# Question Answering

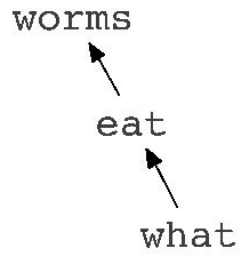
## What is Question Answering?

# Question Answering

One of the oldest NLP tasks (punched card systems in 1961)

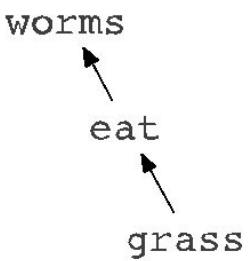
Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

Question: What do worms eat?

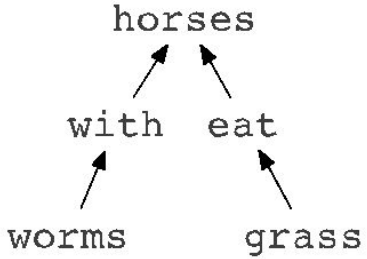



Potential Answers:

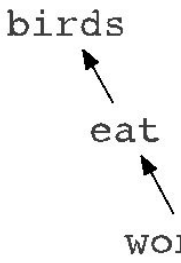
Worms eat grass



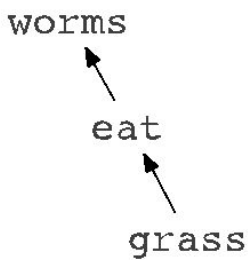

Horses with worms eat grass




Birds eat worms




Grass is eaten by worms



# Question Answering: IBM's Watson

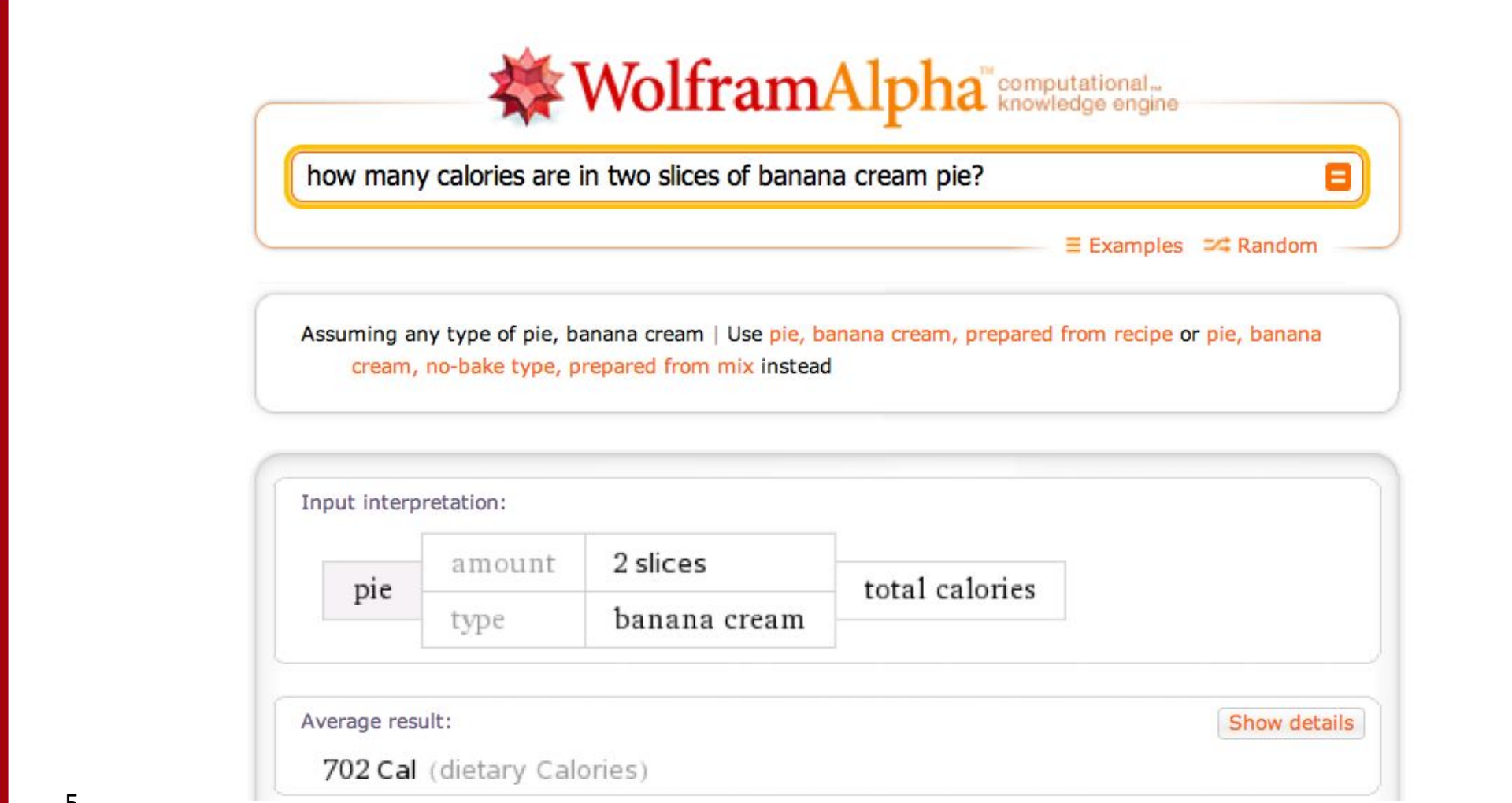
- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES
OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

# Apple’s Siri


AT&T
22:26
65 %
“ Do I need an umbrella tomorrow in San Francisco ”
Yes, San Francisco should get rain tomorrow:
Weekly Forecast
TUES WED THU FRI SAT SUN
14° 16° 17° 17° 18° 18°
9° 7° 6° 6° 7° 8°


WolframAlpha computational knowledge engine
how many calories are in two slices of banana cream pie?
Examples
Random
Assuming any type of pie, banana cream | Use pie, banana cream, prepared from recipe or pie, banana cream, no-bake type, prepared from mix instead
Input interpretation:
pie
amount
2 slices
type
banana cream
total calories
Average result:
Show details
702 Cal (dietary Calories)

