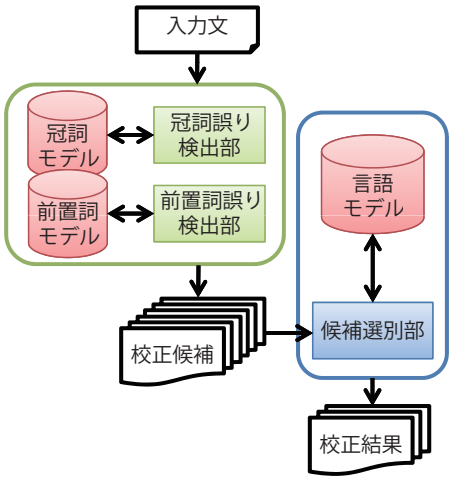

図 -3　Gamon の手法の処理の流れ

ッチングでは困難な文法的誤りの検出が可能となった．河合らの手法[2]では，正常な構文解析ができない入力文に対して，構文規則を緩めて再解析することにより，冠詞や動詞の用法といった統語的誤りの検出が可能である．

その後 1990 年代に入り，計算機の処理能力の向上や言語資源が整備されてきたことが相まって，大規模データに基づく統計的手法が誤り校正においても用いられるようになった．これによって，人手でルールを作成する労力を削減しつつ，複雑な用法を持つ文法の誤りを高い精度で検出することが可能となった．統計的手法による冠詞・前置詞誤りの校正については，次章で Gamon の手法[3]を詳しく説明する．

現在，誤り校正の分野は統計的手法をベースにさまざまな研究が進められているが，統計的手法がすべての誤りの種類にフィットするわけではない．不規則動詞の活用誤りに代表されるような，人手によるルールを用いる方が適している誤りも存在する．さまざまな種類の文法誤りを対象とする自動校正システムを実現しようとした場合は，その種類に適した手法の選択が重要だと考えられる．

冠詞と前置詞の校正手法

本章では，英文の冠詞と前置詞を校正する統計的手法の一例として，Gamon の手法の概要を説明する．

図 -3 に，Gamon の手法の処理過程を示す．入力文はまず，冠詞・前置詞誤り検出部にて事前に作成された統計モデルを用いて，それぞれの誤りを含むかどうかが判定される．誤りを含むと判定された際には，その校正候補を出力する．

誤り検出部が出力した校正候補は候補選別部にて，事前に作成された言語モデルを用いた選別が行われる．選別の結果残ったものだけが，実際の校正結果として出力される．

■冠詞・前置詞モデルの作成

統計的手法を用いた誤り校正では，大量のテキストデータから，対象とする文法項目とその用法を決定づけていると思われる手がかりを抜き出し，推定モデルを作成する必要がある．このような手がかりを素性（そせい）と呼ぶ．Gamon の手法では，冠詞と前置詞の推定に最大エントロピー分類器という分類アルゴリズムの 1 つを用いている．これは大まかに説明すれば，この素性にはこの冠詞が付きやすいといった確率値を事例から学習していくものである．

推定に用いる素性の例を図 -4 に示す．例文 (i) は冠詞，例文 (ii) は前置詞の素性抽出を表している．Gamon の手法では冠詞と前置詞ともに，前後 3 単語とその品詞情報を推定のための手がかり☆1 とし

☆1　実際は単語と品詞に加え，いくつかほかの素性も用いている．

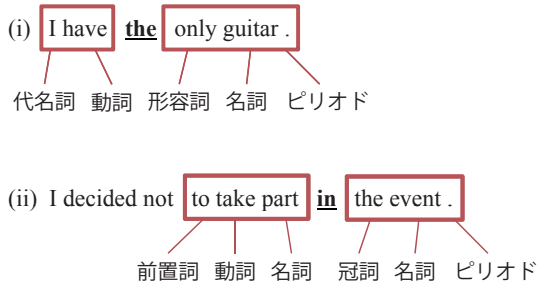

図 -4　推定に用いる特徴の例

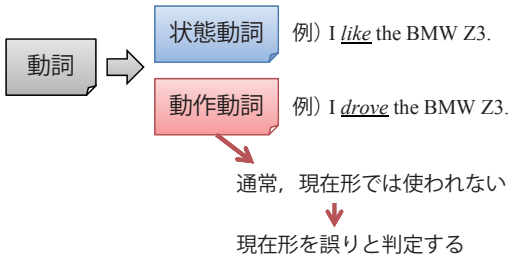

図 -5　永田らの手法の基本的な考え方

て用いている．例文（i）では，定冠詞 *the* を基準として，図の赤枠で示されている前後部分が素性として用いられる．例文（ii）では前置詞 *in* を基準として同様に素性を抽出する．

このような素性抽出を，誤りが含まれないネイティブによって書かれた英文を対象に行い，推定モデルを構築する．Gamon の手法ではニュース記事や百科事典等の約 250 万文をモデル構築に使用している．

