Reference Resolution and other Discourse phenomena

11-711 Algorithms for NLP

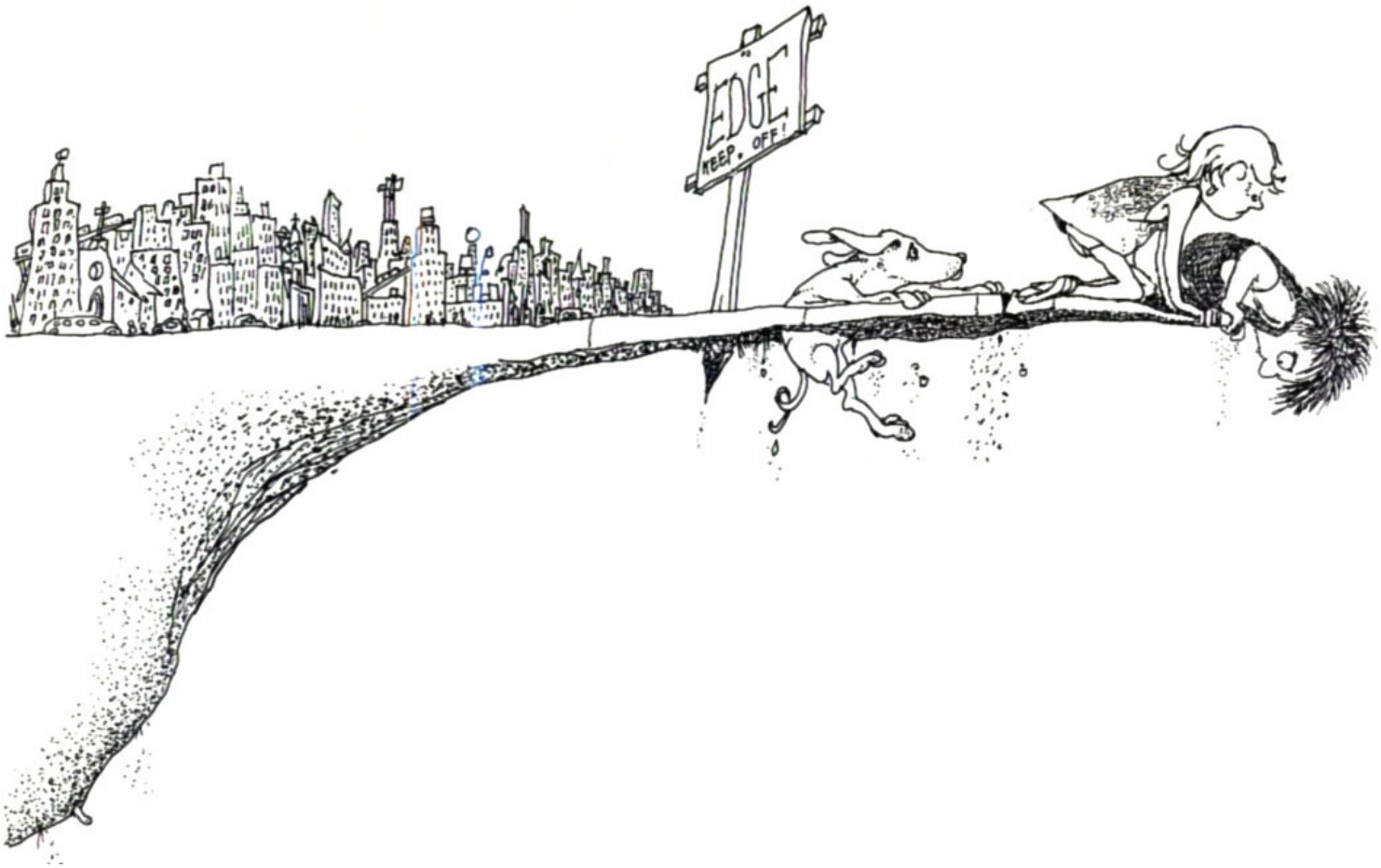
November 2020

What Is Discourse?

