語義曖昧性解消の領域適応のための訓練データの選択法
～複数ドメインからの選択～

堀内浩史郎（東京農工大学　工学部 情報工学科）
古宮嘉那子（東京農工大学　工学研究院）
小谷善行（東京農工大学　工学研究院）

Selection of Training Data for Domain Adaptation of Word Sense Disambiguation
- Selection from Multiple Domains -

Koshiro Horiuchi (Department of Computer and Information Sciences,
Faculty of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institute of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institute of Engineering, Tokyo University of Agriculture and Technology)

１．はじめに

ターゲットデータと異なるドメインのデータ（ソースデータ）で分類器を学習し，ターゲットデータに適応することを領域適応という．しかし，語義曖昧性解消ついて領域適応を行う際，分類したいデータごとに適切なソースデータは異なる．近年では複数のドメインの語義タグつきコーパスが入手できるため，分類したいデータに対して適切なソースデータを選択することが望ましい．

本稿では，ソースデータとして利用できるコーパスを複数を持っている場合を想定し，未知のターゲットデータが現れた際に，「用例全体の素性の平均ベクトルの類似度」「用例全体の出現素性の類似度」「用例全体の素性の分布」「分類器の示す確率」を使って訓練データの選択を試みて，訓練データの選択方法として利用できるものがあるか模索する．

２．関連研究

領域適応についての関連研究として，(Vincent Van Asch，Walter Daelemans(2010))の異なるドメインである訓練データとテストデータのコーパス同士の類似度から，品詞タグづけタスクにおける領域適応時の分類器の正解率を予測する研究がある．この研究では，コーパス同士の類似度と正解率の間に線形相関があることが示されている．(Daumé III(2007))は，素性空間を三倍にすることで，さまざまな supervised の領域適応に併用でき，さらに簡単に実装できマルチドメインに拡張も簡単であることを示した．

（古宮，奥村(2012)）は，語義曖昧性解消について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た．訓練データを選択する研究としては，（Komiya and Okumura(2012)）の訓練データの選択に分類器の確信度を用いる研究がある．全ての訓練データについて分類器をつくり，分類したときの各分類器の示す確信度によって訓練データを選択する手法である．さらに，確信度に加えて LOO-bound という指標を用いる研究も（古宮，小谷，奥村(2013)）によって行われている．本研究でも，確信度と LOO-bound のような指標を用いた実験を行うが，（古宮，小谷，奥村(2013)）では分類器にサポートベクトルマシン（Support Vector Machine，以下 SVM）を用いているのに対して，本研究では最大エントロピー（Maximum Entropy，以下 ME）法（Suarez and Palomar(2002)）の分類器を用いている点が異なっている．

また，（古宮，小谷(2011)）では領域適応が行われる状況によって最も良い手法が異なるとし，与えられたデータの性質を用いて三手法からひとつの手法を選択している．

３．訓練データ選択方法

いくつかのドメイン（ジャンル）のラベルつきデータを全て訓練データとして利用できる際に，ドメインのわからない未知のラベルなしデータを分類したい場合を考える．本稿では，未知のテストデータ（分類対象のターゲットデータ）に合わせた訓練データをいくつかの手法によって選択し，各手法が正解率の向上につながるか調べる．実験は次のようなステップで行う．

I. 各手法による訓練データの選択
II. Iで選択した訓練データでの領域適応（分類実験）
III. 他の手法で選択した訓練データを用いた場合との語義曖昧性解消の正解率の比較

ステップ I において，次に示す手法を試みる．

i. 類似度を用いた訓練データの選択
   - 用例全体の素性平均ベクトルの類似度を利用する手法
   - 用例全体の出現素性ベクトルの類似度を利用する手法
ii. 素性分布の距離を用いた訓練データの選択
iii. 分類器の示す確率を用いた訓練データの選択
   - 分類器の分類確率を利用する手法
   - 分類器の自信度を利用する手法
   - 分類器の分類確率と自信度を利用する手法

なお，語義曖昧性解消の対象単語タイプごとに分類器を作成するため，訓練データの選択は単語のタイプごとに行った．また，iii に関しては，（Komiya and Okumura(2012)）にならい，テストデータの用例ごとに選択を行う実験も行った．

３．１　類似度を用いた訓練データの選択

テストデータと訓練データを表すベクトルを各ひとつずつ，それぞれの素性ベクトル集合を用いて作成し，そのベクトル同士の類似度を訓練データの選択指標として用いる．ベクトルは，各要素を足して用例数で割った「素性平均ベクトル」と，全ての要素の OR を取った「出現素性ベクトル」の二つについて調べる．利用する類似度は，ユークリッド距離（ED），コサイン類似度（CS），ジャッカード係数（JSD），ダイス係数（DSC），シンプソン係数（SSC），ランド類似度（RS）を用いる．なお，全ての類似度についてテストデータとの類似度が最高値であったデータを訓練データにした場合（以後，（最大）と表記）と，最小であったデータを訓練データにした場合（以後，（最小）と表記）の 2 通りを調べる．

