# CCL2016

# CNCC2016


2016中国计算机大会


iscussion议题
CNCC
自然语言问答与聊天机器人
Panel Discussion议题

# CNCC2016：面向互联网大数据的语言与知识计算

- 宗成庆 篇章关系分析与话题深层理解
- 赵军 深度问答：任务、方法和资源
- 刘挺 基于伪数据的自然语言处理方法
- 周国栋 面向互联网大数据的语言理解与认知
- 李涓子 科学网络中的深度知识挖掘
- 孙茂松 基于深度学习的语言与知识计算

# 深度问答的性能

- 限定域深度问答：Geoquery (2001)
  - 问题相对更长更复杂，包含多个概念和关系
- 开放域深度问答：WebQuestion+Freebase (2013)
  - 问题相对更短更简单，一般只包含一个概念和一个关系

| | 方法 | 限定域深度问答 | 开放域深度问答 |
|---|---|---|---|
| 符号表示方法 | CCG | 89.0% (Zettlemoyer et al. 2009) | - |
| | DCS | 91.1% (Liang et, al. 2011) | 39.9% (Berant et, al. 2014) |
| 分布表示方法 | Sum | --- | 39.2% (Bordes et, al. 2014) |
| | CNN | --- | 40.8% (Dong et, al. 2015) |
| | Attention | 84.6%(Dong et, al. 2016) | 42.6% (Zhang et, al. 2016) |
| 混合方法 | NN+Transition | --- | 52.5% (Yih et, al. 2015) |

# Deep Learning, why success?

- **模型**
  - 深度神经网络：更丰富的连接网络可以建模更多的复杂任务
  - 152 layers [ResNet 16], 1000+layers [DN with Stochastic Depth 16]
  - 每2.4年，神经网络中隐含层节点翻番
- **资源**
  - 每个类别5,000个标注样本可以达到理想的性能
  - 超过亿级标注样本就可以超过人类能力
- **计算**
  - 更快的CPU和GPU，更通用的GPU，更快的连接网络
  - 更好的分布式计算平台

Larger networks are able to achieve higher accuracy on more complex tasks with larger datasets.

(Ian Goodfellow, Yoshua Bengio et, al. Deep Learning, 2016)

# 深度学习之NLP

- **深度学习的本质**
  - 从原始数据中学习正确的表示
  - 以分层次的计算方式来学习多步骤的程序
  - 高层（抽象、复杂的）概念建立在底层（简单的）概念之上
- **NLP任务**
  - 资源是否足够多
    - » 知识库：远远不够，现有知识库都存在大量知识缺失，更不用说常识知识库的缺失了
    - » 语料：远远不够，目前单关系简单问题的语料可以达到10万级，但平均到每个任务（关系、属性）和每个概念就非常少了；Freebase包含2千多类概念和3万多关系
  - 模型是否满足任务要求
    - » 自然语言大量存在顺序、树、图等复杂结构，而语音识别没有顺序、图像处理中的结构也相对简单，NLP中还需要探索更合适的模型
    - » 词用向量表示，短语、句子、段落和知识单元等更大语义单元也用向量表示是否足够

**数据和知识如何融合？**

**数值方法和符号方法如何融合？**

# 符号表示（符号匹配） vs 分布表示（数值计算）

- 符号表示的方法
  - 知识表示为符号化的逻辑语义表达式，在符号系统上进行查询和推理
  - 准确度好，覆盖度低，泛化能力弱，不方便扩展，推理效率低
  - 用户可理解

**case-by-case 模块化(过程式)方法**

- 分布表示的方法
  - 知识转化为低维空间中的数值表示，在数值空间中计算知识元素和问句的匹配程度
  - 准确度弱，覆盖度好，泛化能力强，容易扩展和融合到其他系统，效率相对高
  - 不可解释

