What I learned about Semantics from Textual Question-Answering

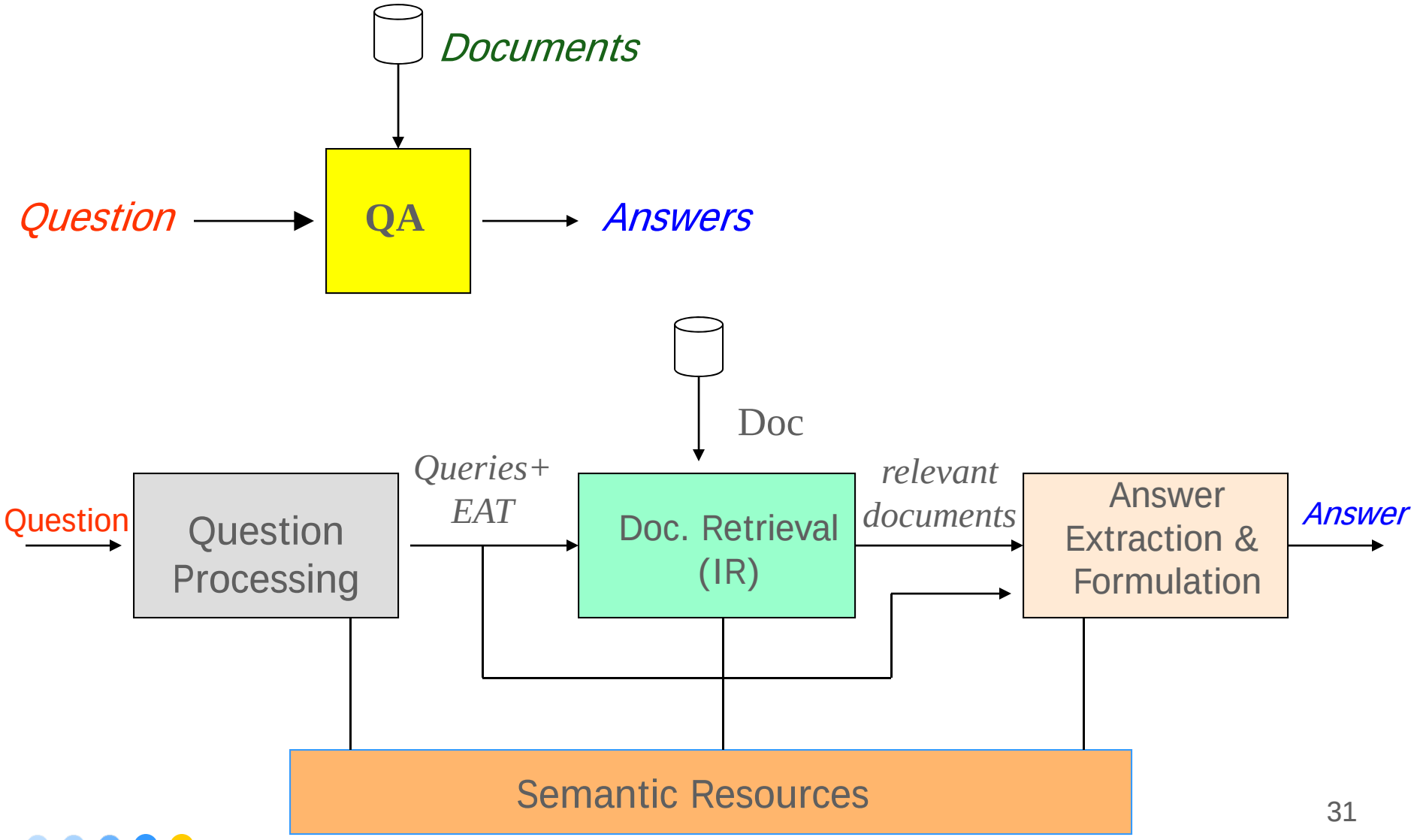
Sanda Harabagiu

University of Texas at Dallas

Outline

1. *The need for Semantic Processing in Textual QA*
   - Detecting the Expected Answer Type
   - Parsing with Predicate Argument Structures
   - Parsing with Semantic Frames
2. *Semantic Decomposition for Complex Questions*
   - A Markov Chain for Semantic Decomposition
   - Incorporating Knowledge into the Random Walk
   - Select the Most Relevant Question Decompositions
3. *Using Semantic Inference for Answering Complex Questions*
   1. Method 1: Using Textual Entailment to Filter Candidate Answers
   2. Method 2: Using Textual Entailment to Rank Candidate Passages
   3. Method 3: Using Textual Entailment to Select Question Decompositions
4. *Conclusions and the Future*

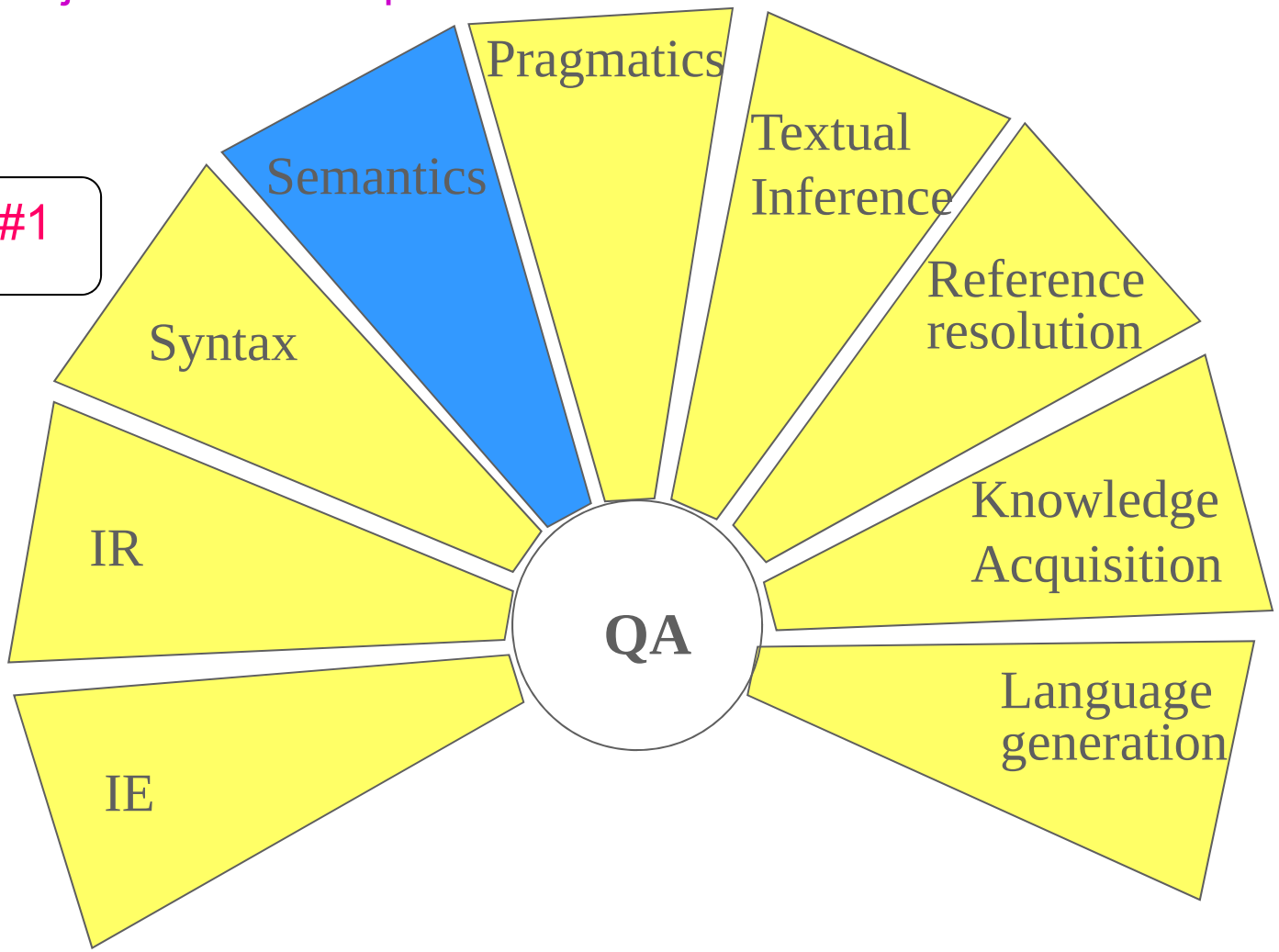
Textual Question Answering



Problem complexity

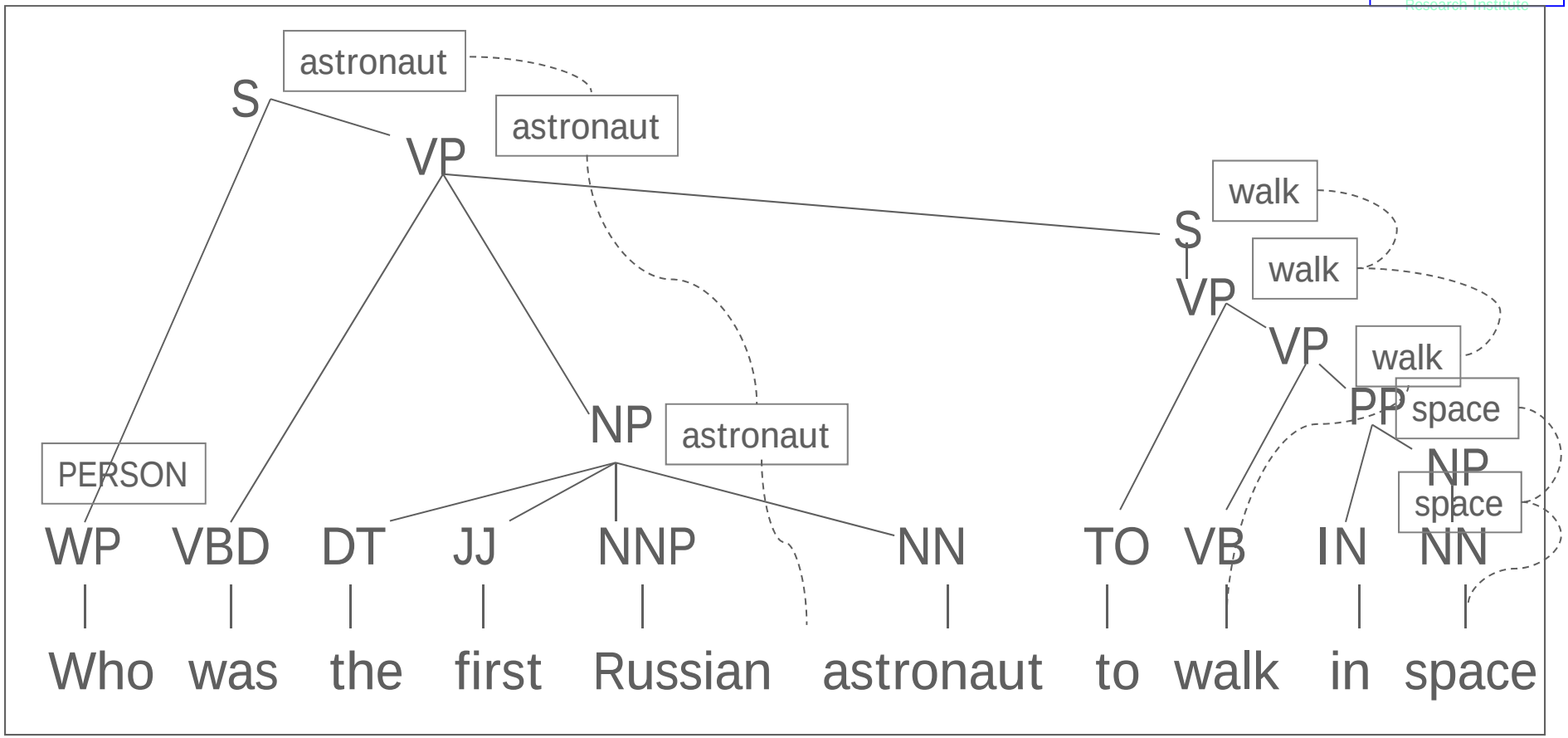
Semantics is just one of the problems! But what kind of semantics do we need?

