IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 14, 2018, accepted November 28, 2018, date of publication December 5, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885032

Syntax-Directed Hybrid Attention Network for Aspect-Level Sentiment Analysis

XINYI WANG[1,2,3], GUANGLUAN XU[1,3], JINGYUAN ZHANG[1,2,3], XIAN SUN[1,3], LEI WANG[1,3], AND TINGLEI HUANG[1,3,4]

[1]Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China
[3]Key Laboratory of Technology in Geospatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China
[4]Institute of Electronics, Chinese Academy of Sciences, Suzhou 215123, China

Corresponding author: Tinglei Huang (tlhuang@mail.ie.ac.cn)

This work was supported by the Foundation of Precision Poverty Alleviation Technology under Grant 42-Y30B12-9001-17/18.

ABSTRACT Aspect-level sentiment analysis is a fine-grained task in sentiment analysis that aims at detecting sentiment polarity towards a specific target in a sentence. Previous studies focus on using global attention mechanism that attends to all words in the context to model the interaction between target and sentence. However, global attention suffers from assigning high-attention score to irrelevant sentiment words in the cases where sentence contains noisy words or multiple targets. To address this problem, we propose a novel syntax-directed hybrid attention network (SHAN). In SHAN, a global attention is employed to capture coarse information about the target, and a syntax-directed local attention is used to take a look at words syntactically close to the target. An information gate is then utilized to synthesize the information from local and global attention results and adaptively generate a less-noisy and more sentiment-oriented representation. The experimental results on SemEval 2014 Datasets demonstrate the effectiveness of the proposed method.

INDEX TERMS Aspect-level sentiment analysis, hybrid attention, syntactic information, gating mechanism.

I. INTRODUCTION

Aspect-level sentiment analysis is a fundamental task in sentiment analysis [1], [2]. Different from sentence-level or document-level sentiment analysis, aspect-level sentiment analysis is more fine-grained. Given a sentence and a target which is a text span in the sentence, aspect-level sentiment analysis aims at predicting sentiment polarity in the sentence towards the specific target. For example, in the sentence "*Great food but the service was dreadful!*", sentiment polarity is positive when target is "food", but sentiment polarity becomes negative if "service" is treated as target.

As an important research topic in natural language processing, aspect-level sentiment analysis is attracting the attention of researchers. Conventional approaches focus on extracting handcraft features such as sentiment lexicon and bag-of words features. These features are then used to train a classifier [3]–[5]. However, designing these features is labor-intensive and these methods are highly dependent on the quality of the selected features. Different from conventional approaches, neural network models [6]–[8] learn sentence representation directly from the text which are more labor-saving and scalable.

With attention mechanism showing promising result in NLP tasks [9]–[12], attention-based methods [13]–[16] are proposed for the task of aspect-level sentiment analysis. In these works, global attention mechanism is operated to attend to all words in the context and get corresponding weight of each word given a specific target. However, one drawback of global attention is that it may introduce noise and downgrade the prediction accuracy in cases where sentence contains noisy words or multiple targets. For example, in sentence "*The **wait staff** is very friendly, if you are not rude or picky.*", global attention approaches tend to assign high attention score to "rude" which is semantically meaningful for the target when they highlight the opinion modifier "friendly". To alleviate this problem, [15], [16] attempt to assign attention weights to context words according to the distance of each word to the target in the sentence based on the intuition that words close to the target should bring more target-specific information than a further one. However, the assumption is kind of arbitrary and not always true in

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

some cases. For example, in sentence "*apple is unmatched in product quality, aesthetics, craftmanship, and* ***customer service***", the opinion modifier "unmatched" lies far from the targets "customer service".

Different from global attention, local attention mechanism, which is firstly proposed by [17] in image caption, only looks at a subset of words in a sentence. As [18] point out, sentiment of a given target is usually determined by key phrase in the context. In other words, the target-specific sentiment information is dominant by a subset of words, which coincides with the spirit of local attention. However, since there is no such labeled sentiment scope information in the training data, it is difficult to accurately predict the sentiment-oriented text span without any supervision. Moreover, selecting local attention words by rules is difficult and tending to leave out useful sentiment information because of diverse expression of human language.

Since both local and global attention have their own strengths and drawbacks in dealing with aspect-level sentiment analysis task, it is natural to consider combining the advantages from either of them. For local attention, it is considered to capture target-orient sentiment information with less noisy words. As [19] put, dependency tree contains rich linguistics information between words, which can potentially capture syntactic dependencies between words and their contexts. Therefore, it is reasonable to select words for local attention based on their syntactic relatedness to target words. For global attention, it has the ability to make up for the information left out by local attention with the help of attending to all words in the context.

The most commonly used method to combine the information from different sources is concatenation [14], [20]. However, concatenating local and global attention results is technically problematic. If local attention result is useful but global attention result is noisy, the global attention information can potentially bias the entire representations. Similarly, if global attention is useful but local attention is useless, the entire representation is also biased. Therefore, a more flexible way is needed to control how much the information from local and global attentions are flowed to the final sentence representation. To this end, gating mechanism, which is used to control the flow of information [21]–[23], is right on the target. Moreover, compared to concatenation, the gating mechanism provides a more interpretable way for identifying the importance of each words in the sentence for the final prediction.

Based on the analysis above, we propose a novel architecture called **S**yntax-directed **H**ybrid **A**ttention **N**etwork (SHAN). The model firstly employs global and local attention to capture coarse sentiment information in the whole context and syntactic information close to the target, respectively. In order to select words for local attention, we define syntax-based words distance upon dependency trees, and select words near the target words based on the defined word distance. Secondly, an information gate is operated to adaptively evaluate local and global attention results based on target-dependent information. Finally, attention score for each word is reweighted by combining global attention weight with local attention weight according to the output of information gate, leading to a more sentiment-specific and target-oriented sentence representation.

The main contributions of this work are summarized as follows:

1) We introduce a hybrid attention mechanism to understand sentence for a given target in both local and global views. Syntax-directed local attention selects words on a dependency tree, which allows the model to incorporate syntactic information into attention results. Global attention attends to all words in the sentence, which is able to make up for the information left out by local attention.
2) We propose a gating mechanism to distill target-oriented information for performing aspect-level sentiment analysis. This gating mechanism can dynamically leverage local and global attention results by evaluating the importance of each attention result with gating mechanism, leading to a less-noisy and more target-specific sentence representation.
3) Experimental results show that our model outperforms both global and local attention based models, indicating that our model can generate powerful context representation by dynamically combining local and global attention results.

