Vol. 44 No. 12 情報処理学会論文誌 Dec. 2003

EMアルゴリズムの最適ループ回数の予測を用いた語義判別規則の教師なし学習

新納浩幸[†] 佐々木稔[††]

本論文では Nigam らが文書分類問題に対して提案した EM アルゴリズムを用いた教師なし学習の手法(EM 法と略す) を語義判別問題に適用する．EM 法はラベル付きの訓練データから学習した分類器の精度を，ラベルなし訓練データにより向上させる．ただし確実に精度が向上する保証はなく，逆に精度が悪化する場合も多い．ここでは EM アルゴリズムの最適な繰返し回数を推定することで精度の低下を防ぐ．この推定のために，2 つの手法 CV-EM と CV-EM2 を提案する．CV-EM は交差検定による推定である．CV-EM2 は交差検定の結果と 2 つの判断を用いた推定であり，EM 法が修正なしで利用できるか，EM 法を用いるべきかどうかを判断することで推定を行う．SENSEVAL2 の日本語辞書タスクで課題となった名詞 50 単語を用いた実験では，ラベル付き訓練データのみから学習した語義判別規則の精度は 0.7678 であり，EM 法は 0.7356 となり精度が下がるが，CV-EM や CV-EM2 は精度を向上させた．しかも CV-EM2 の精度 0.7856 は現在公開されている正解率の最高値に匹敵する．また動詞 50 単語についても CV-EM や CV-EM2 の効果を確認できた．

Unsupervised Learning of Word Sense Disambiguation Rules by Estimating an Optimum Iteration Number in the EM Algorithm

Hiroyuki Shinnou[†,?] and Minoru Sasaki[††,?]

In this paper, we apply an unsupervised method using EM algorithm, which Nigam et al. have proposed, to word sense disambiguation (WSD) problems. That method, which we call as the EM method in this paper, improves the performance of the classifier learned through small labeled data by using huge unlabeled data. However, the EM method does not always improve it, and often gets worse it. To overcome this issue, we estimate an optimum iteration number of the EM algorithm. To estimate that number, we propose two methods, CV-EM and CV-EM2. The CV-EM is the cross validation. The CV-EM2 is a method combining the cross validation and two ad-hoc judgments, which judge whether we can use the EM method without modification, and judge whether the EM method is useful. In experiments, we solved 50 noun WSD problems in Japanese Dictionary Task in SENSEVAL2. The classifier learned through only labeled data produced 0.7678 precision. The EM method got worse 0.7356 precision, but both of CV-EM and CV-EM2 improved the precision. Especially, the precision of CV-EM2 (0.7856) is a match for the best public score. Furthermore, CV-EM and CV-EM2 were confirmed to be also effective for verb WSD problems.

1. はじめに

本論文では EM アルゴリズムを用いた教師なし学習(以下 EM 法と略記する) を語義判別問題に適用する．ただし単純に適用すると精度が低下する場合もある．ここでは EM アルゴリズムの最適な繰返し回数を予測することで精度の低下を防ぐ．

自然言語処理の個々の問題を分類問題に定式化し，帰納学習の手法を用いて解決するというアプローチは非常に大きな成功を収めている．語義判別問題に関しては，多義語の単語数は膨大であるために，人手により語義判別規則を作成することは困難であり，機械学習のアプローチをとるのが現実的である．そのため語義判別問題に対する機械学習の手法は多数提案されており [1)]，それらはどれも語義判別問題を分類問題に定式化して帰納学習の手法を用いている．しかしこのアプローチには帰納学習で必要とされるラベル付き訓練データを作成するコストが高いという問題がある．この問題の1 つの対処方法として教師なし学習が提案されている．ここでいう教師なし学習とは，小さなラベ

† 茨城大学工学部システム工学科
Department of Systems Engineering, Faculty of Engineering,Ibaraki University
†† 茨城大学工学部情報工学科
Department of Computer and Information Sciences, Faculty of Engineering, Ibaraki University

ル付きの訓練データから学習される分類器の精度を，大量のラベルなし訓練データを用いて高める手法をいう☆．代表的な手法として Co-training[2] と EM 法[3] がある．どちらも本来は文書分類問題に対して考案されており，語義判別問題に適用できるかどうかは明らかではない．語義判別問題は自然言語処理の重要な課題であり，それらの手法が語義判別問題に適用できることが望ましい．ここでは EM 法を適用する．

Nigam らが提案した EM 法は Naive Bayes の分類器[4] に EM アルゴリズムを組み合わせたものである．EM アルゴリズムは，本来，部分的に欠損値のある不完全な観測データ $x_1, x_2, \cdots, x_N$ から，そのデータを発生する確率モデル $P_\theta(x)$ を推定する手法である．$P_\theta(x)$ は未知パラメータ θ を含み，$P_\theta(x)$ の推定とは，θ の推定に帰着される．分類問題の教師なし学習では，ラベル付き訓練データが完全な観測データ，ラベルなし訓練データがラベルを欠損値とした不完全な観測データとなる．EM アルゴリズムは，現時点での θ を使って，モデル $P_\theta(c|x_i)$ のもとでの $\log P_{\hat{\theta}}(x_i, c)$ の期待値をとる（ E-step）．次に，この期待値を最大にするような $\hat{\theta}$ を求める（ M-Step）．$\hat{\theta}$ を新たな θ として先の E-step と M-step を繰り返す．ここで c は欠損値となるラベルである．Nigam らは $P_\theta(x)$ を Naive Bayes のモデル，θ をラベル c_i のもとで素性 f_k が起こる条件付き確率 $p(f_k|c_i)$ に設定している．これにより，$P_\theta(c|x_i)$ と $\log P_{\hat{\theta}}(x_i, c)$ が具体的な形で与えられ，計算が可能となる．

ただし EM 法を単純に適用すると，精度が逆に悪化する場合もある．EM アルゴリズムの繰返し回数に比例して精度が向上し，とりうる最高精度で EM アルゴリズムが収束すれば理想的だが，そのような都合の良い問題は少ない．多くの場合，ある地点まで精度が上がっても，最終的にはそれよりも低い精度で収束する．悪くすると，ラベル付き訓練データのみから学習された分類器の精度よりも低い精度に収束する場合もある．

