辞書定義文中の上位概念を用いた頑健な語義曖昧性解消

小川 千隼 (310024)

北陸先端科学技術大学院大学 情報科学研究科

2005 年 2 月 10 日

キーワード: 語義曖昧性解消, 辞書定義文, 上位概念, 分類器の組み合わせ, 頑健性.

語義曖昧性解消 (Word Sense Disambiguation,WSD) は文中に現れる単語の意味 (語義) を決める処理である．現在，語義曖昧性解消の手法として，語義タグ付きコーパスを利用した教師ありの機械学習による手法が主流であるが，訓練データ量を必ずしも十分に確保できないというデータの過疎性の問題がある．

このような問題に対処する手法として，語義タグのないコーパスを用いる教師なし学習を行う手法も提案されているが，本研究では 辞書定義文から語義の上位概念を抽出し，抽出した上位概念を反映した確率モデルを学習することにより，低頻度語の語義曖昧性解消の正解率を向上させる．例えば,「筆者」の 1 つの語義の辞書定義文,「その文章・書画を書いた人」から「人」という上位概念を抽出する．この例のように「人」という上位概念を持つ単語は「筆者」以外にもコーパスに存在する．このため，語義と文脈の共起情報を用いるよりも上位概念と文脈の共起情報を利用することで，語義自体はコーパスにあまり現れない単語でも上手く学習が出来る可能性がある．低頻度語用のモデルとして，上位概念と文脈の共起性を反映した Naive Bayes モデルを用いる．モデルに使う素性は，対象語の前後の表記，品詞，係り受け関係にある単語の基本形，同一文中にある自立語の基本形など，WSD に一般的に用いられるものを用いた．

本研究に近い研究を行った八木は，EDR 概念辞書を用いて上位概念を抽出した．ところが，EDR 概念辞書は機械処理に特化しているため，辞書定義文が単純でわかりづらい．例えば，EDR 概念辞書の「犬」の定義文は「犬という動物」である．これに対し，岩波国語辞典における「犬」の定義文は「古くから人間が家畜として飼い親しむ，いぬ科のけだもの」であり，犬に関してより多くの情報を得られる．定義文の品質が重要視されるアプリケーションにおいては，辞書定義文を理解しやすい一般の国語辞典を用いる方が望ましい．本研究では，より人にとって有益な表現の多い一般の国語辞典，具体的には岩波国語辞典を用いる．

次に，辞書定義文から語義の上位概念を抽出する手法について述べる．一般に，辞書定義文の末尾にある単語がその語義の上位概念を表していることが多い．したがって，原則

として，辞書定義文の末尾の単語を上位概念として抽出する．しかし，末尾の単語では上位概念としてふさわしくない場合もある．例えば,「拝借」の定義文の「借りることをへりくだって言う語」というような「N を V て言う語」で終わる辞書定義文から末尾の「語」を上位概念とするのは適切ではないので，N の部分を取り出すパターンを適用して上位概念「借りること」を取り出す．このような上位概念抽出パターンを 116 個人手で作成し，岩波国語辞典の辞書定義文から上位概念を抽出した．そして，岩波国語辞典の全語義のうち，97.25%の上位概念の抽出に成功した．

また，岩波国語辞典は EDR 概念辞書と異なり，1 つの語義に対して複数の辞書定義文が存在する場合がある．個々の定義文から上位概念を取り出すとすると，1 つの語義に対して複数の上位概念が抽出される．一方，上位概念と文脈の共起性を反映した Naive Bayes モデルでは語義の上位概念は 1 つであることを仮定している．そのため，抽出した複数の上位概念から最適なものを選択する必要がある．具体的には，辞書定義文の第 2 文以降を第 1 文の上位概念や文頭または文末のキーワードを手がかりに 3 つのタイプに分類し，そのタイプにしたがって上位概念を選択する．上記のプロセスのタイプ分類に用いるキーワードは全部で 38 種類ある．1 つの語義に対して複数の辞書定義文が存在している語義の辞書定義文をランダムに 200 語義取り出し，その第 2 文以降の分類タイプが適切なものであるかを人手で確認したところ，全体の 93.5%に相当する 187 語義の辞書定義文の分類タイプが適切であった．したがって，定義文の分類の精度は十分高いと言える．

本研究では，最終的に高頻度語のための教師あり学習モデルの 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．組み合わせる手法は以下の通りである．

- 単語ごとの訓練データにおける出現頻度により分類器を選択
- 単語ごとの調整用データにおける正解含有率により分類器を選択
- スタッキングによる手法により分類器を選択

  スタッキングの手法を用いて，SVM による分類器と上位概念を用いた Naive Bayes モデル分類器を 1 次分類器とし，それらの出力を素性とする 2 次分類器を学習し，語義曖昧性解消を行う．2 次分類器では学習アルゴリズムとして SVM を用いる．また，1 次分類器の作り方や 2 次分類器で用いる素性について，3 つの異なるスタッキング手法を提案した．

また，単語と文脈の共起情報と，上位概念の共起情報を同時に利用する新たな分類器を作成し，分類器を組み合せた場合との比較も行った．

最後に提案手法を評価する実験を行った．単体の分類器同士の SVM，NB，BL(ベースラインモデル) の比較や，混合モデルである 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)，2 つの共起情報を同時に反映した NB モデルの比較を行った．この結果，本研究の提案手法である NB は SVM には及ばなかったものの，BL モ

デルと比べてF値で2%以上の上昇がみられた．また，SVMとNBの混合モデルの中で最もF値の高いモデルは交差検定を用いたスタッキングで，全ての混合モデルの中で最も高いF値が得られた．