Discourse & Anaphora Resolution

B. Senthil Kumar
Asst. Professor, CSE
Natural Language Processing

Agenda

- Discourse
- Cohesion
- Anaphora resolution

Discourse

a. John went to his favorite music store to buy a piano.

b. He had frequented the store for many years.

c. He was excited that he could finally buy a piano.

d. He arrived just as the store was closing for the day.

a. John went to his favorite music store to buy a piano.

b. It was a store John had frequented for many years.

c. He was excited that he could finally buy a piano.

d. It was closing just as John arrived.

Discourse

- A structure that is needed for interpretation of text in a sentence is known as *discourse structure*.
- The collection of interrelated sentences is a discourse.
- Types of discourse:
  - Monologue – communication is unidirectional from speaker to hearer
  - Dialogue – two-way communication

Cohesion & Coherence

- The phenomena of discourse is **cohesion** and **coherence**.
- Cohesion is the grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning.
- There are two main types of cohesion:
  - grammatical cohesion, which is based on structural content
  - lexical cohesion, which is based on lexical content and background knowledge

Cohesion

- Five general categories of cohesive devices that create coherence in texts:
  - reference
  - ellipsis
  - substitution
  - lexical cohesion
  - conjunction

Reference

- There are two referential devices that can create cohesion:
- Anaphoric reference occurs when the writer refers back to someone or something that has been previously identified, to avoid repetition.

  Victoria Chen, CFO of Magabucks, saw her pay jump 20% to...
- Cataphoric reference is the opposite of anaphora: a reference forward as opposed to backward in the discourse.
- Something is introduced in the abstract before it is identified.

  "Here he comes, our award-winning host... it's John Doe!"

Ellipsis

- A form of cohesion where the words are omitted when the phrase must be repeated.
- (A) Where are you going?

  (B) To dance "I am going to dance"
- (A) I know that man. Do you?

  “know that man” - the verb phrase is left out.

Substitution

- A word is not omitted, as in ellipsis, but is substituted for another, more general word.
- Example:

1: "Which ice-cream would you like?"

"I would like the pink one"

2. "I dropped the ice-cream because it was dirty."

Lexical cohesion

- Lexical cohesion refers to the way related words are chosen to link elements of a text.
- There are two forms: repetition and collocation.
- Repetition uses the same word, or synonyms, antonyms, etc.,
- Example:

  "Which dress are you going to wear?"

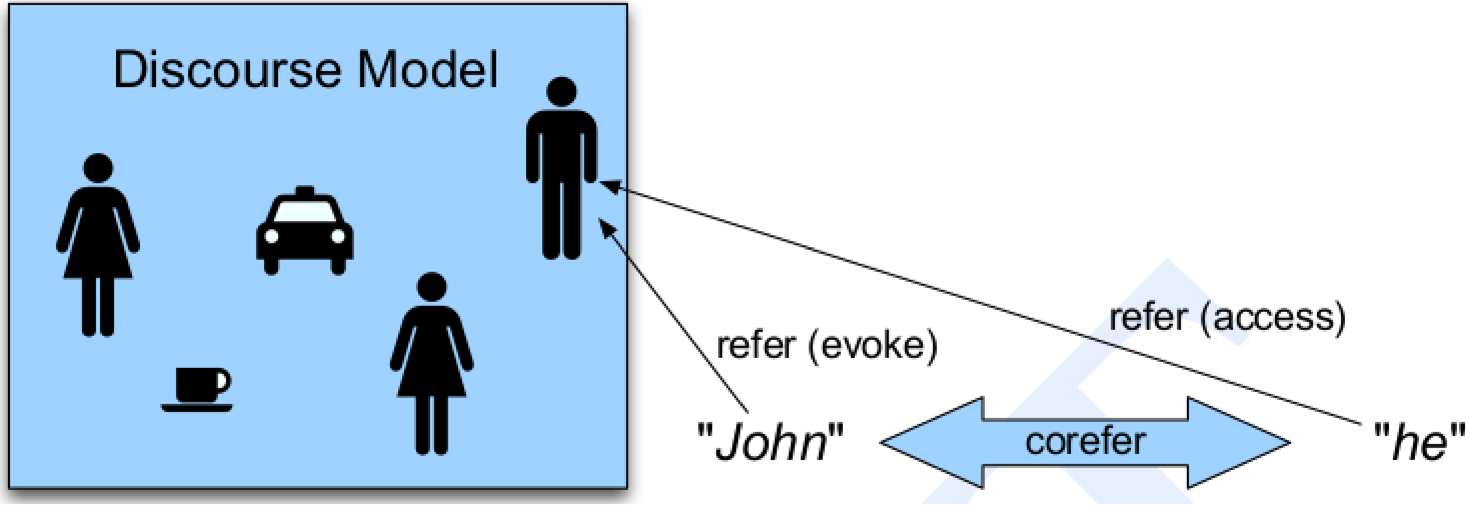
  "I will wear my green frock"
- Collocation uses related words that typically go together.
- Example: “strong tea”, “powerful computer”

