Learn More from Context: Joint Modeling of Local and Global Attention for Aspect Sentiment Classification

Siyuan Wang[1,2][0000−0002−5036−0608], Peng Liu[1,2], Jinqiao Shi[1,3], Xuebin Wang[1,2], Can Zhao[1,2], and Zelin Yin[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Beijing University of Posts and Telecommunications
Beijing, China
{wangsiyuan,pengliu1995,wangxuebin,zhaocan,yinzelin}@iie.ac.cn,
shijinqiao@bupt.edu.cn

Abstract. Aspect sentiment classification identifies the sentiment polarity of the target that appears in a sentence. The key point of aspect sentiment classification is to capture valuable information from sentence. Existing methods have acknowledged the importance of the relationship between the target and the sentence. However, these approaches only focus on the local information of the target, such as the positional relationship and the semantic similarity between the words in a sentence and the target. Moreover, the global information of the interaction of words in sentence and their influence on the final prediction of sentiment polarity are ignored in related works. To tackle this issue, the present paper proposes Joint Modeling of Local and Global Attention(LGAJM), with the following two aspects: (1) the study develops a position-based attention network concentrating on the local information of semantic similarity and position information of the target. (2) In order to fetch global information, such as context information and interaction between words in sentences, the self-attention network is introduced. Besides, a BiGRU-based gating mechanism is proposed to weight the outputs of these two attention networks. The model is evaluated on two datasets: laptop and restaurant from SemEval 2014. Experimental results demonstrate the high effectiveness of the proposed method in aspect sentiment classification.

Keywords: Aspect Sentiment Classification · Attention · Gating · CNN.

1 Introduction

Aspect sentiment classification, with its inherent challenges and wide applications, has been an important task in natural language processing (NLP) and draws wide attention both by industry and academia. Aspect sentiment classification aims to identify the sentiment polarity of targets(positive, neutral, negative) that appear in a given sentence. For example, the second sentence in Figs.1, the sentiment polarity of the target *"menu"* should be identified as negative while it would be positive polarity for *"dishes"*. Therefor, a critical demand is to extract valuable information and precisely recognize sentiment polarity in aspect level sentiment classification.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

The existing methods dealing with aspect sentiment classification can be classified into two categories: feature-based methods, such as support vector machine [5], are labor-intensive and highly depend on the quality of features. Neural-network-based methods, such as Target-Dependent LSTM (TD-LSTM) and Target-Connection LSTM (TC-LSTM) [13], can learn features without feature engineering and have been widely used in the fields of NLP. However, these neural-network-based methods cannot effectively identify the importance of words in a sentence. Consequently, the attention mechanism is introduced to promote the neural-network-based methods. The attention mechanism is used to help the model to pay more attention to essential words in a sentence. For example, AE-LSTM [17], ATAE-LSTM [17] and IAN [7] are all designed with attention networks to improve their model for aspect sentiment classification.

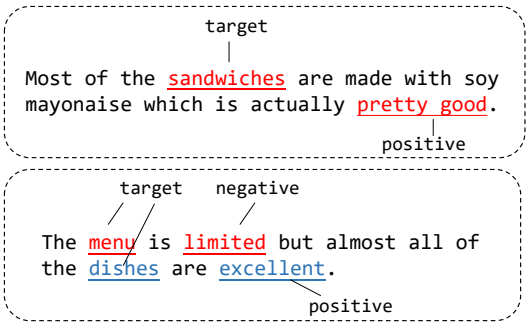

Fig. 1: Examples for aspect-level sentiment classification

These existing methods only considered local information about the target. On the other hand, this paper argues that global information plays an important role in aspect sentiment classification. For instance, considering the first sentence in the Fig.1, words near the target usually get more attention in previous methods, but the sentiment polarity of aspect *"sandwiches"* is determined by opinion words *"pretty good"* which are far away from the aspect and get little attention. Therefore, for the sake of approving the accuracy of the model, global information should be considered in the model.
LGAJM, proposed by the present study, consists of a word representation module, multi-attention module, BiGRU-based gating module, and CNN classifier module and aims at tracking the aforementioned problems. In LGAJM, the word representation module generates different representations of individual sentence words by the target. The multi-attention module contains two attention networks: local attention network and global attention network. These two attention networks can capture not only local information about the aspect but also global information about the sentence, thereby improving classification performance. The BiGRU-based gating module is used to weight the importance of two outputs of the multi-attention module and combine them. The CNN classifier is used to extract n-gram features of the sentence and make sentiment polarity prediction. Experimental results and comparisons on benchmark SemEval2014 datasets have shown the superior performance of the proposed method.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

