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

茨城大学工学部情報工学科
新納浩幸・佐々木稔

2013年5月23,24日
情報処理学会 NL 研

概要

WSD の教師なし領域適応手法の提案

WSD の領域適応の問題

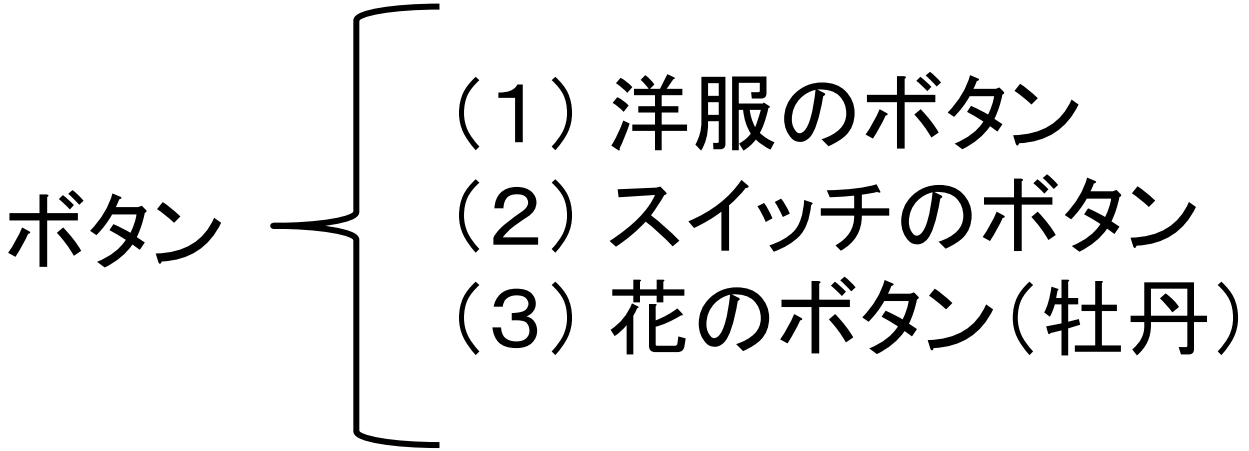
（1）語義分布が異なる

→ K-近傍法

（2）データスパースネス

→ トピックモデル

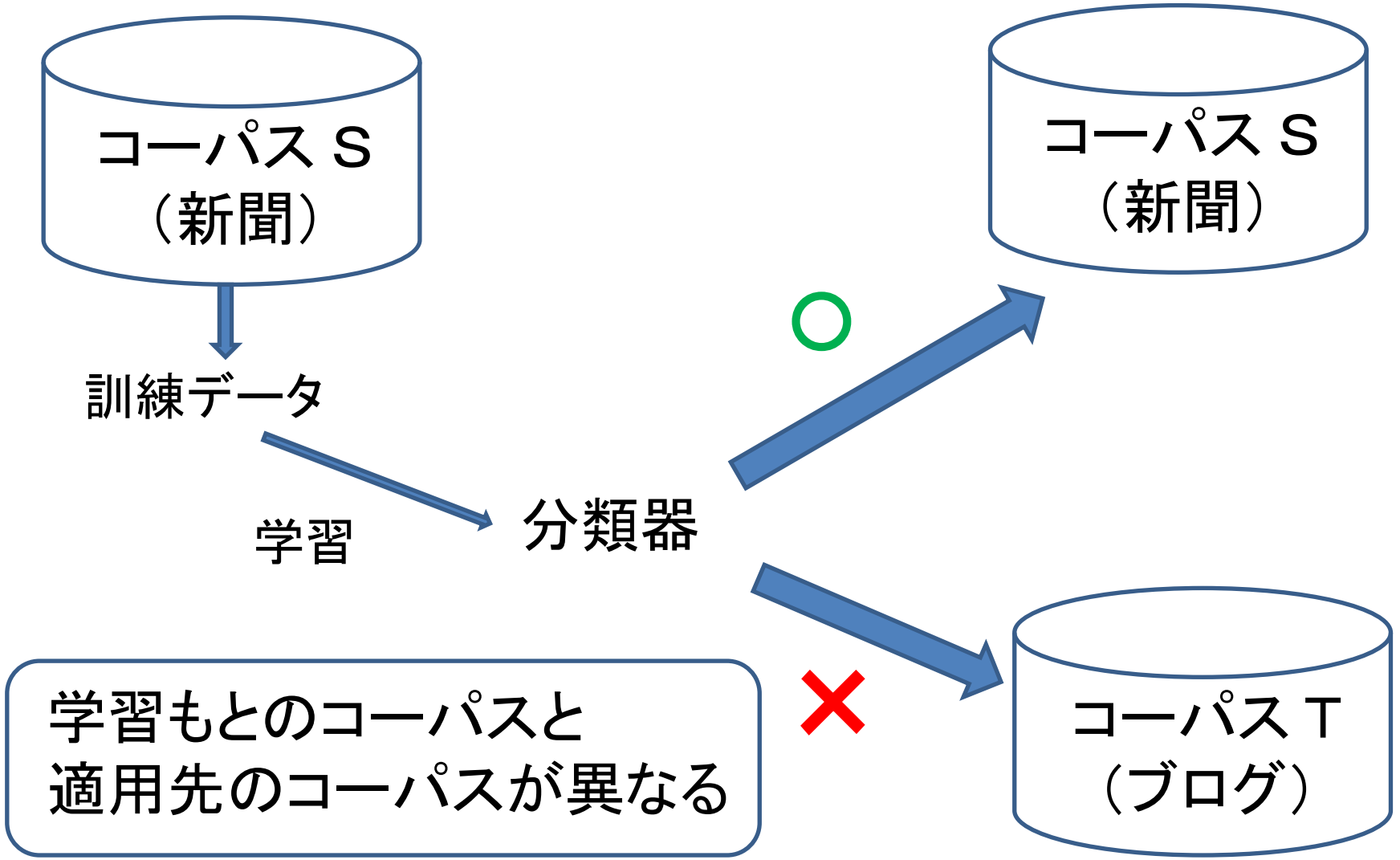
語義曖昧性解消
(WSD: word sense disambiguation)



