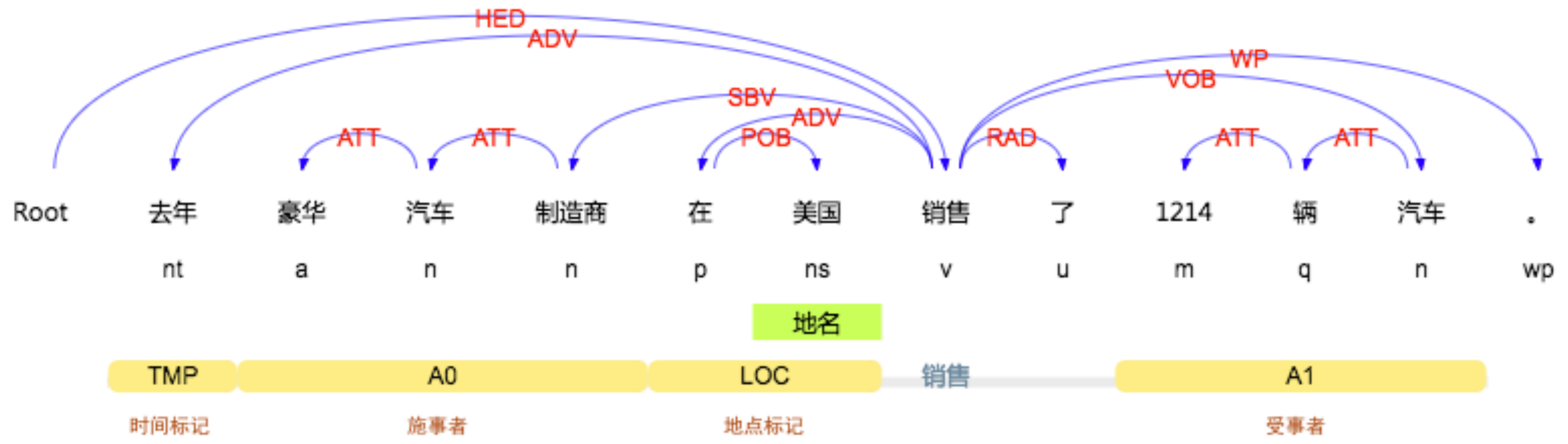
Deep Semantic Role Labeling: What Works and What’s Next

Author: Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer

Repoter: Yang Liu

Task

- semantic role labeling (SRL)
- recover the predicate-argument structure of a sentence, to determine essentially “who did what to whom”,“when”, and “where”



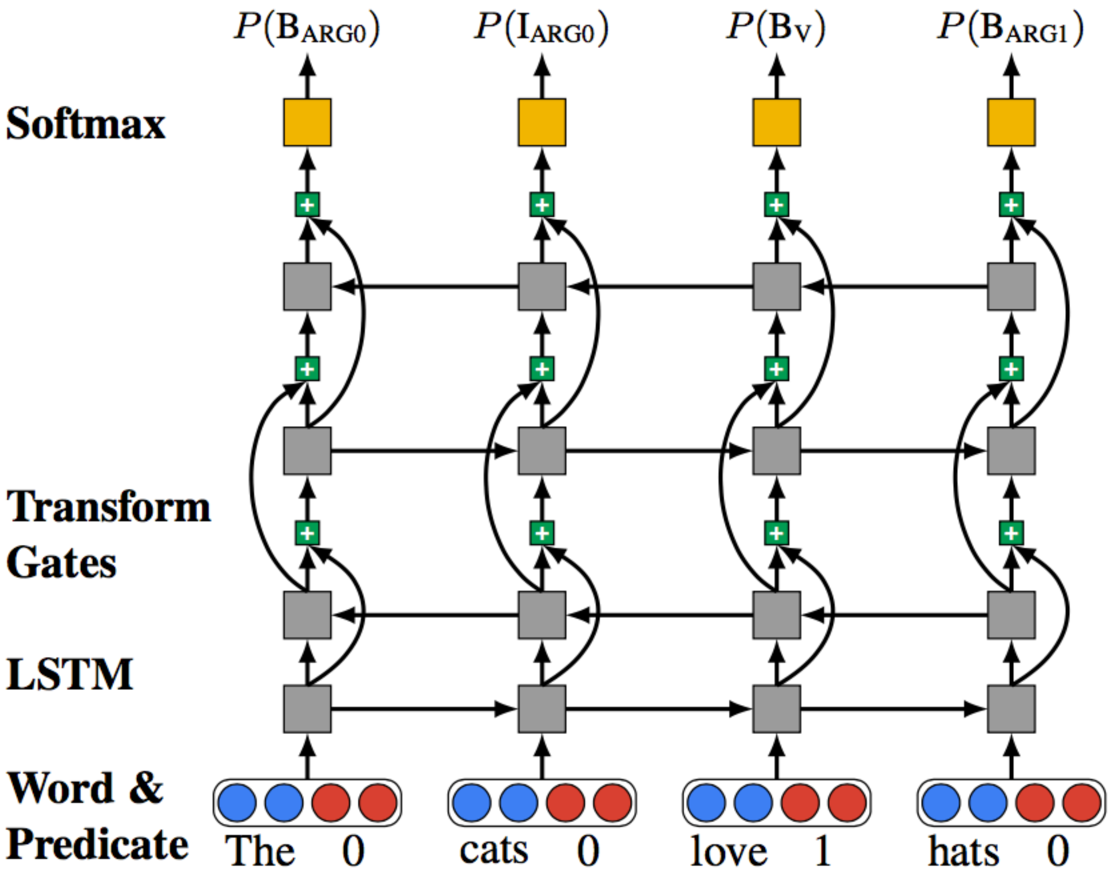
Model (RNN)

