Computational Discourse

Le Anh Cuong

Reading

- Chapter 21 [1]

Outline

- Discourse segmentation
- Text cohrence
- Reference resolution

Discourse

- Consists of collocated, structured, coherent groups of sentences
  - What makes something a discourse as opposed to a set of unrelated sentences?
  - How can text be structured (related)?

- * Monologue: a speaker (writer) and hearer (reader) with communication flow in one direction only
- Dialogue: each participant takes turn being the speaker and the hearer (so 2-way participation)
  - Human-human dialogue
  - Human-computer dialogue (conversational agent)

Discourse Phenomina: Coreference Resolution

- The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard’s response.

Discourse Phenomina: Coreference Resolution

- The Tin Woodman went to the Emerald City to see the Wizard of Oz and ask for a heart. After he asked for it, the Woodman waited for the Wizard's response.

- What do we need to resolve?
- Why is it important?
  - Information extraction, summarization, conversational agents

Coherence Relations: Coreference

- First Union Corp is continuing to wrestle with severe problems. According to industry insiders at Pain Webber, their president, John R. Georgius, is planning to announce his retirement tomorrow.

Coherence Relations: Coreference

- First Union Corp is continuing to wrestle with severe problems. According to industry insiders at Pain Webber, their president, John R. Georgius, is planning to announce his retirement tomorrow.

- First Union Corp is continuing to wrestle with severe problems. According to industry insiders at Pain Webber, their president, John R. Georgius, believes Pain Webber can be instrumental in solving most of First Union’s problems.

Coherence Relations (Discourse Structure)

- First Union Corp is continuing to wrestle with severe problems. According to industry insiders at Pain Webber, their president, John R. Georgius, is planning to announce his retirement tomorrow.

- Reasonable summary:

  First Union President John R. Georgius is planning to announce his retirement tomorrow.

What you need to know: coherence relations between text segment – the first sentence is providing background for the more important 2nd sentence.

Coherence (relation based)

- John hid Bill’s car keys. He was drunk.
- ?? John hid Bill’s car keys. He likes spinach.
- Coherence Relations – relations such as EXPLANATION or CAUSE that exists between two coherent sentences. Connections between utterances.

More Coherence (entity based)

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.

e) John went to his favorite music store to buy a piano.
f) It was a store John had frequented for many years.
g) He was excited that he could finally buy a piano.
h) It was closing just as John arrived.

More Coherence (entity based)

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.

e) John went to his favorite music store to buy a piano.
f) It was a store John had frequented for many years.
g) He was excited that he could finally buy a piano.
h) It was closing just as John arrived.

Discourse Segmentation

- We want to separate a document into a linear sequence of subtopics

- Unsupervised Discourse Segmentation: Marti Hearst's TextTiling (done in early 90's)

Consider a 23 paragraph article broken into segments (subtopics):

- 1-2 Intro to Magellan space probe
- 3-4 Intro to Venus
- 5-7 Lack of craters
- 8-11 Evidence of volcanic action
- 12-15 River Styx
- 16-18 Crustal spreading
- 19-21 Recent volcanism
- 22-23 Future of Magellan

Wants to do this in an unsupervised fashion – how?

