coreference resolution

CS 585, Fall 2018

Introduction to Natural Language Processing

http://people.cs.umass.edu/~miyyer/cs585/

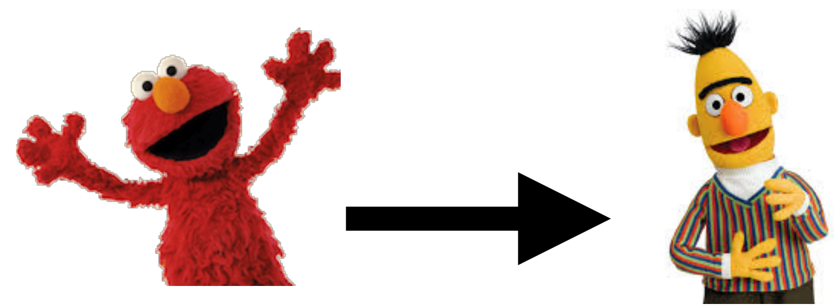
Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

most slides from Richard Socher

questions from last time

- what is BERT? 
- what's up with CS690D (deep learning for NLP)?
- progress reports due Friday!
- project breakdown: 5% proposal, 5% progress report, 25% final report & presentation
- any topics you want to cover? high-level or low-level tasks? more QA? dialog? ethics? something else?

What is 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.

What is 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.

What is 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.

What is 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

What is 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.

What is 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 he
had foreign affairs experience as a former

Applications

- Full text understanding
  - information extraction, question answering, summarization, ...
  - “He was born in 1961”

Applications

- Full text understanding
- Machine translation
- Dialogue Systems

  "Book tickets to see James Bond"

  "Spectre is playing near you at 2:00 and 3:00 today. How many tickets would you like?"

  "Two tickets for the showing at three"

Coreference Resolution is Really Difficult!

- “She poured water from the pitcher into the cup until it was full”
- Requires reasoning /world knowledge to solve

Coreference Resolution is Really Difficult!

- “She poured water from the pitcher into the cup until it was full”
- “She poured water from the pitcher into the cup until it was empty”

- Requires reasoning /world knowledge to solve

Coreference Resolution is Really Difficult!

- "She poured water from the pitcher into the cup until it was full"
- "She poured water from the pitcher into the cup until it was empty"

- The trophy would not fit in the suitcase because it was too big.
- The trophy would not fit in the suitcase because it was too small.

- These are called **Winograd Schema**

Coreference Resolution is Really Difficult!

- "She poured water from the pitcher into the cup until it was full"
- "She poured water from the pitcher into the cup until it was empty"

- The trophy would not fit in the suitcase because it was too big.
- The trophy would not fit in the suitcase because it was too small.

- These are called **Winograd Schema**
  - Recently proposed as an alternative to the Turing test
    - Turing test: how can we tell if we've built an AI system? A human can't distinguish it from a human when chatting with it.
    - But requires a person, people are easily fooled
  - If you've fully solved coreference, arguably you've solved AI

Coreference Resolution in Two Steps

1. Detect the mentions (easy to do in many cases)

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

   - mentions can be nested!

2. Cluster the mentions generally hard to do

   "[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:

what about event coreference?

The president's **speech** shocked the audience. He **announced** several new controversial policies.

1. Pronouns
   - I, your, it, she, him, etc.

2. Named entities
   - People, places, etc.

3. Noun phrases
   - "a dog," "the big fluffy cat stuck in the tree"

Mention Detection

- Span of text referring to some entity
- For detection: use other NLP systems

1. Pronouns
   - Use a part-of-speech tagger

2. Named entities
   - Use a NER system

3. Noun phrases
   - Use a constituency parser

Mention Detection: Not so Simple

- Marking all pronouns, named entities, and NPs as mentions over-generates mentions
- Are these mentions?
  - It is sunny
  - Every student
  - No student
  - The best donut in the world
  - 100 miles
- Some gray area in defining "mention": have to pick a convention and go with it

How to deal with these bad mentions?

- Could train a classifier to filter out spurious mentions
- Much more common: keep all mentions as "candidate mentions"
  - After your coreference system is done running discard all singleton mentions (i.e., ones that have not been marked as coreference with anything else)

Can we avoid a pipelined system?

- We could instead train a classifier specifically for mention detection instead of using a POS tagger, NER system, and parser.

- Or even jointly do mention-detection and coreference resolution end-to-end instead of in two steps
  - Will cover later in this lecture!

On to Coreference! First, some linguistics

- **Coreference** is when two mentions refer to the same entity in the world
  - *Barack Obama traveled to … Obama*

- Another kind of reference is **anaphora:** when a term (anaphor) refers to another term (antecedent) and 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

