Knowledge-based 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?

Knowledge-QA

Barack Obama

★ Get updates

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

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)

Multiple Intelligence in Modern QA

Who is the wife of Barack Obama?

Document-QA

Multiple Intelligence in Modern QA

Who is the wife of Barack Obama?

Social-QA

Quora

YAHOO! ANSWERS

ChaCha Questions

Answers™

Multiple Intelligence in Modern QA

Who is the wife of Barack Obama?

Today's focus!

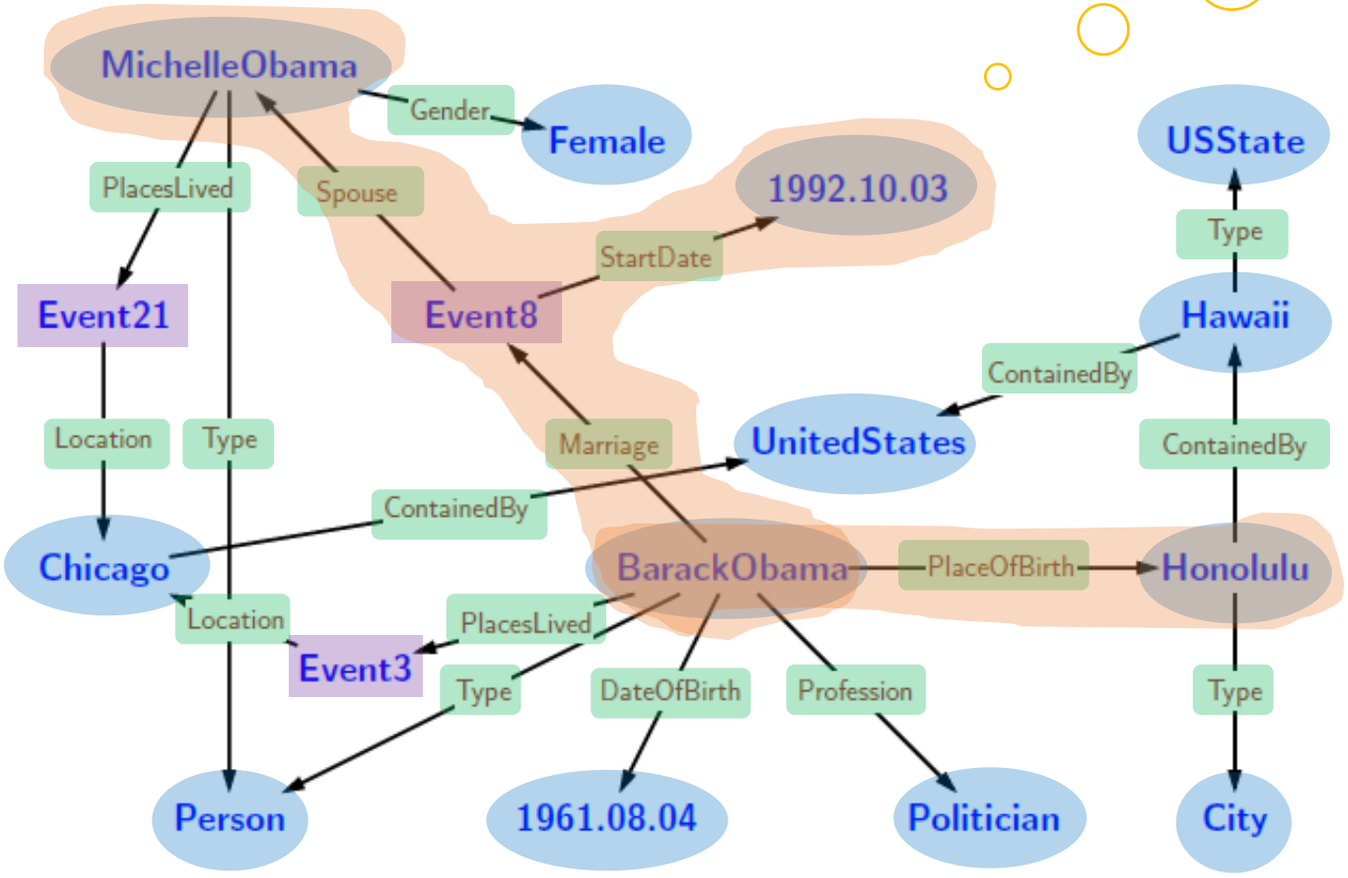
Knowledge-QA

Document-QA

Social-QA

Knowledge-based QA

Knowledge-Base (KB)

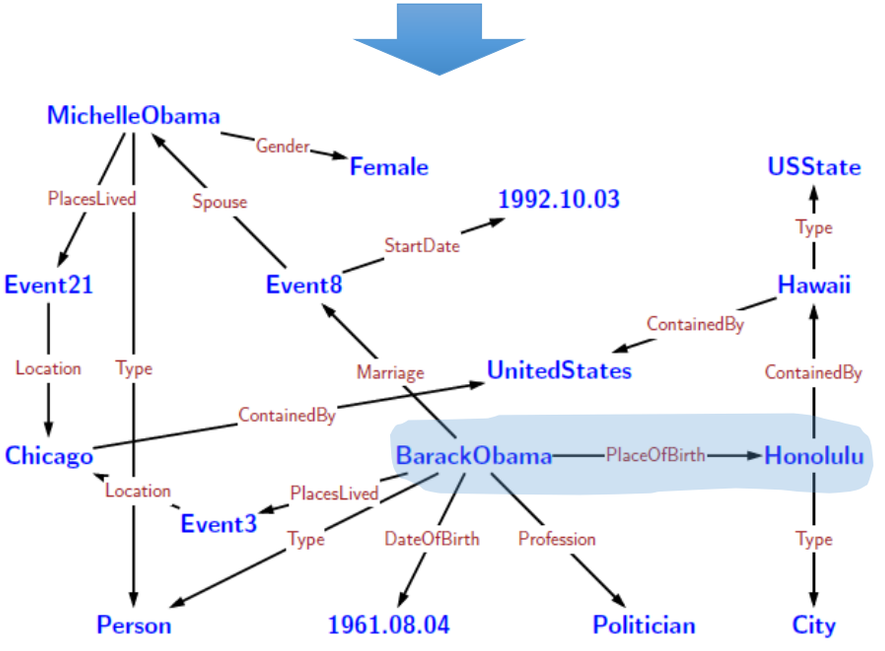


- **Entity**
- **Predicate**
  Relation between two connected entities
- **CVT** (Compound Value Type)
  Not a real-world entity, but is used to collect multiple fields of an event
- **Fact**
  Triple, which connects two entities
  Event, which connects multiple entities via a CVT node

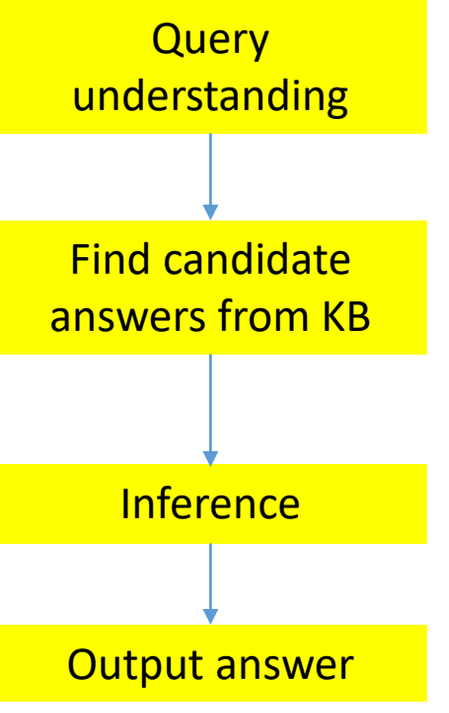
Knowledge-Based Question Answering (KB-QA)

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

Question: *where was Barack Obama born ?*

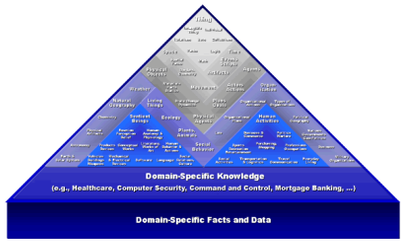


Answer: Honolulu

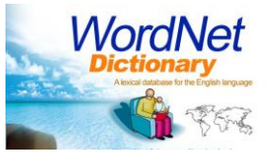


Evolution of KB-QA (1980~)

(isA, Bill Clinton, United States President)



CYC: Ontology + Common Sense Knowledge



guitarist ⊂ {player,musician} ⊂ artist

algebraist ⊂ mathematician ⊂ scientist

Extract structured knowledge from Wikipedia

Google Knowledge Graph

1984 | 1985 | 2001 | 2007 | 2008 | 2010 | 2012 | 2013 | 2015 | …

Template-based KB-QA

KB-QA using Semantic Parsing (CCG, λ-calculus, DCS,…)

Open IE/QA (TextRunner, ReVerb, NELL,…)

KB-QA using Information Retrieval (NLG, Graph Search,…)

KB-QA using Deep Learning

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 ?

Semantic Parsing

CCG | DCS | SMT | ...

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

KB Lookup

<Barack Obama, Place_of_Birth, Honolulu>

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

where was Barack Obama born ?

