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

新納浩幸[†]

本論文では，Nigam らによって提案された EM アルゴリズムを利用した教師なし学習の手法を，SENSEVAL2 の日本語翻訳タスクで出題された名詞の語義の曖昧性解消問題に適用する．この手法は，ラベルなしデータをラベルを欠損値とする観測データ，その観測データを発生させるモデルを Naive Bayes モデル，このモデルの未知パラメータをラベル c のもとで素性 f が起る条件付き確率 $p(f|c)$ に設定して，EM アルゴリズムを用いる．結果として，モデルの識別精度が向上する．ここでは識別のための素性として，対象単語の前後数単語の原型や表記という簡易なものに設定した．実験では，ラベル付き訓練データのみから学習した Naive Bayes の正解率が 58.2%，同データから学習した決定リストの正解率が 58.9%（ Ibaraki の公式成績）であったのに対し，ラベル付き訓練データの他にラベルなし訓練データを用いた本手法では，61.8% の正解率を得た．また訓練データの一部の不具合を修正することで，Naive Bayes の正解率を 62.3% に改善できた．更に本手法によりそれを 68.2% に向上させることができた．

キーワード： EM アルゴリズム，教師なし学習，多義語の曖昧性解消，翻訳タスク，SENSEVAL2

Application of unsupervised learning using EM algorithm to Japanese Translation Task

Hiroyuki Shinnou

In this paper, we apply an unsupervised learning method using the EM algorithm which Nigam et al. have proposed for text classification, to disambiguation problems involving noun meanings taken up in Japanese Translation Task of SENSEVAL2. This method uses the EM algorithm, setting up hidden labels of unlabeled data as missing values of observational data, the Naive Bayes model as the generating model, and the conditional probabilities $p(f|c)$ (where f is a feature and c is a label) as parameters of the model. As the result, the learned classifier is improved. In this study, we use only simple features for the classification, which are some words surrounding a target word. In the experiments, the precision of Naive Bayes classifier learned through only labeled data was 58.2%. The precision of the decision list learned through the same data was 58.9%, which is the Ibaraki record in the Translation Task contest. Our unsupervised learning method improved the precision to 61.8% by using unlabeled data in addition to labeled data. Furthermore, by revising a small part of labeled data, the precision levels of the Naive Bayes classifier and our unsupervised learning method were improved to 62.3% and 68.2% respectively.

KeyWords: *EM algorithm, unsupervised learning, word sense disambiguation, Translation Task, SENSEVAL2*

[†] 茨城大学工学部システム工学科，Department of Systems Engineering, Ibaraki University

1　はじめに

本論文では，Nigam らによって提案された EM アルゴリズムを利用した教師なし学習の手法 (Nigam,McCallum,Thrun and Mitchell 2000) を，SENSEVAL2 の日本語翻訳タスク (黒橋禎夫, 白井清昭 2001) で出題された名詞の語義の曖昧性解消問題に適用する．その結果，通常の教師付き学習で得られる分類規則の精度を向上させ得ることを示す．

自然言語処理では個々の問題を分類問題として定式化し，帰納学習の手法を利用して，その問題を解決するというアプローチが大きな成功をおさめている．しかしこのアプローチには帰納学習で必要とされる訓練データを用意しなければならないという大きな問題がある．この問題に対して，近年，少量のラベル付き訓練データから得られる分類器の精度を，大量のラベルなし訓練データによって高めてゆく教師なし学習が散見される．代表的な手法として，Co-training(Blum and Mitchell 1998) と，EM アルゴリズムを利用した手法 (Nigam et al. 2000) がある．Co-training は 2 つの独立した属性 A と B を設定し，一方の属性 A から構築される分類器を利用して，ラベルなしデータにラベル（クラス）を付与する．その中から信頼性のあるラベルが付与されたデータをラベル付き訓練データに加える．このようにして追加されたラベル付き訓練データは，もう一方の属性 B から見るとランダムなサンプルにラベル付けされたデータとして振る舞うので，属性 B から構築される分類器の精度が高まる．これをお互いに作用し合うことで，分類器の精度が高められる．一方，EM アルゴリズムは，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．EM アルゴリズムはパラメータ θ とモデル $P_\theta(x)$ を適切に設定することで，隠れマルコフモデルや文脈自由文法のパラメータ推定，あるいは名詞と動詞間の関係クラスの教師なし学習 (Rooth,Riezler,Prescher,Carroll and Beil 1999)(Torisawa 2001) などに利用できる．そして，Nigam らは文書分類を題材にモデル $P_\theta(x)$ を Naive Bayes のモデル，θ をラベル c のもとで素性 f が起る条件付き確率 $p(f|c)$ に設定することで，教師なし学習を試みている (Nigam et al. 2000)．

