日本語エモティコンに基く Pre-training を用いた Convolutional Neural Networks の改善

三浦 康秀　大熊 智子
富士ゼロックス株式会社 研究技術開発本部コミュニケーション技術研究所
{yasuhide.miura, ohkuma.tomoko}@fujixerox.co.jp

1 はじめに

近年，画像処理や音声処理での成功を受けて深層学習を用いた自然言語処理への注目が高まっている [6]．深層学習を用いることにより，様々なタスクにおいて特徴量の設計を単純化しつつ高い性能が得られることが報告されている．深層学習では様々なモデルが用いられているが，中でも Convolutional Neural Networks (CNN) は 2〜3 層の単純なモデルにより，評判分析および質問応答のタスクにおいて既存の手法と同等以上の性能が得られている [2]．本稿では，CNN に大規模データを用いた Pre-training を取り入れて評判分析の改善を試みる．評判分析タスクはレビュー，ソーシャルメディア等を対象に長年研究されており，様々な深層学習モデルの適用が行われている．

画像処理においては，大規模な物体認識データセットである ImageNet で Pre-training した CNN が類似したタスクの性能向上に活用できることが知られている [5]．自然言語処理で CNN を用いるときにも同様の Pre-training を行うことが考えられるが，類似したタスクのラベル付きデータを大量に用意することは必ずしも容易ではない．一方，自然言語処理においては弱い手掛かりを用いる Distant Supervision に基づく手法が提案されている [3, 4]．弱い手掛かりとしては大規模知識源への含有 (関係抽出) やエモティコンの出現 (評判分析) が用いられ，人手では困難な大量のラベル付きデータの自動収集が可能になる．

Severyn ら [4] は Distant Supervision に基く Pre-training の CNN への導入を提案し，評判極性が明示的なエモティコンを弱い手掛かりとして用いることにより，英語の評判分析の性能が向上することを報告している．本稿では Severyn らの手法を用いて日本語の評判分析の性能向上に取り組み，以下の 2 点を検証した：

1. エモティコンによる Pre-training の効果はエモティコンへの評判極性を付与しなくても得られる．
2. 人手で作成されたラベル付きデータが少ないときに Pre-training の効果が強まる．

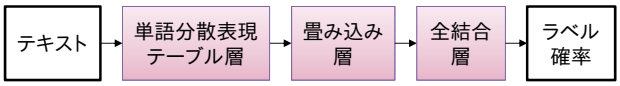

図 1: CNN モデルの概要．影付き要素は学習可能な要素を意味する．

本稿では，まず 2 章において CNN に基く評判分析手法および Distant Supervision の導入手順を説明する．次に 3 章で評判分析の評価実験で利用するデータについて述べ，4 章で提案手法の効果を確認するために実施した実験の詳細を述べる．最後に 5 章では実験結果の考察および今後の展望を述べる．

2 手法

本稿では CNN モデルに Distant Supervision に基く Pre-training を導入した手法を用いる．2.1 節では本手法で用いる Collobert ら [1] の CNN モデルを説明し，2.2 節では Severyn ら [4] の Distant Supervision に基く Pre-training の導入手順を述べる．

2.1 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，図 1 に示す手順で処理する．

単語分散表現テーブル層

入力された各単語を d 次元の単語分散表現に変換する．単語分散表現は乱数で初期化した値もしくは事前に別データで学習したものを用いる．

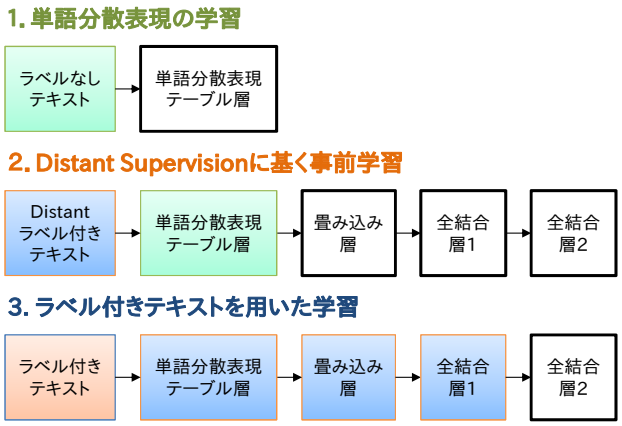

図 2: Distant Supervision に基く Pre-training の導入手順．3 段階の手順であり，緑色が第 1 段階，橙色が第 2 段階，青色が第 3 段階を意味する．第 2 段階および第 3 段階の学習では，ひとつ前の学習結果を部分的に利用する．

畳み込み層

単語分散表現列 $\mathbf{x}$ に対して，ウィンドウ幅 w_c の畳み込み処理を行う．時刻 t における畳み込み処理は式 1 で表せる．

$$\mathbf{o}(t) = \sum_{j=1-t}^{n_w-t} \mathbf{L}_j \cdot \mathbf{x}_{t+j} \tag{1}$$

ここで $\mathbf{L}_j \in \mathbb{R}^{n_f \times d}$ $(-n_w \leq j \leq n_w)$ はパラメータ，n_f は素性フィルタ数，n_w は入力単語数，d は単語分散表現の次元数である．畳み込み処理結果に対しては時間方向の max-pooling 処理および非線形関数の適用を行い，次元 n_f の出力を得る．

