

基于深度学习的知识库问答

刘 康

中国科学院自动化研究所
模式识别国家重点实验室
2016年4月17日

历史

Template-based QA
Expert System

IR-based QA

KB-based QA
Community QA

1960

BaseBall
LUNAR
MACSYMA
SHRDLE

1990

MASQUE
TREC



2000

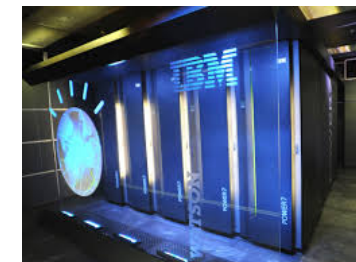


知乎

分享你的知识、经验和见解

搜索问题、话题或人

2010

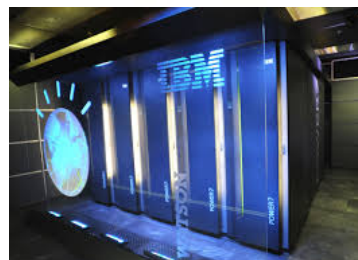


ProBase

问答系统分类

IR-based QA

基于文本检索
+ 信息抽取



Community QA

依赖于网民贡献，问答过程
依赖于检索技术



KB-based QA

Knowledge
Base
语义解析



根据问答形式分类

- 一问一答

姚明个子有多高

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约3,140,000个

搜索工具



姚明身高:

226cm

姚明，1980年生于上海市徐汇区。祖籍吴江震泽。中国篮球运动员。1998年4月，他入选王非执教的国家队，开始篮球生涯。2002年，他以状元秀身份被NBA的休斯敦火箭队选中。20... [详情>>](#)

来自百度百科 | 报错

- 交互式问答




- 阅读理解

Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.
Q: Where was the milk before the den?
A. Hallway

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.
Q: What color is Brian?
A. White

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom

KB-based QA 应用

 **WolframAlpha** computational knowledge engine

Enter what you want to calculate or know about:

how big is China ☆ ☰

[Examples](#) [Random](#)

Assuming "how big" is international data | Use as [referring to socioeconomic data](#) or [referring to species](#) or [referring to administrative divisions](#) instead

Assuming total area | Use [population](#) instead

Input interpretation:

China total area

Result: Show non-metric

$9.597 \times 10^6 \text{ km}^2$ (square kilometers) (world rank: 4th)

Unit conversions:

$9.597 \times 10^{12} \text{ m}^2$ (square meters)

3.705 million mi^2 (square miles)

$1.033 \times 10^{14} \text{ ft}^2$ (square feet)

Comparisons as area:

$\approx 0.96 \times$ total area of Canada ($9.98467 \times 10^6 \text{ km}^2$)

$\approx 0.996 \times$ total area of the United States ($9.63142 \times 10^6 \text{ km}^2$)

\approx largest extent of the Roman Empire ($\approx 9 \text{ Mm}^2$)

姚明个子有多少 🔊

[网页](#) [新闻](#) [贴吧](#) [知道](#) [音乐](#) [图片](#) [视频](#) [地图](#) [文库](#) [更多»](#)

百度为您找到相关结果约3,030,000个 🔍 搜索工具

 姚明身高:
226cm

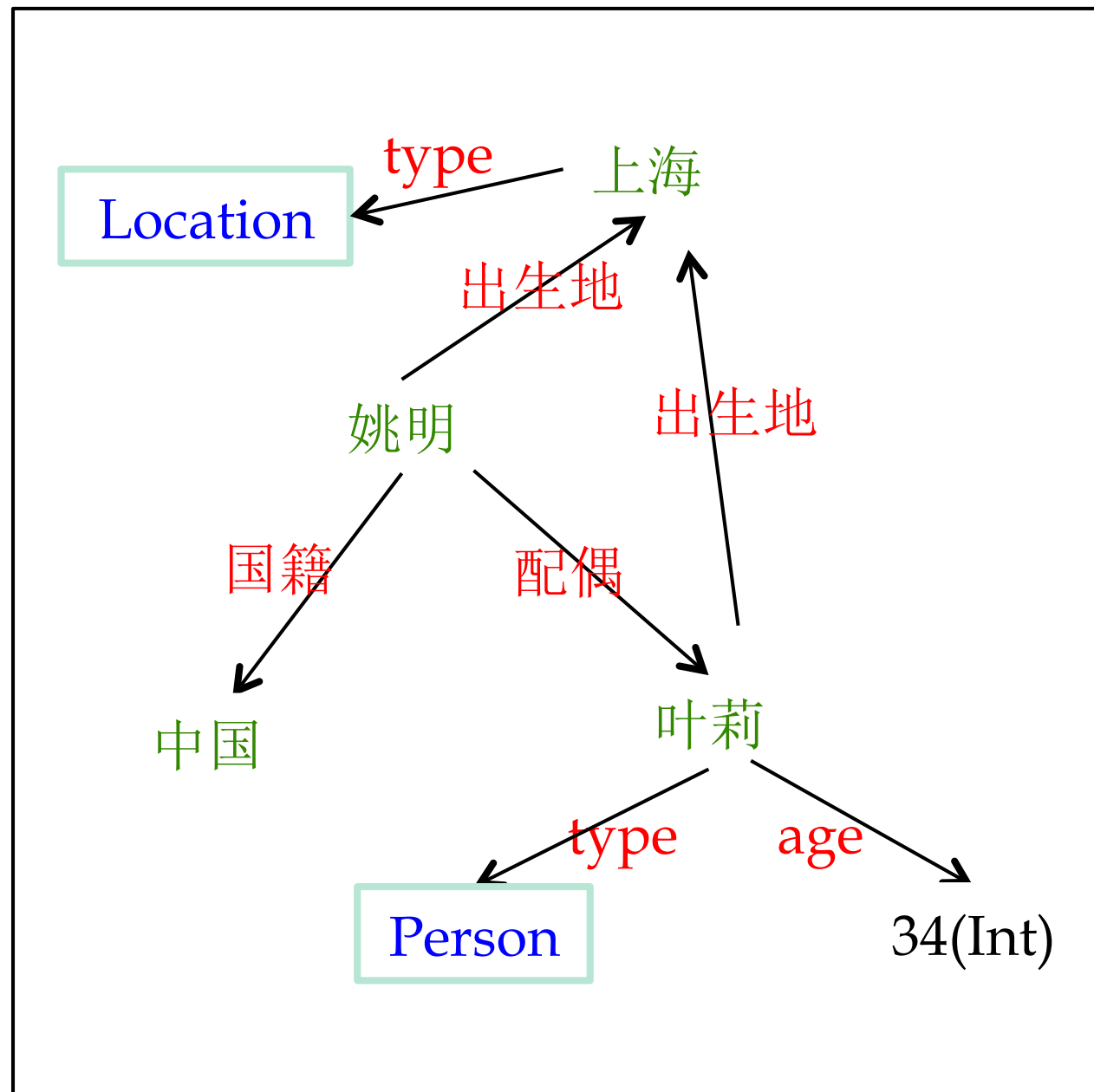
姚明，1980年生于上海市徐汇区，祖籍吴江震泽。中国篮球运动员。1998年4月，他入选王非执教的国家队，开始篮球生涯。2002年，他以状元秀身份被NBA的休斯敦火箭队选中。20... [详情>>](#)

来自百度百科 | [报错](#)

