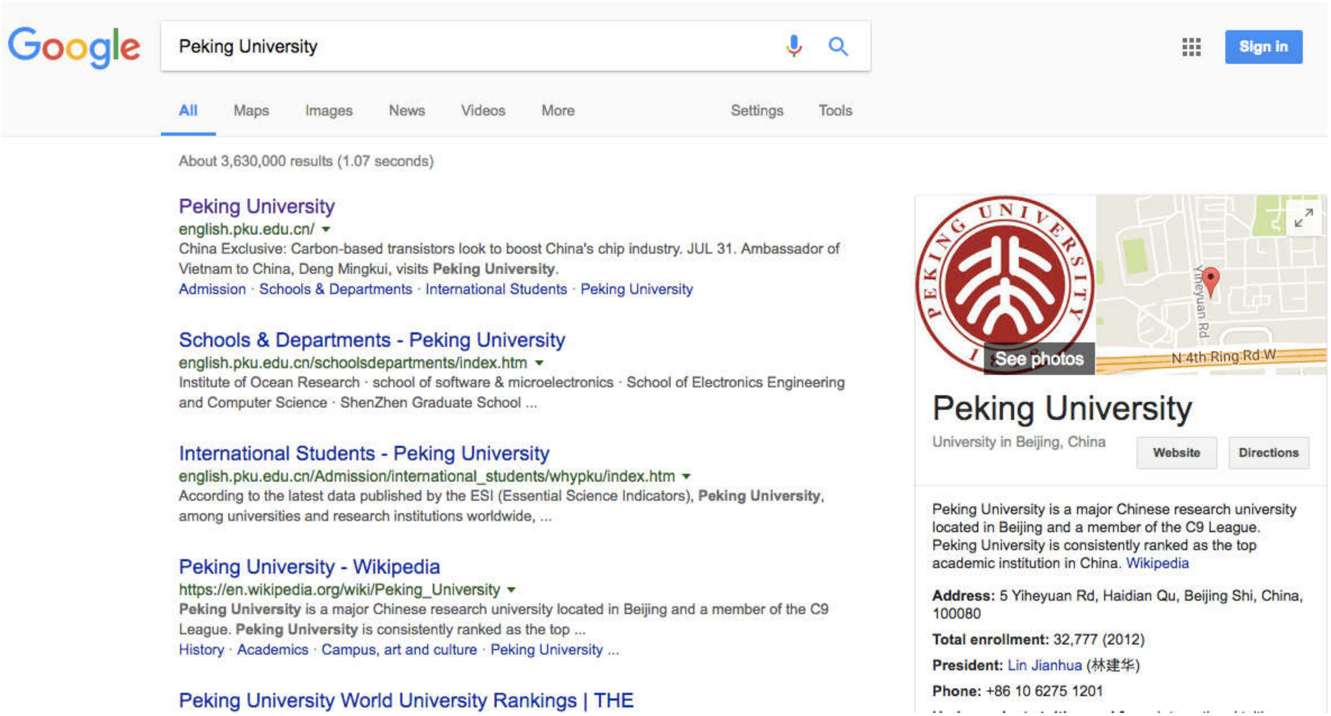
Question Answering Over Knowledge Graph

Lei Zou

Knowledge Graph

Google launches **Knowledge Graph** project at 2012.



About 3,630,000 results (1.07 seconds)

Peking University
english.pku.edu.cn/
China Exclusive: Carbon-based transistors look to boost China's chip industry. JUL 31. Ambassador of Vietnam to China, Deng Mingkui, visits **Peking University**.
Admission · Schools & Departments · International Students · Peking University

Schools & Departments - Peking University
english.pku.edu.cn/schoolsdepartments/index.htm
Institute of Ocean Research · school of software & microelectronics · School of Electronics Engineering and Computer Science · ShenZhen Graduate School ...

International Students - Peking University
english.pku.edu.cn/Admission/international_students/whypku/index.htm
According to the latest data published by the ESI (Essential Science Indicators), **Peking University**, among universities and research institutions worldwide, ...

Peking University - Wikipedia
https://en.wikipedia.org/wiki/Peking_University
Peking University is a major Chinese research university located in Beijing and a member of the C9 League. **Peking University** is consistently ranked as the top ...
History · Academics · Campus, art and culture · Peking University ...

Peking University World University Rankings | THE

Peking University
University in Beijing, China

Peking University is a major Chinese research university located in Beijing and a member of the C9 League. Peking University is consistently ranked as the top academic institution in China. Wikipedia

Address: 5 Yiheyuan Rd, Haidian Qu, Beijing Shi, China, 100080

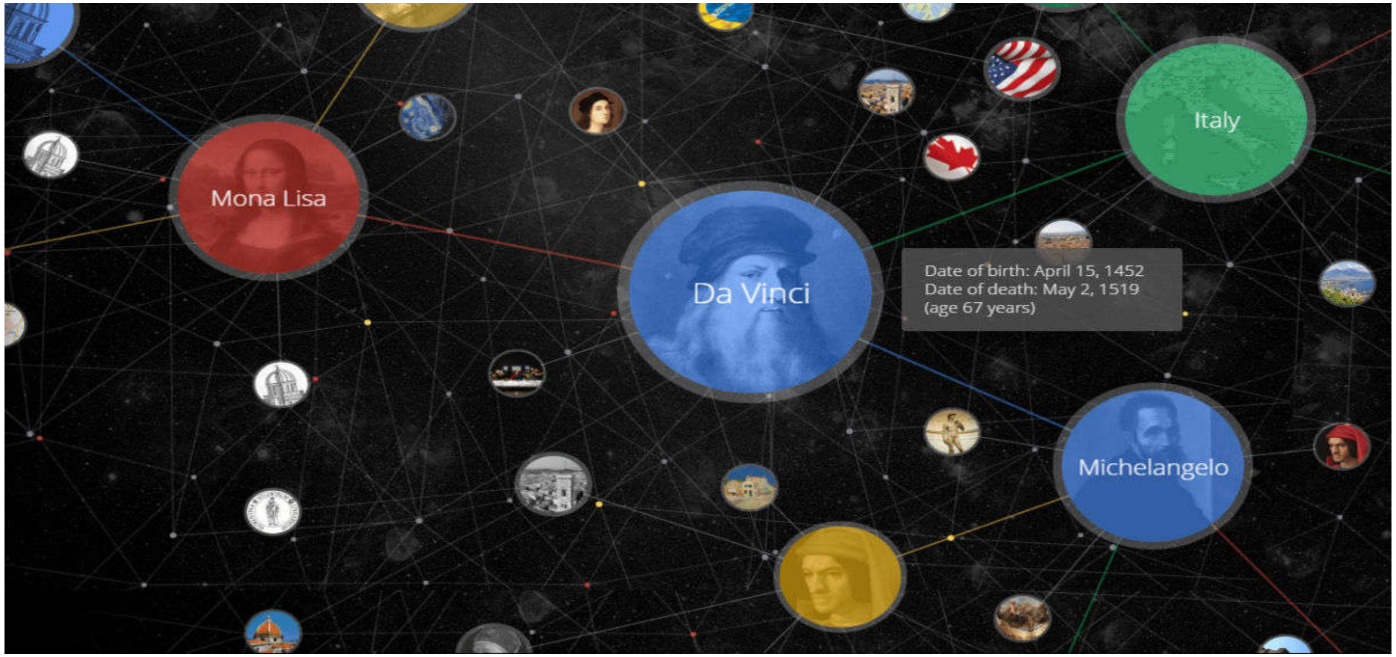
Total enrollment: 32,777 (2012)

President: Lin Jianhua (林建华)

Phone: +86 10 6275 1201

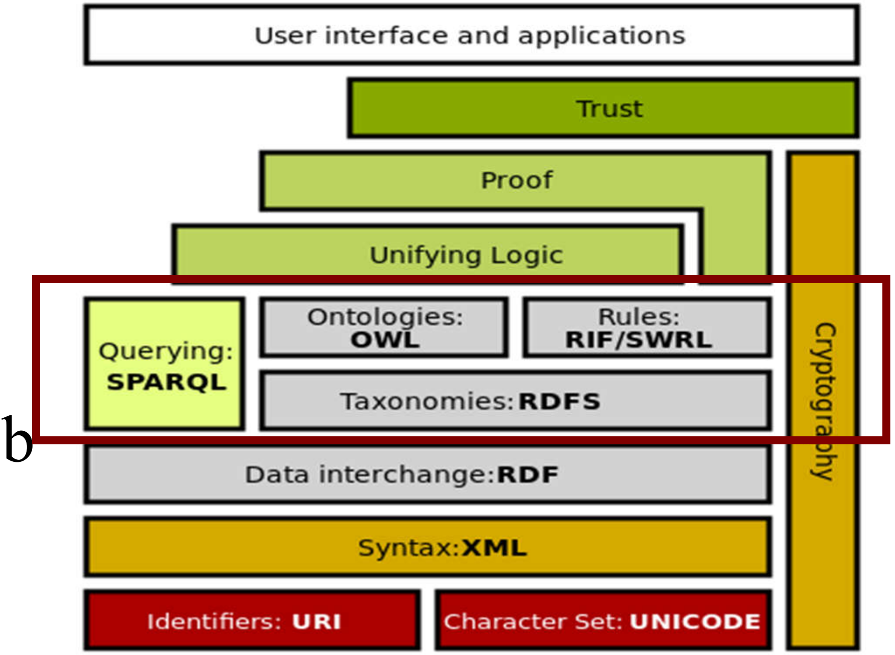
Knowledge Graph

Essentially, KG is a sematic network, which models **the entities (including properties) and the relation between each other.**



Resource Description Framework (RDF)

- RDF is an **de facto standard** for Knowledge Graph (KG).
- RDF is a **language** for the conceptual modeling of information about web resources
- A **building block** of semantic web
- Make the information on the web and the interrelationships among them "**Machine Understandable**"



RDF & SPARQL

RDF Datasets

Subject	Predicate	Object
Resident_Evil:_Retribution	type	film
Resident_Evil:_Retribution	budget	“6.5E7"
Resident_Evil:_Retribution	director	Paul_W._S._Anderson
Paul_W._S._Anderson	type	director
Resident_Evil	director	Paul_W._S._Anderson
Paul_Anderson_(actor)	type	actor
The_Revenant	strarring	Philadelphia
Priestley Medal	awards	Paul S. Anderson
Maclovia_(1948_film)	distributor	Filmex

“What is the budget of the film directed by Paul Anderson ?.”

SPARQL

```
SELECT ?y WHERE
{
?x director Paul_W._S._Anderson .
?x type film .
?x budget ?y.
}
```

Interdisciplinary Research

Database

RDF Database

Data Integration 、 Knowledge Fusion

Natural Language Processing

Information Extraction

Semantic Parsing

Machine Learning

Knowledge Representation (Graph Embedding)

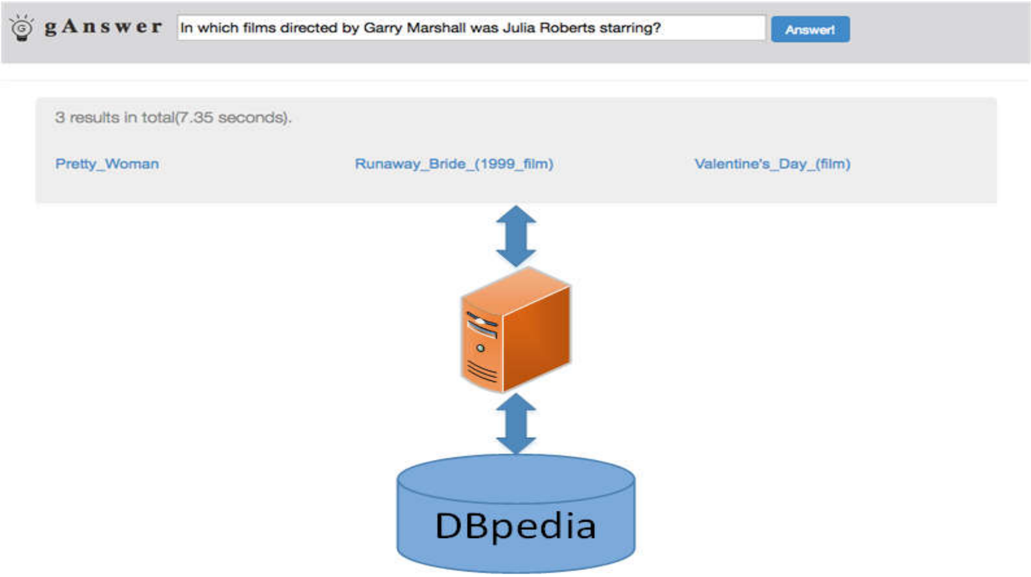
Knowledge Engineering

KB construction

Rule-based Reasoning

KG-based Question/Answering

- SPARQL syntax are too complex for ordinary users
- RDF KG is“schema-less”data, not like schema-first relational database.



In which films directed by Garry Marshall was Julia Roberts starring?

3 results in total(7.35 seconds).

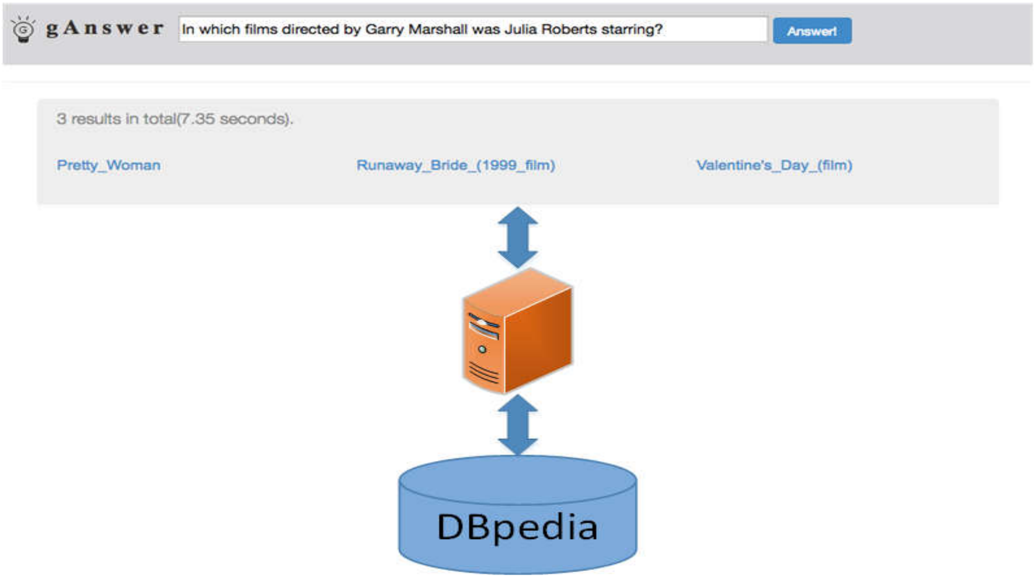
Pretty_Woman

Runaway_Bride_(1999_film)

Valentine's_Day_(film)

KG-based Question/Answering

- An **Easy-to-Use** Interface to Access Knowledge Graph
- It is interesting to both **academia** and **industry**.
- **Interdisciplinary research** between database and NLP (natural language processing) communities.



In which films directed by Garry Marshall was Julia Roberts starring?

3 results in total(7.35 seconds).

Pretty_Woman

Runaway_Bride_(1999_film)

Valentine's_Day_(film)

KG-based Question/Answering

COMMENT

GENETICS Growth in genome screening could cause dangerous meddling p.27

EVOLUTION How genes and culture have shaped our ability to cooperate p.29

CHEMISTRY Debating how life got going on the early Earth p.30

EXHIBITION Wildlife paintings from Yukon to Yellowstone p.32

Search needs a shake-up

On the twentieth anniversary of the World Wide Web's public release, **Oren Etzioni** calls on researchers to think outside the keyword box and improve Internet trawling.

Two decades after Internet pioneer Tim Berners-Lee introduced his World Wide Web project to the world using the alt.hypertext newsgroup, web search is on the cusp of a profound change — from simple document retrieval to question answering. Instead of poring over long lists of documents that contain requested keywords, users need direct answers to their questions. With sufficient scientific and financial investment, we could soon view today's keyword searching with the same nostalgia and amusement reserved for bygone technologies such as electric typewriters and vinyl records.

