Enhancing Fine-grained Sentiment Classification Exploiting Local Context Embedding

Heng Yang  
School of Computer,  
SCNU, China  
yangheng@m.scnu.edu.cn

Biqing Zeng  
School of Software,  
SCNU, China  
zengbiqing@scnu.edu.cn

Abstract

Target-oriented sentiment classification is a fine-grained task of natural language processing to analyze the sentiment polarity of the targets. To improve the performance of sentiment classification, many approaches proposed various attention mechanisms to capture the important context words of a target. However, previous approaches ignored the significant relatedness of a target's sentiment and its local context. This paper proposes a local context-aware network (LCA-Net), equipped with the local context embedding and local context prediction loss, to strengthen the model by emphasizing the sentiment information of the local context. The experimental results on three common datasets show that local context-aware network performs superior to existing approaches in extracting local context features. Besides, the local context-aware framework is easy to adapt to many models, with the potential to improve other target-level tasks.

1 Introduction

Target-oriented sentiment classification (TSC) is fine-grained subtask of sentiment analysis, aiming to infer the sentiment polarities (e.g. positive, negative and neutral) of recognized targets (aka Aspect-based sentiment classification (Pontiki et al., 2014), ABSC). Given a customer’s review "The screen resolution attracts me but its battery is miserable", the sentiment polarity of “screen resolution” is positive while the polarity of “battery” is negative.

Existing works usually adapt the recurrent neural networks (RNNs) to solve TSC and regard the target as important supplemental information to infer the polarities. The attention mechanism introduced in machine translation (Bahdanau et al., 2014) has been adapted to TSC by some approaches. For example, Wang et al. (2016), Ma et al. (2017) and Fan et al. (2018) proposed the attention-based models to learn the relatedness of target and context words. However, the improvement of classical attention is limited, while deploying multiple sophisticated attentions increases computation and reduces inferring efficiency. Meanwhile, RNN-based frameworks are insufficient to learn the semantic features of remote context words of a target: while a word is far away from the target in the context, it’s hard to precisely extract the correlation between the target and context. Therefore, many studies explore to eschew RNNs to build models. Self-attention is a novel attention mechanism with a stronger semantic feature extraction ability. And the self-attention (Vaswani et al., 2017) is effectively to extract the semantic relatedness between remote context words based on parallel matrix calculation.

The importance of local context in TSC has been proved in recent works (Yang et al., 2019). The empirical observation is that the context words which are neighbor to a target are more semantic-relevant to the target. In that case, more sentiment information is possibly contained in the target’s local context rather than the remote context. In this paper, we study the significance of local context on TSC, and propose the local context embedding and local context prediction loss, promoting the model to capture local context semantic features from the input text. Combined with the local context focus mechanism, we propose the local context-aware network (LCA-Net) for the TSC task. Apart from implementing the baseline model based on the self-attention, we also implement an enhanced version integrating with BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). Experimental

The codes for this paper are available at https://github.com/yangheng95/LC-ABSA
results on three commonly used TSC datasets show that the local context embedding and local context prediction loss can significantly improve the performance of TSC, which provides a guideline for using local context to improve other target-level natural language processing tasks.

The main contributions of this paper are as follows:

1. The local context embedding (LCE) is proposed in this paper and provides additional information to identify the local context words. LCE promotes the model to learn the local context features during the learning process.

2. A local context prediction (LCP) layer is deployed apart from the sentiment prediction layer, and the local context prediction loss is introduced to enhance the model learning the features of the local context precisely.

3. Accompanied by the local context focus mechanism, the local context-aware network is proposed and obtains state-of-the-art performance. According to the analysis of the ablation experimental results, the local context embedding and local context prediction loss proposed in this paper are highly adaptive to strengthen other TSC models.

2 Related Works

The TSC was regarded as a fine-grained text classification task in previous studies. Traditional machine learning-based approaches (Kiritchenko et al., 2014; Wagner et al., 2014; Vo and Zhang, 2015) generally relied on manual designed features and lexicon features, etc., which are inefficient and easily reach the performance bottleneck. Neural networks have been proved to be competent for extracting text features and semantic relatedness. Consequently, there are flourishing studies of TSC based on deep neural networks.

