Topics for Today

- Last class: Pragmatics
  - problem of inference
  - knowledge-based methods for inferring text cohesion
    - knowledge about action and causality
    - scripts
- Today: Pragmatics of discourse context
  - reference resolution
  - noun phrase coreference resolution
  - machine learning approach to NP coreference resolution

The problem of reference resolution

Gracie: Oh yeah…and then Mr. And Mrs. Jones were having matrimonial trouble, and my brother was hired to watch Mrs. Jones.

George: Well, I imagine she ws a very attractive woman.

Gracie: She was, and my brothter watched her day and night for six months.

George: Well, what happened?

Gracie: She finally got a divorce.

George: Mrs. Jones?

Gracie: No, my brother's wife.

George Burns and Gracie Allen in *The Salesgirl*

Reference resolution

- **Reference**: the process by which speakers use expressions like "John Simon" and "his" to denote the same real-world entity
  - **Referring expressions**: NL expression used to perform reference
  - **Referent**: the entity that is referred to
  - **Shorthand form**: *his* refers to John Simon

John Simon, Chief Financial Officer of Prime Corp. since 1986, saw his pay jump 20%, to $1.3 million, as the 37-year-old also became the financial-services company's president...

?

Coreference

- **Coreference:** two referring expressions that are used to refer to the same sntity are said to corefer
- *John Simon* is the **antecedent** of *his*.
- Reference to an entity that has been previously introduced into the discourse is called **anaphora**; and the referring expression used is said to be **anaphoric**.

John Simon, Chief Financial Officer of Prime Corp. since 1986, saw his pay jump 20%, to $1.3 million, as the 37-year-old also became the financial-services company's president...

Types of referring expressions

- Indefinite noun phrases
  - Introduce entities that are new to the hearer into the discourse context
    - » I saw *a Subaru WRX* today.
    - » I saw *this awesome Subaru WRX* today.
- Definite noun phrases
  - Refer to an entity that is identifiable to the hearer
    - » It has already been mentioned in the discourse
    - » It is contained in the hearer's set of beliefs about the world
    - » The uniqueness of the object is implied by the description itself
      - ◆ I saw a Subaru WRX today. *The WRX* was blue and needed a wash.
      - ◆ *The Indy 500* is the most popular car race in the US.
      - ◆ *The fastest car in the Indy 500* was a Subaru WRX.

Types of referring expressions

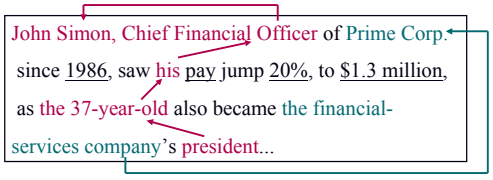
- Pronouns
  - Another form of definite reference
  - Referent must have a high degree of activation or **salience** in the discourse model
    - » John went to Bob's party, and parked next to a beautiful Subaru WRX. He went inside and talked to Bob for more than an hour. Bob told him that he recently got engaged.
      - ◆ ?? He also said that he bought *it* yesterday.
      - ◆ He also said that he bought *the WRX* yesterday.
  - Cataphora: referring expression is mentioned before its referent
    - » Before *he* bought *it*, John checked over the WRX carefully.

Types of referring expressions

- Demonstrative pronouns
  - Behave somewhat differently than simple definite pronouns
    - » Can appear alone or as determiners
    - » Choice of *this* or *that* depends on some notion of spatial proximity
      - ◆ I bought a WRX yesterday. It's similar to the one I bought a year ago. *That one* was really nice, but I like *this one* even better.
- One-anaphora
  - Blends properties of definite and indefinite reference
    - » I saw no fewer than 6 Subaru WRX's today. Now I want *one*.
  - May introduce a new entity into the discourse, but it is dependent on an existing referent for the description of this new entity.

Noun Phrase Coreference

- Identify all phrases that refer to each real-world entity mentioned in the text



Why It's Hard

Many sources of information play a role

- head noun matches
  - » IBM *executives* = the *executives*
  - » Microsoft *executives*
- syntactic constraints
  - » John helped himself to...
  - » John helped him to…
- discourse focus, recency, syntactic parallelism, semantic class, agreement, world knowledge, …

Why It's Hard

No single source is a completely reliable indicator

- semantic preferences
  - » Mr. Callahan = president =? the carrier
- number and gender
  - » assassination (of Jesuit priests) = these murders
  - » the woman = she = Mary =? the chairman

Why It's Hard

Coreference strategies differ depending on the type of referring NP

- definiteness of NPs
  - » … Then Mark saw the man walking down the street.
  - » … Then Mark saw a man walking down the street.
- pronoun resolution alone is notoriously difficult
  - » resolution strategies differ for each type of pronoun
  - » some pronouns refer to nothing in the text

I went outside and it was snowing.

Types of referents: complications

- Inferrables
  - A referring expression does not refer to an entity in the text, but to one that is inferentially related to it.
    - » I almost bought a WRX today, but *a door* had a dent and *the engine* seemed noisy.
    - » Mix the flour, butter, and water. Stir *the batter* until all lumps are gone.
- Discontinous sets
  - Referents may have been evoked in discontinous phrases
    - » John has a Volvo, and Mary has a Mazda. *They* drive *them* all the time.
- Generics – refer to a class of entities
  - I saw no fewer than 6 WRX's today. *They* are the coolest cars.

