Examining the Role of Statistical and Linguistic Knowledge Sources in a General-Knowledge Question-Answering System

Claire Cardie   Vincent Ng   David Pierce

Cornell University

Chris Buckley

SaBIR Research

Question Answering

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.

TREC Q&A Framework

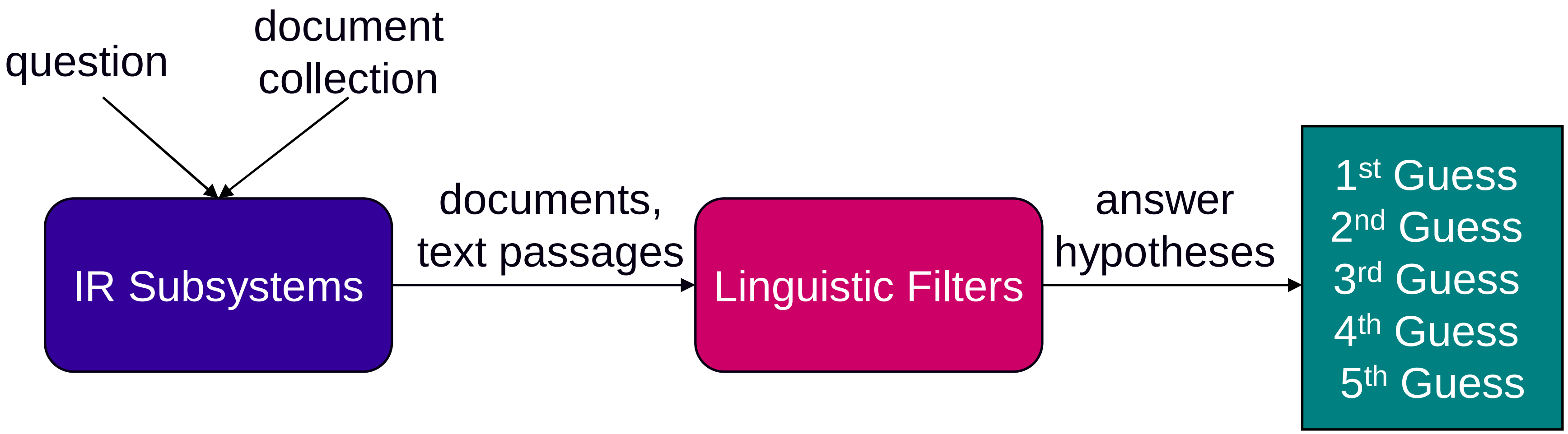
- Restrictions
  - The answer exists in the collection
  - All supporting info can be found in a single document
  - The answer is short (less than 50 bytes)
  - Can return up to 5 guesses per question to the user

Goals of the Research

Investigate the role of statistical and linguistic knowledge sources in a general-knowledge question answering system

- u Knowledge sources
  - – Word co-occurrence information
    - § Standard IR techniques
  - – Syntactic information
    - § Noun phrase bracketing
  - – Semantic information
    - § Semantic type checking