II. RELATED WORK

A. ASPECT BASED SENTIMENT ANALYSIS

Aspect-level sentiment analysis is an important research topic in the field of sentiment analysis. Traditional machine learning approaches try to solve this problem by extracting handcraft features such as sentiment lexicons and bag-of-words features. The extracted features are used to train a sentiment classifier like SVM [4], [5]. However, the performances of these methods usually highly depend on the quality of the handcraft features. Besides, designing these features is labor-intensive.

As neural networks have the capability of learning powerful text representation through multiple hidden layers, neural network methods has been proposed. Reference [6] use Recursive Neural networks to incorporate the target information into the feature learning with dependency trees. Reference [8] propose target-dependent LSTM which uses two LSTMs on left and right side of target words respectively to model the relatedness of a target word with its context words. Reference [24] build a three-way gated neural network to model the interaction between the left context, the right context and the target. Despite the advantages of neural network based models, they are not good at capturing long-range information when target-specific sentiment words are far from the target.

Attention mechanism has been used successfully in machine translation [9], question answering [11] and other tasks [10], [12]. In aspect-level sentiment analysis, atten-

tion mechanism is introduced to assign attention scores to context words by modeling the interaction between target and context. Reference [13] propose attention-based LSTM (AT-LSTM) to get the key part of sentence in response to a given target. [14] propose IAN model which use two LSTM to model sentence and targets, respectively. Hidden states generated from sentence are used to calculate attention scores for target by a pooling operation, and vice versa. Reference [25] extend IAN by modeling word-wise interaction between target and context words. Reference [26] propose a hierarchical attention structure which firstly applies attentions on aspect terms to get aspect representation. Afterwards, aspect representation together with the hidden state of sentence are used for sentence attention. Reference [15] introduce Deep Memory Network (DMN) in which multi-hop attention is applied to learn representation of text with multiple levels of abstraction. In addition, [15] introduce a novel attention mechanism to get location information between context word and aspect. However, while DMN considers content and location information into attention modeling, useful lexical features are ignored. Therefore, [27] propose MNWSI by adding sentiment polarity vectors of words into word representation and incorporating part of speech information of words into attention mechanism. [16] extend Deep Memory Network by introducing a memory module between the attention module and the input module. Besides, a recurrent neural network is utilized to combine the attention result.

However, the models in [13], [14], [25], and [26] employ global attention to get the aspect-specific sentiment information, which may cause mismatching of sentiment words and targets when an irrelevant sentiment word is semantically meaningful for the target. Though [15] and [16] try to solve this problem by incorporating location information, their assumption is somewhat arbitrary and cannot deal with the variety of human language.

B. LOCAL ATTENTION MECHANISM

Different from global mechanism which attends to all words, local mechanism looks at a subset of words in the sentence. [28] firstly introduce local attention mechanism in machine translation. To generate local attention words, an aligned position p_t is firstly predicted in source sentence for each target word in step t. Then attention weight for each word within a fixed window of around p_t is calculated based on word distance to p_t. Reference [29] extend local attention with syntax-distance constraint by focusing on syntactically related words with the predicted target words. For the task of entity disambiguation, [30] define context score to measure the relevance between context words and entity candidates. In order to get local attention candidates, top K words with highest context score are selected. Reference [31] propose syntax-based local attention for aspect-level sentiment analysis by selecting words syntactically close to the target for final sentence representation.

Though [31] employ local attention to exploit syntactic information, the local attentions used in the proposed method and [31] are different. Reference [31] use local attention to select words within a larger window to keep the semantic completeness, but the local attention used in the proposed method prefers a small window size to capture relatively pure sentiment information about the target. In addition, [31] used the distance weights in local attention to highlight the words closer to the target while the proposed method utilizes a gating mechanism to dynamically reweight words in the sentence. More importantly, compared with [31], the proposed method employs both local and global attentions to generate final sentence representation, which can make up for the information left out by local attention and is more robust to parsing errors.

III. PROPOSED METHODOLOGY

We describe the proposed approach for aspect-level sentiment analysis in this section.

The task of aspect-level sentiment classification can be formulated as follows: given a sequence of n words $S = \{w_1, w_2, w_3..., w_n\}$, which is called context, and a target $T = \{w_i..., w_j\}$ which contains one or more consecutive words in the context, the goal is to figure out the sentiment polarity towards the given target in the context.

The architecture of syntax-directed hybrid-attention model is shown in Fig. 1. The model firstly maps context and target word(s) into continuous low dimensional word vectors using pretrained word embeddings in the input layer. Then Bi-LSTM is operated upon these word vectors to preserve sequential information in memory modeling layer. In hybrid-attention layer, both global and local attention are employed to couple the target vector with context vectors. Afterwards, we use an information gate to dynamically evaluate the importance of local and global attention results. A fine-grained attention score will be calculated by combining the local and global attention scores with the information gate. A more accurate sentence representation is obtained as a weighted sum of hidden states with respect to their attention scores. Finally, sentiment polarity is predicted through softmax layer. The details of each layer are described as follows.

A. INPUT LAYER

Given a sentence $S = \{w_1, w_2, w_3..., w_n\}$, each word in S is mapped into a k-dimensional vector $\mathbf{e}_i \in \mathbb{R}^k$ by looking up in a pretrained word embedding matrix $\mathbf{E} \in \mathbb{R}^{k*|V|}$, such as glove [32] and word2vec [33]:

$$\mathbf{e}_i = \mathbf{E}(w_i) \tag{1}$$

where k is the dimension of the word vector and $|V|$ is the vocabulary size. If target T only contains one word, its representation $\mathbf{v}_a$ is the embedding of target word. If target T is a phrase, $\mathbf{v}_a$ is obtained by calculating the average value of target words embeddings.

