COS 484: Natural Language Processing

Coreference Resolution

Fall 2019

(Slides adapted from Christopher Manning)

Announcements

- No lecture on **Dec 12th**.
- All the project meetings scheduled on **Dec 10th**
  - Submit checkpoint reports by Dec 9th (optional but recommended!)
  - Sign up project meetings!

Overview

- What is coreference resolution?
- Four types of coreference models
- End-to-end coreference resolution
- Evaluation and results

Discourse

- Text is more than the sum of its individual sentences/utterances.
- Discourse processing: NLP beyond the sentence/utterance boundary.
    - Monologue
    - Dialogue (multi-party)
- Topics:
    - Topic segmentation
    - **Coreference resolution**
    - Coherence
    - Discourse relations
    - …

Coreference resolution

- Identify all mentions that refer to the same real world entity

> Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Coreference resolution

- Identify all mentions that refer to the same real world entity

> Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Coreference resolution

- Identify all mentions that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Coreference resolution

- Identify all mentions that refer to the same real world entity

Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Coreference resolution

- Identify all mentions that refer to the same real world entity

Input

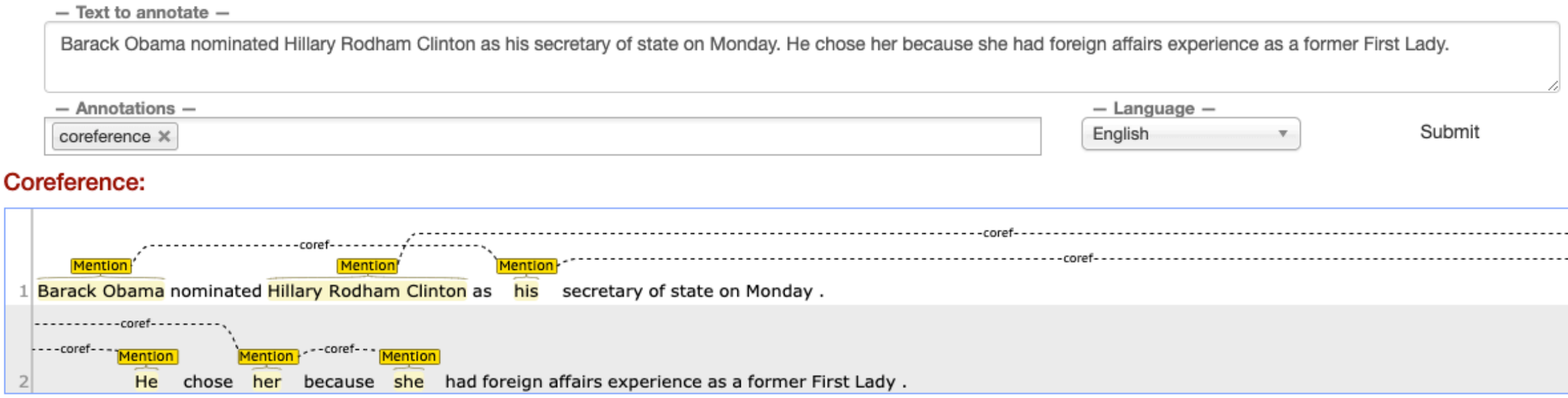
> Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Output

- {Barack Obama, his, He}
- {Hillary Rodham Clinton, secretary of state, her, she, First Lady}

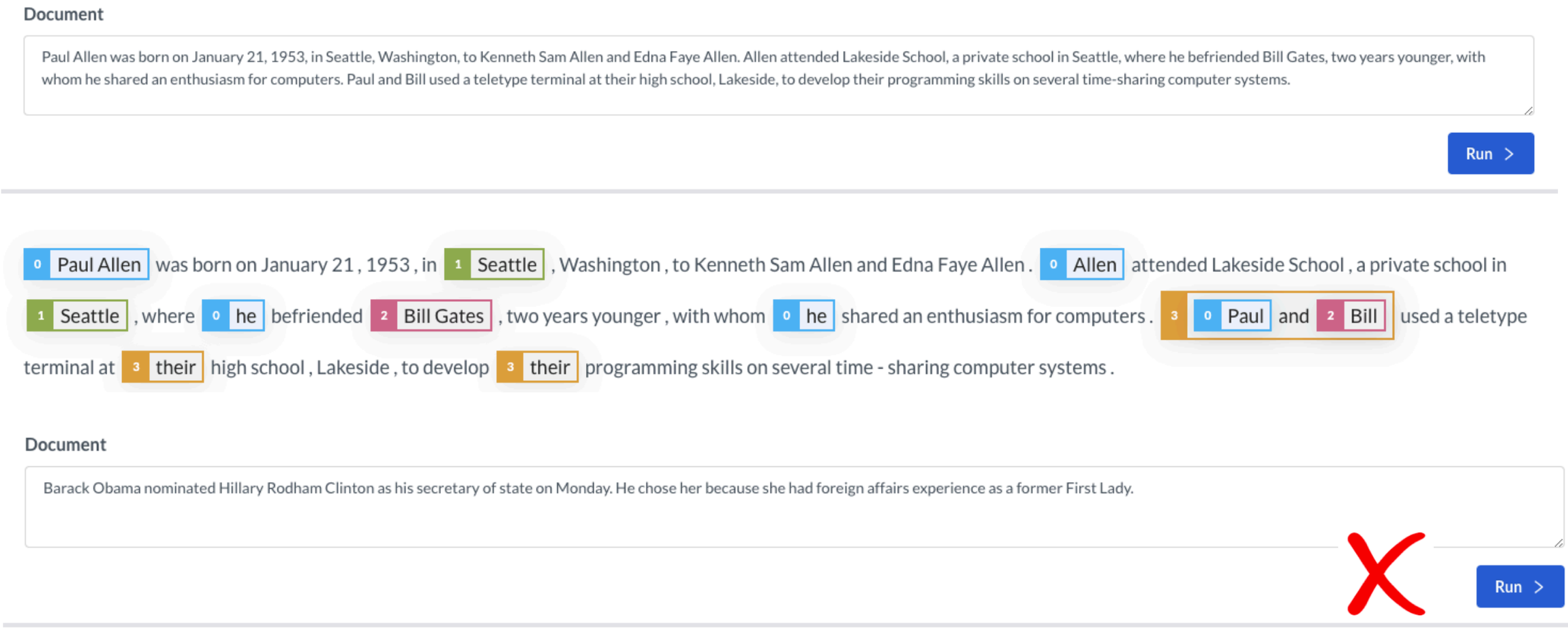
Try coref systems yourself!

- https://corenlp.run



Try coref systems yourself!

- https://demo.allennlp.org/coreference-resolution/



Applications

- Full text understanding
- Information extraction, question answering, summarization..

Barack Obama

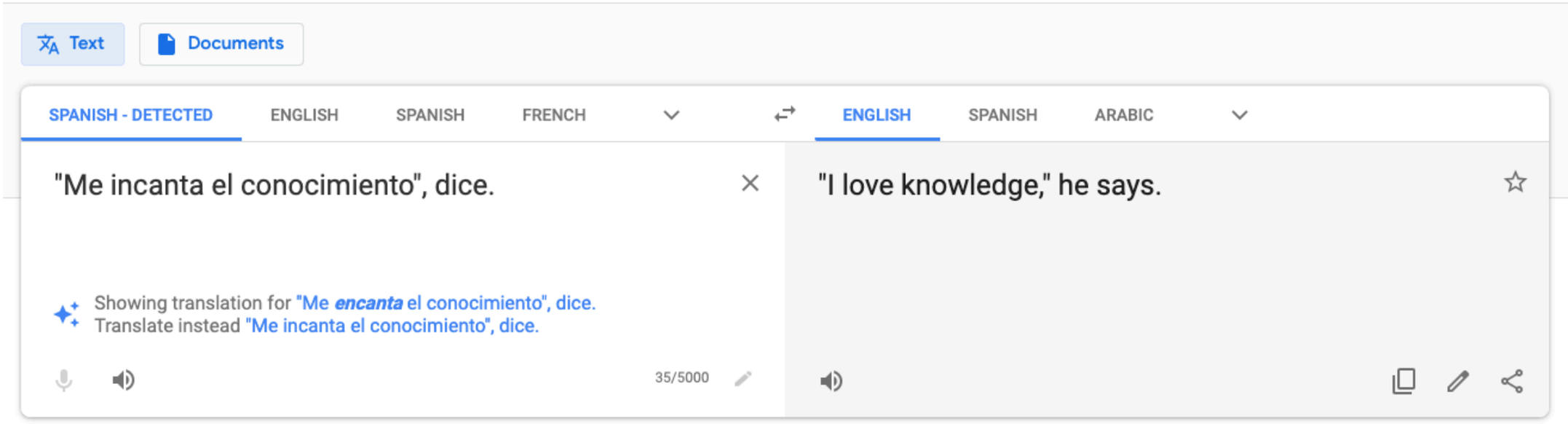
From Wikipedia, the free encyclopedia

Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago. In 1988, **he** enrolled in Harvard Law School, where he was the first black president of the *Harvard Law Review*. After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School from 1992 to 2004.

