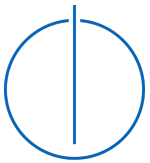


Faculty for Informatics

Technical
University
of Munich

Natural Language Processing
IN2361

PD Dr. Georg Groh

Social Computing
Research Group

Deep NLP
Part F: Coreference Resolution

- content is based on [2] (lecture 13)
- certain elements (e.g. figures, equations or tables) were taken over or taken over in a modified form from [2]
- citations of [2] are omitted for legibility
- errors on these slides are fully in the responsibility of Georg Groh
- BIG thanks to Richard Socher and his colleagues at Stanford for publishing materials [2] of a great Deep NLP lecture

Coreference Resolution

Identify all mentions that refer to the same real world entity

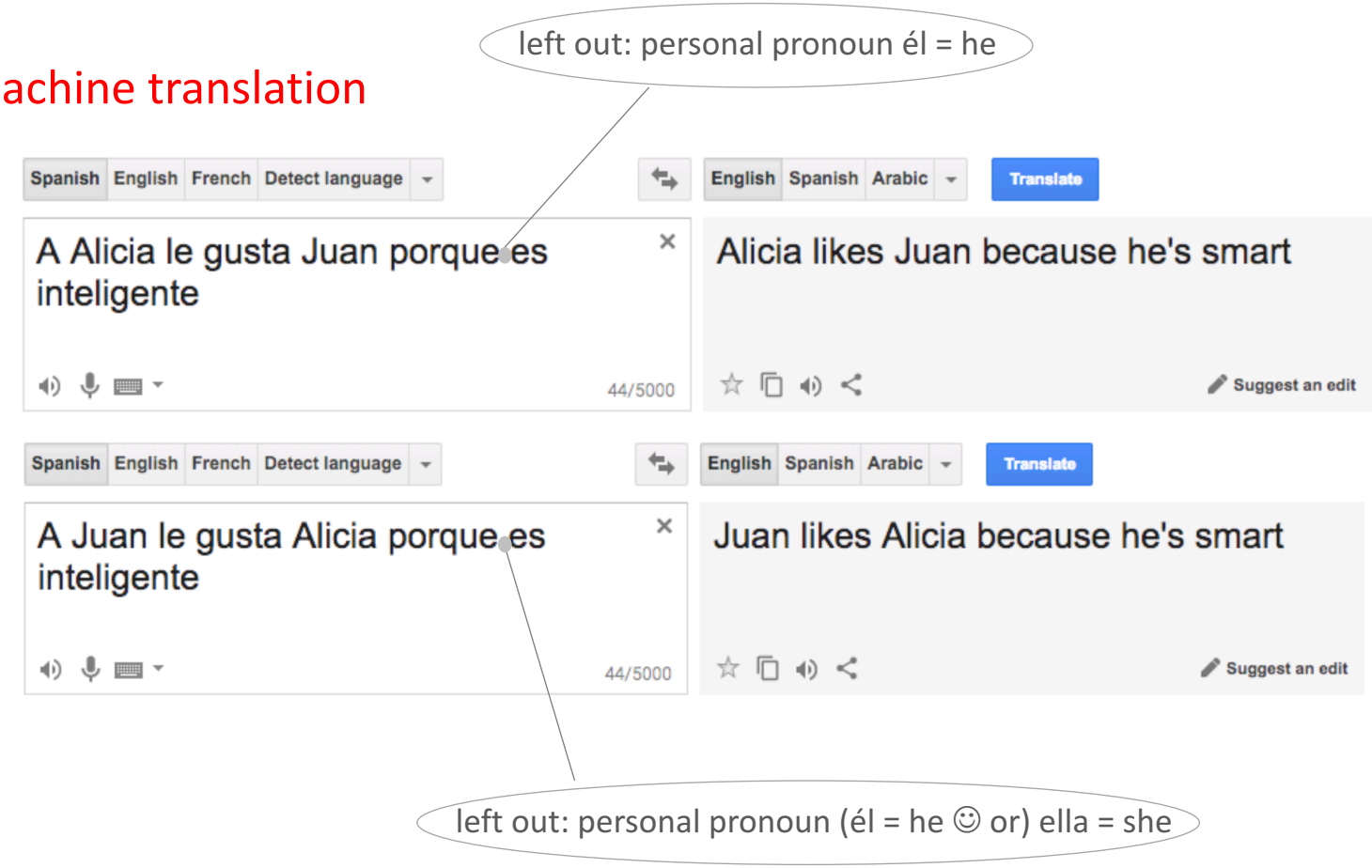
Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

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.

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.

Applications of Coreference Resolution

- full text understanding: information extraction, question answering, summarization, …
- machine translation



Applications of Coreference Resolution

- full text understanding: information extraction, question answering, summarization, …
- 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"

Full Coreference Resolution is AI Hard

"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 Schemata
- requires full scale, AI-hard reasoning to solve

Coreference Resolution: Two Steps

- mention detection (rather easy):

  [*I*] *went for a walk with* [*Hannah*]. [*She*] *also brought* [[*her*] *friend*] [*Helen*].

- mention clustering (rather hard):

  [*I*] *went for a walk with* [*Hannah*]. [*She*] *also brought* [[*her*] *friend*] [*Helen*].

Mention Detection

- Mention: span of text referring to some entity

- three kinds of mentions:
  - Pronouns
    I, your, it, she, him, etc.
  - Named entities
    People, places, etc.
  - Noun phrases
    “a dog,” “the big fluffy cat stuck in the tree”

Mention Detection

- Mention: span of text referring to some entity

- use other NLP systems to detect these
  - Pronouns
    → POS tagger
  - Named entities
    → NER system
  - Noun phrases
    → constituency parser

Difficulties in Mention Detection

- marking all pronouns, named entities, and NPs as mentions may over-generate 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

Difficulties in Mention Detection

how to deal with these "bad" mentions?

- idea: train a classifier to filter out spurious mentions
  - much more common: keep all mentions as "candidate mentions"; after coreference system is done, discard all singleton mentions

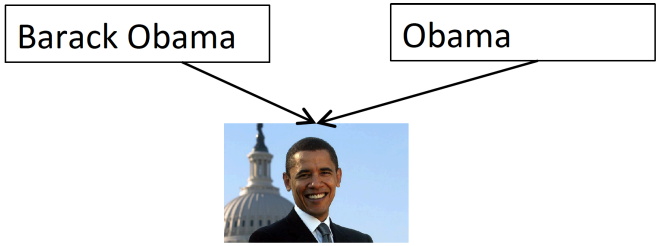
can we avoid a pipelined system altogether?

- idea: train specific mention detection system (instead of POS tagger, NER system etc.)
- other idea: train mention detection + coreference resolution together end-to-end

Anaphora vs Coreference

- coreference: two mentions refer to the same entity

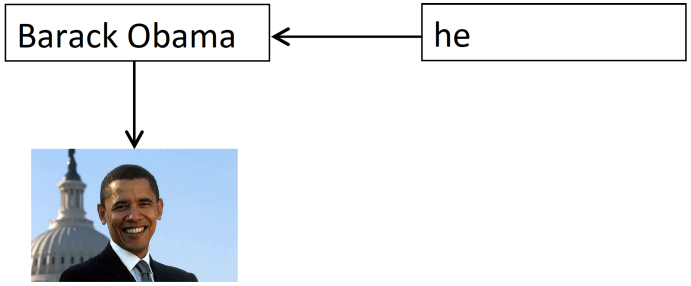
  Barrack Obama travelled to Rome this evening.
  Obama was accompanied by….



- anaphora: *anaphor* refers to *antecedent*; interpretation of anaphora determined by interpretation of antecedent

  Barrack Obama said he would sign the bill.

  antecedent anaphor

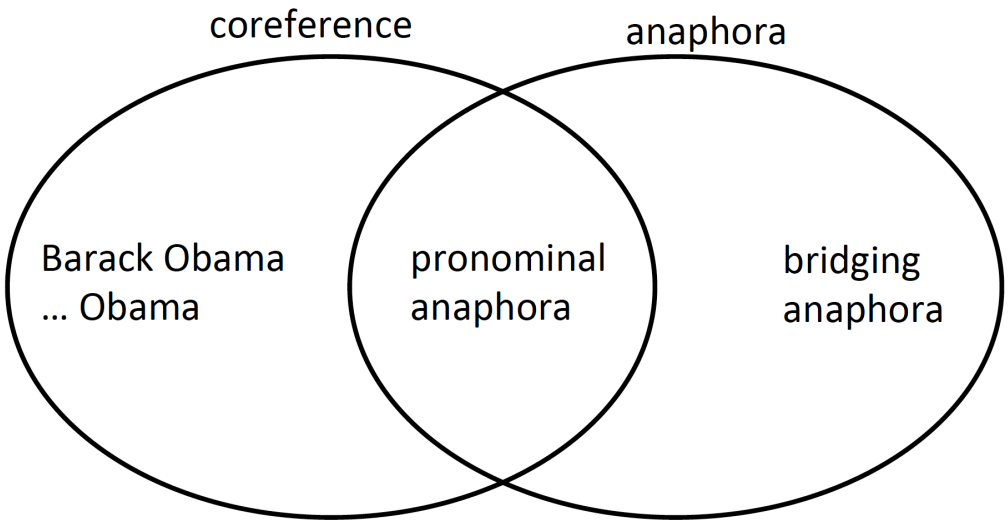


Anaphora vs Coreference

- 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

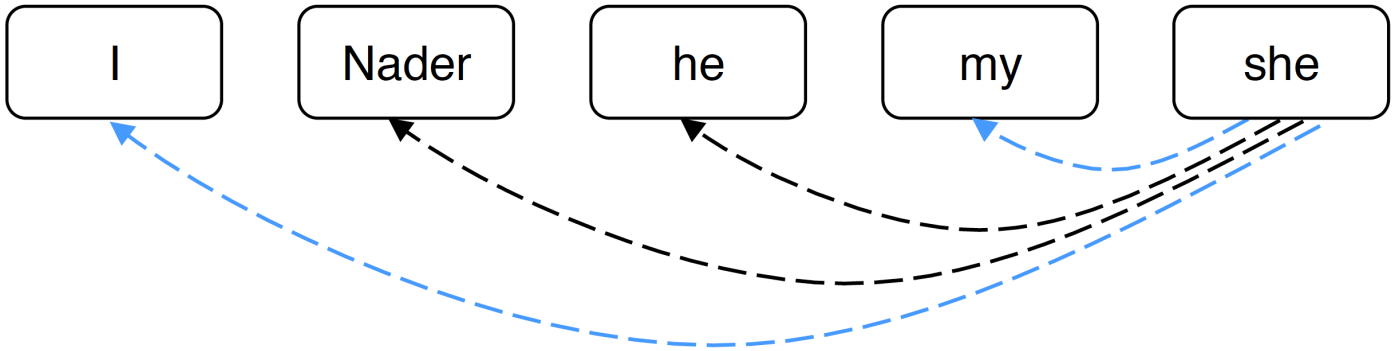
- reverse order of antecendent and anaphora (= “cataphora”)

