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

古宮 嘉那子[1]　　小谷 善行[1]　　奥村 学[2]
[1]東京農工大学　工学研究院，[2]東京工業大学　精密工学研究所
{kkomiya, kotani}@cc.tuat.ac.jp, oku@pi.titech.ac.jp

1　はじめに

テストのターゲットとなるドメインとは異なるドメインのデータ（ソースデータ）を利用して学習を行い，ターゲットのドメインのデータ（ターゲットデータ）に適応することを領域適応といい，近年さまざまな手法が研究されている．

本稿では，あるドメイン（ジャンル）のターゲットデータに対して複数のジャンルのソースデータが混在している場合，ソースデータの全体集合から，ターゲットデータに適した訓練事例の部分集合を自動的に選択する試みについて述べる．なお，ターゲットデータのラベルは未知とし，語義曖昧性解消（Word Sense Disambiguation，WSD）について領域適応を行った．確信度とLOO-boundという指標を利用して訓練事例集合を自動的に選んだところ，入手可能なコーパスを全て使うよりもWSDの正解率が上昇した．

2　関連研究

領域適応は，学習に使用する情報により，supervised，semi-supervised，unsupervisedの三種に分けられる．まずsupervisedの領域適応は，多量なラベルつきのソースデータに加え，少量のラベルつきのターゲットデータを用いて学習を行うもので，訓練事例としてソースデータまたは少量のターゲットデータだけを利用する場合よりも，分類器を改良することを目指す．次のsemi-supervisedの領域適応は，多量なラベルつきのソースデータに加え，多量なラベルなしのターゲットデータを利用し，訓練事例としてソースデータだけを利用する場合よりも，分類器を改良することを目指す．また，最後のunsupervisedの領域適応は，ラベルつきのソースデータで学習後，ターゲットデータで実行する．本研究で扱うのは，semi-supervisedの領域適応である．

[6]はWSDについてsupervisedな領域適応を行った場合，最も効果的な領域適応手法はソースデータとターゲットデータの性質により異なることを示し，最も効果的な領域適応手法を，WSDの対象単語タイプ，ソースデータ，ターゲットデータの三つ組ごとに自動的に選択する手法について述べた．また，[3]は，WSDのsupervisedな領域適応において，本稿でも使用する確信度という尺度を用い，用例ごとに訓練事例を自動的に選択した．最後に，[7]は[3]の手法がsemi-supervisedの領域適応に対しても有効であることを示している．

3　訓練事例の部分集合の選択

あるドメイン（ジャンル）のターゲットデータを対象にWSDを行うことを考える．このターゲットデータのラベル（語義）は未知であるとする．一方，複数のジャンルのソースデータが入手可能であるとすると，本稿ではこれらのソースデータの全体集合から，ターゲットデータに適した訓練事例の部分集合を自動的に選択する．この際，以下の手順で訓練事例の部分集合の選択を行う．なお，WSDの対象単語タイプごとに分類器を作成するため，訓練事例集合の選択は単語のタイプごとに行う．

(1) ソースデータの全体集合から訓練事例をランダムに選択して，訓練事例集合を複数個作成する．

(2) それぞれの訓練事例集合で分類器を学習し，ターゲットデータに適用する．

(3) 分類器が出力する値をもとに分類器ごとにスコアを計算する．

(4) スコアの最も高い分類器を作成した訓練事例集合を選択する．

ここでの分類器のスコアには，以下の三つを利用し，比較する．

- 確信度 [3].
- LOO-bound[5] を基にしたスコア.
- 上記二つのスコアを掛け合わせた値.

ここで，確信度とは分類の確からしさを表す 0〜1 の値であり，active-learning においてラベル付けする用例を選択するのによく利用される．確信度は用例ごとに得られるため，ターゲットデータの全事例の確信度の平均を使用した．そのため，確信度によるスコアは「ある分類器があるターゲットデータ全体に対して，平均的にどのくらい自信をもって分類しているか」を表す.

また，LOO-bound は SVM に対し Leave-One-Out Estimation を行ったときのエラーの期待値の上限であり, 以下の式で表される.

$$LOO - bound = \frac{\text{サポートベクターの数}}{\text{訓練事例の数}} \quad (1)$$

しかしこの値はエラー率であるため，分類器のスコアとする際に，全体を 1 から引いた．また領域適応においては訓練事例中の語義がひとつになる場合があり，そのようなときには訓練事例の数が 0 となる．このような事態を避けるため，最終的には以下のような変形を行って使用した.

$$LOO - bound \text{を基にしたスコア} = 1 - \frac{\text{サポートベクターの数} + 0.5}{\text{訓練事例の数} + 0.5} \quad (2)$$

4 実験

分類器としてはマルチクラス対応の SVM（libsvm）[1] を使用した．また，libsvm の確率として出力される分類の確からしさを確信度として用いた．カーネルは予備実験の結果，線形カーネルが最も高い正解率を示したため，これを採用した．また，WSD の素性には，以下の 17 種類の素性を用いた.

