Coreference

• Two words/phrases are coreferent if they refer to the same object or event.

• An earlier description is called the antecedent and the subsequent referring expression is called the anaphor. The plural term is anaphora.

• There are many different forms of anaphora, including pronouns, relative pronouns, definite references, and appositives.

• Coreference is everywhere! Especially noun phrases, though verb phrases can also be coreferent.

Coreference Resolution

Coreference is a relation between two or more textual entities.

• \( \alpha_1 \) and \( \alpha_2 \) corefer if and only if \( \text{referent}(\alpha_1) = \text{referent}(\alpha_2) \).

• If \( \alpha_i \) is limited to NPs, the problem is \( NP \text{ Coreference} \).

• \( \text{referent}(x) \) is the entity or concept referred to by \( x \).

• Coreference is an equivalence relation: reflexive, symmetric and transitive.

Coreference resolution is the process of determining which \( \alpha \)s refer to one another.

Sample Text

Internet search leader Google is snapping up YouTube for $1.65 billion, brushing aside copyright concerns to seize a starring role in the online video revolution.

THE ALL-STOCK DEAL announced Monday unites one of the Internet’s marquee companies with one of its rapidly rising stars. IT came just a few hours after YouTube unveiled three separate agreements with media companies to counter the threat of copyright-infringement lawsuits. The price makes YouTube Inc., a still-unprofitable startup, by far the most expensive purchase made by Google during its eight-year history.

Why is coreference resolution important?

• Coreference is a pervasive phenomenon in natural language.

• The problem lies at the intersection of syntax, semantics and discourse.

• Coreference resolution is essential for natural language understanding.

• Coreference resolution is necessary for many NLP applications, such as information extraction, question answering, and summarization.
Coreferent noun phrases

Coreferent noun phrases are extremely common and appear in many forms. Consider this short story:

John F. Kennedy was a great president.
John Kennedy was married to Jackie Onassis.
JFK was assassinated.
Mr. Kennedy died very young.
The former president is buried in Arlington National Cemetery.
He is buried next to his brother, Bobby Kennedy.
The country will never forget him.
The favorite son of Massachusetts is a hero to many.

Intrasentential Coreference

Coreference can occur within a sentence:

The actress who won the awards was excited.
I saw the actress who won the awards.
I saw the actress wearing the red and white dress with the diamond earrings in the corner who won the awards.
I saw the actress from Virginia who won the awards.
I saw the actress and actor who won the awards.
Please give me the pen on the table that has three legs.
Please give me the pen on the table that has red ink.

Appositives

An appositive is a pair of coreferent NPs separated by a comma.

- Bill Clinton, the former president, lives in New York.
- The head of Microsoft, Bill Gates, is very rich.
- My two dogs, Rover and Fido, always attack the mailman.
- Rover, Fido, and Fluffy, my three dogs, play together a lot.
- Rover, Fido, Fluffy, my cat, play together a lot.
- Rover, my cat, Fido, my fish, and Fluffy, my hedgehog, play together.
- Rover, Fido, and my turtle play together a lot.

Many objects are mentioned implicitly

- I bought a used car. The horn doesn’t work.
- I finished writing a book, but I still have to write the dedication.
- I went to a concert last night. The intermission was long.
- Henry was flying a kite. The string broke and it blew away.
References to non-entities

S1: Jack went to New Orleans last year.

S2: He loves to go there.
S2: Mardi Gras was on then.
S2: The trip changed his life.
S2: It changed his life.
S2: He does that every year.
S2: Sam did it last year too.
S2: That really surprised Helen.

Rule-based Approaches

- Many anaphora can be resolved with reasonable accuracy using relatively simple rules and a history list. Resolve against the nearest entity that satisfies constraints.

- Hand-built rules typically rely on:
  - String matching
  - Scoping heuristics and recency
  - Number, person, and gender agreement
  - Syntactic heuristics
  - Semantic compatibility (e.g., NER or semantic class types)

String Matching

- String matching can handle many cases, with varying degrees of reliability.
  - Proper Names: exact string match is often quite reliable within a document.
  - Common NPs: exact string match is often effective, but sometimes risky.
  - Partial Matching can capture many cases but adds risk.
  - Added complexity can include titles, abbreviations, and acronym recognition.

Scoping Restrictions

The scope of an anaphor is the region where you expect to find the antecedent. Different types of anaphora require different scopes.

- The scope of reflexive pronouns is the same clause.
- The scope of relative pronouns is the same sentence.
- Genderless pronouns typically have a limited scope (about 1-2 sentences). Gendered pronouns have a somewhat larger scope.
- Common and proper nouns have a wide scope: the antecedent may be very far from the anaphor.
Agreement Heuristics

Anaphora must agree with their antecedent in:

**Number:**
- Jane bought 3 books from the store that was burned down yesterday.
- Jane bought 3 books from the store that were worth a lot of money.

