国語辞典と語義タグ付きコーパスを用いた頑健な語義曖昧性解消

小川　千隼　　白井　清昭
北陸先端科学技術大学院大学 情報科学研究科
{ochihaya, kshirai}@jaist.ac.jp

1 はじめに

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，語義タグ付きコーパスは人手で作成する必要があり，作成コストが高いという問題がある．また，コーパスにおける出現頻度の少ない語義については語義判定のモデルの学習が難しいというデータの過疎性の問題もある．

このような問題に対処する手法として，教師なし学習を行う手法 [2, 3] が提案されている．また，白井らは，辞書定義文から得られる上位概念を利用した頑健な語義曖昧性解消手法を提案している [1]．この手法は，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 WSD の正解率を向上させている．また，高頻度語を対象とした教師あり学習によるモデルと低頻度語を対象とした語義の上位概念を用いたモデルを組み合わせることにより，頑健な語義曖昧性解消を実現している．

本研究は白井らの手法を以下の 2 点について拡張する．

- 辞書定義文として一般の国語辞典を利用する
  白井らは，辞書定義文として EDR 概念辞書を用いている．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ずらい．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場合，例えば語義の定義文を表示することによって日本語学習者の文書の読解を支援するシステムを構築する場合には，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，人にとって有益な情報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．ただし，岩波国語辞典の辞書定義文は EDR 概念辞書よりも複雑であるため，辞書定義文から語義の上位概念を抽出することは難しくなる．
- モデルの組み合わせ手法を洗練する
  白井らは高頻度語用の教師あり学習によるモデルと低頻度語用の語義の上位概念を用いたモデルの組み合わせについては単純な手法しか試していない．本研究では，2 つのモデルの組み合わせ手法を洗練する．具体的にはスタッキングによる組み合わせを試みる．

2 語義曖昧性解消モデル

本手法では高頻度語用のモデルと低頻度語用のモデルの 2 つを用いて語義曖昧性解消を行う．

高頻度語用には教師あり学習の手法を用いる．学習アルゴリズムとして SVM を用いた．単語ごとに分類器を作成し，語義の判別を行った．学習に用いる素性としては，対象語の前後の表記，品詞，係り受け関係にある単語の表記，同一文中にある自立語の基本形や意味クラスなど，語義曖昧性解消に一般的に用いられる素性を用いた．SVM の学習には TinySVM [1]を使用した．使用したカーネルは線形カーネルである．また，SVM は二値分類器であるが，多値分類問題である WSD タスクに適用するために one vs rest 法を用いた．

低頻度語用には式 (1) の Naive Bayes モデルを用いる．

$$P(s)P(F|c) = P(s)\prod_{f_i \in F} P(f_i|c) \qquad (1)$$

式 (1) において，s は語義，c は語義 s の上位概念，F は語義を決めたい単語の文脈を表わす素性の集合である．F として，予備実験によって最適化された SVM の素性のサブセットを用いた．

語義の上位概念 c は，国語辞書の各語義の辞書定義文から抽出される．上位概念を用いる理由は，擬似的に訓練データを増やす効果が得られるためである．例えば，「歳暮 (せいぼ)」という単語に対して語義曖昧性解消を行う場合を考える．「歳暮」の岩波国語辞典の辞書定義文を以下に示す．

【歳暮】(1) としのくれ．年末．

[1] http://chasen.org/%7Etaku/software/TinySVM/

単語	辞書定義文
【歳末】	年のくれ．年末.
【歳晩】	年のくれ．年末.
【中元】	七月十五日．そのころに行う贈物.
【御礼】	その言葉やその気持を表す贈物.

