Speech and Language Processing

Discourse: Anaphora Resolution and Coherence

Dan Jurafsky

Thanks to Diane Litman, Andy Kehler, Jim Martin!!! This material is from J+M, written by Andy Kehler, slides inspired by Diane Litman + Jim Martin

Outline

- Reference
  - Kinds of reference phenomena
  - Constraints on co-reference
  - Preferences for co-reference
  - The Lappin-Leass' algorithm for coreference
- Coherence
  - Hobbs' coherence relations
  - Rhetorical Structure Theory

Reference

Reference Resolution

- Two examples:
  - *John went to Bill's car dealership to check out an Acura Integra. He looked at it for half an hour*
  - *I'd like to get from Boston to San Francisco, on either December 5th or December 6th. It's ok if it stops in another city along they way*
- What is the target of "it"?
  - First example: two possible targets
    - *Bill's car dealership*
    - *An Acura Integra*
  - Second example: where is the target?

Why reference resolution?

- Conversational Agents:
  - See the second example…
  - … airline reservation system needs to know what “it” refers to in order to book correct flight
- Information Extraction:
  First Union Corp. is continuing to wrestle with severe problems unleashed by a botched merger and a troubled business strategy.
  According to industry insiders at Paine Webber, their president, John R. Georgius, is planning to retire by the end of the year.
  - *Their*… what? *First Union Corp.* or *Paine Webber* ?

Some terminology

- *John went to Bill's car dealership to check out an Acura Integra. He looked at it for half an hour*
- Reference: process by which speakers use words John and he to denote a particular person
  - Referring expression: *John*, *he*
  - Referent: the actual entity (but as a shorthand we might call "John" the referent).
  - John and he "corefer"
  - Antecedent: John
  - Anaphor: he
- Cataphora: pronoun before the referent
  - *Before he bought it, John checked over the Integra very carefully*

Many types of reference

- (after Webber, ‘91)
- *According to John, Bob bought Sue an Integra, and Sue bought Fred a Legend*
  - *But that turned out to be a lie* (a speech act)
  - *But that was false* (proposition)
  - *That struck me as a funny way to describe the situation* (manner of description)
  - *That caused Sue to become rather poor* (event)
- But we focus on references to entities
  - The references showed in previous slides

Reference Phenomena (definite / indefinite / inferable)

- Indefinite noun phrases: new to hearer
  - *I saw an Acura Integra today*
  - *Some Acura Integras were being unloaded…*
- Definite noun phrases: identifiable to hearer because
  - Mentioned:
    I saw an Acura Integra today. The Integra was white
  - Identifiable from beliefs (common knowledge):
    The Indianapolis 500
  - Inherently unique:
    The fastest car in Indianapolis 500…
- Inferable
  - *I almost bought an Acura Integra today, but the engine seemed noisy.*
  - The engine of? Easy to infer: *the Acura Integra*

Reference Phenomena (pronouns)

- Pronouns:
  - *I saw an Acura Integra today. It was white*
  - *I saw no less than 6 Acura Integras today. They are the coolest cars.*
- Referent salience, in case of discourse:
  1. *John went to Bob's party, and parked next to a beautiful Acura Integra*
  2. *He went inside and talked to Bob for more than an*
  3. *Bob told him that he recently got engaged.*
  4. *a) He also said that he bought it yesterday.*
     b) He also said that he bought the Acura yesterday.

"it"… what?

Reference phenomena (others)

- Demonstratives
  - *I bought an Integra yesterday.*
    It's similar to the one I bought five years ago.
    ***That** one was really nice, but I like **this** one even better*
  - *This* and *that* often refer metaphorically to time
- A non-pronominal anaphora
  - *I saw no less that 6 Acura Integra today. I want one*
    - … one (of them)

Pronominal Reference Resolution

- Given a pronoun, find the reference
- Constraints to leverage
  - Hard constraints on reference
  - Soft constraints on reference
- Algorithms which use/don’t use these constraints

Hard constraints: syntax

- Number agreement
  - **John has an Acura. They??? are red*
  - *John has an Acura. It is red*
- Person and case agreement
  - **John and Mary have Acuras. We love them???* (who/what???)
  - *John and I have Acuras. We love them.*
- Gender agreement
  - *John has an Acura. He / it is attractive.*
- Syntactic constraints
  - *John bought himself a new Acura* (himself == John)
  - *John bought him a new Acura* (him =/= John)

Soft constraints

Pronoun Interpretation Preferences

- Selectional Restrictions
  - *John parked his Acura in the garage. He had driven it around for hours.*
  - *To drive* needs ”it” to be drivable → *his Acura*
- Recency
  - *John has an Integra. Bill has a Legend. Mary likes to drive it.*
  - *Legend* and *an Integra* are possible targets, but *Legend* is the closest one

Soft constraints

Pronoun Interpretation Preferences

- Syntactic Role: Subject preference
  - *John went to the Acura dealership with Bill. He bought an Integra.*
  - *He* refers to *John* because *John* is the subject

  - *John and Bill went to the Acura dealership. He bought an Integra*
  - Cannot disambiguate…

Soft constraints

Repeated Mention preference

- *John needed a car to get to his new job.*
  He decided that he wanted something sporty.
  Bill went to the Acura dealership with him.
  He bought an Integra.
- *John* is the subject of the previous sentence, and referenced (i.e., repeated) into the second one by means of *him*. Better target than *Bill*.

