Natural Language Processing

Question Answering

Question Answering Systems

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language.

Question answering research attempts to deal with a wide range of question types including: fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions.

Multimodal question answering uses multiple modalities of user input to answer questions, such as text and images.

Is the TV on?

yes

no

How many pets are present?

2

1

What sign is this?

handicap

one way

Is the computer a laptop or a desktop?

desktop

laptop

https://arxiv.org/pdf/1612.00837.pdf

Question Answering Systems

Closed-domain question answering deals with questions under a specific domain (for example, medicine or automotive maintenance), and can exploit domain-specific knowledge frequently formalized in ontologies. Alternatively, closed-domain might refer to a situation where only a limited type of questions are accepted, such as questions asking for descriptive rather than procedural information.

Open-domain question answering deals with questions about nearly anything, and can only rely on general ontologies and world knowledge. On the other hand, these systems usually have much more data available from which to extract the answer.

BASEBALL: AN AUTOMATIC QUESTION-ANSWERER (1961)

https://web.stanford.edu/class/linguist289/p219-green.pdf

Baseball is a computer program that answers questions phrased in ordinary English about stored data. The program reads the question from punched cards. After the words and idioms are looked up in a dictionary, the phrase structure and other syntactic facts are determined for a content analysis, which lists attribute-value pairs specifying the information given and the information requested. The requested information is then extracted from the data matching the specifications, and any necessary processing is done. Finally, the answer is printed. The program's present context is baseball games; it answers such questions as "Where did each team play on July 7?".

LUNAR (1977)

LUNAR answered questions about the geological analysis of rocks returned by the Apollo moon missions.

LUNAR was demonstrated at a lunar science convention in 1971 and it was able to answer 90% of the questions in its domain posed by people untrained on the system.

Factoid questions - questions that can be answered with simple facts expressed in short texts.

Who founded Virgin Airlines?
What is the average age of the onset of autism?
Where is Apple Computer based?

IR-based Factoid Question Answering

Question	Answer
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's the official language of Algeria?	Arabic
How many pounds are there in a stone?	14

Figure 23.1 Some sample factoid questions and their answers.

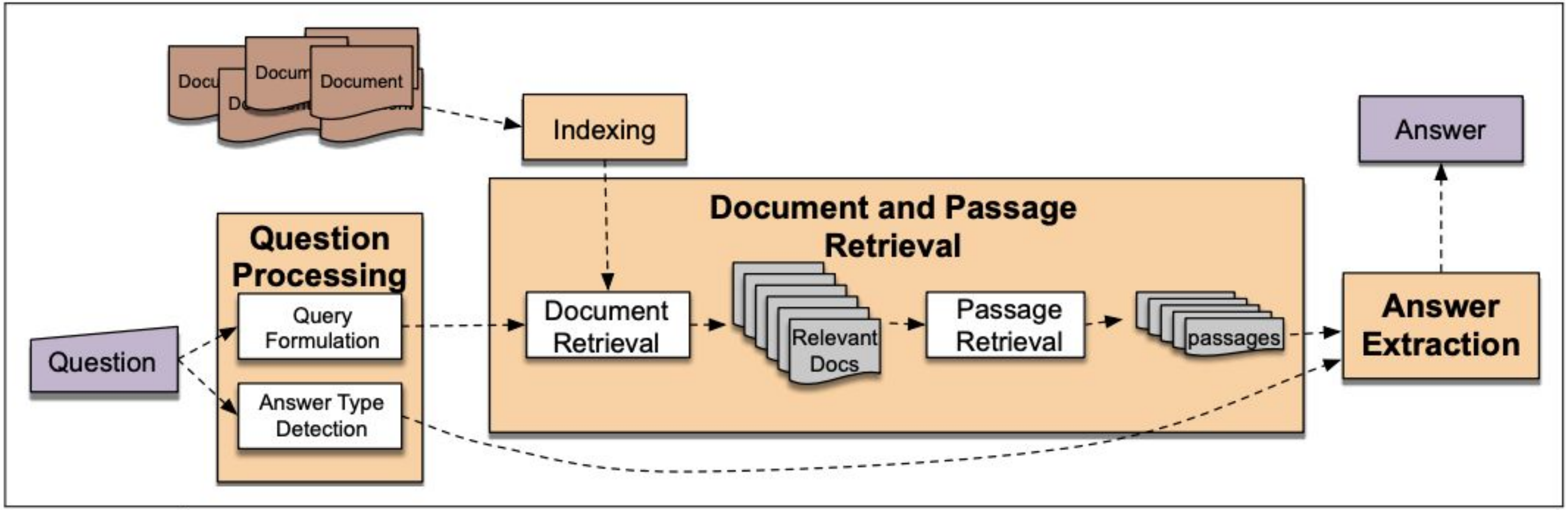

Figure 23.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

Question Processing

The main goal of the question-processing phase is to extract the **query**.

