基于知识图谱的问答系统关键技术

肖仰华（Yanghua Xiao）

复旦大学知识工场实验室(Kw.fudan.edu.cn)

上海数眼科技发展有限公司(shuyantech.com)

Outline

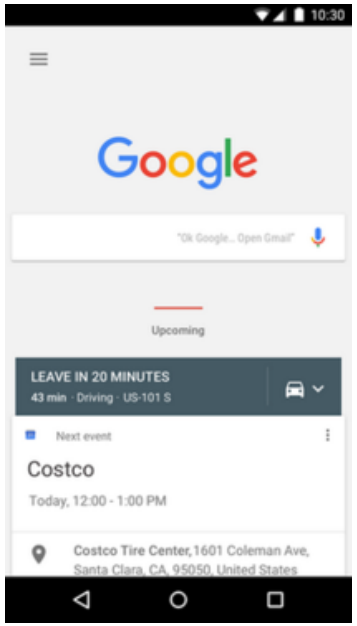
- KBQA background
- 不倒翁问答系统
- Template based KBQA
- Conclusion

Backgrounds

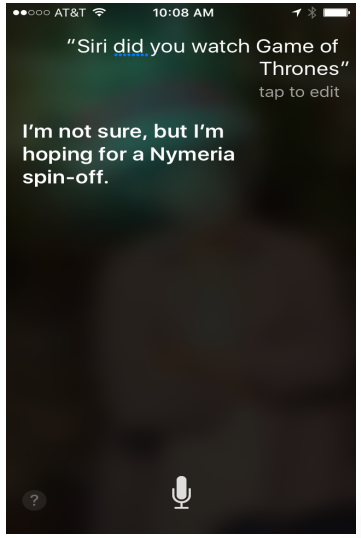
- Question Answering (QA) systems answer natural language questions.

IBM Watson

Google Now

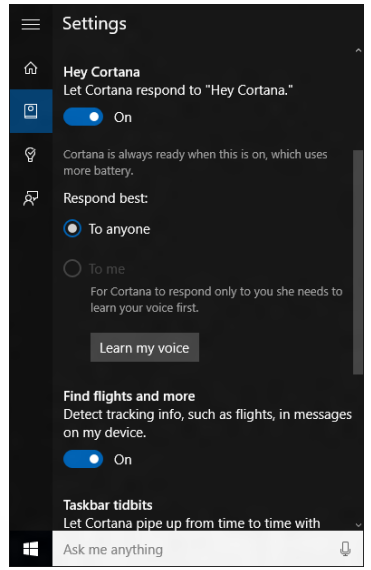


Apple Siri



Amazon Alexa

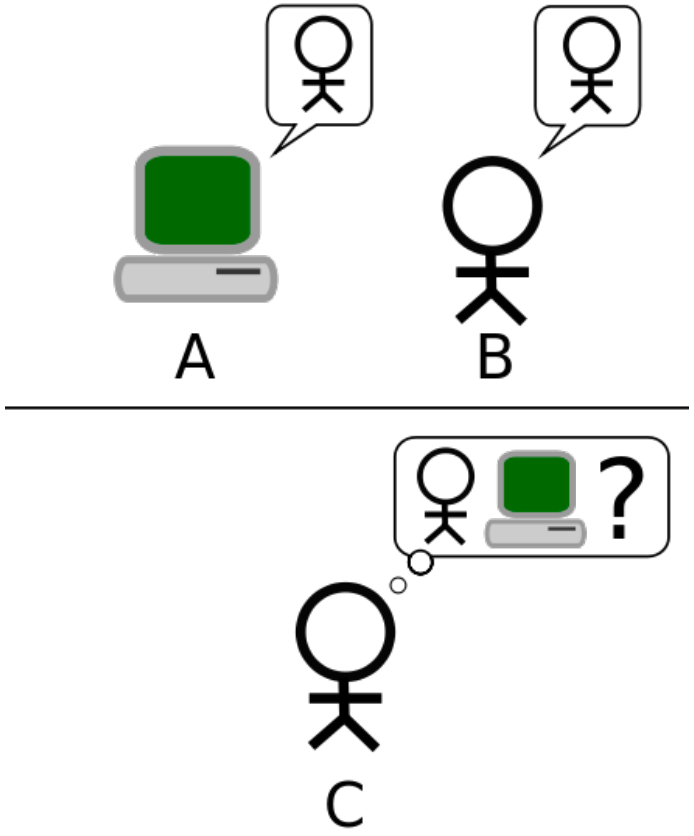
Microsof Cortana



Why QA

- QA application:
  - One of the most natural human-computer interaction
  - Key components of Chatbot, which attracts wide research interests from industries

- QA for AI:
  - One of most important tasks to evaluate the machine intelligence: Turing test
  - Important testbed of many AI techniques, such as machine learning, natural language processing, machine cognition



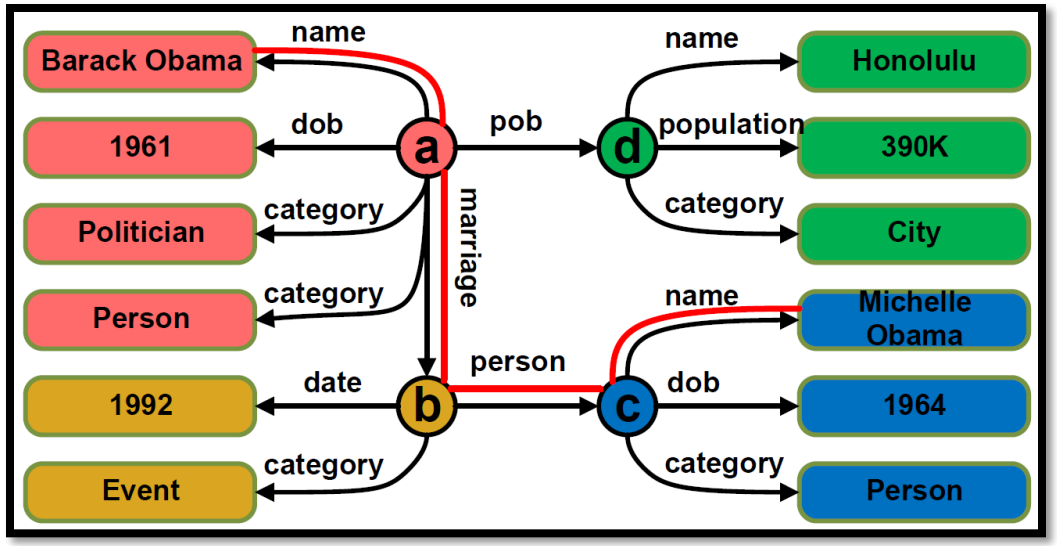
Turing test

Why KBQA?

