Higher-order Coreference Resolution with Coarse-to-fine Inference

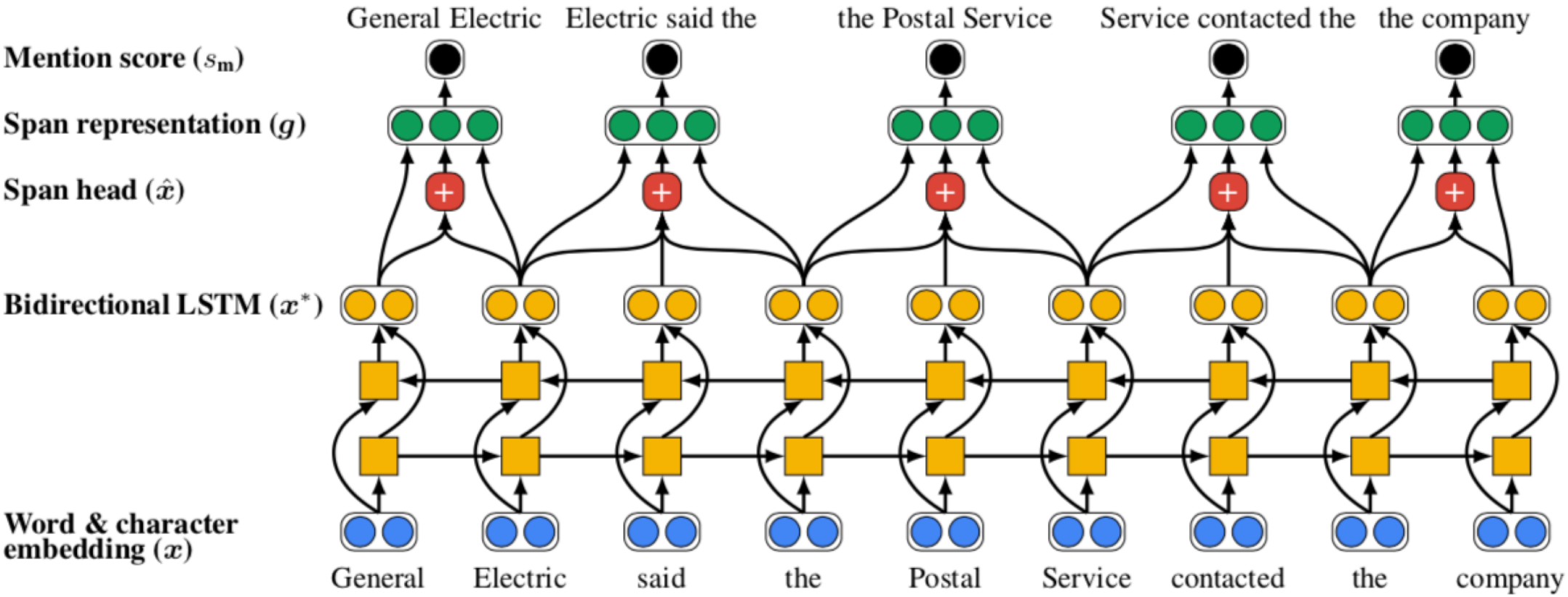
Kenton Lee, Luheng He, Luke Zettlemoyer

NAACL 2018

Presenter: Liyan Xu

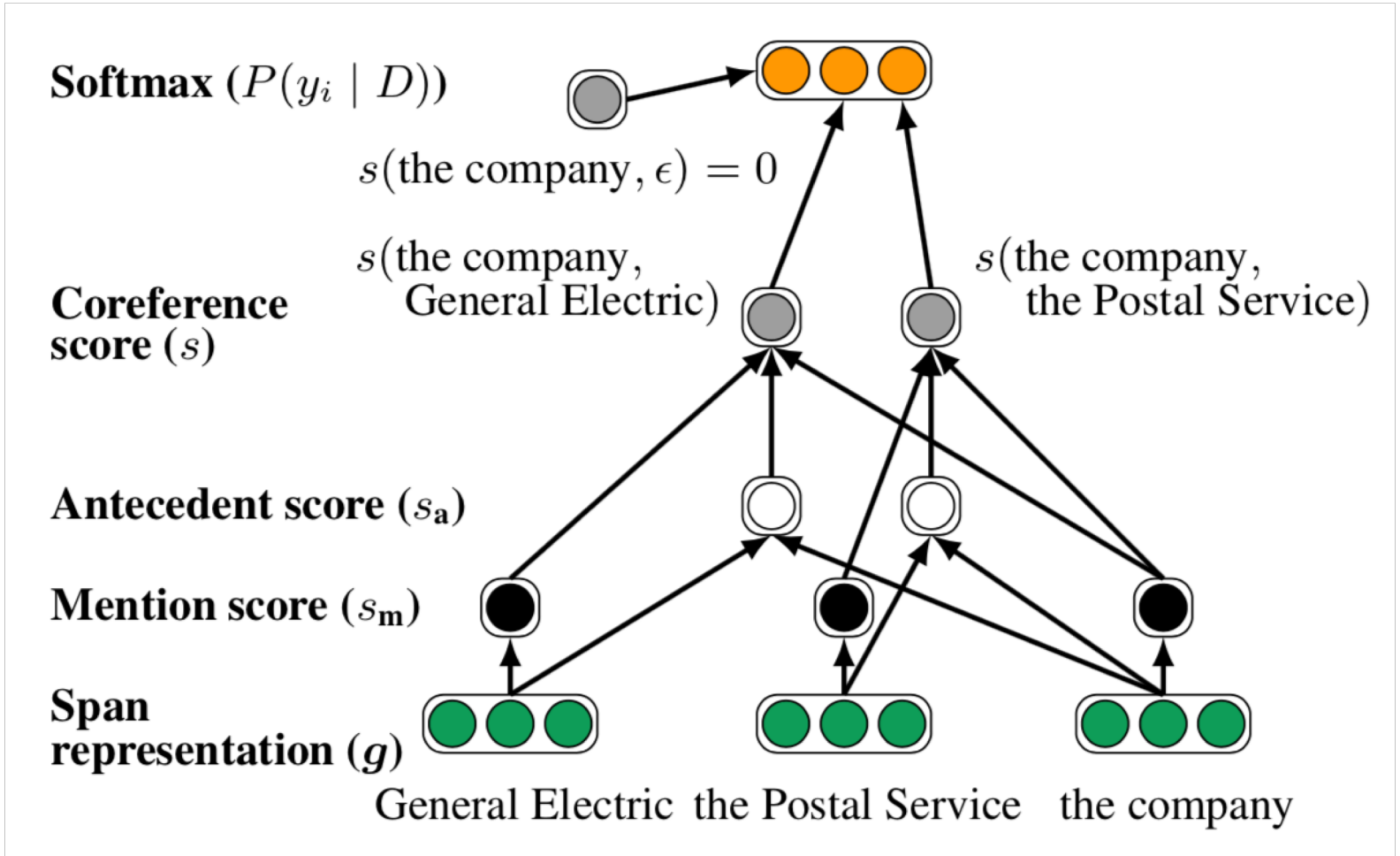
Review:E2E Neural Coref Model

- Span Representation



Review:E2E Neural Coref Model

- Final Antecedent Score



Review:E2E Neural Coref Model

- Score objective: optimize the marginal likelihood of all correct antecedents (best antecedent is latent) implied by the gold clustering.

$$\log \prod_{i=1}^{N} \sum_{\hat{y} \in \mathcal{Y}(i) \cap \text{GOLD}(i)} P(\hat{y})$$

Review:E2E Neural Coref Model

- Pruning: original span size $O(T^2)$ for doc length T
  - Limit max span length with up to L words
  - Only keep ƛT spans with the highest mention scores s_m
  - **Only consider up to K antecedents for each span**
  - Others

Higher-order Coreference Resolution

- Motivation: modeling these decisions at the document-level requires explicit inference due to the potentially very large surface distance between mentions.