Text Cohesion

Text Cohesion

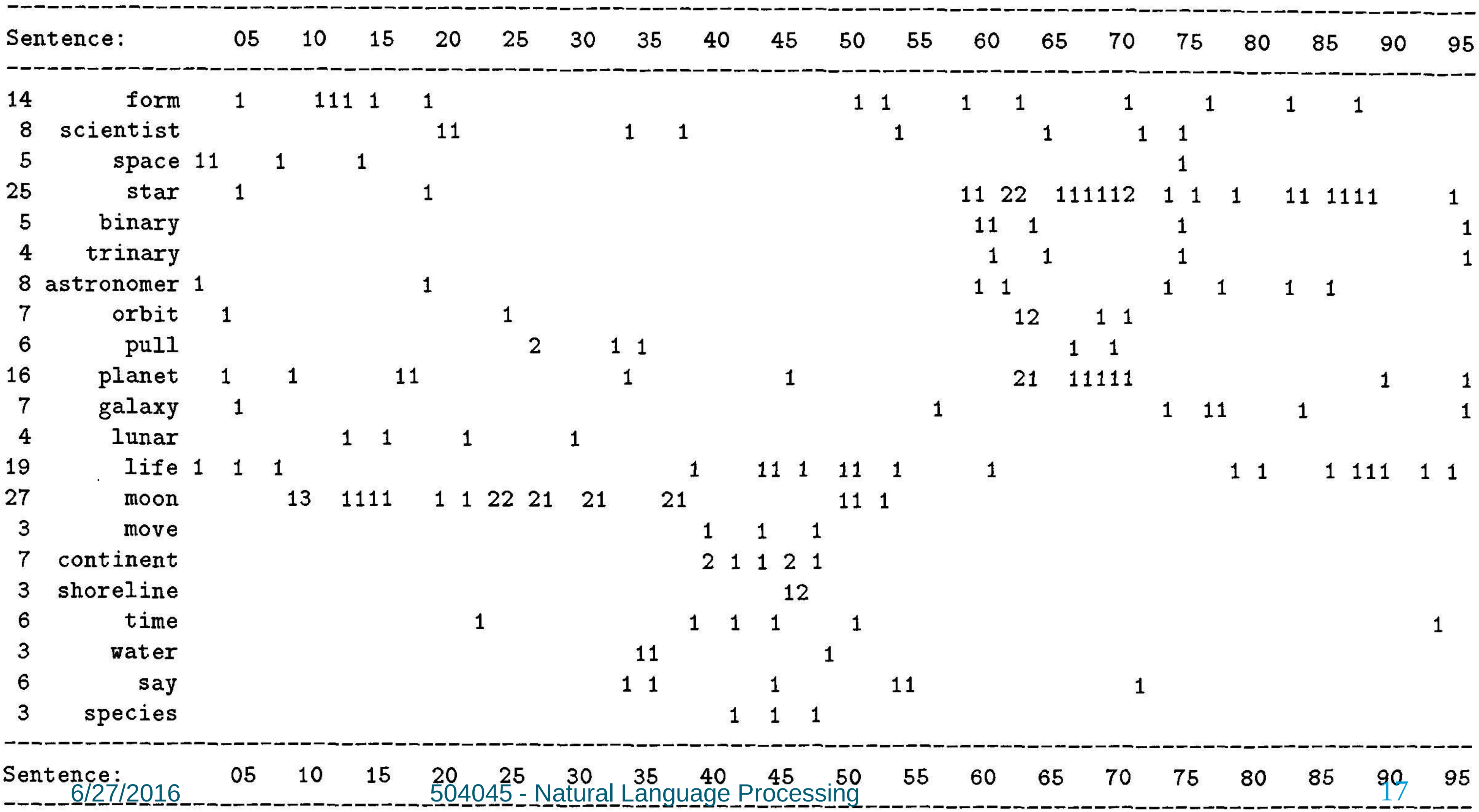
Halliday and Hasan (1976): “The use of certain linguistic devices to link or tie together textual units”

- Lexical cohesion: Indicated by relations between words in the two units (identical word, synonym, hypernym)
  - Before winter I built a chimney, and shingled the sides of my house..
  - I thus have a tight shingled and plastered house.
- Non-lexical cohesion like anaphora
  - Peel, core and slice the pears and the apples.
  - Add the fruit to the skillet.

Intuition to a Cohesion-based approach to segmentation

- Sentences or paragraphs in a subtopic are cohesive with each other, but not with paragraphs in a neighboring subtopic.

