

# CS474 Intro to Natural Language Processing

## Question Answering

## Question answering

- Overview and task definition
- ➔ **History**
- **Open-domain question answering**
- **Basic system architecture**
- **Predictive indexing methods**
- **Pattern-matching methods**
- **Advanced techniques**

## History

- **Closed-domain QA systems**
  - LUNAR [Woods & Kaplan, 1977]
  - WOLFIE [Thompson & Mooney, 1998]
  - Q/A [Lehnert, 1978]
- **Open-domain QA systems**
  - TREC QA evaluations [1999, 2000, 2001, ...]


## LUNAR

- **Answered questions about moon rocks and soil gathered by the Apollo 11 mission**
  - Data base of information for all collected samples
- **Architecture**
  - Parse English question into a data base query
    - Syntactic analysis via augmented transition network parser and heuristics (including some for shallow semantics)
    - Semantic analysis maps parsed request into query language; query denotes unambiguous meaning of the request
  - Run query on data base to produce answer

## LUNAR

- **Resources required**
  - Parser for a subset of English (size unclear)
  - Handled tense, modality, some anaphora, some conjunctions, some relative clauses, some adjective modifiers (dealing with quantification)
  - Vocabulary of about 3,500 words
- **Sample questions**
  - What is the average concentration of aluminum in high alkali rocks?
  - What samples contain P205?
  - Give me the modal analyses of P205 in those samples.

## LUNAR example

- **Do any samples have greater than 13 percent aluminum?**
- **Data base query**  
(TEST (FOR SOME X1 / (SEQ SAMPLES): class to test  
T;  
(CONTAIN X1  
(NPR\* X2 / 'AL203)  
(GREATERTHAN 13 PCT))))  
  
No restriction on class  
Proposition
- **Answer:**
  - Yes

## LUNAR assessment

- **System characteristics**
  - Closed domain (lunar geology and chemistry)
  - Structured data (information contained in a data base)
  - Structured answers (information contained in a data base)
    - Avoided dialogue problems
  - Context: sophisticated users demanding high accuracy
- **Labor intensive to build**
  - Complex system
  - High accuracy required
  - Few general-purpose NLP resources available at the time

## LUNAR assessment

- **Research on systems like LUNAR continued for another decade**
- **Focused on**
  - Syntactic parsing
  - Incorporating domain knowledge
  - Dialogue management
- **Problems**
  - Expensive to build
  - Brittle...prone to unexpected sudden failure

## WOLFIE

- **W**ord **L**earning **F**rom **I**nterpreted **E**xamples  
[Thompson and Mooney, 1998]
  - Closed domain, structured data, structured answers
  - Avoids labor-intensive system development
  - Uses *inductive logic programming* methods to acquire parsers that can map natural language queries into executable logical form (i.e. data base queries)
    - Requires examples of NL queries and their logical form
  - Indirectly evaluate the parser based on the number of queries that system gets right/wrong.

## Lehnert's Q/A system

- **Implemented a broader theory of question answering**
  - motivated by issues of cognitive plausibility
  - relied on the linguistic/cognitive theories of *conceptual dependency*, *scripts*, *plans*, etc.  
[Schank, 1970's]
  - used a question taxonomy
  - somewhat closed domain (actions), unstructured data, generated answers
  - answered questions about an arbitrary input text (usually event-based)

## Q/A example 1

- **Input text**

John threw the baseball to Mary. She missed the ball and it hit her on the head.
- **Questions:**
  - Was Mary happy?
  - Who has the baseball?
  - Why did John throw the baseball to Mary?

## Lehnert's Q/A system

- **“Parse” input text into a semantic representation (*conceptual dependency*)**
- **Generate inferences from that representation**
  - Inferences associated with CD's semantic primitives
- **“Parse” the question, mapping it into one of the predefined question types**
- **Employ the method associated with the question type to answer the question**

## Q/A example 2

- **Input text**

For their first date, John took Mary to McDonald's for burgers. Mary was not impressed.

- **Questions:**

- Did John and Mary pay for the burgers?
- What did John and Mary eat for dinner on their first date?