図 1: 同じ上位概念を持つ単語とその定義文の例

【歳暮】(2) この一年世話になった礼の意味で，年末に贈物をすること．その贈物.「お—」

辞書定義文から上位概念を抽出する手法については 3 節で詳しく述べるが，ここでは語義 (1) の上位概念として「年末」が，(2) の上位概念として「贈物」が抽出されたとする．このとき,「歳暮」の語義と同様に,「年末」「贈物」を上位概念として持つ単語が岩波国語辞典には表 1 のように存在する．例えば「年末」という上位概念を持つ「歳末」「歳晩」という単語の語義は,「歳暮」の語義 (1) とほぼ同じような意味を持つといえる．したがって,「歳暮」という単語がコーパスにあまり出現しなくても，同じ上位概念を持つ単語の語義「歳末」「歳晩」がコーパスに出現していれば，それを「歳暮」の語義曖昧性解消を行うための訓練データとして使うことができる．式 (1) において，上記の効果を直接反映しているのが項 $P(f_i|c)$ である．$P(f_i|c)$ は，上位概念 c と，語義を決める単語の文脈を表わす素性 f_i との共起関係を反映しているとみなせる．このとき，上位概念 c は語義 s そのものよりも訓練コーパスにおける出現頻度が高いため，信頼できる確率を推定するために必要な訓練データ量を確保しやすい．

最終的に，高頻度語用の SVM によるモデルと，低頻度語用の上位概念を考慮した Naive Bayes モデルの 2 つを組み合わせることにより，より多くの単語に対して適用可能な語義曖昧性解消のモデルを構築することができる．両者を組み合わせる手法については 4 節で述べる．

3 定義文からの語義の上位概念抽出

本節では，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．

3.1 上位概念抽出ルール

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．図 1 の例でも，定義文の末尾の単語が語義の上位概念であるとみなせる．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，末尾の単語以外の単語が上位概念としてふさわしい場合もある．例を以下に挙げる．

【言い尽くす】 言い得ることをすべて言ってしまう.

この定義文から単純に末尾にある単語を取り出すと,「しまう」という単語が上位概念として取り出される．これは非自立な動詞であり，明らかに誤った上位概念である．この定義文から正しく取り出すべき上位概念は「言う」である．正しい上位概念を取り出すためには，定義文が「＜動詞＞＋てしまう」で終わるとき,「＜動詞＞」を上位概念として取り出すというルールを用意する必要がある．本研究では，このような 116 個の上位概念抽出ルールを人手で作成した．

3.2 複数の定義文の取り扱い

岩波国語辞典では，EDR 概念辞書とは異なり，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，式 (1) のモデルは語義の上位概念 c 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．まず，辞書定義文の第 2 文以降を以下のように分類する．

- 同語義定義文: 1 つ前の文と同じもしくは似た意味を表わす文.
  例:【全治】 病気や傷が，すっかり直ること．全快.
- 別語義定義文: 1 つ前の文とは異なる意味を表わす文.
  例:【代行】 本人に代わって物事を行うこと．また，その人.
- 非定義文: 例・由来・性質・使い方など表わし，単語の意味を説明するものではない文.
  例 1:【安山岩】火成岩の一種. 暗灰色で，緻密 (ちみつ)．建築・土木用に多く使われる.
  例 2:【蜂】膜翅 (まくし) 目の昆虫. 丈夫な膜質の羽があり，(中略) スズメバチなど種類が多い.

第 2 文以降の文をこれらのタイプに分類するためのアルゴリズムを以下に示す.

- 非定義文の判定
  - 第 1 文から抽出した上位概念が「動物・植物・昆虫・魚・高木」などであれば，その語義の第 2 文以降は全て非定義文とみなす.
  - 第 1 文の定義文の文末が「〜の一つ・一種・名称・単位」などであれば，その語義の第 2 文以降は全て非定義文とみなす.

  - 第 2 文以降の定義文の文頭が「例」であれば，その文以降の文は全て非定義文とみなす．
  - 第 2 文以降の定義文の文末が「〜使う・用いる・言う・読む・用」などであれば，その文以降は全て非定義文とみなす．
- 別語義定義文の判定
  - 第 2 文以降が「また・もと・転じて・比ゆ的に」などで始まれば，その定義文は別語義定義文とみなす．
- 同語義定義文の判定
  - 非定義文，別語義定義文と判定された文以外の文は全て同語義定義文とみなす．

上記のアルゴリズムは，実際の辞書定義文と上位概念を見て考案した．上記のプロセスで用いる判定のためのキーワードは全部で 38 種類である．この手法を評価するために，1 つの語義に対して複数の辞書定義文が存在する語義をランダムに 200 個選択し，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．

第 2 文以降の複数の定義文を分類した後，抽出すべき上位概念を以下のように決定する．

1. 非定義文の場合，その定義文は単語の意味を表現するものではないので上位概念を抽出しない．
2. 同語義定義文 (似た意味の定義文) の場合，個々の定義文から上位概念を抽出する．これらのうち，辞書全体における出現頻度が高い上位概念，すなわちより多くの語義の定義文に出現する上位概念を 1 つ選択する．
3. 別語義定義文 (全く別の意味の定義文) の場合，1 つの定義文の中に複数の語義が定義されているとみなせる．このとき，複数の上位概念を全て抽出する．語義曖昧性解消の際，全ての上位概念の候補について式 (1) の確率を計算し，最も確率の大きい上位概念を選択する．すなわち，1 つの定義文に記述されている複数の語義に対して曖昧性を解消する．

具体的に「解釈」の辞書定義文から上位概念を抽出する例を示す．

> 【解釈】 文章や物事の意味を，受け手の側から 理解すること．また，その理解したところを 説明すること．その 内容．

まず，第 2 文が「また」で始まることから別語義定義文，第 3 文が同語義定義文と判定される．したがって，この定義文中には，第 1 文で定義される語義と，第 2,3 文で定義される語義の 2 つがあるとみなせる．後者からは 2 つの上位概念「説明すること」と「内容」が抽出されるが，辞書中の頻度が高い「内容」が選択される．最終的に「理解すること」と「内容」が【解釈】の上位概念として選択される．語義曖昧性を行う際，2 つの上位概念のそれぞれについて確率を計算し，大きい方を選択する．

本節で述べた手法で，岩波国語辞典の全ての 76,439 の語義のうち，74,336 語義 (97.25%) の上位概念の抽出に成功した．抽出に失敗した辞書定義文のほとんどは，文法情報や用例など，単語の意味を表わさない文であった．

4 混合モデル

本研究では，高頻度語用のモデルと低頻度語用のモデルを組み合わせたモデルを混合モデルと呼ぶ．ここでは混合モデルを作成する 3 つの手法について，4.1，4.2，4.3 項でそれぞれ述べる．白井らの研究 [1] では 4.1，4.2 項の手法を試しているのに対し，本研究ではさらにスタッキングを用いる手法 (4.3 項) を提案する．

4.1 頻度による混合モデル

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，それ以外は上位概念を用いた Naive Bayes モデル分類器を選択する．5 節の実験ではこの閾値を 5 とした．

4.2 正解含有率による混合モデル

共通の調整用データを用意し，それぞれの分類器単体の正解含有率 (式 (2)) を調べる．

$$\text{正解含有率} = \frac{\text{出力した語義に正解が含まれる単語数}}{\text{分類器が語義を一つ以上出力した単語数}} \tag{2}$$

単語毎に調整用データにおける正解含有率を求め，それが高い分類器の出力を最終的な出力として選択する．

4.3 スタッキングによる混合モデル

スタッキングとは Wolpert によって提案された手法で，複数の分類器を同じ訓練データで訓練し，それらの分類器の出力を新たに素性として加え，分類器を再学習することにより，機械学習の精度を向上させる手法である [4]．ここでは，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，それらの出力と 1 次分類器で用

いた素性[2]を新たな素性集合として 2 次分類器を学習し，語義曖昧性解消を行った．また，2 次分類器の学習アルゴリズムとして SVM を用いた．さらに，スタッキングによる混合モデルとして以下の 3 つを試した．