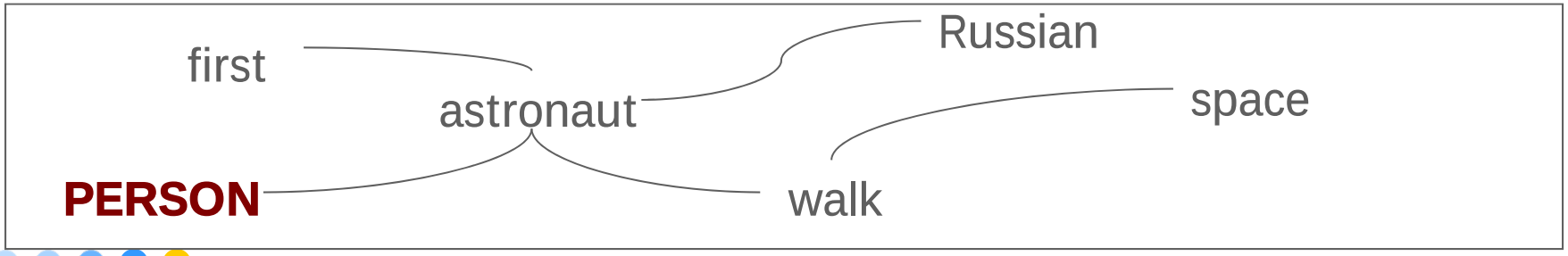
Lesson #1



An Example

Who was the first Russian astronaut to walk in space?

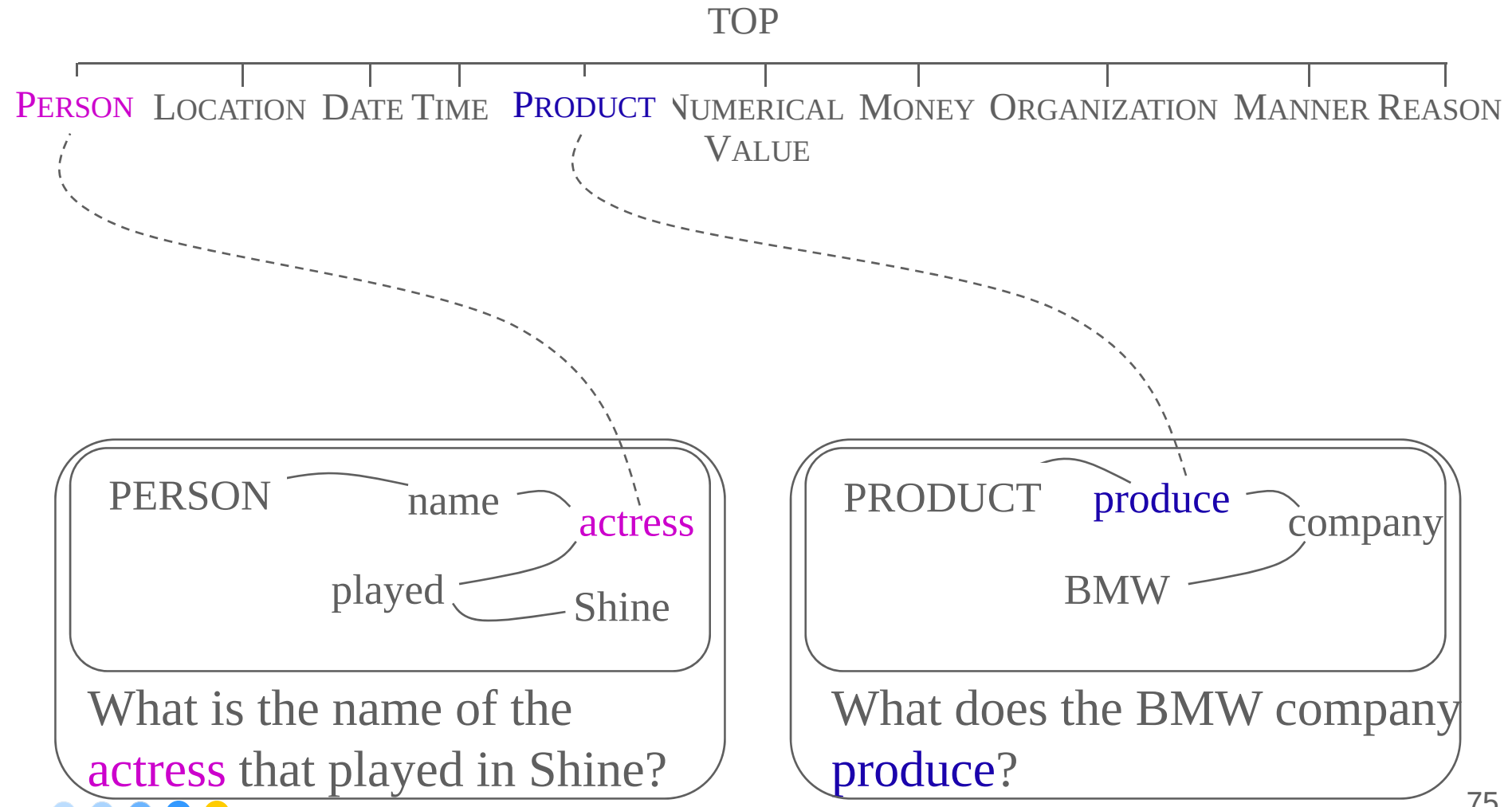




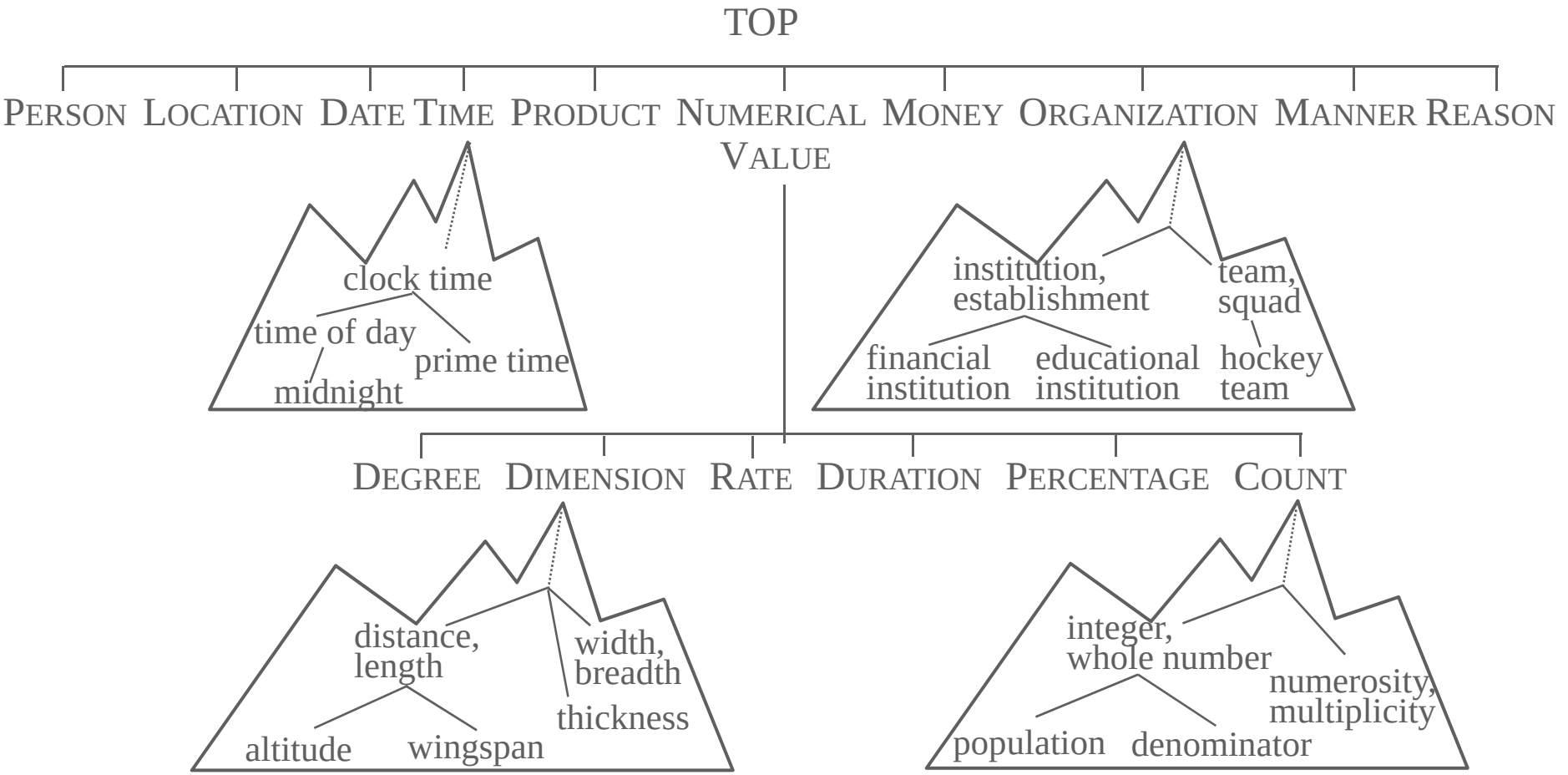
Detecting the Answer Type

1. Determine the category(ies) of the question stem
2. Select answer type nodes {A} having the same category as the question stem
3. Select node N that
   (a) is connected to the question stem
   (b) has highest connectivity in the semantic representation
4. Search for the word in node N along Answer hierarchies
5. Return the answer type as the top of the hierarchy found when N was located

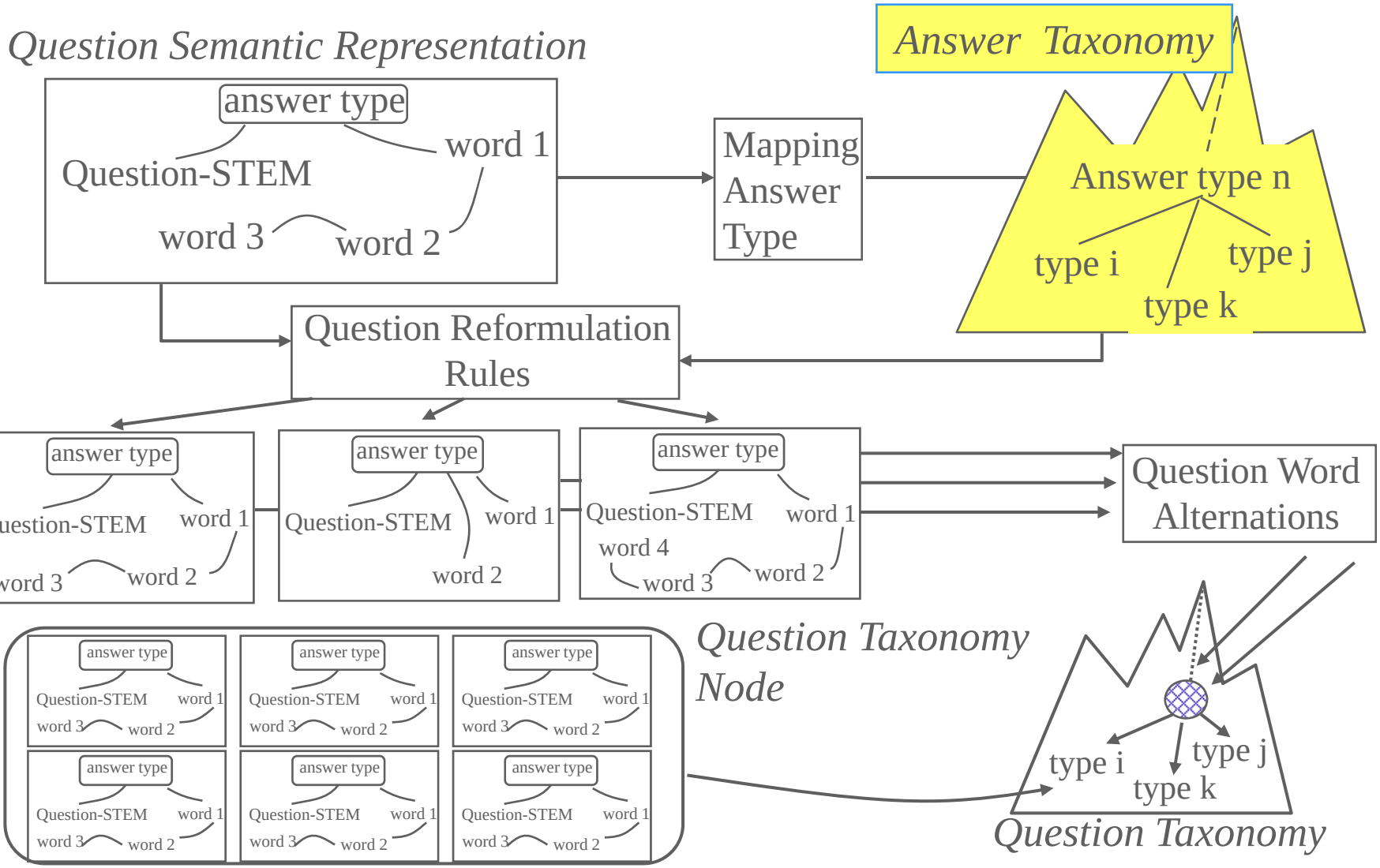
Examples


TOP
PERSON
LOCATION
DATE TIME
PRODUCT
NUMERICAL VALUE
MONEY
ORGANIZATION
MANNER
REASON
PERSON
name
actress
played
Shine
What is the name of the actress that played in Shine?
PRODUCT
produce
company
BMW
What does the BMW company produce?

