単語の分散表現に基づく極性判定のための教師なし分野適応

森谷 一至
北陸先端科学技術大学院大学
s1910217@jaist.ac.jp

白井 清昭
北陸先端科学技術大学院大学
kshirai@jaist.ac.jp

1 はじめに

テキスト中に表明されている書き手の意見が肯定的か否定的かを判定する極性判定は，オピニオンマイニングにおける基礎的な技術のひとつである．近年の極性判定に関する研究は教師あり機械学習を用いた手法が主流であるが，一般に，教師あり機械学習によって得られた分類モデルは，訓練データと異なるドメインのテストデータに適用すると分類の正解率が低下することが知られている．オピニオンマイニングでは，評価対象は多岐に渡るため，ドメインの違いによる極性判定の正解率の低下は重要な問題である．

分野適応 (転移学習) とは，訓練データとテストデータのドメインが異なるときに，テストデータに対する分類性能が落ちないように分類モデルを学習する技術である．このとき，訓練データ，テストデータのドメインはそれぞれソースドメイン，ターゲットドメインと呼ばれる．分野適応の手法には，少量のターゲットドメインの正解ラベル付きデータを用いる教師あり分野適応と，ターゲットドメインについてはラベル付きデータを用いない教師なし分野適応がある．本研究は，単語の分散表現を利用した教師なし分野適応手法を提案する．また，通販サイトに投稿された日本語レビューの極性判定を対象に提案手法の有効性を実験的に検証する [1].

2 関連研究

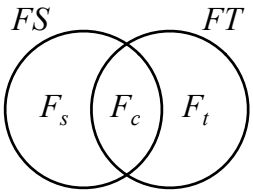
Blitzer らは Structural Correspondence Learning(SCL) という分野適応の手法を提案している [2]．ソースドメインとターゲットドメインの両方で頻出し，かつ分類ラベルとの相互情報量が高い単語 (ピボット素性) と，極性判定に関係する分野特有の単語 (非ピボット素性) の関連性をモデル化することで分野適応を行う．

Glorot らは，Stacked Denoising Autoencoders (SDA) を用いて，利用可能な全てのドメインのレビューから分野に依存しない抽象度の高い素性を教師なしの手法で抽出する方法を提案している [3]．Amazon レビューデータを用いた極性判定の実験で，先行研究の手法である SCL[2], Multi-label Consensus Training (MCT)[4], Spectral Feature Alignment (SFA) algorithm[5] と比べて，彼らの手法の分野適応能力が高いことを確認した．

Ziser と Reichart は，Pivot Based Language Model (PBLM) による教師なし分野適応の手法を提案している [6]．PBLM は，後続するピボット素性あるいは None(ピボット素性以外の素性) を予測する LSTM (Long Short-Term Memory) Language Modeling であり，大量のラベルなしソースドメインとターゲットドメインのデータから学習される．さらに，PBLM を入力とする LSTM (PBLM-LSTM) もしくは CNN (PBLM-CNN) によってテキストの極性を判定する．

Wang は，Bag-of-words を機械学習の素性とし，訓練データを素性ベクトルに変換する際，ソースドメインの訓練データに出現する単語と類似しかつターゲットドメインのみに出現する単語を素性ベクトルに追加することで分野適応を行った [7]．ここで単語間の類似度は事前学習された単語の分散表現を用いて測る．英語で書かれたテキストの著者の性別を推定するタスクについて，マイクロブログ (Twitter) ならびにブログを異なるドメインとする実験を行い，単語の分散表現を用いた素性拡張によって性別推定の正解率が向上することを確認した．

本研究は，Wang による手法 [7] を修正し，ターゲットドメインのみに出現する素性を拡張する際に，クラス分類に大きく貢献する素性のみを拡張する手法を提案する．また，Wang の実験は英語で書かれたテキストの著者の性別を判定していたが，本研究では日本語で書かれたレビューに対する極性判定を対象に分野適応を行う．単語の分散表現に基づく教師なし分野適応の手法がタスク (性別推定もしくは極性判定) や言語 (英語もしくは日本語) が異なる場合でも有効であるかを実験的に検証する．



$F_c = FS \cap FT$　　$F_s = FS \setminus F_c$　　$F_t = FT \setminus F_c$

図 1　素性集合の定義

3　提案手法

3.1　問題設定

ユーザによって書かれた商品レビューの極性を判定する．ここでの極性のクラスは「肯定的」「否定的」の 2 つとする．レビューは評価対象の商品カテゴリ (「食品」「レディースファッション」など) によっていくつかのドメインに分かれているものとする．学習データとテストデータではドメインが異なると仮定する．また，ソースドメインのレビューには正解ラベルが付与されているが，ターゲットドメインのレビューには付与されていないものとする (教師なし分野適応)．

