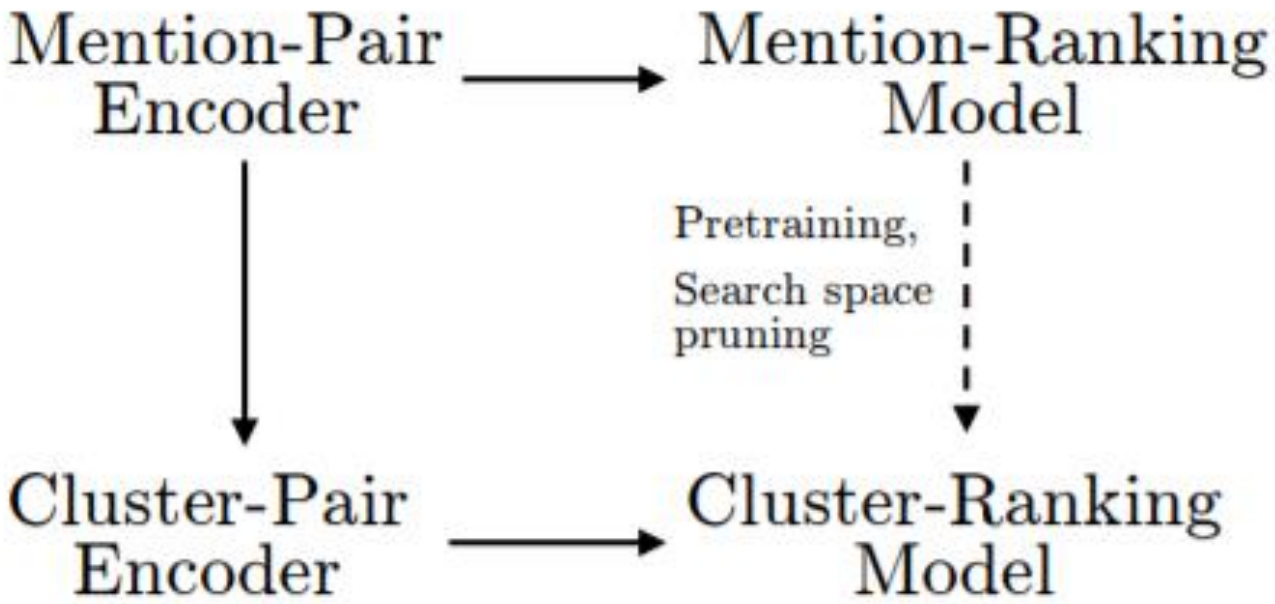
Improving Coreference Resolution by Learning Entity-Level Distributed Representations

K. Clark and C. Manning, ACL 2016

Coreference from clustering – Why?

- - Learns entity-level
  - **Bill Clinton** says…
  - **Clinton**…, **she** is very happy to be home.
  - {Bill Clinton}, {Clinton, she}.

