

Resolução de Co-Referência

Luciano Barbosa

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

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

Spanish English French Detect language ▾

↔

English Spanish Arabic ▾

Translate

A Alicia le gusta Juan porque es inteligente ×

44/5000

Alicia likes Juan because he's smart

☆ 📄 🔊 ➦

✎ Suggest an edit

Spanish English French Detect language ▾

↔

English Spanish Arabic ▾

Translate

A Juan le gusta Alicia porque es inteligente ×

44/5000

Juan likes Alicia because he's smart

☆ 📄 🔊 ➦

✎ Suggest an edit

Componentes

referente



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.

anáfora: refere-se a um
termo antecedente



singleton

Anáfora vs Correferência

Barack Obama traveled to ... Obama

Barack Obama

Obama

Co-referência



Barack Obama said he would sign the bill.
antecedent anaphor

Barack Obama

he

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

No student

The best donut in the world

100 miles

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.

Mention Pair

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

I

Nader

he

my

she

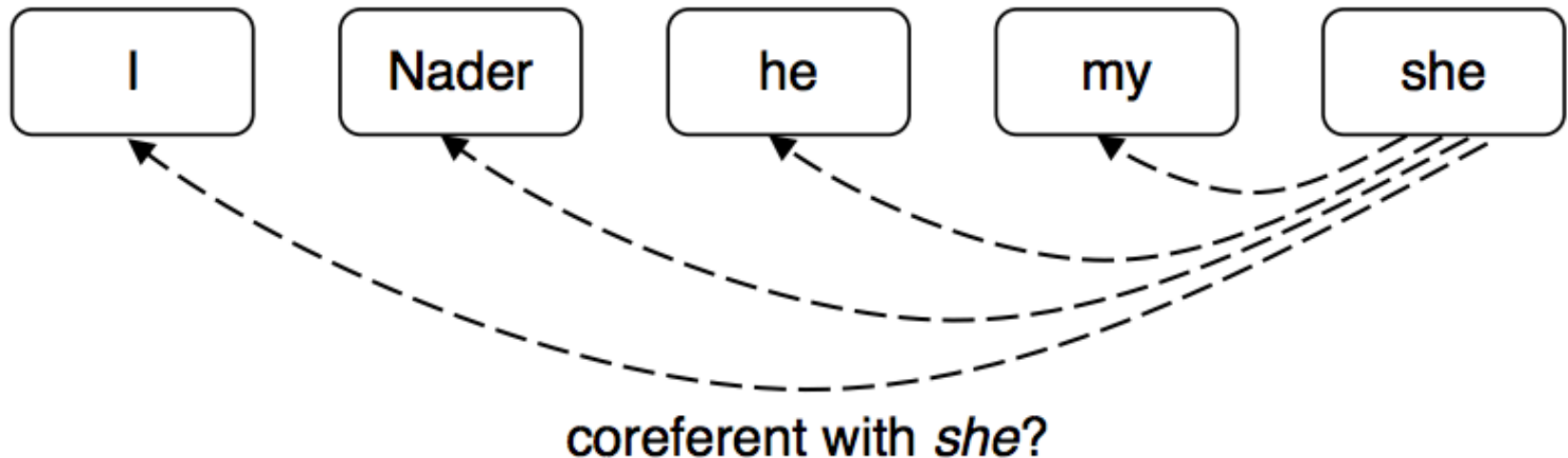
Coreference Cluster 1

Coreference Cluster 2

Mention Pair

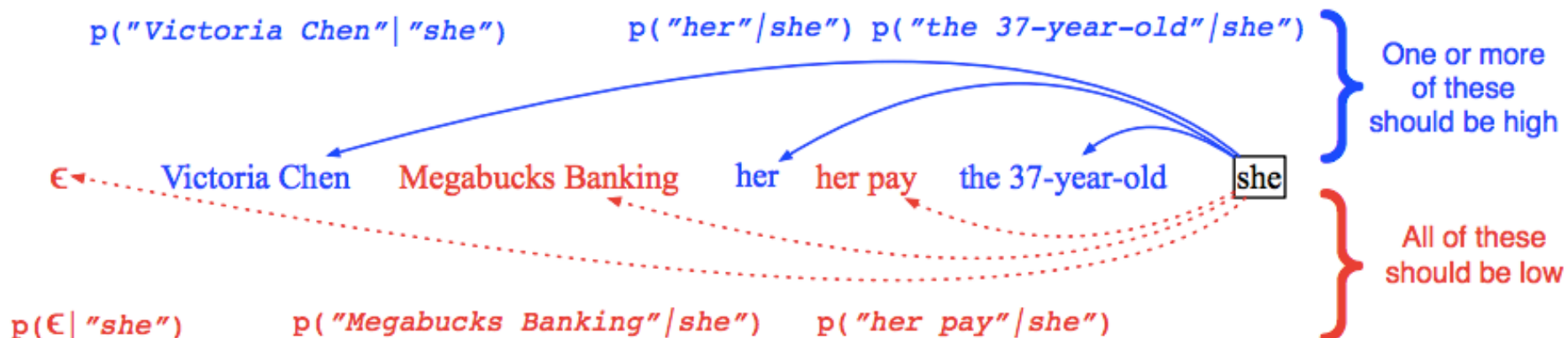
- Dado um par de menções, predizer a probabilidade de ser uma co-referência