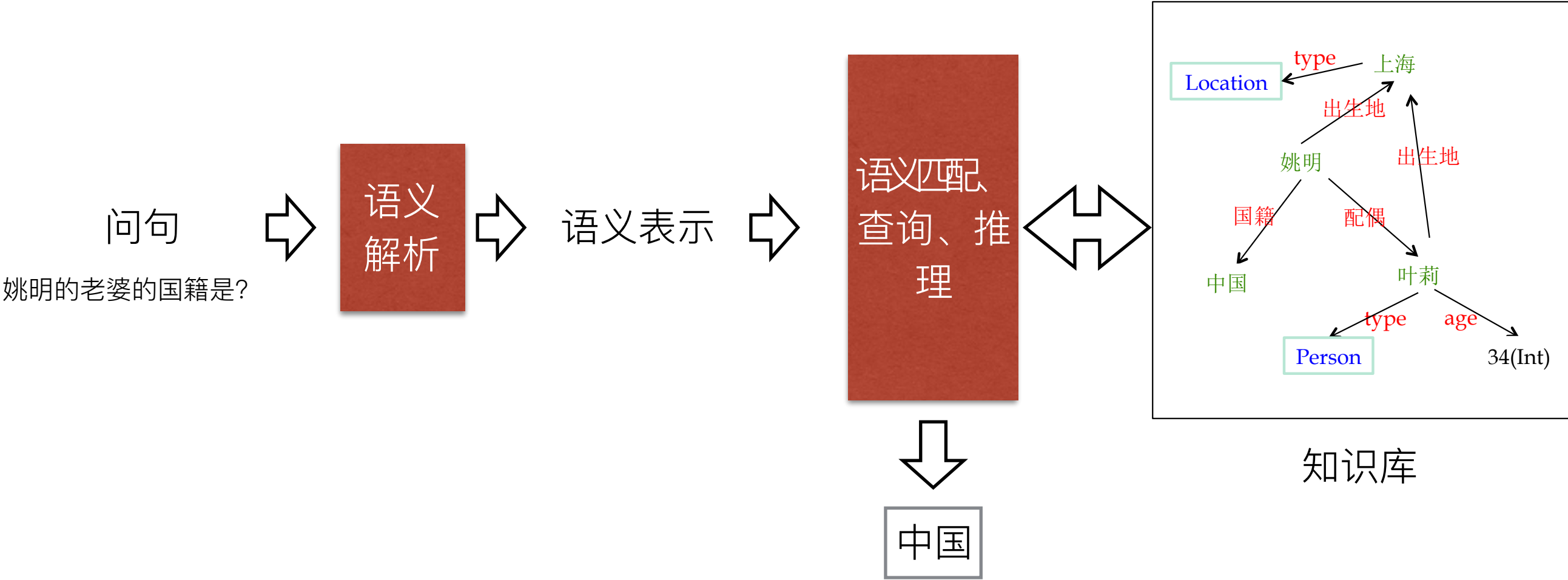
相关人物 展开

 姚沁蕾	 叶莉	 易建联	 迈克尔·乔丹
 沙奎尔·奥尼尔	 勒布朗·詹姆斯	 鲍喜顺	 特雷西·麦克格雷迪
 詹世钗	 梁天云	 方凤娣	 姚德芬

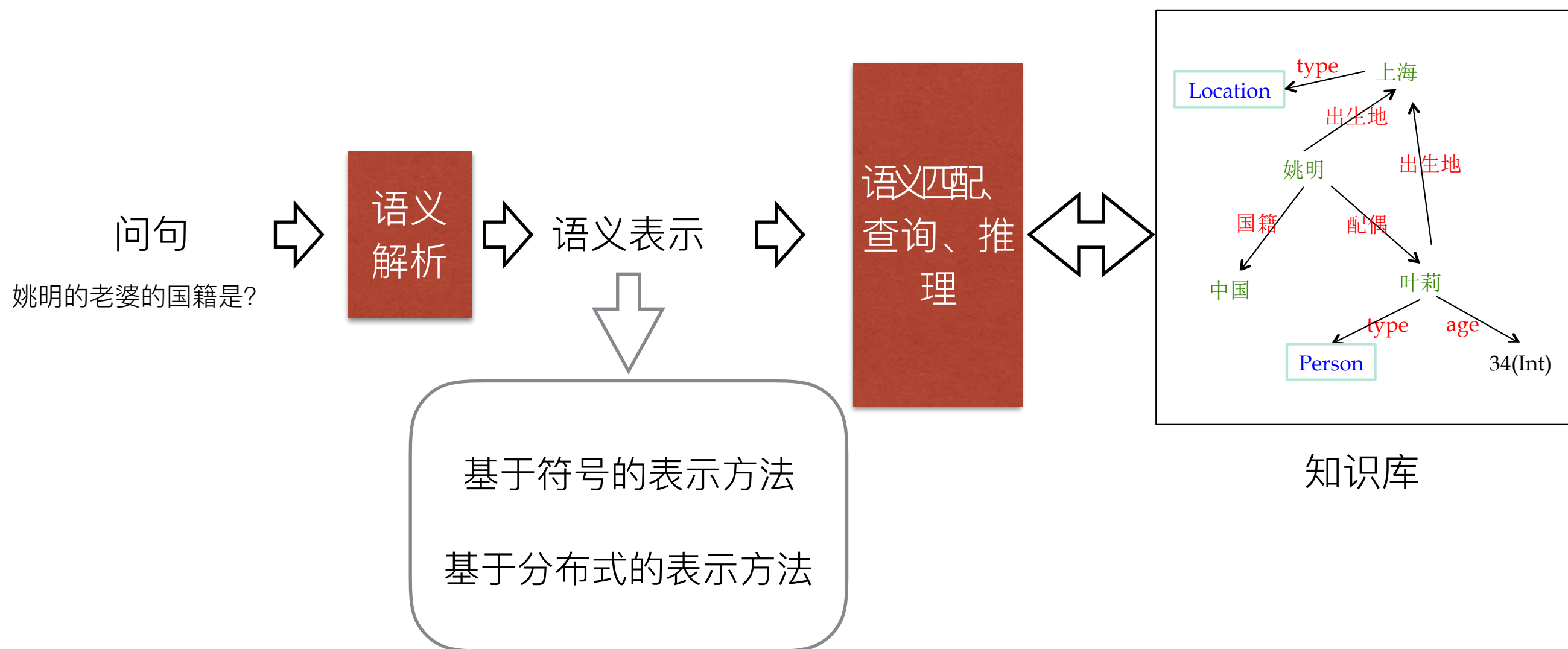
知识库



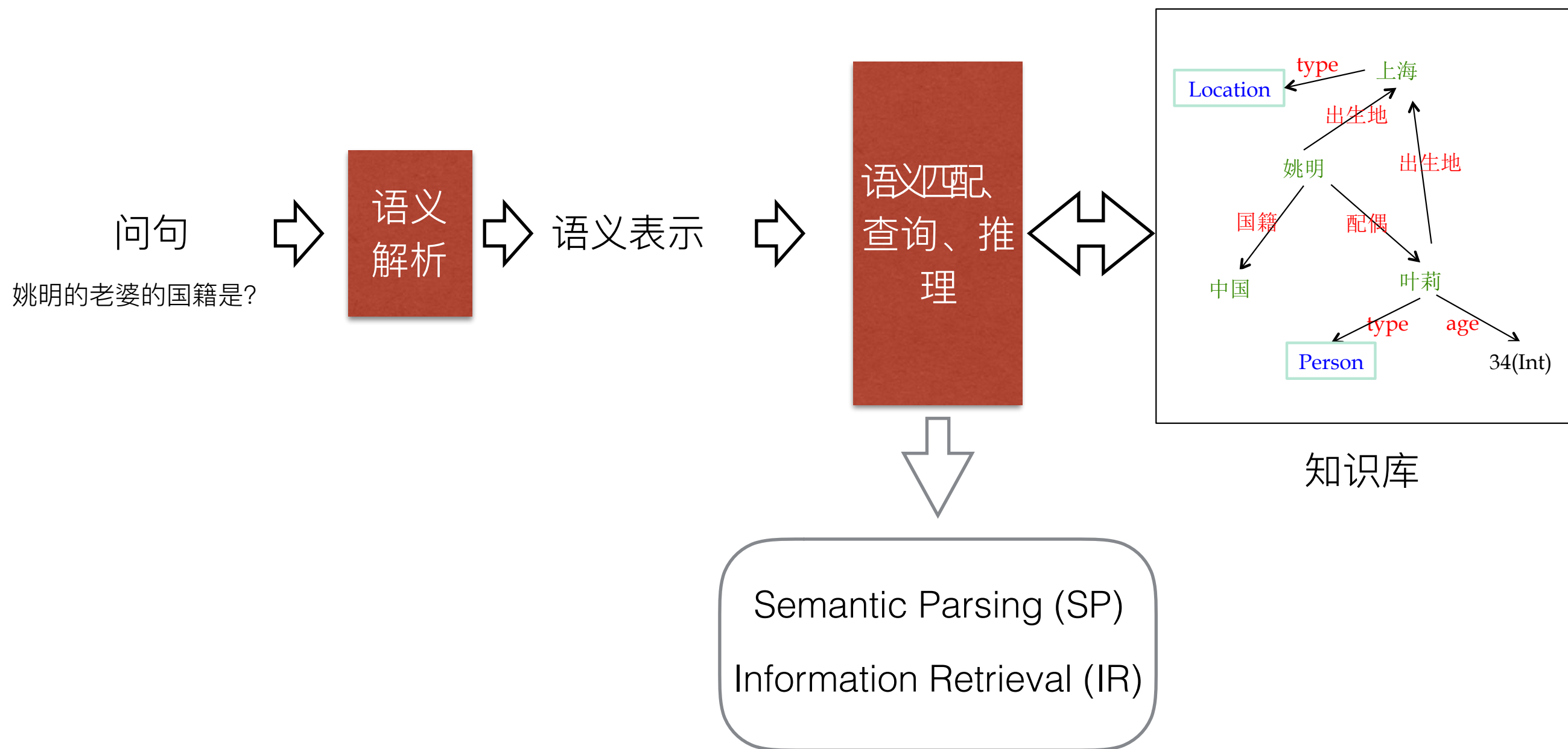
知识库问答



知识库问答



知识库问答



基于符号表示(传统)的知识库问答

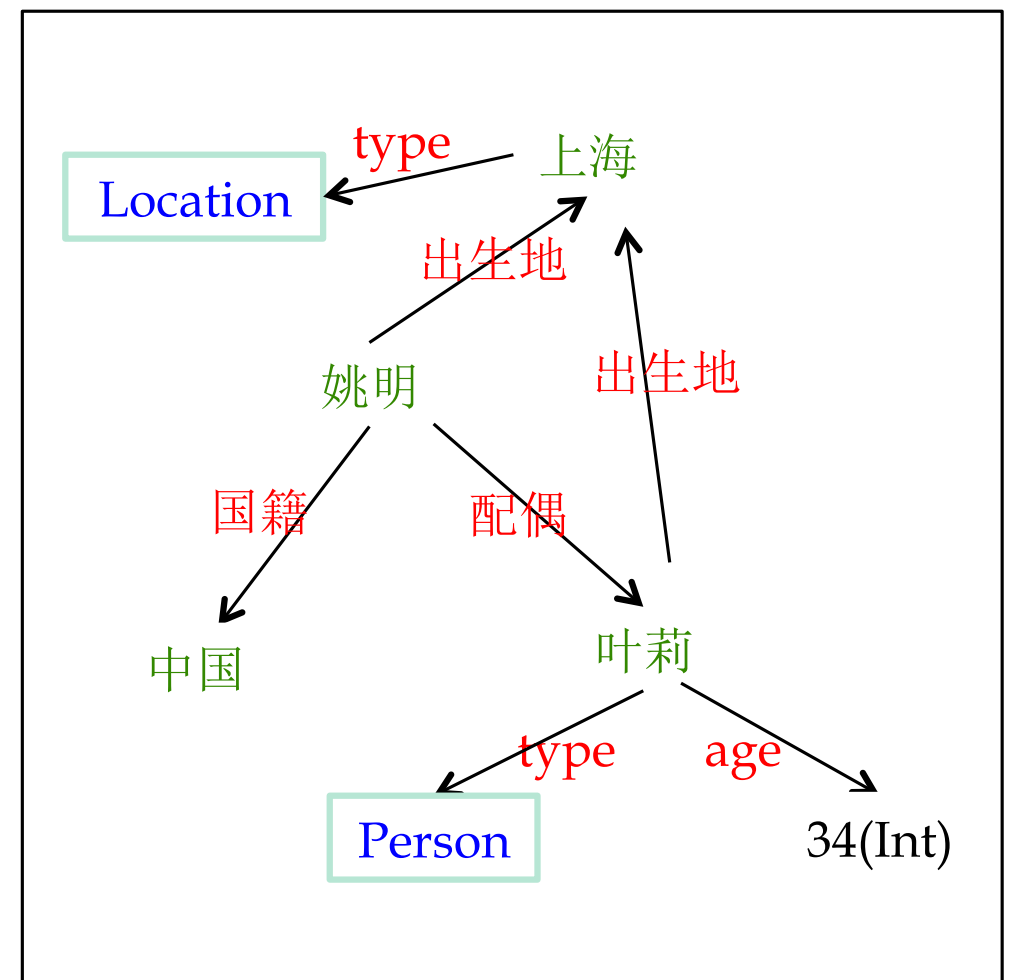
Semantic Parsing

姚明的老婆的国籍是?

