End-to-end Neural Coreference Resolution

Kenton Lee, Luheng He, Mike Lewis and Luke Zettlemoyer

Presented by Wenxuan Hu

Introduction

Coreference Resolution

The task of finding all expressions that refer to the same entity in a text.

"I voted for Nader because he was most aligned with my values," she said.

Introduction

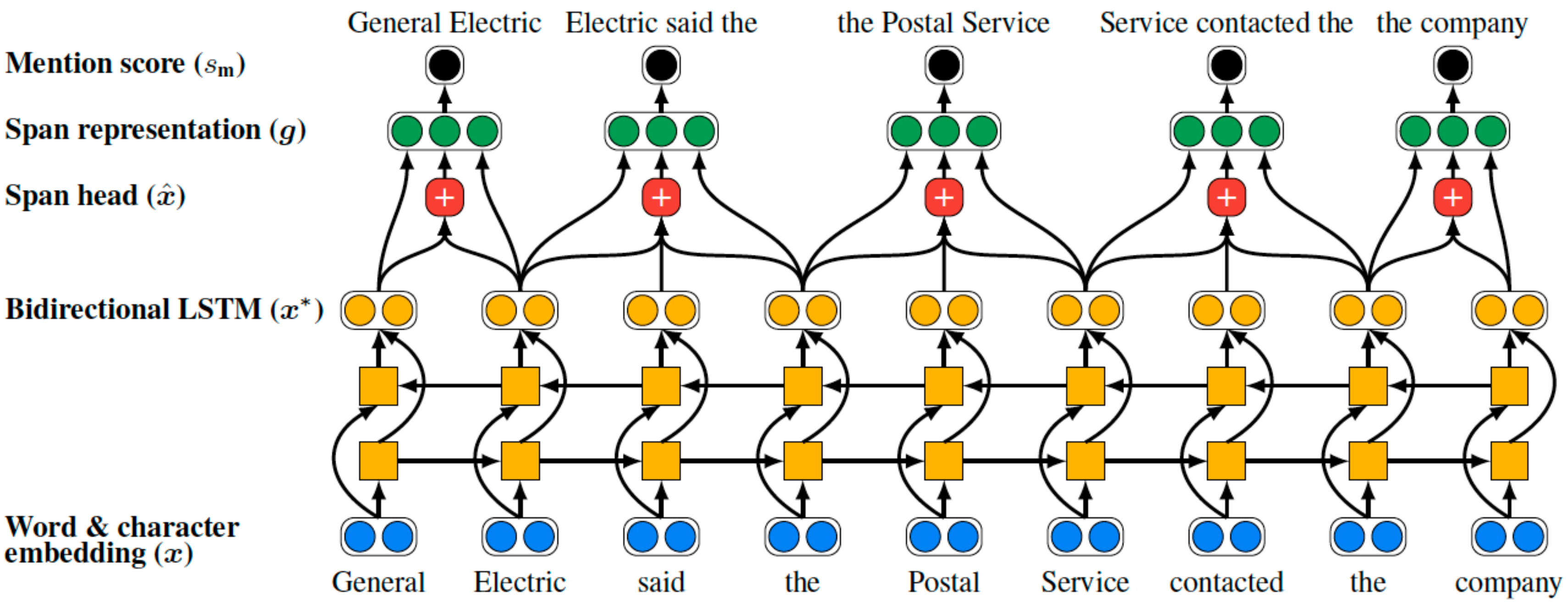
First end-to-end coreference resolution model

- Significantly outperforms all previous work
- Without using a syntactic parser or hand-engineered mention detector
- Instead, used a novel attention mechanism for head words and span-ranking model for mention detection

