

Chatbot in DSNO

Wabywang, Allenzhai, WinsterLin, AndyweiZhao

- 团队介绍
- 问答任务分类
- 通用算法介绍
- Chatbot in DSNO

Our Teams

- 林全郴 (Winsten)
 - 多年互联网从业经历，在系统架构领域和算法均有很深的积淀
- 翟铮 (Allen)
 - 浙大数学博士，发表了 **TPAMI** ^[1] 论文，为计算机专业内影响因子最高的期刊
- 赵伟 (Andy)
 - 发表有计算机顶级会议SIGIR、EMNLP和CIKM。
- 高永强 (Jason)
 - 腾讯绝艺核心算法人员之一，从TEG转来，现在优图

[1] IEEE Transactions on Pattern Analysis and Machine Intelligence

About me [2]

- 王本友

- 社交网络运营部/数据应用中心/知识发现组，腾讯量子企鹅计划首批成员
- 发表有信息检索领域顶级会议SIGIR、CIKM，AI顶级会议AAAI，累积论文数量达到10篇（4篇与QA有关），累积谷歌引用16。
- 开源多个深度学习相关的github项目，其中IRGAN项目获得247颗星，业内有较大影响力

- News

- 已与清华大学出版社签约，明年上半年出版《推荐系统与深度学习》一书
- 最新关于问答的长文被AAAI 2018接受，为AI领域最顶级会议
- 关于问答的长文被SIGIR 2017接受，并获得**最佳论文提名奖**（列第二）

Paper 一览

- [1] Chen Y, Zhang P, Song D, **Wang B**. A Real-Time Eye Tracking Based Query Expansion Approach via Latent Topic Modeling[C]//**CIKM** 2015. ACM, 2015: 1719-1722. [CCF-B]
- [2] **Wang B** et al. Exploration of Quantum Interference in Document Relevance Judgement Discrepancy[J]. Entropy, 2016 [SCI-3]
- [3] **Wang B** et al. A Chinese Question Answering Approach Integrating Count-based and Embedding-based Features[C]NLPCC 2016. [EI]
- [4] Zhang P, Li J, **Wang B**, et al. A Quantum Query Expansion Approach for Session Search[J]. Entropy, 2016, 18(4): 146. [SCI-3]
- [5] Li J, Wu Y, Zhang P, Song D, **Wang B**. Learning to Diversify Web Search Results with a Document Repulsion Model[J] **Information Sciences**. 2017. [SCI-2] [Top]
- [6] Wang J, Yu L, Zhang W, Gong Y, Xu Y, **Wang B** et al. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models[C]. **SIGIR** 2017 long paper. [CCF-A] (three strong-accept reviews and **SIGIR best paper** Honorable Mention)
- [7] Zhang S, Hou Y, **Wang B**, et al. Regularizing Neural Networks via Retaining Confident Connections[C], Entropy, 2017. [SCI-3]
- [8] Shang ZG, et.al How Users Select Query Suggestions Under Different Satisfaction States? CCIR 2017.
- [9] Su Z, **Wang B**, et.al Enhanced Embedding based Attentive Pooling Network for Answer Selection. NLPCC 2017
- [10] Zhang P, Niu J, Su Z, **Wang B**, et.al A End-to-end quantum language model in question answering, **AAAI** 2018

问答系统分类



问答手段分类

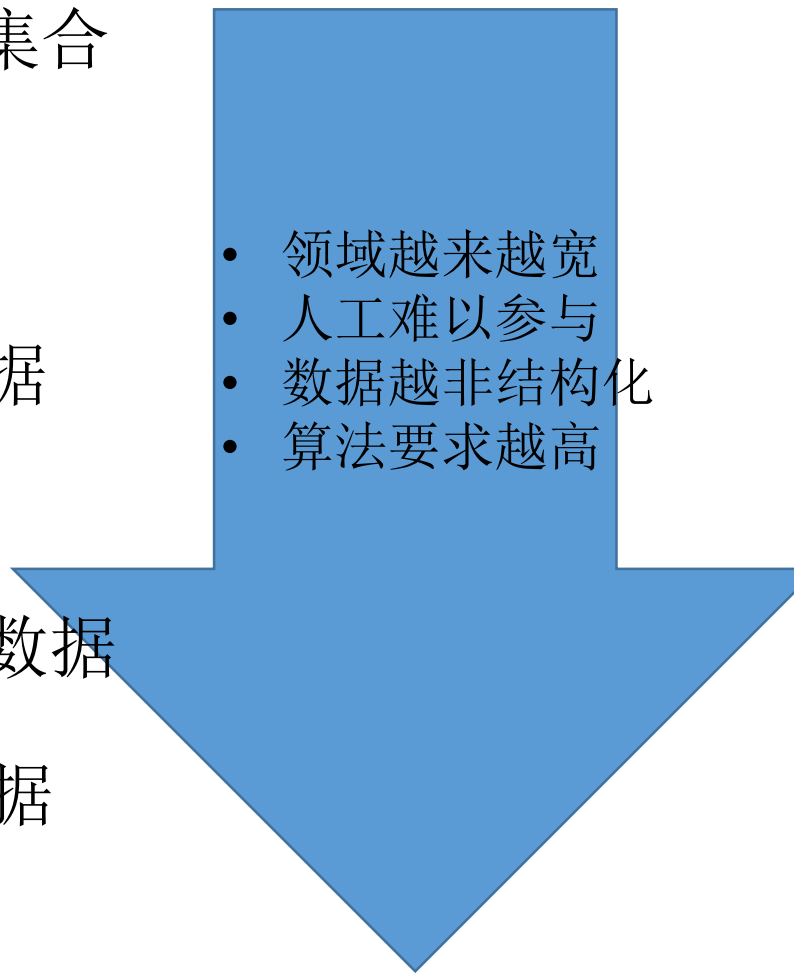
- 基于**候选句子**的问答
 - FAQ 问题和**问题**的匹配
 - 问答匹配 问题和**候选答案**的匹配。
 - 闲聊 大量候选回答里面的匹配
 - 社区问答 知乎最佳答案
- 基于**知识库**的匹配
 - 知识图谱
 - 搜索引擎
 - 基于关系型数据的问答
- 基于**上下文**的匹配
 - 阅读理解 （政企类匹配方式）
- 基于**图像**的问答（VQA）
 - 看图问问题

--候选问答集合

--结构化数据

--非结构化数据

--非文本数据


- 
- 领域越来越宽
 - 人工难以参与
 - 数据越非结构化
 - 算法要求越高

Demo1 : 知识图谱

Who is the wife of Barack Obama?

Web Images Videos Maps News

55,700 RESULTS Any time ▾



Barack Obama · Spouse

Michelle Obama

(m. 1992)

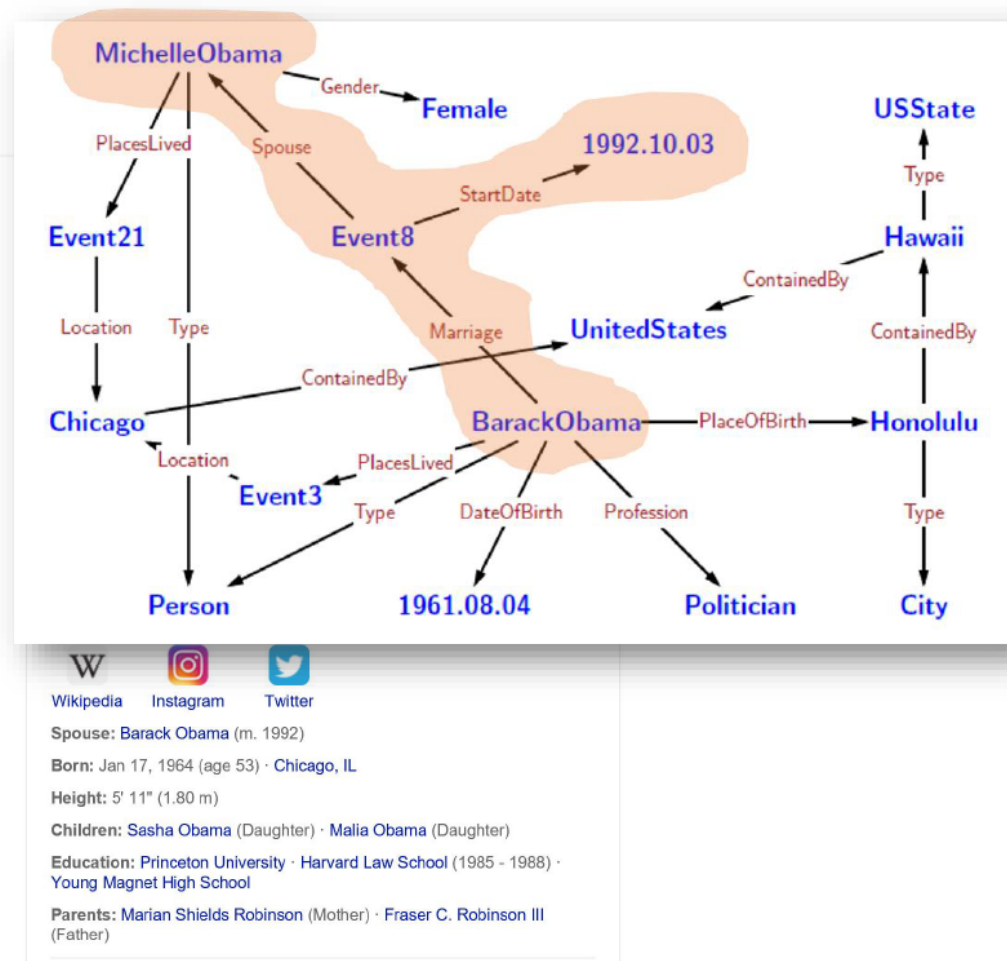

Michelle Obama - Wikipedia
https://en.wikipedia.org/wiki/Michelle_Obama
Barack Obama wrote in his second book, ... She met with Peng Liyuan, the wife of Chinese President Xi Jinping, visited historic and cultural sites, ...

Michelle LaVaughn Robinson Obama (born January 17, 1964) is an American lawyer and writer who was First Lady of the United States from 2009 to 2017. She is married to the ...

They married in October 1992, and have two daughters, Malia Ann (born 1998) and Natasha (known as Sasha, born 2001). Michelle Obama's mother, Marian Robinson was ...

Who is Barack Obama's wife - Answers.com
www.answers.com > ... > **Barack Obama** > **Who is Barack Obama's wife?** ▾
Who is Barack Obama's wife? SAVE CANCEL. already exists. Would you like to merge this ... Both of Barack Obama's parents are deceased. His father ...

Images of who is the wife of barack obama?
bing.com/images



缺点：构建结构性数据库难度大

Cited by Msra 段楠, Search by Bing

Demo2：基于上下文文档问答

 进

声明：百科词条人人可编辑，词条创建和修改均免费，绝不存在官方及代理商付费代编，请勿上当受骗。 [详情>>](#)

[首页](#) [分类](#) [特色百科](#) [用户](#) [权威合作](#) [手机百科](#)

腾讯

 编辑



 收藏

 24117

 1500

深圳市腾讯计算机系统有限公司成立于1998年11月^[1]，由马化腾、张志东、许晨晔、陈一丹、曾李青五位创始人共同创立。^[1]是中国最大的互联网综合服务提供商之一，也是中国服务用户最多的互联网企业之一。^[2]

腾讯多元化的服务包括：社交和通信服务QQ及微信/WeChat、社交网络平台QQ空间、腾讯游戏旗下QQ游戏平台、门户网站腾讯网、腾讯新闻客户端和网络视频服务腾讯视频等。^[3]

2004年腾讯公司在香港联交所主板公开上市（股票代码00700），董事会主席兼首席执行官是马化腾。

2017年11月23日，腾讯公司与香港铁路有限公司正式签署合作协议，双方就微信支付和WeChat Pay HK（微信香港钱包）在港铁的移动支付业务展开合作。^[4]

腾讯什么时候上市？

答案：

2004年腾讯公司在香港联交所主板公开上市（股票代码00700），董事会主席兼首席执行官是马化腾。

通用英文问答数据集

English QA Datasets

- KBQA
 - **WebQuestions** (Stanford)
 - <https://nlp.stanford.edu/software/sempre/>
 - **SimpleQuestions** (Facebook)
 - <https://research.fb.com/downloads/babi/>
- TableQA
 - **WikiTableQuestions** (Stanford)
 - <https://nlp.stanford.edu/blog/wikitablequestions-a-complex-real-world-question-understanding-dataset/>
- PassageQA
 - **WikiQA** (Microsoft Research)
 - <https://www.microsoft.com/en-us/research/publication/wikiqa-a-challenge-dataset-for-open-domain-question-answering/>
- CommunityQA
 - **Question Pairs** (Quora)
 - <https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs>
 - **Task 3: Community QA** (SemEval)
 - <http://alt.qcri.org/semeval2017/index.php?id=tasks>
- Machine Reading Comprehension
 - **SQuAD** (Stanford)
 - <https://rajpurkar.github.io/SQuAD-explorer/>
 - **MS MARCO** (Microsoft)
 - <http://www.msmarco.org/>
 - **CNN/Daily Mail** (DeepMind)
 - <http://cs.nyu.edu/~kcho/DMQA/>

通用中文QA数据集

Chinese QA Datasets

- KBQA
 - **NLPCC2016-KBQA** (NLPCC & Microsoft Research Asia)
 - http://tcci.ccf.org.cn/conference/2016/pages/page05_evadata.html
 - **NLPCC2017-KBQA** (NLPCC & Microsoft Research Asia)
 - <http://tcci.ccf.org.cn/conference/2017/taskdata.php>
- PassageQA
 - **NLPCC2016-DBQA** (NLPCC & Microsoft Research Asia)
 - http://tcci.ccf.org.cn/conference/2016/pages/page05_evadata.html
 - **NLPCC2017-DBQA** (NLPCC & Microsoft Research Asia)
 - <http://tcci.ccf.org.cn/conference/2017/taskdata.php>

我们的成绩

DBQA Submissions	Rank (by MRR)
复旦大学	1
中科院自动化所 (CBrain Team)	2
哈尔滨工业大学 (机器智能与翻译实验室) [Secondary]	3
哈尔滨工业大学 (机器智能与翻译实验室) [Primary]	4
天津大学	5
黑龙江工程学院 [MART]	6
哈尔滨工业大学 (ITNLP Group)	7
黑龙江工程学院 [CA]	8
哈尔滨工业大学 (HIT-SCIR)	9
黑龙江工程学院 [LM]	10
北京航空航天大学	11
山西大学	12
大连理工大学	13
同济大学	14
东北大学	15
武汉科技大学 (NLP@WUST)	16
Harbin ShenZhi Technology Co., Ltd.	17
浙江大学	18

DBQA Submissions	Rank (by MRR)
复旦大学	1
北京邮电大学	2
同济大学	3
DeepIntell-1	4
天津大学	5
DeepIntell-2	6
DeepIntell-3	7
DeepIntell-4	8
华中师范大学	9
北京大学	10
浙江大学	11
网龙网络有限公司	12
北京航空航天大学	13
国防科技大学	14
华东师范大学	15
北京联合大学	16
北京大学	17
大连理工大学-1	18
大连理工大学-2	19
大连理工大学-3	20
重庆理工大学	21

问答的核心—match

- 两个句子对的**匹配**

- 问题-问题匹配
- 问题-答案匹配

Q1:员工在外地就医的单据，能否通过公司商业保险报销？

Q2:您好，我在外地看病，商保怎么报。

Q:员工在外地就医的单据，能否通过公司商业保险报销？

A:您好，可以的。普通员工的医疗保险保障区域为中国大陆境内，不包含港、澳、台地区。只要所产生的费用在社保范围内,医院是社保范围内医院,是可以通过商业保险进行报销的。

问答匹配的算法分类

- 无监督方法

- 文本距离度量
- 检索式指标 (TFIDF, IDF)
- 基于加权词向量的方法
- 基于依存句法树的匹配
- ...