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



Mention Pair

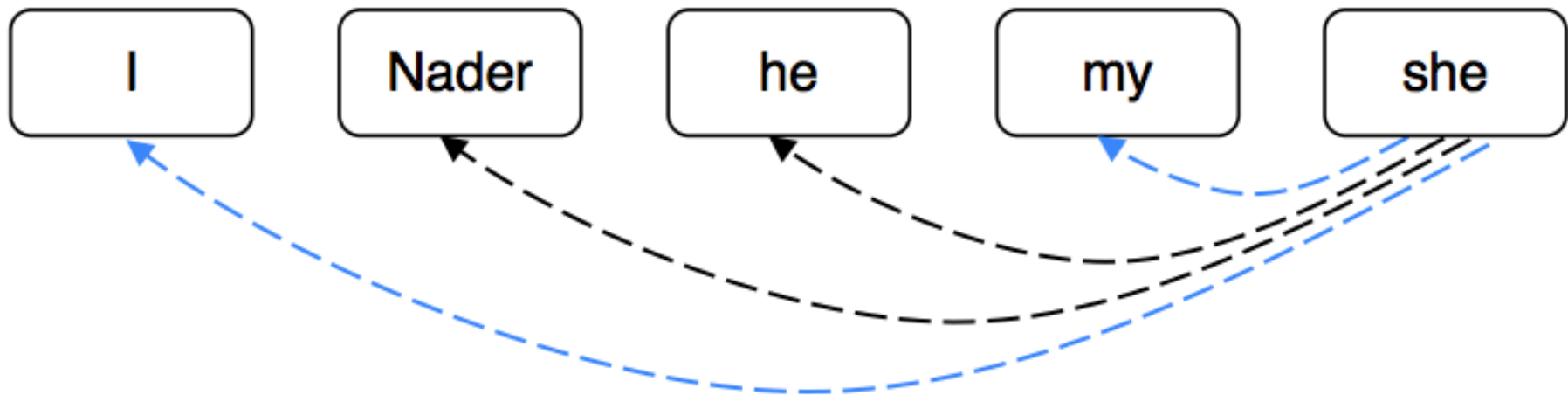
- Dado um par de menções, predizer a probabilidade de ser uma co-referência



Mention Pair

- Exemplos positivos

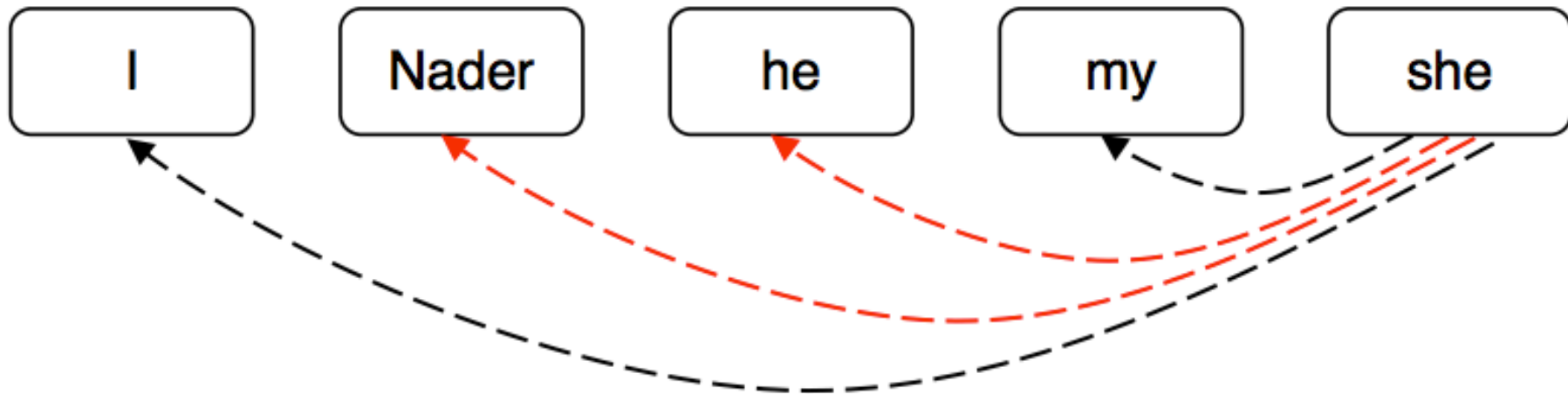
*"I voted for **Nader** because **he** was most aligned with **my** values," **she** said.*