System Architecture



question

document collection

IR Subsystems

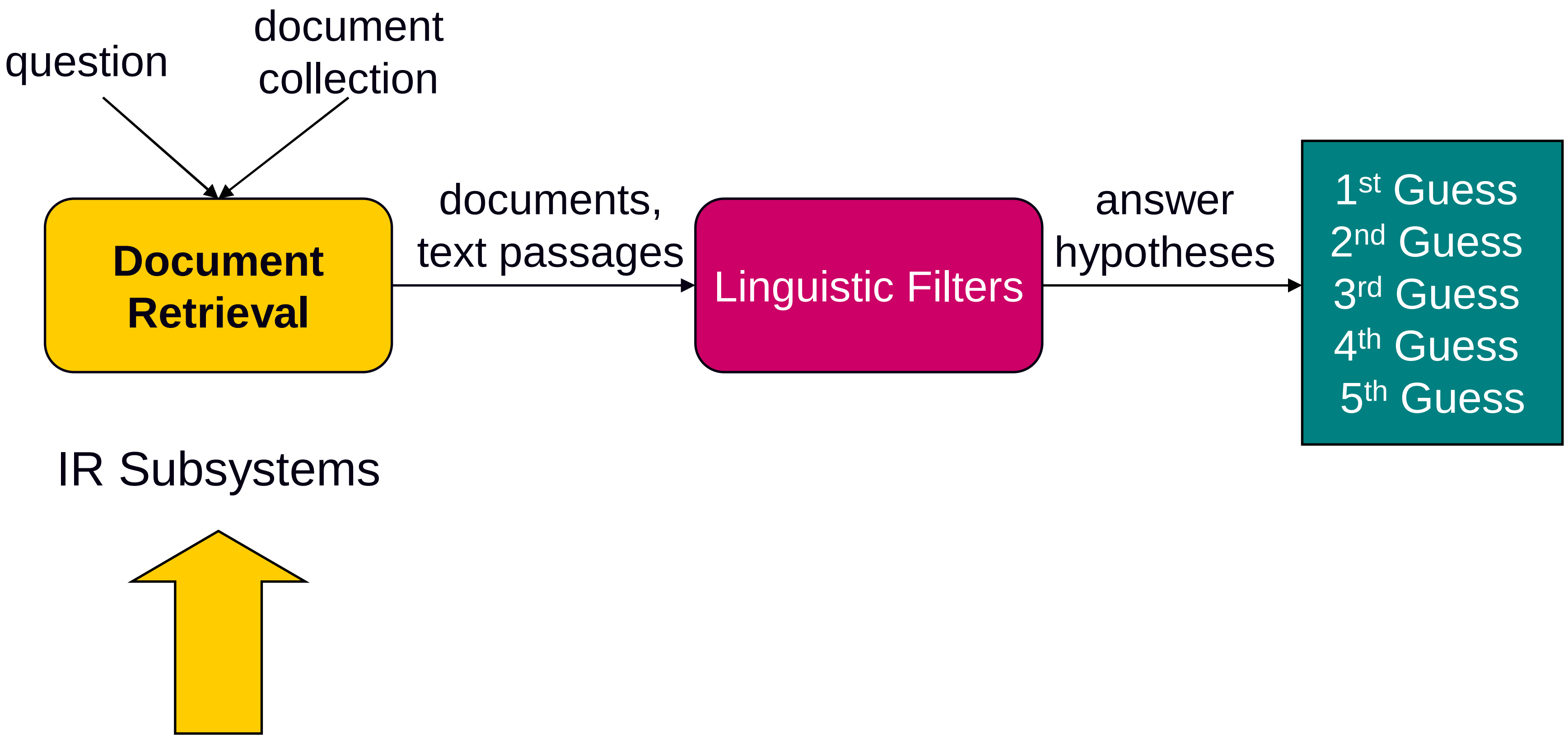
documents, text passages

Linguistic Filters

answer hypotheses

1st Guess
2nd Guess
3rd Guess
4th Guess
5th Guess

System Architecture



Document Retrieval

- Vector space model for document retrieval

- Text retrieval system: Smart
  - Standard term-weighting strategies
  - No automatic relevance feedback

