JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 1419-1433 (2021)
DOI: 10.6688/JISE.202111_37(6).0013

Non-Local Attention Based CNN Model for Aspect Extraction

DANG-GUO SHAO, MING-FANG ZHANG, YAN XIANG+, RONG HU AND TING LU
Faculty of Information Engineering and Automation
Kunming University of Science and Technology
Kunming, 650500 P.R. China
E-mail: {1254116691; 50691012}@qq.com; huntersdg@163.com

Aspect extraction is the basis for aspect-based sentiment analysis, aiming to find out the target of opinions from reviews. The existing neural network for aspect extraction model based on sequence labeling has a poor effect on the extraction of long aspect terms. To solve this problem, we propose a new aspect extraction framework, which uses a three-layer convolutional neural network (CNN) to learn multi-layer semantic features from reviews, and then uses the non-local attention mechanism to obtain dependence feature between different words. Moreover, conditional random field (CRF) has been used to reduce the probability of label conversion errors. CNN filters are good at learning local features without considering long-distance dependencies, while the non-local attention mechanism can strengthen the long-distance dependencies of words and ensure the integrity of long aspect terms. The proposed model is tested on two datasets of SemEval and compared with some baseline models. The experimental results show that the performance of the model is superior.

Keywords: wireless sensor networks, localization, mobile beacon, mobile anchor, RSSI

1. INTRODUCTION

Aspect-based sentiment analysis aims to extract the structured aspect terms and opinion terms from unstructured reviews, and further determines their sentiment polarity (positive, negative, or neutral). For example, in the review "*Boot time is super fast, around anywhere from 35 seconds to 1 minute*", the customer opinion can be summarized as <*Boot time, fast*>, and its sentiment polarity is positive. The aspect terms extraction (ATE) aims to find the object of opinion expression "*Boot time*" from the review, belonging to the attribute of the computer "*operating system*". ATE is one of the most important sub-tasks of the aspect-based sentiment analysis.

At present, ATE models can be roughly classified into unsupervised methods and supervised methods. Unsupervised methods focus on syntactic rules-based extraction [1, 2] and topic models [3], while the supervised methods rely on manually labeled data, mainly using conditional random field (CRF) [4-6], and Neural Networks [7, 8] for feature extraction. However, the existing aspect terms extraction systems still have some problems in practical applications. One of those problems is poor extraction effect on long aspect terms. Long aspect terms are composed of several words to form an opinion entity together. For example, "*build in virus control*" is a long aspect term in the review "*quick and has build in virus control*". We believe that there is a strong dependence between two words in a long aspect term. How to capture this dependence is the key to extract a long aspect term. Xu *et al.* [9] use double-embedding (DE-CNN), and regard general embedding and domain

Received December 26, 2019; revised September 15 & December 25, 2020; accepted January 18, 2021.
Communicated by Hsin-Hsi Chen.
+ Corresponding author.

specific embedding as inputs of the CNN model to alleviate the impact of domain differences. If we only use a CNN-based extractor, the predicted result is the uncompleted aspect term "virus control". Current convolution operators are local operations in space and time, so the effect of long aspect terms extraction is limited. For a CNN-based network, the receptive field is extended by stacking multiple convolutional modules and backpropagating, which can capture long-distance dependencies between words in a review. However, repeating stacking convolution kernels can make the network deeper and the capture efficiency lower.

To solve the problem mentioned above, we use a non-local attention mechanism to directly calculate correlation coefficients between two words, which is used to represent the dependency relationship of words, and is not limited to distance between words. The introduction of long-distance dependencies information between words not only improves the ability of the network to extract long aspect terms significantly, but also enables the network to find some aspect terms that only contain one word and far away from other aspect terms. At the same time, some adjectives with obvious sentimental polarity and verbs with strong relevance to the aspect may be judged as aspect terms with one word, because they are obtained big weights after the calculation of long-distance dependencies. For example, "offer" is more likely to regarded as an aspect term. We use CRF to make a further judgment for avoiding these errors.

The contributions of this paper are summarized as follows:

(1) We propose a model to extract aspect terms in product reviews, which effectively combines the ability of CNN to extract local features with the ability of non-local attention to obtain global feature.
(2) We use a non-local attention mechanism to directly calculate the long-distance dependencies between two words in a review, which helps the extraction of long aspects terms.
(3) We analyze the classification performance of different tags, and focus on evaluating B and I tags. By analyzing the tags of different categories, we can clearly find the effect of extracting aspect terms.

