DEIM2020 F7-3(day2 d10)

極性反転ニューラルネット

伊藤　友貴† 　坪内　孝太†† 　坂地　泰紀† 　山下　達雄†† 　和泉　潔†

† 東京大学大学院工学系研究科
†† ヤフー株式会社
E-mail: †m2015titoh@socsim.org

あらまし　深層学習モデルが強力なモデルである一方でそのブラックボックス性が故に説明責任を伴う場面では利用できない場合が多い．そこで，本研究では、予測結果を説明しながらドキュメントレベルの感情を分析できる NN, SSNN を提案する．SSNN を用いることでモデルが文書の極性分類に関する予測結果を単語単位でのオリジナルセンチメント，極性反転，文脈センチメントを用いて説明可能であり，説明責任を伴う場面でも利用可能であることが期待できる．さらに，本研究では SSNN 実現のための特殊な学習手法 JSP 学習を提案する．実データを用いて本提案手法の性能を評価したところ，JSP 学習によって「高い予測性能」も「高い説明性能」も兼ね備える SSNN を構築可能であることを検証した．

キーワード　Interpretable Neural Network, Sentiment Analysis, Support System

1　はじめに

1.1　研究背景

商品レビューや口コミなどのテキストデータは商品の改善や現在の経済状況などを把握する上で重要な情報である．これらのテキストデータを解析するアプローチの一つとして，感情分析 (= 各レビューがポジティブかネガティブかを判断する) に関する技術がよく使われる．この「感情分析」においては，深層学習モデルが有効である事が知られている [10]. では，ビジネスにおけるありとあらゆる場面において「深層学習」が実際に使えるのだろうか？その答えは恐らく No である．その理由は「深層学習モデルはブラックボックスである」ことにある．確かに，多くの場合において深層学習モデルを用いると高性能なモデルを構築する事が可能である，しかしながら，深層学習モデルは「その予測結果を説明出来ない」が故に「説明責任が伴う場面」ではなかなか使いにくいのである．そのような「説明責任が伴う場面」では例え性能が大きく深層学習モデルに劣るとしても，線形回帰モデルのような予測結果が説明可能なモデルが使われる事が多い．このようなビジネス上における課題を踏まえると,「予測結果を説明可能」かつ「予測性能も十分に高い」ニューラルネットワークモデルの構築は重要な課題の一つであることは想像に難くない．

1.2　目　　的

そこで，本研究では，図 1 のように最終的なポジネガ分類までの流れを単語レベルでの「オリジナルセンチメント」(= 単語そのものが持つセンチメント値)，その「極性反転」(=「〜ではない」のような否定などにによる極性反転)，そして「文脈センチメント (= 極性反転考慮後のセンチメント)」を用いて可視化できるようなニューラルネットワークモデルの構築を目指す．

図 1 のような説明に関する可視化ができれば，説明責任を伴う場面でもニューラルネットワークモデルを使う事ができ，その意味でこのようなニューラルネットワークモデルはビジネス上においても有用であると考えられる．しかし，既存研究のニューラルネットワークモデルに関する解釈性に関する枠組み [2], [7], [11], [15], [21], [22], [4], [9], [19], [20] では「何が重要か」だけの可視化に留まり，上記のような可視化は難しい．確かに,「何が重要か」は有用であるが,「さらなる納得感」を追い求めて，本研究では図 1 のような説明に関する可視化をできるようなニューラルネットワークの構築を目指す．

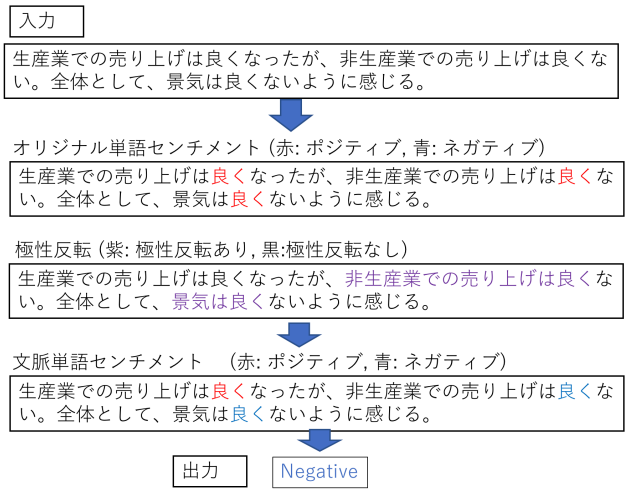

図 1　目標: 予測結果を説明可能なニューラルネットワークモデルの構築

そこで，本研究では上記の課題を達成するために 図 1 のように各単語について「オリジナルセンチメントを意味する層 (WOSL)」,「極性反転を意味する層 (SSL)」，そして「文脈センチメントを意味する層 (WCSL)」を意味する層を持つようなニューラルネットワークモデル, 極性反転ニューラルネット SSNN [8] の構築を目指す．ここで，WOSL は単語センチメン

トに関しての辞書形式での埋め込み層を用いて表現し，SSL は long short-term memories (LSTM) [17] を用いて表現することを考える．さらに， WOSL と SSL の値を掛け合わせることで CWSL を表現し，さらに CWSL における値の和によって最終的な予測結果 (= 文書単位でのポジネガ) を表現する．このような SSNN を実現する上で問題になるのが,「どうやって上記のような各層の解釈性を実現するか」である．当然，通常の Back Propagation 法による学習ではこのような各層の解釈性は実現できない．そこで，本研究では SSNN の各層における解釈性を実現するために Joint sentiment propagation (JSP) learning の提案をする．

