Discourse Coherence and Question Answering

Natalie Parde

UIC CS 421

This Week's Topics

Tuesday

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

Thursday

- Classic QA
- IR- and Knowledge-Based QA
- Evaluating QA Systems

This Week's Topics

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

Tuesday

Thursday

Classic QA
IR- and Knowledge-Based QA
Evaluating QA Systems

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 2 was due the following 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.

This Week's Topics

Discourse Relations

Discourse Parsing

Entity-Based Coherence

Topical Salience and Global Coherence

Thursday

Tuesday

Classic QA

IR- and Knowledge-Based QA

Evaluating QA Systems

RST relations can be hierarchically organized into discourse trees.

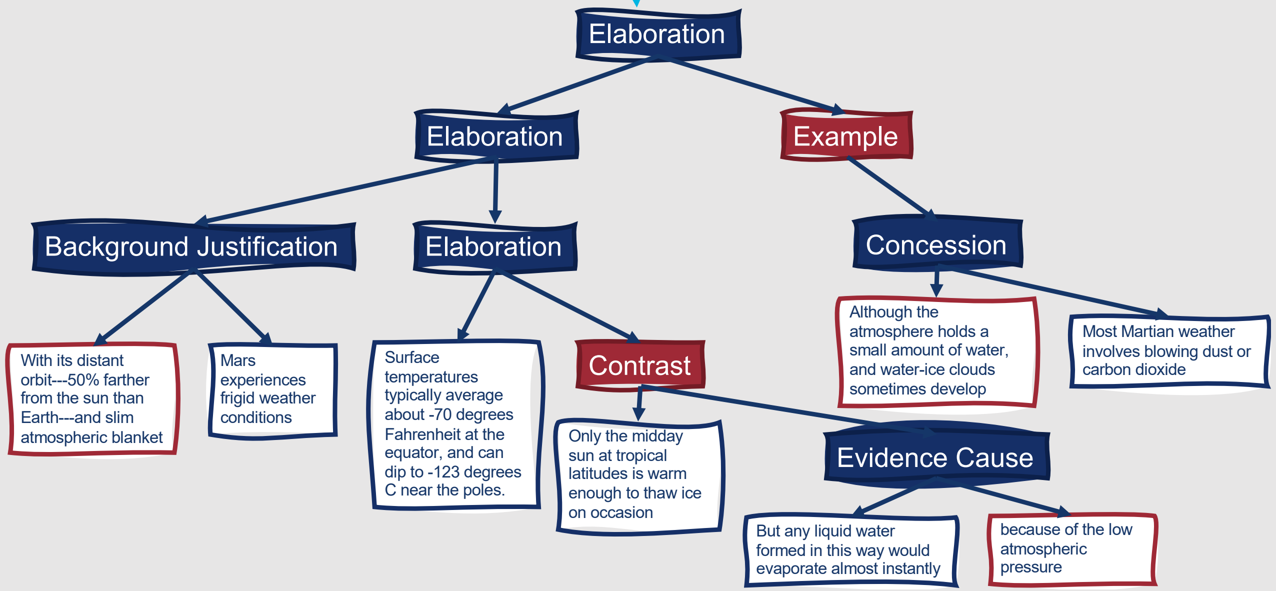
With its distant orbit-–50% farther from the sun than Earth-–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees C near the poles.

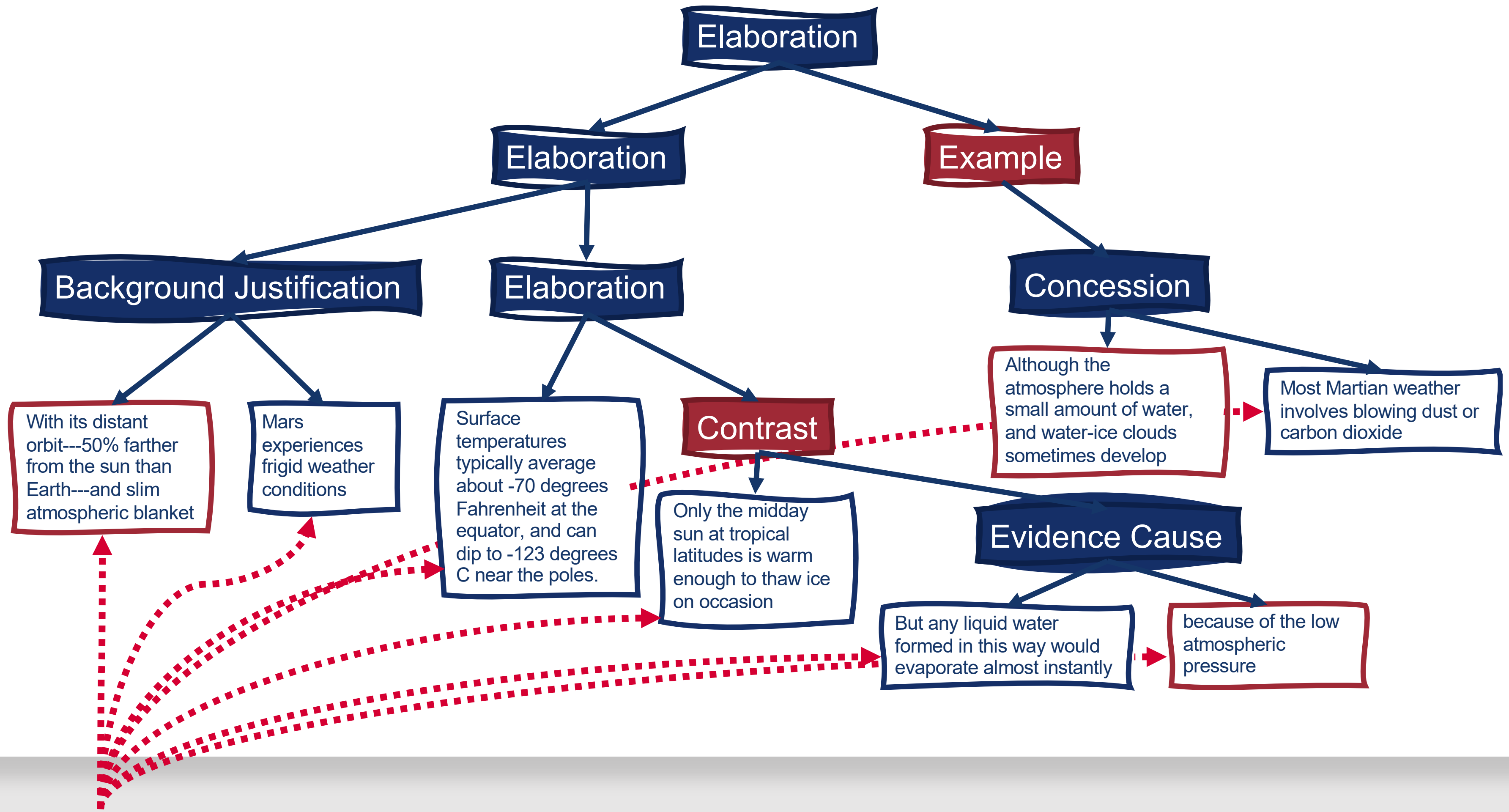
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian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

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$

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 on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$

Queue

Stack

Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$

Queue

[Natalie said]$_{e1}$

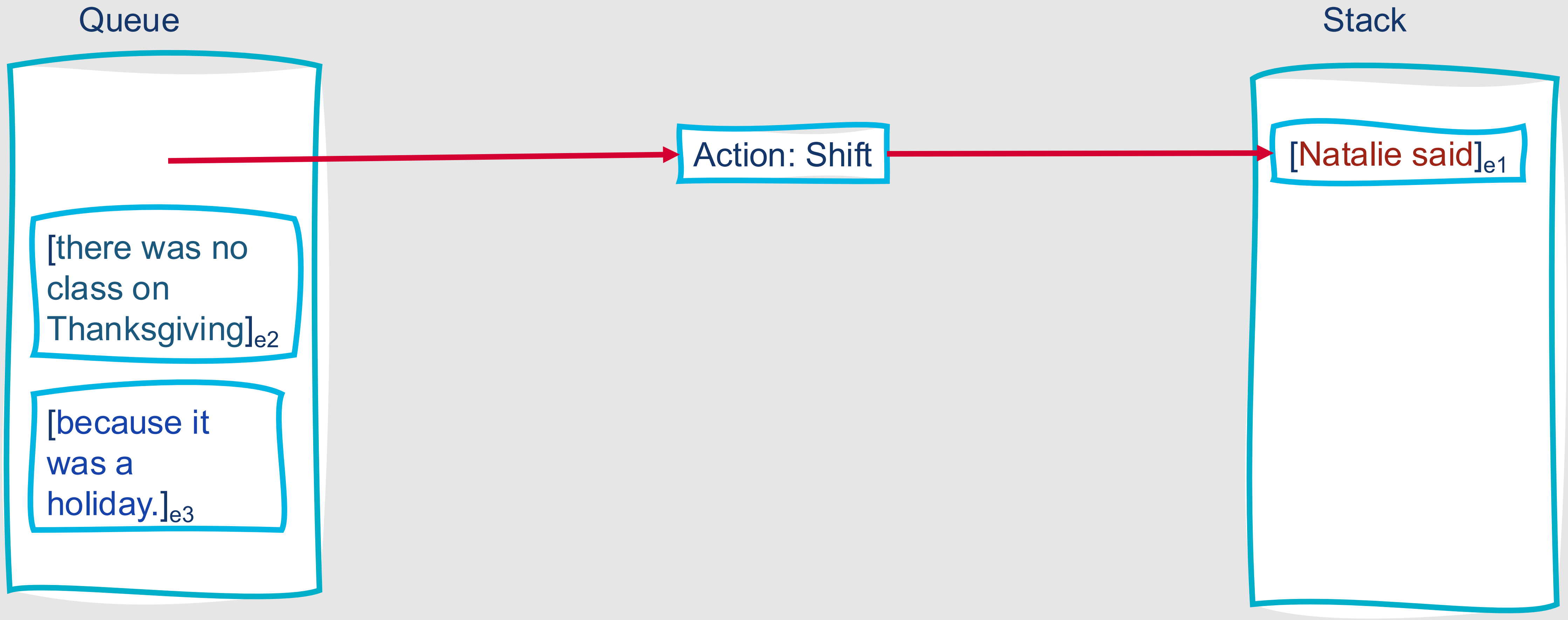
[there was no class on Thanksgiving]$_{e2}$

[because it was a holiday.]$_{e3}$

Stack

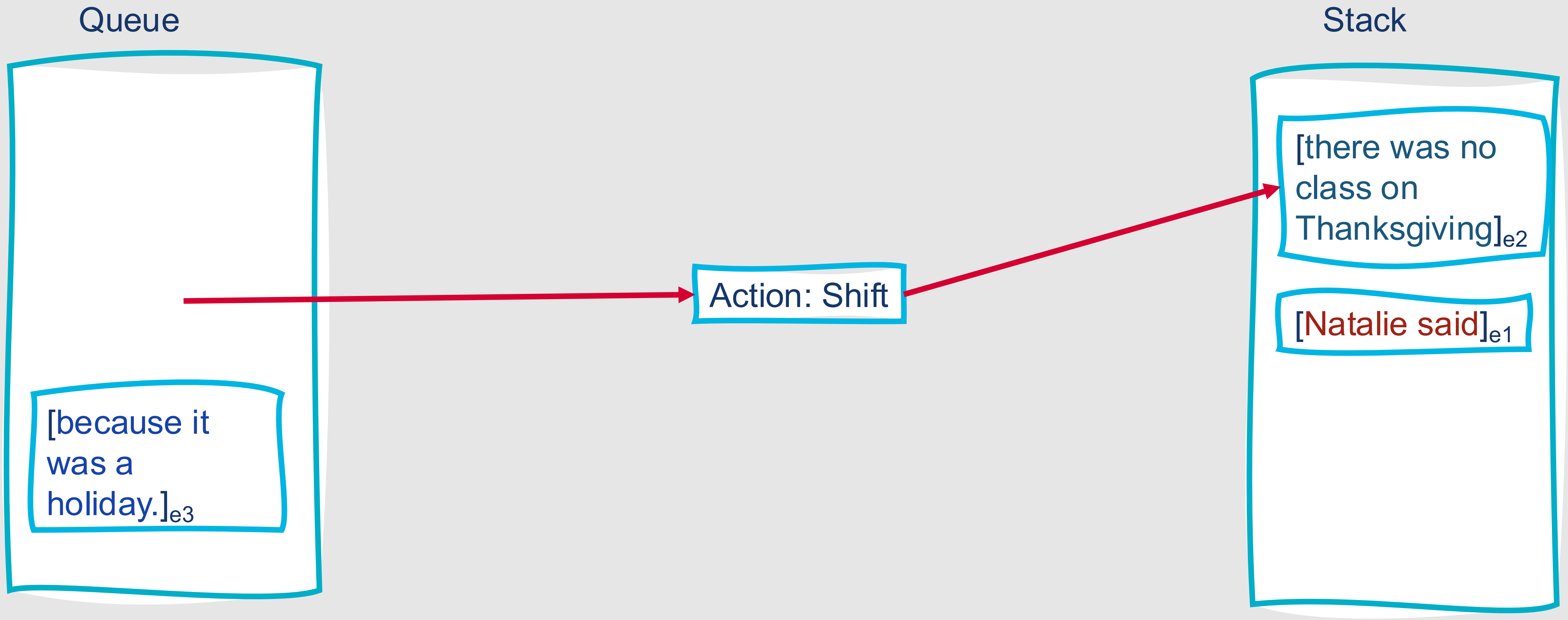
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$