Mention Pair

- Exemplos negativos

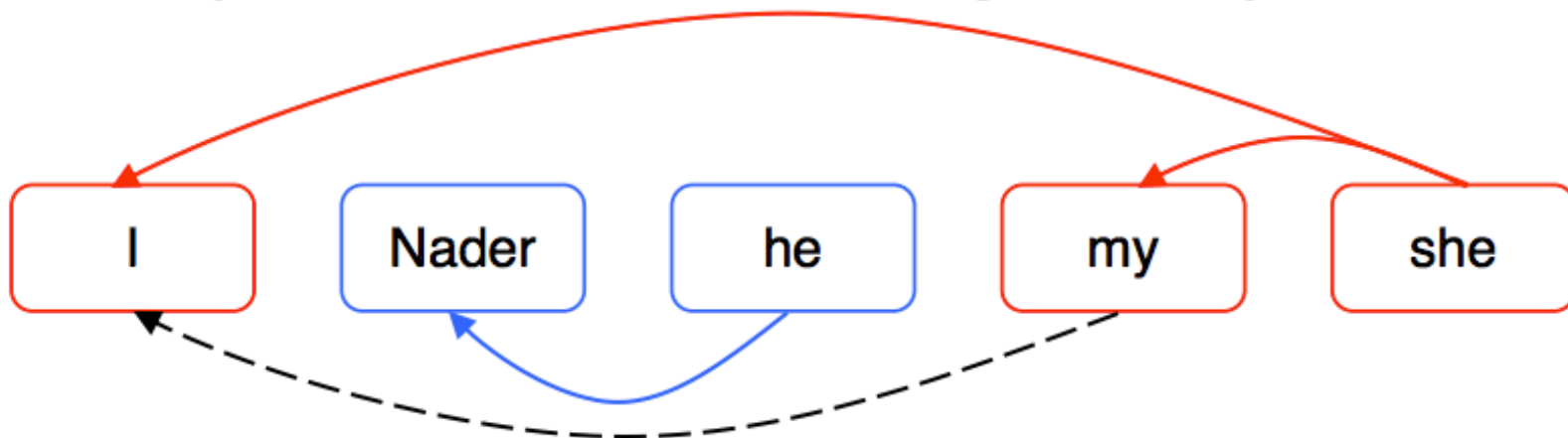
*“I voted for **Nader** because **he** was most aligned with **my** values,” **she** said.*



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 antecedente com probabilidade maior que 0.5
- Best-first: o antecedente com maior probabilidade
- Cria grupos por transitividade

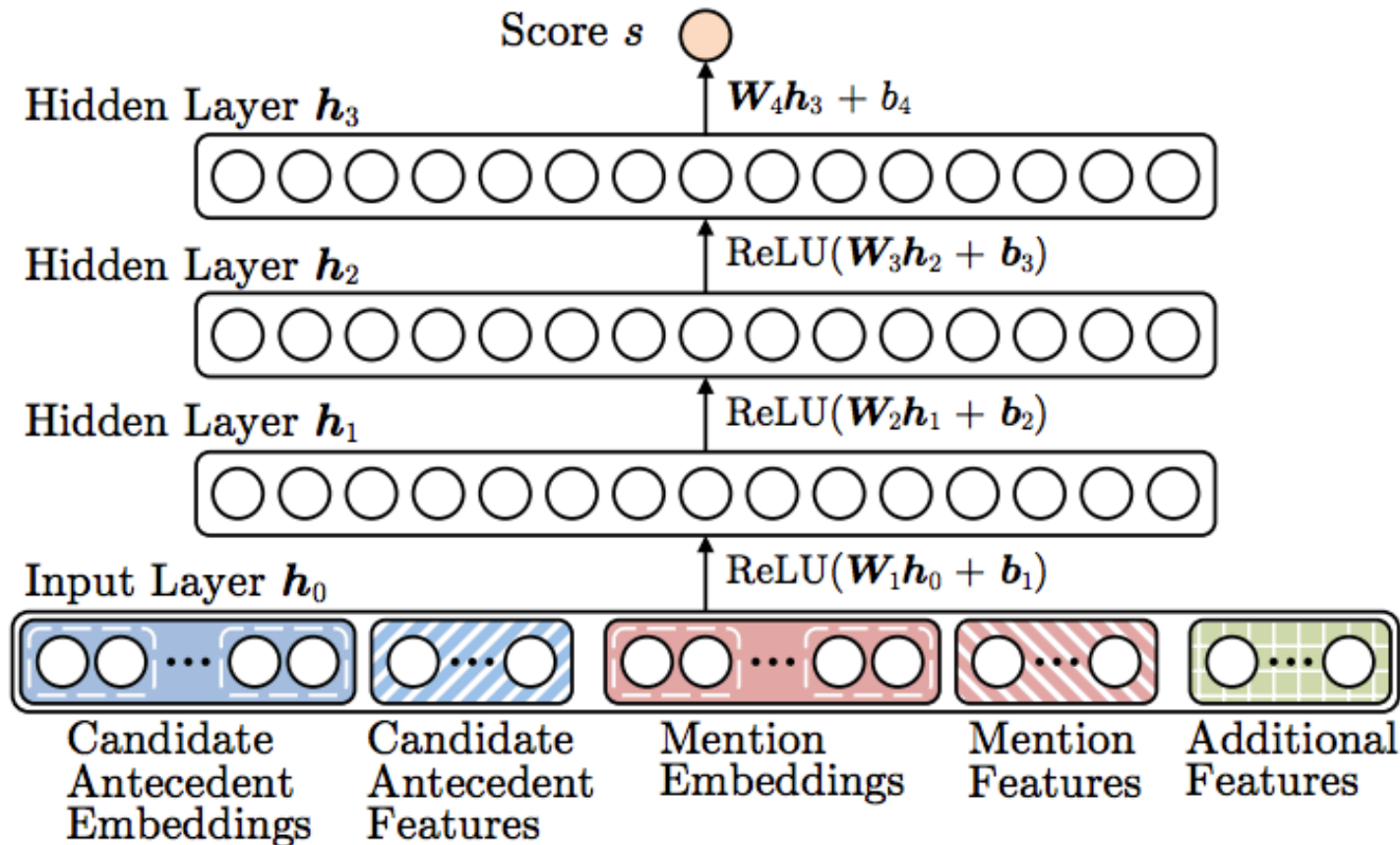
*"I voted for **Nader** because **he** was most aligned with **my** values," **she** said.*



