Constituency Parsing with a Self-Attentive Encoder

Nikita Kitaev, Dan Klein

Berkeley -- ACL18

AntNLP -- Tao Ji

taoji.cs@gmail.com

Outline

- Motivation
- Architecture
  - Self-Attention
  - Position-Wise Feed-Forward Sublayer
  - Span Scores
  - Content vs. Position Attention
  - Lexical Models
- Experiments

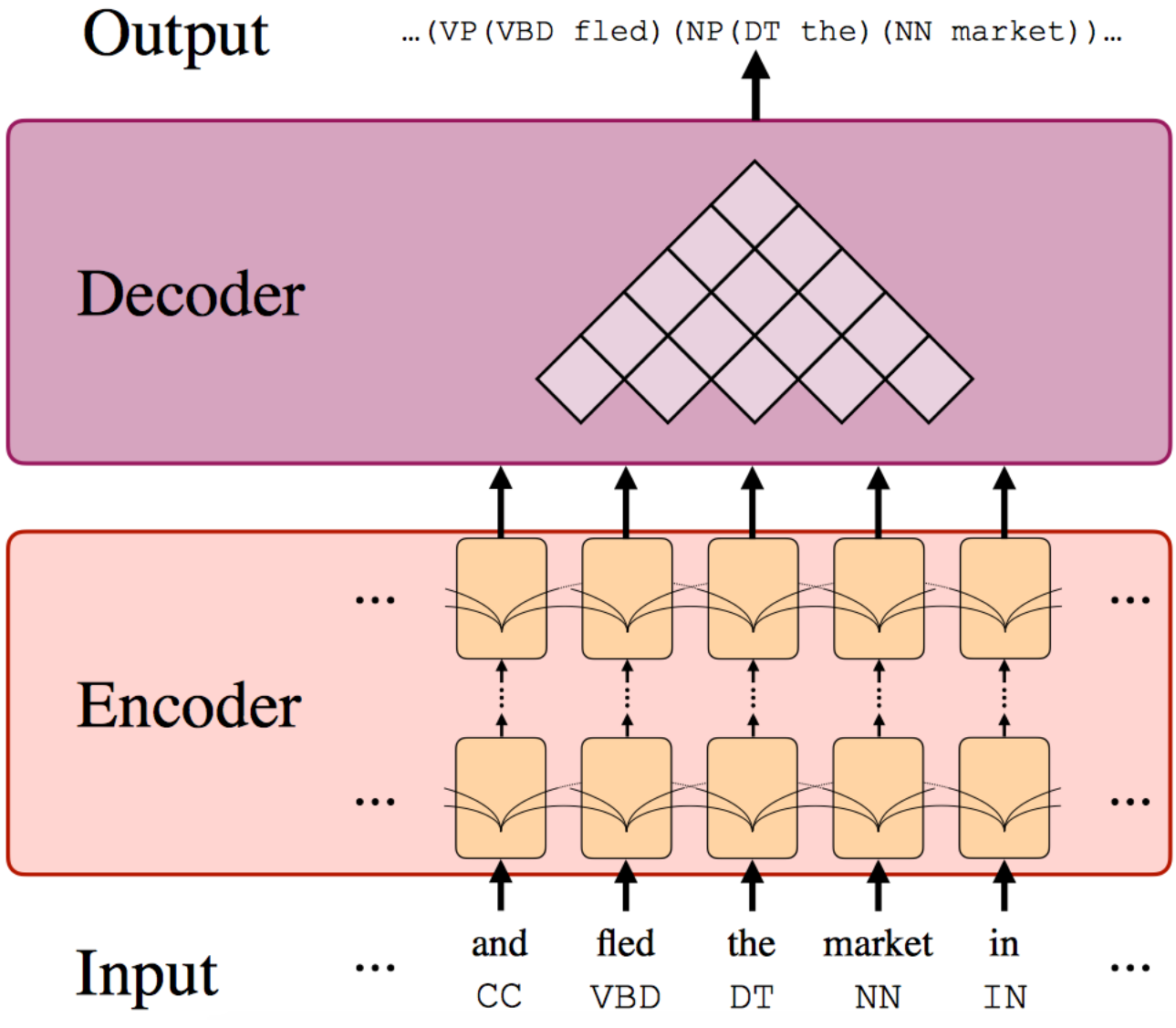
Motivation

- They demonstrate that replacing an LSTM encoder with a self-attentive architecture

