クラスタリングを利用した能動学習による語義曖昧性解消の領域適応

小野寺 喜行 (茨城大学 工学部 情報工学科) [1]
新納 浩幸 (茨城大学 工学部 情報工学科) [2]

Domain Adaptation for Word Sense Disambiguation by Active Learning Using a Clustering Method

Yoshiyuki Onodera (Ibaraki University, Department of Computer and Information Sciences)
Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応の問題に対して，初期段階でデータ選択にクラスタリングを利用した能動学習手法を提案する．

自然言語処理のタスクにおいて帰納学習手法を用いる際，訓練データとテストデータは同じ領域のコーパスから得ていることが通常である．ただし実際には異なる領域である場合も存在する．そこである領域（ソース領域）の訓練データから学習された分類器を，別の領域（ターゲット領域）のテストデータに合うようにチューニングすることを領域適応という[3]．

領域適応の問題をターゲット領域のラベル付きデータの不足からくる問題として見なせば，能動学習 (Settles (2010)) や半教師あり学習 (Chapelle et al. (2006)) を利用することは有効である．ここでは WSD の領域適応の問題に対して，能動学習を利用する．

一般に能動学習はラベルなしデータの集合から学習効果の高いデータを選択し，そのデータにラベルを付けて訓練データに追加することで，徐々に分類器の精度を高めてゆく．能動学習のポイントはどのようにして学習効果の高いデータを選択するかである．通常はその時点で保持しているラベル付きデータを利用してその選択が行われるが，領域適応では能動学習の初期の段階ではソース領域のラベル付きデータで占められるため，上記の選択が有効に行える保証がない．ここでは能動学習の初期の段階ではターゲット領域のデータをクラスタリングし，そこから代表点を選ぶことを提案する．

実験では BCCWJ コーパス (Maekawa (2007)) の 2 つ領域 PB (書籍) と OC (Yahoo! 知恵袋) から共に頻度が 50 以上の多義語 17 単語を対象にして，能動学習を利用した WSD の領域適応の実験を行い，提案手法の有効性を示す．

2　クラスタリングを利用した能動学習

2.1　能動学習

精度の高い分類器を構築するためにはラベル付きデータを増やせばよい．ただしラベルを付けるコストは高いため，少量のラベル付けで精度の高い分類器を構築する方法が望まれている．能動学習もそのような背景から考案された学習手法である．

能動学習では学習効果の高いデータをシステムが選択し，ユーザがそのデータにラベルを付ける．これを繰り返すことで分類器の精度を徐々に向上させてゆく．ランダムに取り出されたデータにラベルを付けるよりも，少量のラベル付けで精度の高い分類器が構築できる．

能動学習には様々な手法が存在するが，ここでは簡易でありながら効果が高い Schohn の手法 (Schohn and Cohn (2000)) を利用する．Schohn の手法を図 1 に示す．

[2] 10t4019s@hcs.ibaraki.ac.jp
[2] shinnou@mx.ibaraki.ac.jp
[3] 領域適応は機械学習の分野では転移学習 (神嶌敏弘 (2010)) の一種と見なされている．

(1) ラベル付きデータから分類器を作成する．

(2) 作成した分類器によりラベルなしデータを識別する．このとき識別の信頼度も求める．

(3) 識別の信頼度が最も低いデータにラベルを付け，ラベル付きデータに追加する．

(4) (1) に戻る

図 1: 能動学習の手順

ここでは分類器として SVM を用いる．また SVM ツールには libsvm[4] を用いた．そこでは `-b` オプションにより識別の信頼度が求められる．

2.2 初期データ選択

領域適応の問題解決のために，ターゲット領域のラベル付きデータを用意する戦略をとれば，能動学習が利用できることは明かである．ただしこの場合，能動学習の初期の段階ではソース領域のラベル付きデータの占める割合が高く，そこから新たにラベルを付けるデータを選択する手法 (図 1) が有効に働かない可能性がある．