この問題への対処として，ここでは EM アルゴリズムの最適な繰返し回数を推定する．実際の学習では，推定された繰返し回数で EM アルゴリズムを終了する．そして最適な繰返し回数を推定する手法として，本論文では 2 つの手法を試みる．1 つは交差検定である．ここではラベル付き訓練データを 3 等分し，2 つを新たなラベル付き訓練データ，残りの 1 つをテストデータとする．新たなラベル付き訓練データと大量のラベルなし訓練データを用いて，EM 法を適用し，テストデータに対して最高精度を記録する繰返し回数を推定値とする手法である．この手法を CV-EM と名付ける．もう 1 つの手法は，交差検定の結果にある判断基準を設けて推定を行う手法である．ある判断基準とは，EM 法を修正なしに単純に適用できるかどうか，と EM 法が利用可能かどうか，という 2 つの判断である．この手法を CV-EM2 と名付ける．

実験では SENSEVAL2 の日本語辞書タスク[5] を用いた．名詞の場合，ラベル付き訓練データのみから得られた Naive Bayes（ 以下 NB と略す）の正解率は 0.7678 であった．単純に EM 法を適用した場合，正解率は 0.7356 に下がってしまった．しかし，CV-EM や CV-EM2 を用いることで，正解率は 0.7788 と 0.7856 に向上した．この CV-EM2 の正解率 0.7856 は，現在公開されている辞書タスクの名詞での最高正解率に匹敵する．また動詞に対しても本手法を適用した．ラベル付き訓練データのみから学習した NB の正解率は 0.7816，単純な EM 法は 0.7874 であったのに対し，CV-EM はそれよりも高い 0.7922，さらに CV-EM2 はそれよりもわずかに高い 0.7926 であった．名詞，動詞ともに最適な繰返し回数の予測を行うことで，NB やオリジナルの EM 法よりも高い精度を達成でき，本手法の有効性が示せた．

2. Naive Bayes による語義判別

事例 x が素性のリストとして，以下のように表現されたとする．

$$x = (f_1, f_2, \cdots, f_n)$$

x の分類先のクラスの集合を $C = \{c_1, c_2, \cdots, c_m\}$ とおく．分類問題は $P(c|x)$ の分布を推定することで解決できる．実際に，x のクラス c_x は以下の式で求まる．

$$c_x = arg \max_{c \in C} P(c|x)$$

ベイズの定理を用いると，

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

なので，結局，以下が成立する．

$$c_x = arg \max_{c \in C} P(c)P(x|c)$$

ここで，$P(c)$ は比較的簡単に推定できる．問題は，$P(x|c)$ の推定だが，これは現実的には難しい．Naive Bayes のモデルは，この推定に以下の仮定を導入する．

☆ 本論文では教師なし学習という用語を用いているが，近年では bootstrap あるいは semi-supervised と呼ばれている．

$$P(x|c) = \prod_{i=1}^{n} P(f_i|c) \quad (1)$$

$P(f_i|c)$ の推定は比較的容易であるために，結果として $P(x|c)$ が推定できる [4]．Naive Bayes を使った分類がうまくゆくかどうかは，式 (1) の仮定をできるだけ満たすような素性を選択することである．文書分類であれば，各素性を各単語の生起に設定することで，Naive Bayes が有効であることが知られている．

ここでは素性を設定するために，まず語義判別の手がかりとなる属性として以下のものを設定した．

e1	直前の単語
e2	直後の単語
e3	前方の内容語2 つまで
e4	後方の内容語2 つまで
e5	e3 の分類語彙表の番号
e6	e5 の分類語彙表の番号

たとえば，語義判別対象の単語を「記録」として，以下の文を考える（形態素解析され各単語は原型に戻されているとする）．

過去/最高/を/記録/する/た/。

この場合，「記録」の直前，直後の単語は「を」と「する」なので，‘e1=を’，‘e2=する’となる．次に，「記録」の前方の内容語は「過去」，「最高」なので，ここから「記録」に近い順に 2 つとり，‘e3=過去’，‘e3=最高’が作られる．またここでは句読点も内容語に設定しているので，「記録」の後方の内容語は「する」と「。」となり，‘e4=する’，‘e4=。’が作られる．次に「最高」の分類語彙表 [6] の番号を調べると，3.1920_4 である．ここでは分類語彙表の 4 桁目と 5 桁目までの数値をとることにした．つまり ‘e3=最高’ に対しては，‘e5=3192’ と ‘e5=31920’ が作られる．同様に「過去」の分類語彙表の番号 1.1642_1 から ‘e5=1164’ と ‘e5=11642’ が作られる．次は「する」の分類語彙表を調べるはずだが，ここでは平仮名だけで構成される単語の場合，分類語彙表の番号を調べないことにした．これは平仮名だけで構成される単語は多義性が高く，無意味な素性が増えるので，その問題を避けたためである．もしも分類語彙表上で多義になっていた場合には，それぞれの番号に対して並列にすべての素性を作成する．

結果として，上記の例文に対しては以下の 10 個の素性が得られる．

e1=を，e2=する，e3=最高，e3=過去，
e4=する，e4=。，e5=3192，e5=31920，
e5=1164，e5=11642

上記の例文をデータ x としておくと，データ x はこの 10 個の素性を要素として持つリストとして表せる．

$x =$ (e1=を, e2=する, e3=最高, e3=過去,
e4=する, e4=。, e5=3192, e5=31920,
e5=1164, e5=11642)

3. EM 法

分類問題の解決に Naive Bayes が有効に使えれば，Nigam らが提案した教師なし学習（EM 法）が利用できる．そこでは EM アルゴリズムを用いることで，ラベルなし訓練データを用いて，ラベル付き訓練データから学習された分類器の精度を向上させる．

ここではポイントとなる式とアルゴリズムだけを示す [3]．

基本となるのは，あるクラス c_j のもとで，素性 f_i が発生する確率 $P(f_i|c_j)$ を求めることである．これは以下の式で求まる．この式は頻度 0 の部分を考慮したスムージングを行っている．

$$P(f_i|c_j) = \frac{1 + \sum_{k=1}^{|D|} N(f_i, d_k) P(c_j|d_k)}{|F| + \sum_{m=1}^{|F|} \sum_{k=1}^{|D|} N(f_m, d_k) P(c_j|d_k)} \quad (2)$$

式 (2) の D はラベル付けされた訓練データとラベル付けされていない訓練データを合わせた訓練データ全体を示す．D の各要素を d_k で表す．F は素性全体の集合である．F の各要素を f_m で表す．また，$N(f_i, d_k)$ は，訓練事例 d_k に含まれる素性 f_i の個数を表す．ここでの設定では1 事例の中に，重複した素性は含まれないので，$N(f_i, d_k)$ は 0 か 1 の値をとる．また事例と素性の組合せは膨大であるが，事例が含む素性は少数なので，$N(f_i, d_k)$ の値はほとんどの場合 0 である．$P(c_j|d_k)$ は訓練データがクラス c_j を持つ確率である．ラベル付けされた訓練データに対しては，0 か 1 の値をとる．ラベル付けされていない訓練データに対しては，最初は 0 であるが☆，EM アル

