自己回帰型言語モデルを活用した Sentiment Interpretable Neural Network の構築

伊藤友貴 [1]
[1] 三井物産株式会社
Tomok.ito@mitsui.com

概要

深層学習モデルは強力なモデルである一方，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．このような問題を解決するアプローチの一つとして，解釈可能なニューラルネットワークモデルの構築が考えられる．このような背景のもと，本研究では「感情分類」のタスクを対象に，予測結果を説明可能かつその予測性能も高いニューラルネットワークの構築を目指す．本目的達成のため，近年提案された，単語レベルでのセンチメントに基づき予測結果を説明可能ななニューラルネットワーク Sentiment Interpretable Neural Network へ自己回帰型学習言語モデルを導入し，予測性能を解釈性を保ちつつ向上させることを試みる．景気ウォッチャー調査に関する日本語データを用いて本手法の性能を検討した結果，本研究にて提案されたアプローチを取ることで，説明性を担保しつつもベースライン手法に比べ予測性能が高いニューラルネットワークを構築することができた．

1 はじめに

深層学習モデルは強力なモデルである一方，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．このような問題を解決するアプローチの一つとして，解釈可能なニューラルネットワークモデルの構築が考えられる．このような背景のもと，本研究では「感情分類」のタスクを対象に，予測結果を説明可能かつその予測性能も高いニューラルネットワークの構築を目指す．

解釈可能なニューラルネットワークの構築に関する研究は近年，いくつか提案されているが，その中でも最近提案された手法の一つが， Sentiment Interoratable Neural Network (SINN) である．本ニューラルネットワークを活用することで，図 1 のように単語レベルでのセンチメントに基づき予測結果を説明可能となる．SINN を構築するには，専門用語に関する極性辞書をなニューラルネットワーク Sentiment Interoratable Neural Network Model を構築する手法 Lexical Initialization Learning (LEXIL) [2] や Joint Sentiment Propagation 学習 [1]．本手法を利用することにより，センチメント分析の結果を各単語のオリジナルセンチメント，極性反転，及び大域的な重要度に分解する形で説明可能なニューラルネットワークを構築することが可能となる．

本学習手法は対象タスクが感情分類であれば，一定の汎用性が見込まれる一方で，先行研究では比較的層の浅いネットワークにしか適用できておらず，GPT を始めとする，近年提案されている大規模言語モデルに適用できるかどうかは不明瞭である．また，SINN では LSTM 及び Self Attention を用いた文脈層により，極性反転，及び大域的な重要度をモデリングしており，結果，Self Attention の層が双方向の情報を活用しており，極性反転が右側の文脈から起きるのか左側の文脈から起きるのかを LSTM の層にて捉えることができても，その強弱が左右どちらの文脈から引き起こすのかについて観測することは難しい．

このような背景のもと，本研究では，SINN へ自己回帰型の学習済み大規模言語モデルを組み込み，Joint Sentiment Propagation 学習を用いて自然な形で学習させることで，(1) ニューラルネットワークの予測性能向上，及び (2) 各単語のセンチメントへの影響をその前後で分離させ，可視化することを試みる．学習済み大規模言語モデルを利用することで性能の向上が期待され，また，自己回帰型のモデルを採用することで各単語のセンチメントへの影響をその前後で分離させることができることが期待される．提案手法では，まず，学習済み大規模言語モデルを使えるようにするために SINN

[2] を LSTM ベースから GPT-2[3] ベースへ切り替える．さらに，Global Word-Level Context 層には GPT-2 の Decoder 層の Attention をそのまま活用し，Global Word-Level Context 層の計算時には前の情報しか利用できない形にする．このようなモデルの構造を採用することで，各単語のセンチメントへの影響をその前後で分離されることが期待する．景気ウォッチャー調査に関する日本語データを用いて本手法の性能を検討した結果，本研究にて提案されたアプローチを取ることで，説明性をある程度担保しつつも従来手法に比べ予測性能が高いニューラルネットワークモデルを構築できることを検証できた．その一方で，提案手法では極性反転は捉えきれないといった課題も見えた．

