語義曖昧性解消の領域適応における Misleading データの存在と検出

吉田 拓夢 (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]

Existence and Detection of Misleading Data in Domain Adaptation for Word Sense Disambiguation

Hiromu Yoshida (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，負の転移を起こす Misleading データの存在と検出について議論する．

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]．

領域適応の問題を扱う場合，ソース領域のラベル付きデータは全て使う方がよいと考えられるが，必ずしもそうではない．ソース領域のラベル付きデータの一部は利用しない方が逆に精度が向上することがある．これは負の転移現象 (Rosenstein et al. (2005)) と呼ばれる現象である．Jiang はこのような悪影響を及ぼすデータを Misleading データと呼び，そのデータを検出・削除してから学習を行うことを提案している (Jiang and Zhai (2007))．一方，新納は WSD の領域適応の問題を，語義分布の推定の問題とスパース性の問題に分けて考え，語義分布の推定に影響を及ぼさなければ，ソース領域のラベル付きデータは全て使う方が良いことを主張している (新納浩幸・佐々木稔 (2013))．

WSD の領域適応の問題は共変量シフトの問題 (Shimodaira (2000) 杉山将 (2006)) と見なせるため，密度比を用いてデータの重要性を測ることができる．ここでは密度比の低いデータを Misleading データとして設定できるかどうかを調べる．

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，WSD の領域適応の実験を行なった．密度比を用いてデータの重要度を測り，その値が低いデータを Misleading データと考え，それらデータを外した場合の平均正解率を調べた．結果，密度比から Misleading データを検出する効果はなかった．総当たり的に各データが Misleading データかどうかを確認したところ，Misleading データの存在は確認できた．また密度比からは Misleading データを検出することが困難であることも示した．有効な Misleading データの検出法を今後の課題とする．

2　語義曖昧性の領域適応

WSD の対象単語 w の語義の集合を $C = \{c_1, c_2, \cdots, c_k\}$，$w$ を含む文（入力データ）を x とする．WSD の問題は最大事後確率推定を利用すると，以下の式の値を求める問題として表現できる．

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

つまり訓練データを利用して語義の分布 $P(c)$ と各語義上での入力データの分布 $P(x|c)$ を推定することで WSD の問題は解決できる．今，ソース領域を S，ターゲット領域を T とした場合，WSD の領域適応の問題は $P_S(c) \neq P_T(c)$ と $P_S(x|c) \neq P_T(x|c)$ から生じている．

[2] 10t4067l@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．

新納は $P_S(x|c) = P_T(x|c)$ は成立していると考え，$P_T(x|c)$ の推定を困難にしているのはコーパスのスパース性の問題として捉えた．つまり $P_T(c)$ の推定に影響を及ぼさなければ，ソース領域のラベル付きデータは全て利用しても問題ないことを示した (新納浩幸・佐々木稔 (2013))．

一般に識別モデルである SVM は語義分布の影響をあまり受けずに識別境界を定める．このため，新納の結果に基づければ，SVM で学習するのであれば，ソース領域のラベル付きデータは全て利用しても精度の低下は生じない，あるいは非常に小さいと考えられる．

本論文では WSD に SVM を用いる．その際に，Misleading データと見なせるものを省くことで精度が向上するかどうかを確認する．

3　密度比による Misleading データの検出

語義が曖昧な単語 w を含む文 s があったとき，この s がどのような領域のコーパスに出現したとしても w の語義が変化するとは考えられない．この点から Kikuchi は WSD の領域適応の問題は共変量シフトの問題と見なせることを示し，共変量シフトの問題の解法手法を利用して WSD の領域適応の解決を図った (Kikuchi and Shinnou (2013))．そこではデータ x に対して密度比 $r = P_T(x)/P_S(x)$ を測り，その r を x の重みとして学習する．密度比はデータ x の重要度を表しているので，ここでは密度比の小さなものを Misleading データと見なすことにする．

密度比の測り方だが，ここでは論文 (Kikuchi and Shinnou (2013)) で示された方法を使う．

対象単語の w の用例 $\boldsymbol{x}$ の素性リストを $\{f_1, f_2, \cdots, f_n\}$ とする．求めるのは領域 $R \in \{S, T\}$ 上の $P_R(\boldsymbol{x})$ である．まず以下を仮定する．

$$P_R(\boldsymbol{x}) = \prod_{i=1}^{n} P_R(f_i)$$

領域 R のコーパス内の w の全ての用例について素性リストを作成し，素性 f の頻度を $n(R, f)$ とおく．また素性の総頻度を $N(R)$ とおく．つまり $N(R) = \sum_{f \in R} n(R, f)$ である．次に領域 S と領域 T における w に関する素性の種類数を M とする．

$P_R(f)$ を以下で定義する．

$$P_R(f) = \frac{n(R, f) + 1}{N(R) + M}$$

4　実験

実験では BCCWJ コーパスの OC(Yahoo!知恵袋) と PB(書籍) を用いる．両領域で頻度 50 以上の表 1 に示す 17 単語を対象単語として，OC から PB，PB から OC の 2 通りの WSD の領域適応を行う．