Soft constraints

Parallelism Preference

- Same structure
  - *Mary went with Sue to the Acura dealership. Sally went with her to the Mazda dealership.*
- But… with similar structure…
  - *Mary went with Sue to the Acura dealership. Sally told her not to buy anything.*

Soft constraints

Verb Semantics Preferences

- *John telephoned Bill. He lost the pamphlet on Acuras.*
- *John criticized Bill. He lost the pamphlet on Acuras.*
- Implicit causality is the best target
  - Implicit cause of criticizing is object
  - Implicit cause of telephoning is subject
- Verbs define such semantic preference

Algorithms for pronoun anaphora resolution

- Knowledge-rich approach
  - Syntactic-based: Hobbs' algorithm
  - Discourse-based: Centering Theory
  - Hybrid approaches: Lappin and Leas
  - Corpus-based: Charniak, Hale, and Ge
- Knowledge-poor approach
  - Machine Learning
- We'll see the Lappin&Leas algorithms

Lappin and Leass

- Lappin and Leass (1994): Given he/she/it, assign antecedent.
- Implements only the soft constraints **recency** and preferences on **syntactic role**
- Two steps
  - Discourse model update
    - When a new noun phrase is encountered, add a representation to discourse model with a salience value
    - Modify saliences.
  - Pronoun resolution
    - Choose the most salient antecedent

Salience Factors and Weights

- Salience given to an NP

1 Recency	100
2 Subject emphasis	80
3 Existential emphasis	70
4 Accusative (direct object) emphasis	50
5 Ind. Obj and oblique emphasis	40
6 Non-adverbial emphasis	50
7 Head noun emphasis	80

Syntactic roles

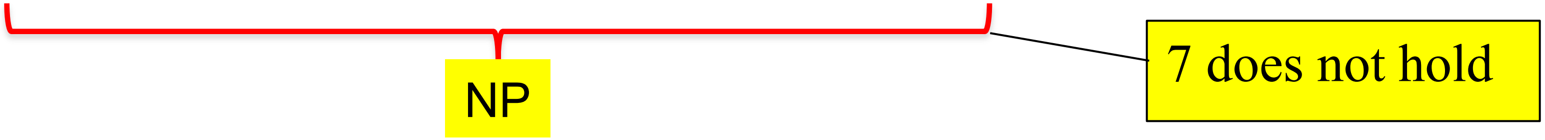
Salience Factors and Weights

1) Give 100 to the latest NP
2) Give 80 to NP acting as the subject of the sentence
3) Give 70 to NP beginning with “there is…” or similar
4) Give 50 to NP acting as the direct object
5) Give 40 to NP acting as indirect object or oblique complements
6) Demarcated adverbial PP:
   - adverbial phrase introduced by coma or adverb (“his”)
   - not a good candidate
   - Thus, give 50 to NP that are **not** a demarcated adverbial PPs
7) Give 80 to NP if NP is **not** part of a larger NP

- Cut in half after each sentence is processed

Example of syntactic roles

- Salience factors 2-6: Syntactic role preference
  - Subject > existential predicate nominal > object > indirect object > demarcated adverbial PP
- Examples for 2-5
  - *An Acura Integra is parked in the lot* (subject)
  - *There is an Acura Integra parked in the lot* (existential pred. nominal)
  - *John parked an Acura Integra in the lot* (object)
  - *John gave his Acura Integra a bath* (indirect obj)
- Add salience if 6 holds (*not* part of demarcated adverbial PP):
  - *Inside his Acura Integra, John showed Susan his new CD player* (here, it is part of demarcated adverbial PP → no salience)
- Add salience if 7 holds (not part of larger NP):
  - *The owner's manual for an Acura Integra is on John's desk*



NP

7 does not hold

Lappin and Leass Algorithm

1. Collect the potential referents (up to 4 sentences back)
2. Remove potential referents that do not agree in number or gender with the pronoun (hard constraints)
3. Compute total salience value of referent from all factors (see table)

- Also, apply the following rules:
  - role parallelism (+35)
  - cataphora (-175).

4. Select referent with highest salience value. In case of tie, select closest.

Example

- ***John*** *saw a beautiful Acura* ***Integra*** *at the* ***dealership****.* *He showed it to Bob. He bought it.*

Sentence 1:

Referent	1 Recency	2 Subject	3 Exist	4 Object	5 Ind-object	6 Non-adv	7 Head N	Total
John	100	80				50	80	310
Integra	100			50		50	80	280
dealership	100					50	80	230

After sentence 1

- Sentence 1 does not contain any pronoun
- So, go to sentence 2
  - Cut all values of sentence 1 in half

Referent	Phrases	Value
John	{John}	155
Integra	{a beautiful Acura Integra}	140
dealership	{the dealership}	115