2. RELATED WORK

Sentiment analysis has been extensive researched at the document level, sentence level, and aspect level [10-12]. This work focuses on sentiment analysis at the aspect level. Aspect terms extraction is one of the main sub-tasks of sentiment analysis. At present, aspect extraction of product reviews mainly includes two modes: aspect terms extraction and aspect-opinion terms co-extraction.

2.1 Aspect Terms Extraction

ATE only considers the opinion targets in reviews, regardless of the sentimental tendency. At present, there are mainly unsupervised methods and supervised methods for this research. Unsupervised methods include frequent pattern mining [13], the syntactic rule-based approach [1, 14], and topic-based methods [15-17]. The syntactic rule-based approach is to manually design some specific rules based on the syntax or dependency structure of reviews to extract aspect terms. Yin *et al.* [18] propose an unsupervised model

named WDEmb, which uses multiple embeddings to enhance CRFs. As for supervised methods, ATE is regarded as a sequence labeling task. In traditional sequential models, researchers pay more attention to CRF [5, 19] for jointly considering the adjacent words. But CRF can't effectively use semantic features of words and the long-distance dependencies between words. Besides, Long Short-Term Memory networks (LSTM) [7, 20] and support vector machines [21] are also used in aspect extraction tasks. Li *et al.* [22] use the aspect detection history to help to predict possible aspect terms at the current time. When there are multiple aspect terms in reviews, it can also be used to identify the following uncommon aspect terms. At the same time, they also use the opinion summary to strengthen the relationship of aspect extraction. Xu *et al.* [9] use both general embedding and domain specific embedding as inputs of the CNN model, and let the CNN model determine which embedding has more useful information. Experiments shows that for aspect extraction, the double embedding mechanism performs better than general domain embedding or domain-specific embedding. Shu *et al.* [23] use a modified CNN modified named Controlled CNN (Ctrl), which has two types of control modules to prevent overfitting through asynchronous parameter updating. Ctrl significantly improves the performance of ATE.

2.2 Aspect-Opinion Terms Co-Extraction

Some models consider the extraction of aspect and opinion terms at the same time. Wang *et al.* [24] propose a recurrent neural network-conditional random field (RNCRF). It is a joint model with a dependency tree based on recurrent neural network and CRF for aspect and opinion terms co-extraction. In addition to opinion annotations, it also uses handcrafted features. One assumption of RNCRF is that dependency resolution could capture the relationships between aspect terms and opinion terms in reviews, to achieve good results in co-extraction. This assumption is usually valid for simple reviews, but is more fragile for some complex structures such as clauses and parentheses. Furthermore, the network construction of RNCRF relies on the input dependency tree, and it has a dependency resolution error. Li and Lam [20] use two LSTMs (MIN) to extract aspect and opinion terms respectively, and they use an additional LSTM network to determine whether exist sentimental words in the review, so as to enhance the relevance of the aspect and opinion terms. Wang *et al.* [25] propose a multi-layer coupled-attention network (CMLA) that also performs aspect and opinion terms co-extraction. Luo *et al.* [26] use two RNNs to generate respective representations of aspect and opinion terms, and then use cross-shared units (DOER) to get interaction between ATE and aspect sentiment classification (ASC). Two auxiliary tasks are used to improve the aspect and opinion terms extraction.

3. THE PROPOSED MODEL

3.1 Task Description

A review is denoted as $S = \{w, w_2, \ldots, w_i, \ldots, w_n\}$, where n is the maximum sentence length. For each word $w_i \in S$, the task of ATE is to find the corresponding tag $t_i \in T$, where $T = \{B, I, O\}$. "B" and "I" represent the beginning word, inside word of aspect terms, and "O" represent the other word in the review. For example, "*Boot/B time/I is/O super/O fast/O, /O around/O anywhere/O from/O 35/O seconds/O to/O 1/O minute/O. /O*" is a tagged sentence, in which the aspect term is "Boot time".

3.2 The Non-local Attention Based CNN Model

We propose an ATE network that can capture long-distance dependencies between words by non-local attention mechanism. As shown in Fig. 1, the model framework consists of four parts: embedded layer, CNN layer, attention layer and CRF layer. Firstly, we input the combination of general word embedding and domain specific word embedding in the embedding layer, and extract features by convolution kernels of different sizes in the CNN layers. Then we use the attention layer to capture the long-distance dependencies between words. Finally, we transfer the extracted results to CRF to make the predicted tag sequence more reasonable.

