Chinese Event Subject Extraction in the Financial Field Integrated with BIGRU and Multi-head Attention

Yangwenhao Liu*, Changhui Liu¹, Liya Wang², Zhe Chen¹ and Yuan Wu¹

¹ School of Computer Science & Engineering, Wuhan Institute of Technology, Wuhan 430205, China
² Artificial Intelligence Academy, Zhejiang Industry & Trade Vocational College, Wenzhou 325003, China

*Email: lywh0117@qq.com

Abstract. Event subject extraction was to extract subjects of specific event types. For the traditional BiLSTM network, the threshold is complicated, the required parameters are many, and the time cost is high. This paper is oriented to the financial field and proposes a method of introducing a multi-head attention mechanism based on the BIGRU network to extract event subjects. First, the text is vectorized, and then the word vector obtained is input into the BIGRU network to learn the context features, and introduce a multi-head attention mechanism to extract the depth feature values of the text. Finally, a comparative experiment is conducted on the data set. The method in this paper achieves an Accuracy value of 80.47% and an F1 value of 89.18%. The result is better than the control group, indicating that the model proposed in this paper can effectively improve the accuracy of extracting Chinese event subjects.

1. Introduction

The extraction of event subject belongs to the category of information extraction[1-3]. The event subject extraction task in this article originated from the CCKS: China Conference on Knowledge Graph and Semantic Computing and Ant Financial, as well as the CCKS 2019 financial field event subject extraction competition[4] jointly organized by the Institute of Automation of the Chinese Academy of Sciences.

The extraction of event subjects is one of the important tasks in the field of public opinion monitoring and the financial field. In the financial field, "events" are an important decision-making reference for investment analysis and asset management. The complexity of the extraction of the event subject lies in the judgment of the event type and the event subject. For example, "Company A has additives in its products, and its subsidiaries B and Company C have been investigated." For the "product violation" event type, the subject in this sentence is "Company A" instead of "Company B" or "Company C". It is said that the subject of a specific event type becomes the subject of the event. The scope of the subject of the event in this task is limited to: companies and institutions. The event type is determined as: ['Funding account risk', 'Suspected of fraud', 'Decline in performance', 'Letter approval violation', 'Suspected MLM', 'Transaction violation', 'Financial fraud', 'Rating adjustment', 'Restructuring Failed', 'Change of actual controller shareholder', 'Unable to perform duties', 'Suspected of illegal', 'Suspected of illegal fund-raising', 'Negative assets', 'Closing of business', 'Difficulty withdrawing', 'Negative executives', 'Complaint against rights protection', "lost connection and run away", "product violation", "company stock market abnormal"].
2. Related Research Work

There are mainly three traditional information extraction methods, which are based on rules [5,6], based on statistics [7], and methods based on a combination of rules and statistics [8-10]. However, these three methods have strong limitations. They rely too much on manual rules, the process is complicated, the human engineering is huge, and the influence of human factors is too large. Later, with the gradual maturity of deep learning technology, deep learning methods were applied to information extraction. Deep learning-based methods realized an end-to-end extraction method, no longer relying on artificial features, and reduced the human cost required by traditional methods. In turn, work efficiency is improved.

Collobert and Weston [11] introduced the neural network framework to the named entity recognition task for the first time in 2008. The feature vector input is composed of English word spelling features, dictionaries and dictionaries. Subsequently, Collobert et al. [12] improved the model again in 2011 and replaced the artificially constructed features with word embeddings, which can display words in n-dimensional space and retain word information. This method is a completely unsupervised learning process, which extracts features of words only through co-occurrence features near the words. Chen [13] proposed a dynamic multi-pool convolutional neural network (DMCNN) model in 2015. Convolutional neural network (CNN) has strong local learning capabilities and can extract deep phrase features, but due to the CNN network itself The limitations of, it is impossible to learn the characteristics of the sentence system and ignore the contextual semantic characteristics. Nguyen [14] proposed the JRNN model based on the C-CBOW language and training model in 2016, and used Bidirectional Long Short-Term Memory (BiLSTM) to encode the resulting vector. The BiLSTM is a combination of two LSTMs in the opposite direction. In order to obtain more comprehensive sentence features than LSTM, it makes up for the shortcomings of CNN.