# Types of Questions in Modern Systems

- Factoid questions
  - *Who wrote "The Universal Declaration of Human Rights"?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
  - *What do scholars think about Jefferson's position on dealing with pirates?*

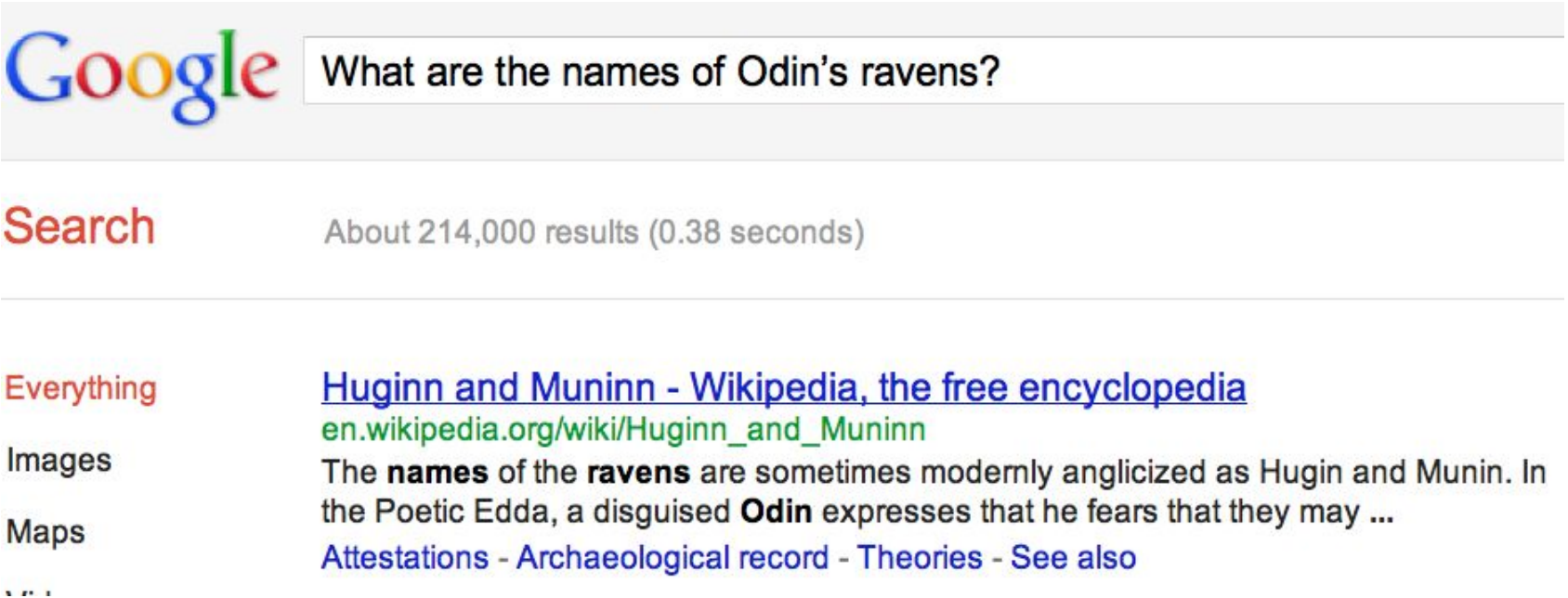
# Commercial systems: mainly factoid questions

| | |
|---|---|
| Where is the Louvre Museum located? | In Paris, France |
| What's the abbreviation for limited partnership? | L.P. |
| What are the names of Odin's ravens? | Huginn and Muninn |
| What currency is used in China? | The yuan |
| What kind of nuts are used in marzipan? | almonds |
| What instrument does Max Roach play? | drums |
| What is the telephone number for Stanford University? | 650-723-2300 |

