LISA

linguistically-informed self-attention for semantic role labeling

서상우

LISA

- EMNLP 2018 best paper

- Emma Strubell
  - Natural Language Processing, Machine Learning
  - SRL, slop filing, relation extraction…

- University of Massachusetts Amhers and Google AI Language

SRL (Semantic Role Labeling) 이란?

- 술어(predicate)-논항(argument) 구조를 발견하는 것을 목표로 한다.
- 목표 동사(술어)에 대해, 동사의 의미역을 취하는 문장의 모든 구성요소가 인식
  - 의미 논항은 행위주, 대상, 도구 등이며 위치, 시간, 방법, 원인 등도 포함

- SRL = detecting basic event structures such as who did what to whom, when and where

SRL : CONLL 2005

- Sentence
  - The $1.4 billion robot spacecraft faces a six-year journey to explore Jupiter and its 16 known moons.

The	-	(A0*	(A0*
$	-	*	*
1.4	-	*	*
billion	-	*	*
robot	-	*	*
spacecraft	-	*)	*)
faces	face	(V*)	*
a	-	(A1*	*
six-year	-	*	*
journey	-	*	*
to	-	*	*
explore	explore	*	(V*)
Jupiter	-	*	(A1*
and	-	*	*
its	-	*	*
16	-	*	*
known	-	*	*
moons	-	*)	*)
.	-	*	*

SRL : CONLL 2005

- WORDS. The words of the sentence.
- NE. Named Entities.
- POS. PoS tags.
- PARTIAL SYNT. Partial syntax, namely chunks (1st column) and clauses (2nd column).
- FULL SYNT. Full syntactic tree. Note that this column represents the following WSJ tree:

```
(S
   (NP (DT The)
     (ADJP
       (QP ($ $) (CD 1.4) (CD billion) ))
     (NN robot) (NN spacecraft) )
   (VP (VBZ faces)
     (NP (DT a) (JJ six-year) (NN journey)
       (S
         (VP (TO to)
           (VP (VB explore)
             (NP
               (NP (NNP Jupiter) )
               (CC and)
               (NP (PRP$ its) (CD 16) (JJ known) (NNS moons) )))))))
   (. .) )
```

- VS. VerbNet sense of target verbs. These are hand-crafted annotations that will be available only for training and development sets (not for the test set).
- TARGETS. The target verbs of the sentence, in infinitive form.
- PROPS. For each target verb, a column reprenting the arguments of the target verb.

