

# Question Answering

Nan DUAN

Microsoft Research Asia

2016-06-01

# Outline

- QA Overview
- Knowledge-based QA
- Latest Trends of QA
- From QA Engines to Conversational Engines

# QA Overview

# Question Answering (QA)

- Definition
  - Answer natural language (NL) questions automatically by machines
- Types of questions
  - Factoid:            "*who is the wife of Barack Obama?*"
  - Definition:        "*what is operating system?*"
  - Yes-No:            "*is Saddam Hussein alive?*"
  - Opinion:           "*what do most Americans think of gun control?*"
  - Comparison:      "*what are the differences between Nokia and iPhone?*"
  - Jeopardy! quiz:    "*On Sept. 1, 1715 Louis XIV died in this city, site of a fabulous palace he built.*"
  - ...

Question-Answering is a key technology to NUI, chat-bot, search engine and personal assistant for mobile users.

# Multiple Intelligence in Modern QA

? Who is the wife of Barack Obama?



**Barack Obama**

Barack Hussein Obama II is the 44th and current President of the United States, and the first African American to hold the office. Born in Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he served as president of the Harvard Law Review. He was a community organizer in Chicago before earnin... +

[en.wikipedia.org](http://en.wikipedia.org)

[Twitter](#) [Facebook](#)

**Born:** Aug 04, 1961 (age 53) · Honolulu, Hawaii

**Net worth:** \$12.20 million USD (2014)

**Spouse:** Michelle Obama (Since 1992)

**Children:** Malia Ann Obama · Natasha Obama

**Office:** President of the United States (2009 - present)

**Previous offices:** United States Senator IL (2005 - 2008) · Illinois State Senator (1997 - 2004)

## Knowledge-QA



Freebase

ProBase



悟 satori

# Multiple Intelligence in Modern QA



? Who is the wife of Barack Obama?

## Document-QA



Google

Who is the wife of barack obama?

网页 图片 新闻 视频 地图 更多 搜索工具

找到约 230,000,000 条结果 (用时 0.25 秒)

[Michelle Obama - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Michelle\\_Obama](http://en.wikipedia.org/wiki/Michelle_Obama) 翻译此页  
Michelle LaVaughn Robinson [Obama](#) (born January 17, 1964) is an American lawyer and writer. She is the [wife](#) of the 44th and current [President](#) of the United States ...  
Craig Robinson - Sidley Austin - Hyde Park, Chicago - Valerie Jarrett

[Family of Barack Obama - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Family\\_of\\_Barrack\\_Obama](http://en.wikipedia.org/wiki/Family_of_Barrack_Obama) 翻译此页  
Michelle Obama, née Robinson, the [wife of Barack Obama](#), was born on January 17, 1964, in Chicago, Illinois. She is a lawyer and was a University of Chicago ...  
Sidwell Friends School - Marian Shields Robinson - Bo - Charles T. Payne

# Multiple Intelligence in Modern QA



Who is the wife of Barack Obama?

Answers™ Ask us anything GO

ENTERTAINMENT TECH LIFESTYLE FOOD HEALTH POLITICS MONEY SPORTS INTERVIEWS ALL SECTIONS

Categories

History History of the United States History, Politics & Society Obamacare (Affordable Care Act) US Presidents

© 2014 Answers About Careers Terms of Use Privacy Policy Consumer Choice IP Issues Disclaimer Write Articles Directory

Share f g t

Who is the wife of Barack Obama?

In BARACK OBAMA

Answer by Malia1998

His wife's name is Michelle LaVaughn Robinson Obama.

Social-QA

Quora

YAHOO! ANSWERS

ChaCha Questions

Answers™

# Multiple Intelligence in Modern QA



Who is the wife of Barack Obama?

Today's focus!

## Knowledge-QA



Freebase



ProBase

悟 SATORI

## Document-QA



## Social-QA

Quora

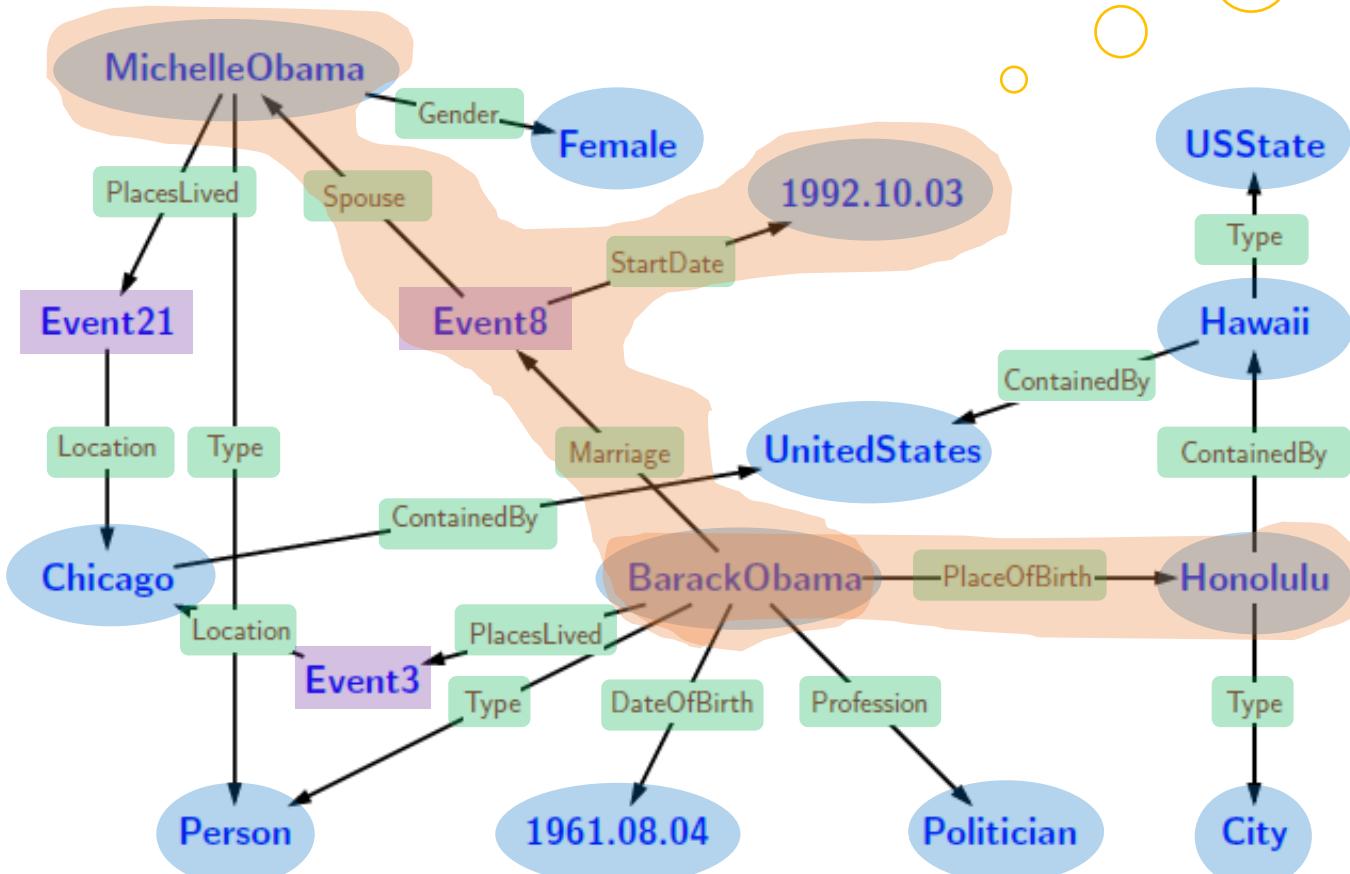
YAHOO! ANSWERS

ChaCha Questions

Answers™

# Knowledge-based QA

# Knowledge-Base (KB)

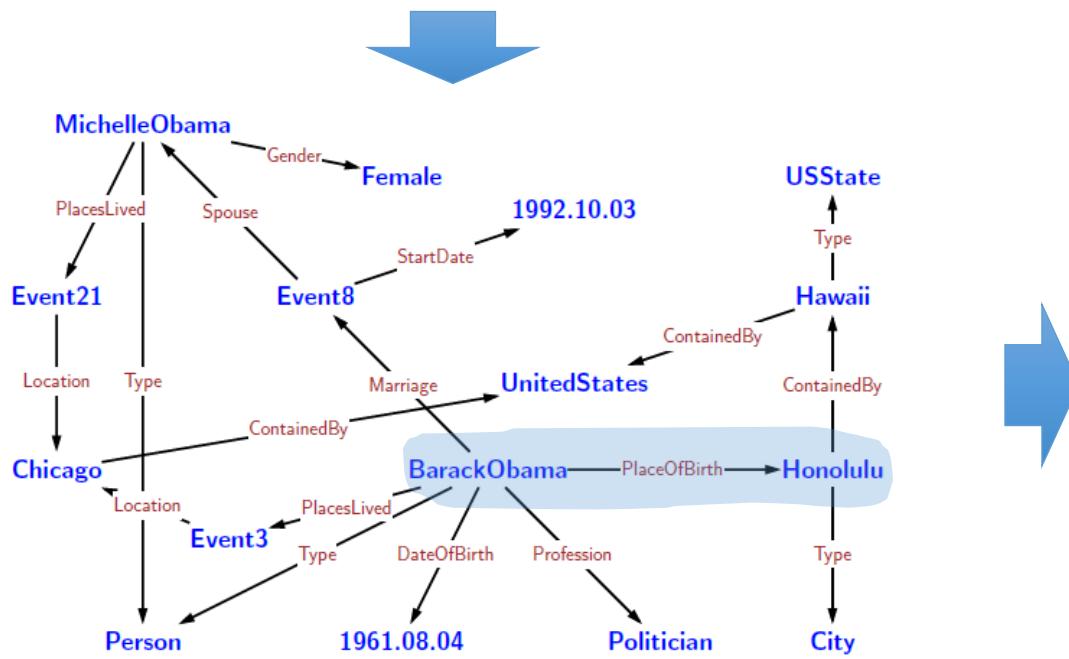


- **Entity**
- **Predicate**  
Relation between two connected entities
- **CVT (Compound Value Type)**  
Not a real-world entity, but is used to collect multiple fields of an event
- **Fact**  
Triple, which connects two entities  
Event, which connects multiple entities via a CVT node

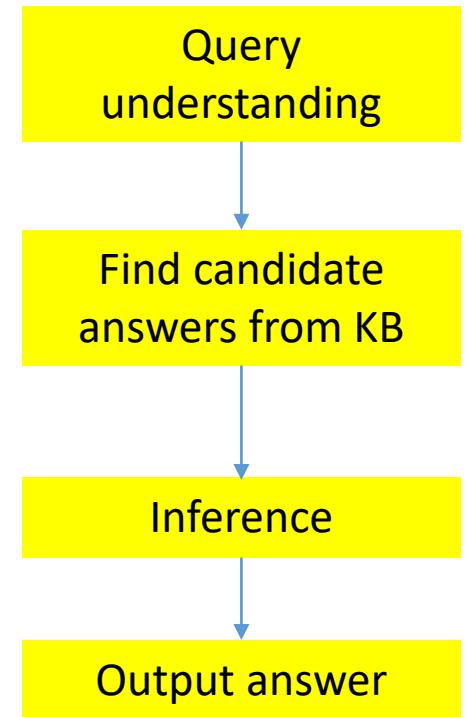
# Knowledge-Based Question Answering (KB-QA)

- Compute answers to natural language questions using existing knowledge bases (KB)

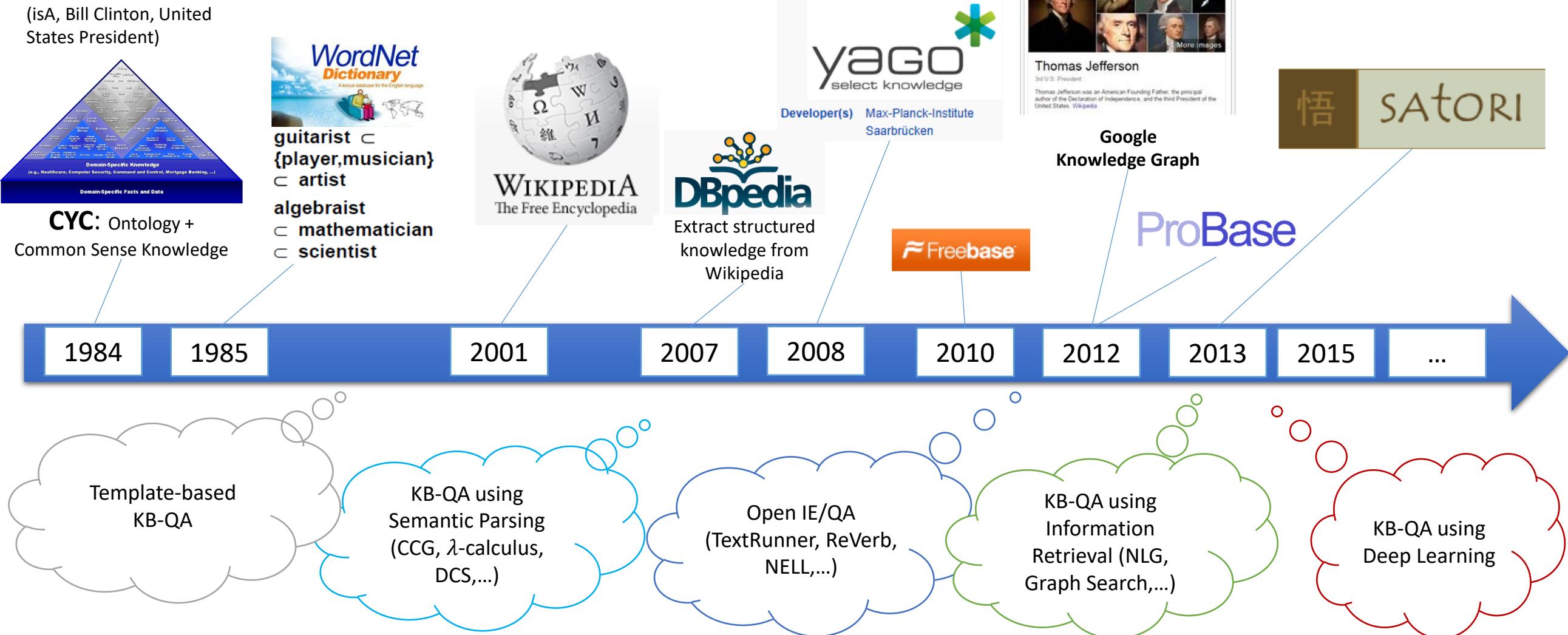
Question: *where was Barack Obama born ?*



Answer: **Honolulu**



# Evolution of KB-QA (1980~)



# Knowledge-based QA (KB-QA)

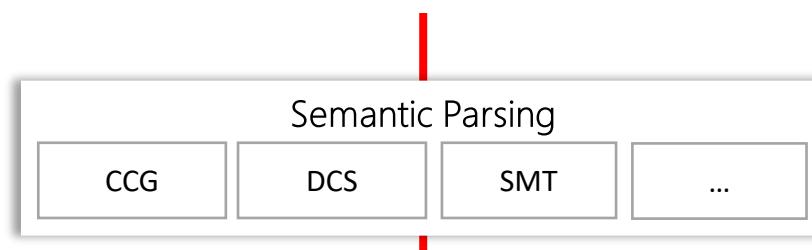
CCG: Combinatory Categorial Grammar

DCS: Dependency-based Compositional Semantics

SMT: Statistical Machine Translation

## Semantic Parsing-based KB-QA(SP-QA)

where was Barack Obama born ?