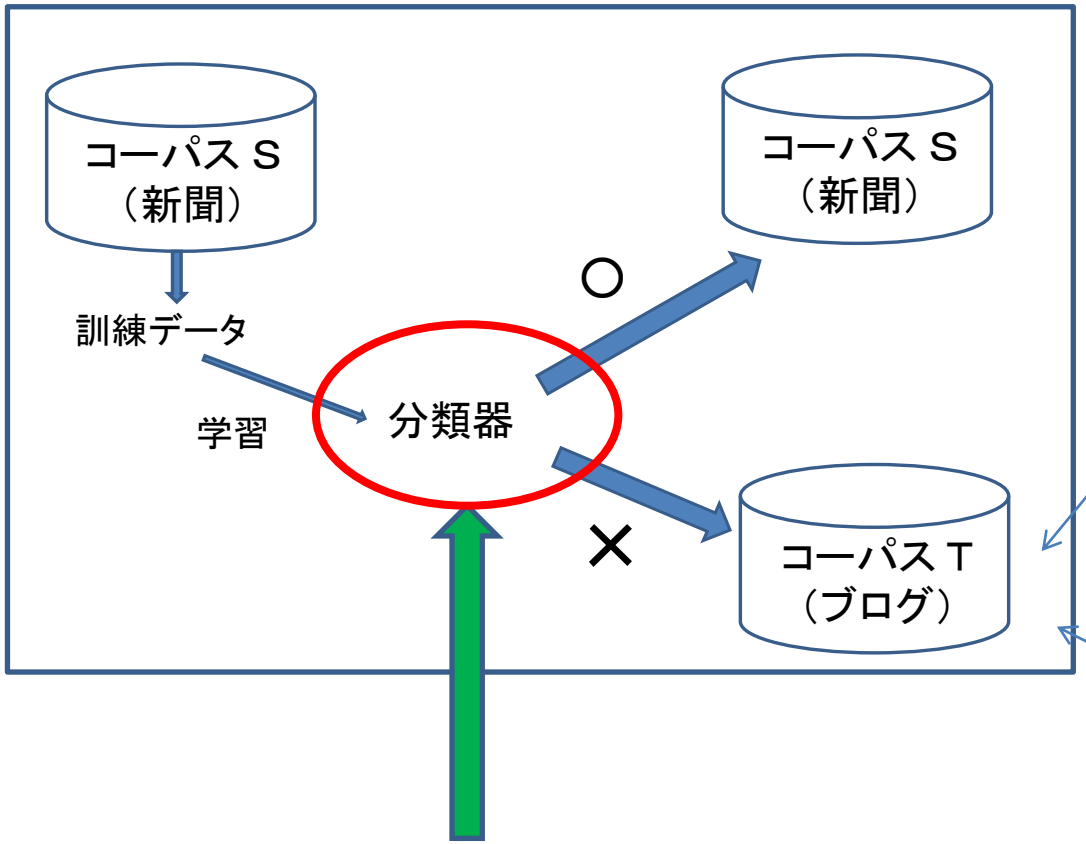
（問）シャツのボタンが取れた

（1）or（2）or（3）？

領域適応（DA: Domain Adaptation)



教師付き DA と教師なし DA



T の少量のラベル付き
データを利用する

教師付き DA

T のラベル付きデータを
全く利用しない

教師なし DA

これをターゲット領域 T に適応
させるのが領域適応（DA）、転移学習

事後確率最大化からの考察

x ：対象単語 w の用例

$C = \{c_1, c_2, \cdots, c_K\}$ ：w の語義の集合

$$\arg\max p(c \mid x) = \arg\max p(c)\, p(x \mid c)$$

領域適応の問題

（1） $p_S(c) \neq p_T(c)$

（2） $p_S(x \mid c) \neq p_T(x \mid c)$

正しい

本当？

$p_S(x|c) \neq p_T(x|c)$ ？

確認することは不可能、両方ともかなり小さな値

⇒ $p_S(x|c) = p_T(x|c)$ を仮定

WSD の領域適応の問題は、
ターゲット領域の語義の分布の推定

[Chan 06], [古宮 13]

