Coreference Resolution and Discourse Coherence

Natalie Parde

UIC CS 421

This Week's Topics

Tuesday

- Coreference Resolution Approaches
- Evaluating Coreference Resolution
- Discourse Relations

Thursday

- Discourse Parsing
- Entity-Based Coherence
- Topical Salience and Global Coherence

This Week's Topics

Tuesday

- Coreference Resolution Approaches
- Evaluating Coreference Resolution
- Discourse Relations

Thursday

- Discourse Parsing
- Entity-Based Coherence
- Topical Salience and Global Coherence

Coreference Tasks

- We can formalize the task of coreference resolution as follows:
  - **Given a text *T*, find all entities and the coreference links between them**
- This requires a few subtasks:
  - **Detect mentions**
    - Likely to be mentions:
      - Pronouns
      - Definite noun phrases
      - Indefinite noun phrases
      - Names
    - Exclude non-referential pronouns or noun phrases
  - **Link those mentions into clusters**

What counts as a mention?

- Depends on the task specifications and dataset
- Some coreference datasets do not include singletons as mentions
  - Makes the task easier
    - Singletons are often difficult to distinguish from non-referential noun phrases, and constitute a majority of mentions

Sample Coreference Task

The University of Illinois at Chicago is an excellent place to study natural language processing. UIC has many faculty currently working in NLP, including but not limited to Natalie Parde, Barbara Di Eugenio, Cornelia Caragea, Bing Liu, and Philip Yu. The school is located in bustling downtown Chicago, and as a bonus it will be opening a snazzy new CS building in 2024.

Sample Coreference Task

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a bonus **it** will be opening a snazzy new **CS building** in 2024.

Detect mentions

Sample Coreference Task

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a bonus **it** will be opening a snazzy new **CS building** in 2024.

Detect mentions

Cluster mentions

Sample Coreference Task

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a bonus **it** will be opening a snazzy new **CS building** in 2024.

Detect mentions

Cluster mentions

Coreference Chains:

- {University of Illinois at Chicago, UIC, The school, it}
- {natural language processing, NLP}
- {faculty}
- {Natalie Parde}
- {Barbara Di Eugenio}
- {Cornelia Caragea}
- {Bing Liu}
- {Philip Yu}
- {Chicago}
- {CS building}

Popular Coreference Datasets

OntoNotes

- Chinese, English, and Arabic texts in a variety of domains (e.g., news, magazine articles, speech data, etc.)
- No singletons
- https://catalog.ldc.upenn.edu/LDC2013T19

ISNotes

- Adds information status to OntoNotes
- https://github.com/nlpAThits/ISNotes1.0

ARRAU

- English texts in a variety of domains
- Includes singletons
- https://catalog.ldc.upenn.edu/LDC2013T22

Moving on to the finer details….

- Mention detection: The process of finding spans of text that constitute a referring expression (mention)
  - It’s common to be very liberal in predicting mentions, and rely on downstream filtering to prune bad predictions

The **University of Illinois at Chicago** is an excellent ~~place~~ to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a **~~bonus~~ it** will be opening a snazzy new **CS building** in 2024.

Mention Detection

- How is filtering performed?
  - Sometimes, rules
  - More often, **classifiers**
- Classifiers for mention filtering often make use of features characterizing the words, their relationship, and their position in the surrounding text

1. Take all predicted mentions
2. Remove numeric quantities, mentions embedded in larger mentions, and stop words
3. Remove non-referential "it" based on regular expression patterns

Mention filtering can be a tricky balance!

- Filter too many → recall suffers
- Filter too few → precision suffers
- Some recent approaches also perform mention detection, filtering, and entity clustering jointly in an end-to-end model

Architectures for Coreference Algorithms

Several different ways to tackle the problem:

- **Entity-based classification**
  - Make decisions based on a given entity in the discourse model as a whole
- **Mention-based classification**
  - Make decisions locally for each mention
- **Ranking models**
  - Compare potential antecedents with one another (can be combined with either entity-based or mention-based approaches)

The Mention-Pair Architecture

Simple premise:

Given:

- Pair of mentions (candidate anaphor and candidate antecedent)

Decide:

- Whether or not they corefer

How does this work?

Compute coreference probabilities for every plausible pair of mentions

Goal: High probability for actual coreferring pairs, and low probability for other pairs

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as **a bonus** **it** will be opening a snazzy new **CS building** in 2024.

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as **a bonus** **it** will be opening a snazzy new **CS building** in 2024.

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a **bonus** **it** will be opening a snazzy new **CS building** in 2024.

The Mention-Pair Architecture

The **University of Illinois at Chicago** is an excellent place to study **natural language processing**. **UIC** has many **faculty** currently working in **NLP**, including but not limited to **Natalie Parde**, **Barbara Di Eugenio**, **Cornelia Caragea**, **Bing Liu**, and **Philip Yu**. **The school** is located in bustling downtown **Chicago**, and as a **bonus** **it** will be opening a snazzy new **CS building** in 2024.

How do we learn these probabilities?

- Select training samples
  - For every one positive instance (m_i, m_j) where m_j is the closest antecedent to m_i,
  - Extract numerous negative instances (m_i, m_k) for each m_k between m_j and m_i
- Extract features
  - Manually engineered features, and/or
  - Implicitly learned representations
- Train classification model

How do we make predictions?

- Apply the trained classifier to each test instance in a clustering step
  - **Closest-first clustering**
    - For mention *i*, classifier is run backwards through prior *i*-1 mentions
    - First prior mention (candidate antecedent) with probability > 0.5 is selected and linked to *i*
  - **Best-first clustering**
    - Classifier is run on all possible *i*-1 antecedents (all mentions prior to mention *i*)
    - Mention with highest probability is selected as the antecedent for *i*

Mention-Pair Architecture

- Advantage:
  - **Simplest** coreference resolution architecture
- Disadvantage:
  - **Doesn't directly compare candidate antecedents** with one another
  - **Considers only mentions**, not overall entities

How can we address these limitations?

- One option: The **Mention-Rank Architecture**
  - Directly compares antecedents with one another
  - Selects the highest-scoring antecedent for each anaphor
- How does this work?
  - For a mention *i*, we have:
    - Random variable y_i ranging over the values $Y(i) = \{1, \ldots, i-1, \varepsilon\}$
      - ε = dummy mention meaning *i* does not have an antecedent
  - When training:
    - Use heuristics to determine the best antecedent for an anaphor (e.g., closest = best)
    - Or, learn more optimal ways to model latent antecedents using machine learning
  - At test time, for *i* the model computes a softmax over all possible antecedents

Another Option: Entity-based Models

- Considers discourse entities, rather than individual mentions
- How does this work?
  - Have the model make decisions over clusters of mentions, where each cluster corresponds to an entity
  - Can be implemented using feature-based or neural classifiers

How can we implement mention-pair, mention-rank, and entity-based architectures?

- Traditional machine learning models using manually-defined features
- Neural models

Feature-based Classification Models

- Common feature types:
  - Features of the candidate anaphor
  - Features of the candidate antecedent
  - Features of the relationship between the pair
- For entity-based models, this can also include:
  - Features of all mentions of the candidate antecedent’s entity cluster
  - Features of the relation between the candidate anaphor and the mentions of the candidate antecedent in the entity cluster

What would be examples of these features?

- First word
- Head word
- Gender
- Named entity type
- Length
- Grammatical role
- Document genre
- …and many more!

Neural Classification Models

- Generally end-to-end without a separate mention detection step
  - Instead, consider every possible text span of length < *k* as a possible mention
- Same overall goal as usual:
  - Assign to each span *i* an antecedent y_i ranging over the values $Y(i) = \{1, \ldots, i-1, \varepsilon\}$

What goes on behind the scenes?

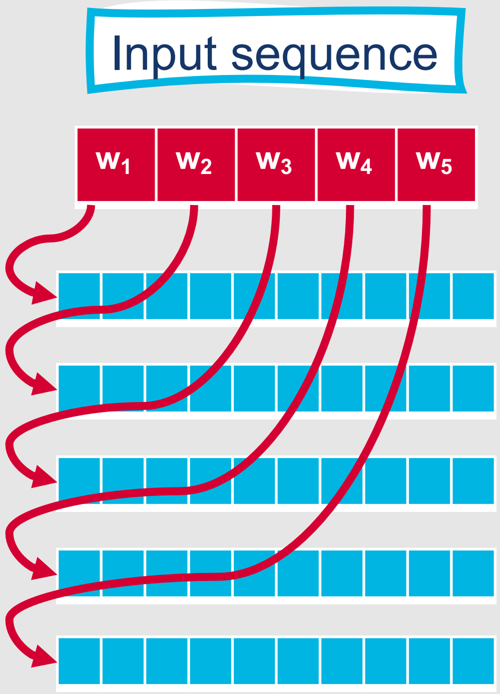
- For each pair of spans *i* and *j*, the system assigns a score $s(i,j)$ for the coreference link between the two
  - $s(i,j) = m(i) + m(j) + c(i,j)$
    - $m(i)$: Whether span *i* is a mention
    - $m(j)$: Whether span *j* is a mention
    - $c(i,j)$: Whether *j* is the antecedent of *i*
- The functions $m(\cdot)$ and $c(\cdot,\cdot)$ are computed using neural models:
  - $m(i) = w_m \cdot NN_m(g_i)$
  - $c(i,j) = w_c \cdot NN_c([g_i, g_j, g_i \circ g_j, \phi(i,j)])$
    - For example, where g_i is a vector representation of span *i* and $\phi(i,j)$ encodes manually-defined characteristics of the relationship between *i* and *j*
    - Exact definition of $c(i,j)$ may differ across models

Altogether, a neural coreference resolution model might look like the following….

