JAIST Repository
https://dspace.jaist.ac.jp/

Title	用例のクラスタリングに基づく単語の新語義の発見
Author(s)	田中，博貴
Citation	
Issue Date	2009-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/8107
Rights	
Description	Supervisor:白井　清昭　准教授，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

用例のクラスタリングに基づく単語の新語義の発見

田中 博貴 (0710043)

北陸先端科学技術大学院大学 情報科学研究科

2009 年 2 月 5 日

キーワード: コーパス, 新語義, 辞書, クラスタリング.

本論文では、コーパスから新しい単語の意味を自動的に発見する手法について述べる。ここでは、新しい単語の意味とは、既存の辞書に記載されていない語義を指す。提案手法における処理の流れは以下の通りである。まず、コーパスから単語のインスタンス（用例）をいくつか収集し、クラスタリングの手法を用いて同じ意味を持つインスタンスをまとめてクラスタを作成する。次に、作成した用例クラスタ、辞書の語義を特徴ベクトルで表現し、ベクトル間の類似度を測ることで、用例クラスタに対応する語義をそれぞれ決定する。最後に、語義の対応付けの際に計算したクラスタと語義の類似度から、用例クラスタの意味が既存の語義である度合いを示す既存語義近接度を計算し、その値を基に新語義の判定を行う。

単語の用例のクラスタリングは、九岡らによって提案された手法を用いる。ここでは、用例をベクトルで表現する際、九岡らが提案する隣接ベクトル、連想ベクトル、LDA 拡張文脈ベクトル、トピックベクトルの 4 種類のベクトルとこれらを組み合わせる 2 通りの手法を試した。また、クラスタリングアルゴリズムとして、九岡らが採用した K-means 法の他に、初期クラスタを KKZ 法で決める K-means 法（K-means+KKZ 法）やトップダウン分割法を適用した。実験の結果、1 つのクラスタに同じ意味の用例がどれだけ集まっているかという purity の指標では K-means+KKZ 法が、新語義の用例がまとまったクラスタがどれだけ作成されたかという新語義識別率の指標ではトップダウン分割法が優れていた。

次に、用例クラスタに対し既存の語義を対応づける手法について述べる。まず、用例クラスタを特徴ベクトルで表現する。始めに、単語間の共起の強さを表わす共起行列を作成し、行列の 1 つの列をある単語の共起ベクトルとする。次に、用例クラスタにおいて、対象語の周辺に現れる自立語に対し、それに対応する共起ベクトルの和をクラスタの特徴ベクトルとする。次に語義の特徴ベクトルの作成方法について述べる。ここでは、辞書における語義の語釈文を用いる。本研究では、辞書の語釈文を定義文、例文、参照見出し、その他の 4 つのタイプ分類し、この中で単語の意味を表わしていると思われる定義文と例文を特徴ベクトルの作成に用いる。具体的には、定義文または例文に出現する自立語につい

Copyright © 2009 by Tanaka Hiroki

て、共起ベクトルの和を求め、語義の特徴ベクトルとする。また、この手法で作成された語義の特徴ベクトルは、定義文や例文の長さに応じて特徴ベクトルのスパースネスの度合いが語義毎に大きく異なり、語義の対応付けの精度を著しく低下させるという問題があることがわかった。これに対し、本研究では、辞書定義文の特徴ベクトルの構築方法を改良する方法、用例クラスタと辞書定義文との類似度の計算方法を改良する方法、特徴ベクトルに対する補正を行う方法の3つの案を提案する。この中で最も効果があったのは特徴ベクトルに対する補正を行う方法であった。評価実験では、コーパスから抽出された用例に対し人手で語義を付与し、同じ語義を持つ用例をまとめてクラスタを作成した完全に正しい用例クラスタと、提案したクラスタリング手法を用いて自動的に作成した用例クラスタとを用いた。一番高い対応付けの正解率は、正しい用例クラスタを用いた場合で61.9%、自動作成したクラスタを用いた場合で59.5% であった。

最後に、用例クラスタが新語義であるかを判定する手法を提案する。ここでは、用例クラスタが既存の語義である度合いを既存語義近接度と定義し、この値を基に判定する。本研究では次の3種類の既存語義近接を比較した。1つ目は用例クラスタと既存語義との類似度の分散、2つ目は用例クラスタと既存語義との類似度の最大値と最小値の差、3つ目は用例クラスタと既存語義との類似度の最大値であり、それぞれを $K\text{-}Var$、$K\text{-}Diff$、$K\text{-}Max$ とした。基本的には、既存語義近接度が小さい用例クラスタは新語義であると判断する。しかし、予備実験の結果、既存語義近接度に対して閾値を設けて新語義か否かを判定する手法は見込みがないことがわかった。そこで、用例クラスタを既存語義近接度の大きい順に並べ、その差が一番大きく、かつ差が十分に大きい所で、用例クラスタを既存語義と新語義に弁別する手法を考案した。実験の結果、3つの既存語義近接度の中では $K\text{-}Max$ が新語義の判定に一番有効であることがわかった。また、そのときの新語義判定のF値は、正しい用例クラスタを用いた場合で0.615であった。