Co-attention Networks for Aspect-level Sentiment Analysis

Haihui Li[1,2], Yun Xue[1](✉), Hongya Zhao[2], Xiaohui Hu[1], and Sancheng Peng[3]

[1] School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou, 510006, China
[2] Industrial Central, Shenzhen Polytechnic, Shenzhen, 518055, China
xueyun@scnu.edu.cn
[3] Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou, 510006, China

Abstract. Aspect-level sentiment analysis has identified its significance in sentiment polarity classification of consumer review. For the purpose of specific target sentiment analysis, we put forward a co-attentive deep learning method in the manner of human processing. To start with, the GRUs are taken to extract the hidden states of the different word embeddings. Further, via the interactive learning of the co-attention network, the representations of the target and the context can be obtained. In addition, the attention weights are determined based on the self-attention mechanism to update the final representations. The experimental results evaluated on the SemEval 2014 and Twitter establish a strong evidence of the high accuracy.

Keywords: aspect-level sentiment analysis · co-attention · GRU.

1 Introduction

Natural language processing (NLP) concerns the interactions between computers and human natural languages, which provides morphologic, syntactic and semantic tools to transform stored text from raw data into useful information [1]. As one major interest of NLP, sentiment analysis refers to the ability to identify both the terms and the positive or negative tone of the text. An increasing amount of available text data, such as opinions, critics and recommendations on the Internet, adopts sentiment analysis approaches for opinion mining and product recommendation. Instead of classifying the overall contextual polarity of a document, recent studies focus on finding the attitudes on certain topics within the texts. For instance, in the sentence *Average to good Thai food, but terrible delivery*, the sentiment polarity is positive when target is *Thai food* while the sentiment polarity becomes negative for the target *delivery*. For this reason, aspect-level sentiment analysis paves a way for greater depth of analysis, which is more fine-grained and more sophisticated due to its distinguishing the specific target together with the corresponding contexts. As such, research is ongoing to design and deploy new algorithms for aspect-level sentiment analysis.

More recently, the progresses in NLP tasks are, however, largely driven by the flourish of deep learning algorithms in line with the increased computational resources. The deep learning models can learn the semantic representations from high dimensional original data automatically without carefully designed features, that is, it can use end-to-end training without any prior knowledge [2]. On this occasion, the recurrent neural networks (RNNs) show their superiority in sentiment analysis because it is capable of tackling variable length of input data. Notably, two of the most commonly used RNNs, namely long short-term memory (LSTM) and gated recurrent unit (GRU), are currently widespread in aspect-level sentiment.

Whereas, according to the complexity of human natural language, issues of sentiment analysis are far from solved. Aiming at teaching a computer to handle the texts that is distinctly human, the attention mechanism is introduced. Attention mechanism is initially inspired by human visual attention, which is both creative and practical in image recognition [3]. There are some research using attention mechanism to resolve aspect-level sentiment analysis tasks as well [4]. Further, aiming to carry out the working flow in a manner of human processing, the deployment of the model is taken as the key point. Above all, the online users pay attention to specific words out of interest instead of reading word by word in most cases. By considering the information delivered by the target word and its contexts, the sentiment of this review is updated in the mind. Since humans consistently outperform computers in the semantic interpretation, researchers tend to revise current algorithms to optimize the procedures [5].

In order to carry out the human analyzing practice, a co-attention based neural network is established to facilitate the aspect-level sentiment analysis in this work, whose major significance is its attending target and context information simultaneously. The attentive representation of each part is established while the interaction between the target and the context is studied, and thus to determine the contribution of different words in the target aspect. The utilization of the co-attention based deep learning model, targets at optimizing the model efficiency and improving the working accuracy.

2 Related Work

2.1 Aspect-level sentiment analysis

Previous work has proved that the RNNs have achieved impressive performance in representation learning. Concerning the sentiment analyzing approaches, the LSTM model integrates the RNN with a memory unit and a set of gate units to transmit long-term information, aiming to extract the representations from input data and store it for long or short time durations [6]. As such, specific models, such as Target-Dependent LSTM (TD-LSTM), are developed for target dependent sentiment analysis. Likewise, GRU, which is very similar to the LSTM unit, is applied to aspect-level sentiment analysis as well. GRU can be considered as a lighter version of RNN with lower computation cost and simpler

model structure. In comparison to LSTM, the GRU shows its strong capability of modeling long-term sequences of texts as well.