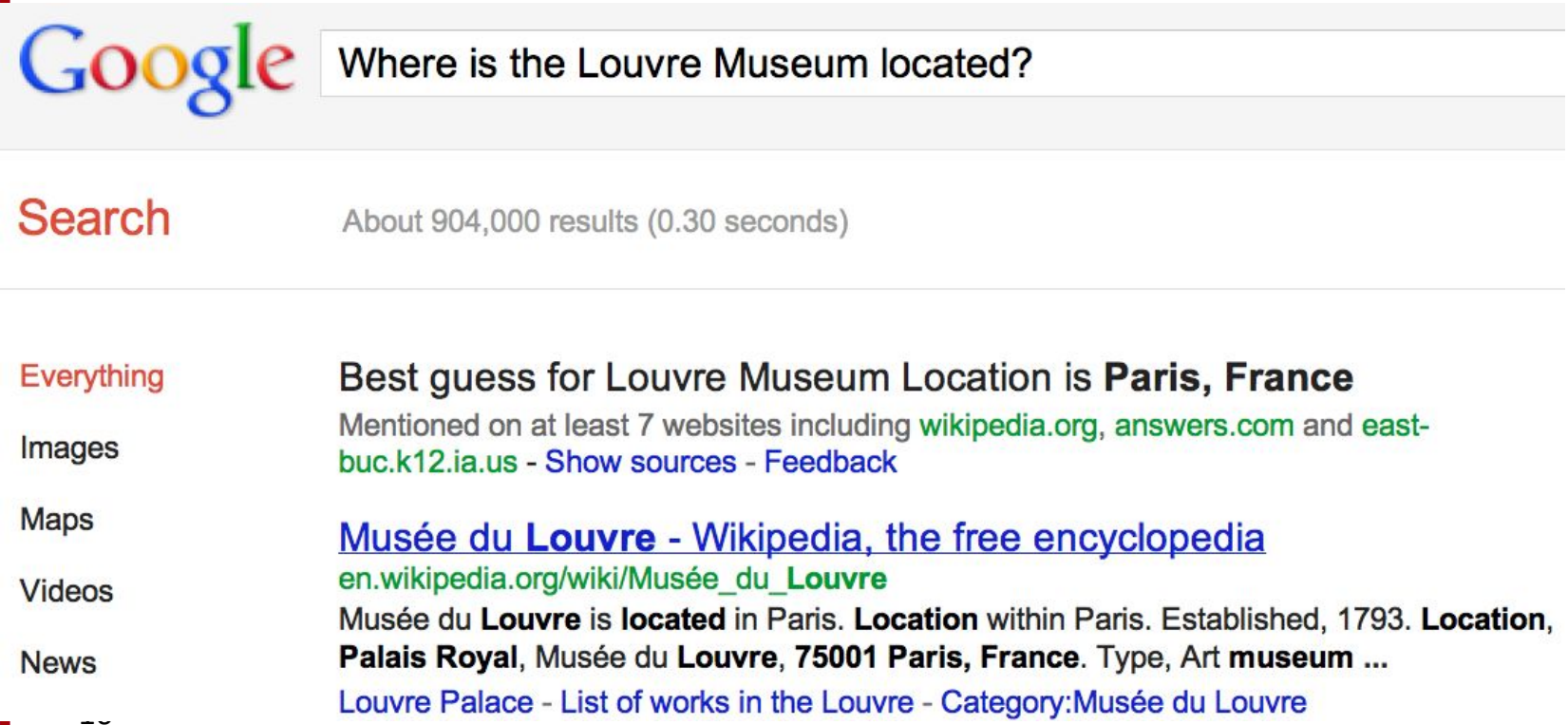
# Paradigms for QA

- IR-based approaches
  - TREC;  IBM Watson; Google
- Knowledge-based and Hybrid approaches
  - IBM Watson; Apple Siri; Wolfram Alpha; True Knowledge Evi

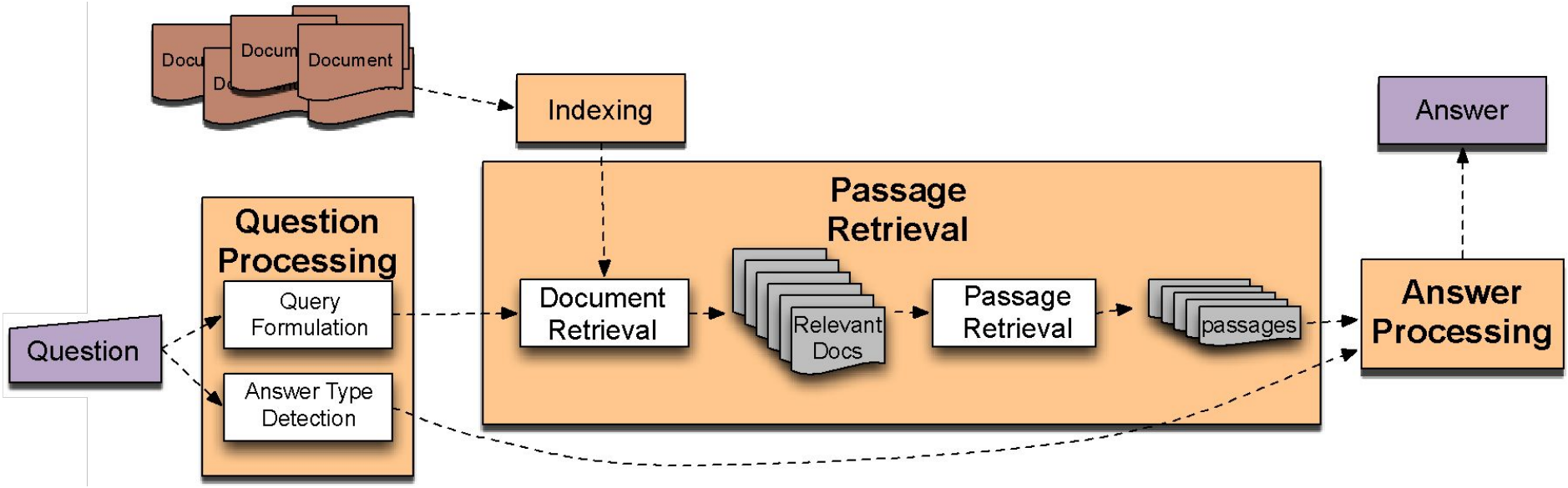
# Many questions can already be answered by web search



# IR-based Question Answering



# IR-based Factoid QA



# IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

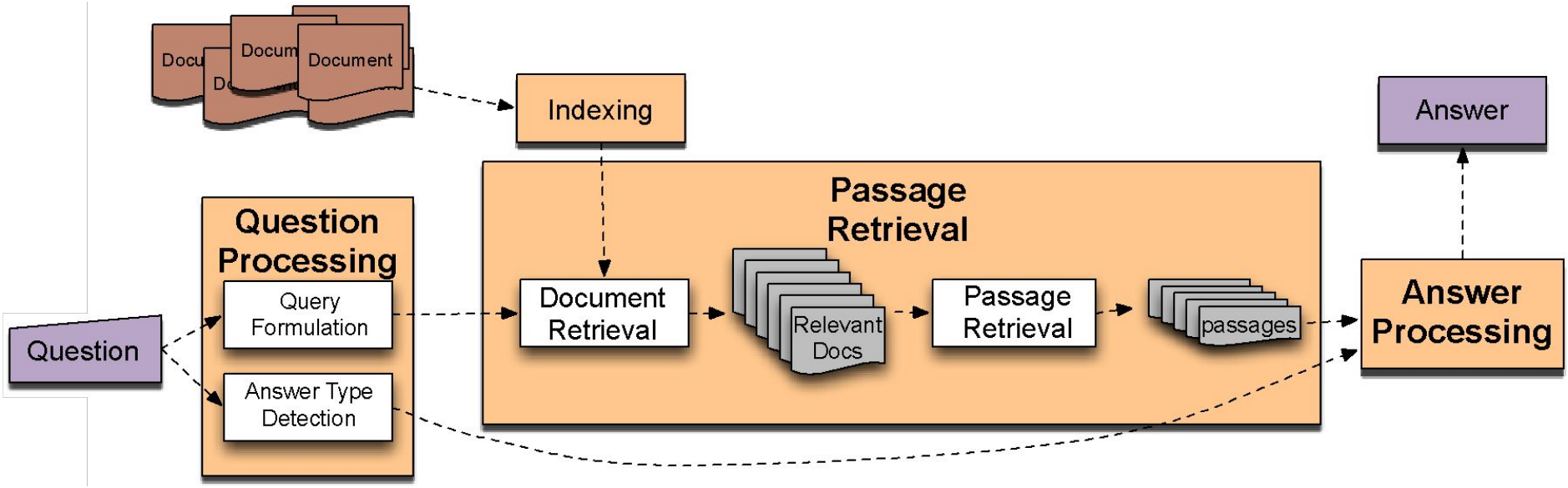
# Knowledge-based approaches (Siri)

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data  or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases

# Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

# Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Question Processing
Passage Retrieval
Question
Query Formulation
Answer Type Detection
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

# Question Processing
# Things to extract from the question

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

# Question Processing

They're the two states you could be reentering if you're crossing Florida's northern border

- Answer Type: US state
- Query: two states, border, Florida, north
- Focus: the two states
- Relations: borders(Florida, ?x, north)

# Answer Type Detection: Named Entities

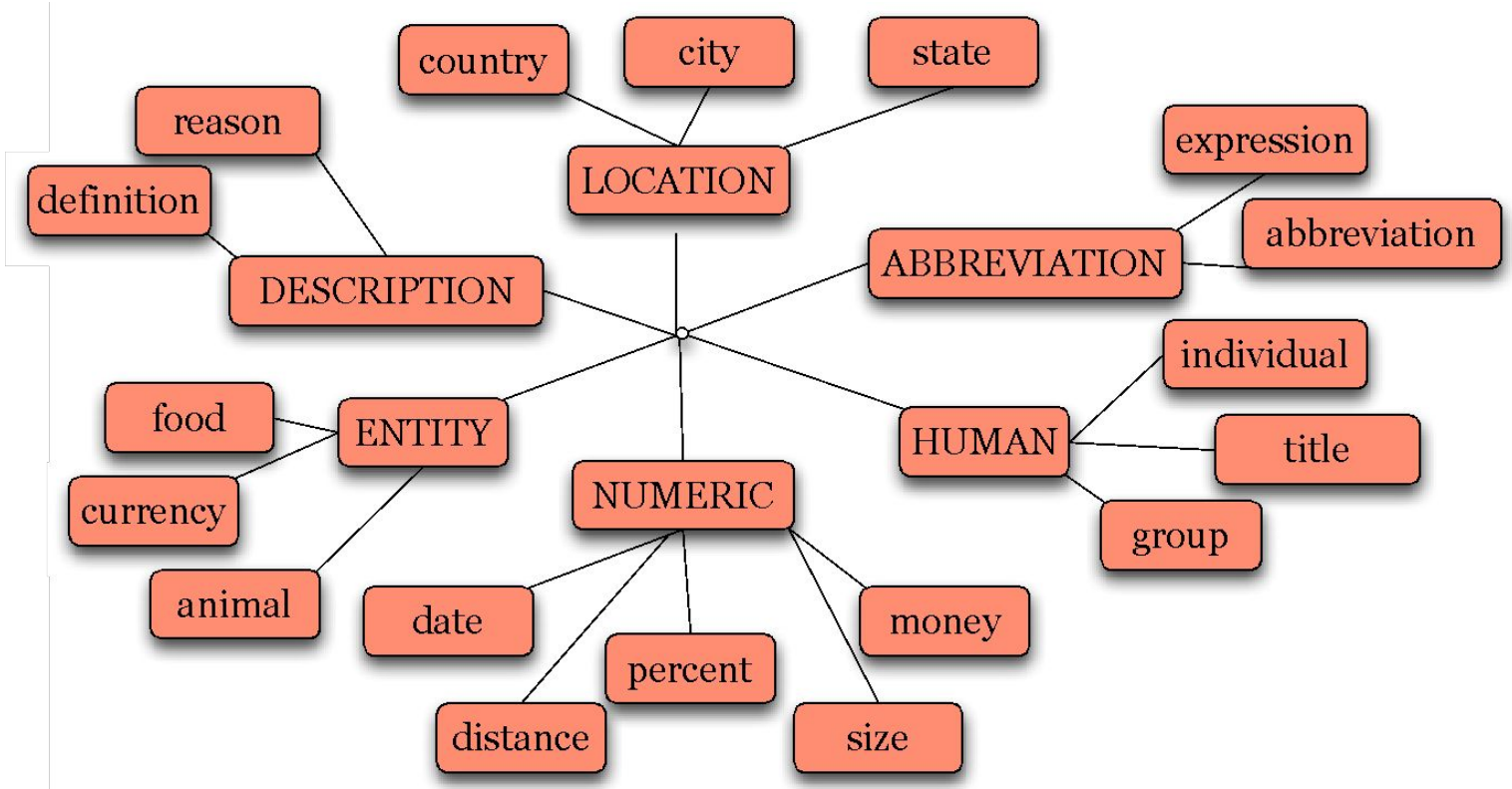
- *Who founded Virgin Airlines?*
  - PERSON
- *What Canadian city has the largest population?*
  - *CITY.*

# Answer Type Taxonomy

Xin Li, Dan Roth. 2002. Learning Question Classifiers. COLING'02

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency…

# Part of Li & Roth's Answer Type Taxonomy



# Answer Types

| 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? |

# More Answer Types

| | |
|---|---|
| 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? |
| speed | What is the speed of the Mississippi River? |
| temp | 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? |

# Answer types in Jeopardy

Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59-79.

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types

he, country, city, man, film, state, she, author, group, here, company, president, capital, star, novel, character, woman, river, island, king, song, part, series, sport, singer, actor, play, team, show, actress, animal, presidential, composer, musical, nation, book, title, leader, game

# Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

# Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who {is|was|are|were} PERSON
  - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:

  (the headword of the first noun phrase after the wh-word)

  - Which **city** in China has the largest number of foreign financial companies?
  - What is the state **flower** of California?

