Natural Language Processing
SoSe 2015

HPI Hasso
Plattner
Institut

IT Systems Engineering | Universität Potsdam

Question Answering

Dr. Mariana Neves   *July 6th, 2015*

(based on the slides of Dr. Saeedeh Momtazi)
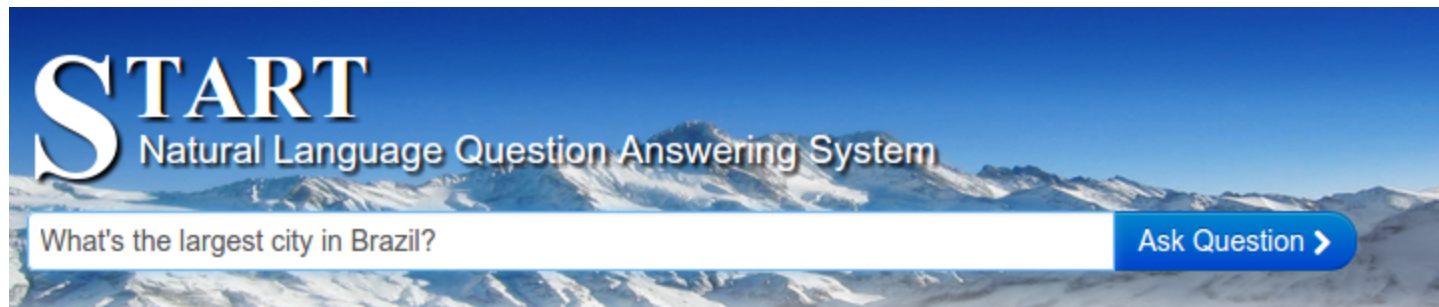
Outline

- Introduction

- History

- QA Architecture

Outline

- Introduction

- History

- QA Architecture

Motivation

- Finding small segments of text which answer users' questions

===> What's the largest city in Brazil?

Brazil

Capital (2009 est.): Brasília, 3,789,000

Largest cities: São Paulo, 19,900,000; Rio de Janeiro, 11,836,000; Salvador, 2,590,400; Belo Horizonte, 5,736,000; Recife, 1,485,500; Porto Alegre, 4,034,000

(http://start.csail.mit.edu/)

Motivation

Motivation

Motivation

- How many presidents do the USA have?

    - 44

    - List of names in chronological order

40	Ronald Reagan[3][4]	Actor & broadcaster, President of the Screen Actors Guild[5]
41	George H. W. Bush	Pilot (Navy Lieutenant, Junior Grade),[6] Businessman (Oil)[7]
42	Bill Clinton	Lawyer, Law lecturer
43	George W. Bush	Businessman (Oil,[8] baseball)[9]
44	Barack Obama	Public Interest professional, Lawyer, Constitutional Law Professor [10]

(http://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States_by_occupation)

Motivation

- Information retrieval

  - Keywords (short input)

  - Document (long output)

- Question Answering

  - Natural language question (long input)

  - Short answer (short output)

QA Types

- Closed-domain
  - Answering questions from a specific domain

- Open-domain
  - Answering any domain independent question

- Closed-domain vs. Open-domain
  - Using a large collection of unstructured data (e.g., the Web) instead of databases

Open-domain QA

- Advantages

  - Many subjects are covered

  - Information is constantly added and updated

  - No manual work is required to build databases

Open-domain QA

- Problems
  - Information is not always up-to-date
  - Wrong information cannot be avoided
  - Much irrelevant information is found

Outline

- Introduction

- History

- QA Architecture

History

- BASEBALL [Green et al., 1963]

  - One of the earliest question answering systems

  - Developed to answer users' questions about dates, locations, and the results of baseball matches

- LUNAR [Woods, 1977]

  - Developed to answer natural language questions about the geological analysis of rocks returned by the Apollo moon missions

  - Able to answer 90% of questions in its domain posed by people not trained on the system

History

- STUDENT

    – Built to answer high-school students' questions about algebraic exercises

- PHLIQA

    – Developed to answer the user's questions about European computer systems

- UC (Unix Consultant)

    – Answered questions about the Unix operating system

- LILOG

    – Was able to answer questions about tourism information of cities in Germany

Open-domain QA

- START [Katz, 1997]

  – Utilized a knowledge-base to answer the user's questions

  – The knowledge-base was first created automatically from unstructured Internet data

  – Then it was used to answer natural language questions

  http://start.csail.mit.edu/index.php

Open-domain QA
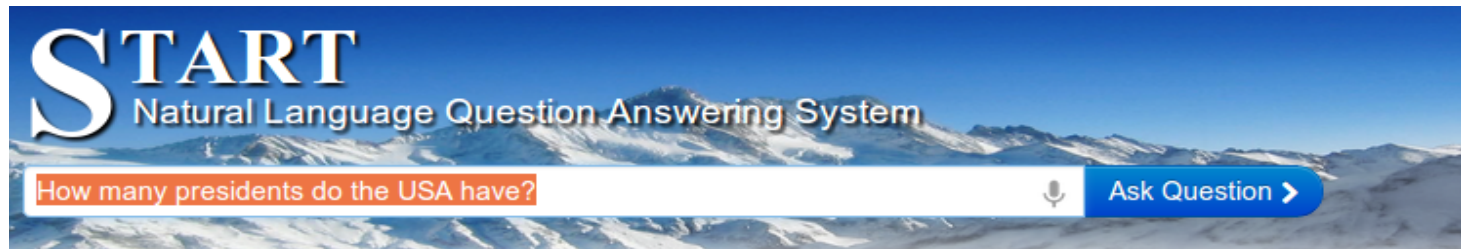
===> How many presidents do the USA have?

United States

Executive branch:
chief of state: President Barack H. OBAMA (since 20 January 2009); Vice President Joseph R. BIDEN (since 20 January 2009); note - the president is both the chief of state and head of government
head of government: President Barack H. OBAMA (since 20 January 2009); Vice President Joseph R. BIDEN (since 20 January 2009)
cabinet: Cabinet appointed by the president with Senate approval

(For more information visit the World Leaders website)
elections: president and vice president elected on the same ticket by a college of representatives who are elected directly from each state; president and vice president serve four-year terms (eligible for a second term); election last held 6 November 2012 (next to be held on 8 November 2016)
election results: Barack H. OBAMA reelected president; percent of popular vote - Barack H. OBAMA 50.6%, Mitt ROMNEY 47.9%, other 1.5%;

IBM Watson

- Playing against two greatest champions of Jeopardy

- Challenges

  - Knowledge

  - Speed

  - Confidence

- Final Jeopardy, Stephen Backer

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

  https://www.youtube.com/watch?v=WFR3lOm_xhE

Outline

- Introduction

- History

- QA Architecture

Architecture

Natural Language Processing

Question Analysis → Question Classification → Query Construction

How many presidents do the USA have?