3.2　素性ベクトルの拡張

極性判定モデルの学習は以下に述べるような基本的な手法を用いる．機械学習の素性として Bag-of-words 素性を用いる．レビューを MeCab を用いて形態素解析し，自立語 (名詞，動詞，形容詞，副詞) を抽出し，その基本形を素性とする．ソースドメインならびにターゲットドメインの個々のレビューを素性ベクトルで表現する．素性ベクトルはバイナリベクトルとする．すなわち，ある素性 (単語) がレビューに出現するときにはその素性の重みを 1 に，それ以外の素性の重みは 0 とする．

ソースドメインのレビュー集合に含まれる素性の集合を FS，ターゲットドメインのレビュー集合に含まれる素性の集合を FT とおく．さらに，図 1 に示すように，ソース・ターゲットの両方に出現する素性の集合を F_c，ソースドメインのみに出現する素性の集合を F_s，ターゲットドメインのみに出現する素性の集合を F_t とおく．

ドメインが異なる場合，実質的には F_c の素性のみで極性判定を行うことになる．F_c の素性の数が少ない場合，極性判定の正解率が低下すると考えられる．ターゲットドメインのレビューの極性を判定するためには，ターゲットドメインに固有の素性，すなわち F_t の素性が重要な役割を果たすことが予想されるが，分類器の学習には F_c と F_s の素性のみが使われ，F_t の素性は使われない．そこで，ソースドメインの訓練データから素性ベクトルを作成する際，F_t の素性を自動的に追加する．以下，この手続きを素性ベクトルの拡張と呼ぶ．

訓練事例 (レビュー) の素性ベクトルの拡張の手続き [7] を説明する．F_t の要素であり，レビューに出現する素性 $f (\in F_c \cup F_s)$ と類似した素性 f' を素性ベクトルに追加する．正確には，素性ベクトルの次元数は $|F_c \cup F_s \cup F_t|$ のままだが，f' に対する重みを 0 から 1 に変更する．素性間の類似度，すなわち単語間の類似度は，大量のテキストから事前学習された単語の分散表現のコサイン類似度で測る．本研究では，単語の分散表現として，ウェブテキストから事前学習された単語埋め込みである NWJC2Vec[8] を用いる．

3.3　有効な素性に限定した素性ベクトルの拡張

前項で述べた素性拡張の手法では，新たに素性ベクトルに追加された素性の中には極性判定に有効でないものが含まれる可能性がある．有効でない素性の拡張は極性判定の正解率の低下を招く原因になりうる．提案手法では，有効な素性に限定して素性ベクトルを拡張する．

提案手法による素性ベクトルの拡張の擬似コードを Algorithm 1 に示す．ここで，i 番目の訓練事例の素性ベクトルを $\vec{v}_i = (w_{i1}, \cdots, w_{in})^T$ とおく．w_{ij} は j 番目の素性 f_j に対する重み，n は素性の総数 $(n = |F_c \cup F_s \cup F_t|)$ を表す．

まず，17〜23 行目に示すように，有効な素性を以下の 2 つの条件を満たす素性と定義する．

- 極性クラスとの相関が強い

  極性判定の 2 つの分類クラスのうち，肯定クラス c と素性 f_j の相関の強さを χ^2 統計量 ($\chi^2(c, f_j)$ と記す)[9] で測る．素性 f_j が $\chi^2(c, f_j)$ の大きい上位 T_{cor} 件の素性のひとつであるとき，有効な素性とする．18 行目の $rank$ は素性を $\chi^2(c, f_j)$ の順に並べたときの順位を表す．
- 素性の出現頻度が大きい

  素性 f_j の訓練データにおける出現頻度 ($Fre(f_j)$ と記す) が閾値 T_{fre} よりも大きいとき，有効な素性とする．

Input: T(original training data)
Output: T'(new training data)

1 $T' \rightarrow \phi$;
2 **for** $\vec{v_i} \in T$ **do**
3 $\quad T' \rightarrow T' \cup \{ExtendFeatureVector(\vec{v_i})\}$
4 **end**
5 **Function** *ExtendFeatureVector*($\vec{v_i}$)**:**
6 $\quad$ **for** $f_j \in F_c \cup F_s$ **do**
7 $\quad\quad$ **if** $w_{ij} = 1$ ***and*** *IsEffective*(f_j) **then**
8 $\quad\quad\quad$ **for** $f_k \in F_t$ **do**
9 $\quad\quad\quad\quad$ **if** $cos(\overrightarrow{we}(f_j), \overrightarrow{we}(f_k)) \geq T_{sim}$ **then**
10 $\quad\quad\quad\quad\quad$ $w_{ik} \rightarrow 1$
11 $\quad\quad\quad\quad$ **end**
12 $\quad\quad\quad$ **end**
13 $\quad\quad$ **end**
14 $\quad$ **end**
15 $\quad$ **Return** $\vec{v_i}$
16 **end**
17 **Function** *IsEffective*(f_j)**:**
18 $\quad$ **if** $rank(\chi^2(c, f_j)) \leq T_{cor}$ ***and*** $Fre(f_j) \geq T_{fre}$ **then**
19 $\quad\quad$ **Return** true
20 $\quad$ **else**
21 $\quad\quad$ **Return** false
22 $\quad$ **end**
23 **end**