表 1: 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
持つ	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

ソース領域の各データについて密度比を測り，その値が閾値 θ 以下のものを Misleading データと見なし，訓練データから Misleading データを除いた．この訓練データから SVM で学習し，語義識別の平均正解率 (%) を調べた．結果を表 2 と表 3 に示す．表 2 は OC から PB の領域適応であり，表 3 は PB から OC の領域適応である．表の「そのまま」は Misleading データの検出を行わずに，訓練データをすべて利用した場合の識別に対応する．

実験は $\theta = 0.1$ と $\theta = 0.05$ で試したが，どちらのケースにおいても Misleading データを除いて学習する効果はなかった．

表 2: 実験結果 (OC → PB)

単語	そのまま	Misleading を削除	
		$\theta = 0.1$	$\theta = 0.05$
言う	78.37	80.61	79.80
入れる	71.43	67.86	71.43
書く	67.74	80.65	85.48
聞く	64.23	65.04	65.04
来る	97.12	97.12	97.12
子供	31.18	30.11	26.88
時間	87.84	85.14	85.14
自分	92.21	84.74	86.69
出る	59.21	59.87	62.50
取る	27.16	27.16	27.16
場合	85.40	82.48	85.40
入る	46.61	45.76	36.44
前	78.13	74.38	73.75
見る	82.78	83.52	82.78
持つ	78.43	69.93	69.93
やる	92.95	92.95	92.95
ゆく	88.72	87.22	87.22
平均	72.32	71.44	71.51

表 3: 実験結果 (PB → OC)

単語	そのまま	Misleading を削除	
		$\theta = 0.1$	$\theta = 0.05$
言う	79.13	79.43	80.63
入れる	72.60	71.23	49.32
書く	73.74	73.74	73.74
聞く	67.74	58.06	56.45
来る	79.89	79.89	79.89
子供	23.38	46.75	25.97
時間	83.02	83.02	83.02
自分	87.50	87.50	87.50
出る	67.18	51.15	66.41
取る	37.70	32.79	34.43
場合	86.51	84.13	85.71
入る	57.35	45.59	45.59
前	86.67	86.67	86.67
見る	55.73	56.49	56.49
持つ	83.87	46.77	59.68
やる	94.02	94.02	93.16
ゆく	68.49	68.49	68.49
平均	70.85	67.40	66.66

5　考察

5.1　Misleading の存在

実験では Misleading データを除いて学習を行う効果は確認できなかった．原因は Misleading データが存在しない，あるいは密度比では Misleading データを検出できない，のいずれかである．

本節では，上記の点を明らかにするために，総当たり的に各データが Misleading データかどうかを調べることで Misleading データの存在を調べる．具体的には先の実験で使用した各々の訓練データ $D = \{d_1, d_2, \ldots, d_N\}$ について i 番目のデータ d_i を除いた訓練データ $D_i = \{d_1, d_2, \ldots, d_{i-1}, d_{i+1}, \ldots, d_N\}$ を用意し，D_i を用いて学習した際の正解率を調べ，正解率が上昇するデータを Misleading データとした．このようにして設定した Misleading データを除いて学習を行った結果を表 4，表 5 に示す．OC から PB また PB から OC のどちらの領域適応においても，平均正解率の向上が確認でき，Misleading データは存在すると結論づけられる．

表 4: Misleading データの存在確認実験 (OC → PB)

単語	そのまま	Misleading を削除	Misleading の個数 (データ数)
言う	78.37	82.50	159 (666)
入れる	71.43	75.00	6 (73)
書く	67.74	82.26	21 (99)
聞く	64.23	73.98	26 (124)
来る	97.12	97.12	0 (189)
子供	31.18	39.78	5 (77)
時間	87.84	90.54	1 (53)
自分	92.21	97.08	13 (128)
出る	59.21	64.47	14 (131)
取る	27.16	32.10	6 (61)
場合	85.40	85.40	0 (126)
入る	46.61	45.76	36 (68)
前	78.13	84.38	8 (105)
見る	82.78	84.98	10 (262)
持つ	78.43	72.55	8 (62)
やる	92.95	92.95	0 (117)
ゆく	88.72	88.72	17 (219)
平均	72.32	75.86	19.41 (150.59)

表 5: Misleading データの存在確認実験 (PB → OC)

単語	そのまま	Misleading を削除	Misleading の個数 (データ数)
言う	79.13	82.43	127 (1114)
入れる	72.60	68.49	19 (56)
書く	73.74	73.74	0 (62)
聞く	67.74	68.55	26 (123)
来る	79.89	80.42	1 (104)
子供	23.38	46.75	12 (93)
時間	83.02	83.02	0 (74)
自分	87.50	87.50	0 (308)
出る	67.18	64.89	39 (152)
取る	37.70	42.62	10 (81)
場合	86.51	91.27	7 (137)
入る	57.35	60.29	38 (118)
前	86.67	89.52	10 (160)
見る	55.73	56.87	3 (273)
持つ	83.87	85.48	2 (153)
やる	94.02	94.02	0 (156)
ゆく	68.49	70.32	15 (133)
平均	70.85	73.31	18.18 (193.94)