Possible Answer Types



Expected Answer Taxonomy



The role of the semantic information encoded in Expected Answer Hierarchies

Human Language Technology Research Institute

Lesson #2

Semantic information encoded in EATs

plays and important role in QA

Distribution of Errors in the modules of a QA system

Module	Module definition	Errors (%)
M1	Keyword pre-processing (split/bind/spell check)	1.9
M2	Construction of internal question representation	5.2
M3	Derivation of expected answer type	36.4
M4	Keyword selection (incorrectly added or excluded)	8.9
M5	Keyword expansion desirable, but missing	25.7
M6	Actual retrieval (limit on passage number or size)	1.6
M7	Passage post-filtering (incorrectly discarded)	1.6
M8	Identification of candidate answers	8.0
M9	Answer ranking	6.3
M10	Answer formulation	4.4

Semantic Problems with EATs

- The problem of assigning EATs
  - E.g. “manner questions”:
  - Example *“How did Hitler die?”*
- The problem of recognizing answer types/structures
  - Should *“manner of death”* be considered an answer type?
  - What other manner of event/action should be considered as answer types?
- The problem of recognizing EATs in texts
  - Should we learn to extract “manner” relations?
  - What other types of relations should we consider?
  - Is relation recognition sufficient for answering all types of questions? Is it necessary?

EAT = Manner-of-death

In TREC evaluations several questions asked about manner of death:

- *"How did Adolf Hitler die?"*

- Solution:
  - *We considered "Manner-of-Death" as an answer type, pointing to a variety of verbs and nominalizations encoded in WordNet*
  - We developed text mining techniques for identifying such information based on lexico-semantic patterns from WordNet
  - Example:
    - [*kill #sense1 (verb) – CAUSE → die #sense1 (verb)*]
  - Source of the troponyms of the [*kill #sense1 (verb)*] concept are candidates for the MANNER-OF-DEATH hierarchy
    - e.g., drown, poison, strangle, assassinate, shoot

Practical Hurdle

Lesson #3

Semantic information for EATs needs to be recognized by text mining techniques

- Not all MANNER-OF-DEATH concepts are lexicalized as verbs
  → we set out to determine additional patterns that capture such cases
- *Goal:* (1) set of patterns
  (2) dictionaries corresponding to such patterns
  → well known IE technique: (IJCAI'99, Riloff&Jones)

X { DIE / be killed } in ACCIDENT	seed: train, accident, car wreck (ACCIDENT)
X { DIE / be killed } {from|of} DISEASE	seed: cancer (DISEASE) AIDS
X { DIE } { after suffering / suffering of } MEDICAL CONDITION	seed: stroke, complications caused by diabetes (ACCIDENT)

- *Results: more than100 patterns were discovered*

How much information does the text wear on its sleeve?

By doing only:

- named entity recognition
- semantic classification of the expected answer type (off-line taxonomy + semantic info from WordNet)
- Text mining

→ 55% accuracy on factual trivia-like questions (*TREC-8, Moldovan et al.)*

What else is needed?

→ *Most questions cannot be processed successfully in this way, as they do not have a simple, conceptual EAT*

We need to consider additional forms of semantic knowledge and semantic processing.

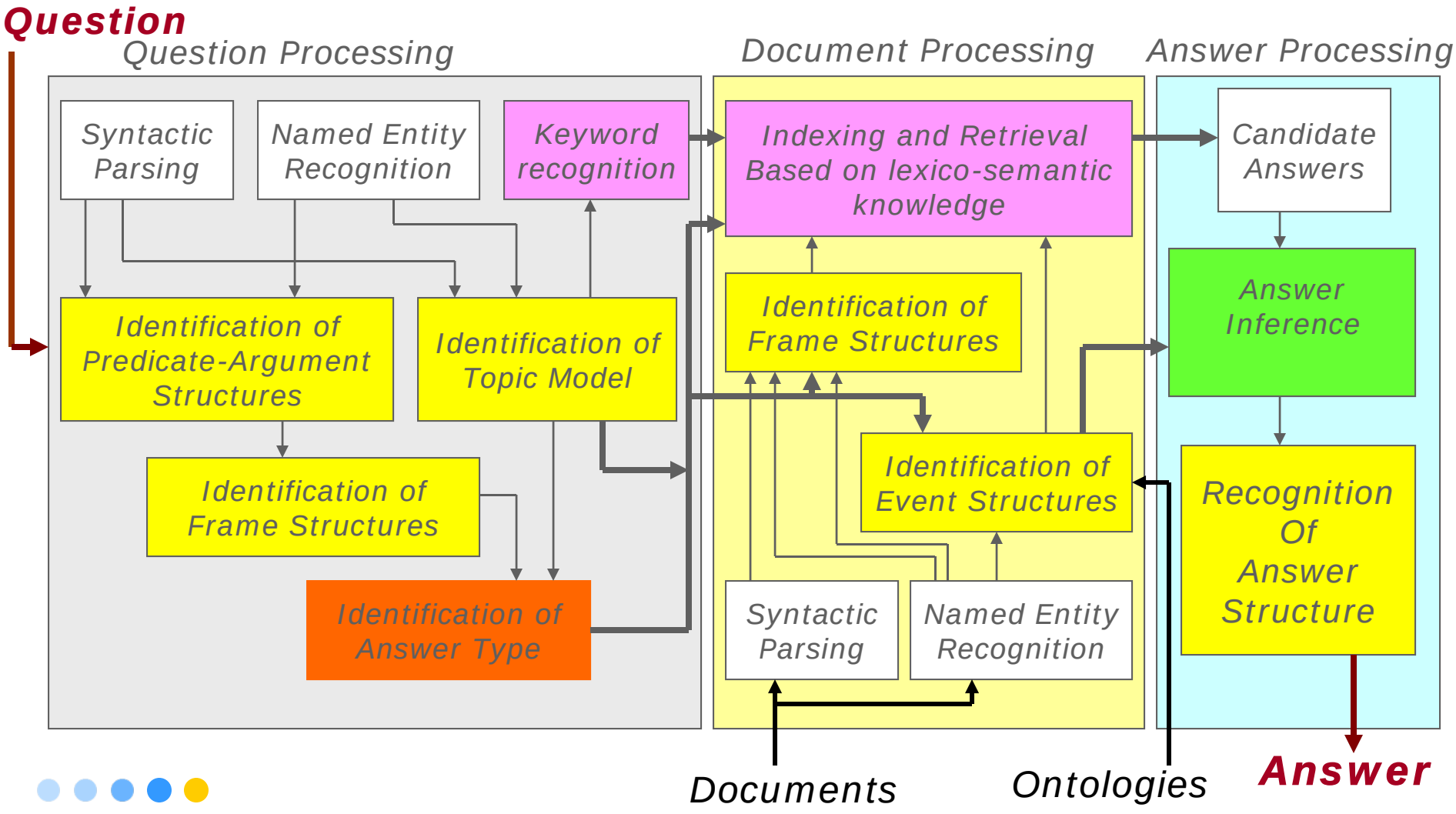
Lesson #4

AQUINAS – *Answering QUestions using INference and Advanced Semantics*

Collaborative Research Project
between UTD,
ICSI Berkeley and Stanford University

The driving rationale for our approach is that humans appear to have limited need for factoid question answering, but rather much more need to have systems that can deal with complex reasoning about causes, effects and chains of hypotheses.

QA architecture based on semantic structures



Semantic structures discovered by shallow semantic parsing

- *Parsers that use shallow semantics encoded as either predicate-argument structures or semantic frames*

- *In the past 8 years, several models for training such parsers have emerged*
- *Lexico-Semantic resources are available (e.g PropBank, FrameNet, TimeBank)*
- *Several evaluations measure the performance of such parsers (e.g. SENSEVAL, CoNNL, Semeval)*

Predicate-argument structures for questions

Lesson #5

Predicate-arguments structures improve the detection of EATs!!!

- *Example 1: From which country did North Korea import its missile launch pad metals?*

```
Predicate: import
Argument 0: (role = importer): North Korea (COUNTRY)
Argument 1: (role = commodity): missile launch pad metals
Argument 2 (role = exporter): ANSWER (COUNTRY)
```

- *Example 2: What stimulated India's missile programs?*

```
Predicate: stimulate
Argument 0: (role = agent): ANSWER (part 1)
Argument 1: (role = thing increasing): India's missile programs
Argument 2 (role = instrument): ANSWER (part 2)
```

Applying Predicate-Argument Structures to QA

Predicate-arguments structures improve answer extraction!!!

Lesson #6

- Parsing Questions

***Q**: What kind of materials were stolen from the Russian navy?*

PAS(Q): What [Arg1: kind of nuclear materials] were [Predicate:stolen] [Arg2: from the Russian Navy]?

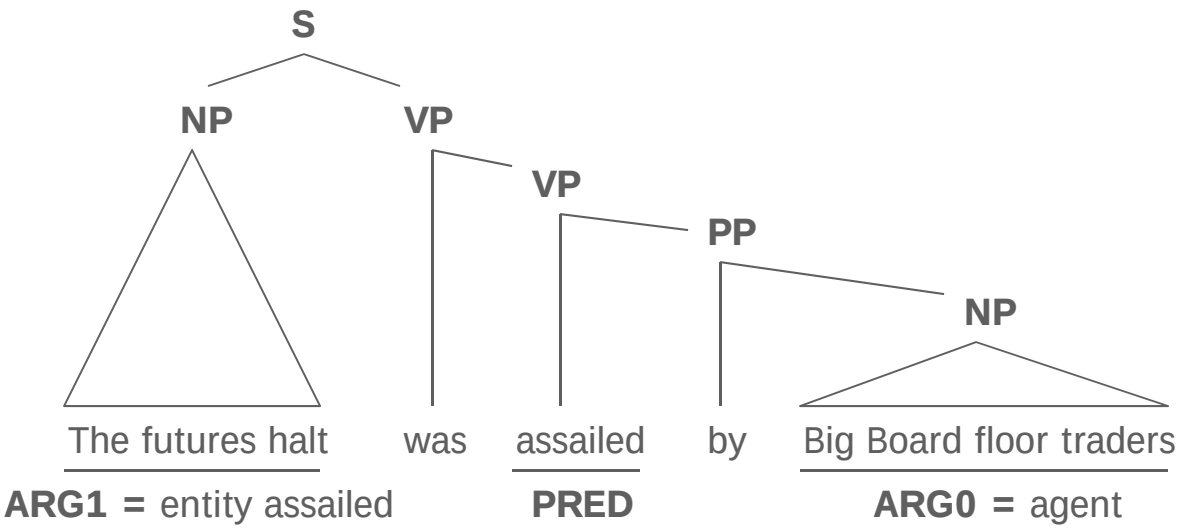
- Parsing Answers

***A(Q)**: Russia's Pacific Fleet has also fallen prey to nuclear theft; in 1/96, approximately 7 kg of HEU was reportedly stolen from a naval base in Sovetskaya Gavan.*

PAS(A(Q)): [Arg1(P1redicate 1): Russia's Pacific Fleet] has [ArgM-Dis(Predicate 1) also] [Predicate 1: fallen] [Arg1(Predicate 1): prey to nuclear theft]; [ArgM-TMP(Predicate 2): in 1/96], [Arg1(Predicate 2): approximately 7 kg of HEU] was [ArgM-ADV(Predicate 2) reportedly] [Predicate 2: stolen] [Arg2(Predicate 2): from a naval base] [Arg3(Predicate 2): in Sovetskawa Gavan]