Anaphora vs Coreference

- Coreference with named entities

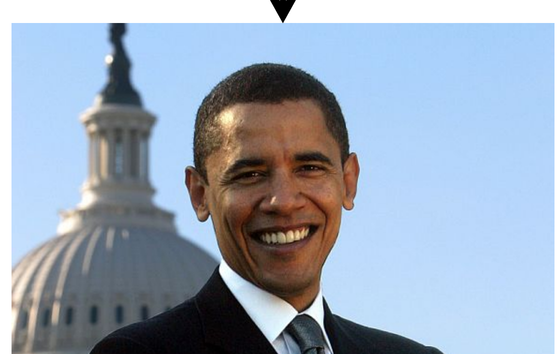
text Barack Obama Obama

world

- Anaphora

text Barack Obama he

world

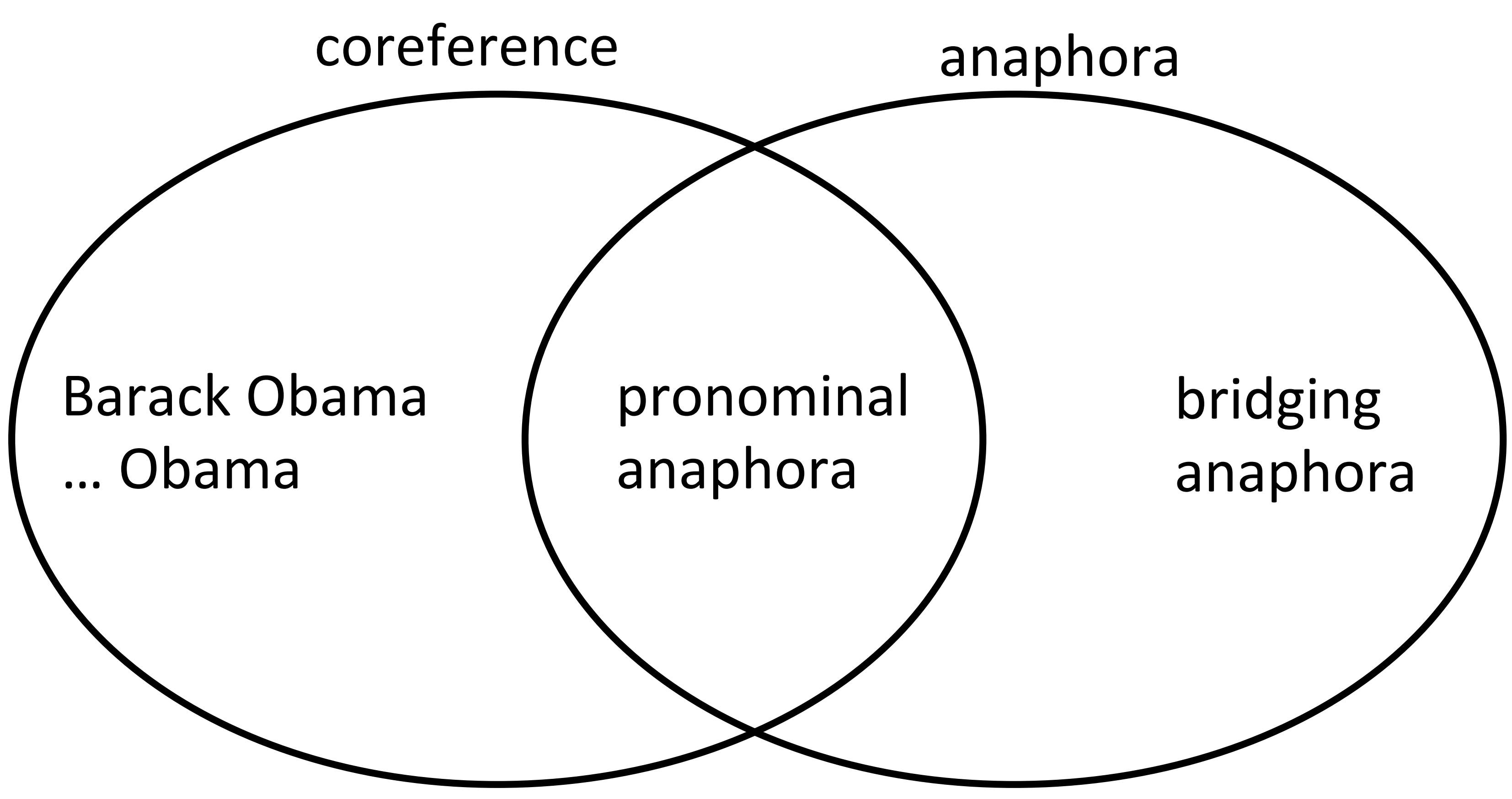


Anaphora vs. Coreference

- Not all anaphoric relations are coreferential

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

- This is referred to as bridging anaphora.



Cataphora

"From the corner of the divan of Persian saddle-bags on which he was lying, smoking, as was his custom, innumerable cigarettes, Lord Henry Wotton could just catch the gleam of the honey-sweet and honey-coloured blossoms of a laburnum…"

(Oscar Wilde – The Picture of Dorian Gray)

Next Up: Three Kinds of Coreference Models

- Mention Pair
- Mention Ranking
- Clustering

Coreference Models: Mention Pair

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

I Nader he my she

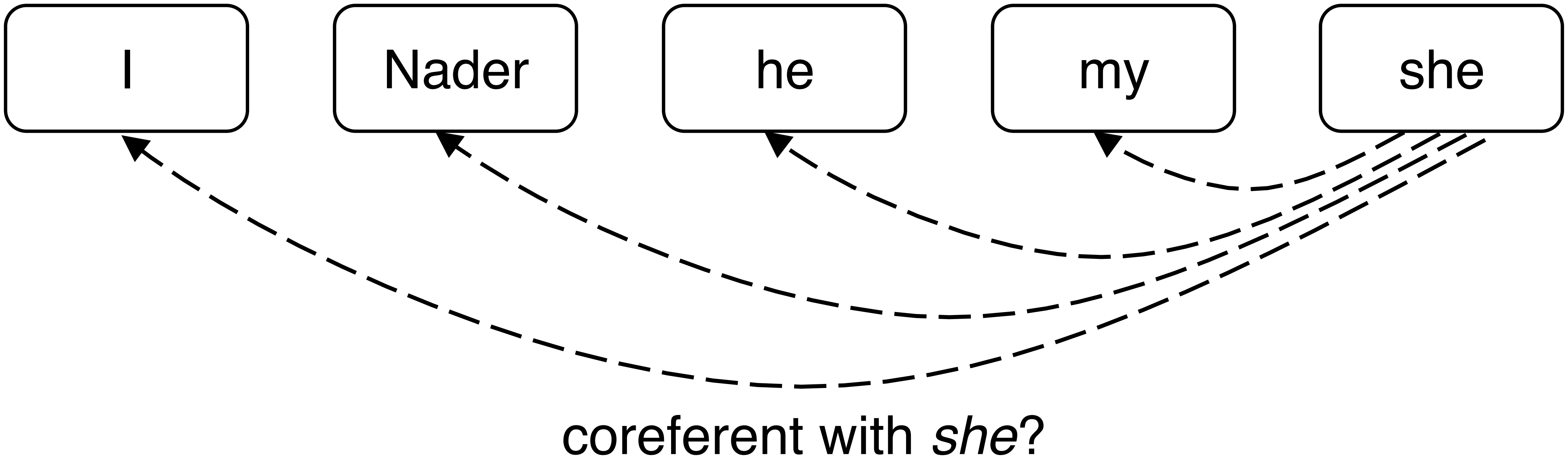
Coreference Cluster 1

Coreference Cluster 2

Coreference Models: Mention Pair

- Train a binary classifier that assigns every pair of mentions a probability of being coreferent: $p(m_i, m_j)$
  - e.g., for “she” look at all **candidate antecedents** (previously occurring mentions) and decide which are coreferent with it

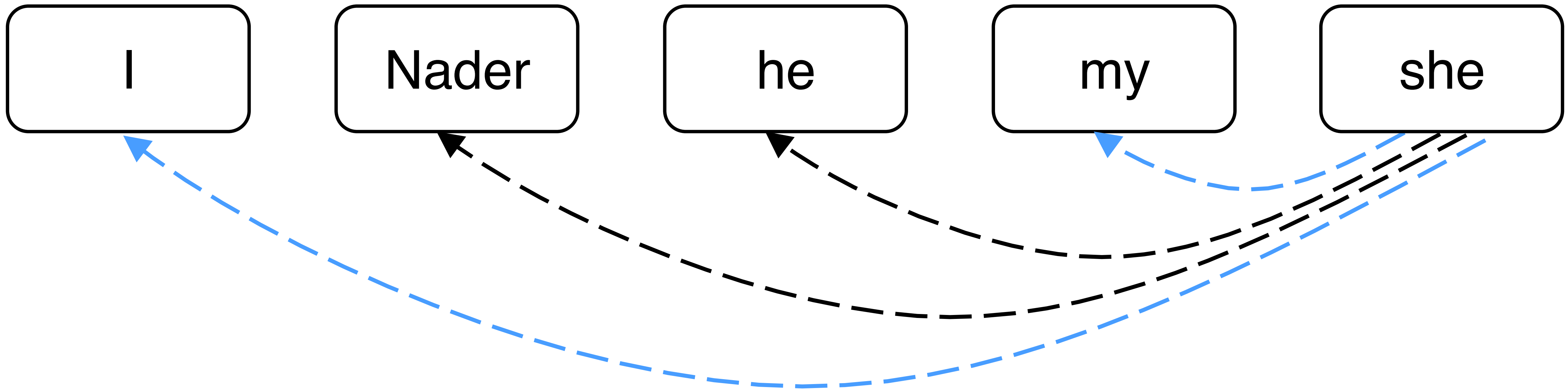
*“**I** voted for **Nader** because **he** was most aligned with **my** values,” **she** said.*



Coreference Models: Mention Pair

- Train a binary classifier that assigns every pair of mentions a probability of being coreferent: $p(m_i, m_j)$
  - e.g., for "she" look at all **candidate antecedents** (previously occurring mentions) and decide which are coreferent with it