More and More Knowledge bases are created

- Google Knowledge graph, Yago, WordNet, FreeBase, Probase, NELL, CYC, DBPedia
- Large scale and high quality

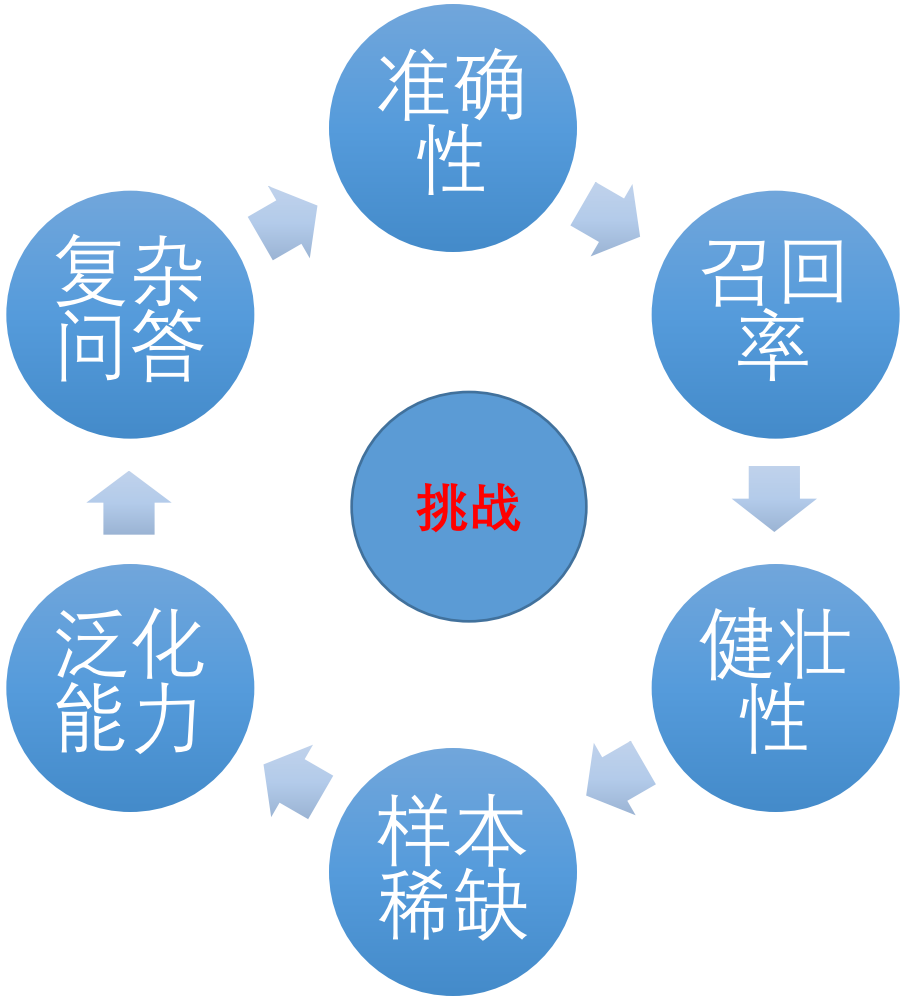
The boost of knowledge bases



A piece of knowledge base, which consist of triples such as (d, population, 390k)

实用化知识问答-机遇与挑战

- 机遇：从人工智障到人工智能
  - 大量的问答模型研究
  - 大量的知识库与语料
- 挑战
  - 如何构建一个实用化问答系统



为什么现在人工智能助理都像人工智障？

2017十大“人工智障”事件 科技巨头无一幸免

从人工智能变成“人工智障”，聊天机器人殆在哪？

Outline

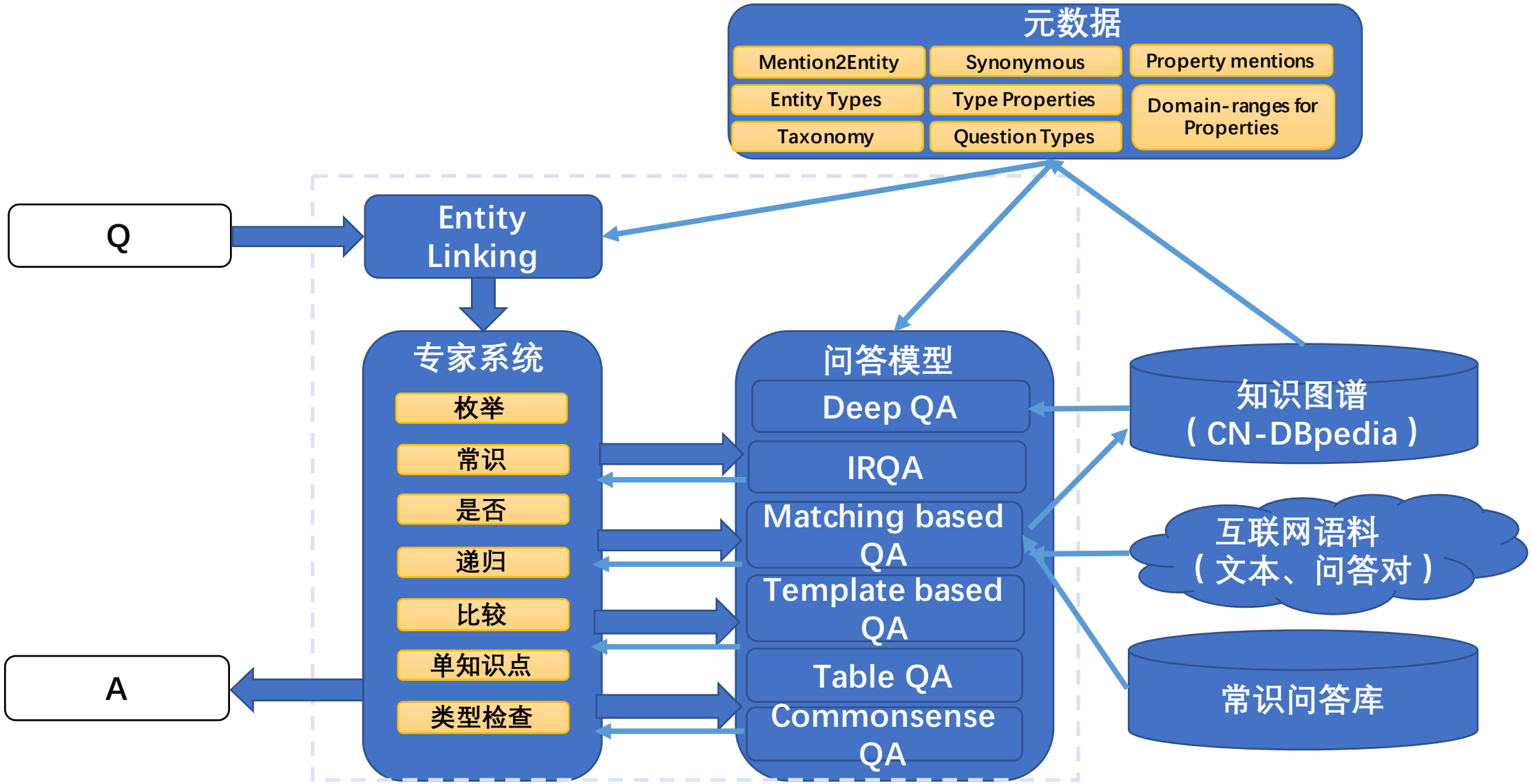
- KBQA background
- 不倒翁问答系统
- Template based KBQA
- Conclusion