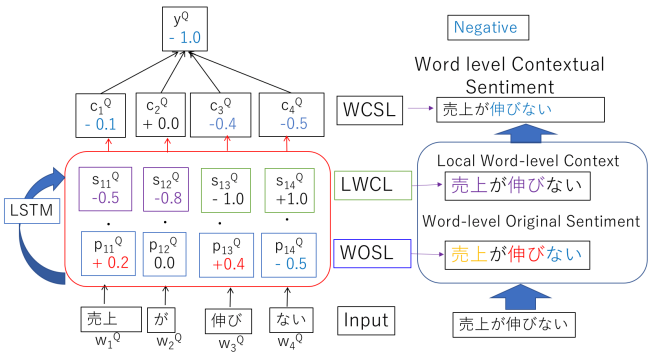

図 2　SSNN

1.3　本研究の貢献

本研究の貢献は以下の通りである．

1) 本研究では極性反転ニューラルネットワークモデル SSNN と呼ばれる，そのポジネガ判定に関する予測結果の説明を各単語のセンチメントの流れという観点から説明可能なニューラルネットワークモデルを提案した．

2) SSNN における各層の解釈性の実現のため，JSP 学習という新しい学習手法を提案した．

3) 実データを用いて本手法の妥当性を検証した結果，本手法が日本語にも英語にも有効な手法であることが検証できた．

2　SSNN: 極性反転ニューラルネット

本節では極性反転ニューラルネットワークモデル SSNN: Sentiment Shift Neural Network について紹介する．SSNN は JSP 学習により構築可能である．ここで， JSP 学習では訓練データ $\{(\mathbf{Q}_i, d^{\mathbf{Q}_i})\}_{i=1}^{N}$ 及び小規模な単語のセンチメントスコア辞書を用いる．ここで， N は訓練データのサイズ,$\mathbf{Q}_i$ はレビュー，$d^{\mathbf{Q}_i}$ はセンチメントタグ (1: ポジティブ, 2: ネガティブ) である．

2.1　SSNN の構成

SSNN は Word-level Original Sentiment layer (WOSL), Sentiment Shift Layer (SSL), そして Word-level Contextual Sentiment layer (WCSL) からなる，レビュー $\mathbf{Q} = \{w_t^{\mathbf{Q}}\}_{t=1}^{n}$ 入力すると，そのポジネガ予測 $y^{\mathbf{Q}} \in \{0(\text{negative}), 1(\text{positive})\}$ を出力する NN である．本論文ではコーパスに出現する語彙数 v の語彙集合を $\{w_i\}_{i=1}^{v}$, 単語 w_i の語彙 ID を $I(w_i)$, $\boldsymbol{w}_i^{em} \in \mathbb{R}^e$ を単語 w_i の次元 e の用意されたコーパスから計算された分散表現 (skip-gram [13] を利用）とし，さらに $\boldsymbol{W}^{em} \in \mathbb{R}^{v\times e} := [\boldsymbol{w}_1^{emT}, \cdots, \boldsymbol{w}_v^{emT}]^T$ とする．

2.1.1　WOSL

この層ではコメント $\mathbf{Q} = \{w_t^{\mathbf{Q}}\}_{t=1}^{n}$ の各単語をその単語が文脈に左右されずに持つセンチメント値，オリジナルセンチメント値に変換する．

$$p_t^{\mathbf{Q}} := w^p_{I(w_t^{\mathbf{Q}})} \tag{1}$$

ここで，$\boldsymbol{W}^p \in \mathbb{R}^v$ は各単語のオリジナルセンチメント値を表す．w_i^p は $\boldsymbol{W}^p$ の i 番目の要素を表し，w_i^p の値が w_i のオリジナルセンチメント値に対応する．

2.1.2　SSL

この層は各単語のセンチメントが反転しているかどうかを表す．まず，レビュー $\mathbf{Q}$ 内の単語 $\{w_t^{\mathbf{Q}}\}_{t=1}^{T}$ を埋め込み表現 $\{\boldsymbol{e}_t^{\mathbf{Q}}\}_{t=1}^{T}$ に変換する．その後 順方向及び逆方向の LSTM によって順方向からの 極性反転 $\overrightarrow{s}_t$ と逆方向からの極性反転 $\overleftarrow{s}_t$ を表す値に変換する．

$$\overrightarrow{\boldsymbol{h}}_t^{\mathbf{Q}} = \overrightarrow{\text{LSTM}}(\boldsymbol{e}_t^{\mathbf{Q}}), \overleftarrow{\boldsymbol{h}}_t^{\mathbf{Q}} = \overleftarrow{\text{LSTM}}(\boldsymbol{e}_t^{\mathbf{Q}}). \tag{2}$$

$$\overleftarrow{s}_t^{\mathbf{Q}} = \tanh(\boldsymbol{v}^{left^T} \cdot \overleftarrow{\boldsymbol{h}}_t^{\mathbf{Q}}), \overrightarrow{s}_t^{\mathbf{Q}} = \tanh(\boldsymbol{v}^{right^T} \cdot \overrightarrow{\boldsymbol{h}}_t^{\mathbf{Q}}). \tag{3}$$