Discourse is the coherent structure of language above the level of sentences or clauses. A **discourse** is a coherent structured group of sentences.

What makes a passage coherent?

A practical answer: It has meaningful connections between its utterances.



Cover of Shel Silverstein's *Where the Sidewalk Ends* (1974)

Applications of Computational Discourse

- Analyzing sentences in context

- Automatic essay grading
- Automatic summarization
- Meeting understanding
- Dialogue systems

Kinds of discourse analysis

- Discourse: monologue, dialogue, multi-party conversation
- (Text) Discourse vs. (Spoken) Dialogue Systems

Discourse mechanisms vs. Coherence of thought

- "Longer-range" analysis (discourse) vs. "deeper" analysis (real semantics):
  - *John bought a car from Bill*
  - *Bill sold a car to John*
  - *They were both happy with the transaction*

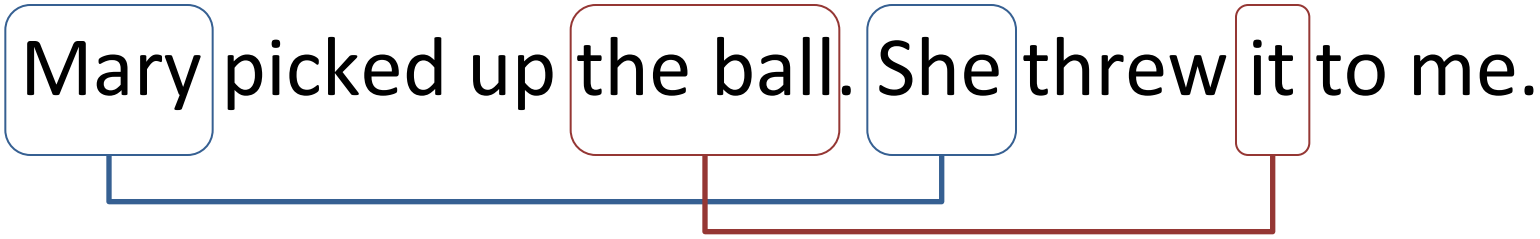
Reference resolution

Reference Resolution: example

- [[Apple Inc] Chief Executive Tim Cook] has jetted into [China] for talks with govt. officials as [he] seeks to clear up a pile of problems in [[the firm]'s biggest growth market] … [Cook] is on [his] first trip to [the country] since taking over…

- *Mentions* of the same *referent* (*entity*)
- Coreference chains (clusters):
  - {Apple Inc, the firm}
  - {Apple Inc Chief Executive Tim Cook, he, Cook, his}
  - {China, the firm's biggest growth market, the country}
  - And a bunch of ***singletons*** (dotted underlines)

Coreference Resolution



Reference resolution (entity linking)

Mary picked up the ball. She threw it to me.

3 Types of Referring Expressions

1. Pronouns
2. Names
3. Nominals

1st type: Pronouns

- Closed-class words like s*he, them, it*, etc. Usually **anaphora** (referring back to **antecedent**), but also **cataphora** (referring forwards):
    - *Although **he** hesitated, **Doug** eventually agreed.*
  - strong constraints on their use
  - can be bound: *Every student improved his grades*
- Pittsburghese: *yinz=yuns=youse=y'all*
- US vs UK: *Pittsburgh is/are undefeated this year.*

- SMASH(?) approach:
  - **Search** for antecedents
  - **Match** against hard constraints
  - **And Select** using **Heuristics** (soft constraints)

Search for Antecedents

- Identify all preceding NPs
  - Parse to find NPs
    - Largest unit with particular head word
  - Might use heuristics to prune
  - What about verb referents? Cataphora?