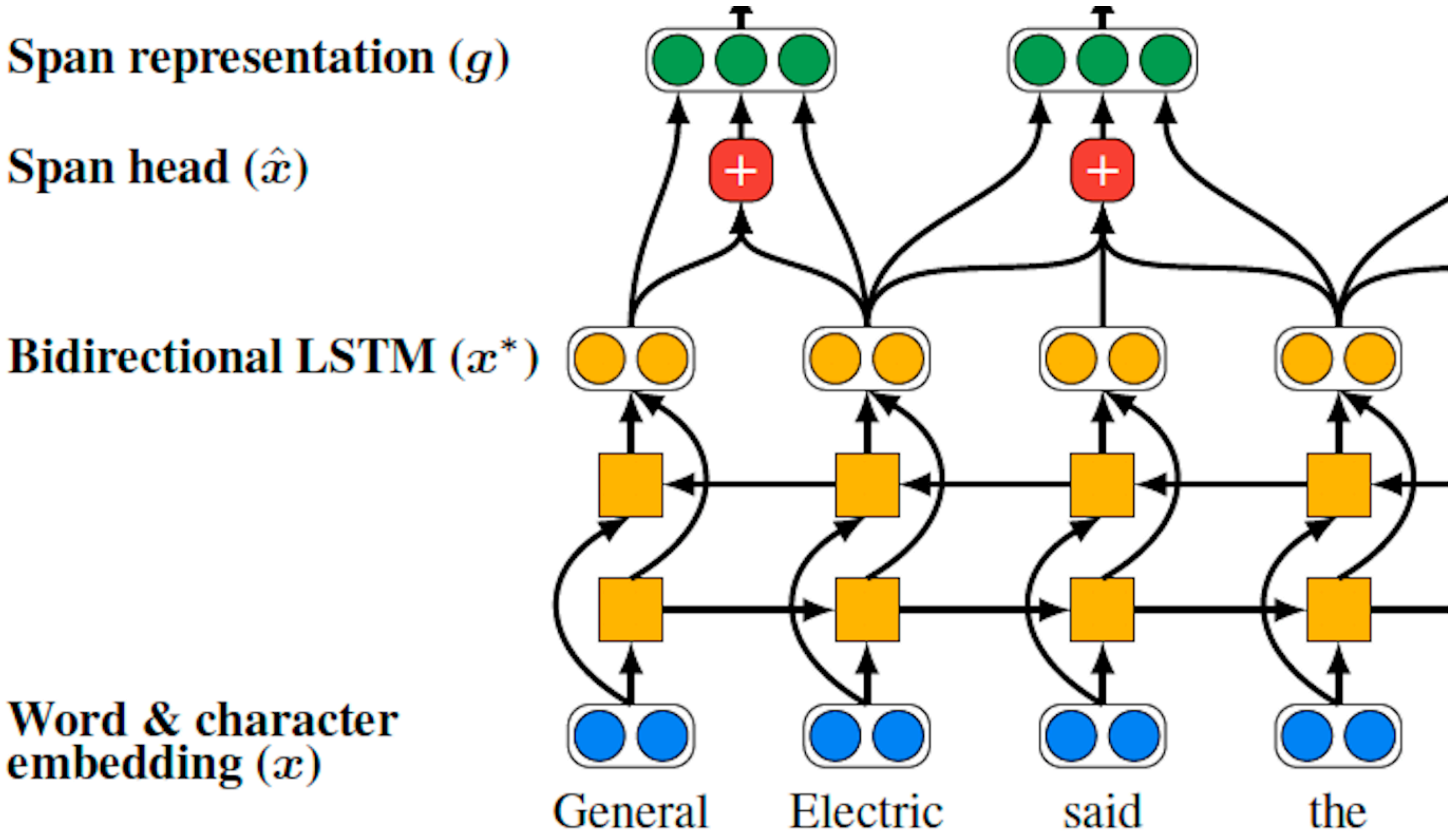
Model: End to End

- Input: Word embedding along with metadata such as speaker and genre information.
- Two steps model:
  - First step computes mention score and encodes span embedding
  - Second step computes the final coreference score by summing antecedent scores from pairs of span representations and the mentions score for each span
- Output:
  - Assign to each span i an antecedent y_i.

Model: Step one



Step one: Span Embeddings



Head-finding Attention

For each span i, for each word t:

$$\alpha_t = \boldsymbol{w}_\alpha \cdot \mathrm{FFNN}_\alpha(\boldsymbol{x}_t^*)$$

$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\mathrm{START}(i)}^{\mathrm{END}(i)} \exp(\alpha_k)}$$

$$\hat{\boldsymbol{x}}_i = \sum_{t=\mathrm{START}(i)}^{\mathrm{END}(i)} a_{i,t} \cdot \boldsymbol{x}_t$$

Span Representation

$$\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$$

$\emptyset(i)$ just encodes the size of span i.