Deep BiLSTM Model

$$\boldsymbol{i}_{l,t} = \sigma(\mathbf{W}_{\mathrm{i}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{i}}^{l})$$
$$\boldsymbol{o}_{l,t} = \sigma(\mathbf{W}_{\mathrm{o}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{o}}^{l})$$
$$\boldsymbol{f}_{l,t} = \sigma(\mathbf{W}_{\mathrm{f}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{f}}^{l} + 1)$$
$$\tilde{\boldsymbol{c}}_{l,t} = \tanh(\mathbf{W}_{\mathrm{c}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{c}}^{l})$$
$$\boldsymbol{c}_{l,t} = \boldsymbol{i}_{l,t} \circ \tilde{\boldsymbol{c}}_{l,t} + \boldsymbol{f}_{l,t} \circ \boldsymbol{c}_{t+\delta_l}$$
$$\boldsymbol{h}_{l,t} = \boldsymbol{o}_{l,t} \circ \tanh(\boldsymbol{c}_{l,t})$$
$$\boldsymbol{x}_{l,t} = \begin{cases} [\mathbf{W}_{\mathrm{emb}}(w_t), \mathbf{W}_{\mathrm{mask}}(t = v)] & l = 1 \\ \boldsymbol{h}_{l-1,t} & l > 1 \end{cases}$$
$$\delta_l = \begin{cases} 1 & \text{if } l \text{ is even} \\ -1 & \text{otherwise} \end{cases}$$

l for layer indexes

t for timesteps



Model (RNN)

Highway Connections, Zhang et al., 2016;
Srivastava et al., 2015

$$\boldsymbol{i}_{l,t} = \sigma(\mathbf{W}_{\mathrm{i}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{i}}^{l})$$

$$\boldsymbol{o}_{l,t} = \sigma(\mathbf{W}_{\mathrm{o}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{o}}^{l})$$

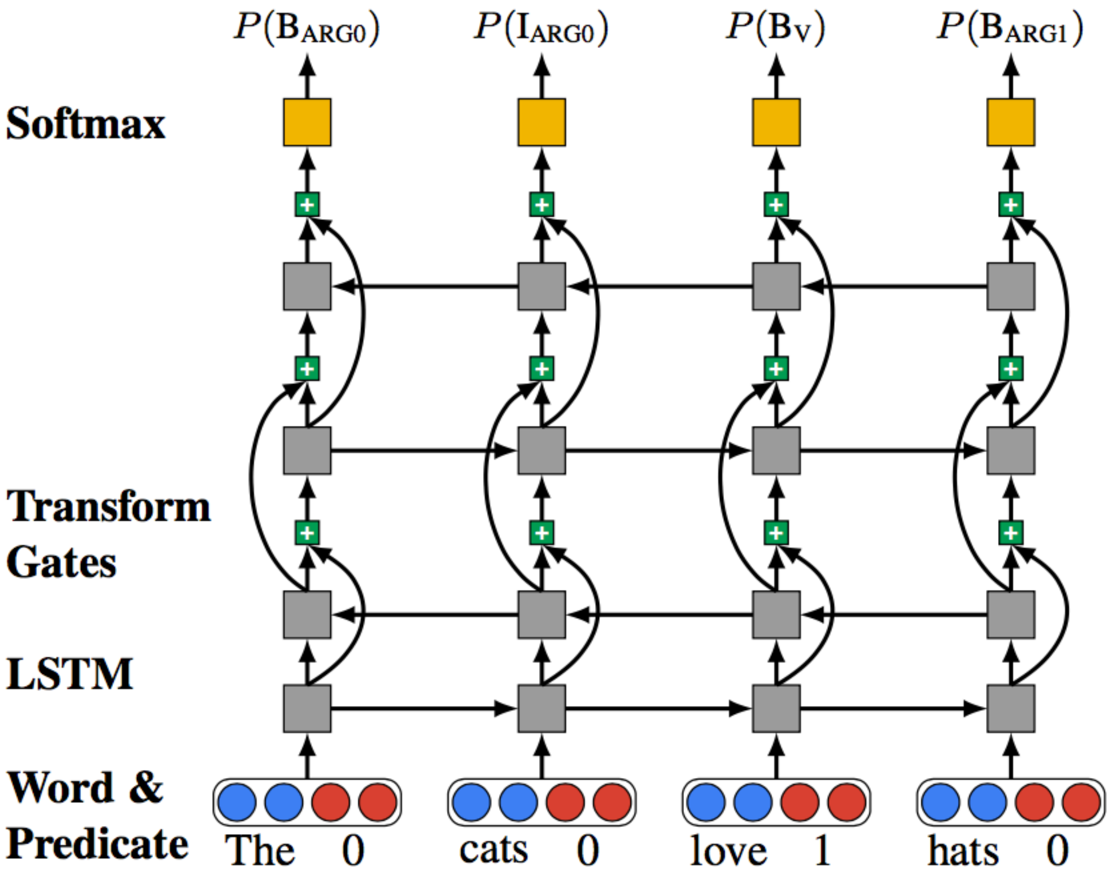
$$\boldsymbol{f}_{l,t} = \sigma(\mathbf{W}_{\mathrm{f}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{f}}^{l} + 1)$$

$$\tilde{\boldsymbol{c}}_{l,t} = \tanh(\mathbf{W}_{\mathrm{c}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{c}}^{l})$$