> *“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)

Coreference Models: Mention Pair Classifier

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

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

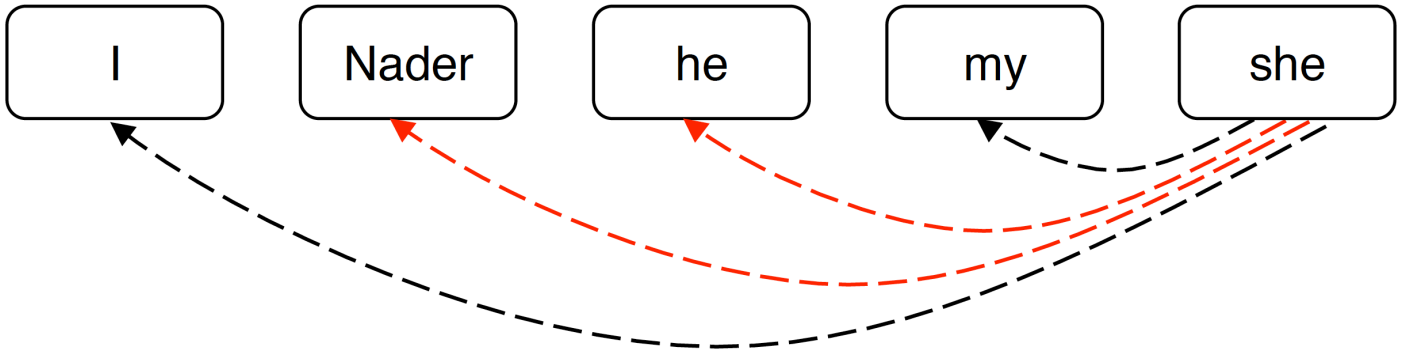


Positive examples: want $p(m_i, m_j)$ to be near 1

Coreference Models: Mention Pair Classifier

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

*"**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

Coreference Models: Mention Pair Classifier

- training: use cross-entropy loss:
  ($y_{ij} = 1$ if (m_i, m_j) coreferent; $y_{ij} = -1$ if (m_i, m_j) not coreferent):

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

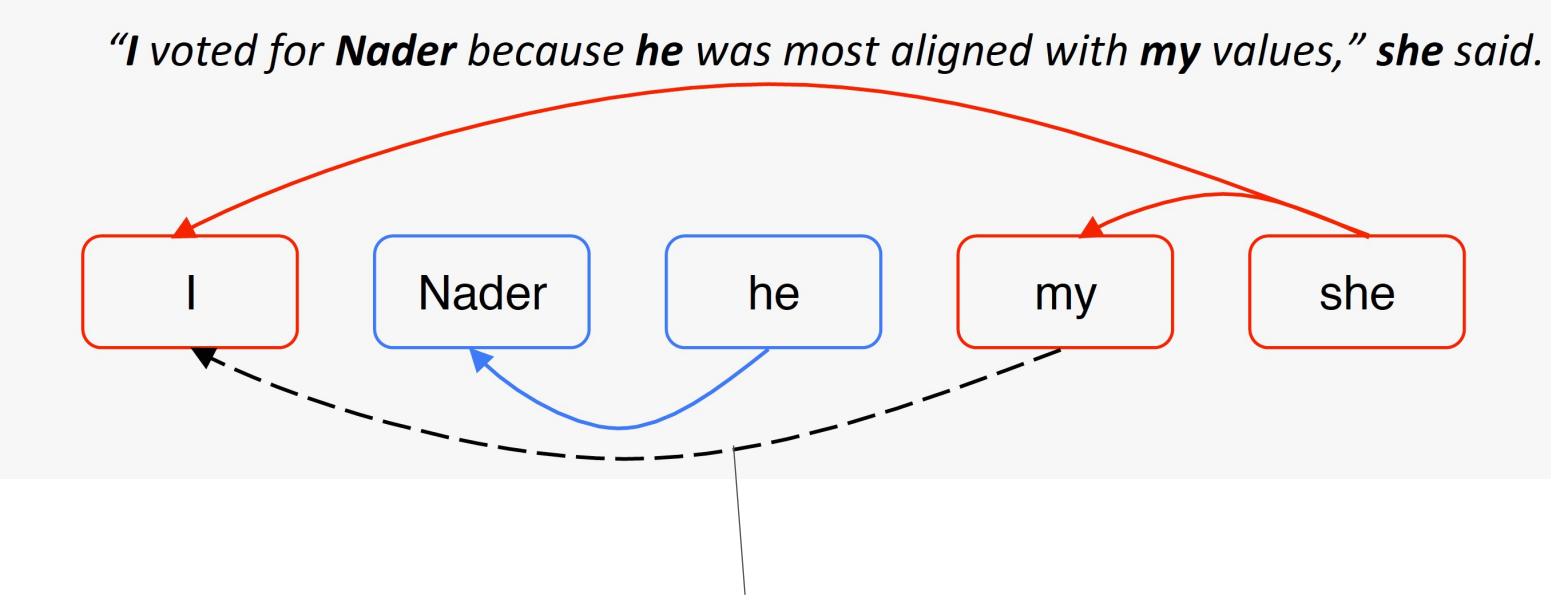
Iterate through mentions

Iterate through candidate antecedents (previously occurring mentions)

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

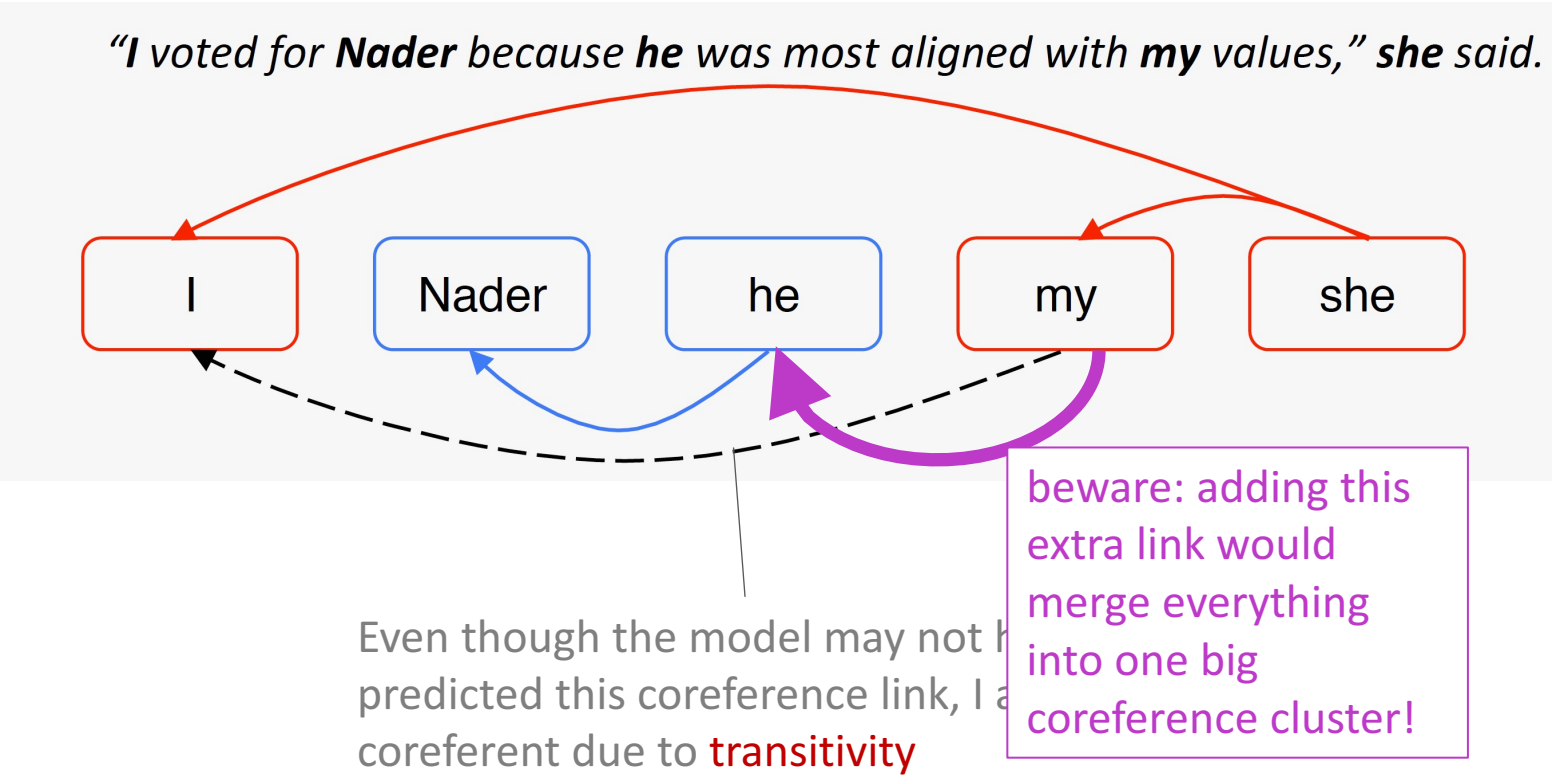
Coreference Models: Mention Pair Classifier: Test Time

- add coreference link if $p(m_i, m_j) >$ threshold
- take transitive closure to complete clustering



Coreference Models: Mention Pair Classifier: Test Time

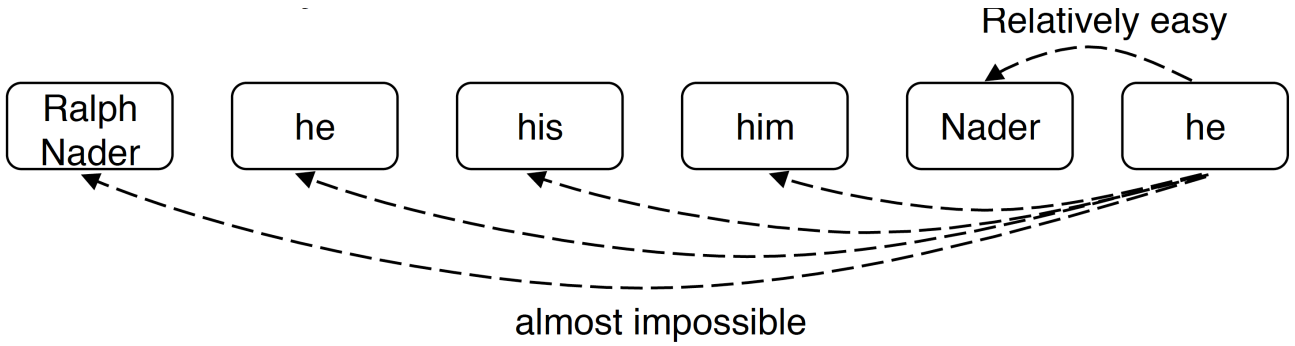
- add coreference link if $p(m_i, m_j) >$ threshold
- take transitive closure to complete clustering