However, due to the limitations of the BiLSTM network, it is too costly to extract working time in real life. This paper is oriented to the financial field and introduces the BiGRU network plus multi-head attention mechanism model on the task of event subject extraction, which strengthens the model's ability to learn text features. The experimental results show that for the data set in this article, the method based on BiGRU introduces the attention mechanism (BiGRU_MA) to obtain an Accuracy value of 80.47% and an F1 value of 89.18%, which is better than the control group.

3. BIGRU_MA Model

The Chinese event subject extraction method based on the BIGRU_MA model is mainly divided into three parts: one is Bidirectional Gated Recurrent Unit (BiGRU), the other is Multi-Head Attention (MA), and the third is decoding function. The model structure is shown in Figure 1.

Figure 1. BIGRU with MA model.
3.1. BIGRU Model

Cho et al. [15] proposed Gated Recurrent Unit (GRU). GRU is an improvement of LSTM. It combines the forget gate and input gate into one update gate, which has a simpler structure, fewer parameters, and reduced Model time cost. BiGRU is a neural network composed of output state connection layers of forward GRU, reverse GRU, and forward and reverse GRU. BiGRU can well solve the shortcomings of GRU while maintaining its advantages. That is, BiGRU can learn relatively global features, and the required parameters are reduced. Compared with BiLSTM, the network structure is relatively simple, so as to improve the model training speed.

If the hidden state of the forward GRU output at time \( t \) is \( \vec{h}_t \) and the hidden state of the reverse GRU output is \( \vec{h}_t \), then the hidden state of the BiGRU output is \( \vec{h}_t \). The specific calculation process is shown in formulas 1~3. The network structure of BiGRU is shown in Figure 2.

\[
\vec{h}_t &= \text{GRU}(\vec{h}_{t-1}, U_t) \\
\vec{h}_t &= \text{GRU}(\vec{h}_{t-1}, U_t) \\
h_t &= w_t \vec{h}_t + v_t \vec{h}_{t-1} + b_t
\]

where \( w_t, v_t \) is the weight matrix, GRU: GRU function, \( U_t \): GRU input at time \( t \), \( b_t \): offset vector.

![Figure 2. BiGRU model.](image)

3.2. Multi-Head Attention

The attention mechanism originated from human visual attention, which simulates that when humans observe information, they will focus on certain specific parts of the information. Currently, Attention has been successfully applied to multiple tasks, such as machine translation [16], text classification [17], image recognition [18], etc. Vaswani et al. [19] proposed the multi-head attention mechanism (MA). Multi-Head Attention can capture the key information of the sequence from many aspects, that is, it can learn deeper text feature information. The Multi-Head Attention structure is shown in Figure 3.

The model takes the output of the BIGRU layer as a vector \( X_B \), and the text input sequence is \( X = (x_1, x_2, ..., x_t) \), where \( x_i \) represents the word vector of the i-th word, and the dimension is \( d \), \( X \in \mathbb{R}^{n \times d} \). The input of the Multi-Head Attention layer is \( [X_B, X_B, X_B, X, X] \), \( Q=K=V=X_B, Q \in \mathbb{R}^{n \times d_B}, K \in \mathbb{R}^{m \times d_k}, V \in \mathbb{R}^{m \times d_v}, Q, K \) and \( V \) pass through three independent linear transformations, as shown in formulas 4-6.

\[
Q = QW^Q_i \\
K = KW^K_i \\
V = VW^V_i
\]

Pass the \( Q, K \) and \( V \) values after three independent linear transformations into Scaled Dot-Product Attention, and repeat the Scaled Dot-Product Attention operation \( h \) times, \( h \) is the number of "heads" of Multi-Head Attention, Take single head attention calculation as an example, as shown in formulas 7-8.

\[
head_i = \text{Attention}(Q, K, V) \\
\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V
\]

Finally, all Attention values (\( head_i \)) are spliced together as the final output of the Multi-Head Attention layer, as shown in formula 9.
\[ MultiHead(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h) \] (9)

Figure 3. Structure of Multi-Head Attention.