全結合層

入力を隠れ変数次元 n_{hu} に線形変換する処理を行う．全結合層は容易に多層化でき，4 章で述べる評価実験においても複数の層を用いる．線形変換処理結果に対しては Softmax 関数 $f(x) = \frac{a_i}{\sum_j \exp(a_j)}$ を適用し，モデルの出力としてラベル毎の確率値を得る．

2.2 Convolutional Neural Networks モデルへの Distant Supervision の導入

CNN モデルへの Distant Supervision の導入を，図 2 に示す 3 段階の学習により行う．

第 1 段階: 単語分散表現の学習

CNN モデルでは入力の単語列を単語分散表現テーブル層により単語分散表現列に変換する．単語分散表現は，事前に大規模データで教師なし学習を行った結果を用いることにより性能向上が得られることが知られている [1].

第 2 段階: Distant Supervision に基く Pre-training

CNN モデルの事前学習として，弱い手掛かりにより収集した大量のラベル付きデータを用いた CNN の学習を行う．CNN モデルのパラメータは乱数により初期化した値を用いるが，単語分散表現テーブル層については第 1 段階で学習した結果で初期化する．図 2 の例であれば，Pre-training においては単語分散表現テーブル層，畳み込み層，全結合層 1〜2 を学習する．

第 3 段階: ラベル付きデータを用いた学習

CNN モデルの学習をラベル付きデータを用いて行う．CNN モデルのパラメータは全結合層の最終層 (図 2 の例であれば全結合層 2) のみ乱数で初期化し，それ以外は第 2 段階の学習で得られた値を用いる．第 2 段階で得られたパラメータを初期化に用いることにより，Distant Supervision に基く Pre-training を取り入れたラベル付きデータの学習を行うことができる．

3 データ

本稿で用いる手法では 3 段階の学習を行う．3.1 節では第 1 段階で用いる単語分散表現の学習データについて説明する．3.2 節では第 2 段階の CNN の Pre-training で用いる Distant Supervision データについて述べ，3.3 節では第 3 段階の CNN の学習で用いる人手によるラベルが付けられたデータについて述べる．なお，4 章で述べる評価実験においてはツイートを対象とした評判分析を行い，3 種類のデータはツイートデータを基に作成している．

3.1 単語分散表現学習データ

ツイートを対象とした単語分散表現を構築するために，Twitter よりツイートデータを以下の手順で収集した．

ステップ 1 2014 年に投稿された日本語ツイートを Streaming APIs を用いて収集．

ステップ 2 リツイートではないかつ bot ではない [1] と判定したツイートを抽出．

結果として，ラベルなしの約 1.46 億ツイートを収集した．

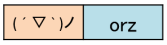

図 3: ポジティブ (橙色) とネガティブ (青色) の評判極性を付与した 2 個のエモティコン.

(　´_ゝ`)	(;∀;)	(^ω^)	(´-ω-`)	(´^ω^`)
(*´∇`*)	(T_T)	(ToT)	(^ν^)	∩^ω^∩

図 4: 評判極性を付与していない 10 個のエモティコン.

3.2 Distant Supervision データ

エモティコンを弱い手掛かりとして，評判分析タスクの Distant Supervision データを以下の手順で作成した.

ステップ 1 ウェブ上で公開されている複数の顔文字辞書を統合し，文字長 10 以下で他の顔文字の部分文字列とならない 1627 エモティコンのリストを作成.

ステップ 2 3.1 節で収集したツイートの中から，単語長が 5 以上でエモティコンが 1 回出現するツイートを抽出.

ステップ 3 出現頻度が上位である 12 エモティコンを抽出.

結果として得られた 12 のエモティコンは，図 3 に示すポジティブとネガティブの評判極性を付与したセット (評判極性付与あり) と，図 4 に評判極性を付与していないセット (評判極性付与なし) に分割した．評判極性付与ありセットでは，エモティコン毎に学習 5 万，開発 5000，評価 5000 の合計 12 万ツイートをステップ 2 のデータから抽出した．評判極性付与なしセットでは，エモティコン毎に学習 1 万，開発 1000，評価 1000 の合計 12 万ツイートを同データから抽出した.

3.3 ラベル付きデータ

評判分析タスクのラベル付きデータとして，クラウドソーシングを利用して以下の手順で作成した.

ステップ 1 2014/2/17 および 2014/2/18 に投稿された日本語ツイートのうち bot ではないと判定したツイートを収集.

ステップ 2 評価極性を持つ語が含まれるツイートを抽出．語の評価極性の判断には SentiWordnet[2] と日本語 WordNet[3] を用い，$score_{positive} \geq 0.3$ *or* $score_{negative} \geq 0.3$ で判定.

ステップ 3 Yahoo!クラウドソーシング [4] のタスクとして，ポジティブ，ネガティブ，ニュートラルのタグを 10 重複で付与.

ステップ 4 アノテーション結果より，各ツイートごとに付与数が最大となるタグを教師信号として設定.