Algorithm 1: 素性拡張アルゴリズム

Algorithm 1 では，2〜4 行目に示すように，個々の訓練事例に対して素性ベクトルを拡張して，新しい訓練データ T' を作成する．素性ベクトルを拡張する際，レビューに出現しかつ有効な素性に対し (7 行目)，それと類似しターゲットドメインのみに出現する素性を素性ベクトルに追加する (8〜12 行目)．このとき，単語の分散表現 ($\overrightarrow{we}$) のコサイン類似度が閾値 T_{sim} 以上の素性を拡張する．

提案手法には 3 つのパラメタ $T_{sim}, T_{cor}, T_{fre}$ が存在する．これらのパラメタは訓練データを用いた交差検定によって最適化する．

4 評価実験

4.1 実験設定

評価実験には，楽天データ [10] のうち楽天市場に投稿されたレビューのデータセットを用いた．楽天市場における 34 の商品ジャンルのうち，レビュー件数の多い上位 5 つのジャンル (表 1 に示す) を選び，これらのジャンルの商品について投稿されたレビューを抽出した．それぞれのレビューに対し，レビューワーによって付与された 1〜5 の評価ポイントが 5 または 4 のときは「肯定」のラベルを，1 または 2 のときは「否定」のラベルを付与し，それ以外のレビューは除去した．肯定と否定のバランスを取るため，またジャンル毎のデータ数をほぼ同じにするため，肯定，否定のラベルがついたレビューをランダムにおよそ 2,000 件サンプリングし，実験データとした．表 1 は楽天データのレビュー数とサンプリング後のレビュー数を示している．

表 1 実験データ

	食	レ	日	美	イ
レビュー	124,646	118,604	101,569	84,671	79,364
肯定	109,564	95,738	93,663	73,923	71,582
否定	5,557	8,945	2,066	2,952	1,995
サンプル	4,000	4,000	4,000	4,000	3,990

食=食品，　レ=レディースファッション，
日=日用品雑貨・文房具・手芸，
美=美容・コスメ・香水，　イ=インテリア・寝具・収納

あらかじめ，それぞれのジャンルのレビューデータを 80%の訓練データと 20%のテストデータに分割した．そして，5 つのジャンルの全ての組み合わせについて，1 つをソースドメイン，もう 1 つをターゲットドメインとして，極性判定の実験を行った．これはソースドメインとターゲットドメインが同じ場合も含む．ソースとターゲットドメインが異なる場合でも，ドメインが同じ場合とほぼ同じ量のテストデータで比較するため，80%の訓練データと 20%のテストデータを使用する．

本実験では以下の 3 つの手法を比較した．

ベースライン (BL) 分野適応の手法を使わない手法
素性拡張 1 (FE1) 全ての素性から素性ベクトルを拡張する手法 (3.2 項)
素性拡張 2 (FE2) 有効な素性のみから素性ベクトルを拡張する手法 (3.3 項)

極性判定の分類器は Naive Bayes モデルを用いて学習した[1)]．

4.2 実験結果と考察

FE1 で実際に拡張された素性の例を表 2 に示す．ほぼ同じ意味を持ち，かつターゲットドメインに固有の素性が拡張されていることが確認された．例え

1) Support Vector Machine も試したが，結果に大きな差はなかった．

ば，美容ドメインにおける「可愛い」という素性から，日用品ドメインに固有の「おもろい」(表 2 の 5 行目) や，インテリアドメインに固有の「映える」(表 2 の 7 行目) が素性ベクトルに追加されている．

表 2　素性拡張の例

ソースドメイン	ターゲットドメイン	元の素性	拡張された素性
日用品	レディース	剥がれる	取り外せる
		粘着	塗布
		弱い	緩い
美容	日用品	可愛い	おもろい
		見える	映る
美容	インテリア	可愛い	映える
		ある	留まる