总览

- “不倒翁“知识问答DEMO
  - http://218.193.131.250:20013/
  - http://shuyantech.com/qa

- 知识问答
  - 知识图谱、深度学习、规则系统

“不倒翁”知识问答系统架构



模块列表

- Entity Linking
  - 实体识别、消歧、链接到知识库
- 专家系统
  - 基于复杂规则解决递归、比较等复杂问题
  - 决定子模块的选择性调用
  - 复杂问题存在规律性且缺少训练数据，规则系统是较优选择
- 元数据
  - 提供问答系统所需要的语言知识、词汇知识、本体等元数据

模块列表

- Matching based QA
  - 基于深度学习对问题库进行模糊匹配
  - 解决常识类问答
    - e.g. *为什么天空是蓝色的*
- Deep QA
  - 基于深度学习将问句匹配到知识库中的某条具体知识
  - 解决单条事实类问答
    - e.g. *七里香是谁唱的*
- IRQA
  - 到互联网上搜索，并使用深度机器阅读理解模型获取答案
  - 获取到的答案可以用于补全知识库
  - 解决知识库缺失问题
    - e.g. *鲁迅的身高是多少*
- Template based QA
  - 基于taxonomy、问答语料对训练生成概念语义模板
  - 利用概念语义模板对问题进行理解，完成回答
  - 适用于对于新实体的问答
    - Eg. *华为P20多大尺寸？*
- Table/list QA
  - 基于table、list回答枚举类问题
    - Eg，*复旦有哪些院士？*
- Commonsense QA
  - 回答常识问答类问题
    - Eg，天空为什么是蓝色的？

Mention2Entity库

- 形式： 字符串（ mention ）→ 实体（ entity ）或 实体列表 （ entity list ）
- 示例
  - “复旦”→“复旦大学”
  - “周董”→“周杰伦”
  - “雨神” → [“萧敬腾（ 华语流行男歌手 ）”, “雨神（ 中国神仙 ）”]
- 作用：提高QA系统识别实体的能力

等价属性库

- 形式：属性 → 其他等价属性列表
- 示例
  - “妻子”→ [“夫人”，“老婆”，“伉俪”，“爱人”，“配偶”]
  - “出生日期”→ [“生日”，“出生年月”，“出生时间”，“诞辰”]
  - “主演”→ [“主要演员”，“影片主演”，“领衔主演”]

- 提高QA系统识别属性的能力

搜索结果

回答：

朱丽倩

搜索结果

回答：

1979年01月18日

搜索结果

回答：

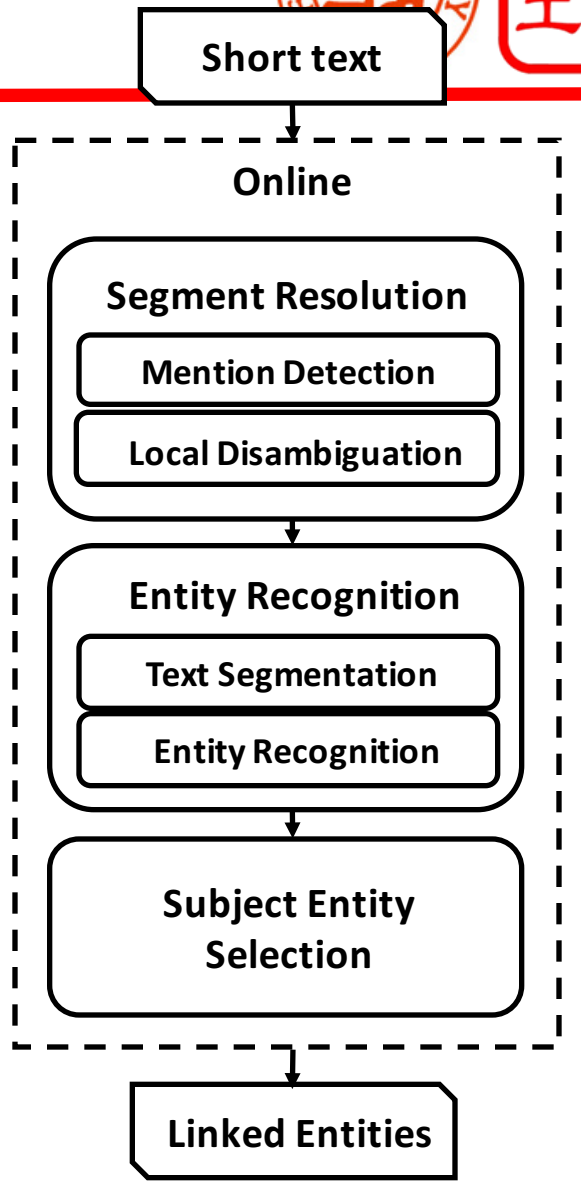
谭卓, 章宇, 徐峥

属性元数据库

- 形式：属性 → domain → range
- 示例
  - “妻子”→ 人物 → 人物
  - “出生日期”→ 人物 → 日期
  - “主演”→ 电影 → 演员
- 提高QA系统的准确率

针对QA的实体链接

- 段解析
  - 检测出所有的mention
  - 在有限的上下文进行局部链接
    - 利用实体概念作为细粒度的主题
    - 用词向量相似度来解决稀疏文本的问题
- 实体识别
  - 利用局部链接的分数进行语义文本划分
  - 根据段的语法语义信息进行实体识别
- 主体实体选择
  - 选择作为询问主体的实体
    - 实体关系 & 实体流行度



Running example

- Input: text sequence: 刘若英语怎么样
- Segment Resolution
  - Detect all possible mention: “刘若”、“刘若英”、“英语”
  - Candidate entity generation: 刘若1_{KB}、刘若2_{KB} 、刘若英$_{KB}$、英语$_{KB}$…
  - Local disambiguation: φ(刘若,刘若1_{KB})=0.4、φ(刘若,刘若2_{KB})=0.3…
- Entity Recognition
  - Text segmentation: 刘若|英语|怎么样
  - Entity recognition: 刘若、英语
- Global linking: rel(刘若,刘若1_{KB})=0.01、rel(刘若,刘若2_{KB})=0.2
- Output: mention and entity {(刘若, 刘若2_{KB}), (英语, 英语$_{KB}$)}

$$\varphi(m,e) = \varepsilon \cdot sim_c(m,e) + (1-\varepsilon) \cdot sim_t(m,e)$$

Topic coherence: $sim_c(m,e)$

Textual similarity: $sim_t(m,e)$

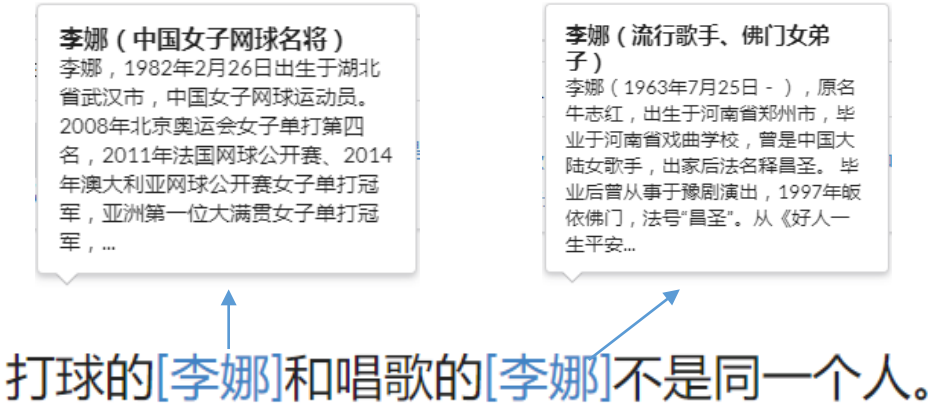
$$W = argmax\, O(W) = \log\big(P(W)\big) = \sum_{i=1}^{l} P(w_i)$$

