

Coreference Resolution

- Different “mentions” in a sentence / paragraph can refer the same underlying entity.
- Identifying these “coreferences” is useful for information extraction, summarization, question answering, and other downstream tasks.

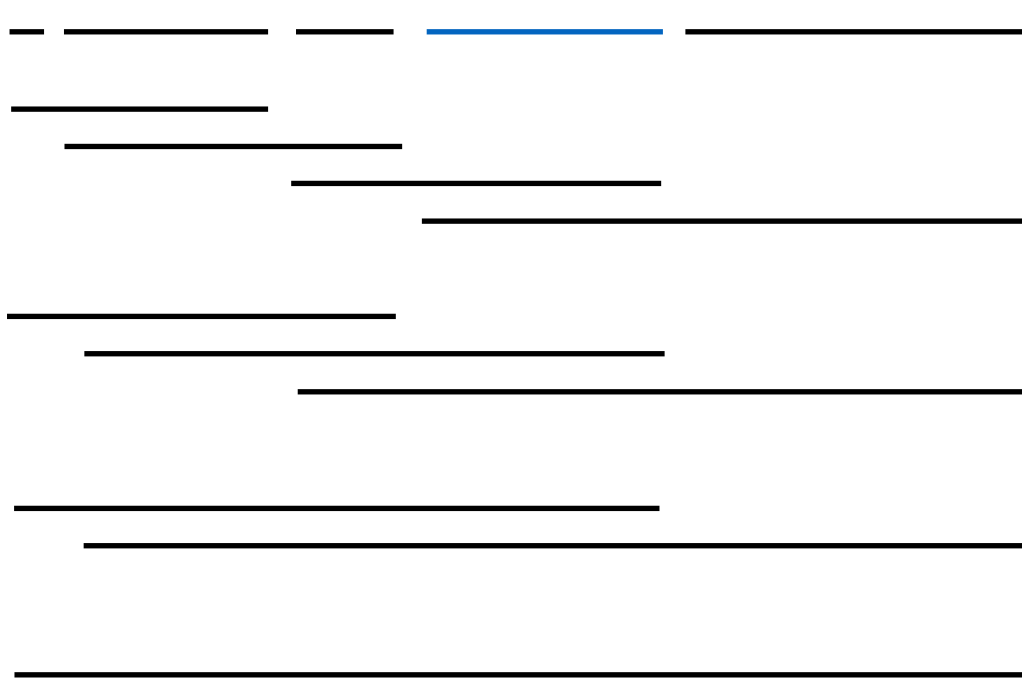
“I voted for Nader because he was most aligned with my values,” she said.

The diagram shows three curved arrows indicating coreferences: one from 'I' to 'she', one from 'Nader' to 'he', and one from 'my' to 'Nader'.

Problem setup



I voted for **Nader** because **he** was most aligned with my values.



“Nader” and “he” refer to the same person.

Nomenclature

- mention

John told **Sally** that **she** should come watch **him** play the **violin**.

- antecedent

John told Sally that **she** should come watch him play the violin.



- coreferent

John told **Sally** that **she** should come watch him play the violin.



- cluster

John told **Sally** that **she** should come watch **him** play the **violin**.


- anaphoric

John told **Sally** that she should come watch him play the violin.



- non-anaphoric

John told Sally that she should come watch him play the **violin**.



John told Sally that she should come watch him play the violin.

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.

We are looking for a region of central Italy bordering the Adriatic Sea. The area is mostly mountainous and includes ~~Mt. Corne~~, the highest peak of the Apennines. It also includes a lot of sheep, good clean-living, healthy sheep, and an Italian entrepreneur has an idea about how to make a little money of them.

