future internet

Article

A Multi-Attention Network for Aspect-Level Sentiment Analysis

Qiuyue Zhang and Ran Lu *

School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
* Correspondence: luran@sdnu.edu.cn; Tel.: +86-138-0640-6182

Received: 17 June 2019; Accepted: 11 July 2019; Published: 16 July 2019

Abstract: Aspect-level sentiment analysis (ASA) aims at determining the sentiment polarity of specific aspect term with a given sentence. Recent advances in attention mechanisms suggest that attention models are useful in ASA tasks and can help identify focus words. Or combining attention mechanisms with neural networks are also common methods. However, according to the latest research, they often fail to extract text representations efficiently and to achieve interaction between aspect terms and contexts. In order to solve the complete task of ASA, this paper proposes a Multi-Attention Network (MAN) model which adopts several attention networks. This model not only preprocesses data by Bidirectional Encoder Representations from Transformers (BERT), but a number of measures have been taken. First, the MAN model utilizes the partial Transformer after transformation to obtain hidden sequence information. Second, because words in different location have different effects on aspect terms, we introduce location encoding to analyze the impact on distance from ASA tasks, then we obtain the influence of different words with aspect terms through the bidirectional attention network. From the experimental results of three datasets, we could find that the proposed model could achieve consistently superior results.

Keywords: aspect-level; sentiment analysis; multi-attention

1. Introduction

Text sentiment analysis, is the process of analyzing, processing, summarizing, and inferring subjective text with sentiment polarity, which is a vital task in natural language processing (NLP) and is also known as opinion mining [1]. With the rapid development of Internet technology and continuous popularity of social networks, people express their opinions on different platforms, such as online social media, product review websites and other media. Thus, accumulating a large amount of text information data. Relevant comments and opinions have become an important reference for users, and even in academic research, they are also interested in the processing of massive data. In recent years, aspect-level sentiment analysis (ASA) is proposed, which could understand reviews better than traditional sentiment analysis and aim at identifying the sentiment polarity of an input sentence in a certain aspect term [2]. For example, given the sentence "Great food but service was dreadful", the polarity of the sentence towards the aspect "food" is positive while the polarity of "service" is negative.

Previously, most of the literatures make use of neural networks to capture the sequence information and automatically learn useful representations of context and aspect terms [3], such as Recurrent Neural Network (RNN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). Xue et al. [4] proposed a method of target extraction based on the gated convolution network, which was easy to confuse the sentiment polarity of aspect terms and could not achieve better fusion with the original sentence. The transduction learning approach proposed by Marcacini et al. [5] is able to extract aspect terms well, but it did not fully analyze the sentiment of a particular aspect term.

However, the existing neural network models are not effective enough in dealing with ASA tasks alone, and they are still in the immature stage of processing fine-grained sentiment analysis tasks.

It can be seen that most of the current models of ASA tasks adopt the method of combining neural networks and attention mechanisms. Attention is widely used in machine translation [6] and reading comprehension [7]. Therefore, a number of researchers have also used attention mechanisms to solve ASA tasks, which could focus on the effect of different contextual words on the target [8]. Tang et al. [9] proposed the Memory Network (MemNet) model, which consists of computational layers of shared parameters. Each layer contains an attention layer and a linear layer. Wang et al. [10] proposed Attention-based LSTM with Aspect Embedding (ATAE-LSTM), which was able to attend to different parts of a sentence when different aspects are concerned. From the experimental results of MemNet and ATAE-LSTM, they all show that the attention mechanism is effective. Considering that neural networks are difficult to parallellize and requires a great deal of computation, this paper chooses to use multiple attention mechanisms to complete the task.

Currently, word embedding and aspect term embedding usually use the original data preprocessing methods, such as Glove [11] and Word2vec. For example, Ma et al. [12] proposed that context embedding and aspect term embedding be used for inputting for ASA, and Glove be used for preprocessing input data. For the ASA task and datasets proposed in this paper, we find the data preprocessed by Bidirectional Encoder Representations from Transformers (BERT) [13] will greatly improve the effect of downstream tasks. Therefore, BERT is adopted to preprocess data onto this paper. Through the experimental comparison of this paper, the method improves the experimental results effectively.

Inspired by the above problems, we propose a Multi-Attention Network (MAN) model for ASA. This paper mainly completes the ASA task, which is a fine-grained sentiment analysis task, and the goal is to infer the corresponding sentiment polarity for given sentences with aspect terms, such as positive, negative, and neutral.