Features

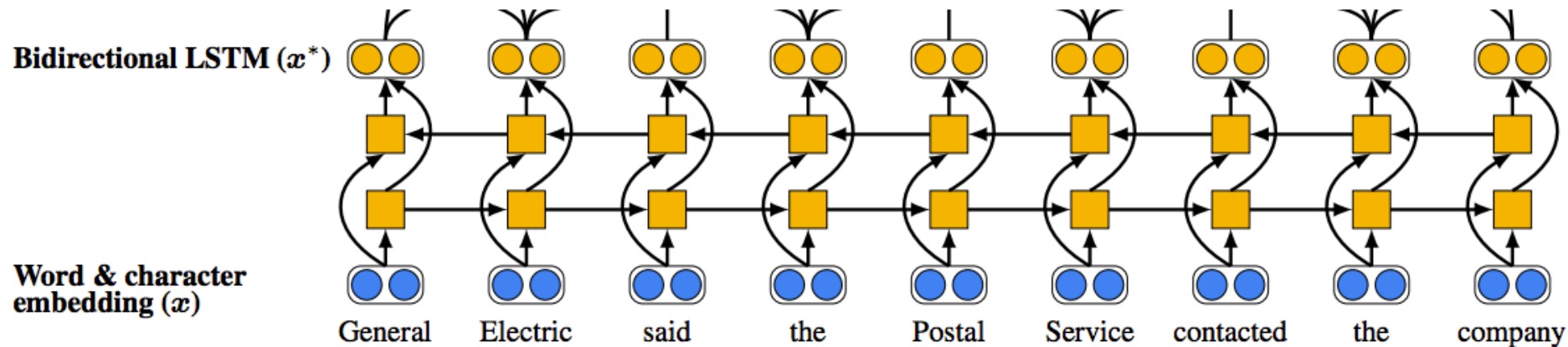
- Concordância de gênero
 - Jack gave Mary a gift. She was excited.
- Compatibilidade semântica
 - ... the mining conglomerate ... the company ...
- Entidade mencionada mais recente
 - John went to a movie. Jack went as well. He was not busy.
- Grupo gramatical: entidades no substantivo
 - John went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

Neural Coref Model: MLP



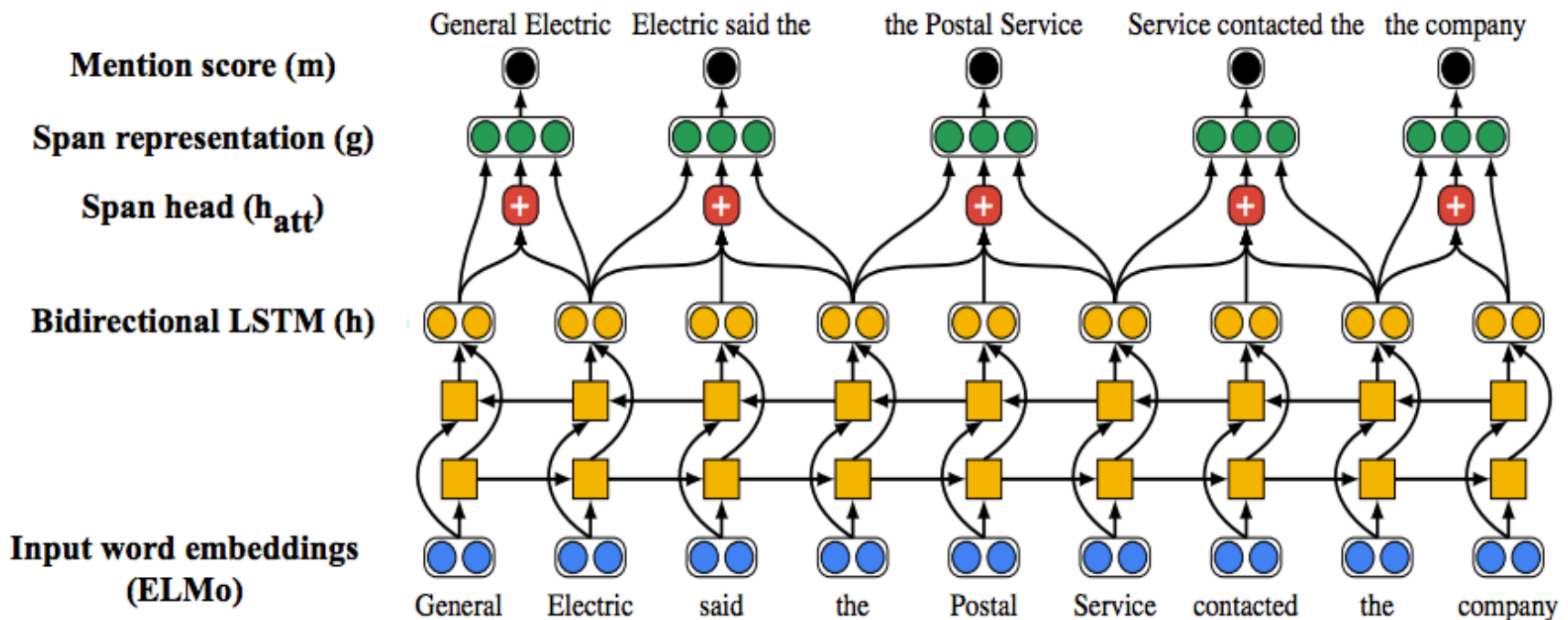
End-to-End

- Detecção de menção e co-referência ao mesmo tempo
- Entrada: embeddings de caracteres e palavras
- Executa BLSTM



