Information Retrieval

Question Answering

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°

IBM's Watson


ΣΚΕΨΟΥ
PIENSE
DENK
THINK
ДУМАЙ
सोचिए
$24,000
$77,147
WATSON
$21,600

Types of Questions

- Factoid questions
  - Who wrote ”The Mahabharata“
  - What is Microsoft’s HQ located
  - Where does Liverpool stand in the EPL

- Complex (narrative) questions
  - What do scholars think about the Jefferson’s position on dealing with pirates
  - People’s opinion on Biden winning the US elections

Paradigms

- Information-retrieval based
  - Uses text datasets
  - Google
- Knowledge based
  - Uses semantic parsers and structured knowledge bases
  - Siri, Wolfram Alpha
- Hybrid
  - Uses multiple information sources
  - IBM Watson

Question Processing

Phases of IR-Based Question Answering

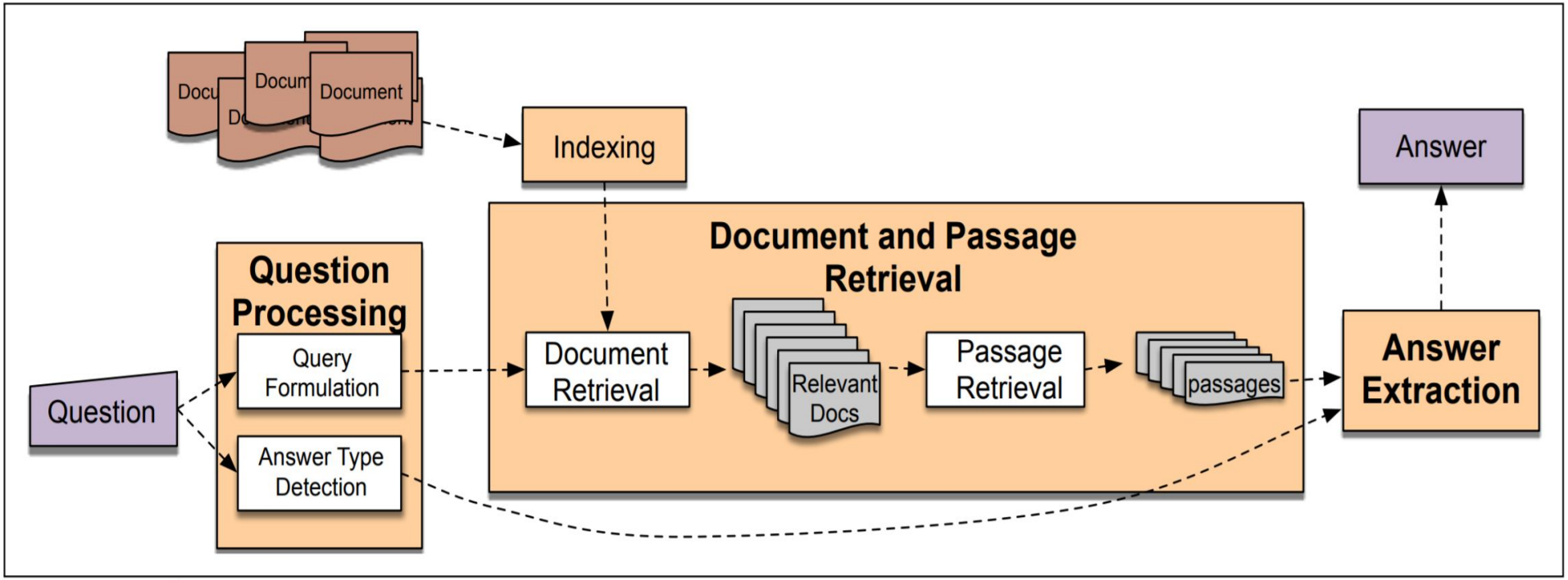


Figure 25.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

Question Processing

Goal : Extract the query and pass the *keywords* to the IR system to match potential documents.

Extracting information from question ,

- ***answer type***: the entity type
- ***focus***: the string of words in the question that is 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?
- ***relation extraction***: finds relation between entities

Query Formulation

Definition - task of creating a query-a list of tokens, to send to an information retrieval system to retrieve documents that might contain answer strings.

Query expansion methods can add query terms in hopes of matching the particular form of the answer as it appears, like adding morphological variants of the content words in the question, or synonyms from a thesaurus.