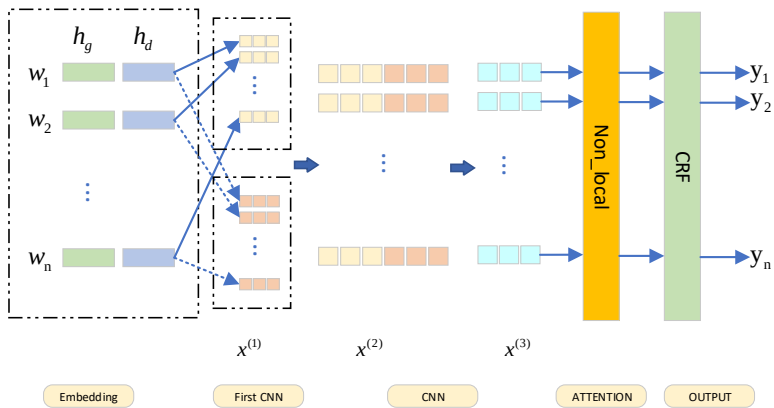


Fig. 1. The non-local attention based CNN model.

3.2.1 The embedding layer

Inspired by Xu *et al.* [9], we use double embedding as the initial input. Double embedding consists of two types: the generic word embedding and the domain word embedding, which differ in whether they were trained by a domain-specific corpus. For the sentence $S = \{w, w_2, \ldots, w_i, \ldots, w_n\}$, word w_i is initialized by a feature matrix $h_{\omega_i} \in \mathbb{R}^{(d_G+d_D)\times|v|}$ here d_G and d_D come from the general word vector embedding $G \in \mathbb{R}^{d_G\times|v|}$ and the specific domain word embedding $D \in \mathbb{R}^{d_D\times|v|}$, $|v|$ is the size of the vocabulary. The word vectors outside the vocabulary are randomly generated. For each sentence, we use 0 padding to align with the longest length n of all sentences.

3.2.2 The CNN layers

CNN can extract the local features of the data, so we use three one-dimensional CNN networks to learn local features. After obtaining the embedding matrix $x^{(l)}$, the shape of $x^{(l)}$ is $n * d$, where d is the sum of the double embedding dimensions. We adjust $x^{(l)}$ and transpose it to $d * n$. Each CNN layer has several one-dimensional convolution kernels. The size of the convolution kernel is $k = 2c + 1$. At the same time, we fill both ends of the output of each layer with 0 to align with the original input length for sequence labeling. The convolution operation is performed as follows:

$$x^{(l+1)} = \sum_{j=-c}^{c} w_j^{(l)} x^{(l)} + b^{(l)}. \quad (1)$$

Where l represents the number of the convolution layer. According to the convolution kernel of $k = 2c + 1$, we can calculate the relationship between the ith word and c context words. For the first layer of CNN, we use the convolution filters with two different sizes to obtain different views and concatenate them together. For the remaining two CNN layers, we use the convolution filters with the same size for feature extraction. At the same time, after convolution of each layer, the linear input data is transformed into nonlinear data by activation layer, which makes the data have a certain discrimination ability in the fitting process. In this model, ReLU is selected as the activation function. $x^{(3)} \in R^{b*h*w}$ represents the output of the third convolution layer in Fig. 1.

3.2.3 The non-local attention layer

The traditional CNN model is limited by the size of convolution kernel, which only considers the local effect on the elements in the convolution kernel. There are always problems of long-distance dependency between words in the ATE task. If we only consider the local relationship of elements, we would lose some useful features. Therefore, we should consider more non-local information for this task. We introduce the non-local attention mechanism proposed by the reference [27], which can obtain the correlation coefficient matrix of the original review. The correlation coefficient matrix represents the relationship between two words with long-distance. The overall structure of the non-local attention extraction layer is shown in Fig. 2.