In addition, the integration of attention mechanisms with deep learning models is devised for settling specific tasks. For the purpose of sentiment analysis of specific target, [7] propose attention-based LSTM with target embedding (ATAE-LSTM). [2] design the Interactive attention networks (IAN) model for well representing a target and its collocative context to facilitate the sentiment classification. [8] establish the attention-over-attention (AOA) based network to model target and context in a joint way and explicitly capture the interaction between them. For describing the relation between target and its left/right contexts, [9] apply a rotatory attention mechanism to the model of three Bi-LSTMs. [10] present GRU has a better working performance than LSTM while integrating with attention mechanism.

2.2 Co-attention network

The main purpose of employing the attention mechanism is that both targets and contexts deserve special treatment. The representations of each part can be learned via self and interactive learning. For example, the phrase *picture quality* will definitely be associated to the expression of *clear-cut* within one text, where the effects on each other are established. On the other hand, there can be more than one word in the target. For example, in the sentence *customer service is terrible*, the word *service* is of more importance than *customer* in delivering the negative sentiment. For this reason, the co-attention network can be carried out to ensure the detection of within and cross-domain interactions [7].

Originally, the co-attention is developed to intimate the eye movements for decision-making, in the same way of human focusing on a series of identical attention regions through repeatedly looking back and forth [11]. Currently, the co-attention network is applied to the image detection tasks, which refers to the mechanism that jointly reasons about the visual attention of different part within one image [12]. Thereby, remarkable evolution is made in the fusion of visual and language representations in images [13]. In aspect-level sentiment analysis, the co-attention mechanism can be used to remove unrelated information and obtain the essential representations. In [14], Zhang et al. preliminary verify co-attention mechanism in capturing the correlation between aspect and contexts. To this end, methods can be applied to update the co-attentive representation of both parts in further steps.

3 Methodology

3.1 Model establishing

We propose a model that aims to get the discriminative representations of both the target and the contexts in aspect-level sentiment analysis tasks. The algorithm is devised by using the GRU together with the co-attention network, as shown in Fig.1.

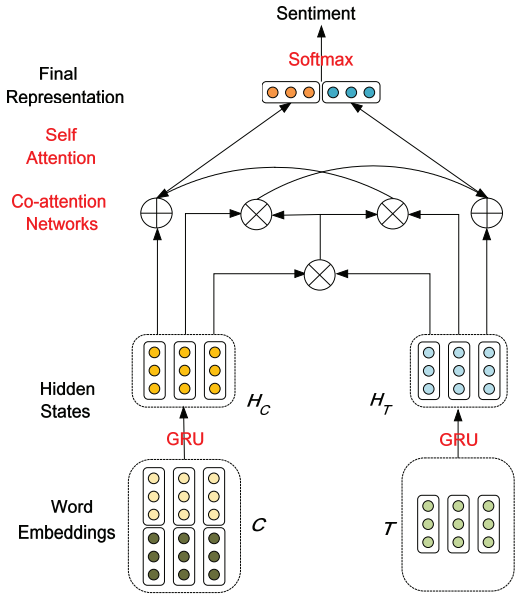

Fig. 1. Overall architecture of model

For a given sequence of n words $[w_{\tilde{c}}^1, w_{\tilde{c}}^2, \ldots, w_{\tilde{c}}^n] \in \mathbb{R}^n$, we call it the context. Within the context, a target with m words $[w_t^1, w_t^2, \ldots, w_t^m] \in \mathbb{R}^m$ contains one or more consecutive words.

