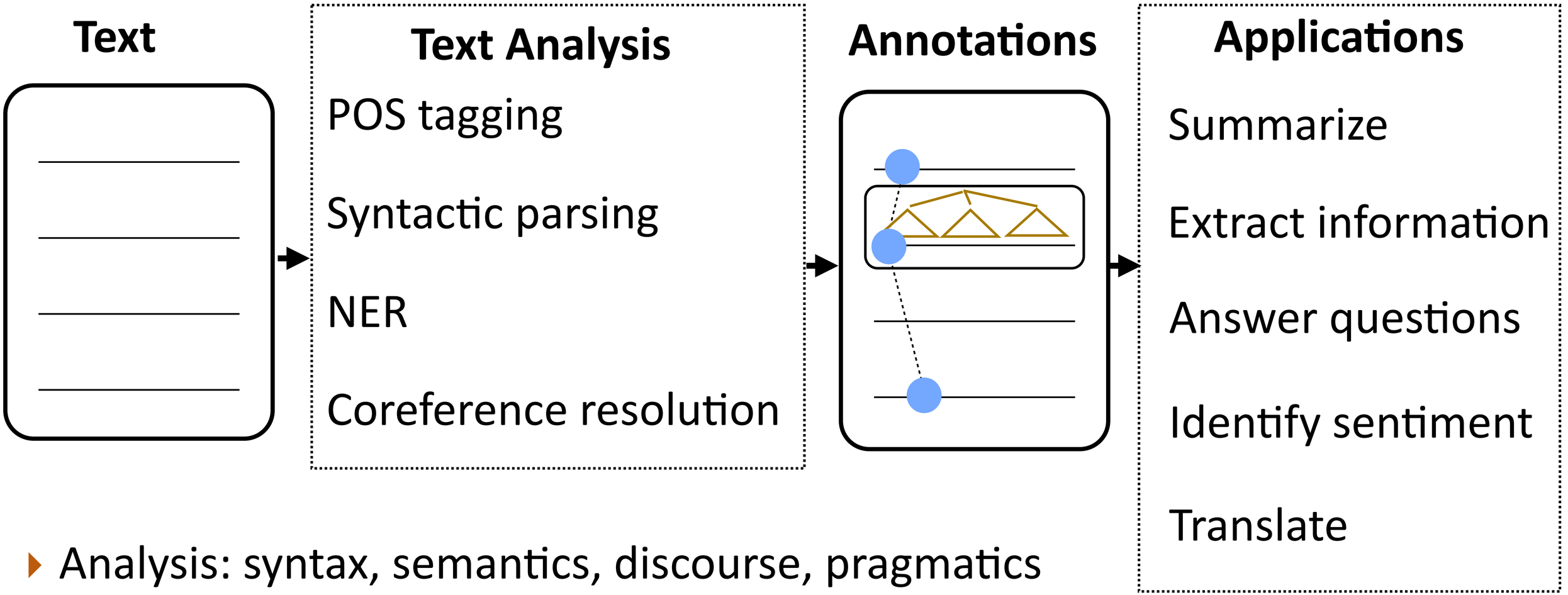
CS388: Natural Language Processing
Coreference Resolution

Greg Durrett

Road Map



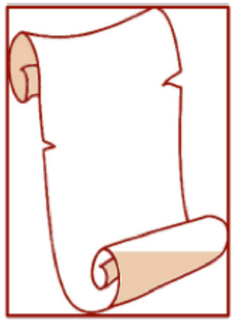
- Analysis: syntax, semantics, discourse, pragmatics
- Coreference: discourse + pragmatics