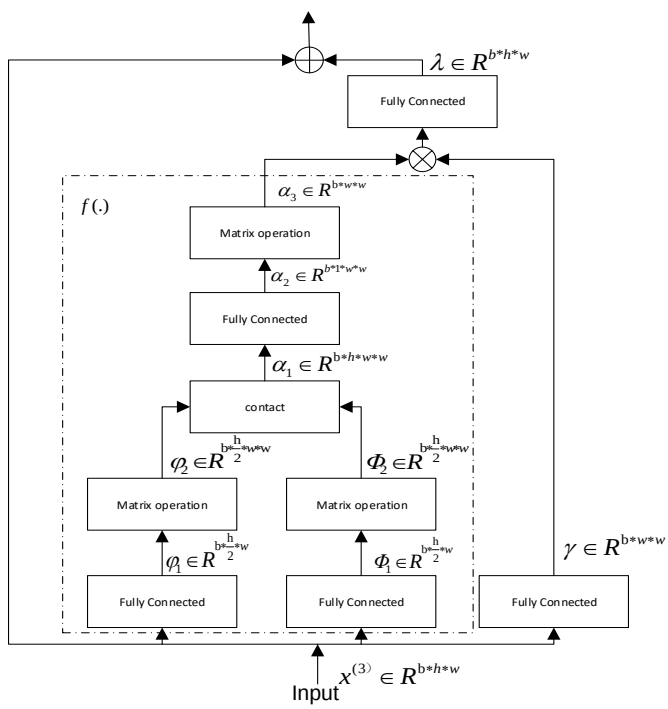


Fig. 2. Non-local attention mechanism.

The main formulas of non-local attention mechanism are as follows:

$$y_i = \frac{1}{D(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{2}$$

$$f(x_i, x_j) = ReLU(w_f^T [\varphi(x_i), \Phi(x_j)]). \tag{3}$$

Where $f(\cdot)$ denotes a function for calculating the correlation coefficient between current elements x_i and other elements x_j, $\varphi(x_i)$ and $\Phi(x_j)$ denotes the matrix transformation. $D(x)$ denotes normalized constant, and $g(x_j)$ denotes a linear layer. We obtain the non-local output by multiplying $f(\cdot)$ and $g(\cdot)$. The specific dimensions of other variables are shown in Fig. 2. Finally, we add the original input $x^{(3)}$ and the non-local output in order to strengthen the ability of the network fitting.

$$Y = y_i + x^{(3)} \tag{4}$$

3.2.4 The CRF layer

For ATE task, there are strong constraint relationships between the labels of adjacent words. CRF can influence the latter label based on the previous label when the training corpus is insufficient. So, we add the linear chain CRF to the attention layer for obtaining the global optimal labeling sequence. We convert the result Y of the attention layer into $P \in R^{n*m}$ by a fully connected layer, where m is the number of labels, then put P into the CRF layer. p_{ij} is defined as the probability of jth label of the ith word in the sentence. For a prediction sequence $y = \{y_1, y_2, \ldots, y_i \ldots, y_n\}$, its probability can be expressed as:

$$s(S, y) = \sum_{i=1}^{n} p_{i,y_i} + \sum_{i=1}^{n-1} A_{y_i, y_{i-1}}. \tag{5}$$

In Eq. (5), $A \in R^{m*m}$ is a transition probability matrix. In other words, A_{ij} represents the probability of being transferred from the label i to j. Therefore, the probability of getting the label y under the condition of the original statement S is as follows:

$$p(y \mid s) = \frac{e^{s(S,y)}}{\sum_{y' \in Y} e^{s(S,y)}}. \tag{6}$$

Where y' represents the true tag and Y is the set of all output sequences. The highest probability tag could be found by negatively minimizing maximum likelihood function:

$$loss(S, y^t) = -\log p(y^t|S). \tag{7}$$

Where y^t denotes predicted label.

4. EXPERIMENTS

4.1 Datasets

To be consistent with the work of Xu *et al.* [9], we conduct experiments on two benchmark datasets of the SemEval Challenge [28, 29] as shown in Table 1. The first dataset

derived from laptop domain on subtask 1 of SemEval-2014 Task 4, which has 2676 sentences. The second dataset derived from restaurant domain on subtask 1 of SemEval-2016 Task 5, which has 3,845 sentences.

Table 1. Statistics of the dataset.

Dataset	The Number of The Training Data		The Number of The Testing Data	
	Sentence/Aspect		Sentence/Aspect	
SemEval-14 Laptop	3045	2358	800	654
SemEval-16 Restaurant	2000	1743	676	622

4.2 Parameters Setting

For the proposed model, we hold out 150 training examples as validation data to decide parameters. The first CNN layer has 128 filters with a convolution kernel size of k = 3, 5. The second CNN layer has 256 filters with a convolution kernel size of k = 5. To reduce the computational complexity of the attention layer, the last CNN layer has 128 filters with a convolution kernel size of k = 5, and we set step size to be 1. The feature number of the input and output of the attention layer is 128. The training batch size is 128 and the dropout rate is 0.55. The learning rate of Adam optimizer [30] is 0.0001 because CNN training tends to be unstable.

4.3 The Baseline Models

In order to verify the effectiveness of the proposed model, we select the representative models with three groups for comparison as follows.