Corpus ← Document Retrieval

Sentence Retrieval

Sentence Annotation

Web

44 ← Answer Validation ← Answer Extraction

Information Retrieval

Information Extraction

19
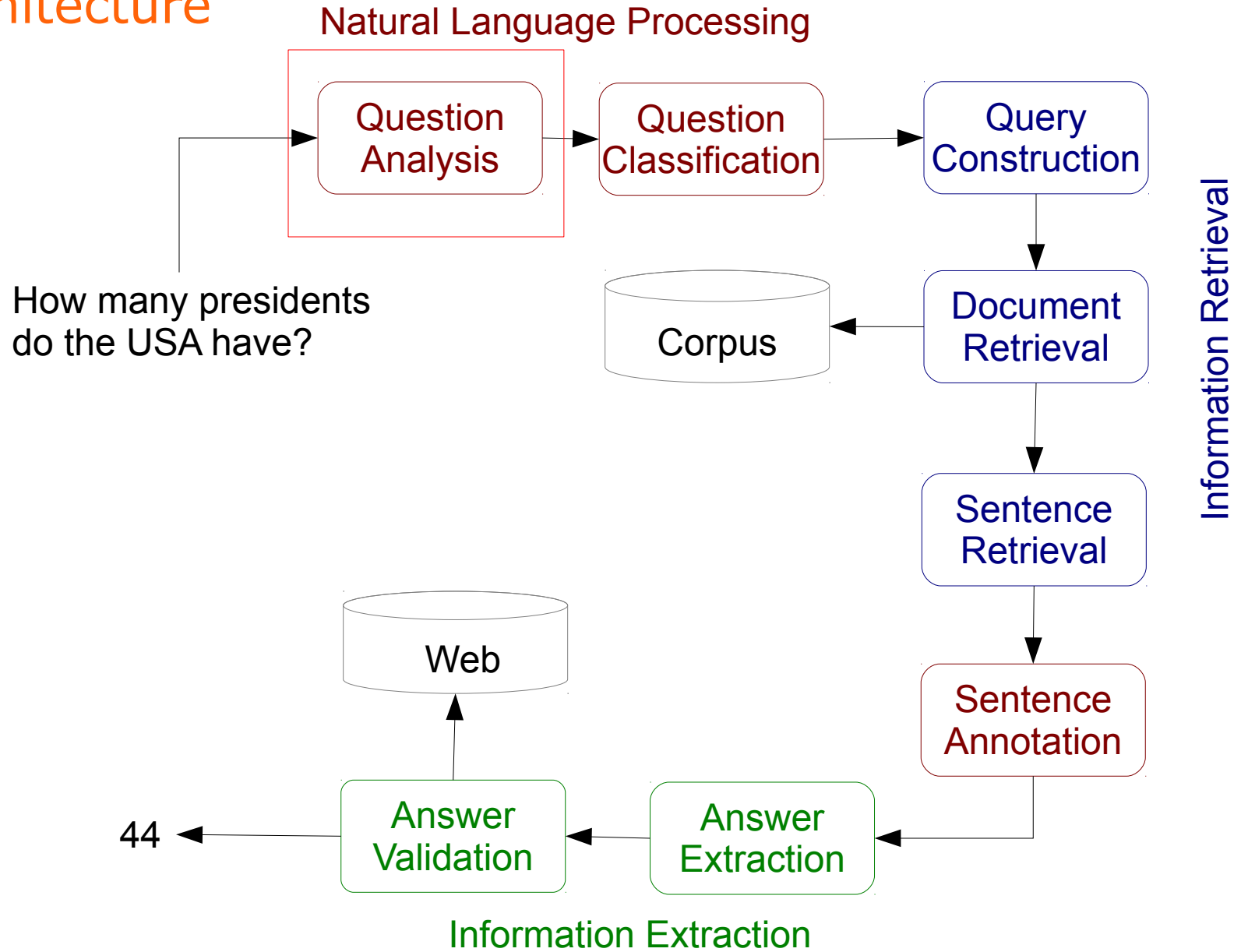
Question Analysis

- Named Entity Recognition

- Surface Text Pattern Learning

- Syntactic Parsing

- Semantic Role Labeling

Question Analysis: Named Entity Recognition

- Recognizing the named entities in the text to extract the target of the question

- Using the question's target in the query construction step

- Example:

    - Question: "In what country was Albert Einstein born?"

    - Target: "Albert Einstein"

Question Analysis: Pattern Learning

- Extracting a pattern from the question

- Matching the pattern with a list of pre-defined question patterns

- Finding the corresponding answer pattern

- Realizing the position of the answer in the sentence in the answer extraction step

Question Analysis: Pattern Learning

- Example:

  - Question: "In what country was Albert Einstein born?"

  - Question Pattern: "In what country was X born?"

  - Answer Pattern: "X was born in Y."

Question Analysis: Syntactic Parsing

- Using a dependency parser to extract the syntactic relations between question terms

- Using the dependency relation paths between question terms to extract the correct answer in the answer extraction step

(http://nlp.stanford.edu:8080/corenlp/process)
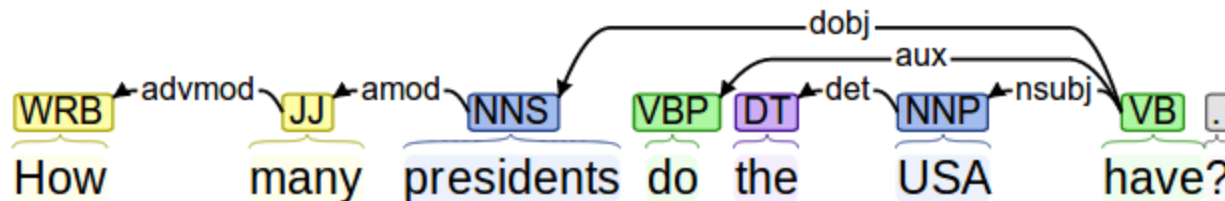
Question Analysis: Semantic Role Labeling

- FrameNet: a lexical database for English

- More than 170,000 manually annotated sentences

- Frame Semantics: describes the type of event, relation, or entity and the participants in it.

- Example:

    – John$_{[Cook]}$ grills a fish$_{[Food]}$ on an open fire $_{[Heating-Instrument]}$

Container [Container]
Semantic Type: Container

The Container holds the Food to which heat is applied.
BOIL the potatoes in a medium-sized pan.
Things that apply the heat directly are Heating_Instruments, e.g. crock-pot, electric skillet.

Cook [Cook]
Semantic Type: Sentient

The Cook applies heat to the Food.
Drew SAUTEED the garlic in butter.

