Position Weight with Muti-Head Attention for Aspect-Based Sentiment Classification

Ping Ji[a], Xiaohong Liu[b]

Beijing University of Posts and Telecommunications

[a]pingji@bupt.edu.cn, [b]xiaohongliu@bupt.edu.cn

Keywords: Aspect based sentiment classification; Position weights; Multi-head attention

Abstract: The purpose of aspect based sentiment classification (ABSC) is to predict emotional tendencies of a sentence or article in certain aspects. Most of the existing methods use coarse-grained attention mechanisms and do not notice the relationship between emotional orientation and contextual content. In this paper, we proposed a model called PW-MHA in which a pre-trained BERT is applied. PW-MHA uses the multi-head attention mechanism to capture the relationships between internal words. Position weights are used for the classification on the idea that a closer context word is more likely to be the importance word for the target. We evaluated the PW-MHA model on three datasets: laptops and restaurants from SemEval2014, and the ACL 14 twitter dataset. It is shown that the PW-MHA has achieved competitive results on the given datasets.

1. Introduction

Aspect-based sentiment classification (ABSC) [1] [2] is a basic task in sentiment analysis designed to predict emotional polarity (e.g, positive, neutral, negative) in various aspects. For example, giving a sentence "I think the food in this restaurant is not good, but the environment is very beautiful." , the emotional polarity of the words "food" and "environment " are positive and negative respectively. In general, traditional sentence-level or chapter-level sentiment analysis methods such as LSTM [3] cannot effectively analyze the sentiment tendencies of the aspects that need to be predicted, because these methods do not involve specific aspects. It is difficult to determine the overall emotion when a sentence has multiple aspects.

Some methods have been proposed to solve aspect-level sentiment classification. There are two main solutions at present. One way is to build sentiment classifiers through some machine learning methods. The other way is to build a deep neural network. Methods based on deep learning can generally learn the word vector representations of context and aspect, and the most representative method is the LSTM network. However, these models do not handle fine-grained target sentiment classification tasks very well. The main reason in these methods is to focus only on feature recognitions based on global context, but long distance information between words is not easy to transfer. We found that words closer to the aspect may have a greater effect on determining the emotional polarity of an aspect.

We proposed a model of multi-head attention mechanism PW-MHA based on position weights. In this model, a pre-trained BERT [4] layer was employed to capture long-term dependencies within the context. It has been found that the shorter the distance between context words and aspect words, the more important it may be. In the PW-MHA model, a multi-head attention mechanism is used to extract features, which has more powerful computing power compared with traditional LSTM or GRU [5].

2. Related Work

Sentiment analysis is an important branch of natural language processing. Text sentiment analysis includes several important tasks: text sentiment polarity classification (such as text-level or

sentence-level sentiment classification), subjective /objectivity recognition, and fine-grained sentiment analysis. In recent years, many scholars have devoted themselves to the study of emotional polarity at the aspect levels.

Traditional machine learning methods use Naive Bayes or support vector machines (SVM) for supervised training, most of which rely on feature engineering. Based on this consideration, a number of sentiment dictionaries need to be constructed.

Because the deep neural network-based methods have the excellent ability to capture original features, many related works are now based on RNN or RNN's deformed body LSTM and the like. The TD-LSTM [6] proposed in 2015 is a very classic model, which uses two LSTM models for predicting the target word based on the surrounding context. In the past two years especially, pre-trained models have become the research trend of ABSC (aspect based sentiment classification).Pre-trained models can be applied to NLP tasks and greatly improve the performance. ELMo [7] and GPT[8] are pre-trained language models which have achieved excellent results. With the emergence of BERT, many records of NLP tasks were broken in one fell swoop, and many scholars applied BERT to ABSC tasks.o

3. Model

In this section, we will detail the specific structure of our model and its training. PW-MHA requires two parts of inputs: one is a text sequence $w^c = [w_1^c, w_2^c, \ldots w_n^c]$ and the other is a target sequence $w^t = [w_1^t, w_2^t, \ldots w_m^t]$.PW-MHA aims to predict emotional sentiment in a given text sequence and target sequence. The specific structure of the model is shown in Figure 1.