The contribution of this paper can be summarized as follows:

- LGAJM employs the local attention network to model semantic information and position relevance to extract local relevant information. LGAJM also utilizes the global attention network to model the interaction of sentence words and obtain global contextual information.
- The study propose a BiGRU-based gating mechanism that effectively combines the local attention network and global attention network.
- Experimental results on two benchmark datasets illustrate that the proposed model significantly outperforms the comparative baselines.

2 Related Work

Conventional methods for aspect sentiment classification are rule-based methods and statistic-based methods. Statistic-based methods, such as SVM [5] and MaxEnt-LDA [19], build features between the target and sentence, and then predict the sentiment polarity of target. However, statistic-based methods are highly dependent on the quality of feature engineering work and laborious job.

Recently studies on neural-network-based methods can automatically encode original features as continuous and low-dimensional vectors without feature engineering. Some LSTM-based methods, such as TD-LSTM [13] and TC-LSTM [13] make prediction based on the relationship between sentence and aspect. However, these models ignore which words are more important in a sentence.

The Attention mechanism is taken advantage of assigning different weights to words of different importance in sentences. For example, AE-LSTM and ATAE-LSTM [17] calculate the attention weights with a standard attention mechanism. However, these methods ignore the position information. EAM [4] and PBAN [3] fist use the position information in their model. Position information is used to select keywords and ignore other unimportant words in EAM [4]. In PBAN [3] position information is used to combine position embedding with word embedding as the inputs. But when position information makes a significant influence in a model, global information is hardly considered in the same model.

In an effective attention mechanism, both local position information and global information should require more attention. Multi-head attention [15] has been recognized as an effective mechanism to advance the existing attention functions [6,8,9] and showed its effect on many tasks such as [12,16]. It can catch global information by carrying different features of the target and sentence in different subspaces.

3 Model Description

This section gives a brief description of the architecture of the model in this paper. Our model consists of four modules: a word representation module, mutli-attention module, BiGRU-based gating module and CNN classifier module, as shown in Fig.2. Given a sentence $sen = (w_1, w_2, w_3, ..., w_n)$ containing n words and target entries $a = (a_1, a_2, a_3..., a_m)$ consisting of m words which is a subsequence of sentence sen,

