Neurocomputing 419 (2021) 344–356

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction

Heng Yang [a,1], Biqing Zeng [b,*,1,2], Jianhao Yang [b,1], Youwei Song [c,1], Ruyang Xu [a,1]

[a] School of Computer, South China Normal University, Guangzhou 510631, China
[b] School of Software, South China Normal University, Foshan 528225, China
[c] Baidu Inc., Beijing 100085, China

ARTICLE INFO

Article history:
Received 12 December 2019
Revised 20 July 2020
Accepted 8 August 2020
Available online 5 September 2020
Communicated by Erik Cambria

Keywords:
Aspect term extraction
Aspect polarity classification
Chinese sentiment analysis
Multi-task learning
Multilingual ABSA
Domain-adapted BERT

ABSTRACT

Aspect-based sentiment analysis (ABSA) task is a fine-grained task of natural language processing and consists of two subtasks: aspect term extraction (ATE) and aspect polarity classification (APC). Most of the related works merely focus on the subtask of Chinese aspect term polarity inferring and fail to emphasize the research of Chinese-oriented ABSA multi-task learning. Based on the local context focus (LCF) mechanism, this paper firstly proposes a multi-task learning model for Chinese-oriented aspect-based sentiment analysis, namely LCF-ATEPC. Compared with other models, this model equips the capability of extracting aspect term and inferring aspect term polarity synchronously. The experimental results on four Chinese review datasets outperform state-of-the-art performance on the ATE and APC subtask. And by integrating the domain-adapted BERT model, LCF-ATEPC achieves the state-of-the-art performance of ATE and APC in the most commonly used SemEval-2014 task4 Restaurant and Laptop datasets. Moreover, this model is effective to analyze both Chinese and English reviews collaboratively and the experimental results on a multilingual mixed dataset prove its effectiveness.

1. Introduction

Aspect-based sentiment analysis [1–3] (ABSA) is a fine-grained task compared with traditional sentiment analysis, which aims to automatically extract the aspect terms and predict the polarities of them. Cambria et al. [4] discussed the importance of sentiment analysis and introduced its application scenarios in many fields. For example, given a restaurant review: "The dessert at this restaurant is delicious but the service is poor." the full-designed model for ABSA needs to extract the aspects "dessert", "service" and correctly inferring about their polarity. In this review, the consumers' opinions on "dessert" and "service" are not consistent, with positive and negative sentiment polarity respectively.

Generally, aspect terms and their polarity need to be manually labeled before running the APC task. However, most of the proposed models for aspect-based sentiment analysis tasks only focus on improving the classification accuracy of aspect polarity and ignore the research of the Chinese ATE subtask. Therefore, when conducting transfer learning on aspect-based sentiment analysis, those proposed models often fall into the dilemma of lacking aspect extraction method on targeted tasks because there is not enough research support.

The aspect term extraction (ATE) and aspect polarity classification (APC) are the subtasks of ABSA. The APC is a kind of text classification task. There are a large number of deep learning-based models have been proposed to solve the APC subtasks, such as the models [5–10] based on long short-term memory (LSTM) and the methodologies [11,12] based on the Transformer network [13]. The purpose of the APC is to predict the exact sentiment polarity of different aspect terms, rather than fuzzily analyz the overall sentiment polarity on the sentence-level or document-level. In the APC task, the polarities are most usually classified into three categories: positive, negative, and neutral. The sentiment polarity classified based on aspects can better mine the fine-grained emotional tendency in reviews or tweets, thus providing a more accurate reference for decision-makers.

Consistent with the named entity recognition [14] (NER) task, the ATE is recognized as a subtask sequence tagging, and aims to extract aspect terms from the reviews or tweets. In most approaches [15–17], the ATE is studied independently, away from

* Corresponding author at: School of Software, South China Normal University, Foshan 528225, China.
E-mail addresses: yangheng@m.scnu.edu.cn (H. Yang), zengbiqing@scnu.edu.cn (B. Zeng), yangjianhao@m.scnu.edu.cn (J. Yang), songyouwei@baidu.com (Y. Song), cs_xuruyang@m.scnu.edu.cn (R. Xu).
[1] ORCID: 0000-0002-6831-196X
[2] ORCID: 0000-0001-9088-4759

https://doi.org/10.1016/j.neucom.2020.08.001
0925-2312/

the APC task. The ATE models first segment a review into separate tokens and then infer whether the tokens belong to any aspect. The tokens may be labeled in different forms in different studies, but most of the works adopts the IOB2 [3] labels to annotate tokens. Table 1 are several cases of joint task of ATE and APC.

To extract aspect terms from the text efficiently and analyze the sentiment polarity of aspects collaboratively, this paper proposes a multi-task learning model for aspect-based sentiment analysis. Multilingual processing is an important research orientation of natural language processing. The LCF-ATEPC[4] model is a novel multilingual and multi-task model. Apart from achieving state-of-the-art performance in commonly used SemEval-2014 task4 datasets, the experimental results in four Chinese review datasets also validate that this model has a strong ability to be satisfy the demands of multilingual ABSA. The proposed model is based on multi-head self-attention (MHSA), integrating the bidirectional encoder representations from transformers (BERT) [18] and the local context focus mechanism. Training on a small amount of annotated data of aspect terms and their polarities, the model can be adapted to a large-scale dataset, automatically extracting the aspect terms and predicting the sentiment polarities. In this way, the model can discover the unknown aspects and avoids the tedious and huge cost of manually annotating all aspects and polarities. It is of great significance for the field-specific ABSA.

The main contributions of this article are as follows:

1. For the first time, we study the multi-task learning for APC and ATE using Chinese and multilingual reviews, and proposes a novel model to solve the APC and ATE synchronously.
2. We adapt the self-attention and local context focus techniques to improve collaborative training of ATE and APC, and experimental results on Chinese and multilingual datasets demonstrate our model significantly outperforms state-of-the-art performance compare to existing approaches.
3. The proposed model integrates domain-adapted BERT to improve both the performance of ATE and APC. The experiments indicate the great potential of domain-adapted pretrained model and bring considerable effect especially the F1 score of ATE task.

2. Related works

Many existing approaches regarded the ATE and APC as independent tasks and studied separately. Accordingly, this section will introduce the related works of ATE, APC, and multi-task learning works in this section.

2.1. Aspect term extraction

The works for ATE are classified into two categories: the early dictionary-based or rule-based approaches, and methodologies based on machine-learning or deep learning. Poria et al. [19] proposed a new rule-based approach to extracting aspects from product reviews using common sense and sentence dependency trees to detect explicit and implicit aspects. Liu et al. [20] adopted an unsupervised and domain-independent aspect term extraction method that relies on syntactic dependency rules and can selects rules automatically.

Compared with manually annotating all aspects in the dataset, the models for the ATE subtask can learn the features of aspects and automatically extract aspects in the text, which greatly saves labor and time. Mukherjee et al. [21] proposed a model that can extract and cluster aspects simultaneously according to the seed words provided by users for several aspect categories. By classification, synonymous aspects can be grouped into the same category. Poria et al. [22] introduced the first aspect-oriented deep learning model in opinion mining, which deploys a multi-layer deep convolutional neural network (CNN) to mark each word in the sentences with opinions as an aspect or non-aspect word. He et al. [23] proposed a new method for ATE, which utilized word embedding to explore the co-occurrence distribution of words and apply the attention mechanism to weaken the irrelevant words and further improves the coherence of all aspects. Wang et al. [24] proposed a deep neural network-based model which does not require any parser or other linguistic resources to be pre-processed and provides an end-to-end solution. Besides, the proposed model is a multi-layer attention network, where each layer deploys a pair of attentions. This model allows the aspect terms and opinion terms learned interactively and dual propagate during the training process.

For the Chinese-oriented ATE, a multi-aspect bootstrapping method [25] was proposed to extract the aspects of Chinese restaurant reviews. Zhao et al. [26] introduced machine learning methods to explore and extract aspect terms from Chinese hotel reviews. They chose the optimal feature-dimension, feature representation, and maximum entropy classifier according to the empirical results, and studied the integral effect of aspect extraction. Up to now, the LCF and pretrained language model has not been applied in the Chinese ATE task.

2.2. Aspect polarity classification

The APC is another important subtask of ABSA. The models proposed for APC can be categorized into traditional machine learning and recent deep learning-based methods. In recent years, the models for APC have been comprehensively turned to the deep neural networks. Therefore, this section mainly introduces approaches based on deep learning techniques.