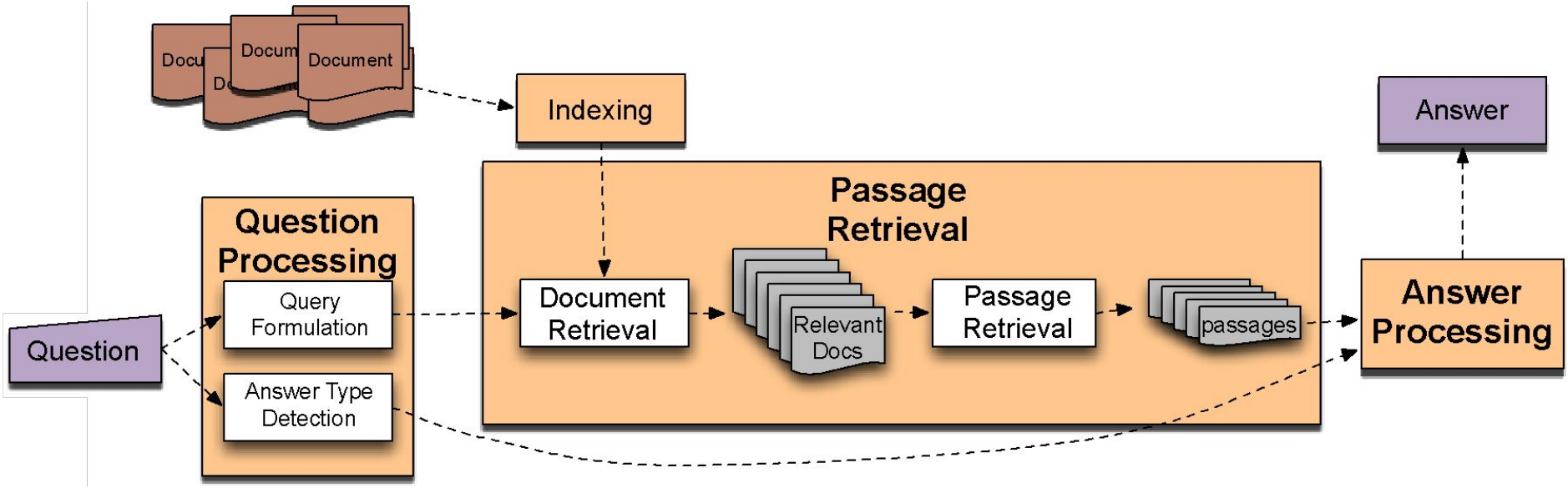
# Answer Type Detection

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - features include those hand-written rules!

# Features for Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

# Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Question Processing
Passage Retrieval
Question
Query Formulation
Answer Type Detection
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

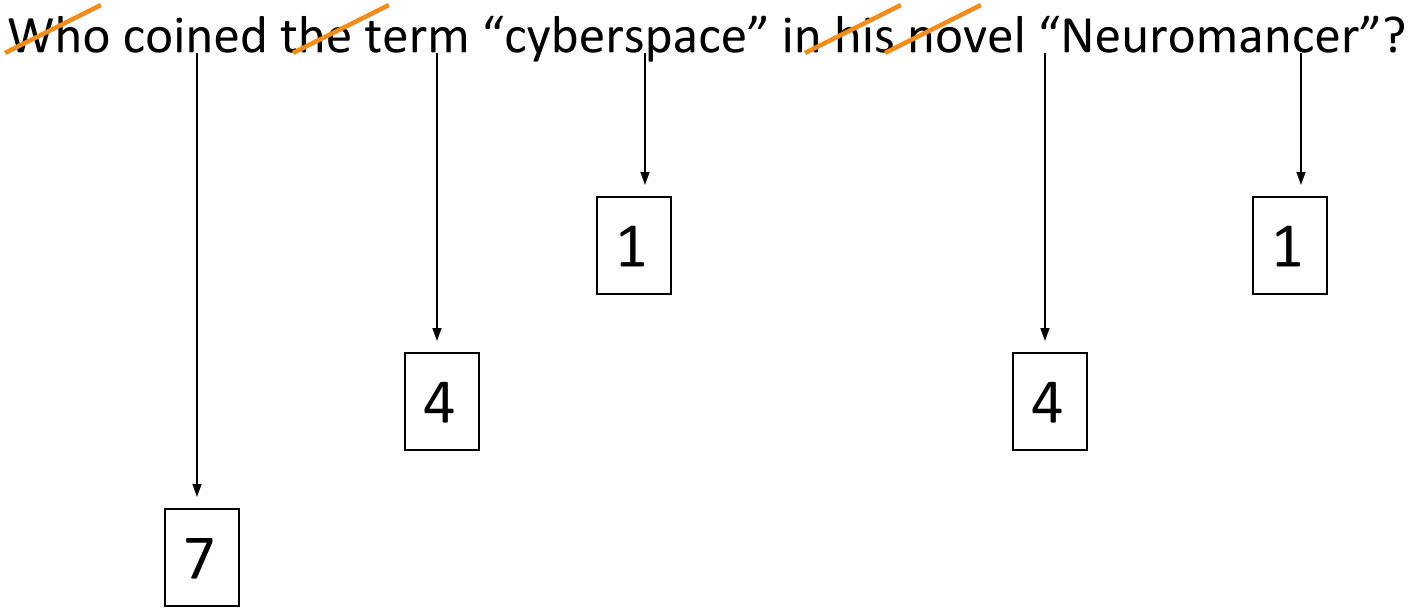
# Keyword Selection Algorithm

Dan Moldovan, Sanda Harabagiu, Marius Paca, Rada Mihalcea, Richard Goodrum, Roxana Girju and Vasile Rus. 1999. Proceedings of TREC-8.