The main contributions of our work are as follows:

1. The method proposed in this paper mainly adopts an attention network. We make use of Multihead Attention (MHA) and Point-wise Feed-Forward Networks (PFFN) to interactively obtain the hidden representation of the context and aspect term embeddings, which are akin to a partial transformer [14]. In addition, for a normal sentence that the distance between context words and aspect terms will have different effects, so this paper includes location encoding into this model.
2. Considering that words in different positions have different influences on aspect terms, the bidirectional attention network is adopted to analyze the influences and correlations of different words on them.
3. Experiments were carried out on three different public authoritative datasets, including multiple baseline models and ablation studies. The experimental results show that the approach proposed in this paper outperforms state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 presents the main methods about ASA, as well as the related work. Section 3 describes in detail our MAN model, especially the description of MHA and PFFN for sequence information extraction. The experimental evaluation results comparing our MAN model and other state-of-the-art methods for ASA are presented and discussed in Section 4, with additional ablation experiments. Finally, Section 5 discusses the conclusion of this work and the direction of future work.

2. Related Work

In this section, we will give a briefly introduction for the latest and authoritative studies on ASA. Traditional research can be divided into three directions: traditional machine learning methods, neural network methods, and attention network methods.

2.1. Traditional Machine Learning Methods

Traditional machine learning methods mainly focus on text representation and feature extraction like sentiment lexicons features and bag-of-words features. Such as support vector machine (SVM), which is employed with well-designed handcrafted features [15]. Jiang et al. [16] proposed the statistic methods, which largely depended on the effectiveness of feature engineering works. Kaji et al. [17] proposed to make use of structural clues that could extract polar sentences from HTML documents, and built lexicon after extracting polar sentences. However, the traditional machine learning methods mentioned above are labor-intensive and tend to lead to high-dimensional and sparse text representation [18].

2.2. Neural Networks Methods

Neural network methods are paid more and more attention as they could learn feature representations in data without paying attention to feature engineering. Tang et al. [19] proposed Target-Dependent LSTM (TD-LSTM) and Target-Connection LSTM (TC-LSTM), which aspect information are taken into account to improve classification accuracy. Li et al. [20] proposed a unified model for opinion target extraction and target sentiment prediction, which uses two stacked recurrent neural networks and a gate mechanism, one recurrent neural network predicts the unified tags and the other performs an auxiliary target boundary prediction. Based on the above neural network methods, we conclude that RNNs is still the primary application object, but it is difficult to parallelize and requires large amounts of data.

2.3. Attention Network Methods

Wang et al. [10] proposed to combine attention mechanism with LSTM to model sentences semantically, which used attention mechanism to capture the importance of different contextual information on a given aspect and solve the problem of ASA. Chen et al. [8] proposed RAM which adopts multi-attention mechanism by bidirectional LSTM and constructs a multi-layer attention through gated recurrent unit (GRU). It can be seen that the deep learning model integrating target and attention mechanism achieves a large number of results. However, these models do not take into account the extent to which words in different locations affect the aspect terms and ignore the bidirectional interaction between the aspect terms and context.

As a consequence, our MAN model chooses a new method, which makes use of multiple attention networks. Furthermore, we leverage the attention mechanism of Aspect2Context (A2C) and Context2Aspect (C2A) to describe word-level interactions and assess how each aspect term/context word affects each context/aspect term word.

3. Multi-Attention Network

In this section, we will introduce a new model proposed in this paper for solving ASA tasks, called Multi-Attention Network (MAN) model, which not only uses new technologies but is more effective than traditional methods. To be specific, the model is shown in Figure 1, which consists of four components: (1) Datasets are preprocessed using BERT as word embedding and aspect term embedding. (2) Computing the hidden states of the input embeddings by MHA and PFFN. (3) Taking advantage of bidirectional attention network to analyze the interaction between context and aspect terms. (4) The output of the sub-structure is average pooled separately and concatenate as input to softmax function. These four components correspond to embedding layer, attentional encoder layer, bidirectional attention layer, and output layer, respectively.

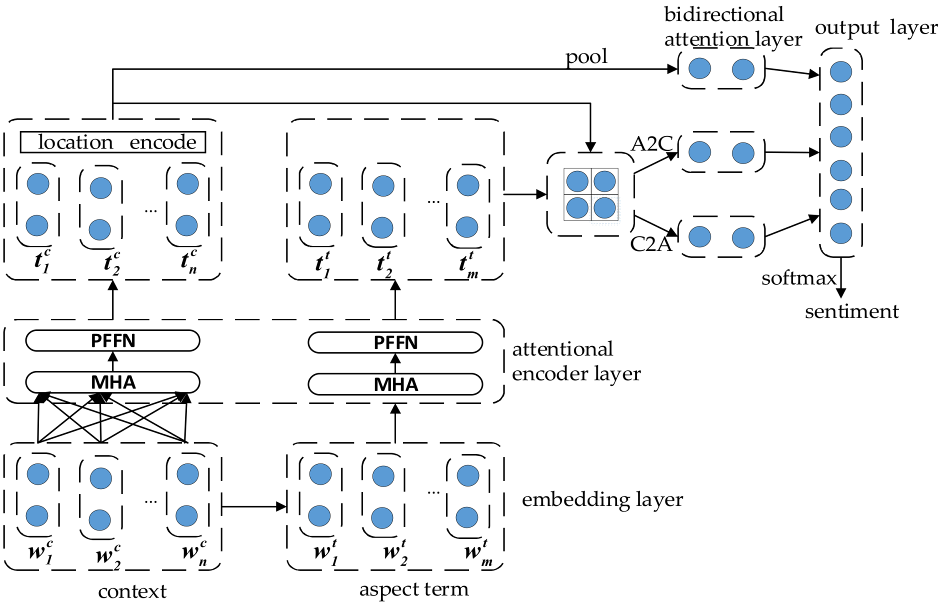


Figure 1. The architecture of MAN model.