Sentence 2:
He showed it to Bob

- He specifies male gender
- So Step 2 reduces set of referents to only John.
  - Referent for *He* found!
- Now update discourse model:
  - He in current sentence (recency=100), subject position (=80), not adverbial (=50) not embedded (=80), so add 310:

Referent	Phrases	Value
John	{John, he_1}	155+310
Integra	{a beautiful Acura Integra}	140
dealership	{the dealership}	115

Sentence 2:
He showed it to Bob

- Targets for “it” can be “Integra” or “dealership” (“John” is not a feasible target)
- Need to add "weights:
  - – Parallelism: “it” and “Integra” are objects (“dealership” is not), so +35 for “Integra”
  - – Integra: 175, dealership: 115
    → pick Integra
    - Referent for *it* found!
- Update discourse model:
  - – “it” is object, gets 100+50+50+80=280

Sentence 2:
He showed it to Bob

Referent	Phrases	Value
John	{John, he_1}	465
Integra	{a beautiful Acura Integra, it_1}	140+280
dealership	{the dealership}	115

Sentence 2:
He showed it to Bob

- Bob is a new referent
- Update discourse model:
  - Bob is oblique argument, weight is 100+40+50+80=270

Referent	**Phrases**	**Value**
John	{John, he_1}	465
Integra	{a beautiful Acura Integra, it_1}	420
Bob	{Bob}	270
dealership	{the dealership}	115

Sentence 3:
He bought it

- Drop weights in half:

Referent	**Phrases**	**Value**
John	{John, he_1}	232.5
Integra	{a beautiful Acura Integra, it_1}	210
Bob	{Bob}	135
dealership	{the dealership}	57.5

Then, He_2 will be resolved to John, and it_2 to Integra

Evaluation

- Referential Rate (Byron, 2001)
- RR = C / (T+E)

  C: # pronouns correctly resolved
  T: all referential pronouns
  E: all excluded referential pronouns

Coherence

Text Coherence

- John hid Bill's car keys. He was drunk
- ??John hid Bill's car keys. He likes spinach

What makes a Discourse coherent?

- Assume that you have collected an arbitrary set of well-formed and independently interpretable utterances
- Do you have a discourse?
  - Usually not
  - In general utterances, when juxtaposed, will not exhibit coherence

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**

Hobbs 1979 Coherence Relations

Result

- Infer that the state or event asserted by S0 causes or could cause the state or event asserted by S1.
- *John bought an Acura. His father was not happy.*
- (S0) *John bought an Acura*
  as a direct consequence
  (S1) *His father was not happy.*

Hobbs 1979 Coherence Relations

Explanation

- Infer that the state or event asserted by S1 causes or could cause the state or event asserted by S0
- *John hid Bill's car keys. He was drunk*
- (S0) *John hid Bill's car keys.*
  because
  (S1) *He was drunk*

Hobbs 1979 Coherence Relations

Parallel

- Infer proposition P(a_1, a_2..) from the assertion of S0 and P(b_1,b_2…) from the assertion of S1, where a_i and b_i are similar, for all i.
- *John bought an Acura. Bill leased a BMW.*
- (S0) *John bought an Acura.*
  → Possession(Person, Car)
  (S1) *Bill leased a BMW.*
  → Possession(Person, Car)

Hobbs 1979 Coherence Relations

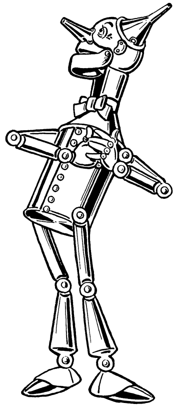
Elaboration

- Infer the same proposition P from the assertions of S0 and S1:
- *John bought an Acura this weekend.*
  He purchased a beautiful new Integra for 20 thousand dollars at Bill's dealership on Saturday afternoon.
- (S0) *John bought an Acura this weekend.*
  (S1) *He purchased a beautiful new Integra ...*
  S1 is just a more precise version of S0

Hobbs 1979 Coherence Relations

Occasion

- A change of state can be inferred from the assertion of S0, whose **final state** can be inferred from S1, or vice versa.
- *Dorothy picked up the oil-can. She oiled the Tin Woodman's joints.*
- (S0) *Dorothy picked up the oil-can.*
  and because of this, at the end
  (S1) *She oiled the Tin Woodman's joints.*



