

《知识图谱: 概念与技术》

第 13 讲

基于知识图谱的问答技术

崔万云

上海财经大学

cui.wanyun@sufe.edu.cn

Wanyun Cui

- Where do I come from?
 - 2017-present, AP, SUFE
 - 2013-2017 PhD, Fudan University
 - Advised by Wei Wang and Yanghua Xiao
- What do I work on?
 - Question answering
 - 2012.1 – 2012.11 Microsoft Research Asia
 - Advised by Haixun Wang, IEEE fellow
 - 2014.7 – 2014.11 Baidu DeepQA project (小度机器人)
 - Knowledge graph

本章大纲

- 知识问答概述
- 深度学习方法
- 文本混合问答
- 如何构建更鲁棒的问答系统
- 总结

知识问答概述

背景

- 问答系统主要功能是回答人提出的**自然语言问题**



IBM Watson

2011年，Watson战胜了其他人类竞争者，并获得答题比赛Jeopardy! 的一百万美金奖金。

背景

- 问答系统主要功能是回答人提出的**自然语言问题**

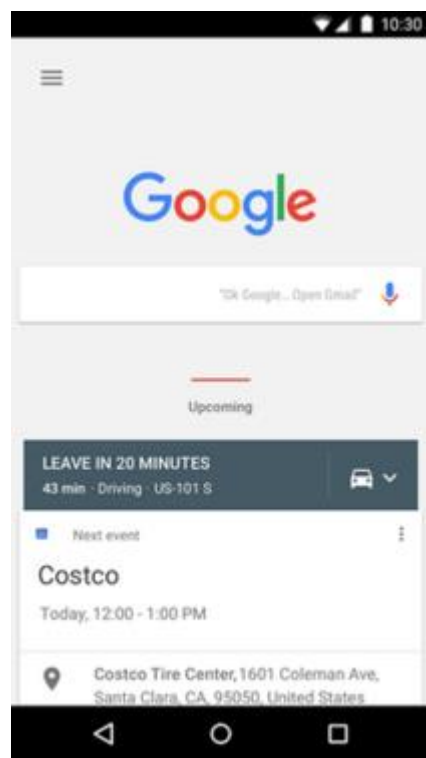


Apple Siri

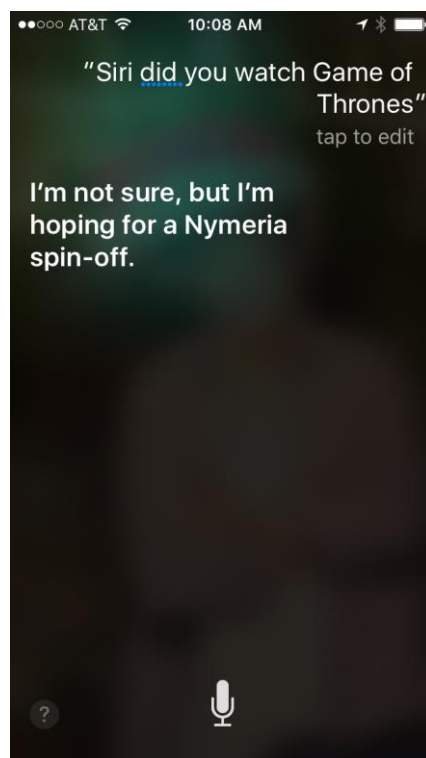
Siri是WWDC 2016展上的明星。

背景：相关产品

Google
Google Now



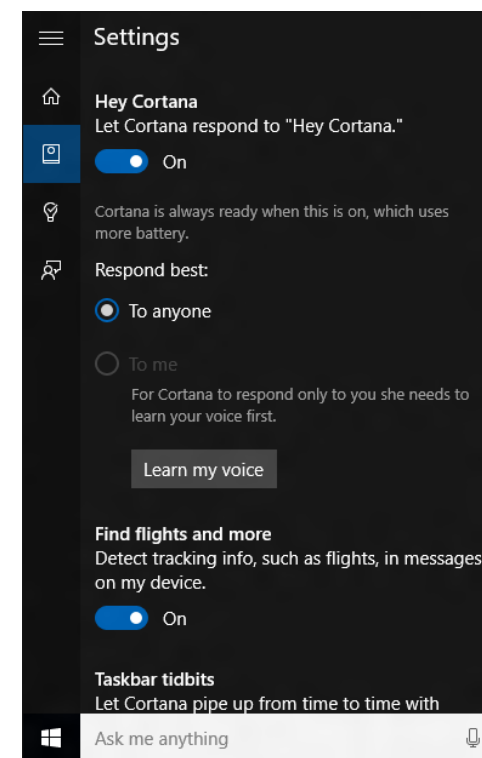
Apple
Siri



Amazon
Alexa

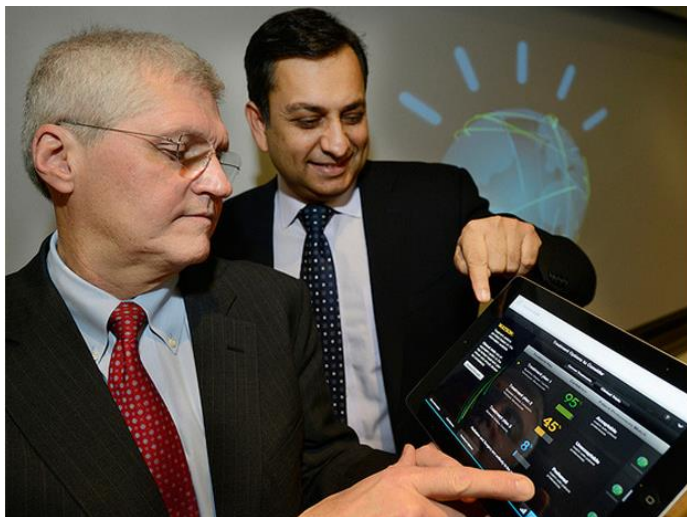


Microsoft
Cortana



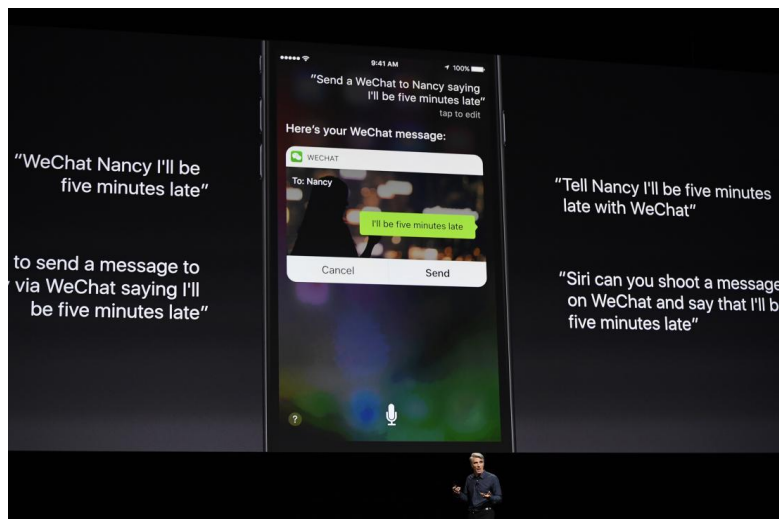
应用场景

健康咨询



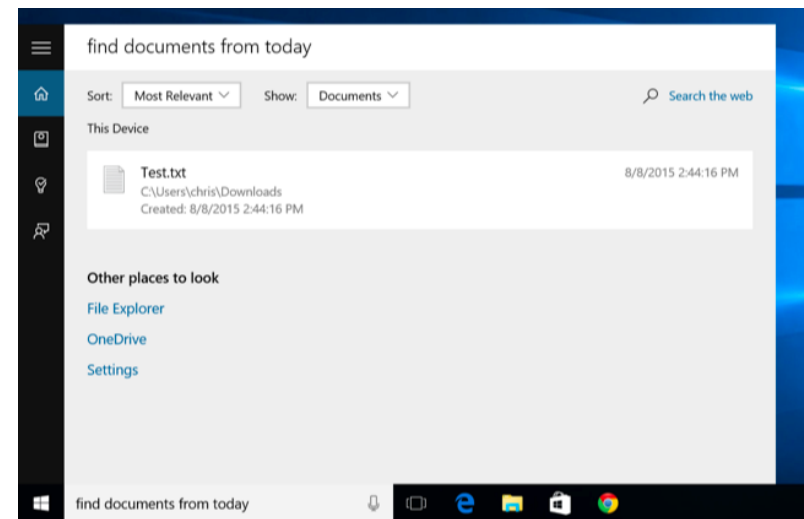
Bumrungrad Hospital and Watson Improve Cancer Care.

数字助理



Craig Federighi speaks during the Apple WWDC 2016.

自然语言搜索



Cortana supports natural language search for files on your computer.

问答系统的研究意义

- 问答系统带来了
 - 为不同语义理解模型的整合提供了**应用出口**
 - 为不同模型的关联分析、数据共享、参数共享等提出了**实际需求**
 - 为多个自然语言语义理解技术模型的整体突破带来了**技术愿景**
- 问答系统应用
 - 降低了人机交互的门槛
 - 提供了访问海量知识的新渠道

QA的知识源

网页

问答社区

搜索引擎

百科

知识图谱

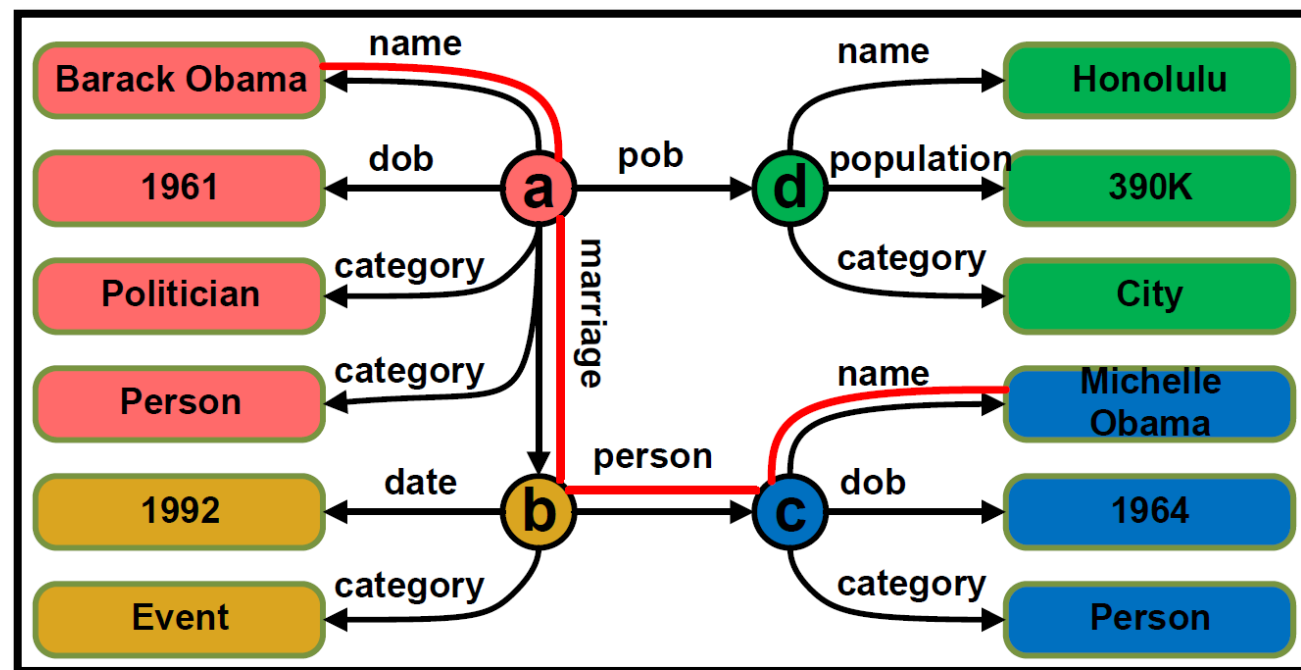


纯文本

结构化数据
(RDF)

一个简单知识图谱

- 结构化、关联化数据
- 每条边表示一条知识
 - (d, population, 390k)



为什么KBQA?

