論 文 の 内 容 の 要 旨

論文題目
Development of Interpretable Neural Networks for Document-level Sentiment Analysis
(文書極性分類タスクにおける解釈可能なニューラルネットワークモデルの構築)

氏名 伊藤友貴

1 Introduction

Deep neural networks (DNNs) are known to be promising methods for the document-level sentiment analysis; however, in the real world, they cannot be used in situations where explanations are required owing to their black-box property. Thus, developing a high predictable neural network (NN) model that can explain the process of its prediction process in a human-like way is a critical problem. One of the human agreeable explanation processes is the explanation using the following four types of sentiments as shown in Fig. 1.

Here, word-level original sentiment represents the sentiment that each word in a document originally has. word-level local contextual sentiment represents the sentiment score of each term in a document after considering its sentiment shift, such as “good” in “not good” and “goodness” in “decrease the goodness.” Word-level global contextual sentiment represents the sentiment score of each term after considering what part is important in the entire document (i.e., the global important point) and its sentiment shift, and Document-level sentiment represents the prediction results for positive or negative sentiment tags of reviews.

However, a method for developing NNs that can explain its predictions using these four types of sentiments is yet to be established. Moreover, the basic learning theory for realizing the interpretability in layers is yet to be established. Therefore, this thesis first aims to develop a basic learning theory for realizing the interpretability of each layer. We then aim to develop practical strategies for developing several kinds of interpretable NNs by applying the proposed basic learning strategies in a practical manner.

To achieve this aim, we first propose two types of basic learning called Lexicon Initialization Learning (LEXIL) and Joint Sentiment Propagation (JSP) learning. We then apply these LEXIL and JSP learning to the development of several interpretable network models using real textual datasets. Here, the original LEXIL and JSP learning are not available in real situations. Therefore, we propose practical learning techniques called PLEXIL and PJSP learning which are the revised versions of the LEXIL and JSP learning, respectively.

The main contributions of this thesis are summarized as follows.

(1) We propose a basic learning strategy for developing interpretable NNs called LEXIL and JSP learning.

(2) We design several interpretable NNs in accordance with the requirements. Moreover, we succeeded to realize them by applying the LEXIL and JSP learning in a practical way.

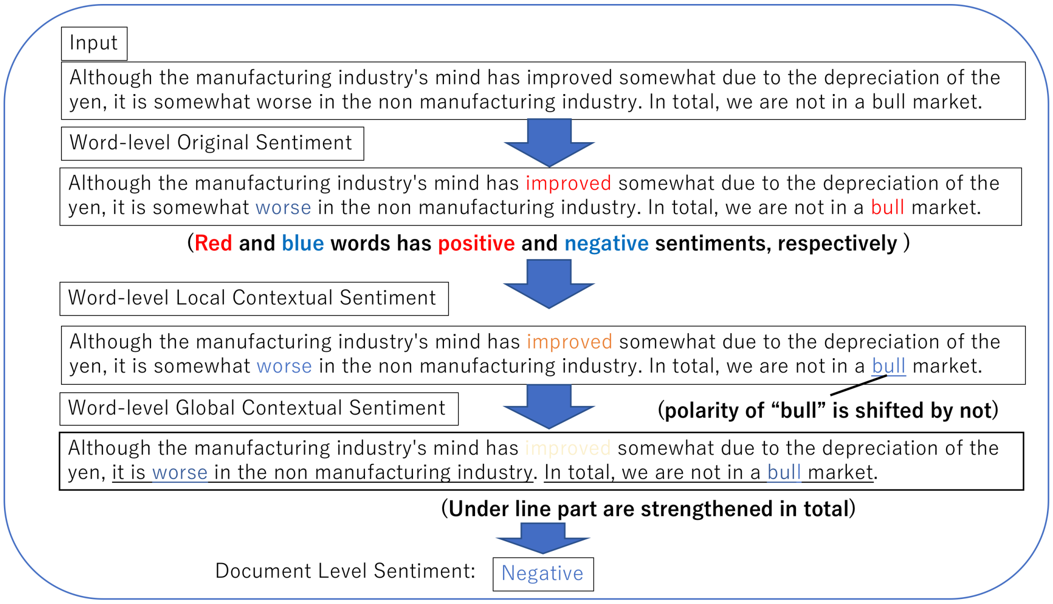

Figure 1: Goal: development of NN that can explain its prediction results using four types of sentiments

(3) As an application of this study, we develop the text-visualization framework called CSCV.

