# 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[†]・佐々木　稔[†]

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領域から構築できるトピックモデルによって，ソース領域の訓練データとターゲット領域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．別種の領域間における本手法の有効性の確認，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 Neighbor Algorithm and Topic Model

Hiroyuki Shinnou[†] and Minoru Sasaki[†]

In this paper, we propose the method of domain adaptation for word sense disambiguation (WSD). This method faces the following problems for WSD. (1) The difference between sense distributions on domains. (2) The sparseness of data caused by changing the domain. In this paper, we discuss and recommend the countermeasure for each problem. We use the k-nearest neighbor algorithm (k-NN) and the topic model for the first and second problems, respectively. In particular, we append topic features developed by the topic model for target domain corpus to to training data in source domain and test data in target domain. Using the extended features of support vector machine (SVM) classifier, we solve WSD. However, when the reliability of decision of the SVM classifier for a test instance is low, we use the decision of the k-NN. In the experiment, we select 17 ambiguous words in both domains, PB (books) and OC (Yahoo! Chie Bukuro) in the balanced corpus of contemporary written Japanese (BCCWJ corpus), which appear 50 times or more in these domains, and conduct the experiment of domain adaptation for WSD using these words to show the effectiveness

[†] 茨城大学工学部情報工学科, Department of Computer and Information Sciences, Ibaraki University

自然言語処理 Vol. 20 No. 5 December 2013

of our method. In the future, we will apply the proposed method to other domains and examine a way to use the topic model considering the universality of a corpus, and an effective ensemble learning for domain adaptation for WSD.

**Key Words**: *Word Sense Disambiguation, Domain Adaptation, Topic Model, k-Nearest Neighbor Algorithm, 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に対する手法を提案する．

まず本論文における「領域」の定義について述べる．「領域」の正確な定義は困難であるが，本論文では現代日本語書き言葉均衡コーパス (BCCWJ コーパス)(Maekawa 2007) におけるコーパスの「ジャンル」を「領域」としている．コーパスの「ジャンル」とは，概略，そのコーパスの基になった文書が属していた形態の分類であり，書籍，雑誌，新聞，白書，ブログ，ネット掲示板，教科書などがある．つまり本論文における「領域」とは，書籍，新聞，ブログ等のコーパスの種類を意味する．

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 (Settles 2010) や半教師あり学習 (Chapelle, Schölkopf, Zien, et al. 2006) 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．また領域適応は転移学習と呼ばれることからも明らかなように，ソース領域の知識（例えば，ラベル付きデータからの知識）をどのように利用するか（ターゲット領域に転移させるか）が解決の鍵であり，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しないことで，その効果が明確になる．このため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．

[1] 領域適応は機械学習の分野では転移学習 (神嶌 2010) の一種と見なされている．
[2] 例えば Daumé の研究 (Daumé 2007) はその簡易性と有効性から広く知られている．

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

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 (Blei, Ng, and Jordan 2003) を利用する．実際の処理は，ターゲット領域から構築できるトピックモデルによって，ソース領域の訓練データとターゲット領域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．

上記の処理を本論文の提案手法とする．提案手法の大きな特徴は，トピックモデルを WSD に利用していることである．トピックモデルの構築には語義のラベル情報を必要としないために，領域適応の教師なし手法が実現される．トピックモデルを WSD に利用した従来の研究 (Li, Roth, and Sporleder 2010; Boyd-Graber, Blei, and Zhu 2007; Boyd-Graber and Blei 2007) はいくつかあるため，それらとの差異を述べておく．まずトピックモデルを WSD に利用するにしても，その利用法は様々であり確立された有効な手法が存在するわけではなく，ここで利用した手法も 1 つの提案と見なせる．また従来のトピックモデルを利用した WSD の研究では，語義識別の精度改善が目的であり，領域適応の教師なし手法に利用することを意図していない．そのためトピックモデルを構築する際に，もとになるコーパスに何を使えば有効かは深くは議論されていない．しかし領域適応ではソース領域のコーパスを単純に利用すると，精度低下を起こす可能性もあるため，本論文ではソース領域のコーパスを利用せず，ターゲット領域のコーパスのみを用いてトピックモデルを構築するアプローチをとることを明確にしている．この点が大きな差異である．