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

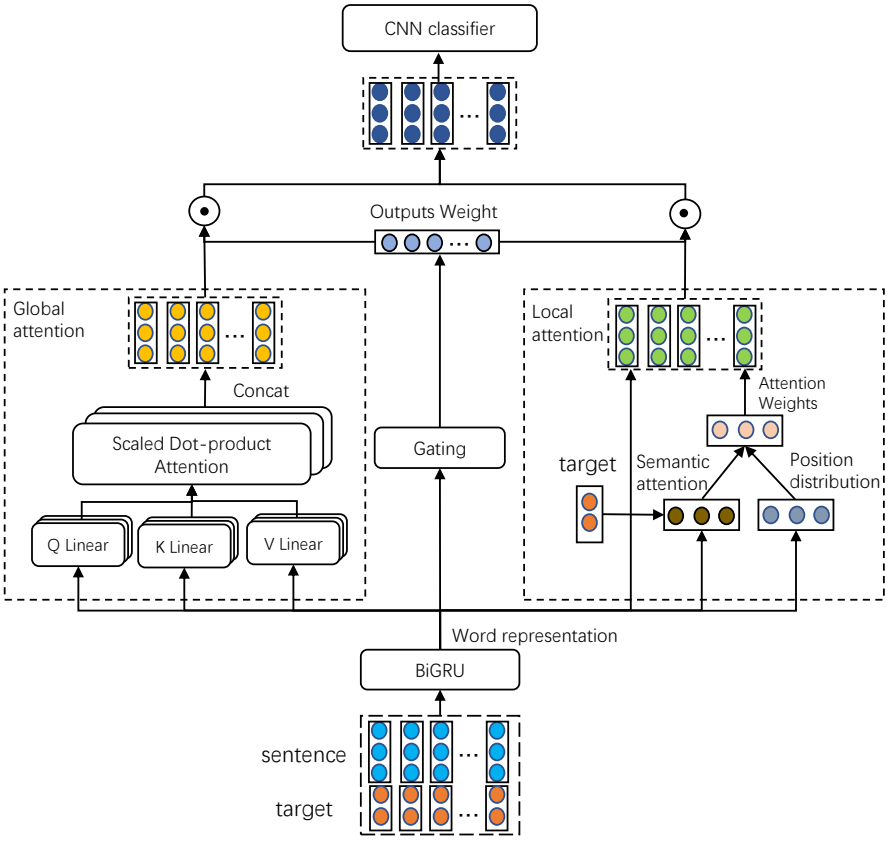

Fig. 2: Architecture of our model.

our model aims to identify the sentiment polarity P of the target. The sentiment polarity of the target belongs to the set $P = \{$"positive", "negative", "neutral"$\}$.

3.1 Word Representation Module

To get high dimensional embedding with abundant information for each word, we employ the pre-trained word vector GloVe [10] to map each word. Define $\mathbb{L} \in \mathbb{R}^{d \times |V|}$ as embedding layer from Glove, where d is the dimension of vectors and $|V|$ is the size of the vocabulary set V. Each sentence word and each target word are required from Glove. Sentence words embeddings are defined as $(e_{w_1}, e_{w_2}, e_{w_3}, ..., e_{w_n})$ and target words embeddings are defined as $(e_{a_1}, e_{a_2}, e_{a_3}, ..., e_{a_n})$.

As targets may have different effects on the final representations of the words in sentence, a sentence should be represented differently based on aspect words. Firstly, a vector a_{avg} is calculated to represent the aspect. Then we generate different represen-

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

tations of each context word by individual target representation. The vector s of words in sentence is defined as follows:

$$a_{avg} = \frac{1}{n}\sum_{i=1}^{n} e_{a_i} \tag{1}$$

$$s_i = [e_{w_i}; a_{avg}] \tag{2}$$

where i indicates the ith word representation, e_{a_i} and e_{w_i} are the ith word embeddings of target and sentence, the dimension of s_i is $2d$. Equation 1 outputs a vector a_{avg} calculated by averagely pooling all word embeddings of target representing the aspect. So s_i is the representation of ith words in sentence under the corresponding target. s_i is generated by concatenating ith sentence word embedding and a_{avg}.
BLSTM models the context dependency with the forward LSTM and the backward LSTM. BLSTM has been widely adopted to NLP tasks [2]. The forward LSTM handles the sentence from left to right, and the backward LSTM handles it from right to left. By BLSTM, model can obtain the bidirectional hidden state of each word. However, to simplify our model, we choose bidirectional GRU which is similar to BLSTM but has fewer parameters and lower computational complexity. LGAJM employs a BiGRU to accumulate the context information for each word of the input sentence.

At each time step t, the new words representation $\overrightarrow{h_t}$ which contains the past information is computed. The update process at time t is as follows:

$$z_t = \sigma(W_z s_t + U_z \overrightarrow{h_{t-1}}) \tag{3}$$

$$r_t = \sigma(W_t s_t + U_z \overrightarrow{h_{t-1}}) \tag{4}$$

$$g_t = tanh(W s_t + U(r_t \circ \overrightarrow{h_{t-1}})) \tag{5}$$

$$\overrightarrow{h_t} = (1 - z_t) \circ \overrightarrow{h_{t-1}} + z_t \circ g_t \tag{6}$$

where σ and $tanh$ are sigmoid and hyperbolic tangent functions, s_t is the representation of word at time t, $W_z, W_t, W \in \mathbb{R}^{d_h \times 2d}$ and $U_z, U_t, U \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters where d_h is the size of $\overrightarrow{h_t}$, z and r simulate binary switches that control to update the information from the current input or not and forget the information in the memory cells or not, respectively.

