Microsoft Research

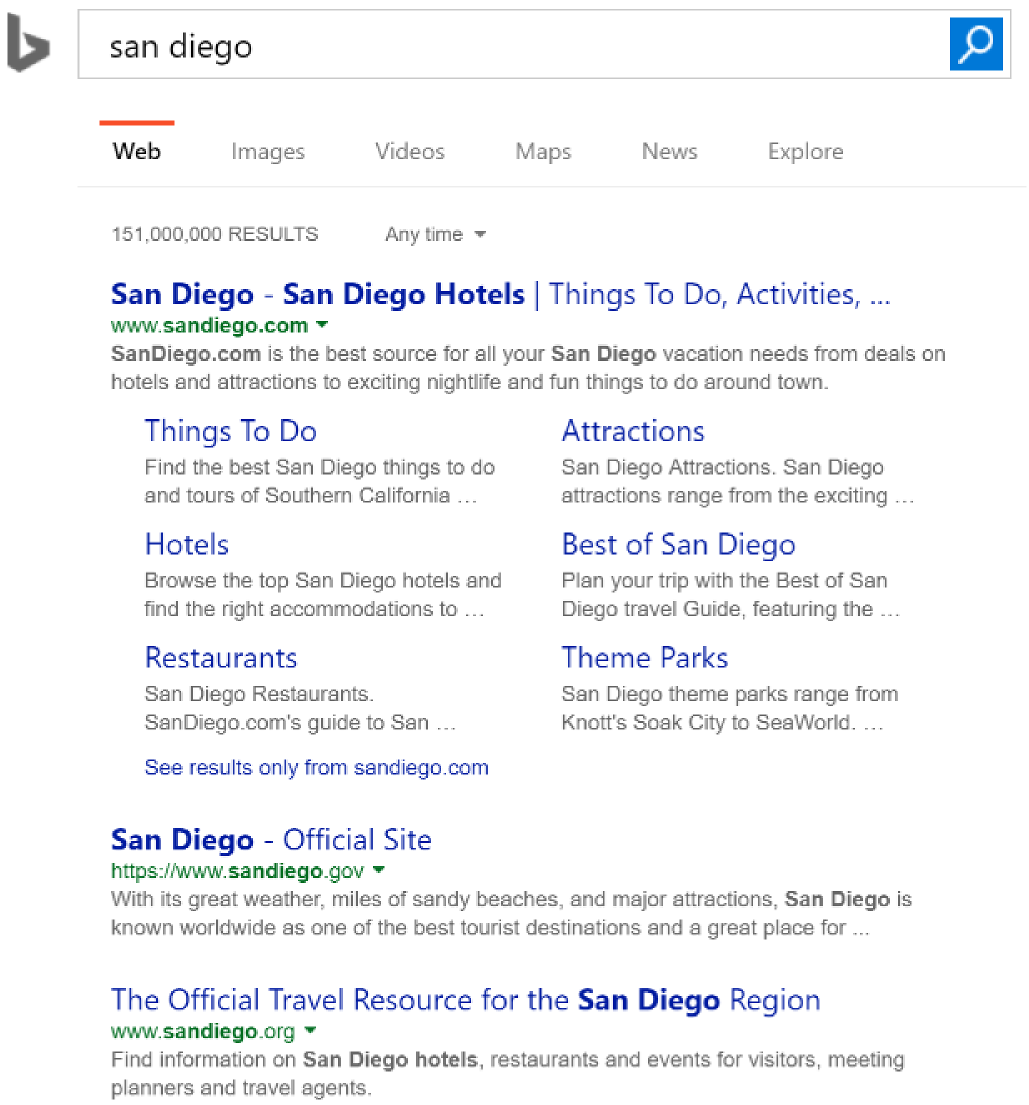
Microsoft

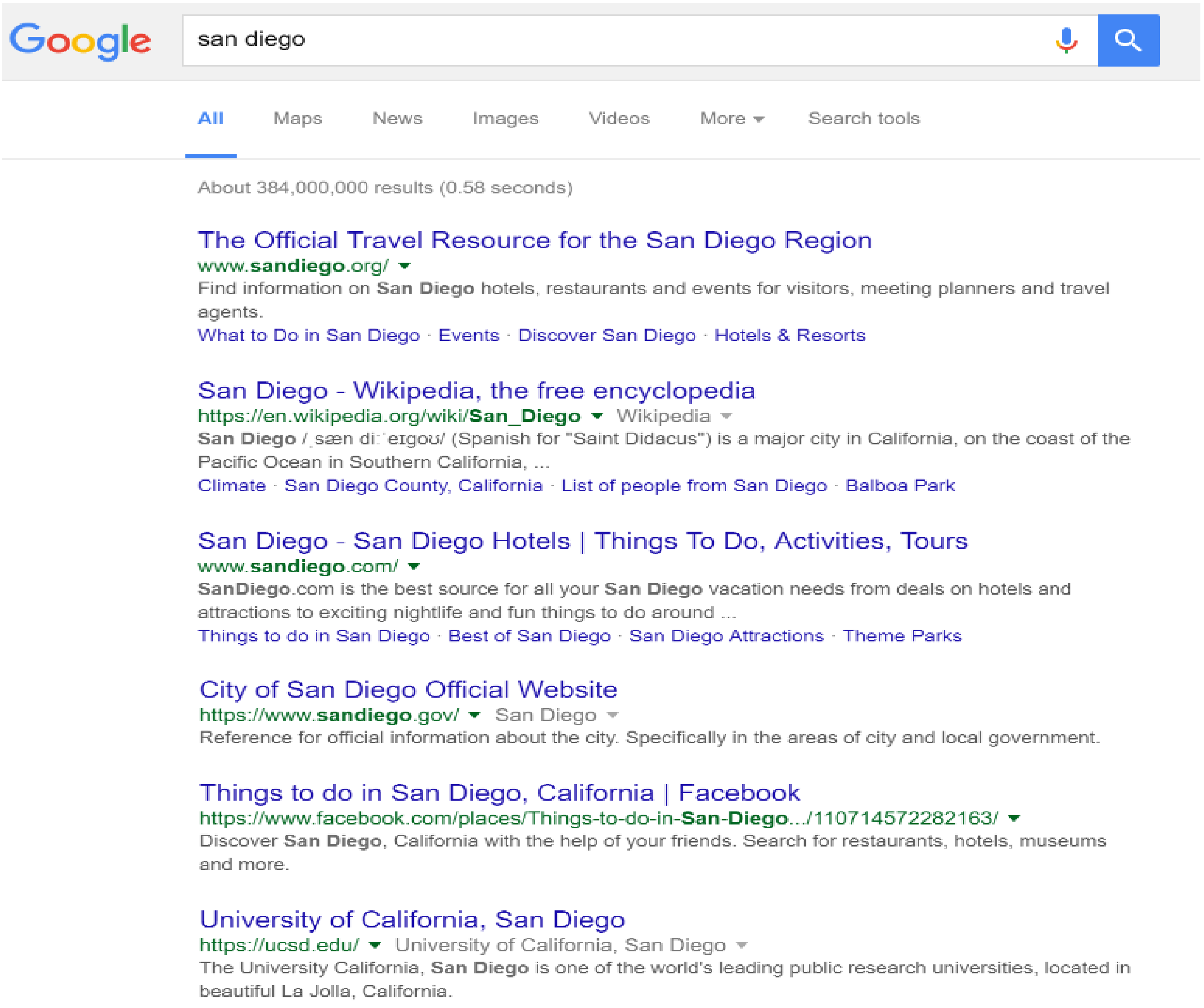
Question Answering with Knowledge Bases, Web and Beyond

Scott Wen-tau Yih & Hao Ma

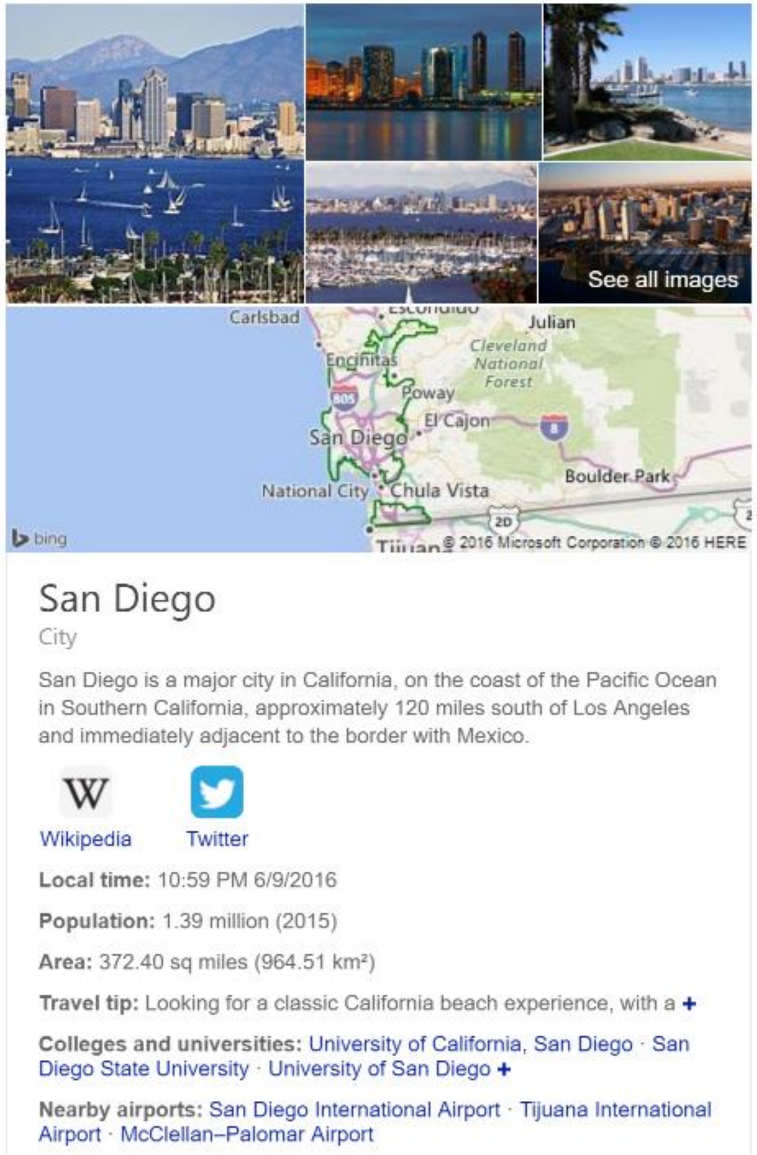
July 17th, 2016

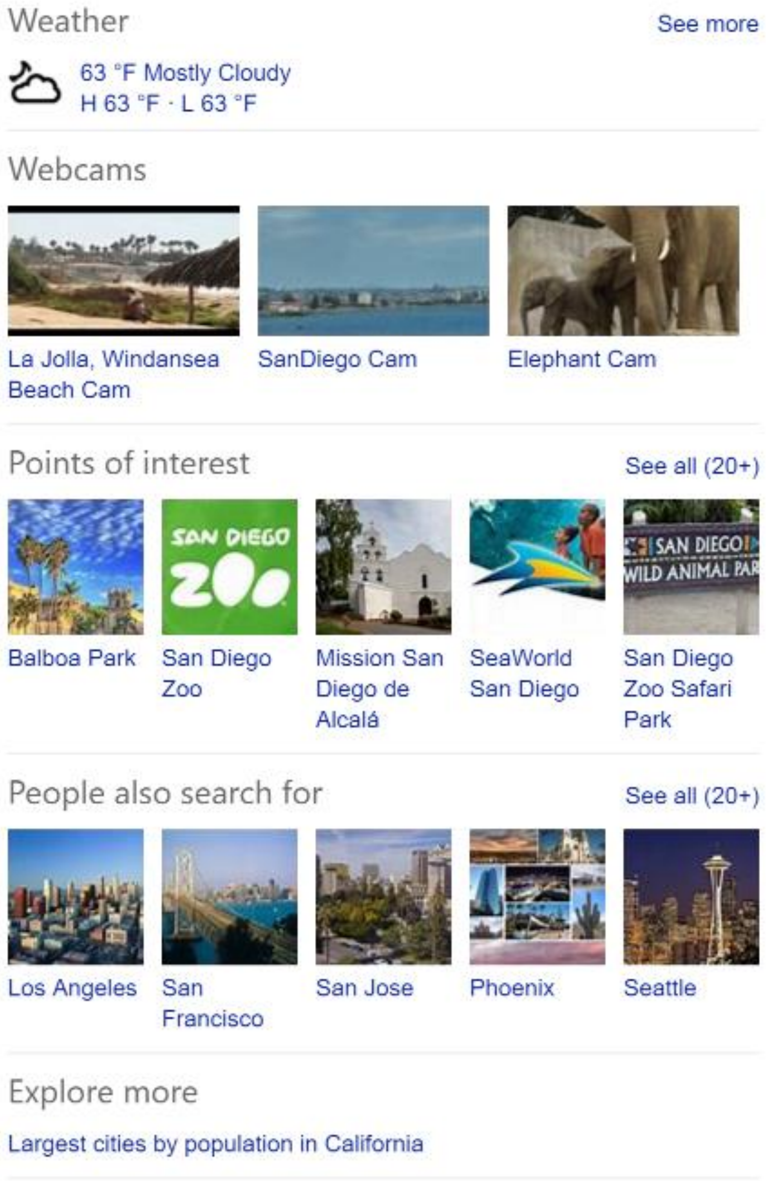
Search Engine Evolves



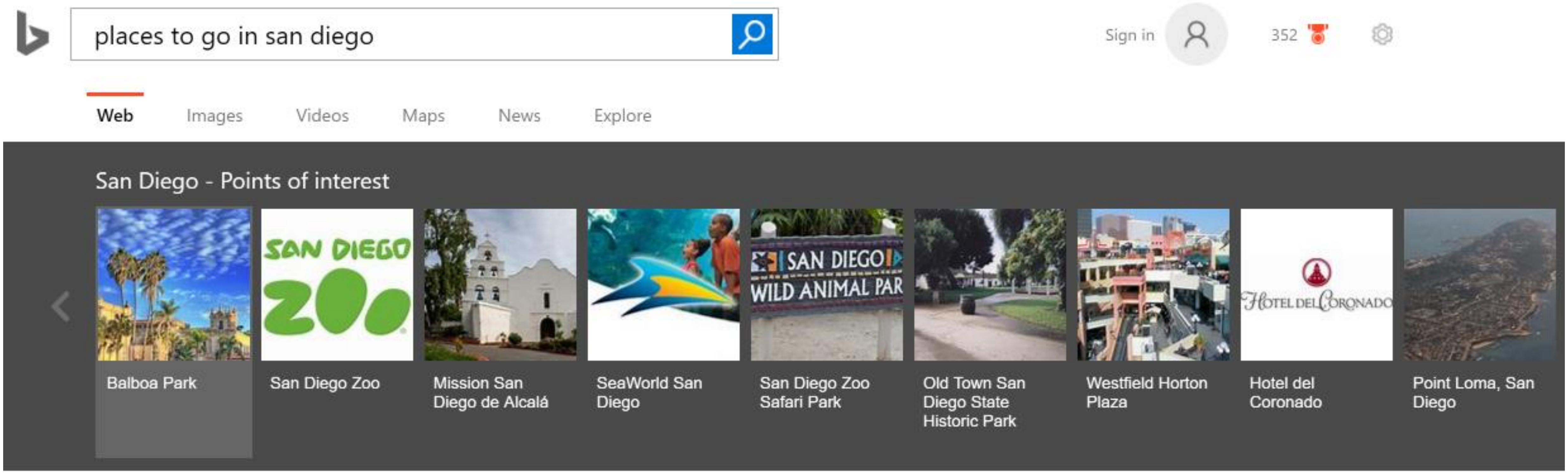


Search Engine Evolves

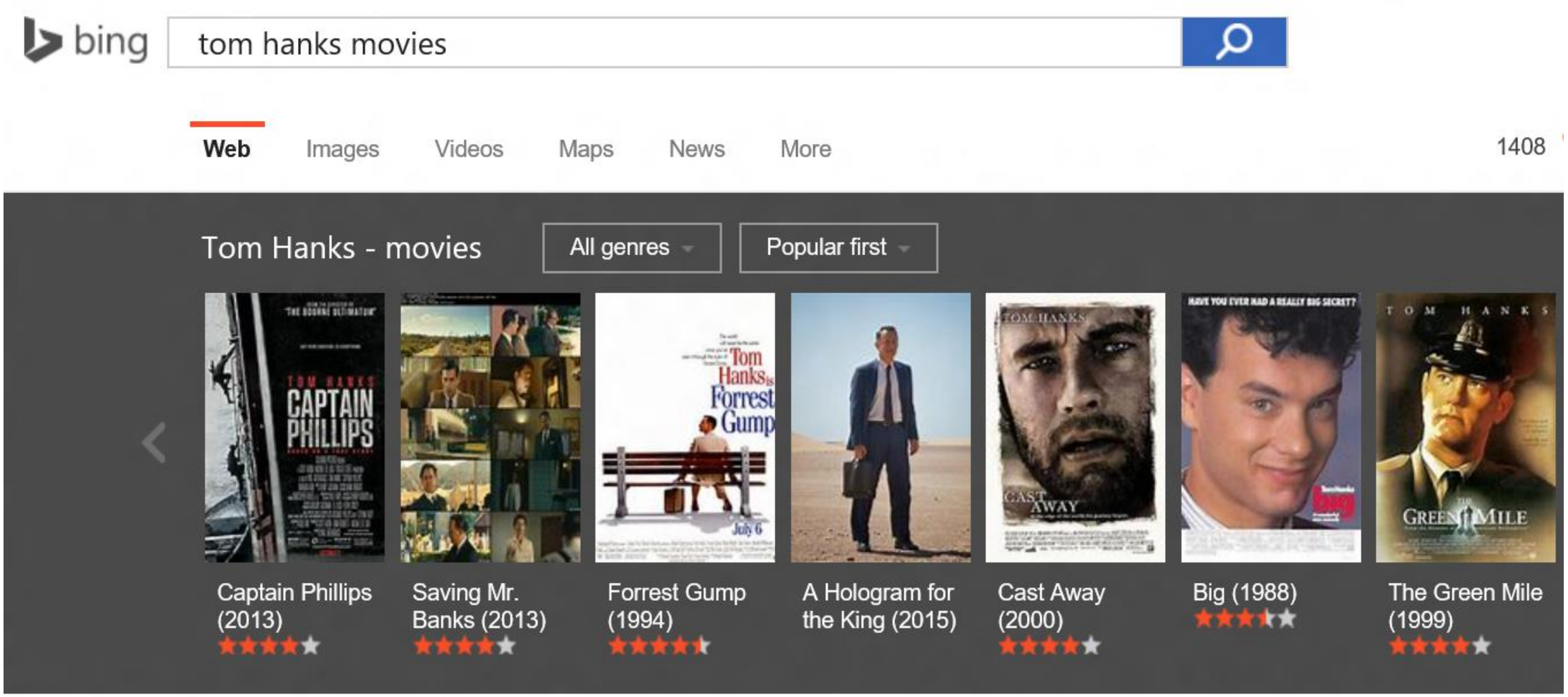




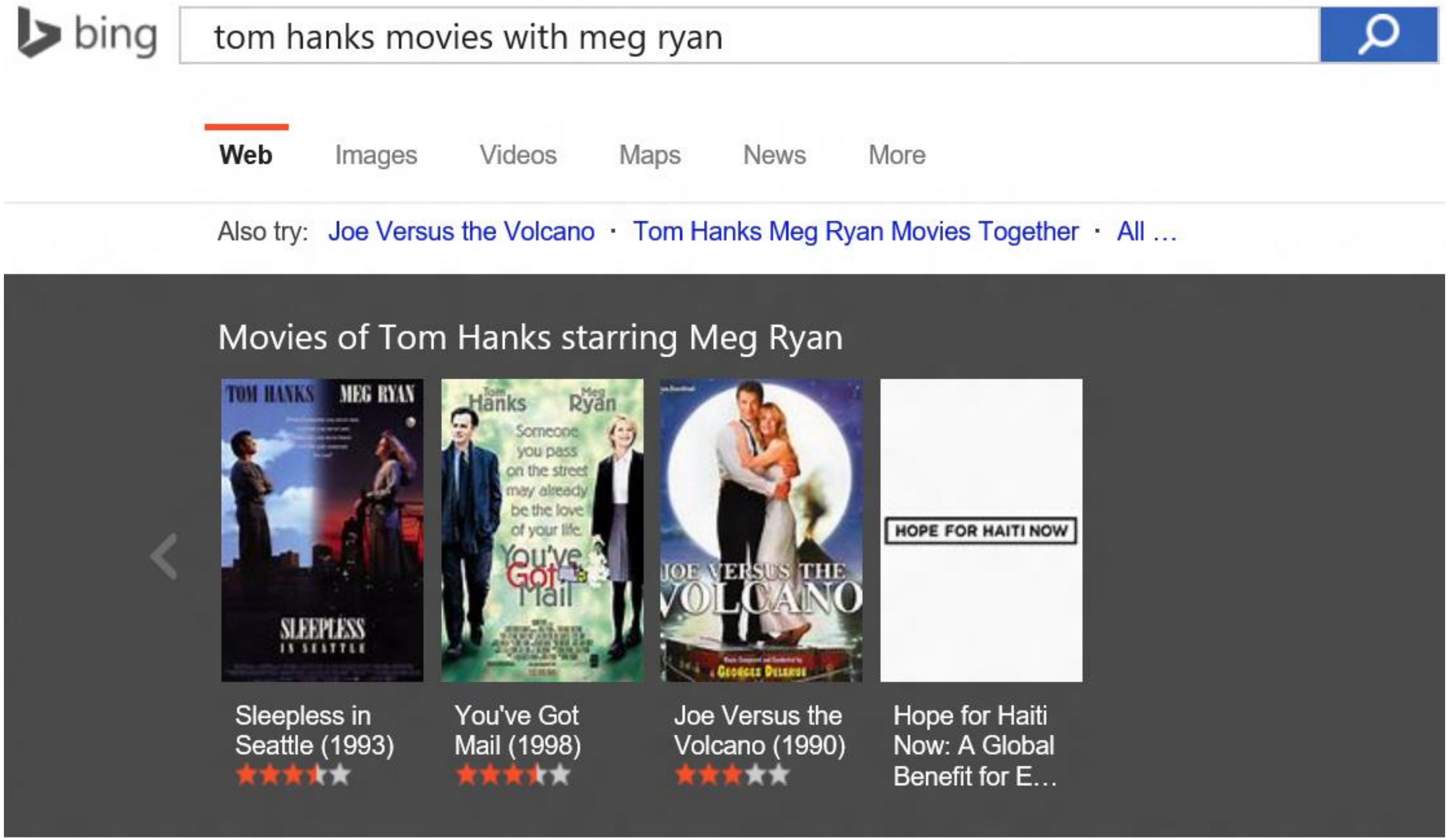
Search Engine Evolves



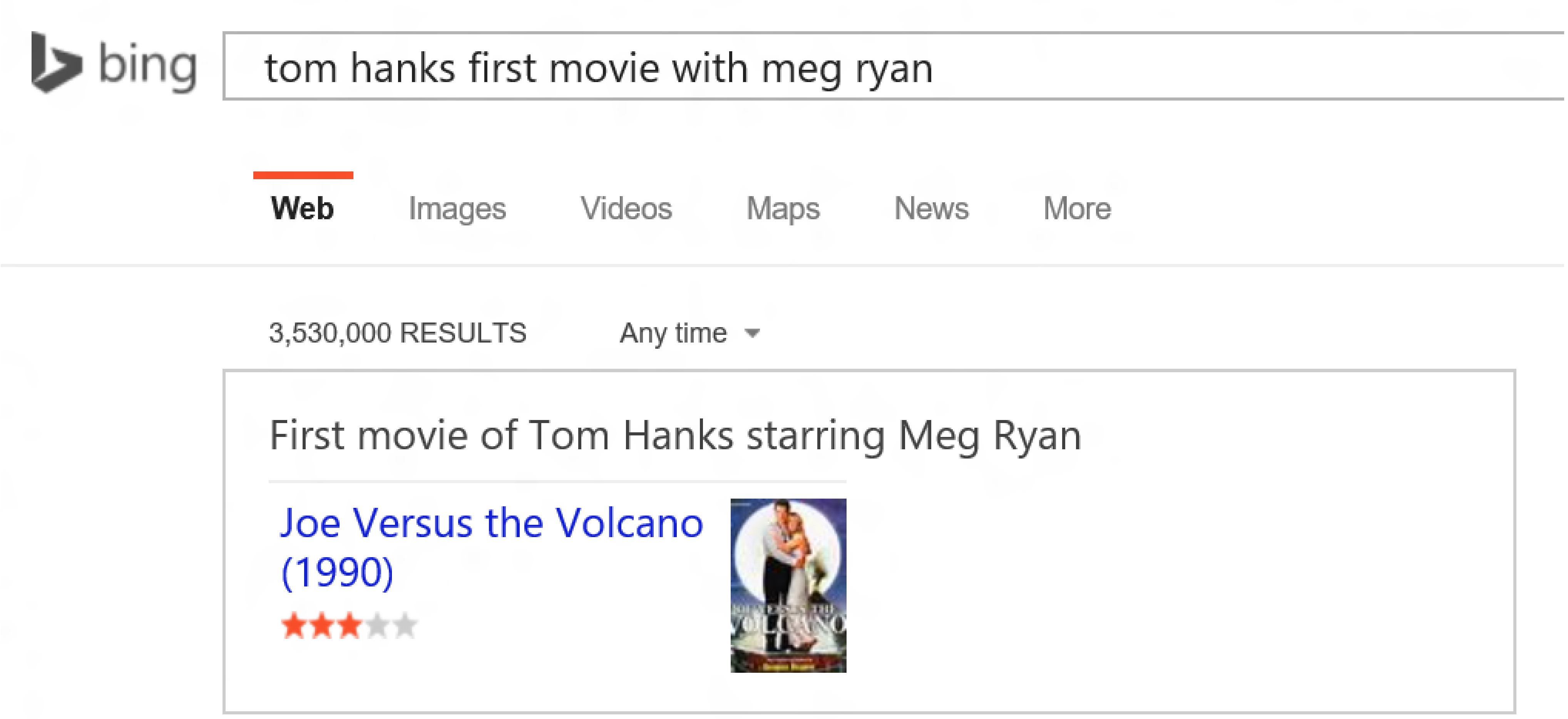
Question and Answering in Modern Search Engines



Question and Answering in Modern Search Engines



Question and Answering in Modern Search Engines



Question and Answering in Modern Search Engines

bing director of tom hanks first movie with meg ryan

Web Images Videos Maps News More

1,620,000 RESULTS Any time

Director of first movie of Tom Hanks starring Meg Ryan

John Patrick Shanley

Joe Versus the Volcano (1990) - IMDb
www.imdb.com/title/tt0099892

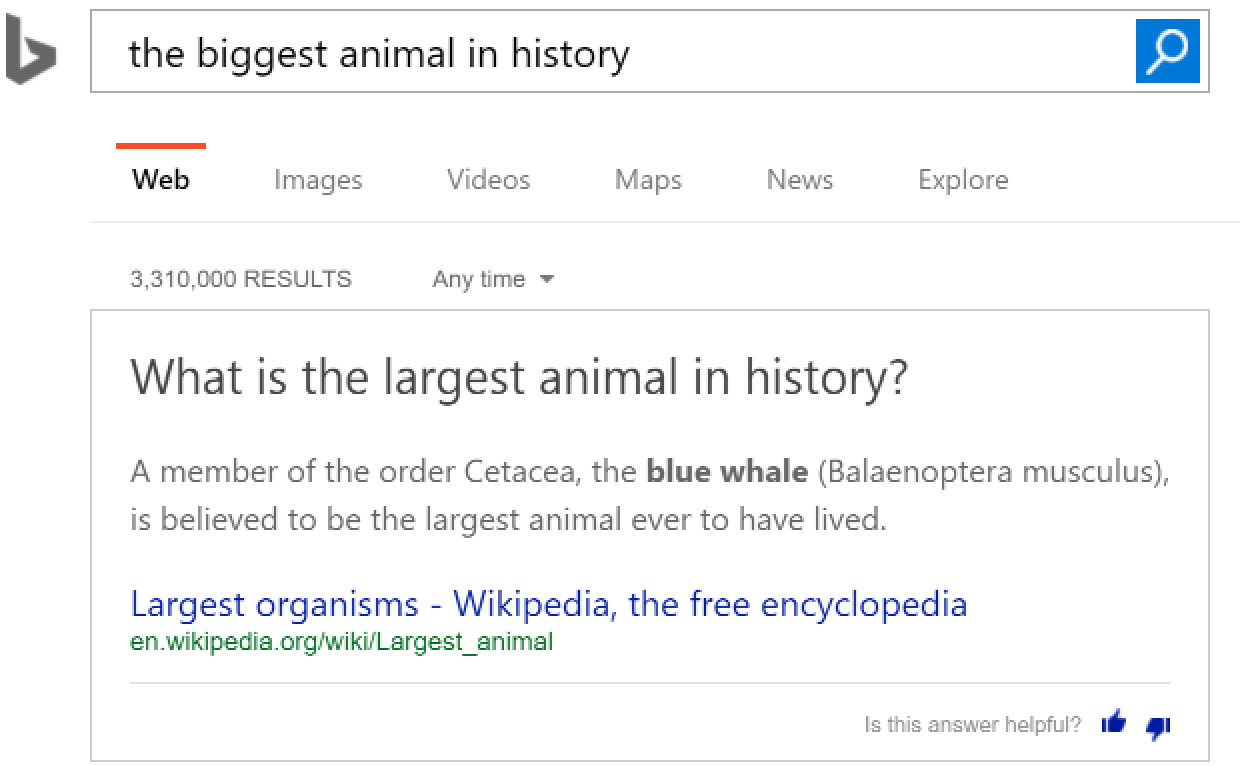
Rating: 5.7/10 · 25,640 ratings · Comedy/Romance · PG · 102 min

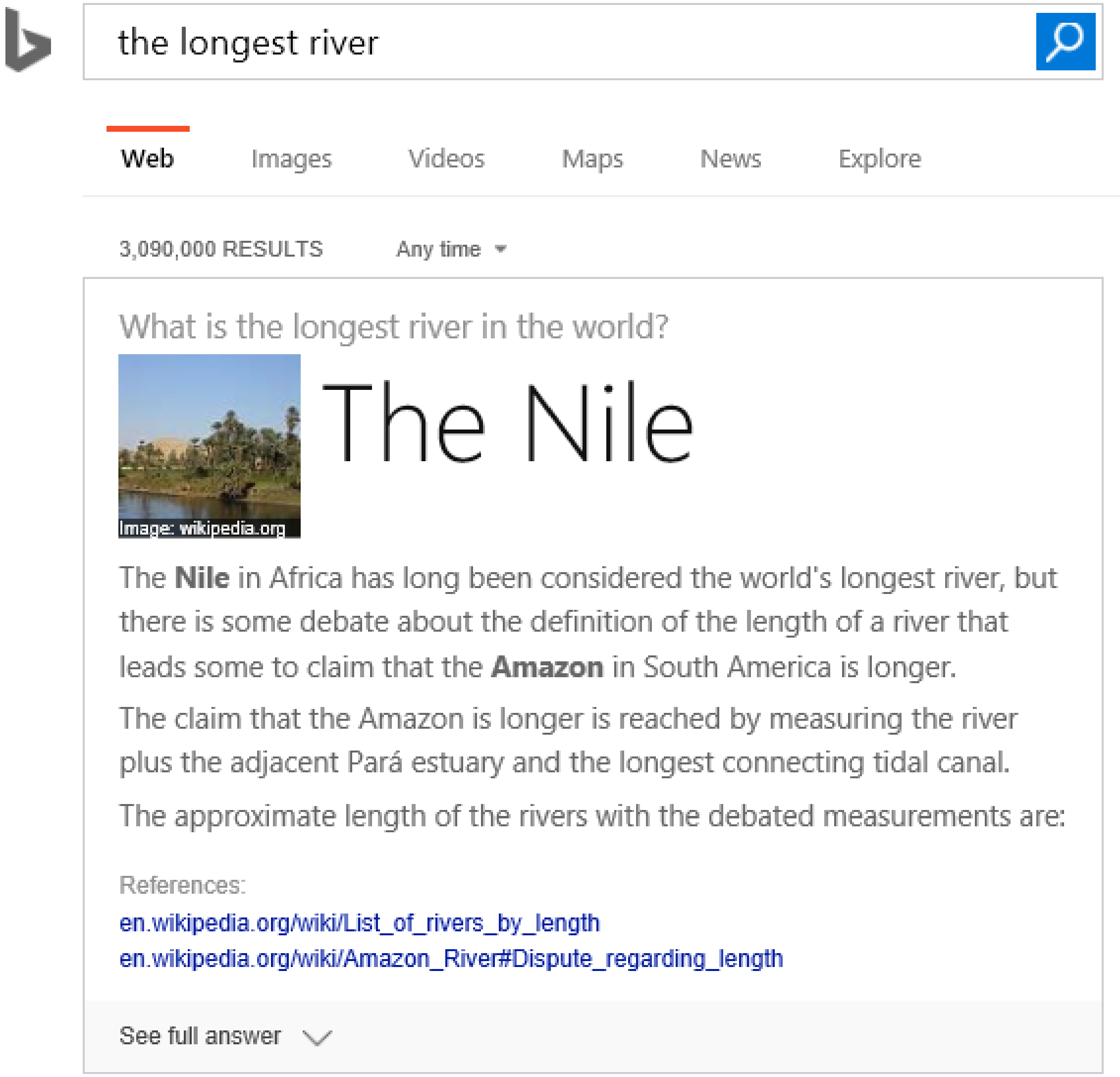
Joe Versus the Volcano PG ... **Director**: **John Patrick Shanley**. Writer: **John Patrick Shanley**. Stars: **Tom Hanks**, **Meg Ryan**, Lloyd Bridges | See full cast and crew »

Meg Ryan Reteams With **Tom Hanks** for Ithaca , Actress Set ...
www.eonline.com/news/505216/meg-ryan-reteams-with-tom-hanks-for...

Jan 29, 2014 · **Meg Ryan** and **Tom Hanks** are teaming ... latest to step into the role of **director**. ... and was instrumental in making the **first film** such a ...