From Hearst 1997



```
--------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
--------------------------------------------------------------------------------------------------------------
14        form    1     111 1   1                               1 1     1   1       1     1     1    1
 8   scientist                   11            1   1               1          1      1  1
 5       space 11    1     1                                                            1
25        star    1             1                                       11 22  111112  1 1  1   11 1111     1
 5      binary                                                            11  1          1                    1
 4     trinary                                                             1   1         1                    1
 8 astronomer 1                1                                        1 1           1   1    1  1
 7       orbit   1                    1                                     12    1 1
 6        pull                          2     1 1                               1  1
16      planet   1    1       11               1           1                21  11111                  1     1
 7      galaxy    1                                                   1                1  11     1           1
 4       lunar            1  1     1       1
19        life 1  1  1                              1    11 1  11  1      1                 1 1    1 111  1 1
27        moon        13  1111   1 1 22 21  21    21           11 1
 3        move                                       1   1   1
 7   continent                                       2 1 1 2 1
 3   shoreline                                             12
 6        time                      1               1  1  1     1                                          1
 3       water                                  11            1
 6         say                                 1 1        1        11                1
 3     species                                         1  1  1
--------------------------------------------------------------------------------------------------------------
Sentence:        05   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95
--------------------------------------------------------------------------------------------------------------
```

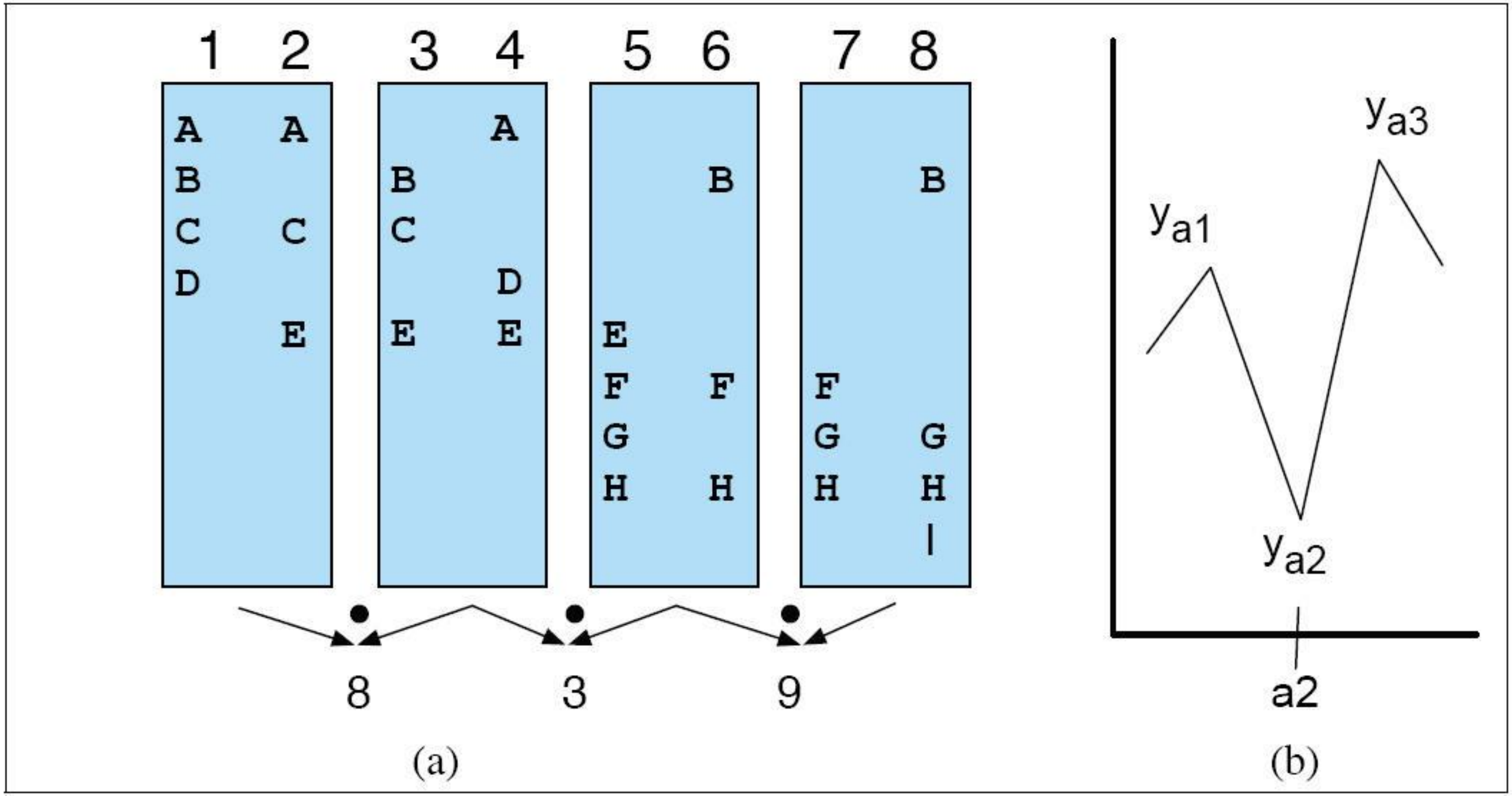
TextTiling (Hearst, 1997)

