Reference Resolution

CSCI-GA.2590

Ralph Grishman
Reference Resolution: Objective

• Identify all phrases which refer to the same real-word entity
  – first, within a single document
  – later, also across multiple documents
Terminology

*referent*: real-world object referred to

*referring expression* [mention]: a phrase referring to that object

Mary was hungry; she ate a banana.
Terminology

• *coreference*: two expressions referring to the same thing

Mary was hungry; she ate a banana.

antecedent anaphor
(prior expression) (following expression)

• So we also refer to process as *anaphora resolution*
Types of referring expressions

- definite pronouns (he, she, it, ...)
- indefinite pronouns (one)
- definite NPs (the car)
- indefinite NPs (a car)
- names
Referring Expressions: pronouns

Definite pronouns: he, she, it, ...

• generally anaphoric
  – Mary was hungry; she ate a banana

• pleonastic (non-referring) pronouns
  – It is raining.
  – It is unlikely that he will come.

• pronouns can represent bound variables in quantified contexts:
  – Every lion finished its meal.
Referring Expressions: pronouns

Indefinite pronouns (one)

• refers to another entity with the same properties as the antecedent
  – Mary bought an IPhone6.
  – Fred bought one too.
  – *Fred bought it too.

• can be modified
  – Mary bought a new red convertible.
  – Fred bought a used one.
    = a used red convertible
    (retain modifiers on antecedent which are compatible with those on anaphor)
Referring Expressions: pronouns

Reflexive pronouns (himself, herself, itself)
• used if antecedent is in same clause
  – I saw myself in the mirror.
Referring expressions: NPs

NPs with definite determiners ("the")
• reference to uniquely identifiable entity
• generally anaphoric
  – I bought a Ford Fiesta. The car is terrific.
• but may refer to a uniquely identifiable common noun
  – I looked at the moon
  – The president announced ...
• or a functional result
  – The sum of 4 and 5 is 9.
  – The price of gold rose by $4.
Referring expressions: NPs

NPs with indefinite determiners (“a”)

• generally introduces a new ‘discourse entity’
• may also be generic:
  – A giraffe has a long neck.
Referring expressions: names

• subsequent references can use portions of name:
  – Fred Frumble and his wife Mary bought a house. Fred put up a hammock.
Complications

- Cataphora
- Bridging anaphora
- Zero anaphora
- Non-NP anaphora
- Conjunctions and collective reference
Cataphora

• Pronoun referring to a following mention:
  – When she entered the room, Mary looked around.
Bridging Anaphora

• Reference to related object
  – Entering the room, Mary looked at the ceiling.
Zero Anaphora

- many languages allow subject omission, and some allow omission of other arguments (e.g., Japanese)
  - these can be treated as zero (implicit) anaphors
    - similar resolution procedures
  - some cases of bridging anaphora can be described in terms of PPs with zero anaphors:
    - "IBM announced the appointment of Fred as president of IBM."
Non-NP Anaphora

- Pronouns can also refer to events or propositions:
  - Fred claimed that no one programs in Lisp. That is ridiculous.
Conjunctions and collective reference

• With a conjoined NP,
  ... Fred and Mary ...
  we can refer to an individual ("he", "she") or
  the conjoined set ("the")

• We can even refer to the collective set if not
  conjoined ...
  “Fred met Mary after work. They went to the
  movies.”
Resolving Pronoun Reference

- Constraints
- Preferences
- Hobbs Search
- Selectional preferences
- Combining factors
Pronouns: constraints

Pronoun must agree with antecedent in:

• animacy
  – Mary lost her husband and her notebook. It was last seen in WalMart.

• gender
  – Mary met Mr. and Mrs. Jones. She was wearing orange pants.
  – needs first-name dictionary
  – some nouns gender-specific: sister, ballerina

• number
  – some syntactically singular nouns can be referred to by a plural pronoun: “The platoon ... they”
Pronouns: preferences

Prefer antecedents that are

• recent
  – at most 3 sentences back

• salient
  – mentioned several times recently

• subjects

Recency and preference for subjects are often captured by Hobbs search order, a particular order for searching the current and preceding parse trees
Hobbs search order

• traverse parse tree containing anaphor, starting from anaphor

then

• traverse trees for preceding sentences, breadth first, left-to-right
  • incorporates subject precedence

• stop at first NP satisfying constraints
  • relatively simple strategy, competitive performance
Pronouns: selectional preferences

• Prefer antecedent that is more likely to occur in context of pronoun
  – Fred got a book and a coffee machine for his birthday. He read it the next day.
  – can get probabilities from a large (parsed) corpus
Pronouns: combining probabilities

\[
P = P \text{(correct antecedent is at Hobbs distance } d) \times P \text{(pronoun } | \text{ head of antecedent)} \times P \text{(antecedent } | \text{ mention count)} \times P \text{(head of antecedent } | \text{ context of pronoun)}
\]

Ge, Hale, and Charniak 1998

83% success
Making a General Resolver
Resolving names

- Generally straightforward: exact match or subsequence of prior name
  - some exceptions for locations
Resolving common noun phrases

• generally difficult

• typical strategies for resolving “the” + N:
  – look for prior NP with same head N
  – look for prior name including token N
    • “the New York Supreme Court” ... the court