Speaker 1: Um and **[I]** think that is what's - Go ahead Linda.
Speaker 2: Well and uh thanks goes to **[you]** and to the media to help us... So our hat is off to **[all of you]** as well.

Figure 1: Example of consistency errors to which first-order span-ranking models are susceptible. Span pairs (**I**, **you**) and (**you**, **all of you**) are locally consistent, but the span triplet (**I**, **you**, **all of you**) is globally inconsistent. Avoiding this error requires modeling higher-order structures.

Higher-order Coreference Resolution

- The learned gate vector **f** determines for each dimension whether to keep the current span in- formation or to integrate new information from its expected antecedent.

$$P_n(y_i) = \frac{e^{s(\boldsymbol{g}_i^n, \boldsymbol{g}_{y_i}^n)}}{\sum_{y \in \mathcal{Y}(i)} e^{s(\boldsymbol{g}_i^n, \boldsymbol{g}_y^n))}}$$

$$\boldsymbol{a}_i^n = \sum_{y_i \in \mathcal{Y}(i)} P_n(y_i) \cdot \boldsymbol{g}_{y_i}^n$$

$$\boldsymbol{f}_i^n = \sigma(\mathbf{W}_{\mathrm{f}}[\boldsymbol{g}_i^n, \boldsymbol{a}_i^n])$$

$$\boldsymbol{g}_i^{n+1} = \boldsymbol{f}_i^n \circ \boldsymbol{g}_i^n + (\mathbf{1} - \boldsymbol{f}_i^n) \circ \boldsymbol{a}_i^n$$

Review: Graph Attention Network

- Node Aggregation

$$\vec{h}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}\mathbf{W}\vec{h}_j\right)$$

- Shared attention vector ***a*** with non-linearity (single-layer FFNN) to compute coefficients

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$

Coarse-to-fine Antecedent Pruning

- Motivation: to reduce computation and keep accuracy.
  - With iterative inference, the antecedent score is the bottleneck.
  - Ability to expand coreference link between any two spans in the document (removing up to K antecedents constraints).

- Key: alternate bilinear scoring function, less accurate but much more efficient to compute. Use it to compute a rough sketch of likely antecedents.