- 有监督方法

- 传统机器学习 (TFIDF+朴素贝叶斯/SVM)
- CNN
- RNN/LSTM/GRU/BiLSTM
- CNN+RNN
- 其他神经网络方法
- ...

更多的人工标注数据，更少人工规则

问答匹配的算法分类

- 无监督方法

- 文本距离度量
- 检索式指标 (TFIDF, IDF)
- 基于加权词向量的方法
- 基于依存句法树的匹配
- ...

- 有监督方法

- 传统机器学习 (TFIDF+朴素贝叶斯/SVM)
- CNN
- RNN/LSTM/GRU/BiLSTM
- CNN+RNN
- 其他神经网络方法
- ...

更多的人工标注数据，更少人工规则

标准问题：怎么提取公积金？

相似问题：我想提取公积金，请问怎么操作？

标注数据类型1：知识库编辑人员

用户实时问题：我想提取公积金，请问怎么操作？

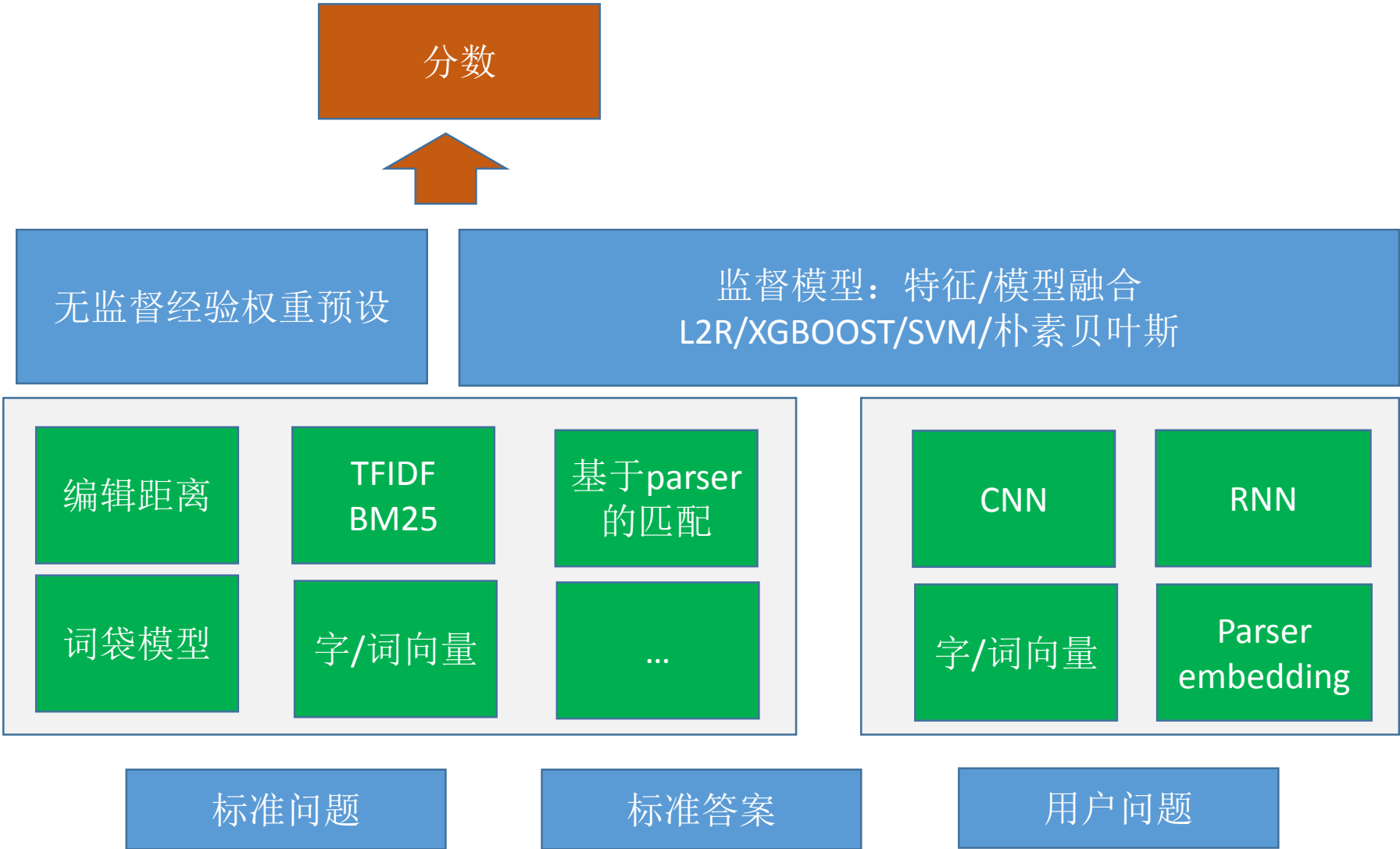
推荐标准问题：怎么提取公积金？

推荐标准答案：公积金提取balabala...，请访问XXX.com

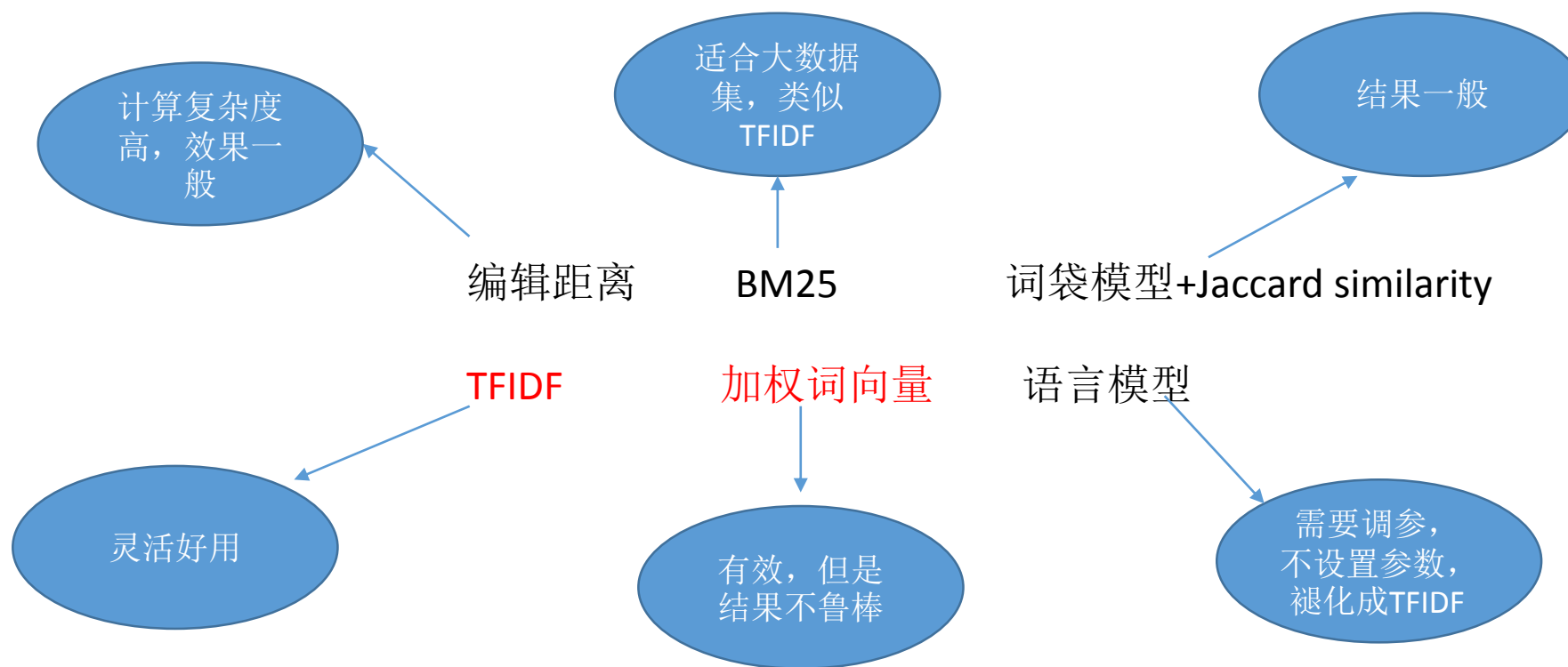
用户打分：非常满意

标注数据类型2：终端用户使用日志，隐式或者显示反馈

匹配算法框架



无监督方法



高度定制化的TFIDF

- IDF定制化

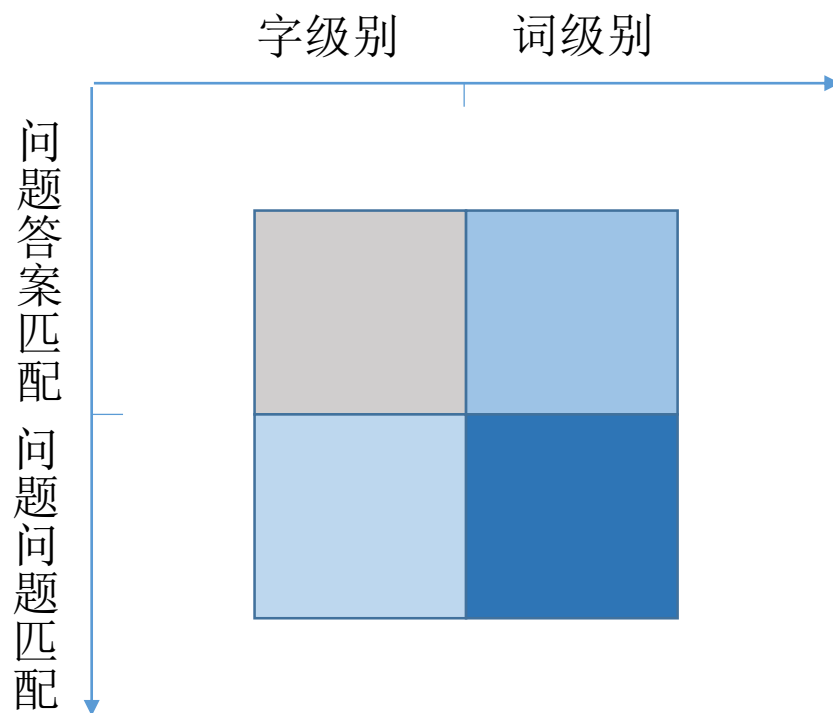
- 更好平滑的IDF，更适应于小语料
- 抽象成内部service，支持增量增删改，并提供给其他算法模块，如词向量模块儿

- TF定制化

- 伯努利假设：假设词语出现次数服从伯努利分布 【出现还是没出现】
- 多项式假设：假设词语出现次数服从多项式分布 【越多越好】
- 高斯假设：假设词语出现次数服从高斯分布 【出现次数不要多也不好少】

多级别的相关性

- 基于字词、与标准问题/答案四个级别的相关性



监督模型

- 传统机器学习（TFIDF+SVM）
- 词向量一览
- CNN/RNN等神经网络模型

词汇鸿沟

- S1:商保通过什么方式报销？
 - 商保 通过 什么 方式 报销
- S2:怎么报商业保险？
 - 怎么 报 商业 保险

词汇鸿沟

- - 商保 通过 什么 方式 报销
- - 怎么 报 商业 保险

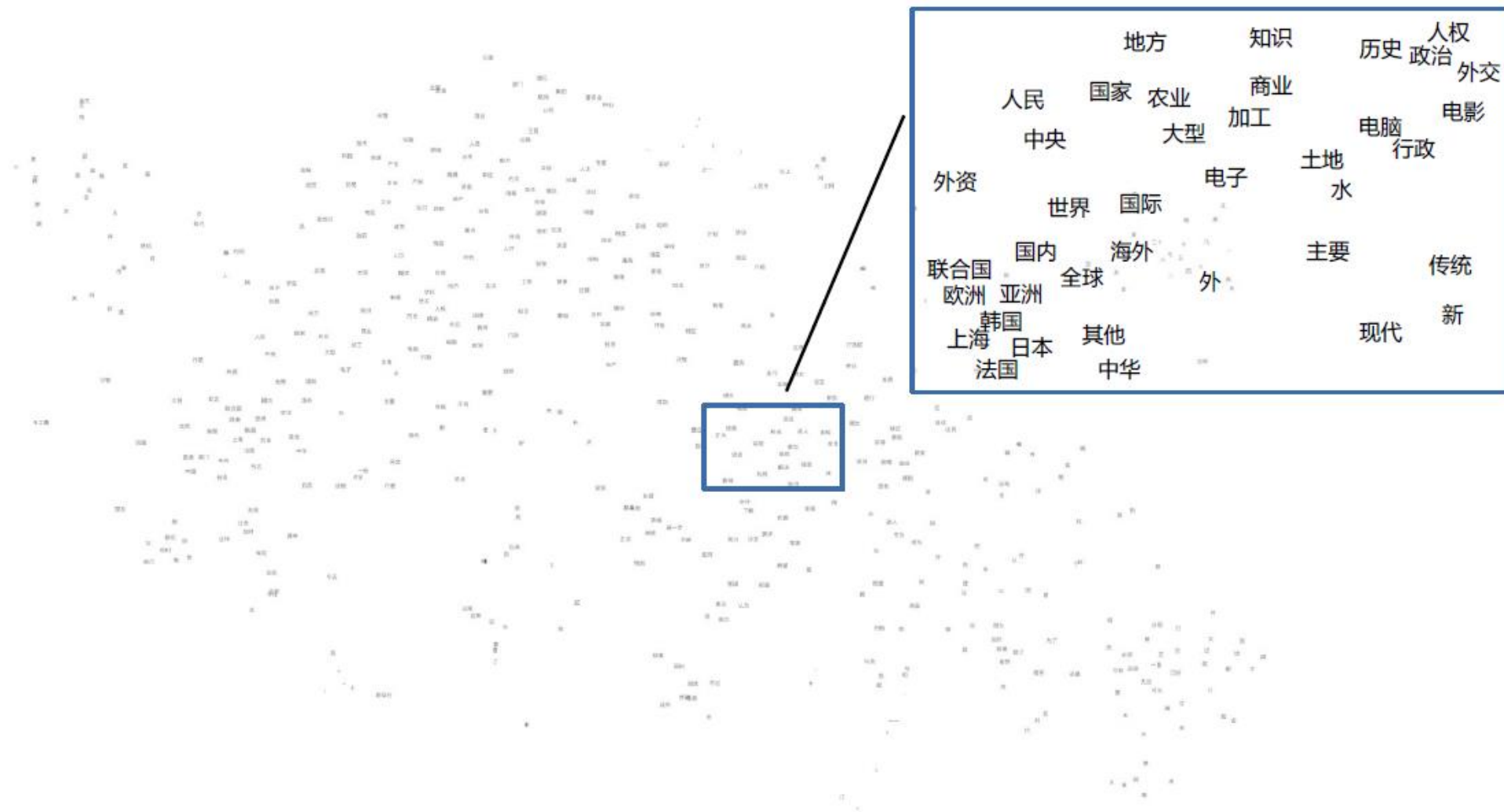
词	S1	S2
商保	1	0
通过	1	0
什么	1	0
方式	1	0
报销	1	0
怎么	0	1
报	0	1
商业	0	1
保险	0	1
...	0	0