Question and Answering in Modern Search Engines





Question and Answering in Modern Search Engines

nlp conference

All Videos News Images Shopping More Search tools

About 13,300,000 results (0.67 seconds)

(The Old) Computational Linguistics / NLP Conferences Calendar

Main NLP/CL 2016 Conference Deadlines

Conference	Submission Date	Notification Date
ACL 2016 (long papers)	Mar 18	May 24
Interspeech 2016	Mar 23	Jun 10
CoNLL 2016	May 06	Jun 01

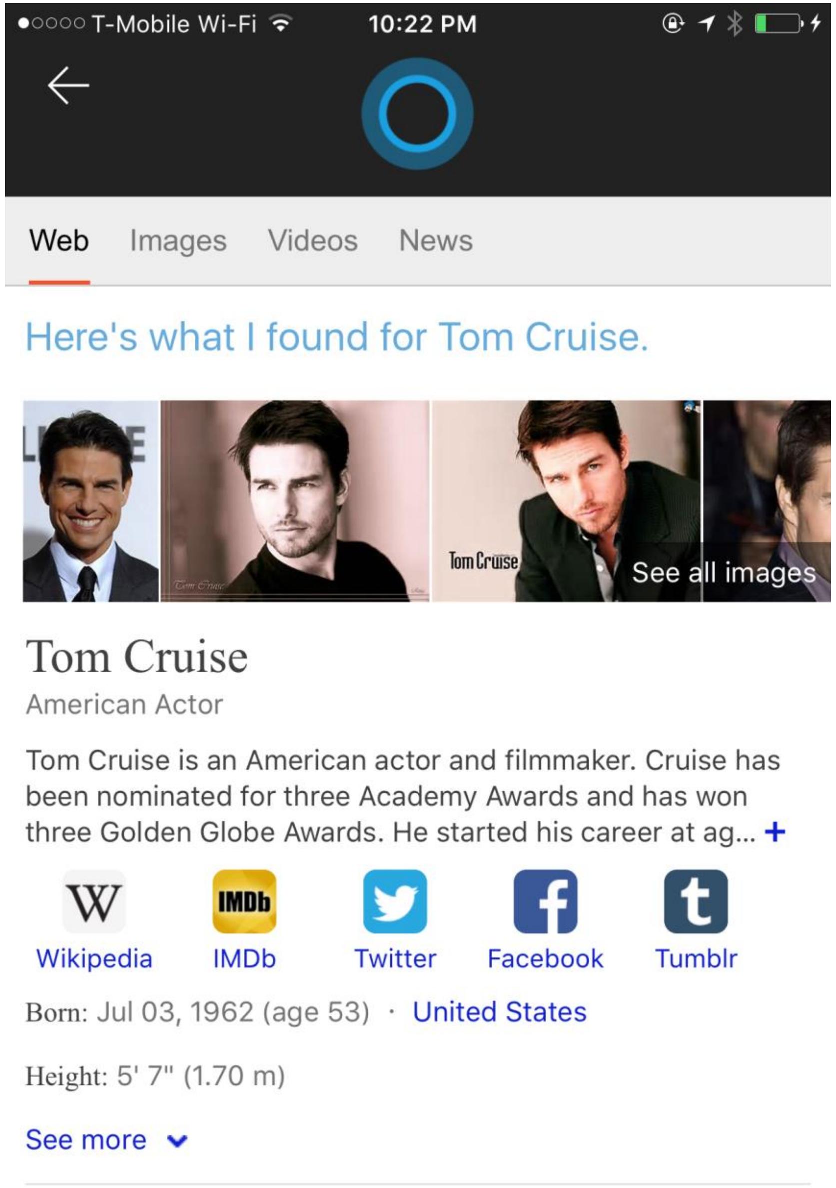
12 more rows, 2 more columns

(The Old) NLP/CL Conference Calendar

https://www.cs.rochester.edu/~tetreaul/conferences.html University of Rochester

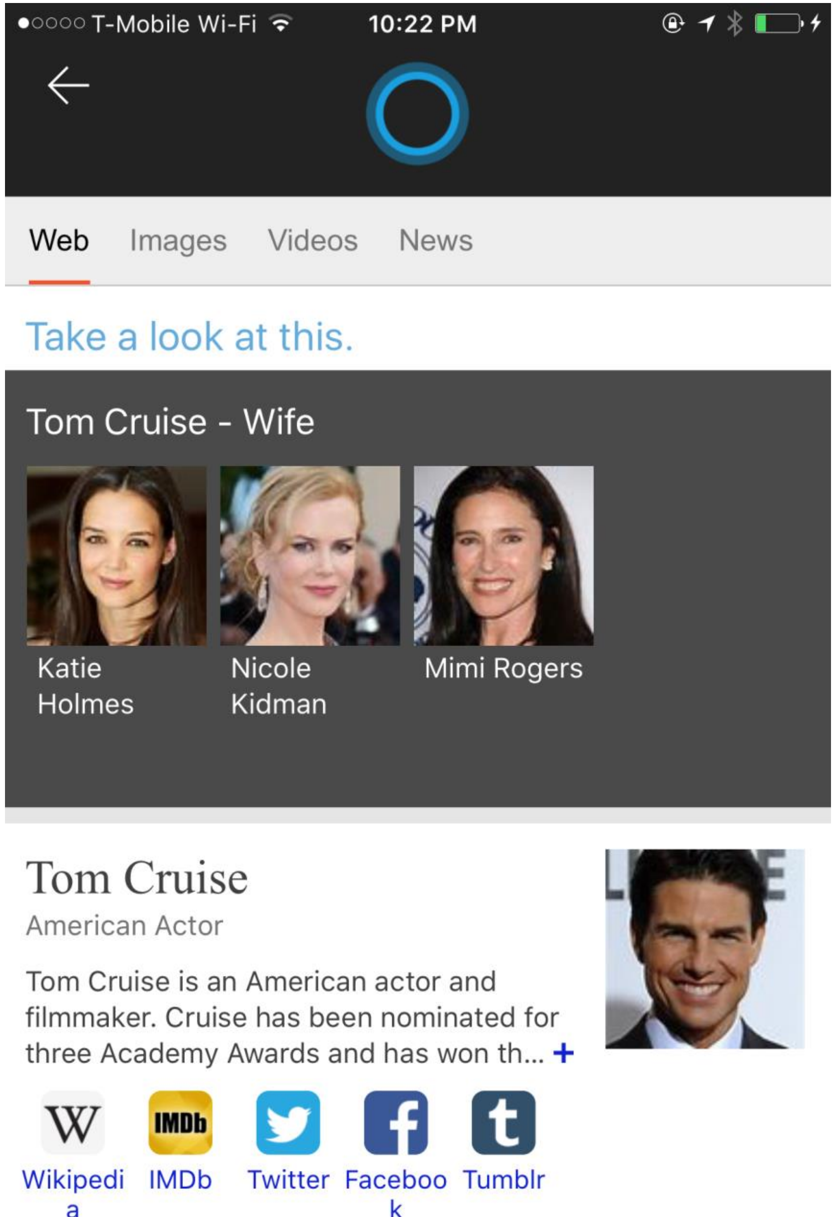
Conversational Question Answering

Tom Cruise



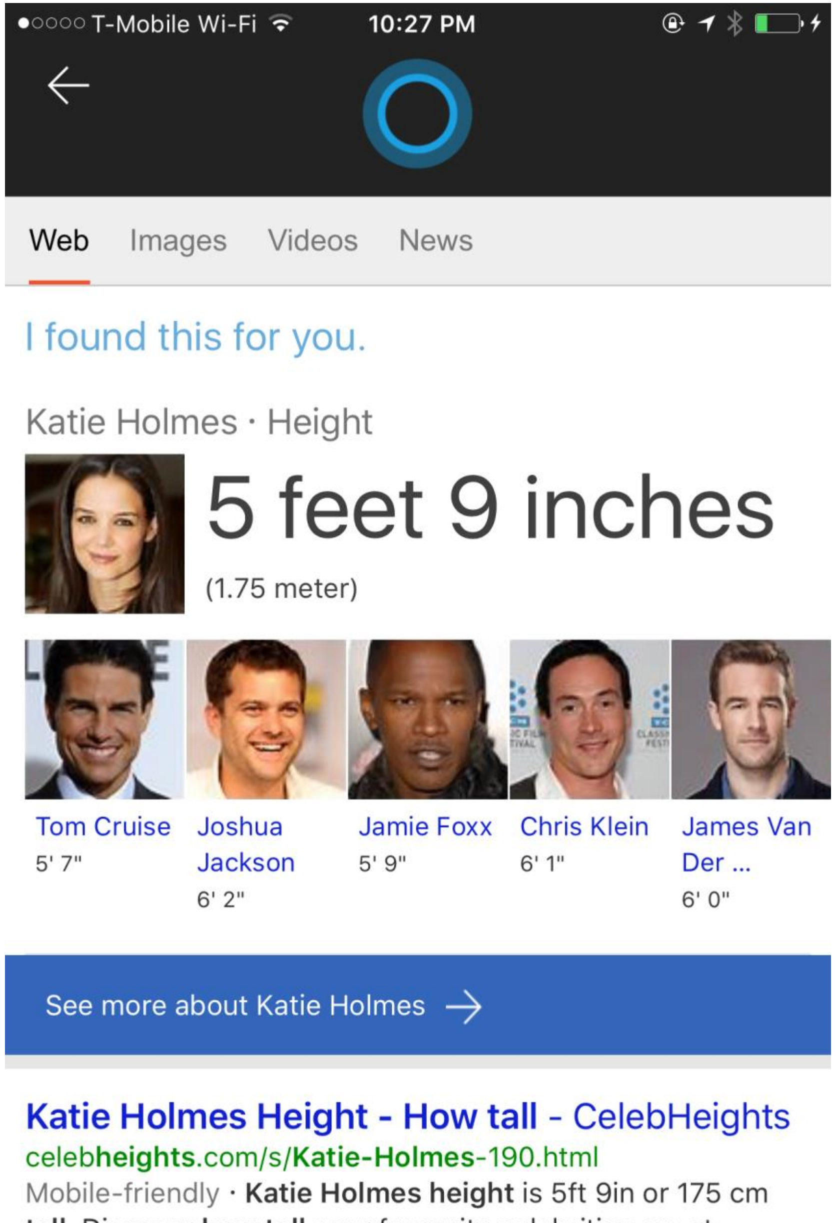
Conversational Question Answering

His Wife



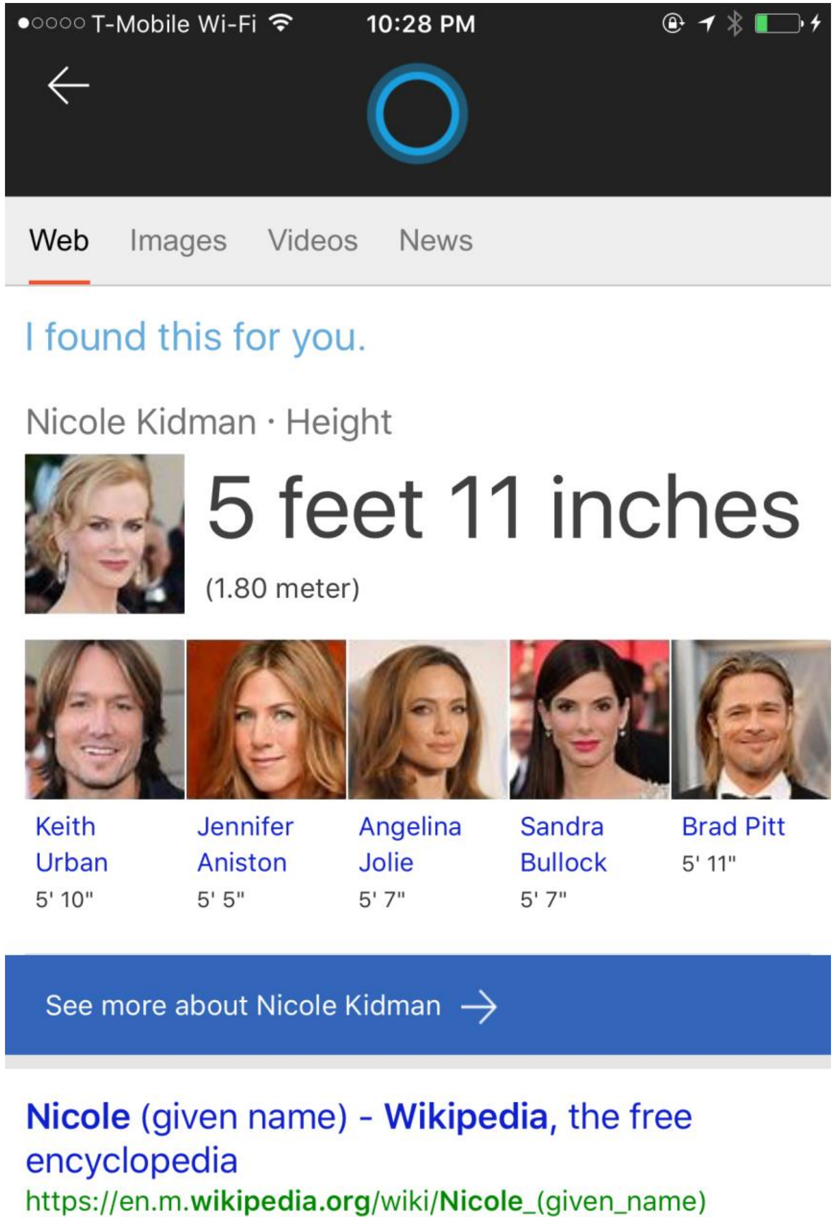
Conversational Question Answering

How tall is Katie



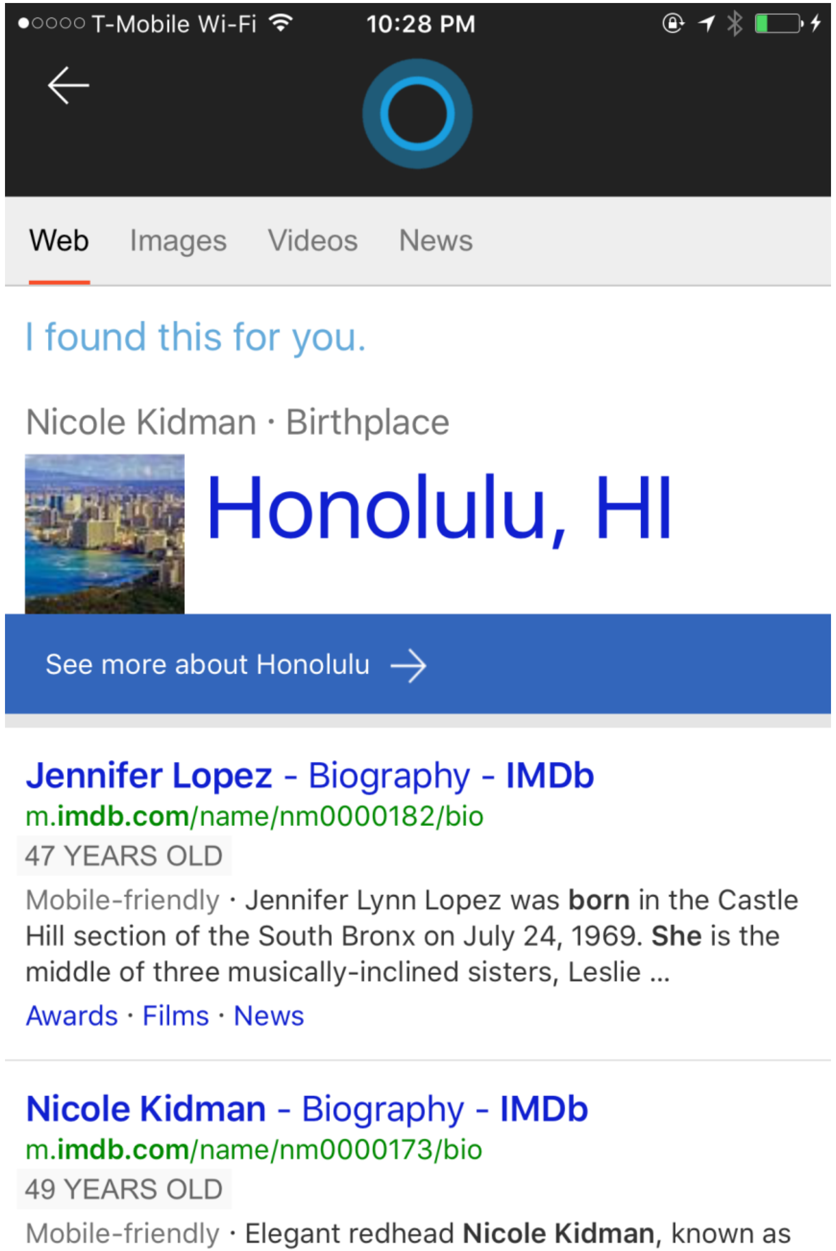
Conversational Question Answering

How about Nicole

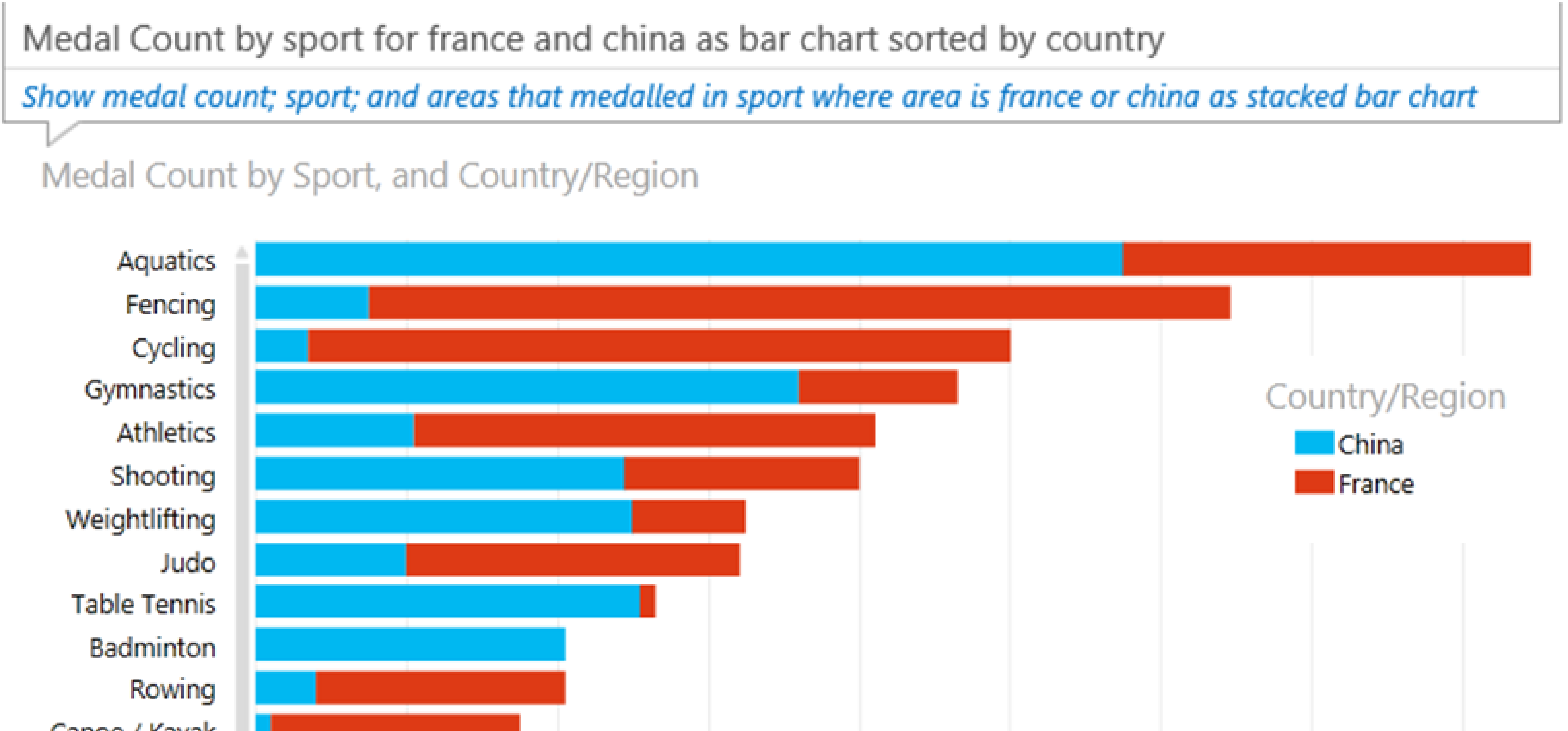


Conversational Question Answering

Where was she born



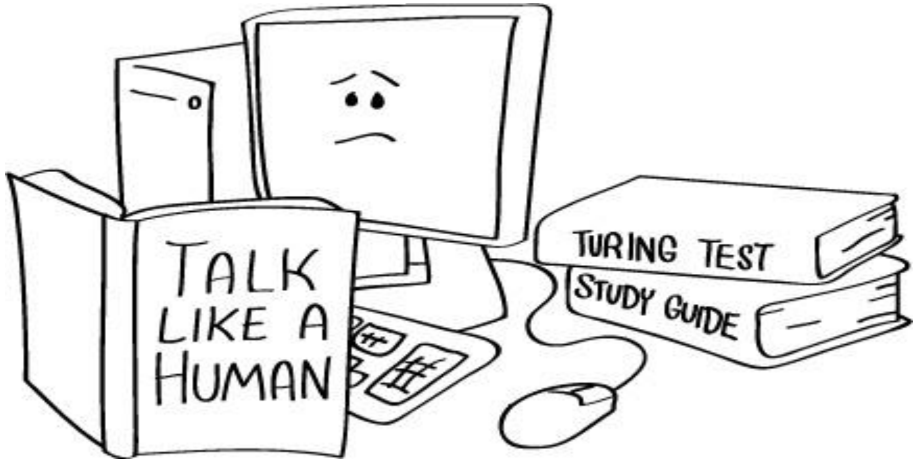
Power BI Natural Language Q&A



Power BI Natural Language Q&A

Natural Language Understanding

- Question-answering machine [Simmons CACM-65]
  - General-purpose language processors that communicate with users in natural language (e.g., English)
  - Deal with statements and/or questions



http://csunplugged.org/turing-test

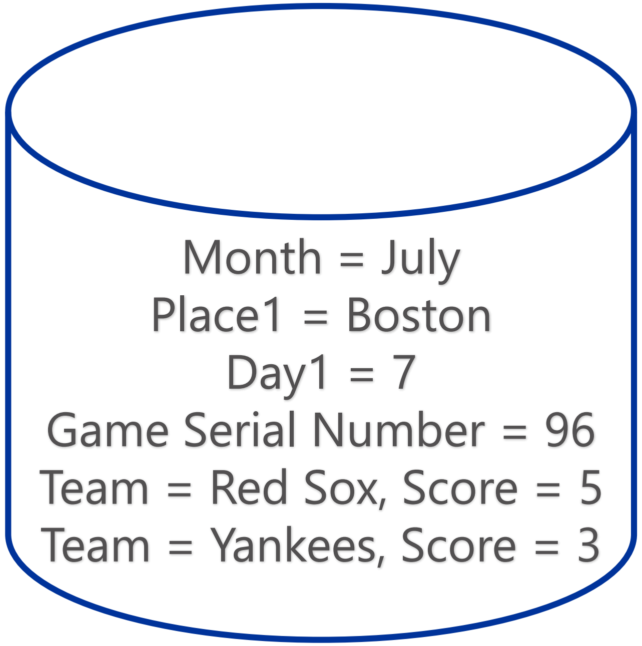
Categories of (Early) QA Systems

- List-structured database systems
  - Organizing knowledge (e.g., kinship) in list DB
- Graphic database systems
  - Map text and graphic data (e.g., pictures, diagrams) to the same logical representations
- Text-based systems
  - Matching questions and text in a corpus to find answers
- Logical inference systems
  - Textual entailment, answering science text book questions & algebra word problems

Baseball [Green, Wolf, Chomsky & Laughery, 1963]

- *How many games did the Yankees play in July?*

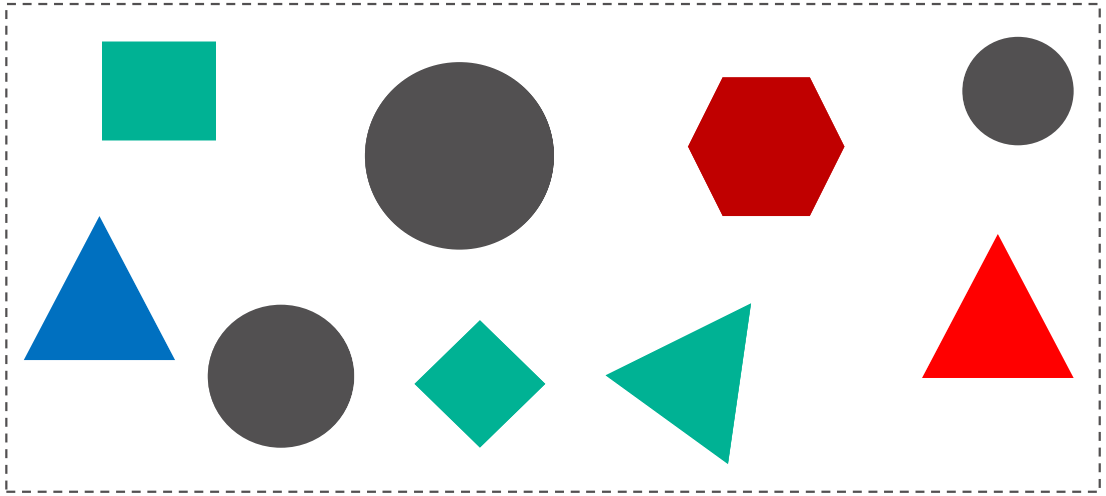
- Step 1: Simple dictionary-based syntactic analysis
  - (How many games) did (the Yankees) play (in (July))?
- Step 2: Semantic analysis that builds "spec"
  - "Who" → ("team" = ?)
  - Conditions (e.g., "winning", "how many") → routines
- Step 3: Execution



Example taken from [Simmons, 1965]

The Picture Language Machine [Krisch, 1964]

- *Is the statement true?*
  All circles are black circles.

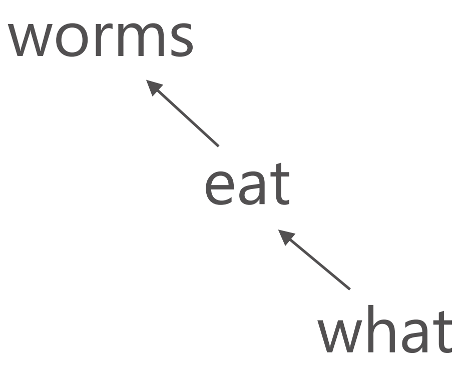


- Both pictures and text are translated into logical language
  - Circle(a), Black(a), Bigger(a, b), Between(a, b, c)
  - $(\forall x)[\mathrm{Circle}(x) \supset (\exists y)[\mathrm{Circle}(y) \wedge \mathrm{Black}(y) \wedge (x = y)]]$

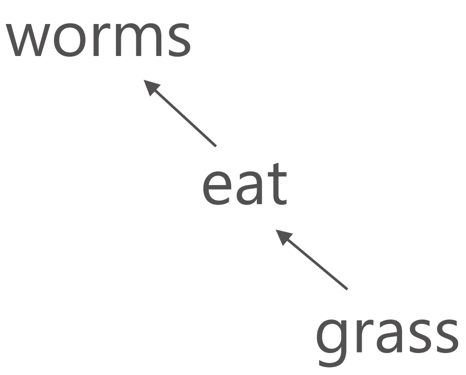
Protosynthex [Simmons+, 1963]
Answer Questions from an Encyclopedia

- Matching questions & text in dependency logic [Hays 1962]

Q: What do worms eat?

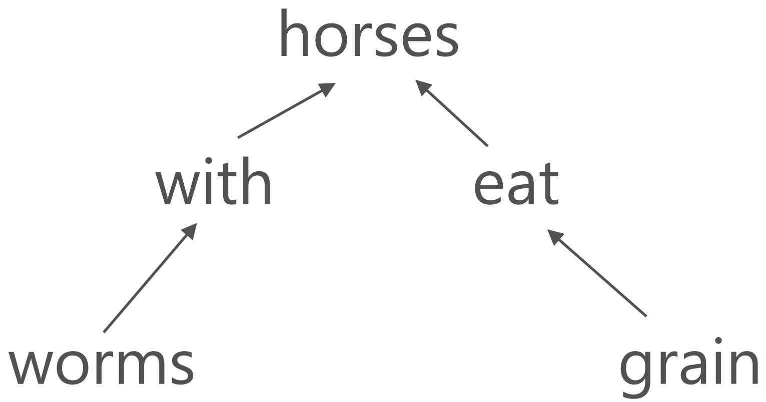


A1: Worms eat grass



Complete Agreement

A2: Horses with worms eat grain



Partial Agreement

Student [Bobrow 1964]

- The first algebra problem solver
  - Translate a set of English statements to mathematical equations
- Step 1: Simplify text and annotate operators
  - “twice” → “two times”, “the square of” → “square”
  - Tag operators like “plus”, “percent”, “times”
- Step 2: Heuristics to break problem into simple sentences
- Step 3: Mapping sentences to equations
  - Rules based on dictionary of words and numbers

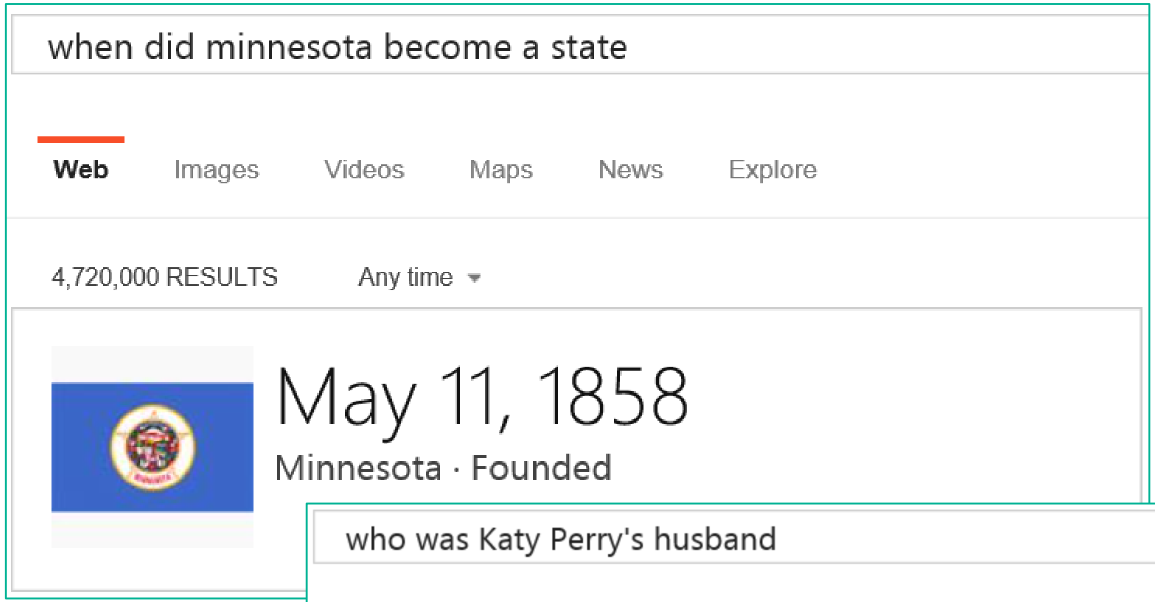
Lessons from Old QA Systems

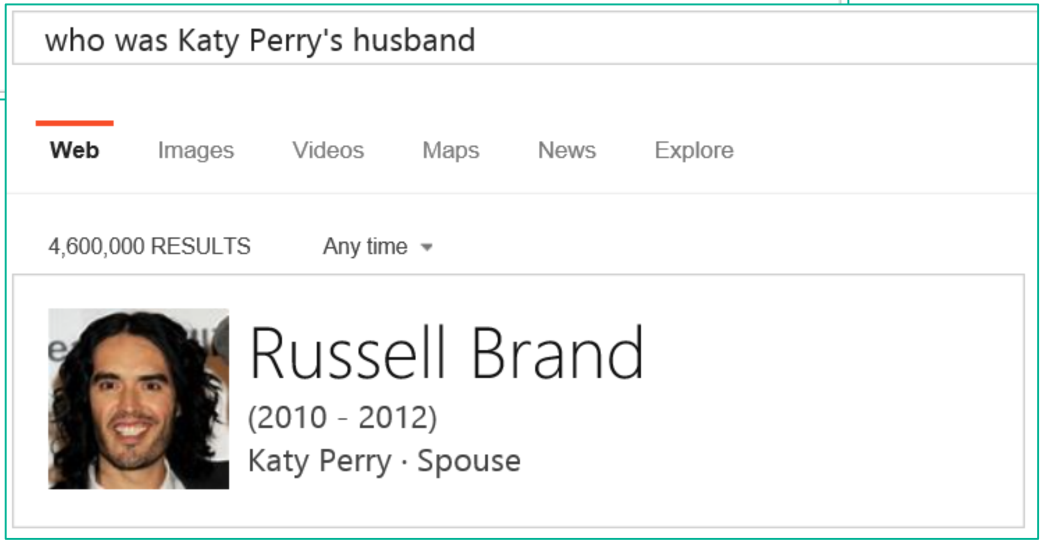
- Limited success
  - Small & limited domains and scopes
    - Often work only on well-controlled, specialized subset of English
  - Not data-driven (e.g., machine learning approaches)
    - Mostly rule-based, potentially brittle
    - Lacks rigorous evaluation
- Open questions [Simmons 1965]
  - Meaning representation & the need of formal languages
  - Syntactic and semantic disambiguation
  - Combine partial answers from various sources

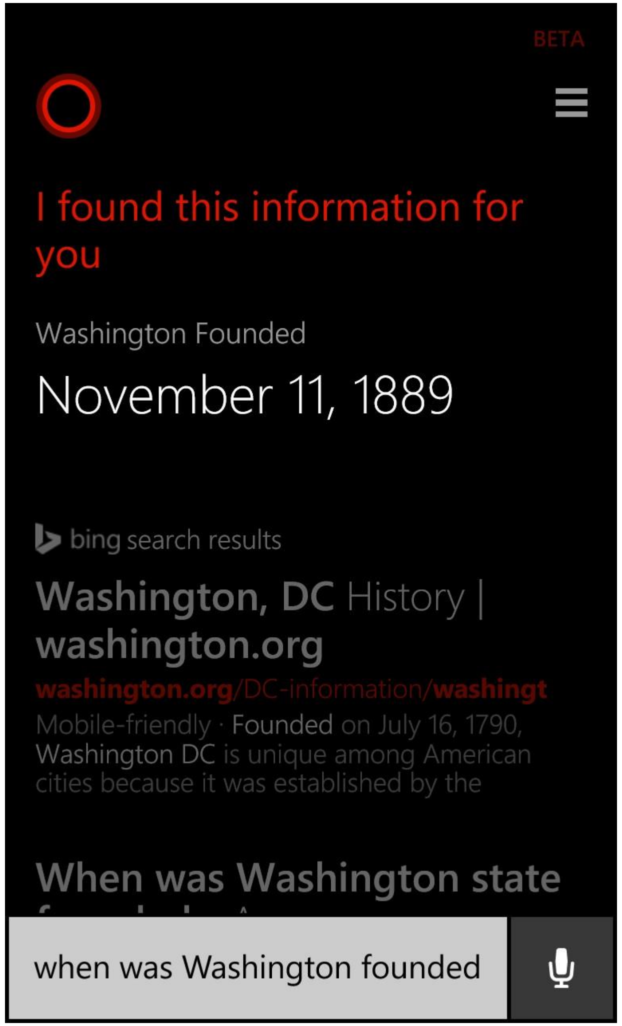
Categories of Modern QA Systems/Problems

- Factoid questions
  - Informational queries about facts of entities
  - Competitions (Jeopardy! & Quiz Bowl)
- Narrative questions
  - Opinion, instructions (how-to questions)
- Multi-modal
  - Visual QA
  - Travel Assistant
- AI ability tests
  - Reading comprehension
  - Elementary School Science and Math Tests

Factoid Questions



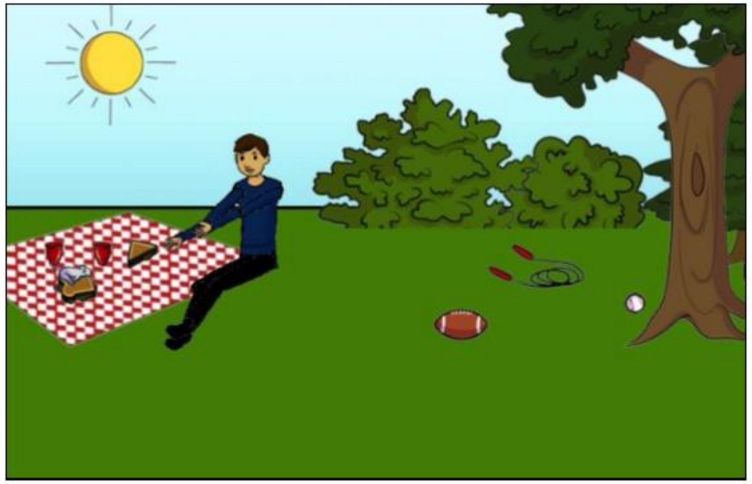




Visual Question Answering [Agrawal et al.]

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Machine Comprehension Test [Richardson+ 2013]

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?

   A) Fries

   B) Pudding

   C) James

   D) Jane

2) What did James pull off of the shelves in the grocery store?

   A) pudding

   B) fries

   C) food

   D) splinters

Data Sources

- Structured data
  - Databases & Knowledge bases
- Semi-structured data
  - Web tables
- Unstructured text
  - Newswire corpora
  - Web

Paradigms

- Semantic parsing
  - Answer questions using knowledge bases
- Information Retrieval
  - Text matching
- Human intelligence
  - Community QA
  - Social QA (I'm an Expert) [Richardson & White, WWW-2011]

General Technological Challenges

- Question analysis
  - Answer type
  - Slot filling
  - Semantic parsing
- Text/Data analysis
- Paraphrasing & Matching
  - Handle variations of questions
  - Ontology matching
- Search complexity

Roadmap

- Question Answering with Knowledge Bases
  - Introduction to modern large-scale knowledge bases
  - Datasets and state-of-the-art approaches
- Question Answering with the Web
  - Problem setting and general system architecture
  - Essential natural language analysis
  - Leveraging additional information sources
- Question Answering for Testing Machine Intelligence
  - Reading comprehension
  - Reasoning questions

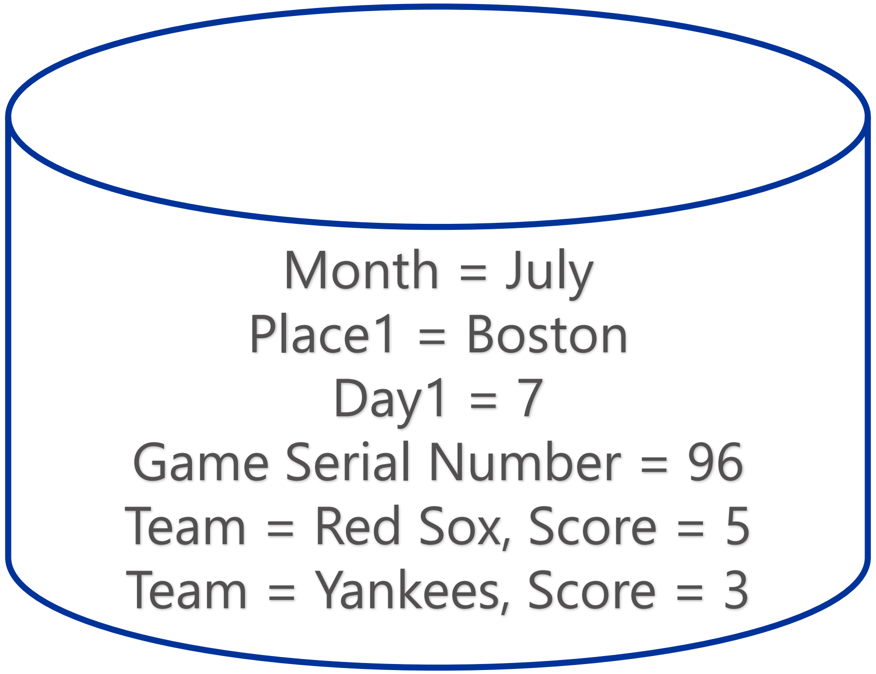
Question Answering with Knowledge Bases

Answer Questions Using Structured Data

- General problem setting
  - Information Source: A “database”
    - Collections of records
    - Tables
    - Large-scale DB with complex schema
  - Input: A natural language question (instead of a formal “query”)
  - Output: Answer

Baseball [Green, Wolf, Chomsky & Laughery, 1963]

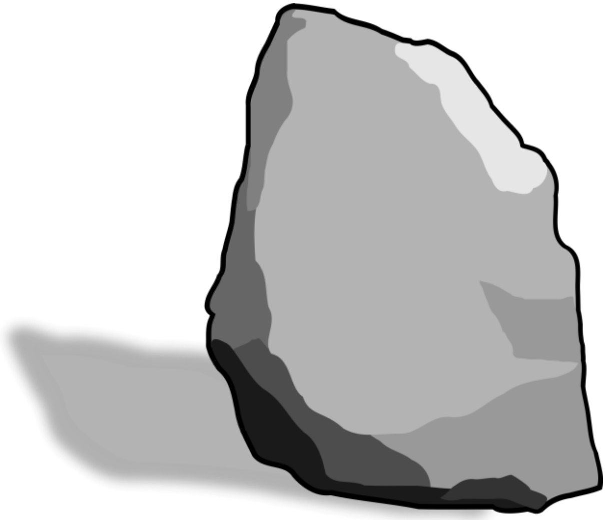
- How many games did the Yankees play in July?

