

Vision and Language

Boris Katz

MIT Computer Science and
Artificial Intelligence Laboratory

August 28, 2015

Multi-modal understanding

- Humans perform many tasks that involve multiple modalities and span established fields: Computer Vision, AI, NLP, and Cognitive Science
- Our goal is to bring to bear techniques from all of these fields, create new models, and shed light on how such tasks operate in the brain
- We are seeking flexible, integrated and grounded AI

What does scene recognition involve?



Slide by Fei-Fei, Fergus, Torralba

Verification: *Is that a street lamp?*



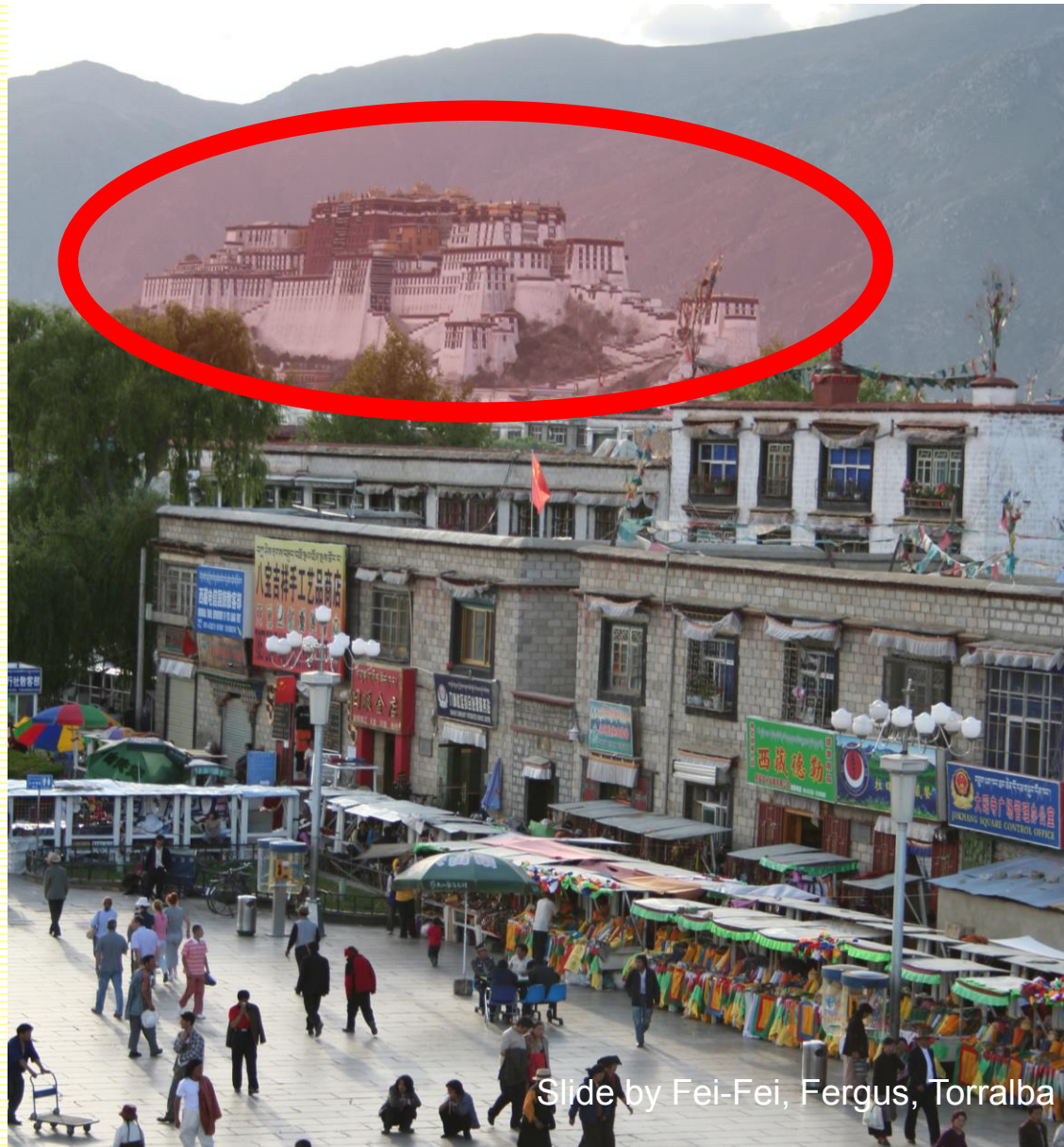
Slide by Fei-Fei, Fergus, Torralba

Detection: *Are there people in the scene?*



Slide by Fei-Fei, Fergus, Torralba

Identification: *Is that Potala Palace?*



Slide by Fei-Fei, Fergus, Torralba

Object categorization



Activity



What is this person doing?

What are these two doing??

Slide by Fei-Fei, Fergus, Torralba

Human understanding of a scene

- *Object verification, detection, identification, categorization ...*
- *Spatial and temporal relationships* between objects
- *Event Recognition*
- *Explanation: What past events caused the scene to look as it does?*
- *Prediction: What future events might occur in the scene?*
- *Gap Filling / Hallucination: What objects – occluded or invisible in the scene – might also be present there? What events – not visible in the scene – could have occurred?*
- ...

Why are machines falling short?

- Our visual system is tuned to process structures typically found in the world.
- But our machines don't know enough about the world and they don't know what structures and events *make sense* and *typically occur* in the world.

Blurry video example





© Fei-Fei Li, Rob Fergus, Antonio Torralba. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Humans make mistakes, too





© Fei-Fei Li, Rob Fergus, Antonio Torralba. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Why are machines falling short?

- Our visual system is tuned to process structures typically found in the world.
- But our machines don't know enough about the world and they don't know what structures and events *make sense* and *typically occur* in the world.

Questions:

- How is this knowledge obtained?
- How can we pass this knowledge to our computer systems?
- How can we determine whether the computer knowledge is correct?
- Our (**partial**) answer – *using language*



Proposal: Combining language and vision processing

- Create a knowledge base containing descriptions of objects, their properties, and relationships among them as they are typically found in the world
- Make the knowledge base available to a scene-recognition system
- Test the performance of the vision system by asking natural language questions

Querying a scene recognition system



Courtesy of Brisbane International Airport. Used with permission.

How many men are in the picture?

What is on the cart?

