情 報 学 基 礎　68－8
自 然 言 語 処 理 151－8
（２００２．９．１８）

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

茨城大学工学部システム工学科　　茨城大学工学部情報工学科
shinnou@dse.ibaraki.ac.jp　　sasaki@cis.ibaraki.ac.jp

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を語義判別問題に適用する．ただし語義判別問題において，彼らの手法は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り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，交差検定を行い，ラベルなし訓練データが正の情報として働くか，負の情報として働くかに注目することで最適な繰り返し回数を推定する．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分類器の精度は 0.7658 であり，その上で Nigam らの手法を用いた場合は 0.7356 となり精度が下がるが，本手法を用いることで 0.7856 まで精度が向上した．この値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 単語についても本手法の効果を確認できた．

Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in EM algorithm

Hiroyuki Shinnou
Department of Systems
Engineering, Ibaraki University
shinnou@dse.ibaraki.ac.jp

Minoru Sasaki
Department of Information
Engineering, Ibaraki University
sasaki@cis.ibaraki.ac.jp

This paper improves an unsupervised learning method using EM algorithm proposed by Nigam et al. for text classification problems. The original method works well, but often causes worse classification for our concerned problems, word sense disambiguation problems. To avoid it, we estimate the optimum iteration number in EM algorithm. The criterion is based on cross validation and judgement that unlabeled training data is used to improve a classification or not. In experiment, we solved 50 noun WSD problems in Japanese Dictionary Task in SENSEVAL2. The classification learned through only labeled training data produces 0.7658 precision. Original unsupervied method got worse 0.7356 precision, but our improved method produced 0.7856 precision. This is a match for the best public score. Furthermore, our method is confirmed to be also effecitve 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り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変形し，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納めている．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．この問題に対する 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，小さなラベル付きの訓練データから学習される判別規則の精度を，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[1] と EM 法 [4] 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中心的課題であり，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 [2] 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 $logP_{\hat{\theta}}(x_i, c)$ の期待値を取る（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 $logP_{\hat{\theta}}(x_i, c)$ が具体的な形で与えられ，計算が可能となる．

ただしこの EM 法を単純に語義判別問題に適用すると，精度が逆に悪くなる場合もある．EM アルゴリズムの繰り返し回数に比例して精度が向上し，取り得る最高精度で 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規則よりも低い精度で収束する場合もある．この問題の解決のために，ここでは交差検定を用いる．交差検定では，ラベルなし訓練データが正の情報として働くか，負の情報として働くかに注目する．それによって，EM アルゴリズムにおける最適な繰り返し回数を予測する．実際の適用では，予測した回数だけ繰り返して得られた判別規則を学習結果とする．また予測した最適な繰り返し回数が 0 回の場合，これは教師なし学習を行わないことを意味する．

実験では SENSEVAL2 の日本語辞書タスク [8] に関して，本手法を試した．名詞の場合，ラベル付き訓練データのみから得られた正解率は 0.7658 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，本手法を用いることで正解率を 0.7856 まで高められた．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 [10] により 0.7822，アダブーストを利用した手法 [11] により 0.7847 などがある[1]．本手法により得られた 0.7856 はそれらの値よりも良い．また動詞に対しても適用してみた．ラベル付き訓練データのみは 0.7816，単純な EM 法は 0.7874 であったが，最適な繰り返し回数の予測を行うと 0.7926 となり，本手法の有効性が示せた．

2 Naive Bayes による語義判別

ある事例 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 $C = \{c_1, c_2, \cdots, c_m\}$ と置く．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．

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

$P(f_i|c)$ 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 [2]．Naive Bayes を使っ

[1] 最高値は論文 [10] の様々な学習手法を融合した手法によるものだと思われるが，名詞，動詞を合わせたものが 0.7933 という値が出ているだけで，名詞のみの値は公開されていない．

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

例えば，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 つ取り，‘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 [7] 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 アルゴリズムを用いた教師なし学習

分類問題の解決に Naive Bayes が使えれば，Nigam らが提案した教師なし学習が利用できる．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 [4].

