Question Answering

Vanessa Murdock

Overview

- Question Answering, in general
- Question Classification
- Information Retrieval
- External Resources
- Answer Extraction
- QA Systems from TREC
- Real-Life QA Systems

Question Answering

- Question answering seeks the token or phrase (or passage, document, document set) that is the exact answer to a question
- Questions have many flavors
- Most research is focused on fact questions
- Answers are assumed to be tokens or phrases
- Complete answers are assumed to be found in a single source

Types of Questions

Fact : Who killed Martin Luther King?

Task : How do I apply for a passport?

Opinion : What was the best movie this year?

Definition : Who is Jane Goodall?

List : What movies was Jude Law in?

Explanation : What was the cause of the Korean war?

Yes-No : Is it legal to turn right on red in Iowa?

Question Examples

Aspartame is also known as what?
At what age did Rossini stop writing operas?
Boxing Day is celebrated on what day?
Define Thalassemia
How big is our galaxy, in diameter?
How cold should a refridgerator be?
CPR is the abbreviation for what?
How long is human gestation?
How long is the Columbia River?

Fact Question Examples

Q: When was Mozart born?

A: 1756

Q: What is a nanometer?

A: a billionth of a meter

A: a millionth of a millimeter

Q: When was The Great Depression?

A: 1930's

A: 1931

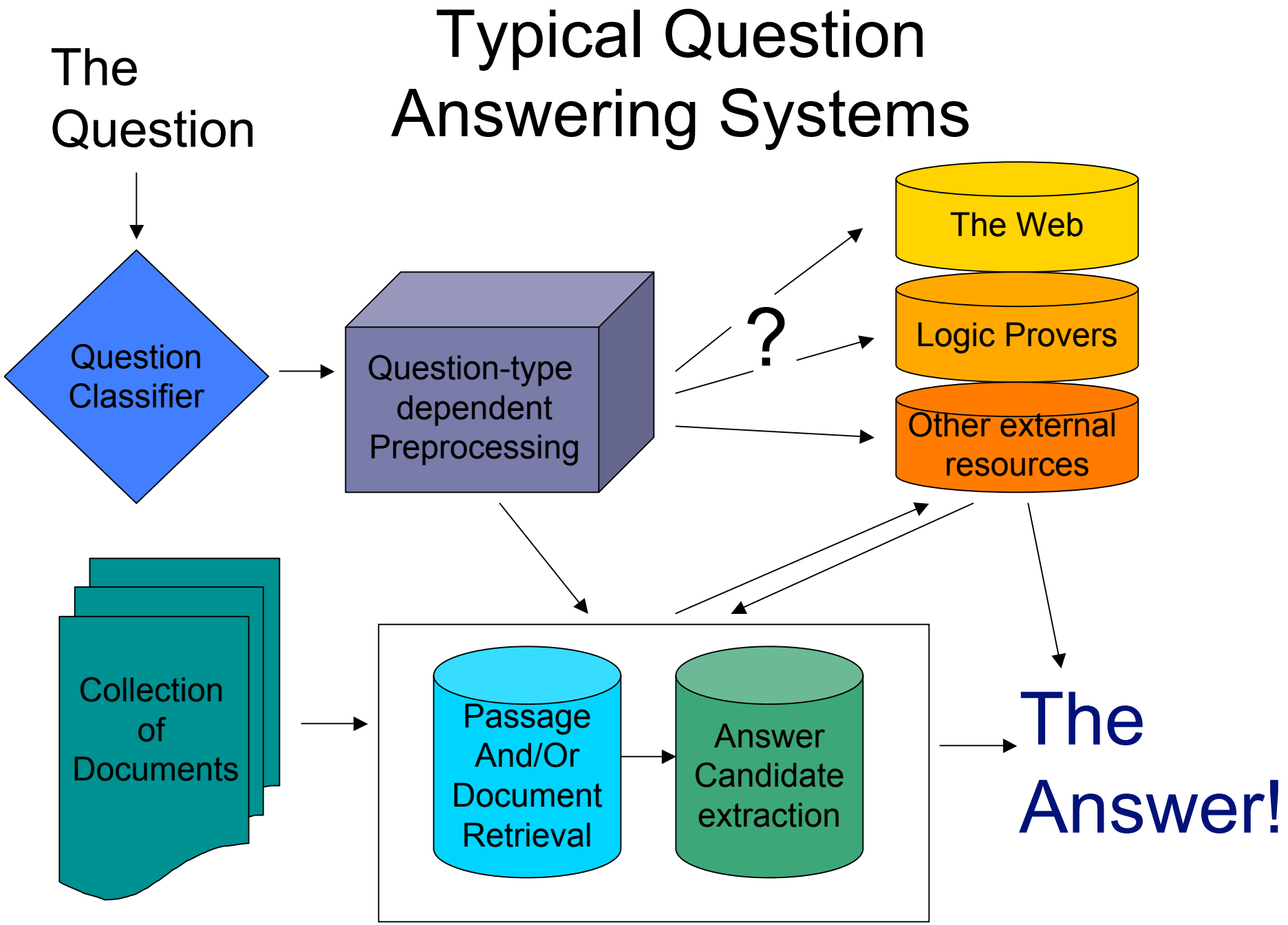
A: 1932

Q: Who is Absalom?

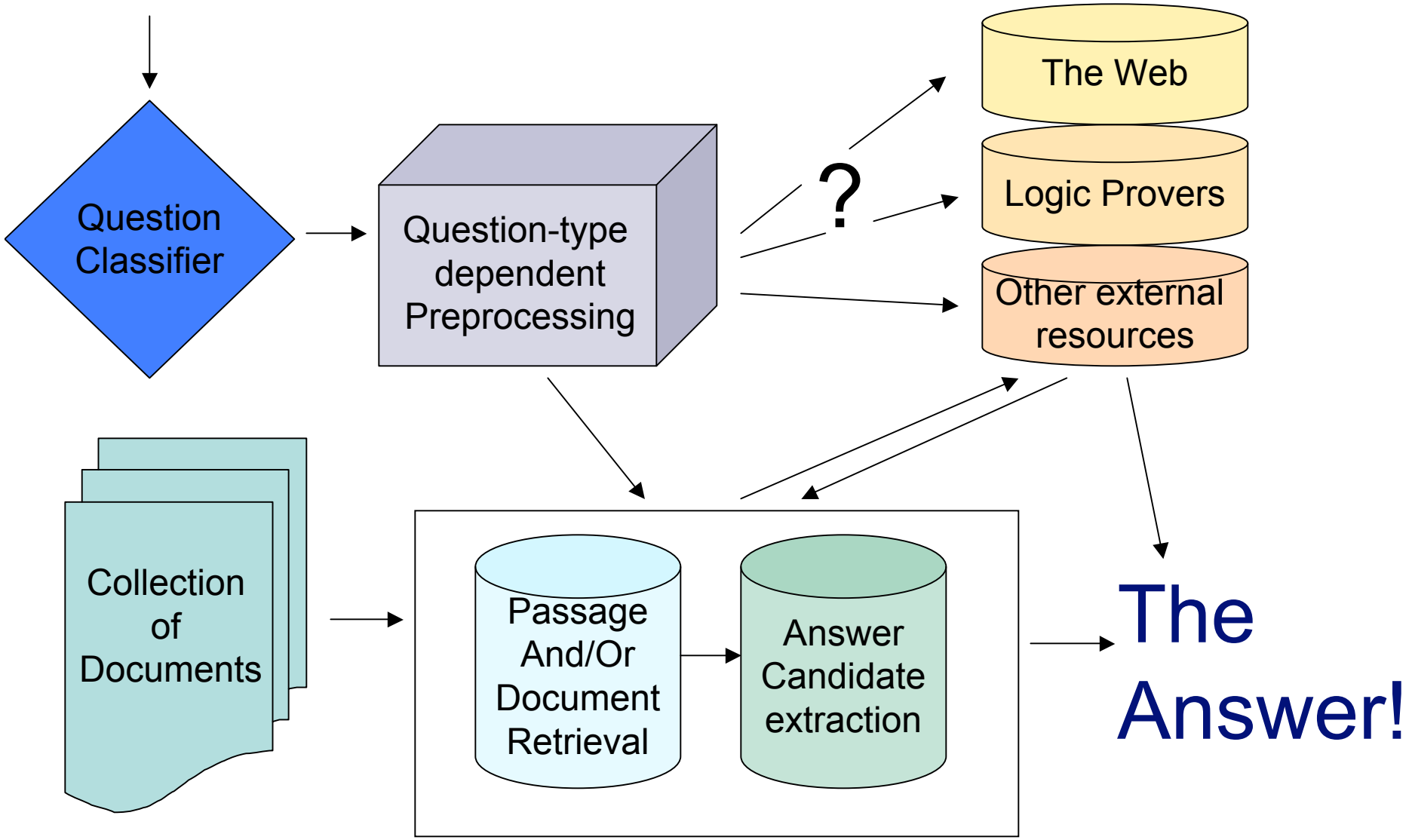
A: African-American leader, first black whaling ship captain, desegregated Nantucket's school system.