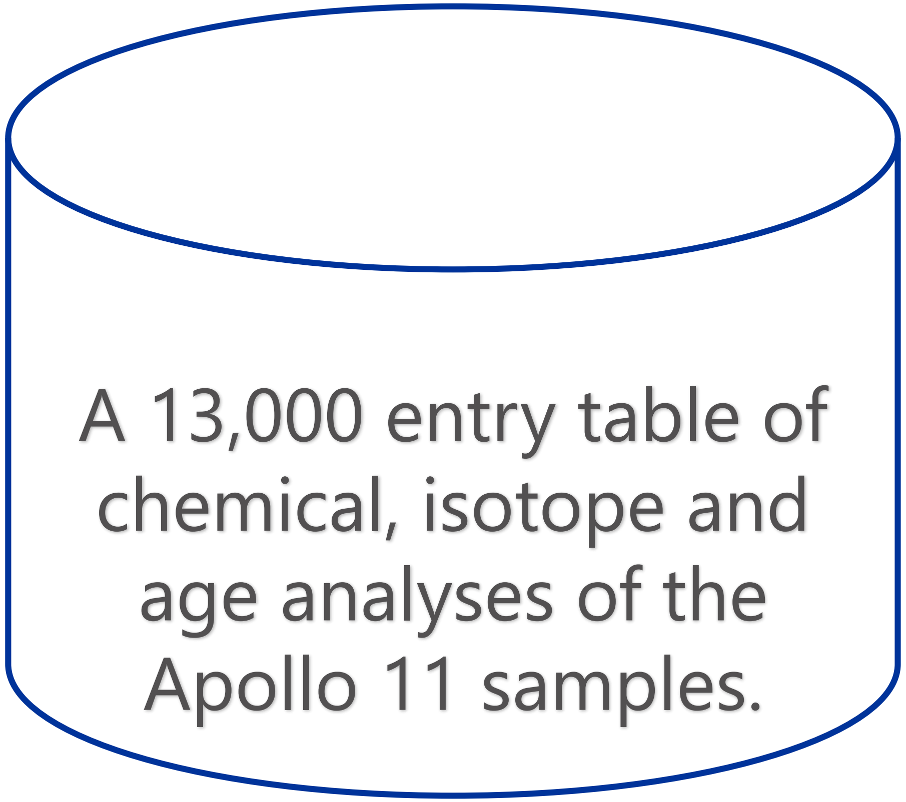


Example taken from [Simmons, 1965]

LUNAR [Woods, 1973]

- Give me all lunar samples with Magnetite.
- How many samples contain Titanium?





Geoquery [Zelle & Mooney, 1996]

- What is the capital of the state with the largest population?
- What are the major cities in Kansas?

Type	Form	Example
country	countryid(Name)	countryid(usa)
city	cityid(Name, State)	cityid(austin,tx)
state	stateid(Name)	stateid(texas)
river	riverid(Name)	riverid(colorado)
place	placeid(Name)	placeid(pacific)

Form	Predicate
capital(C)	C is a capital (city).
city(C)	C is a city.
major(X)	X is major.
place(P)	P is a place.
river(R)	R is a river.
state(S)	S is a state.
capital(C)	C is a capital (city).
area(S,A)	The area of S is A.
capital(S,C)	The capital of S is C.
equal(V,C)	variable V is ground term C.
density(S,D)	The (population) density of S is P
elevation(P,E)	The elevation of P is E.
high_point(S,P)	The highest point of S is P.

Example taken from [Zelle & Mooney, 1996]

Early Work

- Small scale & domain-specific KBs
  - Simple schema
  - Small numbers of entities and relations
  - Limited set of sensible questions
- Approaches
  - Ad-hoc methods (e.g., manually crafting rules) can be quite effective
  - Semantic parsing (of questions)
- Issues
  - Not clear if the methods are scalable
  - Cannot support "open-domain" question answering

Modern Large-scale Knowledge Bases

Google

Knowledge Graph

Microsoft

悟 SATORI

Browse the Knowledge Base!

NELL: Never-Ending Language Learning

Freebase

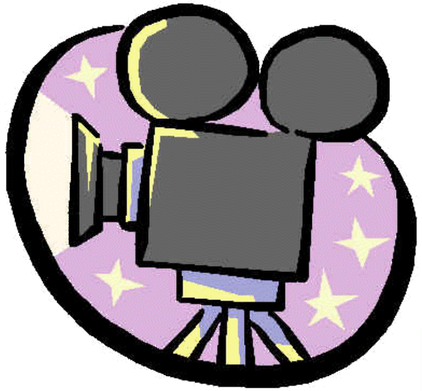
DBpedia

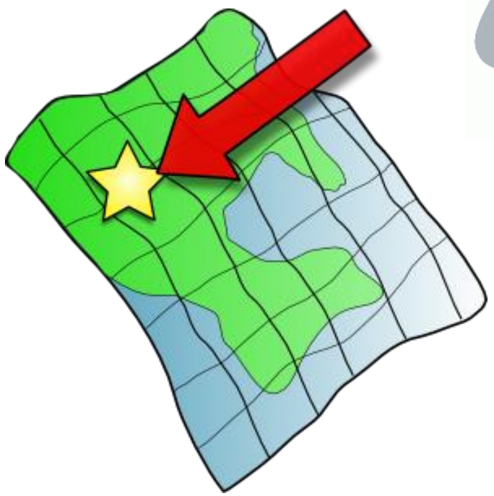
yago select knowledge

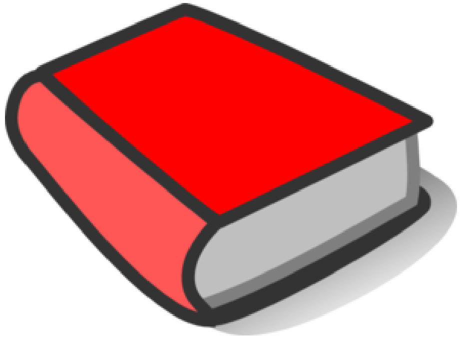
OpenIE
(Reverb, OLLIE)

- Freebase: 46m entities, 2.6b facts
- Microsoft Satori: 852m entities, 18b facts

Entity-centric







Properties & Relations between Entities

Address: 400 Broad St, Seattle, 98109
Phone: (800) 937-9582
Opened: Apr 21, 1962
Height: 605 feet (184.41 m)
Floors: 6

Location

NFL championships: 2013
Head coach: Pete Carroll
Founded: 1976
Division: NFC West

Home Field

Seattle

Population: 652,405 (2013)
Area: 142.55 sq miles (369.20 km²)
Mayor: Ed Murray

Founded: Mar 30, 1971 · Pike Place Market
Customer service: +1 800-782-7282
CEO: Howard Schultz
Founders: Jerry Baldwin · Zev Siegl · Gordon Bowker

Headquarters

Subject-Predicate-Object Triples in Freebase

Seattle Seahawks

Pete Carroll

m.070xg, american_football/football_team/current_head_coach, m.02ttv2

Representing Multi-argument Relations

- Seattle Seahawks – sports.sports_team.roster

Player	Number	Position	From	To
Russell Wilson	3	Quarterback	2012	-
Alan Branch	99	Defensive tackle	2011	2012
Marshawn Lynch	24	Running back	2010	2016
Richard Sherman	25	Cornerback	2011	-
		…		

Representing Multi-argument Relations

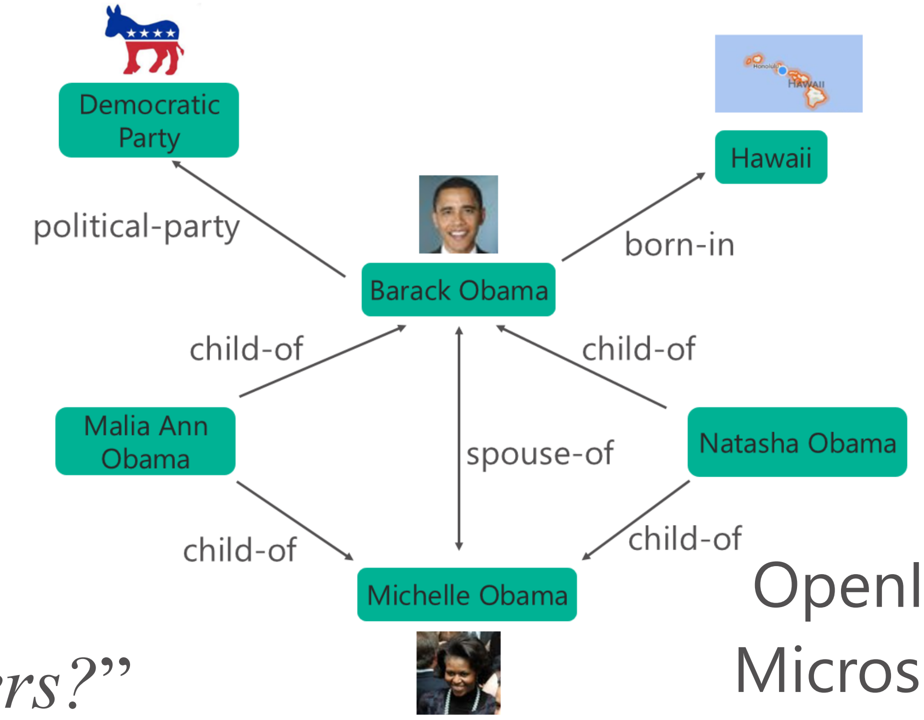
- Seattle Seahawks – sports.sports_team.roster

	Player	Number	Position	From	To
CVT1	Russell Wilson	3	Quarterback	2012	-
CVT2	Alan Branch	99	Defensive tackle	2011	2012
CVT3	Marshawn Lynch	24	Running back	2010	2016
CVT4	Richard Sherman	25	Cornerback	2011	-
			…		

- Compound Value Type (CVT) Nodes
  - Seattle Seahawks – sports/sports_team/roster – CVT1
  - CVT1 – sports/sports_team_roster/player – Russel Wilson
  - CVT1 – sports/sports_team_roster/number – 3

Question Answering with Knowledge Base

- Large-scale Knowledge Base
  - Properties of billions of entities
  - Plus relations among them



Freebase
DBpedia
YAGO
NELL
OpenIE/ReVerb
Microsoft Satori

- Question Answering

"What are the names of Obama's daughters?"

$$\lambda x.parent(Obama, x) \wedge gender(x, Female)$$

WebQuestions Dataset [Berant+ 13]

- *What character did Natalie Portman play in Star Wars?* ⇒ Padme Amidala
- *What currency do you use in Costa Rica?* ⇒ Costa Rican colon
- *What did Obama study in school?* ⇒ political science
- *What do Michelle Obama do for a living?* ⇒ writer, lawyer
- *What killed Sammy Davis Jr?* ⇒ throat cancer

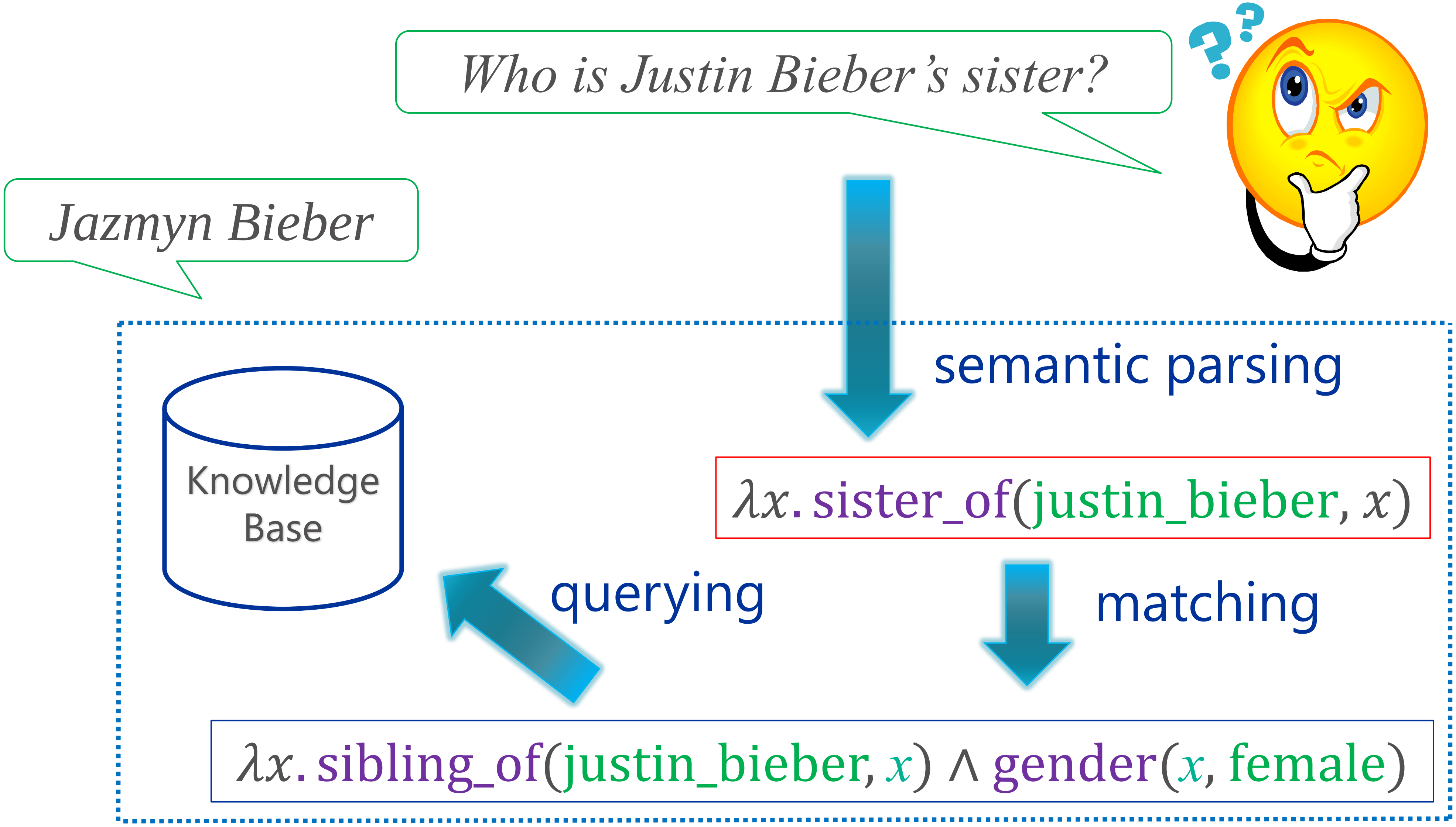
[Examples from Berant]

- 5,810 questions crawled from Google Suggest API and answered using Amazon MTurk
  - 3,778 training, 2,032 testing
  - A question may have multiple answers → using Avg. F1 (~accuracy)

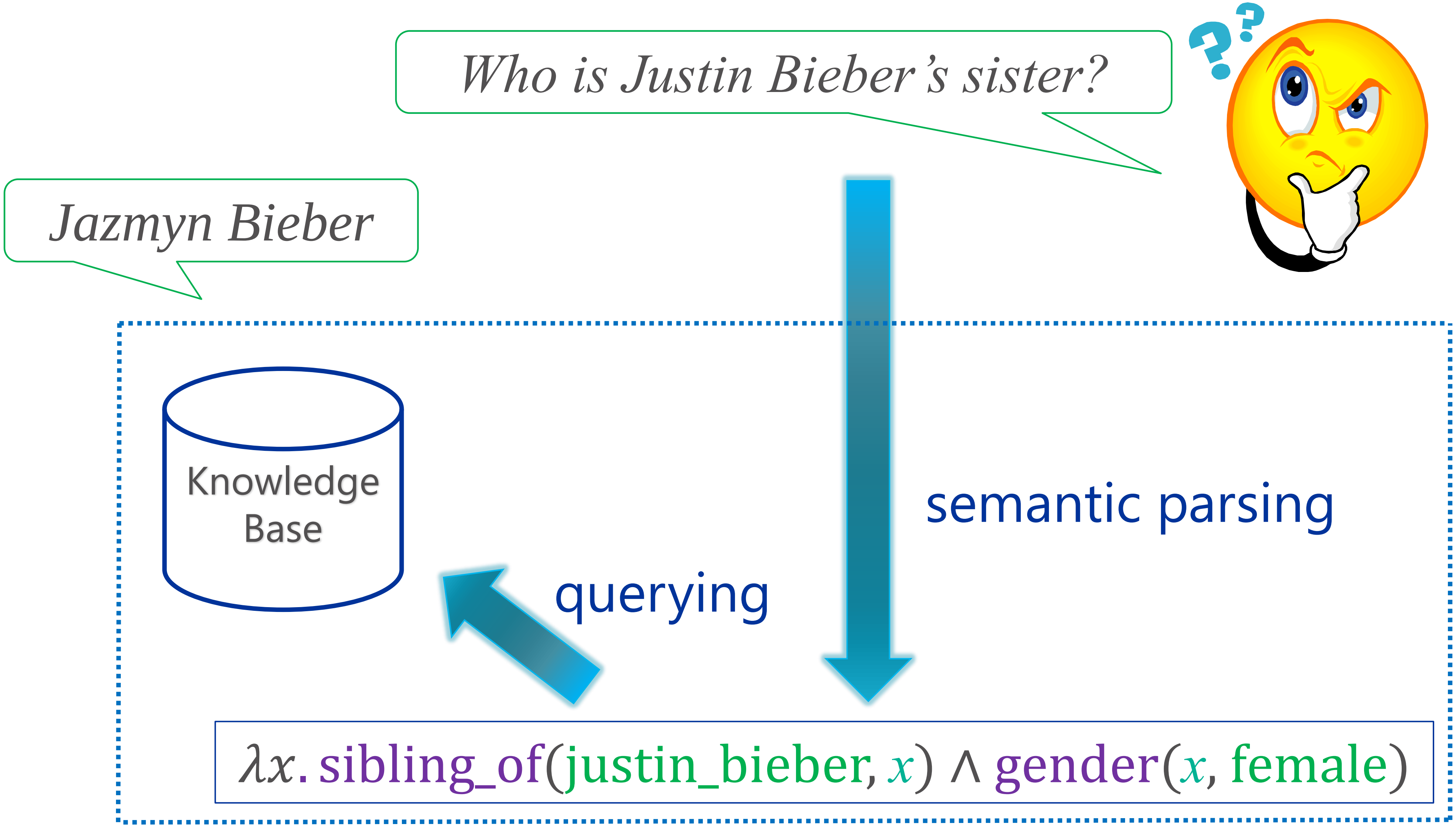
Approaches

- Semantic Parsing
  - Generic semantic parsing and then ontology matching
  - KB-specific semantic parsing
- Information Extraction
- Embedding

Generic Semantic Parsing (e.g., [Kwiatkowski+ 13])



KB-Specific Semantic Parsing (e.g., [Berant+ 13])



Key Challenges

- Language mismatch
  - Lots of ways to ask the same question

    "Who played the role of Meg on Family Guy?"

    "What is the name of the actress for Meg on Family Guy?"

    "In the TV show Family Guy, who is the voice for Meg?"
  - Need to map questions to the predicates defined in KB

    tv.tv_program.regular_cast – tv.regular_tv_appearance.actor
- Large search space
  - Some Freebase entities have >160,000 immediate neighbors
- Compositionality
  - *"What movies are directed by the person who won the most Academy and Golden Globe awards combined?"*

SEMPRE – λ-DCS [Liang, 2013]

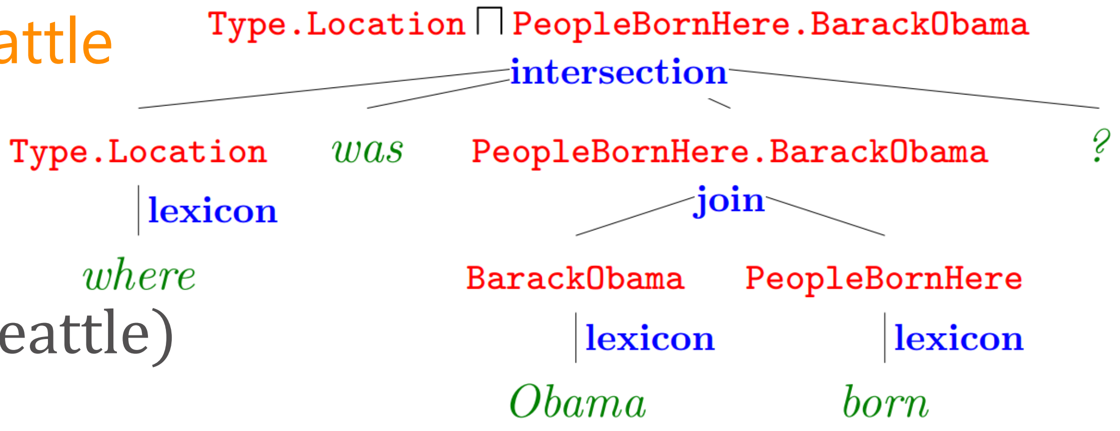
- λ-DCS (lambda dependency-based compositional semantics)
  - Utterance: *"people who have lived in Seattle"*
  - Logical form (lambda calculus): $\lambda x.\exists e.\texttt{PlacesLived}(x,e) \wedge \texttt{Location}(e, \texttt{Seattle})$
  - Logical form (lambda DCS): $\texttt{PlacesLived.Location.Seattle}$

➢ Unary: Seattle $\lambda x.\,[x = \text{Seattle}]$

➢ Binary: PlaceOfBirth $\lambda x.\,\lambda y.\,\text{PlaceOfBirth}(x, y)$

➢ Join: "people born in Seattle" PlaceOfBirth.Seattle
$\lambda x.\,\text{PlaceofBirth}(x, \text{Seattle})$

➢ Intersection: "scientists born in Seattle"
Profession.Scientist ⊓ PlaceOfBirth.Seattle
$\lambda x.\,\text{Profession}(x, \text{Scientist}) \wedge \text{PlaceOfBirth}(x, \text{Seattle})$



SEMPRE – Bridging [Berant et al., EMNLP-2013]

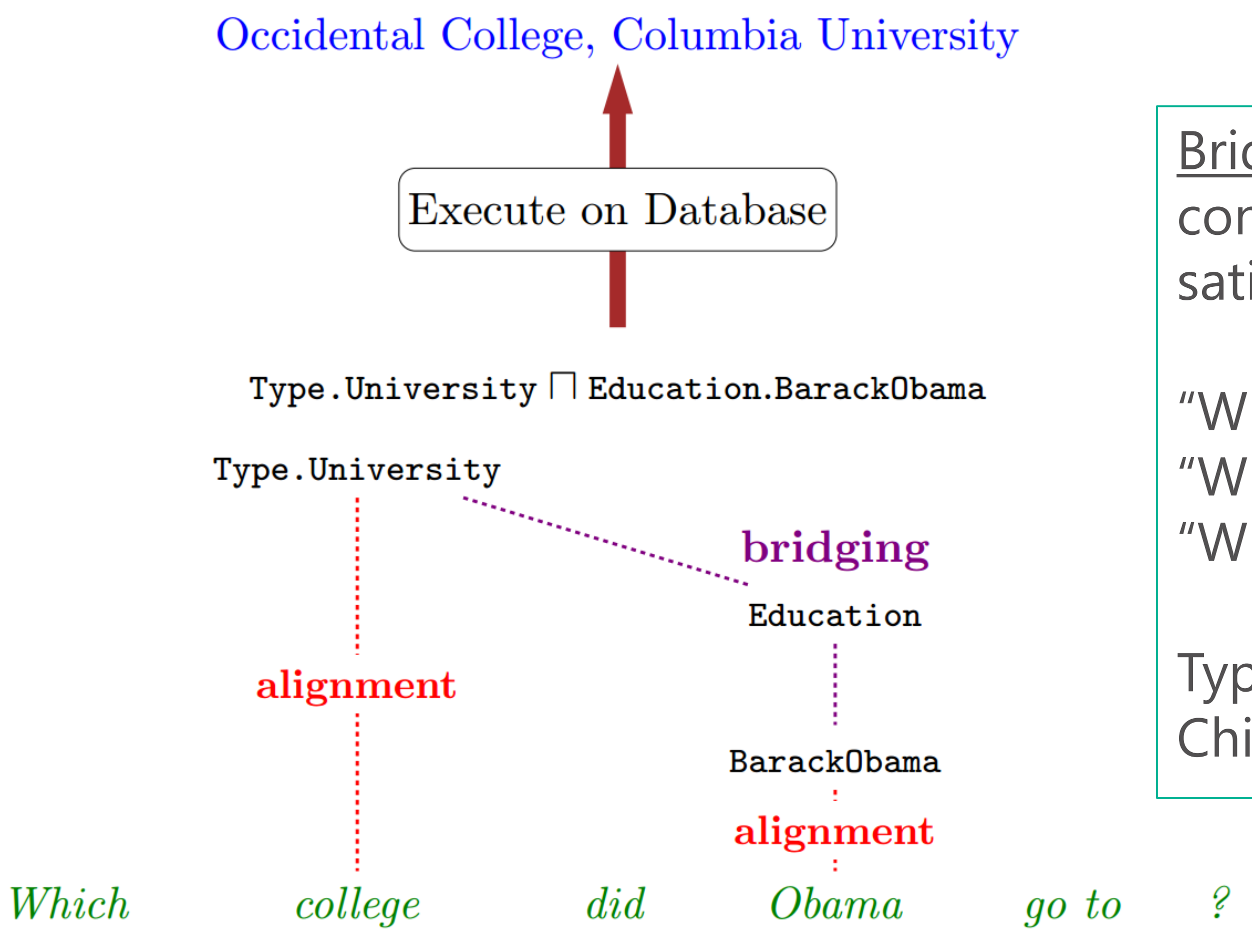


Fig.1 of [Berant et al., 2013]

Bridging: Hypothesizing predicates to be connected when the type constraints are satisfied

"What government does Chile have?"
"What actors are in Top Gun?"
"What is Italy money?"

Type.FormOfGovernment
Chile

SEMPRE – Paraphrasing [Berant & Liang, ACL-14]

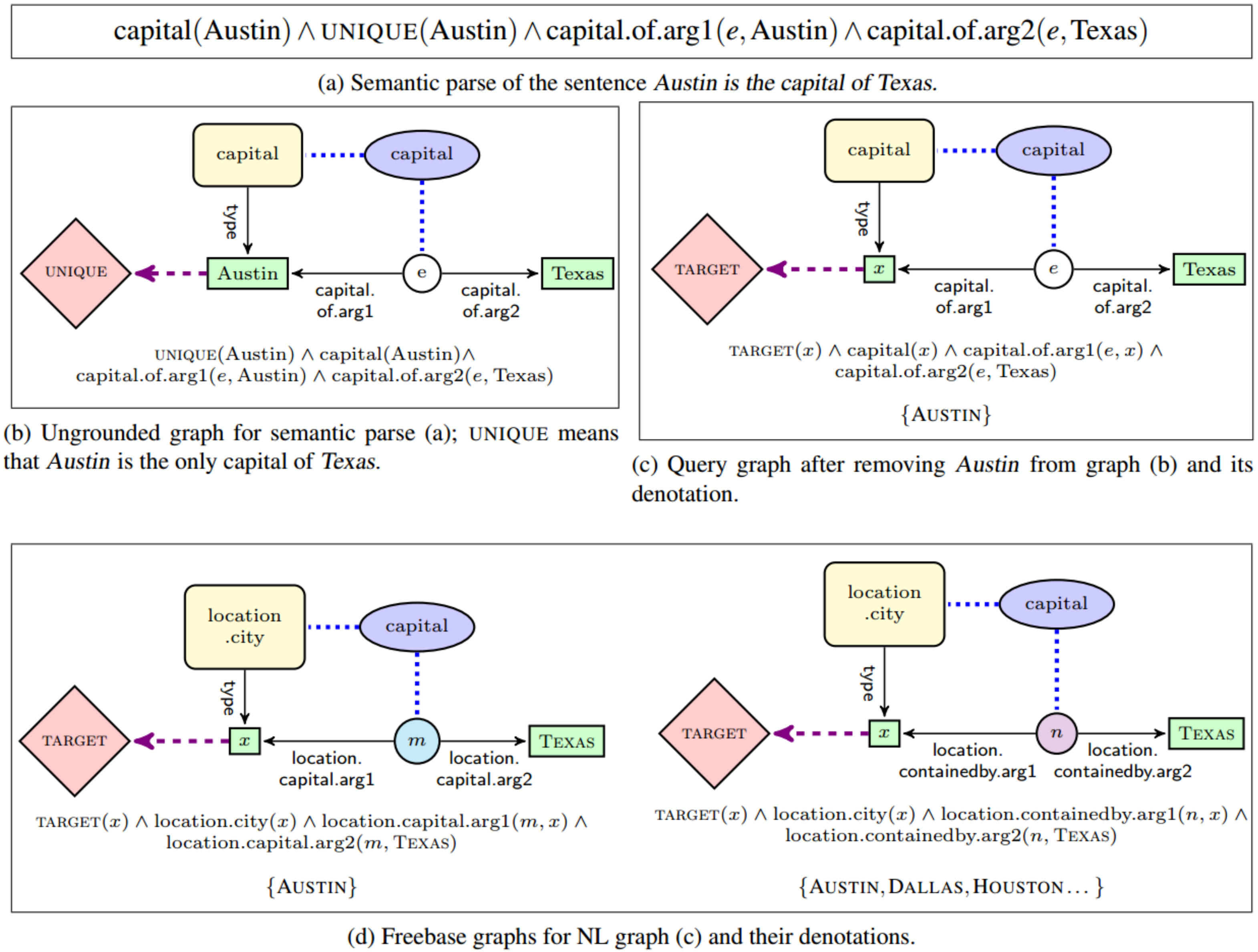
What political party founded by Henry Clay? ... *What event involved the people Henry Clay?*

$\uparrow$ $\uparrow$

`Type.PoliticalParty` $\sqcap$ `Founder.HenryClay` ... `Type.Event` $\sqcap$ `Involved.HenryClay`

Fig.1 of [Berant & Liang, 2014]

“CCG-Graph” [Reddy et al., TACL-2014]



(a) Semantic parse of the sentence *Austin is the capital of Texas.*

(b) Ungrounded graph for semantic parse (a); UNIQUE means that *Austin* is the only capital of *Texas.*

(c) Query graph after removing *Austin* from graph (b) and its denotation.

(d) Freebase graphs for NL graph (c) and their denotations.

Fig.2 of [Reddy et al., 2014]

Austin is the capital of Texas.
What is the capital of Texas?

- Word Nodes (Ovals)
  - word nodes are connected via syntactic dependencies
- Entity Nodes (Rectangles)
- Mediator Nodes (Circles)
  - Represent events
  - Binary predicates
- Type nodes (Rounded rectangles)
  - Unary predicates
- Math nodes (Diamonds)
  - e.g., Aggregation Functions

"Machine Translation" [Bao et al., ACL-14]

CYK Paring

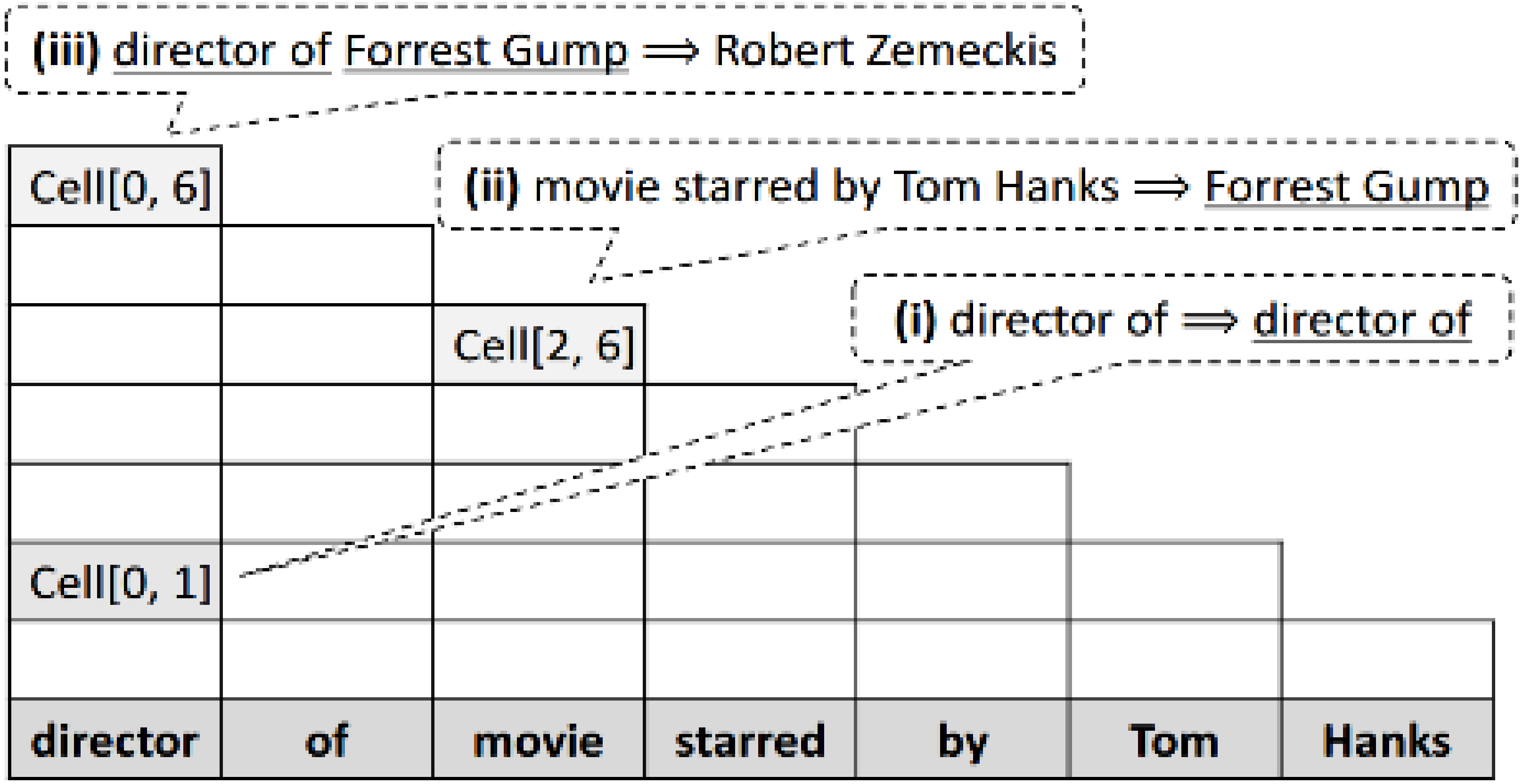


Fig.1 of [Bao et al., 2014]

Each span is a mapping of a single-relation question:

Question Pattern:
"Who is the director of Forrest Gump?"
(Forrest Gump, Director, ?)

- Patterns from mining Bing query logs.

* Few questions in WebQuestions are with a long chain like this.

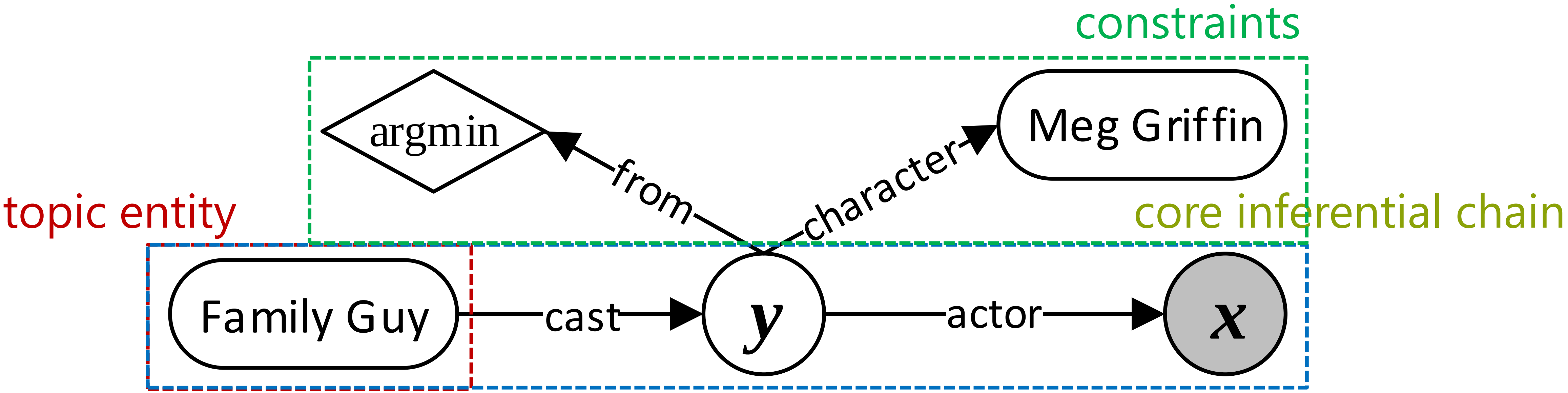
Staged Query Graph Generation [Yih et al. ACL-15]
Core idea

- Proposing a new semantic parse language – query graph
  - Resembles subgraphs of the knowledge base
  - Can be *directly* mapped to an executable query (e.g., SQL, SPARQL)
- Reducing semantic parsing to a search problem
  - Grows the candidate query graph through *staged* state-actions

Query Graph

Who first voiced Meg on Family Guy?

