Research on Opinion Target Extraction and System Information Collection Analysis for Reviews

JianNan Gao¹, *, Yanyan Zhao²

¹Key Laboratory of Interactive Media Design and Equipment Service Innovation, Ministry of Culture and Tourism School of Architecture, Harbin Institute of Technology, Harbin, China
²Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China

*Corresponding author email: jngao@ir.hit.edu.cn

Abstract. Though the study of opinion target extraction of movie review, this paper proposed a method for opinion target extraction of movie review based on deep learning, and analyzed its emotional polarity. Finally, the experimental results were displayed in a visual manner to help users judge the details of the movie content. In this paper, by establishing a bidirectional short-term and short-term memory network model, the opinion target extraction task is translated into a sequence labeling task, which improved the accuracy of opinion object extraction. This paper crawled and organized the Chinese movie review corpus from the Internet and the experimental results proved the effectiveness of the proposed method.

Keywords: opinion target extraction, sentiment analysis, movie review.

1. Introduction
With the rapid development of the Internet in the current era, more and more movie viewers will express their feelings and reviews on movies on different platforms, and make a comprehensive score for the film at the same time. A large number of film review data on different platforms have both universality and diversity. The filming experience and emotional features contained in these spontaneous film reviews are very important for the study of the audience's reflection on the film, the development of the mainstream content of the film, etc., and can be used for business film promotion planning, user viewing selection, and film producer selection. The purpose of this paper is to extract useful information from these massive film reviews with subjective will and conduct sentiment analysis to feed back to users in a more intuitive and readable way.

In the field of natural language processing, chapter-level movie reviews gathered on major movie review websites provide a real data basis for research and developers. Research on movie reviews has also become a research hotspot in recent years, such as the judgment of emotional polarity based on reviews as text content and movie recommendation systems based on movie reviews and movie ratings. [1,2] However, the research on the specific opinion target and emotional tendencies of the evaluation content in the review text is not enough such as: the acting of actors, plot, picture, etc., and the specific evaluation items are not scored by extracting opinion target and judgment of emotional
polarity. At present, there is not enough research on the mining of movie review text content and sentiment analysis.

Opinion target extraction and emotional key sentence extraction are basic tasks in natural language processing.[3] The task of opinion target extraction is to extract the evaluated part of the text content. For example, we get a comment about the movie from the Internet. "The development of the plot in the movie is too slow, especially the three views of the movie are simply incomprehensible, married Is it the best love for a man to have a mentally derailed first girlfriend?", the "story" and the "three views of movies" are all opinion target. Extracting these opinion target from film and television reviews and making a statistics can help us understand The public's focus on film and television content.

This article will use the Bi-directional Long Short-Term Memory (Bi-LSTM) model to extract opinion target from movie reviews. In this paper, the opinion target extraction task is transformed into a sequence tagging task, a Bi-directional Long Short-Term Memory network is used to build the model, and the BIO labeling system ((B-begin, I-inside, O-outside)) is used to extract the opinion target and carry out emotions. The judgment of polarity finally displays the five opinion target and their emotional tendencies that film critics pay most attention to in each movie in a visual way.

2.  Related Work

Many methods and models have been proposed and developed for the realization of the opinion target extraction task. These research methods basically use rule-based extraction and recognition and machine learning-based extraction and recognition.

For the former, various templates are constructed by analyzing the text, summarizing and concluding some characteristic information of the opinion target and the grammatical relationship between words, and finally completing the task of opinion target extraction. Among them, Hu and Li [4] proposed a method of opinion target extraction belonging to different categories using different rules and templates. Although this method can extract opinion target in a targeted manner, the template used The design requires a lot of work. Qiu [5] extended the method based on syntactic dependency to use the dependency relationship between the opinion target and the opinion word, and extract the opinion target through the double-propagation algorithm. Zhao Yanyan [6] proposed an automatic recognition method of sentiment evaluation unit based on syntactic path, which is different from the existing method based on manual templates and rules. This method automatically obtains the syntactic path to describe the modification relationship between the opinion target and its opinion words, and improves the system performance of the emotional evaluation unit extraction by calculating the edit distance of the syntax path. The experimental results show that the syntax path can effectively describe the opinion target and its evaluation. The relationship between words is of great help to the recognition of sentiment evaluation units.

