Resolução de Co-Referência

Luciano Barbosa

(baseado nos slides do curso de PLN de Stanford e livro Speech and Language Processing)

Centro de Informática UFPE

Universidade Federal de Pernambuco

O que é?

- Identificar todas as menções que se referem à mesma entidade (referente)

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.

Aplicações

- Chatbot (Diálogo)

"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"

Aplicações

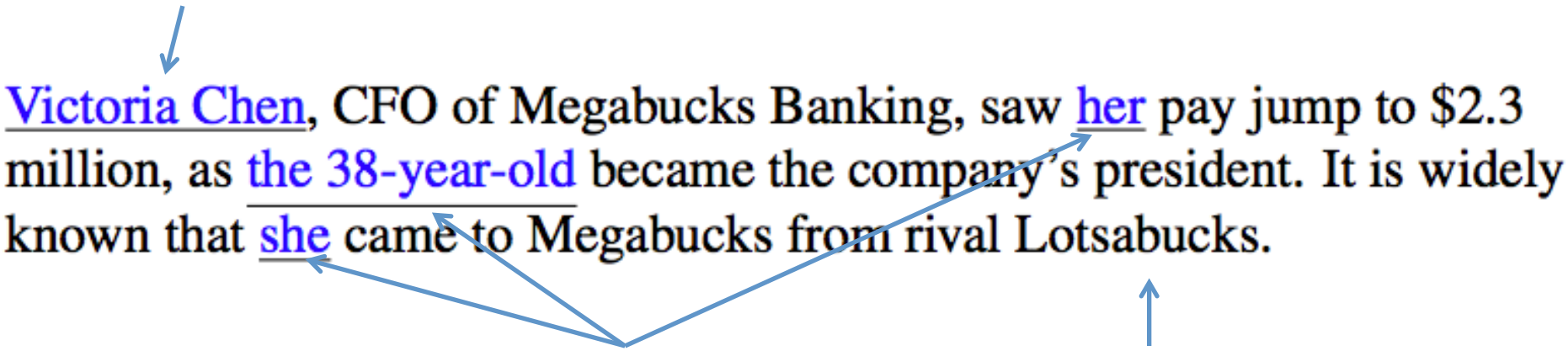
- Q&A
  - Question: “Where Marie Curie was born?”
  - Answer: “She was born in Warsaw”

Aplicações

- Machine Translation

Componentes

referente

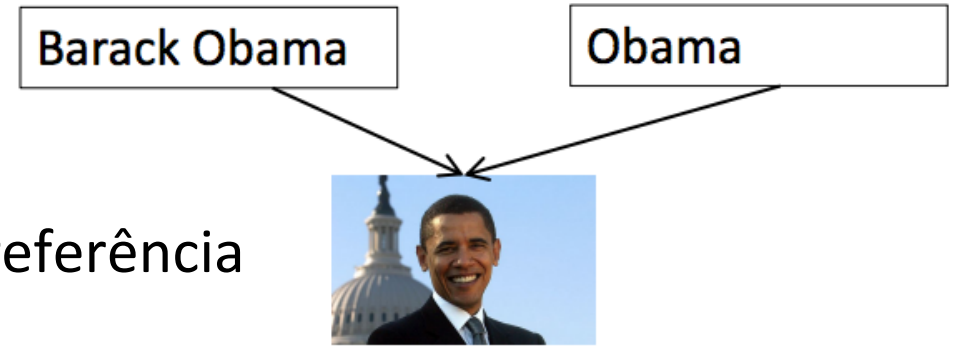


anáfora: refere-se a um termo antecedente

singleton

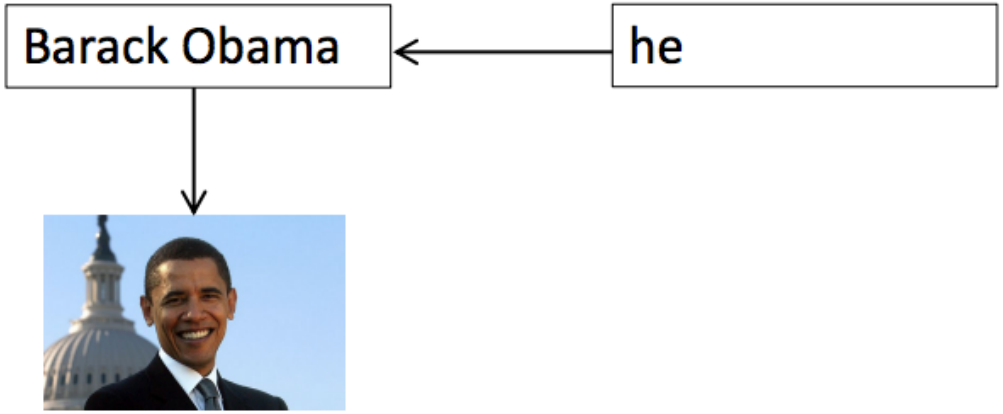
Anáfora vs Correferência

Barack Obama traveled to ... Obama



Co-referência

Barack Obama said he would sign the bill.
antecedent anaphor



Anáfora

Anáforas Podem não Ser Co-referências

- Nem todas noun phrases têm referência

Every dancer twisted *her knee*.

No dancer twisted *her knee.*

Determinando Co-referências

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Entidades

1. {*Victoria Chen, her, the 38-year-old, She*}
2. {*Megabucks Banking, the company, Megabucks*}
3. {*her pay*}
4. {*Lotsabucks*}

Duas tarefas

1. Detectar as menções

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

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

Detecção de Menção

- 3 tipos
  - Pronomes: eu, seu etc
  - Entidades nomeadas: pessoas, lugares, empresas etc
  - Noun phrases: “o cachorro”, “a casa amarela da esquina”

Detecção de Menção

- Para detecção usar modelos de PLN
  - Pronomes: POS tagger
  - Entidades nomeadas: NER
  - Noun phrases: usar um parser

Desafio para a Detecção

- Nem todos pronomes, entidades e NPs são anáforas
- Ex:
  - It is sunny
  - Every student

Exemplo

Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Candidatos

Victoria Chen	$2.3 million	she
CFO of Megabucks Banking	the 38-year-old	Megabucks
Megabucks Banking	the company	Lotsabucks
her	the company's president	
her pay	It	

Algoritmo Baseado em Regra (Lee et al., 2013)

1. Take all NPs, possessive pronouns, and named entities.
2. Remove numeric quantities (100 dollars, 8%), mentions embedded in larger mentions, adjectival forms of nations, and stop words (like *there*).
3. Remove pleonastic *it* based on regular expression patterns.

