Reference Resolution

Adam Meyers
New York University

Outline

- What is Reference Resolution?
- Linguistic Analysis of Coreference
- Coreference Algorithms: Proper Nouns, Pronouns, Common Nouns
- Evaluation Issues
- Summary

Reference Resolution

- Reference Resolution:
  - Which words/phrases refer to some other word/phrase?
  - How are they related?
- Anaphora vs. Cataphora
  - Anaphora: an ***anaphor*** is a word/phrase that refers back to another phrase: the ***antecedent*** of the anaphor
    - ***Mary*** *thought that* ***she*** *lost* ***her*** *keys.*
  - Cataphora (less common): a ***cataphor*** is a word/phrase that refers forward to another phrase: its **precedent**.
    - ***She*** *was at NYU, when* ***Mary*** *realized that she lost her keys.*
  - ***Anaphora*** is often used to refer to all ***Reference Resolution*** (including cataphora) and the term ***antecedent***t is often used instead of ***precedent***.

Types of Anaphora I

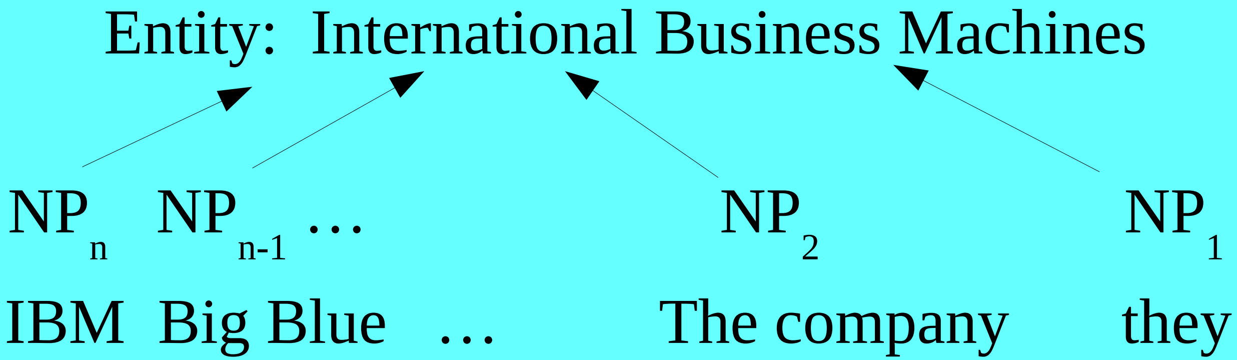
- Coreference: Antecedent = Anaphor
  - Though **Big Blue** won the contract, this official is suspicious of **IBM**.
  - **Mary** could not believe what **she** heard.
- Similar to Coreference
  - Type Coreference (vs. Token)
    - AKA, identify of sense (vs. identify of reference)
    - John ate **a sandwich** and Mary ate **one** also.
  - Bound variable
    - Every **lioness** guards **its** cubs
    - $(\forall \text{lioness}\, L)(\text{L guards L's cubs})$
- Predication and Apposition: some (not all) specs label as coreference
  - ***Mary*** is ***a basketball player***
  - ***Mary***, ***a basketball player from NYU***

Types of Anaphora II

- **Bridging Anaphora:** links between “related” objects
  - ***The amusement park*** *is very dangerous.* ***The gate*** *has sharp edges.* ***The rides*** *have not been inspected for years.*
- Some IE relation instances can be viewed as bridging
  - *When* ***the baby*** *cried,* ***the parents*** *rushed into the room.*
  - ACE Relation: **Per-Social.family**(***the baby***,***the parents***)
- **“Other” Anaphora:** words including *other* and *another* invoke an “other instance of type” relation
  - **This book** is valuable, but **the other book** is not.
- **Non-NP Anaphora**, e.g., events/propositions
  - ***Mary left the room***. ***This*** *upset her parents.*
  - *John* ***read the dictionary***. *Then Mary did* ***it*** *too.*

2 Models of NP Coreference

- **Chains of Coreference**: Which words/phrases co-refer with which other words/phrases, possibly forming a chain of the form:
  - $NP_n \leftarrow NP_{n-1} \leftarrow \ldots \leftarrow NP_2 \leftarrow NP_1$
  - IBM ← Big Blue ←… ← The company ← they
- **Mentions and Entities** (ACE)**:** Which phrases refer to the same object in the real world?