The most commonly applied deep neural network (DNN) architectures for APC are recurrent neural networks [7–9,27,28] (RNNs) and convolutional neural networks (CNNs) [16,17,29]. TD-LSTM [7] is an early RNN-based method that divides the context of aspects into the left and right parts and models for them independently. Attention mechanism [30] is a new technique that successfully improved the APC in the last few years. ATAE-LSTM [8] employs an attention mechanism on the features of aspect terms and context words to dynamically calculate the attention weights of the context words and finally predicts the polarity of aspects according to the weighted context features. IAN [9] is an attention-based LSTM network, with interactively integrating and learning the inner correlation of the features of context and targeted aspects. RAM [15] is based on the bi-directional LSTM (BiLSTM) network, and deploys a multi-layer deep neural network with dedicated memory layers. The multi-layer network utilizes the token features learned based on the attention mechanism and GRUs to finally obtain the global semantic features of the sentence to predict the sentiment polarities of targeted aspects. To retard the loss of context features during the training process, Li et al. introduced a conventional transformation architecture TNet [27] based on context-preserving transformation units. TNet integrates the BiLSTM and CNN, significantly improves the accuracy of sentiment polarity prediction. Multi-grained attention network [10] (MGAN) is a new DNN architecture which equips a variety of fine-grained attention mechanisms and applies these attention mechanisms to interactively learn the token-level features between aspects and context, making great use of the inherent semantic correlation of aspects and context. Peng et al. [31] proposed the methods for the Chinese APC task which conducted the APC at the

[3] The IOB2 labels adopted in this paper are: B_{asp}, I_{asp}, O

[4] The codes for this paper are available at https://github.com/yangheng95/LCF-ATEPC.

Table 1
Several samples from seven ABSA datasets. All the datasets are domain-specific.

No.	Sentence	Aspect	Polarity
1	Great laptop that offers many great **features**!	**features**	positive
2	The **seats** are uncomfortable if you are sitting against the wall on the wooden benches.	**seats**	negative
3	How do you settlers of catan for the **xbox**?	**xbox**	neural
4	**车内顶灯**我觉得灯光效果一般。	车内顶灯	negative
5	高**分辨率**决定了其拥有比以往采用低分辨率的机型更精细逼真的画面效果.	分辨率	positive
6	其致命弱点是**寿命**短而且具有记忆效应。	寿命	negative
7	出色的**制造工艺**加上强金属质感的拉丝工艺面板带来更好的体验。	制造工艺	positive

aspect-level via three granularities. Two fusion methods for the granularities in the Chinese APC are introduced and applied. Empirical results show that the proposed methods achieved promising performance on the most commonly used ABSA datasets and four Chinese review datasets.

Ma et al. [32] presented the Sentic LSTM for ABSA based on the attentive LSTM. In this work, the hierarchical attention of targets and sentences are proposed to capture the important features. By incorporating with the commonsense knowledge of sentiment-related concepts, this approach obtains hopeful performance. The gated alternate neural network (GANN) [33] proposed for APC aimed to solve the shortcomings of traditional RNNs and CNNs. The GANN applied the gate truncation RNN (GTR) to learn the aspect-dependent sentiment clue representations. Tan et al. [34] proposed an aspect embeddings training method according to the correlation between aspect categories and aspect terms, trained a model that can properly represent the relation between aspect-categories and aspect-terms. The studies of Chinese ABSA usually innovate by improving the feature representation, such as constructing auxiliary lexicons [35] and employing radical embedding [36] and word embedding [31] to enhance the extraction ability of Chinese text features. However, the pre-training [18] is another direction of innovations which is focusing on designing universal language models to improve feature-learning ability.

BERT-SPC is the BERT sentence pair classification model, and it was adapted to solve the ABSA task in [11] and achieve considerable performance. LCF-BERT [12] proposed a feature-level local context focus mechanism based on self-attention, which can be applied to ABSA and many other fine-grained natural language processing tasks. BERT-ADA [37] shows that the pretrained model is based on a large universal corpus and is easy to be adapted to most tasks and improve performance. However, the pretrained model is not task-specific. For specific tasks, if the pretrained BERT is adapted to specific tasks through the fine-tuning process on a task-related corpus, the performance can be further improved.

2.3. Multi-task learning works

The ABSA-oriented multi-task learning models generally rely on joint modeling and training. Nguyen et al. [38] proposed a joint model based on BiLSTM and condition random fields (CRFs) to solve ATE and APC subtasks simultaneously. The experimental results on SemEval-2014 task4 datasets show the model achieves competitive performance to several baseline models. Li et al. [39] presented a joint framework of aspect term extraction and aspect polarity classification using two stacked recurrent neural networks. This framework can model the constrained transitions from target boundaries to target sentiment polarities. Chen et al. [40] proposed a joint network aims at APC and aspect-opinion pair identification subtasks. The external knowledge is considered and put into the network to alleviate the problem of insufficient train data. Based on BiLSTMs and CNNs, Akhtar et al. [41] proposed a multi-task learning framework to jointly model for the ATE and APC. This framework employs self-attention and BiLSTM extracting the aspect terms and infer their polarities based on CNN. Hu et al. [42] proposed a span-based extract-then-classify framework which contains three methods for joint ATE and APC task. This framework extracts multiple aspect targets simultaneously under the supervision of target span boundaries and infers the corresponding polarities of aspects exploiting their span representations. IMN [43] is the interactive multi-task learning network that aims to model for multiple related tasks both in the token and document level. Compared with other multi-task learning models, IMN does not depend on learning common features of the relative tasks and employs a delicate message-passing mechanism to deliver information to different tasks using latent variables.

Some scholars also proposed the end-to-end methods to deal with aspect extraction and aspect polarity classification. These methods solve the ATE and APC independently are difficult to learn the correlation between ATE and APC compared with the multi-task learning approaches. Zeng et al. [44] proposed an end-to-end neural network model for the ABSA based on joint learning, and the experimental results on a Chinese review dataset show that the proposed model works fine while conducting ATE and APC simultaneously. Phan et al. [45] introduced another end-to-end ATE and APC solution which explores the grammatical aspect of the sentence and employs the self-attention mechanism for syntactical learning. This work employs the refined LCF to infer aspect polarities. Zhao et al. [46] proposed a multi-task learning framework based on shared spans (SpanMlt) to solve the pair-wise aspect and opinion terms extraction. Compared with existing works, SpanMlt treats the problem from a perspective of joint term and relation extraction and alleviates the sequence tagging patterns. Moreover, the multi-task studies concerning aspect-based sentiment analysis are gradually expanding. e.g., Chauhan et al. [47] proposed a multi-task learning framework for ABSA and sarcasm detection based on sophisticated attention mechanisms. The framework achieves hopeful performance on their self-annotated dataset.

In this paper, we propose a multi-task learning framework for the extraction and classification of aspect terms in a unified model. Compared with other ABSA multi-task learning approaches, our model employs self-attention and LCF mechanism which is more competitive. We conduct sufficient experiments on up to eight datasets (English, Chinese and Multilingual), and experimental results indicate our model outperforms considerable performance on both ATE and APC subtasks. Meanwhile, we adopt the

pre-trained language model (BERT) to solve the poor performance of traditional word embedding on small datasets, and for the first time, the model works more stable and competitive for some complex situations, such as multilingual learning of ATE and APC.

3. Methodology

The methodology of LCF-ATEPC is based on self-attention and the local context focus mechanism. Moreover, the domain-adapted BERT model integrated into the LCF-ATEPC provides an enhancement for model performance. This section introduces the architecture and methodology of LCF-ATEPC. The modules for APC and ATE are introduced independently. The contents are organized by the hierarchy of the network layer.

3.1. Task definition

3.1.1. Aspect term extraction

The ATE is usually regarded as a kind of sequence tagging task, which prepares the input based on IOB labels. We design the IOB labels as B_{asp}, I_{asp}, O, which indicate the beginning, inside and outside of the aspect terms, respectively. e.g., the input of the review "The price is reasonable although the service is poor." will be prepared as $S = \{w_1, w_2 \dots w_n\}$, and w stands for a token after tokenization, $n = 10$ is the total number of tokens. And the sentence will be labeled as $Y = \{O, B_{asp}, O, O, O, O, B_{asp}, O, O, O\}$.

3.1.2. Aspect polarity classification

The APC is a fine-grained subtask of sentiment analysis, aiming to predict the aspect polarity for targeted aspects. e.g., "The price is reasonable although the service is poor." will be prepared as $S = \{w_1, w_2 \dots w_n\}$, and $S^t = \{w_i, w_{i+1} \dots w_j\}$ $(1 <= i < j <= n)$ is the subset of S, represents the words of an aspect. i and j are the beginning and end positions in S, respectively.

3.2. Model architecture

Fig. 1 presents the framework of LCF-ATEPC. The left of the framework is the local context feature generator (LCFG) unit and the right is the global context feature generator (GCFG) unit. The LCF-ATEPC employs the LCFG and GCFG modules based on two independent pretrained BERTs to model the local context and global context, respectively.

Both context feature generator units contain an independent pretrained BERT layer, $BERT^l$, and $BERT^g$ respectively. Feature interactive learning (FIL) layer combines the local context features and global context features and predicts the sentiment polarity of aspects. The aspect extractor identifies aspect terms based on global context features. To conduct the multi-task learning collaboratively, the input sequence is tokenized into tokens, and each token is annotated with ATE and APC labels. The ATE label indicates whether the token belongs to an aspect term and the APC label reveals the polarity of an aspect term.