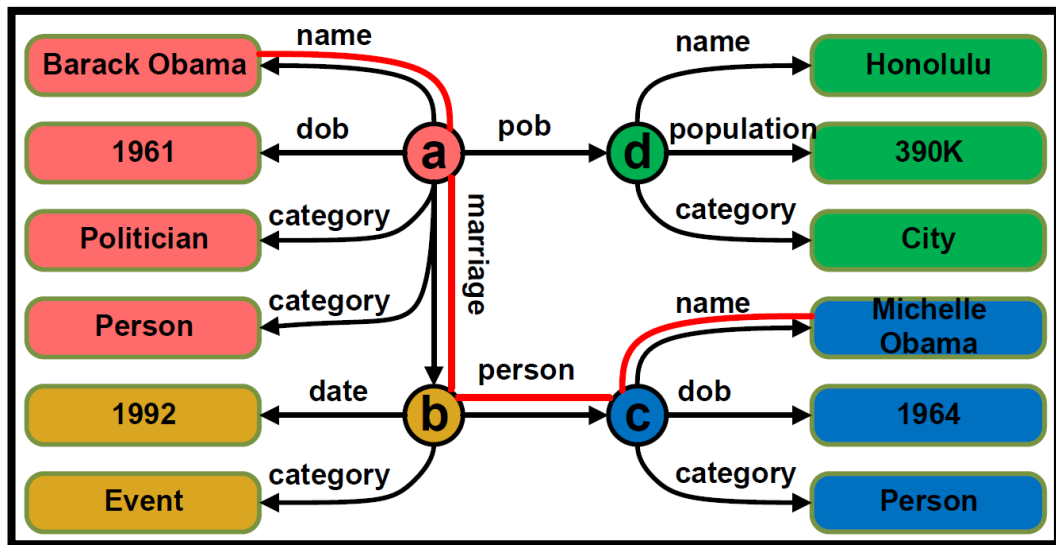
- 关联性数据 – 更丰富知识表示
 - 纯文本: 文本句子的内部理解
 - 知识图谱: 关联性数据, 提供文本理解的语义背景
- 数据质量 – 更高的知识准确率
 - 纯文本: 文本错误或者不同文本的知识矛盾
 - 知识图谱: 人工标注或解析自网页表格的高质量数据
- 结构化数据 – 查询效率
 - 纯文本: 倒排表
 - 知识图谱: 存储于数据库, 使用索引加速查询

这些优势促使我们使用KBQA!

工作方式

- 将自然语言问题转化为知识图谱上的结构化查询
- 核心：属性理解

How many people live in Honolulu?



SPARQL

```
Select ?number
Where {
  Res:Honolulu dbo:population ?num
}
```

SQL

```
Select value
From KB
Where subject='d' and
predicate='population'
```

知识问答方法概述

- 规则模板的方法
- 图相似度方法
- 深度学习方法
- 文本混合问答

知识问答方法概述

- 规则模板的方法
 - 通过人工构造规则模板将问题映射到属性

Pattern matching

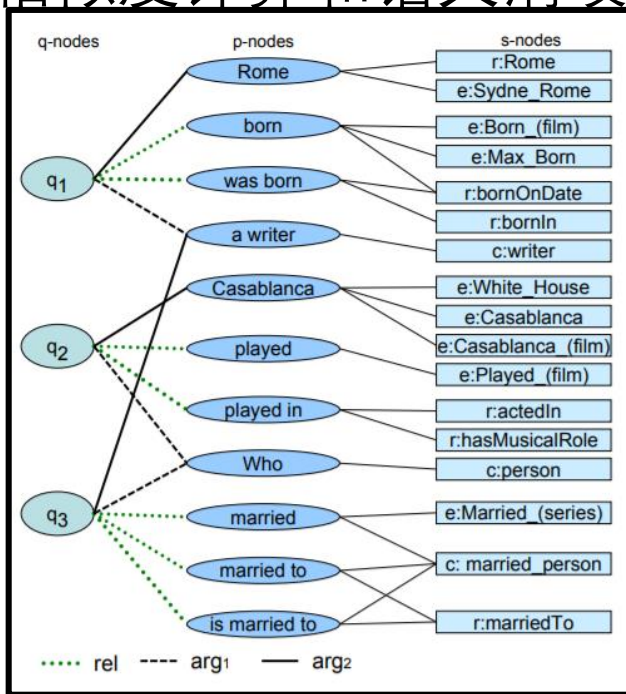
Pattern: ... "capital" ... <country>

Predicate: capital

- 较高的准确率和较低的召回率（对问题多样性的较低覆盖率）
- 可解释性强

知识问答方法概述

- 图相似度方法
 - 基于图做问题-答案的相似度计算和语义消歧

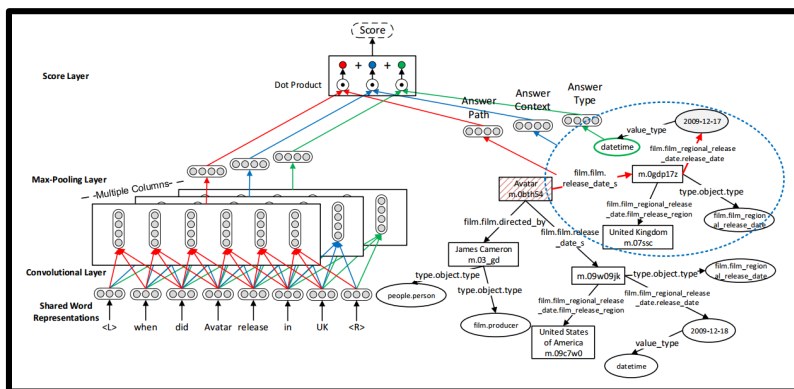


Natural language questions for the web of data,
EMNLP 2012

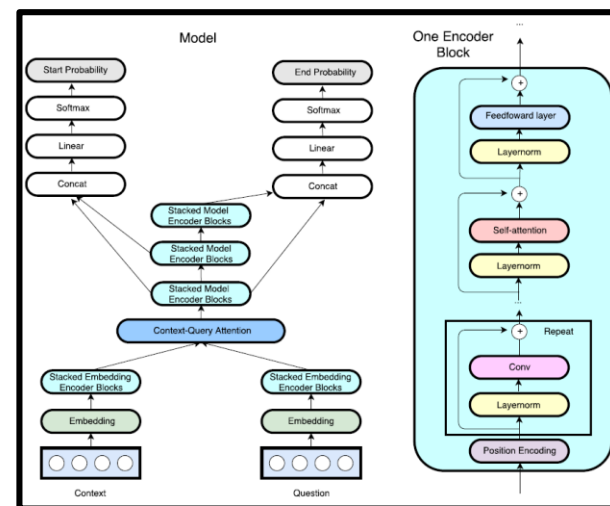
- 召回上升
- 可解释性尚可：来自图的合理性

知识问答方法概述

- 深度学习方法
 - 将离散的问题表示为连续向量
 - 通过深度神经网络理解问题



Question Answering over Freebase with Multi-Column Convolutional Neural Network, ACL 2015

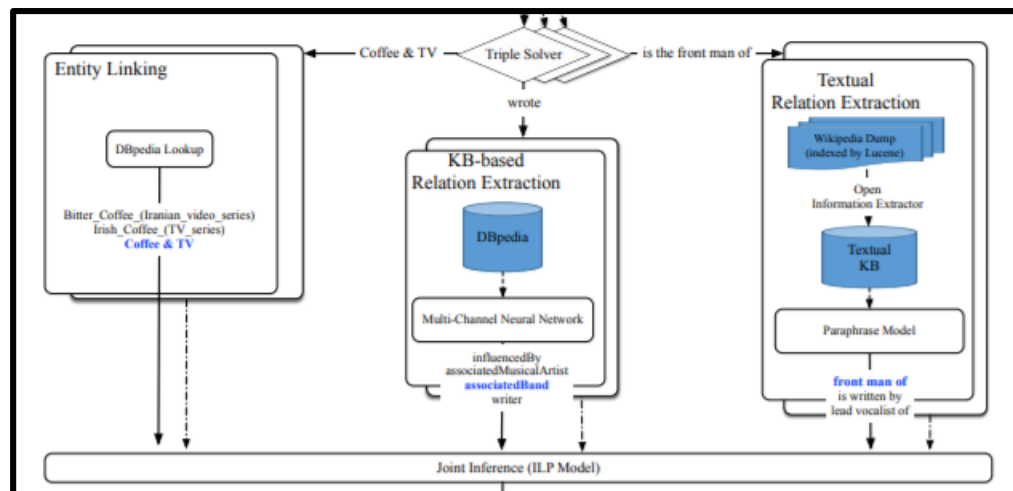


QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension, 2018

- 较高的准确率和召回率
- 可解释性差：人类无法理解和预测神经网络

知识问答方法概述

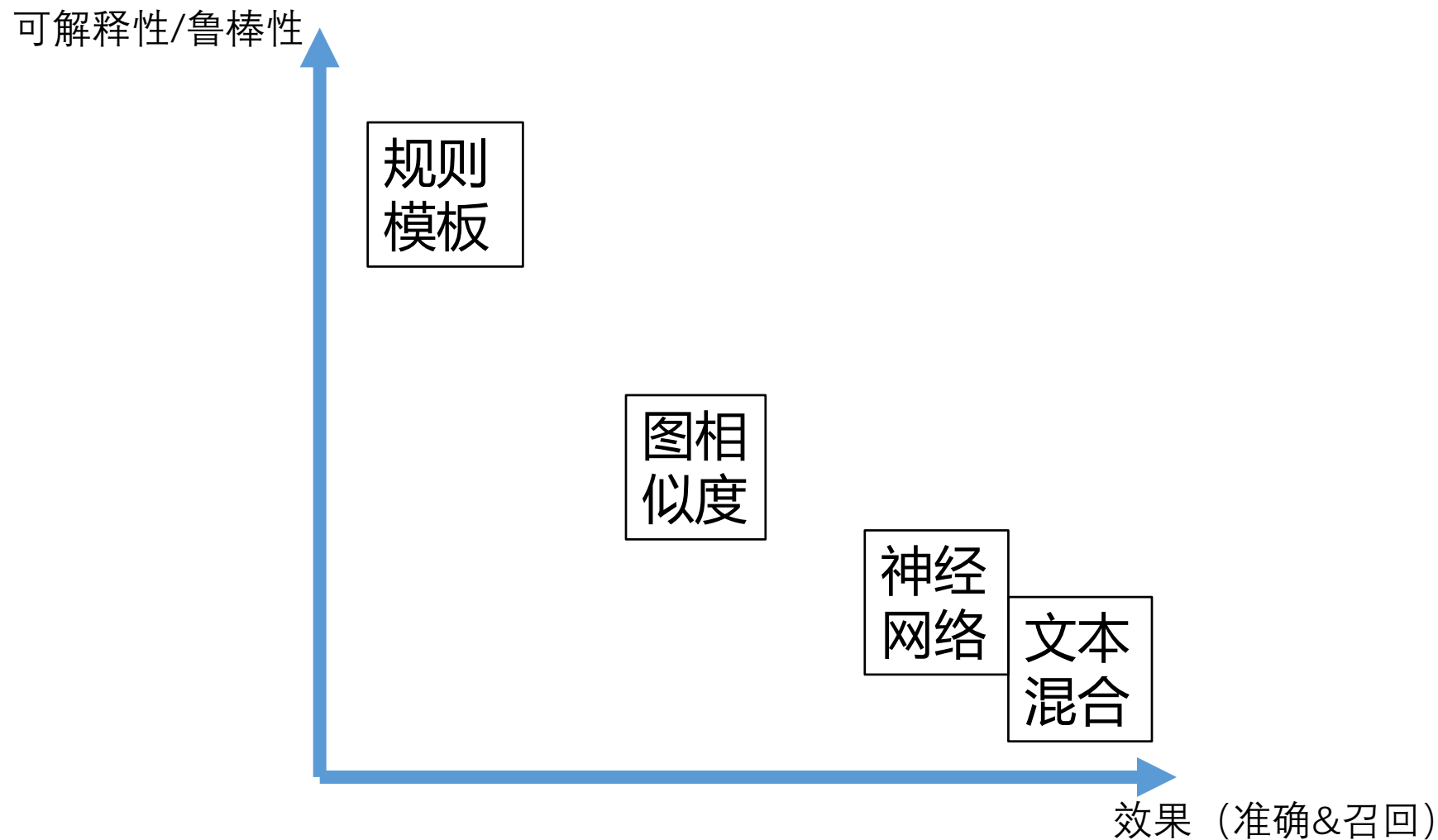
- 文本混合问答
 - 将文本作为补充知识源，解决数据稀疏性问题



Hybrid Question Answering over Knowledge Base and Free Text, COLING 2016

- 更高的准确率和召回率
- 可解释性差
- 适用条件更严格：需要有配对文本

知识问答方法概述: 总结



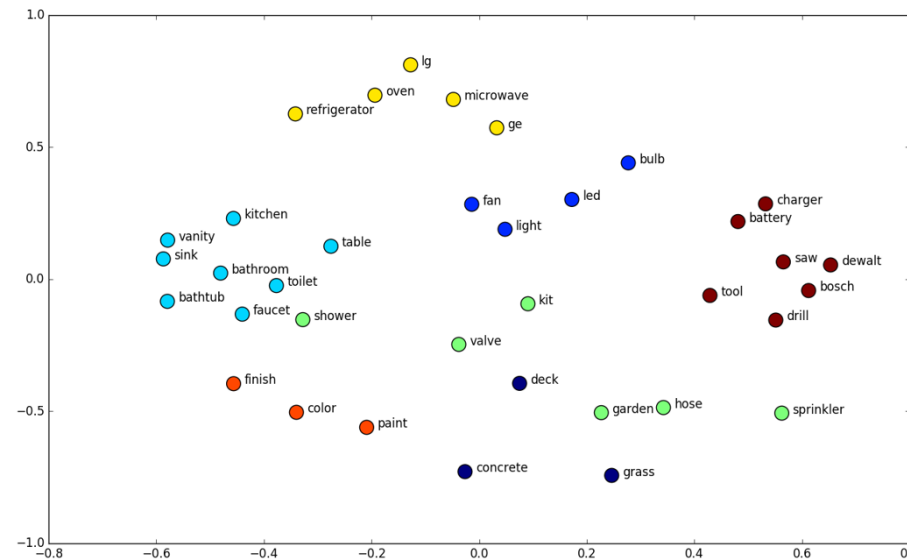
深度学习方法

Problem analysis: intent classification

- Input: question, candidate value
- Output:
 - 0 the value is not the answer
 - 1 the value is the answer

Algorithm: how to understand the questions?