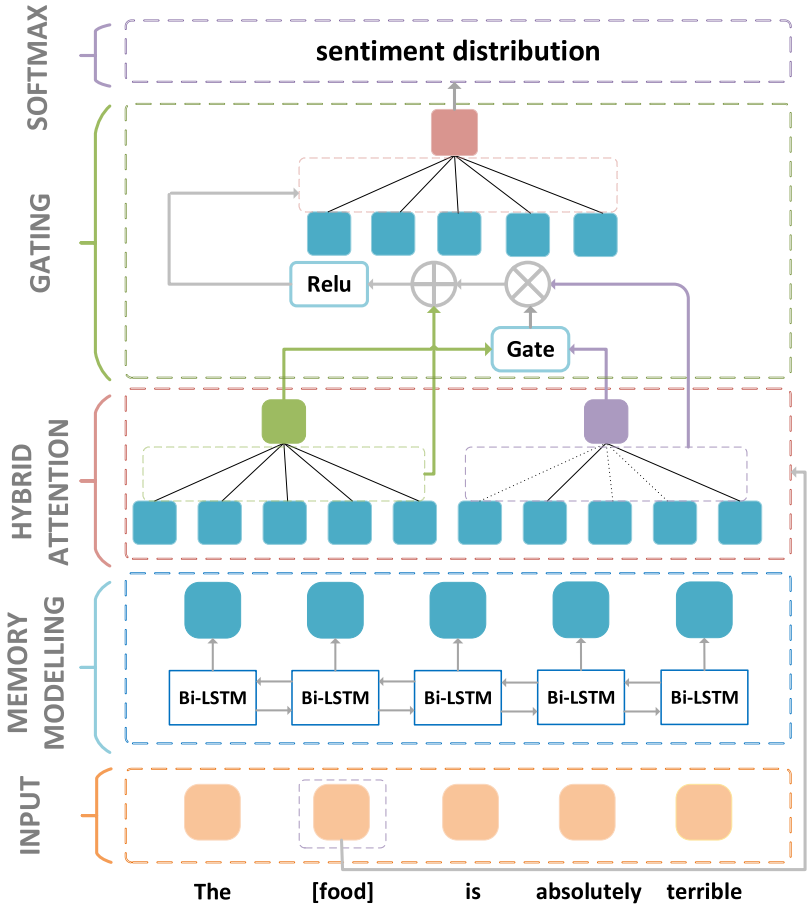


FIGURE 1. The architecture of syntax-directed hybrid attention model. The left and right part of hybrid attention layer denotes global and local attention, respectively.

B. MEMORY MODELLING LAYER

After looking up operation, the word vectors are then fed into a bidirectional long short term memory(Bi-LSTM) network [22] to encode contextual information.

For forward LSTM, given the hidden state $\overrightarrow{\mathbf{h}}_{t-1} \in \mathbb{R}^d$ and the word embedding $\mathbf{e}_t$, hidden state $\overrightarrow{\mathbf{h}}_t$ at time step t can be calculated as:

$$\overrightarrow{\mathbf{h}}_t = \overrightarrow{LSTM}(\mathbf{e}_t, \overrightarrow{\mathbf{h}}_{t-1}). \tag{2}$$

Backward LSTM does the same thing as forward LSTM except that the input sequence is fed in a reversed way. Hidden states of both the forward LSTM and backward LSTM are concatenated and hyperbolic tangent activation function is applied to the concatenation result to form the hidden state $\mathbf{h}_t \in \mathbb{R}^{2d}$ of each word:

$$\mathbf{h}_t = tanh(\left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right]) \tag{3}$$

where [;] stands for concatenate operation, $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$ are the output of forward and backward LSTM at time step t respectively. The output of the Bi-LSTM layer is denoted as $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, ..., \mathbf{h}_n\}$.

C. HYBRID ATTENTION LAYER

Hybrid attention layer is responsible for linking information from context to target and distilling the target information from the sentence. The main purpose of this layer is to understand sentence with respect to the target using information from both global and local views. Therefore, different from the previous works [13]–[16], not only global attention is employed, syntax-directed local attention is also utilized to focus on words close to target.

1) GLOBAL ATTENTION

Given target vector $\mathbf{v}_a \in \mathbb{R}^k$ and sentence representation $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, ..., \mathbf{h}_n\}$, attention score $\alpha_i \in \mathbb{R}$ for each word representation $\mathbf{h}_i$ is calculated as follows:

$$\mathbf{m}_i = \mathbf{w}_{att2}^T tanh(\mathbf{W}_{att1}[\mathbf{h}_i; \mathbf{v}_a] + \mathbf{b}_{att1}) \tag{4}$$

$$\alpha_i = \frac{\exp(\mathbf{m}_i)}{\sum_{j=1}^{n} exp(\mathbf{m}_j)} \tag{5}$$

where $\mathbf{W}_{att1} \in \mathbb{R}^{(2d+k)*(2d+k)}$ and $\mathbf{w}_{att2} \in \mathbb{R}^{2d+k}$ are weight matrices. $\mathbf{b}_{att1} \in \mathbb{R}^{(2d+k)}$ are bias. Afterwards, target-specific representation $\mathbf{r}_{glo} \in \mathbb{R}^{2d}$ is formulated as the weighted sum

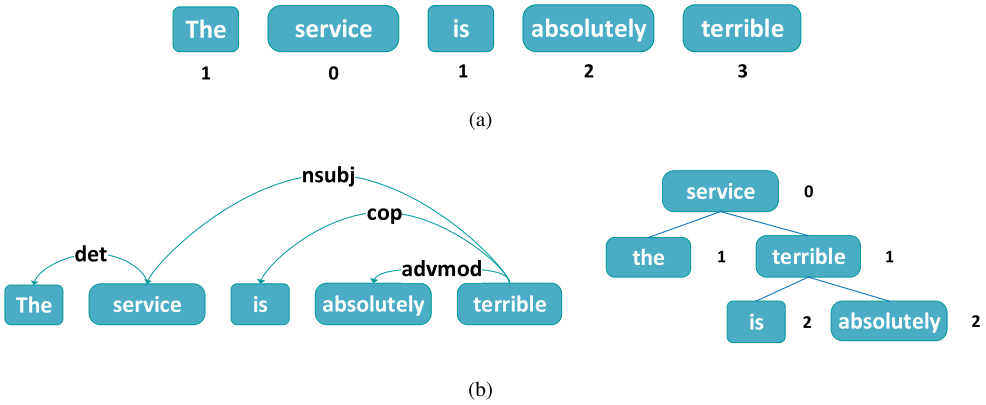


FIGURE 2. (a): Position-based distance of each word to target "service". (b): Dependency tree of sentence "The service is absolutely terrible" and the syntax-based distance of each word to target "service".

of hidden state $\mathbf{h}_i$ with respect to its attention score α_i:

$$\mathbf{r}_{glo} = \sum_{i=1}^{n} \mathbf{h}_i \alpha_i \tag{6}$$

设置一个窗口长度即可

2) SYNTAX-DIRECTED LOCAL ATTENTION

As local attention only focuses on a subset of words in the context, what we need to do first is to select words semantically close to a specific target. As dependency tree contains rich linguistics information between words, it can potentially capture syntactic dependencies between target words and their contexts. Therefore, we introduce a novel word distance defined upon dependency tree which is called syntax-based distance. Specifically, given a sentence S with its dependency tree D, each word is a node in D. The distance between two connected nodes is defined as *one*. We can traverse D to compute the distance of all remaining words to the current target word. Words are selected for local attention based on the syntax-based distance to target word(s).

