School of CIT –
Social Computing
Research Group

Technical
University
of Munich

Natural Language Processing
IN2361

Prof. Dr. Georg Groh

Chapter 26
Coreference Resolution

- content is based on [1] and [2] (lecture 13)
- certain elements (e.g. equations or tables) were taken over or taken over in a modified form from [1] and [2]
- citations of [1] or [2] or from [1] or [2] are omitted for legibility
- errors are fully in the responsibility of Georg Groh
- BIG thanks to Dan and James for a great book!

  BIG thanks to Richard Socher and his colleagues at Stanford for publishing materials [2] of a great Deep NLP lecture

Coreference Resolution

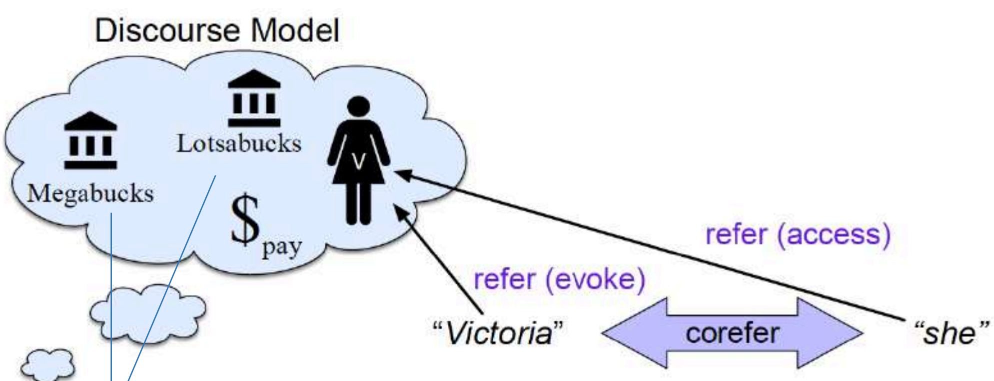
Victoria Chen, CFO of Megabucks Banking, saw her pay jump to \$2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Referent

Mention

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

Discourse Model



Antecedent

Anaphora

Singleton

Coreference Chain or Cluster

Coreference Resolution

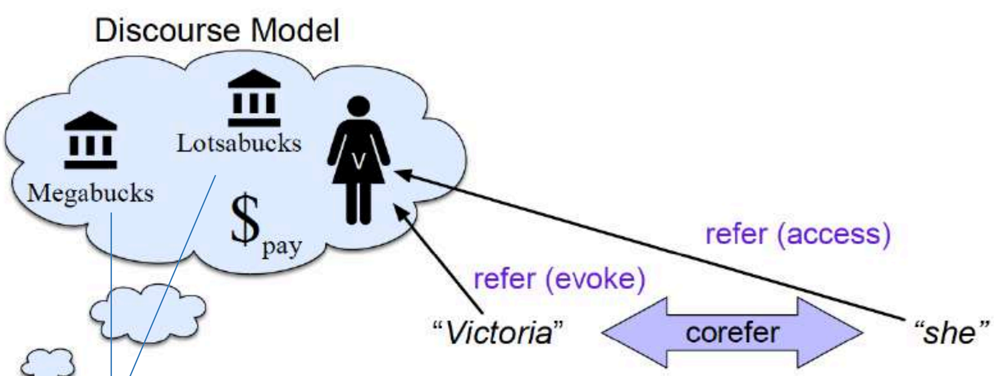
Victoria Chen, CFO of Megabucks Banking, saw her pay jump to $2.3 million, as the 38-year-old became the company's president. It is widely known that she came to Megabucks from rival Lotsabucks.

Referent

Mention

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

Discourse Model



Singleton

Antecedent

Anaphora

Coreference Chain or Cluster

Coreference resolution:

(1) identify the mentions,

(2) cluster them into coreference chains (identify discourse entities)

(3) Entity Linking: map (possibly homonymous) discourse entity (e.g. *Washington*) to real world entity (from an Ontology (e.g. Wikipedia (informal), DBpedia (semi-formal), OWL (formal)))

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

left out: personal pronoun él = he

left out: personal pronoun (él = he ☺ or) ella = she

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

- Winograd Schemata; requires full scale, AI-hard reasoning to solve

"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.

The city council denied the demonstrators a permit because
a. they feared violence.
b. they advocated violence.

- Harder than "normal" coreference resolution: event coreference

AMD agreed to [**buy**] Markham, Ontario-based ATI for around $5.4 billion in cash and stock, the companies announced Monday.
The [**acquisition**] would turn AMD into one of the world's largest providers of graphics chips.

Full Coreference Resolution is AI Hard

- Even harder than event coreference: discourse deixis:

a. But *that* turned out to be a lie. ←→ *Speech Act*
b. But *that* was false. ←→ *proposition*
c. *That* struck me as a funny way to describe the situation.

Linguistic Background: Types of Referring Expressions

- indefinite noun phrases: introduce new entities

  > Mrs. Martin was so very kind as to send Mrs. Goddard *a beautiful goose*.
  > He had gone round one day to bring her *some walnuts*.
  > I saw *this beautiful cauliflower* today.

- definite noun phrases: refer to known entities

  > It concerns a white stallion which I have sold to an officer. But the pedigree of *the white stallion* was not fully established.

  > I read about it in the *New York Times*.
  > Have you seen the car keys?

- pronouns (general case):

  > Emma smiled and chatted as cheerfully as *she* could,

Linguistic Background: Types of Referring Expressions

- pronouns: Cataphora constructions:

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

- bound pronouns:

  Every dancer brought *her* left arm forward.

- clitic pronouns (in some languages e.g. Spanish):

  La intención es reconocer el gran prestigio que tiene la maratón y unir**lo** con esta gran carrera.
  'The aim is to recognize the great prestige that the Marathon has and join|**it** with this great race."

- demonstrative pronouns (*this*, *that*: alone or as determiners):

  I just bought a copy of Thoreau's *Walden*. I had bought one five years ago. *That one* had been very tattered; *this one* was in much better condition.

Linguistic Background: Types of Referring Expressions

- Zero anaphora (e.g. in Romanic or east Asian languages):

EN [John]$_i$ went to visit some friends. On the way [he]$_i$ bought some wine.
IT [Giovanni]$_i$ andò a far visita a degli amici. Per via ϕ_i comprò del vino.
JA [John]$_i$-wa yujin-o houmon-sita. Tochu-de ϕ_i wain-o ka-tta.

[我] 前一会精神上太紧张。[0] 现在比较平静了
[I] was too nervous a while ago. ... [0] am now calmer.

- names: (refer to known and unknown entities)

a. **Miss Woodhouse** certainly had not done him justice.
b. **International Business Machines** sought patent compensation from Amazon; **IBM** had previously sued other companies.