A: Son of (biblical) David, who betrayed his father

Typical Question Answering Systems



Typical Question Answering Systems



Fact Question Classification
classify by expected answer type

- Basically two approaches:
  - Classification
    - Advantage: easy to understand/implement, reliable
    - Disadvantage: Doesn't give information other than class
  - Regular expressions
    - Advantage: can give information in addition to class
    - Disadvantage: very brittle
- Classifier Features: POS tags, words, NE tags, WordNet, wh-words, parse trees etc.
- Regular expressions:
  - Simple: wh-words
  - Complex: QA "typology"
- State of the Art: ~90% accuracy

Question Classification: SVMs

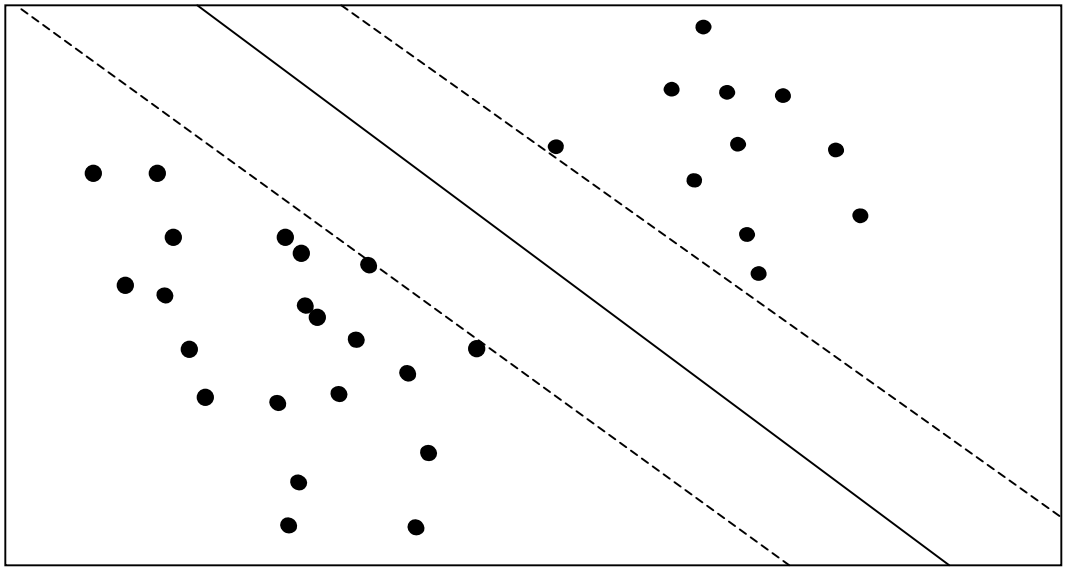
(Zhang & Lee, SIGIR 2003) (tree kernel)

- Defined coarse and fine-grained classes
- binary features: word identity, word n-grams

(Metzler and Croft, *Information Retrieval*, 2005) (linear kernel)

- Fine-grained classes
- Word identities, WordNet synonyms, POS tags

minimize $\langle w \bullet w \rangle + C \sum_i \xi_i$ s.t. $y_i \left(\langle w \bullet x \rangle + b \right) \geq 1 - \xi_i$ for all *i*



Question Classification: regex

- By wh-words, and regular expressions:
  - "Who" => person (Organization? GPE?)
  - "When" => date (Year? Season? Holiday?)
  - "Where" => location (GPE? Organization?)
  - "How"
    - "How many" => cardinal number
    - "How do" => task question
- "Question typology" extensive regex's from patterns

ISI's question typology

Semantic ontology types (I-EN-CITY) **and part of speech labels** (S-PROPER-NAME):

What is the capital of Uganda?

QTARGET: (((I-EN-CITY S-PROPER-NAME)) ((EQ I-EN-PROPER-PLACE)))

Parse tree roles:

Why can't ostriches fly?	QTARGET: (((ROLE REASON)))
Name a film in which Jude Law acted.	QTARGET: (((SLOT TITLE-P TRUE)))

QA Typology nodes:

What are the Black Hills known for?	Q-WHY-FAMOUS
Who was Whitcomb Judson?	Q-WHY-FAMOUS-PERSON
What is Occam's Razor?	Q-DEFINITION

Qargs for additional information:

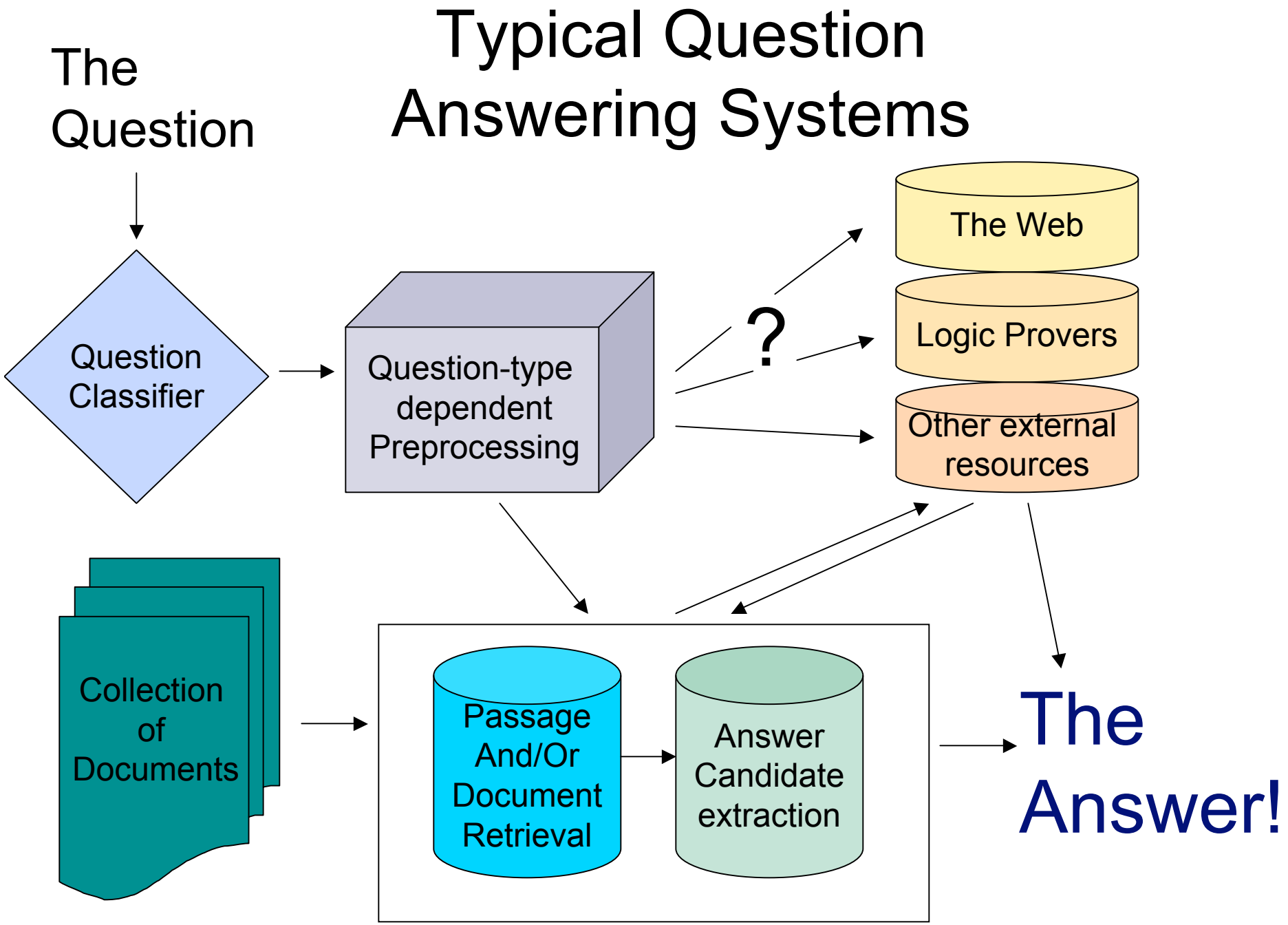
Who was Betsy Ross? QTARGET: (((Q-WHY-FAMOUS-PERSON))) QARGS: (("Betsy Ross"))

How is "Pacific Bell" abbreviated? QTARGET: (((Q-ABBREVIATION))) QARGS: (("Pacific Bell"))

What are geckos? QTARGET: (((Q-DEFINITION))) QARGS: (("geckos" "gecko") ("animal"))

Figure 4. QA-related information, returned in the parse tree of the question.

Typical Question Answering Systems



IR for QA: models

- If we did a good job of IR, QA would be easy
- Passage/sentence retrieval is not just short document retrieval
- Vector Space model
- Query Likelihood
- Translation Models

Vector Space Model

- Represent documents as vectors of term weights
- Term weights are given by tf.idf

$$weight(i,j) = \begin{cases} (1+\log(tf_{i,j}))\log\frac{N}{df_i} & tf_{i,j} \geq 1 \\ 0 & tf_{i,j} = 0 \end{cases}$$

- Rank documents by their cosine of the angle between the query vector and the document vector.

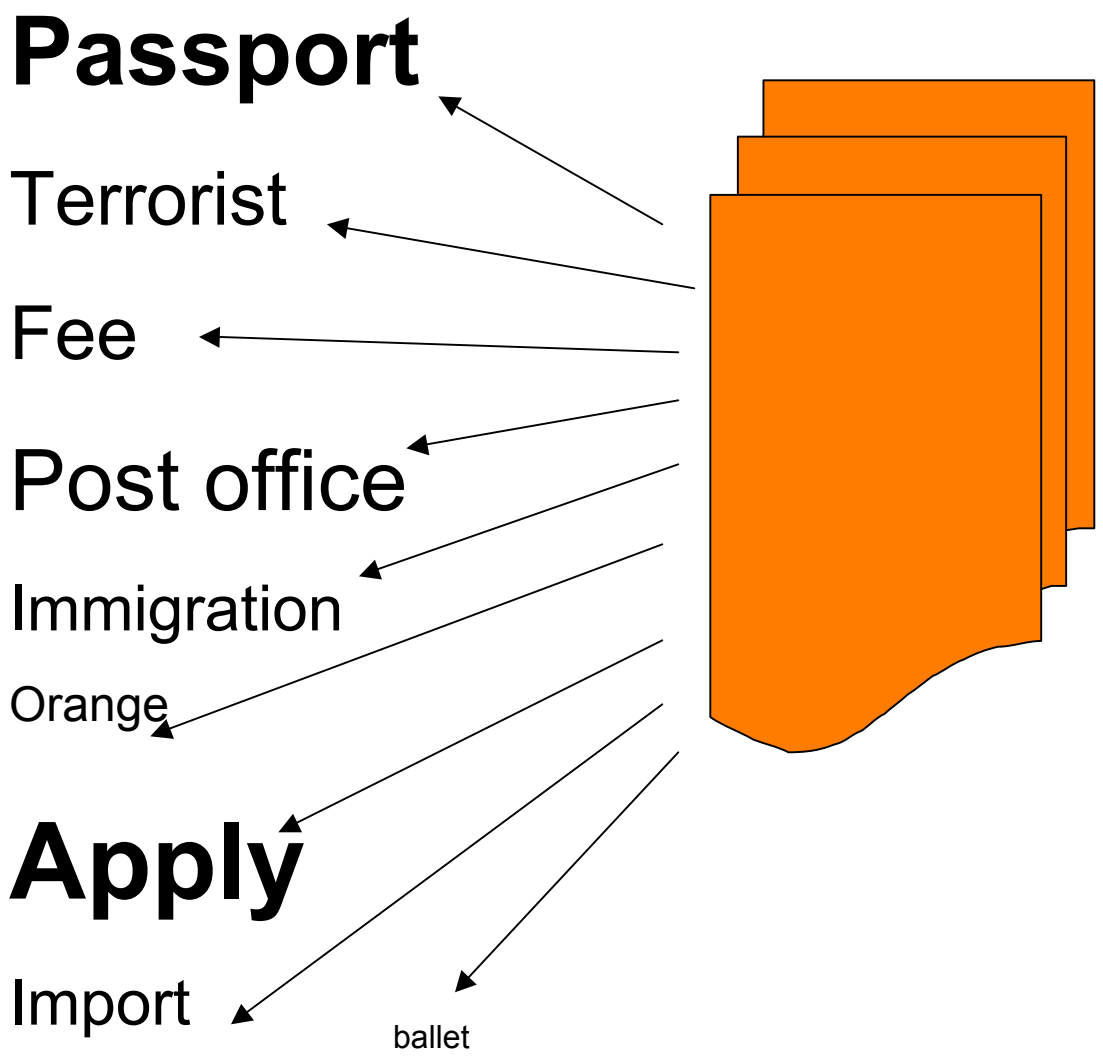
$$\cos(\vec{q},\vec{d}) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2}\sqrt{d_i^2}}$$

Documents are represented by vectors of term weights. Why not other types of features? Other similarity metrics?

Query Likelihood, part 1

Documents can be considered bags of words. We sample words from the document.

Certain words are more frequent than others, thus have higher probability.