３．２　素性分布の距離を用いた訓練データの選択

テストデータと訓練データの素性ベクトル集合において各素性の分布（本稿では 17 個の分布）を作成し，各素性分布同士の距離を測り，その距離の総和が最も小さくなる訓練データを選択する．各素性分布の距離の測定にはジェンセン・シャノン・ダイバージェンスを用いる．

３．３　分類器の示す確率を用いた訓練データの選択

ME 法を用いて分類を行うと，各ラベルに分類される確率が算出される．この確率を「分類確率」と呼び，分類確率の中の最大値を最大分類確率と呼ぶこととする．また，訓練データを 5 分割交差検定した結果を「自信度」と呼ぶこととする．「自信度」は，その分類器

がその訓練データと同じドメインのコーパスをどの程度正確に分類できるかを表す.

訓練データの選択には，最大分類確率の平均値が最大となる訓練データを選択する手法と，自信度を用いて訓練データを選択する手法，そしてこれら二つの値の積を用いて訓練データを選択する手法を試みる.

上記の訓練データ選択実験に加えて，各用例ごとに分類確率を用いて訓練データの選択する手法を試みた．訓練データを変えて分類器を学習し，学習された分類器の中で最高の分類確率を示した分類器の結果を用例ごとに選択する.

４．訓練データ選択実験

４．１　最大エントロピー法

本実験の分類手段として ME 法を用いる．ME モデルの実現には（Le Zhang(2011)）の Maximum Entropy Modeling Toolkit for Python and C++を用いた.

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）(Maekawa (2008))の白書のデータと Yahoo！知恵袋のデータ，またRWCコーパス（Hashida et al. (1998)）の新聞記事を用いた.

単語の語義は岩波国語辞典（西尾ら (1994)）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 1　ドメインごとの単語の最小，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

実験は，三つのドメインのうちひとつのドメインをテストデータとして利用し，他の二つのドメインのデータから訓練データを選択する．たとえば Yahoo！知恵袋をテストデータとした場合は，訓練データの選択肢は「新聞記事」「白書」「新聞記事＋白書」の三通りである.

各ドメインの各単語ごとに選択実験を行うので，それぞれの手法に対して計 66 回の実験を行う.

５．結果

本実験のベースラインは利用できる訓練データ二つのドメインの両方を利用した場合である．表２に各手法の実験結果を示す．なお，全ての手法の中で最も良い正解率を下線に示す．また，テストデータのドメインごとに最も良かった結果に下線を引いた．マクロ平均・マイクロ平均ともに，新聞記事が素性分布の距離，白書が出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小），Yahoo！知恵袋が出現素性ベクトルのコサイン類似度（最大）で訓練データを選択したときに語義曖昧性解消の正解率が最も高くなった．表３にこれらのドメインごとの結果を示す.

なお，出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小）は各ドメインの各単語について全て同じ訓練データを選択したために，同じ結果となっている.

表２　訓練データ選択実験結果

手法		マクロ平均（%）	マイクロ平均（%）
ベースライン		75.04	81.1
素性平均ベクトル	ED（最大）	73.59	74.64
	ED（最小）	71.05	79.09
	CS（最大）	71.82	73.97
	CS（最小）	71.16	78.58
	JSC（最大）	74.21	81.02
	JSC（最小）	69.86	72.92
	DSC（最大）	74.21	81.02
	DSC（最小）	69.86	72.92
	SSC（最大）	73.18	74.79
	SSC（最小）	71.68	78.2
	RS（最大）	67.94	72.68
	RS（最小）	74.37	79.75
出現素性ベクトル	ED（最大）	75.34	81.39
	ED（最小）	68.83	73.74
	CS（最大）	74.38	81.55
	CS（最小）	68.99	73.28
	JSC（最大）	73.33	80.89
	JSC（最小）	71.8	74.31
	DSC（最大）	73.37	80.91
	DSC（最小）	71.8	74.31
	SSC（最大）	71.86	74.2
	SSC（最小）	71.64	79.9
	RS（最大）	68.79	73.55
	RS（最小）	75.34	81.39
素性分布の距離		75.02	81.02
分類確率		70.58	77.53
自信度		72.7	80.54
分類確率と自信度		72.7	80.54
分類確率でラベル予測		74.08	80.41

６．考察

表３が示すように，ドメインごとに適当な訓練データの選択手法は異なる．さらに全体のマクロ平均が最も良かった出現素性ベクトルのユークリッド距離（最大）について，語義曖昧性解消の対象単語のタイプごとに結果を詳しく調べると，訓練データよりもテスト