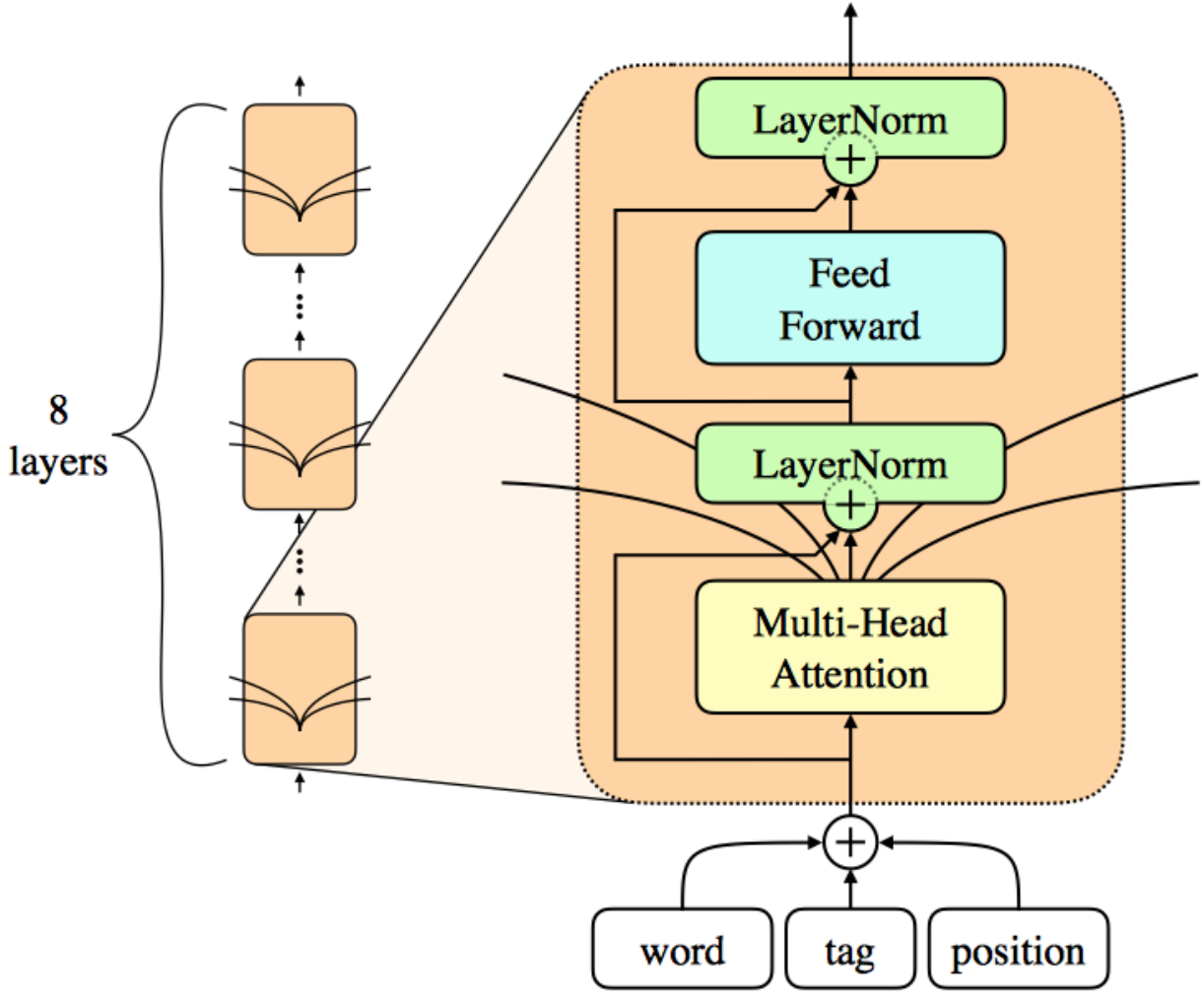
Contribution

- The use of attention makes explicit the manner in which information is propagated between different locations in the sentence.
- Explicitly factoring the two types of attention can noticeably improve parsing accuracy.
- Morphological (or at least sub-word) features to be important to achieving good results.