1. **Tokenization** – convert words to lower case, remove stop words, stem words, group into pseudo-sentences
2. **Lexical Score Determination** – check scores between each pair of sentences = average similarity of the words in the pseudo-sentences before the gap to the pseudo-sentences after the gap

$$\text{sim}_{\text{cosine}}(\vec{b}, \vec{a}) = \frac{\vec{b} \cdot \vec{a}}{|\vec{b}||\vec{a}|} = \frac{\sum_{i=1}^{N} b_i \times a_i}{\sqrt{\sum_{i=1}^{N} b_i^2}\sqrt{\sum_{i=1}^{N} a_i^2}}$$

3. **Boundary Identification** – assign a cut-off distance to identify a new segment.

Figure 21.1


1
2
3
4
5
6
7
8
A A
B
C C
D
E
A
B
C
D
E E
B
E
F F
G
H H
B
F
G G
H H
I
8
3
9
(a)
y_{a1}
y_{a3}
y_{a2}
a2
(b)

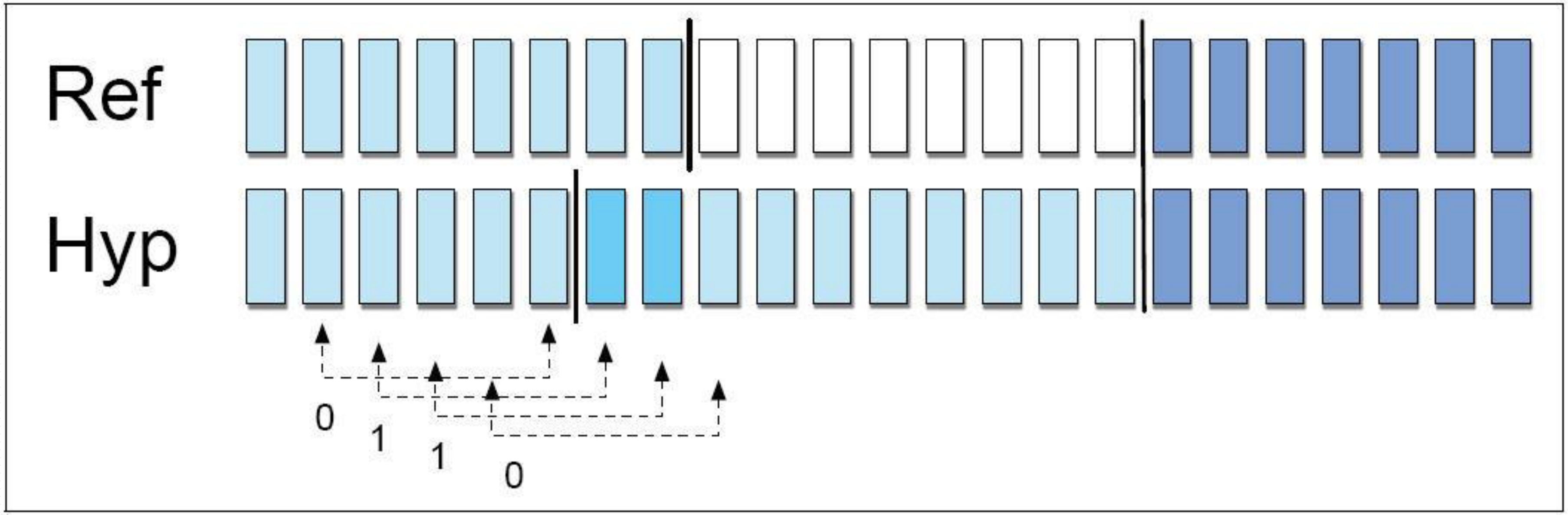
Supervised Discourse Segmentation

- To be used when it is relatively easy to acquire boundary-labeled training data
  - News stories from TV broadcasts
  - Paragraph segmentation
- Lots of different classifiers have been used
  - Feature set; generally a superset of those used for unsupervised segmentation
  - + discourse markers and cue words
- Discourse Markers generally domain specific