Entity: International Business Machines

NP_n NP_{n-1} … NP_2 NP_1

IBM Big Blue … The company they

Chain vs. Entity Model

- Entity model
  - Especially suited for fully spelled out names
  - Instances where coreference is based entirely on the discourse context and is not limited by proximity
    - Instances that are many lines apart
    - Cross-document coreference
- Chain Model
  - Especially suited to pronouns and definite common nouns that refer back to antecedent NPs
  - Instances in which the anaphor abbreviates, or provides a less specific descriptor than the antecedent
  - Instances of coreference where proximity of anaphor and antecedent is a factor

Coreference with different types of Nouns

- Coreference between Proper Nouns (NEs), including abbreviations, nicknames and substrings
  - Focus of most NLP systems: high precision/recall, links most informative NPs, ...
- Coreference between common noun phrases (CNPs) and preceding NPs (NEs and CNPs)
  - Worst system performance, least studied
- Coreference between pronouns and other NPs
  - Focus of largest body of theoretical work
  - Moderate system performance

Coreference between Proper Nouns (NEs) within a Document

- Instances of the same name string in a document usually refers to the same entity
  - ***IBM***, ***IBM***, ***IBM***, ***IBM***, … → ***Entity_IBM***
  - ***George Bush, George Bush, …*** → ***Entity_GB***
- Abbreviations and Nicknames match full name (full name is often first)
  - Abbreviations: mostly rule based (acronyms, subsequences, etc), Nicknames need a lexicon
  - Examples:
    - ***International Business Machines***, ***IBM***, ***Big Blue***… → ***Entity_IBM***
    - ***St. Petersburg*** → ***Saint Petersburg***
    - ***George Bush, George Bush, W, …*** → ***Entity_GB***
    - ***New York Yankees*** ←***New York***, ~~***New York Times***~~ ←***New York*** (place names only match some orgs)
- **S**imple rules work, links most informative NPs, results in high 90s, very little literature
  - Important component of IE systems
- One interesting problem: Name disambiguation
  - Distinguishing multiple individuals with the same name
  - Usually, only dealt with when doing cross-document coreference
    - Exception: *George Bush and his son George W were there*.
  - Abbreviation rules may allow two possible antecedents (*and then George said*)
  - Standardized abbreviations may not be unique,
    - ***AMEX*** → ***American Express*** or ***American Stock Exchange***

Pronouns in English

- **Definite Pronouns**: typically refer to specific NPs
  - 3rd person personal pronouns
    - ***he, him, his, she, her, hers, it, its, they, them, their, theirs***
  - 3rd person Reflexive pronouns
    - ***himself, herself, itself, themselves***
  - ***each other*** – reciprocal pronoun, similar to reflexives
  - *1st and 2nd person pronouns*
    - ***I, me, my, myself, mine, our, ours, ourselves, you, your, yours, yourself***
    - Dialogues between 2 people; or writer/speaker and audience
- **Indefinite Pronouns:**
  - ***one*** – can be used for type coreference
  - Other indefinites – no antecedents in text
    - ***something, someone, everything, everyone, ...***

3rd Def Prons: NonSyntactic Constraints/Preferences

- Usually have an antecedent
- Gender/number/person agreement (language specific)
  - ***Robert ← he, ~~Robert~~ ← she, ~~Robert~~ ← it, ~~Robert~~ ← they***
  - ***~~IBM~~ ← he, ~~IBM~~ ← she, IBM ← it, IBM ← they***
  - ***~~I~~ ← she, ~~me~~ ← her, ~~you~~ ← they***
- Selection Restrictions
  - ***Children*** *have many toys.* ***They*** *love to play.*
  - *Children have* ***many toys****.* ***They*** *are always breaking.*
- Pragmatics
  - *Mary yelled at* ***Alice****.* ***She*** *interrupted the phone call.*
  - ***Mary*** *yelled at Alice.* ***She*** *can be so mean sometimes.*
- Others: closer antecedents preferred, repeated NPs are more likely to be antecedents, etc. (J&M have several more examples)

Binding Theory Constraints

- An Antecedent of personal pronouns cannot be “too close” to the pronoun.
- An Antecedent of a reflexive/reciprocal pronoun cannot be “too far” from the pronoun.
- Definitions of “too close” and “too far”
  - Vary from language to language
  - Vary among different classes of pronouns/reflexives
  - Are defined using different primitive concepts within different linguistic theories
- Binding Theory Constraints are usually defined in terms of syntactic configurations

