A Small amount of Labeled Data Chinese Online Course Review Target Extraction via ALBERT-IDCNN-CRF Model

Liang Min, Xianglin Miao, Peng Bi and Feijuan He*

Department of Computer Science, Xi'an Jiaotong University City College, Xi'an, 710018, China

*Corresponding author's e-mail: hfj@xjtu.edu.cn

Abstract. Aspect sentiment analysis of online course reviews is of great significance in helping users choose courses and improve course quality. Review target extraction is particularly important as the basis of aspect sentiment analysis. Because the current models mostly rely on a large amount of annotation data, there is fewer relevant research on the extraction of online course review targets with higher annotation costs. This paper proposes an ALBERT-IDCNN-CRF review target extraction model for a small amount of labeled data. First, using ALBERT pre-trained sentences obtained dynamic model Chinese word vector coding; Simultaneously; using ALBERT pre-trained model of Transformer obtain sentence abstract features. Then, abstract features are input into the dilated convolutional neural network (IDCNN) to reduce the number of neuron layers and parameters. Finally, conditional random field (CRF) is used to decode and annotate the review sentences to extract the appropriate review objectives. The experimental results on the school online real Chinese online course review data set show that our model has achieved better results than existing models.

1. Introduction
Aspect sentiment analysis of online course reviews is of great significance to students' course selection, course optimization, platform management, etc. Online course review target extraction as a precondition of aspect sentiment analysis has also attracted more and more researchers' attention. The purpose of online course review target extraction is to extract the target entity of the evaluation from the online course review text. E.g. "课程设计的非常合理，老师上课认真负责，感谢学堂在线提供这么好的学习平台! (The curriculum design is very reasonable, and the teachers are serious and responsible in class. Thank you Xuetang for providing such a good learning platform!) " In this sentence, three target entities are designed, namely "课程(course)", "教师(teacher)" and "平台(platform)". The purpose of online course review target extraction is to identify the three target entities involved in this sentence.

In recent years, with the rapid development of deep learning, an end-to-end sequence annotation model that does not rely on language rules and artificial features has become the mainstream solution for review target extraction. In 2015, Huang et al. proposed a fusion of long and short-term memory network and conditional random field sequence labeling model (BiLSTM-CRF), which is currently the most widely used sequence labeling model. Subsequently, the literature [2] proposes a sequence labeling cavity model based on a convolutional neural network (IDCNN-CRF), which uses the training faster IDCC Alternatively the BiLSTM, while improving the efficiency of the model is achieved comparable BiLSTM-CRF Accuracy. The literature [3] pointed out that the Attention-BiLSTM-CRF model combined with the attention mechanism and the conditional random field obtained a very high F1 value in the sequence labeling task. Many scholars have also explored the issue of Chinese sequence labeling.
Literature [4] found that there is a distinction between characters and words in the Chinese sequence tagging task. Literature [5] improved the traditional LSTM unit into a grid and proposed the Lattice LSTM model, which explicitly uses the word and word order information based on the word model to avoid the problem of incorrect transmission of word segmentation. However, the above models all require a large amount of annotation data training. The cost of the manual annotation of Chinese online course reviews is high, and the above model does not work well when the data size is limited. Besides, these methods cannot model the ambiguity of words using traditional word vector models. In 2018, Google proposed the BERT pre-training model, which can achieve good results with a small amount of labeled data. Also, the BERT model can model the ambiguity of words in a more contextual context [6]. Lan, who proposed and BERT model of efficient performance similar version BERT model, reducing the parameters of the model to enhance the training speed [7].

Therefore, an ALBERT-IDCNN-CRF Chinese online course review target extraction model is proposed. The model uses the ALBERT pre-training model to dynamically encode Chinese characters in sentences; uses Transformer to automatically extract semantic features of comments in parallel; uses IDCNN to reduce the number of neuron layers and parameters; finally, CRF generates the optimal label sequence. Experiments were conducted in the online Chinese online course reviews of the school. The experimental results show that the ALBERT-IDCNN-CRF model has a higher F1 score than the baseline model.

2. Task description and model

2.1. Task description

In Chinese online courses Reviews target extraction problem, give a sentence \( s = \{w_1, w_2, \ldots, w_n\} \) containing \( n \) words, Where \( t = \{w_1, w_2, \ldots, w_m\} \) is the target containing \( m \) Chinese characters in the sentence. The target can be a Chinese character or a phrase composed of multiple Chinese characters. The main objective of this paper is to identify a sentence \( S \) are directed to the targets \( t \).

2.2. ALBERT-IDCNN-CRF target extraction model

Figure 1 shows the overall architecture of the ALBERT-IDCNN-CRF target extraction model. It mainly includes four parts: Input layer, Transformer layer, IDCNN layer and CRF layer.