3.2.1. BERT-shared layer

BERT is a fine-tuning based approach which pre-trains the "masked language model" (MLM) using the bidirectional transformer encoder and alleviates the unidirectionality constraint of context. The MLM randomly masks some tokens from the tokenized input context, intending to predict the original masked word merely based on its context. Moreover, the MLM fuse the left and the right context to build a deep bidirectional transformer encoder pretrained model (i.e. BERT). BERT improved performance for most NLP tasks.

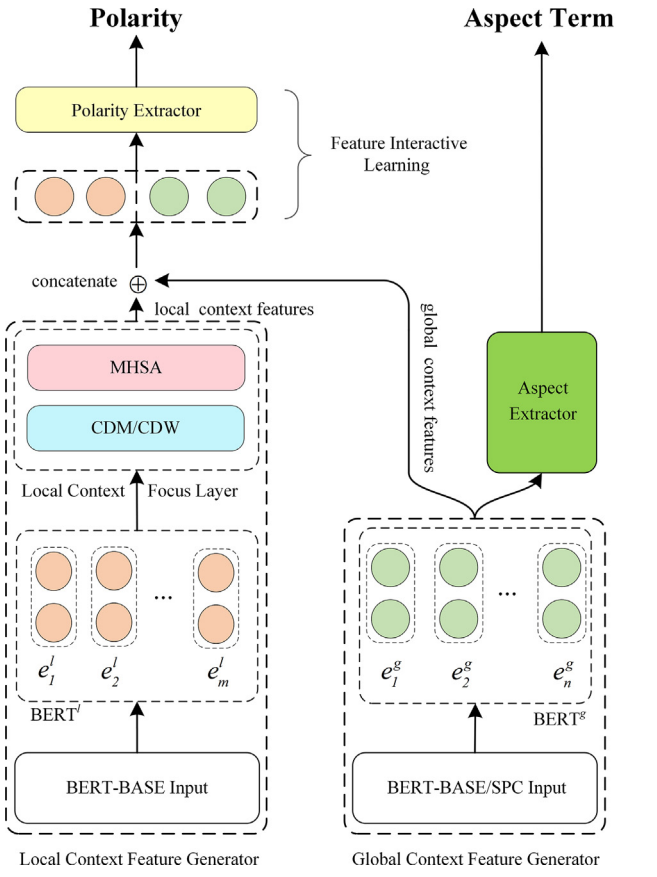


Fig. 1. The multi-task learning framework of our model.

To adapt pretrained BERT to LCF-ATEPC, the fine-tuning process is indispensable. Both BERT-shared layers are regarded as embedded layers, and the fine-tuning process is conducted independently according to the joint loss function during multi-task learning. X^l and X^g are the tokenized inputs of LCFG and GCFG respectively, and we can obtain the preliminary outputs of local and global context features.

$$O^l_{BERT} = BERT^l\left(X^l\right) \tag{1}$$

$$O^g_{BERT} = BERT^g(X^g) \tag{2}$$

O^l_{BERT} and O^g_{BERT} are the preliminary extracted features of the input, respectively. $BERT^l$ and $BERT^g$ are the corresponding BERT-shared layer embedded of LCFG and GCFG.

3.3. Multi-head self-attention

Based on multiple scaled-dot attention (SDA), MHSA can be utilized to extract deep semantic features of the context, and the features are represented in the self-attention score. MHSA can alleviate the negative influence caused by the long-distance dependence of the context. Suppose X_{SDA} is the input features of MHSA, the scaled-dot attention is calculated as follows:

$$\mathrm{SDA}(X_{SDA}) = Softmax\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

$$Q, K, V = f_x(X_{SDA}) \tag{4}$$

$$f_x(X_{SDA}) = \begin{cases} Q = X_{SDA} \cdot W^q \\ K = X_{SDA} \cdot W^k \\ V = X_{SDA} \cdot W^v \end{cases} \quad (5)$$

Q, K and V are the abstract matrices and obtained by multiplying the output features of the upper SDA layer by their weight matrix W^q, W^k, W^v, respectively. For the first SDA layer, they are obtained from the embedded features of the input text. $W^q \in \mathbb{R}^{d_h \times d_q}, W^k \in \mathbb{R}^{d_h \times d_k}, W^v \in \mathbb{R}^{d_h \times d_v}$ are the randomly initialized weight matrices, they will be updated during the training process. $d_h = 768$ is the hidden size, $d_q = d_k = d_v = d_h \div h = 64$. $h = 12$ is the number of attention heads.

$$MHSA(X) = \tanh\left(\{H_1; \ldots; H_h\} \cdot W^{MH}\right) \quad (6)$$

$H_i (1 <= i <= h)$ is the learned feature from each SDA head. MHSA performs multiple scaled-dot attention in parallel and concatenates the output features, then transforms the features by multiplying a vector W^{MH}. The ";" denotes feature vector concatenation of each head. $W^{MH} \in \mathbb{R}^{hd_v \times d_h}$ is the parameter matrices for projection. We apply a tanh activation function for the MHSA encoder, which significantly enhanced feature-capture capability.

3.4. Local context focus

3.4.1. Semantic-relative distance

The identification of local context depends on semantic-relative distance (SRD), which is proposed to determine whether the context word belongs to the local context of a targeted aspect. Local context is a novel concept that can be adapted to many fine-grained NLP tasks. The early approaches generally segment input into aspects and context, and model for them separately. Instead of merely modeling for aspect, LCF-ATEPC takes the local context and global context to learn deep features of the aspect because the empirical result shows the local context of the targeted aspect contains important information. SRD describes how far a token is from a targeted aspect. SRD is calculated as:

$$SRD_i = |i - p_a| - \lfloor \frac{m}{2} \rfloor \quad (7)$$

where i $(1 < i < n)$ is the position of each token, p_a is the central position of each aspect term. m is the length of aspect term, and SRD_i represents for the SRD between i-th token and aspect term.

Figs. 2 and 3 are two implementations of the local context focus mechanism. The bottom and top of the figures represent the feature input and output positions (POS) corresponding to each token. The self-attention treats all tokens equally, each token can generate the self-attention score with other tokens through parallel matrix computation. According to the definition of MHSA, the features of output POSs corresponding to each token are more related to the token itself. After calculating the output of all tokens by MHSA, the output features of each POS will be masked or attenuated, and the features of the local context will be retained intact.

3.4.2. Context-features dynamic mask

According to Fig. 2, the CDM layer masks the non-local context's features learned by the $BERT^l$ layer. Although masking the non-local context words is easier than operating their features, that will completely discard the contribution of non-local context words in the learning process. As the CDM layer deployed, the non-local context words can participate in the MHSA encoding, only the features output on their related POS will be masked.

LCF-ATEPC masks the features of non-local context words by set their feature vectors to zero vectors. Suppose that the O_{BERT^l} is the preliminary output features of $BERT^l$, then we get the local context features output as follows:

$$V_i = \begin{cases} E & SRD_i \leqslant \alpha \\ O & SRD_i > \alpha \end{cases} \quad (8)$$

$$M = \left[V_1^m, V_2^m, \ldots V_n^m\right] \quad (9)$$

$$O_{CDM}^l = O_{BERT^l} \cdot M \quad (10)$$

To mask the features of the non-local context, we define a feature masking matrix M, and V_i^m is the mask vector for each token in the input. α is SRD threshold and n is the length of input sequence including aspect. Tokens whose SRD regarding the targeted aspect is less than the threshold α are local contexts. The $E \in \mathbb{R}_h^d$ represents the ones vector and $O \in \mathbb{R}_h^d$ is the zeros vector. "." denotes the dot-product operation of the vectors.

$$O^l = MHSA\left(O_{CDM}^l\right) \quad (11)$$

Finally, the local context features learned by the CDM layer are delivered as O^l.

3.4.3. Context-features dynamic weighting

We design the CDW layer to explore the potential of the LCF mechanism. The CDW is another implementation of LCF, takes a more modest strategy compared to the CDM which simply discards the features of non-local context completely. While keeping the local context features retained intact, the non-local context features will be weighted decay according to their SRD.

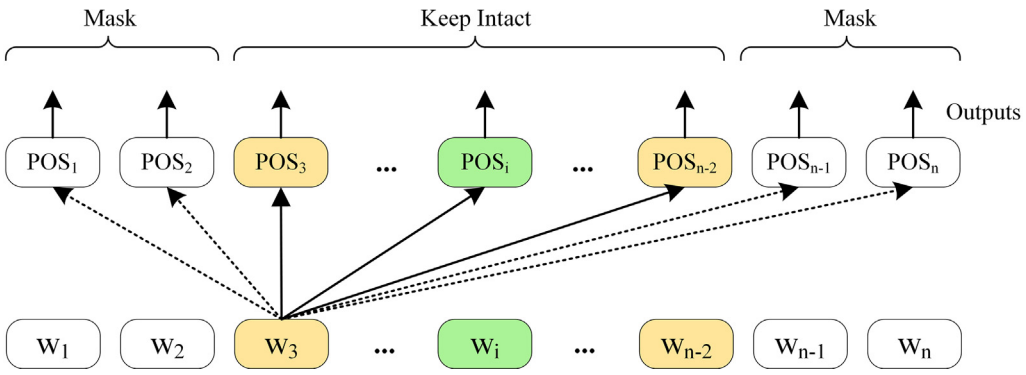


Fig. 2. The simulation of context-feature dynamic mask (CDM) mechanism. The arrows mean the interaction of token in attention computation. And the output features on each dotted-line arrowing position will be masked.

