Semantics

- The study of language meaning

This is linguistics. You can study any aspect of language you want, and you will always find structure

But what is that shadowy place over there?

That is semantics. You must never go there

http://lolphonology.tumblr.com/

Wednesday, November 12, 14
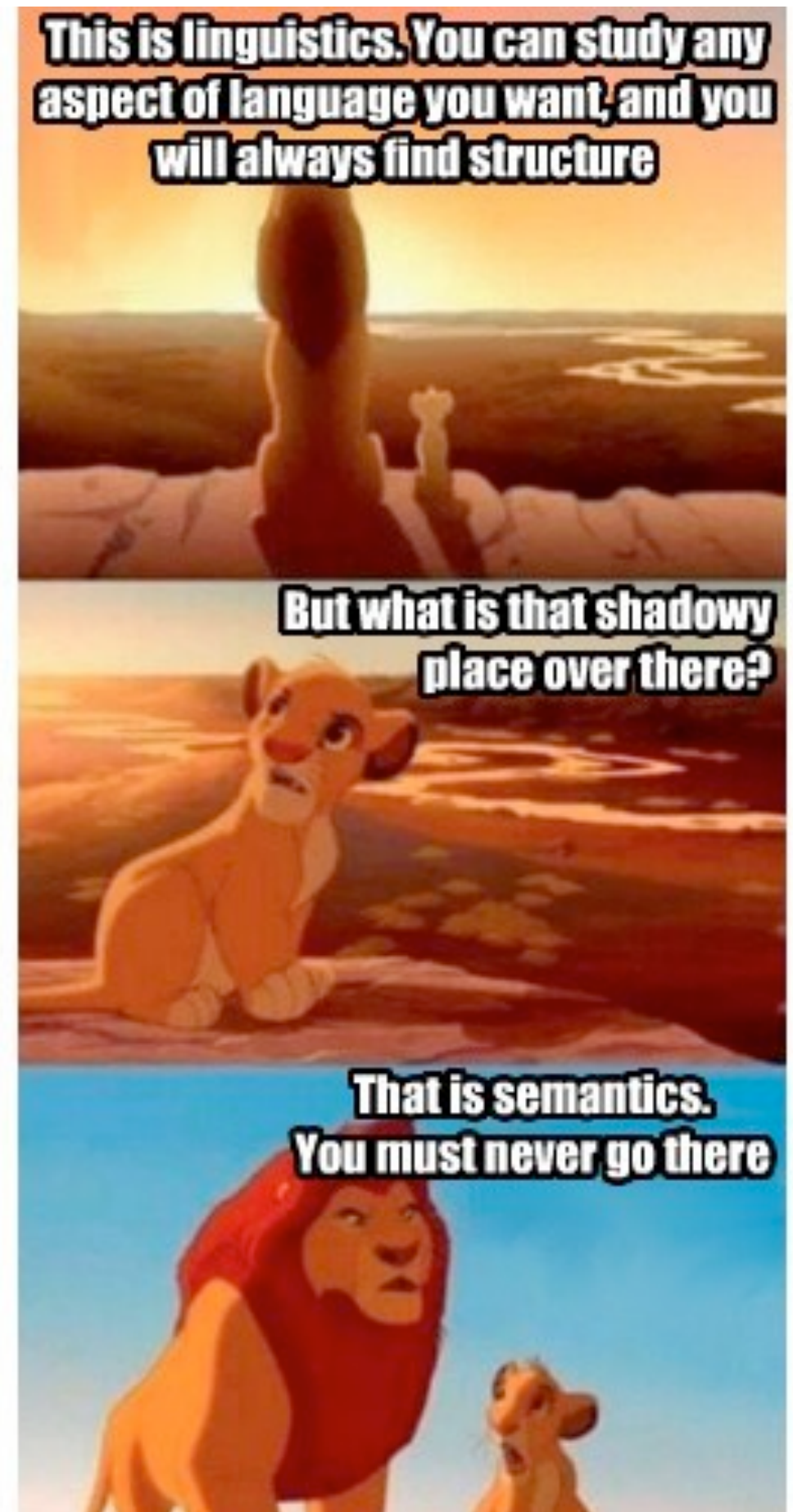
Lecture 19
Coreference and Entity Resolution

Intro to NLP, CS585, Fall 2014
http://people.cs.umass.edu/~brenocon/inlp2014/
Brendan O'Connor

*[Some slides borrowed from Luke Vilnis
and Chris Manning]*

2

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

3

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

4

Noun phrase reference

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)

5

Terminology

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 (discourse entity) ("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.