Some systems additionally extract further information such as:

- **answer type**: the entity type (person, location, time, etc.). of the answer;
- **focus**: the string of words in the question that are likely to be replaced by the answer in any answer string found;
- **question type**: is this a definition question, a math question, a list question?

Question Processing

- **question**: Which US state capital has the largest population?
- **query**: “US state capital has the largest population”
- **answer type**: city
- **focus**: state capital

Query Formulation

Query formulation is the task of creating a query—a list of tokens— to send to an information retrieval system to retrieve documents that might contain answer strings.

A query formulation approach that is sometimes used for questioning the web is to apply **query reformulation** rules to the query. The rules rephrase the question to query reformulation make it look like a substring of possible declarative answers. The question “*when was the laser invented?*” might be reformulated as “*the laser was invented*”.

Answer Types

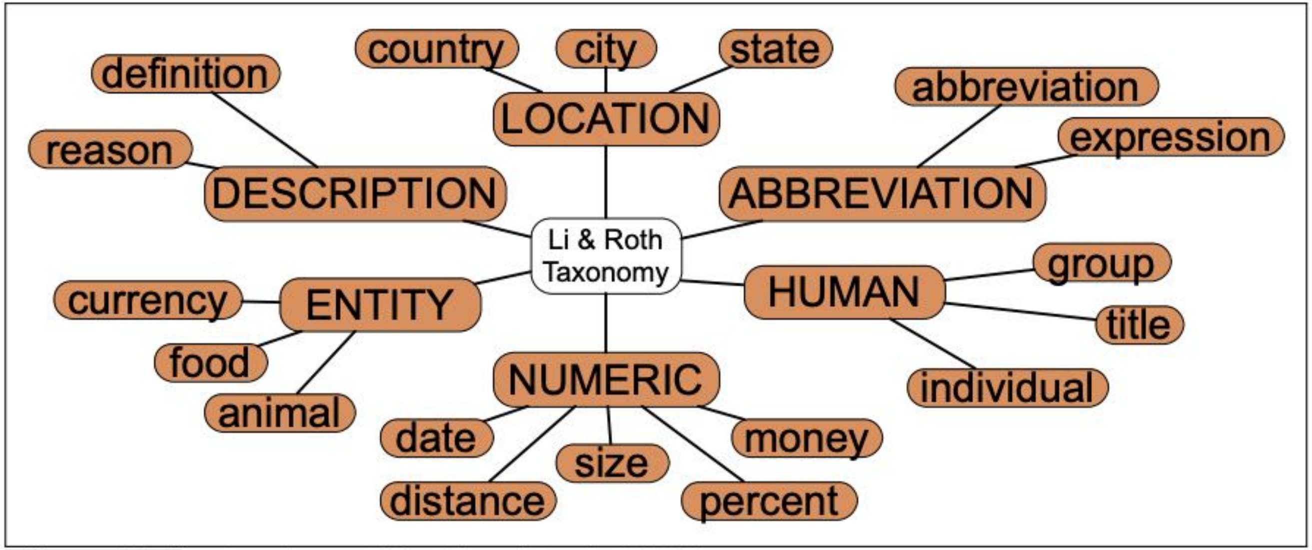

Figure 23.3 A subset of the Li and Roth (2005) answer types.

Document and Passage Retrieval

It's possible to use supervised learning to fully rank the remaining passages, using features like:

- The number of named entities of the right type in the passage;
- The number of question keywords in the passage;
- The longest exact sequence of question keywords that occurs in the passage;
- The rank of the document from which the passage was extracted;
- The proximity of the keywords from the original query to each other;
- The number of n-grams that overlap between the passage and the question