What is the color of the shirt on the lady with the blond hair?

Is anyone walking?

What does the number on the sign say?

Is there any luggage?

Querying a scene recognition system



© Tom Casino. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Q: Who is winning, yellow or red?

A. Red

Knowledge involved:

- The scene depicts a sporting event involving winners and losers
- Here “yellow” and “red” mean the people wearing these colors
- No attention needs to be paid to people in the audience wearing red
- A participant being on the floor likely indicates a loser

Proposal: Combining language and vision processing

- Create a knowledge base containing descriptions of objects, their properties, and relationships among them as they are typically found in the world
- Make the knowledge base available to a scene-recognition system
- Test the performance of the vision system by asking natural language questions

START: Natural language tools

- Providing Machines with New Knowledge:

NL text → *semantic representation*

- Explaining Computer Actions or Describing its Knowledge:

semantic representation → *NL text*

- Testing Computer Understanding by Answering Questions:

NL queries → *semantic representation* → *NL responses*

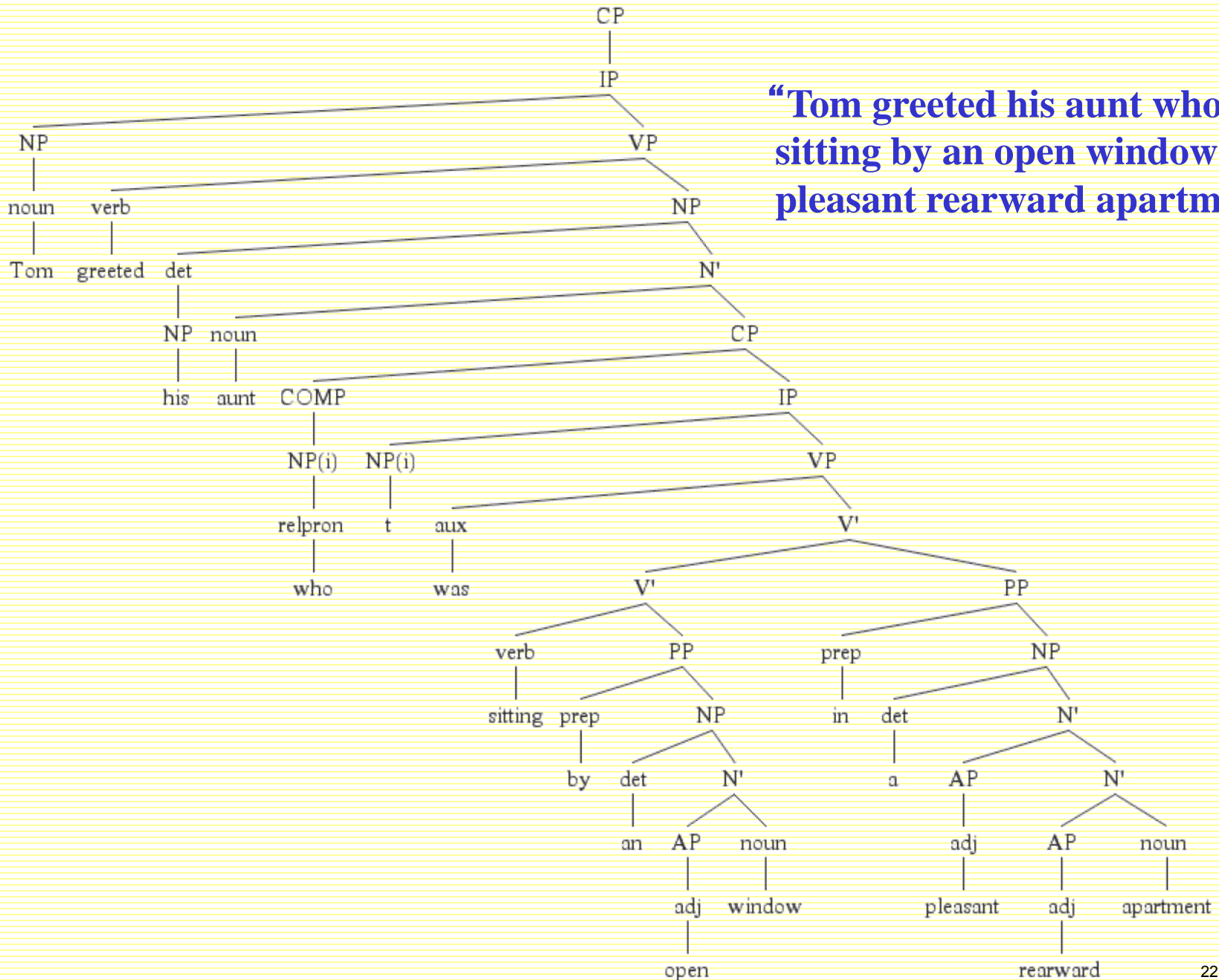
→ *computer actions*

Building blocks in the START system

- Syntactic Analysis: parse trees
- Semantic Representation: ternary expressions
- Language Generation
- Matching
- Replying
- Natural Language Annotations
- Object-Property-Value Data Model
- Question Decomposition

Syntactic analysis: Parse trees

“Tom greeted his aunt who was sitting by an open window in a pleasant rearward apartment.”

