Sentiment Analysis of Bullet-screen Comments Based on AT-LSTM

Xuqiang Zhuang [a], Fangai Liu [b]

School of Information Science and Engineering, Shandong Normal University, Jinan, 250014, China

[a] zhuangxq@sdnu.edu.cn, [b] liufangaisdnu@163.com

Abstract

Bullet-screen Comments can be more accurately and concretely reflect the user's real-time emotion and evaluation when they are watching the video, therefore, we present a AT-LSTM emotion analysis model based on attention model to mine these emotional information. First of all, through the attention model,we can better extract the emotional keywords in the whole bullet-screen comments.At the same time, LSTM model can more effectively combine the emotional dependency relationship between the front and rear bullet-screen comments in the video, and extract the theme based "highlight" video shots. The experimental results show that the proposed method is more accurate than the traditional LDA and LSTM methods. Our model can help users more accurately obtain the emotional information contained in the online bullet-screen video,and then proceeding to provide a new way of video search and video recommendation.

Keywords

Deep learning, bullet-screen comments, sentiment analysis, AT-LSTM.

1. Introduction

Bullet-screen video such as Acfun and Bilibili is a kind of video mode appeared in recent years that combines the comment of the users with online videos. Users can express the comments along the timeline, or so-called bullet-screen comments,will appear as time-sync comments on videos in real time.

Bullet-screen comments contains comment texts and their corresponding video point-in-time information. when compared with the general comments, bullet-screen comments can more accurately and concretely reflect the user's real-time emotion and appraisive evaluation. These emotions and ratings information can provide a reference for other users when choosing a video. Using the sentiment analysis technology to extract sentiment information from the network bullet-screen video, it can help users get the overall emotional orientation of the video comment text, as well as the changing status of comment sentiment over time. With the promotion of bullet-screen feature in the major mainstream video sites, the opinion and emotional expression in the bullet-screen comments will be more universal and referential. On the basis of bullet-screen comments' sentiment analysis, we can create a new way of video retrieval based on the emotional comments to meet the more diversified and personalized search needs. Based on the above analysis, this paper attempts to explore the sentiment analysis of online video.

At present, more and more scholars begin to conduct academic research on videos based on video with bullet-screen comments. For example, automatic tagging technology based on bullet-screen videos, it provides a technology that extracts the bullet-screen keyword from the video shots, and then tags this video shots[1]. Detection method of video highlights shots basing on the bullet-screen comments, mainly analyses spontaneous moments in the video according to the text of bullet-screen and the changing curve of quantity[2].combines collaborative filtering and LSTM network, uses all historical implied feedback as well as users' interest characteristics and bullet-screen comments to recommend video keyframes[3]. Using the influence of uploader and video quality combined with the dynamic herd effect, a prediction model of bullet-screen video's popularity was proposed[4]. In

fact, previous researches often fail to show the emotional trends in the highlights of the video: First of all, the wonderful video shots appear in the intensive bullet-screen comments areas, however, the areas where the bullet-screen comments are intensive are not all wonderful video shots; Second, the video tags extracted from the text of bullet-screen comments usually belong to the keyword and high frequency words in the bullet-screen, so referring to the emotion of video shots by video tags is not accurate.

To this end,as a pilot study, we designed an Attention-based LSTM (AT-LSTM) sentiment analysis model to recommend the highlights video shots:

1)Effectively analyze the anterior and posterior relevance of the bullet-screen comments in the video, thus more accurate access to the theme information bullet-screen comments.

2)We highlighted the impact of key emotional words in bullet-screen comments by adding attention model, further improving the accuracy of the model.

3)According to the topic model, the subjects in the bullet-screen are classified into thematic categories and the emotional similarity between the video shots is calculated, get wonderful video shots based on theme distribution.

2. RELATED WORKS

2.1 Definition of Bullet-screen Comments

Bullet-screen Comments: Bullet-screen comments can be defined as a triplet, which contains users' input, delivery time, user ID. We mainly analyzes the content of the three types of bullet-screen comments in Chinese, English and numbers. Of which, English bullet-screen comments contain part of the Internet language, such as "QAQ" (for cry) and so on; the number section contains common network buzzwords, such as "233" (for laughter) and "666" (for amazing) and so on.

