

# Speech and Language Processing

## Discourse: Anaphora Resolution

**Dan Jurafsky**

# Outline

---

- **Reference**
  - Kinds of reference phenomena
  - Constraints on co-reference
  - Preferences for co-reference
  - The Lappin-Leass algorithm for coreference

# Reference Resolution

---

- John went to Bill's car dealership to check out an Acura Integra. He looked at it for half an hour
- I'd like to get from Boston to San Francisco, on either December 5th or December 6th. It's ok if it stops in another city along they way

# Why reference resolution?

---

- **Conversational Agents:** Airline reservation system needs to know what “it” refers to in order to book correct flight
- **Information Extraction:** **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

---

- John went to Bill's car dealership to check out an Acura Integra. He looked at it for half an hour
- Reference: process by which speakers use words John and he to denote a particular person
  - Referring expression: John, he
  - Referent: the actual entity (but as a shorthand we might call "John" the referent).
  - John and he "corefer"
  - Antecedent: John
  - Anaphor: he

# Many types of reference

---

- (after Webber 91)
- According to John, Bob bought Sue an Integra, and Sue bought Fred a Legend
  - 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)
  - **That** caused them both to become rather poor (combination of several events)

# Reference Phenomena

- **Indefinite noun phrases: new to hearer**
  - I saw an Acura Integra today
  - Some Acura Integras were being unloaded...
  - I am going to the dealership to buy an Acura Integra today.  
(specific/non-specific)
    - I hope they still have it
    - I hope they have a car I like
- **Definite noun phrases: identifiable to hearer because**
  - **Mentioned:** I saw an Acura Integra today. The Integra was white
  - **Identifiable from beliefs:** The Indianapolis 500
  - **Inherently unique:** The fastest car in ...

# Reference Phenomena: Pronouns

---

- I saw an Acura Integra today. **It** was white
- Compared to definite noun phrases, pronouns require more referent **salience**.
  - John went to Bob's party, and parked next to a beautiful Acura Integra
  - 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 Acura** yesterday



# More on Pronouns

---

- **Cataphora: pronoun appears before referent:**
  - Before **he** bought **it**, John checked over the Integra very carefully.

# Inferrables

---

- I almost bought an Acura Integra today, but **the engine** seemed noisy.
- Mix the flour, butter, and water.
  - Knead **the dough** until smooth and shiny
  - Spread **the paste** over the blueberries
  - Stir **the batter** until all lumps are gone.

# Generics

---

- I saw no less than 6 Acura Integras today.  
**They** are the coolest cars.

# Pronominal Reference Resolution

---

- **Given a pronoun, find the reference (either in text or as an entity in the world)**
- **We will approach this today in 3 steps**
  - **Hard constraints on reference**
  - **Soft constraints on reference**
  - **Algorithms which use these constraints**

# Hard constraints on coreference

---

- **Number agreement**
  - John has an Acura. It is red.
- **Person and case agreement**
  - \*John and Mary have Acuras. We love them (where We=John and Mary)
- **Gender agreement**
  - John has an Acura. He/it/she is attractive.
- **Syntactic constraints**
  - John bought himself a new Acura (himself=John)
  - John bought him a new Acura (him = not John)

# Pronoun Interpretation Preferences

---

- **Selectional Restrictions**
  - John parked **his Acura** in the garage. He had driven **it** around for hours.
- **Recency**
  - John has an Integra. Bill has a **Legend**. Mary likes to drive it.

# Pronoun Interpretation Preferences

---

- **Grammatical Role: Subject preference**
  - **John** went to the Acura dealership with Bill. **He** bought an Integra.
  - **Bill** went to the Acura dealership with John. **He** bought an Integra
  - **(?)** John and Bill went to the Acura dealership. **He** bought an Integra

# Repeated Mention preference

---

- **John** needed a car to get to his new job. He decided that he wanted something sporty. Bill went to the Acura dealership with him. **He** bought an Integra.



# Parallelism Preference

---

- Mary went with **Sue** to the Acura dealership. Sally went with **her** to the Mazda dealership.
- **Mary** went with Sue to the Acura dealership. Sally told **her** not to buy anything.

# Verb Semantics Preferences

---

- **John** telephoned Bill. **He** lost the pamphlet on Acuras.
- John criticized **Bill**. **He** lost the pamphlet on Acuras.
- **Implicit causality**
  - Implicit cause of criticizing is object.
  - Implicit cause of telephoning is subject.