Ranker

what is the place of birth of Barack Obama

Barack Obama
Place_of_Birth
Honolulu
...
President

NLG

Embedding

<Barack Obama, Place_of_Birth, Honolulu>

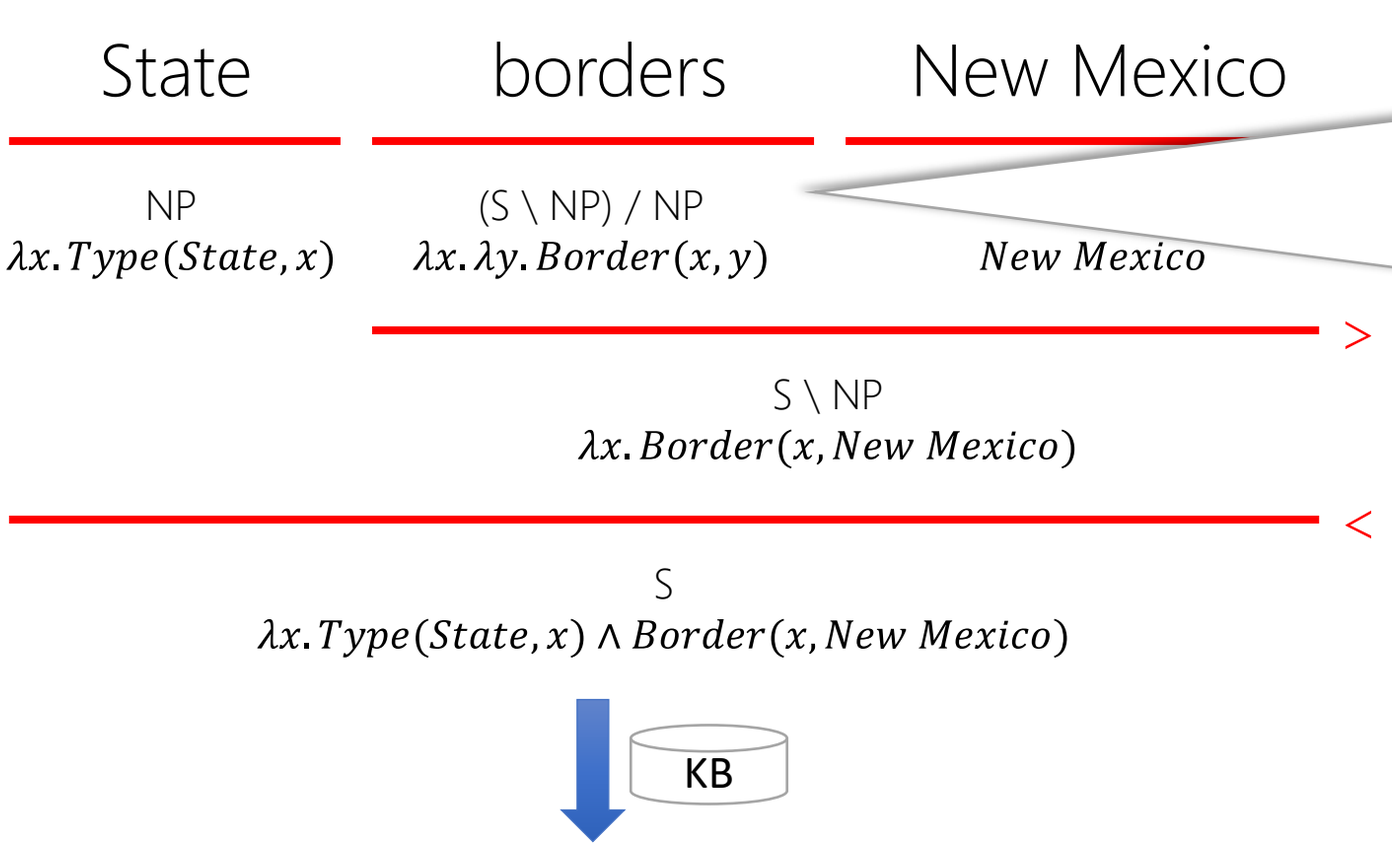
KB-QA Methodology

- 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



Arizona, Colorado, Oklahoma, Texas

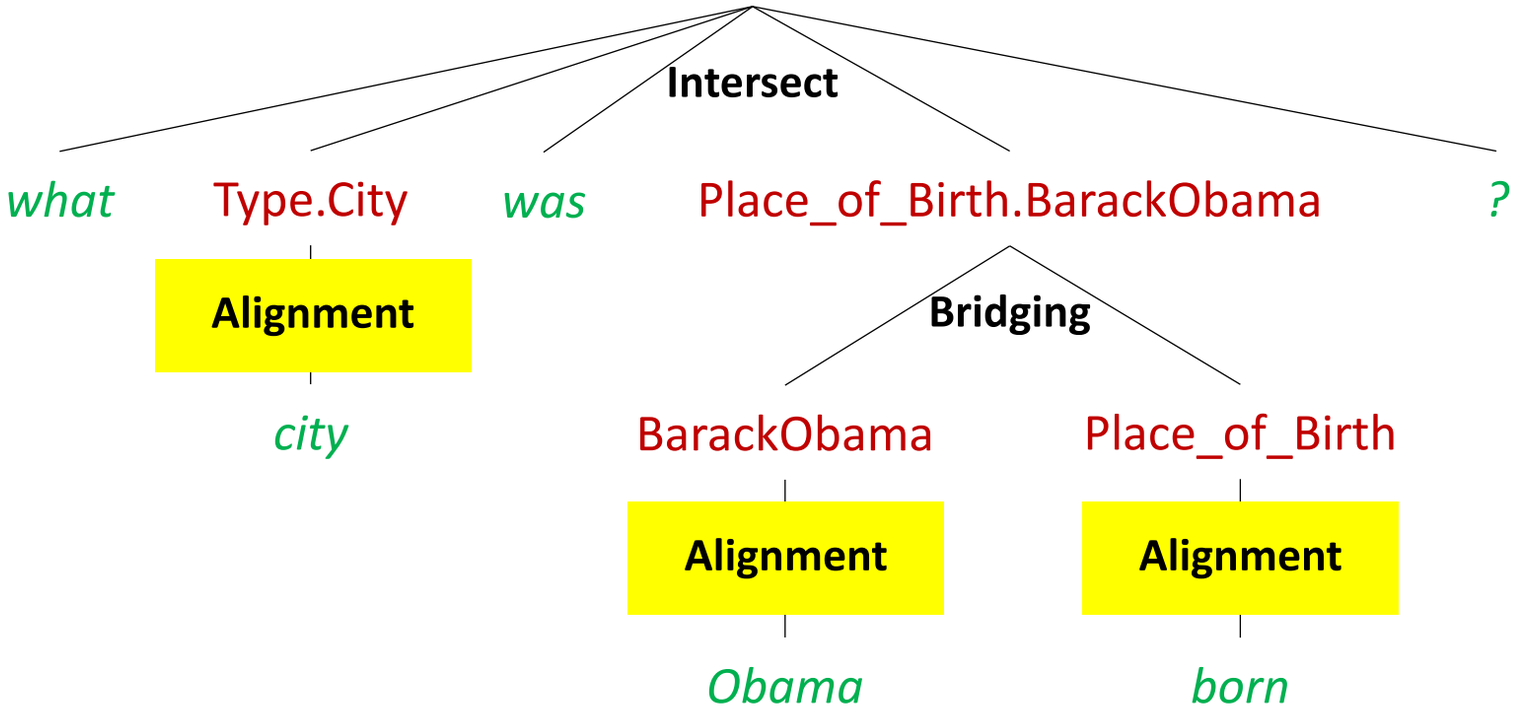
A CCG is defined by

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

Semantic Parsing with DCS

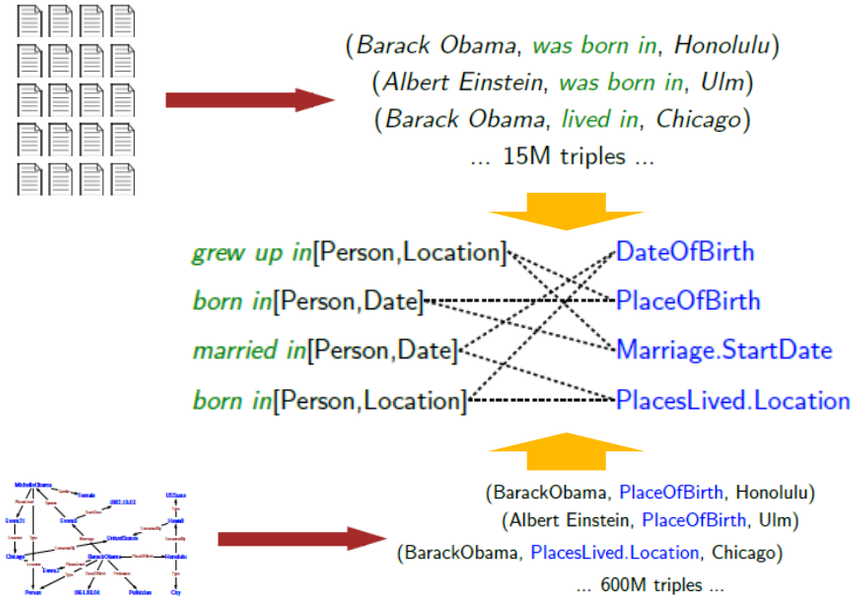
$$\lambda x.Type(City, x) \wedge Place\_of\_Birth(Barack\ Obama, x)$$

Type.City ∩ Place_of_Birth.BarackObama



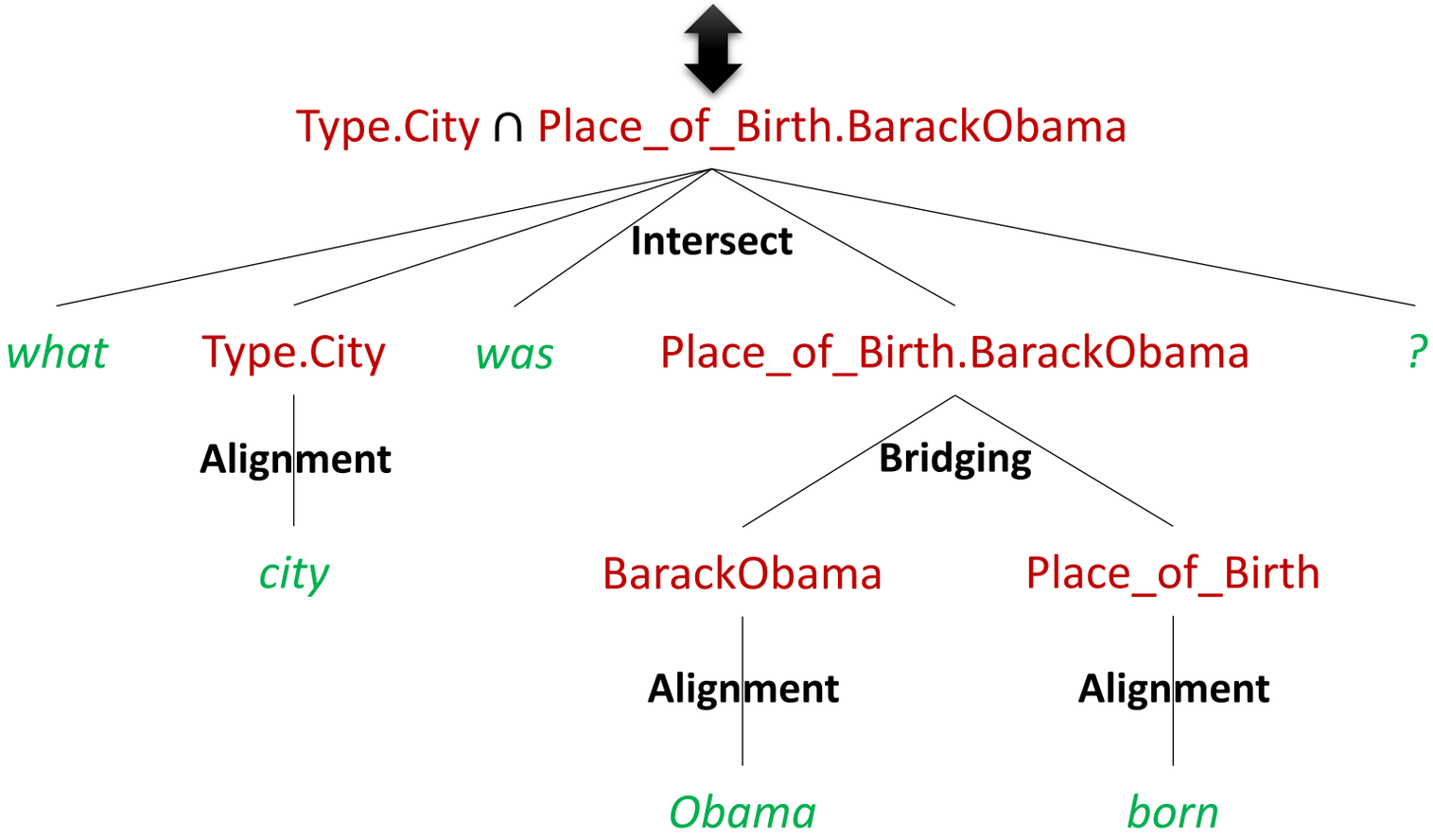
Alignment

- 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 ($\lvert\mathcal{F}(r_1)\rvert$) Log of # entity pairs that occur with the logical predicate r_2 ($\lvert\mathcal{F}(r_2)\rvert$) Log of # entity pairs that occur with both r_1 and r_2 ($\lvert\mathcal{F}(r_1) \cap \mathcal{F}(r_2)\rvert$) Whether r_2 is the best match for r_1 ($r_2 = \arg\max_r \lvert\mathcal{F}(r_1) \cap \mathcal{F}(r)\rvert$)
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 ($\lvert\mathcal{F}(b)\rvert$) 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

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

Semantic Parsing with SMT

- Perform semantic parsing & answer generation in a synchronous manner, based on the machine translation framework

- Given an NL question $\mathcal{Q}$, generate $\mathcal{H}(\mathcal{Q})$, which encodes a set of semantic representation-answer pairs $\{<\mathcal{D}_1, \mathcal{A}_1>, \ldots, <\mathcal{D}_N, \mathcal{A}_N>\}$

Parameter Optimization by Q-A Pairs

Feature Design

$$<\hat{\mathcal{D}}, \hat{\mathcal{A}}> = argmax_{<\mathcal{D},\mathcal{A}>\in\mathcal{H}(\mathcal{Q})} \frac{exp\{\sum_j \lambda_j \cdot h_j(\mathcal{D}, \mathcal{A}, \mathcal{Q})\}}{\sum_{<\mathcal{D}',\mathcal{A}'>\in\mathcal{H}(\mathcal{Q})} exp\{\sum_j \lambda_j \cdot h_j(\mathcal{D}', \mathcal{A}', \mathcal{Q})\}}$$

Search Space Generation by Semantic Parsing

Semantic Parsing with SMT

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

<[Film], Film.Film.Director, [Person]>

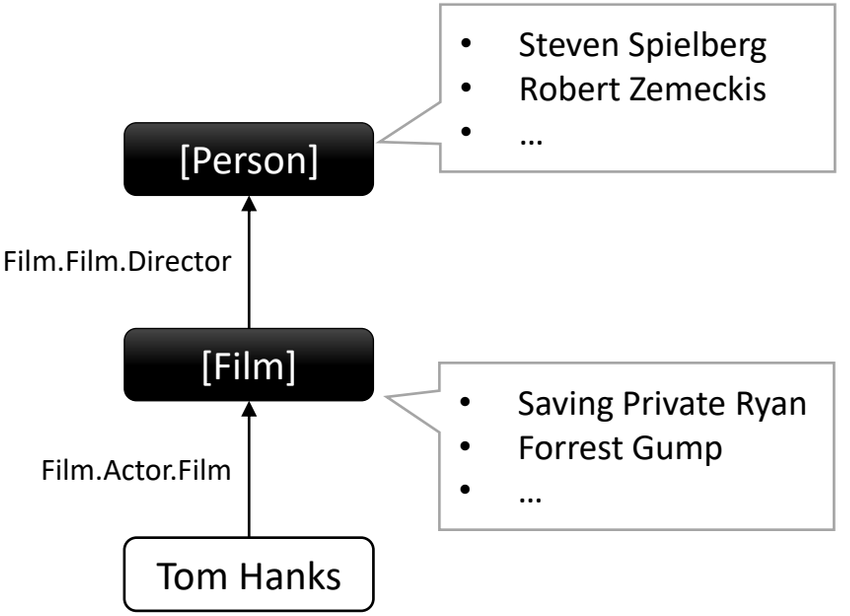
director of [Film]

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

<Tom Hanks, Film.Actor.Film, [Film]>

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
…	…	…

Predicate Grounding

Microsoft was founded by whom

<Microsoft, Organization.Organization.Founder, ?>

Pattern

Organization.Organization.Founder

Corpus with Linked Entities

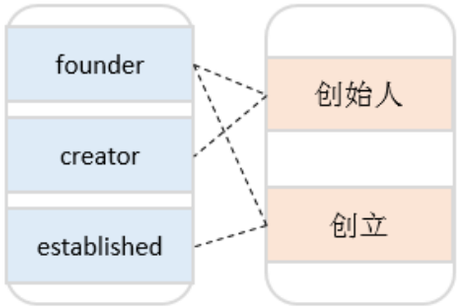
- Bill Gates *is the creator of* Microsoft
- Apple *was established by* Steve Jobs
- Mark Zuckerberg *was the founder of* Facebook
- Google *was founded by* Larry Page and Sergey Brin
- …

[Person] *is the creator of* **[Organization]**
[Organization] *was established by* **[Person]**
[Person] *was the founder* **[Organization]**
[Organization] *was founded by* **[Person]**
…

Paraphrase

Organization.Organization.Founder

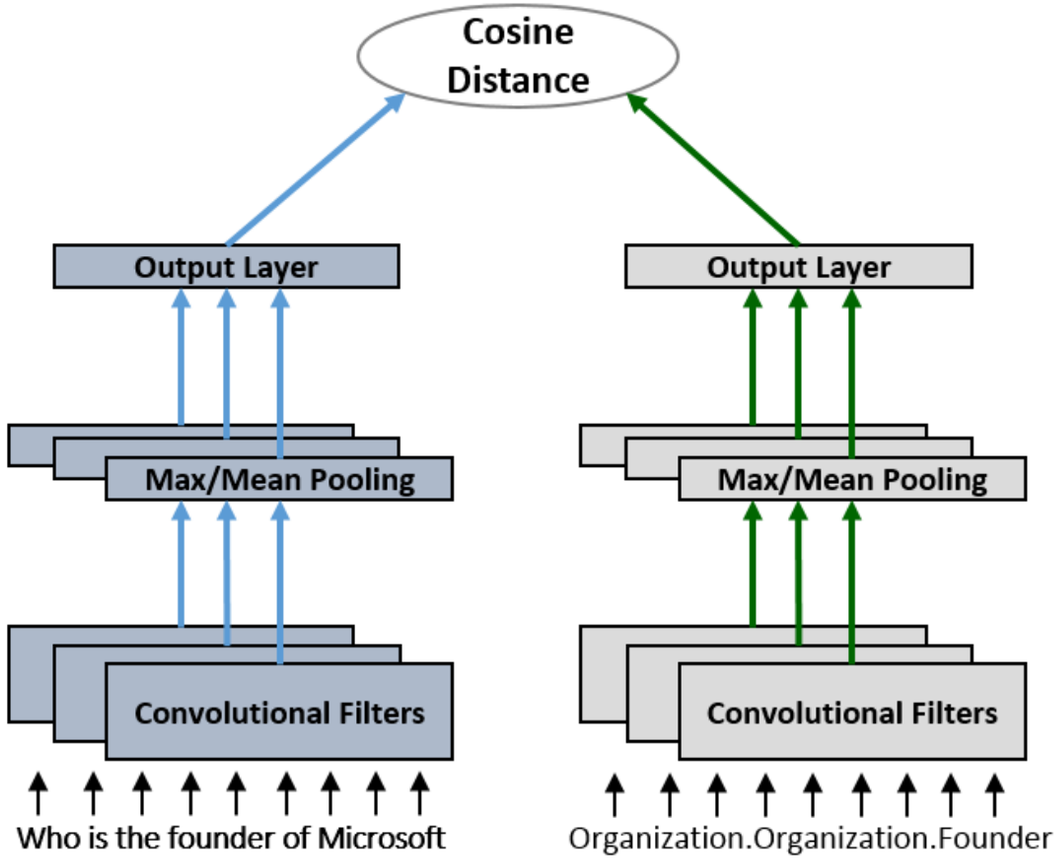
Source Language → **Pivot Language**



$$p(s_j|s_i) = \sum_t p(t|s_i) \cdot p(s_j|t)$$

founder ||| creator ||| 0.01214179
founder ||| founded ||| 0.00694428
founder ||| start ||| 0.00280628
founder ||| set up ||| 0.00088949
founder ||| established ||| 0.00065527
founder ||| pioneer ||| 0.00047020
…

Convolution Neural Network

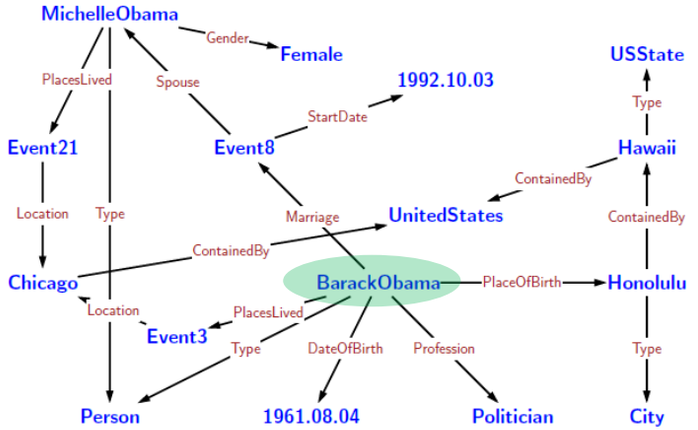


KB-QA Methodology

- 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 ?

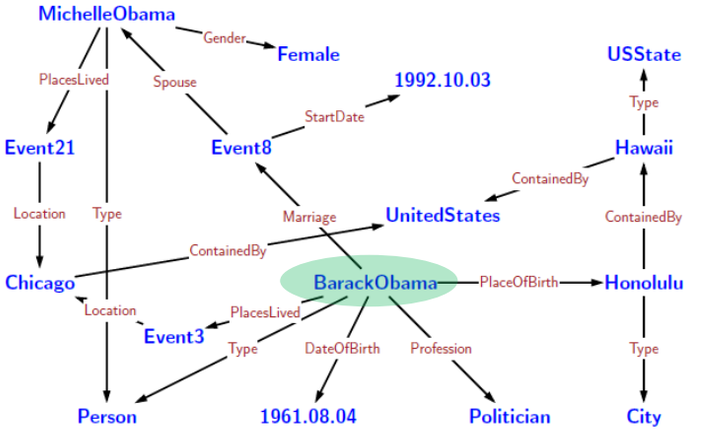


Template	Example
$p.e$	Directed.TopGun
$p_1.p_2.e$	Employment.EmployerOf.SteveBalmer
$p.(p_1.e_1 \sqcap p_2.e_2)$	Character.(Actor.BradPitt⊓Film.Troy)
Type.$t \sqcap z$	Type.Composer⊓SpeakerOf.French
count(z)	count(BoatDesigner.NatHerreshoff)

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 ?

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

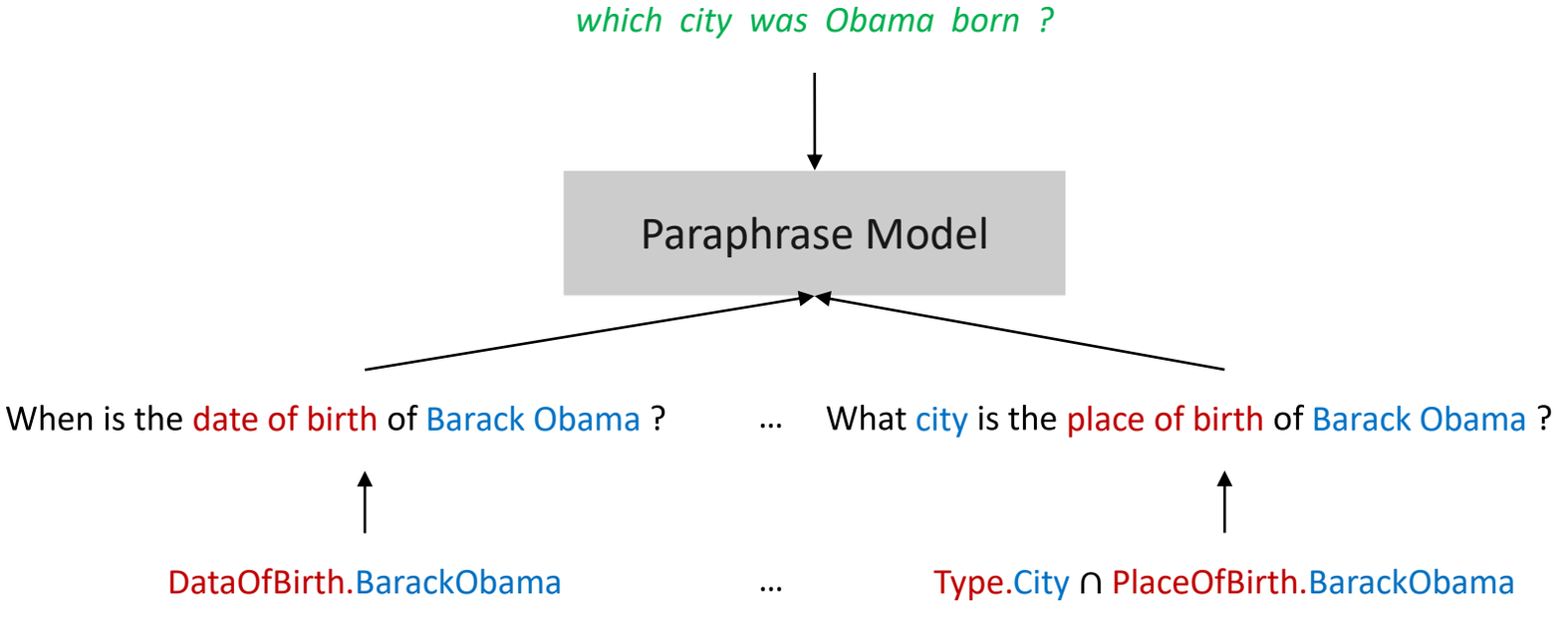
When is the date of birth of Barack Obama ? … What city is the place of birth of Barack Obama ?

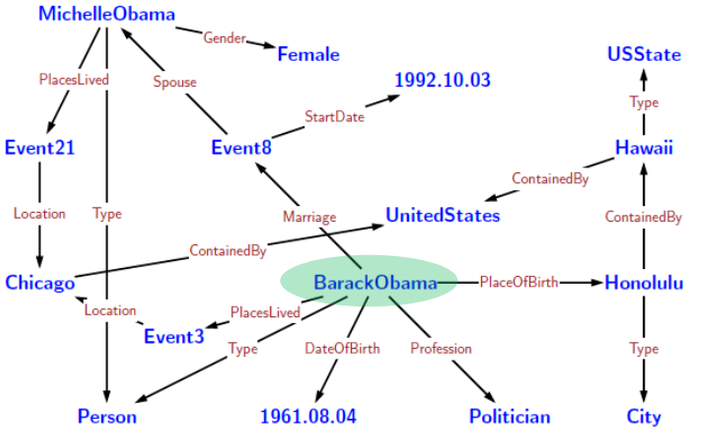
DataOfBirth.BarackObama … Type.City ∩ PlaceOfBirth.BarackObama

	$d(p)$ 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





Association Model

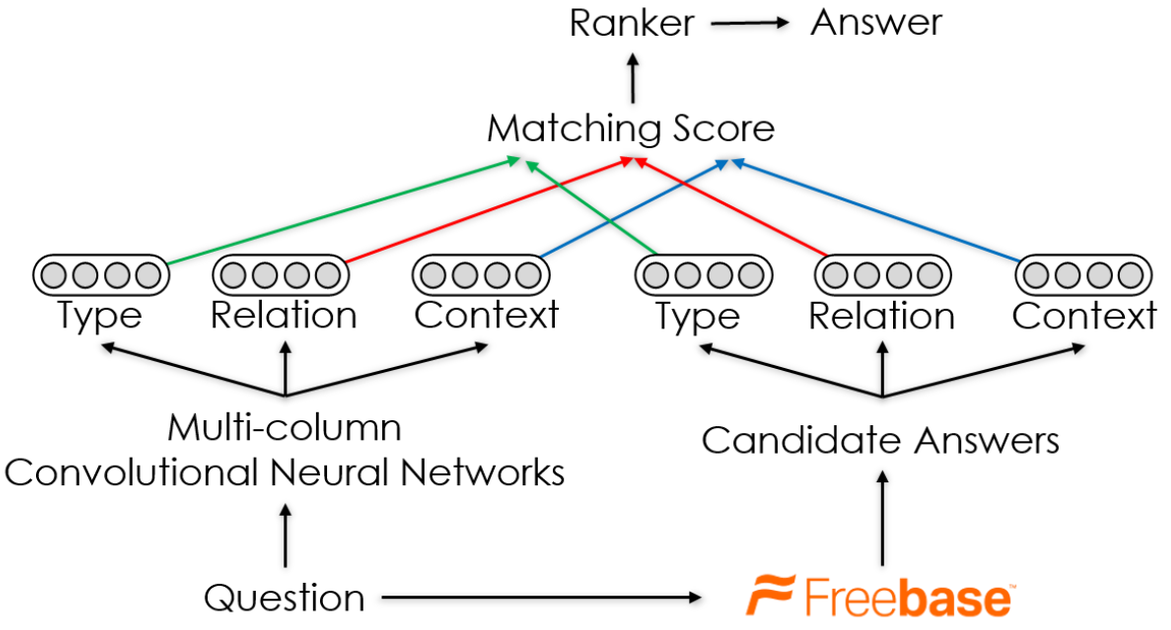
- Learn associated phrase pairs from 18M word-aligned question pairs from WikiAnswers (Fader et al., 2013)

Vector Space Model

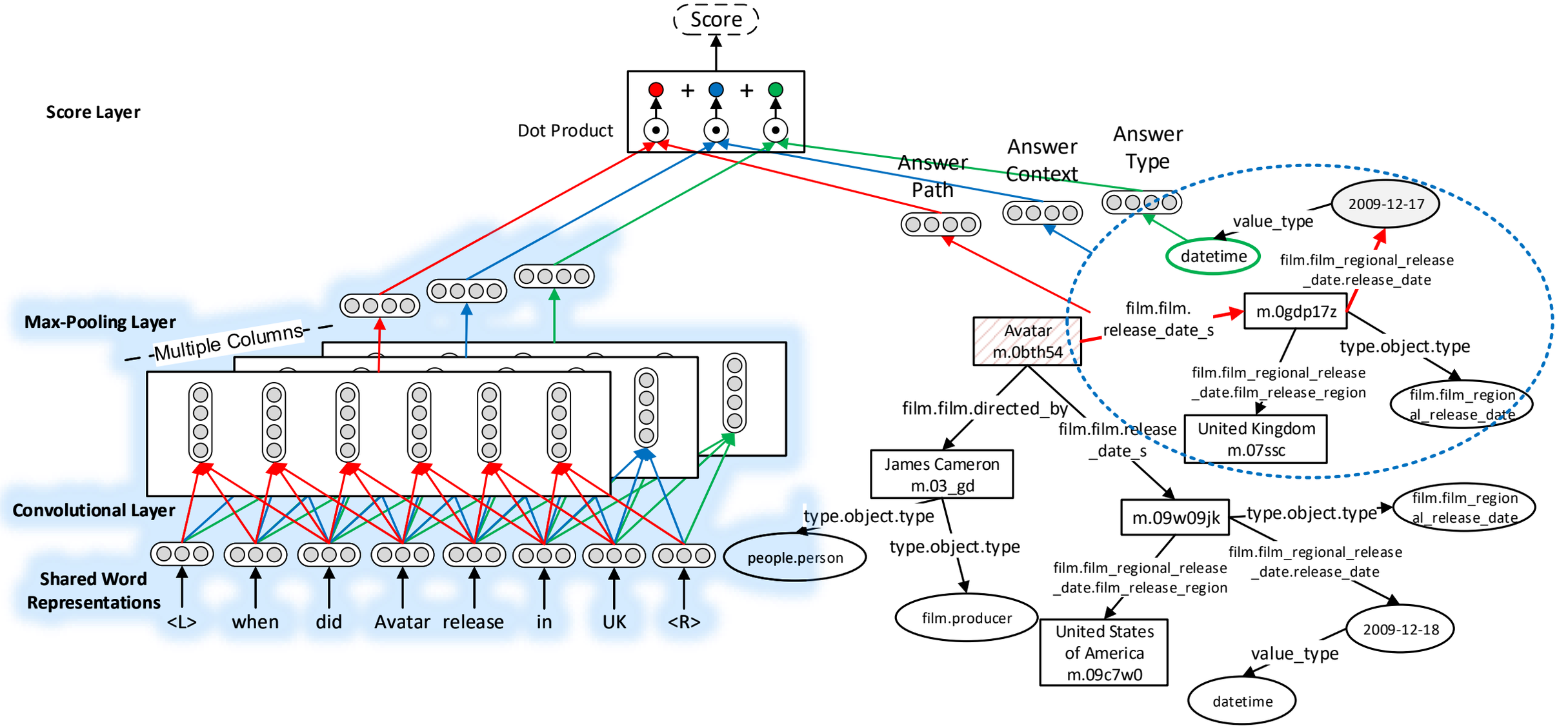
- Learn word embedding vectors from Wikipedia text using the CBOW model (Mikolov et al., 2013)

