Fast(er) Exact Decoding and Global Training for *TBDPs* via a Minimal Feature Set

Tianze Shi, Liang Huang, Lillian Lee

Cornell & Oregon State -- EMNLP17

https://github.com/tzshi/dp-parser-emnlp17

AntNLP -- Tao Ji

taoji.cs@gmail.com

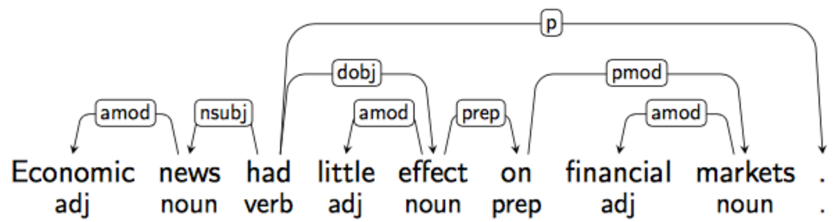
Outline

- Transition-based Dependency Parsing
- Three Transition Systems
- Motivation
- A Minimal Feature Set
- Dynamic Programming for TBDPs
- Practical Optimal Algorithms
- Experiments & CoNLL 2017 Shared Task
- Conclusion

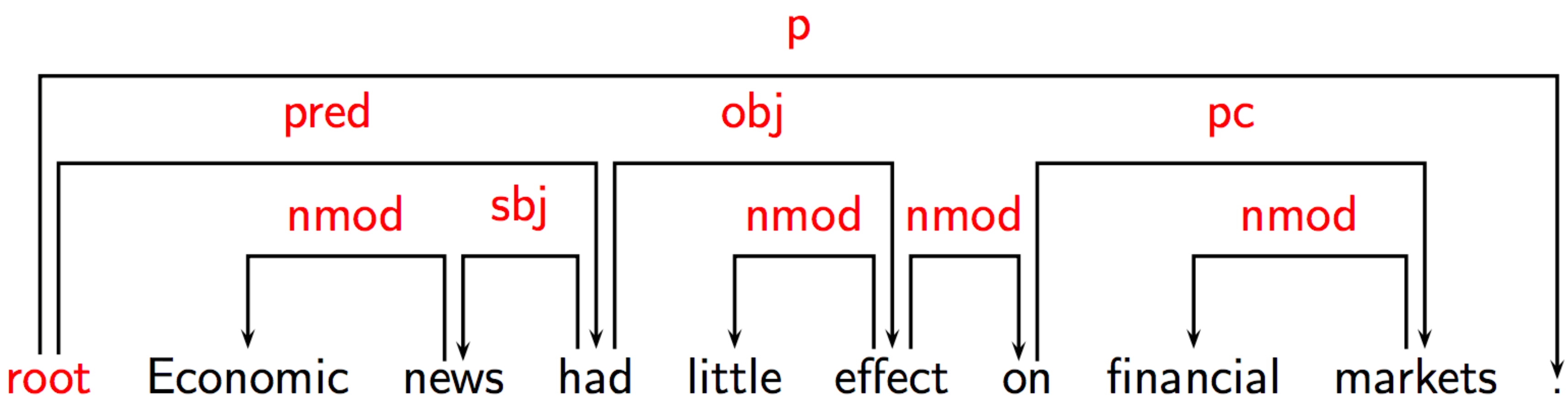
TBDPs

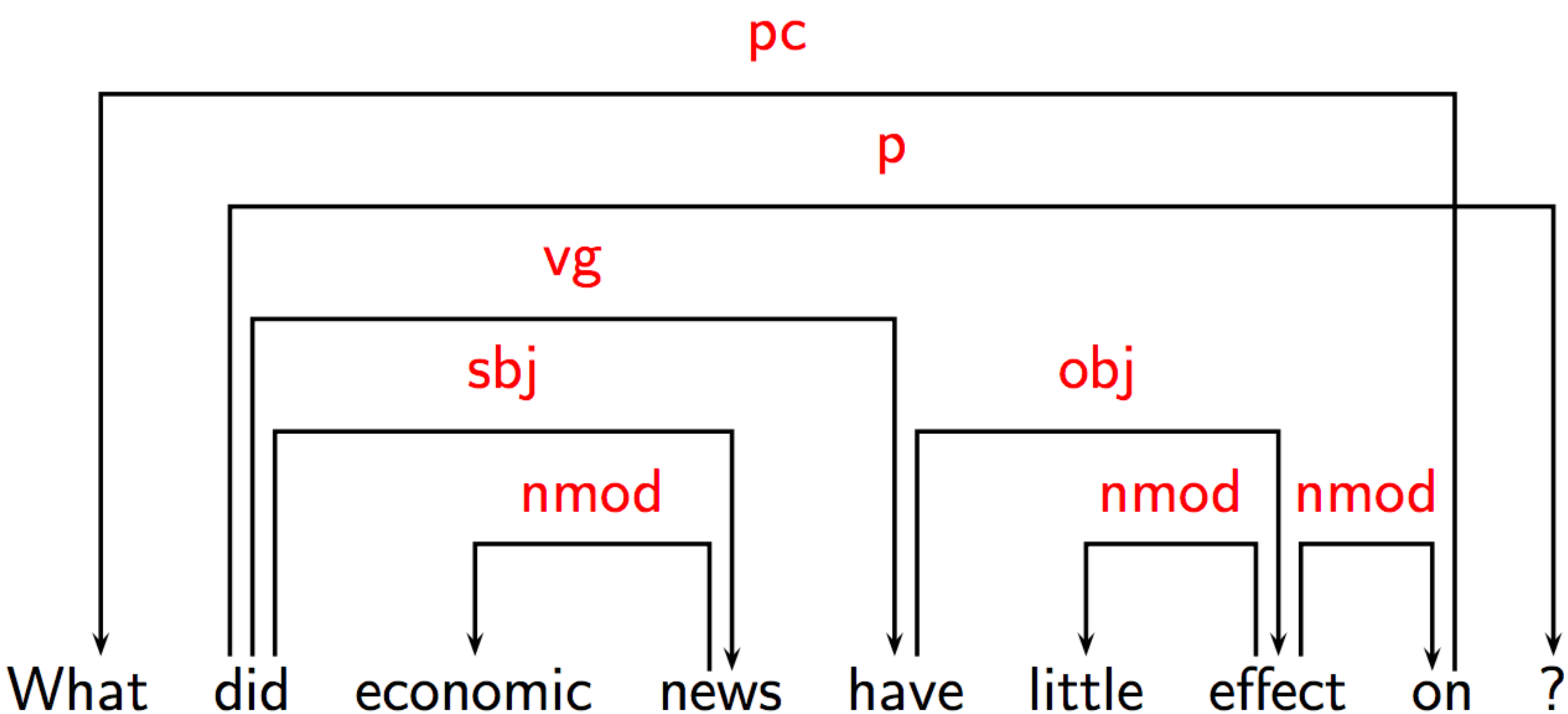
Configuration:	(S, B, A)		
Initial:	$([\ ], [0, 1, \ldots, n], \{\ \})$		
Terminal:	$(S, [\ ], A)$		
Shift:	$(S, i\|B, A)$	$\Rightarrow$	$(S\|i, B, A)$
Reduce:	$(S\|i, B, A)$	$\Rightarrow$	(S, B, A)
Right-Arc(k):	$(S\|i, j\|B, A)$	$\Rightarrow$	$(S\|i\|j, B, A \cup \{(i, j, k)\})$
Left-Arc(k):	$(S\|i, j\|B, A)$	$\Rightarrow$	$(S, j\|B, A \cup \{(j, i, k)\})$