**end-to-end 端到端(黑盒)的方法**

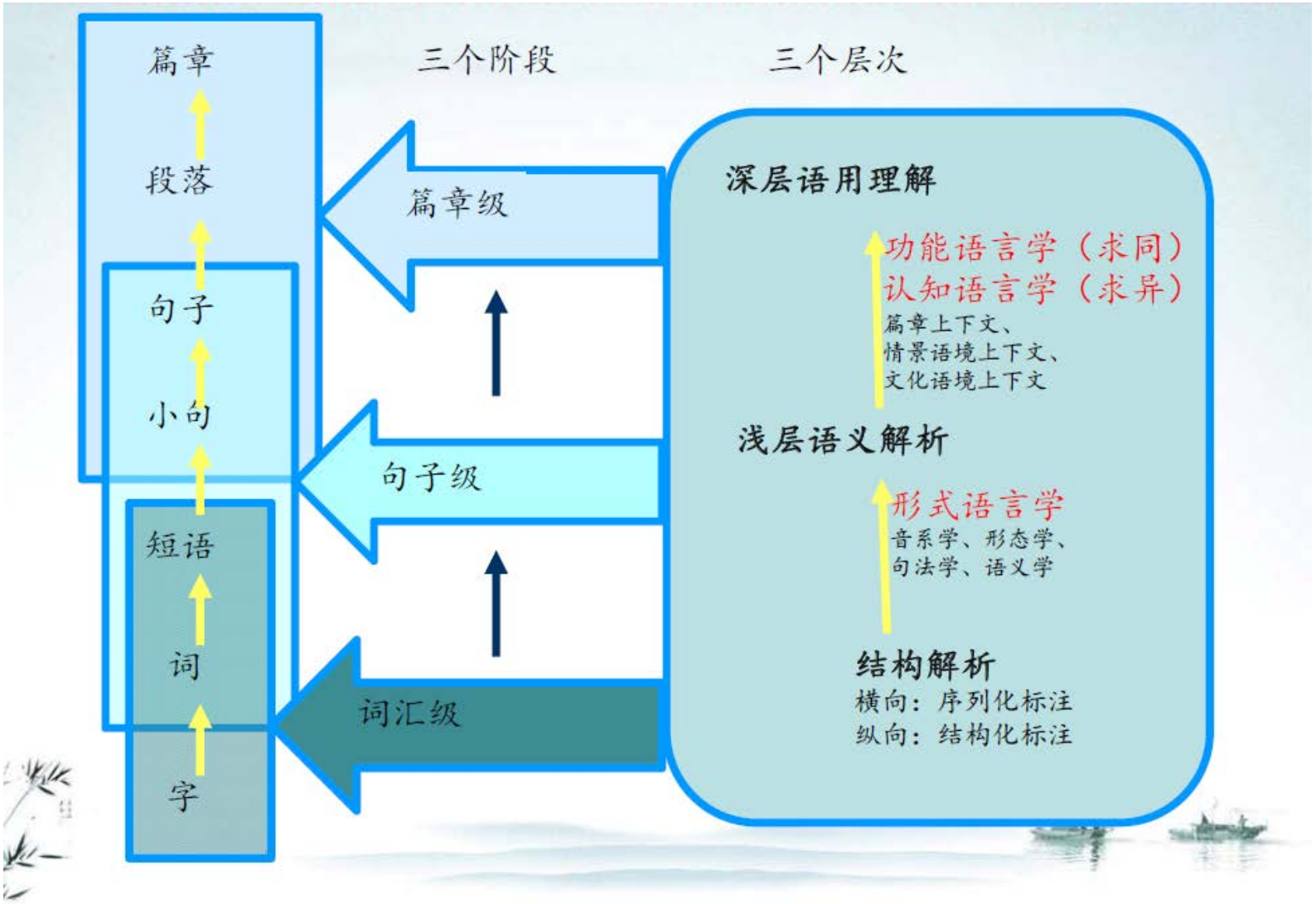
**Deep Learning的方法**

# 未来本实验室NLP研究重点

- **词**
  - 命名实体识别
- **句子**
- **篇章**
  - 主题探测
  - 自动摘要
  - 情感分析
  - 自动问答
- **应用**
  - 问答系统
  - 看图说话：诗词、导航
  - 知识库收集与制作

- **基础**
  - 深度学习模型
  - 主题模型
  - 分词系统、篇章树、知识库

- **创新点**
  - 符号方法与分布式方法的相结合
  - 半监督与无监督的学习
  - 黑盒分析
  - 封闭式——开放式


篇章
段落
句子
小句
短语
词
字
三个阶段
篇章级
句子级
词汇级
三个层次
深层语用理解
功能语言学（求同）
认知语言学（求异）
篇章上下文、
情景语境上下文、
文化语境上下文
浅层语义解析
形式语言学
音系学、形态学、
句法学、语义学
结构解析
横向：序列化标注
纵向：结构化标注

# 2. 篇章关系分析

- **Results and analysis**
  - *More than 20% sharp decrease of F1 in explicit parser on the blind set. This is mainly due to the error propagation of discourse connective identification.*

| | Task | Dev | Test | Blind |
|---|---|---|---|---|
| **Explicit** | *Conn* | 0.8356 | 0.7263 | 0.5627 |
| | *Arg1* | 0.5479 | 0.5587 | 0.3853 |
| | *Arg2* | 0.6849 | 0.6816 | 0.4444 |
| | *Both* | 0.4521 | 0.4916 | 0.2650 |
| | *Sense* | 0.7534 | 0.6480 | 0.4811 |
| | *Parser* | 0.4521 | 0.4859 | 0.2446 |
| **Non-Explicit** | *Conn* | – | – | – |
| | *Arg1* | 0.6282 | 0.6266 | 0.5526 |
| | *Arg2* | 0.6798 | 0.6762 | 0.6017 |
| | *Both* | 0.5341 | 0.5379 | 0.4457 |
| | *Sense* | 0.5068 | 0.4987 | 0.4082 |
| | *Parser* | 0.3982 | 0.3869 | 0.2712 |
| **All** | *Conn* | 0.8356 | 0.7263 | 0.5627 |
| | *Arg1* | 0.6261 | 0.6328 | 0.5439 |
| | *Arg2* | 0.6932 | 0.6921 | 0.5843 |
| | *Both* | 0.5317 | 0.5418 | 0.4178 |
| | *Sense* | 0.5640 | 0.5333 | 0.4326 |
| | ***Parser*** | **0.4120** | **0.4089** | **0.2690** |

- 汉语篇章分析与语料库建设刚刚起步，有太多的理论探索空间，篇章关系分析只是其中的一个基本问题

  (1)我喜欢在春天去观赏桃花，在夏天去欣赏荷花，在秋天去观赏红叶，但更喜欢在冬天去欣赏雪景。

  (2)他整天在外面奔波，根本顾不上家里，孩子疏远他，妻子抱怨他，自己也感到精疲力竭，身体一天天跨了下来。

- 面向具体应用(如机器翻译、自动文摘、问答系统等)的篇章分析方法研究仅摸索阶段

- 归纳推理、常识学习和特定语用场景下的交互学习是文本深层理解的必要手段

# 问答：深度问答

- FAQ、CQA：主观问题，复杂问题，限定域问题
- IR-QA：事实性问题，利用网络冗余信息
- KB-QA：知识抽取、表示与推理，深度问答
- MC-QA：文本理解与推理，深度理解
- 对话：个人助理、多轮交互、会话建模
- …

Baidu 知道 知乎

Google 谷歌 Baidu 百度

WolframAlpha computational knowledge engine

高考

Todai Robot


# 深度问答的信息源

***问题输入***

自然语言问句
(非结构化)

**知识问答系统**

精确答案

***内容输入***

结构化知识

Knowledge Bases
(多)知识库


Triples
抽取事实

(Chicago, be most city in, the United States)
(Charlotte, grandpa, Bill Clinton)
(Chicago, locate in,Illinois)
(Hillary, be politician in, the democratic party)
(Edward H. Allegretti, be born in, Chicago)

非结构化知识

Texts
文本

单知识库 → 多知识库?
简单问题 → 复杂问题?

