CS114 Lecture 17

Discourse

Co-reference

April 23, 2014

Professor Meteer

Thanks for Jurafsky & Martin & Prof. Pustejovksy for slides

Outline

- Discourse Structure
  - Textiling
- Coherence
  - Hobbs coherence relations
  - Rhetorical Structure Theory
- Coreference
  - Kinds of reference phenomena
  - Constraints on co-reference
  - Anaphora Resolution
    - Hobbs
    - Loglinear
  - Coreference

Part III: Coreference

- **Victoria Chen**, **Chief Financial Officer of Megabucks Banking Corp** since 2004, saw **her** pay jump 20%, to $1.3 million, as **the 37-year-old** also became **the Denver-based financial-service company's president**. It has been ten years since **she** came to Megabucks from rival Lotsabucks.

- The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard's response.

Why reference resolution?

- Information Extraction:

“President of which company is retiring?”

First Union Corp. is continuing to wrestle with severe problems unleashed by a botched merger and a troubled business strategy. According to industry insiders at Paine Webber, their president, John R. Georgius, is planning to retire by the end of the year.

Some terminology

- Reference: Process by which speakers use words Victoria Chen and she to denote a particular person
  - **Referring expression**: Victoria Chen, she
  - **Referent**: the actual entity (but as a shorthand we might call "Victoria Chen" the referent).
  - Victoria Chen and she "**corefer**"
  - **Antecedent**: Victoria Chen
  - **Anaphor**: she

Tasks

- Pronominal anaphora resolution
  - Given a pronoun, find its antecedent
- Coreference resolution
  - Find the coreference relations among all referring expressions
  - Each set of referring expressions is a **coreference chain**. What are the chains in our story?
  - {Victoria Chen, Chief Financial Officer of Megabucks Banking Corp, her, the 37-year-old, the Denver-based financial-services company's president, she}
  - {Megabucks Banking Corp., the Denver-based financial-services company, Megabucks}
  - {her pay}
  - {Lotsabucks}

Coreference Example

- **Victoria Chen**, **Chief Financial Officer** of **Megabucks Banking Corp** since 2004, saw **her** pay jump 20%, to $1.3 million, as **the 37-year-old** also became **the Denver-based financial-service company's president**. It has been ten years since **she** came to Megabucks from rival Lotsabucks.

Many types of reference

- (after Webber 91)
- According to Doug, Sue just bought a 1962 Ford Falcon
  - But **that** turned out to be a lie (a speech act)
  - But **that** was false (proposition)
  - **That** struck me as a funny way to describe the situation (manner of description)
  - **That** caused Sue to become rather poor (event)

4 types of referring expressions

1. Indefinite noun phrases: new to hearer
   - Mrs. Martin was so very kind as to send Mrs. **Goddard a beautiful goose**
   - He had gone round one day to bring her **some walnuts**.
   - I am going to the butchers to buy **a goose** (specific/non-specific)
     - I hope they still have **it**
     - I hope they still have **one**

2. Definite noun phrases: identifiable to hearer because
   - Mentioned: It concerns **a white stallion** which I have sold to an officer. But the pedigree of **the white stallion** was not fully established.
   - Identifiable from beliefs or unique: I read about it in **The New York Times**
   - Inherently unique: **The fastest car** in …

Reference Phenomena:
3. Pronouns

Emma smiled and chatted as cheerfully as **she** could.

- Compared to definite noun phrases, pronouns require more **referent salience**.

John went to Bob's party, and parked next to **a classic Ford Falcon**.

He went inside and talked to Bob for more than an hour.

Bob told him that he recently got engaged.

??He also said that he bought **it** yesterday.

OK He also said that he bought **the Falcon** yesterday.

More on Pronouns

- Anaphor: pronoun appears after referent (usual case)
- Cataphora: pronoun appears before referent:
  - Even before she saw it, Dorothy had been thinking about the Emerald City every day.

4. Names

- Miss Woodhouse certainly had not done him justice.
- International Business Machines sought patent compensation from Amazon. In fact, IBM had previously sued a number of other companies.

Complications:
Inferrables and Generics

- Inferrables ("bridging inferences")
  - I almost bought a 1962 Ford Falcon today, but a door had a dent and **the engine** seemed noisy.

- Generics:
  - I'm interested in buying **a Mac laptop**. **They** are very stylish.
  - In March in Boulder you have to wear **a jacket**.

Features for pronominal anaphora resolution

- Number agreement
  - John has a Ford Falcon. It is red.
  - *John has three Ford Falcons. It is red.
  - But note:
    - IBM is announcing a new machine translation product. They have been been working on it for 20 years.
- Gender agreement
  - John has an Acura. He/it/she is attractive.
- Syntactic constraints ("Binding Theory")
  - John bought himself a new Ford (himself=John)
  - John bought him a new Ford (him = not John)

Pronoun Interpretation Features

- Selectional Restrictions
  - John parked his Ford in the garage. He had driven **it** around for hours.

- Recency
  - The doctor found an old map in the captain's chest. Jim found an even older map hidden on the shelf. **It** described an island full of redwood trees and sandy beaches.

