Stacked Denoising Autoencoderを利用した語義曖昧性解消の領域適応

河野 和平　新納 浩幸　佐々木 稔　古宮 嘉那子
茨城大学 工学部情報工学科
11t4077h@hcs.ibaraki.ac.jp, {shinnou, msasaki, kkomiya}@mx.ibaraki.ac.jp

1　はじめに

本稿では深層学習 (Deep Learning) を利用し，語義曖昧性解消 (Word Sense Disambiguation, WSD) の領域適応の問題解決を図る．

WSD において，訓練データとテストデータは異なる領域から得られている場合がある．このような場合，識別精度が低下してしまう．これを解決する方法として領域適応 [1] があり，本研究室においても関連研究が行われている [2][3]．本稿ではこの問題を，異なる領域から得られた訓練データとテストデータは同じ単語に関するデータであっても，その素性に大きな差が生じることによる影響と捉える．

深層学習はニューラルネットワーク (Neural network，NN) を用いて，データの抽象的表現を獲得する手法である． 近年活発に研究されており，画像認識や音声認識の分野でよい結果が得られている [4]．これを WSD の領域適応のタスクに適用し，データの抽象的表現を抽出することで前述した問題を解決する．本稿では複数存在する深層学習のモデルのうち，Stacked Denoising Autoencoder(SdA) を利用した．

実験では，現代日本語書き言葉均衡コーパス (BCCWJ)[5] の領域 OC 及び PB のうち 16 単語を対象とし，領域適応は，OC → PB と PB → OC の双方向を行った．まずはじめに，Latent Semantic Indexing(LSI) により獲得した素性と SdA により獲得した素性とで SVM の識別精度の比較を行った．実験の結果，SdA を利用したデータの識別精度は，各領域適応で LSI よりも良い結果が得られた．次に，SdA で獲得するデータの抽象的表現の次元数に関する実験も行った．その結果，抽出する素性の次元数により，識別精度に大きな変化が見られた．

2　語義曖昧性解消の領域適応

単語には，一般に複数の意味が存在する．語義曖昧性解消は，文中に出現するこのような単語の語義を一意に識別するタスクである．

WSD のタスクにおいて，訓練データとテストデータは異なる領域から得られていることも多い．例えば，訓練データとして書籍から得た文章を利用して分類器を学習し，新聞から得た文章中の単語の語義を識別したいような場合である．このような場合，書籍 (ソース領域) から得たデータで学習した分類器では，新聞 (ターゲット領域) から得たデータを精度良く識別することは困難である．そこで，ソース領域の訓練データによって学習した分類器をターゲット領域の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呼ぶ．

本稿では，この領域適応の問題をソース領域とターゲット領域のデータにおける素性の相違からくる問題と捉え，深層学習によってこれを緩和させることを試みる．

3　深層学習

深層学習は，NN を用いた教師なし学習法で，NN を多層に重ねることにより抽象的なデータ表現を獲得する．前述したように，WSD の領域適応はソース領域とターゲット領域のデータにおける素性の相違により，識別精度が低下してしまう問題がある．そこで，深層学習によりデータの抽象的表現を獲得し，この素性の相違を緩和させる．深層学習のモデルには，確率論的モデルである RBM(Restricted Boltz-

mann Machines) や決定論的モデルの SdA(Stacked Denoising Autoencoder)，RBM の積み重ねからなる DBN(Deep Belief Nets) などが存在する．このうち，本稿では WSD の領域適応に SdA[6] を用いる．

3.1 Denoising Autoencoder

Autoencoder(AE) は，図 1 のように入力層，隠れ層，出力層の 3 層構造からなり，入力層と出力層の差が小さくなるようなモデルを学習する．

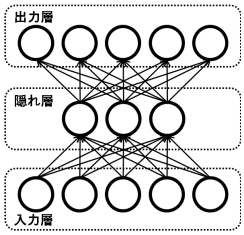

図 1: Autoencoder

Denoising Autoencoder(dA) は AE の入力ベクトル x 対して確率的にノイズを付与したベクトル x' を入力層に与え，ノイズを付与する前の入力ベクトル x と出力層の差が小さくなるようなモデルを学習する．従って dA では，付与したノイズを除去するようなモデルを学習する．入力層 x' から隠れ層 y，隠れ層 y から出力層 z の写像はそれぞれ以下の様に表される．

$$y = sigmoid(Wx' + b) \quad (1)$$

$$z = sigmoid(W'y + b') \quad (2)$$

ここで，b，b' はバイアス，W は重み行列を示し，W' はその転置行列である．また，$sigmoid()$ はシグモイド関数を表す．確率的勾配降下法 (SGD) により，x と z の交差エントロピー誤差が最小となる $W(W')$，b，b' を求める．このようにして得られた隠れ層 y は入力データ x の抽象的表現となる．