Food [Food]

Food is the entity to which heat is applied by the Cook.
Suzy usually STEAMS the broccoli.
In instructional imperatives, this FE, which would be used for the (missing) object, is tagged CNI:
COOK on low heat for two hours. CNI

Heating_instrument [Heat_instr]
Semantic Type: Physical_entity

This FE identifes the entity that directly supplies heat to the Food.
Jim BROWNED the roast in the oven.
This FE will take precedence over Container when both are expressed in the same constituent. For example:
Kate COOKED the rice in a rice-cooker.

Temperature_setting [Temp]
Semantic Type: Temperature

This FE identifies the Temperature_setting of the Heating_instrument for the Food.
He BAKED the cookies at 350 degrees for 11 minutes.

She MICROWAVED the popcorn on high.

You can't COOK popcorn on low heat!
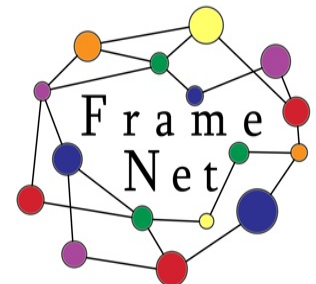
(https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Apply_heat)

Question Analysis: Semantic Role Labeling

- Frame assignment

- Role labeling

- Jim[Driver] flew his plane[Vehicle] to Texas[Goal]   OPERATE–VEHICLE

- Alice[Destroyer] destroys the item[Undergoer] with a plane[Instrument]

  DESTROYING

- Finding the question's head verb

- Example:

  – „Who$_{[Buyer]}$ purchased YouTube$_{[Goods]}$?"

- Commerce_buy

Buyer [Byr]
The Buyer wants the Goods and offers Money to a Seller in exchange for them.
Jess BOUGHT a coat.

Lee BOUGHT a textbook from Abby.

Goods [Gds]
The FE Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.
Only one winner PURCHASED the paintings

(https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Commerce_buy)

Question Analysis: Semantic Role Labeling

- COMMERCE−BUY

  – buy (v), buyer (n), purchase (n), purchase (v), purchaser (n)

- Identifying candidate passages

  – Buyer [Subj,NP] verb Goods [Obj,NP]

  – Buyer [Subj,NP] verb Goods [Obj,NP] Seller [Dep,PP-from]

  – Goods [Subj,NP] verb Buyer [Dep,PP-by]

  – …

- Example:

  – „In 2006, YouTube$_{[Goods]}$ was purchased by Google$_{[Buyer]}$ for $1.65 billion."

Architecture

Natural Language Processing

Question Analysis → Question Classification → Query Construction

How many presidents do the USA have?

Corpus ← Document Retrieval

Sentence Retrieval

Web

Sentence Annotation

44 ← Answer Validation ← Answer Extraction

Information Retrieval

Information Extraction
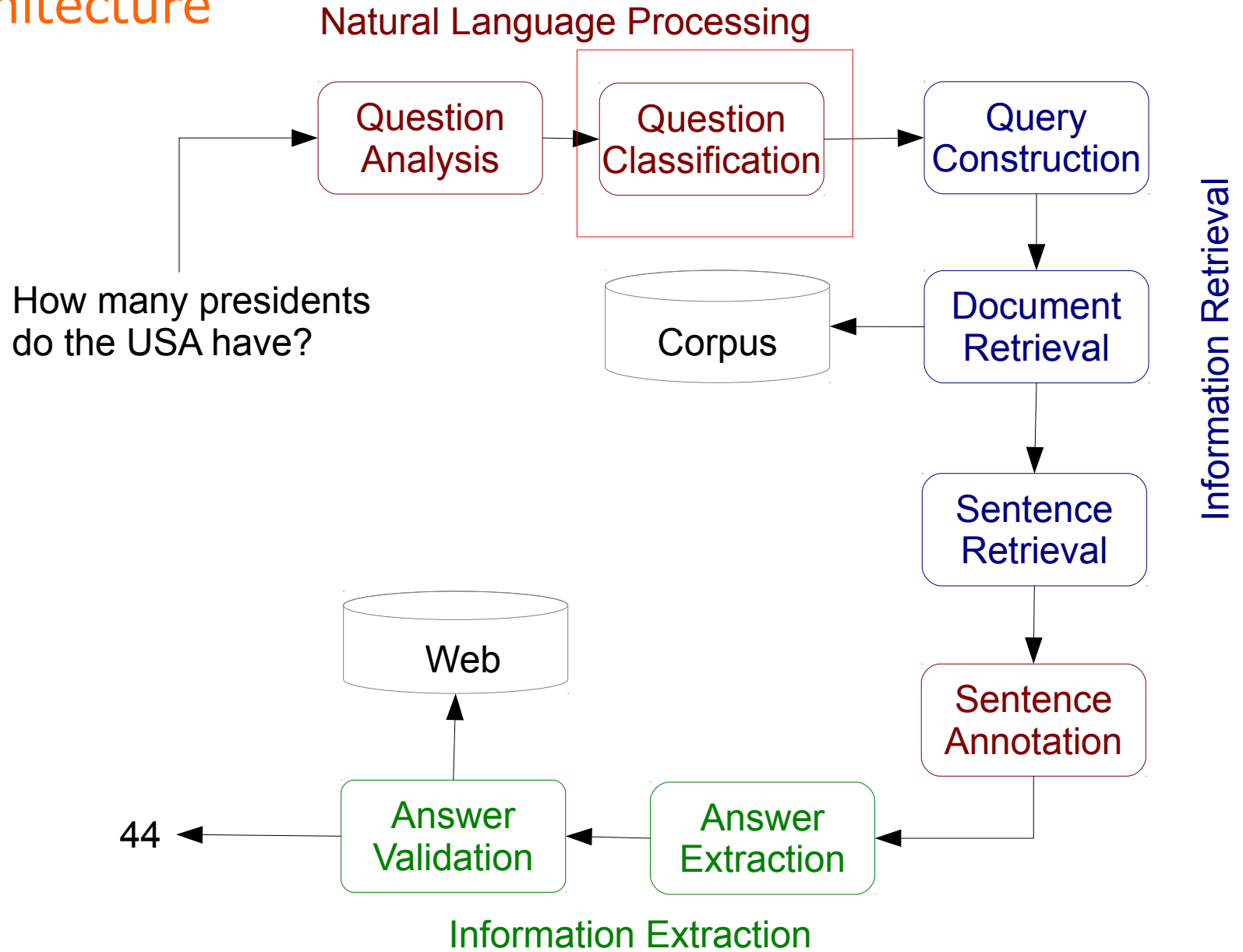
Question Classification

- Classifying the input question into a set of question types

- Mapping question types to the available named entity labels

- Finding strings that have the same type as the input question in the answer extraction step

- Example:

  – Question: "In what country was Albert Einstein born?"

  – Type: LOCATION - Country

Question Classification

- Example (NER):

  - S1: " Albert Einstein was born in 14 March 1879 ."

  - S2: " Albert Einstein was born in Germany ."

  - S3: " Albert Einstein was born in a Jewish family."