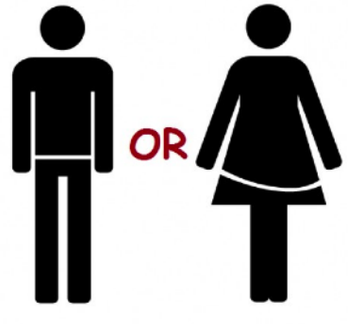
- he = Barack Obama
- (Barack Obama, schools_attended, Harvard University)

Applications

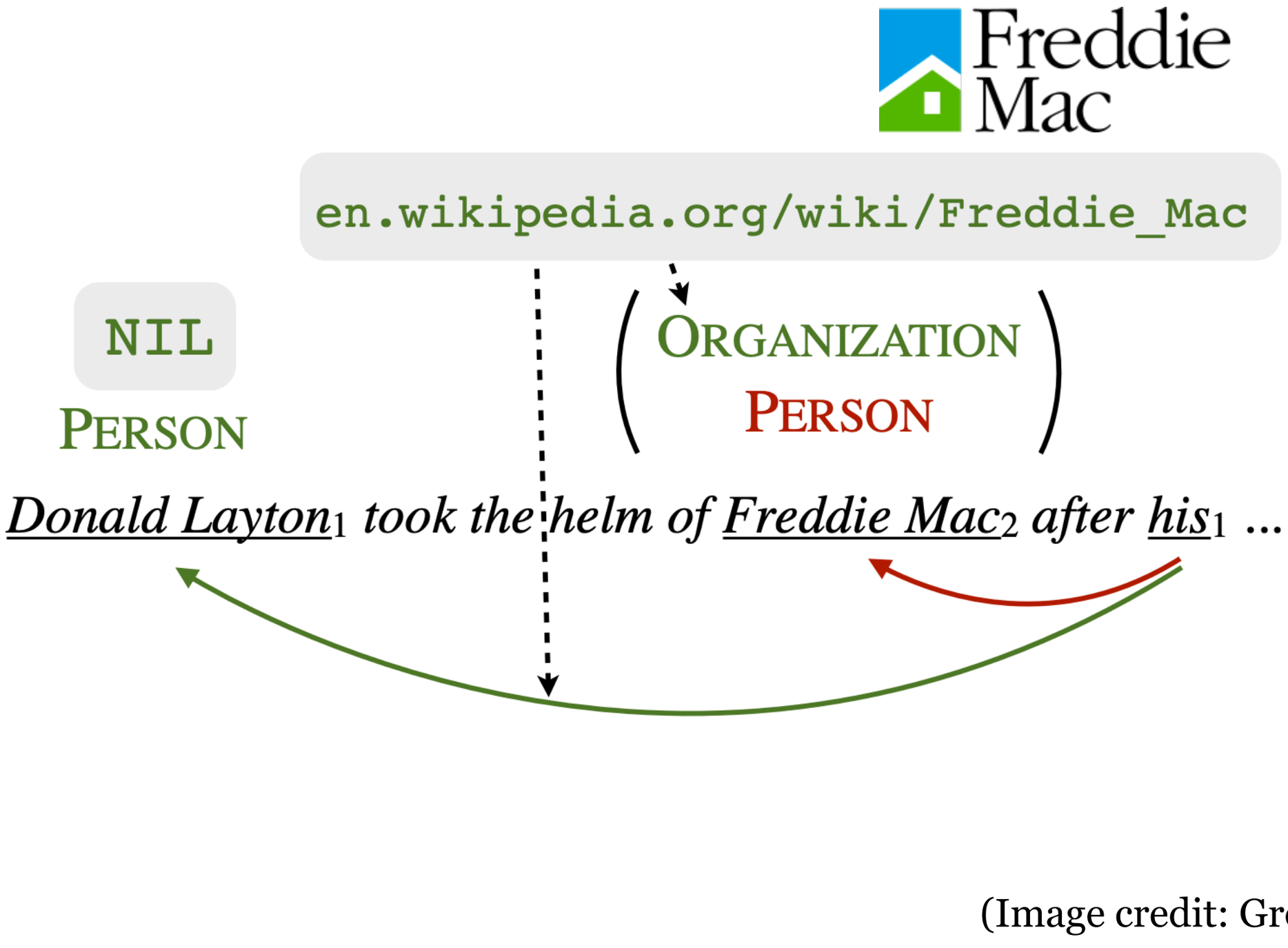
- Languages have different features for gender, number, dropped pronouns, etc.



- “I love knowledge,” he said.
- “I love knowledge,” she said.



Coreference vs Entity Linking



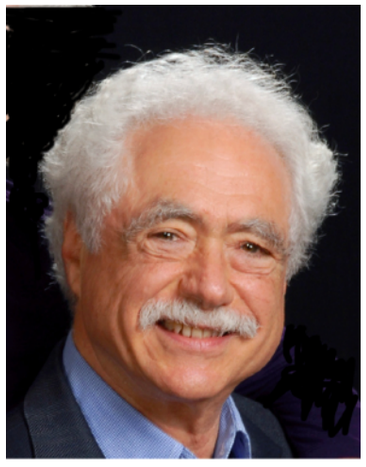
(Image credit: Greg Durrett)

Winograd Schema problems

- Some cases of coreference require world knowledge or common sense reasoning to solve.

They city council denied the demonstrators a permit because they feared violence.

They city council denied the demonstrators a permit because they advocated violence.



Winograd 1972

The trophy didn't fit into the suitcase because it was too large.

The trophy didn't fit into the suitcase because it was too small.

Winograd Schema problems

The Winograd Schema Challenge: recently proposed as an alternative to the Turing test

- (Levesque 2013): "On our best behavior"
- If you've fully solved coreference, arguably you've solved AI

http://commonsensereasoning.org/winograd.html

Coreference: some linguistics

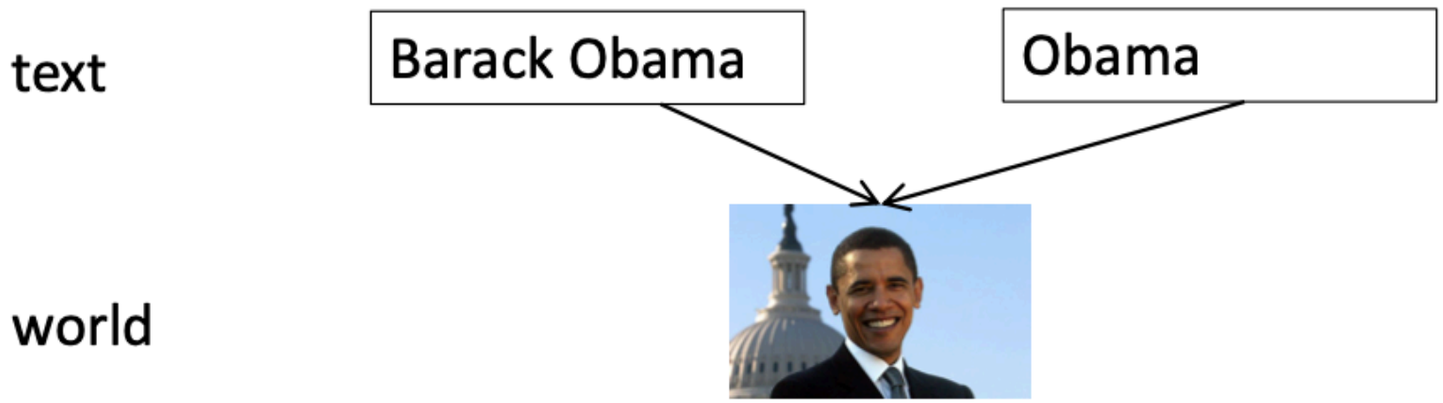
- **Coreference** is when two mentions refer to the same entity in the world
    - **Barack Obama** traveled to.. **Obama**...
- A related linguistic concept is **anaphora**: when a term (anaphor) refers to another term (antecedent)
    - The interpretation of the anaphor is in some way determined by the interpretation of the antecedent
    - **Barack Obama** said **he** would sign the bill.