Tang et al. (2016) proposed TD-LSTM to model the features of the left context and right context of the targets independently and combine these features to analyze polarity. However, the targets are not considered while modeling for the contexts, which may lose the sentiment information of the targets itself. To exploit the potential information of targets, Wang et al. (2016) and Ma et al. (2017) adopted the attention mechanism to help model focus on the important words to the target. Motivated by the obstacles of classical coarse-grained attention, MGAN (Fan et al., 2018) proposed a fine-grained attention mechanism to link and fuse information from the target and the context words. Combining with the coarse-grained and fine-grained attention, MGAN is a multi-attention network and significantly outperforms the coarse-grained attention-based models. For the sentences containing multiple targets, RAM (Chen et al., 2017) and TNet (Li et al., 2018) considers the word position while extract and learn the features of the context, which alleviates the mutual interference between the contextual sentiment information of multiple targets.

There are growing studies that aim to improve the performance of TSC by combining the pre-trained language model (e.g. BERT). BERT-PT and AEN-BERT adapt BERT to solve the TSC and improves its performance. To explore the potential of BERT, Rietzler et al. (2019) argued that, for a specific task, the better performance would be retained if pre-trained BERT is re-fitted on a task-related corpus.

3 Local Context-Aware Network

3.1 Task Definition

Given a sentence $s = \{w_0, w_1, \ldots, w_n\}$ that contains $n$ words including the target, $w_t = \{w'_0, w'_1, \ldots, w'_m\}$ is the target sequence composed of $m$ ($m \geq 1$) words. $s^t$ is a subsequence from $s$, and there could be multiple targets in $s$. Fig 1 is the framework of the local context-aware network.

3.2 Word Embedding

LCA-Net generates the word representations by GloVe (Pennington et al., 2014), which maps each word to a vector space. The embedding lookup matrix is denoted as $E^w \in \mathbb{R}^{d_w \times |V|}$, where $d_w$ is the embedding
Figure 1: The architecture of LCA-Net.

Dimension and $|V|$ is the vocabulary size. And we obtain the global context features by applying multi-head self-attention (MHSA) to learn the raw word representations. The pretrained BERT is an alternative for generating word representations in LCA-Net. The global context features $O^g$ is encoded from the embedded global context $X^g$ by the Global MHSA.

3.3 Multi-Head Self-Attention

We adopt MHSA to encode context features, which perform multiple scaled dot-product attention (i.e. Attention) in parallel. MHSA can alleviate potential features loss of the targets’ remote context words. Suppose $X$ is the feature matrices of input sentences, $K$, $Q$, $V$ are the matrices packed from $X$ by multiplying $W_q \in \mathbb{R}^{d_h \times d_q}$, $W_k \in \mathbb{R}^{d_h \times d_k}$, $W_v \in \mathbb{R}^{d_h \times d_v}$, where $d_h$ is the dimension of hidden size, and $d_q = d_k = d_v = \sqrt{d_h}$. The attention is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax} \left( \frac{Q \cdot K^T}{\sqrt{d_k}} \right) \cdot V$$  \hspace{1cm} (1)

Then we calculated MHSA by assembling scaled-dot attentions:

$$MHSA(X) = \tanh \left( \{H_1; \ldots; H_h\} \cdot W^o \right)$$  \hspace{1cm} (2)

where “;” denotes vector concatenation. $H$ is the output of each attention head, and $h$ is the number of attention heads. $W^o \in \mathbb{R}^{dh \times d_h}$ is the projection matrices. Besides, we deploy a tanh activation.
function for the output of MHSA.

3.4 Local Context-Aware

The framework of the LCA-Net is composed of local context embedding, local context prediction loss, and local context focus mechanism. The local context is identified according to the semantic relative distance (SRD) threshold ($\alpha$), which is proposed to depict the distance between a context word and the target. The SRD ($d_i$) of $i$-th context word relative to a target is calculated as:

$$d_i = |i - p| - \left\lfloor \frac{m}{2} \right\rfloor$$

(3)

where $i$ ($1 \leq i \leq n$) denotes the position of the context word. $p$ is the average position of the target, since a target may contain multiple words. $m$ denotes the length of target.

