PaperDaily — 2017.09.27

Stack-based Multi-layer Attention for Transition-based Dependency Parsing

Zhirui Zhang[1], Shujie Liu[2], Mu Li[2], Ming Zhou[2], Enhong Chen[1]
University of Science and Technology of China, Hefei, China[1]
Microsoft Research Asia, Beijing, China[2]
zr011036@mail.ustc.edu.cn cheneh@ustc.edu.cn[1]
{shujie,muli,mingzhou}@microsoft.com[2]

[EMNLP17]的论文，多层Attention的思想非常有意思，而且效果相当不错，能够在我们的工作中引入。

Abstract

在transition-based dependency parsing任务上简单的应用seq2seq模型并没有像NMT、Text Summarization等任务那样获得有竞争力的结果。本文中，作者提出`stack-based multi-layer attention model`的seq2seq来更好的学习结构化语言信息。他们的方法中，两个binary向量来刻画转移序列中的decoding stack、引入多层attention来捕捉部分树中多个词的依存关系。本文在PTB和CTP上获得了基于seq2seq模型的最好的结果。

Introduction

简单的seq2seq模型的一个问题是，不能明确地使用结构化语言学信息。比如Transition-based parsing使用栈来维护部分子树的根，并基于此选择正确的转移行为。seq2seq的另一个问题是普通的attention机制不能捕捉单词间的依赖关系。传统的特征模板中使用的word unigram, bigram, trigram, ...特征很难被attention捕捉，但是它们对parsing很重要。

作者为了刻画转移系统中的栈，引入了两个binary向量，一个表示单词是否进栈，另一个表示是否出栈。为了捕捉复杂的结构信息，作者提出了基于`栈`、`生成的动作序列`、`输入句子`的多层attention。

作者在PTB和CTB上评估他们的模型，最高的结果是4个模型的ensembled。

English：94.16 (+1.87) UAS Chinese：87.97 (+1.61) UAS

Neural Model for Sequence-to-Sequence Learning

Encoder: 对于输入句子$\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$，使用BiRNN生成hidden states sequence $\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T)$。

Attention Mechanism: context向量$\boldsymbol{c}_i$是$(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T)$的加权求和：

$$\alpha_{i,t} = \frac{\exp(e_{i,t})}{\sum_k \exp(e_{i,k})}$$
$$e_{i,k} = v_a^\top \tanh(W_a z_{i-1} + U_a h_t)$$

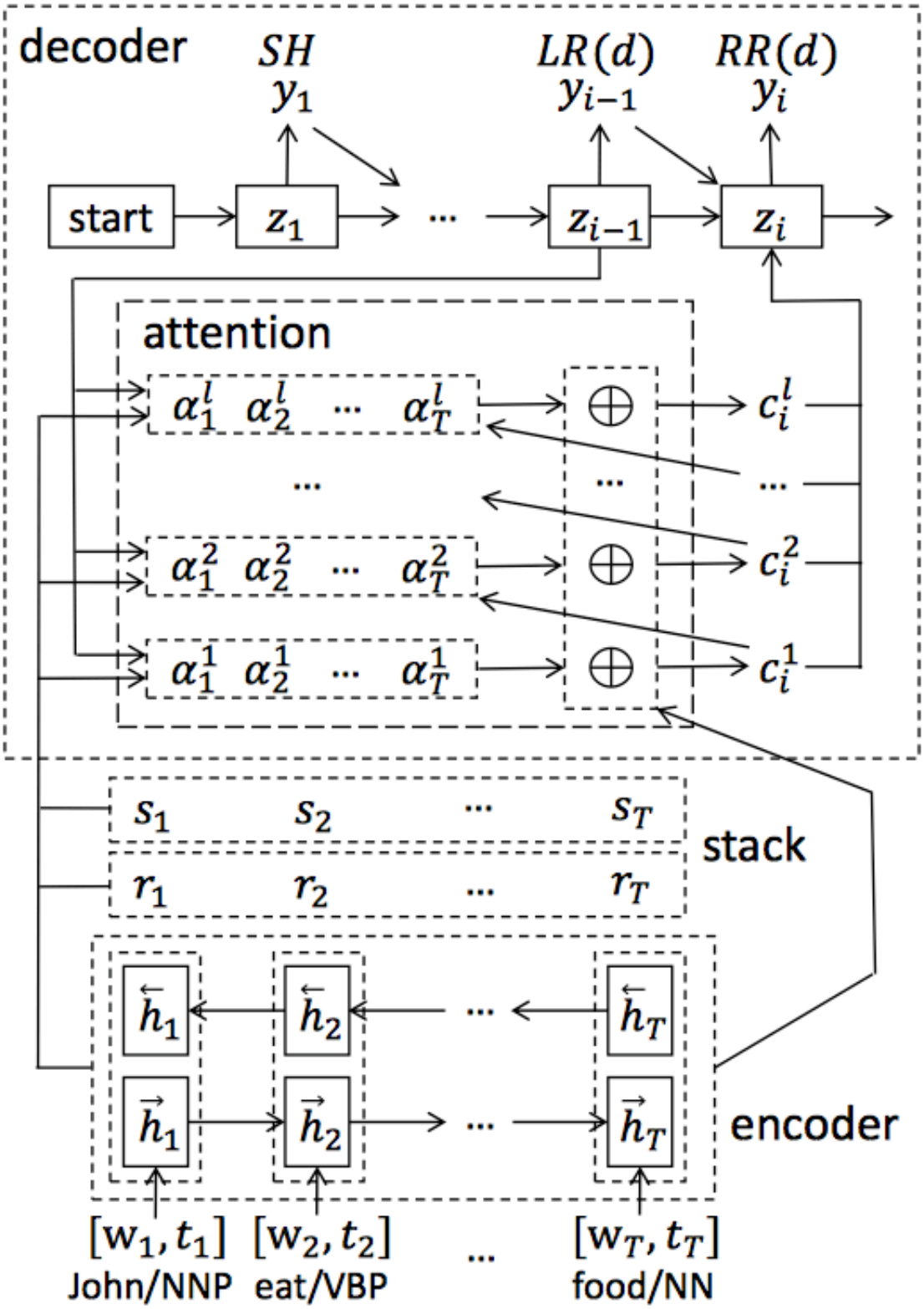
Decoder: 解码网络生成目标序列$Y = (y_1, y_2, \cdots, y_{T'})$，$y_i$的计算如下：

