Pragmatics and Discourse

- **Pragmatics**
  - Defined by Crystal as “The study of the factors influencing a person’s choice of language

- **Discourse**
  - Defined by Crystal as “The study of patterns of linguistic organization in” units of language larger than a sentence

- **A lot of overlap in what they study**
  - Difference not important for this class
  - You may hear both terms, especially in NLP where authors tend to be a bit freer with linguistic terminology (Personal Observation)
In general, reference resolution is the process of determining what
mentions (usually noun phrases) in text refer to in the real world.

Examples

○ This morning on the radio I heard a traffic report about “the Beltway”.
  ■ Was the announcer referring to Interstate 495 or Interstate 695?
  ■ What information would help you determine this
○ What does it refer to in the sentence “I cooked the pizza in the oven and then ate it.”

In NLP these are generally considered two different tasks

○ The first example is called Entity Linking, which is labeling text with a unique identifier
  ■ Could be a wikipedia page or something
○ The second example is what is usually considered Reference Resolution
Coreference Resolution

- Find all entities in a text that refer to the same thing in the real world.

Ben Olsen has been the coach of DC United since 2010. The former midfielder has been with the club for 18 years in total. He played in the 2006 World Cup for the United States, scoring 6 goals in 37 appearances for the team over the course of his career.
Coreference Resolution

- Find all entities in a text that refer to the same thing in the real world.

Ben Olsen has been the coach of DC United since 2010. The former midfielder has been with the club for 18 years in total. He played in the 2006 World Cup for the United States, scoring 6 goals in 37 appearances for the team over the course of his career.
Coreference Resolution

- Find all entities in a text that refer to the same thing in the real world.

Ben Olsen has been the coach of DC United since 2010. The former midfielder has been with the club for 18 years in total. He played in the 2006 World Cup for the United States, scoring 6 goals in 37 appearances for the team over the course of his career.
Ben Olsen has been the coach of DC United since 2010. The former midfielder has been with the club for 18 years in total. He played in the 2006 World Cup for the United States, scoring 6 goals in 37 appearances for the team over the course of his career.
Why is the useful?

● Relation Extraction and Question Answering
  ○ Often times the facts we are looking for are not in the same sentence, but are in adjacent sentences.
  ○ What Would Cup did Ben Olsen play in?

● Vector Space Models
  ○ Vector space models are based on similar words having similar contexts.
  ○ Many nouns don’t actually occur in the contexts we might expect, because co-referents are often used instead.
Referring Expressions

- Referring Expressions are mentions in the text that indicate reference is needed
- What they refer to is known as the referent
  - Technically, this is the entity in the real world and is never present in text
  - Sometimes you might find referent used to describe the words in text that a pronoun points to.
- All noun phrases have the potential to be referring expressions
- Knowing the type of noun phrase can help resolve references
Indefinite Noun Phrases

- Noun Phrases that start with the indefinite determiner (a or an), a quantifier, or sometimes the pronoun this
- Usually signifies a new entity being talked about.
  - May or may not be referenced later
- Examples

  Last night my brother made me a **bowl of soup**.

  A **part of your homework** involves relations extraction.

  **Some food places** on campus are open on the weekend.
Definite Noun Phrases

- Noun Phrases that start with the definite determiner (*the*), or are otherwise identifiable by the hearer/reader
- May signal a new entity or may refer to a previously mentioned entity
- Examples

  Last night my brother made me a bowl of soup. *The soup* was delicious.

  A part of your homework involves relations extraction. *The problem* is about languages

  *The best place* to eat on campus is Chick Fil A
Pronouns

- Pronouns almost never introduce new entities into the text
- Because pronouns refer to something previously mentioned, they require that the entity referred to still be somewhat central to the text.
  - This is known as salience
  - The entity a pronoun refers to is usually only a few sentences back because of this requirement.

- Gotchas
  - Cataphora is like anaphora, but the pronoun is used a little bit before the thing it refers to
    
    They're creepy and they're kooky, Mysterious and spooky, They're altogether ooky, The Addams Family.

- Pleonastic pronouns don’t refer to anything, as see in “It rained yesterday.”
Performing Anaphora Resolution

- **Hobb’s Algorithm**
  - Heuristic Based Algorithm that searches over parse trees

- **Centering Algorithm**
  - Heuristic Based Algorithm rooted in Discourse Theory

- **Supervised Machine Learning**
  - Create a feature vector for each possible anaphora-referent pair in the text.
  - Train a binary classifier to predict if the pair is a reference resolution or not
Hobb’s Algorithm

- Presented in 1978
- Still works well enough to be used a baseline
- Searches over parse trees and only uses two features
  - Morphological Gender
  - Grammatical Number
1. Start a the Noun Phrase which contains the pronoun of interest
2. From there, go up the tree until another NP or S node is encountered
   a. This higher NP/S will be referred to as X
   b. The path between the first NP and X is called $p$
3. Go through all children of X that are to the left of $p$
   a. If any NP is found that also has another NP or S between it X say this is the antecedent and end the algorithm
4. Otherwise, if X is the root of the sentence, search the previous sentences, left-to-right breadth first in the tree for an NP.
Hobb’s Algorithm

5. If $X$ is not at the top of the tree, find the next NP or S up the tree. This is the new $X$ and the path between the old $X$ and the new $X$ is $p$.