相似性为0

词向量

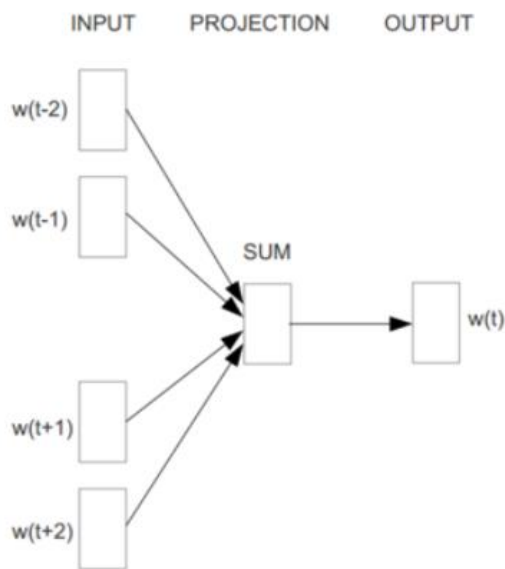
- One-hot representation
 - 商保 [1,0,0,0,0,0,0,0,0]
 - 保险 [0,0,0,0,0,0,0,1,0]
- Distributed representation
 - 商保 [0.792, -0.177, -0.107, 0.109, -0.542]
 - 保险 [0.856, -0.523, 0, 0.2, -0.2]

词向量空间

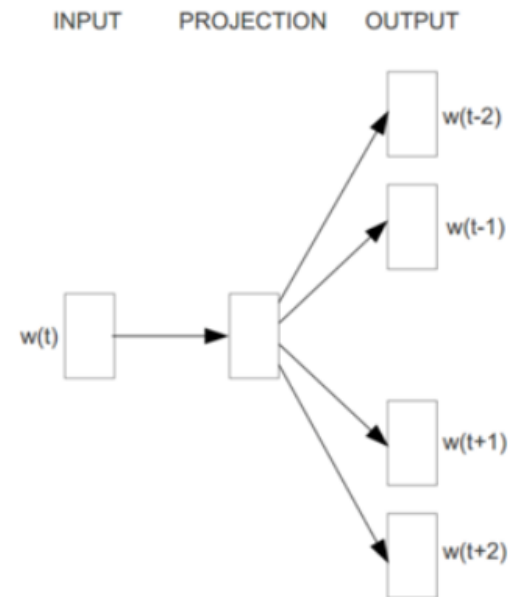


分布式

- 一个词的语义与它的旁边的词相关 [1]



Cbow: 旁边词预测中心词



Skip-gram: 中心词预测旁边词

这样无需标注的自然成句的文本语料成为训练词向量的廉价原料

[1] 跟什么人在一起，自己就是什么水平的人。跟平台部的人一起合作，我就知道我们也是很牛逼的

算命先生-词向量

```
ycc@yc: ~/Documents/NLP_lib/word2Vec
>>> indexes = model.cosine(u'清华大学')
>>> for index in indexes[0]:
...     print (model.vocab[index])
...
北京大学
武汉大学
上海交通大学
南京大学
上海交大
香港浸会大学
台湾大学
中国科技大学
浙江大学
山东大学
>>>
```

同时在找到一些相关词的同时也给模型带来了很多噪音。

【止痛】 的相关词可能有，既相关也不相关，跟算命和星座一样^[1]

【医生】 【消炎】 【止痛药】 【坚强】

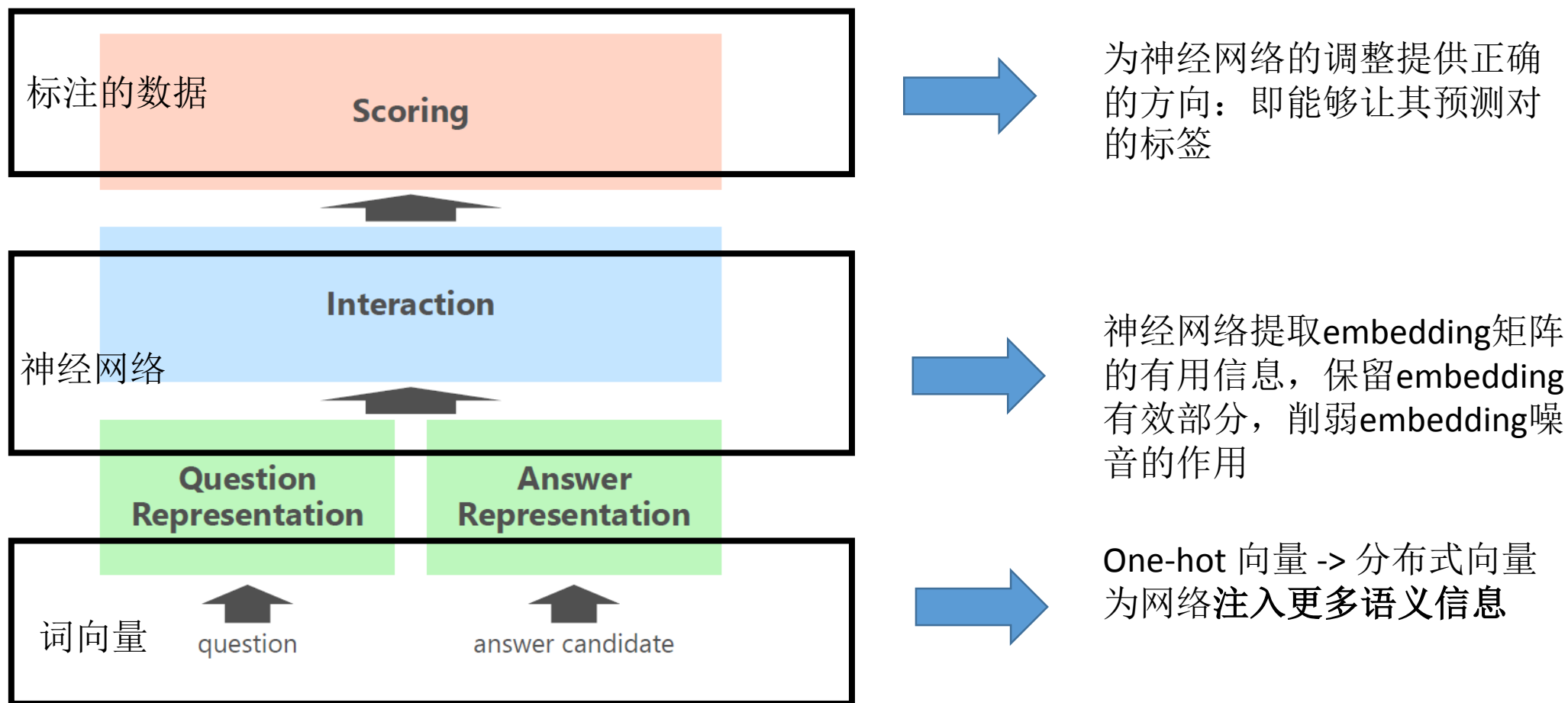
[1] 说法来自蒙特利尔大学信息检索教授聂建云私下学术交流，聂老师跟bengio在一个系

词向量的无监督用法

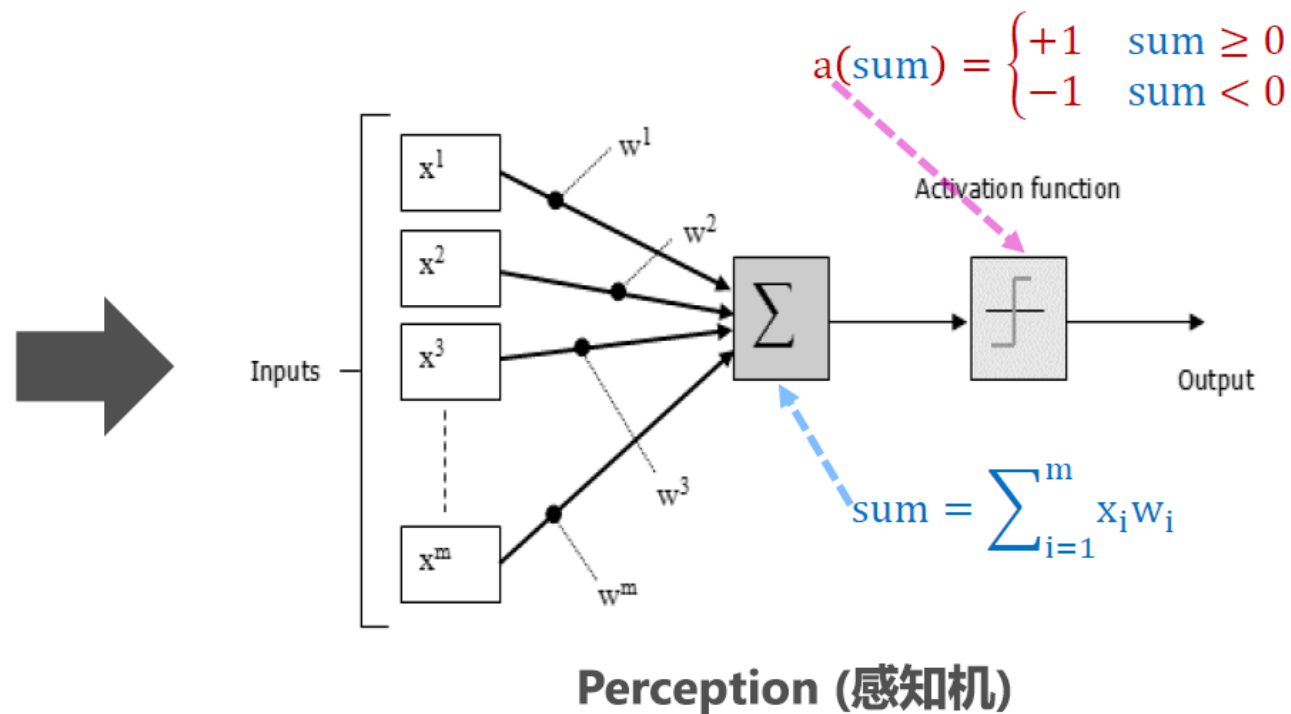
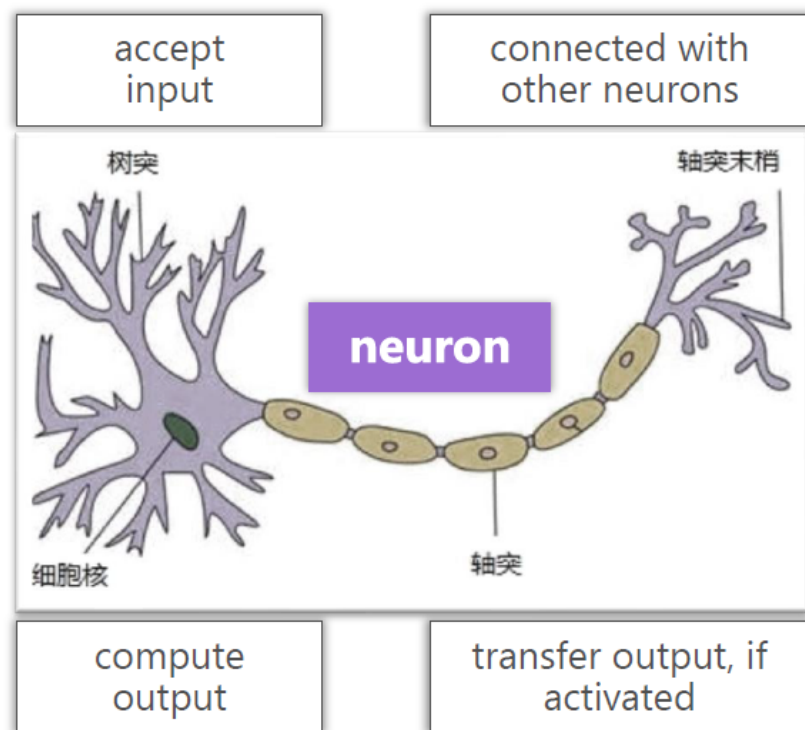
词	S1	S2
商保	1	0
通过	1	0
什么	1	0
方式	1	0
报	1	0
怎么	0	1
报	0	1
商业	0	1
保险	0	1
...	0	0

- 商保 通过 什么 方式 报销
 - $0.5 * V(\text{“商保”}) + 0.3 * V(\text{“报销”}) + 0.2 * V(\text{“通过”})$
- 怎么 报 商业 保险
 - $0.4 * V(\text{“保险”}) + 0.4 * V(\text{“商业”}) + 0.2 * V(\text{“报”})$