3.4.1 Local Context Embedding

We propose a novel local context embedding to enhance the model utilizing the local context features based on the local context tag (LC-tag). The LC-tag flags whether a context word belongs to the local context. Suppose a sentence is $T = \{T_0, T_1, \ldots, T_n\}$, we obtain the LC-tag of $i$-th context word:

$$T_i = \begin{cases} 1, & d_i \leq \alpha \\ 0, & d_i > \alpha \end{cases}$$

(4)

The LC-tags are feed into LCA-Net through a local context embedding matrix $E_{lce} \in \mathbb{R}^{d_v \times 2}$. We apply a dot-product operation between embedded LC-tags $X_t$ and global context features $O^g$ learned by the global MHSA.

$$O_{lce}^g = X_t \odot O^g$$

(5)

We also tried the feature-concatenation between $X_{tag}$ and $O^g$. However, the dot-production plays a better role in exploiting local context embedding according to experimental results (e.g. the performance of LCF-BERT on the Laptop dataset drops approximately 1.5% using vector concatenation between $= X_t$ and $O^g$). It is speculated that the local context embedding is position-wisely associated with the vector of global context features, which can adjust the contribution of the corresponding position’s feature.

3.4.2 Local Context Prediction Loss

The motivation for designing local context-tags prediction loss is that if we can embed LC-tags as auxiliary information to promote model make use of local context features, then we can also deploy an LCP layer to predict LC-tags. An empirical hypothesis is that if the model can precisely predict LC-tags for local context, the model’s ability to extract and learn local contextual features will presumably be enhanced, consequently. It can be seen that LCP is a token-level classification task. The LCP loss is defined as follows:

$$\mathcal{L}_{lcp} = - \sum_i^N \sum_k^k \tilde{t}_i \log t_i$$

(6)

where $N = 2$ is the types of LC-tag, $t_i \in \{0, 1\}$ is the predicted LC-tag, $k$ denotes the sum of the context words.

3.4.3 Local Context Focus

The local context focus (LCF) mechanism is proposed to extracting the local context’s features. Compared with classical attention, the LCF reduces the interference of sentiment information contained in multiple target’s local contexts and significantly improves the TSC performance. We adopt the context-feature dynamic masking (i.e. CDM) to obtain local context features, which means only the features generated at the local context words’ position will be preserved. We mask the features of
non-local context words by set their feature-vectors to zero vectors. The mask vector $V_i$ of $i$-th context word is generated as follow:

$$V_i = \begin{cases} E & d_i \leq \alpha \\ O & d_i > \alpha \end{cases}$$

(7)

where $\alpha$ denotes SRD threshold, $M$ is the mask matrix which contains the mask vector for each word. $E \in \mathbb{R}^{d_h}$ is the ones vector and $O \in \mathbb{R}^{d_h}$ is the zeros vector.

$$M = [V_1, V_2, \ldots V_n]$$

(8)

The local context features $O^l$ are obtained as follow:

$$O^l = O_g \odot M$$

(9)

To rebalance the feature-distribution and learn the inner semantic correlation of local context features, a Post-Local MHSA is deployed to learn the local context features $O^l$.

### 3.5 Output Layer

LCA-Net concatenates the $O^\ell_{loc}$ and $O^l$ and employs linear projections to the global context features and local context features. Then, we take the projected features $O^p$ and the first hidden state $O_{head}$ to predict the LC-tags and polarity.

$$\hat{T} = \text{softmax}(W^t \cdot O^p + b^t)$$

(10)

$$\hat{Y} = \text{softmax}(W^y \cdot O^p + b^y)$$

(11)

where $C$ is the number of polarity categories. $W^t \in \mathbb{R}^{N \times d_h}$, $b^t \in \mathbb{R}^N$, $b^l \in \mathbb{R}^C$ and $W^y \in \mathbb{R}^{C \times d_h}$, $b^y \in \mathbb{R}^C$ are the weight vectors and bias vectors. $\hat{T}$ and $\hat{Y}$ are the predicted LC-tags and sentiment polarities, respectively.

