Vol.2013-NL-211 No.13
Vol.2013-SLP-96 No.13
2013/5/24

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

新納 浩幸[1,a)]　佐々木 稔[1,b)]

概要：本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) の領域適応に対する手法を提案する．WSD の領域適応の問題は，2 つの問題に要約できる．1 つは領域間で語義の分布が異なる問題，もう 1 つは領域の変化によりデータスパースネスが生じる問題である．本論文では上記の点を論じ，前者の問題の対策として学習手法に k-近傍法を補助的に用いること，後者の問題の対策としてトピックモデルを用いることを提案する．具体的にはターゲットドメインから構築できるトピックモデルによって，ソースドメインの訓練データとターゲットドメインのテストデータにトピック素性を追加する．拡張された素性ベクトルから SVM を用いて語義識別を行うが，識別の信頼性が低いものには k-近傍法の識別結果を用いる．BCCWJ コーパスの 2 つの領域 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い，提案手法の有効性を示す．領域の一般性を考慮したトピックモデルを WSD に利用する方法，および WSD の領域適応に有効なアンサンブル手法を考案することを今後の課題とする．

キーワード：語義曖昧性解消, 領域適応, トピックモデル, k-近傍法, 教師なし学習

Domain Adaptation for Word Sense Disambiguation using k-Nearest Neighbors Method and Topic Model

Abstract: In this paper, we propose the method of domain adaptation for Word Sense Disambiguation (WSD). The problem of domain adaptation for WSD is essentially divided into two problems. The first problem is the difference between sense distributions on domains. The second problem is the data sparseness with changing the domain. In this paper, we discuss about it, and propose a countermeasure for each problem. We use the k-nearest neighbor method (k-NN) subsidiarily for the first problem, and use the topic model for the scond problem. In particular, we add topic features made by the topic model built from a target domain corpus to training data in a soruce domain and test data in the target domain. We learn SVM using extended features, and solve WSD. However, when the degree of reliability of judgement of SVM for a test instance is low, we use judgement of k-NN for its instance. In the experiment, we select 17 ambiguous words in both domains, PB (books) and OC(Yahoo! Chie Bukuro) in BCCWJ corpus, which have 50 and more frequent in both domain corppus, and conduct the experimental of domain adaptation for WSD using these data to show the effectiveness of our merhod. In future, we investigate a way to use the topic model in consideration of the universality of a corpus, and an effective ensemble learning for domain adaptation for WSD.

Keywords: Word sense disambiguation, Domain adaptation, Topic model, k-nearest neighbor method, unsupervised learning

1. はじめに

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

[1] 茨城大学 工学部 情報工学科
4-12-1 Nakanarusawa, Hitachi, Ibaraki 316–8511, Japan
[a)] shinnou@mx.ibaraki.ac.jp
[b)] msasaki@mx.ibaraki.ac.jp

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[*1]．本論文では語義曖昧性解消 (Word Sense Disambiguation，WSD) のタスクでの領域適応を行う．

領域適応の手法はターゲット領域のラベル付きデータを利用するかしないかという観点で分類できる．利用する場合を教師付き手法，利用しない場合を教師なし手法と呼ぶ．教師付き手法については多くの研究がある[*2]．また能動学習 [20] や半教師付き学習 [8] は，領域適応の問題に直接利用できるために，それらのアプローチをとる研究も多い．これらに対して教師なし手法の従来研究は少ない．教師なし手法は教師付き手法に比べパフォーマンスが悪いが，ラベル付けが必要ないという大きな長所がある．また教師なし手法を研究することで，領域適応の問題が明確になると考えている．この点から本論文では教師なし手法を試みる．

本論文の特徴は WSD の領域適応の問題を以下の 2 点に分割したことである．

(1) 領域間で語義の分布が異なる
(2) 領域の変化によりデータスパースネスが生じる

実際の領域適応の手法は上記 2 つの問題を同時に解決しているものが多いために，このような捉え方をしていないが，WSD の領域適応の場合，上記 2 つの問題を分けて考えた方が，何を解決しようとしているのかが明確になる．本論文では上記 2 点の問題に対して，ターゲット領域のラベル付きデータを必要としない各々の対策案を提示する．具体的に，(1) に対しては k-近傍法を補助的に利用し，(2) に対しては領域毎のトピックモデル [1] を利用する．実際の処理は，ターゲットドメインから構築できるトピックモデルによって，ソースドメインの訓練データとターゲットドメインのテストデータにトピック素性を追加する．拡張された素性ベクトルから SVM を用いて語義識別を行うが，識別の信頼性が低いものには k-近傍法の識別結果を用いる．