For the latter, opinion target extraction objects can be regarded as a special case of information extraction. Research on information extraction has proposed many supervised learning algorithms. The mainstream method is rooted in sequential learning (Sequential Learning, or Sequential Labeling). Since these methods are supervised learning techniques, they need to have labeled data for training in advance. Currently the best sequence learning algorithms are Hidden Markov Model (HMM) and Conditional Random Field (CRF). Jin and Ho [7] used a lexicalized HMM model to learn the pattern of opinion target extraction and opinion words. Jakob and Gurevych [8] conducted CRF training in different domains to obtain a more domain-independent model, which used features such as part of speech, dependent syntax, sentence spacing and opinion sentences. Li [9] integrated Skip-CRF and Tree-CRF to extract opinion target. The characteristics of these two CRFs are that they can not only learn word sequences, but also find structural features.

In recent years, with the development of deep learning in many different fields, deep learning methods have also been applied to natural language processing tasks. Deep learning uses data for word embedding and uses neural network model processing. Many tasks of NLP, such methods are similar to sequence labeling tasks (such as CWS, POS, NER). The token is mapped from discrete one-hot representation to low-dimensional space to become dense embedding, and then the sentence The
embedding sequence is input into the recurrent neural network, the neural network is used to automatically extract features, and Softmax predicts the label of each token. This method makes the training of the model an end-to-end overall process, instead of a traditional pipeline, does not rely on feature engineering, and is a data-driven method. Compared with traditional machine learning methods, deep learning can actively learn the hidden features in the sequence, and achieve better results without too much manual intervention.

3. Methodology

3.1. Recurrent Neural Network

Deep learning is an emerging method of machine learning. It analyzes and uses different types of data by establishing neural networks and is widely used in various fields. With the rapid development of deep learning methods in the field of natural language, many deep learning models have been applied to opinion target extraction tasks. Recurrent neural networks (RNN) is a deep learning model that can be used to process sequence data. The network structure consists of an input layer, an output layer, and a hidden layer. When processing data, the output value of the current unit in the sequence is related to the output of the historical unit. It will memorize historical information and apply it to the current output calculation. Long short term memory (LSTM) is a special type of recurrent neural network, which alleviates the problems of gradient disappearance and long-term dependence when processing sequence data by ordinary recurrent neural networks. LSTM removes or adds information to the current unit through a structure called "gate". There are input gates, forget gates, and output gates in the LSTM unit, as shown in Figure 1. Among the information entering the unit, only the information that meets the requirements will be left, and the non-compliant information will be forgotten through the forgetting door. [10]

In sequence labeling tasks, the effect of using only historical information to calculate the output value of the current unit is limited. The two-way long and short-term memory network inherits the idea of the two-way cyclic neural network model, and designs two LSTMs forward and backward for each training sequence. This structure provides contextual information about the past and future of each element in the input sequence of the output layer. The forward LSTM captures the above feature information, and the reverse LSTM captures the following feature information, so that more feature information can be captured than a one-way LSTM, which is very important for many sequence labeling tasks.

![Figure 1. LSTM Unit Structure](image-url)
3.2. Opinion Target Extraction Model

The model is divided into three layers, namely the input layer, the BLSTM layer, and the output layer. The model structure is shown in Figure 2. The design of each layer is introduced below.

Input layer: The main function of the input layer is to process the text into a form that the model can accept as input, that is, perform word embedding, and input the words in the text into the BLSTM layer after vectorization.

BLSTM layer: The BLSTM layer constructs a bidirectional LSTM layer, where the forward LSTM can obtain the above feature information, and the reverse LSTM can obtain the following feature information. When the vector xt input by the input layer passes through the LSTM unit, at time t, the values of the input gate, output gate, and forget gate are calculated according to formula (1) to formula (6) and the value of the entire unit State value (output vector).

\[
\begin{align*}
    i_t &= \sigma (W_i h_{t-1} + U_i x_t + b_i) \\
    f_t &= \sigma (W_f h_{t-1} + U_f x_t + b_f) \\
    c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
    o_t &= \sigma (W_o h_{t-1} + U_o x_t + b_o) \\
    h_t &= o_t \odot \tan h (c_t)
\end{align*}
\]

Among them, \(\sigma\) represents the sigmoid activation function, and \(\odot\) represents the element product operation. \(x_t\) represents the input data of the network unit at time \(t\), and \(c_t\) represents the state value of the current unit. \(W_x\) and \(U_x\) (\(x = i, o, f, c\)) respectively represent the weights corresponding to the input gate, output gate, forget gate, and unit state. \(b_x\) (\(x = i, o, f, c\)) respectively represent the input gate, The offset value corresponding to the output gate, forget gate, and unit state. The hidden state of the unit (output vector) \(\tilde{h}_t\) is calculated from the front to the back through the forward LSTM layer \(\rightarrow\), the hidden state of the unit (output vector) \(\bar{h}_t\) is calculated from the back to the forward through the backward LSTM layer \(\leftarrow\), two hidden unit state vectors (The output vector) is combined according to formula (2-7), and input and output layer.

\[
h_t = \tilde{h}_t - \bar{h}_t + \bar{h}_t
\]

Output layer: Splice the obtained hidden layer states through the softmax function to obtain the probability distribution of the labels. The output layer will output the labels of the sequence according to the optimal labeling path for result analysis.