☆ ここで 0 以外の値を付与することも考えられるが，オリジナルの EM 法にならってここでは 0 とした．また 0 以外の適切な値を付与する効果は未知である．

ゴリズムの繰返しによって，徐々に適切な値に更新されてゆく．

式 (2) を利用して，以下の分類器が作成できる．

$$P(c_j|d_i) = \frac{P(c_j)\prod_{f_n \in K_{d_i}} P(f_n|c_j)}{\sum_{r=1}^{|C|} P(c_r)\prod_{f_n \in K_{d_i}} P(f_n|c_r)} \quad (3)$$

ここで，C はクラスの集合である．K_{d_i} は訓練事例 d_i に含まれる素性の集合を示す．$P(c_j)$ はクラス c_j の発生確率であり，以下の式で計算する．

$$P(c_j) = \frac{1 + \sum_{k=1}^{|D|} P(c_j|d_k)}{|C| + |D|} \quad (4)$$

EM アルゴリズムは式 (3) を利用して，ラベル付けされていない事例 d_i に対して，$P(c_j|d_i)$ を求める (E-step)．次に式 (2) を利用して，$P(f_i|c_j)$ を求める (M-step)．この E-step と M-step を交互に繰返して，$P(f_i|c_j)$ と $P(c_j|d_i)$ を収束するまで更新してゆく．ここでは収束の条件として，繰返しに際しての $P(f_i|c_j)$ の差が $8 \cdot 10^{-6}$ 以下になるか，繰返しの回数が 10 回に達した場合に収束したとした．

4. EM 法の問題と最適な繰返し回数の推定

EM 法は必ずしもラベル付き訓練データから学習された分類器の精度を向上させるとは限らない．理論的な背景は不明であるが，論文 [7] でもこの問題は指摘されている．またこの問題への対処は論文 [8] の今後の課題としてもあげられている．

本論文ではこの問題の対処として，EM アルゴリズムの最適な繰返し回数を推定することを試みる．実際の学習では推定された繰返し回数で学習を終了することで，精度の低下を避ける．

本論文では，最適な繰返し回数を推定するために 2 つの手法（CV-EM と CV-EM2）を試みる．

両手法とも，以下の 2 点を注意する．1 点目は，推定された繰返し回数以前に，実際の EM アルゴリズムが収束した場合には，収束回数が実際の停止回数となる点である．2 点目は，推定された繰返し回数が収束回数である場合，推定時における収束回数と実際の学習時の収束回数が異なる場合もある点である．

4.1 CV-EM

CV-EM は交差検定である．ここでは訓練データを 3 分割し，1 つをテストデータ，残りの 2 つを新たなラベル付きの訓練データとする．この 2/3 になったラベル付き訓練データと，ラベルなし訓練データを用いて EM 法を試みる．EM アルゴリズムの各繰返し終了時点で得られた判別規則を用いて，残りの 1/3 から得たテストデータを判別することで，各繰返し終了時点での正解率を得る．3 分割されたラベル付き訓練データのどれをテストデータにするかで，3 通りの組合せがあるので，それらすべての組合せの 3 通りの実験を行い，正解率はそれらの平均とする．そして最高正解率を出した繰返し回数を最適な繰返し回数と推定する．

4.2 CV-EM2

CV-EM を利用しても最適な値を推定できる保証はない．CV-EM2 は別の観点から最適な繰返し回数の推定を行う．

具体的には 2 つの判断を用意する．第 1 の判断は，「EM 法を修正なしに用いることができるかどうか」である．単純に考えれば，EM 法によって精度低下が生じるのは奇異であり，通常は EM 法を単純に適用してもよいと考えられる．このため最初にこの判断を行う．この第 1 の判断が「YES(できる)」であれば，推定値は収束回数となる．「NO(できない)」であれば，第 2 の判断に移る．第 2 の判断は「EM 法が有効かどうか」である．第 1 の判断で EM 法を単純に適用することができないと判断した場合，次に判断すべきことは「何らかの対処を行って EM 法を使う(EM 法は有効である)」か，あるいは「EM 法を使わない(EM 法は有効でない)」かである．この第 2 の判断が「YES(有効である)」の場合は，推定値は CV-EM と同じ推定値とする．「NO(有効でない)」の場合は，推定値は 0 とする☆．

問題は第 1 の判断「EM 法を修正なしに用いることができるかどうか」と第 2 の判断「EM 法が有効かどうか」をどのように判定するかである．ここでは交差検定における繰返し回数とその時点で学習された分類器の精度との関係から，上記 2 つの判断を行うことにした．

まず第 1 の判断「EM 法を修正なしに用いることができるかどうか」を行うために，交差検定において，EM アルゴリズムの繰返しを収束するまで行って得た判別規則の精度と，EM アルゴリズムの繰返しを 1 回だけ行って得た判別規則の精度を比較する．もしも前者(収束回数時の正解率)が後者(繰返し回数 1 回の正解率)以上の場合には「YES(できる)」と判定し，そうでない場合「NO(できない)」と判定する．これはラベルなし訓練データが EM 法によって正の方向に作用するか，負の方向に作用するかに注目した結果である．ラベルなし訓練データが正の方向に作用するか，負の方向に作用するかは，ラベル付き訓練データ

☆ 繰返し回数が 0 であるということは，EM 法を用いないことを意味する．

だけから得られた判別規則の精度には関係ない．そのため比較対象となるのは，EM アルゴリズムの繰返しを収束するまで行って得た判別規則と EM アルゴリズムの繰返しを 1 回だけ行って得た判別規則とした．

次に第 2 の判断「EM 法が有効かどうか」を行うために，交差検定において，EM アルゴリズムの繰返しを 1 回だけ行って得た判別規則の正解率と EM 法を用いない場合の判別規則の正解率とを比較する．もしも前者（繰返し回数 1 回の正解率）が後者（EM 法を用いない場合の正解率）以上の場合には「YES（有効である）」と判定し，そうでない場合「NO（有効でない）」と判定する．これは，EM 法により正解率が向上する場合には，繰返し 1 回目で正解率が向上するというヒューリスティクスを用いた結果である．