Match against hard constraints (1)

- Must agree on number, person, gender, animacy (in English)
- *Tim Cook has jetted in for talks with officials as [he] seeks to…*
  - *he: singular, masculine, animate, 3rd person*
  - *officials: plural, animate, 3rd person*
  - *talks: plural, inanimate, 3rd person*
  - *Tim Cook: singular, masculine, animate, 3rd person*

Match against hard constraints (2)

- Within 1 S, Chomsky government and binding theory:
  - **c-command**: 1st branching node above *x* dominates *y*

- *Abigail speaks with her*. [her != Abigail]
- *Abigail speaks with herself.* [her == Abigail]
- *Abigail's mom speaks with her.* [could corefer]
- *Abigail's mom speaks with herself.* [herself == mom]
- *Abigail hopes she speaks with her*. [she != her]
- *Abigail hopes she speaks with herself.* [she == herself]

Select using **Heuristics**

- Recency: preference for most recent referent
- Grammatical Role: subj>obj>others
  - *Billy went to the bar with Jim. He ordered rum*.
- Repeated mention: *Billy had been drinking for days. He went to the bar again today. Jim went with him. He ordered rum.*
- Parallelism: *John went with Jim to one bar. Bill went with him to another.*
- Verb semantics: *John phoned/criticized Bill. He lost the laptop.*
- Selectional restrictions: *John parked his car in the garage after driving it around for hours.*

Hobbs Algorithm

- Algorithm for walking through parses of current and preceding sentences
- Simple, often used as baseline
  - Requires parser, morph gender and number
    - plus head rules and WordNet for NP gender
- Implements binding theory, recency, and grammatical role preferences

- More complex: Grosz et al: **centering theory**

Semantics matters a lot

From Winograd 1972:

- *[The city council] denied [the protesters] a permit because [they] (advocated/feared) violence.*

Non-referential pronouns

- Other kinds of referents:
  - *According to Doug, Sue just bought the Ford Falcon*
    - *But **that** turned out to be a lie*
    - *But **that** was false*
    - ***That** struck me as a funny way to describe the situation*
    - ***That** caused a financial problem for Sue*
- Generics: *At CMU you have to work hard.*
- Pleonastics/clefts/extraposition:
  - *It is raining. It was me who called. It was good that you called.*
  - Analyze distribution statistics to recognize these.

2nd type: Proper Nouns

- When used as a referring expression, just match another proper noun
  - match syntactic head words
  - in a sequence (in English), the last token in name
    - not in many Asian names: *Xi Jinping* is *Xi*
    - not in organizations: *Georgia Tech* vs. *Virginia Tech*
    - not nested names: *the CEO of Microsoft*
- Use gazetteers (lists of names):
    - Natl. Basketball Assoc./NBA
    - Central Michigan Univ./CMU(!)
    - the Israelis/Israel

3rd type: Nominals

- Everything else, basically
  - {Apple Inc, the firm}
  - {China, the firm's biggest growth market, the country}
- Requires world knowledge, colloquial expressions
  - Clinton campaign officials, the Clinton camp
- Difficult

Learning reference resolution

Ground truth: *Mention* sets

- Train on sets of *markables*:
  - {Apple Inc$_{1:2}$, the firm$_{27:28}$}
  - {Apple Inc Chief Executive Tim Cook$_{1:6}$, he$_{17}$, Cook$_{33}$, his$_{36}$}
  - {China$_{10}$, the firm's biggest growth market$_{27:32}$, the country$_{40:41}$}
  - no sets for singletons

- Structure prediction problem:
  - *identify* the spans that are mentions
  - *cluster* the mentions

Mention identification

- Heuristics over phrase structure parses
  - Remove:
    - Nested NPs with same head: *[Apple CEO [Cook]]*
    - Numerical entities: *100 miles*
    - Non-referential *it*, etc.
  - Favoring recall
- Or, just all spans up to length N