• more ambitious: learn nouns used to refer to particular entities by searching for “name, N” patterns in a large corpus
  – “Lazard Freres, the merchant bank”
Types of models

• mention-pair model
  – train binary classifier: are two mentions coreferential?
  – to apply model:
    • scan mention in text order
      – link each mention to the closest antecedent classified +
      – link each mention to antecedent most confidently labeled +
    • cluster mentions
  – weak model of partially-resolved coreference

• entity-mention model
  – binary classifier: is a mention part of a partially-formed entity?
  – richer model: entity has features from constituent mentions
Anaphora resolution in Jet

• ‘resolve’ operation
• only processes noun groups
• basically an entity-mention model
  – create entity annotations
  – single pass, adding mentions to entities
• display entities in separate window
Using resolver for implicit arguments

Example: extended version of AppointPatterns (see notes)
Diversity of approaches

Two recent systems show range of approaches (see notes):

- Stanford [CL 2013]
  - hand-coded rules
  - 10 passes over complete document, using rules of decreasing certainty
- Berkeley [EMNLP 2013]
  - classifier trained over large corpus with simple feature set
  - single pass
- No system does well on anaphoric NPs
Evaluation

• Coreference key is a set of links dividing the set of mentions into coreference classes
• System response has similar structure
• How to score response?
• MUC scorer
  – based on links ...
    recall error = how many links must be added to system response so that all members of a key set are connected by links
  – Does not give credit for correct singleton sets
Evaluation

• B-cubed metric:
  – Mention-based
  – For each mention m,
    \[ r = \text{size of response set containing m} \]
    \[ k = \text{size of key set containing m} \]
    \[ i = \text{size of intersection of these sets} \]
    \[ \text{recall}(m) = \frac{i}{k} \]
    \[ \text{precision}(m) = \frac{i}{r} \]
  – Then compute average of recall, average of precision
A Coherent Discourse

• A text is not a random collection of facts
• A text will tell a story, make an argument, ...
• This is reflected in the structure of the text and the connections between sentences
• Most of these connections are implicit, but a text without these connections is incoherent

Fred took an NLP course in the Spring.
He got a great job in June.

Fred took an NLP course in the Spring.
He got a great cat in June.
A Coherent Discourse

• Criteria for coherence depend on type of text
• Most intensively studied for narratives
  – causal connections
  – temporal connections
  – scripts (conventional sequences)
Coherence and coreference

Select anaphora resolution more consistent with coherence.

Jack poisoned Sam. He died within a week. vs.
Jack poisoned Sam. He was arrested within a week.

How to do this in practice?
Collect from a large corpus a set of predicate/role pairs, such as:

- subject of poison -- subject of arrest
- object of poison -- subject of die.

Prefer anaphora resolution consistent with such pairs.
Cross-document Coreference

Quite different from within-document coref:

• within document (single author or editor)
  – a single person will be consistently referred to by the same name
  – the same name will consistently refer to the same person

• across documents
  – the same person may be referred to using different names
  – a single name may refer to multiple people ("Michael Collins")
Limitation

• Assume each document separately resolved internally

• Only link entities which are named in each document
  – general NPs very hard to link
  – “Fred’s wife” may refer to different people at different times
  – details may change over time:
    • “the dozens of people killed in the bombing”
    • “the 55 people killed in the bombing”
Two Tasks

• Entity linking: map each document-level entity to an entry in a standard data base
  – e.g., wikification
  – entities not in data base are left unlinked

• Cross-document coreference
  – cluster all document-level entities

• Tasks have a lot in common
  – often cross-doc coreference begins with entity linking against a large knowledge base or Wikipedia
Features

Features for cross-doc coref:

• Internal (name) features
• External (context) features
  – whole-document features
  – local context features
  – semantic features
• Consistency
Internal (name) features

Finding a match:
• exact match suitable for edited text in languages with standard romanization
• use edit distance for informal text
• use edit distance or pronunciation-based measure for other languages (e.g., Arabic)

Estimating probability of coref for exact match:
  – for people, use name perplexity, based on
    • number of family names with same given name
    • number of given names with same family name
External (Context) Features

Names are more likely to be coreferential if:

- documents are similar (using tf-idf cosine similarity)
- local contexts are similar
- values of extracted attributes match (birthplace, religion, employer, ...)

Conversely, distinct values of some attributes (birthplace, birthdate) are strong indicators of non-coreferentiality
Consistent Wikification

• If multiple names are being resolved in a single document, they should preferably be resolved to related entities
  – if “New York” and “Boston” are mentioned in the same sentence, prefer that
    • both resolve to cities
    • both resolve to baseball teams
    • both resolve to hockey teams
  – in ranking referents, include as a factor the number of links connecting the referents
Consistent Wikification

“the Yankees faced Boston yesterday”

- New York Yankees
- Boston [city]
- Boston Red Sox
- Boston Bruins

Link in Wikipedia
Scaling Up

- Potential scale for cross-doc coref much larger
  - collection may have $10^7$ documents with 10-100 entities each: $10^9$ document-level entities
  - computing all pairwise similarities infeasible
  - use hierarchical approach to divide set
    - analog of entity-mention representation within a document
    - potentially with multiple levels (‘sub-entities’)