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，$N(f_i, d_k)$ は 0 か 1 の値であり，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 アルゴリズムの繰り返しによって，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 を交互に繰り返して，$P(f_i|c_j)$ 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，繰り返しに際しての $P(f_i|c_j)$ の差が $8 \cdot 10^{-6}$ 以下になるか，繰り返しの回数が 10 回に達した場合に収束したとした．

4 最適な繰り返し回数の推定

EM アルゴリズムの最適な繰り返し回数を得るために交差検定を用いる．

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り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，各繰り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 通りの組合わせがあるので，それら全ての組合わせの 3 通りの実験を行い，正解率はそれらの平均とする．

単純には，交差検定で得られた最高正解率を出した繰り返し回数を最適な繰り返し回数と推定すれば良い．しかしこの推定法の場合，一部の単語で大きく正解率を下げてしまい，若干頑健性に欠ける．ここでは新たな推定法を提案する．

ここでは，交差検定で得られた結果から，最適な繰り返し回数を推定するために，2 つの判断ポイントを設ける．

第 1 の判断ポイントは，まず単純に EM 法を適用するかどうかである．常識的に考えれば EM 法を適用して精度が悪くなることはない．そこでまず，交差検定において，EM アルゴリズムの繰り返しを収束するまで行って得た判別規則の精度が，EM アルゴリズムの繰り返しを 1 回だけ行って得た判別規則の精度よりも高い場合には，単純に EM 法が適用できると判断する．この場合，推定結果の繰り返し回数は収束するまでの回数となる．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ている．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 アルゴリズムの繰り返しを収束するまで行って得た判別規則」と「EM アルゴリズムの繰り返しを 1 回だけ行って得た判別規則」とした．

もしも上記の判定法で単純に EM 法を適用できないと判断した場合は，EM 法を使わない方が良いのかどうかの判定を第 2 の判断ポイントとした．それには「EM アルゴリズムの繰り返しを 1 回だけ行って得た判別規則の正解率」が「ラベル付き訓練データだけから得られた判別規則の正解率」よりも悪くなっていた場合に，EM 法を使わないという判断を行う．この場合，推定結果の繰り返し回数は 0 回となる．逆に,「EM アルゴリズムの繰り返しを 1 回だけ行って得た判別規則の正解率」が「ラベル付き訓練データだけから得られた判別規則の正解率」よりも良くなっていた場合には，単純に交差検定で得られた最高正解率を出した繰り返し回数を最適な繰り返し回数とする．

ただし推定された最適な繰り返し回数以前に，実際の EM アルゴリズムが収束した場合には，収束したときの値をとる．

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 5000 問，動詞に対しても計 5000 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．これは通常のテキ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 版に納められた毎日新聞 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． `mxn` とあるのは，最高正解率を出した EM アルゴリズムの繰り返し回数である． `est.` とあるのは，本手法により推定した最適な繰り返し回数である．