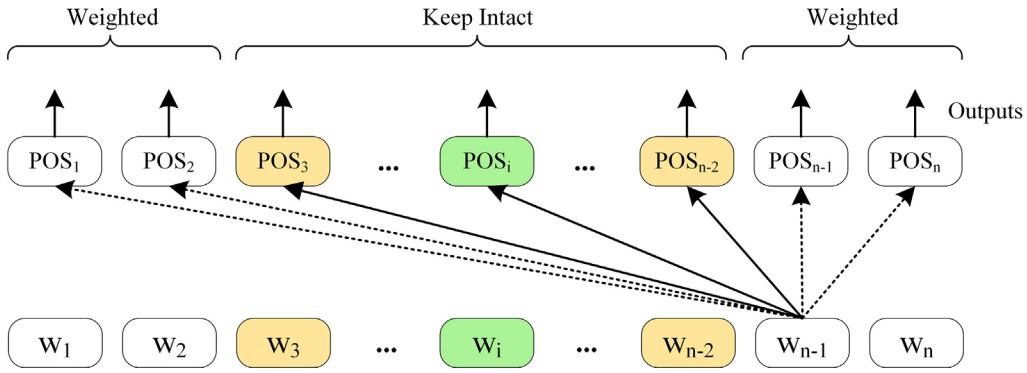


Fig. 3. The simulation of context-feature dynamic weighting (CDW) mechanism. The arrows mean the interaction of token in attention computation. And the output features on each dotted-line arrowing position will be weighted.

$$V_i = \begin{cases} E & SRD_i \leqslant \alpha \\ \frac{n-(SRD_i-\alpha)}{n} \cdot E & SRD_i > \alpha \end{cases} \tag{12}$$

$$W = \left[V_1^w, V_2^w, \ldots V_n^w\right] \tag{13}$$

$$O^l_{CDW} = O_{BERT^l} \cdot W \tag{14}$$

W is the weight matrix, and V_i^w is the weight vector for each non-local context word. Consistently with CDM, SRD_i is the SRD between i-th context token and the targeted aspect. n is the length of the input sequence and "·" denotes the vector dot-product operation.

$$O^l = MHSA\left(O^l_{CDW}\right) \tag{15}$$

O^l_{CDW} is the output of the CDW layer. The CDM and CDW layers are independent, alternative and fusible. Both the output features of CDM and CDW layers are denoted as O^l.

Besides, we tried to concatenate the learned features of CDM and CDW layers and take linear transform as the O^l, which is the LCF-fusion.

$$O^l_{fusion} = \left[O^l_{CDM}; O^l_{CDW}\right] \tag{16}$$

$$O^l_{fusion} = W^f \cdot O^f + b^f \tag{17}$$

$$O^l = MHSA\left(O^l_{fusion}\right) \tag{18}$$

W^f, O^f are weight matrix and b^f is bias vector. LCF-CDM, LCF-CDW, and LCF-Fusion are three modes alternatively to learn local context features.

3.5. Feature interactive learning

The discarded context features may cause some potential information loss, in addition to local context characteristics, LCF-ATEPC takes the global context features as a supplement, combines both to predict aspect polarity by feature interactive learning process (FIL).

$$O^{lg} = \left[O^l; O^g\right] \tag{19}$$

$$O^{lg}_{dense} = W^{lg} \cdot O^{lg} + b^{lg} \tag{20}$$

$$O^{lg}_{FIL} = MHSA\left(O^{lg}_{dense}\right) \tag{21}$$

O^l and O^g are the local context features and global context features, respectively. O^{lg} and O^{lg}_{dense} are the concatenated features and transformed features. $W^{lg} \in \mathbb{R}^{d_h \times 2d_h}$ and $b^{lg} \in \mathbb{R}^{d}_{h}$ are the weight vector and bias vector, respectively. To learn the concatenated features, an MHSA encoding process is performed on the O^{lg}_{dense}. O^{lg}_{FIL} is the output features of FIL.

3.6. Aspect polarity classifier

Aspect polarity classifier performs a head-pooling on the concatenated features. Head-pooling extracts the features of the first output position. Then a Softmax operation is applied to predict the sentiment polarity.

$$X^{lg}_{pool} = POOL\left(O^{lg}_{FIL}\right) \tag{22}$$

$$Y_{polarity} = \frac{\exp\left(X^{lg}_{pool}\right)}{\sum_{k=1}^{C} \exp\left(X^{lg}_{pool}\right)} \tag{23}$$

where C represents the number of sentiment categories, and $Y_{polarity}$ is the predicted polarity of aspect term.

3.7. Aspect term extractor

Aspect term extractor first performs the token-level classification for each token. Suppose T_i is the features on the corresponding position of token T,

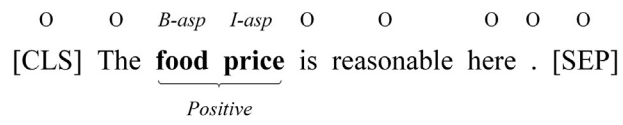
$$Y_{term} = \frac{\exp(T_i)}{\sum_{k=1}^{N} \exp(T_i)} \tag{24}$$

where N is the number of token categories, and Y_{term} represents the token category inferred by aspect polarity classifier.

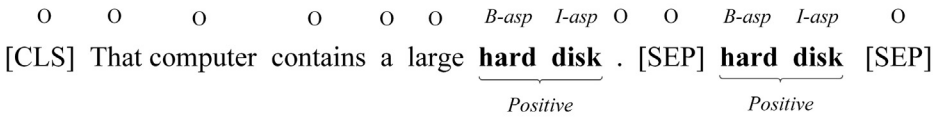
3.8. Training details

LCF-ATEPC employs BERT-BASE[5] in LCFG and GCFG to extract local context features and global context features, respectively. However, BERT-SPC [11] can significantly improve the APC performance on English datasets, such as the Restaurant and Laptop datasets of SemEval-2014 task4. Our APC experimental results of English datasets are obtained by using the BERT-SPC to extracting the global context features. Compared to BERT-BASE, BERT-SPC only reforms the

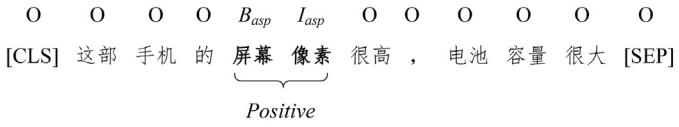
[5] BERT-BASE and BERT-SPC are the variants of BERT.



(a) The English input format of BERT-BASE.



(b) The English input format of BERT-SPC.



(c) The Chinese input format of BERT-SPC.

Fig. 4. The input formats for LCF-ATEPC. The first line depicts the ATE labels, and the second and third lines are the input sentence and APC labels, respectively.

input. For example, the input of BERT-BASE is formed as *"[CLS]" + input (aspect included) + "[SEP]"*, while it is formed as *"[CLS]" + input (aspect included) + "[SEP]" + aspect + "[SEP]"* in BERT-SPC.

Fig. 4 presents the input samples of the BERT-BASE and BERT-SPC models.

The two BERT-shared layers extract the initial features of the input. The LCF layer learns the local context features from the output features of $BERT^l$, and then learns the extracted local context features through MHSA to reblance the distribution of features. The local context features are supplemented to help the model focus on the sentimental information contained in words near the aspect. The ATE module mainly identifies aspect terms based on global context features learned from $BERT^g$. Because the local context features is not applicable since the aim of ATE is to extract aspect terms. The local context information according to the annotated aspects can be given in the training process, however, when extracting aspect terms, the model can not determine the aspects in the input text, so the local context information is not used when extracting aspects.