Calculate from KB

结果

- Entity linking for QA 结果：
  - 准确率：90%，召回率: 94.3% +
  - F1: 92.0% +

Entity Linking已经广泛应用在新闻、出版等领域



李娜（中国女子网球名将）：人物、体育人物、运动员、名将

李娜（流行歌手、佛门女弟子）：人物、演员、歌手、弟子

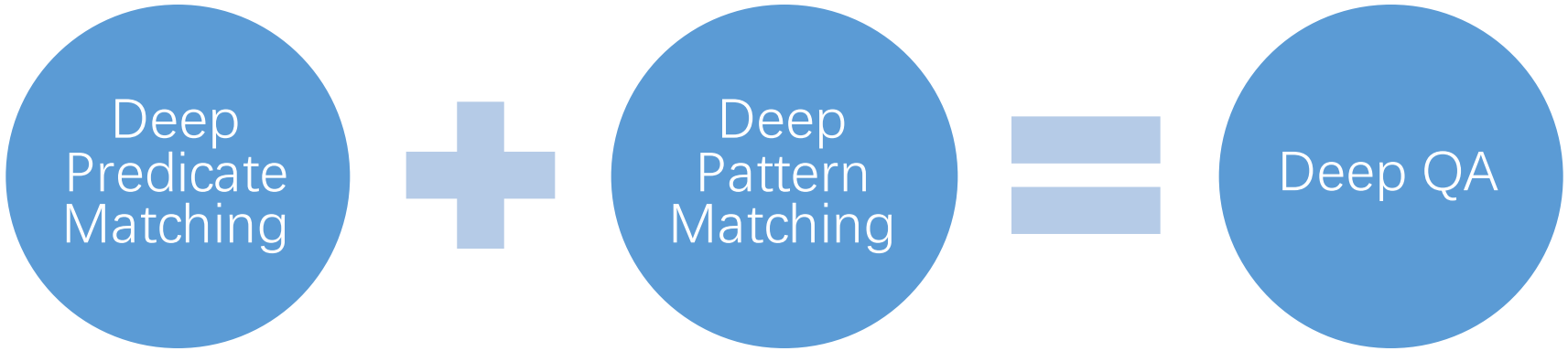
Table 4: Results of entity recognition task

Methods	Prec.	Reca.	F1
Stanford NER [28]	58.7	39.5	47.2
Baidu Entity Annotation [2]	65.5	78.2	71.3
Our Method	**91.0**	**89.4**	**90.2**

中文通用Entity Linking的准确率

Deep QA

- 目标：
  - 单知识点、且三元组知识库包含答案的问题



解决训练数据充足的问法和属性之间的匹配问题，具有强大的错误容忍能力

Input:
请问哪个知道姚明的女的儿是啊啊打发
Output:
姚沁蕾

0.6161 姚明（中国篮球协会主席、中职联公司董事长） 女儿 姚沁蕾

Few-Shot Learning：解决训练数据稀少的问法和属性之间的匹配问题，能处理垂直领域属性和特殊问法，无需训练即可添加规则

Input:
阿里巴巴的波士是谁啊
Output:
马云

1.2954 阿里巴巴集团 创始人 马云 Reason: ~的波士是谁->创始人

专家系统

- 复杂问题存在规律性且缺少训练数据，规则系统是较优选择
- 目标：
  - 基于复杂规则决定子模块的选择和组合调用策略
  - 解决递归、比较等复杂问题

- 规则配置：


```
If Q.contain('有哪些') and
   Q.entity.exists() and
   Q.entity.has_web_tables()
Then
   web_tables =  Q.entity.web_tables()
   answer = GetMaxSimilarity(Q, web_tables)
   return answer
```

测试

- NLPCC-2016 KBQA，Testset，1000 QA-pairs
  - 91.6% Accuracy
- SougouQA Factoid Reading Comprehension Task，验证集
  - 73.3% Exact Match
- IRQA Hand-made Testset
  - Exact Match 83.3%
  - Acceptable 90.0%

2：可解释但不完全正确		
1：正确		
0：错误		
标注	答案	问题
1	忽必烈	哪个皇帝首先定都北京
1	忽必烈	哪个皇帝第一个定都北京
0	玄烨	哪位皇帝第一个定都北京
1	元世祖忽必烈	最早定都北京的皇帝是谁
1	春宵苦短日高起	从此君王不早朝前一句是什么
1	王冶坪	江泽民的老婆是谁
1	王冶坪	江泽民的妻子是哪位
1	朱安	鲁迅的妻子是谁
1	朱安	鲁迅的老婆是谁
1	朱安	鲁迅的夫人是谁
1	贺知章	哪位唐代诗人寿命最长
1	陆游	哪位宋代诗人寿命最长
2	魏小鹏	复旦大学的党委书记是谁
1	严峰	复旦大学图书馆的党委书记是谁
1	冈妈	高达铁血的剧本是谁写的
0	244.99万平方米	复旦大学张江校区的占地面积
1	斯塔夫里阿诺斯	全球通史的作者是谁
0	2012年	兰巴拉尔哪年战死的
1	1949年	中华人民共和国那年成立的
1	2.26米	姚明的身高有多高
2	邓小平	哪位领导人寿命最长
1	德国	2014年世界杯冠军是哪个队
1	阿根廷	2014年世界杯亚军是哪个队
1	小雨转雷阵雨	今天上海天气
1	复旦大学	梁家卿的学校是
1	博学而笃志	复旦大学的校训是什么
1	1米9	特朗普的身高是多少
1	湖南韶山	毛泽东的籍贯在哪里
1	毛泽东	杨开慧是谁的妻子
1	毛泽东	江青是谁的老婆