表 1: 交差検定による最適繰り返し回数の推定

単語	base	1-st	last	max	mxn	est.
aida	0.687	0.687	0.649	0.687	1	1
atama	0.681	0.623	0.623	0.681	0	5
ippan	0.737	0.743	0.707	0.743	1	1
ippou	0.862	0.885	0.885	0.897	3	4
ima	0.642	0.642	0.638	0.642	1	1
imi	0.658	0.658	0.685	0.685	7	6
utagai	0.980	0.980	0.921	0.990	3	3
otoko	0.956	0.956	0.938	0.956	3	3
kaihatsu	0.881	0.890	0.881	0.890	1	1
kaku_n	0.942	0.923	0.923	0.942	0	10
kankei	0.838	0.838	0.825	0.841	4	4
kimochi	0.724	0.712	0.763	0.769	10	6
kiroku	0.721	0.743	0.735	0.750	2	2
gijutsu	0.898	0.878	0.867	0.898	10	0
genzai	0.950	0.971	0.095	0.971	1	1
koushou	0.965	0.965	0.894	0.972	3	3
kokunai	0.870	0.864	0.831	0.870	0	0
kotoba	0.601	0.607	0.601	0.607	3	3
kodomo	0.697	0.705	0.669	0.705	1	1
gogo	0.662	0.669	0.510	0.676	2	2
shijo	0.752	0.745	0.680	0.752	0	0
shimin	0.673	0.589	0.561	0.673	0	0
shakai	0.900	0.900	0.879	0.900	1	1
shonen	0.978	0.978	0.978	0.978	8	7
jikan	0.814	0.820	0.361	0.820	1	1
jigyou	0.797	0.791	0.739	0.797	0	0
jidai	0.631	0.627	0.631	0.658	10	4
jibun	0.920	0.920	0.920	0.920	10	2
joho	0.741	0.692	0.649	0.741	0	0
sugata	0.644	0.644	0.614	0.644	1	1
seishin	0.930	0.930	0.930	0.930	10	5
taishou	0.912	0.912	0.912	0.912	8	3
daihyou	0.910	0.915	0.902	0.926	5	5
chikaku	0.783	0.797	0.819	0.841	10	5
chihou	0.719	0.684	0.637	0.719	0	0
chushin	0.948	0.955	0.955	0.955	10	2
te	0.598	0.606	0.614	0.630	5	6
teido	0.961	0.961	0.961	0.961	2	2
denwa	0.878	0.878	0.683	0.878	1	1
doujitsu	0.582	0.560	0.530	0.597	3	0
hana	0.947	0.947	0.933	0.947	2	2
hantai	0.957	0.957	0.950	0.957	10	2
baai	0.719	0.729	0.771	0.771	4	4
mae	0.874	0.883	0.896	0.902	3	4
minkan	0.959	0.959	0.959	0.959	10	3
musume	0.853	0.853	0.863	0.863	10	4
mune	0.625	0.589	0.625	0.625	10	5
me	0.633	0.602	0.508	0.633	0	0
mono	0.495	0.511	0.498	0.511	1	1
mondai	0.970	0.968	0.968	0.970	0	2

表 2: 実験結果（名詞）

単語	NB	EM 法	CV-EM	理想値	本手法
aida	0.810	0.800	0.820	0.820	0.820
atama	0.600	0.640	0.600	0.660	0.640
ippan	0.880	0.860	0.890	0.890	0.890
ippou	0.820	0.880	0.880	0.890	0.880
ima	0.900	0.900	0.900	0.900	0.900
imi	0.450	0.530	0.530	0.530	0.530
utagai	1.000	0.950	0.980	1.000	0.980
otoko	0.920	0.890	0.920	0.920	0.920
kaihatsu	0.620	0.630	0.620	0.630	0.620
kaku_n	0.710	0.770	0.710	0.810	0.770
kankei	0.850	0.900	0.900	0.900	0.900
kimochi	0.650	0.650	0.650	0.660	0.650
kiroku	0.740	0.710	0.730	0.770	0.730
gijutsu	0.960	0.920	0.960	0.960	0.960
genzai	0.970	0.090	0.980	0.980	0.980
koushou	1.000	0.880	1.000	1.000	1.000
kokunai	0.460	0.580	0.460	0.580	0.460
kotoba	0.450	0.400	0.400	0.450	0.400
kodomo	0.670	0.730	0.720	0.730	0.720
gogo	0.770	0.650	0.860	0.860	0.860
shijo	0.770	0.550	0.770	0.770	0.770
shimin	0.670	0.630	0.670	0.670	0.670
shakai	0.820	0.830	0.830	0.830	0.830
shonen	0.920	0.900	0.900	0.920	0.900
jikan	0.540	0.150	0.540	0.540	0.540
jigyou	0.690	0.700	0.690	0.710	0.690
jidai	0.720	0.770	0.770	0.780	0.770
jibun	1.000	1.000	1.000	1.000	1.000
joho	0.770	0.640	0.770	0.770	0.770
sugata	0.550	0.630	0.610	0.630	0.610
seishin	0.650	0.660	0.660	0.660	0.660
taishou	0.980	0.980	0.980	0.980	0.980
daihyou	0.850	0.950	0.960	0.980	0.960
chikaku	0.740	0.870	0.870	0.870	0.870
chihou	0.700	0.720	0.700	0.720	0.700
chushin	0.980	0.980	0.980	0.980	0.980
te	0.470	0.480	0.470	0.480	0.480
teido	1.000	1.000	1.000	1.000	1.000
denwa	0.840	0.650	0.830	0.850	0.830
doujitsu	0.810	0.510	0.570	0.810	0.810
hana	0.990	0.970	0.990	0.990	0.990
hantai	0.970	0.970	0.970	0.970	0.970
baai	0.820	0.910	0.910	0.920	0.910
mae	0.860	0.910	0.920	0.920	0.910
minkan	1.000	1.000	1.000	1.000	1.000
musume	0.880	0.880	0.880	0.880	0.880
mune	0.710	0.770	0.770	0.790	0.770
me	0.180	0.170	0.180	0.180	0.180
mono	0.310	0.270	0.270	0.310	0.270
mondai	0.970	0.970	0.970	0.970	0.970
平均	0.7678	0.7356	0.7788	0.7964	0.7856