Discourse Analysis

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

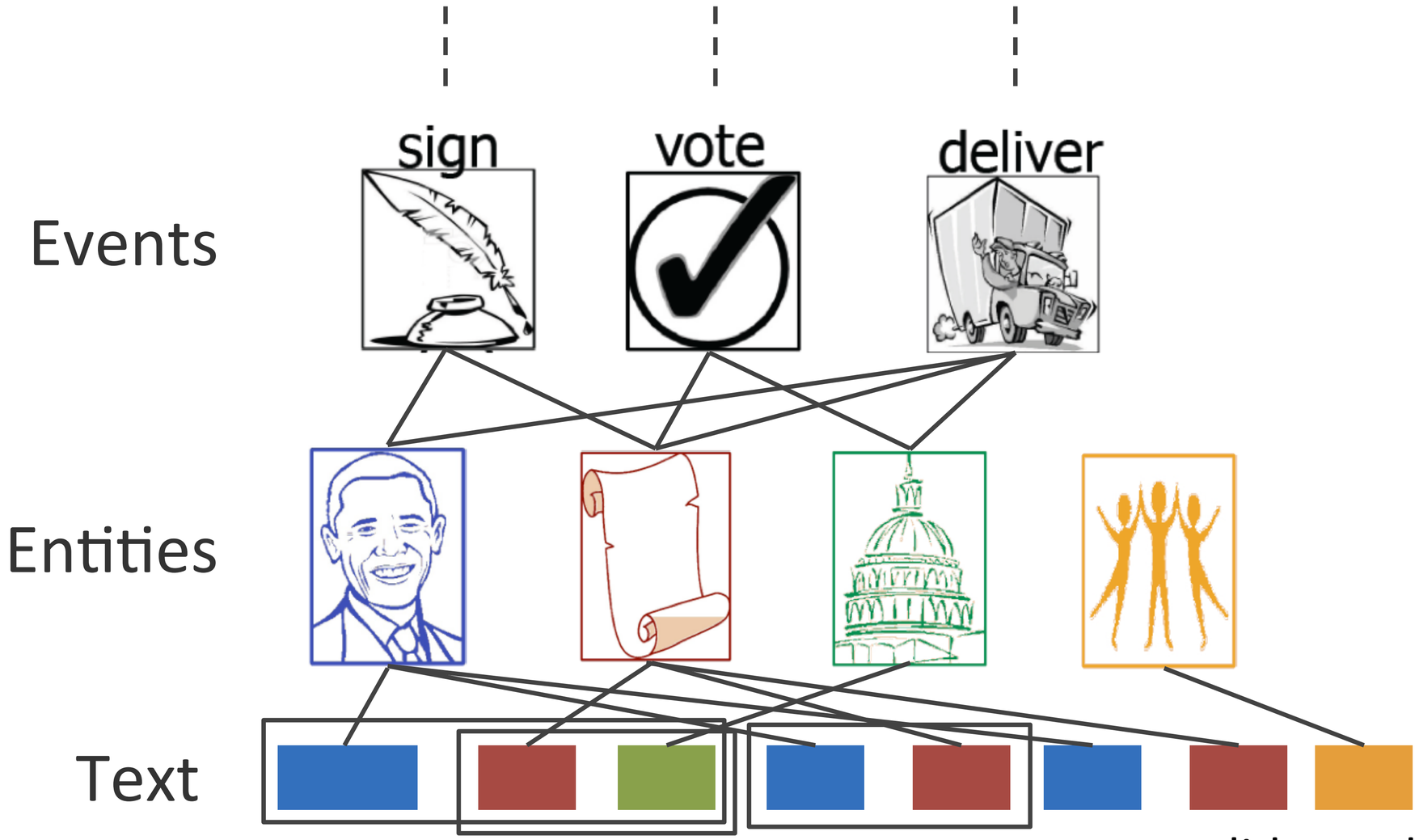
Discourse Analysis

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

slide credit: Aria Haghighi

Discourse Analysis



slide credit: Aria Haghighi

Entities

President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

- Entities are real-world things that can be resolved to an entry in a knowledge base (Wikipedia), can repeatedly reference them in a text

Cluster 1: `en.wikipedia.org/wiki/Barack_Obama`

Cluster 2: `…/wiki/Edward_M.Kennedy_Serve_America_Act`

Cluster 3: `…/wiki/United_States_Congress`

Coreference Resolution

- Input: text with mentions

  President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

- Output: a clustering of those mentions

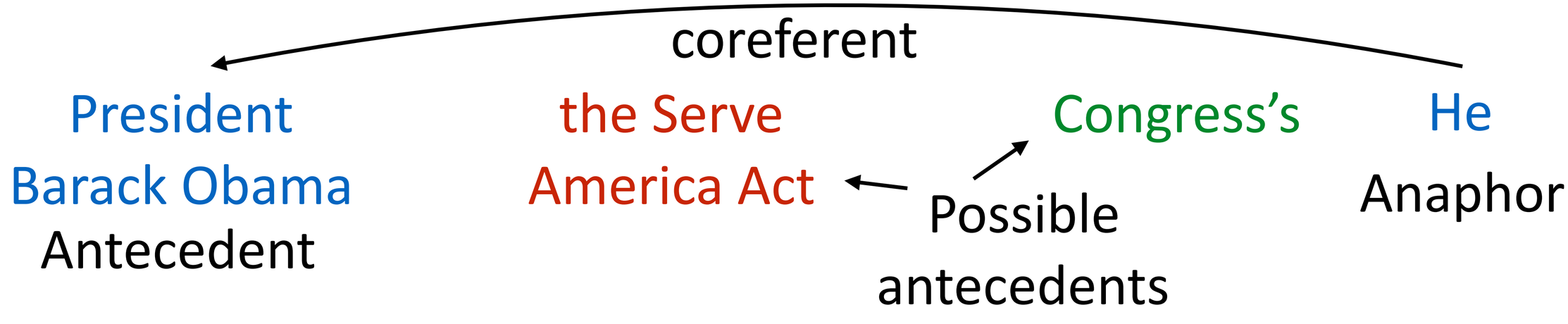
  President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