Modelo de Co-referência: Mention Pair

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

I	Nader	he	my	she

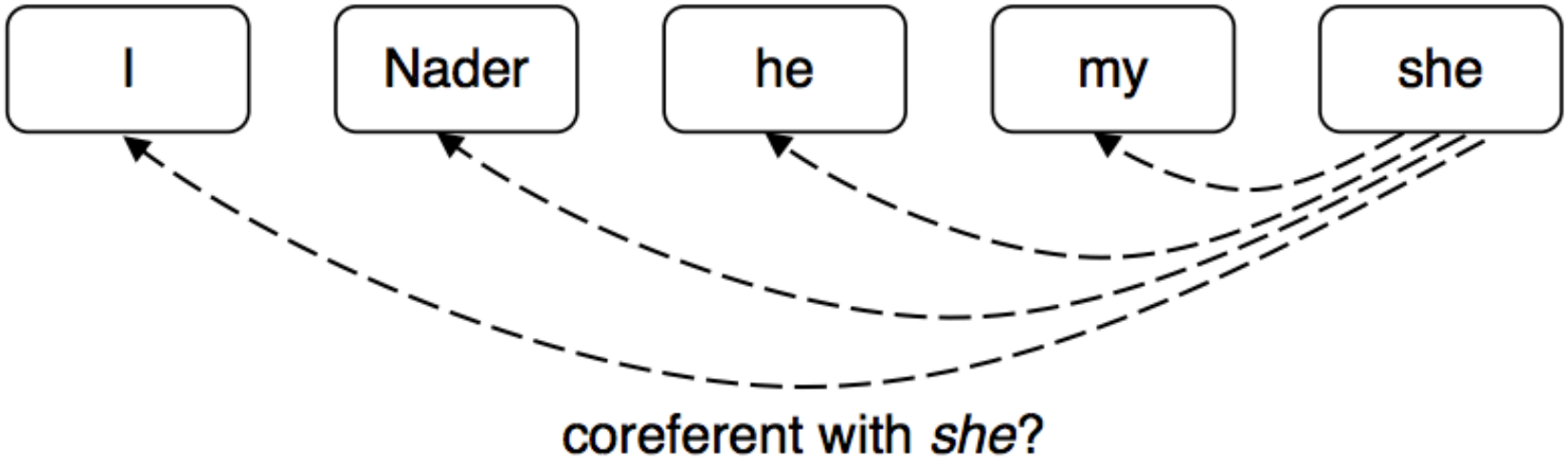
Coreference Cluster 1

Coreference Cluster 2

Modelo de Co-referência: Mention Pair

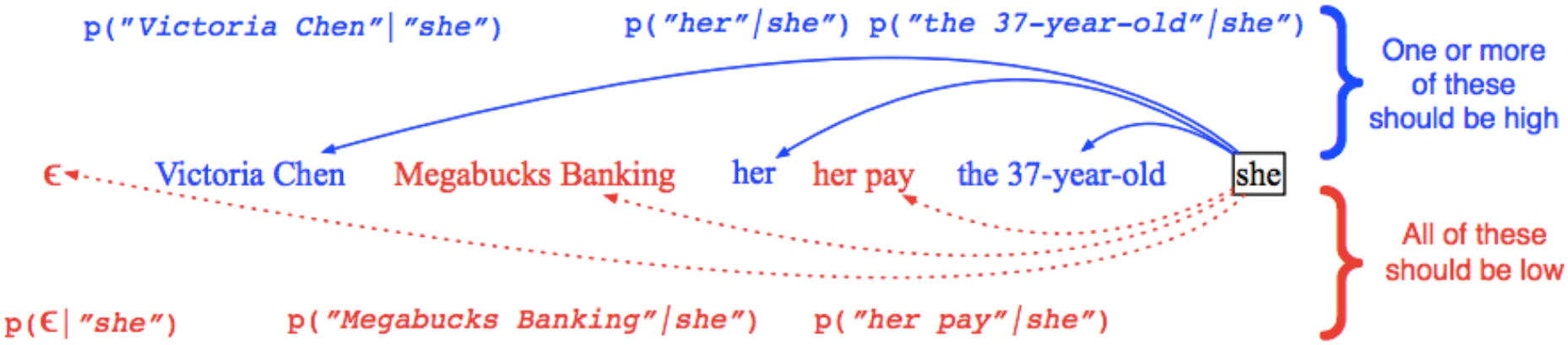
- Dado um par de menções, predizer a probabilidade de ser uma co-referência

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



Modelo de Co-referência: Mention Pair

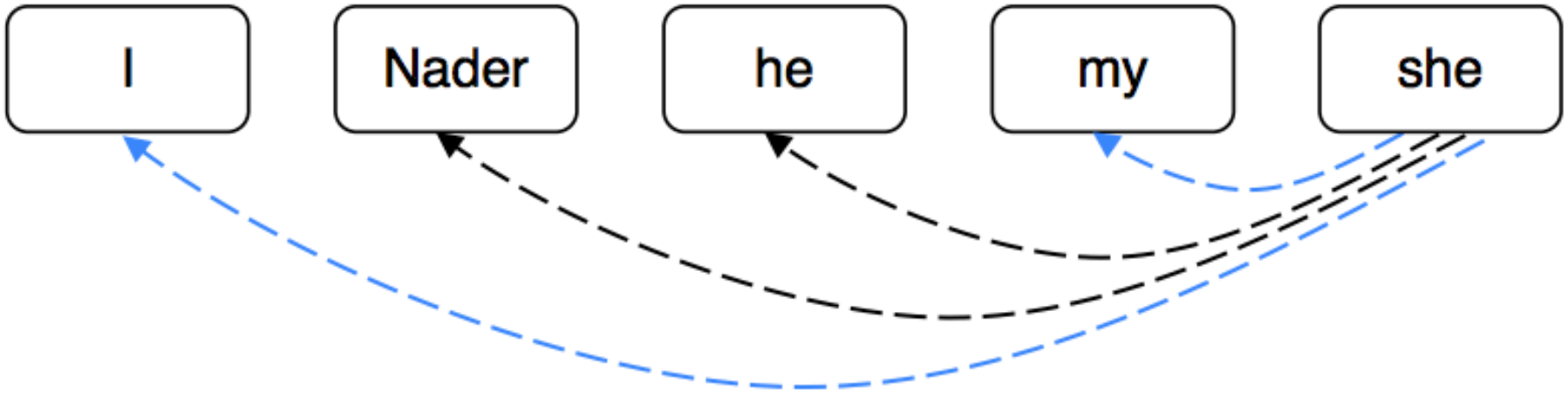
- Dado um par de menções, predizer a probabilidade de ser uma co-referência