Then the other representation containing the future information $\overleftarrow{h_t}$ is obtained in the same way, and we concatenate $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ to generate a new representation h_t for each word:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{7}$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

3.2 Attention Module

We have got the contextual representation of each word in sentence. In this section, we employ two attention networks. One is position-based local attention network to capture the local information of target. The other is the self-attention-based global attention network to obtain the global contextual information and model the word interaction of the whole sentence.

Local Attention Network: In terms of the local information of target, there is no doubt that semantic relationship between the words of sentence and target is a meaningful information. The most relevant opinion words of target are found to be near the target by surveying the datasets. So our model adapts the position-based local attention mechanism to capture semantic information and position relevance. In this attention network, the attention weight is obtained based on the local position information and the semantic relationship between sentence and target.

Firstly, we gain the distance of d_i between the ith sentence words and target words:

$$d_i = \begin{cases} |i - a_s| & i<a_s \\ 0 & a_s \leq i \geq a_e \\ |i - a_e| & i>a_e \end{cases} \tag{8}$$

where i is ith word index in the sentence, a_s is the begin index of the target and a_e is the end index of the target.

Then we map d_i to a value between 0 and 1, in the same interval as on part of attention weight:

$$\widetilde{d}_i = 1 - \frac{d_i}{n} \tag{9}$$

where n is the length of sentence.

Another part of attention k_i is calculated by model input representation h_i and target representation a_{avg}. This part is expected to consider semantic information. The calculation process is as follows:

$$k_i = tanh(h_i W^T a_{avg}) \tag{10}$$

where $tanh$ is the hyperbolic tangent functions.

$\widetilde{d}_i$ and k_i are multiplied and do normalization to get the final attention weight $\widetilde{att}_i^p$:

$$att_i^p = \widetilde{d}_i k_i \tag{11}$$

$$\widetilde{att}_i^p = \frac{e^{att_i^p}}{\sum e^{att_i^p}} \tag{12}$$

The ith output m_i^l of local attention network is:

$$m_i^l = \widetilde{att}_i^p h_i \tag{13}$$

where h_i is the contextual representation of ith word.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

Global Attention Network: When target relevant opinion words distribute in a long sentence, previous attention mechanism based on local information can't perform well. However, the self-attention mechanism allows the model to jointly attend to information from different representation subspaces at different positions and has been proven its effectiveness in the field of machine translation [9]. Therefore the self-attention is utilized to obtain a new representation of the whole sentence, so that our model is capable to extract the global context information from the sentence.
The original self-attention network adds position information into word embeddings which changes the original representation. Expanded dimension of word representation with position embedding is aimed at fitting this task. The original word representation h_i is processed as follows:

$$\widetilde{h}_i = [h_i; pos_i] \tag{14}$$

where position embedding pos_i is obtained by looking up a position embedding matrix $\mathbb{L} \in \mathbb{R}^{d \times V}$ which is according to the distance between the word and target. Position embedding matrix is randomly initialized, and updated during the training process. Then query matrix Q_i^j, key matrix K_i^j, value matrix V_i^j are constructed by new words representation the $\widetilde{h}_i$:

$$Q_i^j, K_i^j, V_i^j = \widetilde{h}_i W_j^Q, \widetilde{h}_i W_j^K, \widetilde{h}_i W_j^V \tag{15}$$

where W_j^Q,W_j^K,W_j^V are trainable parameters and j represents the word representation at the jth semantic subspace.
The ith output of self-attention is calculated as follows:

$$head_j = softmax(\frac{Q_i^j K_i^{j\,T}}{\sqrt{d_k}})V_i^j \tag{16}$$

$$m_i^g = Concat(head_1, ..., head_h)W^O \tag{17}$$

where d_k is the dimension of K, h is the count of head, i represents the index of input in global attention network and m_i^g is the final representation of word in global attention network.