LCF-ATEPC adopts the cross-entropy loss for APC and ATE, and the regularization is $\mathbf{L}_2$, here is the loss function for APC task,

$$\mathcal{L}_{apc} = -\sum_{1}^{C} \widehat{y}_i \log y_i + \lambda_1 \sum_{\theta_1 \in \Theta_1} \theta_1^2 \tag{25}$$

where C is the number of polarity categories, λ is the L_2 regularization parameter, Θ_1 and Θ_2 are the parameter set of the LCF-ATEPC model. The loss function for ATE task is as follow:

$$\mathcal{L}_{ate} = -\sum_{1}^{N} \sum_{1}^{k} \widehat{t}_i \log t_i + \lambda_2 \sum_{\theta_2 \in \Theta_2} \theta_2^2 \tag{26}$$

where N is the number of token classes and k is the sum of the tokens in each input sequence. Accordingly, the loss function of LCF-ATEPC is as follows:

$$\mathcal{L}_{atepc} = \mathcal{L}_{apc} + \mathcal{L}_{ate} \tag{27}$$

4. Experiments

4.1. Datasets and hyperparameters setting

We prepare seven ABSA datasets to comprehensively evaluate the performance of LCF-ATEPC, involving four Chinese review datasets [48,49,31] (Car, Phone, Notebook, Camera) and three most commonly used English ABSA datasets (the Laptops and Restaurant datasets of SemEval-2014 Task4 subtask2 [1], and an ACL Twitter social dataset [50]). The polarity of each aspect term on the Laptops, Restaurants, and Twitter datasets may be "positive", "neutral", or "negative", the conflicting labels of polarity are not considered. The reviews in Chinese datasets have been purged, with each aspect may be "positive" or "negative". To verify the effectiveness and multilingual ATE and APC performance of LCF-ATEPC, we build a multilingual dataset by mixing the seven datasets to conduct multilingual-oriented ATE and APC experiments.

Table 2 demonstrates the details of these datasets.[6] Not all those datasets have a balanced sample distribution. For example, most samples in the Restaurant dataset are positive, while the neutral samples in the Twitter dataset account for the majority.

Generally, the hyperparameters settings referred to previous researches. We also conducted the controlled trials and analyzed the experimental results to optimize the hyperparameters. The fine-tuned hyperparameters are listed in Table 3. The default SRD threshold for all experiments is 5, with additional instructions for experiments with different SRD. All experiment obtains the results by 5 runs, and our experiments are based on Ubuntu 18.04 and GTX 1080Ti.

4.2. Compared methods

We compare the LCF-ATEPC to state-of-the-art approaches. Experimental results show that LCF-ATEPC achieves new state-of-the-art performance both in the ATE and APC tasks.

ATAE-LSTM [8] is a classical LSTM-based network for the APC task, which applies the attention mechanism to focus on the important words in the context. Besides, ATAE-LSTM appends aspect embedding and the learned features to make full use of the aspect features. The ATAE-LSTM can be adapted to the Chinese review datasets.

ATSM-S [31] is a baseline model of the ATSM variant for Chinese language-oriented ABSA task. This model learns the sentence and aspect terms at three perspectives of granularity.

GANN is a novel neural network model for APC aimed to solve the shortcomings of traditional RNNs and CNNs. The GANN applys gate truncation RNN to learn informative aspect-dependent senti-

[6] The dataset processed for joint ATE and APC task are available at https://github.com/yangheng95/LCF-ATEPC

Table 2
The multi-task datasets of ABSA, including three English, four Chinese review datasets and a multilingual dataset.

Datasets	Language	Positive		Negative		Neural	
		Train	Test	Train	Test	Train	Test
Laptop	English	994	339	870	128	463	169
Restaurant		2164	727	807	196	631	196
Twitter		1561	173	1560	173	3126	345
Car	Chinese	708	164	213	66	–	–
Phone		1319	341	667	156	–	–
Notebook		328	88	168	35	–	–
Camera		1197	322	541	112	–	–
Mixed	Multilingual	8271	2154	4826	866	4220	710

Table 3
Global hyperparameters settings of the LCF-ATEPC on Chinese and English datasets.

Hyperparameters	Chinese/English Dataset
learning rate	3×10^{-5}
batch size	32/16
hidden size	768
training epochs	5
max sequence length	40/80
SRD threshold (α)	5
rounds of experiment	5

ment clue representations. GANN obtained state-of-the-art APC performance on the Chinese review datasets.

AEN-BERT [11] is an attentional encoder network based on the pretrained BERT model, which aims to solve the English aspect polarity classification.

BERT-PT [51] is a BERT-adapted model for review reading comprehension (RRC) task, a task inspired by machine reading comprehension (MRC). It could be adapted to the aspect-based sentiment classification.

BERT-BASE [18] is the basic variant of BERT. We adapt it to ABSA multi-task learning, which equips the ability to automatically extract aspect terms and classify aspect polarity.

BERT-SPC [11] is another variant of BERT designed for the sentence-pair classification (SPC) task and improves the APC subtask of LCF-ATEPC model.

BERT-ADA [37] is a domain-adapted BERT-based model proposed for the APC task, which fine-tuned the BERT-BASE model on the task-related corpus. This model obtained state-of-the-art accuracy on the Laptops dataset.

LCF-ATEPC[7] is the multi-task learning model for the ATE and APC, which is based on the BERT-shared layers and local context focus mechanism.

LCF-ATE is the variant of the LCF-ATEPC model which disables the APC module and only optimizes the model for the ATE task.

LCF-APC is another variant of LCF-ATEPC which disables the ATE module and it only optimizes for the APC task during the training process.

4.3. Results analysis

This section introduces experiments results of LCF-ATEPC on seven datasets with multiple languages. Firstly, the baseline performance of LCF-ATEPC on all Chinese and English datasets is tested. Then, the effectiveness of the multi-task learning strategy is demonstrated. Finally, the assistance of domain-adapted BERT in improving performance is evaluated and the sensitivity of different datasets to SRD is studied.

[7] We implement our model based on pytorch-transformers:https://github.com/huggingface/pytorch-transformers.

4.3.1. Performance on Chinese review datasets

Table 4 are the experimental results of LCF-ATEPC models on four Chinese review datasets. Compared with other Chinese datasets, the Notebook dataset is small for the model to learn the characteristics of the data in the notebook domain. Therefore, the ATE effect on the Notebook dataset is not very stable and the APC performance is out of expectation. Both the performance of ATE and APC in the Chinese datasets are superior to the classical English datasets because the Chinese datasets only contain positive and negative polarities, and the Chinese review is generally shorter without interference from emotional information of other aspects. From the experimental results, BERT contributes a lot to the improvement of model effect, therefore LCF-ATEPC obtains better performance than other methods. Compared with the BERT-BASE model, the effect of LCF-ATEPC improves significantly (approximately improves by 1–2%). We speculate that LCF-ATEPC will perform more advanced than BERT-BASE in Chinese datasets with triple polarities, either ATE or APC.

4.3.2. Performance on english datasets

Table 5 lists the main experimental results of LCF-ATEPC and other ABSA-oriented models on the English and multilingual datasets. The multilingual dataset is pretty large so each experiment on this dataset only performed once. We listed representative BERT based APC methods, and implement the multi-task ABSA model using BERT-BASE for comparison. Leave alone the APC effect on the Laptop dataset slightly behind SDGCN-BERT, the LCF-ATEPC almost obtains the best performance on multiple datasets. Moreover, our model is substantially ahead of BERT-BASE in the evaluations of ATE and APC.

Experimental results indicate that not only LCF-ATEPC outperforms state-of-the-art performance on four Chinese review datasets, but the experimental results on the multilingual dataset also show that our model is capable of extracting multilingual aspects and classifying sentiments. Due to the difference in types of sentiment polarity and the capacity between Chinese and English datasets, the experimental effect needs to be further improved. Hence, the multilingual ATE and APC are still a task worthy of attention and research.

4.3.3. Overall performance analysis

There is a bias for each training since the datasets of ABSA are usually small. To address this problem, all experiments in this paper were performed five times, and the best effect and the worst effect are removed before statistics. According to Table 4, it can be seen that LCF-ATEPC obtains better ATE performance in Chinese datasets, which improves by 1–2% compared with the baseline, and English ATE has also improved significantly. However, the achievement of Chinese APC is not good enough. According to our analysis, this reason is Chinese texts are generally shorter than

Table 4
Experimental results (%) of the LCF-ATEPC on four Chinese datasets. $F1_{ate}$, Acc_{apc} and $F1_{apc}$ are the macro-F1 score of ATE subtask, accuracy and macro-F1 score of the APC subtask. The unreported experimental results are denoted as "–". The optimal results are in **bold**.

Model	Car			Phone			Notebook			Camera		
	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$
ATAE–LSTM	–	81.90	76.88	–	85.77	83.87	–	83.47	82.14	–	85.54	84.09
ATSM-S	–	82.94	64.18	–	84.86	75.35	–	75.59	60.09	–	82.88	72.50
GANN	–	83.71	77.66	–	89.17	88.16	–	82.65	82.16	–	87.99	86.75
BERT–BASE	86.32	96.96	96.3	87.89	96.98	96.45	84.05	**93.5**	92.38	84	**96.86**	**95.86**
LCF–ATEPC–CDM	87.52	**97.36**	**96.98**	89.05	97.04	96.17	86.78	92.69	91.35	84.21	96.78	95.77
LCF–ATEPC–CDW	87.27	96.52	95.79	**89.62**	96.98	96.48	86.38	**93.5**	92.32	85.20	96.7	95.76
LCF–ATEPC–Fusion	**87.57**	96.81	96.14	90.2	**97.05**	**96.57**	**87.65**	**93.5**	**92.45**	**86.29**	96.63	95.6

Table 5
Experimental results (%) of the LCF–ATEPC on English and multilingual datasets. The results of APC are obtained by using BERT–SPC to extract global context features and the "†" means the F1 score of ATE is not available for BERT–SPC and the optimal results are in **bold**.