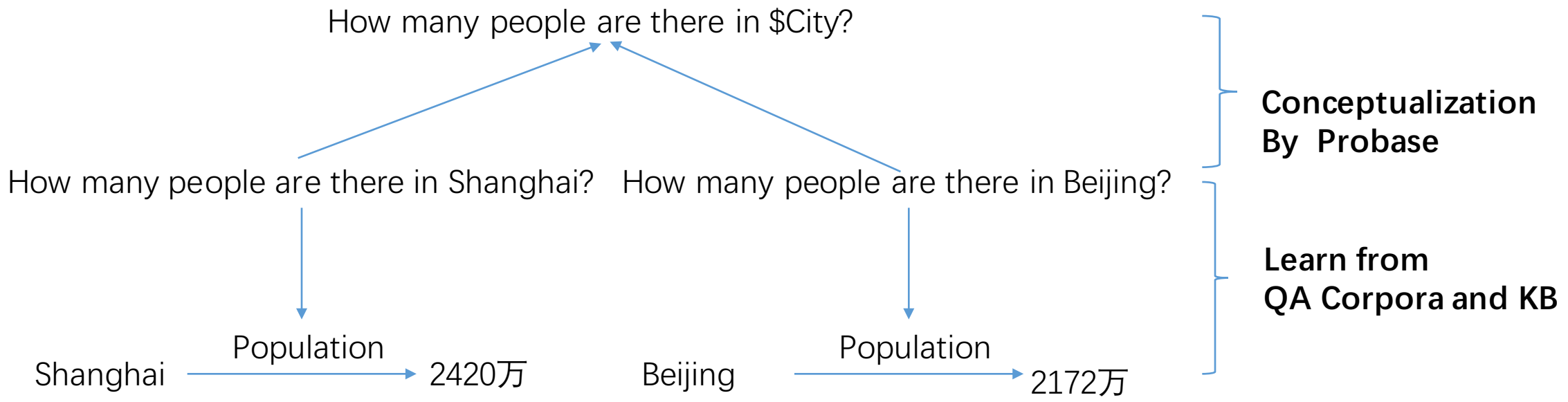
**由于搜索引擎返回结果可能改变，IRQA的返回结果也可能会有不同*

Outline

- KBQA background
- 不倒翁问答系统
- Template based KBQA
- Conclusion

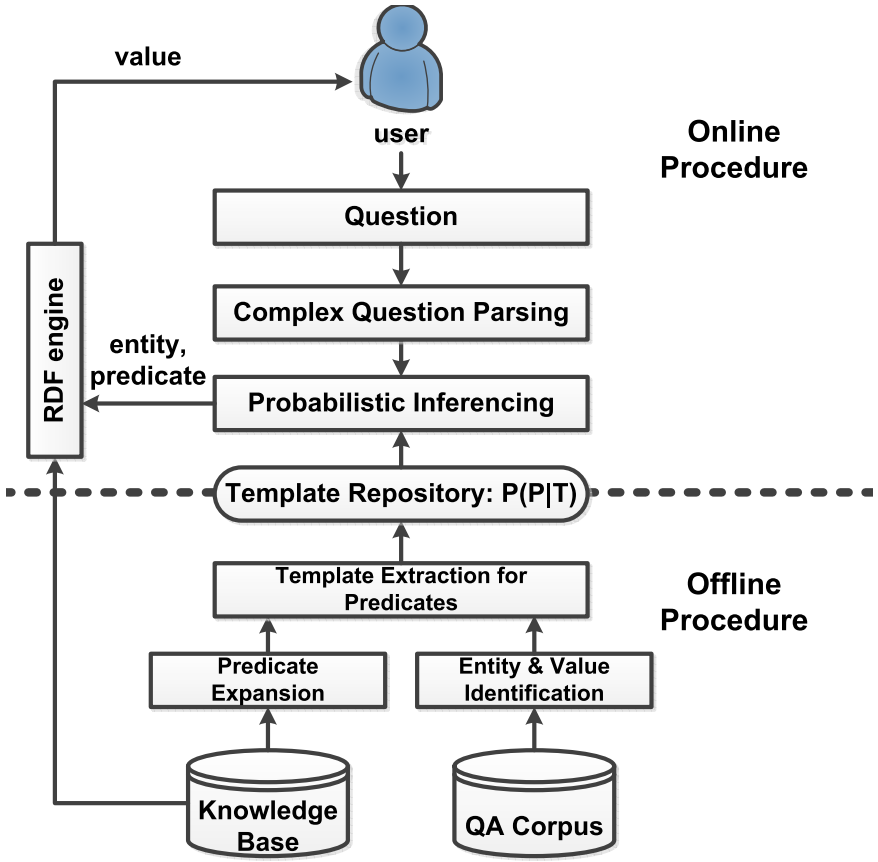
Our approach

- Representation: concept based templates.
  - Questions are asking about entities. The semantic of the question is reflected by its corresponding concept.
  - Advantage: Interpretable, user-controllable
- Learn templates from QA corpus, instead of manfully construction.
  - 27 million templates, 2782 intents
  - Understand diverse questions



System Architecture

- Offline procedure
  - Learn the mapping from templates to predicates: P (p|t),
  - Input: qa corpora, large scale taxonomy, KB
  - Output: P(P|T)
- Online procedure
  - Parsing, predicate inference and answer retrieval
  - Input: binary factoid questions (BFQs)
  - Output: answers in KG



Problem Model

- Given a question q, our goal is to find an answer v with maximal probability (v is a simple value)

$$\arg\max_{v} P(V = v|Q = q) \longrightarrow \arg\max_{v} \sum_{e,t,p} P(v|q, e, t, p)$$

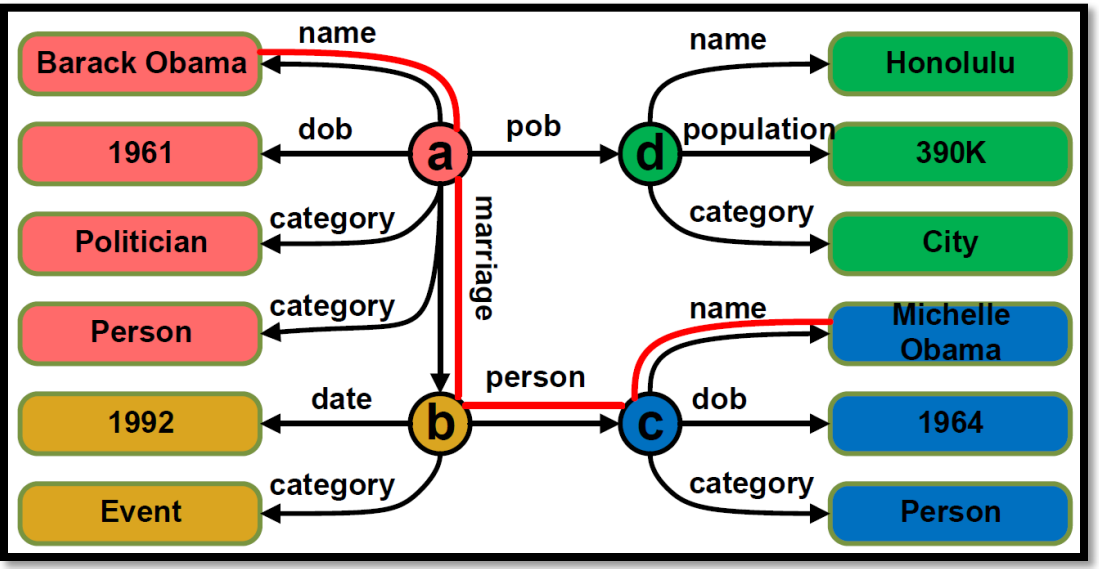
e: entity; t: template; p: predicate

- Basic idea：We proposed a generative model to explain how a value is found for a given question,

- Rationality of probabilistic inference
  - *uncertainty* (e.g. some questions' intents are vague)
  - *Incompleteness* (e.g. the knowledge base is almost always incomplete),
  - *noisy* (e.g. answers in the QA corpus could be wrong)