$$s_c(i,j) = \boldsymbol{g}_i^\top \mathbf{W}_c \, \boldsymbol{g}_j$$

Coarse-to-fine Antecedent Pruning

- Pruning Steps

1. Keep top M span based on mention score.
2. Keep top K span based on the following factors:

$$s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{c}}(i, j).$$

3. The overall coreference s(i,j) is computed based on the remaining span pairs.

Additional Embedding

- Deep contextualized word representations
- https://arxiv.org/abs/1802.05365

Result

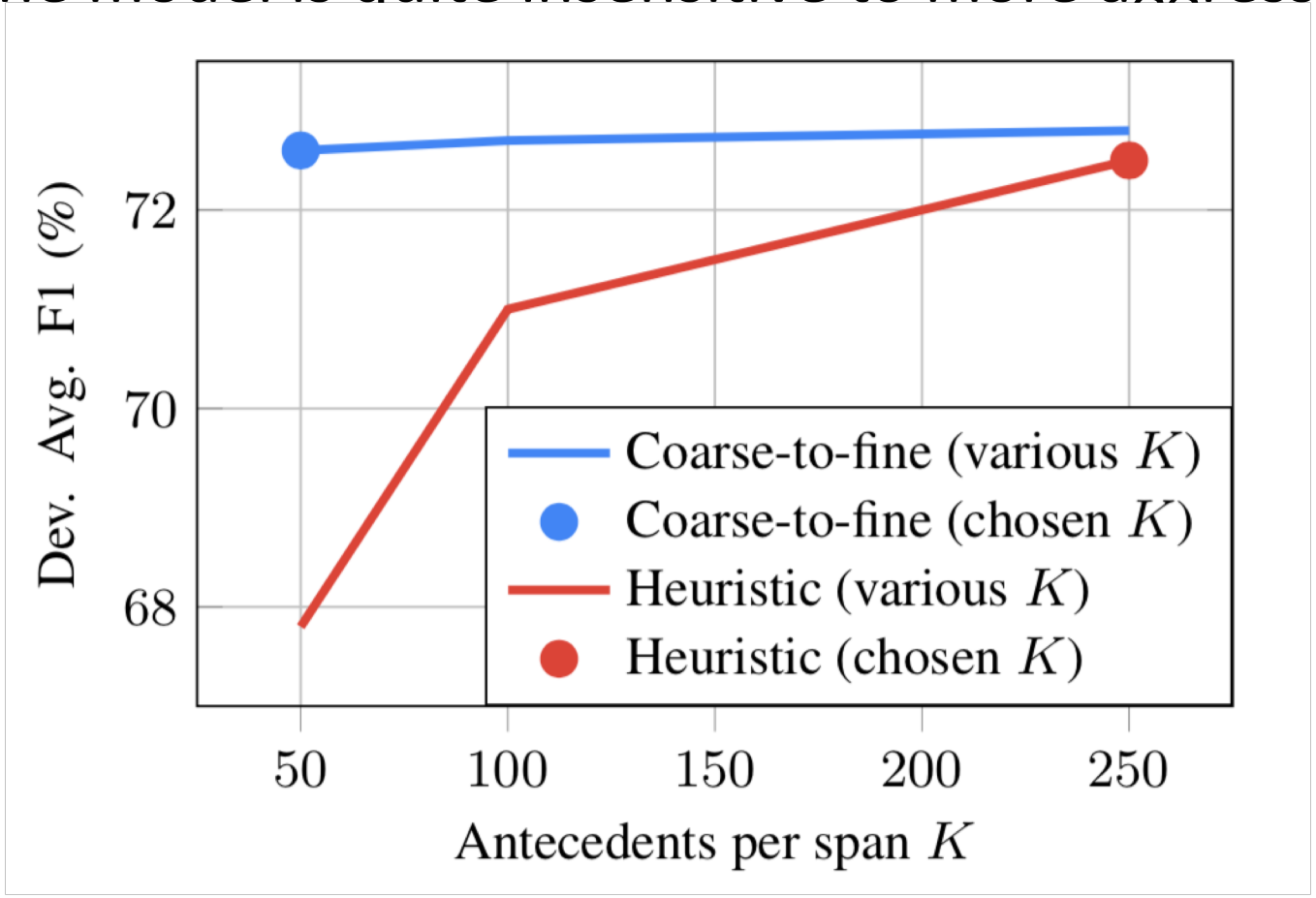
- Compared to the heuristic pruning with up to 250 antecedents, coarse-to-fine model only computes the expensive scores for 50 antecedents.

- Despite less computation, it outperforms heuristic pruning model, because it is able to predict a coreference link between any two spans in the document.

- Further improvement by including the second-order inference.

Result

- Coarse-to-fine model is quite insensitive to more aggressive pruning.



Result

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Lee et al. (2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
+ ELMo (Peters et al., 2018)	80.1	77.2	78.6	69.8	66.5	68.1	66.4	62.9	64.6	70.4
+ hyperparameter tuning	80.7	78.8	79.8	71.7	68.7	70.2	67.2	66.8	67.0	72.3
+ coarse-to-fine inference	80.4	**79.9**	80.1	71.0	**70.0**	70.5	67.5	**67.2**	67.3	72.6
+ second-order inference	**81.4**	79.5	**80.4**	**72.2**	69.5	**70.8**	**68.2**	67.1	**67.6**	**73.0**