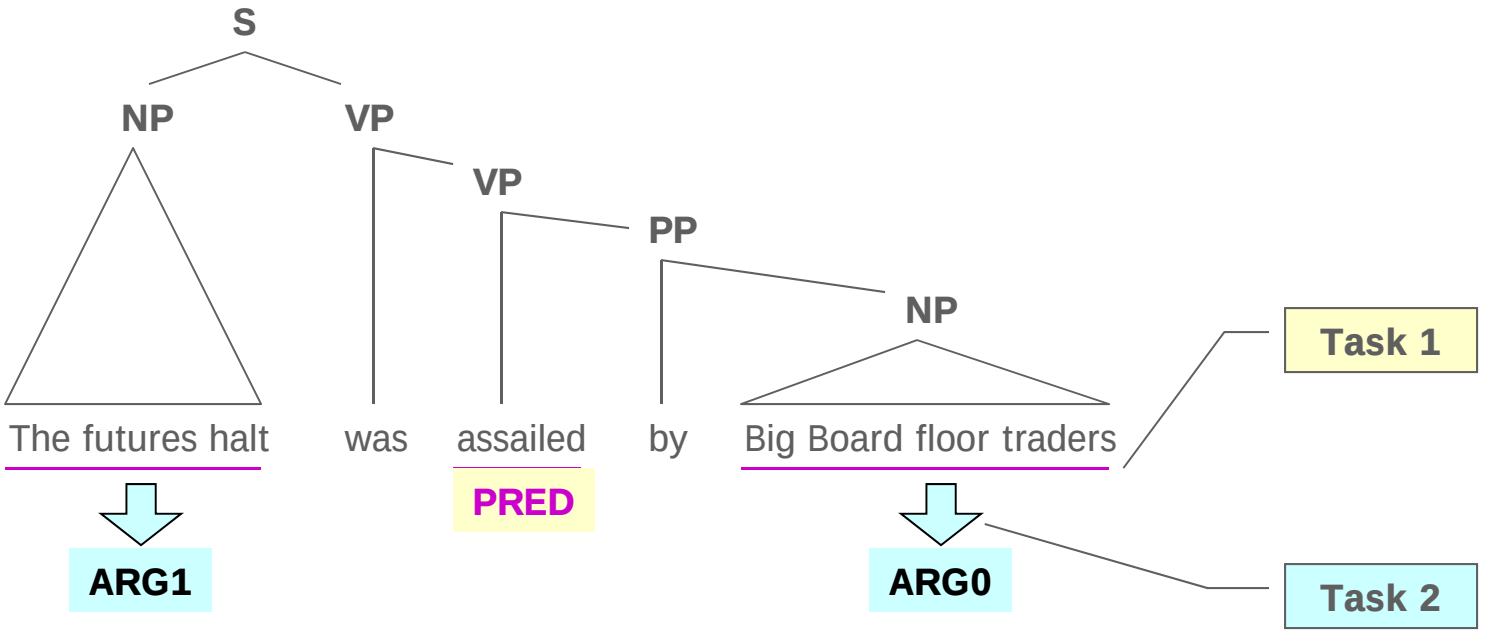
- Result: exact answer= *"approximately 7 kg of HEU"*

Proposition Bank Overview



- *A one million word corpus annotated with predicate argument structures [Kingsbury, 2002]. Currently only predicates lexicalized by verbs.*
- Numbered arguments from 0 to 5. Typically ARG0 = *agent*, ARG1 = *direct object* or *theme*, ARG2 = *indirect object*, *benefactive*, or *instrument*.
- Functional tags: ARMG-LOC = *locative*, ARGM-TMP = *temporal*, ARGM-DIR = *direction*.

The Model



- Consists of two tasks: (1) identifying parse tree constituents corresponding to predicate arguments, and (2) assigning a role to each argument constituent.
- Both tasks casts as classification problems

Additional semantic resources: FrameNet

- Using FrameNet for QA
- *Example: What stimulated India's missile programs?*

FRAME: Stimulate

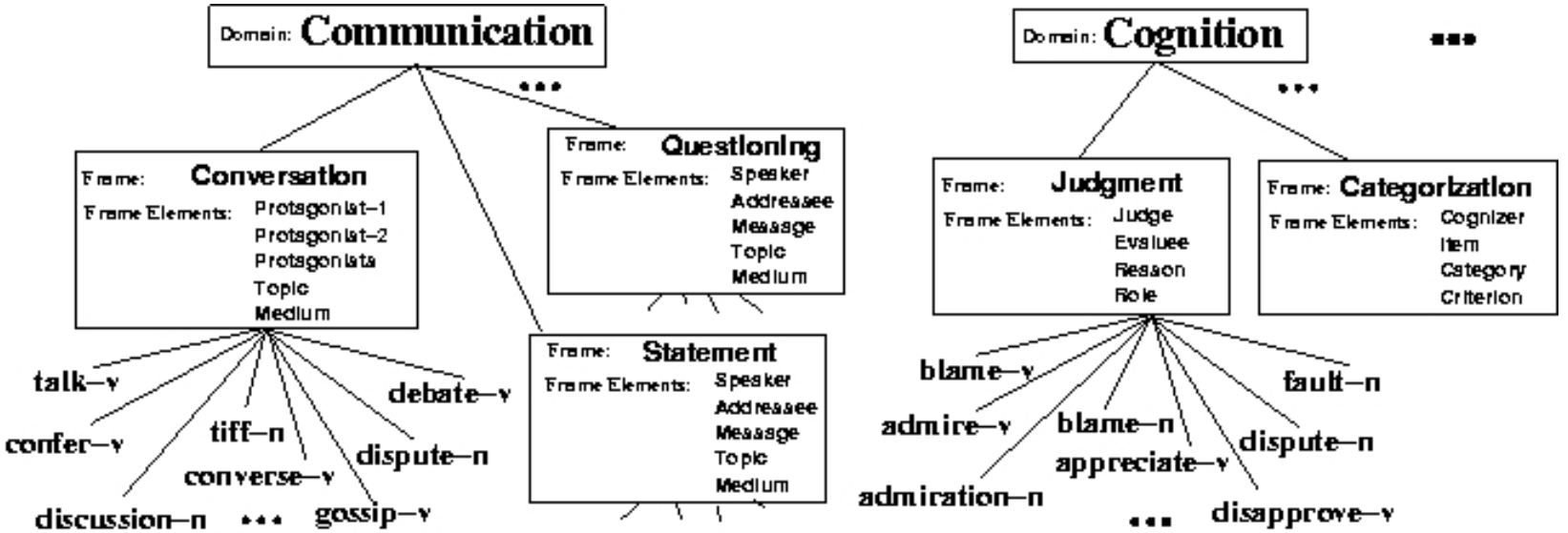
Frame Element CIRCUMSTANCES: ANSWER (part 1)
Frame Element EXPERIENCER: India's missile program
Frame Element STIMULUS: ANSWER (part 2)

FRAME: Subject_Stimulus

Frame Element CIRCUMSTANCES: ANSWER (part 3)
Frame Element COMPARISON SET: ANSWER (part 4)
Frame Element EXPERIENCER: India's missile program
Frame Element PARAMETER: nuclear proliferation

FrameNet

- FrameNet provides database of words with associated semantic roles



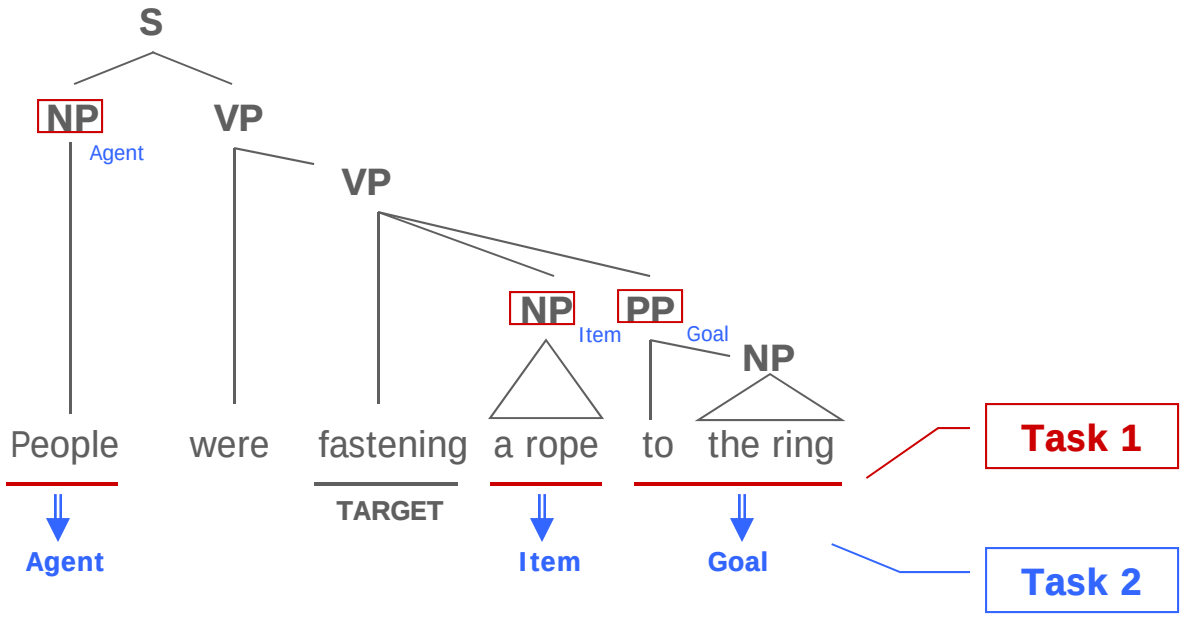
FrameNet Overview

- Produced by ICSI Berkeley [Baker et al.,1998] as a lexico-semantic resource encoding a set of frames (schematic representations of situations)

- Frames are characterized by:
  - target words or lexical predicates whose meaning includes aspects of the frame;
  - frame elements (FEs) which represent the semantic roles of the frame;
  - examples of annotations performed on the British National Corpus for instances of each target word.

- The project methodology was done on a *frame-by-frame basis*:
  - choosing a semantic frame (e.g. Commerce)
  - define the frame and its frame elements (e.g. BUYER, GOODS, SELLER, MONEY)
  - list the various lexical predicates which invoke the frame (buy, sell)
  - finding example sentences of each predicate in a corpus

The Model for Recognizing Frame Structures

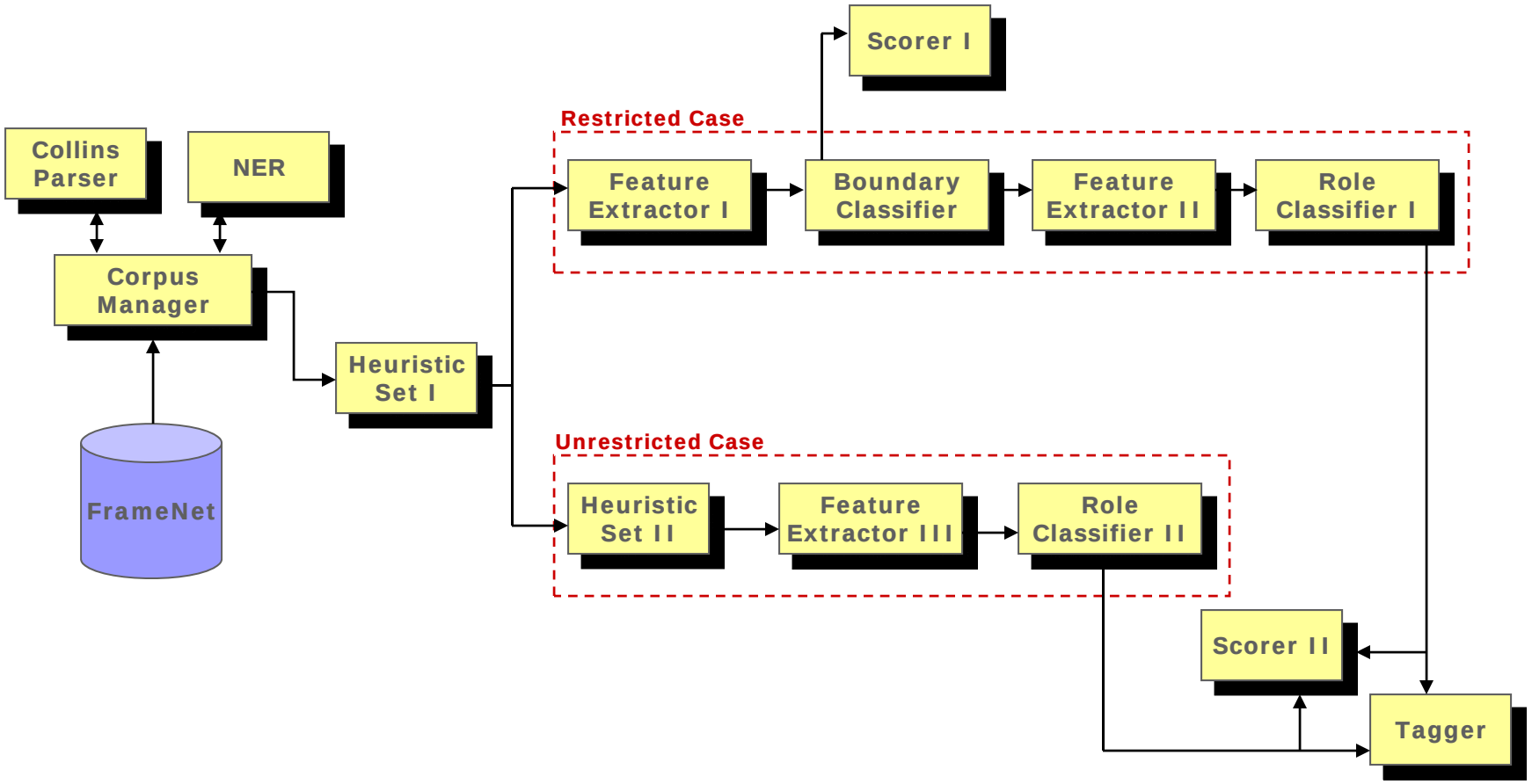


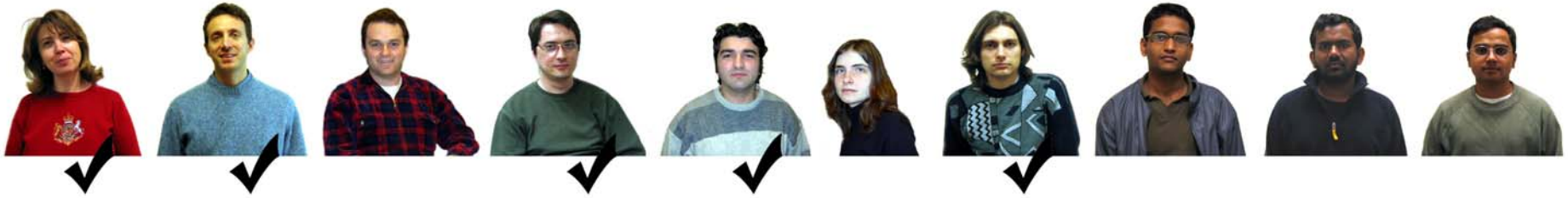
- The basic model consists of two tasks, which can be reformulated as classification problems:
  - Task 1: identifying parse tree constituents corresponding to frame elements.
  - Task 2: assigning a role to each frame element constituent.
- Evaluation of the results of FrameNet-based parsers consider two different cases of automatic labeling of the semantic roles were considered:
  - Unrestricted Case requires systems to assign FE labels to the test sentences for which the boundaries of each frame element were given and the target words identified
  - Restricted Case requires systems to:
    - recognize the boundaries of the FEs for each evaluated frame and
    - assign a label to it.