$$\boldsymbol{c}_{l,t} = \boldsymbol{i}_{l,t} \circ \tilde{\boldsymbol{c}}_{l,t} + \boldsymbol{f}_{l,t} \circ \boldsymbol{c}_{t+\delta_l}$$

$$\boxed{\boldsymbol{h}_{l,t}} = \boldsymbol{o}_{l,t} \circ \tanh(\boldsymbol{c}_{l,t})$$

$$\boxed{\begin{aligned} \boldsymbol{r}_{l,t} &= \sigma(\mathbf{W}_{\mathrm{r}}^{l}[\boldsymbol{h}_{l,t-1}, \boldsymbol{x}_{t}] + \boldsymbol{b}_{\mathrm{r}}^{l}) \\ \boldsymbol{h}'_{l,t} &= \boldsymbol{o}_{l,t} \circ \tanh(\boldsymbol{c}_{l,t}) \\ \boldsymbol{h}_{l,t} &= \boldsymbol{r}_{l,t} \circ \boldsymbol{h}'_{l,t} + (1 - \boldsymbol{r}_{l,t}) \circ \mathbf{W}_{\mathrm{h}}^{l} \boldsymbol{x}_{l,t} \end{aligned}}$$



Model (RNN)

Recurrent Dropout, Gal and Ghahramani (2016)

$$\boldsymbol{i}_{l,t} = \sigma(\mathbf{W}_{\mathrm{i}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{i}}^{l})$$

$$\boldsymbol{o}_{l,t} = \sigma(\mathbf{W}_{\mathrm{o}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{o}}^{l})$$

$$\boldsymbol{f}_{l,t} = \sigma(\mathbf{W}_{\mathrm{f}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{f}}^{l} + 1)$$

$$\tilde{\boldsymbol{c}}_{l,t} = \tanh(\mathbf{W}_{\mathrm{c}}^{l}[\boldsymbol{h}_{l,t+\delta_l}, \boldsymbol{x}_{l,t}] + \boldsymbol{b}_{\mathrm{c}}^{l})$$

$$\boldsymbol{c}_{l,t} = \boldsymbol{i}_{l,t} \circ \tilde{\boldsymbol{c}}_{l,t} + \boldsymbol{f}_{l,t} \circ \boldsymbol{c}_{t+\delta_l}$$

$$\boxed{\boldsymbol{h}_{l,t}} = \boldsymbol{o}_{l,t} \circ \tanh(\boldsymbol{c}_{l,t})$$

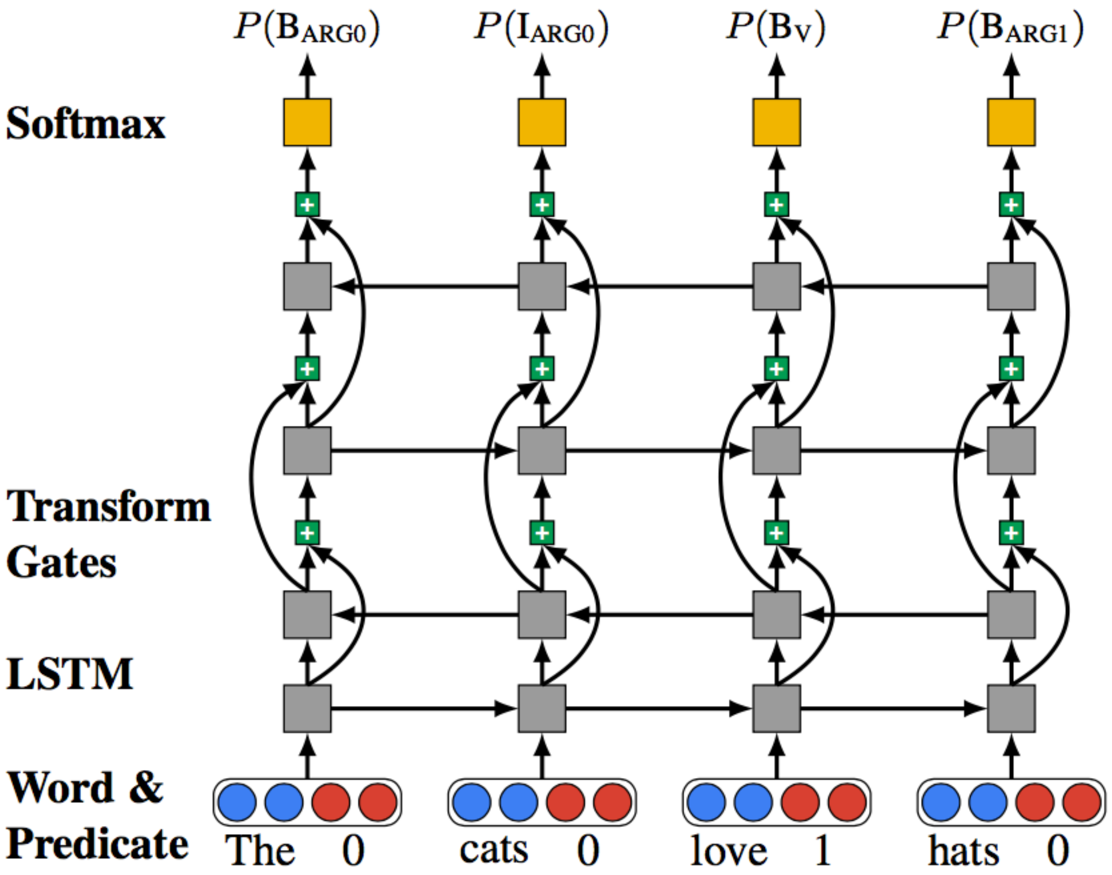
$$\boldsymbol{r}_{l,t} = \sigma(\mathbf{W}_{\mathrm{r}}^{l}[\boldsymbol{h}_{l,t-1}, \boldsymbol{x}_{t}] + \boldsymbol{b}_{\mathrm{r}}^{l})$$