実験では BCCWJ コーパス [17] 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い，提案手法の有効性を示す．

2. WSD の領域適応の問題

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 の問題は事後確率最大化に基づければ以下で表せる．

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

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．

$P_S(c) \neq P_T(c)$ が成立していることは明らかだが，$P_S(x|c) \neq P_T(x|c)$ に対しては一考を要する．一般の領域適応の問題では $P_S(x|c) \neq P_T(x|c)$ であるが，WSD に限れば $P_S(x|c) = P_T(x|c)$ と考えることもできる．実際 Chan らは $P_S(x|c)$ と $P_T(x|c)$ の違いの影響は非常に小さいと考え，$P_S(x|c) = P_T(x|c)$ を仮定し，$P_T(c)$ を EM アルゴリズムで推定することで WSD の領域適応を行っている [6][5]．古宮らは 2 つのソース領域の訓練データを用意し，そこからランダムに訓練データを取り出して WSD の分類器を学習している [23]．論文中では指摘していないが，これも $P_S(c)$ を $P_T(c)$ に近づける工夫である．ソース領域が 1 つだとランダムに訓練データを取り出しても $P_S(c)$ は変化しないが，ソース領域を複数用意することで $P_S(c)$ が変化する．

ただし $P_S(x|c) = P_T(x|c)$ が成立していたとしても，WSD の領域適応の問題が $P_T(c)$ の推定に帰着できるわけでない．仮に $P_S(x|c) = P_T(x|c)$ であったとしても，領域 S の訓練データだけから $P_T(x|c)$ を推定することは困難だからである．これは共変量シフトの問題 [21][26] と関連が深い．共変量シフトの問題とは入力 x と出力 y に対して，推定する分布 $P(y|x)$ が領域 S と T で共通しているが，S における入力の分布 $P_S(x)$ と T における入力の分布 $P_T(x)$ が異なる問題である．$P_S(x|c) = P_T(x|c)$ の仮定の下では，入力 x と出力 c が逆になっているので，共変量シフトの問題とは異なる．ただし WSD の場合，全く同じ文 x が別領域に出現したとしても，x 内の多義語 w の語義が異なることはないので $P_S(c|x) = P_T(c|x)$ が成立している．$P_T(c|x)$ は語義識別そのものなので，WSD の領域適応の問題は共変量シフトの問題として扱えることができる．共変量シフト下では訓練事例 x_i に対して密度比 $P_T(x_i)/P_S(x_i)$ を推定し，密度比を重みとして尤度を最大にするようにモデルのパラメータを学習する．Jiang らは密度比を手動で調整し，モデルにはロジステック回帰を用いている [12]．齋木らは $P(x)$ を unigram でモデル化することで密度比を推定し，モデルには最大エントロピーモデルを用いている [28]．ただしどちらの研究もタスクは WSD ではない．WSD では $P(x)$ が単純な言語モデルではなく，「x は対象単語 w を含む」という条件が付いているので，密度比 $P_T(x)/P_S(x)$ の推定が困難となっている．また教師なしの枠組みで共変量シフトの問題が扱えるのかは不明である．

本論文では $P_S(c|x) = P_T(c|x)$ を仮定したアプローチは取らず，$P_S(x|c) = P_T(x|c)$ を仮定する．この仮定があったとしても，領域 S の訓練データだけから $P_T(x|c)$ を推定するのは困難である．ここではこれをスパース性の問題と

[*1] 領域適応は機械学習の分野では転移学習 [25] の一種と見なされている．
[*2] 例えば Daumé の研究 [11] はその簡易性と有効性から広く知られている．

考える．つまり領域 S の訓練データ D は領域 T においてスパースになっていると考える．スパース性の問題だと考えれば，半教師あり学習や能動学習を領域適応に応用するのは自然である[*3][18]．また半教師あり学習や能動学習のアプローチを取った場合，T の訓練データが増えるので語義の分布の違い自体も同時に解消されていく [7]．

ここで指摘したいのは $P_S(x|c) = P_T(x|c)$ が成立しており $P_T(x|c)$ の推定を困難にしているのがスパース性の問題だとすれば，領域 S の訓練データ D は多いほどよい推定が行えるはずで，D が大きくなったとしても推定が悪化するはずがない点である．しかし現実には D を大きくすると WSD 自体の精度が悪くなる場合もあることが報告されている（例えば [23]）．これは一般に負の転移現象 [19] と呼ばれている．WSD の場合 $P_T(x|c)$ を推定しようとして，逆に語義の分布 $P_T(c)$ の推定が悪化することから生じる．つまり領域 T における WSD の解決には T におけるデータスパースネスの問題に対処しながら，同時に $P_T(c)$ の推定が悪化することを避けることが必要となる．