**Gender:**
- Fred gave Mary a book. He was happy.
- Fred gave Mary a book. She was happy.

**Animacy:**
- I gave a book to Mary that was a NY Times bestseller.
- I gave a book to Mary who was celebrating her 20th birthday.

Pronouns can sometimes appear earlier than the referent (cataphora)

- In its dreams, the goldfish was a shark.
- Before she went to the party, Jill bought some wine.
- Most of his friends think that Joe will win.
- That which we fear most is fear itself.

And definite references may be the first reference.

- The assassination of JFK is still being investigated.
- The truth will never be known.
- Those conspiracy theory rumors will never end.

Non-anaphoric NPs

A major challenge is determining which NPs have a prior referent! Many NPs that you might expect to be anaphoric are not.

- **Pleonastic pronouns** do not have a prior referent.
  - Ex: It is raining... It is true... It is necessary...

- Many definite NPs are non-anaphoric due to:
  - Restrictive pre- and post-modification.
    - Ex: the SLC mayor; the man who shot JFK.
  - World knowledge.
    - Ex: the White House; the truth; the police; the weather.
  - Bridging references.
    - Ex: the ball; the hoop; the score; the referees

The Hobbs Algorithm for Pronoun Resolution

- The **Hobbs algorithm** is a relatively simple and reasonably effective syntactic method for resolving pronominal anaphora.
- This algorithm methodically searches the parse trees of the sentence containing the pronoun and (as needed) previous sentences.
- The search procedure is (roughly) a breadth-first, left-to-right search of the parse tree (see textbook for details).
- The first candidate NP that satisfies gender, number, and person agreement constraints is returned as the antecedent.
- The Hobbs algorithm is commonly used as a baseline when evaluating pronoun resolution methods.
Machine Learning Approaches

Today, most state-of-the-art coreference resolvers use machine learning.

- Supervised learning techniques are used to create a classifier that can identify the antecedent of an anaphor.
- A gold standard corpus labeled with coreference chains is used for training.
- A final clustering stage merges the pairwise coreference decisions made by the classifier into coreference chains.

Typical Coreference Pipeline

- **Preprocessor**: XML removal, tokenization, sentence and paragraph splitting.
- **Part-of-Speech Tagging**
- **Parsing**
- **Named Entity Recognizer**
- **Semantic Class Lookup (usually via WordNet)**
- **Candidate NP extraction**
- **Supervised Learning**
- **Clustering into Chains**

String Matching Features

A lot of resolutions can be made solely on the basis of string-based comparisons!

- **Exact Matching**: exact match of NPs or head nouns.
- **String Matching**: substring and partial matching of the NPs. Ex: “Apple” and “Apple Computer Inc.”
- **Word Overlap**: word overlap comparisons of the NPs. Ex: “John F. Kennedy” and “President John Kennedy”
- **Edit Distance**: edit distance measures between the NPs. Ex: “Creutzfeldt-Jakob Disease” and “Creutzfeldt-Jacob Disease”

Lexical Similarity

Comparing lexical similarity can help find antecedents for many anaphora.

- Ford Motor Company = Ford Co. = Ford
- Apple Computer = Apple
- Silicon Graphics Inc. = SGI
- Federal Bureau of Investigation = FBI
- The United States = The U.S.
- kidnapped ⇒ kidnapping
Semantic Agreement

I adopted a terrier from the animal shelter. The dog was 3 years old.

I bought my son a dog. Jack was thrilled.

I forgot to water the plant on the table yesterday. Its leaves fell off.

3 people were kidnapped yesterday. The kidnapping happened in Bogota.

The assassination of John F. Kennedy stunned the nation. He was murdered by a man near the grassy knoll.

Limitations of Semantic Features

But many candidates have compatible semantic classes, so semantic features are not sufficient to pinpoint the correct antecedent. For example:

• I hired a teacher to tutor my son. He charged $100.
• The tsunami killed 100 people including 30 children. They were in school when the tsunami hit.
• I witnessed the robbery of two tourists by masked men. The perpetrators...
• I witnessed the robbery of two tourists by masked men. The French couple was visiting New York when...
• I witnessed the robbery of two tourists by masked men. Witnesses said that the gunmen ...

Other Features

- **Proximity**: sentence and paragraph-based distance measures between the NPs.
- **Aliases**: titles, acronyms, etc.
- **Number**: number agreement (singular/plural)
- **Gender**: gender agreement (singular/plural)
- **Syntactic**: syntactic properties and comparisons
- **Appositive**: appositive recognition
- **Semantic**: animacy and semantic compatibility comparisons.
- **NER**: Named Entity Recognition
- **Hobbs Distance**

Summary

- Different types of anaphora pose different challenges.
- Hand-built rules can resolve many cases using relatively simple string-matching, scoping heuristics, and agreement checks.
- Identifying non-anaphoric NPs is a substantial problem in its own right.
- Coreference resolution is far from solved, but progress has been made.
  - Current systems achieve about 70-80% F measure scores for full, from-scratch NP coreference.