Coreference Resolution

- Input: text with mentions

  President Barack Obama received the Serve America Act after Congress's vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

- Alternatively: answer "who is my antecedent?" for each anaphor



Outline

- Linguistic phenomena in coreference
- Building coreference models
- Incorporating world knowledge

Phenomena in Coreference

Pragmatics 101

President Barack Obama received the Serve America Act after Congress's vote.

President Barack Obama signed the Serve America Act last Thursday.

President Barack Obama said…

President Barack Obama received the Serve America Act after Congress's vote.

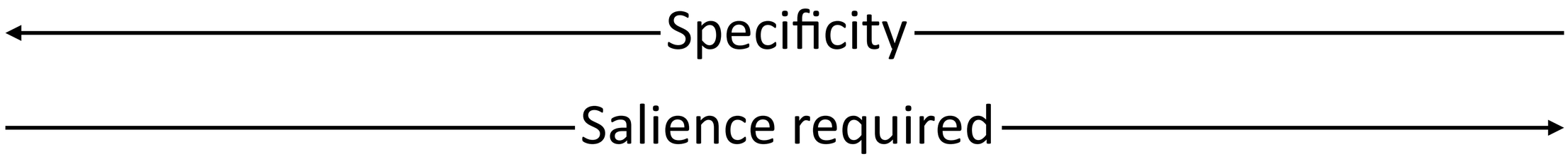
He signed the bill last Thursday.

The president said…

- When we speak/write, we have an idea of what's clear to the listener, and communicate more efficiently as a result

Pragmatics 101

Proper Name	Nominal	Pronoun
President Barack Obama	the president	he



President Barack Obama received the Serve America Act after Congress's vote.
He signed the bill last Thursday.
The president said…

- Proper, nominal, and pronominal mentions all resolve differently

Proper Mentions

- Introduce new entities and give information, identity entities unambiguously (mostly)

  President Barack Obama, 44th president of the United States, …

  President Obama

  Obama

- When might there be ambiguity?

  Dell founded what would become his eponymous company in 1984.

  Dell was later taken private in a leveraged buyout.

- **Main cues: lexical overlap, semantic type agreement**

Pronouns

President Barack Obama received the Serve America Act after Congress's vote. *He* …

President Obama met with Chancellor Merkel. *He* …

The policeman ticketed the driver after *he* noticed a broken taillight

he ran the stop sign

This is the house where the bomb was built into the boat that carried *it*.

- **Main cues: salience, number/gender agreement, event semantics/ commonsense knowledge**

Nominal Mentions

President Obama … *The president* …

Serve America Act … *The bill*

- Basic lexical semantics/hypernymy

NBC … *The network*

- World knowledge

Barack Obama and Angela Merkel … *The leaders*

- Combines the two: Obama is a president, Merkel is a chancellor, the common type of those is leader

- **Main cues: semantic type agreement/world knowledge, salience**

Phenomena

- Salience: distance features
- Semantic compatibility
  - Gender: he vs. she
  - Animacy: he/she vs. it
  - Semantic type: Michael Dell (person) vs. Dell (company)
  - Hypernymy: an act is a bill
  - Commonsense knowledge: a bomb can be carried, a boat cannot be
  - World knowledge: Merkel is a leader
- Coreference is a challenging NLP problem! Several different subproblems, lots of sources of information that we need to consider

Building Coreference Models

Rule-based Systems

- Filter possible antecedents based on syntactic and semantic information, resolve to the closest one

President
Barack Obama

~~the Serve~~
~~America Act~~
- inanimate

~~Congress's~~
- inanimate

He

- Semantic information used: number and gender (automatically scraped), head word / string match, some world knowledge (NBC = network)

Haghighi and Klein (2008)

Entity-centric Ruled-based Systems

FEMALE
Michelle Obama promoted her fitness and nutrition program on Thursday.