3.3. Decoding Function
Two softmax were used to predict the beginning and the end of the results respectively. The softmax function is shown in formula 10.

\[ \text{softmax}(x) = \frac{\exp(x - \max(x))}{\sum \exp(x - \max(x))} \] (10)

4. Experiment
The experiment is oriented to Chinese texts in the financial field. It verifies and analyzes the event subject extraction method based on the BiGRU network that introduces the attention mechanism.

4.1. Experimental Data
The data comes from the CCKS 2019 event subject extraction competition for the financial field, a total of 17,815 entries, but because the event type label of some data in the training set given by the competition is "other", that is, invalid data, it needs to go through simple data cleaning and remove it. There are a total of 3141 pieces of data in this part, and a total of 14,674 pieces of real and effective data sets are found. Among them, it must be added that because the experimental results have not been submitted, the test data given by the competition cannot be used, so the experiment will clean the formal training set of the competition and divide it into a training set at a ratio of about 8:1:1: Validation set: test set. The data set settings are shown in Table 1.

| Data set   | Train set | Verification set | Test set | total  |
|------------|-----------|------------------|----------|--------|
| CCKS 2019  | 11674     | 1500             | 1500     | 14674  |

4.2. Evaluation Index
The evaluation index used in the experiment is Accuracy. The higher the value, the better the model extraction effect. Definition TP: the number of entities that are correctly identified by the model, FP: the number of results that the model can identify entities but the type or boundary determination is wrong, FN: the number of related entities but the model does not detect. Then the evaluation index is shown in formula 11-14.

\[ \text{Accuracy} = \frac{TP}{TP + FP + FN} \] (11)
\[ \text{Precision} = \frac{TP}{TP + FP} \] (12)
\[ \text{Recall} = \frac{TP}{TP + FN} \] (13)
\[ F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \] (14)

4.3. Comparative Experiment Settings

1) BILSTM[14]: A single BILSTM network uses the word vector obtained from word2vec training as the input of the BILSTM network.

2) BIGRU[15]: A single BIGRU network uses the word vector obtained by word2vec training as the input of the BIGRU network.

3) BILSTM_Multi-head attention mechanism(BILSTM_MA): The word vector obtained by word2vec training is used as the input of the BILSTM network and then the multi-head attention mechanism is introduced.

4) BIGRU_Multi-head attention mechanism(BIGRU_MA): The word vector obtained by word2vec training is used as the input of the BIGRU network and then the multi-head attention mechanism is introduced. The experiment is built using the Keras framework and TensorFlow backend. The parameters are set as follows: batch_size=64, char_size=128, max_len=256, gru_dim=64, nb_head=8, size_per_head=16, epoch=115, learning_rate=1e-4, optimizer='Adam'.

4.4. Analysis of Experimental Results

Because the Evaluation Index Score obtained with the test set can reflect the pros and cons of a model better than the score obtained with the training set. The experimental selection is calculated on the test set. The comparison results of the two groups of models are shown in Table 2.

Table 2. Test results of models.

| Model       | Accuracy | Precision | Recall | \( F1 \) |
|-------------|----------|-----------|--------|---------|
| BILSTM      | 0.7847   | 0.9439    | 0.8231 | 0.8793 |
| BIGRU       | 0.7873   | 0.9532    | 0.8190 | 0.8810 |
| BILSTM_MA   | 0.7853   | 0.9546    | 0.8158 | 0.8798 |
| BIGRU_MA    | 0.8047   | 0.9482    | 0.8417 | 0.8918 |

Table 2 shows the comparison results of the two groups of models on the test set. Judging from the evaluation index Accuracy. The BIGRU_MA model is better than the BILSTM_MA model. It shows that the model in this paper can effectively improve the accuracy of subject extraction of Chinese events.

From the detailed analysis of the extraction results of the test set, Table 5 shows the comparison of the results of the third and fourth experimental test sets.

