Chinese Punctuation Prediction with Adaptive Attention and Dependency Tree

Zelong 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:

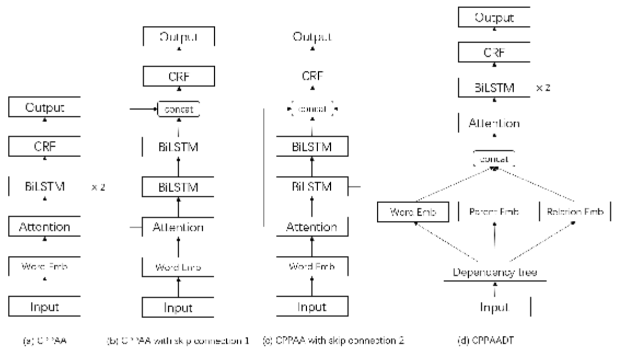

Figure 1: The architectures

▶ Dependency Tree

By dependency tree, the network is able to generate multi-scale embedding by concatenating the word embedding of current word w_t^1 and its parent word w_t^2 and the embedding of relationship r_t between it, which contains more useful information.

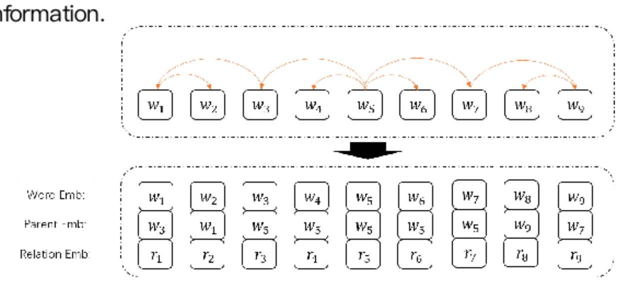

Figure 2: How dependency parsing tree generates multi-scale embedding.

$$u_t = [w_t^1, w_t^2, r_t] \quad (1)$$

$$r_t = Relation(w_t^1, w_t^2) \quad (2)$$

▶ Adaptive Attention

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

$$Multihead(Q, K, V) = Concat(head_1, ...head_n)W^O \quad (3)$$

$$head_i = softmax(\frac{QW_i^Q KW_i^K}{\sqrt{d}} VW_i^V) \quad (4)$$

▶ Conditional random fields

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

$$score(X, Y) = \sum_{i=0}^{T} C_{y_i, y_{i+1}} + \sum_{i=1}^{T} P_{i, y_i} \quad (5)$$

By the Viterbi algorithm, CRF can calculate scores of all possible label sequence and return a 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 1: Details of datasets.

Name	Origin	Word's type	# of train data	# of test data
A	Microsoft Research	88,119	86,925	3,986
B	Peking University	55,303	19,057	1,946

▶ Symbol standardization.

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

Table 2: Rule of symbol standardization.

Before	After
—” “”” (( )) 、 《》	blank( )
,, ; ;: :	comma(,)
.。 . ...! !? ?	period(.)

▶ 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.

Table 3: Experimental results on 2 datasets.

	Dataset A			Dataset B		
Model	acc	F(c)	F(p)	acc	F(c)	F(p)
$LSTM_2$+CRF	91.622	47.939	99.431	88.006	34.487	55.170
Attention+$LSTM_2$+CRF	93.171	63.901	99.446	89.577	47.173	65.022
Attention+$BiLSTM_1$+CRF	93.769	66.899	99.302	90.371	54.302	69.973
$BiLSTM_2$+CRF	93.581	67.401	98.886	90.643	54.673	70.291
$BiLSTM_2$+Attention+CRF	93.947	70.667	99.230	90.663	57.078	69.517
CPPAA	93.816	66.957	99.446	90.935	54.796	71.342
CPPAADT	**98.422**	**91.489**	**99.579**	**95.470**	**77.973**	**71.373**

Table 4: Experimental results with skip connection.

	Dataset A			Dataset B		
Model	acc	F(c)	F(p)	acc	F(c)	F(p)
CPPAA	93.816	66.957	99.446	90.935	54.796	71.342
skip connection 1	93.991	66.648	99.375	90.938	55.086	71.039
skip connection 2	94.101	67.684	99.116	90.985	57.250	**71.592**
CPPAADT	**98.422**	**91.489**	**99.579**	**95.470**	**77.973**	71.373

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 $BiLSTM_2$+Attetnion+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. 2018YFB1003500, No. 2018YFB0204400 and No. 2017YFB1401202.

931766894@qq.com