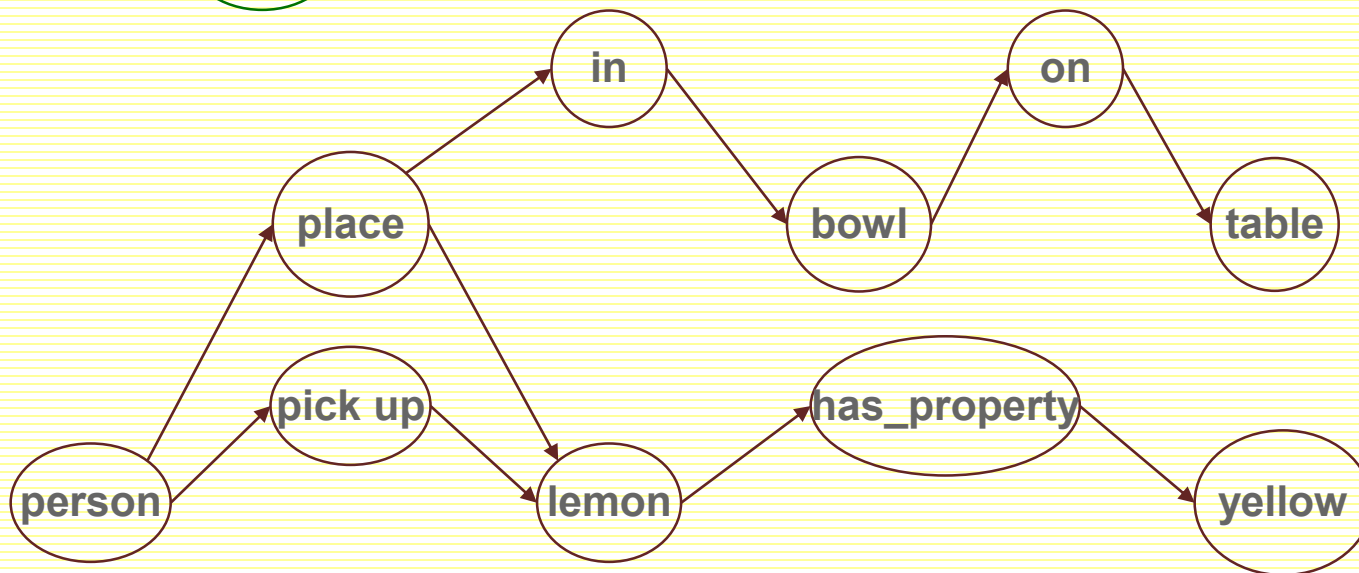
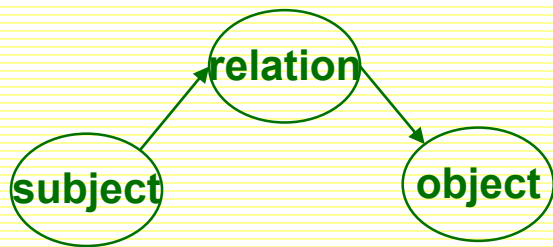


Semantic representation for language: Ternary expression representation

- a versatile syntax-driven representation of language
- highlights significant semantic relations
- very efficient for indexing, matching and retrieval
- reversible representation
- implemented as nested *subject, relation, object* tuples

Language understanding: From sentences to semantic representation

“The person who picked up the yellow lemon placed it in the bowl on the table.”



Language understanding: From sentences to semantic representation

“The person who picked up the yellow lemon placed it in the bowl on the table.”

[subject relation object]



[person place lemon]

[place in bowl]

[bowl on table]

[person pick_up lemon]

[lemon has_property yellow]

Three types of Ternary Expressions

“Tom’s aunt was sitting by an open window”

- Related to the **syntactic structure** of the sentence

[aunt sit nil]

[aunt related_to Tom]

[sit by window]

[window has_property open]

- Related to **syntactic features** that change from sentence to sentence

[sit has_tense past]

[sit is_progressive yes]

[window has_det indefinite]

- Related to **lexical features** of words that don’t change from sentence to sentence

[Tom is_proper yes]

[window has_number singular]

Language generation

As intelligent systems become more mature, they will be expected to...

- Explain their actions
- Answer complex questions
- Keep track of conversation history and state
- Engage in mixed-initiative dialog
- Offer related information of potential interest to the user
- Help users correct and refine their questions
- Indicate incomplete understanding of questions and offer partial responses

Language generation: From semantic representation to sentences

[person place lemon]
[place in bowl]
[bowl on table]
[person pick_up lemon]
[lemon has_property yellow]



“The person who picked up the yellow lemon placed it in the bowl on the table.”

Language generation: Fine-tuning

Semantic knowledge

[person place lemon]
[place in bowl]
[bowl on table]
[person pick_up lemon]
[lemon has_property yellow]
[pick_up has_modifier soon]

Syntactic features

[place has_tense future]
[bowl has_det indefinite]
[pick_up is_main yes]
[pick_up is_question yes]



“Will the person who placed the yellow lemon in a bowl on the table pick it up soon?”

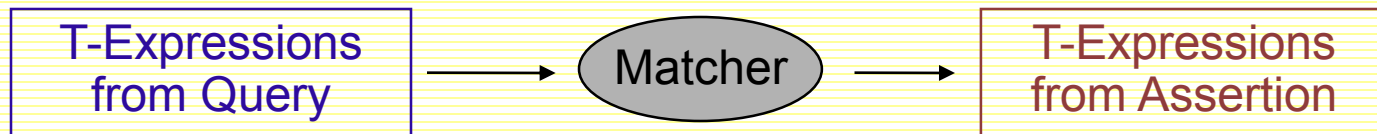
Building blocks in the START system

- Syntactic Analysis: parse trees
- Semantic Representation: ternary expressions
- Language Generation
- Matching
- Replying
- Natural Language Annotations
- Object-Property-Value Data Model
- Question Decomposition

Matching T-Expressions

Assertion: “Tom greeted his aunt who was sitting by an open window in a pleasant rearward apartment.”

Query: “Was anyone sitting by an open window?”