6. If $X$ is an NP and $p$ does not go through a noun immediately under $X$, return $X$ as the antecedent.

7. Otherwise, go through all nodes under $X$ and to the left of $p$ in LR, BFS order. Return the first NP found.

8. If $X$ is an S node, search the nodes under $X$ and to the right of $p$, but do not enter any NP or S nodes. Return first NP found.

Perform Hobb’s Algorithm on the following examples:

Ben Olsen has been the coach of DC United since 2010. He played in the 2006 World Cup for the United States.

The Capitals play in the first playoff game tonight. The Toronto Maple Leafs are their opponent.
To do supervised machine learning we need an annotated corpus that contains examples like

[Ben Olsen\textsuperscript{\text{\texttt{\textsc{ANT}}}}] has been the coach of DC United since 2010. [He\textsuperscript{\text{\texttt{\textsc{PRO}}}}] played in the 2006 World Cup for the United States.

Given these pairs, construct a feature vector for the pair

- Also need to construct negative examples automatically

Pass the feature vector off to a classifier that predicts if a pair of words are in an antecedent-pronoun relation

- Loglinear and Naive Bayes are popular choices, but we aren’t concerned with the internal details.
Supervised Machine Learning Features

- **Boolean Features**
  - Does the grammatical number of the pairs match
  - Does the grammatical gender of the pairs match
  - Is the grammatical number compatible if the number of one part is unknown
  - Is the grammatical gender compatible if the gender of the antecedent is unknown.

- **Numerical Features**
  - Number of sentences between the pair (0 = same sentence and so on)
  - Number of NPs Hobb’s algorithm would have to skip over to get correct answer

- **Categorical Features**
  - Grammatical role (of antecedent usually, but could use pronoun too)
  - Type of referring expression
Ben Olsen has been the coach of DC United since 2010. He played in the 2006 World Cup for the United States.

<table>
<thead>
<tr>
<th></th>
<th># Match</th>
<th>Gen Match</th>
<th># Comp</th>
<th>Gen Comp</th>
<th># Sents</th>
<th># Hobbs</th>
<th>Role</th>
<th>Ref Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>BO-he</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1</td>
<td>0</td>
<td>SUBJ</td>
<td>DEF</td>
</tr>
<tr>
<td>DCU-he</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>1</td>
<td>3</td>
<td>POBJ</td>
<td>DEF</td>
</tr>
</tbody>
</table>
What is the feature vector for the following text?

The Capitals play in the first playoff game tonight. The Toronto Maple Leafs are their opponent.

<table>
<thead>
<tr>
<th># Match</th>
<th>Gen Match</th>
<th># Comp</th>
<th>Gen Comp</th>
<th># Sents</th>
<th># Hobbs</th>
<th>Role</th>
<th>Ref Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Their</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ML-Their</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Coreference resolution

- Coreference resolution can be done many ways, but one of the most common is to use a classifier like before.
- Classification is still done on a pairwise basis, to establish links between particular words or phrases.
- In addition to the previous features, features for coreference often include
  - Edit distance from anaphor to antecedent and vice-versa
  - If the anaphora was found in an appositive
- This method will produce chains of co-referrenants that can be returned.
Coreference resolution

- After determining if pairs of entities are likely co-referents, the set of all entities in a can be viewed as a graph.
  - An edge between nodes represents a predicted link.
Coreference resolution

- After constructing the graphs, standard graph algorithms can be used for post-processing.
- Common methods
  - Clustering
  - Partitioning
- A variation of this would be rather than binary classification, the ML algorithm produces a probability, which is then used as a weight on the graph edge.
A common evaluation method for co-reference is known as B-CUBED.

B-CUBED is just a formulation of precision and recall when dealing with hypothesis chains.

\[
B^3 \text{ Precision} = \sum_i \frac{\# \text{ of correct elements in entity class } i}{\# \text{ of elements in entity class } i}
\]

\[
B^3 \text{ Recall} = \sum_i \frac{\# \text{ of correct elements in entity class } i}{\# \text{ of elements in reference entity class } i}
\]
Conclusions

- Reference resolution is an active area of research in NLP
  - It is a very hard and far from solved problem
- Personal Observation
  - While reference resolution would help many applications and techniques including things like vector semantics, it is not often used
    - It is researched though!
  - I think when developing NLP applications and programs, people want to do the minimum amount of work needed and quickly
    - Reference resolution is hard, and easy to skip because not doing it won’t break things usually