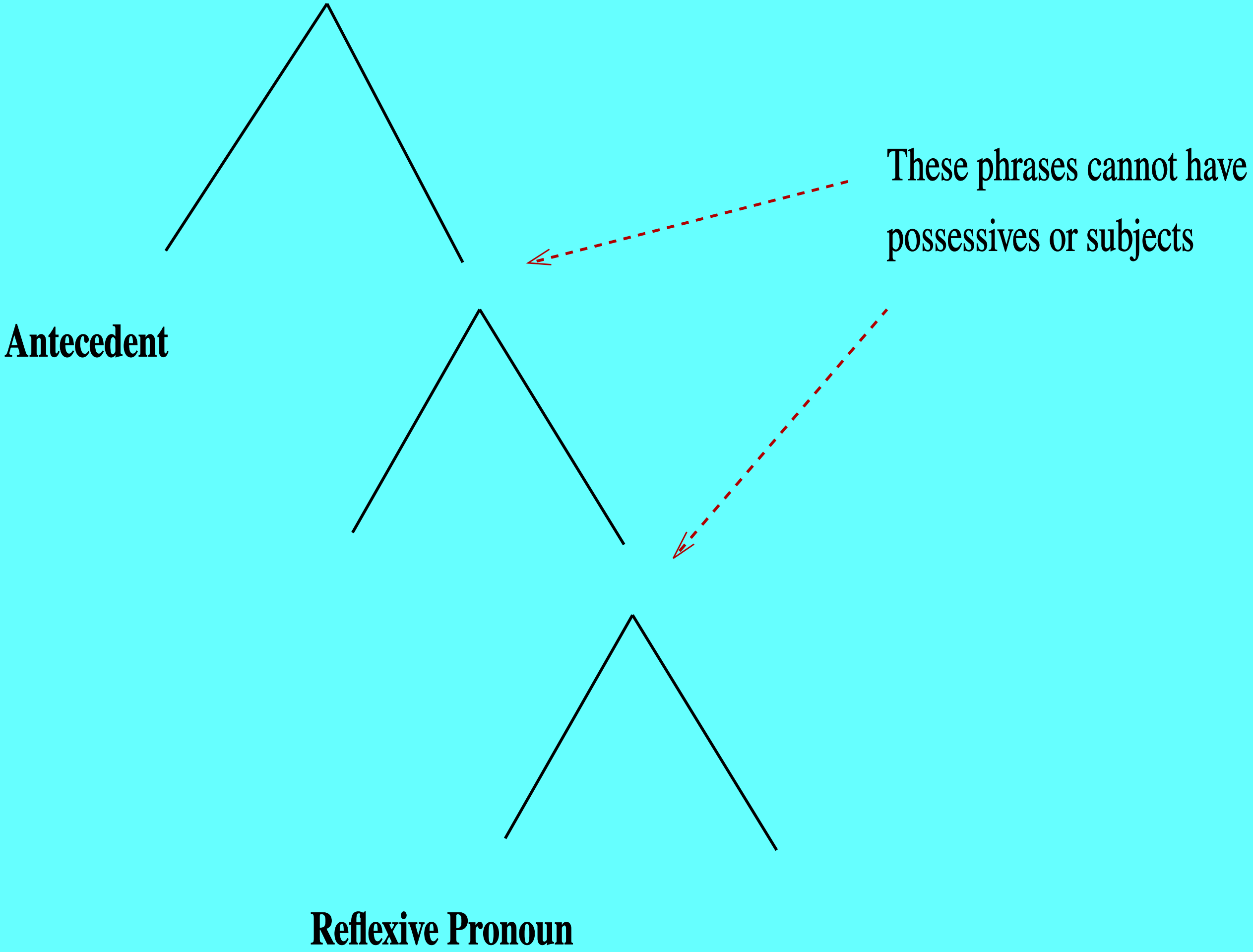
Binding Theory for English 3rd Pers Prons

- Case 1: If the pronoun ***p*** is inside an NP premodified by a possessive, the antecedent needs to be outside of this NP
  - ***John*** *likes* **[***Mary's drawing of* ***him*****]**
  - ***John*** *likes* **[*****his*** *drawing of Mary***]**
  - ***John*** *likes* **[*****Bill's*** *drawing of* ***him*****]**
- Case 2: Otherwise, the antecedent must be outside the immediate **tensed** clause containing the personal pronoun.
  - ***John*** *said that* **[*****he*** *liked pizza.***]**
  - ***John*** *wanted for* ***him*** *to like pizza.*
  - ***John*** *liked* ***him****.*
- Theories of binding vary about how these (and similar) constraints are encoded, but the differences in the final result (quality of system output) is minimal. While the above 2 rules cover most cases, there are also some exceptions:
  - ***John*** *always carries a slice of pizza with* ***him****.*

Binding Theory for English Reflexives/Reciprocals

- The antecedent of a reflexive/reciprocal **must be** the closest subject or possessive such that:
  - The antecedent precedes and "commands" the pronoun
    - **A** commands **B** if **A** is the sibling of a phrase that dominates **B**.
  - There is no possessive or subject for phrases in the path in the phrase structure tree between antecedent and pronoun
- Examples:
  - ___Mary___ _saw_ ___herself___ vs. *___Mary___ _said that John would meet_ ___herself___ ___soon___
  - ___Mary____'s picture of_ ___herself___ vs. *___Mary___ _saw John's picture of_ ___herself___
- These rules covers most cases.
  - Exception: _Pictures of_ ___themselves___ _made_ ___the actors___ _nervous._