$\Leftrightarrow$



Projectivity





Arc-Eager

Configuration:	(S, B, A)	$[S = \text{Stack},\ B = \text{Buffer},\ A = \text{Arcs}]$		
Initial:	$([\ ], [0, 1, \ldots, n], \{\ \})$			
Terminal:	$(S, [\ ], A)$			
Shift:	$(S, i\|B, A)$	$\Rightarrow$	$(S\|i, B, A)$	
Reduce:	$(S\|i, B, A)$	$\Rightarrow$	(S, B, A)	$h(i, A)$
Right-Arc(k):	$(S\|i, j\|B, A)$	$\Rightarrow$	$(S\|i\|j, B, A \cup \{(i, j, k)\})$	
Left-Arc(k):	$(S\|i, j\|B, A)$	$\Rightarrow$	$(S, j\|B, A \cup \{(j, i, k)\})$	$\neg h(i, A) \wedge i \neq 0$

Notation: $S|i$ = stack with top i and remainder S
$j|B$ = buffer with head j and remainder B
$h(i, A)$ = i has a head in A

Arc-Standard

Configuration: (S, B, A) $[S = \text{Stack},\ B = \text{Buffer},\ A = \text{Arcs}]$

Initial: $([\ ], [0, 1, \ldots, n], \{\ \})$

Terminal: $([0], [\ ], A)$

Shift: $(S, i|B, A) \Rightarrow (S|i, B, A)$

Right-Arc(k): $(S|i|j, B, A) \Rightarrow (S|i, B, A \cup \{(i, j, k)\})$

Left-Arc(k): $(S|i|j, B, A) \Rightarrow (S|j, B, A \cup \{(j, i, k)\})$ $i \neq 0$