Anaphora resolution

Anaphora resolution

- 'Anaphora is a cohesion which points back to some previous item.'

  - Halliday, Hassan (1976)
- The '*pointing back*' word or phrase is called an **anaphor** and the entity to which it refers or for which it stands is its **antecedent**.
- The process of determining the antecedent of an anaphor is called **anaphora resolution**.



Types of references

- Types of references
  - indefinite noun phrase
  - definite noun phrase
  - Pronominal reference
  - Quantifier reference (one-anaphora)
  - Inferrables
  - Generic

Indefinite noun phrases

- Introduces entities that are new to the hearer
- Marked with determiner: *a, an, this*; quantifier: *some*
- Example:

  He had gone round one day to bring her *some walnuts*

  I saw *this beautiful Ford* today

  I met *this girl* earlier in a conference

Definite noun phrases

- Refer an entity that is identifiable to the hearer
- The entity has been mentioned previously in the text (discourse)
- Example:

  I read about it in *The New York Times*.

  I bought a printer today. *The printer* didn’t work properly.

Pronominal

Personal pronouns (he, him, she, her, it, they, them)

  - The most difficult for Dalí was to tell her, that *he* loved her.

- Possessive pronouns (his, her, hers, its, their, theirs)
  - But the best things about Dalí are *his* roots and *his* antennae.
- Reflexive pronouns (himself, herself, itself, themselves)
  - Dalí once again locked *himself* in his studio . . .

Pronominal

- Demonstrative pronouns (this, that, these, those)
  - Dalí, used photographic precision to transcribe the images of his dreams. *This* would become one of the constraints. . . .
  - I bought a printer today. I had bought one earlier in 2004. This one cost me Rs.6,000 whereas that one cost me Rs.12,000.
- Relative pronouns (who, whom, which, whose)
  - Dalí, a Catalan *who* was addicted to fame and gold, painted a lot and talked a lot.

Pleonastic – it

- The pronoun *it* can often be non-anaphoric pronouns.
- Non-anaphoric uses of it are also referred to as ***pleonastic***.
- Examples:

  It is raining.

  It is tea time.

  It is a long way from here to Chennai.

  It appears that . . .

Quantifier reference

- Uses the ordinal one, first, etc.,
- Example:

  I visited a computer shop to buy a printer. I have to select one.

Inferrables

- Referring expression does not refer to an enity explicitly.
- But it is inferentially related to an evoked entity.
- Example:

  I bought a printer today. On opening the package, I found the paper tray broken.

  I almost bought an Acura Integra today, but the engine seemed noisy.

Generic

- More complicated reference
- Refers to a whole class instead of an individual or specific entity.
- Example:

  I saw 2 laser printers in a shop. They were the fastest printers available.

  I saw no less than 6 Acura Integra today. They are the coolest cars.

Constraints on referents

- Number agreement
- Person agreement
- Gender agreement
- Binding theory constraints
- Selectional restrictions
- Grammatical role
- Recency
- Repeated mention
- Parallelism
- Verb semantics

Number agreement

- John has a Ford Falcon. It is red.
- ??John has a Ford Falcon. They are red.