Query Likelihood, part 2

How high is Mt. Everest?

Mt Everest is 27,000 ft. high.

How high is Mt. Everest?

Everest's elevation is 27,000 ft.

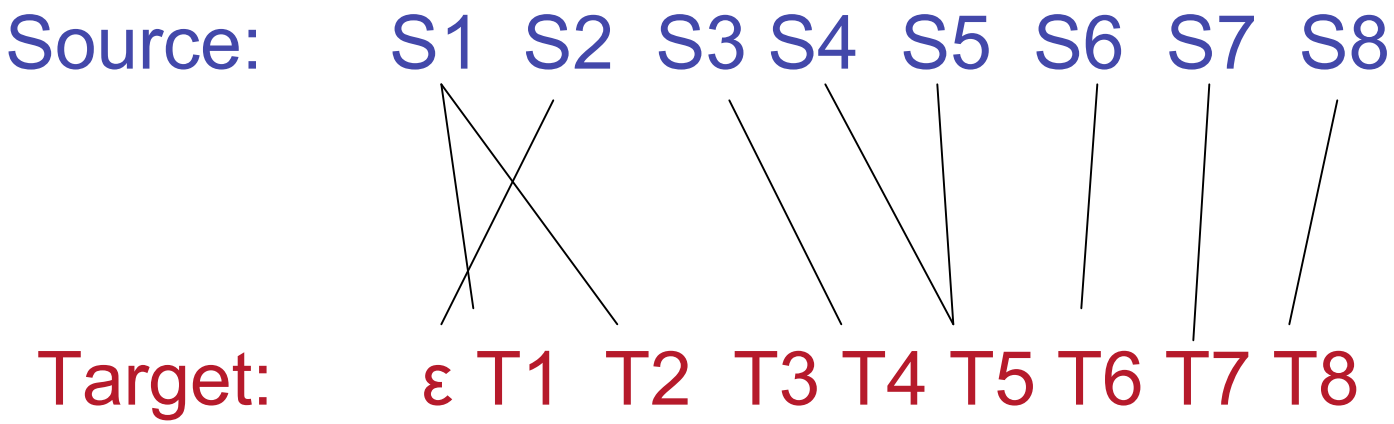
Smoothing constant

$$P(Q \mid D) = \prod_i \lambda P(q_i \mid D) + (1-\lambda) P(q_i \mid GE)$$

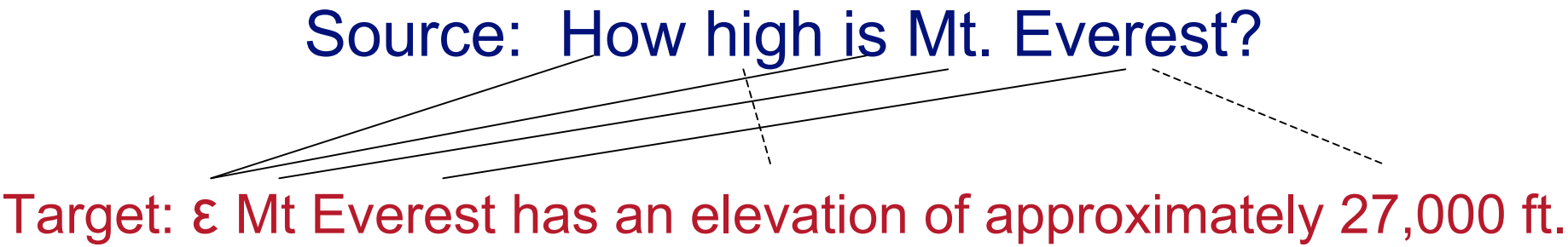
Word probability in the document

Word probability in General English

Translation Models, normally



Translation Models, QA



Translation Models train on a parallel corpus, in our case a set of questions and sentences containing answers.

Translation Models and Query Likelihood

Translation Model:

$$P(Q \mid A) = \prod_i \lambda \sum_j P(q_i \mid a_j) P(a_j \mid A) + (1 - \lambda) P(q_i \mid GE)$$

If there is only one translation

If $qi = aj$ and $p(q_i|a_j) = 1.0$

Query Likelihood:

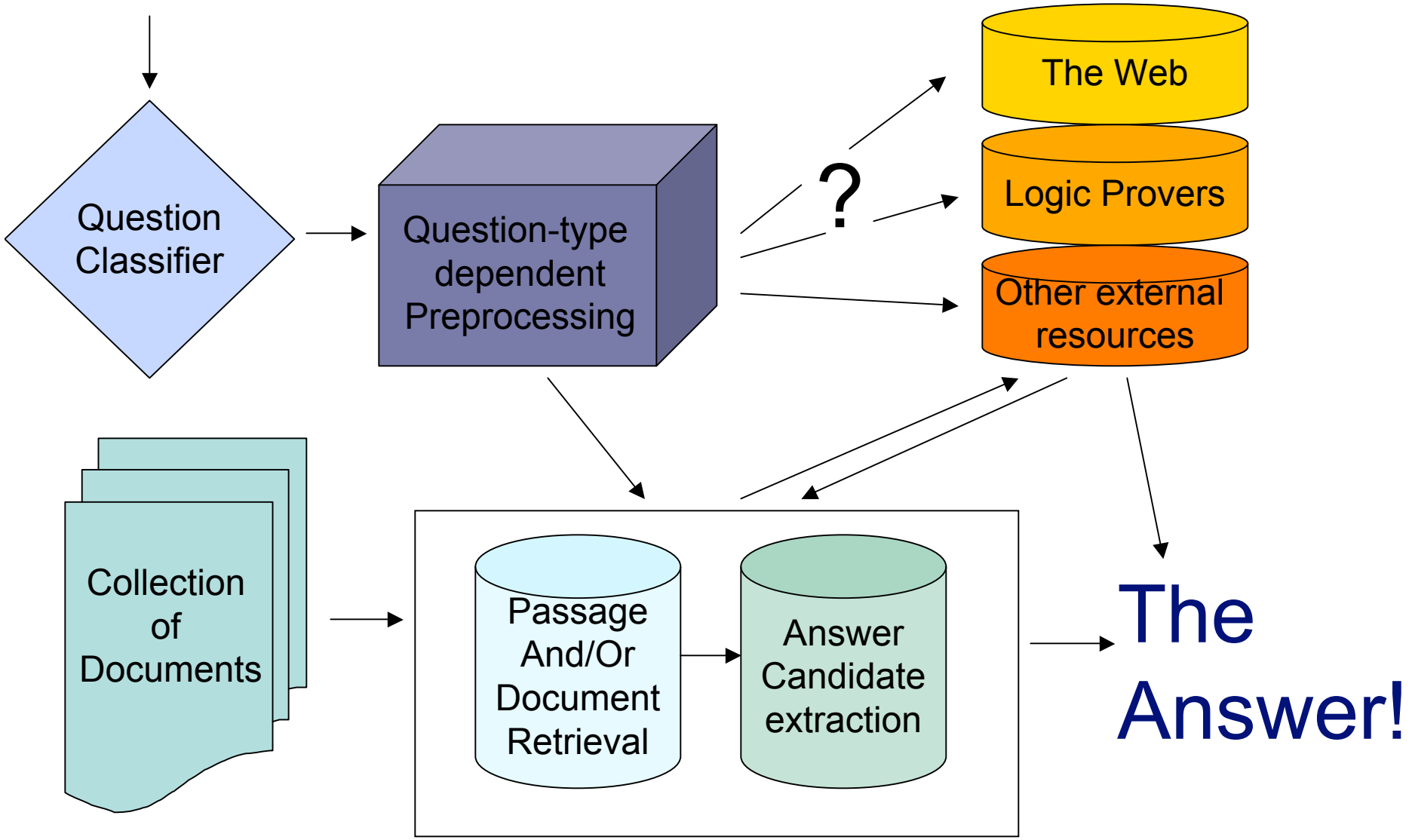
$$P(Q \mid A) = \prod_i \lambda P(q_i \mid A) + (1 - \lambda) P(q_i \mid GE)$$