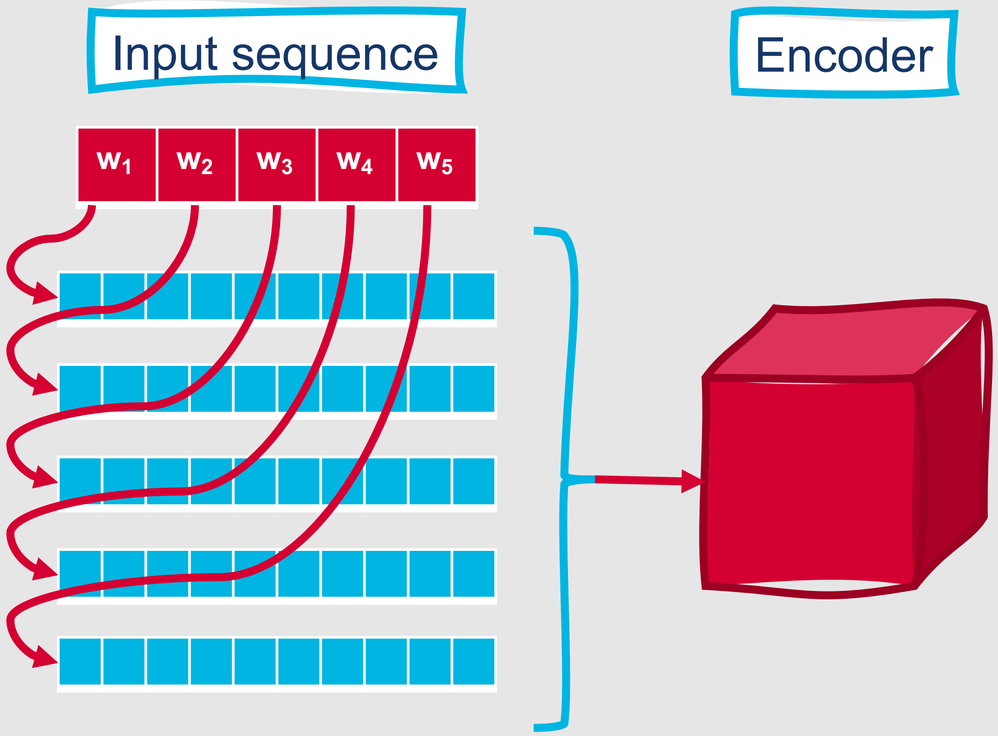
Input sequence

w_1	w_2	w_3	w_4	w_5

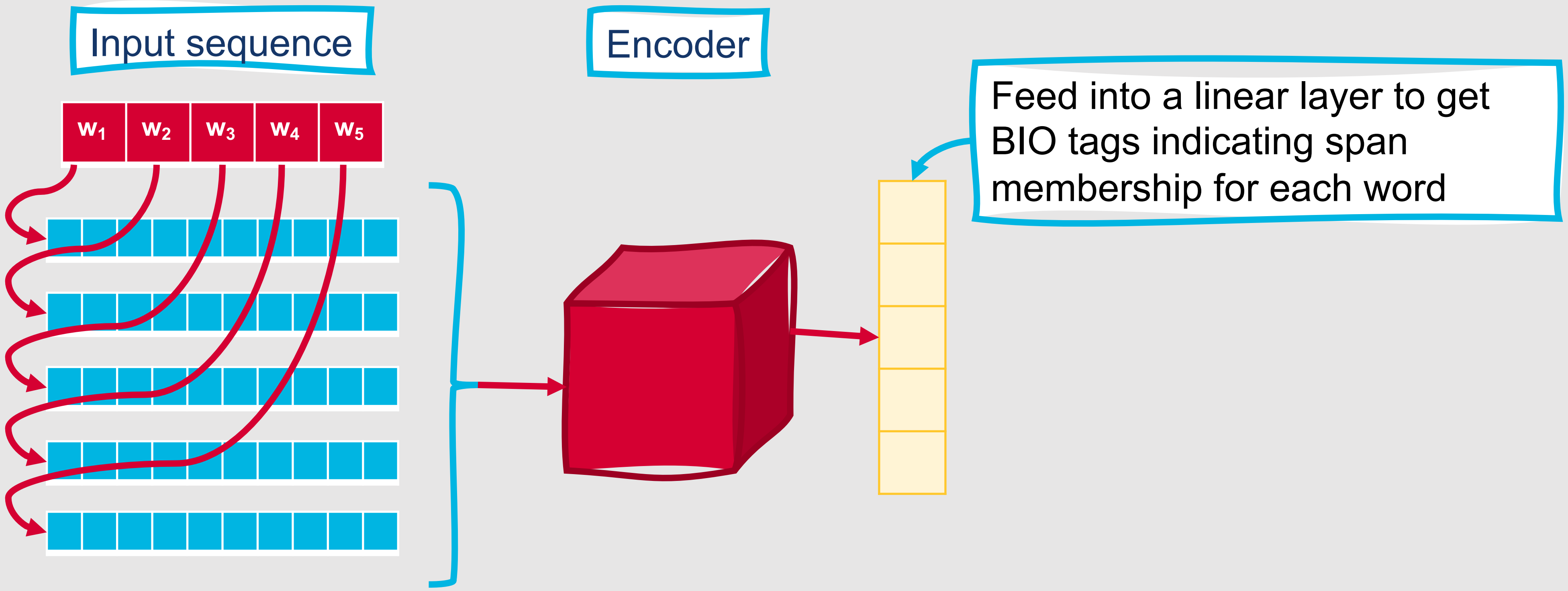
Altogether, a neural coreference resolution model might look like the following….



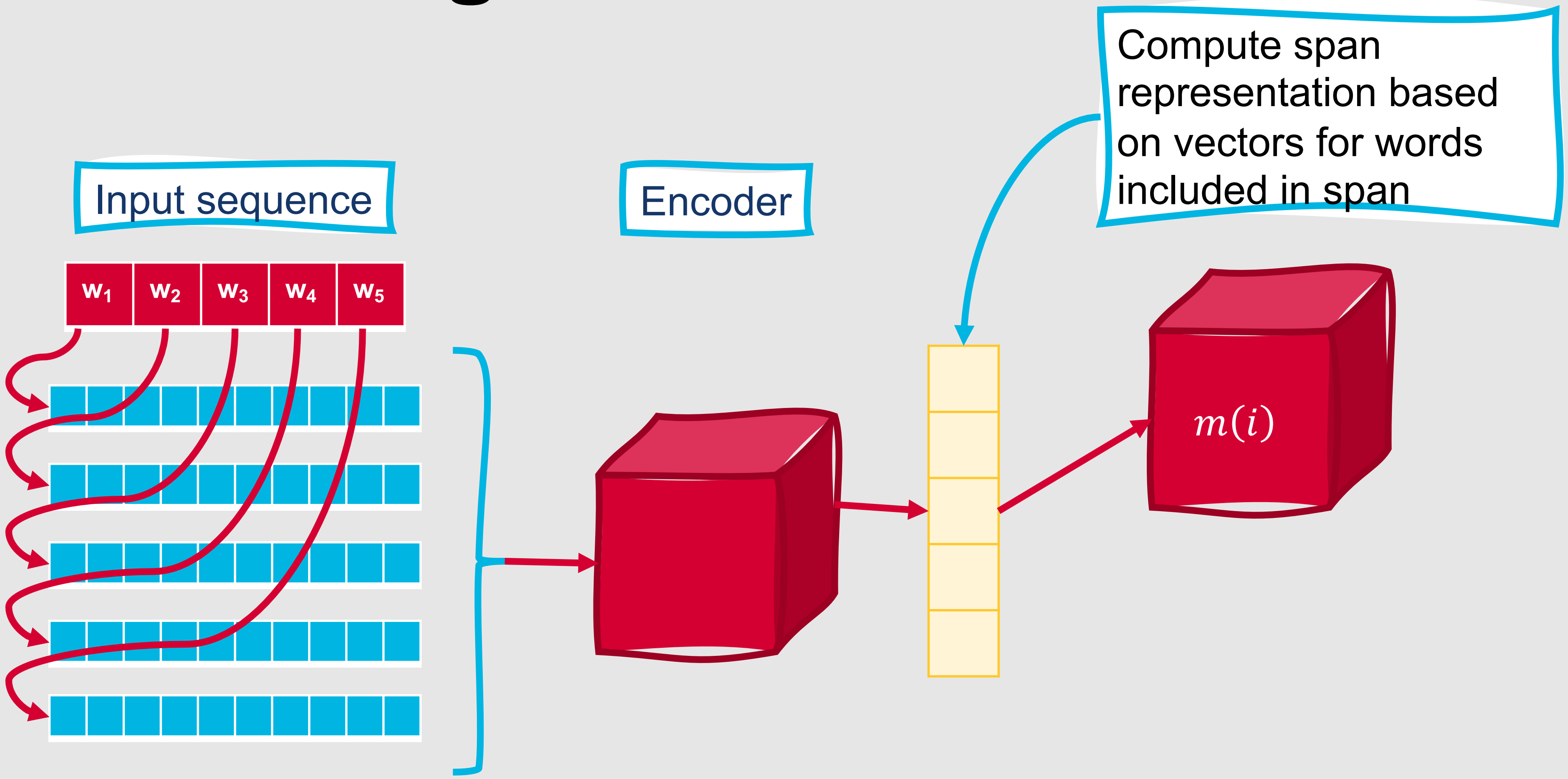
Altogether, a neural coreference resolution model might look like the following….