Question Classification

- Classification taxonomies

  - BBN

  - Pasca & Harabagiu

  - Li & Roth

    - 6 coarse- and 50 fine-grained classes

      - ABBREVIATION
      - ENTITY
      - DESCRIPTION
      - HUMAN
      - LOCATION
      - NUMERIC

Question Classification

HUMAN	human beings
group	a group or organization of persons
ind	an individual
title	title of a person
description	description of a person
LOCATION	locations
city	cities
country	countries
mountain	mountains
other	other locations
state	states

(http://cogcomp.cs.illinois.edu/Data/QA/QC/definition.html)

34

Question Classification

How did serfdom develop in and then leave Russia ?	DESC	manner
What films featured the character Popeye Doyle ?	ENTY	cremat
What fowl grabs the spotlight after the Chinese Year of the Monkey ?	ENTY	animal
What is the full form of .com ?	ABBR	exp
What team did baseball 's St. Louis Browns become ?	HUM	gr

(http://cogcomp.cs.illinois.edu/Data/QA/QC/train_1000.label)

Question Classification
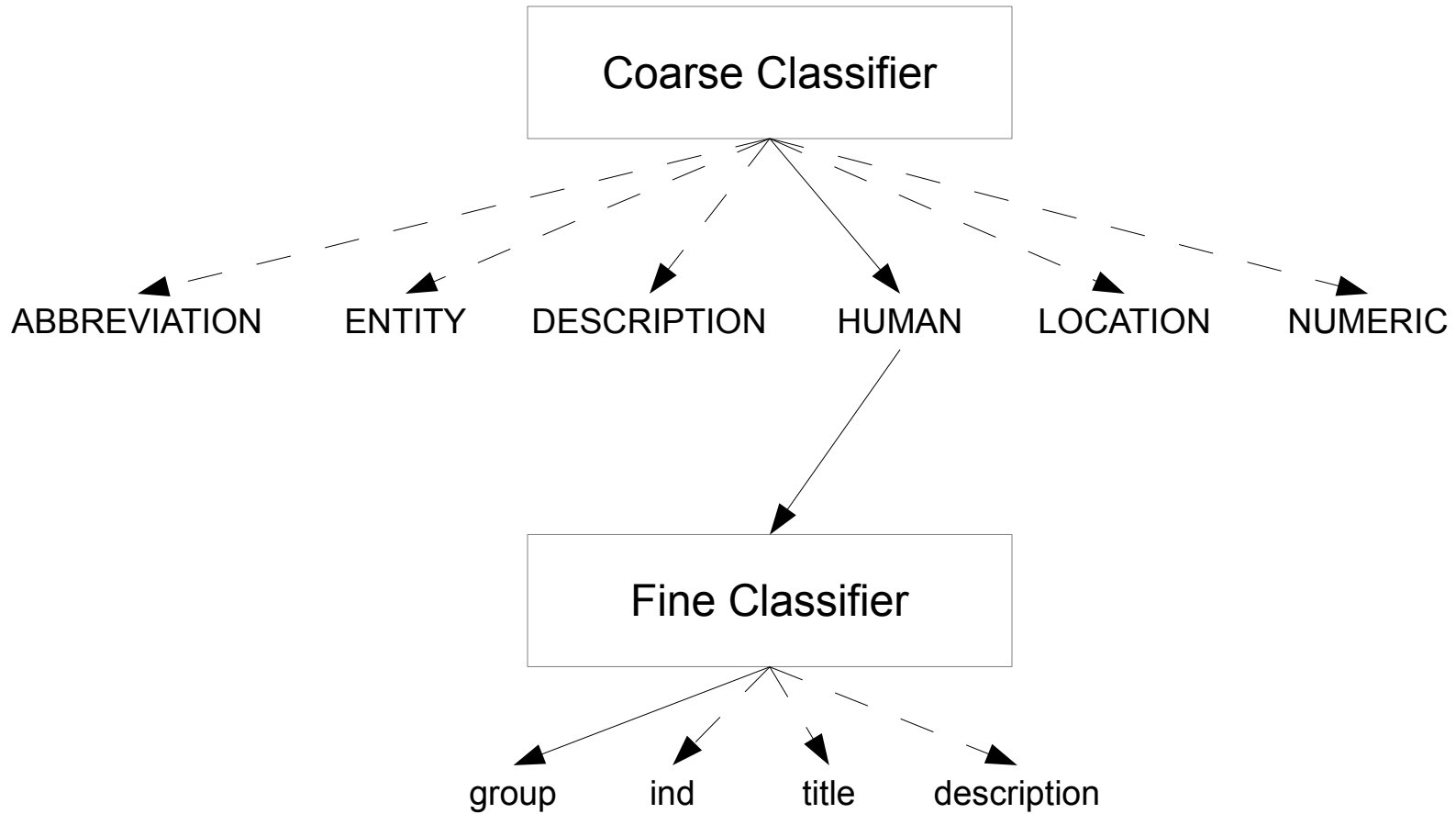
Question Classification

- Biomedical domain
    - UMLS semantic groups and types

T038-Phenomena-Biologic Function	What is the main \<mention>role\<mention> of Ctf4 in dna replication?
T116-Chemicals & Drugs-Amino Acid, Peptide, or Protein	List the \<mention>components\<mention> of a Replisome Progression Complex (RPC).
T067-Phenomena-Phenomenon or Process	What is the \<mention>effect\<mention> of CRD-BP on the stability of c-myc mRNA?
T116-Chemicals & Drugs-Amino Acid, Peptide, or Protein	List GATA-1 \<mention>interacting partners\<mention> as discovered with the help of the biotinylation
T007-Living Beings-Bacterium	Which types of bacterial \<mention>microflora\<mention> are associated with the progression of peri-in
T008-Living Beings-Animal	List \<mention>invertebrates\<mention> where ultraconserved elements have been identified.
T086-Genes & Molecular Sequences-Nucleotide Sequence	How are \<mention>ultraconserved elements\<mention> called when they form clusters?
T047-Disorders-Disease or Syndrome	What is the \<mention>effect\<mention> of the absence of Saccharomyces cerevisiae Rrm3p?
T073-Objects-Manufactured Object	List available \<mention>databases\<mention> containing information about conserved noncoding elen
T067-Phenomena-Phenomenon or Process	What types of DNA \<mention>mutations\<mention> are induced by 2-hydroxy-dATP (2-OH-dATP)?
T081-Concepts & Ideas-Quantitative Concept	What is the typical outer \<mention>diameter\<mention> of microtubules (tubulin heterodimers)?
T067-Phenomena-Phenomenon or Process	In which types of \<mention>DNA repair\<mention> is the UvrAB complex involved?