### 3.6 Model Training

The loss function of polarity classification is cross-entropy, we employ the LCP loss and sentiment classification loss to optimize model. The joint loss function is as follow:

$$\mathcal{L} = - (1 - \sigma) \sum_1^C \hat{y}_i \log y_i + \sigma \mathcal{L}_\text{lc} + \lambda \sum_{\theta \in \Theta} \theta^2$$

(12)

where $\sigma \geq 0$ adjusts the influence of $\mathcal{L}_\text{lc}$. $\lambda > 0$ is the $L_2$ regularization term, and $\Theta$ denotes the parameter set. The optimizer of in LCA-Net is Adam. The LCA-Net is implemented on different architectures, such as LSTM, MHSA, and BERT, respectively. We also adapt the domain-adapted BERT\(^4\)\cite{Rietzler2019} and BERT-SPC \cite{Song2019}, as the tricks to enhance the LCA-Net.

### 4 Experiments

#### 4.1 Dataset and Hyperparameter Setting

To prove the effectiveness of LCA-Net, we conduct sufficient experiments on three TSC datasets: Laptop, Restaurant and Twitter. The Laptop, Restaurant datasets are obtained from SemEval-2014 task4\cite{Pontiki2014}, and the Twitter dataset is released by Dong et al. (2014). Table 1 and Table 2 depicts the details of three datasets and hyperparameters’ setting, respectively. Polarities of the targets in these dataset are categorized into neutral, positive and negative. We employ the Accuracy and macro F1 as the evaluation metrics to evaluate model performance.

\(^3\)For fair comparison the improvement of LCA-Net, the basic BERT was adopted to build LCA-BERT. We implement our models based on https://github.com/huggingface/transformers And all the experiments are conducted on the RTX 2080 GPU.

\(^4\)There is no domain-adapted BERT for the Twitter dataset, we employ the Restaurant domain-adapted BERT, instead.

\(^5\)The datasets can be found at http://alt.qcri.org/semeval2014/task4
Table 1: The details of three English TSC datasets.

| Datasets     | Positive |          |          | Negative |          |          | Neural  |          |          |           |          |
|--------------|----------|----------|----------|----------|----------|----------|---------|----------|----------|-----------|----------|
|              | Train    | Test     | Train    | Test     | Train    | Test     | Train   | Test     | Train    | Test      | Train    | Test     |
| Laptop       | 994      | 341      | 870      | 128      | 463      | 169      |         |          |          |           |          |
| Restaurant   | 2164     | 728      | 807      | 196      | 631      | 196      |         |          |          |           |          |
| Twitter      | 1561     | 173      | 1560     | 173      | 3126     | 345      |         |          |          |           |          |

Table 2: The hyperparameter’s detail of LCA-Net. The “padding length” means all the sentences will be padded to a unique length. The “5, 3, 5” are the $\alpha$ of Laptop, Restaurant, and Twitter datasets, respectively.

| Hyperparameters | LCA-MHSA | LCA-BERT |
|-----------------|----------|----------|
| learning rate   | $2 \times e^{-3}$ | $2 \times e^{-5}$ |
| batch size      | 32       | 16       |
| hidden size ($d_h$) | 300     | 768      |
| dropout         | 0.1      | 0.1      |
| training epoch  | 10       | 5        |
| padding length  | 80       | 80       |
| $h$             | 30       | 12       |
| $\alpha$        | 5, 3, 5  | 5, 3, 5  |
| $\sigma$        | 0.5      | 0.5      |
| $\lambda$       | $1 \times e^{-4}$ | $1 \times e^{-5}$ |

4.2 Compared Models

To comprehensively evaluate the performance of LCF-Net, we compare LCA-Net with the following models.

**LSTM, LCA-LSTM** We implement a baseline model for TSC based on the bi-directional LSTM (BiLSTM). Moreover, we implement the LCA-Net based on the LSTM, namely LCA-LSTM, to prove the effectiveness of the LCA framework.

**IAN** (Ma et al., 2017) employs interactive attention to learn the semantic relatedness of context and targets.

**RAM** (Chen et al., 2017) deploys multiple attention to learn sentiment features combined with recurrent neural network and weighted-memory mechanism.

**BiLSTM-ATT-G** (Liu and Zhang, 2017) learns the features of left and right context using two attention-based LSTMs, and adjusts the contribution of left and right context features for polarity prediction according to gates.

