属性に対する極性判定を対象とした領域適応

LU Binghan　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s2110196,kshirai}@jaist.ac.jp

概要

属性に対する極性判定は教師あり機械学習に基づく手法が主流だが，属性によって有効な素性は異なると考えられる．本論文では，属性をドメインとみなし，ある属性に関する (ソースドメインの) ラベル付きデータから別の属性の (ターゲットドメインの) 極性判定のモデルを学習する領域適応の手法を提案する．ターゲットドメインのラベル付きデータを自動構築するために，ソースドメインのデータから学習した BERT を用いて自動ラベル付けを行う手法と，ソースドメインの文に出現する感情語や特徴語をターゲットドメインのそれに置換する手法を組み合わせて用いる．

1　はじめに

属性に対する感情分析 (極性判定) は，製品などの評価対象の属性に対して表明された意見が肯定的か否定的かを判定するタスクである．教師あり機械学習の手法が用いられることが多いが，訓練データとテストデータでドメインが異なると極性判定の性能が低下する問題が知られている．ドメインとは一般にテキストのジャンルや媒体を指すが，属性に対する極性判定では異なる属性に対して同様の問題が起こりうる．例えば，レストランの属性として「料理」と「価格」があり，「料理」に対するラベル付きデータしか存在しないとき，これから学習した分類器を「価格」に対する極性判定に適用しても高い正解率が得られない．

領域適応は，十分なラベル付きデータが存在するソースドメイン (訓練データのドメイン) から得られた知識を，ラベル付きデータが全くあるいは少量しか存在しないターゲットドメイン (テストデータのドメイン) の分類器の学習に転移する技術である．本研究では，ドメインを属性とみなし，ある属性のラベル付きデータを利用して別の属性の極性を判定することを目的とする．そのため，ソースドメインのラベル付きデータからターゲットドメインのラベル付きデータを自動的に生成する手法を提案する．

2　関連研究

感情分析を対象とした領域適応の手法は多くの先行研究がある．Rietzler らは，ソースドメインのラベル付きデータで BERT[1] を fine-tuning する前に，ターゲットドメインのラベルなしテキストを用いて BERT の事前学習を再度行うことで領域適応を行う手法を提案した [2]．白らは，BERT の最上位層の埋め込み表現ではなく，一つ下の層の埋め込み表現をドメインに特化されていない単語の特徴ベクトルとみなし，この平均ベクトルを入力とする 3 層のニューラルネットワークを学習することで領域適応を行う手法を提案した [3]. Yu らは，Cross-Domain Review Generation(CDRG) という手法を提案した [4]. ソースドメインにおけるラベル付きレビューの属性語・感情語をターゲットドメインのそれに変換することで，ターゲットドメインのラベル付きレビュー文を自動生成し，これを BERT の fine-tuning に用いた．

上記の先行研究ではレビューのジャンルをドメインとした領域適応の手法が提案されていたが，本研究では属性をドメインとした領域適応を扱う．

3　提案手法

3.1　概要

提案手法の概要を図 1 に示す．ラベル付きデータが存在する属性をソースドメイン，存在しない属性をターゲットドメインとする．図中の (S),(T) はそれぞれソースドメイン，ターゲットドメインのデータを表す．

