AdaBoostを用いた語義の曖昧性解消

中野 桂吾
筑波大学 情報学類
hyper@viplab.is.tsukuba.ac.jp

平井 有三
筑波大学 電子・情報工学系
hirai@is.tsukuba.ac.jp

1 はじめに

語義曖昧性解消 (Word Sense Disambiguation; WSD) は, 機械翻訳や情報検索など様々な分野で必要とされる技術である. 本稿では, 語義曖昧性解消のコンテスト SENSEVAL2 日本語辞書タスク [6] に対して AdaBoost を適用した結果を述べる. 辞書タスクは, 岩波国語辞典の語釈の中から目的とする単語の適切な語釈を選ぶことから成る.

自然言語処理の分野では、曖昧性解消問題をクラス分類問題として扱う研究が盛んである. AdaBoost は理論的にも実験的にも幅広く研究されており, 自然言語処理の分野においても文書分類 [9], 品詞タグ付け [10], WSD [2] など幅広い分野で応用され, その有用性が示されている. ブースティングの基本的なアイデアは, 多くの単純な規則を結び付けることによって精度の高い規則を得る, というものである. この単純な規則, すなわち弱仮説は任意のものが使用できるので, 適切な弱仮説を選ぶことにより, 比較的可読性の高い学習結果が得られることもブースティングの利点の一つである.

本稿の構成は以下の通りである. 2節において AdaBoost の解説と, どのような文脈情報を用いるかについて述べる. 3節では, 学習結果を辞書タスクの評価データに適用し, その評価を行う. 最後に 4節では, 本稿のまとめを述べる.

2 学習システムの概要

本節では, AdaBoost を用いた学習システムの構成法について述べる.SENSEVAL2 日本語辞書タスクでは曖昧性解消を行う単語が前もって決められている. 名詞 50 種類, 動詞 50 種類, 計 100 種類の単語に対し, それぞれ 100 事例づつ, 計 10000 事例が評価データとなっている. 本稿で用いるシステムは各単語に対して語義を識別する分類器を学習する必要があるので, 100 種類の分類器を独立に AdaBoost によって学習する. 以下に素性セットの構成法と AdaBoost の多値分類問題への拡張版である AdaBoost.MH アルゴリズムを解説する.

2.1 素性セットの構成

本節では, 語義曖昧性解消に用いる文脈情報, すなわち素性セットの構成法について述べる. 本稿では, 構文解析 (JUMAN[5] と KNP パーサ [4] を用いた) で得られる対象単語と係り受け関係にある語を, 文脈を表す素性として用いる. それらは以下の通りである.

1. 対象単語自身[1]
2. 対象単語を含む文節に共起する単語 (b)
3. 対象単語を含む文節の係り元の文節の付属語列 (dl1) と, その文節の主辞と付属語列 (dl2)
4. 対象単語を含む文節の付属語列＋係り先の主辞 (dr)

例として、「アメリカとの経済関係を早急に改善する必要がある」という文章で, 「関係」が対象単語である場合, 文脈を表す素性として以下のような素性が得られる.

(との dl1),(アメリカとの dl2),(経済 b),(関係 tar),(を b),(を改善 dr)

しかし, このように単語のみを用いた素性セットは非常にスパースであり, 学習セットが小さい場合にはうまく学習できない. 特に名詞は他の品詞と比較して圧倒的に単語の異なり数が多い。そこで素性に用いる名詞を抽象化することが考えられる. 本稿では日本語語彙体系 [3] のシソーラスを用いて, 上記で得られた素性セットを拡張した. 日本語語彙体系は, 名詞を約３０００の階層的な意味クラスに分類している. 個々の単語は複数の意味クラスを持っていることもあるが,

[1] juman と RWC の品詞体系は異なるために実際には目的単語を部分文字列として含む単語となる場合もある.

本稿では可能な全ての意味コードを列挙する．ただし，固有名詞や数詞に関しては（5 人間），(388 場所），(362 組織），(数詞）を与える．上記の例では，「アメリカ」の意味クラスとして「388 場所」、「経済」の意味クラス「1168 制度 (経済)」「1880 節約」と，「改善」の意味クラス「2094 改革」を用いて

(< 388 場所>との dl2),(< 1168 制度 (経済) > b),(< 1880 節約> b),(を< 2094 改革> dr)

