Question Answering and Summarization

Natalie Parde, Ph.D.

Department of Computer Science

University of Illinois at Chicago

CS 421: Natural Language Processing

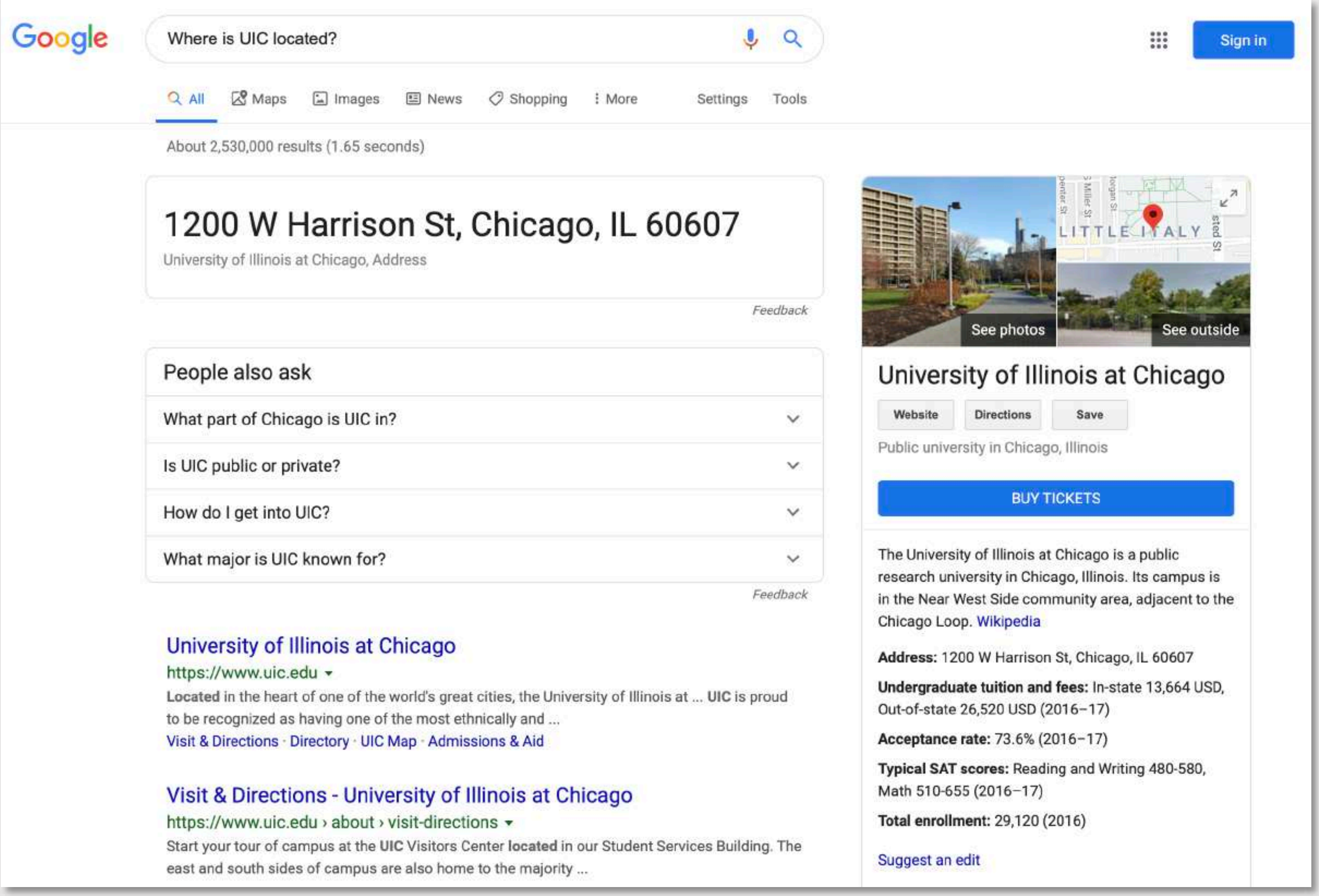
Fall 2019

Many slides adapted from Jurafsky and Martin (https://web.stanford.edu/~jurafsky/slp3/).

What is question answering?

- The process of **automatically retrieving** compact quantities of correct, relevant **information** in response to a user’s **query**

We use question answering systems everyday.



Where is UIC located?

About 2,530,000 results (1.65 seconds)

1200 W Harrison St, Chicago, IL 60607

University of Illinois at Chicago, Address

Feedback

People also ask

What part of Chicago is UIC in?

Is UIC public or private?

How do I get into UIC?

What major is UIC known for?

Feedback

University of Illinois at Chicago

https://www.uic.edu

Located in the heart of one of the world's great cities, the University of Illinois at ... **UIC** is proud to be recognized as having one of the most ethnically and ...

Visit & Directions · Directory · UIC Map · Admissions & Aid

Visit & Directions - University of Illinois at Chicago

https://www.uic.edu › about › visit-directions

Start your tour of campus at the **UIC** Visitors Center **located** in our Student Services Building. The east and south sides of campus are also home to the majority ...

University of Illinois at Chicago

Public university in Chicago, Illinois

The University of Illinois at Chicago is a public research university in Chicago, Illinois. Its campus is in the Near West Side community area, adjacent to the Chicago Loop. Wikipedia

Address: 1200 W Harrison St, Chicago, IL 60607

Undergraduate tuition and fees: In-state 13,664 USD, Out-of-state 26,520 USD (2016–17)

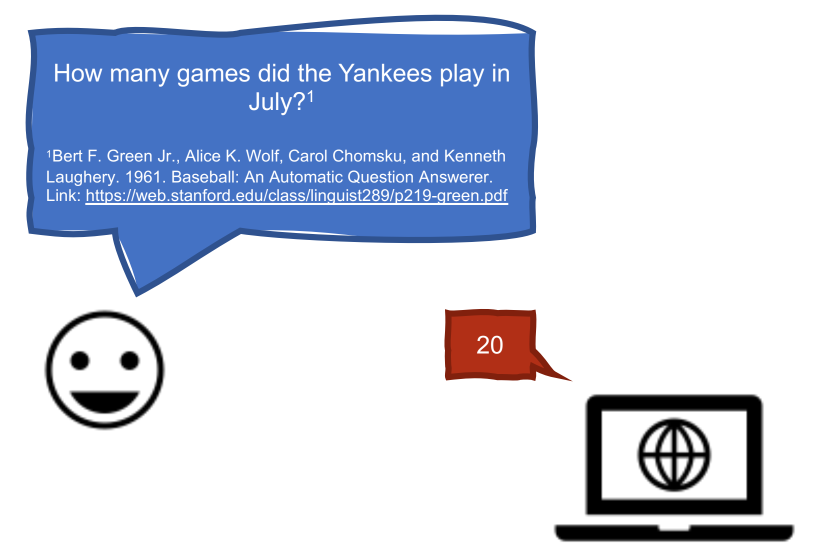
Acceptance rate: 73.6% (2016–17)

Typical SAT scores: Reading and Writing 480-580, Math 510-655 (2016–17)

Total enrollment: 29,120 (2016)

Suggest an edit

People have been interested in question answering systems nearly as long as computers have existed.



TEDBlog

Technology > TEDx

How did supercomputer Watson beat Jeopardy champion Ken Jennings? Experts discuss.

Posted by: Kate Torgovnick May *April 5, 2013 at 1:59 pm EDT*

Question answering systems have even won game shows!

https://blog.ted.com/how-did-supercomputer-watson-beat-jeopardy-champion-ken-jennings-experts-discuss/

Question Answering Systems

- Typically focus on **factoid questions**
  - **Factoid Questions:** Questions that can be answered with simple facts expressed in short texts

When was UIC founded?

What is the average CS class size?

How far is UIC from the University of Chicago?

Question Answering Systems

- Two major paradigms:
  - **Information retrieval-based** question answering
  - **Knowledge-based** question answering

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

Large industrial systems are often hybrids of these two paradigms.

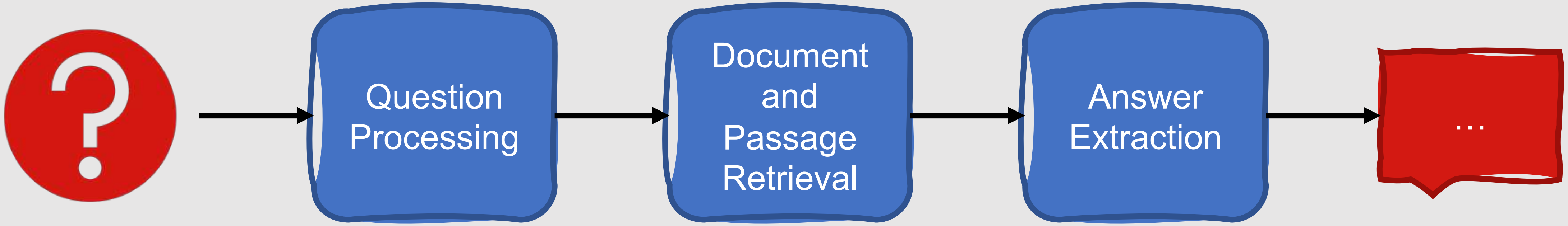
- DeepQA (the question answering system in IBM's Watson):
  - Finds candidate answers in both knowledge bases and text sources
  - Scores each candidate answer
  - Returns the highest scoring answer

Information Retrieval-based Question Answering

Goal: Answer a user's question by finding short text segments containing the requested information

QUESTION	ANSWER
Where is UIC located?	in Chicago, Illinois
What does UIC stand for?	University of Illinois at Chicago
Who taught CS 421 in Fall 2019?	Natalie Parde
How many grad students are in CS 421?	25

Information Retrieval-based Question Answering



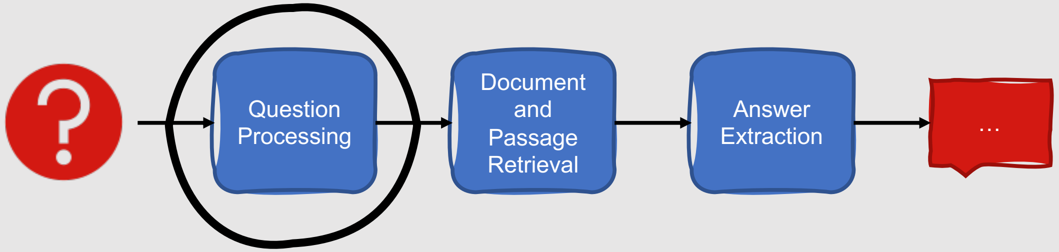
Question Processing

Document and Passage Retrieval

Answer Extraction

…

Question Processing



- **Goal: Extract the query**
  - What keywords are needed to match relevant documents?
  - What type of entity should be in the answer (person, location, etc.)?
  - What is the focus of the question (which string of words will likely be replaced by the answer)?
  - What type of question is this (definition, math, list, etc.)?

Question Processing

- Two most common subtasks involved in question processing:
  - Query formulation
  - Answer type detection

When was UIC's Department of Computer Science created?

Query: UIC Department of Computer Science created
Answer Type: Time

Query Formulation

- The task of creating a query to send to an information retrieval system
  - Should contain keywords necessary to obtain relevant documents
- Simple strategy: Pass the entire question as a query
  - Only works with very large corpora (e.g., the web)
- More complex strategy for smaller corpora (e.g., corporate websites or Wikipedia): Use an IR engine to search and index documents

Common Information Retrieval Techniques

TF*IDF matching

- Which document has the highest cosine similarity with the query?

Query expansion

- Add query terms in hopes of matching an answer in one of its many possible forms

Query reformulation

- Rephrase the question to make it look like a substring of possible answers
  - **When was UIC founded?** → **UIC was founded in**

Answer Type Detection

- The task of determining what type of named entity is needed for the answer
  - Who was the first head of UIC's Department of Computer Science? → PERSON
  - In what city is UIC located? → CITY
- In addition to named entity types, answers can also fall under other categories in a larger, hierarchical, answer type taxonomy
  - PERSON:INDIVIDUAL
  - PERSON:GROUP

Answer Type Detection

- Hierarchical answer type taxonomy
    - Coarse-grained categories:
        - Abbreviation
        - Description
        - Entity
        - Human
        - Location
        - Numeric
    - Finer-grained subcategories of each

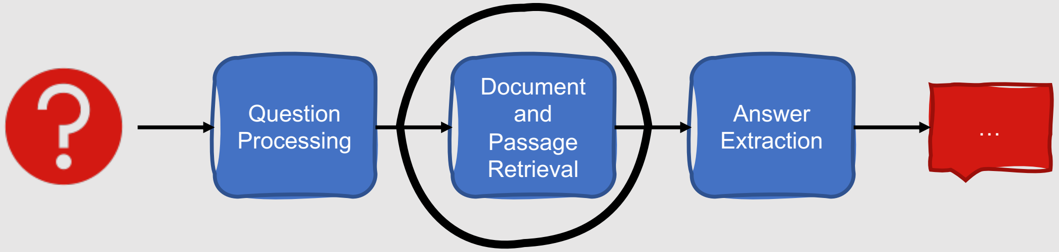
ABBREVIATION	
abb	What's the abbreviation for limited partnership?
exp	What does the "c" stand for in the equation E=mc2?
DESCRIPTION	
definition	What are tannins?
description	What are the words to the Canadian National anthem?
manner	How can you get rust stains out of clothing?
reason	What caused the Titanic to sink?
ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma" in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?
HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
temp	How hot should the oven be when making Peachy Oat Muffins?
speed	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

Figure 25.4 Question typology from Li and Roth (2002), (2005). Example sentences are from their corpus of 5500 labeled questions. A question can be labeled either with a coarse-grained tag like HUMAN or NUMERIC or with a fine-grained tag like HUMAN:DESCRIPTION.

How are answer types detected?

- **Handwritten rules**
  - Who {is | was} the first head of ORGANIZATION → PERSON
- **Supervised machine learning**
- In general, detecting answer types like PERSON, LOCATION, and TIME is easier; detecting other types is more complex

Document and Passage Retrieval

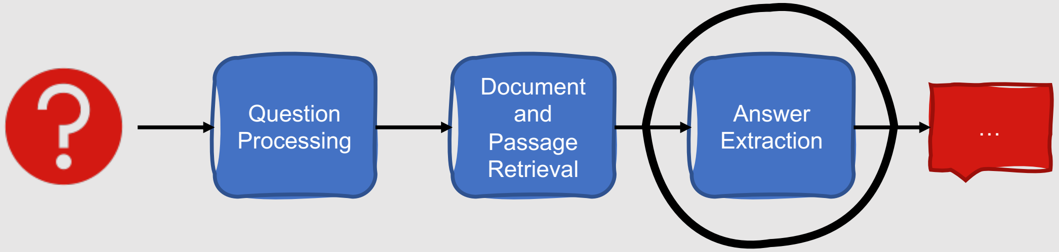


- Ranks a set of documents based on their relevance to the query
- Divides the top *n* documents into smaller passages
- Pass some or all of those passages along to the next stage

Which passages are passed along to the next stage?

- Simplest approach: Pass along every passage from the top *n* documents to the next stage
- More sophisticated approaches:
  - Filter the passages based on whether they contain a named entity of the type specified by the question
  - Rank the passages using supervised machine learning and return the subset of highest-ranked passages

Answer Extraction



- Extracts a specific answer from a passage
  - **Span Labeling:** Given a passage, identify the span of text which constitutes an answer

How can we extract answers from passages?

- Simple approach: Run a named entity tagger on the candidate passage, and return whatever entity corresponds to the desired answer type
- However, the answers to many questions may not require a specific named entity type!
  - **What is natural language processing?** → **The subfield of artificial intelligence that focuses on automatically interpreting and generating natural language**
- Thus, more sophisticated answer extraction systems tend to use supervised machine learning

In what city is UIC located?

UIC, the the largest university in **Chicago**….

Feature-based Answer Extraction

Answer type match	• Does the candidate answer contain a phrase with the correct answer type?
Number of matched keywords	• How many keywords from the question are included in the candidate answer?
Text similarity	• What is the cosine similarity between the candidate answer and the query keywords?
Novelty factor	• Does the candidate answer contain a word that was not in the query?
Apposition features	• Is the candidate answer appositive to a phrase containing many question terms? • The professor, **Natalie Parde**, is in her office making slides.
Punctuation location	• Is the candidate answer immediately followed by punctuation?
Sequences of question terms	• How long is the longest sequence of question terms in the candidate answer?

N-gram Tiling Answer Extraction

Relies on the redundancy of the web

Works by:

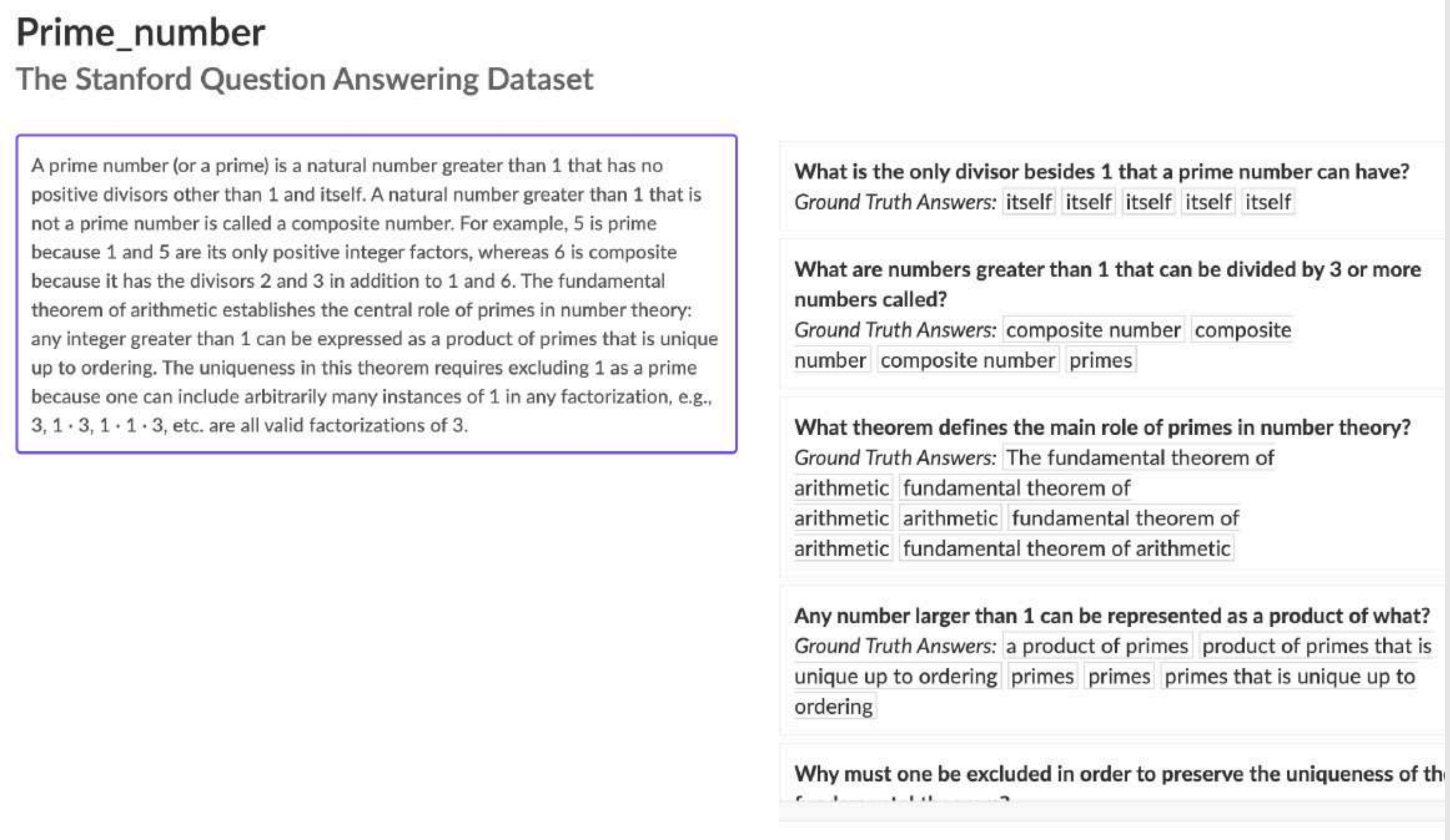
- Starting with the **text snippets returned from a web search engine**
- **Extracting all of the unigrams, bigrams, and trigrams** from each snippet
- **Weighting those n-grams**
  - Based on their frequency and the weight of the patterns that returned them
- **Scoring those n-grams** based on how well they match the predicted answer type
- **Concatenating overlapping n-grams** into longer answers
- **Adding the best concatenation to the list of candidate answers**, and removing lower-scoring candidates