そこで本論文では能動学習の初期の段階のみ，クラスタリングを利用してデータを選択を行う．具体的には，ターゲット領域のデータを K 個のクラスタにクラスタリングする．次に各クラスタ C_i の中から代表点 c_i を選ぶ．得られた $\{c_1, c_2, \cdots, c_K\}$ を初期データとする．代表点の選び方は，そのクラスタ内の他のデータとの類似度の和が最大のデータをとる方法である．

$$c_i = \arg\max_{c \in C_i} \sum_{x \in C_i} sim(c, x)$$

このようにして選択されたデータはターゲット領域のみに依存しているので，ソース領域の影響を受けない．

また本論文では $K = 5$ に設定している．クラスタリングのツールには CLUTO[5] を用いた．

3 実験

実験には BCCWJ コーパスの OC(Yahoo!知恵袋) と PB(書籍) を使った．その中から表 1 に示す 17 単語について OC から PB，PB から OC の 2 通り領域適応の実験を行う．

[4] `http://www.csie.ntu.edu.tw/ cjlin/libsvm/`
[5] `http://glaros.dtc.umn.edu/gkhome/views/cluto`

表 1: 実験対象単語

単語	辞書上の語義数	PB での頻度	PB での語義数	OC での頻度	OC での語義数
言う	3	1114	2	666	2
入れる	3	56	3	73	2
書く	2	62	2	99	2
聞く	3	123	2	124	2
来る	2	104	2	189	2
子供	2	93	2	77	2
時間	4	74	2	53	2
自分	2	308	2	128	2
出る	3	152	3	131	3
取る	8	81	7	61	7
場合	2	137	2	126	2
入る	3	118	4	68	4
前	3	160	2	105	3
見る	6	273	6	262	5
待つ	4	153	3	62	4
やる	5	156	4	117	3
ゆく	2	133	2	219	2
平均	3.35	193.9	2.94	150.6	2.88

実験では図 1 の能動学習手法 (従来手法) と提案手法による語義識別の平均正解率で比較する．提案手法では，クラスタリングから初期の 5 点を選び訓練データに追加する．その後，通常の能動学習で 5 点選び訓練データに追加し，それを用いて SVM の学習を行う．また従来手法では，クラスタリングを用いずに通常の能動学習で 10 点を訓練データに追加する．

従来手法と提案手法の学習の平均正解率を OC から PB の領域適応について表 2 と図 2 に，PB から OC の領域適応について表 3 と図 3 に示す．結果，PB から OC の領域適応では提案手法の効果はあったと言える．しかし OC から PB の領域適応では，8 点目までは提案手法の方が平均正解率が高かったが，最終的な 10 点目での平均正解率は従来手法の方が高かった．

表 2: 平均正解率 (%)(OC → PB)

追加データ数	従来手法	提案手法
1	73.20	
2	73.14	
3	73.96	
4	75.21	
5	76.11	76.32
6	77.05	77.31
7	76.72	77.80
8	78.00	78.58
9	78.88	78.82
10	79.47	79.04

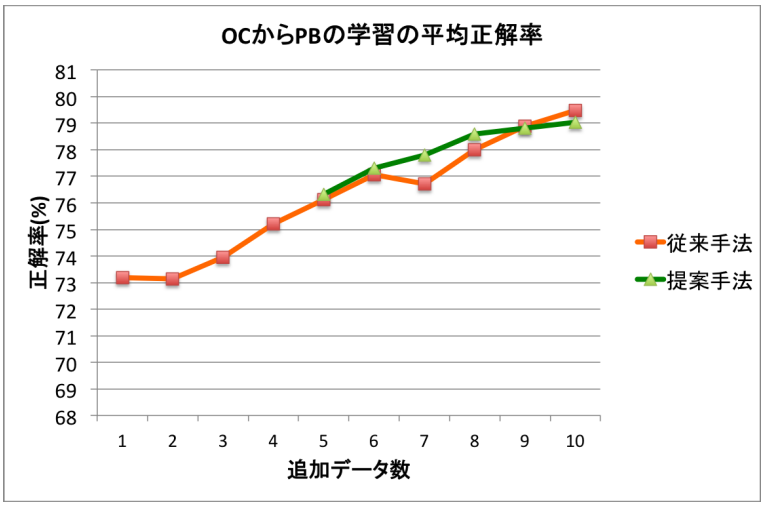

図 2: 平均正解率の変化 (OC → PB)