3. 提案手法

3.1 k-近傍法の利用

領域 T におけるデータスパースネスの問題に対処する際に，$P_T(c)$ の推定が悪化することを避けるために，本論文では識別の際に $P_T(c)$ の情報をできるだけ利用しないという方針をとる．そのために k-近傍法を利用する．どのような学習手法を取ったとしても，何らかの汎化を行う以上，$P_T(c)$ の影響を受けるが，k-近傍法はその影響が少ない．k-近傍法はデータ x のクラスを識別するのに，訓練データの中から x と近いデータ k 個を取ってきて，それら k 個のデータのクラスの多数決により x のクラスを識別する．k-近傍法が $P_T(c)$ の影響が少ないのは $k = 1$ の場合（最近傍法）を考えればわかりやすい．例えば，クラスが $\{c_1, c_2\}$ であり，$P(c_1) = 0.99$，$P(c_2) = 0.01$ であった場合，通常の学習手法であれば，ほぼ全てのデータを c_1 と識別するが，最近傍法では，入力データと最も近いデータ 1 つだけがクラス c_2 であれば，入力データのクラスを c_2 と判断する (図 1 参照)．つまり k-近傍法ではデータ全体の分布を考慮せずに k 個の局所的な近傍データのみでクラスを識別するために，その識別には $P_T(c)$ の影響が少ない．

ただし k-近傍法は近年の学習器と比べるとその精度が低い．そのためここでは k-近傍法を補助的に利用する．具体的には通常の識別は SVM で行い，SVM での識別の信頼度が閾値 θ 以下の場合のみ，k-近傍法の識別結果を利用することにする．

ここで θ の値が問題だが，語義の数が K 個である場合，識別の信頼度（その語義である確率）は少なくとも $1/K$ 以上の値となる．そのためここではこの値の 1 割をプラスし $\theta = 1.1/K$ とした．

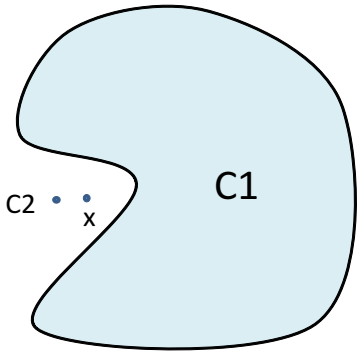

図 1 分布の影響が少ない k-NN

3.2 トピックモデルの利用

領域 T におけるデータスパースネスの問題に対処するために，ここではトピックモデルを利用する．

WSD の素性としてシソーラスの情報を利用するのもデータスパースネスへの 1 つの対策である．シソーラスとしては，分類語彙表などの手作業で構築されたものとコーパスから自動構築されたものがある．前者は質が高いが分野依存の問題がある．後者は質はそれほど高くないが，分野毎に構築できるという利点がある．ここでは領域適応の問題を扱うので，後者を利用する．つまり領域 T からシソーラスを自動構築し，そのシソーラス情報を領域 S の訓練事例と領域 T のテスト事例に含めることで，WSD の識別精度の向上を目指す．注意として，WSD では単語間の類似度を求めるためにシソーラスを利用する．そのため実際にはシソーラスを構築するのではなく，単語間の類似度が測れる仕組みを作っておけば良い．この仕組みが単語のクラスタリング結果に対応する．

この単語のクラスタリング結果を得るためにトピックモデルを利用する，トピックモデルとは文書 d の生起に K 個の潜在的なトピック z_i を導入した確率モデルである．

$$p(d) = \sum_{i=1}^{K} p(z_i)p(d|z_i)$$

トピックモデルの 1 つである Latent Dirichlet Allocation (LDA) [1] を用いた場合，単語 w に対して $p(w|z_i)$ が得られる．つまりトピック z_i をひとつのクラスタと見なすことで，LDA を利用して単語のソフトクラスタリングが可能となる．

領域 T のコーパスと LDA を利用して，T に適した $p(w|z_i)$ が得られる．$p(w|z_i)$ の情報を WSD に利用するいくつかの研究 [16][3][2] があるが，ここでは基本的にハードタグ [4] を利用する．ハードタグとは w に対して最も関連

[*3] ただし D は領域 T 内のサンプルではなく不均衡な訓練データという点には注意すべきであり，この点を考慮した半教師あり学習や能動学習が必要である．

度の高いトピック $z_{\hat{i}}$ を付与する方法である．

$$\hat{i} = \arg\max_i p(w|z_i)$$

まずトピック数を K としたとき，K 次元のベクトル t を用意し，入力事例 x 中の単語 $w_j(j = 1 \sim n)$ に対して最も関連度の高いトピック $z_{\hat{i}}$ を求め，t の $\hat{i}$ 次元の値を 1 にする．これを w_1 から w_n まで行い t を完成させる．作成できた t をここでは**トピック素性**と呼ぶ．トピック素性を通常の素性ベクトル (ここでは**基本素性**と呼ぶ) に結合することで，新たな素性ベクトルを作成し，その素性ベクトルを対象に学習と識別を行う．