5. 実　　験

本手法の有効性を確認するために，SENSEVAL2 の日本語辞書タスクで課題とされた名詞 50 単語に関する語義判別を試みた．

SENSEVAL2 の日本語辞書タスクは，単純な語義判別問題である．対象単語は名詞 50 単語，動詞 50 単語の計 100 単語である．これら 100 単語は語義の頻度分布のエントロピーを考慮して選定されており，語義判別が容易なものから困難なものまでバランス良く選定されている．ラベル付きの訓練データは 1 単語平均して名詞は 177.4 事例，動詞は 172.7 事例用意されている．またテストデータは各単語に対して 100 問のテストが用意されている．つまり名詞に対しては計 5,000 問，動詞に対しても計 5,000 問のテストが行える．ただし，ラベルなし訓練データは SENSEVAL2 からは提供されていない．これは通常のテストが使えそうだが，実際は制限がある．それはラベル付きの訓練データを作成した際に用いた単語辞書や品詞分類を合わせる必要があるからである．そのためここでは RWC テキストデータベース第 2 版に収められた毎日新聞 95 年度版の1 年分の記事を利用して，ラベルなし訓練データを収集した．このデータはラベル付き訓練データのもとになったデータであり，同一の形態素解析システムを用いて形態素解析されている．収集できたラベルなし訓練データの数は 1 単語平均して名詞は 7585.5 事例，動詞は 6571.9 事例である．

次に名詞 50 単語に対する交差検定の結果を表 1 に示す．表 1 において，`base` とあるのは，ラベル付き訓練事例のみから学習した分類器の正解率である．`1-st` とあるのは，EM アルゴリズムを 1 度だけ行った後に得た分類器の正解率である．`last` とあるのは，EM アルゴリズムが収束した後に得た分類器の正解率である．`max` とあるのは，EM アルゴリズムが収束するまでに記録した最高正解率である．`CV-EM` とあるのは，最高正解率を出した EM アルゴリズムの繰返し回数，つまり CV-EM の推定値である．`CV-EM2` とあるのは，CV-EM2 による推定値である．表中の ∞ は収束回数を表す．

表 1 は CV-EM2 の推定値を算出するソース，および CV-EM と CV-EM2 の比較のために提示した．CV-EM と CV-EM2 の比較は考察に述べる．ここでは CV-EM2 の算出の例を示す．たとえば「間」の

表 1 交差検定による最適繰返し回数の推定
Table 1 Estimating an optimum iteration number by cross validation.

単語	base	1-st	last	max	CV-EM	CV-EM2
間	0.687	0.687	0.649	0.687	1	1
頭	0.681	0.623	0.623	0.681	0	∞
一般	0.737	0.743	0.707	0.743	1	1
一方	0.862	0.885	0.885	0.897	3	∞
今	0.642	0.642	0.638	0.642	1	1
意味	0.658	0.658	0.685	0.685	∞	∞
疑い	0.980	0.980	0.921	0.990	3	3
男	0.956	0.956	0.938	0.956	3	3
開発	0.881	0.890	0.881	0.890	1	1
核	0.942	0.923	0.923	0.942	0	∞
関係	0.838	0.838	0.825	0.841	4	4
気持ち	0.724	0.712	0.769	0.769	∞	∞
記録	0.721	0.743	0.735	0.750	2	2
技術	0.898	0.878	0.867	0.898	0	0
現在	0.950	0.971	0.095	0.971	1	1
交渉	0.965	0.965	0.894	0.972	3	3
国内	0.870	0.864	0.831	0.870	0	0
言葉	0.601	0.607	0.601	0.607	3	3
子供	0.697	0.705	0.669	0.705	1	1
午後	0.662	0.669	0.510	0.676	2	2
市場	0.752	0.745	0.680	0.752	0	0
市民	0.673	0.589	0.561	0.673	0	0
社会	0.900	0.900	0.879	0.900	1	1
少年	0.978	0.978	0.978	0.978	∞	∞
時間	0.814	0.820	0.361	0.820	1	1
事業	0.797	0.791	0.739	0.797	0	0
時代	0.631	0.627	0.658	0.658	∞	∞
自分	0.920	0.920	0.920	0.920	∞	∞
情報	0.741	0.692	0.649	0.741	0	0
姿	0.644	0.644	0.614	0.644	1	1
精神	0.930	0.930	0.930	0.930	∞	∞
対象	0.912	0.912	0.912	0.912	∞	∞
代表	0.910	0.915	0.902	0.926	5	5
近く	0.783	0.797	0.841	0.841	∞	∞
地方	0.719	0.684	0.637	0.719	0	0
中心	0.948	0.955	0.955	0.955	∞	∞
手	0.598	0.606	0.614	0.630	5	∞
程度	0.961	0.961	0.961	0.961	∞	∞
電話	0.878	0.878	0.683	0.878	1	1
同日	0.582	0.560	0.530	0.597	3	0
花	0.947	0.947	0.933	0.947	2	2
反対	0.957	0.957	0.950	0.957	2	2
場合	0.719	0.729	0.771	0.771	∞	∞
前	0.874	0.883	0.896	0.902	3	∞
民間	0.959	0.959	0.959	0.959	∞	∞
娘	0.853	0.853	0.863	0.863	∞	∞
胸	0.625	0.589	0.625	0.625	∞	∞
目	0.633	0.602	0.508	0.633	0	0
もの	0.495	0.511	0.498	0.511	1	1
問題	0.970	0.968	0.968	0.970	0	∞

表 2 実験結果(名詞)
Table 2 Experiment results (Noun).

