領域間距離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 the Distance between Domains

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，データ選択に領域間距離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 (Sogaard (2013))[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 WSD の各対象単語毎に識別の信頼度を基準にしてその選択が行われる．つまり WSD を対象とした能動学習では，各対象単語毎に能動学習を適用する形になっている．ただしこの形では WSD の総合的な評価であるマクロ平均やマイクロ平均の評価値を効率的に向上させることができる保証はない．そこで本論文では WSD の全対象単語の全テストデータを対象にして，ラベル付けするデータを選択することを試みる．また，選択する際の信頼度は，領域間の関係に依存していると考えられる．つまり領域間の距離が大きければ，算出した信頼度よりも更に信頼度は低いし，領域間の距離が小さければ，算出した信頼度は妥当と考えられる．このように領域間距離を考慮した信頼度を用いることで更なる精度向上を図る．

実験では現代日本語書き言葉均衡コーパス（BCCWJ コーパス (Maekawa (2007))）における 3 つの領域 OC (Yahoo! 知恵袋)，PB (書籍) 及び PN (新聞) を利用する．SemEval-2 の日本語 WSD タスク (Okumura et al. (2010)) ではこれらのコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．すべての領域である程度の頻度が存在する多義語 16 単語を対象にして，WSD の領域適応の実験を行う．領域適応としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りが存在する．結果 $16 \times 6 = 96$ 通りの WSD の領域適応の問題に対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　領域間距離を利用した能動学習

2.1　全対象単語に対する能動学習

通常，能動学習では図 1 に示す Schohn の手法の様に，WSD の各対象単語毎に識別の信頼度を基準にしてその選択が行われる．つまり WSD を対象とした能動学習では，各対象単語毎に能動学習を適用する形になっている．ただしこの形では WSD の総合的な評価であるマクロ平均やマイクロ平均の評価値を効率的に向上させることができる保証はない．

[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．

そこで WSD の全対象単語の全テストデータを対象にして，その中から信頼度が最も低いデータをラベル付けするデータとして選択することを試みる．

(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 (Schohn の手法)

2.2 領域間距離を用いた能動学習

ラベル付けするデータを選択する際の信頼度は，領域間の関係に依存していると考えられる．つまり領域間の距離が大きければ，算出した信頼度よりも更に信頼度は低いし，領域間の距離が小さければ，算出した信頼度は妥当と考えられる．そこで，領域間距離を考慮した信頼度を用いることで更なる精度向上を図る．

ここでは，確率密度比から求めた重みを領域間類似度と考え領域間距離とし，能動学習においてラベル付けするデータの選択に識別の信頼度と領域間距離から導き出した評価値を用いることにする．WSD の全対象単語の全テストデータを対象にして次に示すように評価値を求め，その値が最も低いデータをラベル付けするデータとして選択する．

まず，分類器によって識別した n 個のラベルなしデータ $x_j\,(j = 1, 2, \ldots, n)$ に対する識別の信頼度の集合 $h_j\,(j = 1, 2, \ldots, n)$ の平均値 m と標本分散値 u^2 を以下の式で求める．

$$m = \frac{\sum_{j=1}^{n} h_j}{n},\ u^2 = \frac{\sum_{j=1}^{n} (m - h_j)^2}{(n-1)}$$

次に，領域間類似度の集合 $s_j\,(j = 1, 2, \ldots, n)$ から同様にして平均値 m' と標本分散値 u'^2 を求める．

$$m' = \frac{\sum_{j=1}^{n} s_j}{n},\ u'^2 = \frac{\sum_{j=1}^{n} (m - s_j)^2}{(n-1)}$$

そして各データ x_j について以下の値を求める．

$$r_j = \frac{(h_j - m)}{u},\ w_j = \frac{u \times (s_j - m')}{u'} + m$$

この r_j と w_j の値を合計したものがデータ x_j の評価値となる．

3 実験

実験では現代日本語書き言葉均衡コーパス（BCCWJ コーパス (Maekawa (2007))）における 3 つの領域 OC (Yahoo! 知恵袋)，PB (書籍) 及び PN (新聞) を利用する．SemEval-2 の日本語 WSD タスク (Okumura et al. (2010)) ではこれらのコーパスの一部に語義タグを付けたデータを公開しており，そのデータを利用する．すべての領域である程度の頻度が存在する多義語 16 単語 (表 1) を対象にして，WSD の領域適応の実験を行う．

表 1: 実験対象単語

単語	辞書上の語義数	OC		PB		PN	
		頻度	語義数	頻度	語義数	頻度	語義数
言う	3	666	2	1114	2	363	2
入れる	3	73	2	56	3	32	2
書く	2	99	2	62	2	27	2
聞く	3	124	2	123	2	52	2
子供	2	77	2	93	2	29	2
時間	4	53	2	74	2	59	2
自分	2	128	2	308	2	71	2
出る	3	131	3	152	3	89	3
取る	8	61	7	81	7	43	7
場合	2	126	2	137	2	73	2
入る	3	68	4	118	4	65	3
前	3	105	3	160	2	106	4
見る	6	262	5	273	6	87	3
持つ	4	62	4	153	3	59	3
やる	5	117	3	156	4	27	2
ゆく	2	219	2	133	2	27	2
平均	3.35	193.9	2.94	150.6	2.88	75.56	2.69