ターゲットドメインのラベル付きデータを 2 つの手法で自動生成する．ひとつは自動ラベル付け

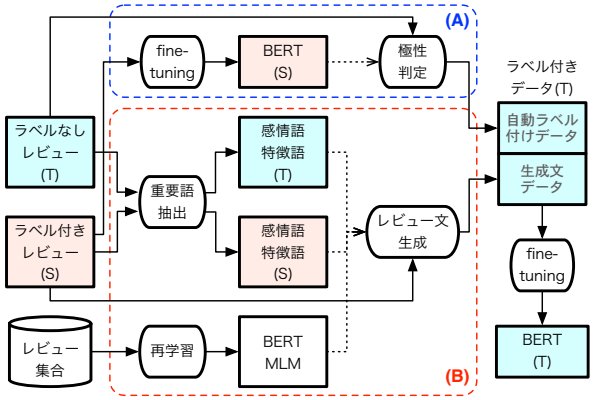

図 1　提案手法の概要

による手法 (図 1 (A)) である．詳細は 3.2 項で述べる．もうひとつは，CDRG[4] を拡張し，ソースドメインのレビュー文から新たにターゲットドメインのラベル付きデータを自動生成する手法 (図 1 (B)) である．本研究ではこれを Cross-Aspect Review Generation (CARG) と呼び，その詳細を 3.3 項で述べる．最後にターゲットドメインの極性判定モデルを学習する．詳細は 3.4 項で述べる．

3.2　自動ラベル付けによる訓練データ構築

ソースドメインの訓練データから事前学習済みの BERT を fine-tuning し，極性判定の分類器 (図 1 の BERT(S)) を得る．次に，これを用いてターゲットドメインの文の極性を判定し，ラベルを付与する．この際，極性の予測確率が T_p 以下のレビュー文は除外する．

3.3　Cross-Aspect Review Generation

3.3.1　感情語と特徴語の抽出

ソースドメインとターゲットドメインのそれぞれについて，そのドメイン (属性) のレビューで使われる感情語を抽出する．感情語とは “excellent”, “bad” など書き手の感情や評価を表す単語である．感情語の抽出には SentiWordNet[5] を用いる．SentiWordNet は，単語のそれぞれの語義について，肯定のスコアと否定のスコア (ともに 0~1 の値) が付与されている．また，語義は出現頻度の順にランク付けされている．単語 w の感情語スコア $SS(w)$ を式 (1) のように定義する．

$$SS(w) = \sum_{i=1}^{n} \frac{1}{n} \cdot |pos(w,i) - neg(w,i)| \qquad (1)$$

i は語義の出現頻度のランク，$pos(w,i)$ と $neg(w,i)$ はそれぞれ単語 w の i 番目の語義の肯定と否定のスコア，n は単語 w の語義の総数を表す．$SS(w)$ が閾値 T_s 以上の単語をドメイン固有の感情語として抽出する．

次に，特定のドメインだけによく使われる単語を「特徴語」と定義し，これを抽出する．例えばドメイン (属性) が food のときには “dinner” や “dessert” などが，service のときは “staff” や “waiter” などがそのドメインの特徴語となる．いま，いくつかの属性について，その属性について言及されたレビュー集合があると仮定する．各属性のレビュー集合を仮想的にひとつの文書とみなし，属性 a における単語 w の TF-IDF を計算する．あるドメイン a のレビュー集合のみに出現し，かつ TF-IDF の上位 T_d 件の単語をドメインの特徴語として抽出する．

3.3.2　レビュー文の生成

ソースドメインのラベル付きデータにおける感情語もしくは特徴語をターゲットドメインの感情語もしくは特徴語に置き換えることにより，ターゲットドメインのラベル付き文を新たに生成する．単語の置き換えには BERT の Masked Language Model (MLM) を利用する．あらかじめ，事前学習済みの BERT の MLM を初期のパラメタとし，大量のラベルなしレビュー文の集合を用いて MLM を再学習する．

レビュー文生成の手続きを以下に示す．

1. ラベル付きのソースドメインの文について，その中にソースドメインの感情語と特徴語が出現していれば，それを [MASK] に置換する．
2. [MASK] に対して左から順に以下の処理を実行する．
   (a) MLM によって [MASK] によって埋めるべき単語を予測する．元の単語が感情語ならターゲットドメインの感情語，特徴語ならターゲットドメインの特徴語のうち，MLM による予測確率が大きい T_k 個の単語を得る．[MASK] をその単語に置換し，T_k 個の新しい文を得る．
   (b) 文の数が組み合わせ的に増大することを避けるため，2 個目以降の [MASK] を置換するときは，$T_k \times T_k$ 個の文のうちスコア[1] が高い T_k 個の文を選択し，次の [MASK] の処理に移る．この処理は最後の [MASK] の