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

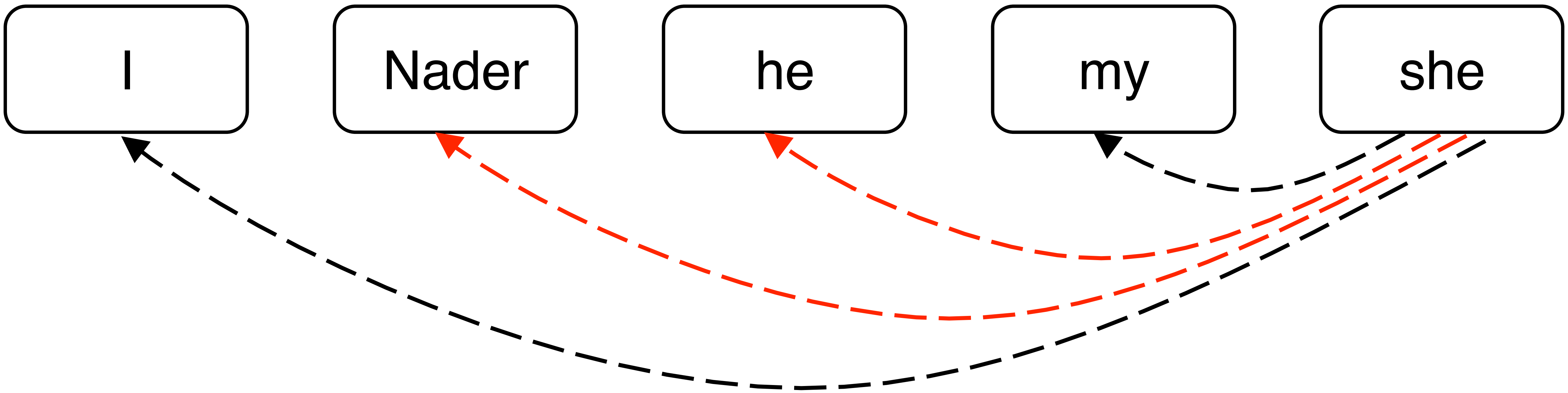


Positive examples: want $p(m_i, m_j)$ to be near 1

Coreference Models: Mention Pair

- Train a binary classifier that assigns every pair of mentions a probability of being coreferent: $p(m_i, m_j)$
  - e.g., for "she" look at all **candidate antecedents** (previously occurring mentions) and decide which are coreferent with it

*"**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

Mention Pair Training

- N mentions in a document
- y_{ij} = 1 if mentions m_i and m_j are coreferent, -1 if otherwise
- Just train with regular cross-entropy loss (looks a bit different because it is binary classification)

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

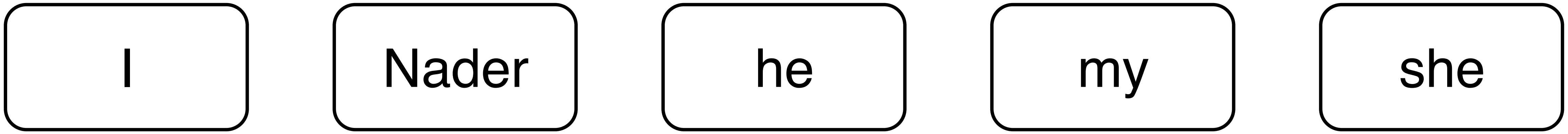
Iterate through mentions

Iterate through candidate antecedents (previously occurring mentions)

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

Mention Pair Test Time

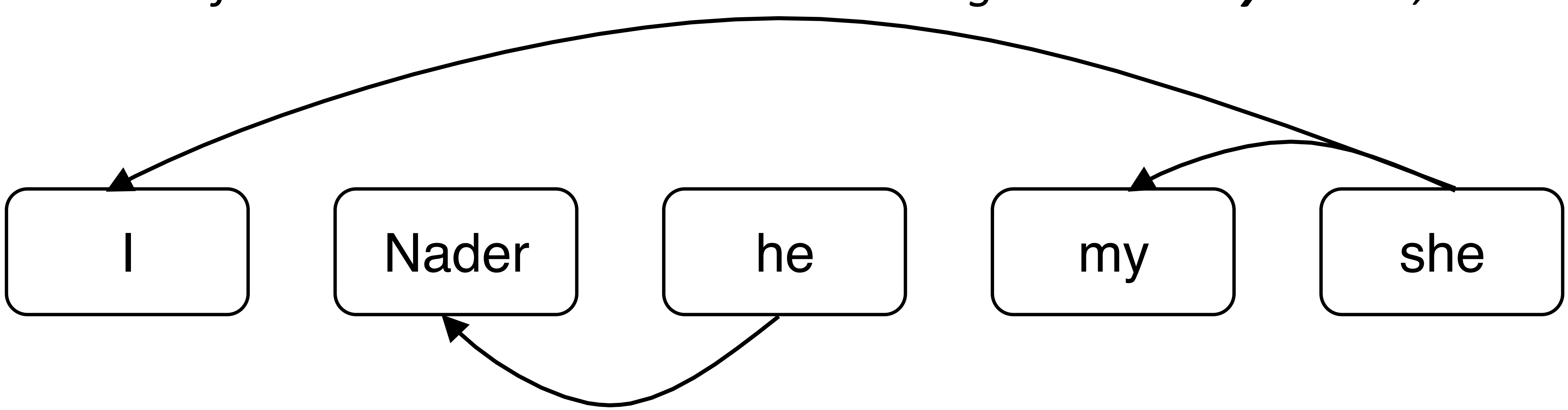
- Coreference resolution is a clustering task, but we are only scoring pairs of mentions… what to do?



Mention Pair Test Time

- Coreference resolution is a clustering task, but we are only scoring pairs of mentions… what to do?
- Pick some threshold (e.g., 0.5) and add **coreference links** between mention pairs where $p(m_i, m_j)$ is above the threshold

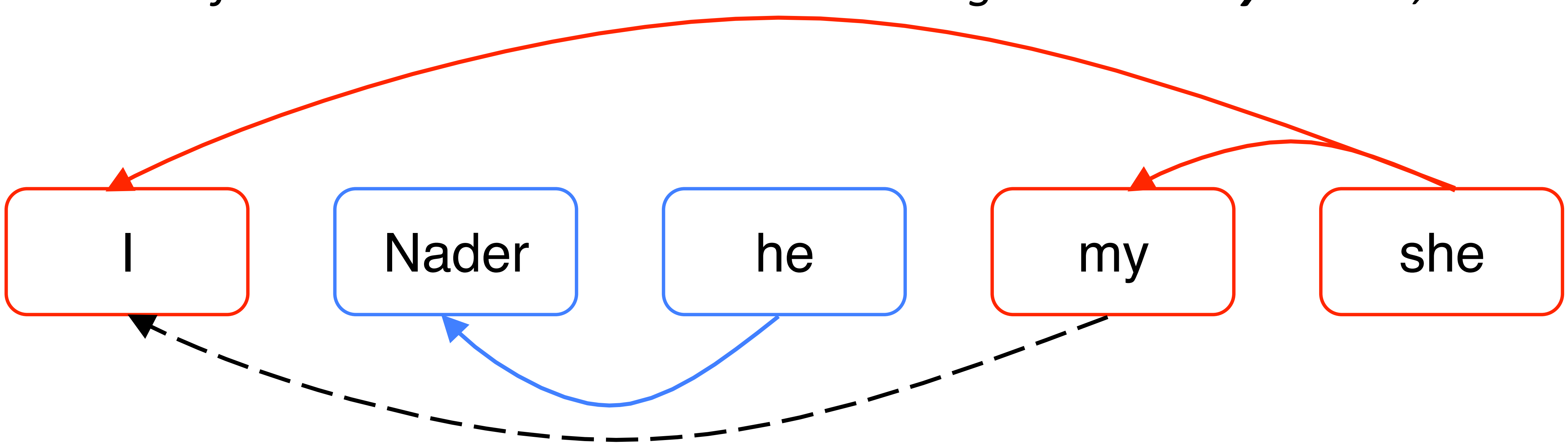
*"**I** voted for **Nader** because **he** was most aligned with **my** values," **she** said.*