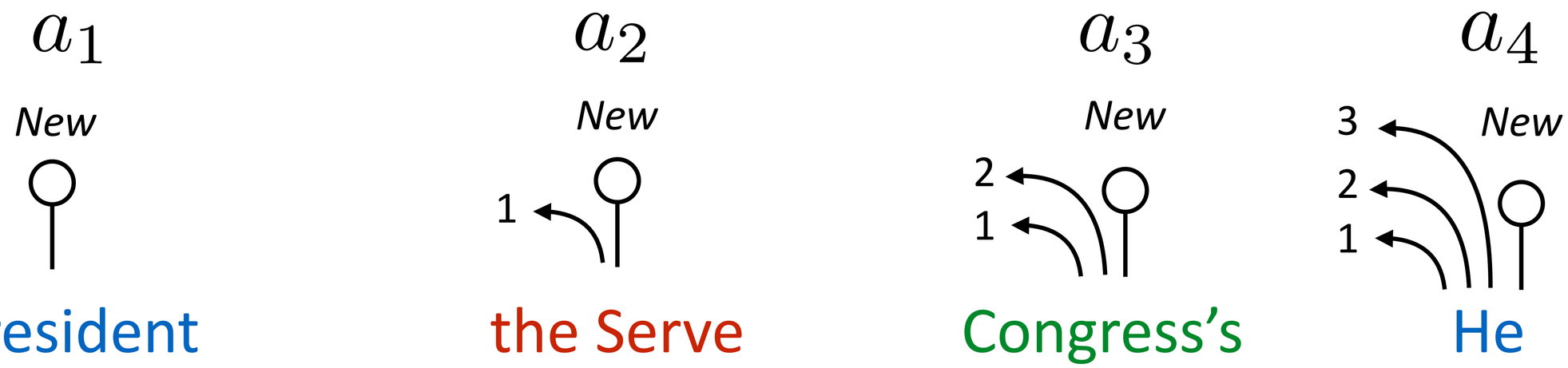
FEMALE
Obama gave a speech on the "Let's Move!" program, praising Sam Kass.

He…

- Coreference depends on identity of Obama, which in turn depends on other coreference links
- Need to make decisions globally: entity-centric, "sieve-based" coreference, "easy-first" systems all rely on earlier decisions to do this

Rahman and Ng (2009), Raghunathan et al. (2010), Lee et al. (2011)

Mention-Ranking Systems



- Log-linear model

$$p(a_{\boxed{i}} = \boxed{j} \,\|\, \boxed{x}) \propto \exp(w^\top \boxed{f(i, j, x)})$$

document

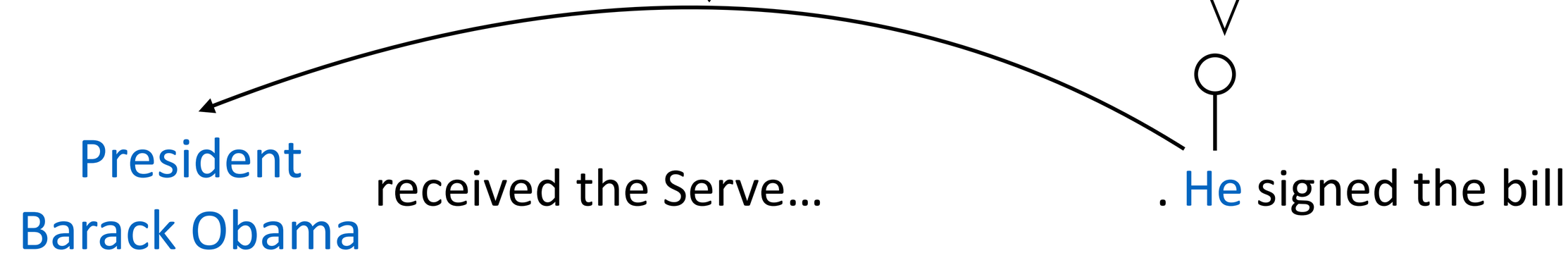
anaphor index: i — antecedent index: j — features of mention pair + document: $f(i, j, x)$

Denis and Baldridge (2008), Fernandes et al. (2012), Durrett and Klein (2013)