領域適応としては OC → PB，PB → PN，PN → OC，OC → PN，PN → PB，PB → OC の計 6 通りが存在する．結果 $16 \times 6 = 96$ 通りの WSD の領域適応の問題に対して通常の能動学習と 2 つの提案手法,「全対象単語に対する能動学習」と「領域間距離を用いた能動学習」による識別の平均正解率で比較する．なお，ラベル付けするデータの数は通常の手法は各単語 5 点ずつ計 80 点，提案手法は全体で 80 点とした．加えて，ランダムにラベル付けするデータを選択する場合も試す．こちらは各単語 5 点ずつ選択する場合と全体から 80 点選択する場合をそれぞれ 5 回行いその平均を求める．分類器は SVM を用い，SVM ツールには libsvm[4] を使用した．libsvm では `-b` オプションにより識別の信頼度が求められる．

実験結果を表 2(マクロ平均) と表 3(マイクロ平均) に示す．結果として,「全対象単語に対する能動学習」は通常の手法よりも主にマクロ平均において良い正解率となり識別精度の向上に効果が有ると言える．また,「領域間距離を用いた能動学習」は低い正解率となり精度の向上に効果は見られなかった．

表 2: 実験結果 (マクロ平均 (%))

	通常の能動学習	ランダムに選択 (各単語 5 点ずつ)	ランダムに選択 (全体から 80 点)	全対象単語に対する能動学習	領域間距離を用いた能動学習
OC → PB	74.80	73.55	72.40	75.49	71.76
PB → PN	79.74	72.29	72.40	81.14	76.94
PN → OC	73.39	71.23	69.89	74.38	68.43
OC → PN	75.23	71.59	69.33	73.86	70.47
PN → PB	78.26	76.40	72.09	79.55	77.33
PB → OC	76.22	74.30	72.09	77.03	72.04

[4] `http://www.csie.ntu.edu.tw/ cjlin/libsvm/`

表 3: 実験結果 (マイクロ平均 (%))

	通常の能動学習	ランダムに選択 (各単語 5 点ずつ)	ランダムに選択 (全体から 80 点)	全対象単語に対する能動学習	領域間距離を用いた能動学習
OC → PB	78.89	77.85	78.33	78.45	77.36
PB → PN	83.37	82.47	78.33	84.62	83.13
PN → OC	74.82	73.94	74.85	74.99	72.88
OC → PN	75.60	74.80	77.49	67.00	65.26
PN → PB	82.09	81.02	74.69	81.65	83.03
PB → OC	78.70	76.35	74.69	78.20	74.57

4 考察

4.1 領域間距離の利用について

もし領域間距離を正しく測れている場合，負の転移の有無を判定できる．負の転移が生じているかどうかを調べるために以下の実験を行った．

単語 w_i についてソース領域 S からターゲット領域 T への領域適応の実験を行う．まずターゲット領域 T のラベル付きデータをランダムに 15 個取り出し，残りを評価データとする．つまり利用できる訓練データはソース領域 S のラベル付きデータとターゲット領域 T からランダムに取り出した 15 個のラベル付きデータとなる．この訓練データを用いて手法 A により分類器を作成し，先の評価データの語義識別の正解率 $P_{i,k}$ を測る．この実験を 5 回行い $P_{i,1}, P_{i,2}, \cdots, P_{i,5}$ を得る．それらの平均 P_i を「単語 w_i の S から T への領域適応における手法 A の平均正解率」とする．上記の単語 w_i を 16 種類の各対象単語 $w_1, w_2, \cdots, w_{16}$ に変えることで，16 個の平均正解率 $P_1, P_2, \cdots, P_{16}$ が得られる．それらの平均 P を「S から T への領域適応における手法 A の平均正解率」とする．

上記の手法 A としては，以下の 3 種類を試す．(1) ソース領域のラベル付きデータのみを用いる手法 (ターゲット領域の 15 個のラベル付きデータの重みを 0 とする手法)（S-Only），(2) ターゲット領域からランダムに取り出した 15 個のラベル付きデータのみを用いる手法 (ソース領域のラベル付きデータの重みを 0 とする手法)（T-Only），(3) ソース領域のラベル付きデータとターゲット領域の 15 個のラベル付きデータを用いる手法（S+T）．

