Paper Reading: Linguistically-Informed Self-Attention for Semantic Role Labeling

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss & Andrew McCallum

University of Massachusetts Amherst & Google AI Language

Motivation

- Semantic role labeling(SRL) uses a deep neural network with no linguistic features. Modeling syntax has proven benefit for SRL task.
- Propose a linguistically-informed self-attention(LISA): a neural network model that combines multi-head self-attention with multi-task learning across dependency parsing, part-of-speech tagging, predicate detection and SRL.

LISA model

Linguistically-informed self-attention(LISA): a model that combines multi-task learning with stacked layers of multi-head self-attention, the model is trained to:

- 1) jointly predict parts of speech and predicates
- 2) perform parsing
- 3) attend to syntactic parse parents
- 4) assigning semantic role labels

Model Architecture

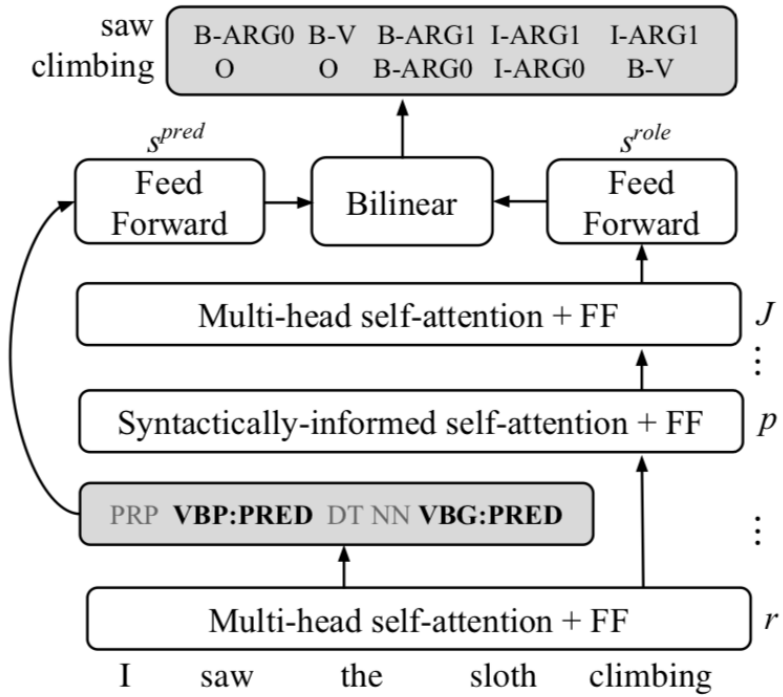

Figure 1: Word embeddings are input to J layers of multi-head self-attention. In layer p one attention head is trained to attend to parse parents (Figure 2). Layer r is input for a joint predicate/POS classifier. Representations from layer r corresponding to predicted predicates are passed to a bilinear operation scoring distinct predicate and role representations to produce per-token SRL predictions with respect to each predicted predicate.

Syntactically-informed self-attention

Replace one attention head with deep bi-affine model(Dozat and Manning et, at. 2017), trained to predict syntactic dependencies. Let A_{parse} be the parse attention weights:

$$A_{parse} = \text{softmax}(Q_{parse} U_{heads} K_{parse}^T)$$

Syntactically-informed self-attention

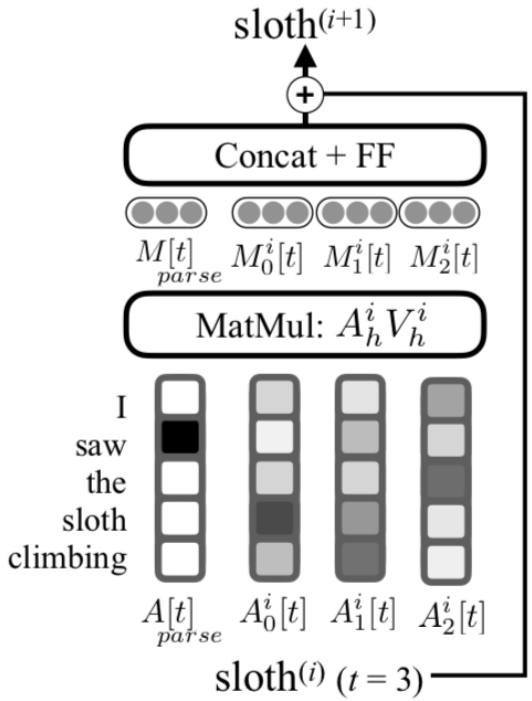

Figure 2: Syntactically-informed self-attention for the query word *sloth*. Attention weights A_{parse} heavily weight the token's syntactic governor, *saw*, in a weighted average over the token values V_{parse}. The other attention heads act as usual, and the attended representations from all heads are concatenated and projected through a feed-forward layer to produce the syntactically-informed representation for *sloth*.