Coreference Models: Mention Pair Classifier: Disadvantage

- Suppose we have a long document with the following mentions

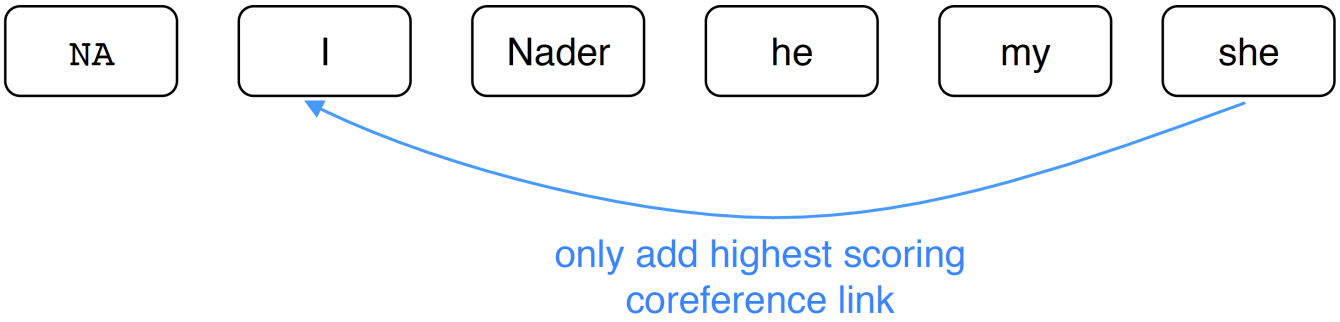
Ralph Nader … **he** … **his** … **him** … <several paragraphs>
… *voted for* ***Nader*** *because* ***he*** …



- → usually one clear antecedent → train the model to predict one antecendent for each mention only → Mention Ranking

Coreference Models: Mention Ranking

- assign each mention its highest scoring antecedent only (use NA antecendent to allow model to decline linking current mention to anything)



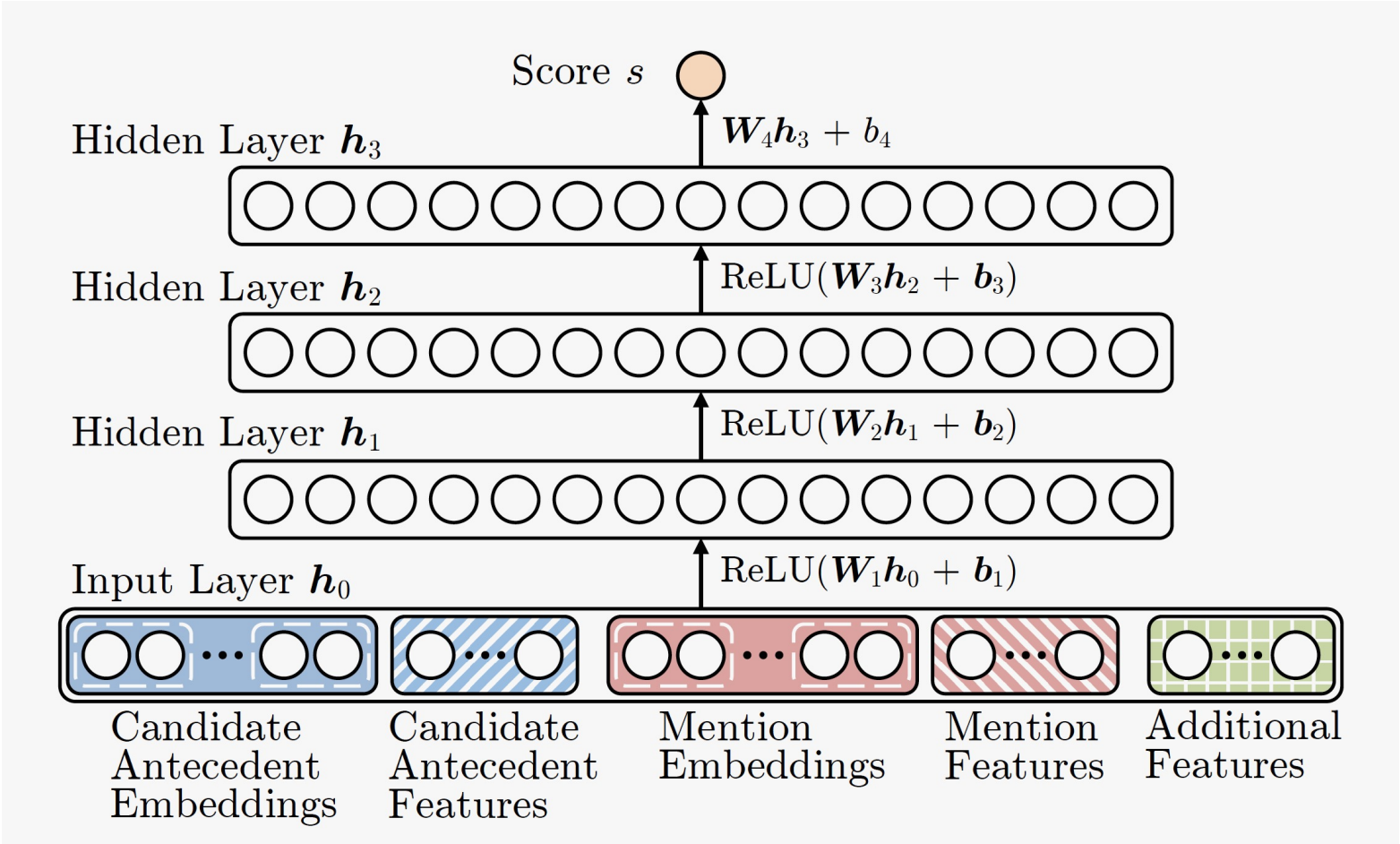
p(NA, she) = 0.1
p(I, she) = 0.5
p(Nader, she) = 0.1
p(he, she) = 0.1
p(my, she) = 0.2

Computing $p(m_i, m_j)$

- classic NLP: use a lot of handcrafted 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.
  - …

Computing $p(m_i, m_j)$: Simple Neural Model [4]

- input: word embeddings (Previous two words, first word, last word, head word, … of each mention
- + some categorical features (e.g. distances, document genre, speaker style, ….)



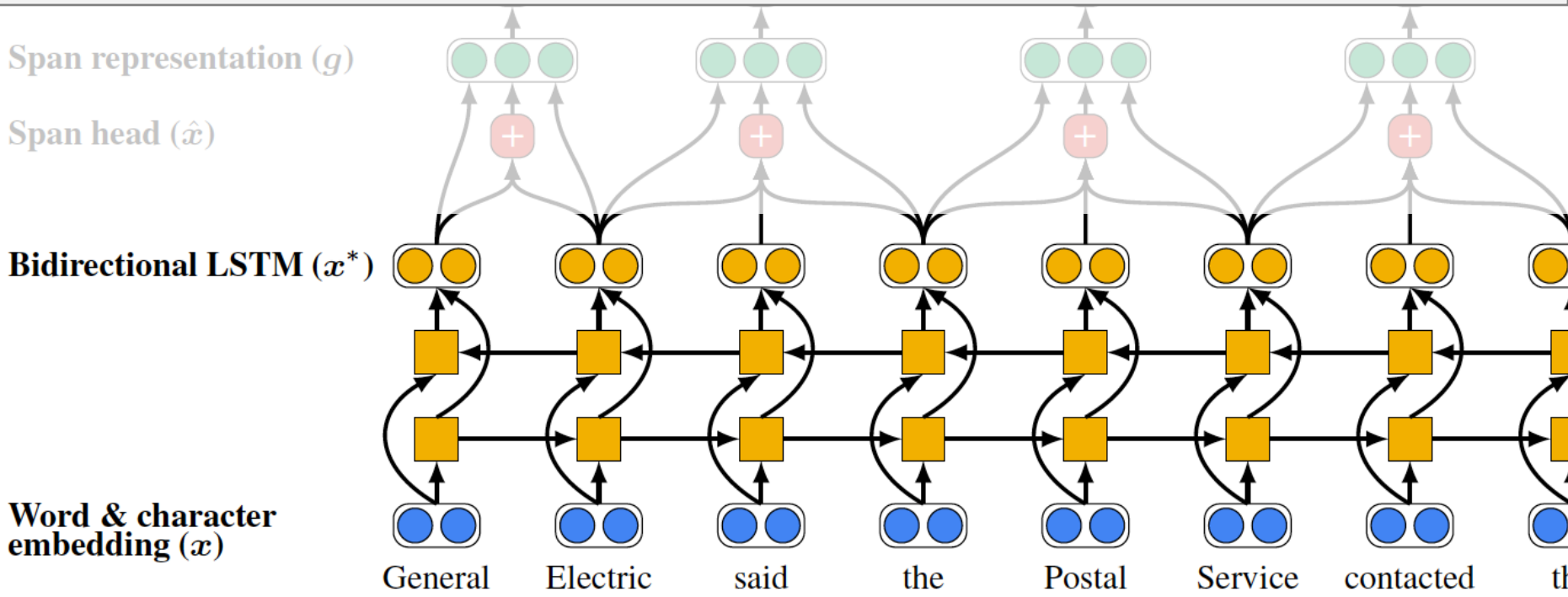
SOTA Model (2017): End-to-End [3]