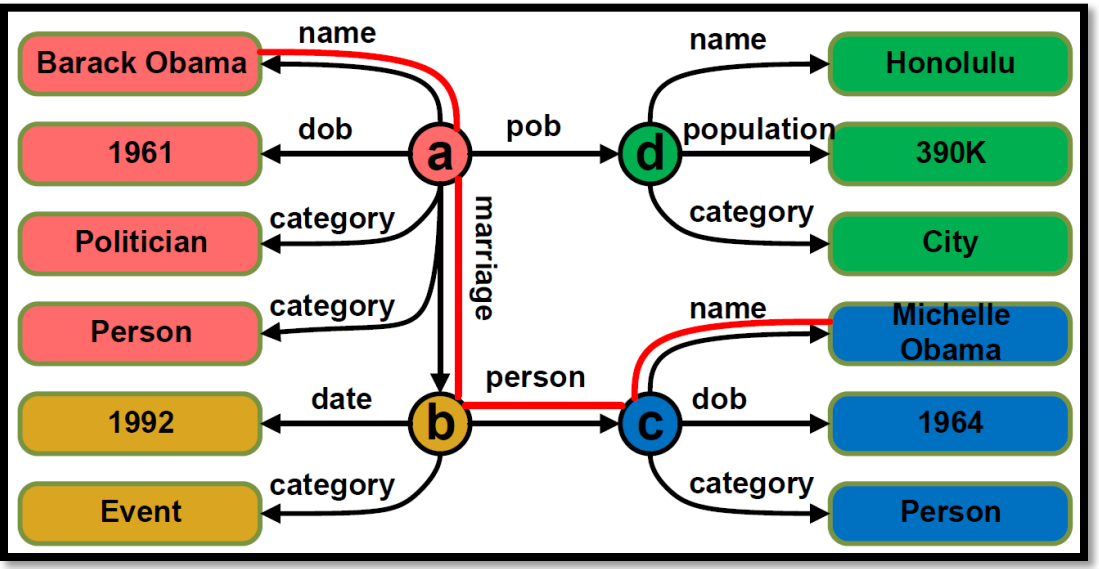
question2answer: a generative process

- A qa pair
  - Q: How many people live in Honolulu?
  - A: It's 390K.



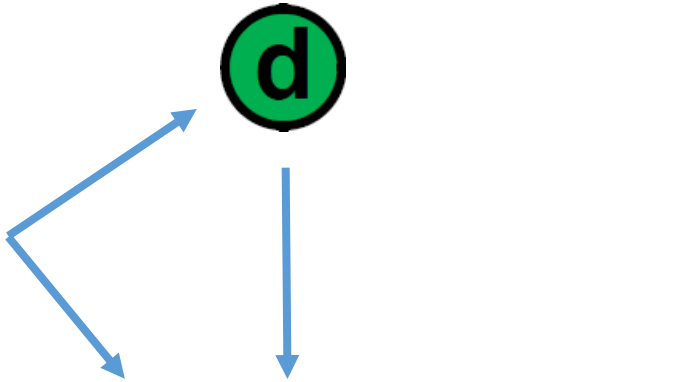
question2answer: entity linking

How many people live in Honolulu?

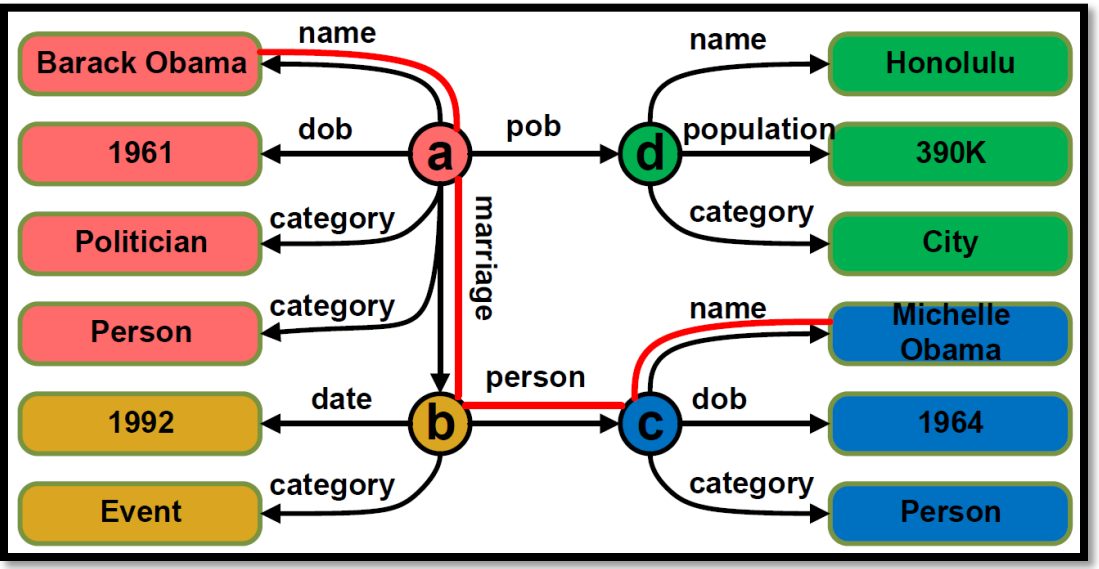


question2answer: conceptualization

How many people live in Honolulu?

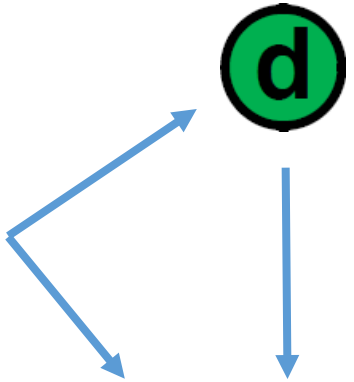


How many people live in $city?

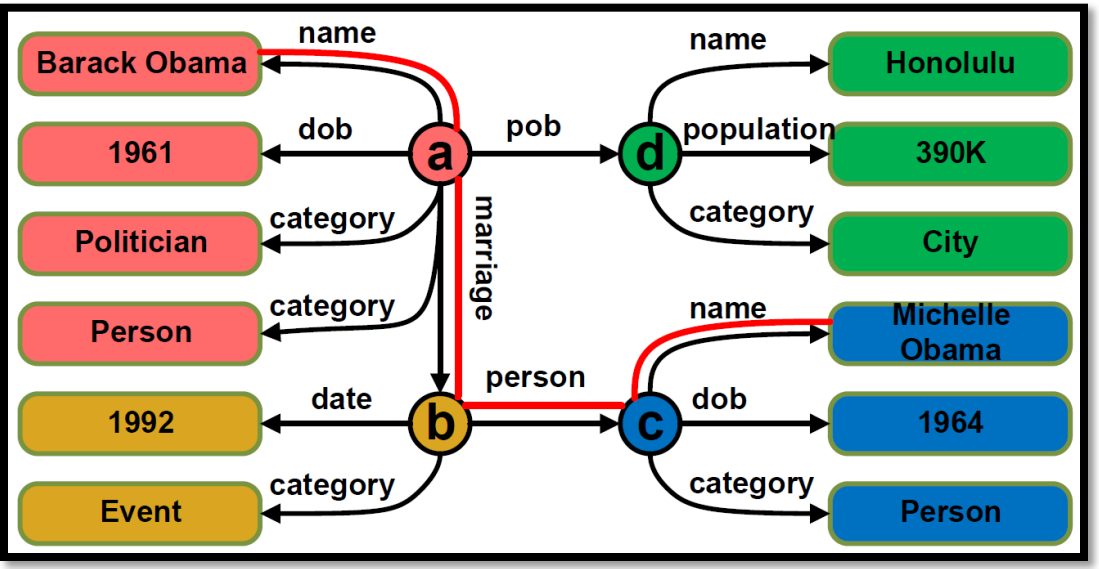


question2answer: predicate inference

How many people live in Honolulu?