语义解析

查询

```
SELECT DISTINCT ?x
WHERE {
  ?y 国籍 ?x.
  res:姚明 配偶 ?y.
}
```



问句的形式化表示

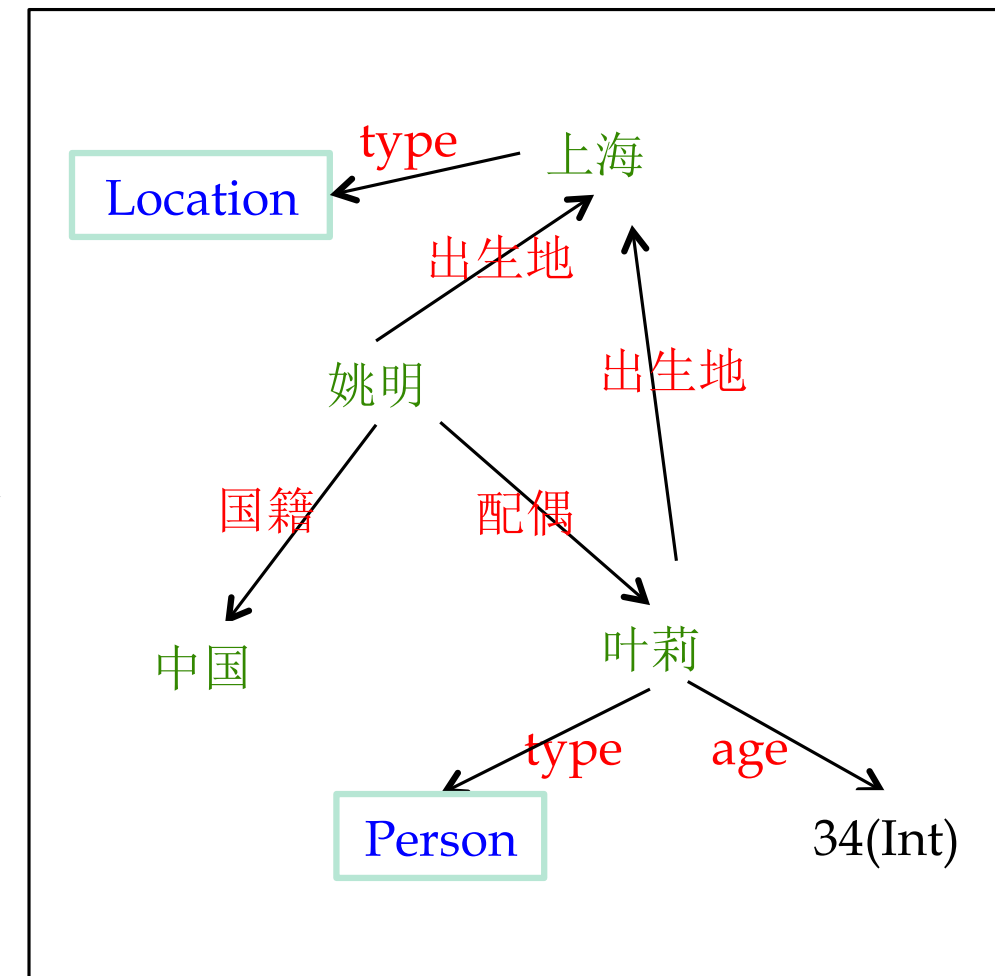
姚明的老婆出生在哪里？

$\lambda x. \text{配偶}(\text{姚明}, y) \wedge \text{出生地}(y, x)$

逻辑表达式
Lambda Calculus
DCS-Tree
Fun-QL
...

```
SELECT DISTINCT ?x
WHERE {
  ?y 出生地 ?x.
  res:姚明 配偶 ?y.
}
```

SQL
SPARQL
Prolog
FunQL
...



任务

What states borders Texas

relation mention

relation mention

entity mention

$\lambda x.state(x) \wedge borders(x, texas)$

Relation

Relation

entity

KB



已有语义解析方法

- 语义解析 (Semantic Parsing)
 - 组合范畴语法 (Combinatory Categorical Grammars) [Zettlemoyer, 1995]
 - “移位-规约”推导 (Shift-reduce Derivations) [Zelle, 1995]
 - 同步语法 (Synchronous Grammars) [Wong, 2007]
 - 混合树 (Hybrid Tree) [Lu, 2008]
 - 类CFG语法 (CFG-like Grammars) [Clarke, 2010]
 - 类CYK方法 (CYK-like Grammars) [Liang, 2011]

基本过程

Which software has been developed by organizations founded in California, USA?

短语检测:

software

developed by

organizations

founded in

California

资源映射:

dbo:Software

dbr:developer

dbo:Company

dbr:foundationPlace

dbo:California

语义组合:

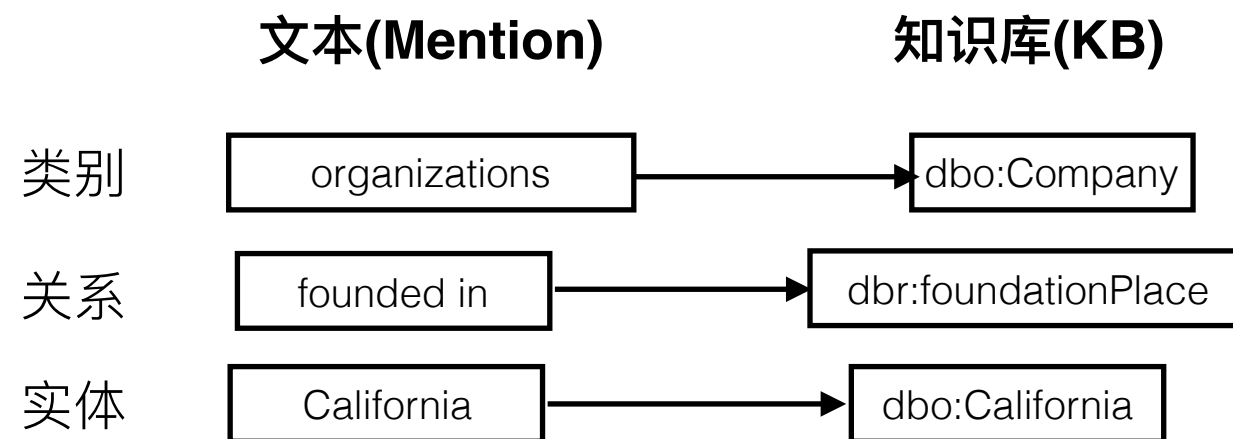
<dbo:Software, dbr:developer, dbo:Company>