Model	Laptop			Restaurant			Twitter			Multilingual		
	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$
BERT–PT	–	78.07	75.08	–	84.95	76.96	–	–	–	–	–	–
AEN–BERT	–	79.93	76.31	–	83.12	73.76	–	74.71	73.13	–	–	–
SDGCN-BERT	–	**81.35**	**78.34**	–	83.57	76.47	–	–	–	–	–	–
BERT–BASE	81.78	78.3	74.29	87.89	82.48	73.91	95.4	75.83	74.45	85.35	81.08	76.63
BERT–SPC	†	78.93	75.38	†	85.25	78.33	†	76.27	**75.16**	†	81.08	77.04
LCF–ATEPC–CDM	82.91	80.03	76.6	88.49	86.06	80.22	95.98	75.11	74.06	85.24	**82.39**	**78.27**
LCF–ATEPC–CDW	83.1	80.03	77.17	88.65	85.97	80.42	96.12	76.56	75.04	85.35	81.43	77.31
LCF–ATEPC–Fusion	**83.82**	80.97	77.86	**89.02**	**86.77**	**80.54**	**96.4**	**76.7**	74.54	**85.4**	81.77	77.38

English texts, and many texts are too short to identify the local context, so LCF cannot fully play its role.

Table 5 lists some joint models based on the traditional neural network to conduct APC and ATE synchronously. We adapt the BERT-BASE model to multi-task learning of ATE and APC to test the effectiveness of LCF. After optimizing the model parameters according to the experiments, this model achieved hopeful performance on some datasets and even surpassed other BERT-refined models, such as BERT-PT, AEN-BERT. Compared with other approaches, experiments on seven datasets show that LCF-ATEPC achieves a leading performance. Generally, LCF-CDM and LCF-CDW perform better in Chinese datasets, while LCF-Fusion works better in English datasets. Compared with BERT-BASE, LCF-ATEPC brings significant improvement of ATE and APC performance on English datasets. For the Chinese review datasets, BERT-BASE performs considerably on the Car and Notebook datasets.

LCF combines the local context and global context features of an aspect to infer the polarity. The empirical explanation is that multiple aspects in the global context usually contain the similar sentiment, either positive, neutral or negative, which may be due to user's bias (e.g. for the Laptop and Restaurant datasets, customers occasionally have a "global" opinion in a review. e.g., if the customer is not satisfied with one aspect, it is likely to criticize other aspects. Things will be the same if a customer prefers a restaurant he would be tolerant of some small disamenity). At the same time, global context focus can avoid the interference of emotional information from multiple aspects, which is also one of the principles of LCF mechanism. In that case, the global context features are indispensable and can promote the APC effect. In the multi-task learning process, the convergence rate of APC and ATE performance is different, so the model does not achieve the optimal effect synchronously. Compared with the BERT-BASE model, BERT-SPC significantly improves the accuracy and F1 score of APC. Besides, for the first time, LCF-ATEPC increases the F1 score of ATE subtask on the Laptop, Restaurant, Twitter dataset up to 83%, 89%, 96%, respectively. ATEPC-Fusion is supplementary of the LCF mechanism, which adopts a moderate strategy to generate local context features. However, experiments indicate that it an important design for English datasets.

4.3.4. Effectiveness of multi-task learning

We ablate LCF-ATEPC by disabling the APC or ATE module to explore the difference between the optimal performance of a single task and the multi-task learning.[8]

Fig. 6 depicts the performance of LCF-ATEPC when performing a single APC or ATE task. Benefiting from the joint training, the LCF-ATEPC obtains a better APC effect on some datasets by multi-task learning, especially the LCF-Fusion. However, LCF-ATE[9] obtains a better F1 score than multi-task learning of APC and ATE. According to our analysis, while optimizing the multi-task model through back-propagation, the model needs to consider multiple loss functions of the different subtasks. So sometimes the multi-task learning cannot achieve as the best effect as single-task learning does, which is possible occurs in many multi-task learning models. Nevertheless, LCF-ATEPC can automatically extract aspects and predicts the aspect polarity, which is unreachable for the single-task model. And the LCF-ATEPC is still superior to other ABSA-oriented multi-task models and even the single-task models aim to APC or ATE Table 6.

4.3.5. Domain-adaption for LCF-ATEPC

The BERT-BASE is pretrained on a large-scale general corpus, so fine-tuning of the BERT-shared layer during the training process is important. Meanwhile, the commonly ABSA datasets are generally small with the dataset-specific characteristic, the effect of LCF-ATEPC on the ABSA datasets can be further improved through domain adaption. Domain adaption is an effective technique while integrating the pretrained BERTs. By further training the BERT in a domain-related corpus related to the target ABSA dataset, then

[8] The loss function of the LCF-ATEPC is set to $\mathcal{L}_{apc}$ and $\mathcal{L}_{ate}$ while optimizing for APC and ATE task, respectively.

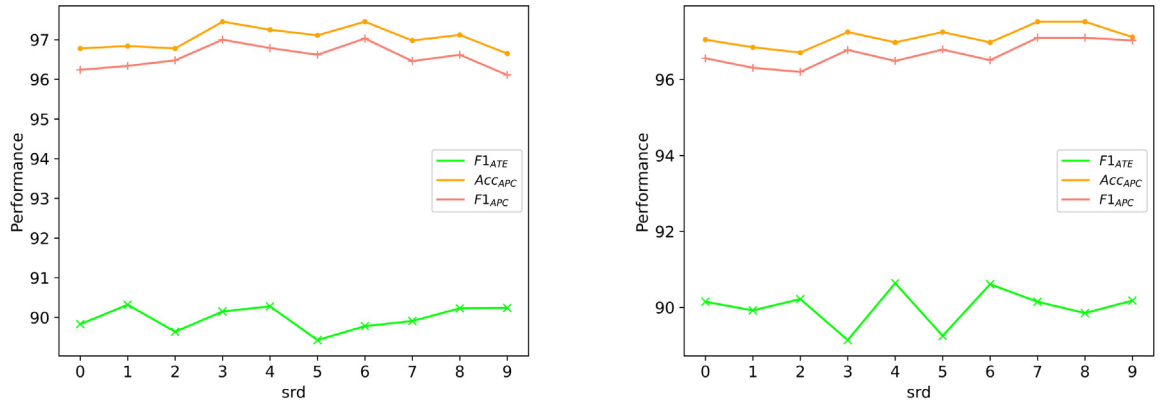
[9] The ATE module does not rely on the LCF mechanisms, so the specific LCF mode is no longer distinguished when the APC module is disabled.

Table 6
The comparison between multi-task and single-task learning approaches. The "‡" means these results are not considered and the optimal results are in **bold**.

Model	Laptop			Restaurant			Twitter		
	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$
LCF–ATEPC–CDM	82.06	80.03	76.6	88.49	86.06	80.22	95.98	75.11	74.06
LCF–ATEPC–CDW	81.61	80.03	77.17	88.65	85.97	80.42	96.12	76.56	**75.04**
LCF–ATEPC–Fusion	83.82	**80.97**	**77.86**	89.02	**86.77**	80.54	96.4	**76.7**	74.54
LCF–ATE	**84.64**	‡	‡	**89.53**	‡	‡	**97.77**	‡	‡
LCF–APC–CDM	‡	80.19	76.52	‡	85.7	79.99	‡	75.4	74.43
LCF–APC–CDW	‡	80.35	76.23	‡	85.79	79.17	‡	75.98	74.49
LCF–APC–Fusion	‡	80.5	77.77	‡	86.15	**80.76**	‡	75.54	74.55

Table 7
Experimental results (%) of LCF–ATEPC on Laptop and Restaurant datasets using domain-adapted pretrained BERT. We calculate the standard deviation of each experiment and list them in brackets.