# 深度问答的计算任务

- **知识表示**：把人类的知识系统，包括对世界的认知、常识等知识用计算机可以存储和计算的形式进行表示。

- **内容理解和知识抽取**：从文本等模态数据中抽取出计算机可以表示的知识的过程（狭义上来讲，就是把非结构化数据转换为结构化的知识的过程）。

- **问句理解**：理解问句的含义，并与已有的知识系统（语义系统）进行匹配。

- **知识推理**：利用常识、现有知识等先验知识，发现未知知识。

# 深度问答的方法

- **符号表示：符号匹配的方法**
  - 将文本等内容和问句转化为知识符号表示，在符号系统上进行语义的查询和推理
- **分布表示：数值计算的方法**
  - 将文本、知识和问句都转化为低维空间中的数值表示，在低维数值空间中计算知识系统中的知识元素和问句语义的匹配程度

# 符号化的知识表示：结构化知识

形式化语义表示方法：

1) 产生式系统

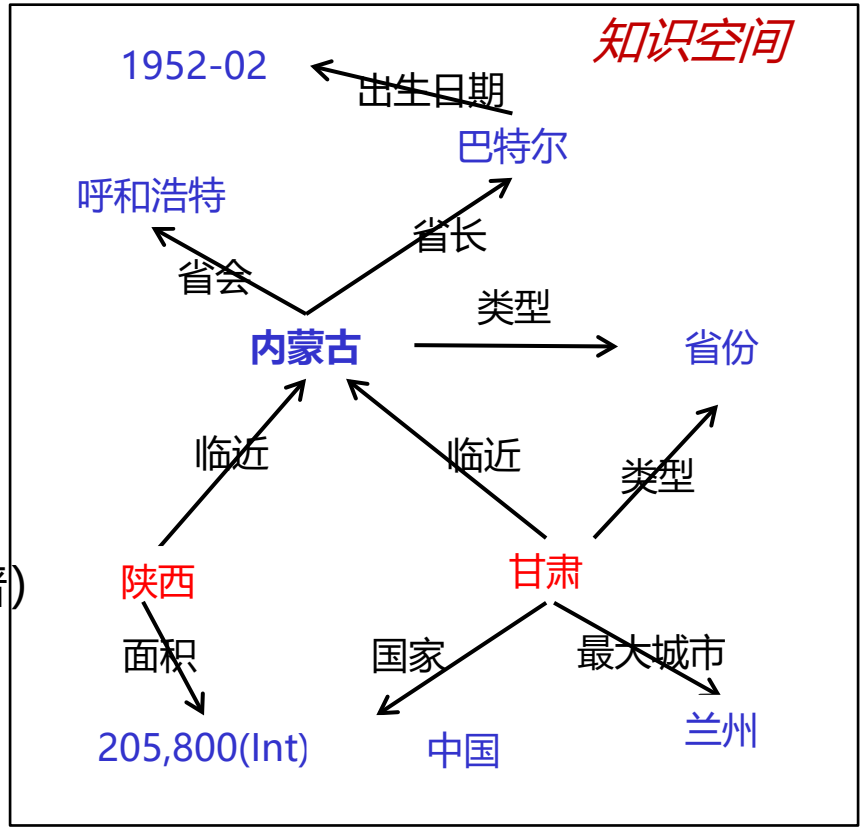
2) 谓词逻辑

3) 框架

4) 脚本

5) 语义网络（知识图谱）

…




数理演算：

1)lambda演算

…

近似推理：

1)归纳逻辑编程

2)马尔科夫逻辑网

3)概率软逻辑

…

概念或关系的语义蕴含在形式化知识结构之中
通过数理演算和近似推理进行语义计算（问句解析和知识推理）

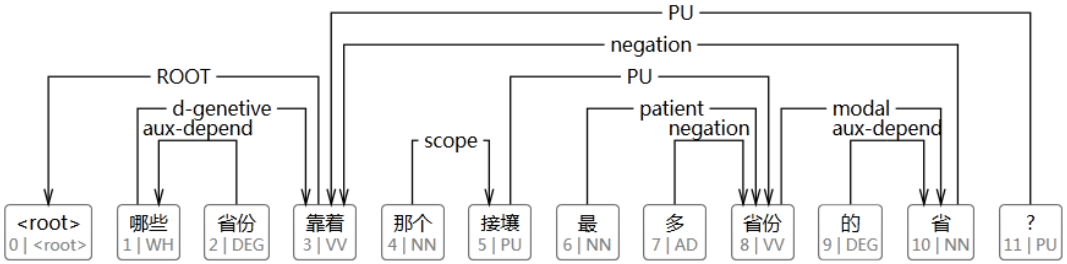
# 符号化的知识表示：非结构化知识

*哪些省份靠着那个接壤最多省份的省?*

- 关键词组合

{靠着、省份、接壤、最多}

- 语义树(图)结构




- 逻辑表达式结构

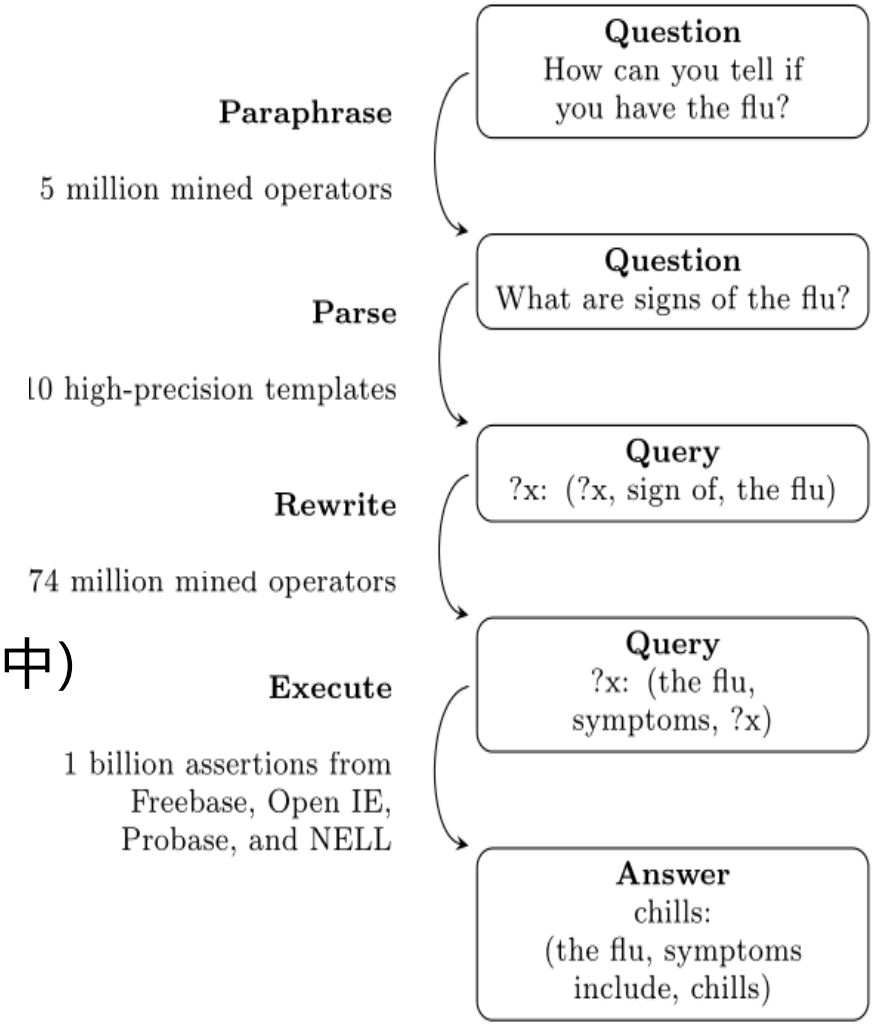
λx. 省 (x) ^ 临近(x, argmax( λy. 省(y), λy. count(λz. 省(z) ^ 临近(y, z))))