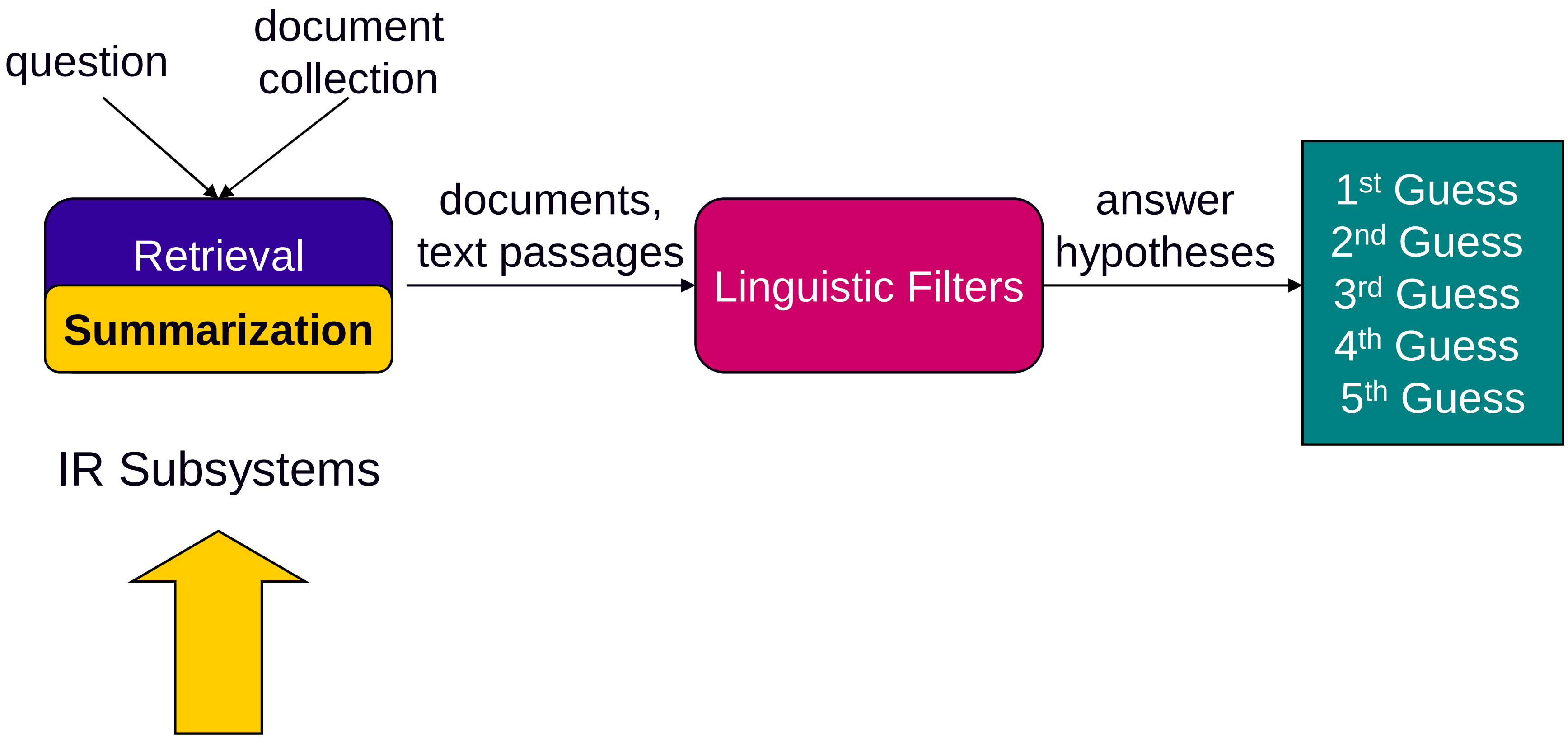
Baseline Evaluation

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

MAR = Mean Answer Rank

- Corpora
  - TREC-8 development corpus (38 questions)
  - TREC-8 test corpus (200 questions)
- Smart performs better than its scores would suggest