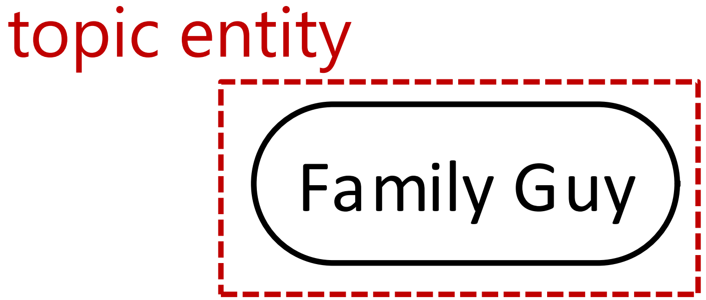
$$\lambda x.\, \exists y.\, \mathrm{cast}(\mathrm{FamilyGuy}, y) \wedge \mathrm{actor}(y, x) \wedge \mathrm{character}(y, \mathrm{MegGriffin})$$



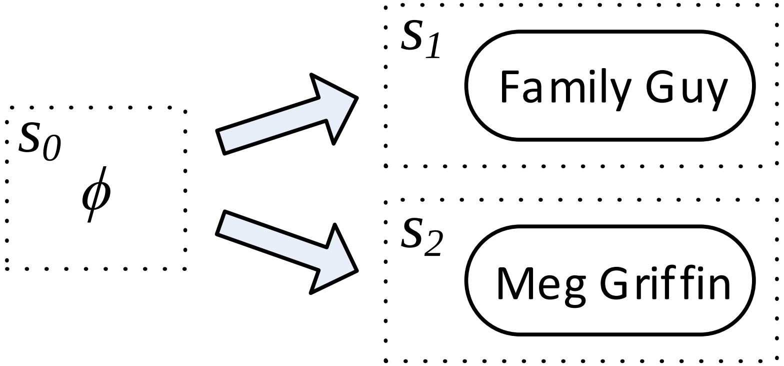
Inspired by [Reddy+ 14], but closer to λ-DCS [Liang 13]

Query Graph – Topic Entity

Who first voiced **Meg** on **Family Guy**?



Link Topic Entity



- An advanced entity linking system for short text

  Yang & Chang, "*S-MART: Novel Tree-based Structured Learning Algorithms Applied to Tweet Entity Linking.*" In ACL-15.

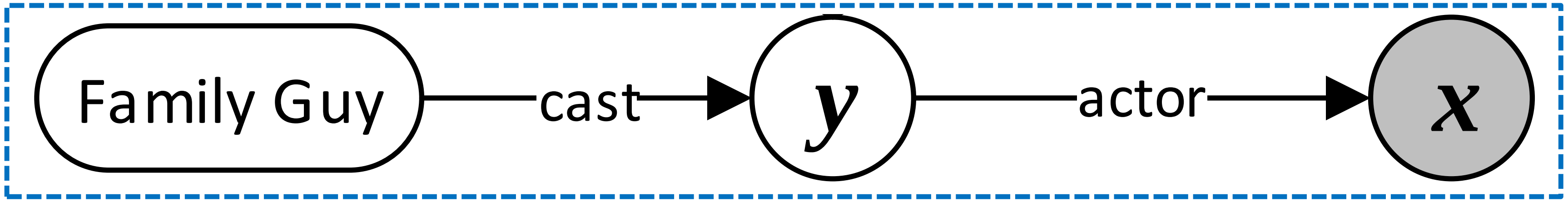
- Prepare surface-form lexicon $\mathcal{L}$ for entities in the KB
- Entity mention candidates: all consecutive word sequences in $\mathcal{L}$, scored by the statistical model
- Up to 10 top-ranked entities are considered as topic entity

Query Graph – Core Inferential Chain

Who first voiced Meg on Family Guy?

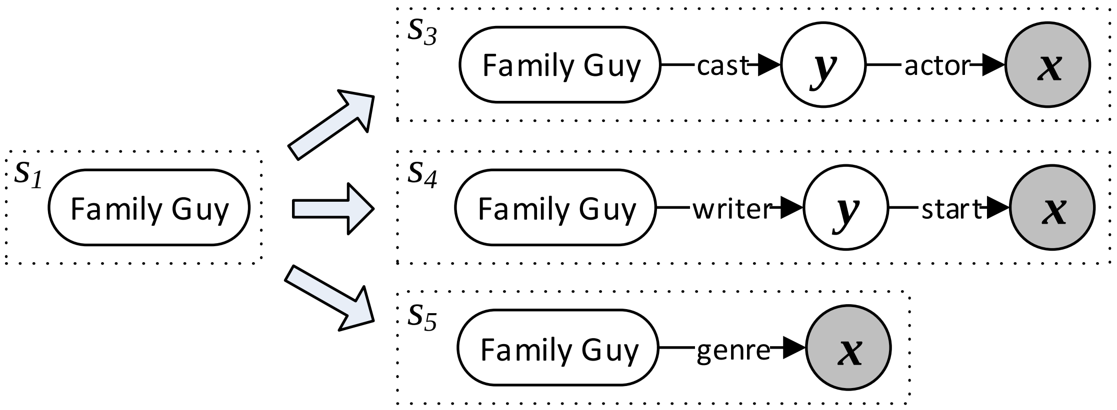
{cast-actor, producer, awards_won-winner}

core inferential chain



Identify Core Inferential Chain

- Relationship between topic and answer (x) entities
- Explore two types of paths
  - Length 1 to non-CVT node
  - Length 2 where y can be grounded to CVT



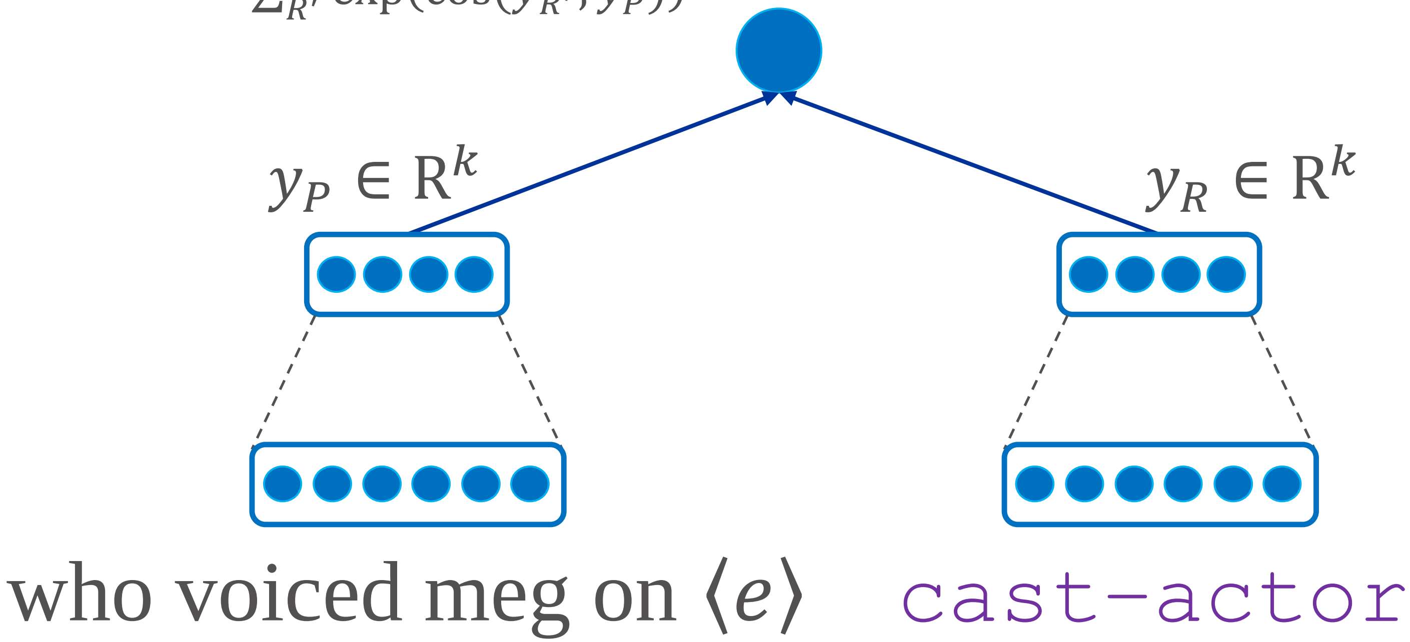
Who first voiced Meg on Family Guy?

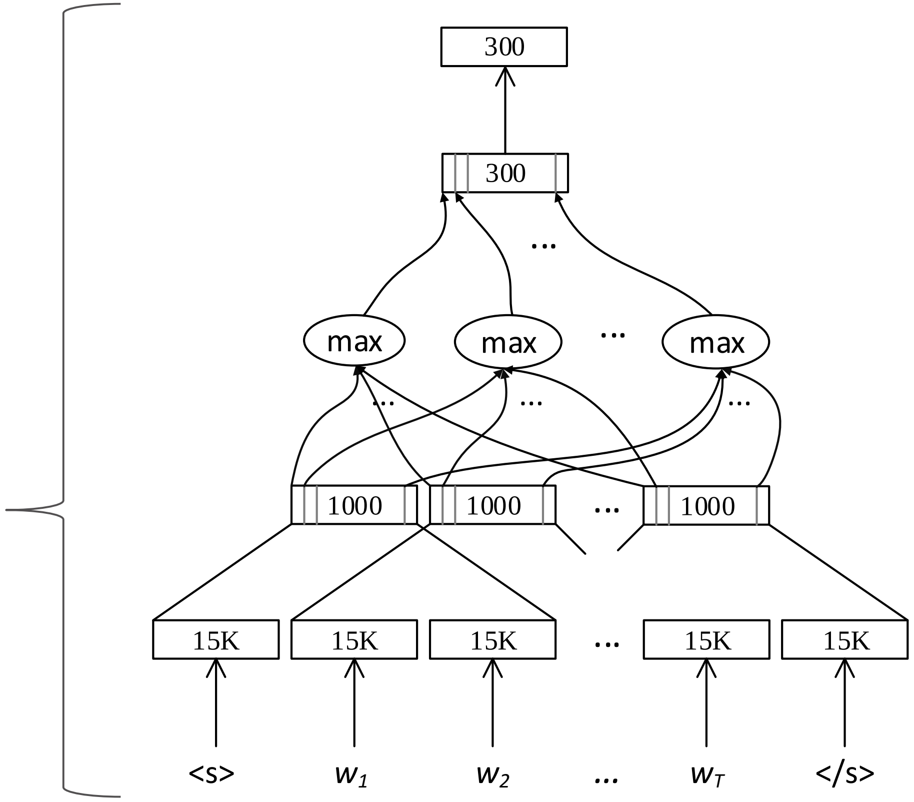
`{cast-actor, writer-start, genre}`

Relation Matching using Deep Convolutional Neural Networks (DSSM [Shen+ 14])

- Input is mapped to two $\boldsymbol{k}$-dimensional vectors
- Probability is determined by softmax of their cosine similarity

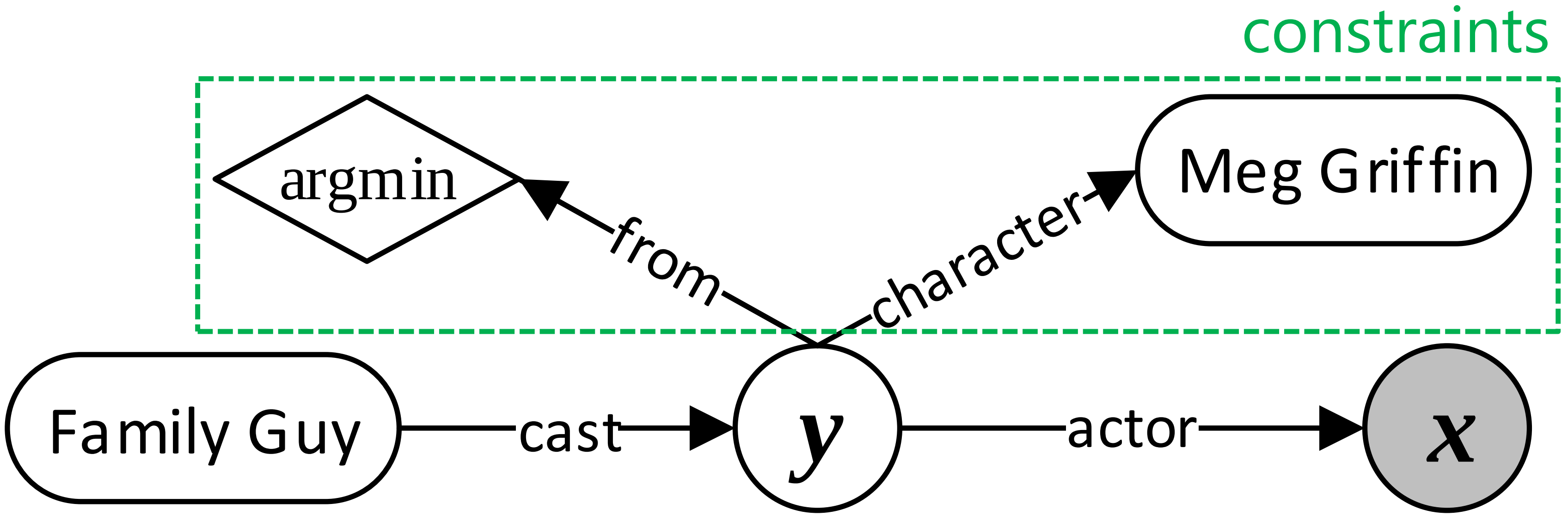
$$P(R|P) = \frac{\exp(\cos(y_R, y_P))}{\sum_{R'} \exp(\cos(y_{R'}, y_P))}$$



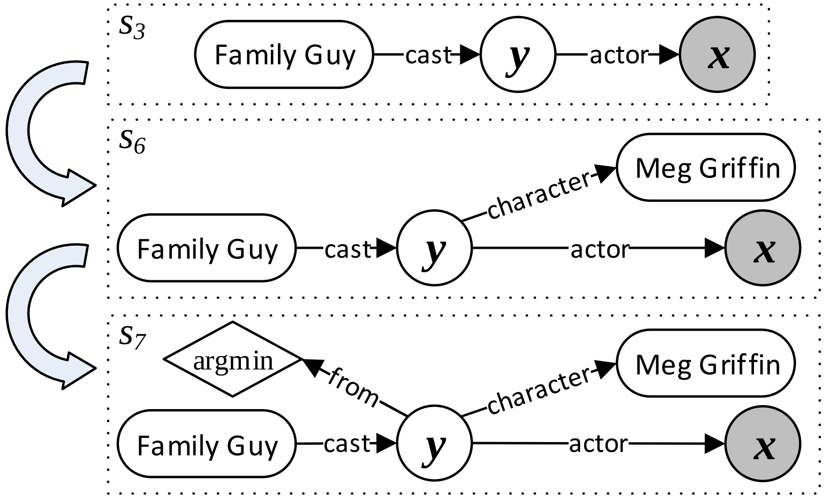


Query Graph - Constraints

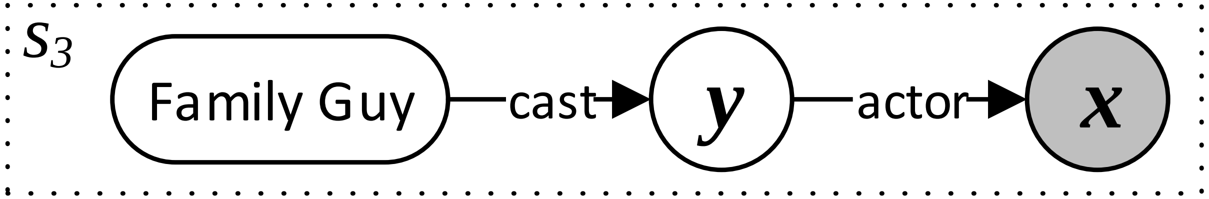
Who first voiced Meg on Family Guy?



Augment Constraints



- Who first voiced Meg on Family Guy?



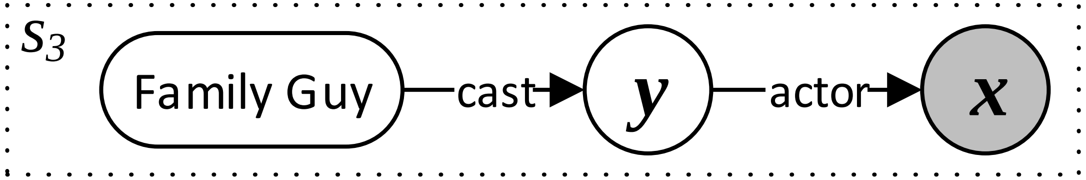
$$\lambda x. \exists y. \, \text{cast}(\text{FamilyGuy}, y) \wedge \text{actor}(y, x)$$

- One or more constraint nodes can be added to y or x
  - y : Additional property of this event (e.g., $\text{character}(y, \text{MegGriffin})$)
  - x : Additional property of the answer entity (e.g., gender)
- Only subset of constraint nodes are considered
  - e.g., entities detected in the question

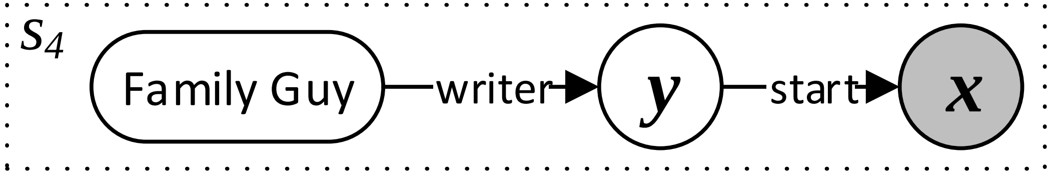
Learning Reward Function γ

- Judge whether a query graph is a correct semantic parse
- Log-linear model with pairwise ranking objective [Burges 10]

Who first voiced Meg on Family Guy?

$\gamma($ 

$)$

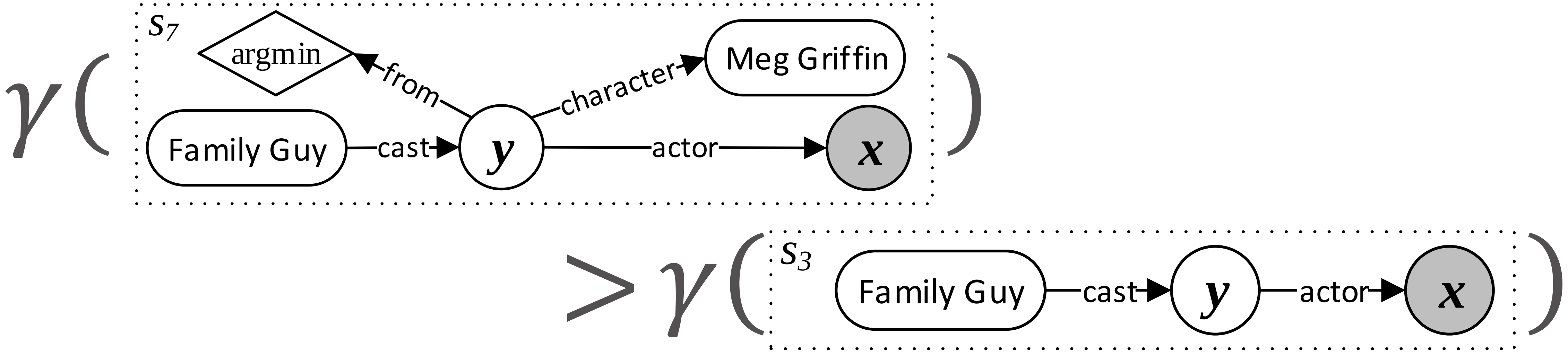
$> \gamma($ 

$)$

Learning Reward Function γ

- Judge whether a query graph is a correct semantic parse
- Log-linear model with pairwise ranking objective [Burges 10]

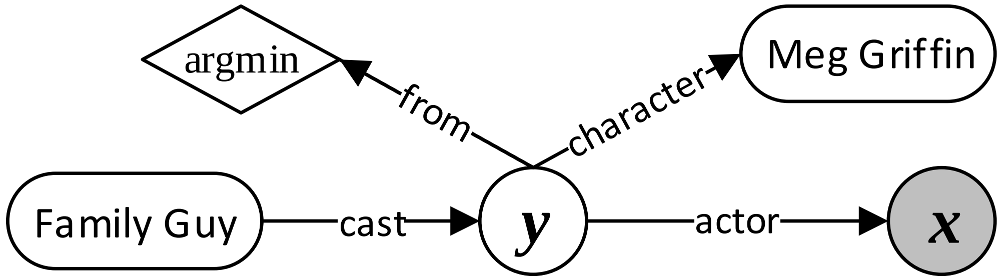
Who first voiced Meg on Family Guy?



Learning Reward Function – Features

q = Who first voiced Meg on Family Guy?

s =



- Topic Entity
  - Entity linking scores
- Core Inferential Chain
  - Relation matching scores (NN models)
- Constraints: Keyword and entity matching
  - ConstraintEntityWord("Meg Griffin", q) = 0.5
  - ConstraintEntityInQuestion("Meg Griffin", q) = 1
- Overall
  - NumNodes(s) = 5
  - NumAnswers(s) = 1

Creating Training Data from Q/A Pairs

Relation Matching (Identifying Core Inferential Chain)

- List all the length 1 & 2 paths from any potential topic entity
- Treat any inferential chain resulting in $F_1 \geq 0.5$ to create positive pairs

Pattern	Inferential Chain
what was <e> known for	people.person.profession
what kind of government does <e> have	location.country.form_of_government
what year were the <e> established	sports.sports_team.founded
what city was <e> born in	people.person.place_of_birth
what did <e> die from	people.deceased_person.cause_of_death
who married <e>	people.person.spouse_s people.marriage.spouse

Creating Training Data from Q/A Pairs

Reward Function γ

- Apply the same best-first search procedure to training data
- Use the F_1 score of the query graph as the reward function
- For each question, create 4,000 candidate query graphs
  - All positive ($F_1 > 0$) examples
  - Randomly selected negative examples

Staged Query Graph Generation
Addresses Key Challenges

- Language mismatch
  - Advanced entity linking [Yang & Chang, ACL-15]
  - Relation matching via deep convolutional NN [Shen et al., CIKM-14]
- Large search space
  - Representation power of a parse controlled by staged search actions
  - Grounding partially the question during search
- Compositionality
  - Possible combinations limited by local subgraphs

Information Extraction [Yao & Van Durme, ACL-2014]

- "What is the name of Justin Bieber brother?"

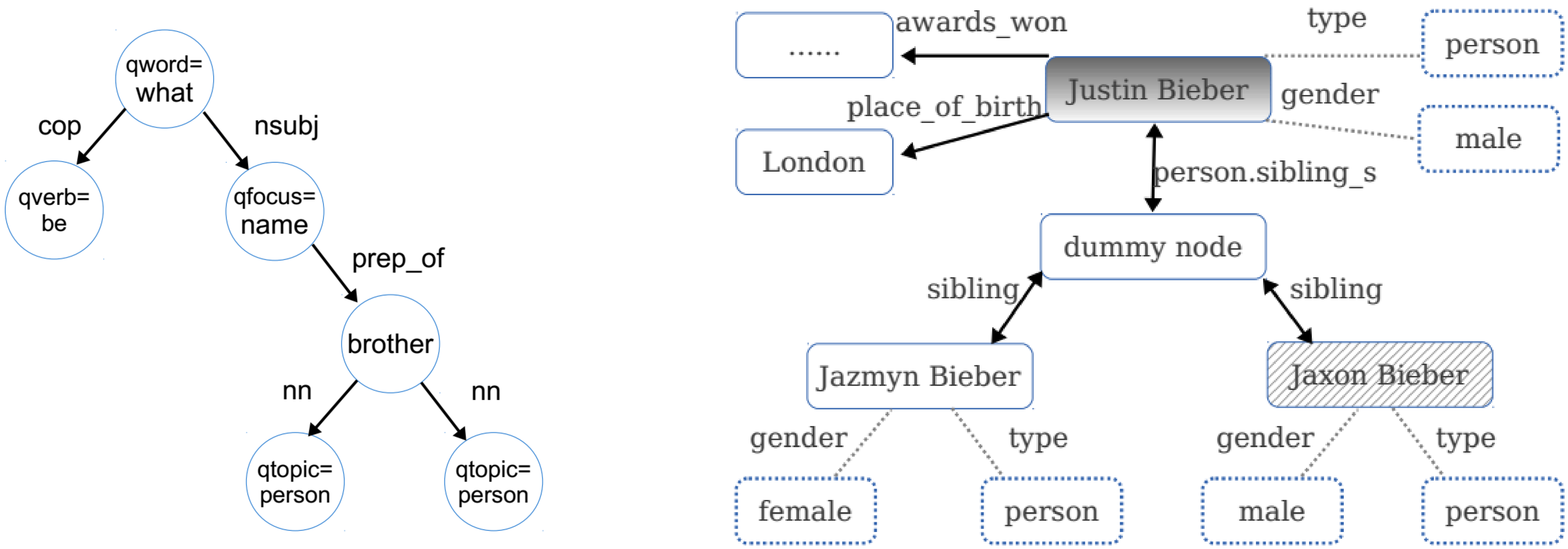


Fig.1 of [Yao & Van Durme, 2014]

- Create lots of features; learn an "answer" classifier (L1-regularized LR)

Embeddings [Bordes et al., EMNLP-2014]

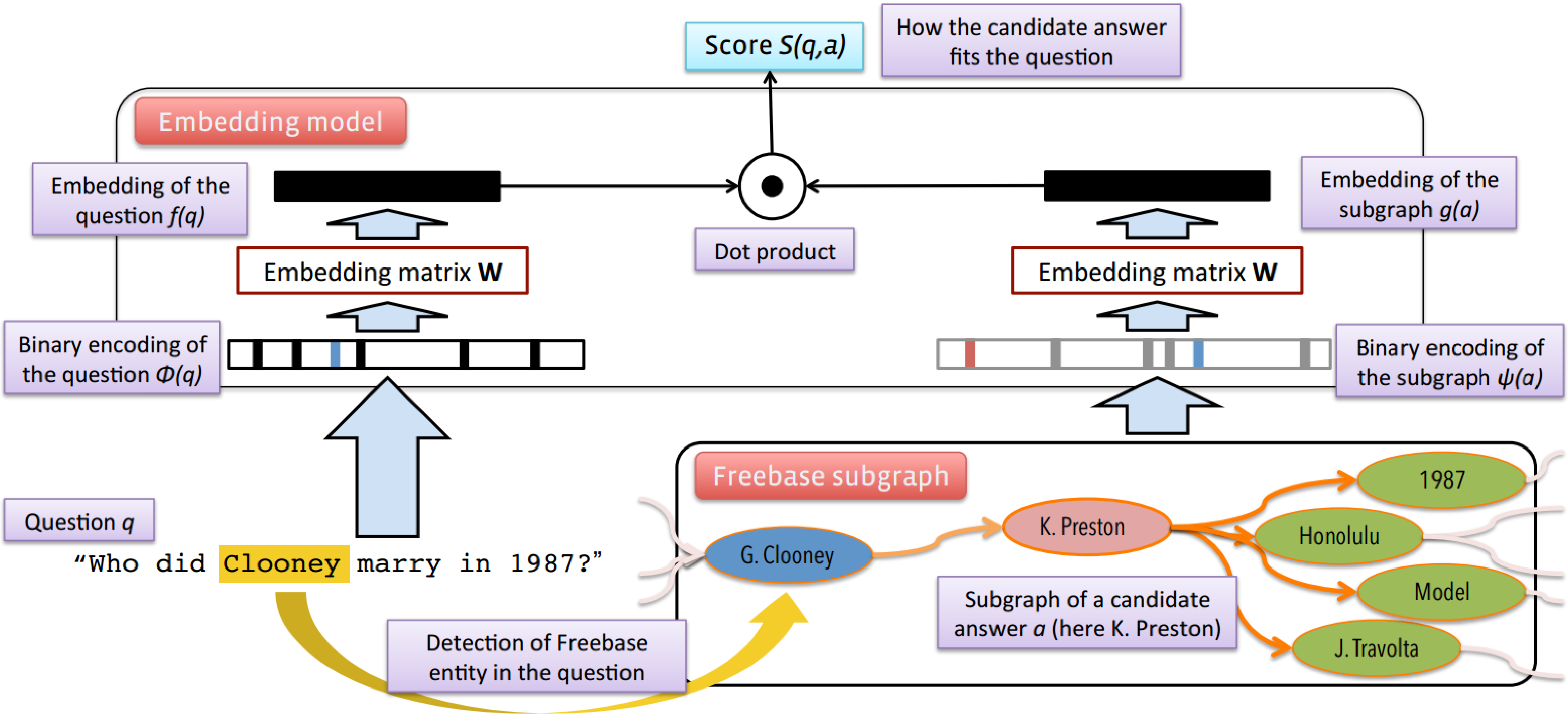
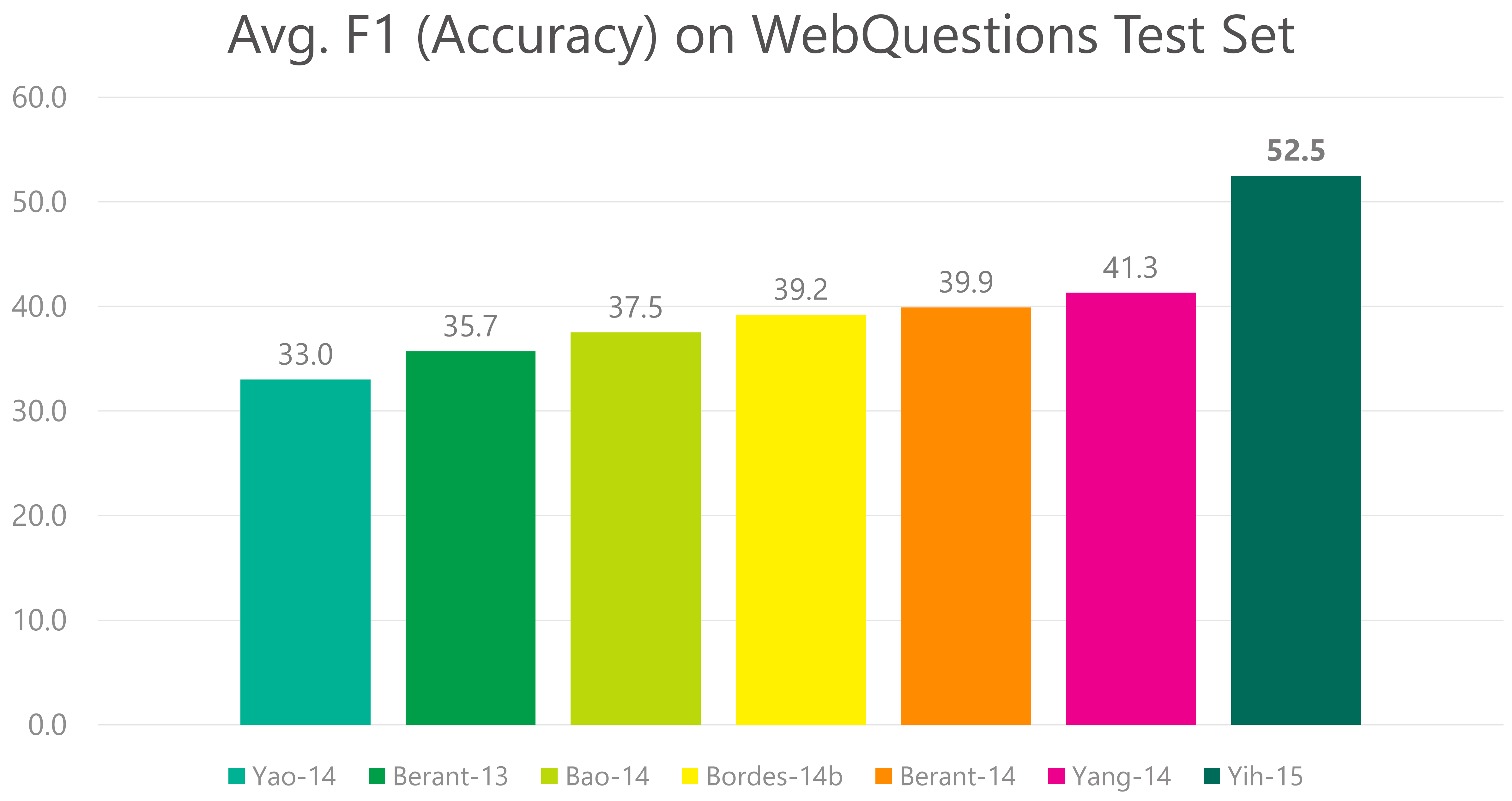
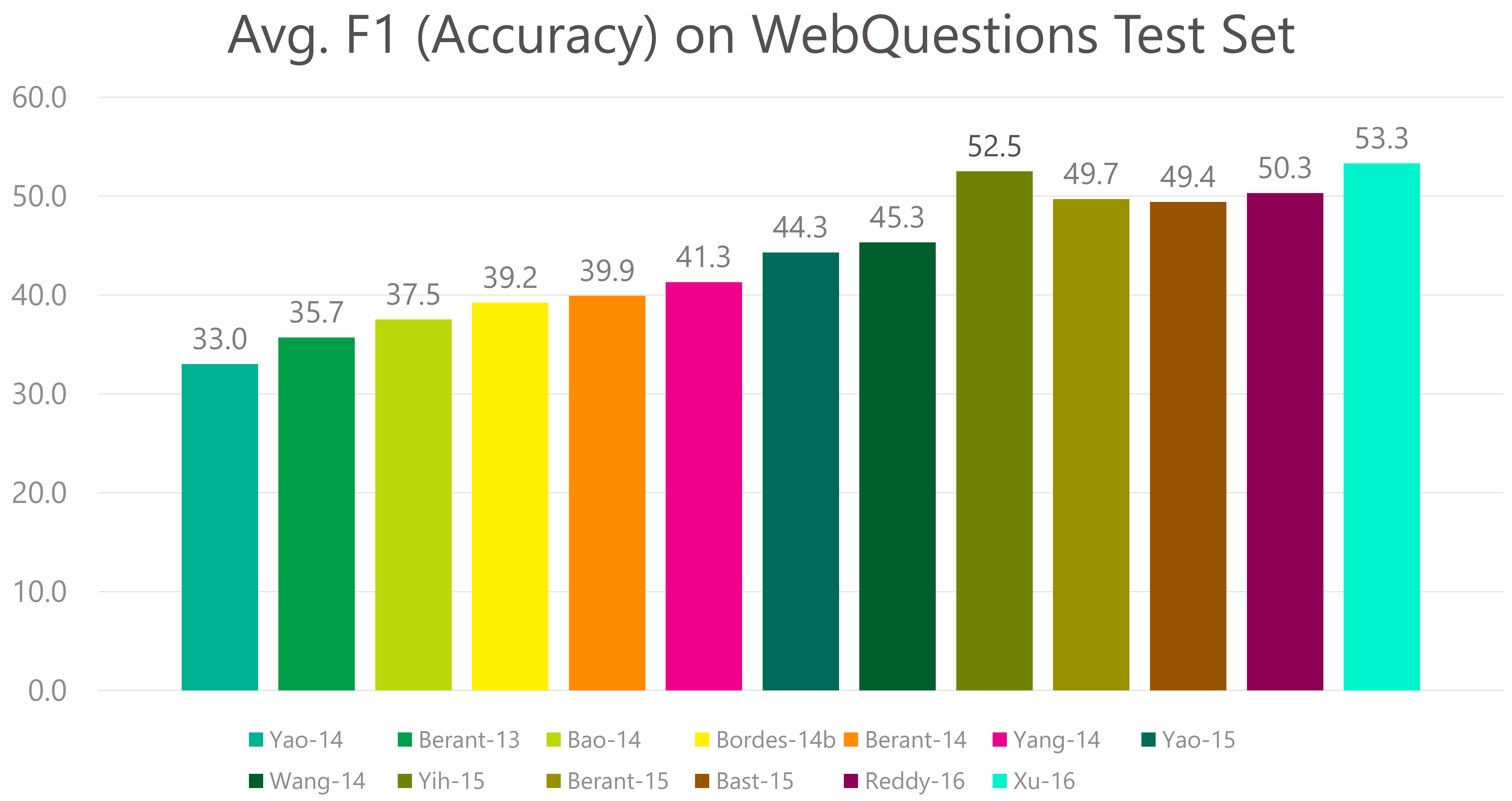


Fig.1 of [Bordes et al., 2014]


Avg. F1 (Accuracy) on WebQuestions Test Set
60.0
50.0
40.0
30.0
20.0
10.0
0.0
33.0
35.7
37.5
39.2
39.9
41.3
52.5
Yao-14
Berant-13
Bao-14
Bordes-14b
Berant-14
Yang-14
Yih-15


Avg. F1 (Accuracy) on WebQuestions Test Set
60.0
50.0
40.0
30.0
20.0
10.0
0.0
33.0
35.7
37.5
39.2
39.9
41.3
44.3
45.3
52.5
49.7
49.4
50.3
53.3
Yao-14
Berant-13
Bao-14
Bordes-14b
Berant-14
Yang-14
Yao-15
Wang-14
Yih-15
Berant-15
Bast-15
Reddy-16
Xu-16

Other Datasets

- Free917 [Cai & Yates, ACL-13]
  - 917 English questions labeled with lambda expressions with predicates & constants defined in Freebase
- Simple Questions [Bordes et al., arXiv:1506.02075]
  - 108,442 questions paired with Freebase triples
  - Multi-argument relations (CVT) don't seem to be included
- WebQuestionsSP (http://aka.ms/WebQSP) [Yih et al., ACL-16]
  - Full semantic parses of WebQuestions in SPARQL, along with updated answers and additional entity/relation information