単語	NB	EM 法	CV-EM	CV-EM2	理想値
間	0.810	0.800	0.820	0.820	0.820
頭	0.600	0.640	0.600	0.640	0.660
一般	0.880	0.860	0.890	0.890	0.890
一方	0.820	0.880	0.880	0.880	0.890
今	0.900	0.900	0.900	0.900	0.900
意味	0.450	0.530	0.530	0.530	0.530
疑い	1.000	0.950	0.980	0.980	1.000
男	0.920	0.890	0.920	0.920	0.920
開発	0.620	0.630	0.620	0.620	0.630
核	0.710	0.770	0.710	0.770	0.810
関係	0.850	0.900	0.900	0.900	0.900
気持ち	0.650	0.650	0.650	0.650	0.660
記録	0.740	0.710	0.730	0.730	0.770
技術	0.960	0.920	0.960	0.960	0.960
現在	0.970	0.090	0.980	0.980	0.980
交渉	1.000	0.880	1.000	1.000	1.000
国内	0.460	0.580	0.460	0.460	0.580
言葉	0.450	0.400	0.400	0.400	0.450
子供	0.670	0.730	0.720	0.720	0.730
午後	0.770	0.650	0.860	0.860	0.860
市場	0.770	0.550	0.770	0.770	0.770
市民	0.670	0.630	0.670	0.670	0.670
社会	0.820	0.830	0.830	0.830	0.830
少年	0.920	0.900	0.900	0.900	0.920
時間	0.540	0.150	0.540	0.540	0.540
事業	0.690	0.700	0.690	0.690	0.710
時代	0.720	0.770	0.770	0.770	0.780
自分	1.000	1.000	1.000	1.000	1.000
情報	0.770	0.640	0.770	0.770	0.770
姿	0.550	0.630	0.610	0.610	0.630
精神	0.650	0.660	0.660	0.660	0.660
対象	0.980	0.980	0.980	0.980	0.980
代表	0.850	0.950	0.960	0.960	0.980
近く	0.740	0.870	0.870	0.870	0.870
地方	0.700	0.720	0.700	0.700	0.720
中心	0.980	0.980	0.980	0.980	0.980
手	0.470	0.480	0.470	0.480	0.480
程度	1.000	1.000	1.000	1.000	1.000
電話	0.840	0.650	0.830	0.830	0.850
同日	0.810	0.510	0.570	0.810	0.810
花	0.990	0.970	0.990	0.990	0.990
反対	0.970	0.970	0.970	0.970	0.970
場合	0.820	0.910	0.910	0.910	0.920
前	0.860	0.910	0.920	0.910	0.920
民間	1.000	1.000	1.000	1.000	1.000
娘	0.880	0.880	0.880	0.880	0.880
胸	0.710	0.770	0.770	0.770	0.790
目	0.180	0.170	0.180	0.180	0.180
もの	0.310	0.270	0.270	0.270	0.310
問題	0.970	0.970	0.970	0.970	0.970
平均	0.7678	0.7356	0.7788	0.7856	0.7964

表 3 実験結果(動詞)
Table 3 Experiment results (Verb).

単語	NB	EM 法	CV-EM	CV-EM2	理想値
与える	0.710	0.780	0.780	0.780	0.780
言う	0.940	0.940	0.940	0.940	0.940
受ける	0.590	0.640	0.590	0.640	0.640
訴える	0.840	0.870	0.870	0.870	0.880
生まれる	0.690	0.830	0.820	0.830	0.830
描く	0.580	0.560	0.560	0.560	0.580
思う	0.900	0.890	0.890	0.890	0.900
買う	0.830	0.830	0.830	0.830	0.830
かかる	0.580	0.570	0.580	0.580	0.580
書く	0.720	0.660	0.720	0.720	0.720
変わる	0.920	0.920	0.920	0.920	0.920
考える	0.990	0.990	0.990	0.990	0.990
聞く	0.560	0.550	0.550	0.550	0.560
決まる	0.960	0.960	0.960	0.960	0.960
決める	0.930	0.930	0.930	0.930	0.930
来る	0.840	0.850	0.860	0.850	0.860
加える	0.890	0.890	0.890	0.890	0.890
超える	0.780	0.820	0.850	0.820	0.880
知る	0.970	0.970	0.970	0.970	0.970
進む	0.490	0.500	0.500	0.500	0.500
進める	0.970	0.950	0.970	0.970	0.970
出す	0.350	0.290	0.350	0.350	0.360
違う	1.000	1.000	1.000	1.000	1.000
使う	0.970	0.970	0.970	0.970	0.970
作る	0.690	0.750	0.780	0.750	0.780
伝える	0.750	0.760	0.760	0.760	0.760
出来る	0.810	0.810	0.810	0.810	0.810
出る	0.590	0.640	0.640	0.640	0.640
問う	0.690	0.790	0.790	0.790	0.790
取る	0.320	0.340	0.320	0.340	0.370
狙う	0.990	0.990	0.990	0.990	0.990
残す	0.790	0.790	0.790	0.790	0.790
乗る	0.540	0.540	0.540	0.540	0.540
入る	0.360	0.360	0.360	0.360	0.360
図る	0.920	0.920	0.920	0.920	0.920
話す	1.000	0.870	1.000	1.000	1.000
開く	0.860	0.940	0.940	0.940	0.940
含む	0.990	0.990	0.990	0.990	0.990
待つ	0.520	0.500	0.510	0.510	0.520
まとめる	0.790	0.800	0.800	0.800	0.800
守る	0.790	0.710	0.700	0.710	0.790
見せる	0.980	0.980	0.980	0.980	0.980
認める	0.890	0.890	0.890	0.890	0.890
見る	0.730	0.710	0.730	0.730	0.730
迎える	0.890	0.890	0.890	0.890	0.890
持つ	0.570	0.620	0.570	0.570	0.620
求める	0.870	0.870	0.870	0.870	0.870
読む	0.880	0.880	0.880	0.880	0.880
よる	0.970	0.970	0.970	0.970	0.970
分かる	0.900	0.900	0.900	0.900	0.900
平均	0.7816	0.7874	0.7922	0.7926	0.7992

場合， 1-st の値 0.687 が last の値 0.649 よりも大きいので，第 1 の判断は NO となり，第 2 の判断に移る． base の値は 0.687 であり， 1-st の値はそれ以上なので，CV-EM2 の推定値は CV-EM と同じとなり，1 となる．また「 頭」の場合， last の値 0.623 は 1-st の値 0.623 以上であるので，第1の判断は YES となり，CV-EM2 の推定値は収束回数(∞) となる．

次に名詞 50 単語に関して，Naive Bayes (表中の NB)，収束するまで EM アルゴリズムを実行した単純な EM 法，提案手法の CV-EM と CV-EM2 および理想的な繰返し 回数の推定が行えた場合(理想値) の各結果を表 2 に示す．またここでの正解率は解答結果に部分点を与える mixed-gained scoring という方式 [5] を用いている．ラベル付き訓練データのみから学習する Naive Bayes の正解率は 0.7678 であった．単純に EM 法を適用すると正解率は 0.7356 となり，正解率が下がってしまう．しかし CV-EM により 0.7788 まで改善される．さらに CV-EM2 では 0.7856 まで改善される☆．この値は現在公開されてい

☆ 検定を行っていないので，ここで得られた改善の程度に有意差があるのかどうかという疑問がある．しかし，本論文の趣旨は EM 法の改善であり，その点は明らかなので，検定は行わなかった．また「 EM 法自体が WSD に対して不向きでは」という査

る辞書タスクの名詞での正解率の最高値に匹敵する．現在公開されている辞書タスクの名詞での正解率の中で好成績のものは，Naive Bayesと様々な属性を利用した手法[1]により0.782，SVMとUDC素性を除いた様々な属性を利用した手法[1]により0.785，決定木とアダブーストを利用した手法[9]により0.7847などがある．CV-EM2により得られた0.7856はそれらの値と同等以上である．また上記にあげた3つの他手法は，どれも素性として構文素性を用いているが，本手法では構文素性を利用しておらず，実際の語義判別時に構文解析を必要としない点も本手法の優位な点である．