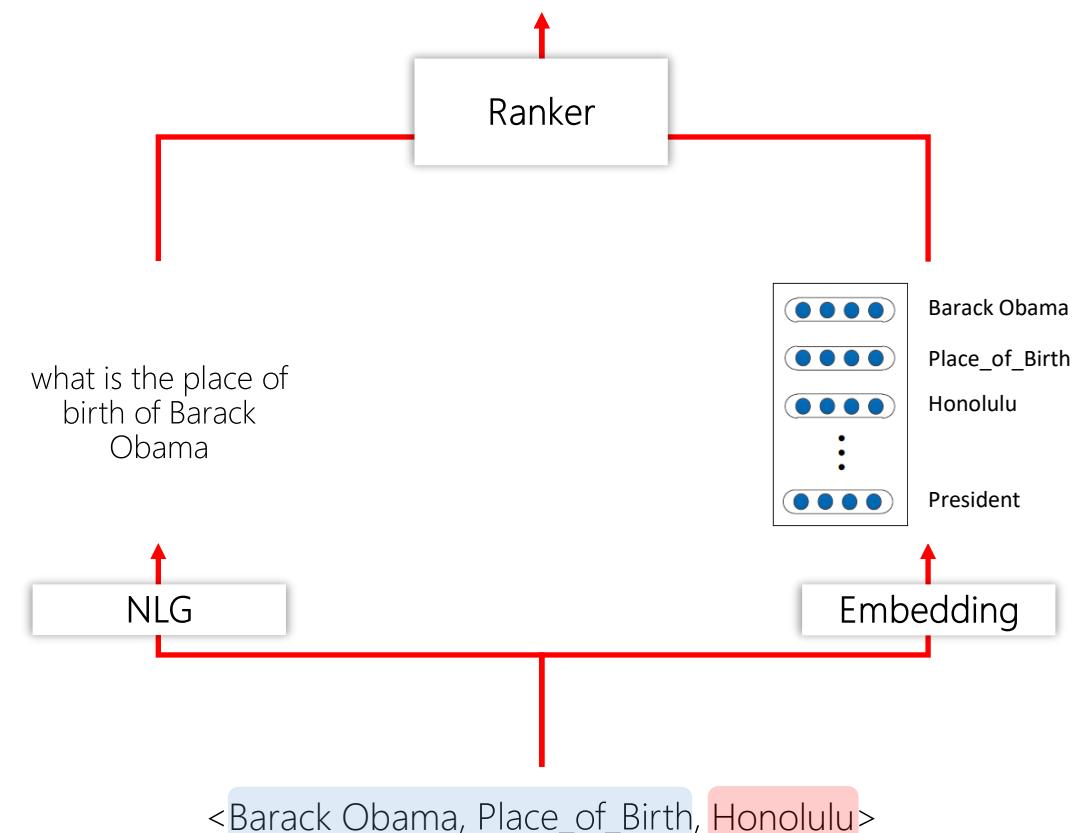
$\lambda x.\text{Place\_of\_Birth}(\text{Barack Obama}, x)$

KB Lookup

<Barack Obama, Place\_of\_Birth, Honolulu>

## Information Retrieval-based KB-QA(IR-QA)

where was Barack Obama born ?



what is the place of  
birth of Barack  
Obama

Barack Obama  
Place\_of\_Birth  
Honolulu  
President

NLG

Embedding

<Barack Obama, Place\_of\_Birth, Honolulu>

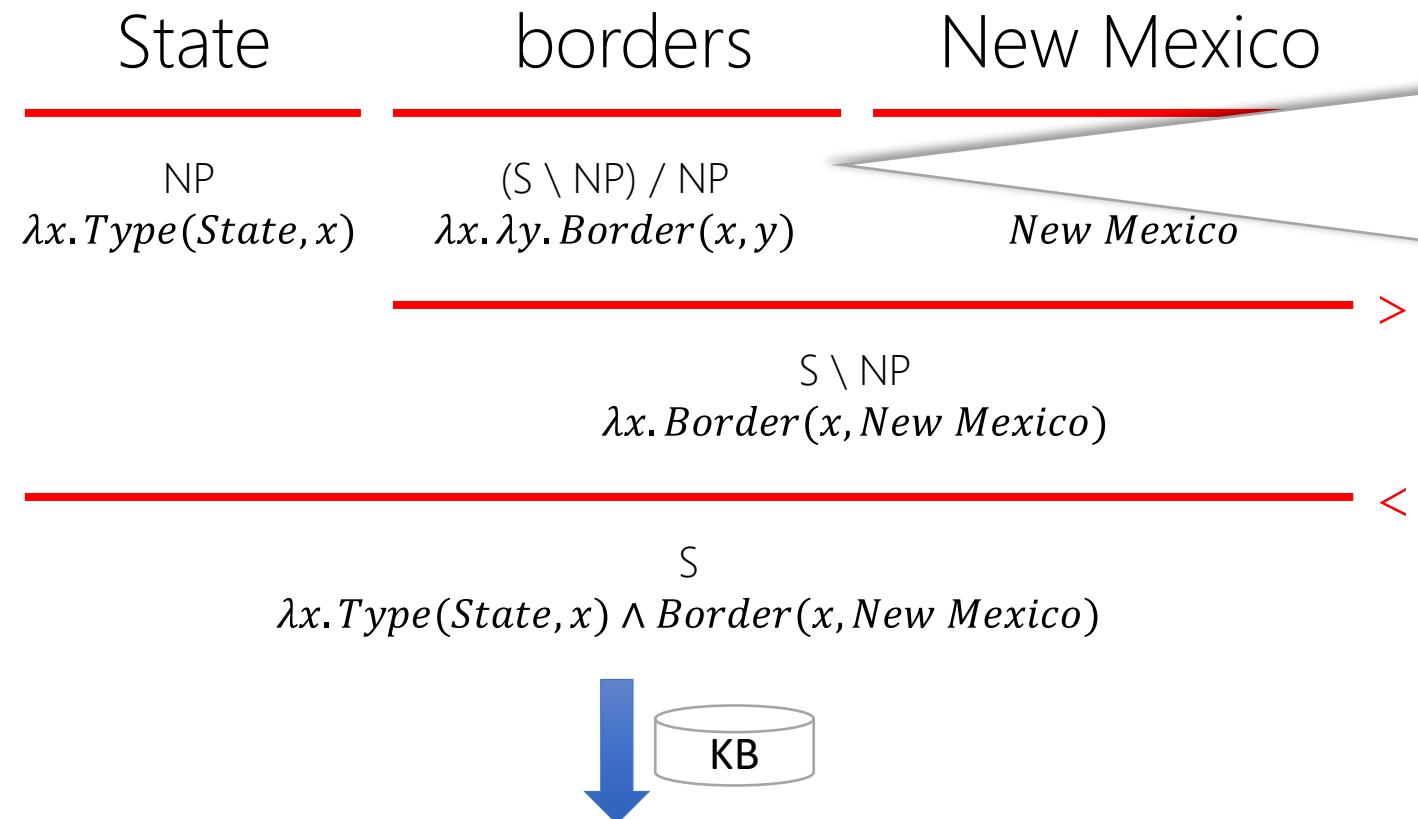
# KB-QA Methodology

- KB-QA by Semantic Parsing
- KB-QA by Information Retrieval
- KB-QA by Open Information Extraction

# KB-QA Methodology

- KB-QA by Semantic Parsing
- KB-QA by Information Retrieval
- KB-QA by Open Information Extraction

# Semantic Parsing with CCG

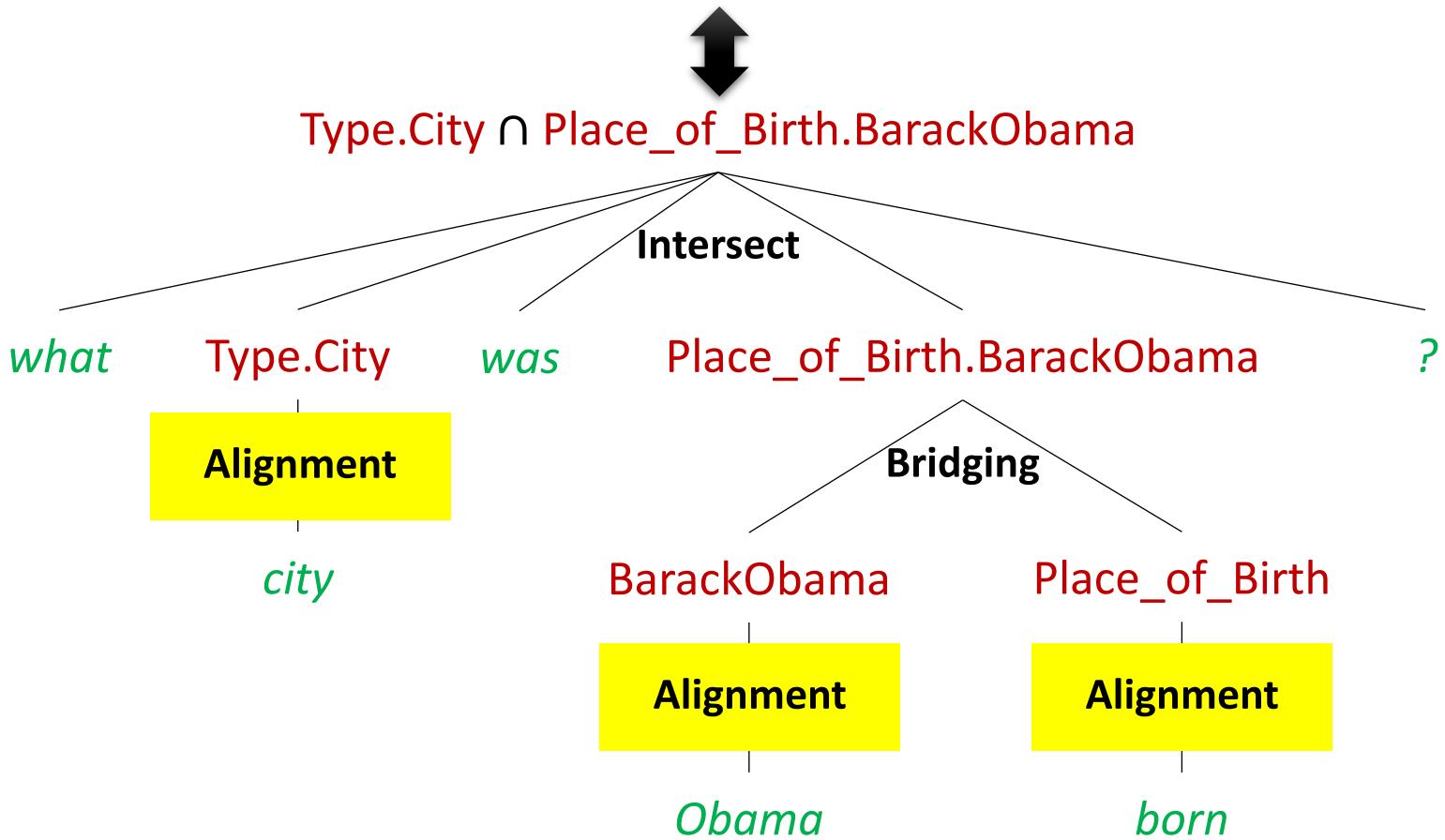


A CCG is defined by

- A lexicon
  - **N-gram:** border
  - **Syntactic category:**  $(S \setminus NP) \setminus NP$
  - **Logical form:**  $\lambda x. \lambda y. Border(x, y)$
- A set of combinators
- Challenge
  - Lexicon acquisition needs  $\langle NL, LF \rangle$  annotations

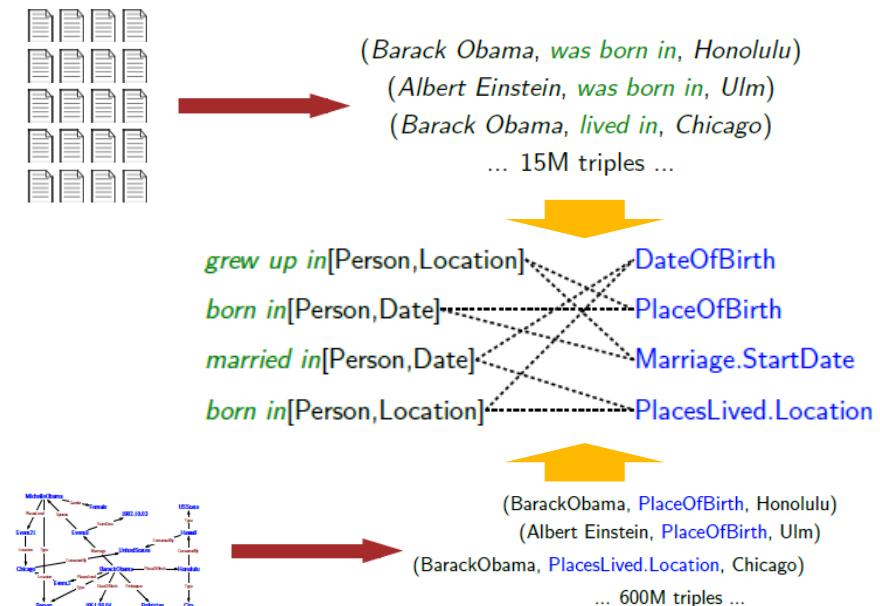
Arizona, Colorado, Oklahoma, Texas

# Semantic Parsing with DCS

$$\lambda x. Type(City, x) \wedge Place\_of\_Birth(Barack\ Obama, x)$$


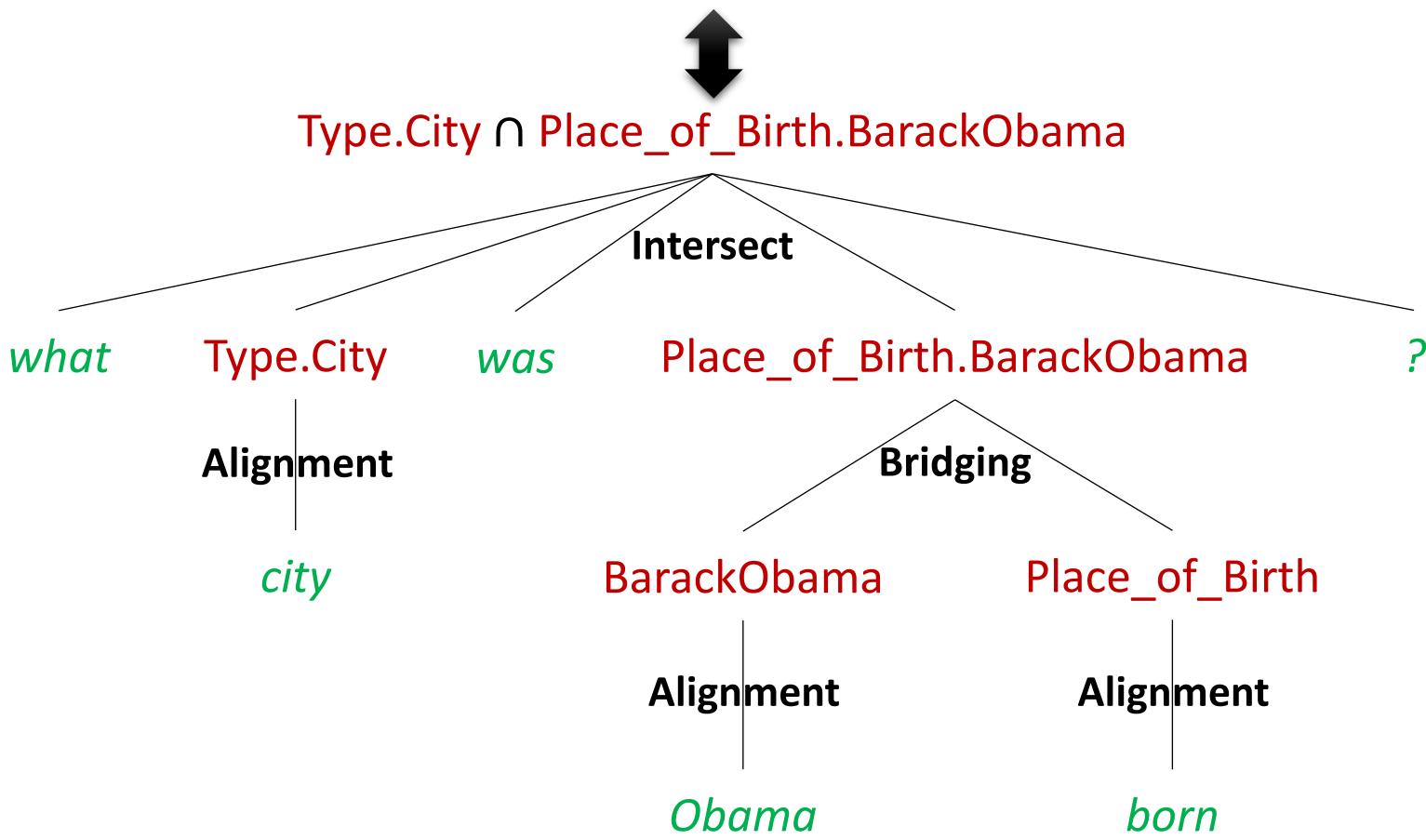
**Alignment**

- Map mentions to KB entities\predicates



# Semantic Parsing with DCS

$\lambda x. Type(City, x) \wedge Place\_of\_Birth(Barack\ Obama, x)$



## Training

- Use Q-A pairs as weak supervision

Category	Description
Alignment	Log of # entity pairs that occur with the phrase $r_1 ( \mathcal{F}(r_1) )$ Log of # entity pairs that occur with the logical predicate $r_2 ( \mathcal{F}(r_2) )$ Log of # entity pairs that occur with both $r_1$ and $r_2 ( \mathcal{F}(r_1) \cap \mathcal{F}(r_2) )$ Whether $r_2$ is the best match for $r_1 (r_2 = \arg \max_r  \mathcal{F}(r_1) \cap \mathcal{F}(r) )$
Lexicalized	Conjunction of phrase $w$ and predicate $z$
Text similarity	Phrase $r_1$ is equal/prefix/suffix of $s_2$ Phrase overlap of $r_1$ and $s_2$
Bridging	Log of # entity pairs that occur with bridging predicate $b ( \mathcal{F}(b) )$ Kind of bridging (# unaries involved) The binary $b$ injected
Composition	# of intersect/join/bridging operations POS tags in join/bridging and skipped words Size of denotation of logical form

$$p(d | x, \theta) = \frac{\exp(\phi(x, d) \cdot \theta)}{\sum_{d' \in \mathcal{D}(x)} \exp(\phi(x, d') \cdot \theta)}$$

# Semantic Parsing with SMT

- Perform semantic parsing & answer generation in a synchronous manner, based on the machine translation framework
- Given an NL question  $Q$ , generate  $\mathcal{H}(Q)$ , which encodes a set of semantic representation-answer pairs  $\{\langle \mathcal{D}_1, \mathcal{A}_1 \rangle, \dots, \langle \mathcal{D}_N, \mathcal{A}_N \rangle\}$

$$\langle \hat{\mathcal{D}}, \hat{\mathcal{A}} \rangle = \operatorname{argmax}_{\langle \mathcal{D}, \mathcal{A} \rangle \in \mathcal{H}(Q)} \frac{\exp\{\sum_j \lambda_j \cdot h_j(\mathcal{D}, \mathcal{A}, Q)\}}{\sum_{\langle \mathcal{D}', \mathcal{A}' \rangle \in \mathcal{H}(Q)} \exp\{\sum_j \lambda_j \cdot h_j(\mathcal{D}', \mathcal{A}', Q)\}}$$

Parameter Optimization by Q-A Pairs

Feature Design

Search Space Generation by Semantic Parsing

# Semantic Parsing with SMT

$\lambda x \lambda y. Directed\_By(y, x) \wedge Starred\_In(Tom\ Hanks, y)$

$\langle [Film], Film.Film.Director, [Person] \rangle$

director of [Film]

$\lambda x. Starred\_In(Tom\ Hanks, x)$

$\langle Tom\ Hanks, Film.Actor.Film, [Film] \rangle$

the movie starred by Tom Hanks

director of the movie starred by Tom Hanks

$\lambda x \lambda y. Directed\_By(y, x) \wedge Starred\_In(Tom\ Hanks, y)$

$\langle [Film], Film.Film.Director, [Person] \rangle$

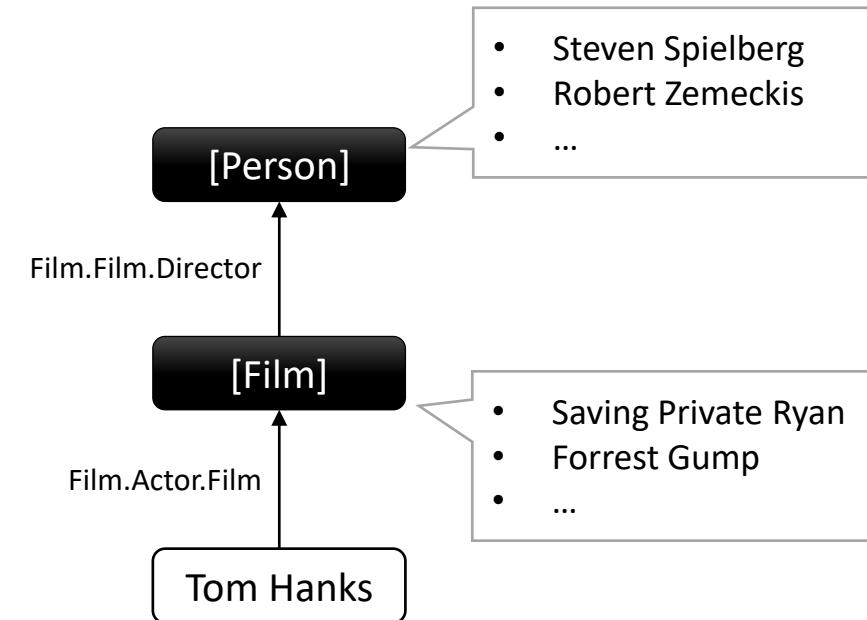
director of [Film]

$\lambda x. Starred\_In(Tom\ Hanks, x)$

$\langle Tom\ Hanks, Film.Actor.Film, [Film] \rangle$

the movie starred by Tom Hanks

director of the movie starred by Tom Hanks



Subject Entity	Predicate	Object Entity
Tom Hanks	Film.Actor.Film	Saving Private Ryan
Tom Hanks	Film.Actor.Film	Forrest Gump
...	...	...