Pronoun Interpretation Preferences

- Grammatical Role: Subject preference

Billy Bones went to the bar with Jim Hawkins.

He called for a glass of rum.

[he=Billy]

Jim Hawkins went to the bar with Billy Bones.

He called for a glass of rum.

[he = Jim]

Repeated Mention Preference

- Billy Bones had been thinking about a glass of rum ever since the pirate ship docked. He hobbled over to the Old Parrot bar. Jim Hawkins went with him. **He** called for a glass of rum.

  [he=Billy]

Parallelism Preference

- Long John Silver went with Jim to the Old Parrot.
- Billy Bones went with **him** to the Old Anchor Inn.

  [him=Jim]

Verb Semantics Preferences

- John telephoned Bill. **He** lost the laptop. [he=John]

- John criticized Bill. **He** lost the laptop. [he=Bill]

- Implicit causality
  - Implicit cause of criticizing is object.
  - Implicit cause of telephoning is subject.

Two algorithms for pronominal anaphora resolution

- The Hobbs Algorithm
- A Log-Linear Model

Hobbs algorithm

1. Begin at NP
2. Go up tree to first NP or S. Call this X, and the path p.
3. Traverse all branches below X to the left of p, left-to-right, breadth-first. Propose as antecedent any NP that has a NP or S between it and X.
4. If X is the highest S in the sentence, traverse the parse trees of the previous sentences in the order of recency. Traverse left-to-right, breadth first. When a NP is encountered, propose as antecedent. If not the highest node, go to step 5.

Hobbs (cont.)

5. From node X, go up the tree to the first NP or S. Call it X, and the path p.
6. If X is an NP and the path to X did not pass through the nominal that X dominates, propose X as antecedent.
7. Traverse all branches below X to the right of the path, in a left-to-right, breadth first manner. Propose any NP encountered as the antecedent.
8. If X is an S node, traverse all branches of X to the right of the path but do not go below any NP or S encountered. Propose any NP as the antecedent.
9. Go to step 4

Hobbs algorithm: walking through an example

John saw a Falcon at the dealership.
He showed it to Bob.
He bought it.

- current sentence: right to left
- previous sentences: left to right

A loglinear model

- Supervised machine learning
- Train on a corpus in which each pronoun is labeled with the correct antecedent
- In order to train: We need to extract
  - Positive examples of referent-pronoun pairs
  - Negative example of referent-pronoun pairs
  - Feature for each one
- Then we train model to predict 1 for true antecedent and 0 for wrong antecedents

Features

- Strict gender (T/F)
  - e.g. male pronoun Pro_i with male antecedent NP_j
- Compatible gender (T/F)
  - e.g. male pronoun Pro_i with antecedent NP_j of unknown gender
- Strict number (T/F)
  - e.g. singular pronoun with singular antecedent
- Compatible number (T/F)
  - e.g. singular pronoun with antecedent of unknown number

Features (cont.)

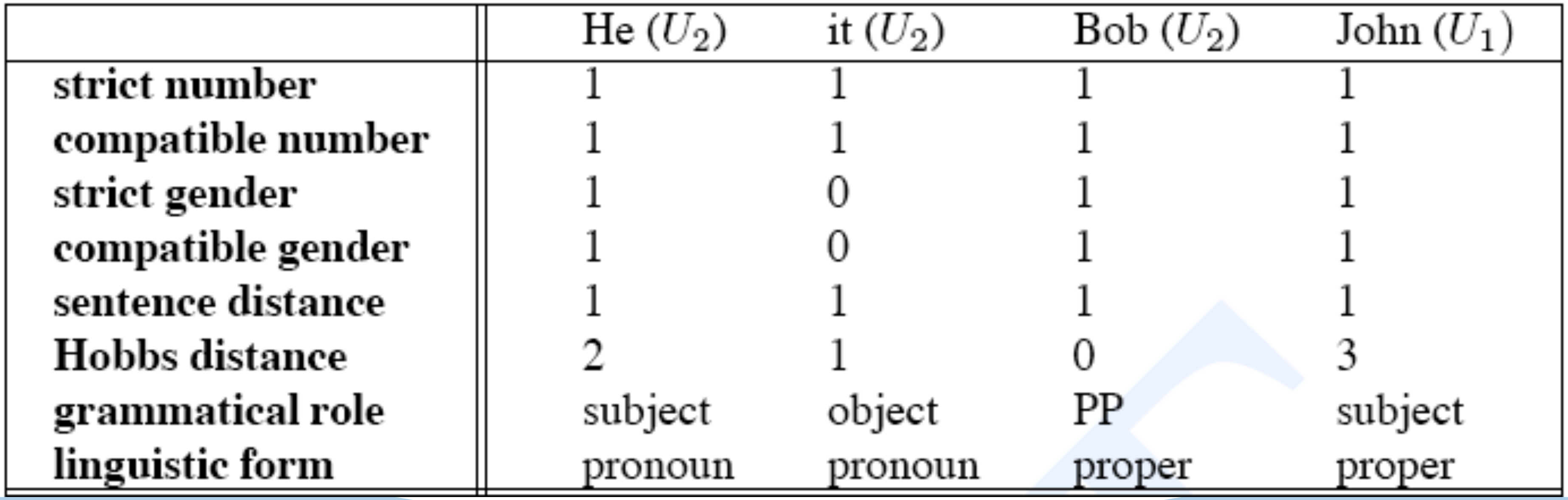
- **Sentence distance**
  - The number of sentences between the pronoun and the potential antecedent
- **Hobbs distance**
  - The number of noun groups that the Hobbs algorithm has to skip, starting backwards from the pronoun, before the potential antecedent id found
- **Grammatical role**
  - Whether the potential antecedent is in the syntactic subject or object, or is embedded in a prepositional phrase
- **Linguist form**
  - Whether the potential antecedent is a proper name, definite description, indefinite NP or a pronoun