EM アルゴリズム

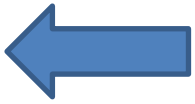
複数領域からのランダムサンプル

$p_S(x|c) = p_T(x|c)$ だとしても・・・

WSD の領域適応の問題は、
ターゲット領域の語義の分布の推定

だけではない

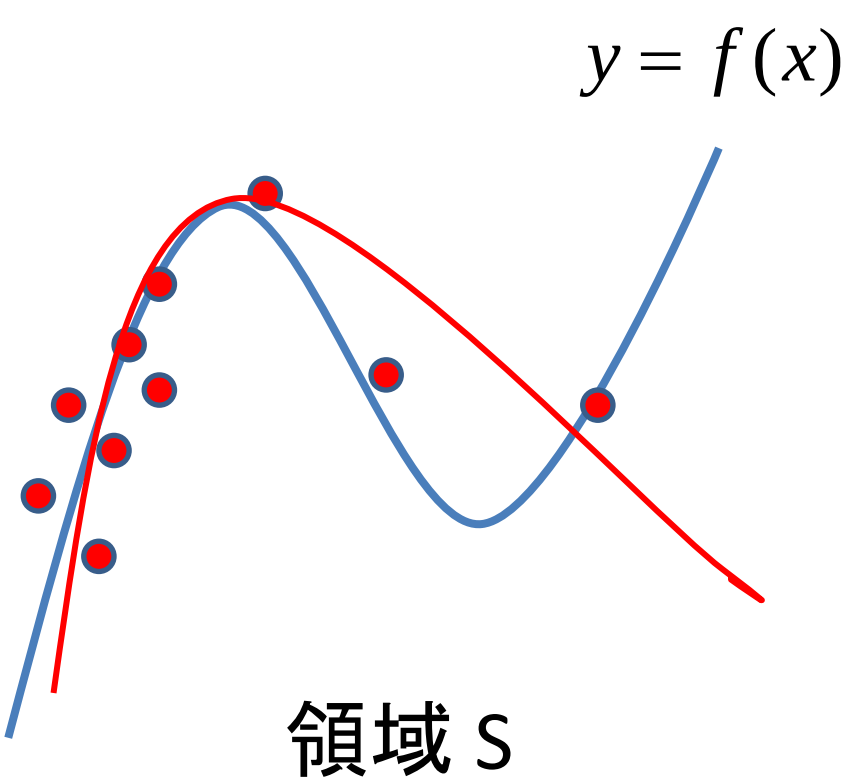
$p_T(x|c)$ の推定は困難

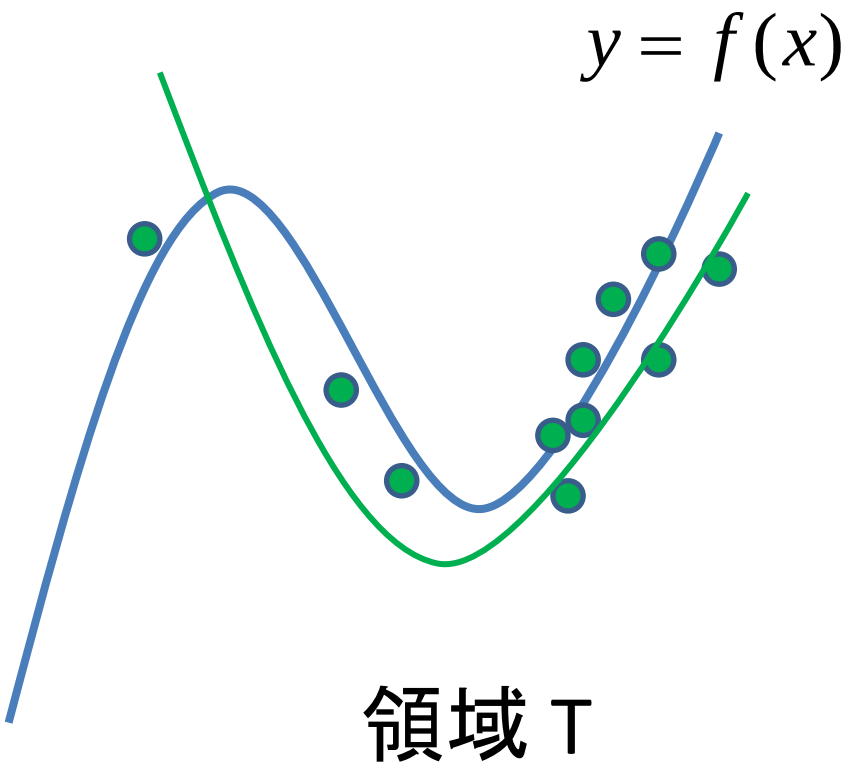


共変量シフトの問題と類似

共変量シフト（1）

$p_S(y|x) = p_T(y|x)$ だけど $p_S(x) \neq p_T(x)$

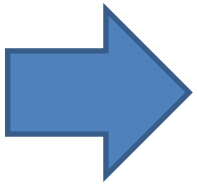




共変量シフト(2)