## Q/A assessment

- **Labor intensive to build**

- Complex system
- Background knowledge needed
  - Data structures to encode scripts, plans, goals, inferences associated with CD primitives
- Few general-purpose resources available at the time

- **Not designed to be a general-purpose Q/A system**

## Question answering

- **Overview and task definition**

- **History**

- ➔ **Open-domain question answering**

- **Basic system architecture**
- **Predictive indexing methods**
- **Pattern-matching methods**
- **Advanced techniques**

## Towards open-domain QA

Which country has the largest part of the Amazon rain forest?

The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by officials of Amazon countries and leading scientists from around the world.

“That’s some of the most encouraging news about the Amazon rain forest in recent years,” said Thomas Lovejoy, a tropical ecologist at the Smithsonian Institution and an Amazon specialist.

“It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon, especially in **Brazil**, except bad news,” Lovejoy said in a recent interview.

**Sixty percent of the Amazon**, the world’s largest tropical rain forest, **lies in Brazil**, but the forest also covers parts of the eight surrounding countries.

Lovejoy was one of the organizers of an unusual workshop held in mid-January in Manaus, **Brazil**, a sprawling city of 1 million people in the heart of the Amazon. It was the center of **Brazil**’s once-thriving rubber trade.

## Question Answering

- **Simplifications**
  - short-answer, fact-based questions
  - answer exists in the collection as a text fragment
  - supporting info can be found in a single document
  - system returns up to 5 guesses per question
- **Sample questions**
  - How many calories are there in a Big Mac?
  - Who is the voice of Miss Piggy?
  - Who was the first American in space?
  - Where is the Taj Mahal?

## TREC QA: evaluation

- **Human assessors judge the answers**
  - Allowed to accept multiple answers
- **Systems scored on *mean reciprocal rank (MRR)* of first correct answer**
  - if first answer correct = 1 point,
  - else if second answer correct = 1/2 point,
  - else if third answer correct = 1/3 point, ...
  - 0 if none of the  $n$  answers are correct
  - Average of MRR across all questions
- **Also reported on the number of questions answered correctly**

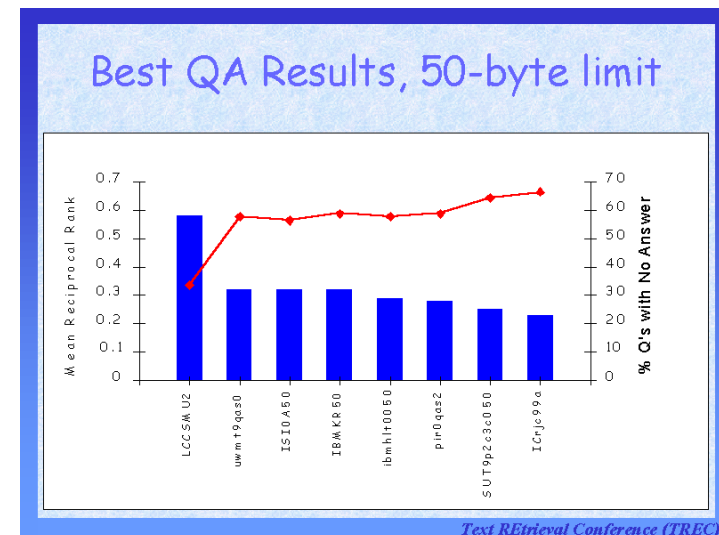
## Question Answering

- **Performance**

	<b>%correct</b>	<b>avg rank 1st</b>
TREC-8 (1999):	70%	1.4
TREC-9 (2000):	65%	1.7
TREC-10 (2001):	70%	1.3
TREC-11 (2002):	83%	1
TREC-12 (2003):	70%	1
TREC-13 (2004):	84%	1

*harder questions  
NIL answers  
1 guess, exact*

## TREC 2000



Courtesy of  
E. Voorhees

## List Questions

- **List questions**

1915: List the names of chewing gums.