# 基于复述的语义匹配：事实库问答

[Fader et al. KDD 2014]

- 回答事实性问题
- 信息抽取工具抽取出事实三元组
- 如何利用这些事实三元组回答问题

- 步骤1：问句复述（问句的不同问法）
- 步骤2：问句解析（问句映射到事实库中）
- 步骤3：查询转换（事实的不同表示）
- 步骤4：执行查询，回答问题

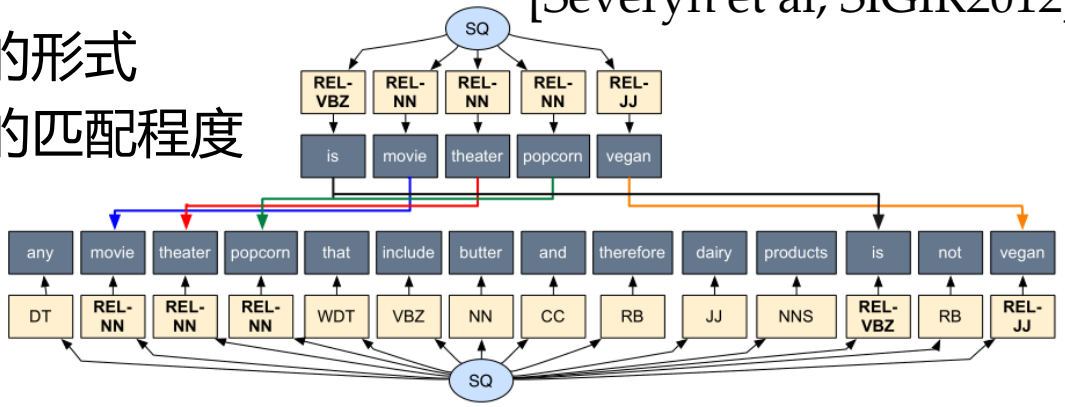


# 基于句法树的语义匹配：文本库问答

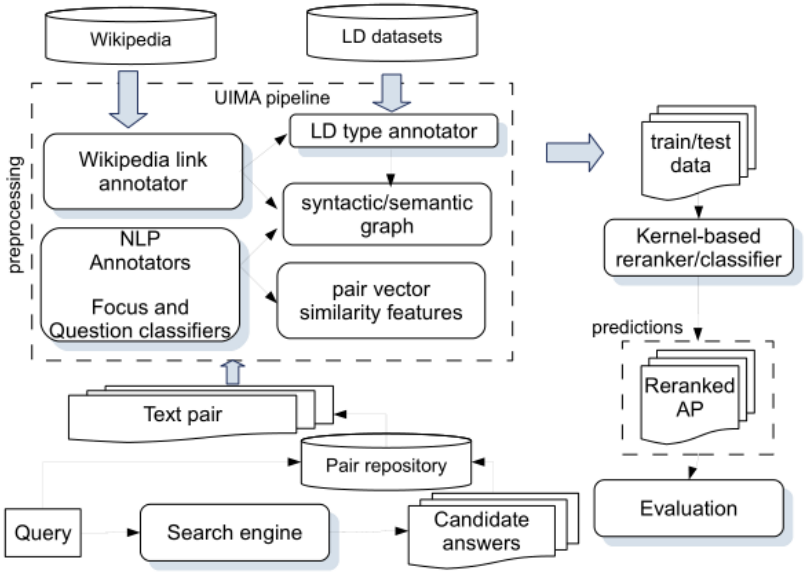
[Severyn et al, SIGIR2012]

- 问句和答案表示为句法树的形式
- 基于树核函数计算树结构的匹配程度

$$TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_2 \in N_{T_2}} \Delta(n_1, n_2)$$