同様に動詞50単語に関して，得られた結果が表3である．動詞による実験では，Naive Bayesの正解率が0.7816であったが，EM法を用いることで0.7874となり，正解率が向上している．さらにCV-EM（0.7922）やCV-EM2（0.7926）でもさらに正解率を向上させることができ，提案手法の効果が確認できる．

6. 考　　察

6.1 EM法がうまくいかない原因

なぜEM法では正解率が下がる場合があるのかは，様々な理由が考えられる．その理由の1つとして，ラベル付き訓練データ，ラベルなし訓練データ，テストデータの3者間の語義の分布の違いが考えられる．今，ラベル付き訓練データの語義の分布をL，ラベルなし訓練データの語義の分布をU，テストデータの語義の分布をTとおく．理想的にはラベル付き訓練データ，ラベルなし訓練データ，テストデータは全体のデータからのランダムサンプルであるので，L，U，Tは同一の分布になるはずである．だが実際は異なる．ラベル付き訓練データにラベルなし訓練データを併用して学習することは，大ざっぱに捉えれば，L+Uの分布から学習していると見なせる．今，分布A，B間の距離を$d(A,B)$で表すと，EM法が有効になるのは，$d(L+U,T) < d(L,T)$の場合であり，逆に$d(L+U,T) > d(L,T)$のときはEM法が逆効果になると考えられる．

上記の点を確認するために，分布間の距離をKL情報量で測る調査を行った．$L+U$の分布は，EMアルゴリズムが収束した後の式(4)を利用して得る．結果を表4にまとめた．表4の行は$d = d(L,T) - d(L+U,T)$の値が正（テストデータに語義分布が近づく）か負（テストデータの語義分布が離れる）の観点でわけ，列はEM法による「精度向上」と「精度悪化」に分けた．「精度向上」とはEM法により5%以上の精度向上があった単語，「精度悪化」とはEM法により5%以上の精度が悪化した単語を意味する．表の要素は該当する名詞23単語中の単語数を表す．

表4 語義分布の正解率への影響
Table 4 Influence of precision through sense distribution.

	テストデータの語義分布に近づく	テストデータの語義分布から離れる
精度向上	6	7
精度悪化	2	8

この結果から，ラベルなしデータを用いてテストデータの語義分布に近付けたかどうかと，EM法による精度向上が起こるかどうかには，緩い相関はありそうだが，完全に関連があると結論付けることは難しい．EM法によって精度悪化が起こるかどうかは，他の要素も影響していると考えられる．ただしEM法で正解率が最も悪化する単語「現在」を見てみると，dの値も50単語中最小の-0.30をとる．この単語についてはテストデータの語義分布が関連していると考えられる．EM法では正解率が下がる原因のさらなる調査は今後の課題である．

6.2 CV-EMとCV-EM2の比較

CV-EM2は現実的にはほぼCV-EMと同じ値を推定する．

表1によると，名詞50単語に関して，CV-EMとCV-EM2が同じ値を推定するのは43単語(86%)である．これはCV-EM2が第1の判断でYESと判断し，収束回数を推定値とした場合，CV-EMも多くの場合，収束回数を推定値とするからである．CV-EM2の第1の判断でYESと判断することは，EM法を修正なしで用いることができることを意味するので，CV-EMにおいても収束回数が推定値となるのは自然である．またCV-EM2が第1の判断でNOと判断し，第2の判断に移り，第2の判断でYESと判断すれば，それはCV-EMと同じ推定値をとることになる．結局，大ざっぱに考えて，CV-EM2がCV-EMと大きく異なるのは，第1の判断でNO，第2の判断でNOとなるケースで，しかも交差検定の結果はEMアルゴリズムの繰返しに対して単調に精度が減少しない図1ようなケースである．

図1は交差検定の結果を示している．横軸がEMの繰返し回数で縦軸が正解率である．EMアルゴリズムの繰返し回数l回目では精度が下がるが，途中精度が

読者からの疑問もあったが，NBが日本語辞書タスクにおいて高い成績を収めている点，EM法では分類に使える情報が増える点などから不向きではないと考えている．

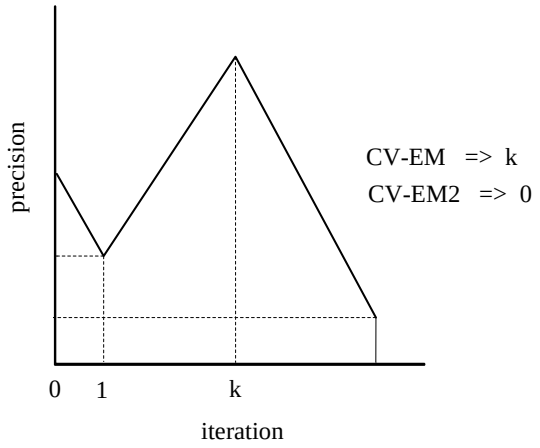


図 1 CV-EM と CV-EM2 が異なる典型的な例
Fig. 1 Typical difference between CV-EM and CV-EM2.

上がり，最終的には低い精度で収束している．このような場合，CV-EM では図中の k を推定値とするが，CV-EM2 の推定値は 0 となる．

実際の学習でも CV-EM2 は CV-EM とほとんど差が生じない．動詞ではその値は +0.0004 と小さいし，名詞では +0.0068 と比較的大きいようだが，個々の単語で見るとほとんど差はない．名詞の場合「同日」に大きな差が生じたために結果的に差が出ているだけである「同日」の場合，実際に図 1 の現象が生じている．交差検定での「同日」の正解率と実際の「同日」の正解率の変化とを図 2 と図 3 に示す．図 2 の交差検定では正解率がある程度まで上昇するが，図 3 に示すように，実際には急激に正解率が落ちている．

また CV-EM が CV-EM2 よりも成績が良かった単語は名詞に対して「前」の 1 単語，動詞に対しては「来る」「超える」「作る」の 3 単語の計 4 単語であったのに対し，CV-EM2 が CV-EM よりも成績が良かった単語は，名詞に対して「頭」「核」「手」「同日」の 4 単語，動詞に対しては「受ける」「生まれる」「取る」「守る」の 4 単語の計 8 単語であった．各々の単語における正解率の差は小さいが，CV-EM2 の推定法の方が若干優れていると言える．