Supervised Discourse Segmentation

- Supervised machine learning
  - Label segment boundaries in training and test set
  - Extract features in training
  - Learn a classifier
  - In testing, apply features to predict boundaries
- Evaluation – usual measures of precision, recall, and F-measure don’t work – need to be sensitive to near-misses.

Figure 21.2



What makes a text coherent?

- Appropriate use of coherence relations between subparts of the discourse --rhetorical structure
- Appropriate sequencing of subparts of the discourse -- discourse/topic structure
- Appropriate use of referring expressions

Coherence Relations

- Possible connections between utterances in a discourse. Such as in Hobbs 1997.

- Result: Infer that the state or event asserted by S_0 causes or could cause the state or event asserted in S_1.
  - The Tin Woodman was caught in the rain. His joints rusted.

- Explanation: Infer that the state or event asserted by S_1 causes or could cause the state or event asserted by S_0.
  - John hid Bill's car keys. He was drunk.

Coherence Relations

- Parallel: Infer p(a_1, a_2,…) from the assertion of S_0 and p(b_1, b_2,…) from the assertion of S_1, where a_i and b_i are similar, for all i.
  - The scarecrow wanted some brains. The Tin Woodman wanted a heart.
- Elaboration: Infer the same proposition P from the assertions of S_0 and S_1.
  - Dorothy was from Kansas. She lived in the midst of the great Kansas prairies.

Coherence Relations

- Occasion: A change of state can be inferred from the assertion of S_0, whose final state can be inferred from S_1, or a change of state can be inferred from the assertion of S_1, whose initial state can be inferred from S_0.
  - Dorothy picked up the oil-can. She oiled the Tin Woodman’s joints.

Hierarchical structures

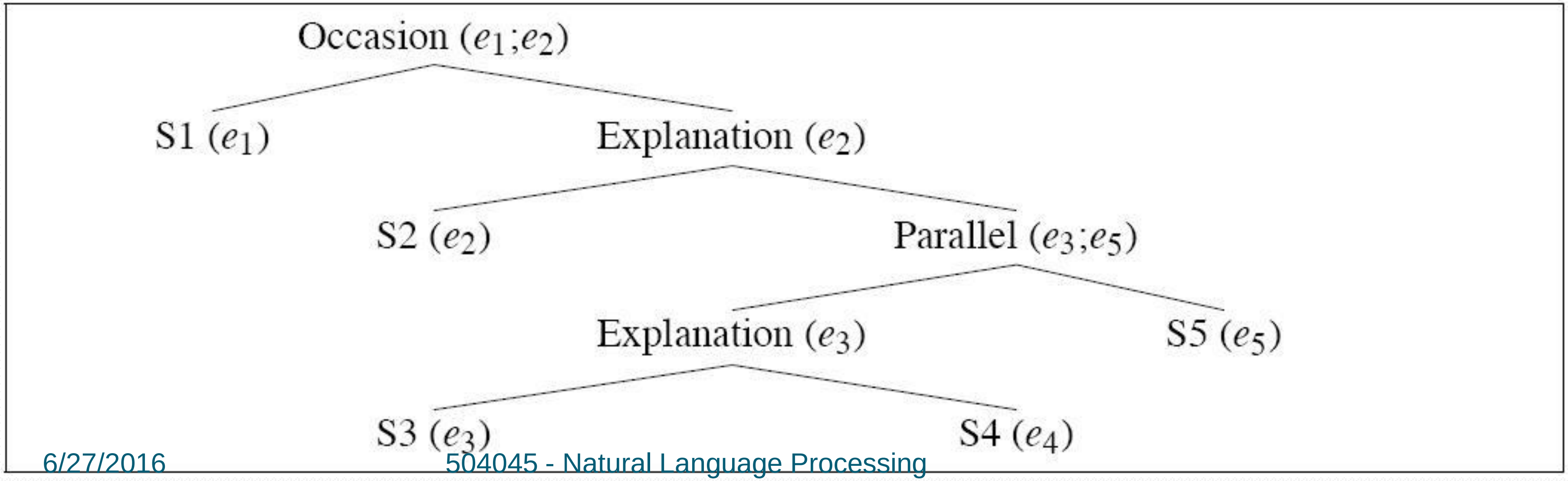
(S1) John went to the bank to deposit his paycheck.