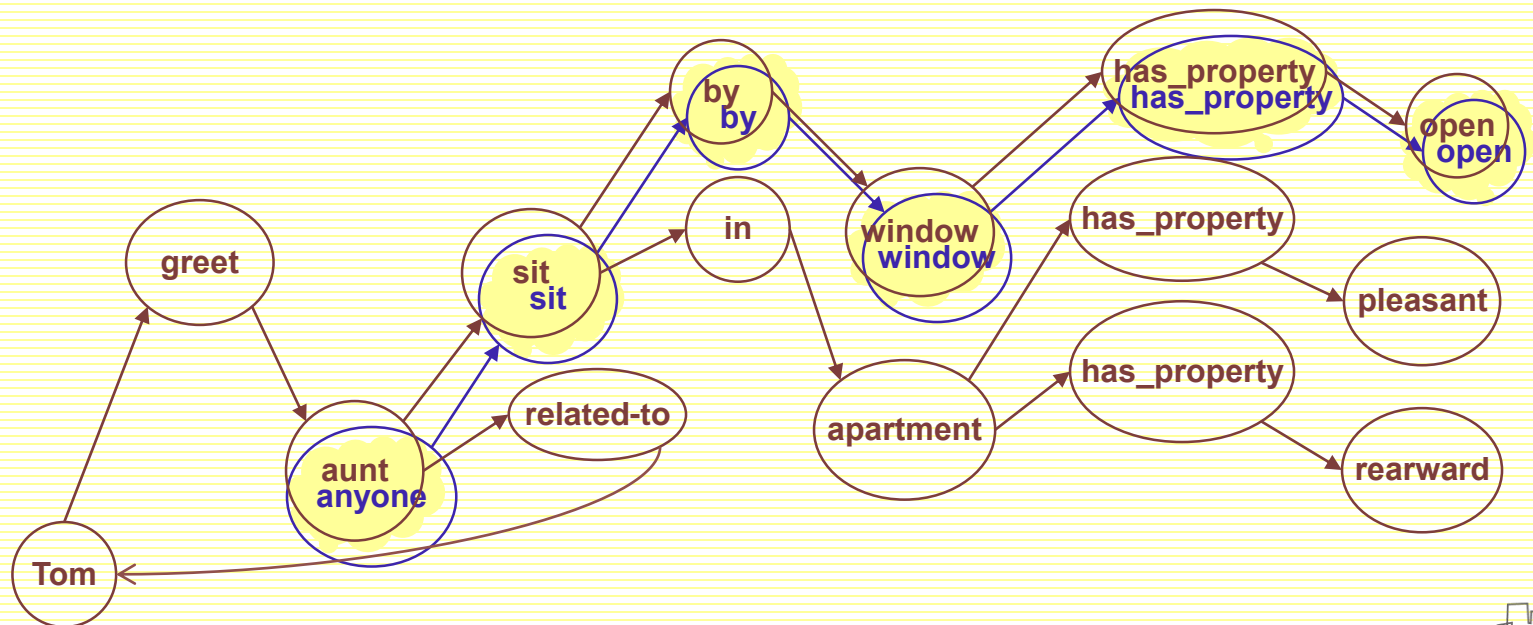
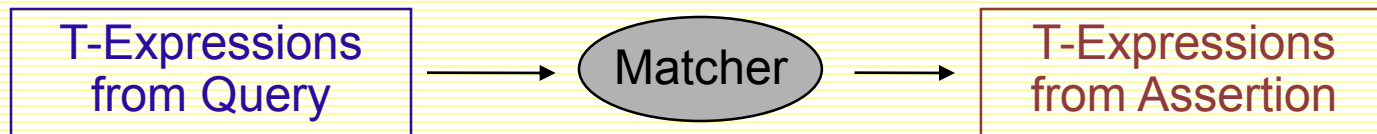


		[Tom greet aunt]	KB
		[aunt related-to Tom]	
[anyone sit ∅]	⇒	[aunt sit ∅]	
[sit by window]	⇒	[sit by window]	
		[sit in apartment]	
[window has_property open]	⇒	[window has_property open]	
		[apartment has_property pleasant]	
		[apartment has_property rearward]	

Matching T-Expressions

Assertion: “Tom greeted his aunt who was sitting by an open window in a pleasant rearward apartment.”

Query: “Was anyone sitting by an open window?”



Matching in START

T-Exps from Questions \longleftrightarrow T-Exps from Assertions

- term matching:
 - lexical match
 - synonym match
 - hyponym match
- structure matching:
 - exact match
 - match via transformational S-rules

Verb argument alternations and paraphrases

“*Surprise*”:

“The patient surprised the doctor with his fast recovery.”

“The patient’s fast recovery surprised the doctor.”

“*Load*”:

“The crane loaded the ship with containers.”

“The crane loaded containers onto the ship.”

“*Provide*”:

“Did Iran provide Syria with weapons?”

“Did Iran provide weapons to Syria?”

Verb argument alternations and paraphrases

“*Surprise*”:

“The patient surprised the doctor with fast recovery.”

“The patient’s fast recovery surprised the doctor.”

* “The patient surprised fast recovery onto the doctor.”

“*Load*”:

“The crane loaded the ship with containers.”

“The crane loaded containers onto the ship.”

* “The crane’s containers loaded the ship.”

“*Provide*”:

“Did Iran provide Syria with weapons?”

“Did Iran provide weapons to Syria?”

* “Did Iran’s weapons provide Syria?”

Verb classes and S-Rules

“**Surprise**”: “The patient surprised the doctor with fast recovery.”

“The patient’s fast recovery surprised the doctor.”

“**Confuse**”: “The patient confused the doctor with slow recovery.”

“The patient’s slow recovery confused the doctor.”

...

Emotional Reaction Verbs (semantic class):

anger, confuse, disappoint, embarrass, frighten, impress, please, *surprise, threaten*, ...

S-Rule:

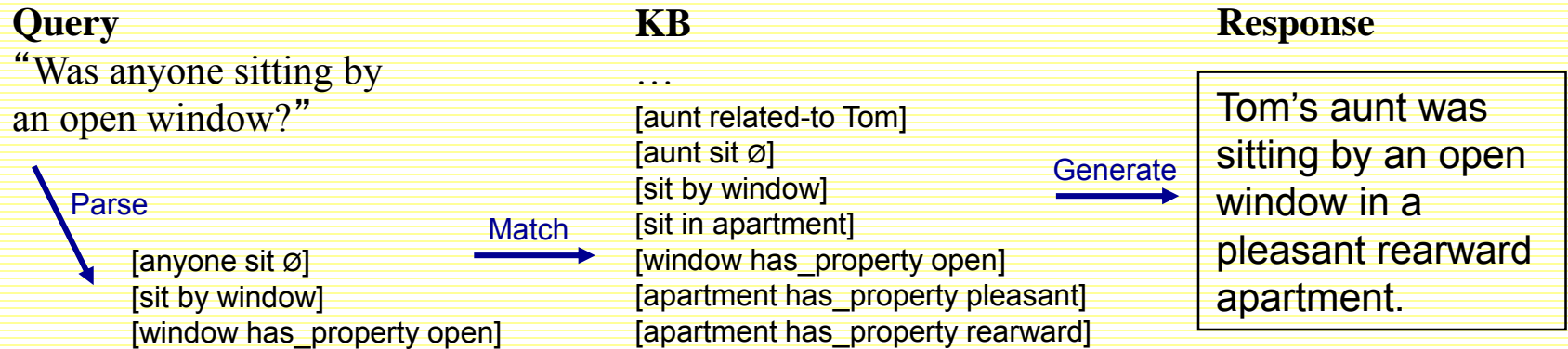
If: $[[\textit{subject1 verb subject2}] \textbf{with object}]$