System Architecture



question

document collection

Retrieval

Summarization

documents, text passages

Linguistic Filters

answer hypotheses

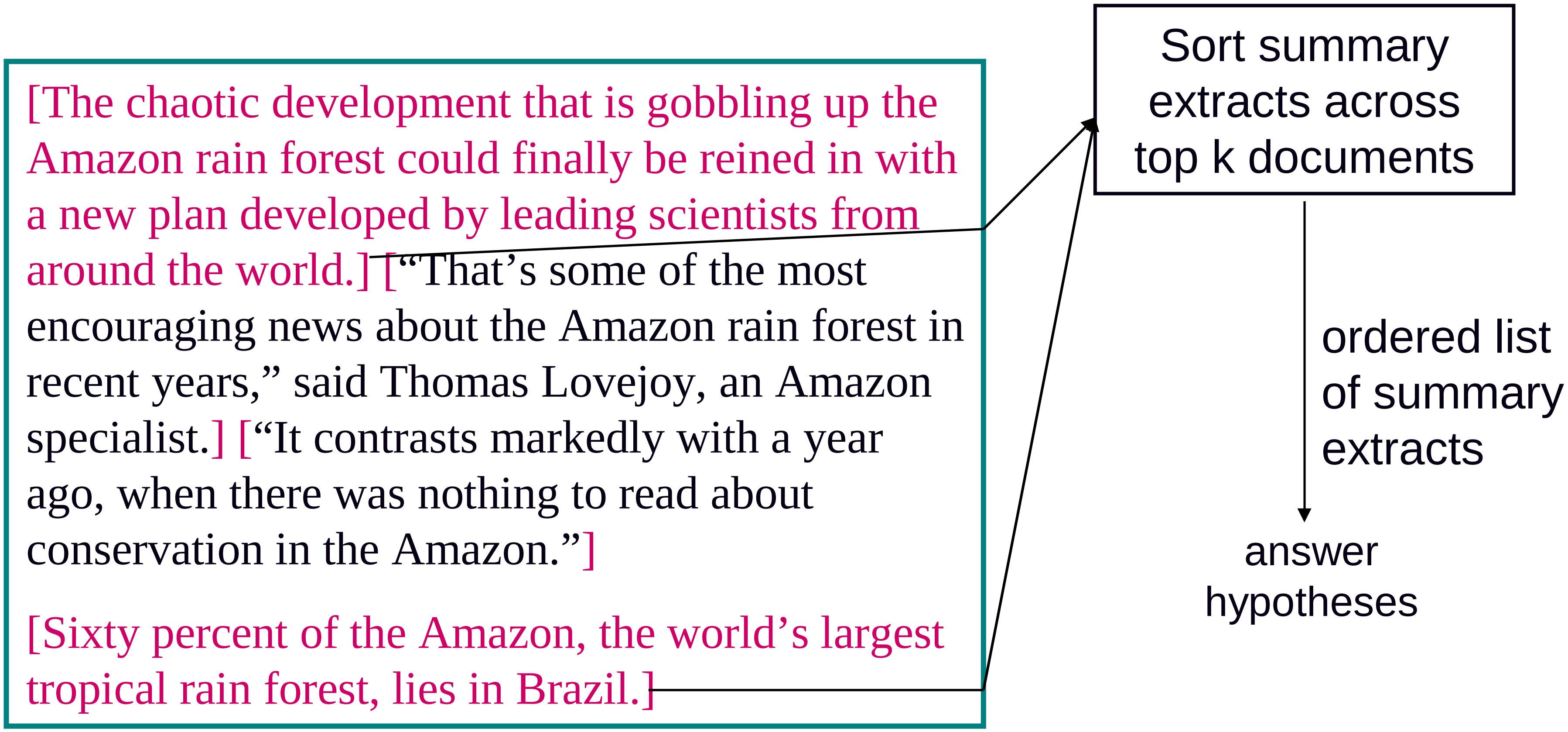
1st Guess
2nd Guess
3rd Guess
4th Guess
5th Guess

IR Subsystems

Query-Dependent Text Summarization

[Salton *et al.*]