Mention clustering

- Two main kinds:
  - Mention-pair models
    - Score each pair of mentions, then cluster
    - Can produce incoherent clusters:
      - *Hillary Clinton, Clinton, President Clinton*
  - Entity-based models
    - Inference difficult, due to exponential possible clusters

Mention-pair models (1)

- Binary labels: If *i* and *j* corefer, *i* < *j*, then $y_{i,j} = 1$

- *[[Apple Inc] Chief Executive Tim Cook] has jetted into [China] for talks with govt. officials as [he] …*
- For mention *he* (mention 6):
  - Preceding mentions: *Apple Inc, Apple Inc Chief Executive Tim Cook, China, talks, govt. officials*
  - $y_{2,6} = 1$, other *y*'s are all 0
- Assuming mention 20 also corefers with *he*:
  - For mention 20: $y_{2,20} = 1$ and $y_{6,20} = 1$, other *y*'s are all 0
- For *talks* (mention 3), all *y* = 0

Mention-pair models (2)

- Can use off-the-shelf binary classifier
  - applied to each mention *j* separately. For each, go from mention *j-1* down to first *i* that corefers with high confidence
  - then use transitivity to get any earlier coreferences
- Ground truth needs to be converted from *chains* to ground truth *mention-pairs*. Typically, only include one positive in each set
- *[[Apple Inc] Chief Executive Tim Cook] has jetted into [China] for talks with govt. officials as [he] …*
- $y_{2,6} = 1$ and $y_{3,6} = y_{4,6} = y_{5,6} = 0$

  $y_{1,6}$ not included in training data

Mention-ranking models (1)

- For each referring expression *i*, identify a single antecedent $a_i \in \{\varepsilon, 1, 2, \ldots, i\text{-}1\}$ by maximizing the score of (a, i)
  - Non-referential *i* gets $a_i = \varepsilon$
    - Might do those in pre-processing
- Train discriminative classifier using e.g. hinge loss or negative log likelihood.

Mention-ranking models (2)

- Again, ground truth needs to be converted from *clusters* to ground truth *mention-pairs*
  - Could use same heuristic (closest antecedent)
    - But *closest* might not be the most *informative* antecedent
  - Could treat *identity* of antecedent as a latent variable
  - Or, score can sum over all conditional probabilities that are compatible with the true cluster

Transitive closure issue

- *Hillary Clinton, Clinton, President Clinton*

- *Post hoc* revisions?
  - but many possible choices; heuristics
- Treat it as constrained optimization?
  - equivalent to graph partitioning
  - NP-hard

Entity-based models

- It is fundamentally a clustering problem
- So entity-based models identify clusters directly
- Maximize over entities: maximize *z*, where
  - z_i indicates the entity referenced by mention *i,* and
  - scoring function is applied to set of all *i* assigned to entity *e*
- Possible number of clusterings is Bell number, which is exponential
- So incremental search, based on local decisions

Incremental cluster ranking

- Like SMASH, but cluster picks up features of its members (gender, number, animacy)
  - Prevents incoherent clusters
  - But may make greedy search errors
  - So, use beam search
  - Or, make multiple passes through document, applying rules (sieves) with increasing recall
    - find high-confidence links first: *Hillary Clinton, Clinton, she*
    - rule-based system won 2011 CoNLL task (but not later)

Incremental perceptron

- Beam search, where states are clusterings
- When nothing in beam is compatible with gold reference, make a perceptron update:

  c^* = {*Abigail, her*}, {*she*}

  $\hat{c}$ = {*Abigail, she*} triggers update
- Or train with margin loss, or neural network

Reinforcement learning

- Think of clustering as a sequence of M *actions* to cluster M mentions
  - each action either: merges *i* into a cluster or starts a new cluster
- Stochastic policy is learned to make decisions
- Can be trained directly on evaluation metric
  - doesn't need to be differentiable or decomposable
- Sample from exponential possible trajectories
- Updates made once action sequence is complete