Document and Passage Retrieval

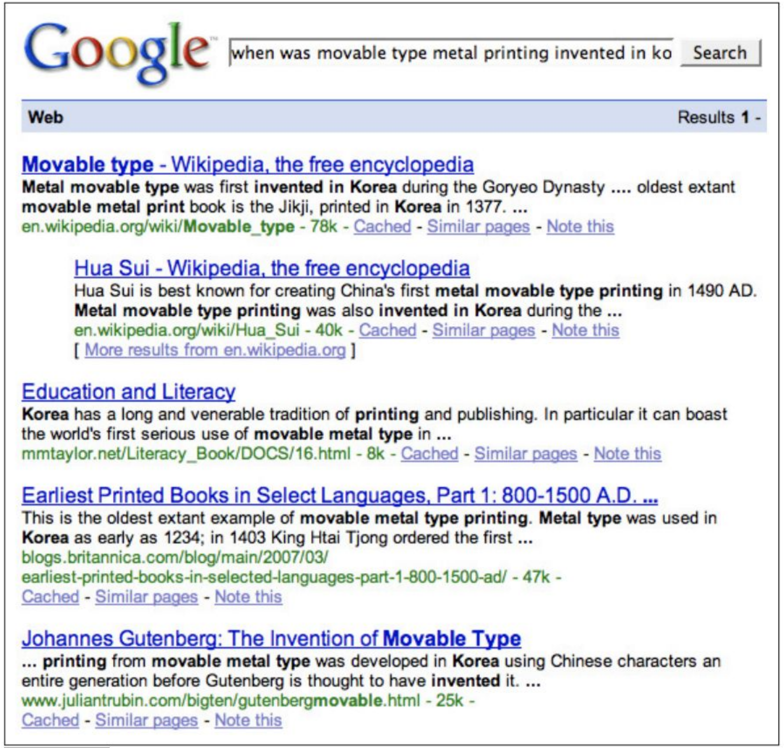

Figure 23.5 Five snippets from Google in response to the query *When was movable type metal printing invented in Korea?*

Feature-based Answer Extraction

Pattern	Question	Answer
<AP> such as <QP>	What is autism?	", developmental disorders such as autism"
<QP>, a <AP>	What is a caldera?	"the Long Valley caldera, a volcanic crater 19 miles long"

Figure 23.6 Some answer-extraction patterns using the answer phrase (AP) and question phrase (QP) for definition questions (Pasca, 2003).

Feature-based Answer Extraction

- **Answer type match**: True if the candidate answer contains a phrase with the correct answer type;
- **Pattern match**: The identity of a pattern that matches the candidate answer;
- **Number of matched question keywords**: How many question keywords are contained in the candidate answer;
- **Keyword distance**: The distance between the candidate answer and query keyword;
- **Novelty factor**: True if at least one word in the candidate answer is novel, that is, not in the query

Neural Answer Extraction

Stanford Question Answering Dataset (**SQuAD**) consists of passages from Wikipedia and associated questions whose answers are spans from the passage, as well as some questions that are designed to be unanswerable; a total of just over 150,000 questions.

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".

Q: "In what city and state did Beyoncé grow up?"
A: "Houston, Texas"

Q: "What areas did Beyoncé compete in when she was growing up?"
A: "singing and dancing"

Q: "When did Beyoncé release Dangerously in Love?"
A: "2003"

Figure 23.7 A (Wikipedia) passage from the SQuAD 2.0 dataset (Rajpurkar et al., 2018) with 3 sample questions and the labeled answer spans.

Neural Answer Extraction

The architecture of the Document Reader component of the **DrQA system**.

DrQA builds an embedding for the question, builds an embedding for each token in the passage, computes a similarity function between the question and each passage word in context, and then uses the question-passage similarity scores to decide where the answer span starts and ends.