を素性セットに加えることになる．意味コードを付与する際の問題は，単語が複数の意味クラスを持っているために逆に精度が落ちる危険があることである．

2.2 AdaBoost.MH

AdaBoost.MH は AdaBoost の多値問題への拡張版である．AdaBoost.MH は各クラス（WSD では対象単語の語義）に対して，「そのクラスか、それ以外のクラスか」に分類することによって k 分類問題を二値分類問題に帰着する（いわゆる one versus rest 方式）．各分類器の予測に対して確信度が得られるので，最も確信度の高いクラスをシステムの予測クラスとする．以下に AdaBoost.MH アルゴリズムを示す．

1. m 個の訓練データ $(x_1, y_1), \ldots, (x_m, y_m)$ を入力として得る．x_i は特徴ベクトル，y_i は $y \in 1, \ldots, k$ をみたす正解ラベルである（k はクラス数）．
2. 訓練データの事例の重みの初期値として $D_1(i,l) = 1/mk$ を与える．ただし $i = 1, \ldots, m, l = 1, \ldots, k$ とする．
3. 各ラウンド $t = 1, \ldots, T$ に対し，4 と 5 を繰り返す．
4. 重み D_t に従って学習し，弱仮説 $h_t(x,l)$ を得る．$h_t(x,l)$ の符合がプラスであれば x はクラス l に属していると予測し，マイナスであれば属していないと予測する．$h_t(x,l)$ の絶対値は予測の確信度を表す．弱仮説については後述する．
5. 次式によって各訓練データの重みを更新する．

$$D_{t+1} = \frac{D_t(i,l)\exp(-Y(i,l)h_t(x_i,l))}{Z_t}$$

ここで，Z_t は $\sum_i \sum_k D_{t+1}$ を 1 にするための正規化項で，$Y(i,l)$ は $y_i = l$ のとき 1，そうでなければ -1 を返す関数である．

6. 最後に以下の線形和で最終的な分類器を得る．

$$H(x) = \arg\max_l \sum_{t=1}^{T} h_t(x,l)$$

重みの更新式により $h_t(x,l)$ によって正しく分類された事例の重みは減らされ，間違って分類された場合には重みが増やされる．このようにブースティングは分類の難しい事例に集中して学習するために過学習しにくいアルゴリズムであると言われている．T ラウンド後に今までに得られた弱仮説の線形和をとり，最終仮説を得る．

2.3 弱仮説の獲得

弱仮説として任意のものが使えるが，本稿では深さ 1 の決定木を用いる．これは「素性 e があれば $h(x,l) = c_{1l}$，なければ $h(x,l) = c_{0l}$ を出力する」というものである[2]．AdaBoost.MH の訓練誤差は高々 $k\prod_t Z_t$ となることが分かっている [2] ので，各ラウンドにおいて Z を最小化する $h_t(x,l)$ を選ぶ，すなわち Z を最小化するような素性を各ラウンドにおいて選択することとなる．以下に素性 e を用いた弱仮説 $h_t(x,l)$ を得る手順を示す．

1. 素性 e を持つ事例を X_1，持たない事例を X_0 に分割する
2. $$W_b^{jl} = \sum_i D_t(i,l)[x_i \in X_j \wedge Y(i,l) = b]$$

   を計算する．ここで $[\pi]$ は π が真ならば 1，そうでなければ 0 を返す関数である．そして

   $$c_{jl} = \frac{1}{2}\log\left(\frac{W_{+1}^{jl} + \epsilon}{W_{-1}^{jl} + \epsilon}\right)$$

   を得る．ϵ は平滑化項で $\epsilon = 1/mk$ とした
3. 素性 e を用いたときの正規化項の上限値

   $$\hat{Z}(e) = 2\sum_{j\in\{1,0\}}\sum_l \sqrt{W_{+1}^{bl}W_{-1}^{bl}}$$

   を計算する

これらの詳細は文献 [8] を参照されたい．c_{jl} は x が素性 e を文脈素性として持つとき ($j = 1$)，あるいは持たないとき ($j = 0$) に x がクラス l に属する “分布 D_t によって重みづけされた” 対数尤度比を表す．各ラウンドにおいて全ての素性の中で $\hat{Z}$ を最小にする素性（すなわち分類性能が最も良い仮説）を一つだけ選ぶ．各ラウンドでは、分布 D_t にのみ依存して弱仮説が決定されるので、同じ素性の有無を調べる仮説（ただし予測値は異なる）が複数回選択されることもある．このような弱仮説の獲得法を用いることの利点の一つは，分類に必要な素性のみを選ぶために，学習結果がコンパクトになることである．

3 実験及び考察

3.1 実験環境

学習セットとして RWC コーパス [11] を用いる．このコーパスは毎日新聞 1994 年の 3000 記事に対し岩波国

[2]直感的には例えば，「受ける」という動詞の目的語として「ボール」があれば，「うけとめる」の意味で使われている可能性が高く，「うけいれる」の意味である可能性は低いと予測するような弱仮説．