3.3 項で述べた手法 FE2 には 3 つのパラメタがある．まず，T_{sim} について，0.4 と 0.5 に設定した場合を比較した．大部分のドメインの組で 0.4 の場合の正解率が高かったので，$T_{sim} = 0.4$ と設定した．次に，訓練データの 5 分割交差検定により，T_{cor} は $\{50, 100, 200\}$，T_{fre} は $\{5, 10, 20\}$ の範囲の中から最適なものを選択した．ドメインの組毎に異なる値が選択されたことを確認した．

BL, FE1, FE2 による極性判定の正解率を表 3 に示す．矢印の左がソースドメインを表す記号，右がターゲットドメインを表す記号である．また，太字はそれぞれのドメインの組で一番正解率の高い値を表す．

まず，ソースとターゲットのドメインが異なる場合のほとんどについて，ドメインが同じときよりも正解率が低下することが確認された．ベースライン (BL) と本研究の手法 (FE1 もしくは FE2) を比べると，ドメインの組によって，本研究の手法の方が正解率が高い場合もあれば，そうでない場合もある．特に，ターゲットドメインが「美容・コスメ・香水」のときは，ベースラインの方が正解率が常に高い．単語の分散表現に基づく素性拡張は常に有効ではないことがわかった．ただし，20 のドメインの組のうち，BL が一番正解率が高いのは 8 組，FE1 もしくは FE2 が一番正解率が高いのは 13 組であるため，全体的には素性拡張は有効であると言える．

FE1 と FE2 を比較すると，同様にドメインの組によって優劣が異なる．ターゲットドメインが「インテリア・寝具・家具」のときは FE1 が，「レディースファッション」のときは FE2 が高い傾向が見られる．

表 3　実験結果

	食→食	レ→食	日→食	美→食	イ→食
BL	0.874	0.785	0.823	0.804	0.808
FE1	—	0.771	0.813	**0.810**	**0.811**
FE2	—	**0.790**	**0.825**	0.806	0.803

	食→レ	レ→レ	日→レ	美→レ	イ→レ
BL	0.795	0.854	0.854	0.821	**0.836**
FE1	0.804	—	0.851	0.820	0.831
FE2	**0.805**	—	**0.856**	**0.835**	0.834

	食→日	レ→日	日→日	美→日	イ→日
BL	0.793	0.781	0.814	**0.813**	0.793
FE1	0.794	**0.779**	—	0.803	**0.801**
FE2	**0.798**	0.776	—	0.804	0.785

	食→美	レ→美	日→美	美→美	イ→美
BL	**0.790**	**0.750**	**0.808**	0.830	**0.784**
FE1	0.779	0.735	0.803	—	0.771
FE2	0.776	0.748	**0.808**	—	0.783

	食→イ	レ→イ	日→イ	美→イ	イ→イ
BL	0.805	**0.820**	**0.830**	0.831	0.854
FE1	0.806	0.819	0.825	**0.835**	—
FE2	**0.814**	0.817	0.809	0.814	—

20 のドメインの組のうち，FE1 の方が正解率が高いのは 8 組，FE2 の方が高いのは 12 組であるため，全体的には，分類クラスとの相関が高い素性のみから新たな素性を拡張した方が効果的な領域適応であると言える．

5　おわりに

本論文は，事前学習した単語の分散表現を用いた素性拡張による教師なし領域適応の手法を提案した．商品レビューを対象とした極性判定のタスクについて，いくつかのソース・ターゲットドメインの組についてその有効性を示した．ただし，提案手法は常に有効であるわけではなく，正解率が向上しないドメインの組も多く存在した．今後は，提案手法が有効に働くドメインの組にはどのような性質があるかを明らかにしたい．例えば，ソース・ターゲットドメインで素性集合の重複の度合が小さいときには，提案手法による素性拡張が有効に働くといった性質があることを予想している．

参考文献

[1]森谷一至. 単語の分散表現に基づく極性判定のための教師なし分野適応. 修士論文, 北陸先端科学技術大学院大学, 3 2021.

[2]John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 120–128, 2006.

[3]Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 513–520, 2011.

[4]Shoushan Li and Chengqing Zong. Multi-domain adaptation for sentiment classification: Using multiple classifier combining methods. In *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering*, 2008.

[5]Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 751–760, 2010.

[6]Yftah Ziser and Roi Reichart. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1241–1251, 2018.

[7]Sitong Wang. Domain adaptation for gender classification of text. 修士論文, 北陸先端科学技術大学院大学, 9 2019.

[8]Masayuki Asahara. NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, Vol. 24, No. 1, pp. 7–22, 2018. https://www.gsk.or.jp/catalog/gsk2020-d.

[9]Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*, chapter 5.3.3, pp. 169–172. The MIT Press, 1999.

[10]Rakuten Institute of Technology. 楽天データ公開. https://rit.rakuten.co.jp/data_release_ja/. (2021 年 1 月閲覧).