「シャツのボタンが取れた」

この文はいかなる領域で出現しようが、
ボタンの語義に変化はない



**WSD の領域適応の問題は
共変量シフトの問題である**

次回 NL 研

でも、共変量シフトの問題を教師なしではうまく解けない・・

スパース性

$p_S(x|c) = p_T(x|c)$　と考える

だけど　$p_S(x|c)$　の推定だけでは不十分

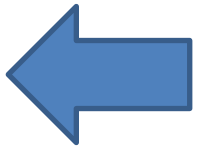
スパース性の問題

語義の分布の推定に悪影響を与えなければ、S のデータをいくら利用しても、負の転移は起きないはず

負の転移

領域適応のポイントはソース領域の
知識の利用方法

下手に使うと、パフォーマンスが下がる

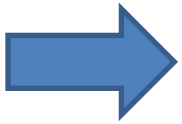


負の転移

WSD の領域適応では語義の分布の推定に
悪影響を与えなければ、負の転移は起きない

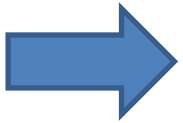
提案手法

（1）語義分布が異なる



K-近傍法

（2）データスパースネス



トピックモデル

K-近傍法の利用

入力 x と距離が近い順に訓練データから
K 個の点を取り出す
それら K 個のラベルの多数決で識別

語義の分布の影響が少ない

SVM で識別、信頼度が低い場合、
K-近傍法の結果を出力、K = 1

トピックモデルの利用

LDA ⇒ $p(w \mid z_i)$

⇒ 単語のソフトクラスタリング

ターゲット領域から LDA 学習

⇒ ターゲット領域に適したシソーラス

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この情報を追加

トピック素性

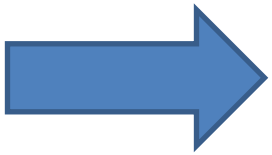
トピック数 K=100

100 次元のOベクトル t を用意

用例中の各単語 w に対して

$$\hat{i} = \arg\max\ p(w \mid z_i)$$

t の $\hat{i}$ 次元の値を 1 にする



トピック素性 t

通常の素性に連結して利用

利用した素性

w_0：対象単語