表 3: 平均正解率 (%)(PB → OC)

追加データ数	従来手法	提案手法
1	72.12	
2	73.26	
3	74.22	
4	76.02	
5	76.46	75.78
6	76.92	77.27
7	77.18	77.49
8	77.61	79.03
9	78.02	79.76
10	78.15	79.54

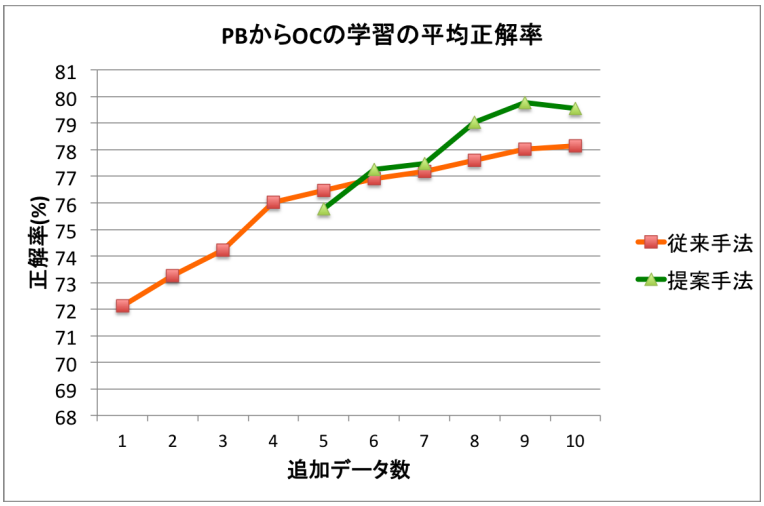

図 3: 平均正解率の変化 (PB → OC)

4　考察

4.1　領域間距離

領域適応ではソース領域とターゲット領域との距離が本質的な役割を担う．距離が近ければ，ソース領域の知識の多くはターゲット領域においても有効である．逆に距離が離れていれば，ソース領域の知識はターゲット領域においてあまり役立たない．本論文における提案手法のアイデアは，ソース領域とターゲット領域との距離がある程度離れていることを想定している．距離が近い場合，従来手法がそのまま使えるはずである．そのため提案手法の効果はソース領域とターゲット領域との距離に相関があると考えられる．ここでは，この点を確認する．

まずソース領域 S とターゲット領域 T との距離の測り方が問題である．これは様々な手法が提案されているが，ここでは以下の形で領域間の類似度 $sim(S,T)$ を測定した．注意としてここでの類似度は方向性があり必ずしも $sim(S,T) = sim(T,S)$ とはなっていないことに注意する．

領域 X 内のデータ x は素性リスト $x = \{f_x^1, f_x^2, \cdots, f_x^{m(x)}\}$ で表せる．集合 F_x はこの素性リストの素性を集めたものである．

$$F_x = \bigcup_{x \in S} \{f_x^1, f_x^2, \cdots, f_x^{m(x)}\}$$

素性 $y \in F_x$ の頻度を $g(y)$ 表す．ソース領域 S とターゲット領域 T と類似度 $sim(S,T)$ を以下で定義する．

$$sim(S,T) = \frac{\sum_{y \in F_s \cap F_T} g(y)}{\sum_{y \in F_s} g(y)}$$

上記の類似度と前述した実験結果との対応関係を表 4 と表 5 に示す．表 4 は OC から PB の領域適応，表 5 は PB から OC の領域適応である．

それぞれ類似度と手法の効果（正解率の差）について相関係数を求めると，OC から PB の学習では 0.17，PB から OC の学習では-0.22 であった．この値からは領域間距離と提案手法の効果との間に関連性は認められない．ただしここで行った領域間の距離の測定は簡易なものであり，適切に測定できていない可能性が高い．提案手法が効果が出るのは，領域間距離が離れている場合であると考えられるので，今後は適切な領域間距離の測定法を考察したい．