Queue

[because it was a holiday.]$_{e3}$

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

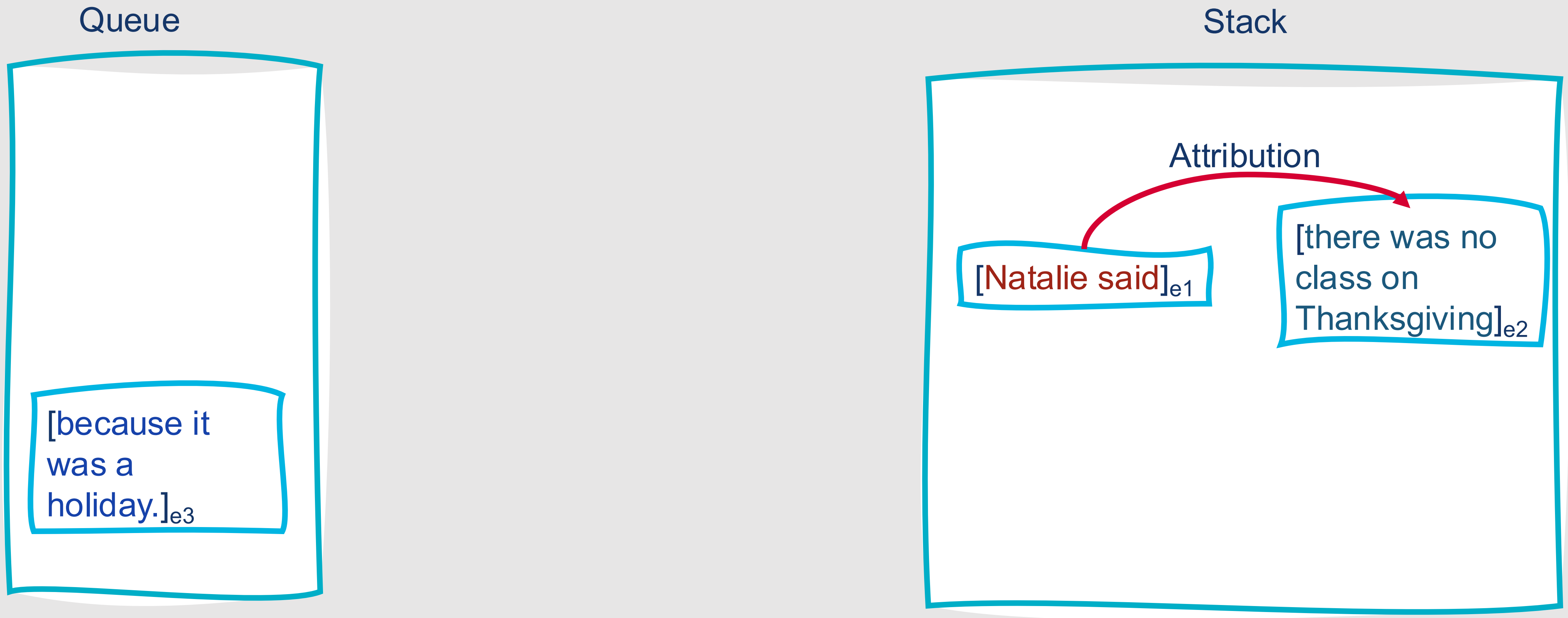
Stack

[there was no class on Thanksgiving]$_{e2}$

[Natalie said]$_{e1}$

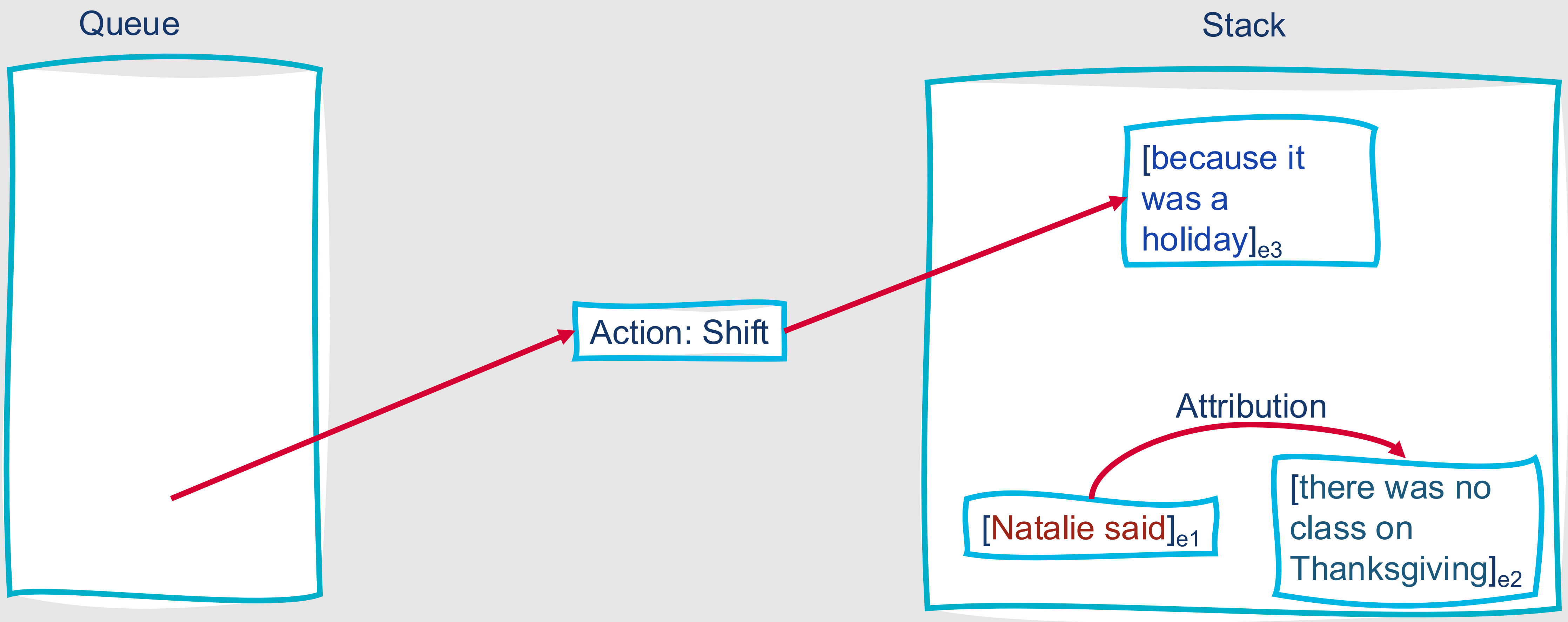
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$

Queue

Stack

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

[because it was a holiday.]$_{e3}$

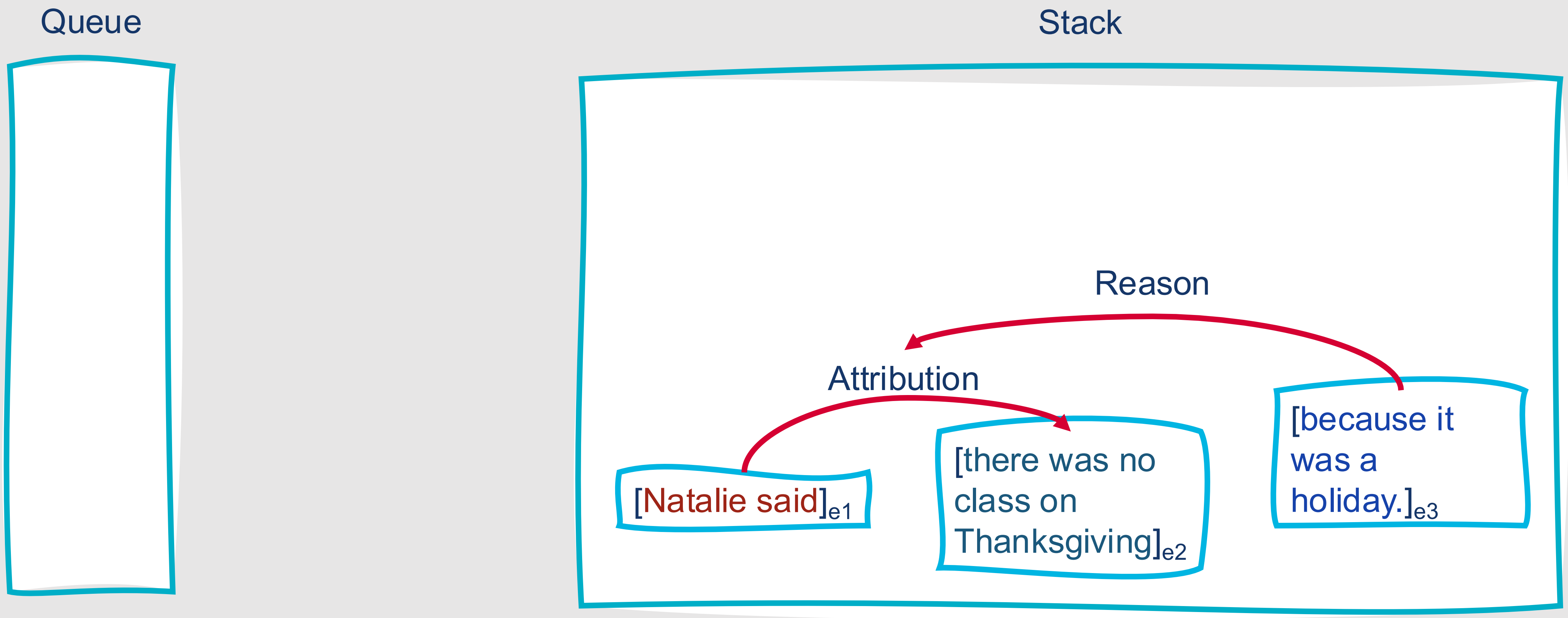
Attribution

[Natalie said]$_{e1}$

[there was no class on Thanksgiving]$_{e2}$

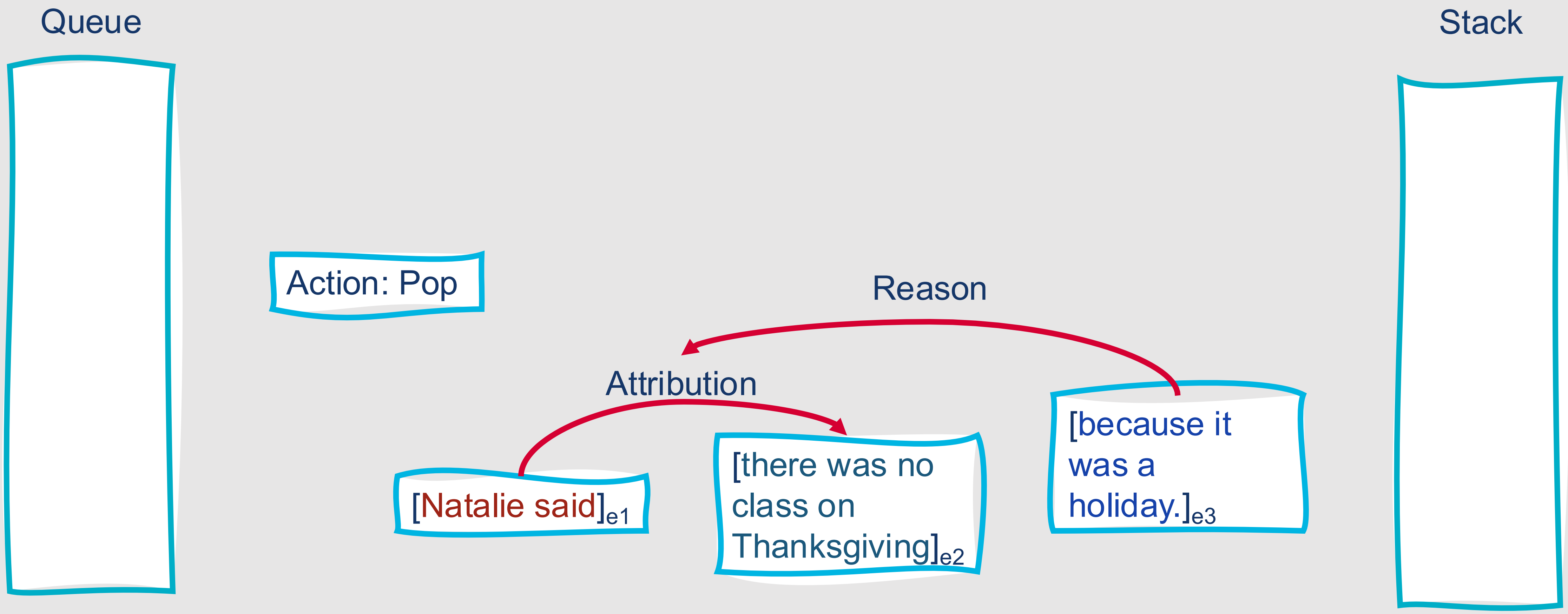
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$



Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no class on Thanksgiving]$_{e2}$ [because it was a holiday.]$_{e3}$



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

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

How does PDTB discourse parsing differ from this?

This Week's Topics

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

Thursday

Tuesday

Classic QA
IR- and Knowledge-Based QA
Evaluating QA Systems

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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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 associate 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	-

{-, o}

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

Example: Entity Grid Model

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

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

Thursday

- Classic QA
- IR- and Knowledge-Based QA
- Evaluating QA Systems

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)
  - $\mathrm{sim}(i,j) = \cos(i,j) = \cos(\sum_{w \in i} \mathbf{w}, \sum_{w \in j} \mathbf{w})$
- The overall coherence of a text is thus the average similarity over all pairs of adjacent sentences s_i and s_{i+1}
  - $\mathrm{coherence}(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathrm{sim}(s_i, s_{i+1})$

Other 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 perturbed 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

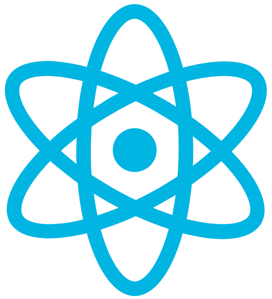
Methods to detect specific argumentation schemes

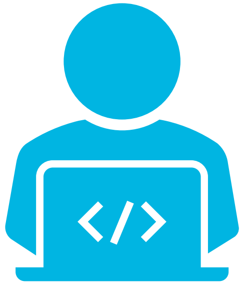
- For example:
  - Argument from example
  - Argument from cause to effect
  - Argument from consequences

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)





Discourse Coherence in Large Language Models

Earlier strategies for enforcing coherence in NLP applications focused on conceptual models (e.g., Rhetorical Structure Theory) or structured measures of entity continuity or lexical cohesion

Modern LLMs are known to produce highly fluent text …but what about coherent discourses?

Discourse (In)coherence in LLMs

- LLMs are great at producing locally fluent discourses due to their focus on next-token prediction as a pretraining task
- However, they often struggle with:
  - **Local topic drift:** Gradual semantic shifts without connective cues
  - **Entity degradation:** Reference errors and forgotten entities
  - **Logical inconsistency:** Reversed or missing causes and effects
  - **Circularity or contradiction:** Forgetfulness of earlier discourse commitments

Why does this happen, and how can we address it?