SRL : CONLL 2005

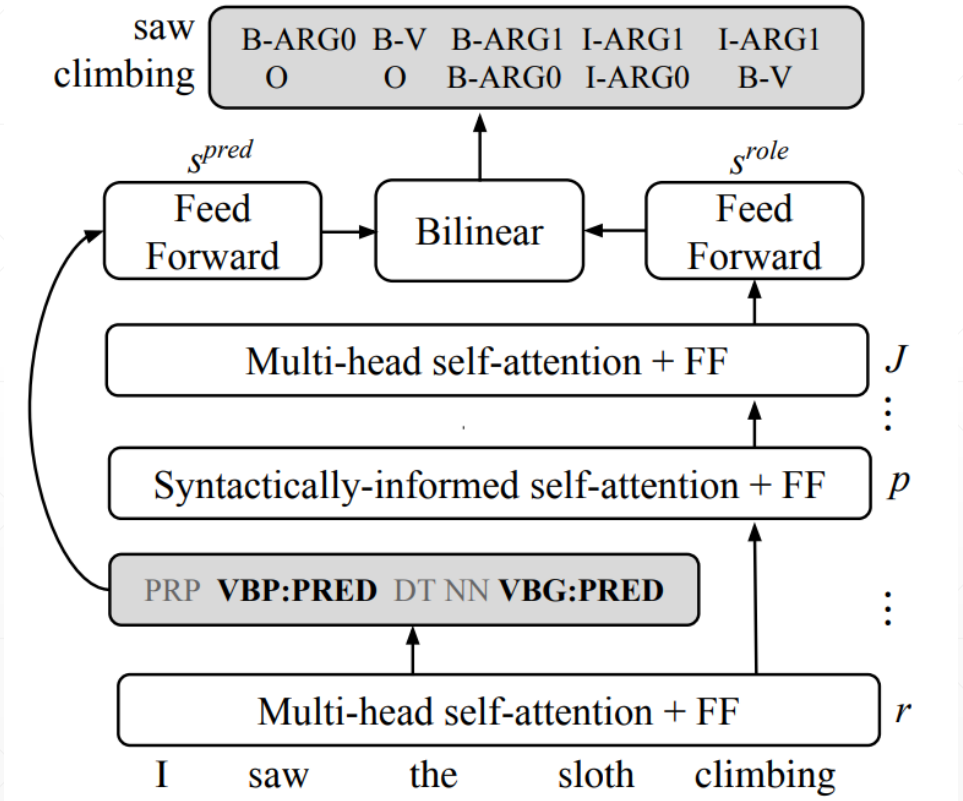
WORDS--->	NE--->	POS	PARTIAL_SYNT		FULL_SYNT---------->	VS	TARGETS	PROPS--------->	
The	*	DT	(NP*	(S*	(S(NP*	-	-	(A0*	(A0*
$	*	$	*	*	(ADJP(QP*	-	-	*	*
1.4	*	CD	*	*	*	-	-	*	*
billion	*	CD	*	*	*))	-	-	*	*
robot	*	NN	*	*	*	-	-	*	*
spacecraft	*	NN	*)	*	*)	-	-	*)	*)
faces	*	VBZ	(VP*)	*	(VP*	01	face	(V*)	*
a	*	DT	(NP*	*	(NP*	-	-	(A1*	*
six-year	*	JJ	*	*	*	-	-	*	*
journey	*	NN	*)	*	*	-	-	*	*
to	*	TO	(VP*	(S*	(S(VP*	-	-	*	*
explore	*	VB	*)	*	(VP*	01	explore	*	(V*)
Jupiter	(ORG*)	NNP	(NP*)	*	(NP(NP*)	-	-	*	(A1*
and	*	CC	*	*	*	-	-	*	*
its	*	PRP$	(NP*	*	(NP*	-	-	*	*
16	*	CD	*	*	*	-	-	*	*
known	*	JJ	*	*	*	-	-	*	*
moons	*	NNS	*)	*)	*))))))	-	-	*)	*)
.	*	.	*	*)	*)	-	-	*	*

LISA

- linguistically-informed self-attention (LISA)
  - SRL Task의 SOTA Model
  - multi-task learning 과 stacked layers of multi-head self-attention 모델의 결합
  - POS tagging + Predicate detection + syntactic dependencies + SRL

전체 모델

- r번째 layer
  - POS tagging과 Predicate detection에 대한 prediction
- p번째 레이어
  - Syntactically-informed self-attention
  - Using deep bi-affine model
  - trained to predict syntactic dependencies.
- SRL을 학습하기 위해 POS tagging과 Predicate detection, syntactic dependencies를 함께 학습