Learning to search

- Policy gradient can have large variance
- Add an *oracle* policy:
  - use it to generate initial path: *roll-in*
  - use it to compute minimum possible loss going forward to goal: *roll-out*
  - or, sample it during both
- Oracle may be noisy

Representations

- Hand-engineered features
- Lexical features
- Distributed representations

Mention Features

- Type: pronoun, name, other.
- Width: in tokens.
- Lexical features: first, last, head word
- Morphosyntactic features: POS, number, gender, dependency ancestors
- Genre type

- Conjoined features

Mention-pair Features

- Distance: in tokens, mentions, sentences; surface or tree traversal
- String match: exact, suffix, head, or complex
- Compatibility: gender, number, animacy
- Nesting (nested NPs cannot corefer)
- Speaker identity
- Gazetteers
- Lexical semantics: WordNet, Knowledge Graphs
- Dependency paths: binding constraints

Semantics

- *China, country, growth market*
- Need meaning? WordNet can provide *China* and *country*
- Also similarity derived from WordNet? *(Use caution here.)*

- Less important for recent systems

Entity features

- Aggregate mention-pair features. Kinds of aggregation:
  - All-True
  - Most-True
  - Most-False
  - None

  - Scalar: min, max, median
  - Number of mentions included, by type, etc.

Distributed representations (1)

- Embed mentions and entities
- Example for embedding mentions:
  - run bidirectional LSTM over whole text
  - concatenate embeddings of first, last, and head words, plus a vector of surface features
    - or use attention to find head word
  - score candidate pair: $\psi_S(a) + \psi_S(i) + \psi_M(a,i)$
    - $\psi_S(a) = \text{FeedFwd}_S(u^{(a)})$ (how likely to be a coreference)
    - $\psi_M(a,i) = \text{FeedFwd}_M([u^{(a)}; u^{(i)}; u^{(a)} \odot u^{(i)}; f(a,i,w)])$
- *blaze/fire*, good. *pilot/flight attendant*, bad.
- Or, embed mention pairs?

Distributed representations (2)

- Embedding entities:
  - Entity represented by its mentions
  - Mention embedding u_i, entity embedding v_e
  - Decision to merge *i* into *e*:
    - $\psi_E(i,e)$ = FeedFwd($[v_e; u_i]$)
    - if yes, $v_e \leftarrow f(v_e, u_i)$
      or $v_e \leftarrow \text{Pool}(v_e, u_i)$

Evaluating coreference

Evaluating coreference

- “Aggravatingly complex”
- Simple metrics too easy to “game”
- CoNLL 2011 practice: average of three:
  - MUC (Message Understanding Conference)
  - B-CUBED
  - CEAF

- CONE (B.Lin, R.Shah, Frederking, Gershman, 2010)
  - for Named Entities, using estimated gold standard

Many other aspects of discourse

- Given/new information
- Coherence/cohesion
- Discourse structure models
- Pragmatics
  - Speech Acts
  - Grice's Maxims (a famous bad idea!)

Information structure: given/new

- *Where are my **shoes**? Your shoes are in the **closet***
- *What's in the **closet**?*
  - *??Your shoes are in the **closet**.*
  - *Your **shoes** are in the closet*.

- Definiteness/pronoun, length, position in S

Coherence, Cohesion

- Coherence relations:
  - *John hid Bill's car keys. He was drunk.*
  - *John hid Bill's car keys. He likes spinach.*

- Entity-based coherence (Centering) and lexical cohesion:
  - *John went to the store to buy a piano*
  - *He had gone to the store for many years*
  - *He was excited that he could finally afford a piano*
  - *He arrived just as the store was closing for the day*

  versus
  - *John went to the store to buy a piano*
  - *It was a store he had gone to for many years*
  - *He was excited that he could finally afford a piano*
  - *It was closing for the day just as John arrived*