1)　文のスコアは，複数箇所の [MASK] を置換した単語の MLM による予測確率の和とする．

ときには行わず，最終的に最大で $T_k \times T_k$ 個の文を得る．

上記の文生成手法は CDRG[4] を元にしているが，CDRG では属性語を置換するのに対し，本研究では特徴語を置換する点，CDRG では 1 つの文からターゲットドメインの文を 1 つ生成するのに対し本研究では最大で $T_k \times T_k$ 個の文を生成する点が異なる．

3.3.3 生成文のフィルタリング

生成された文の自然さを擬似対数尤度スコア PLL(pseudo-log-likelihood score)[6] によって測る．PLL は，文中の各単語 w_t を BERT の MLM で予測した確率の対数尤度の和である．

$$\log P_{MLM}(W) = \sum_{t=1}^{|W|} \log P_{MLM}(w_t | W_{\setminus t}) \qquad (2)$$

自動生成されたレビュー文のうち，$\log P_{MLM}(W)$ が閾値 T_f より小さい文を削除する．

3.4 極性判定モデルの学習

3.2 項と 3.3 項で生成したラベル付きレビューを用いて BERT を fine-tuning し，ターゲットドメインの極性判定モデル (図 1 の BERT(T)) を得る．

後述する実験で用いるデータセットには，極性ラベルの分布に偏りが見られる．この問題に対処するため，本研究では Focal Loss[7] を損失関数として使用する．二値分類のときの Focal Loss の定義を式 (3) に示す．

$$FL(p_t) = -(1 - p_t)^{\gamma} \cdot \log(p_t) \qquad (3)$$

ここで γ はハイパーパラメタである．データセットの中で小さな割合しか占めないデータのクラスは，その予測確率 p_t は小さくなる傾向がある．通常の cross entropy の式に $(1 - p_t)^{\gamma}$ を加えることにより，少数の極性クラスについては損失が大きく見積もられ，その分類エラーが大きく評価される．

4 評価実験

4.1 実験設定

評価実験には 2 つのデータセットを使用する．ひとつは SemEval-2014 Task 4 Aspect Based Sentiment Analysis のレストランのデータセット [8] である．レストランに関するレビュー文に対し，属性のカテゴリとそれに対する極性クラスが付与されている．属性は “service”, “food”, “price”, “ambience”, “anecdotes” の 5 つであり，本実験ではこれをドメインとする．極性クラスは “positive”, “negative”, “conflict”, “neutral” の 4 つである．データセットの詳細を表 1 に示す．positive のレビューが多く，conflict や neutral (ただし “anecdotes” を除く) は少ない．

表 1 レストラン・データセットの詳細

	service	food	price	ambience	anecdotes
positive	127	542	49	127	472
negative	55	138	46	55	167
conflict	33	54	7	33	24
neutral	18	62	6	18	326
total	336	796	108	233	989

もう一つは Laptop ACOS(Aspect-Category-Opinion-Sentiment) のデータセット [9] である．Amazon に投稿されたラップトップパソコンに関するレビューに対し，評価対象の属性と極性が付与されている．極性カテゴリは “postive”, “negative”, “neutral” の 3 つである．本実験では，付与されている属性を “general”, “design”, “performance”, “quality” の 4 つに人手で分類し，これを属性のドメインとみなした．データセットの詳細を表 2 に示す．このデータも neutral のレビューが極端に少ないという極性クラスの偏りが見られる．

表 2 ラップトップ・データセットの詳細

	general	quality	performance	design
positive	194	285	156	85
negative	468	114	152	139
neutral	22	15	11	31
total	684	414	319	255

全ての属性の組について，一方をソースドメイン，もう一方をターゲットドメインとして，後者のデータセットに対する極性判定の正解率を測る．以下の 5 つの手法を比較する．