Nigam らの EM アルゴリズムを利用した手法や Co-training は，どちらも本来は文書分類に対して考案されており，多義語の曖昧性解消に利用できるかどうかは明らかではない．多義語の曖昧性解消は自然言語処理の中心的な課題であり，これらの手法が適用できることが望ましい．ここでは SENSEVAL2 の日本語翻訳タスクで出題された名詞を題材に，EM アルゴリズムを利用した教師なし学習の手法が名詞の語義の曖昧性解消に適用可能であることを示す．

翻訳タスクの出題形式はある単語 w がマークされた（日本語）文書である．翻訳タスクでは予め，単語 w に関する Translation Memory（以下 TM と略す）と呼ばれる日英の対訳例文の

集合が解答者に配られている．そして翻訳タスクの解答形式は，出題された文書内において注目する単語 w を英訳する際に利用できる TM の例文番号である[1]．つまり，翻訳タスクは単語 w の訳を語義と考えた多義語の曖昧性解消問題となっている．また同時に，翻訳タスクは TM の例文番号をクラスと考えた場合の分類問題として扱える．ここで注意すべきは，翻訳タスクは訓練データを作るのが困難な点である．TM は 1 つの単語に対して平均して 21.6 例文がある．今仮にある単語 w の例文として id_1 から id_{20} までの 20 例文が TM に記載されていたとする．新たに訓練データを作成する場合，単語 w を含む新たな文を持ってきて，id_1 から id_{20} のどれか 1 つのラベルを与える必要がある．○か×かの二者択一は比較的容易であるが，20 個のラベルの中から最も適切な 1 つを選ぶのは非常に負荷のかかる作業である．このように，翻訳タスクは訓練データを新たに作るのが困難であるために，教師なし学習を適用する格好のタスクになっている．

実験では SENSEVAL2 の日本語翻訳タスクで出題された全名詞 20 単語を用いて，本手法の評価を行う．各単語に対して，平均 70 事例（ TM の例文も含む）からなるラベル付き訓練データと，新聞記事 1 年分から取り出した平均 3,354 事例からなるラベルなし訓練データを作成し，本手法を適用した．ラベル付き訓練データだけから学習できた決定リストの正解率は 58.9% (コンテストでの Ibaraki の成績) であり，Naive Bayes による分類器の正解率は 58.2% であった．そして本手法を用いて Naive Bayes による分類器の精度を高めた結果 61.8% まで改善された．また一部，訓練データの不具合を修正することで，Naive Bayes による分類器の正解率を 62.3%，決定リストでの正解率を 63.2% に向上できた．更に，本手法を用いて Naive Bayes による分類器の正解率（ 62.3%）を 68.2% まで高めることができた．

2　Naive Bayes による多義語の曖昧性解消

まず，用語の混乱を避けるため，本論文で用いる「 属性」と「 素性」の区別をしておく．本論文では，例えば,「 対象単語の直前の単語」といった識別のための観点を「 属性」と呼び，属性に具体的な値が与えられたものを「 素性」と呼んでいる．例えば「 対象単語の直前の単語」といった属性を `e1` などで表し，対象単語の直前の単語が，例えば,「 日本」であった場合に，`'e1=`日本`'` と表されたものを素性と呼ぶ．

ある事例 x が素性のベクトルとして，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．

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

[1] 厳密には，翻訳システムも参加できるように，英訳自身を返す解答形式も認められているが，ここでは例文番号を返す解答形式のみを考える．

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．

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

$P(f_i|c)$ 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 (Mitchell 1997)．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．

多義語の曖昧性解消でも式 1 の仮定をできるだけ満たすような素性を選択すれば Naive Bayes が利用できる．本論文では以下の 4 つの属性を利用することにした．

