博 士（情報科学） 乙 武 北 斗

学 位 論 文 題 名

英文を対象とした誤りの自動校正手法に関する研究

学位論文内容の要旨

近年，コンピュータの性能は著しく向上しており，それに伴いより大規模なデータをより高速に処理することが可能となりつつある．自然言語処理技術においても，大規模なデータから得られる統計量に基づいた手法が，従来の人手によって作成された規則に基づく解析的手法と比較して成果を上げている．しかしながら，統計的言語処理はその特性上，大量かつ多様なデータを必要とする．我々が普段用いている言葉は一般的な文法規則は存在するものの，実際の用法の明確な規則が存在しなかったり例外的用法が多く存在したりする．このような言葉を研究対象とした場合，統計的言語処理を用いてもその多様な言語現象に対して満足できる性能を得られるとは限らない．これは統計的言語処理が学習データに近いデータに対しては高精度な処理が可能であるが，学習データと類似していないデータに対しては低い精度しか得られないという特徴によるものである．

本研究は，英文を対象とした文法誤りの自動校正手法について，母国語話者が執筆した大規模英語テキストコーパスから自動的に獲得されるルールに基づいて誤りの校正を行うことにより，高精度の誤り自動校正システムを実現することを目的としている．また，著者が独自に提案した，特徴スロットを用いた帰納的学習処理を行うことで，抽象化したルールを自動生成することにより，ルールの汎用性向上を図っている．ここで特徴スロットでは校正対象の用法に影響する文脈情報を要素としている．

第一に，日本人英語学習者が起こしやすい誤りの一つである英語の冠詞誤りを，単語出現状況から帰納的学習を用いて検出及び自動校正する手法の提案を行った．本手法では，電子化コーパス中の英文における名詞句とその周辺の単語を特徴として抽出し，冠詞と組み合わせて冠詞選択ルールとする．本手法における特徴とは，対象名詞の単語や属する句の情報，修飾語句の単語や品詞を要素として持つ特徴スロットのことを指す．このような処理によって，文内の文脈を考慮したルールを獲得することが可能となる．また，特徴スロットを用いた帰納的学習を用いて，抽出されたルール同士から抽象化したルールを新たに自動生成する．次に獲得されたルールに基づいて，冠詞誤りの検出・校正を行う．

本手法の利点としては，まずルールを人手で作成する労力を必要としない点が挙げられる．また，ルールの抽象化を行うことで，冠詞選択に関わる文脈要素を絞り込むことが可能となる．性能評価実験の結果，誤り検出において最も性能が良好なもので，検出の正解率を表す Precision が 67.0%，検出の網羅率を表す Recall が 35.0% となった．また，関連手法と同一の実験データで比較したところ，Precision において関連手法よりも 5 ポイント高い 70.0% という優位性のある結果を示し，本手法の有効性を確認することができた．

前述した冠詞誤り校正手法では, ルールを自動生成するための学習処理に時間を要するため, あらかじめ用意したトレーニングデータ全てからルールを生成することは計算時間の点から現実的には非常に困難であった. そこで, 第二に, ユーザの入力文に応じた適切な量のトレーニングデータを大規模コーパスから自動的に抽出する手法の提案を行った. 本システムは冠詞誤りを含むユーザ入力文を受け取った際に, 入力文中に含まれる名詞や形容詞をクエリとして, コーパスからトレーニングデータを検索する. その結果, ユーザの入力文に適したトレーニングデータが抽出され, それらを用いて誤り校正ルールの生成を行う. 性能評価実験の結果, 誤り検出・校正において 35 ポイントの性能向上を確認した. また,Precision も若干の向上が見られ, 提案したトレーニングデータ抽出手法が著者の提案する冠詞誤り校正手法の性能向上に有効であることを確認した.