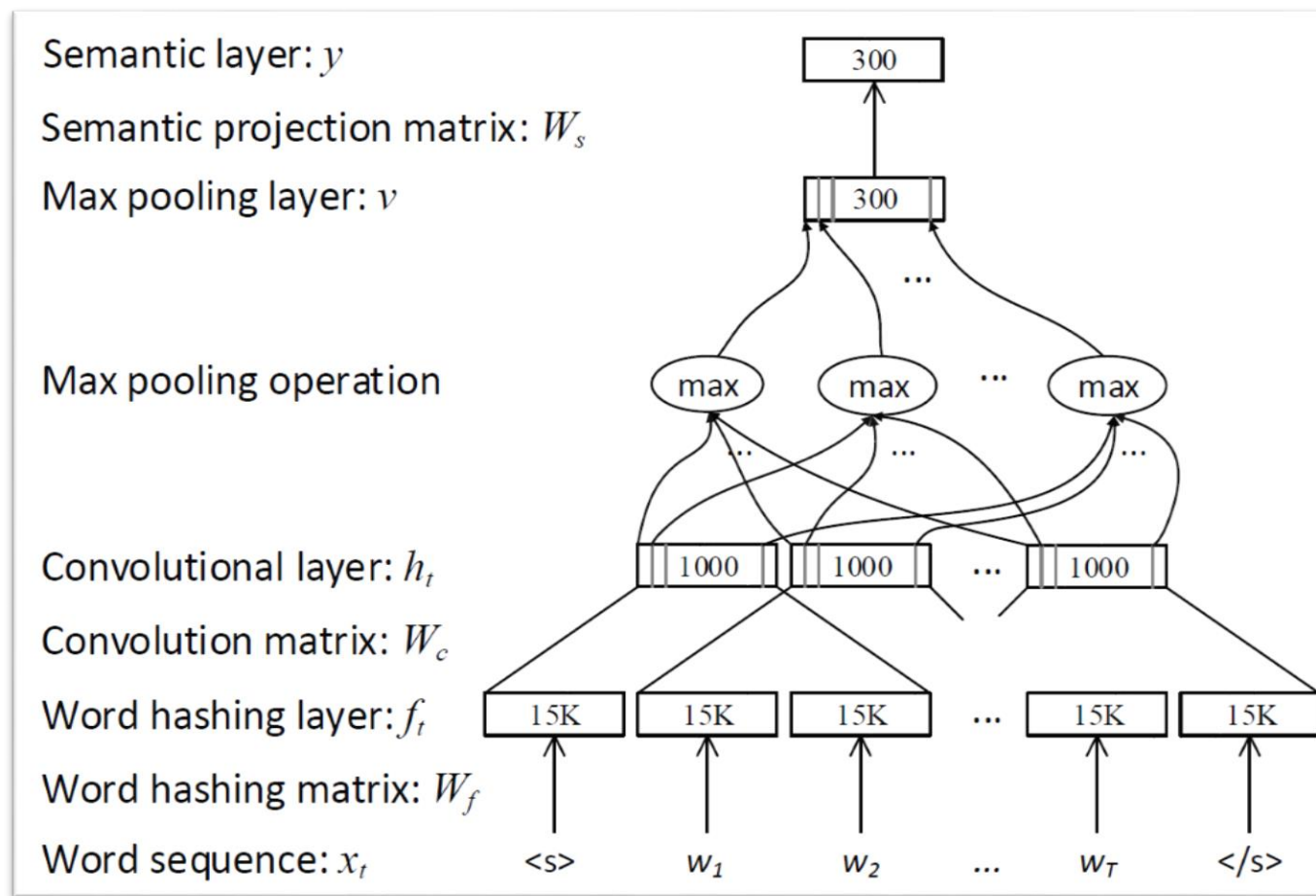
- Word understanding:
 - word embedding
 - Similar semantics have close vectors



Algorithm: how to understand the questions?

- question understanding:
 - Merge the word embeddings
 - Note the sequential feature
 - CNN/RNN

CNN



Multi-granularity DNN: more features hierarchically



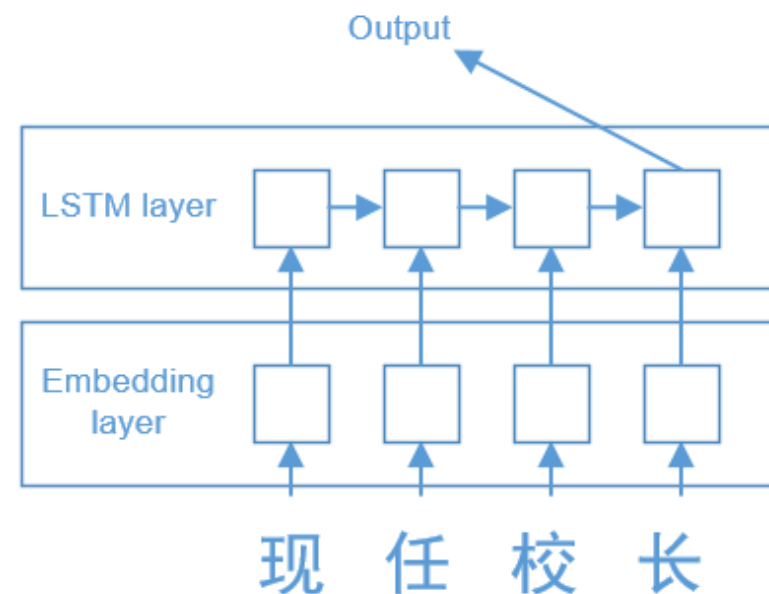
- Fully utilize all the features.
- Hierarchical feature organization.

MGDNN : features

- From questions:
 - Question
 - 上海财经大学的校长是谁
- From knowledge bases:
 - Value
 - 樊丽明
 - Entity description
 - 教授
 - Predicate
 - 现任校长

MGDNN : 3 granularities for 1 feature

- 3 granularities for 1 feature
 - 现任校长
- One hot:
 - 现任校长
- Word list:
 - 现任 校长
- Char list:
 - 现 任 校 长



MGDNN: feature2vector

- 3 representation granularity
- Aggregation

- 现任校长

- One hot:

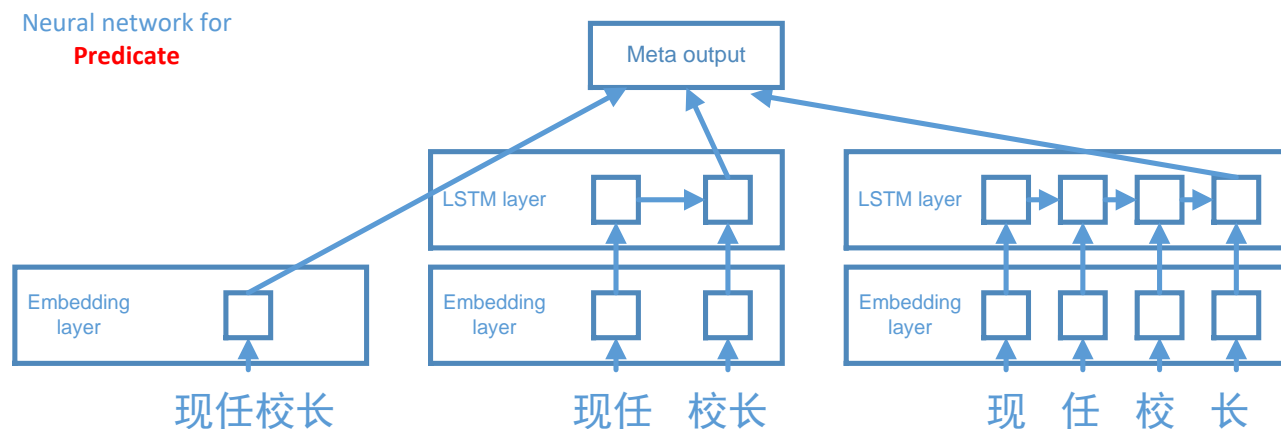
- 现任校长

- Word list:

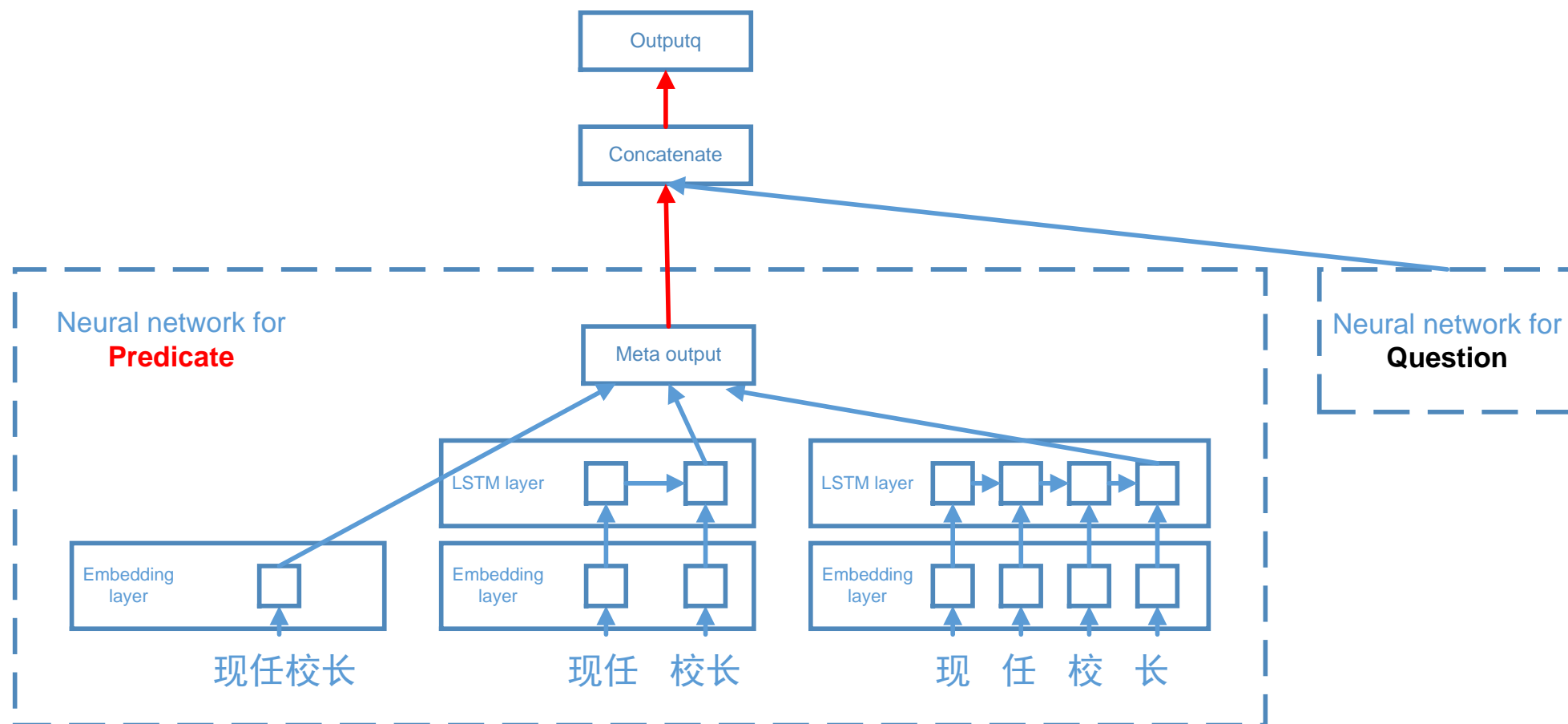
- 现任 校长

- Char list:

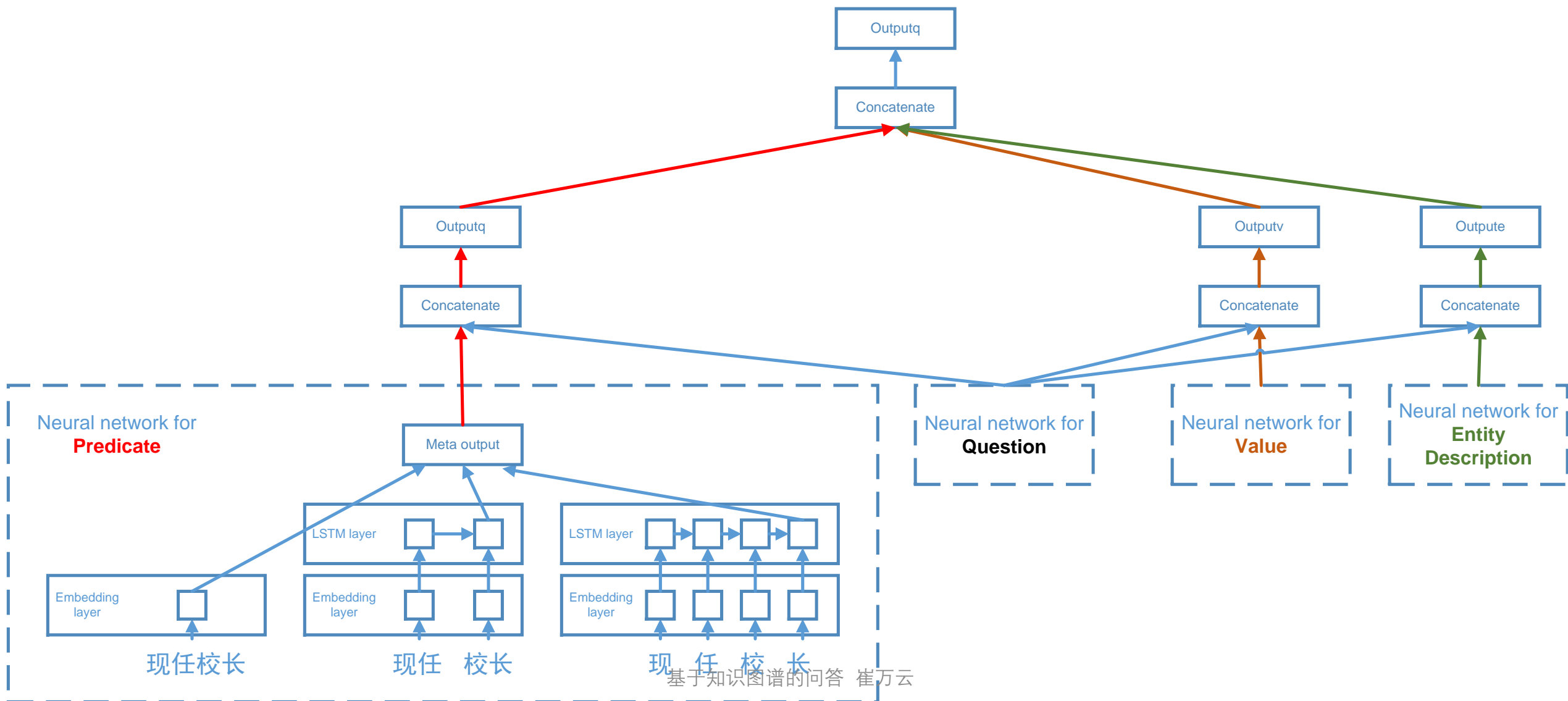
- 现 任 校 长



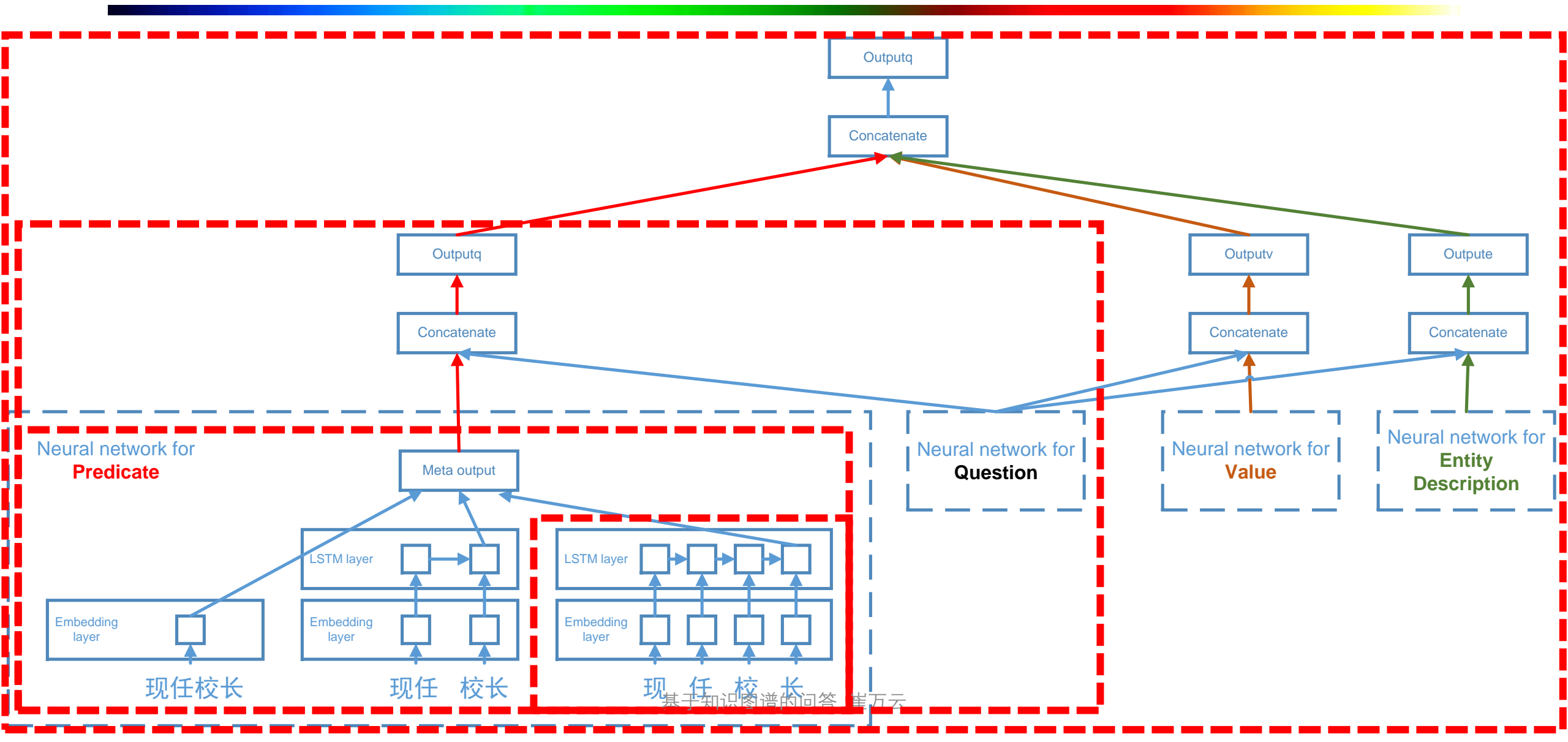
MGDNN: predicate - question similarity



MGDNN



MGDNN



Results



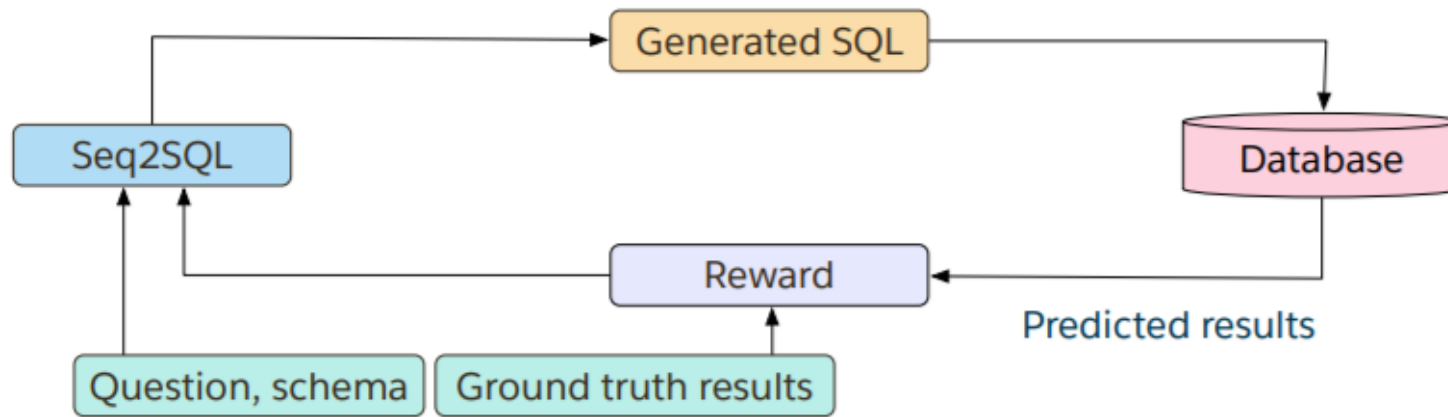
- NLPCC 2016
- Our score: **0.842** (improving)
- KBQA beats all competitors.

KBQA Submissions	F1 Score
北京大学	0.8247
国防科学技术大学	0.8159
华中师范大学	0.7957
哈尔滨工业大学 (HIT-SCIR)	0.7914
东北大学 (自然语言处理实验室)	0.7272
Harbin ShenZhi Technology Co., Ltd.	0.7251

SEQ2SQL: a generative model

- Discriminative model
 - How many people live in USA?
 - Location: 0.04
 - **Population: 0.85**
 - President: 0.02
 - ...
- Generative model
 - How many CFL teams are from York College?
 - **SELECT COUNT CFL Team
FROM CFLDraft WHERE
College='York'**
 - SEQ2SEQ -> SEQ2SQL
 - **PROS: feasible to
complicated intends**

SEQ2SQL: data augmentation by RL



- Seq2SQL
 - Input: a question and the columns of the table
 - Output: generates the corresponding SQL
- Reward: by the results of the SQL against the database

混合问答

Text: extra features for QA

- Data sparsity of KBQA
 - Incompleteness of knowledge graphs
 - Lack of labeled training data
- Text is a good resource to better answer a question

How to adapt the existing algorithms to text features?