Then: $[\textit{object verb subject2}]$
 $[\textit{object related_to subject1}]$

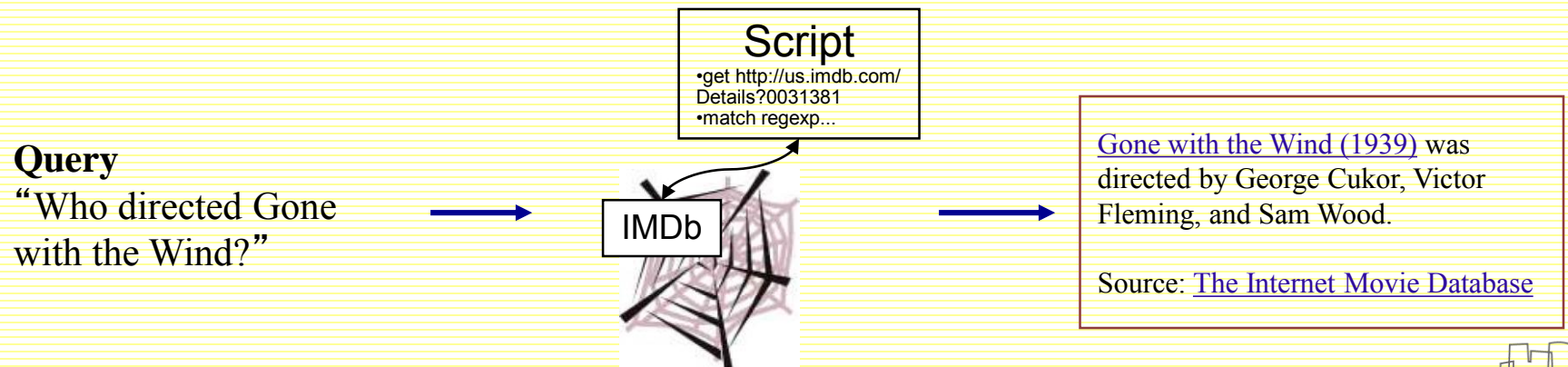
Provided: $\textit{verb} \in \textbf{emotional reaction class}$

Replying to a question after a successful match

1. Generate a sentence from semantic representation



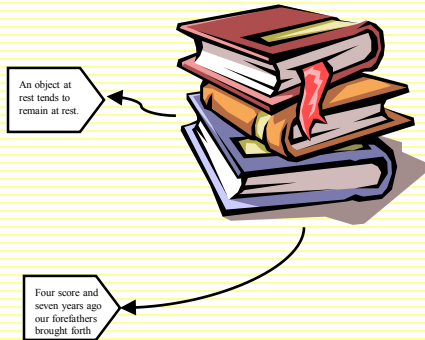
2. Execute a procedure to obtain an answer from the data source



Natural Language Annotations

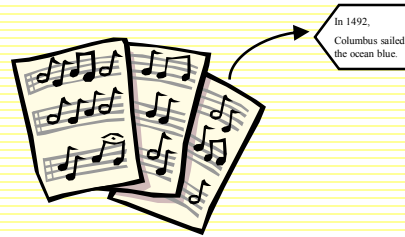


© Source Unknown. All rights reserved.
This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

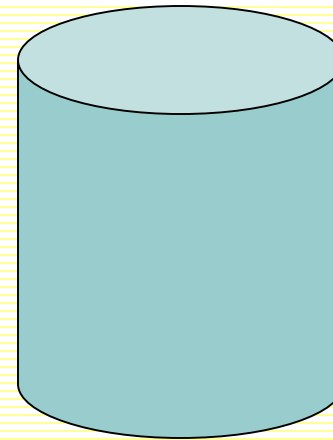


© Source Unknown. All rights reserved.
This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

+



© Source Unknown. All rights reserved.
This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.



Knowledge Base

- annotations = sentences and phrases that describe the content of retrievable information segments
- annotations are matched against submitted queries
- successful match results in retrieval of information

The object–property–value data model

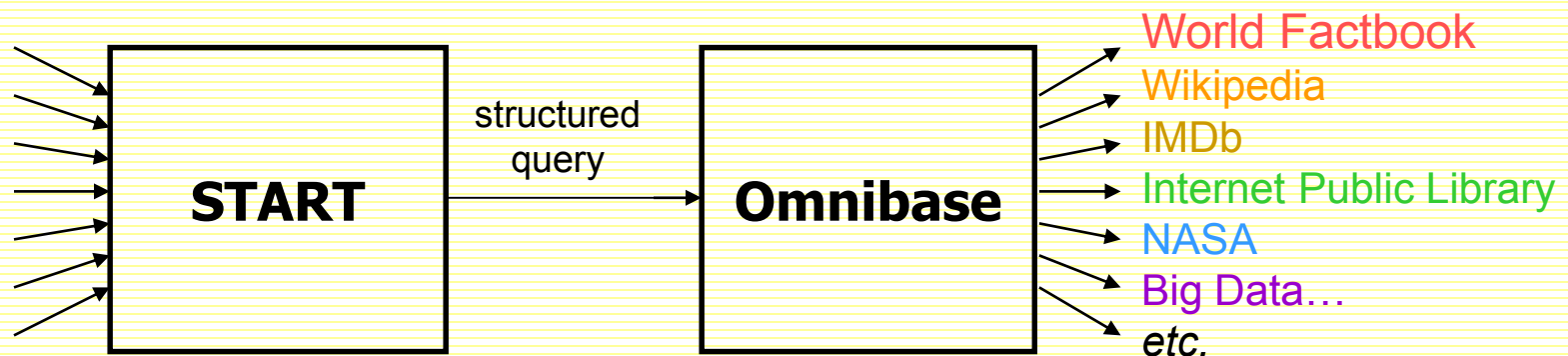
- Many heterogeneous semi-structured information sources on the Web can be modeled using an *object–property–value* (OPV) data model:
 - *countries* and their *capitals*, *areas*, *populations*, ...
 - *individuals* and their *biographies*, *birthdates*, *spouses*, ...
 - *cities* and their *weather reports*, *maps*, *elevations*, ...
- The OPV Model makes it possible to view and use large segments of the Web as a database

Implementing the OPV model: START and Omnibase

Omnibase supports START by providing access to structured and semi-structured information in databases, on the Web, etc.