- シンプルスタッキング
  1 次分類器で学習したデータと同じデータを用いて 2 次分類器を学習する．
- 交差検定を用いたスタッキング
  シンプルスタッキングと同じだが，5 分割の交差検定によって，1 次分類器の学習データとは異なるデータで 2 次分類器を学習する．
- スコアを用いたスタッキング
  2 次分類器を学習する際に，1 次分類器で用いた素性は使わず，1 次分類器の出力のみを用いる．1 次分類器に，複数の語義の候補をスコアとともに出力させ，分類器の種類と語義の組を素性とし，語義のスコアを値とした訓練ベクトルを作成する．SVM のスコアは分類平面との距離，Naive Bayes モデルのスコアは式 (1) の確率である．但し，これらのスコアを最上位の語義のスコアを 1 とした相対スコアに変換する．また，5 分割の交差検定によって，1 次分類器の学習データとは異なるデータで 2 次分類器を学習する．

5 評価実験

RWC コーパスを用いた評価実験を行った．RWC コーパスは 3000 個の新聞記事からなるコーパスであり，各単語に岩波国語辞典の語義 ID が付与された語義タグ付きコーパスである．300 記事をテストデータ，300 記事を調整用データ，2400 記事を訓練データとした．また，4.3 項のスタッキングでは，調整用データと訓練データを合わせた 2700 記事を学習データとした．

まず，単体の分類器の実験結果を表 1 に示す．表 1 において，「SVM」は SVM による分類器，「NB」は語義の上位概念を用いた Naive Bayes モデルによる分類器，「BL」は最も出現頻度が高い語義を選択するベースラインモデルを表わす．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[3]で 2 %以上の上昇がみられた．再現率も高く，また低頻度語を含む全ての単語に適用できた[4]．

[2]SVM による分類器と Naive Bayes モデル分類器では用いる素性が異なる (後者は前者のサブセットである)．本論文では，SVM で用いた素性を 2 次分類器で使用した．

[3]F 値は $2PR/(P+R)$ とした．(P は精度，R は再現率)

[4]適用率とは (語義を出力した単語)/(全単語) を表す．

表 1: 単体分類器の比較

	精度	再現率	F 値	適用率
BL	76.71	76.61	76.66	98.71
NB	78.67	**79.09**	78.88	**100**
SVM	**84.77**	78.46	**81.50**	92.56

表 2: 混合分類器の比較

	精度	再現率	F 値	適用率
頻度	83.35	83.73	83.54	**100**
正解含有率	83.55	83.92	83.73	**100**
シンプル	83.71	84.12	83.91	99.76
交差検定	**84.26**	**84.68**	**84.47**	99.76
スコア	84.00	84.42	84.21	99.02
SVM+BL	83.72	82.92	83.32	98.73

次に，4 節で述べた各種混合モデルの実験結果を表 2 に示す．「SVM+BL」は，高頻度語に SVM，低頻度語に BL を用いたモデルを表わす．これは，辞書を用いずに語義タグ付きコーパスだけを用いて低頻度語に対応するナイーブな手法である．この結果，最も F 値の高い混合モデルは交差検定を用いたスタッキングであった．

6 おわりに

本研究では，WSD におけるデータの過疎性の問題に対処するために，国語辞典から語義の上位概念を抽出する手法を述べた．さらに，高頻度語に有効な教師あり学習による分類器と低頻度語に有効な上位概念を用いた分類器を組み合わせる様々な手法を実験的に比較した．今後の課題として，教師なし学習との比較や上位概念抽出パターンの他の国語辞書への応用，分類器の組み合わせ手法の検討，上位概念抽出手法の洗練などが挙げられる．

参考文献

[1] Kiyoaki Shirai and Tsunekazu Yagi. Learning a Robust Word Sense Disambiguation Model using Hypernyms in Definition Sentences. 20th International Conference on Computational Linguistics, pp. 917-923, 2004.

[2] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. Proceeding on the Annual Meeting of the Association for Computational Linguistics, pp.189-196, 1995.

[3] Seong-Bae Park, Byoung-Tak Zhang and Yung Taek Kim. Word Sense Disambiguation by Learning Decision Trees from Unlabeled Data. Proceedings of Applied Intelligence 19, pp.27-38, 2003.

[4] David H. Wolpert. Stacked Generalization. Neural Networks, v.5 n.2, pp.241-259, 1992.