Summary

- Recent work on question answering with KB
  - Task: Answering WebQuestions using Freebase
  - Most approaches aim for semantic parsing of questions
- Challenges
  - How to leverage multiple resources to handle language mismatch?
  - How to handle compositionality correctly and efficiently?
- Very active research problem
  - Many new methods being proposed (e.g., [Berant & Liang, TACL-15], [Reddy et al., TACL-16], [Xu et al., ACL-16])

Discussion

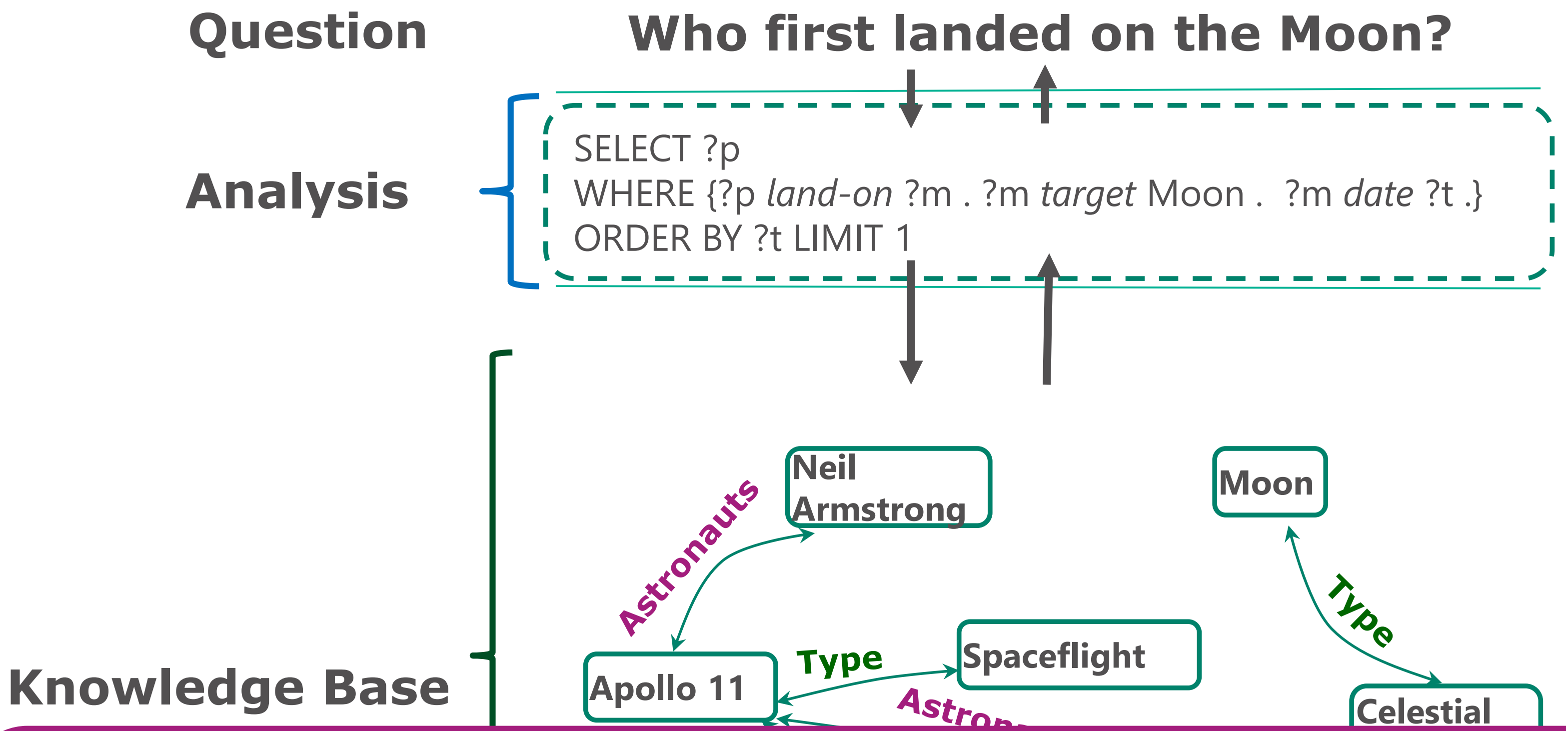
- Why is WebQuestions so successful?
  - "Largest" dataset for evaluating semantic parsing
  - A new direction for open-domain question answering
- Is semantic parsing the right approach for QA?
  - Not many alternatives when the information is stored in the DB
  - The derivation of answers is more interpretable; easier to debug
  - Not necessarily the best approach for factoid question answering

References (Incomplete)

- Unger et al. “Template-based question answering over RDF data.” WWW-2012.
- Yahya et al. “Natural language questions for the web of data.” EMNLP-2012.
- Cai & Yates. “Large-scale semantic parsing via schema matching and lexicon extension.” ACL-2013.
- Kwiatkowski et al. “Scaling semantic parsers with on-the-fly ontology matching.” EMNLP-2013.
- Berant et al. “Semantic parsing on Freebase from question-answer pairs.” EMNLP-2013.
- Zou et al. “Natural language question answering over RDF: a graph data driven approach.” SIGMOD-2014.
- Yih et al. “Semantic parsing for single-relation question answering.” ACL-2014.
- Bao et al. “Knowledge-Based Question Answering as Machine Translation.” ACL-2014.
- Berant & Liang. “Semantic Parsing via Paraphrasing.” ACL-2014.
- Yao & Van Durme. “Information extraction over structured data: Question answering with freebase.” ACL-2014.
- Fader et al. “Open question answering over curated and extracted knowledge bases.” KDD-2014.
- Bordes et al. “Open question answering with weakly supervised embedding models.” ECML-PKDD-2014.
- Bordes et al. “Question answering with subgraph embeddings.” EMNLP-2014.
- Yang et al. “Joint relational embeddings for knowledge-based question answering.” EMNLP-2014.
- Reddy et al. “Large-scale Semantic Parsing without Question-Answer Pairs.” TACL, 2014.
- Yih et al. “Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base.” In Submission.

Question and Answering with the Web

Issues with KB QA



Issues:

- Semantic parsing is difficult due to ontology mismatch
- Knowledge base is incomplete (missing entities/relations/attribute values)

Knowledge Base is largely incomplete

Freebase

Relation	Percentage unknown	
	All 3M	*Top 100K*
PROFESSION	68%	24%
PLACE OF BIRTH	71%	13%
NATIONALITY	75%	21%
EDUCATION	91%	63%
SPOUSES	92%	68%
PARENTS	94%	77%
CHILDREN	94%	80%
SIBLINGS	96%	83%
ETHNICITY	99%	86%

Knowledge Base Completion via Search-Based Question Answering [Robert West, et al., WWW 2014]

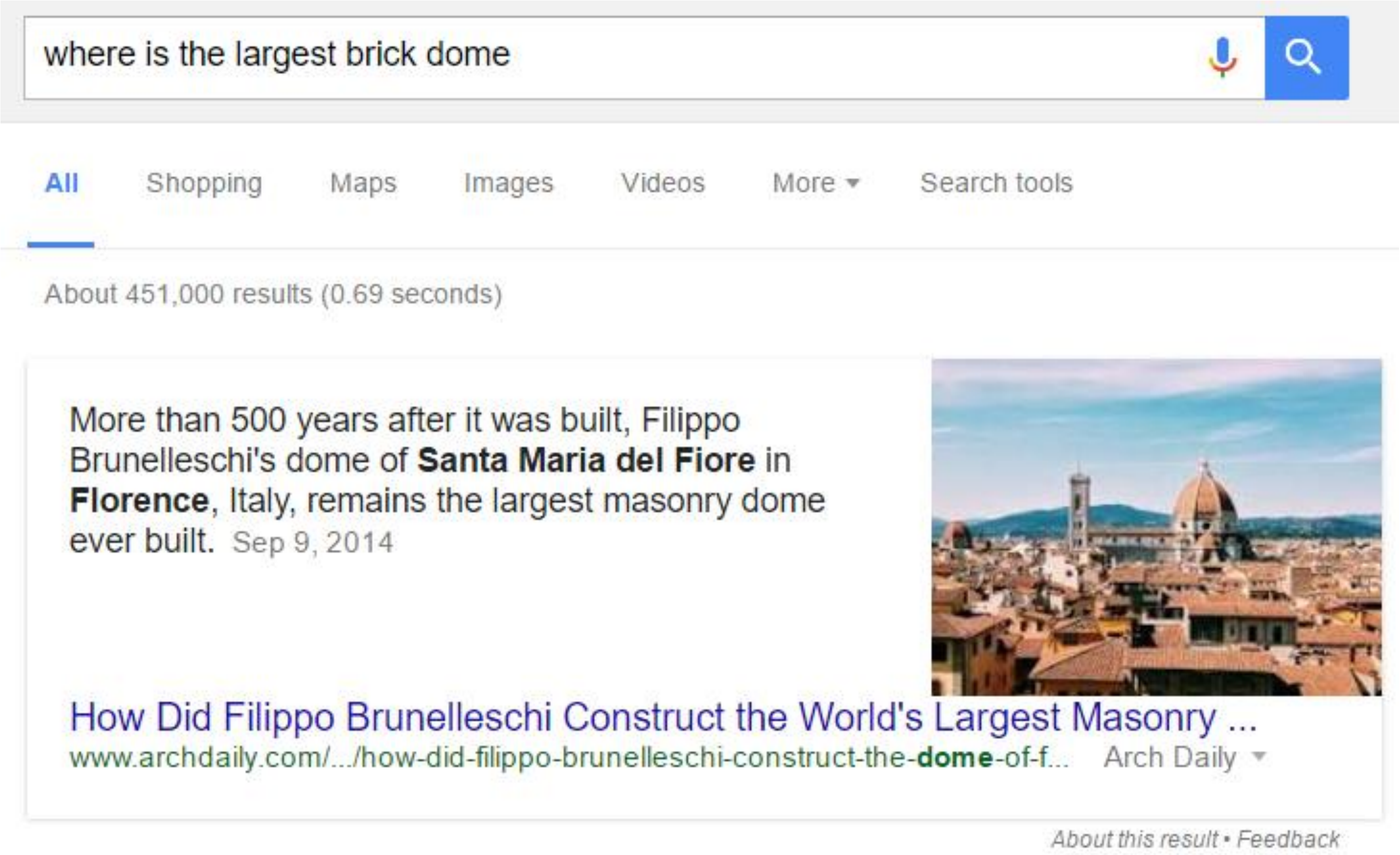
Knowledge Base is largely incomplete

Q: Where is the largest brick dome?

Answer

Florence Cathedral

The Cattedrale di Santa Maria del Fiore is the main church of Florence, Italy. Il Duomo di Firenze, as it is ordinarily called, was begun in 1296 in the Gothic style ... It remains the largest brick dome ever constructed.
en.wikipedia.org



Knowledge Bases

Issues:
- Semantic parsing is difficult due to ontology mismatch
- Knowledge base is incomplete (missing entities/relations/attribute values)

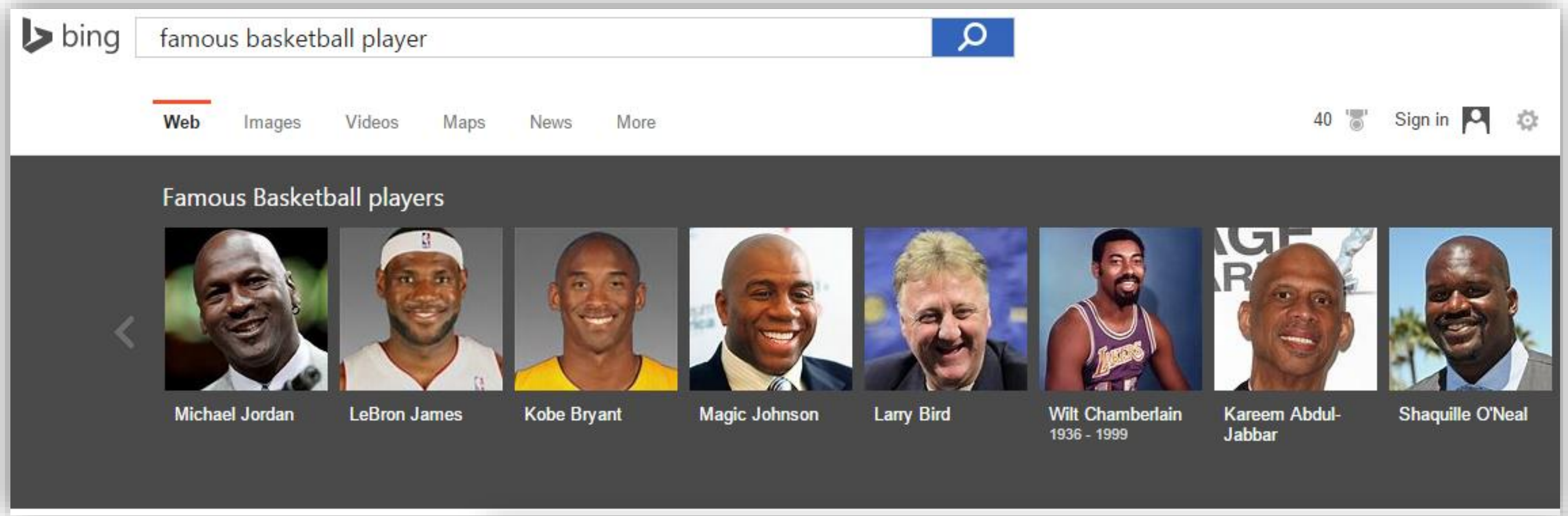
Web

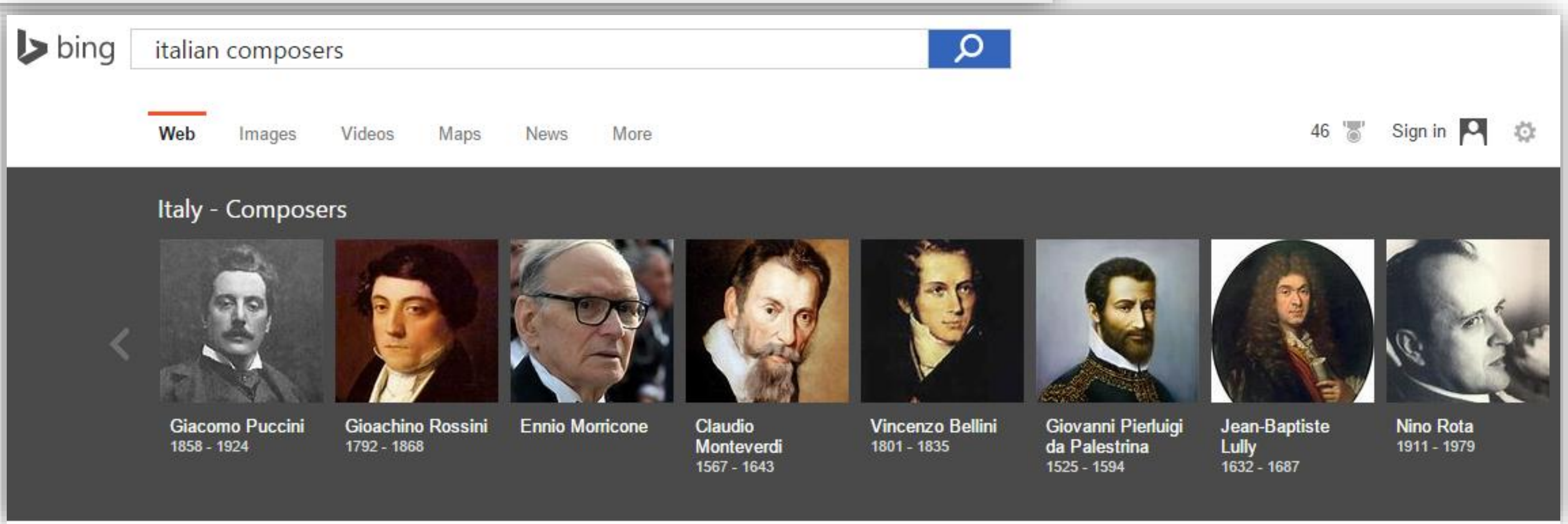
Advantages:
- Contains abundant information
- Redundancy on the Web could help confirm the answers

Web Question and Answering

- Entity Retrieval/Finding
- Factoid Answer based on Web Documents
- Factoid Answer based on Tables

Entity Retrieval/Finding





Entity Retrieval/Finding

- TREC Entity Track (2009 – 2011)
  - Related Entity Finding Task
  - Given
    - Input entity
    - Type of the target entity (PER/ORG/LOC)
    - Narrative (describing the nature of the relation in free text)
  - Return related entities

Entity Retrieval/Finding

Input Entity: Boeing 747
Target Entity Type: Organization
Narrative: Airlines that currently use Boeing 747 planes

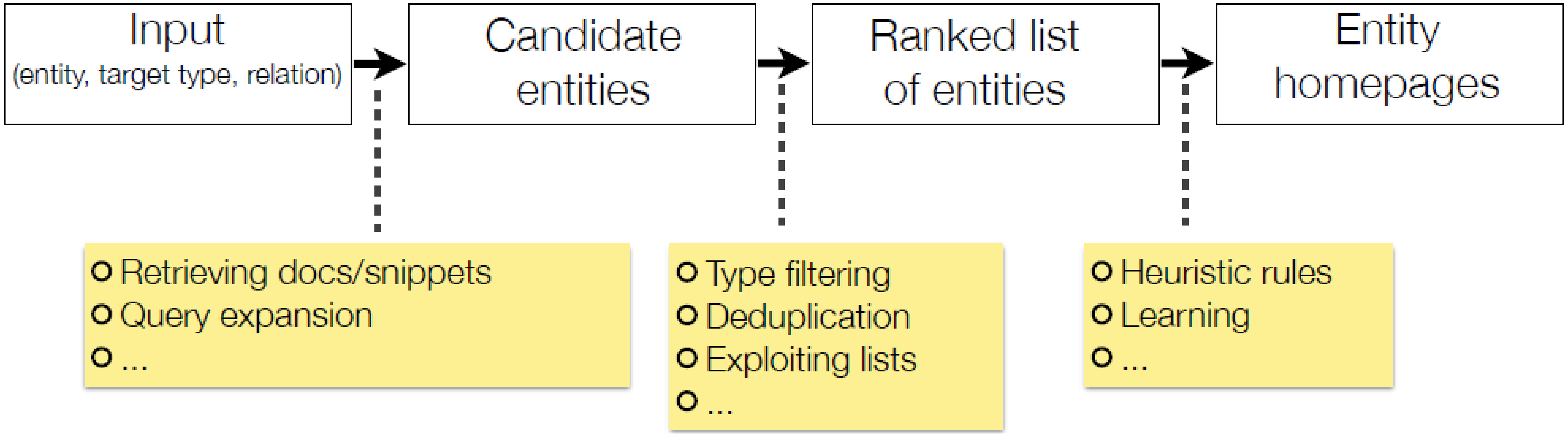
Input Entity: The food network
Target Entity Type: Person
Narrative: Chefs with a show on the food network

Input Entity: Eurail
Target Entity Type: Location
Narrative: What countries does Eurail operate in

Input Entity: Dow Jones
Target Entity Type: Organization
Narrative: Find companies that are included in the Dow Jones industrial average

Entity Retrieval/Finding

- A typical pipeline

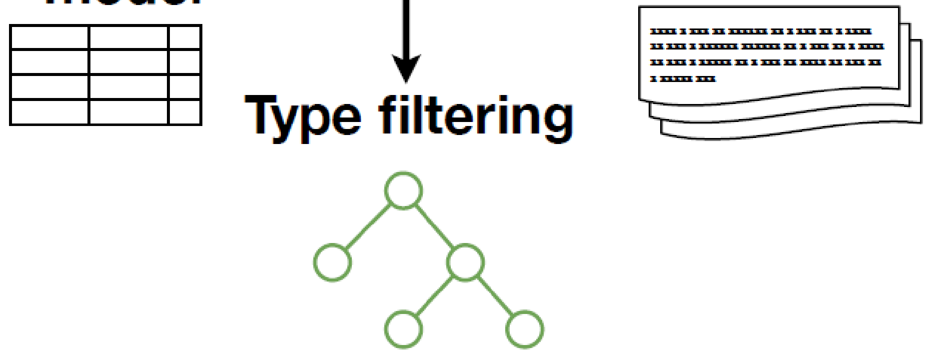


Entity Linking and Retrieval for Semantic Search [Edgar Meij, et al., WSDM 2014]

Entity Retrieval/Finding

- Three component model

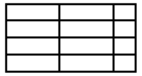
$$p(e|E,T,R) \propto \underbrace{p(e|E)}_{\text{Co-occurrence model}} \cdot \underbrace{p(T|e)}_{\text{Type filtering}} \cdot \underbrace{p(R|E,e)}_{\text{Context model}}$$

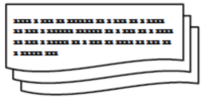


Related Entity Finding Based on Co-Occurrence [Marc Bron, et al., TREC 2009]

Entity Retrieval/Finding

$$p(e|E,T,R) \propto \underbrace{p(e|E)}_{\text{Co-occurrence model}} \cdot \underbrace{p(T|e)}_{\text{Type filtering}} \cdot \underbrace{p(R|E,e)}_{\text{Context model}}$$





$$P(R|E,e) = P(R|\theta_{Ee}) = \prod_{t \in R} P(t|\theta_{Ee})^{n(t,R)}$$

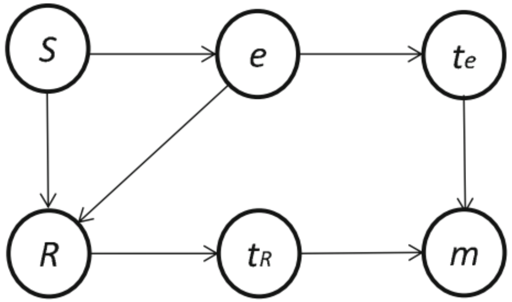
$$P(t|\theta_{Ee}) = \frac{1}{|D_{Ee}|} \sum_{d \in D_{Ee}} P(t|\theta_d)$$

$$P(t|\theta_d) = \frac{n(t,d) + \mu \cdot P(t)}{\sum_t' n(t',d) + \mu}$$

Related Entity Finding Based on Co-Occurrence [Marc Bron, et al., TREC 2009]

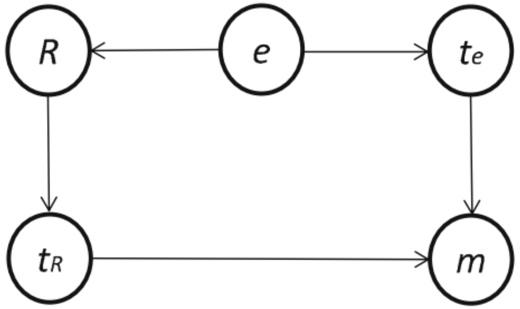
Entity Retrieval/Finding

Model A



$$p(e, m = 1|R, S) \propto p(R|e, S)\, p(e|S) \sum_{t_R} \sum_{t_e} p(m = 1|t_e, t_R)\, p(t_e|e)\, p(t_R|R)$$

Model B



$$p(e, m = 1|R) \propto p(R|e)\, p(e) \sum_{t_R} \sum_{t_e} p(m = 1|t_e, t_R)\, p(t_e|e)\, p(t_R|R)$$

Related Entity Finding by Unified Probabilistic Models [Yi Fang, et al., World Wide Web 2013]

Entity Retrieval/Finding

Input Entity: Dow Jones

Target Entity Type: Organization

Narrative: Find companies that are included in the Dow Jones industrial average

$p(m=1\|e,R)$	$p(R\|e)p(e)$	MA	$p(R\|e,S)p(e\|S)$	MB
nasdaq	**microsoft**	**boeing**	**coca cola**	**boeing**
bloomberg	**boeing**	**ibm**	**boeing**	**coca cola**
ibm	*federal reserve*	**pfizer**	*cnnmoney*	**microsoft**
news corporation	*european*	**coca cola**	*futures*	nasdaq
Yahoo	**coca cola**	**intel**	**microsoft**	**ibm**
atari	*uw*	**alcoa**	**pfizer**	**intel**
washington post	**ibm**	*cnnmoney*	**alcoa**	**merck**
boeing	**intel**	**mcdonald's**	**ibm**	**dupont**
stanford	*futures*	**merck**	*federal reserve*	**caterpillar**
enterprise media group	**merck**	**microsoft**	**mcdonald's**	*stanford*

Related Entity Finding by Unified Probabilistic Models [Yi Fang, et al., World Wide Web 2013]

Entity Retrieval/Finding

- Knowledge base are largely incomplete

Freebase

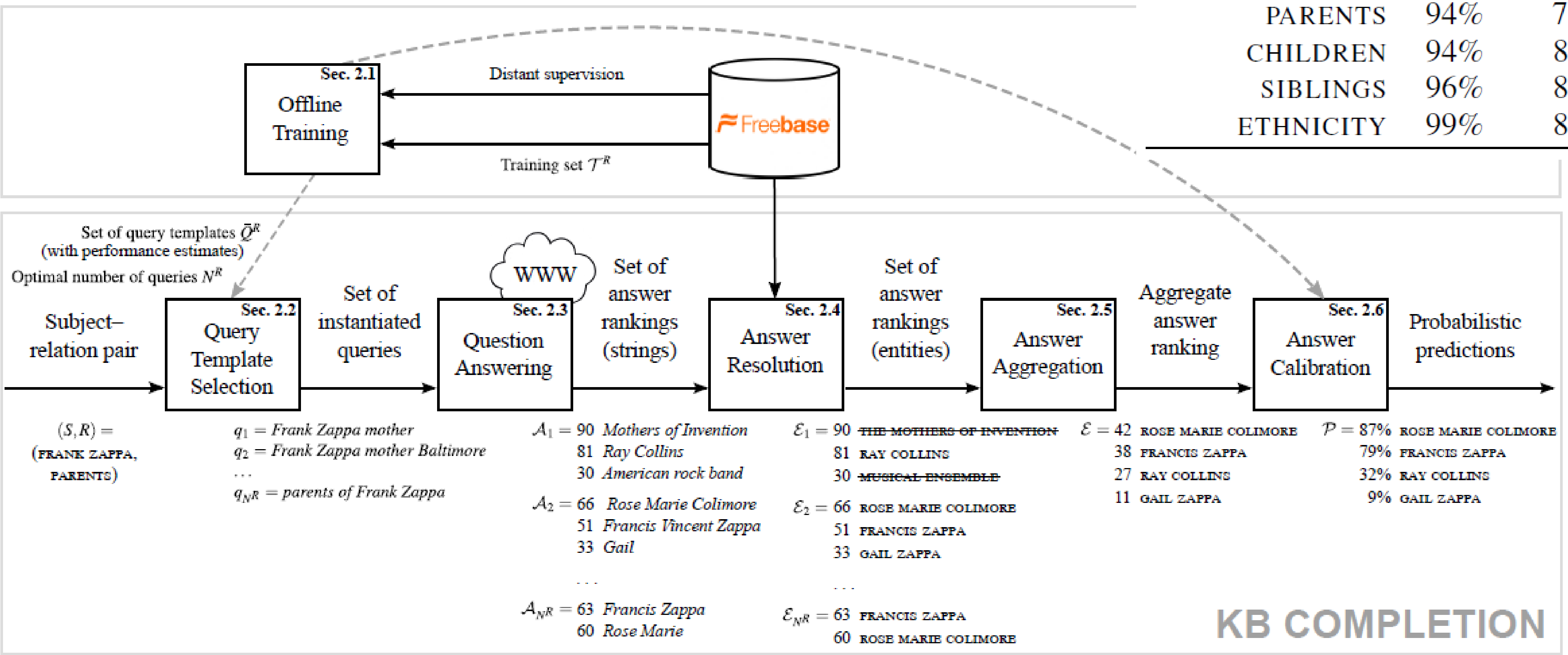
Relation	Percentage unknown	
	All 3M	*Top 100K*
PROFESSION	68%	24%
PLACE OF BIRTH	71%	13%
NATIONALITY	75%	21%
EDUCATION	91%	63%
SPOUSES	92%	68%
PARENTS	94%	77%
CHILDREN	94%	80%
SIBLINGS	96%	83%
ETHNICITY	99%	86%

Entity Retrieval/Finding techniques can be used in Knowledge Base Completion

Knowledge Base Completion via Search-Based Question Answering [Robert West, et al., WWW 2014]

Entity Retrieval/Finding

Relation	Percentage unknown	
	All 3M	*Top 100K*
PROFESSION	68%	24%
PLACE OF BIRTH	71%	13%
NATIONALITY	75%	21%
EDUCATION	91%	63%
SPOUSES	92%	68%
PARENTS	94%	77%
CHILDREN	94%	80%
SIBLINGS	96%	83%
ETHNICITY	99%	86%



Knowledge Base Completion via Search-Based Question Answering [Robert West, et al., WWW 2014]

Entity Retrieval/Finding

- Challenges
  - The TREC's related entity finding track is relatively easy since the "query intent" is known

Input Entity: Dow Jones

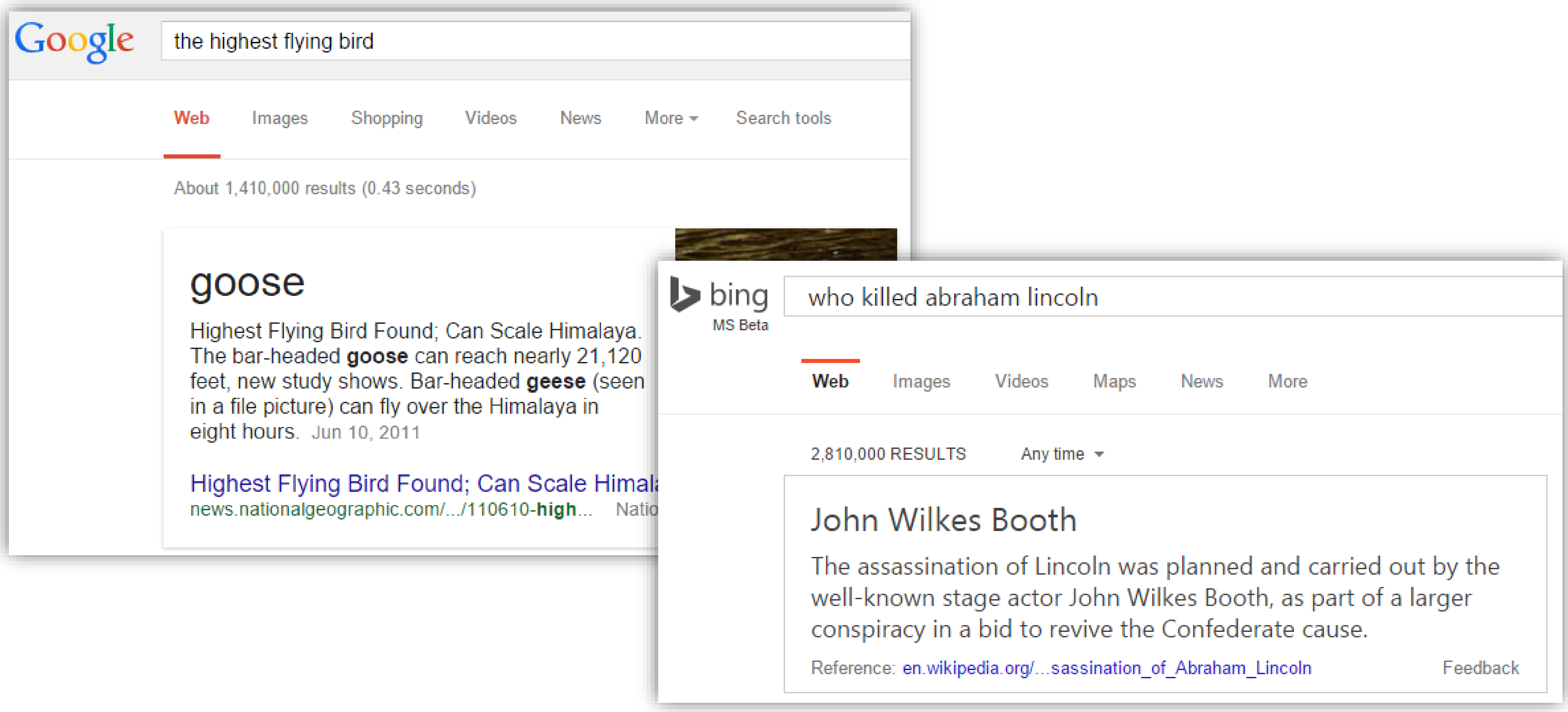
Target Entity Type: Organization

Narrative: Find companies that are included in the Dow Jones industrial average

  - In real world search engines, we need to understand the intent of queries

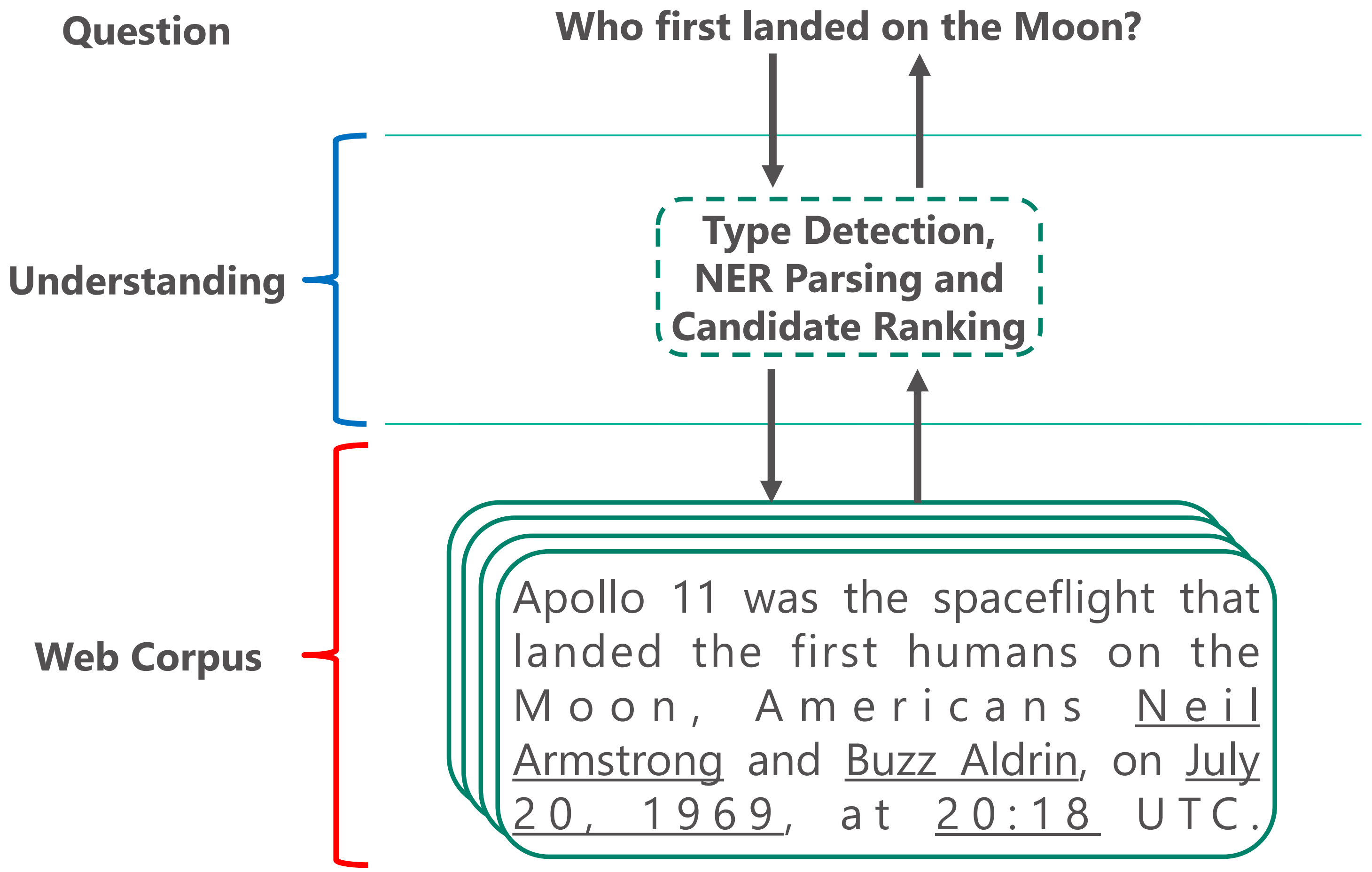
Companies in Dow Jones industrial

Factoid Answer based on Web Documents



Factoid Answer based on Web Documents

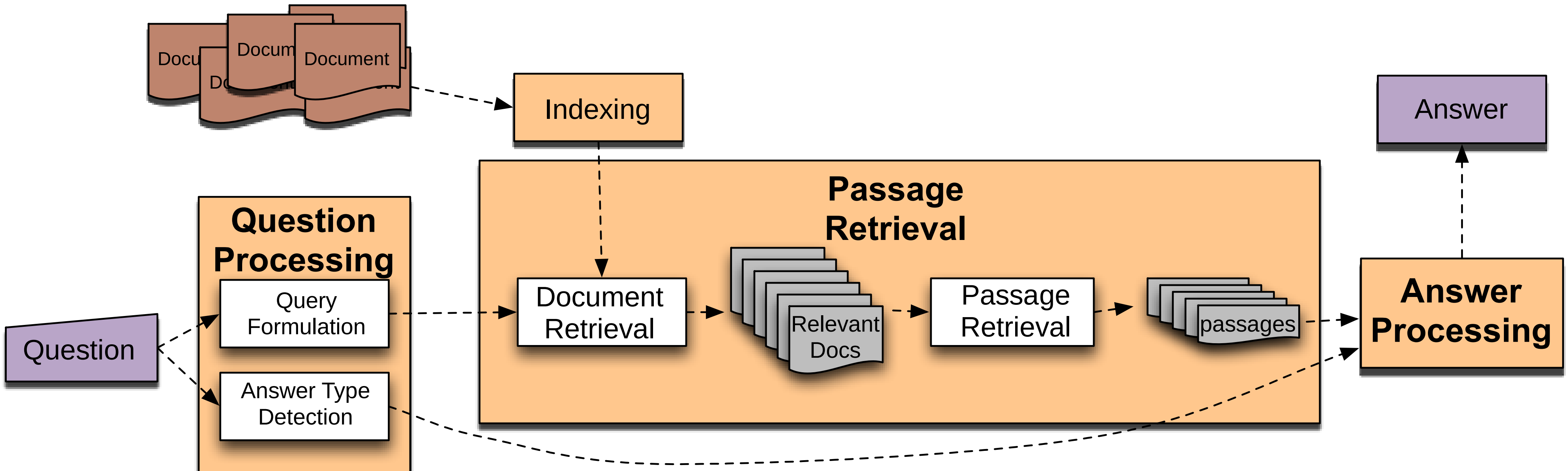
- Typical Architect of Web QnA



Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

Factoid Answer based on Web Documents

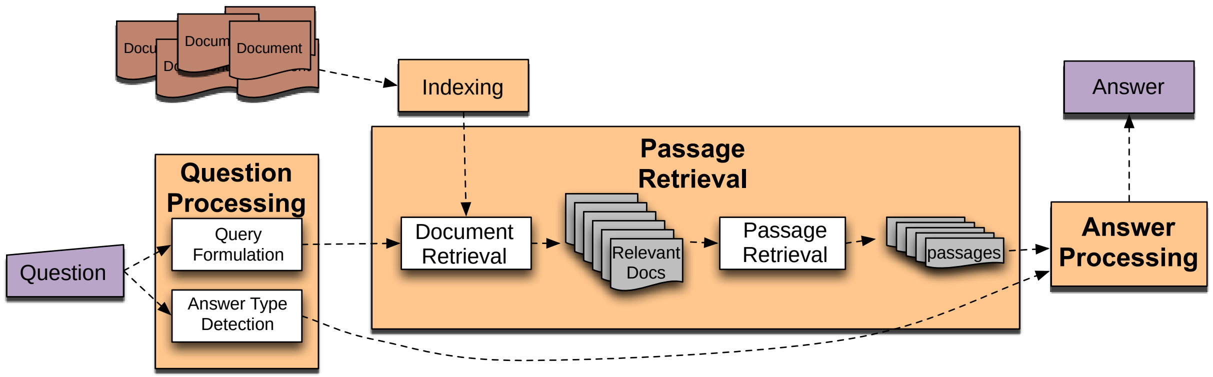
- Detailed Architect



Question Answering [Dan Jurafsky, Stanford]