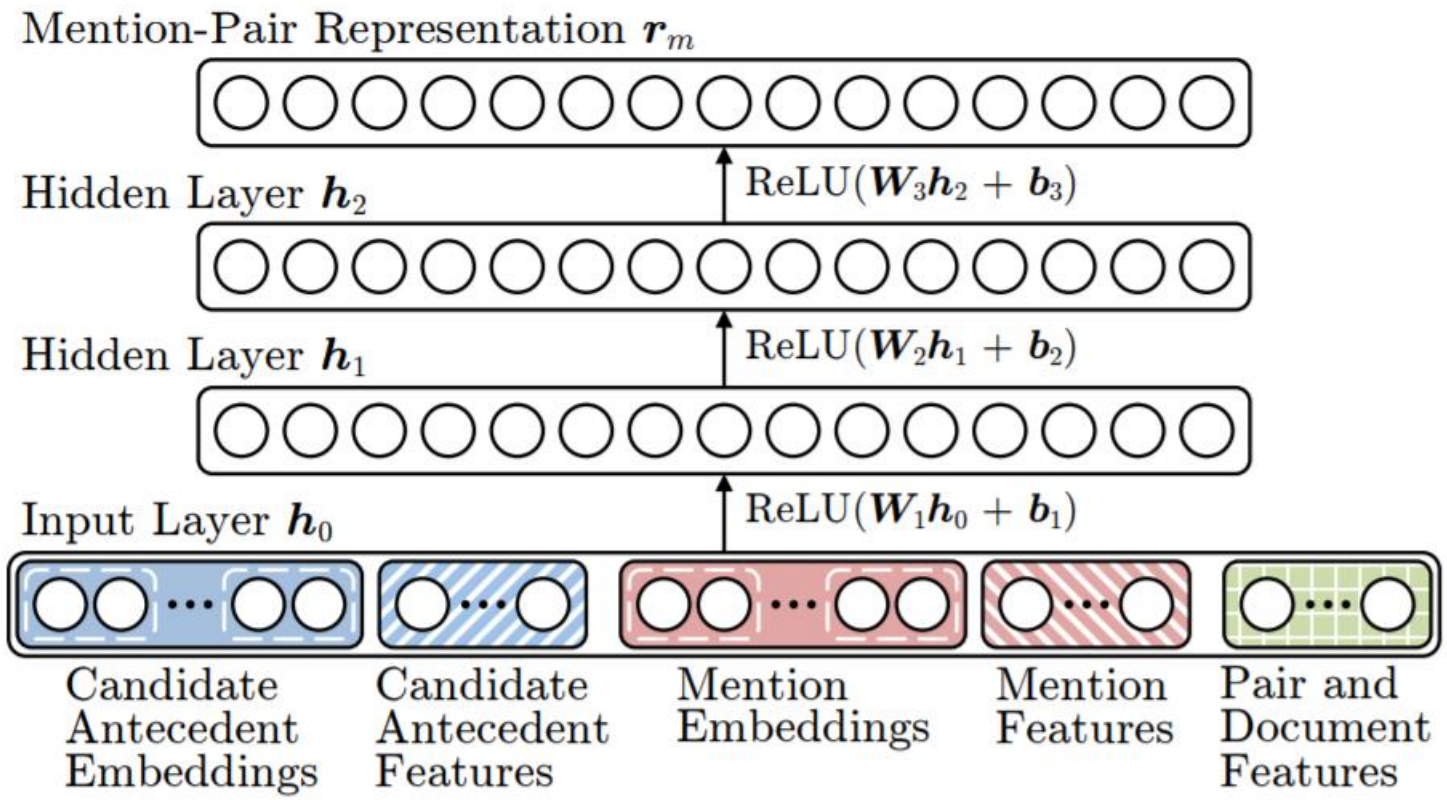
Model – Overall Design



Model – Mention Pair Encoder

- Obama says the U.S. government has killed Bin Laden.
  - Obama: {NA}
  - U.S. government: {Obama}
  - Bin Laden: {U.S. government, Obama}

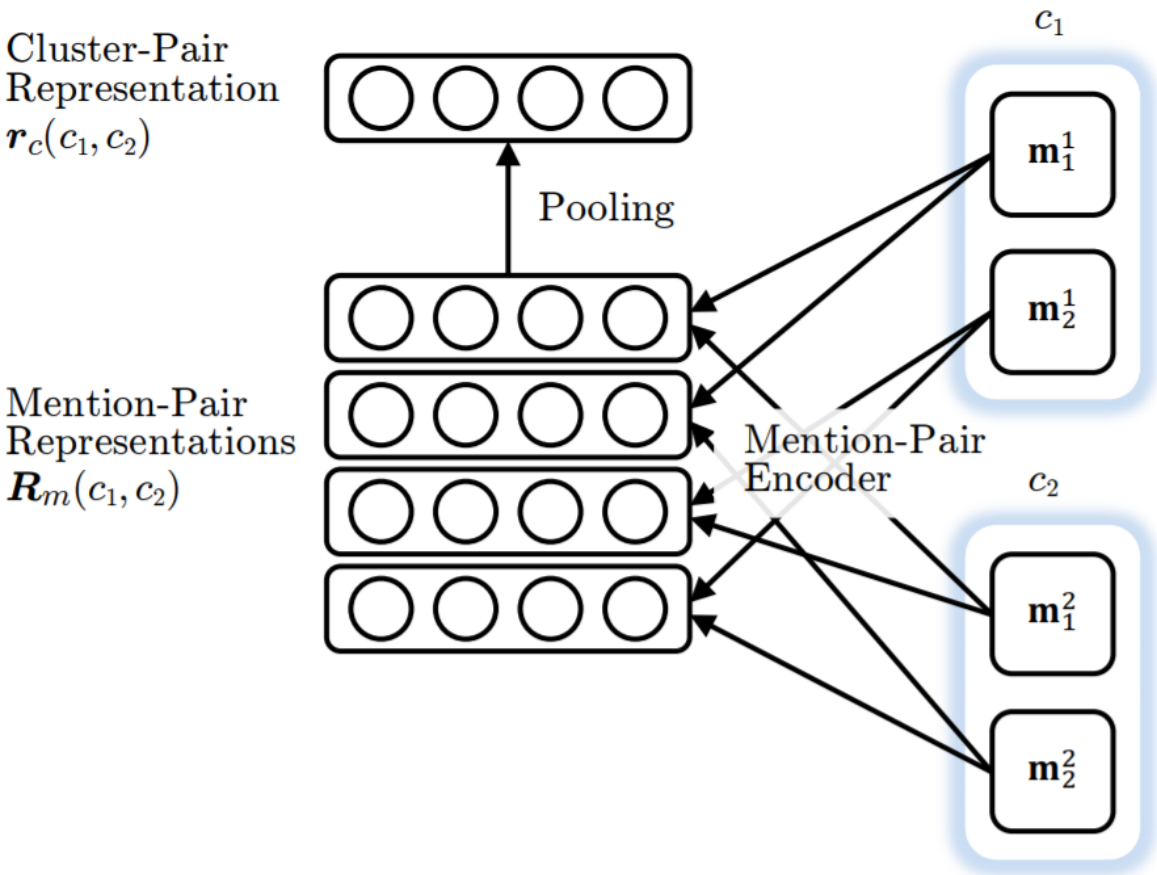
Model – Mention Pair Encoder



Model – Mention Pair Encoder

- Mention Features:
  - Type / position /…
- Pair&Document Features:
  - Genre / Distance / Speaker / String Match /
- Mention Embeddings:
  - head word / dependency parent / first(last word) / two preceding(following) words / averaged five preceding(following) words / averaged all words(mention,sentence,document) /