Neural Answer Extraction

- Relies on the intuition that a question and its answer are semantically similar
- Works by:
  - Computing an embedding for the question
  - Computing an embedding for each token of the passage
  - Selecting spans from the passage whose embeddings are closest to the question embedding
- Often designed in the context of **reading comprehension**

Reading Comprehension

- A task designed to measure natural language understanding performance
- Basic premise: Take children’s reading comprehension tests, and use them to evaluate text comprehension algorithms



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

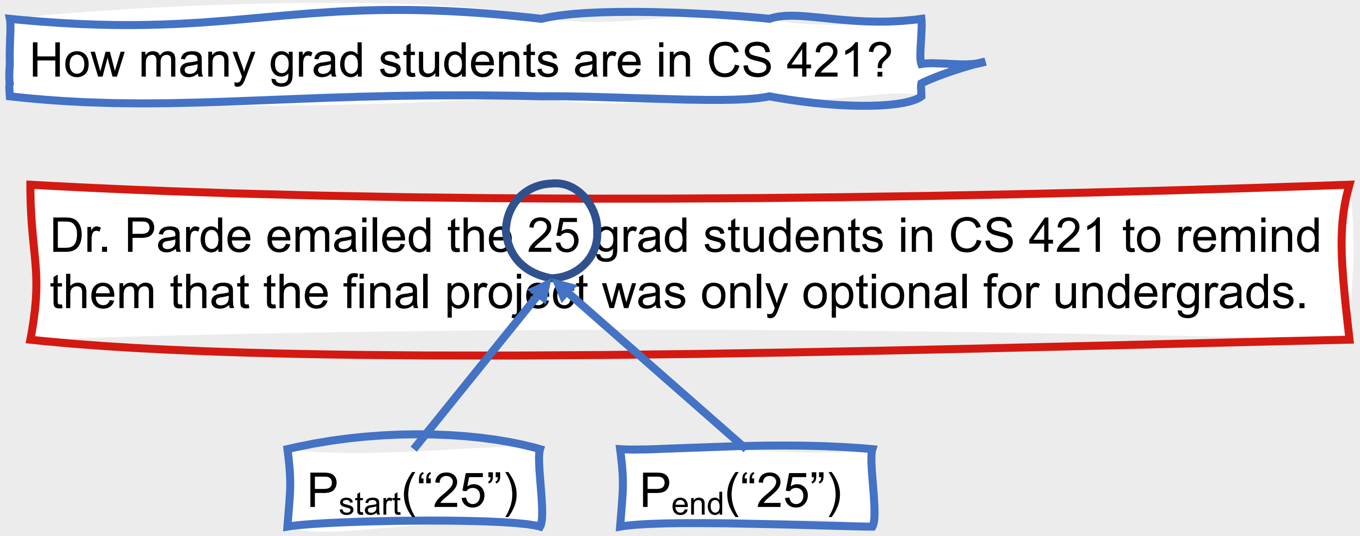
Why must one be excluded in order to preserve the uniqueness of th

Reading Comprehension Datasets

- Stanford Question Answering Dataset (SQuAD)
  - 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/
- NewsQA Dataset
  - Question-answer pairs from CNN news articles

Bidirectional LSTM-based Reading Comprehension

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



How many grad students are in CS 421?

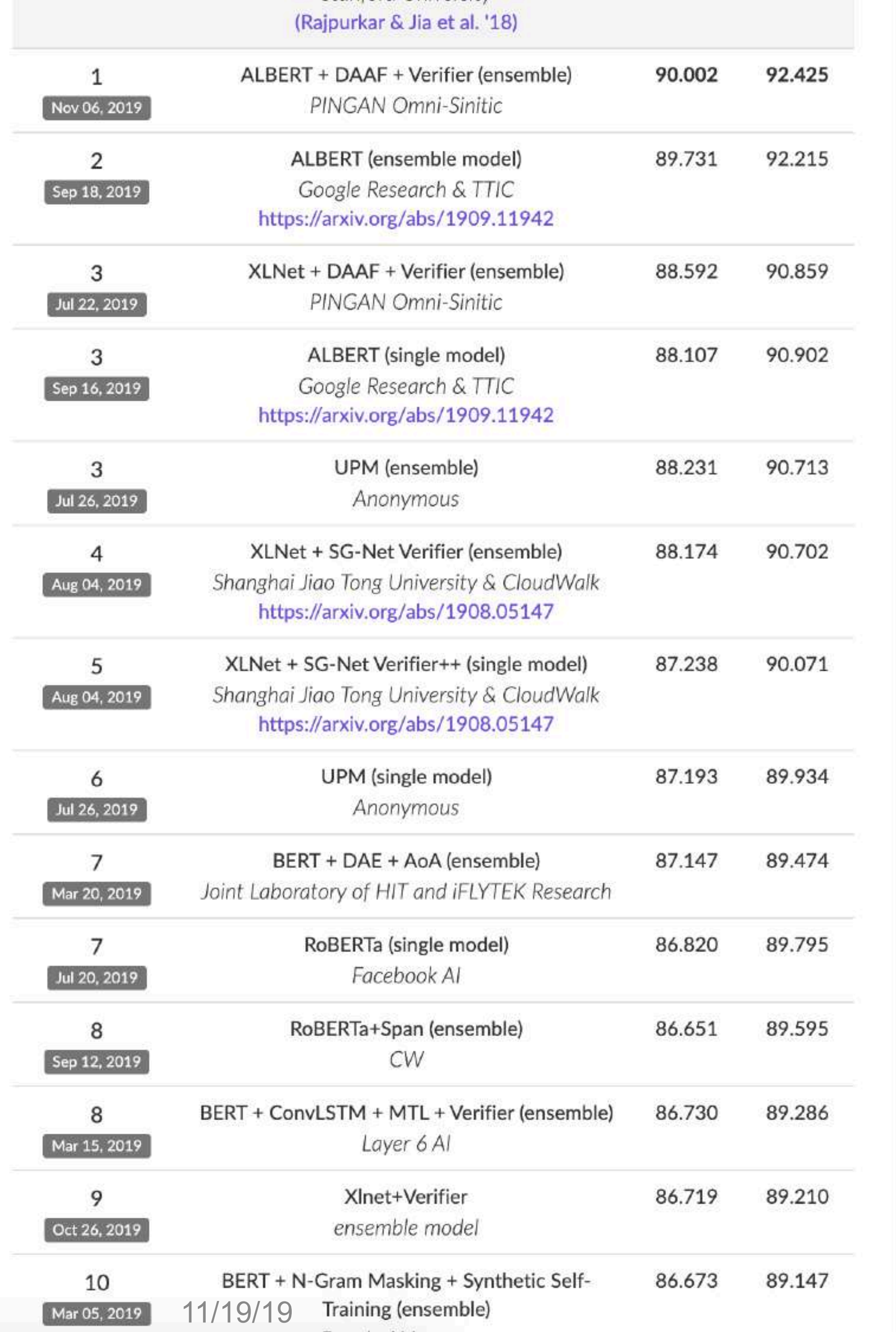
Dr. Parde emailed the 25 grad students in CS 421 to remind them that the final project was only optional for undergrads.

P_{start}(“25”)

P_{end}(“25”)

Bidirectional LSTM-based Reading Comprehension

- Learn representations for each question and each word in a passage using bidirectional LSTMs
- Learn classifiers to predict the two probabilities for each word in the passage



(Rajpurkar & Jia et al. '18)

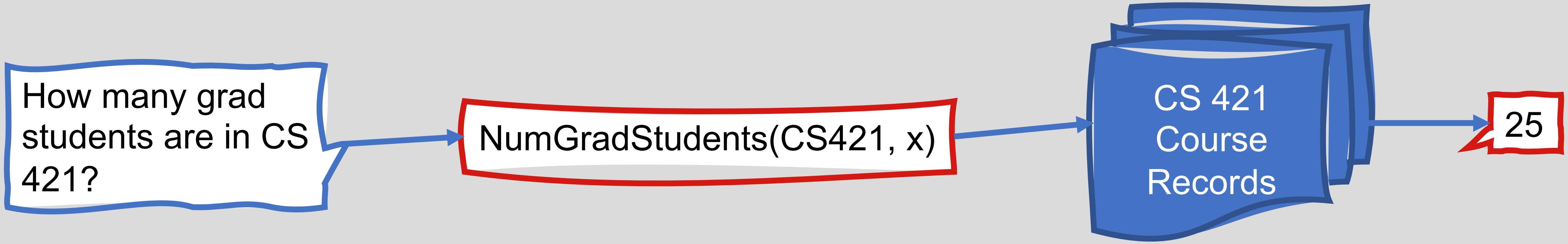
1 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.002**	**92.425**
2 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
3 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
3 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
4 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
5 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071
6 Jul 26, 2019	UPM (single model) *Anonymous*	87.193	89.934
7 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	87.147	89.474
7 Jul 20, 2019	RoBERTa (single model) *Facebook AI*	86.820	89.795
8 Sep 12, 2019	RoBERTa+Span (ensemble) *CW*	86.651	89.595
8 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
9 Oct 26, 2019	Xlnet+Verifier *ensemble model*	86.719	89.210
10 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble)	86.673	89.147

Many other neural approaches to question answering also exist!

- Many recent methods incorporate BERT embeddings
    - Contextual representations learned using Transformers

Knowledge-based Question Answering

- Answers questions by mapping them to queries over structured databases



How many grad students are in CS 421?

NumGradStudents(CS421, x)

CS 421 Course Records

25

How are text strings typically mapped to logical form?

- Semantic parsers
- Typically map text to:
    - Some form of predicate calculus (e.g., first-order logic)
    - Some type of query language
        - SQL
        - SPARQL
- This means that the question ends up either in the form of a database search query, or in a form that can be easily converted to one

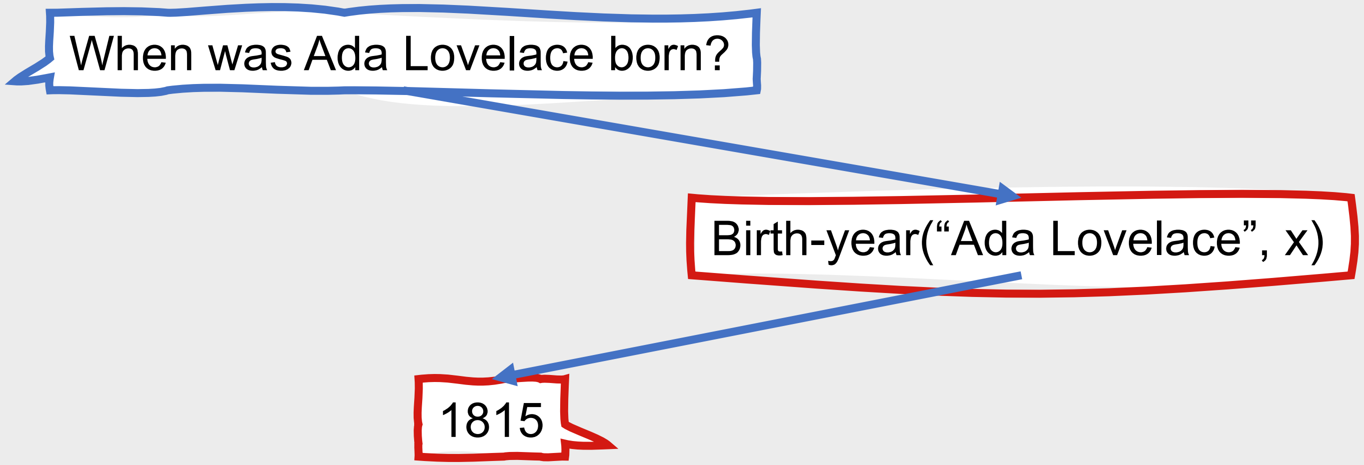
What does the database look like?

- Differs depending on the resource
- Might be:
  - Full relational database
  - Simpler structured database
    - Sets of RDF (subject, predicate, object) triples
- Popular ontologies:
  - Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page
  - DBpedia: https://wiki.dbpedia.org/

Simple Knowledge-based Question Answering Task

- Answer factoid questions that ask about one of the missing arguments in a triple

subject	predicate	object
Ada Lovelace	Birth-year	1815



When was Ada Lovelace born?

Birth-year("Ada Lovelace", x)

1815

Rule-based Methods for Knowledge-based Question Answering

Write patterns to extract frequent relations

- When .+ born → birth-year

Pros:

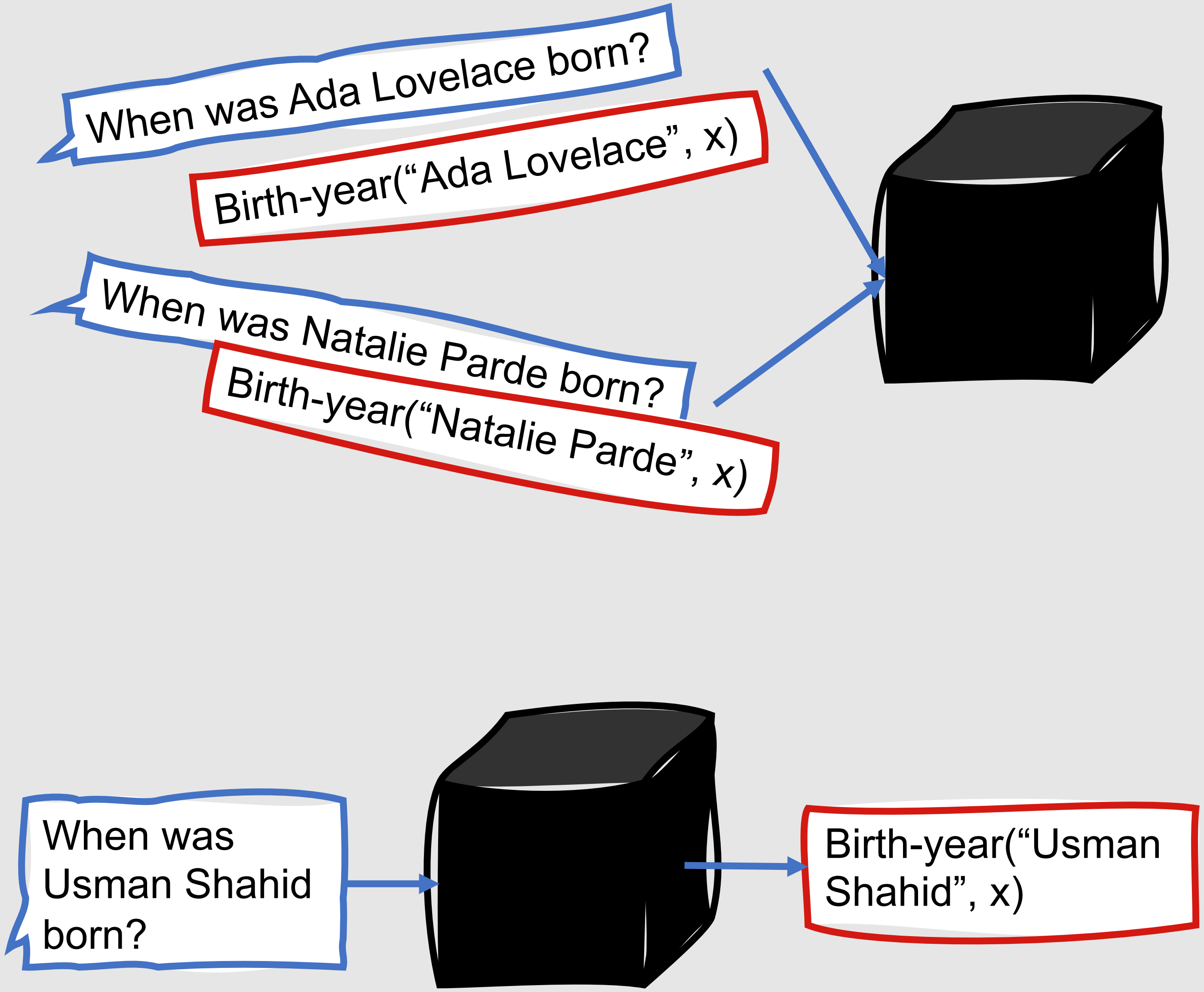
- Simple
- Precise

Cons:

- Not scalable
- Low recall

Supervised Methods for Knowledge-based Question Answering

- Learn from pairs of training questions and their correct logical forms
- Produce a system that maps from new questions to their logical forms


When was Ada Lovelace born?

Birth-year(“Ada Lovelace”, x)

When was Natalie Parde born?

Birth-year(“Natalie Parde”, x)

When was Usman Shahid born?

Birth-year(“Usman Shahid”, x)

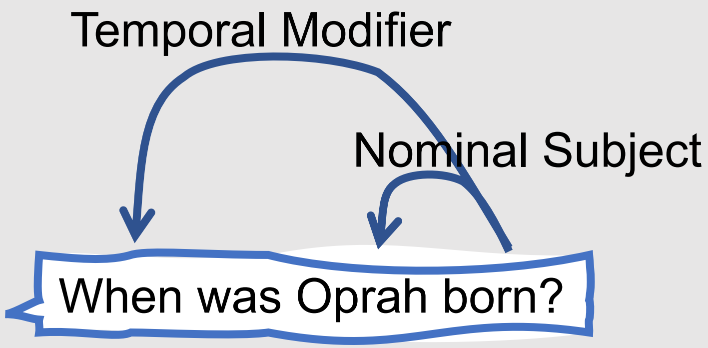
How do most systems do this?

- First, parse the questions
- Then, align the parse trees to a logical form
- Often employ **bootstrapping**
    - Small set of rules for building the mapping
    - Small initial lexicon

What would this look like?

When was Oprah born?

What would this look like?

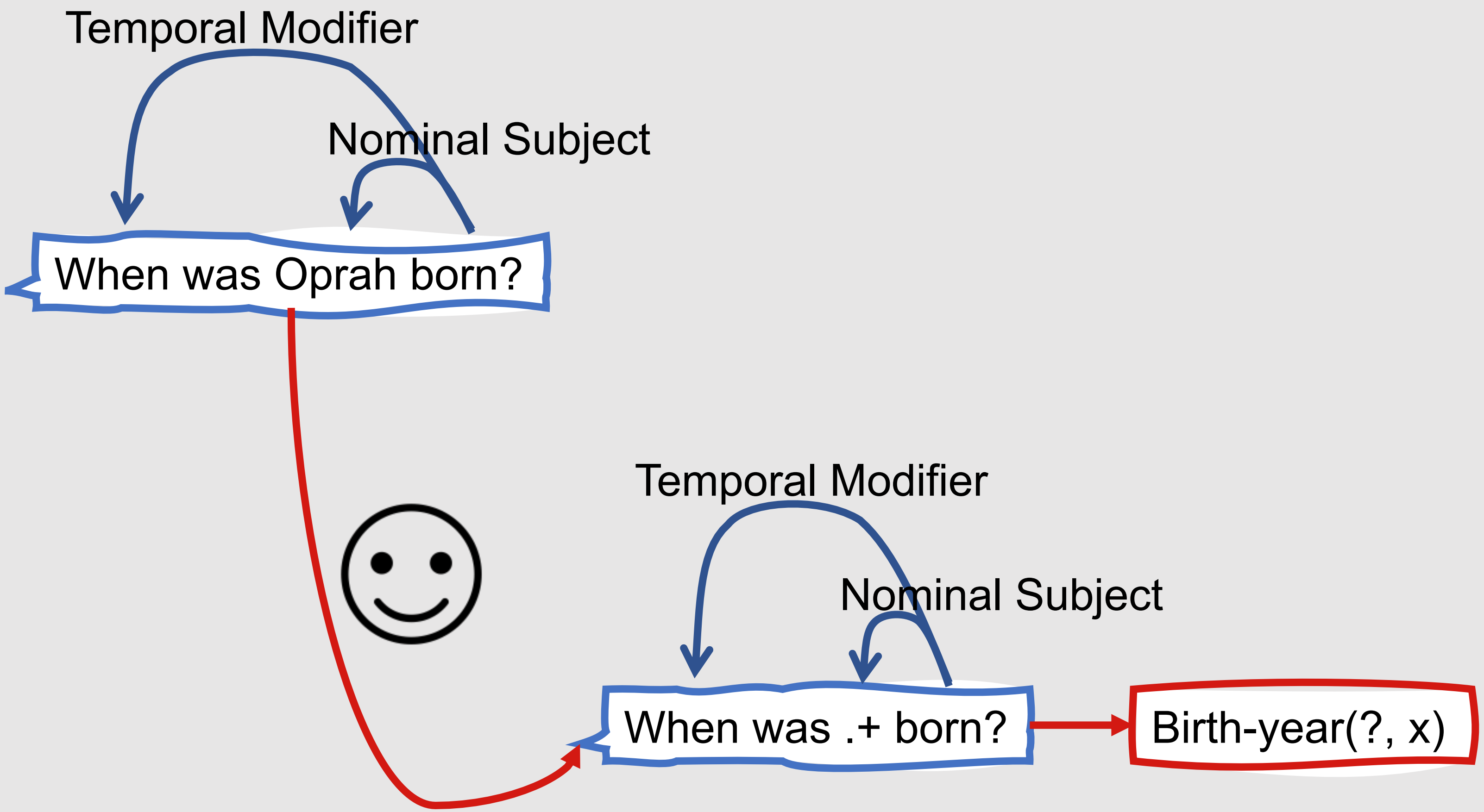


Temporal Modifier

Nominal Subject

When was Oprah born?