1. The first group of models uses CRF and different embeddings as input.
   CRF is conditional random fields with word embedding and full connection layer. It can jointly consider the adjacent words.
   IHS_RD [5] is a model based on the IHS Goldfire language processor. They use a wealth of vocabulary, syntax and statistics in the CRF model. It is a winning system of the ATE subtask in SemEval ABSA challenge.
   WDEmb [18] proposes enhanced CRF with word embedding, linear context embedding, and dependency path embedding as input.
2. The second group of models uses multi-task learning. Except the aspect extraction, the sentiment polarity of the aspect term is also considered.
   RNCRF [24] is a joint model with dependency tree learned from recursive neural network and CRF to extract aspect and opinion terms together. In addition to word tagging, it also uses artificial features.
   MIN [20] uses two LSTMs to extract aspect and opinion terms. It uses a LSTM network to determine whether existed sentimental words in the reviews, because aspect and opinion terms often appear with sentimental words.
   CMLA [26] uses a multi-layer attention network to jointly extract aspect and opinion terms, each of which consists of several attention of tensor operators. A kind of attention

is used to extract aspect, while the other is used to extract opinion terms.

DOER [27] proposes a novel Dual cross-shared RNN framework to simultaneously generate aspect-opinion pairs of input sentences. The DOER uses two bidirectional recursive neural networks to extract the respective representations of each task, and performs a cross-shared unit to consider relationship between them.

3. The third group of models uses neural networks as the basis and can capture more useful features from reviews.

CNN uses different convolutional kernel size to calculate the relationship between *i*th word and context words. It is one of the most basic models used for sequence labeling tasks.

Bert base [31] uses pre-trained BERT with linear layers. We use an open source model to achieve this task, and use the pre-trained "BERT-BASE-UNCASED" model to initialize parameters.

DE-CNN [9] proposes general embedding combined with specific-domain embedding as input, and uses CNN to extract aspect with excellent effect.

LSTM uses Long Short-Term Memory networks for ATE. It is one of the most basic models used for sequence labeling tasks.

DE-LSTM uses double embedding as input the same as DE-CNN.

THA&STN [22] uses aspect detection history to help predict the possible aspect in the current time. When multiple aspect terms appear in a review, detection history can also be used to identify the following uncommon aspect.

Ctrl [23] uses a modified CNN called a controlled CNN (Ctrl) for supervised aspect extraction. The improved CNN has two types of control modules.

DE-SAN uses double embedding, self-attention mechanism (SAN) [32] and CRF.

4.4 Results and Analysis

In this section, we firstly show the experimental results of the models on two datasets. Next, in order to analyze the role of different modules of our model, we give the experimental results of ablation analysis, and finally explore the advantages and disadvantages of the models.

4.4.1 Contrast of the baseline models

Table 2 shows the experimental results of the baseline models and our proposed model on the Laptop and Restaurant datasets, using the F1 score as the evaluation criteria.

In the first group of models of Table 2, HIS_RD uses the combination of vocabulary, syntax, and statistics. From Table 2, F1 score of HIS_RD improves 0.44% on Laptop dataset compared with CRF. WDEmb adds dependency path embedding and improves 0.61% compared with HIS_RD on Laptop dataset.

The second group of models in Table 2 focuses on the aspect-opinion terms co-extraction. RNCRF used CRF and recursive neural networks. A wealth of artificial rules can help the RNCRF model to find the relationship between aspect terms and opinion terms, and improves 3.26% compared with WDEmb on laptop dataset. MIN uses a vanilla LSTM to predict the sentimental of the sentence as additional guidance, which can strengthen the relationship between aspect terms and opinion terms and improves 2.42% compared with

Table 2. F1 scores of different models.

Model	Laptop	Restaurant
CRF	74.01	69.56
HIS_RD	74.55	–
WDEmb	75.16	–
RNCRF	78.42	69.72
MIN	77.58	73.44
CMLA	77.80	72.77
DOER	82.61	–
LSTM	75.25	71.26
CNN	77.67	72.08
THA&STN	79.52	73.61
DE-LSTM	78.73	72.94
Bert base	79.34	74.58
DE-CNN	81.59	74.37
Ctrl	82.73	75.64
DE-SAN	83.71	75.23
Our Model	84.10	76.07

Comparison results in F1 score: numbers in our model are the average score of 5 runs.