更准的算命先生：词向量+神经网络+标注的数据



神经网络1: MLP多层感知机



由万能逼近定理，足够深和宽的神经网络可以无限逼近任何函数

神经网络2: CNN 卷积神经网络

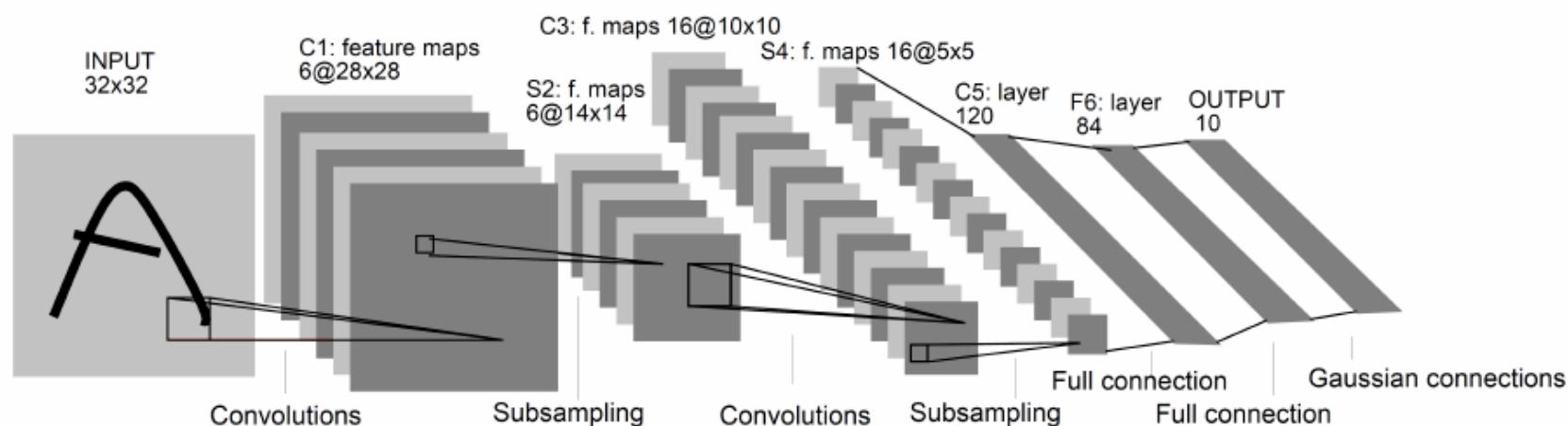
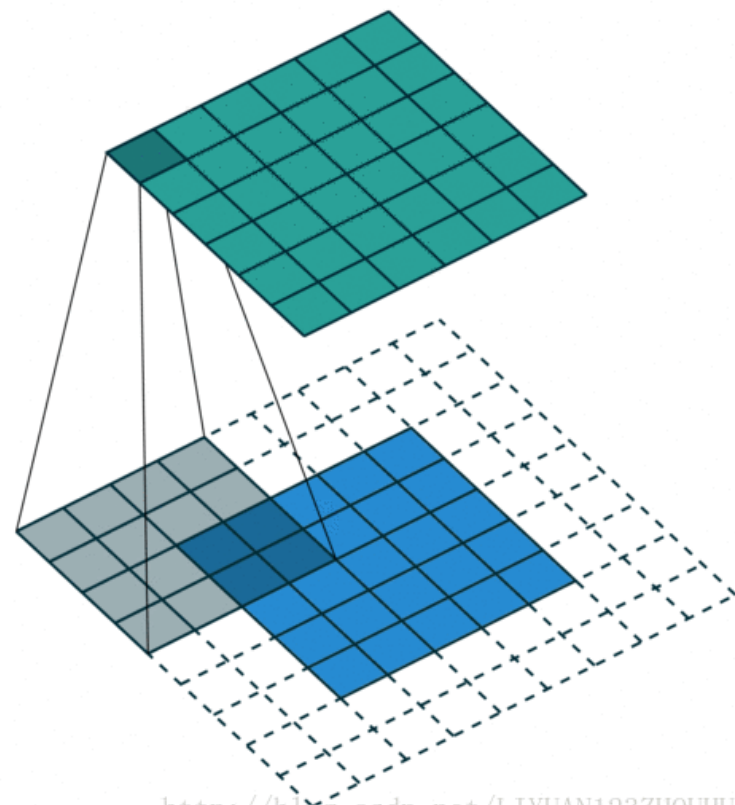


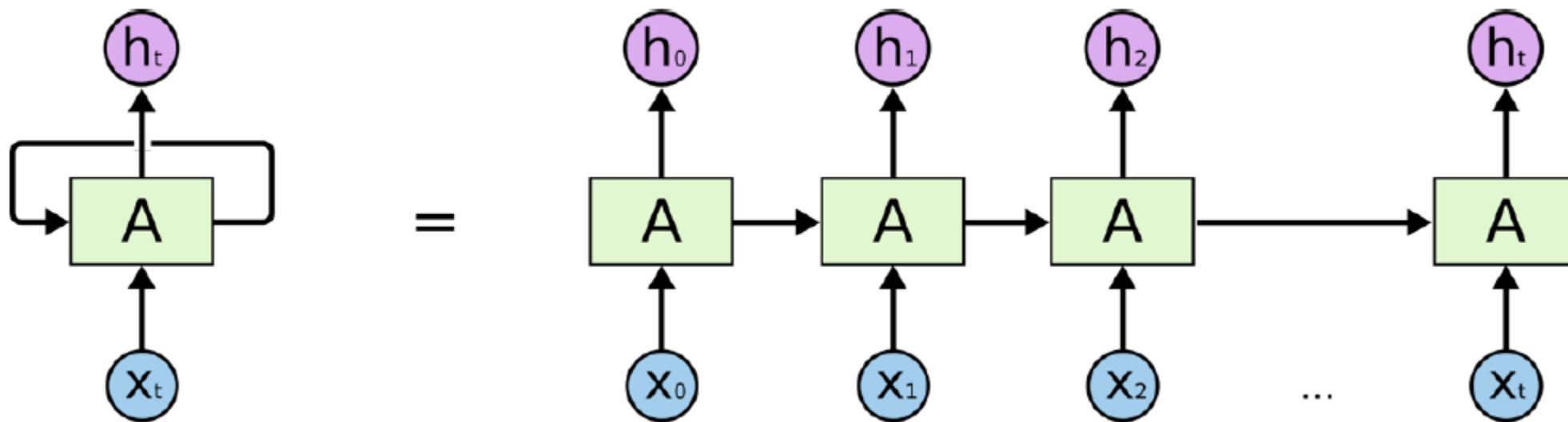
图: LeNet-5 网络结构⁵



<http://blog.csdn.net/LIYUAN123ZHOUHUI>

局部连接、权值共享使得CNN对局部特征更敏感，对建模图像数据更有优势

神经网络3: RNN^[1] 建模序列数据



循环神经网络RNN对序列数据建模更有优势

[1] RNN一般指循环神经网络，ReNN指递归神经网络

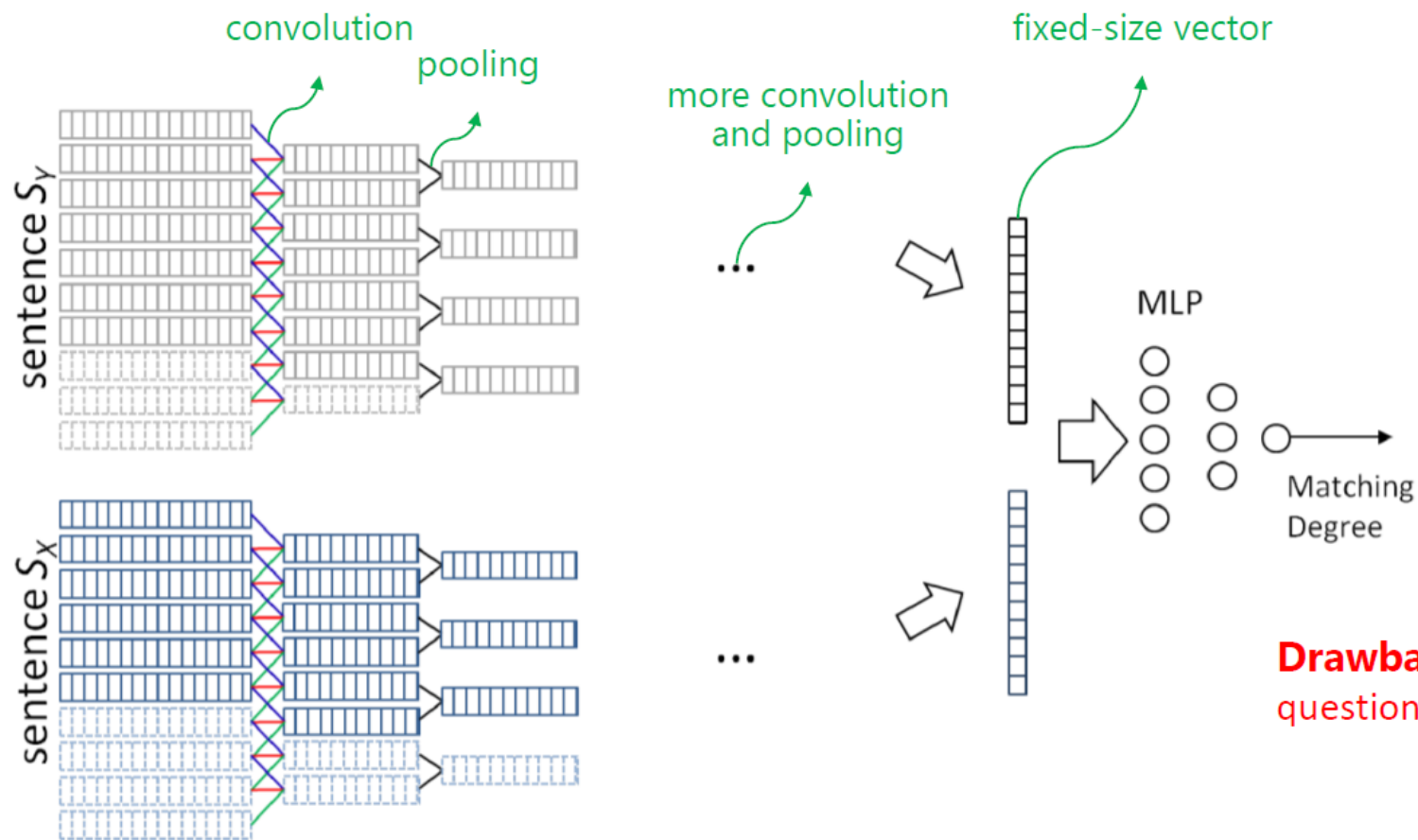
RNN vs CNN

- RNN
 - 序列结构
 - 强调高阶关系
 - 位置跳跃的依赖
- 速度
 - 更慢，串行
 - 方便定长，通过attention

- CNN
 - 两个句子关系
 - N-gram匹配更重要的match 场景
 - 局部依赖关系
- 速度
 - 可以并行，更灵活
 - 输出不定长，跟文本长度有关

CNN : 一维匹配模型

(Hu et al., 2015)

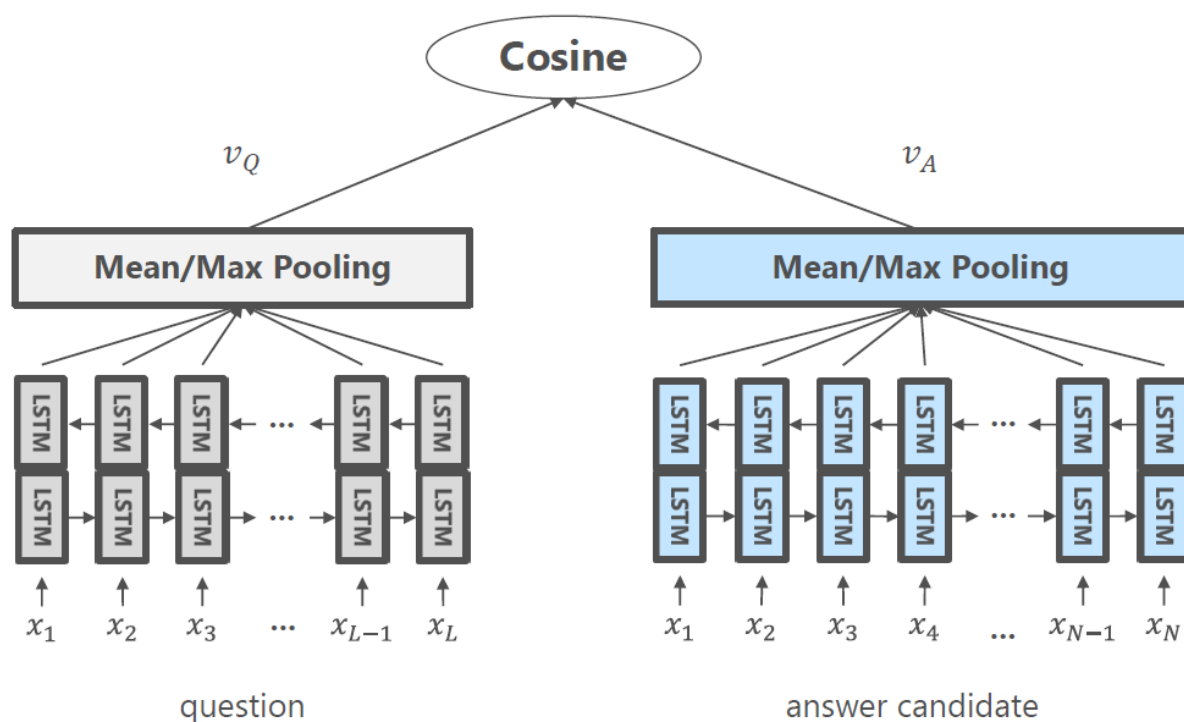


Drawback: defer the interaction between question and answer in the final MLP

RNN 一维匹配

RNN

(Wang and Nyberg, 2015; Tan et al., 2015; Hsu et al., 2016)



$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

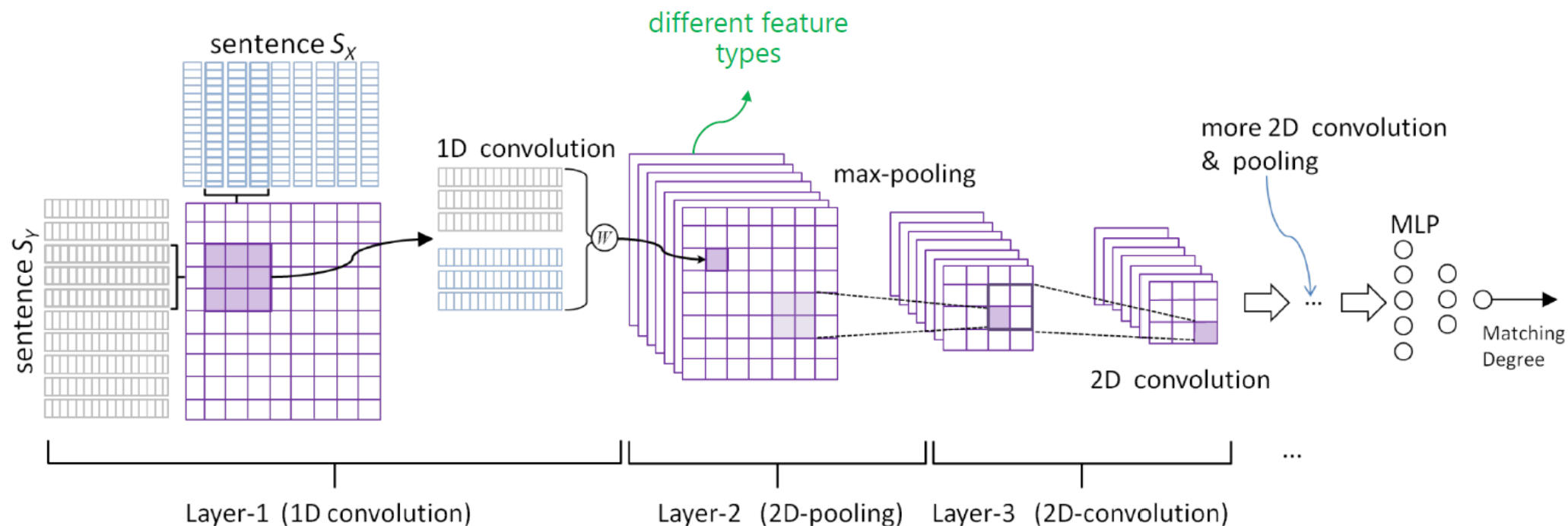
$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1}$$