Architecture for FrameNet-Based Semantic Parsing

Lesson #7

More sophisticated semantic structures can be recognized by more complex systems!!!





Shallow Semantic Parsing Based on FrameNet

Cosmin Adrian Bejan, Alessandro Moschitti, Paul Morărescu, Gabriel Nicolae and Sanda Harabagiu

University of Texas at Dallas

Human Language Technology Research Institute

Shallow semantic parsing for QA

- *Several evaluations measured the performance of shallow semantic parsers (e.g. SENSEVAL, SemEval, CoNNL)*
  - *None of the evaluations used questions or answers !!!*
  - *We created the UTD AnswerBank annotations by considering hand-corrected parses of shallow parsers trained on PropBank and FrameNet*

Lesson #8

Evaluate semantic parsers on

QA data before incorporating the

Semantic parsers in QA systems

AnswerBank

- AnswerBank is a collection of over a 2000 QA annotations from the AQUAINT CNS corpus.

- Questions and answers cover several different domains of the CNS data.

- CNS Domains Covered were
  - WMD related (54%)
  - Nuclear Theft (25%)
  - India's missile program (21%)

- Questions and answers are POS tagged, and syntactically parsed.

- Question and Answer predicates are annotated with PropBank arguments and FrameNet tags.

- Additional semantic information was annotated, including temporal, aspectual information (TIMEML+) and information about event relations and figurative uses of language.

Semantic Parsing Results on AnswerBank

- *Identification of Predicate-Argument Structures*

Corpus	P(Arg)	R(Arg)
PropBank	85.4%	85.6%
AnswerBank-0	54.3%	64.2%
AnswerBank-1	69.1%	73.5%
Corpus	P(Role)	R(Role)
PropBank	88.5%	92.7%
AnswerBank-0	59.2%	66.8%
AnswerBank-1	73.8%	65.2%

Lesson #9

I

- *Identification of Frame Structures*

Need for application-driven semantic parsing and annotations!

Corpus	P(FE)	R(FE)
FrameNet	75.2%	77%
AnswerBank-0	52.5%	56.7%
AnswerBank-1	73.5%	74.2%
Corpus	P(Role)	R(Role)
FrameNet	91.57%	89.13%
AnswerBank-0	68.2%	66.7%
AnswerBank-1	80.2%	78.5%

Processing Complex Questions

Discuss measures that schools and school districts have taken to prevent violent occurrences and shootings, such as those in Littleton, Colorado and Jonesboro, Arkansas.

Keyword Extraction/ Alternation

Extracted: measures, school, district, "school district", prevent, violent, occurrence, shooting, Littleton, Jonesboro, "Littleton, Colorado", "Jonesboro, Arkansas"

Alternations: **measures**, steps, actions, initiatives, **school**, "high school", "elementary school", "middle school", **district**, region, departments, **prevent**, stop, block, **violent**, aggressive, **occurrence**, incident, episode, **shooting**, attack, murder, massacre, rampage, terrorist event

Syntactic Question Decomposition

What measures have **schools** taken to prevent **violent occurrences**?

What measures have **school districts** taken to prevent **violent occurrences**?

What measures have **schools** taken to prevent **shootings**?

What measures have **school districts** taken to prevent **shootings**?

Semantic Question Decomposition

What U.S. schools have had to perform a security lockdown?

Where were drills conducted in which SWAT teams entered schools to put down school shootings?

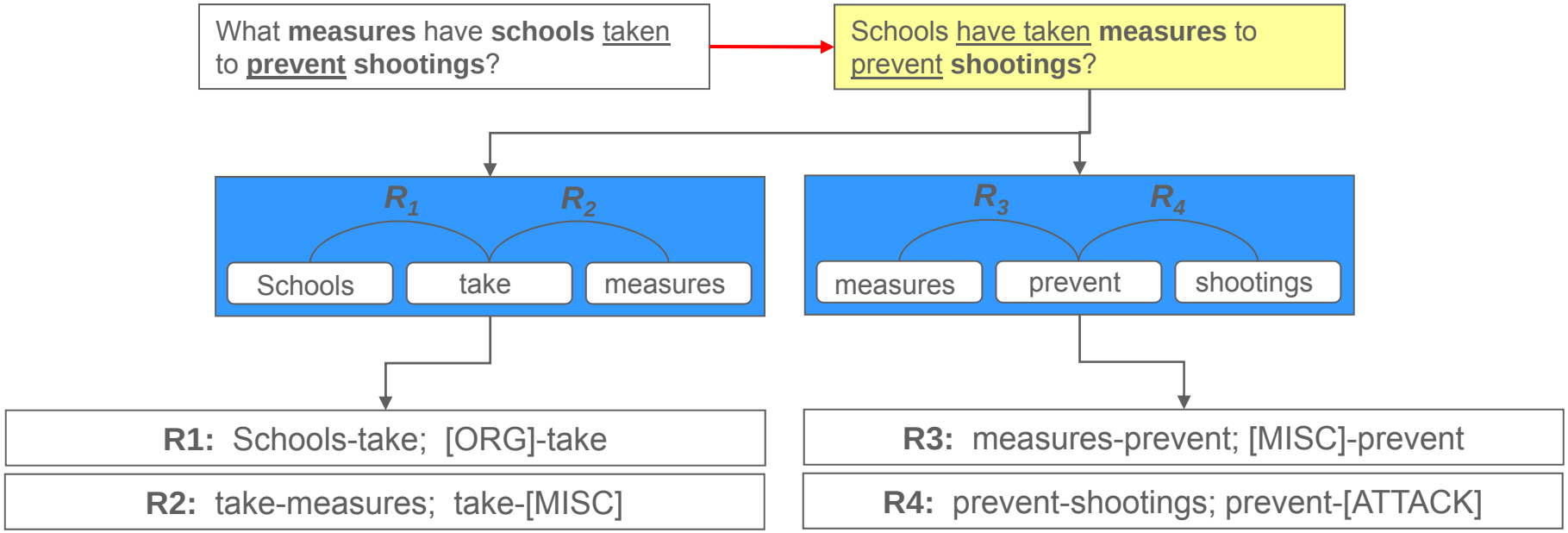
Which schools now use metal detectors at special events?

Which Columbine HS student killed twelve other students?

Which district faced legal action for not preventing violence at its schools?

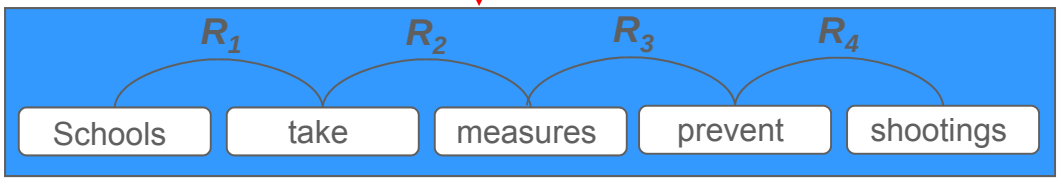
Semantic Question Decomposition

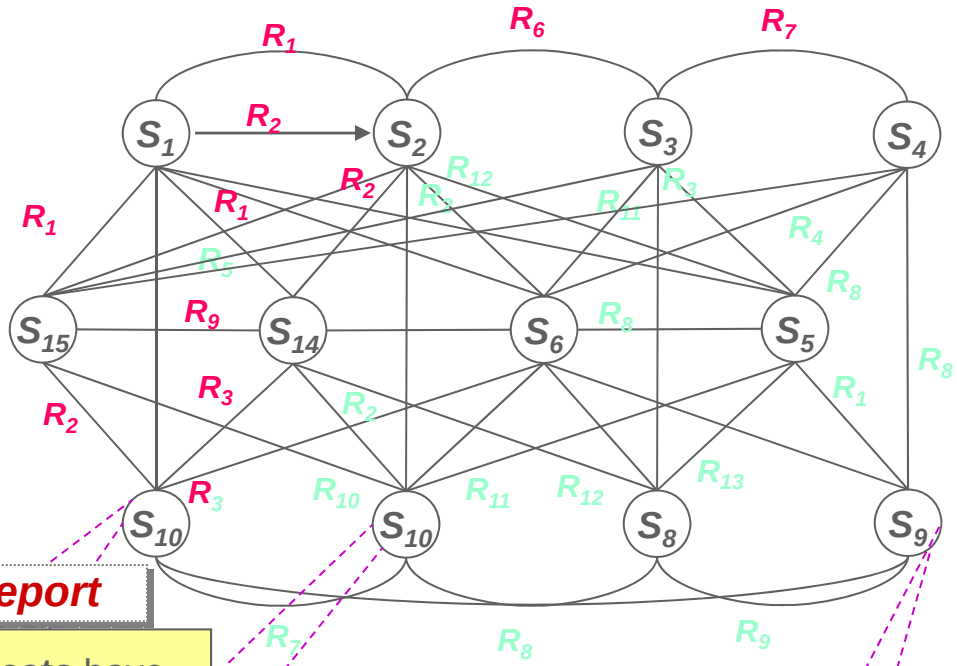
- **Step 1.** The complex question is lexically, syntactically, and semantically analyzed.



Step 2. Create Relational Index (populate a network)

What **measures** have **schools** taken to **prevent** **shootings**?





R_8: *threat-report*

Reports of **threats** have increased in schools...

Bomb **threats** were **reported** in L.A. schools...

Officials have decided to follow up on all new **reports** of **threats** to area schools...

R_1. Safety **measures** **taken** by schools can cost as much as $20 million dollars per year.

R_1. Schools have been quick to **take** security **measures**, citing increased *threats* of *violence.*

R_6. 13 *students* were *suspended* after **threatening violence** against teachers.

R_7. Klebold was **suspended** with two **students** for hacking into the school's computer on the same day that *threats* were *reported*.

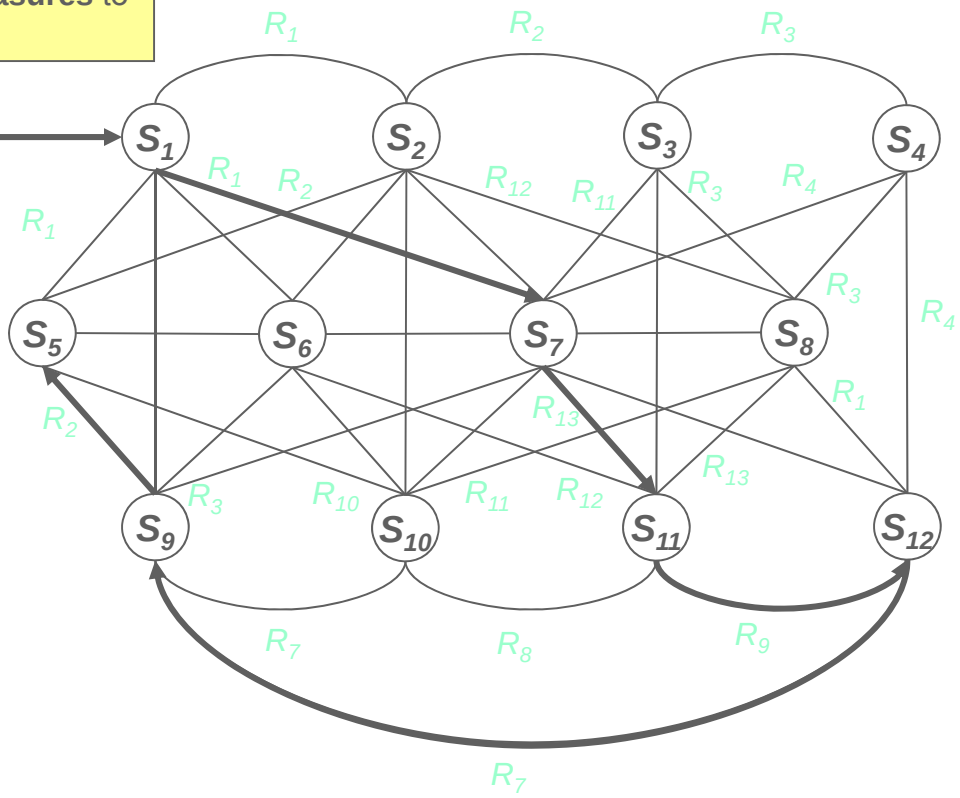
R_8. Reports of **threats** have increased at many U.S. schools, leading some schools to *use metal detectors* at special events.