Fig. 2 shows the comparison of syntax-based and position-based distance to word "service" in sentence "the service is absolutely horrible". Given a sentence S, the position-based distance of two words next to each other in the word sequence is *one*. It can be observed that the syntax-based distance of "service" to itself is *zero*, and syntax-based distance of "the" and opinion modifier "horrible" to "service" is *one*, etc. For position-based word distance, the distance of sentiment word "horrible" is larger than irrelevant words "the", "is" and "absolutely". Compared with position-based distance, syntax-based words distance is more proficient in incorporating semantic information. We can choose words within k-step syntax-based distance to target word(s) which we denote as $LS(k)$ and assign attention scores to these words as follows:

$$\mathbf{n}_i = \mathbf{w}_{att4}^T tanh(\mathbf{W}_{att3}[\mathbf{h}_i; \mathbf{v}_a] + \mathbf{b}_{att2}) \tag{7}$$

$$\beta_i = \frac{exp(\mathbf{n}_i)}{\sum_{j \in LS(k)} exp(\mathbf{n}_j)} \tag{8}$$

where $i \in LS(k)$, $\mathbf{W}_{att3} \in \mathbb{R}^{(2d+k)*(2d+k)}$,$\mathbf{w}_{att4} \in \mathbb{R}^{2d+k}$ and $\mathbf{b}_{att2} \in \mathbb{R}^{2d+k}$ are parameters to be trained. If a target contains more than one word, words within k-step distance of each target word are chosen. For the sake of presentation, we assign the local attention score for each context words as:

$$\beta_i = \begin{cases} \dfrac{exp(\mathbf{n}_i)}{\sum_{j \in LS(k)} exp(\mathbf{n}_j)} & i \in LS(k) \\ 0 & i \notin LS(k) \end{cases} \tag{9}$$

D. GATING LAYER

After global and local attention vectors are derived, in this layer, an information gate is employed to synthesize target specific information from local and global attention results. Firstly, information gate $\mathbf{g} \in \mathbb{R}^{2d}$ is calculated with both local and global attention vectors:

$$\mathbf{g} = tanh(\mathbf{W}_{gate}[\mathbf{r}_{glo}; \mathbf{r}_{loc}]) \tag{10}$$

where $\mathbf{W}_{gate} \in \mathbb{R}^{2d*4d}$ are trainable parameters. $\mathbf{r}_{glo}$ and $\mathbf{r}_{loc}$ denote global and local attention results, respectively. It is worth noticing that we adopt a vector-valued information gate instead of a scalar-valued one. As shown in [34], each dimension in word embedding could reflect different perspectives of word meaning. Similarly, we assume that each dimension of information gate controls different perspectives of the attention vector. Therefore, the representation of the context can be represented as:

$$\begin{aligned} \mathbf{r} &= \mathbf{r}_{glo} + \mathbf{r}_{loc} \odot \mathbf{g} \\ &= \sum_{i=1}^{n} \mathbf{h}_i \alpha_i + (\sum_{i=1}^{n} \mathbf{h}_i \beta_i) \odot \mathbf{g} \\ &= \sum_{i=1}^{n} \mathbf{h}_i \odot \tau_i \end{aligned} \tag{11}$$

where $\tau_i \in \mathbb{R}^{2d}$, and the jth dimension of τ_i is:

$$\tau_{ij} = \alpha_i + \beta_i \cdot g_j \tag{12}$$

where g_j stands for the jth dimension of information gate $\mathbf{g}$. However, the value of g_j can be negative due to the hyperbolic tangent function. As a result, τ_{ij} may be negative, which contradicts with our cognition of attention. To make the reweighted attention score more interpretable and meaningful, we directly combine local and global attention scores with information gate. A rectified linear unit is applied to keep attention score to be non-negative:

$$t_{ij} = relu(\alpha_i + \beta_i \cdot g_j) \tag{13}$$

Afterwards, an normalization function is applied to ensure the sum of attention scores on jth dimension over all words equals to *one*:

$$\begin{aligned} \gamma_{ij} &= normalization(t_{ij}) \\ &= \frac{t_{ij}}{\sum_{k=1}^{n} t_{kj}} \end{aligned} \tag{14}$$

where γ_{ij} is the normalized attention score for w_i on jth dimension. Finally, the representation of the context is calculated as the weighted sum of each hidden state:

$$\mathbf{r}_{final} = \sum_{i=1}^{n} \mathbf{h}_i \odot \gamma_i \tag{15}$$

We can see that information gate provides a reweighting mechanism to adaptively combine local and global attention score according to the information each attention carries. Intuitively speaking, when words used in local attention contain useful information to detect sentiment polarity, the information gate has positive values, the attention score of these words will be higher; when words close the target are meaningless for predicting target-specific sentiment, the information gate has zero or even negative values, the final representation is dominated by the words far from the target. As a result, the gating mechanism can effectively combine local and global attention result by dynamically laying more emphasis on opinion modifier and ignoring unrelated words.

E. SOFTMAX LAYERS

In softmax layer, the final sentiment polarity distribution of the given target is predicted using the target-specific representation obtained from the gating layer:

$$\mathbf{y} = softmax(\mathbf{W}_s \mathbf{r}_{final} + \mathbf{b}_s) \tag{16}$$

where $\mathbf{W}_s \in \mathbb{R}^{c*2d}$, $\mathbf{b}_s \in \mathbb{R}^c$ are parameters to be learned in softmax layer, c is the number of sentiment polarities.

F. MODEL TRAINING

In syntax-directed hybrid attention network, we need to update parameters from memory modeling, hybrid attention, gating and softmax layers. Let $\hat{y}$ denote the estimated probability distribution, and y denote the ground truth. Cross-entropy between y and $\hat{y}$ with L2 regulations are used as loss function:

$$J = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_i^j log(\hat{y}_i^j) + \lambda(\sum_{\theta \in \Theta} \theta^2) \tag{17}$$

where i is the index of sentence, j is the index of class. N is the number of training samples, C is the number of sentiment classes, λ is the L2-regularization term. Θ is the parameter set. Note that word embeddings are fixed during training to avoid overfitting.