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

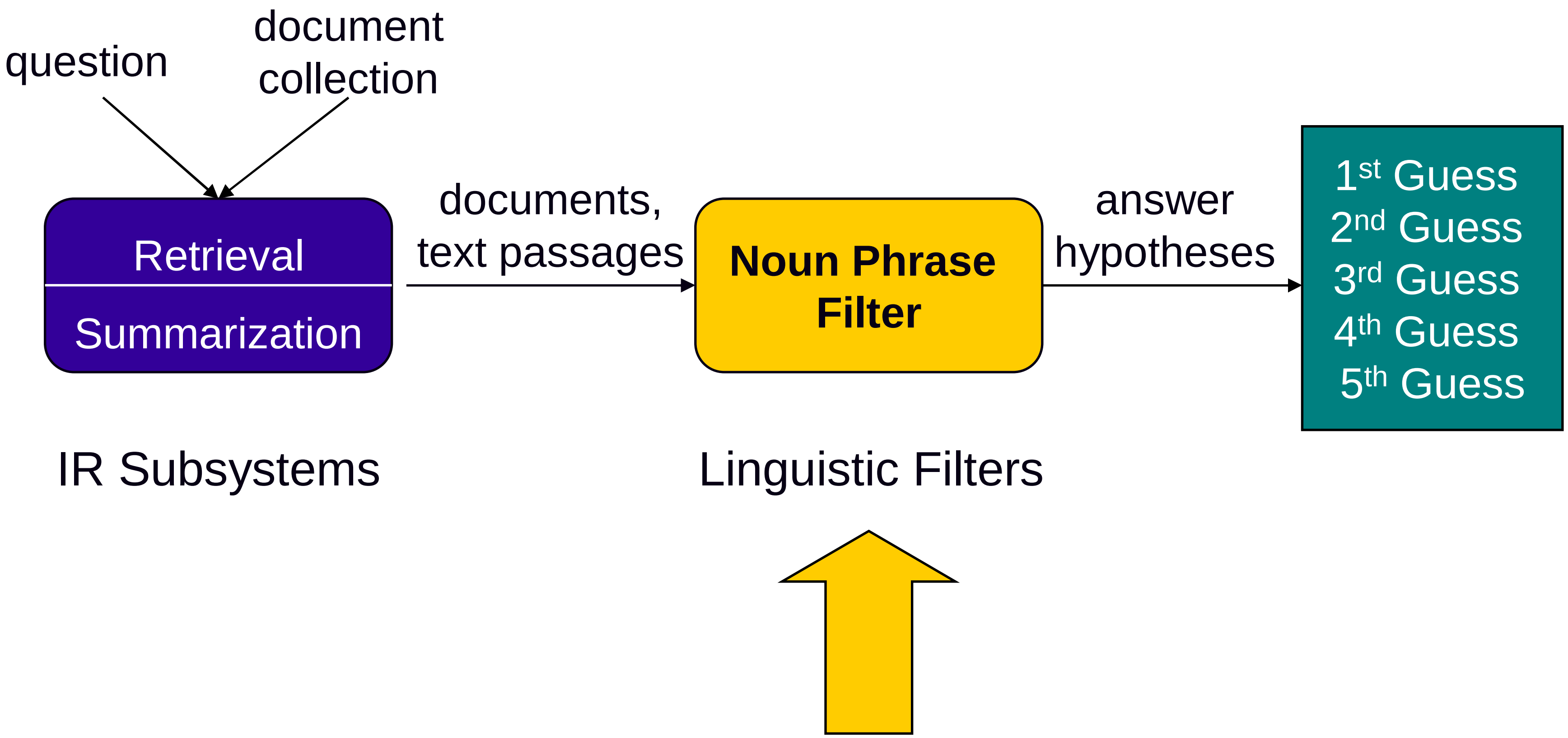
Evaluation (Text Summarization)