Model – Cluster Pair Encoder



Model – Mention Pair Ranker

$$\hat{t}_i = \underset{t \in \mathcal{T}(m_i)}{\operatorname{argmax}}\, s_m(t, m_i)$$

$$\sum_{i=1}^{N} \max_{a \in \mathcal{A}(m_i)} \Delta(a, m_i)(1 + s_m(a, m_i) - s_m(\hat{t}_i, m_i))$$

$$\Delta(a, m_i) = \begin{cases} \alpha_{\text{FN}} & \text{if } a = \text{NA} \ \wedge \mathcal{T}(m_i) \neq \{\text{NA}\} \\ \alpha_{\text{FA}} & \text{if } a \neq \text{NA} \ \wedge \mathcal{T}(m_i) = \{\text{NA}\} \\ \alpha_{\text{WL}} & \text{if } a \neq \text{NA} \ \wedge a \notin \mathcal{T}(m_i) \\ 0 & \text{if } a \in \mathcal{T}(m_i) \end{cases}$$

Model – Cluster Ranking

$$\pi(\textsc{Merge}[c_m, c]|x) \propto e^{s_c(c_m,c)}$$

$$\pi(\textsc{Pass}|x) \propto e^{s_{\textsc{na}}(m)}$$

- Easy First
  - Make easy decisions first
  - Delay hard ones to the last
  - Intuition?
- - Deep Learning to Search
  - Decisions made based on previous decisions

Model – Deep Learning to Search

for $i = 1$ **to** num_epochs **do**
  Initialize the current training set $\Gamma = \emptyset$
  for each example $(x, y) \in \mathcal{D}$ **do**
    Run the policy π to completion from start state x to obtain a trajectory of states $\{x_1, x_2, ..., x_n\}$
    for each state x_i in the trajectory **do**
      for each possible action $u \in U(x_i)$ **do**
        Execute u on x_i and then run the reference policy π^{ref} until reaching an end state e
        Assign u a cost by computing the loss on the end state: $l(u) = \mathcal{L}(e, y)$
      end for
      Add the state x_i and associated costs l to Γ
    end for
  end for
  Update π with gradient descent, minimizing $\sum_{(x,l)\in\Gamma} \sum_{u\in U(x)} \pi(u|x)l(u)$.
end for

Model – Deep Learning to Search

- Run current policy from the start state to end
- Compute loss and update policy with gradient descent
- Expose to mistake, learns how to cope

Results

Model	English F_1	Chinese F_1
Full Model	65.52	64.41
– MENTION	–1.27	–0.74
– GENRE	–0.25	–2.91
– DISTANCE	–2.42	–2.41
– SPEAKER	–1.26	–0.93
– MATCHING	–2.07	–3.44

Table 1: CoNLL F_1 scores of the mention-ranking model on the dev sets without mention, document genre, distance, speaker, and string matching hand-engineered features.

Model	English F_1	Chinese F_1
Full Model	66.01	64.86
– PRETRAINING	–5.01	–6.85
– EASY-FIRST	–0.15	–0.12
– L2S	–0.32	–0.25

Table 3: CoNLL F_1 scores of the cluster-ranking model on the dev sets with various ablations.

Results

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Avg. F_1
CoNLL 2012 English Test Data										
Clark and Manning (2015)	76.12	69.38	72.59	65.64	56.01	60.44	59.44	52.98	56.02	63.02
Peng et al. (2015)	–	–	72.22	–	–	60.50	–	–	56.37	63.03
Wiseman et al. (2015)	76.23	69.31	72.60	66.07	55.83	60.52	59.41	54.88	57.05	63.39
Wiseman et al. (2016)	77.49	69.75	73.42	66.83	56.95	61.50	62.14	53.85	57.70	64.21
NN Mention Ranker	79.77	69.10	74.05	69.68	56.37	62.32	63.02	53.59	57.92	64.76
NN Cluster Ranker	78.93	69.75	**74.06**	70.08	56.98	**62.86**	62.48	55.82	**58.96**	**65.29**
CoNLL 2012 Chinese Test Data										
Chen & Ng (2012)	59.92	64.69	62.21	60.26	51.76	55.69	51.61	58.84	54.99	57.63
Björkelund & Kuhn (2014)	69.39	62.57	65.80	61.64	53.87	57.49	59.33	54.65	56.89	60.06
NN Mention Ranker	72.53	65.72	68.96	65.49	56.87	60.88	61.93	57.11	59.42	63.09
NN Cluster Ranker	73.85	65.42	**69.38**	67.53	56.41	**61.47**	62.84	57.62	**60.12**	**63.66**

Takeaway

- Clustering Coreference – Learns entity level information
- Deep learns policy with easy-first