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

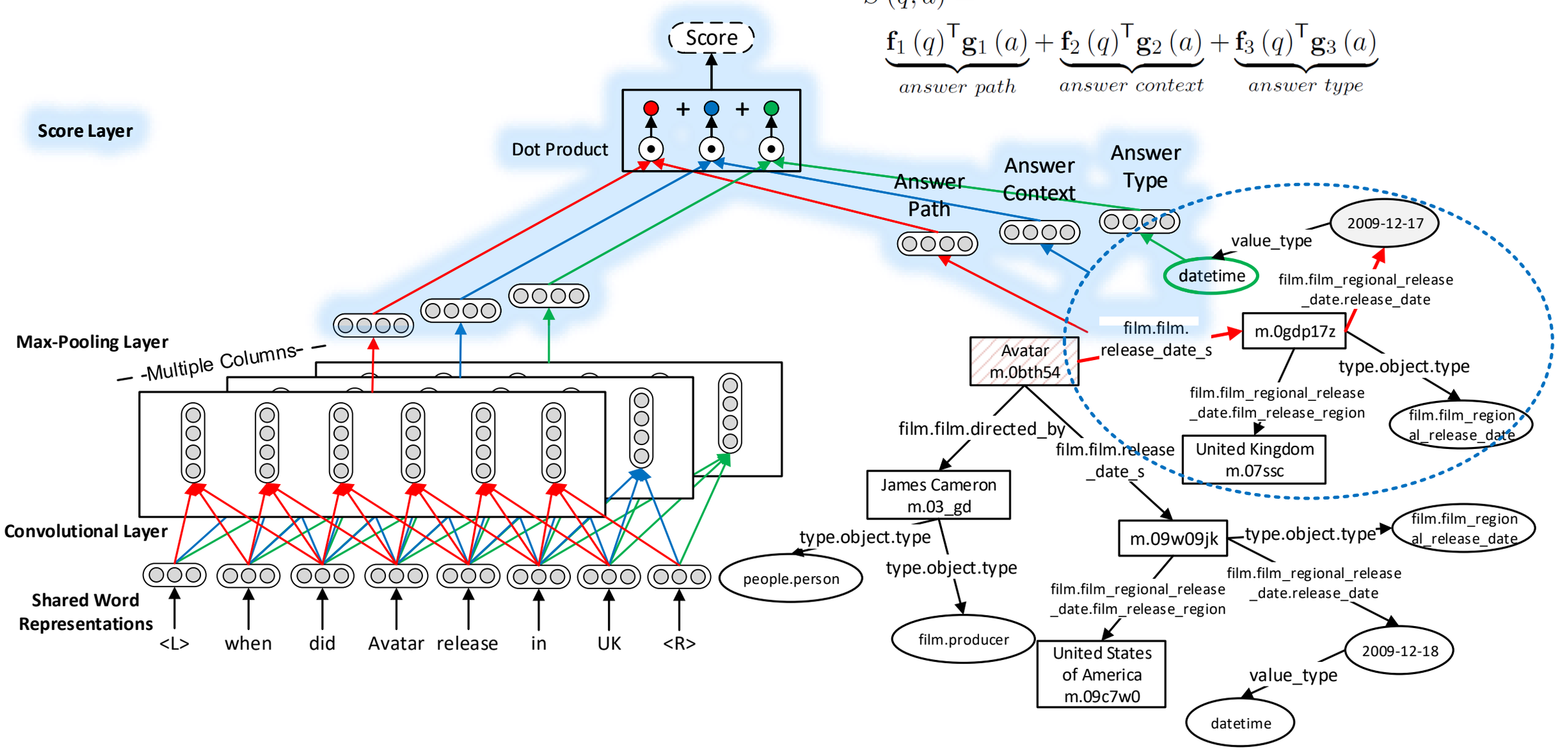


Multi-Column CNN Model

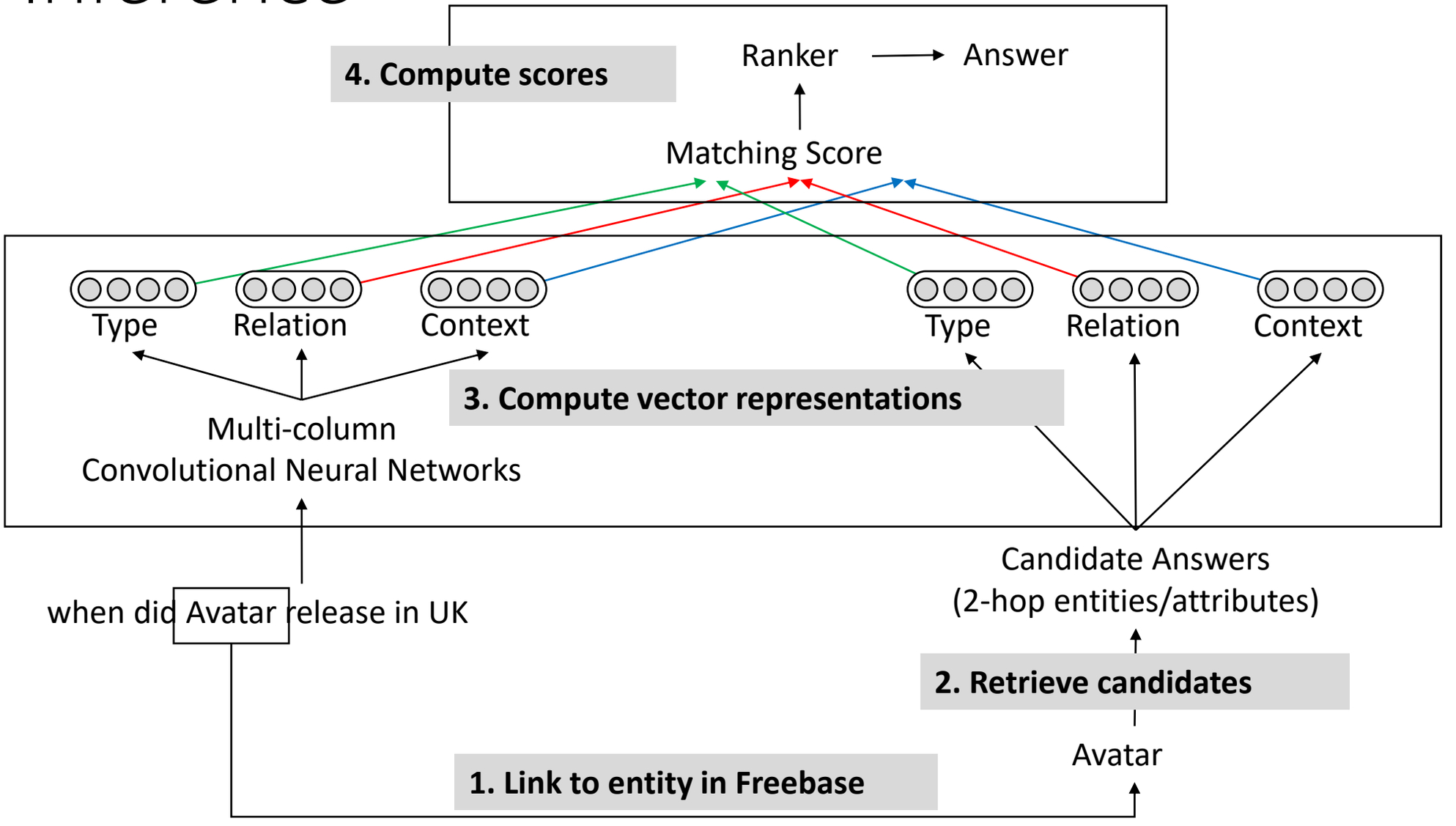


Multi-Column CNN Model

$$S(q,a) = \underbrace{\mathbf{f}_1(q)^{\mathsf{T}}\mathbf{g}_1(a)}_{answer\ path} + \underbrace{\mathbf{f}_2(q)^{\mathsf{T}}\mathbf{g}_2(a)}_{answer\ context} + \underbrace{\mathbf{f}_3(q)^{\mathsf{T}}\mathbf{g}_3(a)}_{answer\ type}$$



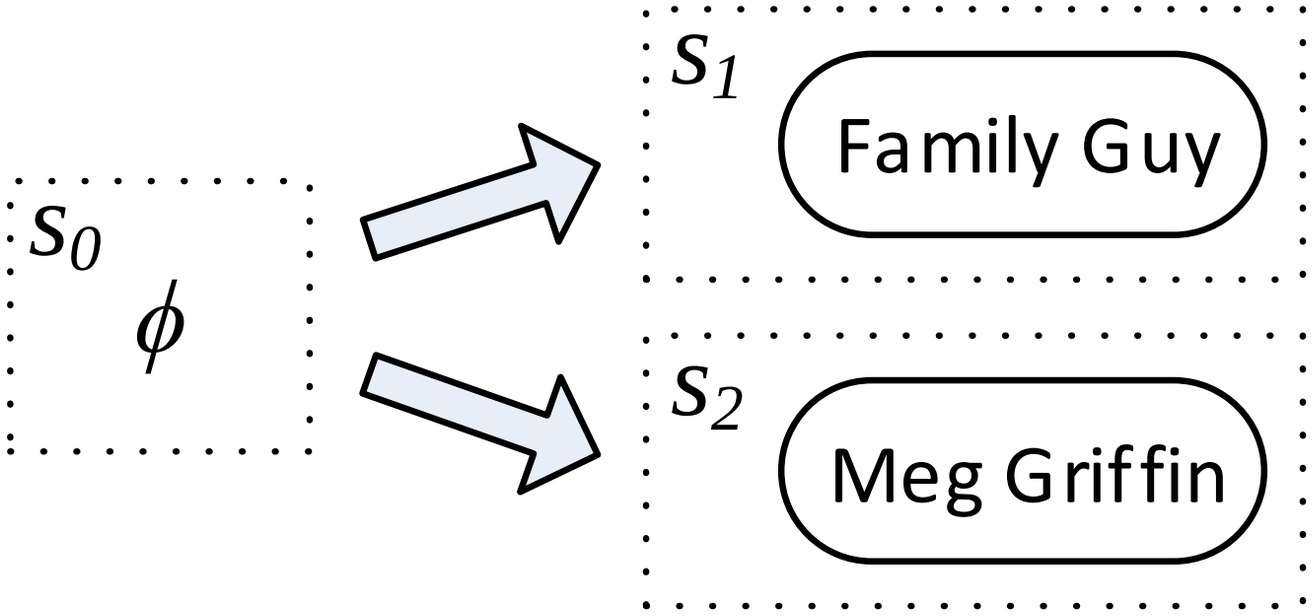
Inference


4. Compute scores
Ranker
Answer
Matching Score
Type
Relation
Context
Type
Relation
Context
3. Compute vector representations
Multi-column
Convolutional Neural Networks
when did Avatar release in UK
Candidate Answers
(2-hop entities/attributes)
2. Retrieve candidates
Avatar
1. Link to entity in Freebase

Query Graph Generation and Multi-Feature Ranking

(1) Link Topic Entity

Who first voiced Meg on Family Guy?