Why it happens:

- Next-token prediction only weakly enforces long-range dependencies, which in turn hinders long-term monitoring of entity salience
- Random sampling can increase the likelihood of topic shifts

How we can address it:

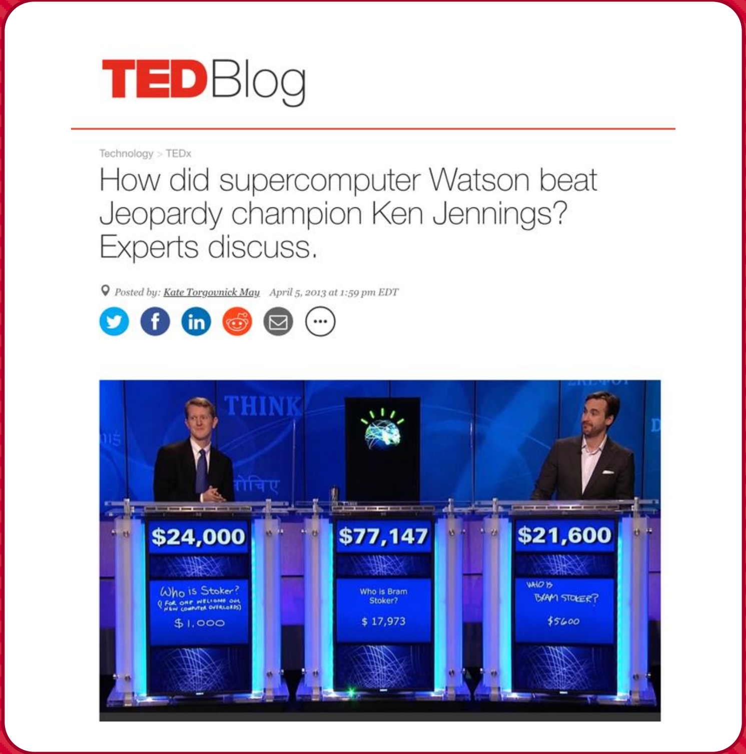
- Incorporate discourse planning (e.g., using clear RST structures) into the generation process
- Perform coherence-aware scoring of hypothesized outputs (e.g., adding coherence scoring into beam search)
- Adding sentence ordering as a pretraining objective
- Evaluating LLMs on discourse coherence datasets

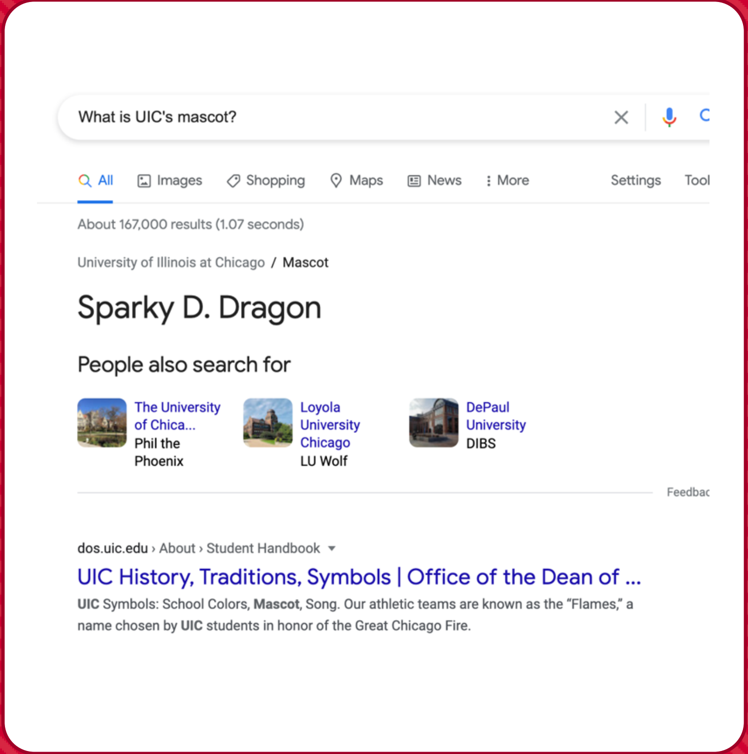
Summary: Discourse Coherence

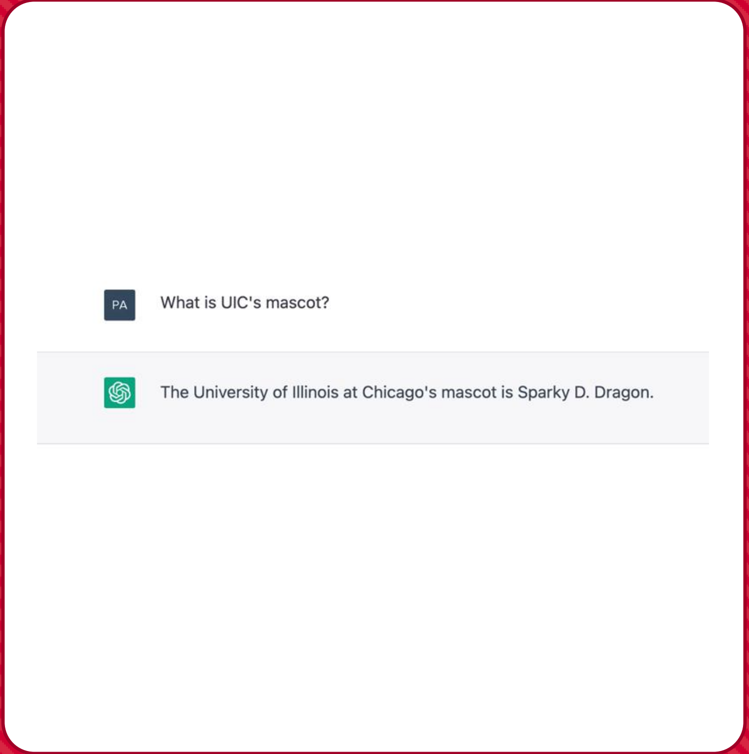
- **Discourse coherence** is the relationship (or lack thereof) between sentences in a discourse
- It is influenced by a variety of factors:
  - **Coherence relations**
  - **Entity salience**
  - **Topical salience**
  - **Global structure**
- Common models of discourse relation include **Rhetorical 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

What is question answering?

- The process of **automatically retrieving** correct, concise, and relevant **information** in response to a user's **query**

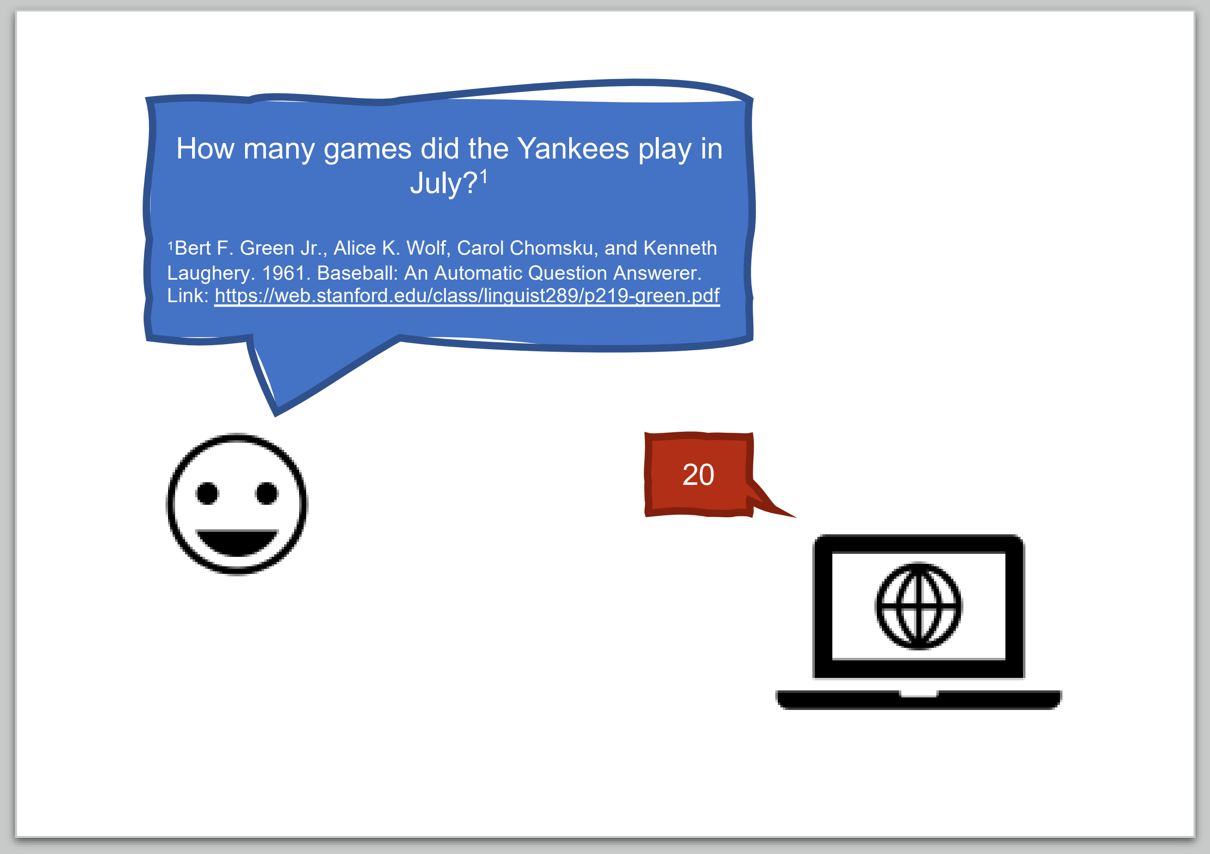


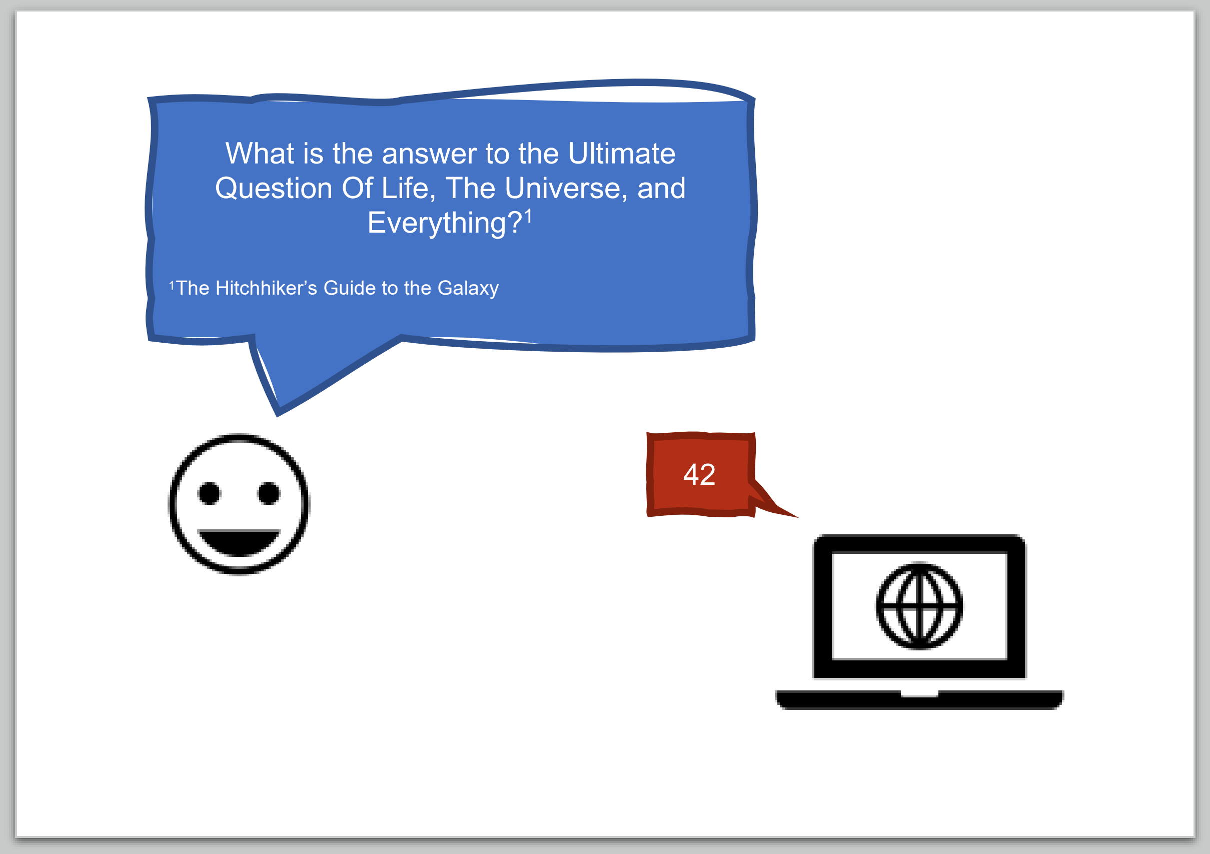




Question answering is an important component for most recent AI systems!

It’s also been a topic of interest for nearly as long as computers have existed.





Question Answering Systems

- Most common focus: **Factoid Questions**
  - Questions that can be answered with simple facts expressed in short texts

When was UIC founded?

How far is UIC from the University of Chicago?

What is the average CS class size?

Question Answering Systems

- Up until recently, QA systems operated under two paradigms:
  - **Information retrieval-based** question answering
  - **Knowledge-based** question answering
- More recently, we've seen many systems using:
  - **Language model-based** question answering
- Further back in time, we also saw:
  - **Classic rule- or feature-based** question answering

This Week's Topics

Tuesday

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

Thursday

- Classic QA
- IR- and Knowledge-Based QA
- Evaluating QA Systems

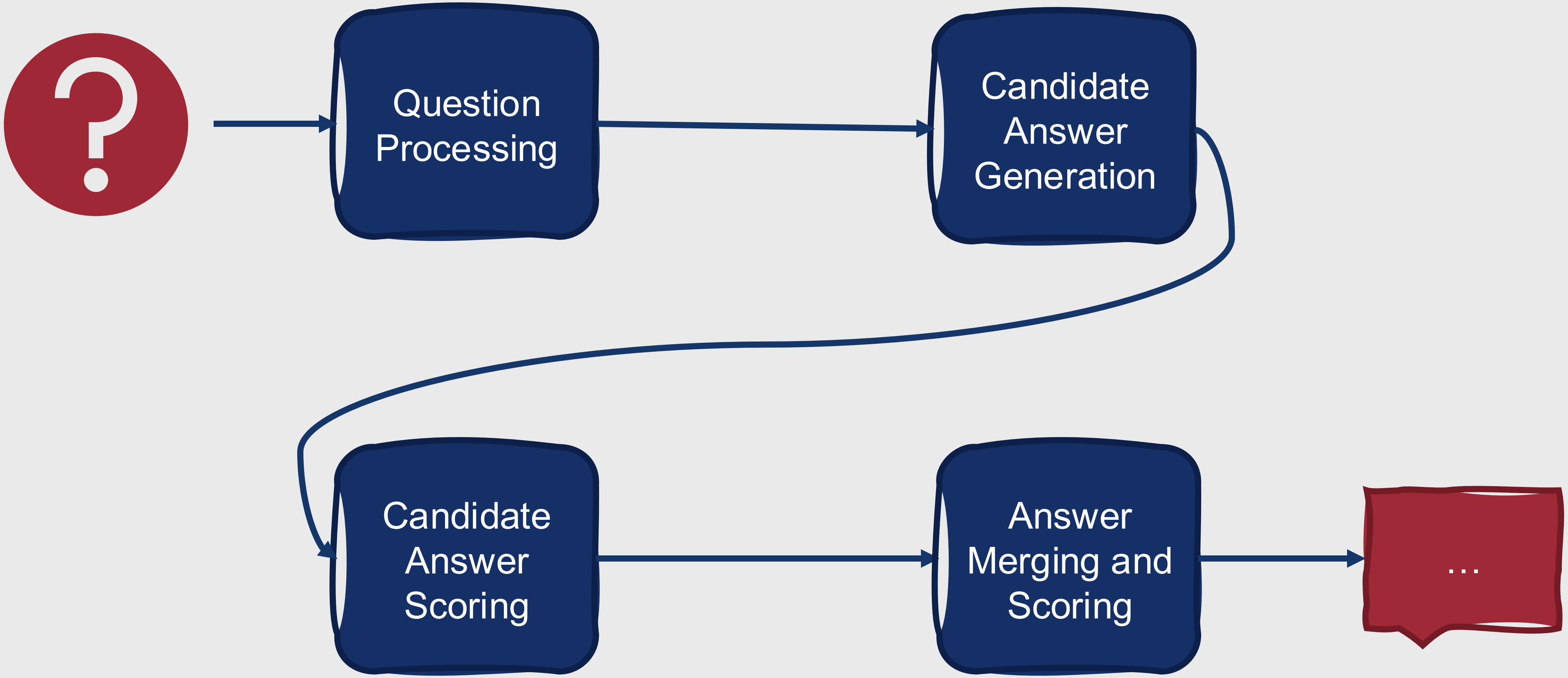
How did classical QA work?