$$\cdots w_{-2} w_{-1} w_0 w_{+1} w_{+2} \cdots$$

w_{-1} の表記と品詞　　w_{+1} の表記と品詞

w_{-n} 自立語3つまで　w_{+n} 自立語3つまで

基本素性　　トピック素性

$$\mathbf{b} = (0,1,0,\cdots,0,0)$$

N 次元

$$\mathbf{t} = (1,0,0,\cdots,1,0)$$

100 次元

$$\mathbf{b} + \mathbf{t} = (0,1,0,\cdots,0,0,1,0,0,\cdots,0,0)$$

N+100 次元

実験設定

BCCWJ コーパス　OC（ Yahoo! 知恵袋）、PB（書籍）

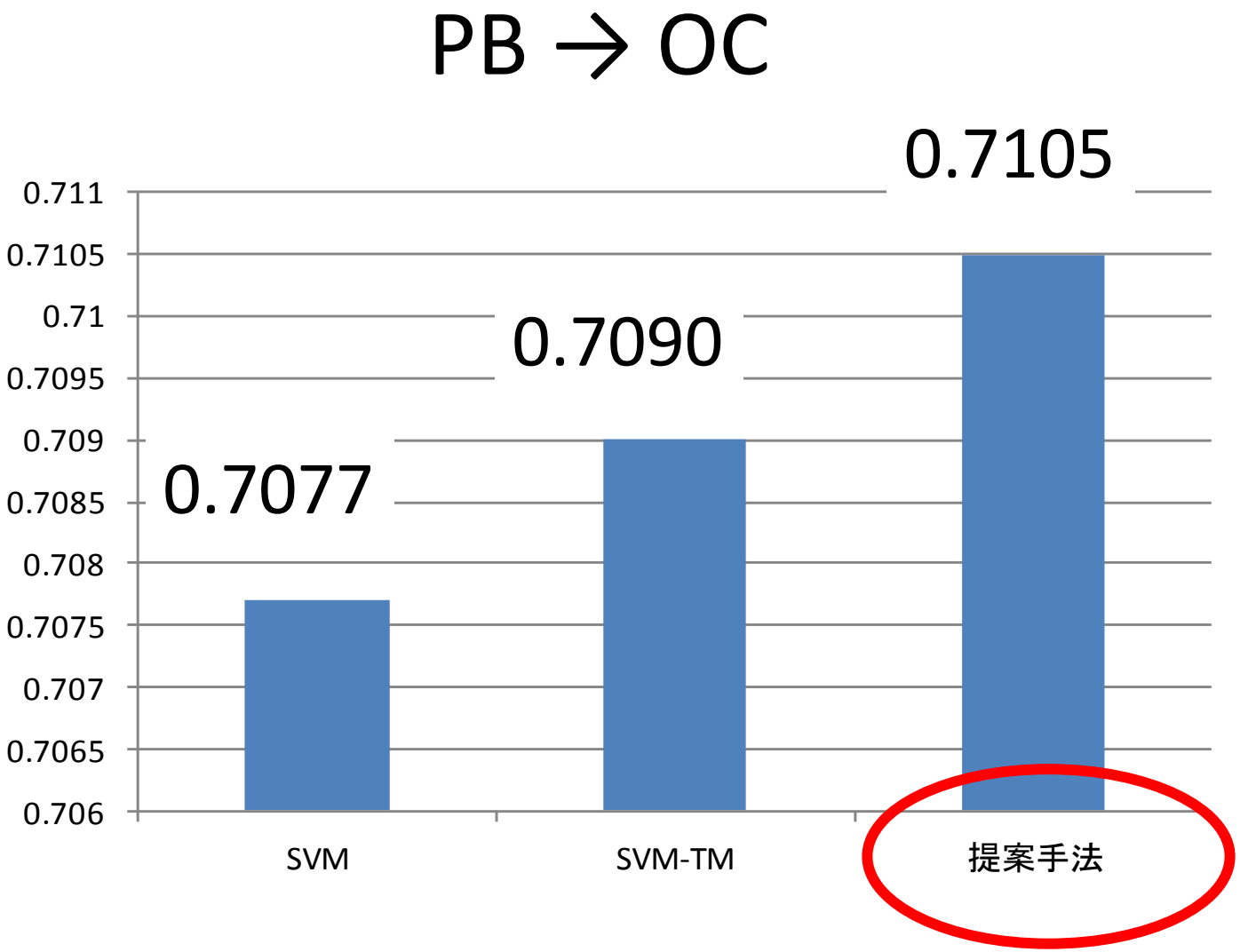
PB → OC　と　OC → PB

頻度 50 以上の 17 単語を対象

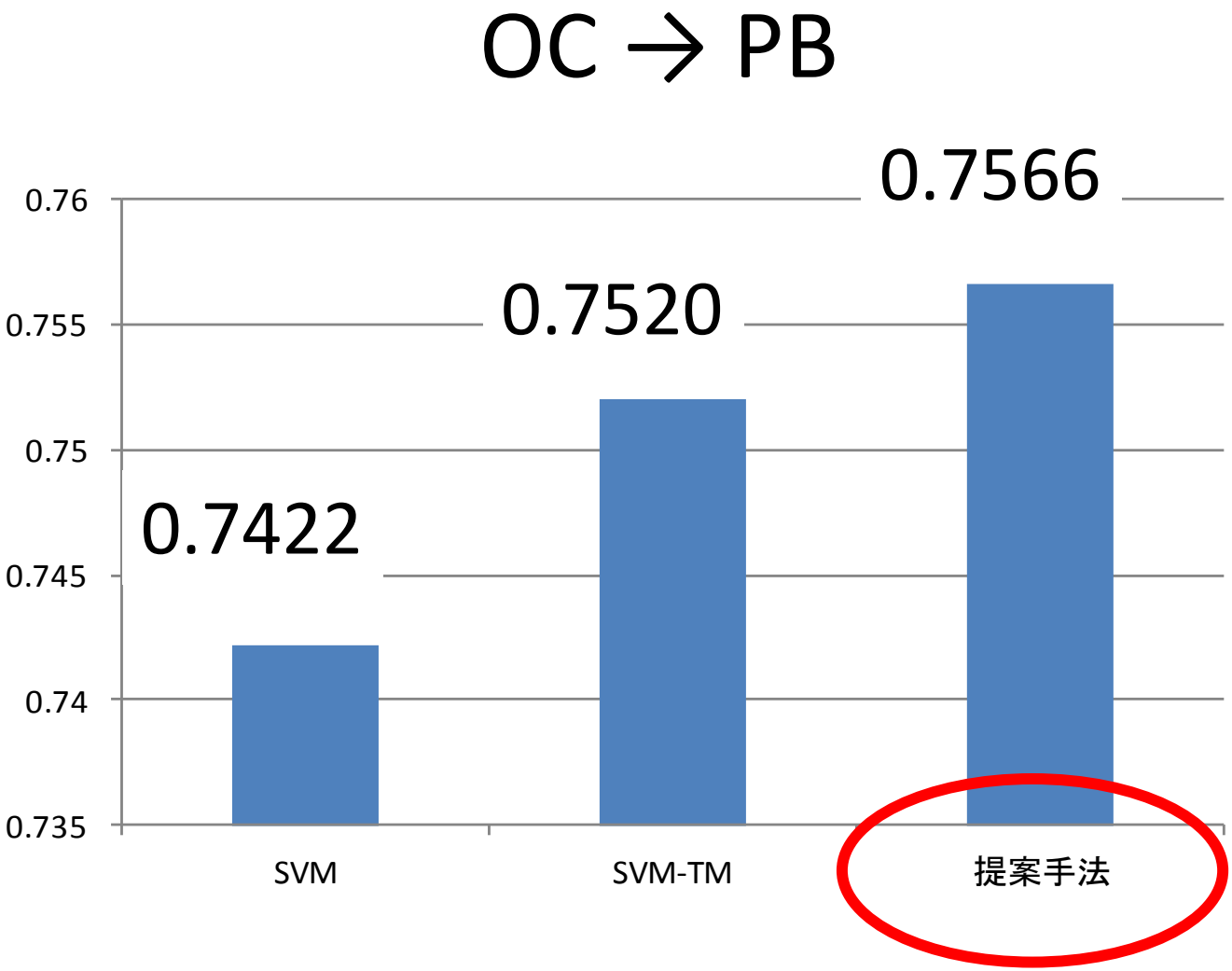
言う、入れる、書く、聞く、来る、
子供、時間、自分、出る、取る、
場合、入る、前、見る、持つ、
やる、ゆく