Information Status

- Entities can be discourse-old or discourse-new,
  hearer-old or hearer-new

  - NP: discourse-new + hearer-new — *a fruit* or *some walnuts*
  - NP: discourse-new + hearer-old — *Hong Kong*, *Marie Curie*, or *the New York Times*.
  - NP: discourse-old + hearer-old — *it* in "*I went to a new restaurant. It was...*".
  - NP: inferable: neither discourse-old nor hearer-old but "inferable" w.r.t. background knowledge:

    I went to a superb restaurant yesterday. *The chef* had just opened it.
    Mix flour, butter and water. Knead *the dough* until shiny.

Complications: Non-Referring Expressions

Janet doesn't have *a car*. ⇔ **It* is a Toyota.
**The car* is red.

- Appositives: Victoria Chen, CFO of Megabucks Banking, saw ...
  United, a unit of UAL, matched the fares.

- Predicative or Prenominal NPs: her pay jumped to *$2.3 million*
  the 38-year-old became *the company's president*
  上海是[中国最大的城市] [Shanghai is *China's biggest city*]

- Expletives: *It* was Emma Goldman who founded *Mother Earth*
  It surprised me that there was a herring hanging on her wall.

- Generics: I love mangos. *They* are very tasty.
  In July in San Francisco *you* have to wear a jacket.

Linguistic Properties of the Coreference Relation

- Number agreement (e.g. singular / plural).
  difficulty:

  > IBM announced a new machine translation product yesterday. *They* have been working on it for 20 years.
  >
  > *It*

- Person agreement (he, she, him, her, …).
  difficulty: citations:

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

- Gender or noun-class agreement.
  difficulty: background knowledge might be required:

  > Maryam has a theorem. She is exciting. (she=Maryam, not the theorem)
  >
  > Maryam has a theorem. It is exciting. (it=the theorem, not Maryam)

- Binding theory constraints:

  > Janet bought herself a bottle of fish sauce. [herself=Janet]
  >
  > Janet bought her a bottle of fish sauce. [her$\neq$Janet]

Linguistic Properties of the Coreference Relation

- Recency:

  The doctor found an old map in the captain's chest. Jim found an even older map hidden on the shelf. It described an island.

- Grammatical role (preference for subjects)

  Billy Bones went to the bar with Jim Hawkins. He called for a glass of rum. [he = Billy]

  Jim Hawkins went to the bar with Billy Bones. He called for a glass of rum. [he = Jim]

- Verb semantics

  John telephoned Bill. He lost the laptop.
  John criticized Bill. He lost the laptop.

- Selectional restrictions / preference

  I ate the soup in my new bowl after cooking it for hours

  edible

Mention Detection

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.

parsing, NER: extract all NPs, possessive pronouns, and named entities

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	

Filtering non-referential mentions:

- rules

```
It is Modaladjective that S
It is Modaladjective (for NP) to VP
It is Cogv-ed that S
It seems/appears/means/follows (that) S
```

- ML: anaphoricity classifier, discourse-new classifier, referentiality with fine tuned thresholds
- ....

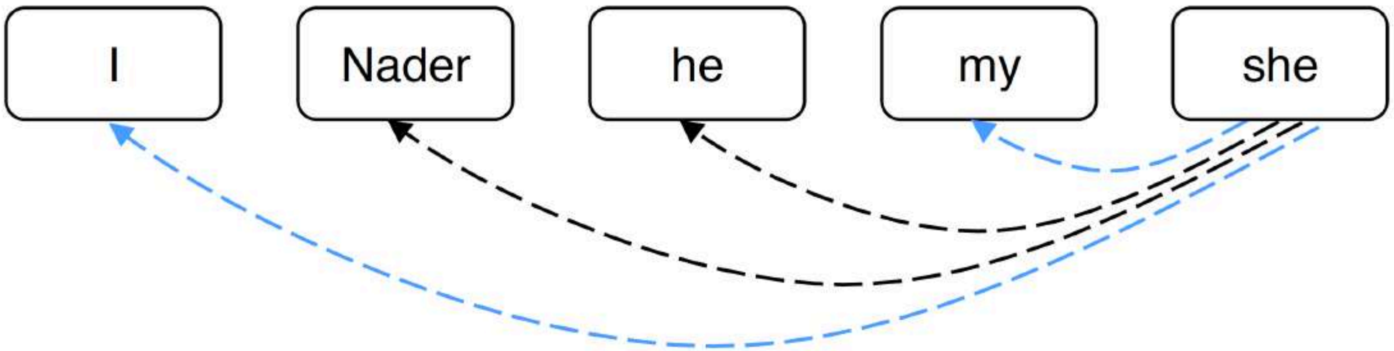
→ Modern CoReference: Supervised (NN or other) End to End

Victoria Chen	her pay	she
Megabucks Bank	the 38-year-old	Megabucks
her	the company	Lotsabucks