	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

- Summarization method can limit performance

System Architecture



question

document collection

Retrieval

Summarization

documents, text passages

Noun Phrase Filter

answer hypotheses

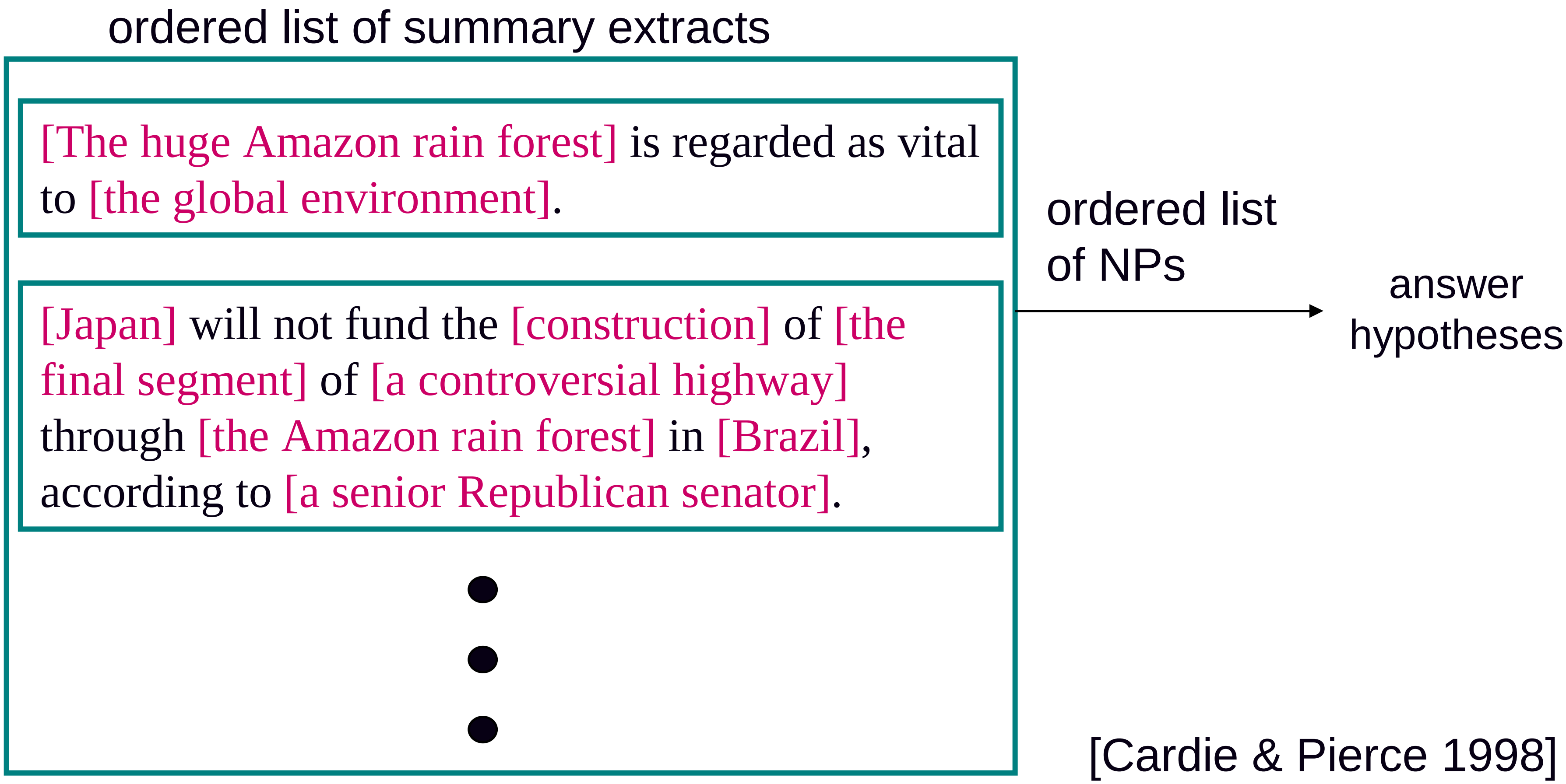
1st Guess
2nd Guess
3rd Guess
4th Guess
5th Guess

IR Subsystems

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

[Cardie & Pierce 1998]

Evaluation (TS + NPs)