$$z_i = \mathrm{RNN}([y_{i-1;c_i}], z_{i-1})$$
$$p(y_i|y_{<i}, h) = \mathrm{softmax}(g(y_{i-1}, z_i, c_i))$$

g是非线性函数（激活函数），z_i是解码网络i_{th}隐层表示，y_{i-1}是i_{th}时刻生成的目标词，c_i是编码网络的上下文向量。

Sequence-to-Sequence Parsing Model

和其他任务相比，Dependency parsing不仅需要考虑输入句子和之前的操作序列，而且还需要许多结构化信息，比如parsing过程中的子树结构。但是普通的seq2seq没有明确结构来刻画这些必要的结构信息。



Encoder: 编码的输入$x_i = [W_e * e(w_i); W_t * e(t_i)]$ 包含词、词性向量。

Attention Mechanism: 栈信息由$s = (s_1, \cdots, s_T)$和$r = (r_1, \cdots, r_T)$构成。当w_i进栈时令$s_i = 1$，出栈时令 $r_i = 1$，其余状态下为0。

$$\alpha_{i,t} = \frac{\exp(e_{i,t}) * (1 - r_t)}{\sum_k \exp(e_{i,k}) * (1 - r_k)}$$
$$e_{i,k} = v_a^\top \tanh(W_a z_{i-1} + U_a h_t + S_a s_t)$$
$$e_{i,k}^m = v_a^\top \tanh(W_a^m [z_{i-1}; c_i^{m-1}] + U_a h_t + S_a s_t)$$

Decoder: 解码转移序列。

Experiments

Setup

- 3-layers GRU 用来编码和解码
- 300 word embedding dimensions
- 32 POS-tag/label embedding dimensions
- 500 hidden units in GRU
- 3-layers attention structure
- 300-dimensional pre-trained GloVe vectors
- SGD algorithm
- 64 mini-batch size
- dropout rate is 0.2
- For testing, beam search with beam size 8

Main Results

Parser	PTB-SD				CTB			
	Dev		Test		Dev		Test	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
Z&N11	-	-	93.00	90.95	-	-	86.00	84.40
C&M14	92.20	89.70	91.80	89.60	84.00	82.40	83.90	82.40
ConBSO	-	-	91.57	87.26	-	-	-	-
Dyer15	93.20	90.90	93.10	90.90	87.20	85.90	87.20	85.70
Weiss15	-	-	93.99	92.05	-	-	-	-
K&G16	-	-	93.99	91.90	-	-	87.60	86.10
DENSE	**94.30**	91.95	94.10	91.90	87.35	85.85	87.84	86.15
seq2seq	92.02	89.10	91.84	88.84	86.21	83.80	85.80	83.53
Our model	93.65	91.52	93.71	91.60	87.28	85.30	87.41	85.40
Ensemble	94.24	**92.01**	**94.16**	**92.13**	**88.06**	**86.30**	**87.97**	**86.18**

- 本文的模型在seq2seq方法中获得了最高的效果
- 4个不同随机初始化的模型做ensemble才超越了之前先进的parser，但是和最近最好的parser仍然有很大差距。
- ensemble之后的性能才和Head selection（DENSE）相当。

	Dev		Test	
	UAS	LAS	UAS	LAS
seq2seq	92.02	89.10	91.84	88.84
$l = 1$	92.85	90.44	92.70	90.40
$l = 2$	93.30	91.13	93.21	90.98
$l = 3$	**93.65**	**91.52**	**93.71**	**91.60**
$l = 4$	93.49	91.29	93.42	91.24

- 增加attention的层数对parsing是有帮助的。3层attention的效果最好。
- attention层数的增加并不是无限制的，会增加训练、预测的时间以及过拟合的风险。

	Dev		Test	
	UAS	LAS	UAS	LAS
Our model	93.65	91.52	93.71	91.60
–pretraining	93.19	90.92	93.22	91.11
–POS	92.73	89.86	92.57	90.05
$-s$ vector	93.18	90.68	93.02	90.89
$-r$ vector	93.16	90.90	93.27	91.02

- 由于没有使用char-embedding，模型比较依赖POS
- 栈信息的引入是有效的