<dbo:Company, dbr:foundationPlace, dbo:California>

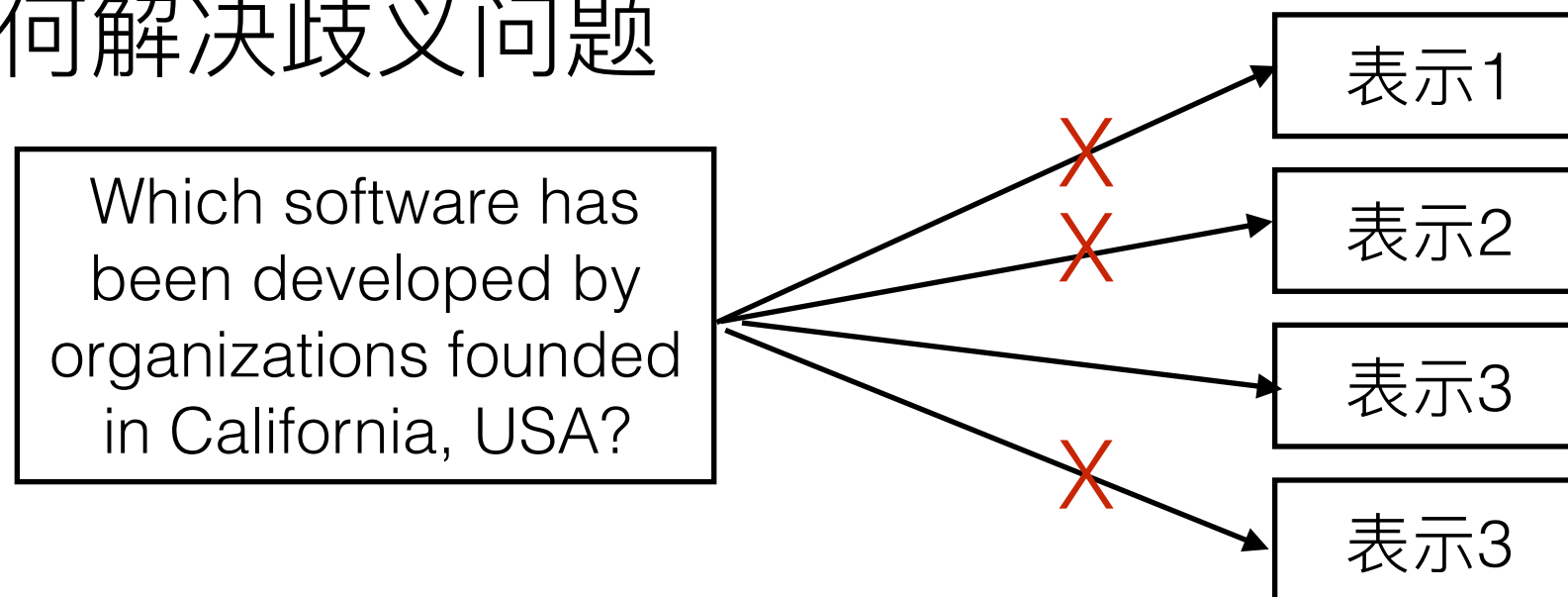
逻辑表达式

两个关键问题

- 获得短语到资源的映射

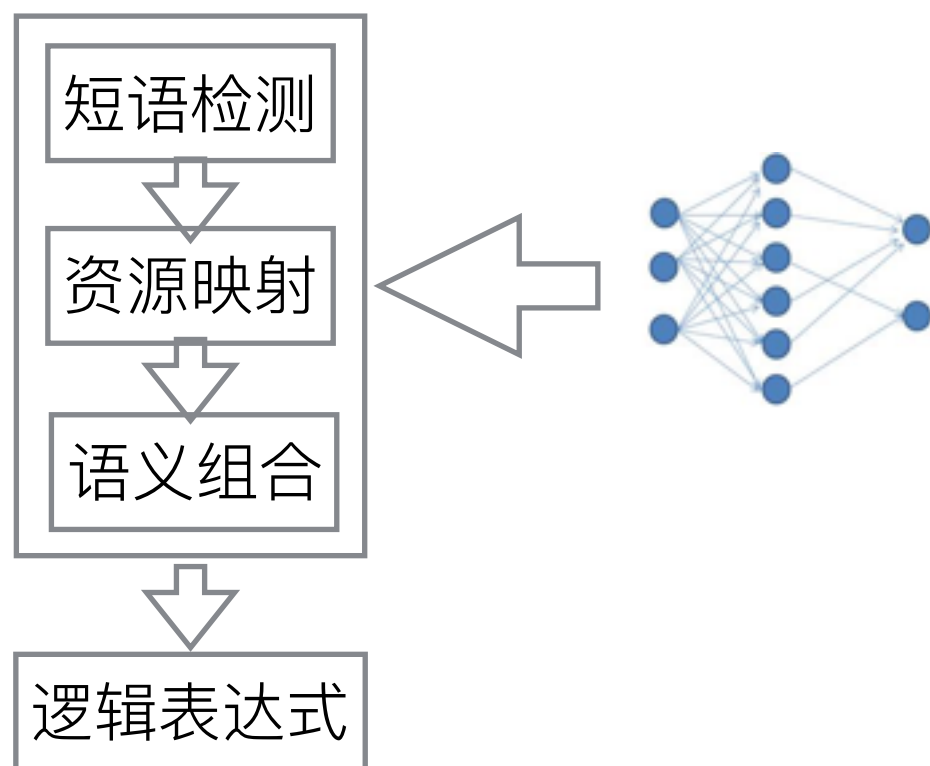


- 如何解决歧义问题



Deep Learning对于传统方法的改进

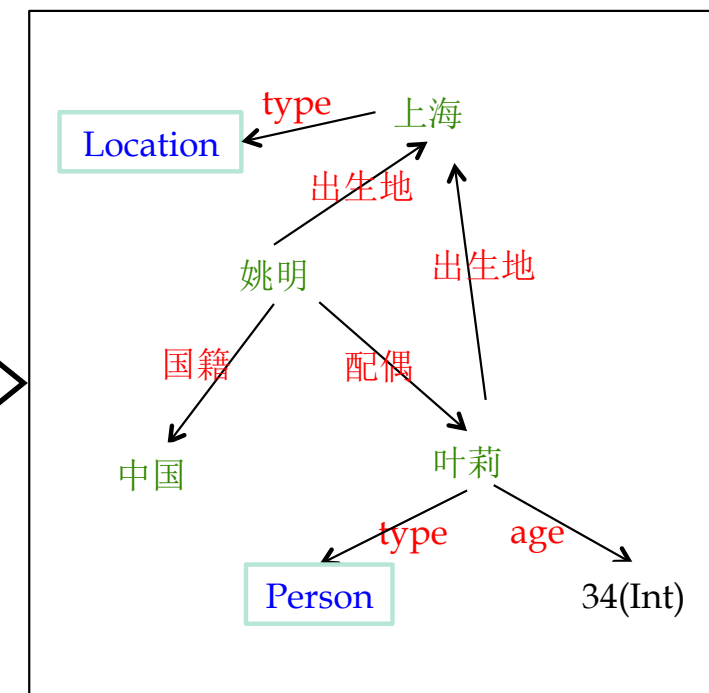
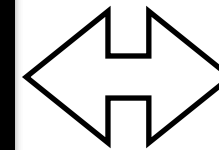
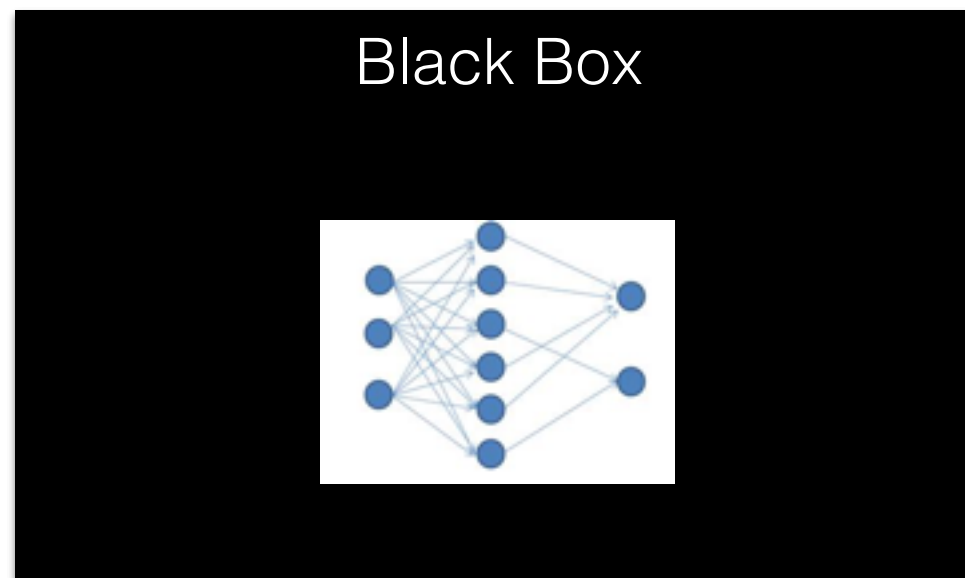
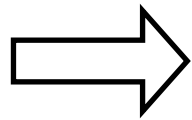
- 利用深度学习对于传统问答方法的改进



- DL-based关系识别
 - Yin et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Bas, In Proceedings of ACL 2015 (**Outstanding Paper**)
 - Zeng et al. Relation Classification via Convolutional Deep Neural Network, in Proceedings of COLING 2014 (**Best Paper**)
 - Zeng et al, Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks, in Proceedings of EMNLP 2015
 - Xu et al. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths, In Proceedings of EMNLP 2015

基于分布式表示(DL,End2End)的知识库问答

姚明的老婆的国籍是?

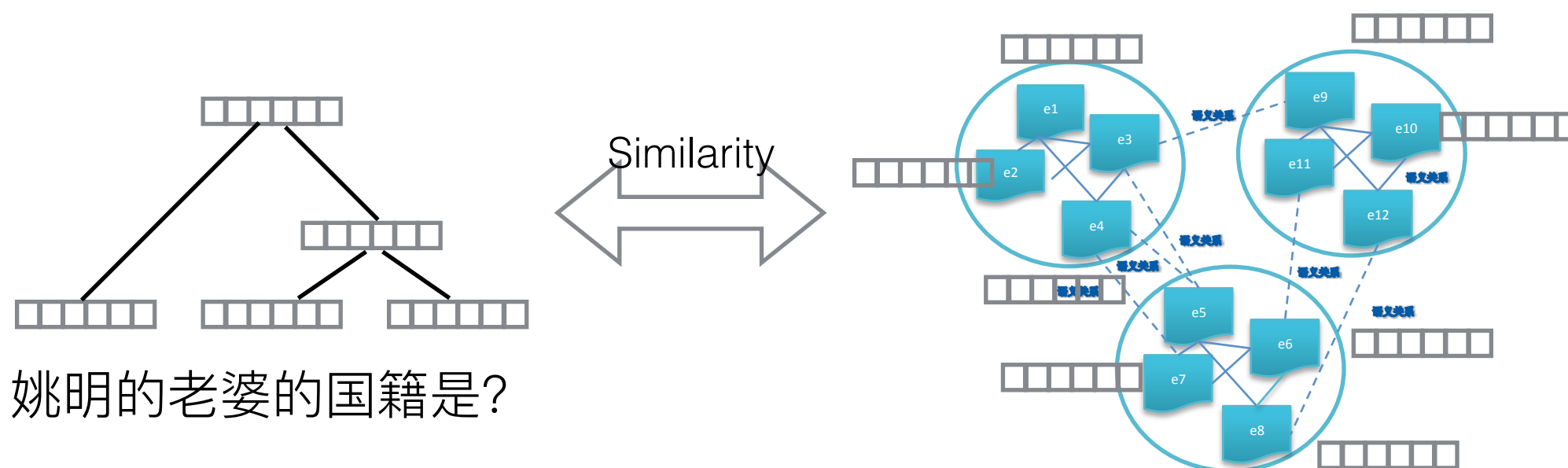


知识库



中国

基于分布式表示(DL,End2End)的知识库问答 IR based Approach



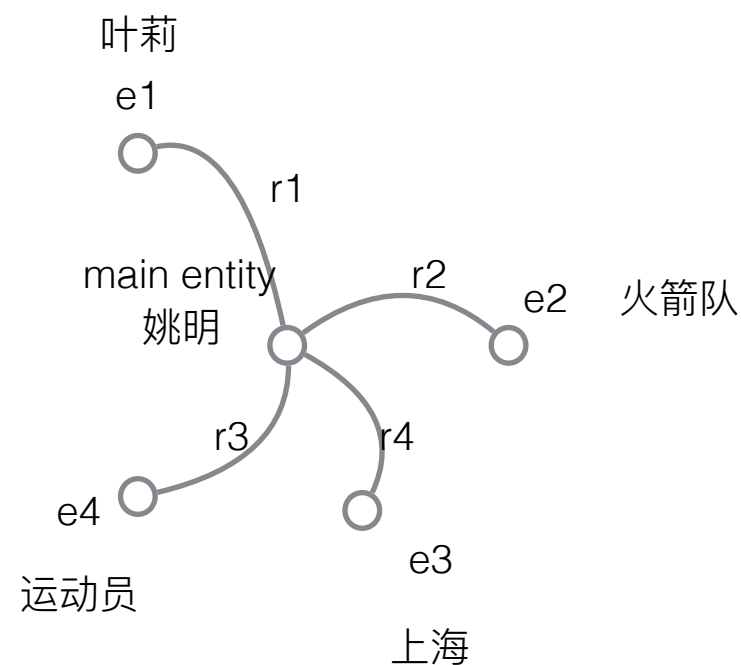
Progress

- Bordes et al. Open Question Answering with Weakly Supervised Embedding Models, In Proceedings of ECML-PKDD 2014
 - Basic System
- Bordes et al. Question Answering with Subgraph Embedding, In Proceedings of EMNLP 2014
 - Contextual Information of Answers
- Yang et al. Joint Relational Embeddings for Knowledge-Based Question Answering, In Proceedings of EMNLP 2014
 - Entity Type
- Dong et al. Question Answering over Freebase with Multi-Column Convolutional Neural Network, In Proceedings of ACL 2015
 - Topic Entity、Relation Path、Contextual Information
- Bordes et al. Large-scale Simple Question Answering with Memory Network, In Proceedings of ICIR 2015
 - Memory Network

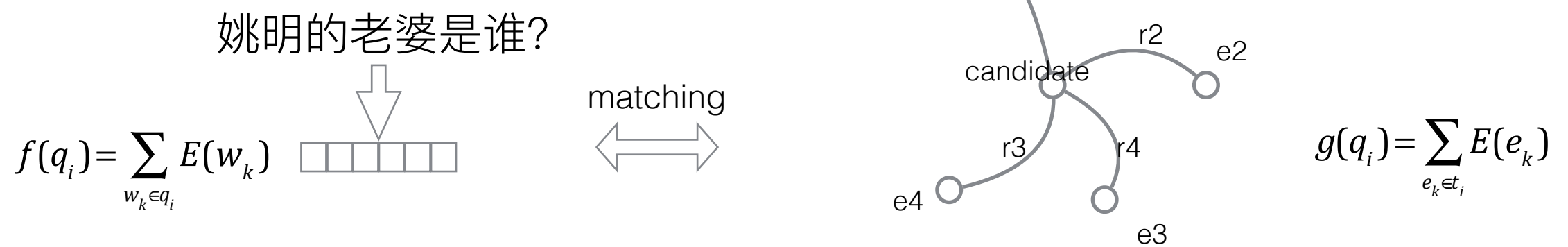
Basic End2End QA System (Bordes et al. 2014)

- 目前只处理单关系 (Single Relation) 的问句 (Simple Question)
- 基本步骤
 - Step1: 候选生成
 - 利用Entity Linking找到main entity
 - 在KB中main entity周围的entity均是候选
 - Step2: 候选排序

姚明的老婆是谁?