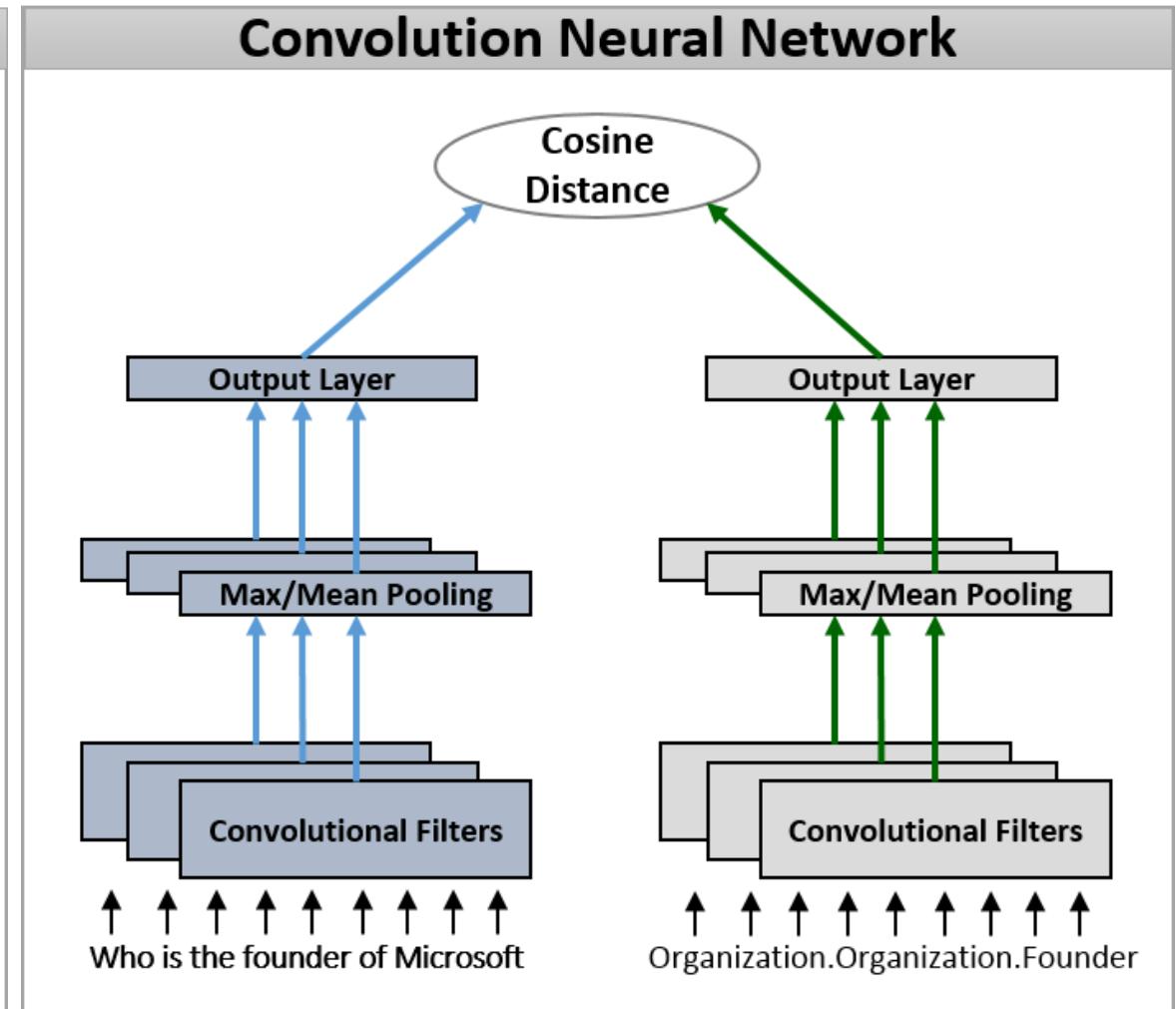
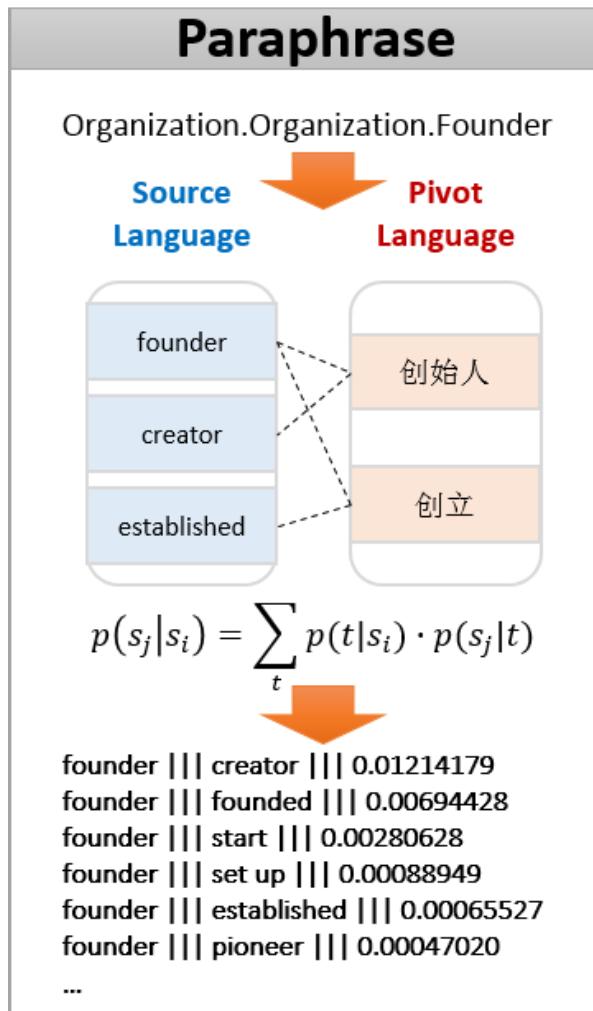
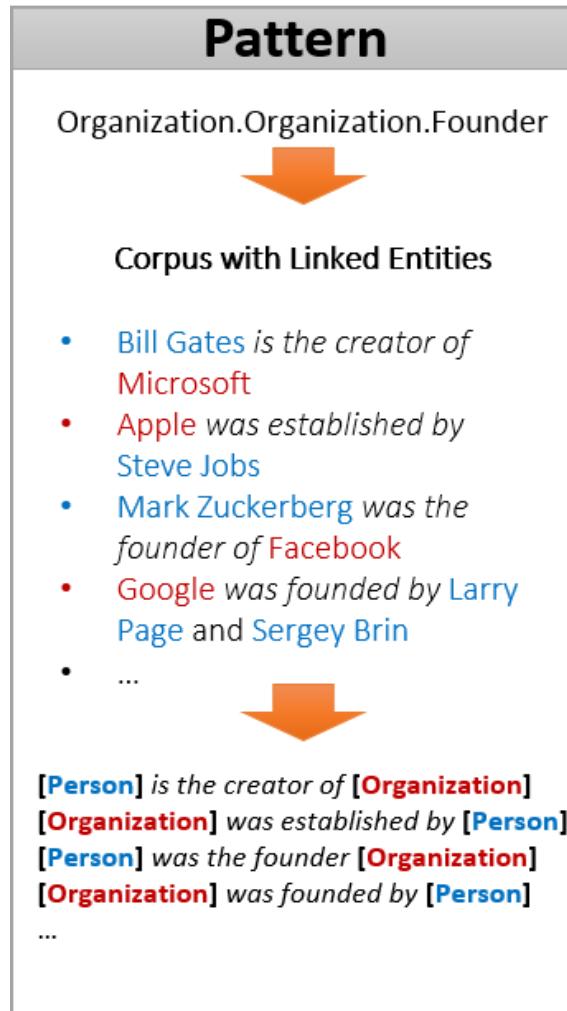
Subject Entity	Predicate	Object Entity
Saving Private Ryan	Film.Film.Director	Steven Spielberg
Forrest Gump	Film.Film.Director	Robert Zemeckis
...	...	...

Microsoft was founded by whom



<Microsoft, Organization.Organization.Founder, ?>

# Predicate Grounding

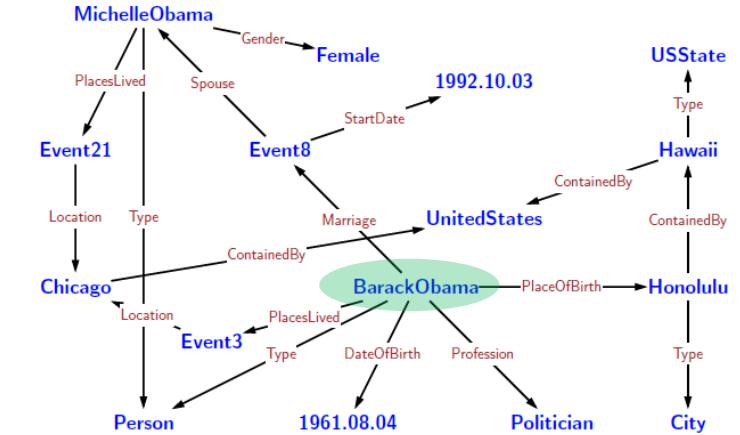


# KB-QA Methodology

- KB-QA by Semantic Parsing
- KB-QA by Information Retrieval
- KB-QA by Open Information Extraction

# Rank by the Distance of Question and Surface Form of Inference Chain

*which city was Obama born ?*



Template	Example
$p.e$	Directed.TopGun
$p_1.p_2.e$	Employment.EmployerOf.SteveBalmer
$p.(p_1.e_1 \sqcap p_2.e_2)$	Character.(Actor.BradPitt \sqcap Film.Troy)
Type. $t \sqcap z$	Type.Composer \sqcap SpeakerOf.French
count( $z$ )	count(BoatDesigner.NatHerreshoff)

## DataOfBirth.BarackObama

• •

Type.City ∩ PlaceOfBirth.BarackObama

## Grow logical forms around entities

# Rank by the Distance of Question and Surface Form of Inference Chain

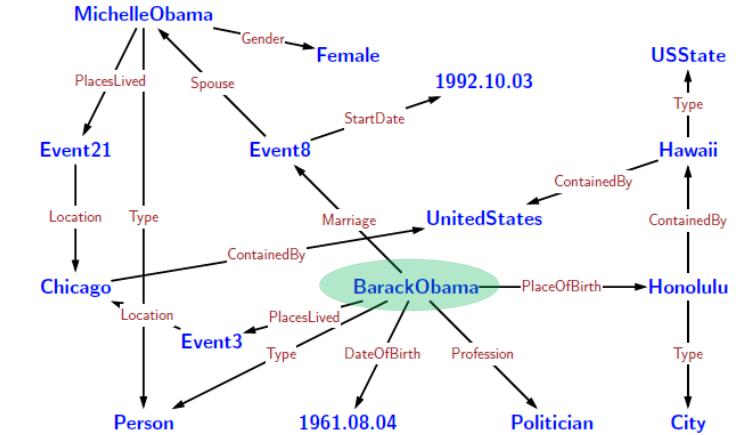
*which city was Obama born ?*

When is the date of birth of Barack Obama ?

... What city is the place of birth of Barack Obama ?

## DataOfBirth.BarackObama

... Type.City  $\cap$  PlaceOfBirth.BarackObama



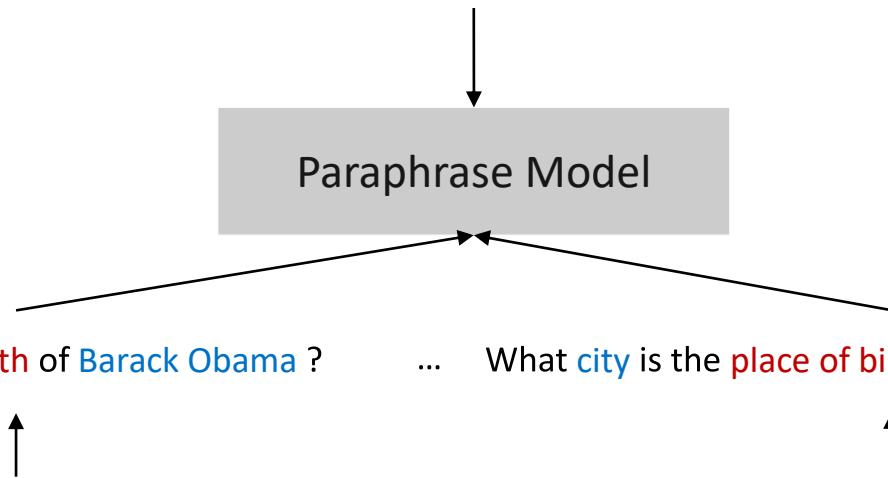
- $d(t)$ : answer type
  - $d(p)$ : predicate
  - $d(e)$ : question entity

	$d(p)$ Categ.	Rule
$p.e$	NP	WH $d(t)$ has $d(e)$ as NP ?
	VP	WH $d(t)$ (AUX) VP $d(e)$ ?
	PP	WH $d(t)$ PP $d(e)$ ?
	NP VP	WH $d(t)$ VP the NP $d(e)$ ?

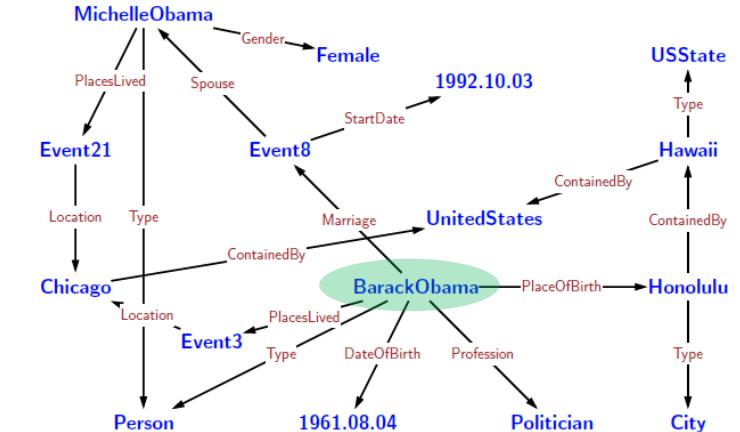
## Generate questions based on templates

# Rank by the Distance of Question and Surface Form of Inference Chain

*which city was Obama born ?*



## Paraphrase Model



## Association Model

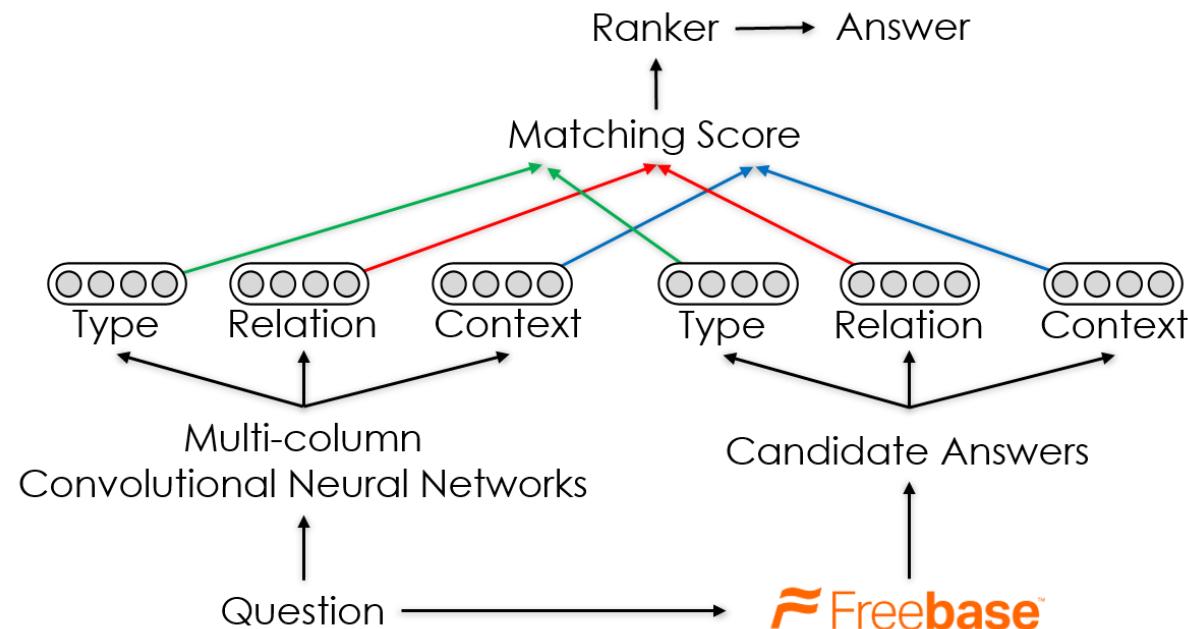
- Learn associated phrase pairs from 18M word-aligned question pairs from WikiAnswers (Fader et al., 2013)