Notation

Document D containing T words

$$N = \frac{T(T+1)}{2} \quad \text{possible spans in the entire document}$$

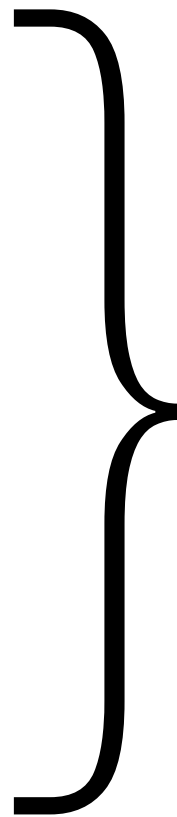
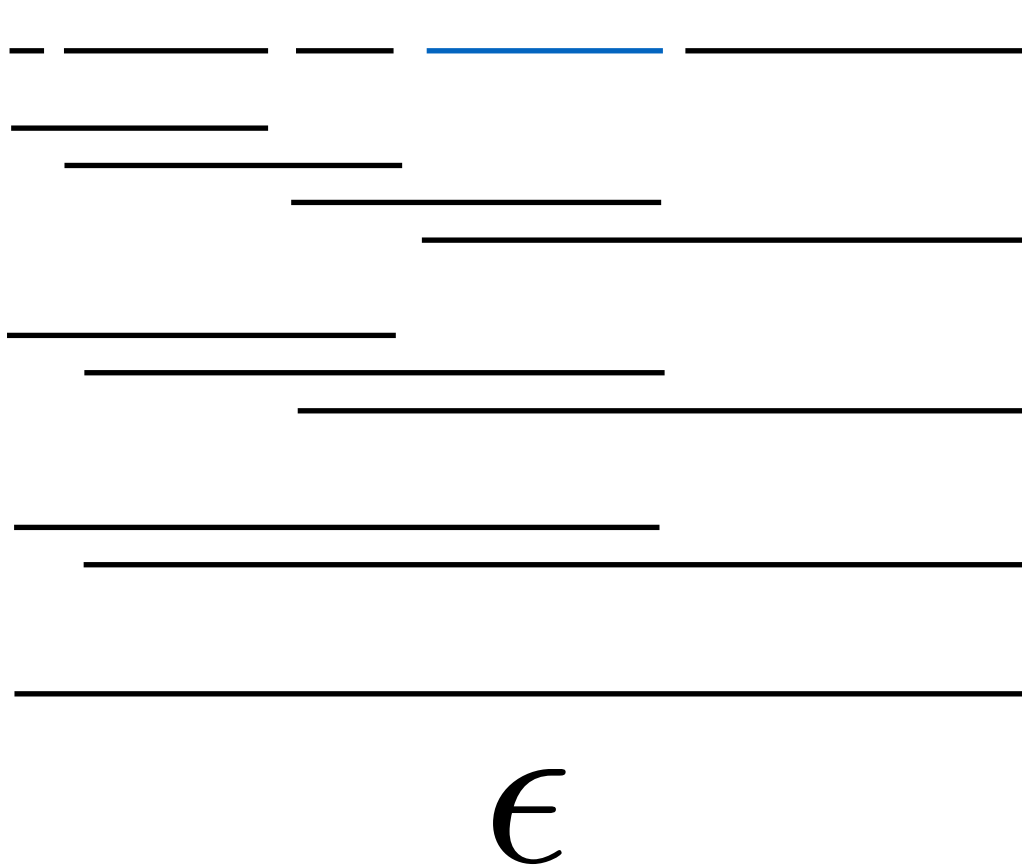
$$y(i) = \{\epsilon, 1, \dots, i-1\} \quad \text{possible antecedents for span at position } i$$

Notation

i



I voted for **Nader** because **he** was most aligned with my values.