Mention Pair Test Time

- Coreference resolution is a clustering task, but we are only scoring pairs of mentions… what to do?
- Pick some threshold (e.g., 0.5) and add **coreference links** between mention pairs where $p(m_i, m_j)$ is above the threshold
- Take the transitive closure to get the clustering

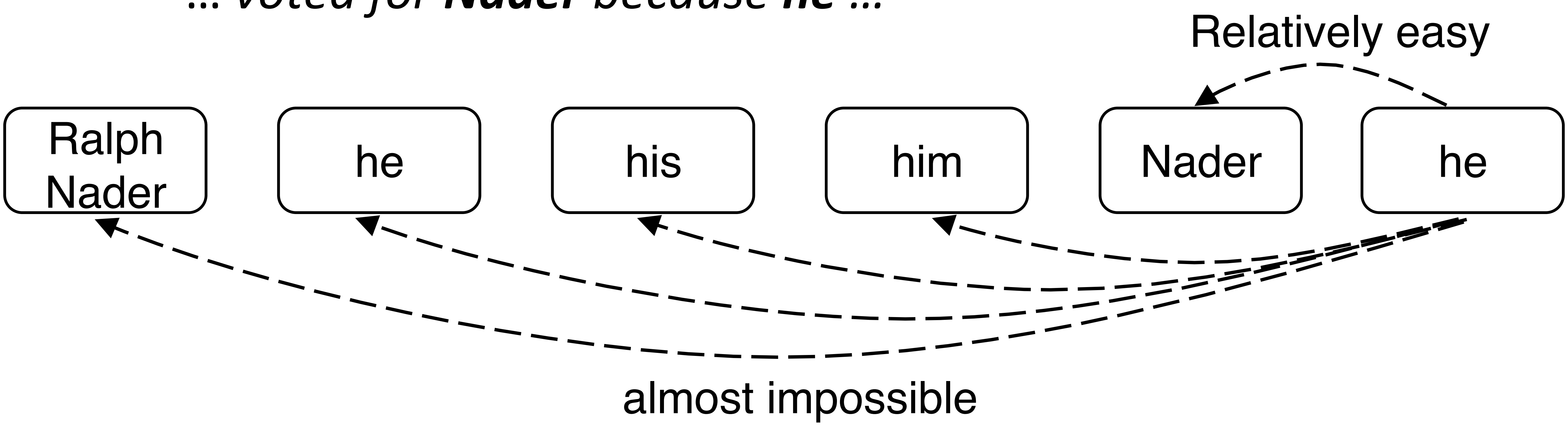
*"**I** voted for **Nader** because **he** was most aligned with **my** values," **she** said.*



Even though the model did not predict this coreference link, *I* and *my* are coreferent due to transitivity

Mention Pair Models: Disadvantage

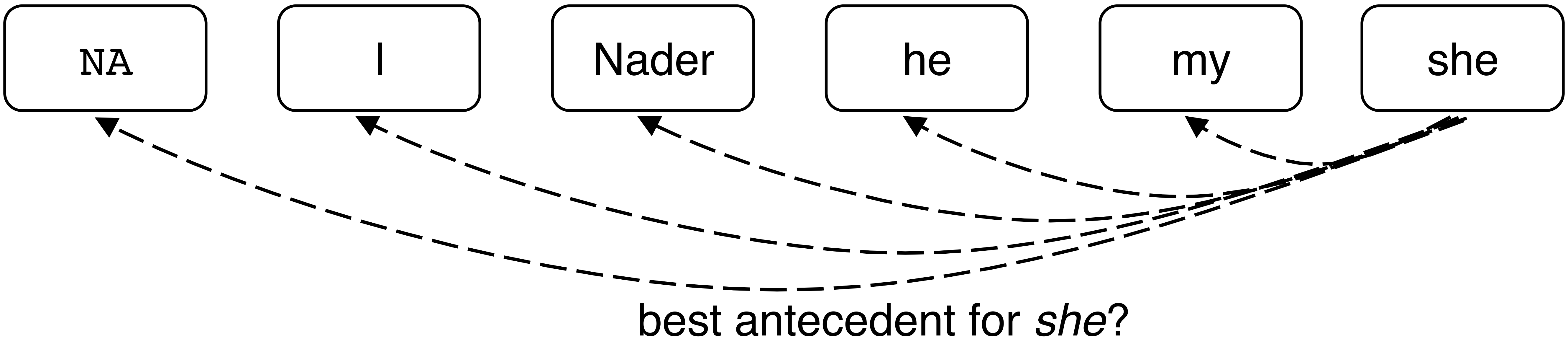
- Suppose we have a long document with the following mentions
  - **Ralph Nader … he … his … him …** <several paragraphs> … *voted for **Nader** because **he** …*



- 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

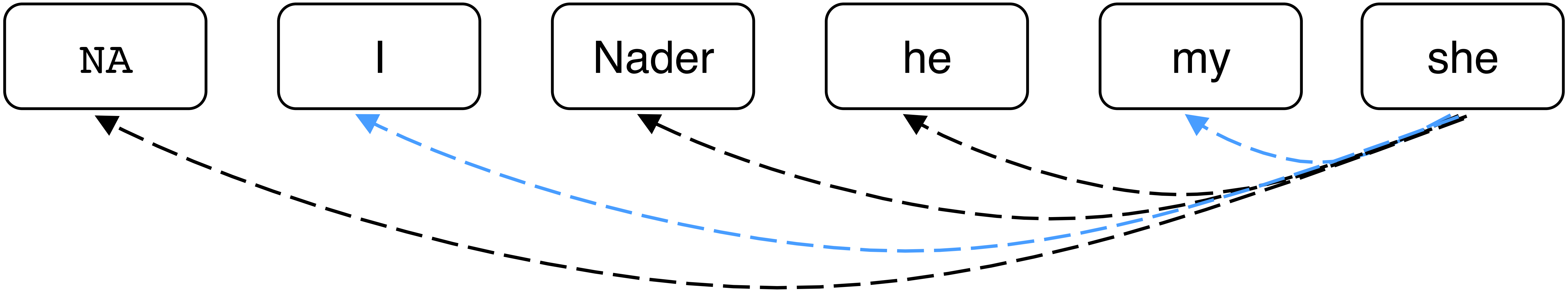
Coreference Models: Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything



Coreference Models: Mention Ranking

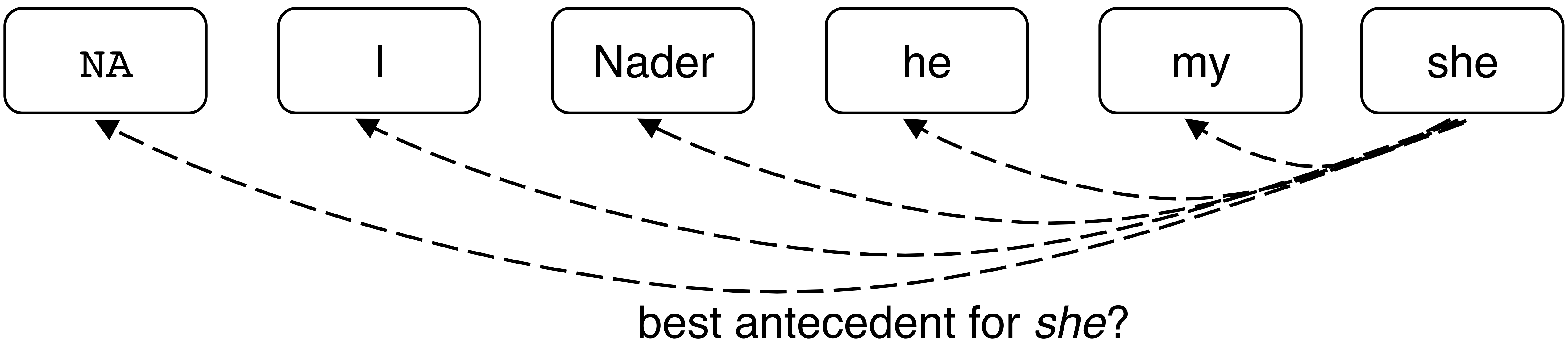
- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything



Positive examples: model has to assign a high probability to either one (but not necessarily both)

Coreference Models: Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything

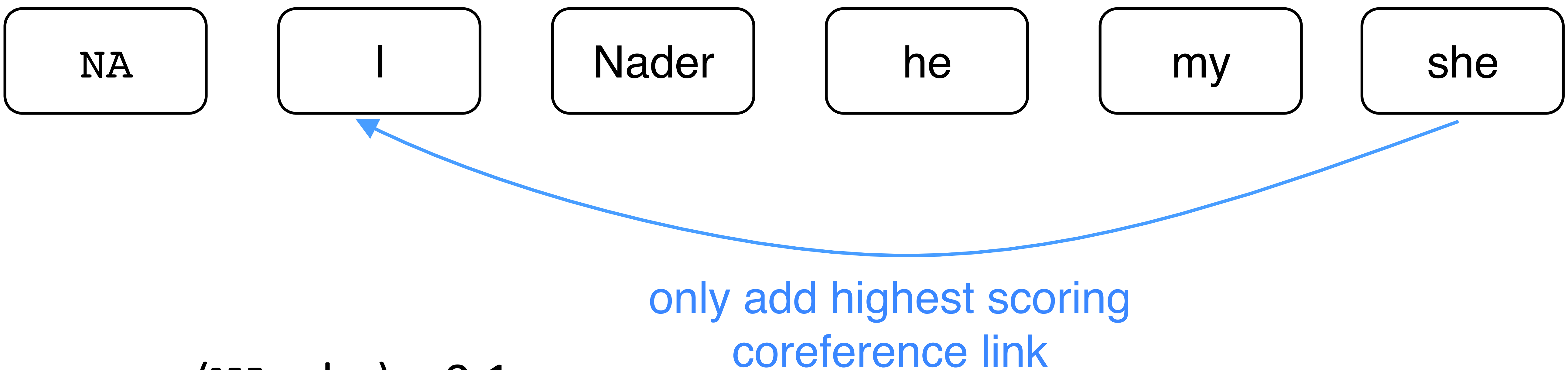


p(NA, she) = 0.1
p(I, she) = 0.5
p(Nader, she) = 0.1
p(he, she) = 0.1
p(my, she) = 0.2

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

Coreference Models: Mention Ranking

- Assign each mention its highest scoring candidate antecedent according to the model
- Dummy NA mention allows model to decline linking the current mention to anything



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

How do we compute the probabilities?

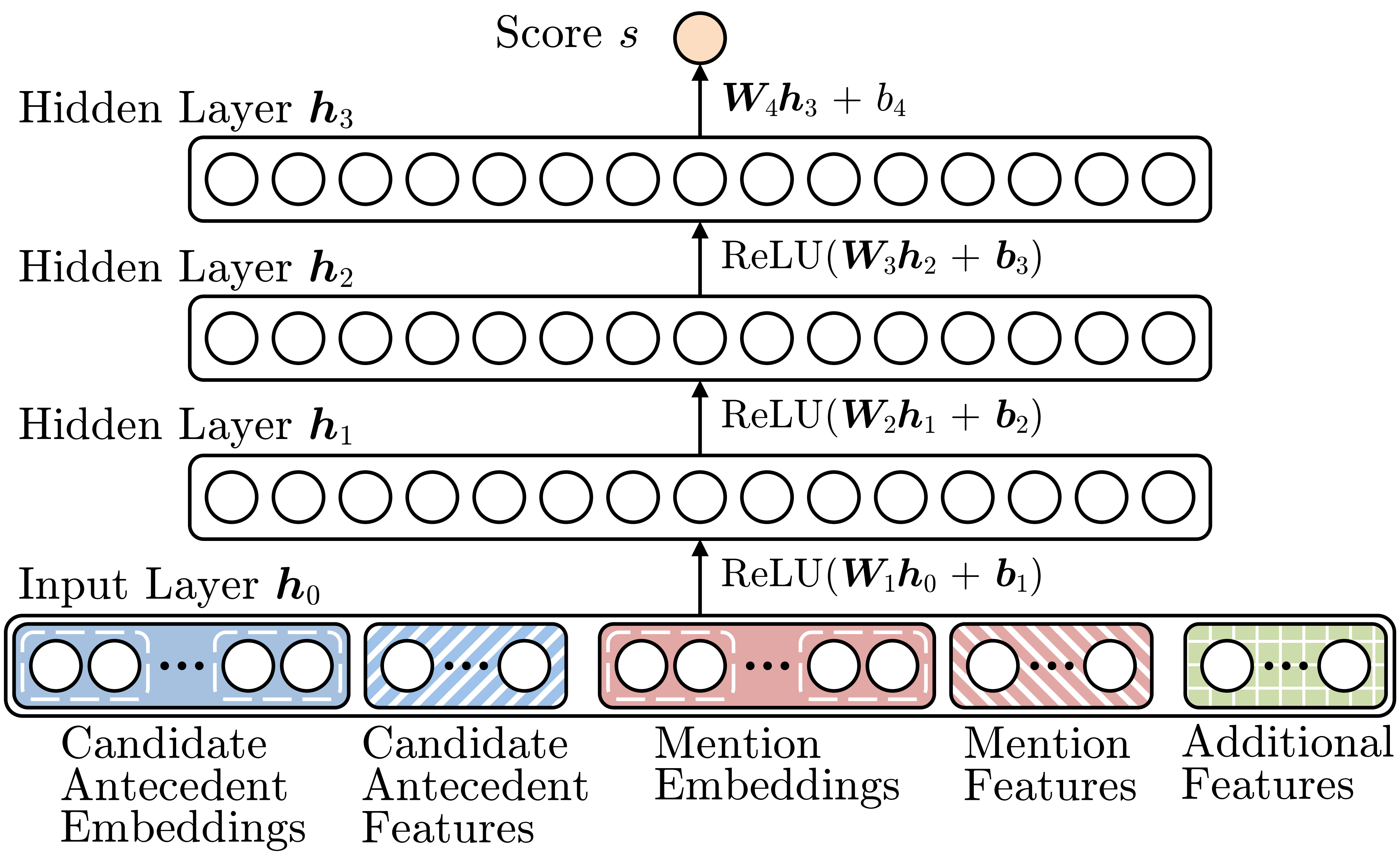
1. Non-neural statistical classifier
2. Simple neural network
3. More advanced model using LSTMs, attention

1. Non-Neural Coref Model: Features

- 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.
- …

2. Neural Coref Model

- Standard feed-forward neural network
  - Input layer: word embeddings and a few categorical features



2. Neural Coref Model: Inputs

- Embeddings
  - Previous two words, first word, last word, head word, … of each mention
    - The **head** word is the "most important" word in the mention – you can find it using a parser. e.g., *The fluffy* ***cat*** *stuck in the tree*