As shown in Fig 1, the texts above the bullet-screen video shot are the bullet-screen comments sent by users.

Fig. 1 Video example of bullet-screen on bilibili

2.2 Highlights video shots.

"Highlights" video shots: The purpose of this paper is to find the "highlights" video shots that highlight the emotional appeal of the bullet-screen video. The so-called "highlights" video shots, ie the highlights of the video, refers to that the users have a strong emotional color discussion with part of the video content. Through the study we found that bullet-screen comments contains a certain "time dependence",that is, when users send bullet-screen comments, they can refer to the current and previous bullet-screen comments. Therefore, adjacent comments may be similar in semantic vector. Based on this phenomenon, we are in a certain area around the bullet-screen comments (the length of the video shots is), then we call the video shots of comments with a certain theme "highlights" video shots.

2.3 LSTM

LSTM (Long Short Term Memory) is an effective chained recurrent neural network (RNN), is widely used in language models, machine translation, speech recognition and other fields.

LSTM consists of intput gate i, output gate o and forget ate f and memory cell c, of which the intput gate, output gate and forget gate are the controller of controlling memory cell reading, writing and losing operation. c_t represents the calculation method of memory cell at time t, h_t is the output of the LSTM cell at time t.See Fig 2.

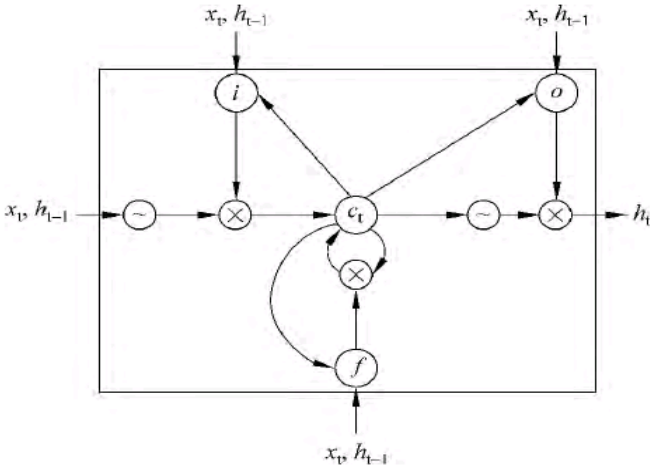

Fig. 2 LSTM model

LSTM consists of input sequence $X = (X_1, X_2, \cdots, X_n)$, implicit vector sequences $h = (h_1, h_2, \cdots h_n)$ and output vector sequences $y = (y_1, y_2, \cdots, y_n)$. At each time increment, the output of the LSTM is controlled by a set of gate functions, the set of gate functions is a input function that consists of a previously hidden state h_{t-1} and inputs at the current time increment x_t and intput gate, output gate, and forget gate. These gate functions together determine the conversion of current memory units and the current hidden state. The LSTM conversion function is defined as follows:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ f_t &= \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \\ l_t &= \tanh(W_l \cdot [h_{t-1}, x_t] + b_l) \end{aligned} \tag{1}$$

$$\begin{aligned} o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot l_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned}$$

Where σ is the sigmoid function with output in [0,1], tanh represents the hyperbolic tangent function with output in [-1,1], the backend is based on the output of the memory unit to control how much information in the old memory unit is discarded, controls how much new information is stored in the current memory location at the same time. Now that LSTM is specifically designed to learn the tasks of long-term dependencies, we chose LSTM to deal with the contextual correlation in bullet-screen comments. During sequence-to-sequence generation, LSTM defines the distribution on the output and uses the softmax function to sequentially predict the output:

$$\mathbf{P(Y|X)} = \prod_{t\in[1,N]} \frac{\exp(g(h_{t-1}, y_t))}{\sum_Y \exp(g(h_{t-1}, Y_t))} \tag{2}$$

In equation (2), g is the activation function. For simplicity, we define x_t, h_{t-1} to represent the LSTM operation on the input x time step t and the previously hidden state h_{t-1}.

2.4 AT-LSTM

The standard LSTM can not detect important emotional commentary in the bullet-screen comments. To solve this problem, we have established a attention model, which can capture the key words of the bullet-screen comments.