Factoid Answer based on Web Documents

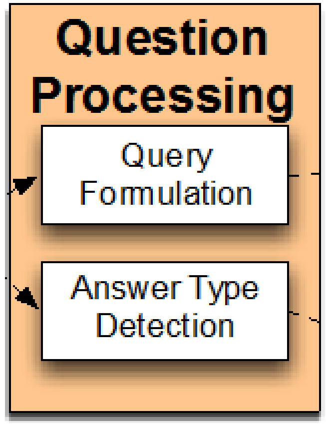
- QUESTION PROCESSING
  - Detect question type, answer type
  - Formulate queries to send to a search engine
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates



Question Answering [Dan Jurafsky, Stanford]

Factoid Answer based on Web Documents

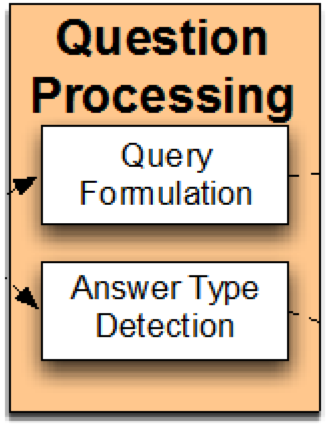
- Answer Type Detection: Name Entities
  - Who first landed on the moon?
    - Person
  - Where is the headquarters of Microsoft?
    - Location
  - What is the largest country in terms of population?
    - Country
  - Highest flying bird
    - Animal/Bird



Question Answering [Dan Jurafsky, Stanford]

Factoid Answer based on Web Documents

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title…
  - ENTITY: animal, body, color, currency…

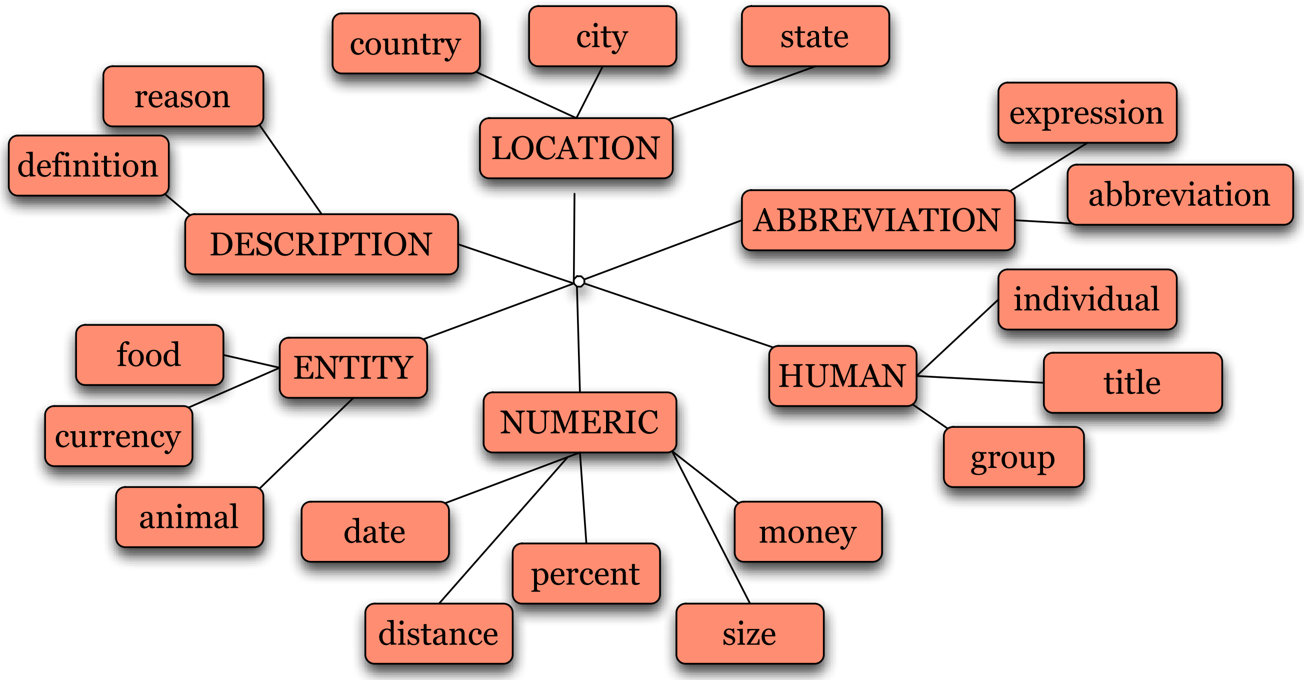


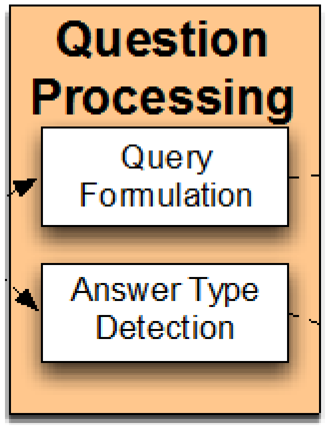
Learning Question Classifiers [Xin Li, et al., COLING 2002]

Question Answering [Dan Jurafsky, Stanford]

Factoid Answer based on Web Documents

- Part of the Answer Type Taxonomy



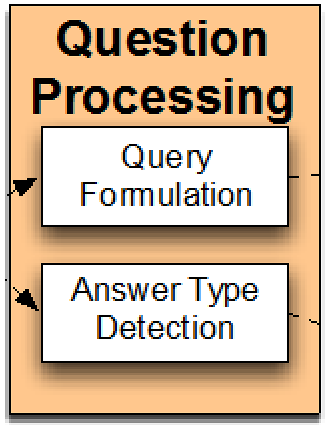


Learning Question Classifiers [Xin Li, et al., COLING 2002]

Question Answering [Dan Jurafsky, Stanford]

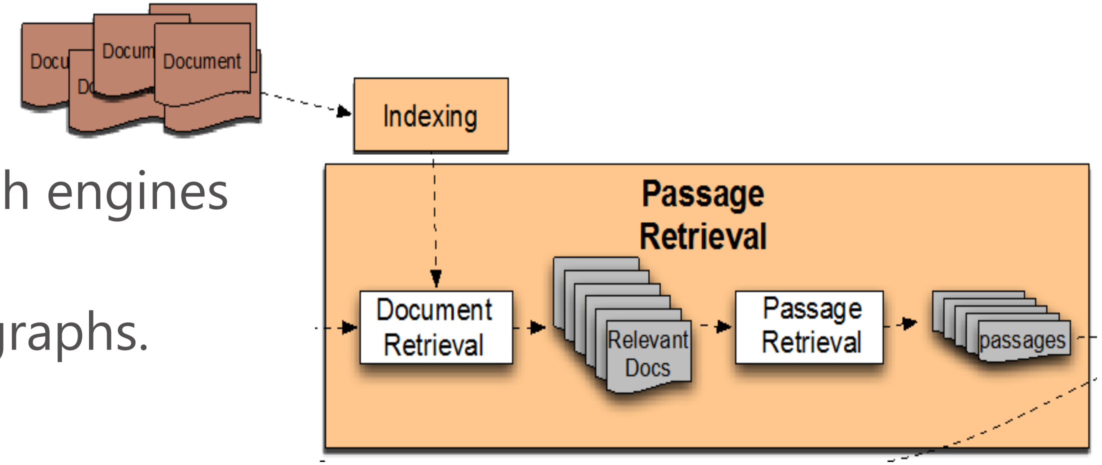
Factoid Answer based on Web Documents

- Answer Type Detection
  - Rules
    - Regular expression based rules
      - Who {is|was|are|were} PERSON
    - Question headword
      - Which **city** in China has the largest number of foreign financial companies?
      - What is the state **flower** of California?
  - Machine Learning
    - **Define** a taxonomy of question types
    - **Annotate** training data for each question type
    - **Train** classifiers for each question class using a rich set of features: Question words and phrases; Part-of-speech tags; Parse features (headwords); Named Entities; Related words



Question Answering [Dan Jurafsky, Stanford]

Factoid Answer based on Web Documents

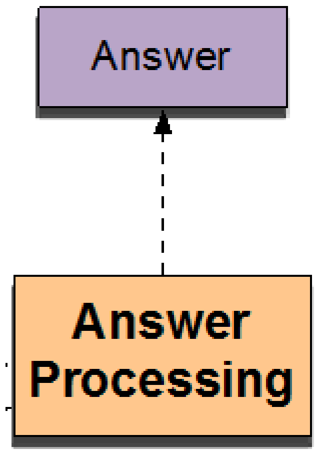


- Passage Retrieval
  - Retrieve documents using query terms through search engines
  - Segment the documents into shorter units, like paragraphs.
  - Passage ranking, features
    - Number of Named Entities of the right type in passage
    - Number of query words in passage
    - Number of question N-grams also in passage
    - Proximity of query keywords to passage
    - Longest sequence of question words
    - Rank of the document containing passage
    - …

Question Answering [Dan Jurafsky, Stanford]

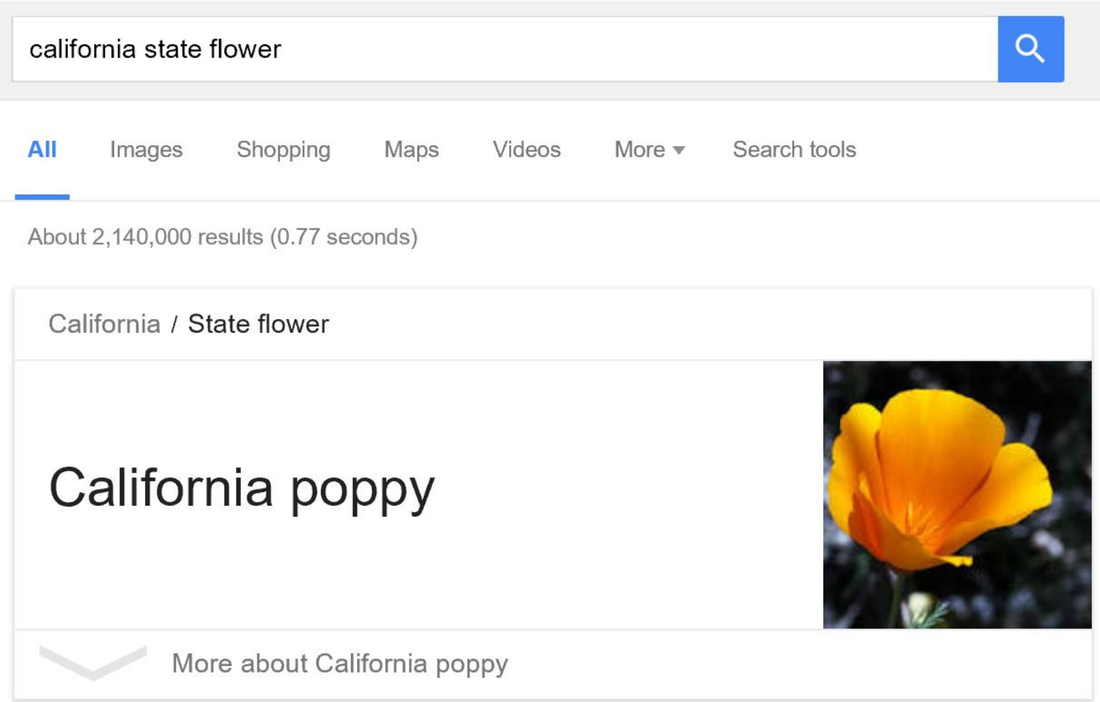
Factoid Answer based on Web Documents

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
- Return the string with the right type:
  - How many bones in an adult human body? (Number)
    - The human skeleton is the internal framework of the body. It is composed of 270 bones at birth – this total decreases to 206 bones by adulthood after some bones have fused together.

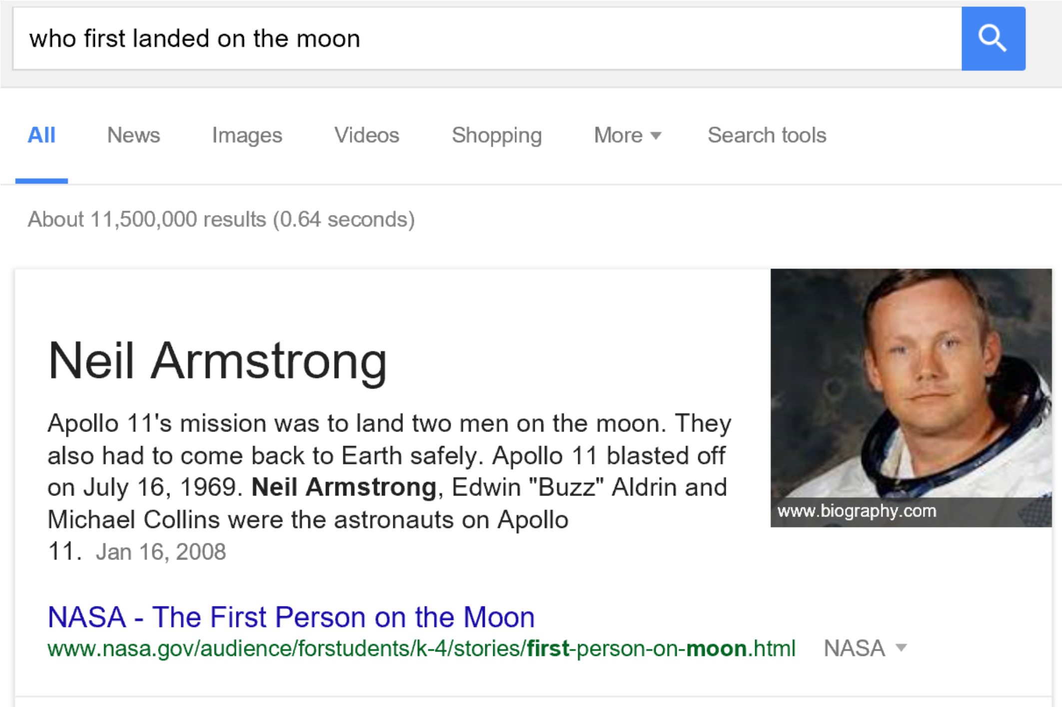


Factoid Answer based on Web Documents

Knowledge Bases based QA



Web Documents based QA



Answer Sentence Selection

Answer Sentence Selection

- Task
  - Input:
    - a question
    - a set of candidate sentences
  - Output:
    - the correct sentence that contains the exact answer
    - can sufficiently support the answer choice

Answer Sentence Selection

- Dataset
  - QASent
    - Created using TREC-QA questions

	Train	Dev	Test	Total
# of ques.	94	65	68	227
# of sent.	5,919	1,117	1,442	8,478
# of ans.	475	205	248	928
Avg. len. of ques.	11.39	8.00	8.63	9.59
Avg. len. of sent.	30.39	24.90	25.61	28.85

What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA [Mengqiu Wang, et al., EMNLP-CoNLL 2007]

Answer Sentence Selection

Algorithm	Reference	MAP	MRR
Punyakanok (2004)	Wang et al. (2007)	0.419	0.494
Cui (2005)	Wang et al. (2007)	0.427	0.526
Wang (2007)	Wang et al. (2007)	0.603	0.685
H&S (2010)	Heilman and Smith (2010)	0.609	0.692
W&M (2010)	Wang and Manning (2010)	0.595	0.695
Yao (2013)	Yao et al. (2013)	0.631	0.748
S&M (2013)	Severyn and Moschitti (2013)	0.678	0.736
Shnarch (2013) - Backward	Shnarch (2013)	0.686	0.754
Yih (2013) - LCLR	Yih et al. (2013)	0.709	0.770
Yu (2014) - TRAIN-ALL bigram+count	Yu et al. (2014)	0.711	0.785
W&N (2015) - Three-Layer BLSTM+BM25	Wang and Nyberg (2015)	0.713	0.791
Feng (2015) - Architecture-II	Tan et al. (2015)	0.711	0.800
S&M (2015)	Severyn and Moschitti (2015)	0.746	0.808
W&I (2015)	Wang and Ittycheriah (2015)	0.746	0.820
Tan (2015) - QA-LSTM/CNN+attention	Tan et al. (2015)	0.728	0.832
dos Santos (2016) - Attentive Pooling CNN	dos Santos et al. (2016)	0.753	0.851
Wang et al. (2016) - Lexical Decomposition and Composition	Wang et al. (2016)	0.771	0.845

Bag of words, Word alignment, Dependency Tree Matching

Deep Neural Networks, LSTM

http://aclweb.org/aclwiki/index.php?title=Question_Answering_(State_of_the_art)

Answer Sentence Selection

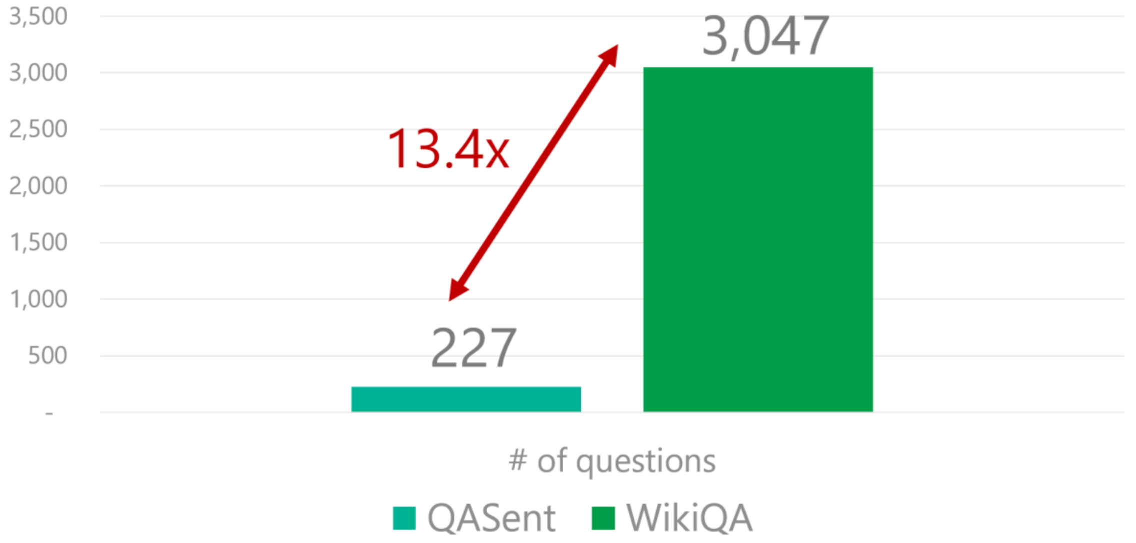
- Dataset

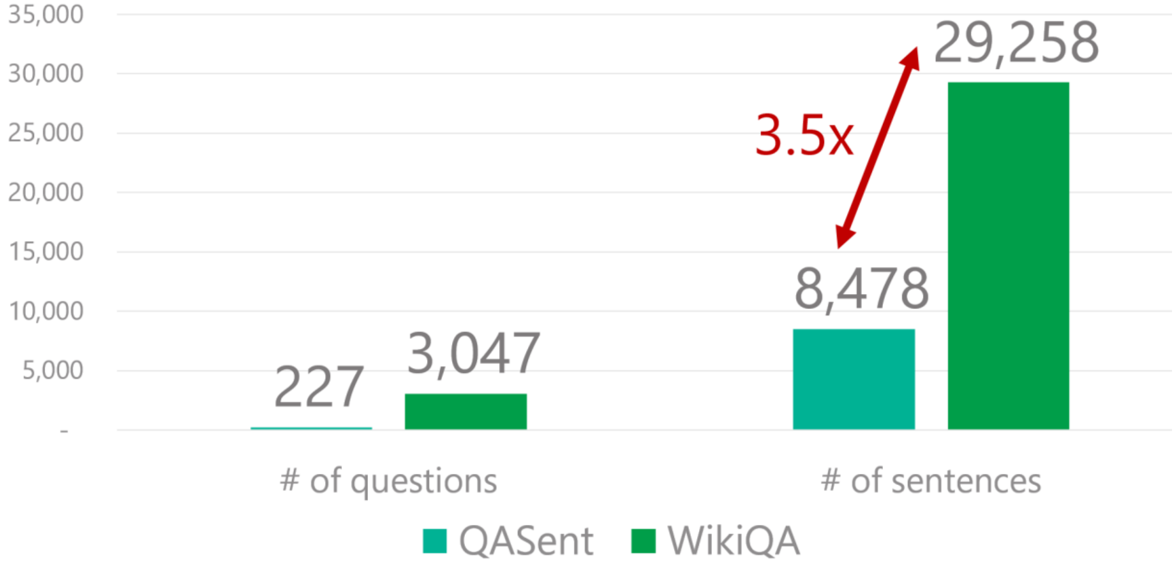
QASent

	Train	Dev	Test	Total
# of ques.	94	65	68	227
# of sent.	5,919	1,117	1,442	8,478
# of ans.	475	205	248	928
Avg. len. of ques.	11.39	8.00	8.63	9.59
Avg. len. of sent.	30.39	24.90	25.61	28.85

WikiQA

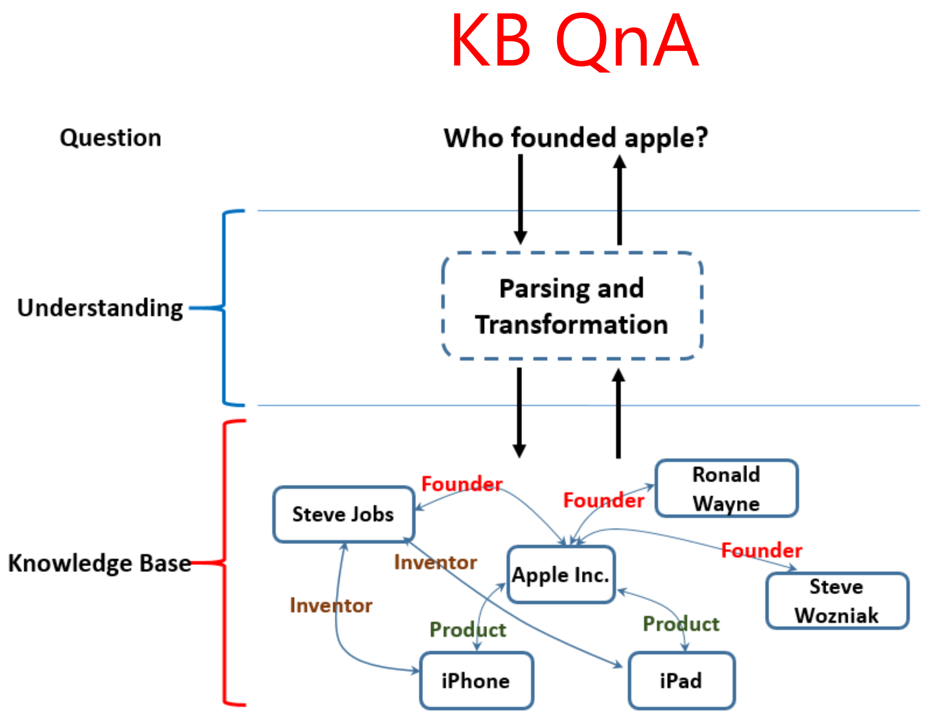
	Train	Dev	Test	Total
# of ques.	2,118	296	633	3,047
# of sent.	20,360	2,733	6,165	29,258
# of ans.	1,040	140	293	1,473
Avg. len. of ques.	7.16	7.23	7.26	7.18
Avg. len. of sent.	25.29	24.59	24.95	25.15
# of ques. w/o ans.	1,245	170	390	1,805

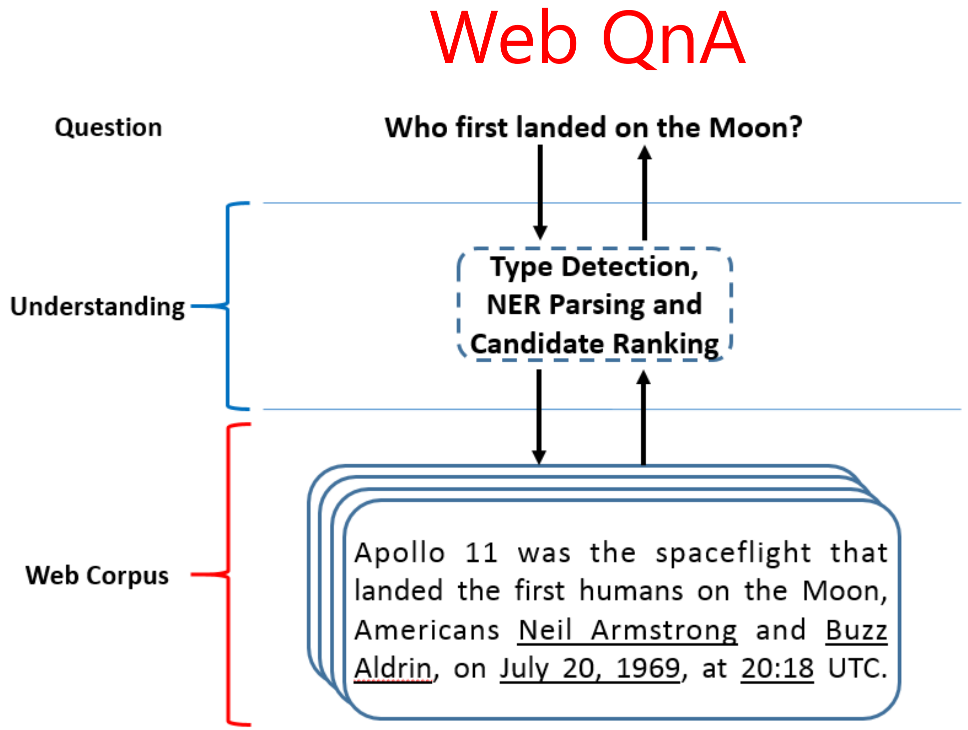




WikiQA: A Challenge Dataset for Open-Domain Question Answering [Yi Yang, et al., EMNLP 2015]

Factoid Answer based on Web Documents





Question Answering via Semantic Enrichment

Question **Who first landed on the Moon?**

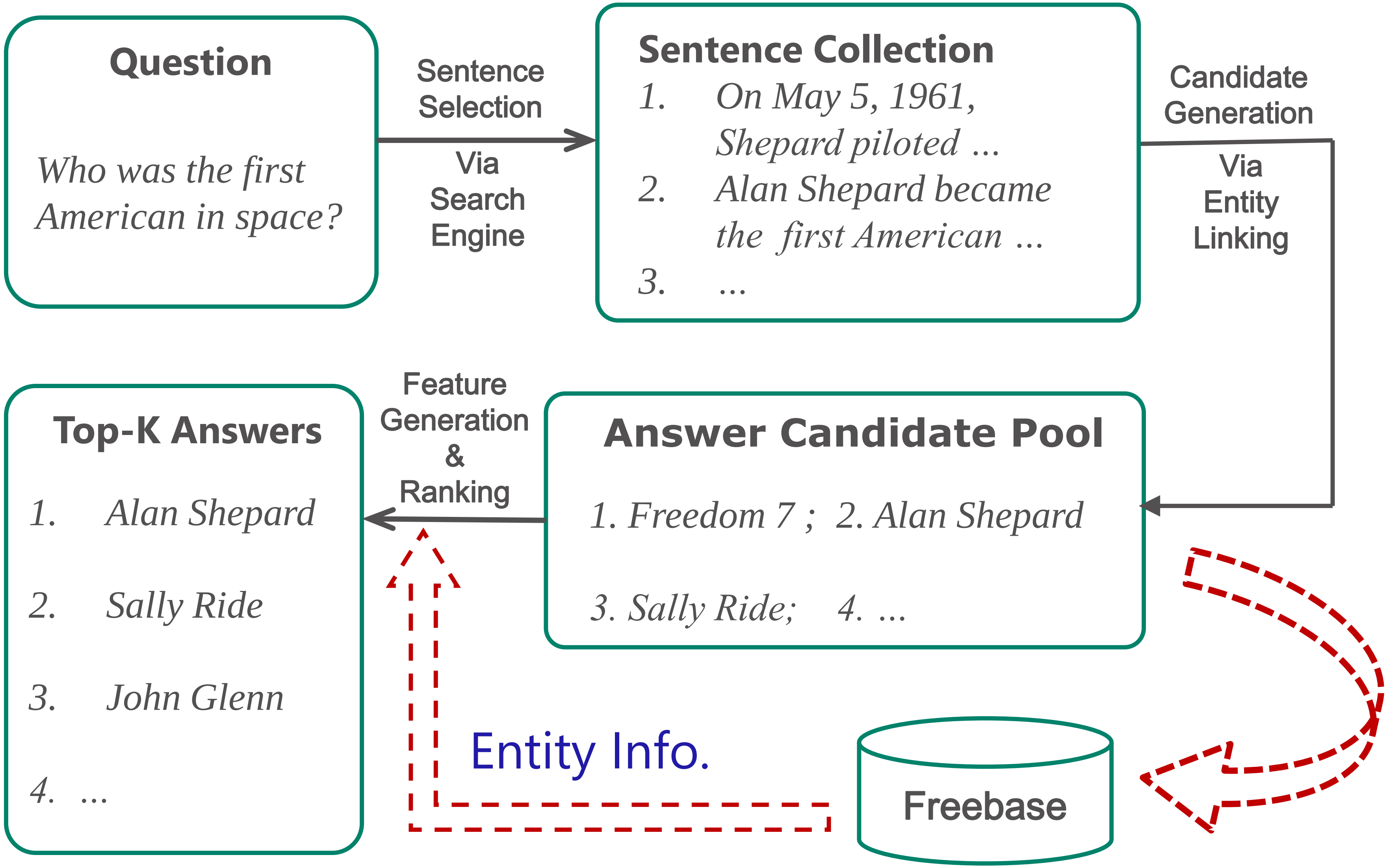
Advantages:

- Generate better answer candidates
  - Entities in Freebase
  - Mentions of the same entity merged to one candidate
- Able to leverage entity information in Freebase
  - Semantic text relevance features for ranking
  - More fine-grained answer type checking

5% ~ 20% improvement in MRR

Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

System Framework



Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

Experiments - Data

- TREC Datasets (well-formed questions)
  - Training: 1,700 (entity) questions (TREC 8-11)
  - Testing: 202 (entity) questions (TREC 12)

Example questions:
1. What are pennies made of?
2. What is the tallest building in Japan?
3. Who sang "Tennessee Waltz"?

- Bing Queries (queries with question intent)
  - Training: 4,725 queries; Testing: 1,164 queries

Example queries:
1. the highest flying bird
2. indiana jones named after
3. designer of the golden gate bridge

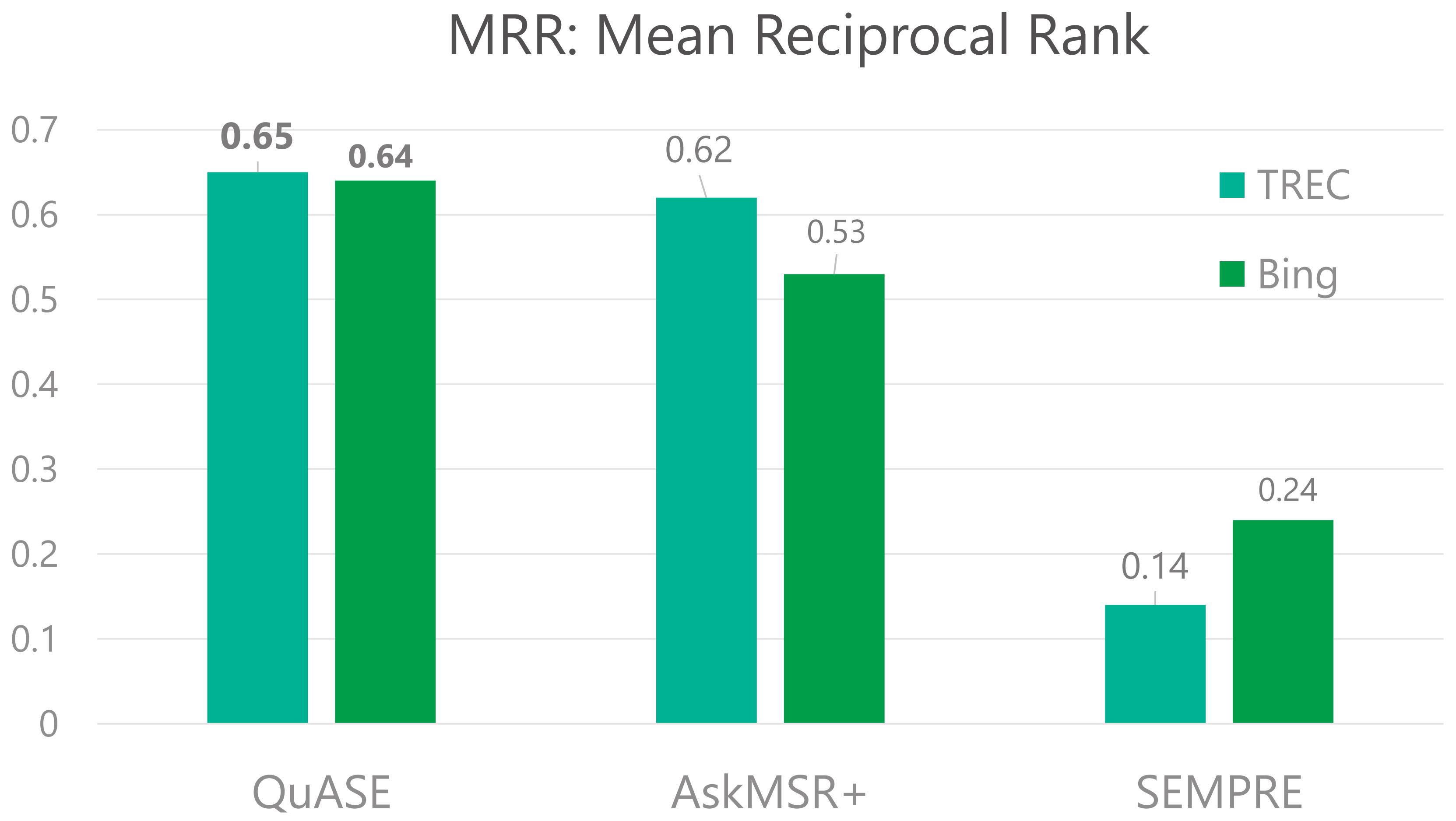
Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

Systems & Evaluation Metrics

- QuASE (Question Answering via Semantic Enrichment)
  - Includes other basic features (e.g., candidate freq.)
  - Ranker learner: MART (Multiple Additive Regression Trees)
- Baselines
  - AskMSR+ [Tsai+ ‘15] – Web-based QA system
  - SEMPRE [Berant+ ‘14] – Semantic parsing QA using Freebase
- Evaluation Metrics
  - MRR: Mean Reciprocal Rank
    - Determined by the top-ranked correct answer

Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

Experiments – Results



Open Domain Question and Answering via Semantic Enrichment [Huan Sun, et al., WWW 2015]

Factoid Answer based on Tables

- What if answers cannot be found through KB and Web Documents?