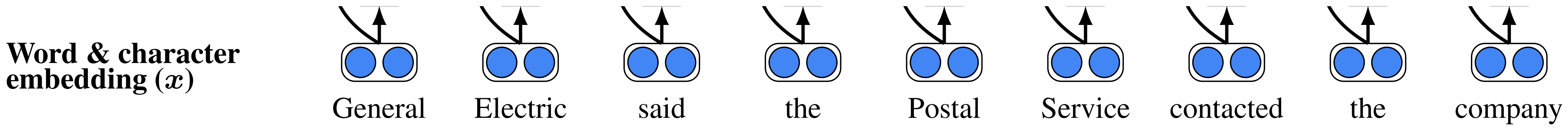
- Still need some other features:
  - Distance
  - Document genre
  - Speaker information

3. End-to-end Model

- Current state-of-the-art model for coreference resolution (Lee et al., EMNLP 2017)
- Mention ranking model
- Improvements over simple feed—forward NN
  - Use an LSTM
  - Use attention
  - Do mention detection and coreference end-to-end
    - No mention detection step!
    - Instead consider every **span** of text (up to a certain length) as a candidate mention
      - a **span** is just a contiguous sequence of words

3. End-to-end Model

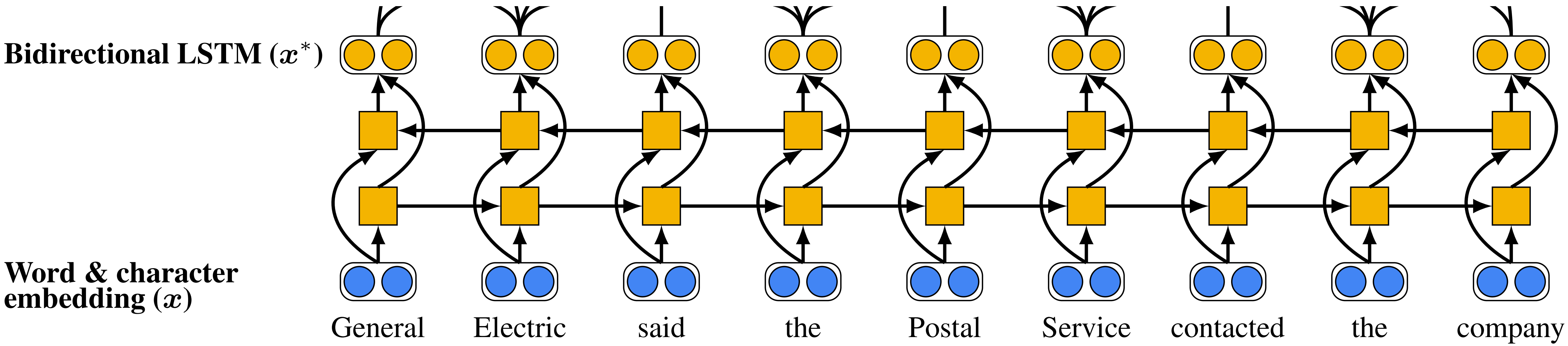
- First embed the words in the document using a word embedding matrix and a character-level CNN



3. End-to-end Model

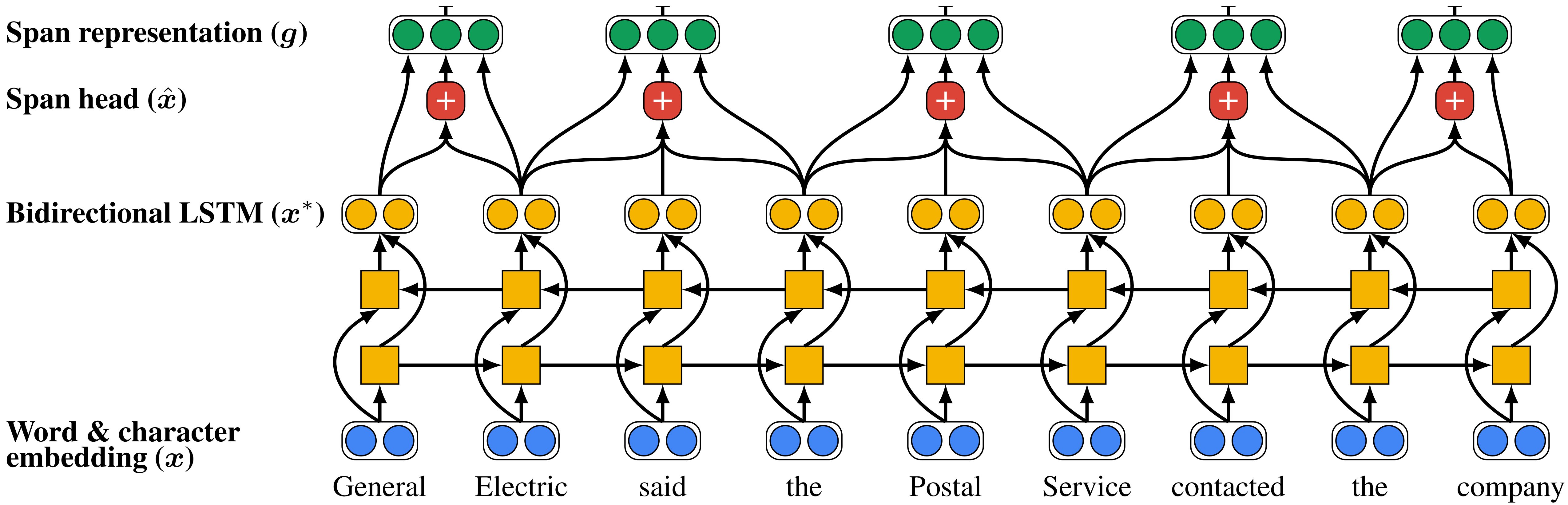
- Then run a bidirectional LSTM over the document