	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

- Comparison with other TREC QA systems
  - One NP per guess

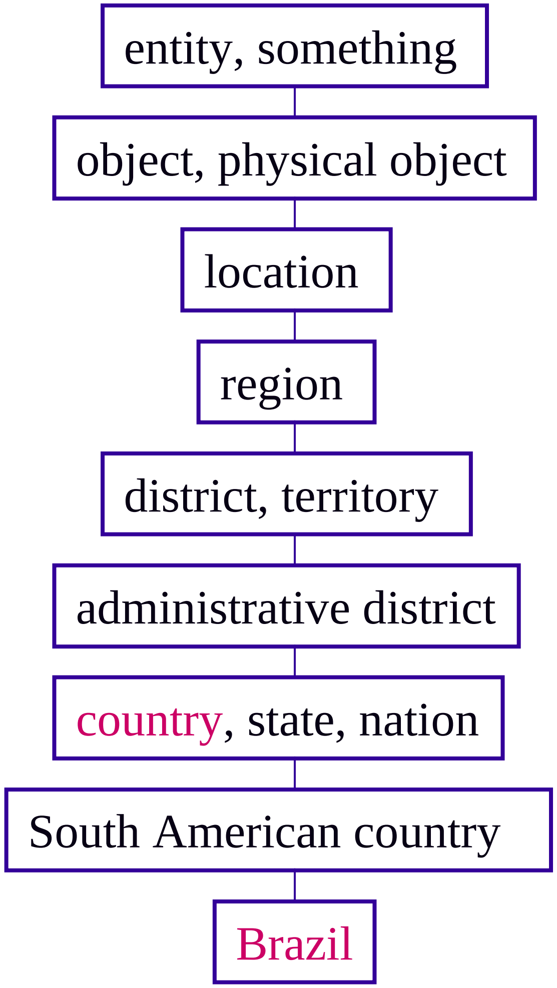
Semantic Type Checking

1. Approximate question type using question word/head noun

- Who is the president of the US?
- Which country has the largest part of the Amazon rain forest?

2. Check that ancestor-descendent relationship holds in the type hierarchy

3. Use heuristics for words not in WordNet



entity, something

object, physical object

location

region

district, territory

administrative district

country, state, nation

South American country

Brazil

Evaluation
(TS + NPs + Semantic Type)

	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

- Heuristics do not generalize well

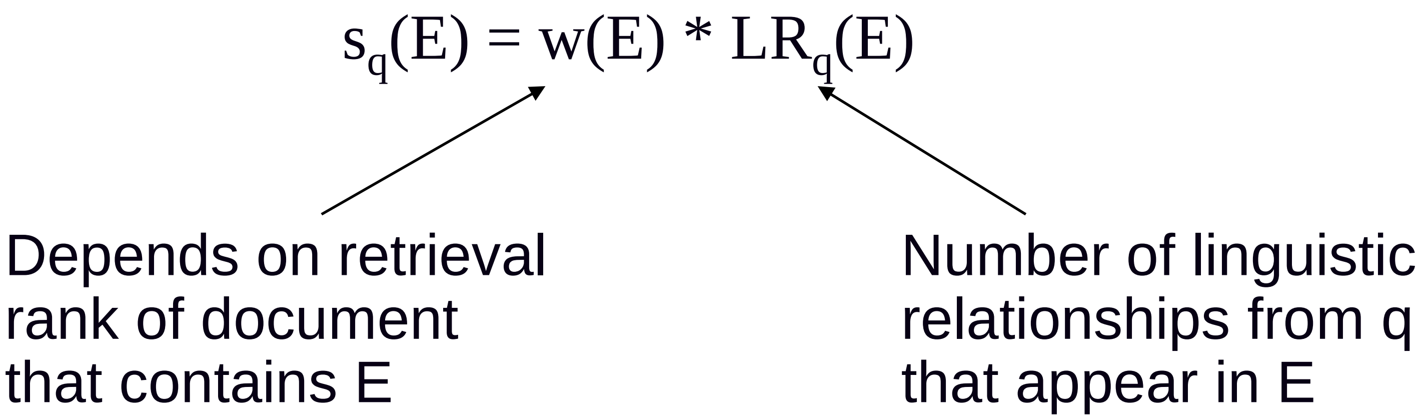
Reordering Summary Extracts

- Approach

  Define a new scoring scheme to reorder summary extracts based on linguistic knowledge