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.*

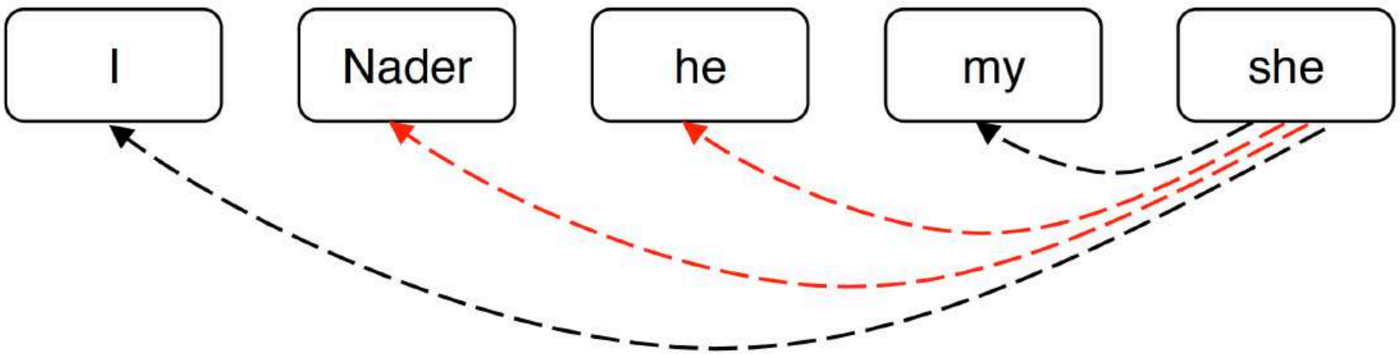


Positive examples: want $p(m_i, m_j)$ to be near 1

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

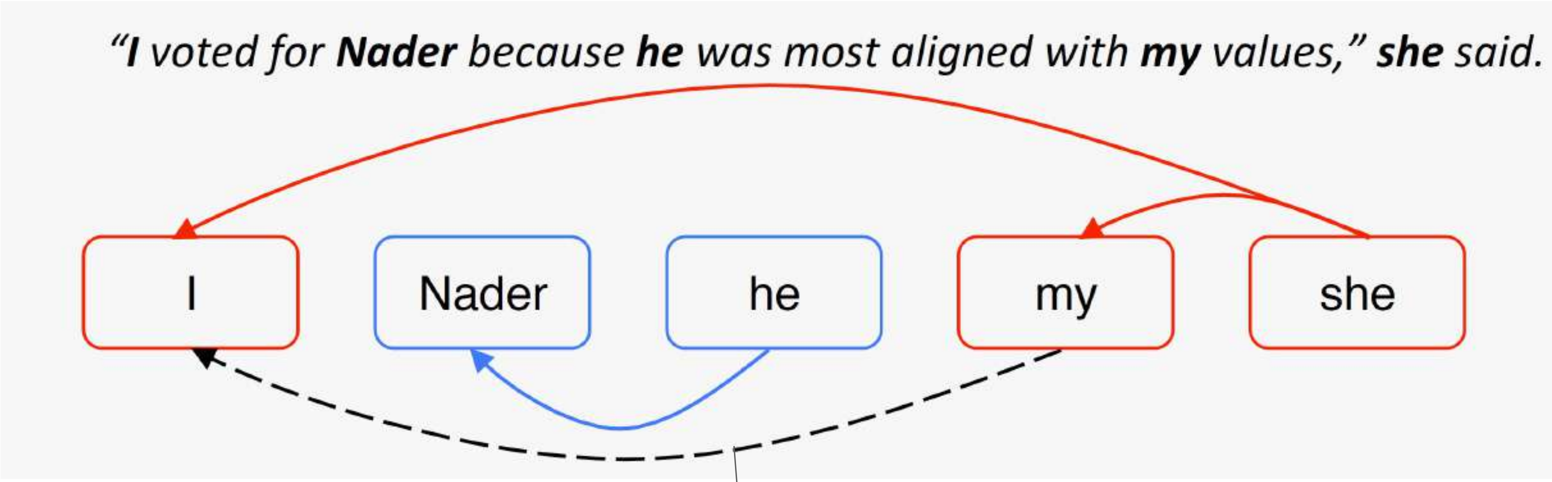
Iterate through candidate antecedents (previously occurring mentions)

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

- select appropriately equal numbers of positive and negative examples.

Coreference Models: Mention Pair Classifier: Test Time

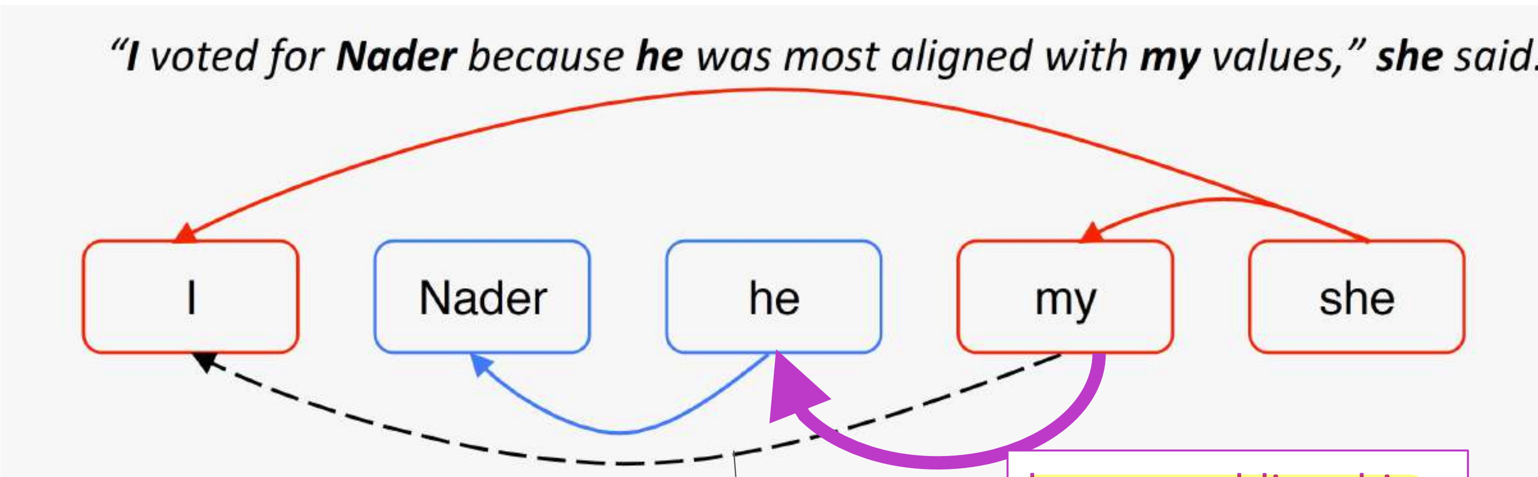
- add coreference link if $p(m_i, m_j) >$ threshold
- take transitive closure to complete clustering



Even though the model may not have predicted this coreference link, I and my are coreferent due to transitivity

Coreference Models: Mention Pair Classifier: Test Time

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



Even though the model may not h
predicted this coreference link, I a
coreferent due to transitivity

beware: adding this extra link would merge everything into one big coreference cluster!

Coreference Models: Mention Pair Classifier: Disadvantages

- yes/no decision is made for each pair locally only, no comparison of antecedents “which one is better”
    - idea: train the model to predict one antecedent (with highest p) for each mention only → Mention Ranking

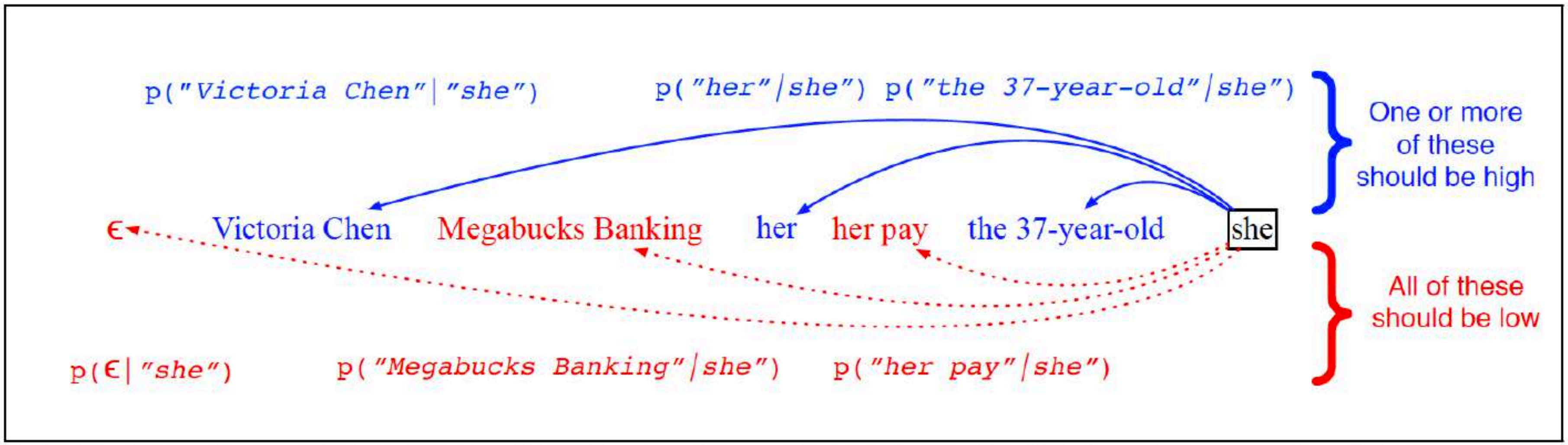
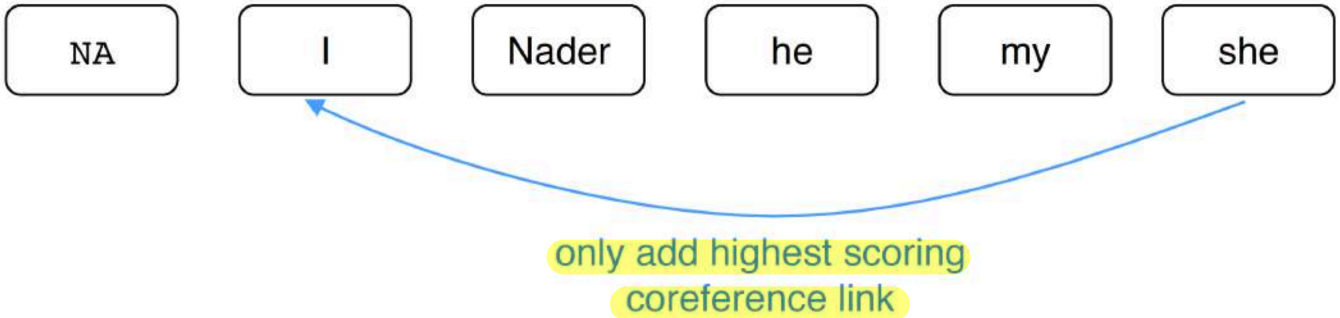