LSTMs are fancy RNNs



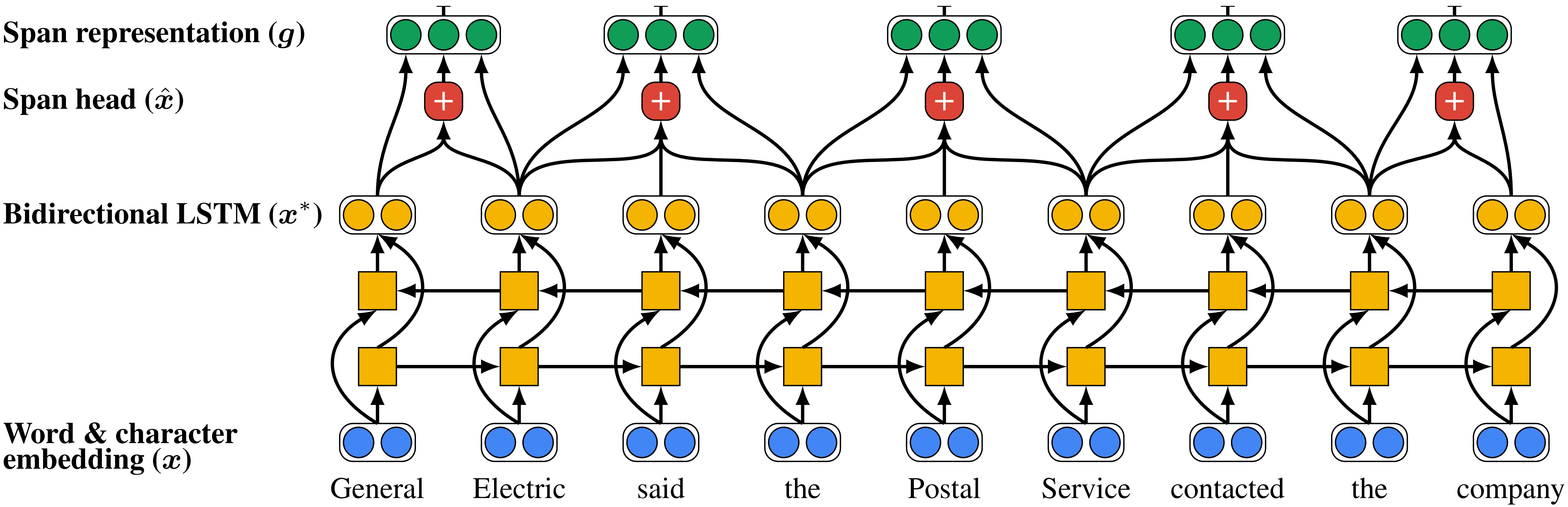
3. End-to-end Model

- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector



3. End-to-end Model

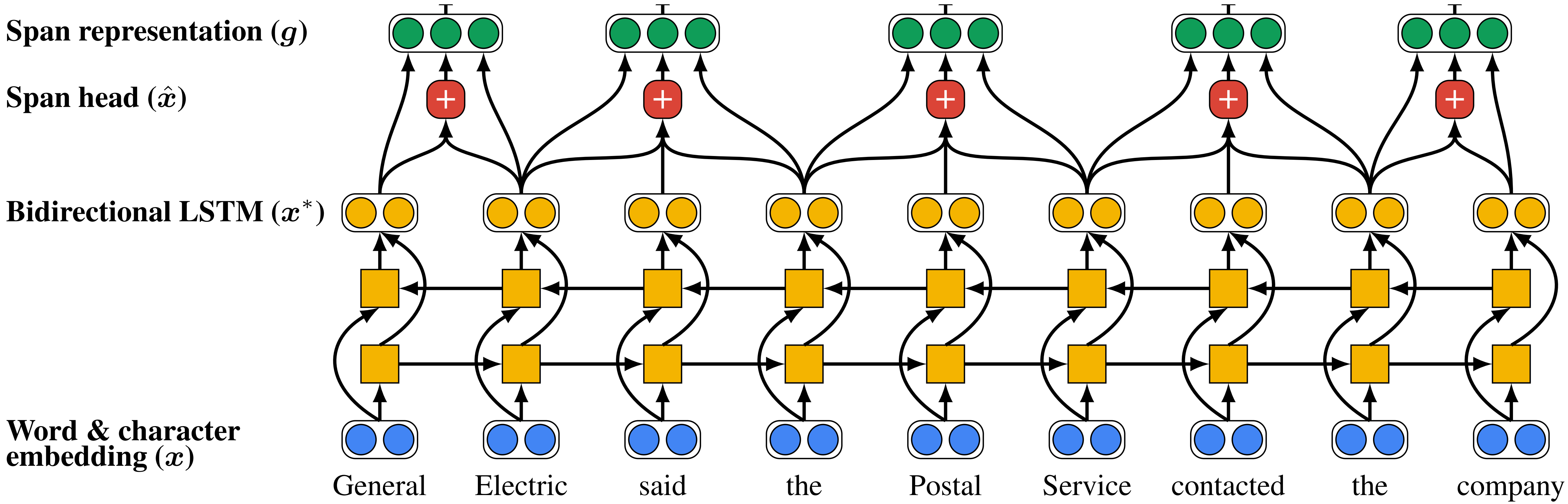
- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector



- *General, General Electric, General Electric said, … Electric, Electric said, …* will all get its own vector representation

3. End-to-end Model

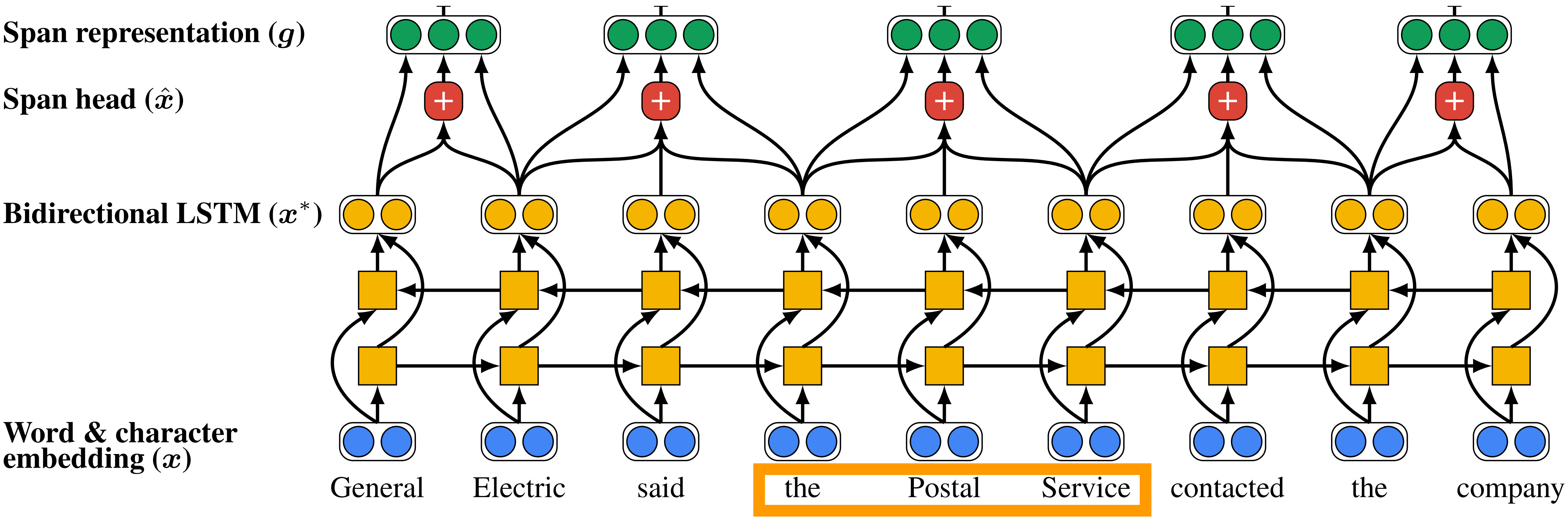
- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector.



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

3. End-to-end Model

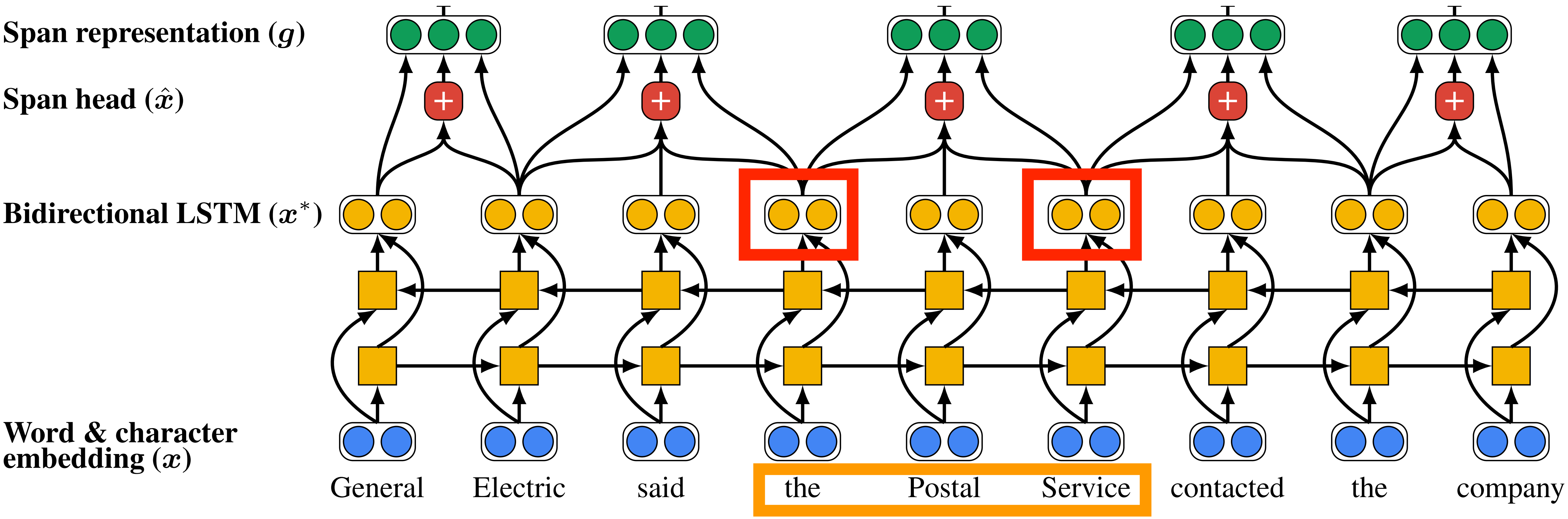
- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector. For example, for "the postal service"