$$h_t = o_t * \tanh(C_t)$$

$$L = \max\{0, 1 - \text{cosine}(q, a_+) + \text{cosine}(q, a_-)\}$$

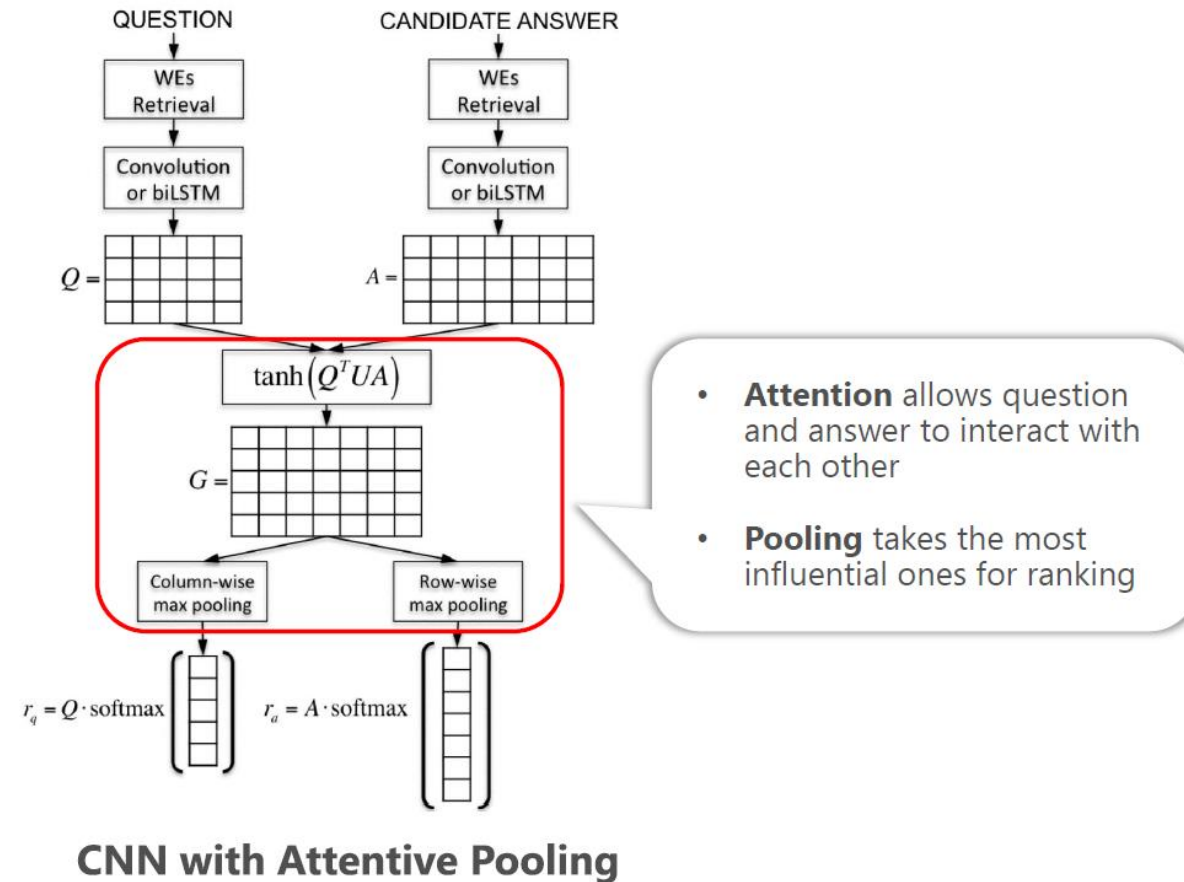
CNN：二维匹配模型

(Hu et al., 2014)



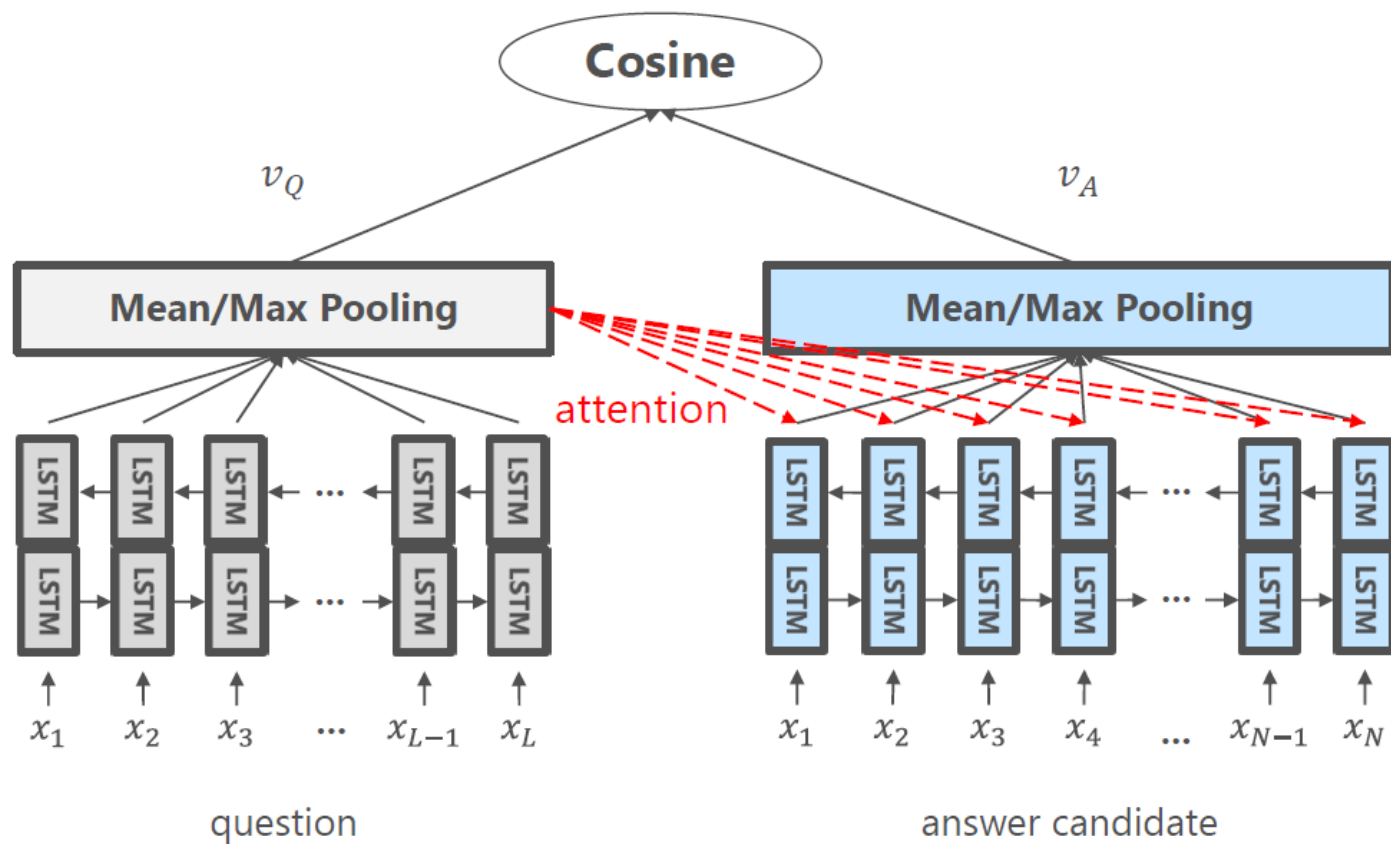
Advantage: let question and answer interact in early stage

CNN+Attentive Pooling



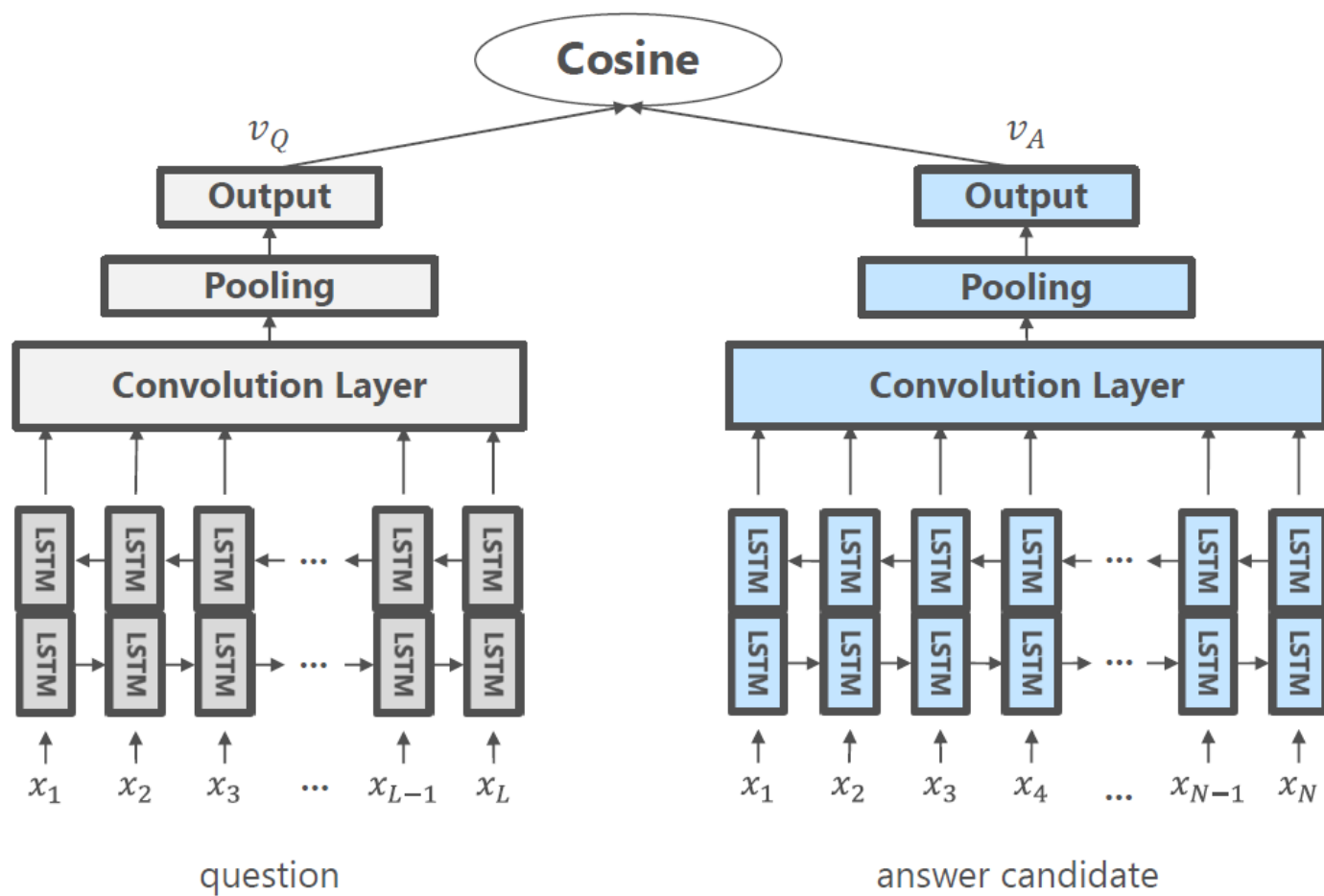
RNN + Attention

(Miao et al., 2016; Tan et al., 2015; Hsu et al., 2016)



RNN+CNN

(Tan et al., 2015)



CNN+RNN

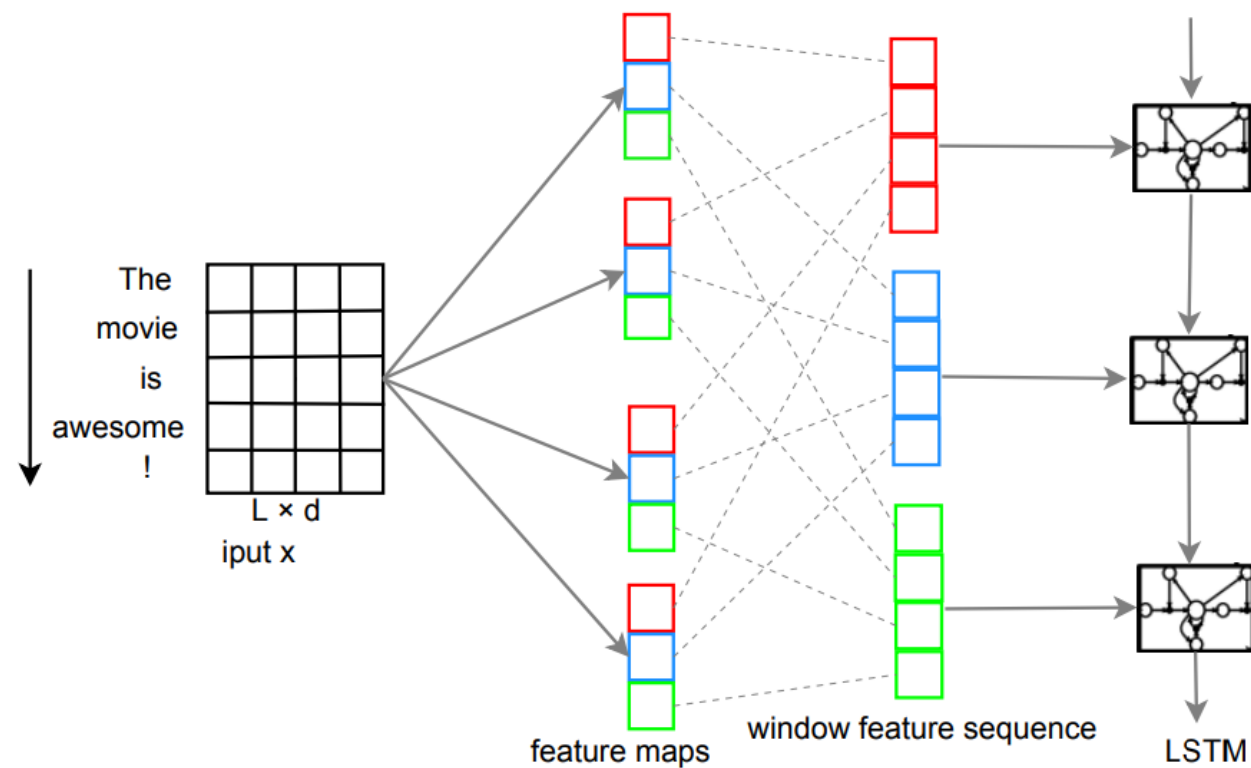
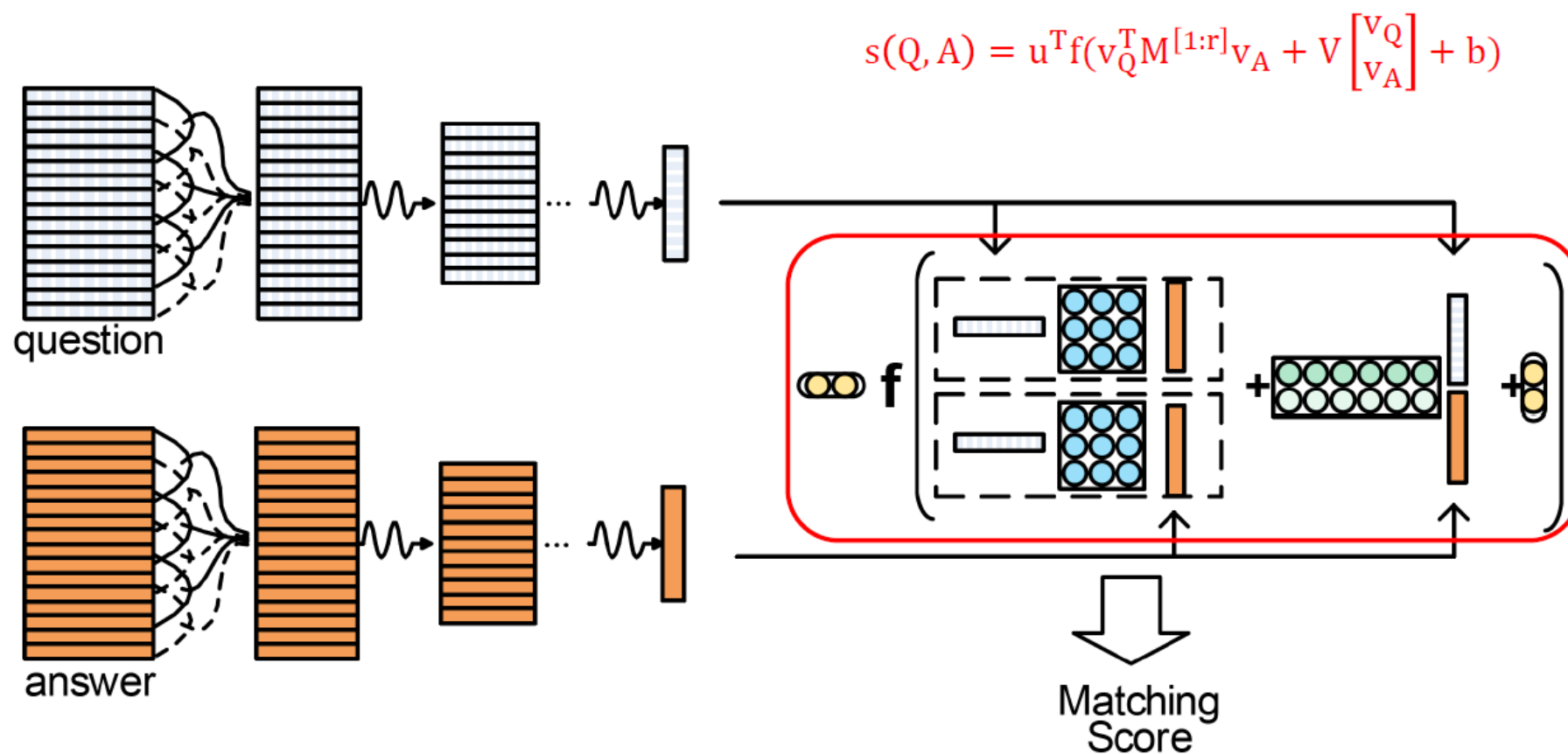


