学位論文題名

Normalization and Generation of Social Media Casual English using a Lexical Base to Support Communication

(コミュニケーション支援のためのソーシャルメディアにおける英語スラング表現を対象とした語彙ベースに基づくテキスト標準化および生成)

学位論文内容の要旨

近年のインターネットの利用の急速な普及に従い、ソーシャルネットワークサイトやブログ、マイクロブログなどメディアのユーザが急増している。これらのユーザは文法・綴り・句読点などの規則から逸脱したカジュアルな英語スラング表現を使用することが多いため、これらのテキストを理解する必要性が増大しつつある。このような状況に対して、カジュアルで例外的な英語スラング表現の標準化が研究の対象となっていることは非常に少なく、これまで自然言語処理の分野では相当労力を必要とする仕事とみなされてきた。関連研究としては全自動、統計的な手法が圧倒的に多く見られる一方、本学位論文では手動と自動の両方を融合した手法を用い、この問題においてはこのような手法がより適切な手法であることを実験により実証している。それゆえ、本研究の目的は自動的トークン化、マッチングや置き換えなどの要素手法と、手動で作成した高品質で網羅性の高いデータベースと組み合わせたシステムに用いられる独自の知見を多くの人々が使えるコミュニケーション支援のための、フレーズを含む語彙データベースとして得ることである。

本学位論文では、独自の語彙ベースを応用したカジュアルな英語スラング表現のテキスト処理を研究課題とし、テキストの標準化および生成を行った結果について述べる。まず自動標準化のための実験システム (CECS: Casual English Conversion System) の提案を行った。CECS の応用は 2 つある。一つは機械翻訳や情報抽出などの自然言語処理タスクにおいてノイズの多いデータの前処理システムである。もう一つは、現在の国境を越えたコミュニケーションへの積極的な参加が困難である実質的に国際標準語となっている英語以外が母国語であるユーザのために、ソーシャルメディアにおける英文の読書理解力を支援するシステムである。CECS を用いて行った評価実験は機械翻訳の前処理実験とコミュニケーションのための評価実験の 2 種類である。機械翻訳の前処理実験では、CECS を用いて前処理を行い、Twitter 上の文章を機械翻訳の入力として評価を行ったが、前処理を行わなかった場合と比較すると前処理後の翻訳されていないトークンが平均 74.3% 減少し、翻訳が誤ったトークンが平均 37.2% 減少した。また、コミュニケーションのための評価実験では、被験者の理解度が SD 法 (Semantic Differential Method: 意味差別法) の 5 段階のスケールにおいて平均 1 ポイント向上した。これらの実験結果から、提案システムの有効性と必要性が確認された。

カジュアルな英語スラング表現の普及に伴う問題として、一般のユーザにとってその文章の理解が困難というのはもちろんだが、もう一つの問題としてその文章に多く見られる独自性のあるスペルを作成し、文章中に入れること自体も一般のユーザにとって、簡単ではないと考えられる。そのため、本研究では標準化だけでなく、さらに、英語スラング表現を自動的に生成するシステムの提案を行い、独自の音節データベースを応用する実験システム (CEGS: Casual English Generation System) の開発を行った。人間のコミュニケーションを支援するという最終的な目的の基に CEGS は CECS

という双方向の機能を有するシステムの開発を行った。本実験システムの科学的な目的は音節ベースを用いることにより人間の独創的の文章創造能力を代行することができるかという人工知能的なものである。また、工学的側面としては、ソーシャルメディアにおいてカジュアルな表現を書くことが苦手な世代を対象としたカジュアル表現の自動生成システムや、テキストを用いる会話システムの文章表現の自然性を向上させることなどが考えられる。CEGS の評価実験を行なった結果、被験者 50 名に対して行なったチューリングテスト型の実験では、CEGS の出力文の人間らしさが人間の書いた文よりやや低く評価された。しかし、この差は 5 点スケールで平均 0.6 点という小さなものであった。また、CEGS の出力文の中で、実際に人間の書いた文より高く評価された事例もあった。この結果から、Twitter などの文章に似た英語のスラング表現を生成するタスクにおいて、CEGS が有効であることが確認された。

英文の電子コミュニケーションを支援するための語彙データベースを応用したテキストの標準化及び生成システムの実験結果より本学位論文で提案する手法が、ソーシャルメディア等に頻繁に出現する英語スラング表現を処理するために有効であり、英語力が低い人や年配の人などを含むユーザのコミュニケーションのサポートを行うためのデータベースの開発に貢献することを確認した。

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

Normalization and Generation of Social Media Casual English using a Lexical Base to Support Communication