We take the word embeddings to represent words from the vocabulary, from which the target and the context can be transformed into $\tilde{C} = [v_{\tilde{c}}^1, v_{\tilde{c}}^2, \ldots, v_{\tilde{c}}^n] \in \mathbb{R}^{(d_w \times n)}$ and $T = [v_t^1, v_t^2, \ldots, v_t^m] \in \mathbb{R}^{(d_w \times m)}$ respectively where d_w represents the dimension of the word embeddings. Thus, the target word embeddings are sent to mean-pooling for getting the average value according to Fig.1.

$$t = \sum_{j=1}^{m} v_t^j / m \tag{1}$$

Hereafter, the outcome t is concatenated into the contexts and a revised word embedding $C = [v_c^1, v_c^2, \ldots, v_c^n] \in \mathbb{R}^{(2d_w \times n)}$ is therefore obtained.

To extract the internal feature of the word embeddings, the GRU is employed to learn the hidden semantics where $H_C \in \mathbb{R}^{(d_h \times n)}$ and $H_T \in \mathbb{R}^{(d_h \times m)}$ are the hidden representations of the context and the target, separately and d_h is the dimension of the hidden layer. The context representation is based on the importance of different words in it and the effect of word sequence in the target. So does the target. The attention weight matrix is computed via the interactive learning of the co-attention network, which is

$$H = relu(H_C^T W H_T) \tag{2}$$

where $W \in \mathbb{R}^{(d_h \times d_h)}$ stands for the parameter matrix and H_C^T is the transposed matrix of H_C.

Apparently, there is a natural symmetry between the contexts and target, based on which the co-attention mechanism is performed. Besides, the word sequences within a fixed part are also taken to express the significances. In this way, the target representation R_T and the context representation R_C are given in eqn.3 and eqn4.

$$R_C = relu(W_C H_C + W_T H_T H^T) \tag{3}$$

$$R_T = relu(W_T H_T + W_C H_C H) \tag{4}$$

where W_C and W_T are the parameter matrices while H^T is the transposed matrix of H.

At this stage, we employed the self-attention to convey the importance of each word in the sequence thoroughly. The self-attention weights of the context and the target are expressed as α and β:

$$\alpha = softmax(w_C R_C) \tag{5}$$

$$\beta = softmax(w_T R_T) \tag{6}$$

where w_C and w_T refer to the parameter vectors and softmax represents the normalization function. The vector representation of the target and the context are determined by using the weighted summation:

$$r_C = \sum_{j=1}^{n} \alpha^i R_C^i \tag{7}$$

$$r_T = \sum_{j=1}^{m} \beta^j R_T^j \tag{8}$$

The final representation $r \in \mathbb{R}^{2d_h}$ for sentiment classification is obtained via concatenating the two parts. By sending r to the softmax classifier, the sentiment distribution of the given target can be identified as

$$x = W_r r + b_r \tag{9}$$

$$y_i = \frac{exp(x_i)}{\sum_{j=1}^{C} exp(x_j)} \tag{10}$$

where W_r is the parametric matrix, b_r is the bias and C is the number of sentiment polarities.

3.2 Model training

The training process is carried out by using the cross entropy with L_2 regularization as the loss function, which is expressed as:

$$J = -\sum_{i=1}^{C} g_i log y_i + \lambda_r (\sum_{\theta \in \Theta} \theta^2) \tag{11}$$

where g_i is the real distribution of sentiment and y_i is the predicted one. Besides, λ_r is the weight of L_2 regularization. The gradients, as well as other parameters are updated through back propagation with the learning rate λ_l:

$$\Theta = \Theta - \lambda_l \frac{\partial J(\Theta)}{\partial \Theta} \tag{12}$$

4 Experiments

4.1 Experimental setting

We carry out our experiments on three public datasets as shown in Table 1. The customer reviews of the laptop and the restaurant are available on SemEval 2014 Task4[4] and the last one is provided in [15]. All the reviews in the experiment are labeled as three different polarities: positive, neutral and negative. In this work we adopt the accuracy as the evaluation metric to demonstrate the working performance. The initialization of all word embeddings is conducted using Glove[5]. All the parameter matrices involved are generated within the distribution $U(-0.1, 0.1)$ randomly and the bias set as 0. The hidden states dimension of GRU is set as 200 with the learning rate of 0.001. In addition the L_2 regularization weight is set as 0.0001. The dropout rate is 0.5 to prevent overfitting.

