科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２３　日現在

機関番号：　３７１１１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７０１０００

研究課題名（和文）理由提示が可能な日本語・英語文法誤りの自動校正システムに関する研究

研究課題名（英文）A research for correcting grammatical errors automatically with suggesting the reasons in Japanese and English

研究代表者
乙武　北斗（Ototake, Hokuto）

福岡大学・工学部・助教

研究者番号：２０５８０１７９

交付決定額（研究期間全体）：（直接経費）　　1,100,000円、（間接経費）　　330,000円

研究成果の概要（和文）：申請者は以前より作成していた英語冠詞誤りの自動校正システムの性能向上を目的としてアルゴリズムの再設計を行い，処理時間の短縮および精度の改善を行った．また，英語の動詞人称誤りや前置詞誤りを対象とした校正システムを作成し，競争型ワークショップにて成果を発表した．本成果はオープンソースライセンスで公開されている．
日本語に関しては，公開されている地方議会会議録データを用いて敬語表現をモデル化することで，日本語敬語の自動校正・サジェストの可能性を確認した．

研究成果の概要（英文）：I redesigned the algorithm of my system for correcting English article errors automatically I implemented before starting this project. The result is that the system runs more accurately and quickly. Also, I implemented two error correction systems. One is for personal verb forms, and another is for prepositions. I introduced the systems at a competitive workshop and released them as open source software.
As for Japanese language, I made a model of Japanese honorific expressions from regional assembly minutes automatically. In assemblies in Japan, speakers usually speak using honorific expression. The model can be used for implementing a system that corrects errors of honorific expressions and suggests why an user input is erroneous.

研究分野：　総合領域

科研費の分科・細目：　教育工学

キーワード：　文書校正　教育応用　冠詞誤り　前置詞誤り　敬語

１．研究開始当初の背景

（1）我々が執筆した文章にはしばしば何らかの誤りが含まれる．特に非母語を用いて文章を執筆する際には，文法誤りが発生する頻度が高まる傾向にある．例えば，英語学習者は冠詞や前置詞の誤りが特に多く，日本語学習者は助詞の誤りが多いことが報告されている．

（2）近年のコンピュータの性能向上により，大規模データからの統計的情報を用いて文法誤りを自動的に検出・校正するシステムが提案されている．しかしながら，それらの精度は 100%ではなく，システムの出力結果から新たな誤りが生じる可能性がある．そのため，校正をシステムに完全に任せるのではなく，校正結果とともに理由を提示してユーザの判断を促す必要がある．また，ユーザは一種類の文法項目について誤用するのではなく，多くの場合はいくつかの文法項目を複合的に誤用する．

２．研究の目的

（1）このような背景から，より高精度かつ多様な項目を対象とした文法誤りの自動校正手法の実現を目指す．

（2）また，ユーザにわかりやすい形での校正理由提示を行う手法の実現を目指す．

（3）加えて，個々の文法誤り校正手法をまとめることで，統合的な文法誤り校正支援システムの構築を行う．

３．研究の方法

（1）本研究では英語の文法誤りを対象とした自動校正システムの実現を目指す．研究成果（2）で述べる競争型ワークショップでテストデータとして用いられた日本人英語学習者による誤りを含んだ英文コーパス（Konan-JIEM learner corpus）を使用し，どのような誤りが含まれるかを調査した結果，図１のような結果となった．本研究で対象とする英語の文法項目は，図１による日本人英語学習者が起こしやすいとされているものを重要視し，以下の項目を対象とした校正アルゴリズムを作成する．

① 冠詞誤り
② 単数・複数形の誤用
③ 前置詞誤り
④ 動詞の人称表現の誤用
⑤ 形容詞–名詞のつながり誤り

（2）本研究では日本語における文法誤りについても校正システムの構築を目指す．対象とする文法項目は助詞誤りのほか，日本語において特徴的な敬語表現や同音異義語誤りとする．