SVM を k-NN にする閾値は　1.1/K　(K = 語義数)

実験結果

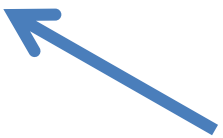

PB → OC
0.7105
0.7090
0.7077
0.711
0.7105
0.71
0.7095
0.709
0.7085
0.708
0.7075
0.707
0.7065
0.706
SVM
SVM-TM
提案手法

実験結果


OC → PB
0.76
0.755
0.75
0.745
0.74
0.735
0.7422
0.7520
0.7566
SVM
SVM-TM
提案手法

語義分布の推定だけでは不十分

$$\arg\max\ p_S(c)\,p_S(x|c)$$



$p_T(c)$ に変更して実験

				SVM
PB → OC	0.6809	⇒	0.6898	(0.7077)
OC → PB	0.7172	⇒	0.7177	(0.7422)

少しは効果あるが、とても小さい

トピックモデルの利用法

本手法

（1）T のトピックモデル

（2）S のトピックモデル

better ?

（3）T＋S のトピックモデル

	SVM+TM(1)		SVM+TM(3)	
PB → OC	0.7077	⇒	0.7175	○
OC → PB	0.7422	⇒	0.7414	×

負の転移

トピック素性の表現方法

$$w \Rightarrow \mathbf{t} = (p(w|z_1), p(w|z_2), \cdots, p(w|z_{100}))$$

ソフトタグ

$$\mathbf{t} = (0,0,\cdots 0,1,0,\cdots,0)$$

ハードタグ

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

$$\mathbf{t} = (0,0.1,\cdots 0.3,0.4,0,\cdots,0.1)$$

ミドルソフトタグ

最適な次元数や表現方法は不明

K-近傍法の効果

正解率

PB → OC

38 用例対象

SVM-TM	K-NN
0.3906	0.6144

OC → PB

107 用例対象

SVM-TM	K-NN
0.3635	0.4413

アンサンブル学習

本手法は SVM と k-近傍法 のアンサンブル

アンサンブル学習はかなり広い概念

バギング、ブースティング、混合分布・・・

領域適応にはよく利用されている、有望そう

[Komiya 11], [Komiya 12] [古宮 12]

WSD の領域適応では、対象単語毎に問題の性質が異なる、単一の手法で解決が難しいことも影響かも・・・

まとめ

- WSD の教師なし領域適応の手法を提案
- WSD の領域適応の問題を2つに要約
  (1) 語義分布の違い
  (2) データスパース
- (1) に対して k-近傍法の補助的利用、
  (2) に対してトピックモデルの利用
- BCCWJ コーパスの OC, PB 17単語で実験、
  提案手法の有効性を示した
- ソース領域のトピックモデルの利用法、
  アンサンブル学習法の検討が課題