コーパスと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

白井 清昭　　八木 恒和
北陸先端科学技術大学院大学 情報科学研究科
{kshirai,t-yagi}@jaist.ac.jp

1 はじめに

語義曖昧性解消は，文中に現われる単語の意味 (語義) を決める処理であり，機械翻訳をはじめとする様々な自然言語処理に必要とされる技術である．近年では，コーパスを利用して語義の曖昧性を解消する研究が主流であり，中でも教師あり学習による手法が比較的良い成果をあげている [1, 2]．しかし，教師あり学習を行うためには正解付きデータ，すなわち語義タグ付きコーパスを必要とする．そのため，コーパスに現われない単語や出現頻度の低い単語については，語義を判定するモデルを学習することができないという，いわゆるデータの過疎性の問題がある．

この問題に対処するために，我々は辞書定義文から得られる上位概念を用いて，低頻度語の語義の曖昧性を解消する手法を提案した [8]．最頻出語義を常に選択するベースラインモデルと比べて，この手法は低頻度語に対する語義曖昧性解消の精度や再現率を向上させることができた．一方，高頻度語についてはあまり向上がみられなかった．そこで，我々は，高頻度語は訓練データ量が十分にあるので，高頻度語については既存の教師あり機械学習アルゴリズムによる手法を用いた方がベースラインモデルよりもはるかに高い精度が得られると考えた．

このような背景から，本研究は，高頻度語を対象とした教師あり機械学習に基づく手法と，低頻度語を対象とした辞書定義文から得られる上位概念を利用する手法を組み合わせ，頑健な語義曖昧性解消を実現することを目的とする．ここでいう頑健性とは，多くの単語について正しい語義を選択することができるという意味であり，実用的な応用を考えるときには重要である．データの過疎性の問題を回避し，頑健性を向上させるためには，教師なし学習を行う手法も考えられるが，本研究はコーパス以外の知識源 (具体的には辞書定義文) を併用するという立場を取る [3, 5, 9, 10]．

以下，本研究では，語義の定義としてEDR概念辞書 [6] を用いる．また，語義曖昧性解消を行うシステムを分類器と呼ぶ．

2 SVMによる分類器

本節では，語義タグ付きコーパスから分類器を機械学習する手法について述べる．機械学習アルゴリズムとしてSupport Vector Machine(SVM) を用いた．学習に用いた素性は以下の通りである．また，これらの素性を得るために，形態素解析器としてJUMAN [1]を，文節の係り受け解析器としてKNP[2] を用いた．

- $S(0), S(-1), S(-2), S(+1), S(+2)$
  対象語及びその周辺にある語の表記．括弧内の数値は対象語からの位置を表わす．
- $P(-1), P(-2), P(+1), P(+2)$
  対象語の周辺にある語の品詞．品詞はJUMANの品詞体系における第一，第二，第四分類の組とした．
- $S(-2)\cdot S(-1)$, $S(+1)\cdot S(+2)$, $S(-1)\cdot S(+1)$
  対象語の周辺にある2つの語の表記の組．
- $P(-2)\cdot P(-1)$, $P(+1)\cdot P(+2)$, $P(-1)\cdot P(+1)$
  対象語の周辺にある2つの語の品詞の組．
- B_{sent}
  同一文中にある自立語の基本形．但し，数字については “NUM” という特別なシンボルを素性として用いた．
- C_{sent}
  同一文中にある自立語の意味クラス．意味クラスは日本語語彙体系 [4] を用いた．ただし，自立語が多義 (複数の意味クラスを持つ) のときには素性に加えないこととした．
- B_{bs_head}, B_{bs_mod}
  対象語が文節の主辞であるとき，その文節の係り先文節の主辞の基本形 (B_{bs_head}) と係り元文節の主辞の基本形 (B_{bs_mod})．
- B_{bs_in}
  対象語が文節の主辞でないとき，その文節の主辞の基本形．

[1] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
[2] http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html

- $(B_{case}; B_{noun})$
  対象語が動詞のとき，その動詞の格 (B_{case}) と格要素 (B_{noun}) の基本形の組.
- $(B_{case}; C_{noun})$
  対象語が動詞のとき，その動詞の格 (B_{case}) と格要素の意味クラス (C_{noun}) の組．意味クラスは日本語語彙体系を用いた．ただし，自立語が多義のときには，その全てを素性に加えた.
- $(B_{case}; B_{verb})$
  対象語が名詞で，かつある動詞の格要素となっているとき，その格 (B_{case}) と動詞 (B_{verb}) の基本形の組.