antecedent

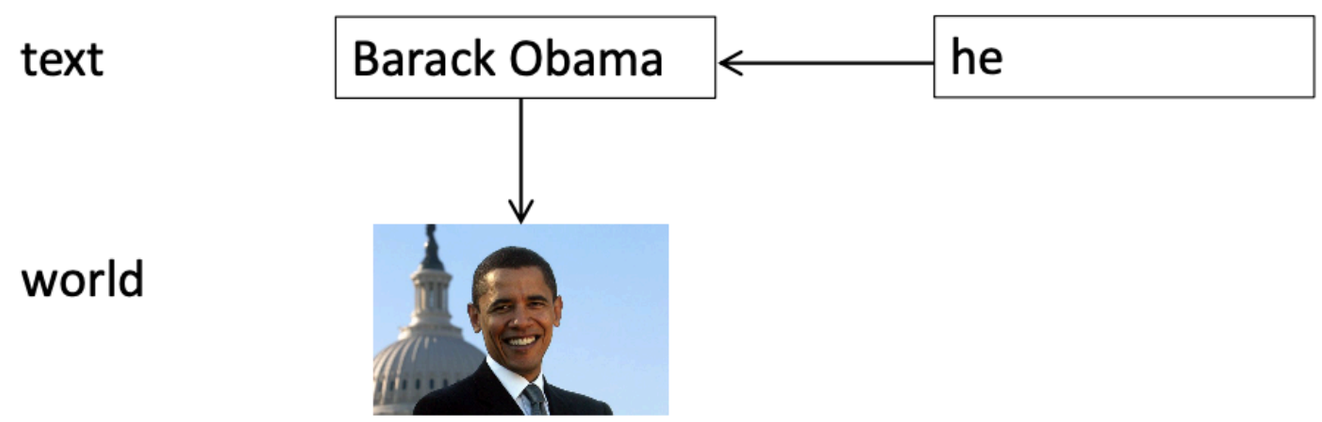
anaphor

Coreference: some linguistics

- Coreference with named entities



- Anaphora

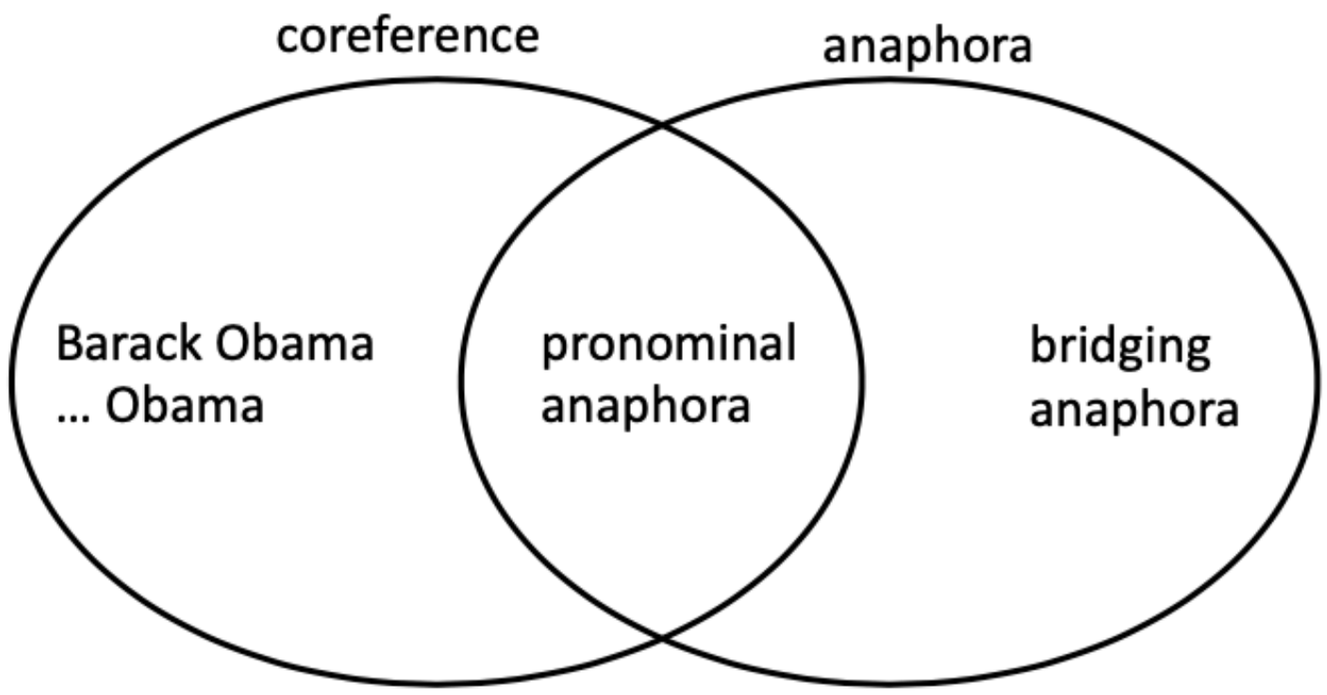


Coreference: some linguistics

- Not all anaphoric relations are coreferential

bridging anaphora

We went to see a concert last night. The tickets were really expensive.



Coreference: some linguistics

Anaphora vs cataphora

- Usually the antecedent comes before the anaphor (e.g., a pronoun) but not always
- Cataphora: mentioned before the their referents

Even before she saw it, Dorothy had been thinking about the Emerald City every day.

More linguistic background in J & M: Chapter 22.1

Coreference resolution in two steps

- **Mention detection (easy)**

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

Mentions can be nested!

- **Mention clustering (hard)**

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

Mention detection

- Mention: span of text referring to some entity
- Three kinds of mentions:
  - #1: Pronouns
    - I, your, it, she, him

Part-of-speech tagger

  - #2: Named entities
    - “Barack Obama” “Princeton University”

named entity recognizer

  - #3: Noun phrases
    - “a dog”, “the big fluffy cat stuck in the tree”

constituency parer

Mention detection: not so simple

- Making all pronouns, named entities, and noun phrases as mentions over-generates mentions

- Are these mentions?
  - It has been ten years
  - Every student
  - No student
  - 100 miles

Mention detection: not so simple

- We can train a classifier to filter out spurious mentions

- Much more common: keep all mentions as "candidate mentions"
  - After your coreference system is done, discard all singleton mentions (ones that have not been marked as coreference with anything else)

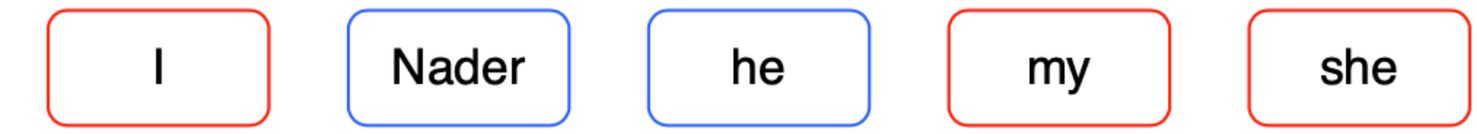
- Recent work: it is possibly to do mention detection and clustering end-to-end instead of in two steps
  - We will cover this later!