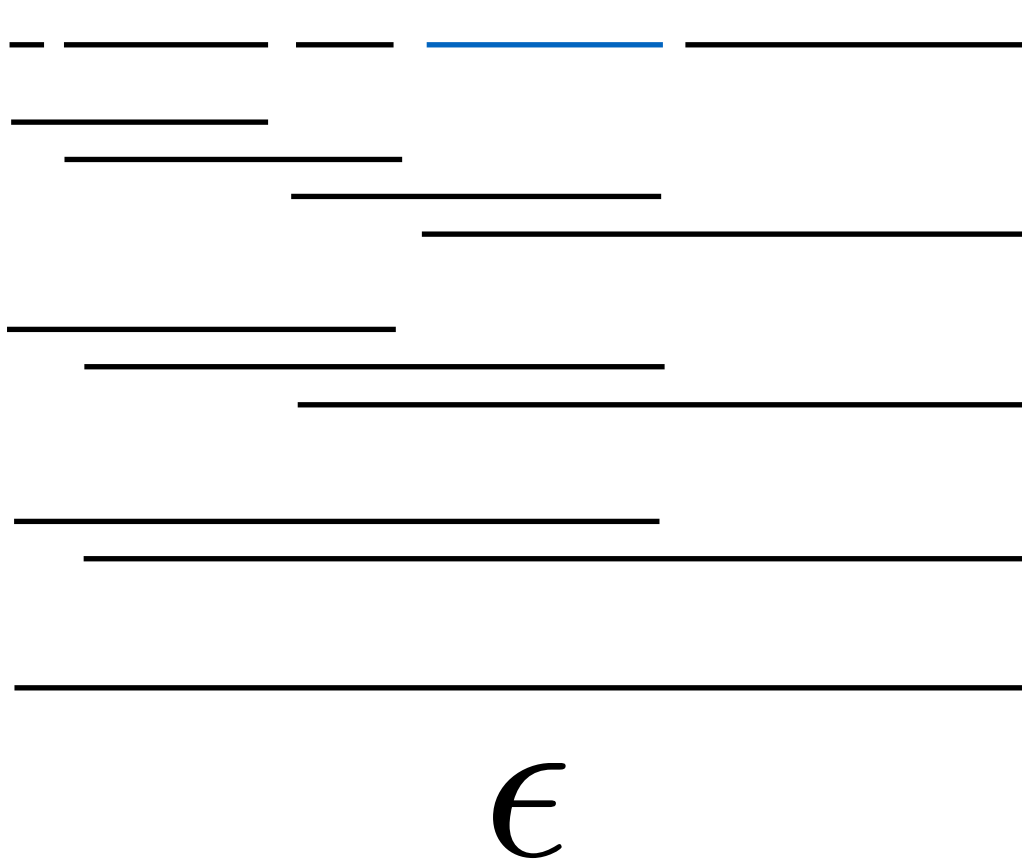
$$y(i) = \{\epsilon, 1, \dots, i - 1\}$$

“Mention ranking” paradigm

i



I voted for **Nader** because **he** was most aligned with my values.



$s(\text{he}, \text{Nader})$

Score each possible antecedent, pick the highest-scoring candidate.

$O(T^4)$ 🤔

How to calculate score function $s(i, j)$?

- Before neural networks:
 - Dependency parse sentences to get candidate matches
 - Embed with hand-engineered features, conjoined in various ways
 - Linear scoring model

$$s_{lin}(x, y) \triangleq \begin{cases} u^\top \begin{bmatrix} \phi_a(x) \\ \phi_p(x, y) \end{bmatrix} & \text{if } y \neq \epsilon \\ v^\top \phi_a(x) & \text{if } y = \epsilon \end{cases}$$

- ϕ_a is a feature function on an individual mention + context
- ϕ_p is a pairwise feature function between a mention and possible antecedent

$$\phi_a, \phi_p$$

Mention Features (ϕ_a)		
Feature		Value Set
Mention Head		\mathcal{V}
Mention First Word		\mathcal{V}
Mention Last Word		\mathcal{V}
Word Preceding Mention		\mathcal{V}
Word Following Mention		\mathcal{V}
# Words in Mention		$\{1, 2, \dots\}$
Mention Synt. Ancestry		see BCS (2013)
Mention Type		\mathcal{T}
+ Mention Governor		\mathcal{V}
+ Mention Sentence Index		$\{1, 2, \dots\}$
+ Mention Entity Type		NER tags
+ Mention Number		$\{\text{sing.}, \text{plur.}, \text{unk}\}$
+ Mention Animacy		$\{\text{an.}, \text{inan.}, \text{unk}\}$
+ Mention Gender		$\{\text{m}, \text{f}, \text{neut.}, \text{unk}\}$
+ Mention Person		$\{1, 2, 3, \text{unk}\}$