3.3 BiGRU-Based Gating mechanism

As attention module has two outputs, $(m_1^l, m_2^l, m_3^l, ..., m_n^l)$ are the outputs of position-based attention network corresponding to the local information, and $(m_1^g, m_2^g, m_3^g, ..., m_n^g)$ are the output of multi-head attention network related to the global information. To effectively combining these outputs, we introduce the gating mechanism. The gating mechanism has proven to be useful for the RNN, and it is used to control the information flowing in a network. Gating mechanism is expected to learn an appropriate weight of two outputs in the process of model training and combine two outputs to the final word representation. As the importance of word global and local representation is related to the contextual information, the final weight is expected to determined based on it. So we make that the representation of the ith word cross the BiGRU network to obtain the weight of the output of local attention module and global attention module.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

A $sigmoid$ activate function which scales a scalar to 0-1 is used to convert ith word representation to a weight f_i:

$$f_i = sigmoid(h_i V^T) \tag{18}$$

where $V^T \in \mathbb{R}^{d_h \times 1}$ is trainable parameter, d_h is the dimension of h_t
So the weight of the ith output of local attention is f_i and the weight of the ith output of global attention is $1 - f_i$. The final representation m_i of the ith word is as follows:

$$m_i = f_i * m_i^l + (1 - f_i) * m_i^g \tag{19}$$

3.4 CNN layer

CNN is introduced to extract n-gram features of sentence. To produce multiple features for CNN, the convolutional layer and the max-pooling layer capture the most crucial features:

$$c_i = \text{relu}(W_{conv}[m_{i-1}, m_i, m_{i+1}] + b_{conv}) \tag{20}$$

$$z = \text{maxpooling}(c_1, c_2, ..., c_n) \tag{21}$$

where $i \in [1, n]$, W_{conv} and b_{conv} are parameters of the convolutional kernel and z is the final representation of the sentence with the target.

Finally, z is fed into a softmax layer to predict sentiment polarity y_i of the target:

$$y_i = softmax(W_s z + b_s) \tag{22}$$

where W_s,b_s are the weight matrix and bias, and y_i is the predicting polarity of the target.
So the training object is minimizing cross-entropy function:

$$loss = -\sum log(y_i pol_i) \tag{23}$$

where pol_i is the real sentiment polarity of target.

4 Experiments

4.1 Experimental Setting

Dataset Settings: Experiments are conducted on two datasets: Restaurant and Laptop from SemEval2014 task4 [11]. These datasets contain lots of user reviews and each review associated with a list of opinion targets and corresponding sentiment polarities. After preprocessing, statistics of the final datasets are given in Table 1.
Hyper-parameters: 20% of the training data is selected randomly as the development set to tune hyperparameters. The network is trained for 200 epochs and select the best model according to the performance on the development set. The final hyperparameters are selected when they produce the highest accuracies on the development set.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

All word embedding are initialized by the pre-trained Glove and the dimension of the word embedding and target embedding are set to 300. The dimension of the BiGRU hidden vectors is 100. The convolutional kernel size is 3. The weight matrices and bias are given the initial value by sampling from the uniform distribution $\bigcup(0.01, 0.01)$. Dropout is applied on the input word embeddings for purpose of avoiding overfitting. Adam with learning rate set to 0.0005 is used for optimizing the models.

Table 1: Statistic results of Laptop and Restaurant.