WDEmb on restaurant dataset. For the first time, CMLA use attention mechanisms in this task. One attention is for extracting aspect terms, while the other is for extracting opinion terms. Two attentions learn interactively to dually propagate information between aspect terms and opinion terms, and the CMLA improves 2.64% compared with WDEmb on restaurant dataset. DOER designs a cross-sharing unit to consider the relationship between aspect and opinion terms. The aspect-opinion terms co-extraction task requires additional annotation data compared with ATE. The ATE task and ASC task should be mutually reinforcing and correctly capture the relationship between them. At the same time, the sentimental expression in reviews has an impact on the extraction of aspect terms (some reviews do not contain aspect terms, but contain sentimental words).

ATE is regarded as a sequence labeling task, and the basic sequence labeling models are usually based on LSTM or CNN. The results in Table 2 show that CNN has better performance on ATE than LSTM, and the same result is also obtained in the experiment of double embeddings. THA&STN uses opinion summary information to strengthen the relationship with aspect terms. At the same time, it uses the historical information of aspect detection to help aspect prediction at the current time, which is better for detecting multiple aspects in the same reviews. THA&STN improves 2-5% on Laptop dataset compared with CNN and LSTM models. As an advanced pre-training word vector model, Bert base can represent the syntactic and semantic information in different contexts to learn better word representations in different domain. The experimental result shows Bert base achieves 0.91% improvement compared to THA&STN. DE-CNN uses the combination of domain word embedding and general word embedding as the input. By eliminating the differences between the domains, DE-CNN achieves 2.07% improvement compared to THA&STN on Laptop dataset. However, it still has the defect of insufficient acquisition of long-distance dependencies between words. Ctrl uses two control modules based on DE-CNN. It can

effectively prevent the network from overfitting by asynchronous updating. Compared with DE-CNN, it has improved 1.14% on the Laptop dataset and achieves the best effect of extracting aspect independently. It should be noted that none of these models focus on long aspect terms.

By introducing non-local attention to our model, we can see that our model is always better than DE-CNN and Ctrl. CNN completely relies on the convolution kernel to capture the long-distance dependencies between words. In contrast, non-local attention neural network can directly calculate the correlation coefficient of two positions to express their dependence, thus the model can obtain more abundant dependence between words. This kind of long-distance dependencies between words is very important for the extraction of long aspect terms. The experimental results show that the F1 scores of our model are 2.51% and 1.7% higher than DE-CNN on laptop and restaurant datasets.

In addition, we also did a comparative experiment of non-local attention and self-attention. When we use the self-attention after the CNN layer, the result is slightly lower than that of non-local attention. Self-attention uses the dot product method to learn simple relationships. In the contrast, non-local expands the dimensions of word representation, and uses different matrix transformations to learn different relationships between words.

4.4.2 Ablation analysis

To further investigate the performance of each module in our proposed method, we designed two additional experiments.

CNN + CRF is the proposed model without the non-local module, which is composed by the three CNN layers and a CRF layer.

CNN + Non-local is the proposed model without the CRF layer, which is composed by the three CNN layers and the non-local attention layer.

Experimental results on laptop dataset are shown in Table 3.

Table 3. F1 scores of ablation models.

Model	Laptop	Restaurant
DE-CNN	81.59	74.37
CNN + CRF	81.10	73.93
CNN + Non-local	83.39	75.35

Numbers in the 3 groups are the average score of 5 runs.

Experimental results show that the average F1 score of "CNN + CRF" is 0.49% lower than DE-CNN. But this is not always the case. It is found in experiments that occasionally (random initialization is better), using only CRF will also get good results (up to F1 83.11%, Table 3 shows the average of 5 runs). This may be that for some reviews, although the model can predict some words in a long aspect term correctly, it cannot predict the integral aspect term. We will analyze it later. When only using three CNN layers and non-local attention, the improvement is very obvious, indicating that the long-distance dependencies relationship between words are effective for the extraction of aspect terms. We consider that some terms far away from the aspect may get more attention because of the calculation of long-distance dependencies, and they are likely to be mistaken as a separate aspect.

These words are usually adjectives with obvious sentimental polarity or verbs with a syntactic relationship with aspect terms. Our model further considers this problem and achieves the best results.

In order to more intuitively compare the functions of the various modules of our model, we measure performance of models by using Precision(P), Recall(R) and F1 value of the different labels on the laptop dataset as evacuation indicators, as shown in Table 4. The results of these detailed labels can help us to further understand the effect of different modules.

Table 4. P, R, F1 values of B I O label.