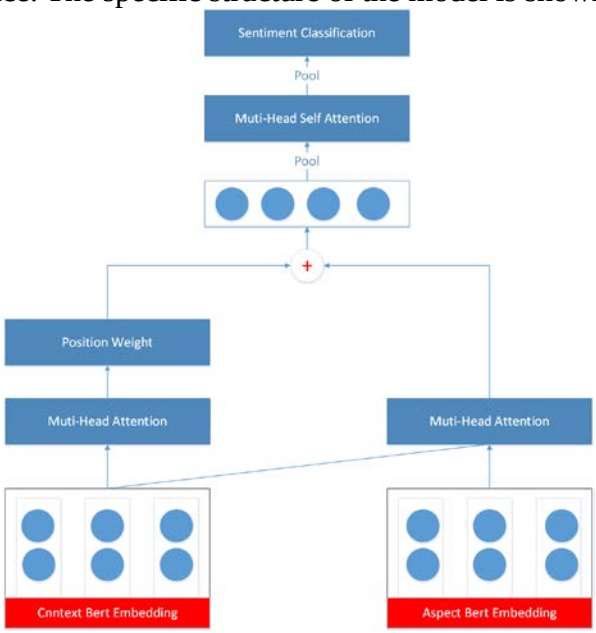

Fig. 1 Specific architecture of PW-MHA design.

3.1. BERT Embedding Layer

In the tradition, the word vector model is a tool that can convert abstract text existing in the real world into vectors that can be manipulated by mathematical formula. BERT is a pre-trained language representation method that users can use to extract high-quality language features from text data. We change the form of the input to “[CLS] + context + [SEP]” and “[CLS] + target + [SEP]” respectively to facilitate fine-tuning later.

3.2. Muti-Head Attention Layer

The essence of multi-head is calculating multiple independent attentions, as an integrated function form to prevent overfitting. The same query sequence Q, key sequence K, value sequence V, are transformed through linear transformation. Each attention mechanism function is only responsible

for a subspace in the final output sequence, which are independent of each other. The calculation of MHA is divided into four steps:

First step: calculate the similarity of Q and K, the result is represented by vector f :

$$f(k_i, q_j) = \tanh([k_i : q_j] * w_{kq}) \quad i, j = 1,2,3 \ldots, n \qquad (1)$$

Where $w_{kq} \in R^{2d_{hidden}}$ is learnable weight and d_{hidden} is hidden layer dimension, n is the length of the input sequences.

Second step: do softmax normalization of the obtained similarity vector f:

$$a_i = \frac{e^{f(k_i,q_j)}}{\sum_j e^{f(k_i,q_j)}} \quad i, j = 1,2,3,..,n \qquad (2)$$

Third step: perform a weighted sum of all normalized similarity values to obtain the attention vector

$$o_i = \sum_i a_i V_i \quad i, j = 1,2,3,..,n \qquad (3)$$

Last step: concatenate all the o_i generated by former step,and do linear product again with weitght w_o.

$$MHA(k, q) = [o_1; o_2; o_3; \ldots; o_{n_{head}}] * w_o \qquad (4)$$

Where $w_o \in R^{d_{hidden} * d_{hidden}}$ is learnable weight vector.

For our task, we compute the multi-head attention vector c and t:

$$c = MHA(w^c, w^c) \qquad (5)$$

Where query sequence Q equals key sequence K.

$$t = MHA(w^c, w^t) \qquad (6)$$

Where query sequence Q is different from key sequence K.

3.3. Position Weight Layer

Our idea is based on the fact that a closer context word is more likely to be the importance word of the target. For example, let us see the sentiments of the sentence "I think their food is not well, but the environment of their service is very good", The word “good” is more important than the word “well” for service. Therefore, in the calculation, aspects of location information are considered in the form of location weights. The weight for word w_i is:

$$dist_{i,a} = |i - \bar{a}| \qquad (7)$$

$$w_i = 1 - \frac{dist_{i,a}}{2dist_{max}} \qquad (8)$$

$$c_i = w_i * c_i \qquad (9)$$

Where i denotes the position of context , $\bar{a}$ denotes the average position of aspect if there are more than one word, $dist_{i,a}$ denotes the distance of $\bar{a}$ and context word. $dist_{max}$ denotes the maximum of $dist_{i,a}$ with different a and i. According to (8), position weights can range from 1.0 to 0.5 and is monotonically decreasing with variable i. As the distance increases, the value of the position weight decreases. We apply the vector generated by (5) to (9).