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った．単純に SVM を利用した手法と提案手法とをマクロ平均により比較した場合，OC をソースデータにして，PB をターゲットデータにした場合には有意水準 0.05 で，ソースデータとターゲットデータを逆にした場合には有意水準 0.10 で提案手法の有効性があることが分かった．

## 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．

$P_S(c) \neq P_T(c)$ 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 (Chan and Ng 2005, 2006)．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 (古宮，小谷，奥村 2013)．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 (Shimodaira 2000; 杉山 2006) 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 $w$ の語義が異なるケースは非常に稀であるため $P_S(c|x) = P_T(c|x)$ が仮定できる．$P_T(c|x)$ 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 (Jiang and Zhai 2007)．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 (齋木，高村，奥村 2008)．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](Rai, Saha, Daumé, and Venkatasubramanian 2010)．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 (Chan and Ng 2007)．

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 (古宮 他 2013)）．これは一般に負の転移現象 (Rosenstein, Marx, Kaelbling, and Dietterich 2005) 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．

また領域適応ではアンサンブル学習も有効な手法である．アンサンブル学習自体はかなり広い概念であり，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る．Daumé らは領域適応のための混合モデルを提案している (Daumé and Marcu 2006)．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 (Dai, Yang, Xue, and Yu 2007)．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 (Kamishima, Hamasaki, and Akaho 2009)．WSD の領域適応については古宮の一連の研究 (Komiya and Okumura 2012, 2011; 古宮，奥村 2012) 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．

[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．

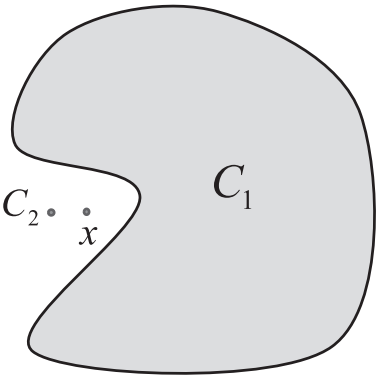



**図 1** 分布の影響が少ない k-NN

## 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 $x$ と最も近いデータ 1 つだけがクラス $c_2$ であれば，$x$ 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 $\theta$ 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 $\theta$ 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．なおこの値は予備実験等から得た最適な値ではないことを注記しておく．

### 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．つまり WSD での利用という観点では，シソーラスと単語クラスタリングの結果は同等である．そのため本論文においてシソーラスと述べている部分は，単語のクラスタリング結果を指している．

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) (Blei et al. 2003) 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 (Li et al. 2010; Boyd-Graber et al. 2007; Boyd-Graber and Blei 2007) があるが，ここではハードタグ (Cai, Lee, and Teh 2007) を利用する．ハードタグとは $w$ に対して最も関連度の高いトピック $z_{\hat{i}}$ 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 $x$ 中に $n$ 種類の単語 $w_1, w_2, \cdots, w_n$ が存在したとき，各 $w_j (j = 1 \sim n)$ 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

## 4.1 実験設定と実験結果

現代日本語書き言葉均衡コーパス（BCCWJ コーパス）(Maekawa 2007) 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．SemEval-2 の日本語 WSD タスク (Okumura, Shirai, Komiya, and Yokono 2010) では PB と OC を含む 4 ジャンルの語義タグ付きコーパスが公開されているので，語義のラベルはこのデータを利用する．

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．これらの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．注意として SemEval-2 の日本語 WSD タスクのデータを用いれば，更に異なった領域間の実験は可能であるが，領域間に共通してある程度の頻度で出現する多義語が少ないことなどから本論文では PB と OC 間の領域適応に限定している．

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

**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 |

[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 SemEval-2 の日本語 WSD タスクは新語義のタグも許しているからである．

**表 2** 各手法による正解率 (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 |
| マクロ平均 | 0.6553 | 0.7077 | 0.7090 | 0.7105 | 0.8453 |