(S2) He then took a train to Bill’s car dealership.

(S3) He needed to buy a car.

(S4) The company he works for now isn’t near any public transportation.

(S5) He also wanted to talk to Bill about their softball league.



Rhetorical Structure Theory

- See old slides

- See old slides on referring and discourse models

5 Types of Referring Expressions

- Indefinite Noun Phrases: Introduces into discourse context entities that are new to the hearer.
  - A man, some walnuts, this new computer

- Definite Noun Phrases: refers to an entity that is identifiable to the hearer (e.g., been mentioned previously or well known, in set of beliefs about the world).
  - “a big dog…. the dog…”, the sun

5 Types of Referring Expressions

- Pronouns: another form of definite reference, generally stronger constraints on use than standard definite reference.
  - He, she, him, it, they…
- Demonstratives: demonstrative pronouns (this, that) can be alone or as determiners.
- Names: Common method of referring including people, organizations, and locations.

Features for Filtering Potential Referents

- Number Agreement: pronoun and referent must agree in number (single, plural)
- Person Agreement: 1st, 2nd, 3rd
- Gender Agreement: male, female, nonpersonal (it)
- Binding Theory Constraints: constraints by syntactic relationships between a referential expression and a possible antecedent noun phrase in the same sentence.
  - John bought himself a new Ford. [himself = John]
  - John bought him a new Ford. [him ≠ John]
  - He said that he bought John a new Ford. [He≠John; he≠John]

Preferences in Pronoun Interpretation

- Recency – entities from most recent utterances more likely
  - The doctor found an old map in the captian’s chest. Jim found an even older map hidden on the shelf. **It** described an island.
- Grammatical Role – salience hierarchy of entities that is ordered by the grammatical position of the expressions that denote them. [subject, object,…]
  - Billy Bones went to the bar with Jim Hawkins. He called for a glass of rum. [He = Billy Bones]
  - Jim Hawkins went to the bar with Billy Bones. He called for a glass of rum. [He =Jim Hawkins]

Preferences (cont.)

- Repeated Mention – keep talking about the same thing.
- Parallelism – subject to subject; object to object.
  - Long John Silber went with Jim to the Old Parrot. Billy Bones went with him to the Old Anchor Inn. [him = Jim]
- Verb Semantics – some verbs seem to place emphasis on one of their argument positions.
  - John telephoned Bill. He lost the laptop.
  - John criticized Bill. He lost the laptop.
- Selectional Restrictions – other semantic knowledge playing a role.
  - John parked his car in the garage after driving it around for hours. [it = garage??]

Algorithms for co-reference resolution

- Hobbs Algorithm
- Centering
- Log-Linear Model (Learning Model)