Singular	Plural	Unspecified
She, her, he, him, his, it	We, us, they, them	you

Person agreement

- John has aFord Falcon. He loves it
- ??John has a Ford Falcon. She loves it

	First	Second	Third
Nominative	I,we	you	he,she,they
Accusative	me,us	you	him,her,them
Genitive	my,our	your	his, her, their

Gender agreement

- John has a Ford Falcon. He is attractive (He=John)
- John has a Ford Falcon. It is attractive (It=Ford)

male	He, him, his
female	She, her
nonpersonal	it

Binding constraint

- Reflexive pronouns corefers with the subject of most immediate clause – binding theory

John bought himself a new Ford [himself = John]

John bought him a new Ford [him ≠ John]

John said that Bill bought him a new Ford [him ≠ Bill]

John said that Bill bought himself a new Ford [himself = Bill]

Selectional restrictions

- Restrictions that a verb places on its arguments removes the referent(s)

John parked his car in the garage after driving it around for hours.

Jim bought a coffee from the store. He drank it quickly.

Recency

- Entities introduced in recent utterances tend to be more salient than those introduced from utterances further back.

John had a pop-tart. Bill had a jelly donut. Mary wanted it.

The doctor found an old map. Jim found an even older map hidden on the shelf. It described an island.

Grammatical role

- Entities mentioned in subject position as more salient than those in object position

Bill went to the bar with Jim. He called for a glass of wine.

[He = Bill]

Jim went to the bar with Bill. He called for a glass of wine.

[He = Jim]

Repeated mention

- Entities that have been **focussed** on in the prior discourse are more likely to be referred in subsequent discourse.

John went to the store to buy coffee.

He loves coffee. He drinks 5 cups a day.

At the store, Bill sold him a cup. He was delighted.

[He=John]

Parallelism

- Even though the grammatical hierarchy ranks *Johnson* as more salient and prefered referent of *him,* the syntactic parallelism prefers [him = Jones]

  Johnson went with Jones to the *Old Parrot*. Billy Bones went with him to the Old Anchor Inn. [him = Jones]

Verb Semantics

- Certain verbs appear to place a *semantically-oriented emphasis* on one of their argument positions.

- John *telephoned* Bill. He lost the laptop. [He=John]

  Implicit cause of *telephoning* is its subject

- John *criticized* Bill. He lost the laptop. [He=Bill]