IV. EXPERIMENT

A. EXPERIMENTAL SETTING

1) DATA PREPARATION AND EVALUATION METRIC

We conduct our experiments on two public datasets which are collected from user reviews in laptop and restaurant domains, respectively. Both of the two datasets are from SemEval 2014 Task4.[1] There are two sentiment classification subtasks in SemEval 2014 Task4, namely Aspect-Term-level classification and Aspect-level classification. The former is intended to predict sentiment polarity towards aspect terms which are consecutive words in the sentence. However, in Aspect-level classification, sentiment analysis is performed to a predefined aspect. An aspect can either appear as different aspect terms in the context, or be mentioned as implicitly without any aspect terms in the text. In this paper, the contributions we made are based on the assumption that sentiment target is a text span in the sentence. Therefore, we focus on the Aspect-Term-level classification Task in this paper. The detail of each dataset is shown in Table 1. The reviews are labeled with three polarities: positive, neutral and negative. We adopt accuracy as evaluation metric for aspect-level sentiment analysis.

TABLE 1. Statistics of datasets.

Dataset		Positive	Negative	Neural
Laptop reviews	Training	994	464	870
	Testing	341	169	128
Restaurant reviews	Training	994	464	870
	Traing	2164	637	807

2) TRAINING DETAILS

In our experiments, the word embeddings for the contexts and targets in both datasets are initialized using 300-dimensional word vectors pretrained by Glove[2] [32]. All out-of-vocabulary words are initialized by sampling from the uniform distribution $U(-0.01, 0.01)$. To generate dependency trees, sentences from both datasets are parsed by the Stanford CoreNLP.[3] All weight matrices and biases are initialized with xavier uniform distribution. The hidden states of LSTM are set to 300, and L2-regulation weight is set to 0.0001. Dropout rate before softmax layer is set to 0.5. We use Adam optimizer with a batch size 32, and initial learning rate is 0.001. Note that it is a well-known issue that the performance fluctuates with different random initializations. To alleviate this problem, the reported accuracy of the

[1] http://alt.qcri.org/semeval2014/task4/
[2] https://nlp.stanford.edu/projects/glove/
[3] https://stanfordnlp.github.io/CoreNLP/

proposed method is obtained as average value over 5 runs with random initializations.

B. MODEL COMPARISONS

To comprehensively evaluate the performance of SHAN, we compare our proposed model against the following models:

- **SVM** [5]: It extracts lexicon features, surface features and parsing features to train SVM. This model achieves the best result in SemEval 2014 task 4.
- **LSTM**: It only uses one LSTM network to model the context. The average value of all hidden states is treated as final context representation.
- **TD-LSTM** [8]: TD-LSTM uses a forward and a backward LSTM to capture the information on the left and right side of the target, respectively. The hidden states of the two LSTMs at last time step are concatenated to represent the context.
- **AT-LSTM** [13]: AT-LSTM applies attention mechanism on the top of hidden states of LSTM. The final representation is the concatenation of attention result and final state of LSTM.
- **IAN** [14]: IAN uses two LSTM networks to model sentences and targets, respectively. Hidden states generated from sentence are used to calculate attention scores for targets by a pooling operation, and vice versa.
- **AOA-LSTM** [25]: Similar to IAN, AOA-LSTM utilizes two bidirectional LSTMs to model sentences and targets. Then bidirectional interaction is employed by modeling word-pairs between sentences and targets to attend the most important part in both sentence and target.
- **BILSTM-ATT**: It employs vanilla global attention mechanism on the output of Bi-LSTM.
- **HEAT-BiGRU** [14]: It firstly applies attention on aspect terms to get aspect representation. Then the aspect representation together with the hidden state of sentence is used to obtain the sentiment attention score.
- **EFFECTIVE-ATT** [31]: EFFECTIVE-ATT uses an autoencoder structure to learn good target representation for a give target. Afterwards, a syntax-based attention model is used for the final prediction.
- **MemNet** [15]: It firstly calculates location attention weights for each word based on their distance to the target. Then multi-hop attention is applied on weighted word embedding. Finally, the output of the last attention is fed to softmax function for prediction.
- **RAM** [16]: It extends MemNet by applying multi-hop attention on the output of Bi-LSTM rather than word embeddings. Moreover, a recurrent function is applied between multiple attentions to model the inner dependencies.

Among all the baseline models, SVM belongs to traditional methods; LSTM and TD-LSTM can be classified into neural network methods; AT-LSTM, HEATB-BIGRU, IAN, AOA-LSTM, BILSTM-ATT, EFFECTIVE-ATT, MemNet, BILSTM-ATT and RAM are part of attention-based methods.

TABLE 2. Experimental results in accuracy.

Model	laptop	restaurant
SVM	70.49	80.16
LSTM	66.50	74.30
TD-LSTM	68.13	75.63
AT-LSTM	68.90	77.20
IAN	72.10	78.60
AOA-LSTM	72.60	79.70
BILSTM-ATT	72.41	78.16
HEAT-BiGRU	73.17	78.68
EFFECTIVE-ATT	73.86	80.54
MemNet	70.33	78.16
RAM	74.49	80.23
SHAN	**74.64**	**81.02**

Table 2 shows the performance of SHAN model compared with other baseline models. Note that Spacy dependency parsing tool is used in EFFECTIVE-ATT, to avoid the effect of different parsing tools, we reproduce the model of EFFECTIVE-ATT using the same parsing tool as the proposed method. Moreover, in our implementation of EFFECTIVE-ATT, we replace LSTM with Bi-LSTM, because the proposed method employs Bi-LSTM for memory modeling. In HEATB-BiGRU, the word embeddings are trained on large domain-specific corpus. For fair comparison, we reproduce HEATB-BiGRU by replacing the domain-specific word embedding with Glove which is commonly used in the aspect-level sentiment analysis.

We can observe that SVM performs worse than SHAN model, indicating that powerful features are difficult to design for such fine-grained aspect-level sentiment analysis task.