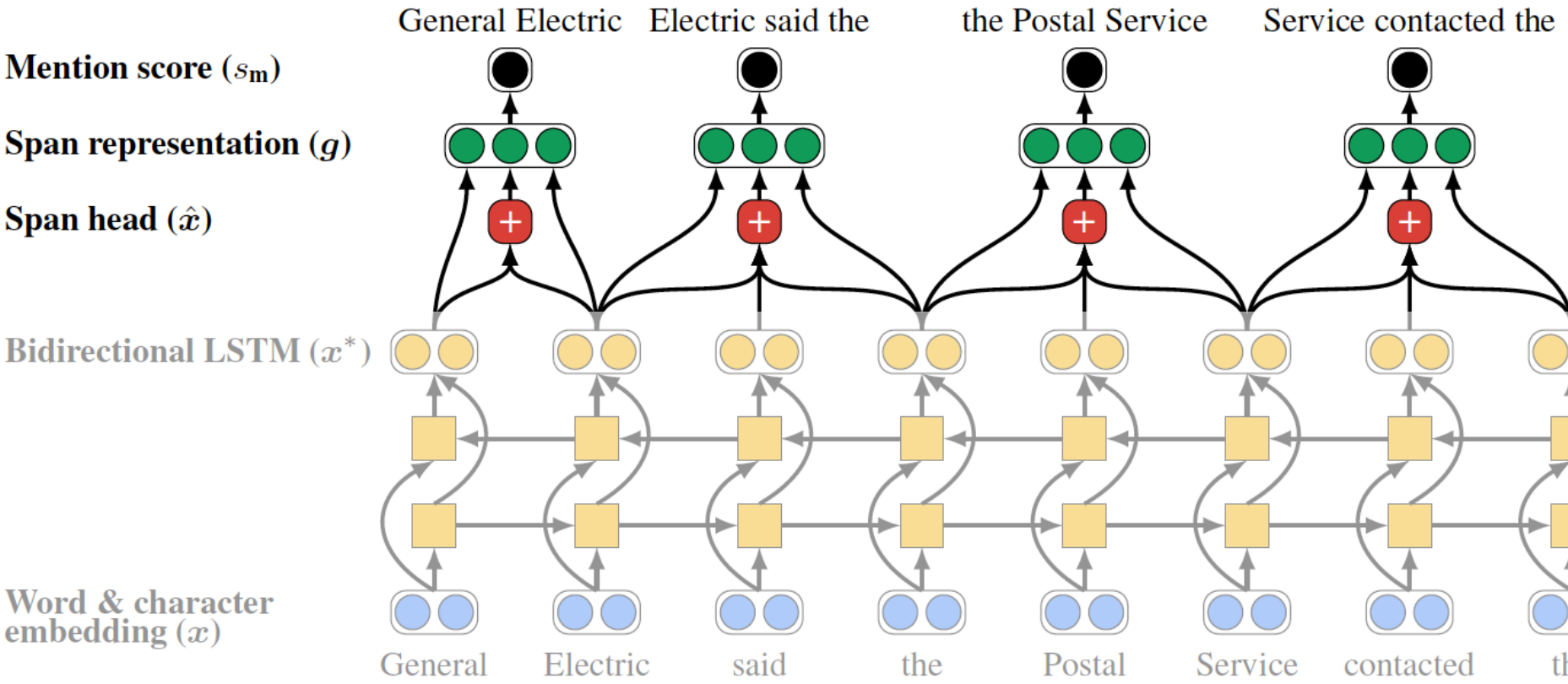
- [3]: 2017 state-of-the-art model for coreference resolution
  - Mention ranking model
  - Bi-LSTM with attention
  - do mention detection & coreference detection end to end
  - no mention detection: consider every span of text (contiguous sequence of words) (up to a certain length) as a candidate mention

if document contains T words → $O(T^2)$ many spans → $O(T^4)$ many possible coreferences → must do aggressive pruning

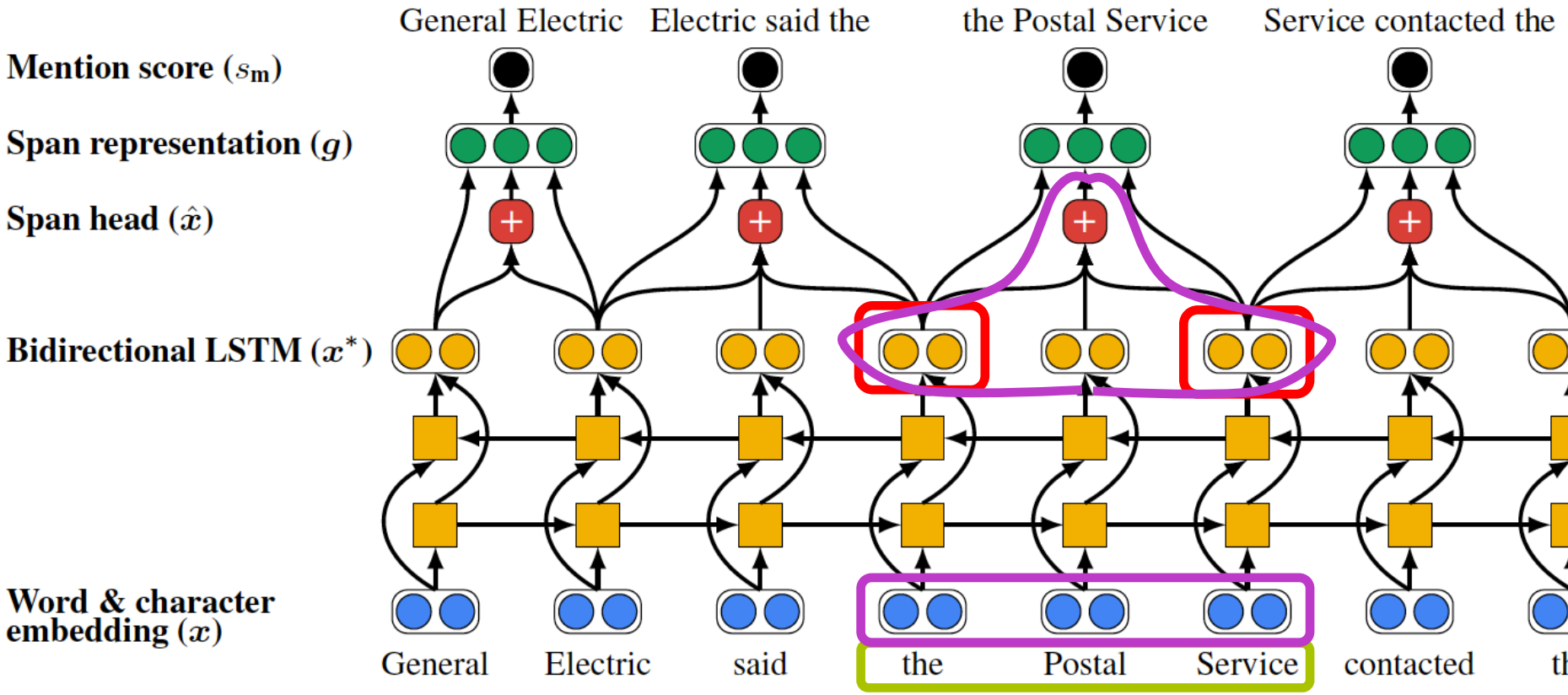
[3]: Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. 2017

SOTA Model (2017): End-to-End [3]





- compute a **representation of each span** i (from START(i) to END(i))
- in principle (←→ pruning) **all possible spans** are considered (here only a couple are depicted (e.g. *said the Postal* is omitted but is in fact also present in the network))



- compute a **representation** of each **span** i (from START(i) to END(i))

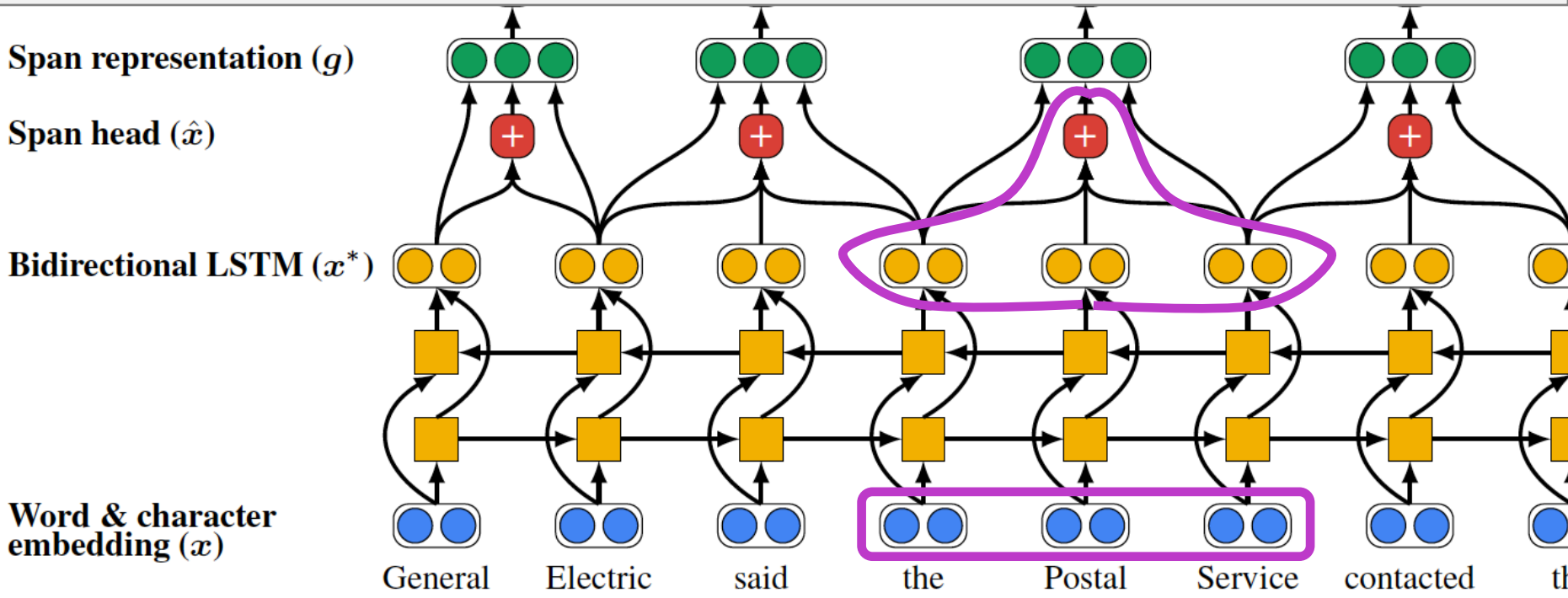
Span representation: $\boldsymbol{g}_i = [\boldsymbol{x}^*_{\text{START}(i)}, \boldsymbol{x}^*_{\text{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$

BILSTM hidden states for span's start and end

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

Additional features

SOTA Model (2017): End-to-End [3]



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

SOTA Model (2017): End-to-End [3]

- why include these elements in the span representation?