  Implicit cause of *criticizing* is its object

Hobbs Algorithm

Centering Theory

Log-linear Model

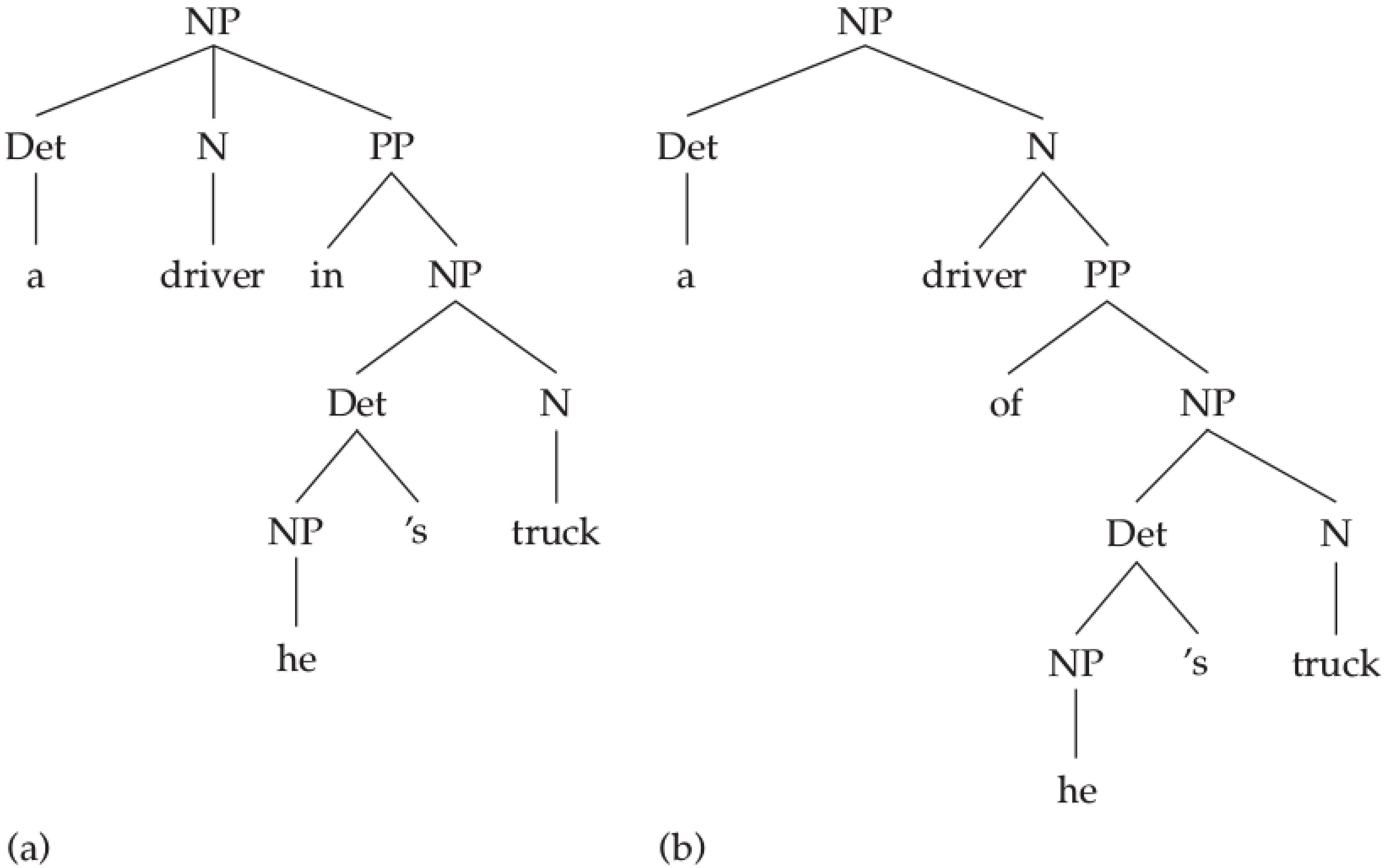
Hobbs Algorithm

- Depends on a syntactic parser plus a morphological gender and number checker.
- Input: pronoun to be resolved + syntactic parse of the sentences
- Start with the target pronoun and walk up the parse tree to root S.
- For each NP or S node, perform breadth-first left-to-right search of node's children to the left of target.
- Assumption: the parse tree represent the correct grammatical structure of the sentence with all adjunct phrases properly attached.