ここで, $\boldsymbol{v}^{right}, \boldsymbol{v}^{left} \in \mathbb{R}^e$ はパラメータであり，$\overrightarrow{s}_t^{\mathbf{Q}}$ 及び $\overleftarrow{s}_t^{\mathbf{Q}}$ はそれぞれ単語 $w_t^{\mathbf{Q}}$ がその右側及び左側の単語群 $w_t^{\mathbf{Q}}$: $\{w_{t'}^{\mathbf{Q}}\}_{t'=1}^{t-1}$ and $\{w_{t'}^{\mathbf{Q}}\}_{t'=t+1}^{n}$ によって反転しているかどうかを表す．

最後に，$\overrightarrow{s}_t$ と $\overleftarrow{s}_t$ から各単語が反転するかどうかを表す値 $\{s_t^{\mathbf{Q}}\}_{t=1}^{T}$ へと変換する．

$$s_t^{\mathbf{Q}} := \overrightarrow{s}_t^{\mathbf{Q}} \cdot \overleftarrow{s}_t^{\mathbf{Q}}. \tag{4}$$

ここで，$s_t^{\mathbf{Q}} < 0$ の場合は単語 $w_t^{\mathbf{Q}} < 0$ が反転している場合を，$s_t^{\mathbf{Q}} > 0$ の場合は反転していない場合をそれぞれ表す．

2.1.3　WCSL

WCSL　では SSL と WOSL の値を用いて各単語の文脈センチメント $\{c_t^{\mathbf{Q}}\}_{t=1}^{T}$ を以下のように表す．

$$c_t^{\mathbf{Q}} = s_t^{\mathbf{Q}} \cdot p_t^{\mathbf{Q}}. \tag{5}$$

2.1.4　出　　力

最後に, SSNN は文書全体でのセンチメント $y^{\mathbf{Q}}$ を以下のように出力する

$$y^{\mathbf{Q}} = \textstyle\sum_{t=1}^{T} c_t^{\mathbf{Q}}.$$

ここで， $y^{\mathbf{Q}} > 0$ は $\mathbf{Q}$ がポジティブであることを表し，また，$y^{\mathbf{Q}} < 0$) は $\mathbf{Q}$ がネガティブであることを表す．

2.2　SSNN の実現における課題及び解決策

本節では，SSNN における各層の解釈性実現のためには何が求められるかについて議論する．また，その上でどのような解決策があり得るのかについて議論する．

SSNN の実現における課題

全ての単語において各層の解釈性を実現させるのは難しいのでここでは以下定義 21 にて定義される単語集合 S^* 内の単語を対象とした各層の解釈性実現を目指すことにする．

［定義 21］ S^* を強いオリジナルセンチメント値を持つ単語の集合とする. このとき，レビュー $\mathbf{Q}$ 内の各単語 $w_t^{\mathbf{Q}} \in \mathbf{Q}$ について， $w_t^{\mathbf{Q}} \in S^*$ であるならば，

$$\begin{cases} d^{\mathbf{Q}} = 1 & (R^*(w_t^{\mathbf{Q}}) \cdot PN^*(w_t^{\mathbf{Q}}) = 1) \\ d^{\mathbf{Q}} = 0 & (R^*(w_t^{\mathbf{Q}}) \cdot PN^*(w_t^{\mathbf{Q}}) = -1) \end{cases} \tag{6}$$

が成り立つものとする．ここで，

$$R^*(w_t^{\mathbf{Q}}) := \begin{cases} -1 & (\text{単語 } w_t^{\mathbf{Q}} \text{の極性が反転している場合}) \\ 1 & (\text{それ以外}) \end{cases},$$

$$PN^*(w_t^{\mathbf{Q}}) := \begin{cases} 1 & (w_t^{\mathbf{Q}} \text{のオリジナルセンチメント値が正}) \\ -1 & (w_t^{\mathbf{Q}} \text{のオリジナルセンチメント値が負}) \end{cases}.$$

とする．

このとき，以下の命題 22 が成り立つ.

［命題 22］ $w_t^{\mathbf{Q}} \in S^*$ かつ 単語 $w_t^{\mathbf{Q}}$ が十分な回数訓練データセット Ω^{tr} に出現するならば, $L_{doc}^{\mathbf{Q}}$ による学習が十分に行われたあと，以下の式 (7) が成り立つ．

$$\begin{cases} c_t^{\mathbf{Q}} > 0 & (d^{\mathbf{Q}} = 1) \\ c_t^{\mathbf{Q}} < 0 & (d^{\mathbf{Q}} = 0) \end{cases} \tag{7}$$

where

$$L_{doc}^{\mathbf{Q}} = CE(\mathrm{sigmoid}(y^{\mathbf{Q}}), d^{\mathbf{Q}}). \tag{8}$$

ここで $CE(a, b)$ は a と b の間の cross-entropy とする.

命題 22 から， 一般的な学習である $L_{doc}^{\mathbf{Q}}$ をロスとする学習によって CWSL の解釈性は 集合 S^* 内の単語についてはある程度保証されることがわかる．しかし， $L_{doc}^{\mathbf{Q}}$ による学習単体では次に述べる問題 23 のために WOSL と SSL における解釈性が保証できないことがわかる．

［問題 23］ $c_t^{\mathbf{Q}}$ の符号が正しく正である場合, 次の二つの場合が状況としてあり得る: (1) $p_t^{\mathbf{Q}} > 0 \wedge s_t^{\mathbf{Q}} < 0$, または (2) $p_t^{\mathbf{Q}} < 0 \wedge s_t^{\mathbf{Q}} > 0$．このとき，$L_{doc}^{\mathbf{Q}}$ を損失とした勾配法による学習では正しい場合を自動的に選択することはできない.