What would this look like?



Temporal Modifier

Nominal Subject

When was Oprah born?

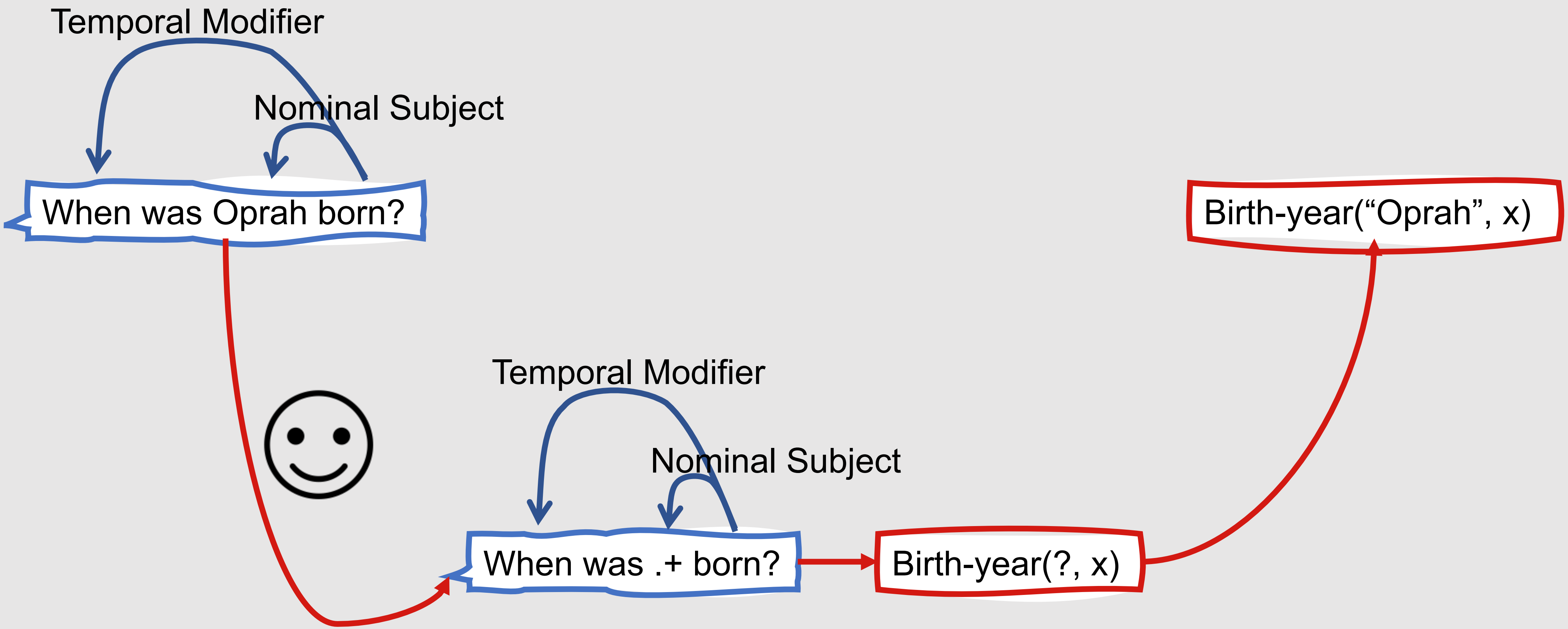
Temporal Modifier

Nominal Subject

When was .+ born?

Birth-year(?, x)

What would this look like?



Temporal Modifier

Nominal Subject

When was Oprah born?

Temporal Modifier

Nominal Subject

When was .+ born?

Birth-year(?, x)

Birth-year(“Oprah”, x)

Supervised approaches can be extended to handle more complex questions.

- More complex default rules can be used
- More complex logical forms can be used
- Training samples can be broken down into smaller tuples and then recombined to parse new sentences

What is the biggest state bordering Illinois?

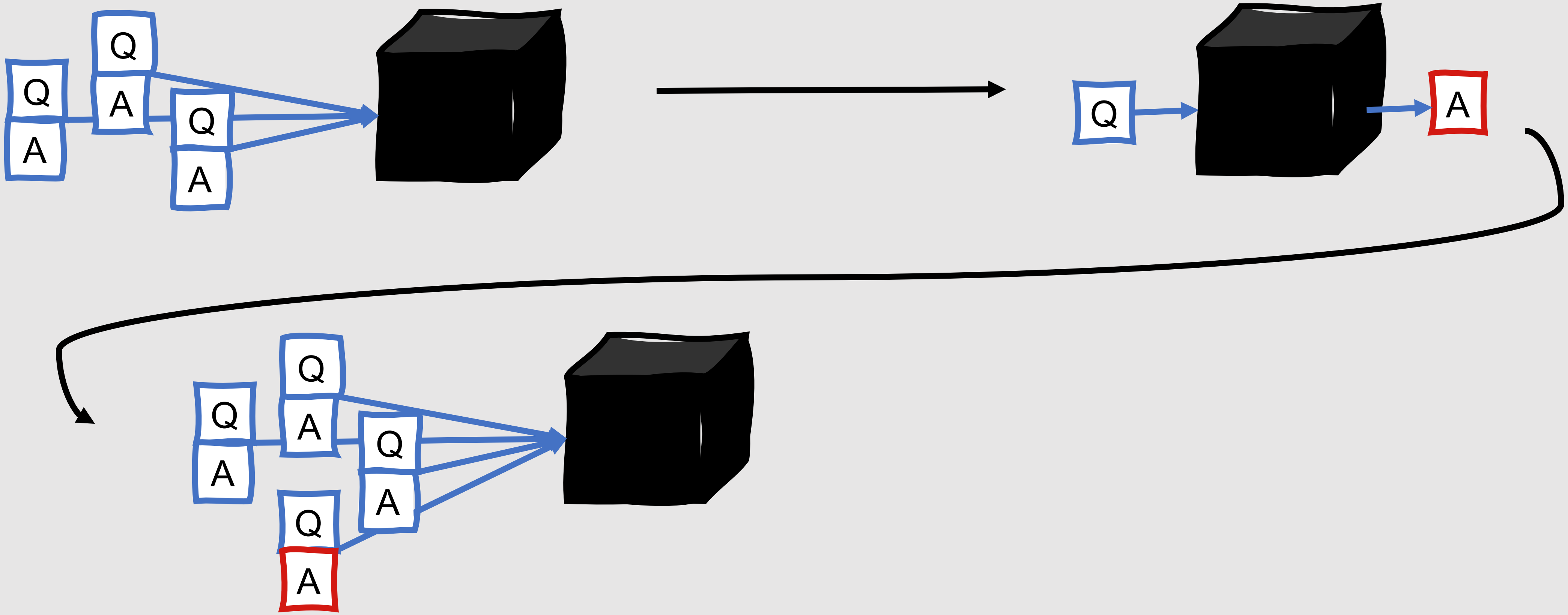
How many more undergrads are there than grad students in CS 421?

Is Chicago closer to Dallas or Denver?

Semi-Supervised Methods for Knowledge-based Question Answering

- What is semi-supervised learning?
    - A form of machine learning that makes use of both labeled and unlabeled data for training
- Example: Bootstrapping

Semi-Supervised Methods for Knowledge-based Question Answering



Why used semi-supervised learning?

- Even though factoid questions may seem simple, it is difficult to build supervised datasets that comprehensively cover all of their different forms!

When was Oprah born?

What is Oprah's birth year?

What year was Oprah born?

In what year was Oprah born?

Semi-supervised methods allow us to efficiently make use of textual redundancy.

phrase	relation	phrase	relation	phrase	relation
Capital of	Country.capital	Capital city of	Country.capital	Become capital of	Country.capital
Capitol of	Country.capital	National capital of	Country.capital	Official capital of	Country.capital
Political capital of	Country.capital	Administrative capital of	Country.capital	Beautiful capital of	Country.capital
Capitol city of	Country.capital	Remain capital of	Country.capital	Make capital of	Country.capital
Political center of	Country.capital	Bustling capital of	Country.capital	Capital city in	Country.capital
Cosmopolitan capital of	Country.capital	Move its capital to	Country.capital	Modern capital of	Country.capital
Federal capital of	Country.capital	Beautiful capital city of	Country.capital	Administrative capital city of	Country.capital

Combining Information Sources

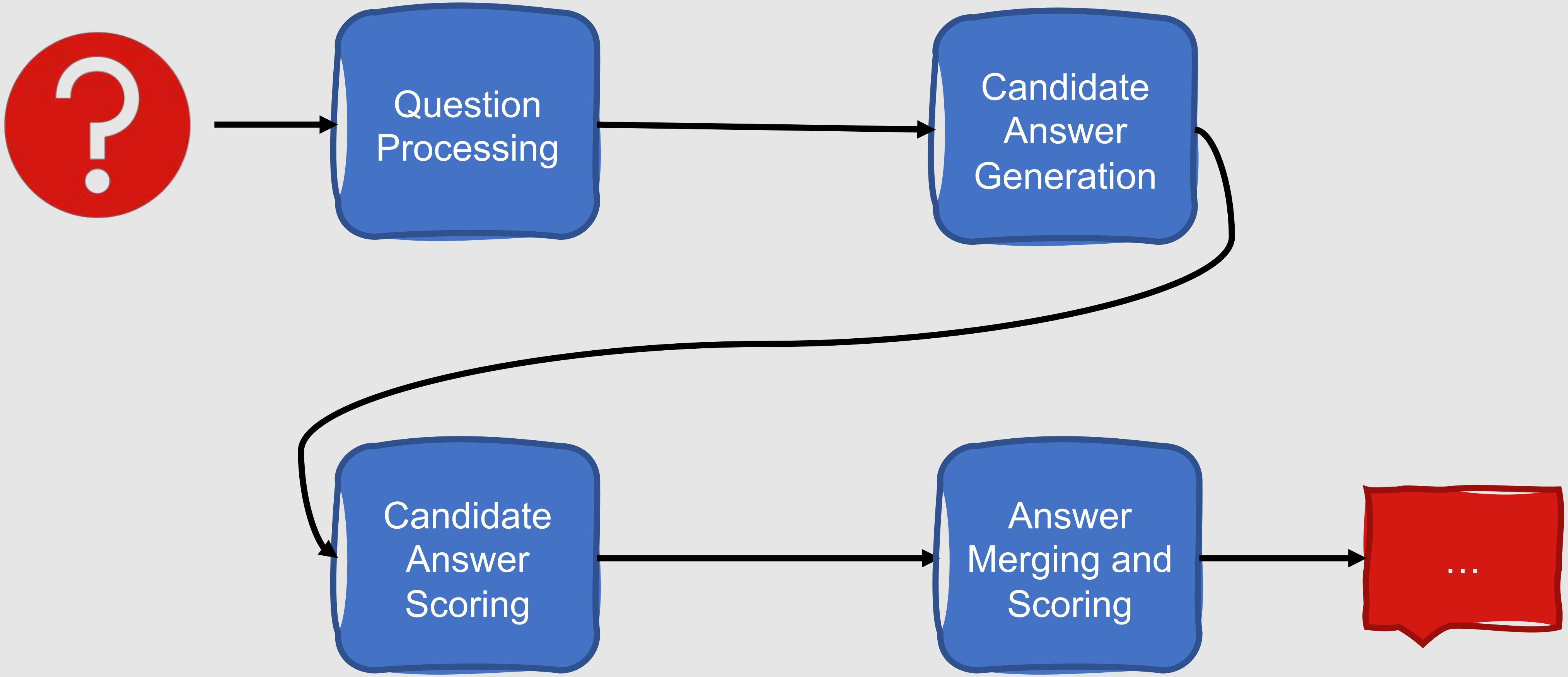
Remember …there’s no need to limit a system to using *only* text-based or *only* knowledge-based methods!