LSTM is the worst method among all the compared ones, because it obtains final representation by treating every word equally. As a result, it can neither distinguish different aspects in the same sentence nor highlight the important sentiment words in sentence. TD-LSTM is better than LSTM but not as good as attention-based methods. The advantage of TD-LSTM over LSTM lies in that TD-LSTM adding target-specific information by extending LSTM to process the left and right contexts with targets. However, the opinion features would be lost while being propagated in the situation where sentiment words are far from target words. For attention-based models, they can generate more reasonable representations for the given target by assigning different attention weights to different words by analyzing the semantic importance of each word rather than depending on the distance to target word.

In attention-based models, BILSTM-ATT exceeds AT-LSTM model, because BILSTM-ATT encodes contextual information from both forward and backward directions while AT-LSTM only contains contextual information from forward direction. Our proposed method outperforms BILSTM-ATT model 2.23% and 2.86% on laptop and

restaurant dataset respectively, which proves the effectiveness of introducing syntactic information.

IAN, AOA-LSTM, HEATB-BIGRU and EFFECTIVE-ATT are better than AT-LSTM, because these models propose different methods to model target representation. IAN performs better than AT-LSTM for reason that IAN strengthens target representation by applying the attention mechanism on target word(s) instead of simply averaging the target words embedding. AOA-LSTM extends IAN with fine-grained attention method by modeling interaction among word-pairs between sentences and targets. As a result, AOA-LSTM obtains better performance than IAN. Similarly, HEATB-BIGRU achieves better result that BILSTM-ATT, because HEATB-BIGRU model firstly employs aspect attention, then uses the aspect attention result for the sentiment attention. This hierarchical structure helps to encode more aspect information into final sentence representation. EFFECTIVE-ATT achieves better result than IAN, AOA-LSTM and HEATB-BIGRU. Because EFFECTIVE-ATT not only models the target representation by an autoencoder structure, but also employs syntax-based local attention to exploit syntactic information.

The proposed method outperforms EFFECTIVE-ATT because the proposed method can make use of both local and global attention results while EFFECTIVE-ATT only takes local attention result into consideration. Therefore, the proposed method can make up for the useful information left out by local attention and alleviate the effect of parsing error in irregular texts. In addition, the distance weight which is determined by the word distance to the target is somehow arbitrary, and the proposed method employs an information gate to reweight the word importance in a more flexible way.

Compared with MemNet and RAM, which use multi-layer attention architecture, our model achieves a better performance. MemNet performs attention on the sequence of word vectors, which cannot synthesize phrase-like features in the original sentence. RAM improves MemNet by modeling contextual information with bidirectional LSTM and combining features from different attentions non-linearly with a recurrent function. Therefore, it stably exceeds the MemNet. Our model achieves better result than RAM since it can make fully of syntax-based local information together with global information rather than assigning attention weight based on word distance to the target, allowing dynamically adjusting of attention weights. Moreover, compared with RAM, SHAN is simpler. Because it doesn't need to perform multi-hop attention and an additional recurrent function to combine these attention results.

C. ANALYSIS OF SHAN MODEL

In order to verify the effectiveness of SHAN model, we design a series of variations of the proposed model. Firstly, we ignore modeling local information and design GA model that only employs global attention. Similarly, we ignore modeling global information and design LA model which applies attention on a subset of words selected based on their syntax-based distance to target. Next, we implement HA-CONCAT model which concatenates local and global attention results for final prediction without gating mechanism. To evaluate the effectiveness of syntactical information, PHAN model is designed by replacing syntax-directed local attention with position-directed local attention which selects attention words within a fixed-window centered around target words. We try 2,3,4,5 for window size on each dataset and report the best result. Finally, we implement SHAN-SCALAR model in which information gate is a scalar instead of a vector.

TABLE 3. Results of SHAN and its variance.

Model	laptop	restaurant
LA	69.87	79.79
GA	72.41	78.16
HA-CONCAT	72.92	80.82
PHAN	73.82	80.73
SHAN-SCALAR	73.82	80.45
SHAN	**74.64**	**81.02**

It can be observed from Table 3 that the SHAN model outperforms both GA and LA models. For GA model, syntax-based local information which may contain pure opinion modifier is ignored. Consequently, it has high possibility to capture irrelevant sentiment words, which will affect the final prediction result to some extent. For LA model, it understands sentence within a narrow view. Though it is capable of extracting more accurate information with respective to the target, it also suffers from leaving out important information outside the local attention boundary due to the diversity of natural language.

As for HA-CONCAT model, it performs better than both GA and LA, but worse than SHAN. Compared with GA and LA, it can benefit from information from both global and local view. However, HA-CONCAT treats local and global information equally, which can't flexibly model various sorts of expression of people's opinion. SHAN model calculates a gate to dynamically combine local and global information, which is more effective and interpretable than simple concatenation.

In contrast to PHAN model which selects local attention words based on position-based distance, SHAN obtains better performance, showing the effectiveness of modeling semantic relatedness by incorporating syntax-based information.

SHAN outperforms SHAN-SCALAR model, which confirms our assumption that similar to word embeddings, each dimension of the attention results also represents a different perspective of the latent semantical space. Therefore, it is beneficial to assign a score for each dimension of the information gate.

D. EFFECT OF DEPENDENCY STEPS

In this section, we evaluate the influence of word distance of syntax-directed local attention on final prediction result.

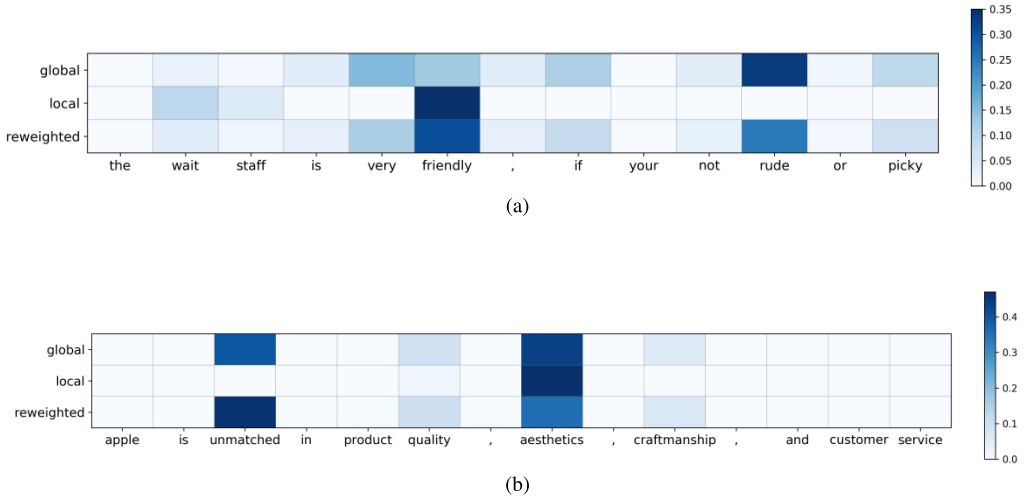


FIGURE 3. Attention Visualizations. "global", "local" and "reweighted" stand for global, local and syntax-directed hybrid attention weight, respectively. (a) Aspect: wait staff, answer: positive. Global prediction: negative, local prediction: positive, proposed model prediction: positive. Information gate > 0. (b) Aspect: aesthetics, answer: positive. Global prediction: positive, local prediction: neural, proposed model prediction: positive. Information gate < 0.