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，繰り返し回数を交差検定時に最高値を記録した回数で終了させた EM 法（CV-EM），理想的な繰り返し回数の推定が行えた場合（理想値），及び本手法により推定された繰り返し回数で終了させた EM 法（本手法），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 [8] 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 の正解率が 0.7658 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．しかし交差検定を行うと，CV-EM では 0.7788 まで改善される．さらに本手法を用いると，0.7856 まで改善される．この値は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は最高値に匹敵する．

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．さらに交差検定を用いると，CV-EM （0.7922）でも本手法（0.7926）でも更に正解率を向上させることができ，交差検定の効果が確認できる．さらに CV-EM と本手法と比較すると，わずか（+0.0004）ではあるが，本手法の方が正解率が高く，最適な繰り返し回数を推定する本手法の効果が確認できる．

表 3: 実験結果（動詞）

単語	NB	EM 法	CV-EM	理想値	本手法
ataeru	0.710	0.780	0.780	0.780	0.780
iu	0.940	0.940	0.940	0.940	0.940
ukeru	0.590	0.640	0.590	0.640	0.640
uttaeru	0.840	0.870	0.870	0.880	0.870
umareru	0.690	0.830	0.820	0.830	0.830
egaku	0.580	0.560	0.560	0.580	0.560
omou	0.900	0.890	0.890	0.900	0.890
kau	0.830	0.830	0.830	0.830	0.830
kakaru	0.580	0.570	0.580	0.580	0.580
kaku_v	0.720	0.660	0.720	0.720	0.720
kawaru	0.920	0.920	0.920	0.920	0.920
kangaeru	0.990	0.990	0.990	0.990	0.990
kiku	0.560	0.550	0.550	0.560	0.550
kimaru	0.960	0.960	0.960	0.960	0.960
kimeru	0.930	0.930	0.930	0.930	0.930
kuru	0.840	0.850	0.860	0.860	0.850
kuwaeru	0.890	0.890	0.890	0.890	0.890
koeru	0.780	0.820	0.850	0.880	0.820
shiru	0.970	0.970	0.970	0.970	0.970
susumu	0.490	0.500	0.500	0.500	0.500
susumeru	0.970	0.950	0.970	0.970	0.970
dasu	0.350	0.290	0.350	0.360	0.350
chigau	1.000	1.000	1.000	1.000	1.000
tsukau	0.970	0.970	0.970	0.970	0.970
tsukuru	0.690	0.750	0.780	0.780	0.750
tsutaeru	0.750	0.760	0.760	0.760	0.760
dekiru	0.810	0.810	0.810	0.810	0.810
deru	0.590	0.640	0.640	0.640	0.640
tou	0.690	0.790	0.790	0.790	0.790
toru	0.320	0.340	0.320	0.370	0.340
nerau	0.990	0.990	0.990	0.990	0.990
nokosu	0.790	0.790	0.790	0.790	0.790
noru	0.540	0.540	0.540	0.540	0.540
hairu	0.360	0.360	0.360	0.360	0.360
hakaru	0.920	0.920	0.920	0.920	0.920
hanasu	1.000	0.870	1.000	1.000	1.000
hiraku	0.860	0.940	0.940	0.940	0.940
fukumu	0.990	0.990	0.990	0.990	0.990
matsu	0.520	0.500	0.510	0.520	0.510
matomeru	0.790	0.800	0.800	0.800	0.800
mamoru	0.790	0.710	0.700	0.790	0.710
miseru	0.980	0.980	0.980	0.980	0.980
mitomeru	0.890	0.890	0.890	0.890	0.890
miru	0.730	0.710	0.730	0.730	0.730
mukaeru	0.890	0.890	0.890	0.890	0.890
motsu	0.570	0.620	0.570	0.620	0.570
motomeru	0.870	0.870	0.870	0.870	0.870
yomu	0.880	0.880	0.880	0.880	0.880
yoru	0.970	0.970	0.970	0.970	0.970
wakaru	0.900	0.900	0.900	0.900	0.900
平均	0.7816	0.7874	0.7922	0.7992	0.7926