解　決　策

よって，上記の [問題 23] を解決するためには，$p_t^{\mathbf{Q}}$ または $s_t^{\mathbf{Q}}$ の値を適切に制限できる学習が SSNN 実現には求められることがわかる．このことを考慮すると，SSNN の各層における解釈性を実現する上では S^* の部分集合 $\Phi(S^*)$ について以下の初期化を学習前に行うことが SSNN 有効な手段であると考えられる.

$$w_i^p \leftarrow \begin{cases} PS(w_i) & (w_i \in S^d) \\ 0 & (\text{otherwise}) \end{cases} \tag{9}$$

上記の初期化によって，まず $\Phi(S^*)$ 内の単語について適切に $p_t^{\mathbf{Q}}$ に制約がかかることが見込まれる．その結果，$\Phi(S^*)$ 内の単語については [問題 23] が解決され，その極性反転も適切に SSL に反映されるようになることが期待できる．このとき，

［条件 24］ $\|\boldsymbol{e}_t^{\mathbf{Q}} - \boldsymbol{w}_j^{em}\| < \delta$ where δ is sufficiently small, then,

$$\|s_t^{\mathbf{Q}} - s_{it}^{\mathbf{Q}(w_{it}^{\mathbf{Q}}, w_j)}\|_2 < T'\delta.$$

が成り立つため， 集合 $\Phi(S^*)$ に十分に近い単語についての極性反転も SSL に反映されるようになる．その結果，この「集合 $\Phi(S^*)$ に十分に近い単語」についても問題 23 が解決され，これらについてのオリジナルセンチメントも WOSL に反映される．このような SSL を媒介とするような伝搬が繰り返されることにより，集合 「$\Omega(\Phi(S^*)) := \min_{w_i \in \Phi(S^*)} \|\boldsymbol{w}_i^{em} - \boldsymbol{w}_j^{em}\|_2 < \delta$ (δ は十分に小さな値) を満たす単語の集合」に含まれ，かつ S* にも含まれる任意の単語について WOSL 及び SSL に関しての解釈性が実現されることが期待される．よって， $\Phi(S^*)$ が十分に大きく，$S^* \subset \Omega(\Phi(S^*))$ が成り立つ場合には S^* 内の任意の単語について SSNN の各層における解釈性が担保されることが期待される．

しかし，このような $\Phi(S^*)$ や $PN^*(\cdot)$ を完璧に知ることが現実的にはできず，実用化のためには式 (9) のような初期化を現実的に行う方法が求められる．

2.3 Joint Sentiment Propagation 学習

以上を踏まえ，本研究では Joint Sentiment Propagation (JSP) 学習 (Algorithm 1) を提案する．JSP 学習は「単語センチメント辞書を用いた初期化 (Lexicon Initialization)」と「SSL への制約付き学習 」により構成される．

Lexicon Initialization

まず， 式 (9) のような初期化を現実的に行うことを目的に，以下のような初期化を学習前に行う.

$$w_i^p \leftarrow \begin{cases} PS(w_i) & (w_i \in S^d) \\ 0 & (\text{otherwise}) \end{cases} \tag{10}$$

ここで、$PS(w_i)$ は単語 w_i のセンチメント辞書値であり， S^d はセンチメント辞書内単語の集合である．これは S^d が S^* の部分集合であり，また，センチメント辞書のセンチメント値が正しい，つまり $PS(w_i)$ の符号が $PN^*(w_i)$ に一致し，かつ S^d が十分に大きく，$S^* \subset \Omega(S^d)$ という条件が成り立つ場合には S^* 内の任意の単語について SSNN の各層における解釈性が担保されることが期待される.

SSL への制約付き学習

さらに， SSL に適切な制限をかけ，SSNN の解釈性を加速させるために以下の $L_{joint}^{\mathbf{Q}}$ を最小化させるよう学習させる.

$$L_{shift}^{\mathbf{Q}} := \sum_{t \in \{t | w_t^{\mathbf{Q}} \in (S^d \cap \mathbf{Q})\}} SCE(s_t^{\mathbf{Q}}, l_{ssl}(PS(w_t^{\mathbf{Q}}))$$

$$L_{joint}^{\mathbf{Q}} := L_{doc}^{\mathbf{Q}} + \lambda \cdot L_{shift}^{\mathbf{Q}}$$

where $l_{ssl}(a) = \begin{cases} 1 & (a > 0 \wedge d^{\mathbf{Q}} = 1) \vee (a < 0 \wedge d^{\mathbf{Q}} = 0) \\ 0 & (a > 0 \wedge d^{\mathbf{Q}} = 0) \vee (a < 0 \wedge d^{\mathbf{Q}} = 1) \end{cases}$.

ここで， λ はハイパーパラメータであり，$L_{shift}^{\mathbf{Q}}$ は SSL への制約に関するコスト関数である.