Pairwise Features (ϕ_p)		
Feature		Value Set
BASIC features on Mention		see above
BASIC features on Antecedent		see above
Mentions between Ment., Ante.		$\{0 \dots 10\}$
Sentences between Ment., Ante.		$\{0 \dots 10\}$
i-within-i		$\{\text{T}, \text{F}\}$
Same Speaker		$\{\text{T}, \text{F}\}$
Document Type		$\{\text{Conv.}, \text{Art.}\}$
Ante., Ment. String Match		$\{\text{T}, \text{F}\}$
Ante. contains Ment.		$\{\text{T}, \text{F}\}$
Ment. contains Ante.		$\{\text{T}, \text{F}\}$
Ante. contains Ment. Head		$\{\text{T}, \text{F}\}$
Mention contains Ante. Head		$\{\text{T}, \text{F}\}$
Ante., Ment. Head Match		$\{\text{T}, \text{F}\}$
Ante., Ment. Synt. Ancestries		see above
+ BASIC+ features on Ment.		see above
+ BASIC+ features on Ante.		see above
+ Ante., Ment. Numbers		see above
+ Ante., Ment. Genders		see above
+ Ante., Ment. Persons		see above
+ Ante., Ment., Entity Types		see above
+ Ante., Ment. Heads		see above
+ Ante., Ment. Types		see above

Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution, Wiseman et al. 2015

Use a FFNN to learn raw feature interactions:

$$s_{lin}(x, y) \triangleq \begin{cases} u^\top \begin{bmatrix} \phi_a(x) \\ \phi_p(x, y) \end{bmatrix} & \text{if } y \neq \epsilon \\ v^\top \phi_a(x) & \text{if } y = \epsilon \end{cases} \quad \rightarrow \quad s(x, y) \triangleq \begin{cases} u^\top g\left(\begin{bmatrix} \mathbf{h}_a(x) \\ \mathbf{h}_p(x, y) \end{bmatrix}\right) + u_0 & \text{if } y \neq \epsilon \\ v^\top \mathbf{h}_a(x) + v_0 & \text{if } y = \epsilon \end{cases}$$

$$\mathbf{h}_a(x) \triangleq \tanh(W_a \phi_a(x) + b_a)$$

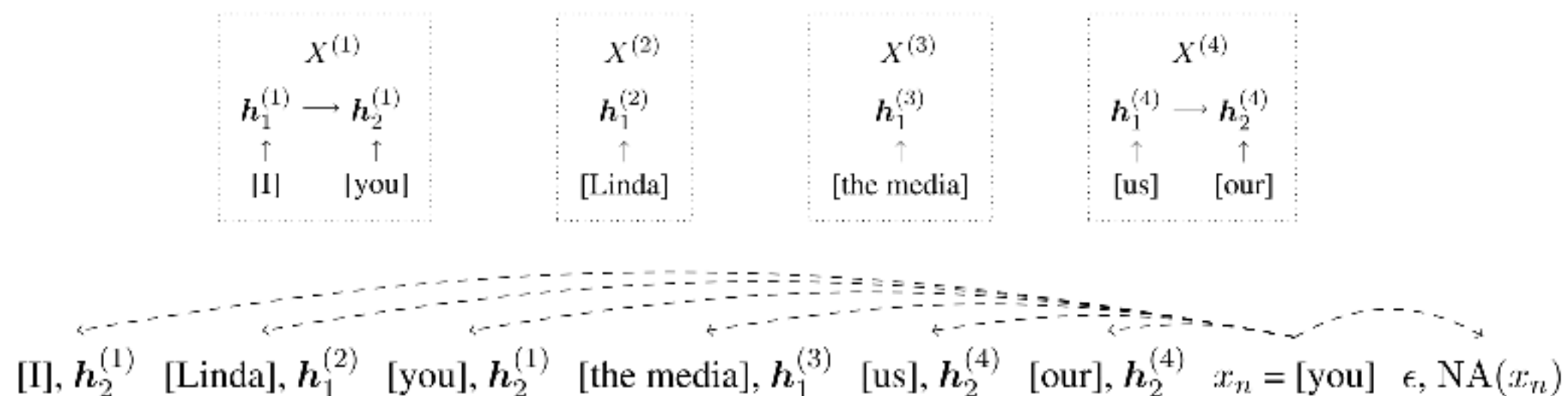
$$\mathbf{h}_p(x, y) \triangleq \tanh(W_p \phi_p(x, y) + b_p)$$

Learning Global Features for Coreference Resolution, Wiseman et al. 2016

- Train LSTMs on mention clusters
- Given a new document — greedily assign mentions to clusters, given current LSTM states and local mention features
- Still using dependency parsers for mention detection + hand-engineered raw features.