## Vector Space Model

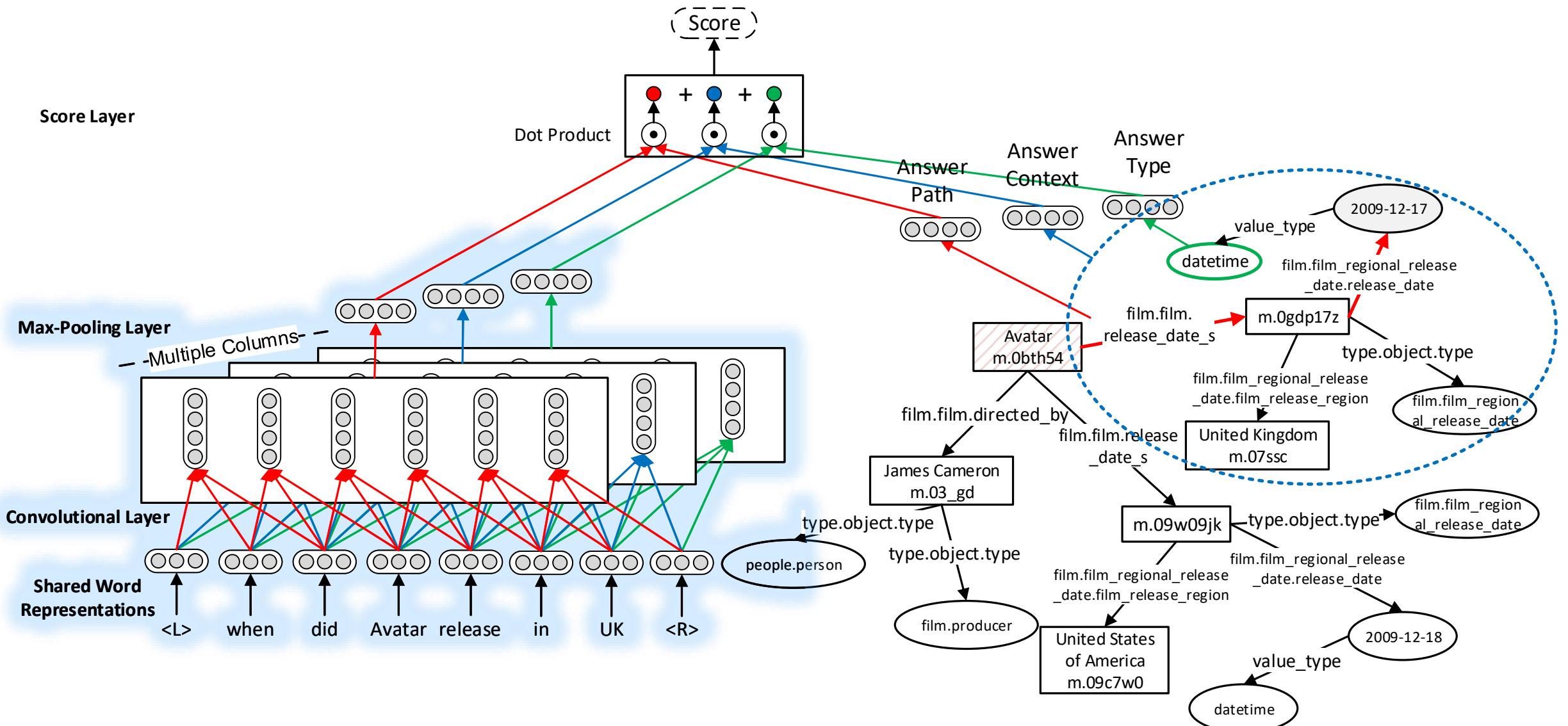
- Learn word embedding vectors from Wikipedia text using the CBOW model (Mikolov et al., 2013)

# Rank by Neural Network

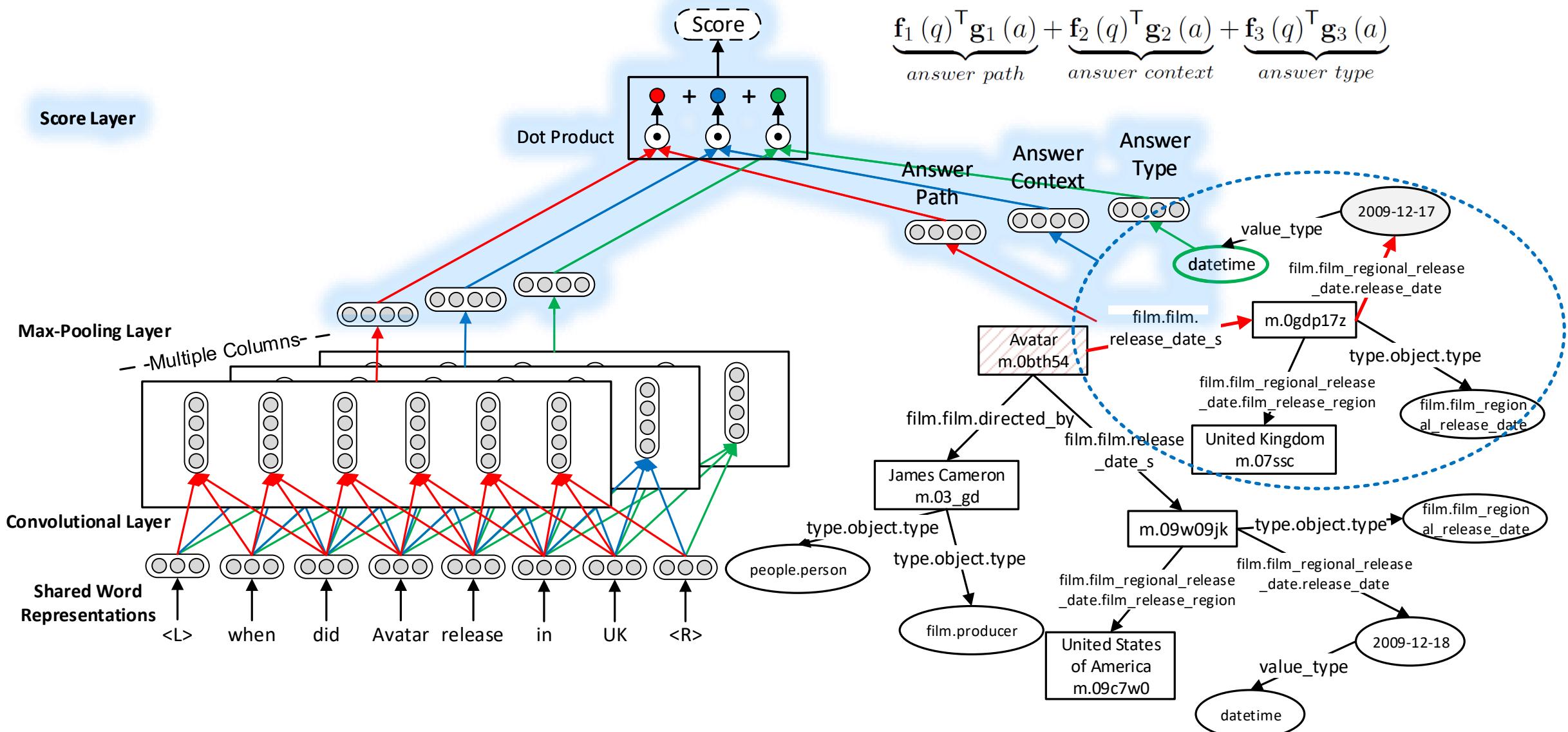
- Question Answering with Subgraph Embeddings (Bordes et al., 2014, EMNLP)
- Question Answering over Freebase with Multi-Column Convolutional Neural Networks (Dong et al., 2015, ACL)
- Question answering -> Constraint matching
  - Answer type, answer path (relation), answer context
- Question understanding with convolutional neural networks



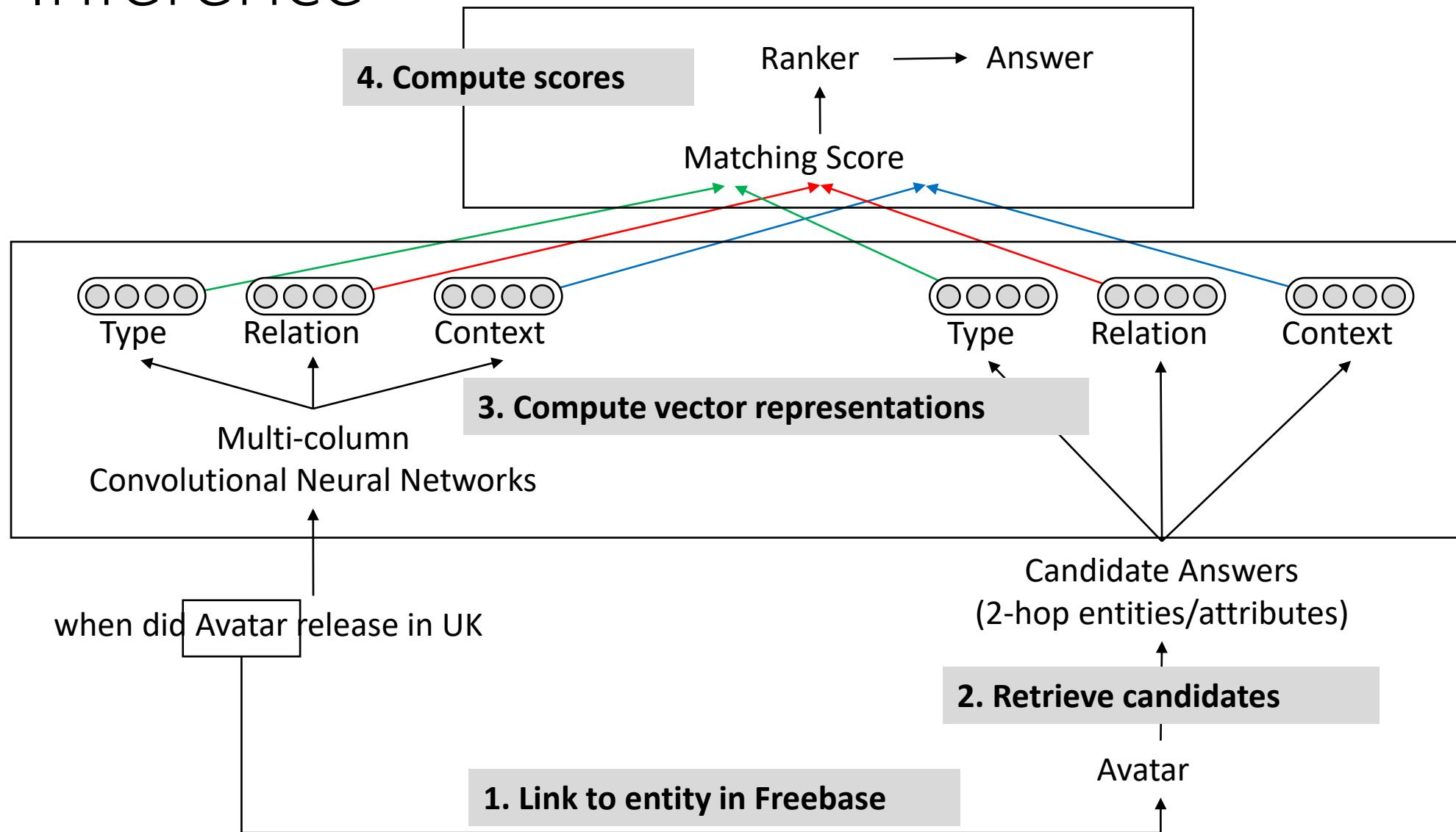
# Multi-Column CNN Model



# Multi-Column CNN Model



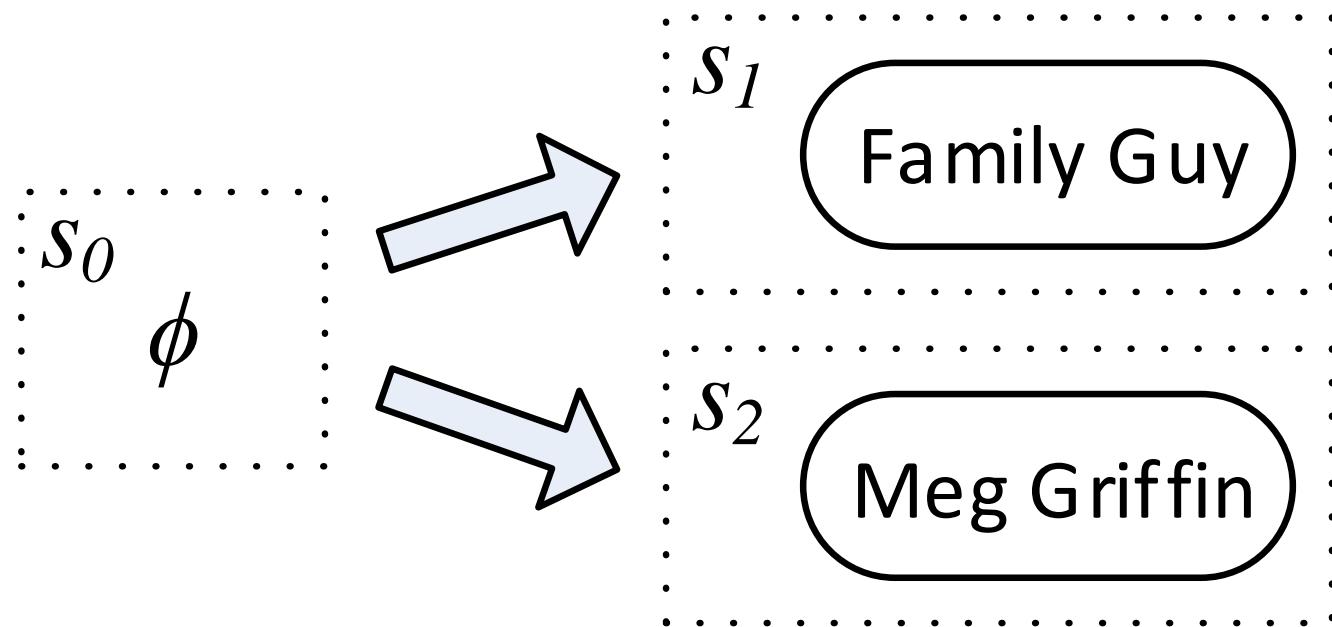
# Inference



# Query Graph Generation and Multi-Feature Ranking

## (1) Link Topic Entity

Who first voiced Meg on  
Family Guy?

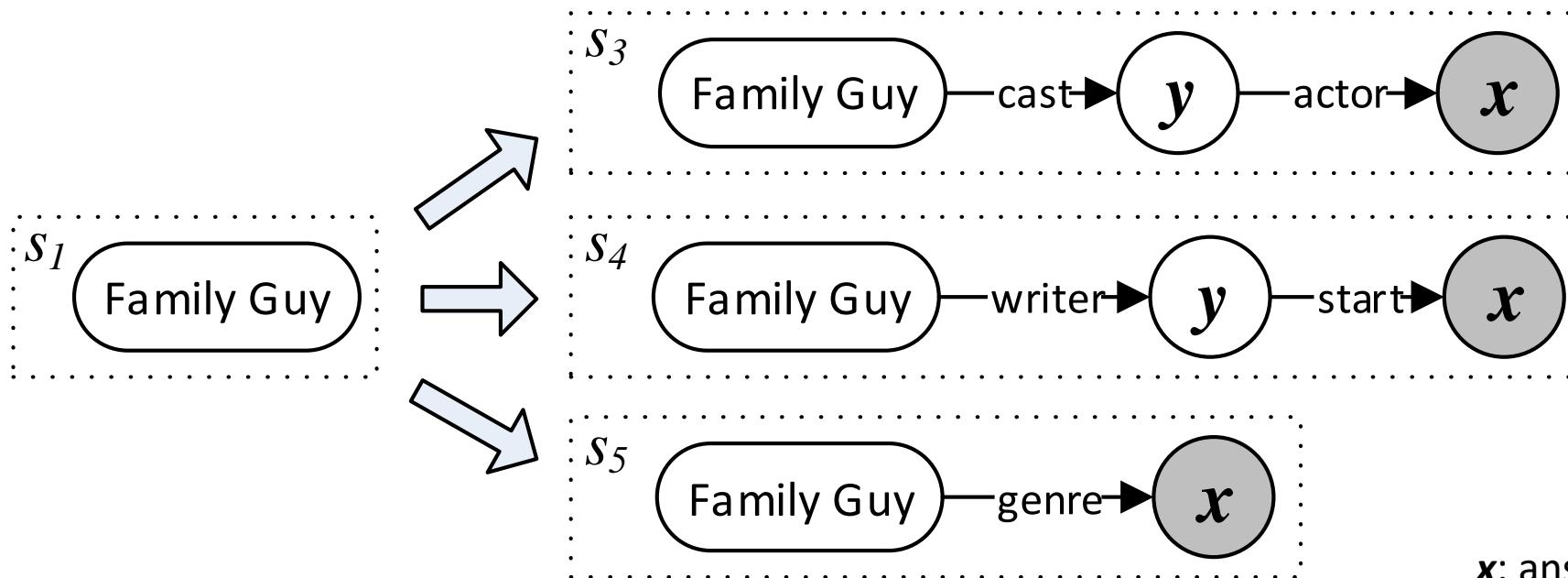


PPT available at  
<http://research.microsoft.com/apps/pubs/default.aspx?id=244749>

