# Discourse Linguistics: Coreference Resolution

# Anaphora / Reference Resolution

- One of the most important NLP tasks for cohesion at the discourse level
- A linguistic phenomenon of abbreviated subsequent reference
  - A cohesive tie of the grammatical and lexical types
    - Includes reference, substitution and reiteration
  - A technique for referring back to an entity which has been introduced with more fully descriptive phrasing earlier in the text
  - Refers to this same entity but with a lexically and semantically attenuated form

# Types of Entity Resolutions

- **Entity Resolution** is an ability of a system to recognize and unify variant references to a single entity.
  - Coreference algorithms usually performed within larger task of entity resolution
- 2 levels of resolution:
  - within document (includes **co-reference resolution**)
    - e.g. *Bin Ladin = he*
    - *his followers = they*
    - *terrorist attacks = they*
    - *the Federal Bureau of Investigation = FBI = F.B.I*
  - across document (or **named entity resolution**)
    - e.g. *Usama Bin Ladin =Osama Bin Ladin = Bin Ladin*
- **Event resolution** is also possible, but not widely used

# Examples from Contexts

1. **The State Department** renewed **its** appeal for **Bin Laden** on Monday and warned of possible fresh attacks by **his** followers against U.S. targets.

…

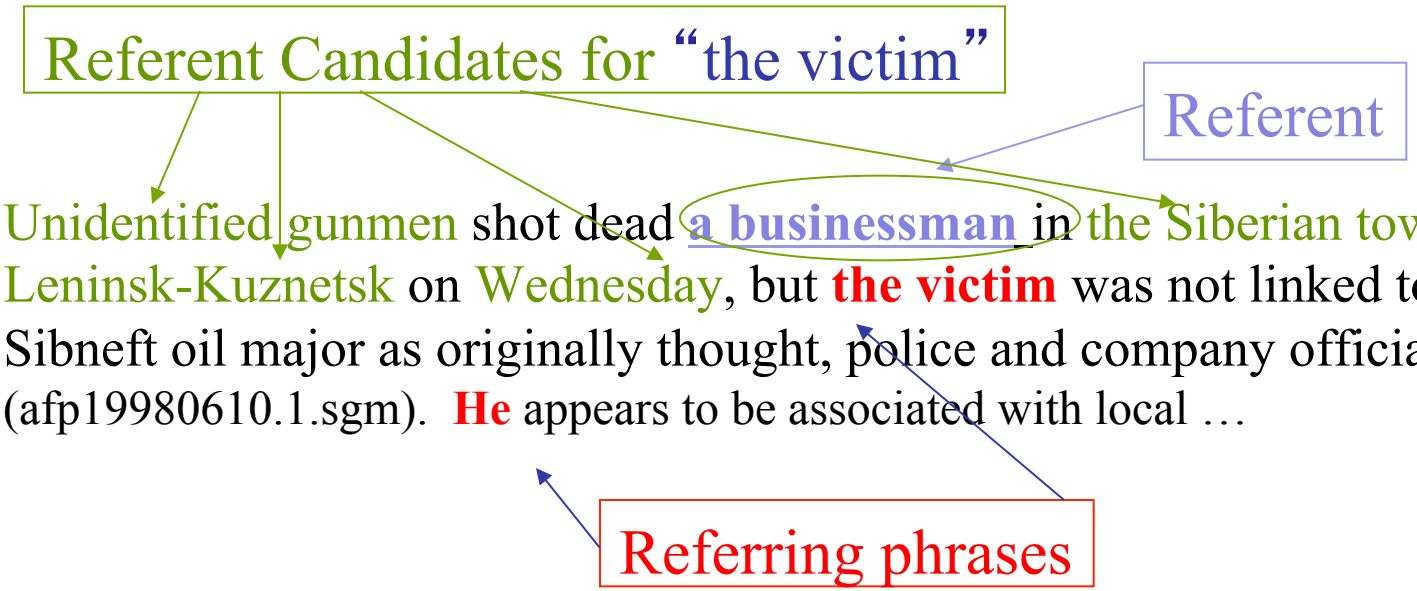
2. One early target of the F.B.I.’s Budapest office is expected to be **Semyon Y. Mogilevich**, **a Russian citizen who** has operated out of Budapest for a decade. Recently **he** has been linked to the growing money-laundering investigation in the United States involving the Bank of New York. **Mr. Mogilevich** is also the target of a separate money laundering and financial fraud investigation by the F.B.I. in Philadelphia, according to federal officials.