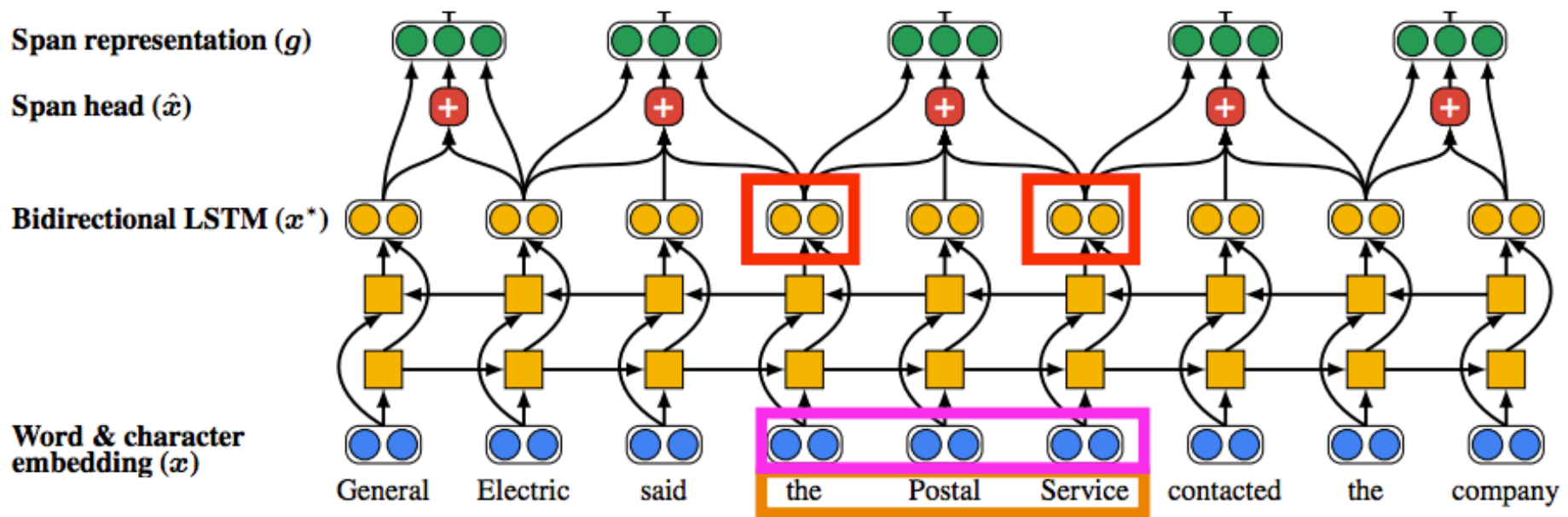
End-to-End

- Cada span do texto é representado por um vetor



End-to-End

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



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

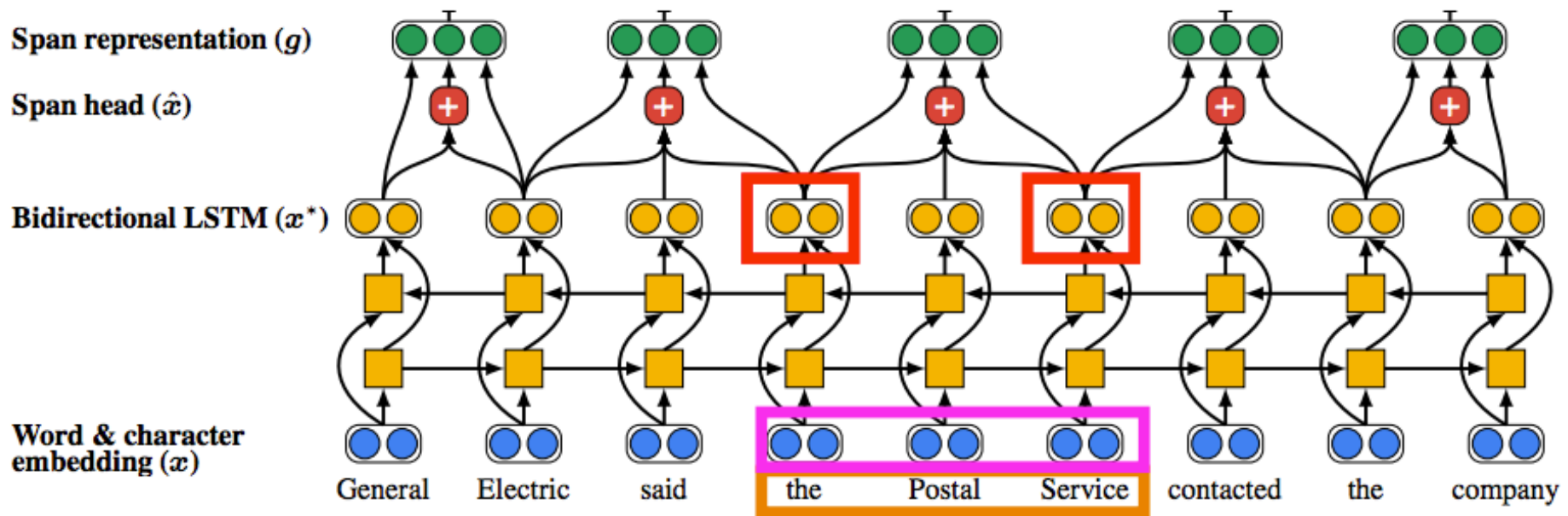
hidden states for
span's start and end

Attention-based representation

Additional features

End-to-End

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



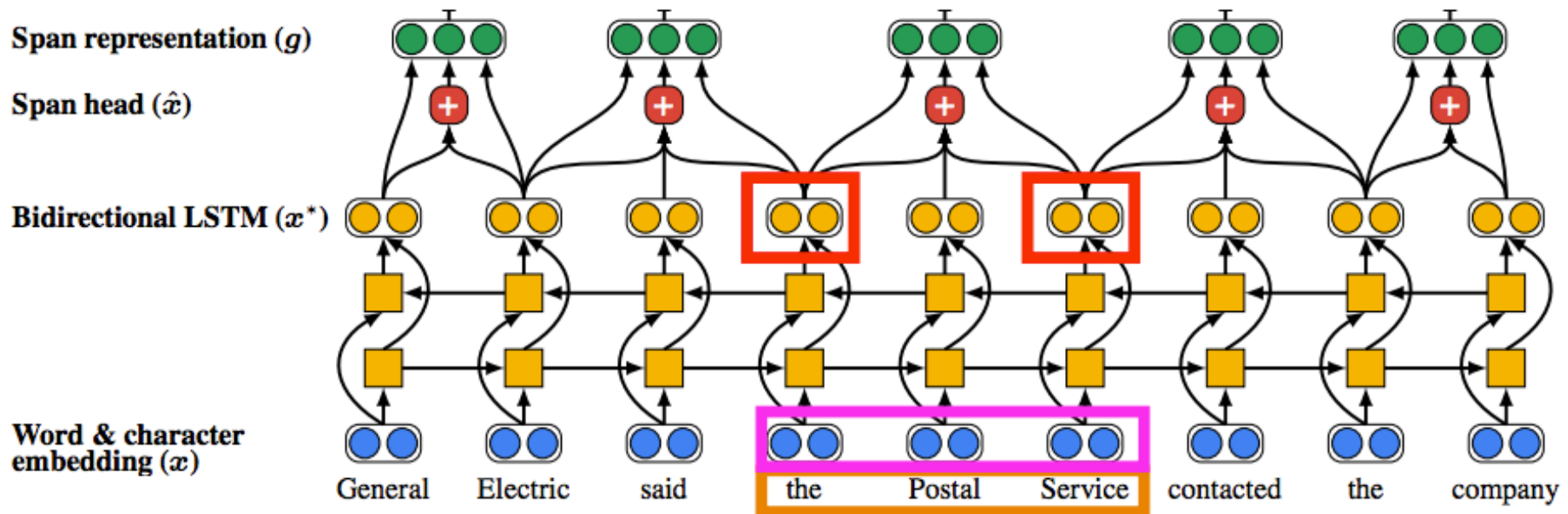
$$g_i = [x_{\text{START}(i)}^*, x_{\text{END}(i)}^*, \hat{x}_i, \phi(i)]$$

hidden states for
span's start and end