第三に, 上記の冠詞誤り校正手法で用いたアルゴリズムをベースに, 英文における前置詞誤りを対象とした誤り自動校正手法の提案を行った. 冠詞と同様, 前置詞の正しい利用も日本人英語学習者にとって難しい問題の一つである. 冠詞誤り校正システムにおけるルールでは, 名詞句とその周辺の単語を要素としていたが, 前置詞誤り校正システムにおけるルールでは, 前置詞とその前後にある句を要素とする. 性能評価実験の結果, 最も誤り校正精度が高かったパラメータの組み合わせで 82.3% の精度,31.3% の網羅率であり, 関連手法の精度 82.1%, 網羅率 14.1% と比較しても同等以上の結果であることが明らかとなった. また, 著者が提案したルール学習アルゴリズムが, 英語冠詞以外に前置詞誤りの校正に対しても有効であることが確認された.

第四に, 著者が提案したルール学習手法をベースに単語の意味情報を考慮した新たな学習手法と, その学習手法を利用した冠詞誤りの自動校正手法の提案を行った. 本システムにとって未知の名詞について冠詞の用法を判断する際に, 同じ意味情報を持つ名詞についてのルールを既に持っていた場合, そのルールを未知名詞に対して適用させることで, 高い精度で誤り校正を行うことが可能であると考えられる. 性能評価実験の結果,6 ポイントの Precision の向上が見られ, 意味情報は Recall を低下させることなく Precision 向上に有効であることが確認された.

また, これまで述べた冠詞・前置詞誤り校正システムを Web アプリケーションとして公開し, 約 300 名の大学生を対象にシステムを実際に利用する実験を行った. そのうえで被験者に対して, システムに関するアンケートを行ったところ, 実用のためには処理時間や校正精度の課題が残るものの, 英文執筆の際には非常に有用なシステムであるという意見を得られた. しかしながら, 最も多かった意見は処理時間の改善に関することであり, 実用的な誤り校正システムの実現のためには, 処理時間改善は優先的に取り組むべき課題であることが確認された.

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　山　本　　強
副　査　教　授　長谷山　美　紀

学位論文題名

英文を対象とした誤りの自動校正手法に関する研究

近年，コンピュータの性能は著しく向上しており，それに伴いより大規模なデータをより高速に処理することが可能となりつつある．自然言語処理技術においても，大規模なデータから得られる統計量に基づいた手法が，従来の人手によって作成された規則に基づく解析的手法と比較して成果を上げている．しかしながら，統計的言語処理はその特性上，大量かつ多様なデータを必要とする．我々が普段用いている言葉は一般的な文法規則は存在するものの，実際の用法の明確な規則が存在しなかったり例外的用法が多く存在したりする．このような言葉を研究対象とした場合，統計的言語処理を用いてもその多様な言語現象に対して満足できる性能を得られるとは限らない．これは統計的言語処理が学習データに近いデータに対しては高精度な処理が可能であるが，学習データと類似していないデータに対しては低い精度しか得られないという特徴によるものである．

本研究は，英文を対象とした文法誤りの自動校正手法について，母国語話者が執筆した大規模英語テキストコーパスから自動的に獲得されるルールに基づいて誤りの校正を行うことにより，高精度の誤り自動校正システムを実現することを目的としている．また，著者が独自に提案した，特徴スロットを用いた帰納的学習処理を行うことで，抽象化したルールを自動生成することにより，ルールの汎用性向上を図っている．ここで特徴スロットでは校正対象の用法に影響する文脈情報を要素としている．

第一に，日本人英語学習者が起こしやすい誤りの一つである英語の冠詞誤りを，単語出現状況から帰納的学習を用いて検出及び自動校正する手法の提案を行った．本手法では，電子化コーパス中の英文における名詞句とその周辺の単語を特徴として抽出し，冠詞と組み合わせて冠詞選択ルールとする．本手法における特徴とは，対象名詞の単語や属する句の情報，修飾語句の単語や品詞を要素として持つ特徴スロットのことを指す．このような処理によって，文内の文脈を考慮したルールを獲得することが可能となる．また，特徴スロットを用いた帰納的学習を用いて，抽出されたルール同士から抽象化したルールを新たに自動生成する．次に獲得されたルールに基づいて，冠詞誤りの検出・校正を行う．