Coherence Relations

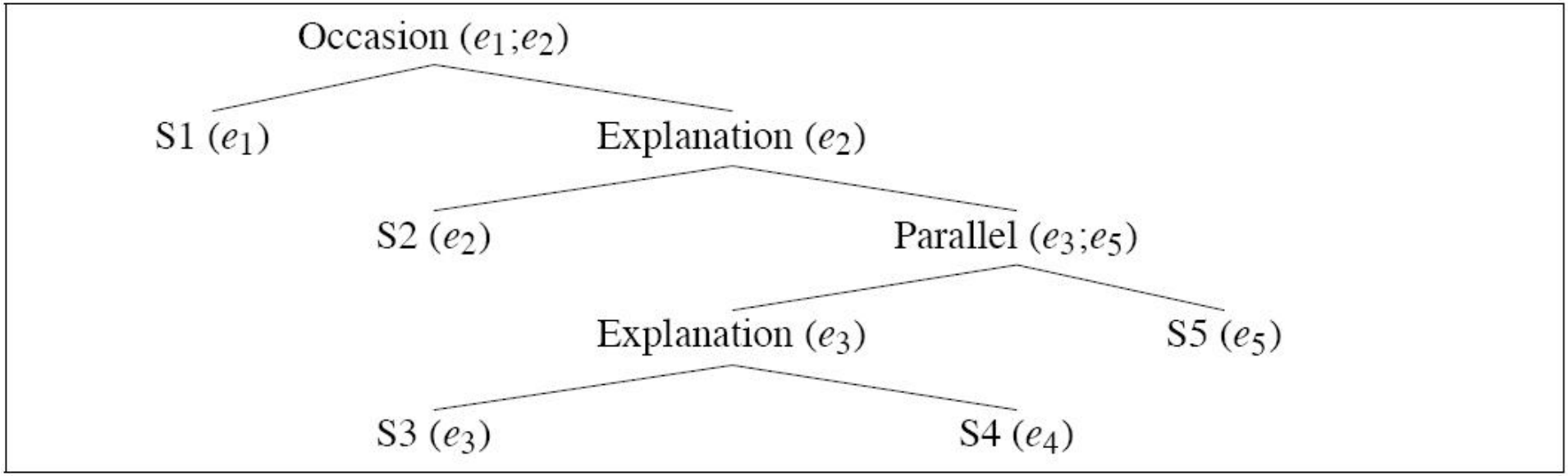
S1: *John went to the bank to deposit his paycheck*

S2: *He then took a bus to Bill's car dealership*

S3: *He needed to buy a car*

S4: *The company he works for now isn't near a bus line*

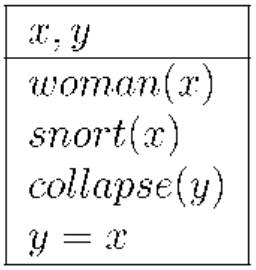
S5: *He also wanted to talk with Bill about their soccer league*



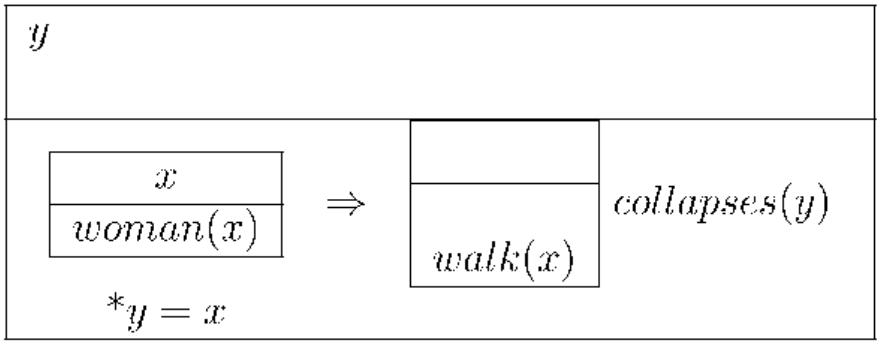
Simple DRS example (DRT by Kamp)

6.12. Example (Preceded by "A woman snorts".)

1. A woman walks. She collapses.



2. Every woman walks. *She collapses.



from Raffaella Bernardi, Trento

Pragmatics

Pragmatics is a branch of linguistics dealing with language use in context.