Exercise

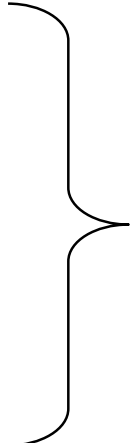Do within-document coreference in the following document by assigning the mentions entity numbers:

[The government]___ said [today]___ [it]___ 's going to cut back on [[[the enormous number]___ of [people]___]___ who descended on [Yemen]___ to investigate [[the attack]___ on [the " USS Cole]___]___]___. " [[[So many people]___ from [several agencies]___]___]___ wanting to participate that [the Yemenis]___ are feeling somewhat overwhelmed in [[their]___ own country]___. [Investigators]___ have come up with [[another theory]___ on how [the terrorists]___ operated]___. [[ABC 's]___ John Miller]___ on [[the house]___ with [a view]___]___. High on [[a hillside]___, in [[a run - down section]___ of [Aden]___]___]___, [[the house]___ with [the blue door]___]___ has [[a perfect view]___ of [the harbor]___]___. [American and Yemeni investigators]___ believe [that view]___ is what convinced [[a man]___ who used [[the name]___ [Abdullah]___]___]___ to rent [the house]___ [several weeks]___ before [[the bombing]___ of [the " USS Cole]___]___. " Early on [investigators]___ theorized [it]___ was [an inside job]___ and [[much]___ of [the focus]___]___ was on [[employees]___ of [[the Mansoon shipping company]___, which was under [[contract]___ by [the Navy]___ to refuel [U.S. warships]___]___ and would have had [[advance information]___ about [[the " Cole 's "]___ arrival]___]___]___]___. Now [the FBI]___ believes [[all]___ [the terrorists]___ needed to do]___ was look out [the window]___, to go through [precisely the same drill]___, well before [the " Cole "]___ [arrived]___. [[The man]___ in [this house]___]___ would have had [[plenty]___ of [[time]___ to signal [[two bombers]___ waiting with [the boat]___ across [the bay]___]___]___]___. [Investigators]___ say [[clues]___ collected over [the last few days]___]___ have already pointed [them]___ to [[locations]___ both near and far outside [[the port city]___ of [Aden]___]___]___, but [they]___ wo n't say [there]___ 's [any indication that [[the plot]___ here]___ goes beyond [[Yemen 's]___ boarders]___]___. Learning [[the true identities]___ of [[those]___ involved in [the bombing]___]___]___ would help answer [that question]___, but [the two suicide bombers]___ died in [the attack]___, and after [the explosion]___, [[the man]___ who lived behind [the blue door]___]___ simply vanished. [John Miller]___, [ABC News]___, [Aden]___.

Do within-document coreference in the following document by assigning the mentions entity numbers:
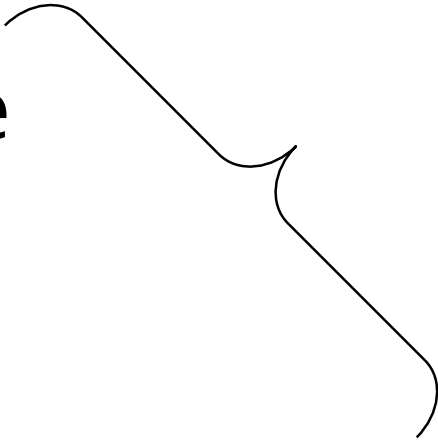
[The government]___ said [today]___ [it]___ 's going to cut back on [[[the enormous number]___ of [people]___]___ who descended on [Yemen]___ to investigate [[the attack]___ on [the " USS Cole]___]___]___. " [[[So many people]___ from [several agencies]___]___]___ wanting to participate that [the Yemenis]___ are feeling somewhat overwhelmed in [[their]___ own country]___. [Investigators]___ have come up with [[another theory]___ on how [the terrorists]___ operated]___. [[ABC 's]___ John Miller]___ on [[the house]___ with [a view]___]___. High on [[a hillside]___, in [[a run - down section]___ of [Aden]___]___]___, [[the house]___ with [the blue door]___]___ has [[a perfect view]___ of [the harbor]___]___. [American and Yemeni investigators]___ believe [that view]___ is what convinced [[a man]___ who used [[the name]___ [Abdullah]___]___]___ to rent [the house]___ [several weeks]___ before [[the bombing]___ of [the " USS Cole]___]___. " Early