Reflexive Pronoun Constraint



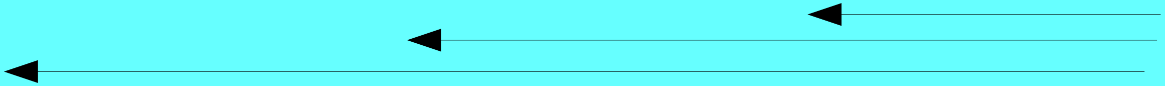
These phrases cannot have possessives or subjects

Antecedent

Reflexive Pronoun

This version of Binding Theory is English Specific

- *zìjǐ* – Chinese reflexive pronoun (example)
  - Ambiguous Example from Choi 1997
    - ***Zhangsan*** *renwei* ***Lisi*** *zhidao* ***Wangwu*** *xihuan* ***zìjǐ***



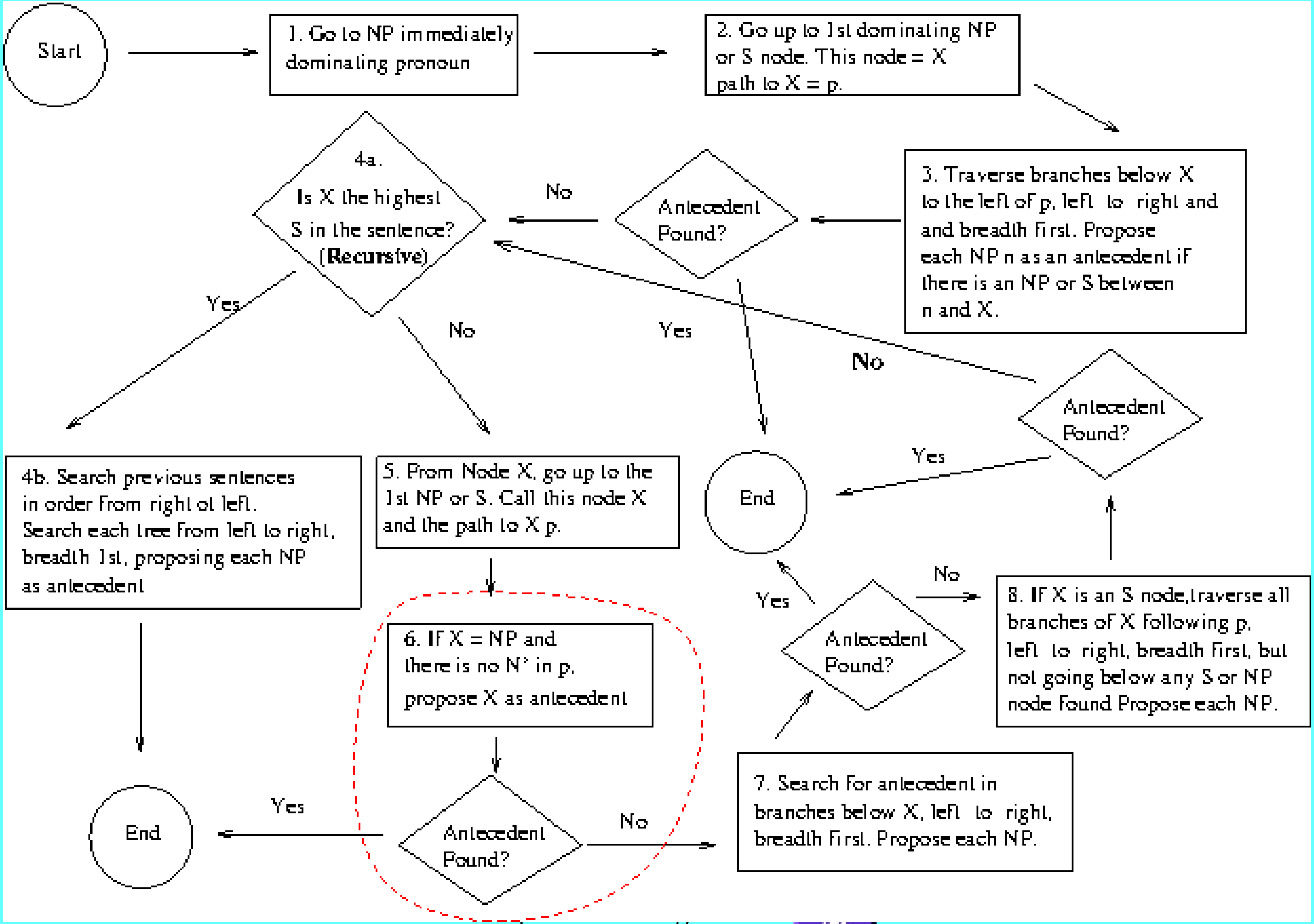
    - ***Zhangsan*** *thinks* ***Lisi*** *knows that* ***Wangwu*** *likes* ***self***
    - ***Zìjǐ*** can be coreferential with ***Zhangsan, Lisi*** or ***Wangwu***
  - In quasi-translated English, Wangwu would be the antecedent
    - ***Zhangsan*** *thinks* ***Lisi*** *knows that* ***Wangwu*** *likes* ***himself***

- Reflexive/Nonreflexive distinction holds across languages, but constraints on how close/far differ across languages: Icelandic, Chinese, etc.