Features for Learning-based Systems

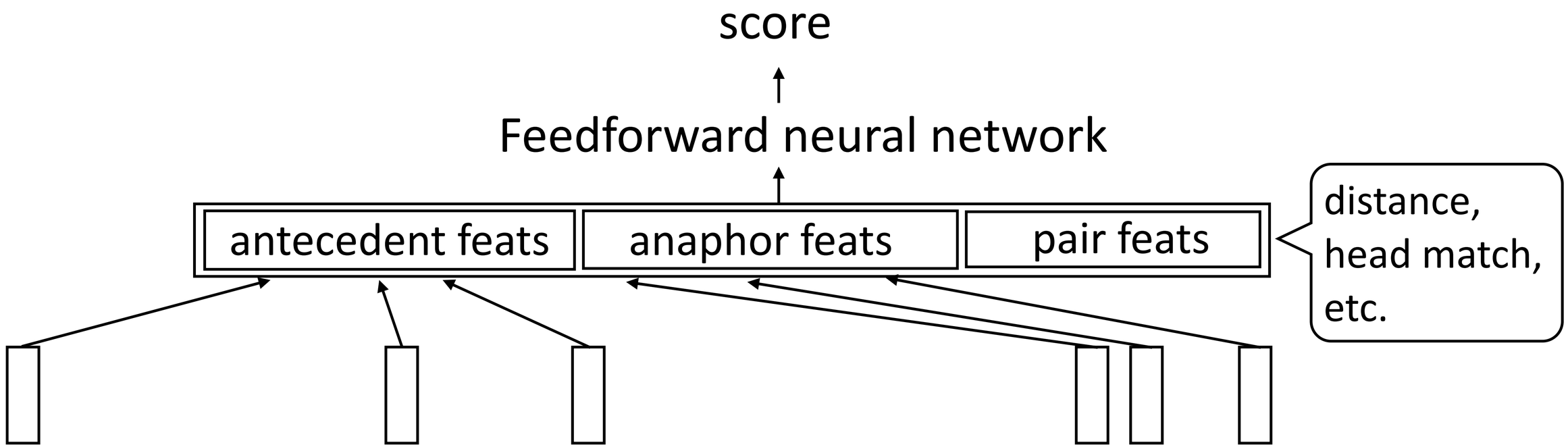
Ment. distance = 3	Sent. distance = 1	Salience
Antecedent length = 3	Anaph length = 1	Pragmatics
No head match	No string match	Semantic compatibility
MALE—he	*Obama*—he	
X *received*—he	PROPER—X *signed*	

[new] PRONOUN
[new] *he*
[new] X *signed*
[new] . X
[new] Length = 1



Denis and Baldridge (2008), Fernandes et al. (2012), Durrett and Klein (2013)