Many high-performing systems combine these two information sources

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



Question Processing

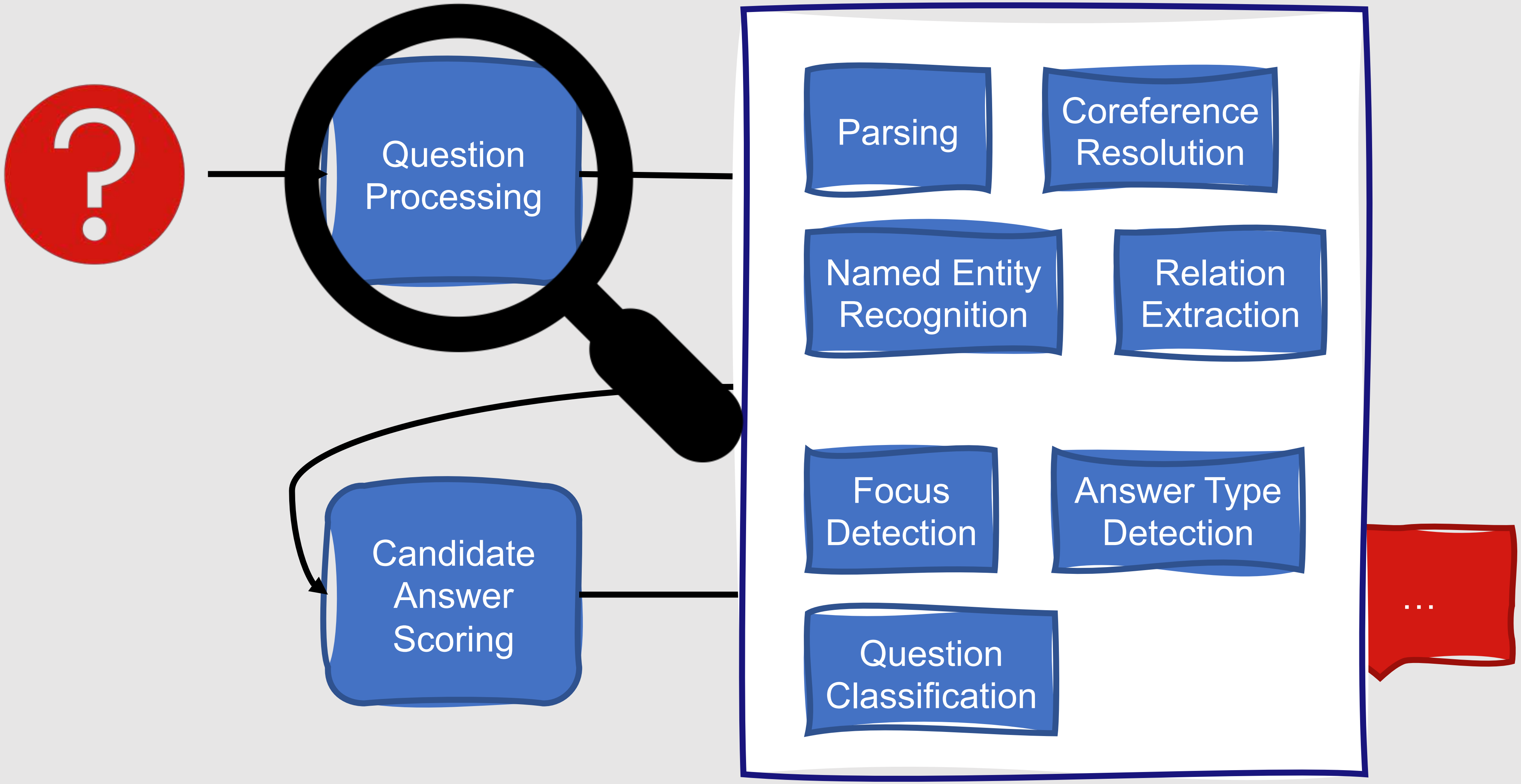
Candidate Answer Generation

Candidate Answer Scoring

Answer Merging and Scoring

…

Stage 1: Question Preprocessing



Question Processing

Candidate Answer Scoring

Parsing

Coreference Resolution

Named Entity Recognition

Relation Extraction

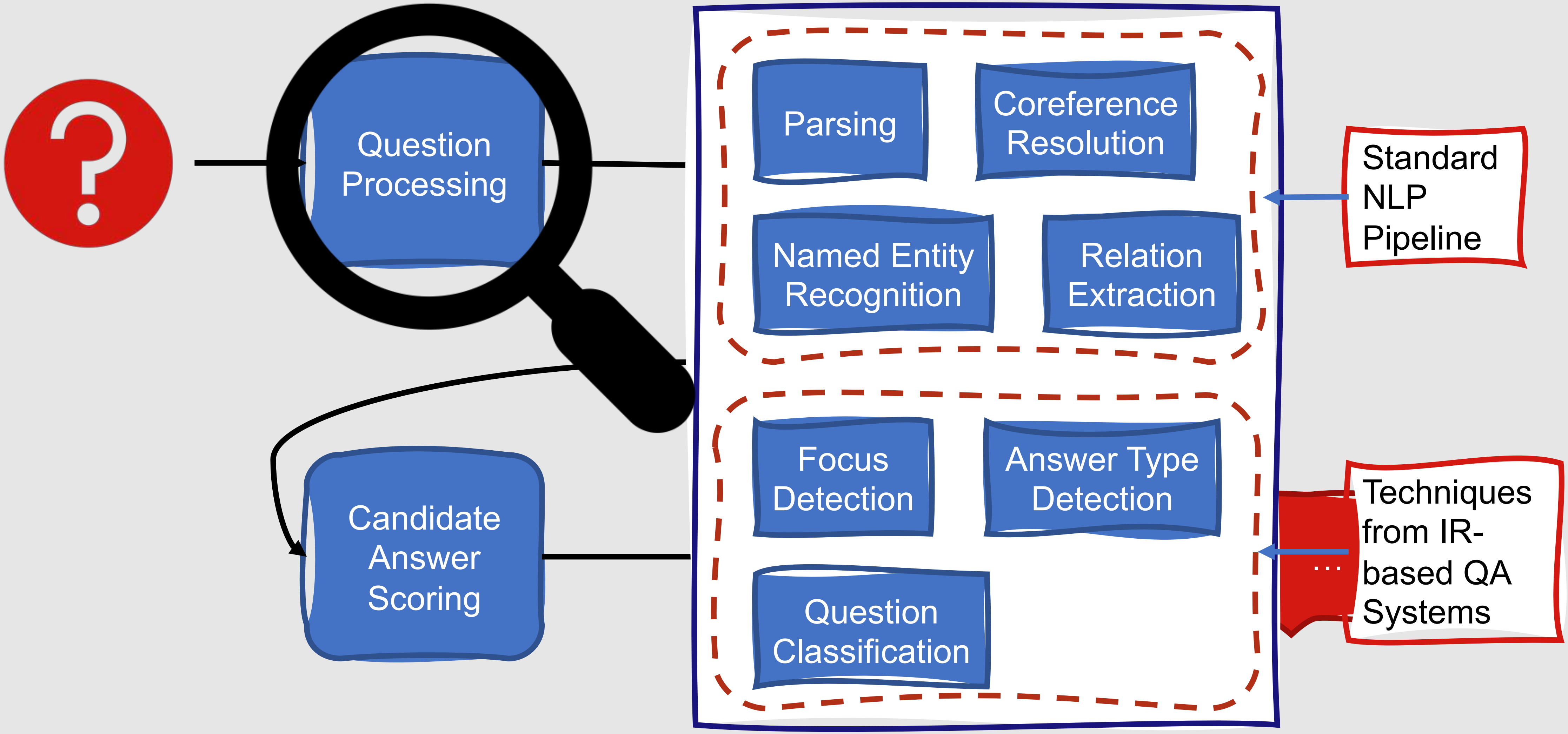
Focus Detection

Answer Type Detection

Question Classification

…

Stage 1: Question Preprocessing



Question Processing

Candidate Answer Scoring

Parsing

Coreference Resolution

Named Entity Recognition

Relation Extraction

Standard NLP Pipeline

Focus Detection

Answer Type Detection

Question Classification

…

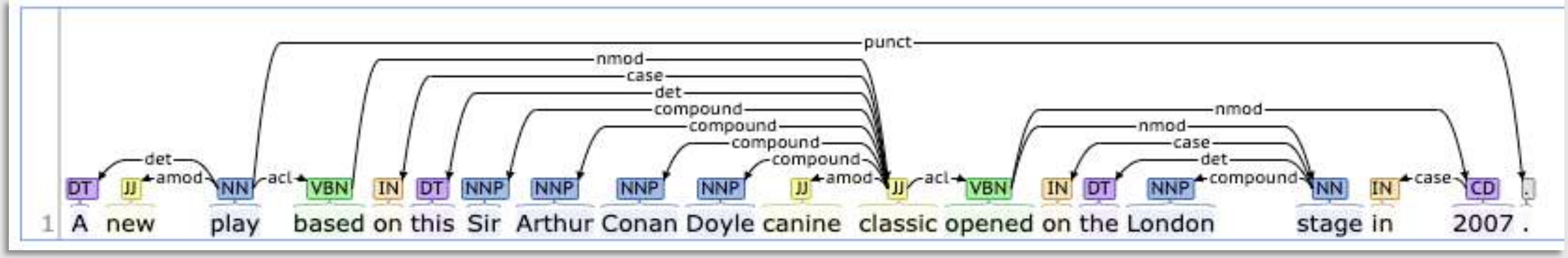
Techniques from IR-based QA Systems

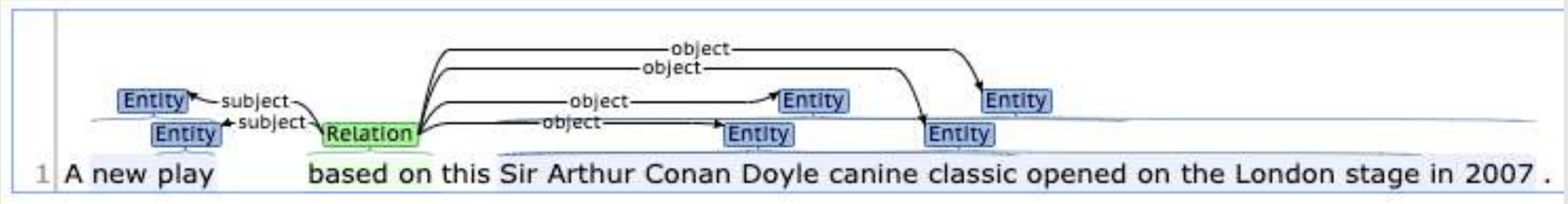
Stage 1: Question Preprocessing

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

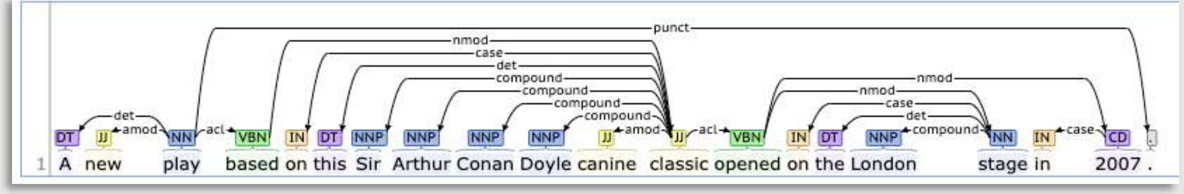
Jeopardy! Example:
A new play based on this Sir Arthur Conan Doyle canine classic opened on the London stage in 2007.





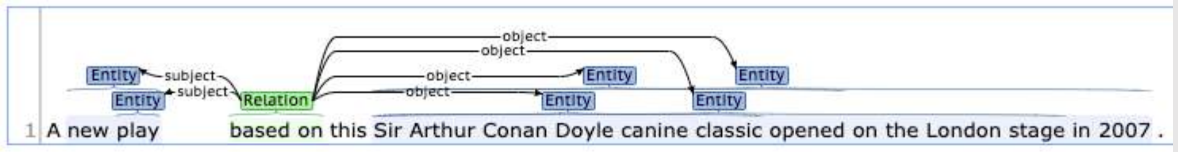
Stage 1: Question Preprocessing

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



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

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



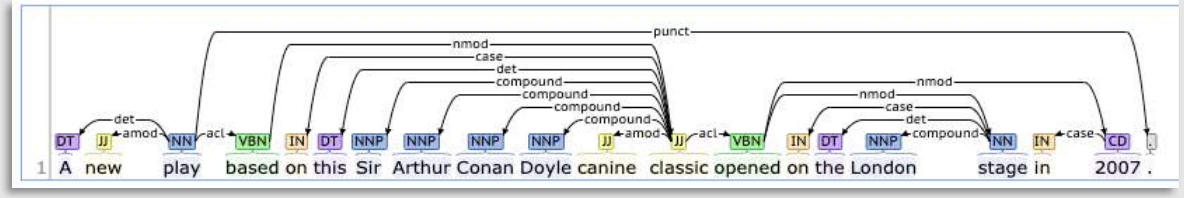
A new play based on this Sir Arthur Conan Doyle canine classic opened on the London stage in 2007 .

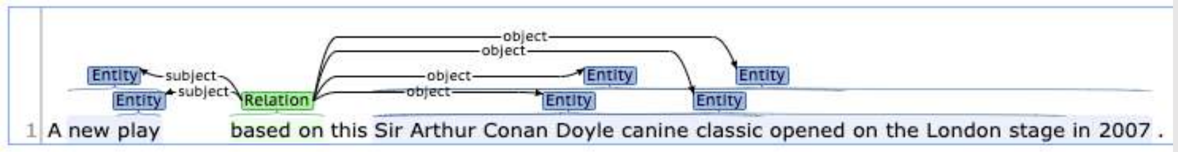
Focus Detection: Which part of the question co-refers with the answer?

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.



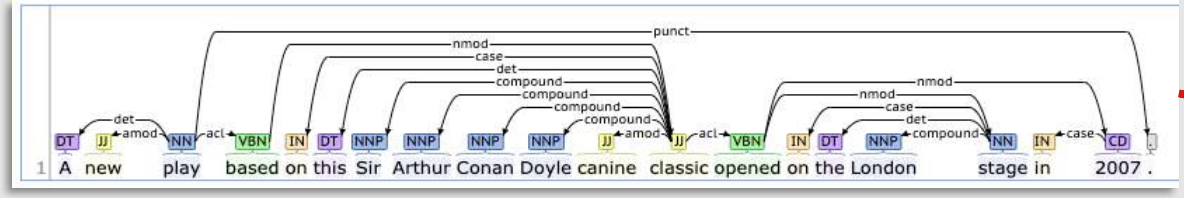


Answer Type Detection: Which word tells us about the semantic type of answer to expect?

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

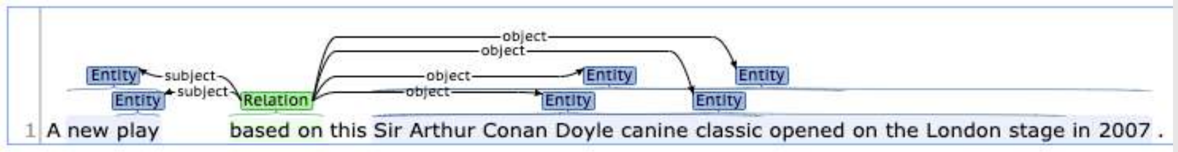
Stage 1: Question Preprocessing

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

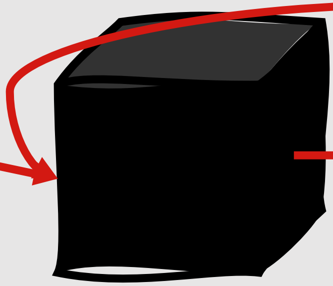


1 A new play based on this Sir Arthur Conan Doyle canine classic opened on the London stage in 2007 .

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



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

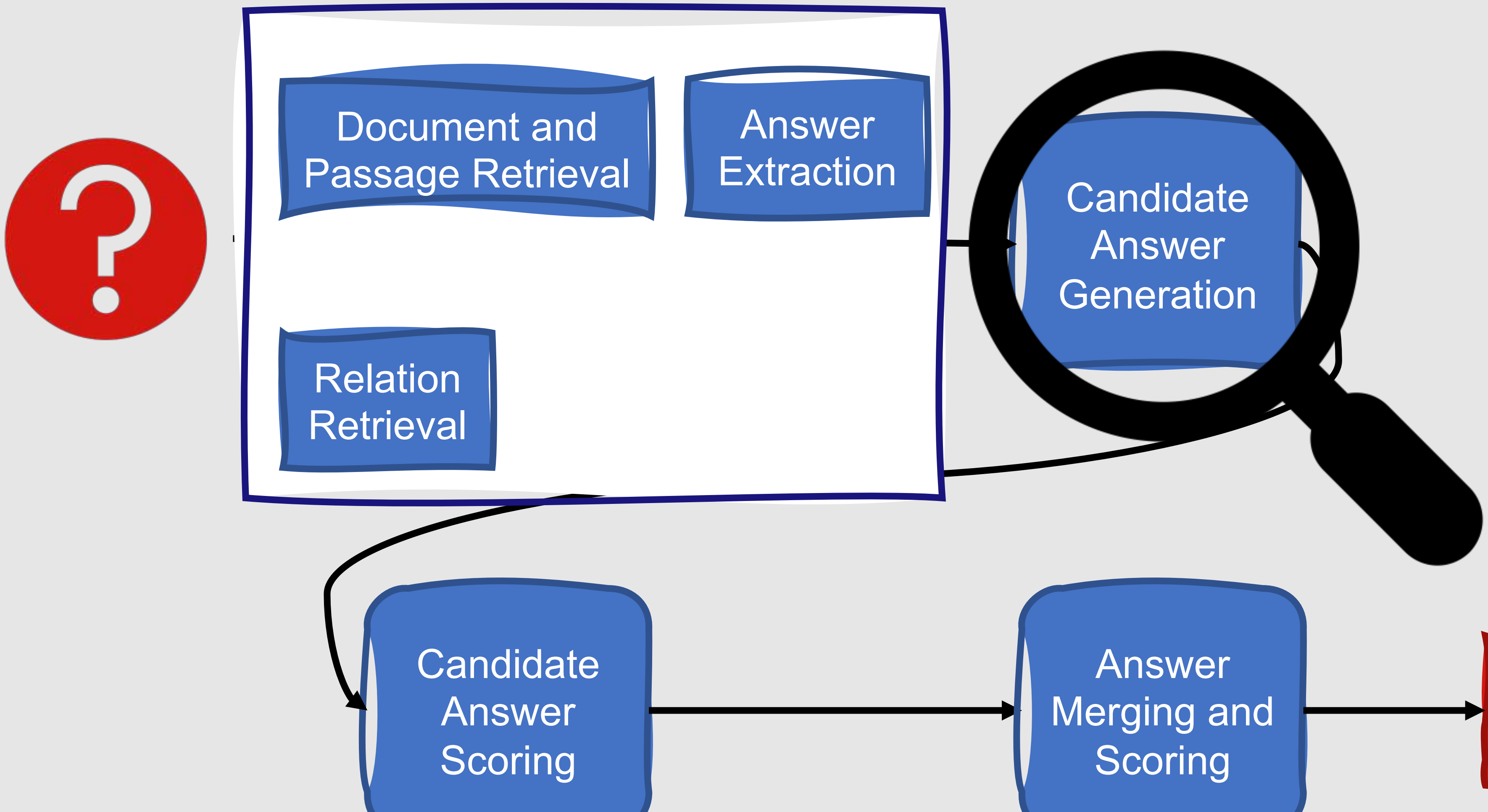


Definition

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



Document and Passage Retrieval

Answer Extraction

Relation Retrieval

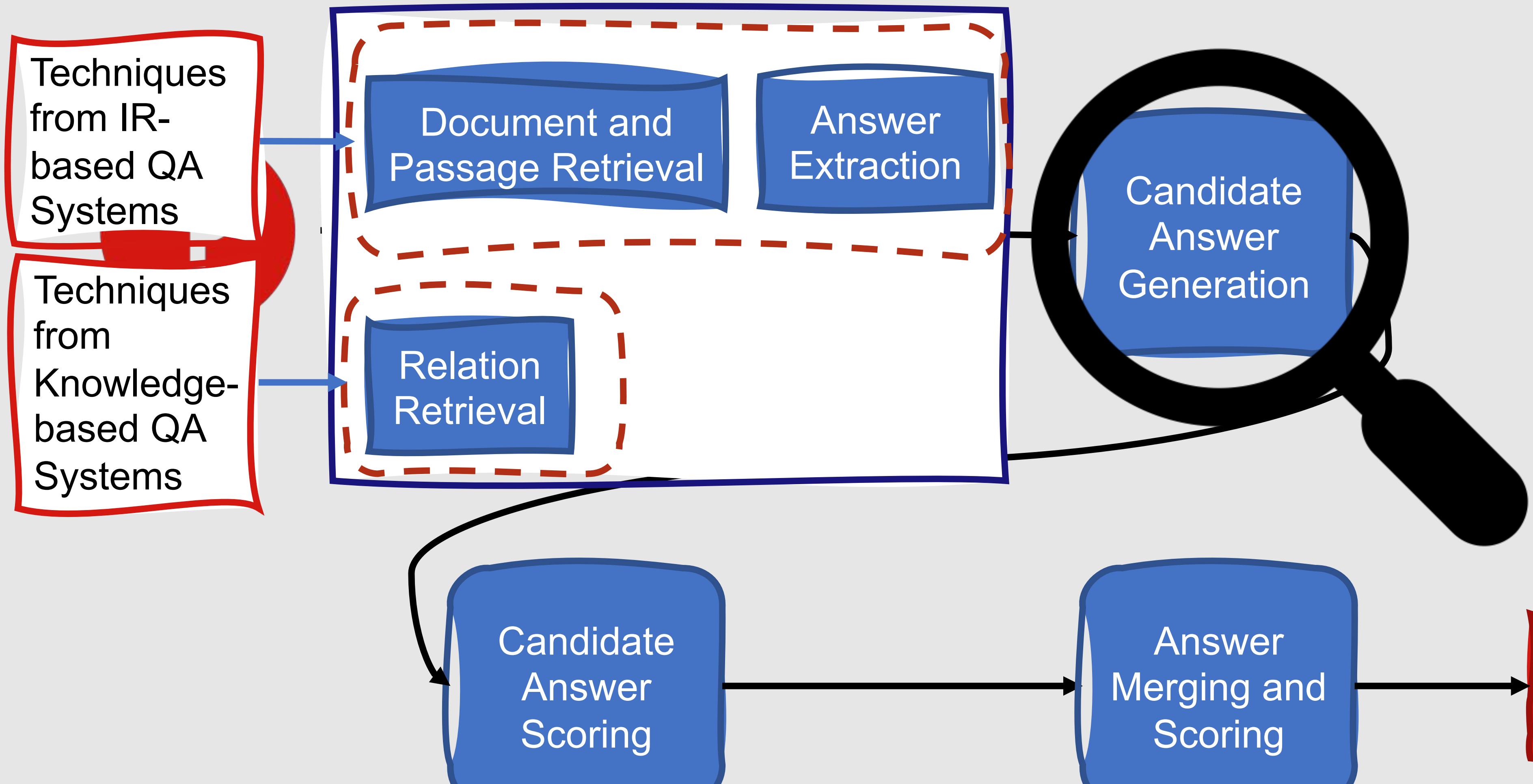
Candidate Answer Generation

Candidate Answer Scoring

Answer Merging and Scoring

…

Stage 2: Candidate Answer Generation



Techniques from IR-based QA Systems

Techniques from Knowledge-based QA Systems

Document and Passage Retrieval

Answer Extraction

Relation Retrieval

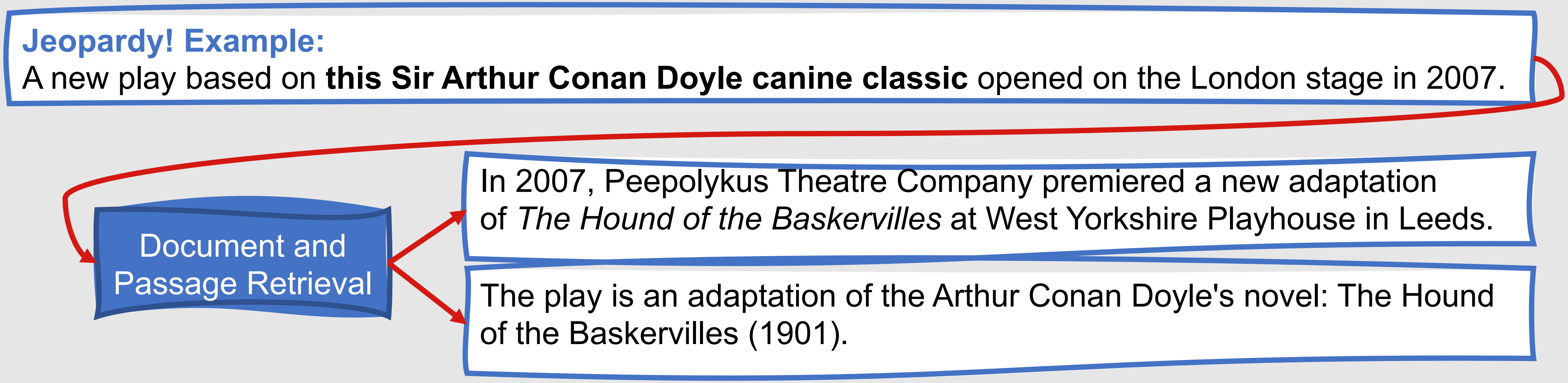
Candidate Answer Generation

Candidate Answer Scoring

Answer Merging and Scoring

…

Stage 2: Candidate Answer Generation



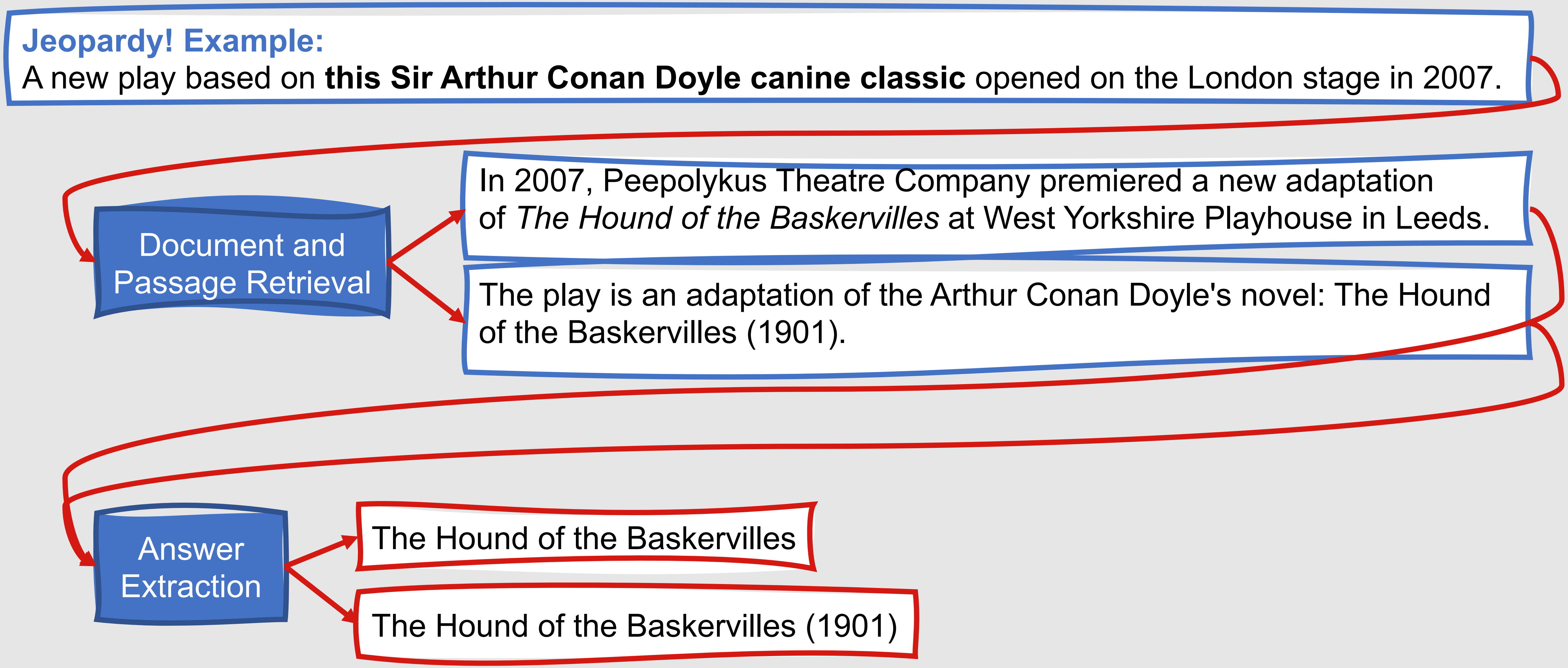
Jeopardy! Example:
A new play based on **this Sir Arthur Conan Doyle canine classic** opened on the London stage in 2007.