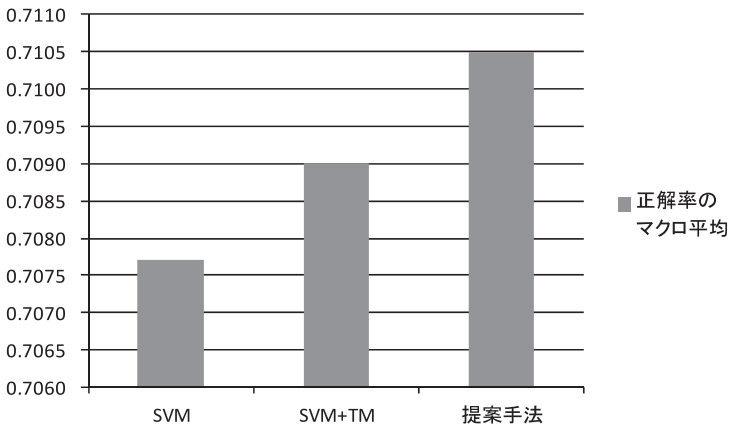



**図 2** 各手法による正解率のマクロ平均 (PB → OC)

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

**表 3** 各手法による正解率 (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 |
| マクロ平均 | 0.6961 | 0.7422 | 0.7520 | 0.7566 | 0.8258 |

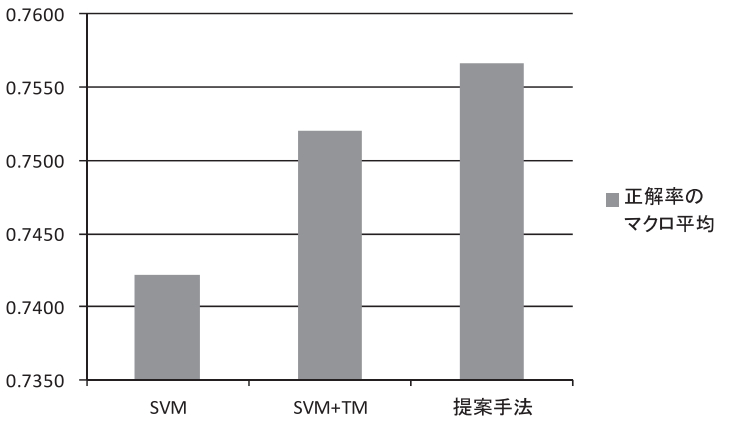



**図 3** 各手法による正解率のマクロ平均 (OC → PB)

ターゲット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．ただし一部の単語で「self」の値が「提案手法」などよりも低い．これはそれらの単語のソース領域のラベル付きデータの情報が，ターゲット領域で有効であったことを意味している．つまり「負の転移」が生じていないため，これらの単語については領域適応の問題が生じていないとも考えられる．

本実験の SVM の実行には，SVM ライブラリの libsvm[5]を利用した．そこで用いたカーネルは線形カーネルである．また識別の信頼度の算出には libsvm で提供されている -b オプションを利用した．このオプションは，基本的には，one vs. rest 法を利用して各カテゴリ（本実験の場合，語義）までの距離（識別関数値）の比較から，信頼度を算出している．識別結果は最も信頼度の高いカテゴリ（語義）となる．また BCCWJ コーパスは形態素解析済みの形で提供されているため，基本素性の単語や品詞は，形態素解析システムを利用せずに直接得ることができる．またトピックモデルの作成には LDA ツール[6]を用い，トピック数は全て 100 として実験を行った．

17 単語の正解率のマクロ平均をみると，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

なお本実験の評価はマクロ平均で行った．マイクロ平均による評価も可能ではあるが，本実験の場合，テストデータの用例数に幅がありすぎ，結果的にテストデータの用例数の多い単語の識別結果がマイクロ平均の値に大きく影響する．このためここではマクロ平均のみによる評価を行っている．マイクロ平均で評価した場合は，わずかではあるが SVM が最も高い評価値を出していた．