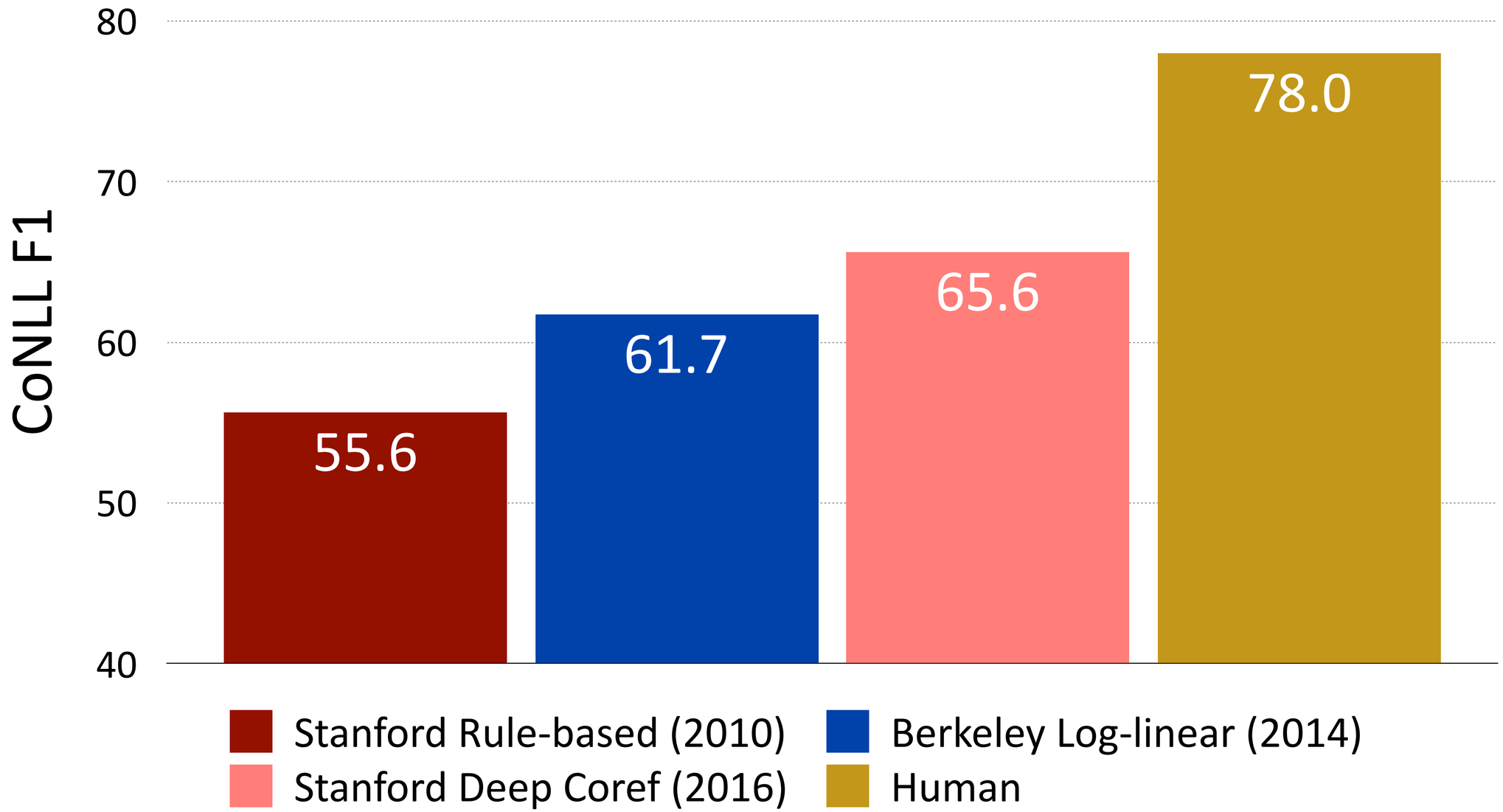
Neural Network Models



- Similar inputs to log-linear model
- Word embeddings + nonlinear layers capture more complex interactions between mention and antecedent

Clark and Manning (2016)