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.

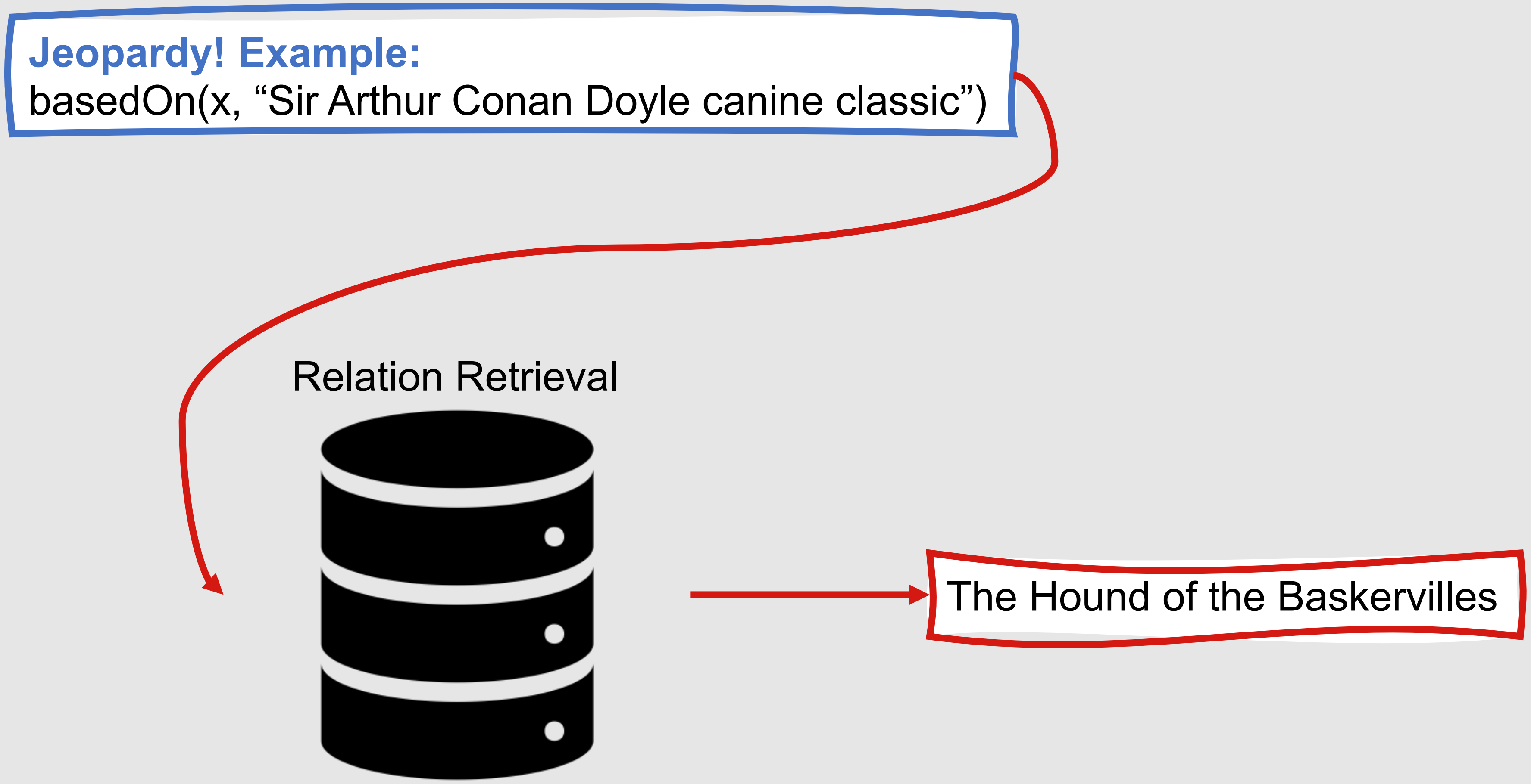
The play is an adaptation of the Arthur Conan Doyle's novel: The Hound of the Baskervilles (1901).

Answer Extraction

The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

Stage 2: Candidate Answer Generation

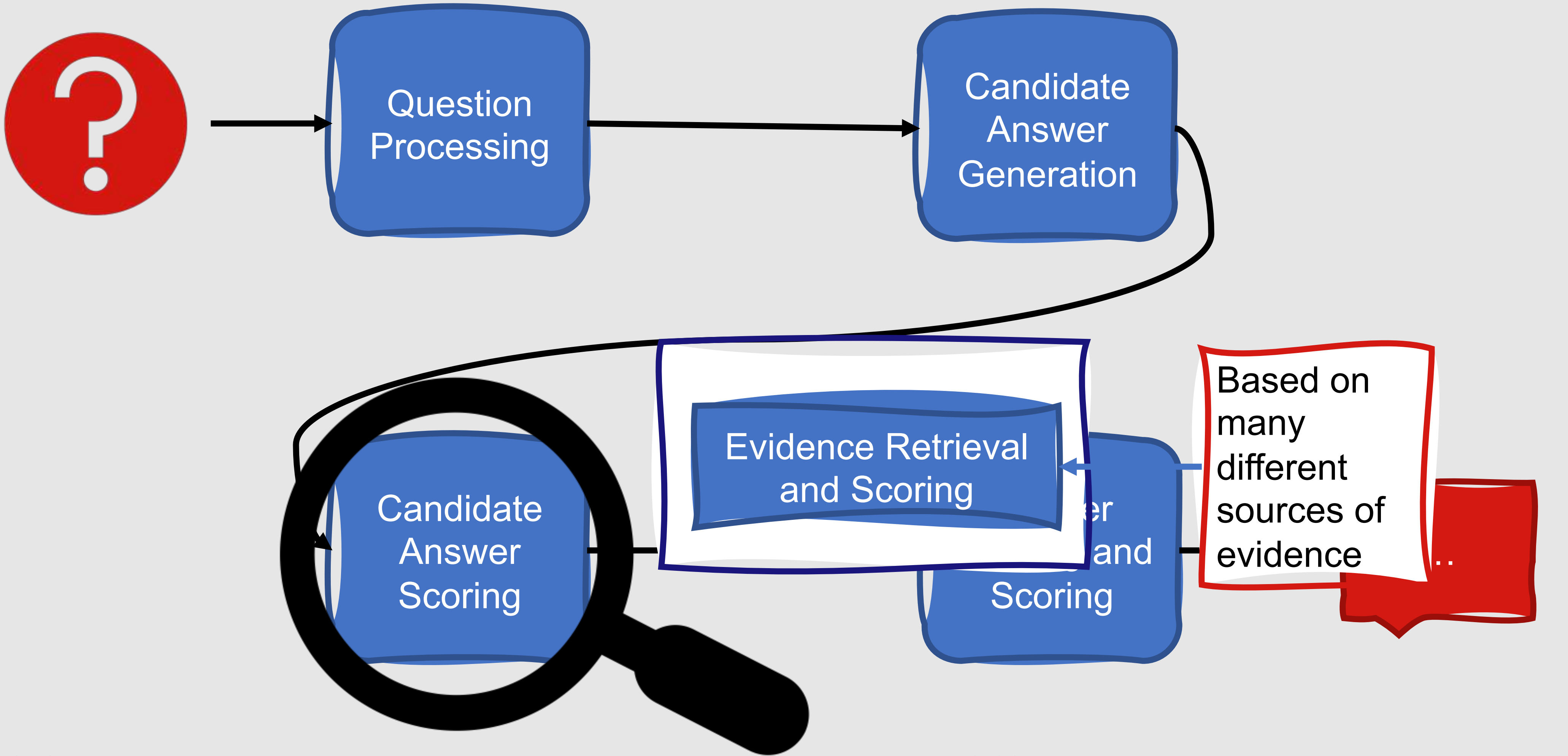


Jeopardy! Example:
basedOn(x, “Sir Arthur Conan Doyle canine classic”)

Relation Retrieval

The Hound of the Baskervilles

Stage 3: Candidate Answer Scoring



Question Processing

Candidate Answer Generation

Candidate Answer Scoring

Evidence Retrieval and Scoring

er
and
Scoring

Based on many different sources of evidence

..

Stage 3: Candidate Answer Scoring

The Hound of the Baskervilles

The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

Stage 3: Candidate Answer Scoring

The Hound of the Baskervilles

The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

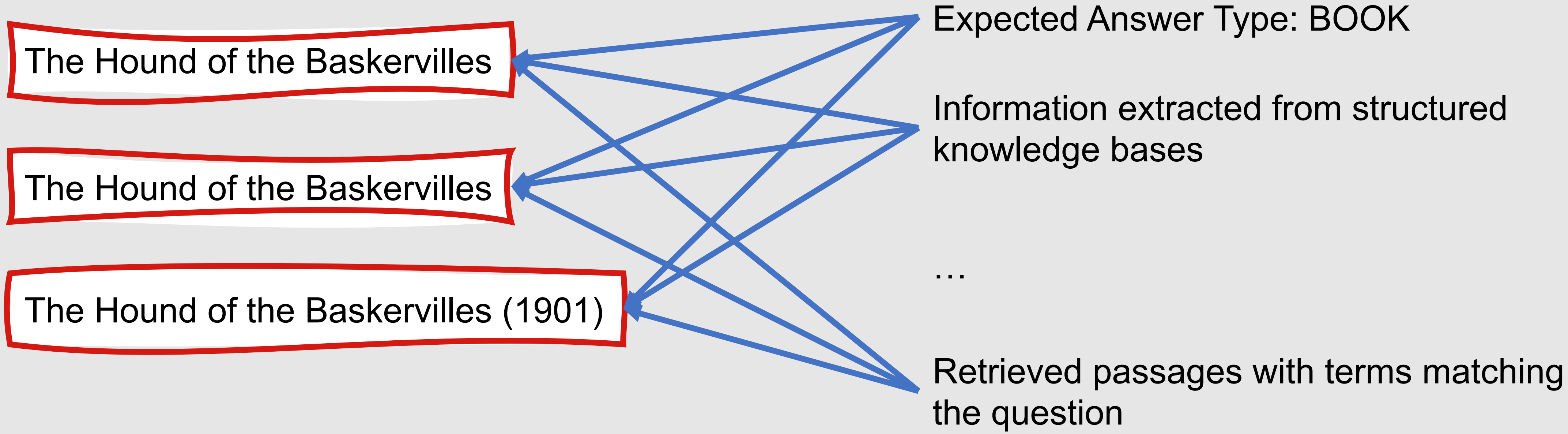
Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Stage 3: Candidate Answer Scoring



The Hound of the Baskervilles

The Hound of the Baskervilles

The Hound of the Baskervilles (1901)

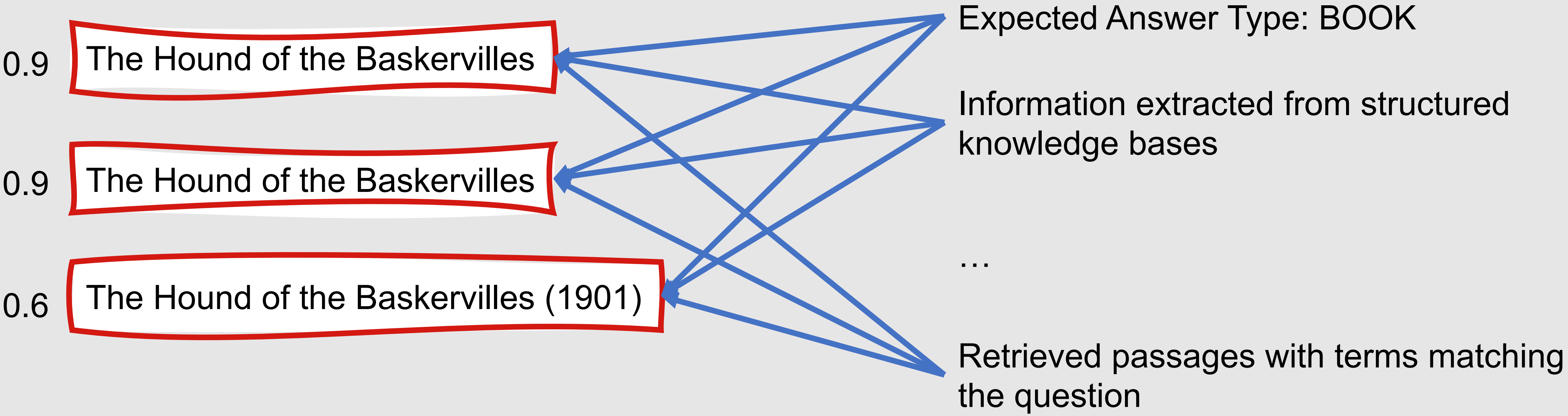
Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Stage 3: Candidate Answer Scoring



0.9 The Hound of the Baskervilles

0.9 The Hound of the Baskervilles

0.6 The Hound of the Baskervilles (1901)

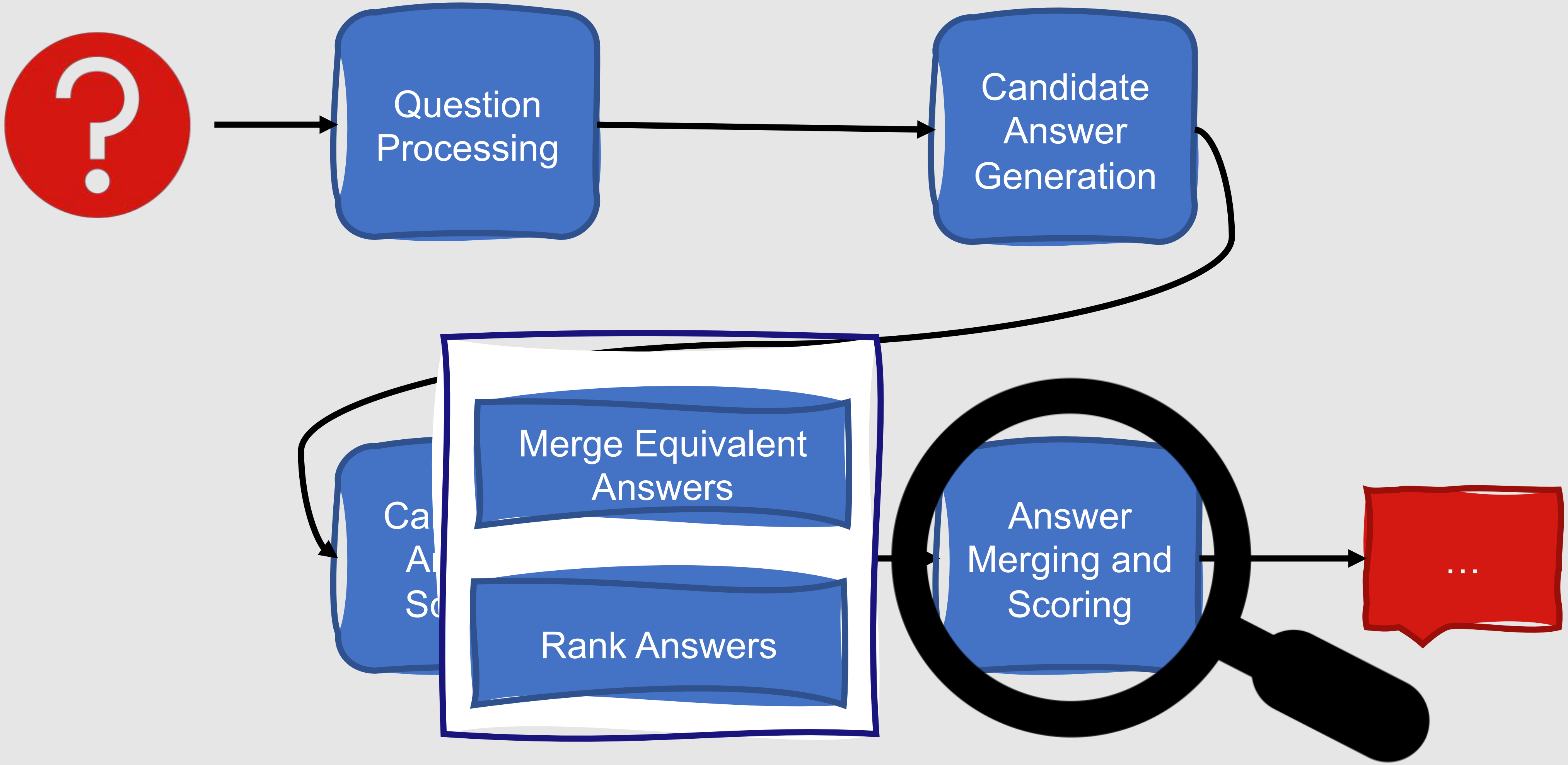
Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Stage 4: Answer Merging and Scoring



Question Processing

Candidate Answer Generation

Ca
A
Sc

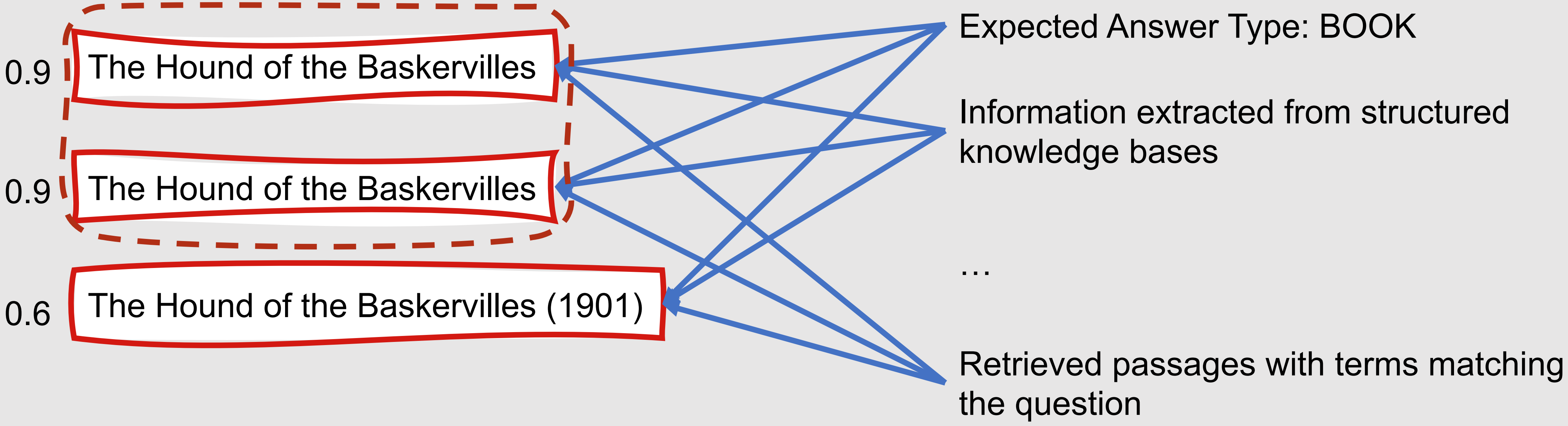
Merge Equivalent Answers

Rank Answers

Answer Merging and Scoring

…

Stage 4: Answer Merging and Scoring



0.9 The Hound of the Baskervilles

0.9 The Hound of the Baskervilles

0.6 The Hound of the Baskervilles (1901)

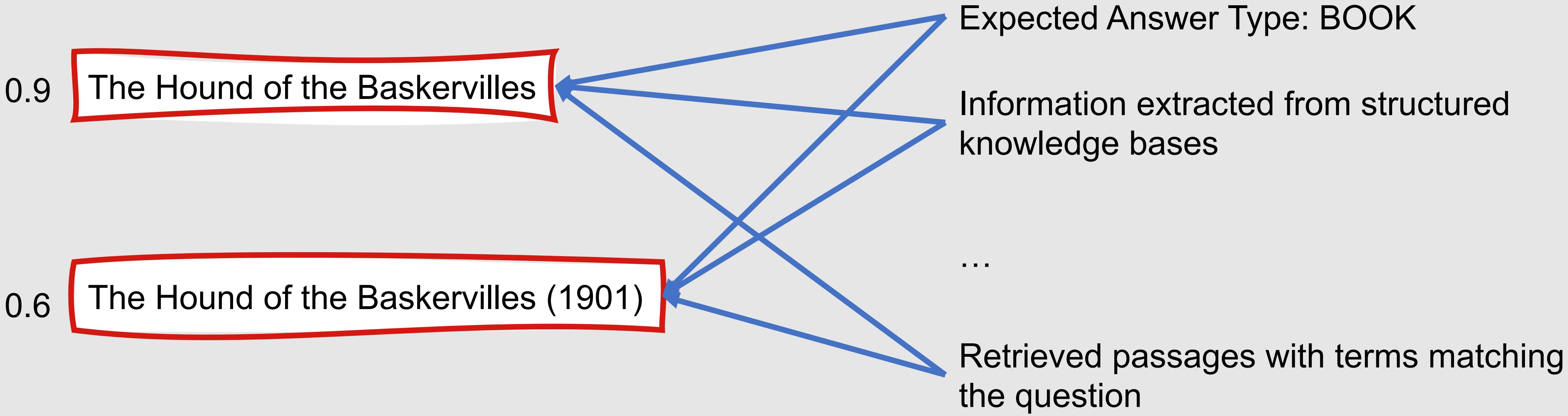
Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Stage 4: Answer Merging and Scoring



