An Effective Gated Recurrent Unit Network Model for Chinese Relation Extraction

Qian-qian ZHANG∗, Meng-dong CHEN and Lian-zhong LIU
School of Computer Science and Engineering, Beihang University, Beijing, China, 100191
∗Corresponding author

Keywords: Relation extraction, Attention, Gated recurrent unit.

Abstract. Relation extraction becomes one important task in information extraction and it is becoming more and more used to discover relations from text. However, it is difficult to combine context information for relation extraction and some important information may appear many times in one sentence. In order to solve these problems, this paper presents Gated Recurrent Unit Networks with attention (CHGRU) model to catch important information in a sentence. In this model, we use the gated unit networks for embedding sentence semantics. Then, we build word-level attention to capture informative sentences with accurate relation patterns. On real-world datasets, our experimental results gains great improvements about relation extraction as compared with others.

Introduction

Researchers create many large knowledge databases to save structured information for the present society in recent years. Knowledge databases have significant effect in many AI and NLP applications, for example, web search and intelligent communication. < Entity1, Relation, Entity2 > is the form of triplets about KBs’s facts. The existing KBs contain a large number of facts, but they are still not complete, meanwhile manual annotation of these knowledge cost a lot of artificial and material resources, many efforts have been invested in automating extraction of novel facts from various network resources. So, extracting relational operations from network resources is a significant work in Natural Language Processing.

The traditional method of extracting manual features from the vocabulary resources is usually based on pattern matching, these methods have gained good improvements [1-3]. One drawback of these traditional methods is that most of the features are derived from Natural Language Processing (NLP) tools, for example, part of speech tagging. These tools may generate errors and these errors will propagate in the methods of relation extraction. Another weakness is that the manual design function is time-consuming, and the generalization ability is poor because of the low coverage of different training sets.

Some of the recent researchers try to make use of deep learning methods for relational extraction, but no existing suitable characteristics [4,5]. The method of establishing the sentence level annotation data based classifier, due to the lack of training data annotated, cannot be applied to large-scale KBS.

We propose a neural network CHGRU for Chinese relation extraction in this paper. The model obtains the most important semantic data in the sentence by using neutral attention mechanism with Gated Recurrent Unit Networks (GRU). Firstly, we embed the semantics of sentences by using a GRU. Then, we express the relation as semantic composition of sentence embeddings to utilize all informative sentences. Finally, we extract the relationship between the relation vectors of the word level concern. We evaluate this method in the task of the real world data set. The final results state clearly that the experiment has made some progress.

The paper’s contribution is to use the neutral attention mechanism with Gated Recurrent Unit Networks (GRU), which can automatically concentrate on the words that play a crucial role on classification in a sentence. The model can take advantage of each entity pair’s all informative sentences.
Related Work

A lot of researchers are doing research on relationship extraction. They used a variety of methods to carry out the experiment. Some of them require large amounts of annotated data and are based on pattern matching to derive lexical features. To address the annotated data, Mintz calibrates pure text with Freebase base [6]. Hoffmann uses multiple examples and multiple label learning relationships to extract [2]. Multi-instance learning takes into account the reliability of the labels for each instance. However, all methods based features are extremely dependent on the level of the features created by the Natural Language Processing tool, which will be affected by the problem of error propagation.

In recent years, deep neural networks have been used to learn the underlying features automatically and have been widely applied in many fields. Zeng combines neural network model with at-least-one multi-instance to extract the factual relations by distant supervision data [7]. But they suppose, for each entity pair, only one sentence is active, that will miss a lot of data.

One related work was proposed by Lin, which utilizes sentence-level attention based convolutional neural networks (CNN) over multiple instance [8]. The model can utilize all informative sentences. While convolutional neural networks is inappropriate to learn long-distance semantic information, so our model uses Gated Recurrent Unit (GRU). Another related work was presented by Bahdanau, which is the first work to use the attention mechanism in NLP [9]. It links the expressions learned by each word in the source language to the words that are currently predicted to be translated, and such links are made through the attention they designed. The attention-based models have been used to many areas, for example, relation extraction, image classification and speech recognition.

Gated Recurrent Unit Network Model

We will propose CHGRU model in this section. We show the model in Figure 1, it contains three main parts:

1. **Embedding layer**: transform raw words of the sentence into low dimensional vectors;
2. **GRU layer**: utilize GRU to get high level features from embedding layer;
(3) Attention layer: calculate the weights of each feature, and then the weights are added to each feature. The greater the weight, the greater the contribution of the feature to the current recognition.