6 考察

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

	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づけたかどうかと，EM 法による精度向上が起るかどうかには，緩い相関はありそうだが，完全に関連があると結論づけることは難しい．EM 法によって精度悪化が起るかどうかは，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 genzai 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更なる調査は今後の課題である．

次に最適な繰り返し回数の推定について考察する．実際に最大正解率を出す繰り返し回数を正しく推定できた割合は 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．この点からほぼ 9 割の推定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，kokunai (-0.05) と kotoba (-0.12) である．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．これらの単語の交差検定での各繰り返し時の EM 法の正解率と，実際の問題での各繰り返し時の EM 法の正解率とを，図 1 と図 2 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，各繰り返し時の EM 法の正解率は始点との差でとっている．

kokunai 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 kokunai だけであった．kotoba の場合，交差検定で 1 回目の繰り返しで正解率が上がるが，以降は徐々

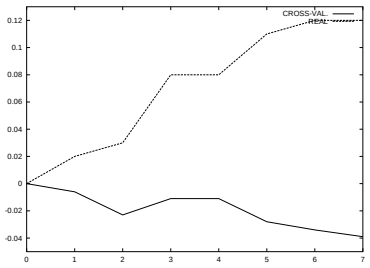

図 1: 交差検定と実際との比較（kokunai）

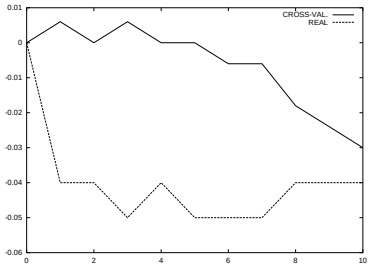

図 2: 交差検定と実際との比較（kotoba）

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 回目の繰り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更に精度の高い最適な繰り返し回数の推定法を考案する必要がある．

また交差検定時に最高値を与えた繰り返し回数を最適な繰り返し回数と推定する手法（CV-EM）と，本手法による推定との差は大きくはない．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い．名詞の場合，doujitu 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．交差検定における doujitu の正解率の変化と，実際の doujitu の正解率の変化とを図 3 と図 4 に示す．交差検定では正解率がある程度まで上昇するが，実際には急激に正解率が落ちる．

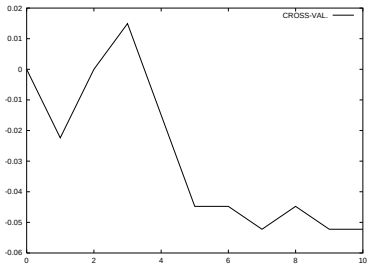

図 3: doujitsu の交差検定

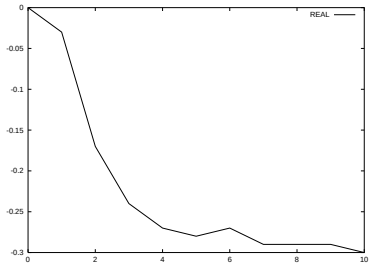

図 4: 実際の doujitsu

また CV-EM が本手法よりも良かったのは名詞に対して mae の 1 単語，動詞に対しては kuru，koeru，tukuru の 3 単語の計 4 単語であるのに対し，本手法が CV-EM よりも良かったのは，名詞に対して atama，kaku_n，te，doujitsu の 4 単語，動詞に対しては，ukeru，umareru，toru，mamortu の 4 単語の計 8 単語である．各々の単語における正解率の差は小さいが，本手法の推定法の方が若干優れていると言える．

