データ拡張による感情分析のアスペクト推定

西本慎之介, 能地 宏, 松本 裕治

{ nishimoto.shinnosuke.nh5, noji, matsu}@is.naist.jp

1 はじめに

自然言語処理においてテキスト分類は, 幅広く研究され, 応用されてきた. とくに感情分析はテキスト分類の応用として著名な分野の一つである. 感情分析とは与えられた文章に対して, その文章の極性 (肯定的か否定的か) を付与するものである. この論文では, 感情分析の中でもアスペクトベースの感情分析について扱う. このタスクでは, 与えられた文章に対して極性を付与するだけでなく, その文章のアスペクトも付与する. たとえば *fried rice here is amazing.* という文が与えられた場合 *positive* という極性に加え, *food* というラベルを付与する.

これまでの研究では, この分野では教師あり機械学習が高い精度を出してきた. 特に近年ではニューラルネットワークベースの研究が高い精度を出している. しかしながら, ニューラルネットワークに基づく手法は高い表現力の代わりに過学習の恐れがあり, 一般に高い精度を得るためには大量の訓練データが必要になることが知られている.

本研究では限られたラベル付きデータしか利用できない状況での感情分析の精度向上を目指し, 擬似的な正解データを生成する手法であるデータ拡張の有効性を検証する. データ拡張は画像処理の分野ではよく研究されているものの, 自然言語処理分野での適用事例はまだ少ない. 本研究ではアスペクトベースの感情分析のタスクである SemEval 2015 の shared task 12 を対象に様々なネットワークの構造とデータ拡張法の組み合わせを検討し, 特に単語ベクトルやシソーラスに基づく類似単語の置き換えによるデータ拡張と, 語順を考慮する LSTM の組み合わせが有効で, 既存研究を上回る性能が得られることを示す.

2 関連研究

2.1 ニューラルネットワークに基づく感情分析

まずはじめに SemEval 2015 shared task 12 のアスペクトベースの感情分析の先行研究について述べる. 最高精度を達成した Toh ら [7] の手法は feedforward ネットワークを用いており, 単語の分散表現を用いず, 言語モデルなどの素性ベクトルを入力として利用している. また, Wang ら [8] は, 2 層の feedfoward ネットワークと単語の分散表現を用いることで高い精度を報告している. これらはどちらも feedforward ネットワークを用いているため, モデルは単純といえる. shared task はデータ量が少なく (3 節参照), LSTM などの強力なモデルでは過学習しやすいことがこの原因と考えられる. 本研究ではデータ拡張によってこの問題を回避することを目指す.

2.2 自然言語処理におけるデータ拡張

自然言語処理のタスクにおいてデータ拡張が用いられた例は, 単語ベクトルの類似度に基づくもの, シソーラスを使用したもの, ルールに基づくものなどが存在する. Wang [9] らは, トピックモデルを用いた文分類を行う際, 単語の分散表現に基づいて訓練データの単語を近傍の単語と入れ替え, 擬似的な訓練データを生成している. 同様に Xiang ら [12] は, シソーラスを用いて訓練データの事例の単語を入れ替えることで, 擬似的な正解データを生成している. これらの研究は離散的な素性を用いて分類を行っており, ニューラルネットのモデルにはなっていない.

また過去の研究では, どのようなデータ拡張が有効であるかを統一的に比較したものは見当たらない. 本研究では, どのようなネットワークの構造に対しデータ拡張が有効か, そしてどのようなデータ拡張の方法が特に有効か, という問いに答えることを目指す.

3 タスクと評価

SemEval において, 2014 年度からアスペクトベースの感情分析が共有タスクとして設定されてきた [6]. Liu [2] と Zhang ら [11] によれば, アスペクト (カテゴリー) は, 対象となる製品やその製品の属性として定義される. 2015 年度の共有タスク [5] では, アスペクトカテゴリーは, エンティタイプ E と属性タイプ A の

表 1: データの統計量

	訓練データ	テストデータ	ラベル数
Laptops	1974	949	198
Restaurants	1654	845	30

組み合わせによって定義される. E は評価の対象とされている製品そのものか, その一部である. たとえば Laptop のドメインにおいては, *laptop* や, *battery* などが E となる. 一方で A は E の属性で, *durability* や *quality* などが挙げられる. E と A は, ドメイン毎にあらかじめ与えらており, 必ずしも, 文の中に名前が出て来るとは限らない. たとえば, *They sent it back with a huge crack in it and it still didn't work; and that was the fourth time I've sent it to them to get fixed.* という文章には, *customer support* (E), *quality* (A) というラベルが付与されている. タスクの目的は, ある文章が与えられた与えられたとき, E と A のペアを当てることである. ドメインとして, restaurant と laptop が与えられており, 表 1 にそれぞれのデータの統計量を示す. このように訓練事例が 2000 弱ほどと多くなく, 複雑なニューラルネットに基づく手法では過学習の恐れがある.