3.4. Feature Interactive Layer

The results of (6) and (9) are concatenated and followed by a pooling layer. Then muti-head self-attention is applied for the output of the pooling layer. The previous methods directly feed the pooling result into the output layer, but we find it is beneficial to get useful internal information using MHA instead of using pooling alone.

$$o = [c; t] \qquad (10)$$

$$o^{pool} = POOL(o) \quad (11)$$

$$o^{MSHA} = MHA(o^{pool}, o^{pool}) \quad (12)$$

3.5. Output Layer

At last, we pool the vectors o^{MSHA} generated by the previous feature interactive layer, and then feed it to a softmax layer for determining the aspect sentiment polarity:

$$O^{final} = POOL(o^{MSHA}) \quad (13)$$

$$Y = softmax(O^{final}) \quad (14)$$

3.6. Model Training

Cross entropy characterizes the distance between two probability distributions. The smaller the cross entropy is, the closer the distributions of the two probabilities are. So we use cross entropy as a loss function for our model. In addition, in order to prevent overfitting the L2 regular term is added:

$$J = \sum_{(x,y)\in D} \sum_{c\in C} P_c^{true}(x,y) \log P_c^{predict}(x,y;\theta) + \lambda\|\theta\|^2 \quad (15)$$

Where C is the number of emotional polarity, D is the training set, p^{true} is the true sentiment probability of the aspect, $p^{predict}$ is the predicted sentiment probability of the PW-MHA, λ is the regularization weight, and θ is a hyperparameter.We used Adam algorithm to speed up the calculation process.

4. Experiments

In this section, we will introduce our datasets, hyperparameters, and comparison methods for the PW-MHA training in detail. We also conducted two sets of controlled experiments to prove that the position weight and the final MHSA (Muti-Head Self Attention) layer is effective. Finally, we also give a single case test, which proves that PW-MHA can solve the situation of other models' judgment errors.

4.1. Datasets

We perform experiments on three datasets: laptops and restaurants from SemEval2014, and the ACL 14 twitter dataset. The SemEval-2014 Task 4 data set is mainly used for fine-grained sentiment analysis. It includes two fields, Laptop and Restaurant. The datasets in each field is divided into training data, verification data (separated from training data), and test data for supervised machine learning algorithms or deep learning algorithms. The specific situation of the data set is shown in Table 1.

Table 1 The statistics of the datasetsz

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

4.2. Hyperparamenters

In PW-MHA, embedding dimension and hidden dimension are set to 768 in BERT. The learning rate is set to $2 * 10^{-5}$,dropout is set to 0.2,L2 regularization is set to $1 * 10^{-5}$, and batch size is set to 16.In addition, for the optimizer, we choose Adam algorithm.

4.3. Model Comparisons

In order to verify the validity of our model, we selected several baseline methods for comparison. We list them as follows:

TD-LSTM[6] is to model separately according to the context before and after the target words, so in fact, two LSTM models , LSTM Left and LSTM Right are used. The input of LSTM Left is the context before the target word plus the target word, that is, input from the first word of the sentence to the last target words; LSTM Right input is the context after the target word plus the target word, that is, from entering the last word of the sentence to the first target word.

ATAE-LSTM [9] combines attention with LSTM, and uses attention to obtain context information that is more important to different aspects to solve the aspect level sentiment analysis.

IAN [10] enters the corresponding word embeddings in the target and context respectively, and inputs the word embedding into the LSTM network to obtain the output of the hidden layer. Then the model uses the average value of the output of the hidden layer of the target and context to combine the attention mechanism to generate attention weights. The final target attention weights are connected with context attention weights as the input of the softmax function to get the classification results.

RAM [11] proposes a framework that can capture long-distance emotion features based on multiple attentions. This framework is more robust to irrelevant information and combines the results of multiple attentions with RNNs in a non-linear manner.

MGAN [12] mainly uses the interaction between word-level target and context to reduce the loss of coarse-grained attention. Aspect alignment loss is added to the objective function to enhance the difference of context weight learning for aspects with the same context but different emotional polarity.