この $L^{\mathbf{Q}}_{shift}$ の活用によって $R^*(w^{\mathbf{Q}}_t)$ と $s^{\mathbf{Q}}_t$ の符号が一致させるような制約が $\Omega(S^d)$ 内の単語についてかかることが期待でき，WOSL や SSL への極性情報の伝搬が促進されることが期待できる．

Algorithm 1 Joint Sentiment Propagation Learning

1: **for** $i \leftarrow 1$ to v **do**

2: $$w^p_i \leftarrow \begin{cases} PS(w_i) & (w_i \in S^d) \\ 0 & (\text{otherwise}) \end{cases};$$

3: Learn SSNN using the gradient values by L^{joint}.;

3 解釈性の評価

本節では実データを用いて JSP 学習によって SSNN の解釈性が実現可能かどうかを A) WOSL, B) SSL, and C) WCSL における解釈性の観点から評価する．

3.1 テキストデータ

レビューとそのセンチメントタグのデータとして 景気ウォッチャー調査の現状に関するコメントのデータセット (EcoRev I), 景気ウォッチャー調査の先行きに関するコメントのデータセット (EcoRev II)，2014 年 9 月の Yahoo Finance 掲示板のコメントのデータセット (Yahoo Rev)，Sentiment 140 [1] を用いた (Table 1 参照)．最初の３つは日本語のデータセットであり，Sentiment 140 は英語のデータセットである．

3.2 実験設定

Lexicon Initialization においては日本語のテキストデータを用いる場合には経済に関する専門家作成のセンチメント辞書 [7] を利用し，Sentiment 140 を用いる場合には Vader word sentiment dictionary (Vader dict) [6] を利用した．

今回の実験では，センチメント辞書の値を 50 個使う場合 (SSNN 50)、100 個使う場合 (SSNN 100), そして 200 単語使う場合 (SSNN 200) について検証した．

3.3 評価指標

本実験では A) WOSL, B) SSL, and C) WCSL の解釈性を以下の基準で評価した.

A) WOSL の評価

本評価では経済単語極性リスト，ヤフーファイナンス掲示板単語極性リスト，及び LEX 単語極性リスト (`http://quanteda.io/reference/data_dictionary_LSD2015.html`) を用いた．これらの単語極性リストには単語とそのポジネガ極性情報に関するリストが含まれている（Table 1 参照）．

本検証では WOSL の妥当性を WOSL から得られるリスト内単語の極性と 単語極性リスト嬢の極性の一致度 (macro F_1 値) をもとに評価した．

B) SSL の評価

本評価では以下のような極性反転に関するタグつきのデータセット（1 は極性反転ありを表し，0 は極性反転なしを表す.）をそれぞれのテキストデータに合わせて 3 種類用意し，SSL の妥当性を評価した （Table 1 参照）．

(1) In total, we are in a $bull^{(0)}$ market.

(2) This room is not $clean^{(1)}$.

(3) Products in this shop are too $expensive^{(1)}$.

このタグ付きデータセットを用いて，タグの値と $s^{\mathbf{Q}}_t$ から得られる極性反転に関する判定結果が一致するかどうかを基準に，SSL の解釈性を評価した．評価指標には macro F_1 値を用いた．

C) WCSL の評価

本評価では以下のような文脈センチメントに関するタグつきのデータセッをそれぞれのテキストデータに合わせて 3 種類用意し，WCSL の妥当性を評価した （表 1 参照）．

(1) In total, we are in a $bull^{+}$ market.

(2) This room is not $clean^{-}$.

(3) Products in this shop are too $expensive^{-}$.

このタグ付きデータセットを用いて，タグの値と $c^{\mathbf{Q}}_t$ から得られる文脈センチメントのポジネガ極性に関する判定結果が一致するかどうかを基準に，WCSL の解釈性を評価した．評価指標には macro F_1 値を用いた．

3.4 ベースライン

「Lexical initialization」と「SSL への制約」の効果を見るために SSNN の結果と以下の $SSNN^{Base}$ 及び $SSNN^{Init}$ の結果を比較した．これらの構造は SSNN と同じだが，以下の点に違いがある．

1) $SSNN^{Base}$ では Lexical initialization を用いず，また，SSL への制約もしない (すなわち $L^{\mathbf{Q}}_{joint}$ ではなく $L^{\mathbf{Q}}_{doc}$ を用いて学習).

2) $SSNN^{Init}$ では SSNN 200 と同様の方法で Lexical initialization は用いるが SSL への制約はしない．

3.5 比較手法

SSNN の各層の解釈性の評価のために各層の評価において，SSNN の評価結果とそれぞれ対応する比較手法の結果を比較する．

A) WOSL WOSL の評価においては PMI [14], logistic fixed weight model (LFW) [21], sentiment-oriented NN (SONN) [11], 及び GINN [7] の結果と比較する．

B) SSL SSL の評価においては NegRNN (frequency odds method [12] と [3] にて提案されている 極性反転を判別する RNN の組み合わせ）の結果と比較する．

C) WCSL WCSL の評価においては PMI, LFW, SONN, GINN, Grad + RNN [9], LRP + RNN [1], and IntGrad + RNN [20] の結果と比較した．

3.6 結　　果

表 2 が評価結果である. SSNN がほとんどの場合において他の手法よりも高い性能を出している．これより，本手法によって SSNN の解釈性を十分に実現可能なことが確認できた．