State of the Art in Passage Retrieval (TREC 2003)

Group	Accuracy
Language Computer Corp.	.685
Nat'l. Univ. of Singapore	.419
Univ. of Waterloo	.351
Univ. of Massachusetts	.201
Macquarie Univ.	.191
Saarland Univ.	.169
IIT Bombay	.133
CL Research	.119
Univ. of Amsterdam	.111
Queens College, CUNY	.097

The Question

The Devil is in the Details…



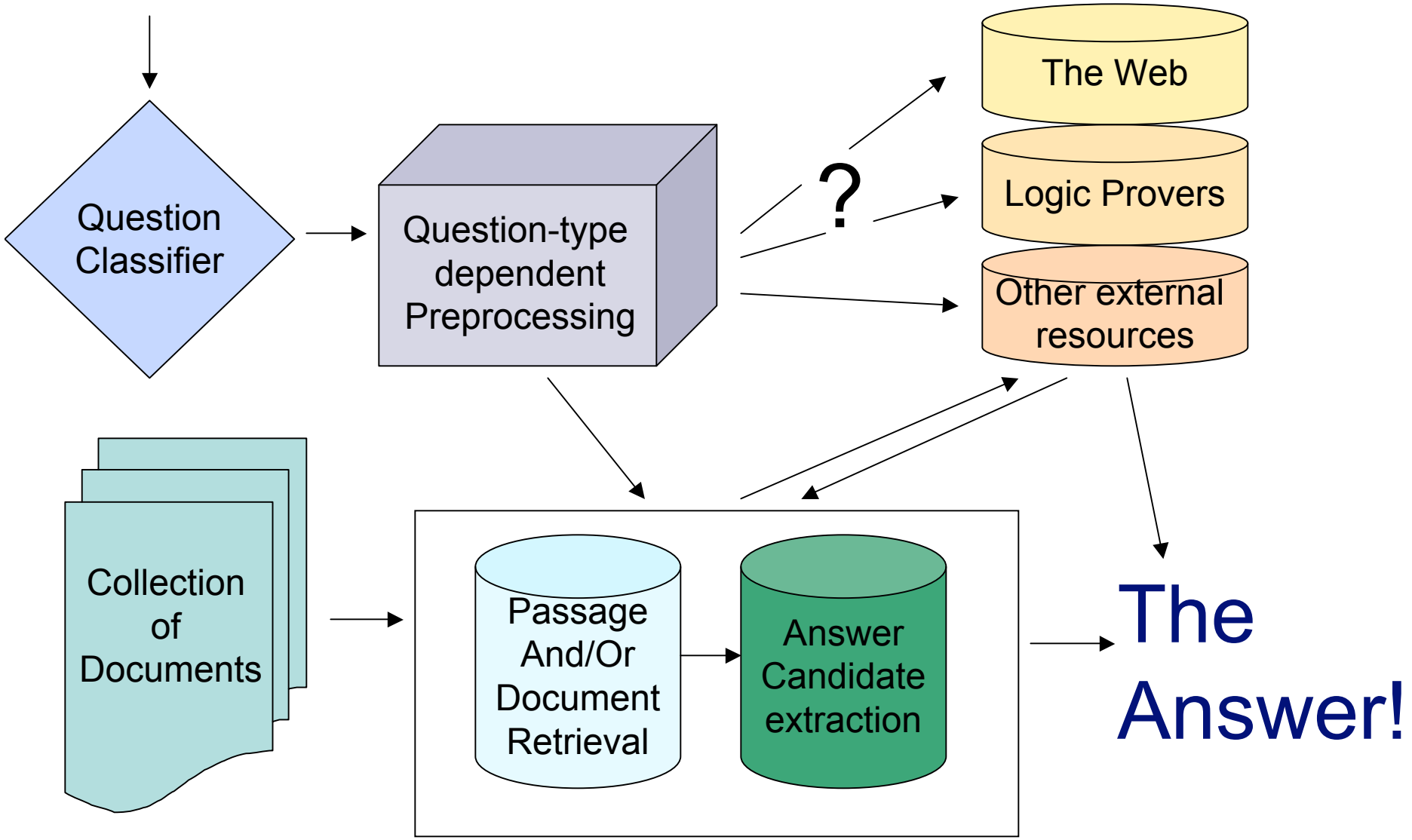
Question Type Dependent Processing

- Question rewrites for the web
  - Turn the question into a query, combine multiple evidence
- Logic provers
  - Attempt to reason about the question
- Answer filtering
  - Rule out answers that look right, but can't be
- Question analysis for patterns
  - Patterns in the question suggest patterns in the answer

External Resources

- The web (problematic)
  - Web summaries
  - Answer validation
  - Increasing training data
- POS taggers, NE extractors, noun-phrase chunkers
- Gazetteers, Ontologies, thesauri
  - WordNet, ConceptNet
- Logic Provers
- Previously answered questions

Typical Question Answering Systems



Answer Extraction (Simplest)

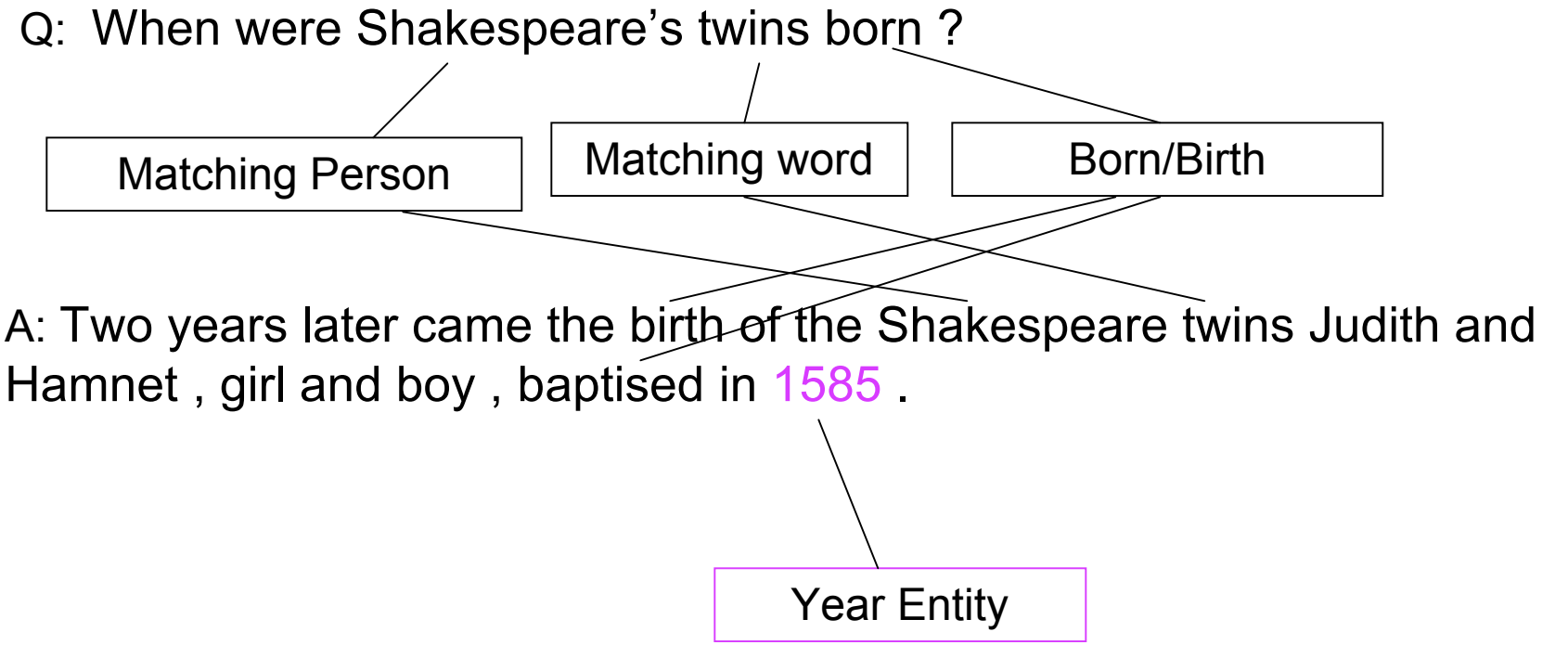
- Extract the answer token that is the correct named entity type from top sentence.
- Extract the answer tokens from top N sentences, and vote.
- Extract answer tokens candidates from top N sentences, validate on the Web