Table 1. Statistics of Dataset

Dataset	Positive	Neutral	Negative
Laptop-Training	994	464	870
Laptop-Testing	341	169	128
Restaurant-Training	2164	637	807
Restaurant-Testing	728	196	196
Twitter-Training	1561	1560	3127
Twitter-Testing	173	173	346

[4] The detail introduction of this task can be seen at:http://alt.qcri.org/semeval2014/task4/

[5] Pre-trained word vectors of Glove can be obtained from http://nlp.stanford.edu/projects/glove/

4.2 Results

Comparative models are presented as follows:

LSTM: The LSTM network is taken to detect the hidden states of the both the target and the context.

TD-LSTM: The contexts information is detected via two LSTM on both left and right contexts of target [16].

ATAE-LSTM: The LSTM, together with the concatenating process, is applied to get the representation of the target and the context. The attention network aims to select the word of sentiment significance [7].

MemNet: This model develops a multiple-attention-layer deep learning approach instead of using sequential neural networks [17].

IAN: The representations are modeled on the foundation of the LSTM based interactive attention networks. Hidden states are taken to compute the attention scores by the pooling process [2].

RAM: The integration of LSTM and Recurrent Attention is established, targeting at solving the multiple words attention issue [18].

AOA-LSTM: The bidirectional LSTMs are used for getting the hidden states of the target and the context, which are sent to attention-over-attention networks for calculating the final representations [8].

LCR-Rot: LCR-Rot performs the left context, target phrase and right context with 3 separated LSTMs. A rotatory attention network is used to model the relation between target and left/right contexts [9].

Specifically, the proposed model with merely self-attention is considered for comparison. The testing accuracy of each dataset is shown in Table 2.

Table 2. Experimental outcomes

	Methods	Restaurant(%)	Laptop(%)	Twitter(%)
Baselines	LSTM	74.30	66.50	66.50
	TD-LSTM [16]	75.60	68.10	70.80
	ATAE-LSTM [7]	77.20	68.70	-
	MemNet [17]	78.16	70.33	68.50
	IAN [2]	78.60	72.10	-
	RAM [18]	80.23	74.49	69.36
	AOA-LSTM [8]	81.20	74.50	-
	LCR-Rot [9]	81.34	75.24	72.69
proposed model	self-attention	81.51	73.82	72.10
	co-attention+self-attention	**81.61**	**75.86**	**73.85**

Among all the baseline models, LSTM shows the lowest accuracy than any other methods, since the representations are computed of each word equally. The TD-LSTM shows a better working performance due to its introducing the target. More reasonable representations are generated by assigning the attentions to weight different parts in the sentence. The ATAE-LSTM concentrates

on identifying the significance of different words, which gets a 77.2% and a 68.7% on the restaurant and laptop reviews, respectively. Similarly, the Memnet model obtains a comparative result. Further, the IAN model, by using the interactive attention mechanism, strengthens the representations and better the working performance. With multiple attention networks applied, RAM improves the accuracy by 1.63% and 2.39% of restaurant and laptop compared to those of IAN. As for AOA-LSTM and LCR-Rot, the interaction between the target and the context is highlighted and the even higher accuracy can be obtained.

Our model takes advantages of the importance of different parts within the sentence. With only self-attention applied, the proposed model is less competitive than LCR-Rot in the average accuracy. Notably, the application of co-attention network updates the representation of the target and the context. As such, the effects on both the target and the context from each other are considered to further determine the word representations via the interactive learning. By combing with the self-attention mechanism, the final representations are addresses precisely. We can observe that the proposed model outperforms other methods in all datasets to prove its capability.

4.3 Visualization of attention

Our model is further evaluated by visualizing the attention vectors in the sentence. According to the review *Granted the space is smaller than most, it is the best service you will find in even the largest of restaurants*, two targets can be identified, which are *space* and *service*. The sentiment polarity for *space* is negative while that for *service* is positive. In Fig.2, the contribution of different words are listed for aspect-level sentiment analysis. The former illustrates the attention weight for target *space* and the latter for *service*. Words in the darker color are of greater weight, and vice versa.