$$\boldsymbol{h}'_{l,t} = \boldsymbol{o}_{l,t} \circ \tanh(\boldsymbol{c}_{l,t})$$

$$\boxed{\boldsymbol{h}_{l,t}} = \boldsymbol{r}_{l,t} \circ \boldsymbol{h}'_{l,t} + (1 - \boldsymbol{r}_{l,t}) \circ \mathbf{W}_{\mathrm{h}}^{l}\boldsymbol{x}_{l,t}$$

$$\widetilde{\boldsymbol{h}}_{l,t} = \boldsymbol{r}_{l,t} \circ \boldsymbol{h}'_{l,t} + (1 - \boldsymbol{r}_{l,t}) \circ \mathbf{W}_{\mathrm{h}}^{l}\boldsymbol{x}_{l,t}$$

$$\boldsymbol{h}_{l,t} = \boldsymbol{z}_{l} \circ \widetilde{\boldsymbol{h}}_{l,t}$$

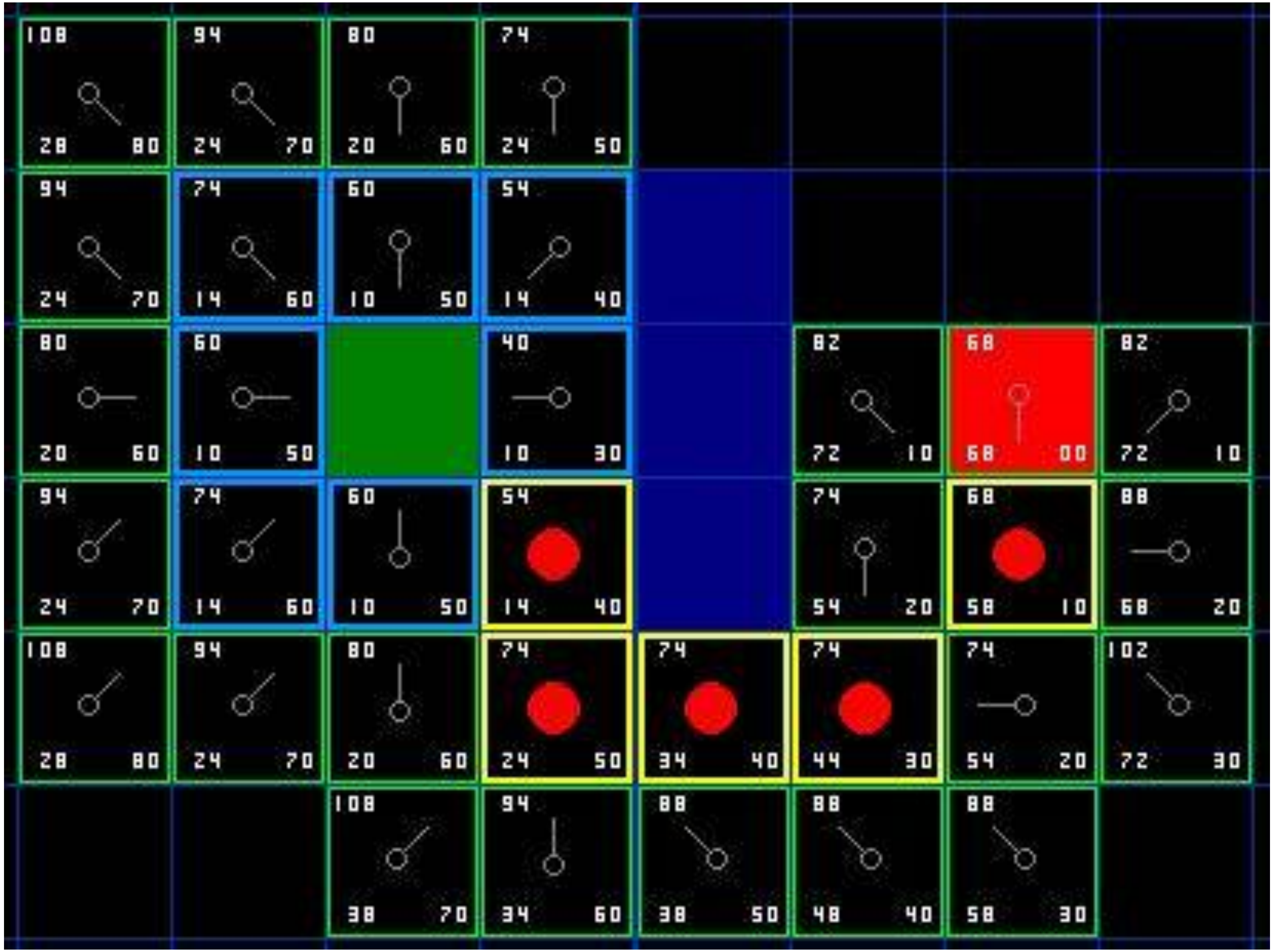


Model (A*)