R_9. Westside HS was one of the first schools to **use** **metal detectors** on a full-time basis following Columbine.

Step 3. Perform a Random Walk in order to generate decompositions

- Uses a Markov Chain which alternates between:
  - selecting a new relation, and
  - generating a new question decomposition.
- The decision to continue / stop the random walk on a bipartite graph of questions and relations is inspired by (Lafferty and Zhai 2001)

Schools have taken **measures** to prevent **shootings**?



R1. Safety **measures** **taken** by schools can cost as much as $20 million dollars per year.

R1. Schools have been quick to **take** security **measures**, citing increased *threats* of *violence.*

R13. 13 *students* were *suspended* after **threatening violence** against teachers.

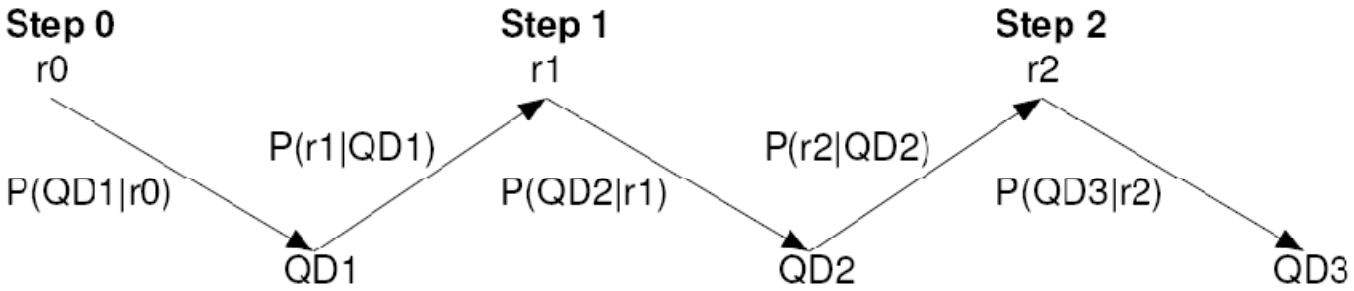
R9. Klebold was **suspended** with two **students** for hacking into the school's computer on the same day that *threats* were *reported*.

R7. **Reports** of **threats** have increased at many U.S. schools, leading some schools to *use metal detectors* at special events.

R2. Westside HS was one of the first schools to **use** **metal detectors** on a full-time basis following Columbine.

Markov Chains for Question Decomposition

- The initial state of the Markov Chain for question decomposition is a relation selected randomly from the list RELATIONS (*time* = 0)



- After selecting a relation r_i at step *i,* the index is consulted to find sentences where r_i is present. We also discover *new relations* r_j that belong to the same sentence.

Step 4. Question Reformulation

Syntactic rewrite rules are then used to transform sentences discovered during the random walk back into questions.

R1. Safety **measures** **taken** by schools can cost as much as $20 million dollars per year.

How much can safety measures taken by schools cost?

R1. Schools have been quick to **take** security **measures**, citing increased *threats* of *violence.*

Who has been quick to take security measures?

R13. 13 *students* were *suspended* after **threatening violence** against teachers.

Who were suspended after threatening violence?

R9. Klebold was **suspended** with two **students** for hacking into the school's computer on the same day that *threats* were *reported*.

Who was suspended for hacking into the school's computer?

R7. **Reports** of **threats** have increased at many U.S. schools, leading some schools to *use metal detectors* at special events.

What has led some schools to use metal detectors?

R2. Westside HS was one of the first schools to **use** **metal detectors** on a full-time basis following Columbine.

When did Westside HS use metal detectors on a full-time basis?

Example:

Q_0: What measures have schools and school districts taken to prevent violence and/or shootings, such as the ones in Littleton, Colorado?

R_1: *schools-take*
R_2: *take-measures*

S_1: Safety measures taken by schools can cost as much as $20 million dollars per year.

R_1: *schools-take*
R_2: *take-measures*

R_5: *cost-AMOUNT*

R_{10}: *install-THING*

Q_1: What safety measures have schools taken?

Q_2: How much can safety measures for schools cost?

Q_3: What have schools installed?

A_1: Almost 75% of U.S. high schools have taken some security measures in response to Littleton, including issuing of ID cards, installing metal detectors and security cameras, and hiring extra security personnel.

A_2: Following the Littleton school shooting, the Houston Independent School District installed security cameras in all of its high schools, a project which cost an estimated $30 million in total.

A_3: A spokesman for the school said that the installation of metal detectors at school entrances was an important step towards ensuring a secure learning environment.

Lesson #10

Translate the semantics of the question into the known semantics of several other, simpler questions.

Which question decompositions should be considered?

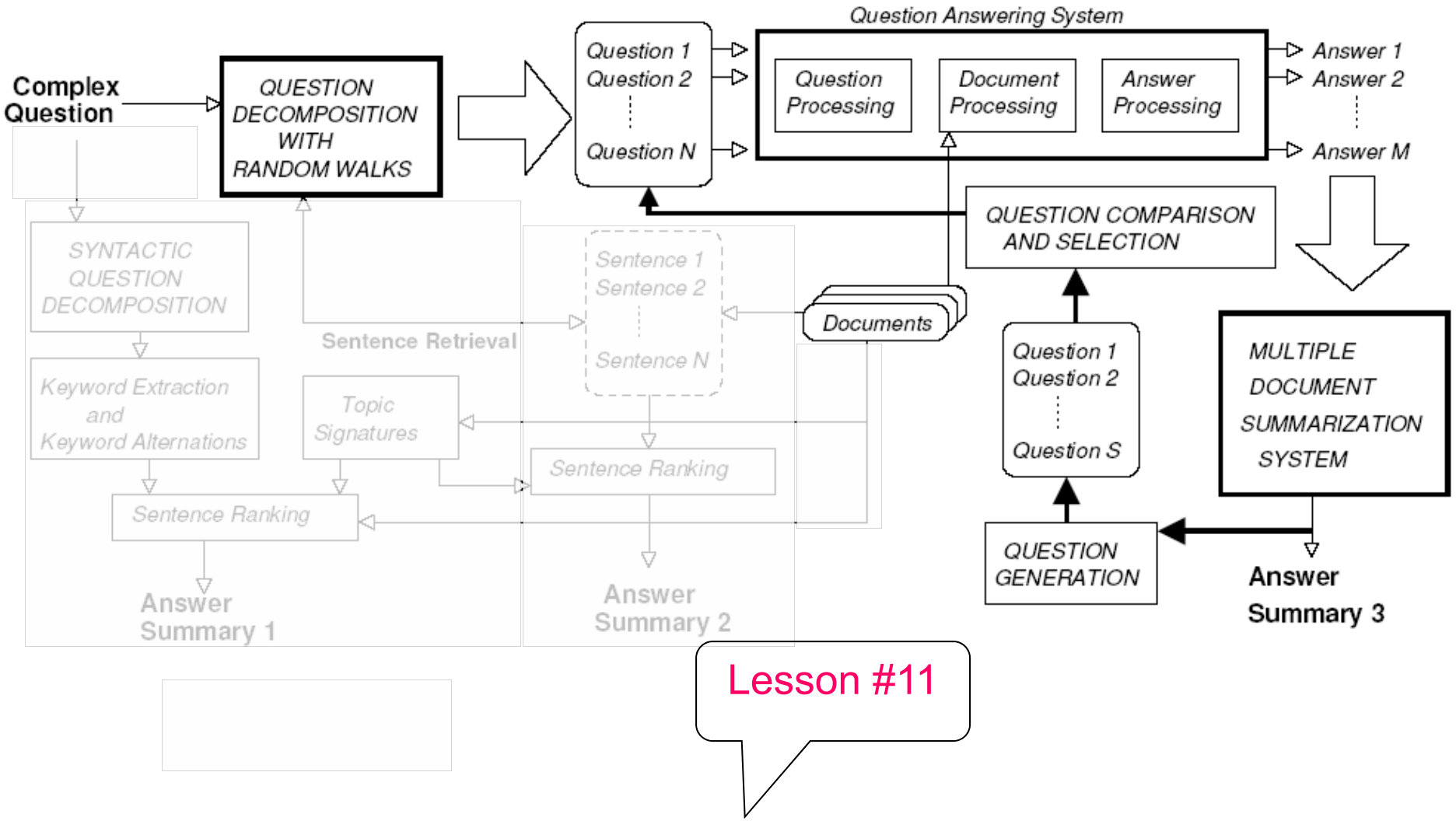
Random walks with Mixture models

- The idea (Otterbacher et al 2005):score each sentence against a complex question and select only the most relevant sentences.
- The sentence rank: produced by a mixture model that combines:
  1. An approximation of the sentence's relevance to a question
  2. Similarity measure to select sentences that are not similar to one another

Use the same idea for selecting question decompositions:

$$relev(QD_i, CQ) = d\frac{sim_a(QD_i, CQ)}{\sum_{QD_j} sim_a(QD_i, CQ)} + (1-d)\sum_{QD_j}\frac{sim_b(QD_i, CD_j)}{\sum_{QD_k} sim_b(QD_k, CD_j)}$$

Generating answers from question decompositions



Lesson #11

Semantic processing without semantic inference????

Role of Textual Entailment in Q/A

- Traditionally, work on the semantics of questions (Groenendijk 1999, Lewis 1998) has argued that the formal answerhood relation between a question and a set of correct answers can be cast in terms of **logical entailment**.

> A proposition *p* is considered to be an **answer** to a question *?q* iff *?q* logically entails the set of worlds in which *p* is true (i.e. *?p*).

Groenendijk (1999): *licensing*; Lewis (1998): *aboutness*

- For example: (taken from Shan and D. ten Cate (2002))
- S_a is an answer to S_q iff $?S_q \models ?S_a$:
  - $\boldsymbol{S_a}$. *Everyone is going to the party*. ($\forall x.Px$)
  - $\boldsymbol{S_q}$. *Who is going to the party?* (*?Px*)
  - $?Px \models ?\forall x.Px$
    - The worlds in which "Everyone is going to the party" are entailed by the partition of worlds denoted by *?Px*.
  - $\therefore$ S_a is an answer to S_q

Role of Textual Entailment in Q/A (1)

- While the notion of **textual entailment** has been defined far less rigorously than **logical entailment**, we believe that the recognition of textual entailment between a question and a set of candidate answers can enable Q/A systems to identify correct answers with greater precision than current techniques.

- For example, if we could accurately recognize textual entailment between a question and each candidate answer, we could learn to filter ***spurious*** answers:

Q. What did Peter Minuit buy for the equivalent of $24.00?

A1. Everyone knows that back in 1626, Peter Minuit bought **Manhattan** from the Indians for $24 worth of trinkets.

A2. In 1626, Minuit flagged down some passing locals, plied them with **beads, cloth and trinkets** worth $24 and walked away with the island.

Role of Textual Entailment in Q/A (2)

- ... Or to re-rank answers that were more (or less likely) to be entailments of the user's original question:

Q. What did Peter Minuit buy for the equivalent of $24.00?

⇧

A3. Peter Minuit, an enterprising Dutch businessman, purchased **Manhattan** while on a hunting trip from the Canarsee Indians for "diverse other wares" valued at 60 guilders, or about $25 U.S. dollars.

⇩

A4. Even though the Dutch East India Company maintained extensive records on their activities in New Amsterdam since 1600, there is no evidence that Minuit bought **Manhattan** for the equivalent of $24.00.