Attention-based representation

Additional features

End-to-End



Attention scores

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

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

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

Attention-weighted sum
of word embeddings

e.br

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”

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

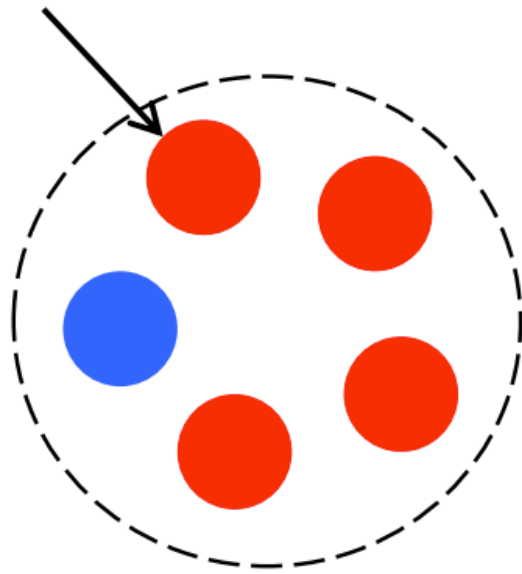
produto interno

Avaliação de Co-referência

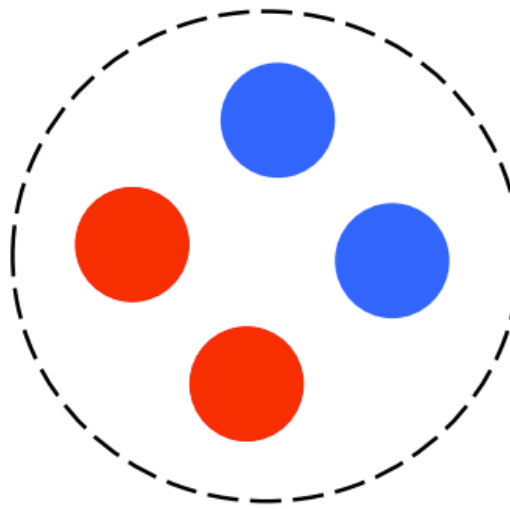
- B-CUBED
 1. Para cada menção, computa precision e recall