Pronoun Resolution Methodology

- Hobbs search:
  - a simple system that provides a high baseline
  - Lappin and Leas (1994) report 82% F-score for Hobbs Search
  - **Ignore Portion in red oval for our purposes**
    - assumes NP → N' PP
      - where N' is larger than N, but smaller than NP
      - more details omitted
- Sets a High Baseline for Pronoun Coreference
- Higher Scoring Systems Tend to be Much More Complex

Hobbs Search Algorithm to Find Antecedent of Anaphors



Start

1. Go to NP immediately dominating pronoun

2. Go up to 1st dominating NP or S node. This node = X path to X = p.

3. Traverse branches below X to the left of p, left to right and and breadth first. Propose each NP n as an antecedent if there is an NP or S between n and X.

Antecedent Found?

4a. Is X the highest S in the sentence? (Recursive)

4b. Search previous sentences in order from right ot left. Search each tree from left to right, breadth 1st, proposing each NP as antecedent

5. From Node X, go up to the 1st NP or S. Call this node X and the path to X p.

6. If X = NP and there is no N' in p, propose X as antecedent

7. Search for antecedent in branches below X, left to right, breadth first. Propose each NP.

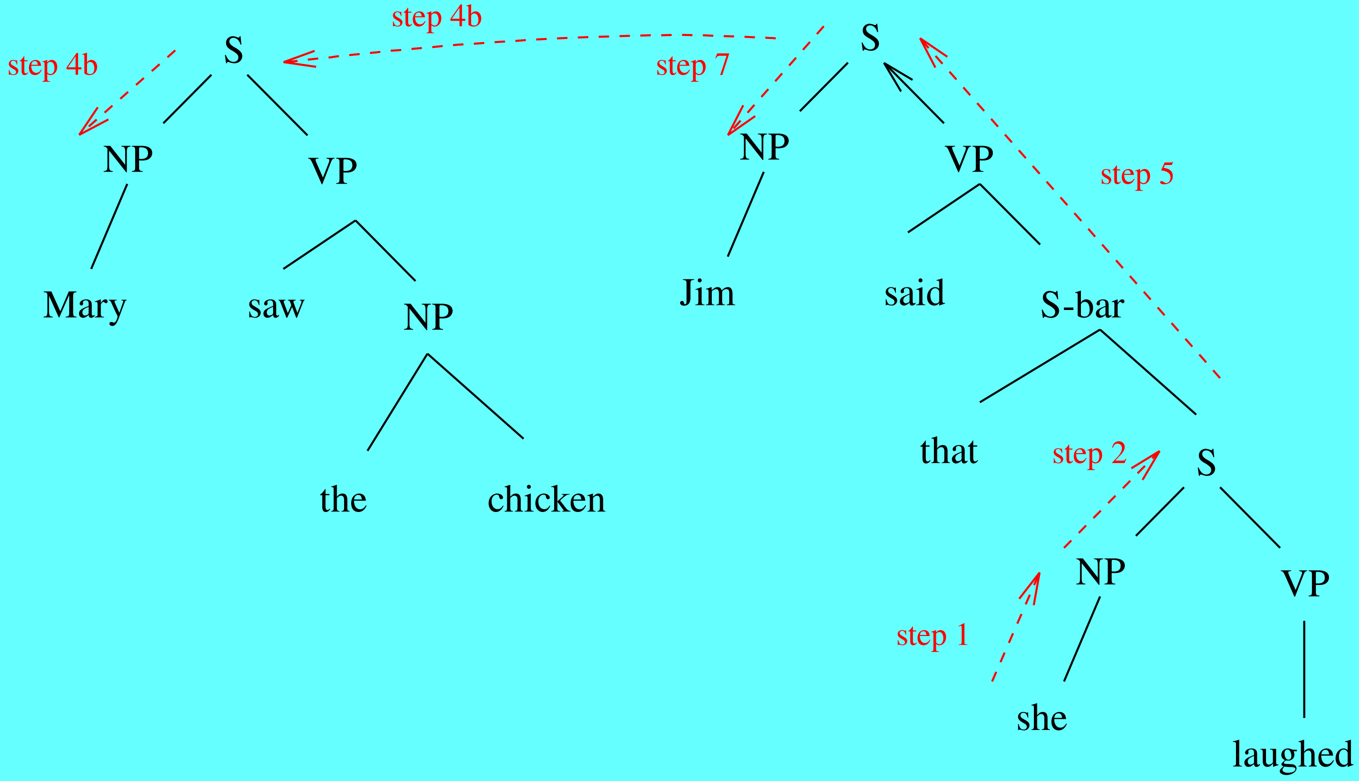
8. If X is an S node, traverse all branches of X following p, left to right, breadth first, but not going below any S or NP node found Propose each NP.