![Figure 1. ALBERT-IDCNN-CRF model framework.](image-url)
2.2.1. **Input layer.** The input of the ALBERT pre-training model is a single Chinese character, and the model is trained through two modes of randomly covering and judging sentence coherence. The specific method of random cover mode is: randomly cover 15% of Chinese characters in the review, and train the model to predict the covered part. The masking rules are as follows: 80% use masked token instead, 10% use a random word to replace, 10% keep the word unchanged. Specific method of determining a coherent sentence is: in the same selected documents in the sentence, the sentence is determined whether the consecutive order. If the sentence is coherent, it is a positive sample, and if it is not coherent, it is a negative sample. The ALBERT pre-training model calculates three features for each Chinese character in the review, which are character feature, sentence feature, and location feature. ALBERT pre-training model input layer is the final output that is:

\[
E = (e_1, e_2, \cdots, e_n) \tag{1}
\]

Where \( e_i = w_i^f + w_i^p + w_i^r \).

2.2.2. **Transformer layer.** The encoder structure of the Transformer model is formed by stacking \( N \times (N = 6) \) identical basic layers, as shown in Figure 2 [8]. Each basic layer is composed of two sub-layers, the first is a multi-headed attention layer, the second is a dense fully connected feed-forward neural network layer, then a residual connection is used in the two sub-layers, and then layer standardization operating.

![Figure 2. The Transformer the basic structural units.](image)

The multi-head attention mechanism can usually be described as follows: The key-value pairs of the query vector and a series of key vectors and value vectors are mapped to the output using weight matrices of different dimensions. It is used to obtain sentence-level semantic information of online course reviews. Residual connection is a commonly used method to solve the problem of deep neural network training. It passes the information of the previous layer to the next layer without error, thereby solving the problem of the disappearance of gradient in the deep neural network. Layer normalization refers to normalizing the activation value of each layer, which can accelerate the model convergence speed. The final output of the Transformer layer of the ALBERT pre-training model is:

\[
H = \text{Transformer}(E) \tag{2}
\]

Where \( H=(h_1, \cdots, h_n) \), \( h_i \) is the hidden states of the Chinese characters in the course review text that have been semantically extracted in the Transformer layer.

2.2.3. **IDCNN layer.** The IDCNN layer has the powerful parallel computing power of the convolutional neural network and as much input information as possible in the LSTM memory network [2]. IDCNN standard CNN the filter after adding a dilation width, when the input matrix, will skip all the dilation width intermediate input data, and the filter size itself remains unchanged, so that filter the acquired data on the wider input matrix. As shown in Figure 3, the dilated width will increase exponentially as the
number of layers increases. The number of parameters in each layer is independent of each other. As the number of layers increases, the number of parameters increases linearly, while the receptive field increases exponentially, which can quickly cover all input data.

![Figure 3. Schematic diagram of IDCNN.](image)

It can be seen in Figure 3 that the receptive field expands at an exponential rate. The original receptive field is a 1x1 area located at the center point. Figure 3-(a) The original receptive domain is spread out in steps of 1 to obtain 8 areas of 1×1 to constitute a new receptive domain, with a size of 3×3. In Figure 3-(b), after the diffusion with a step size of 2, the receptive field of the previous step 3×3 is expanded to 7×7. In Figure 3-(c), after the diffusion with a step size of 4, the original 7×7 receptive field is expanded to a 15×15 receptive field. The final output of the Transformer layer of the ALBERT pre-training model is:

\[ H^* = IDCNN(H) \]  

Where \( H^* = (h^*_1, \ldots, h^*_n) \). \( h^*_i \) is the hidden state of Chinese characters in the course review text after the semantic extraction of the IDCNN layer.

2.2.4. CRF layer. Following the IDCNN layer followed by the CRF layer is a common practice in sequence annotation. The CRF layer can consider the global information of the tag sequence and add constraints to the last predicted tag. Based on these constraints, the probability of occurrence of illegal sequences in the prediction of label sequences is reduced, and labels are better predicted. The CRF layer can learn context information, combine the output layer results and the global probability of the label sequence, and predict the label sequence with the highest probability. Given the input sentence \( X = (x_1, x_2, \ldots, x_n) \) and the output label sequence \( Y = (y_1, y_2, \ldots, y_n) \), the total score of the label sequence is:

\[
S(X, y) = \sum_{i=0}^{n} A_{y_{i},y_{i+1}} + \sum_{i=0}^{n} P_{i,y_{i}}
\]

Where \( A \) and \( P \) are the transfer score matrix and the output score matrix, respectively. Matrix element \( A_{y_{i},y_{i+1}} \) represents the transfer score from label \( i \) to label \( i+1 \). The matrix element \( P_{i,y_{i}} \) represents the output score of the \( i^{th} \) Chinese character under the \( y_{i}^{th} \) label.