Query Reformulation

Query reformulation applies rules to the query to rephrase the question to make it look like a substring of possible declarative answers.

Example:

- when was the laser invented? -> the laser was invented
- where is the Valley of the Kings? -> the Valley of the Kings is located in

Answer Types

Question classification, the task of finding the answer type, the **named-entity categorizing the answer.**

Example**:**

Who founded Virgin Airlines? expects an answer of type PERSON.

What Canadian city has the largest population? expects an answer of type CITY.

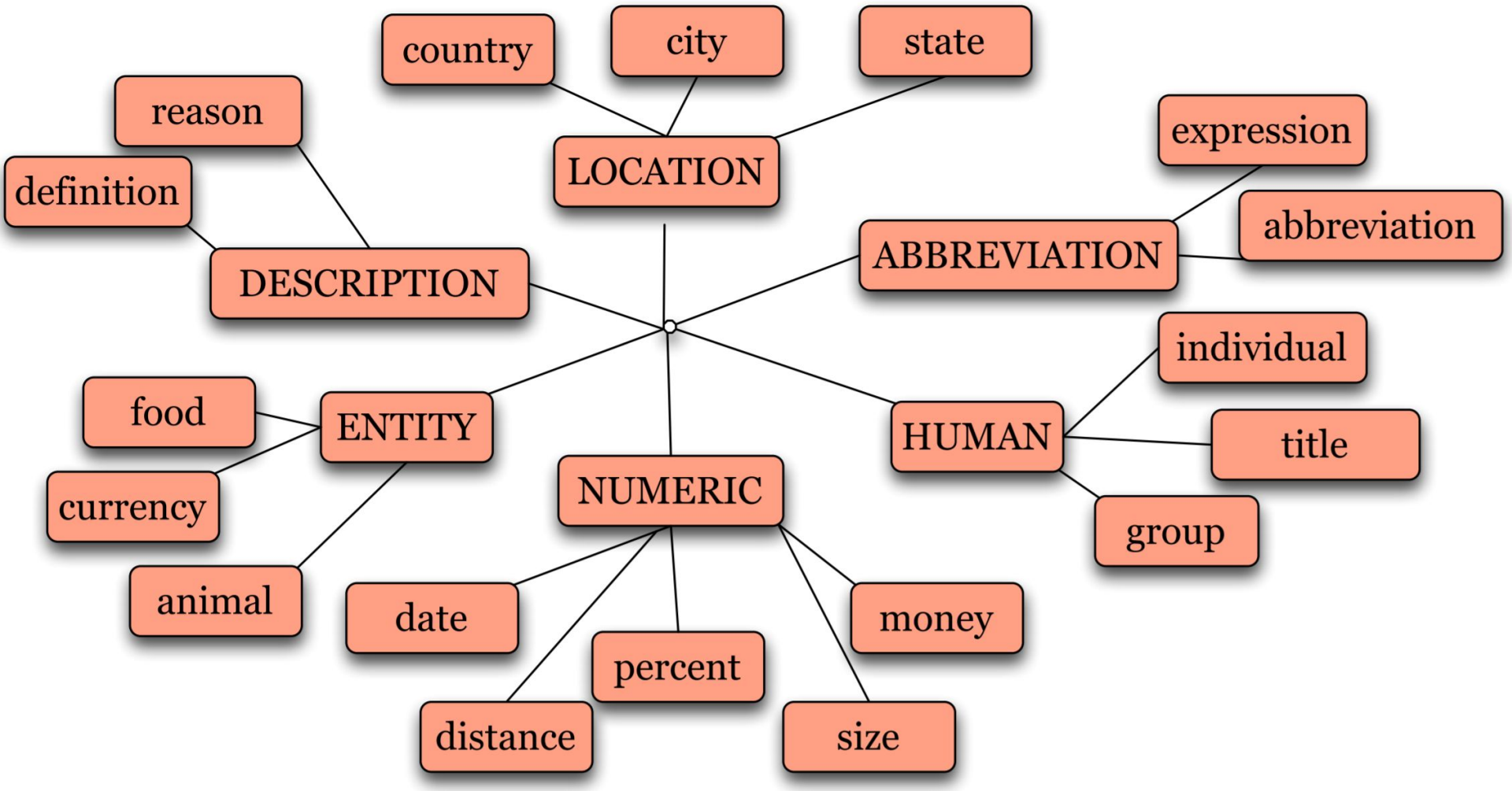
** Helps to avoid scanning whole document for answer.