Modelo de Co-referência: Mention Pair

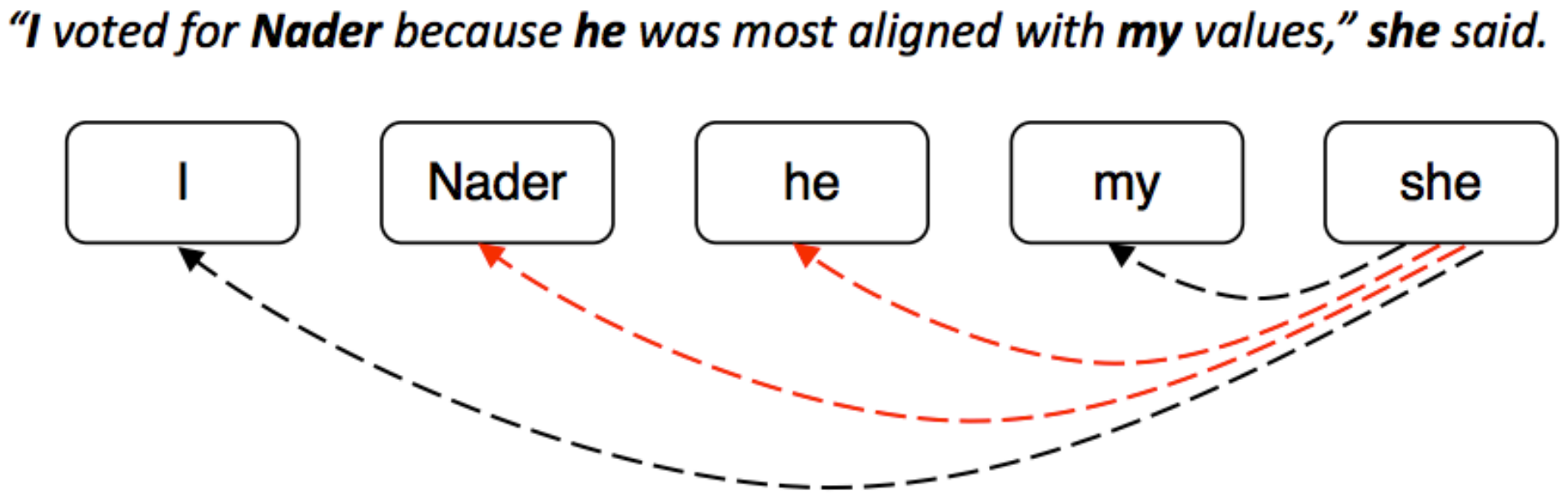
- Exemplos positivos

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



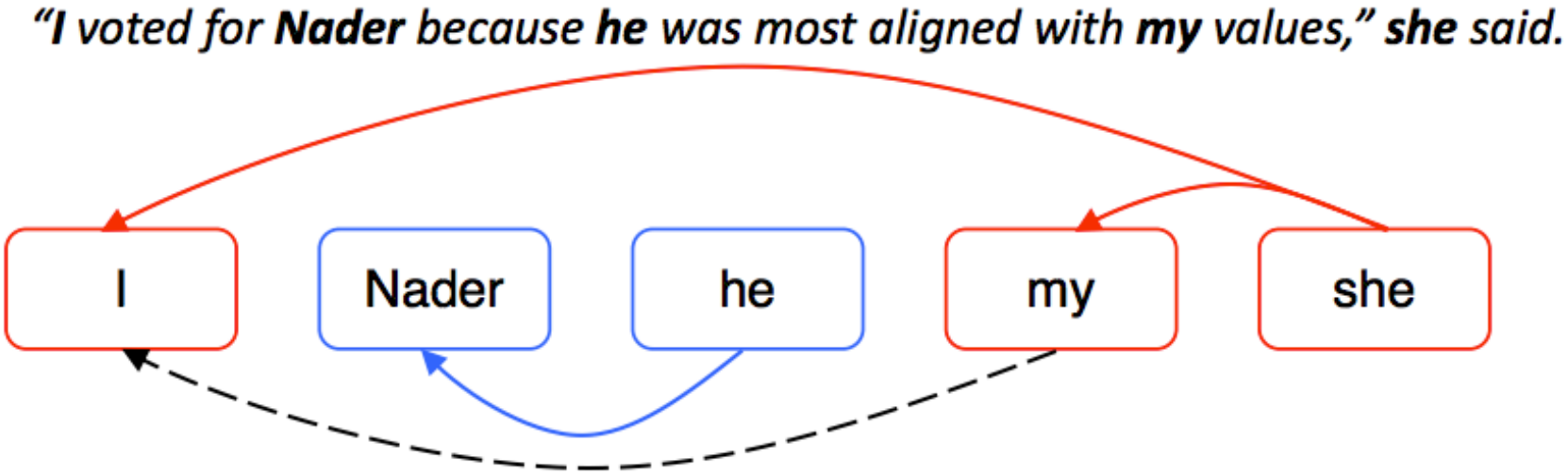
Modelo de Co-referência: Mention Pair

- Exemplos negativos



Mention Pair: Como Construir os Grupos

- Para cada menção *i*, considera a menção i-1, i-2 até 1 (direita pra esquerda)
- Closest-first: primeiro antecendente com probabilidade maior que 0.5
- Best-first: o antecedente com maior probabilidade
- Cria grupos por transitividade