…

3. **The F.B.I.** will also have the final say over the hiring and firing of the 10 Hungarian agents who will work in **the office**, alongside five American agents. **The bureau** has long had agents posted in American embassies

# Terminology Examples

- The referent for a referring phrase is found by the resolution algorithm among the candidates, previous noun phrases.




- Unidentified gunmen shot dead **a businessman** in the Siberian town of Leninsk-Kuznetsk on Wednesday, but **the victim** was not linked to the Sibneft oil major as originally thought, police and company officials said. (afp19980610.1.sgm). **He** appears to be associated with local …

# Reference Types

- An algorithm must first decide which are the referring phrases that must be resolved
  - Pronouns
  - Definite noun phrases (the)
  - Indefinite noun phrases (a, an)
  - Demonstratives
  - Names
  - Others

# Pronouns

- **Pronouns** refer to entities that were introduced fairly recently, 1-4-5-10(?) sentences back.
  - ***Nominative*** (he, she, it, they, etc.)
    - e.g. The German authorities said **a Colombian$_1$** who had lived for a long time in the Ukraine flew in from Kiev. **He$_1$** had 300 grams of plutonium 239 in his baggage.
  - ***Oblique*** (him, her, them, etc.)
    - e.g. Undercover investigators negotiated with **three members of a criminal group$_2$** and arrested **them$_2$** after receiving the first shipment.
  - ***Possessive*** (his, her, their, etc. + hers, theirs, etc.)
    - e.g. **He$_3$** had 300 grams of plutonium 239 in **his$_3$** baggage. The suspected smuggler$_3$* denied that the materials were **his$_3$**. ***(*chain)***
  - ***Reflexive*** (himself, themselves, etc.)
    - e.g. There appears to be a growing problem of **disaffected loners$_4$** who cut **themselves$_4$** off from all groups .

# Definite noun phrases – the X

- Definite reference is used to refer to an entity identifiable by the reader because it is either
  - a) already mentioned previously (in discourse), or
  - b) contained in the reader's set of beliefs about the world (pragmatics), (known entities like "the Grand Canyon") or
  - c) the object itself is unique ("the universe"). (Jurafsky & Martin, 2000)
- E.g.
  - Mr. Torres and his companion claimed **a hardshelled black vinyl suitcase$_1$**. The police rushed **the suitcase$_1$** (a) to **the Trans-Uranium Institute$_2$** (c) where experts cut **it$_1$** open because they did not have the combination to the locks.
  - **The German authorities$_3$** (b) said **a Colombian$_4$** who had lived for a long time in **the Ukraine$_5$** (c) flew in from Kiev. He had **300 grams of plutonium 239$_6$** in his baggage. **The suspected smuggler$_4$** (a) denied that **the materials$_6$** (a) were his.

# Indefinite noun phrases – a X, or an X

- Typically, an indefinite noun phrase introduces a new entity into the discourse and would not be used as a referring phrase to something else
  - The exception is in the case of cataphora:
    *A Soviet pop star was killed at a concert in Moscow last night. Igor Talkov was shot through the heart as he walked on stage.*

# Demonstratives – this and that

- Demonstrative pronouns can either appear alone or as determiners

  *this ingredient, that spice*

- These NP phrases with determiners are ambiguous
  - They can be indefinite
    *I saw this beautiful car today.*
  - Or they can be definite
    *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.*

# Names

- Names can occur in many forms, sometimes called name variants.

  *Victoria Chen, Chief Financial Officer of Megabucks Banking Corp. since 2004, saw her pay jump 20% as the 37-year-old also became the Denver-based financial-services company's president. Megabucks expanded recently . . . MBC . . .*

  – (Victoria Chen, Chief Financial Officer, her, the 37-year-old, the Denver-based financial-services company's president)

  – (Megabucks Banking Corp. , the Denver-based financial-services company, Megabucks, MBC )

- Groups of a referrent with its referring phrases are called a **coreference group or coreference chain**.