表 1: 機械学習の手法と精度（構文情報のみ）

	BL	NB	DL	AB
名詞	71.24%	76.97%	76.94%	**77.14%**
動詞	73.96%	77.98%	78.34%	**78.63%**
総合	7.260%	77.48%	77.64%	**77.89%**

表 2: 機械学習の手法と精度（意味クラスを付与）

	BL	NB	DL	AB
名詞	71.24%	77.20%	77.15%	**78.47%**
動詞	73.96%	78.61%	78.51%	**79.73%**
総合	7.260%	77.91%	77.83%	**79.10%**

語辞典に基づいた語義タグが付与されている. 素性セットの選択法として構文解析から得られる結果のみを用いる方法と, 意味クラス情報を付加する方法の２種類で実験し, 意味クラスを用いることの有用性を検証する. 評価方法は SENSEVAL2 に従った mixed-grained scoring [3]を用いた.

ブースティングのラウンド数は１０分割のクロスバリデーションを用いて各単語に対し最適なラウンド数を推定した. ただし学習に時間がかかるために, ラウンド数の推定には, 各ラウンドにおいて全素性の 10%をランダムに選び, その中から最適な弱仮説を選ぶ LazyBoosting[2] を用いている[4]. また比較対象としてナイーブベイズ, 決定リスト [1] を用いた.

3.2 実験結果

表 1,2 に名詞 (50 単語), 動詞 (50 単語), 及び両者を合わせた結果を示す. BL は各単語で最も頻出する語義に分類するベースライン. NB はナイーブベイズ, DL は決定リスト, AB は AdaBoost を示す.

どの機械学習の手法を用いても、ベースラインよりも精度が高かった。さらに意味クラスを付与しているかどうかに関わらず, AdaBoost が他の機械学習の手法よりも精度が高かった. しかも AdaBoost は意味クラスを素性に追加した方が精度が 1%程度上昇している. 他の手法ではそれほど精度が上昇していないことから, AdaBoost が分類に必要な素性のみを規則化していると考えられる.

[3]正解の語義 ID とシステムの語義 ID が完全に一致していなくても, 語義の階層構造にしたがって部分点を与える評価方式

[4]ただしラウンド数は 50 から 500 までの 50 刻みで推定した.

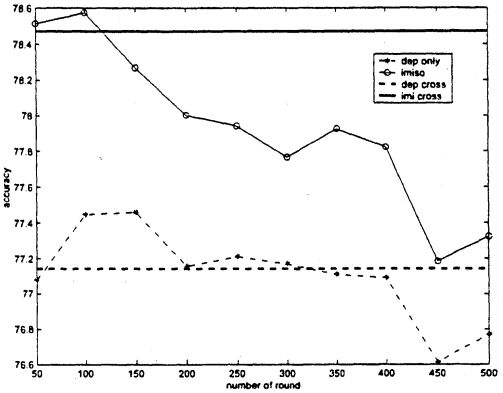

図 1: ラウンド数と精度（名詞）

3.3 AdaBoost のラウンド数と精度

図 1,2 にブースティングのラウンド数と精度の関係を示す. 図中の “dep only” は構文情報のみを用いた場合の精度, “imiso” は意味クラス情報を追加したときの精度, “dep cross”, “imi cross” はそれぞれクロスバリデーションで最適ラウンド数を推定した場合の精度である. この図から名詞に関してはラウンド数が進むにしたがって精度が下がっていることが見て取れる. ブースティングは過学習しにくいアルゴリズムであると言われているが [13], 実際には過学習を起こすことが分かる. また, クロスバリデーションで決定したラウンド数よりもラウンド数を固定した方が精度が良いという場合もあった.

一方で動詞ではラウンド数を重るにしたがって精度が改善されており, 過学習しているようには見えない. しかし実際には名詞でも過学習しない単語もあれば, 動詞であっても過学習する単語もあった. クロスバリデーションによる推定がうまくいかない場合があったり, 過学習する原因の一つは, 分類すべきクラスに対して訓練データが少ないからであると考えられる.