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

hidden states for span's start and end

Represents the context to the left and right of the span

Attention-based representation

Represents the span itself

the (self-)attention mechanism can be seen as "soft"-detecting the span's "head"

Additional features

Represents other information not in the text

SOTA Model (2017): End-to-End [3]

- final step: **scoring**:
  - 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 span representations as input :

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

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

include multiplicative interactions between the representations

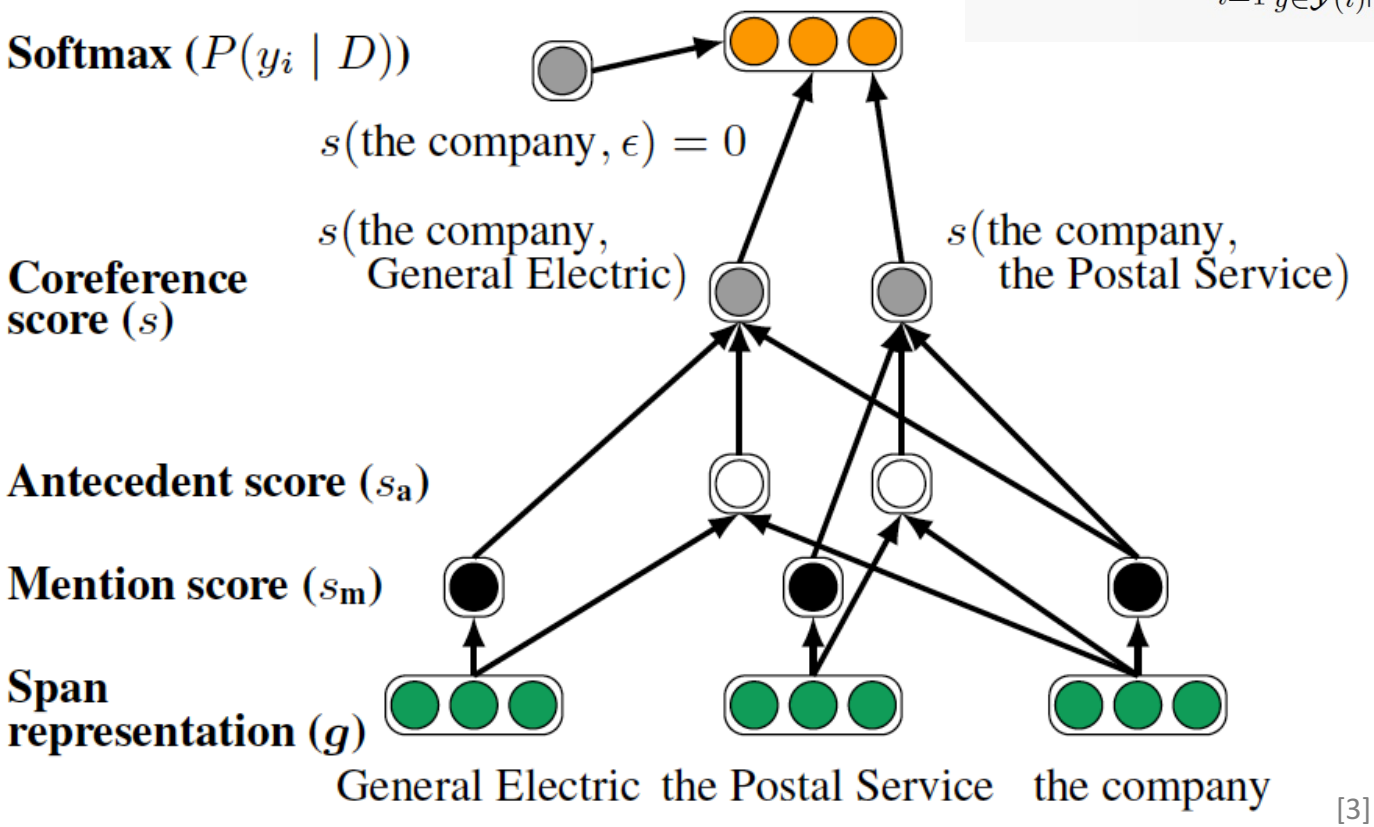
- $\circ$ is element-wise multiplication

again, we have some extra features

SOTA Model (2017): End-to-End [3]

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

Notation: each span i has antecendent y_i (possibly ϵ)



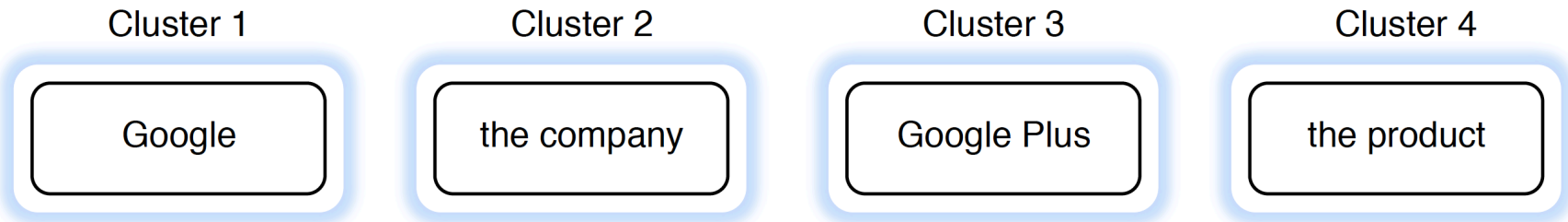
[3]

- Intractabe to score every possible pair of spans → do pruning (only consider spans up to length L and likely to be mentions using mention scores $s_m(.)$ (system's recall of true mentions=0.92))

Coreference Resolution: Clustering Based

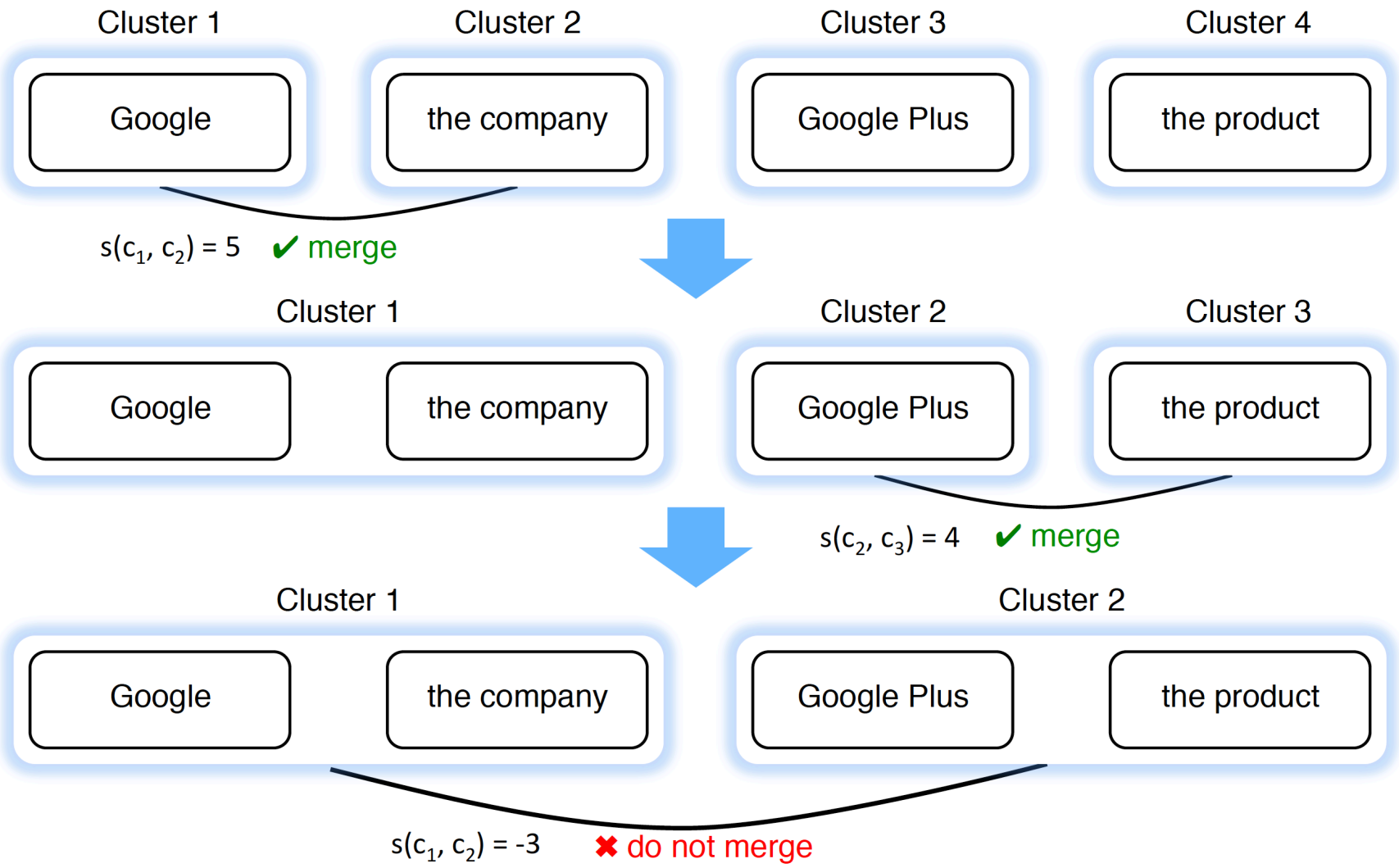
- detect mentions (using POS, NER, parsing etc.)
- agglomerative bottom up clustering:
  - start with each mention in its own cluster
  - merge clusters according to suitable score (e.g. merging costs)

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



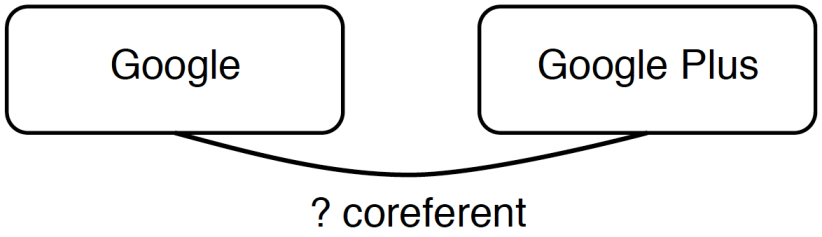
Coreference Resolution: Clustering Based