また，$SSNN^{Init}$ と $SSNN^{Base}$ の結果を比較すると，Lexi-

1 : `https://www.kaggle.com/kazanova/sentiment140`

表 1　Dataset details

Text Corpus	EcoRev I	EcoRev II	Yahoo	Tweets
Training				
positive reviews	20,000	35,000	30,612	650,000
negative reviews	20,000	35,000	9,388	650,000
Validation				
positive reviews	2,000	2,000	3,387	50,000
negative reviews	2,000	2,000	1,613	50,000
Test				
positive reviews	4,000	4,000	7,538	100,000
negative reviews	4,000	4,000	2,462	100,000
vocabulary size v	8,071	11,130	33,080	71,316

Annotated data	EcoRev I	EcoRev II	Yahoo	Tweets
word polarity list				
Positive	348	337	422	1,843
Negative	391	387	372	947
sentiment shift tags				
Shifted tags	872	859	378	429
Non-shifted tags	3,762	3,740	2,391	4,504
word-level contextual polarity tags				
Shifted Negative	776	756	227	169
Shifted Positive	96	96	151	260
Non-shifted Negative	1,491	1483	1,187	1,294
Non-shifted Positive	2,271	2179	1,204	3,210

con Initialization　により各層の解釈性が期待通り，向上していることが確認できた．さらに，$SSNN^{Init}$ と SSNN の結果を比較すると，SSL への制約が WOSL 及び SSL の解釈性の向上において期待通り，有用であることがわかる．特に，$SSNN^{Init}$ では EcoRev II と Yahoo において極性反転の判別に大きく失敗しているが，SSNN では性能を保つことができている．

4　予測性能の評価

本節では SSNN の予測性能を実験的に評価する．

4.1　評 価 指 標

SSNN の 予測性能を各テキストデータにおけるテストデータのポジネガ分類がどの程度できるかを基準に評価した．評価指標には Macro F_1 値を用いた．

比 較 手 法 SSNN の 結 果 と logistic regression model (LR), $SSNN^{Base}$, $SSNN^{Init}$, convolutional NN (CNN) model [10], a bidirectional recurrent NN model with LSTM cells (RNN), word attention network (ATT) model [23], hierarchical attention network (HN-ATT) model [23], sentiment, and negation neural network (SNNN) model [5], and lexicon-based supervised attention (LBSA) model [22] の結果を比較し， SSNN の予測性能の評価を行った．

4.2　Result and Discussion

結果は 表 3 の通りである．SSNN は 多くの場合において LR, CNN, ATT, SNNN，及び LBSA よりも高い性能を出すことができた ($p < 0.05$ in five trials)．また，EcoRevs　において

表 2　解釈性に関する評価結果

(A) Evaluation Result for WOSL

	EcoRev I	EcoRev II	Yahoo	Tweets
PMI	0.734	0.745	**0.793**	0.733
LFW	0.715	0.740	0.766	0.725
SONN	0.702	0.724	0.725	0.705
GINN	0.723	0.755	0.754	0.735
$SSNN^{Base}$	0.525	0.472	0.516	0.493
$SSNN^{Init}$	0.720	0.750	0.755	0.731
SSNN (200)	0.778	0.772	0.776	**0.755**
SSNN (100)	**0.788**	0.777	**0.777**	0.751
SSNN (50)	0.779	**0.813**	0.767	0.754

(B) Evaluation Result for SSL

	EcoRev I	EcoRev II	Yahoo	Tweets
NegRNN	0.536	0.626	0.564	0.558
$SSNN^{Base}$	0.350	0.440	0.495	0.365
$SSNN^{Init}$	0.480	0.800	0.500	0.710
SSNN (200)	0.806	0.804	0.662	0.713
SSNN (100)	0.804	**0.813**	0.668	0.713
SSNN (50)	**0.800**	0.798	**0.690**	**0.729**

(C) Evaluation Result for WCSL

	EcoRev I	EcoRev II	Yahoo	Tweets
PMI	0.578	0.548	0.575	0.631
Grad + RNN	0.578	0.621	0.601	0.681
IntGrad + RNN	0.607	0.621	0.625	0.679
LRP + RNN	0.597	0.518	0.579	0.638
LFW	0.549	0.545	0.578	0.587
SONN	0.555	0.542	0.566	0.600
GINN	0.569	0.555	0.577	0.623
$SSNN^{Base}$	0.538	0.582	0.549	0.716
$SSNN^{Init}$	0.546	0.719	0.566	0.780
SSNN (200)	**0.726**	**0.739**	0.649	0.764
SSNN (100)	0.713	0.727	0.640	0.760
SSNN (50)	0.723	0.720	**0.662**	**0.784**

HN-ATT と同程度の予測性能、そして Yahoo review においては HN-ATT よりも高い性能を出すことができた．これらの結果より，十分に高い予測性能を SSNN は誇ることを確認できた．

表 3　予測性能の評価結果

	EcoRev I	EcoRev II	Yahoo	Tweets
LR	0.878	0.879	0.741	0.785
CNN	0.894	0.911	0.757	0.820
RNN	0.922	0.932	0.749	**0.837**
ATT	0.924	0.937	0.750	0.835
HN-ATT	**0.927**	**0.940**	0.750	**0.837**
SNNN	0.918	0.928	0.752	0.827
LBSA	0.922	**0.940**	0.762	0.832
$SSNN^{Init}$	0.884	0.924	0.753	0.828
$SSNN^{Base}$	0.920	0.928	0.737	0.827
SSNN (200)	**0.927**	**0.940**	**0.779**	0.835
SSNN (100)	0.926	**0.939**	**0.776**	0.834
SSNN (50)	0.925	**0.940**	**0.770**	0.834