Tranformer

- Tranformer 모델의 Encoder 부분을 사용
  - Multi-Head attention과 Feed Forward

$$s_t^{(j)} = LN(s_t^{(j-1)} + T^{(j)}(s_t^{(j-1)}))$$

$$A_h^{(j)} = \text{softmax}(d_k^{-0.5} Q_h^{(j)} {K_h^{(j)}}^T)$$

$$M_h^{(j)} = A_h^{(j)} V_h^{(j)}$$

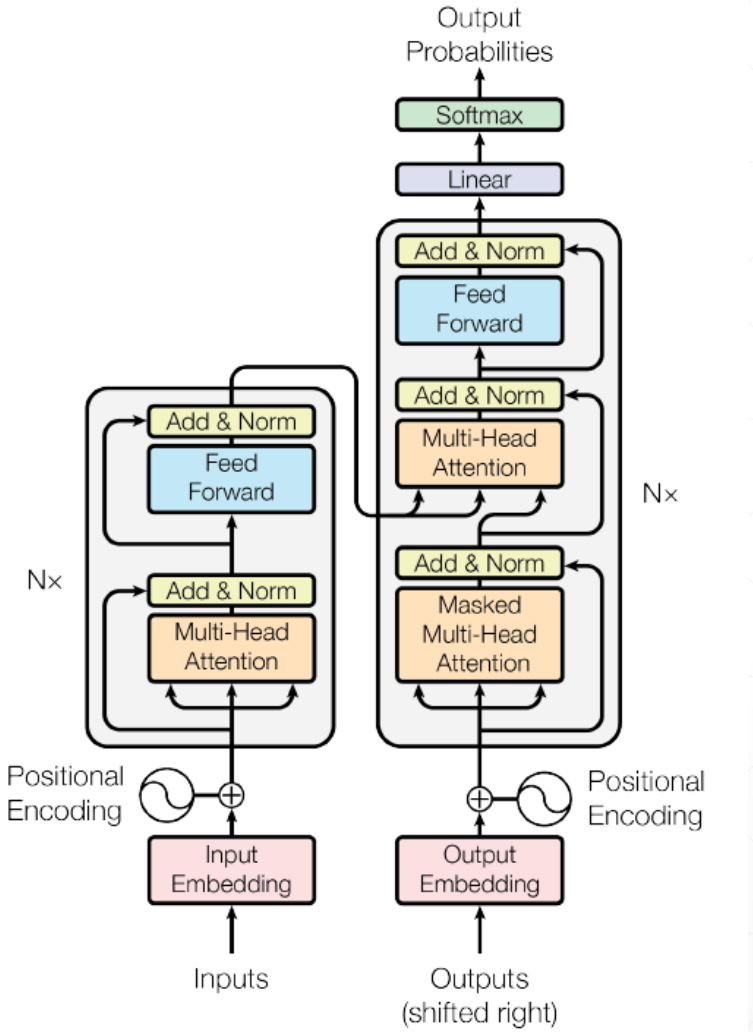


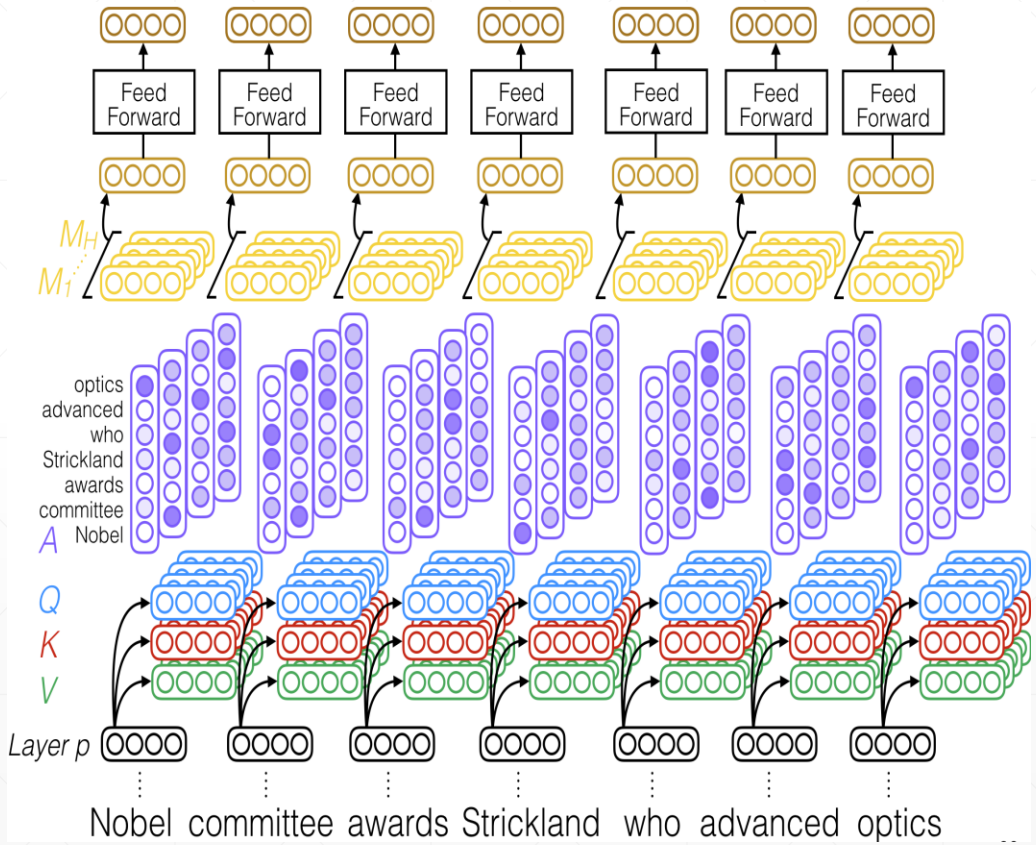
Figure 1: The Transformer - model architecture.