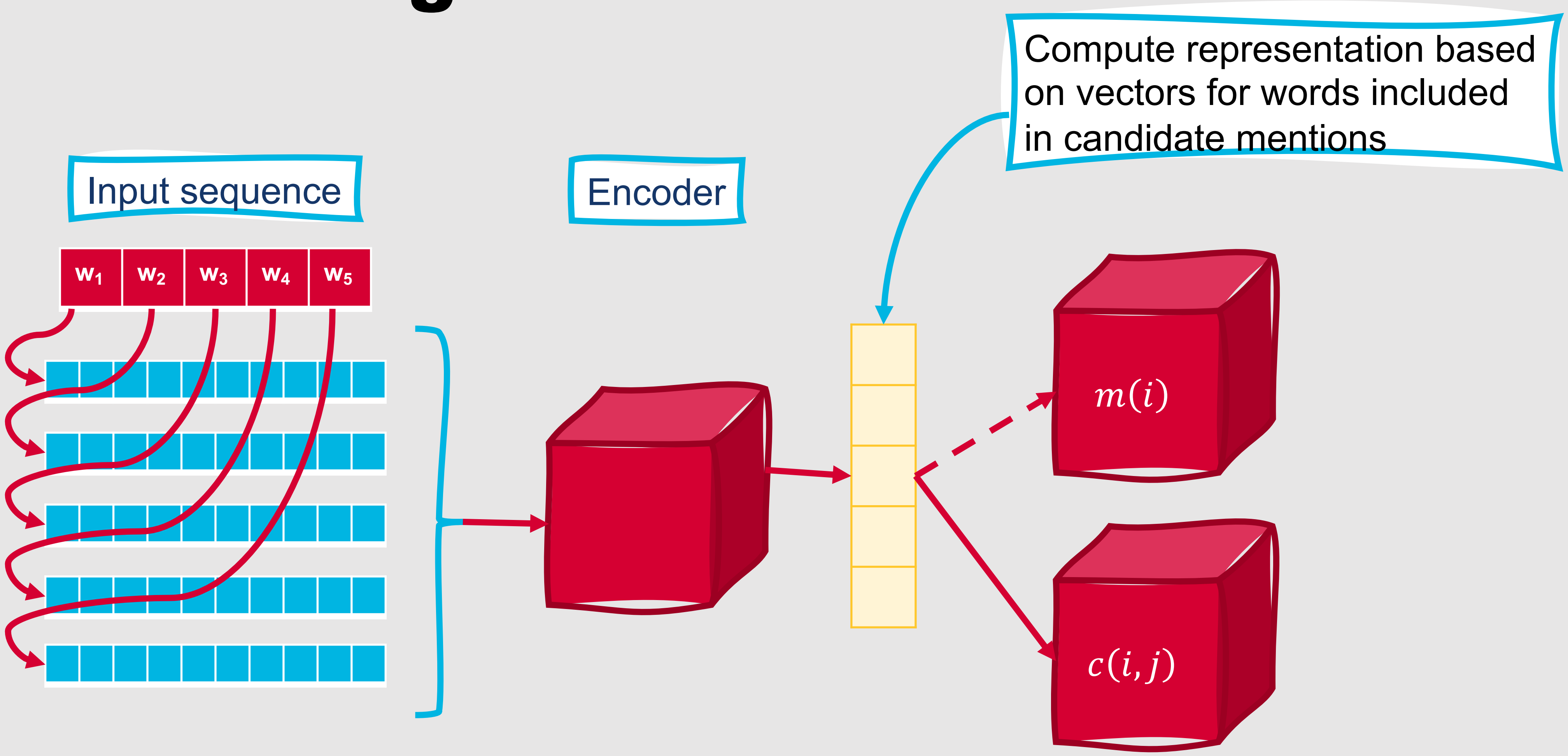
Altogether, a neural coreference resolution model might look like the following….



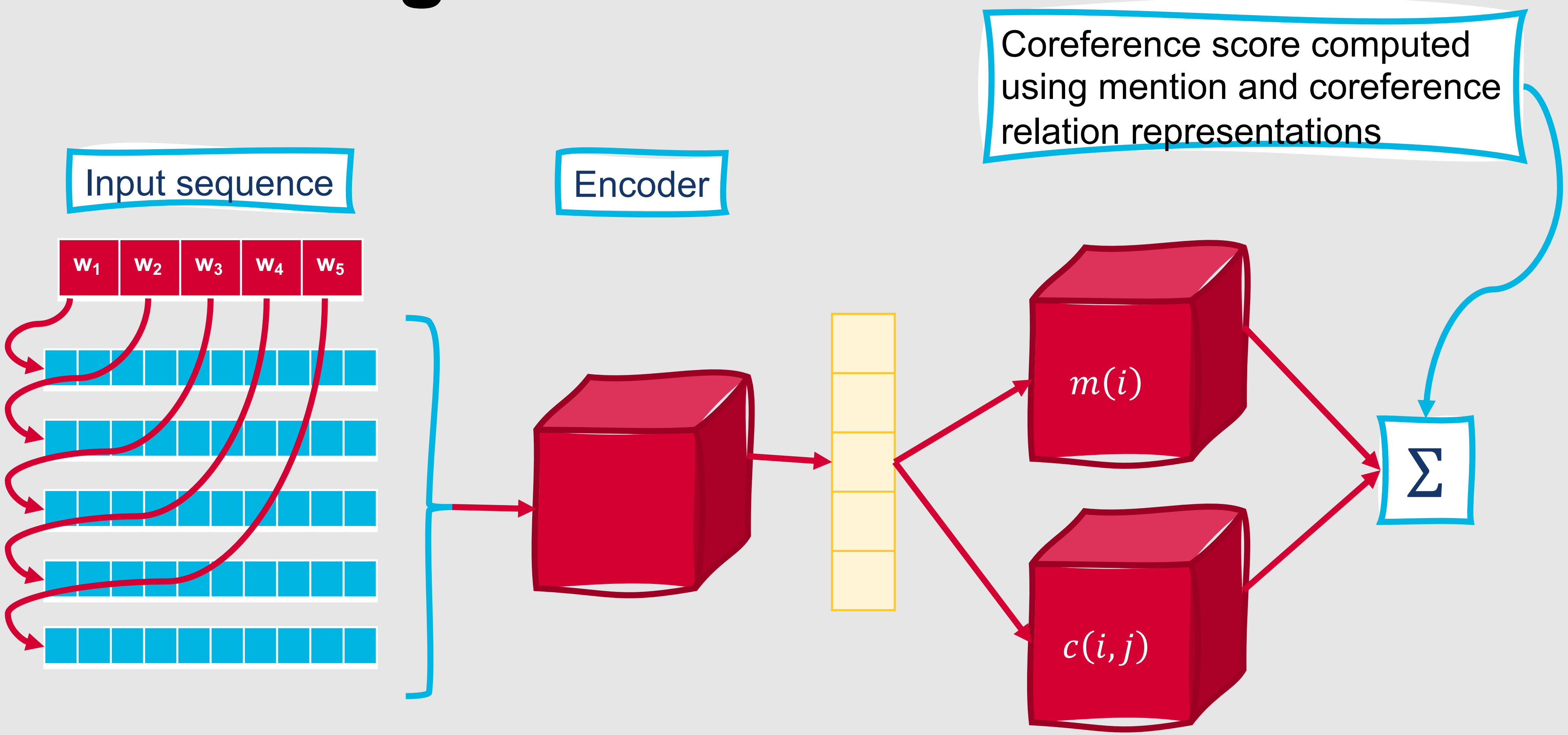
Altogether, a neural coreference resolution model might look like the following....



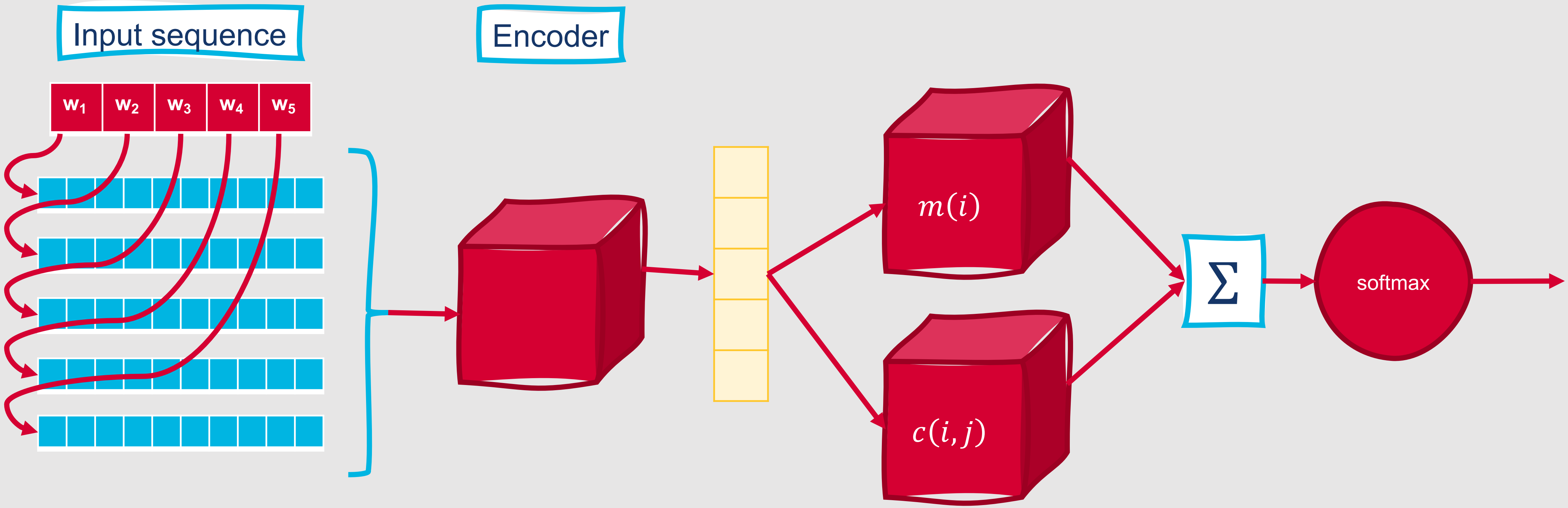
Altogether, a neural coreference resolution model might look like the following….



Altogether, a neural coreference resolution model might look like the following….



Altogether, a neural coreference resolution model might look like the following....



This Week's Topics

Thursday

Tuesday

Tuesday:

- Coreference Resolution Approaches
- Evaluating Coreference Resolution
- Discourse Relations

Thursday:

- Discourse Parsing
- Entity-Based Coherence
- Topical Salience and Global Coherence

How do we evaluate coreference resolution models?

- Compare hypothesis coreference chains or clusters with a gold standard
- Compute precision and recall

How do we compute precision and recall?

- Several approaches:
  - **Link-based:** MUC F-measure
  - **Mention-based:** B^3

MUC F-Measure

- Message Understanding Conference (MUC)
- True positives = Common coreference links (anaphor-antecedent pairs) between hypotheses and gold standard
- Precision = # Common links / # Links in hypotheses
- Recall = # Common links / # Links in gold standard
- A couple downsides to this approach:
  - Biased towards systems that produce large coreference chains
  - Ignores singletons (no links to count)

B^3

- Mention-based
- True positives for a given mention, *i* = # Common mentions in hypothesis and gold standard coreference chain including *i*
- Precision for a given mention, *i* = TP / # Mentions in hypothesis coreference chain including *i*
- Recall for a given mention, *i* = TP / # Mentions in gold standard coreference chain including *i*
- Total precision and recall are the weighted sums of precision and recall across all mentions

So …where are we now?