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

- Bound variables
  - The dancer hurt *herself.*

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

  "anaphora resolution"

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

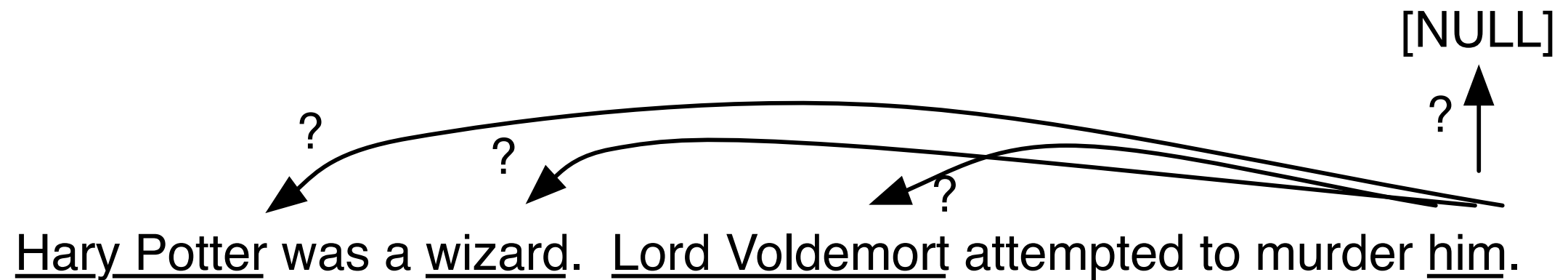Hary Potter was a wizard. Lord Voldemort attempted to murder him.

- 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

[NULL]

? ? ?

Hary Potter was a wizard. Lord Voldemort attempted to murder him.
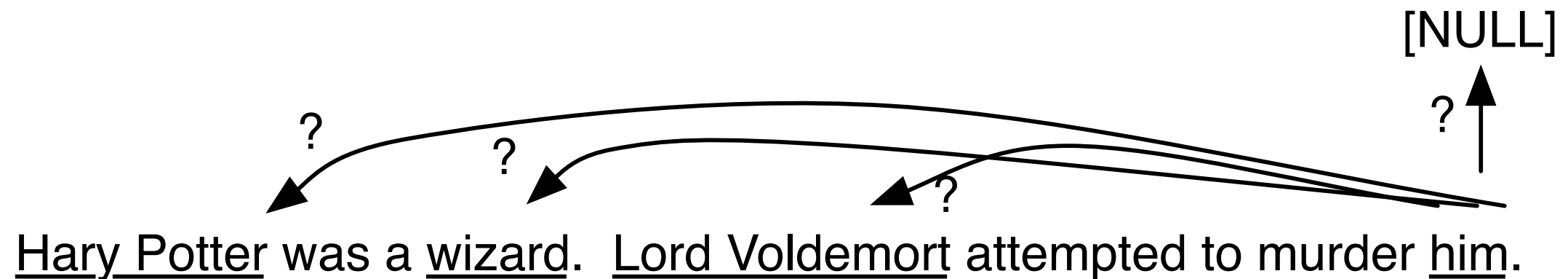
- View as antecedent selection problem: which previous mention do I corefer with?
  - Makes most sense for pronouns, though can use model for all expressions
- Process mentions left to right. For the n'th mention, it's a n-way multi-class classification problem: antecedent is one of the n-1 mentions 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.

13

Antecedent selection model

[NULL]

? ? ?
?

<u>Hary Potter</u> was a <u>wizard</u>. <u>Lord Voldemort</u> attempted to murder <u>him</u>.

- 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]

14

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 <u>himself</u> a book.
  - John knew that Bob bought <u>him</u> a book.

15

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 <u>green house</u>. The <u>house</u> was old.
  - Substring match...