

**Prune & Score Framework Problem Setup Coreference Resolution** is the task of clustering a set of **□ Key Idea:** Divide-and-conquer by learning two functions; mentions in the text such that all mentions in the same cluster refer  $\succ$  A pruning function  $F_{prune}$  to prune all the bad decisions based on the specified pruning parameter *b*. to the same entity.  $\triangleright$  A scoring function  $F_{score}$  to select the best decision from the ary Clinton remaining actions. irst Lady State:  $s = \{C_1, C_2, C_3, C_4, C_5, C_6\}$  Actions:  $A(s) = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$ Her  $\begin{pmatrix} a_2 & a_1 & a_7 \\ 2.5 & 2.2 & 19 \end{pmatrix} \begin{pmatrix} a_5 & a_6 & a_3 & a_4 \\ 1.5 & 1.4 & 0.7 & 0.4 \end{pmatrix} \leftarrow \mathcal{F}_{prune}$  values Pruning: Keeping top b. She nce Output  $A'(s) = \{a_2, a_1, a_7\}$ tput  $\hat{y}$ Scoring: Picking the best.  $\begin{array}{cccc} a_1 & a_2 & a_7 \\ a_5 & a_1 & a_6 \end{array} \leftarrow \mathcal{F}_{score} \text{ values} \end{array}$ **Decision**:  $a_1$  is the best action for state sg, learn a **Representational Power Proposition**: Let  $F_{prune}$  and  $F_{score}$  be in the same function space. For all + (e.g. learning problems,  $\min_{F_{score}} \varepsilon(F_{score}, F_{score}) \geq \min_{(F_{prune}, F_{score})} \varepsilon(F_{prune}, F_{score}).$ Moreover there exist learning problems for which min  $\varepsilon(F_{score}, F_{score})$ can be arbitrarily worse than  $\min_{(F_{prune},F_{score})} \varepsilon(F_{prune},F_{score})$ . Loss Decomposition and Learning right. Loss Decomposition made. Overall expected loss  $\varepsilon$  equals the error due to pruning the target output ( $\varepsilon_{prune}$ ), plus the error due to not selecting the best output nention. within the pruned space ( $\varepsilon_{score}$ ). ter C.  $\mathcal{E} = \mathcal{E}_{prune} + \mathcal{E}_{score|prune}$ m. Pruning and Scoring Function Learning **Stage 1:**  $\hat{\mathcal{F}}_{prune} \approx \arg \min_{\mathcal{F}_{prune} \in \mathbf{F}_{\mathbf{p}}} \epsilon_{prune} \longleftarrow$ d that ist the **Stage 2:**  $\hat{\mathcal{F}}_{score} \approx \arg \min_{\mathcal{F}_{score} \in \mathbf{F}_s} \epsilon_{score} | \hat{\mathcal{F}}_{prune}$ ng Reductions to Rank-learning **1. Pruning Function Learning**  $\rightarrow A(S_{r})$  $NEW(m_7)$ ikh₹∽

"[Barack Obama] nominated [Hillary Clinton] as his [secretary of state] on Monday. [He] chose [her] because [she] had foreign affair experience as a former [First Lady]." Extracted Mentions Input $\chi$	Barack Obama He Secretary of state Coreferen Out
<b>Learning:</b> Given a set of input function $\mathcal{F}: \mathcal{X} \xrightarrow{\longrightarrow} \mathcal{Y}$ to make $\mathfrak{g}$	t-output pairs for training predictions on new inputs
<b>D</b> Evaluation: against a non-ne <i>BCubed</i> ).	gative loss $L(x, y, \hat{y}) \in R$
<b>Greedy Searc</b>	h Formulation
<ul> <li>Greedy Search processes each choose actions greedily according "Processed" means an decision</li> <li>Search Space</li> <li>State S: Partial clustering of all</li> <li>Action: MERGE(m, C): merge NEW(m): start a new constrained to be started to</li></ul>	ach mention from left to a ng to a heuristic. of that mention has been mentions up to current m mention <i>m</i> into the clust cluster that only contains
[Sharm el-Sheikh]] summit to be held [Ramallah], [around 500 people] to slogans denouncing the summit At depth 7	d in [Egypt] tomorrow Monday. In ok to [the town]'s streets chanting Actions
$merge(m_7, C_1)$ $a_2$ $a_3$	
Ramallah (West Bank) C1 C1 C2 C2 C3 Processed mentions and their clu	S Palestinians C 4 M 7 Usters C 4 C 4 C 4 C 4 C 4 C 4 C 4 C 4

Each depth will have a corresponding processing mention; The learned heuristic will pick the best action for that mention.

State

# **Prune-and-Score: Learning for Greedy Coreference Resolution** Chao Ma, Janardhan Rao Doppa, J. Walker Orr, Prashanth Mannem Xiaoli Fern, Tom Dietterich and Prasad Tadepalli

**Current processing mention** 

Generate ranking

in top b

example: ranking  $a^*$ 

 $\mathcal{A}_{10}$ 



## **True best action**

 $a_7$ 

label feature vector

 $\phi(S_t, a_t)$ 

 $\phi(s_t, a_1)$ 

 $\phi(s_t, a_2)$ 

 $\phi(s_t, a_3)$ 

 $\phi(s_t, a_{10})$ 





## **Experiment Results**

## **D** Experiment Setups

- features; and one **NEW indicator** feature.

## Coreference Resolution Results

	OntoNotes 5.0 Test									
	System Mentions				Gold Mentions					
F-1 score	MUC	BCube	CEAF_e	CoNLL	MUC	BCube	CEAF_e	CoNLL		
Prune-Score	72.84	57.94	53.91	61.56	86.96	76.49	77.33	80.26		
Only Scoring	67.98	54.42	53.79	58.73	85.73	74.38	74.62	78.24		
HOTCoref	70.72	58.58	55.61	61.63	-	-	-	-		
Berkeley	70.82	58.14	55.27	61.41	87.46	76.63	76.40	80.16		
UIUC	69.48	57.44	53.07	60.00	84.80	78.74	68.75	77.43		
Stanford	64.71	52.26	49.32	55.43	83.64	74.81	66.98	75.14		

- systems can also benefit from our pruning idea.

## **D** Performance with Different Pruning Parameter *b*





Performance shows Prune-and-Score is robust to the pruning parameter **b**.



# emn p<sub>2014</sub> Doha, Qatar

• **Datasets** OntoNotes 5: Train/Dev/Test: 2802/343/345 documents. • **Base Rank-Learner LambdaMART** implemented in *RankLib*. • Feature Set Employ the same features as *Easyfirst* [Stoyanov et. al., 2012] System, which used 90 mention-pair features; 49 entity-pair

Prune-and-Score performs better than Only-Scoring. This shows the benefit of learning with pruning rules. Other coreference resolution

Prune-and-Score is comparable or better than the state-of-the-art.

Behavior of Prune-and-Score depends on the pruning parameter b: