COMP SCI 7327 Concepts in Artificial Intelligence & Machine Learning -Question Answering

By Dr Wei Zhang

Outline

- Question Answering
  - IR-based
  - Knowledge-based
  - Hybrid

Google

who is australia's prime minister

All News Images Videos Maps More Settings Tools

About 24,500,000 results (0.71 seconds)

Australia / Prime minister

Scott Morrison

Since 2018

pm.gov.au

Scott John Morrison is an Australian politician who has been serving as the 30th Prime Minister of Australia and Leader of Australian Liberal Party since 2018. Wikipedia

Born: 13 May 1968 (age 52 years), Sydney

Spouse: Jenny Morrison (m. 1990)

Party: Liberal Party of Australia

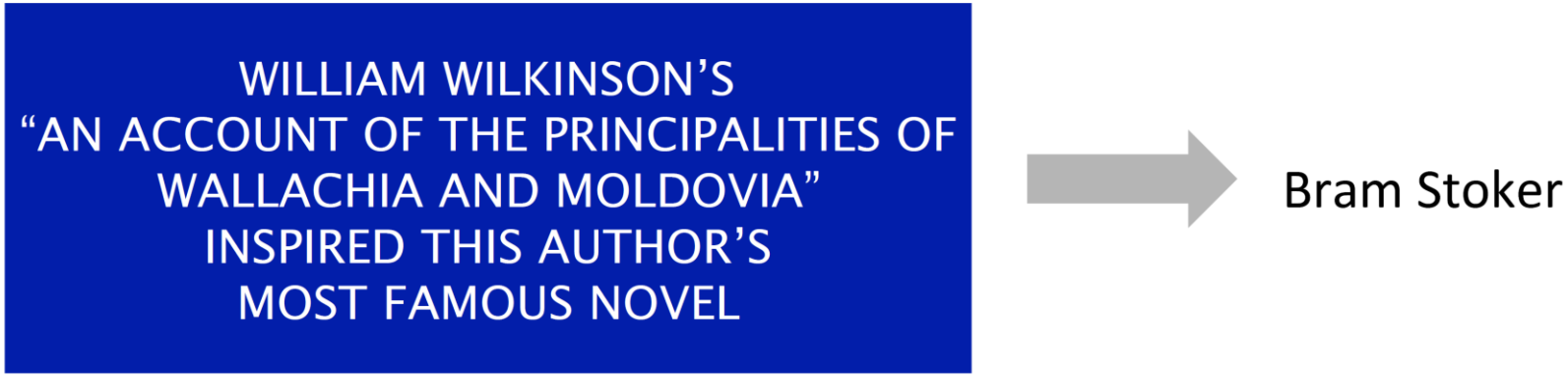
Children: Lily Morrison, Abbey Morrison

Education: UNSW Sydney, Sydney Boys High School

Previous offices: Acting Minister for Home Affairs of Australia (2018–2018), MORE

IBM's Watson

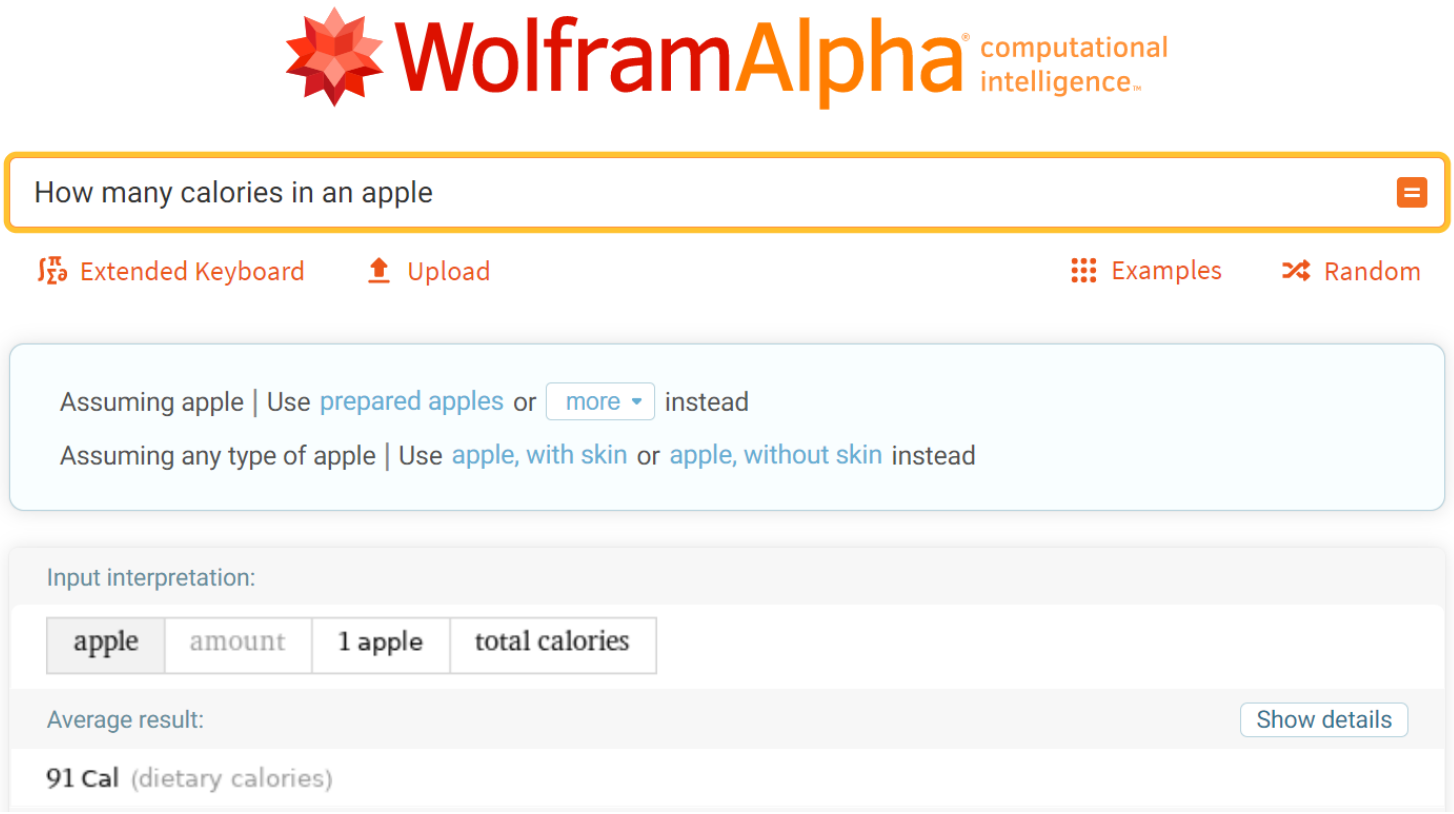
- Won Jeopardy on February 16, 2011



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°

Wolfram Alpha



Motivation

- With massive collections of full-text documents, i.e., the web, simply returning relevant documents is of limited use.
- Rather, we often want answers to our questions
  - Especially on mobile or using a digital assistant device, like Alexa, Google Assistant, …
- We can factor this into two parts:
  - Finding documents that (might) contain an answer, Which can be handled by traditional information retrieval/web search.
  - Finding an answer in a paragraph or a document. This problem is often termed Reading Comprehension.

Reading Comprehension

Passage (*P*) + Question (*Q*) ⟶ Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q Why did Alyssa go to Miami?

A To visit some friends

Stanford Question Answering Dataset (SQuAD)

[Rajpurkar et al., 2016]

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: independent, independent schools, independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: academic, academic, academic

Rather than taxation, what are private schools largely funded by?