Arc-Hybrid

$$\text{sh}[(\sigma, b_0|\beta, A)] = (\sigma|b_0, \beta, A)$$

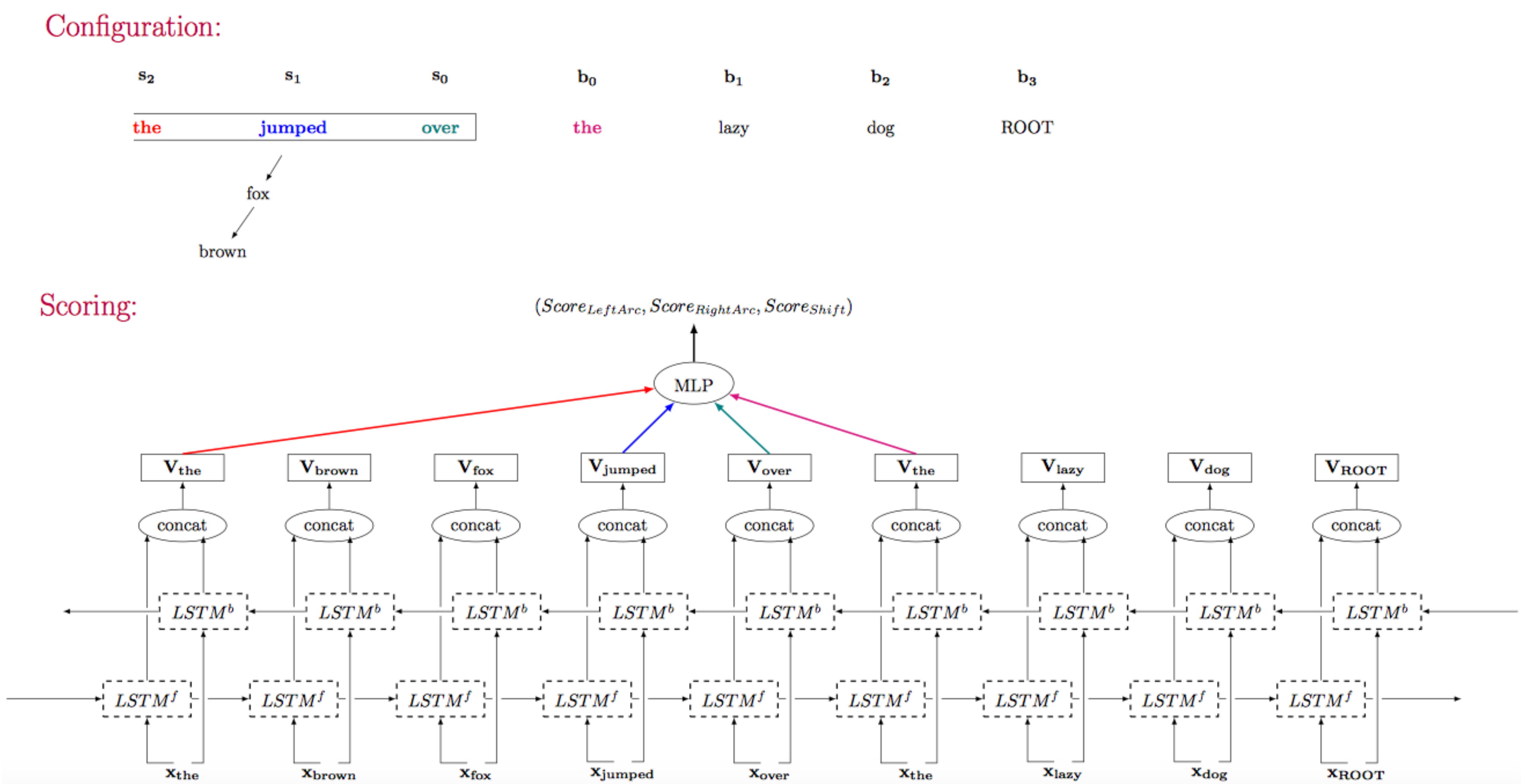
$$\text{re}_{\curvearrowright}[(\sigma|s_1|s_0, \beta, A)] = (\sigma|s_1, \beta, A \cup \{(s_1, s_0)\})$$

$$\text{re}_{\curvearrowleft}[(\sigma|s_0, b_0|\beta, A)] = (\sigma, b_0|\beta, A \cup \{(b_0, s_0)\})$$

Introduction

- Plug their minimal feature set into the dynamic-programming framework.
- Produce the first implementation of worst-case $O(n^3)$ exact decoders for arc-hybrid and arc-eager transition systems.
- With their minimal features, we also present $O(n^3)$ global training methods.
- Their achieve the best UAS reported (to our knowledge) on the CTB and the “second-best-in-class” result on the PTB.
- Their had the top average performance on the four surprise languages and on the small treebank subset.

BIST Parser


Configuration:
s2
s1
s0
b0
b1
b2
b3
the
jumped
over
the
lazy
dog
ROOT
fox
brown
Scoring:
(Score_LeftArc, Score_RightArc, Score_Shift)
MLP
V_the
V_brown
V_fox
V_jumped
V_over
V_the
V_lazy
V_dog
V_ROOT
concat
LSTM^b
LSTM^f
x_the
x_brown
x_fox
x_jumped
x_over
x_the
x_lazy
x_dog
x_ROOT

Minimal Feature Set

Features	Arc-standard	Arc-hybrid	Arc-eager
$\{\overset{\rightarrow\leftarrow}{s}_2, \overset{\rightarrow\leftarrow}{s}_1, \overset{\rightarrow\leftarrow}{s}_0, \overset{\rightarrow\leftarrow}{b}_0\}$	$93.95_{\pm 0.12}$	$94.08_{\pm 0.13}$	$93.92_{\pm 0.04}$
$\{\overset{\rightarrow\leftarrow}{s}_1, \overset{\rightarrow\leftarrow}{s}_0, \overset{\rightarrow\leftarrow}{b}_0\}$	$94.13_{\pm 0.06}$	$94.08_{\pm 0.05}$	$93.91_{\pm 0.07}$
$\{\overset{\rightarrow\leftarrow}{s}_0, \overset{\rightarrow\leftarrow}{b}_0\}$	$54.47_{\pm 0.36}$	$94.03_{\pm 0.12}$	$93.92_{\pm 0.07}$
$\{\overset{\rightarrow\leftarrow}{b}_0\}$	$47.11_{\pm 0.44}$	$52.39_{\pm 0.23}$	$79.15_{\pm 0.06}$

Min positions	Arc-standard	Arc-hybrid	Arc-eager
K&G 2016a	-	4	-
C&H 2016a	3	-	-
our work	3	**2**	**2**

Dynamic Programming

- If beam search reduces search errors, why not exact inference?
- Dynamic programming for transition-based parsers:
  - Using a graph-structured stack [Huang and Sagae 2010]
  - Using push-computations [Kuhlmann et al. 2011]
- Adds constraints on feature representations