But this transformation could be unreasonably delayed. As a community, computer scientists have underinvested in tools that can synthesize sophisticated answers to questions, and have instead focused on incremental progress in lowest-common-denominator search. The classic keyword search box exerts a powerful gravitational pull. Academics and industry researchers need to achieve the intellectual 'escape velocity' necessary to revolutionize search. They must invest much more in bold strategies that can achieve natural-language searching and answering, rather than providing the electronic equivalent of the index at the back of a reference book.

Today, that 'book' is distributed over billions of web pages of uneven quality, and much effort has been directed at ranking the most useful results. Such engines readily index billions of documents, but overwhelm their users with millions of results in response to simple queries. This quandary only worsens as the number of web pages ▸

4 AUGUST 2011 | VOL 476 | NATURE | 25

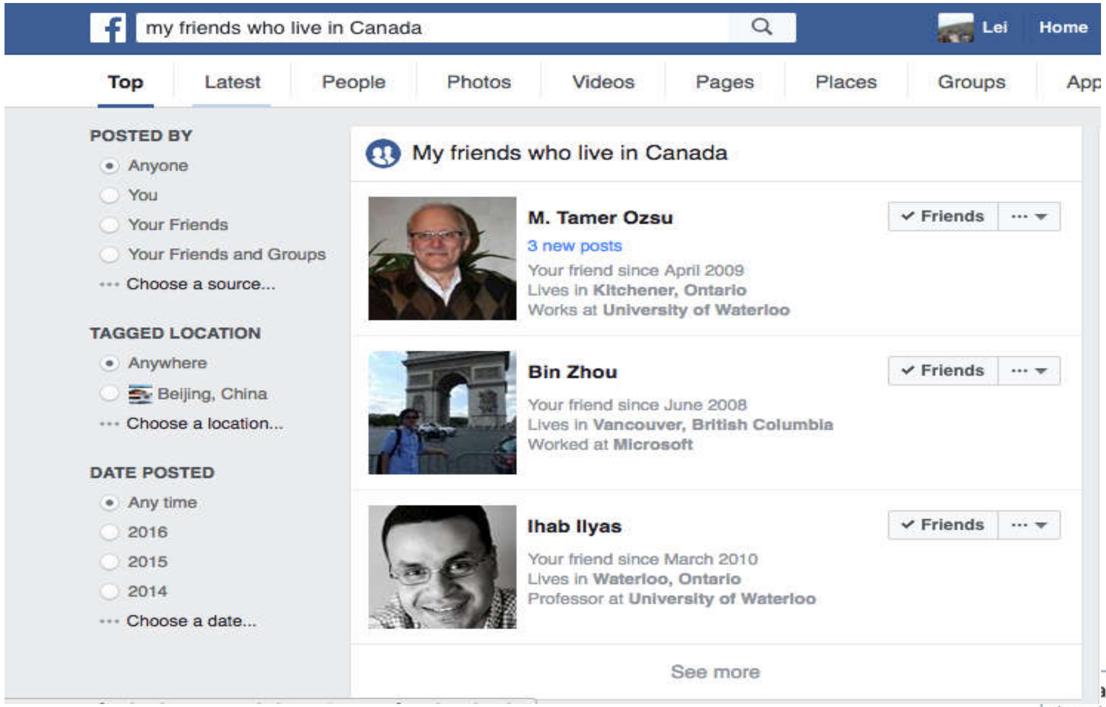
Oren Etzioni, AAAI Fellow

"(Researchers) They must invest much more in bold strategies that can achieve **natural-language searching and answering**"
---Oren Etzioni, Search needs a shake up, **NATURE**, Vol 476, p25-26, 2011.

Facebook Graph Search

“My friends who live in Canada”

“ Facebook Graph Search” ------announced by Mark Zuckerberg on January 16, 2013



my friends who live in Canada

Lei Home

Top | Latest | People | Photos | Videos | Pages | Places | Groups | App

POSTED BY
- ● Anyone
- ○ You
- ○ Your Friends
- ○ Your Friends and Groups
- ··· Choose a source...

TAGGED LOCATION
- ● Anywhere
- ○ Beijing, China
- ··· Choose a location...

DATE POSTED
- ● Any time
- ○ 2016
- ○ 2015
- ○ 2014
- ··· Choose a date...

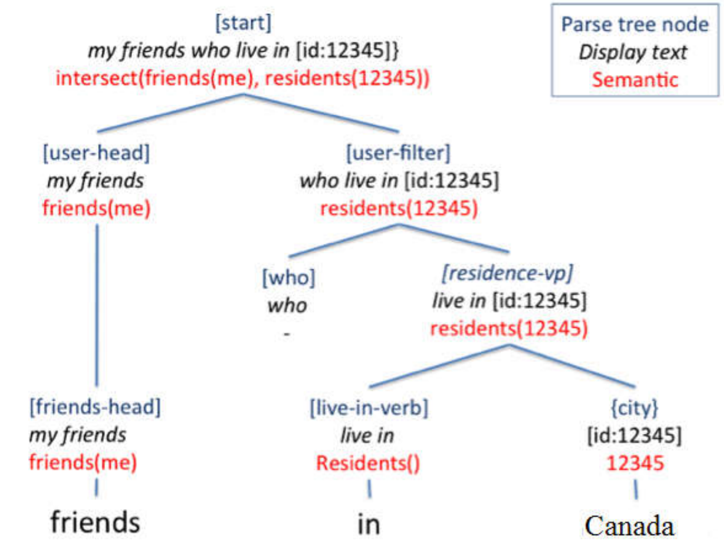
My friends who live in Canada

M. Tamer Ozsu ✓ Friends ···
3 new posts
Your friend since April 2009
Lives in **Kitchener, Ontario**
Works at **University of Waterloo**

Bin Zhou ✓ Friends ···
Your friend since June 2008
Lives in **Vancouver, British Columbia**
Worked at **Microsoft**

Ihab Ilyas ✓ Friends ···
Your friend since March 2010
Lives in **Waterloo, Ontario**
Professor at **University of Waterloo**

See more



Parse tree node
Display text
Semantic

[start]
my friends who live in [id:12345]}
intersect(friends(me), residents(12345))

[user-head]
my friends
friends(me)

[user-filter]
who live in [id:12345]
residents(12345)

[who]
who
-

[residence-vp]
live in [id:12345]
residents(12345)

[friends-head]
my friends
friends(me)
friends

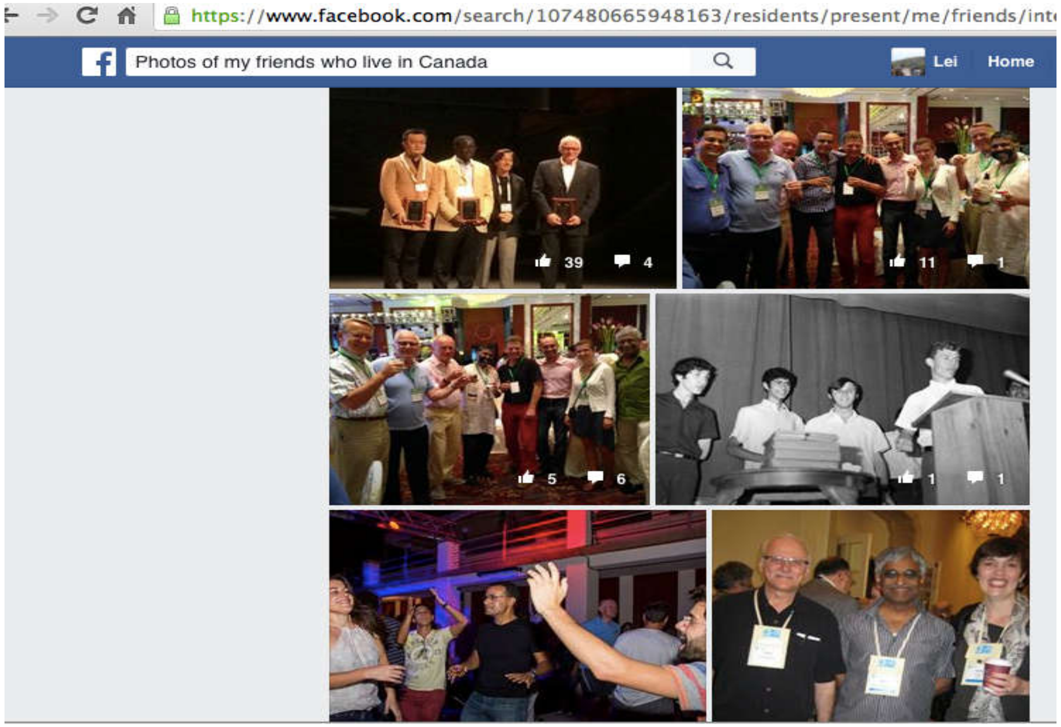
[live-in-verb]
live in
Residents()
in

{city}
[id:12345]
12345
Canada

arse tree, semantic and entity ID used in the above example are for illustration only; they do not represent real information used in Graph Search Beta

Facebook Graph Search

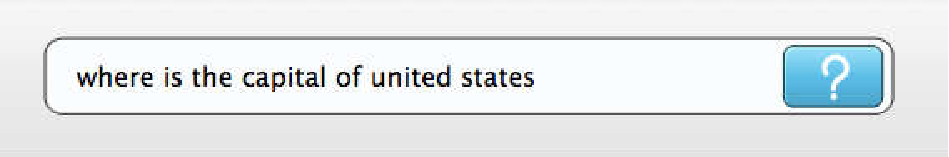
“Photos of my friends who live in Canada”



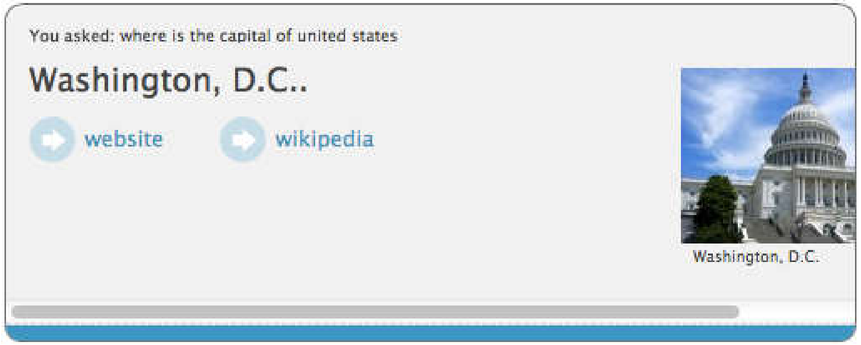
https://www.facebook.com/search/107480665948163/residents/present/me/friends/int

Photos of my friends who live in Canada

EVI---(originally, True Knowledge)



where is the capital of united states



You asked: where is the capital of united states

Washington, D.C..

	Venture Capital
2007-09	1.2 Million USD
2008-07	4 Million USD
2012-01	Acquired by Amazon

William Tunstall-Pedoe: *True Knowledge: Open-Domain Question Answering using Structured Knowledge and Inference*. AI Magazine 31(3): 80-92 (2010)

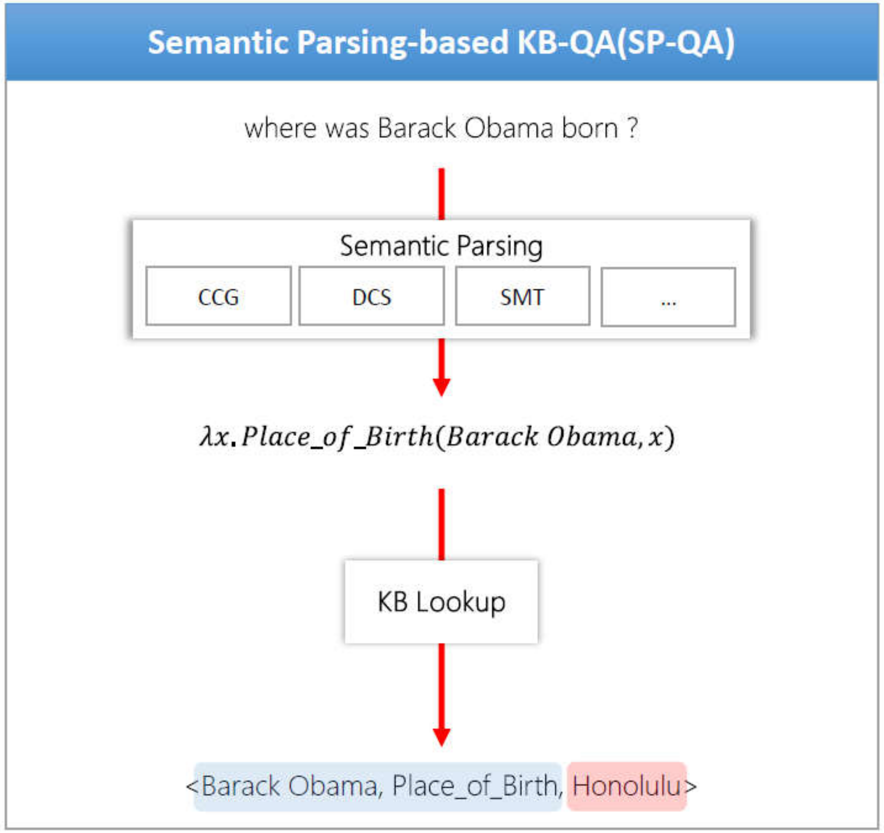
KG-based Question/Answering

- Information Retrieval-based
  - Generate candidate answers
  - Ranking

- Semantic Parsing-based
  - Translate NLQ to logical forms
  - Executing

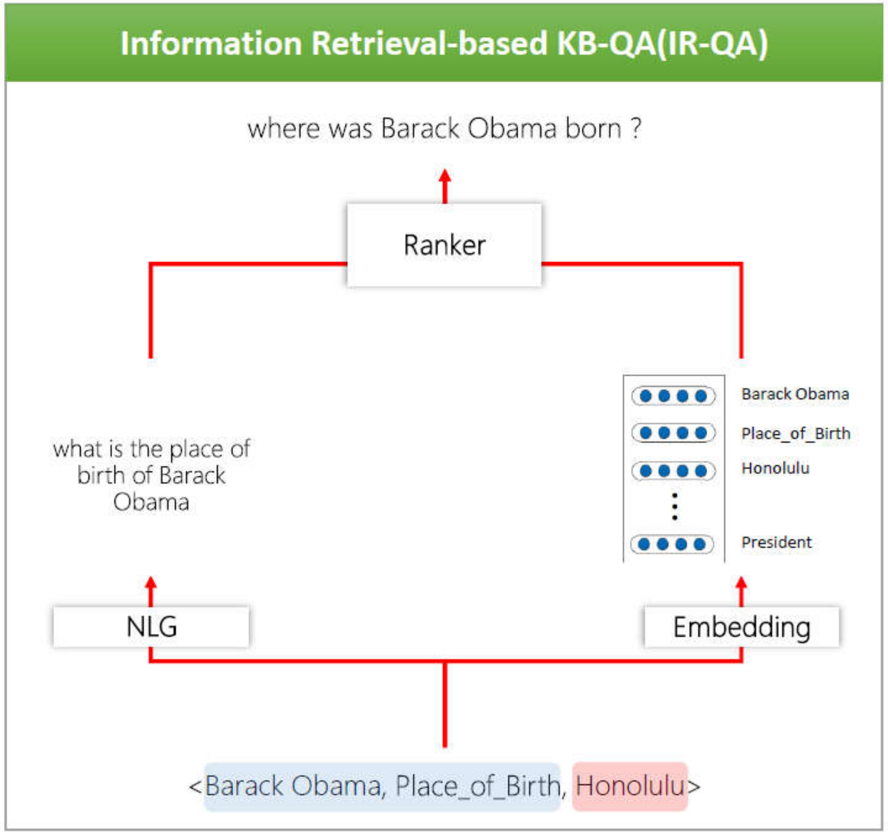
Knowledge-based QA (KB-QA)

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>



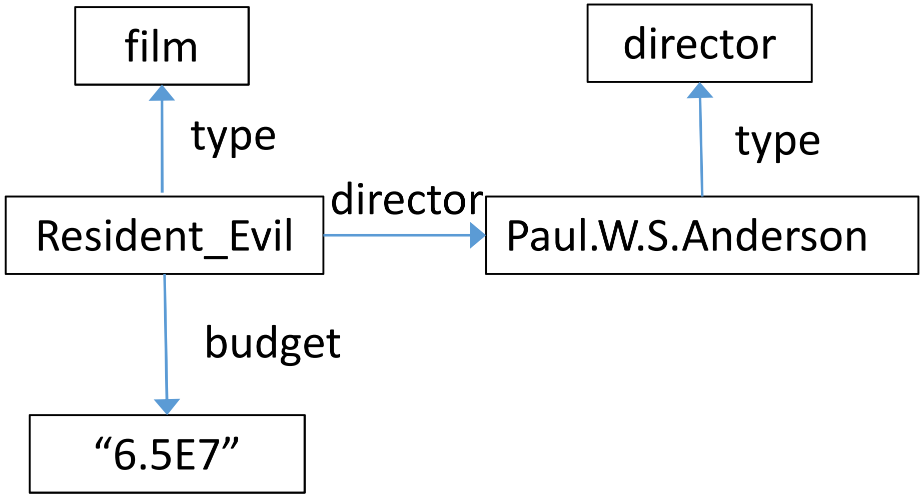
Information Retrieval-based KB-QA(IR-QA)

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>

(Cite: Nan Duan, MSRA)

KG-based Question/Answering

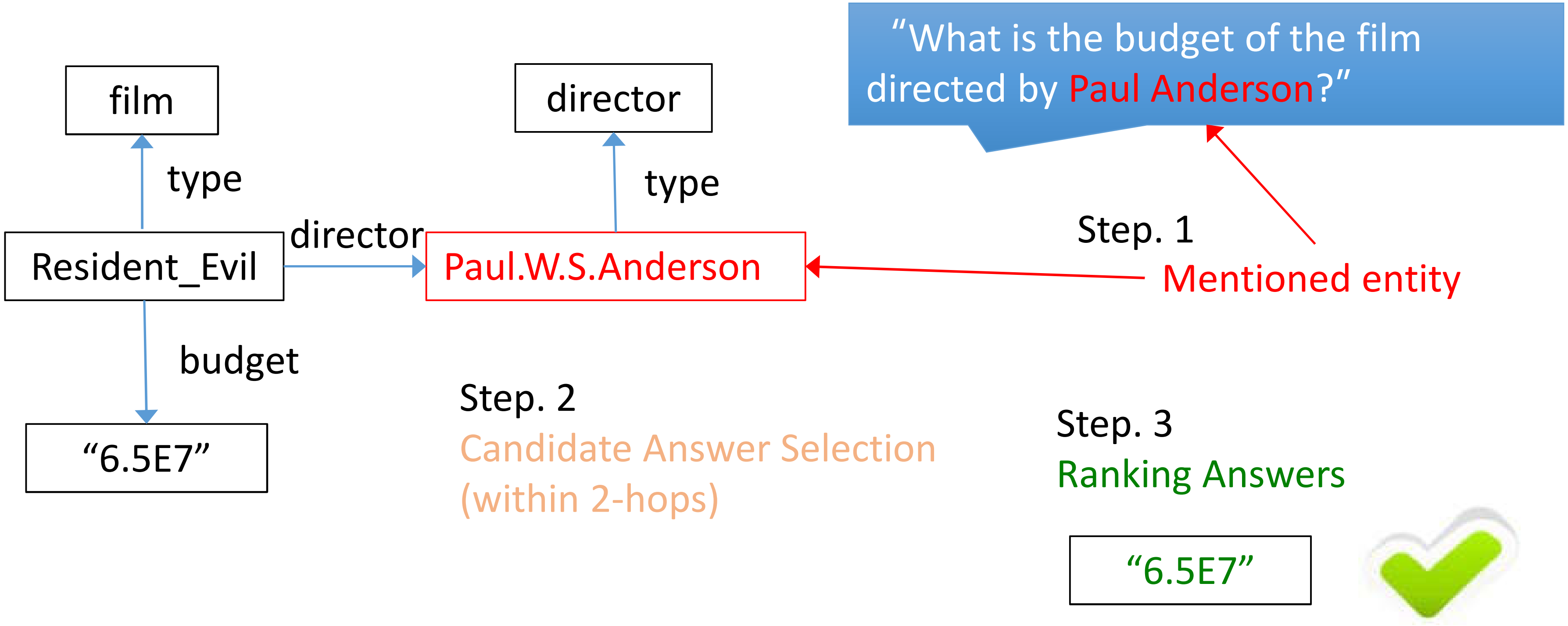
- Information Retrieval-based



"What is the budget of the film directed by Paul Anderson?"

KG-based Question/Answering

- Information Retrieval-based



Question Answering with Subgraph Embeddings [Bordes et al. EMNLP 2014]

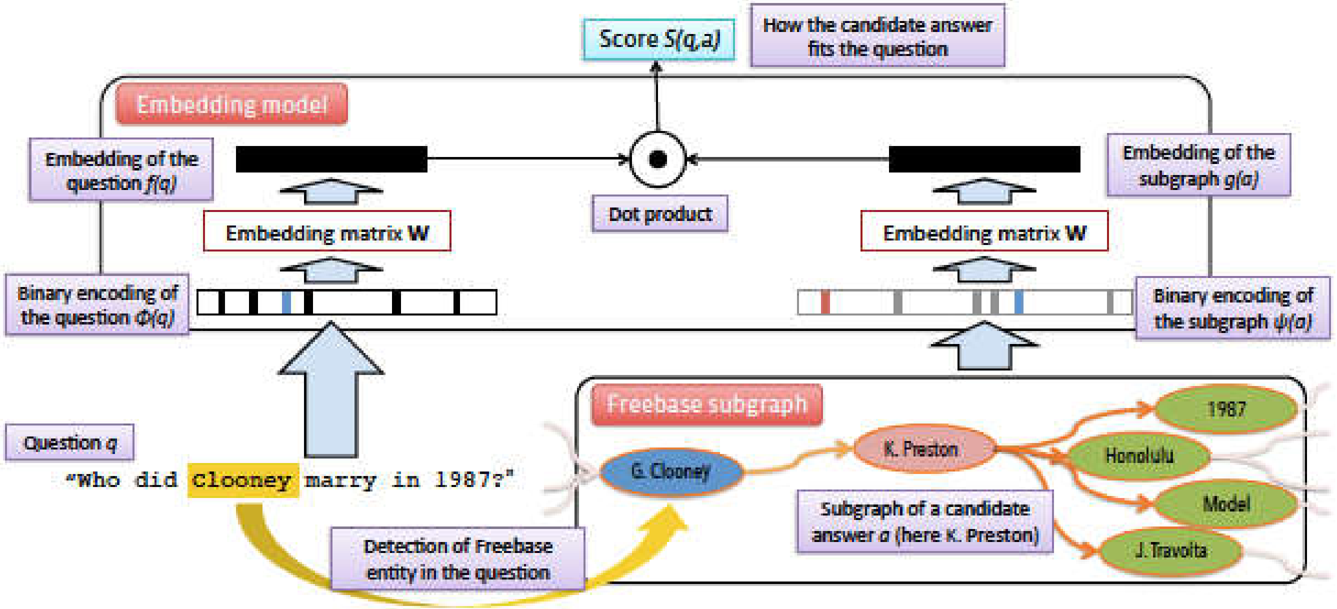

Score S(q,a)
How the candidate answer fits the question
Embedding model
Embedding of the question f(q)
Dot product
Embedding of the subgraph g(a)
Embedding matrix W
Embedding matrix W
Binary encoding of the question Φ(q)
Binary encoding of the subgraph ψ(a)
Freebase subgraph
Question q
"Who did Clooney marry in 1987?"
Detection of Freebase entity in the question
G. Clooney
K. Preston
1987
Honolulu
Model
J. Travolta
Subgraph of a candidate answer a (here K. Preston)

Figure 1: Illustration of the subgraph embedding model scoring a candidate answer: (i) locate entity in the question; (ii) compute path from entity to answer; (iii) represent answer as path plus all connected entities to the answer (the subgraph); (iv) embed both the question and the answer subgraph separately using the learnt embedding vectors, and score the match via their dot product.

Question Answering with Subgraph Embeddings [Bordes et al. EMNLP 2014]

Let W be a matrix $\Re^{k \times N}$

k: the dimension of the embedding space

N: $N = N_W + N_S$

N_W is the number of words

N_S is the number of entities and relation types

Embedding a question q

$$f(q) = W\phi(q)$$

$\phi(q)$ is a sparse vector indicating the presence of words (usually 0 or 1).

Question Answering with Subgraph Embeddings [Bordes et al. EMNLP 2014]

Embedding a candidate answer a

$$g(a) = W\varphi(a)$$

$\varphi(a)$ is a sparse vector representation of the answer *a*

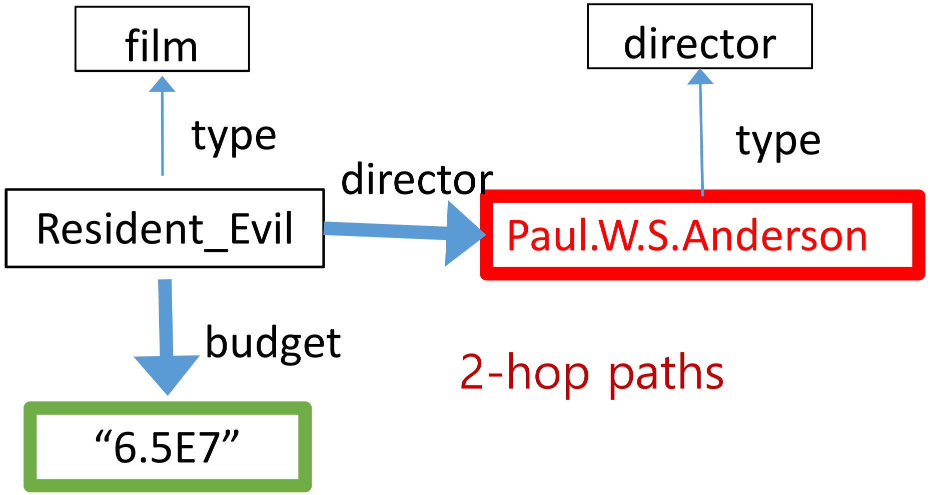
- Single Entity

The answer is represented as a single entity:
$\varphi(a)$ is a 1-of-Ns coded vector with 1 corresponding the answer.

- Path Representation

The answer is represented as a path from the entity mentioned in the question to the answer entity *a*.

$\varphi(a)$ is a 3-of-Ns (or 4-of-Ns) coded vector, expressing the start and the end entities of the path and the relation types (but not entities) in-between.



Candidate Answer

2-hop paths

Question Answering with Subgraph Embeddings [Bordes et al. EMNLP 2014]

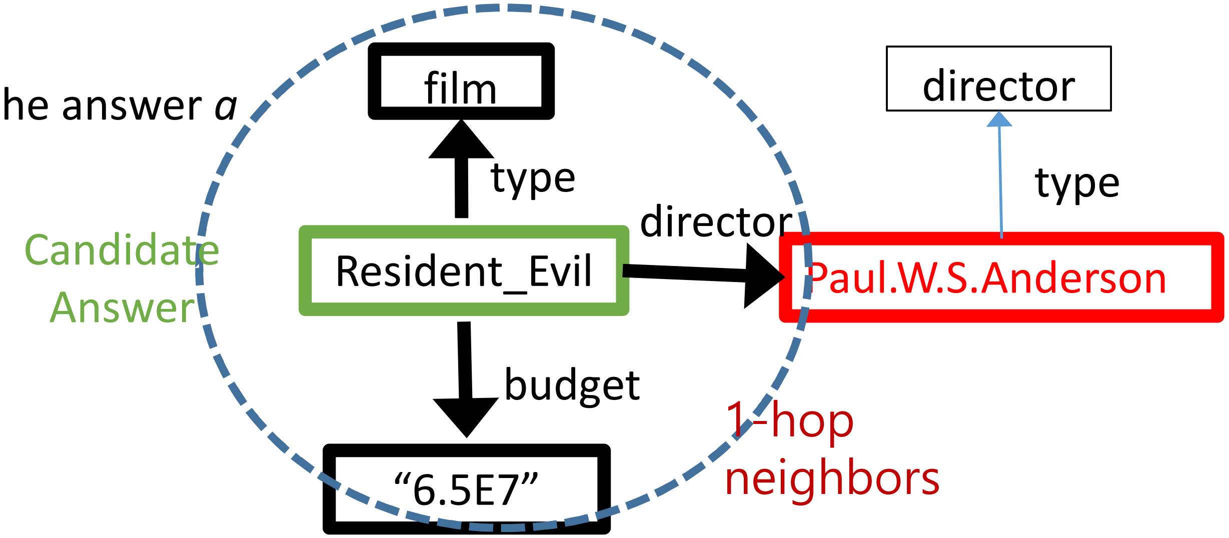
- Subgraph Representation

The answer is represented both the path and 1-hop neighbors around the answer a.

Embedding a candidate answer a

$$g(a) = W\varphi(a)$$

$\varphi(a)$ is a sparse vector representation of the answer *a*



Question Answering with Subgraph Embeddings [Bordes et al. EMNLP 2014]

Scoring Function

candidate answer

$$\boxed{S(q,a) = f(q)^T g(a)}$$

question sentence

The loss function

$$\sum_{i=1}^{|D|} \sum_{a' \in A'(a_i)} \max\{0, m - S(q_i, a_i) + S(q_i, a')\}$$

$A'(a_i)$ is a set of incorrect canidates to question q.

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

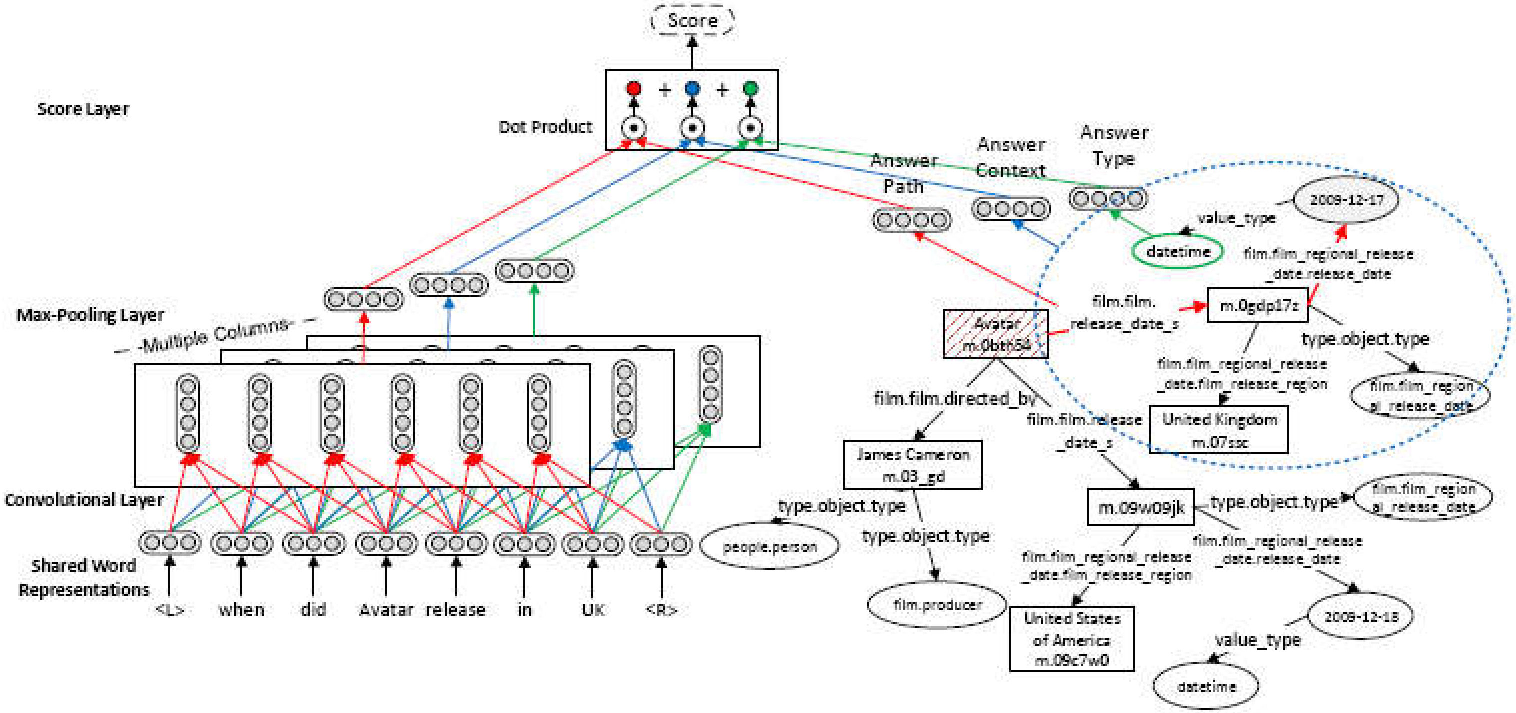


Figure 1: Overview for the question-answer pair *(when did Avatar release in UK, 2009-12-17)*. Left: network architecture for question understanding. Right: embedding candidate answers.