Self-attention

$$s_t^{(j)} = LN(s_t^{(j-1)} + T^{(j)}(s_t^{(j-1)}))$$

$$A_h^{(j)} = \text{softmax}(d_k^{-0.5} Q_h^{(j)} K_h^{(j)^T})$$

$$M_h^{(j)} = A_h^{(j)} V_h^{(j)}$$

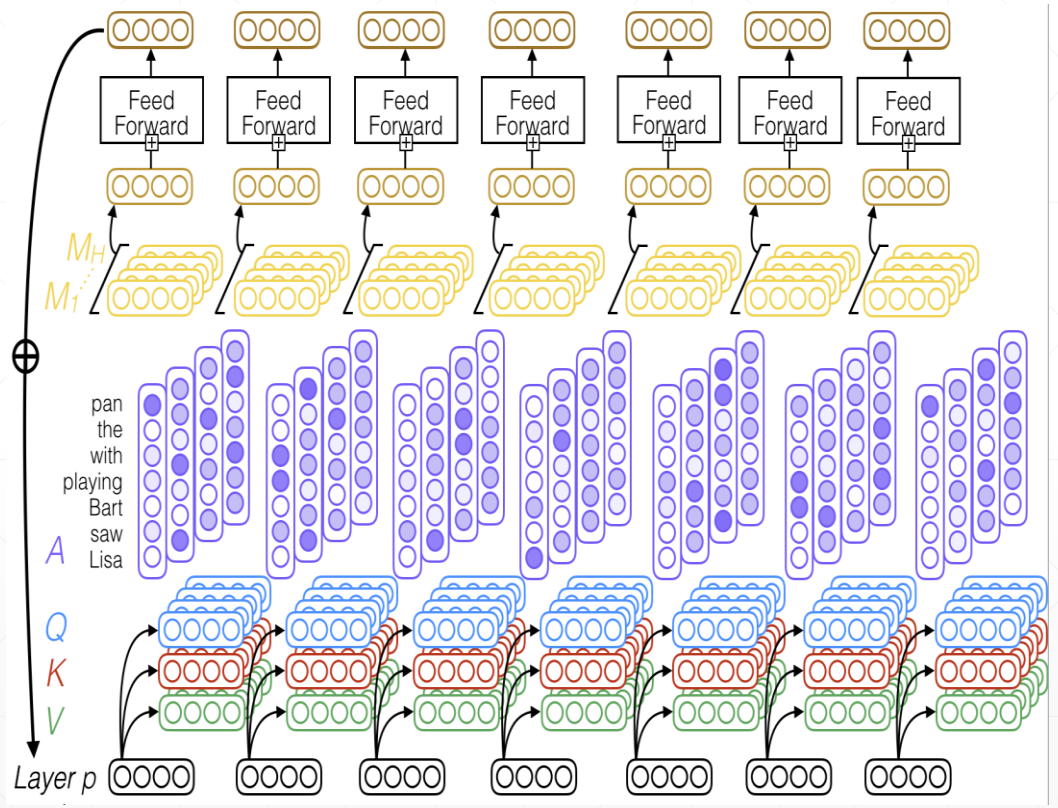


Self-attention

$$s_t^{(j)} = LN(s_t^{(j-1)} + T^{(j)}(s_t^{(j-1)}))$$

$$A_h^{(j)} = \text{softmax}(d_k^{-0.5} Q_h^{(j)} {K_h^{(j)}}^T)$$