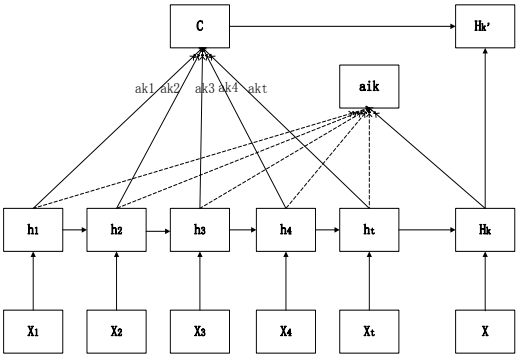

Fig. 3 Attention-based LSTM

The text input sequence $\{x_1, x_2, x_3, \ldots, x_t\}$ is the input sequence as a history node, the cumulative sum of input nodes of historical nodes, get the comment's overall input vector X', X' is the last input for the encoding stage. $\{h_1, h_2, h_3, \ldots, h_t\}$ corresponds to the hidden layer state value of input sequence $\{x_1, x_2, x_3, \ldots, x_t\}$. H_t corresponds to the hidden layer state value of input X'.See Fig 3.

ai in the graph is the attention probability of the historical node to the last node. The larger Ai indicates the more important the emotional sentiment of the word in the bullet-screen comments. The input sequence $\{x_1, x_2, x_3, \ldots, x_t\}$ in the figure is a word representation of the bullet-screen comments, X' is the overall input vector representation of the bullet-screen comments. Calculate the weights of $\{x_1, x_2, x_3, \ldots, x_t\}$ on the overall impact of the bullet-screen comments, can highlight the role of keywords, reduce the impact of non-keywords on the overall semantic of texts.

$$a_{ki} = \frac{\exp(e_{ki})}{\sum_{j=1}^{T} \exp(e_{kj})}$$

$$e_{ki} = v \tanh(Wh_k + Uh_i + b) \tag{3}$$

Which a_{ki} represents the attention probability weight of node i for node K. T is the number of elements of the input sequence. V, W, U represents the weight matrix, h_k is the last input corresponding hidden layer state. h_i represents the state value of the hidden layer corresponding to the i th element of the input sequence. The main idea is to calculate the relationship score between the historical node and the last input node; by this formula we get the attentional probability of each input for the last input.

Calculate semantic coding and eigenvectors of attention distribution probabilities. Calculation formula:

$$C = \sum_{i=1}^{T} a_{ki} h_i$$

$$H_{k'} = H(C, h_k, X') \tag{4}$$

The semantic coding C is mainly obtained by accumulating the product of the attention probability weight and the hidden layer state of the historical input node. The final semantic coding is to take the semantic coding of the probabilistic distribution of historical nodes and the overall vector of the comments as the input of the traditional LSTM module, and then the last node's hidden layer state value H_k is the final eigenvector. This feature vector contains the weight information of the historical input node, highlights the key words of semantic information.

3. SENTIMENT ANALYSIS MODEL BASED ON AT-LSTM

In this section, we will discuss how to perform emotional analysis of bullet-screen comments by using the AT-LSTM model. Specifically, we will first introduce the overall AT-LSTM model architecture, and then explain the details of model learning. See Fig 4.

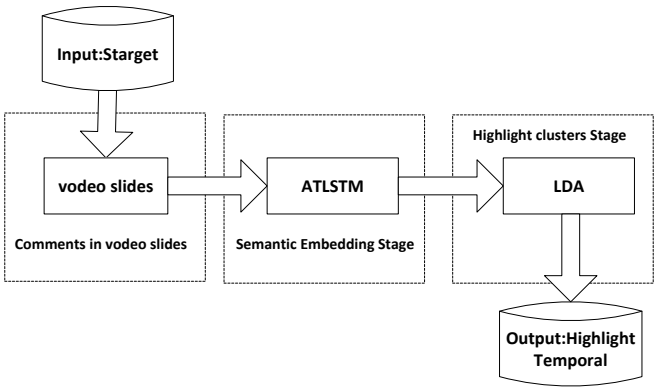

Fig. 4 Emotion Analysis Model Based on AT-LSTM

3.1 Sentiment Analysis Model based on AT-LSTM

Word vector representation part: Mainly using Word Embedding language model to obtain the word vector. Word vectors generated by the Word Embedding language model have a richer semantic representation. The feature of the word vector is that synonyms or approximations have similar vector representations. Specifically, the idea is that each term of the pre-processed bullet-screen comment data is trained into a K-dimensional vector using the word2vec tool. Get the input sequence $x = \{x_1, x_2, x_3, \dots, x_t\}$, where x_t is the K-dimensional input vector. Then through the vector of the cumulative average, get the overall input vector of the comments representation, the same is also a K-dimensional vector.