DA: um and [I]₁ think that is what's - Go ahead [Linda]₂.

LW: Well and thanks goes to [you]₁ and to [the media]₃ to help [us]₄...So [our]₄ hat is off to all of [you]₅...



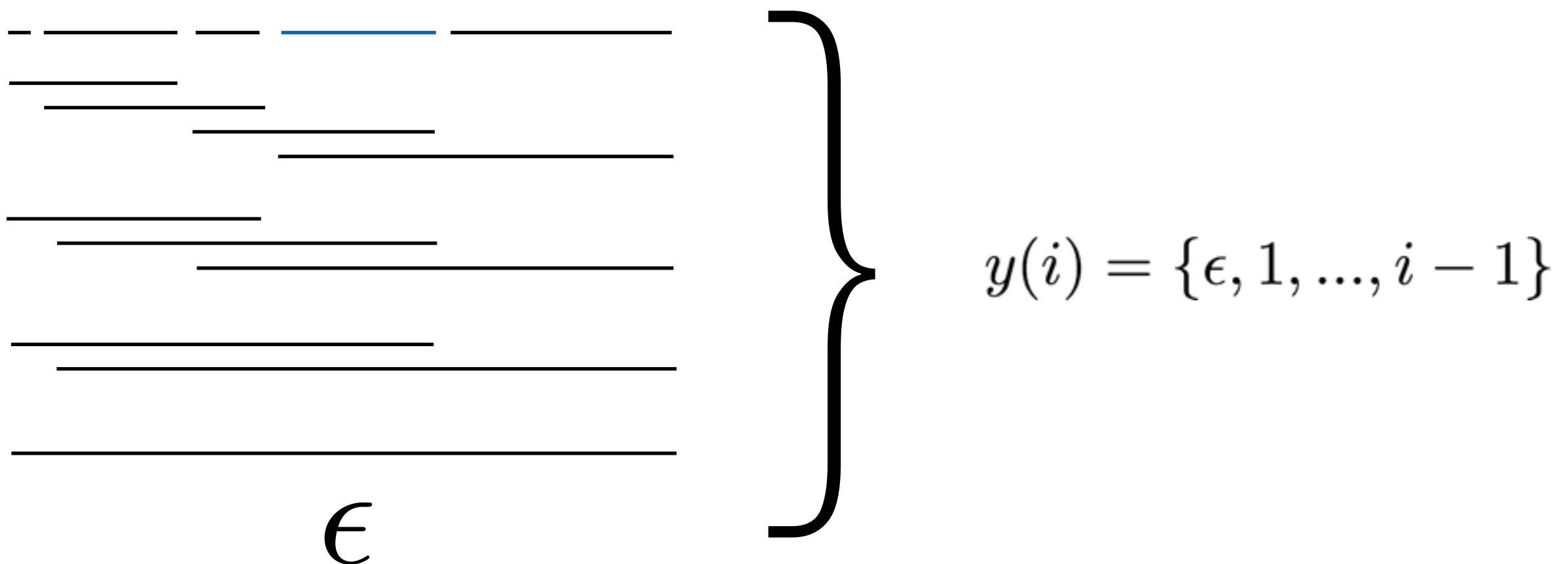
End-to-end Neural Coreference Resolution, Lee et al. 2017

- No dependency parsing — model just learns to predict distribution over set of possible antecedents $y(i)$.

y_1 y_2 y_3 y_4 y_5 y_6


↓ ↓ ↓ ↓ ↓ ↓

I voted for **Nader** because **he** was most aligned with my values.