- WSD の対象単語の前後二語までの形態素の表記（4 種類）
- WSD の対象単語の前後二語までの品詞（4 種類）
- WSD の対象単語の前後二語までの品詞の細分類（4 種類）
- WSD の対象単語の前後二語までの分類コード（4 種類）
- 係り受け（1 種類）

表 1: それぞれの領域における単語ごとの最小，最大，平均用例数

コーパスの種類	最小	最多	平均
BCCWJ 白書	58	7610	2240.14
BCCWJ Yahoo！知恵袋	130	13976	2741.95
RWC　新聞	56	374	183.36

  - 対象単語が名詞の場合はその名詞が係る動詞
  - 対象単語が動詞の場合はその動詞のヲ格の格要素

分類語彙表の分類コードには [8] を使用した.

また，実験は五分割交差検定を用いた.

訓練事例集合は 1 単語タイプにつき，100 個作成した．また，それぞれの訓練事例集合の用例数は，WSD の対象単語タイプごとに，1 件から入手可能な全用例数までのうちからランダムに選択した．ランダム性が高いので，実験は 10 回行い，その平均の正解率を求めた.

5 実験データ

実験には，現代日本語書き言葉均衡コーパス（BCCWJ コーパス）[4] の白書のデータと Yahoo! 知恵袋のデータ，また RWC コーパスの毎日新聞コーパス [2] の三つのデータを利用した．これらのデータには岩波国語辞典 [9] の語義が付与されている．三つのコーパスのうち，ひとつをターゲットデータにし，残りの二つを利用可能なソースデータとして利用することで，全部で 3 通りの領域適応を行った．これらのコーパス中の多義語のうち，三つのコーパス中全てに 50 トークン以上存在する単語を実験対象としたところ，全体で 22 種類となった.

それぞれのコーパスにおける WSD の対象単語タイプごとの最小，最大，平均用例数を表 1 に示す.

また，実験には岩波国語辞典の小分類の語義を採用した．語義数ごとの単語の内訳は，2 語義 :「場合」,「自分」，3 語義 :「事業」,「情報」,「地方」,「社会」,「思う」,「子供」，4 語義 :「考える」，5 語義 :「含む」,「技術」，6 語義 :「関係」,「時間」,「一般」,「現在」,「作る」，7 語義 :「今」，8 語義 :「前」，10 語義 :「持つ」，12 語義 :「見る」，14 語義 :「入る」，16 語義 :「言う」，22 語義 :「手」である.

表 2: 全体の適応手法別の実験結果

手法	マイクロ平均	マクロ平均
Self	93.29%	85.97%
平均的なコーパス	76.92%	71.20%
大きい方のコーパス	81.99%	74.25%
全てのコーパス	81.76%	75.86%
確信度	73.20%	70.65%
確信度× LOO-bound	**82.10%**	**76.37%**
LOO-bound	81.92%	76.35%

6 結果

表 2 に全体の適応手法別の実験結果を，表 3 にコーパスと適応手法別の実験結果を示す.

これらの表において,「Self」は，タグつきターゲットデータが手に入ったと仮定して，supervised の学習を 5 分割交差検定を用いて行った結果である.

「平均的なコーパス」は，ふたつのジャンルのソースデータそれぞれをジャンルごとに分けて訓練事例とした場合の結果の平均である．入手可能なコーパスをそれぞれソースデータとして使用した場合の平均的な結果を示している．例えば，Yahoo！知恵袋のデータがターゲットデータの時のソースデータは白書と新聞であるが，このときの「平均的なコーパス」は，白書の全データで訓練した Yahoo！知恵袋のデータの正解率と，新聞の全データで訓練した Yahoo！知恵袋のデータの正解率の平均となる.

また,「大きい方のコーパス」は，ふたつのジャンルのソースデータのうち，用例数が多いジャンルのソースデータをすべて訓練事例とした場合の結果である．例えば，Yahoo！知恵袋のデータがターゲットデータの時の「大きい方のコーパス」は，白書よりも新聞のほうが全単語タイプで比較したときに用例数が多かったため，新聞の全データで訓練した Yahoo！知恵袋のデータの正解率の平均となる.

最後に,「全てのコーパス」とは，ふたつのジャンルのソースデータ全て（つまり全ソースデータ）を訓練事例とした際の結果である．例えば、Yahoo！知恵袋のデータがターゲットデータの時の「全てのコーパス」は，白書と新聞のコーパス全てを訓練事例として利用した際の結果である.

このとき,「Self」は upper bound であり,「平均的なコーパス」,「大きい方のコーパス」,「全てのコーパス」はベースラインである．表において Self 以外でコーパスごとに一番高い正解率を太字で示した.

7 考察

まず，表 2 と表 3 においてマイクロ平均を比べると，Yahoo！知恵袋コーパスがターゲットデータの時と全体で比較した際には,「全てのコーパス」の正解率より「大きい方のコーパス」の正解率の方が高い．このことから，訓練事例は必ずしも多ければ良いわけではないことが分かる.