Figure 26.3 For each candidate anaphoric mention (like *she*), the mention-ranking system assigns a probability distribution over all previous mentions plus the special dummy mention ϵ.

Coreference Models: Mention Ranking

- assign each mention its highest scoring antecedent only (use “NA” antecedent 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

more difficult to train because may true antecedents for a mention may exist, which one is „best“ is a latent information

Classifiers: Hand-Built Features

Features of the Anaphor or Antecedent Mention		
First (last) word	Victoria/she	First or last word (or embedding) of antecedent/anaphor
Head word	Victoria/she	Head word (or head embedding) of antecedent/anaphor
Attributes	Sg-F-A-3-PER/ Sg-F-A-3-PER	The number, gender, animacy, person, named entity type attributes of (antecedent/anaphor)
Length	2/1	length in words of (antecedent/anaphor)
Grammatical role	Sub/Sub	The grammatical role—subject, direct object, indirect object/PP—of (antecedent/anaphor)
Mention type	P/Pr	Type: (P)roper, (D)efinite, (I)ndefinite, (Pr)onoun) of antecedent/anaphor
Features of the Antecedent Entity		
Entity shape	P-Pr-D	The 'shape' or list of types of the mentions in the antecedent entity (cluster), i.e., sequences of (P)roper, (D)efinite, (I)ndefinite, (Pr)onoun.
Entity attributes	Sg-F-A-3-PER	The number, gender, animacy, person, named entity type attributes of the antecedent entity
Ant. cluster size	3	Number of mentions in the antecedent cluster
Features of the Pair of Mentions		
Longer anaphor	F	True if anaphor is longer than antecedent
Pairs of any features	Victoria/she, 2/1, P/Pr, etc.	For each individual feature, pair of type of antecedent+ type of anaphor
Sentence distance	1	The number of sentences between antecedent and anaphor
Mention distance	4	The number of mentions between antecedent and anaphor
i-within-i	F	Anaphor has i-within-i relation with antecedent
Cosine		Cosine between antecedent and anaphor embeddings
Appositive	F	True if the anaphor is in the syntactic apposition relation to the antecedent. Useful even if appositives aren't mentions (to know to attach the appositive to a preceding head)
Features of the Pair of Entities		
Exact String Match	F	True if the strings of any two mentions from the antecedent and anaphor clusters are identical.
Head Word Match	F	True if any mentions from antecedent cluster has same headword as any mention in anaphor cluster
Word Inclusion	F	All words in anaphor cluster included in antecedent cluster
Features of the Document		
Genre/source	N	The document genre— (D)ialog, (N)ews, etc,

Figure 21.4 Feature-based coreference: sample feature values for anaphor "she" and potential antecedent "Victoria Chen".

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

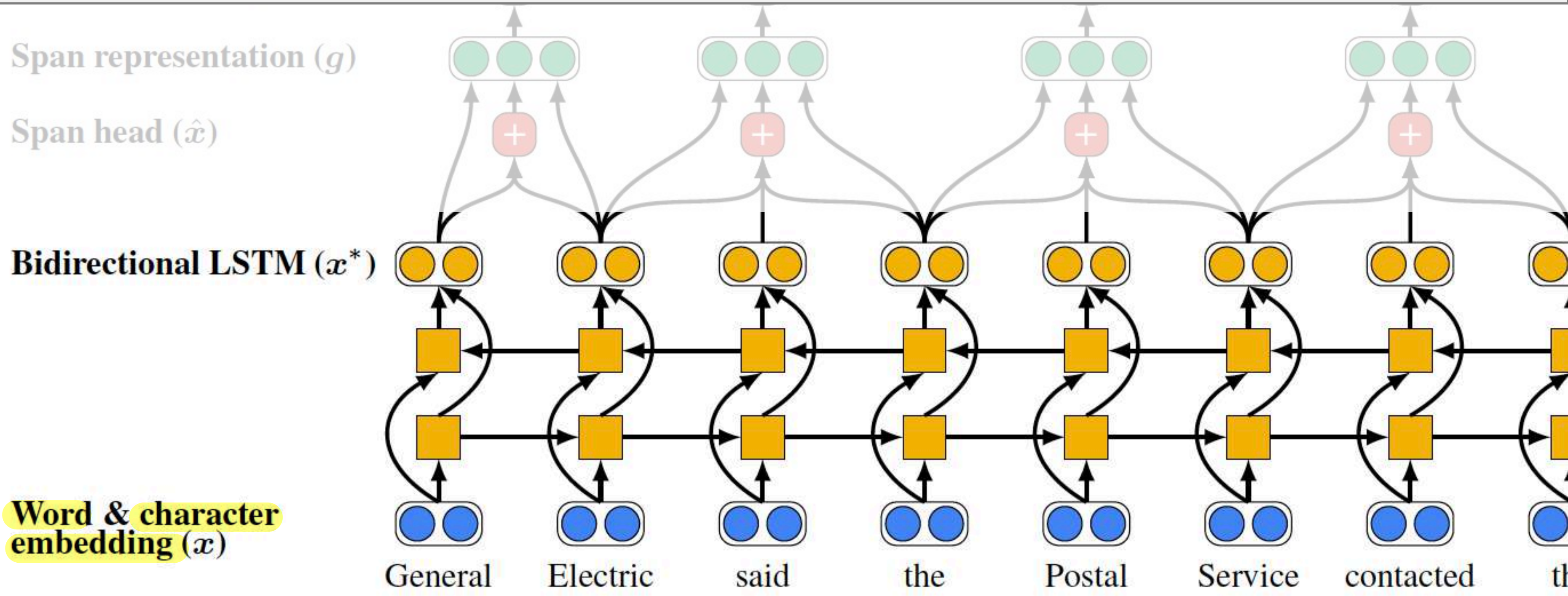
- [3]: 2017 state-of-the-art model for coreference resolution
  - Mention ranking model
  - Bi-LSTM with attention
  - do mention detection & coreference detection end to end
  - no mention detection: consider every span of text (contiguous sequence of words) (up to a certain length) as a candidate mention