3.2 Stacked Denoising Autoencoder

Stacked Denoising Autoencoder(SdA) は，dA を複数回積み重ねたものである．まず，dA によって学習を行う．次に，得られた隠れ層を入力として再度 dA によって学習する．これを繰り返し行うことで，dA による学習を積み重ね，生のデータからデータの抽象的な表現の素性を得る．従って，dA によって得られた出力層は交差エントロピー誤差を求めるためだけに使用し，SdA ではこの過程で得られた隠れ層を利用する．

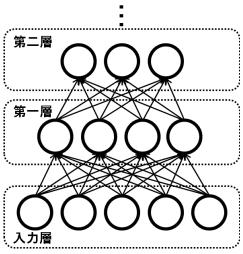

図 2: Stacked Denoising Autoencoder

本稿では，SdA によって得られたデータの抽象的な素性を生のデータに付加することで，領域適応のタスクにおけるソース-ターゲット間の素性の相違を緩和させる．具体的には，SdA を利用して n_x 次元の入力データ x から n_{abst} 次元の抽象的表現 x' を抽出する．入力データ x 及び抽象的表現 x' をそれぞれ正規化し，これらを結合した $n_x + n_{abst}$ 次元のデータを得る．

4 実験

4.1 比較実験

WSD の領域適応のタスクに SdA を利用することの有効性を検証する．SdA によるデータの抽象的表現の獲得は，入力データの素性に対して少ない次元数の素性を獲得するという観点から見れば一種の次元縮約とみなすことができる．そこで実験では，代表的な次元縮約法の 1 つである特異値分解を利用した次元削減法 (Late Semamtoc Indexing，LSI) との比較を行う． 実験は BCCWJ の 2 つの領域 OC(Yahoo! 知恵袋) 及び PB(書籍) のデータのうち，それぞれの領域である程度の頻度で存在する 16 単語を対象とし，双方向 (OC → PB 及び PB → OC) の領域適応を行った．

SdA の学習には，Deep Learning ライブラリとして公開されている Pylearn2[1] を利用する．SdA による dA の積み重ね回数は 2 回とする．第一の dA による学習は，n_x 次元の入力データ x に対して，隠れ層のノード数は $2/3 * n_x$ 個に設定し，第二の dA による学習では第一の dA の隠れ層のノードを入力とし，この学習でノード数は変化させない．この第二の dA の隠れ層のノードが入力データの抽象的表現の素性となる．従って，SdA 全体で得られる素性は入力データの次元数に対して 2/3 次元のデータとなる．

LSI では得られる素性の次元数が SdA の場合と同様になるように n_x 次元の入力データ x に対して 1/3 次元分の素性を削減し，$2/3 * n_x$ 次元の素性を得る．

それぞれの方法によって獲得した素性と入力素性を正規化し，結合したものを Support Vector Machines(SVM) によって識別する．SVM による識別には libsvm[2] を用いる．ソース領域→ターゲット領域を OC → PB としたときの結果を表 1 に，PB → OC の結果を表 2 に示す．ここで，Normal は入力データの素性に対して何も処理をせずに SVM によって識別した結果を表す．比較すると，単語により識別精度が改善されたものとそうでないものが存在するものの，全体としては SdA によって抽象的表現を抽出したデータが最も良い結果が得られた．

表 1: 実験結果 (OC → PB(%))

Word	Normal	LSI	SdA
言う	82.05	81.78	82.14
入れる	62.5	66.07	66.07
書く	83.87	77.42	75.81
聞く	66.67	68.29	67.48
子供	24.73	30.11	27.96
時間	90.54	87.84	87.84
自分	97.40	94.81	95.13
出る	55.92	57.89	58.55
取る	23.46	25.93	24.69
場合	84.67	84.67	84.67
入る	43.22	45.76	47.46
前	69.38	77.50	77.50
見る	83.88	83.88	83.88
持つ	77.12	79.09	77.78
やる	92.95	92.95	92.95
行く	88.72	88.72	88.72
Total	77.20	77.70	77.73

[1] http://deeplearning.net/software/pylearn2/
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

表 2: 実験結果 (PB → OC(%))

Word	Normal	LSI	SdA
言う	83.78	83.48	83.63
入れる	73.97	71.23	76.71
書く	73.74	73.74	73.74
聞く	66.13	65.32	70.16
子供	12.99	25.97	23.38
時間	83.02	83.02	83.02
自分	87.50	87.50	87.50
出る	64.89	66.41	66.41
取る	24.59	31.15	32.79
場合	91.27	90.48	90.48
入る	63.24	58.82	58.82
前	89.52	90.48	89.52
見る	56.11	56.49	56.49
持つ	64.52	82.26	82.26
やる	94.02	94.02	94.02
行く	68.95	68.95	68.95
Total	73.09	73.94	74.31

4.2 ノード数の調整

SdA による学習の精度は，各 dA における隠れ層のノード数に大きく依存すると考えられる．そこで，隠れ層のノード数による識別精度の比較を行う．前述の実験において，SdA は入力データの次元数に対して 2/3 次元の素性を得た．隠れ層のノード数の変化による識別精度の比較を行うため，これを 1/3 次元に変更し，再実験を行った．すなわち，本実験では，n_x 次元の入力データ x に対して第一，第二の dA の隠れ層のノード数を $n_x/3$ 個とした．実験の結果を表 3 に示す．