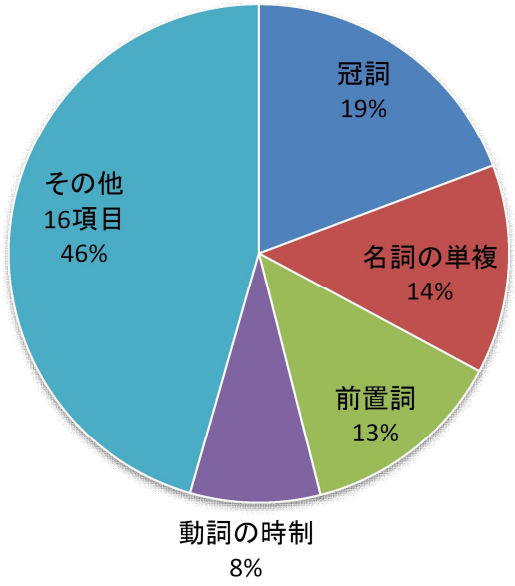

図 1　英語誤りの分析結果

４．研究成果

（1）申請者が以前より開発していた英語冠詞誤りの自動校正システムについて，性能向上を図ることを目的としてアルゴリズムの再設計を行った．その際，計算用のサーバマシンを利用し，従来よりも冠詞推定のための規則の抽出量を大幅に増加させた上で，システムの構築を行った．また，冠詞推定の規則には従来よりも深い意味解析による単語意味情報を付与し，精度の改善を図った．具体的には，表１に示すような素性を冠詞推定に用いた．性能評価実験の結果，従来よりもデータ量を増加させたことによる優位性は認められなかった．しかしながら，冠詞推定規則に意味情報を付与したことによって，定冠詞の推定性能の向上（Precision が 60%で従来と比較して 2 ポイントの改善）を確認した．また，大量の規則の保存・検索を行うために高速ストレージデバイスを用いたことで，従来のシステムよりも実用的な時間で校正結果を出力することが可能となった．

（2）申請者は英語の動詞人称誤りおよび前置詞誤りの自動校正に関する競争型ワークショップに参加し，成果発表および成果物の公開を行った．従来までは英語冠詞誤りに着目していたが，本ワークショップへの参加を通して英語の動詞変化と前置詞の誤りに対応する校正手法の開発を行った．動詞の人称誤り校正手法に関しては，申請者が過去に開発した経験がなかったため，一からルールベースの実装を行った．しかしながら，F 値が 31.4%という結果となり，多様な構文への対応に課題が残る結果となった．前置詞誤り校正手法に関しては，申請者が過去に開発した手法を知見として，新たに固有表現抽出の結果を取り入れた機械学習ベースの手法を実装した．その結果，F 値が 30.5%という結果となり，絶対値こそ動詞の人称誤りの結果よ

表 1　冠詞推定規則に用いる素性リスト

分類	素性名
対象名詞	主名詞 Synset
	単数／複数
	一般／固有名詞
	主名詞
	主名詞以外の名詞
	句の種類
	前置詞
	前に位置する動詞
	前に位置する動詞 Synset
	後ろに位置する動詞
	後ろに位置する動詞 Synset
前置修飾	修飾詞 Synset
	品詞
	所有格をもつ
	修飾詞
後置修飾	名詞句直後の語が属する句
	名詞句直後の語
	名詞句直後の語の品詞
	修飾句
	修飾句の主名詞
	修飾句の主名詞 Synset
	修飾句の主名詞以外の名詞
	修飾句の前置修飾詞
	修飾句の前置修飾詞 Synset
	修飾句の前置修飾詞の品詞
既出	既出かどうか

りも低いものの，ワークショップ参加チームの中では第 2 位となり，比較的高精度な結果となった．本ワークショップのタスクでは，テストデータとして実際に日本人英語学習者が執筆した，誤りを含む英文を用いた．そのため，本タスクで対象の 2 種類の誤り以外にも様々な誤りが複合的に含まれている．そのような英文を対象に高精度な誤り検出を行うにはまだまだ課題が多いこと，および実用のためにはその課題を解決しなければならないことを改めて認識できる良い機会となった．本ワークショップに向けて開発したシステムは，以下の URL よりソースコードを入手可能である．
https://sites.google.com/site/edcw2012/

（3）申請者は議会会議録から日本語の敬語表現をモデル化する手法を提案した．本成果

表 2　敬語モデルの素性リストの例

用言の基本形	“思う”
用言が文末に位置するか？	Yes
用言の品詞	動詞
用言の時制	過去ではない
ト格	“申し上げ”
主格	一人称

の段階では誤り校正までは実現しておらず，その前段階である敬語表現のモデル化までが実現済みである．使用した素性は以下のとおりである．

- 用言の基本形
- 用言が文末に位置するかどうか
- 用言の品詞
- 用言の時制
- 様相（疑問，依頼など）
- ガ格
- ガ格に敬称があるかどうか
- ヲ格
- ヲ格に敬称があるかどうか
- ニ格
- ニ格に敬称があるかどうか
- ト格
- デ格
- カラ格
- ヨリ格
- 主格

