

# Coreference resolution

Borrowing from Roger Levy, Christopher Manning, Dan Klein,  
and Andy Kehler

# Reference Resolution

- Noun phrases refer to entities in the world, many pairs of noun phrases co-refer:

John Smith, CFO of Prime Corp since 1986,

saw his pay jump 20% to \$1.3 million

as the 57-year-old also became

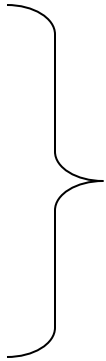
the financial services co.'s president.

# Applications

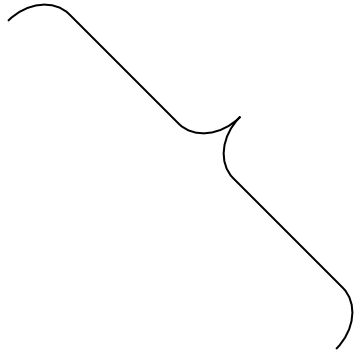
- detecting more named entity mentions
- entity-specific sentiment
- topic modeling / document clustering
- document summarization
- question answering

# Kinds of Reference

- Referring expressions
  - *John Smith*
  - *President Smith*
  - *the president*
  - *the company's new executive*
- Free variables
  - Smith saw *his pay* increase
- Bound variables
  - The dancer hurt *herself*.



More common in  
newswire, generally  
harder in practice



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

# Not all NPs are referring!

- *Every dancer* twisted *her knee*.
- (*No dancer* twisted *her knee*.)
- There are three NPs in each of these sentences; because the first one is non-referential, the other two aren't either.

# Features for Pronominal Anaphora Resolution

- Constraints:
  - Number agreement
    - Singular pronouns (it/he/she/his/her/him) refer to singular entities and plural pronouns (we/they/us/them) refer to plural entities
  - Person agreement
    - He/she/they etc. must refer to a third person entity
  - Gender agreement
    - He → John; she → Mary; it → car
    - Jack gave **Mary** a gift. **She** was excited.
  - Certain syntactic constraints
    - John bought **himself** a new car. [himself → John]
    - John bought **him** a new car. [him can not be John]

# 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.

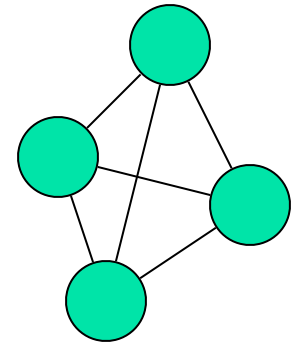
# Features for Pronominal Anaphora Resolution

- 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



# Model

- Mention Pair models
  - Treat coreference chains as a collection of pairwise links
  - Make independent pairwise decisions and reconcile them in some way (e.g. clustering or greedy partitioning)



# Mention Pair Models

- Most common machine learning approach
- Build classifiers over pairs of NPs
  - For each NP, pick a preceding NP or NEW
  - Or, for each NP, choose link or no-link
- Clean up non transitivity with clustering or graph partitioning algorithms
  - E.g.: [Soon et al. 01], [Ng and Cardie 02]
  - Some work has done the classification and clustering jointly [McCallum and Wellner 03]

# Pairwise Features

1. **strict gender** [true or false]. True if there is a strict match in gender (e.g. male pronoun  $Pro_i$  with male antecedent  $NP_j$ ).
2. **compatible gender** [true or false]. True if  $Pro_i$  and  $NP_j$  are merely compatible (e.g. male pronoun  $Pro_i$  with antecedent  $NP_j$  of unknown gender).
3. **strict number** [true or false] True if there is a strict match in number (e.g. singular pronoun with singular antecedent)
4. **compatible number** [true or false]. True if  $Pro_i$  and  $NP_j$  are merely compatible (e.g. singular pronoun  $Pro_i$  with antecedent  $NP_j$  of unknown number).
5. **sentence distance** [0, 1, 2, 3,...]. The number of sentences between pronoun and potential antecedent.
6. **Hobbs distance** [0, 1, 2, 3,...]. The number of noun groups that the Hobbs algorithm has to skip, starting backwards from the pronoun  $Pro_i$ , before the potential antecedent  $NP_j$  is found.
7. **grammatical role** [subject, object, PP]. Whether the potential antecedent is a syntactic subject, direct object, or is embedded in a PP.
8. **linguistic form** [proper, definite, indefinite, pronoun]. Whether the potential antecedent  $NP_j$  is a proper name, definite description, indefinite NP, or a pronoun.

# Coherence relations

- Adjacent sentence pairs come in multiple flavors:
- John hid Bill's car keys.
  - He was drunk. [Explanation; He=Bill?]
  - He was mad. ["Occasion" ; He=Bill?]
  - ??He likes spinach.