Model	Laptop			Restaurant		
	$F1_{ate}$	Acc_{apc}	$F1_{apc}$	$F1_{ate}$	Acc_{apc}	$F1_{apc}$
BERT–BASE	84.1 (0.86)	79.72 (0.27)	75.18 (0.13)	89.76 (0.37)	85.97 (0.67)	77.68 (0.29)
BERT–SPC	†	81.03 (0.75)	77.47 (0.61)	†	88.83 (0.30)	83.52 (0.38)
BERT–ADA	–	80.23	75.77	–	87.14	80.09
LCF–ATEPC–CDM	**85.29** (0.29)	**81.81** (0.86)	**79.84** (1.10)	89.78 (0.33)	**90.18** (0.08)	**85.88** (0.18)
LCF–ATEPC–CDW	85.24 (0.43)	81.76 (0.71)	78.06 (0.57)	**89.99** (0.43)	88.65 (0.15)	83.70 (0.14)
LCF–ATEPC–Fusion	85.06 (0.60)	81.76 (0.20)	78.65 (0.57)	89.94 (0.61)	89.45 (0.48)	84.76 (0.53)



(a) Phone (b) Phone

Fig. 5. (a) and (b) depict the performance of LCF–ATEPC–CDM and LCF–ATEPC–CDW on the Chinese Phone dataset under different α, respectively.

domain-related pretrained BERTs can be obtained. We adopted the method proposed in [37] to obtain the domain-adapted pretrained BERTs based on the corpus of the Yelp Dataset Challenge Reviews[10] and the Amazon Laptops review dataset [52]. Table 7 shows that the performance of the APC task significantly improved by domain-adapted BERT-BASE. The accuracy in Restaurant dataset achieves 90.18%, which means that the LCF-ATEPC is the first ABSA model obtained up to 90% accuracy on the Restaurant dataset. Besides, experiments on the Laptop dataset also indicates the effectiveness of domain adaption in multi-task learning.

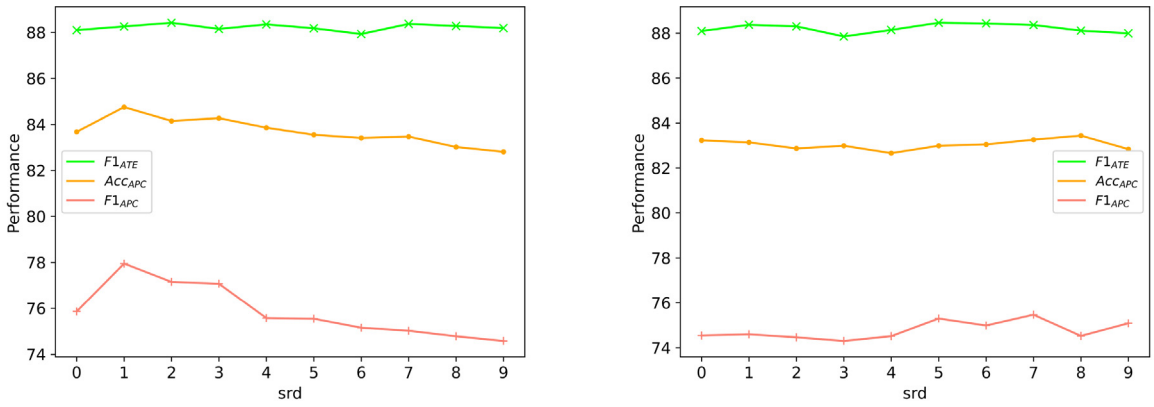
4.3.6. SRD sensitivity on different datasets

Although, the LCF mechanism improves the model's ability to extract local context features and achieves considerable results on most datasets. The extraction of local context features depends on the SRD threshold (α). Based on the experimental results, the optimal α for different datasets (e.g., Chinese and English) is slightly different. This section aims to explore the sensitivity of α on the typical Chinese and English ABSA datasets: the Phone dataset and the Restaurant dataset, respectively. Besides, we adopt the domain-adapted BERT-BASE as the shared layer of the LCF-ATEPC on the Restaurant dataset. The experimental results of Figs. 5 and 6 are obtained in multi-task learning process.

Fig. 5 shows that the superior accuracy and F1 score of APC of LCF–ATEPC–CDM on Phone dataset achieves when $\alpha = 3$ or $\alpha = 6$. And the F1 score of ATE reaches the peak while $\alpha = 1$. The LCF–ATEPC–CDW obtains the optimal APC performance on the Phone dataset while α is 4 or 6, while the competent F1 score of ATE is obtained under $\alpha = 7$ or $\alpha = 8$.

As for the Restaurant dataset, the LC-F-ATEPC–CDM achieves optimal accuracy and F1 score of APC while the $\alpha = 2$. While $\alpha = 7$ the CDW obtains better APC accuracy on the Restaurant dataset. However, the ATE F1 score and APC accuracy is not very sensitive to α in Restaurant dataset with CDW. Beyond our expectations, the optimal SRD curve is not convex or concave. This is probably because the commonly used datasets of ABSA are small

[10] This corpus is available at https://www.yelp.com/dataset/challenge.



(a) Restaurant (b) Restaurant

Fig. 6. (a) and (b) present the performance of LCF–ATEPC–CDM and LCF–ATEPC–CDW on the Restaurant dataset under different α, respectively.

and unbalanced. This problem will be alleviated when there is sufficient data.

5. Conclusion and future works

The ATE and APC were usually treated as independent tasks in previous studies since multi-task learning of ATE and APC failed to attract enough attention. Besides, the studies about Chinese language-oriented ABSA are not sufficient and urgent to be proposed and developed. To address the above problems, this paper proposes a multi-task learning model LCF-ATEPC, which is based on self-attention and local context focus and applies the pretrained BERT to the Chinese ABSA for the first time. Not only for the Chinese language, but the LCF-ATEPC is multilingual and applicable to English, such as the SemEval-2014 task4. The LCF-ATEPC can automatically extract aspect terms from reviews and infer aspects' polarity. The experimental results on three commonly English datasets and four Chinese review datasets show the LCF-ATEPC achieves new state-of-the-art performance on both ATE and APC. For future works, we are exploring the potential of our model on some fine-grained natural language processing tasks, especially sarcasm detection. The preliminary experimental results indicate that our model works competently sarcasm detection, which usually regarded as a shared task of ABSA.

CRediT authorship contribution statement

Heng Yang: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing. **Biqing Zeng:** Conceptualization, Data curation, Writing - original draft. **Jianhao Yang:** Investigation, Resources, Visualization, Validation. **Youwei Song:** Data curation, Writing - review & editing. **Ruyang Xu:** Formal analysis, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

Thanks to the anonymous reviewers and the scholars who helped us. This research is 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; The Guangdong General Colleges and Universities Special Projects in Key Areas of Artificial Intelligence of China, Research and Application of Key Techniques of Sentiment analysis, project number: 2019KZDZX1033. And this research is supported by the Innovation Project of Graduate School of South China Normal University, project number: 2019LKXM038.

References

[1] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 27–35, https://doi.org/10.3115/v1/S14-2004.

[2] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, SemEval-2015 task 12: Aspect based sentiment analysis, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 486–495, https://doi.org/10.18653/v1/S15-2082.

[3] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, G. Eryigit, SemEval-2016 task 5: Aspect based sentiment analysis, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 19–30. doi:10.18653/v1/S16-1002. URL https://www.aclweb.org/anthology/S16-1002.

[4] E. Cambria, Affective computing and sentiment analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107, https://doi.org/10.1109/MIS.2016.31.

[5] D.-T. Vo, Y. Zhang, Target-dependent twitter sentiment classification with rich automatic features, in: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, AAAI Press, 2015, pp. 1347–1353. URL http://dl.acm.org/citation.cfm?id=2832415.2832437.

[6] J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, L. Tounsi, Dcu: Aspect-based polarity classification for semeval task 4 (2014). doi:10.3115/v1/s14-2036.

[7] D. Tang, B. Qin, X. Feng, T. Liu, Effective LSTMs for target-dependent sentiment classification, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 3298–3307. URL https://www.aclweb.org/anthology/C16-1311.

[8] Y. Wang, M. Huang, X. Zhu, L. Zhao, Attention-based lstm for aspect-level sentiment classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 606–615, https://doi.org/10.18653/v1/d16-1058.

[9] D. Ma, S. Li, X. Zhang, H. Wang, Interactive attention networks for aspect-level sentiment classification, in, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 4068–4074, https://doi.org/10.24963/ijcai.2017/568.

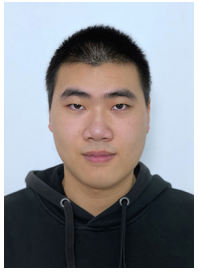
[10] F. Fan, Y. Feng, D. Zhao, Multi-grained attention network for aspect-level sentiment classification, in, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3433–3442, https://doi.org/10.18653/v1/D18-1380.

[11] Y. Song, J. Wang, T. Jiang, Z. Liu, Y. Rao, Attentional encoder network for targeted sentiment classification, arXiv preprint arXiv:1902.09314 (2019). https://arxiv.org/abs/1902.09314.