(Lee et al, 2017): End-to-end Neural Coreference Resolution

Four kinds of coreference models

- Rule-based (pronominal anaphora resolution)
- Machine learning models:
  - Mention pair
  - Mention ranking ← Current state-of-the-art
  - Clustering-based

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



Hobbs algorithm (1978)

- An algorithm that searches parse trees for antecedents of a pronoun
  - Starting at the NP node immediately dominating the pronoun
  - Search previous trees, in order of recency, left-to-right, breadth-first
  - Looking for the first match of the correct gender and number (male vs female, singular vs plural)

- 72.7% accuracy based on the 100 examples he analyzed

Hobbs algorithm (1978)

The castle in Camelot remained the residence of the king until he moved it to London.

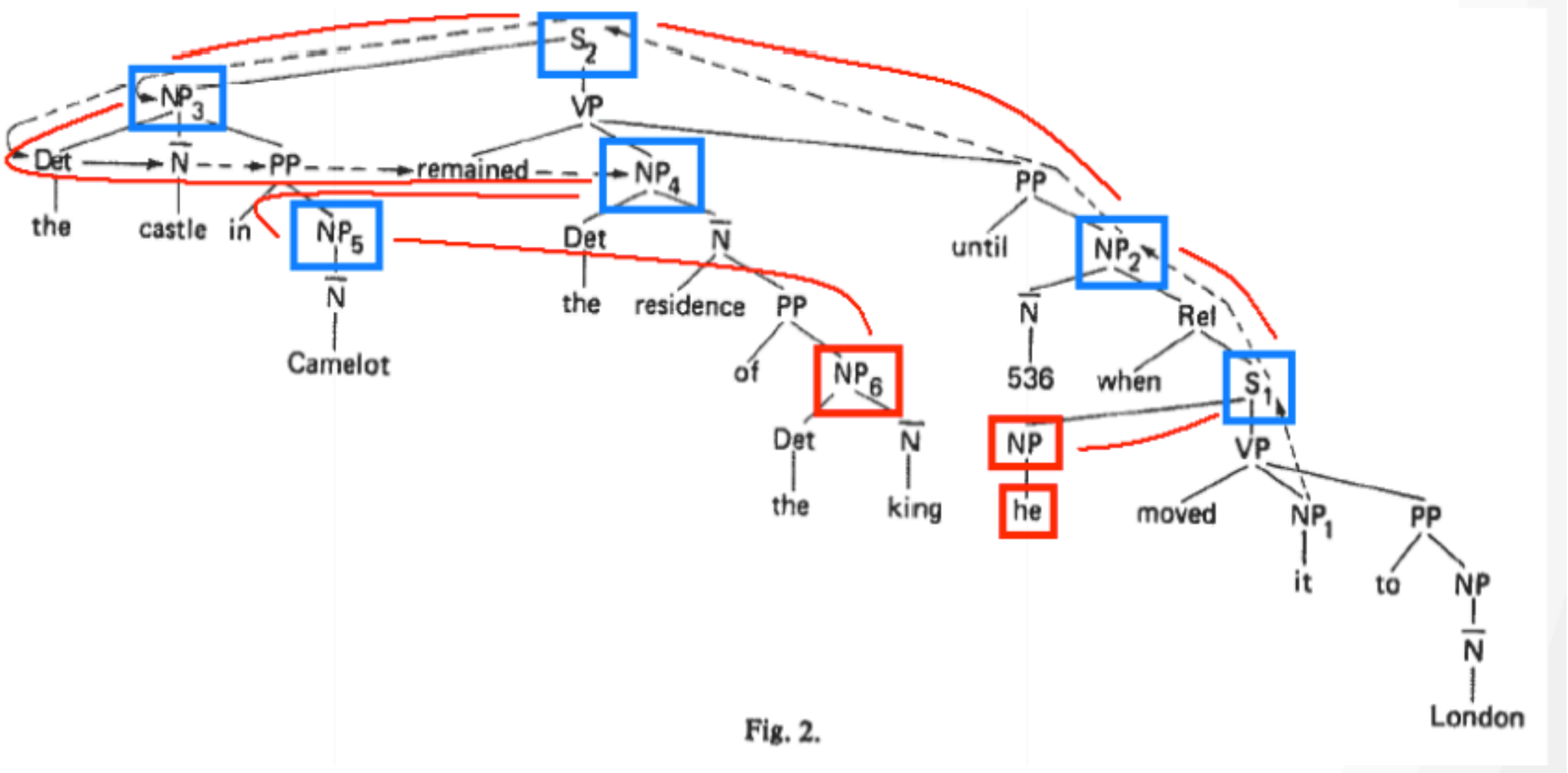
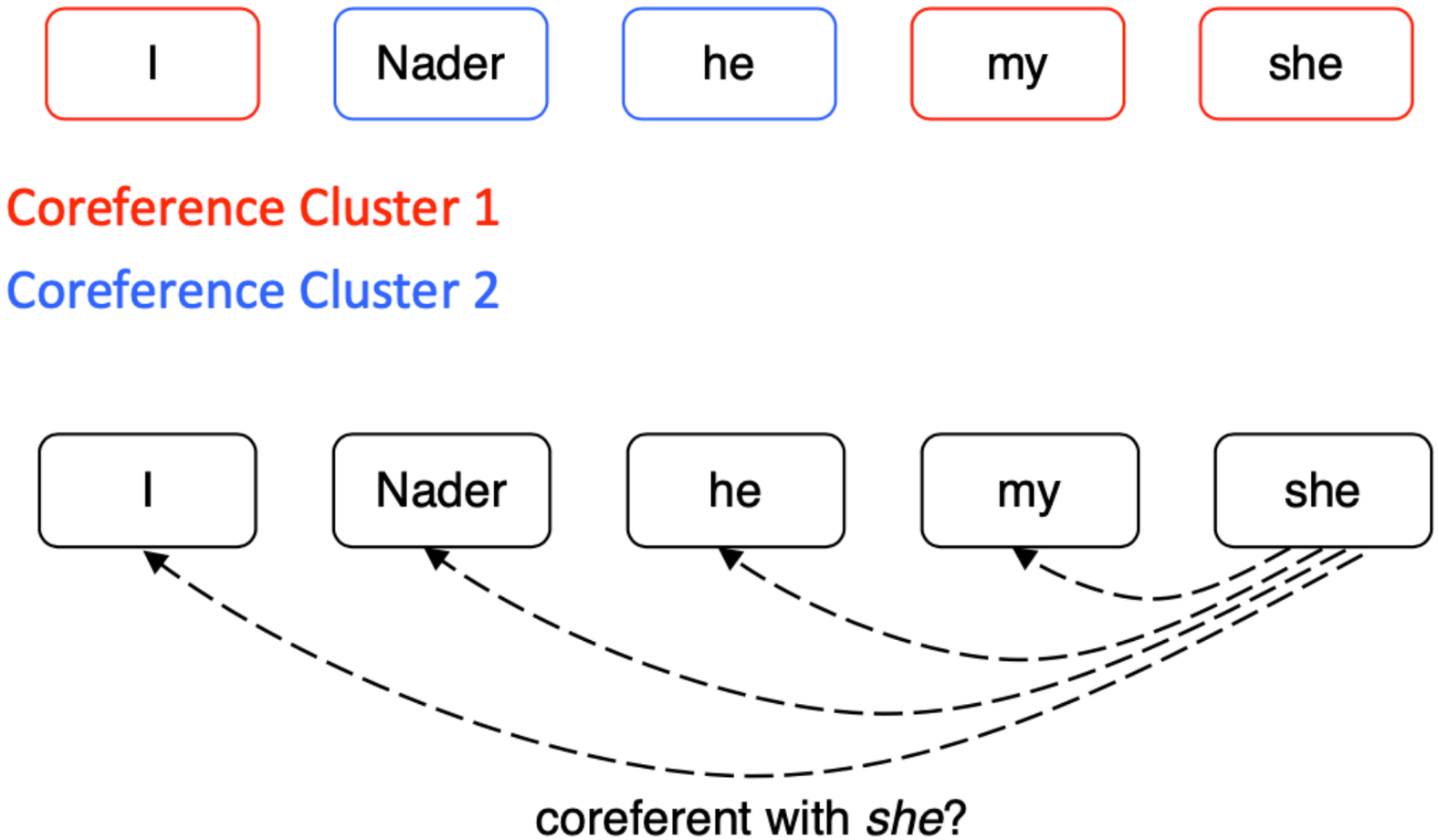


Fig. 2.