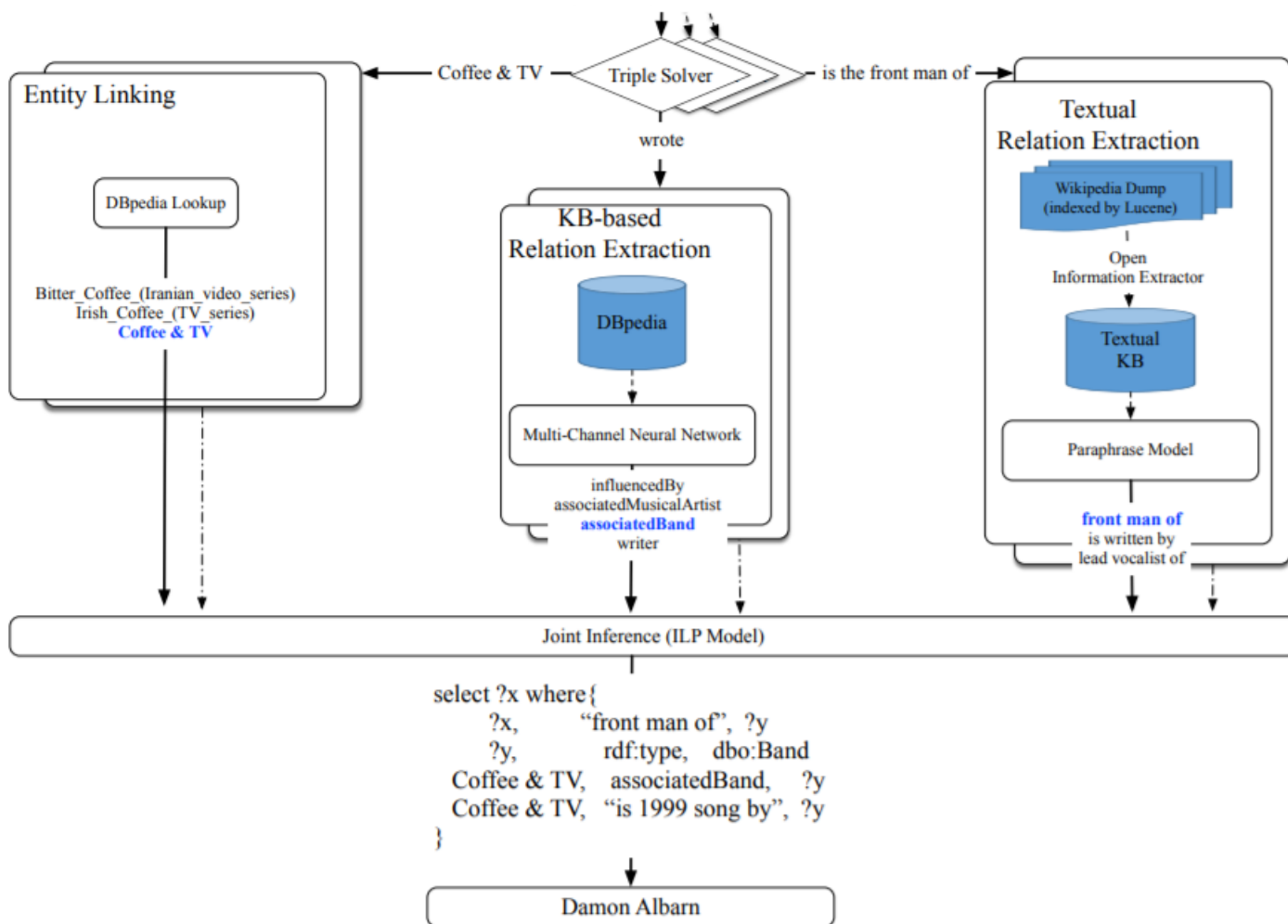
- **Together with KBQA**

- Do KBQA, text understanding simultaneously
- Use joint inference to compute the final answer

- **After KBQA: Answer reranking**

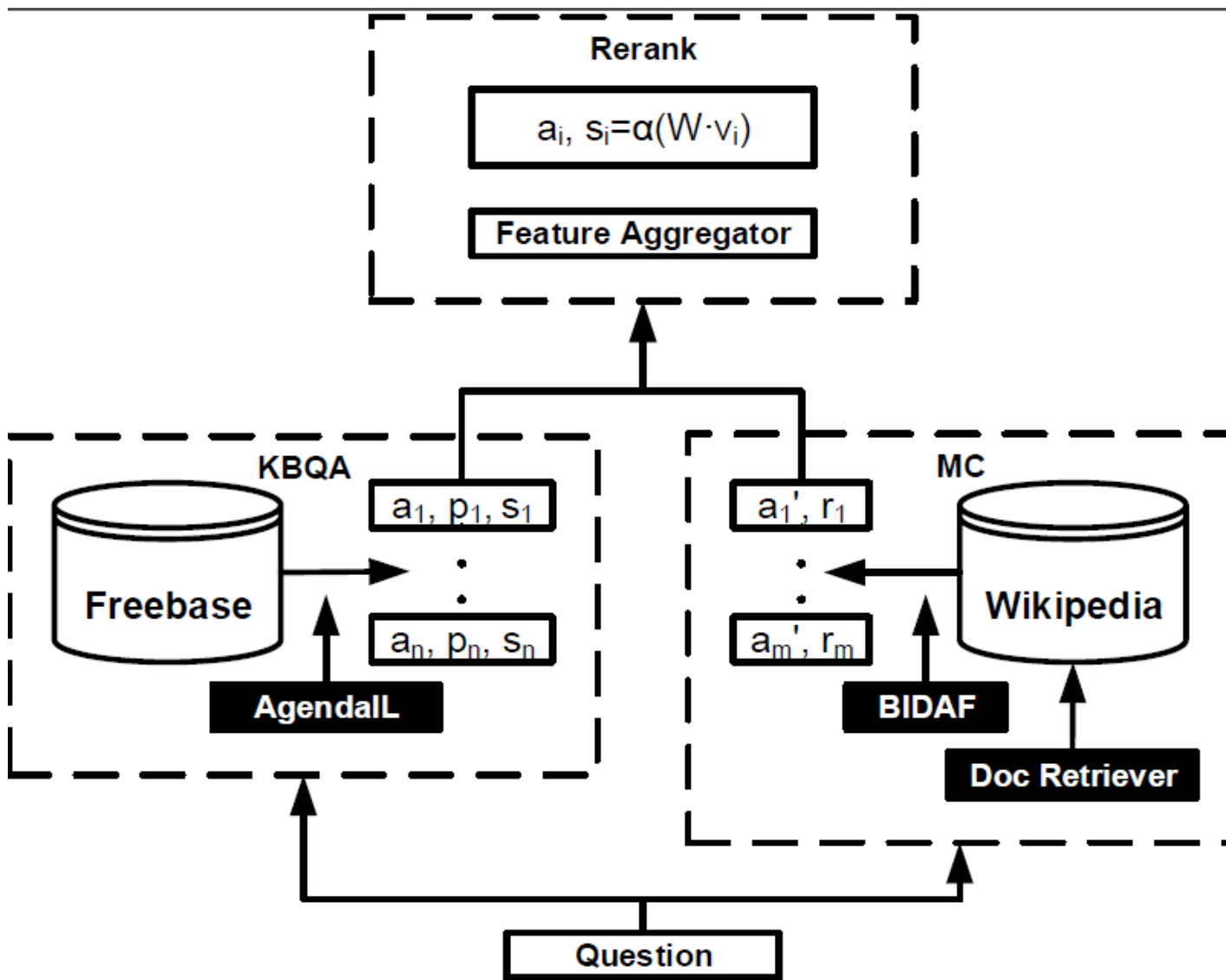
- First do KBQA
 - Get $(\text{answer}_1, \text{score}_1), \dots, (\text{answer}_n, \text{score}_n)$
- Then rerank the answers
 - Features: score_i , **text**

Together with KBQA



- Entity link
- KB-based Relation Extraction
- Textual Relation Extraction
- **Simultaneously**
- Joint inference for heterogeneous features

KBQAr—— Reranking KBQA via Machine Comprehension



- **KBQA**: generates candidate answers and scores by KBQA
- **MC**: provides answers by MC as the extra features
- **Rerank**: reranks answers by KBQA with hybrid features

Experimental Results



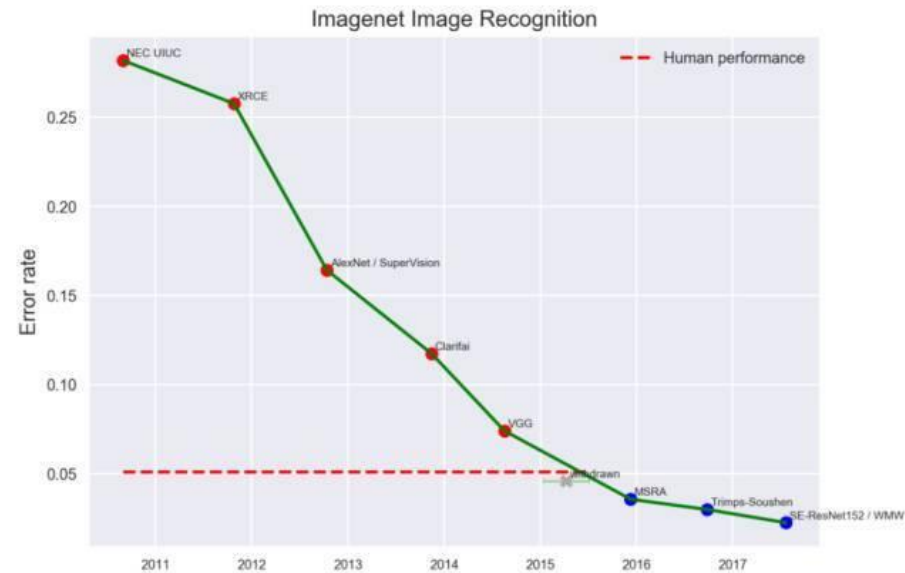
AgendaIL	KBQAr
49.7	50.7 (+1.0)

The f1-score increases of 1.0% for WebQuestions.

This verifies the effectiveness of the reranking.

如何构建更鲁棒的问答系统

Background



SQuAD1.1 Leaderboard

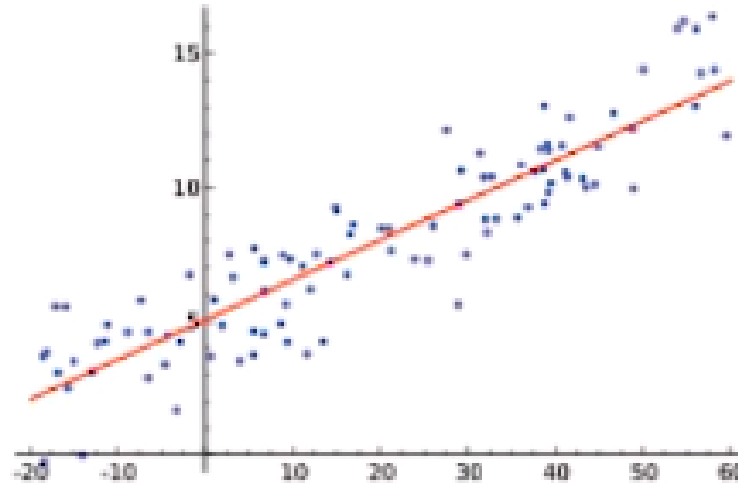
Since the release of SQuAD1.0, the community has made rapid progress, with the best models now rivaling human performance on the task. Here are the ExactMatch (EM) and F1 scores evaluated on the test set of v1.1.

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Jul 12, 2018	QANet (ensemble) Google Brain & CMU	84.454	90.490
2 Jul 09, 2018	r-net (ensemble) Microsoft Research Asia	84.003	90.147
3 Jun 21, 2018	MARS (ensemble) YUANFUDAO research NLP	83.982	89.796
4 Mar 20, 2018	QANet (ensemble) Google Brain & CMU	83.877	89.737
5 Jun 21, 2018	MARS (single model) YUANFUDAO research NLP	83.122	89.224

- State-of-the-art models beat human

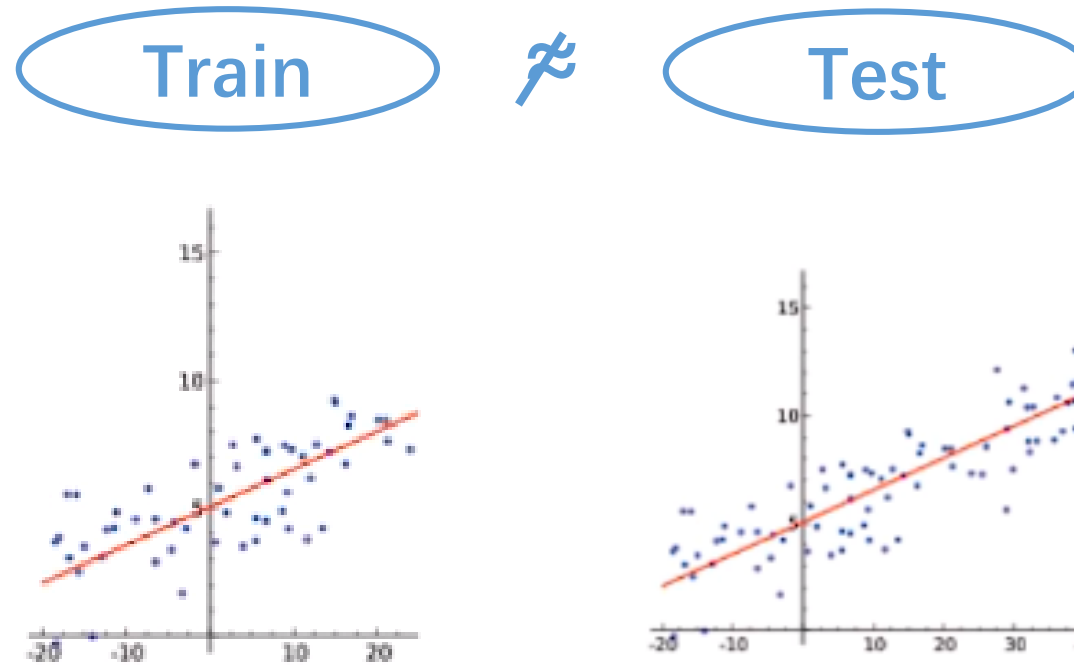
Are they robust?

Are they robust enough?