Given the context sequence w_c and the aspect term sequence w_t, where w_t is a sub-sequence of w_c. The purpose of our proposed model is to predict the sentiment polarity of the sentence w_c over the aspect term w_t.

3.1. Embedding Layer

The input of our model presented in this paper is mainly composed of two parts: word embedding and aspect term embedding. All data are pre-trained using BERT to generate sequence of word vectors. For now, BERT is widely used in a wide range of tasks, which could be a good feature representation for word learning by running self-supervised learning method on the basis of massive corpus. In this paper, BERT is adopted to preprocess all data and the process is shown in Figure 2.

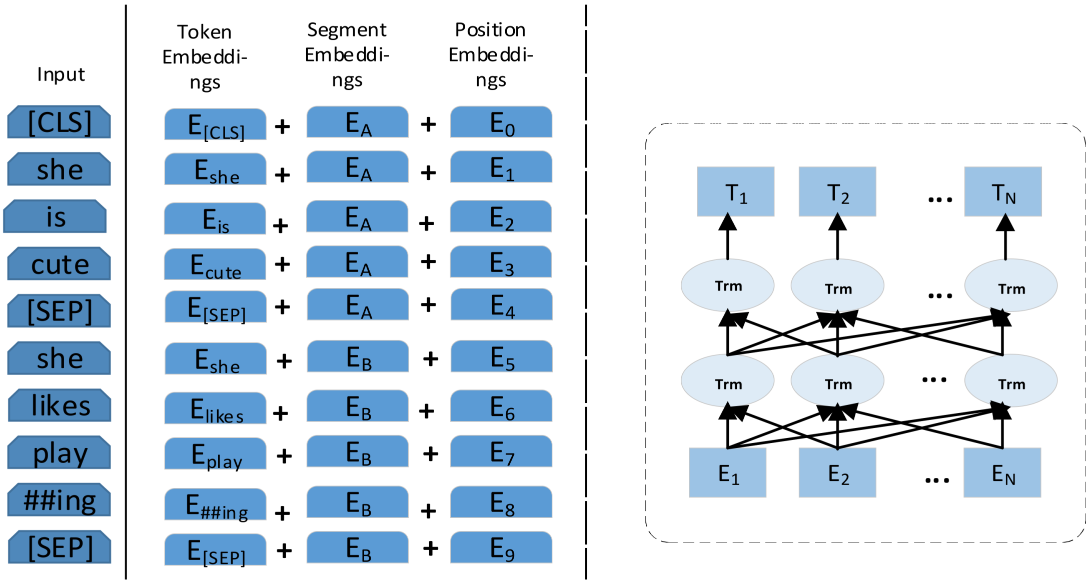

Input
Token Embeddi-ngs
Segment Embeddi-ngs
Position Embeddi-ngs
[CLS]
she
is
cute
[SEP]
she
likes
play
##ing
[SEP]
$E_{[CLS]}$ + E_A + E_0
E_{she} + E_A + E_1
E_{is} + E_A + E_2
E_{cute} + E_A + E_3
$E_{[SEP]}$ + E_A + E_4
E_{she} + E_B + E_5
E_{likes} + E_B + E_6
E_{play} + E_B + E_7
$E_{\#\#ing}$ + E_B + E_8
$E_{[SEP]}$ + E_B + E_9
T_1
T_2
...
T_N
Trm
E_1
E_2
E_N

Figure 2. BERT input representation and the architecture of BERT.

By Equation (1), word embedding and aspect term embedding are calculated, respectively, where x generally refers to the input data, which is processed by BERT to generate embeddings H:

$$H = BERT(x) \tag{1}$$

Pre-trained by BERT, word embedding is interactively input into the corresponding MHA and combined with the aspect term embedding to input into another MHA. They all effectively extract the semantic feature of the text and enhance the performance of downstream tasks.

3.2. Attentional Encoder Layer

Attentional encoder layer can parallelize and compute the hidden state of the input embeddings, and its function is similar to LSTM. This layer consists two submodules: the MHA and the PFFN. Next, this paper will cover these two parts in detail.

3.2.1. Multi-Head Attention

MHA is the attention that performs multiple attention functions simultaneously. For context embedding, input its interaction into MHA to achieve the information interaction between contexts. Meanwhile, the context embedding processed by BERT is transmitted to the aspect term embedding, and the two are input into MHA together.