End

Hobbs Search Example

1. Mary saw the chicken.

2. Jim said that she laughed.



Testing the Hobbs Algorithm

- Try Hobbs on instances of PRP in wsj_0003 from WSJ Penn Treebank
- How many cases does the Hobbs algorithm get correct?
- How many incorrect?
- Are there some tweaks that would give better results?
- Or would these tweaks hurt other cases?

No-Parse Hobbs-like Search

- Only Consider Nouns/NGs satisfying constraints
- Continue searching until antecedent or loop exits

  1. Initialize sentence_counter to 0 and search current sentence from left to right, ending before pronoun.
  2. Repeat the following step until an antecedent is found or sentence_counter reaches the maximum (e.g., 3)
     - i. Search previous sentence from left to right
     - ii. Increment sentence_counter by 1

More Pronoun Coreference Systems

- Lappin and Leass (1994): Hobbs-Search-like procedure, Morphological filter, Binding Theory, Pleonastic Pronoun Handler, preferences based on grammatical role hierarchy (subject > object > ind-object), preference for same grammatical role, frequency of noun, recency, decision procedure for finding pronoun coreference
  - – 4% over Hobbs Search
- Other Systems Using Statistical Weights or Machine Learning Score a Little Bit Better, e.g., Dagan et. al. (1995) score another 3% better (89% vs. 82%).

Gender Agreement Based on Data

- Most Publications to date (e.g., WSJ from 1990s)
  - John, Adam, … → he, him, his
  - Mary, Jenny, … → she, her, her
  - sentient singular → I, me, my, you, your, …
  - Other singular → it, its
  - Plural → we, our, us, you, your, they, them, their
- Newer data
  - Some sentient singular → they, them, their
- Data Bias – a chicken and egg problem
  - Suppose we use Supervised ML to predict an arbitrary person's pronouns? Similar to issue about predicting gender of occupations?

Common Noun Coreference

- Definite Common Nouns
  - Poessio and Veira (2000) baseline:
    - A common noun phrase NP_1 with determiner “the” can be coreferential with a preceding NP_2 if:
      - NP_1 and NP_2 have the same head
      - And (ignoring determiners) NP_1 has a subset of the modifiers of NP_2
- There has been very little improvement on this baseline and very few systems that correctly identify the other cases with any large degree of accuracy
- Other factors:
  - Distance between NP_1 and NP_2
  - Other determiners, modifiers, possessives, etc.

Why is Common Noun Coreference Difficult?

- Only some common noun phrases are anaphoric
  - Definite vs. Generic
    - ***The officers*** vs. ***officers*** vs. ***an officer***
  - Limit to ***the*** phrases is a conservative decision
    - ***this***, ***that***, ***those***, possessives, ... improves recall, lowers precision
- When can a common noun corefer to another noun?
  - Limit to identical nouns is a conservative decision
  - Other choices improve recall, lower precision
    - My experience: a hand-crafted list of matches to NE classes
      - Ex: PERSON matches: man, human, person, individual, woman, .., officer, attorney, ...
      - Hurts approximately as much as it helps (paper wasn't accepted to conference)

Scoring Coreference in MUC-6

- Basics: $Precision = \frac{Correct}{\text{System Output}}$ $Recall = \frac{Correct}{\text{Answer Key}}$ $\text{F-Score} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$
- Problem: How do you measure number of correct?
- MUC-6:
  - – Coreference Chains = Partitions of NPs
  - – Recall and Precision are based on mismatches (edit distance) between partitions: numbers of links added and/or subtracted to change incorrect partitions to correct ones
    - Given 7 NPs in a system output chain: A_1, A_2,A_3,A_4,A_5,B_1,B_2 and 1 NP by itself B_3
    - The sets {A_1, A_2,A_3,A_4,A_5 } and {B_1,B_2,B_3} belong to separate chains in the Answer Key
    - System has 5 **Correct** links: $A_1 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5$ and $B_1 \rightarrow B_2$
    - System has 1 **Incorrect** link: $A_5 \rightarrow B_1$ (No penalty for including B_2 in chain with As)
    - System has 1 **Missed** link: $B_2 \rightarrow B_3$
    - Correct/System Output = Precision = 5/6 = 83%
    - Recall = 5/6 = 83%
    - F-Score = 83%