## 4.2 有意差の検定

t 検定を用いて各手法間の正解率のマクロ平均値の有意差を検定する．

対象単語 $w$ のソース領域でのラベル付きデータからランダムにその 9 割を取り出し，その 9 割のデータから前述した WSD の実験（「SVM」,「SVM + TM」,「提案手法」）を行う．この際，「提案手法」では k-NN の結果を用いるが，そこでも 9 割のデータしかないことに注意する．これを 1 セットの実験とし，50 セットの実験を行い，その正解率のマクロ平均を求めた．PB から OC への領域適応の結果を表 4 に示す．また OC から PB への領域適応の結果を表 5 に示す．

t 検定を行う場合，まず分散比の検定から 2 つのデータが等分散と見なせることを示す必要がある．自由度 (49,49) の F 値を調べることで，有意水準 0.10 で等分散を棄却するためには，分散比が 0.6222 以下か 1.6073 以上の値でなければならない．表 4 と表 5 から，各領域適応でどの手法間の組み合わせを行っても，正解率の分散が等しいことを棄却できないことは明らかであり，ここでは t 検定を行えると判断できる．

t 検定の片側検定を用いた場合，ここでの自由度は 48 なので有意水準 0.05 で有意差を出す t 値は 1.6772 以上，有意水準 0.10 で有意差を出す t 値は 1.2994 以上の値となる．このため有

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[6] http://chasen.org/~daiti-m/dist/lda/

**表 4** 9 割データでの実験結果 (PB → OC)

| 単語 | SVM | SVM+TM | 提案手法 |
|---|---|---|---|
| 言う | 0.8074 | 0.8188 | 0.8214 |
| 入れる | 0.7370 | 0.7658 | 0.7636 |
| 書く | 0.7374 | 0.7374 | 0.7374 |
| 聞く | 0.6556 | 0.6550 | 0.6711 |
| 来る | 0.7989 | 0.7989 | 0.7989 |
| 子供 | 0.2135 | 0.2177 | 0.2294 |
| 時間 | 0.8302 | 0.8302 | 0.8302 |
| 自分 | 0.8755 | 0.8753 | 0.8752 |
| 出る | 0.6647 | 0.6545 | 0.6544 |
| 取る | 0.1964 | 0.2656 | 0.2656 |
| 場合 | 0.8900 | 0.9084 | 0.9054 |
| 入る | 0.5903 | 0.5638 | 0.5632 |
| 前 | 0.8990 | 0.8722 | 0.8695 |
| 見る | 0.5803 | 0.5735 | 0.5735 |
| 持つ | 0.8568 | 0.7890 | 0.7900 |
| やる | 0.9350 | 0.9386 | 0.9386 |
| ゆく | 0.6801 | 0.6851 | 0.6860 |
| マクロ平均 | 0.7028 | 0.7029 | 0.7042 |
| 分散 | $0.2749 \cdot 10^{-4}$ | $0.2302 \cdot 10^{-4}$ | $0.2427 \cdot 10^{-4}$ |

**表 5** 9 割データでの実験結果 (OC → PB)

| 単語 | SVM | SVM+TM | 提案手法 |
|---|---|---|---|
| 言う | 0.8143 | 0.8139 | 0.8161 |
| 入れる | 0.6539 | 0.6554 | 0.6739 |
| 書く | 0.8394 | 0.8800 | 0.8690 |
| 聞く | 0.6681 | 0.6231 | 0.6315 |
| 来る | 0.9712 | 0.9708 | 0.9712 |
| 子供 | 0.3129 | 0.2903 | 0.2849 |
| 時間 | 0.8943 | 0.9019 | 0.9027 |
| 自分 | 0.9361 | 0.9639 | 0.9671 |
| 出る | 0.5634 | 0.5592 | 0.5592 |
| 取る | 0.2217 | 0.2353 | 0.2353 |
| 場合 | 0.8467 | 0.8467 | 0.8467 |
| 入る | 0.4753 | 0.4875 | 0.4875 |
| 前 | 0.8016 | 0.8110 | 0.8115 |
| 見る | 0.8355 | 0.8415 | 0.8415 |
| 持つ | 0.7933 | 0.7902 | 0.7902 |
| やる | 0.9295 | 0.9295 | 0.9295 |
| ゆく | 0.8926 | 0.8967 | 0.9035 |
| マクロ平均 | 0.7323 | 0.7351 | 0.7365 |
| 分散 | $0.2649 \cdot 10^{-4}$ | $0.2276 \cdot 10^{-4}$ | $0.2519 \cdot 10^{-4}$ |