Label		DE-CNN	CNN + CRF	CNN + Non-local	Our Model
B	P	82.84	**86.12**	82.27	84.32
	R	83.18	78.75	84.40	**84.71**
	F1	82.80	82.27	83.32	**84.51**
I	P	80.80	**81.97**	79.90	80.53
	R	65.89	67.99	**71.50**	**71.50**
	F1	72.59	74.33	75.47	**75.75**
O	P	99.68	99.64	99.72	99.72
	R	99.79	99.82	99.77	99.79
	F1	99.73	99.73	99.74	99.75
Overall	P	87.63	**89.24**	87.30	88.19
	R	82.95	82.19	85.22	**85.33**
	F1	85.23	85.57	86.25	**86.74**

Results of data selection closest to the average value.

CRF can consider the labels of adjacent words to give correct results, thus the precision of *BIO* is improved. For example, CRF can rectify "graphic/O design/B" to "graphic/B design/I", and "new/B OS/B" to "new/B OS/I". However, in strict evaluation criteria (for aspect term), CRF does not always work. This is because the CRF often judges it as *O* when the network has insufficient inter-relationship or feature capture between words. It is also the reason that although the *B* tag (mostly a single aspect term in the experiment) has an obvious increase on precision, the recall rate drops significantly.

When only the non-local attention mechanism is used, it can be seen that the recall rate of tag *I* has improved significantly. Compared with DE-CNN, the recall rate of tag *I* significantly increases by 5.61%. Simultaneously, F1 score of the whole results increases by 1.8%. This part of the improvement comes from the extraction of long aspect terms. After the introduction of non-local attention, the relationship between different words can be calculated, and the range of aspect terms can be better distinguished, so that more long aspect terms can be predicted. At the same time, the recall of the tag *B* increases by 1.22%, which is due to the introduction of long-distance dependencies between words, and it enables the network to find some aspect terms that contain only one word and far away from other aspect terms. At the same time, we find the model pays much attention to some adjectives or verbs after the introduction of non-local attention and regards them as aspect terms incorrectly, such as "implicit", "offers", "works", and so on. These words are directly related to aspect terms. In order to solve this problem, we use CRF to make further estimation, which can solve the obvious errors. The precision of the tag *B* has increased by 2.05% and the precision of the tag *I* increases by 0.63% after using CRF layer.

4.4.3 Case study

In this section, we give an example that can reflect the effects of different modules of our model, as shown in Table 5.

Table 5. A review whose labels of the aspect term are marked bold.

Review	*I opted for the **SquareTrade 3-Year Computer Accidental Protection Warranty** ($1500-2000) ...*
Label	*O O O O **B** **I** **I** **I** **I** **I** O O O O O O...*
DE-CNN	*O O O O **B** **I** **I** **I** **O** **I** **I** O O O O...*
CNN+CRF	*O O O O **O** **I** **I** **I** **I** **I** **I** O O O O...*
CNN+Non-local	*O O O O **B** **I** **B** **I** **I** **I** O O O O O O...*
Our Model	*O O O O **B** **I** **I** **I** **I** **I** O O O O O O...*

As shown in Table 5, when we only use a single CNN layer, the extraction of such a long aspect term is very difficult, and there are obvious errors. After the CRF is used, the model fixes the internal labels by considering the front and back labels. At the same time, feature capturing of "SquareTrade" is insufficient to misjudge it as *B*. When non-local attention is used, the model can better capture the long-distance dependence, thereby correctly obtain the range of the aspect term. At the same time, the problem of incorrect internal labeling still occurs (regarding the label of "Computer" as *B*). As for our model, the correct result is obtained by using CRF in the last layer.

5. CONCLUSIONS

In this article, we introduce a model for aspect extraction using a non-local attention mechanism. Non-local attention mechanism has been used to focus on extracting the interdependence between words for improving the extraction of long aspect term. At the same time, non-local attention may bring out the model pay excessive attention to some adjectives and verbs. We used CRF to adjust this deviation. Experimental results on two benchmark datasets validate the effectiveness of our model and show that our model is significantly better than the baselines in terms of aspect term extraction.

REFERENCES

1. G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational Linguistics*, Vol. 37, 2011, pp. 9-27.
2. S. Poria, E. Cambria, L.-W. Ku, C. Gui, and A. Gelbukh, "A rule-based approach to aspect extraction from product reviews," in *Proceedings of the 2nd Workshop on Natural Language Processing for Social Media*, 2014, pp. 28-37.