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）

例えば,「胸」の語義は『体の一部としての胸』という語義と『心の中』という語義がある．そして,「その無力感は今も原告たちの胸に染み付いている」という文中の「胸」の語義は『心の中』なので，この事例のクラスは『心の中』となる．また，この文は以下のように形態素解析される．各行が分割された単語であり，第 1 列が表記，第 2 列が原型，第 3 列が品詞を表す．

その	その	連体詞
無力	無力	名詞-形容動詞語幹
感	感	名詞-接尾-一般
は	は	助詞-係助詞
今	今	名詞-副詞可能
も	も	助詞-係助詞
原告	原告	名詞-一般
たち	たち	名詞-接尾-一般
の	の	助詞-連体化

胸	胸	名詞-一般
に	に	助詞-格助詞-一般
染み付い	染み付く	動詞-自立
て	て	助詞-接続助詞
いる	いる	動詞-非自立

この結果から以下の 4 つの素性が抽出できる．

e1=の， e2=に， e3 ={原告，たち}， e4={染み付く，いる}

属性 e3 と e4 の値は集合になるが，学習の際に以下のように分割して，素性として表す．

e3=原告， e3=たち， e4=染み付く， e4=いる

3　EM アルゴリズムによる教師なし学習

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 (Nigam et al. 2000)．

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．

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

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．

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

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．

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

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．

最終的には収束した $P(f_i|c_j)$ を使って，式 3 から分類が行える．

4　実験

SENSEVAL2 の日本語翻訳タスクで課題として出題された全名詞 20 単語に対して本手法を適用する．

翻訳タスクのコンテストでは，手作業で訓練データを作成し，それを用いて学習するというオーソドックスな戦略を用いたシステムは Ibaraki だけであった．ここではそこで用意された訓練データを借用し，Ibaraki の結果と比較することで本手法を評価する．Ibaraki では，TM の他に毎日新聞'95 年度版から該当単語を含む文を適当な数だけ取りだし，ラベルを付けることで訓練データを増やしている．名詞に対しては各単語に対して約 50 事例を追加している．結果として，各単語に対して平均 70 事例がラベル付き訓練データとして用意された．そのラベル付き訓練データから決定リスト (Yarowsky 1994) を作成し，課題の曖昧性解消問題を解いている．名詞 20 単語に対する Ibaraki の翻訳タスクに対する公式成績を表 1 に示す (新納浩幸 2001)．

Ibaraki で利用した訓練データを借用し，それを本手法のラベル付き訓練データとした．次に，毎日新聞'96 年度版から該当単語を含む文を取りだし，それをラベルなし訓練データとした．

表 2 に，名詞 20 単語の各単語に対するラベル付き訓練データ L の数，ラベルなし訓練データ U の数，ラベル付き訓練データから学習できた決定リスト（ DL と略す）による正解率（ Ibaraki の結果），ラベル付き訓練データのみから学習できた Naive Bayes（ NB と略す）による正解率，NB を EM アルゴリズムにより改善させた分類器（ NB+ EM と略す）の正解率を示す．

表 2 から分るようにラベル付き訓練データ L のみから学習できた DL も NB もほぼ同等の正解率（ 58.9% と 58.2%）である．一方，NB+EM の正解率は 61.8% であり，本手法の効果が確認できる．特に教師なし学習が効果的に行えた kokunai と kiroku の 2 単語について，その学習のグラフを図 1 と図 2 に示す．このグラフの横軸は EM アルゴリズムの繰り返しの回数，縦軸はテスト文に対する正解率を示す．

ラベルなし訓練データを用いることで全体の正解率は向上したが，個々の単語をみると，本手法を利用することで精度が大きく下がる単語が存在する．具体的には表 3 に示す 2 単語である．調査したところ，これは最初に用意しているラベル付き訓練データ中の誤りが原因であっ

表 1: Ibaraki(決定リスト) の正解率

見出し	訓練事例数	決定リストのサイズ	正解率
ippan	87	174	0.467
ippou	63	101	0.567
ima	67	135	0.267
imi	69	181	0.700
kaku_n	58	121	0.800
kiroku	65	159	0.467
kokunai	62	144	0.733
kotoba	79	183	0.800
shimin	64	157	0.733
jigyou	66	186	0.400
jidai	89	249	0.800
sugata	77	206	0.367
chikaku	64	165	0.600
chushin	61	157	0.500
hana	64	139	0.533
hantai	73	176	0.733
baai	73	194	0.733
mae	62	161	0.700
mune	79	179	0.567
mondai	81	204	0.500