SVM の学習には LIBSVM を用いた[3]．ν−SVM [7] によって学習を行い，カーネルは線形カーネル，$\nu = 0.0001$ とした．SVM は二値分類器であるのに対し，本研究における語義曖昧性解消問題は多値問題である．そこで，pairwise 法を用いて SVM を多値問題に適用した.

3 辞書定義文を利用した分類器

本節では，主に低頻度語を対象とし，語義タグ付きコーパスと辞書定義文から抽出された上位概念を用いて語義曖昧性解消を行う分類器について簡単に述べる．なお，この分類器の詳細については文献 [8] を参照されたい.

以下は，EDR 概念辞書に記載されている「漫談」の語義の定義文である．6 桁の英数字は概念 ID と呼ばれる語義の識別子である.

3c5631　こっけいな話をしながら, その中で社会批評や風刺をする演芸
1f66e3　とりとめもない話

ここでは，辞書定義文に含まれる語義の上位概念に着目し，それぞれの辞書定義文の最後に現われる名詞をその語義の上位概念とみなす．例えば，3c5631 の語義の上位概念は「演芸」であり，1f66e3 の語義の上位概念は「話」である．もし,「漫談」が訓練コーパス中に一度も現われない単語であるならば,「漫談」の語義が 3c5631 か 1f66e3 を決定するモデルを機械学習することは不可能である．しかし,「漫談」の上位概念が「演芸」か「話」であるかを選択するモデルを学習することは可能である．なぜなら，EDR 概念辞書中には，他にも演芸や話を上位概念とする語義が存在し，これらの語義が訓練コーパス中に存在する可能性があるためである．そのような語義の例を図 1 に挙げる．一般に，機械学習による語義曖昧性解消は，語義と語義を決める対象語の周辺に出現する情報 (素性) との共起関係を学習することにより行われる.「漫談」が訓練コーパスに存在しない場合,「漫談」の語義 (3c5631 または 1f66e3) と素性との共起性は学習できないが，図 1 に挙げた単語と語義が訓練コーパスにある場合には，演芸や話といった上位概念と素性との共起性は学習することができる．例えば,「落語の世界」「猿楽の世界」のように，“の世界” の前に演芸を上位概念とする語義はよく現われるが，話を上位概念とする語義はあまり現われない，といった傾向が学習できる．このように，上位概念は一般に複数の単語の語義で共有されるため，上位概念のコーパスにおける出現頻度は語義そのものの出現頻度に比べて高い．したがって，辞書定義文から抽出された上位概念を利用することにより，訓練データの量を増やす効果が期待できる.

【落語】 10d9a4 こっけいな話を続け, 最後に落ちをつける寄席演芸
【猿楽】 3c3fbb 猿楽という中世の民衆演芸
【伝説】 3cf737 昔から民間に語り伝えられた話
【実話】 0f73c1 実際にあった本当の話

図 1: 演芸または話を上位概念とする語義の例

分類器の構築は以下のように行う．まず，語義タグ付きコーパスに付与されている語義をその上位概念に置き換えたコーパスを作成する．語義の定義文からの上位概念の抽出は，人手によって作成された 64 個の抽出パタンとのパタンマッチにより行う．次に，Naive Bayes モデルに基づく式 (1) の確率モデルを学習する.

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

式 (1) において，s は語義，c は語義 s の語釈文から抽出された上位概念，F は学習に用いる素性の集合である．直観的に言えば，式 (1) の第 1 項 $P(s)$ は語義の出現頻度を学習するモデル，第 2 項 $\prod P(f_i|c)$ は語義の上位概念 c と素性 f_i の共起性を学習するモデルである．これらのパラメタは，語義タグ付きコーパス及び前述の語義を上位概念に置き換えたコーパスから推定する．語義曖昧性解消は，式 (1) が最大となる語義 s を選択することにより行う．また，式 (1) が最大となる語義が複数存在するときは，その全てを出力する.