Basic End2End QA System (Bordes et al. 2014)



Object: $L = 0.1 - S(f(q), g(t)) + S(f(q), g(t'))$

$$S(f(q), g(t)) = f(q)^T g(t)$$

$$S(f(q), g(t)) = f(q)^T M g(t)$$

Multitask learning with paraphrases:

$$S_p(f(q), f(q_p)) = f(q)^T f(q_p)$$

Considering More Info (Bordes et al. 2014)

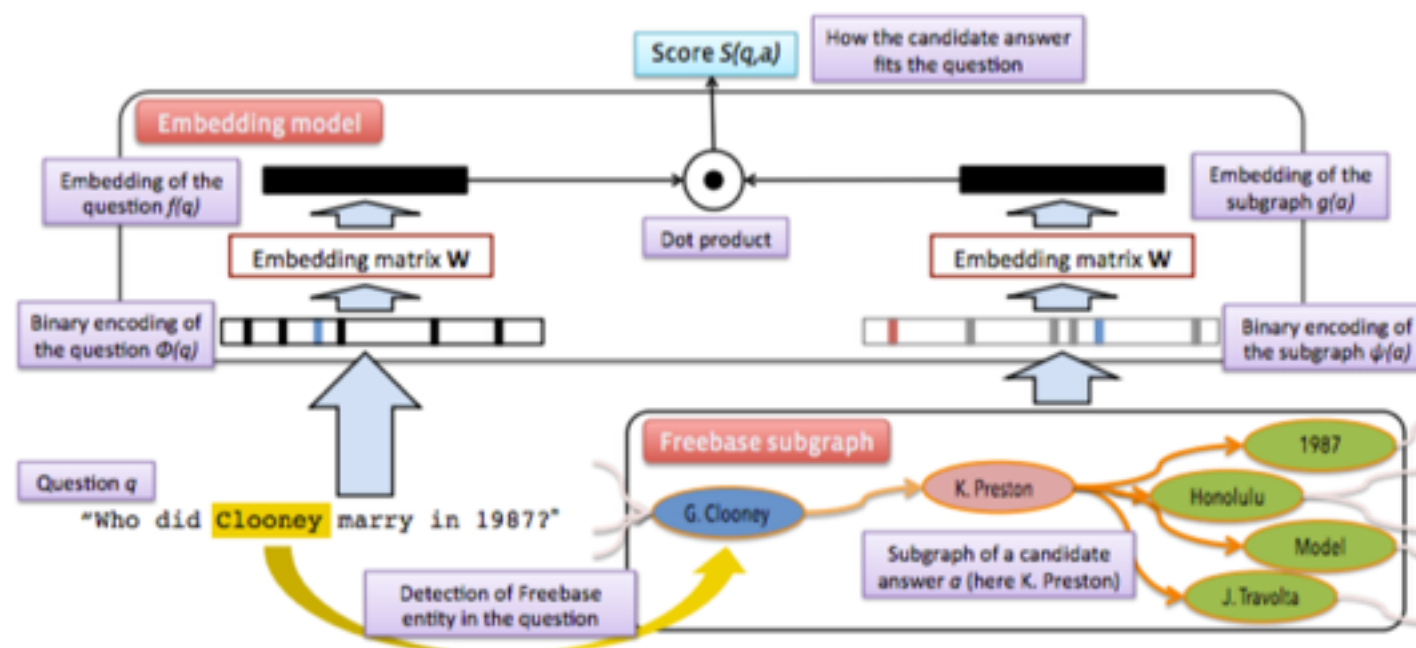
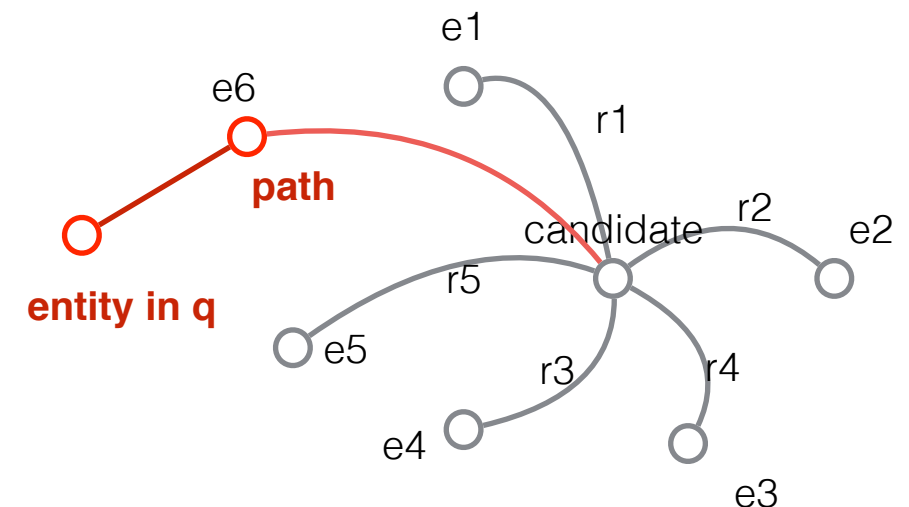
- Entity Information $E(e_k)$

- Path to entities in question

$$g(q_i) = \sum_{e_k \in \text{Path}(t_i)} E(e_k) + \sum_{r_k \in \text{Path}(t_i)} E(r_k)$$

- Subgraph of the answers (contextual Info)

$$g(q_i) = \sum_{e_k \in \text{Context}(t_i)} E(e_k) + \sum_{r_k \in \text{Context}(t_i)} E(r_k)$$



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

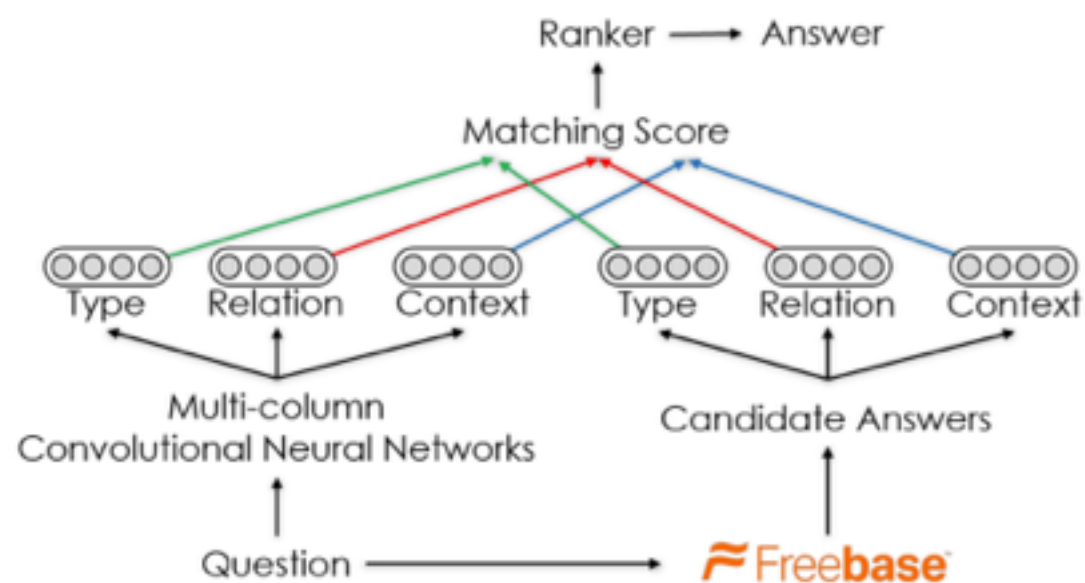
$$S(q, a) = f(q)^T g(a)$$

Results

Method	P@1 (%)	F1 (Berant)	F1 (Yao)
Baselines			
(Berant et al., 2013) [1]		31.4	
(Bordes et al., 2014) [5]	31.3	29.7	31.8
(Yao and Van Durme, 2014) [14]	—	33.0	42.0
(Berant and Liang, 2014) [2]	—	39.9	43.0
Our approach			
Subgraph & $\mathcal{A}(q) = C_2$	40.4	39.2	43.2
Ensemble with (Berant & Liang, 14)	—	41.8	45.7
Variants			
Without multiple predictions	40.4	31.3	34.2
Subgraph & $\mathcal{A}(q) = \text{All 2-hops}$	38.0	37.1	41.4
Subgraph & $\mathcal{A}(q) = C_1$	34.0	32.6	35.1
Path & $\mathcal{A}(q) = C_2$	36.2	35.3	38.5
Single Entity & $\mathcal{A}(q) = C_1$	25.8	16.0	17.8

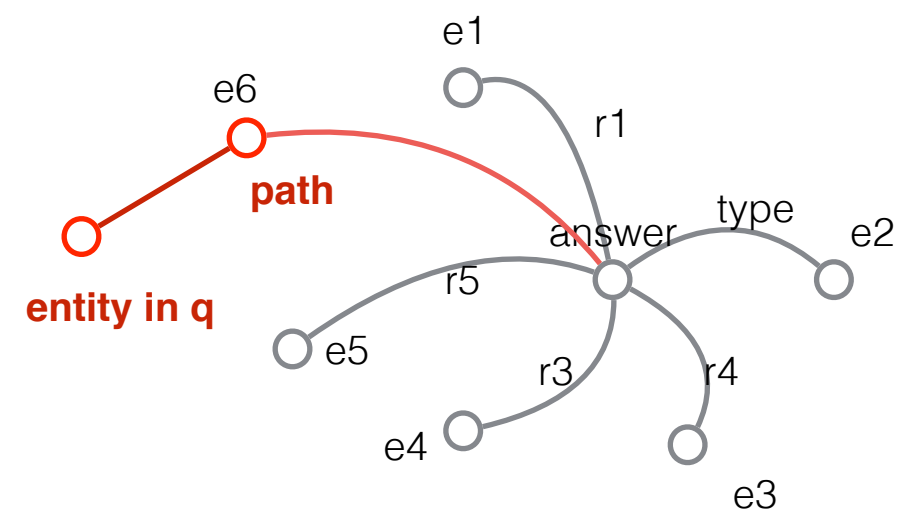
Multi-Column CNN (Dong et al. 2015)