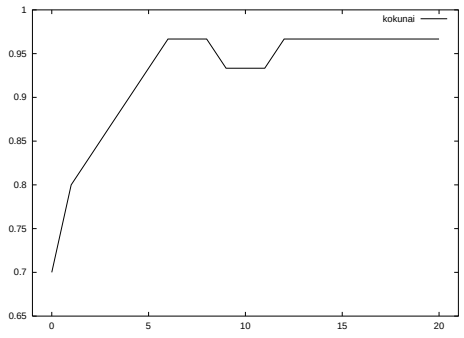

図 1: kokunai の学習

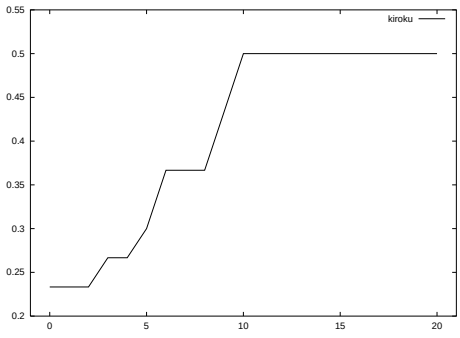

図 2: kiroku の学習

表 2: 実験結果

見出し	L	U	DL	NB	NB+EM
ippan	87	2170	0.467	0.467	0.400
ippou	63	4033	0.567	0.633	0.700
ima	67	5081	0.267	0.200	0.033
imi	69	1761	0.700	0.467	0.467
kaku_n	58	1135	0.800	0.767	0.700
kiroku	65	1726	0.467	0.233	0.500
kokunai	62	2468	0.733	0.700	0.967
kotoba	79	2225	0.800	0.900	0.967
shimin	64	2069	0.733	0.567	0.500
jigyou	66	3500	0.400	0.367	0.467
jidai	89	4397	0.800	0.867	0.833
sugata	77	1971	0.367	0.367	0.333
chikaku	64	1944	0.600	0.600	0.667
chushin	61	3194	0.500	0.600	0.633
hana	64	851	0.533	0.633	0.667
hantai	73	2103	0.733	0.900	0.967
baai	73	3413	0.733	0.833	0.900
mae	62	10931	0.700	0.633	0.667
mune	79	676	0.567	0.633	0.500
mondai	81	11424	0.500	0.500	0.500
平均	70	3354	0.589	0.582	0.618

表 3: 大きく精度が下がる単語

見出し	NB	NB+EM
ima	0.200	0.033
mune	0.633	0.500

た．Ibaraki で用意されたラベル付き訓練データは，一部の単語で必要以上に語義を細かく分けている．上記の 2 単語はその例であり，特に ima では UNASSIGNABLE のラベル（適切な例文がないことを意味するラベル）を付けている事例が 67 事例中 20 事例も存在する．実際は UNASSIGNABLE のラベルを与えた事例には default の語義（この場合,『重要性』の意味で使われている例文番号）を与えるべきであった．mune でも慣用的な表現が多く細かく語義を分け

すぎている．正解を見れば,『 体の一部としての胸』と『 心の中』の 2 つに分類できればよいだけである．これらを考慮して，この 2 単語に関しては，ラベル付き訓練データを修正した．具体的には，ima に対しては UNASSIGNABLE を default の語義に変更し，mune では語義を 2 値に変更した．修正して得られた訓練データに対して，本手法をもう一度試した．またこれらの 2 単語に対しては，修正したラベル付き訓練データを利用した Ibaraki による決定リスト DL の正解率も調べた．修正して得られた結果を表 4 に示す．結果的にラベル付き訓練データ L のみから学習できた NB の正解率 62.3% を本手法により 68.2% まで高めることができた．