(コミュニケーション支援のためのソーシャルメディアにおける英語スラング表現を対象とした語彙ベースに基づくテキスト標準化および生成)

近年のインターネットの利用の急速な普及に従い、ソーシャルネットワークサイトやブログ、マイクログログなどメディアのユーザが急増している。これらのユーザは文法・綴り・句読点などの規則から逸脱したカジュアルな英語スラング表現を使用することが多いため、これらのテキストを理解する必要性が増大しつつある。このような状況に対して、カジュアルで例外的な英語スラング表現の標準化が研究の対象となっていることは非常に少なく、これまで自然言語処理の分野では相当労力を必要とする仕事とみなされてきた。関連研究としては全自動、統計的な手法が圧倒的に多く見られる一方、本学位論文では手動と自動の両方を融合した手法を用い、この問題においてはこのような手法がより適切な手法であることを実験により実証している。それゆえ、本研究の目的は自動的トークン化、マッチングや置き換えなどの要素手法と、手動で作成した高品質で網羅性の高いデータベースと組み合わせたシステムに用いられる独自の知見を多くの人々が使えるコミュニケーション支援のための、フレーズを含む語彙データベースとして得ることである。

本学位論文では、独自の語彙ベースを応用したカジュアルな英語スラング表現のテキスト処理を研究課題とし、テキストの標準化および生成を行った結果について述べる。まず自動標準化のための実験システム (CECS: Casual English Conversion System) の提案を行った。CECS の応用は 2 つある。一つは機械翻訳や情報抽出などの自然言語処理タスクにおいてノイズの多いデータの前処理システムである。もう一つは、現在の国境を越えたコミュニケーションへの積極的な参加が困難である実質的に国際標準語となっている英語以外が母国語であるユーザのために、ソーシャルメディアにおける英文の読書理解力を支援するシステムである。CECS を用いて行った評価実験は機械翻訳の前処理実験とコミュニケーションのための評価実験の 2 種類である。機械翻訳の前処理実験では、CECS を用いて前処理を行い、Twitter 上の文章を機械翻訳の入力として評価を行ったが、前処理を行わなかった場合と比較すると前処理後の翻訳されていないトークンが平均 74.3% 減少し、翻訳が誤ったトークンが平均 37.2% 減少した。また、コミュニケーションのための評価実験では、被験者の理解度が SD 法 (Semantic Differential Method: 意味差判別法) の 5 段階のスケールにおいて平均 1 ポイント向上した。これらの実験結果から、提案システムの有効性と必要性が確認された。

カジュアルな英語スラング表現の普及に伴う問題として、一般のユーザにとってその文章の理解が困難というのはもちろんだが、もう一つの問題としてその文章に多く見られる独自性のあるスベ

ルを作成し、文章中に入れること自体も一般のユーザにとって、簡単ではないと考えられる。そのため、本研究では標準化だけでなく、さらに、英語スラング表現を自動的に生成するシステムの提案を行い、独自の音節データベースを応用する実験システム (CEGS: Casual English Generation System) の開発を行った。人間のコミュニケーションを支援するという最終的な目的の基に CEGS は CECS という双方向の機能を有するシステムの開発を行った。本実験システムの科学的な目的は音節ベースを用いることにより人間の独創的文章創造能力を代行することができるかという人工知能的なものである。また、工学的側面としては、ソーシャルメディアにおいてカジュアルな表現を書くことが苦手な世代を対象としたカジュアル表現の自動生成システムや、テキストを用いる会話システムの文章表現の自然性を向上させることなどが考えられる。CEGS の評価実験を行なった結果、被験者 50 名に対して行なったチューリングテスト型の実験では、CEGS の出力文の人間らしさが人間の書いた文よりやや低く評価された。しかし、この差は 5 点スケールで平均 0.6 点という小さなものであった。また、CEGS の出力文の中で、実際に人間の書いた文より高く評価された事例もあった。この結果から、Twitter などの文章に似た英語のスラング表現を生成するタスクにおいて、CEGS が有効であることが確認された。

英文の電子コミュニケーションを支援するための語彙データベースを応用したテキストの標準化及び生成システムの実験結果より本学位論文で提案する手法が、ソーシャルメディア等に頻繁に出現する英語スラング表現を処理するために有効であり、英語力が低い人や年配の人などを含むユーザのコミュニケーションのサポートを行うためのデータベースの開発に貢献することを確認した。

これを要するに、著者は、ソーシャルメディア等に頻繁に出現する英語スラング表現を処理するためのテキスト処理手法を提案し、その標準化と生成について実験システムを開発し、評価実験によって本手法の有効性を示したものであり、自然言語処理工学の発展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。