Word Embedding

Each input word in the sentence is transformed into a vector via word embeddings and position embeddings.

Word embeddings capture the word’s syntactic meanings. Given a sentence \( m \) that contains \( N \) words, \( T = \{m_1, m_2, ..., m_N\} \), every word \( m_i \) is converted into a real-valued vector \( x_i \). Column vectors in an embedding matrix \( Q \in P^{d_w \times |v|} \) encodes word representations, where \( Q \) is a fixed-sized vocabulary, and \( d_w \) is a parameter of word embedding.

Position embeddings specify the location information of two target entities. It is defined as the relative distance from the current word to the head or tail of the combined entity. By this way, it can represent the input sentence as a vector sequence \( s = \{s_1, s_2, ..., s_T\} \) with \( s_t \in P^T \), where \( t = t^a + t^b \times 2 \) (\( t^a \) and \( t^b \) are the dimensions of word embeddings and position embeddings respectively).

GRU Layer

The recursive neural network recently in Artificial intelligence field shows promising results, especially when the output and input are fixed length. To a certain extent, RNN can’t overcome gradient vanishing problem. We know that although RNNs can theoretically support a long sequence, it is very difficult to train this network. The gated recurrent units allows RNNs to have more persistent memory so that it can support longer sequences. Cho propose a gated recurrent unit (GRU)[9]. In the same corresponding unit, the GRU gate control unit adjusts the inside of the device, the information flow, but there is no single memory cell.

The Four basic operational states of GRU are described by the following formula:

\[
\begin{align*}
\tilde{z}_t^j &= \sigma(W_z^j x_t + U_z^j h_{t-1}^j), \quad \text{(Update gate)} \quad (1) \\
\tilde{r}_t^j &= \sigma(W_r^j x_t + U_r^j h_{t-1}^j), \quad \text{(Reset gate)} \quad (2) \\
\tilde{h}_t^j &= \tanh(W_x^j x_t + U(\tilde{r}_t^{1:3} \odot h_{t-1}^{1:3})), \quad \text{(New memory)} \quad (3) \\
h_t^j &= (1 - \tilde{z}_t^j)h_{t-1}^j + \tilde{z}_t^j \tilde{h}_t^j, \quad \text{(Hidden state)} \quad (4)
\end{align*}
\]

New memory: A new memory \( \tilde{h}_t^j \) is derived from the implicit state of the past \( h_{t-1}^j \) and the new input \( x_t \). That is to say, at this stage, we can reasonably merge the newly observed information (word) and the hidden state of history \( h_{t-1}^j \), and a context vector \( h_{t-1}^j \) is used to sum up the state of fusion of the new word.

Reset gate: The reset signal \( \tilde{r}_t^j \) will determine the importance of the result \( h_{t-1} \). If \( h_{t-1} \) is not related to the calculation of the new memory, the reset gate can completely eliminate the hidden layer information (state) of the past.

Update gate: The update signal \( \tilde{z}_t^j \) will determine \( h_{t-1} \) the extent to which the next state is passed. For example, if \( \tilde{z}_t^j \approx 1 \), \( h_{t-1} \) is almost completely passed to \( h_t \). On the contrary, if \( \tilde{z}_t^j \approx 0 \), the new forward direction \( \tilde{h}_t^j \) is passed to the next layer of hidden layer.

Hidden state: The implicit layer state \( h_t \) is eventually produced using the past hidden layer input \( h_{t-1} \). The new memory will produce \( \tilde{h}_t^j \) based on the decision area of the update door.

In this model, we use a bidirectional Gated Recurrent Unit network whose hidden states form the contextual word embeddings, that is \( h_t = \tilde{h}_t \oplus \tilde{h}_t \), where \( \oplus \) denotes vector concatenation and \( \tilde{h}_t \) and \( \tilde{h}_t \) denote forward and backward contextual embeddings from the respective recurrent networks. At each time node \( t \), the network has two layers of neurons, one spreading from the left to the right, and
the other from right to left. In order to ensure that t has two layers of hidden layers at any time, the
network needs to consume two times the amount of storage to store parameters such as weight and
bias.