Gold answers: tuition, charging their students tuition, tuition

SQuAD Evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1=$\frac{2P*R}{P+R}$
    - Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one.
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1

SQuAD 2.0 Example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft NLNet]

SQuAD 2.0 Leaderboard, 09 Oct 2021

Leaderboard

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

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948

Question Types

- Could be factual vs opinion vs summary
- Factual questions:
  - Yes/no; wh-questions
  - Vary dramatically in difficulty
    - Factoid, List
    - Definitions
    - Why/how..
    - Open ended: ‘What happened?’
  - Affected by form
    - Who was the first president? vs Name the first president

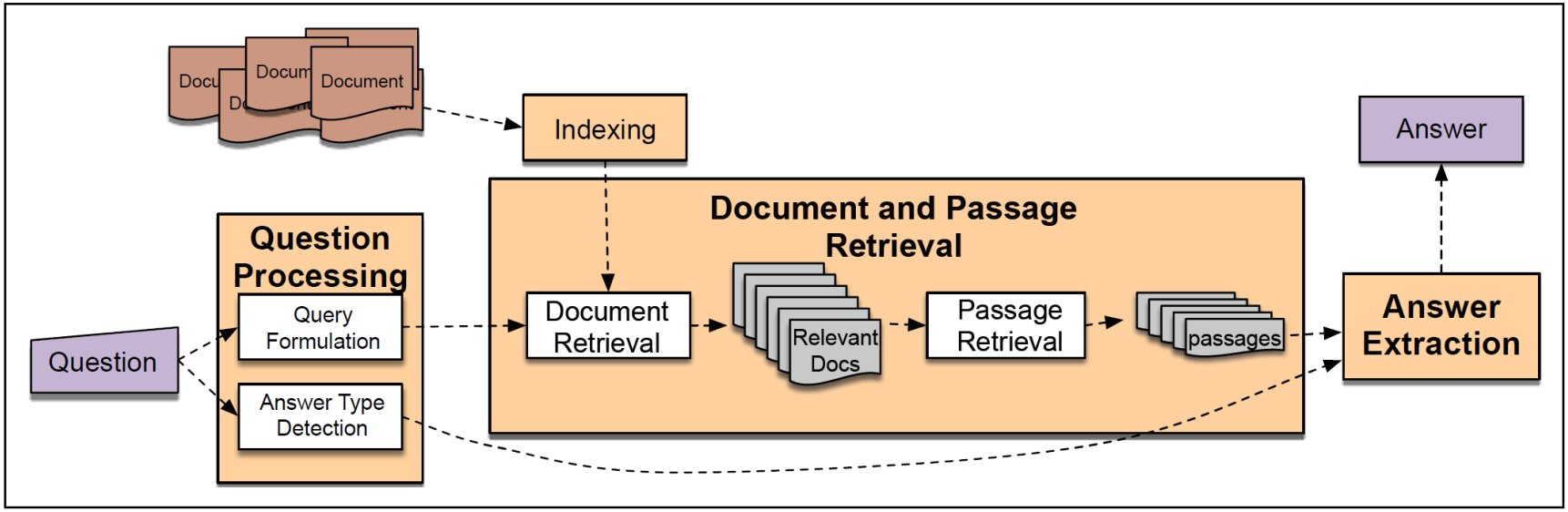
Paradigms for QA

- IR-based
  - Google before 2012
- Knowledge-based
  - Apple Siri
- Hybrid
  - IBM Watson, Google after 2012

IR- based Question Answering

- Google was a pure IR–based QA, but in 2012 **Knowledge Graph** was added to Google's search engine.
  - The Knowledge Graph is a **knowledge base** used by Google to enhance its search engine's search results with semantic--search information gathered from a wide variety of sources.