Feature Extraction Part:Using AT-LSTM network model as a coding model, from the foregoing, we know that AT-LSTM can effectively make use of the input sequence of historical information. This model uses the sum of equation (1) as the input sequence. Of which X' is the last input. As shown in above figure: $\{h_1, h_2, h_3, \dots, h_t\}$ corresponds to the hidden layer state value of input sequence $x = \{x_1, x_2, x_3, \dots, x_t\}$. H_t corresponds to the hidden layer state value of input X'. The AT-LSTM model designed in this paper semantically encodes the accumulation of the product of the attention probability weight and the hidden layer state of the historical input node, the final eigenvector is calculated.

3.2 Analysis of emotional similarity

3.2.1 Preparations

The temporal tag contains not only the type information, but also the temporal range in the video. We first set up a temporal window of “seconds” and divide the video into video shots. For each video shots, we consider it as the basic unit and extract its time tag.

We tend to highlight video shots with annotation topics, because we can get more "highlights" video shots, and we segment the segmented video fragments into thematic clusters. Therefore, we annotate each comment with the corresponding topic. For each video fragment, we can simply calculate whether a topic exists in the comment and its frequency is expressed as f. And if we have a fragment with a subject, the video shots can be recognized as a "highlights" video shot.

If the video segment S is recognized as a "highlights" video shot, it can be used to represent the temporal range of the fragment, represent the subject frequency of the comment. In a topic, the smaller weight of the word, the lower the correlation between the word and the topic. So we only consider the valid words for each topic.

For all video shots to find the maximum value and minimum value corresponding to the subject frequency f, we can set a threshold $\theta = \alpha * \min + (1 - \alpha) * \max$ to calculate whether the fragment is a video fragment (where α is called the qualified rate and $0 \leq \alpha \leq 1$).

3.2.2 Topic Clustering

We deal with "high light" video shots in a supervised way. A training set consisting of bullet-screen comments and a series of existing tags will be divided into equal length video segments, and then the subject frequency of each segment is calculated. In this paper, we use the LDA algorithm as a classifier, and we can get the topic set by clustering each topic fragment.

3.2.3 Theme merged

Finally, for any adjacent segment$< t_{s1}, t_{e1}, T_1 >$and $< t_{s2}, t_{e2}, T_2 >$, We will merge the two shots to get a new $S_{highlight}$ shots $< t_{s1}, t_{e2}, T_1 >$. After combining all the fragments, we got the final set of "highlights" video shots.

4. Experiments

4.1 Experimental Data

The experimental data in our model are composed of videos and bullet-screens comments downloaded from the domestic bullet-screen's website Bilibili. Data include different types of bullet-screen video, filtering out less than 40 bullet-screen videos, each comment contains comments text and comment time. We uses NLPIR as Chinese segmentation tool, after segmentation, cleaning and denoising to bullet-screen comments, setting time window m=100s, split the video into video shots, and then get 1600 video shots and 132850 bullet screen comments, and half of them will be randomly trained data C_{train}, half as the test data C_{test}.

4.2 Experimental setup

In this first step of experiment, we choose LDA and LSTM algorithm as the contrast algorithm. Table 1 shows comparison of experimental results between the model and the other methods. The experimental procedure includes the following steps:

Segment all videos and compute the emotion vectors for each video segment separately.

Using AT-LSTM model to analyze the emotional similarity of video shots. The semantic feature dimension of output is 50, and the dropout strategy is adopted in the training process, and the value of dropout is 0.5.

Using LDA algorithm to calculate the emotional similarity of each video shots, and take video shots with the highest score of N_{top} into the list of recommendations.

4.3 Evaluation index

Next,we use three criteria as model performance evaluation indexes. The indexes are calculated as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{EP}} \\ \text{recall} &= \frac{\text{TP}}{\text{TP} + \text{EP}} \\ \text{mean}F_1 &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \end{aligned} \tag{5}$$

Among them, TP represents the number of correct judgments of output, and (TP+EP) represents the number of all relations of the output, (TP+EP) representing the number C_{train} of all the relations in the test set.