Role of Textual Entailment in Q/A (3)

- … Or to select question decompositions that best fulfilled the information need of the user's original question

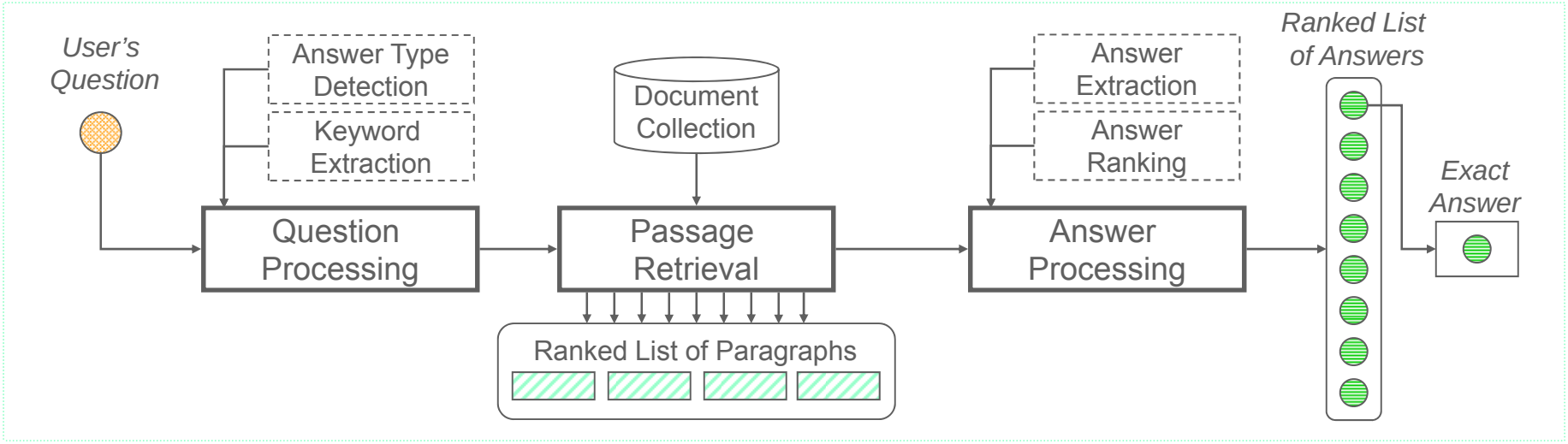
Q. How hot does the inside of an active volcano get?

A1. Tamagawa University volcano experts said lava fragments belched out of the mountain on Jan 31 were as hot as [**300 degrees Fahrenheit**]$_{Answer1}$. The intense heat from a second eruption on Tuesday forced the government's rescue operations to stop after [**90 minutes**]$_{Answer2}$.

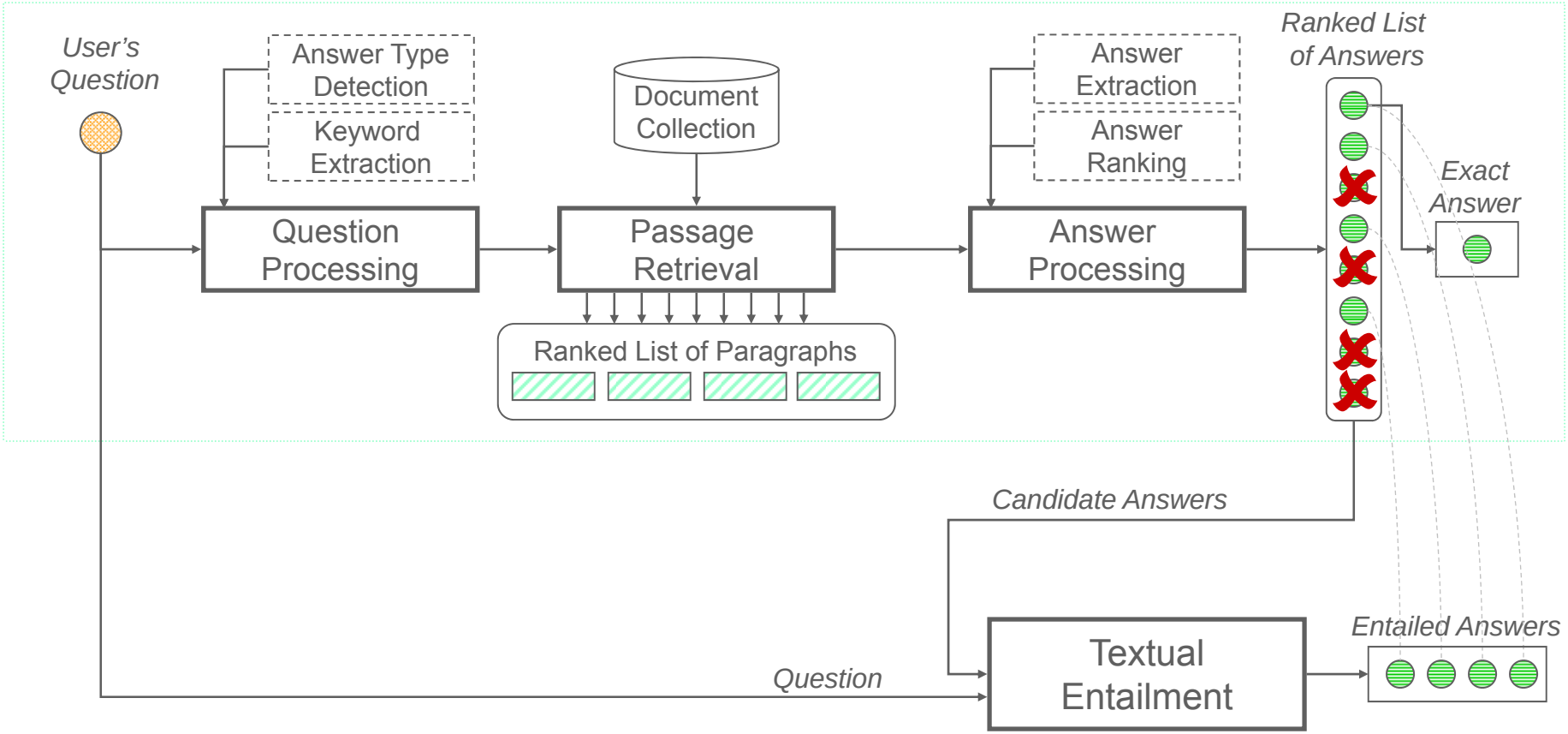
Q_1. What temperature were the lava fragments belched out of the mountain?

Q_2. How long did rescue operations last when the volcano erupted ?

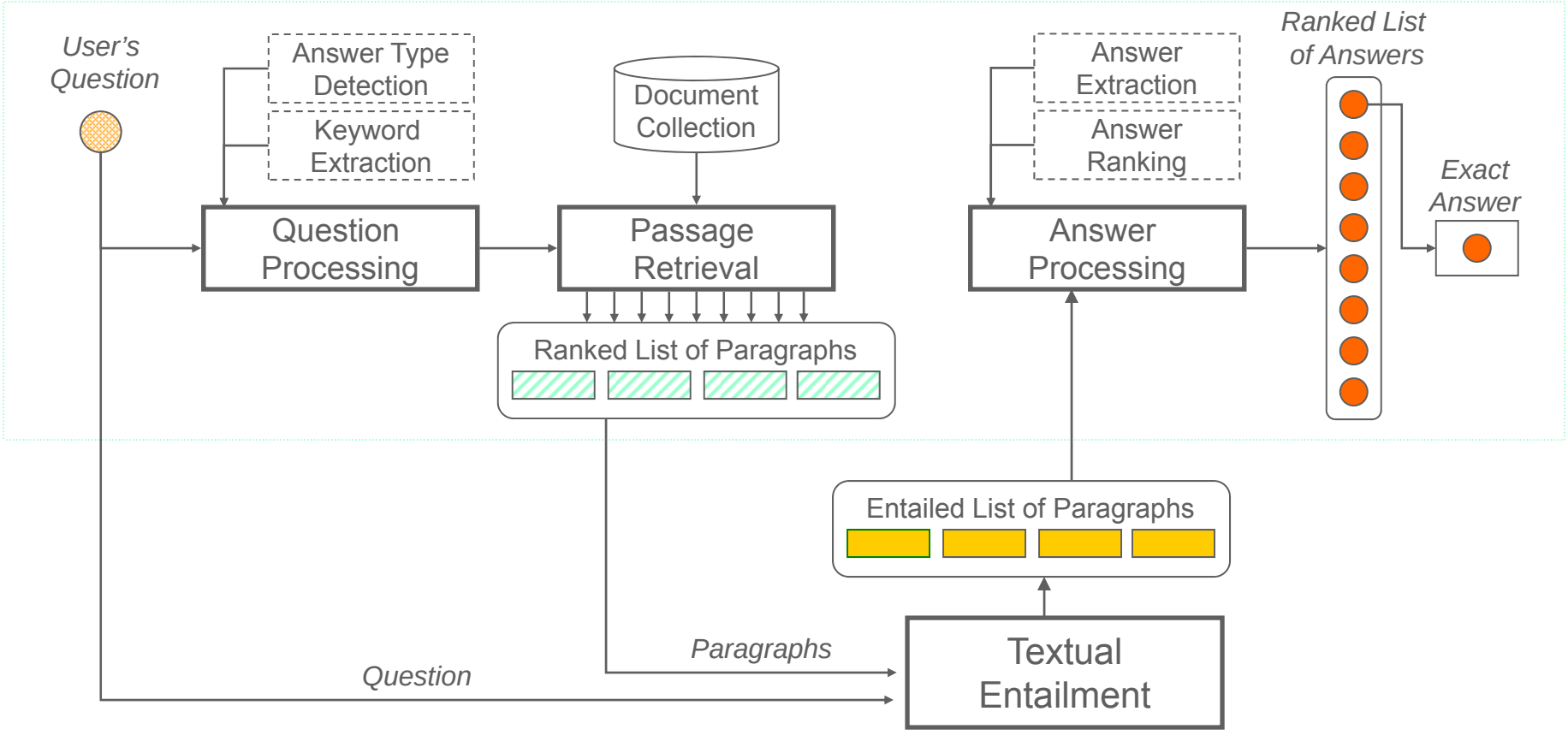
Basic Question-Answering Architecture



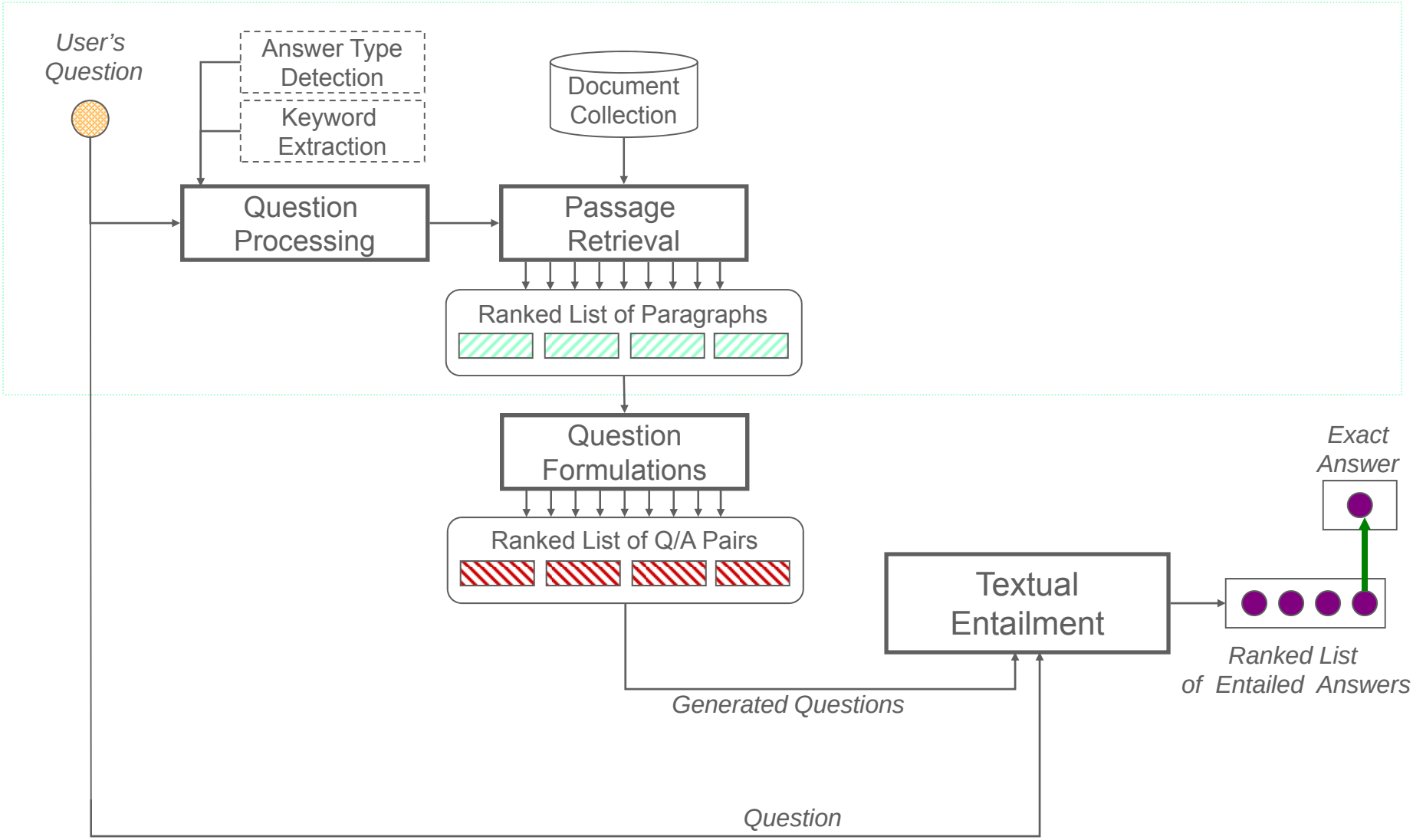
Integrating Textual Entailment into Q/A [1]