2 Proposed Approach

To consider the learning theory, concretely, we define the layers of a BINN in the following way.

We first define several symbols. Let $\Omega^{tr} = \{(\mathbf{Q}_n, d^{\mathbf{Q}_n})\}_{n=1}^{N}$ be a training dataset where N is the training data size, $\mathbf{Q}_n$ is a review, and $d^{\mathbf{Q}_n}$ is its sentiment tag (1 is positive and 0 is negative). Assume that each review $\mathbf{Q}_n$ has L sentences and each sentence contains T words. $w_{it}^{\mathbf{Q}_n}$ represents the tth word in the ith sentence. Let $\{w_i\}_{i=1}^{v}$ be the terms that appear in a text corpus, v be the vocabulary size, and $I(w_i)$ be the vocabulary index of word w_i where $I(w_i) = i$. Let $\boldsymbol{w}_i^{em} \in \mathbb{R}^e$ be an word embedding word w_i where $\|\boldsymbol{w}_i^{em}\|_2 = 1$, and the embedding matrix $\boldsymbol{W}^{em} \in \mathbb{R}^{v \times e}$ be $[\boldsymbol{w}_1^{emT}, \cdots, \boldsymbol{w}_v^{emT}]^T$ where e is the dimension size of the word embeddings.

2.1 Structure of BINN

We first consider the BINN that includes the following WOSL, SSL, WLCSL, GIL, and WGCSL.

WOSL. Given a review $\mathbf{Q} = \{\{w_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$, this layer converts the words $\{\{w_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}$ to word-level original sentiment representations $\{\{p_{it}^{\mathbf{Q}}\}_{t=1}^{n}\}_{i=1}^{L}$ in a word sentiment dictionary form as

$$p_{it}^{\mathbf{Q}} = w_{I(w_{it}^{\mathbf{Q}})}^{p}. \tag{1}$$

$\boldsymbol{W}^p \in \mathbb{R}^v$ is the original sentiment scores of words, and w_i^p is the ith element of $\boldsymbol{W}^p$. The value of w_i^p corresponds to the original sentiment score of word w_i.

SSL. This layer represents their word-level sentiment shift scores $s_{it}^{\mathbf{Q}}$ using terms and contexts as

$$s_{it}^{\mathbf{Q}} := SSL(\boldsymbol{e}_{it}^{\mathbf{Q}}, \{\boldsymbol{e}_{it}^{\mathbf{Q}}\}_{t=1}^{T}) \tag{2}$$

where $\boldsymbol{e}_{it}^{\mathbf{Q}}$ is the embedding representation of word $w_{it}^{\mathbf{Q}}$, $SSL(\cdot) \in [-1, 1]$ and $s_{it}^{\mathbf{Q}}$ denotes whether the sentiment of $w_{it}^{\mathbf{Q}}$ is shifted ($s_{it}^{\mathbf{Q}} < 0$) or not ($s_{it}^{\mathbf{Q}} \geq 0$).

WLCSL. This layer represents the the word-level local contextual sentiments $c_{it}^{\mathbf{Q}}$ as follows:

$$c_{it}^{\mathbf{Q}} := p_{it}^{\mathbf{Q}} \cdot s_{it}^{\mathbf{Q}}. \tag{3}$$

GIL. This layer represents their word-level sentiment shift scores $s_{it}^{\mathbf{Q}}$ using terms and their contexts:

$$\alpha_{it}^{\mathbf{Q}} := GIL(\boldsymbol{e}_{it}^{\mathbf{Q}}, \{\{\boldsymbol{e}_{it}^{\mathbf{Q}}\}_{t=1}^{T}\}_{i=1}^{L}). \tag{4}$$

where $GIL(\cdot) \in [0, \infty]$ and $\alpha_{it}^{\mathbf{Q}} (> 0)$ represents the scores for global important.