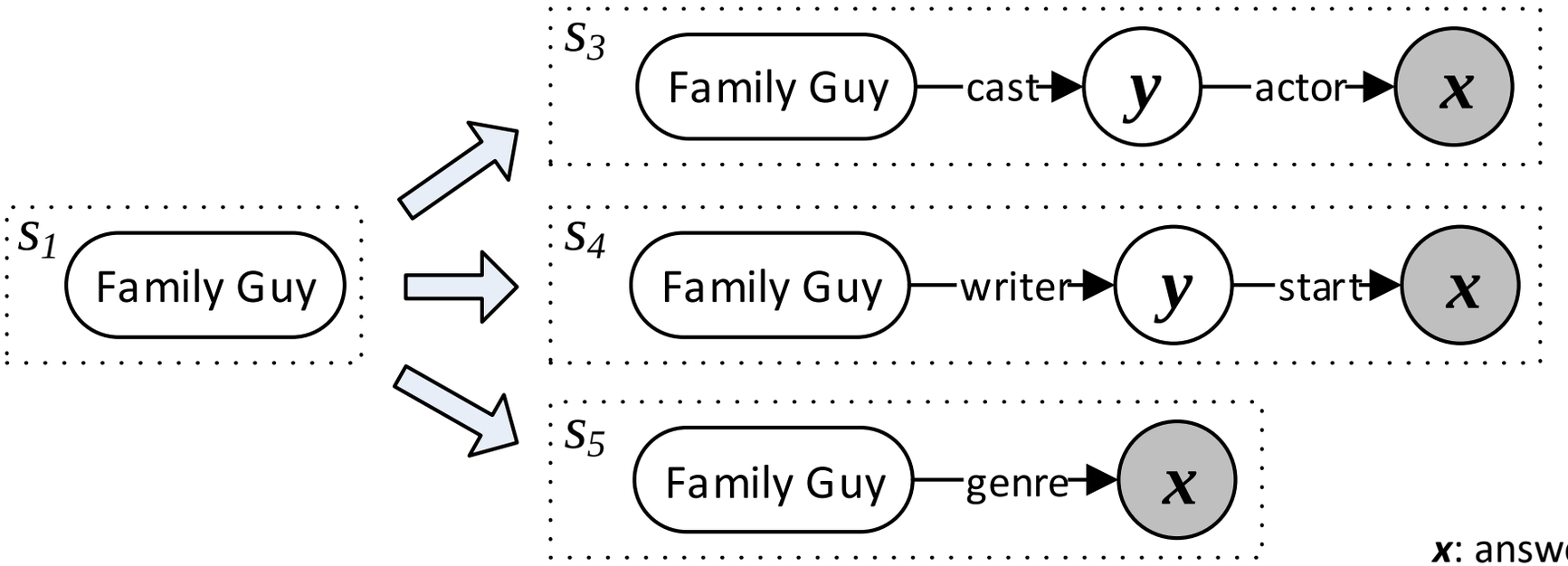


PPT available at
http://research.microsoft.com/apps/pubs/default.aspx?id=244749

Query Graph Generation and Multi-Feature Ranking

(2) Identify Core Inferential Chain

Who first voiced Meg on Family Guy?

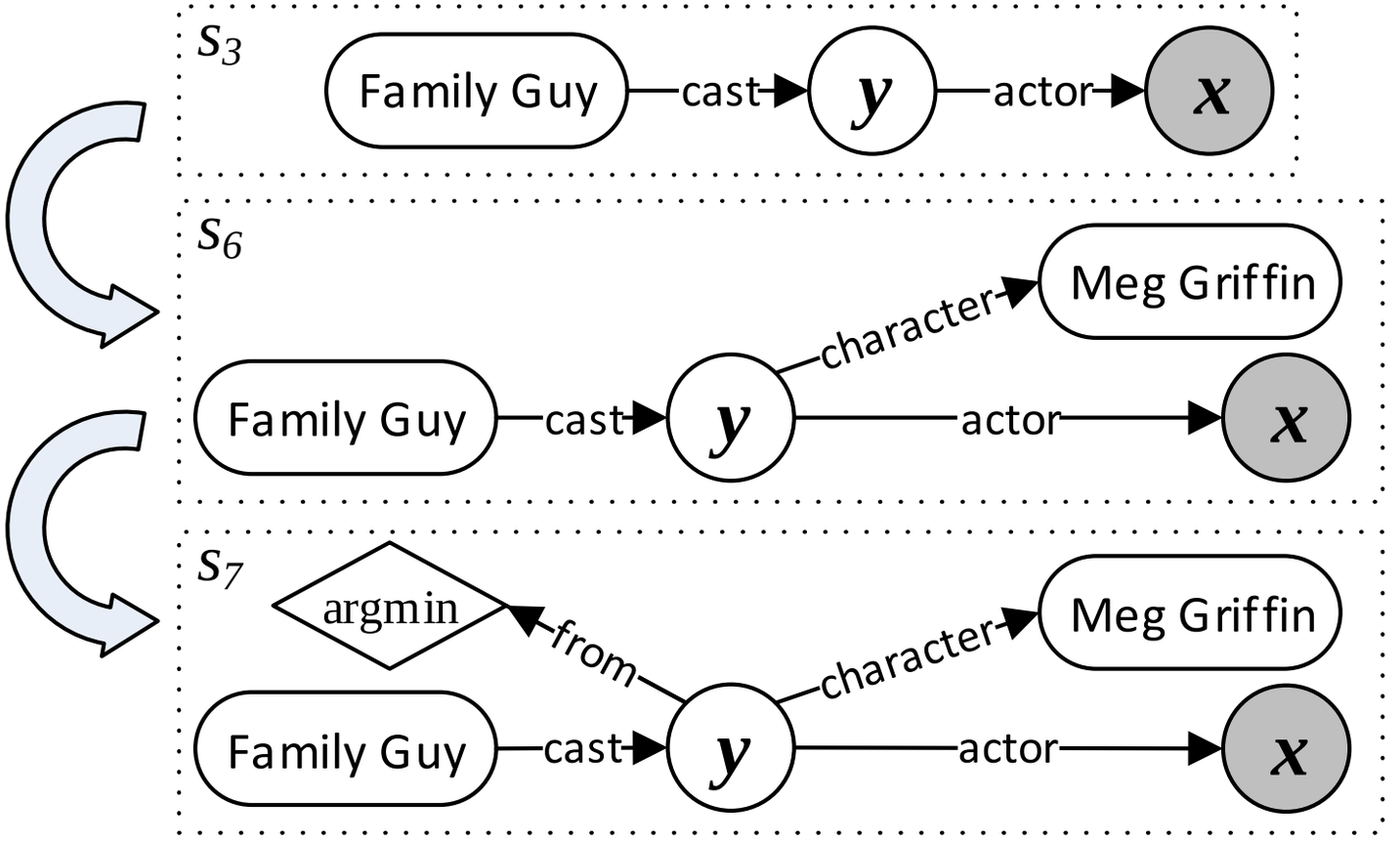


x: answer entity
y: CVT (compound value type), which is not a real-world entity, but is used to collect multiple fields of an event

Query Graph Generation and Multi-Feature Ranking

(3) Augment Constraints

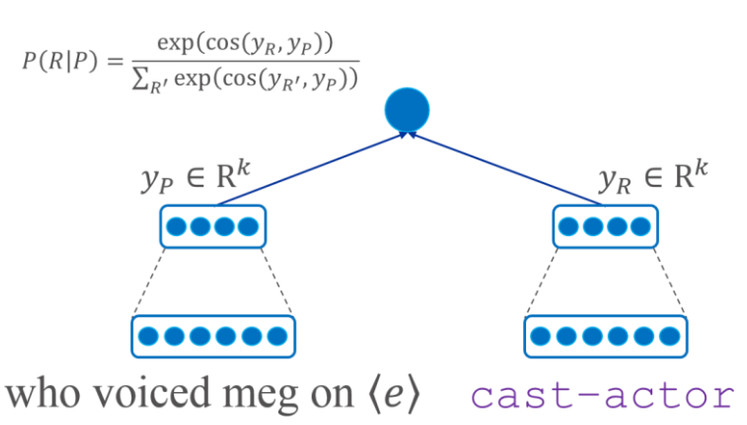
Who first voiced Meg on Family Guy?

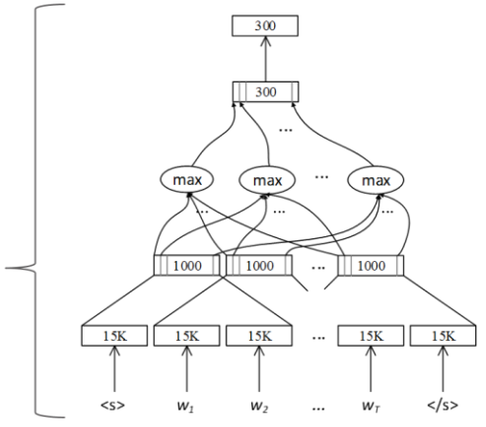


x: answer entity
y: CVT (compound value type), which is not a real-world entity, but is used to collect multiple fields of an event

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)

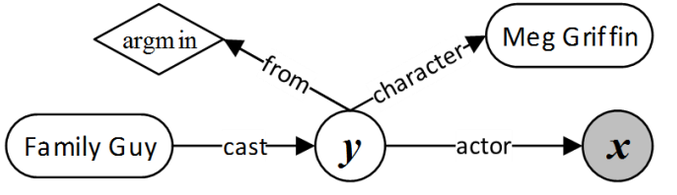




  - Constraints: Keyword and entity matching
    - ConstraintEntityWord("Meg Griffin", q) = 0.5
    - ConstraintEntityInQuestion("Meg Griffin", q) = 1
  - Overall
    - NumNodes(s) = 5
    - NumAnswers(s) = 1

q = Who first voiced Meg on Family Guy?

s =

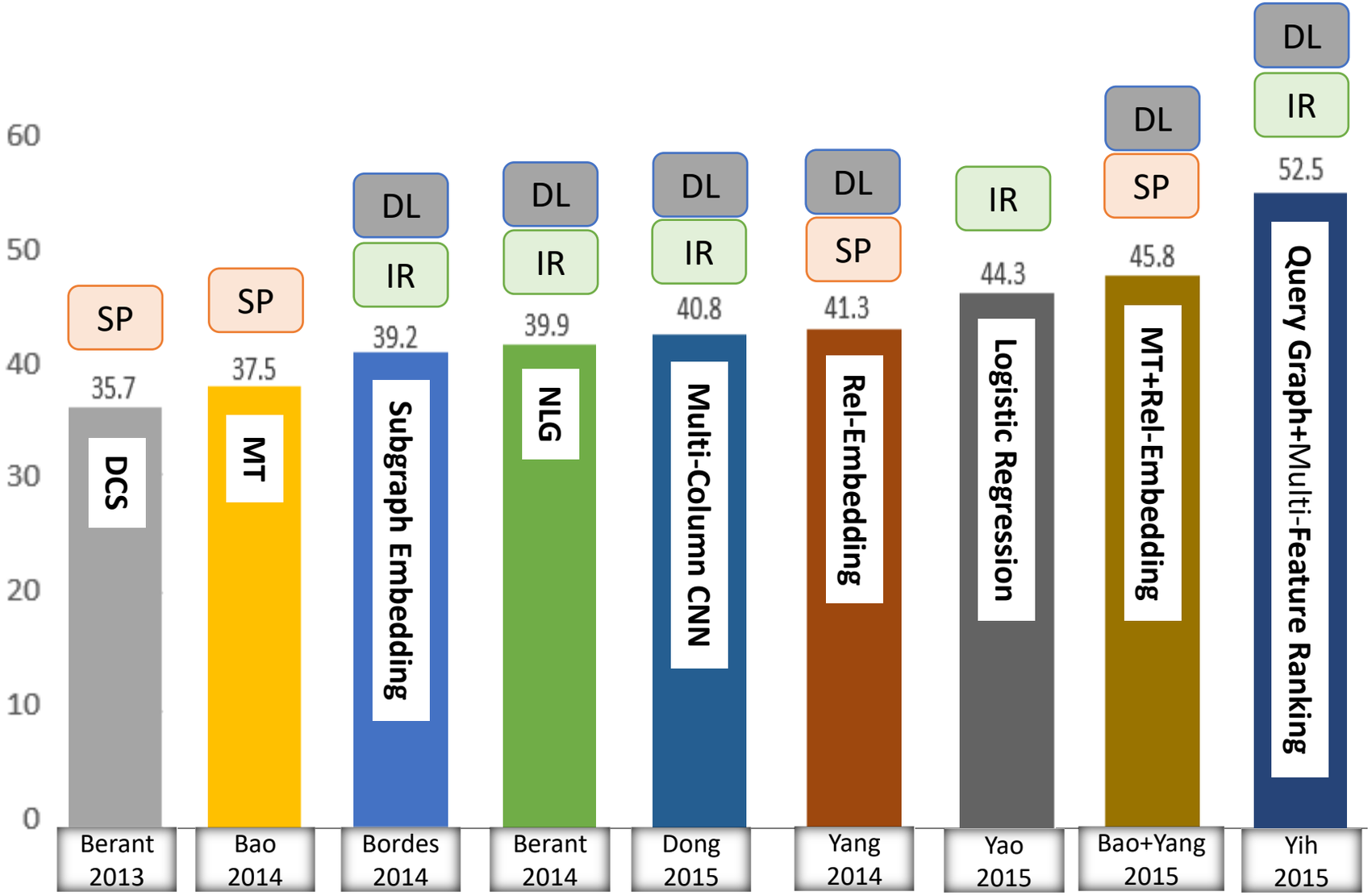


Benchmark

- **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

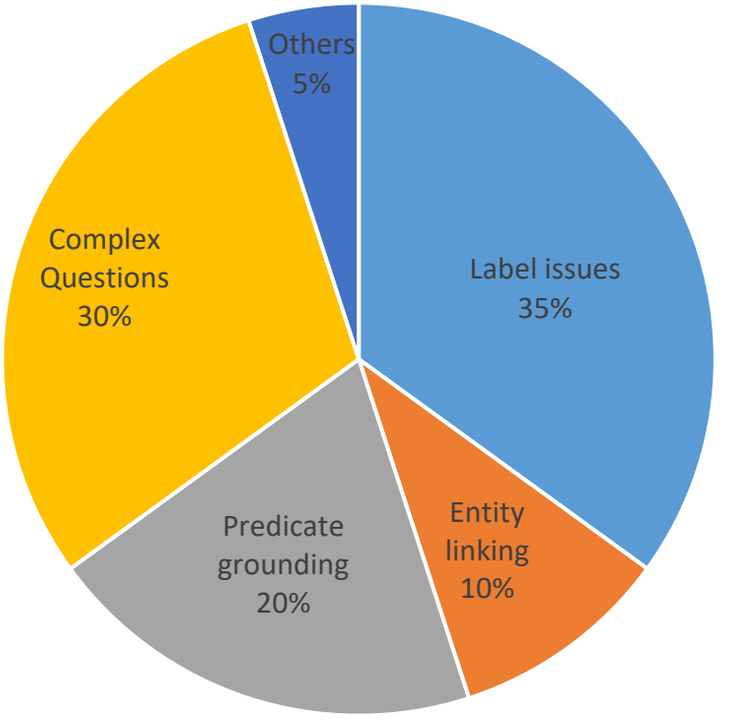
Evaluation Results on WebQuestions



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-constrains, 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