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

Scoring Function

$$S(\overset{\text{question sentence}}{q},\overset{\text{candidate answer}}{a}) =$$

$$\underbrace{f_1(q)^T g_1(a)}_{\text{answer path}} + \underbrace{f_2(q)^T g_2(a)}_{\text{answer context}} + \underbrace{f_3(q)^T g_3(a)}_{\text{answer type}}$$

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

MCCNNs for
Question Understanding

Let the question $q = w_1 w_2 \ldots w_n$

The look layer transform every word into a vector

$$w_j = W_v u(w_j)$$

$W_v \in \Re^{d_v \times |V|}$,

d_v is the word embedding dimention and

|V| is the vocabulary size

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

MCCNNs for
Question Understanding

Let the question $q = w_1 w_2 \dots w_n$

The convolutional layer computes representation of the words in sliding windows.

$$x_j = h(W[w_{j-s}^T \dots w_j^T \dots w_{j+s}^T] + b)$$

The max-pooling layer

$$f(q) = \max_{j=1,\dots,n} \{x_j\}$$

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

Embedding Candidate Answers

Answer Path

$$g_1(a) = \frac{1}{\|u_p(a)\|_1} W_p u_p(a)$$

$u_p(a)$ is a length-|R| binary vector, indicating the presence or absence of every relation in the answer path.

$W_p \in \Re^{d_q \times |R|}$ is the parameter matrix

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

Embedding Candidate Answers

Answer Context

The 1-hop entities and relations connected to the answer path are regarded as the *answer context*.

$$g_2(a) = \frac{1}{\left\| u_c(a) \right\|_1} W_c u_c(a)$$

$u_c(a)$ is a length-|C| binary vector, indicating the presence or absence of every entity or relation in the context.

$W_c \in \Re^{d_q \times |C|}$ is the parameter matrix

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

Embedding Candidate Answers

Answer Type

Type information is an important clue to score candidate answers.

$$g_3(a) = \frac{1}{\|u_t(a)\|_1} W_t u_t(a)$$

$u_t(a)$ is a length-|T| binary vector, indicating the presence or absence of answer type.

$W_t \in \Re^{d_t \times |T|}$ is the parameter matrix

Question Answering over Freebase with Multi-Column Convolutional Neural Networks [Dong et al., ACL 2015]

Model Training

For every correct answer a of the question q, we randomly sample k wrong a’ from the set of candidate answers, and use them as the negative instances to estimate parameters.

$$l(q,a,a') = (m - S(q,a) + S(q,a'))_{+}$$

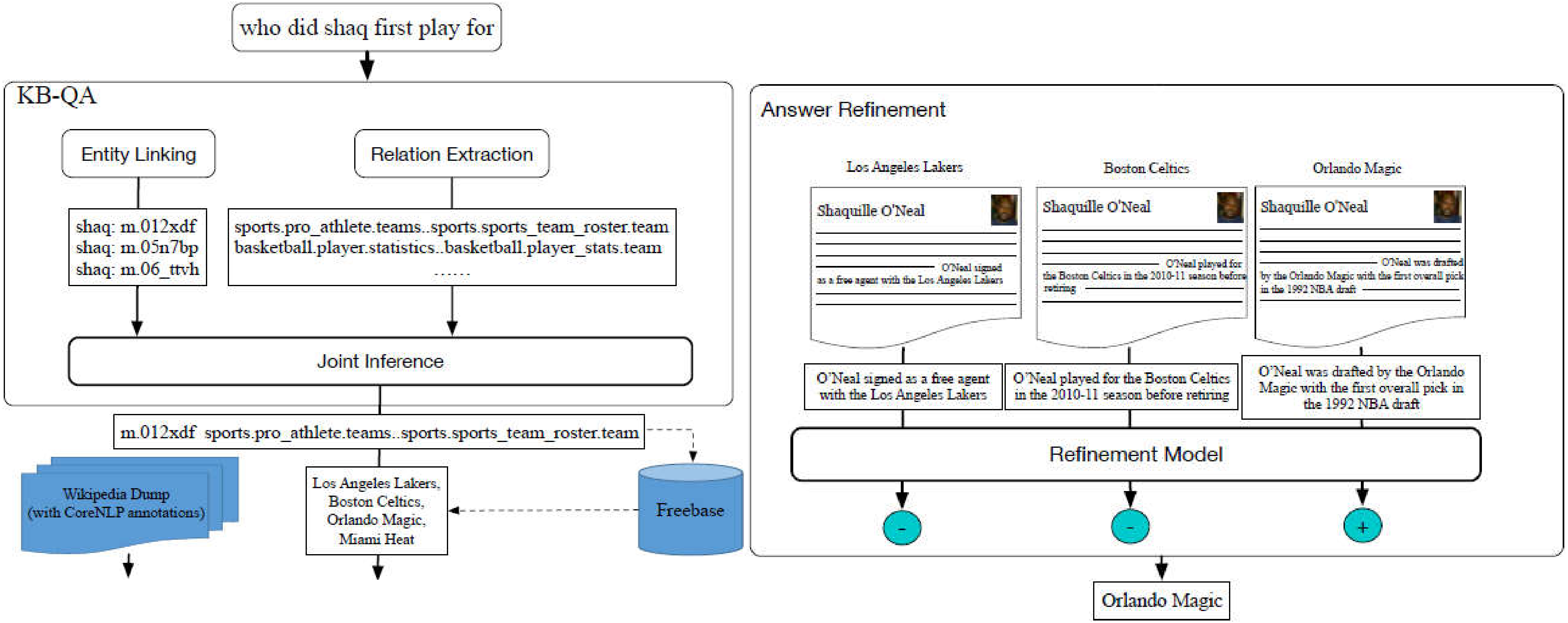
$$\min \sum_{q} \frac{1}{|A_q|} \sum_{a \in A_q} \sum_{a' \in R_q} l(q,a,a')$$

$$R_q \subseteq C_q \setminus A_q$$

A_q is the correct answer set to question q.

C_q is the set of canidate answer set to question q.

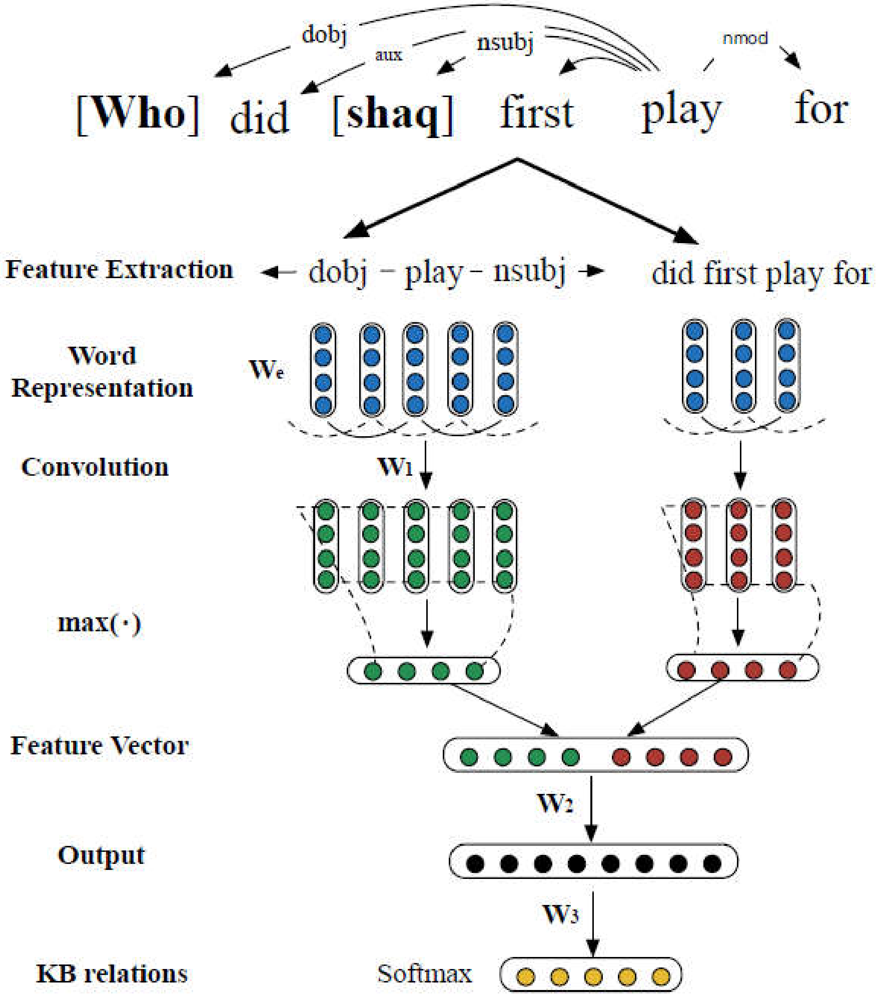
Question Answering on Freebase via Relation Extraction and Textual Evidence [Xu et al., ACL 2016]


who did shaq first play for
KB-QA
Entity Linking
Relation Extraction
shaq: m.012xdf
shaq: m.05n7bp
shaq: m.06_ttvh
sports.pro_athlete.teams..sports.sports_team_roster.team
basketball.player.statistics..basketball.player_stats.team
......
Joint Inference
m.012xdf sports.pro_athlete.teams..sports.sports_team_roster.team
Wikipedia Dump
(with CoreNLP annotations)
Los Angeles Lakers,
Boston Celtics,
Orlando Magic,
Miami Heat
Freebase
Answer Refinement
Los Angeles Lakers
Boston Celtics
Orlando Magic
Shaquille O'Neal
O'Neal signed as a free agent with the Los Angeles Lakers
Shaquille O'Neal
O'Neal played for the Boston Celtics in the 2010-11 season before retiring
Shaquille O'Neal
O'Neal was drafted by the Orlando Magic with the first overall pick in the 1992 NBA draft
O'Neal signed as a free agent with the Los Angeles Lakers
O'Neal played for the Boston Celtics in the 2010-11 season before retiring
O'Neal was drafted by the Orlando Magic with the first overall pick in the 1992 NBA draft
Refinement Model
-
-
+
Orlando Magic

胡森

Question Answering on Freebase via Relation Extraction and Textual Evidence[Xu et al., ACL 2016]

Relation Extraction



Question Answering on Freebase via Relation Extraction and Textual Evidence[Xu et al., ACL 2016]

Question Decomposition

"who plays ken barlow in coronation street? “

decompose

“who plays ken barlow”

+

“who plays in coronation street”

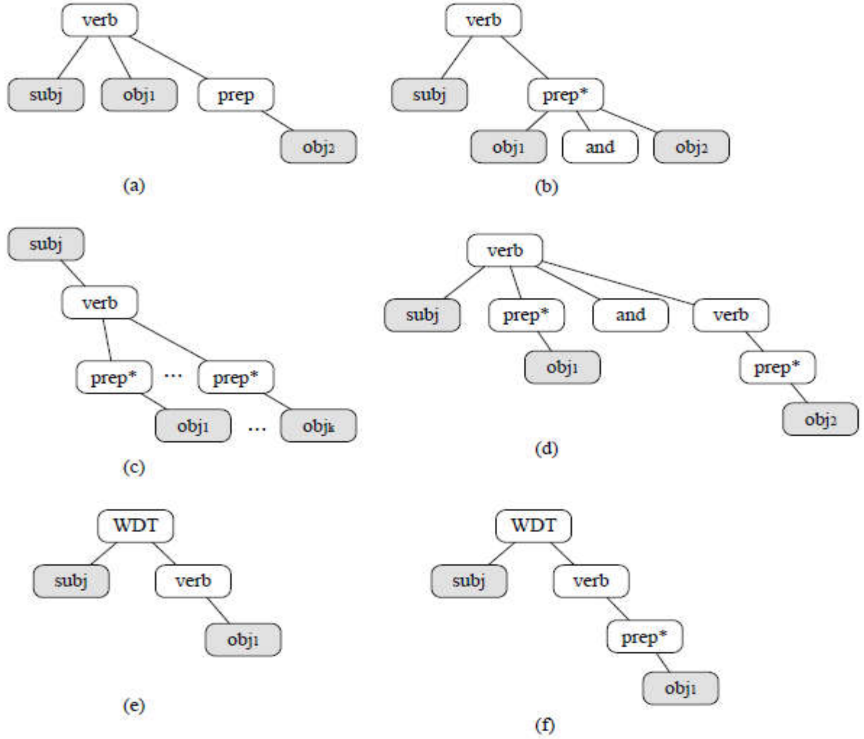


Figure 3: Syntax-based patterns for question decomposition.

KG-based Question/Answering

- Information Retrieval-based
  - Generate candidate answers
  - Ranking

- Semantic Parsing-based
  - Translate NLQ to logical forms
  - Executing

Semantic Parsing

[Zettlemoyer et al., UAI 05]

Transforming natural language (NL) sentences into computer executable complete meaning representations (MRs)
for domain-specic applications.

E.g., “Which states borders New Mexico ?”