# Query Graph Generation and Multi-Feature Ranking

## (2) Identify Core Inferential Chain

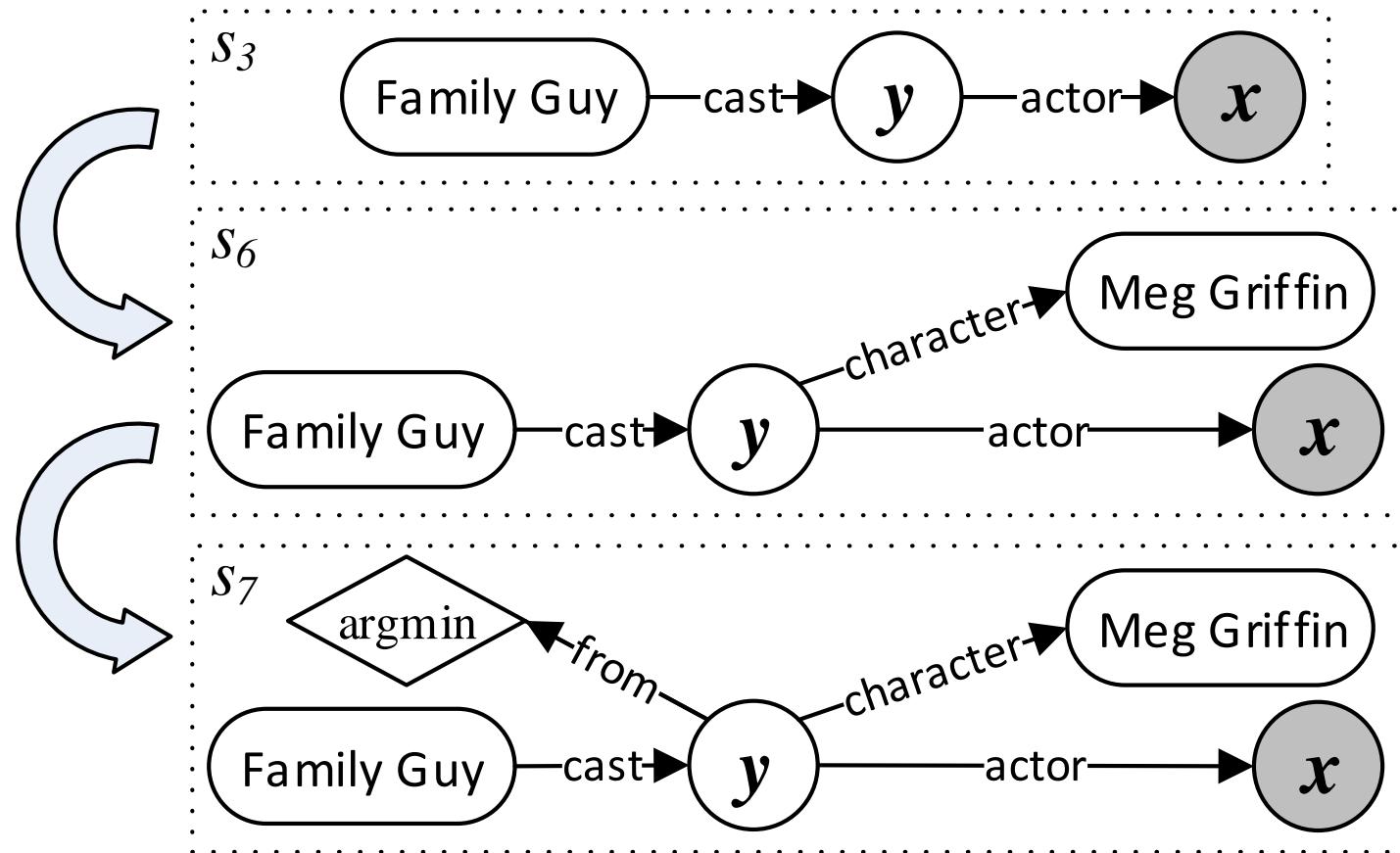
Who first voiced Meg on Family Guy?



$x$ : answer entity  
 $y$ : CVT (compound value type), which is not a real-world entity, but is used to collect multiple fields of an event

# Query Graph Generation and Multi-Feature Ranking

## (3) Augment Constraints

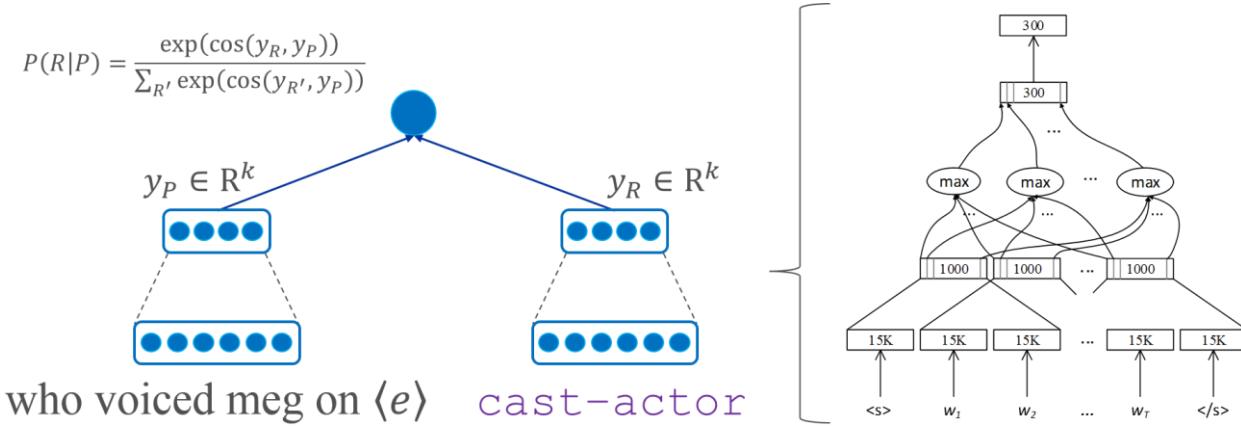


Who first voiced **Meg** on  
**Family Guy?**

$x$ : answer entity  
 $y$ : CVT (compound value type), which  
is not a real-world entity, but is used  
to collect multiple fields of an event

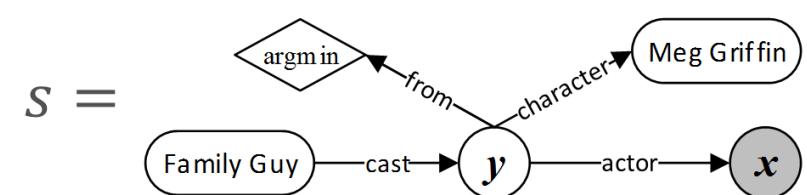
# Query Graph Generation and Multi-Feature Ranking

- Features of the *linear* reward function
  - 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

$q$  = Who first voiced Meg on Family Guy?



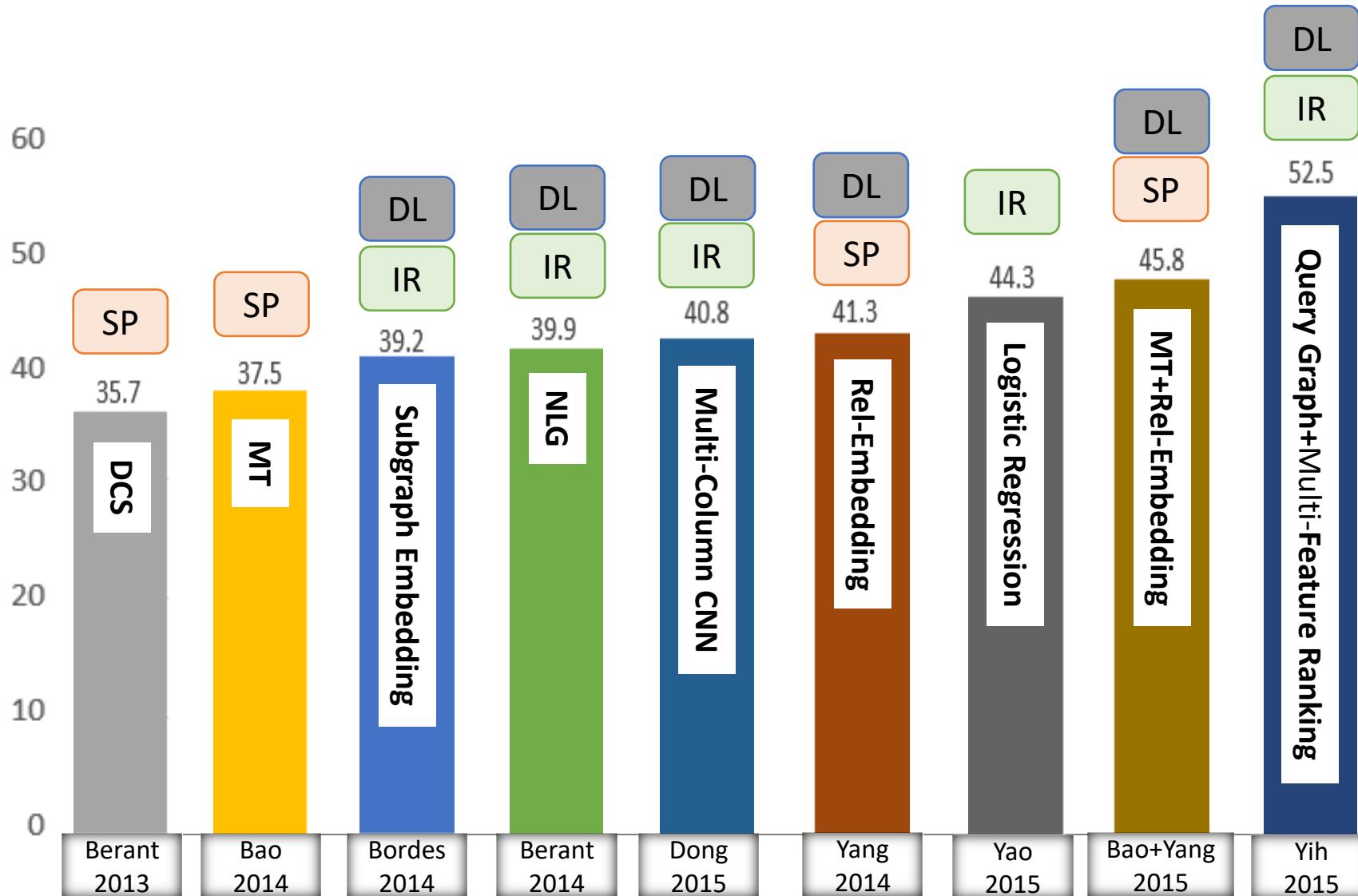
# Benchmark

- WebQuestions
  - Data statistic
    - 5,810 Q-A pairs (English) (questions are sampled from Google query log)
    - Most of them are one-hop factoid questions
  - Citation
    - Jonathan Berant, Andrew Chou, Roy Frostig, Percy Liang, Semantic Parsing on Freebase from Question-Answer Pairs, EMNLP, 2013
  - Link
    - <http://nlp.stanford.edu/software/sempre/>

# Evaluation Results on WebQuestions



# Evaluation Results on WebQuestions

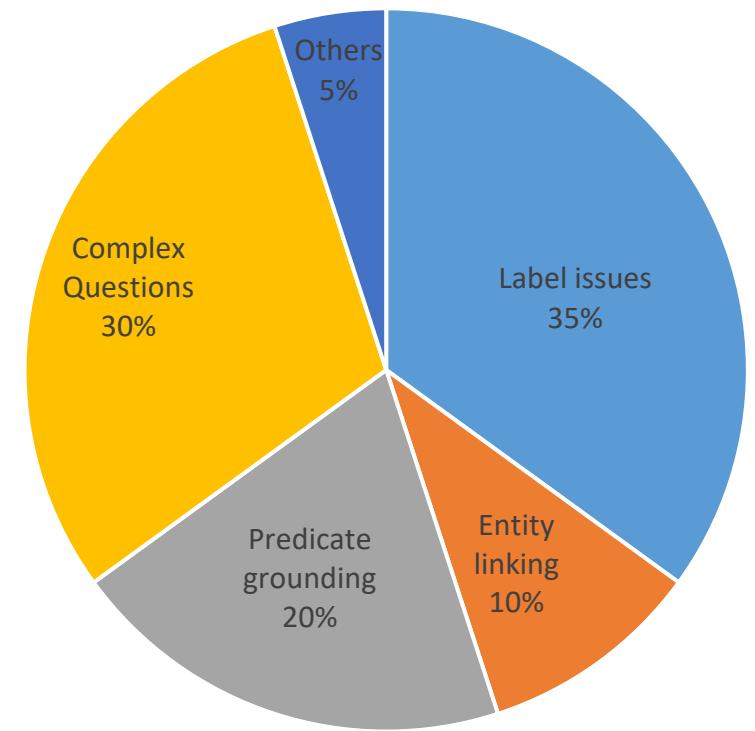


## Key Findings

- IR-based methods becomes popular in recent 2 years
- Deep Learning (DL) brings further gains, upon baselines using surface form features
- IR+DL achieves current best result
- However, there is still big room to improve

# Top Mistake Reasons

- Label issues (35%)
  - *what songs did Bob Dylan sing?*
  - *what kind of currency does Cuba use?*
- Entity linking (10%)
  - *what country did germany invade first in ww1? //World War I*
  - *what did ben franklin invent? //benjamin franklin*
- Predicate grounding (20%)
  - *what vegetables can i plant in november in southern california?*
  - *when did MT st Helens first erupt?*
- Complex questions (multi-hop, multi-constraints, etc.) (30%)
  - *what did james k polk do before he was president?*
  - *when was the last time the new England patriots won the super bowl?*
- Others (5%)



# KB-QA Methodology

- KB-QA by Semantic Parsing
- KB-QA by Information Retrieval
- KB-QA by Open Information Extraction

# Open-KB



Massive Web Pages



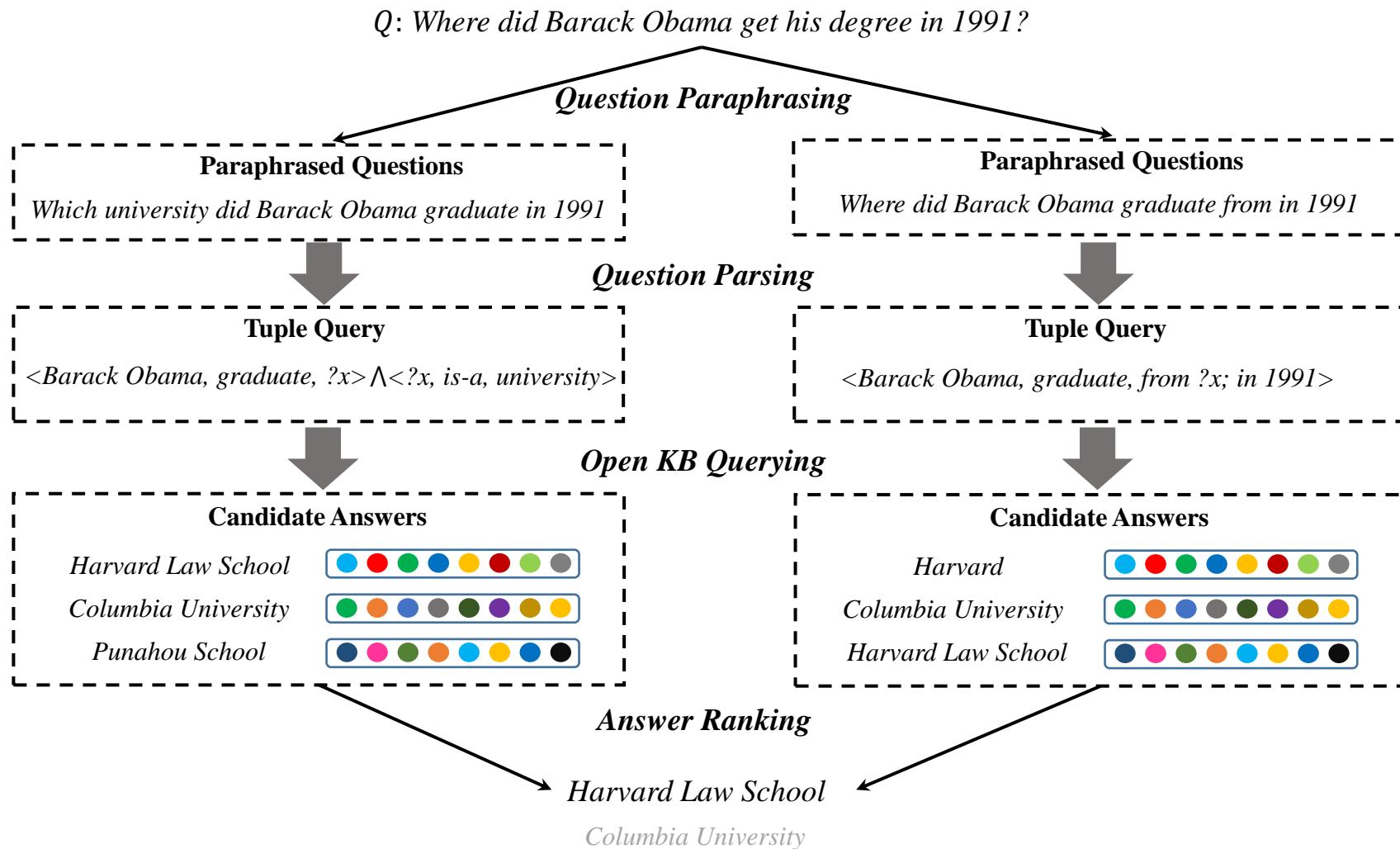
Open Information Extraction

## Open KB by Open IE

- Task  
Extract (semi-)structured information from unstructured texts
- Methods
  - POS pattern-based
    - ReVerb (Fader et al., EMNLP, 2011)
  - Dependency pattern-based
    - WOE (Wu and Weld, ACL, 2010)
    - OLLIE (Mausam et al., EMNLP, 2012)
    - ClausIE (Corro and Gemulla, WWW, 2013)

Subject	Relation	Arguments	Freq.	Conf.
James K. Polk	was	a governor; before he was president	2	0.87
the currency of Spain	was	the Peseta; before 2002	3	0.95
Peseta	was replaced	by Euro; as official tender of Spain; in 2002	3	0.81
Barack Obama	graduated	from Harvard Law School; in 1979 and 1991	4	0.77
Obama	graduated	from Harvard Law School; in 1991	5	0.93
Barack Obama	attended	Harvard university; from 1988	3	0.90
...	...	...	...	...