- Still plenty of room for growth in coreference resolution!
- Recently, lots of interest in **Winograd Schema** problems
  - Coreference resolution problems that are:
    - Easy for humans to solve
    - Particularly challenging for computers to solve, due to their reliance on world knowledge and commonsense reasoning

Winograd Schema Problems

- Winograd Schema problems are characterized by the following:
  - There are two statements that differ by only one word or phrase
  - There are two entities that remain the same across statements
  - A pronoun preferentially refers to one of the entities, but could grammatically also refer to the other
  - A question asks to which entity the pronoun refers
  - If one word/phrase in the question is changed, the human-preferred answer changes to the other entity

Example Winograd Schema Problem

Nikolaos lost the race to Giuseppe because he was **slower**.

Who was “he”?

Nikolaos

Example Winograd Schema Problem

Nikolaos lost the race to Giuseppe because he was **slower**.

Who was “he”?

Nikolaos

Nikolaos lost the race to Giuseppe because he was **faster**.

Who was “he”?

Giuseppe

Example Winograd Schema Problem

Nikolaos lost the race to Giuseppe because he was **slower**.

Who was “he”?

Nikolaos

Nikolaos lost the race to Giuseppe because he was **faster**.

Who was “he”?

Giuseppe

Best way to solve Winograd Schema problems computationally?

- Currently, a mix of language modeling and external knowledge bases

Gender Bias in Coreference Resolution

- As with language modeling, coreference resolution systems can exhibit harmful gender biases
- How can we avoid these issues?
  - One solution: Increase sample size for underrepresented genders
    - Artificially: Generate gender-swapped versions of existing training corpora
    - Manually: Collect new, gender-balanced corpora
  - Other solutions?
    - Still very much an active research question!

This Week's Topics

Tuesday

- Coreference Resolution Approaches
- Evaluating Coreference Resolution
- Discourse Relations

Thursday

- Discourse Parsing
- Entity-Based Coherence
- Topical Salience and Global Coherence

What is discourse coherence?

- The relationship (or lack thereof) between sentences in a **discourse**

I really like my class, CS 421. UIC is in Chicago. It's about natural language processing.

UIC is in Chicago, and I'm taking a class there called CS 421. I really like the class. It's about natural language processing.

What counts as a discourse?

- Discourses in NLP are structured, collocated groups of sentences
  - Chapter of a book
  - News article
  - Conversation
  - Twitter thread
  - Wikipedia page
- Discourses should be coherent, rather than random combinations of sentences

What makes a discourse coherent?

- Local and global factors
  - Relations between text units
  - Degree to which the next text unit is anticipated or can be inferred
  - Entity salience
  - Topical salience
  - Overall structure

I really like my class, CS 421. **UIC is in Chicago.** 😕 **It's** 😕 about natural language processing.

UIC is in Chicago, **and I'm taking a class there** 🙂 called CS 421. I really like **the class** 🙂. **It's** 🙂 about natural language processing.

Why do we care whether a discourse is coherent?

- Measuring discourse coherence is important for measuring the quality of a given text
- Also helpful for:
  - Automated essay grading
  - Determining which sentences to include in automatically-generated summaries
  - Measuring mental or cognitive health

How do we measure discourse coherence?

- Some key techniques:
  - Identify coherence relations
  - Determine entity salience
  - Measure lexical cohesion
  - Identify argument structure

Coherence Relations

- Connections between spans of text in a discourse
- Two commonly-used models:
  - **Rhetorical Structure Theory (RST)**
  - **Penn Discourse Treebank (PDTB)**

Rhetorical Structure Theory

- Based on a set of 23 **rhetorical relations** that can hold between spans of text within a discourse
- Most relations are between two spans:
  - **Nucleus**
    - More central to the writer's purpose
    - Interpretable independently
  - **Satellite**
    - Less central to the writer's purpose
    - Only interpretable with respect to the nucleus

Rhetorical Structure Theory

- Relations are **asymmetric**
  - Represented graphically with arrows pointing from the satellite to the nucleus
- Relations are defined by a **set of constraints** on the nucleus and satellite
- Constraints are based on:
  - **Goals and beliefs** of the writer and reader
  - **Effect** on the reader

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the accept the information provided in the

Natalie told the class that there was nothing due on Friday next week, reminding them that Project Part 3 was due on Wednesday instead.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Natalie told the class that there was nothing due on Friday next week.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Outside was freezing, but **inside was uncomfortably warm**.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the accept the information provided in the nucleus

In the fall, Natalie taught CS 421; in the spring, Natalie taught CS 521; in the summer, Natalie worked on research.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along s
List	A series of nuclei is given, without c
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Natalie spent a lot of time walking around the campus on Monday. She had meetings in many different buildings.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along s
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Natalie must be here. Her office door is cracked open.

Summary: Coreference Resolution and Discourse Relations

- Architectures for coreference resolution systems may be **mention-based** or **entity-based**, and may or may not compare potential **antecedents** with one another
- Models for coreference resolution may learn based on **manually defined features**, **neural features**, or a combination of the two
- Computing precision and recall for coreference resolution systems may be done using either **link-based** or **mention-based** methods
- **Winograd Schema** problems are particularly challenging coreference resolution tasks that rely on world knowledge and commonsense reasoning
- Care should be taken to avoid introducing harmful **gender biases** into coreference resolution systems
- **Discourse coherence** is the relationship (or lack thereof) between sentences in a discourse
- It is influenced by a variety of factors:
    - **Coherence relations**
    - **Entity salience**
    - **Topical salience**
    - **Global structure**

This Week's Topics

Tuesday

- Coreference Resolution Approaches
- Evaluating Coreference Resolution
- Discourse Relations

Thursday

- Discourse Parsing
- Entity-Based Coherence
- Topical Salience and Global Coherence

RST relations can be hierarchically organized into discourse trees.

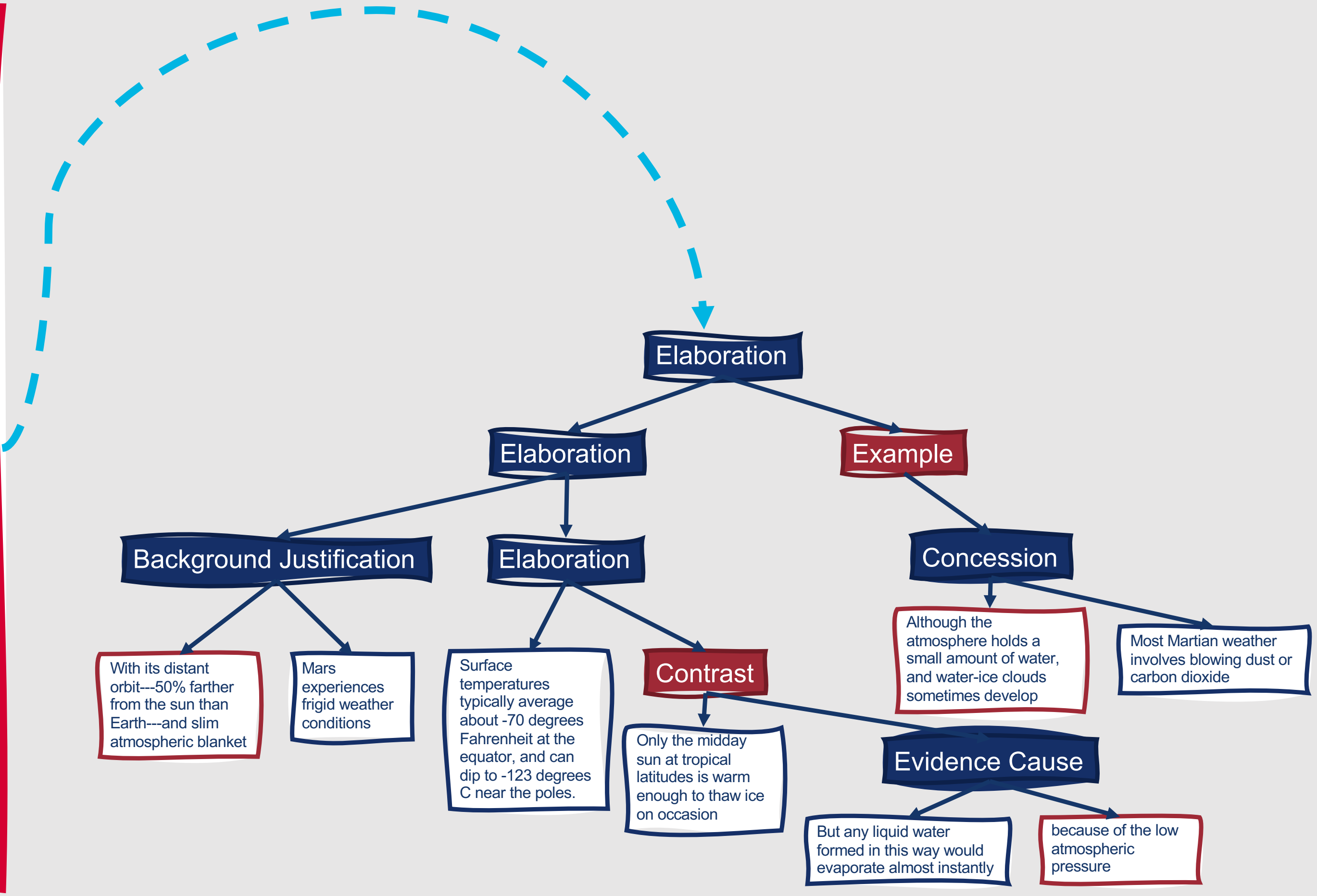
With its distant orbit-–50% farther from the sun than Earth-–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees C near the poles.