**表 6** 手法間の有意差 (PB → OC)

| | SVM+TM と SVM | 提案手法 と SVM+TM | 提案手法 と SVM |
|---|---|---|---|
| t 値 | 0.0954 | 1.4124 | 1.4443 |
| 有意水準 0.05 | 有意差なし | 有意差なし | 有意差なし |
| 有意水準 0.10 | 有意差なし | **有意差あり** | **有意差あり** |

**表 7** 手法間の有意差 (OC → PB)

| | SVM+TM と SVM | 提案手法 と SVM+TM | 提案手法 と SVM |
|---|---|---|---|
| t 値 | 2.7535 | 1.4546 | 4.0890 |
| 有意水準 0.05 | **有意差あり** | 有意差なし | **有意差あり** |
| 有意水準 0.10 | **有意差あり** | **有意差あり** | **有意差あり** |

意差の検定結果は表 6 と表 7 のようにまとめられる.

結論的には提案手法と SVM との正解率のマクロ平均の差は OC から PB の領域適応では有意だが，PB から OC の領域適応では有意ではない．ただし有意水準を 0.10 に緩和した場合には，PB から OC の領域適応でも有意であると言える.

細かく手法を分けて調べた場合，トピックモデルを利用すること（SVM+TM と SVM の差）

と k-NN を併用すること（提案手法と SVM+TM の差）についての有意性はまちまちであった．ただし有意水準を 0.10 に緩和した場合，トピックモデルを利用する手法について PB から OC の領域適応以外の組み合わせについては全て有意性が認められた．

## 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 8 に示す．

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．

**表 8** 理想的語義分布の推定による識別

| 単語 | PB → OC<br>PB の事前分布 | PB → OC<br>OC の事前分布 | OC → PB<br>OC の事前分布 | OC → PB<br>PB の事前分布 |
|---|---|---|---|---|
| 言う | 0.7886 | 0.7901 | 0.7991 | 0.8027 |
| 入れる | 0.7297 | 0.7297 | 0.6140 | 0.6140 |
| 書く | 0.7300 | 0.7300 | 0.8889 | 0.8889 |
| 聞く | 0.6640 | 0.6560 | 0.7500 | 0.7500 |
| 来る | 0.7947 | 0.7947 | 0.9619 | 0.9619 |
| 子供 | 0.1282 | 0.1667 | 0.2447 | 0.2447 |
| 時間 | 0.8148 | 0.8148 | 0.8933 | 0.8933 |
| 自分 | 0.8682 | 0.8682 | 0.9709 | 0.9709 |
| 出る | 0.7046 | 0.7046 | 0.5556 | 0.5556 |
| 取る | 0.0968 | 0.2097 | 0.2195 | 0.2439 |
| 場合 | 0.9528 | 0.9685 | 0.8406 | 0.8406 |
| 入る | 0.6522 | 0.6522 | 0.4286 | 0.4454 |
| 前 | 0.9057 | 0.8962 | 0.6398 | 0.6398 |
| 見る | 0.5589 | 0.5589 | 0.8358 | 0.8321 |
| 持つ | 0.5714 | 0.5714 | 0.7597 | 0.7273 |
| やる | 0.9322 | 0.9322 | 0.9236 | 0.9236 |
| ゆく | 0.6818 | 0.6818 | 0.8657 | 0.8657 |
| マクロ平均 | 0.6809 | 0.6898 | 0.7172 | 0.7177 |

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らかである．ただしその具体的な利用方法は確立されていない．

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 (Cai et al. 2007)．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 (國井，新納，佐々木 2013)．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ら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 (森 2012)．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 9 に示す．