次に，同じ二つの表から,「確信度」は三つのベースラインより正解率が低いことが分かる．この結果は,「確信度」により用例ごとに訓練事例を選択するのが有効であるとしている [7] や [3] の結果と対照的である．この原因として，本実験ではランダムに訓練集合の事例数を決めているため，訓練事例数がとても小さな数になることがあり，これが確信度の正確さを低めていることが考えられる．たとえば，訓練集合の用例数が 1 件になると，他に選択肢がないため，その分類器の確信度は最高値の 1 となる．しかし 1 件の訓練事例による分類器が最も良い分類器であるとは考え難く，訓練事例数の少ない場合の確信度はあまり信用できないことが分かる．また，本実験では全ターゲットデータの事例の確信度の平均を使用しているため，平均することで確信度の正確性が鈍った可能性がある.

また，表 2 と表 3 から,「LOO-bound」は全体のマイクロ平均と Yahoo！知恵袋コーパスがターゲットデータの時のマイクロ平均以外は，三つのベースラインの正解率を上回っていることが分かる.「LOO-bound」はその分類器自体がどれだけ信用できるかを表す値であるため，分類器の選択に有用であると考えられる．しかし，Yahoo！知恵袋コーパスがターゲットデータの時の「大きい方のコーパス」の正解率のマイクロ平均の方が高いため，全体のマイクロ平均では「大きい方のコーパス」を超えることが出来なかった.

これに対して「確信度× LOO-bound」は，Yahoo！知恵袋コーパスがターゲットデータの時のマイクロ平均は「大きい方のコーパス」に敵わないものの，全体でいえばマイクロ平均，マクロ平均両方において，Self を抜かせば最も高い正解率を示した．この結果は「大きい方のコーパス」に対しては有意ではなかったものの,「全部のコーパス」に対してはカイ二乗検定により 0.05 水準で有意な差があった.「確信度× LOO-bound」が良かったのは,「LOO-bound」はどんなターゲットデータに対しても同じ値を返すが,「確信度」は訓練事例とターゲットデータの各事例の組に対して値が決まるため，よりターゲットデータに合った分類器を選択したためだと考えられる.

最後に，サンプル数が少なくなるため有意ではな

表 3: コーパスと適応手法別の実験結果

	マイクロ平均			マクロ平均		
ターゲットデータ	白書	新聞	Yahoo！知恵袋	白書	新聞	Yahoo！知恵袋
Self	96.07%	79.57%	91.93%	91.53%	78.59%	87.80%
平均的なコーパス	73.54%	72.94%	79.95%	70.80%	71.23%	71.57%
大きい方のコーパス	80.72%	74.86%	**83.50%**	75.64%	74.39%	72.73%
全てのコーパス	81.80%	75.95%	82.11%	76.91%	74.91%	75.76%
確信度	71.95%	73.45%	74.19%	69.29%	72.00%	70.67%
確信度× LOO-bound	**82.39%**	**76.12%**	82.27%	**77.58%**	**75.21%**	76.32%
LOO-bound	82.13%	76.04%	82.19%	77.45%	75.18%	**76.37%**

かったが,「確信度× LOO-bound」と「LOO-bound」はマクロ平均においてベースラインより高い正解率を示した．このことから，少ない事例のターゲットデータに関してもこれらの指標がより良い訓練事例集合を選択することが分かる．

8 おわりに

本稿では，semi-supervised な領域適応において，あるターゲットデータに対して複数のジャンルのソースデータが混在した場合，確信度と LOO-bound という指標を利用して，領域適応のための訓練事例の部分集合を WSD の対象単語タイプごとに自動的に選択する手法について述べた.「確信度」,「確信度× LOO-bound」,「LOO-bound」を指標としてよりよい訓練事例集合を選択したところ，マイクロ平均，マクロ平均ともに「確信度× LOO-bound」が最高の正解率を示した．また，入手可能なふたつのジャンルのコーパスのうち大きい方のコーパスを訓練事例にした場合との差は有意ではなかったが，全てのソースデータを訓練事例にする場合との差はカイ二乗検定で有意であった.

また，本稿では，ソースデータの全体集合からランダムに選んだ 100 個の訓練事例集合のうち，より良いものを選択しているが，今後，本稿の手法で選択した訓練事例集合に事例を足したり引いたりして，最終的に最適なものに近付けることを目指す予定である．

謝辞

文部科学省科学研究費補助金 [若手 B（No : 24700138）] の助成により行われた．ここに，謹んで御礼申し上げる．

参考文献

[1] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] Koichi Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, and Wakako Kashino. The rwc text databases. In *LREC 1998*, pp. 457–461, 1998.

[3] Kanako Komiya and Manabu Okumura. Automatic domain adaptation for word sense disambiguation based on comparison of multiple classifiers. In *PACLIC 2012*, pp. 77–85, 2012.

[4] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *ALR 2008*, pp. 101–102, 2008.

[5] V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Comput.*, Vol. 12, No. 9, pp. 2013–2036, 2000.

[6] 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.

[7] 古宮嘉那子，奥村学，小谷善行． 分類器の確信度を用いた合議制による語義曖昧性解消の semi-supervised な領域適応．第三回コーパス日本語学ワークショップ予稿集, In Press, 2013.

[8] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[9] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典 第五版. 岩波書店, 1994.