When a diplomat says yes, he means ‘perhaps’;
When he says perhaps, he means ‘no’;
When he says no, he is not a diplomat.

(Variously attributed to Voltaire, H. L. Mencken, and Carl Jung)

Quote from http://plato.stanford.edu/entries/pragmatics/

In Context?

- Social context
  - Social identities, relationships, and setting
- Physical context
  - Where? What objects are present? What actions?
- Linguistic context
  - Conversation history
- Other forms of context
  - Shared knowledge, etc.

(Direct) Speech Acts

- ***Mood*** of a sentence indicates relation between speaker and the concept (proposition) defined by the LF
- There can be operators that represent these relations:
  - ASSERT: the proposition is proposed as a fact
  - YN-QUERY: the truth of the proposition is queried
  - COMMAND: the proposition describes a requested action
  - WH-QUERY: the proposition describes an object to be identified

Indirect Speech Acts

- Can you pass the salt?

- It's warm in here.

Task-Oriented Dialogue

- Making travel reservations (flight, hotel room, etc.)
- Scheduling a meeting.
- Task oriented dialogues that are frequently done with computers:
  - Finding out when the next bus is.
  - Making a payment over the phone.

Ways to ask for a room

- I’d like to make a reservation
- I’m calling to make a reservation
- Do you have a vacancy on ...
- Can I reserve a room
- Is it possible to reserve a room

Task-oriented dialogue acts related to negotiation

- Suggest
  - I recommend this hotel.
- Offer
  - I can send some brochures.
  - How about if I send some brochures.
- Accept
  - Sure. That sounds fine.
- Reject
  - No. I don’t like that one.

"No, Thursday's out. How about never—is never good for you?"

Now, a famous bad idea
(linked to a good idea)

Grice's Maxims

- Why do these make sense?
  - *Are you 21?*
  - *Yes. I'm 25.*

  - *I'm hungry.*
  - *I'll get my keys.*

  - *Where can I get cigarettes?*
  - *There is a gas station across the street.*

Grice's Maxims

- Why are these strange?
  - (The students are all girls.)
  - *Some students are girls.*
  - (There are seven non-stop flights.)
  - *There are three non-stop flights.*
    - Jurafsky and Martin, page 820
  - (In a letter of recommendation for a job)
  - *I strongly praise the applicant's impeccable handwriting.*

Grice's *Cooperative Principle*

- "Make your contribution such as it is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."

- The Cooperative Principle is good and right.

- On the other hand, we have the Maxims:

Grice's actual Maxims

- Maxim of Quality
  - Try to say something true; do not say something false or for which you lack evidence.
- Maxim of Quantity
  - Say as much as is required to be informative
  - Do not make your contribution more informative than required
- Maxim of Relevance
  - Be Relevant
- Maxim of Manner
  - Be perspicuous
  - Avoid ambigtuity
  - Be brief
  - Be orderly

Flouting the Cooperative Principle

- "Nice throw." *(said after terrible throw)*
- "If you run a little slower, you'll never catch up to the ball." *(during mediocre pursuit of ball)*
- You *can* indeed imply something by clearly violating the principle.
  - The Maxims *still* suck.

Flout ≠ *Flaunt*

- *Flout*: openly disregard (a rule, law or convention).

- *Flaunt*: display (something) ostentatiously, especially in order to provoke envy or admiration or to show defiance.
  - Source: Google

My paper on the Maxims

- Grice's Maxims: "Do the Right Thing" by Robert E. Frederking. Argues that the Gricean maxims are too vague to be useful for natural language processing. [from Wikipedia article]

- "I used to think you were a nice guy."
  - Actual quote from a grad student, after reading the paper

Questions?