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みると，名詞の 0.7856 は非常によい値だが，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 アルゴリズムの繰り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．例えば，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もつ動詞となる．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よい解釈が見つからない [5]．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とも非常に判別

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．例えば，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．

最後に語義判別問題に対する他の教師なし学習との比較を述べておく．

まず Co-training[1] であるが，Co-training は独立な２つの属性させ設定できれば，ベースとなる学習手法を問わないために，応用範囲が広い．また完全に独立な２つの素性が設定できた場合，Co-training は EM 法よりも優れていることが報告されている [3]．しかし Co-training には独立な２つの素性という条件の他に，素性の一貫性という条件も必要になる．この条件のために，実際は Co-training を多値の分類問題に適用することは難しい [9]．一方，EM 法およびそれをベースとした本手法は Naive Bayes の学習を基本とするという制限はあるが，分類問題が多値であっても，原理的に問題はない．そのために，より頑健性の高い現実的な手法と言える．

また語義判別問題に教師なし学習を利用した Yarowsky の研究 [6] との比較についても述べておく．Yarowsky の教師なし学習も，実は Co-training の特殊ケースと見なせる [1]．２つの独立した属性として，１つは前後の文脈，もう１つは「同じ文書内で使われている曖昧な単語の語義は１つに固定される」というヒューリスティクスである．このヒューリスティクスが辞書タスクで設定している語義の細かさに対して，どれほど成立しているかは未知である．またこの手法では，必要とされるラベルなし訓練データは文書，しかも対象単語が複数含まれているような文書となる．これはいかにラベルなしと言えども収集は容易ではない．このため比較対象の実験も困難である．一方，EM 法およびそれをベースとした本手法はその対象単語を含む文が訓練データとなるので，収集は容易であり，より現実的な手法と言える．

今後の課題としては３つある．１つは動詞に対して更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つ目は最適な繰り返し回数の推定法の改善である．本実験で最適な繰り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．推定手法を更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り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，交差検定を行い，ラベルなし訓練データが正の情報として働くか，負の情報として働くかに注目することで最適な繰り返し回数を推定した．SENSEVAL2 の日本語辞書タスクを用いた実験では，本手法は EM 法を大きく改善した．特に名詞では現在公開されている正解率の最高値に匹敵する値を出した．今後の課題は，動詞に対して別種の素性を導入すること，最適な繰り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] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *11th Annual Conference on Computational Learning Theory (COLT-98)*, pp. 92–100, 1998.

[2] Tom Mitchell. *Machine Learning.* McGraw-Hill Companies, 1997.

[3] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-trainig. In *9th International Conference on Information and Knowledge Management*, pp. 86–93, 2000.

[4] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 2000.

[5] Hiroyuki Shinnou. Learning of word sense disambiguation rules by Co-training, checking co-occurrence of features. In *3rd international conference on Language resources and evaluation (LREC-2002)*, pp. 1380–1384, 2002.

[6] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33th Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pp. 189–196, 1995.

[7] 国立国語研究所. 分類語彙表. 秀英出版, 1994.

[8] 黒橋禎夫, 白井清昭. SENSEVAL-2 日本語タスク. 電子情報通信学会言語とコミュニケーション研究会, NLC-36〜48, pp. 1–8, 2001.

[9] 新納浩幸. SENSEVAL2 日本語翻訳タスクに向けて作成した語義判別学習システム Ibaraki. 電子情報通信学会言語とコミュニケーション研究会, NLC-36〜48, pp. 25–30, 2001.

[10] 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. 電子情報通信学会言語とコミュニケーション研究会, NLC-36〜48, pp. 31–38, 2001.

[11] 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.