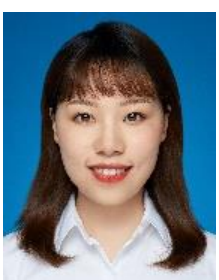
3. S. Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 167-176.
4. N. Jakob and I. Gurevych, "Extracting opinion targets in a single- and cross-domain setting with conditional random fields," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 1035-1045.
5. M. Chernyshevich, "Ihs R&D Belarus: Cross-domain extraction of product features using CRF," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 309-313.
6. L. Shu, H. Xu, and B. Liu, "Lifelong learning crf for supervised aspect extraction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vol. 2, 2017, pp. 148-154.
7. P. Liu, S. Joty, and H. Meng, "Fine grained opinion mining with recurrent neural networks and word embeddings," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1433-1443.
8. T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Systems with Applications*, Vol. 72, 2017, pp. 221-230.
9. H. Xu, B. Liu, L. Shu, and P. S. Yu, "Double embeddings and CNN-based sequence labeling for aspect extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 592-598.
10. R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, 2013, Vol. 56, 2013, pp. 82-89.
11. A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *Proceedings of International Conference on Collaboration Technologies and Systems*, 2012, pp. 546-550.
12. T. T. Thet, J.-C. Na, and C. S. G. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *Journal of Information Science*, Vol. 36, 2010, pp. 823-848.
13. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp 168-177.
14. Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "Automated rule selection for aspect extraction in opinion mining," in *Proceedings of International Joint Conferences on Artificial Intelligence Organization*, 2015, pp. 1291-1297.
15. Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 171-180.
16. Y. He, C. Lin, and H. Alani, "Automatically extracting polarity-bearing topics for cross-domain sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 123-131.
17. Z. Chen, A. Mukherjee, and B. Liu. "Aspect extraction with automated prior knowledge learning," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, pp. 347-358.

18. Y. Yin, F. Wei, L. Dong, K. Xu, M. Zhang, and M. Zhou, "Unsupervised word and dependency path embeddings for aspect term extraction," in *Proceedings of International Joint Conferences on Artificial Intelligence Organization*, 2016, pp. 2979-2985.
19. Z. Toh and W. Wang, "Dlirec: Aspect term extraction and term polarity classification system," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 235-240.
20. X. Li and W. Lam, "Deep multi-task learning for aspect term extraction with memory interaction," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2017. pp. 2886-2892.
21. A. S Manek, P. D. Shenoy, M. C. Mohan, and K. R. Venugopal, "Aspect term extraction for sentiment analysis in large movie reviews using Gini index feature selection method and SVM classifier," *World Wide Web Journal*, Vol. 20, 2016, pp. 135-154.
22. X. Li, L. Bing, P. Li, W. Lam, and Z. Yang, "Aspect term extraction with history attention and selective transformation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence Main Track*, 2018, pp. 4194-4200.
23. L. Shu, H. Xu, and B. Liu, "Controlled CNN-based sequence labeling for aspect extraction," *arXiv Preprint*, 2019, arXiv:1905.06407 [cs.CL].
24. W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Recursive neural conditional random fields for aspect-based sentiment analysis," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 616-626.
25. W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, "Coupled multi-layer attentions for co-extraction of aspect and opinion terms," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 3316-3322.
26. H. Luo, T. Li, B. Liu, and J. Zhang, "DOER: Dual cross-shared RNN for aspect term-polarity co-extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 591-601.
27. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794-7803.
28. M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. de Clercq, *et al.* "Semeval-2016 task 5: Aspect based sentiment analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation*, 2016, pp. 19-30.
29. M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 27-35.
30. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations*, 2015.
31. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171-4186.
32. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *arXiv Preprint*, 2017, arXiv:1706.03762.

Dang-Guo Shao (邵党国) received the Ph.D. degree in Sichuan University and is now an Associate Professor. His main research fields include data mining and NLP. E-mail: huntersdg@163.com

Ming-Fang Zhang (张名芳) received the B.S. degree in Henan Polytechnic University an is now a Postgraduate Students. His main research fields include text mining and NLP. E-mail: 1254116691@qq.com

Yan Xiang (相艳), Ph.D. Student, Associate Professor, her main research fields include text mining and NLP. E-mail: 50691012@ qq.com



Ting-Lu (陆婷) received the B.S. degree in Kinning University of Science and Technology in 1996 and is now a Postgraduate student. Her main research fields include text mining and natural language processing. E-mail: 1530584820@qq.com

Rong Hu (胡蓉) received the M.S. degree in Tsinghua University in 1974 and is now an Associate Professor. Her main research fields include optimization methods and decision support systems. E-mail: ronghu@vip.163.com