- How to ensure the system works?
 - Controllability

Are they robust enough?



- How to adapt to new data?
 - Adaptation

Percy Liang's Question

- IRQA

Article: Super Bowl 50 (from SQuAD dataset)

Paragraph: "Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. **The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38** and is currently Denver's Executive Vice President of Football Operations and General Manager.

Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."

Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

BiDAF Prediction under adversary: Jeff Dean

BiDAF Original Prediction: John Elway

*Adversarial Examples for Evaluating Reading
Comprehension Systems Robin Jia et al. 2017*

Percy Liang's Question

Reading comprehension

Model	Original F1	Adversarial F1
Humans	92.6	89.2
SLQA+	88.6	64.2
r-net+	88.5	63.4
ReasoNet-E	81.1	49.8
SEDT-T	80.1	46.5
BiDAF-E	80	46.9
Mnemonic-E	79.1	55.3
Ruminating	78.8	47.7
jNet	78.6	47
Mnemonic-S	78.5	56
ReasoNet-S	78.2	50.3
MPCM-S	77	50

If retrain model on
adversarial examples
(appended)

F1: 74.3 ->
70.0

If test this model on
prepended sentences

F1: 70.0 ->
36.9

*Adversarial Examples for Evaluating Reading
Comprehension Systems Robin Jia et al.
2017*

Are they robust?- **NO**

Fixes



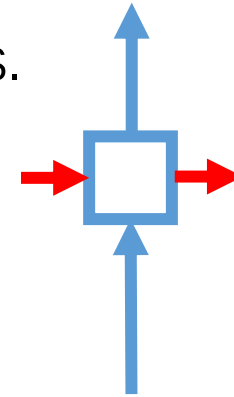
Adaptation to new data

Controllability

Adaptation by Transfer Learning

Introduction

- Deep learning (DL) has shown great values and potentials in question understanding.
- The input sentence is a word sequence.
 - The orders of the words affect their semantics.
- Modeling:
 - recurrent neural network
 - Use sequential memory from the former unit

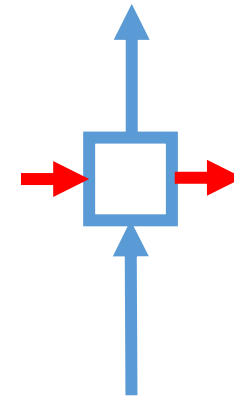


Introduction

- Models for the open domain usually incur more errors in specific domains.
- The insufficiency of the training data for specific domains
- Solution: **transfer knowledge** from the source domain to the target domain

Challenge

- Modeling:
 - recurrent neural network
 - Use sequential memory from the former unit
- How to transfer knowledge?
- How to reflect the sequential memory in the transfer?



How to transfer knowledge?

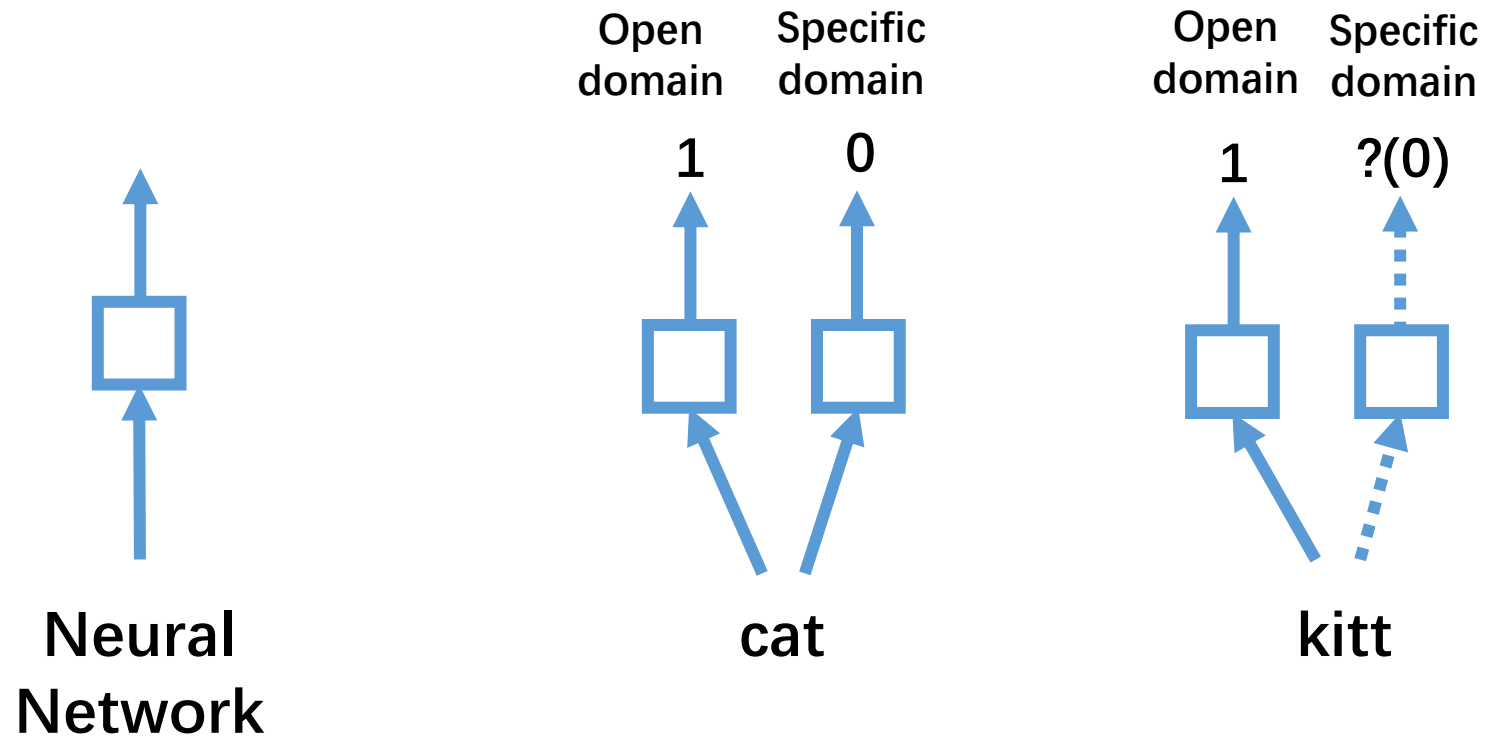
- When to transfer?
 - The target domain's parameters are well-trained
 - No
 - The target domain's parameters are trained insufficiently.
 - Yes
- The target domain's label is the same as the source domain's.
 - No -> directly use the source domain's prediction
- The target domain's label is different from the source domain's.
 - Yes

How to transfer knowledge?

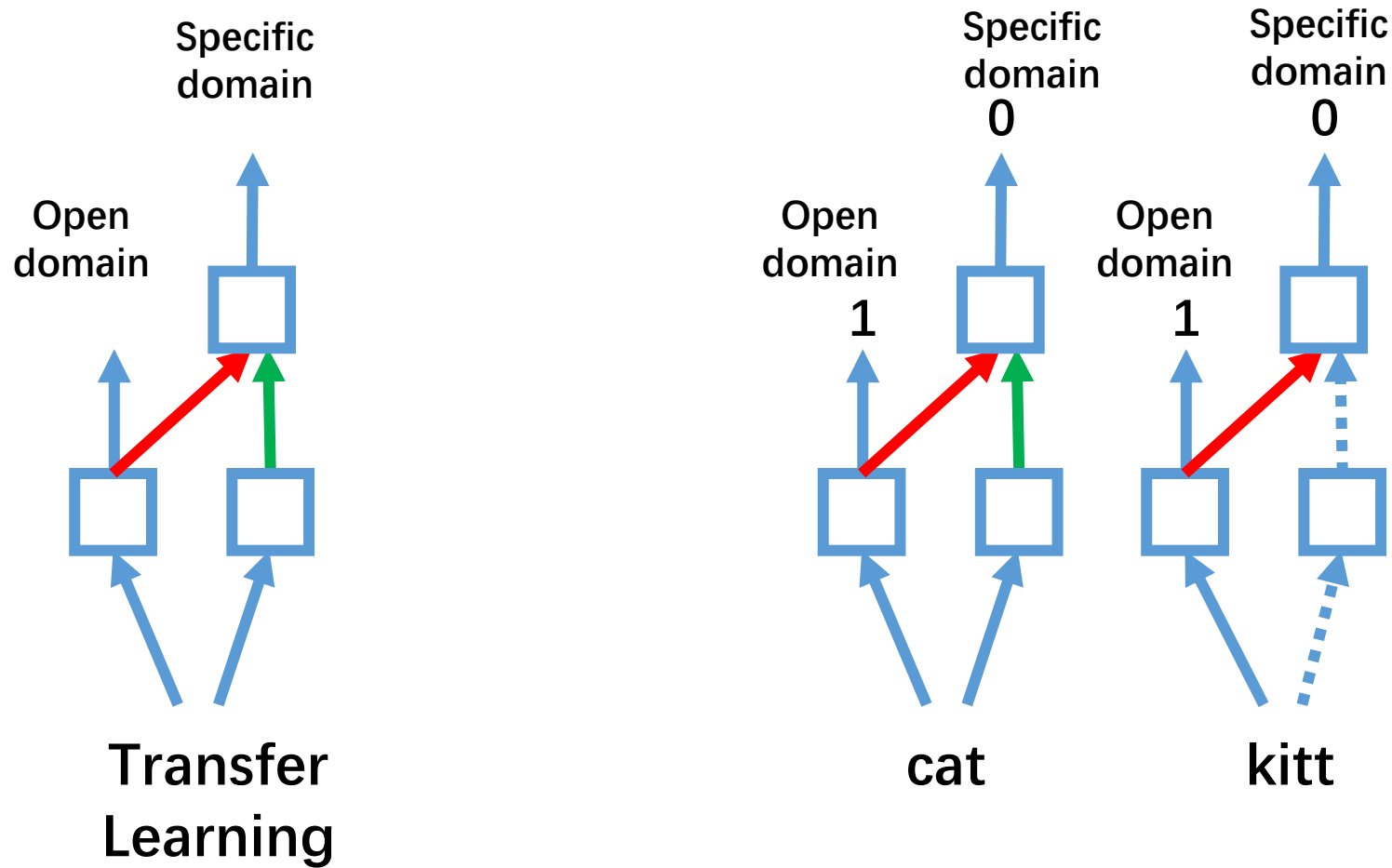


- When to transfer?
 - The target domain's parameters are trained insufficiently.
 - The target domain's label is different from the source domain's.

How to transfer knowledge?



How to transfer knowledge?

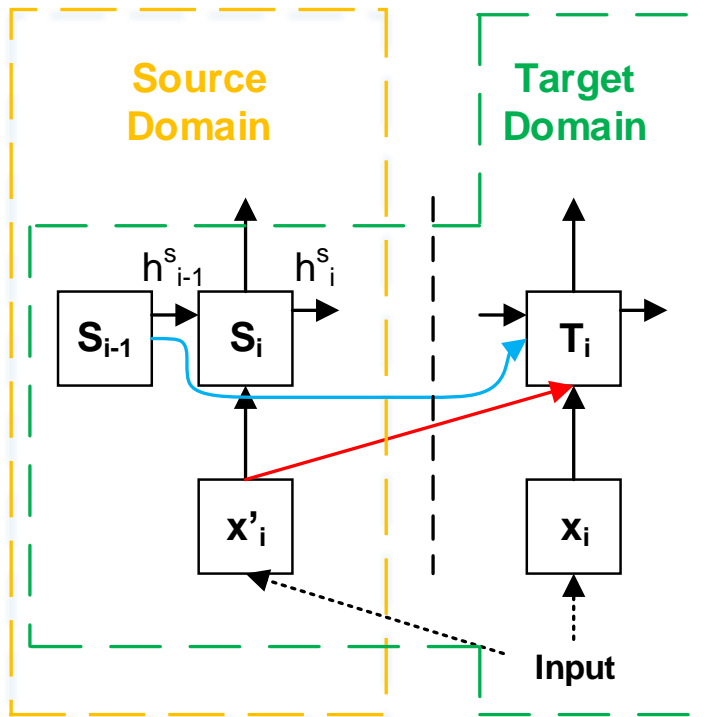


Open domain knowledge transferred through the red edge.

How to transfer sequential memory?

- Transferable RNN

- Adding the source domain's corresponding sequential memory to the target domain.