Question Classification

- Using any kinds of supervised classifiers

    – K Nearest Neighbor

    – Support Vector Machines

    – Naïve Bayes

    – Maximum Entropy

    – Logistic Regression

    – …

- Considering the confidence measure of the classification to filter the result

Architecture

Natural Language Processing

Information Retrieval

Information Extraction

```
[Question Analysis] → [Question Classification] → [Query Construction]
                                                          ↓
                              [Corpus] ← [Document Retrieval]
                                                          ↓
                                               [Sentence Retrieval]
                                                          ↓
                                               [Sentence Annotation]
                                                          ↓
[Web] ← [Answer Validation] ← [Answer Extraction]
44 ← [Answer Validation]
```

How many presidents
do the USA have?

44
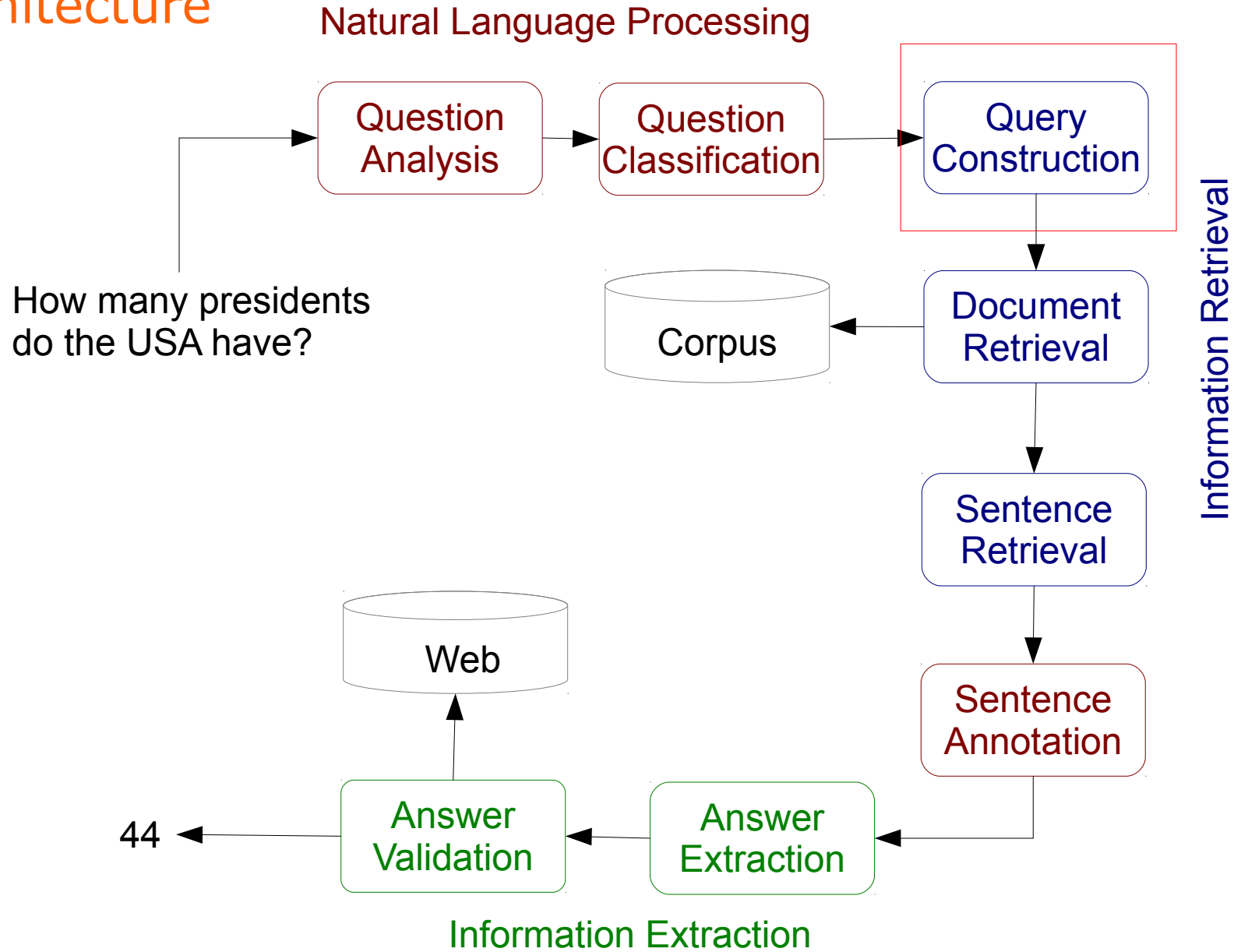
Query Construction

- Goal:

  – Formulating a query with a high chance of retrieving relevant documents

- Task:

  – Assigning a higher weight to the question's target

  – Using query expansion techniques to expand the query

Architecture

Question Analysis → Question Classification → Query Construction

How many presidents do the USA have?