Scoring Coreference with B-Cubed

- B-Cubed (Bagga and Baldwin 1998) – evaluates one to many links
- Example:
    - 8 NPs in 2 system output chains: A_1, A_2,A_3,A_4,A_5,B_1,B_2 and B_3
    - 8 NPs in 2 answer key chains: A_1, A_2,A_3,A_4,A_5 and B_1,B_2 ,B_3
    - Total Links Assumed by System (red is incorrect):
        - $A_1 \rightarrow A_1$, $A_1 \rightarrow A_2$,$A_1 \rightarrow A_3$,$A_1 \rightarrow A_4$, $A_1 \rightarrow A_5$,$\mathbf{A_1 \rightarrow B_1}$,$\mathbf{A_1 \rightarrow B_2}$
        - $A_2 \rightarrow A_1$, $A_2 \rightarrow A_2$,$A_2 \rightarrow A_3$,$A_2 \rightarrow A_4$, $A_2 \rightarrow A_5$,$\mathbf{A_2 \rightarrow B_1}$,$\mathbf{A_2 \rightarrow B_2}$
        - …
        - $\mathbf{B_1 \rightarrow A_1}$, $\mathbf{B_1 \rightarrow A_2}$,$\mathbf{B_1 \rightarrow A_3}$,$\mathbf{B_1 \rightarrow A_4}$, $\mathbf{B_1 \rightarrow A_5}$,$B_1 \rightarrow B_1$,$B_1 \rightarrow B_2$
        - $\mathbf{B_2 \rightarrow A_1}$, $\mathbf{B_2 \rightarrow A_2}$,$\mathbf{B_2 \rightarrow A_3}$,$\mathbf{B_2 \rightarrow A_4}$, $\mathbf{B_2 \rightarrow A_5}$,$B_2 \rightarrow B_1$,$B_2 \rightarrow B_2$
        - $B_3 \rightarrow B_3$
- Precision calculated for each system chain and averaged
    - 5/7 of each of the 5 A chains are correct
    - 2/7 of each of the B_1 and B_2 chains are correct
    - 100% of the B3 chain is correct
- Precision = $((5\times(\frac{5}{7}))+(2\times(\frac{2}{7}))+1)*(\frac{1}{8})\approx .64$

B-cubed Recall & F-measure Same Example

- Answer Key – red missing from system output
  - $A_1 \rightarrow A_1$, $A_1 \rightarrow A_2$,$A_1 \rightarrow A_3$,$A_1 \rightarrow A_4$, $A_1 \rightarrow A_5$
  - $A_2 \rightarrow A_1$, $A_2 \rightarrow A_2$,$A_2 \rightarrow A_3$,$A_2 \rightarrow A_4$, $A_2 \rightarrow A_5$
  - …
  - $B_1 \rightarrow B_1$, $B_1 \rightarrow B_2$, $\mathbf{B_1 \rightarrow B_3}$
  - $B_2 \rightarrow B_1$, $B_2 \rightarrow B_2$, $\mathbf{B_2 \rightarrow B_3}$
  - $\mathbf{B_3 \rightarrow B_1}$, $\mathbf{B_3 \rightarrow B_2}$, $B_3 \rightarrow B_3$
- Recall calculated for each answer key chain (and averaged)
  - 100% recall for each of the 5 A chains
  - 2/3 recall for B_1 and B_2
  - 1/3 recall for B_3
  - Recall = $((5\times(\frac{5}{5}))+(2\times(\frac{2}{3}))+(1\times(\frac{1}{3})))*(\frac{1}{8})\approx.83$
- F-measure: $\frac{2}{((\frac{1}{.64})+(\frac{1}{.83}))}=.72$

Comparison of Coreference Scoring Metrics

- B-cubed vs MUC-6
  - B-cubed assumes entity model (pairing all coreferential items)
  - MUC-6 chain, linking each item to a previous one
    - a link is correct even if it misses an intermediate link
  - B-cube penalizes incorrect links more for precision (links to all members of the entity are incorrect)
  - B-cube gives credit for NPs that are not coreferential with other NPs (No links evaluated under MUC)
- ACE Scorer: complex weighted average designed to count names more than other types of NPs and Person names most of all
  - Participants in ACE used this measure at ACE government meetings, but published using B-cubed.