$$P = 4/5$$

$$R = 4/6$$



System Cluster 1



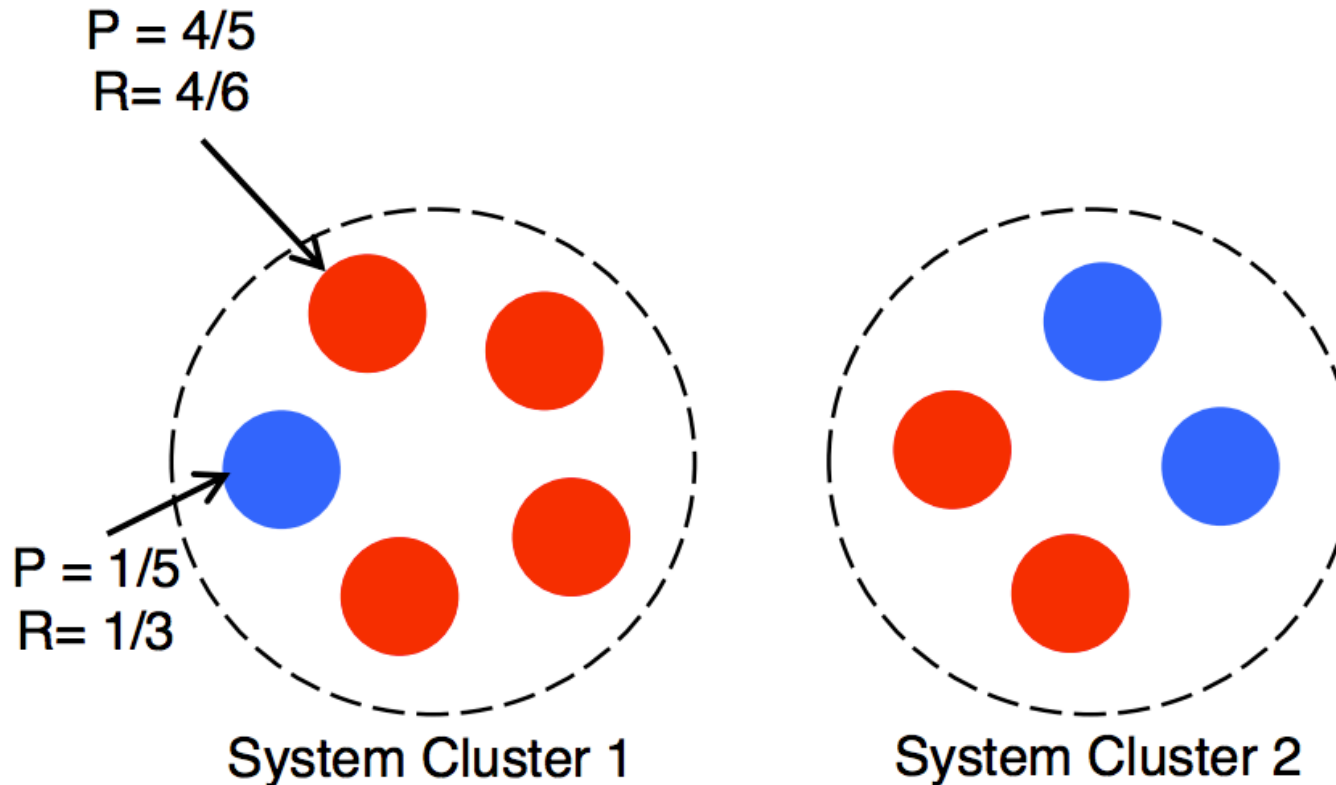
System Cluster 2

Gold Cluster 1

Gold Cluster 2

Avaliação de Coreference

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



Gold Cluster 1

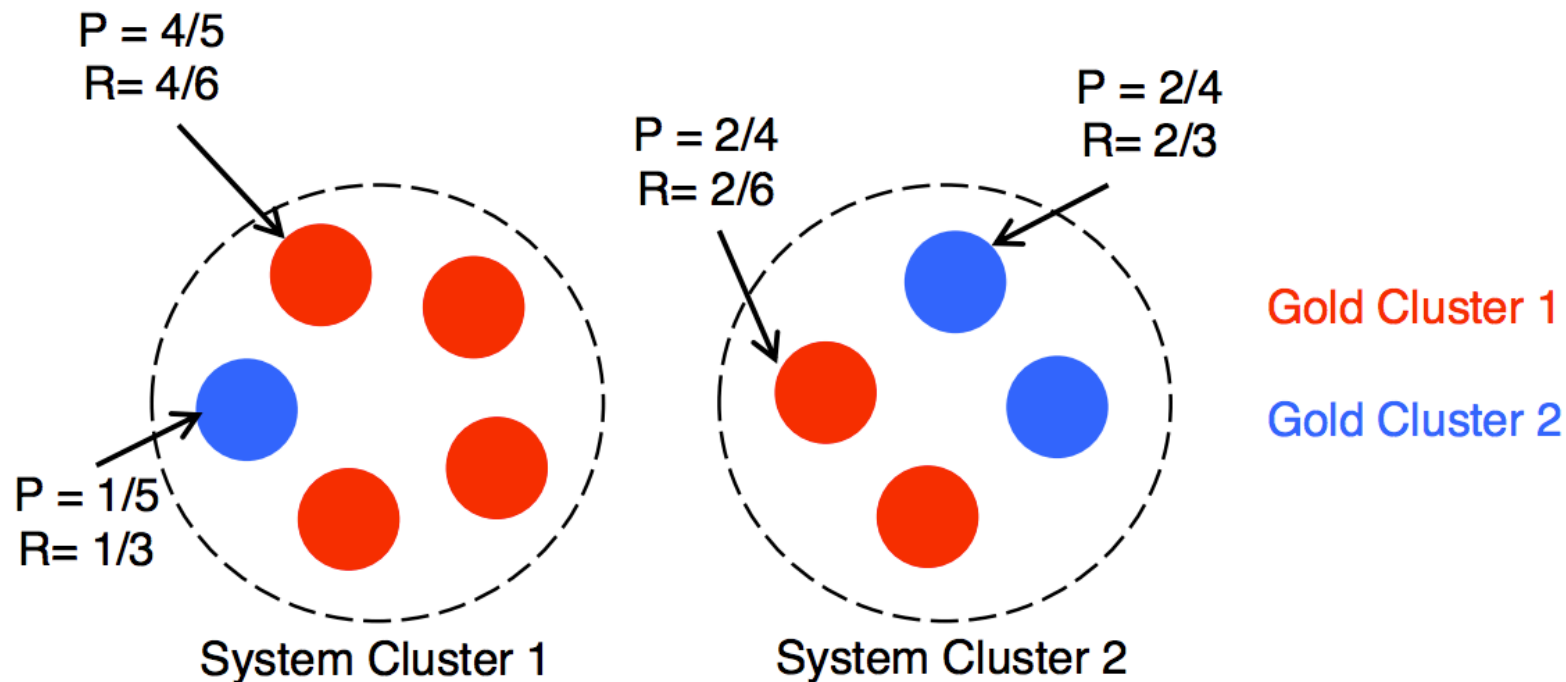
Gold Cluster 2

Avaliação de Coreference

- B-CUBED

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