Table 3. Comparison of test set extraction results.

| Model       | TP   | FP   | FN   | Right | Wrong |
|-------------|------|------|------|-------|-------|
| BILSTM      | 1177 | 70   | 253  | 1177  | 323   |
| BIGRU       | 1181 | 64   | 261  | 1181  | 319   |
| BILSTM_MA   | 1178 | 56   | 266  | 1178  | 322   |
| BIGRU_MA    | 1207 | 66   | 227  | 1207  | 293   |

It can be seen from Table 5 that the five index values of the extraction results on the test set are shown. Among them, the Right value is the number of completely correct extraction results, and the Wrong value is the number of incorrect extraction results, including the FP and FN values, which quantifies the performance of the comparison model. The BIGRU_multi-head attention method is used to extract event subjects. The total number of test samples is 1348, the errors are 264, and the correctness is 1084. The result is better than the control group.
The following analysis is carried out from the training stage, and the change graph of the accuracy of the training set during the entire training is drawn.

**Figure 4.** Comparison of BILSTM and BILSTM_MA models.

Figure 4 and Figure 5 are the comparison of BILSTM and BILSTM_MA models, and the comparison of BIGRU and BIGRU_MA models, where Accuracy is the accuracy on the ordinate, and Epochs on the abscissa is the number of iterations during training. From the accuracy changes of the training set in Figures 4 and 5, it can be seen that after the multi-head attention mechanism is added, the model converges faster on the data set, and the accuracy is higher than the single BILSTM and single BIGRU models.

**Figure 5.** Comparison of BIGRU and BIGRU_MA models.

**Figure 6.** Changes in the accuracy of the training set of the 4 comparison models.

Figure 6 also plots the accuracy changes of the training set of the four groups of comparison models. It can be seen that the model in this paper adds a multi-head attention mechanism to the BIGRU network.

**Table 4.** Time cost.

| Model        | s/Step | s/Epoch | s/Total |
|--------------|--------|---------|---------|
| BILSTM       | 1      | 224     | 26880   |
| BIGRU        | 0.515  | 94      | 11280   |
| BILSTM_MA    | 3      | 585     | 70200   |
| BIGRU_MA     | 2.5    | 471     | 56520   |

Table 4 shows the average time cost of the comparison model on the training set, where s/Step represents the time for the model to complete each step, that is, each batch, s/Epoch represents the time for the model to complete each iteration, s/Total represents the total time for the model to complete training, that is, the time for 120 iterations, and the unit of value in the table is seconds (s). The data in Table 4 shows that the time cost of the BILSTM_multi-head attention mechanism model is about 585s/epoch, and the time cost of
the BIGRU_multi-head attention mechanism model is about 471s/epoch, indicating that the latter adds multi-head attention to the BIGRU network. The time cost of the mechanism is small. In life, the company needs to process a large amount of text, and the time cost is the focus of attention. Therefore, BIGRU's method of introducing a multi-head attention mechanism has more practical value under the premise of improving the accuracy of the experiment. Table 5 shows the actual extraction results of detailed samples from the BIGRU_multi-head attention model.

Table 5. Result of sample extraction.

| Sample                                                                 | Label 1: Event type | Label 2: Event body | Extraction result |
|-----------------------------------------------------------------------|---------------------|---------------------|-------------------|
| ‘东芝旗下西屋电气将申请破产’保护长航凤凰(000520)78 亿元重组“告吹”公司称暂无再次重组计划’ | ‘重组失败’          | ‘长航凤凰’         | ‘长航凤凰’       |
| ‘2019年1月1日，天津市公安机关对权健自然医学科技发展有限公司涉嫌组织、领导传销活动罪和虚假广告罪立案侦查’ | ‘涉嫌欺诈’          | ‘权健自然医学科技发展有限公司’ | ‘权健自然医学科技发展有限公司’ |
| ‘出售电池隔膜公司股权’比亚迪(002594)获4.3亿元投资收益联合信下调中城建评级’     | ‘评级调整’          | ‘中城建’            | ‘中城建’         |
| ‘刚泰控股(600687):资产重组境内项目终止 不影响境外项目推进交大昂立副总裁葛剑秋违规减持 遭上交所通报披露索具(002342)大股东两个交易日套现2.5亿元’ | ‘重组失败’          | ‘刚泰控股’          | ‘刚泰控股’       |
| ‘出售思捷中国套现30亿 华创中期纯利劲升2.7倍上交所通报天目药业(600671)信批违规 部分印章两被罢独董质疑’ | ‘信批违规’          | ‘天目药业’          | ‘天目药业’       |
| ‘英力特(000635):实控人拟转让51.25%股份 控制权或变更通用汽车公司将走出破产保护华联控股11名高管减持刹车改增持 二股东提议“罢免董事长”初衷不改’ | ‘实控人股东变更’   | ‘力合股份’          | ‘英力特’         |