4 実験で用いるニューラルネットワークのモデル

本研究では, 以下に述べる三種類の代表的なニューラルネットワークのモデルに対し, データ拡張 (5 節) を適用し影響を調べる.

4.1 フィードフォワードニューラルネットワーク

実験では, 二層の feedforward 型ニューラルネットワーク (以下, FFNN) を用いる. 第一層は, 全結合層であり, 第二層はソフトマックス関数で. 確率値を出力とする. トレーニングデータが 2000 文弱と少ないため, 単語分散表現を学習することは難しい. そこで, Google News corpus（約 1000 億単語）から word2vec [3] で学習された, 著者のウェブサイトで配布されている 300 次元の単語ベクトル[1]を用いる. 隠れ層の次元は 19 次元とする.

入力となる文ベクトルは単語ベクトルの平均値とする. 出力される確率値は 19 個のラベルに対応する. 学習に際してマルチクラスの negative entropy loss を用いる.

[1]https://code.google.com/archive/p/word2vec/

4.2 LSTM

Recurrent Neural Network(RNN) は系列データを扱うためのモデルであり，前時刻の中間層を現時刻の入力としても用いることで，内部状態を保持しながら学習を行う．しかし，通常の RNN は逆誤差伝播による学習を行う際，勾配が減衰するという問題 (勾配消失) が存在する.

Long short-term memory(LSTM) [1] は勾配消失の問題を解決するために提案されたネットワークの 1 つである．本研究では各単語を one-hot ベクトルで表したものを入力とした. 隠れ層の次元は 128 とした.

4.3 Convolutional Neural Networks モデル

本稿で用いる CNN モデルではテキストを単語の出現順序に応じた時系列データと見なし, 処理する. 本研究では, ウィンドウ幅を 3 とする．素性フィルタには 250，全結合層は 2 層用意し，第 1 層の隠れ変数の次元には 250 を用い，第 2 層の次元には学習ラベル数を用いた.

5 実験で用いるデータ拡張の手法

5.1 単語の分散表現を用いるデータ拡張

単語の分散表現を用いたデータ拡張の方法では, あらかじめ学習された単語ベクトルを利用する. 具体的には, 訓練データ中の単語をそれと近いベクトルを持つ他の単語に置き換える. ベクトル間の類似度はコサイン類似度で計算する.

単語を置き換えた文と元のラベルのペアを, 新しい訓練事例とする. 例えば, *Being late is terrible* という文のそれぞれの単語について近傍の単語を探索し, 置き換える. それにより *Be behind are bad* というような新しい訓練データを生成する. このように, 単語の分散表現に基づき生成されたデータは一般に非文である.

この手法を用いて, 訓練データを 3 倍に拡張する場合, 文中の全ての単語について近傍の単語 2 個 (それぞれ,k=1,2 とする) を探索し, 訓練データの単語を k=1 の単語と入れ替えたもの,k=2 の単語と入れ替えたものをそれぞれ元のラベルと組み合わせて新しい訓練データする.

5.2 シソーラスを用いるデータ拡張

シソーラスを用いる方法では, 訓練データの単語を入れ替えるために, WordNet [4] を用いる. 訓練データの単語からランダムに選ばれた単語が入れ替えられる. 入れ替える単語について WordNet に登録されている synset の中からランダムに選択して入れ替えることで新たな訓練データを生成する.

5.3 ルールを用いるデータ拡張

最後に, ルールに基づいたデータ拡張の手法について説明する. 本手法では Young ら [10] の研究を参考に, 以下の 3 つのルールにもとづいて訓練データの単語を入れ替えることで, 正解データを生成する.

1. **形容詞表現の削除:“red shirt” → “shirt”**
2. **形容詞表現の削除:“run quickly” → “run”**
3. **wordnet を用いて名詞を上位語で置き換える: “red shirt” → “red clothing”**

6 実験

6.1 タスクの前処理

3 節で説明した公式タスクにおいては, データセットの中のラベルの出現には偏りがある. たとえば Laptop のドメインでは, 訓練データ中の 80.7%のアスペクトが最頻出頻出の 17 個のラベルに限定されている. トレーニングデータ中に頻度が小さいものを予測するのは難しいため, 本論文では, ラベル数を 19 値に縮減して実験を行う. 最頻出 17 ラベルに加えて, それ以外のラベルを全て *Other* というラベルにする. また, 一部の事例はアスペクトを含んでおらず, これらには *NONE* というアスペクトを付与する. これは shared task の元の設定とは異なるが, 現在の最新の研究である Wang ら [8] と同一のものである.