ターゲット領域が PB の場合，ソース領域の OC のコーパスを追加することで正解率は低下するが，ターゲット領域が OC の場合，ソース領域の 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．

**表 9** 両領域コーパスを利用した識別

| 単語 | PB → OC<br>OC の TM | PB → OC<br>OC+PB の TM | OC → PB<br>PB の TM | OC → PB<br>OC+PB の TM |
|---|---|---|---|---|
| 言う | 0.7958 | 0.8033 | 0.7953 | 0.8079 |
| 入れる | 0.7671 | 0.7808 | 0.7142 | 0.7679 |
| 書く | 0.7373 | 0.7374 | 0.8064 | 0.7903 |
| 聞く | 0.6612 | 0.6452 | 0.6829 | 0.6667 |
| 来る | 0.7989 | 0.7989 | 0.9711 | 0.9712 |
| 子供 | 0.2207 | 0.2338 | 0.4193 | 0.3548 |
| 時間 | 0.8301 | 0.8302 | 0.8918 | 0.8784 |
| 自分 | 0.8750 | 0.8750 | 0.9577 | 0.9416 |
| 出る | 0.7099 | 0.7023 | 0.6118 | 0.5855 |
| 取る | 0.2950 | 0.2623 | 0.2716 | 0.2716 |
| 場合 | 0.9127 | 0.9206 | 0.8467 | 0.8467 |
| 入る | 0.5735 | 0.6324 | 0.5254 | 0.4576 |
| 前 | 0.8857 | 0.9048 | 0.8312 | 0.8125 |
| 見る | 0.5954 | 0.6069 | 0.8424 | 0.8352 |
| 持つ | 0.7741 | 0.8387 | 0.7777 | 0.7909 |
| やる | 0.9401 | 0.9402 | 0.9294 | 0.9295 |
| ゆく | 0.6803 | 0.6849 | 0.9097 | 0.8947 |
| マクロ平均 | 0.7090 | 0.7175 | 0.7520 | 0.7414 |

## 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 での正解数を比較した．結果を表 10 と表 11 に示す．

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．

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．

**表 10** 識別結果の変更 (PB → OC)

| 単語 | 変更数 | k-NN | SVM+TM |
|---|---|---|---|
| 言う | 17 | 11 | 6 |
| 入れる | 0 | — | — |
| 書く | 0 | — | — |
| 聞く | 5 | 4 | 3 |
| 来る | 0 | — | — |
| 子供 | 10 | 5 | 5 |
| 時間 | 0 | — | — |
| 自分 | 0 | — | — |
| 出る | 0 | — | — |
| 取る | 0 | — | — |
| 場合 | 0 | — | — |
| 入る | 0 | — | — |
| 前 | 2 | 1 | 0 |
| 見る | 0 | — | — |
| 持つ | 0 | — | — |
| やる | 0 | — | — |
| ゆく | 4 | 2 | 2 |
| 総数 | 38 | 23 | 16 |

**表 11** 識別結果の変更 (OC → PB)

| 単語 | 変更数 | k-NN | SVM+TM |
|---|---|---|---|
| 言う | 63 | 37 | 37 |
| 入れる | 7 | 4 | 3 |
| 書く | 6 | 4 | 1 |
| 聞く | 9 | 5 | 4 |
| 来る | 1 | 0 | 0 |
| 子供 | 9 | 4 | 4 |
| 時間 | 0 | — | — |
| 自分 | 10 | 8 | 7 |
| 出る | 0 | — | — |
| 取る | 0 | — | — |
| 場合 | 0 | — | — |
| 入る | 0 | — | — |
| 前 | 0 | — | — |
| 見る | 0 | — | — |
| 持つ | 0 | — | — |
| やる | 0 | — | — |
| ゆく | 2 | 1 | 1 |
| 総数 | 107 | 63 | 57 |