Neural Answer Extraction

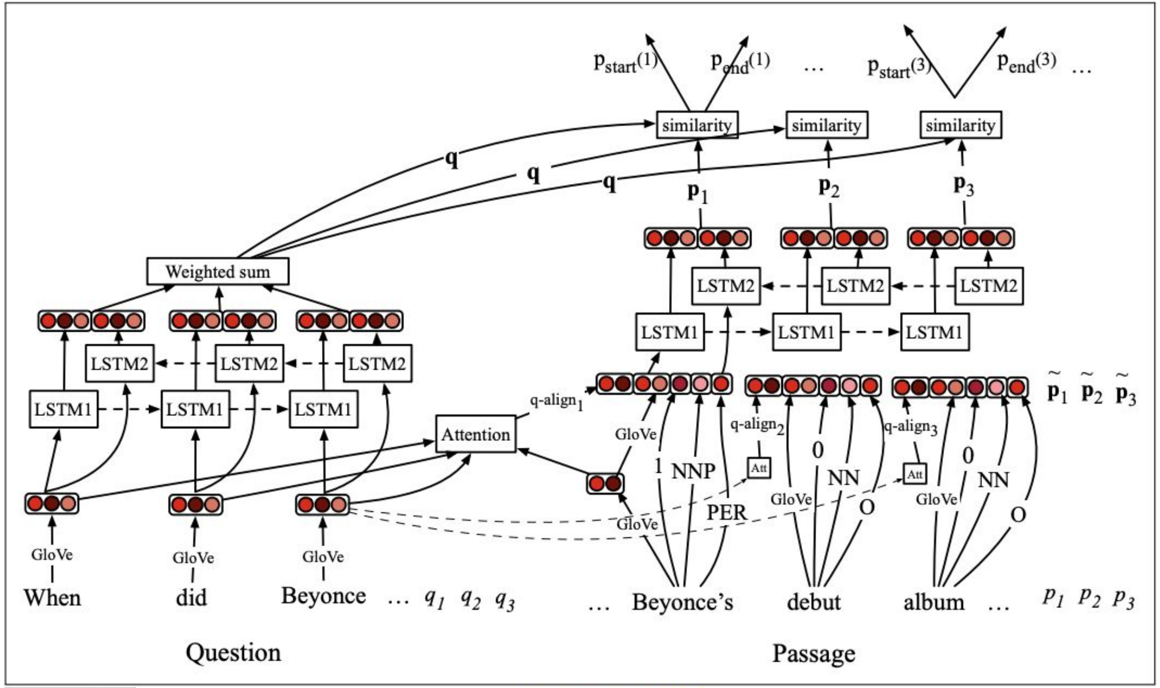

Figure 23.8 The question answering system of Chen et al. (2017), considering part of the question *When did Beyoncé release Dangerously in Love?* and the passage starting *Beyoncé's debut album, Dangerously in Love (2003)*.

Neural Answer Extraction

The question is represented by a single embedding q, which is a weighted sum of representations for each question word q_i .

It is computed by passing the series of embeddings $E(q_1),...,E(q_l)$ of question words through an RNN (such as a bi-LSTM).

The resulting hidden representations $\{q_1,...,q_l\}$ are combined by a weighted sum.

Neural Answer Extraction

To compute the passage embedding $\{p_1,...,p_m\}$ we first form an input representation by concatenating four components:

- An embedding for each word $E(p_i)$ such as from GLoVE;
- Token features like the part of speech of p_i , or the named entity tag of p_i ;
- Exact match features representing whether the passage word p_i occurred in the question
- Aligned question embedding: many QA systems use an attention mechanism to give a more sophisticated model of similarity between the passage and question words, such as similar but non identical words like ***release*** and ***singles***.

Neural Answer Extraction

The result of the previous two steps is a single question embedding q and a representations for each word in the passage $\{p_1,...,p_m\}$.

In order to find the answer span, we can train two separate classifiers, one to compute for each p_i the probability $p_{start}(i)$ that p_i is the start of the answer span, and one to compute the probability $p_{end}(i)$.

Knowledge-based Question Answering

Knowledge-based question answering is for the idea of answering a natural language question by mapping it to a query over a structured database.

Systems for mapping from a text string to any logical form are called **semantic parsers**.

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))`

Figure 23.9 Sample logical forms produced by a semantic parser for question answering. These range from simple relations like `birth-year`, or relations normalized to databases like Freebase, to full predicate calculus.

Knowledge-based Question Answering: RDF-triples

subject	**predicate**	**object**
Ada Lovelace	birth-year	1815

“When was Ada Lovelace born?” → `birth-year (Ada Lovelace, ?x)`
“What is the capital of England?” → `capital-city(?x, England)`

