

Title	コーパスからの単語の意味の発見
Author(s)	九岡, 佑介
Citation	
Issue Date	2008-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/4343
Rights	
Description	Supervisor: 白井 清昭, 情報科学研究科, 修士

コーパスからの単語の意味の発見

九岡 佑介 (0610032)

北陸先端科学技術大学院大学 情報科学研究科

2008 年 2 月 7 日

キーワード: 単語の語義, 語義識別, クラスタリング.

文中の単語の意味を判別する語義曖昧性解消 (Word Sense Disambiguation; WSD) は, 機械翻訳を始めとする自然言語処理に必要となる基盤技術である. しかしながら, 従来の WSD の手法は, 一般に, 岩波国語辞典などの辞書においてあらかじめ定義された語義のいずれかを選択するという問題設定を前提とする. ところが, 語義は時を経るに従って変化するため, 実際に WSD を適用する時点において, 単語が辞書にない語義で使われている場合もありうる.

本論文では, 辞書に依存せず単語の意味を識別する語義識別の手法について研究する. 語義識別は単語の出現 (インスタンス) を同じ意味で使われたものがまとまるようにクラスタリングするという問題であり, 既存の辞書に依存せずに単語の意味を判別できる. 語義識別によって, コーパスにおいて同じ意味で使われている複数のインスタンスを見つけることができる可能性がある. また, 同じ意味で使われたインスタンスを含む用例を自動収集することもできる. これにより, WSD における辞書にない語義を判別できないという問題を解消でき, また異なる意味ごとに用例を提示することにより, 人手による辞書編纂を補助することもできる.

本研究における語義識別は次の手順で行う. まず, コーパスを用意し, 対象語のインスタンスを抽出する. 次に対象語のインスタンスを特徴ベクトルで表現する. そして, 似ている特徴ベクトルが 1 つのクラスにまとまるようにクラスタリングする. 最後に, 同じクラスに属するインスタンスは同じ語義であるとみなしてインスタンスの意味を判別する.

この提案手法の特徴は次の 2 つである. まず, インスタンスを様々な素性に基づく複数の特徴ベクトルで表現する. 具体的には, インスタンスの周辺語やその関連語を素性とする文脈ベクトル, 対象語を含む連語を素性とする隣接ベクトル, インスタンスの周辺語を他の単語や PLSI・LDA により推定したトピックとの関連度で特徴付ける連想ベクトル, インスタンスの出現した文書のトピックを PLSI で推定し素性とするトピックベクトルなどを提案する. もう 1 つの特徴は, 従来研究の多くが単語インスタンスを 1 つの特徴ベクトルで表現しクラスタリングを行っていたのに対し, 本研究では, インスタンスを複数の

特徴ベクトルで表現し、様々なタイプの素性を同時に考慮に入れてクラスタリングを行う。これは、対象語毎に語義識別に有効な特徴ベクトルは異なると考えられるからである。組み合わせ方式は次の2通りである。1つは、異なる特徴ベクトルにおける類似度の重み付き和によってクラスタ間の類似度を測り、クラスタリングする手法である。もう1つは、クラスタリングの良さを測る評価関数を導入し、クラスタリング結果が最も良いと思われる特徴ベクトルを単語毎に選択する手法である。クラスタリングの良さは、クラスタの要素が互いにどれだけ似ているか、異なるクラスタがどれだけ互いに似ていないか、などの観点から測る。また、特徴ベクトルによって類似度の大きさにばらつきがあることを考慮し、クラスタ内の要素間の類似度やクラスタ同士の非類似度を相対的に測る評価関数も提案した。

毎日新聞コーパスから抽出した10単語、Yahoo! 知恵袋コーパスから抽出した23単語を対象語として、提案手法を評価する実験を行った。まず、対象語のインスタンスを毎日新聞については70個、Yahoo! 知恵袋については100個ずつランダムに選択した。これらを Spherical k-means 法やセントロイド法によってクラスタリングした。クラスタリングの結果は Purity, Entropy, Inverse Purity という指標で評価した。

特徴ベクトルを単独で用いてクラスタリングを行ったところ、隣接・連想・トピック・文脈ベクトルの順にクラスタリングの結果が良いことがわかった。さらに対象語によってクラスタリングに有効な特徴ベクトルが異なり、隣接・連想・トピックベクトルによるクラスタリング結果の良さの順位が対象語によって異なることがわかった。また、文脈・隣接・連想ベクトル、文脈・隣接・トピックベクトル、文脈・隣接ベクトルを組み合わせ、これらのベクトル間の類似度の重み付き和を2つのベクトル間の類似度としてクラスタリングを行うことで、特徴ベクトルを単独で用いる手法よりクラスタリングの結果が改善した。さらに、文脈・隣接・連想・トピックベクトルの中から単語毎に最適な特徴ベクトルを選択することで、さらにクラスタリング結果が改善することがわかった。また、クラスタリングの良さを測る評価関数としては、クラスタ内の要素間の類似度を相対的に測る関数が最も有効であった。ただし、Purity や Inverse Purity が最も高い特徴ベクトルを常に選択できたわけではない。そのため、評価関数の改善によりクラスタリング結果のさらなる改善が期待できる。