[12] B. Zeng, H. Yang, R. Xu, W. Zhou, X. Han, Lcf: A local context focus mechanism for aspect-based sentiment classification, Applied Sciences 9 (16) (2019) 3389, https://doi.org/10.3390/app9163389.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., USA, 2017, pp. 6000–6010. URL http://dl.acm.org/citation.cfm?id=3295222.3295349.
[14] E. F. T. K. Sang, F. De Meulder, Introduction to the conll-2003 shared task: language-independent named entity recognition (2003) 142–147 doi:10.1016/j.eswa.2016.10.065.
[15] P. Chen, Z. Sun, L. Bing, W. Yang, Recurrent attention network on memory for aspect sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2017, pp. 452–461, https://doi.org/10.18653/v1/D17-1047.
[16] T. Chen, R. Xu, Y. He, X. Wang, Improving sentiment analysis via sentence type classification using bilstm-crf and cnn, Expert Systems with Applications 72 (2017) 221–230, https://doi.org/10.1016/j.eswa.2016.10.065.
[17] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2514–2523. doi:10.18653/v1/P18-1234. https://www.aclweb.org/anthology/P18-1234.
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.
[19] S. Poria, E. Cambria, L.-W. Ku, C. Gui, A. Gelbukh, A rule-based approach to aspect extraction from product reviews, in: Proceedings of the second workshop on natural language processing for social media (SocialNLP), 2014, pp. 28–37. doi:10.3115/v1/w14-5905..
[20] Q. Liu, Z. Gao, B. Liu, Y. Zhang, Automated rule selection for aspect extraction in opinion mining, in: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, AAAI Press, 2015, pp. 1291–1297. URL http://dl.acm.org/citation.cfm?id=2832415.2832429.
[21] A. Mukherjee, B. Liu, Aspect extraction through semi-supervised modeling, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – vol. 1, ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA, 2012, pp. 339–348. http://dl.acm.org/citation.cfm?id=2390524.2390572.
[22] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, Knowledge-Based Systems 108 (2016) 42–49, https://doi.org/10.1016/j.knosys.2016.06.00.
[23] R. He, W.S. Lee, H.T. Ng, D. Dahlmeier, An unsupervised neural attention model for aspect extraction, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), 2017, pp. 388–397. doi:10.18653/v1/p17-1036.
[24] W. Wang, S.J. Pan, D. Dahlmeier, X. Xiao, Coupled multi-layer attentions for co-extraction of aspect and opinion terms, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, pp. 3316–3322.
[25] J. Zhu, H. Wang, M. Zhu, B.K. Tsou, M. Ma, Aspect-based opinion polling from customer reviews, IEEE Transactions on Affective Computing 2 (1) (2011) 37–49, https://doi.org/10.32657/10356/61830.
[26] Y. Zhao, S. Dong, J. Yang, Effect research of aspects extraction for chinese hotel reviews based on machine learning method, International Journal of Smart Home 9 (2015) 23–34, https://doi.org/10.14257/ijsh.2015.9.3.03.
[27] X. Li, L. Bing, W. Lam, B. Shi, Transformation networks for target-oriented sentiment classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers), 2018, pp. 946–956. doi:10.18653/v1/p18-1087.
[28] B. Huang, Y. Ou, K.M. Carley, Aspect level sentiment classification with attention-over-attention neural networks, in: R. Thomson, C. Dancy, A. Hyder, H. Bisgin (Eds.), Social, Cultural, and Behavioral Modeling, Springer International Publishing, Cham, 2018, pp. 197–206, https://link.springer.com/chapter/10.1007/978-3-319-93372-6_22.
[29] Z. Zhang, Y. Zou, C. Gan, Textual sentiment analysis via three different attention convolutional neural networks and cross-modality consistent regression, Neurocomputing 275 (2018) 1407–1415, https://doi.org/10.1016/j.neucom.2017.09.080, http://www.sciencedirect.com/science/article/pii/S0925231217316090.
[30] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014). https://arxiv.org/abs/1409.0473.
[31] H. Peng, Y. Ma, Y. Li, E. Cambria, Learning multi-grained aspect target sequence for chinese sentiment analysis, Knowledge-Based Systems 148 (2018) 167–176, https://doi.org/10.1016/j.knosys.2018.02.034.
[32] Y. Ma, H. Peng, E. Cambria, Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm (2018). https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16541.
[33] N. Liu, B. Shen, Aspect-based sentiment analysis with gated alternate neural network, Knowledge-Based Systems (2019), https://doi.org/10.1016/j.knosys.2019.105010, 105010.
[34] X. Tan, Y. Cai, J. Xu, H.-F. Leung, W. Chen, Q. Li, Improving aspect-based sentiment analysis via aligning aspect embedding, Neurocomputing 383 (2020) 336–347, https://doi.org/10.1016/j.neucom.2019.12.035, http://www.sciencedirect.com/science/article/pii/S0925231219317382.
[35] H. Peng, E. Cambria, A. Hussain, A review of sentiment analysis research in chinese language, Cognitive Computation 9 (4) (2017) 423–435, https://doi.org/10.1007/s12559-017-9470-8.
[36] H. Peng, E. Cambria, X. Zou, Radical-based hierarchical embeddings for chinese sentiment analysis at sentence level (2017). https://aaai.org/ocs/index.php/FLAIRS/FLAIRS17/paper/view/15460.
[37] A. Rietzler, S. Stabinger, P. Opitz, S. Engl, Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification, arXiv preprint arXiv:1908.11860 (2019). https://arxiv.org/abs/1908.11860.
[38] H. Nguyen, K. Shirai, A joint model of term extraction and polarity classification for aspect-based sentiment analysis, in: 2018 10th International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2018, pp. 323–328. doi:10.1109/kse.2018.8573340.
[39] X. Li, L. Bing, P. Li, W. Lam, A unified model for opinion target extraction and target sentiment prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 6714–6721. doi: 10.1609/aaai.v33i01.33016714.
[40] F. Chen, Y. Huang, Knowledge-enhanced neural networks for sentiment analysis of chinese reviews, Neurocomputing 368 (2019) 51–58, https://doi.org/10.1016/j.neucom.2019.08.054, http://www.sciencedirect.com/science/article/pii/S0925231219311920.
[41] M.S. Akhtar, T. Garg, A. Ekbal, Multi-task learning for aspect term extraction and aspect sentiment classification, Neurocomputing (2020), https://doi.org/10.1016/j.neucom.2020.02.093, http://www.sciencedirect.com/science/article/pii/S0925231220302897.
[42] M. Hu, Y. Peng, Z. Huang, D. Li, Y. Lv, Open-domain targeted sentiment analysis via span-based extraction and classification, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 537–546.
[43] R. He, W.S. Lee, H.T. Ng, D. Dahlmeier, An interactive multi-task learning network for end-to-end aspect-based sentiment analysis, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 504–515.
[44] Z. Zeng, J. Ma, M. Chen, X. Li, Joint learning for aspect category detection and sentiment analysis in chinese reviews, in: China Conference on Information Retrieval, Springer, 2019, pp. 108–120. doi:10.1007/978-3-030-31624-2_9.
[45] M.H. Phan, P.O. Ogunbona, Modelling context and syntactical features for aspect-based sentiment analysis, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3211–3220.
[46] H. Zhao, L. Huang, R. Zhang, Q. Lu, et al., Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction, in, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3239–3248.
[47] D.S. Chauhan, S. Dhanush, A. Ekbal, P. Bhattacharyya, Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4351–4360.
[48] W. Che, Y. Zhao, H. Guo, Z. Su, T. Liu, Sentence compression for aspect-based sentiment analysis, IEEE/ACM Transactions on audio, speech, and language processing 23 (12) (2015) 2111–2124, https://doi.org/10.1109/TASLP.2015.2443982.
[49] Y. Zhao, H. Pan, C. Du, Y. Zheng, Principal curvature for infrared small target detection, Infrared Physics & Technology 69 (2015) 36–43, https://doi.org/10.1016/j.infrared.2014.12.014, http://www.sciencedirect.com/science/article/pii/S1350449514002825.
[50] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, K. Xu, Adaptive recursive neural network for target-dependent twitter sentiment classification, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (vol. 2: Short papers), 2014, pp. 49–54, https://doi.org/10.3115/v1/p14-2009.
[51] H. Xu, B. Liu, L. Shu, P. Yu, 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, vol. 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2324–2335, https://doi.org/10.18653/v1/N19-1242.
[52] R. He, J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: Proceedings of the 25th International Conference on World Wide Web, WWW '16, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2016, pp. 507–517. doi:10.1145/2872427.2883037. doi: 10.1145/2872427.2883037.

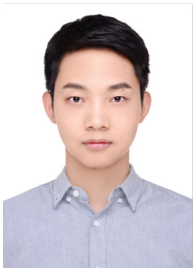
Heng Yang is pursuing a master's degree in computer science and technology from South China Normal University. His research interests are natural language processing, sentiment classification, and entity extraction.

Biqing Zeng is a professor of the School of Software at South China Normal University. He obtained his Ph.D. in Computer Science at Central South University. His interests are natural language processing, artificial intelligence, big data.



Jianhao Yang is pursuing a master's degree in the Software School of South China Normal University. His research interests are natural language processing, sentiment analysis, and recommendation systems.

Youwei Song obtained a master's degree in the School of Data and Computer science at Sun Yat-Sen University. His research interests are sentiment analysis and natural language processing.



Ruyang Xu is pursuing a master's degree in the School of Software at South China Normal University. His research interests include natural language processing, artificial intelligence.