- Rule-based question answering
- Feature-based question answering
- Hybrid techniques that incorporated both approaches

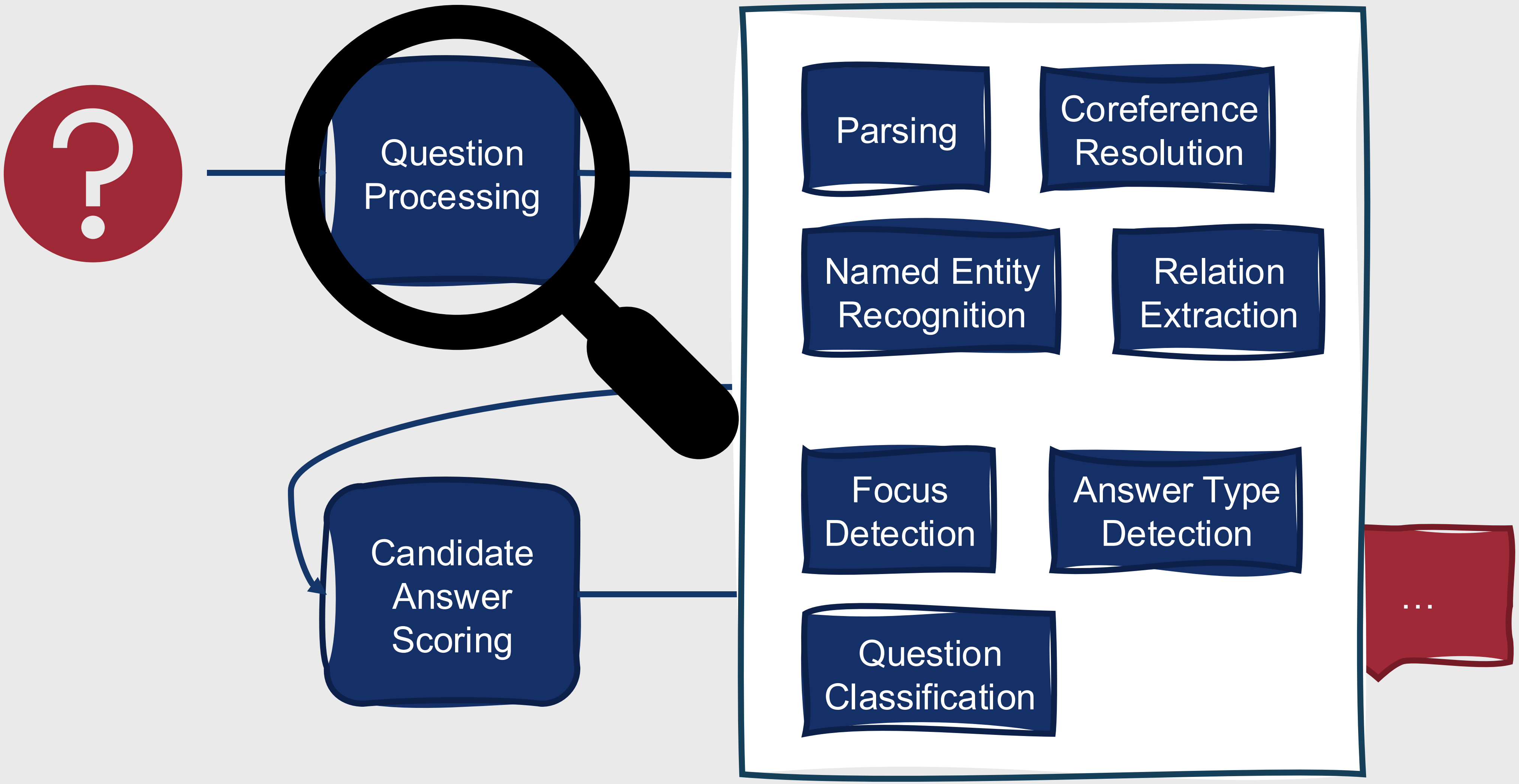
Case Example: DeepQA

- Question answering component of Watson
- Four stages:
  1. **Question processing**
  2. **Candidate answer generation**
  3. **Candidate answer scoring**
  4. **Answer merging and scoring**

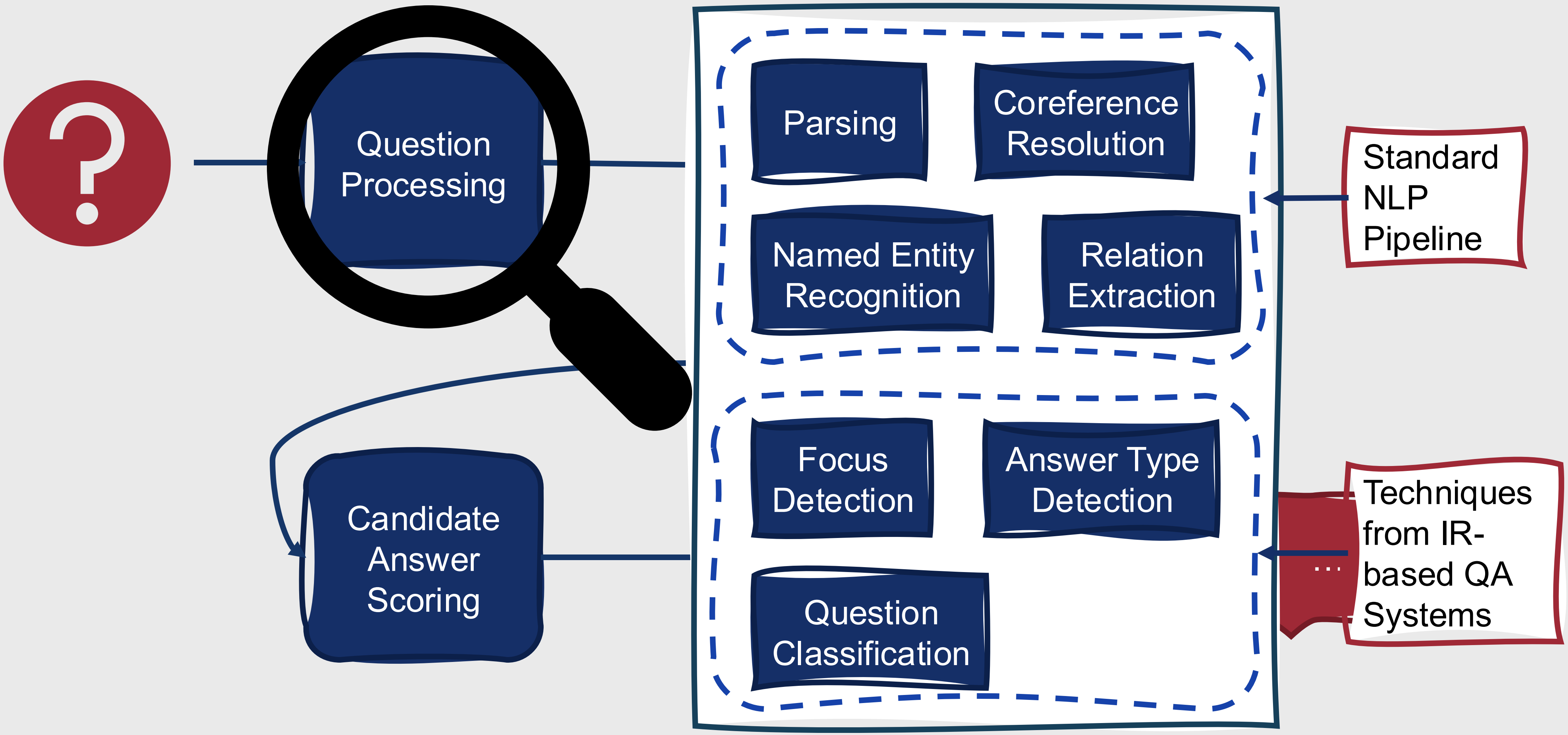
Case Example: DeepQA



Stage 1: Question Preprocessing



Stage 1: Question Preprocessing



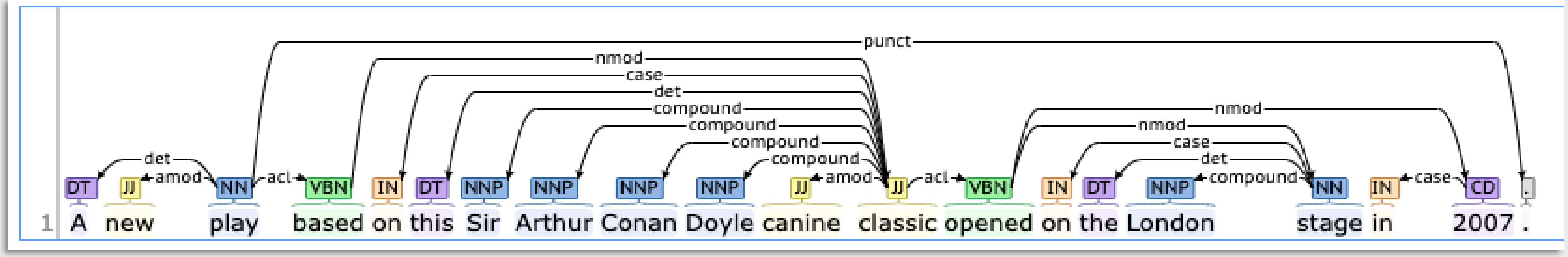
Stage 1: Question Preprocessing

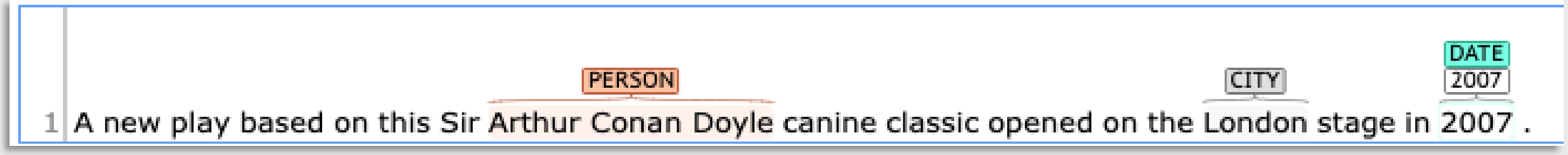
Jeopardy! Example:
A new play based on this Sir Arthur Conan Doyle canine classic opened on the London stage in 2007.

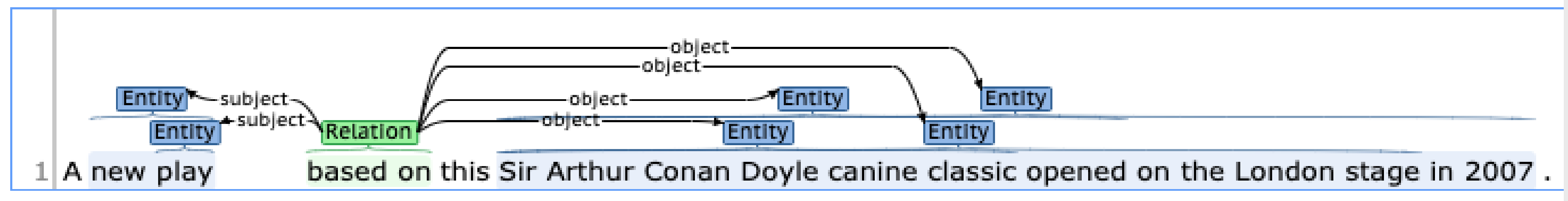
Stage 1: Question Preprocessing

Jeopardy! Example:

A new play based on this Sir Arthur Conan Doyle canine classic opened on the London stage in 2007.

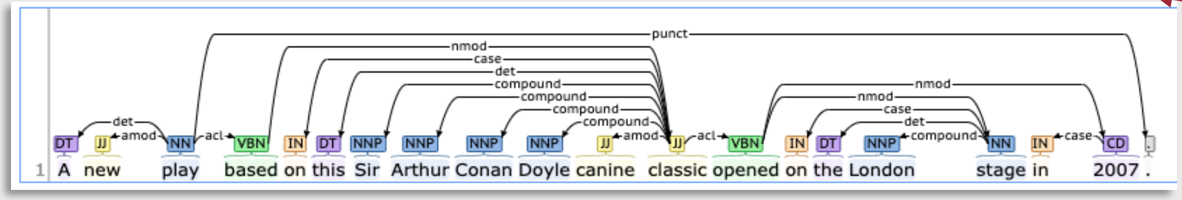


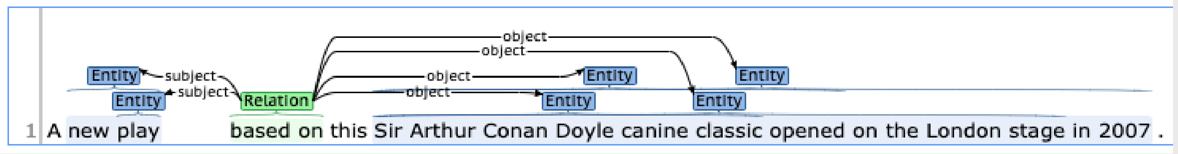




Stage 1: Question Preprocessing

Jeopardy! Example:
A new play based on **this Sir Arthur Conan Doyle canine classic** opened on the London stage in 2007.





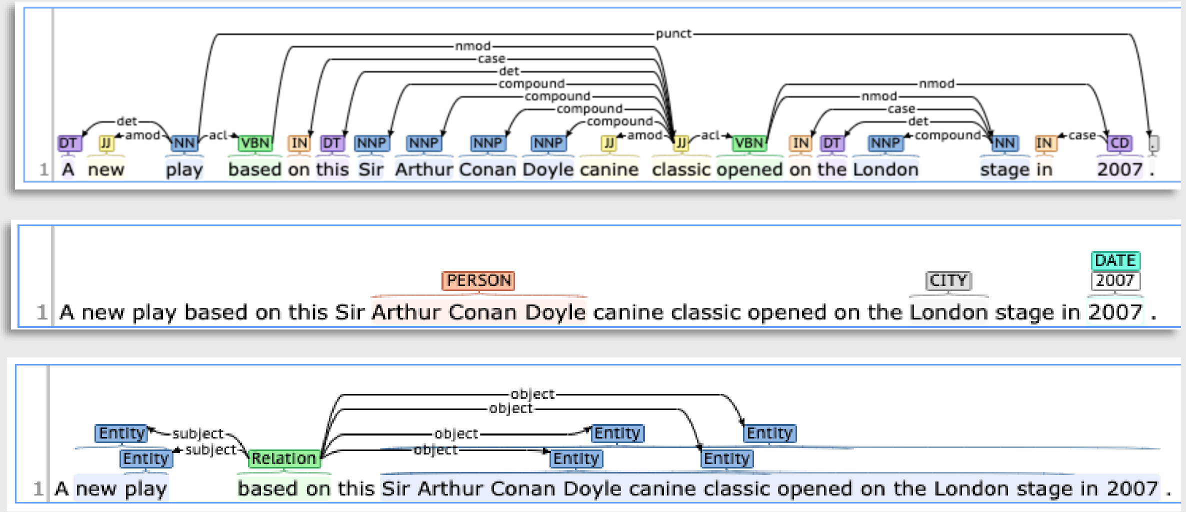
Focus Detection: Which part of the question co-refers with the answer?

Extracted using handwritten rules in DeepQA

Stage 1: Question Preprocessing

Jeopardy! Example:

A new play based on **this Sir Arthur Conan Doyle canine classic** opened on the London stage in 2007.



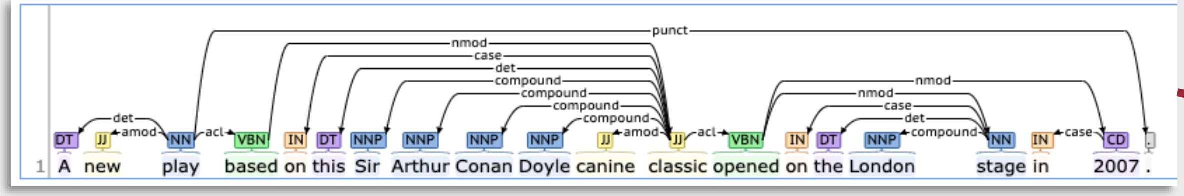
Answer Type Detection: Which word tells us about the semantic type of answer to expect?