A* 搜索算法

$$f(n) = g(n) + h^*(n)$$

启发式搜索算法



Model (A*)

$$f(n) = g(n) + h^*(n)$$

$$A^* cost(w, y_{1:i}) = f(w, y_{1:i}) + g(w, y_{1:i})$$

$$f(\boldsymbol{w}, y_{1:t}) = \sum_{i=1}^{t} \log p(y_i \mid \boldsymbol{w}) - \boxed{\sum_{c \in \mathcal{C}} c(\boldsymbol{w}, y_{1:i})}$$

$$g(\boldsymbol{w}, y_{1:t}) = \sum_{i=t+1}^{n} \max_{y_i \in \boldsymbol{T}} \log p(y_i \mid \boldsymbol{w})$$

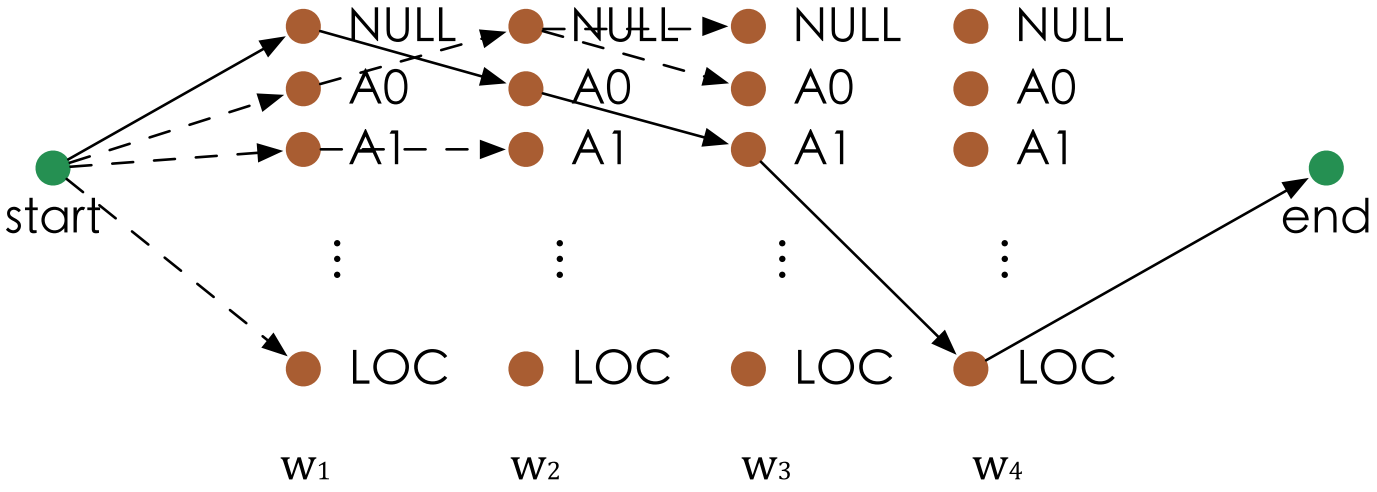
BIO Constraints

- Such as B_{ARG0} followed by $I_{ARG1}$is illegal

SRL Constraints (not use in PoE)

- Core roles (A0-A5) appear once
- Continueation role exist only when its base role realized before it
- Reference role. Same as above.(not use)

Syntactic Constraints



Predicate Identification (PI)

- bidirectional LSTM then softmax
- maximize the likelihood of the gold labels

Experiments Setting

- 8 LSTM layers (4 BiLSTM layers)
- word tokens (lower-cased) initialized with 100-dimensional GloVe embeddings (updated during trainning)
- Ensembling with 5-folds

Results on CoNLL 2005

	Development				WSJ Test				Brown Test				Combined
Method	P	R	F1	Comp.	P	R	F1	Comp.	P	R	F1	Comp.	F1
Ours (PoE)	**83.1**	**82.4**	**82.7**	**64.1**	**85.0**	**84.3**	**84.6**	**66.5**	**74.9**	**72.4**	**73.6**	**46.5**	**83.2**
Ours	81.6	81.6	81.6	62.3	83.1	83.0	83.1	64.3	72.9	71.4	72.1	44.8	81.6
Zhou	79.7	79.4	79.6	-	82.9	82.8	82.8	-	70.7	68.2	69.4	-	81.1
FitzGerald (Struct.,PoE)	81.2	76.7	78.9	55.1	82.5	78.2	80.3	57.3	74.5	70.0	72.2	41.3	-
Täckström (Struct.)	81.2	76.2	78.6	54.4	82.3	77.6	79.9	56.0	74.3	68.6	71.3	39.8	-
Toutanova (Ensemble)	-	-	78.6	58.7	81.9	78.8	80.3	60.1	-	-	68.8	40.8	-
Punyakanok (Ensemble)	80.1	74.8	77.4	50.7	82.3	76.8	79.4	53.8	73.4	62.9	67.8	32.3	77.9

Table 1: Experimental results on CoNLL 2005, in terms of precision (P), recall (R), F1 and percentage of completely correct predicates (Comp.). We report results of our best single and ensemble (PoE) model. The comparison models are Zhou and Xu (2015), FitzGerald et al. (2015), Täckström et al. (2015), Toutanova et al. (2008) and Punyakanok et al. (2008).