3.4 関連研究との比較

SENSEVAL2 のデータを用いて語義曖昧性解消を行った研究として文献 [7],[12] が挙げられる. 文献 [7] では, 様々な素性（同一文中の単語や, 品詞情報など）を用いて, ナイーブベイズで学習した結果, 79.3%という精度を得ている. 本手法は文献 [7] で用いられた素性セットよりもサイズが小さく, 精度も最高で 79.1%と

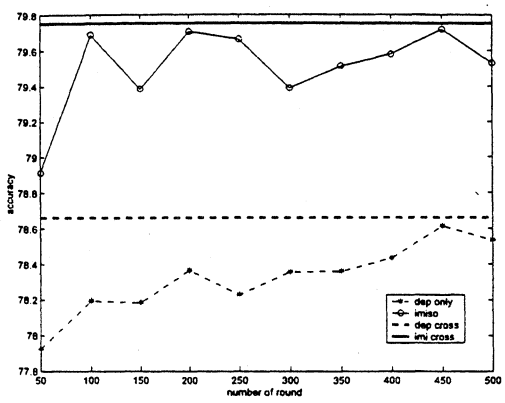

図 2: ラウンド数と精度（動詞）

わずかに劣る程度であったので, 文献 [7] での最高精度と十分比較可能であると考える. 文献 [12] では, 学習アルゴリズムとして決定リストを用い, 学習データに岩波国語辞典の語釈の例文を加えることによって精度を向上させている（最高精度は 77.91%）.

AdaBoost の学習結果は, 分類に必要な素性のみを選択するため, ナイーブベイズや決定リストに比べてコンパクトである. 各単語に対して分類器が必要となることを考えると, 分類規則はできるだけシンプルなほうが良い. この点でも AdaBoost が語義曖昧性解消に対し有効な点である.

3.5 今後の課題

AdaBoost は分類が困難な事例に集中して学習するために, 訓練データに誤りがある場合には, 誤りのあるデータに集中して学習してしまうために精度が落ちることがある. 逆に AdaBoost が重みを集中した事例は訓練データに誤りがある可能性が高く, それらを人手で修正することで精度の向上を期待できる.

また本稿では, 構文解析によって素性セットを構成したが, 得られた素性は全く同じであっても異なる意味で使われている例があった. このような例では別の素性を追加しなければ正しい分類はできない. 品詞レベル（あるいは単語レベル）で曖昧性解消のために最適な素性セットは異なることも予想される. このため係り受け関係だけでなく, 語義曖昧性解消に有効な他の素性の発見が今後の課題である.

4 まとめ

本稿では, AdaBoost を SENSEVAL2 日本語辞書タスクに適用した結果を報告し, 意味クラスを導入することによって精度が向上することを示した. 本稿で用いた素性セットの中では AdaBoost が最も良い精度を示し, AdaBoost が日本語の語義曖昧性解消に対して有効であることが分かった.

参考文献

[1] David Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restroration in Spanish and French. *32th Annual Meeting of the Computational Linguistics,(1994),p.88-95*

[2] Gerard Escudero, Lluís Màrquez, and German Rigau. Boosting applied to word sense disambiguation. In *LNAI 1810: Proceedings of the 12th European Conference on Machine Learning, ECML, pages 129-141, Barcelona, Spain, 2000.*

[3] NTT コミュニケーション技術研究所, 日本語語彙体系, 岩波書店

[4] 黒橋 禎夫, 日本語構文解析システム KNP version 2.0b6 使用説明書

[5] 黒橋 禎夫, 長尾 真, 日本語形態素解析システム JUMAN 使用説明書 version 3.6

[6] 黒橋 禎夫, 白井 清昭,SENSEVAL2 日本語タスク, 電子情報通信学会 自然言語処理研究会 NLC2001-36

[7] 村田 真樹 ほか,SENSEVAL2J 辞書タスクでの CRL の取り組み, 電子情報通信学会 自然言語処理研究会 NLC2001-40

[8] Robert E. Shapire and Yorman Singer. Improved boosting algorithms using confidence-rated prediction. *Machine Learning, 37(3):297-336, December 1999.*

[9] Robert E. Shapire and Yorman Singer. BoosTexter: a boosting-based system for text categorization. *Machine Learning, 39(2/3):135-168, May/June 2000.*

[10] Steven Aveney, Rober E. Shapire and Yorman Singer. Boosting Applied to Tagging and PP Attachment. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999.*

[11] 白井 清昭 ほか, 岩波国語辞典を利用した語義タグ付きテキストデータベースの作成, 情報処理学会自然言語処理研究会, Vol.2001,No.9,pp.117-122

[12] 八木 豊 ほか, 決定リストを用いた語義曖昧性解消, 電子情報通信学会 自然言語処理研究会 NLC2001-40

[13] Yoav Freund and Robert E. Schapire. (翻訳：安倍 直樹), ブースティング入門, 人工知能学会誌 Vol.14 No.5 pp.771-779 (1999)