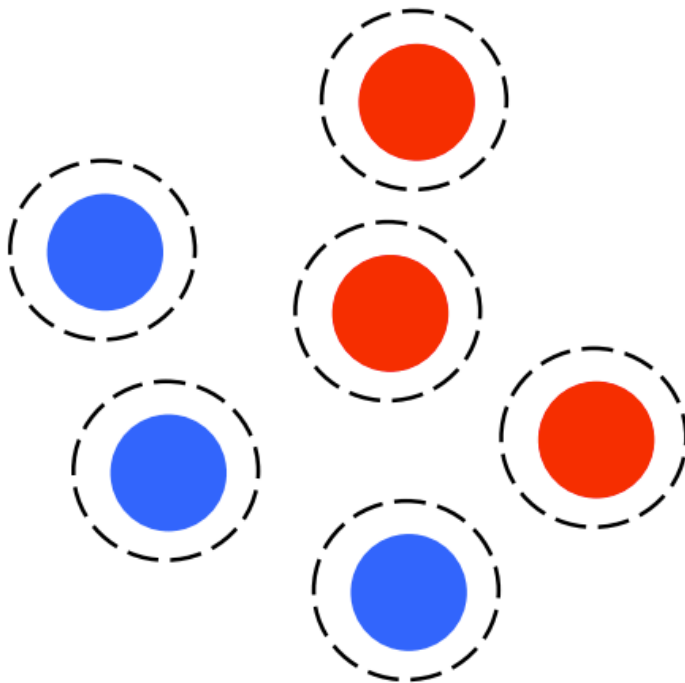
$$P = [4(4/5) + 1(1/5) + 2(2/4) + 2(2/4)] / 9 = 0.6$$



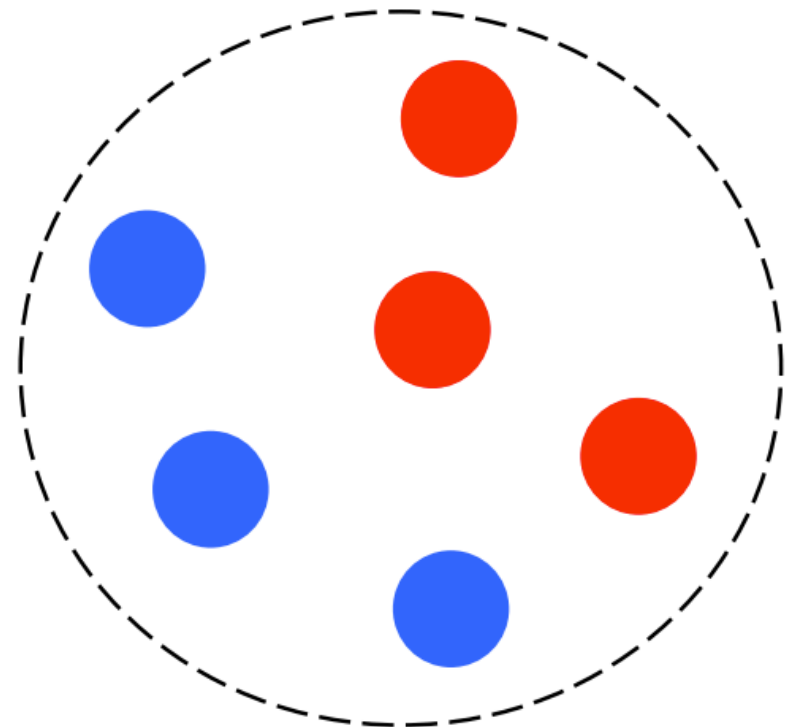
Avaliação de Coreference

- 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	
Fernandes (2012) [CoNLL 2012 English winner]	60.7	51.6	Non-neural machine learning models
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