Mention-pair models

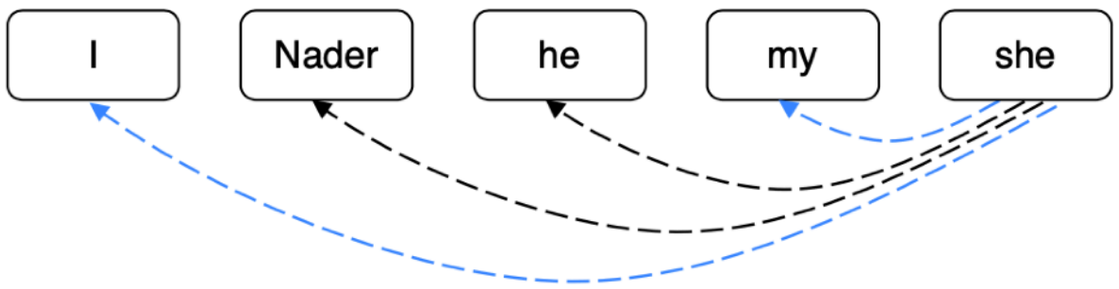
- Train a binary classifier that assigns every pair of mentions a probability of being coreferent: $p(m_i, m_j)$

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



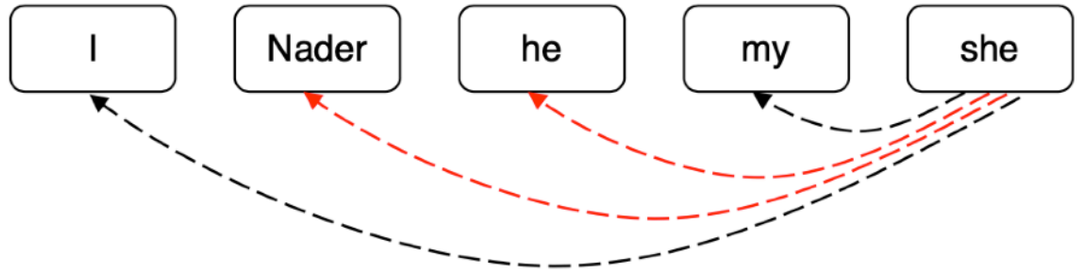
Mention-pair models

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



Positive examples: want $p(m_i, m_j)$ to be near 1

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



Negative examples: want $p(m_i, m_j)$ to be near 0

$$J = -\sum_{i=2}^{N}\sum_{j=1}^{i} y_{ij} \log p(m_j, m_i)$$

Iterate through mentions

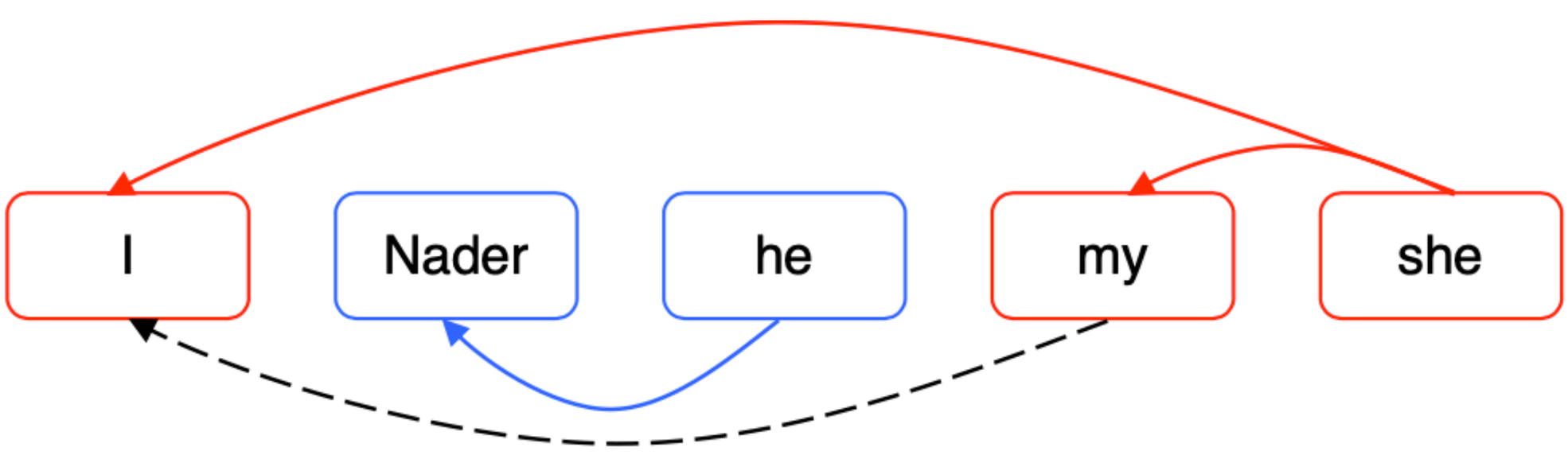
Iterate through candidate antecedents (previously occurring mentions)

Coreferent mentions pairs should get high probability, others should get low probability

Mention-pair models

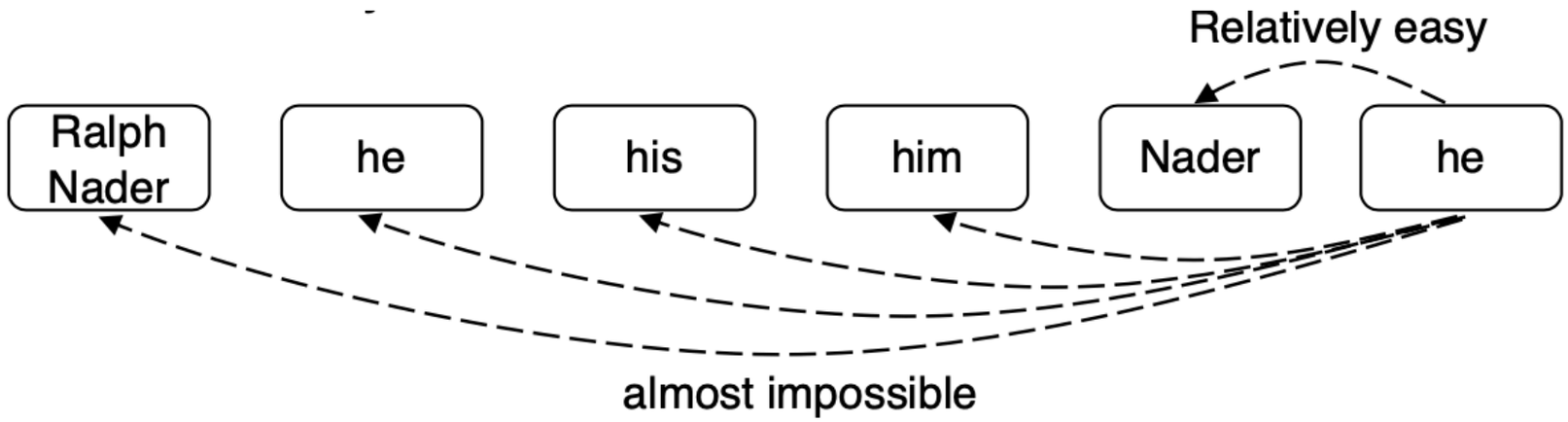
Testing time:

- Run your own mention detector
- Pick some threshold (e.g., 0.5) and add coreference links
- Take the transitive closure to get the clustering



Mention-pair models

Disadvantage: suppose we have a long document with the following mentions: Ralph Nader… he… his.. him.. <several paragraphs>… voted for Nader because he..