式 (1) の推定に用いた素性集合は，2 節で述べた素性のうち，以下の素性を除いたものである．これは，予備実験で検討したところ，SVM の素性を全て用いるよりも下記の素性を除いた方が結果が良かったためである.

[3] http://www.csie.ntu.edu.tw/%7Ecjlin/libsvm/

- $S(-2)$, $S(+2)$, $P(-2)$, $P(+2)$
  対象語の前後 2 語目の位置にある単語の素性
- $S(-2)\cdot S(-1)$, $S(+1)\cdot S(+2)$, $S(-1)\cdot S(+1)$
  $P(-2)\cdot P(-1)$, $P(+1)\cdot P(+2)$, $P(-1)\cdot P(+1)$
  対象語の周囲にある 2 つの語の組に関する素性
- C_{sent}, $(B_{case}; C_{noun})$
  意味クラスを用いた素性

4 混合モデル

1 節で述べたように，本研究は，高頻度語のための SVM による分類器 (2 節) と低頻度語のための上位概念を用いた分類器 (3 節) の 2 つを組み合わせる．組み合わせ手法として以下の 2 つを試みた．

4.1 頻度に基づく混合モデル

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，それ以外は辞書定義文中の上位概念を用いた分類器を選択する．5 節の実験ではこの閾値を 20 とした．

4.2 調整用データでの正解率に基づく混合モデル

共通のテストデータ (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．

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

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．また，調整用データにおける頻度が Oh 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比較する．5 節の実験では Oh=10 とした．

正解含有率ではなく精度 (適合率) を比較することも考えられる．しかし，本研究は再現率の向上を第一の目的としている．そのため，分類器が出力する語義の中に正解が含まれていれば効果があるとして，正解含有率の比較を行った．

5 評価実験

提案手法を評価する実験を行った．実験には EDR コーパス [6] を用いた．EDR コーパスは約 20 万文からなるコーパスであり，各単語に EDR 概念辞書の概念 ID が付与された語義タグ付きコーパスである．EDR コーパスのうち，20,000 文をテストデータ，20,000 文を調整用データ，残りの 161,332 文を訓練データとした．テストデータに含まれる評価単語数は 91,986 である．

表 1: 実験結果

	再現率	精度	F 値	適用率
BL	.5782	.5957	.5868	.9706
NB	.6248	.6530	.6386	.9568
SVM	.6352	.7064	.6689	.8992
SVM+NB(頻度)	.6985	.6990	.6988	.9993
SVM+NB(調整)	.7024	.7029	.7026	.9993

テストデータに対する語義曖昧性解消の再現率，精度，F 値[4]，適用率を表 1 に示す．適用率は分類器が語義を 1 つでも出力することができた単語の割合である．表 1 において，SVM は 2 節で述べた SVM による分類器を，NB は 3 節で述べた (辞書定義文中の上位概念を用いた)Naive Bayes モデルによる手法を表わす．BL は最頻出語義を常に選択するベースラインモデルを表わす．ただし，最頻出語義が複数ある場合にはその全てを答えとして選択する．一方，SVM+NB(頻度) と SVM+NB(調整) は，それぞれ 4.1，4.2 項の手法に基づく SVM と NB の混合モデルを表わす．

まず，SVM による分類器と混合モデルを比較すると，精度以外の評価値において混合モデルは SVM を上回る．特に再現率や適用率の向上が大きい．これは，SVM による分類器は語義タグ付きコーパスの高頻度語しか対象にしていなかったが，辞書定義文を用いた分類器と併用することにより，語義の曖昧性を解消できる単語が増加したことが主な要因であると考えられる．一方，SVM と NB の 2 つの組み合わせ手法，SVM+NB(頻度) と SVM+NB(調整) を比べると，後者の方がわずかに結果が良かったものの，ほとんど差は見られない．調整用データでの正解率に基づく混合モデルにおいて，高頻度語に対しては SVM の方が正解含有率が高いために SVM が選択されやすいのに対し，低頻度語に対しては SVM は語義を出力しないので NB が選択されると予想される．したがって，この組み合わせ方式は，単純に訓練データにおける頻度の大きさで分類器を選択する手法と大差がなかったと思われる．