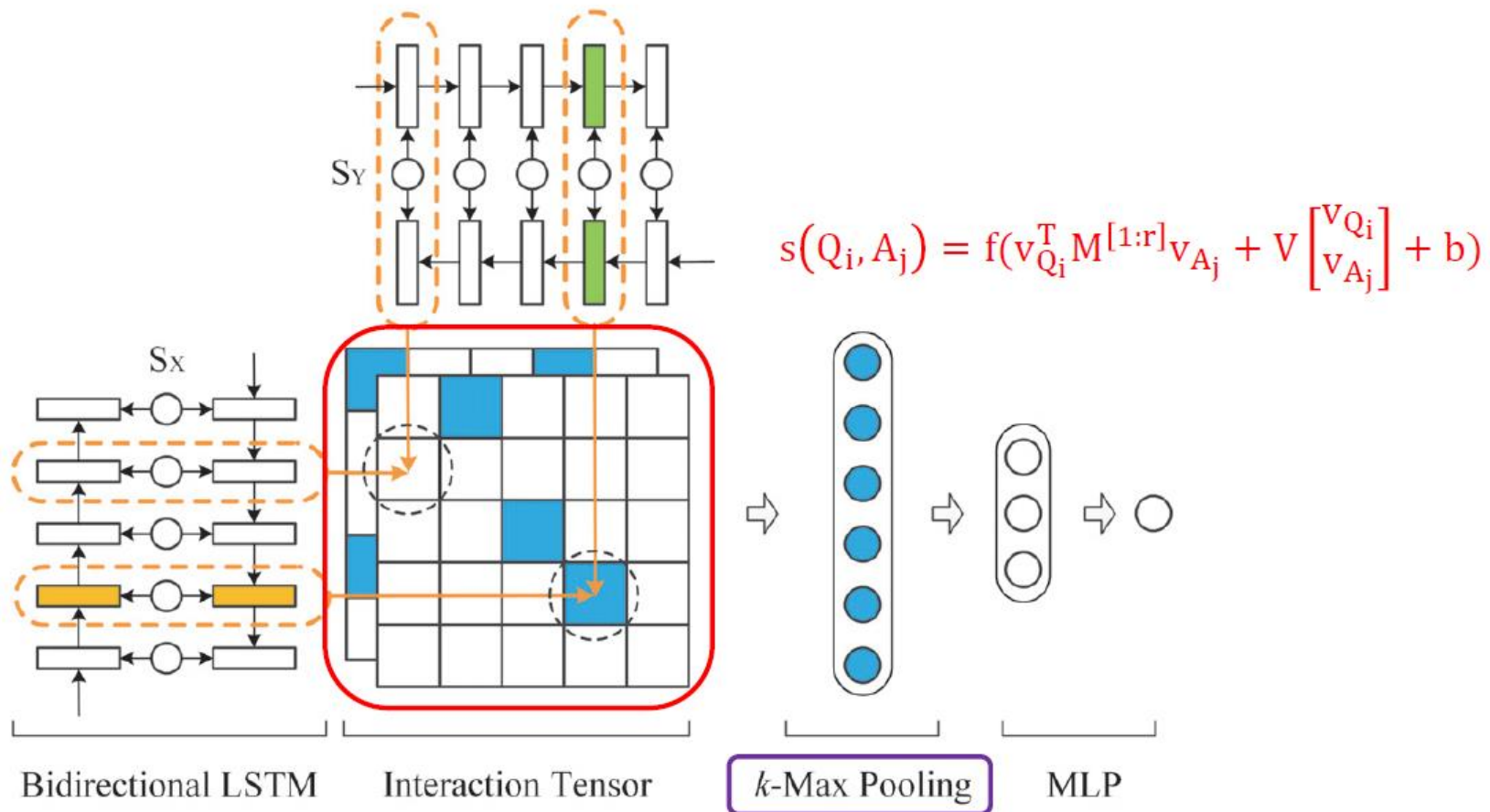
Figure 1: The architecture of C-LSTM for sentence modeling.

CNN+Tensor



RNN with Tensor

(Wan et al., 2016)



未来的方向

- 数据更多，层次更深
- Embedding层融入尽可能多的特征和知识

Evaluation Results

- Dataset
 - WikiQA
- Metric
 - MAP, MRR, P@1

	WikiQA		
	Train	Dev	Test
# Question	2,118	296	633
# Sentence	20,360	2,733	6,165
# Answer Sentence	1,040	140	293
Ave. S / Q	9.61	9.23	9.74
# Question (Have Answer)	873	126	243
	41.22%	42.57%	38.39%
Ave. Q Length	7.16	7.23	7.26
Ave. S Length	25.29	24.59	24.95

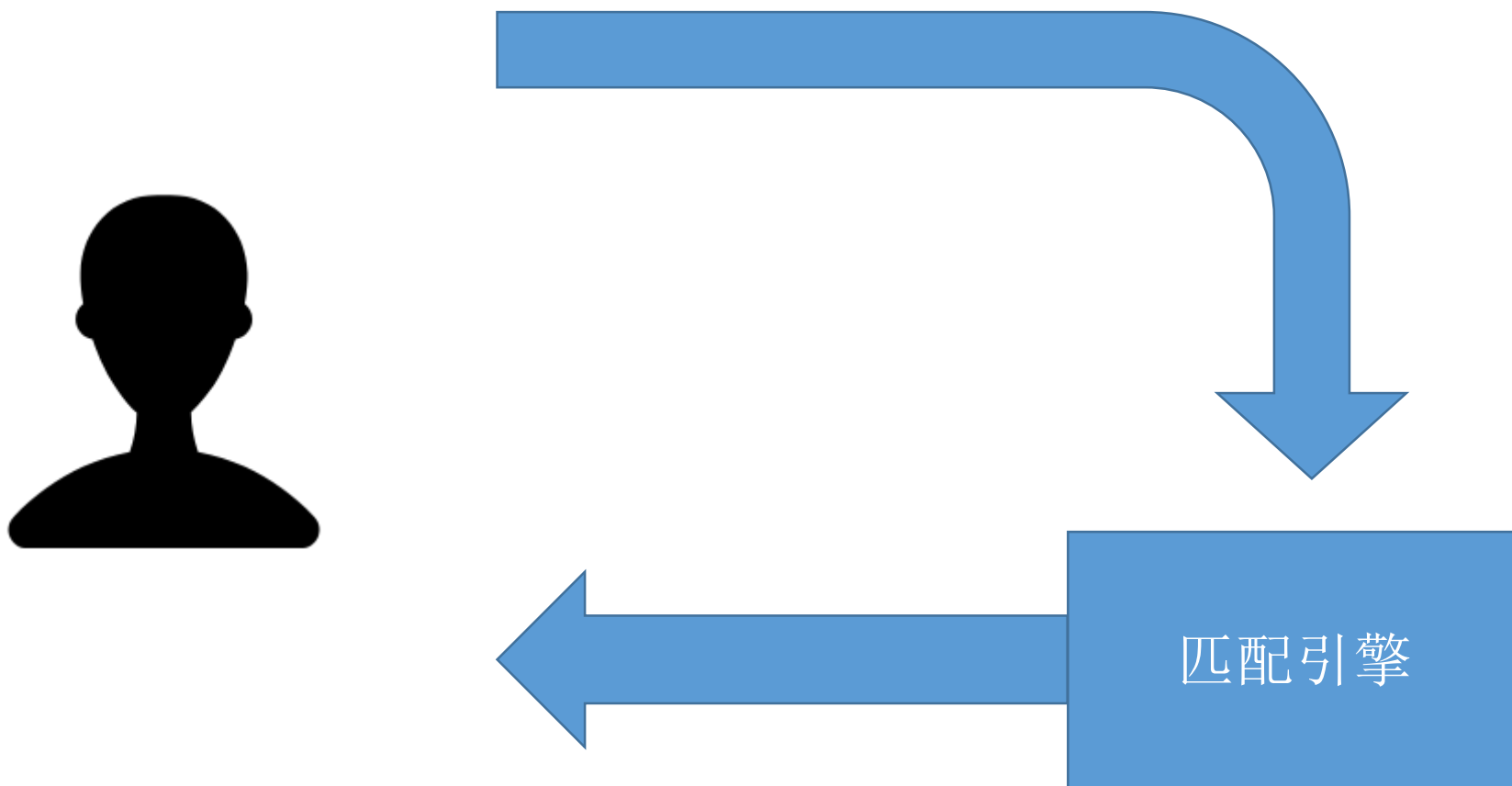
Type	Method		MAP	MRR
CNN	bi-gram CNN	Yang et al., 2015	0.6520	0.6652
	CNNr	Severyn and Moschitti., 2016	0.6951	0.7107
	Attentive pooling CNN	Santos et al., 2016	0.6886	0.6957
	Attention-based CNN	Yin et al., 2016	0.6914	0.7127
	DocChat	Yan et al., 2016	0.7008	0.7222
	CNN + Features	Tymoshenko et al., 2016	0.7417	0.7588
RNN	LSTM+Attention	Miao et al., 2016	0.6855	0.7041
	NASM	Miao et al., 2016	0.6886	0.7069
	IARNN-Occam (context)	Wang et al., 2016 (a)	0.7341	0.7418
CNN+RNN	conv-RNN	Wang et al., 2017	0.7427	0.7504
	RNN+CNN with Mult Interaction	Wang and Jiang., 2016	0.7433	0.7545
Other	Key-Value Memory Network	Miller et al., 2016	0.7069	0.7265
	Pairwise Rank	Rao et al., 2016	0.7010	0.7180
	L.D.C	Wang et al., 2016 (b)	0.7058	0.7226
	CubeCNN (Pairwise Rank)	He and Lin., 2016	0.7090	0.7234

深度学习搭积木， So easy !

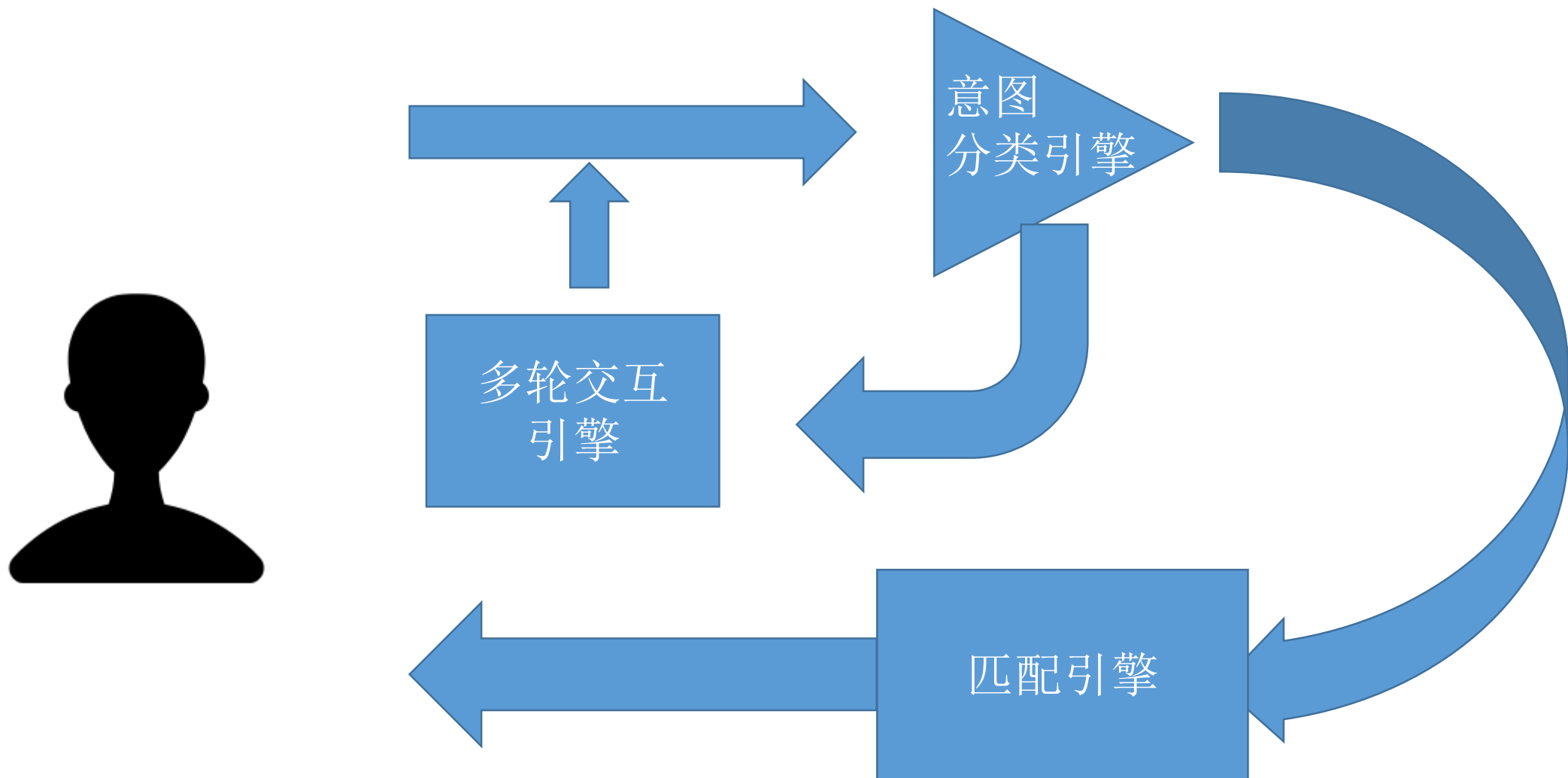
框架	公司	特点
Tensorflow	Google	迭代快，流行
Torch/PyTorch	Facebook	方便调试
Theano	蒙特利尔大学 bengio等	早，编译慢
Caffe		图像领域更好用