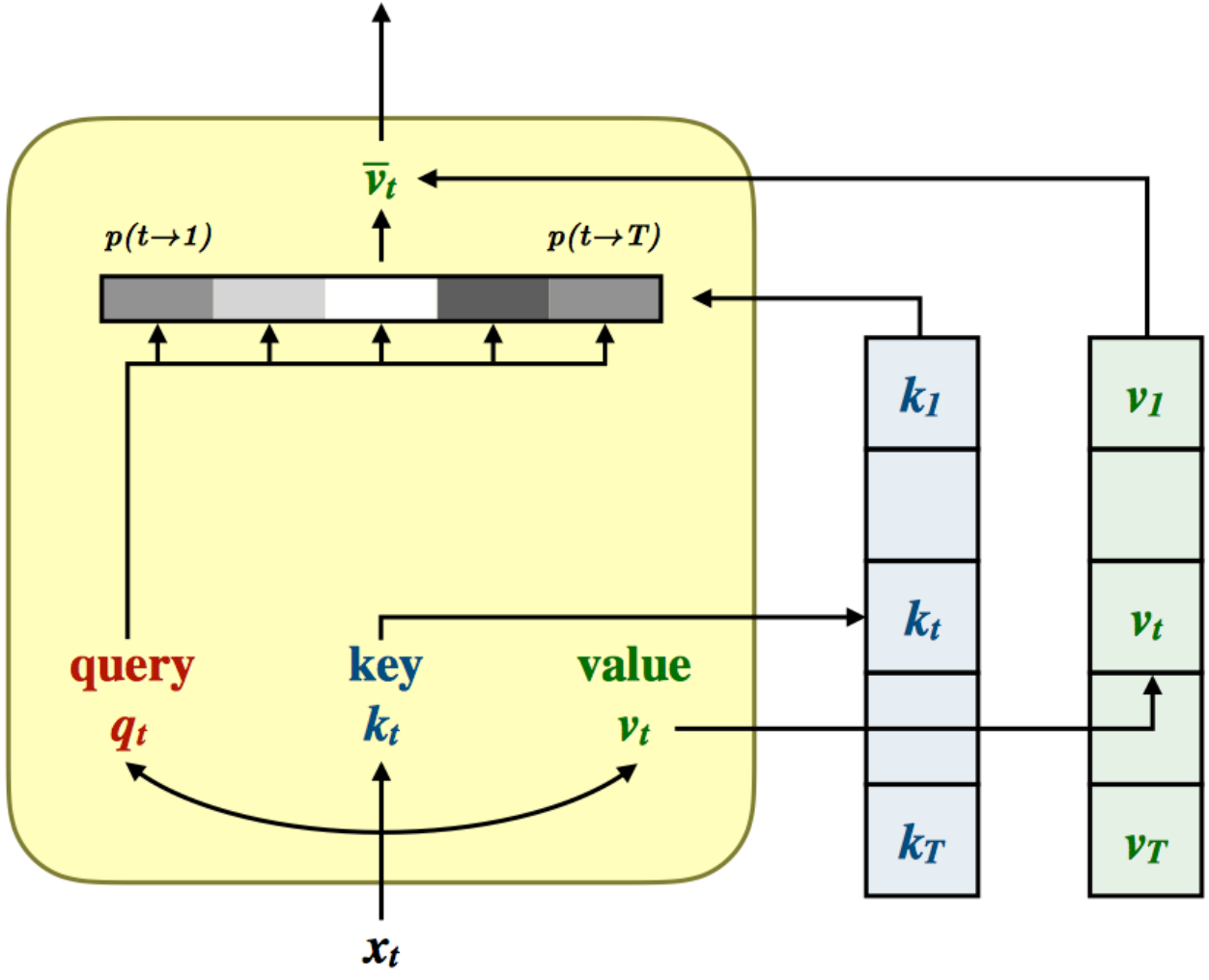
Architecture



Architecture



Architecture



Architecture

Self-Attention

$$\text{SingleHead}(X) = \left[\text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V\right]W_O$$