Table 2 Experimental results of performance

Model		**Laptop**		**Restaurant**		**Twitter**	
		Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
Baselines	TD-LSTM	0.7183	0.6843	0.7800	0.6673	0.6662	0.6401
	ATAE-LSTM	0.6870	-	0.7720	-	-	-
	IAN	0.7210	-	0.7860	-	-	-
	RAM	0.7449	0.7135	0.8023	0.7080	0.6936	0.6730
	MGAN	0.7539	0.7247	0.8125	0.7194	0.7254	0.7081
	BERT-SPC	0.8025	0.7741	0.8598	0.7879	0.7529	0.7363
PE-MHA	PW-MHA	0.8093	0.7781	0.8613	0.7890	0.7591	0.7421
	PW-MHA w/o PE	0.7902	0.7691	0.8436	0.7812	0.7421	0.7311
	PW-MHA w/o MHSA	0.8040	0.7753	0.8562	0.7843	0.7513	0.7356

BERT-SPC [4] uses multi-head attention and position embedding to replace the recurrent neural network, and break the best record of 11 different problems in the field of natural language processing.

4.4. Main Results

As shown in Table 2, compared with several baselines, PW-MHA has got better results.PW-MHA can handle all kinds of comments well, such as relatively formal user product reviews in Laptop and Restaurant, and tweets with more non-grammatical sentences in Twitter.

The PW-MHA model further emphasizes the importance of contextual distance and target through position weighting. It can be seen that PW-MHA achieves the best performance in the baselines we selected. We use a pre-trained BERT word vector, which is also improved compared to BERT-spc, indicating that our downstream model is effective.

We also conducted two sets of comparative experiments in the bottom two rows of Table 2. One

group deleted the position weight layer on the PW-MHA, and the other group deleted the MHSH layer before the classification layer. The results show that both precision and F1 values have decreased after deletions of certain layers, which proves that the idea for PW-MAH is reliable.

5. Case Study

In order to better verify the superiority of the PW-MHA, we use a sample to identify its polarity in sentiment. The sample statement is “The staff should be a bit more friendly.” We used MGAN and IAN for the experiments. The specific results are shown in Table 3.

Table 3 Singleton test results

Model	Label	Prediction
MGAN	negative	positive
IAN	negative	positive
PW-MHA	negative	negative

The reason why MGAN and IAN predict examples as positive is because they give friendly more weight, and in our model, friendly will reduce the weight because it is farther away, reducing intervention in the final prediction result.

6. Conclusion and Future Works

We propose a multi-headed attention model based on position weights, focusing on capturing the distance between context and aspect. In the case of targeted aspect supervision, the PW-MHA model works more accurately at the aspect level. Tests have been done on three datasets and achieved competitive results. Controlled experiments are conducted for two sets to show that the position weight and the MHSA layers are effective. The sample in the section case study can be predicted by PW-MHA, and similar results cannot be given by IAN like models. In the future, we can further consider the representation and calculation of location weights and evaluate the portability of the model.

Acknowledgments

This work is supported by NSFC 61773037.

References

[1] Zhang L , Wang S , Liu B . Deep learning for sentiment analysis: A survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018:e1253.

[2] Young T , Hazarika D , Poria S , et al. Recent Trends in Deep Learning Based Natural Language Processing[J]. 2017.

[3] Hochreiter S, Schmidhuber J. LSTM can solve hard long time lag problems[C]//Advances in neural information processing systems. 1997: 473-479.

[4] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[5] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[6] Tang D , Qin B , Feng X , et al. Effective LSTMs for Target-Dependent Sentiment Classification[J]. Computer Science, 2015.

[7] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[8] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. URL https://s3-us-west-2. amazonaws.com/openaiassets/researchcovers/languageunsupervised/language understanding paper. pdf, 2018.

[9] Wang Y, Huang M, Zhao L. Attention-based LSTM for aspect-level sentiment classification[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 606-615.

[10] Ma D , Li S , Zhang X , et al. Interactive Attention Networks for Aspect-Level Sentiment Classification[J]. 2017.

[11] Chen P, Sun Z, Bing L, et al. Recurrent attention network on memory for aspect sentiment analysis[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. 2017: 452-461.

[12] Fan F, Feng Y, Zhao D. Multi-grained attention network for aspect-level sentiment classification[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3433-3442.

[13] Song Y, Wang J, Jiang T, et al. Attentional encoder network for targeted sentiment classification[J]. arXiv preprint arXiv:1902.09314, 2019.

[14] Zeng B, Yang H, Xu R, et al. LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification[J]. Applied Sciences, 2019.