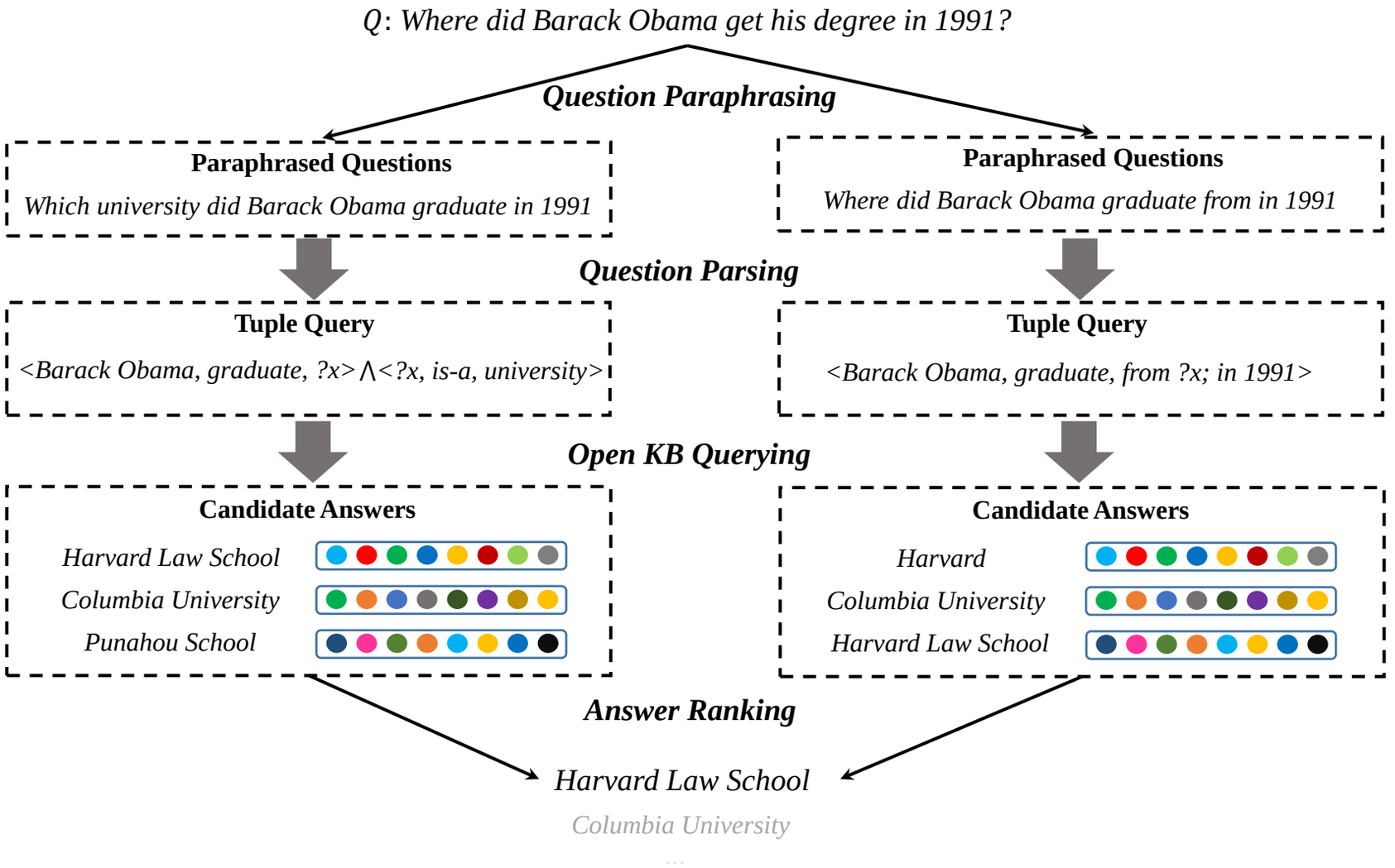
Subject	Relation	Arguments	Freq.	Conf.
James K. Polk	was	a governor; before he was president	2	0.87
the currency of Spain	was	the Peseta; before 2002	3	0.95
Peseta	was replaced	by Euro; as official tender of Spain; in 2002	3	0.81
Barack Obama	graduated	from Harvard Law School; in 1979 and 1991	4	0.77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

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)

Answering Questions with Complex Semantic Constraints on Open Knowledge Bases



	ComplexQuestions
Yin et al., 2015	**39.3%**
Berant et al., 2014	9.7%

$$acc = \frac{number\ of\ correctly\ answered\ questions}{total\ number\ of\ questions}$$

ComplexQuestions Dataset

- Data statistic
  - 300 English Q-A pairs
    - 80 questions from WebQuestions
    - 220 questions from Google query log
- Citation
  - Pengcheng Yin, Nan Duan, Ben Kao, Junwei Bao, Ming Zhou, *Answering Questions with Complex Semantic Constraints on Open Knowledge Bases*, CIKM, 2015
- Link
  - To be released
- An example
  - Q: what did Germany lost after the Treaty of Versailles?
  - A: Northern Schleswig; Alsace Lorraine

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*	↦	*Where did* **[None Phrase]** *graduate from*
Where did **[None Phrase]** *get his degree*	↦	*Which university did* **[None Phrase]** *graduate*
Where did **[None Phrase]** *get his degree*	↦	…

Where did Barack Obama graduate from in 1991?

Which university did Barack Obama graduate in 1991?

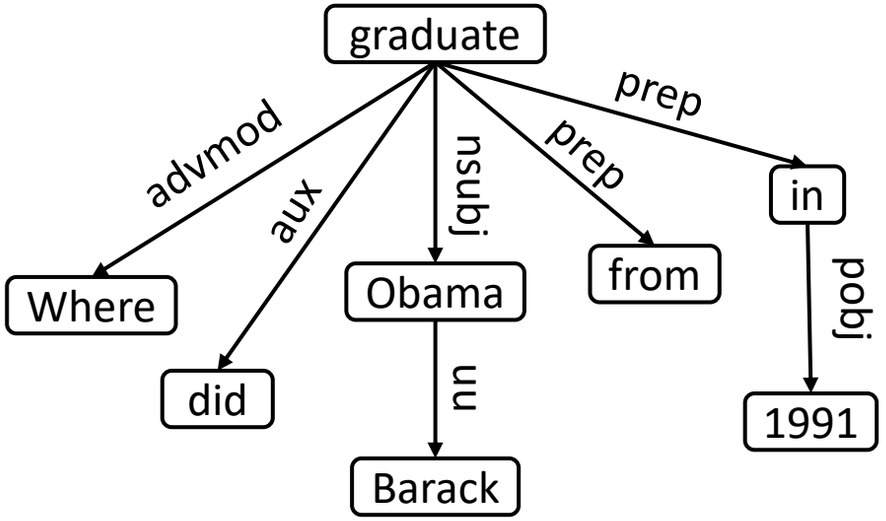
…

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?

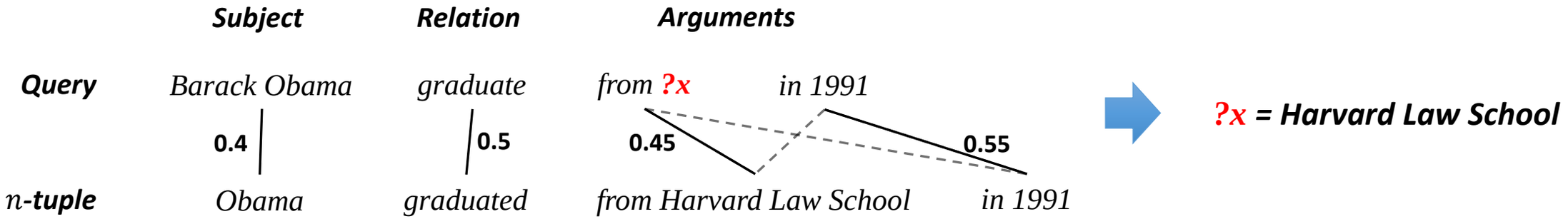


Dependency Parse Tree

<Barack Obama, graduate; from ?x; in 1991>

Open KB Querying

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



Define pairwise similarity between fields:

$$similarity(TQ.i, r.j) = \alpha \cdot text\_similarity(TQ.i, r.j) + (1-\alpha) \cdot pattern\_similarity(TQ.i, r.j)$$

Solve the optimization problem:

$$\max : \sum_i \sum_j x_{ij} \; similarity(TQ.i, r.j)$$

$$\text{subject to: } (x_{11} = 1), \quad (x_{22} = 1), \quad (x_{ij} \in \{0,1\})$$

$$(0 \le \sum_i x_{ij} \le 1), \quad (\sum_j x_{ij} = 1).$$

Answer Ranking

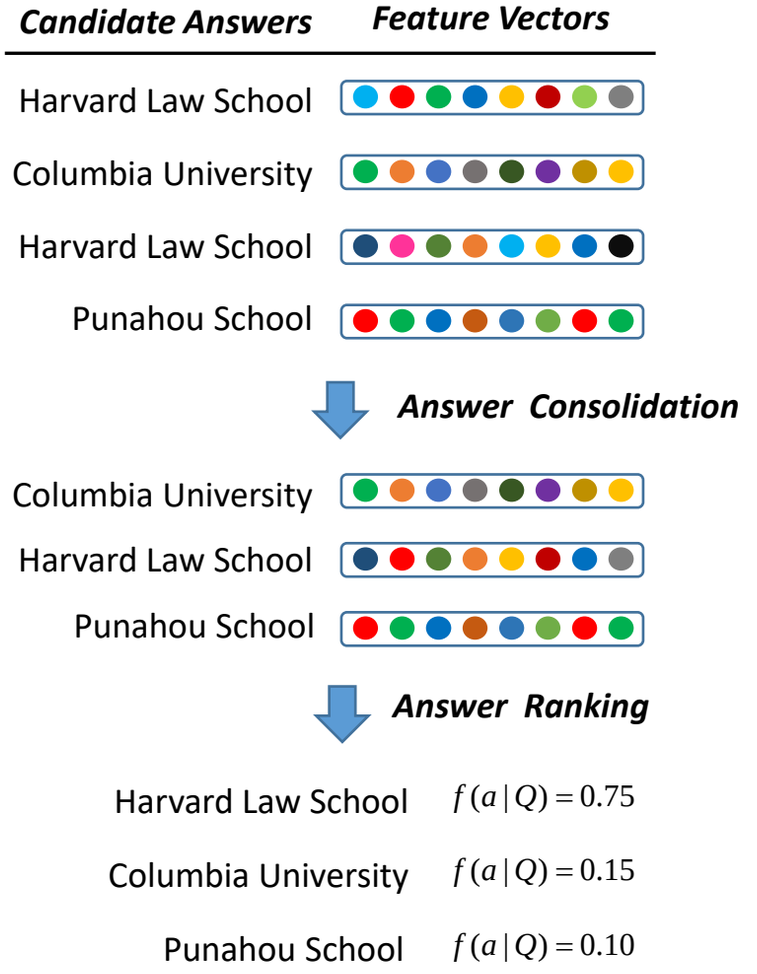
- Answer Consolidation
  - Merge the feature vectors of the answers with the same surface text form

- Ranking Model
  - Use 20K+ features to measure a candidate a is one answer of Q

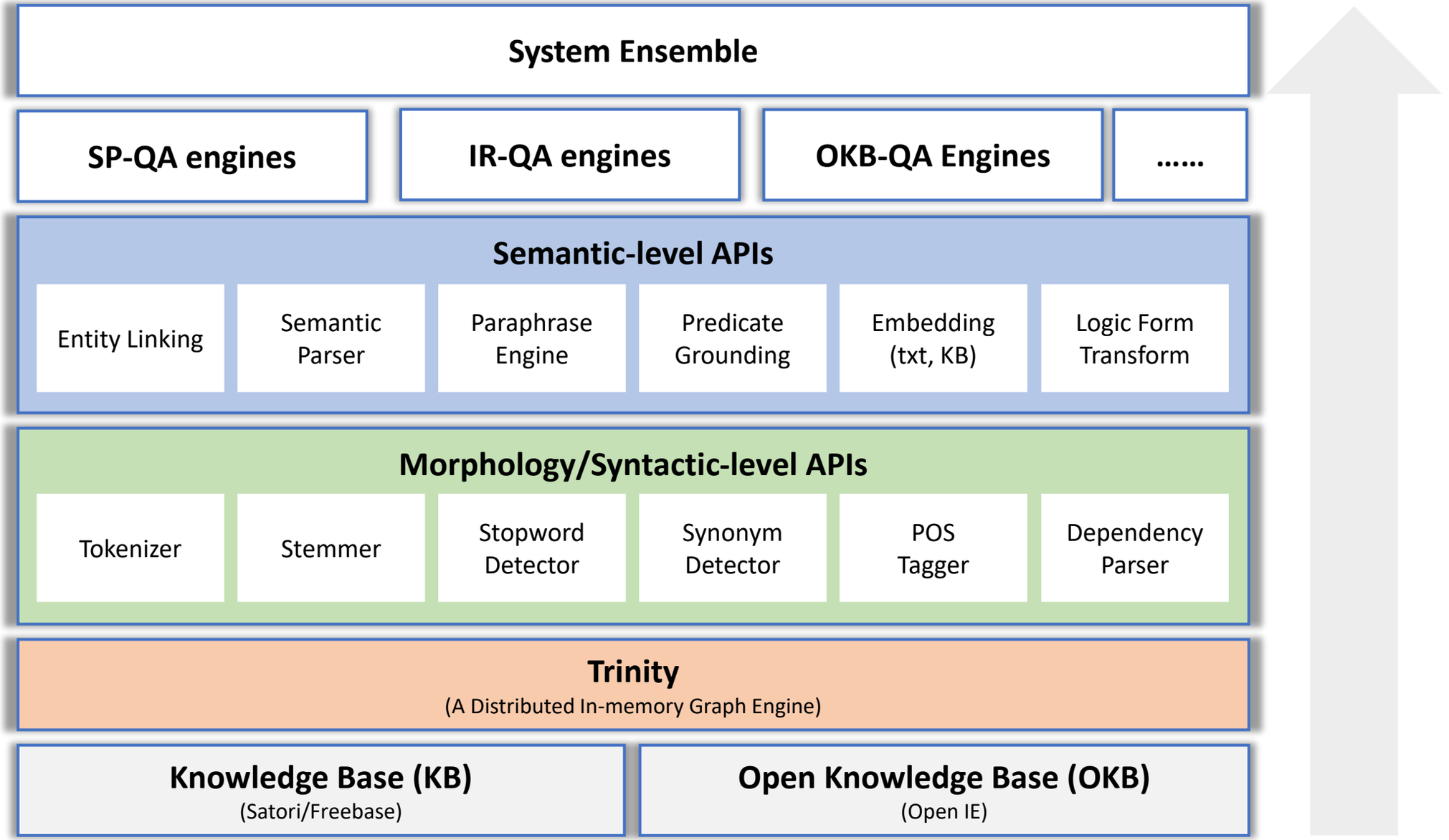
$$p(a|Q) = \frac{exp\{\sum_{k=1}^{M} \lambda_k \cdot f_k(a)\}}{\sum_{a' \in \mathcal{A}} exp\{\sum_{k=1}^{M} \lambda_k \cdot f_k(a')\}}$$

