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領域適応

古宮 嘉那子　　奥村 学
東京農工大学　工学研究院，東京工業大学　精密工学研究所
komiya@cc.tuat.ac.jp, oku@pi.titech.ac.jp

1 はじめに

通常，機械学習とは，新聞データを用いて新聞用の分類器を学習するなど，ドメイン A のデータを用いてドメイン A 用の分類器を学習するものであった．しかし一方，ドメイン B についての分類器を学習したいのに，ドメイン A のデータにしかラベルがついていないことがあり得る．このとき，ドメイン A（ソースドメイン）のデータによって分類器を学習し，ドメイン B（ターゲットドメイン）のデータに適応することを考える．これが領域適応であり，さまざまな手法が研究されている．図 1 はソースドメインを新聞，ターゲットドメインを小説にした際の領域適応の様子を示している．

語義曖昧性解消（WSD: Word Sense Disambiguation）の領域適応の手法はさまざまあるが，我々は用例によって適切な手法は異なると考えた．本稿では，少量のターゲットデータにラベル付けして学習を行う方式と，他のコーパスを訓練事例に加える方式を使って二つの分類器を学習し，学習された分類器の出力する確信度の高い方の答えを採用することにより，分類の精度を向上させる手法を示す.

本稿の構成は以下のようになっている．まず 2 章で領域適応の関連研究について紹介する．3 章では用例ごとの領域適応手法の自動選択について説明し，4 章では本研究で用いた領域適応手法とデータについて述べる．5 章に結果を，6 章に考察を，7 章にまとめを述べる．

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，supervised の領域適応である．

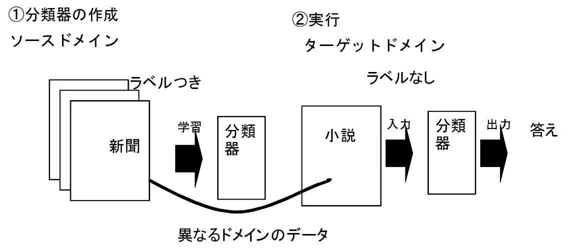

図 1: 領域適応時の機械学習

領域適応の研究は自然言語処理の分野の内外においてさまざまなされており，supervised のものには [2], [4]，[6] などがある．

また，共学習を用いた適応に関する研究に [8] がある．[8] は co-training において適応を行った co-adaptation の研究である．boosting による線形補完により適応を行い，両方の分類器においてエラー率が低下したことを報告している．

本稿では，分類器の確信度により領域適応に用いる手法を選択する手法について述べる．これに関連した研究として [9] や [1]，[10] がある．[9] は，構文解析において，分野間距離をはかり，より適切なコーパスを利用して領域適応を行えるようにした．また，[1] は，構文解析において，自動的にタグ付けされたコーパスを用いて，ソースデータとターゲットデータの類似度から性能を予測できることを示した．これらの研究で

は，領域間の距離からソースデータとして利用できるコーパスを選択するという立場をとっているが，[10] はソースデータとターゲットデータの性質から，適切な領域適応手法を自動選択するという立場をとった．本研究では，分類器の確信度から，用例ごとに手法を選択する．

3 用例ごとの領域適応手法の自動選択

[10] において，我々は WSD のための領域適応において，ターゲットデータやソースデータの性質により，ソースデータ/ターゲットデータ/単語の組み合わせごとに最も効果的な領域適応手法が異なることを示した．本稿では，ソースデータ/ターゲットデータ/単語の組み合わせだけではなく，一例一例，用例ごとに効果的な手法が異なると仮定する．そのため，以下のように用例ごとに領域適応の手法を選択する．

(1) 複数の手法により分類器を学習する．

(2) 用例ごとに，複数の手法による分類器の確信度を比較する．

(3) 分類器の確信度の最も高い手法による結果を採用する．

ここでの分類器の確信度は，分類の確からしさの度合いの予測値であり，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．本手法ではこの確信度が確率として出力されることに注目し，確信度を比較することで，複数の分類器の合議を行う．

4 実験

4.1 WSD のための領域適応手法

WSD のための領域適応手法として，本研究では以下に示す二つ (Target Only，Random Sampling) を用いる．

- Target Only : ソースデータを用いず，ランダムに選んだ少量のターゲットデータにラベル付けしたものだけを訓練事例にする．
- Random Sampling : ランダムに選んだ少量のターゲットデータの用例にラベル付けしたものとソースデータの両方を訓練事例にする．

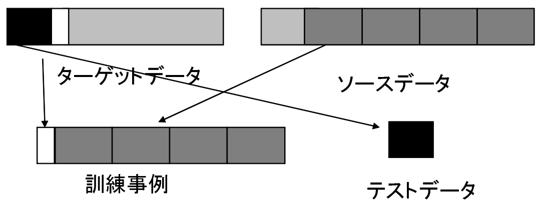

図 2: 領域適応の五分割交差検定

なお，使用するターゲットデータは常に 10 件とした．

分類器としてはマルチクラス対応の SVM（libsvm）[3] 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．本実験では，分類器を二つ学習したため，合議の際には二つのうちより高い確信度である分類器の結果を採用する．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，学習の素性には，以下の 17 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 素性）
- WSD の対象単語の前後二語までの品詞（4 素性）
- WSD の対象単語の前後二語までの品詞の細分類（4 素性）
- WSD の対象単語の前後二語までの分類コード（4 素性）
- 係り受け（1 素性）
  - 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 [11] を使用した．