Pruning

Time complexity: complete model requires $O(T^4)$ in the document length T.

Aggressive Pruning:

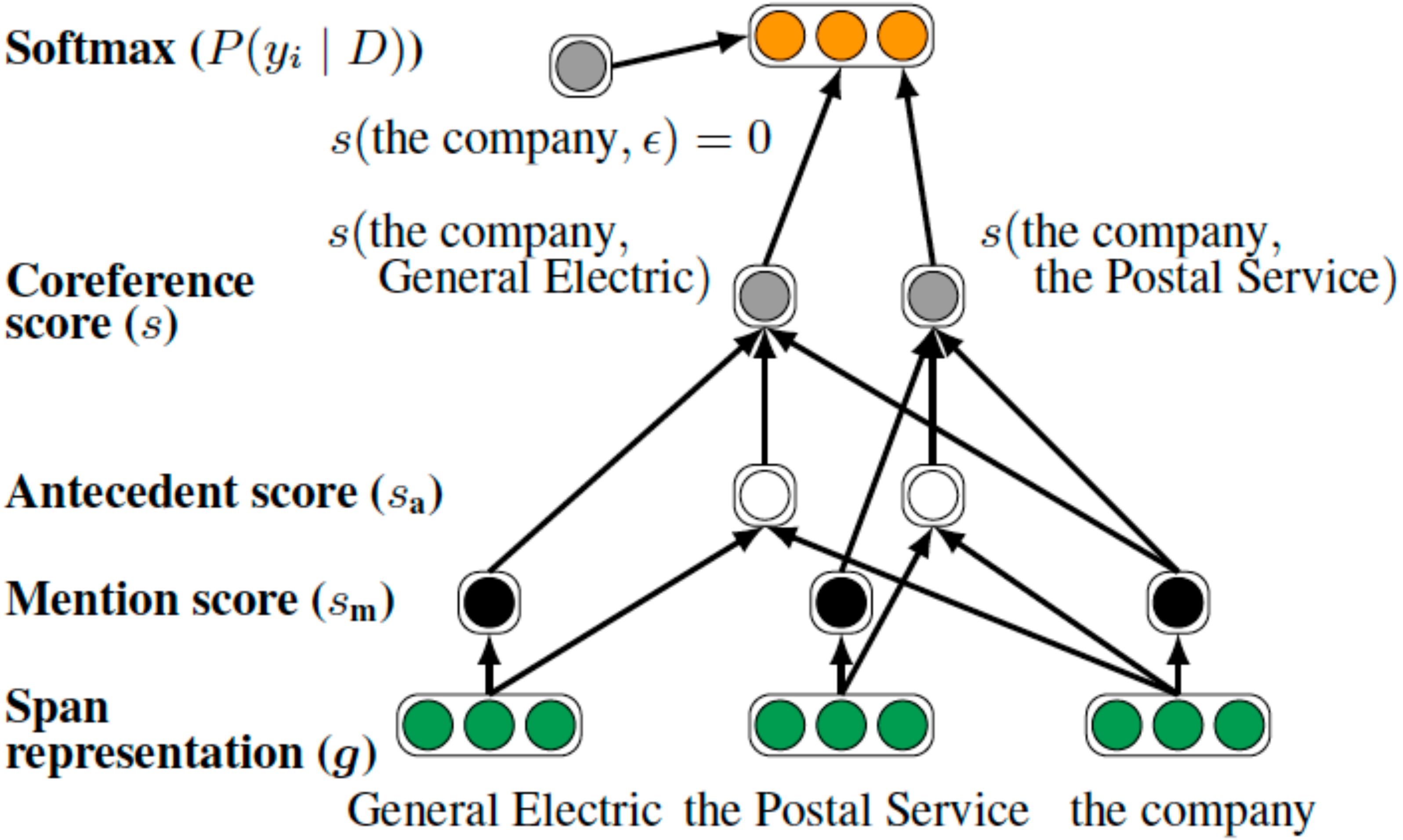
- only consider spans with up to L words
- only keep up to λT spans with the highest mention scores
- only consider up to K antecedents for each.