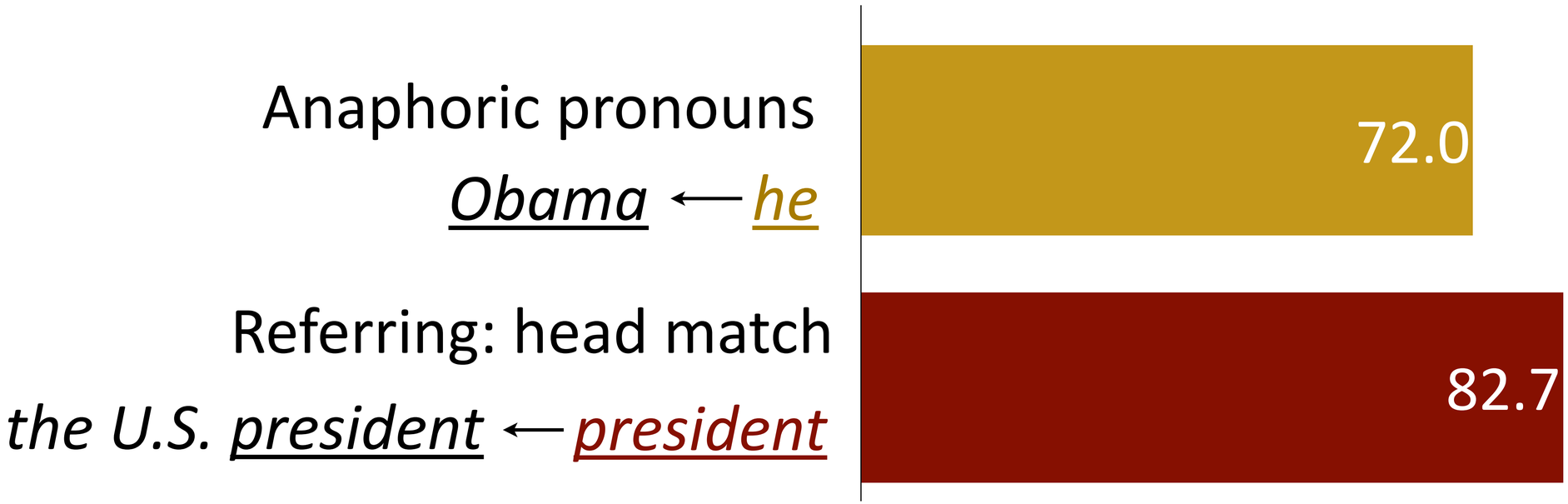
Performance



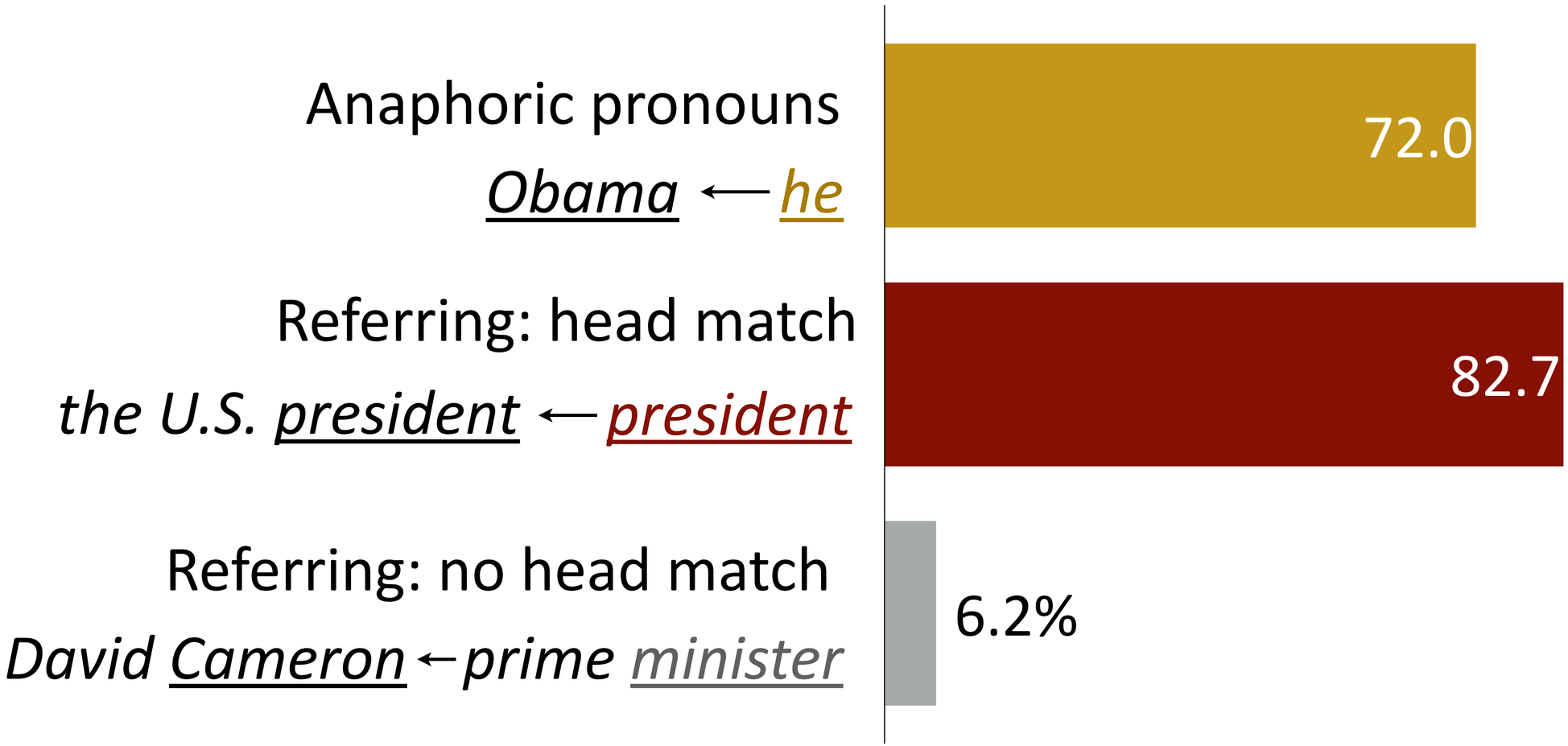
Incorporating World Knowledge

Accuracy Per Mention Class (Berkeley)

Accuracy Per Mention Class (Berkeley)



Accuracy Per Mention Class (Berkeley)



Phenomena

- Salience ✓
- Semantic compatibility
  - Gender ✓
  - Animacy ✓
  - Semantic type (✓)
  - Hypernymy (✓)
  - Commonsense knowledge
  - World knowledge