An attention function can be described as mapping a query sequence $q = \{q_1, q_2, \ldots, q_n\}$ and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. Currently, in NLP research, key and value are often the same, that is, key = value. The key sequence is $k = \{k_1, k_2, \ldots, k_n\}$. To be honest, both q and k are embedding results obtained by different linear transformations. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a function of the query with the corresponding key:

$$Attention(q, k, v) = softmax(\frac{f(q, k)}{\sqrt{d_k}})v \tag{2}$$

The $\sqrt{d_k}$ plays a regulatory role, so that the above inner product is not too large to affect the softmax function [14], where f denotes the similarity function between q_i and k_j is calculated by concatenating, that is, the semantic relevance between the two:

$$f(q_i, k_j) = w_a[q_i; k_j] \tag{3}$$

where $W_a \in R^{2d}$ denotes learnable weights.

The MHA allows the model to learn relevant information in different representation subspaces. The different representations obtained are concatenated together and the values obtained by another linear transformation are taken as the results of MHA:

$$head_i = Attention(q, k, v) \tag{4}$$

$$Multihead = Concate(head_1, head_2, \ldots, head_h)W^O \tag{5}$$

where $W^O \in R^{hd \times d}$ and $h \in [1, n_{head}]$.

Multihead self-attention is a typical $q = k$ mechanism under special situation. The context embedding w^c are input into MHA, the context representations can be obtained by Equation (6). The complete context is expressed as $c = \{c_1, c_2, \ldots, c_n\}$.

$$c = Multihead(w^c, w^c, w^c) \tag{6}$$

In normal attention mechanism, q is different from k. When the context embedding w^c and the aspect term embedding w^t are input into MHA at the same time, we can get the context and target representation:

$$t = Multihead(w^c, w^t, w^t) \tag{7}$$

After MHA calculation, each aspect term embedding is corresponding to context embedding. Then get the context and aspect term are expressed as $t = \{t_1, t_2, \ldots, t_n\}$.

3.2.2. Point-Wise Feed-Forward Networks

A PFFN can transform contextual information gathered by the MHA, which consists of two linear transformations with a ReLU activation in between:

$$PFFN(h) = max(0, hW_1 + b_1)W_2 + b_2 \tag{8}$$

where h denotes input sequence. The process of PFFN is similar to the convolution operation, that two convolutions of kernel size are 1 and apply the same transformation to every token belonging to the input. From Equation (8), $W_1 \in R^{d \times d}$ and $W_2 \in R^{d \times d}$ are the learnable weights, $b_1 \in R^d$ and $b_2 \in R^d$ are biases. Given c and t, PFFN are applied to get the output hidden states of the attentional encoder layer $h^c = \left\{h_1^c, h_2^c, \ldots, h_n^c\right\}$ and $h^t = \left\{h_1^t, h_2^t, \ldots, h_m^t\right\}$ by:

$$h^c = PFFN(c) \tag{9}$$

$$h^t = PFFN(t) \tag{10}$$

Processed by attentional encoder layer, we can obtain the context and aspect term hidden representations, which is useful for further analysis. In addition, considering that the context words close to the aspect term have a great influence on the aspect, we adopt the location encoding. For example, in the sentence "Great food but service was dreadful!" the word "dreadful" should describe the aspect term "service" not "food". The weight for a context word w_i, which the distance from the aspect term is l [21]. It is defined as:

$$w = 1 - \frac{1}{n - m + 1} \tag{11}$$

When the context word is the aspect term, the weight value is set to 0 to focus on the context words in the sentence. Where $m \in [1, n]$, the weight of each position is multiplied by the corresponding PFFN processed context output as the final context word.

3.3. Bidirectional Attention Layer

Different attention should be given to different words, so this paper makes use of the similarity matrix method, which is widely used in sentiment analysis tasks. Similarity is often used in aspect-level sentiment analysis task, so this paper adopts matrix method similar to Fan et al. [21]. Concatenating the processed context with the aspect term hidden representation to generate a similarity matrix $U \in R^{m \times n}$, where U_{ij} represents the similarity between i-th context word and j-th aspect term word:

$$U_{ij} = W_u([w_i^c; w_j^t]) \tag{12}$$

where $W_u \in R^{4d}$ is the weight matrix, [;] is the vector concatenation. Then we use U to calculate the attention vectors in the directions of A2C and C2F. For A2C, the higher the similarity is, the more important the sentiment influence is. Therefore, we get the maximum value of U through the Equation (13). The weight of context word plays a role in determining the ultimate aspect sentiment. With s_i^a, we could calculate the attention weight α_i^a on the context words. The vector n^a processed by A2C is part of the results of the output layer:

$$s_i^a = max(U_{i,}) \tag{13}$$

$$\alpha_i^a = softmax(s_i^a, s_r^a) \tag{14}$$

$$n^a = \sum\nolimits_{i=1}^{n} \alpha_i^a \cdot w_i^c \tag{15}$$

For C2A, it is mainly used to measure the correlation between aspect terms and context words, so as to find out which aspect terms have the greatest correlation with context words. Therefore, we need to calculate the attention weight of the aspect term output for any context word. The above process is completed by Equation (16). Similarly, the vector treated with C2A is also used as part of the output layer result:

$$\alpha_{ij}^c = softmax(U_{ij}, U_{ir}) \tag{16}$$

$$m_i^c = \sum\nolimits_{j=1}^{m} \alpha_{ij}^c \cdot w_j^t \tag{17}$$

$$n^c = averagepooling(m_1^c, m_2^c, \ldots, m_n^c) \tag{18}$$

where the size of aspect terms is r. Once again, the processed context representation is average pooled to obtain text features, which are input to the output layer, called n^l.

3.4. Output Layer

The final representation of each substructure is pooled averagely and connected as the final output $\tilde{O}$, then a fully connected layer is used for projecting the connected vectors into the space of the target classes C. The sequence representation X is obtained by using a linear layer.

$$\tilde{O} = [n^l; n^a; n^c] \tag{19}$$

$$X = \tilde{W}_O^T + \tilde{b}_O \tag{20}$$

$$y = softmax(X) = \frac{exp(X)}{\sum_{k=1}^{C} exp(X)} \tag{21}$$

where $y \in R^C$ is the predicted sentiment polarity distribution, $\tilde{W} \in R^C$ and $\tilde{b}_O \in R^C$ are learnable parameters.

3.5. Regularization and Model Training

For text sentiment analysis, neutral polarity is a vague sentimental state. Thus, we employ Label Smoothing Regularization (LSR) term in the loss function, which penalizes low entropy output distributions [22]. LSR implements model constraints and reduces the degree of over-fitting by adding noise to the output.

For a training sample x with the original ground-truth label distribution $G(g|x)$, we compute $G'(g|x)$ with:

$$G'(g|x) = (1-e)G(g|x) + eu(g) \tag{22}$$

where *u(g)* denotes a known distribution of label k independent of training samples, which generally obeys a simple uniform distribution, then $u(k) = \frac{1}{c}$; $e \in [0,1]$.

LSR corresponds to *KL* distance between the known label distribution *u(g)* and the predicted distribution p_θ. LSR term is defined as:

$$L_{lsr} = -D_{KL}(u(g)\|p_\theta) \tag{23}$$

The proposed module could be trained in an end-to-end manner by optimizing the cross-entropy loss as much as possible with L_{lsr} and L_2 regularization. In our work, y_i denotes the correct sentiment polarity, and $\hat{y}_i$ denotes the predicted sentiment polarity for the given sentence. In addition, λ is the L_2 regularization factor and θ represents all parameters. Based on them, the training loss is constructed as:

$$loss = -\sum_{i=1}^{s} y_i log(\hat{y}_i) + L_{lsr} + \lambda \left\|\theta\right\|^2 \quad (24)$$

4. Experiments

4.1. Datasets

We conduct experiments on public and authoritative datasets: SemEval 2014 Task4 dataset [23], which composed of restaurant reviews and laptop reviews, respectively. The third one is a collection of tweets gathered by Dong et al. [24]. For the three datasets, each review contains a list of aspect terms and corresponding polarities, which are labeled with {positive, negative, neutral}. Since the dataset may contain conflict polarity, which means that a sentence express both positive and negative in an aspect. Therefore, we remove the conflict polarity in case they affect the final result. Table 1 shows the number of training and test samples of each sentiment polarity on different datasets.

Table 1. Statistics of the datasets.

Dataset	Positive		Neutral		Negative	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Restaurant	2164	728	807	196	637	196
Laptop	994	341	870	128	464	169
Twitter	1561	173	3127	346	1560	173

4.2. Experimental Settings

In our experiments, the dimension of word embedding and aspect term embedding are initialized by BERT. The dimension of word embedding and aspect term embedding are set to 768. The weight matrices and bias are initialized by sampling from a uniform distribution $U(-0.01, 0.01)$. During training, we set label smoothing regularization parameter e to 0.1, the coefficient λ of L_2 regularization item is 10^{-5} and dropout rate is 0.5. In addition, the learning rate is set to 5^{-5}. We implement all models in the Tensorflow environment, which adopt accuracy and macro-F1 metrics to evaluate the performance of the model.

4.3. Model Comparsion

In order to evaluate the performance of proposed model, we compare five baseline models and design four ablations of MAN-BERT model.