Cross Document Coreference

- So far: coreference with a single discourse
  - within a single document, names usually are unambiguous
  - Disambiguation strategies for ambiguous names
    - George Bush Sr. vs George W. Bush
      - Bush, George Bush, Mr. Bush, President Bush
    - New York City, New York State
      - New York
    - Disambiguation strategies: favor closer antecedent, favor antecedent with more references, etc.
- Reference Independent of Invidual Documents
  - Same person name, abbreviation, organization name
  - How do we know when they have different referents?

Baseline Strategy For CrossDoc

- Do single document coreference in each document
- Entity = set of “mentions” that are coreferential
- Select only those Entities which include Name mentions
- Choose longest name string as representative label
  - (don't use abbreviations as label)
- Compare representative labels across documents
  - Merge if labels match exactly
  - Merge if labels match modulo minor modifications
    - Delete middle initial or match middle names
    - Possibly delete titles
  - Similar to name coreference, but more conservative

Hard Cases: Ambiguity and Aliases

- Same name, different middle initial, e.g., *George Bush*
- Ambiguous abbreviations
  - *AMEX: American Stock Exchange* or *American Express*
- People famous in specific domains, or epochs
  - *Michael Jackson:* Musician, basketball player, football player, executive, …
  - Ann Hatheway: 16th Century vs 21st Century
- Places
  - *New York* (City vs State)
  - *Paris* in (France, Texas, Ontario, Denmark)
- Metonymy
  - *New York*: Rangers, Mets, Yankees, Giants, Jets, …
- Spelling Variation Across Documents (typos, transliteration, etc.)
  - *Osama bin Ladin, Usama ibn Ladin*
  - *(Moammar|Muammar) (Gadaffi|Gaddafi|Gathafi|Kadafi|Kaddafi|Khadafy|Qadhafi|Qathafi)*
- Name Changes over time
  - *Beijing, Beiping*
  - *Leningrad, Saint Petersburg*

Entity Linking Tasks

- TAC KBP Entity Linking Tasks
  - http://nlp.cs.rpi.edu/kbp/2014/
  - http://www.nist.gov/tac/2015/KBP/ColdStart/
  - http://nlp.cs.rpi.edu/kbp/2016/task.html
  - https://tac.nist.gov/2017/KBP/index.html
  - https://tac.nist.gov/2018/SM-KBP/index.html
  - http://nlp.cs.rpi.edu/kbp/2019/
  - https://tac.nist.gov/2020/KBP/RUFES/
  - Do within document coreference
  - For each people, organization, GPE entity E, either
    - Link E to an entry in the existing wikipedia-based database OR
    - Link E with a cross-document cluster of entities that your system created
    - Or create a new cross-document entity
- Database created semi-automatically from Wikipedia (and other sources)
  - Database entries correspond to Wikipedia pages
    - Ex: there are several *Paris* pages, one for each “sense” of *Paris*

Strategies Researchers Use

- Machine Learning with lots of features
- Baseline strategies as described
- Contextual features: similar contexts/diff docs
  - Ngrams, relations, vocabulary distribution of whole document
- Extract from Wikipedia Info Boxes
- Other features of documents
  - News articles from the same date and similar location
  - Genre or topic of article

Summary

- Reference Resolution Covers a Wide Area
  - Most Studied Area is Coreference
    - Proper Noun Coreference
      - Easiest to find correct answer
      - Most important for many applications
    - Pronoun Coreference
      - Most thoroughly studied in linguistics
  - Opportunities for research:
    - common noun coreference, other types of reference resolution, connection with relation extraction
  - **Current Emphasis in NLP community is on Entity Linking**
- Simple hard-to-beat baselines:
  - Hobbs
  - Poessio and Veira
- Evaluation is Non-Trivial

Short Homework 1:
Due April 11

- https://cs.nyu.edu/courses/spring23/CSCI-UA.0480-057/priv/b-cubed-hw.pdf

Readings and Corpora

- J&M: Chapter 21:3-8, 21:9
- Optional: Lappin and Leas (1994)
  - http://www.aclweb.org/anthology/J94-4002
- I also can make available some coref corpora
  - The one used for MUC-6
  - Penn Treebank WSJ corpus with pronoun coref
  - ACE – training corpus currently in NYUClasses
  - Look on LDC site for more corpora that I can make available

Possibly Interesting Research Question

- Some instances of bridging anaphora could be guided by NomBank
- Consider the relational noun **client** on the next slide (from wsj) – look for the ARG2 (beneficiary)

Client + ARG2

- ***She*** *was untrained and, in one botched job killed a* ***client***
- *The Internal Revenue Service has threatened criminal sanctions against lawyers who fail to report detailed information about* ***clients*** *who pay them more than $10,000 in cash.*
- *Mr. Sonnett said there also may be other circumstances under which individuals wouldn't want the government to know they had retained* ***criminal defense lawyers****.*

  Filling out detailed forms about these individuals would tip the IRS off and spark action against the ***clients****, he said.*