Corpus ← Document Retrieval

Information Retrieval

Sentence Retrieval

Web

Sentence Annotation

44 ← Answer Validation ← Answer Extraction

Information Extraction

41
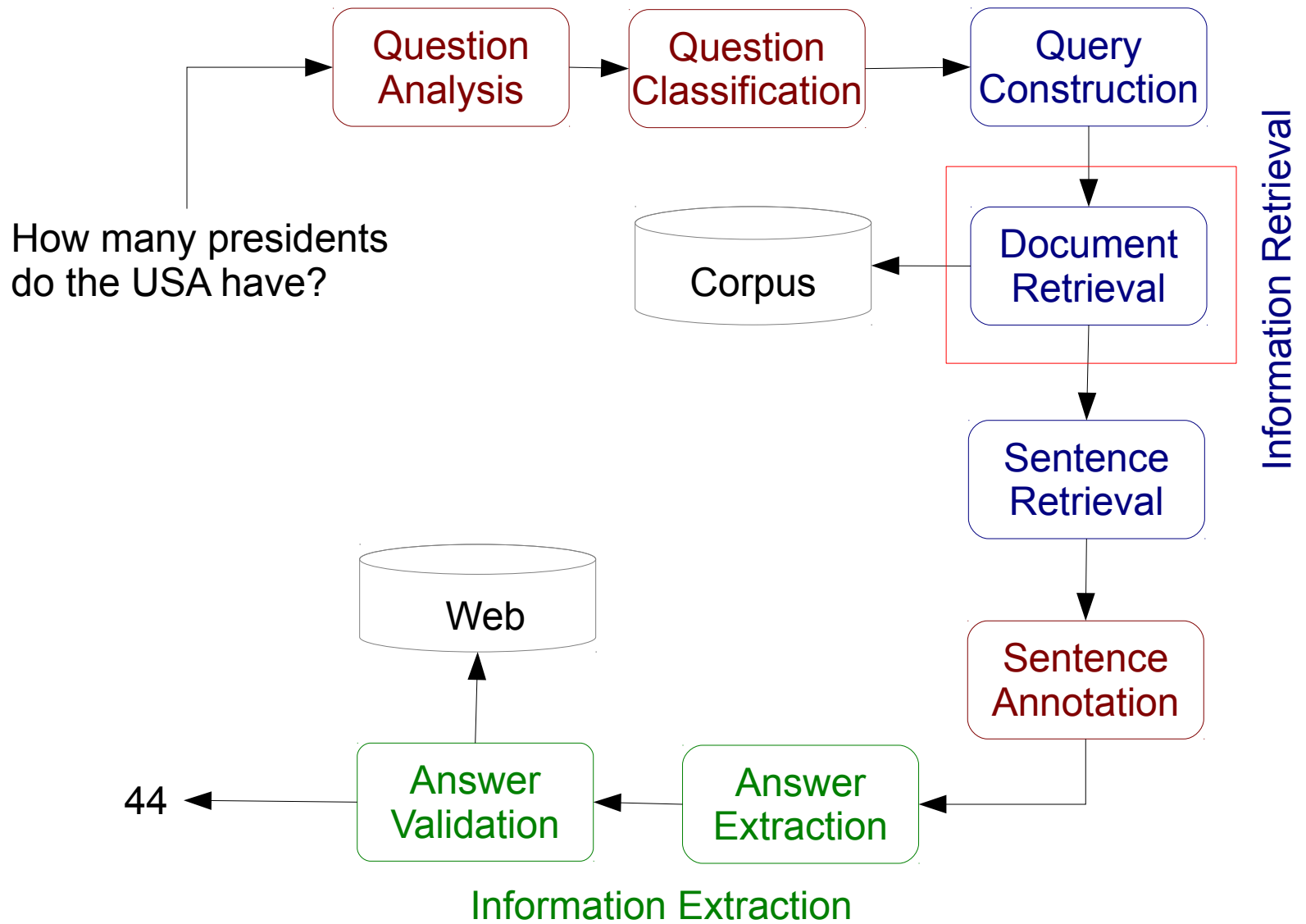
Document Retrieval

- Importance:

  - QA components use computationally intensive algorithms

  - Time complexity of the system strongly depends on the size of the to be processed corpus

- Task:

  - Reducing the search space for the subsequent components

  - Retrieving relevant documents from a large corpus

  - Selecting top n retrieved document for the next steps

Document Retrieval

- Using available information retrieval models

  – Vector Space Model

  – Probabilistic Model

  – Language Model

- Using available information retrieval toolkits: Lucene, Lemur, SAP HANA, etc.

Architecture

Natural Language Processing

Information Retrieval

Information Extraction

```
How many presidents
do the USA have?
```

Question Analysis → Question Classification → Query Construction → Document Retrieval → Corpus

Document Retrieval → Sentence Retrieval → Sentence Annotation → Answer Extraction → Answer Validation → Web

Answer Validation → 44
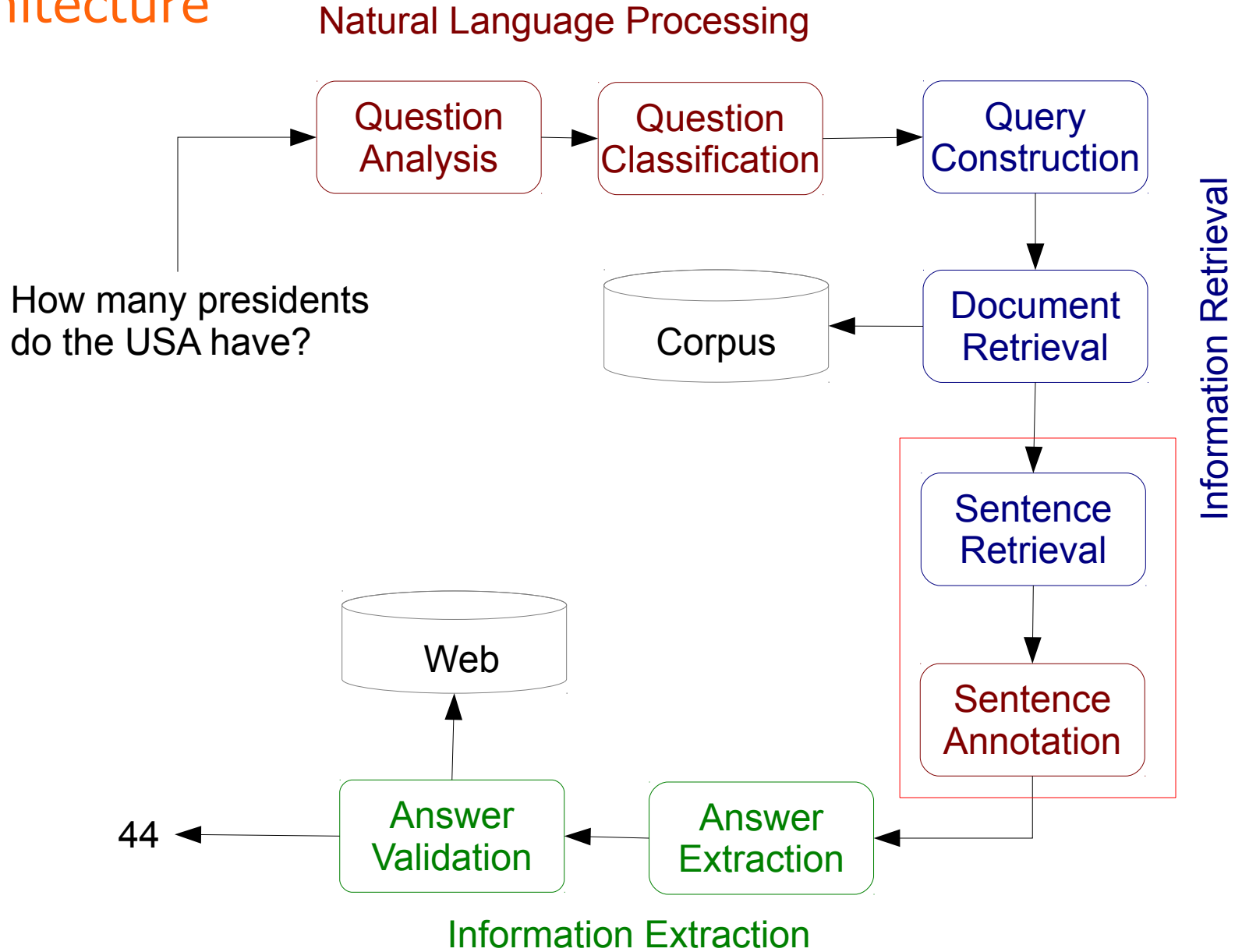
Sentence Retrieval

- Task:

  - Finding small segments of text that contain the answer

- Benefits beyond document retrieval:

  - Documents can be very large

  - Documents span different subject areas

  - The relevant information is expressed locally

  - Retrieving sentences simplifies the answer extraction step

- Main problem: sentence brevity

Sentence Retrieval

- Information retrieval models for sentence retrieval

  - Vector Space Model

  - Probabilistic Model

  - Language Model

    - Jelinek-Mercer Linear Interpolation

    - Bayesian Smoothing with Dirichlet Prior

    - Absolute Discounting

Sentence Retrieval

- Support from named-entity recognition and semantic role labeling

  - " Albert Einstein was born in 14 March 1879 ."

