未定義語義を含む語義曖昧性解消

菊田　篤史 (410036)

北陸先端科学技術大学院大学 情報科学研究科

2006 年 2 月 9 日

キーワード: 語義曖昧性解消, 未定義語義, EM アルゴリズム，Naive Bayes モデル.

文中の単語が，辞書に定義されている複数の意味のうち，どの意味で使用されているかを自動的に判別する処理を語義曖昧性解消 (Word Sense Disambiguation:WSD) という．この処理は，機械翻訳や情報検索などの様々なタスクに幅広く応用することができる．しかし，辞書に未定義の語義を持つ単語も存在する．例えば,「電話」という単語については，岩波国語辞典には「電話機による通話」,「電話機の略」という 2 つの語義が定義されている．ところが，文中で電話という単語が「電話番号」という語義で使用されている場合がある．このとき，従来の辞書に定義されている語義の中から適切なものを選択するという手法では，必ず間違った語義を選択してしまい機械翻訳や情報検索のタスクの誤りの原因となる．そこで，本研究では，文中の単語の語義を判別する際に，辞書に定義されている語義に加え，辞書に未定義であるということも判別できる語義曖昧性解消システムを構築することによりこの問題の解決を図る．

語義曖昧性解消は，コーパスと呼ばれる例文集を用いて語義を判別するモデルを学習する機械学習の手法が主流である．中でも例文中に単語の正しい語義がタグ付けされている語義タグ付きコーパスを用いて語義曖昧性解消のためのモデルを学習する教師あり学習が良い成果を挙げている．しかし，未定義語義が付与された語義タグ付きコーパスは存在しないため，この手法をそのまま適用することはできない．よって，機械学習のもう 1 つの手法で，何も情報が付加されていないプレーンテキストコーパスを用いる教師なし学習を行う．

本研究では，語義曖昧性解消に用いる Naive Bayes モデルを EM アルゴリズムと呼ばれる学習アルゴリズムを用いて学習する．さらに，EM アルゴリズムの初期値設定の際，語義タグ付きコーパスの統計情報を利用する．まず，語義タグ付きコーパスから語義が s_k であるときに素性 f_i が生起する条件付き確率 $P(f_i|s_k)$ を求め，$P(f_i|s_k)$ の上位 n 個の素性を語義ごとに抽出する．そして，n 個の素性の $P(f_i|s_k)$ にそれ以外の素性の $P(f_i|s_k)$ の r 倍の値を与える．また，上位 n 個の素性に対する $P(f_i|s_k)$ ならびにそれ以外の素性に対する $P(f_i|s_k)$ に対しては，全て等しい値を与える．$P(f_i| <$ 未定義 $>)$ については，$P(f_i|s_k)$ の上位 m 個の素性を抽出し，その m 個の素性に対する $P(f_i| <$ 未定義 $>)$ には低い値，そ

の他には高い値を与え，初期値を設定した．r, n, m の値は，未定義語義の判別の正解率が最大となるように実験的に求めた．

提案手法の評価実験を行った．未定義語義を持つ 10 単語について，提案手法によって語義曖昧性解消を行い，その正解率を調べた．Naive Bayes モデルを教師あり学習し，1 位の語義に対する確率がある閾値より低い場合に未定義と判別する手法 (BL) と EM アルゴリズムの初期値を全て一様分布として，Naive Bayes モデルを学習する手法 (EM_{uni}) と比較をした．それぞれの手法の中で辞書に定義されている語義と未定義の語義を合わせた全体の正解率 (Cor_{total}) が最大となるパラメタ設定のときを比較したところ，提案手法の Cor_{total} は 48.5%で，BL より 1.9%上回り，EM_{uni} には 2.7%劣っていた．F 値では，BL より 2.8%劣り，EM_{uni} より 26%上回った．未定義語義の適合率で見ると，BL より 51.4%，EM_{uni} は，全く未定義を判別できなかったので 81.6%上回った．また，未定義語義判別の F 値が最大となるパラメタ設定のときの比較では，Cor_{total} で，BL より 11.8%上回り，EM_{uni} より 4%劣っていた．提案手法の F 値は 36.8%となり，BL より 0.9%，EM_{uni} より 36.8%，適合率では，BL より 37.8%，EM_{uni} より 63.2%上回った．

提案手法では，ほとんどの単語において未定義語義と判別する数が少なく，未定義語義の再現率が低かった．しかし,「朝」,「電話」という単語は，再現率，適合率ともに高い値を得ることができた．r, n, m の値を変えて曖昧性解消の正解率が一番良いときに，これら２語については，未定義語義の適合率が 100%であった．この２語は，辞書に定義されている語義，あるいは未定義語義と共起する素性の違いが明確で，学習データにも多く含まれており，語義判別が比較的容易であると考えられる．以上から，提案手法は改善の必要があるものの，未定義語義を判別する手法として有望であることが確認された．