Features

Overlapping Subproblems

Optimal Substructure

Deduction System for Arc-Eager Parsing

Items: $[i^b, j] \quad \Leftrightarrow \quad (S, i|B, A) \Rightarrow^* (S|i, j|B', A')$

$$b = \begin{cases} 1 & \text{if } [\![h(i) \in A']\!] \\ 0 & \text{otherwise} \end{cases}$$

Goal: $[0^0, n+1]$

Axiom: $[0^0, 1]$

Rules:

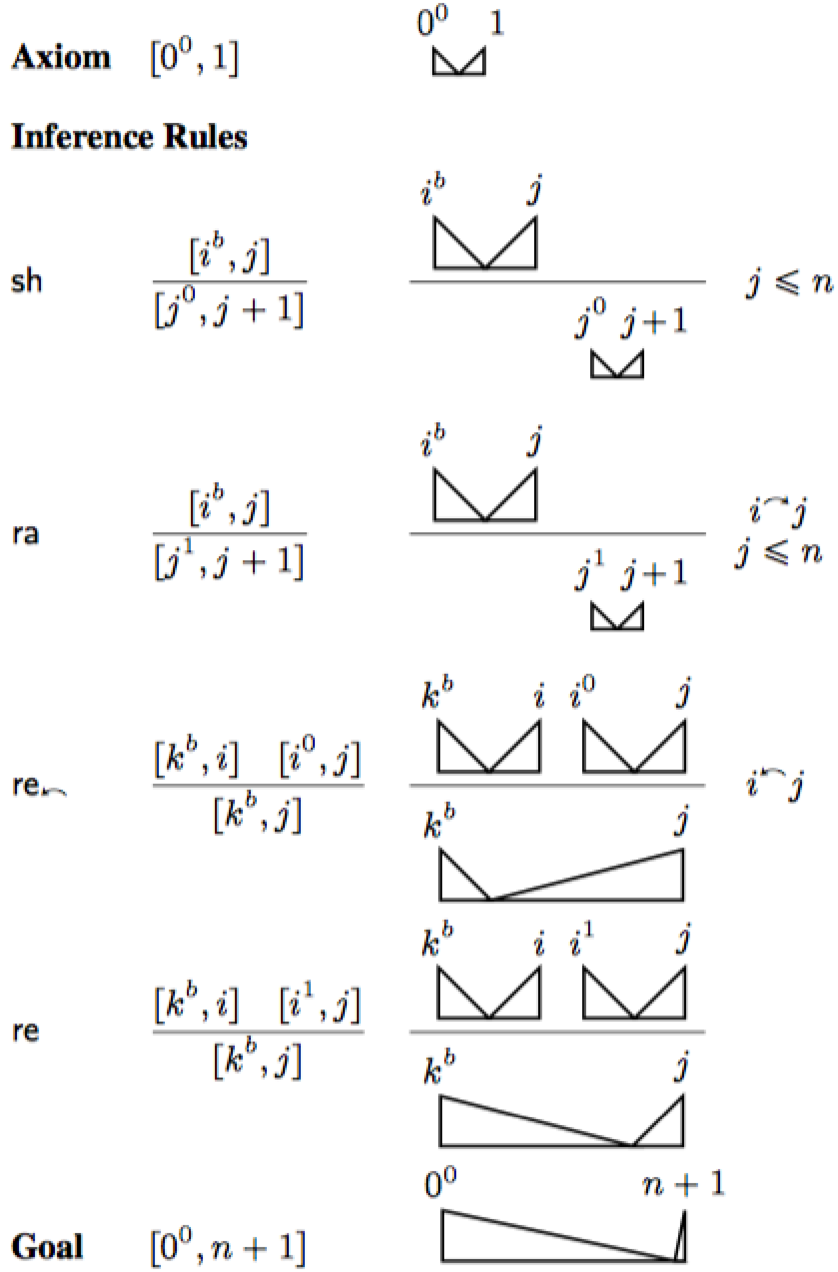
Shift: $[i^b, j] \Rightarrow [j^0, j+1]$