An example

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

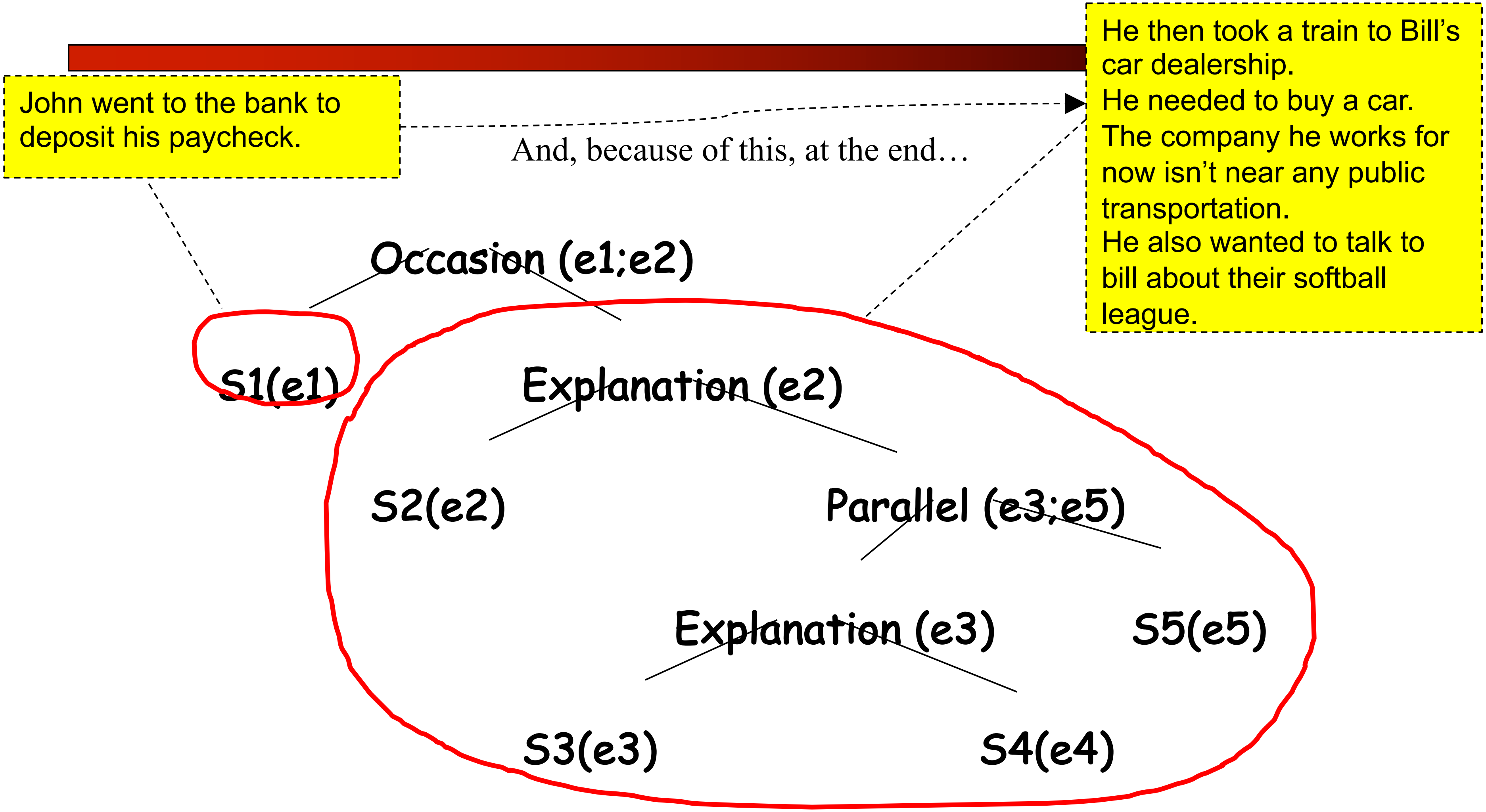
He then took a train to Bill’s car dealership. (S2)

He needed to buy a car. (S3)

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

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

The discourse structure



John went to the bank to deposit his paycheck.

And, because of this, at the end…

He then took a train to Bill's car dealership.
He needed to buy a car.
The company he works for now isn't near any public transportation.
He also wanted to talk to bill about their softball league.

Occasion (e1;e2)

S1(e1)

Explanation (e2)

S2(e2)

Parallel (e3;e5)