IR-based Factoid QA

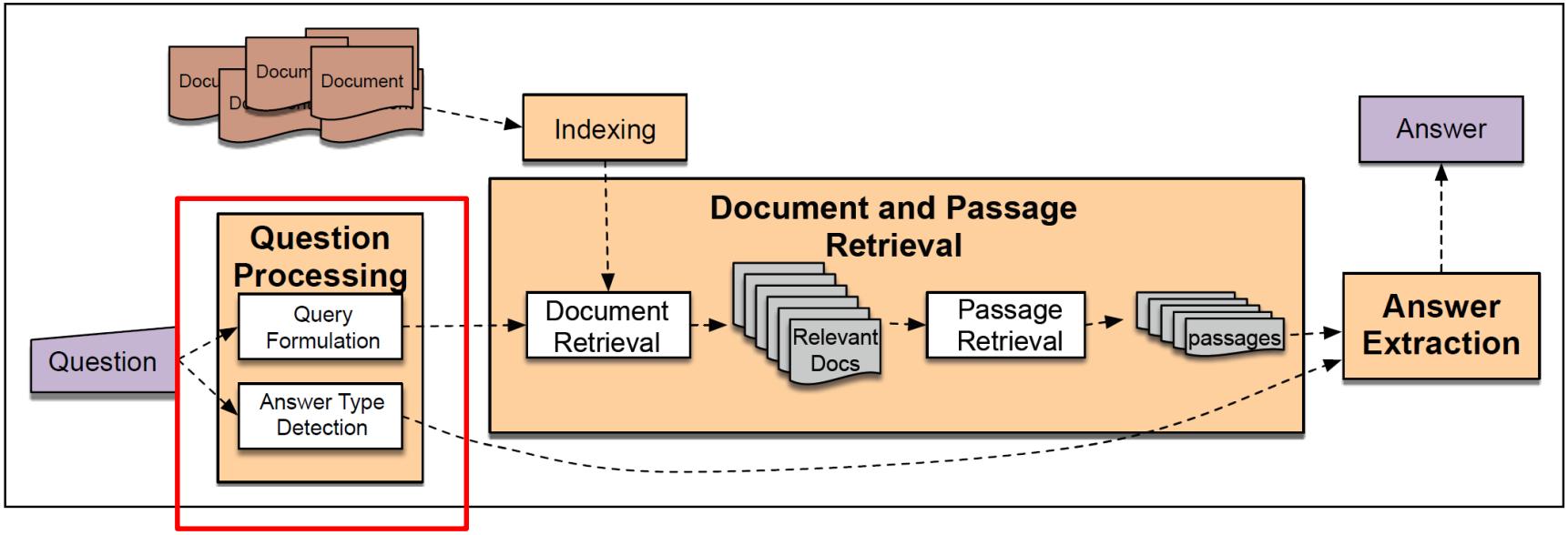


Source: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)

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

IR-based Factoid QA



Source: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)

Question Processing

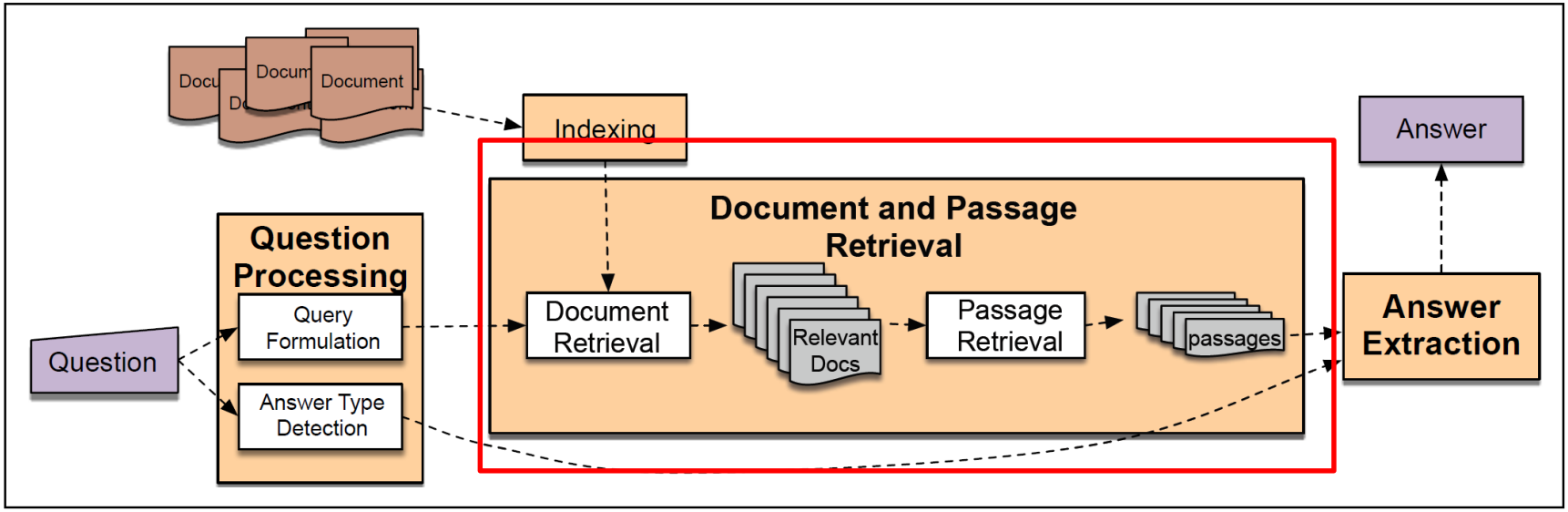
- Answer Type Detection
  - Decide the named entity type (person, place) of the answer
- Query Formulation
  - Create query for the IR system, and make it a declarative sentence
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer

  e.g. *When hit by electrons, a phosphor gives off electromagnetic energy in this form (which form…).*
- Relation Extraction
  - Find relations between entities in the question

  e.g., *They're the two states you could be reentering if you're crossing Florida's northern border.*

  borders(Florida, ?x, north)

IR-based Factoid QA

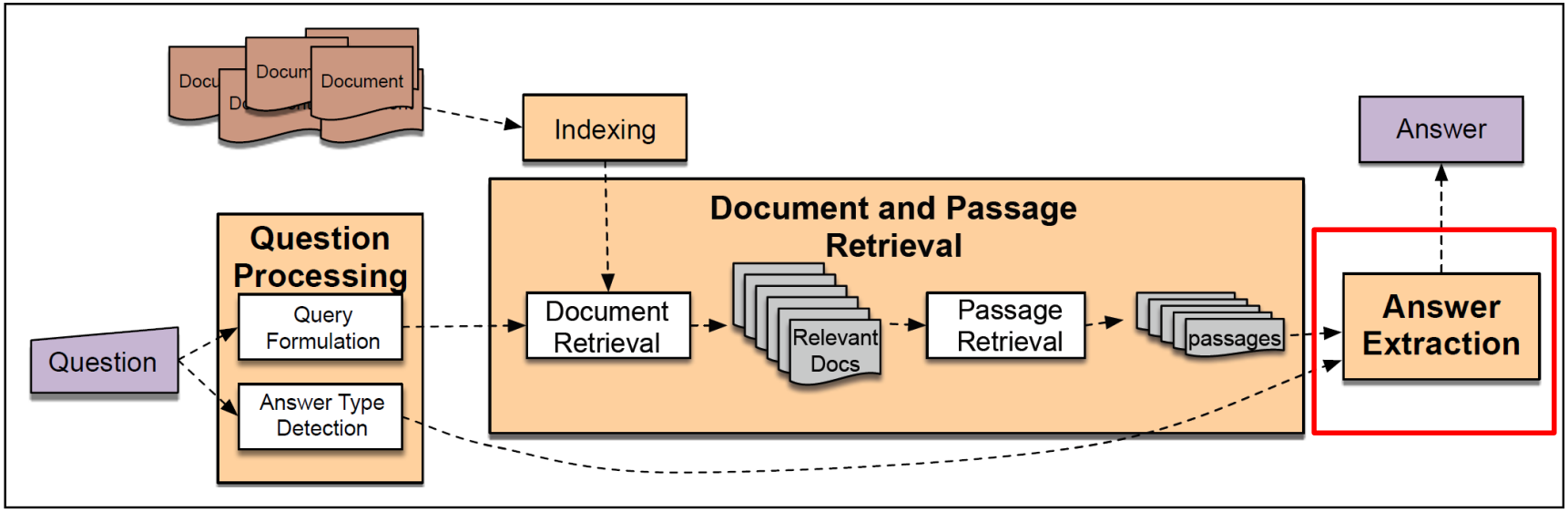


Source: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)