Reduce: $[i^b, m] \wedge [m^1, j] \Rightarrow [i^b, j]$

Right-Arc: $[i^b, j] \Rightarrow [j^1, j+1]$

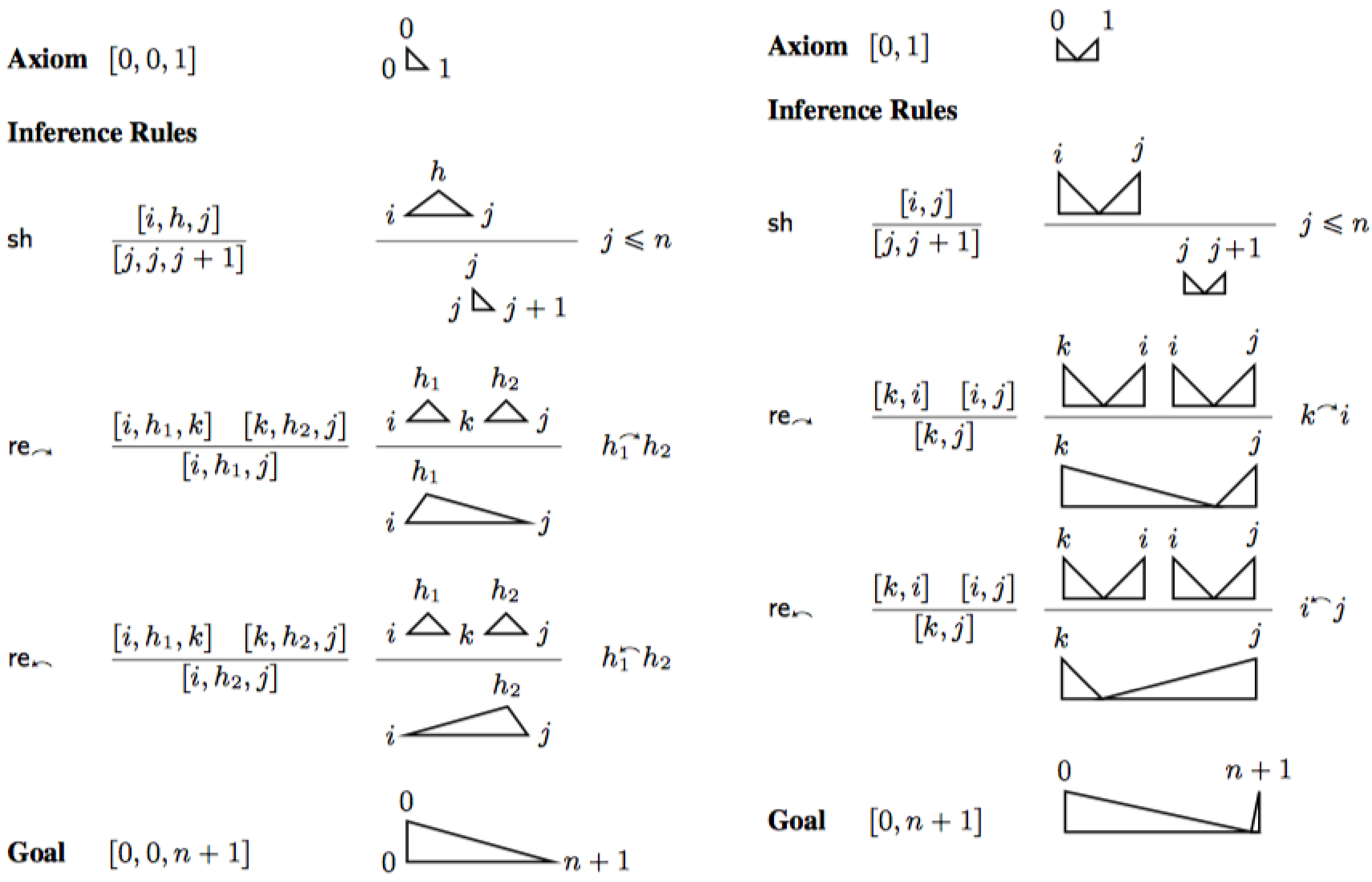
Left-Arc: $[i^b, m] \wedge [m^0, j] \Rightarrow [i^b, j]$

[Kuhlmann et al. 2011]


Axiom $[0^0, 1]$
Inference Rules
sh $\dfrac{[i^b, j]}{[j^0, j+1]}$ $j \leqslant n$
ra $\dfrac{[i^b, j]}{[j^1, j+1]}$ $i \rightarrow j$ $j \leqslant n$
re$_{\leftarrow}$ $\dfrac{[k^b, i] \quad [i^0, j]}{[k^b, j]}$ $i \leftarrow j$
re $\dfrac{[k^b, i] \quad [i^1, j]}{[k^b, j]}$
Goal $[0^0, n+1]$