本手法の利点としては，まずルールを人手で作成する労力を必要としない点が挙げられる．また，ルールの抽象化を行うことで，冠詞選択に関わる文脈要素を絞り込むことが可能となる．性能評価実験の結果，誤り検出において最も性能が良好なもので，検出の正解率を表す Precision が 67.0%，検出の網羅率を表す Recall が 35.0% となった．また，関連手法と同一の実験データで比較したところ，Precision において関連手法よりも 5 ポイント高い 70.0% という優位性のある結果を示し，本手

法の有効性を確認することができた.

前述した冠詞誤り校正手法では,ルールを自動生成するための学習処理に時間を要するため,あらかじめ用意したトレーニングデータ全てからルールを生成することは計算時間の点から現実的には非常に困難であった.そこで,第二に,ユーザの入力文に応じた適切な量のトレーニングデータを大規模コーパスから自動的に抽出する手法の提案を行った.本システムは冠詞誤りを含むユーザ入力文を受け取った際に,入力文中に含まれる名詞や形容詞をクエリとして,コーパスからトレーニングデータを検索する.その結果,ユーザの入力文に適したトレーニングデータが抽出され,それらを用いて誤り校正ルールの生成を行う.性能評価実験の結果,誤り検出・校正において 35 ポイントの性能向上を確認した.また,Precision も若干の向上が見られ,提案したトレーニングデータ抽出手法が著者の提案する冠詞誤り校正手法の性能向上に有効であることを確認した.

第三に,上記の冠詞誤り校正手法で用いたアルゴリズムをベースに,英文における前置詞誤りを対象とした誤り自動校正手法の提案を行った.冠詞と同様,前置詞の正しい利用も日本人英語学習者にとって難しい問題の一つである.冠詞誤り校正システムにおけるルールでは,名詞句とその周辺の単語を要素としていたが,前置詞誤り校正システムにおけるルールでは,前置詞とその前後にある句を要素とする.性能評価実験の結果,最も誤り校正精度が高かったパラメータの組み合わせで 82.3% の精度,31.3% の網羅率であり,関連手法の精度 82.1%,網羅率 14.1% と比較しても同等以上の結果であることが明らかとなった.また,著者が提案したルール学習アルゴリズムが,英語冠詞以外に前置詞誤りの校正に対しても有効であることが確認された.

第四に,著者が提案したルール学習手法をベースに単語の意味情報を考慮した新たな学習手法と,その学習手法を利用した冠詞誤りの自動校正手法の提案を行った.本システムにとって未知の名詞について冠詞の用法を判断する際に,同じ意味情報を持つ名詞についてのルールを既に持っていた場合,そのルールを未知名詞に対して適用させることで,高い精度で誤り校正を行うことが可能であると考えられる.性能評価実験の結果,6 ポイントの Precision の向上が見られ,意味情報は Recall を低下させることなく Precision 向上に有効であることが確認された.

また,これまで述べた冠詞・前置詞誤り校正システムを Web アプリケーションとして公開し,約 300 名の大学生を対象にシステムを実際に利用する実験を行った.そのうえで被験者に対して,システムに関するアンケートを行ったところ,実用のためには処理時間や校正精度の課題が残るものの,英文執筆の際には非常に有用なシステムであるという意見を得られた.しかしながら,最も多かった意見は処理時間の改善に関することであり,実用的な誤り校正システムの実現のためには,処理時間改善は優先的に取り組むべき課題であることが確認された.

これを要するに,著者は,英語冠詞および前置詞誤りの自動校正において,自動獲得されるルールを用いた手法により高精度な誤り校正を実現するとともに,より実用的な誤り校正システムのための有益な知見を得ており,自然言語処理に関する学術分野に貢献するところ大なるものがある.よって著者は,北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.