# Unusual Cases

- Compound phrases

  *John and Mary got engaged. They make a cute couple.*
  *John and Mary went home. She was tired.*

- Singular nouns with a plural meaning

  *The focus group met for several hours. They were very intent.*

- Part/whole relationships

  *John bought a new car. A door was dented.*

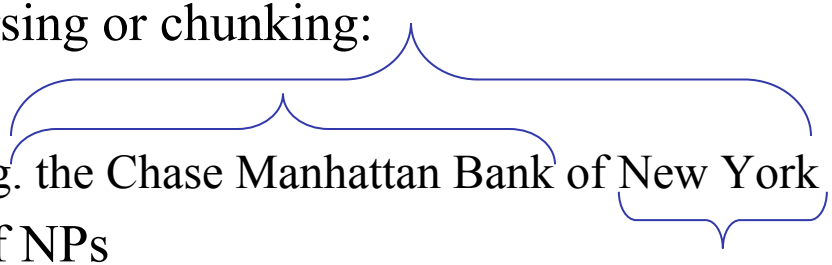
  *Four of the five surviving workers have asbestos-related diseases, including three with recently diagnosed cancer.*

# Approach to coreference resolution

- Naively identify all referring phrases for resolution:
  - all Pronouns
  - all definite NPs
  - all Proper Nouns
- Filter things that look referential but, in fact, are not
  - e.g. geographic names, *the United States*
  - Pronouns without actual meaning:
    - pleonastic "it", e.g. *it's 3:45 p.m., it was cold*
    - non-referential "it", "they", "there"
      - e.g. *it was essential, important, is understood,*
      - *they say,*
      - *there seems to be a mistake*

# Identify Referent Candidates

- All noun phrases (both indef. and def.) are considered potential referent candidates.
- A referring phrase can also be a referent for a subsequent referring phrases,
  - Example: (omitted sentence with name of suspect)
    He had 300 grams of plutonium 239 in his baggage. The suspected smuggler denied that the materials were his.
    (chain of 4 referring phrases)
- All potential candidates are collected in a table collecting feature info on each candidate.
- Requires either parsing or chunking:
  - chunking
    - e.g. the Chase Manhattan Bank of New York
  - Note nesting of NPs



# Features

- Define features between a refering phrase and each candidate
  - Number agreement: plural, singular or neutral
    - He, she, it, etc. are singular, while we, us, they, them, etc. are plural and should match with singular or plural nouns, respectively
    - Exceptions: some plural or group nouns can be referred to by either it or they

      *IBM announced a new product. They have been working on it …*
  - Gender agreement:
    - Generally animate objects are referred to by either male pronouns (he, his) or female pronouns (she, hers)
    - Inanimate objects take neutral (it) gender
  - Person agreement:
    - First and second person pronouns are “I” and “you”
    - Third person pronouns must be used with nouns

# More Features

- Binding constraints
  - Reflexive pronouns (himself, themselves) have constraints on which nouns in the same sentence can be referred to:
    *John bought himself a new Ford.* (John = himself)
    *John bought him a new Ford.* (John cannot = him)
- Recency
  - Entities situated closer to the referring phrase tend to be more salient than those further away
    - And pronouns can't go more than a few sentences away
- Grammatical role, sometimes approcimated by Hobbs distance
  - Entities in a subject position are more likely than in the object position

# Even more features

- Repeated mention
  - Entities that have been the focus of the discourse are more likely to be salient for a referring phrase
- Parallelism
  - There are strong preferences introduced by parallel constructs
    *Long John Silver went with Jim. Billy Bones went with him.*
    (him = Jim)
- Verb Semantics and selectional restrictions
  - Certain verbs take certain types of arguments and may prejudice the resolution of pronouns
    *John parked his car in the garage after driving it around for hours.*

# Example: rules to assign gender info

- Assign gender to "masculine",
  - if it is a pronoun "he, his, him"
  - if it contains markers like "Mr."
  - if the first name belongs to a list of masculine names
- Same for "feminine" and "neutral" (except for latter use categories such as singular, geo names, company names, etc.)
- Else, assign "unknown"
  - A phrase with unknown gender can match other phrases known as either masculine or feminine.

# Approach

- Train a classifier over an annotated corpus to identify which candidates and referring phrases are in the same coreference group
  - Evaluation results (for example, Vincent Ng at ACL 2005) are on the order of F-measure of 70, with generally higher precision than recall
  - Evaluation typically uses the B-Cubed scorer introduced by Bagga and Baldwin, which compares coreference groups
  - Pronoun coreference resolution by itself is much higher scoring, usually over 90%.

# Summary of Discourse Level Tasks

- Most widely used task is coreference resolution
  - Important in many other text analysis tasks in order to understand meaning of sentences
- Dialogue structure is also part of discourse analysis and will be considered separately (next time as part of pragmatics)
- Document structure
  - Recognizing known structure, for example, abstracts
  - Separating documents according to known structure
- Named entity resolution across documents
- Using cohesive elements to make fluent text in language generation and machine translation