Example: target = He (U3)

John saw a beautiful 1961 Ford Falcon at the used car dealership (U1)

He showed it to Bob (U2)

He bought it (U3)



	He (U_2)	it (U_2)	Bob (U_2)	John (U_1)
strict number	1	1	1	1
compatible number	1	1	1	1
strict gender	1	0	1	1
compatible gender	1	0	1	1
sentence distance	1	1	1	1
Hobbs distance	2	1	0	3
grammatical role	subject	object	PP	subject
linguistic form	pronoun	pronoun	proper	proper

Coreference resolution

- **Victoria Chen**, **Chief Financial Officer of Megabucks Banking Corp** since 2004, saw **her** pay jump 20%, to $1.3 million, as **the 37-year-old** also became **the Denver-based financial-service company's president**. It has been ten years since **she** came to Megabucks from rival Lotsabucks.
  - {Victoria Chen, Chief Financial Officer of Megabucks Banking Corp, her, the 37-year-old, the Denver-based financial-services company's president, she}
  - {Megabucks Banking Corp., the Denver-based financial-services company, Megabucks}
  - {her pay}
  - {Lotsabucks}

Coreference resolution

- **Victoria Chen**, **Chief Financial Officer of Megabucks Banking Corp** since 2004, saw **her** pay jump 20%, to $1.3 million, as **the 37-year-old** also became **the Denver-based financial-service company's president**. It has been ten years since **she** came to Megabucks from rival Lotsabucks.
- Have to deal with
  - Names
  - Non-referential pronouns
  - Definite NPs

Algorithm for coreference resolution

- Based on: a binary classifier given an anaphor and a potential antecedent
  - Returns true or false
- Process a document from left to right
  - For each NP_j we encounter
    - Search backwards through document at NPs
    - For each such potential antecedent NP_i
      - Run our classifier
      - If it returns true, coindex NP_i and NP_j and return
    - Terminate when we reach beginning of document

Features for coreference classifier

- Anaphor edit distance [0,1,2…]
  - The character minimum edit distance from the potential antecedent to the anaphor
- Antecedent edit distance [0,1,2…]
  - The character minimum edit distance from the potential anaphor to the antecedent
- Alias (T/F)
  - A multipart feature which required a named entity tagger. Returns T if both named entities are of the same type and NP1 is an alias of NP2
  - Dr. House, House or IBM, International Business Machines

More features

- Appositive (T/F)
  - True if anaphor is in the syntactic apposition relationship to the antecedent
    - The CFO, Vicoria Chen, was …
- Linguistic form
  - Whether the potential anaphor NP_j is a proper, definite, indefinite, or pronoun

Evaluation: Vilain et al 1995

- Suppose A, B, and C are coreferent
- Could represent this as A-B, B-C
  - Or as A-C, A-B
  - Or as A-C, B-C
- Call any of these sets of correct links the reference set.
- The output of coref algorithm is the hypothesis links.
- Our goal: compute precision and recall from the hypothesis to the reference set of links

Evaluation: Vilain et al 1995

- Clever algorithm to deal with the fact that there are multiple possible referent links.
  - Suppose A,B,C,D coreferent and (A-B,B-C,C-D) is referent.
  - Algorithm returns A-B, C-D
  - Precision should be 1, recall should be 2/3

    (since need 3 links to make 4 things coreferent, and we got 2 of them)

Coreference: further difficulties

- Lots of other algorithms and other constraints
  - Hobbs: reference resolution as by-product of general reasoning

    The city council denied the demonstrators a permit because

    they feared violence.

    they advocated violence.
  - An axiom: for all X,Y,Z,W fear(X,Z)&advocate(Y,Z)&enable_to_cause(W,Y,Z)-> deny(X,Y,W)
  - First clause is: deny(city_council,demonstrators,permit)
  - Second clause: Explanation

Coreference: further difficulties

The city council denied the demonstrators a permit because

they feared violence.

they advocated violence.

- An axiom:
  - for all X,Y,Z,W fear(X,Z) & advocate(Y,Z)& enable_to_cause(W,Y,Z)-> deny(X,Y,W)
- from "they=city_council" we could correctly infer deny(X,Y,W) in the "feared violence" example
- from "they=demonstrators" we could correctly infer deny(X,Y,W) in the "advocated violence" example