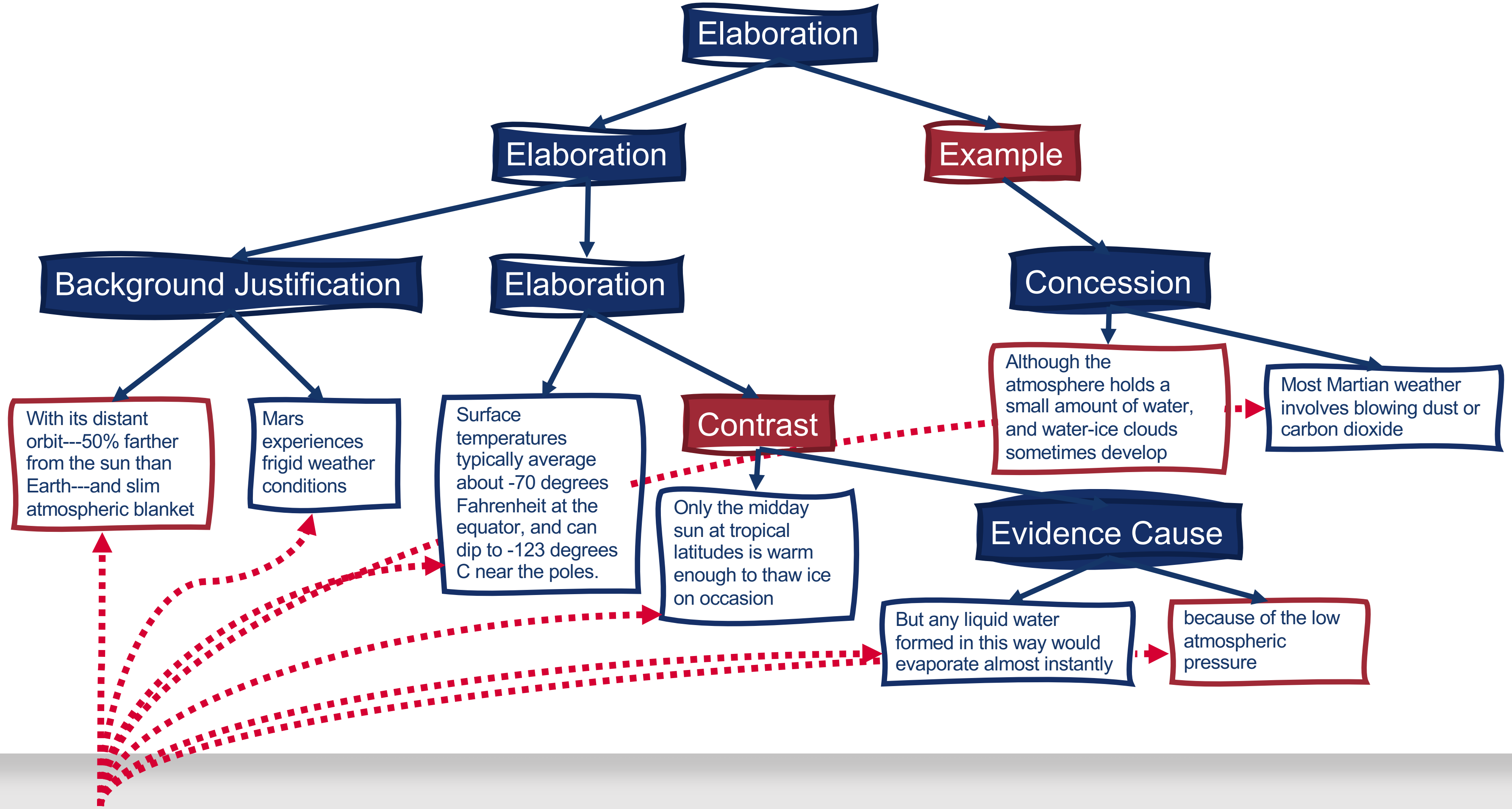
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.

Example Discourse Tree

With its distant orbit-–50% farther from the sun than Earth-–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees 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. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.





Elementary Discourse Units (EDUs)

- Leaves in a discourse tree
  - Also referred to as discourse segments
- Determining the boundaries of EDUs is important for extracting coherence relations

RST Corpora

- **RST Discourse Treebank**
  - 385 English-language documents with full RST parses
  - https://catalog.ldc.upenn.edu/LDC2002T07
- **RST Treebanks for Non-English Data:**
  - CST-News (Brazilian Portuguese): http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/
  - Rhetalho and CorpusTCC (Brazilian Portuguese): https://sites.icmc.usp.br/taspardo/Projects.htm
  - Spanish RST DT (Spanish): http://corpus.iingen.unam.mx/rst/index_en.html
  - Potsdam Commentary Corpus (German): http://angcl.ling.uni-potsdam.de/resources/pcc.html
  - Basque RST DT (Basque): http://ixa2.si.ehu.es/diskurtsoa/en/

Penn Discourse Treebank

- **Lexically-grounded** model of coherence relations
  - Given a list of **discourse connectives** (e.g., *because*, *although*, *when*, *since*, or *as a result*) and an unlabeled document, annotators labeled:
    - Those connectives
    - The spans of text that they connected
  - In some cases, these connectives may be implicit

Arg1

Natalie spent most of her weekend catching up on work. **As a result**, she had a lighter workload on Monday.

Arg2

PDTB Semantic Hierarchy

- Four main classes:
  - Temporal
  - Contingency
  - Comparison
  - Expansion
- Numerous subtypes of each

PDTB Annotations

- Only at the span-pair level!
- No hierarchical tree structure

PDTB Corpus

- 50k+ annotated relations
- Built on top of the Wall Street Journal section of the Penn Treebank
- https://catalog.ldc.upenn.edu/LDC2019T05

Given a specified discourse model (e.g., RST), how do we automatically assign discourse relations to text?

- **Discourse structure parsing:** Given a sequence of text, automatically determine the coherence relations between spans within it
- Discourse structure parsing can be performed similarly to constituency parsing
    - Break text into meaningful subunits
    - Organize those subunits into a set of directed (and, depending on model type, hierarchical) relations

What does this look like for RST parsing?

- **Step #1: EDU Segmentation**
  - Extract the start and end of each elementary discourse unit

Natalie said there was no class next Thursday because it was Thanksgiving.

↓

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$

EDU Segmentation

- EDUs roughly correspond to clauses
- Early EDU segmentation approaches:
  - Run a syntactic parser
  - Post-process the output
- More modern EDU segmentation approaches:
  - Usually apply supervised neural sequence models

What does this look like for RST parsing?

- **Step #1: EDU Segmentation**
  - Extract the start and end of each elementary discourse unit
- **Step #2: Parsing Algorithm**
  - Build representations for each EDU, and apply some method to connect them using RST relations

RST Parsing

- Generally based on syntactic parsing algorithms
- Common syntactic parsing approach that also works well for discourse parsing: **Shift-reduce parser**
  - **Shift:** Push an EDU from the queue onto the stack, creating a single-node subtree
  - **Reduce:** Merge the top two subtrees (either single-node or more complex) on the stack, assigning a coherence relation label and a nuclearity direction
  - **Pop:** Remove the final tree from the stack

Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$

Queue

Stack

Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$

Queue

[Natalie said]$_{e1}$

[there was no class next Thursday]$_{e2}$

[because it was Thanks-giving.]$_{e3}$

Stack

Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$

Queue

[there was no class next Thursday]$_{e2}$

[because it was Thanks-giving.]$_{e3}$

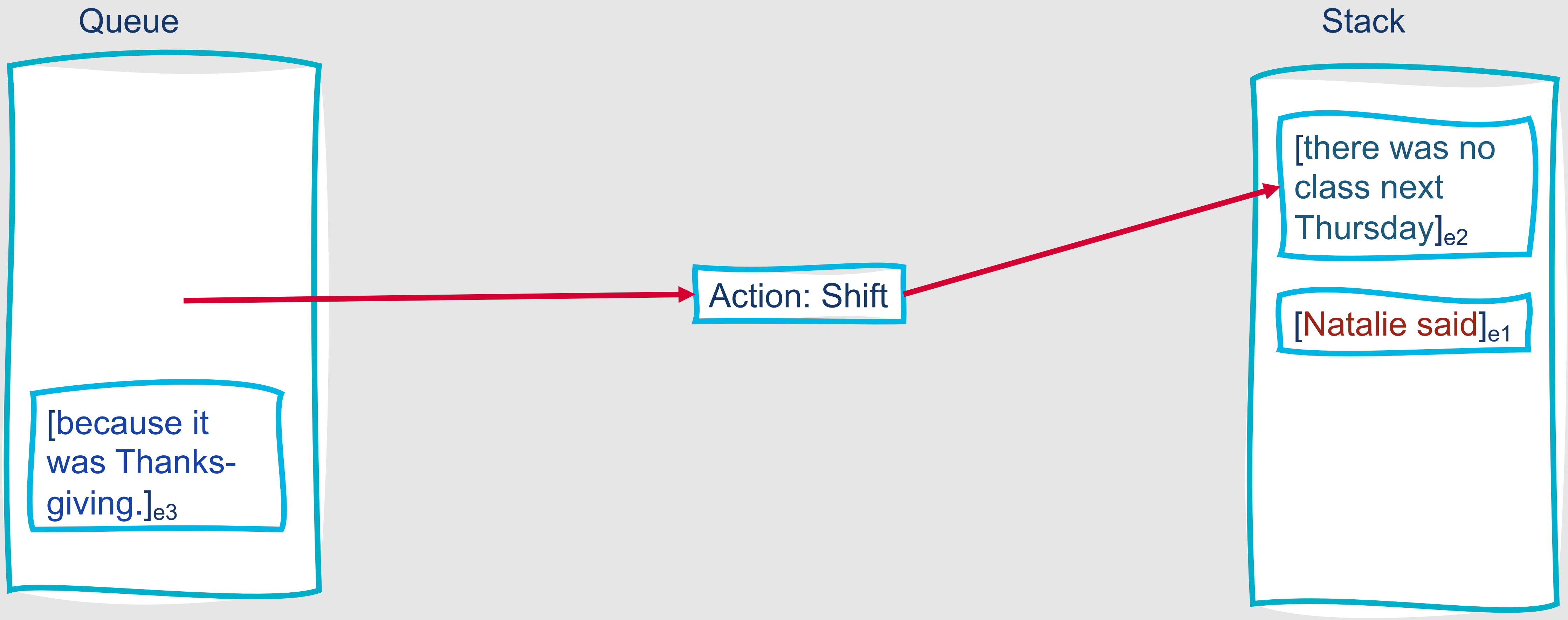
Action: Shift

Stack

[Natalie said]$_{e1}$

Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$

Queue

[because it was Thanks-giving.]$_{e3}$

Action: Reduce(Attribution, (Satellite, Nucleus))