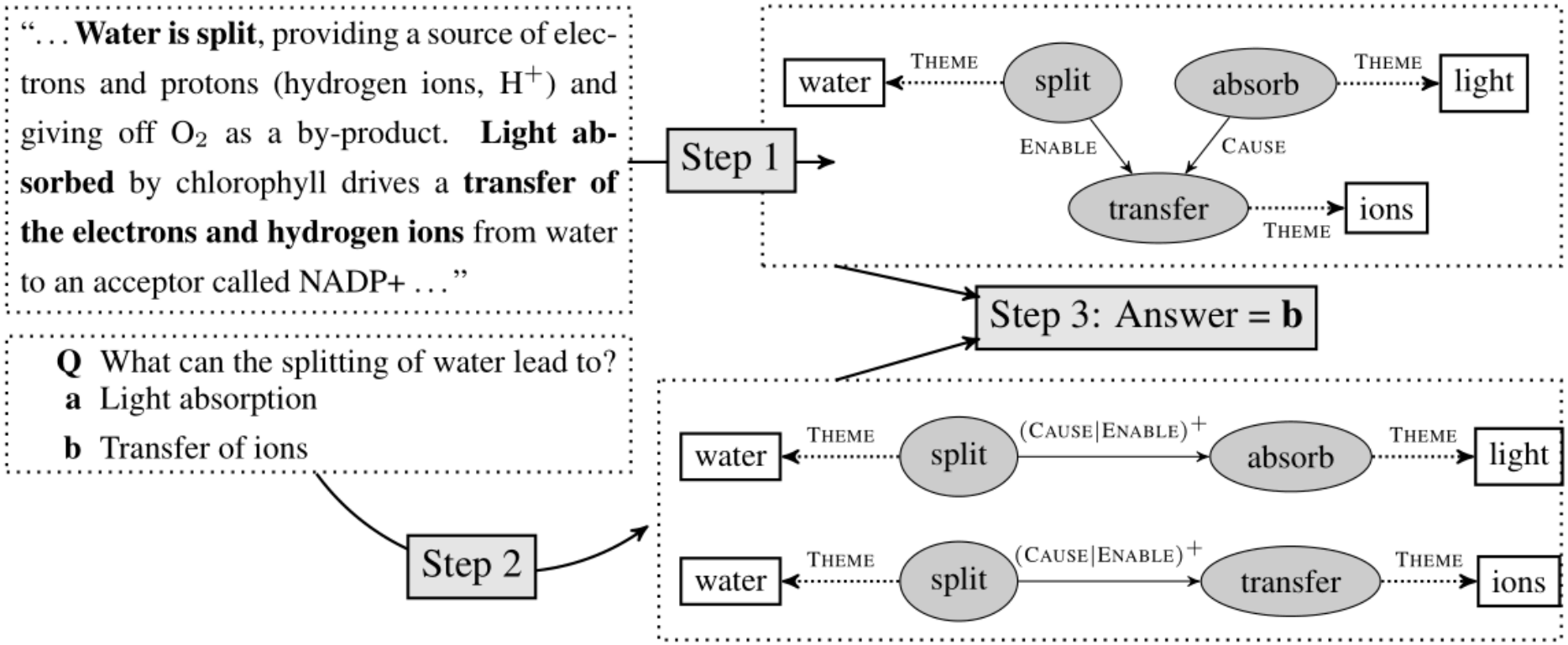

- 答案篇章重排序
- 利用基于浅层和深度句法分析器的句法和语义结构的树核函数
- 在表示中利用链接到开放知识库中的资源进行匹配




[Tymoshenko et al, CIKM 2015]

# 基于篇章语义图的语义匹配：阅读理解

[Berant et al. EMNLP 2014]



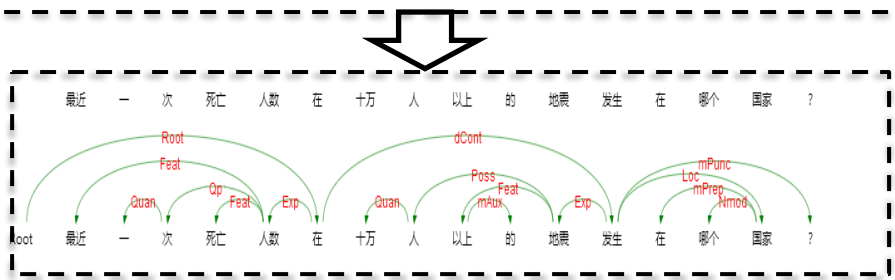

- 步骤1：分析输入段落的结构，用实体、事件关系图表示其含义
- 步骤2：对于每个答案，把问题-答案组合分析为同样的结构，形成查询
- 步骤3：在语义结构图中进行匹配，得到最终答案

# 基于逻辑表达式的语义匹配：知识库问答

[Shizhu He et al. EMNLP 2014]