Answer Tagging

- Treat answer tagging as named-entity tagging
- Answers are frequently not a named entity type (ex. why-famous questions)
- Answer tokens are not predictable and do not always have predictable indicators
- Features of answer tokens are not directly sequential and are often long-range
- Features of one question type may not generalize to other question types

Answer Tagging (Easy)

- Determine the answer type of the question
- Retrieve a good sentence
- Return the appropriate named entity



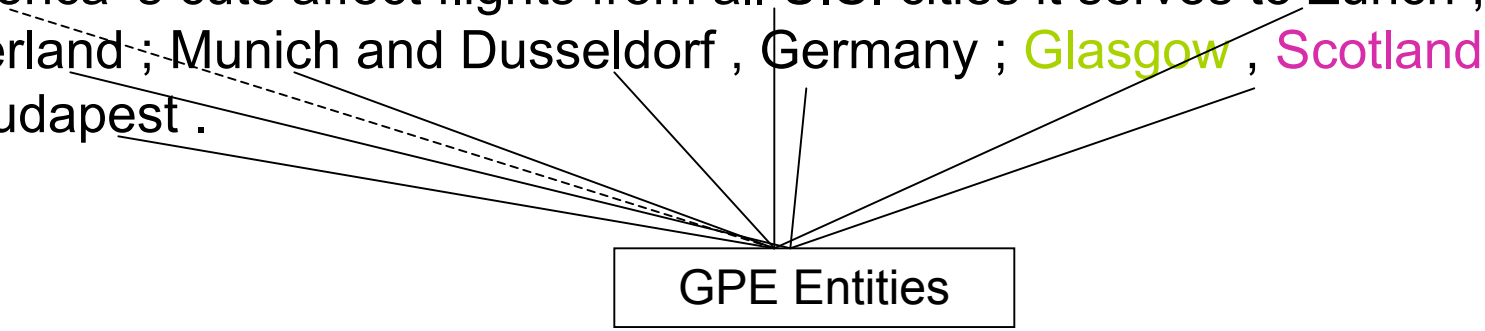
Answer Tagging (Harder)

- Determine the answer type of the question
- Retrieve a good sentence
- Return the appropriate named entity

Q: Where is Glasgow ?

A: The recession came late to Glasgow , as it did to the rest of Scotland.

A: America 's cuts affect flights from all U.S. cities it serves to Zurich , Switzerland ; Munich and Dusseldorf , Germany ; Glasgow , Scotland and Budapest .



Answer Patterns

- Answer Patterns are the text immediately surrounding an answer to a fact question
- Dependent on the question type
- Independent of the specific question

Answer Pattern Examples

"Inventor" pattern examples:

<NAME> , invented by <ANSWER>

<ANSWER>'s <NAME>

<NAME> was invented by <ANSWER>

<ANSWER> invented the <NAME>

Example:

And the demonstration just happens to come 115 years to the day after Edison invented the light bulb.

Answer Pattern Examples

"Discoverer" pattern examples:

When <ANSWER> discovered <NAME>

<ANSWER>'s discovery of <NAME>

<ANSWER> discovers <NAME>.

<NAME> was discovered by <ANSWER>

<ANSWER> discovered <NAME>

Example:

Gene Shoemaker discovered the comet which will hit Jupiter starting in about eight hours.

Birth-Year Pattern Examples

Dodi Fayed was born in 1956.

Moments after Samantha Crystal was born on July 17, 1996, doctors knew...

In her brief life and tragic death, Jessica Dubroff (1988 – 1996) became a metaphor for everything from youthful idealism to New-Age excess.

Hard Birth-Year Pattern Examples:

Born Israel Baline to a poor rural Russian family in 1888, Berlin taught himself to play the piano.

Born in 1924, almost exactly contemporary with Norman Mailer, he was brought up in New York City.