# Pronoun Resolution Algorithm

---

- Lappin and Leass (1994): Given **he/she/it**, assign antecedent.
- Implements only recency and syntactic preferences
- Two steps
  - Discourse model update
    - When a new noun phrase is encountered, add a representation to discourse model with a salience value
    - Modify saliences.
  - Pronoun resolution
    - Choose the most salient antecedent

# Salience Factors and Weights

- From Lappin and Leass

<b>Subject recency</b>	<b>100</b>
<b>Subject emphasis</b>	<b>80</b>
<b>Existential emphasis</b>	<b>70</b>
<b>Accusative (direct object) emphasis</b>	<b>50</b>
<b>Ind. Obj and oblique emphasis</b>	<b>40</b>
<b>Non-adverbial emphasis</b>	<b>50</b>
<b>Head noun emphasis</b>	<b>80</b>

# Recency

---

- Weights are cut in half after each sentence is processed
- This, and a sentence recency weight (100 for new sentences, cut in half each time), captures the recency preferences

# Lappin and Leass (cont)

- Grammatical role preference
  - Subject > existential predicate nominal > object > indirect object > demarcated adverbial PP
- Examples
  - An Acura Integra is parked in the lot (subject)
  - There is an Acura Integra parked in the lot (ex. pred nominal)
  - John parked an Acura Integra in the lot (object)
  - John gave his Acura Integra a bath (indirect obj)
  - In his Acura Integra, John showed Susan his new CD player (demarcated adverbial PP)
- Head noun emphasis factor gives above 80 points, but followed embedded NP nothing:
  - The owner's manual for an Acura Integra is on John's desk

# Lappin and Leass Algorithm

---

- Collect the potential referents (up to 4 sentences back)
- Remove potential referents that do not agree in number or gender with the pronoun
- Remove potential references that do not pass syntactic coreference constraints
- Compute total salience value of referent from all factors, including, if applicable, role parallelism (+35) or cataphora (-175).
- Select referent with highest salience value. In case of tie, select closest.

# Example

- John saw a beautiful Acura Integra at the dealership. He showed it to Bob. He bought it.

Sentence 1:

	rec	Subj	Exist	Obj	Ind-obj	Non-adv	Head N	Total
John	100	80				50	80	310
Integra	100			50		50	80	280
dealership	100					50	80	230



# After sentence 1

- Cut all values in half

Referent	Phrases	Value
John	{John}	155
Integra	{a beautiful Acura Integra}	140
dealership	{the dealership	115

## He showed it to Bob

- **He** specifies male gender
- So Step 2 reduces set of referents to only **John**.
- Now update discourse model:
- **He** in current sentence (recency=100), subject position (=80), not adverbial (=50) not embedded (=80), so add 310:

Referent	Phrases	Value
John	{John, he1}	155+310
Integra	{a beautiful Acura Integra}	140
dealership	{the dealership}	115

## He showed **it** to Bob

- Need to add it, which can be Integra or dealership.
- Need to add weights:
  - Parallelism: it + Integra are objects (dealership is not), so +35 for integra
  - Integra 175 to dealership 115, so pick Integra
- Update discourse model: it is nonembedded object, gets  $100+50+50+80=280$ :

# He showed it to Bob

Referent	Phrases	Value
John	{John, he1}	465
Integra	{a beautiful Acura Integra, it1}	420
dealership	{the dealership}	115

## He showed it to Bob

- Bob is new referent, is oblique argument, weight is  $100+40+50+80=270$

Referent	Phrases	Value
John	{John, he1}	465
Integra	{a beautiful Acura Integra, it1}	420
Bob	{Bob}	270
dealership	{the dealership}	115

# He bought it

- Drop weights in half:

Referent	Phrases	Value
John	{John, he1}	232.5
Integra	{a beautiful Acura Integra, it1}	210
Bob	{Bob}	135
dealership	{the dealership}	57.5

He2 will be resolved to John, and it2 to Integra

# Reference Resolution: Summary

---

- **Lots of other algorithms and other constraints**
  - **Centering theory: constraints which focus on discourse state, and focus.**
  - **Hobbs: ref. 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  $X, Y, Z, W$   
 $\text{fear}(X, Z) \& \text{advocate}(Y, Z) \& \text{enable\_to\_cause}(W, Y, Z) \rightarrow \text{deny}(X, Z, W)$**
  - **Hence  $\text{deny}(\text{city\_council}, \text{demonstrators}, \text{permit})$**