また，実験は五分割交差検定を用いた．Random Sampling の場合には，ソースデータの 4/5（ソースデータの濃い灰色の部分）に加え，ターゲットデータの 4/5（ターゲットデータの白の部分と薄い灰色の部分）から 10 件（白い部分）を訓練事例とする．テストデータは，ターゲットデータの残りの 1/5（黒い部分）である．この様子を図 2 に示す．

4.2 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）[7] の白書のデータと Yahoo! 知恵袋

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	2074.50
BCCWJ Yahoo！知恵袋	82	13976	2300.43
RWC　新聞	50	374	164.46

のデータ，また RWC コーパスの毎日新聞コーパス [5] の三つのデータを利用し，ひとつの単語につきソースデータとターゲットデータを変えることで，全部で 6 通りの領域適応を行った．これらのデータには岩波国語辞典 [12] の語義が付与されている．これらのコーパス中の多義語のうち，ソースデータおよびターゲットデータ中に存在する用例がともに 50 用例以上の単語を実験対象とした．単語の異なり数は，白書⇔ Yahoo! 知恵袋：24　白書⇔新聞：22　Yahoo! 知恵袋⇔新聞：26 であり，全体で 28 単語となった．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：「分かる」，「考える」，5 語義：「含む」，「使う」，「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」，11 語義：「進む」，12 語義：「見る」，14 語義：「入る」，16 語義：「言う」，21 語義：「出す」，22 語義：「手」，「出る」 である．

5　結果

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

表 2: 全体の適応手法別の実験結果

領域適応手法	正解率
Random Sampling	79.85%
Target Only	79.66%
確信度による合議	**83.49%**

これらの表で，コーパスごとに一番高い正解率を太字で示した．またその値を二番目に高い正解率と比較した際，0.05 水準で有意である場合にはその値に下線を引いた．

6　考察

表 3 から，Yahoo！知恵袋をソースデータとして新聞をターゲットデータとした領域適応と，白書をソースデーとしてタ Yahoo！知恵袋をターゲットデータとした領域適応を除いた 4 方向の領域適応において，提案手法である分類器の確信度を用いた合議が最も高い正解率を示すことが分かる．また，表 2 から，全ての方向の領域適応の平均をとった場合には，提案手法である分類器の確信度を用いた合議が最も高い正解率を示し，その値は二番目に高い正解率を示した Random Sampling の結果と比べて有意差が認められたことが分かる．これらのことから，本手法はどのようなコーパスの組み合わせに対しても有効であるわけではないが，一般的に有効な手法であると言えるだろう．

本稿では，Target Only と Random Sampling の二つの手法だけを比較し，この二つのうちより確信度の高い手法による分類器の分類結果を採用した．比較対象の分類手法が変わったとき，また増えた場合の提案手法の有効性の検証は今後の課題である．

7　おわりに

分類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を利用して分類器をつくり，ターゲットドメインの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．語義曖昧性解消（WSD: Word Sense Disambiguation）の領域適応の手法はさまざまあるが，我々は用例によって適切な手法は異なると考えた．本稿では，少量のターゲットデータにラベル付けして学習を行う方式と，他のコーパスを訓練事例に加える方式を使って二つの分類器を学習し，学習された分類器の出力する確信度の高い方の答えを採用することにより，分類の精度を向上させる手法を示した．自動的に選択された手法を用いて領域適応を行うことで，もともとの手法を一括的に使った時に比べ，WSD の平均正解率が有意に向上した．

謝辞

文部科学省科学研究費補助金特定領域研究「現代日本語書き言葉均衡コーパス」の助成により行われた．ここに，謹んで御礼申し上げる．

表 3: コーパスと適応手法別の実験結果

ソースデータ	Yahoo！知恵袋	Yahoo！知恵袋	白書	白書	新聞	新聞
ターゲットデータ	白書	新聞	Yahoo！知恵袋	新聞	Yahoo！知恵袋	白書
領域適応手法	正解率					
Random Sampling	87.21%	**73.95%**	**83.97%**	72.09%	76.61%	72.66%
Target Only	88.35%	66.46%	75.74%	67.75%	74.46%	84.57%
確信度による合議	**88.54%**	72.80%	83.03%	**72.48%**	**78.10%**	**87.81%**

参考文献

[1] Vincent Van Asch and Walter Daelemans. Using domain similarity for performance estimation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010*, pp. 31–36, 2010.

[2] Yee Seng Chan and Hwee Tou Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 89–96, 2006.

[3] 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.

[4] Hal Daumé, III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263, 2007.

[5] Koichi Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, and Wakako Kashino. The rwc text databases. In *Proceedings of The First International Conference on Language Resource and Evaluation*, pp. 457–461, 1998.

[6] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 264–271, 2007.

[7] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102, 2008.

[8] Gokhan Tur. Co-adaptation: Adaptive co-training for semi-supervised learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009.*, pp. 3721–3724, 2009.

[9] 張本佳子, 宮尾祐介, 辻井潤一. 構文解析の分野適応における精度低下要因の分析及び分野間距離の測定手法. 言語処理学会　第 16 回年次大会発表論文集, pp. 27–30, 2010.

[10] 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自動選択. 情報処理学会研究報告, Vol. 2010-NL-198, No. 5, pp. 1–6, 2010.

[11] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[12] 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.