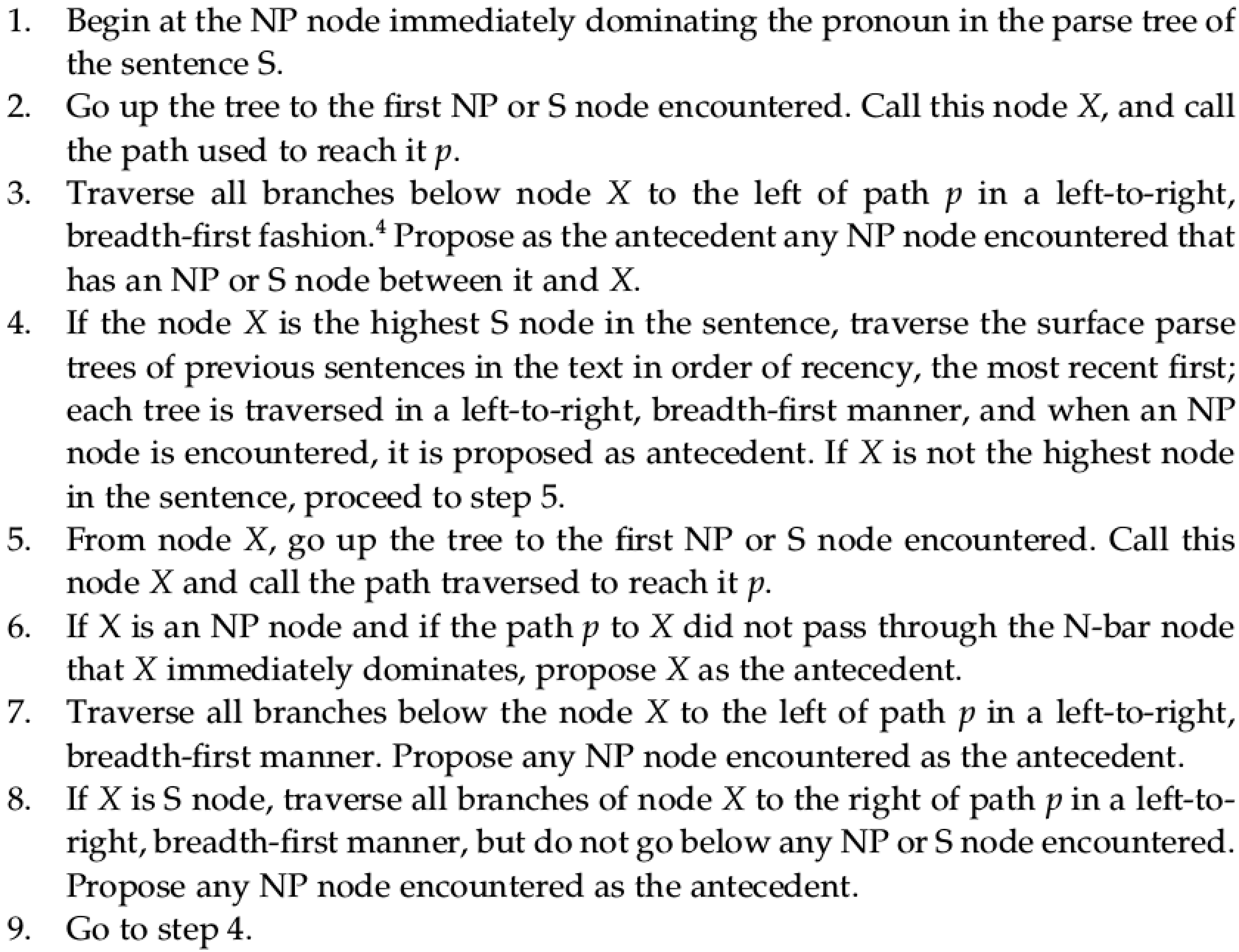
Hobbs Algorithm

a. Mr. Smith saw a *driver* in *his* truck.

b. Mr. Smith saw a driver of *his* truck.