$$\text{where } Q = XW_Q;\ K = XW_K;\ V = XW_V$$

Rather than using a single head, our model sums together the outputs from multiple heads:

$$\text{MultiHead}(X) = \sum_{n=1}^{8} \text{SingleHead}^{(n)}(X)$$

Architecture

Position-Wise Feed-Forward Sublayer

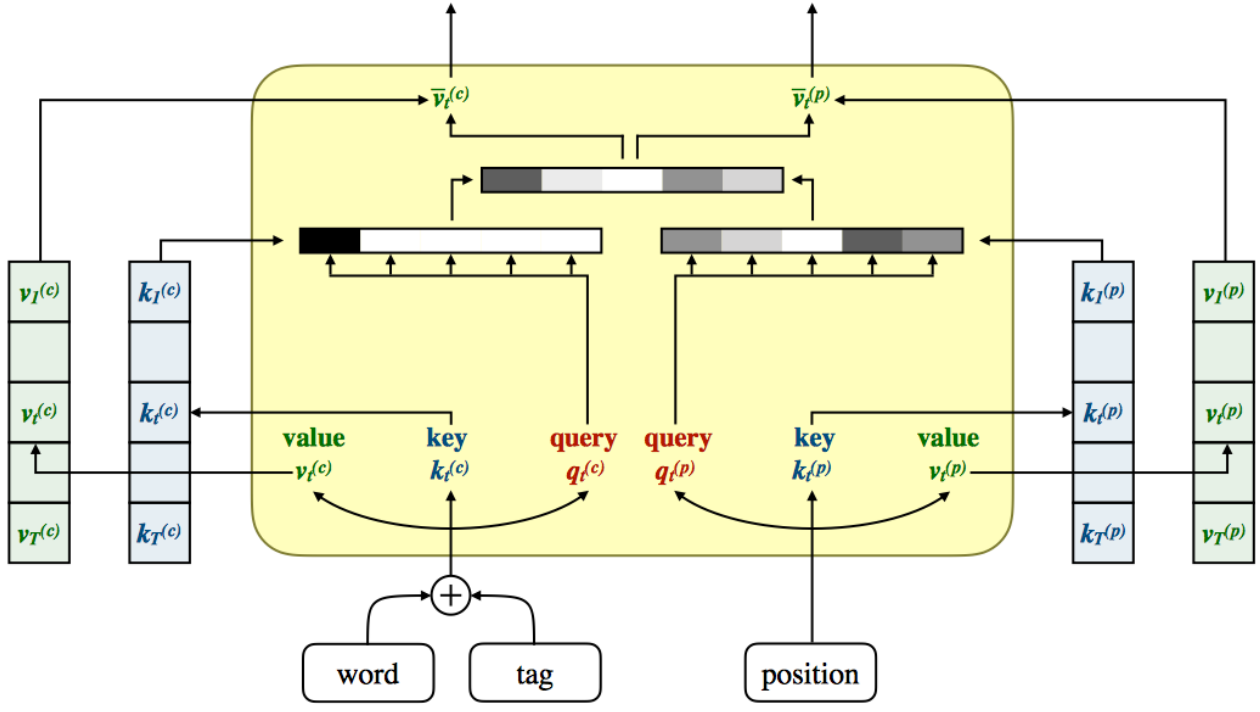
$$\text{FeedForward}(x) = W_2\textbf{relu}(W_1 x + b_1) + b_2$$

Span Scores

$$s(i, j, \cdot) = M_2\textbf{relu}(\text{LayerNorm}(M_1 v + c_1)) + c_2$$

$$[\overrightarrow{y}_j - \overrightarrow{y}_i; \overleftarrow{y}_{j+1} - \overleftarrow{y}_{i+1}]$$