Multi-task learning:

For each POS tag TAG which is observed co-occurring with a predicate, we add a label of the form TAG: PREDICATE.

Predicting semantic roles:

$$s_{ft} = (s_f^{pred})^T U s_t^{role}$$

Training

In order to maximize the model's ability to leverage syntax, during training we clamp $\mathcal{P}$ to gold parse($\mathcal{P}_\mathcal{G}$) and $\mathcal{V}$ to gold predicates $\mathcal{V}_\mathcal{G}$ when passing parse and predicate representations to later layers.

$$\frac{1}{T}\sum_{t=1}^{T}\Big[\sum_{f=1}^{F}\log P(y_{ft}^{role} \mid \mathcal{P}_G, \mathcal{V}_G, \mathcal{X}) + \log P(y_t^{prp} \mid \mathcal{X}) + \lambda_1 \log P(\text{head}(t) \mid \mathcal{X}) + \lambda_2 \log P(y_t^{dep} \mid \mathcal{P}_G, \mathcal{X})\Big] \quad (7)$$

where λ_1 and λ_2 are penalties on the syntactic attention loss.

Experiment Result

	Dev			WSJ Test			Brown Test		
GloVe	P	R	F1	P	R	F1	P	R	F1
He et al. (2017) PoE	81.8	81.2	81.5	82.0	83.4	82.7	69.7	70.5	70.1
He et al. (2018)	81.3	81.9	81.6	81.2	83.9	82.5	69.7	71.9	70.8
SA	83.52	81.28	82.39	84.17	83.28	83.72	72.98	70.1	71.51
LISA	83.1	81.39	82.24	84.07	83.16	83.61	73.32	70.56	71.91
+D&M	**84.59**	**82.59**	**83.58**	**85.53**	**84.45**	**84.99**	**75.8**	**73.54**	**74.66**
+*Gold*	*87.91*	*85.73*	*86.81*	—	—	—	—	—	—
ELMo									
He et al. (2018)	84.9	**85.7**	85.3	84.8	**87.2**	86.0	73.9	**78.4**	76.1
SA	85.78	84.74	85.26	86.21	85.98	86.09	77.1	75.61	76.35
LISA	**86.07**	84.64	**85.35**	86.69	86.42	86.55	78.95	77.17	78.05
+D&M	85.83	84.51	85.17	**87.13**	86.67	**86.90**	**79.02**	77.49	**78.25**
+*Gold*	*88.51*	*86.77*	*87.63*	—	—	—	—	—	—

Table 1: Precision, recall and F1 on the CoNLL-2005 development and test sets.

Parsing, POS and predicate detection

Data	Model	POS	UAS	LAS
WSJ	D&M$_E$	—	96.48	94.40
	LISA$_G$	96.92	94.92	91.87
	LISA$_E$	97.80	96.28	93.65
Brown	D&M$_E$	—	92.56	88.52
	LISA$_G$	94.26	90.31	85.82
	LISA$_E$	95.77	93.36	88.75
CoNLL-12	D&M$_E$	—	94.99	92.59
	LISA$_G$	96.81	93.35	90.42
	LISA$_E$	98.11	94.84	92.23

Table 4: Parsing (labeled and unlabeled attachment) and POS accuracies attained by the models used in SRL experiments on test datasets. Subscript G denotes GloVe and E ELMo embeddings.

Parsing, POS and predicate detection

	Model	P	R	F1
WSJ	He et al. (2017)	94.5	98.5	96.4
	LISA	98.9	97.9	98.4
Brown	He et al. (2017)	89.3	95.7	92.4
	LISA	95.5	91.9	93.7
CoNLL-12	LISA	99.8	94.7	97.2

Table 5: Predicate detection precision, recall and F1 on CoNLL-2005 and CoNLL-2012 test sets.

Conclusion

- Linguistically-informed self-attention
- End to end framework: multi-task learning
- A natural way to introduce linguistic features