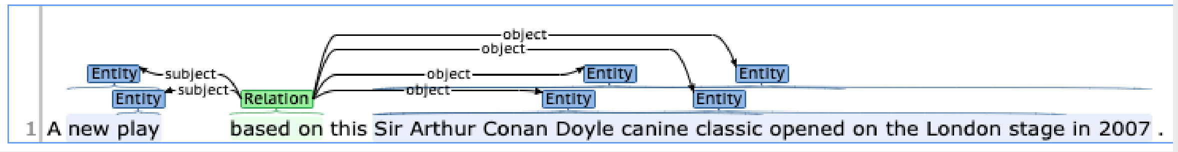
DeepQA extracts roughly 5000 possible answer types (some questions may take multiple answer types), using a rule-based approach

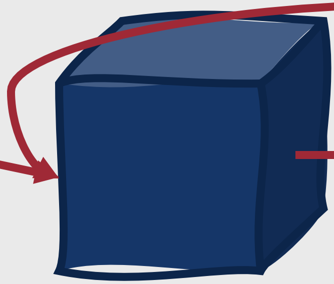
Stage 1: Question Preprocessing

Jeopardy! Example:

A new play based on **this Sir Arthur Conan Doyle canine classic** opened on the London stage in 2007.





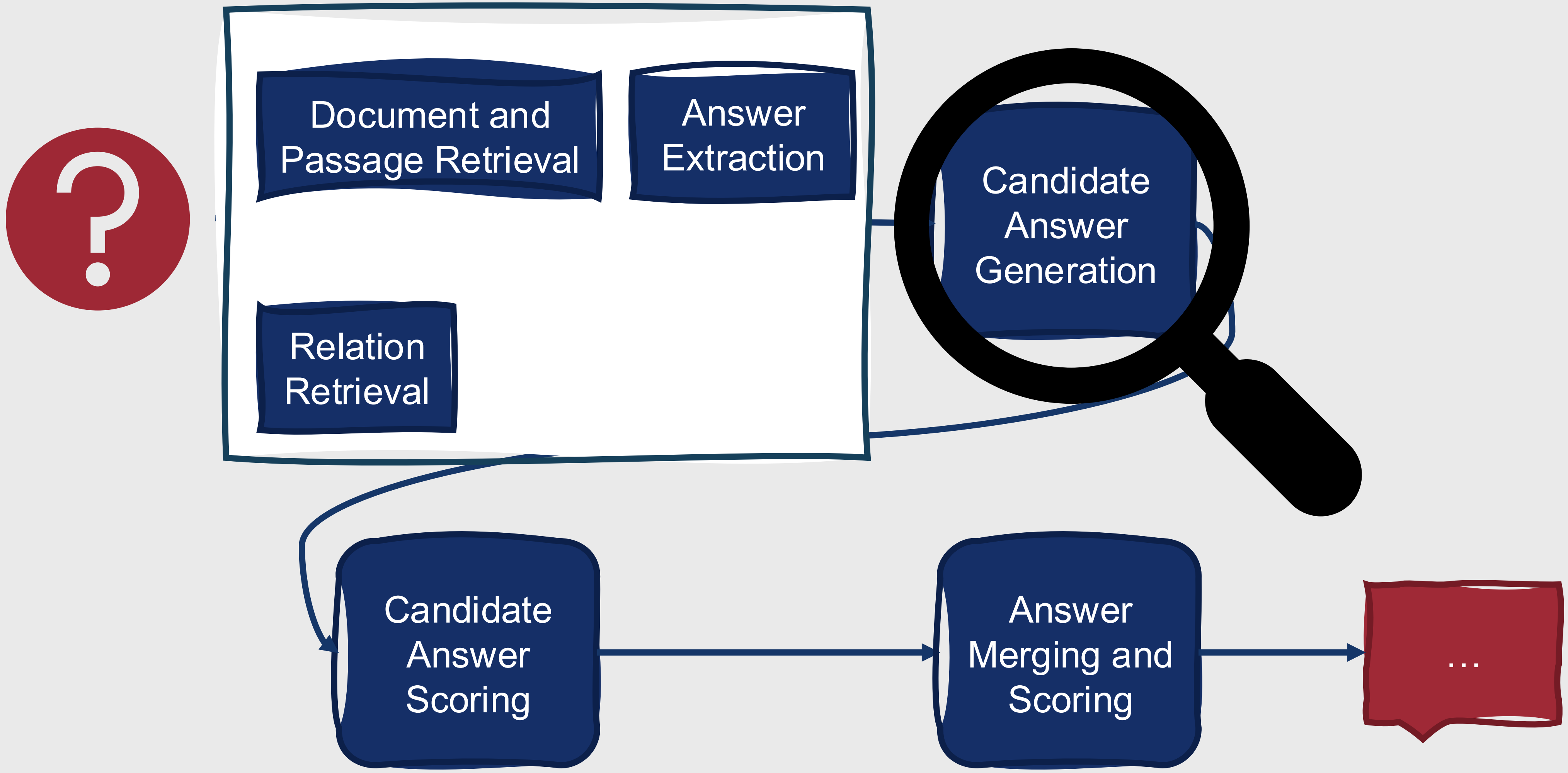


Definition

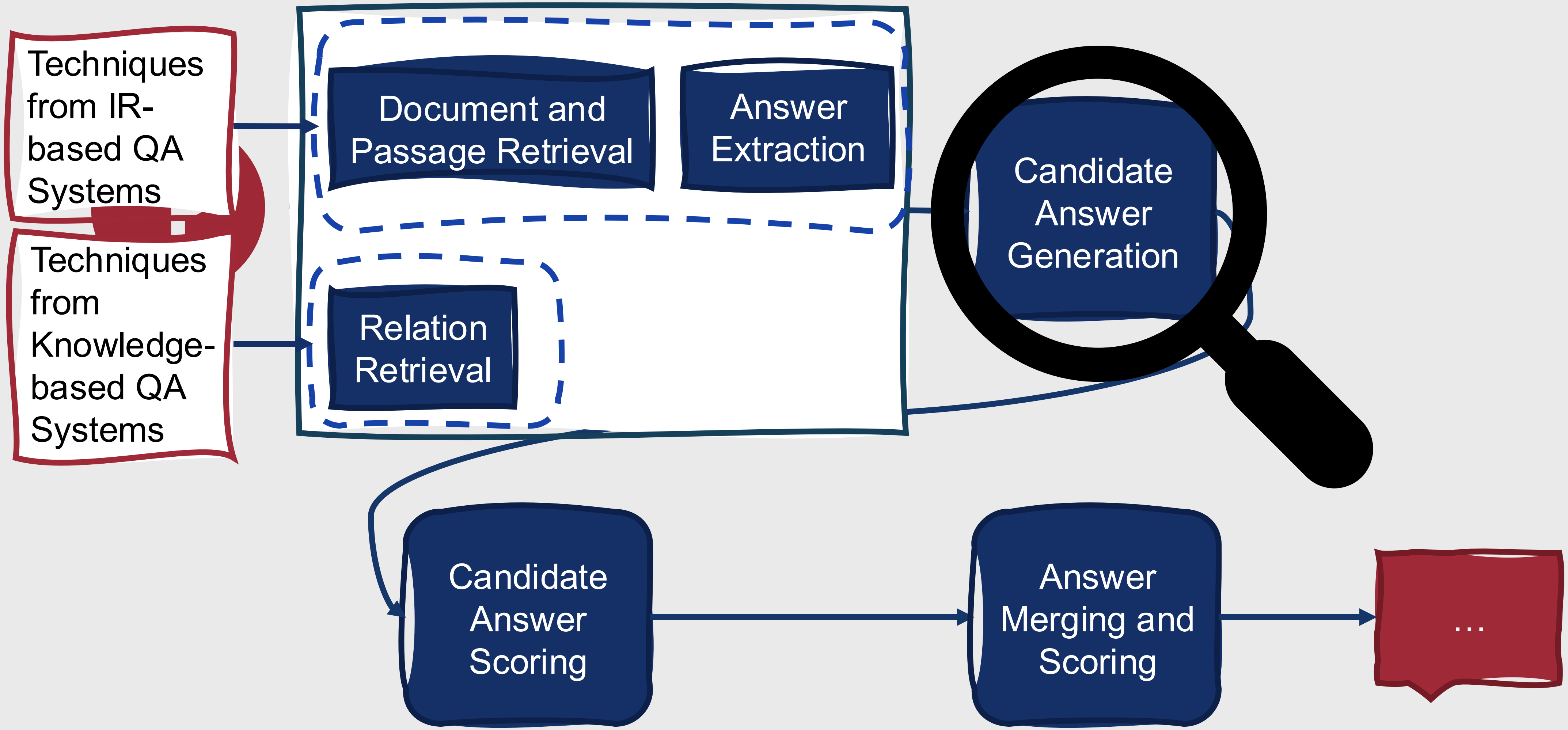
Question Classification: What type of question is this (multiple choice, fill-in-the-blank, definition, etc.)?

Generally done using pattern-matching regular expressions over words or parse trees

Stage 2: Candidate Answer Generation



Stage 2: Candidate Answer Generation



Stage 2: Candidate Answer Generation

Jeopardy! Example:
A new play based on **this Sir Arthur Conan Doyle canine classic** opened on the London stage in 2007.

Document and Passage Retrieval

In 2007, Peepolykus Theatre Company premiered a new adaptation of *The Hound of the Baskervilles* at West Yorkshire Playhouse in Leeds.

The play is an adaptation of the Arthur Conan Doyle's novel: The Hound of the Baskervilles (1901).

Stage 2: Candidate Answer Generation

Jeopardy! Example:
A new play based on **this Sir Arthur Conan Doyle canine classic** opened on the London stage in 2007.

Document and Passage Retrieval

In 2007, Peepolykus Theatre Company premiered a new adaptation of *The Hound of the Baskervilles* at West Yorkshire Playhouse in Leeds.

The play is an adaptation of the Arthur Conan Doyle's novel: The Hound of the Baskervilles (1901).

Answer Extraction

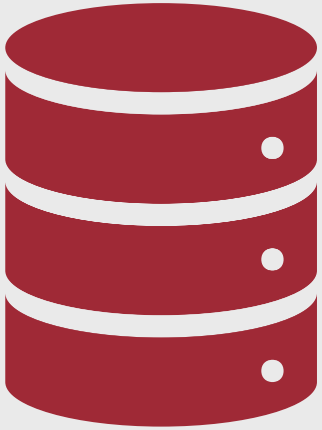
The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

Stage 2: Candidate Answer Generation

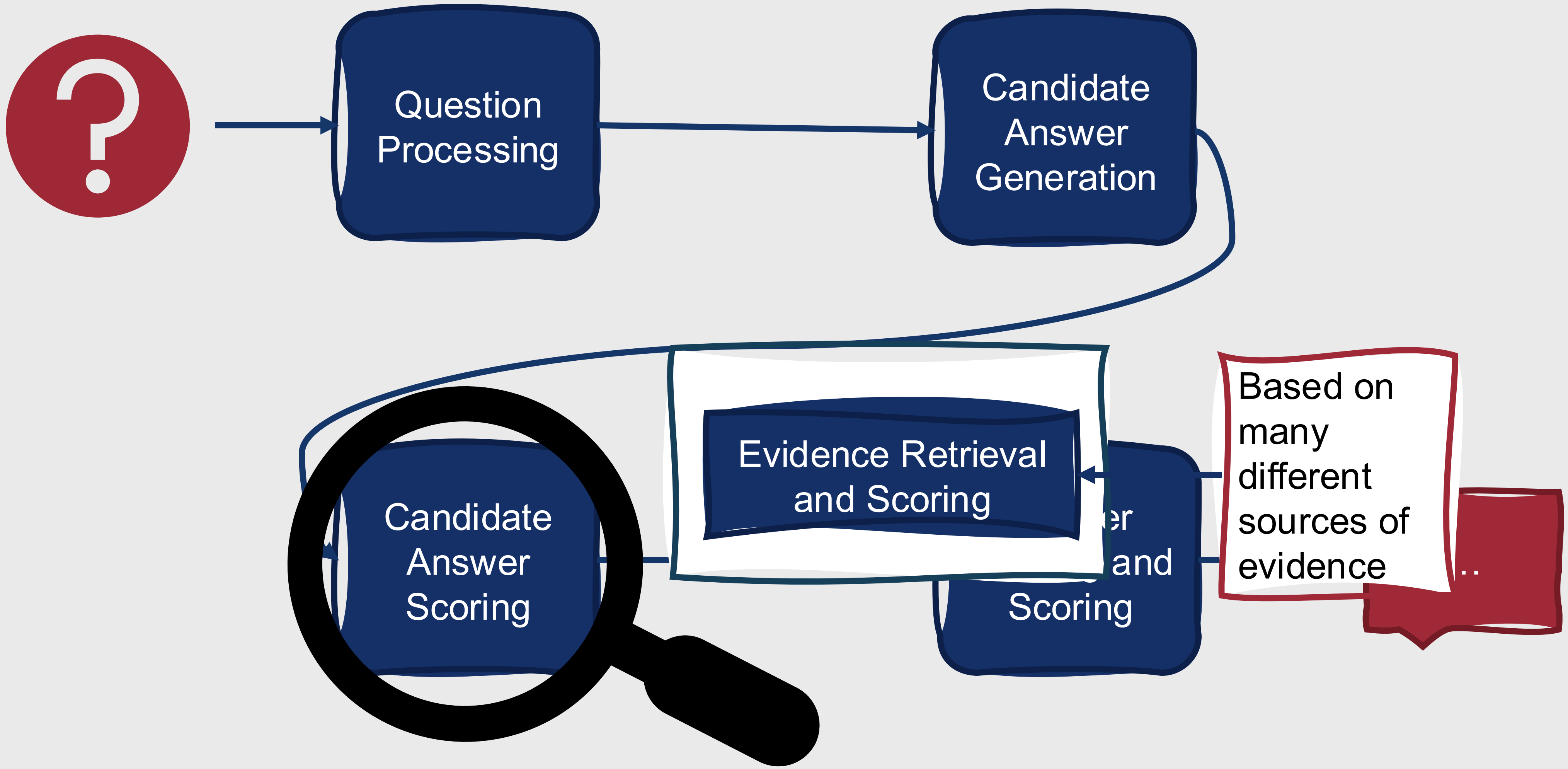
Jeopardy! Example:
basedOn(x, “Sir Arthur Conan Doyle canine classic”)