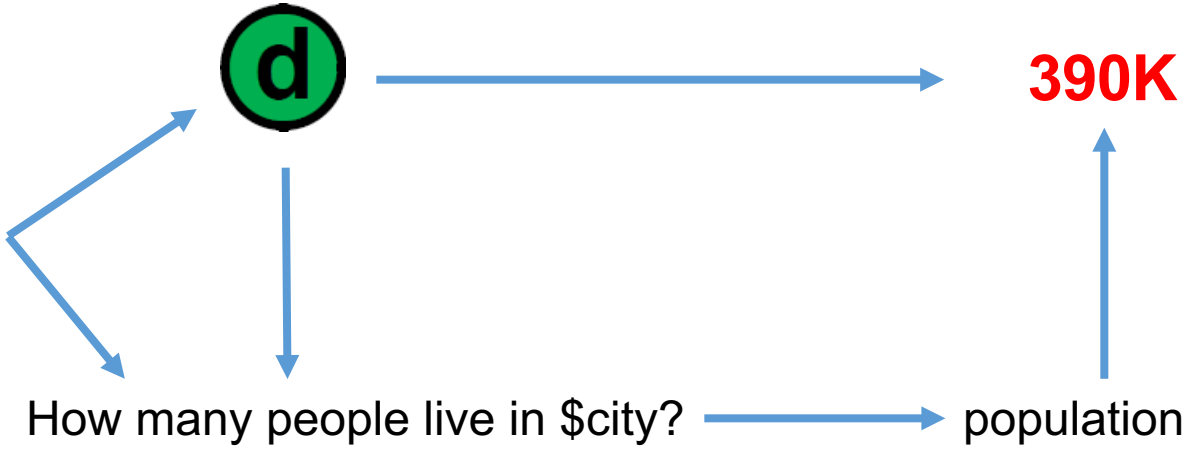


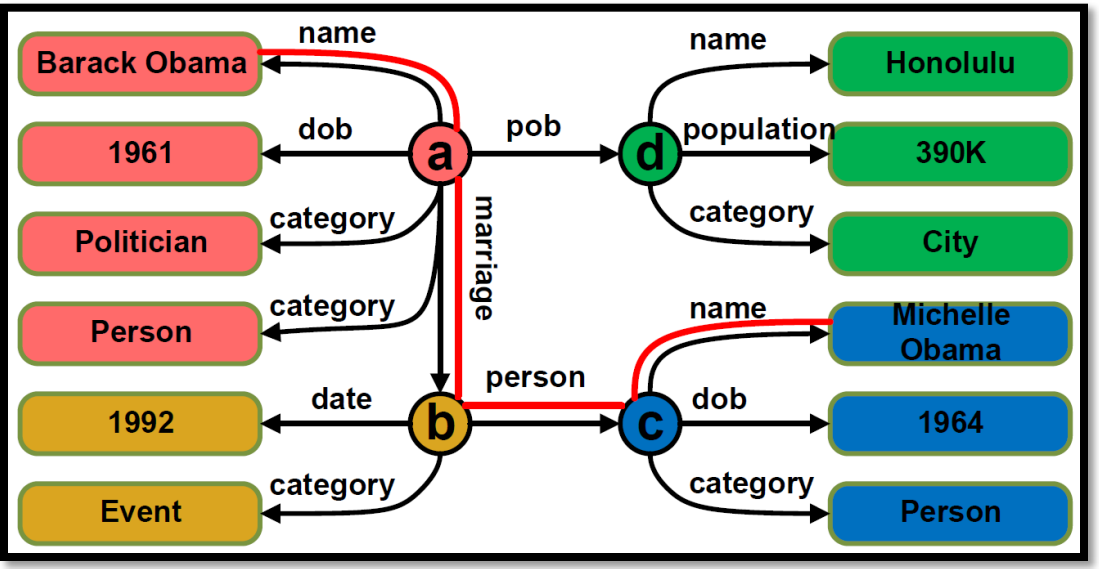
How many people live in $city? → population



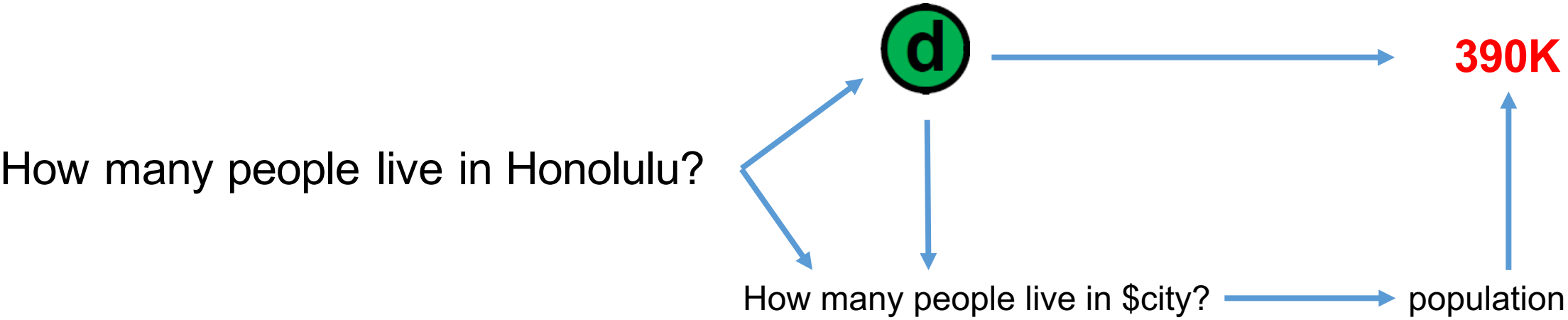
question2answer: value lookup

How many people live in Honolulu?



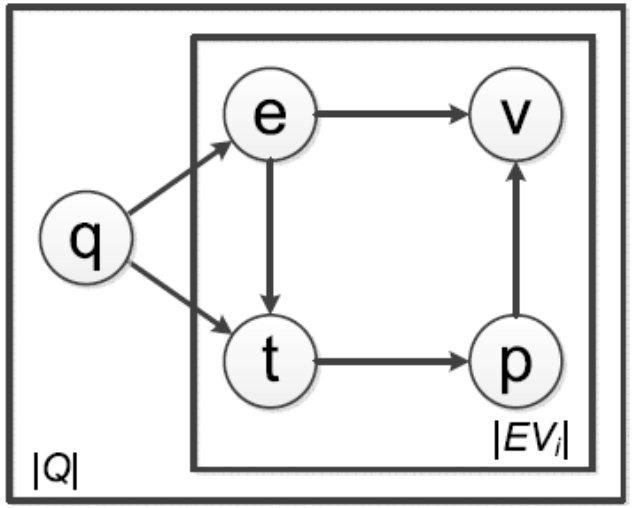


Probabilistic graph model



$$P(q, e, t, p, v) = P(q)P(e|q)P(t|e, q)P(p|t)p(v|e, p)$$

$$\arg\max_{v} \sum_{e,t,p} P(v|q, e, t, p)$$



Probability Computation

- Source
  - QA corpora (42M Yahoo! Answers)
  - Knowledge base such as Freebase
  - Probase(a large scale taxonomy)

- Directly estimated from data
  - Entity distribution P (e|q)
  - Template distribution P(t|q,e)
  - Value (answer) distribution P(v|e,p)

Question	Answer
When was Barack Obama born?	The politician was born in 1961.
When was Barack Obama born?	He was born in 1961.
How many people are there in Honolulu?	It's 390K.