ベースライン 1 (BL1) ソースドメインのラベル付きデータを用いて BERT を fine-tuning する手法．領域適用をしない手法である．

ベースライン 1 (BL2) 3.2 項で述べた手法で得られたターゲットドメインの擬似ラベルを用いて BERT を fine-tuning する手法．

CARG1 BL2 の擬似ラベル文と CARG によって生成されたレビュー文を訓練データとして BERT を fine-tuning する手法．

CARG2 CARG1 に加え，3.3.3 で述べた生成文のフィリタリングを行う手法．

CARG3 CARG2 に加え，3.4 で述べた Focal Loss を使って BERT を fine-tuning する手法．

表 3　レストラン・データセットにおける極性判定の正解率

(source)	S				F				P				Am				An				平均
(target)	F	P	Am	An	S	P	Am	An	S	F	Am	An	S	F	P	An	S	F	P	Am	
BL1	.776	.667	.687	.533	.765	.574	.674	**.607**	.679	.697	.472	.421	.771	.758	.620	.526	**.577**	.707	.537	**.614**	.633
BL2	.795	.676	.687	.537	.771	.620	.687	**.607**	.711	.735	.459	.446	.762	.794	.639	.540	.574	.730	.528	.597	.645
CARG1	.798$^{*}_{+}$	.694	.682	.540$^{*}_{+}$	.783$^{*}_{+}$	.630*	**.704***	.585	.708	.731*	.472	**.461**	.786	.789$^{*}_{+}$	.630	.531$^{*}_{+}$	.571	**.756***	.546	.597	.650
CARG2	.791$^{*}_{+}$	**.731**	**.691**	.541$^{*}_{+}$	.802*	**.657***	.691*	.590	.705	**.754**$^{*}_{+}$	.464	.433	**.798**$^{*}_{+}$	.794$^{*}_{+}$	**.639**	**.558**$^{*}_{+}$	.563$^{*}_{+}$	.727	**.574**$^{*}_{+}$	.597	.655
CARG3	**.799**$^{*}_{+}$	.713	**.691**	**.542**$^{*}_{+}$	**.815***	**.657***	.687*	.602	**.717**	.747$^{*}_{+}$	**.476**	.451	**.795**$^{*}_{+}$	.795$^{*}_{+}$	**.639**	.556*	.574	.731*	**.574**$^{*}_{+}$	.605	**.658**

S: service, F: food, P: price, Am: ambience, An: anecdotes.
*, + は，それぞれ BL1, BL2 と比べて，マクネマー検定によって $p < 0.05$ で有意差があることを表す．

表 4　ラップトップ・データセットにおける極性判定の正解率

(source)	G			Q			P			D			平均
(target)	Q	P	D	G	P	D	G	Q	D	G	Q	P	
BL1	.807	.784	.765	.741	.837	.769	**.789**	.850	.773	.737	.821	.809	.790
BL2	.804	**.809**	.757	.731	.850	.765	.773	.862	.788	.756	.831	.803	.794
CARG1	.807	.803	**.780***	.744+	.850*	.769	.768	**.865***	.788	.770*	**.833***	.806	.799
CARG2	.812	.800	**.780***	**.746+**	**.853***	.765	.779	.862*	**.796***	**.772**$^{*}_{+}$	.829*	**.815**	**.801**
CARG3	**.816**	.800	**.780***	.744+	.846	**.773**	.779	.855	.788	.768*	.829*	**.815**	.800

G: general, Q: quality, P: performance, D: design　*, + は表 3 と同じ．

レストラン・データセットについては，CARG を用いる際，BERT の MLM を再学習するために 3 万件の Yelp のレストランレビュー [10] を使用した．ラップトップ・データセットについては事前学習済みの BERT の MLM をそのまま用いた．実験パラメタの詳細を付録 A に示す．