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

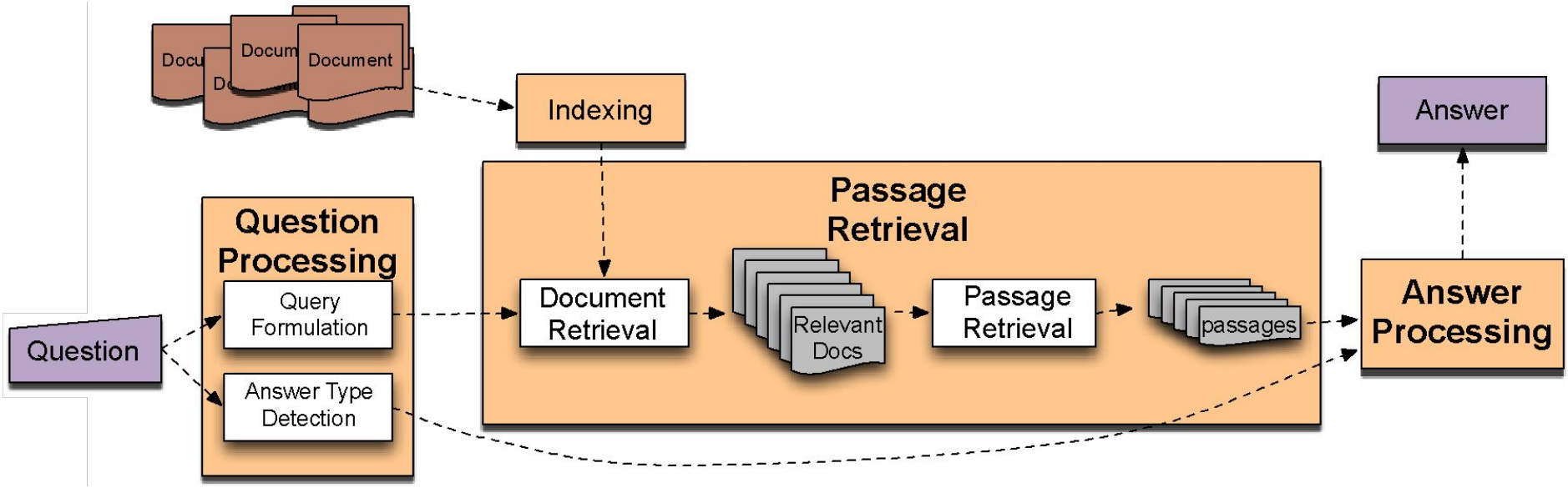
# Choosing keywords from the query

Slide from Mihai Surdeanu




cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

# Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Passage Retrieval
Question Processing
Query Formulation
Answer Type Detection
Question
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

# Passage Retrieval

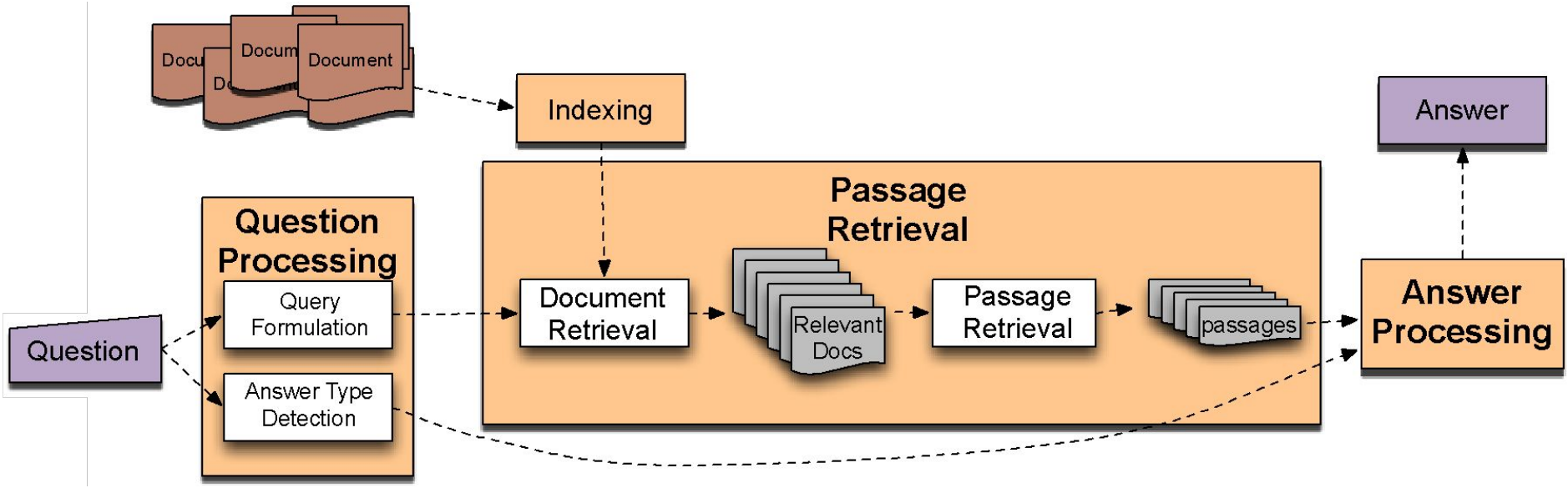
- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help rerank passages

# Features for Passage Ranking

Either in rule-based classifiers or with supervised machine learning

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

# Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Question Processing
Passage Retrieval
Question
Query Formulation
Answer Type Detection
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

# Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - Who is the prime minister of India (PERSON)

**Manmohan Singh**, Prime Minister of India, had told left leaders that the deal would not be renegotiated.

  - How tall is Mt. Everest? (LENGTH)

The official height of Mount Everest is **29035 feet**

# Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**

- Passage:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

# Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**

- Passage:

The Marie biscuit is named after **Marie Alexandrovna**, the daughter of **Czar Alexander II of Russia** and wife of **Alfred**, the second son of **Queen Victoria** and **Prince Albert**

# Use machine learning: Features for ranking candidate answers

**Answer type match:** Candidate contains a phrase with the correct answer type.

**Pattern match**: Regular expression pattern matches the candidate.

**Question keywords**: # of question keywords in the candidate.

**Keyword distance**: Distance in words between the candidate and query keywords

**Novelty factor**: A word in the candidate is not in the query.

**Apposition features**: The candidate is an appositive to question terms

**Punctuation location**: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

**Sequences of question terms**: The length of the longest sequence of question terms that occurs in the candidate answer.