4.4 Parameters Setting

We mainly studies four main parameters in our model:

$\mathbf{N_t}$, the number of iteration times in AT-LSTM model.

$\mathbf{N_{top}}$, the number of implicit topics in LDA process.

α, the number of iterations of theme sampling.

P, the number of valid words in the topic.

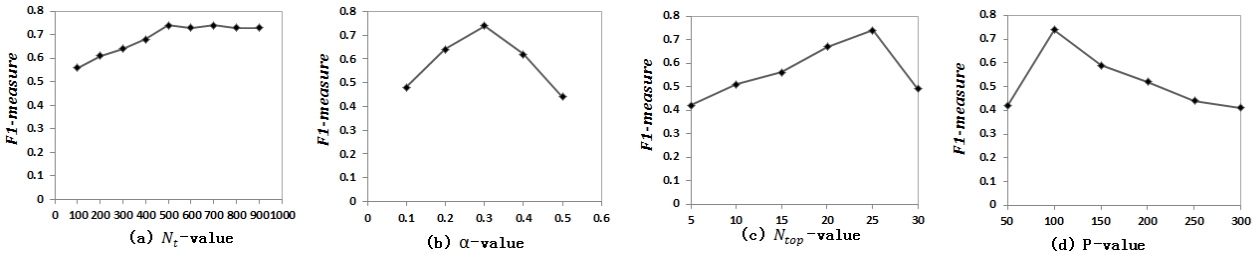

Fig. 5 Parameters Setting

The parameter learning result is shown in Fig 5. When the iteration times N_t of AT-LSTM is 500, the model achieves the best effect as shown in Fig 5(a).

For the qualification rate α for identifying "highlights" video shots are shown in figure 5(b), we set the α from 0.1 to 0.5, and the results are shown in the picture. It can be seen that with the increase of α from 0.1 to 0.5, Precision decreases and Recall increases. Our framework can produce more "highlight" video shots when the α is relatively small, and when the α is set to 0.3, our model can achieve the best performance on the F1 scores.

When the number of themes N_{top} is about 25, the optimal value of the F1 scores is reached. When N_{top}>25, it begins to drop rapidly.when N_{top}<25, the probability of video shots with different styles to the same theme will increase, which also fails to give full play to the advantage of the theme model in emotion recognition. The parameter learning results are shown in Fig 5(c).

When P take about 100, to reach the optimal value, the use of more quantity of words to participate in the calculation will reduce the accuracy of the theme emotion as shown in Fig 5(d).

4.5 Experimental Results

The experimental results of all models are shown in the graph. It can be seen that the experimental results of LDA and ordinary LSTM without attention model are not ideal. In the experiment, the optimal AT-LSTM model is obtained at N_{top} is 25 and the best result of AT-LSTM model is 0.741, AT-LSTM greatly increased compared with LDA and LSTM as shown in Table 1. When N_{top}<25 or N_{top}>25, the difference between their results is not big. LSTM is better than LDA and LSTM in three different values of N_{top}. The main reason for this is that LDA calculates the similarity between shots according to the topic distribution of video shots, however there are many on-line language in bullet-screen comments. Words with high appearing probabilities in every video shots may have greater weight in different themes, thus the effect of the topic model in emotion recognition is reduced.

Table 1. Experiment results under different setting of N_{top}

Model	N_{top}=10	N_{top}=15	N_{top}=20	N_{top}=25	N_{top}=30
LDA	0.483	0.495	0.504	0.507	0.471
LSTM	0.497	0.512	0.587	0.672	0.487
AT-LSTM	0.518	0.547	0.653	0.741	0.506

Through the above experimental analysis, it can be found that the recommendation method of video shots based on AT-LSTM is better than that based on LDA and LSTM as shown in Table 2. The main reason is that it combined with the attention model, highlighted the influence to the model by

LSTM critical input, and obtained more semantic information of pre vs. post correlative bullet-screen comments, accordingly proved this method has the advantage of precision in emotional analysis of Bullet-screen comments.