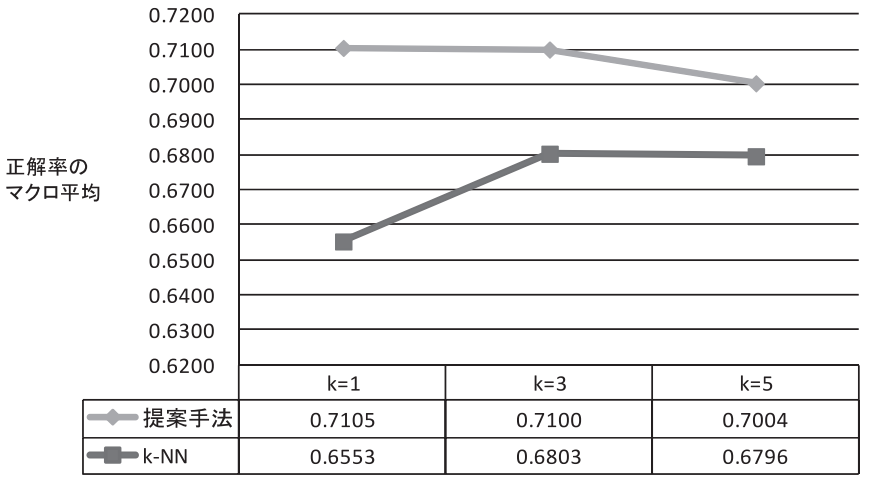
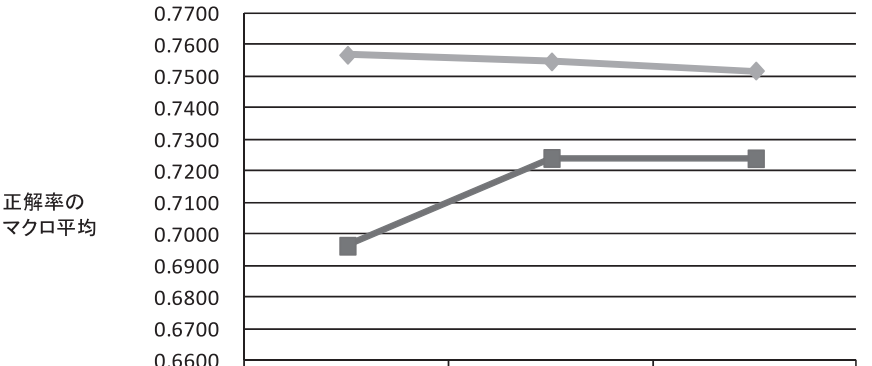



**図 4** k による変化 (PB → OC)




| | k=1 | k=3 | k=5 |
|---|---|---|---|
| 提案手法 | 0.7566 | 0.7545 | 0.7516 |
| k-NN | 0.6961 | 0.7239 | 0.7238 |

**図 5**　k による変化 (OC → PB)

## 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．ただし領域は OC と PB に限定しており，提案手法が他の領域間で有効であるかは確認できていない．この点は今後の課題である．ま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も今後の課題である．

## 参考文献

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent dirichlet allocation.” *Machine Learning Reseach*, **3**, pp. 993–1022.

Boyd-Graber, J. and Blei, D. (2007). “Putop: Turning Predominant Senses into a Topic Model for Word Sense Disambiguation.” In *Proceedings of SemEval-2007*.

Boyd-Graber, J., Blei, D., and Zhu, X. (2007). "A Topic Model for Word Sense Disambiguation." In *Proceedings of EMNLP-CoNLL-2007*, pp. 1024–1033.

Cai, J. F., Lee, W. S., and Teh, Y. W. (2007). "Improving Word Sense Disambiguation using Topic Features." In *Proceedings of EMNLP-CoNLL-2007*, pp. 1015–1023.

Chan, Y. S. and Ng, H. T. (2005). "Word Sense Disambiguation with Distribution Estimation." In *Proceedings of IJCAI-2005*, pp. 1010–1015.

Chan, Y. S. and Ng, H. T. (2006). "Estimating class priors in domain adaptation for word sense disambiguation." In *Proceedings of COLING-ACL-2006*, pp. 89–96.

Chan, Y. S. and Ng, H. T. (2007). "Domain adaptation with active learning for word sense disambiguation." In *Proceedings of ACL-2007*, pp. 49–56.