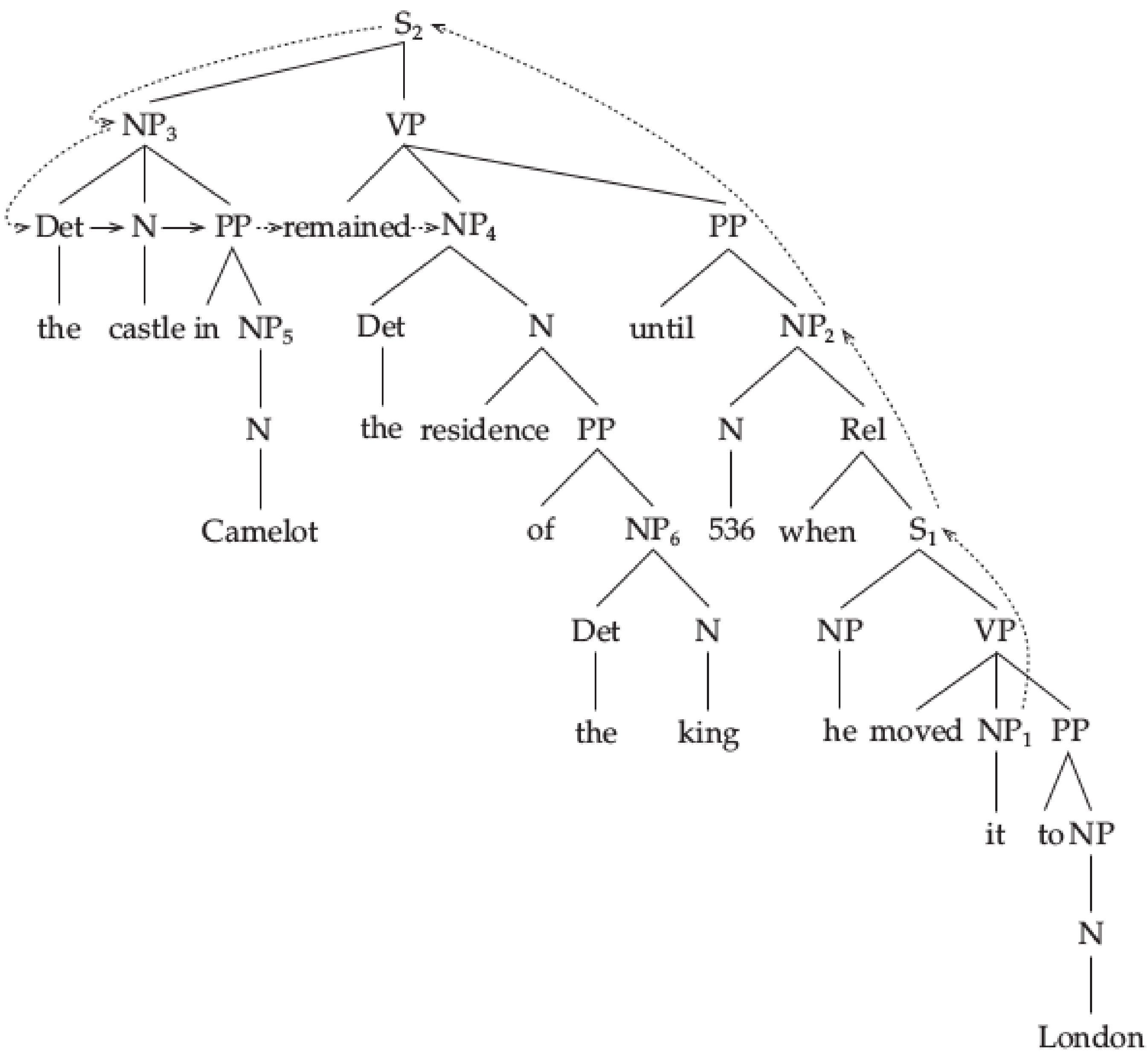


Hobbs Algorithm



1. Begin at the NP node immediately dominating the pronoun in the parse tree of the sentence S.
2. Go up the tree to the first NP or S node encountered. Call this node *X*, and call the path used to reach it *p*.
3. Traverse all branches below node *X* to the left of path *p* in a left-to-right, breadth-first fashion.[4] Propose as the antecedent any NP node encountered that has an NP or S node between it and *X*.
4. If the node *X* is the highest S node in the sentence, traverse the surface parse trees of previous sentences in the text in order of recency, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP node is encountered, it is proposed as antecedent. If *X* is not the highest node in the sentence, proceed to step 5.
5. From node *X*, go up the tree to the first NP or S node encountered. Call this node *X* and call the path traversed to reach it *p*.
6. If X is an NP node and if the path *p* to *X* did not pass through the N-bar node that *X* immediately dominates, propose *X* as the antecedent.
7. Traverse all branches below the node *X* to the left of path *p* in a left-to-right, breadth-first manner. Propose any NP node encountered as the antecedent.
8. If *X* is S node, traverse all branches of node *X* to the right of path *p* in a left-to-right, breadth-first manner, but do not go below any NP or S node encountered. Propose any NP node encountered as the antecedent.
9. Go to step 4.

Hobbs Algorithm



Centering Theory

- There is a single entity being “centered” on at any given point in the discourse.
- This entity need to be distinguished from all other entities that have been evoked.

Centering Algorithm

- Discourse A

(3.1) John works at Barclays Bank.

(3.2) He works with Lisa.

