Multi-task Learning Network based on Attention for Aspect-Based Sentiment Analysis

Yang Dong*, Jing Wang and Junwei Wang

1School of computer science and Engineering, Northeastern University, Shenyang, 110000, China

*Corresponding author’s e-mail: 1971842@stu.neu.edu.cn

Abstract. Aspect-based sentiment analysis is to analyse the sentiment polarity of the aspect in a specific context. The previous methods mostly use the working mode of pipeline model, and internally use Recurrent Neural Network to predict the sentiment polarity of the aspect. This way clearly defines the order of aspect extraction and aspect-level sentiment classification, which will lead to the problem of error transmission. And traditional models have great limitations. In this paper, we propose a multi-task learning network based on attention (MNA), which uses shared word embedding to transmit to downstream tasks, allowing extraction and classification tasks to be processed simultaneously. We design a network architecture based on attention mechanism, and use the improved multi-head attention mechanism to transfer information from extracted tasks to classified tasks. Experimental results on three benchmark datasets show the effectiveness of MNA model.

1. Introduction

Aspect-based sentiment analysis (ABSA) is a significant task in Natural Language Processing[1]. It analyses the people's comments on the product or event and shows the people's attitude towards specific aspects. ABSA consists of two subtasks, aspect term extraction (AE) and aspect-level sentiment classification (AS)[2]. AE is to mark the aspect in the context, and AS is to judge the sentiment polarity of the aspect. For example, in the sentence "the price is reasonable although the service is poor", the aspect terms are "price" and "service", which conveys positive and negative sentiment polarity respectively.

Recently, the research on ABSA has been gradually advanced, and many models and methods have been proposed. Traditional methods are mainly based on machine learning, but such methods often require a large number of feature engineering, high cost and poor portability. Today, some excellent deep learning models have gradually emerged, such as Recurrent Neural Network (RNN)[3], Long Short-Term Memory network (LSTM)[4] and Transformer[5] have proposed in recent years, which have achieved improved results in ABSA. Based on this, we propose a multi-task learning network based on attention (MNA), which makes AE and AS run and output at the same time, it avoids the error conduction caused by the pipeline model. And we implement the information transmission from AE to AS, which makes the AS benefit from AE. In MNA, we use the word embedding layer and Encoder part of Transformer for feature extraction, and we use Bi-directional Long Short-Term Memory network (BiLSTM)[6] and multi-head attention mechanism in AE. BiLSTM can better contact context information. The attention mechanism is used to integrate the information from AE and AS. Finally, the whole network structure iteratively updates the shared sequence of vectors.
through the message passing mechanism. MNA makes full use of the information from AE and AS, and achieves better experimental results under the interaction of the internal network structure.

2. Related work

In this part, we will briefly review the related work of aspect extraction, aspect-level sentiment classification and aspect-based sentiment analysis.

The task of AE is to extract aspect terms or opinion terms in sentences. Recently, deep learning is widely used in AE task. Wang et al.[7] proposed a multi-layer attention network for interactive learning to extract aspect terms and opinion terms. Xu et al.[8] proposed a CNN model with general-purpose embeddings and domain-specific embeddings for supervised aspect extraction. The task of AS is to obtain the corresponding sentiment polarities of aspects. Wang et al.[9] designed an attention-based LSTM for aspect-level sentiment classification. Li et al.[10] proposed a component to generate a specific aspect representation, which is between the CNN layer and the bi-directional RNN layer.

Aspect-based sentiment analysis has developed rapidly in recent years. Wang et al.[11] proposed a multi-task neural network framework, which processes aspect extraction and sentiment prediction tasks simultaneously, and learning aspect-sentiment relationship by attention mechanism. Li et al. [12] proposed a model using a unified annotation scheme, which consists of two stacked RNNs assisted with sentiment analysis. He et al.[13] designed an interactive multi-task learning network, which can use the information of the extraction task and learn from the document-level knowledge.

3. Model

In this paper, we focus on the problem setting of a multi-task learning network based on attention mechanism in ABSA. This task can be formulated by three modules. The overall structure of model is shown in figure 1. Firstly, we input the token sequence \( x_1, x_2, ..., x_n \) into the shared word embedding module (SE), which maps the input token sequence to the sequence of latent vectors \( h_1, h_2, ..., h_n \). Then, it is simultaneously fed to AE and AS for processing. To extract the aspect terms and opinion terms in the sentence in AE, and classify the sentiment polarities of the extracted aspect terms in AS. Finally, the output of different modules is combined through message transmission. We denote that \( h_i^{(t)} \) is the value of the shared latent vector corresponding to the \( i \)th word \( x_i \) in the token sequence after \( t \)-round message delivery. Now we describe each module of the model in detail.