	Set	Total	Positive	Negative	Neutral
Laptop	Train	2328	994	870	464
	Test	638	341	128	169
Restaurant	Train	3608	2164	807	637
	Test	1120	728	196	196

4.2 Compared Methods

To evaluate the effectiveness of our method, LGAJM is compared with the following methods:

- **SVM**: SVM is a traditional support vector machine based model with extensive feature engineering [5].
- **TD-LSTM**: TD-LSTM uses two LSTMs to model the left and right contexts of the target separately, then predicts the sentiment polarity of target based on concatenated context representations [13].
- **ATAE-LSTM**: ATAE-LSTM learns attention embeddings and combines them with the LSTM hidden states to predict the sentiment polarity of target [18].
- **MemNet**: MemNet employs multiple attention layers over the word embeddings and bases on the output of the last layer predicting the sentiment polarity of target. And it introduces the position information [14].
- **IAN**: IAN adapts two LSTMs to model context and target separately, then interactively get the final representation of context and aspect [7].
- **RAM**: RAM employs multi-hop attention on the hidden states of a positon-weighted bidirectional LSTM to extract the relevant information of target, and nonlinearly combine the result from memory with a GRU [1].
- **PBAN**: PBAN incorporates position embedding and word embedding, then use bidirectional attention mechanism between target and sentence to get the final representation of context and target[3].

To investigate the effectiveness of our model, the variants of the LGAJM model are listed as follows: Local attention+CNN(LAM), Global attention+CNN(GAM). The previous two models are set to verify the effectiveness of two attention mechanisms and prove that two attention mechanisms complement each other well. These two models make up our ablation experiments.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

4.3 Result Analysis

Table 2: Average accuracies and Macro-F1 scores over 5 runs with random initializations. The best results are in bold. The marker * indicates that our full model LGAJM is significantly better than TD-LSTM, ATAE-LSTM, MemNet, IAN, RAM, and PBAN with $p < 0.05$ based on one-tailed unpaired t-test.

	Models	LAPTOP		REST	
		Acc.	Macro-F1	Acc.	Macro-F1
Baselines	SVM [5]	70.49	NA	80.16	NA
	ATAE-LSTM [18]	68.70	NA	77.20	NA
	MemNet [14]	72.21	NA	80.95	NA
	IAN [7]	72.10	NA	78.60	NA
	PBAN[3]	74.12	NA	81.16	NA
	TD-LSTM [13]	71.83	68.43	78.00	66.73
	RAM [1]	74.49	**71.35***	80.23	70.80
Ablated models	GAM	73.35	69.51	79.65	70.75
	LAM	74.14	70.16	80.0	70.80
Full model	LGAJM	**74.92***	**71.35***	**81.16***	**71.73***

The accuracy and Macro-F1 of our proposed model is evaluated and compare with others basic model on Laptop and Restaurant datasets. As shown by the result in Table2, the final reported numbers are obtained as the best value over 5 runs with random initialization. We can get the following observations:
(1) LGAJM achieves the best performance among all methods on both accuracy and macro-F1 for all datasets. This proves the effectiveness of our method.
(2) LAM performs better than SVM, IAN, TD-LSTM, and ATAE-LSTM which not follow position information on both Laptop and Restaurant datasets, which suggests the effectiveness of local position information. GAM also outperforms than SVM, IAN, TD-LSTM, MemNet and ATAE-LSTM on Laptop datasets and outperforms than IAN, TD-LSTM, and ATAE-LSTM on Restaurant datasets, which verifies global information can help final prediction. The reason for LAM outperforming than GAM is that more related opinion words of target are near the target words. So usually local position information can play a better role.
(3) In conclusion, LGAJM performs best on all evaluating indicators and all datasets because LGAJM concentrates on not only local position information but also global information. Two attention networks outputs are combined by BiGRU-based gating mechanism which shows local information and global information is complemented and BiGRU-based gating can combine these effectively.

Analysis of Local Attention The relevant opinion words of target are usually found near the target. The proposed model thus introduces position information to model a