**MGAN** (Fan et al., 2018) adopts the fine-grained attention and coarse-grained attention mechanisms to learn the features of context and targets, and integrates a target-alignment loss to predict the polarity.

**BERT-PT** (Xu et al., 2019) adapts the pretrained BERT to improve the performance of TSC based on post-training and fine-tuning.

**AEN-BERT** (Song et al., 2019) proposes an attentional encoder network based on the pretrained BERT to model for context and target.

**BERT-SPC** (Devlin et al., 2019) is the pretrained BERT for sentence-pair classification which regards the context and target as sentence-pair.

**BERT-ADA** (Rietzler et al., 2019) proposes the domain-adapted BERT for the Laptop and Restaurant datasets and obtains promising performance.

**LCF-BERT** (Zeng et al., 2019) employs the local context focus mechanism to extract local context features, which models for global context and local context using dual BERTs. We merge the dual BERTs in LCF-BERT to compare with LCA-BERT and analyze the effectiveness of LCE and LCP.
Table 3: The experimental results (%) of LCA-Net. The results of comparative models are retrieved from the previous papers, and “-” denotes the not reported result. “†” means the results are obtained by our implementations. “‡” indicates the model that adopts domain-adapted BERT to improve performance.

| Model          | Laptop       | Restaurant    | Twitter      |
|----------------|--------------|---------------|--------------|
|                | Accuracy     | macro F1      | Accuracy     | macro F1      | Accuracy     | macro F1      |
| LSTM models    |              |               |              |               |              |               |
| LSTM           | 70.22        | 64.36         | 77.50        | 67.17         | 69.49        | 67.64         |
| LCA-LSTM       | 73.04        | 67.79         | 80.89        | 71.81         | 72.25        | 70.05         |
| Baselines      |              |               |              |               |              |               |
| TD-LSTM        | 71.83        | 68.43         | 78.00        | 66.73         | 66.62        | 64.01         |
| BiLSTM-ATT-G   | 73.12        | 69.80         | 79.73        | 69.25         | 70.38        | 68.37         |
| RAM            | 74.49        | 71.35         | 80.23        | 70.80         | 69.36        | 67.3          |
| MGAN           | 75.39        | 72.47         | 81.25        | 71.94         | 72.54        | 70.81         |
| T-Net-LF       | 76.01        | 71.47         | 80.79        | 70.84         | 74.68        | 73.36         |
| LCA-MHSA       | 75.39        | 70.30         | 82.05        | 73.97         | 72.83        | 71.09         |
| BERT models    |              |               |              |               |              |               |
| BERT-BASE†     | 79.00        | 75.59         | 82.59        | 75.36         | 74.13        | 72.19         |
| BERT-PT        | 78.07        | 75.08         | 84.95        | 76.96         | -            | -             |
| AEN-BERT       | 79.93        | 76.31         | 83.12        | 73.76         | 74.71        | 73.13         |
| BERT-SPC†      | 80.09        | 76.39         | 85.62        | 78.94         | 75.58        | 74.35         |
| LCF-BERT†      | 80.72        | 78.05         | 89.11        | 83.86         | 75.72        | 74.34         |
| BERT-ADA†      | 79.19        | 74.18         | 87.14        | 80.05         | -            | -             |
| LCA-BERT†      | 82.45        | 79.22         | 88.93        | 83.96         | 77.46        | 76.17         |

4.3 Overall Performance Analysis

Table 3 shows the main experimental results of LCA-Net. Although LSTM is the basic neural network, the LSTM equipped with LCA techniques is competitive, which indicates that our framework is network-independent and easy to be integrated with other approaches. Compared with the Laptop and Restaurant datasets, all methods perform worse in the Twitter dataset. This is because there are lots of grammatical and spelling errors in the Twitter dataset. LCA-MHSA achieves the superior performance on the Restaurant dataset but underperforms T-Net in the Twitter dataset. The reason is that the convolutional neural network (CNN) is more competent to accurately extract features from ungrammatical sentences (Li et al., 2018). Compared to BERT-SPC and AEN-BERT, the experimental results indicate that the LCA-BERT obtains considerable performance on three datasets, especially the Laptop and Restaurant datasets (almost up to 3-4%). Compared with BERT-PT and BERT-BASE, the performance of LCA-BERT on the Laptop dataset improved nearly 2-3%. The LCF-BERT performs inferior to the LCA-Net on the Laptop and Twitter datasets. Experimental results show that LCA-BERT almost achieves state-of-the-art performance on three datasets. With the same resource occupation and training time, LCA-Net can achieve better results compared with other BERT-based models.