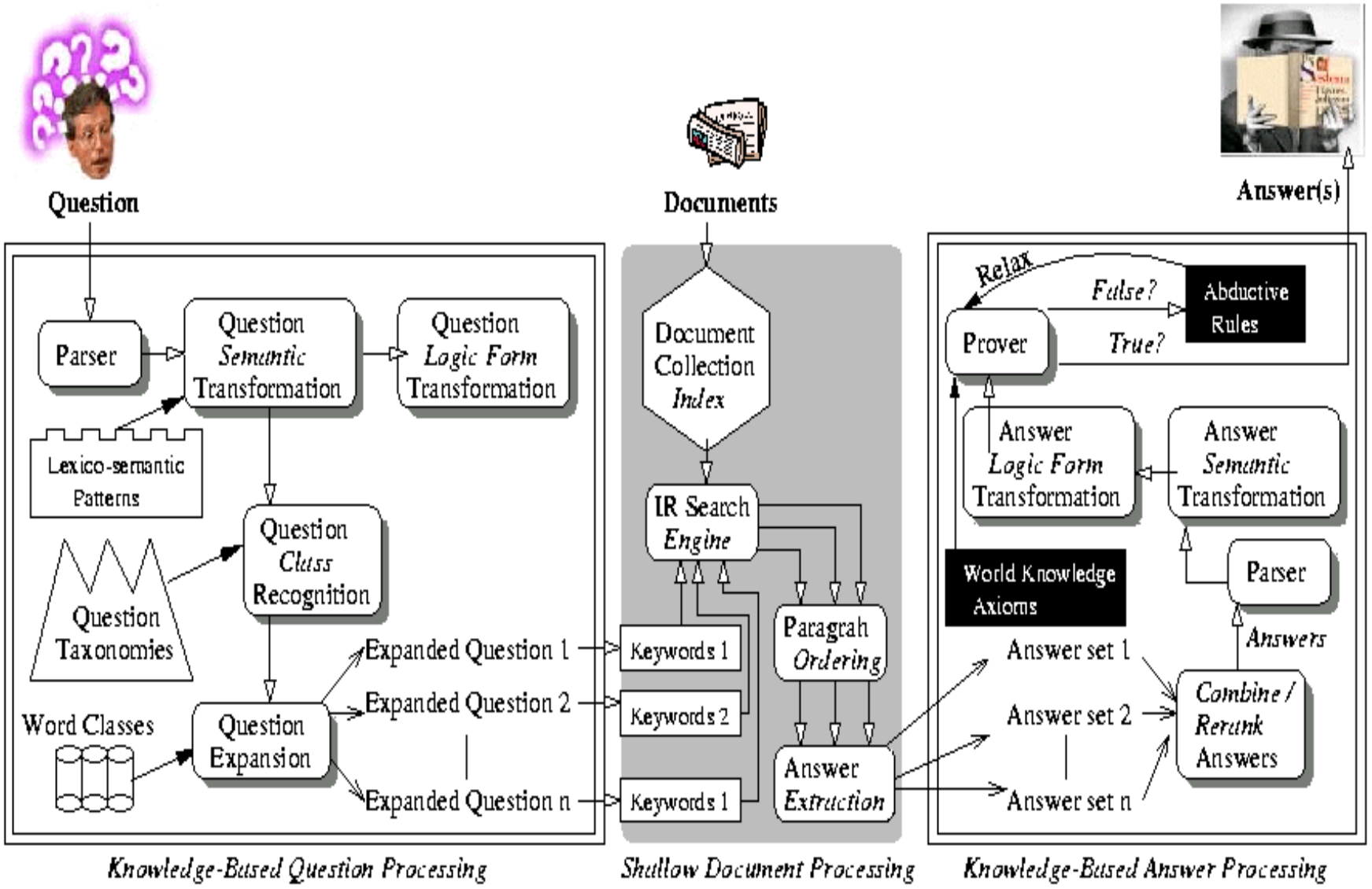
Some things to consider…

- For any given question type, there are potentially hundreds of ways to express the answer.
- Learning patterns depends on multiple unique examples of the same pattern.
- Newswire data has a limited number of examples of any given pattern.
- Newswire data is repetitive: there are many identical examples with different doc ids.

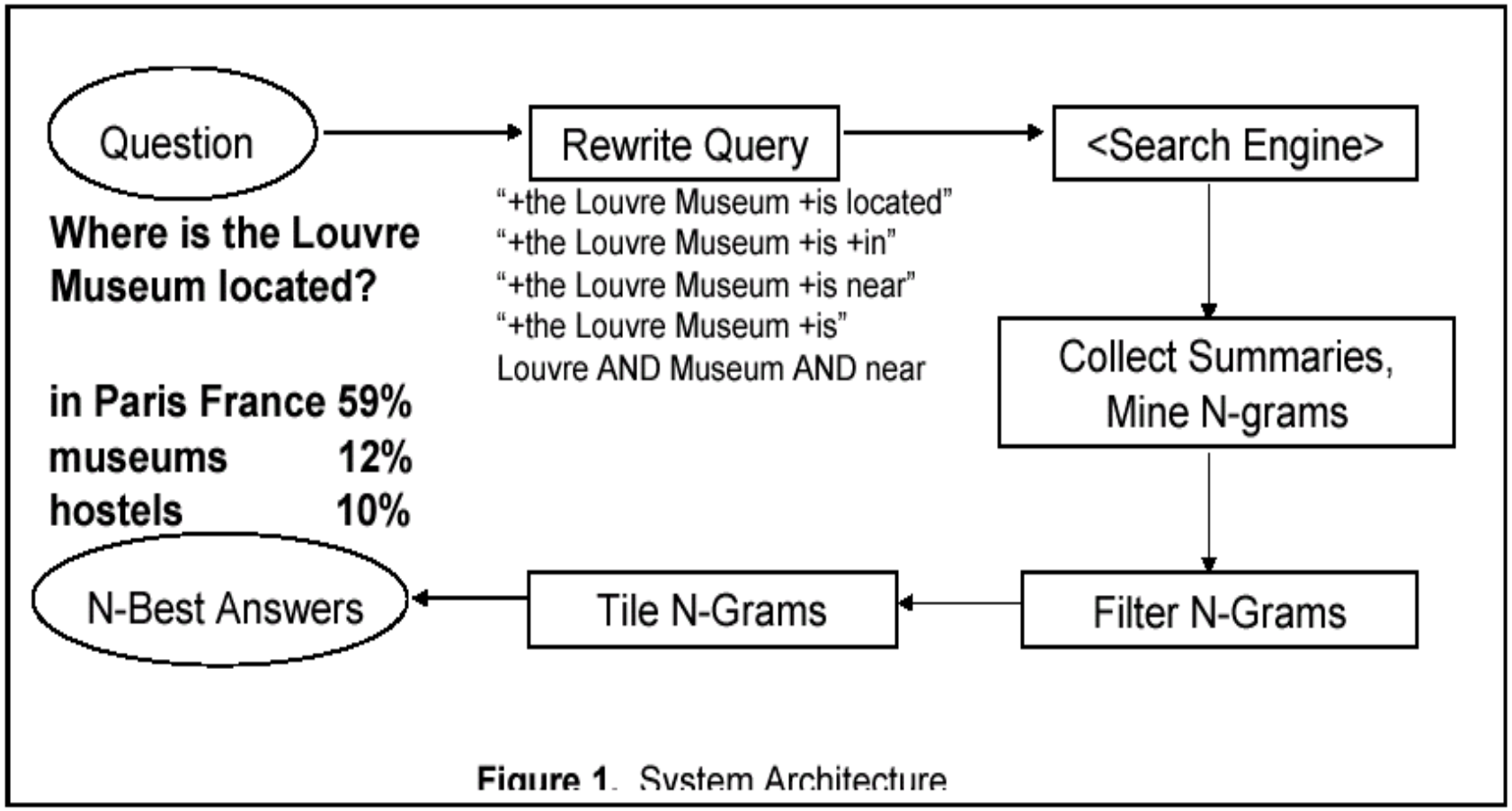
Overview

- Question Answering, in general
- Question Classification
- Information Retrieval
- External Resources
- Answer Extraction
- QA Systems from TREC
- Real-Life QA Systems

FALCON: Boosting Knowledge for Answer Engines. Sanda Harabagiu et al. The Ninth Text REtrieval Conference (TREC 9), 2000.



An Analysis of the AskMSR Question-Answering System. Eric Brill, Susan Dumais, Michele Banko. EMNLP 2002.



N-grams weighted by reliability of pattern retrieving, and frequency.

"Tiling" means voting on most frequent terms.