Stimorol	Orbit	Winterfresh	Double Bubble
Dirof	Trident	Spearmint	Bazooka
Doublemint	Dentyne	Freedent	Hubba Bubba
Juicy Fruit	Big Red	Chiclets	Nicorette

- **Can't just rely on a single document**

- **Performance**

- TREC 2003: F .40
- TREC 2004: F .62

## Definition Questions

- **Who is Colin Powell?**

- **What is mold?**

- **Hard to evaluate**

- Who is the audience?
- Evaluation requires matching *concepts* in the desired response to *concepts* in a system response
  - TREC 2003:
    - Audience: questioner is an adult, a native speaker of English, and an “average” reader of US newspapers
    - Results: F .55

## Context Task

- **Track a target discourse object through a series of questions**

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	

- **Performance**

- TREC 2004
  - Factoids: .84 initial; .74 non-initial
  - Lists: .62 F
  - Other: .46 F

## Question answering

- **Overview and task definition**

- **History**

- **Open-domain question answering**

- **Basic system architecture**

[Cardie et al., ANLP 2000]

- **Predictive indexing methods**

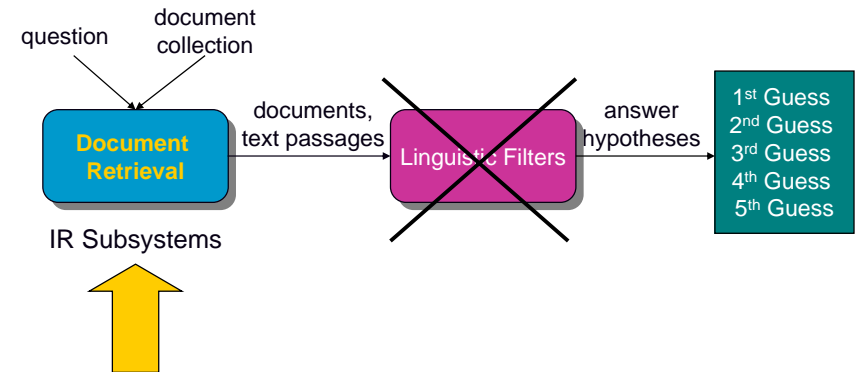
- **Pattern-matching methods**

- **Advanced techniques**

## Basic system architecture



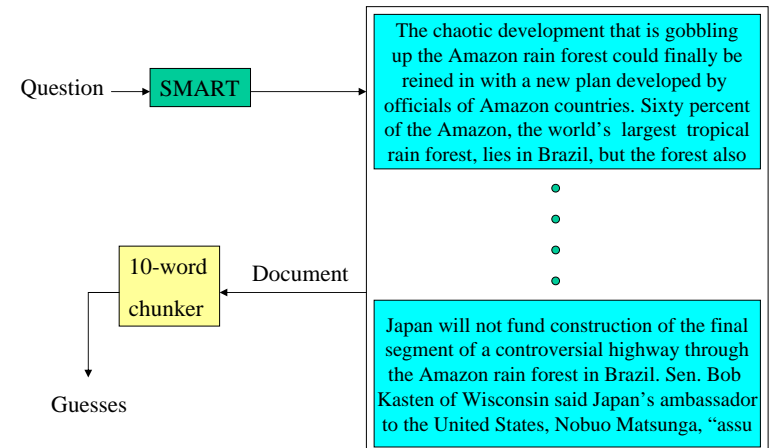
## System architecture: document retrieval



## Document retrieval

- **Standard ad-hoc IR using full-text indexing**
- **Example QA system uses**
  - vector space model
  - text retrieval system: Smart
  - standard term-weighting strategies (tfidf)
  - no automatic relevance feedback

## QA as document retrieval



## Baseline evaluation

- Document retrieval only
- Corpus
  - TREC-8 development corpus (38 questions)
  - TREC-8 test corpus (200 questions)