Mention Score and Antecedent score

Unary mention scores and pairwise antecedent scores

$$s_{\mathrm{m}}(i) = \boldsymbol{w}_{\mathrm{m}} \cdot \mathrm{FFNN}_{\mathrm{m}}(\boldsymbol{g}_i)$$
$$s_{\mathrm{a}}(i,j) = \boldsymbol{w}_{\mathrm{a}} \cdot \mathrm{FFNN}_{\mathrm{a}}([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \circ \boldsymbol{g}_j, \phi(i,j)])$$

Model: Step two



Learning:

Conditional probability distribution

$$
\begin{aligned}
P(y_1, \ldots, y_N \mid D) &= \prod_{i=1}^{N} P(y_i \mid D) \\
&= \prod_{i=1}^{N} \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))}
\end{aligned}
$$

$$
s(i, j) = \begin{cases} 0 & j = \epsilon \\ s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{a}}(i, j) & j \neq \epsilon \end{cases}
$$

Learning: Optimization

Marginal log-likelihood of all correct antecedents implied by the gold clustering:

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

Experiment

- **Dataset:** English coreference resolution data from the CoNLL-2012 shared task
- **Word representations:** 300-dimensional GloVe embeddings and 50-dimensional embeddings from Turian
- **Feature encoding:**
  - encode speaker information as a binary feature
  - the distance feature are binned into the following buckets [1, 2, 3, 4, 5-7 , 8-15, 16-31, 32-63, 64+]

Result: Performance

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Our model (ensemble)	**81.2**	**73.6**	**77.2**	**72.3**	**61.7**	**66.6**	**65.2**	**60.2**	**62.6**	**68.8**
Our model (single)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