Document and Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units, i.e., **passages**
  - Could be sections, paragraphs, or sentences
- Step 3: Passage ranking
  - Use answer type to help re-rank passages.
  - Use features for ranking, e.g:
    - 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
    - …

IR-based Factoid QA



Source: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)

Answer Extraction

- Type-based Answer Extraction
- Feature-based Answer Extraction
- Neural Answer Extraction

Answer Extraction: Type-based

- 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
- 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!

Answer Extraction: Feature-based

- 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
  - e.g., *John, the dentist*
- 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.

Answer Extraction: Neural Approach

- Neural network approaches to answer extraction draw on the intuition that a question and its answer are semantically similar in some appropriate way. This intuition can be fleshed out by
  - computing an embedding for the question and an embedding for each token of the passage, and
  - then selecting passage spans whose embeddings are closest to the question embedding.

Knowledge-based QA

- Answering a natural language question by mapping it to a query over a structured database.
  - this approach dates back to the earliest days of NLP.
- Systems for mapping from a text string to any logical form are called semantic parsers.
  - Semantic parsers for question answering usually map either to some version of predicate calculus or a query language like SQL or SPARQL.

Question	Logical form
When was Ada Lovelace born?	`birth-year (Ada Lovelace, ?x)`
What states border Texas?	λ x.state(x) $\wedge$ borders(x,texas)
What is the largest state	argmax(λx.state(x), λx.size(x))
How many people survived the sinking of the Titanic	`(count (!fb:event.disaster.survivors fb:en.sinking_of_the_titanic))`

Source: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)

Knowledge-based QA

- The logical form of the question is thus either in the form of a query or can easily be converted into one.
  - The database can be a full relational database, or simpler structured databases like sets of RDF triples :

    e.g., (subject, predicate, object)

    (Ada Lovelace, birth-year, 1815)
  - Freebase, Dbpedia are knowledge bases that use RDF.
    - Freebase is part of Google's Knowledge Graph
    - Dbpedia derives from Wikipedia infoboxes

Knowledge-based QA

- Rule-based approach
  - Write handwritten rules to extract relations from the question (for frequent relations)

  e.g., to extract the *birth-year* relation, we could write patterns that search for the question word *When*, a main verb like *born*, and then extract the named entity argument of *born*, which is *Ada Lovelace*.
- Machine learning approach
  - Given a set of questions paired with their correct logical form (ground truth), the task is then to take those pairs of training tuples and produce a system that maps from new questions to their logical forms.
    - Usually used for extracting relations. Then form the logical forms.