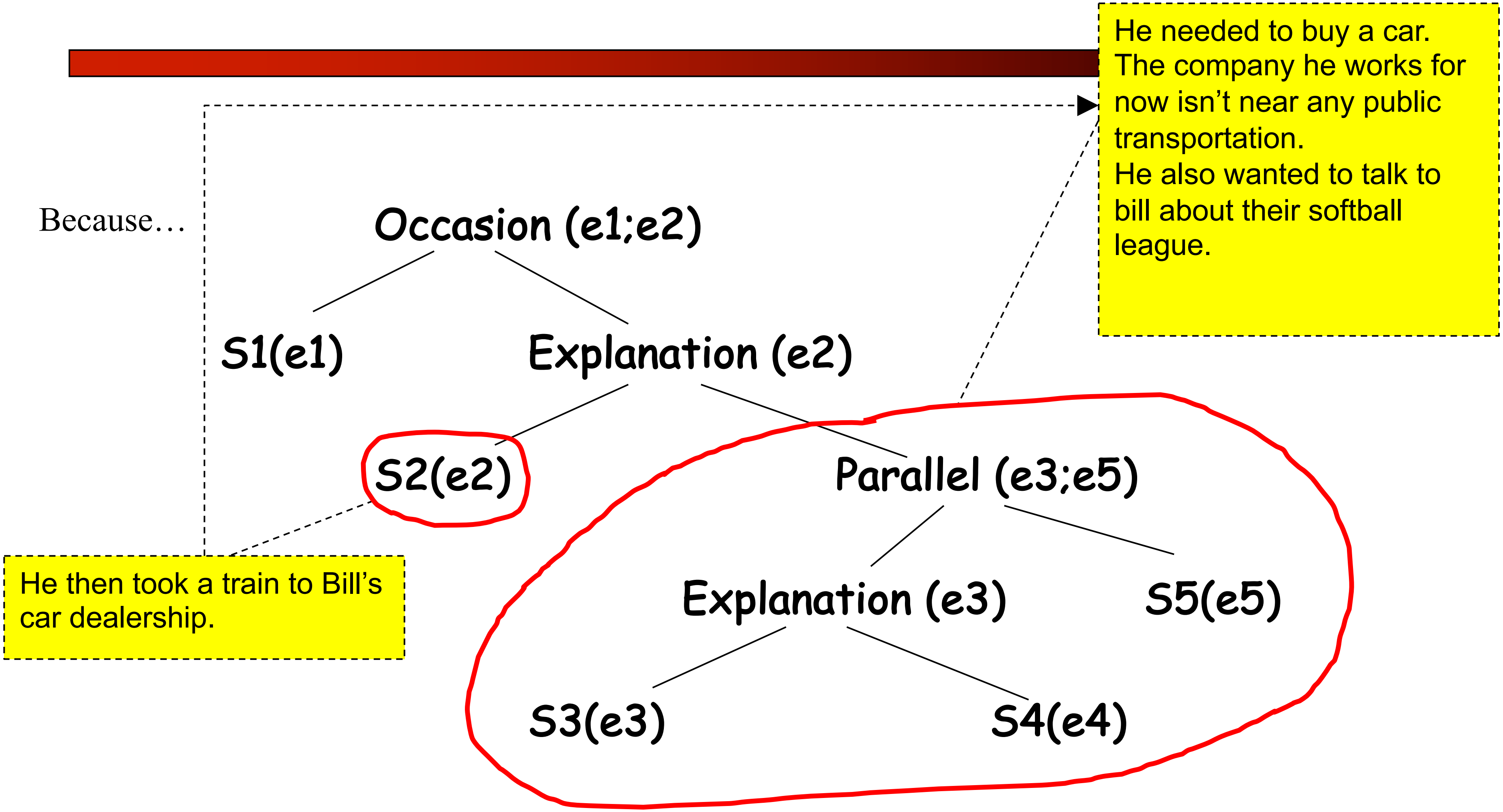
Explanation (e3)

S5(e5)

S3(e3)

S4(e4)

The discourse structure



He needed to buy a car. The company he works for now isn't near any public transportation. He also wanted to talk to bill about their softball league.

Because…

Occasion (e1;e2)

S1(e1)

Explanation (e2)

S2(e2)

Parallel (e3;e5)

Explanation (e3)

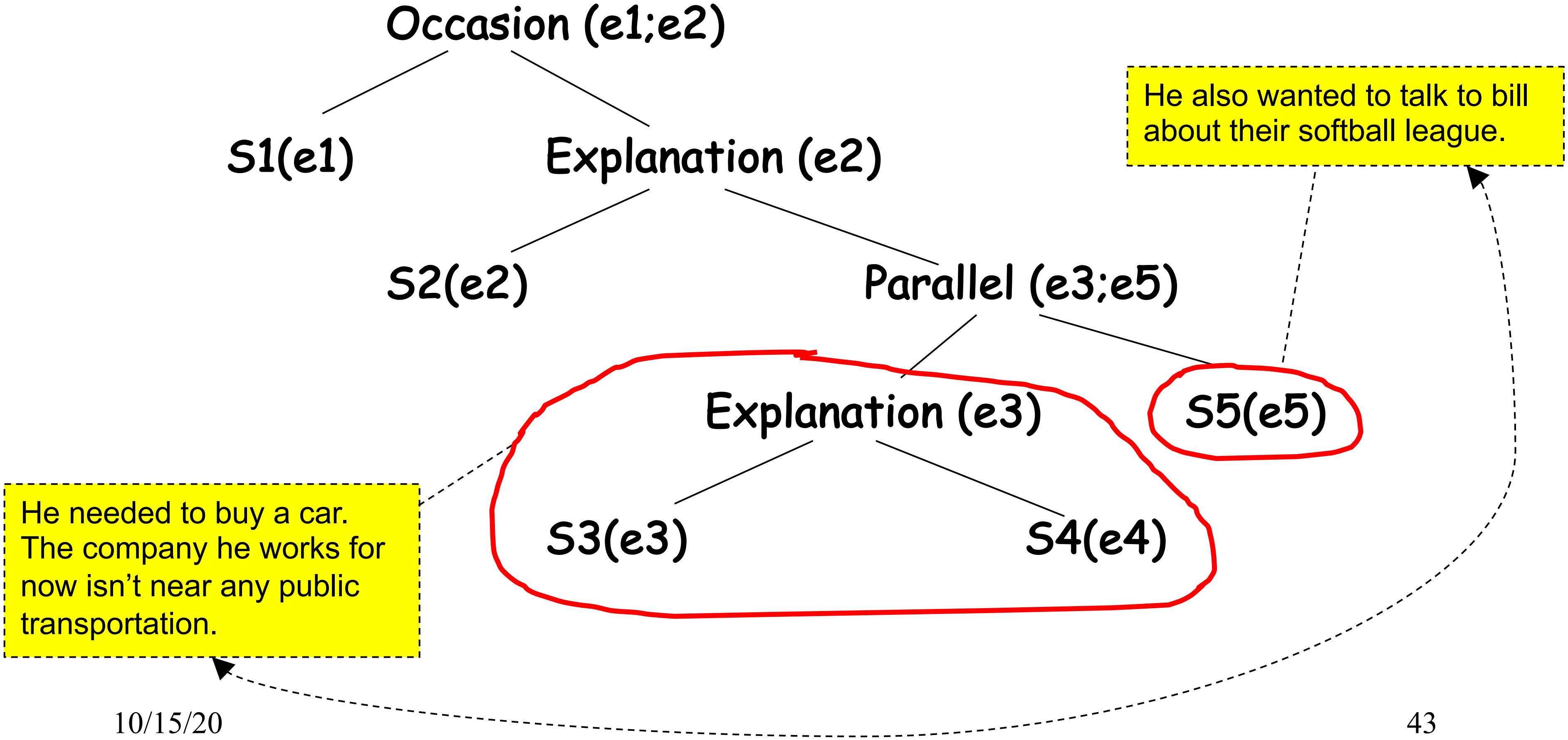
S5(e5)