  - " Albert Einstein was born in Germany ."

  - " Albert Einstein was born in a Jewish family."

  - Buyer [Subj,NP] verb Goods [Obj,NP]

  - Buyer [Subj,NP] verb Goods [Obj,NP] Seller [Dep,PP-from]

  - Goods [Subj,NP] verb Buyer [Dep,PP-by]

Sentence Retrieval

- Supervised learning

    - Features

        - Number of entity types of the right type

        - Number of question keywords

        - Longest exact sequence of question keywords

        - Rank of the document

        - Proximity of the keywords (shortest path)

        - N-gram overlap between passage and question

Sentence Annotation

- Similar to Question Analysis

- Annotating relevant sentences using linguistic analysis

  - Named entity recognition

  - Syntactic parsing

  - Semantic role labeling

Architecture

Natural Language Processing

Information Retrieval

Information Extraction

How many presidents do the USA have?

Question Analysis → Question Classification → Query Construction → Document Retrieval → Corpus

Document Retrieval → Sentence Retrieval → Sentence Annotation → Answer Extraction → Answer Validation → 44

Answer Validation → Web
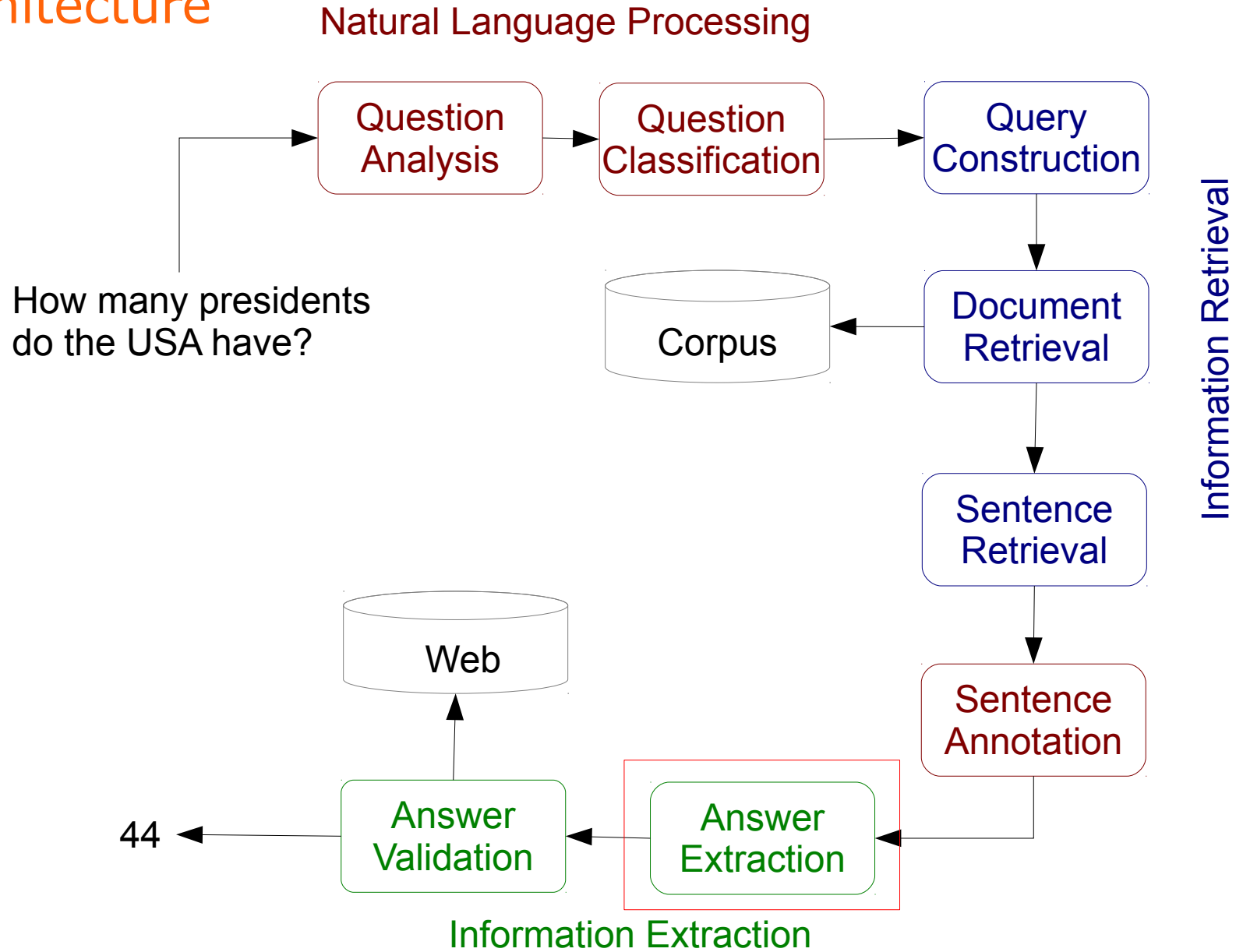
Answer Extraction

- Extracting candidate answers based on various information

  - Question

    - Question Analysis: patterns
    - Question Analysis: syntactic parse
    - Question Analysis: semantic roles
    - Question Classification: question type

  - Sentence

    - Sentence Annotation: all annotated data

Answer Extraction

- Using extracted patterns

- Example:

  - Question: "In what country was Albert Einstein born?"

- Question Pattern: In what country was X born?

- Answer Pattern: X was born in Y.

- Example (Pattern):

  - S1: "Albert Einstein was born in 14 March 1879."
  - S2: "Albert Einstein was born in Germany."
  - S3: "Albert Einstein was born in a Jewish family."

Answer Extraction

- Using question type and entity type

- Example:

  - Question: "In what country was Albert Einstein born?"
  - Question Type: LOCATION - Country

- Example (NER):

  - S1: "Albert Einstein was born in 14 March 1879."
  - S2: "Albert Einstein was born in Germany."
  - S3: "Albert Einstein was born in a Jewish family."