Knowledge-based QA (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

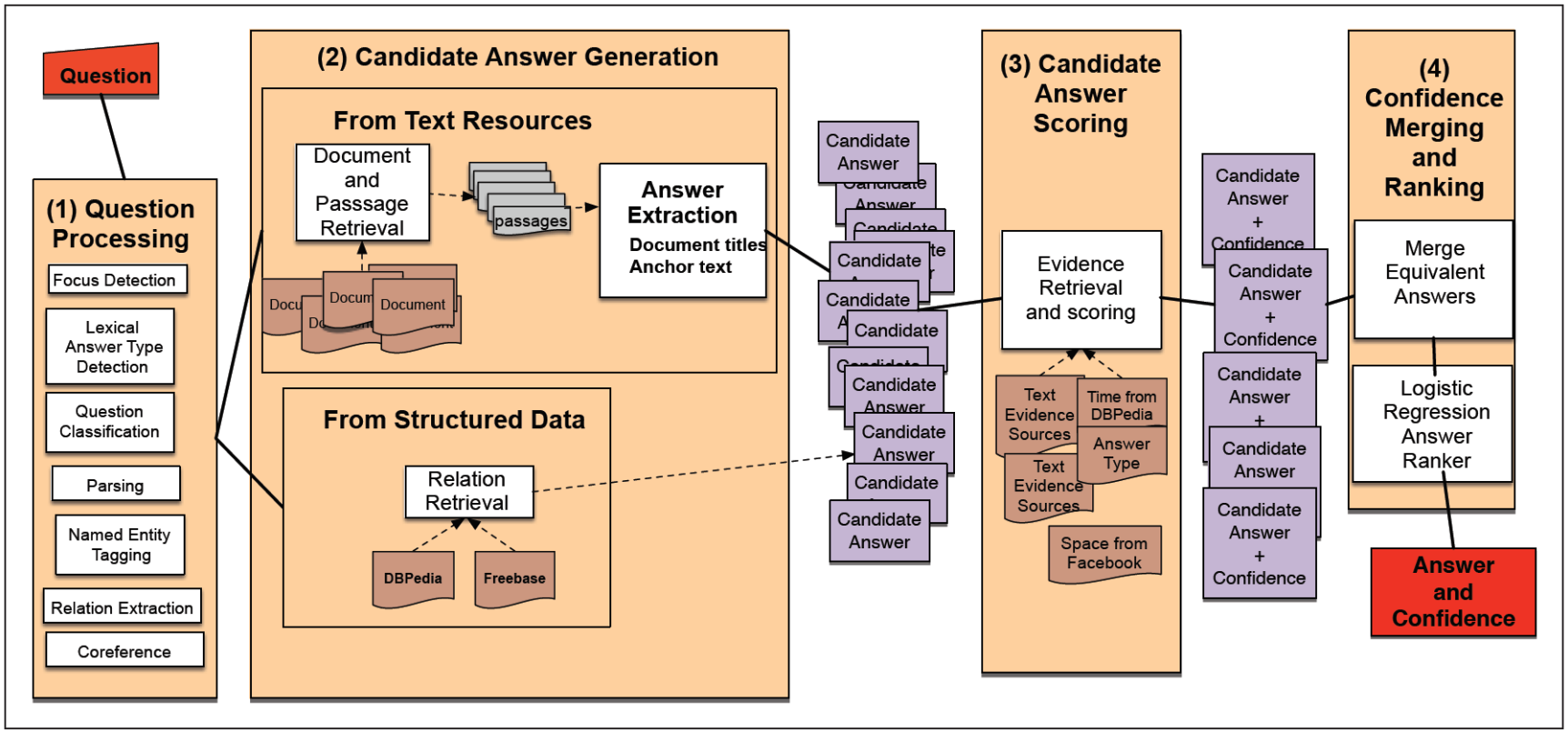
SIRI's main tasks at a high level

- Using ASR (Automatic speech recognition) to transcribe human speech into text.
- Using natural language processing to translate transcribed text into "parsed text":
  - POS tagging, noun--phrase chunking, dependency & constituent parsing
- Using question & intent analysis to analyze parsed text, detecting user commands and actions.
- Using data technologies to interface with 3rd--party web services such as OpenTable, WolframAlpha, to perform actions, search operations, and question answering.
- If SIRI has identified a question, that it cannot directly answer, it will forward to more general question--answering services such as WolframAlpha
- Transforming output of 3rd party web services back into natural language text
- Using TTS (text--to--speech) technologies to transform the natural language text from step 5 above into synthesized speech.

Hybrid QA - 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

Hybrid (QA) – IBM Watson



Common Evaluation Metrics

- Accuracy (F1 score)
  - Does answer match gold--labelled answer
- Mean Reciprocal Rank
  - The *reciprocal rank* of a query response is the inverse of the rank of the first correct answer.
  - The *mean reciprocal rank* is the average of the reciprocal ranks of results for a sample of queries Q.

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

$rank_i$ refers to the rank position of the *first* correct answer for the i-th query.

References

- Dan Jurafsky. CS124, Stanford University.
- Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed.)
- Christopher Manning. CS224N, Stanford University.