Answer Type Taxonomy

- Taxonomy can be built from WordNet or can be hand built.
- Li and Roth tagset is one of the popular ones where each question can be labeled with coarse grained or fine grained tags.
- Question classifiers can be built by hand-writing rules , for ex:

  For detecting the answer type BIOGRAPHY:

  who {is | was | are | were} PERSON


country
city
state
reason
LOCATION
expression
definition
abbreviation
ABBREVIATION
DESCRIPTION
individual
food
ENTITY
HUMAN
title
currency
NUMERIC
group
animal
date
money
percent
distance
size

Question Classifiers

- **Supervised learning** - trained on databases of questions that have been hand-labeled with an answer type.
- **Feature Based** - rely on words in the questions and their embeddings, the part-of-speech of each word, and named entities in the questions.
- **Feature** - a single word in the question gives extra information about the answer type, and its identity is used as a **feature.**

Features for Answer Type Detection

Question words and phrases

Part-of-speech tags

Parse features (headwords)

Named Entities

Semantically related words

- **answer type word/question headword** - the headword of the first NP after the question's wh-word.

Example:

When will **college** reopen?

Which **country** has the highest corona death rate?

Answer Extraction

- **Span Labeling:** given a passage, identifying the span of text which constitutes an answer
- A simple baseline algorithm is to run a named entity tagger on the candidate passage and return whatever span in the passage is the correct answer type
- ***Example:-***

  'How tall is Mt. Everest?"

  The official height of Mount Everest is 29029 feet

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"

Features in such classifiers include:-

1. *Answer type match*
2. *Pattern match*
3. *Number of matched question keywords*
4. *Keyword distance*
5. *Novelty factor*
6. *Apposition features*
7. *Punctuation location*
8. *Sequence of question terms*

N-gram tiling answer extraction

- Every unigram, bigram and trigram occuring in the snippet is extracted and weighted.
- In n-gram filtering, n-grams are scored by how well they match the predicted answer type.
- N-gram tiling algorithm concatenates overlapping n-gram fragments into longer answers.
- The best-scoring concatenation is added to the set of candidates, the lower-scoring candidate is removed, and the process continues until a single answer is built.

Neural Answer Extraction

- A bi-LSTM-based Reading Comprehension Algorithm
- BERT-based Question Answering