# Candidate Answer scoring in IBM Watson

- Each candidate answer gets scores from >50 components
  - (from unstructured text, semi-structured text, triple stores)

  - logical form (parse) match between question and candidate
  - passage source reliability
  - geospatial location
    - California is ”southwest of Montana”
  - temporal relationships
  - taxonomic classification

# Common Evaluation Metrics

1. *Accuracy* (does answer match gold-labeled answer?)
2. *Mean Reciprocal Rank*
   - For each query return a ranked list of M candidate answers.
   - Query score is 1/Rank of the first correct answer
     - *If first answer is correct: 1*
     - *else if second answer is correct: ½*
     - *else if third answer is correct: ⅓, etc.*
     - *Score is 0 if none of the M answers are correct*
   - Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

# SQuAD

- MRC (Machine Reading Comprehension) tasks are challenging
  - Understanding
  - Knowledge

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

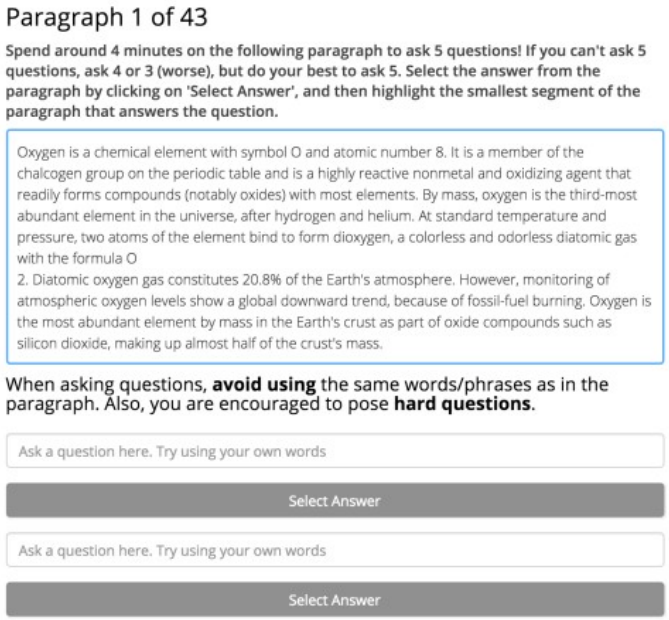
What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

# SQuAD

- Stanford Question Answering Dataset v1.0 (SQuAD)
  - 107,785 Question from 536 articles
  - Short-answer Question
- Collection
  - Crowdsourcing
- Answers
  - Segmentation of text, span from paragraph



# Survey

- Size
  - 2X on size of MCTest
- Quality
  - Better quality than cloze style (CNN, CBT)

| Dataset | Question source | Formulation | Size |
|---|---|---|---|
| **SQuAD** | **crowdsourced** | **RC, spans in passage** | **100K** |
| MCTest (Richardson et al., 2013) | crowdsourced | RC, multiple choice | 2640 |
| Algebra (Kushman et al., 2014) | standardized tests | computation | 514 |
| Science (Clark and Etzioni, 2016) | standardized tests | reasoning, multiple choice | 855 |
| WikiQA (Yang et al., 2015) | query logs | IR, sentence selection | 3047 |
| TREC-QA (Voorhees and Tice, 2000) | query logs + human editor | IR, free form | 1479 |
| CNN/Daily Mail (Hermann et al., 2015) | summary + cloze | RC, fill in single entity | 1.4M |
| CBT (Hill et al., 2015) | cloze | RC, fill in single word | 688K |

# Diversity

| Answer type | Percentage | Example |
|---|---|---|
| Date | 8.9% | 19 October 1512 |
| Other Numeric | 10.9% | 12 |
| Person | 12.9% | Thomas Coke |
| Location | 4.4% | Germany |
| Other Entity | 15.3% | ABC Sports |
| Common Noun Phrase | 31.8% | property damage |
| Adjective Phrase | 3.9% | second-largest |
| Verb Phrase | 5.5% | returned to Earth |
| Clause | 3.7% | to avoid trivialization |
| Other | 2.7% | quietly |

# Reasoning

| Reasoning | Description | Example | Percentage |
|---|---|---|---|
| Lexical variation (synonymy) | Major correspondences between the question and the answer sentence are synonyms. | Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle. | 33.3% |
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve. | Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process. | 9.1% |
| Syntactic variation | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington. | 64.1% |
| Multiple sentence reasoning | There is anaphora, or higher-level fusion of multiple sentences is required. | Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance. | 13.6% |
| Ambiguous | We don't agree with the crowdworkers' answer, or the question does not have a unique answer. | Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment. | 6.1% |

# Evaluation

| Feature Groups | Description | Examples |
|---|---|---|
| Matching Word Frequencies | Sum of the TF-IDF of the words that occur in both the question and the sentence containing the candidate answer. Separate features are used for the words to the left, to the right, inside the span, and in the whole sentence. | Span: $[0 \leq \text{sum} < 0.01]$<br>Left: $[7.9 \leq \text{sum} < 10.7]$ |
| Matching Bigram Frequencies | Same as above, but using bigrams. We use the generalization of the TF-IDF described in Shirakawa et al. (2015). | Span: $[0 \leq \text{sum} < 2.4]$<br>Left: $[0 \leq \text{sum} < 2.7]$ |
| Root Match | Whether the dependency parse tree roots of the question and sentence match, whether the sentence contains the root of the dependency parse tree of the question, and whether the question contains the root of the dependency parse tree of the sentence. | Root Match = False |
| Lengths | Number of words to the left, to the right, inside the span, and in the whole sentence. | Span: $[1 <= \text{num} < 2]$<br>Left: $[15 \leq \text{num} < 19]$ |
| Span Word Frequencies | Sum of the TF-IDF of the words in the span, regardless of whether they appear in the question. | Span: $[5.2 \leq \text{sum} < 6.9]$ |
| Constituent Label | Constituency parse tree label of the span, optionally combined with the wh-word in the question. | Span: NP<br>Span: NP, wh-word: "*what*" |
| Span POS Tags | Sequence of the part-of-speech tags in the span, optionally combined with the wh-word in the question. | Span: [NN]<br>Span: [NN], wh-word: "*what*" |
| Lexicalized | Lemmas of question words combined with the lemmas of words within distance 2 to the span in the sentence based on the dependency parse trees. Separately, question word lemmas combined with answer word lemmas. | Q: "*cause*", S: "*under*" $\xleftarrow{\text{case}}$<br>Q: "*fall*", A: "*gravity*" |
| Dependency Tree Paths | For each word that occurs in both the question and sentence, the path in the dependency parse tree from that word in the sentence to the span, optionally combined with the path from the wh-word to the word in the question. POS tags are included in the paths. | VBZ $\xrightarrow{\text{nmod}}$ NN<br>what $\xleftarrow{\text{nsubj}}$ VBZ $\xrightarrow{\text{advcl}}$<br>+ VBZ $\xrightarrow{\text{nmod}}$ NN |