ISI Webclopedia

The External Use of Knowledge in Factoid QA
(Hovy,Hermjakob,Lin. TREC 10)

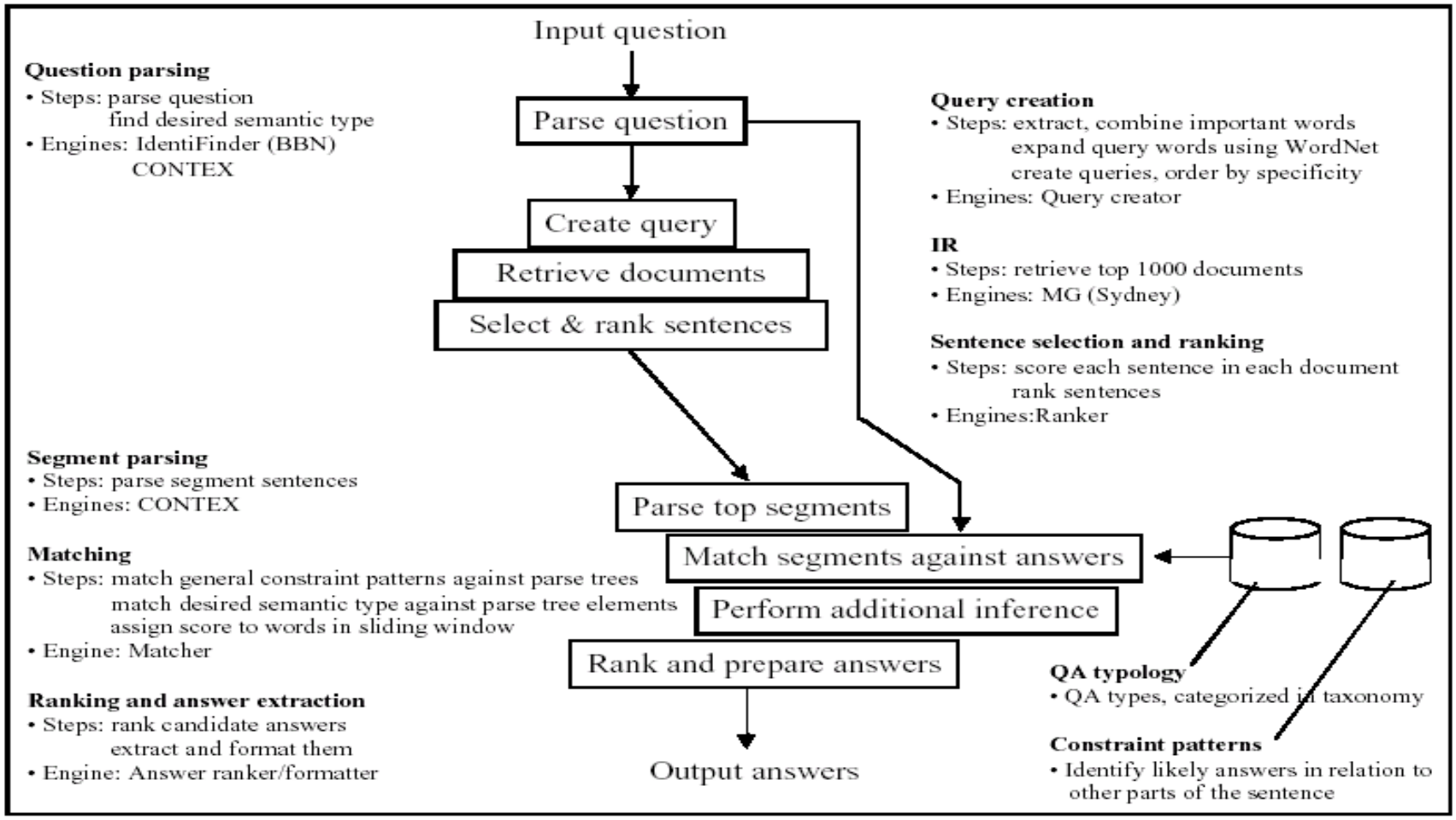


Figure 1. Webclopedia architecture.

Evaluation

- Precision at rank one
  - Percent of questions answered correctly
- MRR
  - 1/(rank of first correct answer)
- F-measure: $\frac{2PR}{P+R}$
- TREC:

  Fact score is precision

  Def score is approximately F-measure for definition components

  List score is F-measure for list components

Fact Questions from TREC 2003

Group	Fact Score	Def Score	Final Score
Language Computer Corp.	.700	.442	.559
Nat'l. Univ. of Singapore	.562	.473	.479
LexiClone	.622	.159	.363
USC (ISI)	.337	.461	.313
BBN	.206	.555	.266
MIT	.293	.309	.256
ITC-irst	.235	.317	.216
IBM Research	.298	.175	.212
Univ. of Albany	.240	.146	.178
Fudan University	.191	.192	.165

Final Score = ½ Fact score + ¼ list score + ¼ def score

http://www.trec.nist.gov

TREC 2004

3	Hale Bopp comet		
	3.1	FACTOID	When was the comet discovered?
	3.2	FACTOID	How often does it approach the earth?
	3.3	LIST	In what countries was the comet visible on its last return?
	3.4	OTHER	
21	Club Med		
	21.1	FACTOID	How many Club Med vacation spots are there worldwide?
	21.2	LIST	List the spots in the United States.
	21.3	FACTOID	Where is an adults-only Club Med?
	21.4	OTHER	
22	Franz Kafka		
	22.1	FACTOID	Where was Franz Kafka born?
	22.2	FACTOID	When was he born?
	22.3	FACTOID	What is his ethnic background?
	22.4	LIST	What books did he author?
	22.5	OTHER	

TREC 2004 Results

Table 2: Evaluation scores for runs with the best factoid component.

Run Tag	Submitter	Accuracy			NIL Prec	NIL Recall
		All	Initial	Non-Initial		
lcc1	Language Computer Corp.	0.770	0.839	0.744	0.857	0.545
uwbqitekat04	Univ. of Wales, Bangor	0.643	0.694	0.625	0.247	0.864
NUSCHUA1	National Univ. of Singapore	0.626	0.710	0.595	0.333	0.273
mk2004qar1	Saarland University	0.343	0.419	0.315	0.177	0.500
IBM1	IBM Research	0.313	0.435	0.268	—	0.000
mit1	MIT	0.313	0.468	0.256	0.083	0.045
irst04higher	ITC-irst	0.291	0.355	0.268	0.167	0.091
FDUQA13a	Fudan University (Wu)	0.257	0.355	0.220	0.167	0.091
KUQA1	Korea University	0.222	0.226	0.220	0.042	0.045
shef04afv	University of Sheffi eld	0.213	0.177	0.226	0.071	0.136

Real-Life QA Systems

- AskJeeves (is not a QA system):
  http://www.ask.com
- BrainBoost:
  http://www.brainboost.com/
- START:
  http://start.csail.mit.edu/
- LCC:

  http://www.languagecomputer.com/solutions/question_answering/index.html

Summary

- QA systems are comprised of components that are cobbled together – it's not always obvious how (or why) they work.
- Failure in one component propagates through the system
- Answers have a high degree of variation
- Most systems are brittle – unexpected types of questions fail

Open Problems

- Is there a unifying mathematical framework?
- What are the uses for question answering?
- Non-fact questions
- Generating natural language answers
  - Sentences provide context
  - We would like exact answer sentences, but how?
- Interactive Question Answering

Open Problems

- Combining multiple sources
  - How do we compare results from multiple sources?
  - Stitch together sentences?
  - Combine sentences into paragraphs?
  - Do we choose similar or novel pieces of information?
  - How do we combine structured and unstructured data?

Open Problems

- Answer granularity
  - Fact questions:
    - Token?
    - Phrase?
    - More than a phrase?
  - Which questions require what size answer?
  - Questions not answerable with facts:
    - What are they answerable with?
      - Forms?
      - A narrative document?
      - A table?

Resources

- TREC QA Track
  - Data: http://trec.nist.gov/data/qa.html
  - Publications: http://trec.nist.gov/pubs.html
- Conferences such as ACL, EMNLP, SIGIR frequently have QA tracks or workshops:
  - NAACL/HLT Workshop on Interactive QA: http://www.ils.albany.edu/IQA06/
  - ACL Workshop on Task-Focused Summarization and Question-Answering: http://research.microsoft.com/~lucyv/WS7.htm