なお，本論文で利用した基本素性は，対象単語の前後の単語と品詞及び対象単語の前後 3 単語までの自立語である．

4. 実験

現代日本語書き言葉均衡コーパス（BCCWJ コーパス [17]）の PB(書籍) と OC(Yahoo! 知恵袋) を異なった領域として実験を行う．PB と OC から共に頻度が 50 以上の多義語 17 単語を WSD の対象単語とする．これら単語と辞書上での語義数及び各コーパスでの頻度と語彙数を表 1 に示す[*4]．領域適応としては PB をソース領域，OC をターゲット領域としたものと，OC をソース領域，PB をターゲット領域としたものの 2 種類を行う．

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

PB から OC への領域適応の実験結果を表 2 に示す．また OC から PB への領域適応の実験結果を表 3 に示す．表 2 と表 3 の数値は正解率を示している．「k-NN」の列は k-近傍法の識別結果を示す．ここでは $k = 1$ としている．「SVM」の列は基本素性だけを用いて学習した SVM の識別結果を示し，「SVM + TM」の列は基本素性にターゲット領域から得たトピック素性を加えた素性を用いて学習した SVM の識別結果を示し，「提案手法」の列は「SVM + TM」の識別で信頼度の低い結果を k-近傍法の結果に置き換えた場合の識別結果を示す．また「self」は対象領域の訓練データに対して 5 分割交差検定を行った場合の平均正解率であり，理想値と考えて良い．

17 単語の正解率の平均をみると，PB から OC への領域適応と OC から PB への領域適応のどちらにおいても，以下の関係が成立しており，提案手法が有効であることがわかる．

k-NN < SVM < SVM+TM < 提案手法

また「提案手法」の中には一部 「self」の値よりも高いものも存在する．これはトピックモデルを利用した効果である．

表 2 実験結果 (PB → OC)

単語	k-NN	SVM	SVM+TM	提案手法	self
言う	0.8318	0.8093	0.7958	0.8033	0.8859
入れる	0.6438	0.7534	0.7671	0.7671	0.7266
書く	0.6767	0.7373	0.7373	0.7373	0.7900
聞く	0.6371	0.6451	0.6612	0.6693	0.7503
来る	0.7883	0.7989	0.7989	0.7989	0.8890
子供	0.3766	0.1818	0.2207	0.2207	0.9108
時間	0.7924	0.8301	0.8301	0.8301	0.8709
自分	0.8671	0.8750	0.8750	0.8750	0.8978
出る	0.5877	0.7022	0.7099	0.7099	0.7111
取る	0.1475	0.2459	0.2950	0.2950	0.6217
場合	0.7460	0.8968	0.9127	0.9127	0.9760
入る	0.4117	0.5735	0.5735	0.5735	0.7494
前	0.7904	0.9142	0.8857	0.8952	0.8952
見る	0.5343	0.5839	0.5954	0.5954	0.9119
持つ	0.7419	0.8709	0.7741	0.7741	0.8871
やる	0.9145	0.9316	0.9401	0.9401	0.9652
ゆく	0.6529	0.6803	0.6803	0.6803	0.9316
平均	0.6553	0.7077	0.7090	0.7105	0.8453

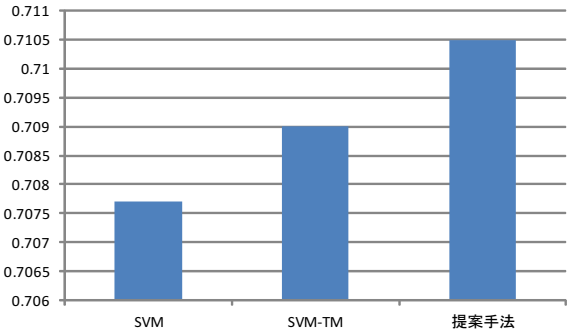

図 2 正解率 (PB → OC)

表 3 実験結果 (OC → PB)