0.9 The Hound of the Baskervilles

0.6 The Hound of the Baskervilles (1901)

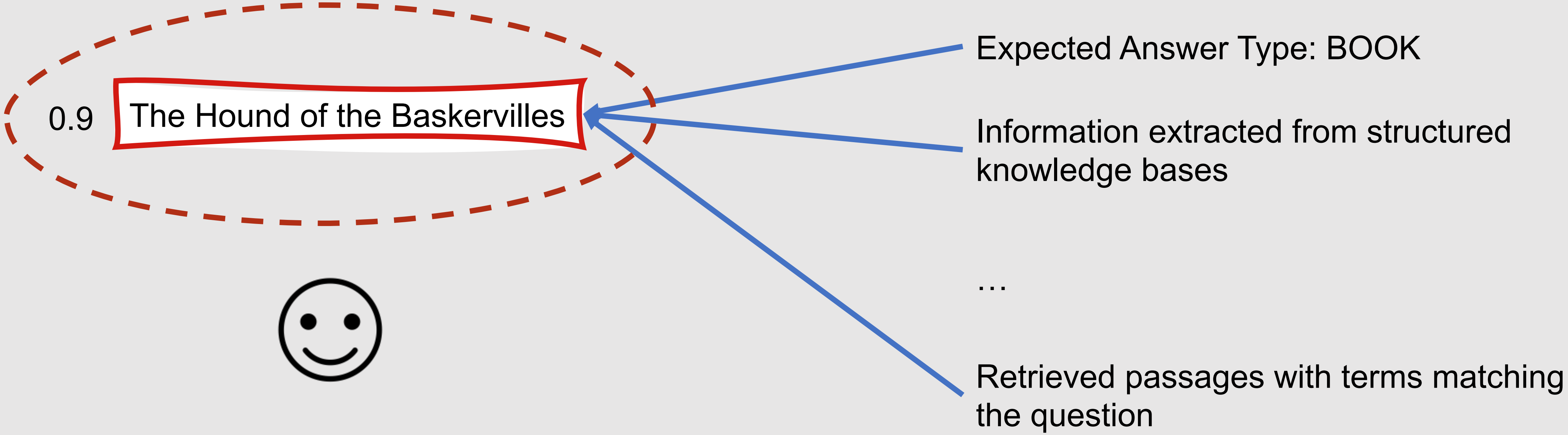
Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Stage 4: Answer Merging and Scoring



0.9 The Hound of the Baskervilles

☺

Expected Answer Type: BOOK

Information extracted from structured knowledge bases

…

Retrieved passages with terms matching the question

Watson is just one of many question answering architectures!

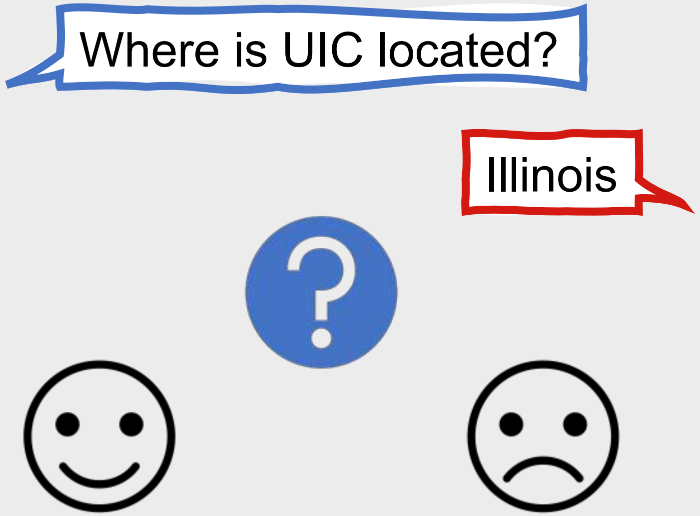
- Most high-performing QA systems will follow the same intuition:
  - Propose a large number of candidate answers using both IR-based and knowledge-based techniques
  - Develop a variety of IR-based and knowledge-based features to score the candidates

Summary: Question Answering (Part 1)

- **Question answering** is the process of automatically retrieving short spans of correct, relevant information in response to a user's **query**
- Most question answering systems focus on **factoid** questions
- There are two major types of question answering systems:
  - **Information retrieval-based**
  - **Knowledge-based**
- These two types of question answering systems are often combined, as seen in Watson's DeepQA architecture

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
    - $\text{MRR} = \frac{1}{N} \sum_{i=1 \text{ s.t. } rank_i \neq 0}^{N} \frac{1}{rank_i}$

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

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 → Robert Sloan

Prediction	Rank
Peter Nelson	1
Robert Sloan	2
Natalie Parde	3
Usman Shahid	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 → Robert Sloan

Prediction	Rank
Peter Nelson	1
Robert Sloan	2
Natalie Parde	3
Usman Shahid	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 → Robert Sloan

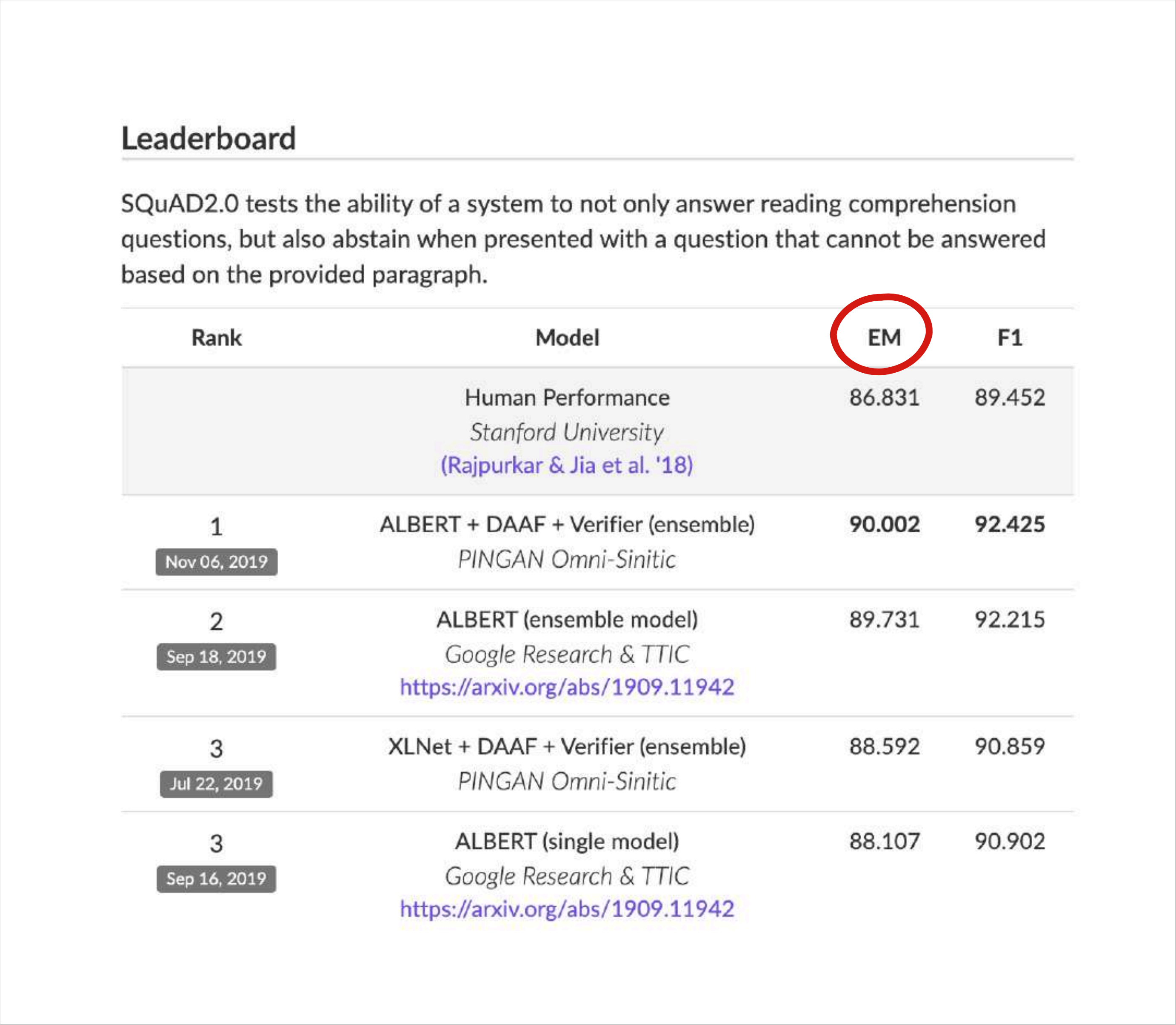
Prediction	Rank
Peter Nelson	1
Robert Sloan	2
Natalie Parde	3
Usman Shahid	4

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



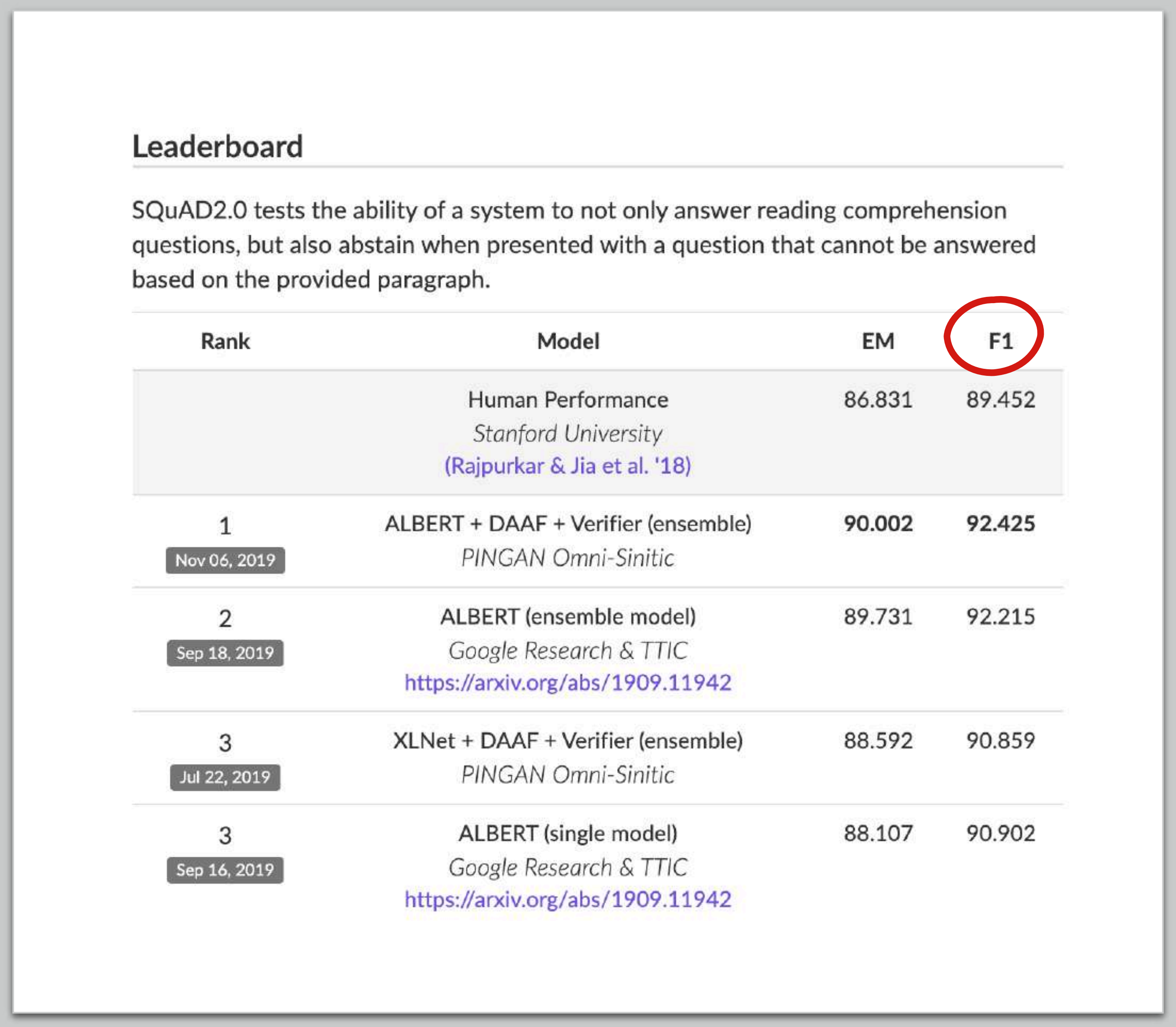
Leaderboard

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 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.002**	**92.425**
2 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
3 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902

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 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	**90.002**	**92.425**
2 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
3 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902

Computing F_1 for Question Answering Systems

Where is UIC located? ← Question

Gold Standard → Chicago

Prediction → Chicago, Illinois

	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

More Question Answering Datasets

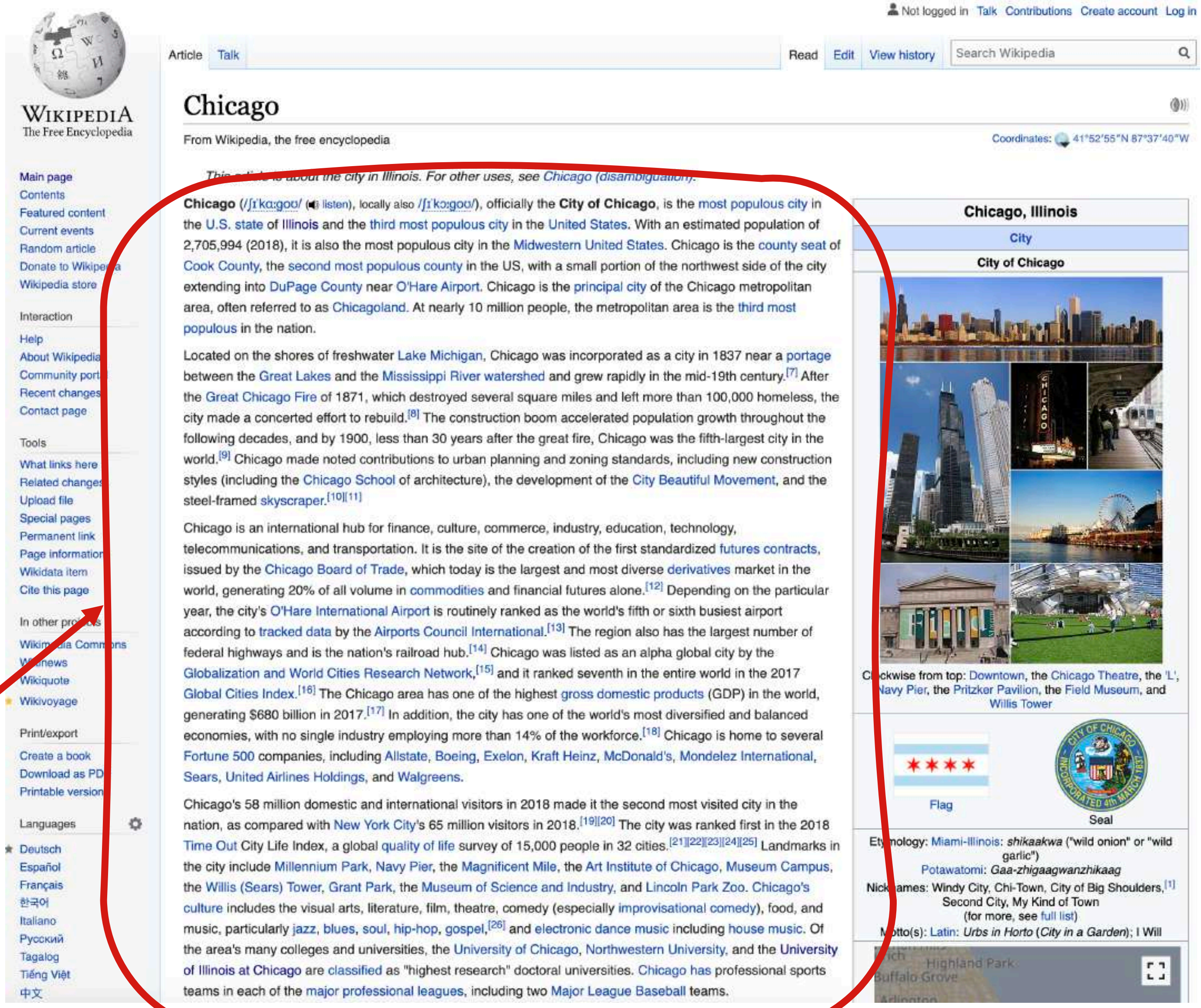
TREC QA Dataset	https://trec.nist.gov/data/qa.html
TriviaQA Dataset	https://nlp.cs.washington.edu/triviaqa/
WebQuestions Dataset	https://worksheets.codalab.org/worksheets/0xba659fe363cb46e7a505c5b6a774dc8a
NarrativeQA Dataset	https://github.com/deepmind/narrativeqa
Question Answering in Context Dataset	https://quac.ai/
MCTest Dataset	https://github.com/mcobzarenco/mctest/tree/master/data/MCTest
AI2 Reasoning Challenge	http://data.allenai.org/arc/

What is text summarization?

- The process of **automatically extracting the most important information** from a text to create an abridged version of it

Summarization

Chicago is one of the largest cities in the United States. It is located in Illinois, and is bordered by Lake Michigan. It is an international cultural, financial, and transportation hub.



Chicago

From Wikipedia, the free encyclopedia

This article is about the city in Illinois. For other uses, see Chicago (disambiguation).

Chicago (/ʃɪˈkɑːɡoʊ/ (listen), locally also /ʃɪˈkɔːɡoʊ/), officially the **City of Chicago**, is the most populous city in the U.S. state of Illinois and the third most populous city in the United States. With an estimated population of 2,705,994 (2018), it is also the most populous city in the Midwestern United States. Chicago is the county seat of Cook County, the second most populous county in the US, with a small portion of the northwest side of the city extending into DuPage County near O'Hare Airport. Chicago is the principal city of the Chicago metropolitan area, often referred to as Chicagoland. At nearly 10 million people, the metropolitan area is the third most populous in the nation.

Located on the shores of freshwater Lake Michigan, Chicago was incorporated as a city in 1837 near a portage between the Great Lakes and the Mississippi River watershed and grew rapidly in the mid-19th century.[7] After the Great Chicago Fire of 1871, which destroyed several square miles and left more than 100,000 homeless, the city made a concerted effort to rebuild.[8] The construction boom accelerated population growth throughout the following decades, and by 1900, less than 30 years after the great fire, Chicago was the fifth-largest city in the world.[9] Chicago made noted contributions to urban planning and zoning standards, including new construction styles (including the Chicago School of architecture), the development of the City Beautiful Movement, and the steel-framed skyscraper.[10][11]

Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. It is the site of the creation of the first standardized futures contracts, issued by the Chicago Board of Trade, which today is the largest and most diverse derivatives market in the world, generating 20% of all volume in commodities and financial futures alone.[12] Depending on the particular year, the city's O'Hare International Airport is routinely ranked as the world's fifth or sixth busiest airport according to tracked data by the Airports Council International.[13] The region also has the largest number of federal highways and is the nation's railroad hub.[14] Chicago was listed as an alpha global city by the Globalization and World Cities Research Network,[15] and it ranked seventh in the entire world in the 2017 Global Cities Index.[16] The Chicago area has one of the highest gross domestic products (GDP) in the world, generating $680 billion in 2017.[17] In addition, the city has one of the world's most diversified and balanced economies, with no single industry employing more than 14% of the workforce.[18] Chicago is home to several Fortune 500 companies, including Allstate, Boeing, Exelon, Kraft Heinz, McDonald's, Mondelez International, Sears, United Airlines Holdings, and Walgreens.