Table 2. Experiment results under setting of N_{top}=25

Model	Precision	Recall	F1-Score
LDA	0.455	0.571	0.507
LSTM	0.661	0.683	0.672
AT-LSTM	0.757	0.726	0.741

The main reason is that the recommendation algorithm based on AT-LSTM takes into account how to calculate the emotion value of the word through the thematic model when it encounters the word without emotion tagging, that is, a completely unknown new word. Compared with the other three methods, AT-LSTM considers that an unfamiliar word may have emotion difference in different scenes; the emotion vector of the word in the current video shots can be calculated in real time according to the current topic distribution of the word. However, the LDA-based experimental method only considers the quantitative relationship between the words in each shots and the existing emotion word vector to evaluate the emotion of the shots. When the data in the emotion dictionary is lacking, the emotion of the unknown word cannot be explored. Therefore, the recommended method based on AT-LSTM is superior to other comparison methods.

5. Concluding

The emotional characteristics and trends of online video acquired after emotional analysis and visual processing to the information of bullet-screen comments can be used as the emotional tag of the video. On this basis we can establish a video retrieval mode based on the emotion of comments. We propose an Attention-based LSTM model(AT-LSTM) to experiment on the network bullet-screen comments, and analyzes the topic of bullet-screen comments clustering. By experimenting with our model and other algorithmic models, the performance of our model is superior to that of other models. The "highlights" video shots obtained from the emotion analysis model can be used to recommend the user to watch their interested bullet-screen video and help the user acquire the emotional information contained in the online video, thereby providing a new way of video retrieval.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (61373148, 61502151), Shandong Province Natural Science Foundation (ZR2012FM038, ZR2014FL010), Shandong Province Outstanding Young Scientist Award Fund (BS2013DX033), Science Foundation of Ministry of Education of China (14YJC860042), Project of Shandong Province Higher Educational Science and Technology Program (No.J15WB37, No.J15LN02) and Social Sciences Project of Shandong Province (No.15CXWJ13).

References

[1] Wu B, Zhong E, Tan B, et al. Crowdsourced time-sync video tagging using temporal and personalized topic modeling, International Conference on Knowledge Discovery and Data Mining, 2014: 721-730.

[2] Xian Y, Li J, Zhang C, et al. Video Highlight Shot Extraction with Time-Sync Comment[C]// International Workshop on Hot Topics in Planet-Scale Mobile Computing and Online Social NETWORKING. ACM, 2015:31-36.

[3] Chen X, Zhang Y, Ai Q, et al. Personalized Key Frame Recommendation, International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017: 315-324.

[4] He, M, Ge Y, et al. Predicting the Popularity of DanMu-enabled Videos: A Multi-factor View, Database Systems for Advanced Applications, Springer International Publishing, 2016:351-366.
[5] Zhao Yan-Yan, Qin Bing, Liu Ting. Sentiment Analysis, Journal of Software, Vol.21 (2010) No.8, p.1834−1848.
[6] Zheng YY, Xu J, xiao Z. Utilization of Sentiment Analysis and Visualization in Online Video Bullet-screen Comments, New Technology of Library and Information Service, 2015 (11): 82-90.
[7] Huang S, Niu ZD, Shi CY. Automatic construction of domain-speciflc sentiment lexicon based on constrained label propagation, Knowledge-Based Systems, 2014, 56: 191-200.
[8] Wang CH, Wang F. Extracting sentiment words using pattern based Bootstrapping method, Computer Engineering and Applications, 2014, 50(1): 127-129.
[9] Li RJ, Wang XJ, Zhou YQ. Semantic orientation computing using PageRank model, Journal of Beijing University of Posts and Telecommunications, 2010, 33(5): 141-144.
[10] Song YX, Zhang SW, Lin HF. Sentence Sentiment Analysis Based On Ambiguous Words, Journal of Chinese Information Processing, Vol.26 (2012) No.3, p. 38-43.
[11] Wang SG, Wu SH. Feature-Opinion Extraction in Scenic Spots Reviews Based on Dependency Relation, Journal of Chinese Information Processing, Vol.26 (2012) No.3, p. 116-121.
[12] Jiang TJ, Wan CX, Liu DX.Extracting Target-Opinion Pairs Based on Semantic Analysis, Chinese Journal of Computers, Vol.39 (2016) No.15.