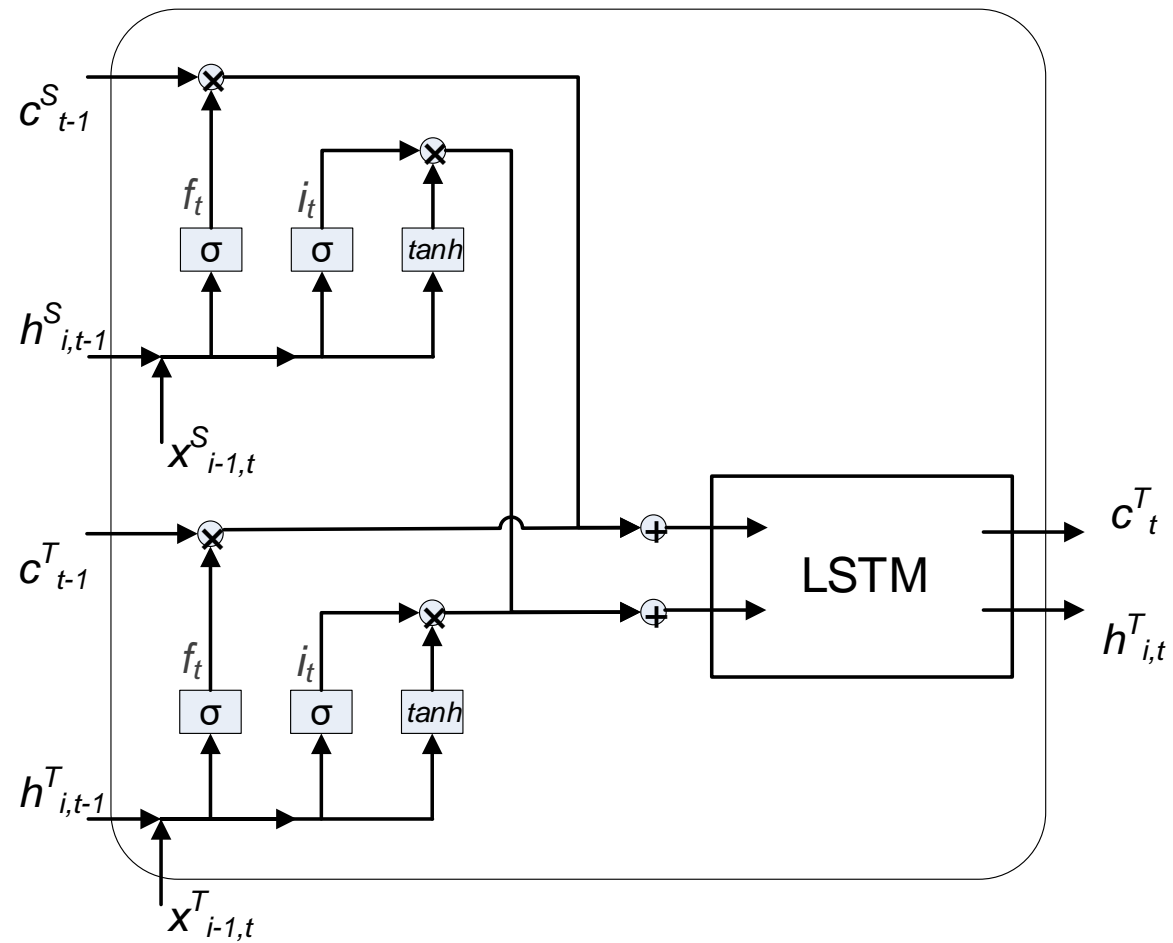
The sequential memory is transferred through the blue edge.

The input knowledge is transferred through the red edge.

T-LSTM


- A specific implementation of the transferable RNN under LSTM
- Key
 - Allow the sequential memory transferring
 - Transferable RNN framework
 - To what extend the sequential memory is transferred
 - If the source domain's parameters are trained sufficiently, use the source domain more - and vice versa.
 - Use gates to decide to what extend the knowledge is retained.

T-LSTM



QA Controllability

by Learning Question Answering with Templates and Knowledge Bases



Weakness of previous works

- Template / rule based approaches

- Represent sentences by templates
- By human labeling
- PROs:
 - User-controllable
 - Applicable to industry use
- CONs:
 - Relies on manpower. Too costly.
 - Cannot handle the **diversity** of questions.

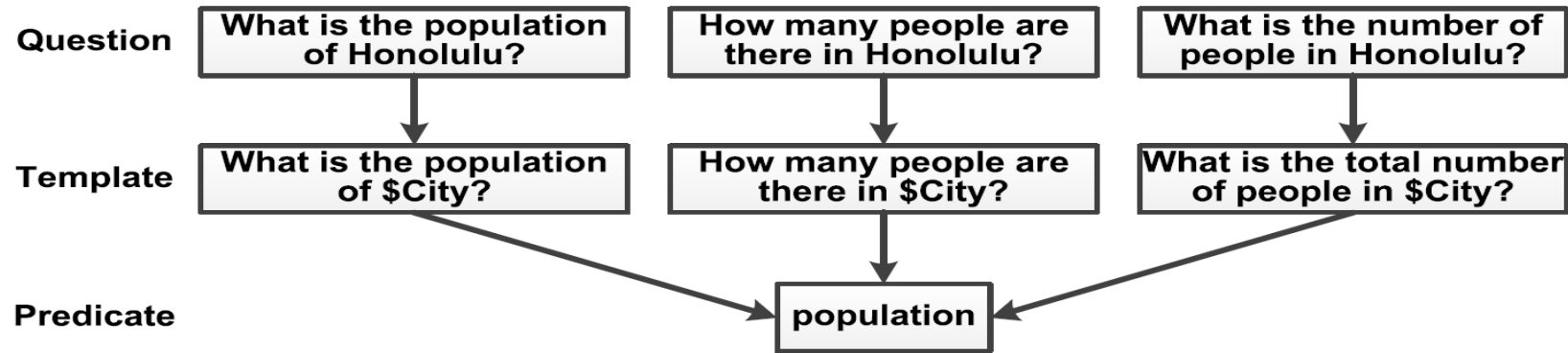
- Neural network based approaches

- Represent sentences by embeddings
- By learning from corpus
- PROs:
 - Feasible to understand diverse questions
- CONs:
 - Poor interpretability
 - **Not controllable.** Unfriendly to industrial application.

Our approach

- Represent natural language questions by templates.
 - E.g.
 - How many people are there in \$city?
 - Interpretable
 - User-controllable
- Learn templates from QA corpus.
 - Understand diverse questions
 - 27 million templates, 2782 intents

QA by templates



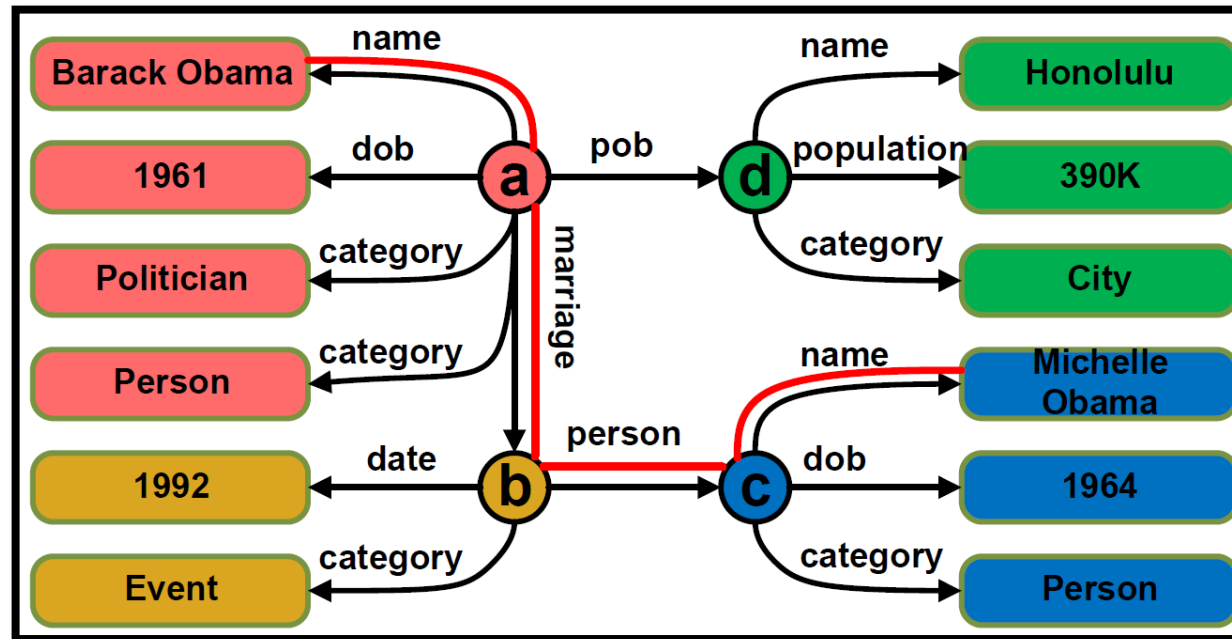
Template: replace the entity of question by its concept

A template represents complete intent of the question.

Key problem: collect templates and identify their corresponding predicates

Q2A: a generative process

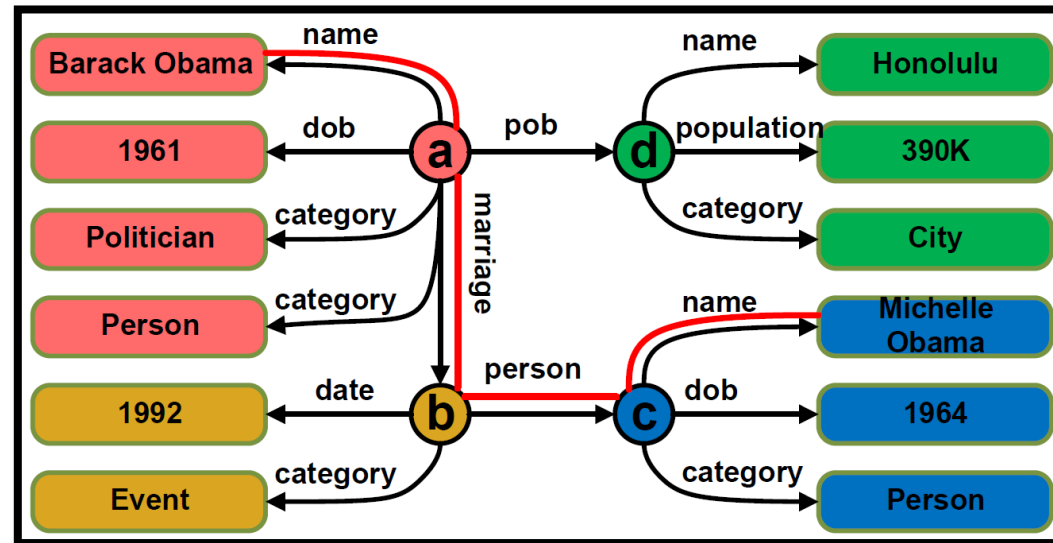
- A qa pair
 - Q: How many people live in Honolulu?
 - A: It's 390K.



Q2A: entity linking

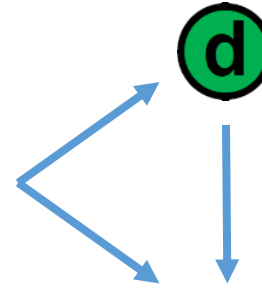
How many people live in Honolulu?

d

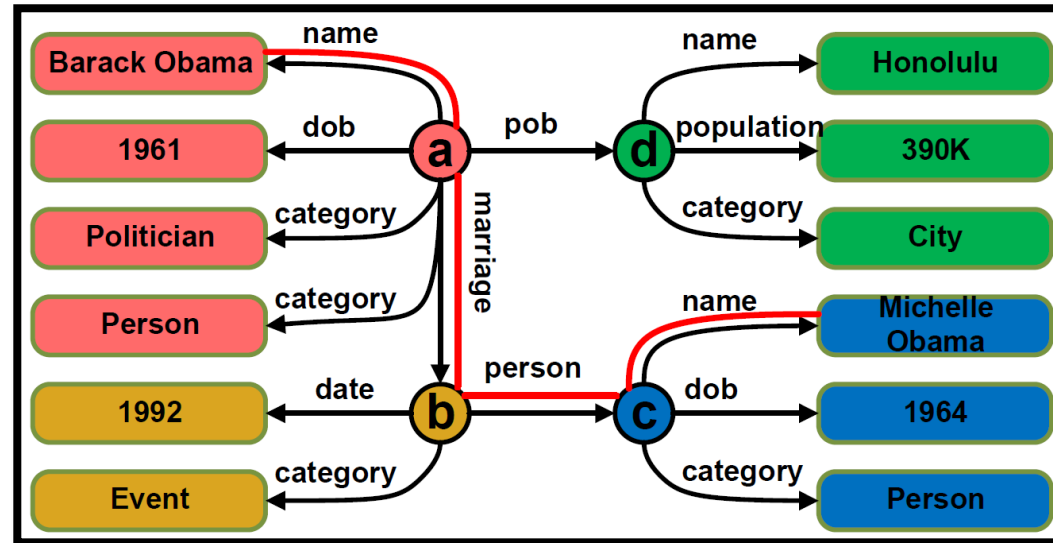


Q2A: conceptualization

How many people live in Honolulu?