(c) Arc-eager

Deduction System



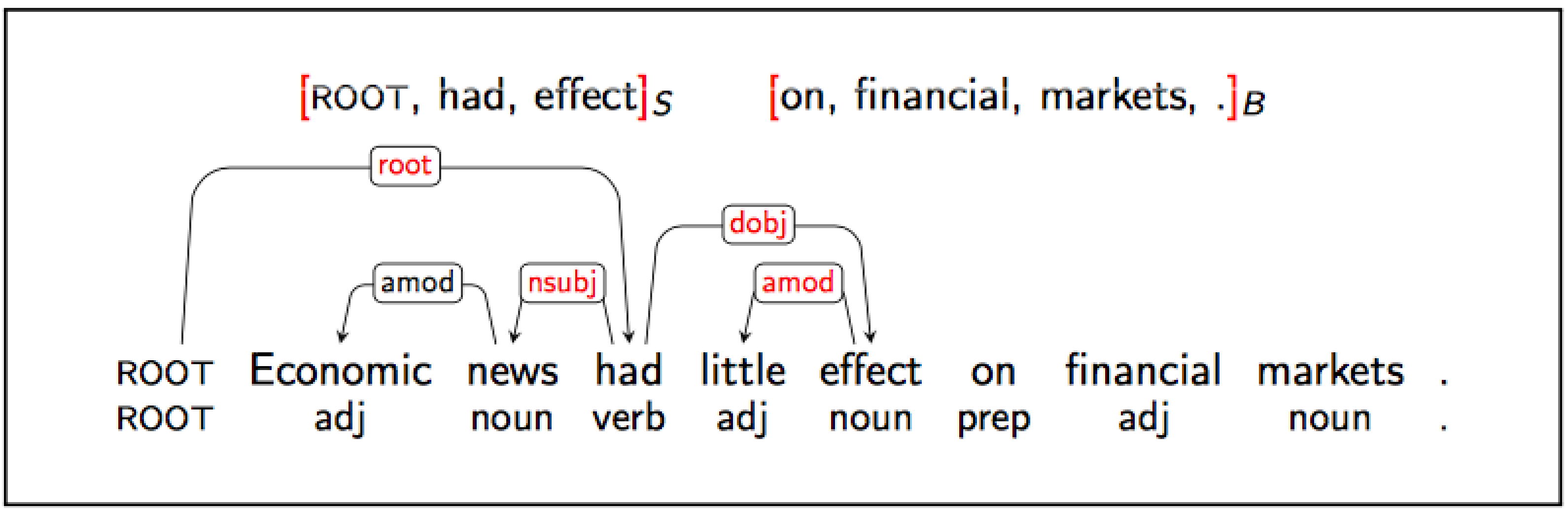
(a) Arc-standard

(b) Arc-hybrid

Exact Decoding

$$\frac{[i^b, j] : v}{[j^0, j+1] : 0}\,(\mathsf{sh}) \qquad \frac{[k^b, i] : v_1 \quad [i^0, j] : v_2}{[k^b, j] : v_1 + v_2 + \Delta}\,(\mathsf{re}_{\curvearrowleft})$$

Configuration



Global Training

$$\max_{\mathbf{t}} \left(F(\mathbf{t}) + cost(\mathbf{t}^{\mathrm{gold}}, \mathbf{t}) - F(\mathbf{t}^{\mathrm{gold}}) \right)$$

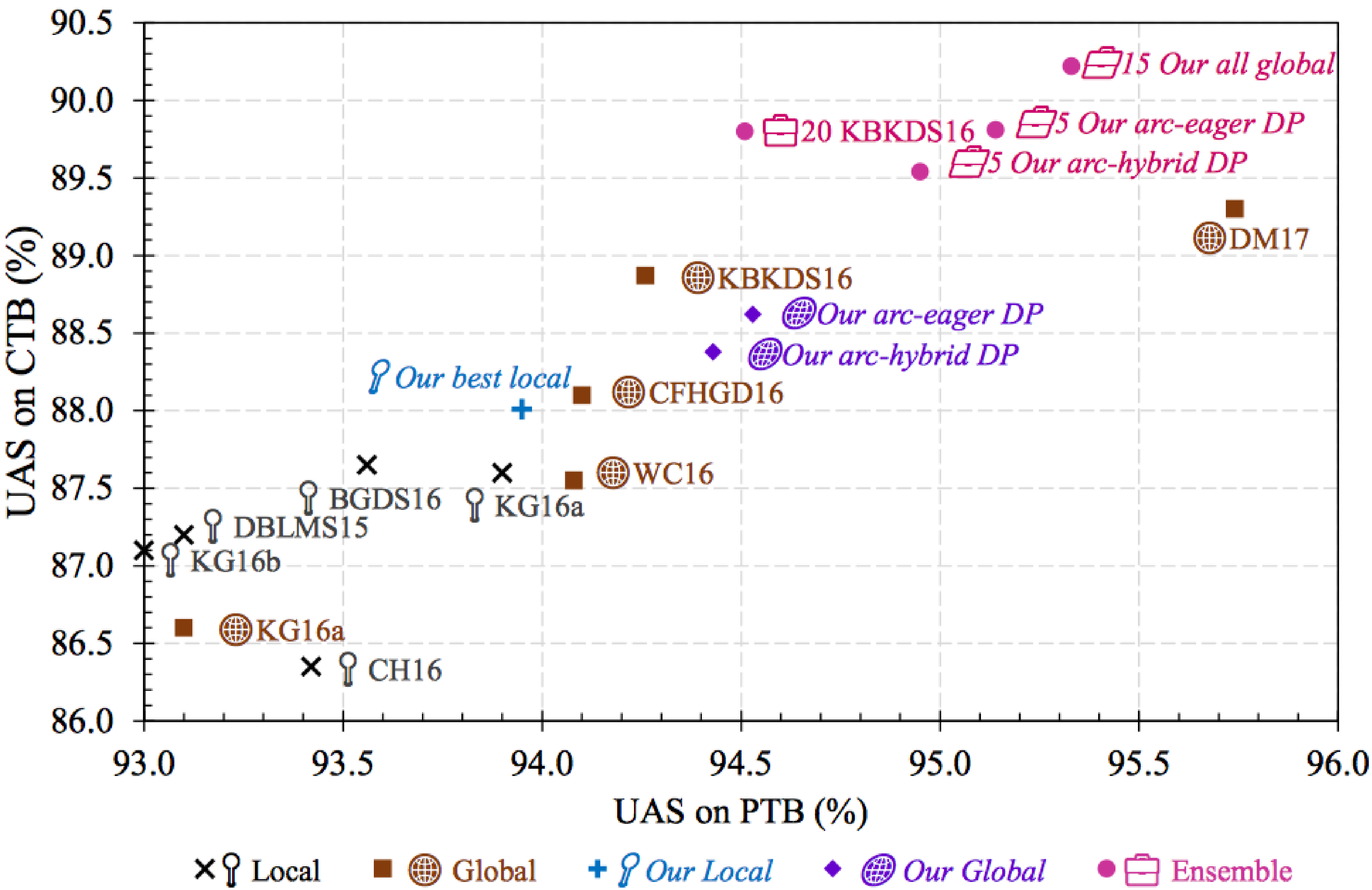
$$\frac{[k^b, i] : v_1 \quad [i^0, j] : v_2}{[k^b, j] : v_1 + v_2 + \Delta'} (\mathsf{re}_{\curvearrowleft})$$