表 4: 一部修正後の実験結果

見出し	L	U	DL	NB	NB+EM
ippan	87	2170	0.467	0.467	0.400
ippou	63	4033	0.567	0.633	0.700
ima	67	5081	0.700	0.833	1.000
imi	69	1761	0.700	0.467	0.467
kaku_n	58	1135	0.800	0.767	0.700
kiroku	65	1726	0.467	0.233	0.500
kokunai	62	2468	0.733	0.700	0.967
kotoba	79	2225	0.800	0.900	0.967
shimin	64	2069	0.733	0.567	0.500
jigyou	66	3500	0.400	0.367	0.467
jidai	89	4397	0.800	0.867	0.833
sugata	77	1971	0.367	0.367	0.333
chikaku	64	1944	0.600	0.600	0.667
chushin	61	3194	0.500	0.600	0.633
hana	64	851	0.533	0.633	0.667
hantai	73	2103	0.733	0.900	0.967
baai	73	3413	0.733	0.833	0.900
mae	62	10931	0.700	0.633	0.667
mune	79	676	0.800	0.633	0.800
mondai	81	11424	0.500	0.500	0.500
平均	70	3354	0.632	0.623	0.682

5　考察

ここでは本手法を名詞のみに適用した．同じ処理によって，動詞に対しても適用することができるが，ここではその実験を行わなかった．教師なし学習を利用するには，本質的に，識別のための冗長性のある情報が必要である．名詞の場合，その名詞を修飾する語句（左文脈）は，その名詞の語義を特定できる可能性が高いし，その名詞を格にもつ動詞（右文脈）もその名詞の語義を特定できる可能性が高いので，一方の文脈から名詞の語義が識別できれば，もう一方の文脈は識別のための冗長性のある情報となる．このため，設定した属性は教師なし学習に適していると考えられる．一方，動詞の語義を識別するのは，格要素になる名詞，つまり左文脈が重要であり，右文脈は語義の識別の助けになることは少ない．連体修飾の用法にしても，左右が逆になるだけである．つまり，どちらかの文脈を利用して語義を識別した場合に，もう一方の文脈は識別に寄与する情報にならない．このため，動詞に対しては，本手法を利用する効果は低いと考えた (Shinnou 2002)．ただし「効果がない」ということでもないことを注意しておく．本手法はラベル付き訓練データのみから得られた分類器の精度を必ずしも向上するとは言えず，逆に精度を落す危険性もある．そのために，本手法を利用する効果があまり期待できない場合には，危険性を犯してまで本手法を試みる必要はないと判断した．動詞に対して実際にどの程度の精度向上，あるいは精度低下があるのか，あるいは動詞に対してはどのような属性を設定するのが良いのかを調べることは今後の課題である．

先ほども述べたが，本手法により必ずしも精度が向上するとは限らない．実際に，実験では表 5 の 5 単語に関して，わずかではあるが精度が低下している．

表 5: 精度が下がる単語

見出し	NB	NB+EM
ippan	0.467	0.400
kaku_n	0.767	0.700
shimin	0.567	0.500
jidai	0.867	0.833
sugata	0.367	0.333

精度低下の原因を一般的に論じるのは難しい．この実験の場合，偶然的な要素が強かった．NB による分類器では正解したが，NB+EM による分類器では誤るようなテスト文を調査すると，NB による分類器で正解したのは，たまたま default の規則が適用できて，正解になったというように，偶然的な要素が強い．EM による学習が進むと，default から少しずれてくるために，誤ってしまう．精度低下の原因に関しては，ラベル付き訓練データ，ラベルなし訓練データおよびテストデータの関係を詳しく調査する必要がある．

本手法による更なる精度向上をはかるための最も有効な手段は，最初のラベル付き訓練デー

タを見直すことである．今回利用したラベル付き訓練データは，コンテストの正解が提示される以前に作成されたものであり，出題者が想定した語義と微妙に違う部分がある．概して，出題者が想定した語義は荒く，Ibaraki で用意された語義は細かい．語義が細かいと，結果として訓練データが小さいものになり，学習から得られる規則の精度が悪く，無用な部分で識別が誤る．ima や mune でもラベル付きの訓練データを見直すことで精度が改善された．

またラベルなし訓練データの量の問題が指摘されるかも知れない．ラベルなし訓練データは多ければ多いほど精度が向上すると言われている．今回，精度低下のあった ippan，shimin, jidai の 3 単語に関して，ラベルなし訓練データの量を約 4 倍に増やして実験を行った．このデータは別年度の毎日新聞記事から取り出した[2]．結果を表 6 に示す．