(3.3) John is going to marry Lisa.

(3.4) He is looking forward to the wedding.

- Discourse B

(3.1) John works at Barclays Bank.

(3.2) He works with Lisa.

(3.3) John is going to marry Lisa.

(3.5) She is looking forward to the wedding.

Centering Algorithm

- Centering predicts that Discourse B is less coherent than Discourse A.
- The shift in center and the use of a pronominal form to realise the new center (3.5) contribute to making B less coherent than A.
- A discourse segment D consists of a sequence of utterances $U_1, U_2, \ldots U_n$.
- Each utterance U is assigned a set of potential next centers known as **forward-looking centers** $C_f(U)$ which correspond to the discourse entities evoked by the utterance.
- Each utterance (other than the first) is assigned a single center as the **backward-looking center** $C_b(U)$.

Centering Algorithm

- The C_b entity connects the current utterance to the previous discourse: it focuses on *an entity that has already been introduced*.
- A central claim of centering is that each utterance has exactly one backward-looking center.
- $C_f(U)$ is partially ordered according to their discourse salience.
- The highest-ranked element in $C_f(U)$ is called the **preferred center** $C_p(U)$.
- $C_p(U_N)$ is the most likely backward-looking center of the following utterance $C_b(U_{N+1})$

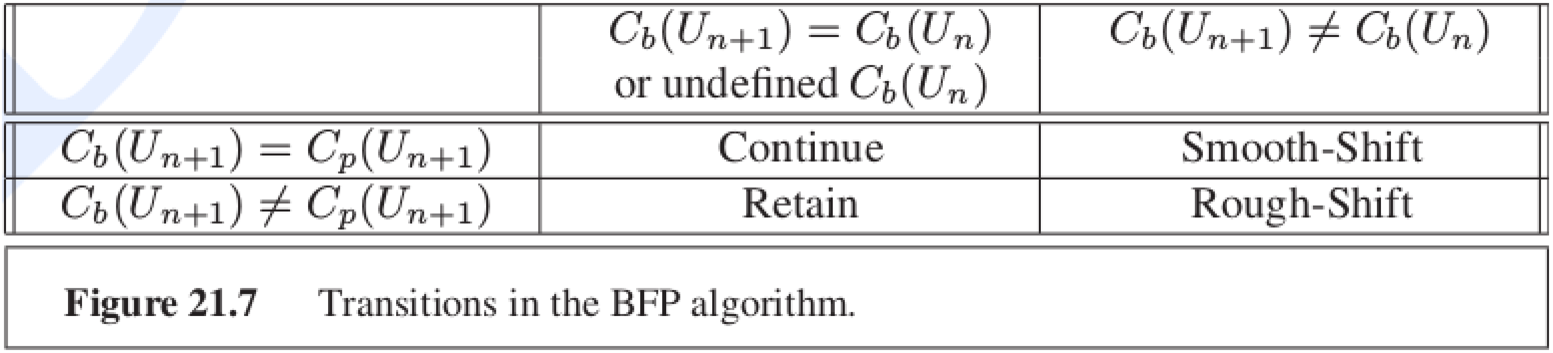
Centering Algorithm

- Basic Steps
- 1. Generate possible C_b -C_f combinations for each possible set of reference assignments .
- 2. Filter by constraints, e.g., syntactic coreference constraints, selectional restrictions, centering rules and constraints.
- 3. Rank by transition orderings.

Centering Algorithm

- Rule 1: If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.
- Rule 2: Transition states are ordered:

  Continue > Retain > Smooth-Shift > Rough-Shift.



	$C_b(U_{n+1}) = C_b(U_n)$ or undefined $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Figure 21.7 Transitions in the BFP algorithm.

Centering Algorithm

U_1: John saw a beautiful 1961 Ford Falcon at the used car dealership.

U_2: He showed it to Bob.

U_3: He bought it.

- Using the grammatical role hierarchy:

  C_f (U_1): {John, Ford, dealership}

  C_p (U_1): John

  C_b (U_1): undefined

Centering Algorithm

- Sentence U_2 contains two pronouns:

  he [John], and it [Ford or the dealership].