Baseline models:

- LSTM: LSTM uses the sentence as input to get the hidden representation of each word. Then it takes the average value of all hidden states as the representation of sentence, and puts it into a softmax layer to predict the sentiment polarity [18].
- ATAE-LSTM: Standard LSTM could not detect important information in text for ASA. To solve this problem, attention mechanism is introduced and AT-LSTM model is proposed, which could obtain the key information of a given target in text. ATAE-LSTM strengthens the effect of target embedding which extended AT-LSTM by appending the aspect embedding to each word embedding to highlight the role of aspect embedding.
- IAN: IAN has designed a model for interactive computing of aspect terms and sentences, which leverages attention in context and aspect terms to generate representations of aspect terms and context, respectively. Finally, the sentiment polarity of aspect term in context is predicted by combining aspect term and context representation.

- MemNet: The deep memory network with three computational hops and the results of the attention mechanism are iterated many times for ASA.
- RAM: The framework of recurrent attention on memory, which uses deep bidirectional LSTM to build memory record all information. In addition, position weights are introduced to expect a better prediction accuracy [8].

MAN-BERT ablations:

- MAN-BERT w/o MHA ablates MHA module.
- MAN-BERT w/o LSR ablates label smoothing regularization.
- MAN-BERT w/o BERT ablates the pre-trained BERT pattern.
- MAN-BERT-Bi-LSTM replaces the MHA and PFFN modules with bidirectional LSTM.

4.4. Main Results

According to the datasets, the performance of MAN and other models are shown in Table 2, which represents the classification accuracy and macro-F1. From Table 2, the accuracy indicates the probability of being divided into three polarities of sentiment {positive, negative, neutral}. We could observe that our proposed MAN model obtains the best performance.

Table 2. Comparison with accuracy and macro-F1. "-" denotes not reported. Best results are in bold.

	Models	**Restaurant**		**Laptop**		**Twitter**	
		Accuracy	**Macro-F1**	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**
Baseline models	LSTM	74.28	-	66.45	-	-	82.55
	ATAE-LSTM	77.20	90.90	68.70	87.60	66.62	85.63
	IAN	78.60	-	72.10	-	-	-
	MemNet	78.16	65.83	70.33	64.09	68.50	66.91
	RAM	80.23	70.80	74.49	71.35	69.36	67.30
MBTN-BERT ablations	MAN-BERT w/o MHA	80.02	88.99	73.62	88.49	70.24	87.21
	MAN-BERT w/o LSR	81.54	90.69	75.90	89.92	70.11	87.68
	MAN-BERT w/o BERT	80.00	85.88	71.78	87.83	69.25	86.45
	MAN-BERT-Bi-LSTM	81.30	91.19	73.59	87.92	71.29	87.65
Ours	MAN-BERT	**81.43**	**91.22**	**76.35**	**88.13**	**71.35**	**87.92**

Comparing the baseline models with LSTM, ATAE-LSTM and IAN perform better than LSTM because they all make full use of the information about aspect term and attention mechanism. It is obvious that LSTM cannot obtain the target information in the sentence, even if given different targets, when the corresponding sentiment polarity is the same. ATAE-LSTM uses aspect embedding twice and an attention mechanism by giving different weights to different words in the aspect term. IAN strengthens the interaction between aspect term and context which uses connected attention networks. Thus, IAN improves 1.4 points and 3.4 points on restaurant and laptop datasets in accuracy, respectively.

For MemNet and RAM, they achieve better results of the three datasets, among which RAM introduces position weight, which is similar to the MAN model. They all take into account the influence of location. At present, there are few methods to solve ASA tasks by applying location encoding, so we can apply it further in the later stage. Although the performance of RAM is better than other methods, it performs less competitively than our MAN model on all datasets. RAM uses Bi-LSTM to learn aspect term and context representations. In addition, RAM employs multiple recurrent attention models to obtain weights in different context words. Compared with the model proposed in this paper, which adopts a bidirectional attention structure, and the influence of feature extraction and the analysis context on the aspect term is better than that of RAM. The experimental results also confirm this point. According to the final results of the experiments, we can find that MAN-BERT's results are far better than RAM, and the maximum difference value is more than 20 points. Therefore, the method in this paper is much better than the traditional methods.

As shown in Table 2, the performance of MAN-BERT ablations is not as good as MAN-BERT in accuracy and macro-F1, which shows that our proposed model is indispensable in structure and parameter settings. When MHA and PFFN in the MAN model is replaced by Bi-LSTM, the accuracy and macro-F1 of the two models are close because they have the same influence. At present, most sentiment analysis tasks are solved by neural networks. In this paper, from the perspective of replacing neural networks with a transformer, the hidden information is obtained in depth. However, when the MHA is ablated and the Bi-LSTM substitution is not used, the final results are lower than that of the proposed model due to the lack of feature extraction in advance. Comparing the results of MAN-BERT w/o LSR and MAN-BERT, we observe that the MAN-BERT w/o LSR drops significantly, we attribute this phenomenon to the neutral sentiment. Neutral polarity is similar to conflicting polarity, which makes the results extremely unbalanced. More significantly, when BERT used for pre-training data is ablated, the effect of Glove on the conventional pre-training word vector is not good, which indicates that preprocessing of data by BERT will have a greater impact on the experimental results in ASA. Although BERT is rarely used at present, its effect is much better than that of traditional data processing methods.