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

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



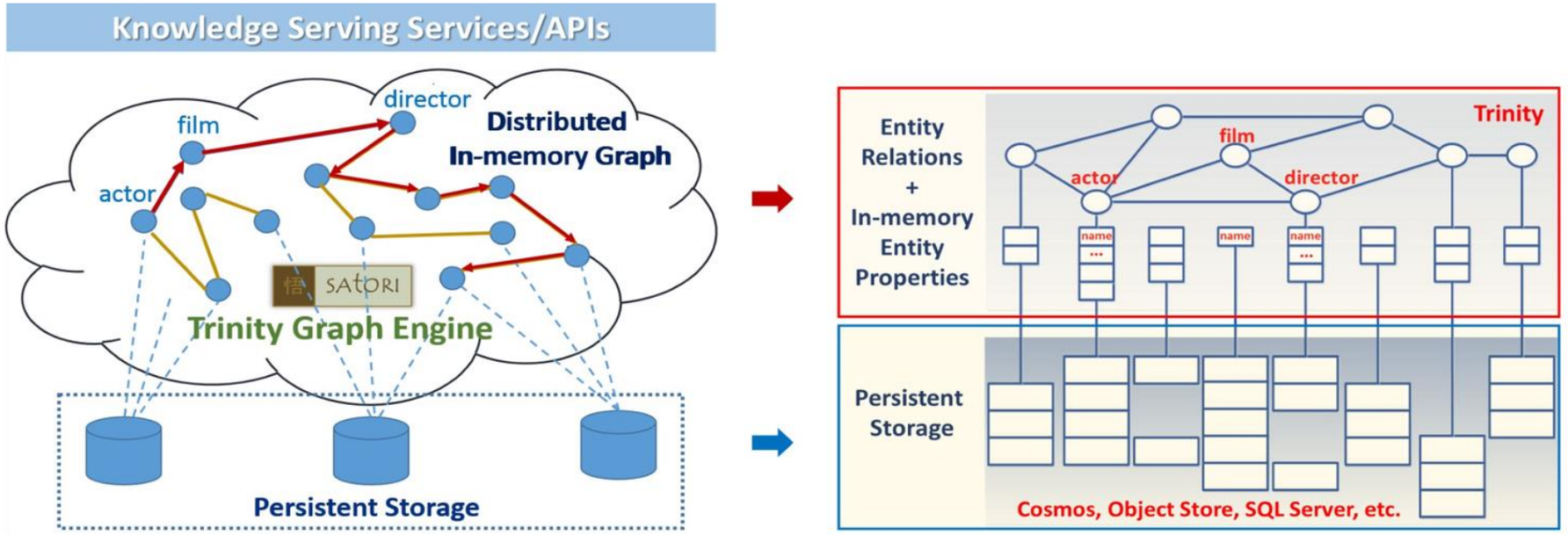
MSRA KB-QA Architecture



Bin Shao, Haixun Wang, Yatao Li, Trinity: ***A Distributed Graph Engine on a Memory Cloud***, SIGMOD, 2013

Trinity: Distributed In-memory Graph Engine

http://research.microsoft.com/graphengine



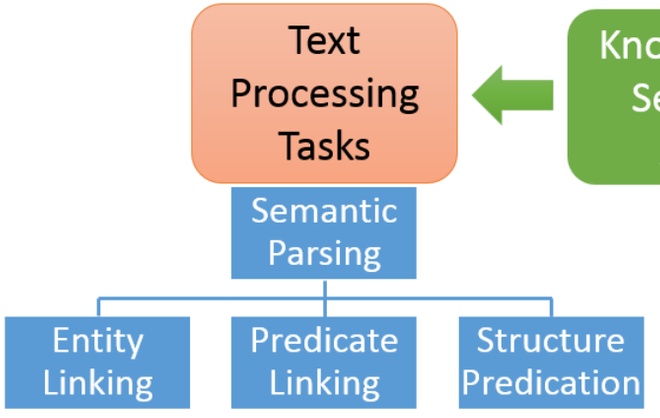
Fast Graph Traversal and Modularized Architecture

Deployment

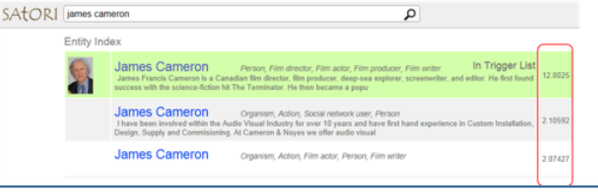
- **Fast building pipeline:** 6 hours to build from a cosmos Satori snapshot
- **Economical:** 16 machines for the whole Satori

Trinity Serving KB Access

Satori Access APIs
List<string[]> GetEntityIdByName (string entityName)
List<string> GetPredicatesByEntityId (string entityId)
List<string[]> GetValuesByEntityPredicate (string entityId, string predicate)
……



- 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

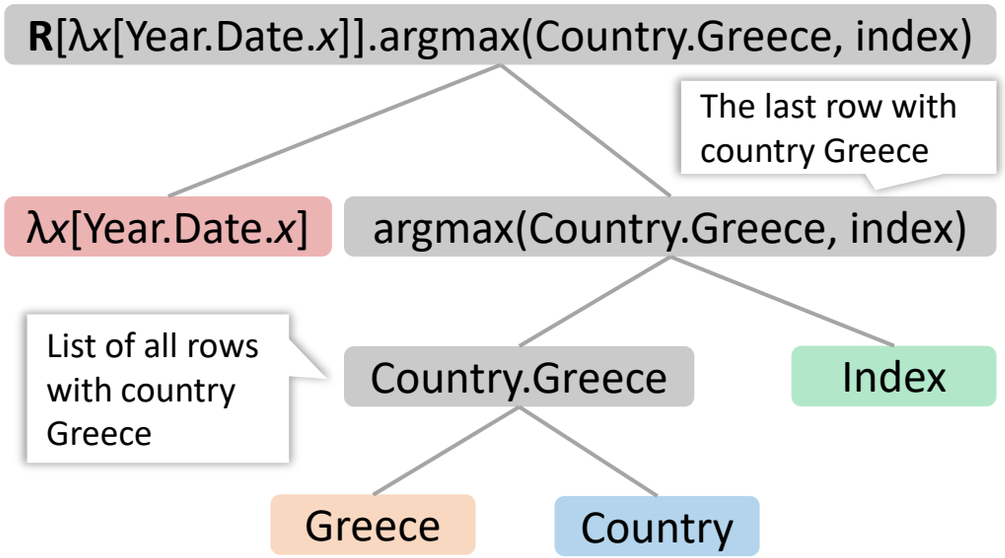
Table-based QA (TBQA)

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

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.

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

	City	Country	Continent	Summer	Winter	Year	Opening Ceremony	Closing Ceremony
1	Athens	Greece	Europe	I	—	1896	April 6	April 15
	Paris	France	Europe	II	—	1900	May 14	October 28
	St. Louis[a]	United States	North America	III	—	1904	July 1	November 23
	London[b]	United Kingdom	Europe	IV	—	1908	April 27	October 31
	Stockholm	Sweden	Europe	V	—	1912	May 5	July 22
	...							
28	Athens	Greece	Europe	XXVIII	—	2004	August 13	August 29
	Turin	Italy	Europe	—	XX	2006	February 10	February 26
	Beijing[e]	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[d]	—	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

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

how many person died in the Oklahoma city bombing?

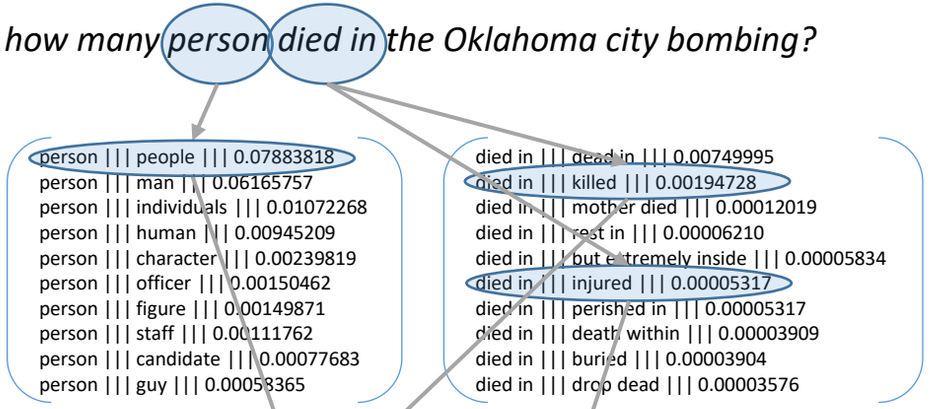
Oklahoma City bombing

From Wikipedia, the free encyclopedia
(Redirected from 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[1] and injured more than 680 others.[2] The blast destroyed or damaged 324 buildings within a 16-block radius, destroyed or burned 86 cars, and shattered glass in 258 nearby buildings,[3][4] causing an estimated $652 million worth of damage.[5] 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.[6][7]

Paraphrasing Model

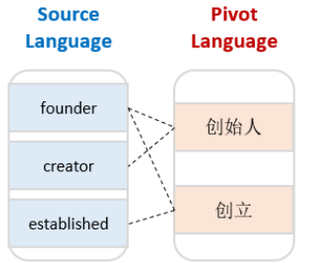
- An approach to match language pieces with different expressions but similar/identical semantic meanings



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 and injured more than 680 others.* The blast destroyed or damaged 324 buildings within a 16-block radius, destroyed or burned 86 cars, and shattered glass in 258 nearby buildings, causing an estimated $652 million worth of damage.

Paraphrasing Model

- Training data
  - SMT bilingual sentence pairs (Chinese-English)
- Training method
  - Pivot-based paraphrase extraction
    1. **43.8M** paraphrase pairs (English)
    2. **11.3M** paraphrase pairs (Chinese)
- Feature function
  - Weighted BLEU score

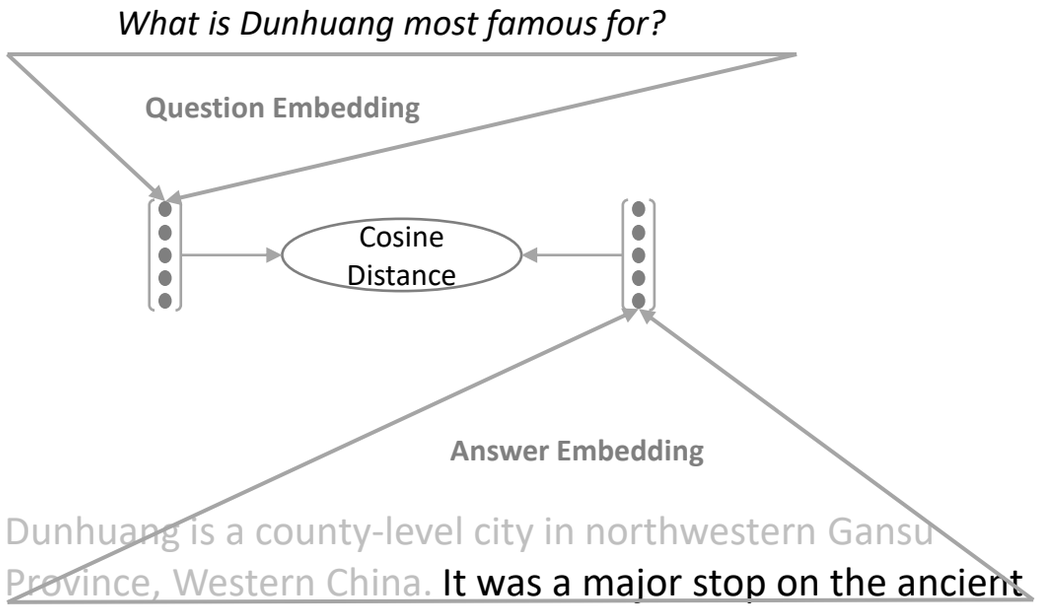


$$p(s_j|s_i) = \sum_t p(t|s_i) \cdot p(s_j|t)$$

founder ||| creator ||| 0.01214179
founder ||| founded ||| 0.00694428
founder ||| start ||| 0.00280628
founder ||| set up ||| 0.00088949
founder ||| established ||| 0.00065527
founder ||| pioneer ||| 0.00047020
...