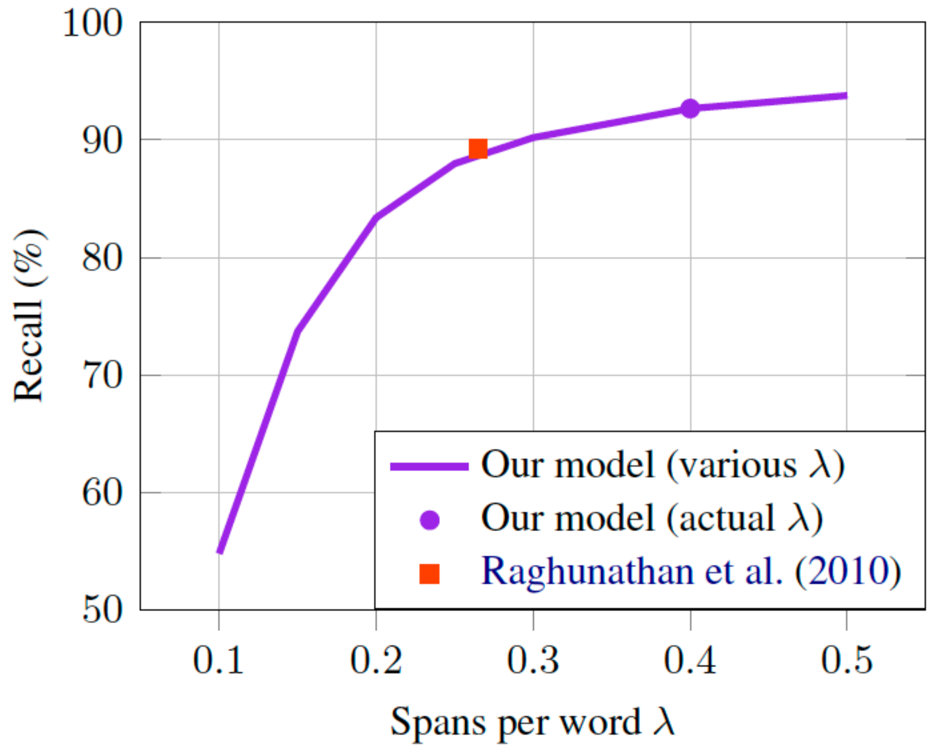
Ablations

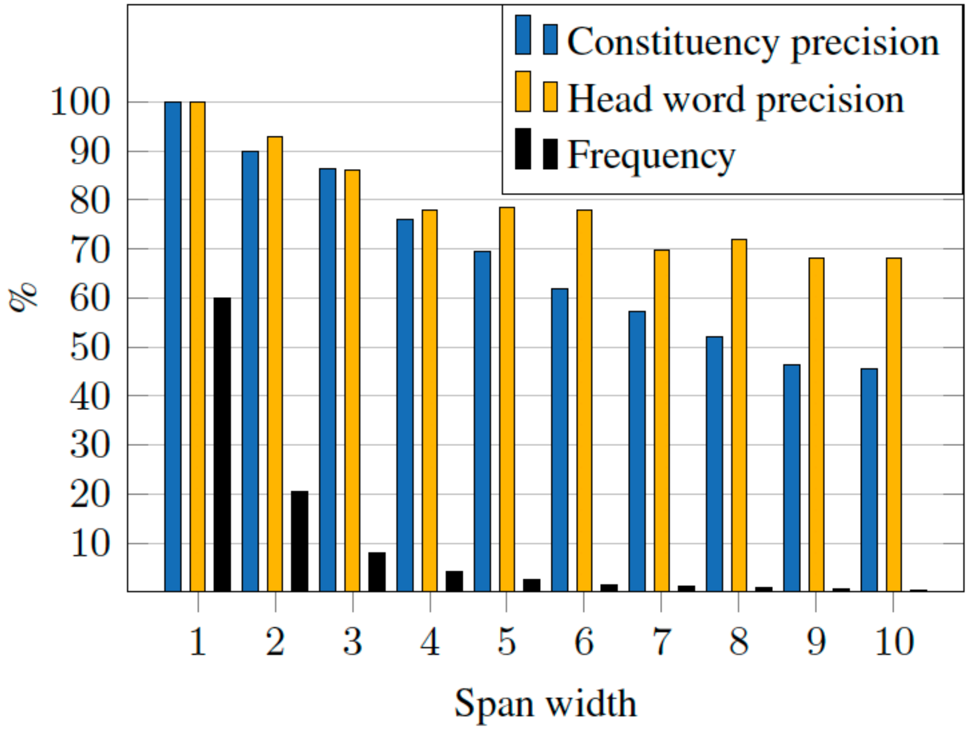
How the ablation of different parts of this model will affect the performance?

	Avg. F1	Δ
Our model (ensemble)	69.0	+1.3
Our model (single)	67.7	
– distance and width features	63.9	-3.8
– GloVe embeddings	65.3	-2.4
– speaker and genre metadata	66.3	-1.4
– head-finding attention	66.4	-1.3
– character CNN	66.8	-0.9
– Turian embeddings	66.9	-0.8

	Avg. F1	Δ
Our model (joint mention scoring)	67.7	
w/ rule-based mentions	66.7	-1.0
w/ oracle mentions	85.2	+17.5

Span Pruning Strategies





Strength and Weakness

Strength

- Novel head-finding attention mechanism detects relatively long and complex noun phrases
- Word embeddings to capture similarity between words

Weakness

- Prone to predicting false positive links when the model conflates paraphrasing with relatedness or similarity
- Does not incorporate world knowledge

Strength and Weakness: Example

(**Prince Charles and his new wife Camilla**) have jumped across the pond and are touring the United States making (**their**) first stop today in New York. It's Charles' first opportunity to showcase his new
4 wife, but few Americans seem to care. Here's Jeanie Mowth. What a difference two decades make. (**Charles and Diana**) visited a JC Penney's on the prince's last official US tour. Twenty years later here's the prince with his new wife.

Summary

- New model: State-of-the-art coreference resolution model
- New mechanism: A novel head-finding attention mechanism
- New insight: Proves that syntactic parser or hand-engineered mention detector isn’t necessary