

# Lecture 19

## Coreference and Entity Resolution

Intro to NLP, CS585, Fall 2015  
Brendan O'Connor

- **Within-sentence NLP we've seen so far**
  - **Parts of speech, named entities, syntactic parse trees, sentence-level machine translation**
- **Cross-sentence NLP? (*Discourse*)**
  - **Next: noun phrase coreference, just one issue in discourse**

# Noun phrase reference



Barack Obama nominated Hillary Rodham Clinton as his secretary of state. He chose her because she had foreign affairs experience.

Referring expressions reference discourse entities  
e.g. real-world entities

# Noun phrase reference



Barack Obama nominated Hillary Rodham Clinton as his secretary of state. He chose her because she had foreign affairs experience.

Referring expressions reference discourse entities  
e.g. real-world entities

# Noun phrase reference

[http://harrypotter.wikia.com/wiki/Harry\\_Potter](http://harrypotter.wikia.com/wiki/Harry_Potter)

Harry James Potter (b. 31 July, 1980) was a half-blood wizard, the only child and son of James and Lily Potter (née Evans), and one of the most famous wizards of modern times ... Lord Voldemort attempted to murder him when he was a year and three months old ...

Referring expressions reference discourse entities  
e.g. real-world entities  
(... or non-real-world)

# Terminology

[http://harrypotter.wikia.com/wiki/Harry\\_Potter](http://harrypotter.wikia.com/wiki/Harry_Potter)

Harry James Potter (b. 31 July, 1980) was a half-blood wizard, the only child and son of James and Lily Potter (née Evans), and one of the most famous wizards of modern times ... Lord Voldemort attempted to murder him when he was a year and three months old ...

an **Entity** or **Referent** is a ~real-world object  
(“HARRY\_POTTER\_CONCEPT”)

**Referring expressions** a.k.a. **Mentions**

14 NPs are underlined above (are they all referential?)

**Coreference**: when referring mentions have the same referent.

**Coreference resolution**: find which mentions refer to the same entity.

I.e. cluster the mentions into **entity clusters**.

Applications: text inference, search, etc.

- Who tried to kill Harry Potter?

# Reference Resolution

- Noun phrases refer to entities in the world, many pairs of noun phrases co-refer, some nested inside others

John Smith, CFO of Prime Corp. since 1986,

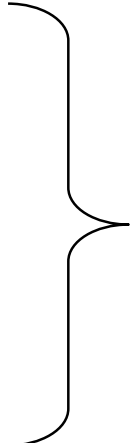
saw his pay jump 20% to \$1.3 million

as the 57-year-old also became

the financial services co.'s president.

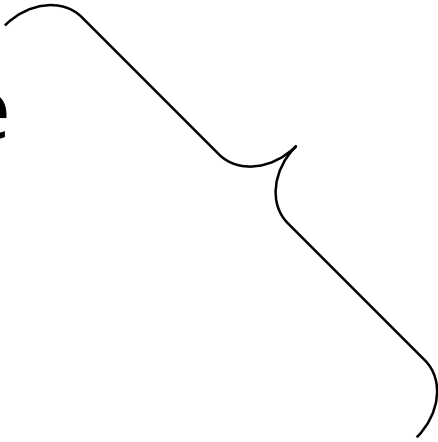
# Kinds of Reference

- Referring expressions
  - *John Smith*
  - *President Smith*
  - *the president*
  - *the company's new executive*



More common in  
newswire, generally  
harder in practice

- Free variables
  - Smith saw *his pay* increase



More interesting  
grammatical  
constraints,  
more linguistic  
theory, easier in  
practice

- Bound variables
  - The dancer hurt *herself*.

“anaphora  
resolution”



# Syntactic vs Semantic cues

- State-of-the-art coref uses with the first three






# Syntactic vs Semantic cues

- Lexical cues
    - I saw a house. The house was red.
    - I saw a house. The other house was red.
  - Syntactic cues
    - John bought himself a book.
    - John bought him a book.
  - Shallow semantic cues
    - John saw Mary. She was eating salad.
    - John saw Mary. He was eating salad.
- 
- State-of-the-art coref uses with the first three

# Syntactic vs Semantic cues

- Lexical cues
  - I saw a house. The house was red.
  - I saw a house. The other house was red.
- Syntactic cues
  - John bought himself a book.
  - John bought him a book.
- Shallow semantic cues
  - John saw Mary. She was eating salad.
  - John saw Mary. He was eating salad.
- Deeper semantics (world knowledge)
  - The city council denied the demonstrators a permit because they feared violence.
  - The city council denied the demonstrators a permit because they advocated violence.
- State-of-the-art coref uses with the first three