セット	ポジティブ	ネガティブ	ニュートラル
学習	10770	9000	19159
開発	827	683	1490
評価	865	676	1459

表 1: ラベル付きデータの各セットにおける評判極性ラベルの数.

結果として 44915 のラベル付きテキストが得られ，38929 ツイートを学習セット，3000 ツイートを開発セット，3000 ツイートを評価セットに設定した．表 1 に各セットのポジティブ，ネガティブ，ニュートラルのデータ数を示す.

4 実験

4.1 モデルの設定

単語分散表現の学習

評価実験においては，3.1 節で述べたデータを用いて事前に単語分散表現を学習した．単語分割には Kuromoji[5] を用い，ユーザ名および URL の固有文字列への正規化 (それぞれ USER および URL)，ハッシュタグを 1 語として切り出すための前処理を行った．学習には word2vec[6] を用い，skip-gram アルゴリズムで分散表現の次元数 300，ウィンドウ幅 10，ネガティブサンプル数 15 のパラメータを用いた.

CNN モデルのパラメータ

ウィンドウ幅には $w_c = 3$ を用いた．素性フィルタには $n_f = 600$，全結合層は 2 層用意し，第 1 層の隠れ変数の次元には $n_{hu1} = 300$ を用い，第 2 層の次元には学習ラベル数を用いた.

[1] 人手で作成した 80 件の投稿クライアントのリストと照合して判定.

[2] http://sentiwordnet.isti.cnr.it/
[3] http://compling.hss.ntu.edu.sg/wnja/
[4] http://crowdsourcing.yahoo.co.jp/
[5] http://www.atilika.org/
[6] https://code.google.com/p/word2vec/

ラベル付き データ量	Pre-train なし	Pre-train 評判極性 付与あり	Pre-train 評判極性 付与なし
1000	55.20	65.60	67.00
2000	62.97	65.93	67.93
5000	69.10	69.97	70.07
10000	69.50	70.50	70.83
20000	72.30	71.73	72.60
38929	72.60	73.20	73.50

表 2: 2 種類のエモティコンセットにおいてラベル付きデータの量を変化させた場合の評判分析性能の評価結果．評価値は Accuracy である．

CNN の学習アルゴリズム

CNN モデルのパラメータ最適化は確率的勾配降下法と誤差逆伝播法により行う．パラメータの更新規則には AdaDelta を定数 $c = 3$ の max-norm 規格化と合わせて用いた．AdaDelta のパラメータは $\rho = 0.95, \epsilon = 1e^{-6}$ に設定した．過学習の影響を抑制するため，Dropout を畳み込み層および全結合層 1 の出力に対して適用した．学習は 20 epoch 実施し，開発セットで最も高い精度が得られた epoch のモデルを以降の処理で用いた．

4.2 評価

Distant Supervision に基く Pre-training を導入した CNN モデルの性能を，3 章のデータで学習および評価した．ラベル付きデータの量と Pre-training の関係を調べるため，ラベル付きデータ量が 1000，2000，5000，10000，20000，38929 の場合でそれぞれ実験を行った．Pre-training は 3.2 節の 2 種類のエモティコンセットで行い，各セットで Pre-training を行ったときおよび Pre-training を行わなかったときの性能を確認した．表 2 に各設定での評判分析の Accuracy を示す．

Distant Supervision に基く Pre-training を導入することにより，学習データ量 20000 で評判極性付与ありの場合を除いたすべての設定で性能向上を確認した．2 種類のエモティコンセットについては，評判極性を付与しなかった設定がすべての場合で評判極性を付与した場合を上回った．ラベル付きデータ量については，特に量が少ないときには Pre-training の効果が大きく (1000 で 10〜12%，2000 で 3〜5%)，量が多いときには 0.5〜1.0% 程度の限定的な性能向上しか確認できなかった．

5 おわりに

Distant Supervision に基く Pre-training を CNN モデルに導入することにより，日本語の評判分析の性能が向上することを確認した．エモティコンの評判極性を付与した場合と付与しなかった場合を比較したが，付与しない場合の方が僅かではあるが高い性能が得られた．また，ラベル付きデータ量と Pre-training の関係を調査し，データ量が少ない場合には Pre-training の効果が大きいことを確認した．

今後の課題としては，CNN の各層で学習している特徴を調査し，Distant Supervision に基く Pre-training で得られる性能向上の要因を確認する予定である．画像においては CNN の各層において段階的にタスク依存な特徴が学習されていると考えられており [5]，言語においても各層で学習できている特徴を明らかにしたい．

参考文献

[1] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, Vol. 12, pp. 2493–2537, 2011.

[2] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP 2014*, pp. 1746–1751, 2014.

[3] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL/IJCNLP 2009*, pp. 1003–1011, 2009.

[4] Aliaksei Severyn and Alessandro Moschitti. UNITN: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of SemEval 2015*, pp. 464–469, 2015.

[5] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*, pp. 3320–3328. Curran Associates, Inc., 2014.

[6] 坪井祐太. 自然言語処理におけるディープラーニングの発展. オペレーションズ・リサーチ, Vol. 4, pp. 205–211, 2015.