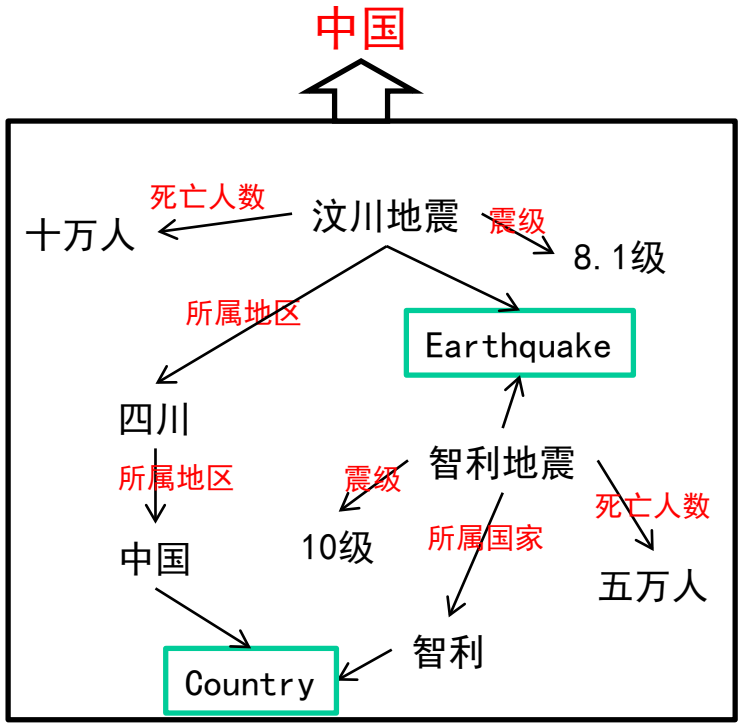
- 将自然语言问句转换为结构化查询语句（如SPARQL语句）

最近一次死亡人数在十万人以上的地震发生在哪个国家？



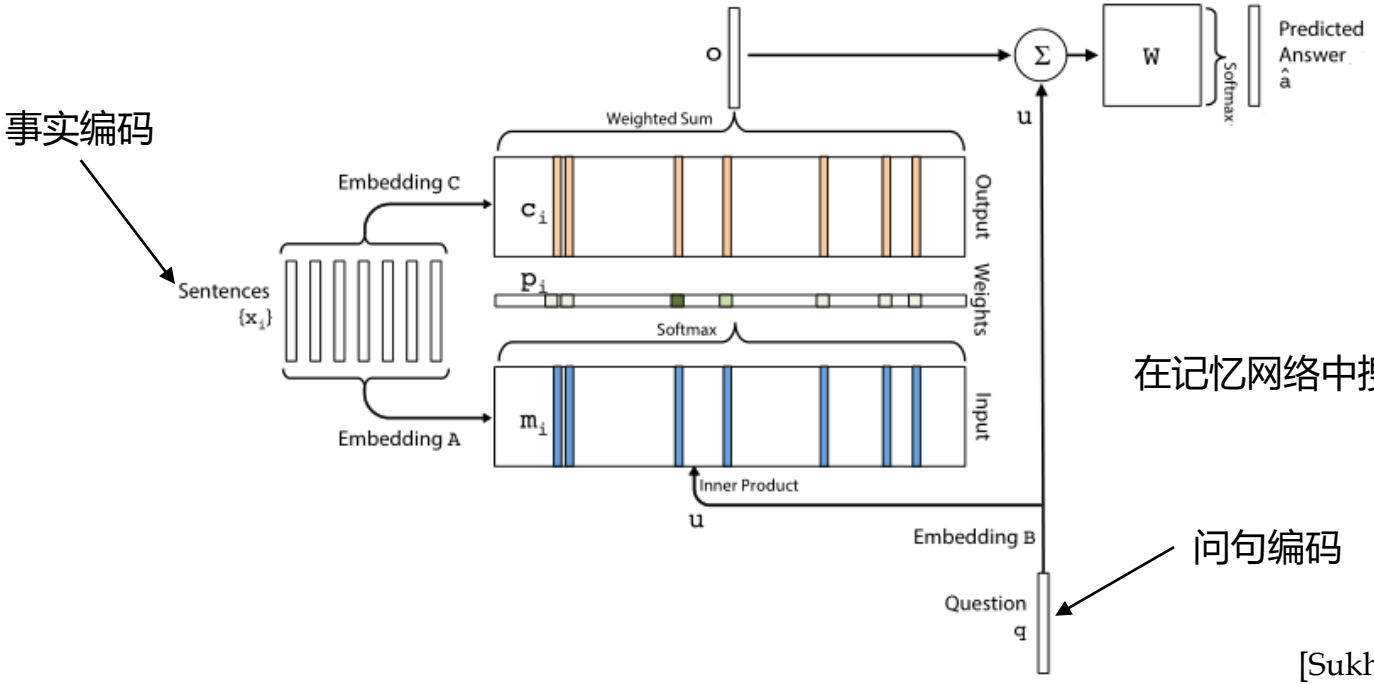

语义解析

```
SELECT DISTINCT ? X
WHERE{
    ?y 所属国家 ?x; ?y 死亡人数 ?p;
    ?y 时间 ?t; argmin(?t-now);
    max(?t-now, 0); max(?p, 十万);
}
```



# 在潜在空间中表示含义

- 基于神经网络对问句、文本、知识图谱和答案直接编码，表示为向量、矩阵等形式
  - 不需要NLP操作
  - 不需要人工提取特征
  - 在潜在空间中学习问句和答案的匹配

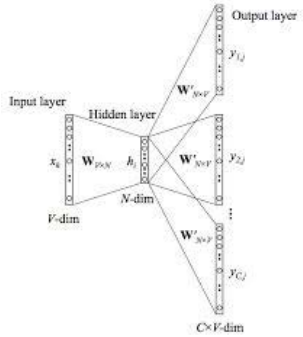



在记忆网络中搜索支撑事实再进行推理

[Sukhbaatar et al, 2015]

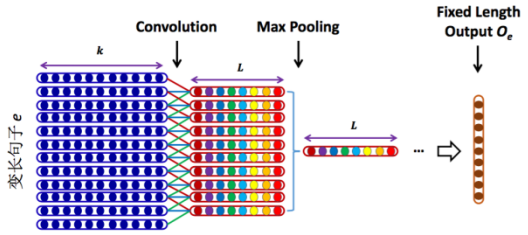
# 文本的向量化表示

- **词的向量化**
  - NNLM, C&W, Skip-gram

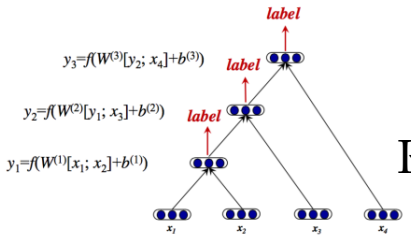



Word Embedding

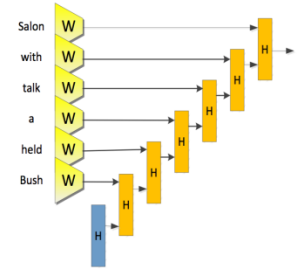
- **句子（文本）的向量化**
  - 直接累加
  - 卷积神经网络
    (*Convolutional Neural Network*)
  - 递归神经网络
    ( *Recursive neural network* )
  - 循环神经网络
    ( *Recurrent neural network* )