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

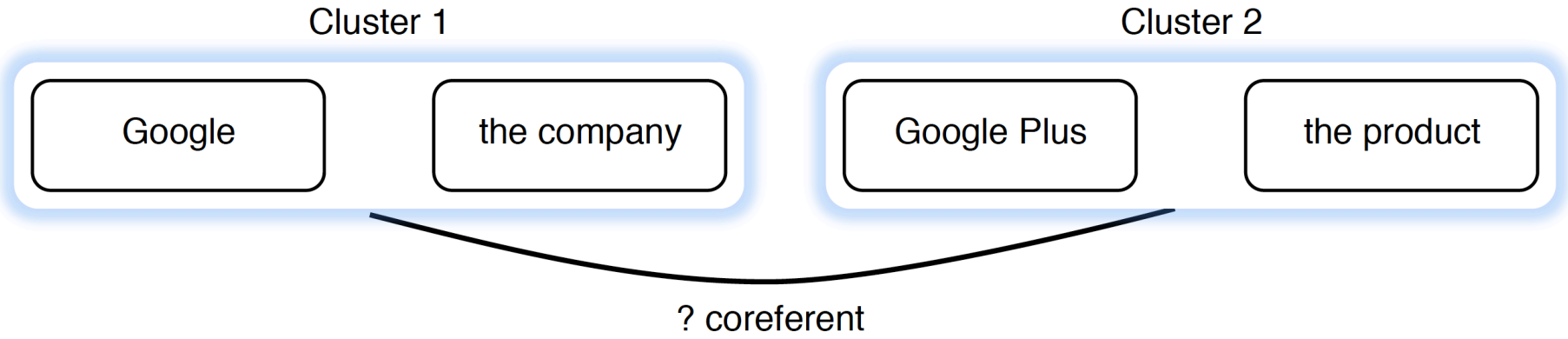


Coreference Resolution: Clustering Based

Mention-pair decision is difficult

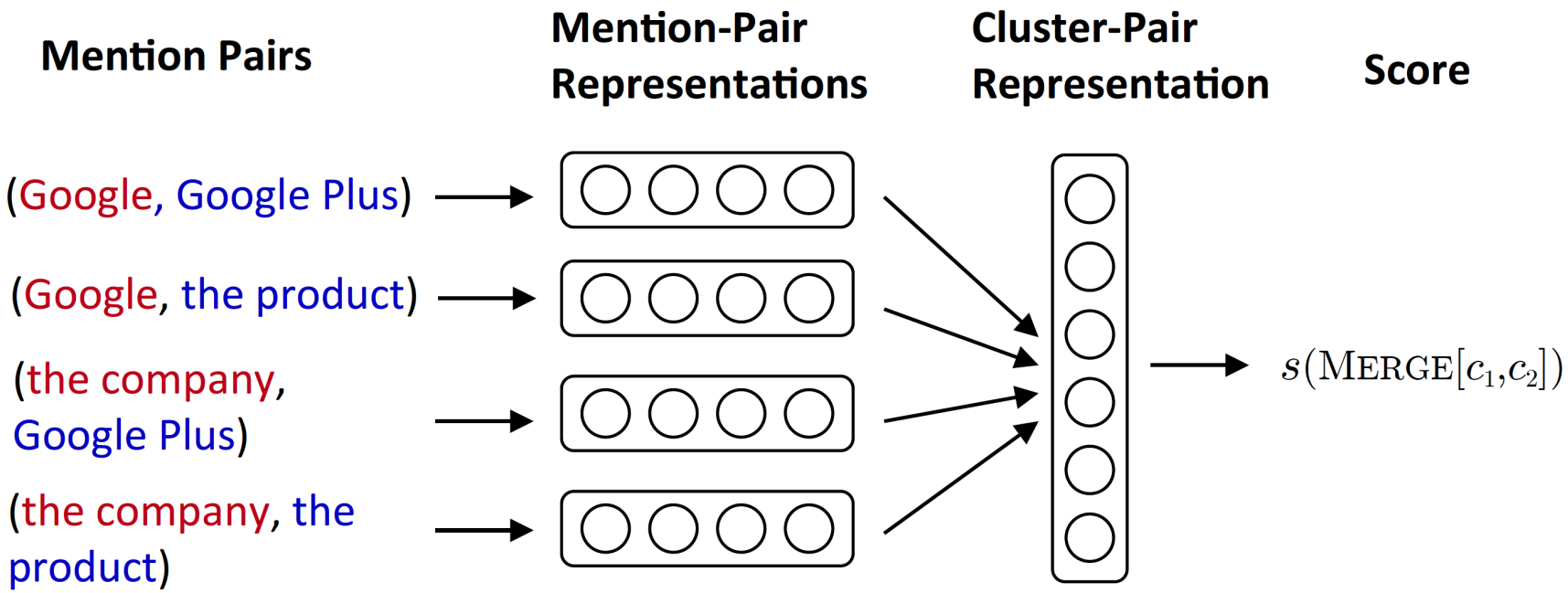


Cluster-pair decision is easier



Coreference Resolution: Clustering Based [4], [5]

Merge clusters c_1 = {Google, the company} and c_2 = {Google Plus, the product}?

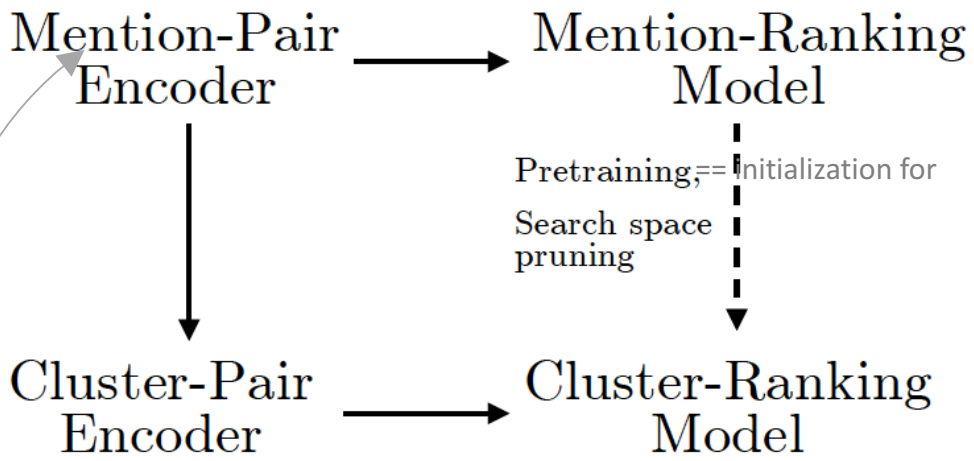


[4]: Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations.

[5]: Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models

Coreference Resolution: Clustering Based [4], [5]

simple 3 layer FFNN



single FFNN layer

on next slides

on next slides

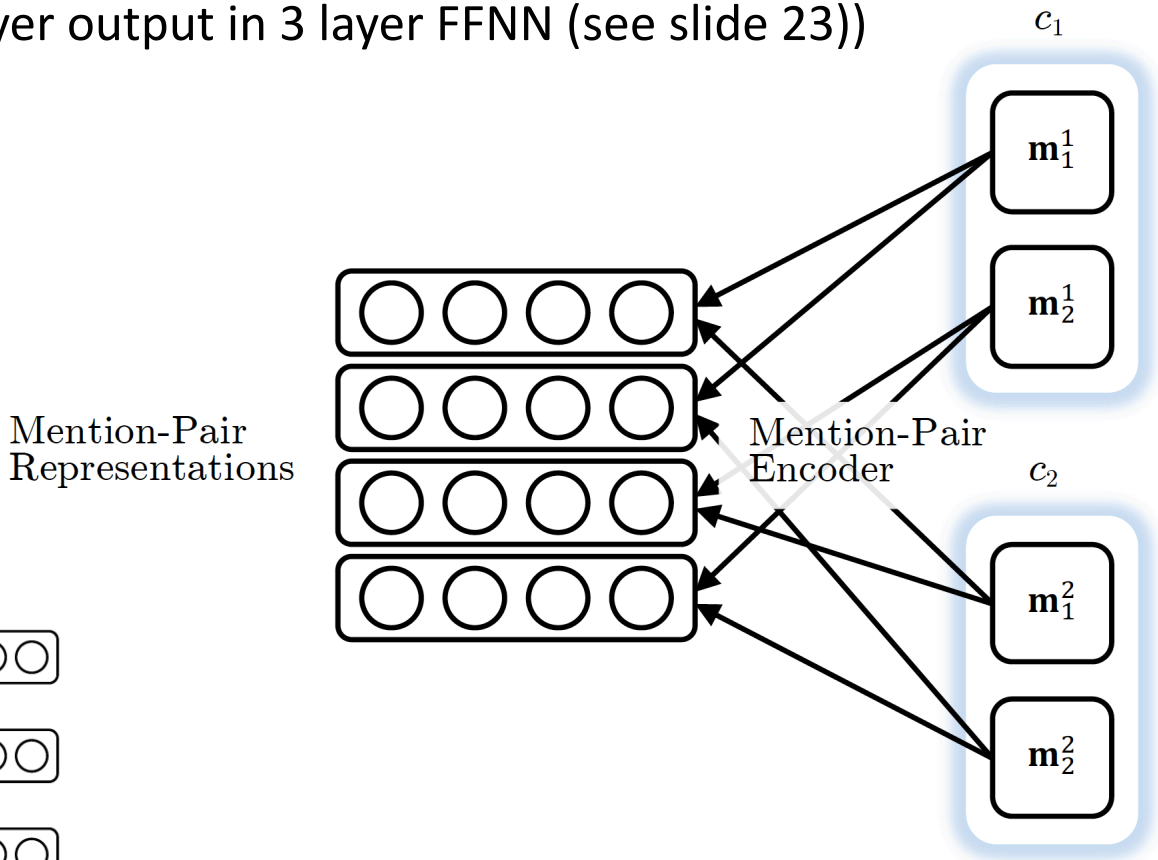
representation for groups of words: average of respective word embeddings + additional features (first + last word, distance features, dependency features, genre and speaker information etc.)

[4]: Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations.

[5]: Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models

Coreference Resolution: Clustering Based [4], [5]

- for each pair of mentions from a cluster pair: produce a vector representation (hidden layer output in 3 layer FFNN (see slide 23))