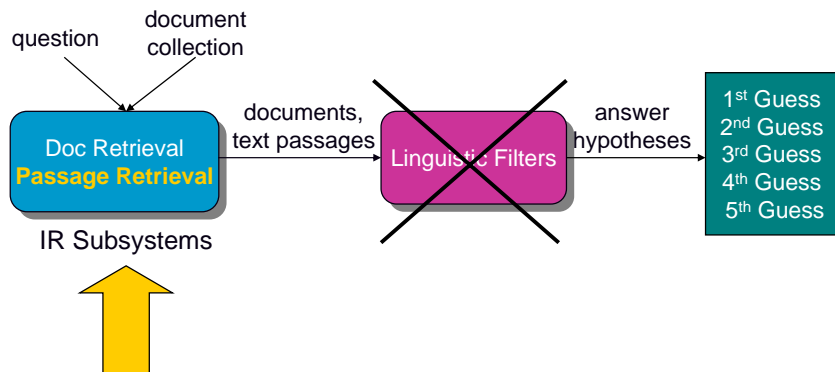
	Development (38)		Test (200)	
	Correct	MAR	Correct	MAR
Smart	3	3.33	29	3.07

MAR = Mean Answer Rank

## Baseline evaluation

- Smart performs better than its scores would suggest
  - Development corpus
    - For 18 of 38 questions, the answer appears in the top-ranked document
    - For 33 of 38 questions, the answer appears in one of the top 7 documents
    - For only 2 questions does Smart fail to retrieve the answer in one of the top 25 documents
  - Test corpus
    - Over half (110) of the questions are answered in the top-ranked document
    - Over 75% of the questions (155) are answered in the top 5 documents
    - 19 questions were not answered in the top 20 documents

## System architecture: passage retrieval



## Passage retrieval

[Salton *et al.*]

### Query-dependent text summarization

Which country has the largest part of the Amazon rain forest?

[The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by leading scientists from around the world.] [That's some of the most encouraging news about the Amazon rain forest in recent years," said Thomas Lovejoy, an Amazon specialist.] [It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon.]

[Sixty percent of the Amazon, the world's largest tropical rain forest, lies in Brazil.]

Extract passages that best summarize each document w.r.t. the query



## Query-dependent text summarization

- **Basic algorithm**

1. Decide on a summary length (10% of document length).
2. Use standard ad-hoc retrieval algorithm to retrieve top documents.
3. *Treat each sentence/paragraph in top N documents as a document itself.*

Use standard document similarity equations to assign a similarity score to the sentence/paragraph.

4. Return highest-scoring sentences/paragraphs as the summary, subject to the length constraint.

## Passage retrieval

[Salton *et al.*]

### Query-dependent text summarization

Which country has the largest part of the Amazon rain forest?

[The chaotic development that is gobbling up the Amazon rain forest could finally be reined in with a new plan developed by leading scientists from around the world.] [“That’s some of the most encouraging news about the Amazon rain forest in recent years,” said Thomas Lovejoy, an Amazon specialist.] [“It contrasts markedly with a year ago, when there was nothing to read about conservation in the Amazon.”]

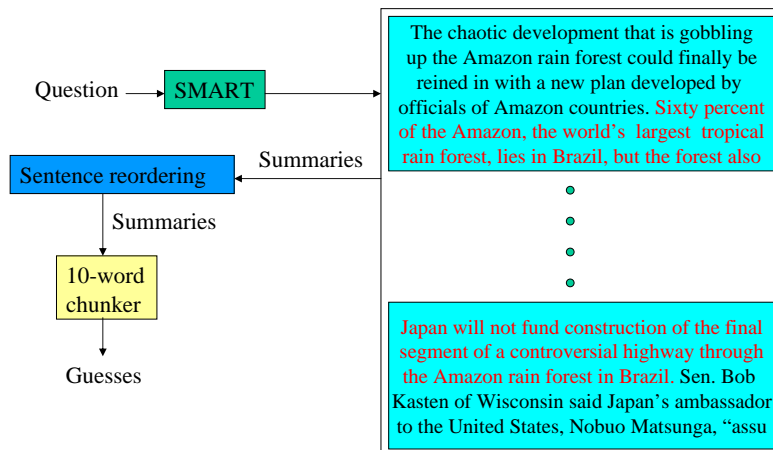
[Sixty percent of the Amazon, the world’s largest tropical rain forest, lies in Brazil.]