Model operation process: The model first needs to input the pre-processed data in batches and vectorize the data; process the input vectorized data through the BLSTM layer, and calculate according to formula (2-1) ~ formula (2-6) One-way hidden layer state, according to formula (2-7) to get the hidden layer state splicing including context; finally, the input sequence is sequenced through the output layer; after the above training process is looped, the training model is obtained.
4. Experiments

4.1. Experiment Data
In the experiment in this chapter, we used a processed dataset containing 7376 comment texts that were manually labeled, including "My First Half of Life", "Ancient Swords and Strange Tan", "Ode to Joy", "A Slight Smile is Allure", etc. The scale of the data set is shown in the table 1 for multiple film and television dramas and movie reviews.

| Comment Statement | Aspect | polarity-expression | polarity | other-polarity-word |
|-------------------|--------|---------------------|----------|---------------------|
| 7376              | 7534   | 3688                | 7376     | 1576                |

Figure 3. shows the labeling information of a sentence. The opinion target corresponding to each sentence in the data set is individually marked. The aspect module marks the opinion target in the sentence, the polarity-expression module marks the target words of the opinion target, and the polarity module marks the emotional tendency of the opinion target. The other-polarity-word module marks other words that judge the emotional tendency of the opinion target. The opinion target extraction task uses the opinion target information marked in the sentence.

I feel that Jin Dong plays the same role, I can't get his acting.
<aspect>Jin Dong's role</aspect>
<polarity-expression>follow the same pattern</polarity-expression>
<polarity>N</polarity>
<other-polarity-word>NULL</other-polarity-word>

4.2. Experiment Process

4.2.1. Data Preprocessing. The original label data used in the experiment is in txt format, which needs to be converted into the sequence label format needed in this article. Use the BIO marking system to mark the segmented sentences, using the number "2" for "B", the number "1" for "I", and the
number "0" for "O". "B" marks the first word of the opinion target, "I" marks the remaining words of the opinion target, and "O" marks the part that is not the opinion target. The data is preprocessed as shown in Figure 4. (Assuming the maximum length of the sentence is 10)

Figure 4. Data Preprocessing

4.2.2. *Train Optimization Model.* Set the model parameters, select 50% of the data as the training set to train the model, and 20% of the data as the validation set and adjust the parameters on the validation set. The final selected model parameters are shown in Table II. Save the model parameters that perform well on the training set and test set. Use the data of the training model to test the saved model to check whether the accuracy rate reaches 99%, to ensure that the model we trained converges.

| state size | epoch size | dropout rate | batch size | learning rate |
|------------|------------|--------------|------------|---------------|
| 50         | 100        | 0.5          | 32         | 0.01          |

In Table II, the state size represents the initial state size of the LSTM unit, the epoch size represents the number of times the training set is fully trained, the dropout value is used to prevent the occurrence of overfitting, and the batch size represents each time during training. The number of samples taken in the training set, the learning rate represents the learning rate, and the learning rate guides us how to adjust the network weight through the gradient of the loss function.

4.2.3. *Test Model.* Select the remaining data as the test set, use the model saved after training to test on the test set, generate labeling results, and count the experimental results.

4.3. *Experiment Results*  
The experiment uses the accuracy rate, recall rate, and F1 value of the model annotation results as evaluation criteria. The results are shown in Table III. From the experimental results in Table III, we can see the accuracy, recall, and F value of the words in the review text that are labeled with their corresponding labels.