- 依据问答特点，考虑答案不同维度的信息



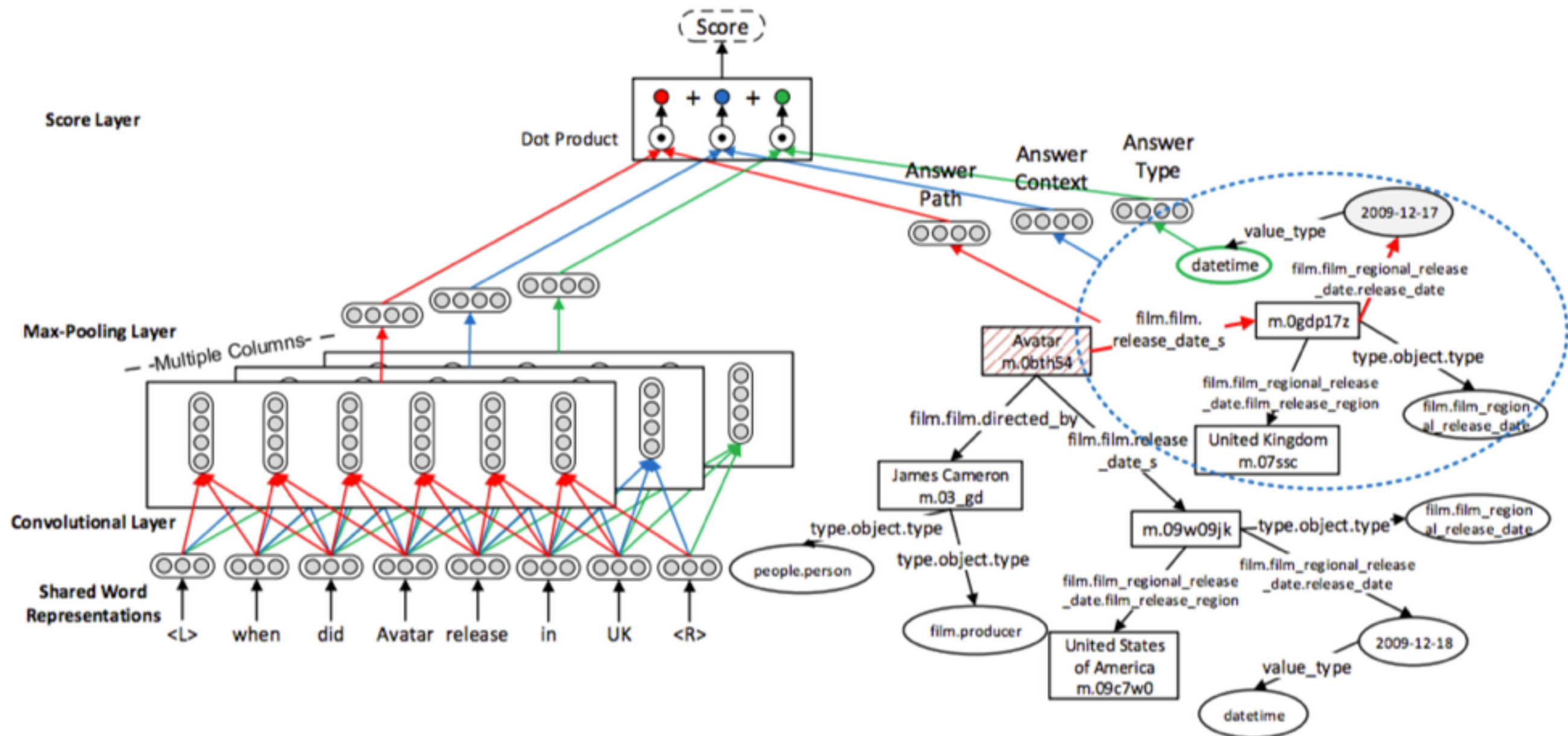
$$S(q, a) = \underbrace{\mathbf{f}_1(q)^T \mathbf{g}_1(a)}_{\text{answer path}} + \underbrace{\mathbf{f}_2(q)^T \mathbf{g}_2(a)}_{\text{answer context}} + \underbrace{\mathbf{f}_3(q)^T \mathbf{g}_3(a)}_{\text{answer type}}$$

Answer Type
Answer Context
Answer Path



Framework

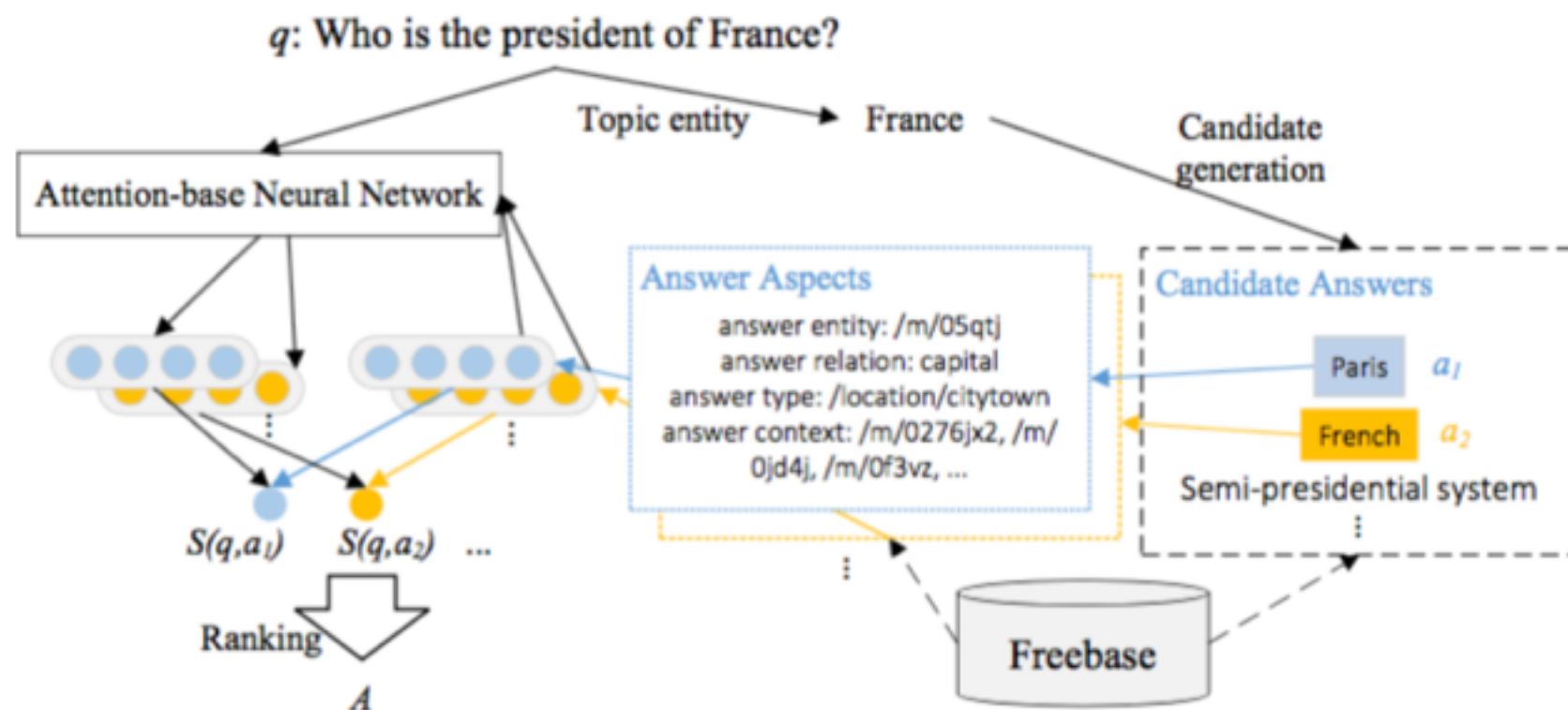
when did Avatar release in UK



Results

Method	F1	P@1
(Berant et al., 2013)	31.4	-
(Berant and Liang, 2014)	39.9	-
(Bao et al., 2014)	37.5	-
(Yao and Van Durme, 2014)	33.0	-
(Bordes et al., 2014a)	39.2	40.4
(Bordes et al., 2014b)	29.7	31.3
MCCNN (our)	40.8	45.1

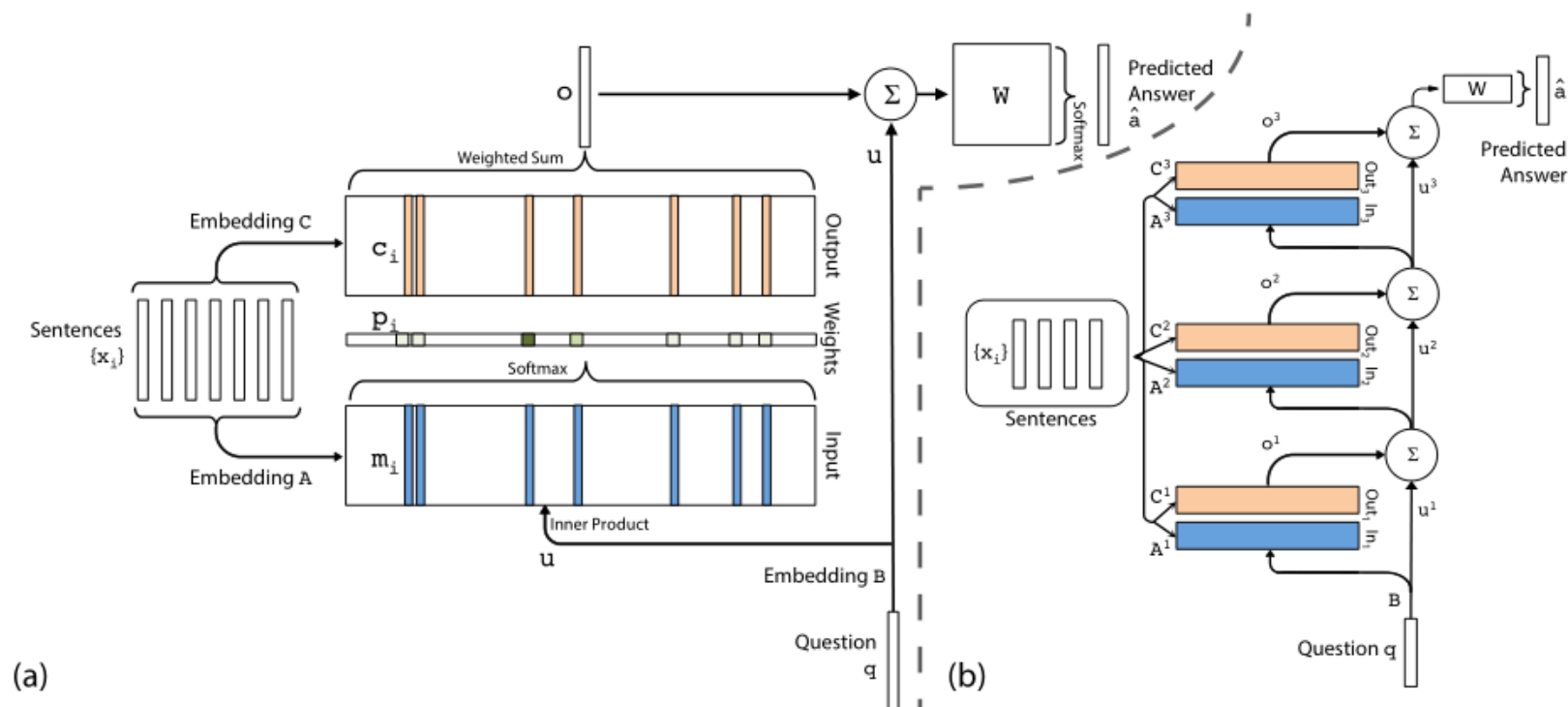
Neural Attention-based Neural Model for QA with Combining Global Knowledge Information (Zhang et al. 2016)