一方，敬語表現を表すラベルは，対象とする用言の形態素解析結果や語尾表現から推定されたものとする．例えば，文「そのことをまずお詫びを申し上げたいと思います」における素性値は表 2 に示すものになり，対象用言「思う」の語尾に「ます」があるため，ラベルは謙譲語となる．モデルの妥当性評価実験の結果，公開されている地方議会会議録データから自動的に作成した敬語表現のモデルが有効である可能性を見出すことができた．また，地方のデータを用いることから，地方独特の敬語表現に対応できる利点がある．しかしながら，方言やオノマトペなどは形態素解析や構文解析において正しく判別できず，誤った解析結果を出力する場合が多い．正しい解析結果をどのようにして得るかが課題として残る．

（5）申請者は英語冠詞について，特に固有名詞に付与される定冠詞に着目した研究成果の発表を行った．多くの固有名詞は冠詞を伴わずに用いると思われがちだが，実際には定冠詞 the が付与される要因が存在する．このような the＋固有名詞を正確に推定することも，定冠詞 the 全体の推定性能を左右する要因となる．固有名詞とその周辺の文脈の特徴を利用した機械学習ベースの手法を適用し，性能の向上を確認した．また，成果発表の際には Wikipedia を情報源として用いる追加実験についても述べた．追加実験の結果，Wikipedia の情報も併せて利用することで，

より多くの固有名詞に対する定冠詞の有無を判別可能になった一方で，判別結果の信頼性の低下が確認された．具体的には，Wikipediaの利用によりRecallが3ポイント向上したものの，Precisionが13ポイント低下する結果となった．ただし，本成果発表時のWikipedia情報の抽出方法は，Wikipediaから対象固有名詞の項目を検索し，見つかった場合は冒頭の説明文で用いられている冠詞の用法を推定結果として用い，見つからなかった場合は機械学習による判別結果を用いるという非常にシンプルなものである．そのため，Wikipediaに記載されている本文全体を考慮するような方法を用いることで，評価結果が変わる可能性がある．また，機械学習による判別結果とWikipediaによる推定結果を適切にマージする方法を考える必要がある．

（6）今後の展望として，研究成果として得られた知見や実装システムをまとめ，種々の誤りに対応した統合的な誤り校正システムをWeb上に公開することが挙げられる．また，本成果の（2）で述べた競争型ワークショップでテストデータとして用いられた日本人英語学習者による実際の誤りを含む英文を材料として，校正システムの実用化に必要な精度向上や複合的な誤りへの対応方法を検討する必要がある．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計０件）

〔学会発表〕（計７件）

① 乙武 北斗，吉村 賢治，英文における固有名詞を対象とした定冠詞の自動付与手法の提案，第２９回ファジィシステムシンポジウム，２０１３年９月９日～２０１３年９月１１日，大阪
② Hokuto Ototake and Kenji Yoshimura, Development and Evaluation of a Model for Japanese Honorific Expressions Using Assembly Minutes, 2012 International Conference on Asian Language Processing，査読あり，２０１２年１１月１３日～２０１２年１１月１５日，Hanoi，Vietnam
③ 乙武 北斗，ハンドメイドルールシステム（動詞誤り検出トラック），誤り検出・訂正ワークショップ２０１２（EDCW2012），２０１２年９月３日，仙台
④ 乙武 北斗，ME分類ベースシステム（前置詞誤り検出トラック），誤り検出・訂正ワークショップ２０１２（EDCW2012），２０１２年９月３日，仙台
⑤ 乙武 北斗 他，事例の自動抽象化に基づくルールを用いた英語冠詞誤りの自動付与手法の提案，言語処理学会第１８回年次大会，２０１２年３月１３日～２０１２年３月１６日，広島

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等
福岡大学研究者情報
http://resweb2.jhk.adm.fukuoka-u.ac.jp/FukuokaUniv/R101J_Action.do

６．研究組織
(1)研究代表者
乙武　北斗　（OTOTAKE，Hokuto）
福岡大学・工学部・助教
研究者番号：２０５８０１７９