WGCSL. This layer represents the word-level global contextual sentiment scores as follows:

$$g_{it}^{\mathbf{Q}} := c_{it}^{\mathbf{Q}} \cdot \alpha_{it}^{\mathbf{Q}}. \tag{5}$$

Output. Finally, the document-level sentiment score of this review $\mathbf{Q}$ is output as follows:

$$y^{\mathbf{Q}} := \sum_{i=1}^{L} \sum_{t=1}^{T} c_{it}^{\mathbf{Q}} \tag{6}$$

where $y^{\mathbf{Q}} > 0$ means that a review $\mathbf{Q}$ is positive and $y^{\mathbf{Q}} < 0$ means that a review $\mathbf{Q}$ is negative.

2.2 Basic Learning Strategy

This section describes the proposed Lexical Initialization learning (LEXIL) and JSP learning, which are the learning strategy for developing a BINN. We propose them motivated by the following assumption.

Assumption 2.1 *Let S^* be a set of terms which have strong original sentiment. For each $w_{it}^{\mathbf{Q}} \in \mathbf{Q}$, if $w_{it}^{\mathbf{Q}} \in S^*$,*

$$\begin{cases} d^{\mathbf{Q}} = 1 & (R^*(w_t^{\mathbf{Q}}) \cdot PN^*(w_t^{\mathbf{Q}}) \cdot G^*(w_t^{\mathbf{Q}}) = 1) \\ d^{\mathbf{Q}} = 0 & (R^*(w_t^{\mathbf{Q}}) \cdot PN^*(w_t^{\mathbf{Q}}) \cdot G^*(w_t^{\mathbf{Q}}) = -1) \end{cases} \tag{7}$$

is satisfied where

$R^*(w_t^{\mathbf{Q}}) := \begin{cases} -1 & (\textit{sentiment of } w_t^{\mathbf{Q}} \textit{ is shifted}) \\ 1 & (\textit{otherwise}) \end{cases}$, $G^*(w_t^{\mathbf{Q}}) := \begin{cases} 1 & (\textit{term } w_t^{\mathbf{Q}} \textit{ is important in } \mathbf{Q}) \\ 0 & (\textit{otherwise}) \end{cases}$,

and $PN^*(w_t^{\mathbf{Q}}) := \begin{cases} 1 & (\textit{original sentiment of } w_t^{\mathbf{Q}} \textit{is positive}) \\ -1 & (\textit{otherwise}) \end{cases}$.

2.2.1 LEXIL

The proposed learning strategy for BINN called LEXIL is described as follows.

Lexical Initialization LEXIL first initializes the values in $\boldsymbol{W}^p$ using a subset of S^*, $\Phi(S^*)$, as follows:

$$w_i^p \leftarrow \begin{cases} PN^*(w_i) & (w_i \in \Phi(S^*)) \\ 0 & (\text{otherwise}) \end{cases} \tag{8}$$

This Lexical Initialization is effective for improving the iterpretability in GIL, WOSL, and SSL.

Then, LEXIL learns the BINN using the following $L_{doc}^{\mathbf{Q}}$ as a loss:

$$L_{doc}^{\mathbf{Q}} := SCE(\sum_{i=1}^{L} \sum_{t=1}^{T} c_{it}^{\mathbf{Q}}, d^{\mathbf{Q}})$$

where $SCE(a, b)$ means the sigmoid cross-entropy between a and b. Through LEXIL, the layers of BINN learn to represent the corresponding scores in an ideal case where (1) the size of $\Phi(S^*)$ is large enough to satisfy $S^* \in \Omega(\Phi(S^*))$, and (2) Eq (7) is satisfied for all the terms in S^*, and (3) following Condition 2.2 is satisfied:

Condition 2.2 $\|\boldsymbol{e}_{it}^{\mathbf{Q}} - \boldsymbol{w}_j^{em}\| < \delta$ *where δ is sufficiently small, then,*

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

where $\mathbf{Q}(w_{it}^{\mathbf{Q}}, w_j)$ represents the review where word $w_{it}^{\mathbf{Q}}$ is replaced by w_j in $\mathbf{Q}$, $T' > 0$, $T'' > 0$ are established.