Method	F ₁
Bordes et al., 2014b	29.7
Bordes et al., 2014a	39.2
Yang et al., 2014	41.3
Dong et al., 2015	40.8
Bordes et al., 2015	42.2
ours	42.6

Memory Network

- 4 Components:
 - I (Input feature map): 将输入转换为内部的特征表示
 - G (Generalization): 根据输入更新Memory
 - O (Output feature map): 根据输入和当前Memory状态, 生成输出 (在特征空间中)
 - R (Response): 根据输出的特征表示, 生成答案



过程

- Input Component

- 知识库:

- 三元组 (head, r, tail)

- 三元组表示:

$$V(triple) = \frac{1}{N} \sum_{item \in triple} V(item)$$

- 问句:

- 句子中出现的词（与知识库中的实体、关系可以链接）的词向量的平均

$$V(q) = \frac{1}{N} \sum_{w \in q} V(w)$$

- Generalization Component

- 加入其它知识库信息（知识库对齐），更新Memory

- 相似度计算、词典等

过程

- Output Component

$$S_{QA}(q, y) = \cos(\mathbf{W}_V g(q), \mathbf{W}_S f(y))$$

- Response Component
 - 输出三元组的object

训练

- Multitask Learning

$$\ell_{QA}(q, y, y') = [\gamma - S_{QA}(q, y) + S_{QA}(q, y')]_+$$

$$\ell_{QQ}(q, q', q'') = [\gamma - S_{QQ}(q, q') + S_{QQ}(q, q'')]_+$$

实验结果

						WebQuestions F1-SCORE (%)	SimpleQuestions ACCURACY (%)	Reverb ACCURACY (%)
BASELINES								
Random guess						1.9	4.9	35
(Berant et al., 2013)						31.3	n/a	n/a
(Fader et al., 2014)						n/a	n/a	54
(Bordes et al., 2014b)						29.7	n/a	73
(Bordes et al., 2014a) – <i>using path</i>						35.3	n/a	n/a
(Bordes et al., 2014a) – <i>using path + subgraph</i>						39.2	n/a	n/a
(Berant and Liang, 2014)						39.9	n/a	n/a
(Yang et al., 2014)						41.3	n/a	n/a
(Weston et al., 2015) – <i>the original MemNN</i>						n/a	n/a	72
MEMORY NETWORKS (<i>never trained on Reverb – only transfer</i>)								
KB	TRAIN SOURCES			CANDS	ENSEMBLE			
	WQ	SIQ	PRP	AS NEGS				
FB2M	yes	yes	yes	–	–	36.2	62.7	n/a
FB5M	–	–	–	–	–	18.7	44.5	52
FB5M	–	–	yes	–	–	22.0	48.1	62
FB5M	–	yes	–	–	–	22.7	61.6	52
FB5M	–	yes	yes	–	–	28.2	61.2	64
FB5M	yes	–	–	–	–	40.1	46.6	58
FB5M	yes	–	yes	–	–	40.4	47.4	61
FB5M	yes	yes	–	–	–	41.0	61.7	52
FB5M	yes	yes	yes	–	–	41.0	62.1	67
FB5M	yes	yes	yes	yes	–	41.2	62.2	65
FB5M	yes	yes	yes	yes	5 models	41.9	63.9	68
FB5M	yes	yes	yes	yes	Subgraph	42.2	62.9	62

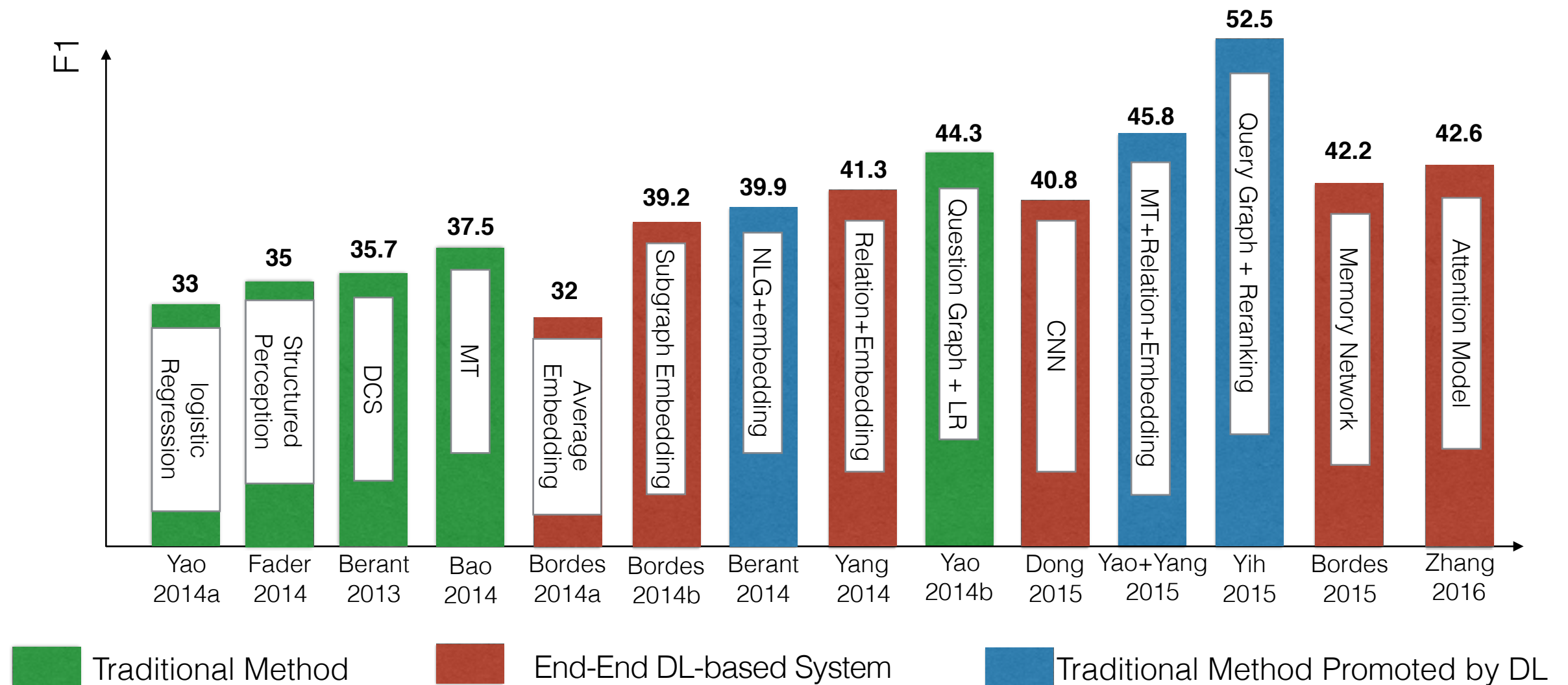
更加适合阅读理解

Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.
Q: Where was the milk before the den?
A. Hallway

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.
Q: What color is Brian?
A. White

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom

Comparison in Benchmark of WebQuestion



Problems

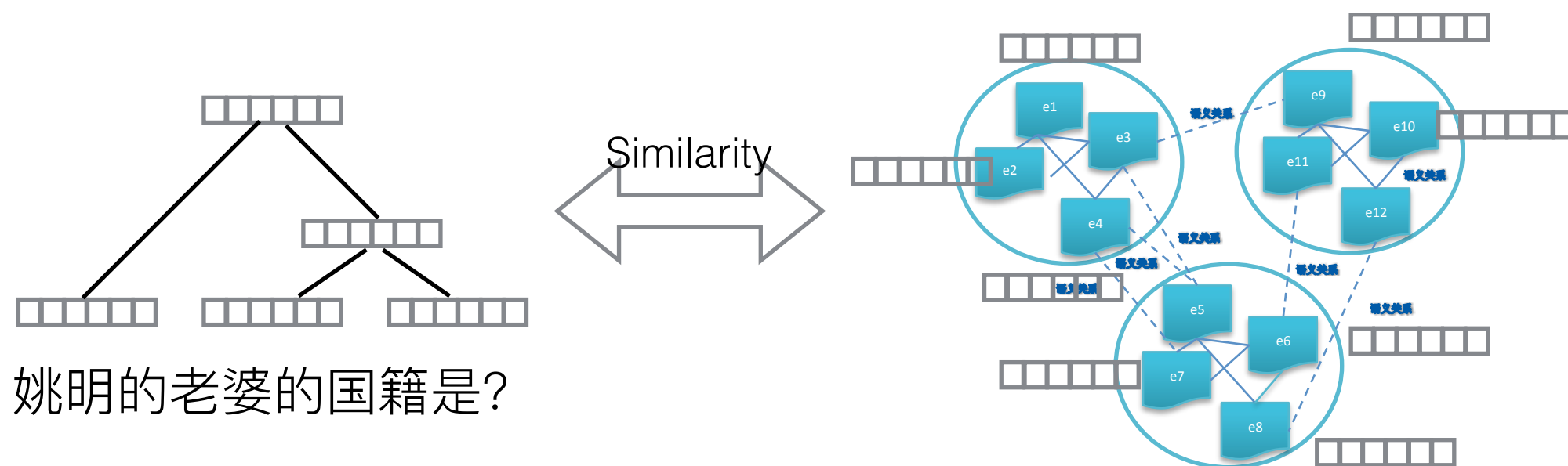
复杂问句

- 目前，End2End Model 只能解决Simple（单关系）类型的问题
- 面对复杂问句（包含多种关系）

Which song **are performed by** person who **was born in** New York and **played a role** in Valentine's Day

