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

Title	EDR概念辞書とコーパスを用いた語義曖昧性解消
Author(s)	八木, 恒和
Citation	
Issue Date	2004-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1888
Rights	
Description	Supervisor:白井 清昭, 情報科学研究科, 修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

EDR 概念辞書とコーパスを用いた語義曖昧性解消に関する研究

八木恒和 (210096)

北陸先端科学技術大学院大学 情報科学研究科

2004 年 8 月 13 日

キーワード: 語義曖昧性解消、教師あり学習、辞書定義文、上位概念、頑健性.

自然言語処理を行う際の重要な問題の一つに語義曖昧性解消 (WSD) がある。語義曖昧性解消の過去の研究として教師あり学習がよく行われている。しかし教師あり学習は正解付きデータを必要とし、正解付きデータの中で良く出現する単語に関しては良い結果が得られるが、低頻度語については学習が困難である。また本研究は、人間の文章理解を支援する読解支援システムでの使用を前提としているが、このような場合は低頻度語を含めた多くの単語に対して語義曖昧性解消を行う手法が求められる。この問題を解決するために、語義タグ付きコーパスと辞書定義文を利用した分類器を作成する手法を提案する。そして、高頻度語には教師あり機械学習の手法を使用し、低頻度語には辞書の定義文を用いた分類器を使用する。この 2 つを組み合わせることにより頑健な WSD システムの構築を行う。

まず、辞書定義文を用いた分類器について説明する。例えば、「犬」という単語があり、その語義として「犬という動物」と「スパイという役割の人」の 2 つがあるとする。そして、「犬にえさをあげる」という文の「犬」の語義を判定する場合を考える。「犬」が語義タグ付きコーパスに一度も出現しない場合は、「犬」の語義を判定するモデルを学習することはできない。ここで辞書定義文に含まれる語義の上位概念に着目する。語義の上位概念は辞書定義文の末尾から取出せる。例えば、「犬という動物」から「動物」という上位概念が取出すことできる。ここで、コーパスに「象にえさをあげる」「猫にえさをあげる」などの文があり、象、猫の上位概念は「動物」であるとする。このとき上位概念「動物」と「えさ」が共起することを学習できるので、「犬にえさをあげる」という文では「犬」の語義が「犬という動物」であることがわかる。このように、辞書定義文から語義の上位概念を抽出し、上位概念と周辺語との共起性を学習することにより、低頻度語でも単語の語義を正確に判定できる。

次に、上位概念を用いて語義曖昧性解消を行う手法について説明する。ここでは上位概念を用いた Naive Bayes モデル $P(s)\prod_i P(f_i|c)$ を学習し、その確率が最大となる語義を選

Copyright © 2004 by TSUNEKAZU YAGI

択する。この式において、f_i は素性、c は上位概念、s は語義を表わしている。また、語義を決めたい単語 (w) の直前・直後に存在する単語の表記や品詞、w の前後 20 単語以内に現れる自立語の基本形、w と係り受け関係にある単語の基本形などを素性 f_i として使用する。

次に、辞書定義文から上位概念を抽出する手法について述べる。基本的には、辞書定義文の末尾にある単語を上位概念として取り出す。また、上位概念が辞書定義文の末尾以外になるときもある。そのような場合を考慮して抽出パターンを作成した。例えば、「欠点をさがしだして言う悪口のこと」というような「N のこと」で終わる辞書定義文から N(この場合は「悪口」) を取出すパターンを作成した。64 個の抽出パターンを作成し、EDR 概念辞書の辞書定義文から上位概念を抽出した。上位概念を抽出できた語義数を EDR 概念辞書中の全ての語義数で割った値は 98%であった。また、抽出した上位概念をランダムで 200 個選択し、人手で判定したところ、185 個が適切な上位概念であった。

次に、高頻度語の WSD を行う教師あり機械学習の手法について説明する。ここでは、教師あり機械学習アルゴリズムとして、Support Vector Machine(SVM) を使用する。SVM で使用した素性として、Naive Bayes モデルで使用した素性以外に、w の二つ前または二つ後の単語または品詞、w の直前または直後に現れる 2 つの単語の表記または品詞の組を追加した。

本研究では、最終的に高頻度語のための SVM による分類器と低頻度語のための上位概念を用いた分類器の 2 つを組み合わせた。組みあわせる手法は以下の通りである。語義を決めたい単語の訓練データにおける出現頻度がある閾値以上なら SVM による分類器を、それ以外は Naive Bayes モデルを用いて判別を行う。本研究ではこの閾値を 5 とした。

最後に提案手法を評価する実験を行った。SVM、NB、BL(ベースラインモデル)、SVM+BL (SVM とベースラインモデルの組み合わせ)、SVM+NB (提案手法) の 5 つの手法の比較を行った。提案手法である SVM+NB と、3 つの単独の分類器の中で最も高い SVM を比較すると、再現率、F 値、適用率においては SVM+NB は SVM を上回るが、精度は劣る。特に再現率や適用率の向上が大きい。これは、SVM による分類器が語義タグ付きコーパスにおける高頻度語のみを対象としているのに対し、SVM+NB は低頻度語に対しても語義を出力しているからである。すなわち、語義曖昧性解消が行われる単語が増加するので再現率や適用率が向上したと考えられる。一方、SVM+NB は SVM+BL と大きな差は無かった。頻度 20 以下の単語だけを対象に両者を比較すると、頻度が小さい単語ほど SVM+NB は SVM+BL を大きく上回った。したがって、提案手法は低頻度語の語義曖昧性解消に有効であるということが明らかになった。