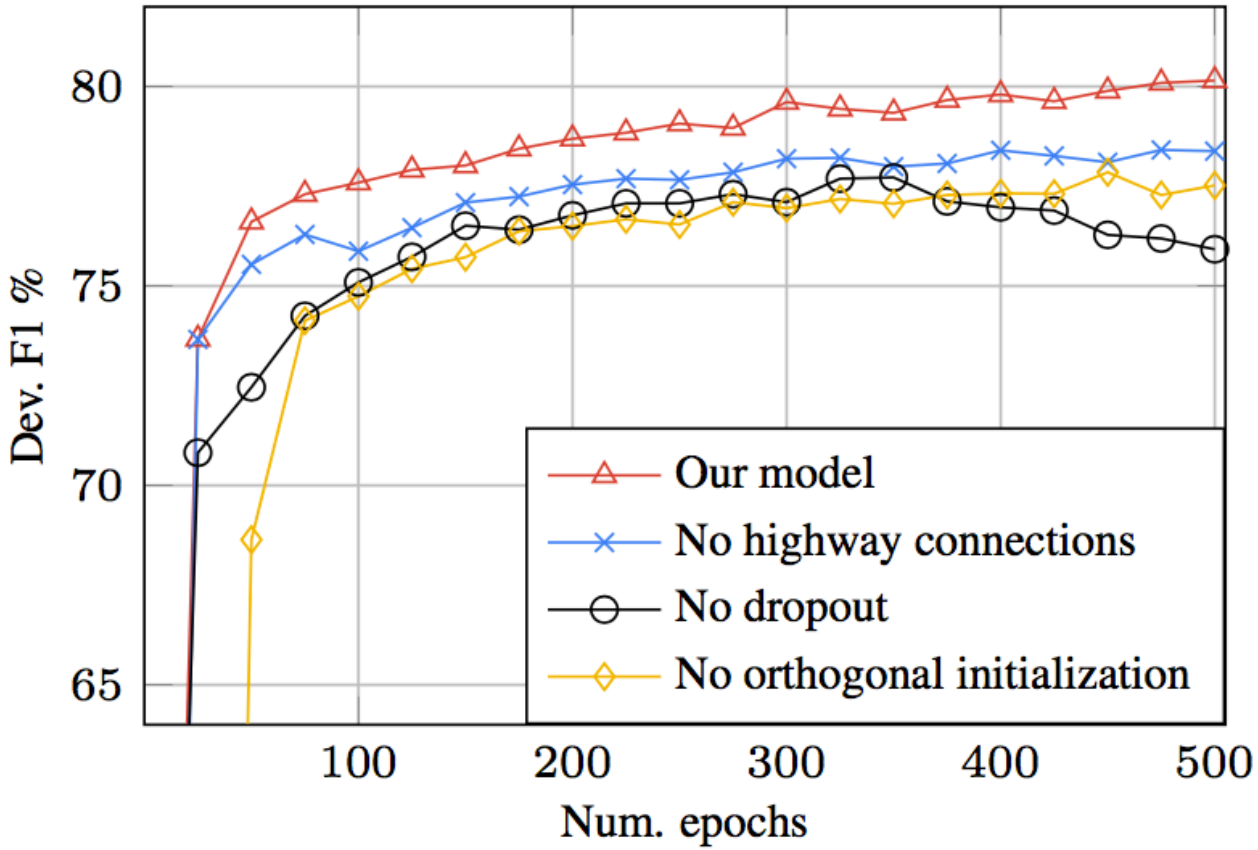
Results on CoNLL 2012

Method	Development				Test			
	P	R	F1	Comp.	P	R	F1	Comp.
Ours (PoE)	**83.5**	**83.2**	**83.4**	**67.5**	**83.5**	**83.3**	**83.4**	**68.5**
Ours	81.8	81.4	81.5	64.6	81.7	81.6	81.7	66.0
Zhou	-	-	81.1	-	-	-	81.3	-
FitzGerald (Struct.,PoE)	81.0	78.5	79.7	60.9	81.2	79.0	80.1	62.6
Täckström (Struct.)	80.5	77.8	79.1	60.1	80.6	78.2	79.4	61.8
Pradhan (revised)	-	-	-	-	78.5	76.6	77.5	55.8

Table 2: Experimental results on CoNLL 2012 in the same metrics as above. We compare our best single and ensemble (PoE) models against Zhou and Xu (2015), FitzGerald et al. (2015), Täckström et al. (2015) and Pradhan et al. (2013).

Trainning

- No highway
- No dropout
- And no orthogonal initialization



End-to-end Results

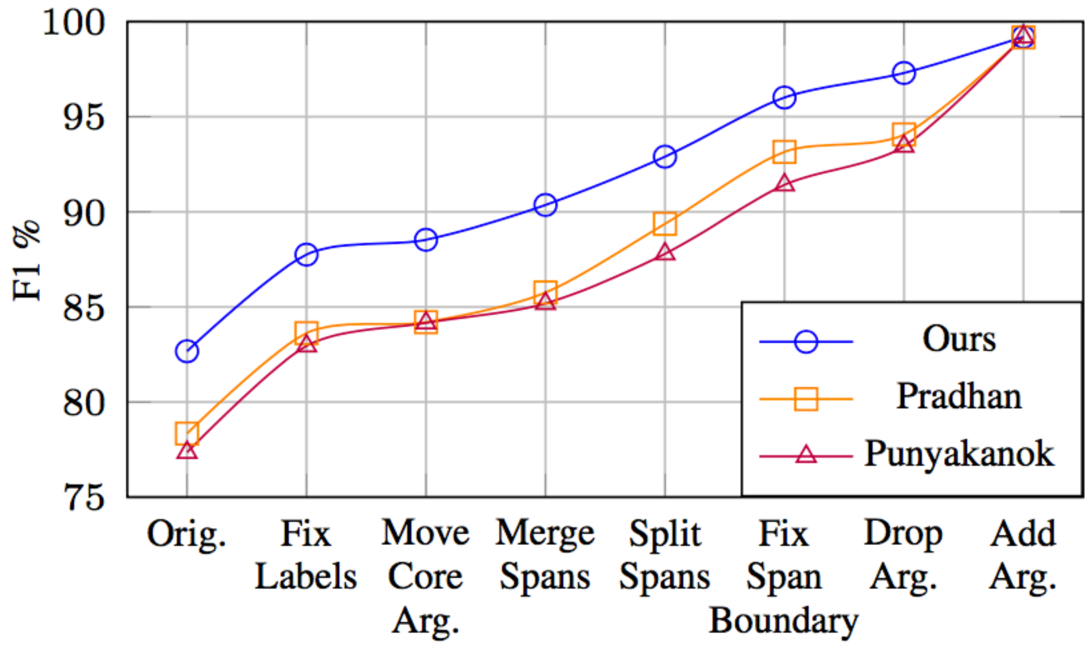
	Predicate Detection			End-to-end SRL (Single)			End-to-end SRL (PoE)			
Dataset	P	R	F1	P	R	F1	P	R	F1	Δ F1
CoNLL 2005 Dev.	97.4	97.4	97.4	80.3	80.4	80.3	81.8	81.2	81.5	-1.2
WSJ Test	94.5	98.5	96.4	80.2	82.3	81.2	82.0	83.4	82.7	-1.9
Brown Test	89.3	95.7	92.4	67.6	69.6	68.5	69.7	70.5	70.1	-3.5
CoNLL 2012 Dev.	88.7	90.6	89.7	74.9	76.2	75.5	76.5	77.8	77.2	-6.2
CoNLL 2012 Test	93.7	87.9	90.7	78.6	75.1	76.8	80.2	76.6	78.4	-5.0