単語	k-NN	SVM	SVM+TM	提案手法	self
言う	0.7737	0.8249	0.7953	0.7953	0.9075
入れる	0.5535	0.7500	0.7142	0.7321	0.7681
書く	0.7258	0.8387	0.8064	0.8548	0.9051
聞く	0.6178	0.6585	0.6829	0.6910	0.7543
来る	0.9519	0.9711	0.9711	0.9711	0.9804
子供	0.3978	0.3333	0.4193	0.4193	0.8192
時間	0.6351	0.8918	0.8918	0.8918	0.8895
自分	0.9480	0.9318	0.9577	0.9610	0.9772
出る	0.5526	0.5789	0.6118	0.6118	0.7303
取る	0.1851	0.2345	0.2716	0.2716	0.4330
場合	0.8613	0.8467	0.8467	0.8467	0.8910
入る	0.4067	0.4915	0.5254	0.5254	0.6094
前	0.8500	0.8062	0.8312	0.8312	0.9250
見る	0.8168	0.8388	0.8424	0.8424	0.8498
持つ	0.7777	0.8039	0.7777	0.7777	0.7907
やる	0.8846	0.9294	0.9294	0.9294	0.9360
ゆく	0.8947	0.8872	0.9097	0.9097	0.8717
平均	0.6961	0.7422	0.7520	0.7566	0.8258

[*4] 語義は岩波国語辞書がもとになっている．そこでの中分類までを対象にした．また「入る」は辞書上の語義が 3 つだが，PB や OC では 4 つの語義がある．これは BCCWJ コーパスでは新語義のタグも許しているからである．

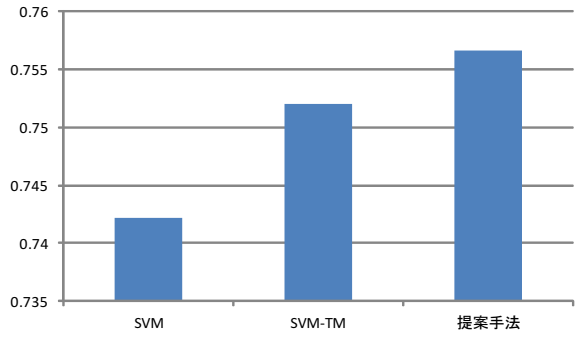

図 3 正解率 (OC → PB)

5. 考察

5.1 語義分布の違い

本論文では，WSD の領域適応は語義分布の違いの問題を解決するだけは不十分であることを述べた．Naive Bayes を利用して，この点を調べた．Naive Bayes の場合，以下の式で語義を識別する．

$$\arg\max P_S(c)P_s(x|c)$$

ここで事前分布 $P_S(c)$ の代わりに領域 T の訓練データから推定した $P_T(c)$ を用いる．これは語義分布を正確に推定できたという仮定での仮想的な実験である．結果を表 4 に示す．

表 4 理想的語義分布の推定による識別

単語	PB -> OC PB の事前分布	PB -> OC OC の事前分布	OC -> PB OC の事前分布	OC -> PB PB の事前分布
言う	0.78861	0.79010	0.79910	0.80269
入れる	0.72973	0.72973	0.61404	0.61404
書く	0.73000	0.73000	0.88889	0.88889
聞く	0.66400	0.65600	0.75000	0.75000
来る	0.79474	0.79474	0.96190	0.96190
子供	0.12821	0.16667	0.24468	0.24468
時間	0.81481	0.81481	0.89333	0.89333
自分	0.86822	0.86822	0.97087	0.97087
出る	0.70455	0.70455	0.55556	0.55556
取る	0.09677	0.20968	0.21951	0.24390
場合	0.95276	0.96850	0.84058	0.84058
入る	0.65217	0.65217	0.42857	0.44538
前	0.90566	0.89623	0.63975	0.63975
見る	0.55894	0.55894	0.83577	0.83212
持つ	0.57143	0.57143	0.75974	0.72727
やる	0.93220	0.93220	0.92357	0.92357
ゆく	0.68182	0.68182	0.86567	0.86567
平均	0.68086	0.68975	0.71715	0.71766

全体として理想的な語義分布を利用すれば，正解率は改善されるが，効果はわずかしかない．また PB から OC の「前」や OC から PB の「見る」「持つ」は逆に精度が悪化している．更に理想的な語義分布を利用できたとしても，通常の SVM よりも正解率が劣っている．これらのことから，語義分布の正確な推定のみでは WSD の領域適応の解決は困難であることがわかる．

5.2 トピックモデルの領域依存性の度合い

WSD においてデータスパースネスの問題の対処として，シソーラスを利用することは一般に行われてきている．LDA から得られるトピック z_i のもとで単語 w が生起する確率 $p(w|z_i)$ は，単語のソフトクラスタリング結果に対応しており，これは LDA の処理対象となったコーパスに合ったシソーラスと見なせる．このためトピックモデルが WSD に利用できることは明かである．ただしその具体的な利用方法は未解決である．