# Answering Questions with Complex Semantic Constraints on Open Knowledge Bases



ComplexQuestions	
Yin et al., 2015	39.3%
Berant et al., 2014	9.7%
$acc = \frac{\text{number of correctly answered questions}}{\text{total number of questions}}$	

ComplexQuestions Dataset	
• Data statistic	<ul style="list-style-type: none"><li>• 300 English Q-A pairs</li><li>• 80 questions from WebQuestions</li><li>• 220 questions from Google query log</li></ul>
• Citation	<ul style="list-style-type: none"><li>• Pengcheng Yin, Nan Duan, Ben Kao, Junwei Bao, Ming Zhou, <i>Answering Questions with Complex Semantic Constraints on Open Knowledge Bases</i>, CIKM, 2015</li></ul>
• Link	<ul style="list-style-type: none"><li>• To be released</li></ul>
• An example	<ul style="list-style-type: none"><li>• Q: what did Germany lost after the Treaty of Versailles?</li><li>• A: Northern Schleswig; Alsace Lorraine</li></ul>

# Question Paraphrasing

- Rewrite input question into multiple similar questions

*Where did [Barack Obama] get his degree in 1991 ?*



## Paraphrase Template

*Where did [None Phrase] get his degree*  $\mapsto$  *Where did [None Phrase] graduate from*

*Where did [None Phrase] get his degree*  $\mapsto$  *Which university did [None Phrase] graduate*

*Where did [None Phrase] get his degree*  $\mapsto$  ...



*Where did Barack Obama graduate from in 1991?*

*Which university did Barack Obama graduate in 1991?*

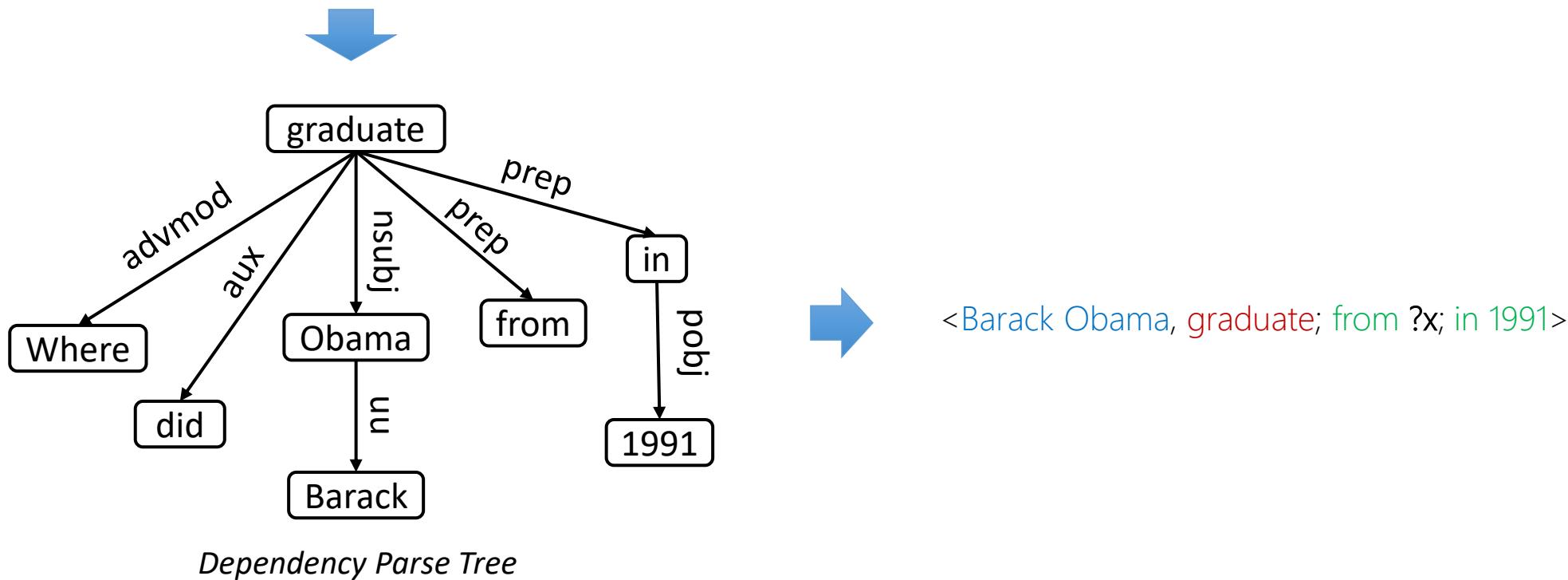
...

5M paraphrase templates extracted from 23M  
WikiAnswers question clusters (Fader et al., 2014)

# Question Parsing

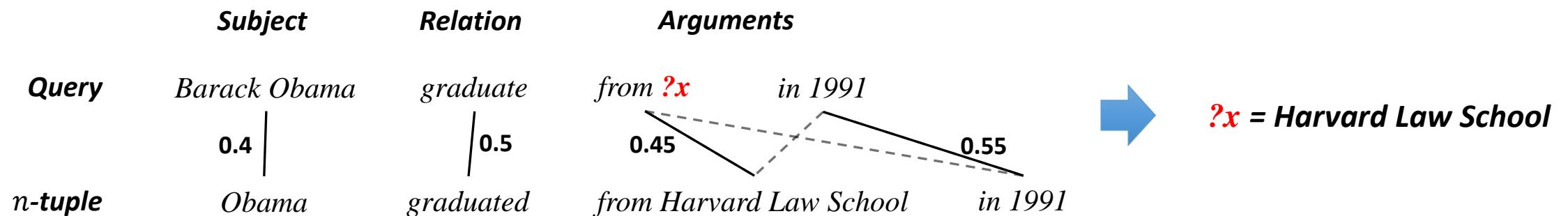
- Parse question into tuple query by traveling its dep-tree

*Where did **Barack Obama** graduate from **in 1991**?*



# Open KB Querying

- Given an  $n$ -tuple query, query open KB to extract answers



Define pairwise similarity between fields:

$$\begin{aligned} \text{similarity}(TQ.i, r.j) &= \alpha \cdot \text{text\_similarity}(TQ.i, r.j) \\ &\quad + (1 - \alpha) \cdot \text{pattern\_similarity}(TQ.i, r.j) \end{aligned}$$

Solve the optimization problem:

$$\begin{aligned} \max : & \sum_i \sum_j [x_{ij}] \text{similarity}(TQ.i, r.j) \\ \text{subject to: } & (x_{11} = 1), (x_{22} = 1), (x_{ij} \in \{0, 1\}) \\ & (0 \leq \sum_i x_{ij} \leq 1), (\sum_j x_{ij} = 1). \end{aligned}$$

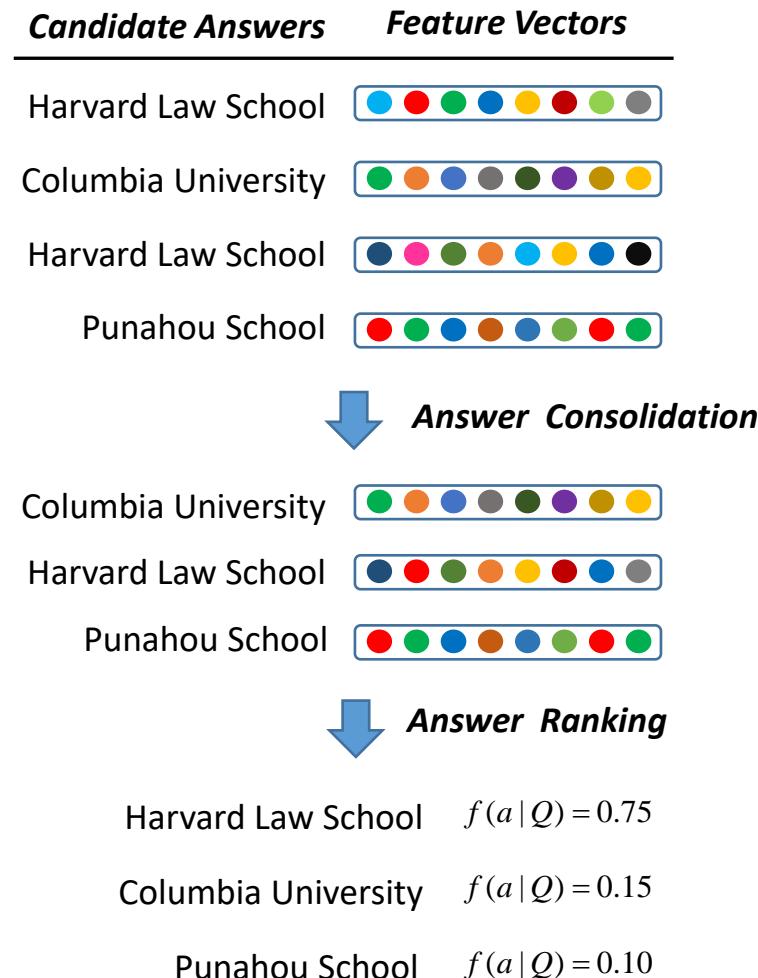
# Answer Ranking

- Answer Consolidation
  - Merge the feature vectors of the answers with the same surface text form
- Ranking Model
  - Use 20K+ features to measure a candidate  $a$  is one answer of  $Q$

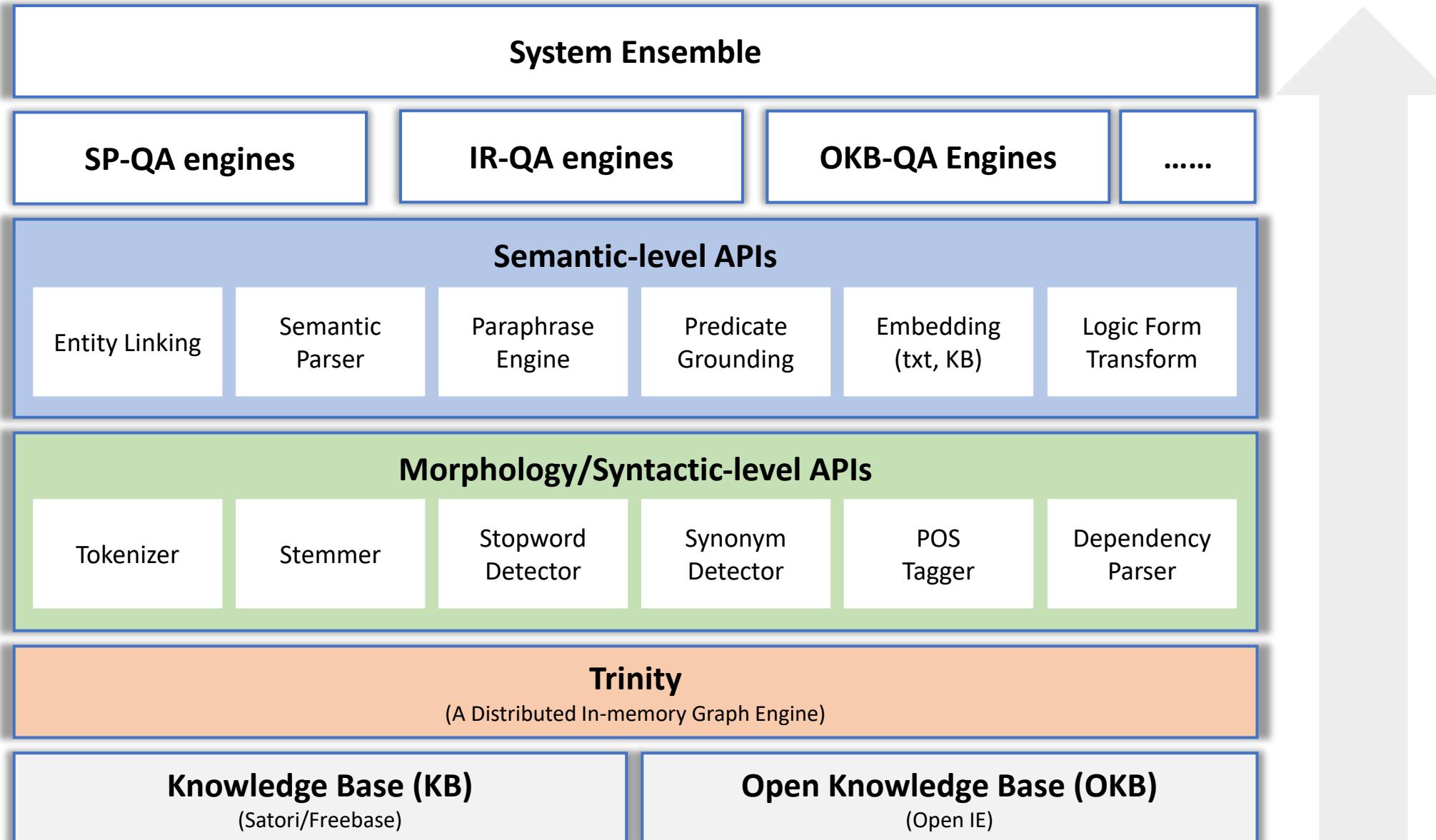
$$p(a|Q) = \frac{\exp\{\sum_{k=1}^M \lambda_k \cdot f_k(a)\}}{\sum_{a' \in \mathcal{A}} \exp\{\sum_{k=1}^M \lambda_k \cdot f_k(a')\}}$$

- Maximize the log-likelihood on a set of question-answer pairs

$$\mathcal{L}(\mathcal{D}; \lambda) = \sum_{t=1}^N \log p(a_t|Q_t; \lambda)$$

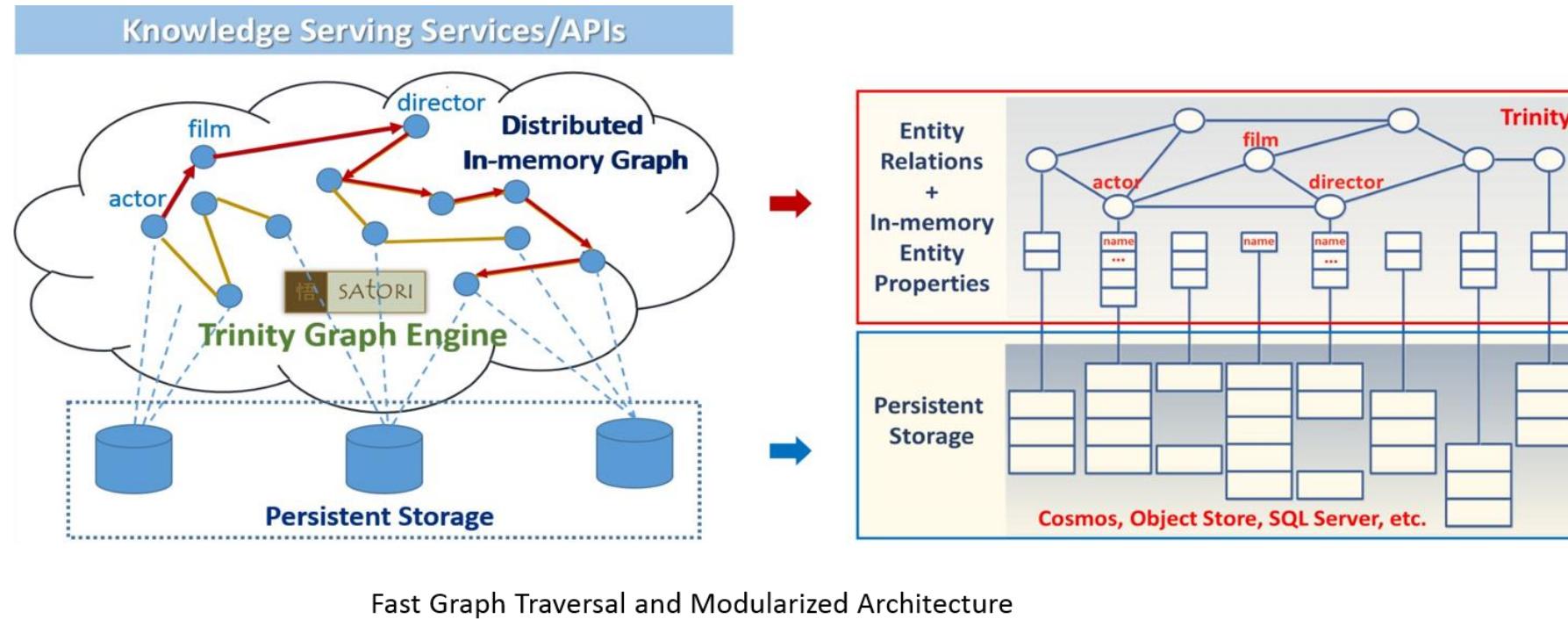


# MSRA KB-QA Architecture



# Trinity: Distributed In-memory Graph Engine

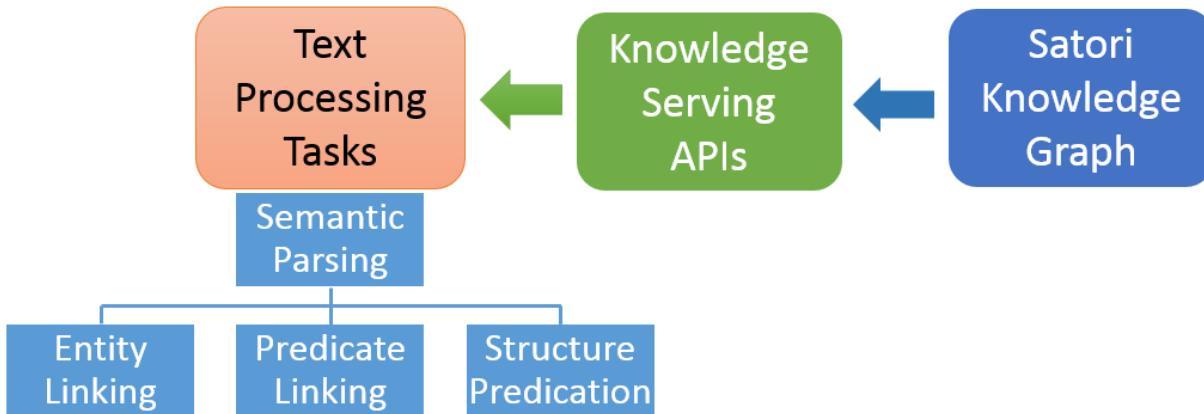
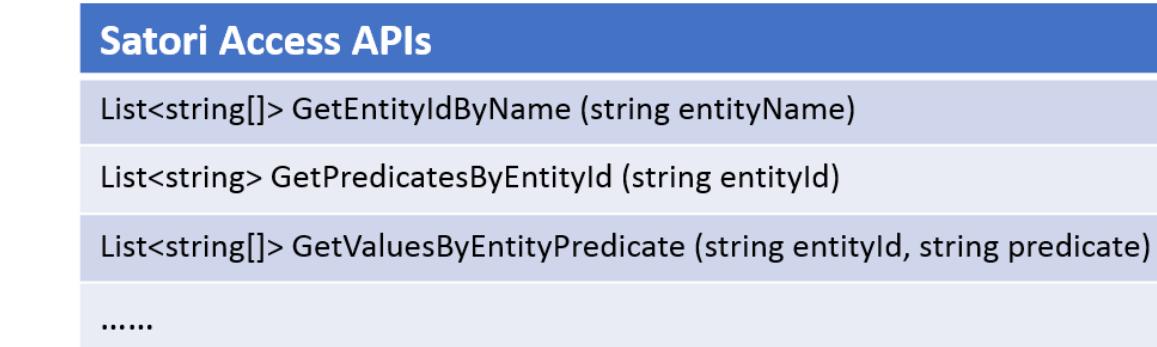
<http://research.microsoft.com/graphengine>