![Figure 1. The architecture of MNA.](image-url)
3.1. SE module
First, the pre-trained GloVe word vector is used to transform the token sequence into a word embedding matrix in the word embedding layer, and then further feature extraction is performed in the Encoder part of the Transformer model.

3.2. AE module
We use the BIO marking scheme to express the extraction task as a sequence marking problem. For the aspect terms, we use three types of tags \{BA, IA, O\} to denote the beginning, interior of the aspect terms and other words; for the opinion terms, we also use three types of tags \{BP, IP, O\} to express the beginning, interior of the opinion terms and other words. First, we use the output of SE as the input of BiLSTM, which can fully contact the context information. Then, we use the multi-head attention mechanism with \(n\) heads, it can combine the representations of different subspaces. Finally, we use the fully-connected layer and the softmax activation function, which map the final word representation to a probability distribution \(y^{ae}_1, y^{ae}_2, ..., y^{ae}_n\), which represents the probability that each token in the token sequence is an aspect term or opinion term.

3.3. AS module
We execute aspect-level sentiment classification through AS, we formulate AS as a sequence tagging problem with labels \{pos, neg, neu\}, indicating positive, negative, and neutral sentiment respectively. We process the output sequence from SE in AS. Firstly, the sequence of vectors \(h_1^{as}, h_2^{as}, ..., h_n^{as}\) is obtained by preliminary feature extraction from Convolutional Neural Networks (CNN). Then, in order to enable the prediction of AS to make use of the information of opinion terms extracted from AE, we calculate the output of AE \(\hat{y}_1^{ae}, \hat{y}_2^{ae}, ..., \hat{y}_n^{ae}\) and \(h_1^{as}, h_2^{as}, ..., h_n^{as}\) by attention score. Specifically, the improved multi-head self-attention matrix \(A\) is calculated as follows:

\[
\text{score}_{ij}^{(i\neq j)} = \frac{\exp (\text{score}_{ij})}{\sum_{k=1}^{d_k} \exp (\text{score}_{ij})}
\]

\[
A^{(i\neq j)} = \frac{\exp (\text{score}_{ij})}{\sum_{k=1}^{d_k} \exp (\text{score}_{ij})}
\]

where the first term in equation (1) indicates the semantic correlation between the \(i\)th token and the \(j\)th token; the second term indicates the relationship between the aspect term and opinion term marked in the AE; [:] indicates the concatenation operation; the third term indicates the relative position of two tokens; and the fourth term \(P^{op}_j\) indicates the probability of predicting that the \(j\)th token is part of any opinion term. \(W^{as}\) is parameter matrix. Finally, using the fully-connected layer and the softmax activation function, the final token representation is mapped to the probability distribution \(\hat{y}_1^{as}, \hat{y}_2^{as}, ..., \hat{y}_n^{as}\). The final probability distribution we get in the AS module is the final prediction result of our sentiment classification.

3.4. Learning and training
The training is a process of iteratively updating the sequence. To take advantage of the interactive information between different tasks, we integrate the predictions of different tasks from each iteration and use this information to update \(h_i^{s(t)}\). It is expressed as:

\[h_i^{s(t)} = f(h_i^{s(t-1)}; \hat{y}_1^{ae}, \hat{y}_2^{ae}, ..., \hat{y}_n^{ae})\]

where \(t > 0\) is required in the above formula, [:] indicates the concatenation operation. We employ the fully-connected layer with the ReLu activation function as the encoding function \(f\). We use the existing aspect-level sentiment analysis dataset for training, and its loss function is:

\[
L = \frac{1}{N_a} \sum_{i=1}^{N_a} \frac{1}{n_i} \sum_{j=1}^{n_i} (L(y_i^{ae}, \hat{y}_i^{ae(T)}) + L(y_i^{as}, \hat{y}_i^{as(T)}))
\]

where \(N_a\) denotes the total number of aspect-level training instances, \(n_i\) denotes the number of token contained in the \(i\)th training instance, and \(y_i^{ae}, \hat{y}_i^{ae(T)}\) represents the one-hot of the standard tags of the
dataset in the AE(AS) task, $T$ denotes the maximum number of iterations of the message passing mechanism, $l$ is the cross entropy loss on each token in the token sequence.