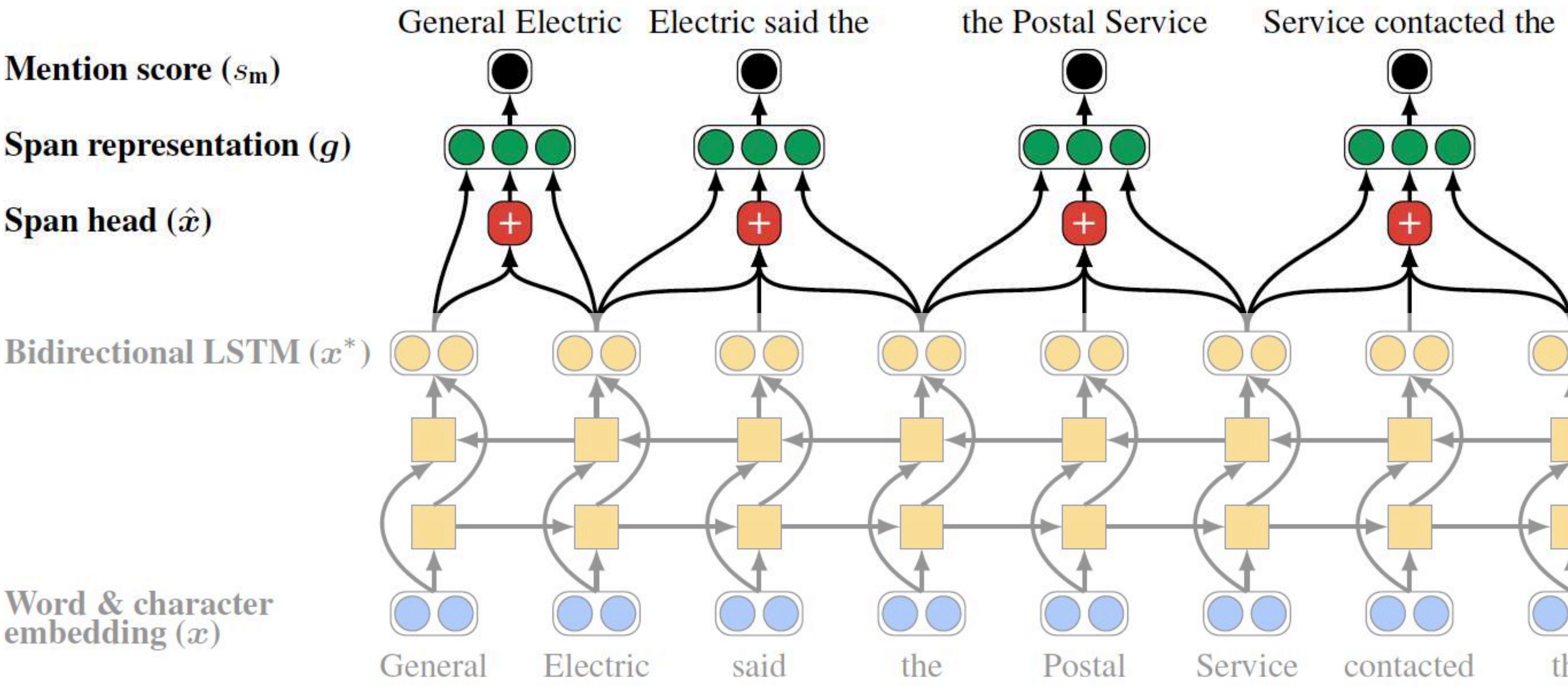
    if document contains T words → O(T²) many spans → O(T⁴) many possible coreferences → must do aggressive pruning

  - for each span i learn probability for previous span $y_i \in \{1, .., i-1, \epsilon\}$ being its antecedent: $P(y_i) = softmax\{s(i, y_i)\}$ where $s(i, j)$ is a score for the corefentiality of spans i and j

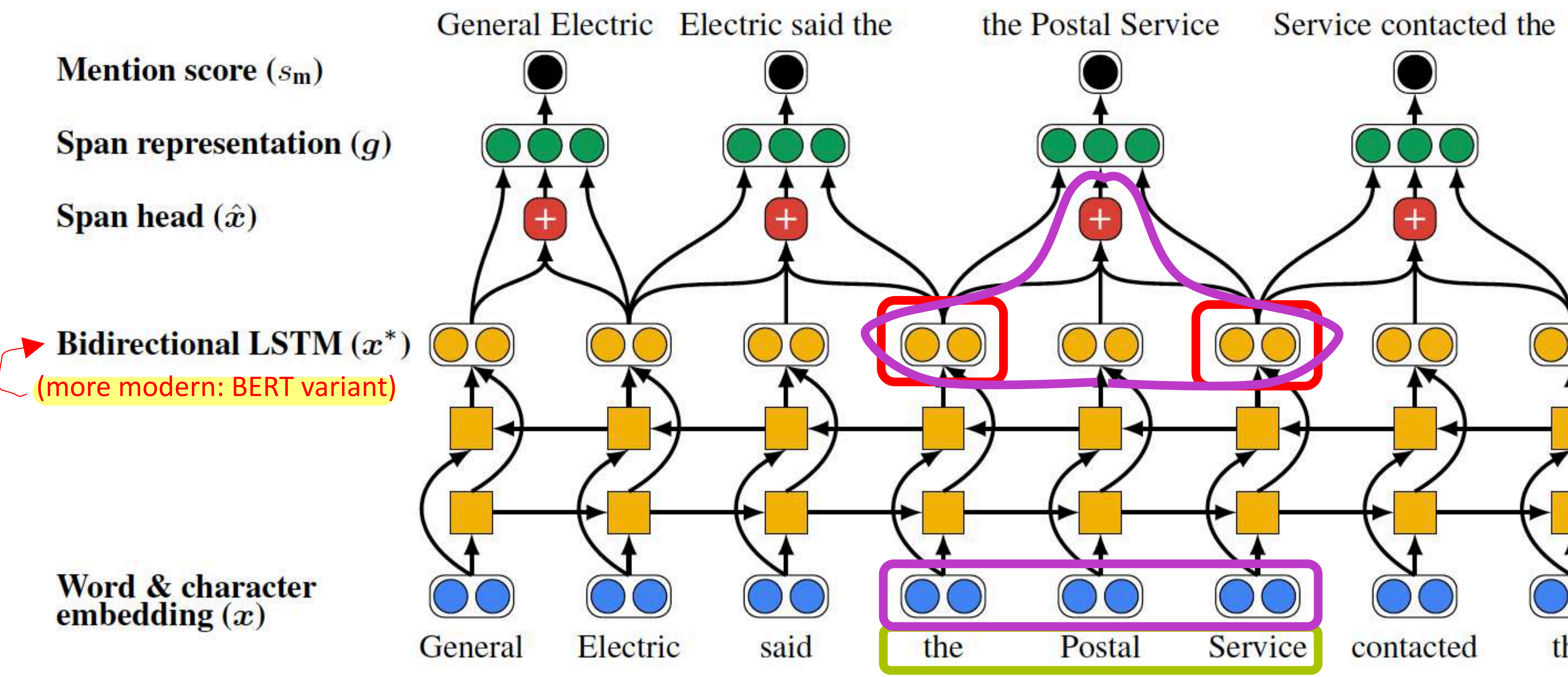
[3]: Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end neural coreference resolution. 2017