Chapelle, O., Schölkopf, B., Zien, A., et al. (2006). *Semi-supervised learning*, Vol. 2. MIT press Cambridge.

Dai, W., Yang, Q., Xue, G.-R., and Yu, Y. (2007). "Boosting for transfer learning." In *Proceedings of ICML-2007*, pp. 193–200.

Daumé, H. III (2007). "Frustratingly Easy Domain Adaptation." In *Proceedings of ACL-2007*, pp. 256–263.

Daumé, H. III and Marcu, D. (2006). "Domain adaptation for statistical classifiers." *Journal of Artificial Intelligence Research*, **26** (1), pp. 101–126.

Jiang, J. and Zhai, C. (2007). "Instance weighting for domain adaptation in NLP." In *Proceedings of ACL-2007*, pp. 264–271.

Kamishima, T., Hamasaki, M., and Akaho, S. (2009). "Trbagg: A simple transfer learning method and its application to personalization in collaborative tagging." In *Proceedings of the 9th IEEE International Conference on Data Mining*, pp. 219–228.

神嶌敏弘 (2010). 転移学習. 人工知能学会誌, **25** (4), pp. 572–580.

古宮嘉那子，奥村学 (2012). 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, **19** (3), pp. 143–166.

古宮嘉那子，小谷善行，奥村学 (2013). 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 回年次大会発表論文集, pp. C6–2.

Komiya, K. and Okumura, M. (2011). "Automatic Determination of a Domain Adaptation Method for Word Sense Disambiguation using Decision Tree Learning." In *Proceedings of IJCNLP-2011*, pp. 1107–1115.

Komiya, K. and Okumura, M. (2012). "Automatic Domain Adaptation for Word Sense Disambiguation Based on Comparison of Multiple Classifiers." In *Proceedings of PACLIC-2012*, pp. 75–85.

國井慎也，新納浩幸，佐々木稔 (2013). 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 回年次大会発表論文集, pp. P3–9.

Li, L., Roth, B., and Sporleder, C. (2010). "Topic Models for Word Sense Disambiguation and Token-based Idiom Detection." In *Proceedings of ACL-2010*, pp. 1138–1147.

Maekawa, K. (2007). "Design of a Balanced Corpus of Contemporary Written Japanese." In *Proceedings of the Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

森信介 (2012). 自然言語処理における分野適応. 人工知能学会誌, **27** (4), pp. 365–372.

Okumura, M., Shirai, K., Komiya, K., and Yokono, H. (2010). "SemEval-2010 Task: Japanese WSD." In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 69–74.

Rai, P., Saha, A., Daumé, H. III, and Venkatasubramanian, S. (2010). "Domain adaptation meets active learning." In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pp. 27–32.

Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. (2005). "To transfer or not to transfer." In *Proceedings of the NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*, Vol. 2, p. 7.

Settles, B. (2010). "Active Learning Literature Survey." Tech. rep., University of Wisconsin, Madison.

Shimodaira, H. (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of statistical planning and inference*, **90** (2), pp. 227–244.

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, **13** (3), pp. 111–118.

齋木陽介，高村大也，奥村学 (2008). 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, **2008** (33), pp. 61–67.

## 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 4 月茨城大学工学部システム工学科助手. 1997 年 10 月同学科講師，2001 年 4 月同学科助教授，現在，茨城大学工学部情報工学科准教授. 博士（工学）. 機械学習や統計的手法による自然言語処理の研究に従事. 言語処理学会，情報処理学会，人工知能学会 各会員.

**佐々木　稔**：1996 年徳島大学工学部知能情報工学科卒業. 2001 年同大学大学院

博士後期課程修了．博士（工学）．2001 年 12 月茨城大学工学部情報工学科助手．現在，茨城大学工学部情報工学科講師．機械学習や統計的手法による情報検索，自然言語処理等に関する研究に従事．言語処理学会，情報処理学会各会員．

（2013 年 5 月 25 日　受付）
（2013 年 7 月 30 日　再受付）
（2013 年 9 月 5 日　再々受付）
（2013 年 10 月 4 日　採録）