Table 3: Predicate detection performance and end-to-end SRL results using predicted predicates. Δ F1 shows the absolute performance drop compared to our best ensemble model with gold predicates.

Analysis

- What is the model good at and what kinds of mistakes does it make?
- How well do LSTMs model global structural consistency, despite conditionally independent tagging decisions?
- Is our model implicitly learning syntax, and could explicitly modeling syntax still help?

Error Types Breakdown



Operation	Description	%
Fix Labels	Correct the span label if its boundary matches gold.	29.3
Move Arg.	Move a unique core argument to its correct position.	4.5
Merge Spans	Combine two predicted spans into a gold span if they are separated by at most one word.	10.6
Split Spans	Split a predicted span into two gold spans that are separated by at most one word.	14.7
Fix Boundary	Correct the boundary of a span if its label matches an overlapping gold span.	18.0
Drop Arg.	Drop a predicted argument that does not overlap with any gold span.	7.4
Add Arg.	Add a gold argument that does not overlap with any predicted span.	11.0

Table 4: Oracle transformations paired with the relative error reduction after each operation. All the operations are permitted only if they do not cause any overlapping arguments.

Label Confusion

- The model of- ten confuses ARG2 with AM-DIR, AM-LOC and AM-MNR. These confusions can arise due to the use of ARG2 in many verb frames to represent semantic relations such as direction or location.
- For example, ARG2 in the frame *move.01* is defined as *Arg2-GOL: destination*.

pred. \ gold	A0	A1	A2	A3	ADV	DIR	LOC	MNR	PNC	TMP
A0	-	55	11	13	4	0	0	0	0	0
A1	78	-	46	0	0	22	11	10	25	14
A2	11	23	-	48	15	56	33	41	25	0
A3	3	2	2	-	4	0	0	0	25	14
ADV	0	0	0	4	-	0	15	29	25	36
DIR	0	0	5	4	0	-	11	2	0	0
LOC	5	9	12	0	4	0	-	10	0	14
MNR	3	0	12	26	33	0	0	-	0	21
PNC	0	3	5	4	0	11	4	2	-	0
TMP	0	8	5	0	41	11	26	6	0	-

Table 5: Confusion matrix for labeling errors, showing the percentage of predicted labels for each gold label. We only count predicted arguments that match gold span boundaries.

Attachment Mistakes

Sumitomo ***financed*** *the acquisition from Sears*

ARG2

Sumitomo ***financed*** *the acquisition from Sears*

ARG2(gold)

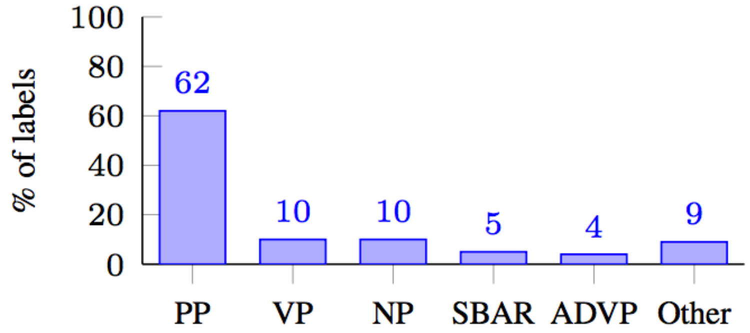


Figure 4: For cases where our model either splits a gold span into two ($Z \rightarrow XY$) or merges two gold constituents ($XY \rightarrow Z$), we show the distribution of syntactic labels for the Y span. Results show the major cause of these errors is inaccurate prepositional phrase attachment.

Long-range Dependencies

- neural model performance deteriorates less severely on long-range dependencies than traditional syntax-based models.