推定モデル構築後は，そのモデルを使用して冠詞や前置詞の推定が可能となる．誤りを検出したい入力文の特徴を利用することで，その素性に相応しい前置詞や冠詞をスコア（確率値）付きで出力することができる．

■言語モデルによる選別

Gamon の手法では，誤り検出部での出力を最終的な結果とはせず，さらに言語モデルによる選別処理を行っている．

ここでいう言語モデルとは単語列の確率分布を表している．つまり，ある単語列がどのくらいの確率で出現するかといった情報が収められている．この値が大きいということは，英語として流暢な単語の並びであることを意味する．ただ，あまりに長い単語列は出現頻度がきわめて低く，信頼できる確率の計算が不可能である．そのため，Gamon の手法では連続する 7 単語までを用いる単語 7-gram を言語モデルとして用いている．単語 7-gram の場合，ある単語の出現確率は直前の 6 単語に依存するという仮定のもとで，単語の出現確率が計算される．

言語モデルは，先ほど述べた冠詞および前置詞モデルの構築に用いた英文よりも，さらに大量の英文（10 億語規模）から構築される．誤り検出部による校正候補によって書き換えられた入力文は，書き換え前の入力文とともに，この言語モデルを用いた確率値に基づくスコア付けが行われる．このように計算された言語モデルのスコア値，および冠詞／前置詞の推定モデルのスコア値から，機械学習を用いた校正候補の選別を行う．以上の処理を行うことで，より精度の高い校正候補だけをユーザに見せることが可能になる．

■性能の評価

Gamon は約 6,000 文（その半数に 1 つ以上の冠詞・前置詞誤りを含む）を対象に性能評価を行っている．それによれば，誤りを検出して校正結果も正解と一致したものは，冠詞・前置詞ともに約 33%となり，誤り検出のみの精度は前置詞で 85%，冠詞で 76%と報告されている．正しい校正候補を一意に決めることは難しいことが分かる．

動詞の時制の校正手法

日本人英語学習者は冠詞や前置詞の用法が苦手だと冒頭で述べたが，それらに次いで動詞の時制誤りも少なくない．本章では，動詞の時制誤りを校正する手法の 1 つとして，永田らの手法[4] を紹介する．

図 -5 に，永田らの手法の基本的な考え方を示す．図 -5 では，まず動詞が状態動詞と動作動詞の 2 つに分類できることを表している☆2．動作動詞と現在時制の組合せは通常，発話時点で起こっている動

☆2 *stativity* と呼ばれる動詞の性質に基づいた分類．

図 -6　動詞の分類に用いる特徴の例

作を表現するのには用いられない．そのため，動詞が動作動詞でかつ現在形で用いられている場合にそれを誤りと判定することで，学習者の基本的な時制誤りを検出することができる．

問題は動詞が状態動詞なのか動作動詞なのかを判定することであるが，永田らは前章で述べた最大エントロピー分類器を用いて，動詞の判定を行っている．その際に用いている特徴は図 -6 に示されるように，動詞を基準とした左右の単語，および副詞としている．動詞の分類モデルを構築するための訓練データにはこれらの特徴に加え，その特徴に応じた動詞の分類情報も必要となる．永田らはそのような訓練データを人手で作成している．

永田らは訓練データとして約 2,500 事例を使用し，本稿の冒頭で紹介した KJ learner corpus 中の時制誤り 155 個を対象に性能評価実験を行っている．その結果，誤り検出の精度は約 55%と報告されている．

英語学習支援への応用

これまで，日本人英語学習者が誤りがちな文法項目の誤り校正手法について述べてきた．実際に使用できる英語学習支援システムを考えたときに，これら誤り校正手法を組み合わせたり発展させたりすることで，ユーザが型にはまらない文章を入力できるという点において，自由度の高い学習支援システムが実現できる．

冠詞・前置詞の校正手法の紹介で述べた Gamon の手法を実装している Microsoft Research ESL Assistant は，統計的手法だけでなくルールに基づく手法も取り入れたさまざまな種類の誤り校正モジュールを組み合わせて構成されている☆3．ESL Assistant は，入力文の校正結果だけでなく，校正前・校正後の各事例が含まれる例文を出力する機能も持つ．この機能は，誤りの判定が微妙である場合には，ユーザは例文を判断材料とすることができる．また，校正結果の実例と併せ見ることで，より深い用法の理解につながりやすくなるメリットがあると考えられる．

このように，校正結果とともに例文を出力するシステムはいくつか提案されている．筆者が提案した手法[5]を実装した冠詞・前置詞誤り校正システム☆4においても，図 -7 に示すように校正結果と例文を同時に表示している．図 -7 において，入力文は最上部の文で，赤色で示される項目はシステムが誤りで

☆3 http://www.eslassistant.com/
Web インタフェースの公開は，残念ながら 2011 年 4 月で終了したとのこと．
☆4 http://hkt.tl.fukuoka-u.ac.jp/