S3(e3)

S4(e4)

He then took a train to Bill's car dealership.

The discourse structure



Occasion (e1;e2)

S1(e1)

Explanation (e2)

S2(e2)

Parallel (e3;e5)

Explanation (e3)

S5(e5)

S3(e3)

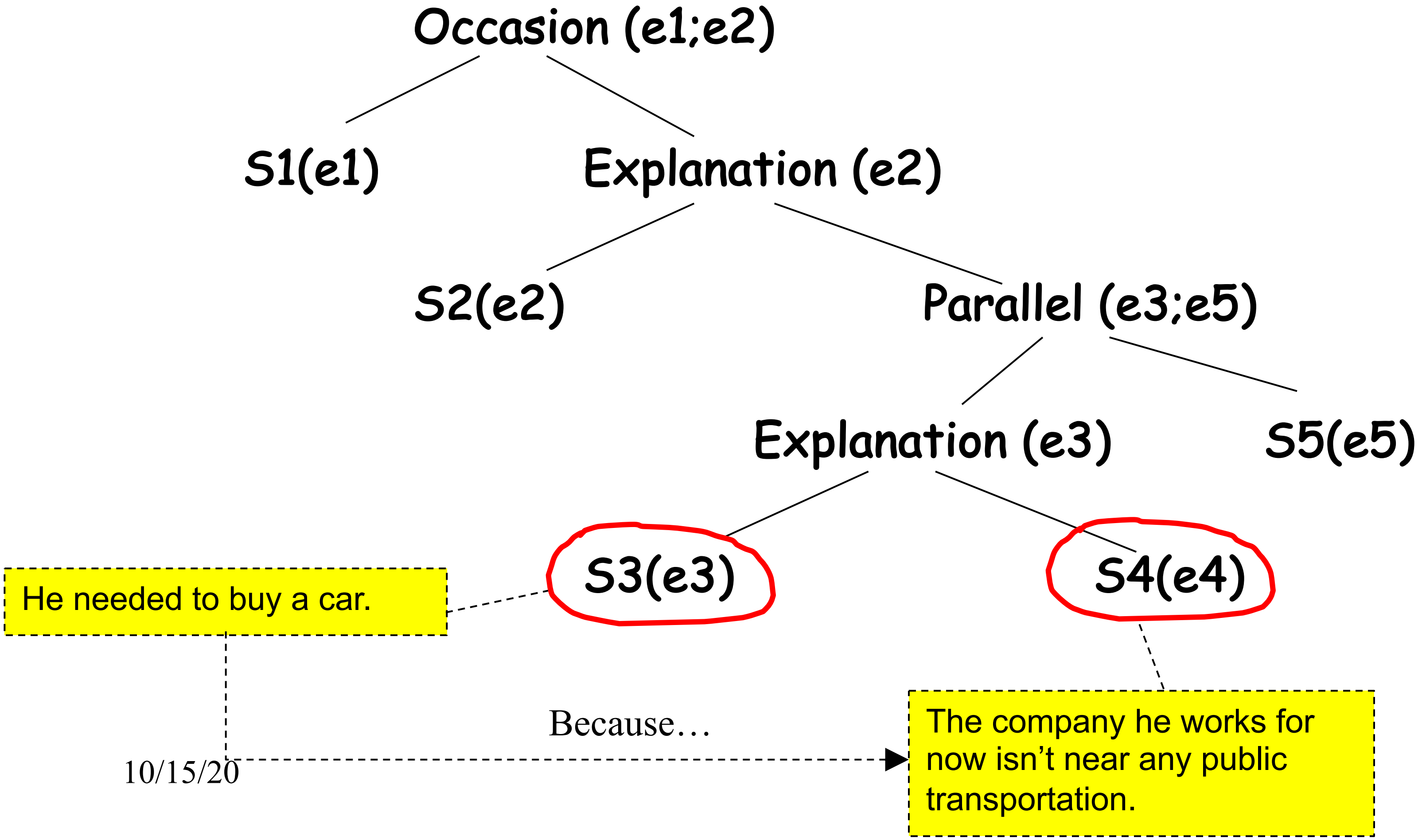
S4(e4)

He also wanted to talk to bill about their softball league.