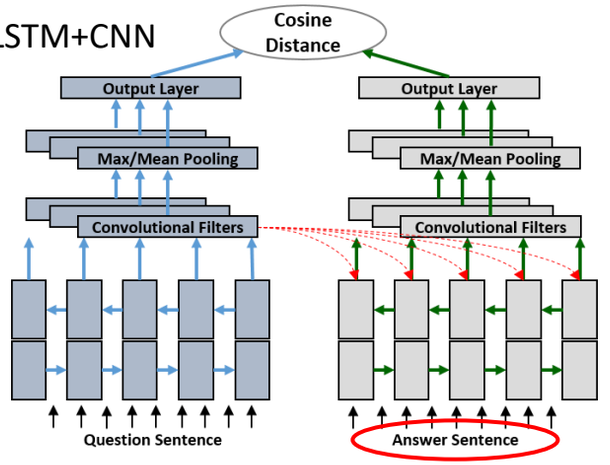
Answer Sentence Selection Model

- A deep learning approach to *address question queries*



Answer Sentence Selection Model

- Training data
  - **10M** <question, answer> pairs from WikiAnswers (English)
  - **10M** <question, answer> pairs from Baidu Zhidao (Chinese)
- Training method
  - Attention-based LSTM+CNN



- Feature function
  - Cosine distance

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

- *Each entity in Wikipedia has a content table.*
- *Each content table item is a topic of the given entity.*
- *Each content table item has a corresponding paragraph.*

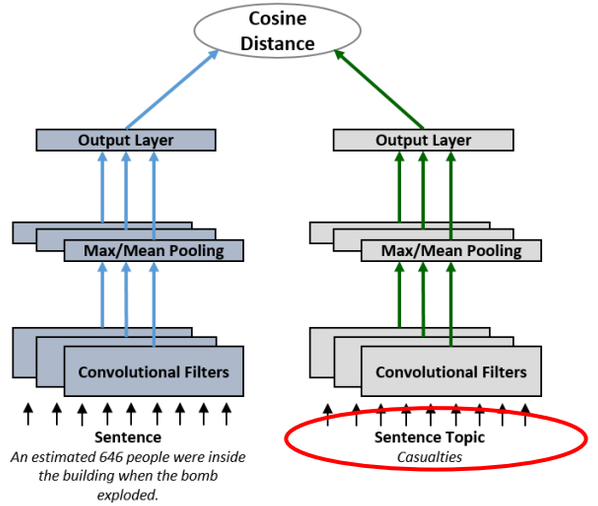
training instance

History [edit]

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[who?] have

Topic Embedding Model

- Training data
  - **1M** <sentence, sentence topic> pairs (English)
  - **1M** <sentence, sentence topic> pairs (Chinese)
- Training method
  - CNN
- Feature function
  - Cosine distance



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% (state-of-the-art result)

  - 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

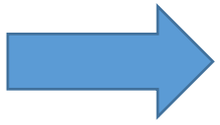
- KB-based Conversation
  - JD XiaoIce (京东小冰)
  - CEAir XiaoIce (东航小冰)
- Document-based Conversation
  - Dunhuang XiaoIce (敦煌小冰)

JD XiaoIce (京东小冰)

- Platform
  - JD App
- Function
  - ChitChat
  - Product Card
  - Product Chat


NOKIA
●●●○○ 中国联通 上午9:26 55%
小冰(京东)
你好
你好，我是微软小冰，史上最强人工智能聊天机器人。我爱卖萌，爱耍小聪明，爱大闹企鹅村。和我聊的越多，我越聪明哦~.有好多独家秘技，等你探索哪~
明天天气如何
听我老爸说...北京明天阴转多云20°C~33°C。微风小于3级O(∩_∩)O~伴着清风晨练棒棒哒，如果你能早起的话。。。心飞扬~
发送
NOKIA
●●●○○ 中国联通 上午9:26 55%
小冰(京东)
我好无聊，陪我聊天吧
出来聊天?
懒得出来，你讲个笑话吧
说笑话说的舌头都抽筋了，要不调戏调戏还猪格格去?[yeah]~~~ 还猪格格向学霸室友请教如何能成为一个学习天才，室友拍了拍还猪格格的肩说“哪里有什么天才，我只是把你睡觉的时间用来学习而已。”还猪格格听完之后才恍然大悟怪不得老子总是觉得睡觉时间不够，原来是被这贱人偷偷拿去用来学习了!
发送


NOKIA
●●●○○ 中国联通 上午9:26 55%
小冰(京东)
哈哈，你真有意思
那里有意思拉
你们这卖东西么
论吨卖
那你给我推荐个手机吧
(⊙v⊙)这个觉得好不?
发送




NOKIA
●●●○○ 中国联通 上午9:27 55%
小冰(京东)
联想 乐檬 K3 Note（K50-t5）16G 珍珠白 移动联通4G...
￥899.00
第二次购买了 性价比不错
要苹果的
(⊙v⊙)这个觉得好不?
Apple iPhone 6 Plus (A1524) 16GB 金色 移动联通电信4G手机
￥5583.00
发送


●●●●○ 中国联通 上午9:33 57%
商品详情
Apple iPhone 5s (A1530) 16GB 金色 移动联通4G手机 自营
先进的 iPhone 5s，满载出色功能和创新技术，现更值得拥有。
￥3588.00
降价通知
关注
购物车
加入购物车


NOKIA
●●●○○ 中国联通 上午9:32 57%
小冰(京东)
现在买很划算啊
在国内买好怕买到假货啊泪目不知道什么时候才会再有折扣
有合适的手机壳么
某问题呀，这个肿么样。
【京东自营】dostyle PC303 苹果iphone6 4.7寸 手机壳/手机...
￥49.00
发送


NOKIA
●●●●○ 中国联通 上午9:42 61%
小冰(京东)
对了，京东最近有啥优惠活动不吗
真会问，问到专家了~京东上优惠信息我最知道了，到这里逛逛吧，样样物美价廉…… http://sq.jd.com/BuT3rK
看了看，挺多商品的，你快成售货员了
不过挺好吃的
行了，不和你说了，回头再找你聊天吧
我的小冰
常用技能
薅羊毛


sale.jd.com
抢！天天9块9
今日专享价 就比PC省
限时9块9
永安康健 Kefir克菲尔益生菌 成人 开菲尔颗粒7袋装
￥29.00
卖光啦
18渡铁观音 乌龙茶清香型 安溪铁观音250g 八马茶业
￥19.90
卖光啦
喜龙牌三叶减肥茶2g*30

Dunhuang XiaoIce (敦煌小冰)

- Platform
  - WeChat
- Function
  - ChitChat
  - DocCard
  - DocChat


NOKIA
中国联通 下午12:35 100%
微信(2)
敦煌研究院
下午12:30
莫高窟为啥叫莫高窟啊
一说，在此初凿石窟的乐僔，道行高超，其余僧众“莫高于此僧”；一说，是在汉代之前的游牧民族的语言遗留下的音；一说，古代敦煌莫高窟所在的山也叫“漠高山”，所以开凿的石窟也叫“莫高窟”..
莫高窟在哪个省
甘肃省


中国联通 下午12:48 100%
微信(1)
小冰
下午12:45
莫高窟为啥叫莫高窟
位于甘肃省酒泉敦煌
莫高窟在哪个省
没有去看
莫高窟有多少个洞窟
4千多
莫高窟彩塑真漂亮啊


NOKIA
中国联通 下午12:36 100%
微信(2)
敦煌研究院
莫高窟有多少洞窟
总共有735个，其中有壁画、塑像等文物信息的是492个。
莫高窟彩塑真漂亮啊
~→_→
真想吃一碗敦煌的驴肉黄面
驴肉黄面 黄面也是一种麦面制品。

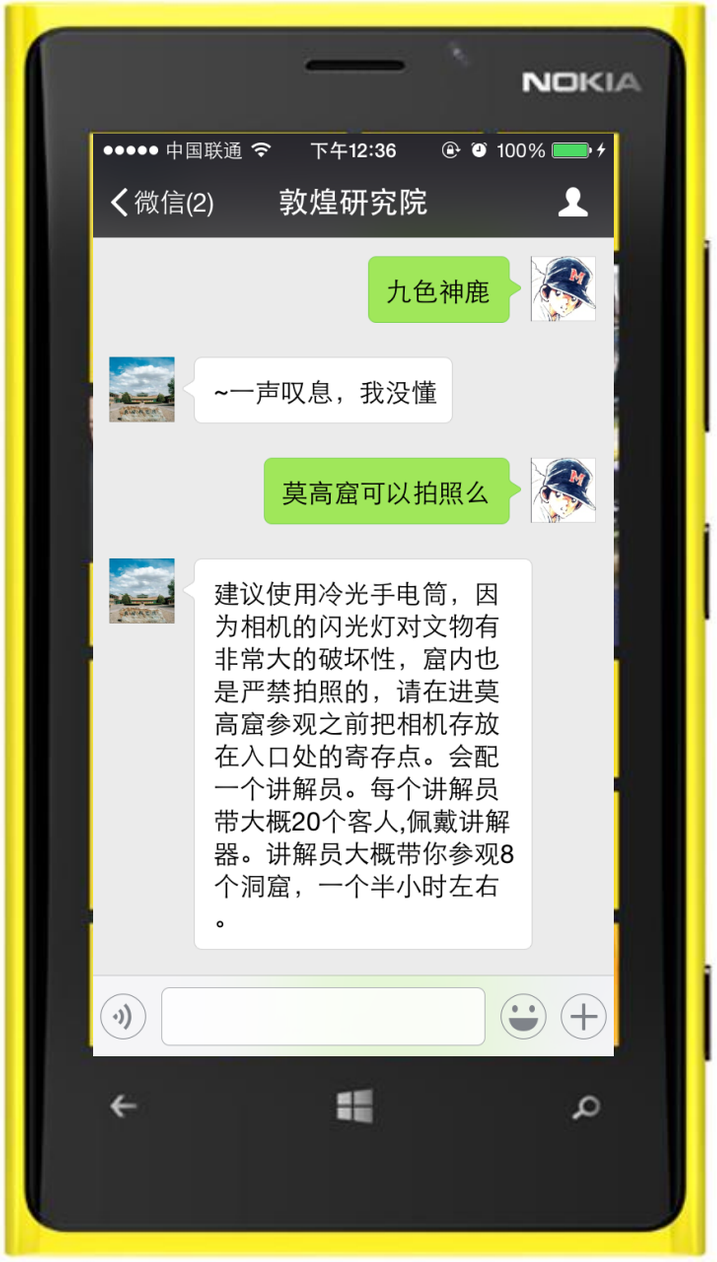

中国联通 下午12:48 100%
微信(1)
小冰
莫高窟有多少个洞窟
4千多
莫高窟彩塑真漂亮啊
鸣沙山
真想吃一碗敦煌的驴肉黄面
我们想到一起去了！
谁发现了藏经洞


NOKIA
中国联通 下午12:36 100%
微信(2) 敦煌研究院
谁发现了藏经洞
~哎呦，你好像懂得挺多。。。
飞天是啥意思
飞天,意为飞舞的天人.飞天多画在佛教石窟壁画中.飞天是歌神乾闼婆和乐神紧那罗的化身,原是古印度神话中的歌舞神和娱乐神,他们是一对夫妻,后被佛教吸收为天龙八部众神之内.
敦煌有啥小吃


中国联通 下午12:48 100%
微信(1) 小冰
谁发现了藏经洞
发现你了
飞天是啥意思
精辟
敦煌有啥小吃
小吃是什么
什么时候去敦煌旅游最合适

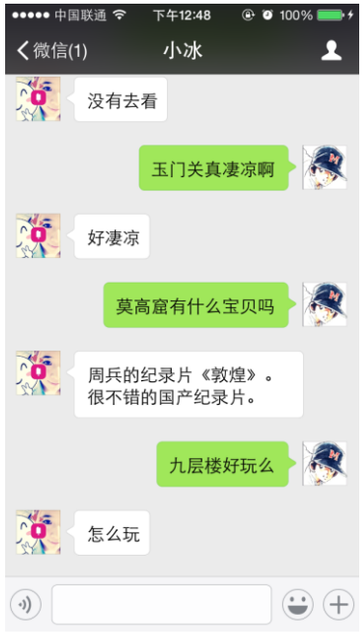

NOKIA
中国联通 下午12:36 100%
微信(2) 敦煌研究院
敦煌有啥小吃
驴肉黄面驴肉黄面是敦煌八大怪之四,敦煌顺张驴肉黄面馆,在文庙巷第84号,2005年夏,著名华裔物理学家李政道在敦煌旅游时,曾经来光临过顺张黄面馆.敦煌最著名的驴肉黄面馆是达记驴肉黄面馆.
什么时候去敦煌旅游最合适
~呵呵
九色神鹿


中国联通 下午12:48 100%
微信(1) 小冰
敦煌有啥小吃
小吃是什么
什么时候去敦煌旅游最合适
8-9月去看塔里木胡杨林最合适
九色神鹿
不理你
莫高窟可以拍照么


NOKIA
中国联通 下午12:36 100%
微信(2) 敦煌研究院
九色神鹿
~一声叹息，我没懂
莫高窟可以拍照么
建议使用冷光手电筒，因为相机的闪光灯对文物有非常大的破坏性，窟内也是严禁拍照的，请在进莫高窟参观之前把相机存放在入口处的寄存点。会配一个讲解员。每个讲解员带大概20个客人,佩戴讲解器。讲解员大概带你参观8个洞窟，一个半小时左右。


中国联通 下午12:48 100%
微信(1) 小冰
九色神鹿
不理你
莫高窟可以拍照么
没有去看
玉门关真凄凉啊
好凄凉
莫高窟有什么宝贝吗


NOKIA
中国联通 下午12:36 100%
微信(2) 敦煌研究院
玉门关真凄凉啊
嗯啊.
莫高窟有什么宝贝吗
敦煌莫高窟是我国著名的四大石窟之一,也是世界上现存规模最宏大,保存最完好的佛教艺术宝库。
九层楼好玩么
好玩啊


中国联通 下午12:48 100%
微信(1) 小冰
没有去看
玉门关真凄凉啊
好凄凉
莫高窟有什么宝贝吗
周兵的纪录片《敦煌》。很不错的国产纪录片。
九层楼好玩么
怎么玩

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

Red-colored papers are from MSR