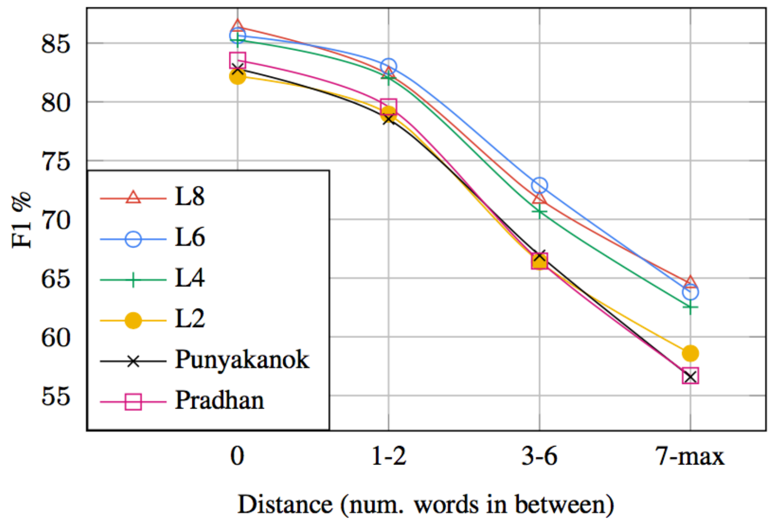


Figure 5: F1 by surface distance between predicates and arguments. Performance degrades least rapidly on long-range arguments for the deeper neural models.

BIO Violations

- Using BIO-constrained decoding can resolve ambiguity and result in a structurally consistent solution.

	Accuracy		Violations	Avg. Entropy	
Model (no BIO)	F1	Token	BIO	All	BIO
L8+PoE	81.5	91.5	0.07	0.02	0.72
L8	80.5	90.9	0.07	0.02	0.73
L6	80.1	90.3	0.06	0.02	0.72
L4	79.1	90.2	0.08	0.02	0.70
L2	74.6	88.4	0.18	0.03	0.66

Table 6: Comparison of BiLSTM models without BIO decoding. We compare F1 and token-level accuracy (Token), averaged BIO violations per token (BIO), overall model entropy (All) model entropy at tokens involved in BIO violations (BIO). Increasing the depth of the model beyond 4 does not produce more structurally consistent output, emphasizing the need for constrained decoding.

SRL Structure Violations

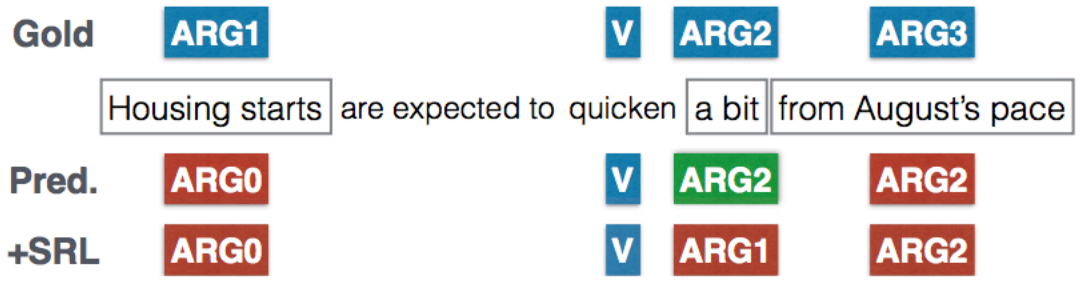


Figure 6: Example where performance is hurt by enforcing the constraint that core roles may only occur once (+SRL).

Model or Oracle	F1	Syn %	SRL-Violations U	C	R
Gold	100.0	98.7	24	0	61
L8+PoE	82.7	94.3	37	3	68
L8	81.6	94.0	48	4	73
L6	81.4	93.7	39	3	85
L4	80.5	93.2	51	3	84
L2	77.2	91.3	96	5	72
L8+PoE+SRL	82.8	94.2	5	1	68
L8+PoE+AutoSyn	83.2	96.1	113	3	68
L8+PoE+GoldSyn	85.0	97.6	102	3	68
Punyakanok	77.4	95.3	0	0	0
Pradhan	78.3	93.0	84	3	58

Table 7: Comparison of models with different depths and decoding constraints (in addition to BIO) as well as two previous systems. We compare F1, unlabeled agreement with gold constituency (Syn%) and each type of SRL-constraint violations (**U**nique core roles, **C**ontinuation roles and **R**eference roles). Our best model produces a similar number of constraint violations to the gold annotation, explaining why deterministically enforcing these constraints is not helpful.

Can Syntax Still Help SRL?

- if the decoded sequence contains k arguments that do not match any unlabeled syntactic constituent, it will receive a penalty of kC, where C is a single parameter dictating how much the model should trust the provided syntax.
- C = 10000 on CoNLL 2005 and C = 20 on CoNLL 2012

			SRL-Violations		
Model or Oracle	F1	Syn %	U	C	R
Gold	100.0	98.7	24	0	61
L8+PoE	82.7	94.3	37	3	68
L8	81.6	94.0	48	4	73
L6	81.4	93.7	39	3	85
L4	80.5	93.2	51	3	84
L2	77.2	91.3	96	5	72
L8+PoE+SRL	82.8	94.2	5	1	68
L8+PoE+AutoSyn	83.2	96.1	113	3	68
L8+PoE+GoldSyn	85.0	97.6	102	3	68
Punyakanok	77.4	95.3	0	0	0
Pradhan	78.3	93.0	84	3	58

Table 7: Comparison of models with different depths and decoding constraints (in addition to BIO) as well as two previous systems. We compare F1, unlabeled agreement with gold constituency (Syn%) and each type of SRL-constraint violations (**U**nique core roles, **C**ontinuation roles and **R**eference roles). Our best model produces a similar number of constraint violations to the gold annotation, explaining why deterministically enforcing these constraints is not helpful.

Thanks and Q&A.