別紙２

論文審査の結果の要旨

論文提出者氏名　増山毅司

本学位請求論文は日本語において外来語の表記のために用いられるカタカナ表記が揺れる、すなわち同じ意味であるにもかかわらず異なる表記（以下、これをカタカナ異表記と呼ぶ）になる現象が引き起こす問題の解決方法に関するものである。

カタカナ異表記が存在することは、情報検索、日本語から他の言語への機械翻訳、文書分類など多岐にわたる日本語を扱う応用分野において根本的な問題である。例えば、「ロシア」と「ロシヤ」を同じ意味を示すカタカナ異表記であることが認識できないと、情報検索などの精度は劣化する。本論文はこの問題を人間の言語的直感に頼らず、計算機プログラムによって解決する手法を提案したもので、７章からなる。

第１章では、カタカナ異表記の問題提起を行っている。第２章では関連研究を概観している。第３章では、ふたつのカタカナ語が異表記であるか否かを判定するための尺度となる表記ペナルティをシステマティックな統計処理によって求める方法を提案しており、本論文の主要な提案のひとつである。まず、Web上の英日辞書等を検索エンジンで検索し、一つの英単語から派生したカタカナ表記を収集する。次に収集したカタカナ異表記を統計的に処理して、二つのカタカナ表記が「ア」と「ァ」、「ヴァ」と「バ」などの文字列を含むとき異表記になるかどうかの尺度となる表記ペナルティと呼ぶ文字列ペア間の重みを導出した。第４章は、二つのカタカナ語の異表記か否かの判別を、3章の結果である表記ペナルティとそれらのカタカナ語の出現する文脈の類似度を併用して行う方法を提案している。この方法によって、新聞記事38年分のコーパスから得たカタカナ語の異表記を高い精度で認識することに成功した。第5章は提案手法の実験的評価であり、これまでに例がなかった新聞記事38年分という大規模実データを用いた実験で再現率91．5%、適合率91．7%という結果を示している。これは、既存のワープロや検索エンジンの提供している表記の揺れの訂正機能にくらべてはるかに高い性能であり実用性があることが分かった。第６章は、抽出した映画評論のテキストにカタカナ異表記を応用し、映画ジャンルへの文書分類を行った。この実験においても精度の向上を確認している。第７章はまとめである。

提案した方法は、従来、自然言語処理研究者が人手で行っていた異表記ペナルティの開発および言語学者などが人手で構築していたカタカナ異表記収集の作業を、属人性を排した自動的な方法に改善することに成功しており、今後のカタカナ語の増大にも機動的に対応できる優れた手法である。

これらの研究に関して、申請者は、査読ジャーナル論文、国際会議論文などで発表を行い、高い評価を受けている。

したがって、本審査委員会は博士（学術）の学位を授与するにふさわしいものと認定する。