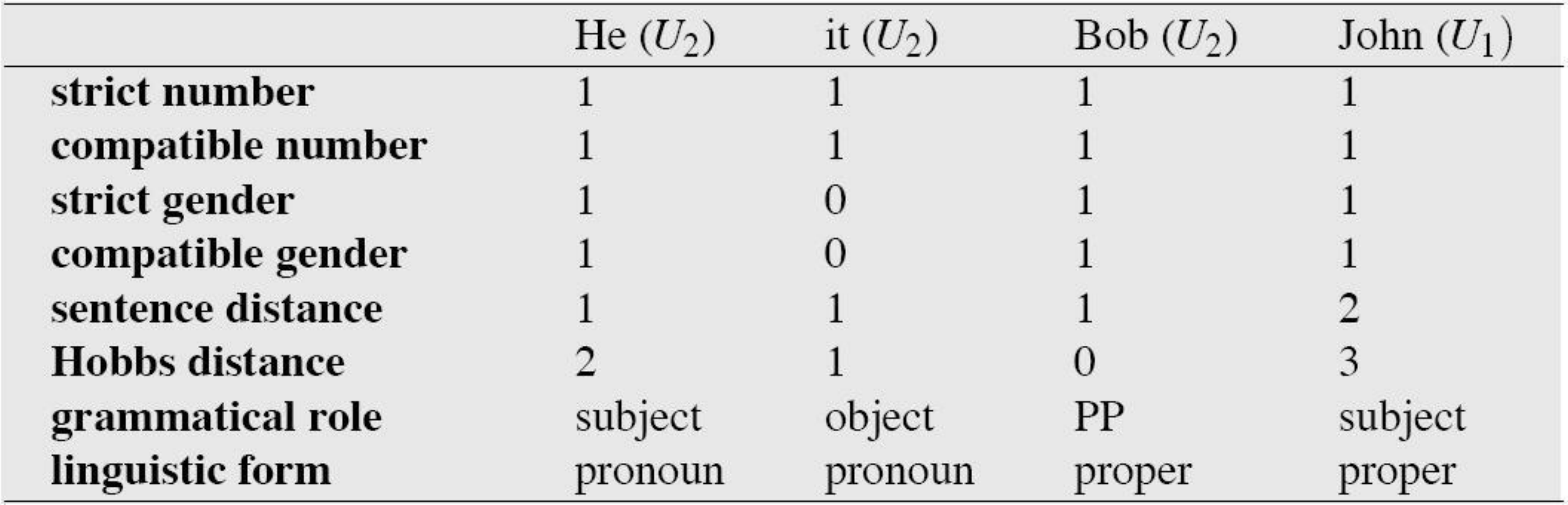
Log-Linear Model for Pronominal Anaphora Resolution

- Simple supervised machine learning approach
- Train classifier on a hand-labeled corpus in which are marked
  - Positive examples – antecedents marked with each pronoun
  - Negative examples (derived) – pairing pronouns with non-antecendent NPs
- Train on set of features
- Given a pro-antecedent pair predict 1 if they co-refer and 0 otherwise.

Features for Pronominal Anaphora Resolution

- Strict number [true or false]
- Compatible number [true or false]
- Strict gender [true or false]
- Compatible gender [true or false]
- Sentence distance [0, 1, 2, 3,…]
- Hobbs distance [0, 1, 2, 3,…] (noun groups)
- Grammatical role [subject, object, PP] – taken by potential antecedent
- Linguistic form [proper, definite, indefinite, pronoun] – of the pronoun

(U1) **John** saw a beautiful 1961 Ford Falcon at the used car dealership.
(U2) **He** showed **it** to Bob.
(U3) He bought it.



	He (U_2)	it (U_2)	Bob (U_2)	John (U_1)
strict number	1	1	1	1
compatible number	1	1	1	1
strict gender	1	0	1	1
compatible gender	1	0	1	1
sentence distance	1	1	1	2
Hobbs distance	2	1	0	3
grammatical role	subject	object	PP	subject
linguistic form	pronoun	pronoun	proper	proper

Figure 21.8

Coreference Resolution

- Harder – must decide if any 2 noun phrases co-refer.

New Features:

- Anaphor edit distance
- Antecedent edit distance
- Alias [true or false] – use named entity tagger
- Appositive [true or false]
- Linguistic form [proper, definite, indefinite, pronoun]

Evaluation

- Look at coreference chains as forming a set
- We represent the fact that A, B, and C corefer by having a class with A, B, and C in it.
- Reference Chain – True Chain – correct or true coreference chain an entity occurs in.
- Hypothesis chain – chain/class assigned to the entity by a coreference algorithm.
- Precision (weighted sum of correct # in hypothesis chain/# in hyp chain) and recall (# correct in hyp chain/# of elements in reference chain)