Relation Retrieval



The Hound of the Baskervilles

Stage 3: Candidate Answer Scoring



Stage 3: Candidate Answer Scoring

The Hound of the Baskervilles

The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

Information extracted from structured knowledge bases

Retrieved passages with terms matching the question

Stage 3: Candidate Answer Scoring

The Hound of the Baskervilles

The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

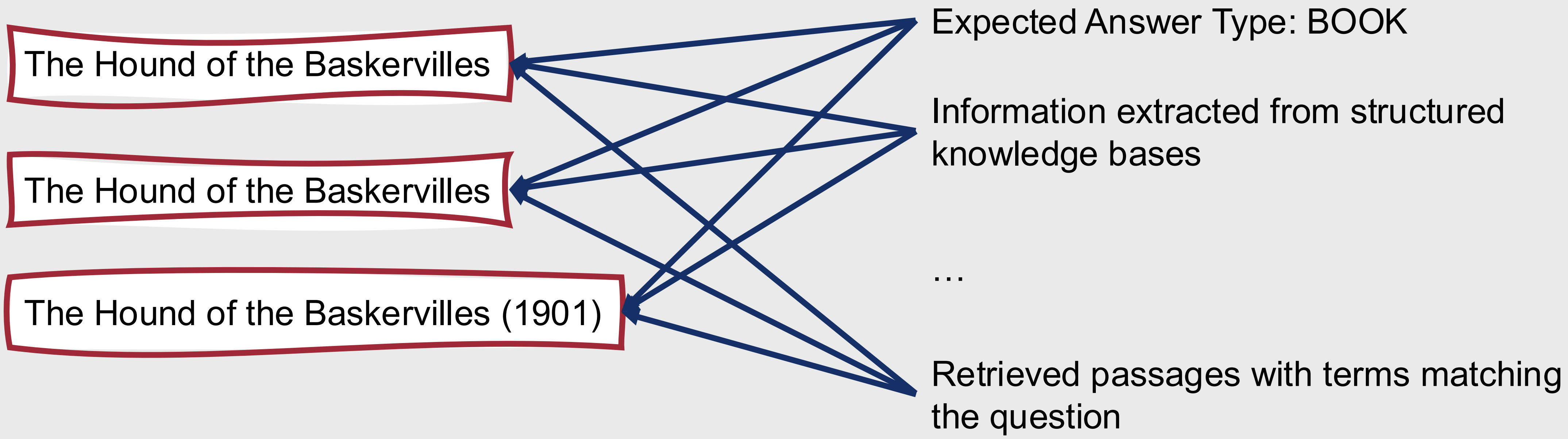
Expected Answer Type: BOOK

Information extracted from structured knowledge bases

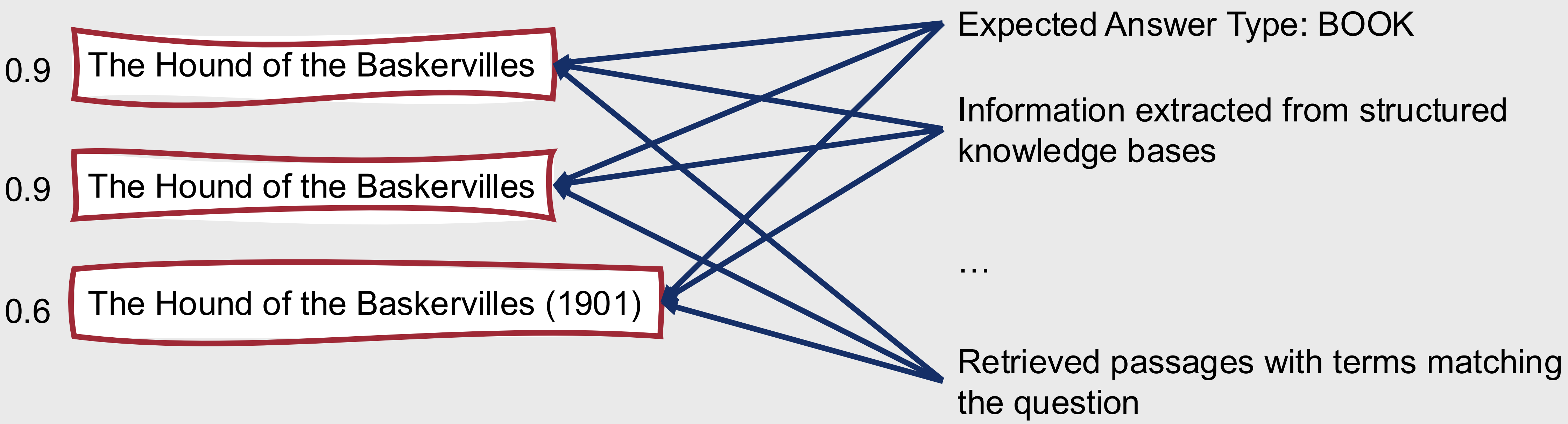
…

Retrieved passages with terms matching the question

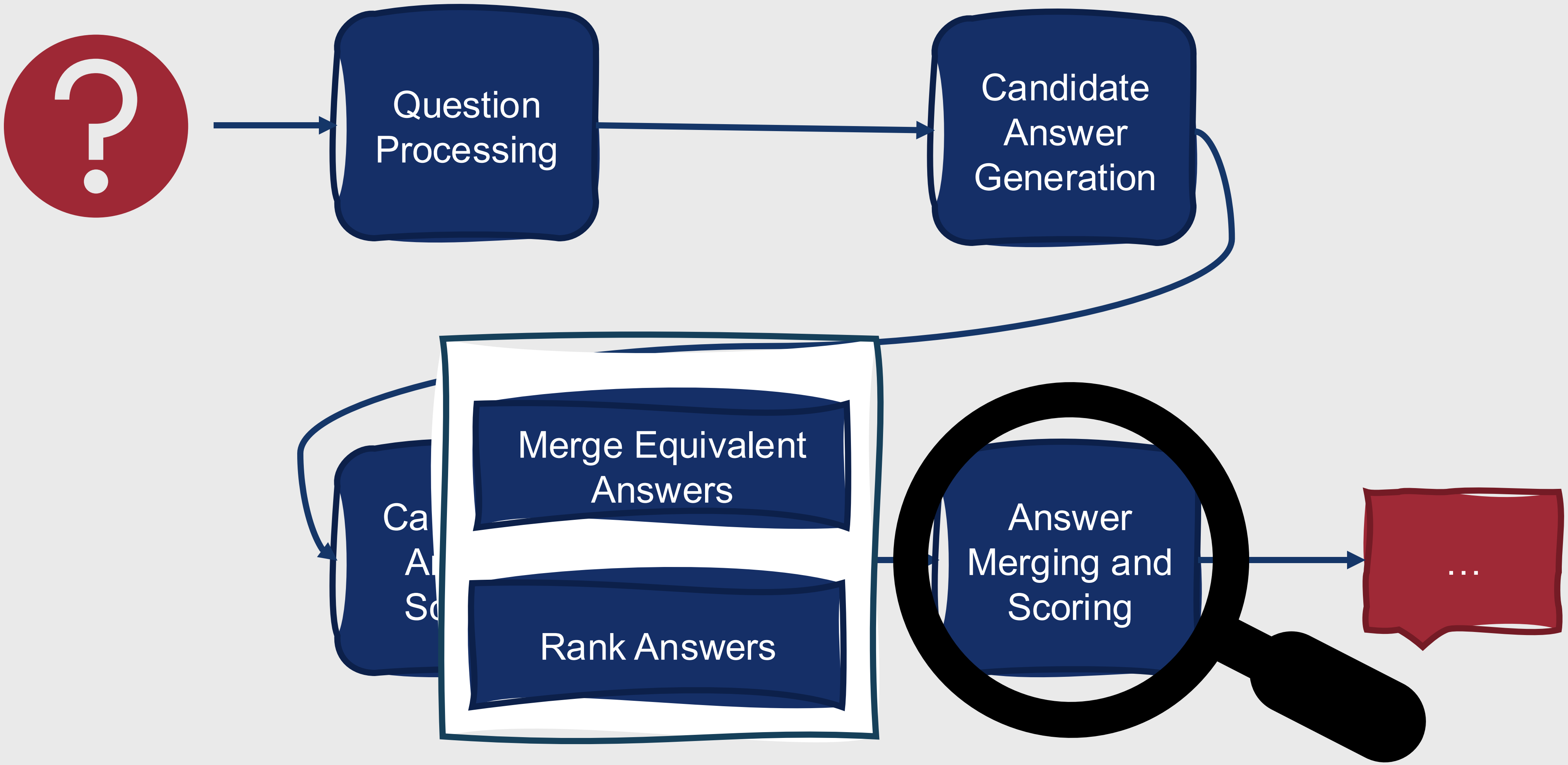
Stage 3: Candidate Answer Scoring



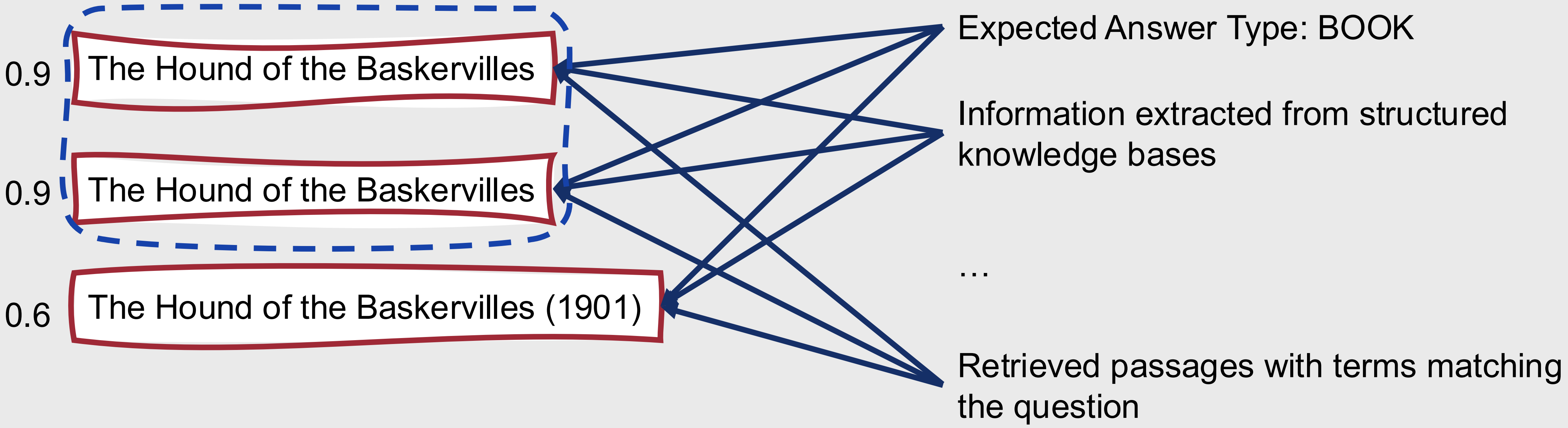
Stage 3: Candidate Answer Scoring



Stage 4: Answer Merging and Scoring



Stage 4: Answer Merging and Scoring



Stage 4: Answer Merging and Scoring

0.9 The Hound of the Baskervilles

0.6 The Hound of the Baskervilles (1901)

Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Stage 4: Answer Merging and Scoring

Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

This Week's Topics

Tuesday

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

Thursday

- Classic QA
- IR- and Knowledge-Based QA
- Evaluating QA Systems

Information Retrieval-based Question Answering

- Relies on text from the web or from large corpora
- Given a user question:
  1. Find relevant documents and passages of text
  2. Read the retrieved documents or passages
  3. Extract an answer to the question directly from spans of text

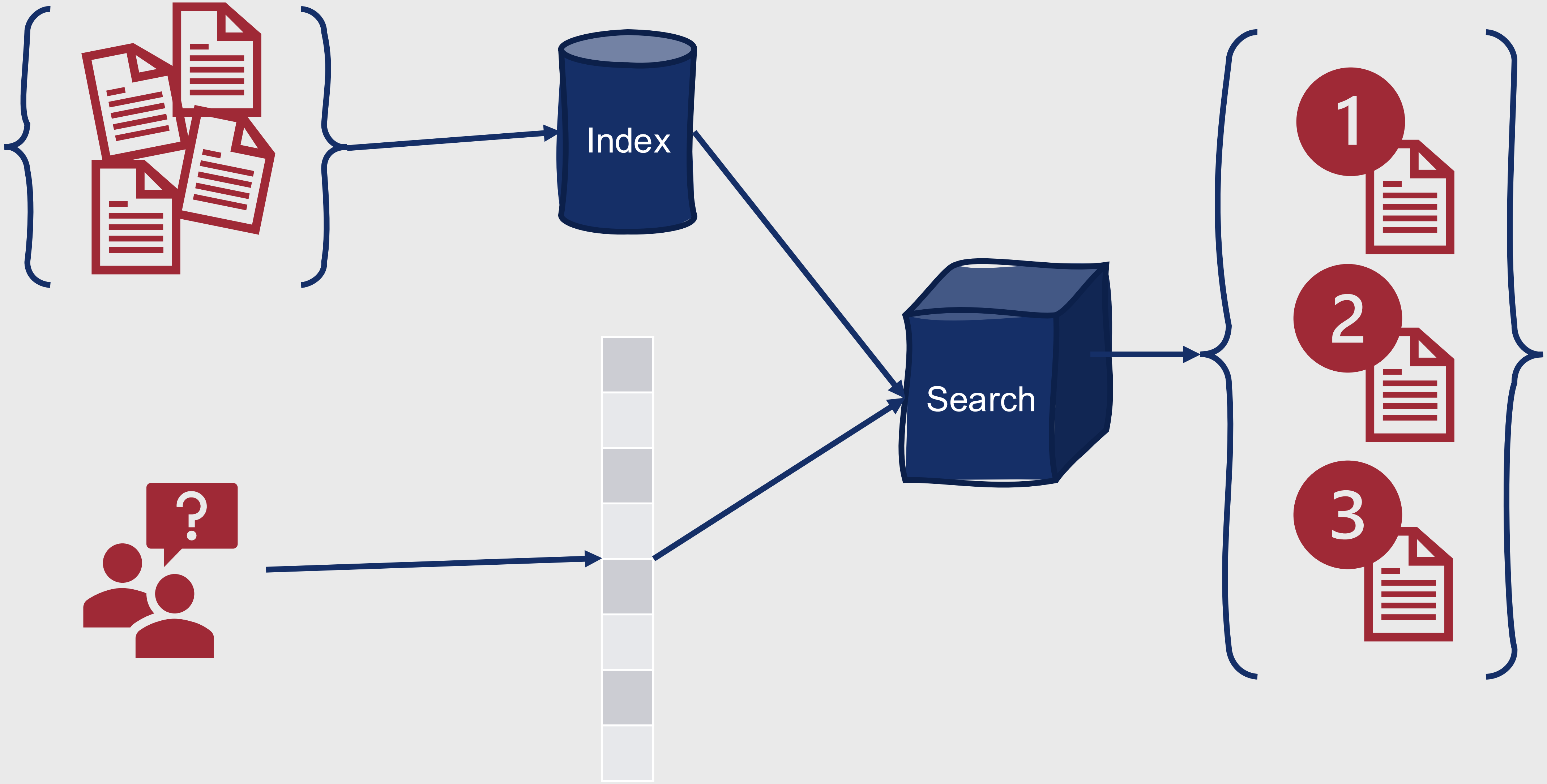
Knowledge-based Question Answering

Builds a semantic representation of the user's query

When was UIC founded? → founded(UIC, x)

Uses these representations to query a database of facts

How does information retrieval work?



