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

Title	複数の特徴ベクトルを同時に考慮した語義識別
Author(s)	中西，隆一郎
Citation	
Issue Date	2011-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/9619
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

複数の特徴ベクトルを同時に考慮した語義識別

中西隆一郎 (0910041)

北陸先端科学技術大学院大学 情報科学研究科

2011 年 2 月 8 日

キーワード: 語義識別, コーパス, クラスタリング.

単語の意味は日々変化している。新しい単語の意味 (新語義) が出現した場合には、辞書にその語義を追加する必要がある。本論文では、コーパスから新語義を発見することを目標とし、そのための重要な要素技術である用例クラスタリングの新しい手法を提案する。用例クラスタリングとは、ある単語の用例の集合が与えられたとき、同じ語義を持つ用例のクラスタを作成する処理を指す。なお、用例のクラスタを作成した後、そのクラスタが新語義であるかを判定すれば、コーパスから新語義を自動的に発見することが可能である。本論文では、複数の特徴ベクトルを同時に考慮することで用例クラスタリングの精度向上を目指す。用例クラスタリングの先行研究として、九岡の研究がある。九岡は用例を 4 つの特徴ベクトルで表現し、それらを用いてクラスタリングを 4 回行い、生成されたクラスタリングの結果から最良と考えられるクラスタ集合を 1 つ選択するといった手法を提案している。これは、単語によって語義を特徴づけやすい観点が異なるという考えに基づいている。しかし、一般に、単語だけでなく語義によっても特徴づけられやすい観点が異なる場合があるため、同じ語義を持つ用例を 1 つのクラスタにまとめる事を目的とした場合、クラスタリングの過程で複数の特徴ベクトルを同時に用いることが望ましい。本論文ではその一手法を提案する。

以下、提案手法の概要について述べる。まず用例を特徴ベクトルで表現する。特徴ベクトルの作成方法は九岡が用いたものとほぼ同じ方法を用いた。彼は、異なる 4 つの観点から隣接、文脈、連想、トピックといった特徴ベクトルを作成している。隣接ベクトルについて、九岡は前後 1 語の単語をベクトルの素性としていたが、これでは本来は違う意味である語義の組に対しても前後の単語が一致していただけで高い類似度を与えてしまう。そこで、本研究では前後 2 語を用いて隣接ベクトルを作成している。

次にクラスタリングアルゴリズムについて述べる。本研究では凝集型クラスタリングアルゴリズムによって用例クラスタリングを行う。ただし、クラスタ間の類似度は 4 つの特徴ベクトルのうち最大のものと定義する。これは、4 つの特徴ベクトルで考慮されている様々な観点のうち、どれか 1 つでも似ていれば、語義が同じである可能性が高いという考えに基づく。また、特徴ベクトルの種類によって類似度の平均にばらつきがあり、類似度

Copyright © 2011 by Nakanishi Ryuichiro

の比較を行った場合に 1 つのベクトルのみが選択される可能性が高いという問題がある。そこで、ベクトル間類似度を正規化する 2 つの手法を提案する。1 つ目の手法は、あらかじめ全ての用例の組について類似度を計算し、各特徴ベクトル毎に類似度の最大値と最小値を求め、両者の範囲内における相対的な大きさをベクトル間類似度と定義する。もう 1 つは、同様に全ての用例の組について類似度を求め、その標本における偏差値を類似度と定義する手法である。

複数の特徴ベクトルを同時に考慮するにあたって、クラスタがどのような観点で注目されクラスタリングされたかを把握出来た方が望ましい。したがって、同じ種類の特徴ベクトルの類似度が高い場合にしかクラスタをマージしないという制約を与える。2 つのクラスタをマージして新しくクラスタを作成する際、マージするときに注目した (4 つのうち類似度が最大であった) 特徴ベクトルをクラスタラベルとして記録する。さらに、異なるクラスタラベルを持つクラスタはマージしないという制約を設ける。この制約から、同じクラスタに属する用例は同じ観点で注目してまとめられることになる。

凝集型クラスタリングの停止条件は、全クラスタ数が Tc 以下となり、かつ大きさが最大のクラスタ内に含まれる要素数の全用例数に対する比が Tr を超えたときの両方を満たした場合とした。

評価実験は SemEval-2 日本語タスクの訓練データを用いた。同タスクの 40 語を対象に、1 単語につき 40〜50 の用例を対象にクラスタリングを行った。作成されたベクトルに対してクラスタリングを行い、提案手法と先行研究との比較を行う。クラスタリングの停止条件として Tc=10 と 15 の 2 つの条件で実験を行った。比較のための評価指標は 9 つ用いており、その中でも Purity,Homogeneity,Paired Precision の 3 つの指標に注目した。実験の結果、提案手法の中では偏差値で正規化を行ったものが高い評価値を示しており、先行研究の九岡の手法よりも精度が高いことがわかった。しかし、全体の中で最も高い評価値を示したのは隣接ベクトルのみを用いてクラスタリングを行う手法であった。その原因を調査したところ、隣接ベクトルを用いてクラスタリングを行った場合に生成されるクラスタ集合には、初期状態のまま 1 度もマージされないクラスタが多く含まれていることがわかった。このようなクラスタは明らかに有用でない。そこで、1 つのクラスタに 2 つ以上の要素を含むクラスタについて、クラスタ内で最多の語義が占める割合を求めて比較した結果、提案手法は 1 種類のベクトルを用いる手法を上回った。これらの結果から、複数の特徴ベクトルを同時に考慮すること、その際に特徴ベクトルの類似度を正規化することが、用例クラスタリングの性能の向上に有効であることがわかった。