- New scoring measure

  For each summary extract E for question q



$$s_q(E) = w(E) * LR_q(E)$$

Depends on retrieval
rank of document
that contains E

Number of linguistic
relationships from q
that appear in E

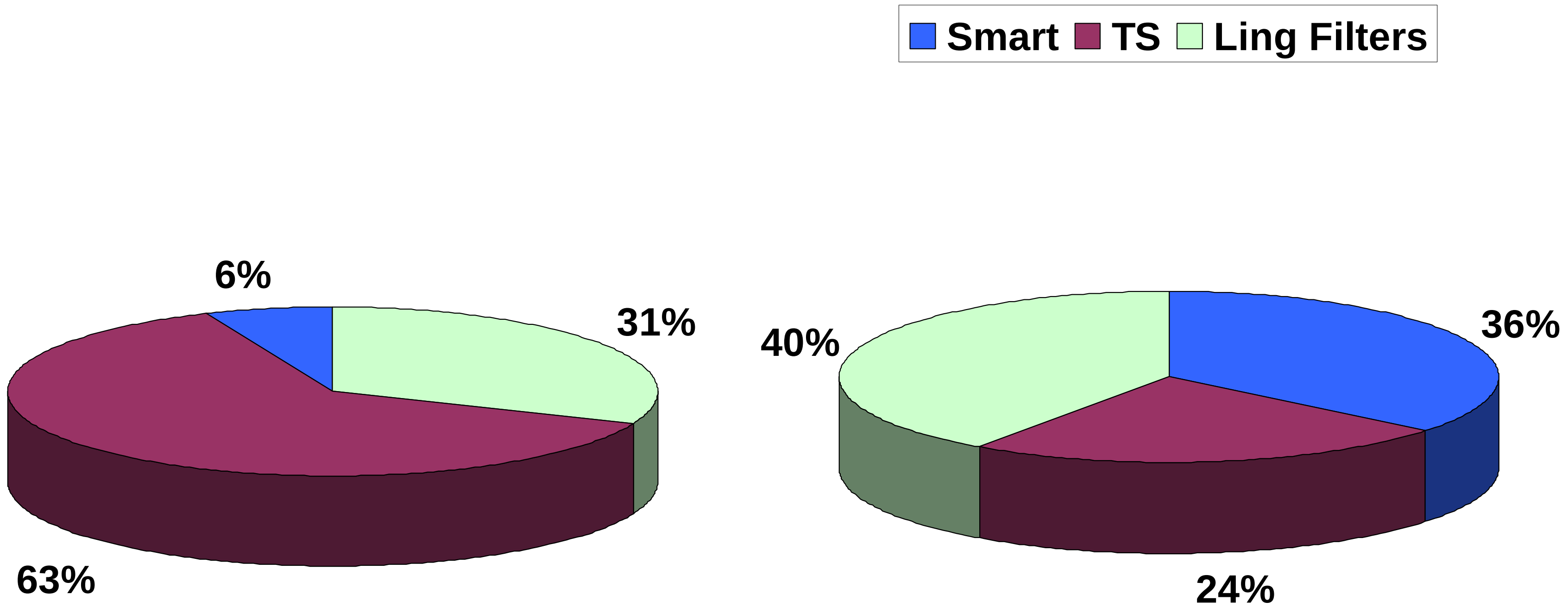
Evaluation

	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
TS with Syntactic Ordering + NPs + Semantic Type	22	1.32	91	1.82

MAR = Mean Answer Rank

- One NP per guess: 65 test questions answered correctly

Sources of Error



Development questions
(38)

Test questions
(200)

Related Work

- Some previous research that addressed QA scenarios
  - Story understanding (Lehnert, 1978)
  - Frequently-asked questions (Burke *et al*., 1995)

- TREC QA systems: combination of IR and shallow NLP
  - Cymphony (Srihari and Li, 1999)
  - AnSel (Prager *et al.*, 1999)
  - Qanda (Breck *et al.*, 1999)
  - SyncMatcher (Oard *et al.*, 1999)
  - LASSO (Moldovan *et al.*, 1999)
  - …

Conclusion

- Investigated uses of a handful of linguistic and statistical knowledge sources for question answering
  - Word co-occurrence information
  - Syntactic information
  - Semantic information

- Weak linguistic knowledge can offer substantial improvements over purely IR-based techniques