Document Scoring

- Weight terms in documents to create document vectors
- Weight terms in queries to create query vectors
- Compute cosine similarity between a given document, d, and query, q
  - $\text{score}(q, d) = \cos(\mathbf{q}, \mathbf{d}) = \frac{\mathbf{q} \cdot \mathbf{d}}{|\mathbf{q}||\mathbf{d}|}$
- This can be slightly simplified since queries are usually short and query normalization won't have an impact on document ranking
  - $\text{score}(q, d) = \sum_{t \in q} \frac{\text{tf}-\text{idf}(t,d)}{|d|}$

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

$$\text{tf–idf}(t, d) = \log_{10}(\text{count}(t, d) + 1) * \log_{10} \frac{N}{\text{df}_t}$$

CS 421

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t, d) = \log_{10}(\text{count}(t, d) + 1) * \log_{10} \frac{N}{\text{df}_t}$$

Document 1

word	count	TF
CS	1	0.301
is	1	0.301
the	1	0.301
best	1	0.301
topic	1	0.301
421	0	0
covers	0	0
NLP	0	0
class	0	0

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d) + 1) * \log_{10}\frac{N}{\text{df}_t}$$

Document 1

word	count	TF	# docs	IDF
CS	1	0.301	2	0.176
is	1	0.301	2	0.176
the	1	0.301	2	0.176
best	1	0.301	2	0.176
topic	1	0.301	1	0.477
421	0	0	2	0.176
covers	0	0	1	0.477
NLP	0	0	1	0.477
class	0	0	1	0.477

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d)+1) * \log_{10}\frac{N}{\text{df}_t}$$

Document 1

word	count	TF	# docs	IDF	TF-IDF
CS	1	0.301	2	0.176	0.053
is	1	0.301	2	0.176	0.053
the	1	0.301	2	0.176	0.053
best	1	0.301	2	0.176	0.053
topic	1	0.301	1	0.477	0.144
421	0	0	2	0.176	0
covers	0	0	1	0.477	0
NLP	0	0	1	0.477	0
class	0	0	1	0.477	0

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d)+1) * \log_{10}\frac{N}{\text{df}_t}$$

word	count	TF	# docs	IDF	TF-IDF
					0.053

word	count	TF	# docs	IDF	TF-IDF
CS	1	0.301			
is	0	0			
the	0	0			
best	0	0			
topic	0	0			
421	1	0.301			
covers	1	0.301			
NLP	1	0.301			
class	0	0			

word	count	TF	# docs	IDF	TF-IDF
CS	0	0	2	0.176	0
is	1	0.301	2	0.176	0.053
the	1	0.301	2	0.176	0.053
best	1	0.301	2	0.176	0.053
topic	0	0	1	0.477	0
421	1	0.301	2	0.176	0.053
covers	0	0	1	0.477	0
NLP	0	0	1	0.477	0
class	1	0.301	1	0.477	0.144

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d) + 1) * \log_{10}\frac{N}{\text{df}_t}$$

Doc. 1	Doc. 2	Doc. 3
0.053	0.053	0
0.053	0	0.053
0.053	0	0.053
0.053	0	0.053
0.144	0	0
0	0.053	0.053
0	0.144	0
0	0.144	0
0	0	0.144

$$\text{score}(q,d) = \sum_{t \in q} \frac{\text{tf–idf}(t,d)}{|d|}$$

Doc.	|d|	TF-IDF("CS")	TF-IDF("421")	Score

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d)+1) * \log_{10}\frac{N}{\text{df}_t}$$

Doc. 1	Doc. 2	Doc. 3
0.053	0.053	0
0.053	0	0.053
0.053	0	0.053
0.053	0	0.053
0.144	0	0
0	0.053	0.053
0	0.144	0
0	0.144	0
0	0	0.144

$$\text{score}(q,d) = \sum_{t \in q} \frac{\text{tf–idf}(t,d)}{|d|}$$

Doc	|d|	TF-IDF("CS")	TF-IDF("421")	Score
1	0.179	0.053	0	0.296

$$\sqrt{0.053^2 + 0.053^2 + 0.053^2 + 0.053^2 + 0.144^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2} = 0.179$$

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d)+1) * \log_{10}\frac{N}{\text{df}_t}$$

Doc. 1	Doc. 2	Doc. 3
0.053	0.053	0
0.053	0	0.053
0.053	0	0.053
0.053	0	0.053
0.144	0	0
0	0.053	0.053
0	0.144	0
0	0.144	0
0	0	0.144

$$\text{score}(q,d) = \sum_{t \in q} \frac{\text{tf–idf}(t,d)}{|d|}$$

Doc	|d|	TF-IDF(“CS”)	TF-IDF(“421”)	Score
1	0.179	0.053	0	0.296
2	0.260	0.053	0.053	0.408

$$\sqrt{0.053^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0.053^2 + 0.144^2 + 0.144^2 + 0.144^2 + 0^2} = 0.260$$

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d)+1) * \log_{10}\frac{N}{\text{df}_t}$$

Doc. 1	Doc. 2	Doc. 3
0.053	0.053	0
0.053	0	0.053
0.053	0	0.053
0.053	0	0.053
0.144	0	0
0	0.053	0.053
0	0.144	0
0	0.144	0
0	0	0.144

$$\text{score}(q,d) = \sum_{t \in q} \frac{\text{tf–idf}(t,d)}{|d|}$$

Doc	|d|	TF-IDF("CS")	TF-IDF("421")	Score
1	0.179	0.053	0	0.296
2	0.260	0.053	0.053	0.408
3	0.179	0	0.053	0.296

$$\sqrt{0^2 + 0.053^2 + 0.053^2 + 0.053^2 + 0^2 + 0.053^2 + 0^2 + 0^2 + 0^2 + 0.144^2} = 0.179$$

Document Scoring: Case Example

CS is the best topic!

CS 421 covers NLP.

421 is the best class.

CS 421

$$\text{tf–idf}(t,d) = \log_{10}(\text{count}(t,d)+1) * \log_{10}\frac{N}{\text{df}_t}$$

Doc. 1	Doc. 2	Doc. 3
0.053	0.053	0
0.053	0	0.053
0.053	0	0.053
0.053	0	0.053
0.144	0	0
0	0.053	0.053
0	0.144	0
0	0.144	0
0	0	0.144

$$\text{score}(q,d) = \sum_{t\in q}\frac{\text{tf–idf}(t,d)}{|d|}$$

Doc	|d|	TF-IDF(“CS”)	TF-IDF(“421”)	Score
1	0.179	0.053	0	0.296
2	0.260	0.053	0.053	0.408
3	0.179	0	0.053	0.296

Evaluating IR Systems

Regular NLP metrics

- Precision
- Recall

More sophisticated IR-specific metrics

- Precision-recall curve
- Mean average precision

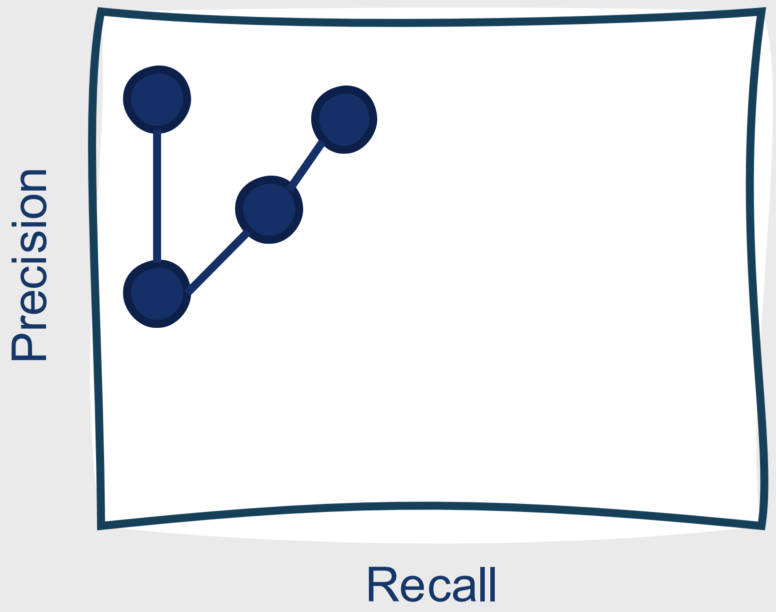
Precision-Recall Curve

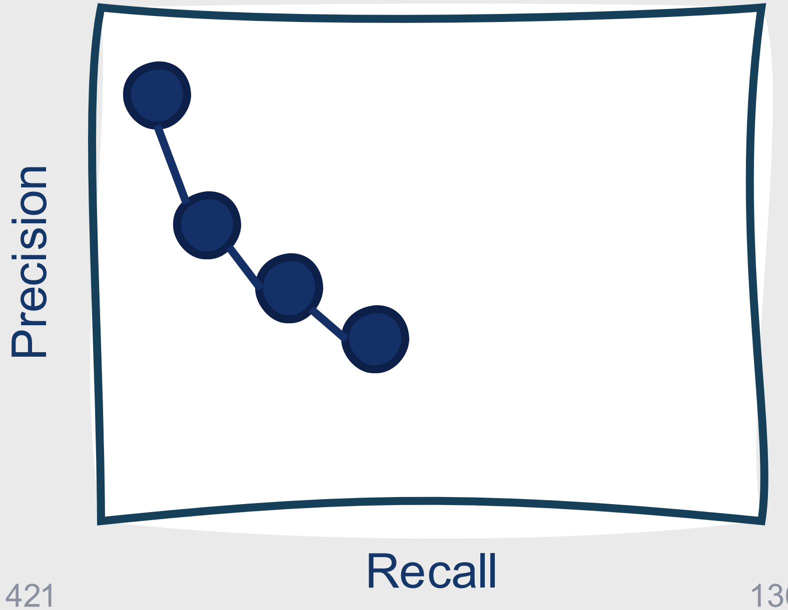
Rank	Precision at Rank	Recall at Rank
1	1.0	0.1
2	0.5	0.1
3	0.67	0.2
4	0.75	0.3

$$\text{IntPrecision}(r) = \max_{i \geq r} \text{Precision}(i)$$

Interpolated Precision	Recall
1.0	0.0
1.0	0.1
0.75	0.2
0.75	0.3

Avg. Interpolated Precision	Recall
1.0	0.0
0.75	0.1
0.5	0.2
0.4	0.3





Mean Average Precision

- Single score across multiple queries
- Given a set of queries Q and a set, R_r, of relevant documents d at or above rank r:
  - $\mathrm{AP} = \frac{1}{|R_r|} \sum_{d \in R_r} \mathrm{Precision}_r(d)$
  - $\mathrm{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{AP}(q)$

Information Retrieval-based Question Answering

- Relies on text from the web or from large corpora
- Given a user question:
  1. Find relevant documents and passages of text
  2. Read the retrieved documents or passages
  3. Extract an answer to the question directly from spans of text

Answer Span Extraction

- Goal: Compute, for each token, the probability that it is:
  - The start of the answer span
  - The end of the answer span

How many floors are in the Science and Engineering Offices building?

Although there are 13 floors in SEO, the elevator only goes to the 12th floor since the architect didn't like how elevator boxes look on the top of buildings.

P_{start}("13")

P_{end}("13")

Answer Span Extraction

- Common approach
  - Concatenate the query and passage, separated by a [SEP] token
  - Pass the concatenated sequence to an encoder
  - Add a linear layer that learns span-start (*S*) and span-end (*E*) embeddings
  - Compute span-start and span-end probabilities for each token p_i in a passage P
    - $P_{\text{start}_i} = \frac{e^{S \cdot p_i}}{\sum_{j=0}^{|P|} e^{S \cdot p_j}}$
    - $P_{\text{end}_i} = \frac{e^{E \cdot p_i}}{\sum_{j=0}^{|P|} e^{E \cdot p_j}}$
  - Select the highest-scoring passage
- This is **extractive** QA approach

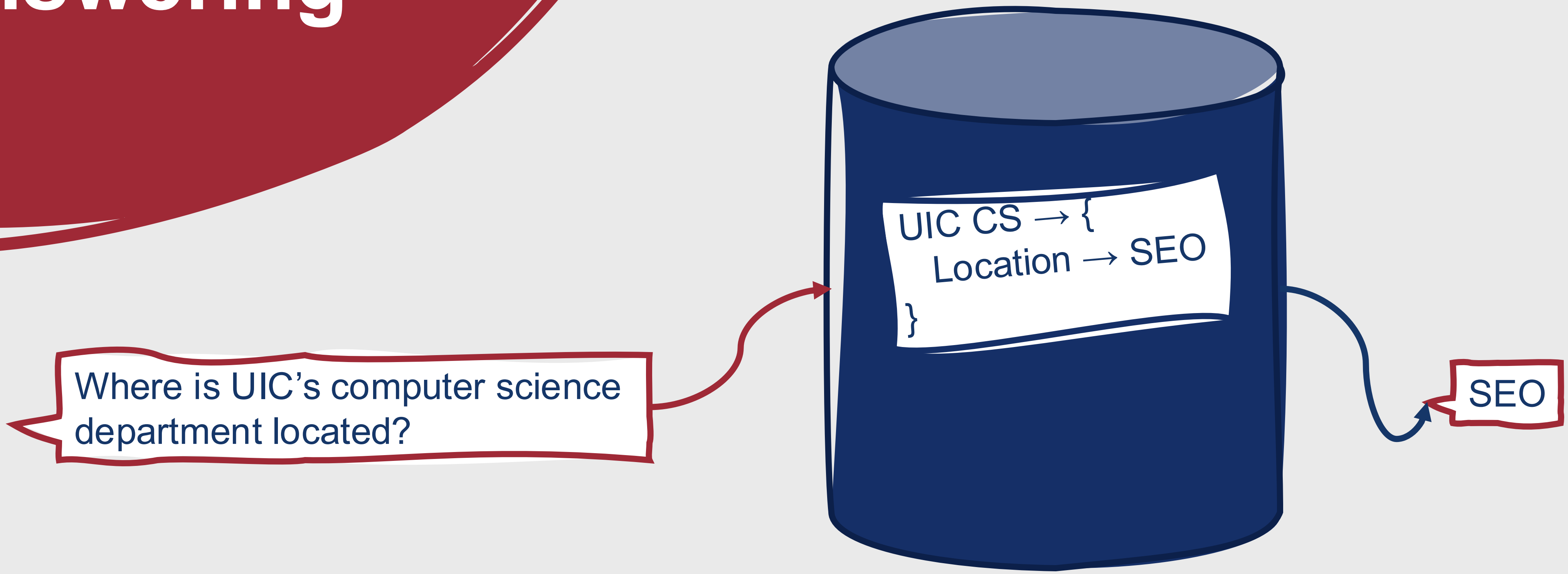
Entity Linking

- Modern approaches often make use of **bidirectional Transformer encoders**
  - One encoder is trained to encode a candidate mention
  - One encoder is trained to encode an entity (e.g., a Wikipedia page)
  - The dot product between the two encoded representations is computed
- Require annotated data indicating mention boundaries and corresponding entity links
  - **WebQuestionsSP:** https://www.microsoft.com/en-us/download/details.aspx?id=52763
  - **GraphQuestions:** https://github.com/ysu1989/GraphQuestions

The coolest department at UIC is the Department of Computer Science.

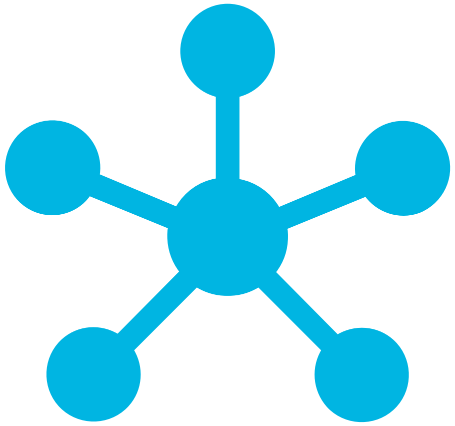
Knowledge-based Question Answering

- Answers questions by mapping them to formal queries over structured knowledge sources



Knowledge-based Question Answering

- Two common paradigms:
  - **Graph-based** question answering
  - Question answering by **semantic parsing**
- Both require entity linking



Graph-based Question Answering

- Facts are stored as (subject, predicate, object) triples
- Entity mentions are linked to entities in a knowledge graph
- Queries are mapped to canonical relations
  - *"Where is UIC's computer science department located?"* → LOCATIONOF("UIC CS", ?x)
  - Can be done using similar methods to neural entity linking
- Triples matching the canonical relations are identified and ranked
  - Can be done based on entity graph structure

Question Answering by Semantic Parsing

- Maps questions directly to logical form using a semantic parser
  - First-order logic
  - SQL
- Logical form is used to query a knowledge base directly

Where are we today?