Modelo de Co-referência: Mention Pair

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

I	Nader	he	my	she

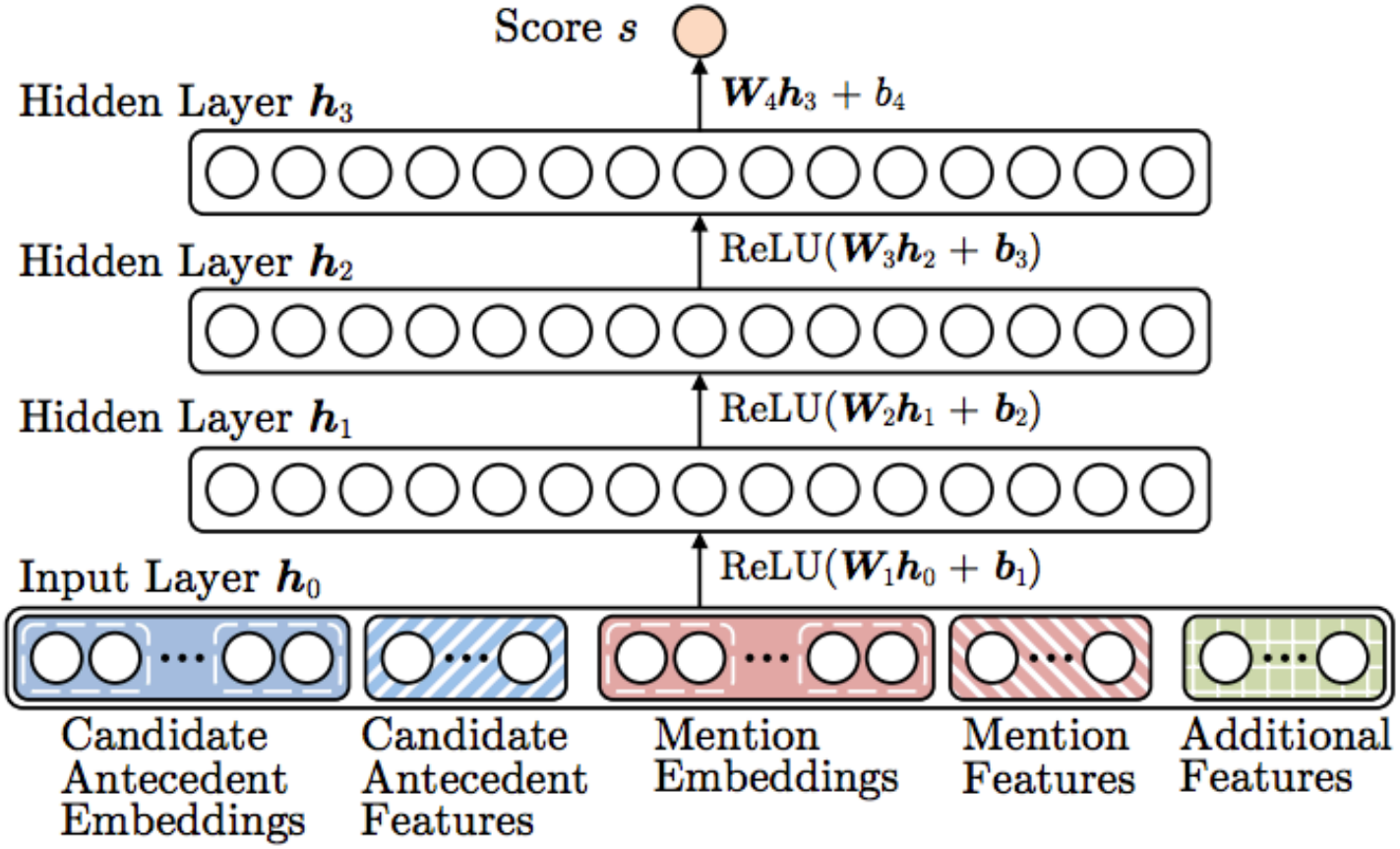
Coreference Cluster 1

Coreference Cluster 2

Features

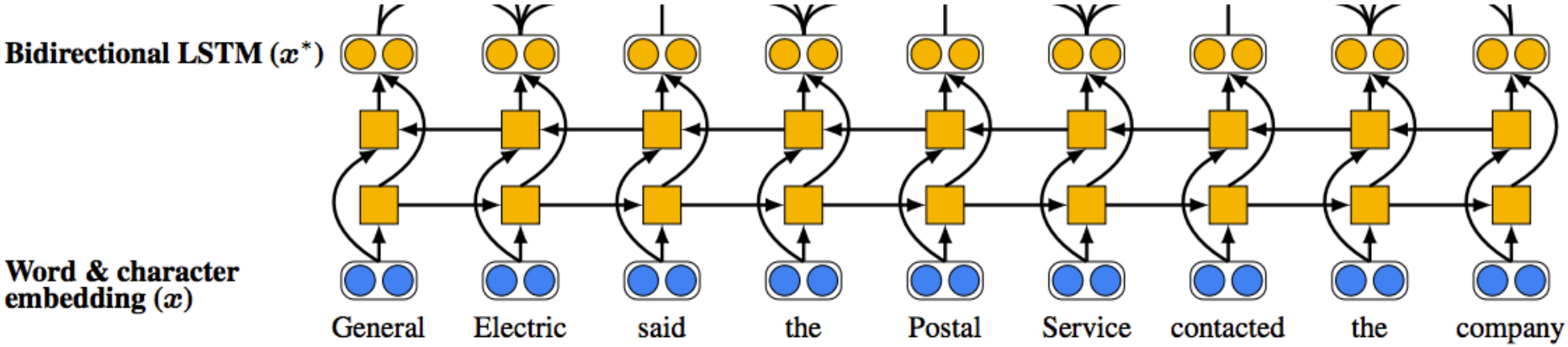
- Concordância de gênero
  - Jack gave Mary a gift. She was excited.
- Compatibilidade semântica
  - … the mining conglomerate … the company …
- Entidade mencionada mais recente
  - John went to a movie. Jack went as well. He was not busy.
- Grupo gramatical: prefirir entidades no sujeito da oração
  - John went to a movie with Jack. He was not busy.

Neural Coref Model: MLP



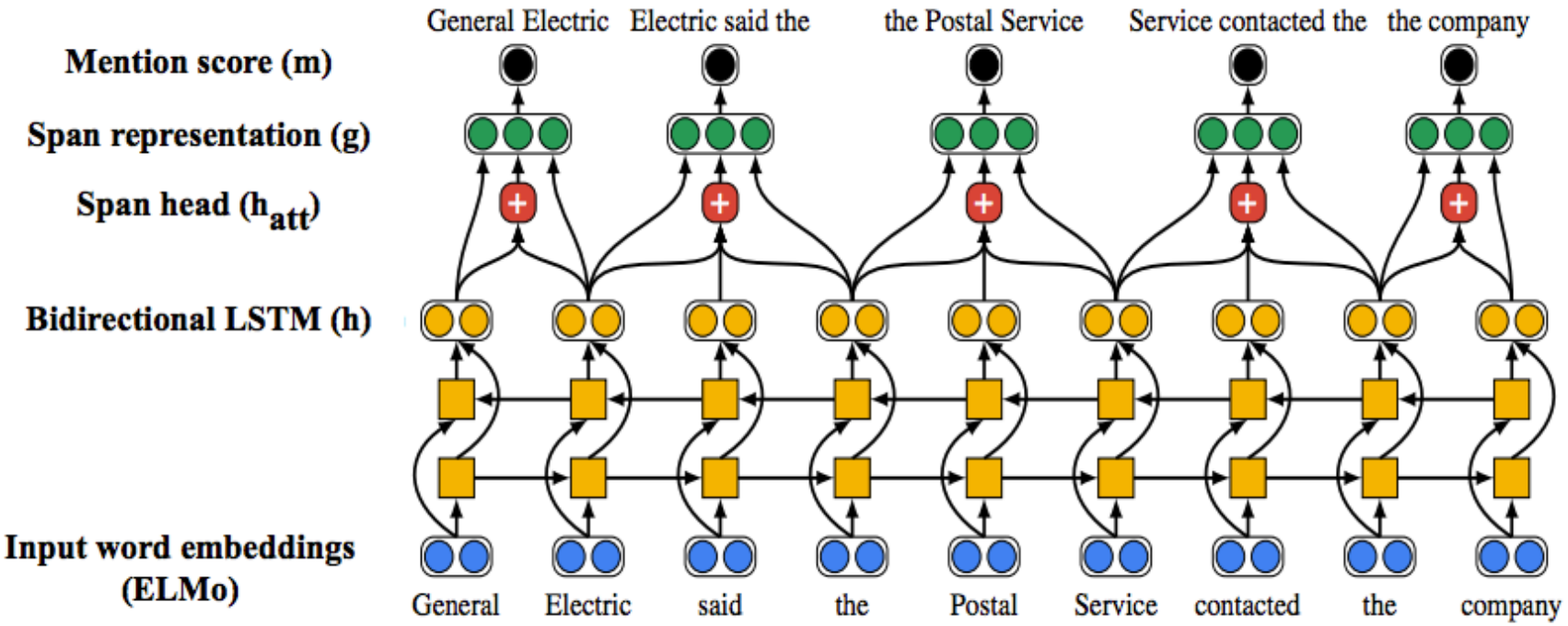
End-to-End

- Detecção de menção e co-referência ao mesmo tempo
- Entrada: embeddings de caracteres e palavras
- Executa BLSTM
- Fetures adicionais: distância, tópico do documento etc