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))

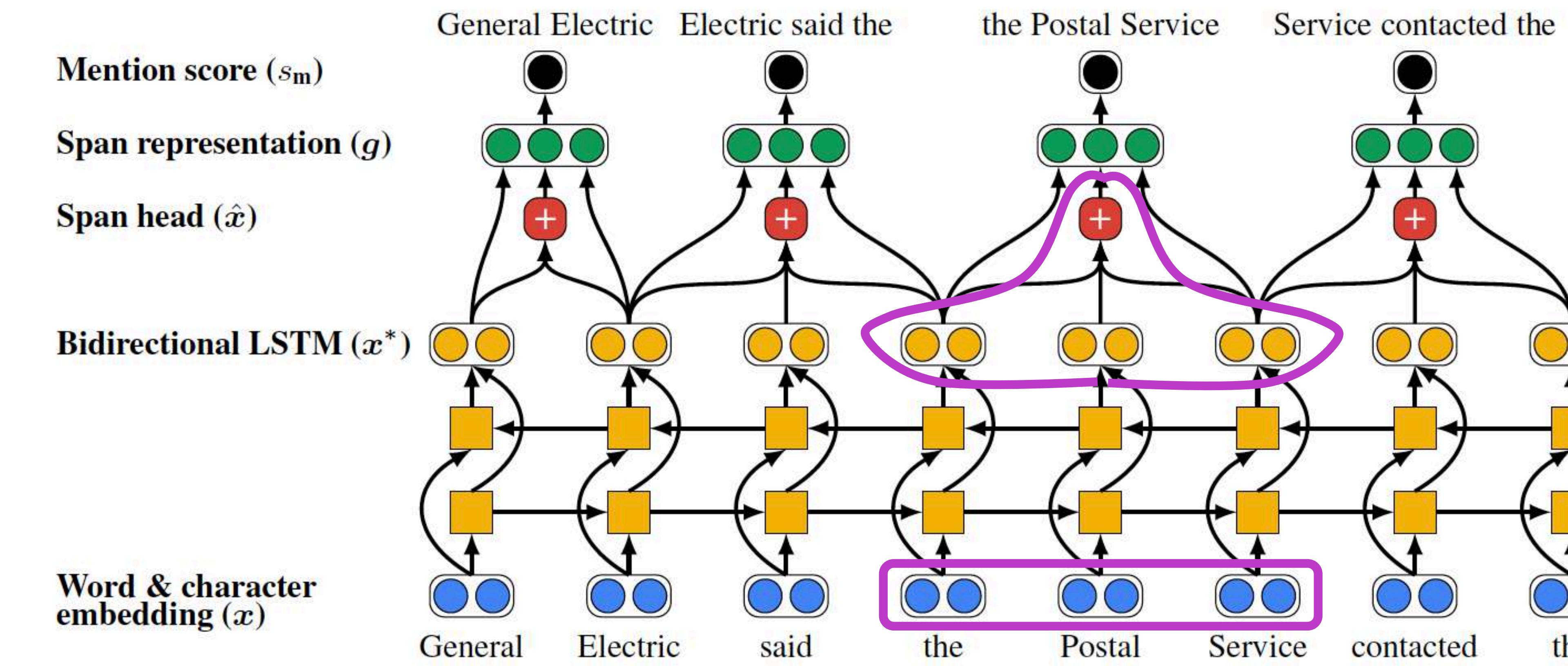
Span representation: $\boldsymbol{g}_i = [\boldsymbol{x}^*_{\mathrm{START}(i)}, \boldsymbol{x}^*_{\mathrm{END}(i)}, \hat{\boldsymbol{x}}_i, \phi(i)]$

or contextual embeddings from encoder)

BILSTM hidden states for span's start and end

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

Additional features



filter-style attention

Attention scores

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

dot product of weight vector and transformed hidden state

or contextual embeddings of token

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?