Knowledge Bases

Structured

Tables

Category	Structure	Country	City	Height (metres)	Height (feet)
Mixed use	Burj Khalifa	United Arab Emirates	Dubai	829.8	2,722
Self-supporting tower	Tokyo Skytree	Japan	Tokyo	634	2,080
Mixed use	Shanghai Tower	China	Shanghai	632	2,073
Clock building	Abraj Al Bait Towers	Saudi Arabia	Mecca	601	1,972
Military structure	Large masts of INS Kattabomman	India	Tirunelveli	471	1,545
Mast radiator	Lualualei VLF transmitter	United States	Lualualei, Hawaii	458	1,503
Twin towers	Petronas Twin Towers	Malaysia	Kuala Lumpur	452	1,482
Residential	432 Park Avenue	United States	New York	425.5	1,396
Chimney	Ekibastuz GRES-2 Power Station	Kazakhstan	Ekibastuz	419.7	1,377
Radar	Dimona Radar Facility	Israel	Dimona	400	1,312
Lattice tower	Kiev TV Tower	Ukraine	Kiev	385	1,263
Electricity pylon	Zhoushan Island Overhead Powerline Tie	China	Zhoushan	370	1,214

Semi-Structured

Web Documents

Unstructured

Factoid Answer based on Tables

- What if answers cannot be found through KB and Web Documents?

Q: Where is the largest brick dome?

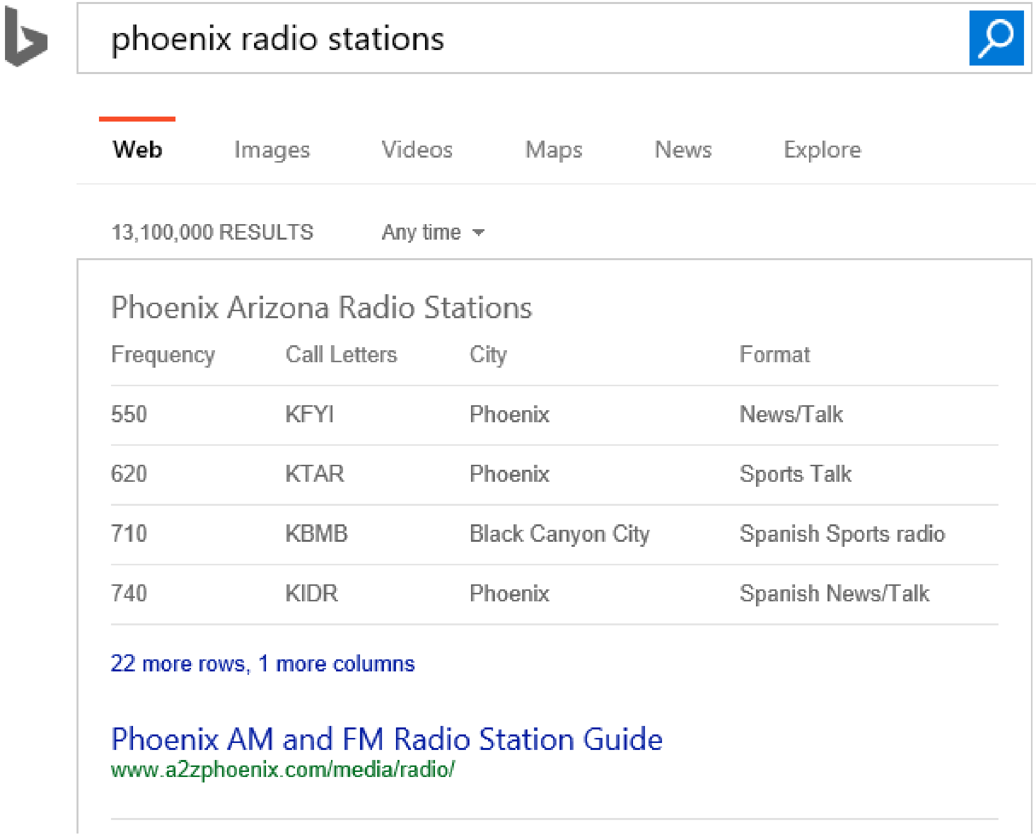
Below is a list of buildings that have held the title of the largest dome on their continent.

Europe [edit]

Held record	Diameter	Name	Location	Builder	Comment
1250 BC–1st century BC	14.5 m[1]	Treasury of Atreus	Mycenae, Greece	City state of Mycenae	Corbel dome
1st century BC–19 BC	21.5 m[2]	Temple of Mercury	Baiae, Italy	Roman Empire	First monumental dome[3]
1436–1881	45.52	Santa Maria del Fiore	Florence, Italy	Roman Catholic Archdiocese of Florence	Largest brick and mortar dome in the world till present. Octagonal dome.

Factoid Answer based on Tables

- What if answers cannot be found through KB and Web Documents?


phoenix radio stations

Web Images Videos Maps News Explore

13,100,000 RESULTS Any time

Phoenix Arizona Radio Stations

Frequency	Call Letters	City	Format
550	KFYI	Phoenix	News/Talk
620	KTAR	Phoenix	Sports Talk
710	KBMB	Black Canyon City	Spanish Sports radio
740	KIDR	Phoenix	Spanish News/Talk

22 more rows, 1 more columns

Phoenix AM and FM Radio Station Guide
www.a2zphoenix.com/media/radio/

Factoid Answer based on Tables

- Knowledge Bases/Graphs
  - Structured but incomplete

- Unstructured Texts
  - Completely no structure

- Semi-Structured Tables
  - Rich: hundreds of millions tables [Lehmberg et al, WWW'16]
  - Schema
    - Table caption
    - Column names
    - Table cells

University	City	Province	Established
University of Alberta	Calgary	Alberta	1906
University of Toronto	Toronto	Ontario	1827
University of Montreal	Montreal	Quebec	1878

List of universities in Canada

Factoid Answer based on Tables

Given:

Question: What languages do people in France speak?

Table Database:

Country	Capital	Location	Main Language	Currency
Algeria	Algiers	Africa	Arabic, French	Dinar
France	Paris	Europe	French	Euro
Hungary	Budapest	Europe	Hungarian	Forint
Singapore	Singapore	Asia	Malay, Chinese, Tamil	Singapore Dollar

The goal: to find a table cell containing answers.

Answer: French

Evidence:

Country	Main Language
France	French

Source: http://hasibul.info/gk/countries.php

Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

- Too many tables! How to find related ones?

  “What languages do people in France speak?”

  More than 100K tables contain “France” !

- How to precisely identify the answer cell?

  “What languages do people in France speak?”

  Capital? Main Language? Currency?

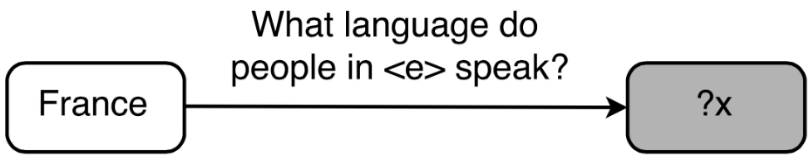
Country	Capital	Currency	Main Language
Algeria	Algiers	Dinar	Arabic
Egypt	Cairo	Pound	Arabic
France	Paris	Euro	French
...	...	...	...

A list of countries and their capital, language etc.

Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

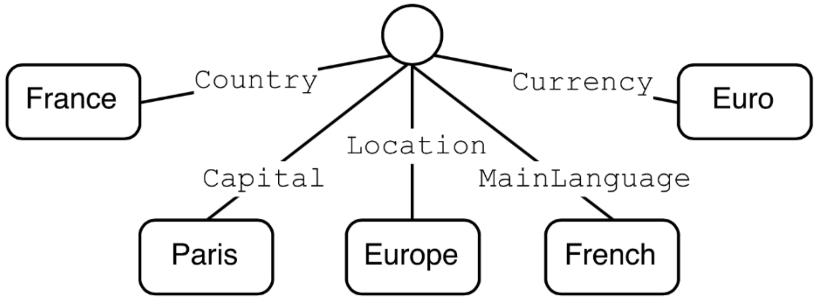
- Question chain



Chain representation for “What languages do people in France speak?”:
entity + question pattern

- Table cell chain

Graph representation
of a table row:



Relational chain between “France” and “French”:

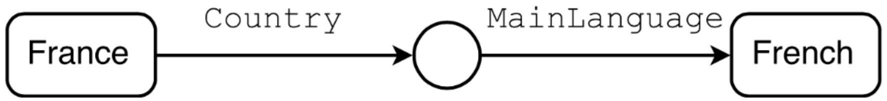


Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

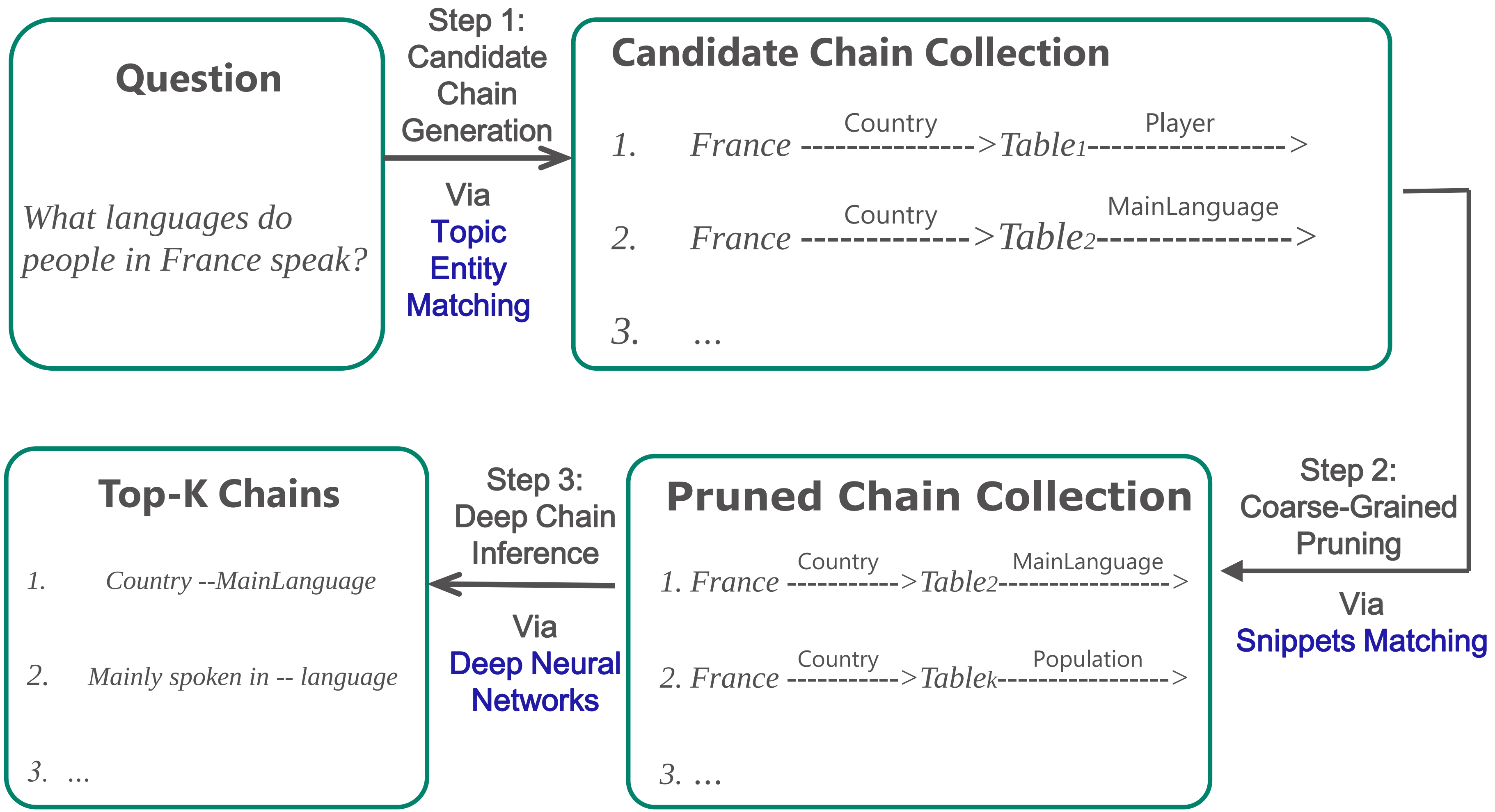


Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

What languages do people in France speak?

Topic entity: France

Step 1:
Candidate
Chain
Generation

String match with table cells

Country	Capital	Currency	Main Language
Algeria	Algiers	Dinar	Arabic
Egypt	Cairo	Pound	Arabic
France	Paris	Euro	French
...	...	...	...

Generate an initial set of chains

{ France $\xrightarrow{\text{Country}}$ Table ID $\xrightarrow{\text{MainLanguage}}$? ;

France $\xrightarrow{\text{Country}}$ Table ID $\xrightarrow{\text{Capital}}$?;...}

Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

Step 2:
Coarse-grained Pruning via Snippets Matching

- Shallow features for each candidate chain

  (1) Candidate chain side
  - Word vector using table title, caption, column names etc.

  (2) Question side
  - Word vector using Bing snippets

- Select top-k candidate chains using shallow features

- Most irrelevant chains can be removed

Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

Step 3:
Deep Chain Inference

Question

What languages do people in France speak?

Candidate chain

1. *France* $\xrightarrow{\text{Country}}$ *Table*2 $\xrightarrow{\text{MainLanguage}}$

2. *France* $\xrightarrow{\text{Country}}$ *Tablek* $\xrightarrow{\text{Population}}$

3. …

Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

Step 3: Deep Chain Inference

Question

What languages do people in France speak?

Candidate chain

1. $\textit{France} \xrightarrow{\text{Country}} \textit{Table}_2 \xrightarrow{\text{MainLanguage}}$
2. $\textit{France} \xrightarrow{\text{Country}} \textit{Table}_k \xrightarrow{\text{Population}}$
3. …

What languages do people in <e> speak

Deep Neural Networks

- Answer type: Column name w.r.t the answer cell
  e.g., “Main Language”
- Pseudo-predicate: Column name pair
  e.g., <country, main language>
- Entity pairs: entity pairs in the two columns
  e.g., { <Spain, Spanish> ; <Italy, Italian>; … }

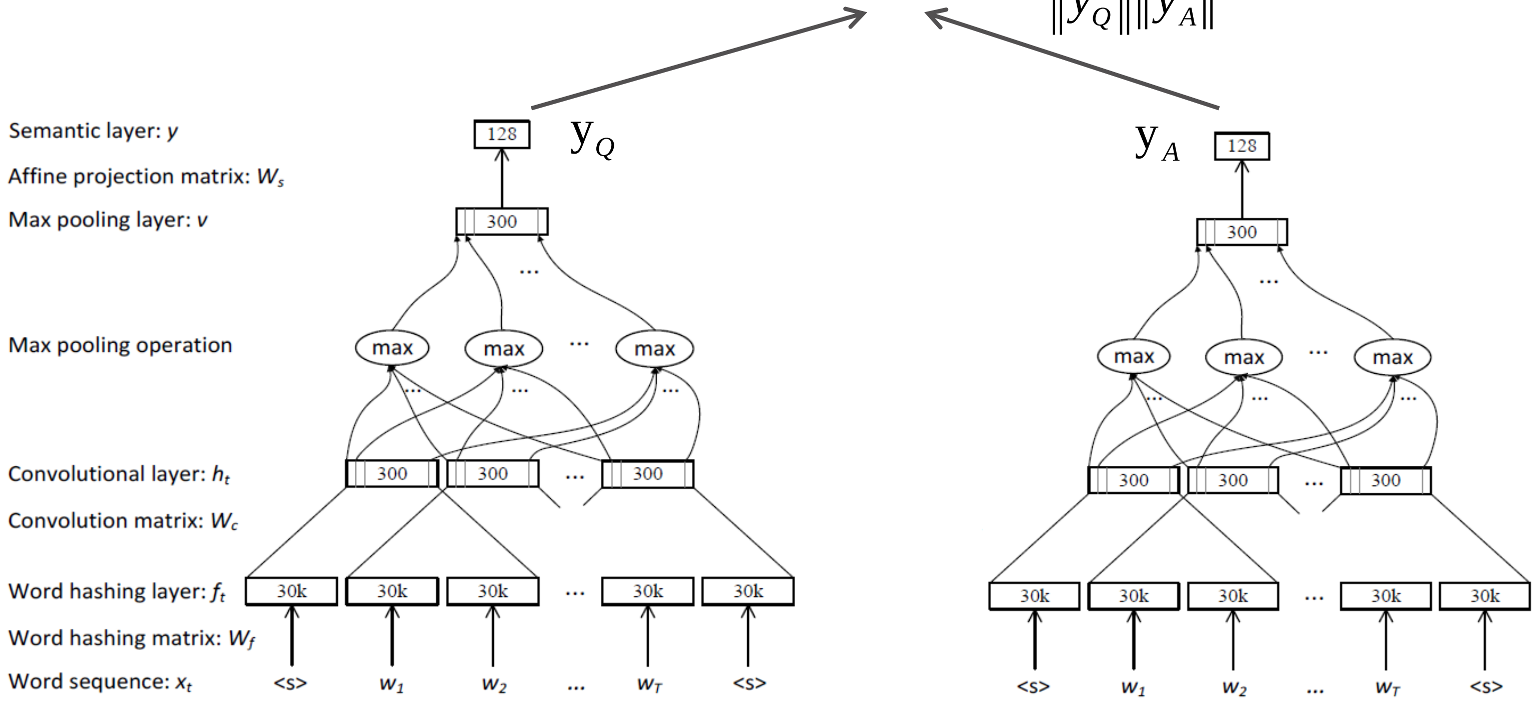
Table Cell Search for Question Answering [Huan Sun, et al., WWW 2016]

Factoid Answer based on Tables

- Deep features
  - <question pattern, answer type>: DSSM-Type
  - <question pattern, pseudo-predicate>: DSSM-Predicate
  - <question pattern, entity pairs>: DSSM-EntityPairs

e.g., DSSM-Type:

$$R(\mathrm{Q},\mathrm{A}) = \cos\mathrm{ine}(\mathrm{y}_Q, \mathrm{y}_A) = \frac{y_Q{}^T y_A}{\|y_Q\|\|y_A\|}$$



Input: Question pattern Answer type

Question Sets

- WebQuestions: WebQ
  - Training: 3,778 (entity) questions
  - Testing: 2,032 (entity) questions

Example questions:
1. who did the voice for lola bunny?
2. in what countries do people speak danish?

- Bing Queries: BingQ
  - Training: 4,725 queries
  - Testing: 1,164 queries

Example queries:
1. cherieff callie voice
2. boeing charleston sc plant location

Table Sets

- WikiTables
  - Tables from Wikipedia and Wikipedia Infoboxes
  - ~5M Tables

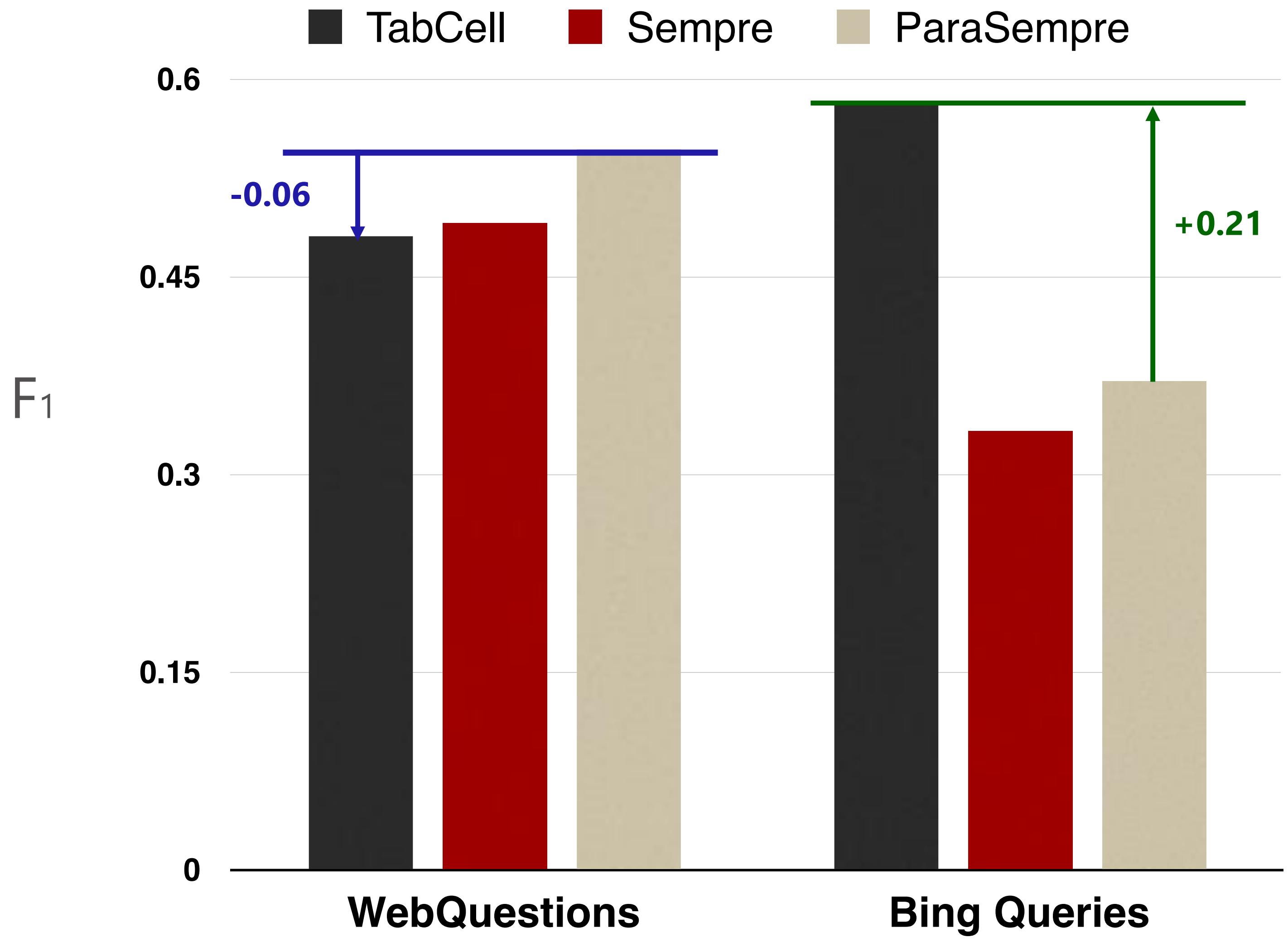
Baselines and Metrics

- TabCell: Table Cell Search
  - Feature set: shallow features, deep features
  - Algorithm: MART (Multiple Additive Regression Trees)
- Baselines: Semantic parsing on Freebase
  - Sempre [Berant et al, EMNLP'13]
  - ParaSempre [Berant et al, ACL'14]
- TabCell + ParaSempre: simply combine their Top-1 results

- Evaluation Metrics
  - Precision, Recall, F_1
    - # of answers in ground truth: N
    - # of true answers contained in top-1 table cell: M
    - Recall = M / N
    - Precision = 0 if M=0; 1 otherwise (b/c, only 1 table cell returned)

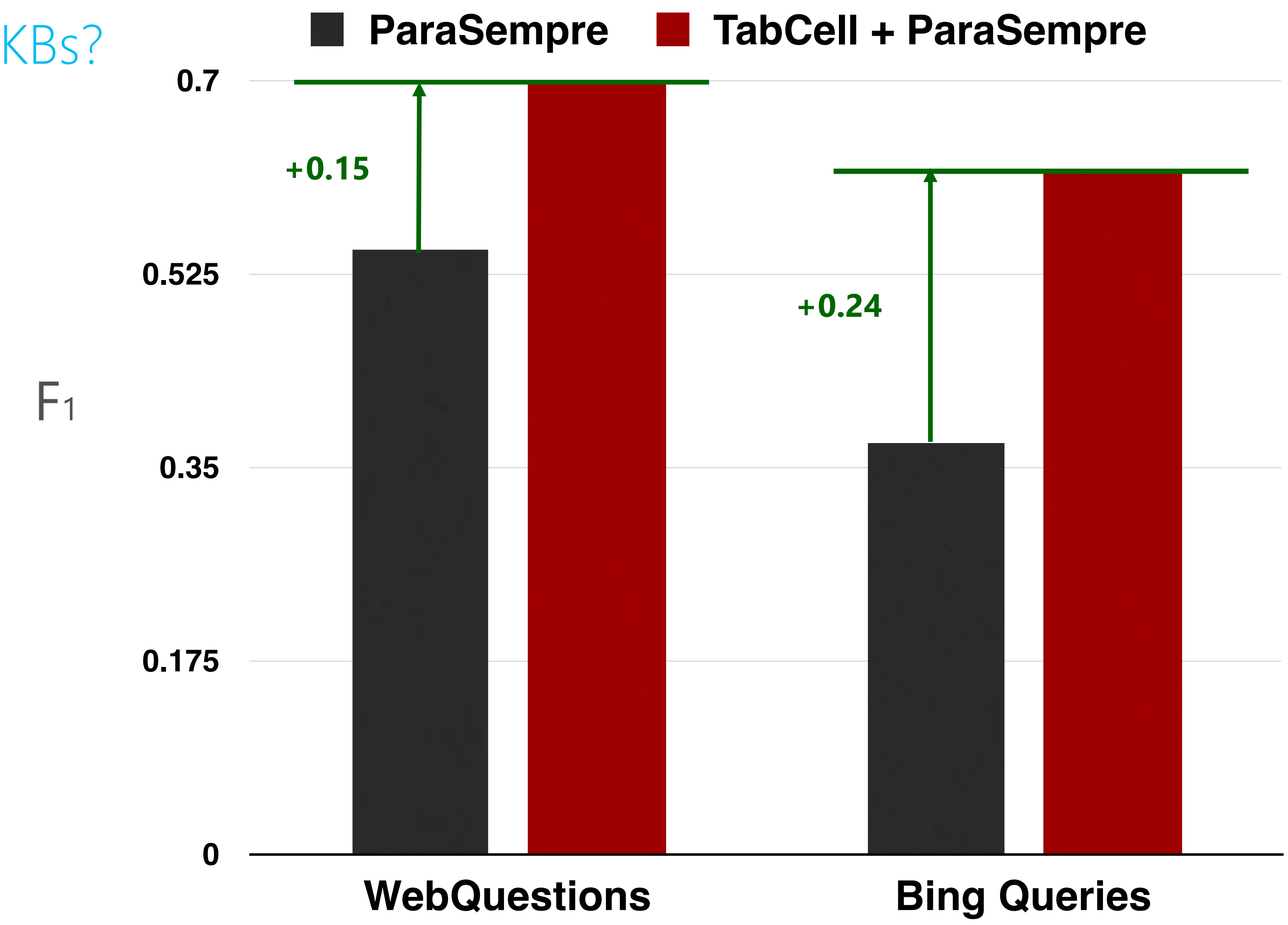
Factoid Answer based on Tables

- How Does TabCell Compare with ParaSempre?



Factoid Answer based on Tables

- Do Tables Complement KBs?



Factoid Answer based on Tables

- Take-away Messages
  - Tables contain rich knowledge to complement knowledge bases.
  - QA based on tables calls for deep understanding table semantics, e.g., column meaning and relations among columns.

Challenges in Web-based QA

when will be the end of the world

All News Images Videos Maps More Search tools

About 1,140,000,000 results (0.60 seconds)

4) An asteroid will hit on May 16, **2016** – followed by a black hole created by CERN.
The end of the world is nigh, appaz (Picture Alamy)
The world will be over by October 25, according to Pastor Ricardo Salazar, who's behind a series of very, very odd YouTube rants. Jan 4, 2016

5 reasons the world is going to end this year, probably on February 14 ...
metro.co.uk/.../5-reasons-the-**world**-is-going-to-**end**-this-year-probably-on-februa... Metro

Challenges in Web-based QA

who will be the president in 2016

All News Images Videos Shopping More Search tools

About 445,000,000 results (0.67 seconds)

Who will win the 2016 U.S. presidential election?

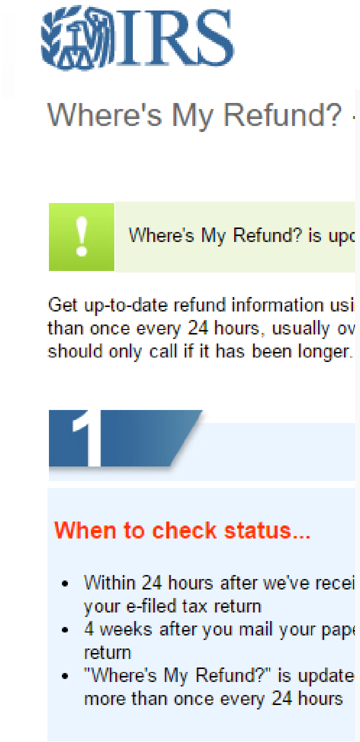
USPREZ16	Latest	Buy Yes
Hillary Clinton CLINTON.USPREZ16	65¢ 1¢	65¢
Donald Trump TRUMP.USPREZ16	34¢ 1¢	35¢
Joe Biden BIDEN.USPREZ16	4¢ NC	4¢
Bernie Sanders SANDERS.USPREZ16	3¢ 2¢	4¢

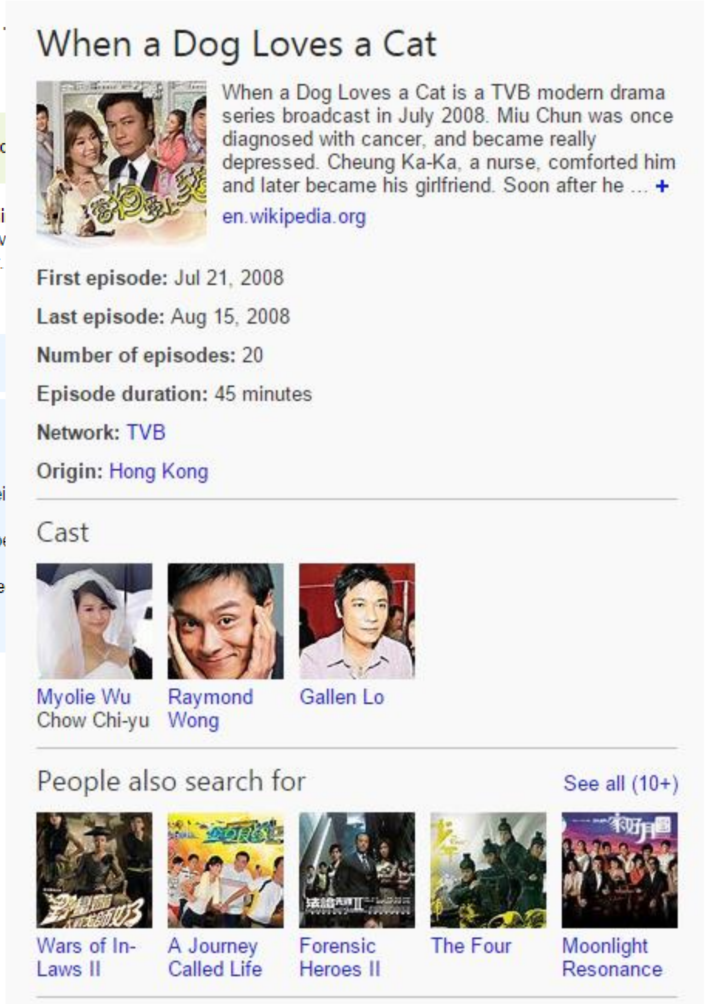
28 more rows, 3 more columns

PredictIt | Who will win the 2016 U.S. presidential election?
https://www.predictit.org/market/1234/**who-will**-win-the-**2016**-us-**presidential**-election

Challenges in Web-based QA

- Question Understanding
  - Rules are not always correct
  - "where is my refund"
    - location?
    - When and how to get refund
  - "when a cat loves a dog"
    - Date Time?
    - TV series





An Incomplete List of Academic Papers on Web QnA

- Dumais et al. “Web Question Answering: Is More Always Better? ” SIGIR-2002.
- Brill et al. “An analysis of the AskMSR question-answering system.” EMNLP-2002.
- Chu-Carroll et al. “A multi-strategy and multi-source approach to question answering.” Technical report-2006.
- Ko et al. “A probabilistic graphical model for joint answer ranking in question answering.” SIGIR-2007.
- Schlaefer et al. “A pattern learning approach to question answering within the ephyra framework.” TSD-2006.
- Azari et al. “Web-Based Question Answering: A Decision-Making Perspective.” UAI-2003.
- Ravichandran et al. “Learning surface text patterns for a Question Answering system.” ACL-2002.
- Kwok et al. “Scaling question answering to the web.” TOIS-2001.
- Brill et al. “Data-intensive question answering.” TREC-2001.
- Bian et al. “Finding the Right Facts in the Crowd: Factoid Question Answering over Social Media.” WWW-2008.
- Cheng et al. “EntityRank: Searching Entities Directly and Holistically.” VLDB-2007.
- Lin et al. “Question answering from the web using knowledge annotation and knowledge mining techniques.” CIKM-2003.
- Chaturvedi et al. “Joint question clustering and relevance prediction for open domain non-factoid question answering.” WWW-2014
- Huan et al. “Open Domain Question Answering via Semantic Enrichment”. WWW-2015
- Huan et al. “Table Cell Search for Question Answering”. WWW-2016

Question Answering for Testing Machine Intelligence

A Different Kind of Question Answering…

- Story comprehension (MCTest)
- Fill-in-the-blank questions (MSR sentence completion, DeepMind Q&A Dataset, Facebook Children Stories)
- Commonsense reasoning (Facebook bAbI)
- Quiz competition (Quiz Bowl, Jeopardy!)
- Standard test for measuring AI (AI2)
- Visual Question Answering

A Different Kind of Question Answering…

- Story comprehension (MCTest)
- Fill-in-the-blank questions (MSR sentence completion, DeepMind Q&A Dataset, Facebook Children Stories)
- Commonsense reasoning (Facebook bAbI)
- Quiz competition (Quiz Bowl, Jeopardy!)
- Standard test for measuring AI (AI2)
- Visual Question Answering

Story Comprehension – Early Work

- *Charniak. "Toward A Model of Children's Story Comprehension." PhD Dissertation. 1972.*
  - Model the world knowledge
  - Understand natural language

- *Hirschman et al. "Deep Read: A Reading Comprehension System." ACL-1999.*
  - A small reading comprehension dataset (3^{rd} to 6^{th} grade stories)
  - Find sentences to answer "who/what/when/where/why" questions
  - Simple BoW approach reaches 40% accuracy (~5% random)

MCTest: Reading Comprehension Test [Richardson+, EMNLP-13]

- 660 children’s stories, 2,640 comprehension questions
- Data collection: Crowdsourcing via Amazon MTurk
  - **No copyright issues, freely downloadable**
- Fictional: Answers are found only in the story
- Grade-school level: limited vocabulary (8,000 words)
- Multiple-choice: objective/offline evaluation
- Open-Domain

Sample Story

Timmy liked to play games and play sports but more than anything he liked to collect things. He collected bottle caps. He collected sea shells. He collected baseball cards. He has collected baseball cards the longest. He likes to collect the thing that he has collected the longest the most. He once thought about collecting stamps but never did. His most expensive collection was not his favorite collection. Timmy spent the most money on his bottle cap collection.