Topics for today

- Pragmatics of discourse
  - reference resolution
  - noun phrase coreference resolution
  - machine learning approach to NP coreference resolution
    - » a high-performing machine learning solution
    - » two extensions (if time)

A Machine Learning Approach

- Classification
  - given a description of two noun phrases, NP_i and NP_j, classify the pair as *coreferent* or *not coreferent*

[John Simon], [Chief Financial Officer] of [Prime Corp.]

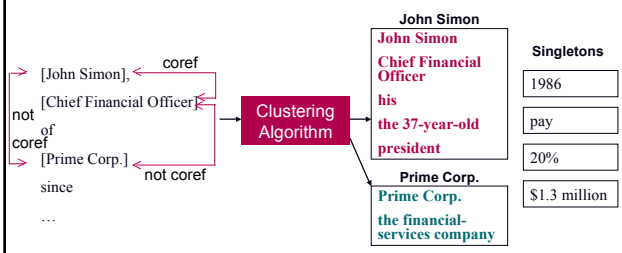
since 1986, saw his pay jump 20%, to $1.3 million,

as the 37-year-old also became the ….

Aone & Bennett [1995]; Connolly et al. [1995]; McCarthy & Lehnert [1995]; Soon, Ng & Lim [2001]; Ng & Cardie [2002]

A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



Training Data

- Creating training instances
  - texts annotated with coreference information
  - one instance for each pair of noun phrases
    - » feature vector: describes the two NPs and context
    - » class value:
      - *coref* pairs on the same coreference chain
      - *not coref* otherwise

$f1_1, f1_2, \ldots, f1_n$ (NP 1) $f2_1, f2_2, \ldots, f2_n$ (NP 2) $r_1, r_2, \ldots, r_n$ (relations) C (class)

Training Instance Selection

- all NP pairs produces highly skewed data set
- create
  - *positive instance* for each anaphoric noun phrase, NP_j, and its closest preceding antecedent, NP_i
  - *negative instance* for NP_j and each intervening noun phrase, NP_{i+1}, NP_{i+2} , … , NP_{j-1}

Soon et al. Instance Representation

	Soon_str	Pronoun_1 Pronoun_2 Definite_2 Demonst_2	Number	Gender	Both_ proper_ nouns	Appos- itive	WordNet_ class	Alias	Sent_ dist	Class
NP2/ NP1	Incom- patible	No; No; No; No	Com- patible	Com- patible	Com- patible	Com- patible	Com- patible	Incom- patible	0	Coref
NP3/ NP2	Incom- patible	No; No; No; No	Com- patible	Incom- patible	Com- patible	Incom- patible	Incom- patible	Incom- patible	0	Not coref

NP1 NP2 NP3

[John Simon], [Chief Financial Officer] of [Prime Corp.] since …

Soon et al. Clustering Algorithm

CREATE-COREF-CHAINS (NP_1, NP_2, …, NP_n; *doc*)

Mark each NP_j as belonging to its own class: $NP_j \in c_j$

Proceed through the NPs in left-to-right order. For each NP_j encountered, consider each preceding NP_i:

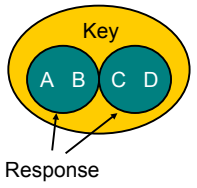
Let c_i = class of NP_i and c_j = class of NP_j

Let *coref-likelihood* = dtree (feat_vec (NP_i, NP_j, *doc*))

If *coref-likelihood* > *0.5* then $c_j = c_j \cup c_i$

Evaluation

- MUC-6 and MUC-7 coreference data set
- documents annotated w.r.t. coreference
- 30 + 30 training texts (dry run)
- 30 + 20 test texts (formal evaluation)
- scoring program
  - recall
  - precision
  - F-measure: 2PR/(P+R)



Results

	C4.5						RIPPER					
	MUC-6			MUC-7			MUC-6			MUC-7		
	R	P	F	R	P	F	R	P	F	R	P	F
Original Soon	58.6	67.3	62.6	56.1	65.5	60.4	-	-	-	-	-	-
Duplicated Soon Bsln	64.0	67.0	65.5	55.2	68.5	61.2	62.4	65.0	63.7	54.0	69.5	60.8
Soon Str Match	49.3	68.8	57.4	46.0	69.2	55.2	-	-	-	-	-	-
Single Cluster	93.8	33.4	49.2	90.7	33.5	48.9	-	-	-	-	-	-
Top System	59	72	64.9	-	-	-	-	-	-	-	-	-

Classifier for MUC-6 Data Set

```
ALIAS = C: +
ALIAS = I:
| SOON_STR_NONPRO = C:
| | ANIMACY = NA: -
| | ANIMACY = I: -
| | ANIMACY = C: +
| SOON_STR_NONPRO = I:
| | PRO_STR = C: +
| | PRO_STR = I:
| | | PRO_RESOLVE = C:
| | | | EMBEDDED_1 = Y: -
| | | | EMBEDDED_1 = N:
| | | | | PRONOUN_1 = Y:
| | | | | | ANIMACY = NA: -
| | | | | | ANIMACY = I: -
| | | | | | ANIMACY = C: +
| | | | | PRONOUN_1 = N:
| | | | | | MAXIMALNP = C: +
| | | | | | MAXIMALNP = I:
| | | | | | | WNCLASS = NA: -
| | | | | | | WNCLASS = I: +
| | | | | | | WNCLASS = C: +
| | | PRO_RESOLVE = I:
| | | | APPOSITIVE = I: -
| | | | APPOSITIVE = C:
| | | | | GENDER = NA: +
| | | | | GENDER = I: +
| | | | | GENDER = C: -
```

Summary of the ML approaches

- Perform better than the best non-learning approaches on two standard data sets
- Still lots of room for improvement
  - Generic noun resolution remains a major limiting factor