本章及び前章における評価実験の結果より，JSP 学習によっ

て予測性能も説明性能も併せ持つ SSNN を実現することができることが検証された．

4.3 SSNN による可視化例

最後に．SSNN による可視化結果の結果を紹介する．図 3 が日本語及び英語のレビューに対しての SSNN による可視化例である．このような SSNN による可視化結果をもとに，ユーザーはその予測結果について説明することが可能となる．

例えば，英語の例を見てみると，”great” がポジティブからネガティブに反転しているわけだがこれが ”not” によるものであることが順方向の SSL に関する可視化結果から推察できる．逆に日本語の例を見てみると 「売る」がネガティブからポジティブへ反転しているわけだが，これが「煽り」によるものであることが逆方向の SSL に関する可視化結果から推察できる．

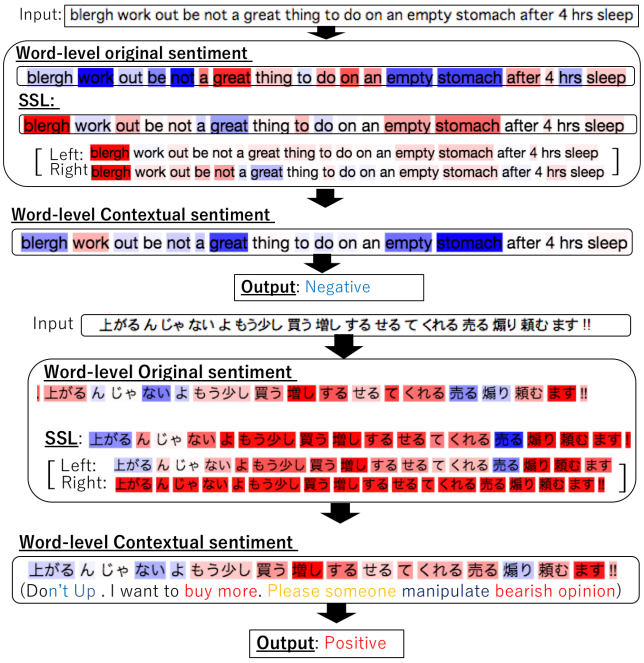

図 3　SSNN's による可視化結果 （上: Yahoo，下: Tweets）. 赤はポジティブ，青はネガティブの値を表す．

5 関 連 研 究

深層学習モデルの ブラックボックス性に関する取り組みとしていくつかの関連研究いが挙げられる．深層学習モデルの予測結果を説明する取り組みとしていくつかの「ニューラルネットワークモデルの解釈」に関する研究が過去に行われてきた [2], [4], [9], [16], [18], [20]. これらの手法を用いると，出力に対する入力の寄与度を Back Propagation 法に近い形で計算することによって，入力要素のうちどこが出力に大きな影響を与えたかを可視化することができる．また，その他の有用なアプローチとして「各層の意味が解釈可能なニューラルネットの構築」 [11], [15], [21], [22] も挙げられる. しかし，これらの既存手法単体では予測結果をオリジナルセンチメント，極性反転，そして文脈センチメントに分解した形で予測結果を説明することができないため，今回の目的を達成することは難しい．

6 お わ り に

本研究では、予測結果を説明しながらドキュメントレベルの感情を分析できる NN, SSNN を提案した．さらに，SSNN を実現するために、JSP 学習を提案した．日本語と英語のデータセットを含むいくつかのデータセットを利用した実験により，JSP 学習によって SSNN が実現可能であること，そして JSP 学習によって構築された SSNN は十分に高い説明能力と予測性能を誇ることを実験的に示しました．今後の課題として，英語・日本語以外の言語への SSNN の適用や JSP 学習の他のタスクへの応用などが挙げられる．また，今回は単語レベルのセンチメントの可視化に焦点を当てて「説明性」について議論したが，この考え方がそもそも妥当なのかについては別途より深く議論する必要がある．なお，本論文は [8] の再公表論文である．

7 謝　　辞

本研究は JSPS 科研費 JP17J04768 の助成を受けたものである．

文　　献

[1] L. Arras, G. Montavon, K. R. Muller, and W. Samek. Explaining recurrent neural network predictions in sentiment analysis. In *EMNLP Workshop*, 2017.

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, and W. Samek. On pixel-wise explanations for nonlinear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, Vol. 10, No. 7, pp. 1–46, 2017.

[3] F. Fancellu, A. Lopez, and B. Webber. Neural networks for negation scope detection. In *ACL 2016*, 2016.

[4] Y. Hechtlinger. Interpretation of prediction models using the input gradient. In *arXiv:1611.07634*, 2016.

[5] Q. Hu, J. Zhou, Q. Chen, and L. He. Snnn: Promoting word sentiment and negation in neural sentiment classification. In *AAAI 2018*, 2018.

[6] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM-14*, 2014.

[7] T. Ito, H. Sakaji, K. Tsubouchi, K. Izumi, and T. Yamashita. Text-visualizing neural network model: Understanding online financial textual data. In *PAKDD 2018*, 2018.