Architecture



Experiments

Attention		
Content	Position	F1
All 8 layers	All 8 layers	93.15
All 8 layers	Disabled	72.45
Disabled	All 8 layers	90.84
First 4 layers only	All 8 layers	91.77
Last 4 layers only	All 8 layers	92.82
First 6 layers only	All 8 layers	92.42
Last 6 layers only	All 8 layers	92.90

Experiments

Distance	F1 (strict)	F1 (relaxed)
5	92.74	92.94
10	92.92	93.00
20	93.06	93.17
∞	93.15	

	Word embeddings	
	✓	✗
None	92.20	–
Tags	93.15	–
CharLSTM	93.40	93.61
CharConcat	93.32	93.35

Experiments

Encoder Architecture	F1 (dev)	Δ
LSTM (Gaddy et al., 2018)	92.24	-0.43
Self-attentive (Section 2)	92.67	0.00
+ Factored (Section 3)	93.15	0.48
+ CharLSTM (Section 5.1)	93.61	0.94
+ ELMo (Section 5.2)	95.21	2.54

Experiments

	LR	LP	F1
Single model, WSJ only			
Vinyals et al. (2015)	–	–	88.3
Cross and Huang (2016)	90.5	92.1	91.3
Gaddy et al. (2018)	91.76	92.41	92.08
Stern et al. (2017b)	92.57	92.56	92.56
Ours (CharLSTM)	**93.20**	**93.90**	**93.55**
Multi-model/External			
Durrett and Klein (2015)	–	–	91.1
Vinyals et al. (2015)	–	–	92.8
Dyer et al. (2016)	–	–	93.3
Choe and Charniak (2016)	–	–	93.8
Liu and Zhang (2017)	–	–	94.2
Fried et al. (2017)	–	–	94.66
Ours (ELMo)	94.85	95.40	**95.13**

Experiments

	Arabic	Basque	French	German	Hebrew	Hungarian	Korean	Polish	Swedish	Avg
Dev (all lengths)										
Coavoux and Crabbé (2017)	83.07	88.35	82.35	88.75	90.34	91.22	86.78[b]	**94.0**	79.64	87.16
Ours (CharLSTM only)	**85.94**	**90.05**	84.27	91.26	90.50	92.23	**87.90**	93.94	79.34	**88.38**
Ours (CharLSTM + word embeddings)	85.59	89.31	**84.42**	**91.39**	**90.78**	**92.32**	87.62	93.76	**79.71**	88.32
Test (all lengths)										
Björkelund et al. (2014), ensemble	81.32[a]	88.24	82.53	81.66	89.80	91.72	83.81	90.50	**85.50**	86.12
Cross and Huang (2016)	–	–	83.31	–	–	–	–	–	–	–
Coavoux and Crabbé (2017)	82.92[b]	88.81	82.49	85.34	89.87	92.34	86.04	93.64	84.0	87.27
Ours (model selected on dev)	**85.61**	**89.71**	**84.06**	**87.69**	**90.35**	**92.69**	**86.59**	**93.69**	84.45	**88.32**
Δ: Ours - Best Previous	**+2.69**	**+0.90**	**+0.75**	**+2.35**	**+0.48**	**+0.35**	**+0.55**	**+0.05**	-1.05	

Experiments

Symbol	Description	Value
N	Number of layers (when not using ELMo embeddings)	8
N	Number of layers (when using ELMo embeddings)	4
d_{model}	Model dimensionality	1024
h	Number of attention heads	8
d_k	Size of attention query/key vectors	64
d_v	Size of attention value vectors	64
d_{ff}	Size of intermediate vectors in the feed-forward sublayer	2048
	Size of character embeddings (CharConcat)	32
	Size of character embeddings (CharLSTM)	64
	Attention dropout probability; see Vaswani et al. (2017)	0.2
	ReLU dropout probability in feed-forward sublayer	0.1
	Residual dropout probability (at all residual connections)	0.2
	Word embedding dropout probability	0.4
	Dropout probability for part-of-speech tag embeddings	0.2
	Dropout probability for CharConcat/CharLSTM morphological representations	0.2
	Character embedding dropout probability at the inputs to CharLSTM	0.2

Thank you!

Q&A