2 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「ニューラルネットワークモデルの解釈」に関する研究いくつかが過去に行われてきた [5, 6, 7]．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」も挙げられる．その中でも近年，提案されたのがセンチメント分析の結果をオリジナルセンチメント，極性反転，極性反転，大域的な重要度に分解する形で説明可能なニューラルネットワーク SINN である．また，本手法ではこのような解釈可能なニューラルネットワーク実現のため，極性辞書を用いた初期化 Lexical Initialization を利用した学習手法を提案している．また，[1] では，その改良版アルゴリズム Joint Sentiment（JSP） Propagation 学習が提案されている．

3 自己回帰言語型モデルベース SINN

本節では今回検討する自己回帰型言語モデルベース SINN を紹介する．SINN は訓練データ $\{(\mathbf{Q}_n, d^{\mathbf{Q}_n})\}_{n=1}^{N}$ 及び小規模な単語のセンチメントスコア辞書を用いた学習 Joint Sentiment Propagation (JSP) 学習により構築可能である．ここで， N は訓練データのサイズ，$\mathbf{Q}_n$ はレビュー，$d^{\mathbf{Q}_n}$ はセンチメントタグ (1: ポジティブ, 2: ネガティブ) である．

3.1 モデル

SINN は Token-level Original Sentiment layer (WOSL), Token-level Contextual layer (WCL)，Token level Contextual Sentiment layer (WCSL)，そして出力層から構成される，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．

3.1.1 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．

3.1.2 WCL

この層は各単語 $w_{t'}^{\mathbf{Q}}$ へのセンチメントに関する影響（反転や強弱）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自己回帰型言語モデル CLM （本研究における検証では CLM として GPT2[3] を採用）によって順方向からのセンチメントへの影響 $\overrightarrow{s}_t^{\mathbf{Q}}$ と逆方向からのセンチメントへの影響 $\overleftarrow{s}_t^{\mathbf{Q}}$ を表す値に変換する．

$$\overrightarrow{\boldsymbol{h}}_t^{\mathbf{Q}} := \overrightarrow{\text{CLM}^{DEC}}(w_1^{\mathbf{Q}}, w_2^{\mathbf{Q}}, ..., w_t^{\mathbf{Q}}), \tag{2}$$

$$\overleftarrow{\boldsymbol{h}}_t^{\mathbf{Q}} := \overleftarrow{\text{CLM}^{DEC}}(w_t^{\mathbf{Q}}, w_{t+1}^{\mathbf{Q}}, ..., w_n^{\mathbf{Q}}), \tag{3}$$

$$\overrightarrow{\alpha}_t^{\mathbf{Q}} = \tanh(\boldsymbol{v}^{left^T} \cdot \overleftarrow{\boldsymbol{h}}_t^{\mathbf{Q}}), \overleftarrow{\alpha}_t^{\mathbf{Q}} = \tanh(\boldsymbol{v}^{right^T} \cdot \overrightarrow{\boldsymbol{h}}_t^{\mathbf{Q}}). \tag{4}$$

$$\overrightarrow{\beta}_t^{\mathbf{Q}} = \tanh(\overrightarrow{\text{CLM}^{att}}(w_1^{\mathbf{Q}}, w_2^{\mathbf{Q}}, ..., w_t^{\mathbf{Q}})), \tag{5}$$

$$\overleftarrow{\beta}_t^{\mathbf{Q}} := \tanh(\overleftarrow{\text{att}^{DEC}}(w_t^{\mathbf{Q}}, w_{t+1}^{\mathbf{Q}}, ..., w_n^{\mathbf{Q}})). \tag{6}$$

ここで，$\boldsymbol{v}^{right}, \boldsymbol{v}^{left} \in \mathbb{R}^e$ はパラメータであり，$\overrightarrow{\text{CLM}^{DEC}}$ 及び $\overleftarrow{\text{CLM}^{DEC}}$ はそれぞれ順方向及び逆方向の CLM のデコーダーによって出力される最終層への変換，また，$\overrightarrow{\text{CLM}^{att}}$ 及び $\overleftarrow{\text{CLM}^{att}}$ は順方向及