[8] Tomoki Ito, Kota Tsubouchi, Hiroki Sakaji, Tatsuo Yamashita, and Kiyoshi Izumi. Ssnn: Sentiment shift neural network. In *SDM 2020*, 2020.

[9] S. Karen, Ve. Andrea, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

[10] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP 2014*, 2014.

[11] Q. Li. Learning stock market sentiment lexicon and sentiment-oriented word vector from stocktwits. In *CoNLL 2017*, pp. 301–310, 2017.

[12] S. Li, S. Yat, M. Lee, Y. Chen, C. R. Huang, and G. Wang. Sentiment classification and polarity shifting. In *COLING 2010*, pp. 635–643, 2010.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*, 2013.

[14] S. Mohammad, S. Kiritchenko, and X. D. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *SemEval-2013*, 2013.

[15] Z. Quanshi, Y. N. Wu, and S. C. Zhu. Interpretable convolutional neural networks. In *CVPR 2018*, 2018.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. ”why should i trust you?”

explaining the predictions of any classifier. In *KDD*, 2016.
[17] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681, 1997.
[18] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.
[19] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller. In *Striving for simplicity: The all convolutional net*. ICLR Workshop, 2015.
[20] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.
[21] D. T. Vo and Y. Zhang. Don't count, predict! an automatic approach to learning sentiment lexicons for short text. In *ACL 2016*, pp. 219–224, 2016.
[22] Q. Zhang X. Huang Y. Zou, T. Gui. A lexicon-based supervised attention model for neural sentiment analysis. In *COLING 2018*, 2018.
[23] Zichao Yang, Diyer Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL 2016*, 2016.

付　　録

本節では JSP 学習について理論解析を行う．まず，$PN(\cdot)$ 及び 条件 01 を次のように定義する．

$$PN(w_t^{\mathbf{Q}}) := \begin{cases} 1 & (\mathrm{sign}(d^{\mathbf{Q}} - 0.5) \neq R(w_t^{\mathbf{Q}})) \\ -1 & (\mathrm{sign}(d^{\mathbf{Q}} - 0.5) = R(w_t^{\mathbf{Q}})) \end{cases}.$$

［条件 01］

$$w_i^p \begin{cases} > 0 & (OS(w_i^p) > 0) \\ < 0 & (OS(w_i^p) < 0) \end{cases} \quad \text{(A·1)}$$

が成り立つ．ここで，$OS(w_j^p) := E[PN(w_t^{\mathbf{Q}})|w_t^{\mathbf{Q}} = w_j^p, \mathbf{Q} \in \Omega^{tr}]$ であり，Ω^{tr} は訓練データ内のレビューの集合である．

ここで，Here, $PN(w_t^{\mathbf{Q}}) = 1$ は $w_t^{\mathbf{Q}}$ がネガティブなレビュー内で極性反転している場合またはポジティブなレビュー内で極性反転していない場合を指し，$PN(w_t^{\mathbf{Q}}) = -1$ はそれ以外の場合を指す．ここで，S^* 内の任意の単語が $PN^*(w) = \mathrm{sign}(OS(w))$ を満たすことを想定する．このとき，次の命題 02 が成りたつ．

［命題 02］ Eq (7) is satisfied.

［命題 03］ 条件 01　が S^d 内の任意の単語 w_i について成り立つならば， $\Omega(S^d)$ 内の任意の単語 w_t について

$$\begin{cases} E[w^p_{I(w_t)}] > 0 & (OS(w^p_{I(w_t)} > 0) \\ E[w^p_{I(w_t)}] < 0 & (OS(w^p_{I(w_t)}) < 0) \end{cases} \quad and \quad \text{(A·2)}$$

$$\begin{cases} E[s_t^{\mathbf{Q}}] > 0 & (R(w_t^{\mathbf{Q}}) > 0) \\ E[s_t^{\mathbf{Q}}] < 0 & (R(w_t^{\mathbf{Q}}) < 0) \end{cases} \quad \text{(A·3)}$$

が十分な回数の JSP 学習における更新後に $\lambda = 0$ の場合であっても成り立つ．

この命題により，WOSL, SSL, 及び WCSL が理想的な場合には対応するセンチメントを表すようになることがわかる．また，この解析により，センチメント単語辞書の質と量が重要であり，$|S^d|$ は十分に大きく，かつ $S^d \subset S^*$ である必要があることがわかる．

なお，命題 03 は次の命題を用いて説明可能である．

［命題 04］ 条件 01 が単語 $w_t^{\mathbf{Q}}$ について成り立つならば, 式 (A·3) が単語 $w_t^{\mathbf{Q}}$ について成り立つ．

［命題 05］ w_i について条件 01 及び式 (A·3) が成り立つならば, 式 (A·3) が $w_j \in \Theta(w_{it}^{\mathbf{Q}}, \delta)$ (δ は十分に小さい) について成り立つ．ここで, $\Theta(w_t^{\mathbf{Q}}, \delta)$ は $\|\boldsymbol{e}_t^{\mathbf{Q}} - \boldsymbol{w}_j^{em}\|_2 < \delta$ (δ は十分に小さい) を満たす単語の集合である．

［命題 06］ 式 (A·3) が w_i について成り立つならば, 式 (A·2) が w_i について成り立ち，その結果， 条件 01 が成り立つ．