- Assume *it* refers to the *Ford*, the assignments would be:

  C_f (U_2): {John, Ford, Bob}

  C_p (U_2): John

  C_b (U_2): John

- Result: Continue ((C_p (U_2)=C_b (U_2); C_b (U_1) undefined)

Centering Algorithm

- Assume *it* refers to the *dealership*:

  C_f (U_2): {John, dealership, Bob}

  C_p (U_2): John

  C_b (U_2): John

- Result: Continue ($C_p(U_2) = C_b(U_2)$; $C_b(U_1)$ undefined)

- Ties are broken in terms of ordering of C_f list, take *it* to refer to the *Ford* instead of *dealership*.

Centering Algorithm

- Assume *he* refers to *John*, then *John* is C_b (U_3) and the assignments would be:

  C_f (U_3): {John, Ford}

  C_p (U_3): John

  C_b (U_3): John

- Result: Continue (C_p (U_3) = C_b(U_3) = C_b(U_2))

Centering Algorithm

- Assume *he* refers to *Bob*, then *Bob* is C_b (U_3) and the assignments would be:

  C_f (U_3): {Bob, Ford}

  C_p (U_3): Bob

  C_b (U_3): Bob
- Result: Smooth-Shift ($C_p(U_3) = C_b(U_3); C_b(U_3) \neq C_b(U_2)$)
- Continue is preferred to a Smooth-Shift, John is taken as referent.

Mitkov's knowledge-poor algorithm

Mitkov's Algorithm

- Mitkov’s approach avoids complex syntactic, semantic and discourse analysis, relying on *antecedent indicators*.
- It works from the output of a text processed by a part-of-speech tagger and an NP extractor.
- Locate noun phrases which precede the anaphor within a distance of two sentences.
- Check for gender and number agreement with the anaphor.
- Apply the indicators to the remaining candidates by assigning a positive or negative score (2, 1, 0 or −1).
- The noun phrase with the highest composite score is proposed as antecedent.

Antecedent indicators

- The boosting indicators are:
- *First noun phrases*:

  A score of +1 is assigned to the first NP in a sentence.
- *Indicating verbs*: A score of +1 is assigned to those NPs immediately following a set of verbs.

  (*analyse, assess, check, consider, cover, define, describe,* etc.,)

  Above verbs usually carry more salience.
- *Lexical reiteration*: A score of +2 is assigned to those NPs repeated twice or more, and a score of +1 is assigned to those NPs repeated once in the paragraph

Antecedent indicators

- Collocation match: A score of +2 is assigned to those NPs that have an identical collocation pattern to the pronoun.
- Immediate reference: A score of +2 is assigned to those NPs appearing in constructions of the form:

  '. . . (You) V_1 NP . . . con (you) V_2 it (con (you) V_3 it)',

  Example: you can stand the printer up or lay it flat.
- Sequential instructions get score of +2.

  Ex: To turn on *the video recorder*, press ...... To programme *it*,
- Term preference: A score of +1 is applied to those NPs identified as terms in the genre of the text.

Antecedent indicators

- The impeding indicators are:
- Indefiniteness: Indefinite NPs are assigned a score of −1.
- Prepositional noun phrases: NPs appearing in prepositional phrases are assigned a score of −1.
- Referential distance: (may impede/boost)

  distance between the NP from the pronoun

  NPs in the previous clause to (but in same sentence as) the pronoun: +2,

  NPs in the previous sentence to the pronoun: +1

  NPs in the sentence prior to that: 0

  NPs with more distant pronouns: −1

Antecedent indicators

- Raise the original cover. Place the original face down on the original glass so that *it* is centrally aligned.

Original cover

1(first NP)+1(term preference)+1(referential distance) = 3

Original face

Preferred

1(first NP)+1(term prefernce)+2(referential distance) = 4

Original glass

1(term preference)-1(PP)+2(referential distance) = 2

References

- *Speech and Language Processing*, Daniel Jurafsky, Martin, Pearson, 2006.
- *Anaphora Resolution*, Ruslan Mitkov, Pearson Education, 2002.