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

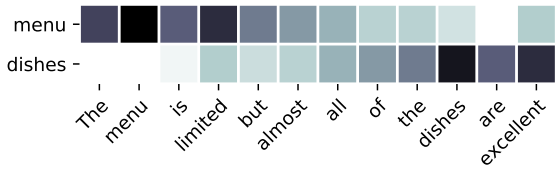

Fig. 3: LAM ablation model's attention distribution

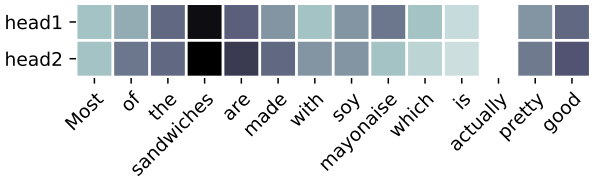

Fig. 4: GAM ablation model's target word attention distribution

position-based attention network that can pay more attention to the adjacency words of the target. As shown in Table 2, experimental results show that the LAM model usually performs better than base methods. To further illustrate the contribution of the local attention mechanism, some sentences are picked from the datasets tested by the ablated model LAM. Fig.3 demonstrates the different distribution of attention. For the sentence in Fig.3, the target wordw are *"menu"* and *"dishes"* and the opinion words are *"limited"* and *"excellent"*. These opinion words all near the target. The local attention network can precisely assign a significantly higher weight to the real opinion word *"limited"* and *"excellent"* than to other words separately. Experimental results show the advantage of the local position especially when there is a short distance between opinion words and target words.

Analysis of Global Attention The global attention mechanism is required to capture the global information, especially when the relevant opinion words of the target are distributed in a long sentence. Similarly to previous experiments, the ablation experiments are also set up and examples are picked to check the effectiveness of the global attention mechanism. Fig.4 illustrates the difference. There is a long-distance between the opinion word *"good"* and target words, and as expected, the ablated model LAM assigns low weight to the related opinion word *"good"*. However, the ablated model GAM assigns obviously higher weight to the word *"good"*. Experimental results show the ability of self-attention to capture global information as well as the effectiveness of global information.

Analysis of Gating Mechanism In the proposed model, the gating mechanism is chosen to control information flow in the model and combine the two outputs of the attention module to ensure that the model considers both local and global information. When the LAM model assigns low weight to the opinion word *"good"* and GAM

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

model assigns higher weight, the gating mechanism resets high weight to the output of GAM. The existing model can thus make the correct prediction. Results show that two attention mechanisms are complementary and the BiGRU-based-gating mechanism demonstrates its strengths.

5 Conclusion

This paper verified the importance of local position information and global context information. In order to ensure that the proposed attention network focuses on relevant words of the target no matter if they are near or far away from the aspect, the study presents a novel model that contains a local attention mechanism and a global attention mechanism. Besides, a BiGRU-based gating mechanism is come up with for learning the weight of these two attention networks' outputs and it adjusts the word representations based on context information of sentence words. The final representation of each word in sentence incorporating the global information and local information is obtained. CNN is used to extract local n-gram features for sentiment classification. Experimental results on two datasets of SemEval 2014 prove the high effectiveness of the proposed model.

References

1. Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *EMNLP*, pages 452–461. Association for Computational Linguistics, 2017.
2. Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.
3. Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 774–784, 2018.
4. Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. Effective attention modeling for aspect-level sentiment classification. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1121–1131, 2018.
5. Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *SemEval@COLING*, pages 437–442. The Association for Computer Linguistics, 2014.
6. Jian Li, Zhaopeng Tu, Baosong Yang, Michael R. Lyu, and Tong Zhang. Multi-head attention with disagreement regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2897–2903, 2018.
7. Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. Interactive attention networks for aspect-level sentiment classification. In *IJCAI*, pages 4068–4074. ijcai.org, 2017.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20

8. Fandong Meng and Jinchao Zhang. DTMT: A novel deep transition architecture for neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 224–231, 2019.
9. Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 1–9, 2018.
10. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL, 2014.
11. Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *SemEval@COLING*, pages 27–35. The Association for Computer Linguistics, 2014.
12. Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4929–4936, 2018.
13. Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. In *COLING*, pages 3298–3307. ACL, 2016.
14. Duyu Tang, Bing Qin, and Ting Liu. Aspect level sentiment classification with deep memory network. In *EMNLP*, pages 214–224. The Association for Computational Linguistics, 2016.
15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.
16. Patrick Verga, Emma Strubell, and Andrew McCallum. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 872–884, 2018.
17. Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615, 2016.
18. Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In *EMNLP*, pages 606–615. The Association for Computational Linguistics, 2016.
19. Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 56–65, 2010.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50436-6_20