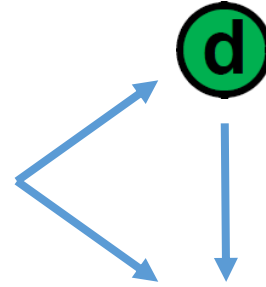


How many people live in \$city?



Q2A: predicate inference

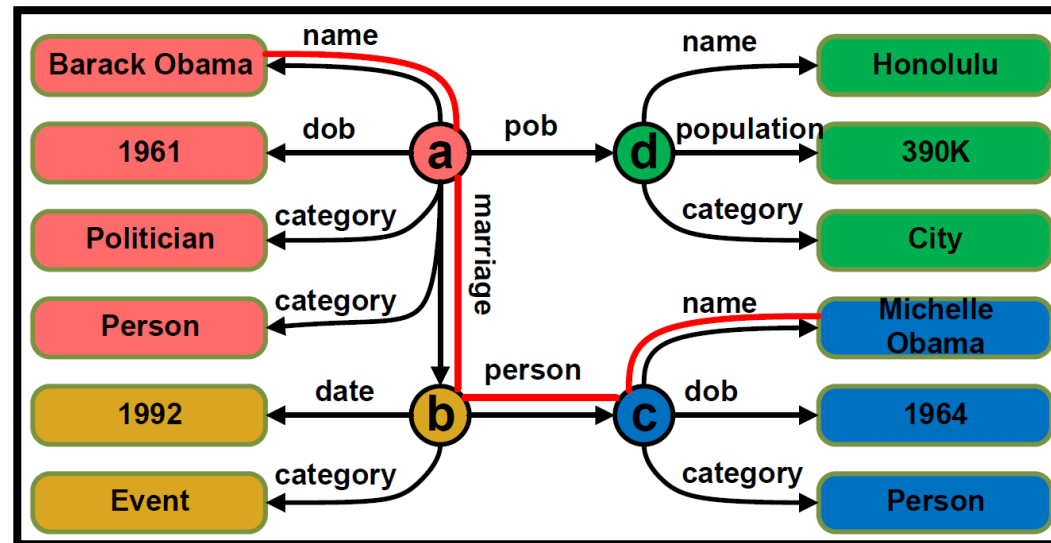
How many people live in Honolulu?



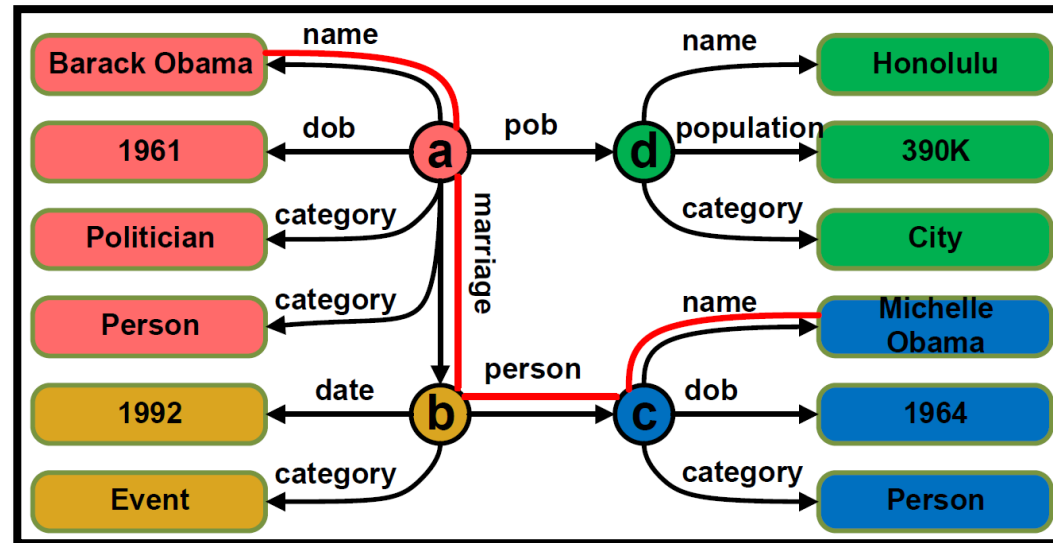
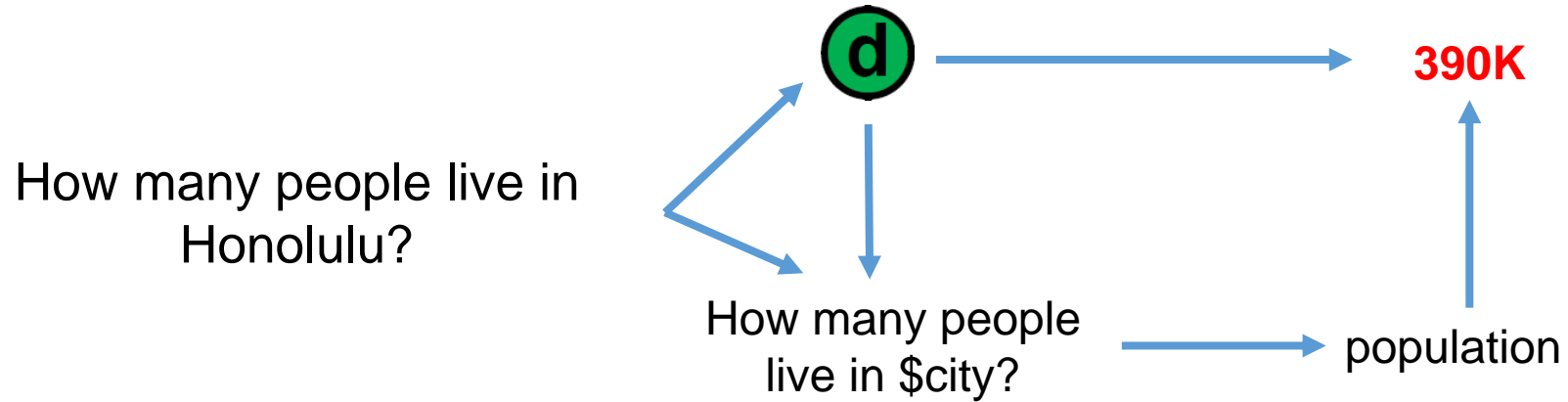
How many people
live in \$city?



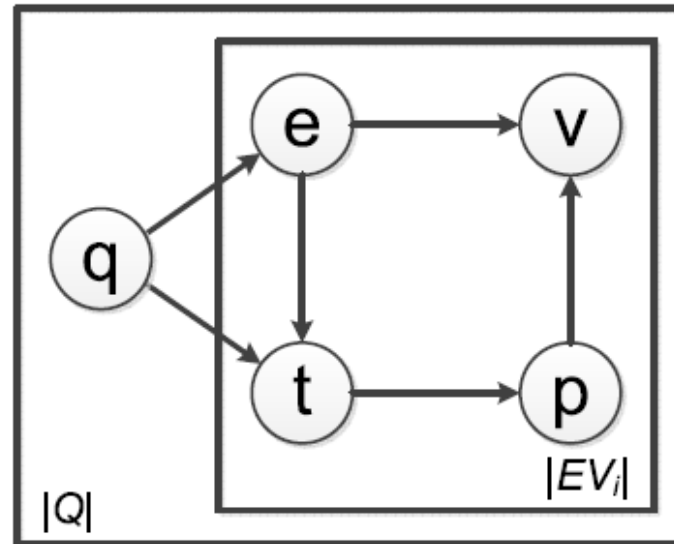
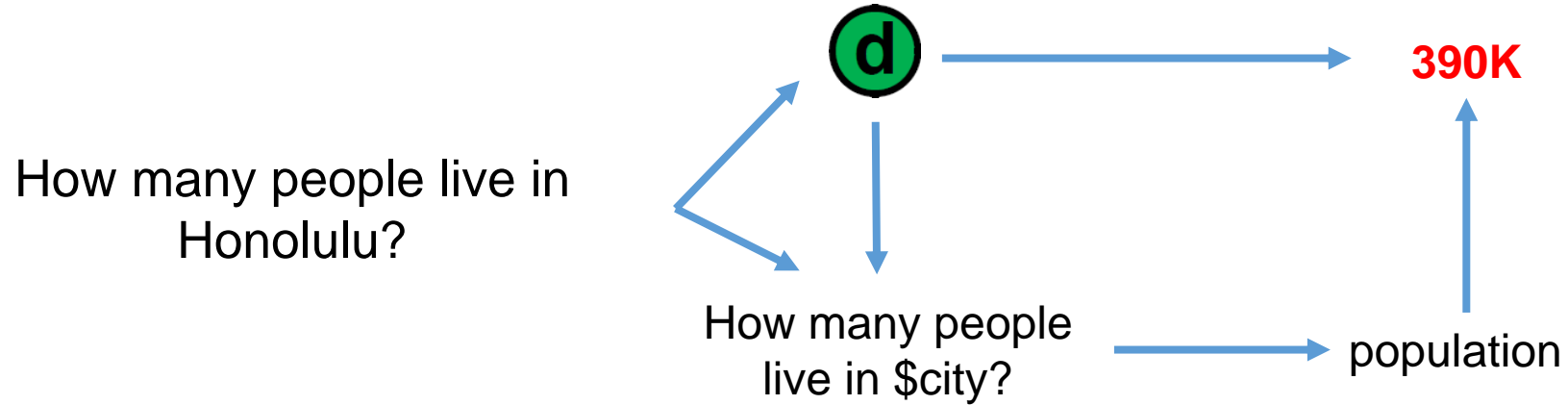
population



Q2A: value lookup

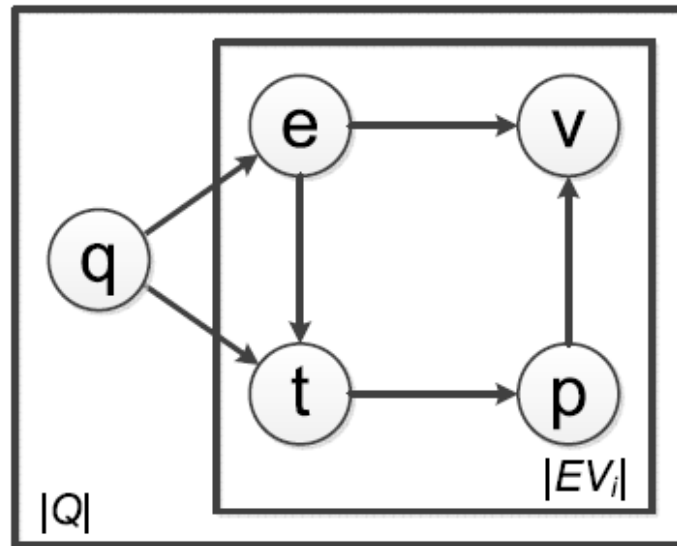


Probabilistic graph model



Probabilistic inferencing

- Learning parameters from QA corpora (42M Yahoo! Answers)
 - Intuition: **maximize the likelihood** of observing such QA corpora



Results

- 27126355个问题模板
- 2782个问题意图

	#pro	#ri	#par	R	R _{BFQ}	R*	R* _{BFQ}	P	P _{BFQ}	P*	P* _{BFQ}
squall2sparql	96	80	13	.78	.81	.91	.94	.84	.95	.97	.95
SWIP	21	14	2	.14	.24	.16	.24	.67	.77	.76	.77
CASIA	52	29	8	.29	.56	.37	.61	.56	.79	.71	.86
RTV	55	30	4	.30	.56	.34	.56	.55	.72	.62	.72
gAnswer [38]	76	32	11	.32	.54	.43	-	.42	.54	.57	-
Intui2	99	28	4	.28	.54	.32	.56	.28	.54	.32	.56
Scalewelis	70	32	1	.32	.41	.33	.41	.46	.50	.47	.5
KBQA+KBA	25	17	2	.17	.42	.19	.46	.68	.68	.76	.76
KBQA+FB	21	15	3	.15	.37	.18	.44	.71	.71	.86	.86
KBQA+DBp	26	25	0	.25	.61	.25	.61	.96	.96	.96	.96

QALD-3上的结果 (KB-based)

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



微信端界面

展望

Future work

- How to build a better QA system?
 - Precision/recall
 - Better model
 - More data (text/images etc.)
 - Adaptation
 - Transfer learning
 - More
 - Controllability
 - Learning question templates
 - More