また CV-EM2 による正解率が理想値と一致するのは，名詞 50 単語中 29 単語の 58% である．また残り 21 単語のうち，15 単語は理想値との差が 0.02 以下である．この点から CV-EM2 の推定のほぼ 9 割は有効であったと考えられる．また理想値との差が 0.05 以上あるのは「国内」(−0.05) と「言葉」(−0.12) である．CV-EM および CV-EM2 とも，この 2 単語が全体の正解率を下げる大きな原因になっている．これ

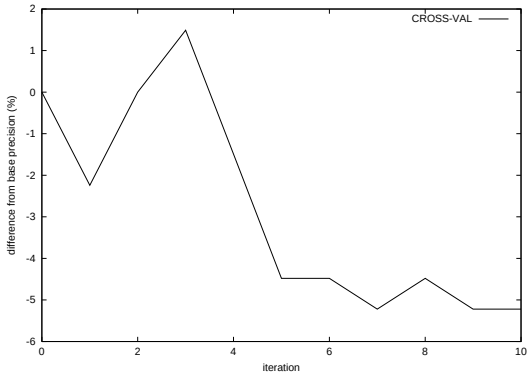


図 2 「同日」の交差検定
Fig. 2 Cross validation for 'doujitsu.'

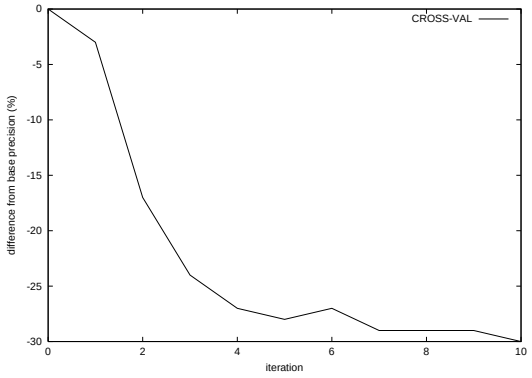


図 3 実際の「同日」
Fig. 3 Actual evaluation for 'doujitsu'.

らの単語の交差検定での各繰返し時の EM 法の正解率と，実際の問題での各繰返し時の EM 法の正解率とを，図 4 と図 5 に示す．ただし図ではグラフの形を見るためにグラフの始点を 0 に設定し，各繰返し時の EM 法の正解率は始点との差でとっている．

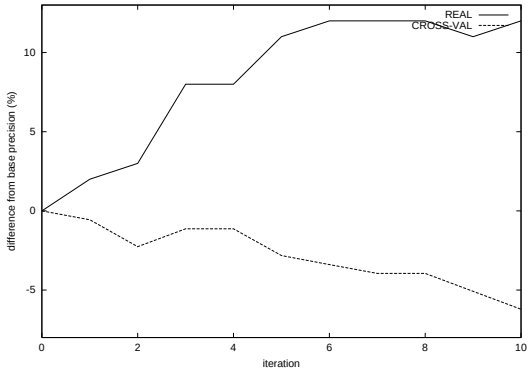


図 4 交差検定と実際との比較(「国内」)
Fig. 4 Comparison of cross validation and actual evaluation ('kokunai').

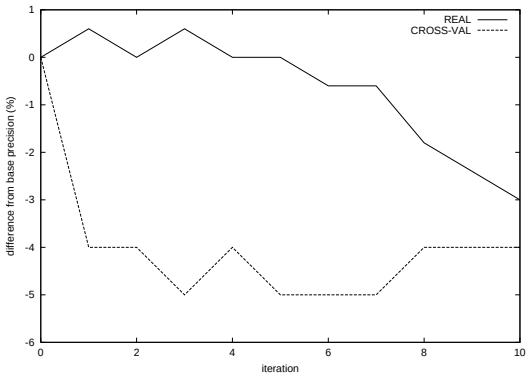


図 5 交差検定と実際との比較(「言葉」)
Fig. 5 Comparison of cross validation and actual evaluation ('kotoba').

図 4 の「国内」の場合，交差検定では EM 法を用いると正解率が下がってゆく．一方，実際の問題では EM 法により正解率が向上してゆく．このようにまったく逆のパターンが生じると推定が大きく誤る．ただし，このように交差検定ではまったく効果がなく，実際の問題では大きな効果があるようなケースは「国内」だけであった．図 5 の「言葉」の場合，交差検定で 1 回目の繰返しで正解率が上がるが，以降は徐々に下がる．実際の問題では 1 回目から正解率が下がる．この 1 回目の繰返しの結果の違いが推定回数を誤らせている．この 2 つの単語は特異なケースである．細かい調査を行ってさらに精度の高い最適な繰返し回数の推定法を考案する必要がある．

6.3 動詞の語義判別問題に対する教師なし学習

今回の実験では，名詞も動詞もどちらに対しても，本手法により EM 法を大きく改善できた．ただし最終的に得られた正解率の値だけ見ると，名詞の 0.7856 は非常に良い値だが，動詞の 0.7926 はそこそこの値でしかない．この原因の1 つとして，動詞に対しては教師なし学習というアプローチ自体が難しいことが考えられる．もしも完璧に最適な EM アルゴリズムの繰返し回数を推定できたとすれば，名詞の正解率は 0.7964，動詞の正解率は 0.7992 となる．ベースとなるラベル付き訓練データのみから学習された分類器の正解率に対して，名詞は 1.037 倍であるが，動詞は 1.022 倍である．これは動詞の方が名詞よりも EM 法の効果が低いことを示している．

教師なし学習では，素性の独立性が影響していると予想している．たとえば，データ x が 2 つの素性 f_1 と f_2 から構成されているとする．もしデータ x のクラス c_x が素性 f_1 から判別されたとき，素性 f_2 からクラス c_x を判別する確率 $P(c_x|f_2)$ が高くなる．問題は $P(c_x|f_2)$ を高くすることが本当に妥当かどうかである．妥当であれば教師なし学習はうまく機能するが，妥当でなければ教師なし学習は破綻する．直感的には，この妥当性を保証するのが素性の独立性と思われる．名詞の場合，多くの場合，対象単語の左文脈がその名詞を修飾する単語列となり，対象単語の右文脈がその名詞を格に持つ動詞となる．これらはそれ自身で語義を判別できる情報を持ち，しかもそれらはある程度独立している．一方，動詞に対してはそのような都合の良い解釈が見つからない [10]．そのため教師なし学習がうまく機能する保証がない．Naive Bayes でも素性の独立性を仮定している．しかし現実問題では独立でなくても非常に判別力の高い分類器が構築できる．動詞に対してある程度 EM 法および本手法がうまく機能したのは，そのような Naive Bayes の頑健性に負うところが大きいと思われる．逆に，教師なし学習がうまく機能する保証がありそうな名詞に対して EM 法で正解率が下がったのは，ラベル付き訓練データ，ラベルなし訓練データ，テストデータの関係のある種のアンバランスさが原因だと予想している．たとえば，ラベル付き訓練データ中の誤りや，テストデータとラベル付き訓練データのラベル付けの不統一性などが考えられる．この確認は今後の課題である．