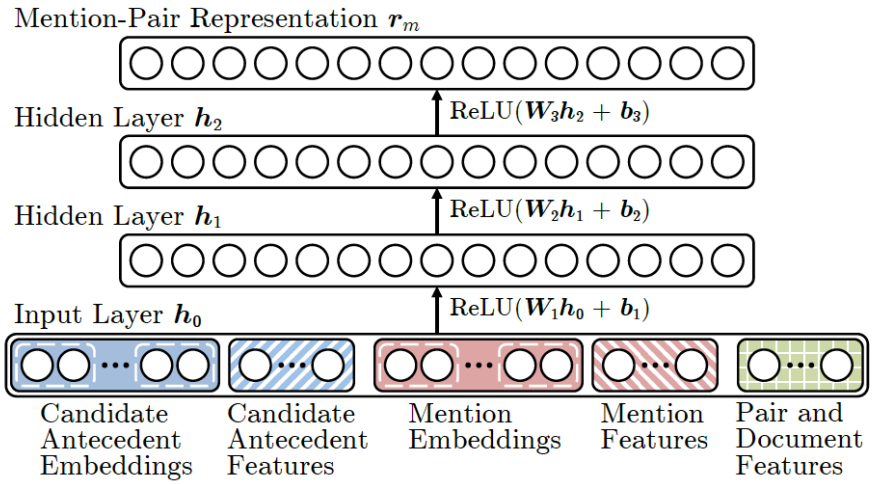
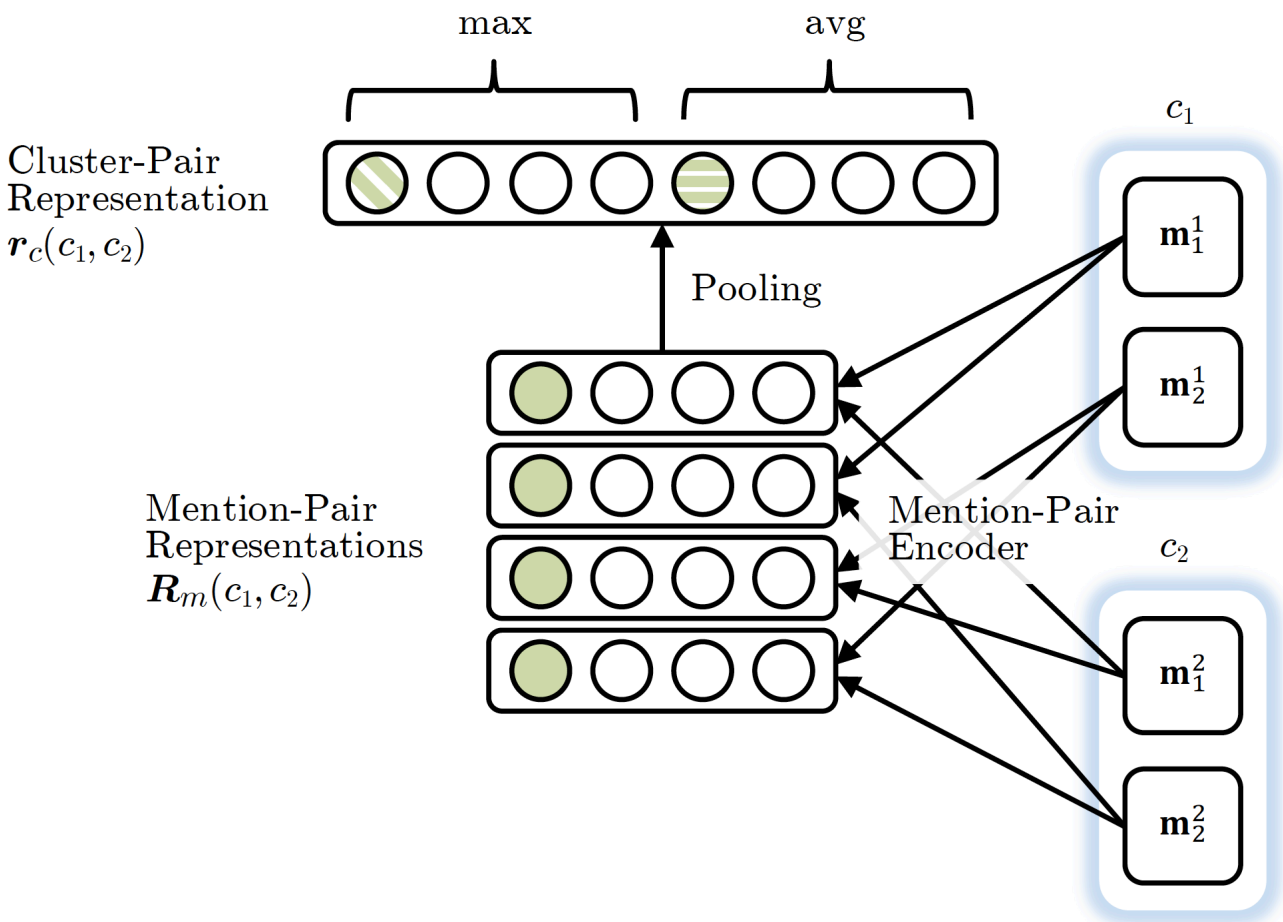


Figure 2: Mention-pair encoder.

Coreference Resolution: Clustering Based [4], [5]

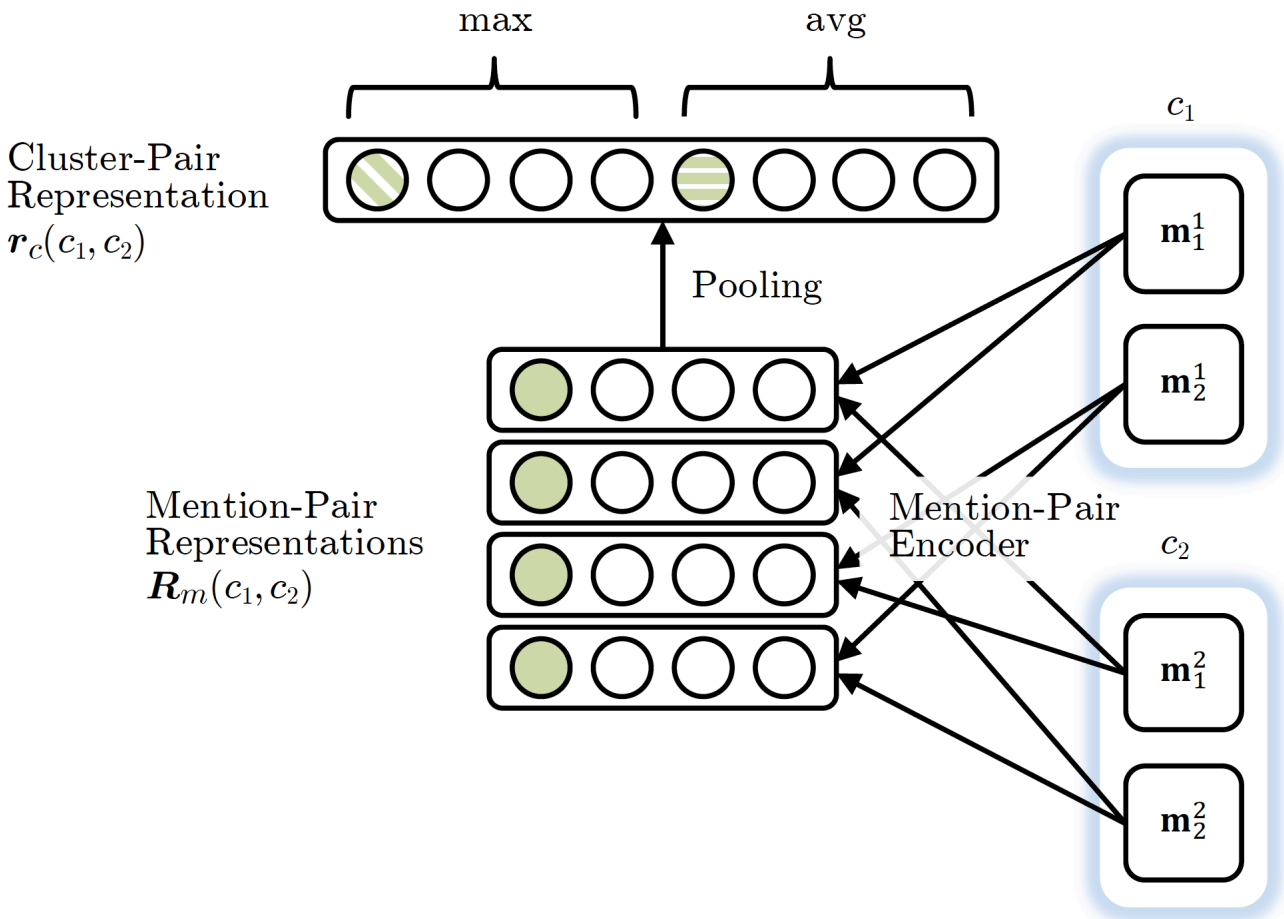
- then apply pooling over resulting mention-pair representation matrix → cluster pair representation



Coreference Resolution: Clustering Based [4], [5]

- compute merge score for cluster pair via dot-product with weight vector u :

$$s(\text{MERGE}[c_1, c_2]) = u^T r_c(c_1, c_2)$$

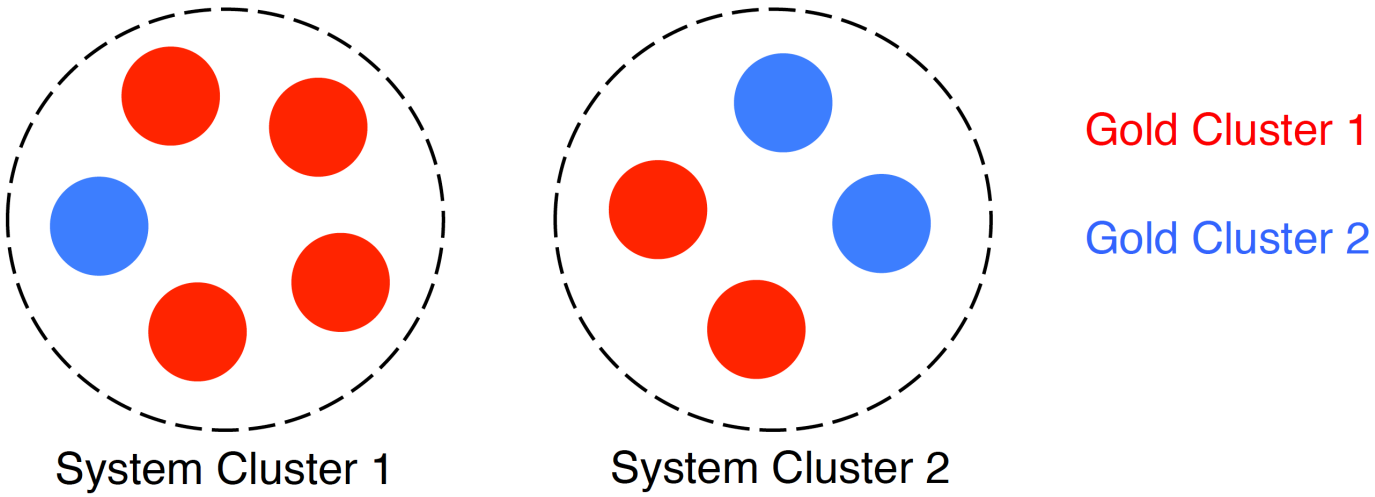


Coreference Resolution: Clustering Based: Training [4], [5]

- mention ranking: use “normal” loss concept
- BUT: for cluster ranking, current candidate cluster merges depend on previous ones it already made
- → can’t use regular supervised learning for cluster ranking sub-task in a straightforward way
- → instead use something like Reinforcement Learning to train the model (Learning to Search)
- → reward for each merge: the change in a coreference evaluation metric