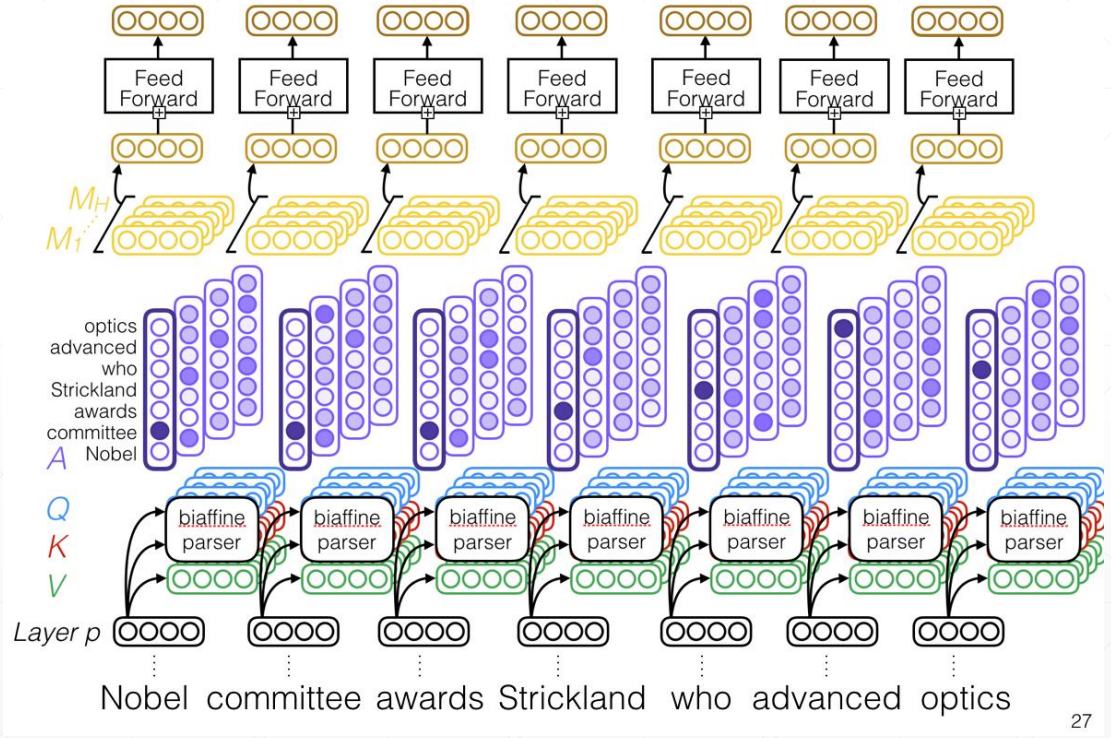
$$M_h^{(j)} = A_h^{(j)} V_h^{(j)}$$



Syntactically-informed self-attention

$$P(q = \text{head}(t) \mid \mathcal{X}) = A_{parse}[t, q]$$

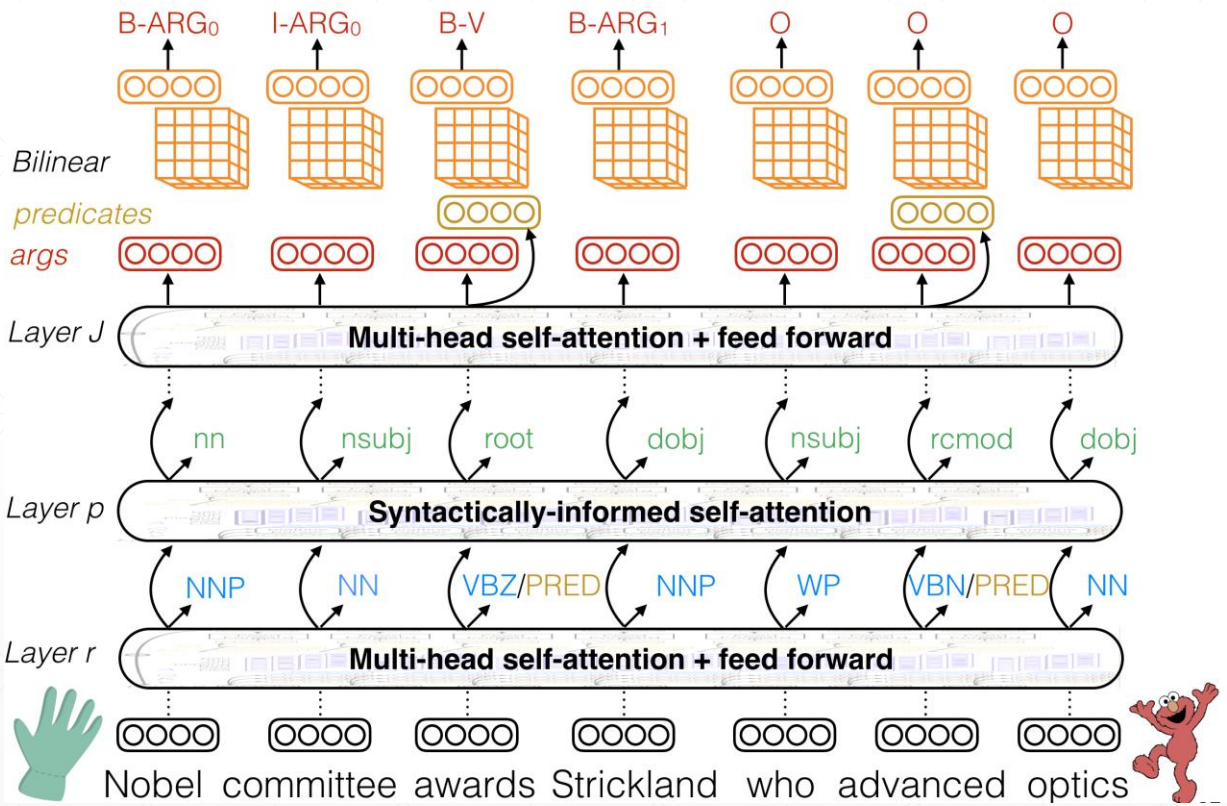
$$A_{parse} = \text{softmax}(Q_{parse} U_{heads} K_{parse}^{T})$$