4. Experiments

4.1. Experimental settings
Dataset statistics are shown in table 1. We run experiments on three benchmark datasets, which from SemEval2014[14] and SemEval2015[15]. The tags of Datasets contain positive, negative and neutral.

| Datasets | Train | Test |
|----------|-------|------|
| aspect   | opinion | aspect | opinion |
| D1 | Restaurant14 | 3699 | 3484 | 1134 | 1008 |
| D2 | Laptop14 | 2373 | 2504 | 654 | 674 |
| D3 | Restaurant15 | 1199 | 1210 | 542 | 510 |

In the word embedding layer of SE, we use pre-trained Glove word vectors with 300 dimensions. In AE, the number of heads of the multi-head attention is set to 6. In AS, the number of CNN layer is set to 2, about the number of heads $n$ of the improved multi-head attention, we made $n$ equals 6, and the number of iterations to update the shared sequence of vectors is set to 2. We use Adam optimizer with learning rate set to $10^{-4}$ and set batch size to 32. In the training phase, we randomly select 20% of the training data from the datasets as the verification set, and only use the remaining 80% for training.

In our experiment, two indicators were used to evaluate the accuracy and F1.

4.2. Models comparison
To comprehensively evaluate the performance of our proposed model, we choose the following baseline models.

- CMLA-ALSTM: The model uses the CMLA model to extract aspects and the ALSTM model to aspect-level sentiment classification. It is a pipeline model.
- DECNN-ALSTM: The model uses DECNN and ALSTM for aspect extraction and sentiment classification respectively. It is a pipeline model.
- MNN: It is a multi-task neural learning framework for simultaneous aspect extraction and sentiment prediction subtasks.
- INABSA: It is an end-to-end unified model with a unified label scheme, including two stacked recurrent neural networks.
- IMN: It is an interactive multi-task learning network, and it can jointly learn multiple related tasks at the token level and document level at the same time.

4.3. Results and analysis
We verify the performance of the model through experiments. Table 2 shows the comparison of the experimental results of different models. F1-a and F1-o indicate the F1 extracted from aspect terms and opinion terms, acc-s and F1-s indicate the accuracy rate and F1 of aspect-level sentiment classification task respectively, and F1-I indicate the F1 of the whole task. It can be found that MNA has been improved on F1-I. We select two better extraction models and a classification model to form two pipeline models. The experimental results show the integrated approach model is better than pipeline approach. Compared with the optimal integrated model, the performance of MNA in the three datasets is improved by 3.31%, 4.28% and 3.89% respectively. We can see that it is indeed beneficial for us to use multi-head attention mechanism to emphasize aspect terms and opinion terms in AE and AS, as well as the relevance of classification and extraction to realize the information transmission from AE to AS.
Table 2. Model comparison.

|      | CMLA-ALSTM | DECNN-ALSTM | MNN | INABSA | IMN | MNA |
|------|------------|-------------|-----|--------|-----|-----|
| D1   |            |             |     |        |     |     |
| F1-a | 82.45      | 83.94       | 83.05 | 83.92  | 83.33 | 83.65 |
| F1-o | 82.67      | 85.60       | 84.55 | 84.97  | 85.61 | 85.71 |
| acc-s| 77.46      | 77.79       | 77.17 | 79.68  | 83.89 | 83.75 |
| F1-s | 68.70      | 68.50       | 68.45 | 68.38  | 75.66 | 75.39 |
| F1-I | 63.87      | 65.26       | 63.87 | 66.60  | 69.54 | 69.77 |
| D2   |            |             |     |        |     |     |
| F1-a | 76.80      | 78.38       | 76.94 | 77.34  | 77.96 | 78.63 |
| F1-o | 77.33      | 78.81       | 77.77 | 76.62  | 77.51 | 78.71 |
| acc-s| 70.25      | 70.46       | 70.40 | 72.30  | 75.36 | 75.32 |
| F1-s | 66.67      | 66.78       | 65.98 | 68.24  | 72.02 | 72.14 |
| F1-I | 53.68      | 55.05       | 53.80 | 53.88  | 58.37 | 58.62 |
| D3   |            |             |     |        |     |     |
| F1-a | 68.55      | 68.32       | 70.24 | 69.40  | 70.04 | 70.19 |
| F1-o | 71.07      | 71.22       | 69.38 | 71.43  | 71.94 | 71.86 |
| acc-s| 81.03      | 80.32       | 80.79 | 82.56  | 85.64 | 85.78 |
| F1-s | 58.91      | 57.25       | 57.90 | 58.81  | 71.76 | 71.87 |
| F1-I | 54.79      | 55.10       | 56.57 | 57.38  | 59.18 | 59.41 |