To sum up, similar to the experiment on ASA, our model achieves state-of-the-art performance, and it could effectively judge the sentiment polarity of different aspect terms in its corresponding sentence, so as to improve the classification accuracy and macro-F1.

4.5. Model Analysis

In this section, in order to further demonstrate the effectiveness of the MAN model, we analyze the training process of the model in the restaurant dataset, including the change trend charts of accuracy and Macro-F1. The loss function results of the training dataset and test dataset are shown in Figure 3. As can be seen from the training set function curve, the loss function value decreases with the increase of the training epoch. However, the loss function curve of the test dataset fluctuates less. This demonstrates that the addition of L_{lsr} and L_2 regularization in the cross-entropy loss function could reduce the value of the loss function to a lower level.

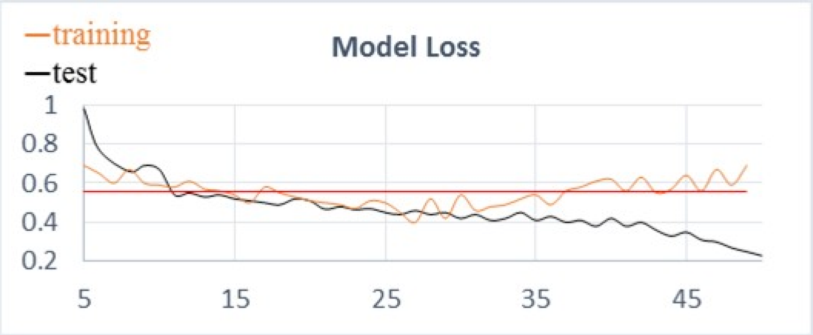
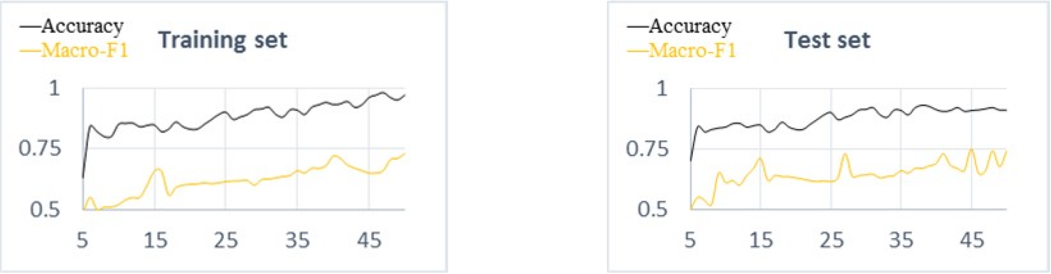


Figure 3. Model loss.

In order to verify that the method proposed in this paper could achieve better sentiment analysis results on restaurant reviews, a comparative experiment of accuracy and macro-F1 values is carried out on training set and test set. The experimental results are shown in Figure 4a,b. As is shown in Figure 4a, we could observe that with the increase of training epoch, the accuracy is also increasing. After training, the accuracy and macro-F1 value of our model are relatively stable. The performance of the test dataset is shown in Figure 4b, and its performance tends to be stable.



(**a**) (**b**)

Figure 4. (**a**) The training dataset comparison. (**b**) The test dataset comparison.

Through the above analysis, the model proposed in this paper could obtain better performance. The main reason is to use BERT to preprocess data, while performing in-depth analysis of the relationship between aspect terms and contexts. The experimental results show that this method is indeed effective in implementing ASA tasks.

4.6. Case Study

In order to give an intuitive understanding of our MAN model, we visualize the attention weights on the aspect term and sentence in Figure 5. The deeper of the color is, the greater the effect on the sentence.

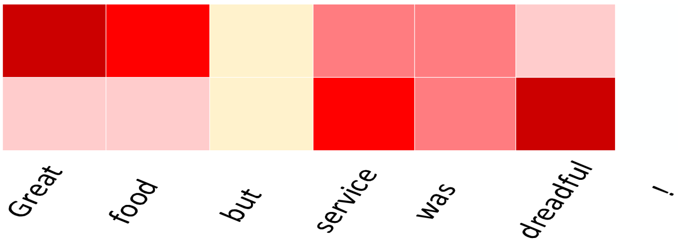


Figure 5. The visualized attention weights for the sentence and aspect term by MAN.

As shown in Figure 5, the sentence is "Great food but the service was dreadful!" for aspect terms food and service, the corresponding polarity are positive and negative, respectively. The attention mechanism could dynamically supervise the sentiment words in the text against known aspect terms, even if the text contains other aspect terms and sentiment words. We can see that the attention mechanism enables the model to focus more on the important words associated with aspect terms. For example, in terms of "food" and "service", they have higher weights than other words. Therefore, the MAN model is able to judge the sentiment polarity by modeling the interactive relationships between the aspect term and sentence.