Predicting & Training

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

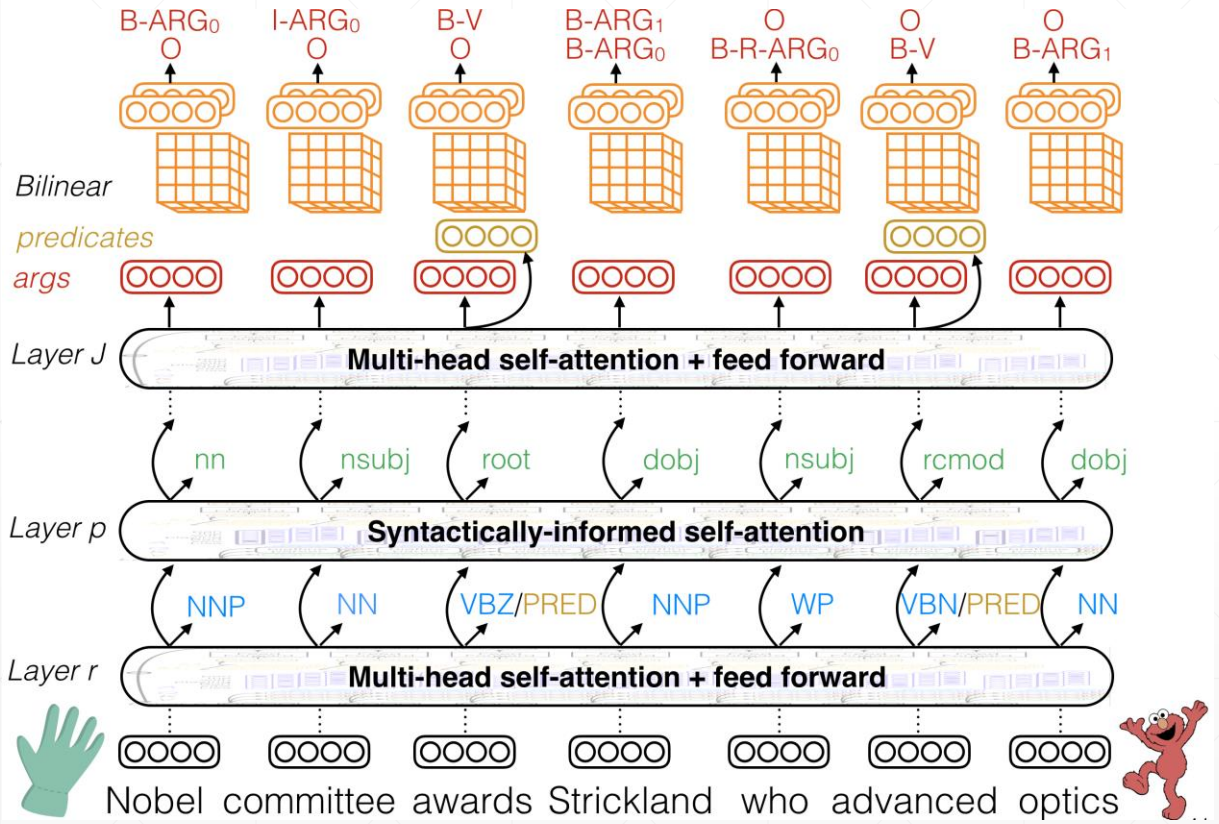
$$\frac{1}{T}\sum_{t=1}^{T}\Big[\sum_{f=1}^{F}\log P(y_{ft}^{role} \mid \mathcal{P}_G, \mathcal{V}_G, \mathcal{X}) + \log P(y_t^{prp} \mid \mathcal{X}) + \lambda_1 \log P(\text{head}(t) \mid \mathcal{X}) + \lambda_2 \log P(y_t^{dep} \mid \mathcal{P}_G, \mathcal{X})\Big]$$