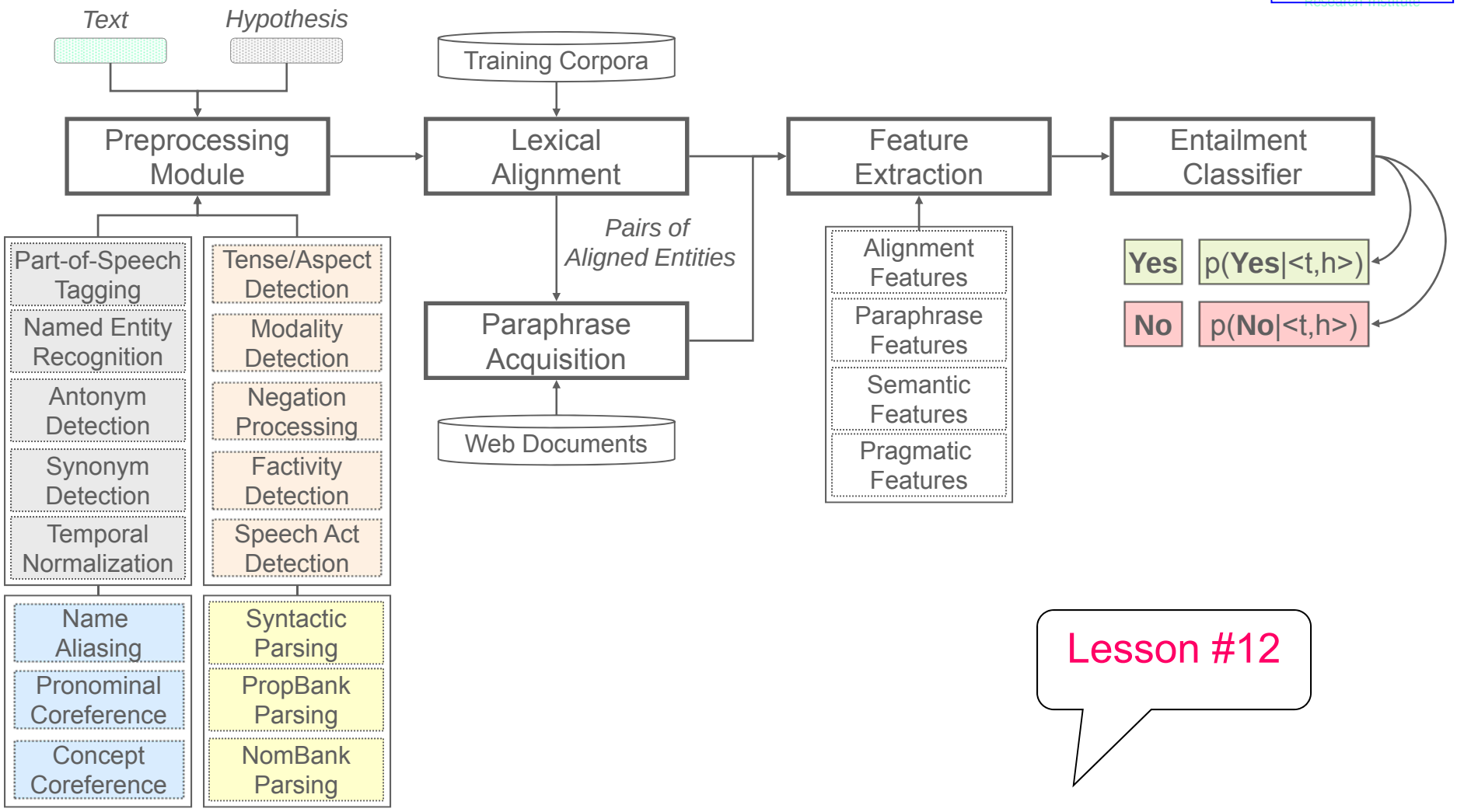
Integrating Textual Entailment into Q/A [2]



Integrating Textual Entailment into Q/A [3]



How does a TE System work?



Lesson #12

Textual Entailment operates like a semantic inference method

Q/A Baseline

- We evaluated impact of TE in Q/A using a corpus of **600** factoid questions selected randomly from past TREC evaluations:
  - **55.0%** of questions (335/600) were assigned a correct answer type from a hierarchy of over 150+ different answer types
  - **45.0%** of questions (265/600) were assigned an incorrect answer type or were assigned no answer type from the answer type hierarchy
- In order to provide a baseline, questions were then submitted to an interactive Q/A system (Harabagiu et al. 2005) that does **not** make use of TE information in answering questions:

Question Set	Questions	Correct	Accuracy	MRR
Known Answer Types	335	107	32.0%	0.3001
Unknown Answer Types	265	81	30.6%	0.2987
All Questions	**600**	**188**	**31.3%**	**0.2999**

RTE Performance

- Performance of a TE system was evaluated in the 2006 PASCAL Recognizing Textual Entailment Challenge
- Four types of sentence pairs were considered:
  - **QA**: Output from an Automatic Question-Answering System
  - **IE**: Output from an Information Extraction System
  - **IR**: Output from an Information Retrieval Engine
  - **SUM**: Output from a Multi-Document Summarization System

	Training Data		**Difference**
	Development Set	**Additional Corpora**	
Number of Examples	800	201,000	200,200
QA-Test	0.5750	0.6950	0.1200
IE-Test	0.6450	0.7300	0.0850
IR-Test	0.6200	0.7450	0.1250
SUM-Test	0.7700	0.8450	0.0750
Overall Accuracy	0.6525	**0.7538**	0.1013

Validating Candidate Answers with TE

- Method 1: Using TE to Filter Candidate Answers

Q. What did Peter Minuit buy for the equivalent of $24.00?

A1. Everyone knows that back in 1626, Peter Minuit bought **Manhattan** from the Indians for $24 worth of trinkets.

A2. In 1626, Minuit flagged down some passing locals, plied them with **beads, cloth and trinkets** worth $24 and walked away with the island.

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	**44.1%**	**0.4114**	**39.5%**	**0.3833**
Method 2	52.4%	0.5558	42.7%	0.4135
Method 3	41.5%	0.4257	37.5%	0.3575
Hybrid	53.9%	0.5640	41.9%	0.4010

Validating Candidate Answers with TE

- Method 2: Using TE to Rank Candidate Passages

Q. What did Peter Minuit buy for the equivalent of $24.00?

A3. Peter Minuit, an enterprising Dutch businessman, purchased **Manhattan** while on a hunting trip from the Canarsee Indians for "diverse other wares" valued at 60 guilders, or about $25 U.S. dollars.

A4. Even though the Dutch East India Company maintained extensive records on their activities in New Amsterdam since 1600, there is no evidence that Minuit bought **Manhattan** for the equivalent of $24.00.

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	44.1%	0.4114	39.5%	0.3833
Method 2	**52.4%**	**0.5558**	**42.7%**	**0.4135**
Method 3	41.5%	0.4257	37.5%	0.3575
Hybrid	53.9%	0.5640	41.9%	0.4010

Validating Candidate Answers with TE

- Method 3: Using TE to Select Question Decompositions

Q. How hot does the inside of an active volcano get?

A1. Tamagawa University volcano experts said lava fragments belched out of the mountain on Jan 31 were as hot as [**300 degrees Fahrenheit**]$_{Answer1}$. The intense heat from a second eruption on Tuesday forced the government's rescue operations to stop after [**90 minutes**]$_{Answer2}$.

Q_1. What temperature were the lava fragments belched out of the mountain?

Q_2. How long did rescue operations last?

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	44.1%	0.4114	39.5%	0.3833
Method 2	52.4%	0.5558	42.7%	0.4135
Method 3	**41.5%**	**0.4257**	**37.5%**	**0.3575**
Hybrid	53.9%	0.5640	41.9%	0.4010

Validating Candidate Answers with TE

- Hybrid Method 1:
  - Candidate answers ranked using features from entailment classification
    - Original Question – Candidate Answer (Method 1)
    - Original Question – Decomposed Questions (Method 3)
  - After ranking, filtered answers from top 25 answers were not entailed by original question

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	44.1%	0.4114	39.5%	0.3833
Method 2	52.4%	0.5558	42.7%	0.4135
Method 3	41.5%	0.4257	37.5%	0.3575
Hybrid M1	**53.9%**	**0.5640**	**41.9%**	**0.4010**

Validating Candidate Answers with TE

- Hybrid Method 2:
  - Candidate answers ranked using features from entailment classification
    - Original Question – Candidate Answer (Method 1)
    - Original Question – Retrieved Passages (Method 2)
  - After ranking, filtered answers from top 25 answers were not entailed by original question

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	44.1%	0.4114	39.5%	0.3833
Method 2	52.4%	0.5558	42.7%	0.4135
Method 3	41.5%	0.4257	37.5%	0.3575
Hybrid M1	**53.9%**	**0.5640**	**41.9%**	**0.4010**
Hybrid M2	**66.5%**	**0.6921**	**55.4%**	**0.5435**

Validating Candidate Answers with TE

- Hybrid Method 3:
  - Candidate answers ranked using features from entailment classification
    - Original Question – Retrieved Passages (Method 2)
    - Original Question – Decomposed Questions (Method 3)
  - After ranking, filtered answers from top 25 answers were not entailed by original question

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	44.1%	0.4114	39.5%	0.3833
Method 2	52.4%	0.5558	42.7%	0.4135
Method 3	41.5%	0.4257	37.5%	0.3575
Hybrid M1	**53.9%**	**0.5640**	**41.9%**	**0.4010**
Hybrid M2	**66.5%**	**0.6921**	**55.4%**	**0.5435**
Hybrid M3	**64.2%**	**0.6735**	**52.7%**	**0.5120**

Validating Candidate Answers with TE

- Hybrid Method 4:
  - Candidate answers ranked using features from entailment classification
    - Original Question – Candidate Answer (Method 1)
    - Original Question – Retrieved Passages (Method 2)
    - Original Question – Decomposed Questions (Method 3)
  - After ranking, filtered answers from top 25 answers were not entailed by original question

	Known Answer Type		Unknown Answer Type	
	Accuracy	MRR	Accuracy	MRR
Baseline	32.0%	0.3001	30.6%	0.2978
Method 1	44.1%	0.4114	39.5%	0.3833
Method 2	52.4%	0.5558	42.7%	0.4135
Method 3	41.5%	0.4257	37.5%	0.3575
Hybrid M1	**53.9%**	**0.5640**	**41.9%**	**0.4010**
Hybrid M2	**66.5%**	**0.6921**	**55.4%**	**0.5435**
Hybrid M3	**64.2%**	**0.6735**	**52.7%**	**0.5120**
Hybrid M4	**72.1%**	**0.7420**	**60.1%**	**0.5984**

How much semantic inference is needed in QA?

By doing only:

- named entity recognition
- semantic classification of the expected answer type (off-line taxonomy + semantic info from WordNet)
- Text mining

55% accuracy on factual trivia-like questions (*TREC-8, Moldovan et al.)*

What else is needed?

By incorporating a TE system into a QA architecture: we have obtained 72% accuracy (for known AT) and 60% accuracy for Unknown AT!

Lesson #13

Semantic inference has more impact than semantic processing in QA.

Conclusions

- **What did I learn about semantics from Textual QA?**
  1. There are different forms of semantic knowledge required by QA systems : EATs, semantic structures (predicate argument structures, frames), semantic relations.
  2. Semantic information encoded in EATs plays and important role in QA processing.
  3. Semantic information for EATs needs to be recognized by text mining techniques.
  4. Predicate-arguments structures improve the detection of EATs!!!
  5. Predicate-arguments structures improve answer extraction!!!
  6. More sophisticated semantic structures can be recognized by more complex systems. Implicit and explicit semantic knowledge of significance to QA processing needs to be uncovered !!!
  7. Evaluate semantic parsers on QA data before incorporating the Semantic parsers in QA systems
  8. Need for application-driven semantic parsing and annotations!

Conclusions

- **What did I learn about semantics from Textual QA?**
  9. Relational semantics translated the unknown semantic structures of a question into a network that corresponds to question decompositions.
  10. Semantic processing serves textual inference.
  11. Textual Entailment is a very useful form of semantic inference for QA.
  12. Semantic inference has more impact on the results of QA systems than semantic processing alone.
  13. Additional semantic information needs to be considered, e.g. event structures, temporal structure of events,
  14. cQA relies on identifying the intentions of questions.
  15. Users organized in social networks mutually re-enforce the answers provided to questions
  16. Semantic processing needs to consider contexts that are larger than one sentence.
  17. Semantic knowledge is not all explicit. We need to infer the implicit semantic information as well.
  18. Semantic processing for textual QA is complex, effort-intensive, and still an exploratory activity.