He needed to buy a car. The company he works for now isn't near any public transportation.

Two parallel actions: to buy and to talk

The discourse structure



Occasion (e1;e2)

S1(e1)

Explanation (e2)

S2(e2)

Parallel (e3;e5)

Explanation (e3)

S5(e5)

S3(e3)

S4(e4)

He needed to buy a car.

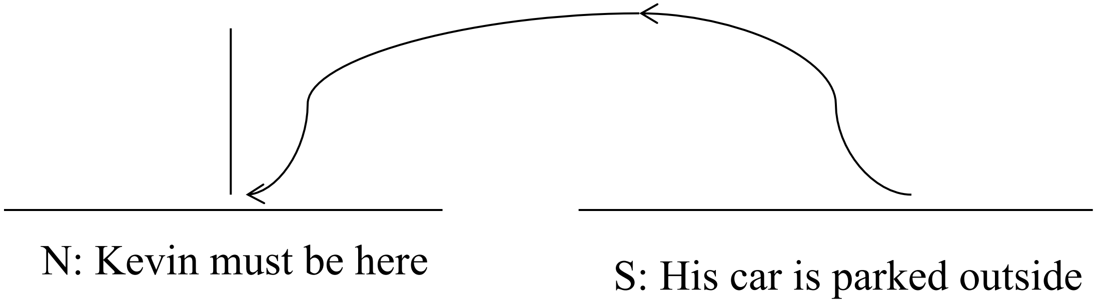
Because…

The company he works for now isn't near any public transportation.

Rhetorical Structure Theory

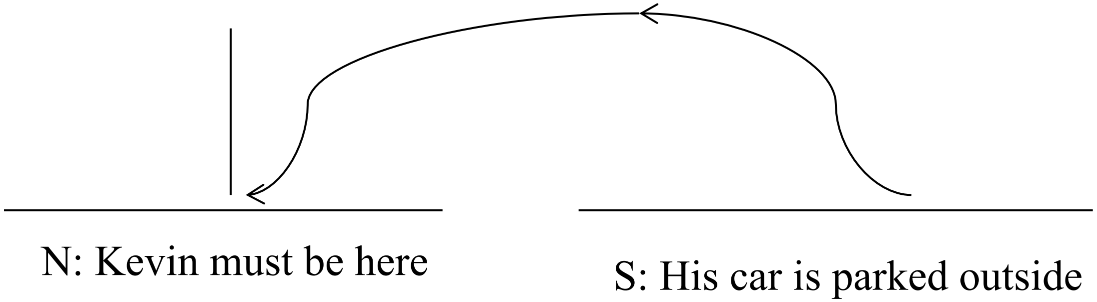
- One theory of discourse structure, based on identifying relations between segments of the text
  - Nucleus/satellite notion encodes asymmetry
  - Some rhetorical relations:
    - Elaboration (set/member, class/instance, whole/part…)
    - Contrast: multinuclear
    - Condition: Satellite presents precondition for N
    - Purpose: Satellite presents goal of the activity in N
    - Background: Satellite gives context for interpreting N
    - Attribution: multinuclear
    - List: multinuclear
    - Evidence: (see in the following)

Relations



- In the original (Mann & Thompson 1987) formulation. An RST relation is formally defined by
  - a set of constraints on the Nucleus (**N**) and satellite (**S**),
  - having to do with the goals and the beliefs of the writer (**W**) and reader (**R**),
  - and by the effect on the reader (**R**)

Relations



- A sample definition
  - Relation: evidence
  - Constraints on Nucleus: Reader might not believe Nucleus to a degree satisfactory to Writer (→ so, evidence is needed)
  - Constraints on Satellite: Reader will believe Satellite or will find it credible
  - Constraints on Nucleus+Satellite: Reader’s comprehending Satellite increases Reader’s belief of Nucleus
  - Effects: Reader’s belief of Nucleus is increased

An example

Mars

With its distant orbit –50 percent farther from the sun than Earth– and slim atmospheric blanket, Mars experiences frigid weather conditions.

Surface temperatures typically average about -60 ˚C at the equator and can dip to -123 ˚C near the poles.

Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure.

An example

- Split discourse in units:

(1) **Mars**
(2) With its distant orbit –50 percent farther from the sun than Earth– and slim atmospheric blanket,
(3) Mars experiences frigid weather conditions.
(4) Surface temperatures typically average about -60 ˚C at the equator
(5) and can dip to -123 ˚C near the poles.
(6) Only the midday sun at tropical latitudes is warm enough
(7) to thaw ice on occasion,
(8) but any liquid water formed in this way would evaporate almost instantly
(9) because of the low atmospheric pressure.