問題は 2 つある．1 つはトピック素性の表現方法である．ここではハードタグを利用したが，ソフトタグの方が優れているという報告もある [4]．國井はハードタグとソフトタグの中間にあたるミドルソフトタグを提案している [27]．いずれにしても，トピック素性の有効な表現方法はトピック数やコーパスの規模にも依存した問題であり，どういった表現方法で利用すれば良いかは未解決である．

もう 1 つの問題はトピックモデルから得れるシソーラスの領域依存性の度合いである．本論文でも LDA から領域依存のトピックモデルが作成できることに着目してトピックモデルを領域適応の問題に利用した．ただし領域 A のコーパスと領域 B のコーパスがあった場合，各々のコーパスから各々の知識を獲得するよりも，両者のコーパスを合わせて両領域の知識を獲得した方が，一方のコーパスから得られる知識よりも優れていることがある．例えば森は単語分割のタスクにおいて，各々の領域のタグ付きデータを使うことで精度を上げることができたが，全ての領域のタグ付きデータを使えば更に精度を上げることができたことを報告している [24]．領域の知識を合わせることは，その知識をより一般的にしていることであり，領域依存の知識はあまり領域に依存しすぎるよりも，ある程度，一般性があった方がよいという問題と捉えられる．本実験で言えば PB のコーパスと OC のコーパスと両者を合わせて学習したトピックモデルは，各々のコーパスから学習したトピックモデルよりも優れている可能性がある．以下その実験の結果を表 5 に示す．

表 5 両領域コーパスを利用した識別

単語	PB -> OC OC の TM	PB -> OC OC+PB の TM	OC -> PB PB の TM	OC -> PB OC+PB の TM
言う	0.80931	0.80330	0.82496	0.80790
入れる	0.75342	0.78082	0.75000	0.76786
書く	0.73737	0.73737	0.83871	0.79032
聞く	0.64516	0.64516	0.65854	0.66667
来る	0.79894	0.79894	0.97115	0.97115
子供	0.18182	0.23377	0.33333	0.35484
時間	0.83019	0.83019	0.89189	0.87838
自分	0.87500	0.87500	0.93182	0.94156
出る	0.70229	0.70229	0.57895	0.58553
取る	0.24590	0.26230	0.23457	0.27160
場合	0.89683	0.92063	0.84672	0.84672
入る	0.57353	0.63235	0.49153	0.45763
前	0.91429	0.90476	0.80625	0.81250
見る	0.58397	0.60687	0.83883	0.83516
持つ	0.87097	0.83871	0.80392	0.79085
やる	0.93162	0.94017	0.92949	0.92949
ゆく	0.68037	0.68493	0.88722	0.89474
平均	0.70770	0.71750	0.74223	0.74135

領域 PB の場合，OC のコーパスを追加することで正解率は低下するが，領域 OC の場合，PB のコーパスを追加することで正解率が向上する．これは OC (Yahoo!知恵袋) のコーパスの領域依存が強いが，その一方で，PB (書籍) のコーパスの領域依存が弱く，より一般的であることから

生じていると考える．一般性の高い領域に領域依存の強い知識を入れると性能が下がるが，より特殊な領域には，その領域固有の知識に一般的知識を組み入れることで性能が更に向上すると考えられる．これらの詳細な分析と対策は今後の課題である．

5.3 k-近傍法の効果とアンサンブル手法

本論文では SVM での識別の信頼度の低い部分を k-近傍法の識別結果に置き換えるという処理を行った．置き換えが起こったものだけを対象にして，k-近傍法と SVM での正解率を比較した．結果を表 6 と表 7 に示す．

表 6 識別結果の変更 (PB → OC)

単語	変更数	k-NN	SVM-TP
言う	17	0.6471	0.3529
入れる	0	-	-
書く	0	-	-
聞く	5	0.8000	0.6000
来る	0	-	-
子供	10	0.4000	0.5000
時間	0	-	-
自分	0	-	-
出る	0	-	-
取る	0	-	-
場合	0	-	-
入る	0	-	-
前	2	0.5000	0.0000
見る	0	-	-
持つ	0	-	-
やる	0	-	-
ゆく	4	0.7250	0.5000
平均		0.6144	0.3906

表 7 識別結果の変更 (OC → PB)

単語	変更数	k-NN	SVM-TP
言う	63	0.6190	0.5873
入れる	7	0.2857	0.4286
書く	6	0.6667	0.1667
聞く	9	0.4444	0.4444
来る	1	0.0000	0.0000
子供	9	0.5556	0.4444
時間	0	-	-
自分	10	0.9000	0.7000
出る	0	-	-
取る	0	-	-
場合	0	-	-
入る	0	-	-
前	0	0.5000	0.0000
見る	0	-	-
持つ	0	-	-
やる	0	-	-
ゆく	2	0.0000	0.5000
平均		0.4413	0.3635