↓

Lambda-**calculus** [Alonzo Church, 1940]

$$\lambda x.state(x) \wedge borders(x, new\_mexico)$$

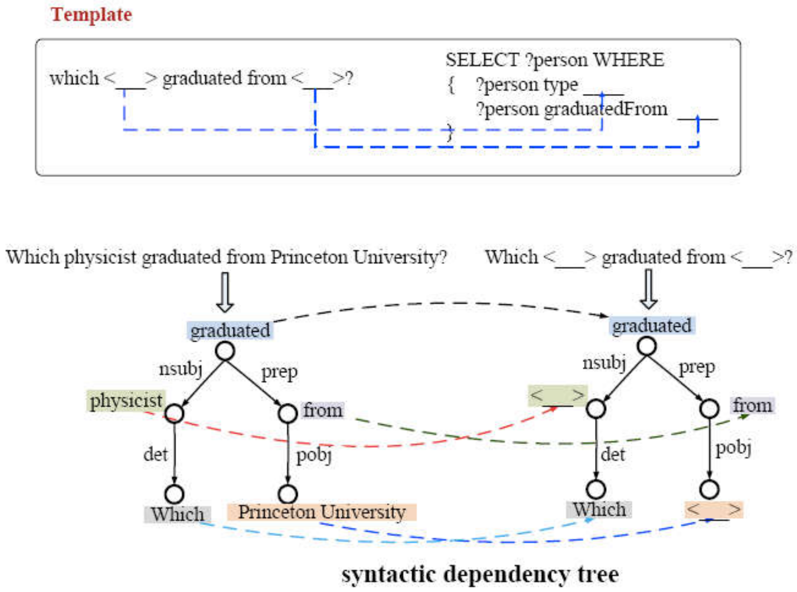
“**Simply typed Lambda-calculus** can express varies database query languages such as **relational algebra**, fixpoint logic and the complex object algebra." [Hillebrand et al., 1996]

Semantic Parsing

- **Manually constructed rules**
  [Pedoe, AI magazine 2010]

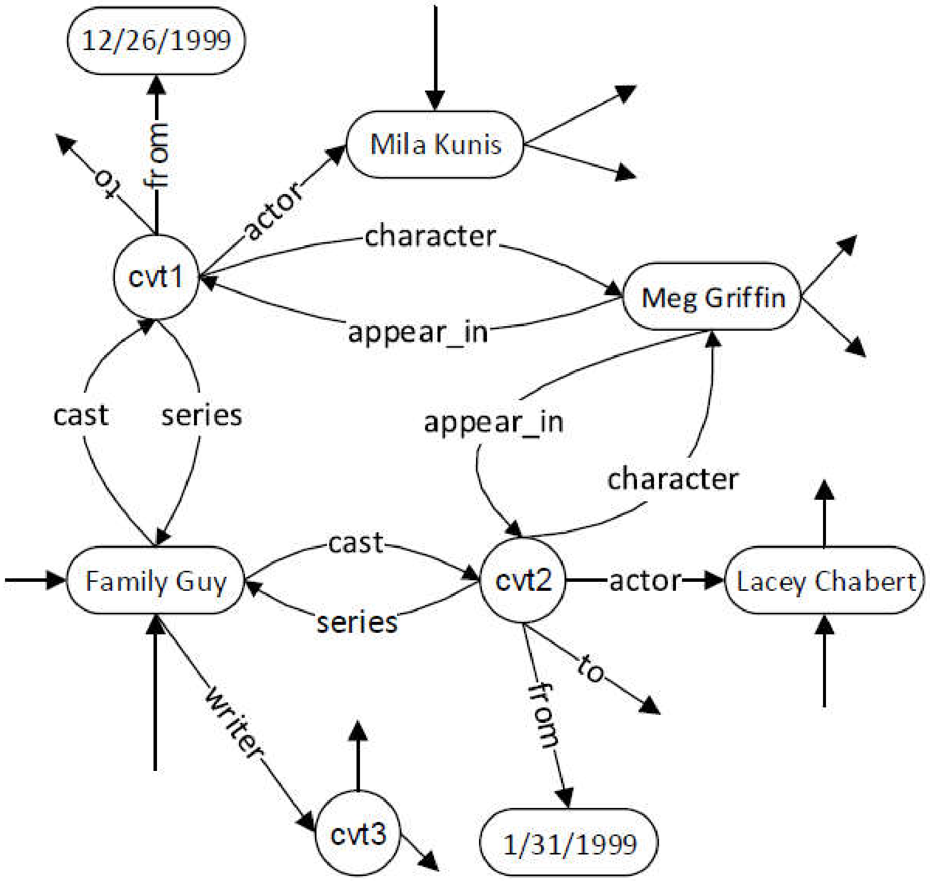
- **Grammar-based, e.g.,**
  Combinatory Categorial Grammar
  [Zettlemoyer and Collins, UAI 2005]

- **Supervised Learning**
  [Berant and Liang, ACL 2014]


Template
which <___> graduated from <___>?
SELECT ?person WHERE
{ ?person type ___
?person graduatedFrom ___
}
Which physicist graduated from Princeton University?
Which <___> graduated from <___>?
graduated
nsubj
prep
physicist
from
det
pobj
Which
Princeton University
graduated
nsubj
prep
<___>
from
det
pobj
Which
<___>
syntactic dependency tree

Template-based Approach [cite: Weiguo Zheng, Lei Zou, et al., SIGMOD 15]

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base [Yih et al., ACL 2015]



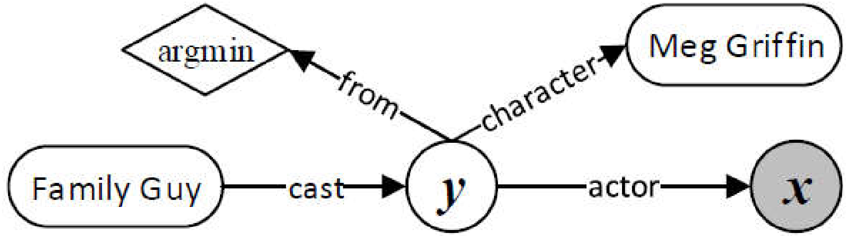


Figure 2: Query graph that represents the question "Who first voiced Meg on Family Guy?"

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base [Yih et al., ACL 2015]

Query Graph Generation

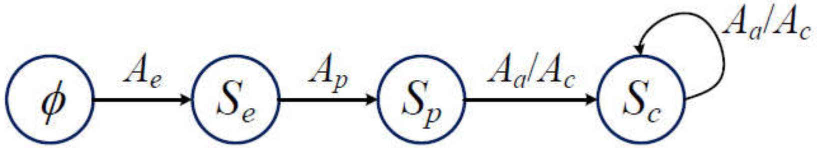


Figure 3: The legitimate actions to *grow* a query graph. See text for detail.

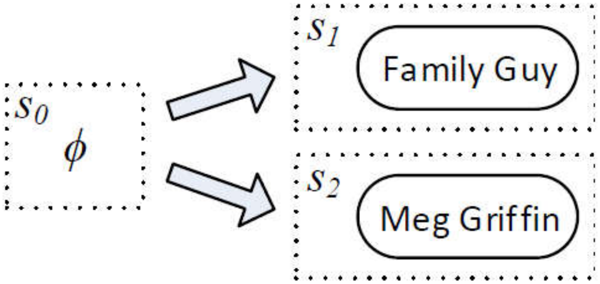


Figure 4: Two possible topic entity linking actions applied to an empty graph, for question “Who first voiced [Meg] on [Family Guy]?”

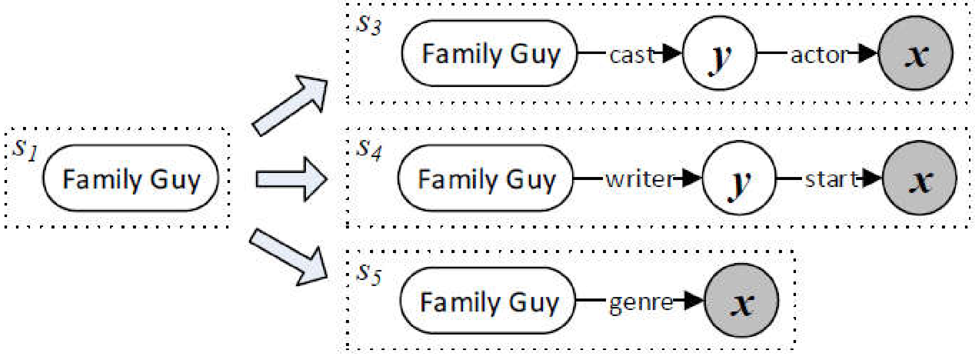


Figure 5: Candidate core inferential chains start from the entity `FamilyGuy`.

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base [Yih et al., ACL 2015]

Query Graph Generation

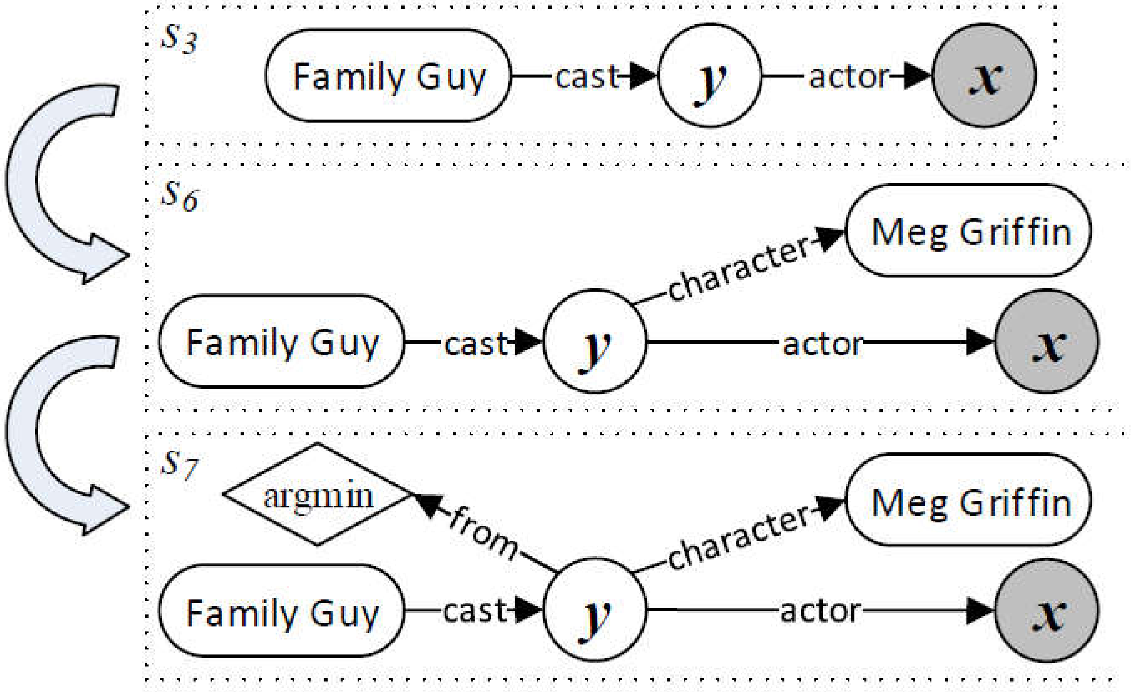
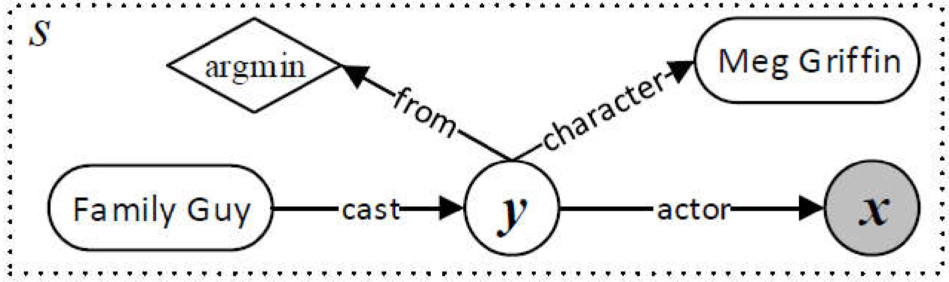


Figure 7: Extending an inferential chain with constraints and aggregation functions.

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base [Yih et al., ACL 2015]

Reward Function

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



(1) EntityLinkingScore(FamilyGuy, "Family Guy") = 0.9

(2) PatChain("who first voiced meg on <e>", cast-actor) = 0.7

(3) QuesEP(q, "family guy cast-actor") = 0.6

(4) ClueWeb("who first voiced meg on <e>", cast-actor) = 0.2

(5) ConstraintEntityWord("Meg Griffin", q) = 0.5

(6) ConstraintEntityInQ("Meg Griffin", q) = 1

(7) AggregationKeyword(argmin, q) = 1

(8) NumNodes(s) = 5

(9) NumAns(s) = 1

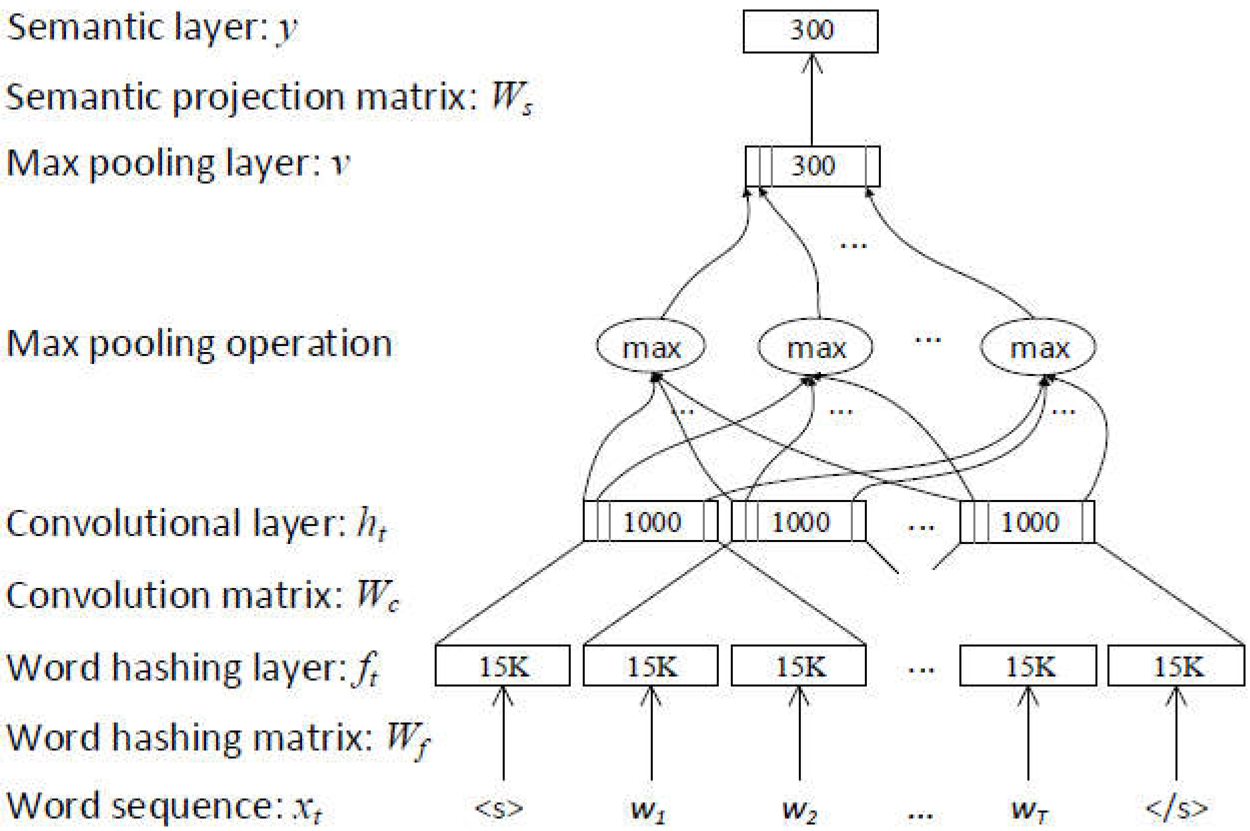
Figure 8: Active features of a query graph s. (1) is the entity linking score of the topic entity. (2)-(4) are different model scores of the core chain. (5) indicates 50% of the words in "Meg Griffin" appear in the question q. (6) is 1 when the mention "Meg" in q is correctly linked to MegGriffin by the entity linking component. (8) is the number of nodes in s. The knowledge base returns only 1 entity when issuing this query, so (9) is 1.

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base [Yih et al., ACL 2015]

Identifying Core Inferential Chain (Relation Extraction)

two neural networks
1) question
2) inferential chain

Compute Similarity
(e.g. cosine)



Language to Logical Form with Neural Attention [Dong et al., ACL 2016]

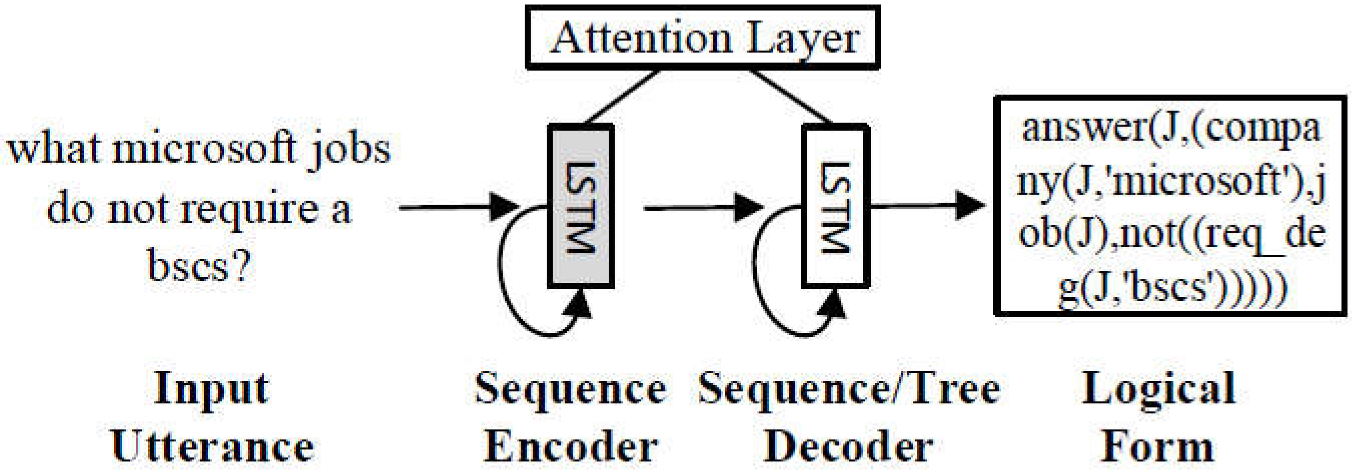

Attention Layer
what microsoft jobs do not require a bscs?
LSTM
LSTM
answer(J,(company(J,'microsoft'),job(J),not((req_deg(J,'bscs')))))
Input Utterance
Sequence Encoder
Sequence/Tree Decoder
Logical Form

Figure 1: Input utterances and their logical forms are encoded and decoded with neural networks. An attention layer is used to learn soft alignments.

