辞書の語義立てに基づく語義曖昧性解消に関する研究

玉垣 隆幸 (110076)

北陸先端科学技術大学院大学 情報科学研究科

2004 年 2 月 13 日

キーワード: 語義多義性解消, 語義タグ付きコーパス, 分類器の組み合わせ, 機械可読辞書.

単語の意味を決める語義曖昧性解消は、自然言語処理の中でも重要なタスクの一つである。本研究では、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た語義曖昧性解消のための分類器を作成する。読解支援システムでの使用が前提なので、より多くの単語を扱える再現率を重視した分類器が必要である。そのために、２つの異なる知識源を用いることにより、この問題の解決を試みた。一つ目の知識源は、注釈付きコーパスである。注釈付きコーパスとは、新聞記事などに人手で様々な付加情報を付け加えたテキストデータである。注釈付きコーパスから機械学習を行い、分類器を作成する。コーパスを使用した教師あり学習によってつくられた分類器の利点として、一般的に精度が良く、データ量が豊富であれば再現率も高いとされている。しかし、欠点もある。コーパス中に出現回数の少ない語義や文脈は学習に反映されづらいという、データの過疎性の問題がある。この欠点を克服するために、もう一つの異なる知識源 (国語辞典) を用いた分類器と組み合わせることにした。

注釈付きコーパスから機械学習を行うアルゴリズムとして、Support Vector Machine(SVM) を用いた。SVM は二値分類のアルゴリズムで、汎化性が強く、過学習を起こしにくいと言われている。学習を行う素性として以下のものを用意し、様々な素性に対して学習を行い、最もマッチした素性を実験的に求めた。

- 多義語の前後 n 語に含まれる自立語の基本型を抜き出す。n を可変にして、最適な文脈の大きさを調査した 。
- 多義語の直前、直後にある m 語の品詞情報と表記を抜き出す。m を可変にして、最適な m の大きさを調査した。

多義語の前後 n 語以内に現れる自立語の意味クラスを抜き出す。意味クラスはシソーラスの ID を用いた。意味クラスを用いる場合は、以下の２つの点で最適化を試みた。

- 一つは分類語彙表の桁数に関する最適化である。分類語彙表の ID を上位 3 桁から 7 桁まで変化させた

- もう一つは、一つの単語が複数の意味クラスをもつ場合の処理に関する最適化である。複数の ID が存在する場合は展開して素性に加える場合と単独の ID のみを加える場合を考慮した。

本研究では、コーパスから学習をして作成した分類器の他に、岩波国語辞典に記述されている情報を使用して 2 種類の分類器を作成した。

岩波国語辞典では、定義文中に用例が記述されていることがある。用例を用いた分類器は、入力文と語釈文中の用例の類似度を計算し、最も類似度の高い用例を持つ語義を選択する。類似度はシソーラスを使い求めた。一方、岩波国語辞典では、ある語義が出現する条件が文法情報として記述されていることがある。そこで、語義の文法情報を用いて語義曖昧性解消を行う分類器を作成した。この分類器は、候補となる全ての語義について、入力文がその語義の文法情報を満たすかどうかを調べる。そして、文法情報を満たす語義があれば、これを正しい語義として出力する。

さらに、SVM、用例、文法情報を用いたの 3 つの分類器を組み合わせる方法を提案した。最初に、共通のテストデータ (ヘルドアウトデータ) を用意し、それぞれの分類器単体の正解含有率を調べる。正解含有率は、出力した語義に正解が含まれる単語数の分類器によって語義が一つ以上出力された単語対する割合と定める。そして、ヘルドアウトデータにおける正解含有率の一番高い分類器の出力を最終的な出力として選択する。但し、SVM については単語毎に正解含有率を測定し、他の分類器の正解含有率との比較を行った。さらに、ヘルドアウトデータにおける頻度が 10 以下の単語については、正解含有率の信頼性が低いので、全単語の平均の正解含有率を SVM の正解含有率とした。

ヘルドアウトデータ、テストデータを用いて分類器の作成、評価を行った。SVM 分類器の作成・評価では、ベースライン精度 0.7877 に比べ、最高精度が 0.8059 と 1 %強しか上昇しなかった。そのときに用いた素性は、多義語の前後 7 語の読みと表記であった。また、シソーラスの意味クラスなど素性を加えると、かえって精度が落ちた。これは、過学習が起きたためと思われる。

一方、組み合わせの手法を用いた分類器は SVM 分類器と比べて精度は 8 %強、F 値は約 3 %落ちた。これに対し、再現率は 2 %強、適用率は 2 %近く上昇した。本研究の目的は、読解支援システムでの使用を前提とし、再現率を上げ、より多くの単語について語義の曖昧性を解消することにある。本結果から、この目的がある程度達成されたことが確認された。