6.4 過学習とメタパラメータの推定

EM 法がうまくいかない原因は未知であることは述べたが，過学習が生じているからとも考えられる．論文 [7] では単語のクラスの割合が大きく変化することを原因としてあげているが，これも一種の過学習であろう．

過学習への対策としては精巧なスムージングの手法を導入することが考えられる．ここでは EM アルゴリズムのターゲット関数として式 (2) を用いているが，この式で導入しているスムージング法は，未知の素性の頻度に 1 を与えるという簡易な方法である☆．この部分をもっと精緻なものにすることで過学習を避けられる可能性はある．

また過学習への対策としては学習を早期に打ち切る手法も有望である．一般に学習回数の増加に従って，学習誤差は小さくなってゆくが，予測誤差がある時点から増大する現象が生じる．これが過学習である．この予測誤差が増大する付近で学習を終了させればよい．ここで提案した CV-EM や CV-EM2 は，このような手法としてもとらえられる．そして学習を停止する回

☆ 予期尤度推定法と呼ばれている．

数を推定することは，メタパラメータの推定問題である．メタパラメータの推定には一般に Leave-one-out 法が用いられる．これは交差検定の一種であり，ラベル付き訓練データが n 個ある場合に，$n-1$ 個を学習データ，残りの 1 個をテストデータとする方法である．本論文の CV-EM や CV-EM2 ではラベル付き訓練データを 3 等分しているので，Leave-one-out 法よりも精度が低い可能性もある．ただし，今回の実験ではラベル付き事例数が約 170 事例あり比較的多い．そのため 3 等分程度でも十分と考えた．また Leave-one-out 法の場合，必要とする計算時間が膨大となる問題もある．またメタパラメータの推定にはブートストラップ法 [11] という手法もあり，今回の実験では Leave-one-out 法よりも処理時間が少なくて済むと思われるので，今後検討したい．

6.5 今後の課題

今後の課題としては 3 つある．1 つは動詞に対してさらに正解率をあげることである．ここでの素性を使う限りは，これ以上の改善は難しいので，別種の素性を導入する必要があるだろう．2 つ目は最適な繰返し回数の推定法の改善である．本実験で最適な繰返し回数を完璧に予想できれば，名詞の正解率は 0.7964 となる．推定手法をさらに高度化することで，正解率はまだまだ改善できる余地がある．3 つ目は EM 法により正解率が下がる原因の解明である．このことが頑健性の高い教師なし学習の実現の鍵だと考える．

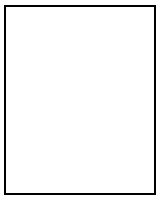
7. おわりに

本論文では EM 法を語義判別問題に適用した．ただし単純に適用すると精度が低下する場合もあるので，EM アルゴリズムの最適な繰返し回数を推定することで精度の低下を防いだ．この推定のために，CV-EM と CV-EM2 という 2 つの推定方法を提案した．CV-EM は交差検定であり，CV-EM2 は交差検定に 2 つの判断規則を組み合わせたものである．SENSEVAL2 の日本語辞書タスクを用いた実験では，CV-EM と CV-EM2 は EM 法を大きく改善した．特に CV-EM2 の名詞の語義判別に対する成績は，現在公開されている正解率の最高値に匹敵する．今後の課題は，動詞に対して別種の素性を導入すること，最適な繰返し回数の推定法を改善すること，EM 法により正解率が下がる原因を解明することの 3 点である．

参考文献

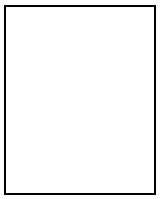
1) 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均: SENSEVAL2J 辞書タスクでの CRL の取り組み - 日本語単語の多義性解消における種々の機械学習手法と素性の比較 -, 自然言語処理, Vol. 10, No. 3, pp. 115–133 (2003).
2) Blum, A. and Mitchell, T.: Combining Labeled and Unlabeled Data with Co-Training, *11th Annual Conference on Computational Learning Theory (COLT-98)*, pp. 92–100 (1998).
3) Nigam, K., McCallum, A., Thrun, S. and Mitchell, T.: Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134 (2000).
4) Mitchell, T.: *Machine Learning*, McGraw-Hill Companies (1997).
5) 白井清昭: SENSEVAL-2 日本語辞書タスク, 自然言語処理, Vol. 10, No. 3, pp. 3–24 (2003).
6) 国立国語研究所: 分類語彙表, 秀英出版 (1994).
7) Tsuruoka, Y. and Tsujii, J.: Training a Naive Bayes Classifier via the EM Algorithm with a Class Distribution Constraint, *7th Conference on Computational Natural Language Learning (CoNLL-03)*, pp. 127–134 (2003).
8) 新納浩幸: EM アルゴリズムを用いた教師なし学習の日本語翻訳タスクへの適用, 自然言語処理, Vol. 10, No. 3, pp. 61–73 (2003).
9) 中野桂吾, 平井有三: AdaBoost を用いた語義の曖昧性解消, 言語処理学会第 8 回年次大会, pp. 659–662 (2002).
10) Shinnou, H.: Learning of word sense disambiguation rules by Co-training, checking co-occurrence of features, *3rd international conference on Language resources and evaluation (LREC-2002)*, pp. 1380–1384 (2002).
11) 石井健一郎, 上田修功, 前田英作, 村瀬洋: わかりやすいパターン認識, オーム社出版局 (1998).

(平成 14 年 9 月 26 日受付)
(平成 15 年 10 月 16 日採録)



新納 浩幸(正会員)

昭和 36 年生．昭和 60 年東京工業大学理学部情報科学科卒業．昭和 62 年同大学大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，平成 5 年 4 月茨城大学工学部システム工学科助手．平成 9 年 10 月同学科講師，平成 13 年 4 月同学科助教授，現在に至る．博士 (工学)．



佐々木稔(正会員)

昭和 48 年生．平成 8 年徳島大学工学部知能情報工学科卒業．平成 13 年同大学大学院博士後期課程修了．博士 (工学)．現在，茨城大学工学部情報工学科助手．機械学習や統計的手法による情報検索，自然言語処理等に関する研究に従事．言語処理学会会員．