Chicago's 58 million domestic and international visitors in 2018 made it the second most visited city in the nation, as compared with New York City's 65 million visitors in 2018.[19][20] The city was ranked first in the 2018 Time Out City Life Index, a global quality of life survey of 15,000 people in 32 cities.[21][22][23][24][25] Landmarks in the city include Millennium Park, Navy Pier, the Magnificent Mile, the Art Institute of Chicago, Museum Campus, the Willis (Sears) Tower, Grant Park, the Museum of Science and Industry, and Lincoln Park Zoo. Chicago's culture includes the visual arts, literature, film, theatre, comedy (especially improvisational comedy), food, and music, particularly jazz, blues, soul, hip-hop, gospel,[26] and electronic dance music including house music. Of the area's many colleges and universities, the University of Chicago, Northwestern University, and the University of Illinois at Chicago are classified as "highest research" doctoral universities. Chicago has professional sports teams in each of the major professional leagues, including two Major League Baseball teams.

Clockwise from top: Downtown, the Chicago Theatre, the 'L', Navy Pier, the Pritzker Pavilion, the Field Museum, and Willis Tower

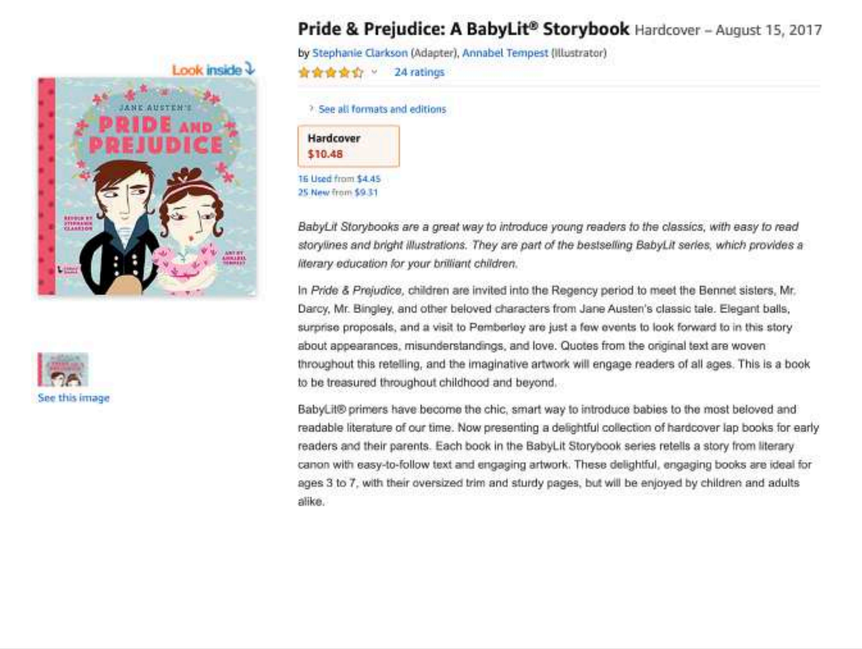
Etymology: Miami-Illinois: *shikaakwa* ("wild onion" or "wild garlic")
Potawatomi: *Gaa-zhigaagwanzhikaag*
Nicknames: Windy City, Chi-Town, City of Big Shoulders,[1] Second City, My Kind of Town
(for more, see full list)
Motto(s): Latin: *Urbs in Horto* (*City in a Garden*); I Will

Summarization

- Summaries are shorter than the full documents returned using information retrieval algorithms
- Summaries are longer than the short answer phrases returned by question answering systems

Summaries in the Real World

- Document outlines
- Abstracts for academic articles
- News article headlines
- Website snippets on search results pages
- Meeting minutes
- "Child-friendly" versions of text



Pride & Prejudice: A BabyLit® Storybook Hardcover – August 15, 2017

by Stephanie Clarkson (Adapter), Annabel Tempest (Illustrator)

24 ratings

See all formats and editions

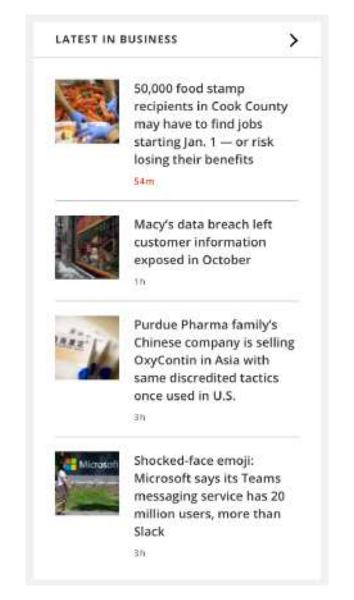
Hardcover
$10.48

16 Used from $4.45
25 New from $9.31

BabyLit Storybooks are a great way to introduce young readers to the classics, with easy to read storylines and bright illustrations. They are part of the bestselling BabyLit series, which provides a literary education for your brilliant children.

In *Pride & Prejudice*, children are invited into the Regency period to meet the Bennet sisters, Mr. Darcy, Mr. Bingley, and other beloved characters from Jane Austen's classic tale. Elegant balls, surprise proposals, and a visit to Pemberley are just a few events to look forward to in this story about appearances, misunderstandings, and love. Quotes from the original text are woven throughout this retelling, and the imaginative artwork will engage readers of all ages. This is a book to be treasured throughout childhood and beyond.

BabyLit® primers have become the chic, smart way to introduce babies to the most beloved and readable literature of our time. Now presenting a delightful collection of hardcover lap books for early readers and their parents. Each book in the BabyLit Storybook series retells a story from literary canon with easy-to-follow text and engaging artwork. These delightful, engaging books are ideal for ages 3 to 7, with their oversized trim and sturdy pages, but will be enjoyed by children and adults alike.



LATEST IN BUSINESS

50,000 food stamp recipients in Cook County may have to find jobs starting Jan. 1 — or risk losing their benefits
54m

Macy's data breach left customer information exposed in October
1h

Purdue Pharma family's Chinese company is selling OxyContin in Asia with same discredited tactics once used in U.S.
3h

Shocked-face emoji: Microsoft says its Teams messaging service has 20 million users, more than Slack
3h

University of Illinois at Chicago - Wikipedia
https://en.wikipedia.org › wiki › University_of_Illinois_at_Chicago ▾
The **University of Illinois at Chicago** (UIC) is a public research **university** in **Chicago, Illinois**. Its campus is in the Near West Side community area, adjacent to the **Chicago** Loop. ... UIC competes in NCAA Division I Horizon League as the UIC Flames in sports.

Campus: Urban, 244 acres (98.7 ha) **Mascot**: Sparky D. Dragon
Students: 33,390 **Colors**: Indigo blue and Flame red;

History · Academics · Campus · Student life

University of Illinois - University of Illinois at Chicago
https://www.uillinois.edu › about › Chicago ▾
The **University of Illinois at Chicago** is an acclaimed research center and a vital partner in the educational, technological, and cultural fabric of one of the nation's ...

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

1 Introduction

Types of Summarization

Number of documents summarized

- Single-document summarization
- Multiple-document summarization

Nature of the summary

- Generic summarization
- Query-focused summarization

Single-Document Summarization

- Given a single document, produce a summary
- Best for situations where the end goal is to characterize the content of a single document
- Example use cases:
    - Generating a headline for a news article
    - Producing an outline for a document

Multiple-Document Summarization

- Given a group of documents, produce a summary
- Best for situations when content from multiple sources needs to be synthesized
- Example use cases:
  - Summarizing a series of news stories covering the same event
  - Reviewing a cluster of similar prior work in a research area

Generic vs. Query-focused Summarization

Generic Summaries

- Provide the important information in a document
- Do not consider a specific user or a specific information need

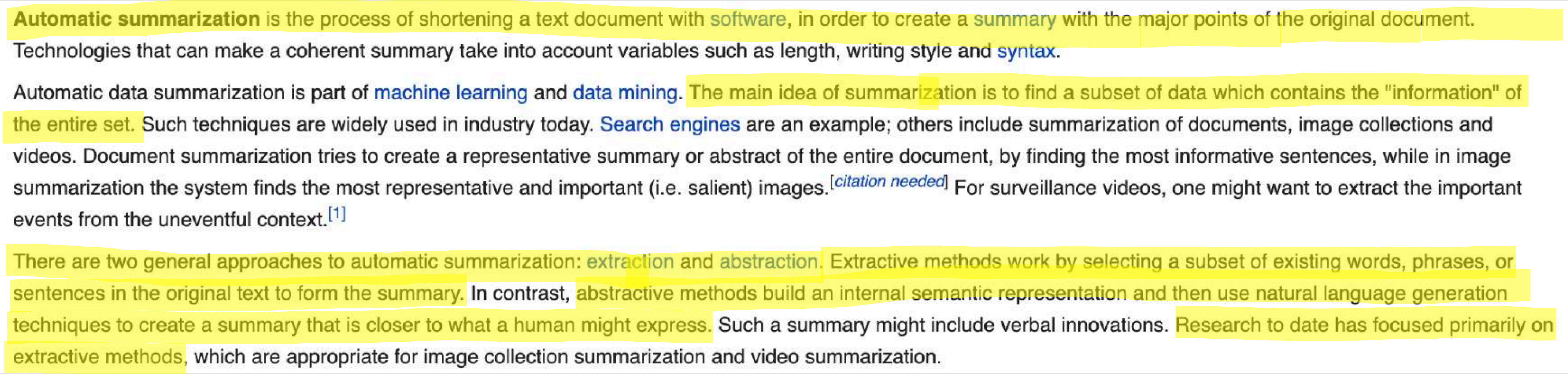
Query-focused Summaries

- Provide a specific set of information in response to a user's query
- Can be viewed as a longer, non-factoid answer to a question

Text Summarization Paradigms

Extractive Summarization

Abstractive Summarization



Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax.

Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a subset of data which contains the "information" of the entire set. Such techniques are widely used in industry today. Search engines are an example; others include summarization of documents, image collections and videos. Document summarization tries to create a representative summary or abstract of the entire document, by finding the most informative sentences, while in image summarization the system finds the most representative and important (i.e. salient) images.[citation needed] For surveillance videos, one might want to extract the important events from the uneventful context.[1]

There are two general approaches to automatic summarization: extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express. Such a summary might include verbal innovations. Research to date has focused primarily on extractive methods, which are appropriate for image collection summarization and video summarization.

Extractive Summarization

- Simplest form of text summarization
- Extract phrases or sentences from the source document(s) and combine them

Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax.

Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a subset of data which contains the "information" of the entire set. Such techniques are widely used in industry today. Search engines are an example; others include summarization of documents, image collections and videos. Document summarization tries to create a representative summary or abstract of the entire document, by finding the most informative sentences, while in image summarization the system finds the most representative and important (i.e. salient) images.[*citation needed*] For surveillance videos, one might want to extract the important events from the uneventful context.[1]

There are two general approaches to automatic summarization: extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express. Such a summary might include verbal innovations. Research to date has focused primarily on extractive methods, which are appropriate for image collection summarization and video summarization.

Automatic summarization is the process of transforming a full text document into a concise summary containing the same key information. The two general approaches to automatic summarization are extraction and abstraction. Extractive methods select subsets of text from the original document to form the summary, whereas abstractive methods generate new text that conveys the same core content. Most research to date has focused on extractive summarization.

Abstractive Summarization

- Much more complex
- Summarizes the underlying content in the text using different words
- Key goal in recent research is to move toward better abstractive summarization techniques

In general, summarization approaches need to focus on three main problems.

- **Content Selection**
  - What information should be selected from the document(s) being summarized?
- **Information Ordering**
  - How should the extracted information be ordered?
- **Sentence Realization**
  - What changes need to be made to the resulting summary to ensure that it is grammatically correct and natural-sounding?

Single-Document Summarization

Key focus:

- Content selection
- Sentence realization

Information ordering is often unnecessary!

- Original order from the source document can be used

How is content selected?

- Classification task
  - Predict whether each sentence in a document is **important** or **unimportant**
- This can be done using either **supervised** or **unsupervised** methods

Unsupervised Content Selection

- Often determine whether sentences are informative based on different **characteristics of their individual words**
- Sometimes detect **representative sentences** by computing each sentence's similarity with all other sentences in the document
- Sometimes rely on **rhetorical parsing**
  - **Rhetorical Parsing:** Identifying a hierarchical **discourse structure** for a passage of text

Rhetorical Structure Theory

- Text organization model
- 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


Natalie must be here.

Her coat is on the swivel chair.

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

Common RST Relations

Elaboration

- Satellite gives further information about the content of the nucleus

Attribution

- Satellite gives the s
  of reported speech

Natalie told the class not to come on November 28th, reminding them that it would be Thanksgiving.

Contrast

- Two or more nuclei contrast along some important dimension

List

- A series of nuclei is given, without contrast or explicit comparison

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 nuc
  dimension

Natalie pointed out that her students preferred to work the day before the deadline.

List

- A series of nuclei is given, without contrast or explicit comparison

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 nuc
  comparison

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 mor
dimension

In the fall, Natalie taught CS 421; **in the spring, Natalie taught CS 521**.

List

- A series of nuclei is given, without contrast or explicit comparison

RST relations can be hierarchically organized into discourse trees.

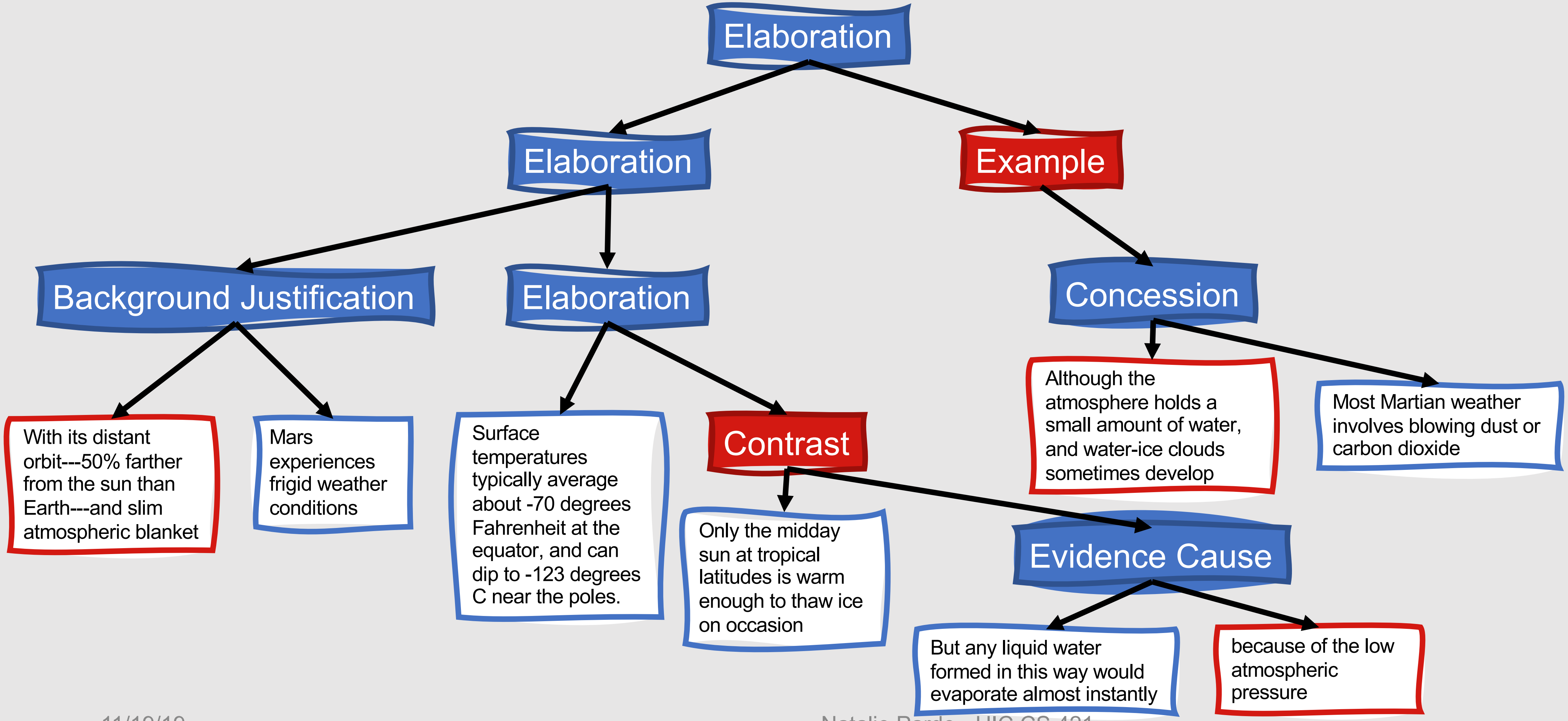
- This structure can in turn be used to determine which information to extract for a summary
- Simple strategy:
    - Keep nuclei
    - Discard satellites

Summarization based on Rhetorical Parsing

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.

Summarization based on Rhetorical Parsing



Elaboration

Elaboration

Example

Background Justification

Elaboration

Concession

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.

Contrast

Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop

Most Martian weather involves blowing dust or carbon dioxide

Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion

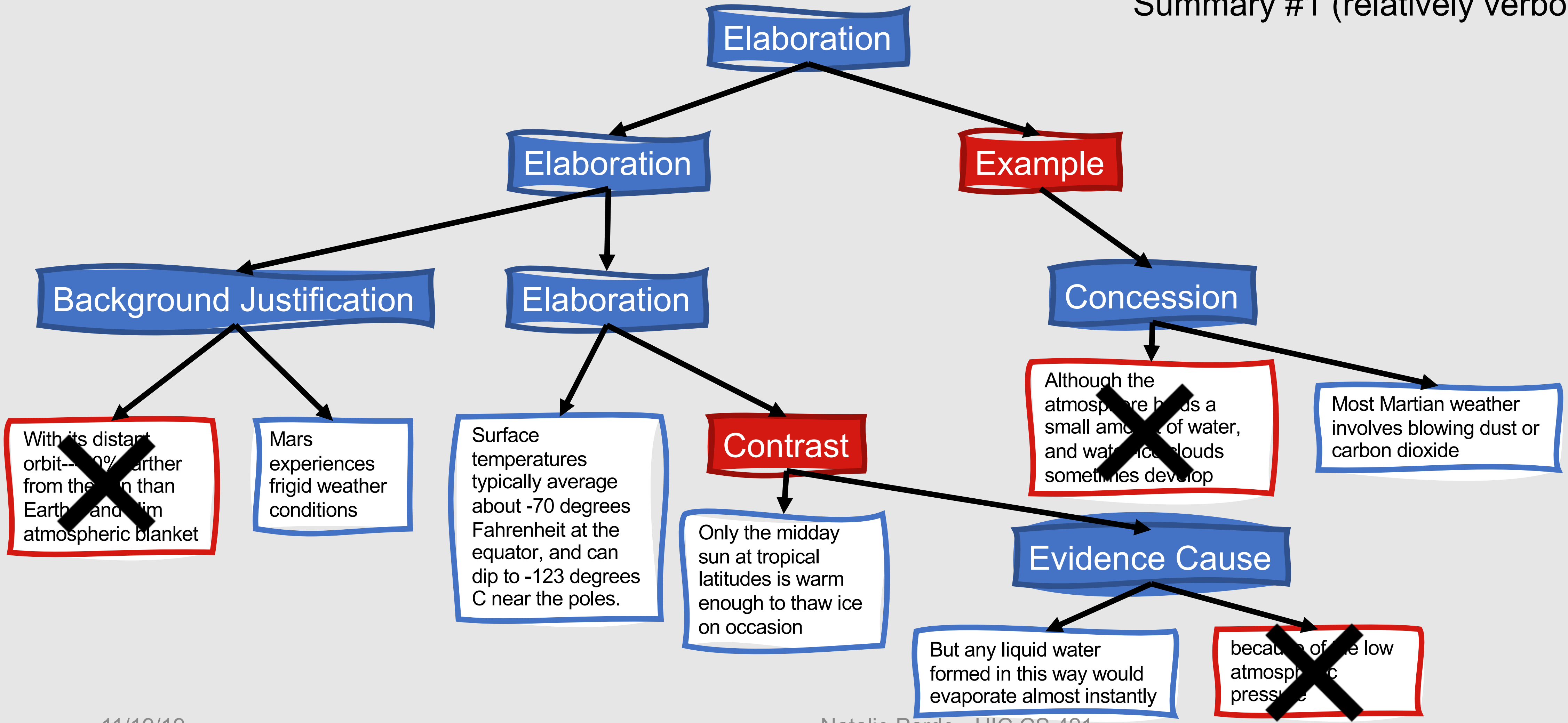
Evidence Cause

But any liquid water formed in this way would evaporate almost instantly

because of the low atmospheric pressure

Summarization based on Rhetorical Parsing

Summary #1 (relatively verbose)



Elaboration

Elaboration

Example

Background Justification

Elaboration

Concession

~~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.