び逆方向の CLM による各単語へのアテンションへの変換を表す．

$\overrightarrow{s}_t^{\mathbf{Q}}$ 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 $\{w_{t'}^{\mathbf{Q}}\}_{t'=t+1}^{n}$ によるセンチメントに関する影響に関するスコアを表す．次に，$\overrightarrow{s}_t^{\mathbf{Q}}$ と $\overleftarrow{s}_t^{\mathbf{Q}}$ から各単語への両方向からのセンチメントへの影響値 $\{s_t^{\mathbf{Q}}\}_{t=1}^{T}$ へと変換する．

$$s_t^{\mathbf{Q}} := \overrightarrow{s}_t^{\mathbf{Q}} \cdot \overleftarrow{s}_t^{\mathbf{Q}}. \tag{7}$$

先行研究では CLM を LSTM，$\beta_t^{\mathbf{Q}}$ を LSTM から出力される隠れ層をベースとするアテンションを利用しているが，本研究では事前気隔週ズム言語モデル活用による性能向上，及び左右からの影響を見えるようにするため，CLM には GPT2 を採用する．

3.1.3 WCSL

WCSL では WOSL 及び WCL 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．

$$c_{it}^{\mathbf{Q}} := p_{it}^{\mathbf{Q}} \cdot s_{it}^{\mathbf{Q}} \cdot \alpha_{it}^{\mathbf{Q}}. \tag{8}$$

3.1.4 出力

最後に SINN は文の極性を $y^{\mathbf{Q}}$ 以下のように出力する

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

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．

3.2 JSP 学習

SINN は JSP 学習によって学習する．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，以下のような初期化を学習前に行う．

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

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 SINN 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

学習時には次の $L_{joint}^{\mathbf{Q}}$ を最小化させるよう学習させる．

$$\begin{aligned} L_{shift}^{\mathbf{Q}} &:= \sum_{t \in \{t | w_t^{\mathbf{Q}} \in (S^d \cap \mathbf{Q})\}} SCE(s_t^{\mathbf{Q}}, l_{ssl}(PS(w_t^{\mathbf{Q}})) \\ L_{joint}^{\mathbf{Q}} &:= L_{doc}^{\mathbf{Q}} + \lambda \cdot L_{shift}^{\mathbf{Q}} \end{aligned}$$

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 $L_{shift}^{\mathbf{Q}}$ の活用によって $R^*(w_t^{\mathbf{Q}})$ と $s_t^{\mathbf{Q}}$ 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．

4 評価実験

5 解釈性の評価

本節では実データを用いて本手法の評価を予測性能及び解釈性の二点から評価する．

5.1 実験設定

景気ウォッチャー調査の現状に関する日本語コメントのデータセット [2] を用いて実験を実施した．本データセットは訓練データ，検証データ，テストデータから構成され，各データセットにポジティブコメント及びネガティブコメントがそれぞれ 20,000 件，2,000 件，4,000 件格納されている．Lexicon Initialization においては「" 上がる", " 回復", " 上方", " 増加", " 上昇"」及び「[" 減少", " 低下", " 損失", " 遅れ", " リスク"]」についてそれぞれ +1 と -1 を入れる形で初期化を実施した．また，Tokenizer は rinna/japanese-roberta-base を利用した．

5.2 評価指標

予測性能についてはテストデータに対するポジネガ分類制度により評価した．解釈性については，[2] にて提供されている人手で作成された，単語レベルでのポジネガリスト及び極性反転に関するデータセットをもとに (A) WOSL 及び (B) WCSL の評価を実施した．(A) WOSL の評価では，WOSL の妥当

性を WOSL から得られるリスト内単語の極性（ポジティブ 189 件，ネガティブ 198 件）と単語極性リスト内の極性の一致度 (macro F_1 値) をもとに評価した．また，(B) WCSL の評価では，まず極性反転に関するデータセットに記載される各コメントをモデルに読み込ませ， CWL 層から各トークンの文脈スコア $s_t^{\mathbf{Q}}$ を抽出する．その後，単語極性リスト内に含まれるトークンを対象に，$s_t^{\mathbf{Q}} < 0$ の場合，または $\{s_t^{\mathbf{Q}}\}$ の平均値よりも小さい場合にはセンチメントが「反転または軽減されている」とみなし，そうでない場合には「反転または軽減されていない」とみなした．この予測結果とデータセットにつけられる反転タグが一致するかどうかによって評価した．評価指標には Macro F1 値を利用した．尚，本評価データには「反転または軽減されている」のラベルが 660 件，「反転または軽減されていない」のラベルが 3,000 件付与されていた．