CNN



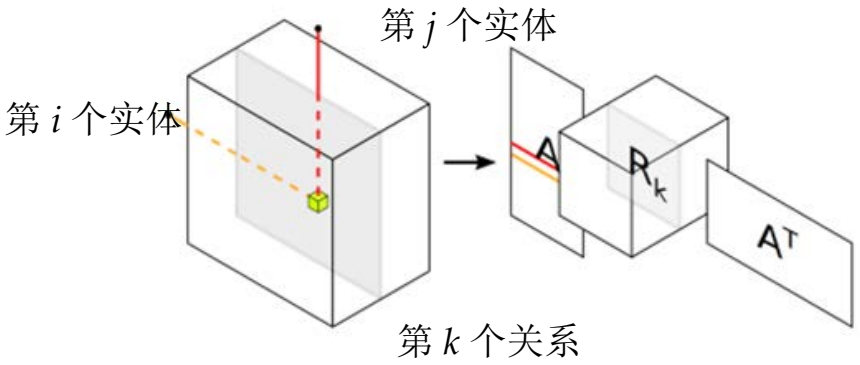

Recursive NN




Recurrent NN

# 知识库的向量化表示（知识库的表示学习）

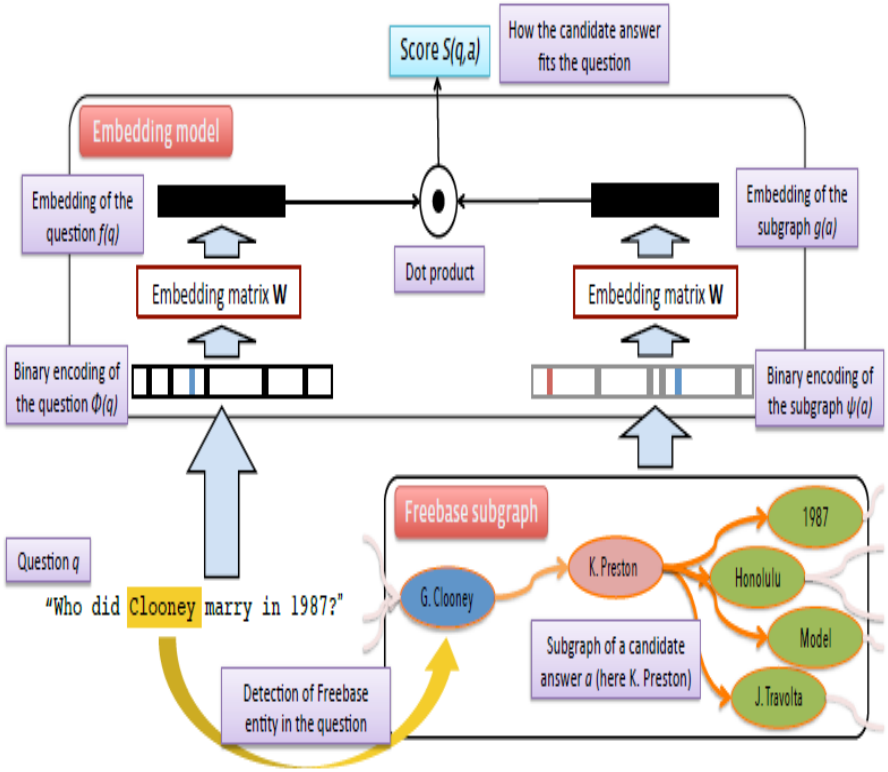
- 把知识库中的实体和关系表示为低维空间的对象（向量）及其它们的操作（空间转换）；
- 该表示能够蕴涵其在知识库中的性质，即具有类似上下文的对象，在低维空间中更接近。



# 问句和知识库的联合分布表示学习

[Borders et al. EMNLP 2014]

- 对问句和答案（在知识图谱中对应的子图）在同一个空间中进行联合表示学习，学习问句中的词语以及知识库中的实体和关系的表示和匹配




– 问句和子图进行向量表示

» 问句：$f(q) = \sum_{w \in q} vec(w)$

» 子图：$g(a) = \sum_{l \in g} vec(l)$

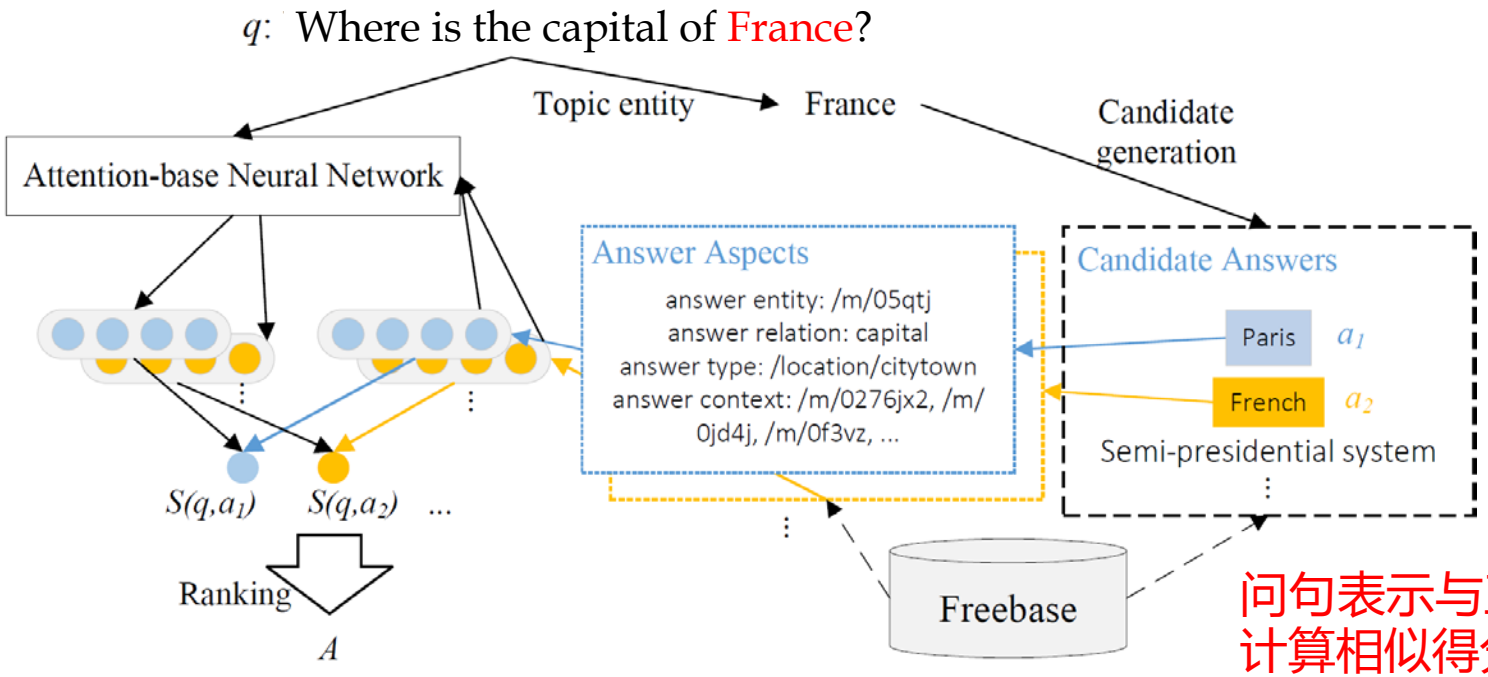
– 计算向量相似度对子图进行排序

» $s(q,a) = f(q)^T \cdot g(a)$

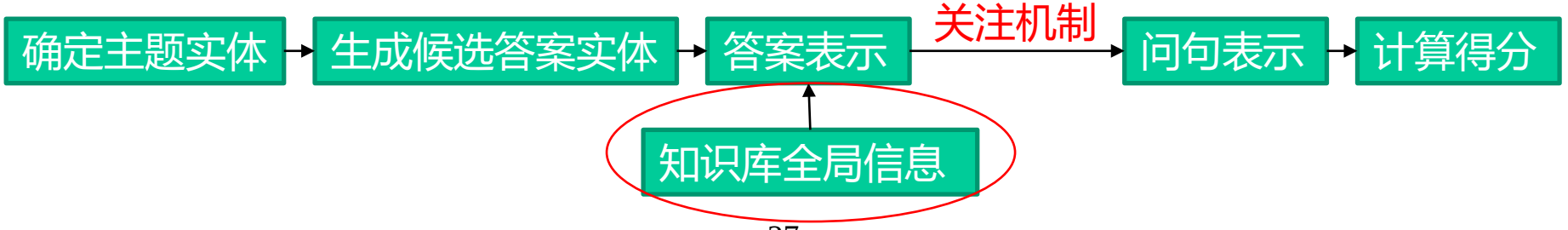
» $a' = \sum_{a \in A(q)} s(q,a)$

# 基于关注机制的问句表示和知识库问答

[Yuanzhe Zhang et al. AAAI 2016]