User Questions

Data Resources



1. What does the question mean?
2. Where can the answer be found?
3. What are the object and property?

1. Go to the specific data source or Web page containing the answer.
2. Extract the answer from the data source.

START in action


The START Natural Language Question Answering System

START
Natural Language Question Answering System


Does Russia border on Moldova? [Ask Question >](#)

==> Does Russia border on Moldova?
Moldova does not border on Russia.

Moldova



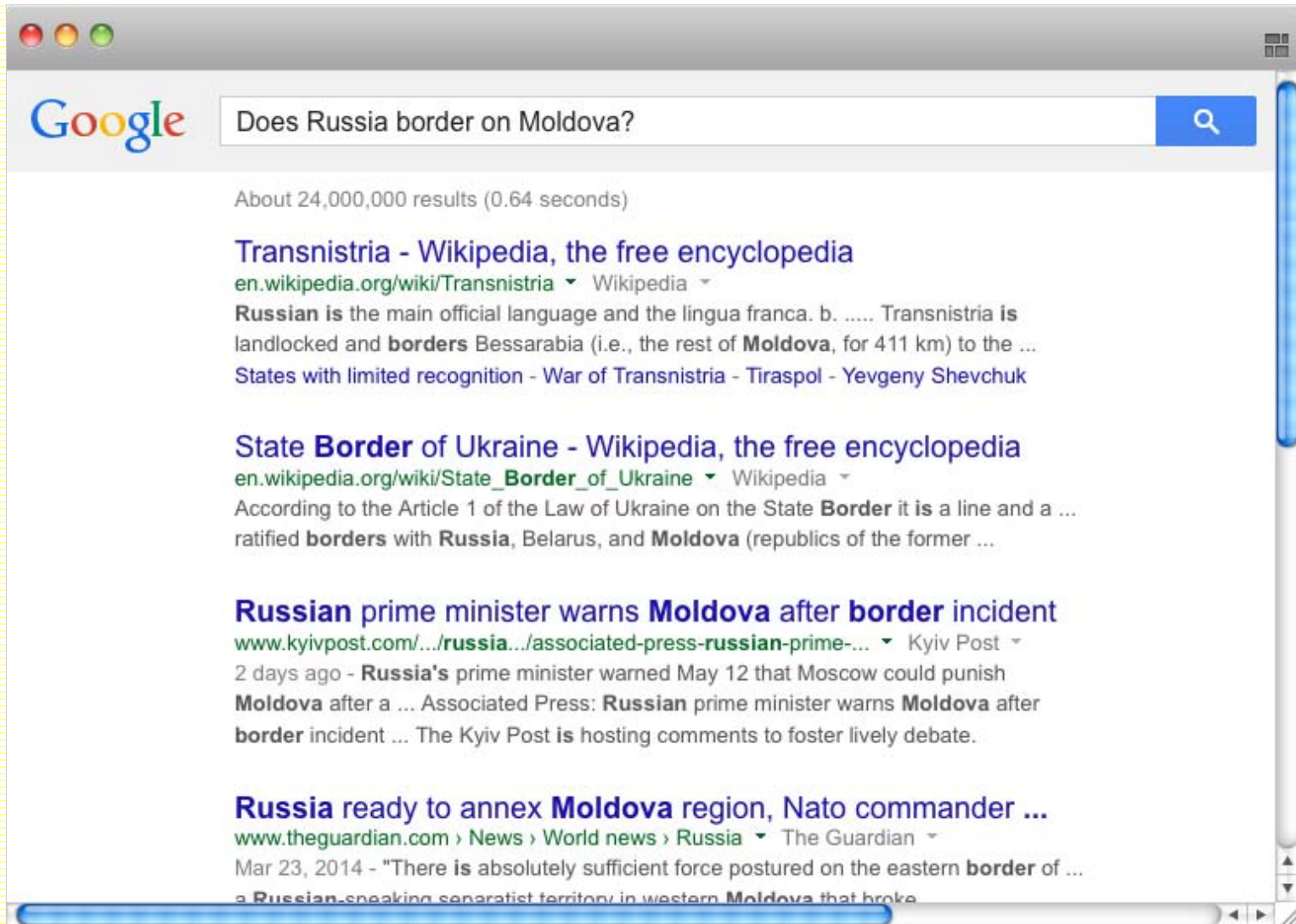
Land boundaries:
total: 1,390 km
border countries: Romania 450 km, Ukraine 940 km



Source: [The World Factbook](#)

Courtesy of START Natural Language Question Answering System. Used with permission.

A search engine in action



The screenshot shows a Google search interface. The search bar contains the text "Does Russia border on Moldova?". Below the search bar, it indicates "About 24,000,000 results (0.64 seconds)". The search results are listed as follows:

- Transnistria - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Transnistria - Wikipedia
Russian is the main official language and the lingua franca. b. Transnistria is landlocked and borders Bessarabia (i.e., the rest of Moldova, for 411 km) to the ... States with limited recognition - War of Transnistria - Tiraspol - Yevgeny Shevchuk
- State Border of Ukraine - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/State_Border_of_Ukraine - Wikipedia
According to the Article 1 of the Law of Ukraine on the State Border it is a line and a ... ratified borders with Russia, Belarus, and Moldova (republics of the former ...
- Russian prime minister warns Moldova after border incident**
www.kyivpost.com/.../russia.../associated-press-russian-prime-... - Kyiv Post
2 days ago - Russia's prime minister warned May 12 that Moscow could punish Moldova after a ... Associated Press: Russian prime minister warns Moldova after border incident ... The Kyiv Post is hosting comments to foster lively debate.
- Russia ready to annex Moldova region, Nato commander ...**
www.theguardian.com > News > World news > Russia - The Guardian
Mar 23, 2014 - "There is absolutely sufficient force postured on the eastern border of ... a Russian-speaking separatist territory in western Moldova that broke

Answering complex questions

- Syntactically decompose a complex question into a set of nested ternary expressions
- Successively resolve groups of ternary expressions containing variables
 - Answer sub-questions by replacing variables with values

“Who is the president of the fourth largest country married to?”

What is the fourth largest country?

Who is its president?

Who is he married to?

The “under the hood” view of syntactic decomposition

“Who is the president of the fourth largest country married to?”