Language to Logical Form with Neural Attention [Dong et al., ACL 2016]

dallas to san francisco leaving after 4 in the afternoon please
(lambda \$0 e (and (>(departure_time \$0) 1600:ti) (from \$0 dallas:ci) (to \$0 san_francisco:ci)))

Algorithm 1 Decoding for SEQ2TREE

```
Input: $q$: Natural language utterance
Output: $\hat{a}$: Decoding result
▷ Push the encoding result to a queue
$Q.init(\{hid : \mathsf{SeqEnc}(q)\})$
▷ Decode until no more nonterminals
while $(c \leftarrow Q.pop()) \neq \varnothing$ do
    ▷ Call sequence decoder
    $c.child \leftarrow \mathsf{SeqDec}(c.hid)$
    ▷ Push new nonterminals to queue
    for $n \leftarrow$ nonterminal in $c.child$ do
        $Q.push(\{hid : \mathsf{HidVec}(n)\})$
$\hat{a} \leftarrow$ convert decoding tree to output sequence
```

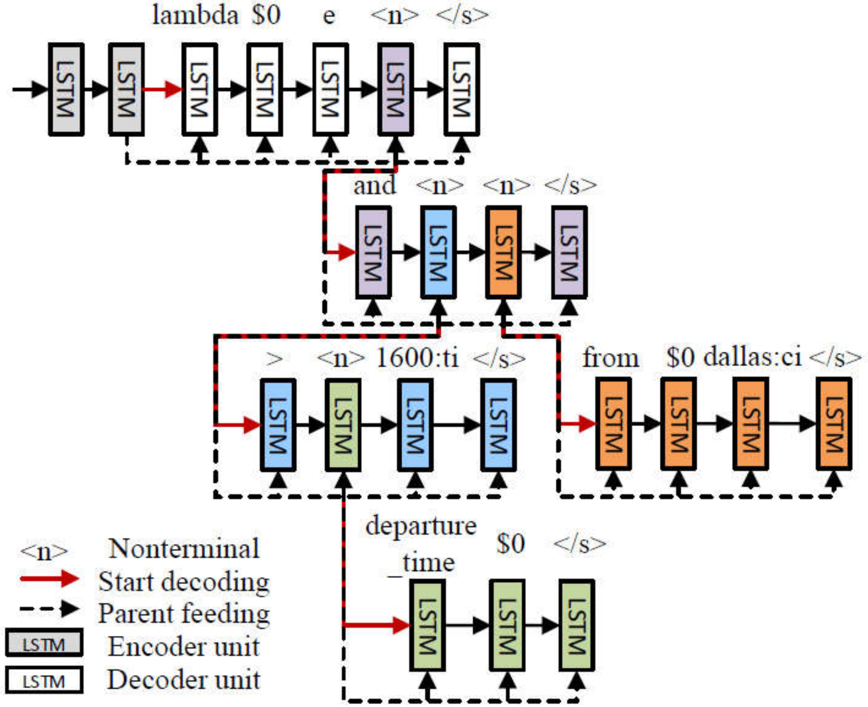
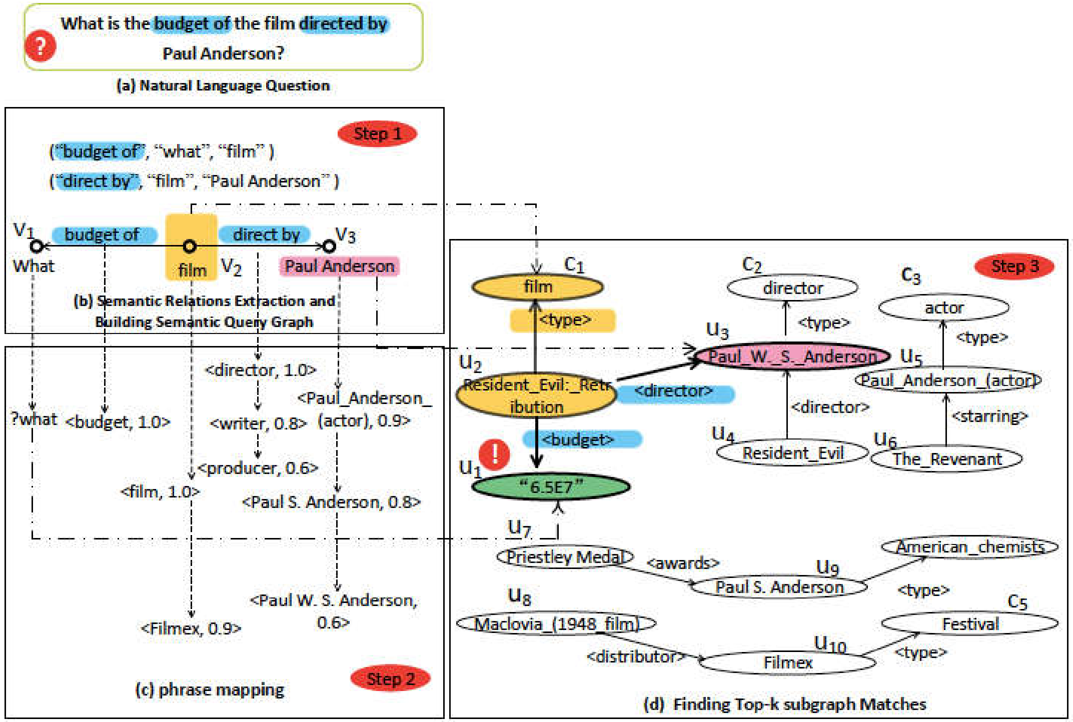


Figure 3: Sequence-to-tree (SEQ2TREE) model with a hierarchical tree decoder.

Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

- Using graph matching-based method
- Graph Matching-based Disambiguation
- Combing Disambiguation and Query together



What is the budget of the film directed by Paul Anderson?

(a) Natural Language Question

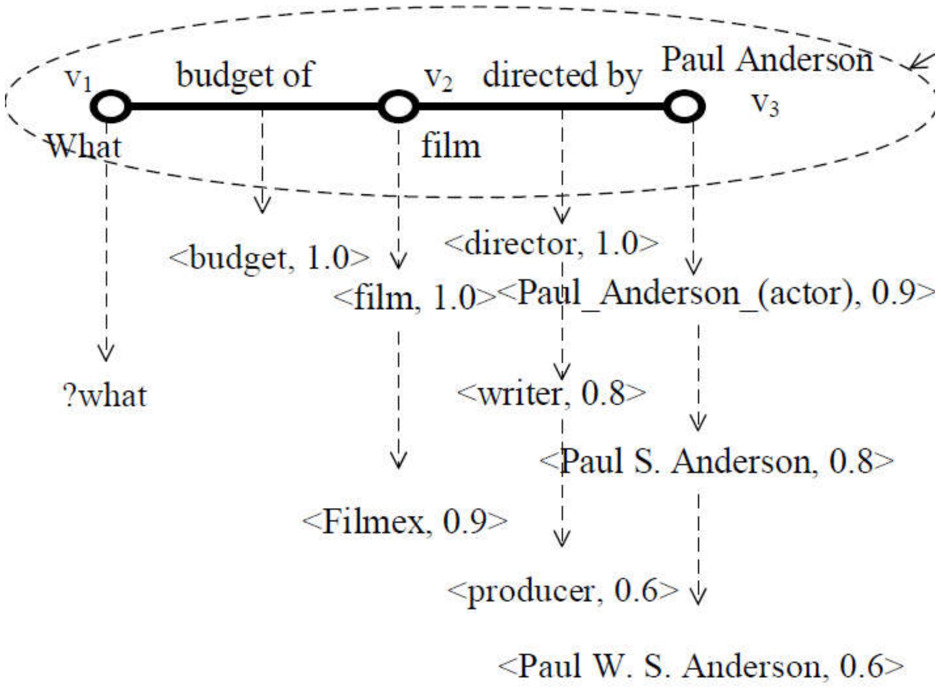
(b) Semantic Relations Extraction and Building Semantic Query Graph

(c) phrase mapping

(d) Finding Top-k subgraph Matches

Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

Semantic Query Graph



Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

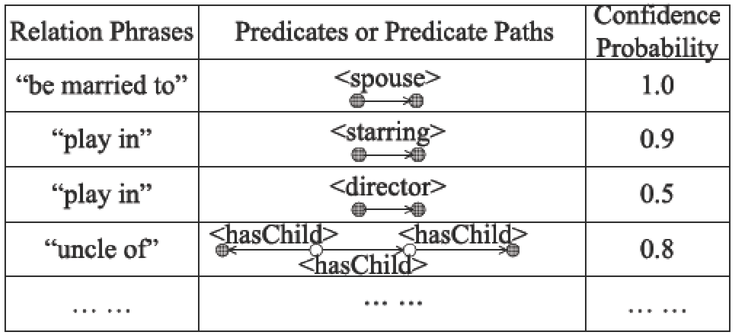
Besides KG, we require two dictionaries.

- Entity Mention Dictionary

  It helps the entity linking task
  [Spitkovsky et al., LERC 12; Chisholm et al, TACL 15].

- Relation Mention Dictionary

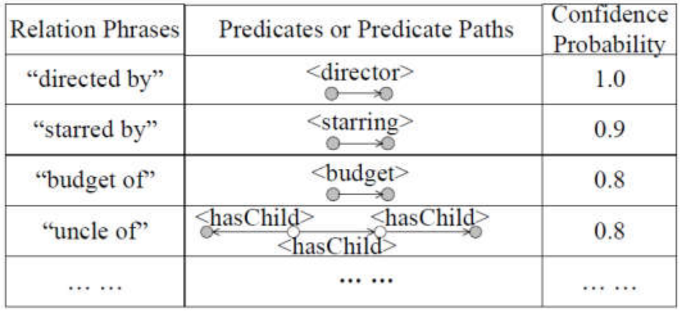
  Mapping the natural language relation phrases to predicate in RDF dataset.
  [Nakashole et al., EMNLP-CoNLL 2012]



Relation Phrases	Predicates or Predicate Paths	Confidence Probability
"be married to"	<spouse>	1.0
"play in"	<starring>	0.9
"play in"	<director>	0.5
"uncle of"	<hasChild> <hasChild> <hasChild>	0.8
… …	… …	… …

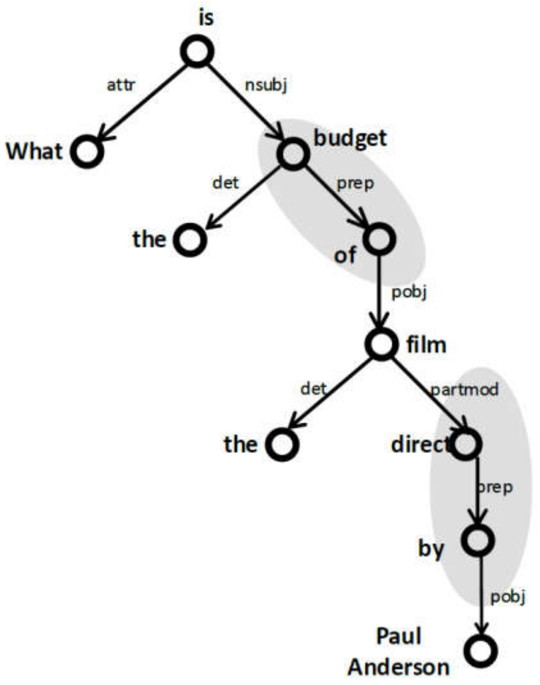
Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

- Question Understanding
  - Relation extraction



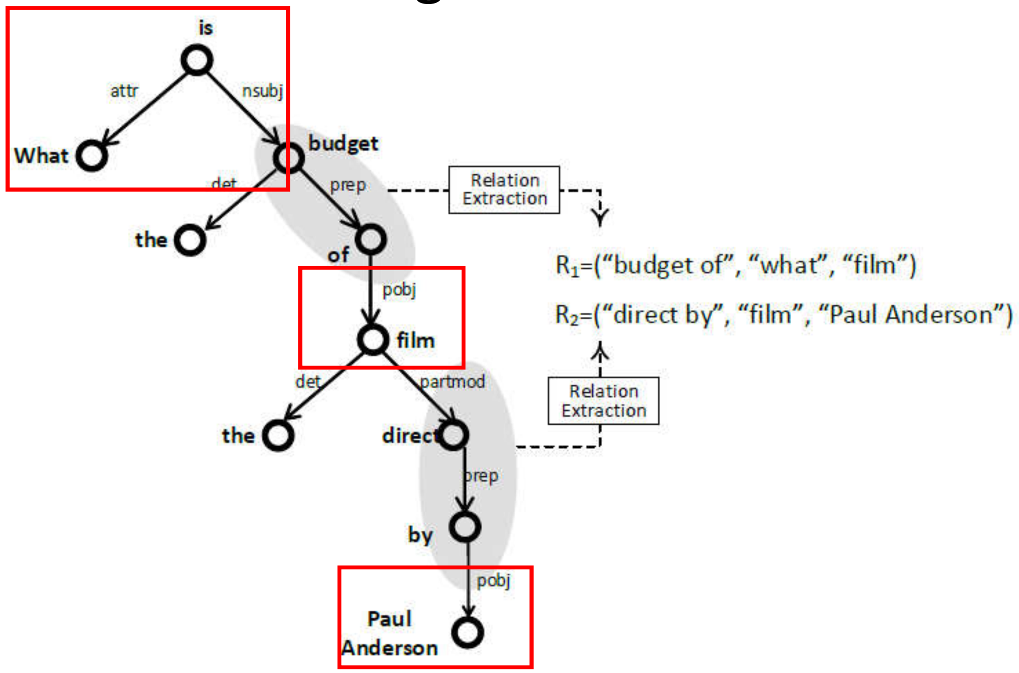
Relation Phrases	Predicates or Predicate Paths	Confidence Probability
"directed by"	<director>	1.0
"starred by"	<starring>	0.9
"budget of"	<budget>	0.8
"uncle of"	<hasChild> <hasChild> <hasChild>	0.8
... ...	... ...	... ...