IBM's Watson

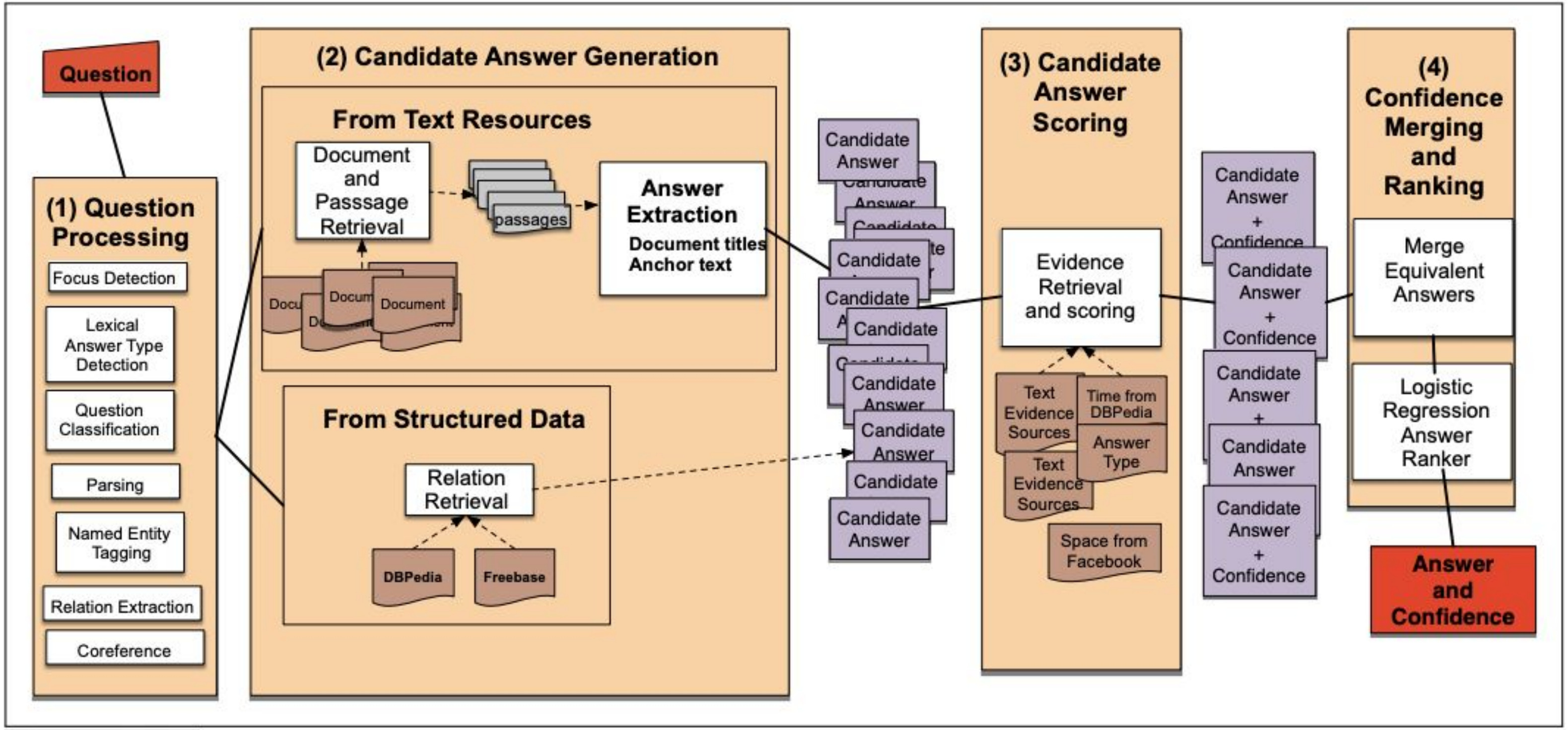

Figure 23.11 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

IBM’s Watson

The first stage is **question processing**. The DeepQA system runs parsing, named entity tagging, and relation extraction on the question.

Then the DeepQA system extracts the **focus**, the **answer type**, and performs **question classification** and **question sectioning**.

IBM’s Watson

In the second candidate **answer generation stage**, we combine the processed question with external documents and other knowledge sources to suggest many candidate answers. These candidate answers can either be extracted from text documents or from structured knowledge bases.

IBM’s Watson

The third **candidate answer scoring stage** uses many sources of evidence to score the candidates.

One of the most important is the lexical answer type. DeepQA includes a system that takes a candidate answer and a lexical answer type and returns a score indicating whether the candidate answer can be interpreted as a subclass or instance of the answer type.

IBM's Watson

The final **answer merging and scoring step** first merges candidate answers that are equivalent.

Thus if we had extracted two candidate answers *J.F.K.* and *John F. Kennedy*, this stage would merge the two into a single candidate.

IBM’s Watson

Now we have a set of candidates, each with a feature vector.

A **classifier** takes each feature vector and assigns a confidence value to this candidate answer.

The classifier is trained on thousands of candidate answers, each labeled for whether it is correct or incorrect, together with their feature vectors, and learns to predict a probability of being a correct answer.

IBM’s Watson

Evaluation of Factoid Answers

Reading comprehension systems on datasets like SQuAD are often evaluated using two metrics, both ignoring punctuations and articles (a, an, the):

- **Exact match**: The percentage of predicted answers that match the gold answer exactly;
- **F1 score**: The average overlap between predicted and gold answers. Treat the prediction and gold as a bag of tokens, and compute F1, averaging the F1 over all questions.

Demo

https://demo.ipavlov.ai/#en