$$\text{where } \Delta' = \Delta + \mathbf{1}\,(head(w_i) \neq w_j).$$

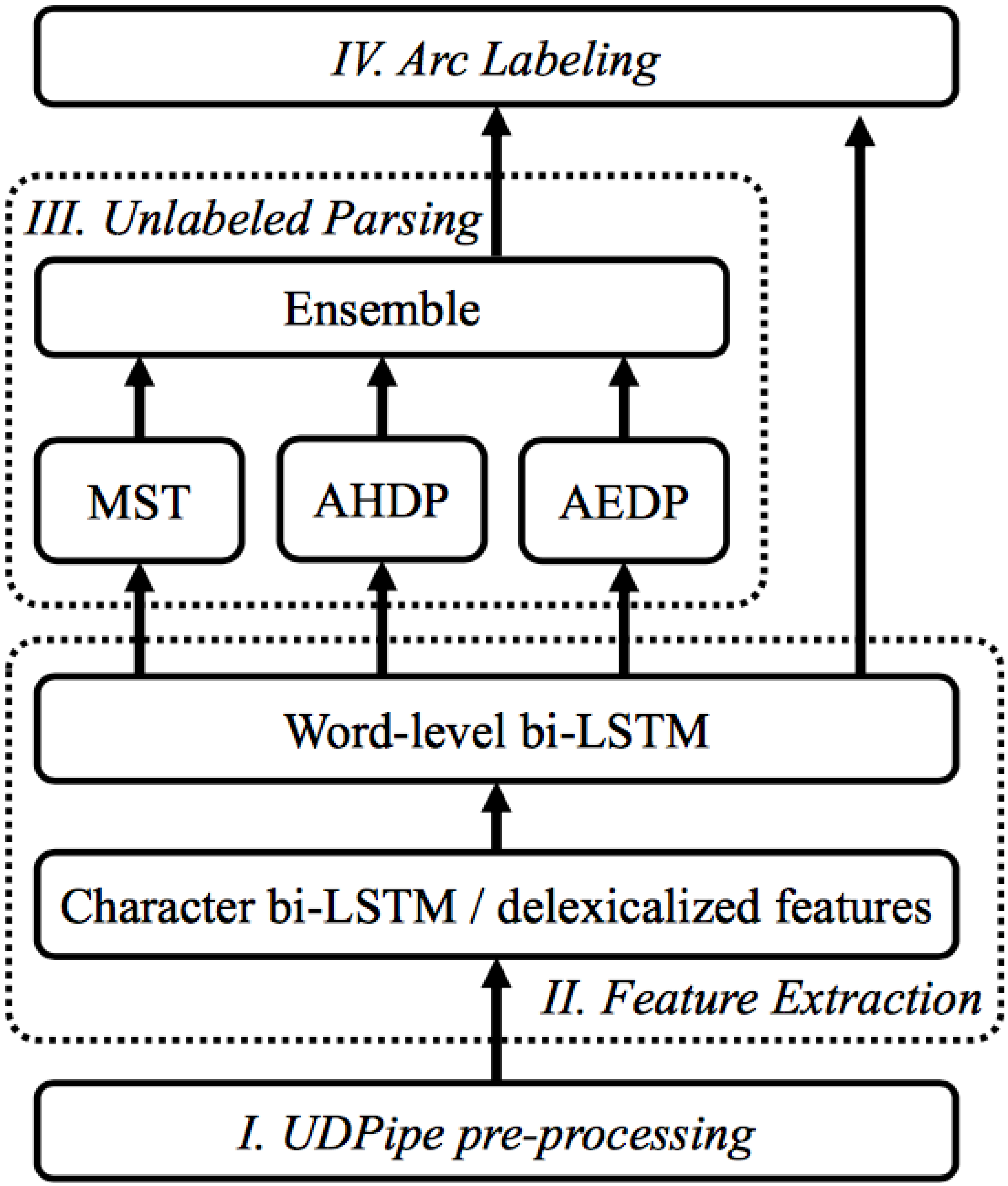
Experiments

Model	Training	Features	PTB		CTB	
			UAS (%)	UEM (%)	UAS (%)	UEM (%)
Arc-standard	Local	$\{\overleftrightarrow{s}_2, \overleftrightarrow{s}_1, \overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$93.95_{\pm 0.12}$	$52.29_{\pm 0.66}$	$88.01_{\pm 0.26}$	$36.87_{\pm 0.53}$
Arc-hybrid	Local	$\{\overleftrightarrow{s}_2, \overleftrightarrow{s}_1, \overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$93.89_{\pm 0.10}$	$50.82_{\pm 0.75}$	$87.87_{\pm 0.17}$	$35.47_{\pm 0.48}$
	Local	$\{\overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$93.80_{\pm 0.12}$	$49.66_{\pm 0.43}$	$87.78_{\pm 0.09}$	$35.09_{\pm 0.40}$
	Global	$\{\overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$94.43_{\pm 0.08}$	$53.03_{\pm 0.71}$	$88.38_{\pm 0.11}$	$36.59_{\pm 0.27}$
Arc-eager	Local	$\{\overleftrightarrow{s}_2, \overleftrightarrow{s}_1, \overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$93.80_{\pm 0.12}$	$49.66_{\pm 0.43}$	$87.49_{\pm 0.20}$	$33.15_{\pm 0.72}$
	Local	$\{\overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$93.77_{\pm 0.08}$	$49.71_{\pm 0.24}$	$87.33_{\pm 0.11}$	$34.17_{\pm 0.41}$
	Global	$\{\overleftrightarrow{s}_0, \overleftrightarrow{b}_0\}$	$\mathbf{94.53}_{\pm 0.05}$	$53.77_{\pm 0.46}$	$\mathbf{88.62}_{\pm 0.09}$	$\mathbf{37.75}_{\pm 0.87}$
Edge-factored	Global	$\{\overleftrightarrow{h}, \overleftrightarrow{m}\}$	$94.50_{\pm 0.13}$	$\mathbf{53.86}_{\pm 0.78}$	$88.25_{\pm 0.12}$	$36.42_{\pm 0.52}$