表 4: 領域間類似度と手法 (%) の効果 (OC → PB)

単語	sim(OC,PB)	正解率の差	従来手法	提案手法
言う	0.53	2.42	79.26	81.68
入れる	0.13	-1.67	80.36	78.69
書く	0.20	-3.76	90.32	86.57
聞く	0.32	-0.77	79.67	78.91
来る	0.26	-1.79	99.04	97.25
子供	0.34	-4.95	56.99	52.04
時間	0.27	0.60	90.54	91.14
自分	0.36	-0.28	97.73	97.44
出る	0.26	-0.65	60.53	59.87
取る	0.25	6.35	30.86	37.21
場合	0.23	0.49	86.13	86.62
入る	0.32	-1.67	61.02	59.35
前	0.23	0.97	88.13	89.09
見る	0.33	0.28	84.62	84.89
持つ	0.34	-0.62	79.74	79.11
やる	0.31	-0.42	93.59	93.17
ゆく	0.29	-1.90	92.48	90.58

表 5: 領域間類似度と手法 (%) の効果 (PB → OC)

単語	sim(PB,OC)	正解率の差	従来手法	提案手法
言う	0.45	-0.76	82.58	81.82
入れる	0.16	2.69	78.08	80.77
書く	0.30	1.26	73.74	75.00
聞く	0.29	-1.17	70.16	68.99
来る	0.32	1.02	80.42	81.44
子供	0.27	11.86	45.45	57.32
時間	0.18	1.30	84.91	86.21
自分	0.28	0.44	88.28	88.72
出る	0.25	3.36	68.70	72.06
取る	0.17	5.61	45.90	51.52
場合	0.23	-1.44	97.62	96.18
入る	0.17	1.91	72.06	73.97
前	0.21	-1.47	92.38	90.91
見る	0.34	0.63	86.26	86.89
持つ	0.24	-1.01	93.55	92.54
やる	0.28	-0.61	94.87	94.26
ゆく	0.34	0.14	73.52	73.66

4.2　スペクトラルクラスタリングの利用

提案手法ではクラスタリングを利用して，初期のデータ選択を行う．このためクラスタリングの精度が正解率に影響を与えている可能性がある．ここでは精度の高いクラスタリング手法として知られるスペクトラルクラスタリング (Von Luxburg (2007)) を利用して提案手法を試す.

従来手法，提案手法およびスペクトラルクラスタリングを用いた手法の 3 手法による平均正解率を，OC から PB の学習について表 6 と図 4 に，PB から OC の学習について表 7 と図 5 に示す.

実験の結果，OC から PB の領域適応では提案手法より平均正解率が向上し，従来手法の結果も上回っている．また，PB から OC の領域適応において平均正解率は提案手法より下がったものの，従来手法の結果より高い正解率となった．つまり提案手法はクラスタリングの精度とも関連しており，

精度の高いクラスタリング手法を利用することで，効果が現れると言える．

表 6: スペクトラルクラスタリングの利用 (OC → PB)

追加データ数	従来手法	提案手法	スペクトラルクラスタリングの利用
1	73.20		
2	73.14		
3	73.96		
4	75.20		
5	76.11	76.32	75.74
6	77.05	77.31	76.34
7	76.72	77.80	77.80
8	78.00	78.58	78.20
9	78.88	78.82	78.87
10	79.47	79.04	79.49