表 6: ラベルなし訓練データを増やした実験

見出し	L	U	new U	NB	NB+EM (using L+U)	NB+EM (using L+ new U)
ippan	87	2170	8048	0.467	0.400	0.400
shimin	64	2069	7912	0.567	0.500	0.533
jidai	89	4397	15858	0.867	0.833	0.833

精度は悪くなることはなかったが，ほとんど変化は生じなかった．おそらく今回実験で利用した程度のラベルなし訓練データの量でも，このタスクでは十分であったと考えられる．

またもう一つの代表的な教師なし学習の手法である Co-training(Blum and Mitchell 1998) との比較について述べておく．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 つの属性が設定できた場合，Co-training は EM アルゴリズムを利用した手法よりも優れていることが報告されている (Nigam and Ghani 2000)．しかし Co-training には独立な 2 つの属性という条件の他に，属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 (新納浩幸 2001)．一方，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．

また多義語の曖昧性解消問題に教師なし学習を利用した Yarowsky の研究 (Yarowsky 1995) 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 (Blum and Mitchell 1998)．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翻訳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，本手法はその対象

[2] ただしテスト文が 94 年度版から取られることが分っているので，94 年度版は利用していない．

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2つある．1つは名詞以外の単語への適用である．教師なし学習が機能するような属性をどのように設定するかが課題である．2つ目は教師なし学習による精度低下の原因の調査，およびその回避策の検討である．これによってより頑健な教師なし学習が可能となる．

6　おわりに

本論文では，Nigam らによって提案された EM アルゴリズムを利用した教師なし学習の手法を，SENSEVAL2 の日本語翻訳タスクで出題された名詞に適用した．識別のための属性としては，対象単語の前後数単語の原型や表記という簡易なものを利用した．ラベル付き訓練データだけから学習できた決定リストの正解率は 58.9% (コンテストでの Ibaraki の成績) であり，Naive Bayes による分類器の正解率は 58.2% であった．そして本手法を用いて Naive Bayes による分類器の正解率を 61.8% まで改善できた．また一部，訓練データの不具合を修正することで，Naive Bayes による分類器の正解率 62.3%（ 決定リストでの正解率は 63.2%）を，本手法により 68.2% まで高めることができた．問題点としては名詞のみの適用である点と，精度が低下するケースも存在する点である．これら問題の解決が今後の課題であり，より頑健性の高い教師なし学習手法の構築を目指す．

参考文献

Blum, A. and Mitchell, T. (1998). “Combining Labeled and Unlabeled Data with Co-Training.” In *11th Annual Conference on Computational Learning Theory (COLT-98)*, pp. 92–100.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Companies.

Nigam, K. and Ghani, R. (2000). “Analyzing the effectiveness and applicability of co-training.” In *9th International Conference on Information and Knowledge Management*, pp. 86–93.

Nigam, K.,McCallum, A.,Thrun, S.,and Mitchell, T. (2000). “Text Classification from Labeled and Unlabeled Documents using EM.” *Machine Learning*, **39** (2/3), 103–134.

Rooth, M.,Riezler, S.,Prescher, D.,Carroll, G.,and Beil, F. (1999). “Inducing a Semantically Annotated Lexicon via EM-Based Clustering.” In *37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 104–111.

Shinnou, H. (2002). “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.

Torisawa, K. (2001). “An Unsupervised Method for Canonicalization of Japanese Postposi-

tions." In *6th Natural Language Processing Pacific Rim Symposium (NLPRS-2001)*, pp. 211–218.

Yarowsky, D. (1994). "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French." In *32th Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp. 88–95.

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

黒橋禎夫, 白井清昭 (2001). "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.

新納浩幸 (2001). "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.

略歴

新納 浩幸: 1985 年東京工業大学理学部情報科学科卒業. 1987 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, 1993 年茨城大学工学部システム工学科助手. 1997 年同学科講師，2001 年同学科助教授. 情報処理学会，人工知能学会，言語処理学会，ACL 各会員. 博士 (工学).

(2002 年 4 月 9 日 受付)
(2002 年 7 月 10 日 再受付)
(2002 年 8 月 20 日 採録)