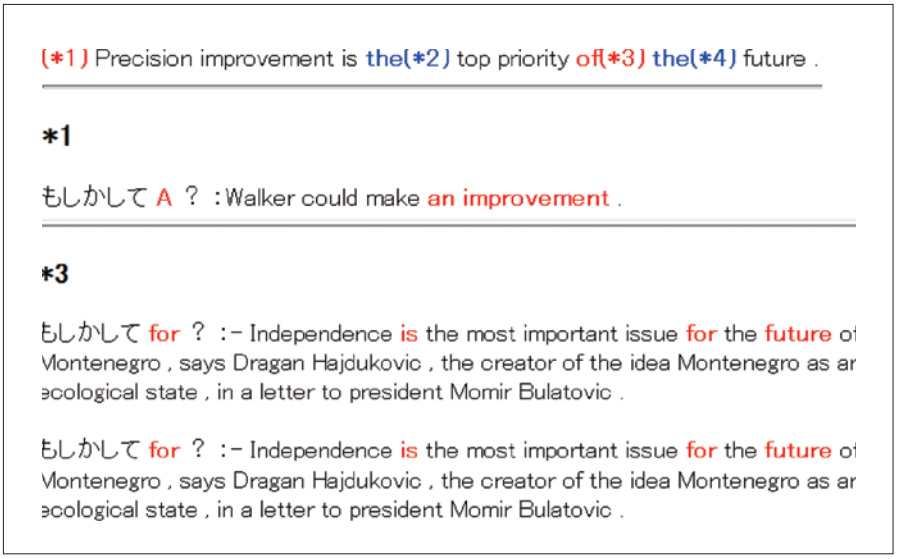

図 -7　校正結果と例文がともに出力される例

あると判断したことを示す．青色で示される項目は，正しい用法と判断していることを意味する．そして検出された各誤りに対して校正候補と例文を示している．

今後の課題

これまで述べたように，さまざまな誤り校正手法が提案されており，それらの応用として英語学習支援システムも開発されている．しかしながら，より実用的な誤り校正や英語学習支援を実現するためには，解決しなければならない課題も少なくない．そのような課題について少しではあるが述べたい．

課題の 1 つとして，誤り検出精度の向上が挙げられる．Gamon の手法の冠詞誤り検出では，冠詞の前後 3 単語を特徴として用いていたが，実際の冠詞の用法はさらに広範囲の文脈に影響される場合がある．たとえば，前の文と同じ名詞が出現し，かつ同じ内容を表している場合は定冠詞 the を用いるといったものである．さらに，名詞が多義語の場合，意味によって可算名詞か不可算名詞かどちらとして用いるかが変わるものがある．意味による影響は前置詞や他の文法項目にも当てはまる．このような用法に対応するために，意味や文脈を考慮することが必要となるが，それらをどう考慮するかが難しい課題である．

また，日本人英語学習者による多種多様な誤りを含む文章を正確に解析できるかといった課題も存在する．多くの誤り校正手法は，前処理として品詞情報抽出に代表される何らかの解析を必要とする．依存構造を抽出するような深い解析をする場合，多様な誤りを含む文を対象とすると正確な解析ができない場合が多い．近年開発された，KJ learner corpus のように誤り情報と品詞・句構造情報が付加された言語資源が準備されてきており，誤りを含む文章の解析精度向上が期待されている．

以上，簡単ではあるが，英語学習支援と誤り校正手法について説明した．より詳しい内容は文献 6）がまとめている．興味を持たれた方は，参考にしていただきたい．

参考文献

1）Nagata, R., Whittaker, E. and Sheinman, V. : Creating a Manually Error-tagged and Shallow-parsed Learner Corpus, Proceedings of the 49th ACL, pp.1210-1219 (2011).
2）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，情報処理学会論文誌，Vol.25, No.6, pp.1072-1079 (June 1984).
3）Gamon, M. : Using Mostly Native Data to Correct Errors in Learners' Writing : A Meta-Classifier Approach, Proceedings of NAACL 2010, pp.163-171 (2010).
4）永田　亮，Sheinman, V. : Stativity 判定に基づいた時制誤り検出，言語処理学会第 17 回年次大会発表論文集，pp.1055-1058 (2011).
5）Ototake, H. and Araki, K. : English Article Correction System Using Semantic Category Based Inductive Learning Rules, Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI), 5866, pp.597-606 (2009).
6）Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J. : Automated Grammatical Error Detection for Language Learners, Morgan & Claypool Publishers, San Francisco (2010).

（2011 年 11 月 17 日受付）

乙武 北斗（正会員）ototake@fukuoka-u.ac.jp

福岡大学工学部助教．2010 年北海道大学大学院情報科学研究科博士後期課程修了．同年より現職．博士（情報科学）．主に自然言語処理に関する研究に従事．