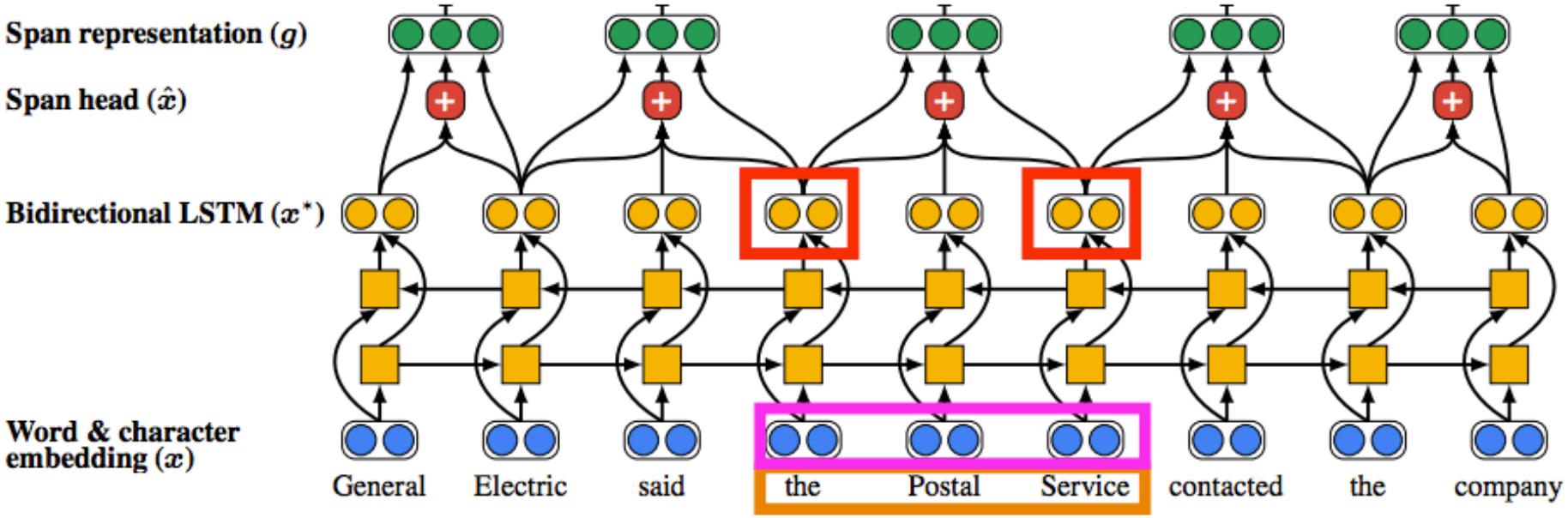
End-to-End

- Cada span do texto é representado por um vetor



End-to-End

- Cada span do texto é representado por um vetor
- Para “the Postal Sevice”



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

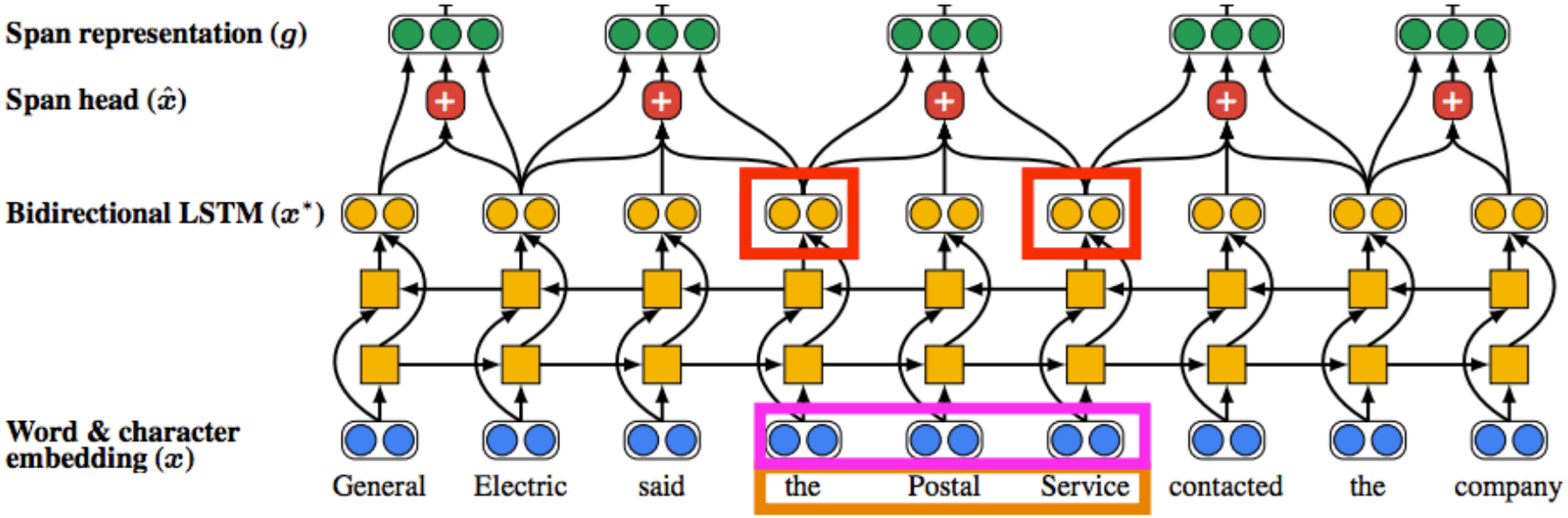
hidden states for span’s start and end

Attention-based representation

Additional features

End-to-End

- $\hat{\boldsymbol{x}}_i$ é o vetor médio usando attention dos word embeddings no span