## Deployment

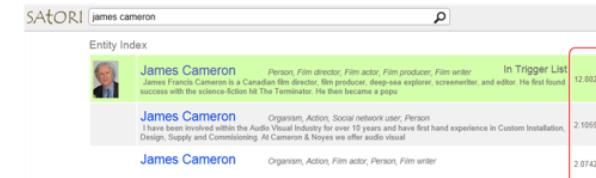
- **Fast building pipeline:** 6 hours to build from a cosmos Satori snapshot
- **Economical:** 16 machines for the whole Satori

# Trinity Serving KB Access



- `List<string[]> GetEntityIdByName (string entityName)`

- Input
  - An entity name
- Output
  - A list of <entity id, entity ranking score> pairs



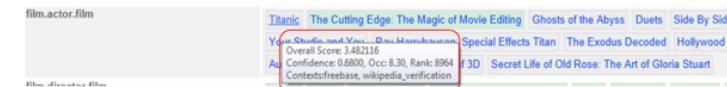
- `List<string> GetPredicatesByEntityId (string entityId)`

- Input
  - An entity id
- Output
  - A set of predicates



- `List<string[]> GetValuesByEntityPredicate (string entityId, string predicate)`

- Input
  - An entity id and a predicate
- Output
  - A set of <entity id, confidence score> pairs



# Latest Trends of QA

# Table-based QA (TBQA)

Reply user utterances (esp. questions) based on  
**tables together with their captions/surrounding texts.**

- Question: Greece held its last Summer Olympics in which year?
- Answer: 2004

City	Country	Continent	Summer	Winter	Year	Opening Ceremony	Closing Ceremony
Athens	Greece	Europe	I	—	1896	April 6	April 15
Paris	France	Europe	II	—	1900	May 14	October 28
St. Louis <sup>[a]</sup>	United States	North America	III	—	1904	July 1	November 23
London <sup>[b]</sup>	United Kingdom	Europe	IV	—	1908	April 27	October 31
Stockholm	Sweden	Europe	V	—	1912	May 5	July 22
...							
Athens	Greece	Europe	XXVIII	—	2004	August 13	August 29
Turin	Italy	Europe	—	XX	2006	February 10	February 26
Beijing <sup>[e]</sup>	China	Asia	XXIX	—	2008	August 8	August 24
Vancouver	Canada	North America	—	XXI	2010	February 12	February 28
London	United Kingdom	Europe	XXX	—	2012	July 27	August 12
Sochi	Russia	Europe <sup>[d]</sup>	—	XXII	2014	February 7	February 23
Rio de Janeiro	Brazil	South America	XXXI	—	2016	August 5	August 21

Host cities for Olympic Games in Summer and Winter

[https://en.wikipedia.org/wiki/List\\_of\\_Olympic\\_Games\\_host\\_cities](https://en.wikipedia.org/wiki/List_of_Olympic_Games_host_cities)

WikiTableQuestions dataset, each question comes with a table from Wikipedia. **Given the question and the table, the task is to answer the question based on the table.** The dataset contains **2,108** tables and **22,033** questions.

$R[\lambda x[Year.Date.x]].argmax(Country.Greece, index)$

The last row with country Greece

$\lambda x[Year.Date.x]$

$argmax(Country.Greece, index)$

List of all rows with country Greece

Country.Greece

Index

Greece

Country

1. Table selection based on caption and surrounding text
2. Topic entity detection
3. Logical form generation with flexible semantic parsing
4. Logical form ranking
5. Answer extraction based on generated logical form and the table

# Document-based QA (DBQA)

- Answer NL questions based on unstructured documents

## Oklahoma City bombing

From Wikipedia, the free encyclopedia  
(Redirected from Oklahoma city bombing)

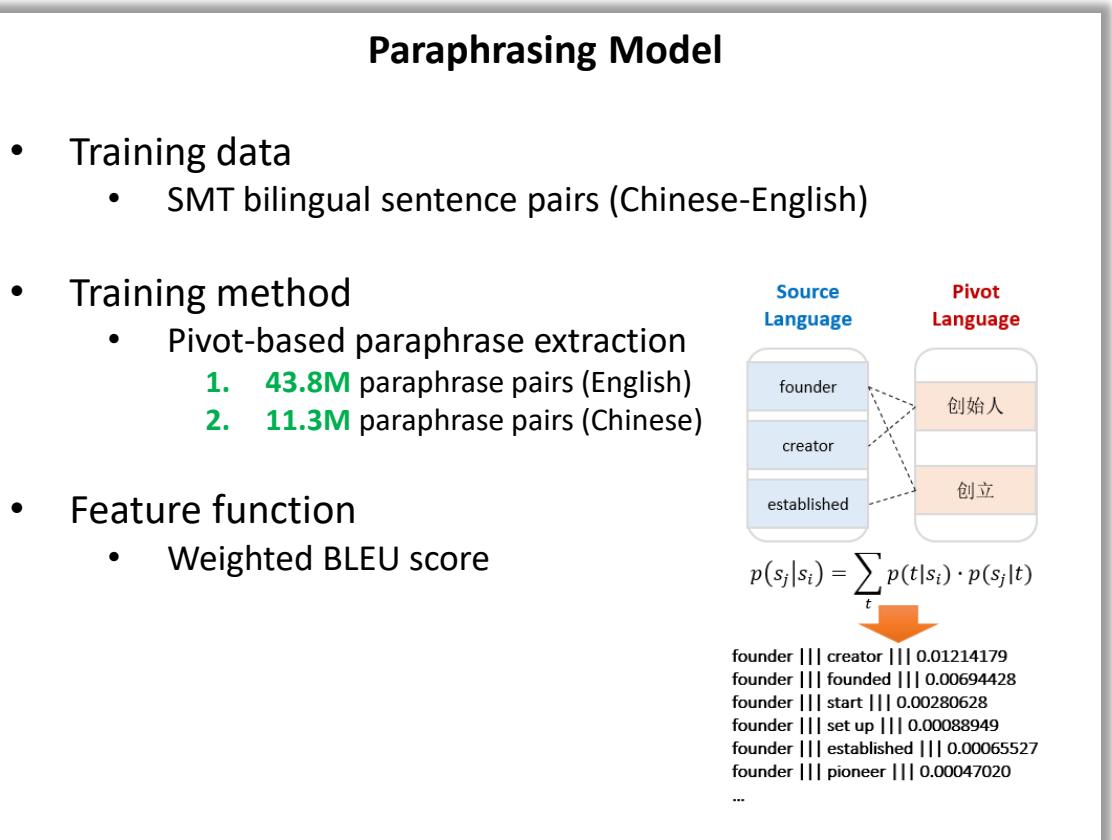
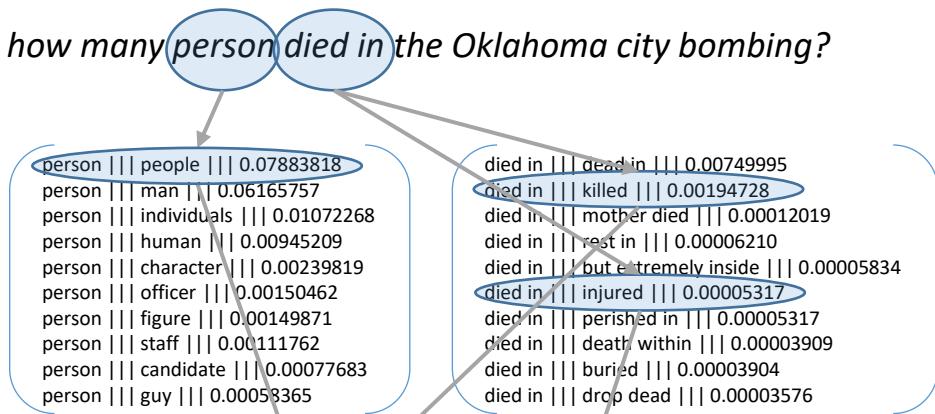
*how many person died in the  
Oklahoma city bombing?*



The **Oklahoma City bombing** was a domestic terrorist bomb attack on the Alfred P. Murrah Federal Building in downtown Oklahoma City on April 19, 1995. Carried out by Timothy McVeigh and Terry Nichols, the bombing killed 168 people<sup>[1]</sup> and injured more than 680 others.<sup>[2]</sup> The blast destroyed or damaged 324 buildings within a 16-block radius, destroyed or burned 86 cars, and shattered glass in 258 nearby buildings,<sup>[3][4]</sup> causing an estimated \$652 million worth of damage.<sup>[5]</sup> Extensive rescue efforts were undertaken by local, state, federal, and worldwide agencies in the wake of the bombing, and substantial donations were received from across the country. The Federal Emergency Management Agency (FEMA) activated eleven of its Urban Search and Rescue Task Forces, consisting of 665 rescue workers who assisted in rescue and recovery operations.<sup>[6][7]</sup>

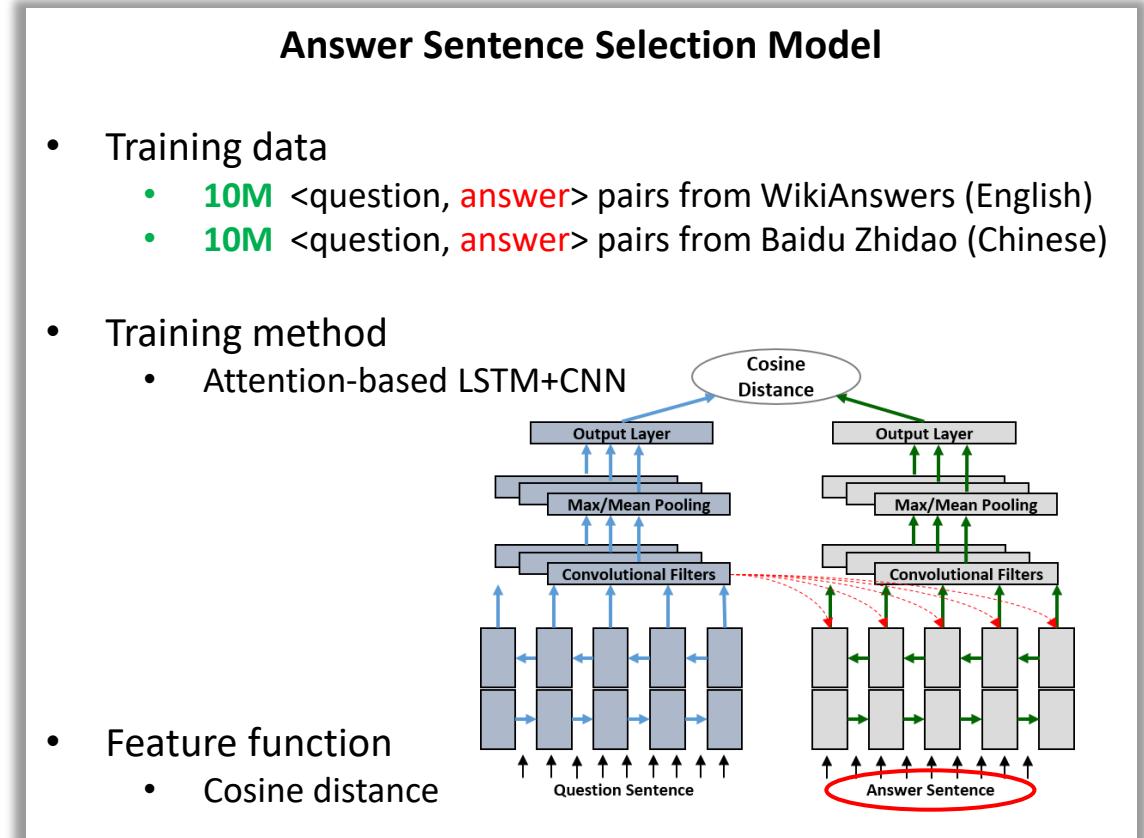
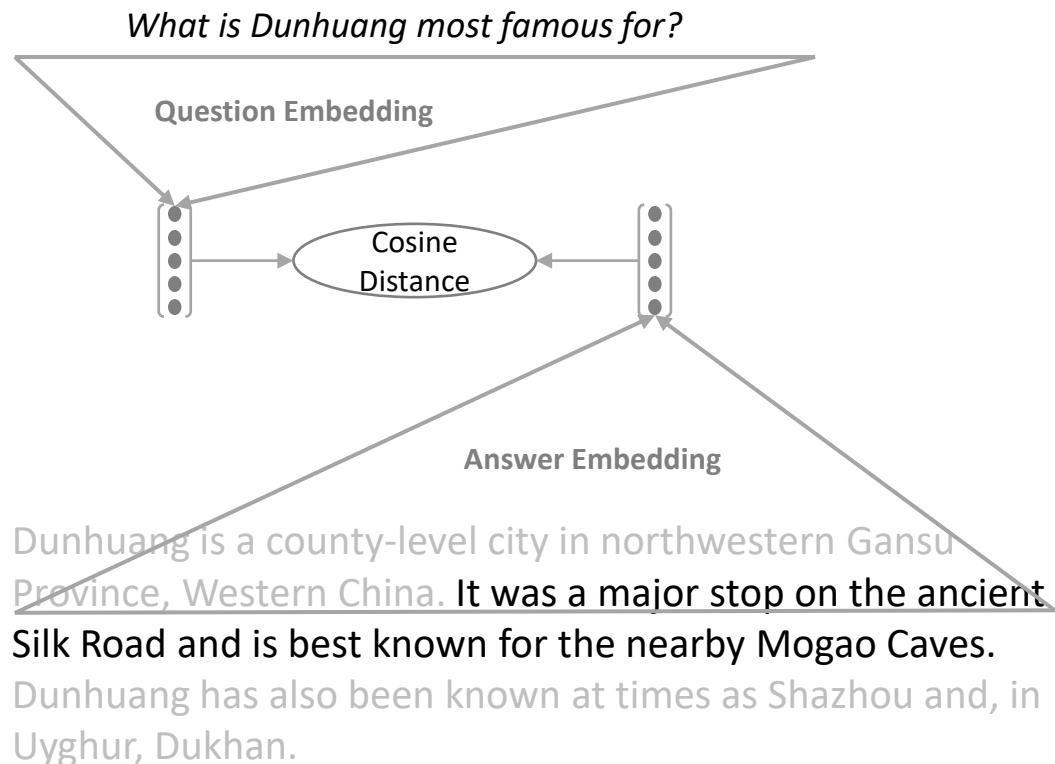
# Paraphrasing Model

- An approach to match language pieces with different expressions but similar/identical semantic meanings



# Answer Sentence Selection Model

- A deep learning approach to address question queries



# Topic Embedding Model

- A deep learning approach to *predict the topic of a given sentence*

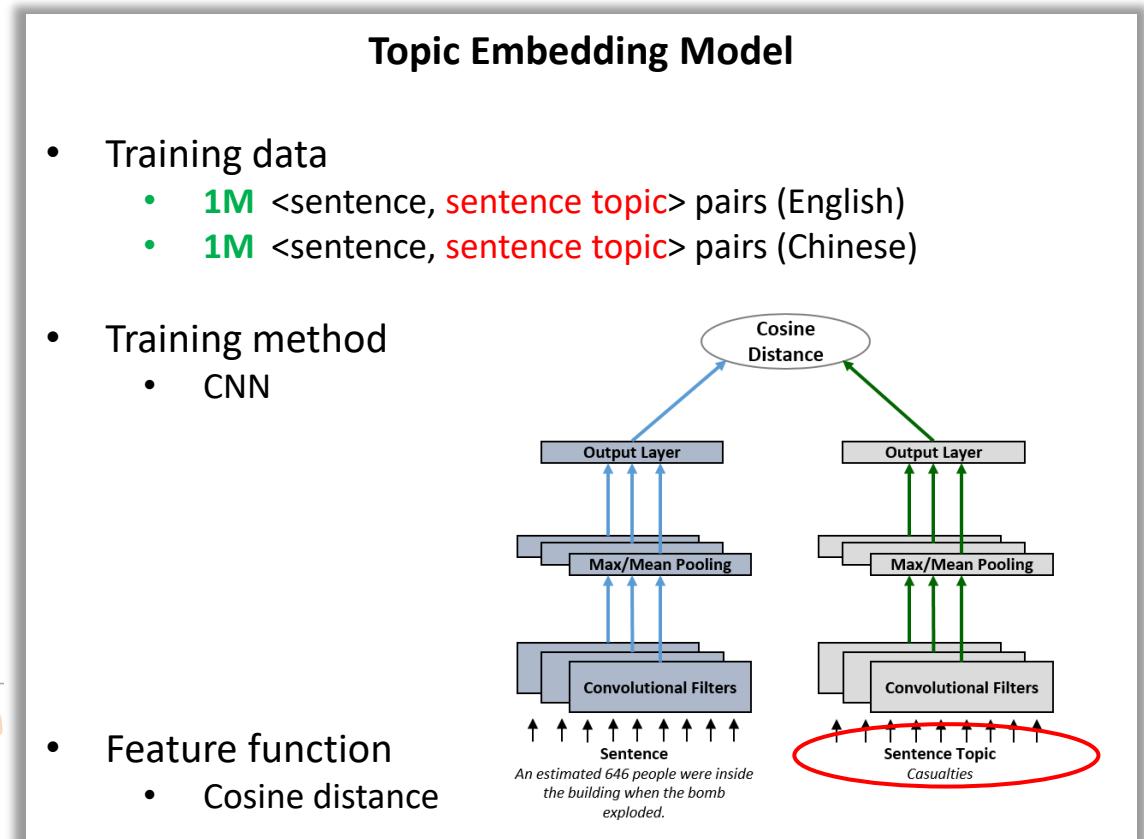
Contents [hide]

- 1 History
- 2 Culture
  - 2.1 Buddhist caves
  - 2.2 Other historical sites
  - 2.3 Museums
  - 2.4 Night market
- 3 Climate
- 4 Transportation
- 5 See also
- 6 Gallery
- 7 Footnotes
- 8 References
- 9 External links

History [edit]

training instance

There is evidence of human habitation in the Dunhuang area as early as 2,000 BC, possibly by people recorded as the [Qiang](#) in Chinese history. Its name was also mentioned as part of the homeland of the [Yuezhi](#) in the [Records of the Grand Historian](#). While some<sup>[who?]</sup> have



# More Features (Yan et al., ACL 2016)

- Word Embedding Model
- Translation Model
- Discourse Model
- Relation Embedding Model
- Type Embedding Model
- ...