# Supervised ML: Mention pair model

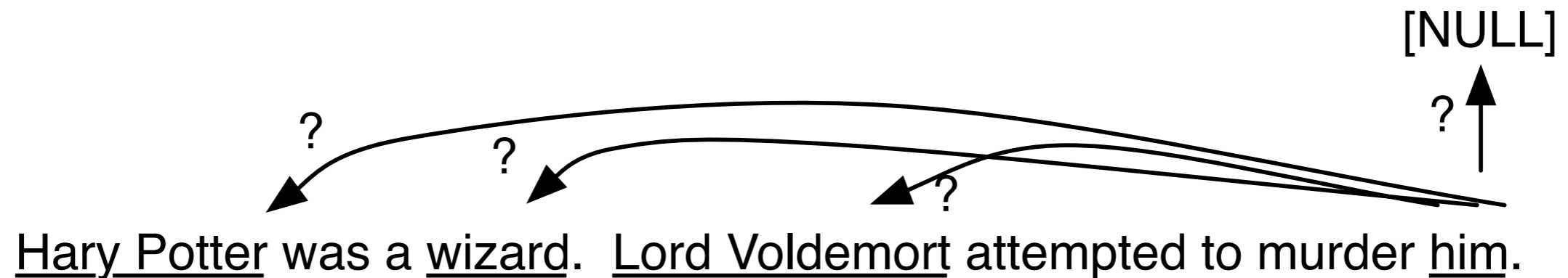


Hary Potter was a wizard. Lord Voldemort attempted to murder him.

The diagram consists of three arcs above the text. The first arc connects 'Hary Potter' to 'wizard'. The second arc connects 'Lord Voldemort' to 'him'. The third arc connects 'Hary Potter' to 'him', spanning across the first two arcs.

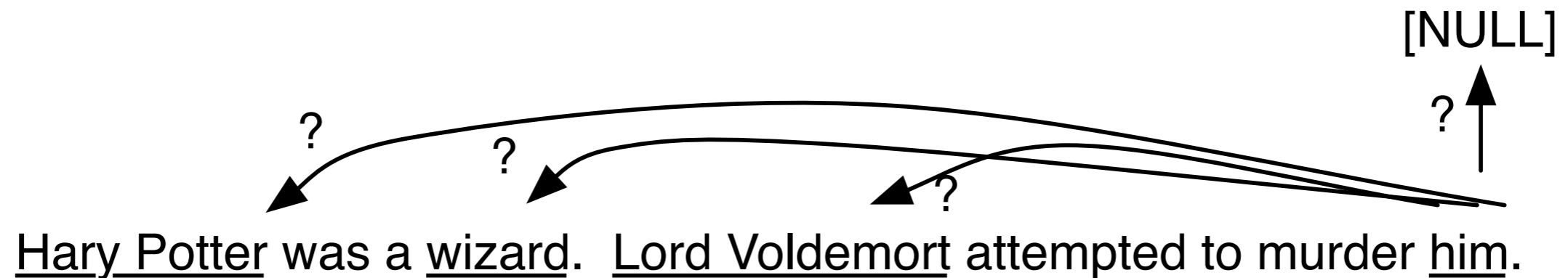
- View gold standard as defining links between mention pairs
- Think of as binary classification problem: take random pairs as negative examples
- Issues: many mention pairs. Also: have to resolve local decisions into entities

# Antecedent selection model



- View as antecedent selection problem: which previous mention do I corefer with?
- Makes most sense for pronouns, though can use model for all expressions
- Process mentions left to right. For the  $n$ 'th mention, it's a  $n$ -way multi-class classification problem: antecedent is one of the  $n-1$  mentions to the left, or NULL.
- Features are asymmetric!
- Use a limited window for antecedent candidates, e.g. last 5 sentences (for news...)
- Score each candidate by a linear function of features. Predict antecedent to be the highest-ranking candidate.

# Antecedent selection model



- Training: simple way is to process the gold standard coref chains (entity clusters) into positive and negative links. Train binary classifier.
- Prediction: select the highest-scoring candidate as the antecedent. (Though multiple may be ok.)
- Using for applications: take these links and form entity clusters from connected components [whiteboard]



# Features for pronoun resolution

- English pronouns have some grammatical markings that restrict the semantic categories they can match. Use as features against antecedent candidate properties.
  - Number agreement
    - he/she/it vs. they/them
  - Animacy/human-ness? agreement
    - it vs. he/she/him/her/his
  - Gender agreement
    - he/him/his vs. she/her vs. it
- Grammatical person - interacts with dialogue/discourse structure
  - I/me vs you/y'all vs he/she/it/they
- Reflexives
  - Bob knew that John bought himself a book.
  - John knew that Bob bought him a book.

# Other syntactic constraints

- High-precision patterns
  - Predicate-Nominatives: “X was a Y ...”
  - Appositives: “X, a Y, ...”
  - Role Appositives: “president Lincoln”

# Features for Pronominal Anaphora Resolution

- Preferences:
  - Recency: More recently mentioned entities are more likely to be referred to
    - John went to a movie. Jack went as well. He was not busy.
  - Grammatical Role: Entities in the subject position is more likely to be referred to than entities in the object 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.

# Features for Pronominal Anaphora Resolution

- Preferences:
  - Verb Semantics: Certain verbs seem to bias whether the subsequent pronouns should be referring to their subjects or objects
    - John telephoned Bill. He lost the laptop.
    - John criticized Bill. He lost the laptop.
  - Selectional Restrictions: Restrictions because of semantics
    - John parked his car in the garage after driving it around for hours.
- Encode all these and maybe more as features

# Features for non-pronoun resolution

- Generally harder!
  - String match
  - Head string match
    - I saw a green house. The house was old.
  - Substring match...