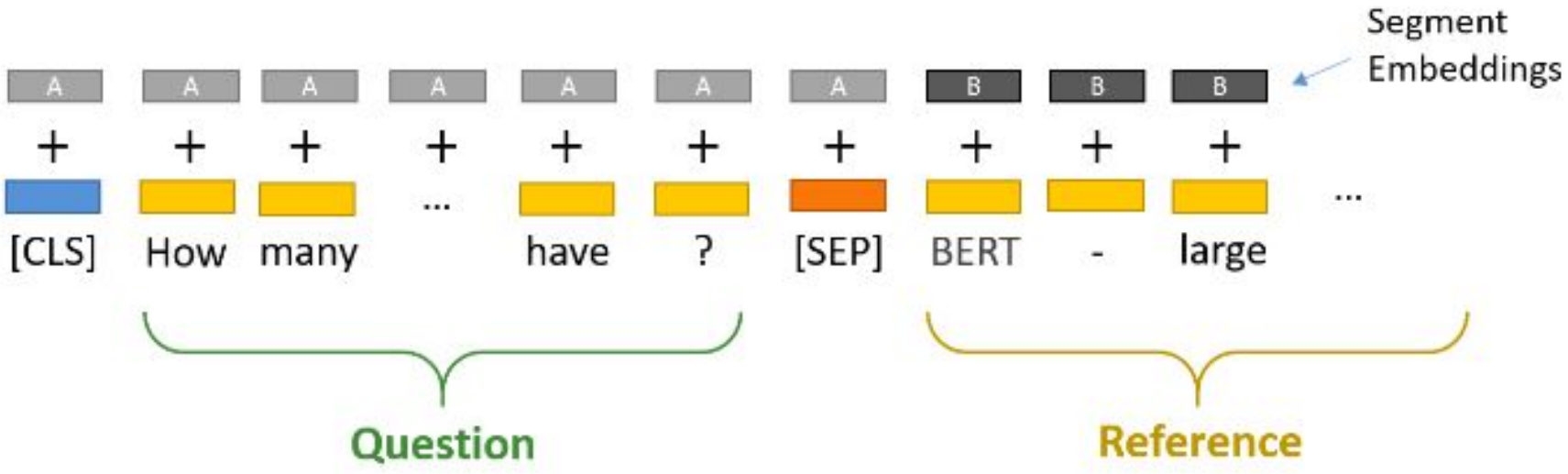
Bi-LSTM based Question Answering

- Neural algorithms for reading comprehension are given a question *q* of *l* tokens $q_1,...,ql$ and a passage p of *m* tokens $p_1,..., p_m$.
- Their goal is to compute, for each token *pi* the probability $p_{start}(i)$ that p_i is the start of the answer span, and the probability $p_{end}(i)$ that p_i is the end of the answer span.

BERT-based Question Answering

The input is processed in the following way:

1. Token embeddings
2. Segment embeddings



Document and Passage Retrieval

Answer-extraction methods are designed to apply to smaller regions such as paragraphs.

Knowledge-based Question Answering

Definition: Answering a natural language question by mapping it to a query over a structured database

Dates back to the earliest days of NLP, with systems like BASEBALL (Green et al., 1961) that answered questions from a db of baseball games and stats

Semantic parsers: Systems for mapping from a text string to any logical form

Semantic Parsing

Question	**Logical Form**
When was Ada Lovelace born?	Birth-year (Ada Lovelace, ?)
What states border TN?	λ x.state(x) Λ borders(x, TN)
What is the largest state?	Argmax(λ x.state(x), λ x.size(x))

Rule-based methods

Definition: Handwritten rules to extract relations from the question, especially for relations that are very frequent.

Supervised Methods

Definition: Given a set of questions paired with their correct logical forms, produce a system that maps from new questions to their most logical forms.

Semi-Supervised Methods

Definition: Semi-supervised machine learning is a combination of supervised and unsupervised machine learning methods.

It is difficult to create training sets with questions labeled with their meaning representations.

Semantic Parsing

capital of	capital city of	become capital of
capitol of	national capital of	official capital of
political capital of	administrative capital of	beautiful capital of
capitol city of	remain capital of	make capital of
political center of	bustling capital of	capital city in
cosmopolitan capital of	move its capital to	modern capital of
federal capital of	beautiful capital city of	administrative capital city of