表 3: ノード数による識別精度の変化 (%)

	OC → PB		PB → OC	
Word	2/3 次元	1/3 次元	2/3 次元	1/3 次元
言う	82.14	81.51	83.63	82.88
入れる	66.07	64.29	76.71	72.60
書く	75.81	74.19	73.74	73.74
聞く	67.48	63.41	70.16	66.94
子供	27.96	26.88	23.38	40.26
時間	87.84	86.49	83.02	83.02
自分	95.13	94.16	87.50	87.50
出る	58.55	59.21	66.41	62.60
取る	24.69	23.46	32.79	26.23
場合	84.67	84.67	90.48	91.27
入る	47.46	47.46	58.82	54.41
前	77.50	76.88	89.52	87.62
見る	83.88	83.88	56.49	56.87
持つ	77.78	75.82	82.26	74.19
やる	92.95	92.95	94.02	94.02
行く	88.72	89.47	68.95	70.78
Total	77.73	77.04	74.31	73.81

SdA によって入力データに対して 2/3 次元の素性を獲得する場合と比べ，一部の単語で僅かに精度が改善されたケースも存在するが，大半の単語において精度が低下或いは変化が見られず，大幅に悪化してしまうケースも存在した．また，OC → PB の領域適応において，入力データに対して 1/3 次元の素性を得る SdA では，そのままのデータを SVM によって識別する場合よりも精度が悪くなった．

5 考察

本稿では WSD の領域適応のタスクにおいて SdA を用いてデータの抽象的表現を抽出し，素性に加えることで識別精度の改善を試みた．領域適応の主問題として，同じ単語に関するデータであってもソース-ターゲット間には素性に大きな差があり，識別精度が低下してしまうことが挙げられる．SdA によってデータの抽象的な表現を獲得することができれば，この素性の相違をうまく緩和させることができると考えた．

実験では，なにも施さないそのままのデータや LSI によって獲得したデータと比べて，SdA によって獲得したデータに対する SVM の識別精度の方が良い結果が得られた．これは，SdA によってデータの抽象的な表現をうまく獲得でき，これを素性に加えることで，意図したようにソース領域とターゲット領域の素性の相違を緩和できたと考えられる．

次に，dA の隠れ層のノード数を変更し，精度の比較を行った．結果として，ノード数をより少なくした場合の結果は，そうでない場合よりも精度が悪化した．更には，領域によっては，SdA を実施せず，そのままのデータを SVM によって識別した方が良い結果となった．dA において，隠れ層のノード数を少なく設定すると，各ノードはより抽象的な情報を保持していると言える．しかし，極端に少ないノード数の場合，正確な情報を表現しきれなくなってしまい，識別精度が低下してしまうと考えられる．今後，この dA の隠れ層のノード数に関する追実験を行い，適切なノード数を考察する予定である．

dA の積み重ね回数に関しても実験が必要である．本稿では，dA の積み重ね回数を 2 回としたが，これを増やし，より抽象的な表現を獲得した場合，識別精度がどのように変化するか考察を行う必要がある．

6 おわりに

本稿では WSD の領域適応において，深層学習を利用し，データに抽象的な表現の素性を付加することの有効性を検証した．領域適応の問題をソース-ターゲット間のデータの素性の相違からくる識別精度の低下にあると捉え，SdA によりデータの抽象的な表現を抽出することでこれを緩和させることを試みた．

実験では，BCCWJ の OC 及び PB のデータを利用した．その結果，SdA の識別精度は LSI と比べて良い結果が得られた．しかし，dA の隠れ層のノード数に関する実験では，これが識別精度に大きな影響を与え，適切な設定ができなければ，精度が悪化してしまうことが分かった．今後は，dA の隠れ層のノード数に関する追実験と dA の積み重ね回数に関する実験を行い，適切なパラメータを考察したい．

参考文献

[1] Anders Sogaard. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool, 2013.

[2] 新納浩幸, 佐々木稔. 共変量シフト下の学習による語義曖昧性解消の教師なし領域適応. 自然言語処理, Vol. 21, No. 5, pp. 1011–1035, 2014.

[3] 新納浩幸, 佐々木稔. 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. 自然言語処理, Vol. 20, No. 5, pp. 707–726, 2013.

[4] Quoc Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg Corrado, Jeff Dean, and Andrew Ng. Building high-level features using large scale unsupervised learning. In *ICML-2012*, 2012.

[5] Kikuo Maekawa. Design of a Balanced Corpus of Contemporary Written Japanese. In *LKR-2007*, pp. 55–58, 2007.

[6] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain Adaptation for Large-Scale Sentiment Classification:A Deep Learning Approach. In *ICML-2011*, Vol. 27, pp. 97–110, 2011.