After normalizing all possible sequences, the probability distribution of the generated output sequence is shown in equation (5).

\[
P(y | X) = \frac{e^{S(X, y)}}{\sum_{j \in Y} e^{S(X, j)}} \sum_{i=0}^{n} A_{y_{i},y_{i+1}} + \sum_{i=0}^{n} P_{i,y_{i}}
\]

Where \( \tilde{y} \) represents the true mark value. The likelihood function of the labeled sequence during training is:
\[
\log(P(y^* \mid X)) = S(X, y^*) - \log(\sum_{y \in Y_X} e^{S(X, y)})
\]  \hspace{1cm} (6)

Where \( Y_X \) represents all possible mark sets, including mark sequences that do not conform to beginning-inside-outside (BIO) ternary labeling rules. Through equation (6) training, an effective and reasonable output sequence is obtained. When predicting, a set of sequences with the highest overall probability is output by equation (7):

\[
\log(P(y^* \mid X))y^* = \arg \max_{y \in Y_X} S(X, y)
\]  \hspace{1cm} (7)

3. Experiment

3.1. Dataset

In order to evaluate the ALBERT-IDCNN-CRF Chinese online course review target extraction model, we conducted experiments on the school online data set manually marked by the BIO ternary labeling method. The detailed information of the data set is shown in Table 1.

| Dataset   | Corpus | Teacher | Course | Platform |
|-----------|--------|---------|--------|----------|
| Training set | 25480  | 18263   | 15802  | 6681     |
| Test set  | 6627   | 5356    | 2068   | 1020     |

3.2. Evaluation index and experiment settings

The parameter settings of the ALBERT-IDCNN-CRF model are shown in Table 2. The parameters of the baseline model selected in the experiment all adopt the default parameters of the model.

| Parameter name       | Parameter value |
|----------------------|-----------------|
| Word vector          | ALBERT-Base     |
| Word vector dimension| 768             |
| Maximum sequence length | 128            |
| Batch size           | 64              |
| Learning rate        | 1e-4            |
| Training steps       | 5               |
| Number of layers     | 4               |

3.3. Experimental results and analysis

In order to analyze the performance of the ALBERT-IDCNN-CRF model, we compare it with four baseline models. The experimental results are shown in Table 3.

| Model          | P   | R   | F1  |
|----------------|-----|-----|-----|
| CRF            | 0.8543 | 0.8241 | 0.8389 |
| IDCNN-CRF      | 0.8652 | 0.8489 | 0.8570 |
| ALBERT         | 0.8845 | 0.8578 | 0.8709 |
| ALBERT-CRF     | 0.9042 | 0.8628 | 0.8830 |
| ALBERT-IDCNN-CRF | 0.9172 | 0.9275 | 0.9223 |

It can be seen from the experimental results in Table 3 that the ALBERT-IDCNN-CRF model proposed in this paper has the highest F1 score among the five models. This is because the ALBERT pre-training model adopted by ALBERT-IDCNN-CRF carries out a large number of semantic
representation learning on a large number of unsupervised libraries, and can dynamically encode the same Chinese characters with different meanings according to the context semantics, which is better than the words used in the remaining models. The vector coding model works better.

4. Conclusion
Aiming at the problem that the existing Chinese online course review target extraction model is overly dependent on the scale of the annotation data. An ALBERT-IDCNN-CRF model for a small amount of annotated data is proposed, which can represent the semantic information of input sentences through a large number of pre-training corpora, which greatly reduces the model's dependence on the scale of annotated data. In addition, the model reduces the number of neuron layers and parameters through IDCNN; uses CRF to grammatically constrain the sequence to further extract the rationality of the target. Experimental results show that ALBERT-IDCNN-CRF has a better effect than other models in data sets with limited annotation data. In future research, it is planned to migrate the model to the XLNet pre-training model [9] in order to further improve the performance of the model.

Acknowledgments
This work was supported by National Social Science Fund of China (18XXW005); Ministry of Education Humanities and Social Sciences Fund (17YJA860028); Natural Science Basic Research Program of Shaanxi (2020JM-711); Social Science Planning Fund of Xi'an (WL141); Xi'an Jiaotong University City College (2020Y11).

References
[1] Huang Z, Xu W, Yu K. (2015) Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint, 8: 01991.
[2] Trubell E, Verga P, Belanger D, et al. (2017) Fast and accurate entity recognition with iterated dilated convolutions. arXiv preprint, 2: 02098.
[3] Luo L, Yang Z, Yang P, et al. (2018) An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34: 1381-1388.
[4] Verhaeghe H, Lecoutre C, Deville Y, et al. (2017) Extending compact-table to basic smart tables. In: Principles and Practice of Constraint Programming. Stamford. pp. 297-307.
[5] Zhang Y, Yang J. (2018) Chinese ner using lattice lstm. arXiv preprint, 5: 02023.
[6] Devlin J, Chang M W, Lee K, et al. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 10: 04805.
[7] Lan Z, Chen M, Goodman S, et al. (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint, 9: 11942.
[8] Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: Advances in neural information processing systems. Long Beach. pp. 5998-6008.
[9] Yang Z, Dai Z, Yang Y, et al. (2019) Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. Vancouver. pp. 5753-5763.