Goal: Learn distribution whose most likely configuration produces the correct clustering

$$P(y_1, \dots, y_N | D) = \prod_{i=1}^N P(y_i | D) = \prod_{i=1}^N \frac{\exp(s(i, y_i))}{\sum_{y' \in \mathcal{Y}(i)} \exp(s(i, y'))}$$



For each span, distribution over all possible antecedents, including ε

3-part score function

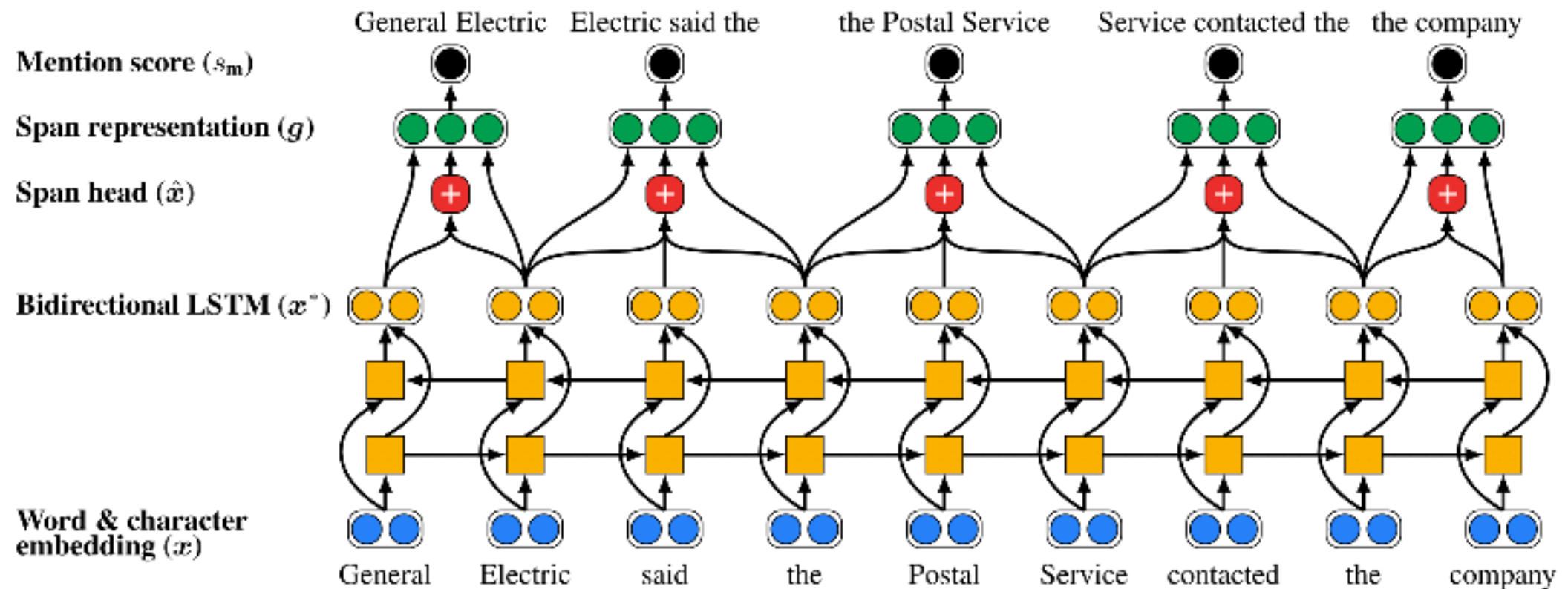
$$s(i, j) = \begin{cases} 0 & j = \epsilon \\ s_m(i) + s_m(j) + s_a(i, j) & j \neq \epsilon \end{cases}$$

Is i a mention? $s_m(i) = w_m \cdot FFNN_m(g_i)$

Is j the antecedent of i ? $s_a(i, j) = w_a \cdot FFNN_a([g_i, g_j, g_i \circ g_j, \phi(i, j)])$

$$g_i = \text{encoding of span } y_i$$

Span encoding



$$g_i = [x_{START(i)}^*, x_{END(i)}^*, \hat{x}_i, \phi(i)]$$

(Left LSTM state, right LSTM state, attention, size)

$$\alpha_t = \mathbf{w}_\alpha \cdot FFNN_\alpha(\mathbf{x}_t^*)$$

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

$$\hat{x}_i = \sum_{t=START(i)}^{END(i)} \alpha_{i,t} \cdot \mathbf{x}_t$$

Objective

Maximize the log-likelihood of all correct antecedents:

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

Dataset

The authors used the CoNLL 2012 shared task dataset.