Sort summary extracts across top k documents

ordered list of summary extracts

answer hypotheses

## QA as query-dependent text summarization



## Evaluation: text summarization

	Development (38)		Test (200)	
	Correct	MAR	Correct	MAR
Smart	3	3.33	29	3.07
Text Summarization	4	2.25	45	2.67

**MAR = Mean Answer Rank**

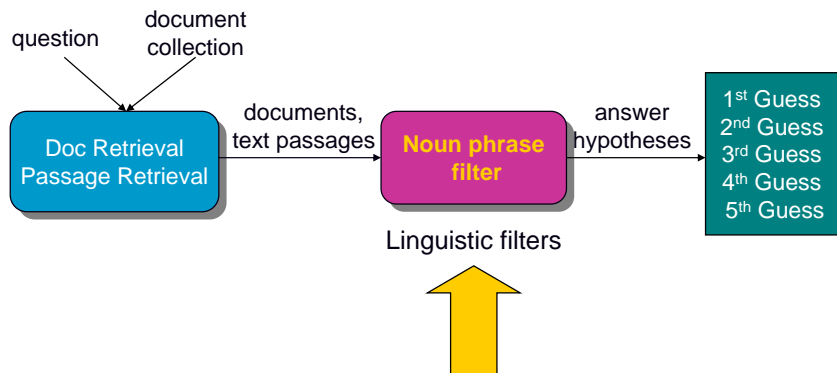
## Evaluation: text summarization

- **Summarization method can limit performance**
  - Development corpus
    - In only 23 of the 38 developments questions (61%) does the correct answer appear in the summary for one of the top  $k=7$  documents
  - Test corpus
    - In only 135 of the 200 developments questions (67.5%) does the correct answer appear in the summary for one of the top ( $k=6$ ) documents

## Linguistic filters

- **50 byte answer length effectively eliminates *how* or *why* questions**
- **almost all of the remaining question types are likely to have **noun phrases** as answers**
  - development corpus: 36 of 38 questions have noun phrase answers
- **consider adding at least a simple linguistic filter that considers only noun phrases as answer hypotheses**

## System architecture: linguistic filters



## The noun phrase filter

Which country has the largest part of the Amazon rain forest?

ordered list of summary extracts

[The huge Amazon rain forest] is regarded as vital to [the global environment].

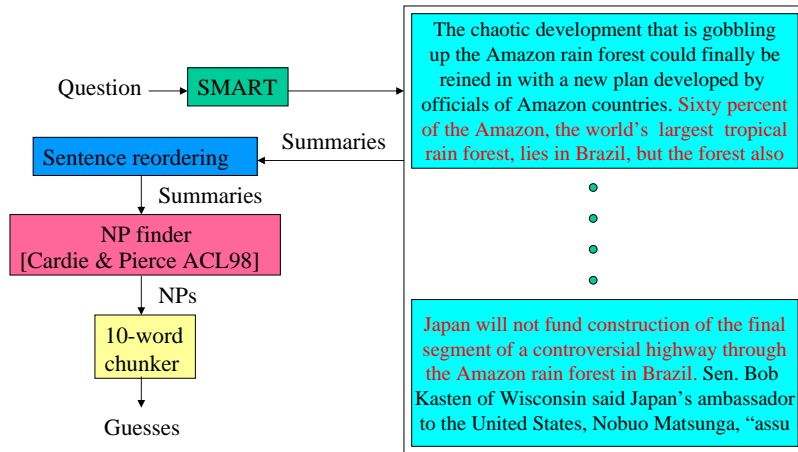
[Japan] will not fund [the construction] of [the final segment] of [a controversial highway] through [the Amazon rain forest] in [Brazil], according to [a senior Republican senator].