表３　ドメインごとの実験結果

	マクロ平均（%）			マイクロ平均（%）		
手法	新聞記事	白書	Yahoo！知恵袋	新聞記事	白書	Yahoo！知恵袋
出現素性ベクトルの ED（最大）	74.37	76.74	74.9	75.56	80.32	82.66
出現素性ベクトルの RS（最小）	74.37	76.74	74.9	75.56	80.32	82.66
出現素性ベクトルの CS（最大）	71.26	74.28	77.6	73.6	79.66	83.63
素性分布の距離	74.96	74.16	75.95	75.83	79	83.03
ベースライン	74.37	76.57	74.17	75.56	79.86	82.48

データの方が用例数が多い場合に正解率が上がっているものが多いことが分かった．このことから，適当な訓練データの選択手法は語義曖昧性解消の対象単語のタイプごとにも異なることが分かる．

訓練データよりもテストデータが少ない場合について調べると，素性平均ベクトルのユークリッド距離（最大）を用いた場合が最も良い結果となった．ここで，訓練データよりもテストデータが多い場合は出現素性ベクトルのユークリッド距離（最大）を，少ない場合は素性平均ベクトルのユークリッド距離（最大）を用いて訓練データを選択した結果を表４に示す．

表４　二手法を組み合わせたときの正解率

	マクロ平均			マイクロ平均		
手法	新聞記事	白書	Yahoo！知恵袋	新聞記事	白書	Yahoo！知恵袋
二手法組み合わせ	74.37	76.65	77.53	75.56	80.25	83.87
ベースライン	74.37	76.57	74.17	75.56	79.86	82.48

表４より，白書と Yahoo！知恵袋でベースラインよりも語義曖昧性解消の正解率が良くなり，新聞記事でもベースラインと同じ正解率となった．今回の類似度の組み合わせは本研究で利用したデータの特徴から手法を選択しているため，よりデータの性質から適した手法の組み合わせを考える必要があるだろう．

分類確率を用いた実験については，SVM で有効に働いていたが，本実験の ME 法では語義曖昧性解消の正解率を上げることができなかった．

７．まとめ

語義曖昧性解消における領域適応の正解率を向上させるために，「素性平均ベクトルの類似度」「出現素性の類似度」「素性分布の距離」「分類器の分類確率と自信度」を用いて訓練データの選択を行い，どの選択手法が最も優れているかを調べた．全体の平均を見ると，マクロ平均で出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小），マイクロ平均で出現素性ベクトルのコサイン類似度（最大）で選んだ際に，語義曖昧性解消の正解率が最も高くなった．

各ドメインの結果を見ると，マクロ平均・マイクロ平均ともに，新聞記事が素性分布の距離, 白書が出現素性ベクトルのユークリッド距離（最大）とランド類似度（最小）, Yahoo！知恵袋が出現素性ベクトルのコサイン類似度（最大）で訓練データを選択したときに語義曖昧性解消の正解率が最も高くなった．

それぞれの手法は訓練データとテストデータの性質によって異なる様相を見せたため，データサイズによって二手法を組み合わせた実験を行った．二手法を組み合わせた結果，語義曖昧性解消の正解率が全てのドメインでベースライン以上となった．

分類確率を用いた手法については，本研究では ME 法を用いて実験したが，異なる分類器である SVM を用いた関連研究のように語義曖昧性解消の正解率を上げることができなかった．

謝　辞

本研究は，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．

文　献

Vincent Van Asch，Walter Daelemans(2010)「Using Domain Similarity for Performance Estimation」，DANLP 2010，pp.31-36.

H. Daumé III(2007)「Frustratingly Easy Domain Adaptation」，ACL 2007，pp.256-263.

H. Daumé III，Abhishek Kumar，Avishek Saha(2010)「Frustratingly Easy Semi-Supervised Domain Adaptation」，ACL 2010，pp.53-59.

Le Zhang(2011)「Maximum Entropy Modeling Toolkit for Python and C++」，http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Kanako Komiya and Manabu Okumura(2011)「Automatic determination of a domain adaptation method for word sense disambiguation using decision tree learning」, IJCNLP 2011, pp.1107-1115.

古宮嘉那子，奥村学(2011)「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」，言語処理学会第 17 回年次大会発表論文集，pp.552-555.

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選択―三手法からの決定―」，言語処理学会第 18 回年次大会発表論文集，pp.1288-1291.

古宮嘉那子，小谷善行(2011)「階層型クラスタリングを利用した文脈によるオノマトペの分類」，NLP 若手の会第 6 回シンポジウム.

Michel Marie Deza, Elena Deza(2012)「Encyclopedia of Distances」，Springer-Verlag.

西尾実，岩淵悦太郎，水谷静夫(1994)「岩波国語辞典第五版」，岩波書店.

Koichi Hashida，Hitoshi Isahara，Takenobu Tokunaga，Minako Hashimoto，Shiho Ogino，and Wakako Kashino(1998)「The rwc text databases」，LREC 1998，pp.457-461.

Kikuo Maekawa(2008)「Balanced corpus of contemporary written japanese」, ALR 2008, pp.101-102.

Armándo Suarez，Manuel Palomar(2002)「A Maximum Entropy-based Word Sense Disambiguation system」，COLING 2002，Vol.1，pp.1-7.