- New paradigm: **language model-based question answering**
  - In pretraining, train an encoder-decoder architecture to fill in masked spans of text
  - In finetuning, train the decoder to output an answer for a given question

PA How do question answering systems work today?

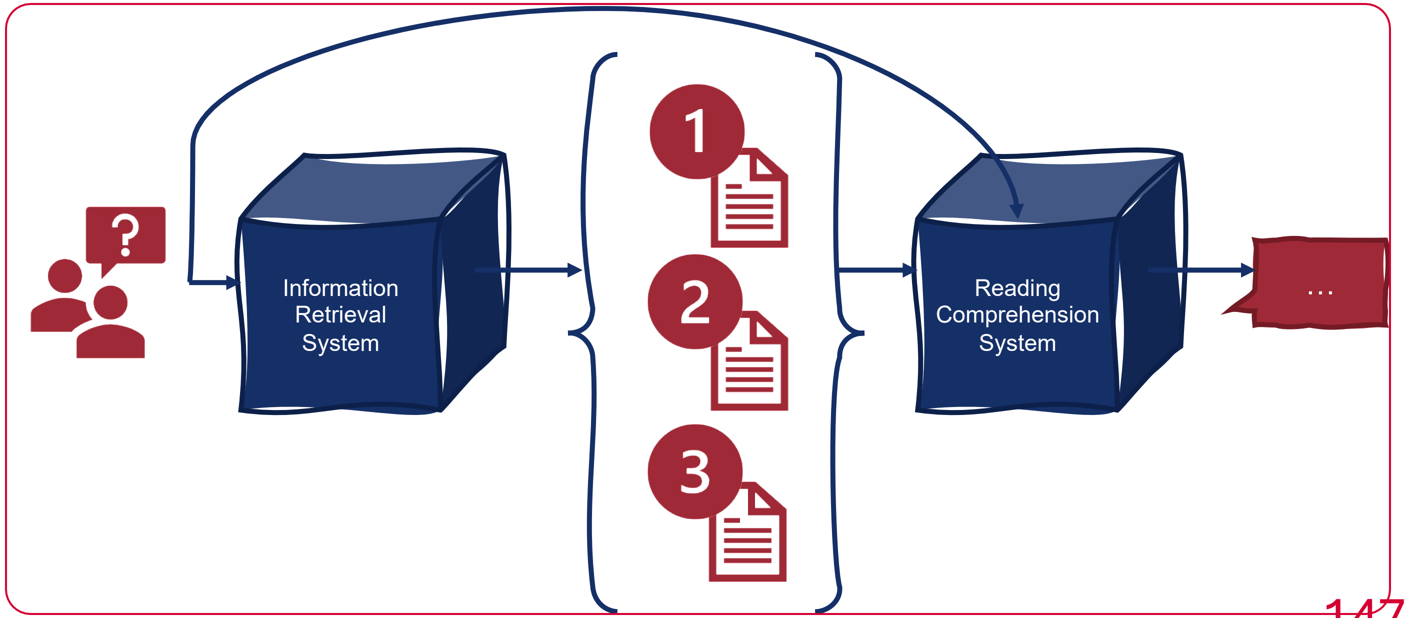
Today's question answering systems typically use a combination of natural language processing (NLP) and machine learning (ML) techniques to understand the user's question and generate an answer. The process usually involves the following steps:

1. Question understanding: The system converts the text of the user's question into a machine-readable format and analyzes it to identify the intent and extract relevant information.
2. Information retrieval: The system searches a pre-defined knowledge source (such as a database, a web page, or a large corpus of text) to find information that may answer the user's question.
3. Answer generation: The system selects the most relevant information from the retrieved data and generates an answer. In some systems, this may involve additional NLP and ML techniques, such as summarization or text generation, to produce a concise and coherent answer.
4. Answer ranking: The system uses various methods, such as word overlaps, word embeddings, or neural networks, to rank the answers and select the most likely candidate.

Overall, the accuracy and effectiveness of today's question answering systems depend on the quality and scope of their training data, the sophistication of their NLP and ML models, and the ability to handle complex and ambiguous questions.

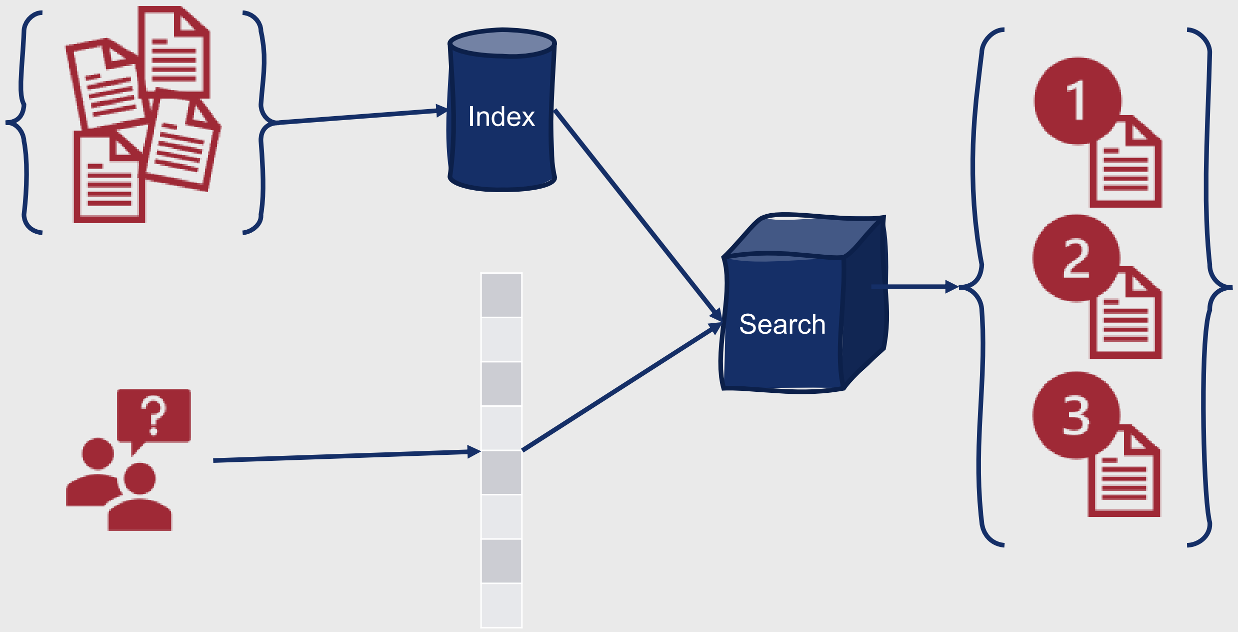
Retrieval-Augmented Generation (RAG)

- Retrieval-augmented generation: The process of generating answers based on retrieved documents
- RAG uses a "retrieve and read" model:
  - **Retrieve** relevant documents for the given query
  - **Read** those documents to find text segments that answer the query



Step #1: Retrieve

- Performed using a standard information retrieval architecture



Step #2: Read

- Performed using a **reading comprehension** model
- **Reading comprehension:** Given a document and a query, select (if available) the span of text from the document that answers the query
  - Provides a way to measure natural language understanding based on children's reading comprehension tests

Prime_number
The Stanford Question Answering Dataset

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, 1 · 3, 1 · 1 · 3, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?
Ground Truth Answers: itself itself itself itself itself

What are numbers greater than 1 that can be divided by 3 or more numbers called?
Ground Truth Answers: composite number composite number composite number primes

What theorem defines the main role of primes in number theory?
Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic

Any number larger than 1 can be represented as a product of what?
Ground Truth Answers: a product of primes product of primes that is unique up to ordering primes primes primes that is unique up to ordering

Why must one be excluded in order to preserve the uniqueness of th

Reading Comprehension Datasets

- Stanford Question Answering Dataset (SQuAD)
  - English
  - Passages from Wikipedia
  - Associated questions
    - Many have answers that are spans from the passage
    - Some are designed to be unanswerable
  - https://rajpurkar.github.io/SQuAD-explorer/
- HotpotQA
  - English
  - Question-answer pairs based on multiple context documents
  - https://hotpotqa.github.io/
- Natural Questions
  - English
  - Based on real, anonymized queries to Google Search
  - https://ai.google.com/research/NaturalQuestions
- TyDi QA
  - Question-answer pairs from typologically diverse languages
  - https://ai.google.com/research/tydiqa

How do we implement RAG?

- Conditional (autoregressive) generation, using question-answer pairs for pretraining and/or fine-tuning
  - $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p([\mathtt{Q:}]; q; [\mathtt{A:}]; x_{<i})$

Q: Who wrote the book "Funny Story"? A: Emily Henry

Q: Who wrote the book "The Extinction of Irena Rey"? A:

This Week's Topics

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

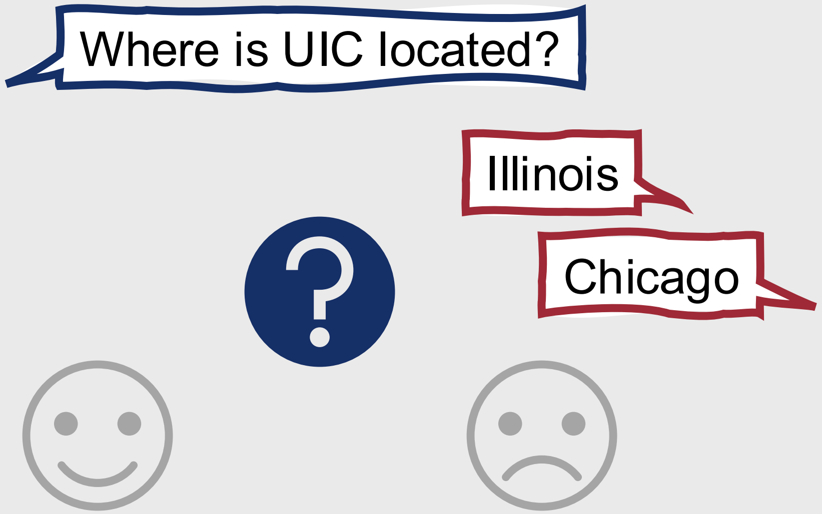
Thursday

Tuesday

Classic QA
IR- and Knowledge-Based QA
Evaluating QA Systems

How are question answering systems evaluated?

- Common metric for factoid question answering: **Mean Reciprocal Rank**
  - Assumes that gold standard answers are available for test questions
  - Assumes that systems return a short ranked list of answers



Mean Reciprocal Rank

- Scores each question according to the reciprocal of the rank of the first correct answer
  - Highest ranked correct answer is ranked fourth → reciprocal rank = ¼
- Assigns a score of 0 to questions with no correct answers returned
- System's overall score is the average of all individual question scores
  - $\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{r_i}$

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction	Rank
Illinois	1
West Loop	2
Chicago	3
Little Italy	4

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction	Rank
Illinois	1
West Loop	2
Chicago	3
Little Italy	4

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction	Rank
Illinois	1
West Loop	2
Chicago	3
Little Italy	4

Reciprocal
Rank = 1/3

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Who is the head of UIC's Department of Computer Science? ← Question

Gold Standard → Baoxin Li

Prediction	Rank
Illinois	1
West Loop	2
Chicago	3
Little Italy	4

Reciprocal Rank = 1/3

Prediction	Rank
Robert Sloan	1
Baoxin Li	2
Natalie Parde	3
Grace Hopper	4

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction	Rank
Illinois	1
West Loop	2
Chicago	3
Little Italy	4

Reciprocal Rank = 1/3

Who is the head of UIC's Department of Computer Science? ← Question

Gold Standard → Baoxin Li

Prediction	Rank
Robert Sloan	1
Baoxin Li	2
Natalie Parde	3
Grace Hopper	4

Reciprocal Rank = 1/2

Mean Reciprocal Rank

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction	Rank
Illinois	1
West Loop	2
Chicago	3
Little Italy	4

Reciprocal Rank = 1/3

Who is the head of UIC's Department of Computer Science? ← Question

Gold Standard → Baoxin Li

Prediction	Rank
Robert Sloan	1
Baoxin Li	2
Natalie Parde	3
Grace Hopper	4

Reciprocal Rank = 1/2

$$\text{MRR} = \frac{\frac{1}{3}+\frac{1}{2}}{2} = 0.417$$

Other Evaluation Metrics for Question Answering Systems

- **Exact Match**
  - Remove punctuation and articles
  - Compute the percentage of predicted answers that match the gold standard answer exactly

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011

Other Evaluation Metrics for Question Answering Systems

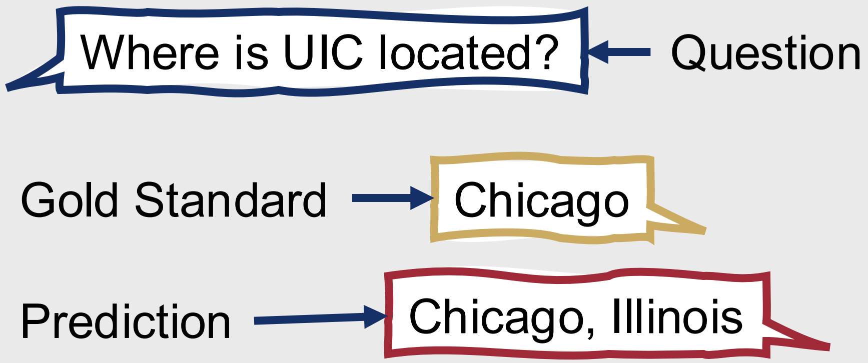
- **F_1 Score**
  - Remove punctuation and articles
  - Treat the predicted and gold standard answers as bags of tokens
  - True positives: Tokens that exist in both the gold standard and predicted answers
  - Average F_1 over all questions

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011

Computing F_1 for Question Answering Systems



	Actual True	Actual False
Predicted True		
Predicted False		

Computing F_1 for Question Answering Systems

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction → Chicago, Illinois

	Actual True	Actual False
Predicted True	1	1
Predicted False	0	

Computing F_1 for Question Answering Systems

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction → Chicago, Illinois

	Actual True	Actual False
Predicted True	1	1
Predicted False	0	

Precision = $\frac{TP}{TP+FP} = \frac{1}{1+1}$ = 0.5

Recall = $\frac{TP}{TP+FN} = \frac{1}{1+0}$ = 1

F_1 = $\frac{2*P*R}{P+R} = \frac{2*0.5*1}{0.5+1}$ = 0.67

Summary: Question Answering

- **Question answering** is the process of retrieving relevant information and fluently presenting it to users in response to their queries
- QA systems often use **knowledge-based** or **information retrieval** methods to formulate answers to questions
- Some systems also use **language modeling** or **rule-/feature-based approaches**
- Many recent high-performing approaches leverage **retrieval-augmented generation**
- QA systems are often evaluated using **mean reciprocal rank**, **exact match**, or **F_1 metrics**