4.4 Ablation Analysis

We have listed the performance of LCA-LSTM, LCA-MHSA, and LCA-BERT in Table 3, proving that LCA-Net which employs LCE, LCP, and CDM improves the neural network-based methods. Next, we discuss the contribution of LCE, LCP, and CDM to the performance improvement through ablation analysis. Table 4 are experimental results of the ablated LCA-Net. It can be seen that LCA-Net based on MHSA almost outperforms all the ablated models, which illustrates that the LCE and LCP proposed for LCA-BERT improves the model’s capability of extracting local context features. According to our analysis, CDM contributes the most in LCA-Net, followed by LCE and LCP. The performance of LCA-BERT on Laptop and Restaurant data machines is not as good as that of some ablated models (such as LCA-BERT w/o LCP and LCA-BERT w/o SPC), but it is worth noting that the training time of LCA-Net is shorter and the convergence is faster compared to ablated LCA-Net. LCA-BERT w/o ADA abandons the domain-adapted BERT to explore the performance of LCA-Net which is based on BERT-BASE, and LCA-BERT w/o SPC removes the BERT-SPC trick. For the LCA-BERT, the LCA-BERT w/o SPC performs better than the baseline of LCA-BERT in the Laptop and Restaurant datasets (82.60% and 89.38% of accuracy, respectively), unexpectedly. Besides, the LCA-BERT outperforms all the ablated
models in the Twitter dataset with a considerable accuracy of 77.46%. In the absence of domain-adapted BERT, LCA-BERT is approximately 1-2% ahead of other BERT-based models on the Laptop dataset.

Table 4: The experimental results (%) of ablated LCA-Nets. “w/o” means “without”. “LCE”, “LCP” and “CDM” denote the local context embedding, local context tags prediction and local context focus, respectively. “SPC” and “ADA” are the BERT-SPC and domain-adapted BERT tricks, respectively.

| Model       | Laptop Accuracy | Laptop macro F1 | Restaurant Accuracy | Restaurant macro F1 | Twitter Accuracy | Twitter macro F1 |
|-------------|-----------------|-----------------|---------------------|---------------------|-----------------|-----------------|
| LCA-MHSA    | 75.39           | 70.30           | 82.05               | 73.97               | 72.83           | 71.09           |
| w/o LCE     | 72.10           | 67.38           | 81.16               | 71.91               | 71.53           | 70.49           |
| w/o LCP     | 74.29           | 70.06           | 81.34               | 73.19               | 72.25           | 70.53           |
| w/o CDM     | 73.51           | 68.49           | 79.73               | 69.57               | 70.81           | 68.79           |
| LCA-BERT    | 82.45           | 79.20           | 88.93               | 83.96               | 77.46           | 76.17           |
| w/o LCE     | 81.03           | 77.88           | 89.11               | 84.52               | 75.87           | 73.90           |
| w/o LCP     | **82.60**       | **79.42**       | 88.39               | 83.07               | 76.73           | 74.10           |
| w/o CDM     | 80.72           | 77.45           | 85.62               | 77.81               | 75.29           | 73.84           |
| w/o SPC     | 82.45           | 79.12           | **89.38**           | **84.66**           | 75.43           | 73.91           |
| w/o ADA     | 81.66           | 78.63           | 86.07               | 79.12               | 76.59           | 75.48           |

4.5 Discussion on Sigma ($\sigma$)

LCA-Net introduces an extra hyperparameter $\sigma$. To explore the influence of $\sigma$ on the performance, we set different $\sigma$ in the LCA-MHSA model to analyze $\sigma$’s impact. Fig 2 is the experimental results on three datasets.