Attention scores

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

dot product of weight vector and transformed hidden state

Attention distribution

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

just a softmax over attention scores for the span

Final representation

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

Attention-weighted sum of word embeddings

Computando o Score

- Para cada par de spans

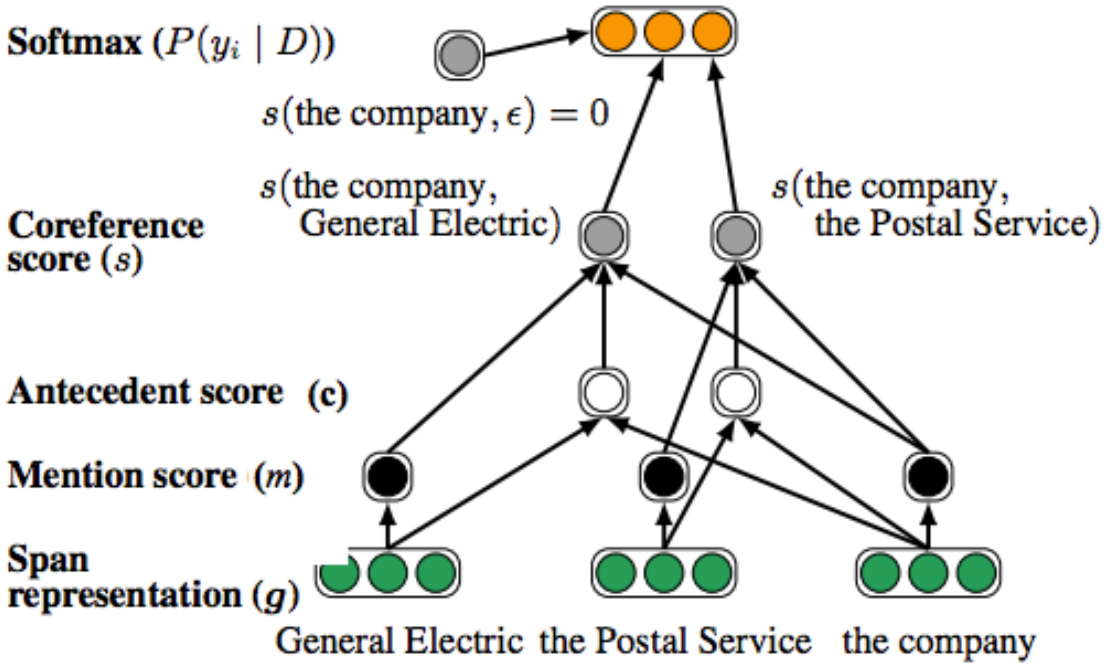
$$s(i,j) = m(i) + m(j) + c(i,j)$$

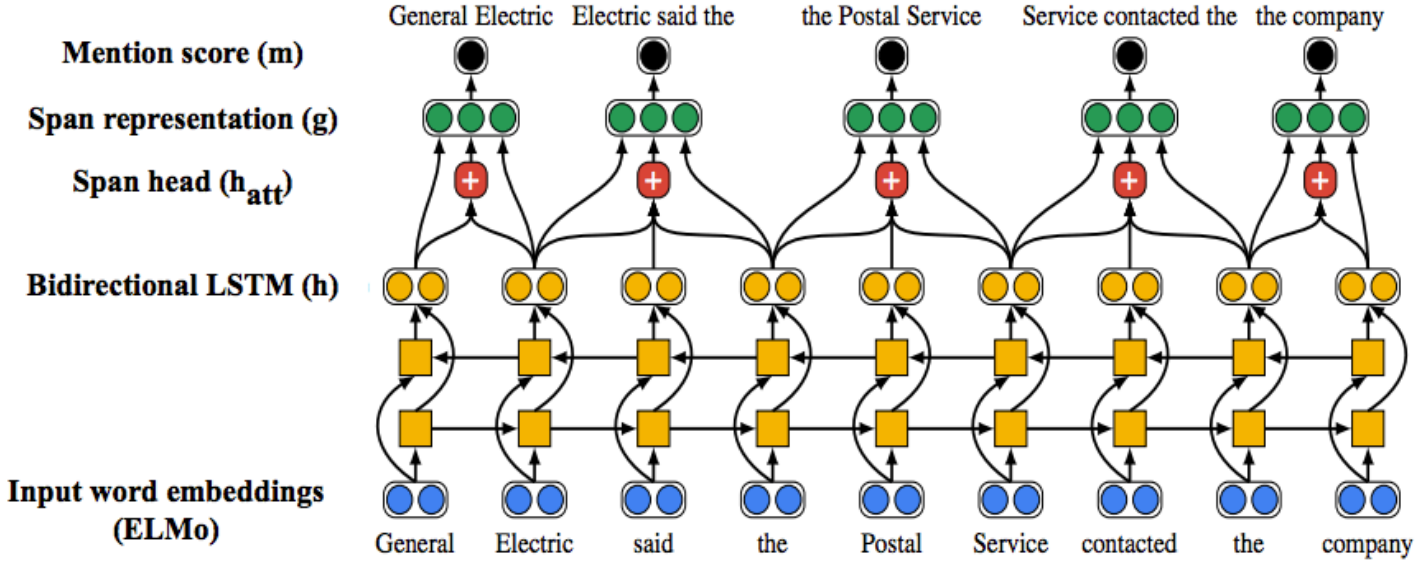
score de menção

score de "co-referência"

$$m(i) = w_m \cdot \text{FFNN}_m(\mathbf{g}_i)$$
$$c(i,j) = w_c \cdot \text{FFNN}_c([\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_i \circ \mathbf{g}_j, \phi(i,j)])$$