团队同时熟练支持 **Tensorflow/PyTorch/Theano/Keras** 框架开发！ ！ ！

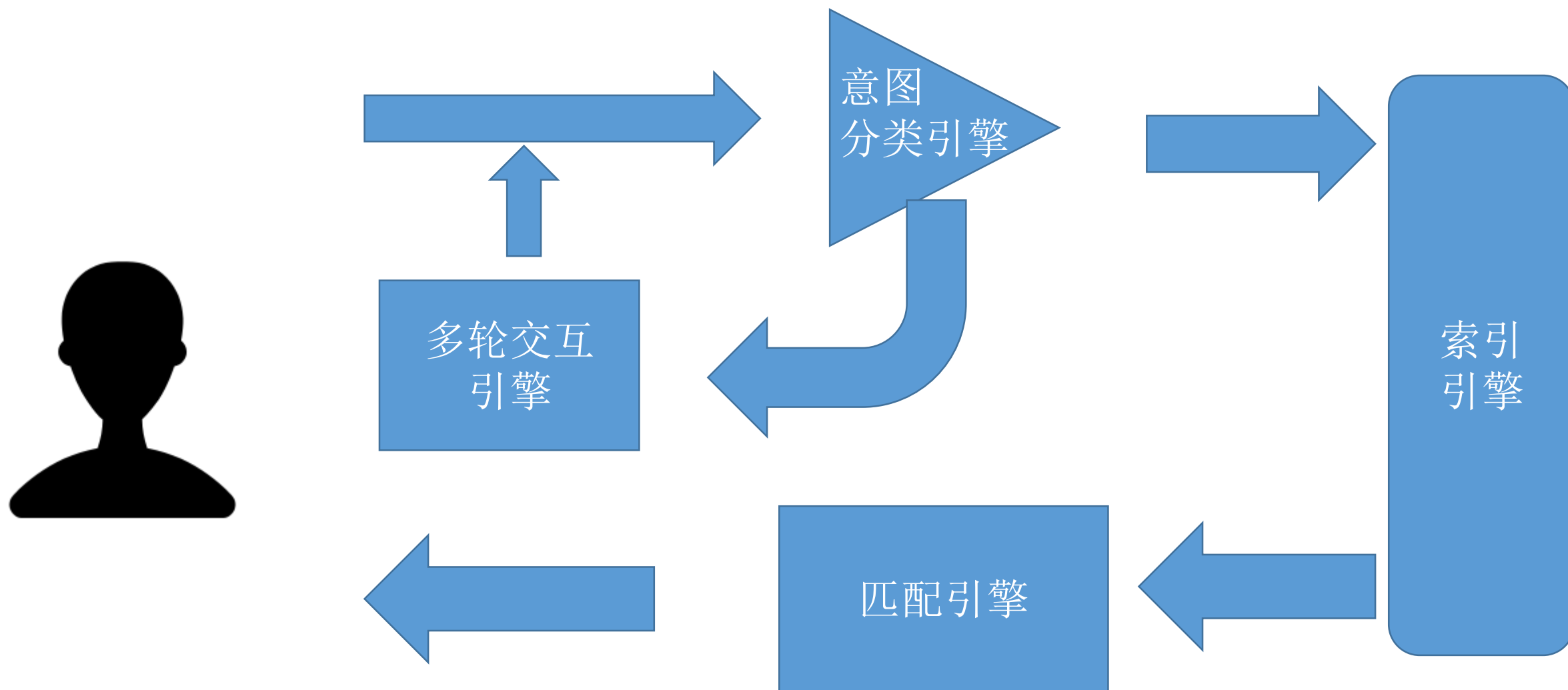
智能客服系统-V0



智能客服系统-V1

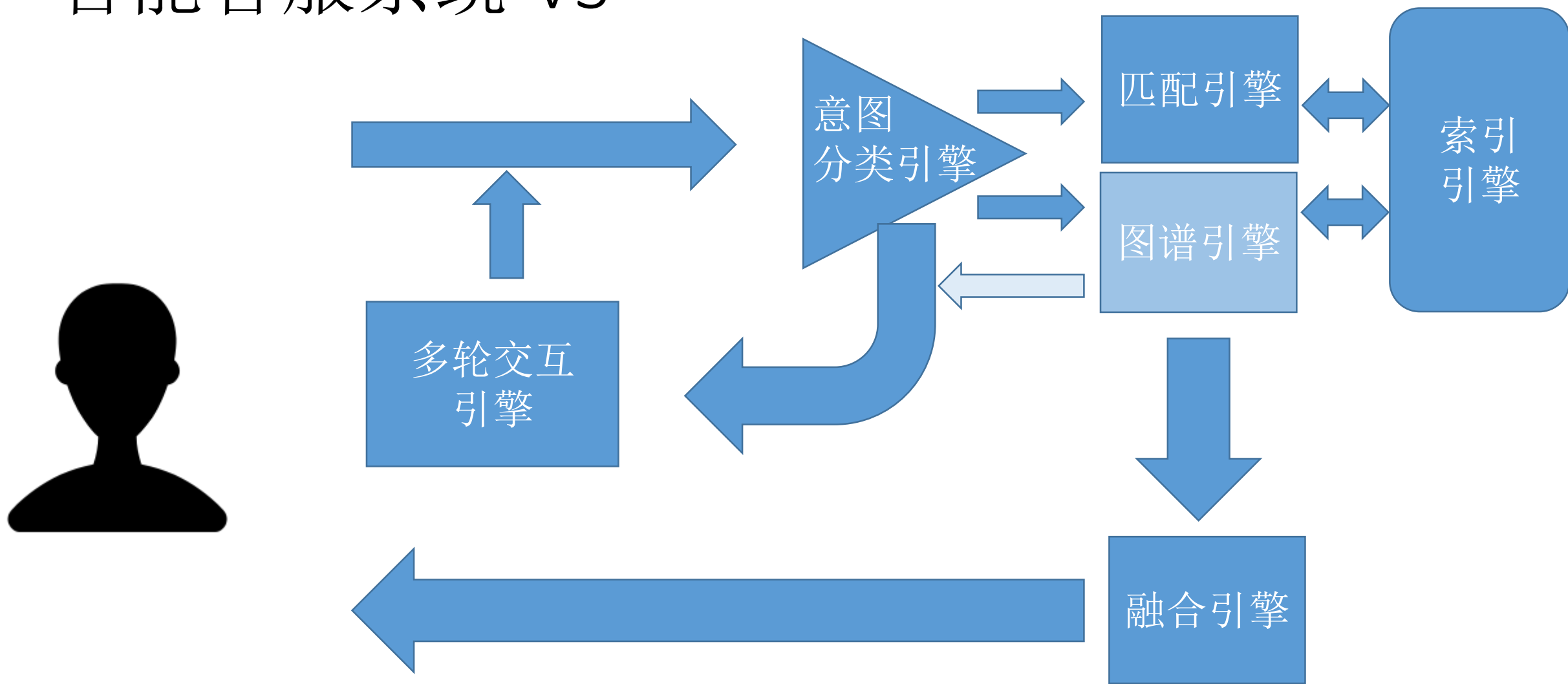


智能客服系统-V2



后台检索算法耗时在1-2ms内，即时增删改知识库和索引。融入稀疏矩阵、索引、cache、lazy load等多种机制

智能客服系统-V3



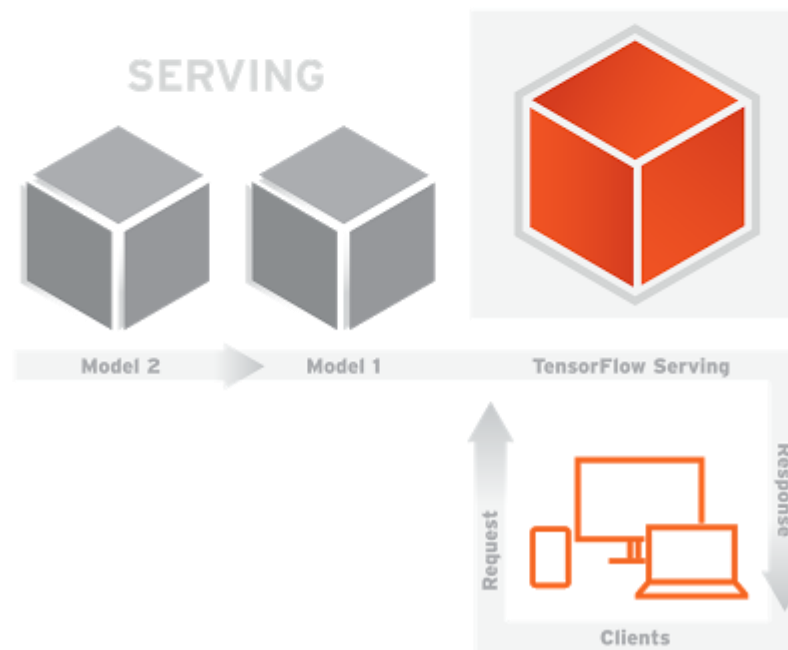
性能问题

- 基于索引初筛，再算法精排
- 增量增删改知识库
- FAQ的统计特征cache存起来，索引检索出来再实时计算
- 稀疏矩阵减少计算量

从6S 到 0.001S，检索时间不随FAQ线性增加，实时增删改

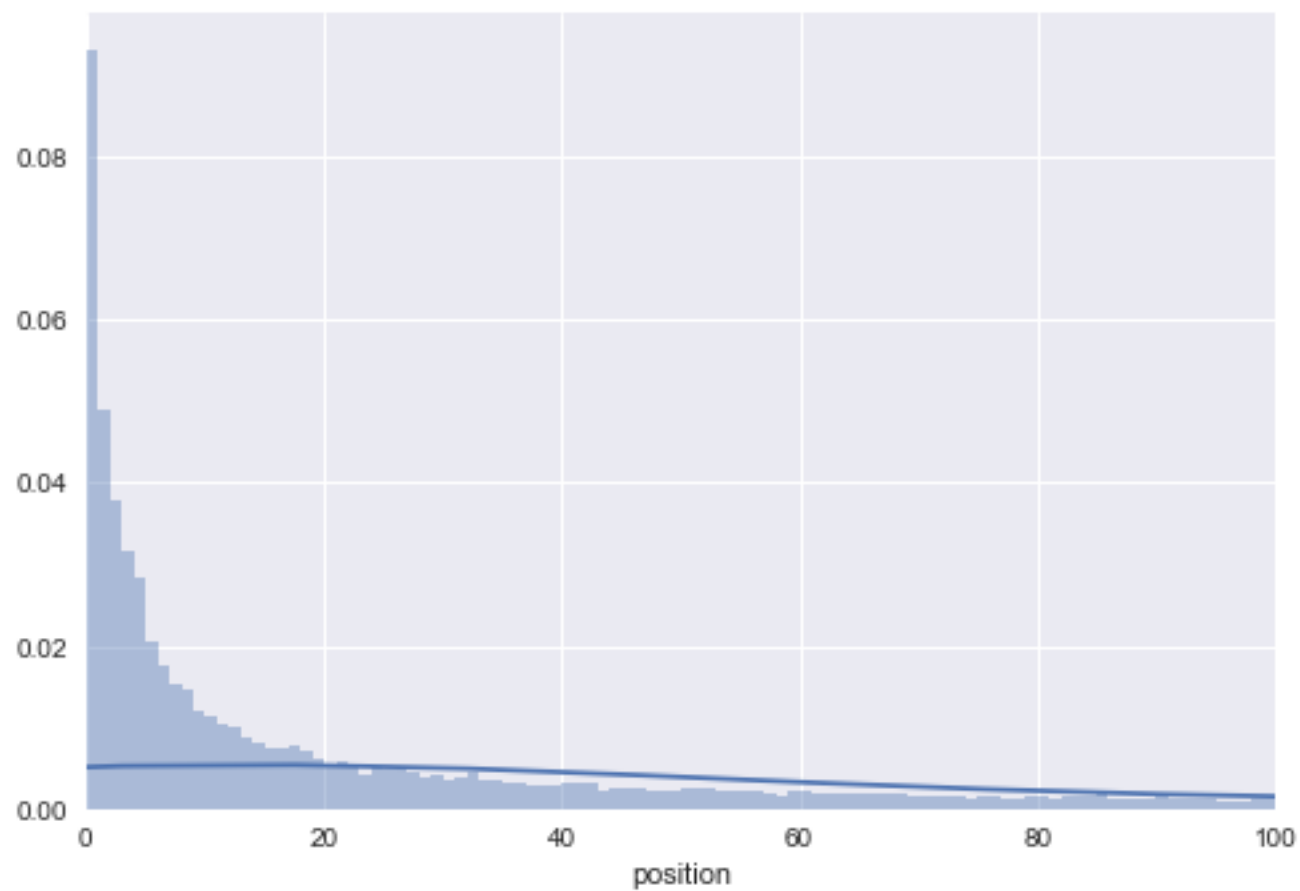
神经网络的部署

- Tensorflow Serving
 - 线上线下载耦
 - 热切换
 - RPC调用，跨平台，语言无关
 - C/C++编写 速度更快
 - GPU/CPU资源的高效利用