1) Timmy liked to do which of these things the most?
   A) Collect things
   B) Collect stamps
   C) Play games
   D) Play sports

2) Which is Timmy's most expensive collection?
   A) Stamps
   B) Baseball Cards
   C) Bottle Cap
   D) Sea Shells

3) Which item did Timmy not collect?
   A) Bottle caps
   B) Baseball cards
   C) Stamps
   D) Sea shells

4) Which item did Timmy like to collect the most?
   A) Stamps
   B) Baseball cards
   C) Bottle caps
   D) Sea shells

Courtesy of Matt Richardson

Baselines

- Window Algorithm:
  - S = question + hypothesized answer
  - Score: best matching $|S|$-sized window in story
  - Answer with best score wins
- Distance Algorithm:
  - For each word in question, find distance in story to the nearest word in answer
  - Answer with lowest average distance wins

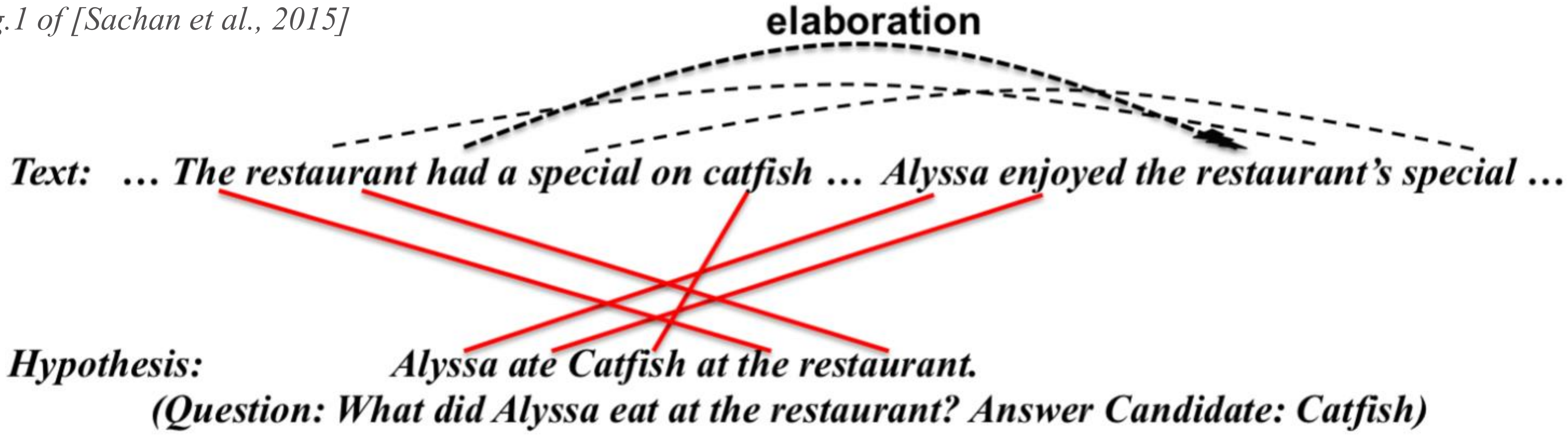
- MC500 Test Questions: W+D: 60.26% Accuracy

Fostered Research on a Variety of Approaches

- Lexical matching [Smith et al., 2015]
- Discourse processing [Narasimhan and Barzilay, 2015]
- Rules [Chen et al., 2015]
- Semantic frames [Wang et al., 2015]
- Memory Networks [Kapashi et al., 2015]
- Answer-entailing structures [Sachan et al., 2015]
- Attention-based CNNs [Yin et al., 2016]
- Parallel-Hierarchical NN [Trischler et al., 2016]

Answer-entailing structures [Sachan et al., 2015]

Fig.1 of [Sachan et al., 2015]



- Latent structured SVMs with rich features
  - Lexical semantic features based on SENNA word vectors & WordNet
  - RST (Rhetorical Structure Theory) tags for cross-sentence relations
- Best accuracy: 67.83% (with multitask learning)

Parallel-Hierarchical NN [Trischler et al., 2016]

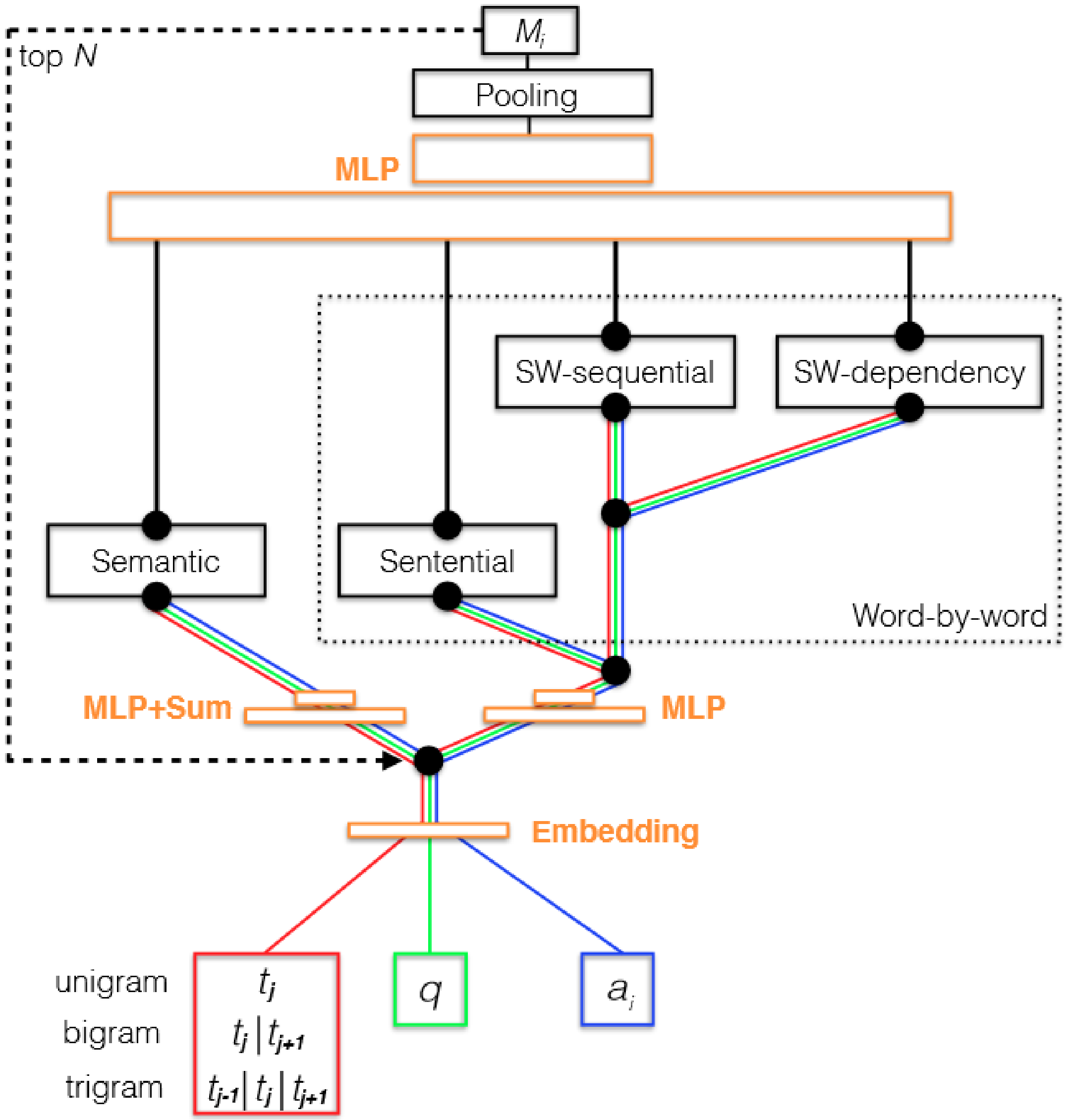


Fig.1 of [Trischler et al., 2016]

- Embed document and question/answer
- Combine multiple *perspectives*
  - Text semantic vectors
  - Sentential vectors
  - Sliding window based on words and dependency trees
- Accuracy: 71.0%

Story Comprehension – Summary

- Simple baselines are strong (~60% vs. 25% random)
- ML-based “text matching” approaches are winning
  - LSSVMs + multitask leaning → 67.8%
  - Neural networks + word embedding → 70.0%

- Reasoning process is not easily interpretable
  - No explicit world knowledge or model has been used
  - Cannot provide explanations on why the answers are chosen
- Still room for improvement (the ceiling is 100%)

Fill-in-the-blank Quiz Questions

“At last she looked up with something ______ and defiant in her manner.”

a) *reckless*
b) *solid*
c) *pallid*
d) *jovial*
e) *warm*

Example from [MSR Sentence Completion Challenge]()

Fill-in-the-blank Quiz Questions

- Motivation
  - Same high-level goal as MCTest
  - Seeking a more scalable way to collect data (e.g., vs. crowdsourcing)
    - MCTest dataset might be too small for supervised learning, especially for NN approaches
- High-level process
  - Pick a large corpus (e.g., news articles, stories)
  - Develop an (almost) automatic way to generate (fill-in-the-blank) questions

DeepMind Q&A Dataset [Hermann et al., NIPS-15]

- 93k CNN & 220k Daily Mail articles
- Bullet points (summary / paraphrases) → Cloze questions
  - Replacing one entity with a placeholder
  - ~4 questions per document
  - ~1M document / query / answer triples

- Datasets recreated by Kyunghyun Cho
  - http://cs.nyu.edu/~kcho/DMQA/

Example [Hermann et al., NIPS-15. Table 3]

	Original Version	Anonymised Version
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer		
	Oisin Tymon	*ent193*

Word Counting Baselines

- Majority
  - Pick the most frequently observed entity in the D
- Exclusive majority
  - Same as Major, but the entity is not observed in Q

Symbolic Matching Models

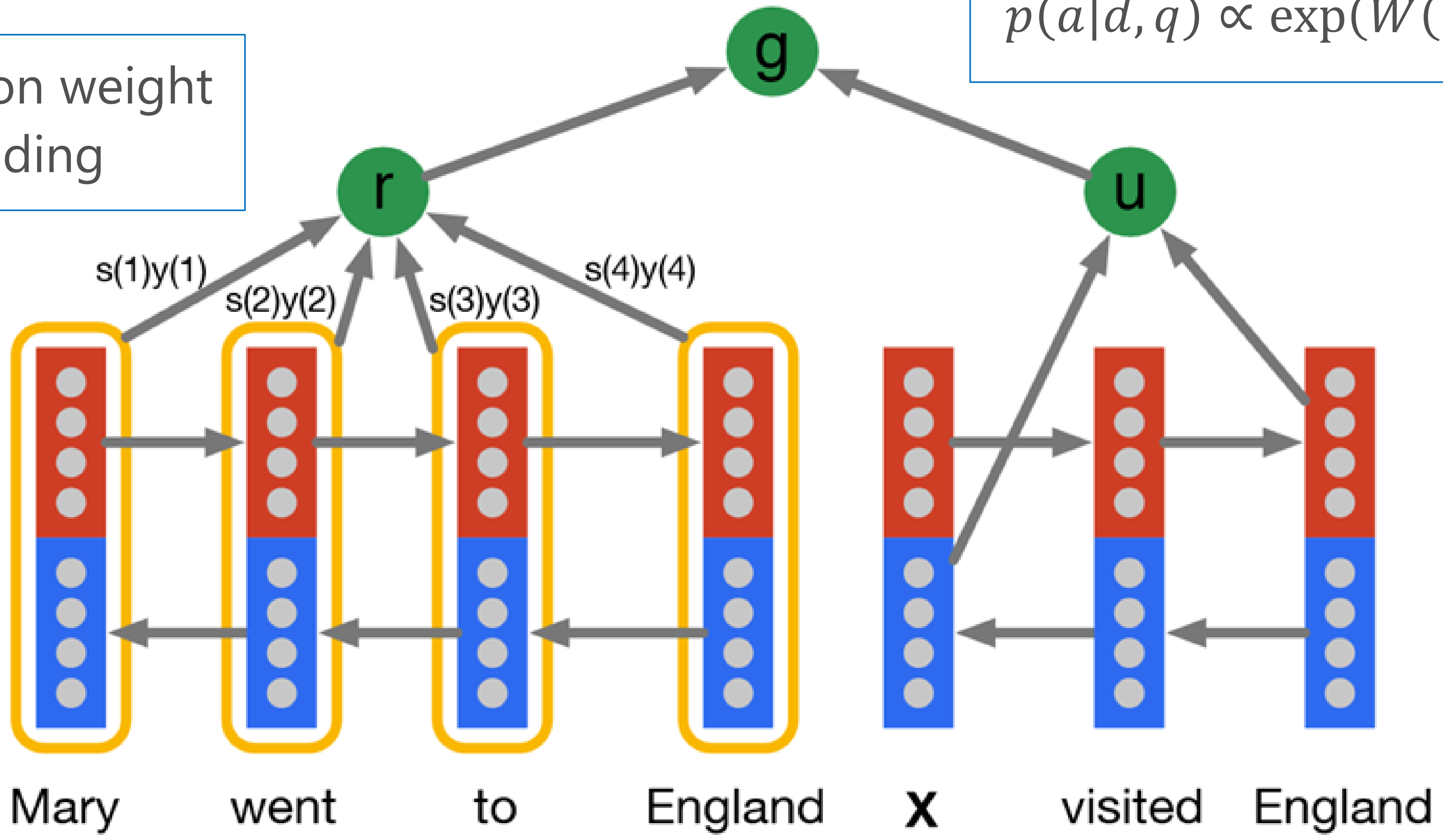
- Frame-semantic parsing
  - Match PropBank triples (x, V, y)
  - Q: "X loves Sue" vs. D: "Kim loves Sue"

- Word distance benchmark
  - Align the placeholder in Q with each possible entity in D
  - Sum the distances of each word in Q to nearest aligned words in D

Neural Network Models – Attentive Reader

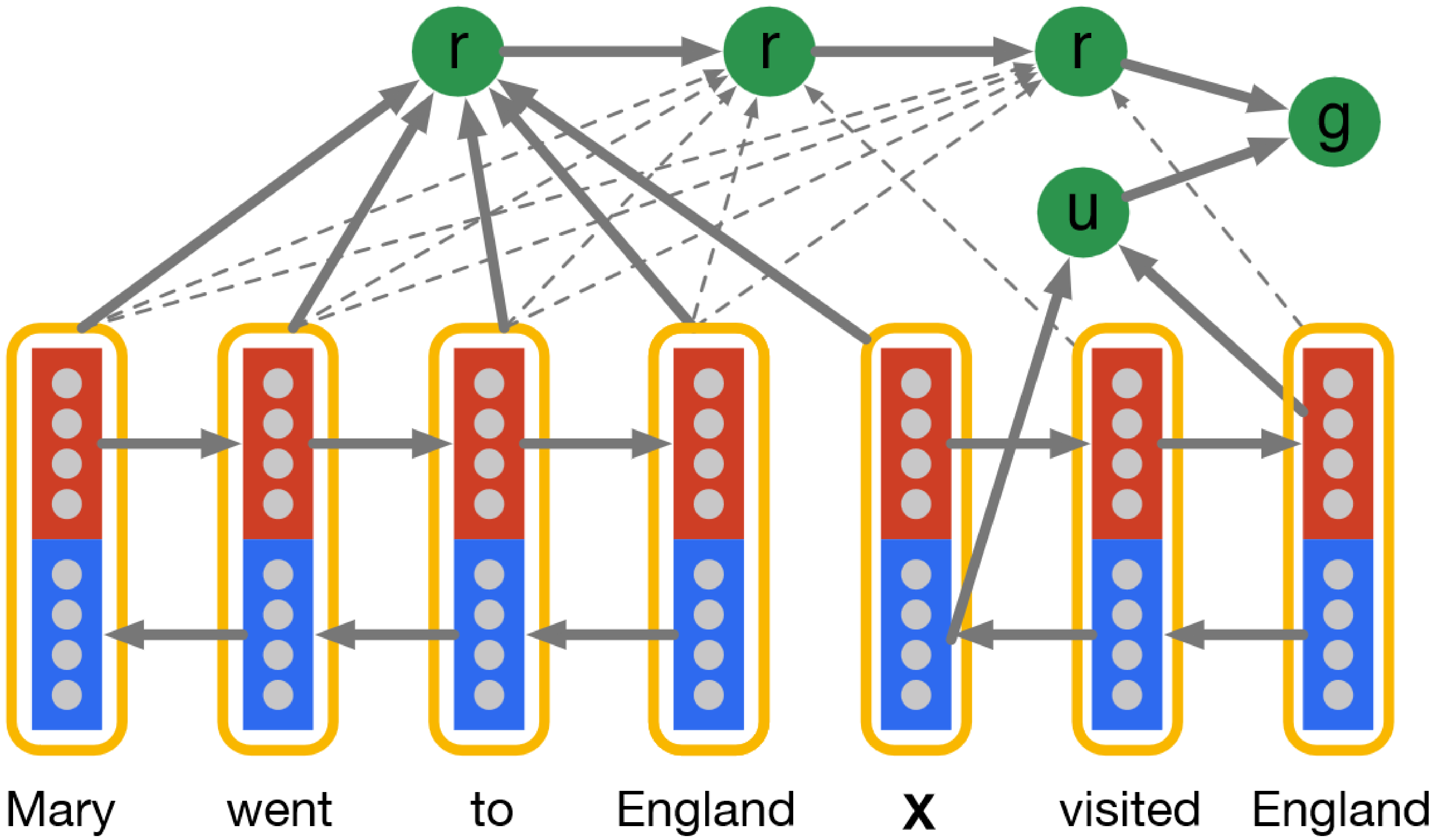
s: attention weight
y: embedding

$p(a|d,q) \propto \exp(W(a)g(d,q))$

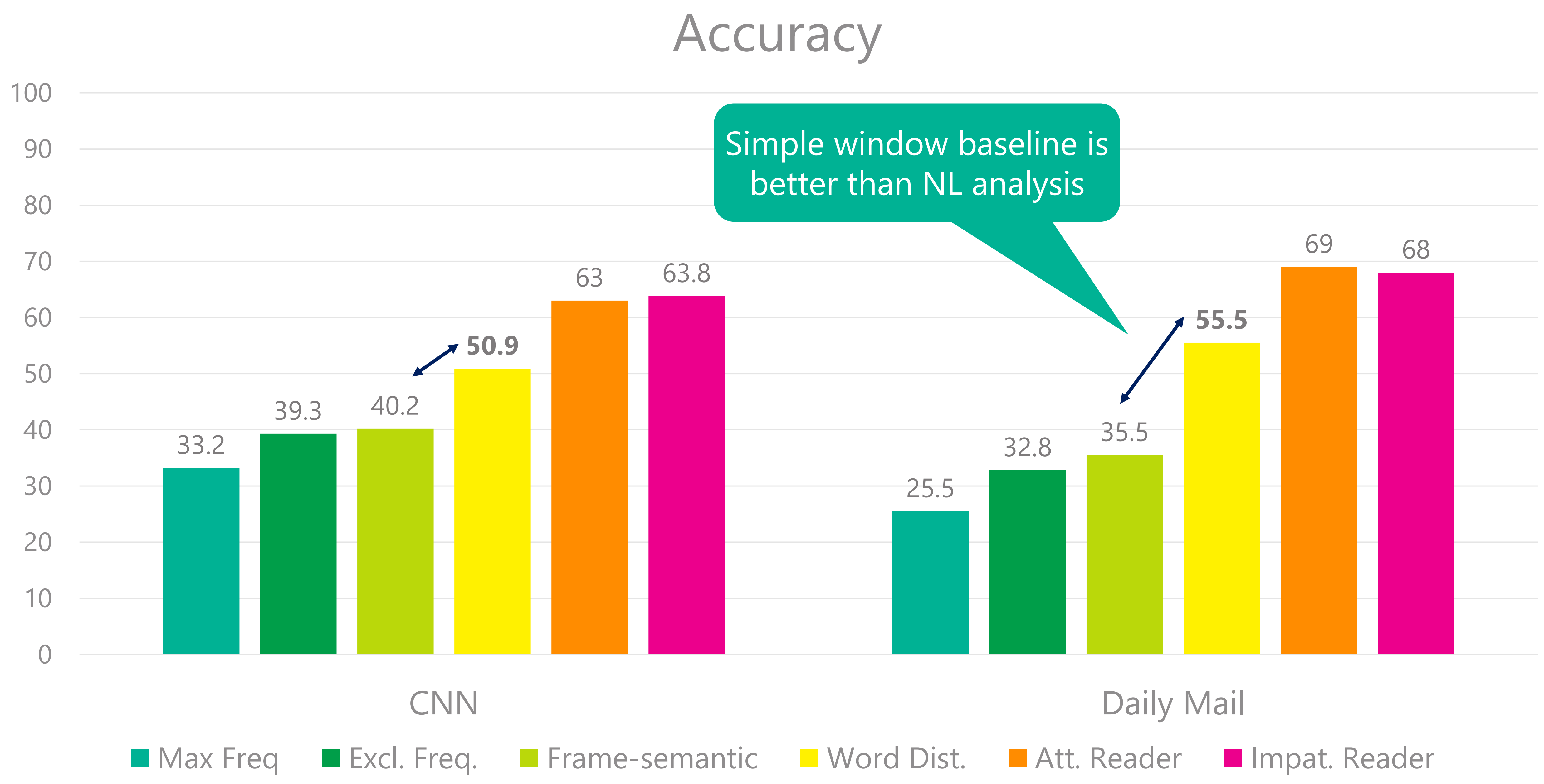


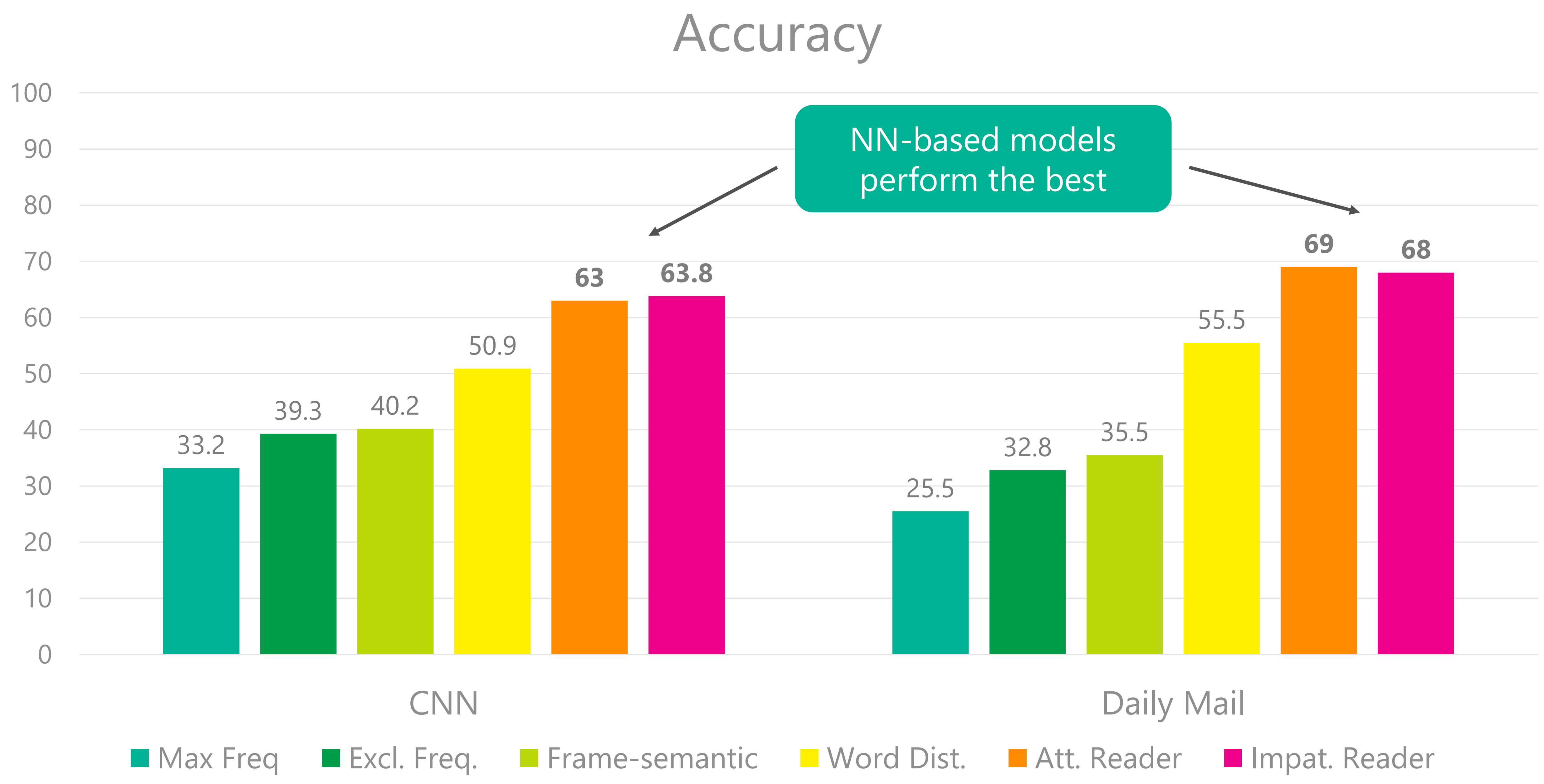
[Hermann et al., NIPS-15. Fig 1a]

Neural Network Models – Impatient Reader



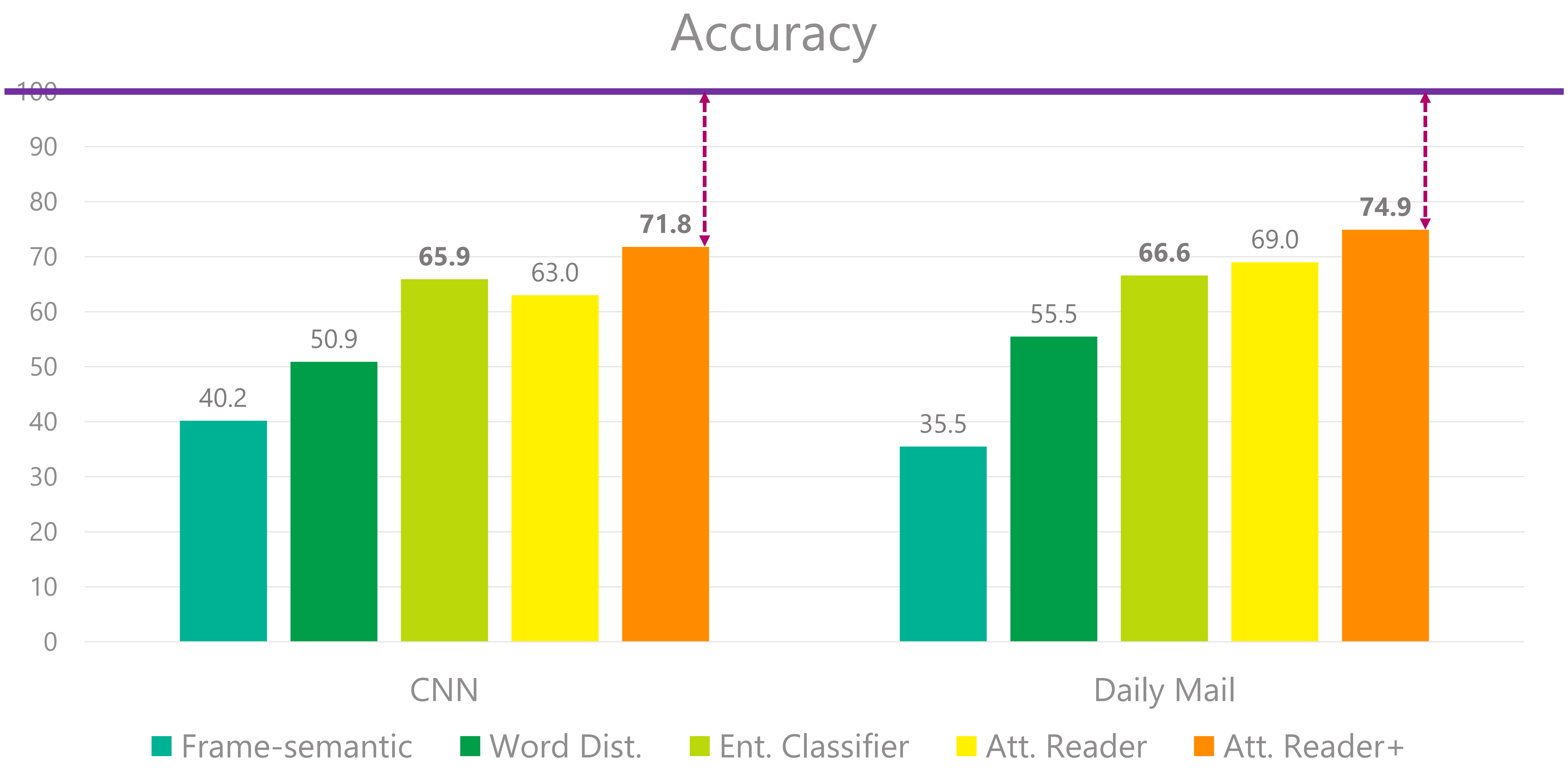
[Hermann et al., NIPS-15. Fig 1b]


Accuracy
Simple window baseline is better than NL analysis
100
90
80
70
60
50
40
30
20
10
0
33.2
39.3
40.2
50.9
63
63.8
25.5
32.8
35.5
55.5
69
68
CNN
Daily Mail
Max Freq
Excl. Freq.
Frame-semantic
Word Dist.
Att. Reader
Impat. Reader


Accuracy
NN-based models perform the best
100
90
80
70
60
50
40
30
20
10
0
33.2
39.3
40.2
50.9
63
63.8
25.5
32.8
35.5
55.5
69
68
CNN
Daily Mail
Max Freq
Excl. Freq.
Frame-semantic
Word Dist.
Att. Reader
Impat. Reader

A Thorough Examination… [Chen et al. ACL-16]

- Challenges & Questions
  - A clever way of creating large supervised data, but an artificial task
  - Unclear what level of reading comprehension needed
- Good News – The task is not really difficult!
  - An entity-centric classifier with simple features works fine
  - A variant of the Attentive Reader model achieves the new best result
- Bad News – The task is not really difficult!
  - Not much "comprehension" is needed
  - Probably have reached the ceiling



Accuracy
100
90
80
70
60
50
40
30
20
10
0
40.2
50.9
65.9
63.0
71.8
35.5
55.5
66.6
69.0
74.9
CNN
Daily Mail
Frame-semantic
Word Dist.
Ent. Classifier
Att. Reader
Att. Reader+

Analysis on 100 Examples from CNN

Category	Ratio
Exact match	13%
Paraphrasing	41%
Partial clue	19%
Multiple sentences	2%
Coreference errors	8%
Ambiguous / hard (to human)	17%

- 25% questions are not answerable!

Analysis on 100 Examples from CNN

Category	Ratio	Classifier	NN
Exact match	13%	13 (100.0%)	13 (100.0%)
Paraphrasing	41%	29 (70.7%)	39 (95.1%)
Partial clue	19%	14 (73.7%)	17 (89.5%)
Multiple sentences	2%	1 (50.0%)	1 (50.0%)
Coreference errors	8%	3 (37.5%)	3 (37.5%)
Ambiguous / hard (to human)	17%	2 (11.8%)	1 (5.9%)

- 25% questions are not answerable!
- NN handles paraphrases and lexical variations better.

Other Related Tasks & Datasets (1/2)

- MSR Sentence Completion Challenge [Zweig & Burges, 2011]
  - 1,040 sentences from five Sherlock Holmes novels
  - An infrequent word is chosen as the focus of the question
  - 4 alternates chosen by hand from 30 words suggested by LM
  - Random: 25%. Human: 91%. Current Best: 56% [Liu et al., ACL-15]

- Quiz Bowl: paragraph factoid questions [Iyyer et al., EMNLP-14]
  - Predict the entity described by the short paragraph

Other Related Tasks & Datasets (2/2)

- Facebook Children's Book Test [Hill et al., ICLR-16]
  - 20 sentence as context
  - 21st sentence → Cloze question with 10 candidates

- ROCStories and Story Cloze Test Corpora [Mostafazadeh et al., NAACL-HLT-16]
  - 50k five-sentence commonsense stories
  - Given the first 4 sentences, select the correct ending
  - Designed to be 100% answerable by human judges

Facebook bAbI Tasks [Weston et al., ICLR-16]

- 20 categories of simple commonsense reasoning tasks
  - A short description of agents moving around & passing objects
  - Followed by a simple question that can be answered based on the description
  - 1,000/1,000 questions for training/testing

Task 3: Three Supporting Facts

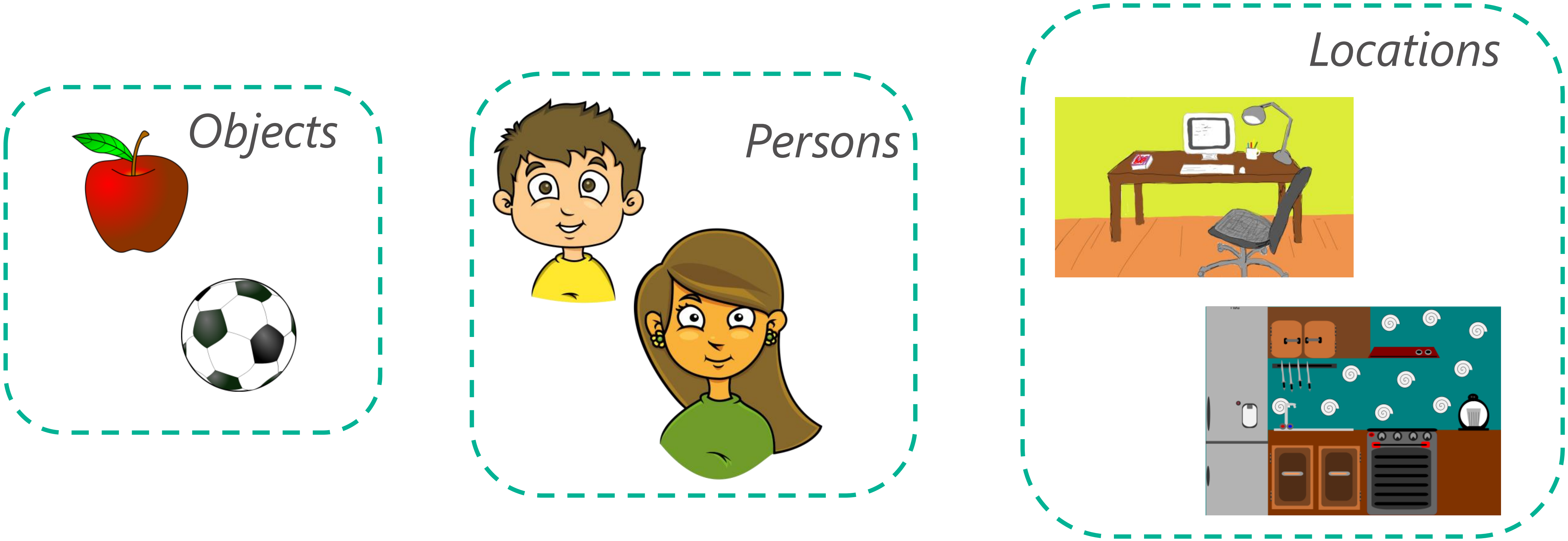
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Arguments for Creating bAbI Tasks

- Categorize different reasoning questions into skill sets

- Claims / Hopes:
  - Analyze model performance on different *skills* to study the strengths and weaknesses
  - Simple language and problems make the results easy to interpret
  - Each task checks a skill that a system should have
  - Mastering all the tasks is a prerequisite for any system with full text understanding and reasoning ability

Task Generation via a Simulated World



- States & properties of entities
- Actions an actor can take (e.g., go <loc>, get <obj>)

Memory Networks [Weston et al. 2014]

- Class of models instead of one model
- Key concepts
  - Explicit memory storage and index
  - Select memory for matching
- Basic components
  - Input feature map: sentence x to an internal representation $I(x)$
  - Generalization: update memory $\boldsymbol{m}$: $\boldsymbol{m}_i = G(\boldsymbol{m}_i, I(x), \boldsymbol{m}), \forall i$.
  - Output feature map: compute output o: $O(I(x), \boldsymbol{m})$
  - Response: decode o to give a textual response $r = R(o)$
- Implementation could be very simple

Memory Networks for bAbI

- Input: embedding of simple bag of words
- Generalization: store embedding of sentences sequentially
- Output: find two supporting facts
  - 1st supporting fact s_1 (max match score [dot product] with question q)
  - 2nd supporting fact s_2 (max match score with s_2 & q)
- Response: rank possible answer words given the facts
  - Based on dot products of the word vector and the embedding of facts

75% accuracy; advanced variation achieves 93% accuracy

Unsolved Tasks

- Counting, Lists/Sets, Positional Reasoning, Path Finding

Task 19: Path Finding

The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

Reasoning in Vector Space [Lee et al., ICRL-16]

- Decouple semantic parsing & logical reasoning
- Two vector-space reasoning models, inspired by Tensor Product Representation [Smolensky 1990/2006]
  - All entities are represented in d-dimensional unit vectors
  - Relation between two entities is described by matrix product (binding)
  - Inference (answering questions) is done by inner product
- 100% accuracy except Categories 5 & 16
  - Incorrect answers & ambiguity in facts

#	Statements/Questions	Encodings
1	Mary went to the kitchen.	mk^T
2	Mary got the football there.	fm^T
3	Mary travelled to the garden.	mg^T
4	Where is the football?	

- Left-multiply by f^T all statements prior to the current time ($f^T \cdot mk^T$, $f^T \cdot fm^T$, $f^T \cdot mg^T$)
- Pick the most recent container where 2-norms are ~1.0 (m^T)
- If the container is an actor
  - Find the most recent container of the actor by left-multiplying by m^T (Yields g^T)
  - Answer by the most recent container. ⇒ garden
- If the container is a location, return it as answer

Some Observations – Dataset Creation

- Synthetic or semi-synthetic
  - ✓ Relatively easy to create large-scale datasets
  - ✗ Datasets may have unexpected issues and thus more *breakable*

- Human generated or validated
  - ✓ Datasets are more natural and real
  - ✓ Could design specific reasoning tasks
  - ✗ Less scalable, even with the help of crowdsourcing

Some Observations – Current Results

- Simple methods often provide strong baselines (vs. random)
- New methods give incremental improvement
- SOTA from statistical methods, but still far behind human

- Reasoning process is hard to interpret
  - For the ease of evaluation, being able to explain the decision process to human is not part of the metric
  - Not clear whether the solutions are *general*

Tutorial Summary – Part 1

- Modern question answering applications
  - Search engines evolve to handle question queries
  - Digital assistants address multi-turn QA
  - Business analytics service adopt natural language QA interface
- Pioneer work on question answering machines
  - Similar problems & applications
  - Limited success, often ad-hoc solutions
  - Constrained by data size, computational power & models

Tutorial Summary – Part 2

- Open-domain factoid question answering with KB
  - Large-scale knowledge bases as the sole information source
  - Find entities or properties of entities in KB to answer questions
- Mainstream approach – semantic parsing of questions
  - Map natural language questions to logical forms / structured queries
  - Accurate answers when parse & KB is complete and correct
  - Able to explain how the answers are derived
  - Challenges: language mismatch, large search space, compositionality

Tutorial Summary – Part 3

- Open-domain factoid question answering with the Web
  - Leverage Web redundancy – commonly asked facts stated frequently in various Web documents
  - Recent approaches to incorporate structured (KB) and semi-structured (Web tables) information sources

- Challenges
  - Difficult in handling domain-specific or tail questions
  - Deeper understanding of questions

Tutorial Summary – Part 4

- Question answering for testing machine intelligence
  - Designed to test AI; Not to fulfill users' information need
  - A long-standing research strategy
- Introduced recently proposed tasks
  - Story comprehension (multiple-choice questions)
  - Fill-in-the-blank questions (find entities)
  - Commonsense reasoning (find answer words)
- Challenges
  - Having a well-designed and large dataset/task

Future

- Conversational intelligence supported by QA
  - No longer an independent task
  - Integrated naturally in a conversational system

- Multi-modal interaction
  - Visual question answering
  - Virtual tour guide