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

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

again, we have some extra features

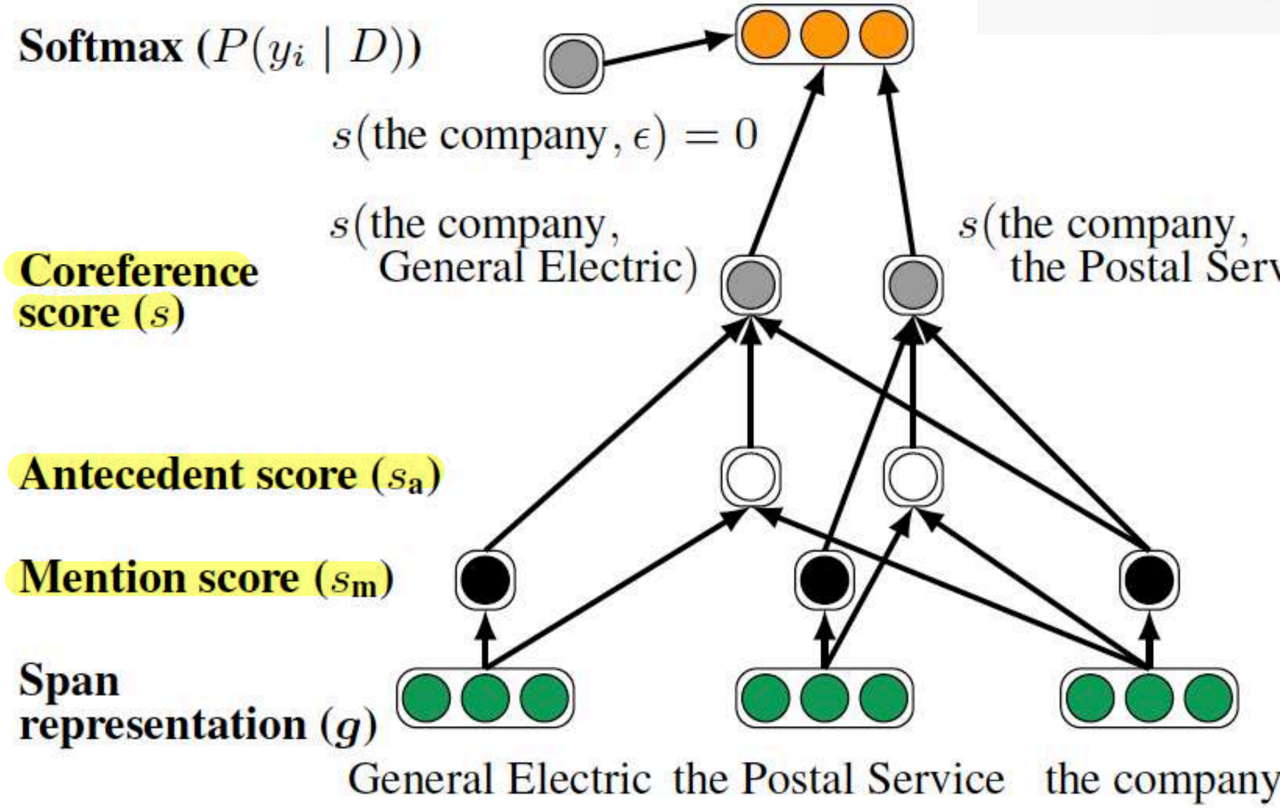
$\circ$ is element-wise multiplication

[3]

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

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

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

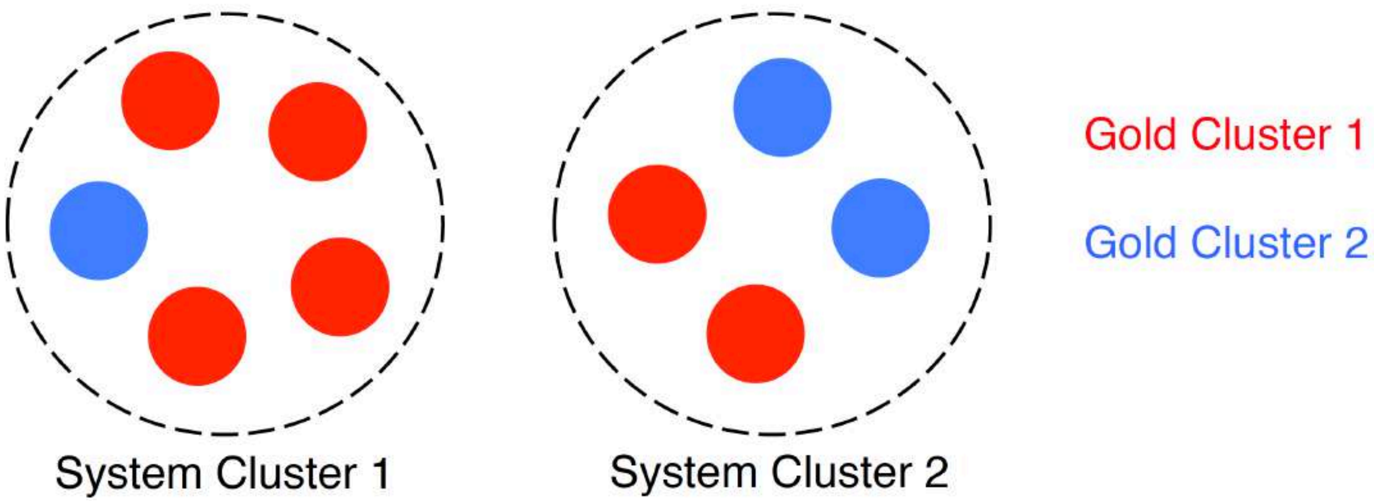


[3]

- Intractable 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 Evaluation

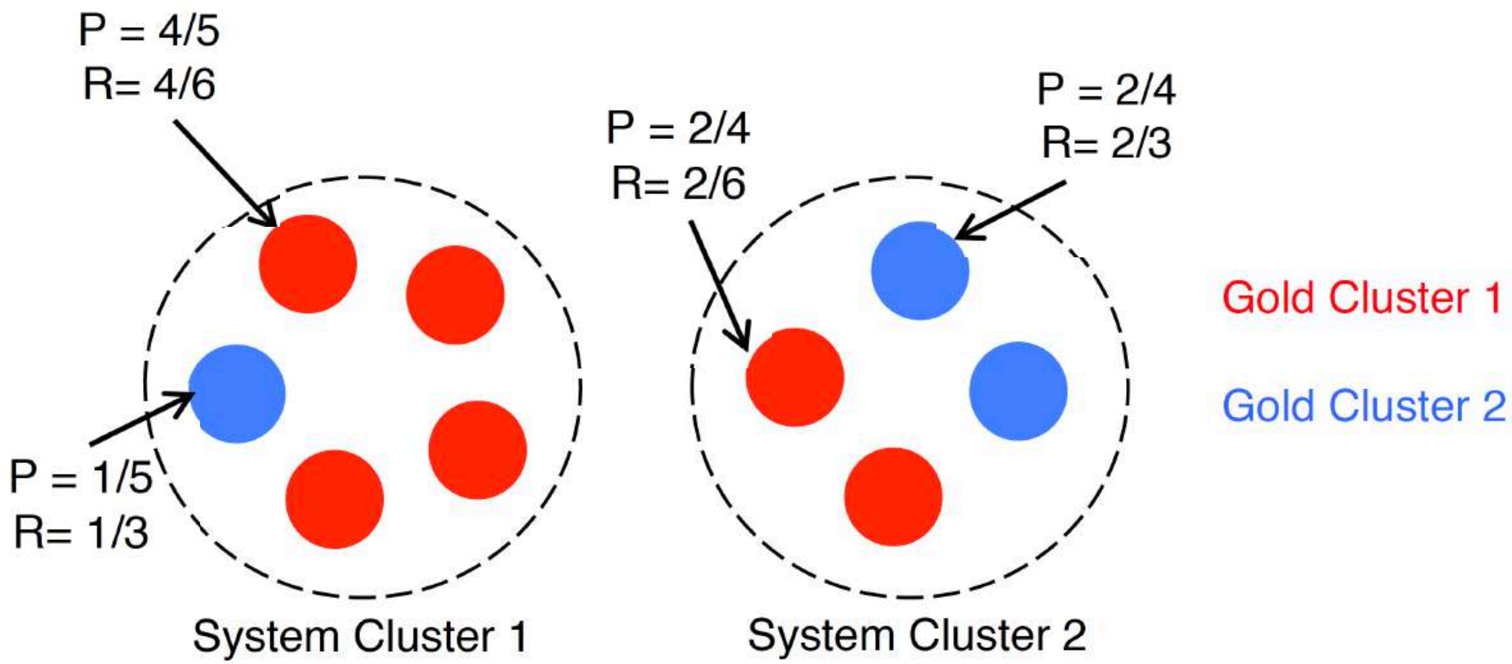
- many different metrics: MUC, CEAF, LEA, B-CUBED, BLANC → often report the average over a few different metrics