我々は頑健な語義曖昧性解消システムを構築することを目的としている．教師あり学習に基づく手法の頑健性を向上させるナイーブな方法として，ベースラインモデルとの併用が考えられる．そこで，SVM による分類器とベースラインモデルによる混合モデルを作成し，提案

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

表 2: 混合モデルの比較

	再現率	精度	F 値	適用率
SVM+NB	.7024	.7029	.7026	.9993
SVM+BL	.6939	.6968	.6953	.9958
SVM+NB+BL	.7052	.7056	.7054	.9995

表 3: 混合モデルで選択された分類器の割合

	SVM	NB	BL	no ans.
SVM+NB	.7369	.2623	—	.0007
SVM+BL	.7606	—	.2352	.0042
SVM+NB+BL	.6449	.2467	.1078	.0005

手法との比較を行った．また，本論文で提案した 2 つの分類器とベースラインモデルを組み合わせる混合モデルも作成した．分類器の組み合わせ手法は調整用データでの正解率に基づく手法 (4.2 項) である．結果を表 2 に示す．

一番成績が良いのは 3 つの分類器の混合モデルであった．また，SVM+NB と SVM+BL を比較すると，全ての評価基準で前者が後者を上回った．これは，語義タグ付きコーパスの他に辞書定義文を語義曖昧性解消の知識源として利用していることの効果であるとみなせる．しかし，両者の差はわずかであり，その効果はそれほど大きいとはいえない．したがって，低頻度語を対象とした分類器 (NB) の更なる改善が必要である．また，SVM+NB は SVM+BL と比べてわずかではあるが適用率が向上していることに注目していただきたい．これは，ベースラインモデルは，訓練コーパスに現われない語義を正解とする単語に対しては正解を出力することができないのに対し，辞書定義文中の上位概念を用いた分類器では，そのような単語に対しても正しい語義を選択できることがあるためである．このように，コーパスに存在しない語義に対応できるという点は，語義タグ付きコーパス以外の知識源を利用することのひとつの大きな利点である．

表 3 は，混合モデルにおける個々の分類器が選択された回数の割合である．表中の “no ans.” はどの分類器も語義を出力しなかったことを表わす．辞書定義文を用いた手法 (NB) が選択された回数の割合は 25%程度であった．

6 おわりに

本論文では，語義曖昧性解消の頑健性を改善することを目的に，語義タグ付きコーパスを用いた教師あり学習に基づく分類器と，辞書定義文中に含まれる上位概念を利用した分類器を組み合わせる手法を提案した．実験の結果，コーパス以外の知識源を併用することにより，システムの頑健性が改善されたことを確認した．しかし，その効果は決っして大きいとはいえない．辞書定義文を用いる手法の更なる改善や，辞書定義文以外の知識源を用いた分類器をさらに組み合わせることなどが今後の課題である．

参考文献

[1] *ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, 2002.

[2] *Natural Language Engineering – Special Issue on Evaluating Wo rd Sense Disambiguation Systems*, Vol. 8, 2002.

[3] E. Agirre, G. Rigau, L. Padró, and J. Atserias. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities*, Vol. 34, No. 1,2, pp. 103–108, 2000.

[4] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙体系 — 全 5 巻 —. 岩波書店, 1997.

[5] Kenneth C. Litkowski. Sense information for disambiguation: Confluence of supervised and unsupervised methods. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Success and Future Direction*, pp. 47–53, 2002.

[6] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR–045, 1995.

[7] Bernhard Schölkopf, Alex J. Smola, R. Williamson, and P. Bartlett. New support vector algorithms. *Neural Computation*, Vol. 12, pp. 1083–1121, 2000.

[8] 白井清昭, 八木恒和. 辞書定義文を用いた低頻度語のための語義曖昧性解消モデルの学習. 情報処理学会情報処理学会自然言語処理研究会 (NL-158-20), pp. 127–132, 2003.

[9] Mark Stevenson and Yorick Wilks. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, Vol. 27, No. 3, pp. 321–349, 2001.

[10] 玉垣隆幸, 白井清昭. 読解支援システムのための語義曖昧性解消に関する研究. 言語処理学会第 9 回年次大会, pp. 481–484, 2003.