Paraphrase Databases

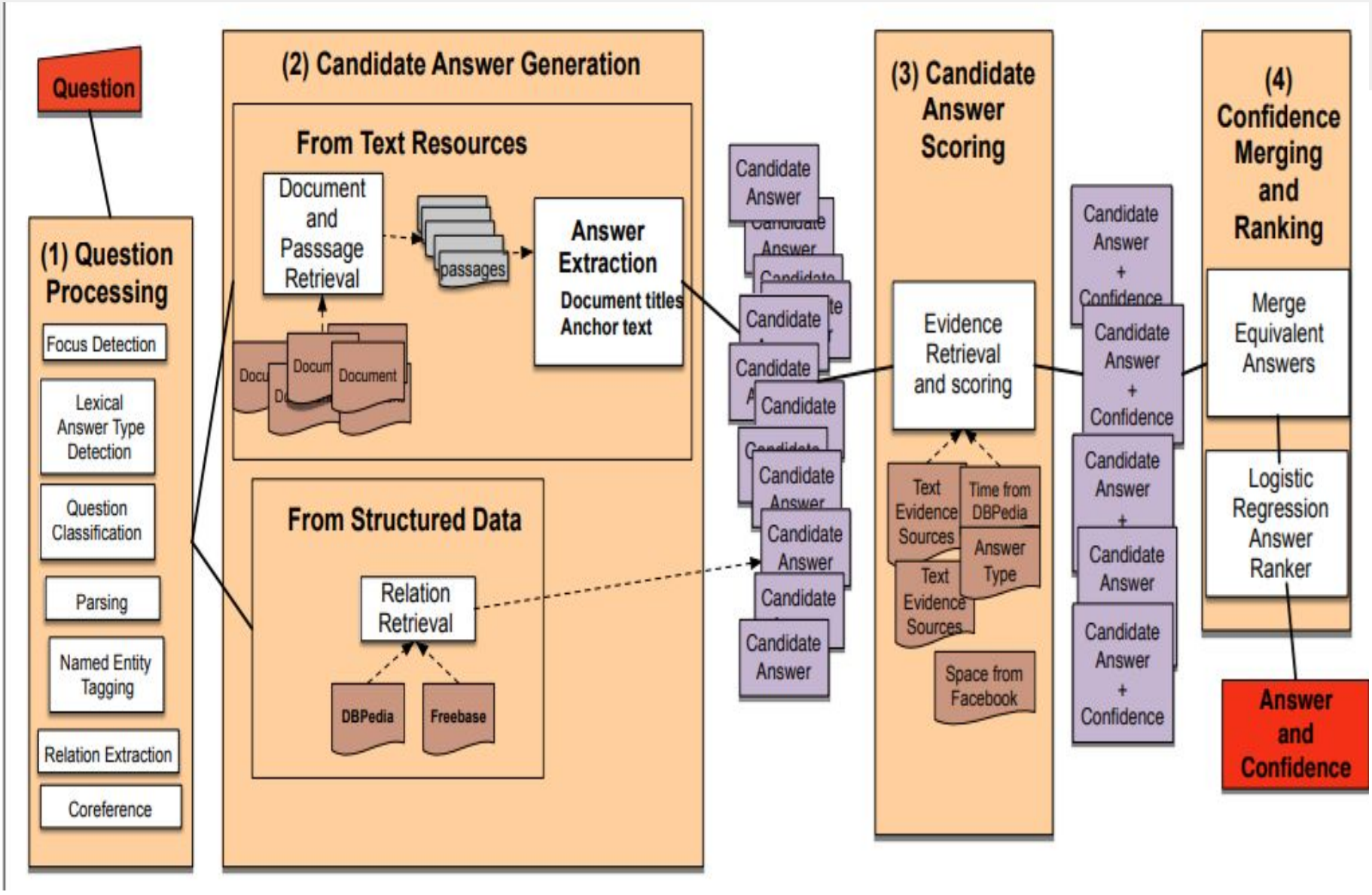
Q: What are the green blobs in plant cells?
Lemmatized synonyms from PARALEX:
what be the green blob in plant cell?
what be green part in plant cell?
what be the green part of a plant cell?
what be the green substance in plant cell?
what be the part of plant cell that give it green color?
what cell part do plant have that enable the plant to be give a green color?
what part of the plant cell turn it green?
part of the plant cell where the cell get it green color?
the green part in a plant be call?
the part of the plant cell that make the plant green be call?

IBM Watson: Question Answering

- Won Jeopardy on February 16, 2011!
- This system relies on a wide variety of resources to answer questions.

What is Jeopardy!?

Jeopardy! is an American television game show created by Merv Griffin. The show features a quiz competition in which contestants are presented with general knowledge clues in the form of answers, and must phrase their responses in the form of questions.


Question
(1) Question Processing
Focus Detection
Lexical Answer Type Detection
Question Classification
Parsing
Named Entity Tagging
Relation Extraction
Coreference
(2) Candidate Answer Generation
From Text Resources
Document and Passsage Retrieval
passages
Document
Answer Extraction
Document titles
Anchor text
From Structured Data
Relation Retrieval
DBPedia
Freebase
Candidate Answer
(3) Candidate Answer Scoring
Evidence Retrieval and scoring
Text Evidence Sources
Time from DBPedia
Answer Type
Text Evidence Sources
Space from Facebook
Candidate Answer + Confidence
(4) Confidence Merging and Ranking
Merge Equivalent Answers
Logistic Regression Answer Ranker
Answer and Confidence

Jeopardy! example

- Poets and Poetry: He was a bank clerk in the Yukon before he published “Songs of a Sourdough” in 1907.

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

Relations

authorof(focus,“Songs of a sourdough”)

publish (e1, he, “Songs of a sourdough”)

in (e2, e1, 1907)
temporallink(publish(...), 1907)

Evaluation of Factoid Answers

- A common evaluation metric for factoid question answering, introduced in the TREC Q/A track in 1999, is mean reciprocal rank.
- MRR assumes a test set of mean reciprocal rank MRR questions that have been human-labeled with correct answers.

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

References

Speech and Language Processing. Daniel Jurafsky & James H. Martin.