Relation Paraphrase Dictionary



Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

- Question Understanding
  - Find associated arguments

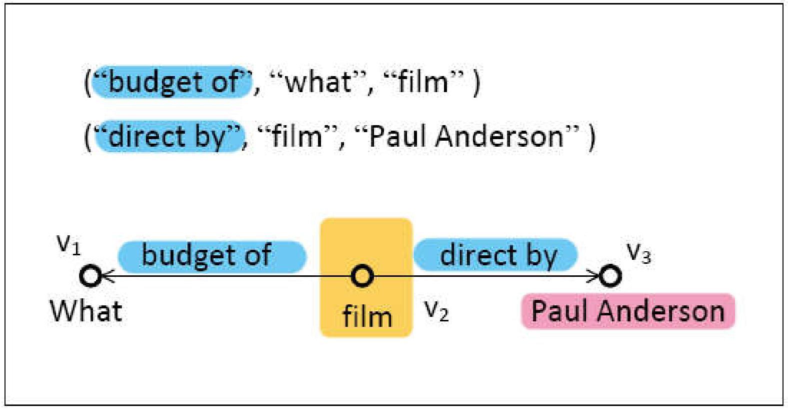


R_1=("budget of", "what", "film")

R_2=("direct by", "film", "Paul Anderson")

Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

- Question Understanding
  - Query Graph Assembly

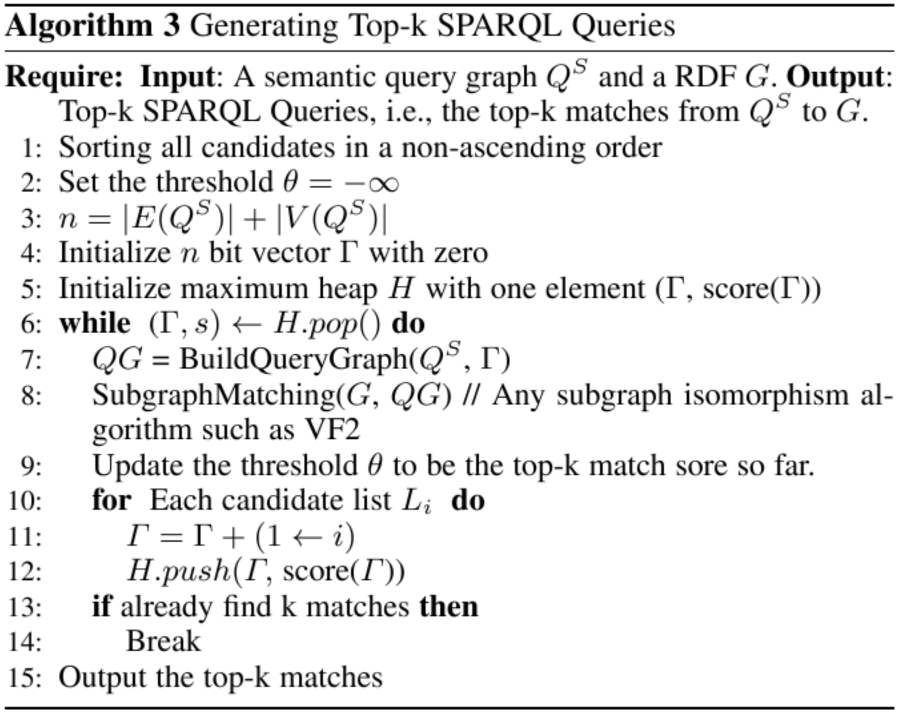


(“budget of”, “what”, “film”)
(“direct by”, “film”, “Paul Anderson”)
v1
budget of
What
film
v2
direct by
v3
Paul Anderson

Semantic Relations Extraction and Building Semantic Query Graph

Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

- Query Execution



Algorithm 3 Generating Top-k SPARQL Queries

```
Require: Input: A semantic query graph $Q^S$ and a RDF $G$. Output:
    Top-k SPARQL Queries, i.e., the top-k matches from $Q^S$ to $G$.
Sorting all candidates in a non-ascending order
Set the threshold $\theta = -\infty$
$n = |E(Q^S)| + |V(Q^S)|$
Initialize $n$ bit vector $\Gamma$ with zero
Initialize maximum heap $H$ with one element ($\Gamma$, score($\Gamma$))
while ($\Gamma, s) \leftarrow H.pop()$ do
    $QG$ = BuildQueryGraph($Q^S$, $\Gamma$)
    SubgraphMatching($G$, $QG$) // Any subgraph isomorphism algorithm such as VF2
    Update the threshold $\theta$ to be the top-k match sore so far.
    for Each candidate list $L_i$ do
        $\mathit{\Gamma} = \Gamma + (1 \leftarrow i)$
        $H.push$($\mathit{\Gamma}$, score($\mathit{\Gamma}$))
    if already find k matches then
        Break
Output the top-k matches
```

Our Approach- Data Driven & Relation-first framework gAnswer [Zou et al, SIGMOD 14]

- Limitations
  - Still highly relied on parser and heuristic rules
  - Can not handle implicit relations

What is the budget of the film directed by Paul Anderson and starred by a **Chinese girl**

<?girl, dbo:country, dbr:China>

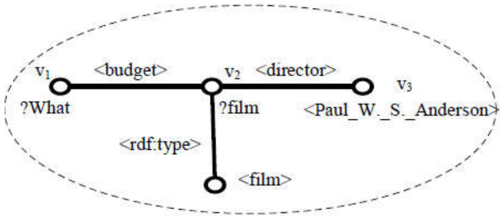
Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Data Driven!
  - The structure of query graph can be modified in execution stage.
  - First recognize nodes.

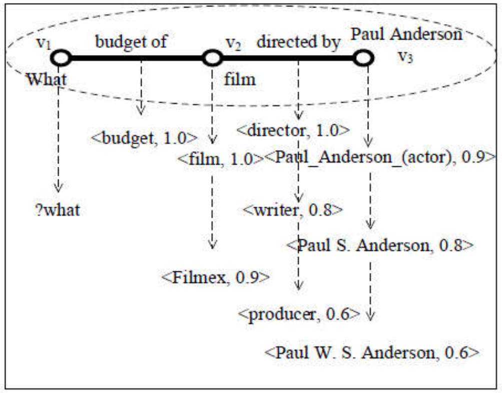
Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

Hyper Query Graph

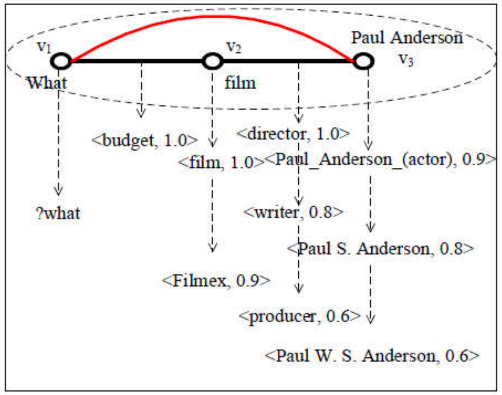
- Extend SQG by *allowing false edges*.



query graph



semantic query graph



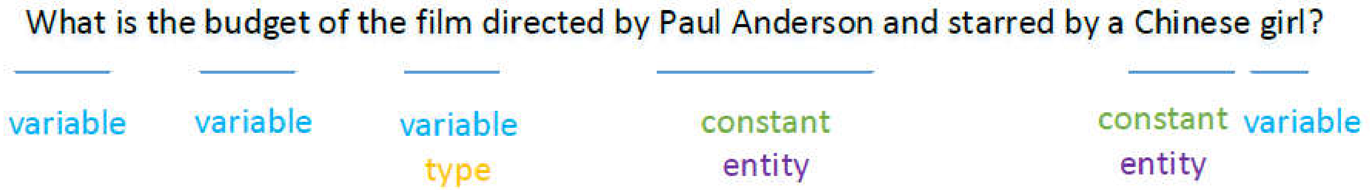
hyper query graph

Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Question Understanding
  - Node recognizing

entity extraction + conflict resolution

- entity, type, literal, wildcard
- constant, variable
- modified, hidden information

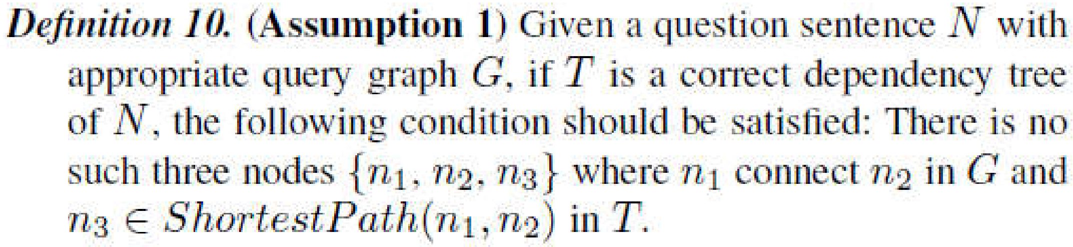


What is the budget of the film directed by Paul Anderson and starred by a Chinese girl?

Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

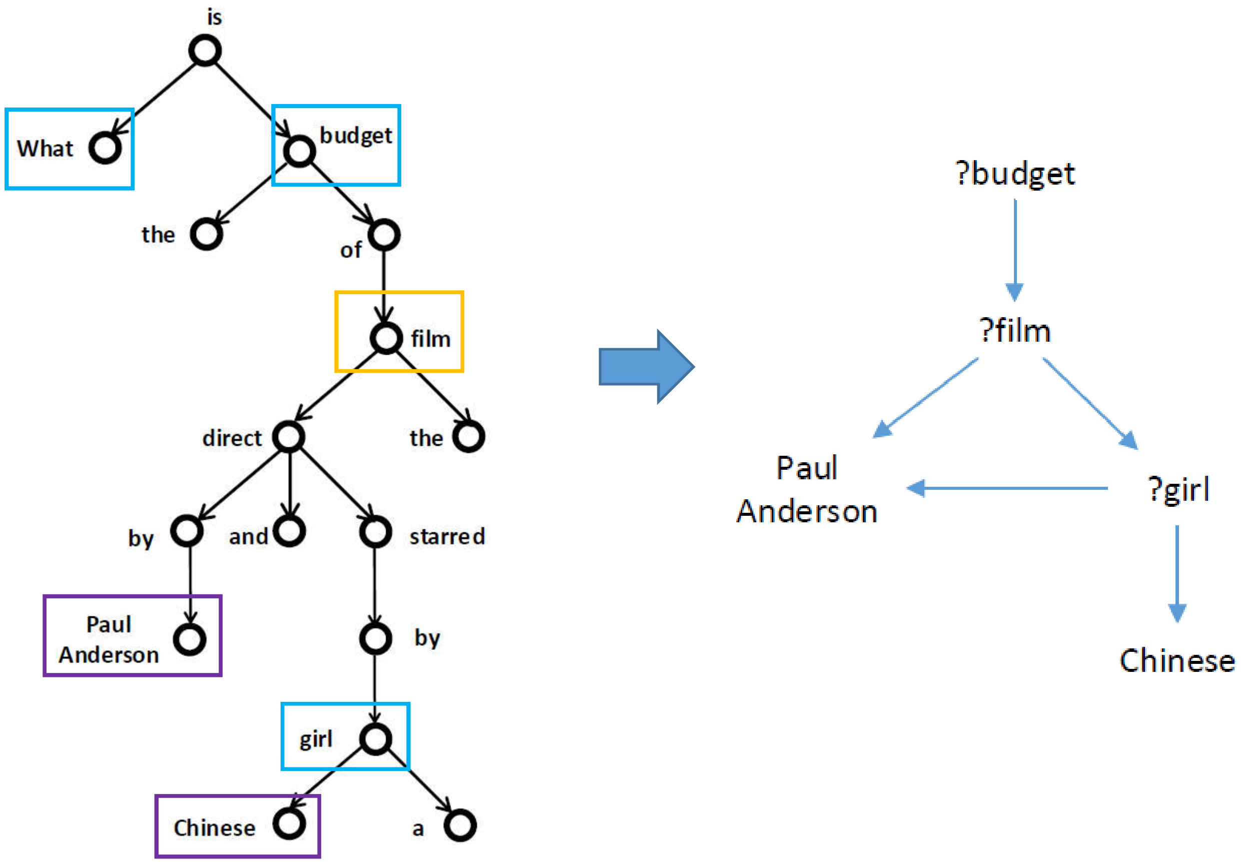
- Question Understanding
  - Build structure of HQG

connect which two nodes?



Definition 10. **(Assumption 1)** Given a question sentence N with appropriate query graph G, if T is a correct dependency tree of N, the following condition should be satisfied: There is no such three nodes $\{n_1, n_2, n_3\}$ where n_1 connect n_2 in G and $n_3 \in ShortestPath(n_1, n_2)$ in T.

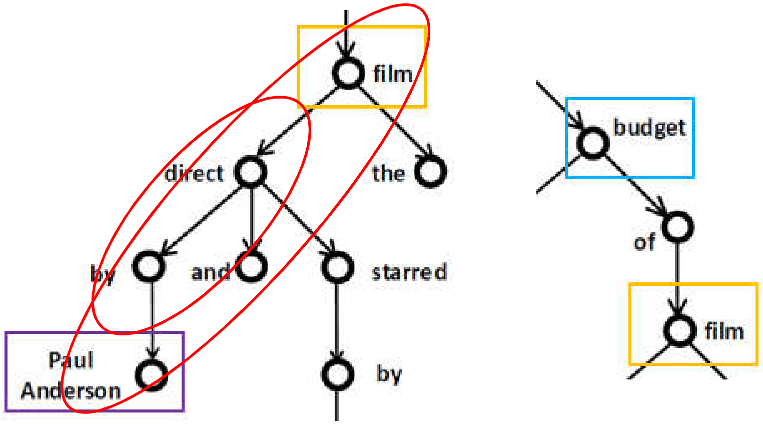
Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]



Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Question Understanding
  - Finding relations

Explicit relation



"directed by"	<director>
"starred by"	<starring>
"budget of"	<budget>

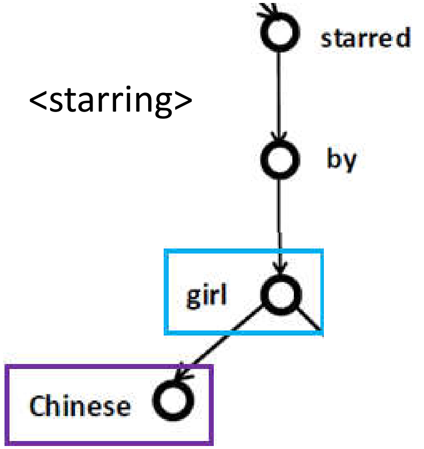
Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Question Understanding
  - Finding relations

Implicit relation

- Locating the two nodes in KG and finding the frequent predicate between them.

<starring>



<China>

Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Query Executing
  - A top-down algorithm

- Naïve method
  (1) Enumerate spanning subgraph of HQG,
  (2) Call algorithm SQG executing algorithm
  (3) Sort and select top-k matches
- Advanced method
  (1) Add <drop, 0> to the candidate list of unsteady edges
  (2) Call algorithm 3

Our Approach- Data Driven & Node-first framework gAnswer+[Hu and Zou et al, TKDE 17]

- Query Executing
  - A top-down algorithm

Drawbacks

- Query graphs with higher scores may have no matches

s	p	o
e_1	p_1	var
…	…	
e_n	p_m	

Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

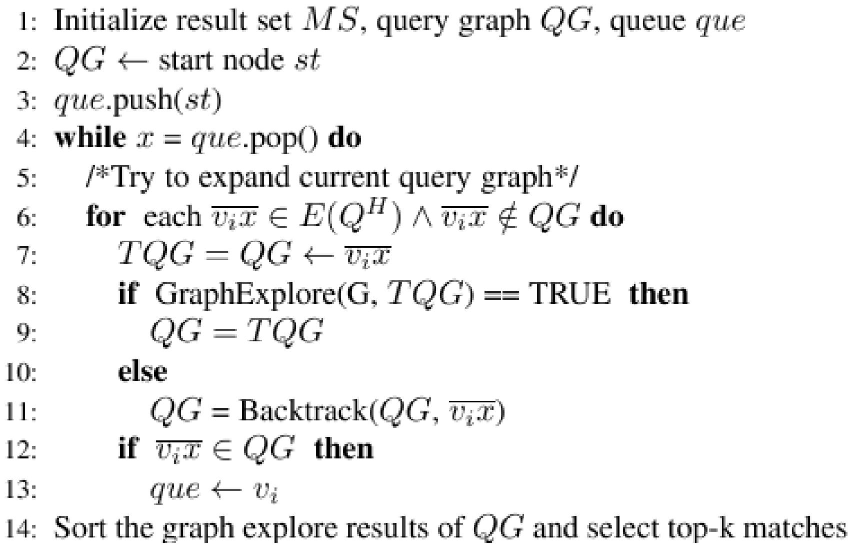
- Query Executing
  - A bottom-up algorithm

Intuition

- Growing structures step by step
- Keep correct structures when growing
- Find matches of multi-label query graph (SQG)
- Drop useless candidates as early as possible

Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Query Executing
  - A bottom-up algorithm



```
Initialize result set $MS$, query graph $QG$, queue $que$
$QG \leftarrow$ start node $st$
$que$.push($st$)
while $x$ = $que$.pop() do
    /*Try to expand current query graph*/
    for each $\overline{v_i x} \in E(Q^H) \wedge \overline{v_i x} \notin QG$ do
        $TQG = QG \leftarrow \overline{v_i x}$
        if GraphExplore(G, $TQG$) == TRUE then
            $QG = TQG$
        else
            $QG$ = Backtrack($QG$, $\overline{v_i x}$)
        if $\overline{v_i x} \in QG$ then
            $que \leftarrow v_i$
Sort the graph explore results of $QG$ and select top-k matches
```

Our Approach- Data Driven & Node-first framework gAnswer+ [Hu and Zou et al, TKDE 17]

- Query Executing
  - A bottom-up algorithm

Optimization

- Call GraphExplore() only when adding unsteady edges
- Design cost model to estimate the best explore order

Experiments

QALD is a series of evaluation campaigns on question answering over linked data.