Answer Extraction

- Using syntactic parsing

    - Different wordings possible, but similar syntactic structure

Q: Who founded the Black Panthers organization ?
    subj   obj

S1: Bobby Seale, a student at Merritt College, founded the Black Panther Party for self-defense .
    subj   obj

S2: The Black Panther Party, co-founded by Seale and Newton, flourished...
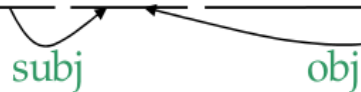   obj   subj

(based on the work by Dan Shen)

Answer Extraction

- Using syntactic parsing

  - Many syntactic variations → need robust matching approach
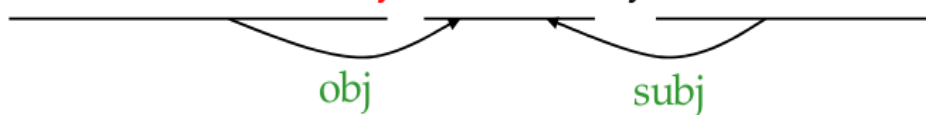
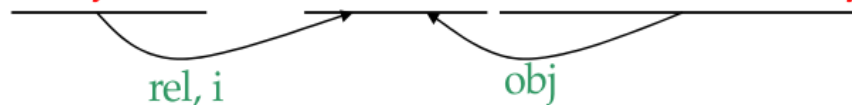

Q: Who founded the Black Panthers organization ?

S1: Bobby Seale, a student at Merritt College, founded the Black Panther Party for self-defense .

S2: The Black Panther Party, co-founded by Seale and Newton, flourished...

S3: Hilliard introduced Bobby Seale, who co-founded the Black Panther Party here .

S4: Black Panthers Co-founder Bobby Seale visits UMM.
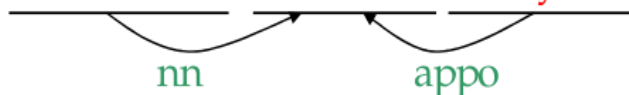
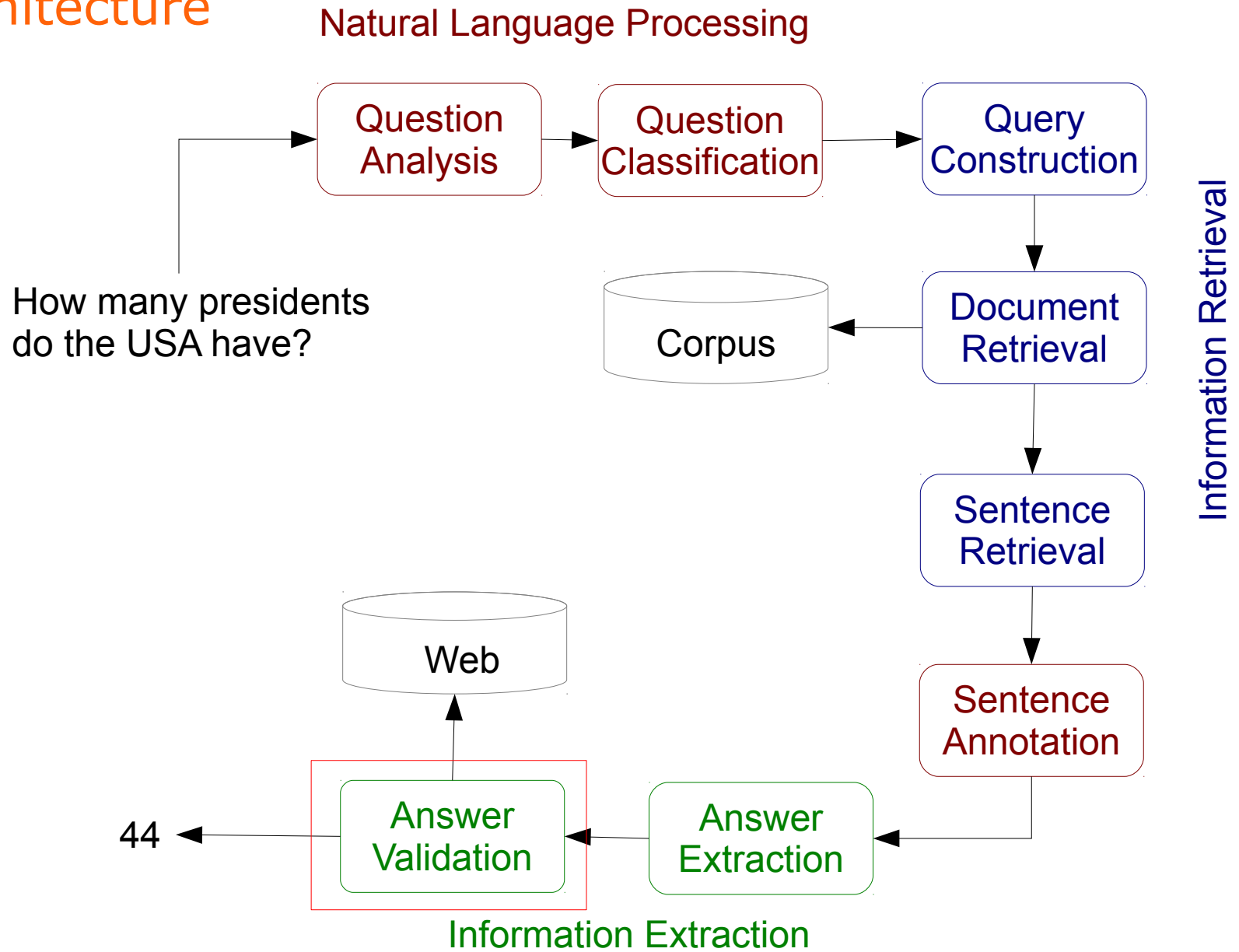(based on the work by Dan Shen)

Answer Extraction

- Using semantic roles

- Example:

  - "Who$_{[Buyer]}$ purchased YouTube?$_{[Goods]}$"

- Example:

  - " In 2006, YouTube$_{[Goods]}$ was purchased by Google$_{[Buyer]}$ for $1.65 billion."

Architecture

Natural Language Processing

Question Analysis → Question Classification → Query Construction

How many presidents do the USA have?

Information Retrieval

Query Construction → Document Retrieval

Corpus ← Document Retrieval

Document Retrieval → Sentence Retrieval

Sentence Retrieval → Sentence Annotation

Web

Sentence Annotation → Answer Extraction

Answer Validation ← Answer Extraction

44 ← Answer Validation

Answer Validation → Web

Information Extraction

Answer Validation

- Using Web as a knowledge resource for validating answers

- Required steps
  - Query creation
  - Answer rating

Answer Validation

- Query creation

    - Combining the answer with a subset of the question keywords

    - Choosing different combinations of subsets

        - Bag-of-Word

        - Noun phrase chunks

        - Declarative form

Answer Validation

- Example:

  – Question: "In what country was Albert Einstein born?"

  – Answer Candidate: Germany

- Queries:

  – Bag-of-Word:

    - Albert Einstein born Germany

  – Noun-Phrase-Chunks:

    - "Albert Einstein" born Germany

  – Declarative-Form:

    - "Albert Einstein born Germany"

Answer Validation

- Answer rating

  - Passing the query to a search engine

  - Analyzing the result of the search engine

    - Counting the results
    - Parsing the result snippets

  - Other possibilities:

    - Using knowledge bases to find relations between the question keywords and the answer

Further Reading

- Speech and Language Processing

  - Chapters 23.1, 23.2