![Figure 2: The performance of LCA-MHSA on under different sigma($\sigma$).](image)

LCA-Net introduces an extra hyperparameter $\sigma$. To explore the influence of $\sigma$ on the performance, we set different $\sigma$ in the LCA-MHSA model to analyze $\sigma$’s impact. Fig 2 indicates that the optimal $\sigma$ in three datasets is different. The LCA-MHSA achieves better performance on the Laptop dataset when $\sigma \approx 0.4$. The optimal effect on the Restaurant dataset obtained when $\sigma \approx 0.2$. For the Twitter dataset, the preferable performance obtained when $\sigma = 0$ or $\sigma \approx 0.6$. Due to the limited computation resource, we do not conduct this experiment in LCA-BERT, and the $\sigma$ used by LCA-BERT in three datasets is 0.5.

4.6 Case Study

Table 5 shows the predictions of some cases in LCA-Nets. “P”, “N”, “O” means positive, negative, and neutral, respectively. It can be seen that the local context-aware framework has a strong capability of target sentiment feature extraction. LCA-Net achieves high accuracy in predicting local context.
Table 5: Several cases of the TSC task. The words in **bold** and *italic* are the targets and its local context \((\alpha = 3)\), respectively. The “✓” and “×” indicate the correct and error prediction of the LC-tags.

| No. | Sentences                                                                 | Polarity |
|-----|---------------------------------------------------------------------------|----------|
| 1   | The food was extremely tasty, creatively presented and the wine excellent.  | P ✓ ✓ ✓ ✓ |
| 2   | It feels cheap, the keyboard is not very sensitive.                       | N ✓ ✓ ✓ ✓ |
| 3   | The food is surprisingly good, and the decor is nice.                    | P ✓ ✓ ✓ ✓ ✓ |
| 4   | Windows 7 can get 8 out of 10 viruses. Merry Christmas.                   | P × (O) ✓ |
| 5   | I hope the songs of britney spears will be good just like right now the next 20 years. | P ✓ ✓ ✓ ✓ ✓ |

tags. Consistent with other approaches, LCA-Net performs slightly poor in predicting neutral sentiment compared to positive and neutral sentiment. More than half of the targets in the Twitter dataset contain neutral sentiment, which potentially results in low performance. According to the experimental results, the more LC-tags are correctly identified, the more likely the polarity of the aspect is to be accurately inferred. Generally.

5 Conclusion and Future Works

To exploit the significance of potential target-related information of local context, the local context embedding and local context prediction are proposed in this paper. Combining with the local context focus mechanism, we propose a novel LCA-Net framework, which obtains state-of-the-art performance on three datasets. Besides, We conducted extensive ablation experiments to demonstrate the importance of LCE and LCP. To validate the transferability of the LCA-Net framework, we implement the LCA-Net based on LSTM, MHSA, and BERT, respectively, which indicates that it is a network-independent framework and can be easily adapted to other approaches. In the future, we will study the promotion of local context-aware techniques on other target-level NLP tasks, such as word sense disambiguation and part-of-speech tagging.

Acknowledgments and Funding

Thanks to the anonymous reviewers and the scholars who helped us. This research is supported by the Innovation Project of Graduate School of South China Normal University and funded by National Natural Science Foundation of China, Multi-modal Brain-Computer Interface and Its Application in Patients with Consciousness Disorder, Project approval number: 61876067.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In Proceedings of the 2017 conference on empirical methods in natural language processing, pages 452–461.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers), pages 49–54.
Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3433–3442.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), pages 437–442.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 946–956.

Jiangming Liu and Yue Zhang. 2017. Attention modeling for targeted sentiment. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 572–577.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, pages 4068–4074.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. arXiv preprint arXiv:1908.11860.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Targeted sentiment classification with attentional encoder network. In International Conference on Artificial Neural Networks, pages 93–103. Springer.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In Twenty-Fourth International Joint Conference on Artificial Intelligence, pages 1347–1353. AAAI Press.

Joachim Wagner, Piyush Arora, Santiago Cortés Vaillo, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Toumi. 2014. Deu: Aspect-based polarity classification for semeval task 4. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 223–229.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2324–2335, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Heng Yang, Biqing Zeng, JianHao Yang, Youwei Song, and Ruyang Xu. 2019. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. Applied Sciences, 9(16):3389.