5. Conclusions and Future Work

In this paper, we propose an MAN model based on attention network of ASA, which employs MHA and PFFN to capture the hidden information and uses bidirectional attention network mechanism to obtain the mutual influence of different words and corresponding aspect terms in the text. We also take advantage of pre-trained BERT in this task and location encoding is introduced to analyze the extent to which different positional words affect aspect terms. Experimental results on datasets demonstrate that compared with other methods, the accuracy of our proposed model has been further improved, which could better solve ASA problem.

Since ASA is a fine-grained and complex task, there are other directions that can be explored. For example, the part-of-speech is not considered and too few datasets are selected in this paper. Furthermore, the experimental results show that there is still a great deal of room for improvement.

Therefore, the next step will focus on them. We believe that there will be more effective solutions in the near future.

Author Contributions: Conceptualization: Q.Z. and R.L.; methodology: Q.Z.; software: Q.Z.; validation: R.L.; resources: Q.Z. and R.L.; writing—original draft preparation: Q.Z.; writing—review and editing: Q.Z. and R.L.; supervision: R.L.; visualization: Q.Z. and R.L.

Funding: This work was supported by the National Natural Science Foundation of China (grant no. 61373148), the Science Foundation of the Ministry of Education of China (grant no. 14YJC860042), and the Shandong Provincial Social Science Planning Project (grant no. 18CXWJ01/18BJYJ04/17CHLJ33/17CHLJ30/17CHLJ18).

Acknowledgments: Thanks to all commenters for their valuable and constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, L.; Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2016**, *30*, 167.
2. Heyz, R.; Lee, W.S.; Ng, H.T. An Unsupervised Neural Attention Model for Aspect Extraction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 388–397.
3. Zheng, L.; Wei, Y.; Zhang, Y.; Zhang, X.; Li, X.; Yang, Q. Exploiting Coarse-to-Fine Task Transfer for Aspect-level Sentiment Classification. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
4. Xue, W.; Li, T. Aspect Based Sentiment Analysis with Gated Convolutional Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2514–2523.
5. Marcacini, R.M.; Rossi, R.G.; Matsuno, I.P.; Rezende, S.O. Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. *Decis. Support Syst.* **2018**, *114*, 70–80. [CrossRef]
6. Luong, T.; Pham, H.D.; Manning, C. Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421.
7. Gui, Y.M.; Chen, Z.P.; Wei, S.; Wang, S.J.; Liu, T.; Lu, G.P. Attention-over-Attention neural networks for reading comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 30 July–4 August 2017; pp. 593–602.
8. Chen, P.; Sun, Z.Q.; Bing, L.D.; Yang, W. Recurrent attention network on memory for aspect sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 452–461.
9. Tang, D.Y.; Qin, B.; Liu, T. Aspect Level Sentiment Classification with Deep Memory Network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
10. Wang, Y.Q.; Huang, M.L.; Zhao, L.; Zhu, X.Y. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
11. Pennington, J.; Socher, R.D.; Manning, C. Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
12. Ma, D.H.; Li, S.J.; Zhang, X.D.; Wang, H.F. Interactive Attention Networks for Aspect-Level Sentiment Classification. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 4068–4074.
13. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available online: https://arxiv.org/abs/1810.04805 (accessed on 13 June 2019).
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. Available online: https://arxiv.org/pdf/1706.03762 (accessed on 13 June 2019).

15. Prez-Rosas, V.; Banea, C.; Mihalcea, R. Learning Sentiment Lexicons in Spanish. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, 21–27 May 2012; pp. 3077–3081.
16. Jiang, L.; Yu, M.; Zhou, M.; Zhao, T.J. Target-dependent Twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, USA, 19–24 June 2011; pp. 151–160.
17. Kaji, N.; Kitsuregawa, M. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, 28–30 June 2007; pp. 1075–1083.
18. Gu, S.Q.; Zhang, L.P.; Hou, Y.X.; Song, Y. A Position-aware Bidirectional Attention Network for Aspect-Level Sentiment Analysis. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 774–784.
19. Tang, D.Y.; Qin, B.; Feng, X. Effective LSTMs for Target-Dependent Sentiment Classification. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–17 December 2016; pp. 3298–3307.
20. Li, X.; Bing, L.D.; Li, P.J.; Lam, W. A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
21. Fan, F.F.; Feng, Y.S.; Zhao, D.Y. Multi-grained Attention Network for Aspect-Level Sentiment Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3433–3442.
22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
23. Pontiki, M.; Galanis, D.; Pavlopoulos, J.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S. Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23–24 August 2014; pp. 27–35.
24. Dong, L.; Wei, F.R.; Tan, C.Q.; Tang, D.Y.; Zhou, M.; Xu, K. Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, MD, USA, 23–25 June 2014; pp. 49–54.