Yahoo! Answers QA pairs

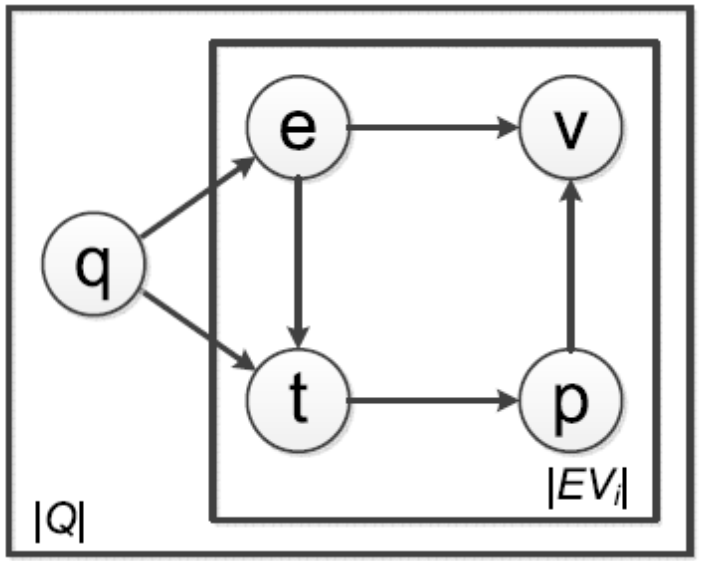
P(P|T) estimation

- We treat P(P|T) as parameters, and learn the parameter using maximum likelihood estimator, maximizing the **likelihood** of observing QA corpora
- An EM algorithm is used for parameter estimation

$$\hat{\theta} = \arg\max L(\theta)$$

$$L(\theta) = \sum_{i=1}^{m} \log P(x_i) = \sum_{i=1}^{m} \log P(q_i, e_i, v_i)$$

$$= \sum_{i=1}^{m} \log[\sum_{p \in P, t \in T} P(q_i) P(e_i|q_i) P(t|e_i, q_i) \theta_{pt} P(v_i|e_i, p)]$$



Answering complex questions

- When was Barack Obama's wife born?
  - (Who is) Barack Obama's wife?
  - When was Michelle Obama born?
- How to decompose the question into a series of binary questions?

$$\arg\max_{\mathcal{A}\in\mathbb{A}(q)} P(\mathcal{A})$$

- A binary question sequence is meaningful, only if each of the binary question is meaningful.

$$P(\mathcal{A}) = \prod_{\check{q}\in\mathcal{A}} P(\check{q})$$

- A dynamic programming (DP) algorithm is employed to find the optimal decomposition.

Experiments

	KBQA	Bootstrapping
Corpus	41M QA pairs	256M sentences
Templates	27,126,355	471,920
Predicates	2782	283
Templates per predicate	9751	4639

KBQA finds significantly more templates and predicates than its competitors despite that the corpus size of bootstrapping is larger.

$$marriage \rightarrow person \rightarrow name$$

Who is $person marry to?
Who is $person's husband?
What is $person's wife's name?
Who is the husband of $person?
Who is marry to $person?

Concept based templates are meaningful

Experiments

	#pro	#ri	#par	R	R*	P	P*
Xser	42	26	7	0.52	0.66	0.62	0.79
APEQ	26	8	5	0.16	0.26	0.31	0.50
QAnswer	37	9	4	0.18	0.26	0.24	0.35
SemGraphQA	31	7	3	0.14	0.20	0.23	0.32
YodaQA	33	8	2	0.16	0.20	0.24	0.30
				R R_{BFQ}	R^* R^*_{BFQ}		
KBQA+KBA	7	5	1	0.10 0.42	0.12 0.50	0.71	0.86
KBQA+Freebase	6	5	1	0.10 0.42	0.12 0.50	0.83	1.00
KBQA+DBpedia	8	8	0	0.16 0.67	0.16 0.67	**1.00**	**1.00**

Results over QALD-5. The results verify the effectiveness of KBQA over BFQs.

Experiments

Hybrid systems

- First KBQA
- If KBQA gives no reply, then baseline systems.

System	R	R*	P	P*
SWIP	0.15	0.17	0.71	0.81
KBQA+SWIP	0.33(+0.18)	0.35(+0.18)	0.87(+0.16)	0.92(+0.11)
CASIA	0.29	0.37	0.56	0.71
KBQA+CASIA	0.38(+0.09)	0.44(+0.07)	0.66(+0.10)	0.76(+0.05)
RTV	0.3	0.34	0.34	0.62
KBQA+RTV	0.39(+0.09)	0.42(+0.08)	0.66(+0.32)	0.71(+0.09)
gAnswer	0.32	0.43	0.42	0.57
KBQA+gAnswer	0.39(+0.07)	-	-	-
Intui2	0.28	0.32	0.28	0.32
KBQA+Intui2	0.39(+0.11)	0.41(+0.09)	0.39(+0.11)	0.41(+0.09)
Scalewelis	0.32	0.33	0.46	0.47
KBQA+Scalewelis	0.44(+0.12)	0.45(+0.12)	0.60(+0.14)	0.62(+0.15)

Results of hybrid systems on QALD-3 over DBpedia. The results verify the effectiveness of KBQA for a dataset that the BFQ is not a majority.

Outline

- KBQA background
- 不倒翁问答系统
- Template based KBQA
- Conclusion

Conclusion

- 样本增强可以显著提升问答模型的健壮性
- 基于阅读理解的IRQA可以显著提升召回率
- Pattern+DL可以显著提升问答模型的泛化能力
- 规则系统可以提升问答系统的复杂问题的回答能力
- 支撑性数据是确保问答系准确性的前提

- 如何实现一个实用化的知识问答系统？
  - 实用化问答系统的关键技术研究十分缺乏
- 如何有效评测一个知识问答系统？
  - 有效的知识问答评测数据集、评测指标的研究仍然十分缺乏

Reference

- ***Wanyun Cui**, et al., KBQA: Learning Question Answering over QA Corpora and Knowledge Bases, **(VLDB 2017)***
- ***Wanyun Cui**, et al., KBQA: An Online Template Based Question Answering System over Freebase, **(IJCAI 2016),** demo*
- Lihan Chen, et al, Entity Linking for Short Text, Technique report of KW.
- Jiaqing Chen, et al, Incorporating complicated rules in deep generative network. Technique report of KW

Thank you!

DEMO

http://218.193.131.250:20013/

http://shuyantech.com/qa