TABLE 4. The impact of dependency distance.

distance	laptop	restaurant
1-step	**74.64**	**81.02**
2-step	74.08	80.48
3-step	74.33	80.34
4-step	73.92	80.16
5-step	73.61	80.05

We choose local attention words from 1 to 5 steps word distance for syntax-directed local attention in SHAN model. Note that when we select words within k-step distance, if there are "NEG" relation in $(k+1)$-step, we also add the words corresponding to "NEG" relation in $(k+1)$-step distance into local attention to keep the semantic completeness. Because negative relation is important in sentiment analysis. Table 4 shows the result of SHAN model on different syntax-based distances.

From Table 4 we notice that local attention with dependency distance 1-step achieves best result among all the dependency distances. For one thing, this result is in line with our common sense that in reviews that people tend to deliver their opinion using phrases like "T are my favorite", "great T" and "enjoy T" where "T" stands for sentiment target which can be captured using one-step dependency relation. For another, it is worth noticing that one step syntax distance corresponds to multiple syntax-related words. Thus, adding one step syntax distance may introduce more irrelevant words in, which tends to introduce more noise in local attention result. This may add more difficulty for information gate to decide how to combine local and global results. As a result, it can be observed that as the dependency distance becomes larger, the performance of the model tends to become worse.

E. CASE STUDY

In this section, we give some examples and visualize their attention weight of each attention mechanism to explain how the information gate works. Note that the examples are picked from the experiment result with relation to one-step syntax-directed local attention. Generally speaking, the output of information gate indicates the importance of local attention result. If the output of information gate is positive, the local attention result contains useful information which needs to be paid more attention. Otherwise, the local attention result is useless, and more focus should be put on other words in global attention.

Different from conventional attention mechanism which outputs a scalar-valued attention weight for each word, the attention score obtained by syntax-directed hybrid-attention is a vector. Each dimension of the vector represents the relationship of target word(s) with each dimension of corresponding context word. As a result, it is challenging to visualize the attention score of syntax-directed hybrid attention mechanism. To tackle with the problem, instead of visualizing each dimension of attention score, we choose to visualize the average over each dimension of attention score for each word. We deal with the information gate in the same way. The visualization results are shown in Fig. 3. The first row of each subfigure shows the attention score of global attention, the second and third row represent local and syntax-directed hybrid attention score, respectively. The color depth indicates the importance of corresponding words, the deeper means more important.

Fig. 3(a) presents a case that there are noisy sentiment words in the sentence. In this case, the opinion modifier is "friendly". However, we can observe from Fig. 3(a) that global attention highlights "rude" by mistake. As a result, the global attention model incorrectly predicts sentiment

TABLE 5. Examples for each error category.

error category	examples
words attention error	1. There restaurant is very casual, but perfect for **[lunch]***neu*.
words understanding error	2. The portions of the food that came out were **[mediocre]***neu*. 3. **[Usb3 peripherals]***pos* are noticeably less expensive than the thunderbolt ones 4. I then upgraded to **[mac os x 10.8 mountain lion]***neu*.
sentence understanding error	5. Could have better for 1/3 the **[price]***neg* in Chinatown. 6. **[Waiting]***neg* three hours before getting our entrees was treat as well.

polarity towards ''waiter'' as negative. For local attention model, with the help of syntactic information, the opinion modifier is accurately emphasized without the distraction of uncorrelated words. For the proposed model, the information gate outputs a positive value after synthesizing the local and global attention result. Then, it increases the attention weight of ''friendly'' and cuts down the attention weight of ''rude''. Consequently, the proposed model is able to correctly figure out the final sentiment polarity towards ''waiter'' as positive.

Fig. 3(b) gives an example that local attention doesn't contain useful information for the final sentence representation. In this example, opinion modifier ''unmatched'' lies outside the local attention boundary. As a result, local attention can't get any useful sentiment information by attending to words ''quality, aesthetics,'', and it wrongly predicts the sentiment polarity as neural. For global attention, it successfully captures the sentiment words ''unmatched''. In this situation, information gate outputs a negative value, subtracting the share of words used in local attention, at the same time boosting the attention weight of sentiment words ''unmatched'' larger, which leads to a more sentiment-oriented representation.

F. ERROR ANALYSIS

In order to figure out the limitations of the proposed method, we carefully analyzed the wrongly predicted samples in the test sets of the two datasets. Based on the analysis, we categorize these errors into three levels: words attention error, words understanding error and sentence understanding error. Representative examples for each category are listed in Table 5.

In the first level, the prediction error is caused by attending noisy sentiment words. For example, in table 5 sentence 1, the proposed method assigns large attention weight to word ''perfect'' even though there is no opinion modifier for target ''lunch''. In spite of the introduction of syntactic information, it is difficult for the proposed model to understand the difference between ''perfect for lunch'' and ''perfect lunch''. We think the dependency relations between words may help improve this problem.

The words understanding error stands for the case where the model succeeds in attending related sentiment words, but wrongly predicts the sentiment polarity. It is the most common error occurred in test set. As shown in table 5, the error may occur in the following three cases. **Case 1:** the opinion modifier is infrequent sentiment words (sentence 2, ''mediocre''), as there may be no such words in the training set. **Case 2:** the opinion modifier is negation phrase (sentence 3, ''less expensive'') because meaning the phrase is not correctly represented by LSTM. **Case 3:** the sentence describes a fact without any opinion words (sentence 4) as more positive samples in the training may bias the neutral sentence to positive.