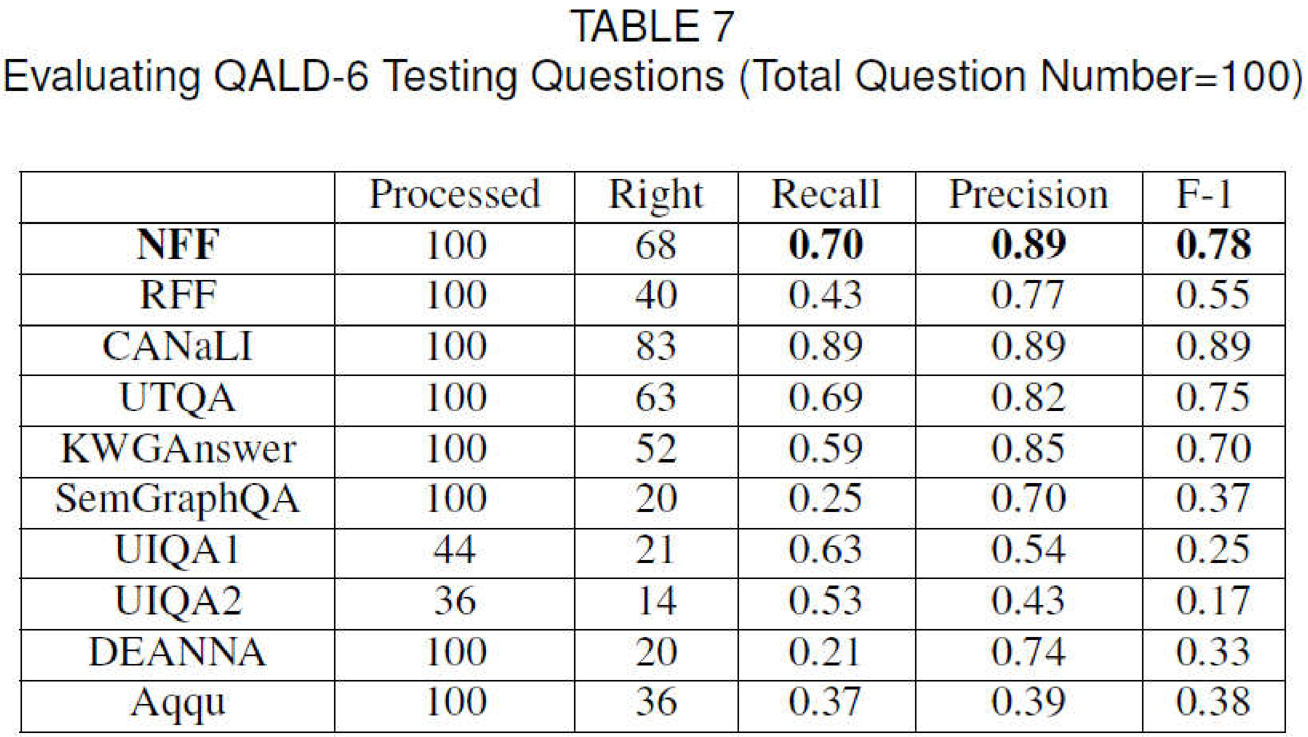


TABLE 7
Evaluating QALD-6 Testing Questions (Total Question Number=100)

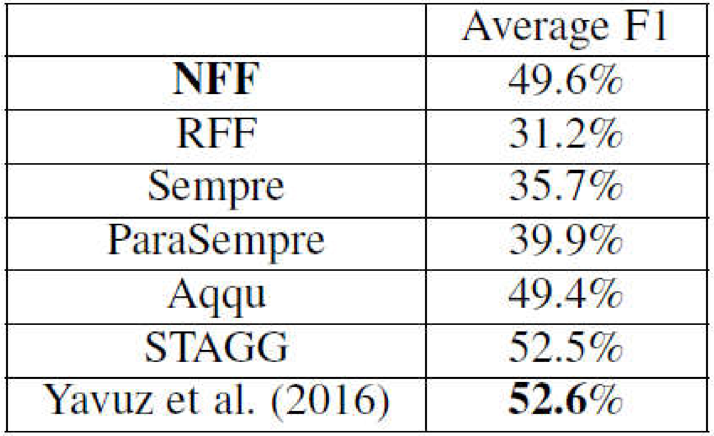
	Processed	Right	Recall	Precision	F-1
NFF	100	68	**0.70**	**0.89**	**0.78**
RFF	100	40	0.43	0.77	0.55
CANaLI	100	83	0.89	0.89	0.89
UTQA	100	63	0.69	0.82	0.75
KWGAnswer	100	52	0.59	0.85	0.70
SemGraphQA	100	20	0.25	0.70	0.37
UIQA1	44	21	0.63	0.54	0.25
UIQA2	36	14	0.53	0.43	0.17
DEANNA	100	20	0.21	0.74	0.33
Aqqu	100	36	0.37	0.39	0.38

QALD-6 Competition Results

Experiments

WebQuestions is widely used in Question Answering literatures and does not contain golden SPARQL queries.

TABLE 8
Evaluating WebQuestions Testing Questions

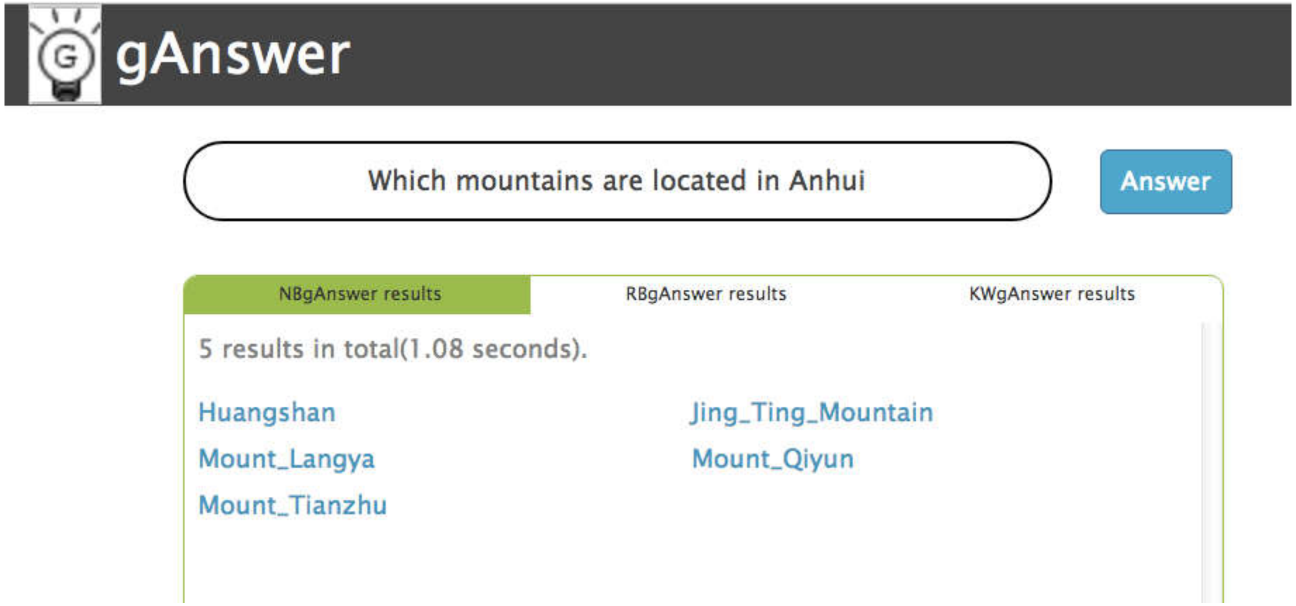


	Average F1
NFF	49.6%
RFF	31.2%
Sempre	35.7%
ParaSempre	39.9%
Aqqu	49.4%
STAGG	52.5%
Yavuz et al. (2016)	**52.6%**

WebQuestions Results

Online Demo

URL: http://ganswer.gstore-pku.com/



Which mountains are located in Anhui

5 results in total(1.08 seconds).

Huangshan

Jing_Ting_Mountain

Mount_Langya

Mount_Qiyun

Mount_Tianzhu

Is it Possible ?

Semantic Parsing （NLP） +Query Evaluation （DB）

Where is the nearest post office?

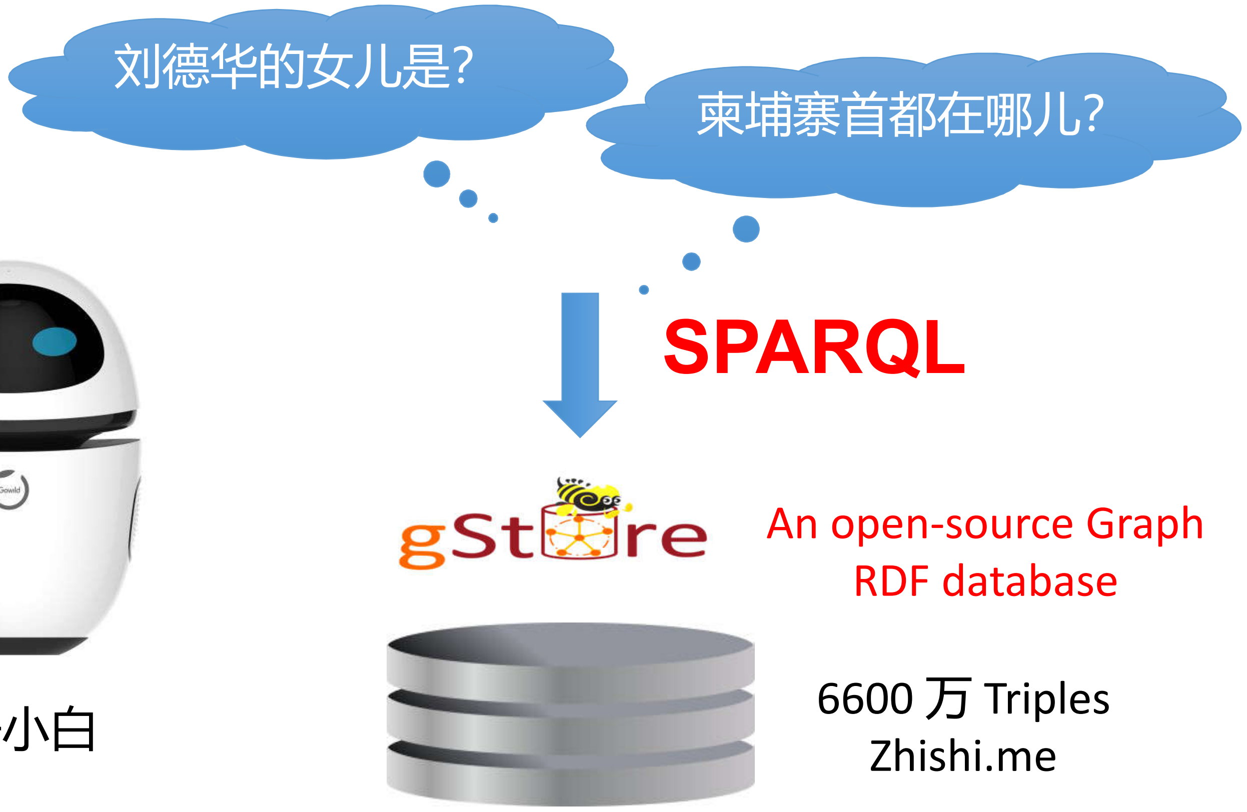
$$\arg\min(\lambda x.POST(x) \wedge dis(HERE, x))$$

SPARQL

```
SELECT ?x WHERE {
?x rdf:type Post.
?x :longitude ?o.
?x :latitude ?a. }
ORDERY BY Dist(HERE, [?o, ?a])
LIMIT 1
```

与深圳狗尾草公司合作.

公子小白



刘德华的女儿是?

柬埔寨首都在哪儿?

SPARQL

gStore

An open-source Graph RDF database

6600 万 Triples
Zhishi.me

参考文献：

- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, Oksana Yakhnenko: Translating Embeddings for Modeling Multi-relational Data. NIPS 2013: 2787-2795
- Luke S. Zettlemoyer, Michael Collins: Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. UAI 2005: 658-666
- Pablo N. Mendes, Max Jakob, Christian Bizer: DBpedia: A Multilingual Cross-domain Knowledge Base. LREC 2012: 1813-1817
- Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum, Yago - A Core of Semantic Knowledge, 16th international World Wide Web conference (WWW 2007)
- Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, Jamie Taylor: Freebase: a collaboratively created graph for structuring human knowledge. SIGMOD Conference 2008: 1247-1250
- Peter Buneman, Gao Cong, Wenfei Fan, Anastasios Kementsietsidis: Using Partial Evaluation in Distributed Query VLDB 2006: 211-222
- Yuk Wah Wong, Raymond J. Mooney: Learning for Semantic Parsing with Statistical Machine Translation. HLT-NAACL 2006
- C. Unger, L. Bühmann, J. Lehmann, A.-C. N. Ngomo, D. Gerber, and P. Cimiano. Template-based question answering over data. In WWW, pages 639–648, 2012
- Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He and Dongyan Zhao, Natural Language Question Answering over RDF ---- A Graph Data Driven Approach , SIGMOD (2014)
- Lei Zou, Jinghui Mo, Lei Chen,M. Tamer Özsu, Dongyan Zhao, gStore: Answering SPARQL Queries Via Subgraph Matching, in Proceedings of 37th International Conference on Very Large Databases (VLDB), 2011.
- Peng Peng, Lei Zou, Tamer Ozsu, Lei Chen, Dongyan Zhao, Processing SPARQL queries over distributed RDF graphs, accepted by VLDB Journal
- Church, A. “A Formulation of the Simple Theory of Types”. Journal of Symbolic Logic 5: 1940. doi:10.2307/2266170

参考文献：

- C. Unger, L. Bühmann, J. Lehmann, A.-C. N. Ngomo, D. Gerber, and P. Cimiano. Template-based question answering over data. In WWW, pages 639–648, 2012
- Lei Zou, Ruizhe Huang, Haixun Wang, Jeffrey Xu Yu, Wenqiang He and Dongyan Zhao, Natural Language Question Answering over RDF ---- A Graph Data Driven Approach , SIGMOD (2014)
- Lei Zou, Jinghui Mo, Lei Chen,M. Tamer Özsu, Dongyan Zhao, gStore: Answering SPARQL Queries Via Subgraph Matching, in Proceedings of 37th International Conference on Very Large Databases (VLDB), 2011.
- Antoine Bordes, Sumit Chopra, Jason Weston: Question Answering with Subgraph Embeddings. EMNLP 2014: 615-620
- William Tunstall-Pedoe: True Knowledge: Open-Domain Question Answering using Structured Knowledge and Inference. AI Magazine 31(3): 80-92 (2010)
- Luke S. Zettlemoyer, Michael Collins: Learning Context-Dependent Mappings from Sentences to Logical Form. ACL/IJCNLP 2009: 976-984
- Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL’14), Baltimore, USA
- Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, Dongyan Zhao: Question Answering on Freebase via Relation Extraction and Textual Evidence. ACL (1) 2016
- Sen Hu, Lei Zou, Jeffrey Xu Yu, Haixun Wang, Dongyan Zhao, Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs, accepted by IEEE Transactions on Knowledge and Data Engineering, 2017
- Wen-tau Yih, Ming-Wei Chang, Xiaodong He, Jianfeng Gao: Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. ACL (1) 2015: 1321-1331
- Li Dong, Mirella Lapata: Language to Logical Form with Neural Attention. ACL (1) 2016

Thanks！

zoulei@pku.edu.cn