4.2　実験結果と考察

レストラン・データセットの実験結果を表 3 に示す．太字はそれぞれのドメインの組において正解率が最も高いことを表す．ベースライン手法 (BL1,BL2) と提案手法 (CARG1~3) を比較すると，20 のドメインの組のうち 17 組について，提案手法の方が優れている．このことから，ターゲットドメインのラベル付きデータを自動生成する CARG の有効性が確認できる．CARG1 と CARG2 を比較すると，全 20 組のうち 12 組について CARG2 の正解率が高いことから，自動生成された文のうち生成確率の低い文をフィルタリングする手法はある程度有効であると言える．CARG2 と CARG3 を比較すると，全 20 組のうち 15 組については CARG3 の正解率が同じもしくは高い．今回の実験では Focal Loss の導入が効果的であった．

ラップトップ・データセットの実験結果を表 4 に示す．レストラン・データセットの結果と同様に，ベースラインよりも提案手法の方が優れている．ただし，統計的に有意差があるドメインの組はレストラン・データセットに比べて少ない．レストランの実験では Yelp のレビューを用いて MLM を再学習したのに対し，ラップトップでは事前学習した MLM をそのまま用いたことが原因のひとつとして考えられる．また，CARG1 よりも CARG2 の方が正解率が高い傾向にあることが確認できるが，CARG3 は CARG2 と比べて同等もしくはやや劣る．2 つの異なるデータセットでおおよそ同様の結果が得られたことから，提案手法がどのようなジャンルのレビューにも適用できるという意味での汎用性を有することが示唆される．

CARG による文生成の例を付録 B に示す．

5　おわりに

本論文は，属性に対する極性判定において，異なる属性のラベル付きデータを利用して極性判定の分類器を学習する領域適応の手法を提案した．評価実験では，全体的には提案手法はベースラインを上回ったが，属性の組によっては極性判定の正解率が向上しなかった．今後の課題として，この原因を精査し，任意のドメインの組において自動生成する文の品質を向上させる手法を探究したい．例えば，属性に関連する語として特徴語を抽出しているが，属性の特徴を明示的に示す単語だけでなく暗黙的に示す語も抽出することを検討する．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[2] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In **Proceedings of the 12th Conference on Language Resources and Evaluation(LREC 2020)**, pp. 4933–4941, 2020.

[3] 白静, 田中裕隆, 曹類, 馬ブン, 新納浩幸. Bert の下位階層の単語埋め込み表現列を用いた感情分析の教師なし領域適応. 情報処理学会研究報告, Vol. 2019-NL-240, No. 17, pp. 1–6, 2019.

[4] Jianfei Yu, Chenggong Gong, and Rui Xia. Cross-domain review generation for aspect-based sentiment analysis. In **Findings of the Association for Computational Linguistics**, pp. 4767–4777, 2021.

[5] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In **Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)**, pp. 2200–2204, 2010.

[6] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2699–2712, 2020.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. **arXiv:1708.02002**, 2017.

[8] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 task 4: Aspect based sentiment analysis. **Association for Computational Linguistics**, Vol. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 27–35, 2014.

[9] Hongjie Cai, Rui Xia, and Jianfei Yu. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 340–350, 2021.

[10] Yelp dataset, (2022 年 12 月閲覧). https://www.yelp.com/dataset.

A　実験時のパラメタ

実験時に設定したパラメタは以下の通りである．

- 自動ラベル付けにおける予測確率の閾値 T_p を 0.8(レストラン) または 0.5(ラップトップ) に設定した．
- 感情語抽出の極性スコアの閾値 T_s を 0.3 に設定した．
- 特徴語抽出の際の件数 T_d を 100 に設定した．
- CARG における文生成数のパラメタ T_k を 100 に設定した．
- 生成文のフィルタリングの閾値 T_f を -30 に設定した．
- Focal Loss のハイパーパラメタ γ を 2 に設定した．

B　CARG による文生成の例

CARG によって生成された文の例を表 5 に示す．下線は置換された単語を表す．評価語が “good” や “great” など food の評価に良く使われるものに置換され，また “service” が “food” に置換されており，ターゲットドメイン (food) のレビューらしい文が生成されている．

表 5　CARG による生成文の例

Domain	Review sentence
service (source)	The service was attentive and her suggestions of menu items was right on the mark
food (target)	the food was good and her choice of menu items was great on the menu