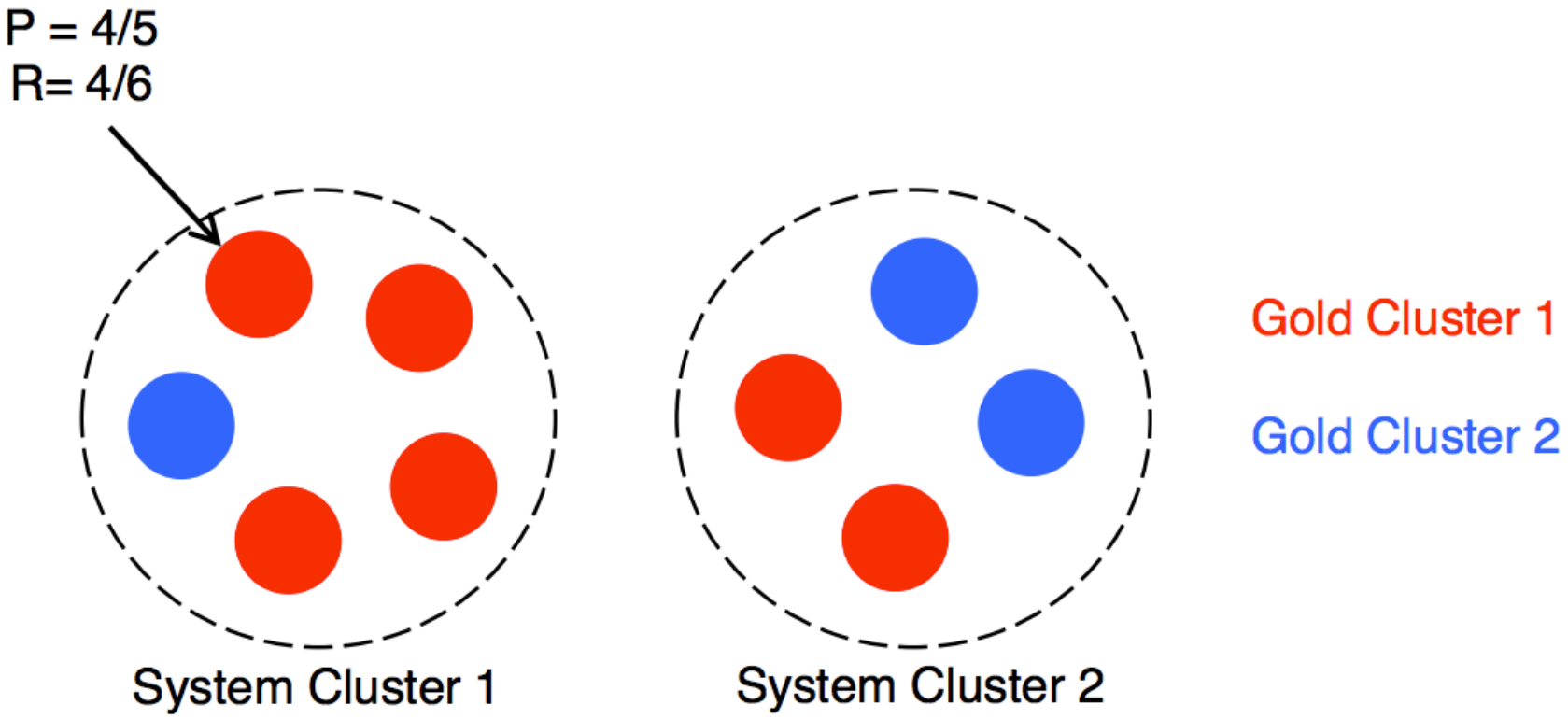
element-wise multiplication


Softmax ($P(y_i \mid D)$)
$s(\text{the company}, \epsilon) = 0$
Coreference score (s)
$s(\text{the company}, \text{General Electric})$
$s(\text{the company}, \text{the Postal Service})$
Antecedent score (c)
Mention score (m)
Span representation (g)
General Electric
the Postal Service
the company


General Electric
Electric said the
the Postal Service
Service contacted the
the company
Mention score (m)
Span representation (g)
Span head (h_{att})
Bidirectional LSTM (h)
Input word embeddings (ELMo)
General
Electric
said
the
Postal
Service
contacted
the
company