Train: 2802 documents

Dev: 343 documents

Test: 348 documents

Avg Length: 454 words

Maximum Length: 4009 words

In each document, coreference links are indicated by nested parentheses within sentences.

Pruning / Tractability

- Only spans of up to **L** words have their mention scores computed.
- Of these, only the top **λT** highest mention score spans are kept.
- For these, we consider only up to **K** antecedents for each.
- Spans are accepted in decreasing order of the mention scores, unless there exists a previously accepted span that **crosses** with the current span.

The authors used $L = 10$, $\lambda = 0.4$, and $K = 250$. Despite these pruning methods, their recall for gold mentions was over 92%.

Metrics

MUC: Link-based metric for the minimum number of links between mentions that need to be inserted or deleted when evaluating a response. Cannot represent “singleton” entities.

B³: Mention-based metric that is the fraction of the correct mentions that are included in the response of an entity. Needs extensions to handle “twinless” entities.

CEAF_{φ4}: Key-based metric which uses the *CEAF* algorithm to align entities in the key and response by computing similarity between pairs to find the best total similarity. Not very intuitive as mistakes are not counted, and the scores can be skewed by output size.

Hyperparameters

Word Representations: 300-dimensional GloVe embeddings, 50-dimensional Turian embeddings.

OOV: 8-dimensional char embeddings. CNN filter sizes of 3, 4, 5 with 50 filters each.

Hidden Dimensions: LSTMs have hidden state sizes of 200, FFNNs consist of two hidden layers with 150 dimensions and ReLUs.

Learning: Adam optimizer, learning rate $1e-3$ with a scheduled decay rate of 0.1% per 100 steps. Trained for 150 “epochs” with early stopping.

Regularization: Mini-batch size 1, dropout $p = 0.50$ to word embeddings and char-CNN, $p = 0.20$ to hidden layers.

Other Features

An additional 20-dimensional vector (dropout $p = 0.20$) for:

Speaker: binary feature indicating if the mentions in a span have the same speaker

Genre: broadcast news, newswire, web data, etc.

Span distance: bucketed distances ([1, 2, 3, 4, 5- 7, 8-15, 16-31, 32-63, 64+]) that tell how far two spans of text are from one another

Mention width: how long (in words) a given mention span is

Results

	MUC			B ³			CEAF _{ϕ_4}			Avg. F1
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	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

	Avg. F1	Δ		
Our model (ensemble)	69.0	+1.3		
Our model (single)	67.7		Avg. F1	
– distance and width features	63.9	-3.8	Our model (ensemble)	68.8
– GloVe embeddings	65.3	-2.4	Our model (single)	67.2
– speaker and genre metadata	66.3	-1.4	Clark and Manning (2016a)	65.7
– head-finding attention	66.4	-1.3	Clark and Manning (2016b)	65.3
– character CNN	66.8	-0.9	Wiseman et al. (2016)	64.2
– Turian embeddings	66.9	-0.8	Wiseman et al. (2015)	63.4

Ablations

(A **fire** in a **Bangladeshi garment factory**) has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee (**the blaze**) in the four-story building.

A fire in (**a Bangladeshi garment factory**) has left at least 37 people dead and 100 hospitalized. Most of the deceased were killed in the crush as workers tried to flee the blaze in (**the four-story building**).

We are looking for (**a region of central Italy bordering the Adriatic Sea**). (**The area**) is mostly mountainous and includes Mt. Corno, the highest peak of the Apennines. (**It**) also includes a lot of sheep, good clean-living, healthy sheep, and an Italian entrepreneur has an idea about how to make a little money of them.

(**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 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.

Possible Improvements

Idea 1: Incorporate non-static word embeddings so that the model makes fewer errors when predicting head words that may be semantically similar.

(The flight attendants) have until 6:00 today to ratify labor concessions. (The pilots') union and ground crew did so yesterday.

(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 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.

Possible Improvements

Idea 2: Enable the attention mechanism to output more than one possible syntactic head to pick up on plurality, grammaticality in small datasets.

Also such location devices, (some ships) have smoke floats (they) can toss out so the man overboard will be able to use smoke signals as a way of trying to, let the rescuer locate (them).

Possible Improvements

Idea 3: Propose alternative ways to reduce the computational complexity from $O(T^4)$.