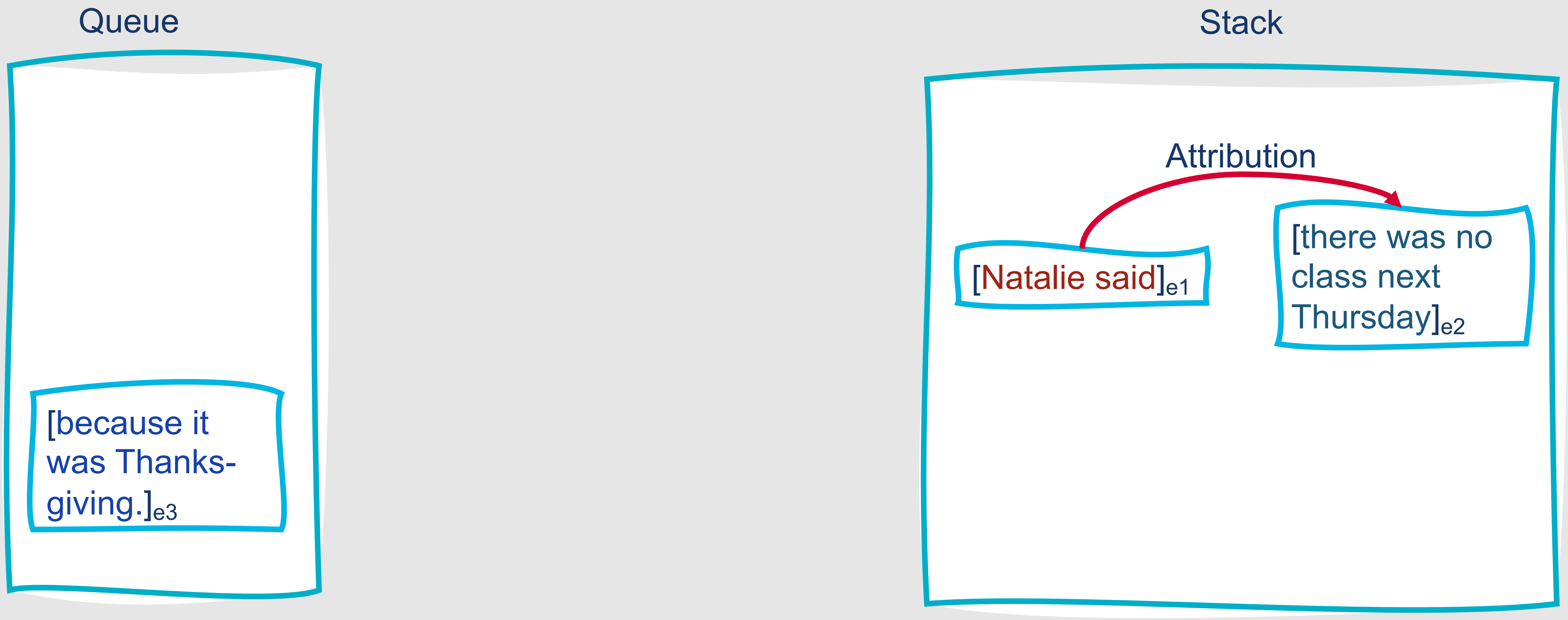
Stack

[there was no class next Thursday]$_{e2}$

[Natalie said]$_{e1}$

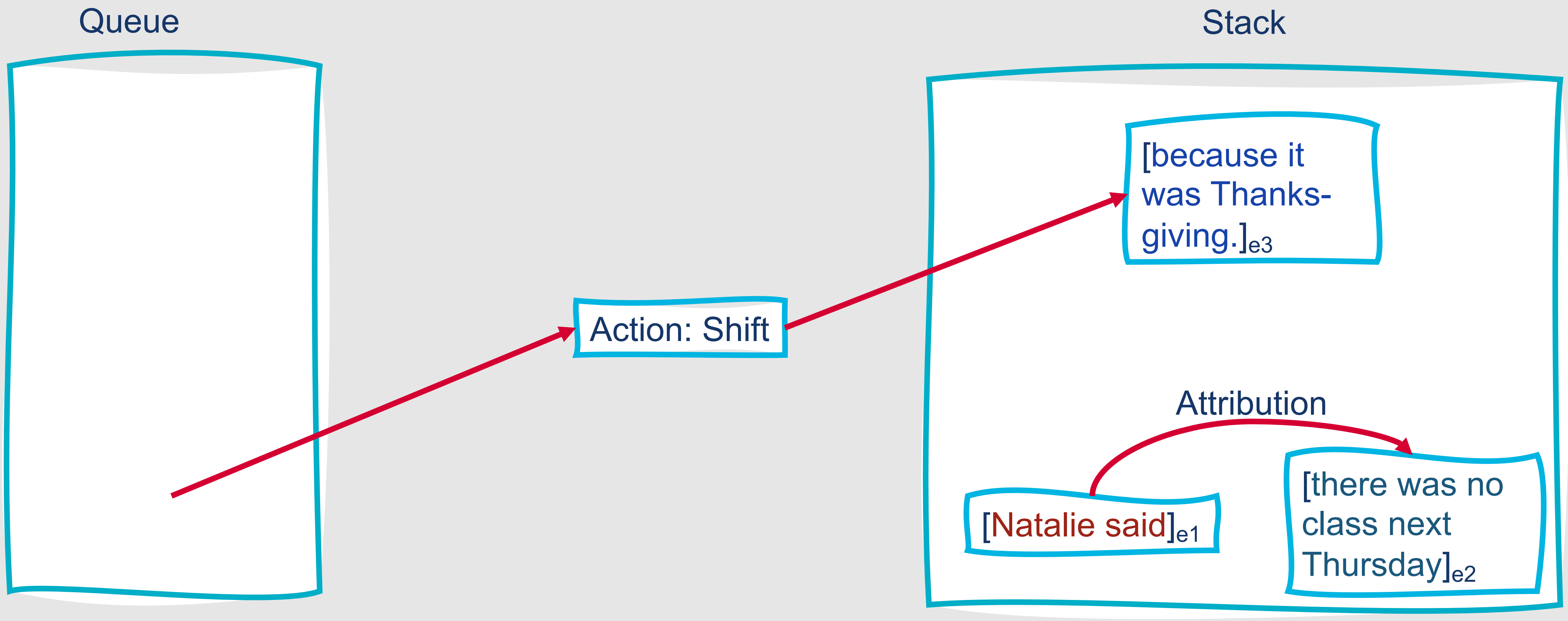
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$



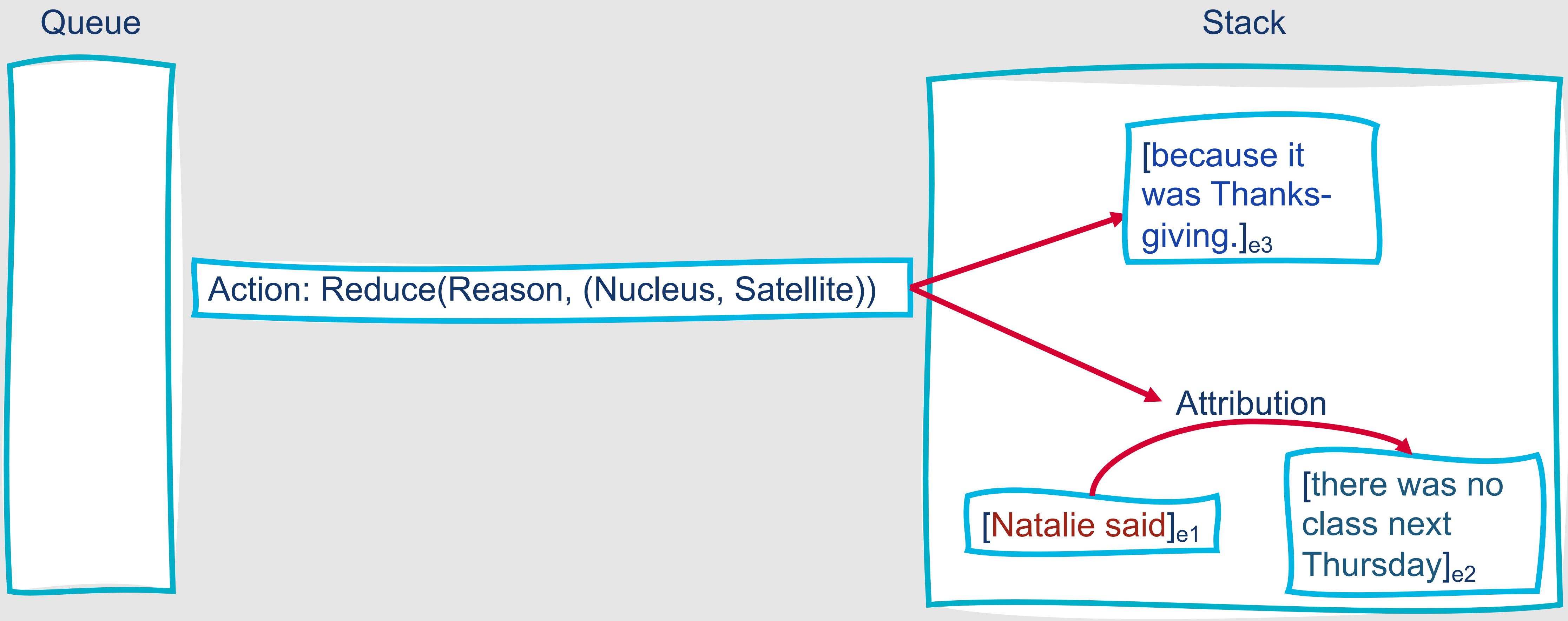
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$