2.2.2 Joint Sentiment Propagation (JSP) Learning

In addition, we propose Joint Sentiment Propagation (JSP) Learning as the improved LEXIL. Motivated by Assumption 2.1, after the Lexical Initialization, JSP learns the BINN using the following $L_{joint}^{\mathbf{Q}}$ as a loss:

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

where λ is the hyper-parameter value. $L_{doc}^{\mathbf{Q}}$ corresponds to the loss for document-level sentiment and $L_{shift}^{*\mathbf{Q}}$ corresponds to the loss for regularizing the SSL, and $L_{shift}^{*\mathbf{Q}}$ is expected to accelerate the learning.

2.3 Practical Usage

In practice, $\Phi(S^*)$ and $PN*(\cdot)$ are not available. Therefore, we utilize S^d and $PS(\cdot)$ insead where S^d is a set of sentiment dictionary and $PS(w)$ the sentiment score of word w provised by the dictionary. Experimental evaluations demonstrate that the proposed LEXIL and JSP learning are effective for developing several interpretable NNs, namely, Sentiment Interpretable NN (SINN) [1], Sentiment Shift NN (SSNN), Gradient Interpretable NN (GINN) [2], and Contextual Sentiment NN (CSNN) [3], and that the developed NNs have both the explanation ability and prediction ability.

2.4 Application to Text Visualization

As an application of this study, we develop the text visualization framework called CSCV. Using this CSCV, we can visualize the reviews of shops as shown in Figure 3.

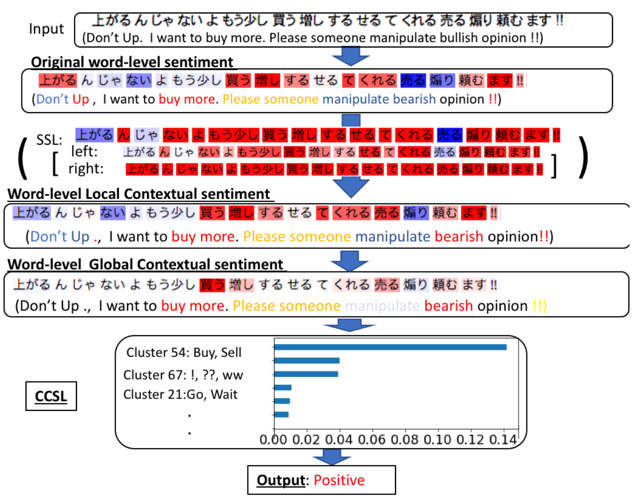

Figure 2: Explanation Example by CSNN

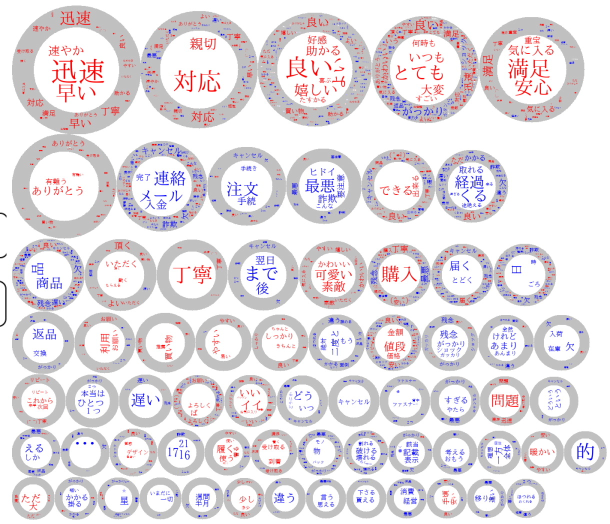

Figure 3: Text Visualization Example by CSCV

3 Conclusion

This thesis proposes two basic learning strategies for developing interpretable NNs called LEXIL and JSP learning. We then applied them to the development of several interpretable NNs in a practical way using real textual datasets. We experimentally demonstrate that the developed NNs with our learning strategy have both the high explanation ability and high predictability. In addition, as an application of this study, we developed the text-visualization framework called CSCV.

References

[1] T. Ito, K. Tsubouchi, H. Sakaji, T. Yamashita, and K. Izumi, "Word-level contextual sentiment analysis with interpretability," in *AAAI 2020*, 2020.

[2] 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.

[3] T. Ito, K. Tsubouchi, H. Sakaji, T. Yamashita, and K. Izumi, "Csnn: Contextual sentiment neural network," in *IEEE ICDM 2019*, 2019.