阈值

- 卡高分
- 卡低分
- 卡差值

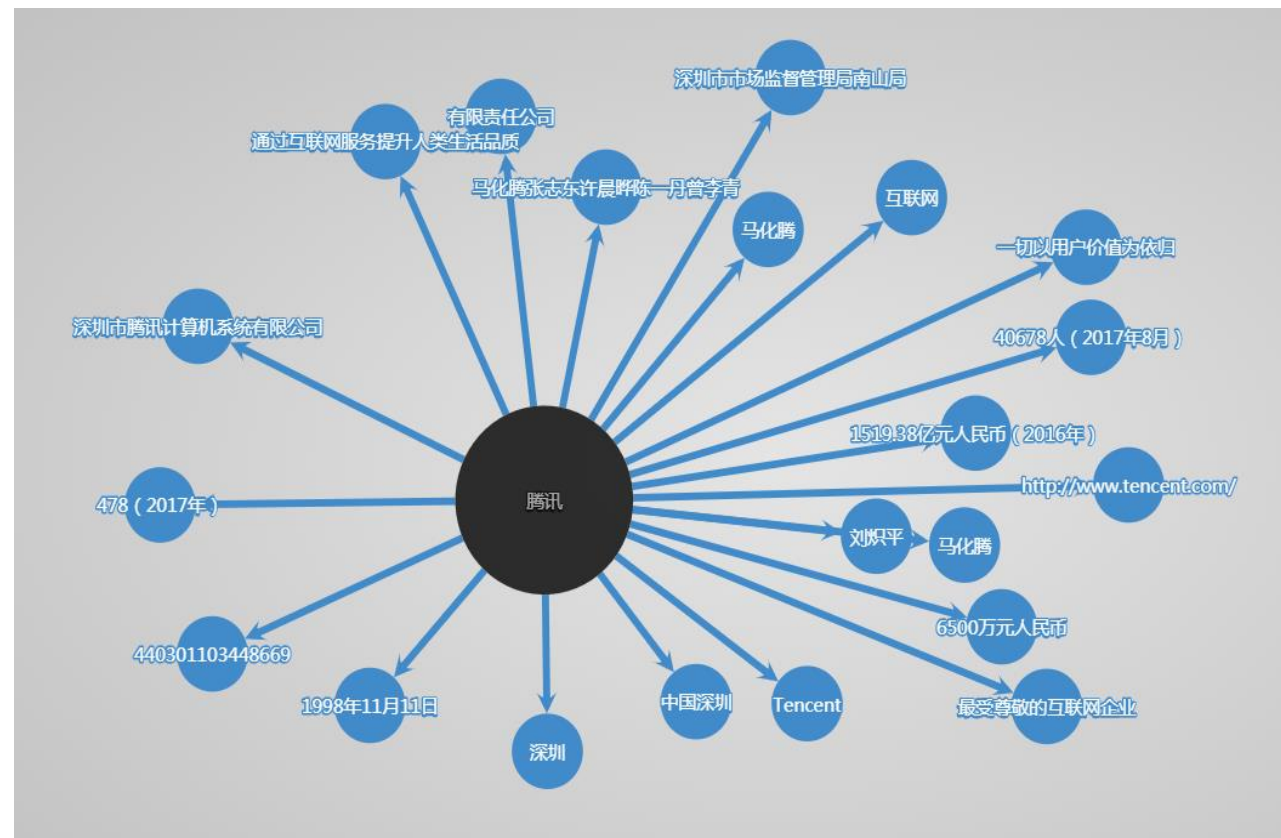


其他外围开发

- 哈工大同义词林
 - 基于规则的通用同义词匹配
- 外围开发：
 - 基于Trie树和概率语言模型拼音纠错器

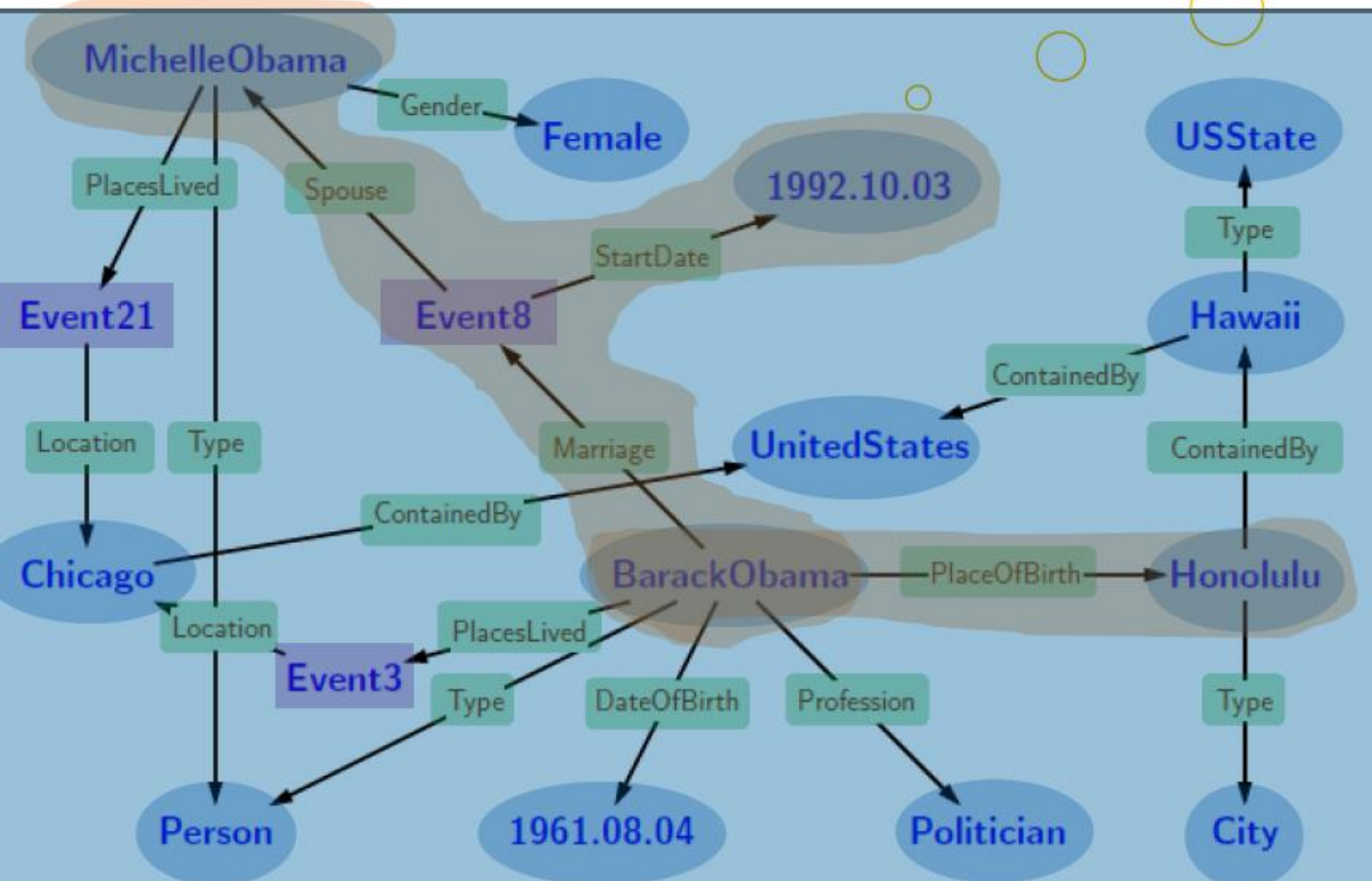
知识图谱

- 特定领域知识库
 - 用户自己定义的知识库
 - 对检索的效果提升很显著
 - 获取成本很大
- 通用知识图谱
 - Wiki中文、百度百科、互动百科
 - 复旦CN-Probase等其他知识库
 - 对检索效果提升有限
 - 一般开放获取，也可自己
 - ...



Knowledge Base (KB)

- Curated database with well-defined schema



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

知识图谱能回答的问题举例

- 信息类

- 螺纹钢今日的最重要的10条资讯?
- 显示1个月内 中钢协发布资讯?
- 与焦炭相关的公司有哪些?
- 深天马A的主营业务产品是什么?

- 按重要性
- 按来源
- 公司节点
- 公司基本

- 数据类

- 铁矿石的价格/产量/库存/进口量?
- 计算焦炭价格与M2之间的相关性吗?
- 10月螺纹钢情绪指数的最大值出现在哪一天?
- 列出与冷轧板卷产量相关性最高的10个数据指标

- 具体数据
- 数据+简单计算
- 数据+简单分析
- 数据+简单处理

领域知识库 来自互联网爬取数据，经过人工整理！！

实体-意图 方式

- 简单将FAQ结构化，多两个字段，用模板构建了知识库
- FAQ与知识库三元组一一对应，检索的是某一条三元组。
 - 实际知识图谱会由一个和多个条知识三元组组合推理得到
- 基本不是图谱，三元组的末节点没有出度
- 实体意图拆分不需要实体抽取，必须要求字段匹配

My paper

[3] **Wang B** et al. A Chinese Question Answering Approach Integrating Count-based and Embedding-based Features[C]NLPCC 2016. [EI]

[6]Wang J, Yu L, Zhang W, Gong Y, Xu Y, **Wang B** et al. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models[C]. **SIGIR** 2017 long paper. [CCF-A] (three strong-accept reviews and SIGIR **best paper** Honorable Mention)

[9] Su Z, **Wang B**, et.al Enhanced Embedding based Attentive Pooling Network for Answer Selection. NLPCC 2017

[10]Zhang P, Niu J, Su Z, **Wang B**, et.al A End-to-end quantum language model in question answering, **AAAI** 2018

Count-based and embedding-based

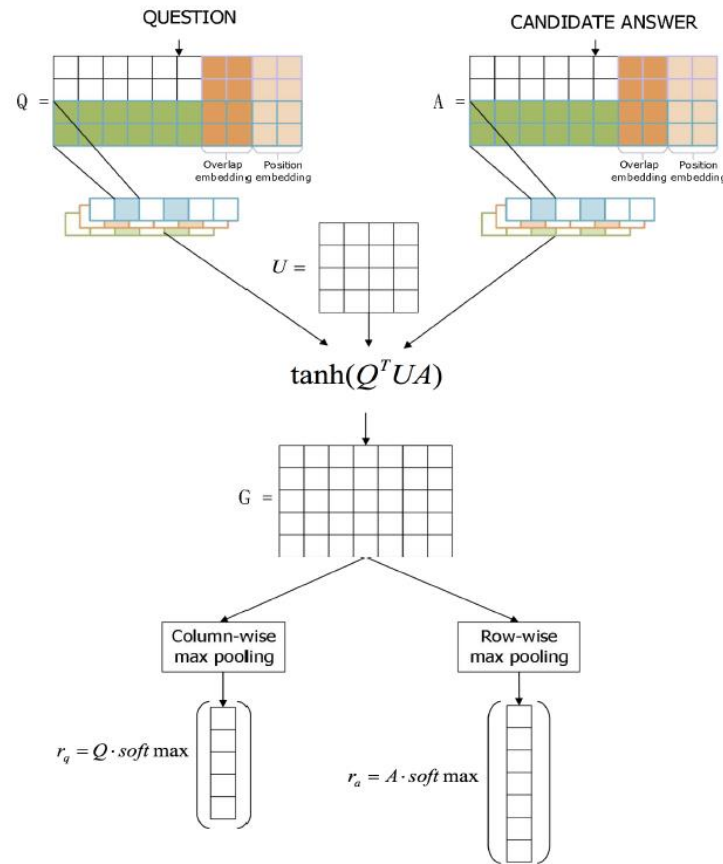
Table 1. The result of our approach.

method	MAP	MRR
Average Word Embedding	0.4610	0.4610
Machine Translation	0.2410	0.2412
Paraphrase	0.4886	0.4906
Word Overlap	0.5114	0.5134
Count-based features	0.7750	0.7756
Embedding-based features	0.7467	0.7470
All features	0.8005	0.8008

基于词向量的计算没有计数方法鲁棒！！！！

Wang, B., Niu, J., Ma, L., Zhang, Y., Zhang, L., Li, J., ... & Song, D. A Chinese Question Answering Approach Integrating Count-based and Embedding-based.

CNN with Enhanced Embedding



基于对抗网络的问答系统

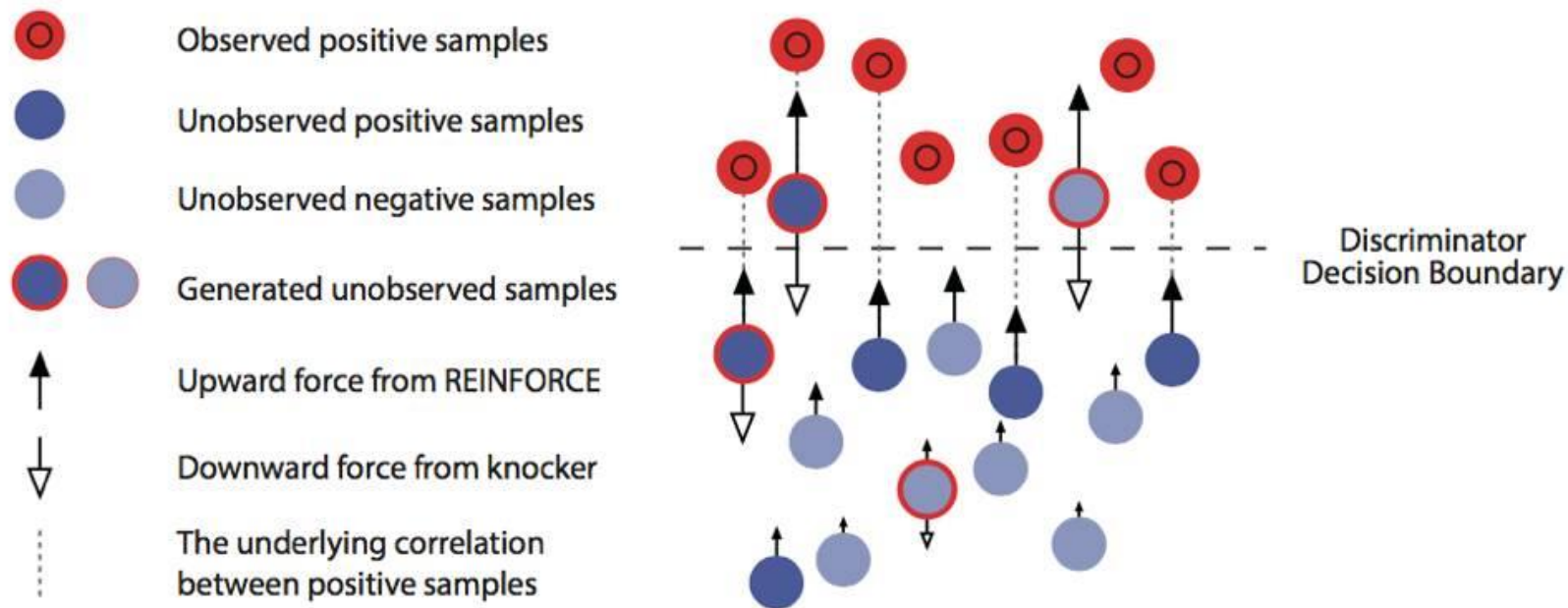


Figure 1: An illustration of IRGAN training.

Wang J, Yu L, Zhang W, Gong Y, Xu Y, **Wang B** et al. IRGAN: A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models[C]. **SIGIR** 2017 long paper. [CCF-A] (three strong-accept reviews and SIGIR **best paper** Honorable Mention)

问答中的量子语言模型