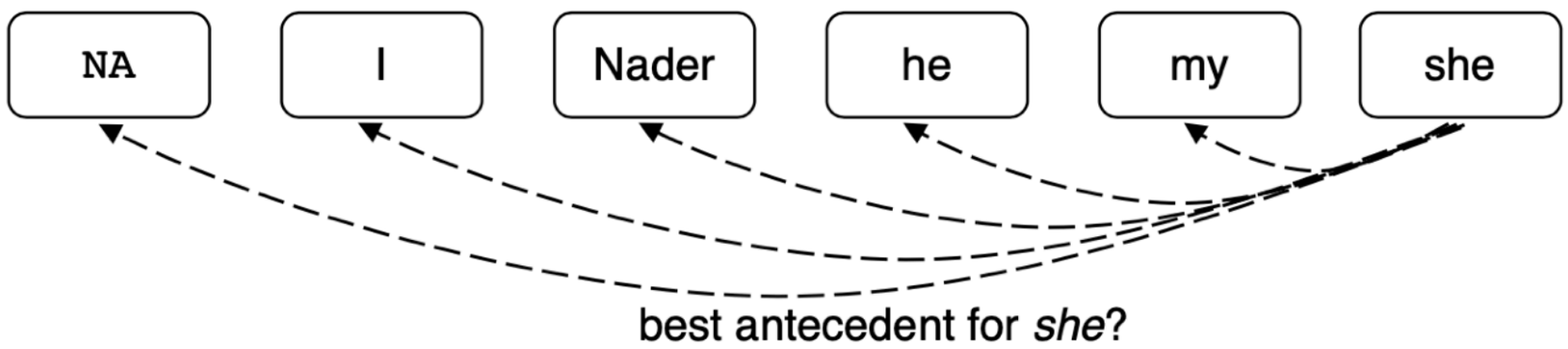


Many mentions only have one clear antecedent but we are asking the model to predict all of them.

Solution: instead train the model to predict only one antecedent for each mention (more linguistically plausible)

Mention-ranking models

- Assign each mention its highest scoring candidate antecedent according to the model
- Add a dummy NA mention to decline linking the current mention to anything (“singleton” or “first” mention)



p(NA, she) = 0.1
p(I, she) = 0.5
p(Nader, she) = 0.1
p(he, she) = 0.1
p(my, she) = 0.2

Apply a softmax over the scores for candidate antecedents so probabilities sum to 1

Mention-ranking models

- Training time: only clustering information is observed (no annotation of "antecedent"), so we optimize the marginal log-likelihood of all the correct antecedents.

$$J = \sum_{i=2}^{N} -\log \left(\sum_{j=1}^{i-1} \mathbb{1}(y_{ij} = 1) p(m_j, m_i) \right)$$

Iterate over all the mentions in the document

Usual trick of taking negative log to go from likelihood to loss

- Testing time: same as mention-pair but we only pick one antecedent for each mention

Clustering-based models

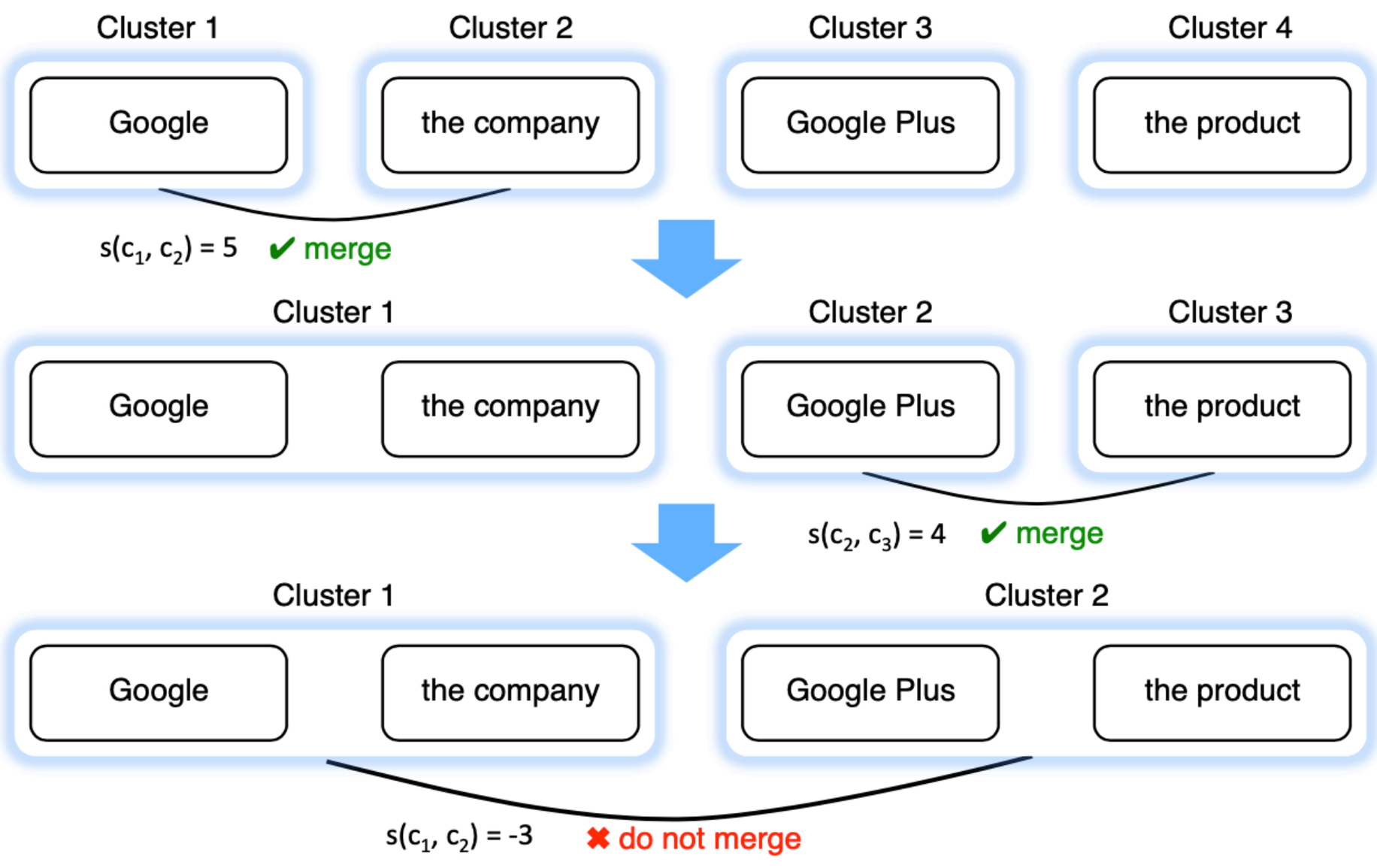
- Coreference is a clustering task, so let's use a clustering algorithm directly.

- Start with each mention in its own singleton cluster

- Merge a pair of clusters at each step and use a model to score which cluster merges are good.

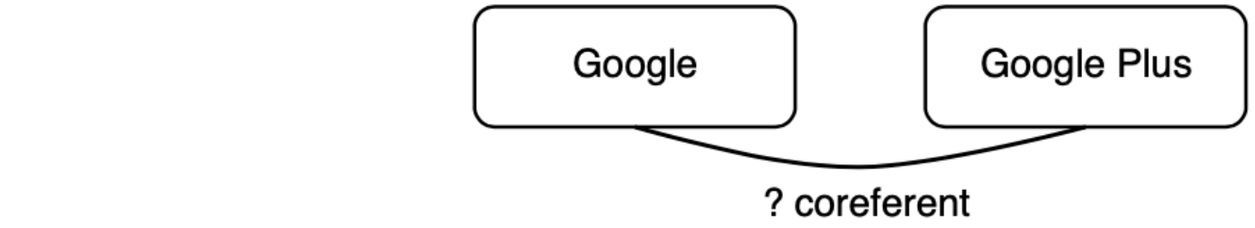
Clustering-based models

Google recently … **the company** announced **Google Plus** … **the product** features …