5. Summary

This paper, oriented to the financial field, proposes a method of extracting Chinese event subjects with BIGRU introducing a multi-head attention mechanism. Using the BIGRU network to learn contextual semantic features reduces the time cost of the model and improves the work efficiency of the model. By introducing the multi-head attention mechanism, the key feature information in the sequence is captured from multiple aspects, and the extraction further improves the accuracy of the extraction work. The experiment was trained and tested on Chinese financial texts. The results show that BIGRU's method of introducing a multi-head attention mechanism can effectively improve the accuracy of the extraction of Chinese event subjects, but because the model language pre-training part uses a simple embedding layer (Embedding), Affect extraction accuracy. In the future, it is the goal of the next step to study how to improve the accuracy of the extraction work and the model with less time cost.
References

[1] Guo X, He T 2015 Overview of Research on Information Extraction Computer Science 2015,42(02) 14-17+38

[2] Du Q 2017 Research on the spatiotemporal information extraction method of events in unstructured text (Shandong Normal University)

[3] Wikipedia: Message Understanding Conference [EB/OL]. 2013-12-27. Available at: http://en.wikipedia.org/wiki/Message_Understanding_Conference

[4] https://www.biendata.com/competition/ccks_2019_4/

[5] Feng C, Hao A 2006 Natural language recognition based on pattern matching Computer Engineering and Applications 42(19) 144-6

[6] Wen Y, Tan H, Zheng J 2009 Research on time standardization based on rules 2009 International Information Technology and Application Forum

[7] Liu Z, Du Y, Shi S 2019 Key Information Extraction of Web News Based on Improved Hidden Markov Model Data Analysis and Knowledge Discovery 3(03)120-8

[8] Cheng Z 2015 Research on Chinese Named Entity Recognition Method Based on Rule and Conditional Random Field (Wuhan, China: Central China Normal University)

[9] Yan Z, Ji D 2015 Chinese time information extraction based on CRF and semi-supervised learning Computer Engineering and Design 2015(6) 1642-6

[10] Wu Q 2015 Research on Automatic Recognition of Chinese Time Expression (Dalian, China: Dalian University of Technology)

[11] Collobert R and Weston J 2008 A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning,” in Proceedings of the 25th International Conference on Machine Learning (New York, NY, USA) 160–167

[12] Kavukcuoglu K and Kuksa P 2011 Natural Language Processing (Almost) from Scratch J. Mach. Learn. Res 12(1) 2493–2537

[13] Chen Y, Xu L, Liu K, et al. 2015 Event extraction via dynamic multi-pooling convolutional neural networks. In ACL/JCNLP

[14] Nguyen T H, Cho K, Grishman R 2016 Joint Event Extraction via Recurrent Neural Networks Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies pp 300-309

[15] Cho K, Van Merrienboer B, Gul Cehre C, et al. 2014 Learning phrase representations using RNN encoder-decoder for statistical machine translation Proceedings of the 2014 Conference on Empirical Methods in Natural Language (Doha: CoRR) pp 1724-1734

[16] Bahdanau D, Cho K, Bengio Y. 2014 Neural Machine Translation by Jointly Learning to Align and Translate Computer Science arXiv:1409.0473

[17] Yang Z, Yang D, Dyer C, et al. 2016 Hierarchical Attention Networks for Document Classification Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

[18] Mnih V, Heess N, Graves A, et al. 2014 Recurrent Models of Visual Attention Advances in Neural Information Processing Systems, arXiv:1406.6247v1

[19] Vaswani A, Shazeer N, Parmar N, et al 2017 Attention Is All You Need Advances in Neural Information Processing Systems 2017 5998-6008