< < president is married > to who >

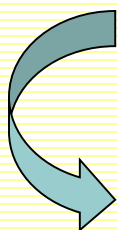
< president related-to country >

< country is largest >

< largest mod fourth >

country = China

The fourth largest country is China.



< < president is married > to who >

< president related-to **China** >

president = Xi Jinping

The president of China is Xi Jinping.




< < **Xi Jinping** is married > to who >

who = Peng Liyuan

Xi Jinping is married to Peng Liyuan.

Answering Complex Questions



The START Natural Language Question Answering System

START
Natural Language Question Answering System

In what city was the 5th president of the US born? [Ask Question >](#)

====> In what city was the 5th president of the US born?

I know that the 5th president of the USA is James Monroe (source: Internet Public Library).

Using this information, I determined where James Monroe was born:

James Monroe

Place of origin: Monroe Hall, Virginia

I know about two more persons called "James Monroe": [James Monroe \(congressman\)](#) and [James Monroe \(New York politician\)](#)

Source: [Wikipedia](#)

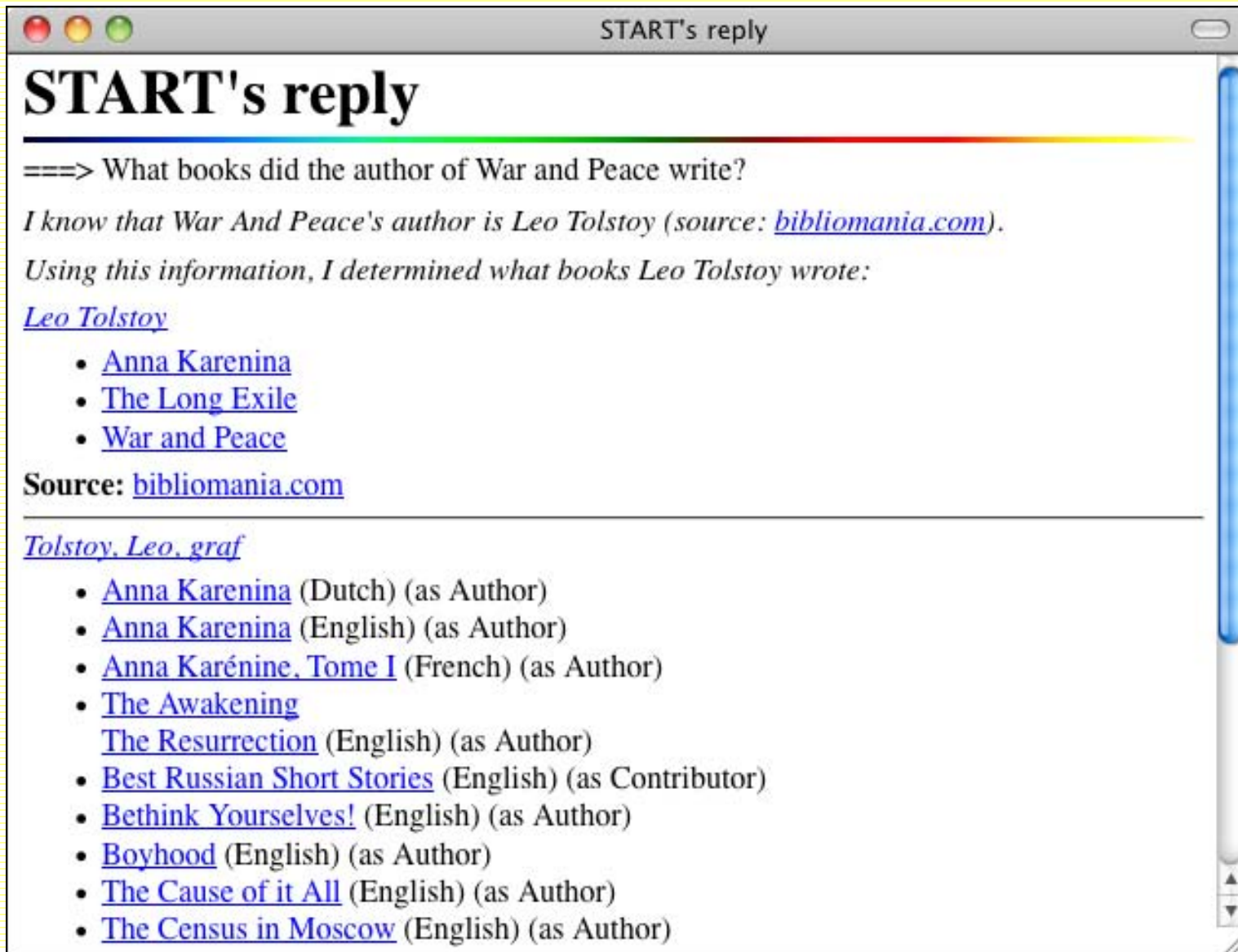
[James Monroe](#) was born April 28, 1758, in Westmoreland County, Virginia.

Source: [Internet Public Library](#)

[James Monroe \(II\)](#)'s location of birth: [Westmoreland County, Virginia, USA](#)

Source: [The Internet Movie Database](#)

Replying: Syntactic Decomposition



START's reply

====> What books did the author of War and Peace write?

I know that War And Peace's author is Leo Tolstoy (source: bibliomania.com).

Using this information, I determined what books Leo Tolstoy wrote:

[Leo Tolstoy](#)

- [Anna Karenina](#)
- [The Long Exile](#)
- [War and Peace](#)

Source: bibliomania.com

[Tolstoy, Leo, graf](#)

- [Anna Karenina](#) (Dutch) (as Author)
- [Anna Karenina](#) (English) (as Author)
- [Anna Karénine, Tome I](#) (French) (as Author)
- [The Awakening](#)
- [The Resurrection](#) (English) (as Author)
- [Best Russian Short Stories](#) (English) (as Contributor)
- [Bethink Yourselves!](#) (English) (as Author)
- [Boyhood](#) (English) (as Author)
- [The Cause of it All](#) (English) (as Author)
- [The Census in Moscow](#) (English) (as Author)

Technologies inspired by START

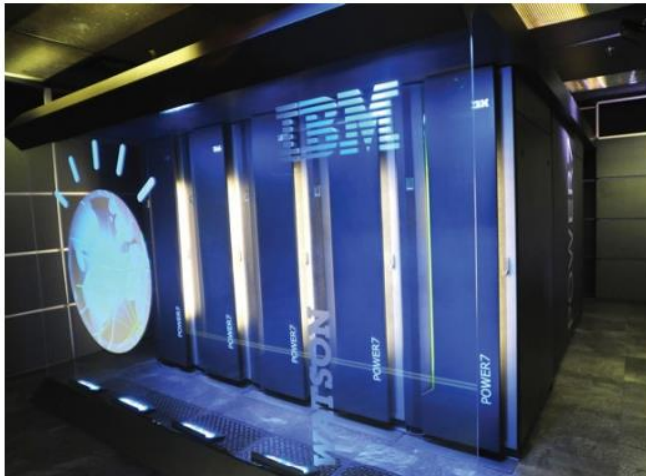
- Ask Jeeves / Ask.com
- Wolfram Alpha
- Google QA
- IBM's Watson
- Apple's Siri

...