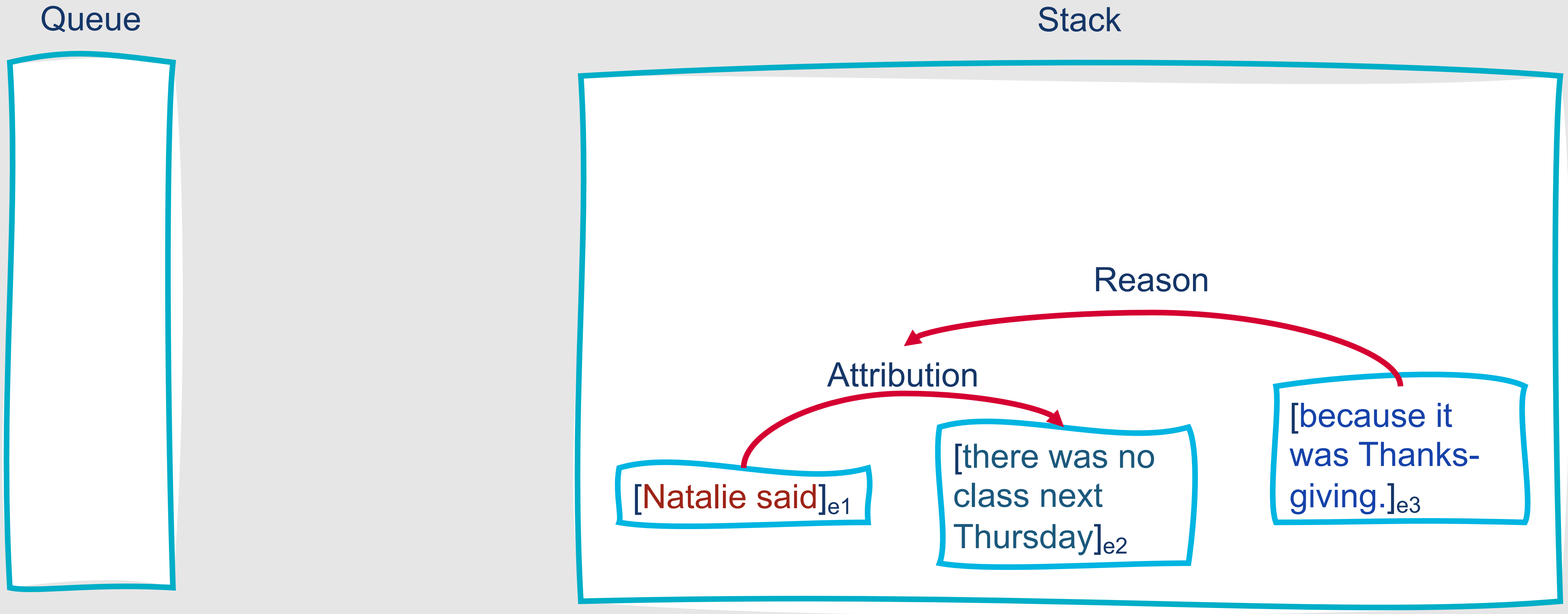
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$



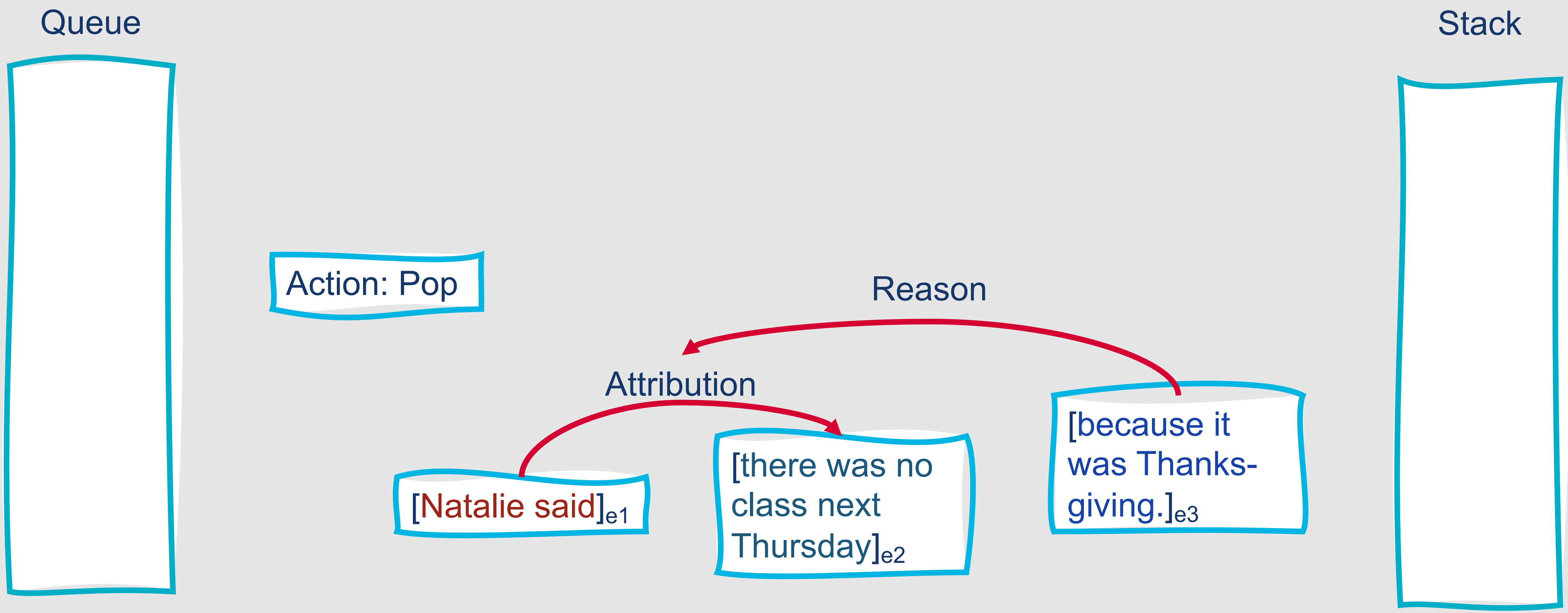
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class next Thursday]$_{e2}$ [because it was Thanksgiving.]$_{e3}$



Modern RST parsers generally select actions using neural networks.

$$\mathbf{o} = \mathbf{W}(\underbrace{\mathbf{h}^{t}_{s0}, \mathbf{h}^{t}_{s1}, \mathbf{h}^{t}_{s2}}, \mathbf{h}^{e}_{q0})$$

Action

Hidden representations of the top three subtrees on the stack

Hidden representation of the first EDU on the queue

How does PDTB discourse parsing differ from this?

- **Shallow discourse parsing:** Identifying relationships between text spans only, rather than full hierarchical discourse trees

This Week's Topics

Coreference Resolution Approaches

Evaluating Coreference Resolution

Discourse Relations

Thursday

Tuesday

Discourse Parsing

Entity-Based Coherence

Topical Salience and Global Coherence

Identifying discourse relations is one way to model discourse coherence….

- Another?
  - Determine **entity salience**

Entity-Based Coherence

- At each point in the discourse, some entity is salient
- A discourse remains coherent by continuing to discuss the salient entity
- Two key models for entity-based coherence:
  - **Centering Theory**
  - **Entity Grid Model**

Centering Theory

- At any point in the discourse, one of the entities in the discourse model is salient (**being “centered” on**)
- Discourses in which adjacent sentences **continue** to maintain the same salient entity are more coherent than those which **shift** back and forth between multiple entities

Centering Theory: Intuition

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

Centering Theory: Intuition

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

Same propositional content, difference entity saliences

Centering Theory: Intuition

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

Much more coherent!

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

How does Centering Theory realize this intuition?

- Maintain two representations for each utterance U_n
  - $C_f(U_n)$: Forward-looking centers of U_n
    - Set of potential future salient entities (potential $C_b(U_{n+1})$)
  - $C_b(U_n)$: Backward-looking center of U_n
    - The highest-ranked element of $C_f(U_{n-1})$ that is realized in U_n
- Set of $C_f(U_n)$ are ranked based on a variety of factors (e.g., grammatical role)
- Highest-ranked $C_f(U_n)$ is the preferred center C_p

There can be four intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

	$C_b(U_{n+1}) = C_b(U_n)$ **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

There can be four intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

The same entity is centered as in the previous utterance, and it is anticipated that this will continue

	$C_b(U_{n+1}) = C_b(U_n)$ **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

There can be four intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

The same centered entity is retained as in the previous utterance, although it is not anticipated that this will continue

	$C_b(U_{n+1}) = C_b(U_n)$ **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

There can be four intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

The center has shifted to a new entity

	$C_b(U_{n+1}) = C_b(U_n)$ **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

Based on these relationships, we can define two rules.

- Centered entities should be realized as pronouns when they are continued
- Transition states are ordered such that 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

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

$C_f(U_1)$: {Natalie, UIC}
$C_p(U_1)$: Natalie
$C_b(U_1)$: undefined

$C_f(U_2)$: {Natalie, UIC, class}
$C_p(U_2)$: Natalie
$C_b(U_2)$: Natalie

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

$C_f(U_1)$: {Natalie, UIC}
$C_p(U_1)$: Natalie
$C_b(U_1)$: undefined

$C_f(U_2)$: {Natalie, UIC, class}
$C_p(U_2)$: Natalie
$C_b(U_2)$: Natalie

	$C_b(U_{n+1}) = C_b(U_n)$ 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

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

$C_f(U_1)$: {Natalie, UIC}
$C_p(U_1)$: Natalie
$C_b(U_1)$: undefined

$C_f(U_2)$: {UIC, class, Natalie}
$C_p(U_2)$: UIC
$C_b(U_2)$: Natalie

	$C_b(U_{n+1}) = C_b(U_n)$ **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

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

☺

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

Entity Grid Model

- Alternative way to capture entity-based coherence
- Learns **patterns of entity mentioning** that can be used to train a supervised learning model to predict coherence
- Based on an **entity grid**
  - Two-dimensional array representing the **distribution of entity mentions across sentences**
    - Rows = sentences
    - Columns = discourse entities
    - Values in cells = Whether the entity appears in the sentence, and its grammatical role (subject, object, neither, or absent)

	Natalie	UIC	class	NLP
S1				
S2				
S3				
S4				

Example: Entity Grid Model

- [Natalie]$_s$ was an assistant professor at [UIC]$_x$.
- [Natalie]$_s$ taught a [class]$_o$ at [UIC]$_x$ called CS 421.
- [Natalie]$_s$ enjoyed teaching the [class]$_x$ and liked [NLP]$_o$ a lot.
- [Natalie]$_s$ was planning to teach the [class]$_x$ once per year.