PB から OC への領域適応では「子供」，OC から PB への領域適応では「入れる」については SVM の方が k-近傍法の方よりもよい正解率だが，それ以外は k-近傍法の正解率は SVM の正解率と等しいかそれ以上であった．つまり SVM で識別精度が低い部分に関しては，k-近傍法で識別する効果が確認できる．

また k-近傍法の k をここでは $k = 1$ とした．この k の値を 3 や 5 に変更した実験結果を図 4 と図 5 に示す．

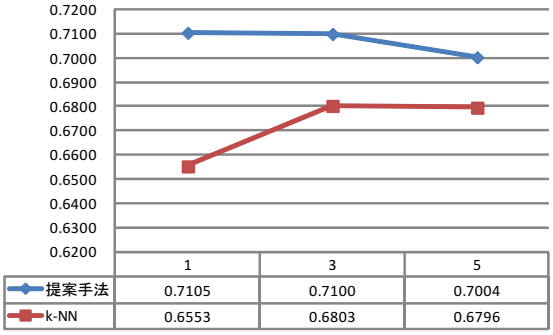

図 4 k による変化 (PB → OC)

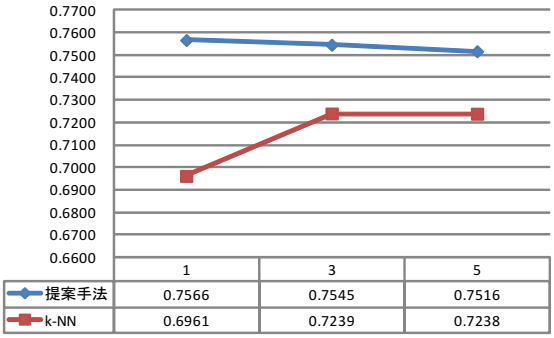

図 5 k による変化 (OC → PB)

複数の分類器を組み合わせて利用する学習手法をアンサンブル学習というが，本論文の手法もアンサンブル学習の一種と見なせる．k-近傍法自体は $k = 1$ よりも $k = 3$ や $k = 5$ の方が正解率が高いが，本手法のように SVM の識別の信頼度の低い部分のみに限定すれば，$k = 1$ の k-近傍法を利用した方がよい．これはアンサンブル学習では高い識別能力の学習器を組み合わせるのではなく，互いの弱い部分を補強し合うような形式が望ましいことを示している．

アンサンブル学習自体はかなり広い概念である．実際，バギング，ブースティングまた混合分布もアンサンブル学習の一種であり，アンサンブル学習を領域適応に応用した研究も多い．Daumé らは領域適応のための混合モデルを提案している [10]．そこでは，ソース領域のモデル，ターゲット領域のモデル，そしてソース領域とターゲット領域を共有したモデルの 3 つをの混合モデルの構成要素としている．Dai らは代表的なブースティングアルゴリズムの AdaBoost を領域適応の問題に拡張した TrAdaBoost を提案している [9]．また Kamishima らはバギングを領域適応の学習用に拡張した TrBagg を提案している [13]．

WSD の領域適応については古宮の一連の研究 [15][14][22] があるが，そこではターゲット領域のラベルデータの使い

方に応じて学習させた複数の分類器を用意しておき，単語や事例毎に最適な分類器を使い分けることで，WSD の領域適応を行っている．これらの研究もアンサンブル学習の一種と見なせる．アンサンブル学習は領域適応において有効な手法と考えられる．

6. おわりに

本論文では WSD の領域適応に対する手法を提案した．まず WSD の領域適応の問題を，以下の 2 つの問題に要約できることを示し，関連研究との位置づけを示した．

- 領域間で語義の分布が異なる
- 領域の変化によりデータスパースネスが生じる

次に上記の 2 つの問題それぞれに対処する手法を提案した．1 点目の問題に対しては k-近傍法を補助的に用いること，2 点目の問題に対してはトピックモデルを利用することである．BCCWJ コーパスの 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い，提案手法の有効性を示した．領域の一般性を考慮したトピックモデルを WSD に利用する方法，および WSD の領域適応に有効なアンサンブル手法を考案することを今後の課題とする．