Contrast

~~Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop~~

Most Martian weather involves blowing dust or carbon dioxide

Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion

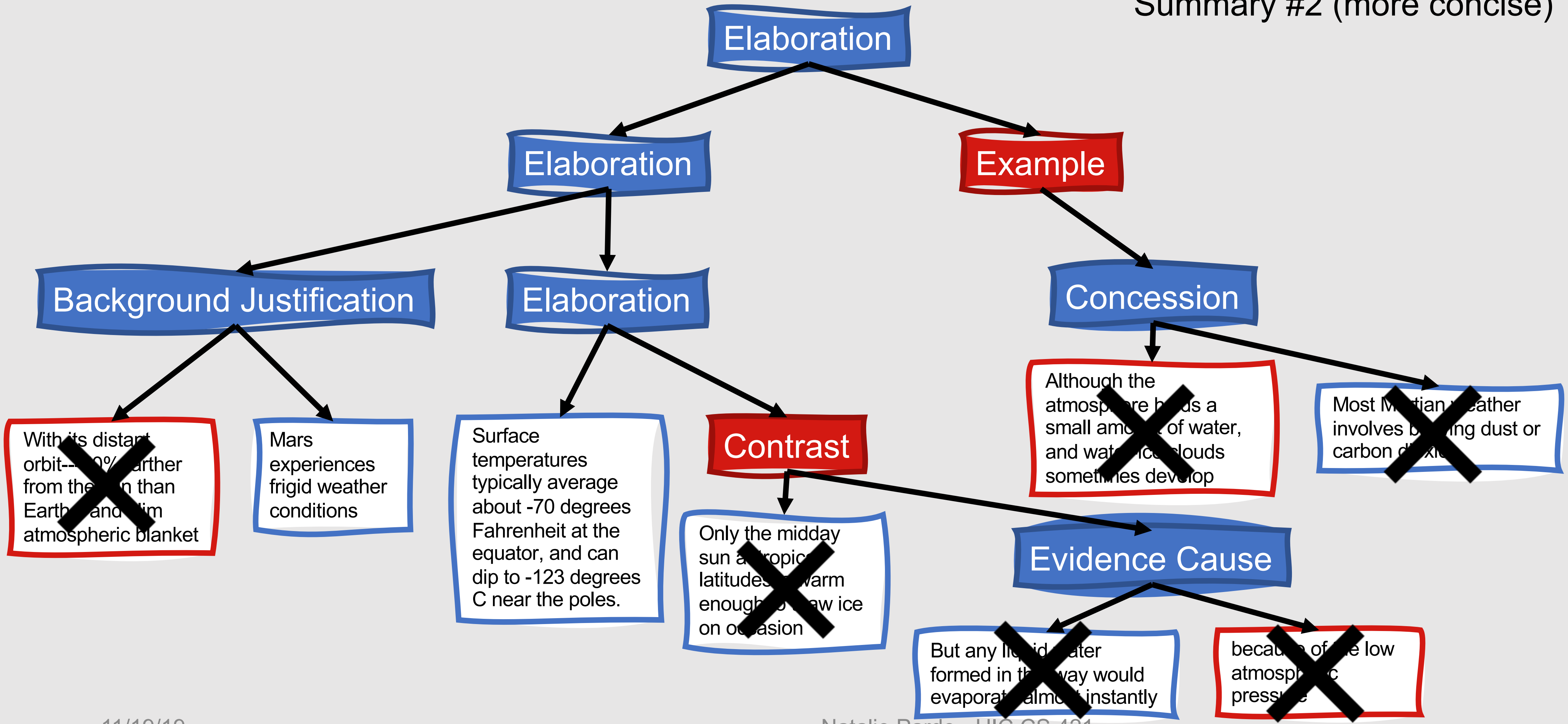
Evidence Cause

But any liquid water formed in this way would evaporate almost instantly

~~because of the low atmospheric pressure~~

Summarization based on Rhetorical Parsing

Summary #2 (more concise)



Elaboration

Elaboration

Example

Background Justification

Elaboration

Concession

~~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.

Contrast

~~Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion~~

Evidence Cause

~~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~~

Supervised Content Selection

- **Supervised machine learning**
    - Train a model based on various characteristics of the data to predict whether individual sentences should be included in a summary
- Requires that an **alignment** is found between source and summary content during the training phase
- Common training corpora:
    - Academic articles and their abstracts
    - Wikipedia and Simple Wikipedia (https://simple.wikipedia.org/wiki/Main_Page) articles

Sentence Simplification

Simplest approaches use rules to determine which parts of a sentence should be retained or discarded

Common rules:

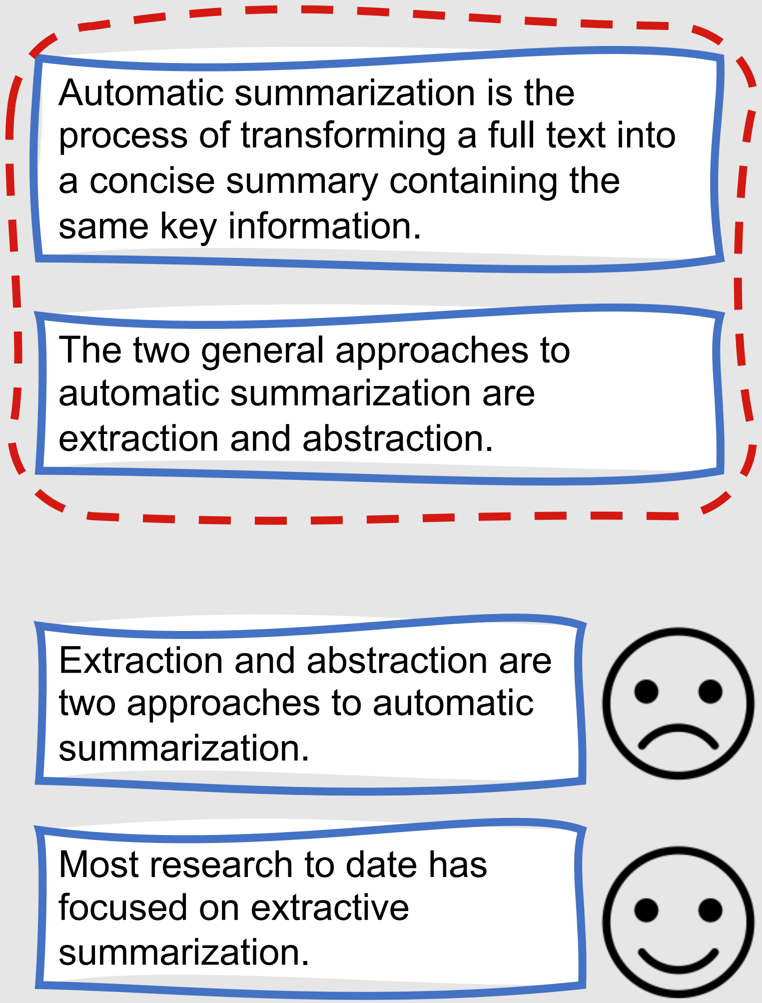
- Remove appositives
- Remove attribution clauses
- Remove prepositional phrases without named entities
- Remove initial adverbials
  - For example
  - As a matter of fact
  - On the other hand

Multiple-Document Summarization

- Requires **content selection** and **sentence realization** techniques, just like with single-document summarization
- Additionally, **information ordering** is important!

How is content selected in multi-document summarization tasks?

- Main difference: Greater risk of selecting **redundant information**
- The most important sentences in individual documents may overlap substantially with one another
  - We don't want a summary to consist of sets of identical sentences!
- How to address this?
  - Penalize sentences that are similar to those that have already been extracted into a summary



Automatic summarization is the process of transforming a full text into a concise summary containing the same key information.

The two general approaches to automatic summarization are extraction and abstraction.

Extraction and abstraction are two approaches to automatic summarization. ☹

Most research to date has focused on extractive summarization. ☺

Information Ordering

- One option: **Chronological order**
  - Can only be used if each sentence can be mapped to some location on a timeline
- However, placing sentences from multiple documents in chronological order can result in summaries with **low cohesion**
  - Summaries can seem like a collection of jumbled sentences rather than a unified block of text

Information Ordering

- Most important factor: **Coherence**
  - Is the information presented in a logical, consistent order?
- Simple way to maximize coherence:
  - Check the cosine similarity between each pair of sentences
  - Order the sentences in a way that maximizes the average cosine similarity between neighboring sentences
- Although good approximation approaches exist, finding an optimal order of sentences is challenging
  - Technically an **NP-complete** problem (Cohen et al., 1999)

Sentence Realization

- In multi-document summarization, entity names may need to be normalized
- Can be addressed by:
  - Applying **coreference resolution** to the summary
  - Extracting all possible names for each entity
  - Selecting one for the first mention, and a shorter one for all subsequent mentions

Natalie Parde is an assistant professor at the University of Illinois at Chicago. Dr. Natalie Parde joined University of Illinois (Chicago) in Fall 2018.…

Dr. Natalie Parde is an assistant professor at the University of Illinois at Chicago. Parde joined UIC in Fall 2018.…

Query-Focused Summarization

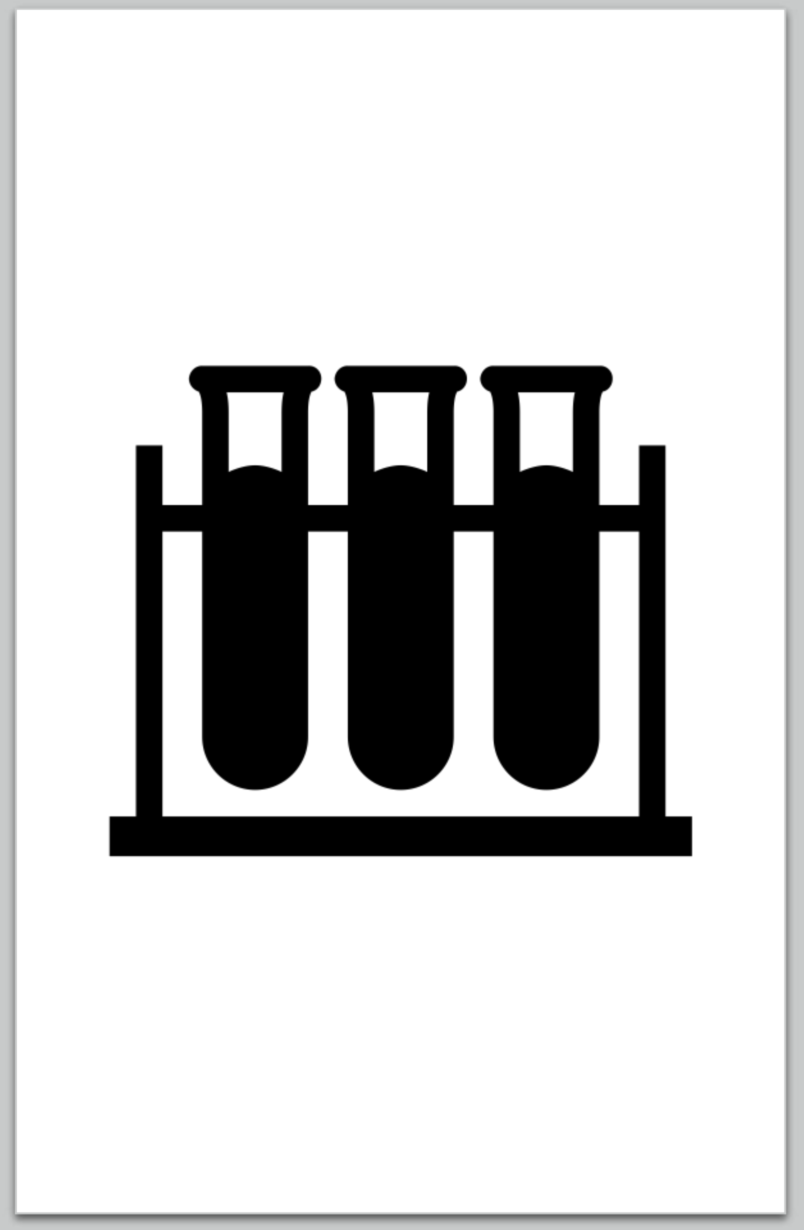
- Main difference from general summarization: Produced summary needs to **be relevant** to a user's question
- Thus, query-focused summarization may be viewed as a **long-form question answering** task

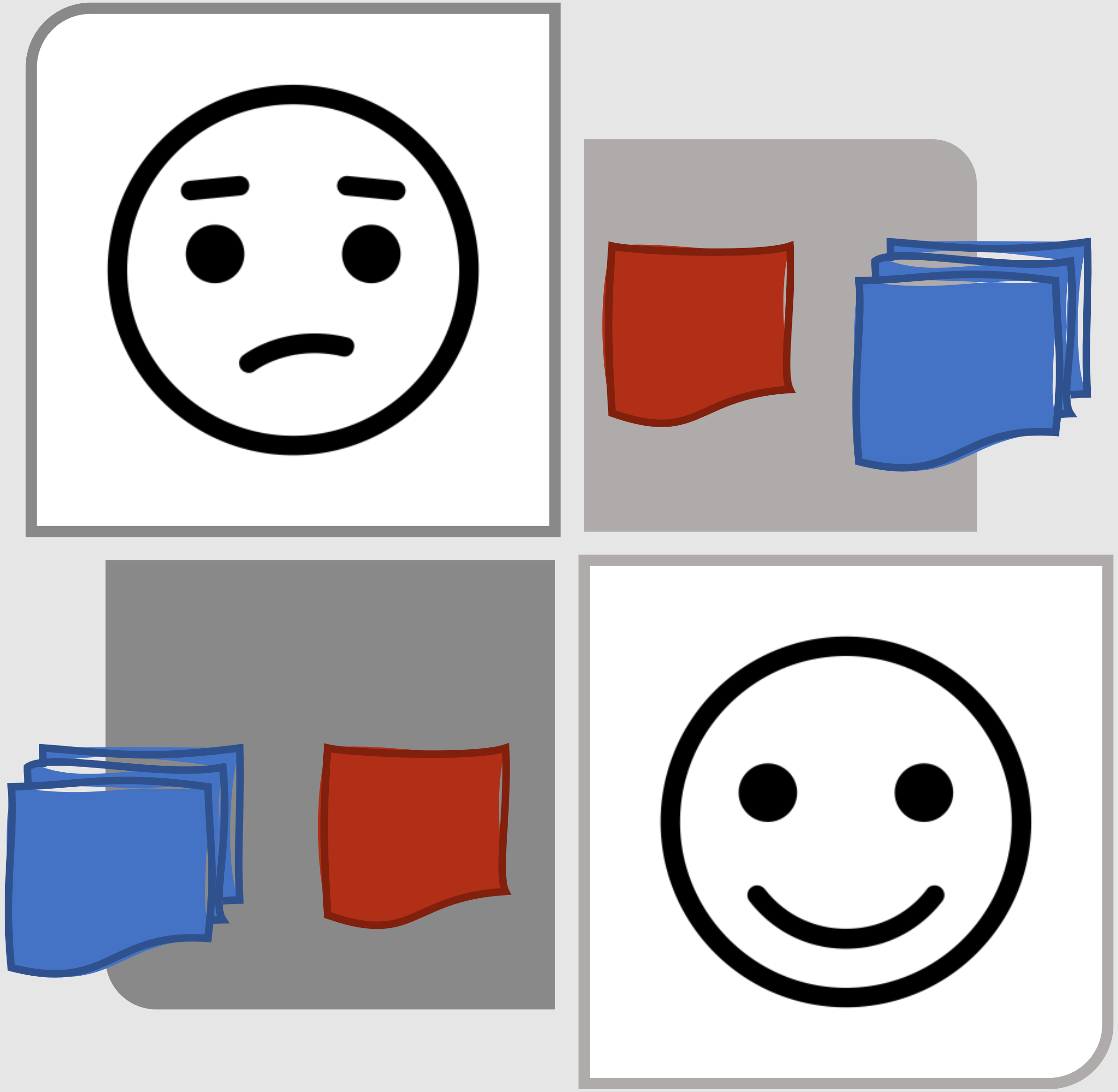
How can we modify general summarization methods for query-focused settings?

- When ranking sentences during content selection, **require a minimum amount of overlap with the query**
- **Add the cosine similarity with the query as a feature** for supervised content selection approaches
- Domain-specific approaches can **incorporate external knowledge about what factors are likely to interest people**
    - People asking biographical questions are likely to want to know about birth date, education, and nationality
    - People asking medical questions are likely to want to know symptoms, interventions, and outcomes

How do we determine the quality of our summarization approaches?

- Extrinsic methods
- Intrinsic methods





Extrinsic Evaluation

- Give automatically-generated summaries to humans to use while performing some task
- Evaluate their performance at the task relative to others using manually-generated summaries

Intrinsic Evaluation

- Recall-Oriented Understudy for Gisting Evaluation (**ROUGE**)
- Automatically scores a machine-generated candidate summary by measuring its **n-gram overlap with human-generated reference summaries**

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

- **Fixed n-gram length**
  - ROUGE-1 uses unigram overlap
  - ROUGE-2 uses bigram overlap
  - ROUGE-4 uses four-gram overlap
- Can be viewed as a form of **n-gram recall**

Computing ROUGE Scores

- Extract all n-grams from the candidate summary
- Extract all n-grams from the reference summary
- Find the intersection of the two lists
  - You can view these as true positives
- Divide the number of n-grams in the intersection (TP) by the total number of n-grams in the reference summary
- Formal equation:
  - $\text{ROUGE}-\text{N} = \frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_s\in S} Count_{match}(gram_s)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_s\in S} Count(gram_s)}$

Example: Computing ROUGE-2

Chicago is the third largest city in the country. ← Candidate Summary

Chicago is the third most populous city in the country. ← Reference Summary

Example: Computing ROUGE-2

Chicago is the third largest city in the country. ← Candidate Summary

Chicago is the third most populous city in the country. ← Reference Summary

- Chicago is
- is the
- the third
- third largest
- largest city
- city in
- in the
- the country.

- Chicago is
- is the
- the third
- third most
- most populous
- populous city
- city in
- in the
- the country.

Example: Computing ROUGE-2

Chicago is the third largest city in the country. ← Candidate Summary

Chicago is the third most populous city in the country. ← Reference Summary

- Chicago is
- is the
- the third
- third largest
- largest city
- city in
- in the
- the country.

∩

- Chicago is
- is the
- the third
- third most
- most populous
- populous city
- city in
- in the
- the country.

=

- Chicago is
- is the
- the third
- city in
- in the
- the country.

Example: Computing ROUGE-2

Chicago is the third largest city in the country. ← Candidate Summary

Chicago is the third most populous city in the country. ← Reference Summary

- Chicago is
- is the
- the third
- third largest
- largest city
- city in
- in the
- the country.

∩

- Chicago is
- is the
- the third
- third most
- most populous
- populous city
- city in
- in the
- the country.

=

- Chicago is
- is the
- the third
- city in
- in the
- the country.

ROUGE-2 = 6/9 = .67

Many variations of ROUGE exist!

ROUGE-L

- Longest common subsequence between the candidate and reference summaries

ROUGE-S

- Allows skip bigrams (any pair of words in their sentence order)

ROUGE-SU

- Uses both skip bigrams and unigrams

ROUGE isn't perfect….

- Measuring word overlap is only one (relatively poor) way to measure the similarity between a candidate and reference sentence
- Plus, human summarizers tend to disagree about which sentences to include in a summary, even with one another

Other Evaluation Metrics

Some metrics instead check the overlap between **summary content units (SCUs)** in candidate and reference sentences

- **Summary Content Unit:** Semantic units that roughly correspond to propositions or coherent pieces of propositions

However, identifying SCUs can be a very difficult task in itself

Baselines for Comparison

Random sentences

- Choose N random sentences from the full document to use as the summary

Leading sentences

- Choose the first N sentences from the full document to use as the summary

The leading sentences baseline is surprisingly strong!

- People tend to put the most important information early in a document

Summary: Question Answering and Summarization

- Question answering systems are often evaluated using **mean reciprocal rank**
  - Scores each question according to the reciprocal of the rank of the first correct answer
- Other common evaluation metrics are **exact match** and **F_1**
- **Text summarization** is the process of extracting the most important content from a text and presenting it in a concise, coherent manner
- Text summarization can be:
  - Performed on **one or more documents**
  - **Abstractive** or **extractive**
- Summaries can be **generic** or **query-focused**
- The three key processes involved in summarization are:
  - **Content selection**
  - **Information ordering**
  - **Sentence realization**
- Content selection is sometimes performed using **rhetorical parsing**
- Text summarization techniques are usually evaluated using **ROUGE**