# Evaluation

| | Exact Match | | F1 | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| Random Guess | 1.1% | 1.3% | 4.1% | 4.3% |
| Sliding Window | 13.2% | 12.5% | 20.2% | 19.7% |
| Sliding Win. + Dist. | 13.3% | 13.0% | 20.2% | 20.0% |
| Logistic Regression | 40.0% | 40.4% | 51.0% | 51.0% |
| Human | 80.3% | 77.0% | 90.5% | 86.8% |

# **Feature** Ablation

- Feature Ablation of Ours
  - To find important features

 Lexicalized, Dpendency Tree Path are most important

| | $F_1$ | |
|---|---|---|
| | Train | Dev |
| Logistic Regression | 91.7% | 51.0% |
| – Lex., – Dep. Paths | 33.9% | 35.8% |
| – Lexicalized | 53.5% | 45.4% |
| – Dep. Paths | 91.4% | 46.4% |
| – Match. Word Freq. | 91.7% | 48.1% |
| – Span POS Tags | 91.7% | 49.7% |
| – Match. Bigram Freq. | 91.7% | 50.3% |
| – Constituent Label | 91.7% | 50.4% |
| – Lengths | 91.8% | 50.5% |
| – Span Word Freq. | 91.7% | 50.5% |
| – Root Match | 91.7% | 50.6% |

**Table 6:** Performance with feature ablations. We find that lexicalized and dependency tree path features are most important.

# **Performance** Stratified by Answer Type

- Performance by answer type
  - Human is uniform

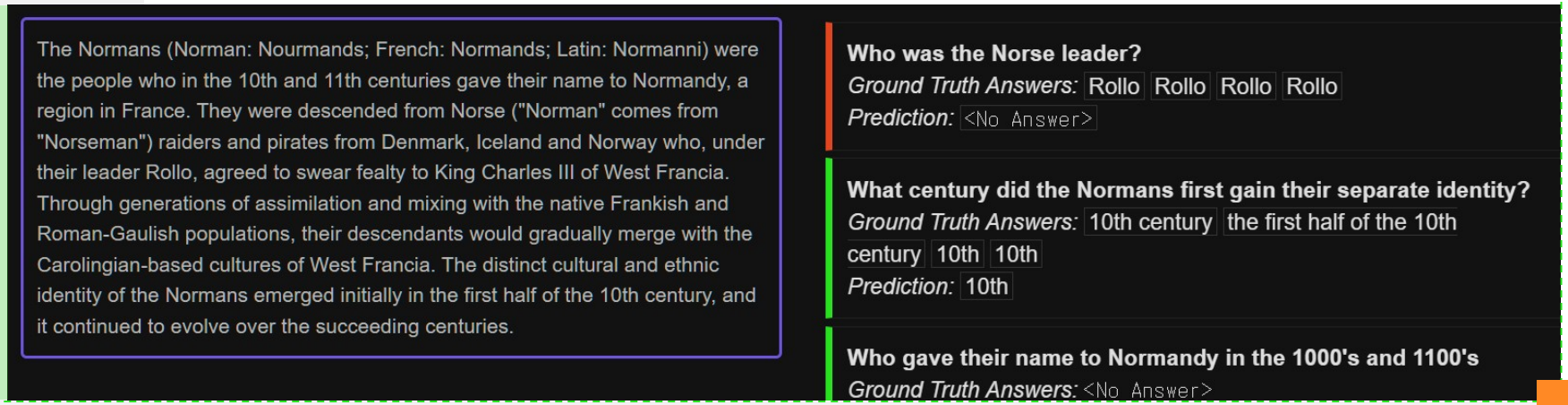
Ours are better at certain type

| | Logistic Regression Dev F1 | Human Dev F1 |
|---|---|---|
| Date | 72.1% | 93.9% |
| Other Numeric | 62.5% | 92.9% |
| Person | 56.2% | 95.4% |
| Location | 55.4% | 94.1% |
| Other Entity | 52.2% | 92.6% |
| Common Noun Phrase | 46.5% | 88.3% |
| Adjective Phrase | 37.9% | 86.8% |
| Verb Phrase | 31.2% | 82.4% |
| Clause | 34.3% | 84.5% |
| Other | 34.8% | 86.1% |

**Table 7:** Performance stratified by answer types. Logistic regression performs better on certain types of answers, namely numbers and entities. On the other hand, human performance is more uniform.

# Recent

- SQuAD 2.0
  - Adversarial attack ( unanswerable question)



# Relation Extraction

- Answers: Databases of Relations
  - born-in(“Emma Goldman”, “June 27 1869”)
  - author-of(“Cao Xue Qin”, “Dream of the Red Chamber”)
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

  Whose granddaughter starred in E.T.?

```
(acted-in ?x “E.T.”)
   (granddaughter-of ?x ?y)
```

# Temporal Reasoning

- Relation databases
  - (and obituaries, biographical dictionaries, etc.)
- IBM Watson

  ”In 1594 he took a job as a tax collector in Andalusia”

  Candidates:
    - Thoreau is a bad answer (born in 1817)
    - Cervantes is possible (was alive in 1594)

# Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for ”Asian city”
- California is ”southwest of Montana”
- geonames.org:

# Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities

  U: “Book a table at Il Fornaio at 7:00 with **my mom**”

  U: “Also send **her** an email reminder”

- Clarification questions:

  U: “Chicago pizza”

  S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”