Our model assigns the importance to *the* and *is small* to carry out the negative sentiment of *space*. However, the word *the* has no direct sentiment towards the target, which can be taken as an error by separating the idiom *the space*. For the target *service*, the word *find* is paid the most attention with *it is* and *best* following. Hence, the model can identify the sentiment polarity towards *service* as positive. In this case, the compositional expression *it is* deliver affirmation of the sentence while the comma before it does not have the same function. In this way, the proposed model misunderstands the meaning of the comma and identifies its significance. Yet we cannot have a 100% accurate rate exactly in a

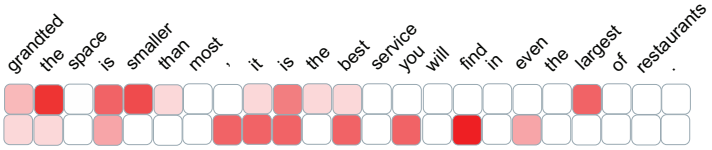

Fig. 2. Attention weights assignment

human processing way. In spite of the aforementioned errors, our model is still able to analyze the sentiment of different target aspects in one sentence properly.

5 Conclusion

In this work, the GRU based co-attention network, in line with human analyzing practice, was developed for aspect-level sentiment analysis task. Aiming to obtain the representations of the word embeddings, the co-attention mechanism was performed to effectively detect the interaction between the target and the context and further identify the words with greater importance. Further, the self-attention was employed to facilitate the determination of the attention weight. As a manner of human processing, the proposed model benefits from the precise and essential presentations of the target and the context. Experiments are conducted on open data sets of review to validate that our model stably outperforms other widely-used models. As such, a better working performance was achieved in aspect-level sentiment analysis.

Further work should be addressed to the relation among different word embeddings to explore the effects on the attentive representation. Although it seems clear that the proposed model can accurately classify the sentiment polarity, it is still an open question whether it is distinctively in accordance with human focusing.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61876205, and the Science and Technology Plan Project of Guangzhou under Grant Nos. 201802010033 and 201804010433.

References

1. Ira Goldstein. *Automated classification of the narrative of medical reports using natural language processing*. Citeseer, 2011.
2. Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074. AAAI Press, 2017.
3. Jiangfeng Zeng, Xiao Ma, and Ke Zhou. Enhancing attention-based lstm with position context for aspect-level sentiment classification. *IEEE Access*, 7:20462–20471, 2019.
4. J. Du, L. Gui, Y. He, R. Xu, and X. Wang. Convolution-based neural attention with applications to sentiment classification. *IEEE Access*, 7:27983–27992, 2019.
5. X. Hu, K. Li, J. Han, X. Hua, L. Guo, and T. Liu. Bridging the semantic gap via functional brain imaging. *IEEE Transactions on Multimedia*, 14(2):314–325, April 2012.

6. Junliang Wang, Jie Zhang, and Xiaoxi Wang. Bilateral lstm: a two-dimensional long short-term memory model with multiply memory units for short-term cycle time forecasting in re-entrant manufacturing systems. *IEEE Transactions on Industrial Informatics*, 14(2):748–758, 2018.
7. Yequan Wang, Minlie Huang, Li Zhao, et al. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.
8. Binxuan Huang, Yanglan Ou, and Kathleen M Carley. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer, 2018.
9. Shiliang Zheng and Rui Xia. Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention. *arXiv preprint arXiv:1802.00892*, 2018.
10. Lishuang Li, Yang Liu, and AnQiao Zhou. Hierarchical attention based position-aware network for aspect-level sentiment analysis. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 181–189, 2018.
11. Lan Lin, Huan Luo, Renjie Huang, and Mao Ye. Recurrent models of visual co-attention for person re-identification. *IEEE Access*, 7:8865–8875, 2019.
12. Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.
13. Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018.
14. Peiran Zhang, Hongbo Zhu, Tao Xiong, and Yihui Yang. Co-attention network and low-rank bilinear pooling for aspect based sentiment analysis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6725–6729. IEEE, 2019.
15. Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 49–54, 2014.
16. Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, 2016.
17. Duyu Tang, Bing Qin, and Ting Liu. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, 2016.
18. Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461, 2017.