5.3 ベースライン

性能評価のため，解釈性に関しては，SINN[2] に加え，SINN における WGCL の層に RoBERTa

5.4 結果・考察

表 1 が評価結果である.SINN+GPT の結果が提案手法の結果である．SINN に比べ，予測性能をあげることには成功した．また，WOSL や WCL に関する解釈性についてもある程度性能を保っていることが見て取れる．また，図 1 で見て取れるように，本提案手法 (SINN+GPT) センチメントへの影響を後ろから受けていることが見て取れる．例えば、「イベントが何もないので、」という記載においてイベント（ポジティブワード）が「何もない」によって後ろから反転させられていることが可視化結果をもとに理解できる。一方，SINN+GPT 内の WCL に関しては，極性反転されている部分について，マイナスの値を取ることができていない様子が見られ，この部分を改善し，SINN (ベースライン) のように極性反転に対し，WCL 内の値をマイナスに取ることができれば，WOSL に関しても更なる性能向上がされることが期待される．

6 結論

SINN へ自己回帰型の学習済み大規模言語モデルを組み込み，Joint Sentiment Propagation 学習を用いて自然な形で学習させることで，(1) ニューラルネットワークの予測性能向上，及び (2) 各単語のセンチメントへの影響をその前後で分離させ，可視化することを試みた．予測性能を解釈性をある程度保ちつつも向上させることには成功したが，まだ改善の余地があると言える．今後の展開としてはマルチリンガルモデルの活用による性能向上や Few Shot 学習手法の提案，あるいはより効率の良い学習手法の提案等が挙げられる．また，より大規模なデータセットを用いた本手法の評価やもう一歩踏み込んだ可視化結果の評価等も今後の展開としては考えられる．

表 1 評価結果

	予測性能	(解釈性) WOSL	(解釈性) WCL
BERT	0.946	–	–
RoBERTA	0.955	–	–
SINN	0.930	0.882	0.779
SINN + Transformer	0.922	0.808	0.684
SINN +ReBerta	0.947	0.898	0.480
SINN+GPT	0.939	0.866	0.654

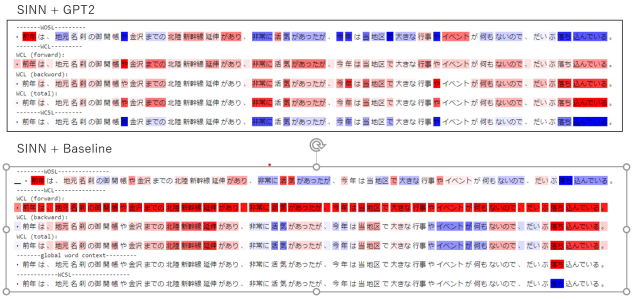

図 1 Visualzation example by SINNs

参考文献

[1] Tomoki Ito, Kota Tsubouchi, Hiroki Sakaji, Tatsuo Yamashita and Kiyoshi Izumi, SSNN: Sentiment Shift Neural Network, SDM 2020, 2020.

[2] Tomoki Ito, Kota Tsubouchi, Hiroki Sakaji, Tatsuo Yamashita and Kiyoshi Izumi, Word-level Contextual Sentiment Analysis with Interpretability, AAAI 2020.

[3] Radford, Alec and Wu, Jeff and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya, Language Models are Unsupervised Multitask Learners, 2019.

[4] Zhuang et al., A Robustly Optimized BERT Pre-training Approach with Post-training, CCL 2021

[5] S. Bach and A. Binder and G. Montavon and F. Klauschen and 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. 2017.

[6] L. Arras and G. Montavon and K. R. Muller and W. Samek,

Explaining Recurrent Neural Network Predictions in Sentiment Analysis, EMNLP Workshop 2017.
[7] M. T. Ribeiro and S. Singh and C. Guestrin, Why Should I Trust You?" Explaining the Predictions of Any Classifier, KDD 2016.