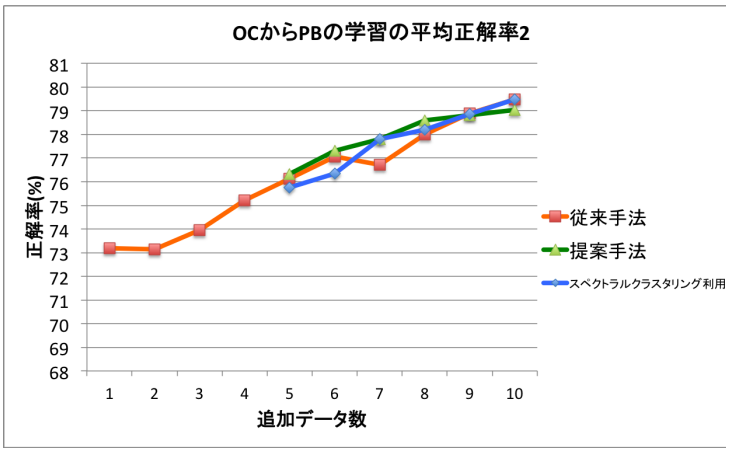

図 4: スペクトラルクラスタリングの利用による平均正解率の変化 (OC → PB)

表 7: スペクトラルクラスタリングの利用 (PB → OC)

追加データ数	従来手法	提案手法	スペクトラルクラスタリングの利用
1	72.12		
2	73.26		
3	74.22		
4	76.02		
5	76.46	75.78	74.48
6	76.92	77.27	76.09
7	77.18	77.49	76.84
8	77.61	79.03	77.25
9	78.02	79.76	78.38
10	78.15	79.54	79.14

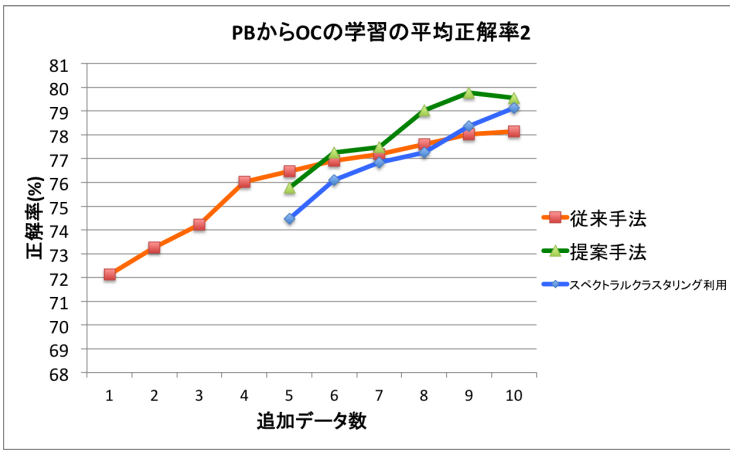

図 5: スペクトラルクラスタリングの利用による平均正解率の変化 (PB → OC)

5　おわりに

本論文では WSD の領域適応の問題に対して能動学習を試みた．領域適応に能動学習を利用する場合，初期の段階ではソース領域のデータが占める割合が高いため，データ選択が適切に行えない可能性がある．そこでここでは能動学習の初期の段階ではターゲット領域のデータをクラスタリングすることでデータ選択を行うこと提案した．

BCCWJ コーパスの OC(Yahoo!知恵袋) と PB(書籍) の 2 つの領域を用いた実験では，PB から OC の領域適応では提案手法の効果はあったが，OC から PB の領域適応では 8 点目までは提案手法の効果はあったが，最終的な 10 点目では従来手法の方が平均正解率は高かった．ただし精度の高いクラスタリング手法を使うことで，提案手法の効果を確認できた．

提案手法を効果的に利用するには，領域間距離を適切に測ることが重要だと考えている．この測定法を考案することが今後の課題とする．

文献

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien et al. (2006) *Semi-supervised learning*, Vol. 2: MIT press Cambridge.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Greg Schohn and David Cohn (2000) “Less is more: Active learning with support vector machines,” in *ICML*, pp. 839–846.

Burr Settles (2010) “Active learning literature survey,” *University of Wisconsin, Madison*.

Ulrike Von Luxburg (2007) “A tutorial on spectral clustering,” *Statistics and computing*, Vol. 17, No. 4, pp. 395–416.

神嶌敏弘 (2010) 「転移学習」，人工知能学会誌，第 25 巻，第 4 号，pp.572–580.