The accuracy and recall rate for label 3B (the first word of the opinion target) are both high, reaching over 97%; but the accuracy and recall rate of label I (the remaining words of the opinion target) are relatively low. This is because most of the target words annotated in the data set we use are opinion target of a single word, such as "story", "screenwriter", "Shanghai Auntie", etc., and "Luo Jin's acting", "female image" There are few opinion target composed of multiple words, such as "Tang Jingren Design", etc., which leads to all opinion target need to be labeled with B label, and only a small part of opinion target need to be labeled with I label. Learning is not enough, the accuracy rate is relatively low when testing. But overall, this model can still complete the task of opinion target extraction.

| Precision | 0.971 | 0.478 | 0.592 |
| Recall    | 0.973 | 0.516 | 0.570 |
| F         | 0.972 | 0.496 | 0.580 |

Table 3. Table Type Styles
We built a set of models for opinion target extraction based on two-way LSTM. The opinion target extraction task is regarded as a sequence labeling task, the two-way LSTM is used as the hidden layer, and the Softmax regression model is used as the output layer, so that each word is on the label set. The value of the probability distribution. During the training process, we use the backpropagation algorithm to obtain the partial derivative value of the parameter, and use the AdaGrad optimization algorithm (Adaptive Gradient) to adjust the learning rate for each different parameter, and to make the parameter that changes frequently smaller The step size of is updated, and the sparse parameter is updated with a larger step size. This paper uses 7376 comment texts marked with multiple films and TV dramas and movie reviews as a data set for training and testing. The experimental results use accuracy, recall, and F1 values as evaluation criteria. The experimental results show that the above model can be very good Complete the extraction task.

5. Visual Display
This paper uses the above model to process movie review data crawled on the Internet, extracts opinion target and judges emotional polarity for each film and television work, and extracts the 5 opinion target with the highest frequency from them, and visualizes the experimental results Way to display. An example of the display results is shown in the figure 5. The area of the rectangle represents the frequency of the target words in all the reviews of each film, and the shade of the color represents the emotional polarity. In the color matching of geometric figures, the change of emotional tendency is expressed by changing the purity and brightness of the color. The higher the score, the lower the purity of the color block, the lower the brightness, and the lower the score, the higher the purity of the color block, the lower the brightness.

![Image](example_image.png)

Figure 5. Visual Display

6. Conclusions And Future Work
Based on the information contained in a large number of movie reviews accumulated on the Internet, it is very important for merchants’ movie promotion planning, user viewing choices, and movie producers’ choice of theme content. There is an urgent need for technology that can process movie review text data. Driven by deep learning, many tasks in natural language processing have implementation methods that achieve better results and faster completion. This paper uses a deep learning model to extract opinion target from movie review texts, builds a two-way long and short-term memory network to convert opinion target extraction tasks into sequence labeling tasks, and uses BIO labeling system to extract opinion target. The experimental results show that The method can well complete the task of opinion target extraction.

The deep learning-based model built and trained in this article completes the task of processing movie review data, visualizing the information in a large number of movie reviews, helping users quickly obtain the main review content of a movie, and enabling users to directly obtain movie reviews of related movies The most frequent opinion target and corresponding score information.

In the next step, we can consider using a more complex model than the one used in this article to achieve the task of opinion target extraction and identifying key emotional sentences; in addition, the visual display of the film completed in this article can be further developed into a film review app. Features.
References

[1] Saranya Maneeroj, Yuka Kato, Katsuya Hakozaki. An Advanced Movie Recommender System Based on High-Quality Neighbors. 2005, 1:181-192.

[2] Mala Saraswat, Shampa Chakraverty, Atreya Kala. Analyzing emotion based movie recommender system using fuzzy emotion features[J]. International Journal of Information Technology: An Official Journal of Bharati Vidyapeeth's Institute of Computer Applications and Management, 2020, 12(6).

[3] Liu Bing. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions [M]. Cambridge: Cambridge University Press, 2015:137-138.

[4] Hu M, Liu B. Mining and summarizing customer reviews[C]// Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2004:168-177

[5] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through double propagation[J]. Computational linguistics, 2011, 37(1):9-27

[6] Zhao Yanyan, Qin Bing, Che Wanxiang, et al. AppraisalExpression Recognition Based on Syntactic Path [J]. Journal of Software, 2011, 22(05): 887-898.

[7] Jin W, Ho H H, Srihari R K. A novel lexicalized HMM-based learning framework for web opinion mining[C]// Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 465-472.

[8] Jakob N, Gurevych I. Extracting opinion targets in a single-and cross-domain setting with conditional random fields[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 1035-1045.

[9] Li F, Han C, Huang M, et al. Structure-aware review mining and summarization[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 653-661.

[10] P Zhang, T Chen, C Chen, Y Wang. Opinion targets extraction method for Chinese microblogs based on deep learning [J]. Computer Engine and Design, 2018, 39(08):2638-2642+2693