4.4. Impact of $n$

We study the effect of hyperparameter $n$ on the final performance of the model. We set the value of $n$ to 0, 6 and 12, where 0 means the model without attention mechanism, and the result of F1-I is shown in table 3. It can be seen that the best result can be obtained when $n$ is set to 6, so we set the value of the hyperparameter $n$ to 6 in the model. When $n$ equals 0, the experimental result is relatively poor, which proves the effectiveness of the attention mechanism in our model.

Table 3. Comparison experiment of $n$

| n   | D1  | D2  | D3  |
|-----|-----|-----|-----|
| 0   | 67.12 | 56.74 | 57.80 |
| 6   | 69.77 | 58.62 | 59.41 |
| 12  | 69.23 | 57.95 | 59.08 |

5. Conclusion

In this paper, we propose a multi-task learning model based on attention, which includes information transmission extraction and classification tasks, and is applied to the research of aspect extraction and aspect-level sentiment classification tasks. MNA fuses the information from the aspect extraction task into the attention mechanism of the aspect-level sentiment classification task, which greatly increases the accuracy of the classification task and further improves the performance of the model. In addition, MNA could also make adaptive changes in other similar downstream tasks in the NLP field, which is suitable for the application.

Acknowledgements

I sincerely thank my teachers and friends for taking the time to help me after their busy study and work. I am grateful for the support provided by my laboratory. They gave me encouragement and guidance when I encountered difficulties in my research.

References

[1] Zhang, L., Wang, S., Liu, B. (2018) Deep learning for sentiment analysis: A survey. Wiley Interdiscip Rev Data Min Knowl Discov 8(4).

[2] Hu M, Liu B (2004) Mining and summarizing customer reviews. In: {ACM} 2004. Washington. pp 168-177.
[3] Usama, M., Ahmad, B., Song, E., Hossain, M., Alrashoud, M. (2020) Attention-based sentiment analysis using convolutional and recurrent neural network. Future Gener. Comput. Syst, 113: 571-578

[4] Hochreiter, S., Schmidhuber, J. (1997) Long short-term memory. Neural Computation 9(8):1735-1780.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser L., Polosukhin, I. (2017) Attention is all you need. In: {NIPS} 2017. Long Beach. pp. 5998-6008.

[6] Nguyen, N., Le, A., Pham, H. (2016) Deep Bi-directional Long Short-Term Memory Neural Networks for Sentiment Analysis of Social Data. In: {IUKM} 2016. Da Nang. pp. 255-268.

[7] Wang, W., Pan, S., Dahlmeier, D., Xiao, X. (2017) Coupled Multi-Layer Attentions for Co-Extraction of Aspect and Opinion Terms. In: {AAAI} 2017. San Francisco. pp. 3316-3322.

[8] Xu, H., Liu, B., Shu, L., Yu, P. (2018) Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. In: {ACL} 2018. Melbourne. pp. 592-598.

[9] Wang, Y., Huang, M., Zhu, X., Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In: {EMNLP} 2016. Austin. pp. 606-615.

[10] Li, X., Bing, L., Lam, W., Shi, B. (2018) Transformation networks for target-oriented sentiment classification. In: {ACL} 2018. Melbourne. pp. 946-956.

[11] Wang, F., Lan, M., Wang, W. (2018) Towards a One-stop Solution to Both Aspect Extraction and Sentiment Analysis Tasks with Neural Multi-task Learning. In: {IJCNN} 2018. Rio de Janeiro, pp. 1-8.

[12] Li, X., Bing, L., Li, P., Lam, W. (2019) A Unified Model for Opinion Target Extraction and Target Sentiment Prediction. In: {AAAI} 2019. Hawaii. pp. 6714-6721.

[13] He, R., Lee, W., Ng, H., Dahlmeier D. (2019) An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In: {ACL} 2019. Florence. pp. 504–515.

[14] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androustopoulos, I., Manandhar, S. (2014) Semeval-2014 task 4: Aspect based sentiment analysis. In: SemEval 2014. Dublin. pp 27-35.

[15] Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androustopoulos, I. (2015) Semeval-2015 task 12: Aspect based sentiment analysis. In: SemEval 2015. Denver. pp 486-495.