The Brains Behind Watson

CSAIL professor helped computer win at *Jeopardy!*

By Larry Hardesty on April 19, 2011



In February, the game show *Jeopardy!* pitted its two most successful contestants against an IBM computer system called Watson, which defeated them soundly. Several of the strategies that Watson used were based on research by Boris Katz, a principal research scientist in the Computer Science and Artificial Intelligence Laboratory.

In the early 1980s, Katz began developing a natural-language question-answering system called START, which went online in 1993. But automatic question answering ultimately gave way to search engines like Yahoo and Google, which provided less precise answers but were much easier to implement.

Nokia Phones Go to Natural Language Class

Nokia and MIT researchers are teaching cell phones to take commands in natural language.

By Katherine Bourzac on April 27, 2006



As part of a research collaboration with MIT computer scientists, the Nokia Research Center Cambridge, in Cambridge MA, is developing cell phones that can understand and respond to written commands typed in English.

Using the MobileStart system on the phone on the right, you can remind your mother to take her medication. The phone on the left shows a new calendar event created in "mom's" phone by MobileStart. (Courtesy of Boris Katz and Federico Mora, MIT.)

Robert Iannucci, head of Nokia's research centers, says the company wants to transform phones from simple calling terminals to "information gateways" – to the Internet, GPS and sensors, MP3s, desktop computers, iPods, and other devices. And, he says, that requires rethinking the entire interface between people and handhelds. For both Nokia and MIT, that means using text interaction.

"Humans are good with language," says [Boris Katz](#), lead research scientist at MIT's Computer Science and Artificial Intelligence Laboratory, the principle group working with Nokia. "We want language to be a first-rate citizen" on cell phones, he says.

StartMobile: Using Language to Connect People to Mobile Devices

- An intelligent phone assistant in 2005-2006 (*before the iPhone even existed!*)
- Retrieving general-purpose information
- Providing access to available computational services
- Performing an action on another mobile device
- Triggering specific apparatus (*e.g. camera*) on the user's mobile device
- Retrieving instructional videos
- Combined use of information, context, and available services

STARTMobile: An Intelligent Phone Assistant in 2005-2006



IBM's Watson: Jeopardy! Challenge

Can we create a computer system to compete against the best humans at a task thought to require high levels of human intelligence?

Sample of Jeopardy! Questions (“clues”):

Q: “To push one of these paper products is to stretch established limits”

A: “envelope”

Q: “The chapels at Pembroke & Emmanuel Colleges were designed by this architect”

A: “Sir Christopher Wren”

Question Decomposition in Watson

Clue: “Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that’s farthest north”

Inner sub-clue: “the 4 countries in the world that the U.S. does not have diplomatic relations with”

Outer sub-clue: “Of Bhutan, Cuba, Iran, and North Korea, the one that is farthest north”

Answer: North Korea

Question Decomposition in START

“Who is the president of the fourth largest country married to?”

- Syntactically decompose a complex question into a set of nested ternary expressions
- Successfully resolve groups of ternary expressions containing variables
 - Answer sub-questions by replacing variables with values

“Who is the president of the fourth largest country married to?”

What is the fourth largest country?

Who is its president?

Who is he married to?

From semi-structured to unstructured data

- Watson incorporates ideas from START:
 - Ternary Expressions representation
 - Natural language annotations
 - Object-Property-Value data model
 - Question Decomposition model

and applies them as appropriate when:

- The clue is fully analyzed and understood
 - The semi-structured resource is available for finding the answer
- Watson uses statistical machine learning approaches if the clue is very convoluted and the answer is not available from well-understood data resources.

Watson: steps in the pipeline

- Content acquisition
- Question analysis
- Document search
- Candidate answer generation
- Soft filtering
- Obtaining new evidence
- Scoring
- Final ranking and merging

Speed

- Early implementation of Watson:
 - Ran on a single processor
 - Took 2 hours to answer a single question
- Now:
 - Scaled up to over 2,500 compute cores
 - Reduced the time to about 3 seconds

Jeopardy! Challenge

Can we create a computer system to compete against the best humans at a task thought to require high levels of human intelligence?

- Watson melds the state of the art in many disciplines - NLP, question answering, information retrieval, machine learning
- Great piece of engineering
- Remarkable performance
- It re-ignited the public's interest in artificial intelligence
- It brought new talented people to our field

Watson blunders

Category: “Letters”

Clue: “In the late 40s a mother wrote to this artist that his picture *Number Nine* looked like her son’s finger painting”

Correct answer: “Jackson Pollock”

Watson’s answer: “Rembrandt”

Reason:

- Watson failed to recognize that the phrase “late 40s” referred to the 1940s.

Watson blunders

Category: “U.S. city”

Clue: “Its largest airport is named for a World War II hero;
its second largest, for a World War II battle”

Correct answer: “Chicago” <Edward O’Hare; Midway>

Watson’s answer: “Toronto”

Reasons:

- By studying previous competitions, Watson “learned” to pay less attention to the category part of the clue
- Watson knew that a Toronto team is in the U.S. baseball league
- One of Toronto airports is named for a WWI hero

Jeopardy! Challenge

Can we create a computer system to compete against the best humans at a task thought to require high levels of human intelligence?

- IBM has not created a machine that Thinks Like Us
- Watson's success does not bring us closer to understanding human intelligence
- Watson's occasional blunders should remind everyone that this problem is waiting to be solved
- This should be our next big challenge

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Brains, Minds and Machines Summer Course
Tomaso Poggio and Gabriel Kreiman

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.