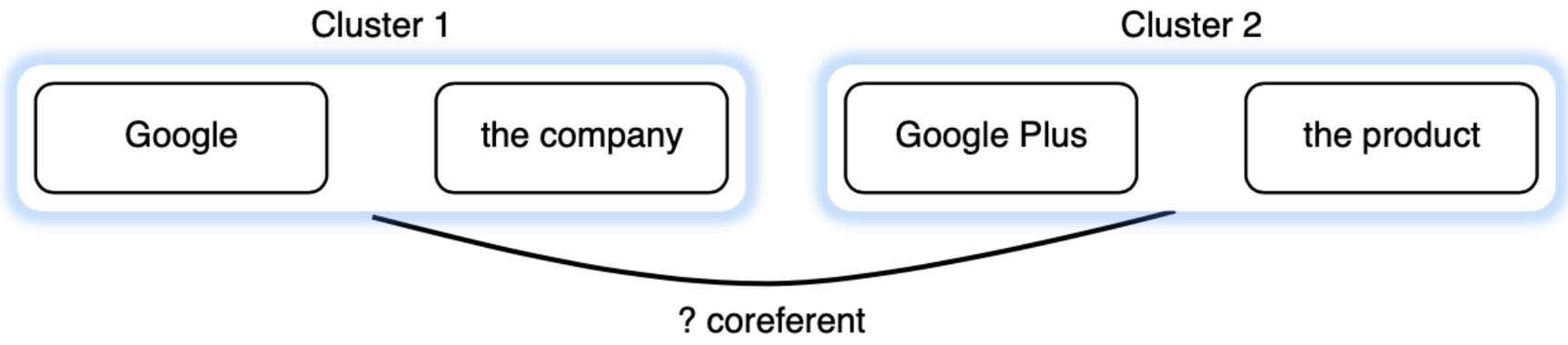


Clustering-based models

Mention-pair decision is difficult



Cluster-pair decision is easier



Clustering-based models

- Current candidate cluster merges depend on previous ones it already made
    - We can't use supervised learning.
    - Researchers used reinforcement learning to train the model
        - Reward function is the change in a coreference evaluation metric

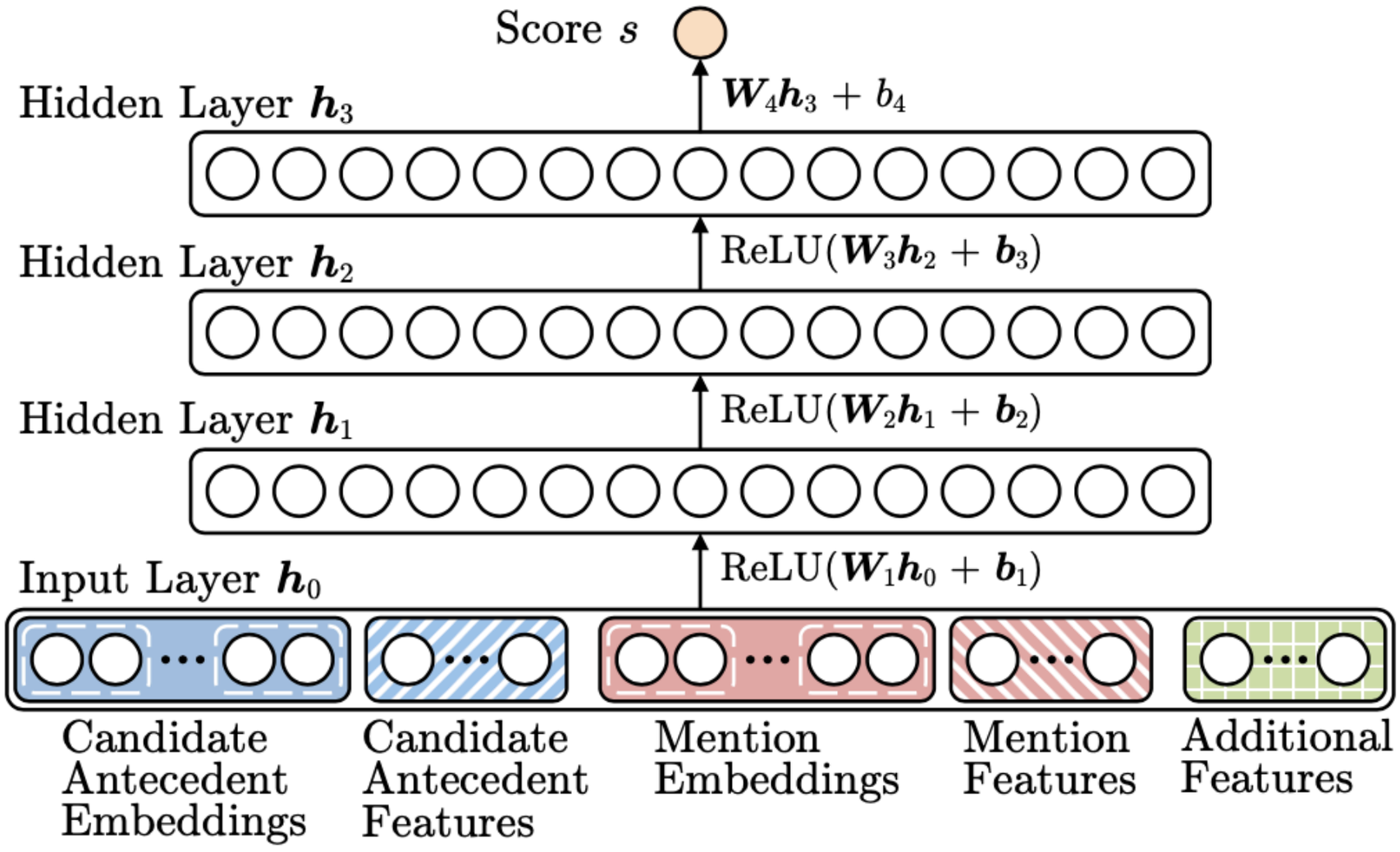
- Some nice attempts but NOT the dominating approach

Features

- Non-neural coreference models:
  - Person/Number/Gender agreement
    - Jack gave Mary a gift. She was excited.
  - Semantic compatibility
    - … the mining conglomerate … the company …
  - Certain syntactic constraints
    - John bought him a new car. [him can not be John]
  - More recently mentioned entities preferred for referenced
    - John went to a movie. Jack went as well. He was not busy.
  - Grammatical Role: Prefer entities in the subject position
    - John went to a movie with Jack. He was not busy.
  - Parallelism:
    - John went with Jack to a movie. Joe went with him to a bar.

Features

- Neural coreference models:



End-to-end coreference resolution

- A mention-ranking model
- Joint mention detection and clustering — so you don't need an additional mention detector (parser/part-of-speech tagger)

$$J = \sum_{i=2}^{N} -\log \left(\sum_{j=1}^{i-1} \mathbb{1}(y_{ij} = 1) p(m_j, m_i) \right)$$

Iterate over all the mentions in the document

Usual trick of taking negative log to go from likelihood to loss

We consider all the possible spans + {NA}

$$p(m_j, m_i) = \frac{\exp(s(m_j, m_i))}{\sum_{j' < i} \exp(m_{j'}, m_i)}$$

$$N = \frac{T(T+1)}{2} + 1$$

T: number of words

(Lee et al, 2017): End-to-end Neural Coreference Resolution

End-to-end coreference resolution

$$s(i,j) = s_{\mathrm{m}}(i) + s_{\mathrm{m}}(j) + s_{\mathrm{a}}(i,j)$$

Are spans *i* and *j* coreferent mentions?

Is *i* a mention?

Is *j* a mention?

Do they look coreferent?

$$s_{\mathrm{m}}(i) = \boldsymbol{w}_{\mathrm{m}} \cdot \mathrm{FFNN}_{\mathrm{m}}(\boldsymbol{g}_i)$$