Predicting & Training

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

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



ELMO embedding

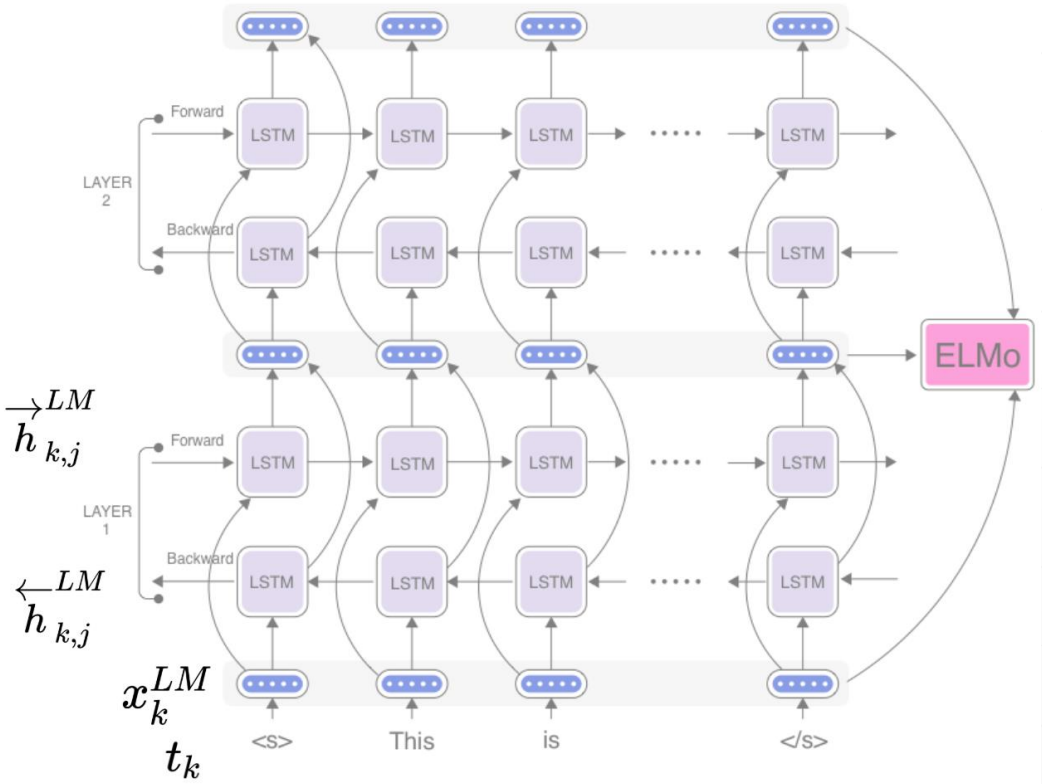
Structure

Each token t_k

L-layer biLM computes 2L+1 representations

k is the k-th token

j is the j-th biLM layer



CONLL 2005

	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.

CONLL 2005

WSJ Test	P	R	F1
He et al. (2018)	84.2	83.7	83.9
Tan et al. (2018)	84.5	85.2	84.8
SA	84.7	84.24	84.47
LISA	84.72	84.57	84.64
+D&M	**86.02**	**86.05**	**86.04**

Brown Test	P	R	F1
He et al. (2018)	74.2	73.1	73.7
Tan et al. (2018)	73.5	74.6	74.1
SA	73.89	72.39	73.13
LISA	74.77	74.32	74.55
+D&M	**76.65**	**76.44**	**76.54**

CONLL 2012

Dev	P	R	F1
	GloVe		
He et al. (2018)	79.2	79.7	79.4
SA	82.32	79.76	81.02
LISA	81.77	79.65	80.70
+D&M	**82.97**	**81.14**	**82.05**
+*Gold*	*87.57*	*85.32*	*86.43*
	ELMo		
He et al. (2018)	82.1	**84.0**	83.0
SA	84.35	82.14	83.23
LISA	**84.19**	82.56	**83.37**
+D&M	84.09	82.65	83.36
+*Gold*	*88.22*	*86.53*	*87.36*

Test	P	R	F1
	GloVe		
He et al. (2018)	79.4	80.1	79.8
SA	82.55	80.02	81.26
LISA	81.86	79.56	80.70
+D&M	**83.3**	**81.38**	**82.33**
	ELMo		
He et al. (2018)	81.9	**84.0**	82.9
SA	**84.39**	82.21	83.28
LISA	83.97	82.29	83.12
+D&M	84.14	82.64	**83.38**

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