训练语料

- 知识库中实体、关系embedding的学习与训练语料密切相关
- 训练语料之外的实体、关系、词？



$$L = 0.1 - S(f(q), g(t)) + S(f(q), g(t'))$$

面对多知识库如何进行问句语义解析

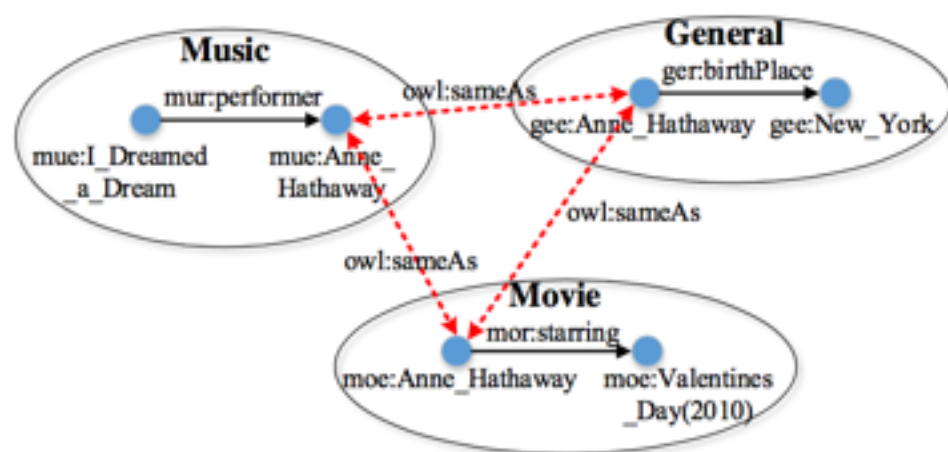
- 开放域环境下，用户的问题复杂多样，很多场景下，单单只用一个知识库的信息不能完全回答用户的问题

Which song are performed by person who was born in New York and played a role in Valentine's Day

music KB

people KB

movie KB



```
SELECT ?v1 WHERE {  
  <?v1, mur:performer1, ?v2>  
  <?v2, owl:sameAs, ?v3>  
  <?v3, mor:starring, moc:Valentines_Day(2010)>  
  <?v3, owl:sameAs, ?v4>  
  <?v4, ger:birthPlace, gee:New_York> }
```

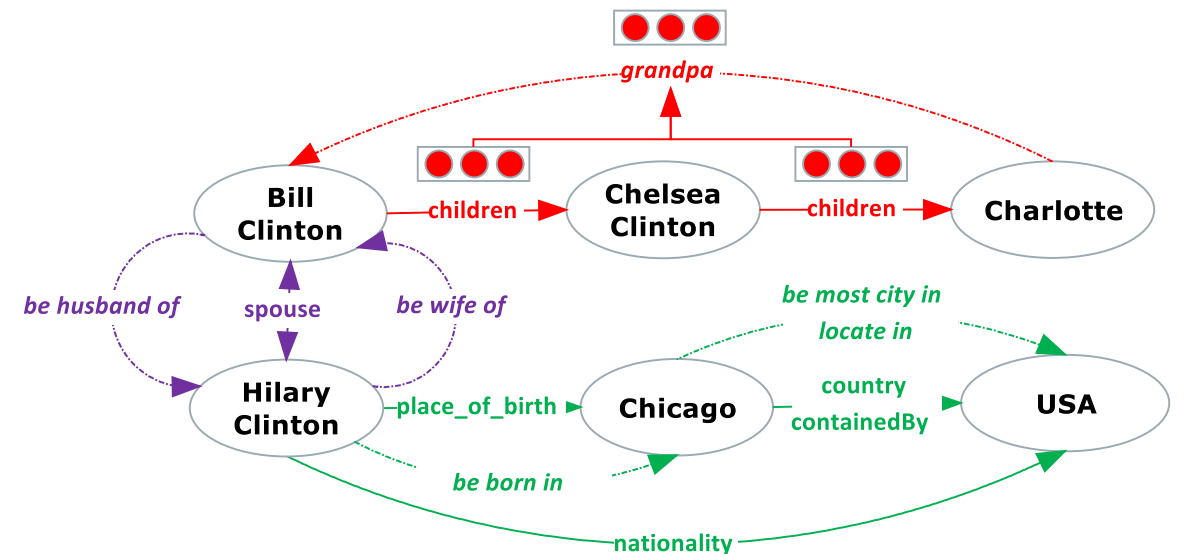
- 多知识库间冗余、异构
- 需要对齐

问题

- 基于End2End框架的多源异构知识库问答
- 如何利用DL进行多源异构知识库的对齐
- 如何在DL框架下，进行多源异构知识库对齐、问句语义解析的联合学习

大规模知识推理

- Link Prediction
 - 缺乏显式的推理解释
 - 召回率高，准确率低
- 大规模推理规则学习



公开的评测数据集

数据集	# 训练集	# 测试集	知识库	形式	发布时间
ATIS	8297	3211	ATIS	答案	1994
Geo880	880		GeoBase	逻辑形式	2001
QALD-1	50	50	DBpedia	逻辑形式 & 答案	2011
QALD-2	100	99	DBpedia & YAGO	逻辑形式 & 答案	2012
QALD-3	100	99	DBpedia & YAGO	逻辑形式 & 答案	2013
Free917	641	276	Freebase	答案	2013
WebQuestion	3782	2037	Freebase	答案	2013
WikiAnswers	2.4M	698	Reverb	答案	2013
QALD-4	100	50	DBpedia & YAGO	逻辑形式 & 答案	2014
QALD-5	170	59	DBpedia & YAGO	逻辑形式 & 答案	2015
SimpleQuestions	86755	21687	Freebase & Reverb	答案	2015

Reference

- [Zelle, 1995] J. M. Zelle and R. J. Mooney, “Learning to parse database queries using inductive logic programming,” in Proceedings of the National Conference on Artificial Intelligence, 1996, pp. 1050–1055.
- [Wong, 2007] Y. W. Wong and R. J. Mooney, “Learning synchronous grammars for semantic parsing with lambda calculus,” in Proceedings of the 45th Annual Meeting-Association for computational Linguistics, 2007.
- [Lu, 2008] W. Lu, H. T. Ng, W. S. Lee, and L. S. Zettlemoyer, “A generative model for parsing natural language to meaning representations,” in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 783–792.
- [Zettlemoyer, 2005] L. S. Zettlemoyer and M. Collins, “Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars,” in Proceedings of the 21st Uncertainty in Artificial Intelligence, 2005, pp. 658–666.
- [Clarke, 2010] J. Clarke, D. Goldwasser, M.-W. Chang, and D. Roth, “Driving semantic parsing from the world’s response,” in Proceedings of the 14th Conference on Computational Natural Language Learning, 2010, pp. 18–27.
- [Liang, 2011] P. Liang, M. I. Jordan, and D. Klein, “Learning dependency-based compositional semantics,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 2011, pp. 590–599.
- [Cai, 2013] Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In ACL.
- [Berant, 2013] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In EMNLP.
- [Yao, 2014] Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In ACL.
- [Unger, 2012] Christina Unger, Lorenz B. Uhm, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. 2012. Template-based question answering over rdf data. In WWW.
- [Yahya, 2012] Mohamed Yahya, Klaus Berberich, Shady Elbas-suoni, Maya Ramanath, Volker Tresp, and Gerhard Weikum. 2012. Natural language questions for the web of data. In EMNLP.
- [He, 2014] S. He, K. Liu, Y. Zhang, L. Xu, and J. Zhao. 2014. Question answering over linked data using first-order logic. In EMNLP.
- [Artzi, 2011] Y. Artzi and L. Zettlemoyer. 2011. Bootstrapping semantic parsers from conversations. In Empirical Methods in Natural Language Processing (EMNLP), pages 421–432. Dan Goldwasser, Roi Reichart, James Clarke and Dan Roth. Confidence Driven Unsupervised Semantic Parsing. Proceedings of ACL (2011)
- [Krishnamurthy, 2012] Jayant Krishnamurthy and Tom M. Mitchell. Weakly Supervised Training of Semantic Parsers. In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2012
- [Reddy, 2014] S. Reddy, M. Lapata, and M. Steedman. 2014. Large-scale semantic parsing without question-answer pairs. Transactions of the Association for Computational Linguistics (TACL), 2(10):377–392.
- [Fader, 2013] A. Fader, L. S. Zettlemoyer, and O. Etzioni, “Paraphrase-driven learning for open question answering,” in Proceedings of the 51th Annual Meeting-Association for computational Linguistics, 2013, pp. 1608–1618.

Reference

- [Bordes, 2014a] Antoine Bordes, Jason Weston, and Nicolas Usunier, "Open Question Answering with Weakly Supervised Embedding Models", In Proceedings of ECML-PKDD, 2014.
- [Bordes, 2014b] Antoine Bordes, Sumit Chopra, and Jason Weston, "Question Answering with Subgraph Embedding", In Proceedings of EMNLP 2014.
- [Yang, 2014] Min-Chul Yang, Nan Duan, Ming Zhou, and Hae-Chang Rim, "Joint Relational Embeddings for Knowledge-Based Question Answering", In Proceedings of EMNLP 2014.
- [Dong, 2015] Li Dong, Furu Wei, Ming Zhou, and Ke Xu, "Question Answering over Freebase with Multi-Column Convolutional Neural Network", In Proceedings of ACL 2015.
- [Bordes, 2015] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston, "Large-scale Simple Question Answering with Memory Network", In Proceedings of ICIR 2015.
- [Yih, 2015] Wen-tau Yih, Ming Wei Chang, Xiaodong He, and Jianfeng Gao, "Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base", In Proceedings of ACL 2015. (Outstanding Paper)
- [Zeng, 2014] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou and Jun Zhao, "Relation Classification via Convolutional Deep Neural Network", in Proceedings of COLING 2014. (Best Paper)
- [Zeng, 2015] Daojian Zeng, Kang Liu, Yubo Chen and Jun Zhao, "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks", in Proceedings of EMNLP 2015.
- [Xu, 2015] Xu Yan, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin, "Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths", In Proceedings of EMNLP 2015.
- [Berant, 2014] Jonathan Berant, and Percy Liang, "Semantic Parsing via Paraphrasing", In Proceeding of ACL 2014. (Best long paper honorable mention)
- [Zhang, 2016] Yuanzhe Zhang, Shizhu He, Kang Liu and Jun Zhao. A Joint Model for Question Answering over Multiple Knowledge Bases, To appear in Proceedings of AAAI 2016, Phoenix, USA, Phoenix.

谢谢! Q&A!