A discurse tree (Marcu 2000)

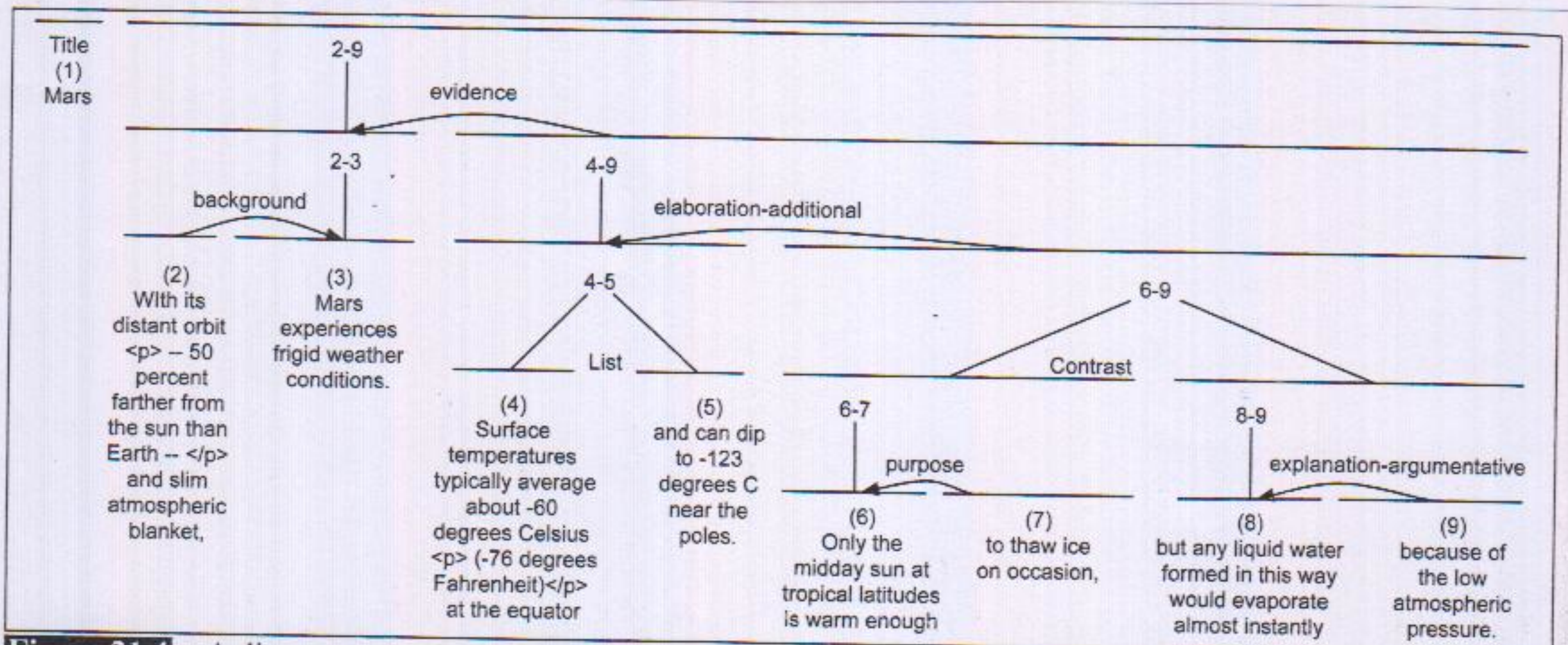

Title
(1)
Mars
2-9
evidence
2-3
background
4-9
elaboration-additional
(2)
With its distant orbit <p> -- 50 percent farther from the sun than Earth -- </p> and slim atmospheric blanket,
(3)
Mars experiences frigid weather conditions.
4-5
List
6-9
Contrast
(4)
Surface temperatures typically average about -60 degrees Celsius <p> (-76 degrees Fahrenheit)</p> at the equator
(5)
and can dip to -123 degrees C near the poles.
6-7
purpose
8-9
explanation-argumentative
(6)
Only the midday sun at tropical latitudes is warm enough
(7)
to thaw ice on occasion,
(8)
but any liquid water formed in this way would evaporate almost instantly
(9)
because of the low atmospheric pressure.

Figure 21.4 A discourse tree for the *Scientific American* text in (21.23), from Marcu (2000a). Note that asymmetric relations are represented with a curved arrow from the satellite to the nucleus.

Automatic Rhetorical Structure Labeling

- Supervised machine learning
  - Get a group of annotators to assign a set of RST relations to a text
  - Extract a set of surface features from the text that might signal the presence of the rhetorical relations in that text
  - Train a supervised ML system based on the training set
- Very difficult!

Features

- Explicit markers: *because, however, therefore, then, etc.*
- Tendency of certain syntactic structures to signal certain relations: Infinitives are often used to signal purpose relations: *Use rm* ***to delete files.***
- Ordering
- Tense/aspect
- Intonation (if text is the transcription of an utterance)

Some Problems with RST

- How many Rhetorical Relations are there?
- How can we use RST in dialogue as well as monologue?
- Difficult to get annotators to agree on labeling the same texts
  → very difficult to create good corpora