**Attention Layer**

The Attention model is originally applied to image recognition. When a person looks at an image, the
focus of the eye moves on a different object. When the neural network identifies the image or
language, the recognition is more accurate each time it is concentrated on the part of the feature. How
to measure the importance of characteristics? The most intuitive way is weight. Therefore, the result
of Attention model is to calculate the weight of each feature at first and then add the weight to each
feature. The larger the weight, the greater the contribution of the feature to the current
recognition[10]. We use the attention mechanism like Zhou [11] in this section.

The expression of relation s is made up of a weighted sum of these output vectors:

\[ M = \tanh(H) \quad (5) \]

\[ \alpha = \text{softmax}(w^TM) \quad (6) \]

\[ s = H\alpha^T \quad (7) \]

where \( H \in \mathbb{R}^{d_w \times n} \), \( w \) is trained parameter vector, \( d_w \) is the word vector dimensions and \( w^T \) is a
transpose.

We obtain the final sentence for relation extraction of representation:

\[ h^* = \tanh(s) \quad (8) \]

Lastly, we use a softmax layer to define the conditional probability \( p(y|S, \theta) \):

\[ p(y|S) = \text{softmax}(W^S h^* + b^{(S)}) \quad (9) \]

**How to optimize?**

We will introduce the optimization and learning details in this section. The cost function is the
negative log-likelihood of the true class labels \( p(y|S) \):

\[ J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} t_i \log(y_i) + \lambda||\theta||_F^2 \quad (10) \]

Where \( y \) is the estimated probability for each class by softmax, \( t \) is the one-hot represented ground
truth, and \( \lambda \) is learning rate.

We use dropout on the output layer to prevent overfitting in the implementation [12]. The dropout
layer is defined as an meta multiplication of a vector \( h \) with Bernoulli random variable \( p \).

**Experiments**

We introduce the experiment used the datasets and evaluation firstly. Then, we apply contrast
validation to determine the parameters of the model. Finally, we compare with the existing methods.

We use a Chinese dataset which is developed by Lin [13] to evaluate this model. This dataset was
generated by aligning Wikidata with Baidu Baike. Wikidata relational fact of the data set consists of
three parts, one is for training, another is for validation and the last one is for testing. Our model only
use testing and training dataset. The training data contains 940,595 sentences and 42,536 relational
facts. the testing set contains 167,224 sentences and 4,326 relational facts.

We evaluate this method in the held-out evaluation like previous work. The method was evaluated
by comparing the facts found in the test system and the facts in KB. In experiments, we record both
F1 value and the aggregate curves precision/recall curves.
Our model was trained by using a sliding window size 3, a learning rate of 0.01, the sentence embedding size 230 and a batch size 160. The other parameters in this model are random initialization.

In order to evaluate this approach, we choose the five characteristics based methods for comparative evaluation:

Mintz is a conventional distant supervised method [6].

Hoffman proposes a new method of learning overlapping relations with multi instance. This method combines sentence level extraction model with simple corpus level components to summarize individual facts [2].

MIMLRE learns all instances which a pair of entities include to train. This model can get different results by different expressions [3].

CNN+ATT, PCNN+ATT [8] proposes sentence-level attention over multiple instances. We get these above data from [8].

Table 1. Comparison F1 with previous results.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM (2010, Rink and Harabagiu)</td>
<td>82.2</td>
</tr>
<tr>
<td>MVRNN (2012, Socher et al.)</td>
<td>82.4</td>
</tr>
<tr>
<td>CNN+ Softmax (2014, Zeng et al.)</td>
<td>82.7</td>
</tr>
<tr>
<td>CRCNN (2015, Santos et al.)</td>
<td>84.1</td>
</tr>
<tr>
<td>ATT-BLSTM(2016, Zhou et al.)</td>
<td>84.0</td>
</tr>
<tr>
<td>ATT-CHGRU</td>
<td>86.7</td>
</tr>
</tbody>
</table>

Table 1 compares our CHGRU with other relation extraction’s methods. The CHGRU model achieves an F1-score of 86.7%. It has a higher level of functionality than many existing competitive methods.

**Conclusions**

In this paper, we propose a gated recurrent unit model with attention mechanism, named CHGRU, to extract relation. In the network, word features and position features are successfully proposed to specify the pairs of entities to a certain extent. The experimental results show that this model has achieved good performance.
References


