Chinese Punctuation Prediction with Adaptive Attention and Dependency Tree
Zhelong Yan, Jianzong Wang, Ning Cheng, Tianbo Wu and Jing Xiao
Ping An Technology (Shenzhen) Co., Ltd., China

Introduction
Punctuation prediction is one of the most important and fundamental tasks in natural language processing (NLP). Current methods are devoted to learning based algorithms and have been achieved the good performance. Now BiLSTM+CRF is one of the most popular methods in sequence labelling. However, the lack of ability of capturing long-distance interactions among words and extracting useful semantic information makes it still a difficult problem.

In our proposed network CPPAADT, adaptive multi-head self-attention and dependency parsing tree are utilized to tackle this problem thus enrich word representation largely.

Method

The architectures we mentioned in this paper can be shown as follows:

![Figure 1: The architectures](image)

**Dependency Tree**

By dependency tree, the network is able to generate multi-scale embedding by concatenating the word embedding of current word \( w_t \) and its parent word \( w_{p(t)} \) and the embedding of relationship \( r_t \) between it, which contains more useful information.

![Figure 2: How dependency parsing tree generates multi-scale embedding.](image)

\[ u_t = [w_t, w_{p(t)}, r_t] \]
\[ r_t = \text{Relation}(w_t, w_{p(t)}) \]

**Adaptive Attention**

Introducing adaptive attention in front of LSTM to enhance the ability of capturing long-distance interaction among words.

\[ \text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_d)W^O \]

\[ \text{head}_i = \text{softmax}\left(\frac{QW_i^O K^W_i}{d}\right) V^W_i \]

**Conditional random fields**

CRF layer models the relation between neighbouring labels which results better than simply predicting each label separately.

\[ \text{score}(X, Y) = \sum_{i=0}^{T} C_{y_t, y_i} + \sum_{i=1}^{T} P_{y_t, y_i} \]

By the Viterbi algorithm, CRF can calculate scores of all possible label sequence and return an optimal one with highest score.

Experiments

**Punctuation prediction**

The task of Chinese punctuation prediction can be illustrated briefly as follows. Given an input sentence of words, we label each word based on the punctuation after this word. In detail, we label each word with comma, period and blank (non-punctuation).

**Dataset**

<table>
<thead>
<tr>
<th>Name</th>
<th>Origin</th>
<th>Word's type</th>
<th># of train data</th>
<th># of test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Microsoft Research</td>
<td>88,119</td>
<td>88,925</td>
<td>3,986</td>
</tr>
<tr>
<td>B</td>
<td>Peking University</td>
<td>55,303</td>
<td>19,057</td>
<td>1,946</td>
</tr>
</tbody>
</table>

**Symbol standardization.**

Depending on rule of symbol standardization, all punctuation marks are sorted into three conventional symbols by substituting.

<table>
<thead>
<tr>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>,</td>
</tr>
<tr>
<td>;</td>
<td>;</td>
</tr>
<tr>
<td>!</td>
<td>!</td>
</tr>
<tr>
<td>(</td>
<td>)</td>
</tr>
<tr>
<td>,</td>
<td>(</td>
</tr>
<tr>
<td>;</td>
<td>)</td>
</tr>
<tr>
<td>!</td>
<td>(</td>
</tr>
<tr>
<td>(</td>
<td>(</td>
</tr>
<tr>
<td>)</td>
<td>)</td>
</tr>
</tbody>
</table>

**Results**

Our proposed CPPAADT outperforms existing methods with a gap of above 4.5% of accuracy and reaches state-of-the-art performance in two datasets.

**Conclusions**

**BiLSTM works much better than LSTM.**

The networks with adaptive attention outperform that without adaptive attention.

Different positions of adaptive attention accompanies with different characteristics.

CPPAA works better in predicting label period and BiLSTM+Attention+CRF works better in label common.

Dependency tree makes a great deal of improvement.

CPPAADT outperforms other existing methods with a gap of above 4.5% of accuracy in two datasets mentioned above.

Skip connection don’t make a significant difference.

CPPAADT, the combination of adaptive attention and dependency tree, achieves consistently best performance.

Acknowledgement

This paper is supported by National Key Research and Development Program of China under grant No. 2018YFB1003900, No. 2018YF0204400 and No. 2017YFB100202.

November 4, 2020
931766894@qq.com