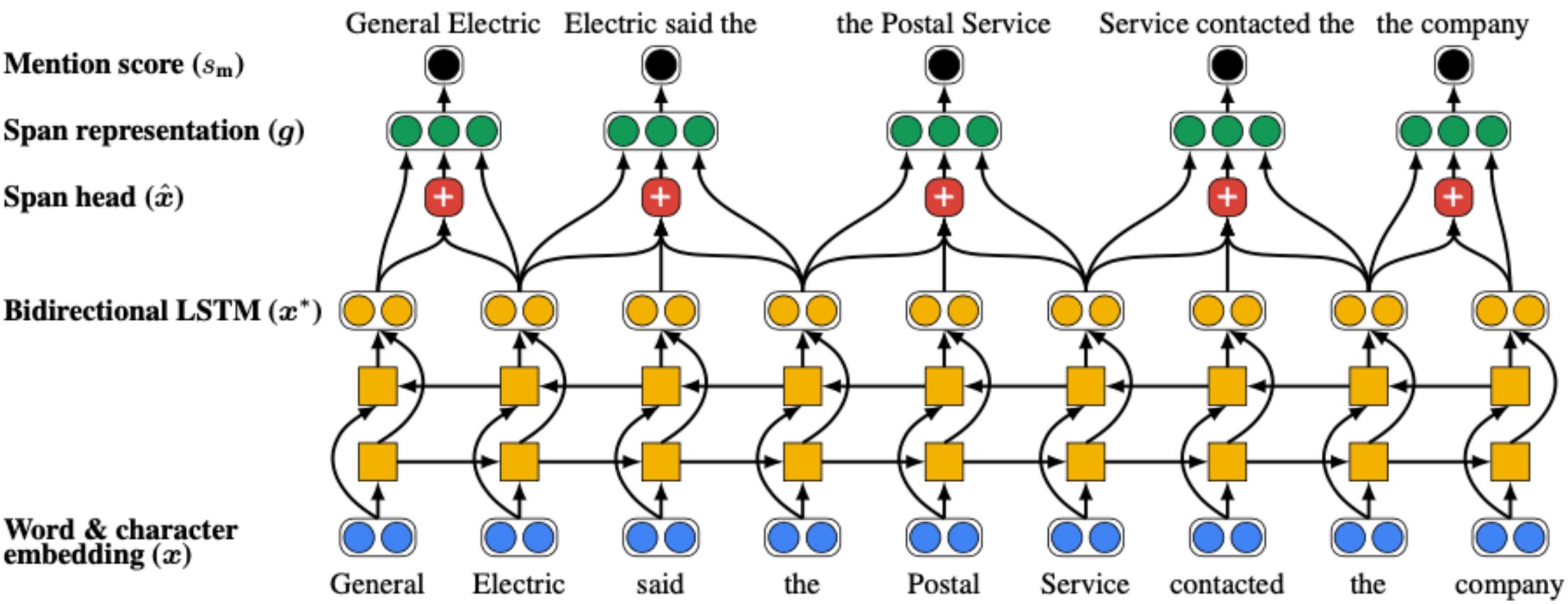
$$s_{\mathrm{a}}(i,j) = \boldsymbol{w}_{\mathrm{a}} \cdot \mathrm{FFNN}_{\mathrm{a}}([\boldsymbol{g}_i, \boldsymbol{g}_j, \boldsymbol{g}_i \circ \boldsymbol{g}_j, \phi(i,j)])$$

Let's compute a vector representation $\mathbf{g}_i \in \mathbb{R}^d$ for each span i

$\phi(i,j)$: manual features such speaker/gender information

(Lee et al, 2017): End-to-end Neural Coreference Resolution

End-to-end coreference resolution



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

BILSTM hidden states for span's start and end

Attention-based representation (details next slide) of the words in the span

Additional features

(Lee et al, 2017): End-to-end Neural Coreference Resolution

End-to-end coreference resolution

Attention scores

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

dot product of weight vector and transformed hidden state

Attention distribution

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

just a softmax over attention scores for the span

Final representation

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

Attention-weighted sum of word embeddings

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

hidden states for span's start and end

Attention-based representation

Additional features

Represents the context to the left and right of the span

Represents the span itself

Represents other information not in the text

(Lee et al, 2017): End-to-end Neural Coreference Resolution

End-to-end coreference resolution

Testing time:

- $O(T^2)$ spans of text in a document
- $O(T^4)$ possible pairs of spans — too expensive!
- Use $s_m(i)$ for aggressive pruning

(Lee et al, 2017): End-to-end Neural Coreference Resolution

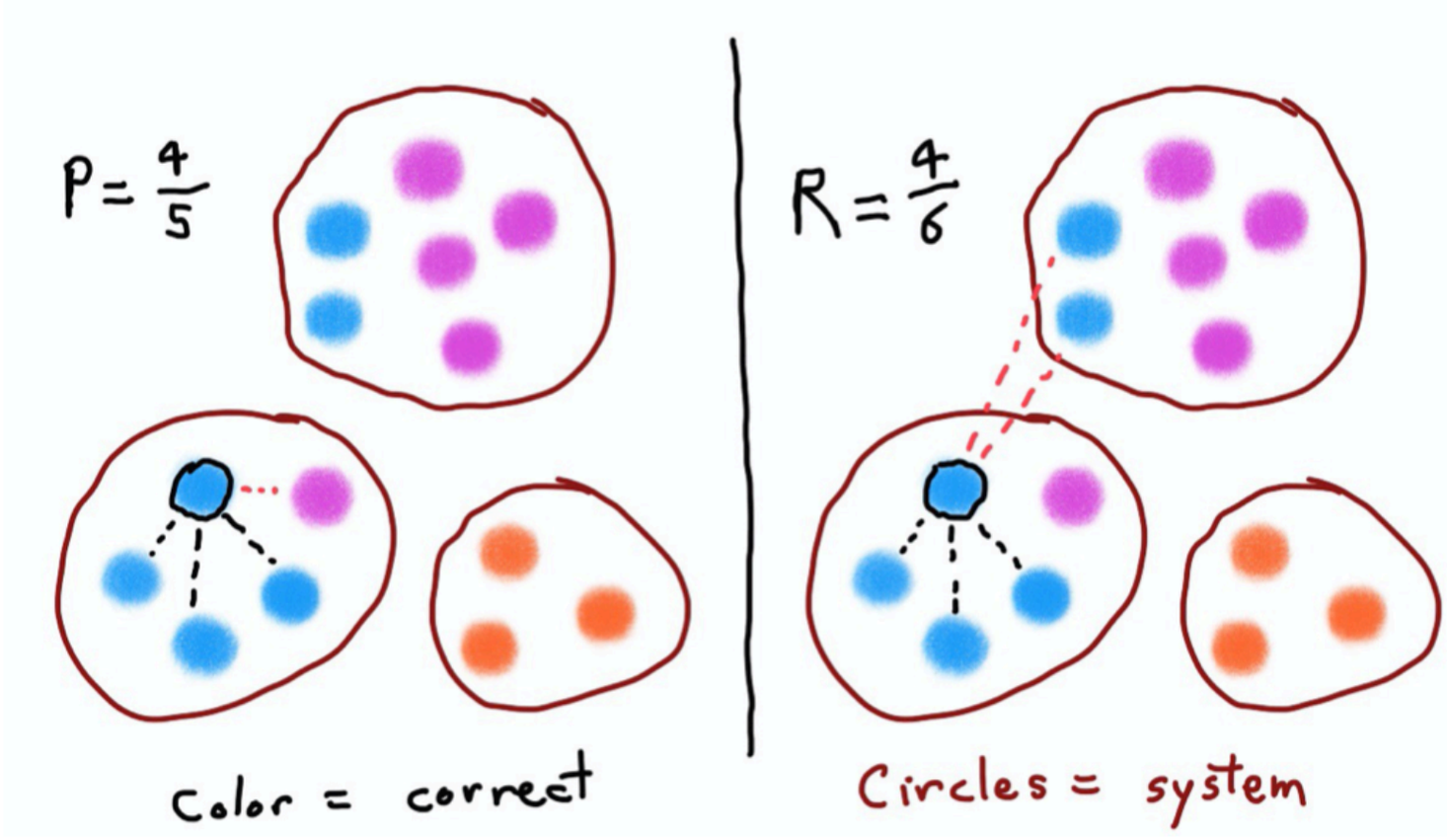
Coreference evaluation

- You need to get both “mentions” and “clusters” correctly.
- Standard practice: we use 3 types of metrics
  - B^3: mention-based
  - MUC: link-based (pair of mentions)
  - CEAF: entity-based
- .. and finally take the average of these 3 F1 scores

https://github.com/conll/reference-coreference-scorers

B-cubed evaluation metric

- For each mention in the reference chain, compute a precision and a recall (e.g., # of mentions in the same reference chain with the current mention)
- The final precision/recall is an average of all the mentions



System performance

- Evaluation on English Ontonotes (CoNLL-2012 Shared Task)
- #Train: 2,802 / #Dev: 343 / #Test: 348 documents

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
Lee et al. (2017) (single model)	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
Fernandes et al. (2014)	75.9	65.8	70.5	77.7	65.8	71.2	43.2	55.0	48.4	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

"**Eash Victories** and **Uphill Battles** in Coreference Resolution"

System performance