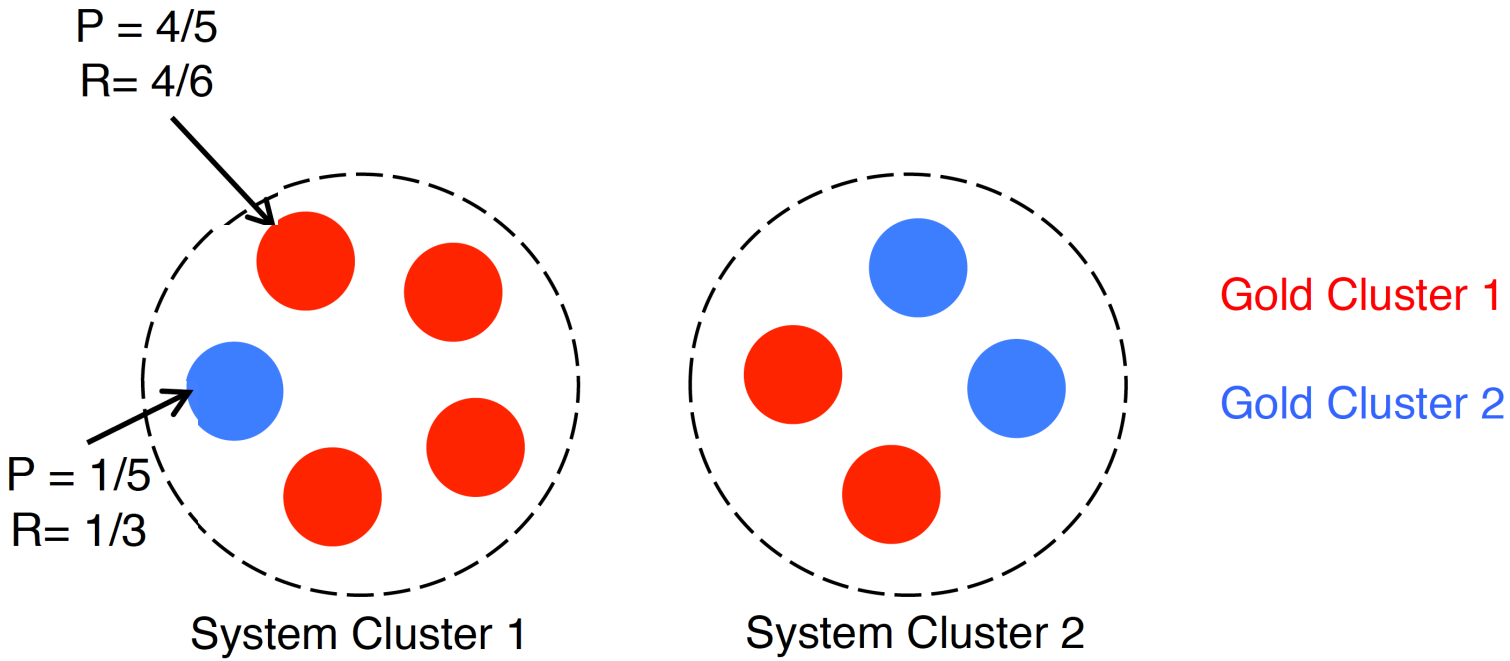
Coreference Evaluation

- many different metrics: MUC, CEAF, LEA, B-CUBED, BLANC → often report the average over a few different metrics



Coreference Evaluation

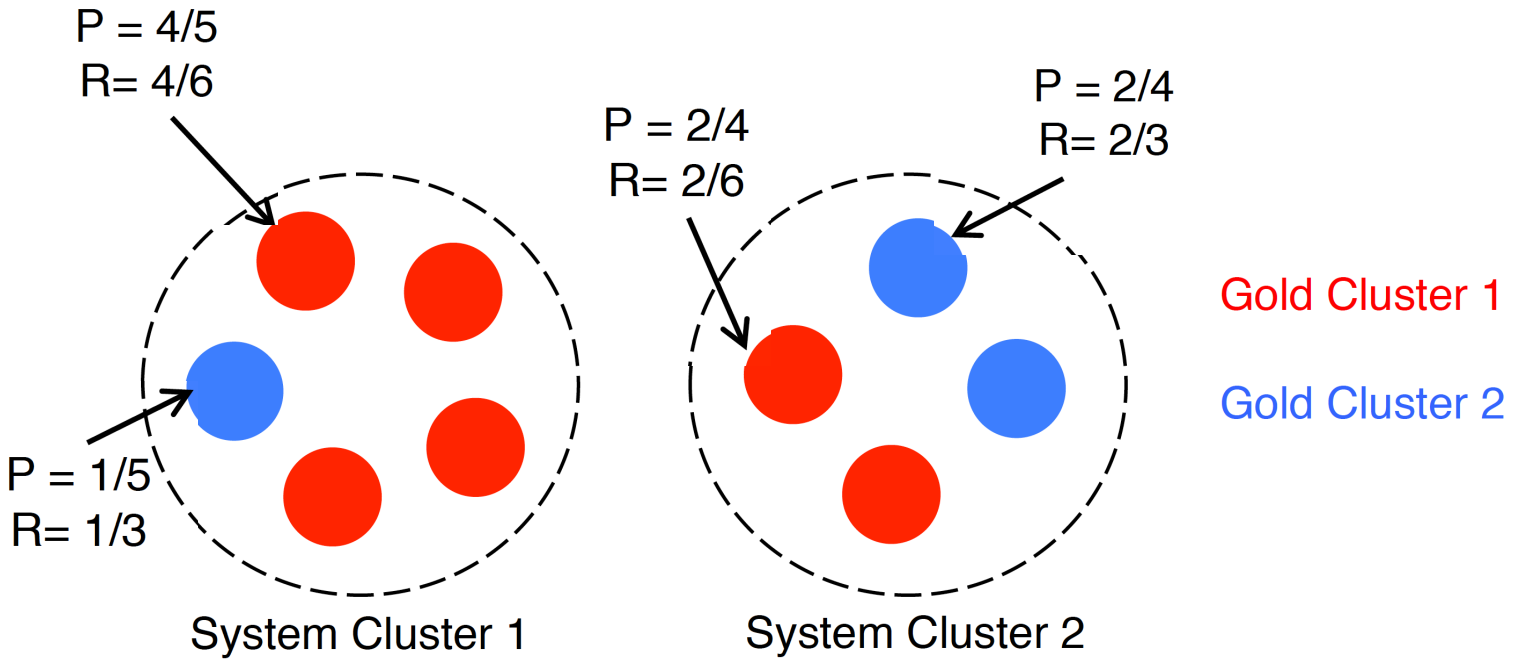
- example B-CUBED
  - for each mention: compute precision & recall



Coreference Evaluation

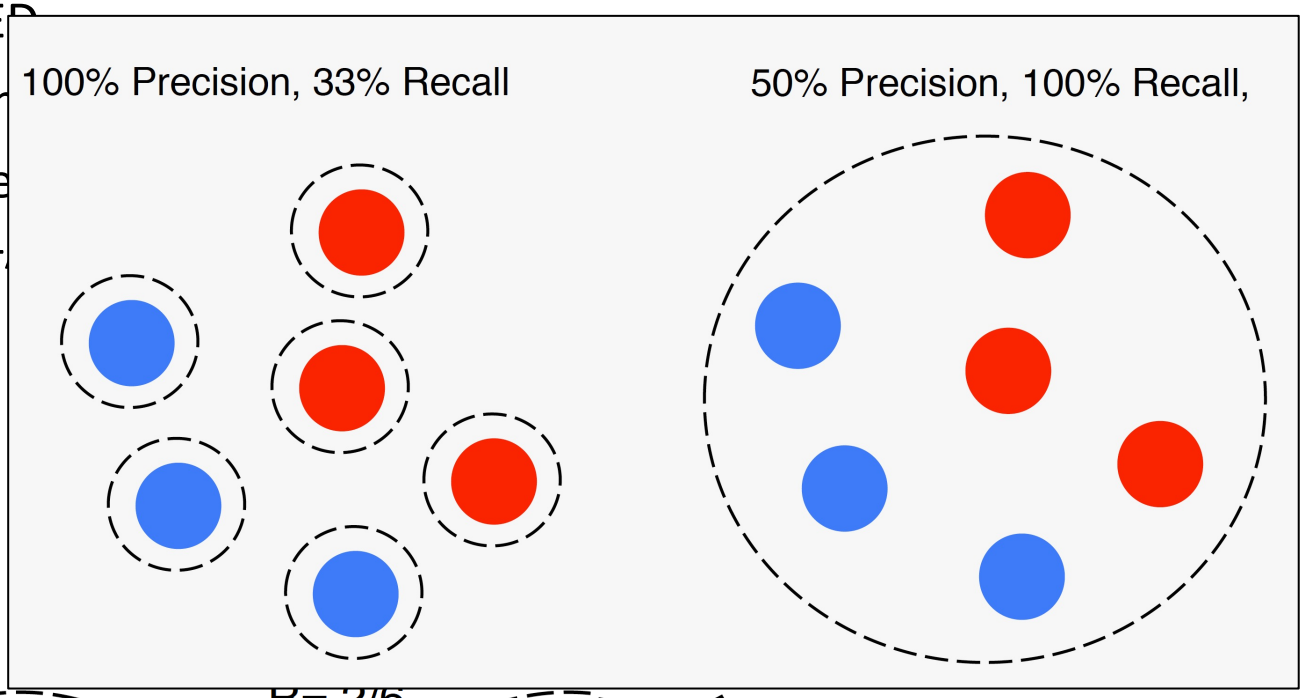
- example B-CUBED
  - for each mention: compute precision & recall
  - then average the individual Ps and Rs:

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



Coreference Evaluation

- example B-CUBE
  - for each mer
  - then average
    P = [4(4



P = 4/5
R= 4/6

R= 2/6

P = 1/5
R= 1/3

Gold Cluster 1

Gold Cluster 2

System Cluster 1

System Cluster 2

Coreference Resolution: Clustering Based: Performance [4], [5]

OntoNotes dataset: ~3000 documents labeled by humans (English and Chinese data): F1 score averaged over 3 coreference metrics

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

Bibliography

(2) Richard Socher et al: “CS224n: Natural Language Processing with Deep Learning”, Lecture Materials (slides and links to background reading) http://web.stanford.edu/class/cs224n/ (URL, May 2018), 2018

(3) Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045.

(4) Clark, K., & Manning, C. D. (2016). Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

(5) Clark, K., & Manning, C. D. (2016). Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

Recommendations for Studying

- minimal approach:

  work with the slides and understand their contents! Think beyond instead of merely memorizing the contents

- standard approach:

  minimal approach + read the corresponding paragraphs of a choice of the papers corresponding to the discussed example systems:

  - [3] sections 1-4
  - [4] sections 1-6

- interested / deeply interested student's approach:

  standard approach + study the few omitted elements of the corresponding lecture slides from [2] + read all of the recommended background reading of [2] for lecture 13