Coreference Evaluation

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

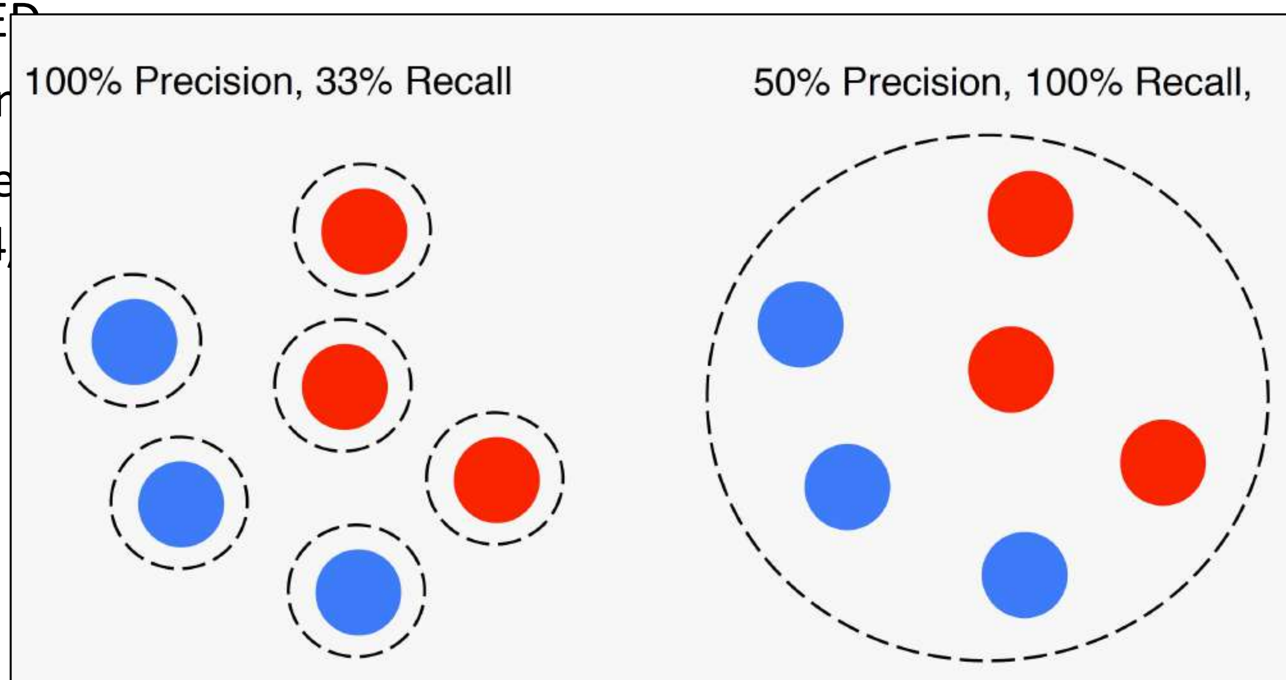
    P = [4(4/5) + 1(1/5) + 2(2/4) + 2(2/4)] / 9 = 0.6

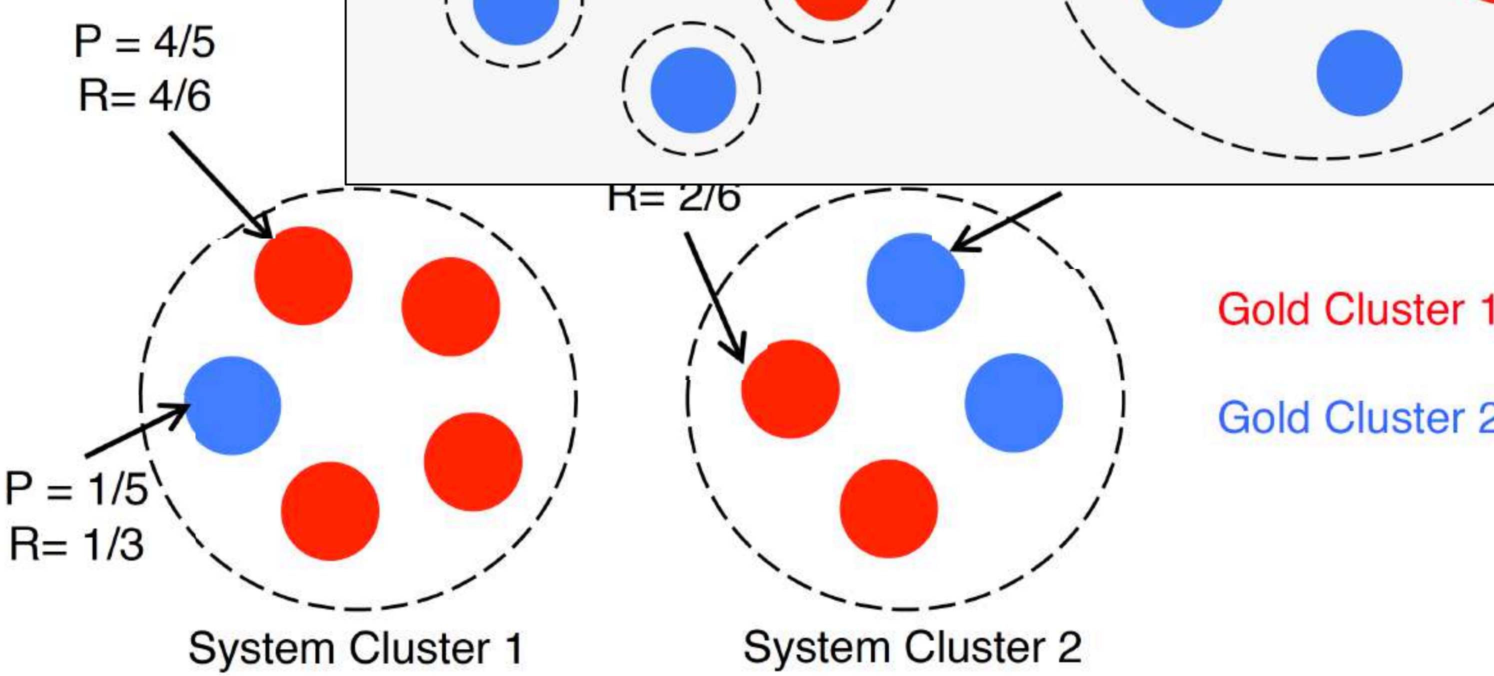


Coreference Evaluation

- example B-CUBED
  - for each men
  - then average

    P = [4(4





Gender Bias

The secretary called the physician$_i$ and told him$_i$ about a new patient [pro-stereotypical]

The secretary called the physician$_i$ and told her$_i$ about a new patient [anti-stereotypical]

Bibliography

(1) Dan Jurafsky and James Martin: Speech and Language Processing (3rd ed. draft, version Jan 2023); Online: https://web.stanford.edu/~jurafsky/slp3/ (URL, Oct 2023); this slide-set is especially based on chapter 26

(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.

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 pages in Jurafsky [1]

- interested students

  == standard approach