



HOKKAIDO UNIVERSITY

Title	Automatic Metaphor Detection in Japanese and English Using Machine Learning [an abstract of dissertation and a summary of dissertation review]
Author(s)	Babieno, Mateusz
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第15119号
Issue Date	2022-06-30
Doc URL	https://hdl.handle.net/2115/86388
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Mateusz_Babieno_abstract.pdf, 論文内容の要旨



学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称 博士（情報科学） 氏名 Mateusz Babieno

学 位 論 文 題 名

Automatic Metaphor Detection in Japanese and English Using Machine Learning

（日本語及び英語における機械学習を用いたメタファーの自動検出）

比喩的な表現は日常会話の至る所で見られるにもかかわらず、現在の自然言語処理の分野ではメタファーの処理が困難となっている。その原因の一つは、人間と異なり、コンピュータには曖昧な表現を扱うための背景知識が不足していることである。

世界中の言語における比喩的な表現の裏に同じメタファーが見られるケースが多い。その一例として「良いは上・悪いは下」(GOOD IS UP・BAD IS DOWN) という概念メタファーが挙げられる。「下品」, “a fallen country”, 「墮落した人格」, “I feel down lately”, 「結果が下がった」, “low standards”, 「最低なやつ」などは全てネガティブな感情を持ち、「悪いは下」のメタファーに基づく表現である。そのような表現は言語を問わず日常的に使われ、そのメタファーの普遍性を示していると考えられる。したがって、メタファーをめぐる研究を、1つの対象言語に限定する必要はないと考えている。そこで、本研究では対象言語を日本語及び英語の2ヶ国語とした。

本学位論文の第一章では、自然言語処理の立場より考慮した比喩表現をめぐる研究の重要性を示し、本研究による本研究分野への貢献について述べる。全体を理解しやすくするために、比喩的な言語について述べ、それに関わる術語の定義を挙げる。比喩的な言語の種類を羅列し、その定義の曖昧さを指摘しながら、それぞれの特徴を説明する。

第二章では、メタファーに対する言語哲学・認知言語学・情報科学的なアプローチについて述べ、それぞれの優位性や制限を指摘する。現在に至るまで、各分野において提案されてきた本研究の着想となった関連研究を紹介する。

第三章では、日本語を対象に行ってきた実験について述べる。メタファー・データとして比喩辞典の項目を、その反対であるリテラルなデータとしてウイキペディアの記事などを用いた。評価データとして、それぞれのクラスに属する文が含まれる青空文庫の小説文章を用い、以下に述べる2つの実験を行った。1つ目の実験では、メタファーをめぐる認知言語学の仮説に基づいた特徴量の7種類を使い、それぞれのパフォーマンスへの影響を比較した。2つ目の実験では、各文を32,231次元の単語ベクトルで表し、3つの分類アルゴリズム、すなわち SVM, Naïve Bayes, Random Forest の性能を比較した。

機械学習を用いた二値分類では、両方のクラスを代表する学習データを収集することは重要な第一歩である。ところが、リテラルであるはずのデータに比喩表現も少なからず入っていることを確認した。この原因はメタファーはいかなる使用域にも見られるためである。人手でラベルを付与された大量のデータが必要だと考えられるが、本研究の著者の知る限りでは、メタファー検出のために構築した公開されているそのような日本語のデータセットは実験を行なっていた当時に存在しなかった。その上、提案手法のスコアを他の既存手法のスコアと客観的に比較することが不可能であった。

第四章では、英語を対象に行ってきた実験について述べている。メタファー自動検出研究で現時点で存在する主要なデータセットを3つ使用し、既存手法との比較を行った。その大量の公開されて

いるデータセット (VUAMC, TroFi, MOH-X) にはメタファー・非メタファーという人手で付与したラベルが用意されているため, 世界の研究者の既存手法のスコアと提案手法によるスコアを比較し, その有意性を客観的に確認した. 提案手法は既存アルゴリズムを基に, 理論上適切な改善を 3 種類加え, 最終的にモデルを 3 つ提案した. その 1 つ目はある単語のリテラルな意味を表すよう, その単語の辞書の定義を導入したモデルである. 2 つ目のモデルではそれに加え, ある単語が比喩的に使われているか否かを確認するために, コサイン類似度を使用し, その単語のリテラルな意味と文脈内の意味との差を計るモデルである. 3 つ目のモデルでは他の機械学習手法よりも良質な文ベクトルを出力することが確認されている Sentence-BERT を 1 つ目・2 つ目のモデルで使用した RoBERTa 言語モデルの代わりに使用するモデルである. 合計 13 回の評価実験を通し, そのうち 8 回では, 提案手法が既存手法の F 値を上回ることを確認した. その主な理由は辞書の定義を導入したことによるものであると考えられる.

第五章では, 上述の内容のまとめとそれに基づいた結論や今後の課題について述べる. 英語を対象言語にした本研究の後半において, 辞書の定義が表す単語の基本的な意味と文脈上の意味とのギャップは, その単語の比喩的な使用を示唆することが多いと確認した. それを参考に, 必要に応じて新しいデータセットを用意し, データ・アノテーションを加え, 深層学習に基づいた現在日本語でも利用可能になった最新言語モデルを用い, 日本語に対して実験を再度行うことを今後の課題とする.