	Natalie	UIC	class	NLP
S1	S	X	-	-
S2				
S3				
S4				

Example: Entity Grid Model

- **[Natalie]$_s$ was an assistant professor at [UIC]$_x$.**
- [Natalie]$_s$ taught a [class]$_o$ at [UIC]$_x$ called CS 421.
- [Natalie]$_s$ enjoyed teaching the [class]$_x$ and liked [NLP]$_o$ a lot.
- [Natalie]$_s$ was planning to teach the [class]$_x$ once per year.

	Natalie	UIC	class	NLP
S1	S	X	-	-
S2	S	X	O	-
S3				
S4				

Example: Entity Grid Model

- [Natalie]$_s$ was an assistant professor at [UIC]$_x$.
- **[Natalie]$_s$ taught a [class]$_o$ at [UIC]$_x$ called CS 421.**
- [Natalie]$_s$ enjoyed teaching the [class]$_x$ and liked [NLP]$_o$ a lot.
- [Natalie]$_s$ was planning to teach the [class]$_x$ once per year.

	Natalie	UIC	class	NLP
S1	S	X	-	-
S2	S	X	O	-
S3	S	-	X	O
S4				

Example: Entity Grid Model

- [Natalie]$_s$ was an assistant professor at [UIC]$_x$.
- [Natalie]$_s$ taught a [class]$_o$ at [UIC]$_x$ called CS 421.
- **[Natalie]$_s$ enjoyed teaching the [class]$_x$ and liked [NLP]$_o$ a lot.**
- [Natalie]$_s$ was planning to teach the [class]$_x$ once per year.

	Natalie	UIC	class	NLP
S1	S	X	-	-
S2	S	X	O	-
S3	S	-	X	O
S4	S	-	X	-

Example: Entity Grid Model

- [Natalie]$_s$ was an assistant professor at [UIC]$_x$.
- [Natalie]$_s$ taught a [class]$_o$ at [UIC]$_x$ called CS 421.
- [Natalie]$_s$ enjoyed teaching the [class]$_x$ and liked [NLP]$_o$ a lot.
- **[Natalie]$_s$ was planning to teach the [class]$_x$ once per year.**

Entity Grid Model

- Dense columns indicate entities mentioned often
- Sparse columns indicate entities mentioned rarely
- Coherence is thus measured by patterns of **local entity transition**
- Each transition ends up with a probability

	Natalie	UIC	class	NLP
S1	s	x	-	-
S2	s	x	o	-
S3	s	-	x	o
S4	s	-	x	-

{x, x, -, -}

Example: Entity Grid Model

	Natalie	UIC	class	NLP
S1	s	x	-	-
S2	s	x	o	-
S3	s	-	x	o
S4	s	-	x	-

Example: Entity Grid Model

{x, x, -, -}

$$p(\{x, x, -, -\}) = \frac{1}{4}$$

	Natalie	UIC	class	NLP
S1	S	X	-	-
S2	S	X	O	-
S3	S	-	X	O
S4	S	-	X	-

Example: Entity Grid Model

{-, o}

$$\boxed{p(\{-, o\}) = \frac{2}{12} = \frac{1}{6}}$$

Entity Grid Model

- These transitions and their probabilities can be used as features for a machine learning model that is trained to predict coherence scores
- These models can be trained in a **self-supervised** manner:
  - Learn to distinguish the natural order of sentences in a discourse (expected to be coherent) from a modified order (e.g., randomized order)

How do we evaluate entity-based coherence models?

- Best option: Compare human coherence ratings with predicted coherence ratings from the model
- However, collecting human labels is expensive!
- Alternate option:
  - Similar strategy to self-supervised training process
  - Evaluate the frequency with which the model predicts the naturally-occurring document to be more coherent than other randomized or otherwise perturbed version(s)

This Week's Topics

Tuesday

- Coreference Resolution Approaches
- Evaluating Coreference Resolution
- Discourse Relations

Thursday

- Discourse Parsing
- Entity-Based Coherence
- Topical Salience and Global Coherence

We’ve talked about identifying coherence relations and entity salience …what about topical salience?

- Discourses are more coherent when they discuss a consistent set of topics
- This can be modeled using measures of **lexical cohesion**
  - **Lexical cohesion:** The sharing of identical or semantically-related words across nearby sentences

Latent Semantic Analysis (LSA)

- Early model of lexical cohesion
  - Still used by many humanities and social science researchers
- First approach using word embeddings for measuring cohesion
- Models the coherence between two sentences *i* and *j* as the cosine between their embedding vectors (traditionally, dimensionality-reduced TF*IDF vectors)
  - $\text{sim}(i, j) = \cos(i, j) = \cos(\sum_{w \in i} \mathbf{w}, \sum_{w \in j} \mathbf{w})$
- The overall coherence of a text is thus the average similarity over all pairs of adjacent sentences s_i and s_{i+1}
  - $\text{coherence}(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} \text{sim}(s_i, s_{i+1})$

More modern models make use of this intuition as well.

- **Local coherence discriminator (LCD)**
  - Computes the coherence of a text as the average of coherence scores between adjacent sentences
  - Learns to discriminate between naturally-occurring adjacent sentences and those in a messed-up order using a self-supervised neural model

Coherence relations, entity salience, and topical salience all focus on local coherence.

- However, discourses must be globally coherent as well!
  - Stories have an overall narrative structure
  - Persuasive essays follow specific argument structure
  - Scientific papers are characterized by a structure common across research publications

Argumentation Structure

- **Argumentation mining:** The computational analysis of rhetorical strategy
- Persuasive arguments generally contain well-defined argumentative components:
  - **Claim:** The central, controversial component of the argument
  - **Premise:** A persuasive support or attack of the claim or another premise

Example: Argumentation Structure

CS 421 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers lectures on a variety of core techniques and NLP application areas. This mix is nice because you can learn fundamental principles but also get up to speed on how they are used.

Example: Argumentation Structure

CS 421 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers lectures on a variety of core techniques and NLP application areas. This mix is nice because you can learn fundamental principles but also get up to speed on how they are used.

Claim

Example: Argumentation Structure

CS 421 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers lectures on a variety of core techniques and NLP application areas. This mix is nice because you can learn fundamental principles but also get up to speed on how they are used.

Claim

Premises supporting the claim

Example: Argumentation Structure

CS 421 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers lectures on a variety of core techniques and NLP application areas. This mix is nice because you can learn fundamental principles but also get up to speed on how they are used.

Claim

Premises supporting the claim

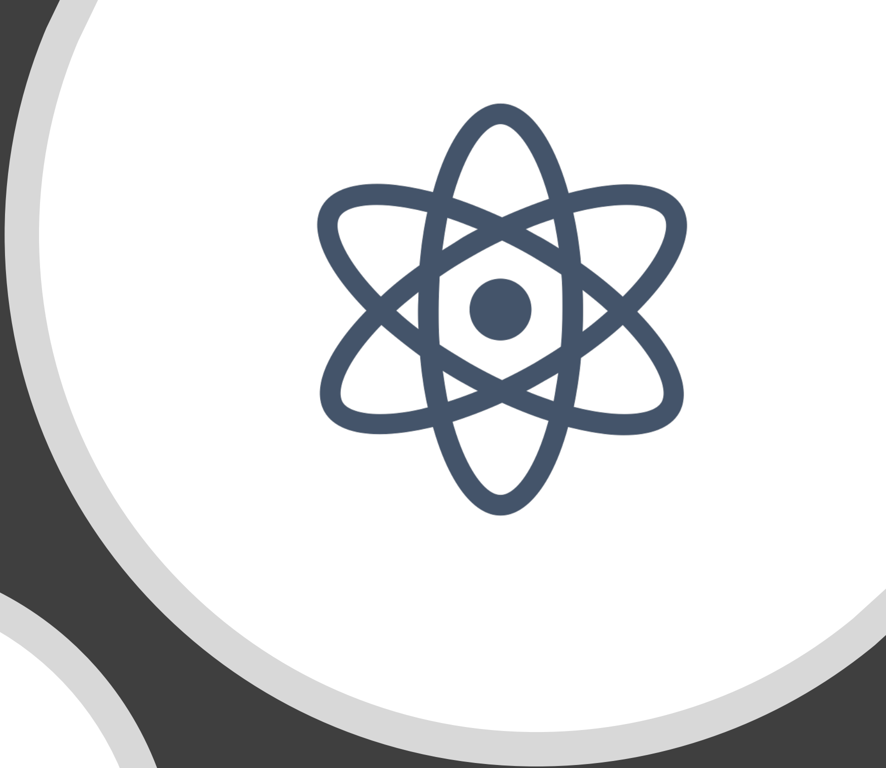
Premise supporting the second premise

How can we detect argumentation structure?

- Classifiers to identify claims, premises, and non-argumentation
- Methods to detect specific argumentation schemes
  - For example:
    - Argument from example
    - Argument from cause to effect
    - Argument from consequences
- Related research: Studying how components of argument structure are associated with persuasive success

We can apply similar methods to scientific discourse!

- In scientific papers, authors need to:
  - Indicate a scientific goal
  - Develop a method for reaching that goal
  - Provide evidence for the solution
  - Compare to prior work
- Parallel to argumentation structure: Each paper tries to make a **knowledge claim**!
- Modeling scientific discourse is an active research problem, as is modeling other global discourse structures (e.g., stories)



Summary: Discourse Coherence

- Common models of discourse relation include **Rhetorical Structure Theory** and the **Penn Discourse Treebank**
- **Discourse parsing** can be performed using techniques that are also common for other structured language parsing tasks
- **Entity salience** can be modeled using **Centering Theory** or the **Entity Grid Model**
- **Lexical cohesion** may be measured using **latent semantic analysis** or other word embedding-based methods
- **Argumentation structure** captures **global coherence,** and may be applied to a variety of domains including persuasive essays and scientific discourse