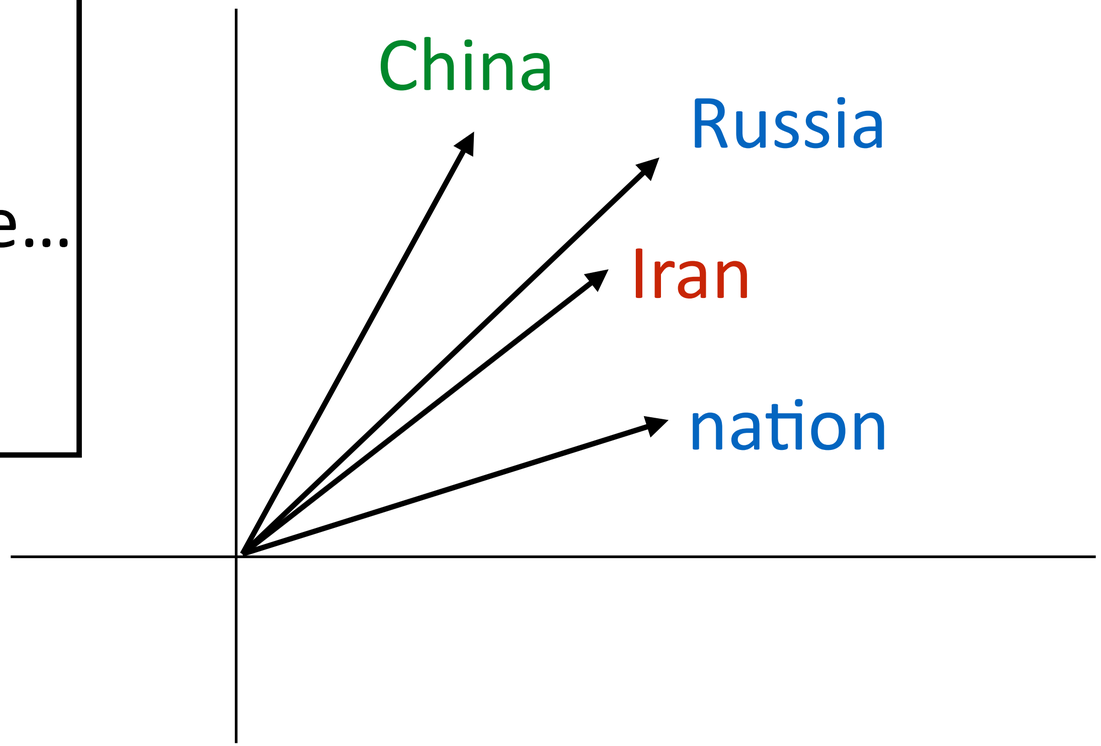
- Basic features get these
- Word embeddings sort of do these

Word Embeddings

Russia 's economy has been sluggish…

…suspected collusion with Russia . The…

…a trip to Russia in the springtime



- Russia is not Iran! Possibly compatible pairs are less similar than many incompatible pairs
- Word vectors capture *topical similarity*, are not trained to capture *referential identity*

Phenomena

- Salience ✓
- Semantic compatibility
  - Gender ✓
  - Animacy ✓
  - Semantic type (✓)
  - Hypernymy (✓)
  - Commonsense knowledge ✗
  - World knowledge ✗

- Basic features get these
- Word embeddings sort of do these
- …but they don't do these

Leveraging External Resources

- How do we figure out what kind of thing NBC is?
- Use an external knowledge base like Wikipedia
- Knowledge can import the features needed to make difficult coreference decisions

NBC

From Wikipedia, the free encyclopedia

This article is about the television and radio netwo
Broadcasting Corporation. For other uses, see NE

The **National Broadcasting Company** (**NBC**) is an American English language commercial broadcast television network that is a flagship property of NBCUniversal, a subsidiary of Comcast. The

Joint Entity Linking and Coreference

- There are many things NBC could mean!
- Need to tackle *entity linking* as well: figuring out what entity a given occurrence of NBC refers to
- Joint models resolve entities to Wikipedia and simultaneously place coreference links (Durrett and Klein, 2014)
- Improvement from entity linking is small: ~1% on CoNLL metric

NBC (disambiguation)

From Wikipedia, the free encyclopedia

NBC is a television broadcast network in the United States and the Philippines.

NBC may also stand for:

Art, entertainment, and media [edit]

Broadcasting [edit]

- NBC PNG, Papua New Guinea
- Namibian Broadcasting Corporation, Namibia
- Nation Broadcasting Corporation, Philippines
- Newfoundland Broadcasting Company, Canada, now CJON-DT
- Norwegian Broadcasting Corporation, Norway

Challenge: Need Complex Inferences

Russia's economy has been sluggish… The Eastern European nation …

Article Talk

Russia

From Wikipedia, the free encyclopedia

PRINCETON UNIVERSITY

WordNet
A lexical database for English

Russia …is a country in northeast Eurasia.

country — state — land — nation

country — rural area

Conclusion

- Coreference is a challenging NLP problem
- Many phenomena to capture, including salience and semantic compatibility
- Mention-ranking classifiers work pretty well (non-neural or neural)
- World knowledge is needed to solve many remaining errors, but is hard to incorporate