•  
•  
•

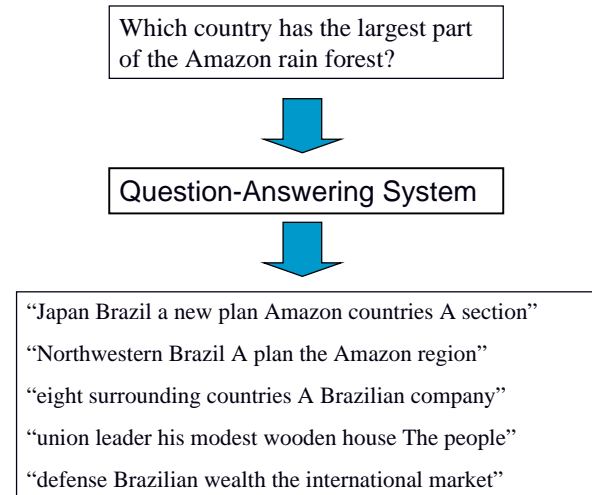
ordered list of NPs

answer hypotheses

## QA using the NP filter



## Chunking answer hypotheses: BAD



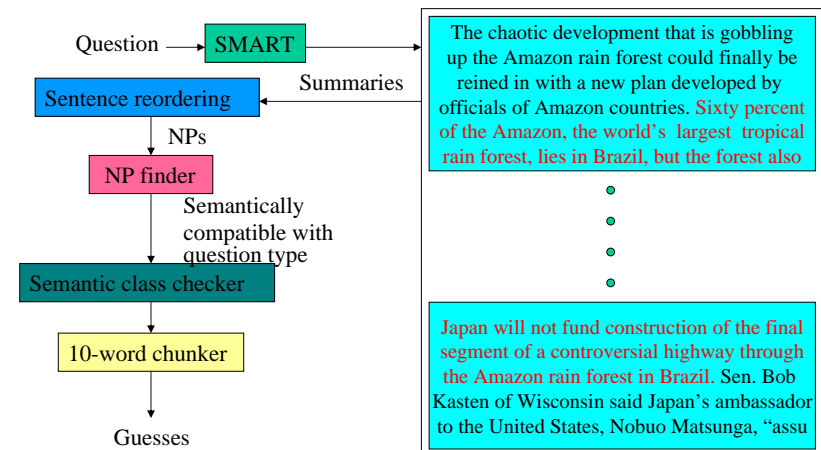
## Evaluation: NP filter

	Development (38)		Test (200)	
	Correct	MAR	Correct	MAR
Smart	3	3.33	29	3.07
Text Summarization	4	2.25	45	2.67
TS + NPs	7	2.29	50	2.66

**MAR = Mean Answer Rank**

- Using NP finder of Cardie & Pierce (1998)
  - ~94% precision and recall on Wall Street Journal text
- How much does the (unnatural) NP “chunking” help?
  - Without it, only 1 and 20 questions answered for each corpus, respectively
  - NP filter is extracting good guesses, but better linguistic processing is needed to promote the best guesses to the top of the ranked guess list

## Semantic class checking

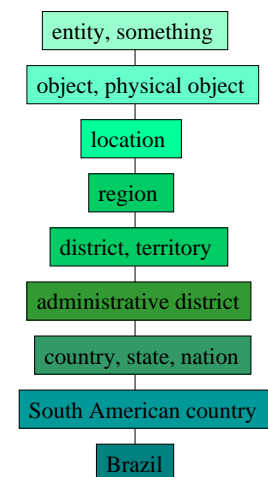


## Semantic class checking

- **Approximate question type using question word**
  - Who is the president of the U.S.?  
person
  - Which country has the largest part of the Amazon rain forest?  
country
  - Where is the Connecticut River?  
state? county? country? location?
  - What fabric should one use to make curtains?  
fabric???
- **Check that head noun (i.e. the last noun) of answer NP is of the same type**
  - a man = person
  - Massachusetts = state, location

## Semantic type checking

- **Use lexical resource to determine semantic compatibility**
  - WordNet!
- **Proper names handled separately since they are unlikely to appear in WordNet**
  - Small set (~20) rules



## Evaluation: semantic class filter

	Development (38)		Test (200)	
	Correct	MAR	Correct	MAR
Smart	3	3.33	29	3.07
Text Summarization	4	2.25	45	2.67
TS + NPs	7	2.29	50	2.66
TS + NPs + Semantic Type	21	1.38	86	1.90

MAR = Mean Answer Rank

- **Weak syntactic and semantic information allows large improvements**
- **Problems?**

## Sources of error