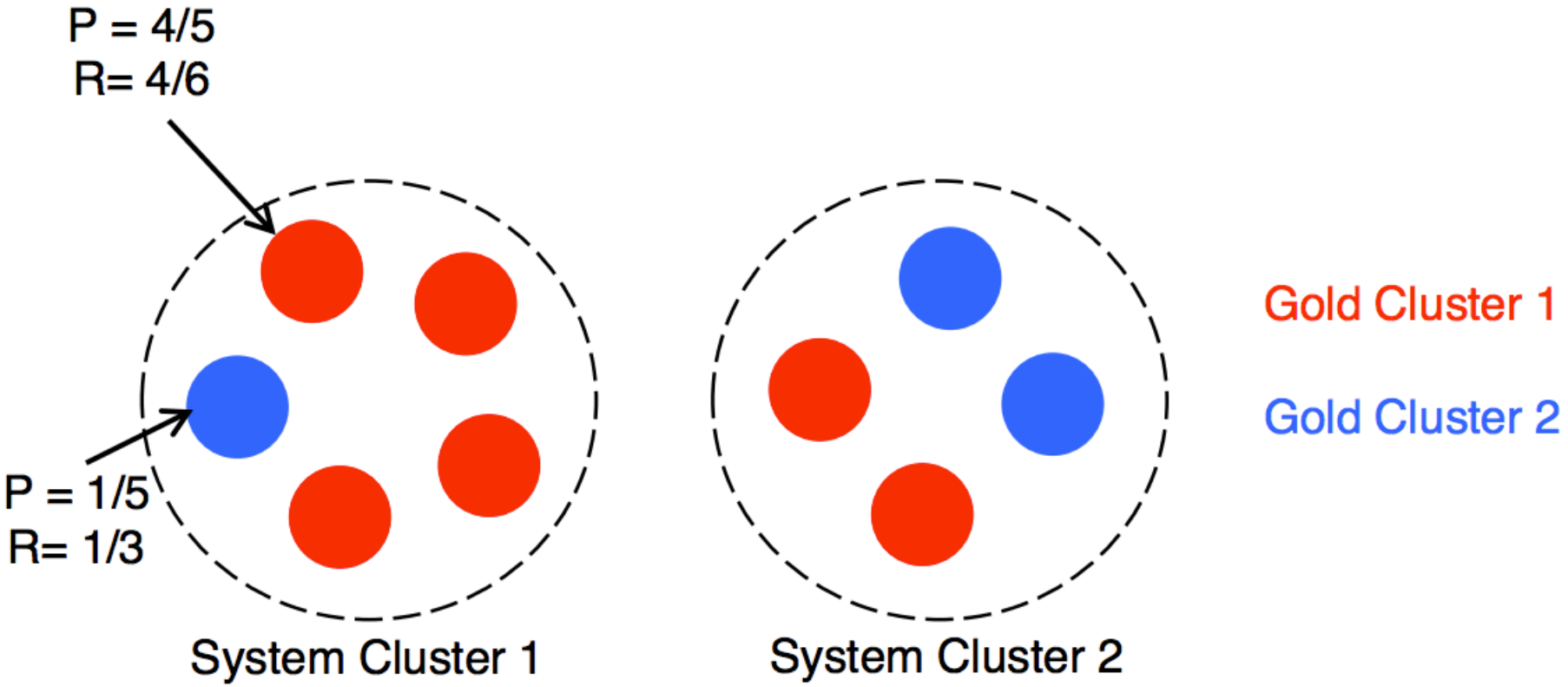
Avaliação de Co-referência

- B-CUBED
  1. Para cada menção, computa precision e recall



Avaliação de Coreference

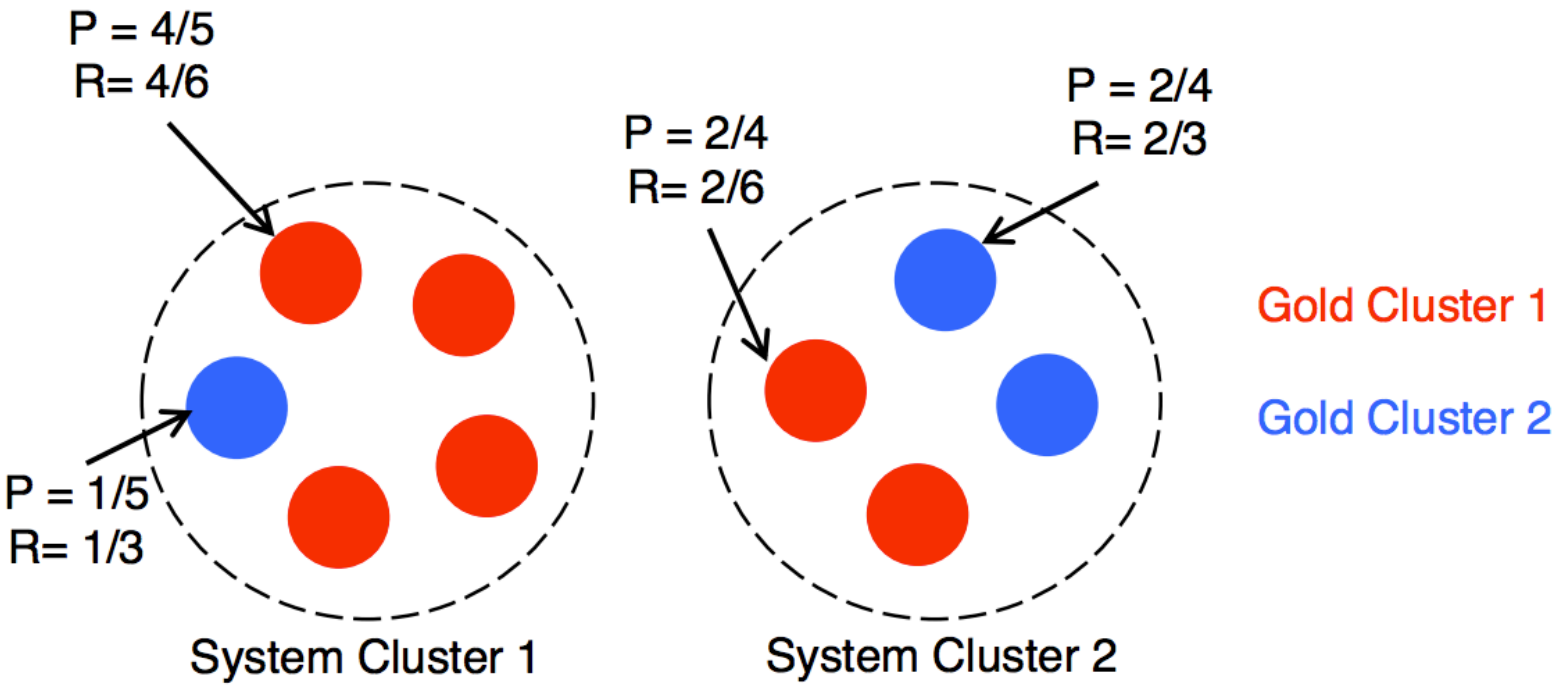
- B-CUBED
  1. Para cada menção, computa precision e recall



Avaliação de Coreference

- B-CUBED
  1. Para cada menção, computa precision e recall
  2. Calcula a média individual de precision e recall

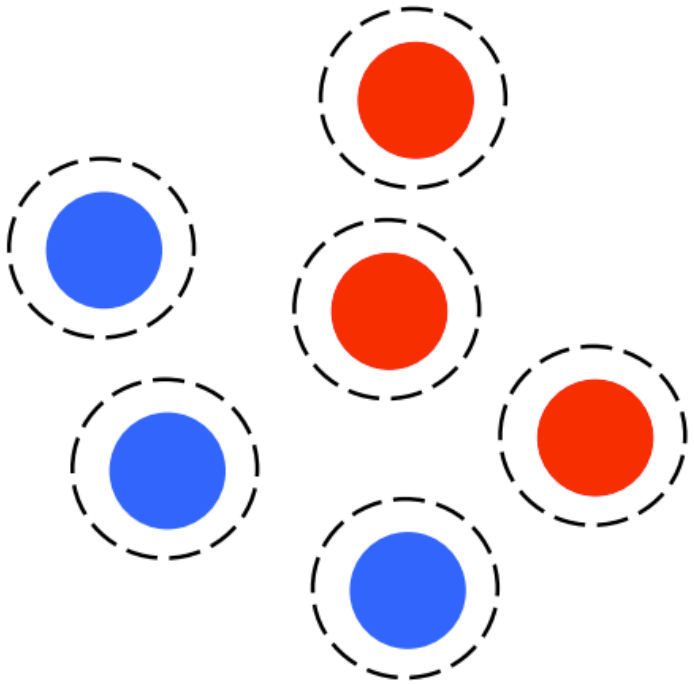
$$P = [4(4/5) + 1(1/5) + 2(2/4) + 2(2/4)] / 9 = 0.6$$



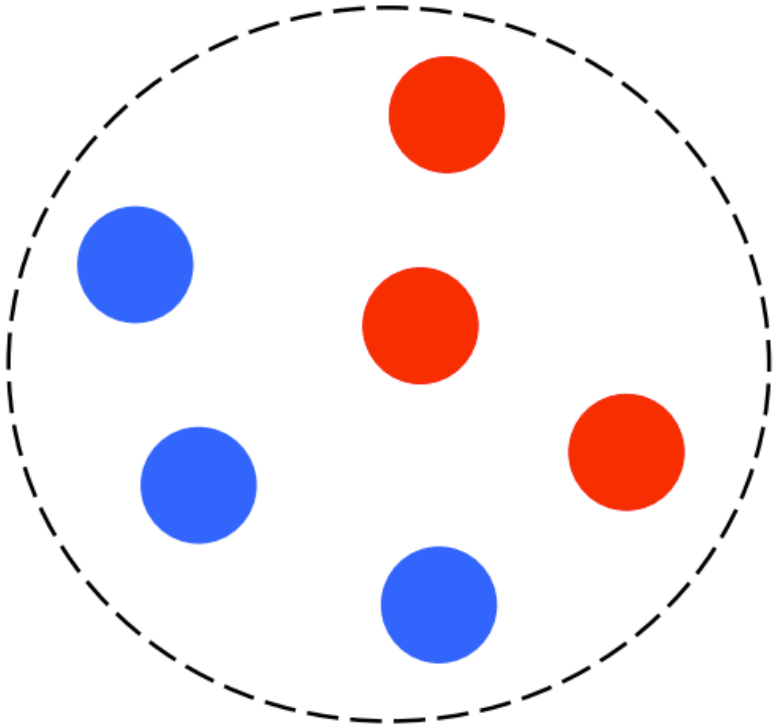
Avaliação de Coreference

- B-CUBED

100% Precision, 33% Recall



50% Precision, 100% Recall,



Resultados de Sistemas

- Dataset:
  - OntoNotes
  - Aprox. 3000 docs
  - Inglês e Chinês
- F1 médio de 3 métricas de correferência

Resultados de Sistemas

Model	English	Chinese	
Lee et al. (2010)	~55	~50	Rule-based system, used to be state-of-the-art!
Chen & Ng (2012) [CoNLL 2012 Chinese winner]	54.5	57.6	Non-neural machine learning models
Fernandes (2012) [CoNLL 2012 English winner]	60.7	51.6	
Wiseman et al. (2015)	63.3	—	Neural mention ranker
Clark & Manning (2016)	65.4	63.7	Neural clustering model
Lee et al. (2017)	67.2	--	End-to-end neural mention ranker

Onde Redes Neurais Ajudam

- Quando não há matching exato de string com NPs e entidades nomeadas

Example Wins

Anaphor	Antecedent
the country's leftist rebels	the guerillas
the company	the New York firm
216 sailors from the ``USS cole''	the crew
the gun	the rifle