参考文献

[1] Blei, D. M., Ng, A. Y. and Jordan., M. I.: Latent dirichlet allocation, *Machine Learning Reseach*, Vol. 3, pp. 993–1022 (2003).
[2] Boyd-Graber, J. and Blei, D.: Putop: Turning Predominant Senses into a Topic Model for Word Sense Disambiguation, *SemEval-2007*.
[3] Boyd-Graber, J., Blei, D. and Zhu, X.: A Topic Model for Word Sense Disambiguation, *EMNLP-CoNLL-2007*, pp. 1024–1033 (2007).
[4] Cai, J. F., Lee, W. S. and Teh, Y. W.: Improving Word Sense Disambiguation using Topic Features, *EMNLP-CoNLL-2007*, pp. 1015–1023 (2007).
[5] Chan, Y. S. and Ng, H. T.: Word sense disambiguation with distribution estimation, *IJCAI-05* (2005).
[6] Chan, Y. S. and Ng, H. T.: Estimating class priors in domain adaptation for word sense disambiguation, *COLING-ACL-2006*, Association for Computational Linguistics, pp. 89–96 (2006).
[7] Chan, Y. S. and Ng, H. T.: Domain adaptation with active learning for word sense disambiguation, *ACL-2007*, Vol. 45, No. 1, p. 49 (2007).
[8] Chapelle, O., Schölkopf, B., Zien, A. et al.: *Semi-supervised learning*, Vol. 2, MIT press Cambridge (2006).
[9] Dai, W., Yang, Q., Xue, G.-R. and Yu, Y.: Boosting for transfer learning, *ICML-2007*, pp. 193–200 (2007).
[10] Daumé III, H. and Marcu, D.: Domain adaptation for statistical classifiers, *Journal of Artificial Intelligence Research*, Vol. 26, No. 1, pp. 101–126 (2006).
[11] Daumé III, Hal: Frustratingly Easy Domain Adaptation, *ACL-2007*, pp. 256–263 (2007).
[12] Jiang, J. and Zhai, C.: Instance weighting for domain adaptation in NLP, *ACL-2007*, pp. 264–271 (2007).
[13] Kamishima, T., Hamasaki, M. and Akaho, S.: Trbagg: A simple transfer learning method and its application to personalization in collaborative tagging, *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, IEEE, pp. 219–228 (2009).
[14] Komiya, K. and Okumura, M.: Automatic Determination of a Domain Adaptation Method for Word Sense Disambiguation using Decision Tree Learning, *IJCNLP-2011*, pp. 1107–1115 (2011).
[15] Komiya, K. and Okumura, M.: Automatic Domain Adaptation for Word Sense Disambiguation Based on Comparison of Multiple Classifiers, *PACLIC-2012*, pp. 75–85 (2012).
[16] Li, L., Roth, B. and Sporleder, C.: Topic Models for Word Sense Disambiguation and Token-based Idiom Detection, *ACL-2010*, pp. 1138–1147.
[17] Maekawa, K.: Design of a Balanced Corpus of Contemporary Written Japanese, *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58 (2007).
[18] Rai, P., Saha, A., Daumé III, H. and Venkatasubramanian, S.: Domain adaptation meets active learning, *NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, Association for Computational Linguistics, pp. 27–32 (2010).
[19] Rosenstein, M. T., Marx, Z., Kaelbling, L. P. and Dietterich, T. G.: To transfer or not to transfer, *NIPS 2005 Workshop on Transfer Learning*, Vol. 898 (2005).
[20] Settles, B.: Active learning literature survey, *University of Wisconsin, Madison* (2010).
[21] Shimodaira, H.: 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 (2000).
[22] 古宮嘉那子， 奥村学：語義曖昧性解消のための領域適応手法の決定木学習による自動選択，自然言語処理，Vol. 19, No. 3, pp. 143–166 (2012).
[23] 古宮嘉那子，小谷善行， 奥村学：語義曖昧性解消の領域適応のための訓練事例集合の選択，言語処理学会第 19 回年次大会，pp. C6–2 (2013).
[24] 森信介：自然言語処理における分野適応，人工知能学会誌， Vol. 27, No. 4, pp. 365–372 (2012).
[25] 神嶌敏弘：転移学習，人工知能学会誌， Vol. 25, No. 4, pp. 572–580 (2010).
[26] 杉山将：共変量シフト下での教師付き学習，日本神経回路学会誌， Vol. 13, No. 3, pp. 111–118 (2006).
[27] 國井慎也，新納浩幸，佐々木稔：ミドルソフトタグのトピック素性を利用した語義曖昧性解消，言語処理学会第 19 回年次大会，pp. P3–9 (2013).
[28] 齋木陽介，高村大也， 奥村学：文の感情極性判定における事例重み付けによるドメイン適応，情報処理学会研究報告. 自然言語処理研究会報告，Vol. 2008, No. 33, pp. 61–67 (2008).