Experiments



CoNLL 2017 Shared Task



CoNLL 2017 Shared Task

Following Dozat and Manning (2017), we use a deep bi-affine scoring function:

$$\text{score}^{\text{MST}}(h, m) = v_h^\intercal U v_m + b_h \cdot v_h + b_m \cdot v_m + b$$

where

$$v_h = \text{MLP}^{\text{MST-head}}(\overleftrightarrow{h})$$
$$v_m = \text{MLP}^{\text{MST-mod}}(\overleftrightarrow{m})$$

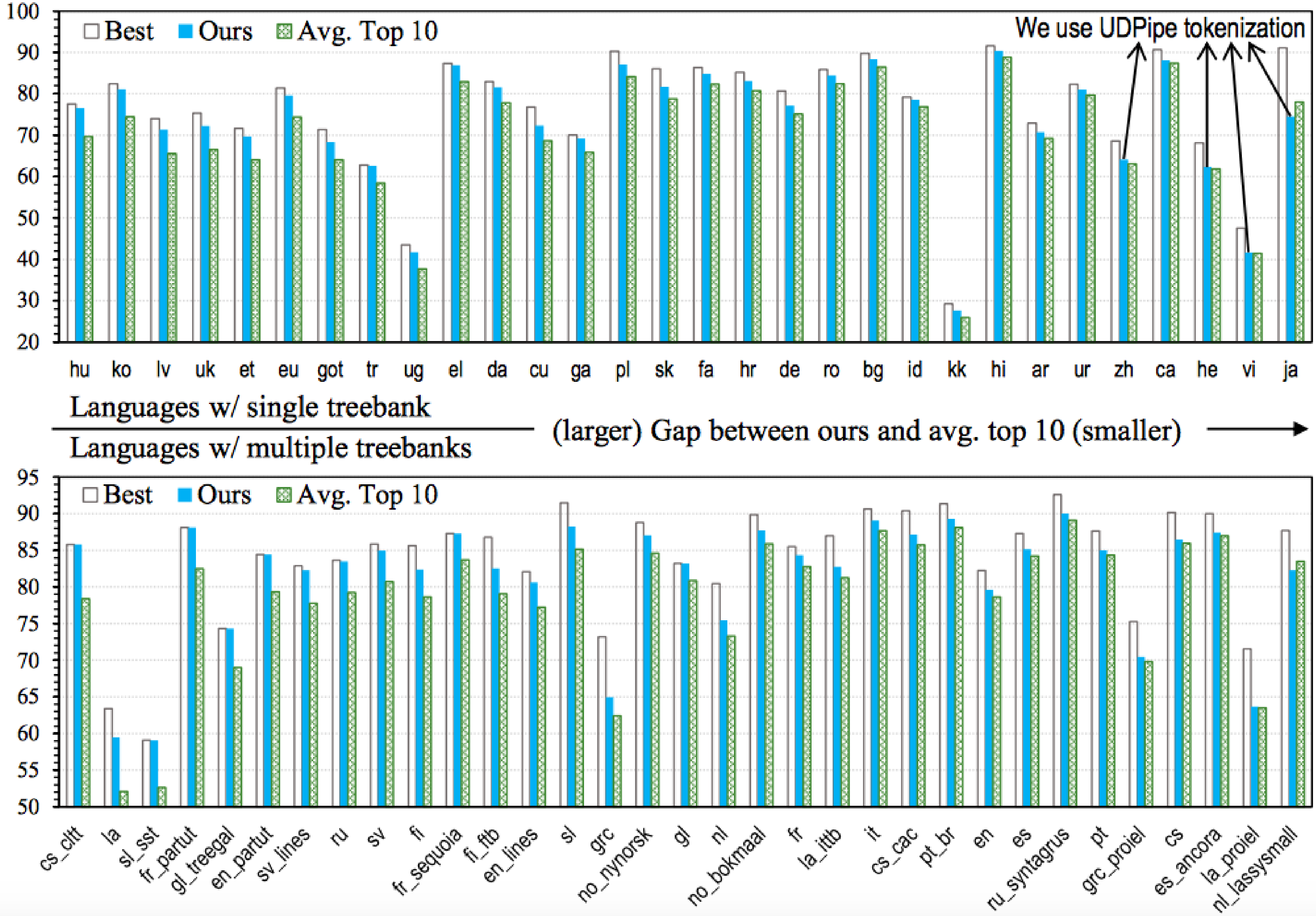
CoNLL 2017 Shared Task

	UAS F1	LAS F1	Official Ranking
Big Treebanks	85.16	79.85	2
Small Treebanks	70.59	61.49	1
PUD Treebanks	80.17	71.49	2
Surprise Languages	58.40	47.54	1
Overall	80.35	75.00	2

CoNLL 2017 Shared Task

Target	Source	UAS F1	LAS F1	Official Ranking
`bxr`	`hi`	50.79	31.98	2
`hsb`	`cs`	69.45	61.70	1
`kmr`	`fa`	54.51	47.53	1
`sme`	`fi`	58.85	48.96	1
Average		58.40	47.54	1

CoNLL 2017 Shared Task



Thank you!

Q&A