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

3. End-to-end Model

- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector. For example, for "the postal service"

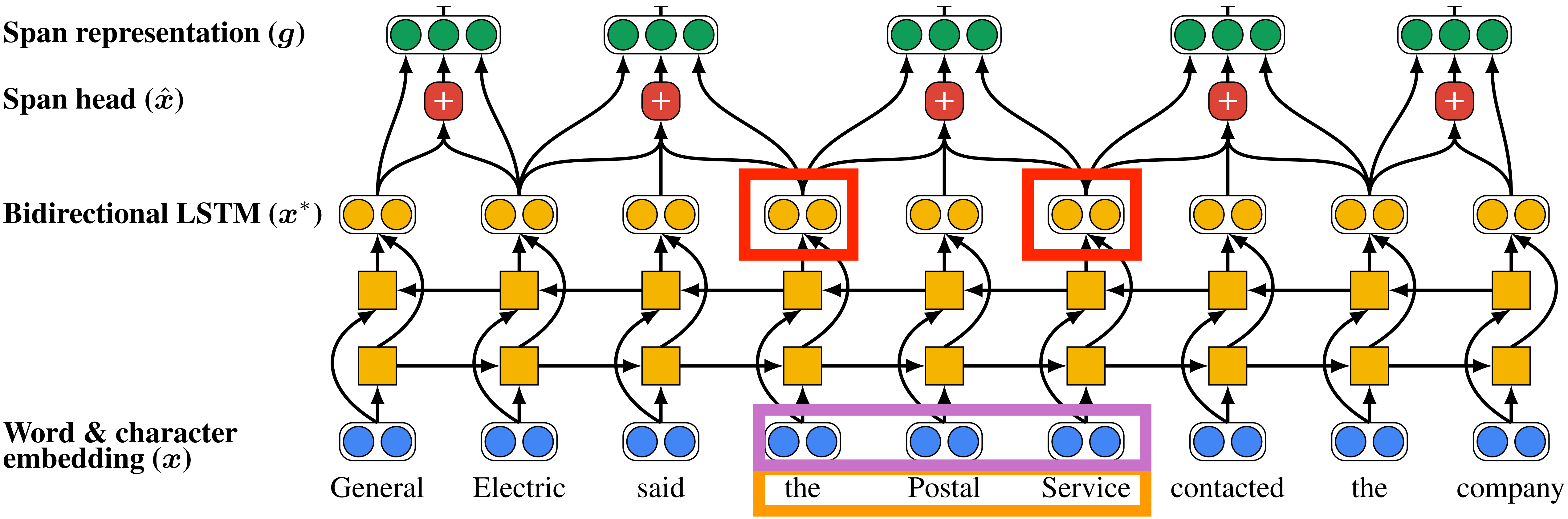


Span representation: $\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$

BILSTM hidden states
for span's start and end

3. End-to-end Model

- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector. For example, for "the postal service"



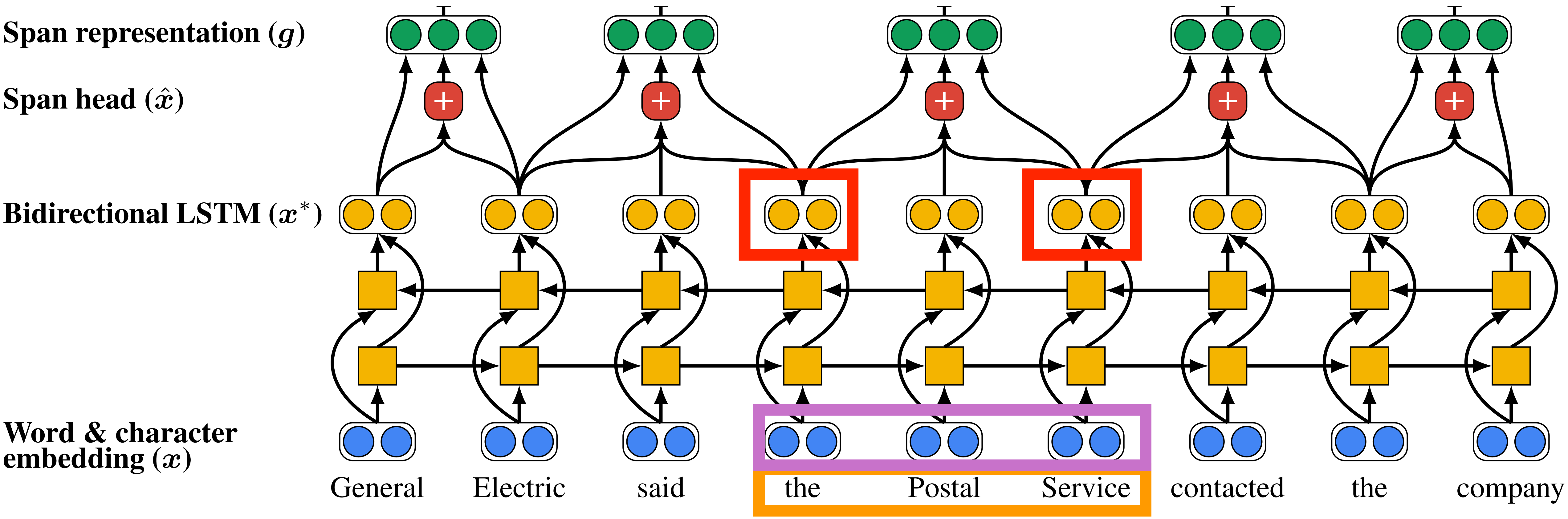
Span representation: $\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$

BILSTM hidden states for span's start and end

Attention-based representation of the words in the span

3. End-to-end Model

- Next, represent each span of text *i* going from START(*i*) to END(*i*) as a vector. For example, for “the postal service”



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

BILSTM hidden states for span’s start and end

Attention-based representation of the words in the span

Additional features

3. End-to-end Model

- Why include all these different terms in the span?

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

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

3. End-to-end Model

- Lastly, score every pair of spans to decide if they are coreferent mentions

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

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

Is *i* a mention?

Is *j* a mention?

Do they look coreferent?

3. End-to-end Model

- Lastly, score every pair of spans to decide if they are coreferent mentions

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

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

Is *i* a mention?

Is *j* a mention?

Do they look coreferent?

- Scoring functions take the span representations as input

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

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

3. End-to-end Model

- Intractable to score every pair of spans
  - O(T^2) spans of text in a document (T is the number of words)
  - O(T^4) runtime!
  - So have to do lots of pruning to make work (only consider a few of the spans that are likely to be mentions)

- Attention learns which words are important in a mention (a bit like head words)

(**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.

exercise!