6.2 実験設定

Keras[2]を用いて 4 節で説明した FFNN, LSTM と CNN のモデルを実装した. 前節で説明したようにタスクの前処理を行い 19 値分類のマルチクラス分類問題として Laptop のドメインのみを用いる.

実験 1 では, どのようなネットワークに対しデータ拡張が有効かを検証するため, データ拡張として単語の分散表現を用いた場合の各ネットワークの性能を調べる. ここでは, 5.1 節で説明した単語分散表現の手法を使用し, 訓練データを 10 倍に増やし, 元の 1974 文のデータを 19740 文にした. そして増加させたデータに対して FFNN, LSTM, CNN のモデル毎の違いを比較する. 評価は F 値を用いて結果を測定する.

なおデータ拡張の際の単語の分散表現については, 4.1 節で説明したのと同様に Google News corpus（約 1000 億単語）から word2vec [3] を使用し, 学習する.

実験 2 では, ネットワークを固定し, 異なるデータ拡張の手法を比較を行う. アーキテクチャとして LSTM を用いる. 5 節で説明した, 単語の分散表現, シソーラス, ルールを用いて, 6800 文にデータ拡張を行う.

表 2: 実験 1 の結果

	FFNN	LSTM	CNN
1974 文	51.3	66.7	62.1
19740 文	48.9	72.0	65.1

6.3 実験結果

実験 1 と実験 2 の結果を表 2, 表 3 にそれぞれ示す. FFNN で精度の悪化が見られ, LSTM と CNN で精度の向上が確認された. 特に LSTM では大きな精度向上が見られた.FFNN で精度が向上しなかった原因は, FFNN は語順を考慮しない bag-of-words のモデルであり, また今回置き換えがネットワークへの入力と同じ単語ベクトルであるため, 新たに生成された文でも入力ベクトルは元の文とあまり変わらないためであると考えられる.

それに対し LSTM と CNN では精度が向上しており, 語順を考慮することで類似単語への置き換えが有効に働くようになることが見てとれる.

なお, FFNN の 1974 文の結果は先行研究である Wang らの数値と同一であり, 本研究はそれよりも大幅な精度向上を果たしている.

実験 2 ではシソーラスを用いた場合に最も高い F 値を得た. 5.1 節で述べたように, 単語の分散表現とシソーラスの一つの違いは, 単語の分散表現では非文が生成されやすいということで, 正しい文が生成されやすいシソーラスの方が学習がうまくいきやすい, ことが考えられる.

[2]https://keras.io/

表 3: 実験 2 の結果

	F 値
ベースライン	66.7
単語の分散表現	68.2
シソーラス	70.3
ルール	67.5

7 おわりに

本研究では，感情分析における文カテゴリー推定タスクに対して，有効なニューラルネットワークとデータ拡張の手法の組み合わせを比較検討した．同じデータ拡張手法を用いた場合, 他のネットワークに比べて LSTM モデルの有効性が確認された．また同じネットワークを用いた場合, シソーラスを用いたデータ拡張の有効性が確認された．今後の課題としては, 別の自然言語処理のタスクの応用などが考えられる．

参考文献

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780, November 1997.

[2] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[4] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, Vol. 3, pp. 235–244, 1990.

[5] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pp. 486–495, 2015.

[6] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014.*, pp. 27–35, 2014.

[7] Zhiqiang Toh, Fusionopolis Way, and Fusionopolis Way. NLANGP : Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction. Vol. 14, No. SemEval, pp. 496–501, 2015.

[8] Bo Wang and Min Liu. Deep Learning for Aspect-Based Sentiment Analysis. pp. 1–9, 2015.

[9] William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2557–2563, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[10] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations : New similarity metrics for semantic inference over event descriptions. Vol. 2, pp. 67–78, 2014.

[11] Lei Zhang and Bing Liu. Aspect and entity extraction for opinion mining. In *Data mining and knowledge discovery for big data*, pp. 1–40. Springer, 2014.

[12] Xiang Zhang and Yann LeCun. Text Understanding from Scratch. *APL Materials*, Vol. 3, No. 1, p. 011102, 2015.