S から T への領域適応における各手法の平均正解率を表 4 に示す．

表 4: 各手法の平均正解率

領域適応	S-Only	T-only	S+T
OC → PB	0.7137	0.7559	0.7511
PB → PN	0.7678	0.7206	0.7801
PN → OC	0.6926	0.7716	0.7630
OC → PN	0.6829	0.7300	0.7324
PN → PB	0.7543	0.7561	0.7863
PB → OC	0.6988	0.7766	0.7533
平均	0.7184	0.7518	0.7611

負の転移が生じているかどうかの判定には，上記の実験でより得られた `T-Only`，`S-Only` 及び `S+T` の正解率を利用する．もしも正解率で以下の関係が成立しているなら，負の転移が生じていないと考えられる．

T-Only, S-Only < S+T

結果を表 5 に示す．チェックがつけられた箇所が負の転移が生じていない領域適応の問題である．96 種類の領域適応の問題の中で 44 種類において負の転移が生じていない．

表 5: 負の転移が生じていない領域適応

単語	OC → PB	PB → PN	PN → OC	OC → PN	PN → PB	PB → OC
言う		✓	✓	✓	✓	
入れる		✓	✓	✓	✓	✓
書く	✓			✓	✓	
聞く	✓					
子供			✓		✓	
時間	✓		✓		✓	
自分	✓	✓				
出る				✓		✓
取る			✓		✓	✓
場合		✓	✓		✓	✓
入る		✓	✓		✓	✓
前		✓				
見る	✓					
持つ	✓	✓				✓
やる		✓		✓		✓
ゆく		✓		✓	✓	

この表をもとに，負の転移が生じていない a 個の単語については従来手法と同様の手順で各単語 5 点ずつラベル付きデータを追加し，負の転移が生じている 16-a 個の単語はその中で信頼度最低のものをラベル付きデータに追加を計 (16-A)*5 回行う手法を試す．

実験結果を表 6(マクロ平均) と表 7(マイクロ平均) に示す．通常の能動学習よりも良い結果を得ることが出来た．よって，領域間距離の情報を用いてラベル付けするデータを選択する手順を制御することは効果が有ると言える．

提案手法の平均正解率が低い原因は領域間距離を利用すること自体ではなく，「確率密度比から領域間距離を測る」という方法が正しく領域間距離を測ることが出来なかった事や，領域間距離をラベル付けするデータ選択に上手く取り入れられなかった事が考えられる．したがって，今後は領域間距離を正しく測る方法と領域間距離をラベル付けするデータ選択に取り入れる方法を考えていきたい．

表 6: 実験結果 (マクロ平均 (%))

	通常の能動学習	ランダム (各単語 5 点)	ランダム (全体から)	全対象単語に対する	領域間距離を利用	負の転移を考慮
OC → PB	74.80	73.55	72.40	75.49	71.76	75.92
PB → PN	79.74	72.29	72.40	81.14	76.94	80.71
PN → OC	73.39	71.23	69.89	74.38	68.43	74.17
OC → PN	75.23	71.59	69.33	73.86	70.47	75.01
PN → PB	78.26	76.40	72.09	79.55	77.33	78.08
PB → OC	76.22	74.30	72.09	77.03	72.04	75.66

表 7: 実験結果 (マイクロ平均 (%))

	通常の能動学習	ランダム (各単語 5 点)	ランダム (全体から)	全対象単語に対する	領域間距離を利用	負の転移を考慮
OC → PB	78.89	77.85	78.33	78.45	77.36	78.86
PB → PN	83.37	82.47	78.33	84.62	83.13	83.95
PN → OC	74.82	73.94	74.85	74.99	72.88	75.66
OC → PN	75.60	74.80	77.49	67.00	65.26	76.18
PN → PB	82.09	81.02	74.69	81.65	83.03	81.99
PB → OC	78.70	76.35	74.69	78.20	74.57	78.95

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．WSD を対象とした能動学習では，各対象単語毎に能動学習を適用する形になっているが，この形では WSD の総合的な評価であるマクロ平均やマイクロ平均の評価値を効率的に向上させることができる保証はない．

そこで，WSD の全対象単語の全テストデータを対象にしてラベル付けするデータを選択する手法と，領域間距離を考慮した信頼度を用いる手法を試した．BCCWJ コーパスの OC(Yahoo!知恵袋)，PB(書籍) 及び PN(新聞) の 3 つの領域を用いた実験の結果，全対象単語から選択する手法は効果があったが，領域間距離を考慮した手法には効果は見られなかった．

そこで，仮に領域間距離を正しく測れているとして負の転移の有無を判定し，その判定を用いてデータ選択の手順を変える手法を試した結果，領域間距離を利用するという手法には効果が有ることが確認できた．

提案手法を効果的に利用するには，今後は領域間距離を正しく測る方法と領域間距離をラベル付けするデータ選択に取り入れる方法を考えることが今後の課題である．

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.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono (2010) “SemEval-2010 Task: Japanese WSD,” in *The 5th International Workshop on Semantic Evaluation*, pp. 69–74.

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

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

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.