The sentence understanding error usually occurs when the opinions are expressed in an implicit way. For example, in sentence 5 the opinion is conducted in a subjunctive style. Sentence 6 could only be understood with common sense. In these cases, the sentiment polarity cannot be figured out by detection of explicit opinion words but a deep understanding of the whole sentence. Therefore, how to get better understanding of implicit opinion remains a challenge in deep learning methods.

V. CONCLUSION

In this paper, we propose an effective syntax-directed hybrid attention network for aspect-level sentiment analysis task. SHAN can adaptively leverage the information from global and local attention results by gating mechanism. When selecting words for local attention, instead of position-based distance, syntax-directed distance is employed for better modeling semantic relatedness. Experiments on SemEval2014 verify that SHAN stably exceeds single-layer local and global attention-based models, and even better than complex multilayer global attention-based model, indicating that given a specific target, our proposed model is able to generate powerful context representation by dynamically combining local and global attention results.

ACKNOWLEDGEMENTS

The authors would like to thank Xiao Liang and Hongzhi Zhang for the helpful discussions and suggestions to help improve this paper.

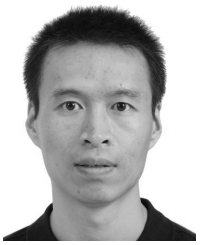
REFERENCES

[1] B. Pang and L. Lee, ''Opinion mining and sentiment analysis.'' *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.
[2] B Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool, 2012
[3] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, ''SemEval 2014 task 4: Aspect based sentiment analysis,'' in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, Dublin, Ireland, Aug. 2014, pp. 27–35.

[4] J. Wagner *et al.*, "DCU: Aspect-based polarity classification for semeval task 4," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, Dublin, Ireland, Aug. 2014, pp. 223–229.
[5] S. Kiritchenko, X. Zhu, C. Cherry, and S. M. Mohammad, "NRC-Canada-2014: Detecting aspects and sentiment in customer reviews," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, Dublin, Ireland, Aug. 2014, pp. 437–442.
[6] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 49–54.
[7] D. T. Vo and Y. Zhang, "Target-dependent Twitter sentiment classification with rich automatic features," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 1347–1353.
[8] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *26th Int. Conf. Comput. Linguistics, Tech. Papers*, 2016, pp. 3298–3307.
[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[10] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389.
[11] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in *Advances in Neural Information Processing Systems*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2015, pp. 2440–2448.
[12] P. Zhou *et al.*, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, vol. 2, Aug. 2016, pp. 207–212.
[13] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.
[14] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 4068–4074.
[15] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 214–224.
[16] P. Chen, Z. Sun, L. Bing, and W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 452–461.
[17] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[18] X. Li, L. Bing, W. Lam, and B. Shi, "Transformation networks for target-oriented sentiment classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 946–956.
[19] J. Duan, X. Ding, and T. Liu, "Learning sentence representations over tree structures for target-dependent classification," in *Proc. Conf. North Amer. Ch. Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, 2018, pp. 551–560.
[20] Z. Yang, R. Salakhutdinov, and W. Cohen, "Multi-task cross-lingual sequence tagging from scratch," *CoRR*, vol. abs/1603.06270, Aug. 2016.
[21] Z. Yang, B. Dhingra, Y. Yuan, J. Hu, W. W. Cohen, and R. Salakhutdinov, "Words or characters? Fine-grained gating for reading comprehension," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[23] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Statist. Transl. (SSST-8)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 103–111. [Online]. Available: http://aclweb.org/anthology/W14-4012, doi: 10.3115/v1/W14-4012.
[24] M. Zhang, Y. Zhang, and D.-T. Vo, "Gated neural networks for targeted sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 3087–3093.
[25] B. Huang, Y. Ou, and K. M. Carley, "Aspect level sentiment classification with attention-over-attention neural networks," *CoRR*, vol. abs/1804.06536, 2018.
[26] J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, and H. Wang, "Aspect-level sentiment classification with heat (hierarchical attention) network," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 97–106.
[27] X.-F. Wang, L. Wang, A. Hawbani, and F.-Y. Miao, "Aspect level sentiment classification with memory network using word sentiment vectors and a new attention mechanism AM-PPOSC," in *Proc. 20th IEEE Int. Conf. High Perform. Comput. Commun. (HPCC)*, Jun. 2018.
[28] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[29] K. Chen, R. Wang, M. Utiyama, E. Sumita, and T. Zhao, "Syntax-directed attention for neural machine translation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 4779–4792.
[30] O.-E. Ganea and T. Hofmann, "Deep joint entity disambiguation with local neural attention," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2619–2629.
[31] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Effective attention modeling for aspect-level sentiment classification," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1121–1131.
[32] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, vol. 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2013, pp. 3111–3119.
[34] H. Choi, K. Cho, and Y. Bengio, "Context-dependent word representation for neural machine translation," *Comput. Speech Lang.*, vol. 45, pp. 149–160, Sep. 2017.

XINYI WANG received the B.Sc. degree from Xidian University, Xi'an, China, in 2016. She is currently pursuing the Ph.D. degree with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China.

Her research interests include deep learning, natural language processing, and sentiment analysis, especially on fine-grained sentiment analysis.



GUANGLUAN XU received the B.Sc. degree from the Beijing Information Science and Technology University, Beijing, China, in 2000, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2005.

He is currently a Professor with the Institute of Electronics, Chinese Academy of Sciences. His research interests include remote-sensing image understanding, and geospatial data mining and visualization.

JINGYUAN ZHANG received the B.Sc. degree from the Wuhan University of Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China.

His research interests include deep learning, natural language processing, and information extraction.

XIAN SUN received the B.Sc. degree from Beihang University, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2006 and 2009, respectively. He is currently an Associate Professor with the Institute of Electronics, Chinese Academy of Sciences. His research interests include computer vision and remote-sensing image understanding.

LEI WANG received the B.Sc. and M.Sc. degrees from the Shandong University of Technology, Jinan, China, in 1995 and 1998, respectively. He received the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2001, where he is currently an Associate Professor. His research interests include data mining and geospatial information application technology.

TINGLEI HUANG received the B.Sc. degree from the Wuhan University of Technology, Wuhan, China, in 1993, and the M.Sc. and Ph.D. degrees from the University of Shanghai for Science and Technology, Shanghai, China, in 1997 and 2000, respectively. He is currently an Associate Professor with the Institute of Electronics, Chinese Academy of Sciences, Beijing, China. His research interests include data mining, knowledge graph, and geospatial information application technology.