5.2 密度比による Misleading の検出

前述の実験により Misleading データの存在は確認できた．次に密度比によりどの程度の Misleading データが検出できるのかを調べた．まず前述の実験により得られた Misleading データを検出の正解データと見なした場合，密度比による Misleading データの検出での再現率を表 7 に正解率を表 8 に示す．

表 7 より，PB から OC への領域適応の場合，密度比による Misleading データの検出の再現率は

高いが，OC から PB への領域適応の場合はあまり高くないことが確認できる．次に表 8 からは，どちらの方向の領域適応においても検出の正解率は低い．

検出の F 値は表 6 となる．

表 6: 密度比による Misleading データ検出の F 値

	OC → PB	PB → OC
$\theta = 0.1$	0.2087	0.2281
$\theta = 0.05$	0.1984	0.2185

本論文の実験結果（表 4 と表 5）から考えて，上記程度の検出能力では WSD の領域適応には効果が出ない．結論としては，密度比による Misleading データの判別は有用ではないと考える．

表 7: 密度比による Misleading データ検出の再現率 (%)

単語	OC → PB		PB → OC	
	$\theta = 0.1$	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.05$
言う	70.44	64.15	83.46	71.65
入れる	50.00	50.00	89.47	89.47
書く	85.71	80.95	-	-
聞く	57.69	53.85	88.46	65.38
来る	-	-	100.00	100.00
子供	80.00	60.00	83.33	75.00
時間	0.00	0.00	-	-
自分	69.23	61.54	-	-
出る	78.57	78.57	89.74	87.18
取る	66.67	66.67	100.00	100.00
場合	-	-	71.43	71.43
入る	75.00	69.44	78.95	71.05
前	62.50	50.00	90.00	80.00
見る	80.00	50.00	100.00	100.00
持つ	50.00	25.00	100.00	100.00
やる	-	-	-	-
ゆく	100.00	100.00	66.67	53.33
平均	66.13	57.87	87.81	81.88

表 8: 密度比による Misleading データ検出の正解率 (%)

単語	OC → PB		PB → OC	
	$\theta = 0.1$	$\theta = 0.05$	$\theta = 0.1$	$\theta = 0.05$
言う	25.63	26.02	11.76	11.45
入れる	4.76	5.17	32.69	34.00
書く	20.22	20.99	-	-
聞く	16.13	15.73	23.23	19.10
来る	-	-	1.03	1.08
子供	5.56	4.29	12.05	11.25
時間	0.00	0.00	-	-
自分	10.23	9.64	-	-
出る	10.38	11.22	25.36	25.37
取る	7.55	7.55	13.70	13.89
場合	-	-	4.24	4.50
入る	46.55	45.45	28.85	27.00
前	6.17	5.26	5.96	5.67
見る	3.77	2.58	1.25	1.38
持つ	7.69	4.08	1.44	1.50
やる	-	-	-	-
ゆく	8.81	9.60	8.93	7.69
平均	12.39	11.97	13.11	12.61

5.3　語義分布推定への影響

ここでは「Misleading データは存在する」と結論づけたが，これは新納の結果 (新納浩幸・佐々木稔 (2013)) と矛盾するものではない．論文 (新納浩幸・佐々木稔 (2013)) の主張からは「Misleading データは存在しない」ことになるが，これは「語義分布の推定に影響を与えなければ」という前提が存在する．本論文で発見できた Misleading データを除くことで訓練データ中の語義分布が変化している可能性がある．この点は早急に確認したい．

6　おわりに

本論文では WSD の領域適応における Misleading データの存在と検出について論じた．密度比の低いものを Misleading データと見なす方法を試したが，効果はなかった．総当たり的に各データが Misleading データと見なせるかどうかを調べることで，Misleading データの存在を確認できた．また密度比を利用した Misleading データの検出については，その検出能力が低いことも示した．今後はまず Misleading データの削除により語義分布に変化が生じているかどうかの確認する必要がある．また精度の高い Misleading データの検出法も考えたい．

文献

Jing Jiang and Chengxiang Zhai (2007) “Instance weighting for domain adaptation in NLP,” in *ACL-2007*, pp. 264–271.

Hironori Kikuchi and Hiroyuki Shinnou (2013) “Domain Adaptation for Word Sense Disambiguation under the Problem of Covariate Shift,” 情報処理学会自然言語処理研究会報告，pp.NL–212–4.

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

Michael T Rosenstein, Zvika Marx, Leslie Pack Kaelbling, and Thomas G Dietterich (2005) “To transfer or not to transfer,” in *NIPS 2005 Workshop on Transfer Learning*, Vol. 898.

Hidetoshi Shimodaira (2000) “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, Vol. 90, No. 2, pp. 227–244.

新納浩幸、佐々木稔 (2013)「k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応」，情報処理学会自然言語処理研究会報告，pp.NL–211–13.

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.

杉山将 (2006)「共変量シフト下での教師付き学習」，日本神経回路学会誌，第 13 巻，第 3 号，pp.111–118.