# Evaluation on Answer Sentence Selection

- Dataset
  - WikiQA (English)
    - 2,118 question-document pairs in training set
    - 633 question-document pairs in testing set

	MRR (on Testing Set)
WikiQA	72.2% <small>(state-of-the-art result)</small>

- DBQA (Chinese)
  - Homework!!!
  - NLPCC 2016 QA shared task

What bird family is the owl? [from WikiQA]

1. Owls are a group of birds that belong to the order Strigiformes, constituting 200 extant bird of prey species.
2. Most are solitary and nocturnal, with some exceptions (e.g., the Northern Hawk Owl).
3. Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish.
4. Earth except Antarctica, most of Greenland and some remote islands.
5. Owls are characterized by their small beaks and wide faces, and are divided into two families: the typical owls, Strigidae; and the barn-owls, Tytonidae.

# From QA Engines to Conversational Engines

# Practical Applications

- KB-based Conversation
  - JD Xiaoice (京东小冰)
  - CEAir Xiaoice (东航小冰)
- Document-based Conversation
  - Dunhuang Xiaoice (敦煌小冰)

# JD Xiaoice (京东小冰)

- Platform

- JD App

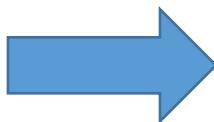


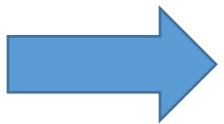
- Function

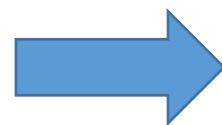
- ChitChat

- Product Card

- Product Chat







# Dunhuang Xiaoice (敦煌小冰)

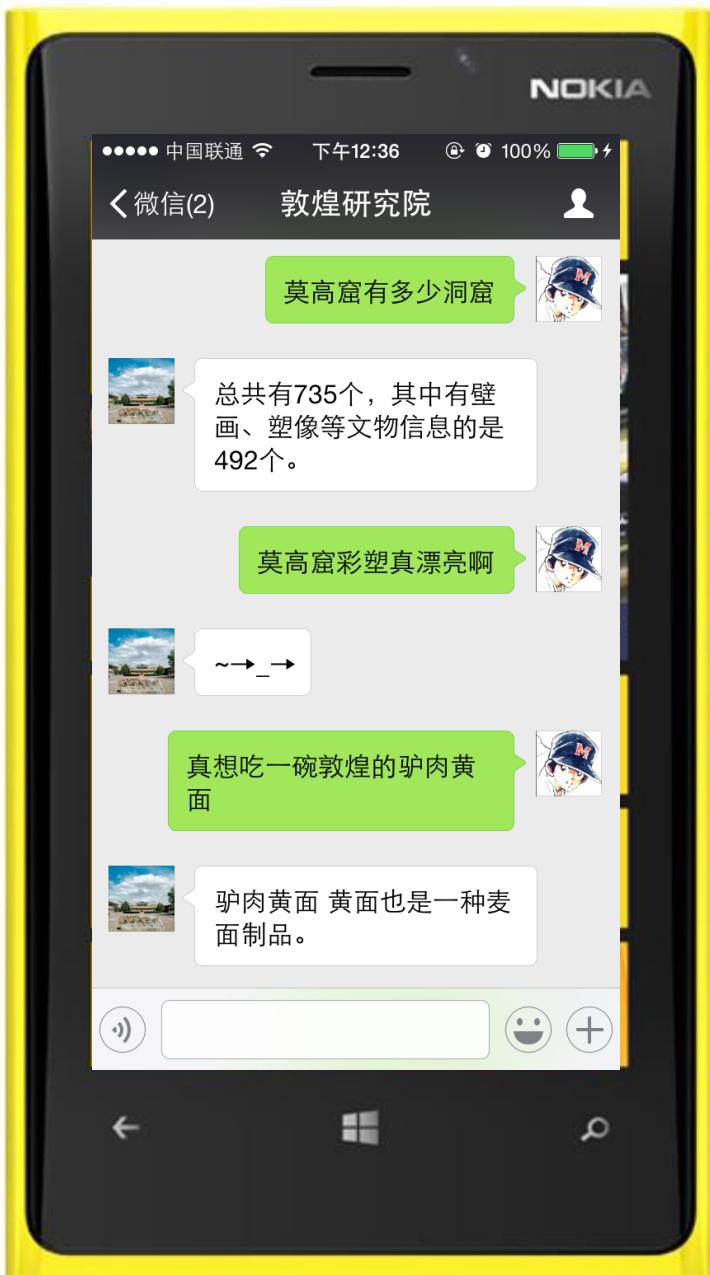
- Platform

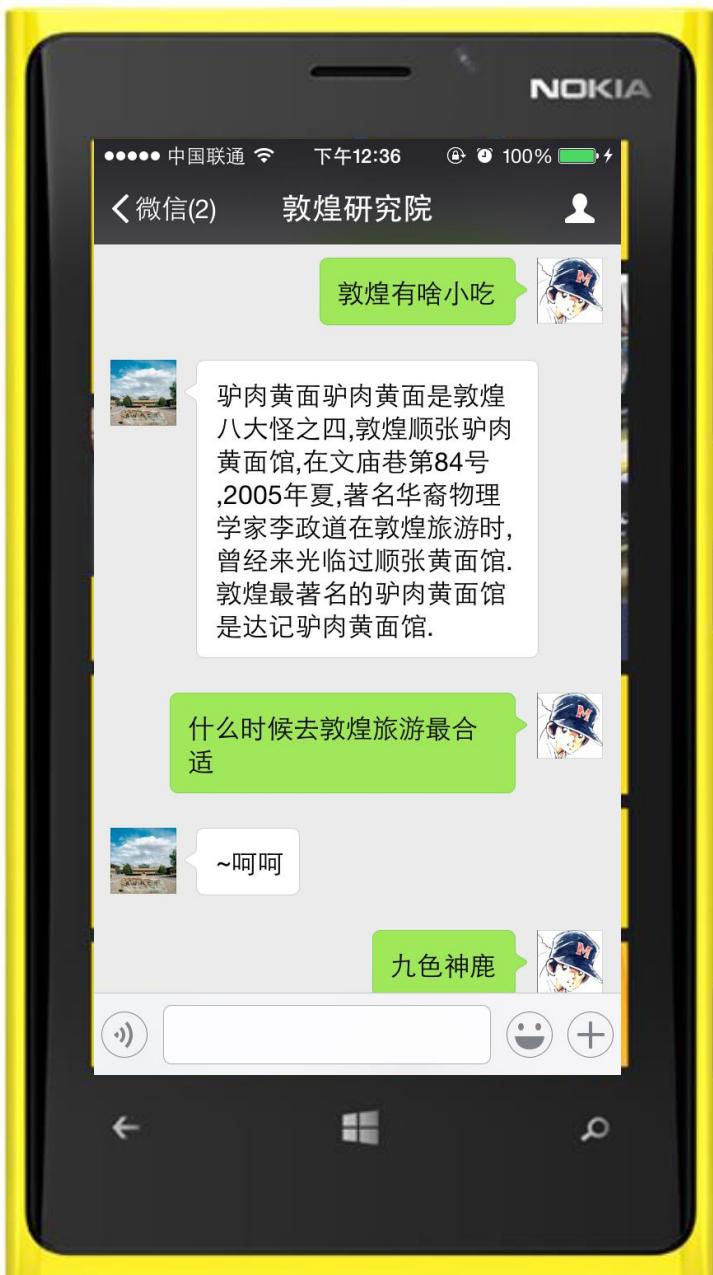
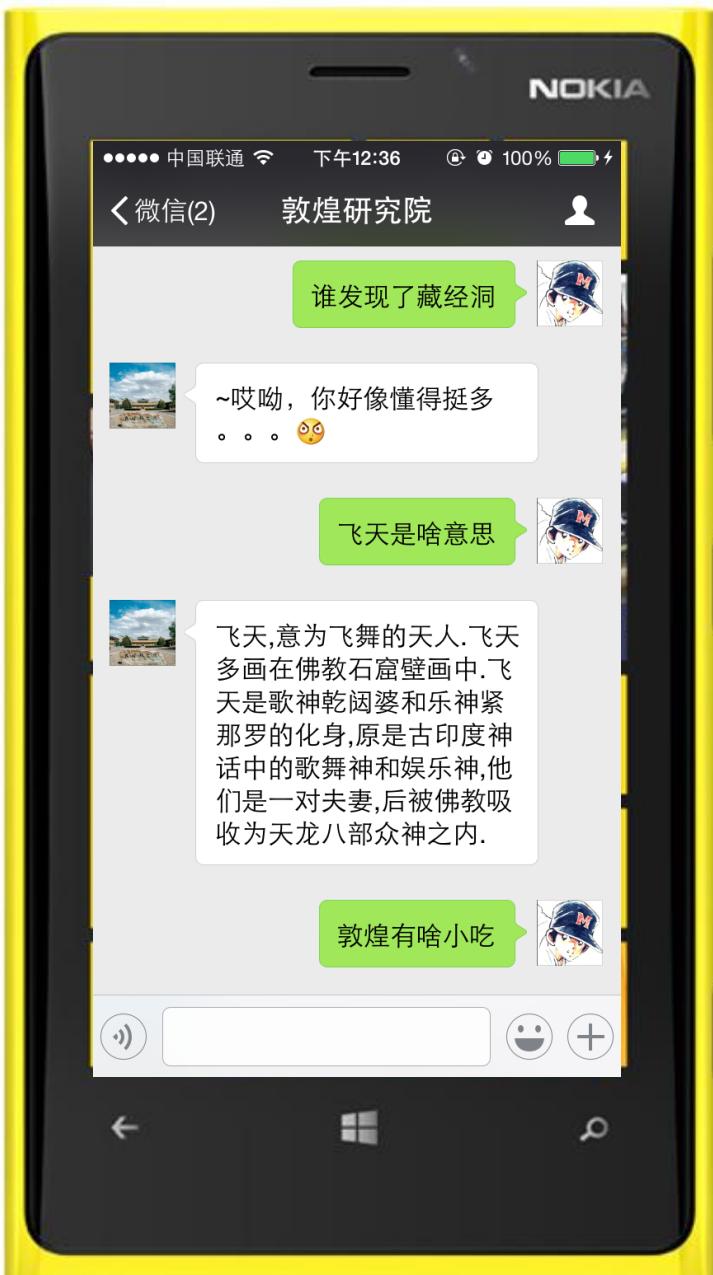
- WeChat

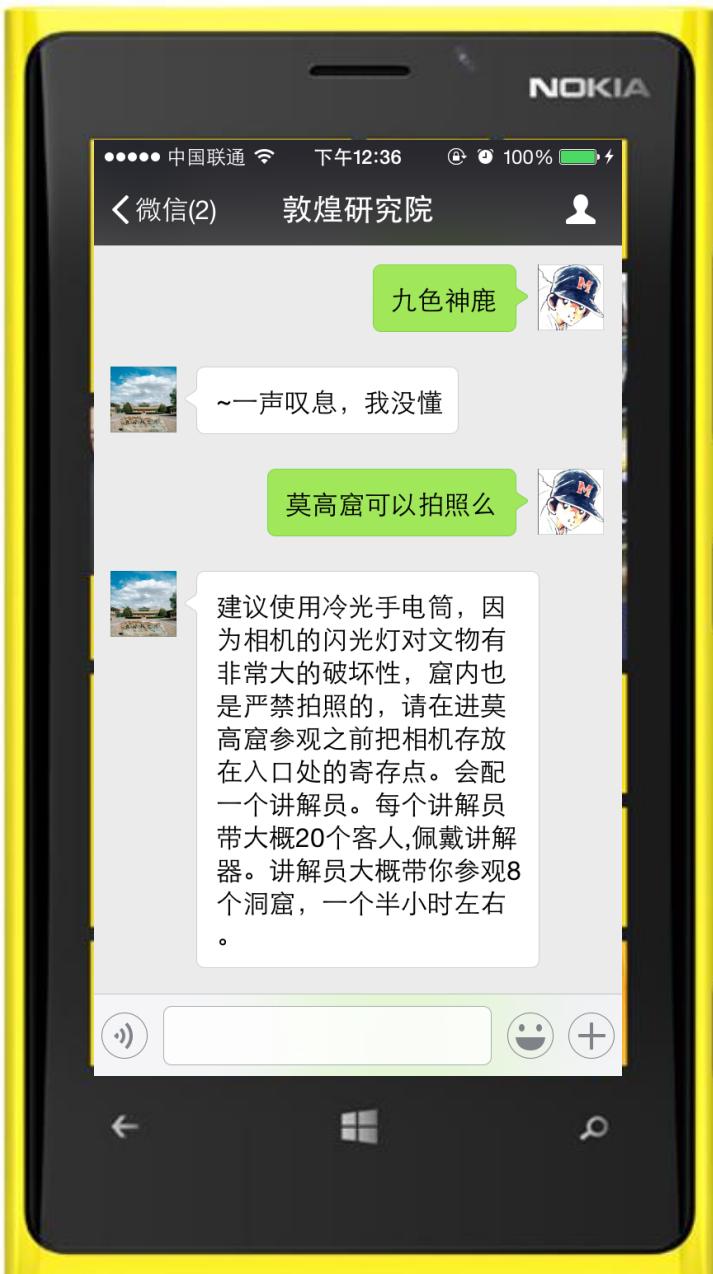


- Function

- ChitChat
  - DocCard
  - DocChat







# Key Challenges/Future Work

- Understanding
- Trigger
- Style
- Multi-turn
- Proactive

# Thanks

- nanduan@microsoft.com

# Reference (1): SP-QA

- Building a Semantic Parser Overnight
  - Wang et al., 2015, ACL
- Compositional Semantic Parsing on Semi-Structured Tables
  - Pasupat and Liang, 2015, ACL
- Knowledge-Based Question Answering as Machine Translation
  - Bao et al., 2014, ACL
- Joint Relational Embeddings for Knowledge-Based Question Answering
  - Yang et al., 2014, EMNLP
- Large-scale Semantic Parsing without Question-Answer Pairs
  - Reddy et al., 2014, TACL
- Semantic Parsing on Freebase from Question-Answer Pairs
  - Berant et al., 2013, EMNLP
- Scaling Semantic Parsers with On-the-fly Ontology Matching
  - Kwiatkowski et al., 2013, EMNLP
- Lexical generalization in CCG grammar induction for semantic parsing
  - Kwiatkowski et al., 2011, EMNLP
- Online Learning of Relaxed CCG Grammars for Parsing to Logical Form
  - Zettlemoyer and Collins, EMNLP, 2007

Red-colored papers are  
from MSR

# Reference (2): IR-QA

- Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base
  - Yih et al., 2015, ACL
- Large-scale Simple Question Answering with Memory Network
  - Bordes et al., 2015, ICLR
- Towards AI-complete Question Answering: A Set of Prerequisite Toy Tasks
  - Bordes et al., 2015, arXiv
- Memory Networks
  - Weston et al., 2015, ICLR
- Question Answering over Freebase with Multi-Column Convolutional Neural Networks
  - Dong et al., 2015, ACL
- Lean Question Answering over Freebase from Scratch
  - Yao, 2015, NAACL
- Semantic parsing via paraphrasing
  - Berant and Liang, 2014, ACL
- Information extraction over Structured Data: Question Answering with Freebase
  - Yao and Van Durme, 2014, ACL
- Question Answering with Subgraph Embeddings
  - Bordes et al., 2014, EMNLP
- Open Question Answering with Weakly Supervised Embedding Models
  - Bordes et al., 2014, ECML-PKDD

Red-colored papers are  
from MSR

# Reference (3): Open-KB-QA

- Answering Questions with Complex Semantic Constraints on Open KBs
  - Yin et al., 2015, CIKM
- Open Question Answering over Curated and Extracted Knowledge Bases
  - Fader et al., 2014, KDD
- Paraphrase-Driven Learning for Open Question Answering
  - Fader et al., 2013, ACL
- ClausIE: Clause-based Open Information Extraction
  - Corro et al., 2013, WWW
- Open Language Learning for Information Extraction
  - Mausam et al., 2012, EMNLP
- Natural Language Questions for the Web of Data
  - Mohamed, 2012, EMNLP-CoNLL
- Identifying Relations for Open Information Extraction
  - Fader et al., 2011, EMNLP
- Open Information Extraction using Wikipedia
  - Wu et al., 2010, ACL
- Open Information Extraction from the Web
  - Banko et al., 2007, IJCAI

Red-colored papers are  
from MSR

# Reference (4): Our Work on Web-QA, Social-QA

- Web-QA
  - Answer Extraction with Multiple Extraction Engines for Web-based Question Answering, Hong Sun, Furu Wei, Ming Zhou, NLPCC 2014
  - Answer Extraction from Passage Graph for Question Answering, Hong Sun ,Nan Duan, Yajuan Duan, MingZhou, IJCAI 2013
- Social-QA
  - Improving Search Relevance for Short Queries in Community Question Answering, Haocheng Wu, Wei Wu, Ming Zhou, Enhong Chen, Lei Duan, Heung-Yeung Shum, WSDM 2014
  - Mining Query Subtopics from Questions in Community Question Answering, Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou, AAAI 2015
  - Question Retrieval with High Quality Answers in Community Question Answering, Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou, CIKM 2014