问句表示与正确答案计算相似得分更高。



# 基于神经网络的阅读理解

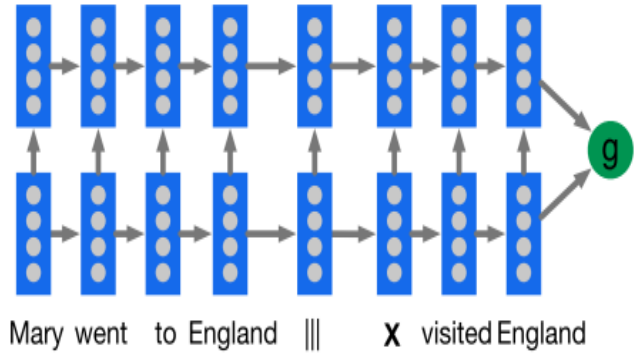
Teaching machines to read and comprehend ., Google DeepMind（ Karl Moritz Hermanny et al.) NIPS 2015

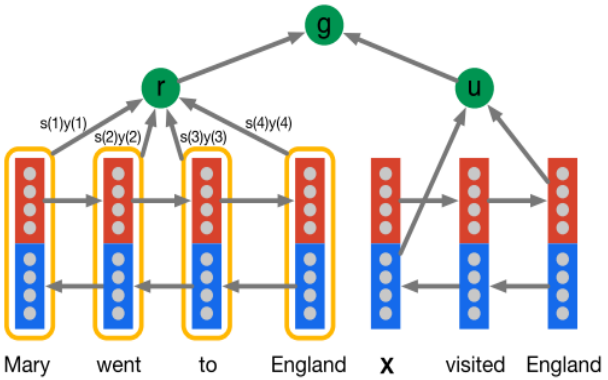
- 在基于循环神经网络对文档和问题进行表示的基础上，计算出文档中每个实体相对于问题的概率作为答案输出

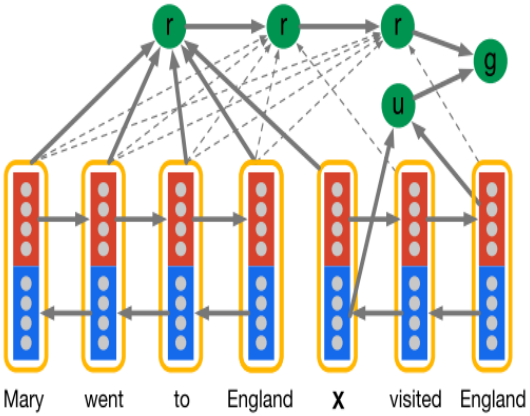
***MODEL1:Deep LSTM：*** 依靠Long-Short Term Memory（LSTM）分别表示背景文档和问题，然后计算相似度。LSTM能够捕捉句子中的背景文档中的长距离语义信息。

***MODEL2:Attentive Reader：*** 在表示背景文档的时候，针对不同的问题，文档的表示也应该不同，这就是关注机制（Attention Mechanism）。在整个问句表示完毕，才进行关注过程。

***MODEL2:Attentive Reader：*** 在表示背景文档的时候，针对不同的问题，文档的表示也应该不同，这就是关注机制（Attention Mechanism）。在整个问句表示完毕，才进行关注过程。



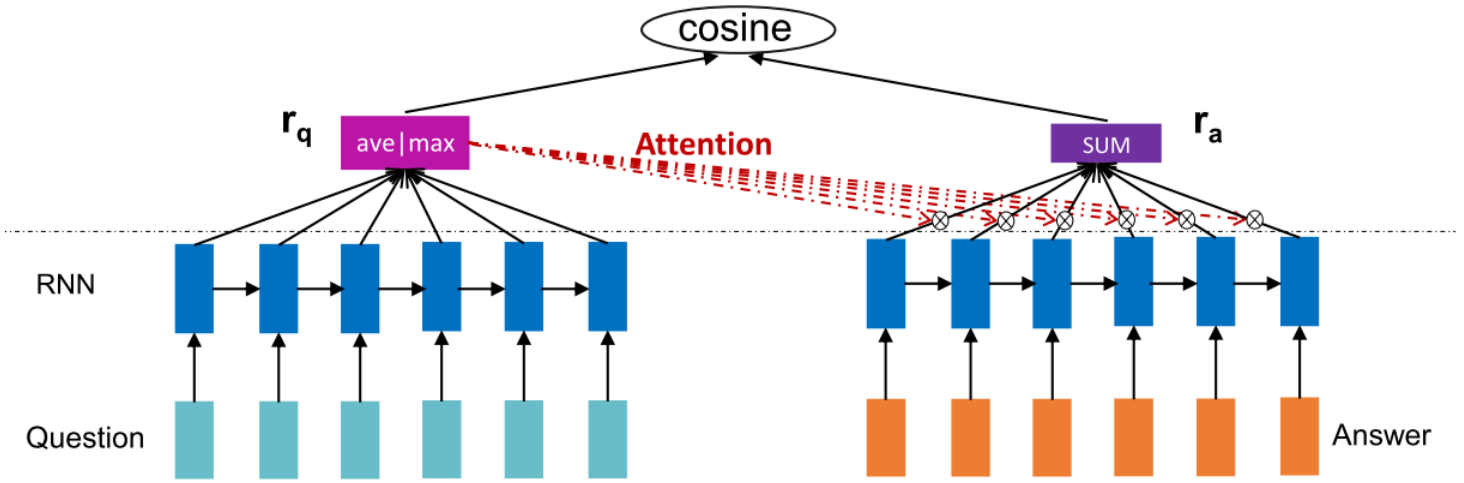
# 基于神经网络的答案选择

Inner Attention based Recurrent Neural Network for Answer Selection . Wang et al. ACL2016

## ■Motivation

✓传统方法的问题：一般的基于关注机制的答案选择过程都是在隐含层上进行关注，但是在循环神经网络中当前隐含层包含了这次词之前所有的信息，因而关注的权重不是这个词的权重，而是这个词之前的句子的权重。

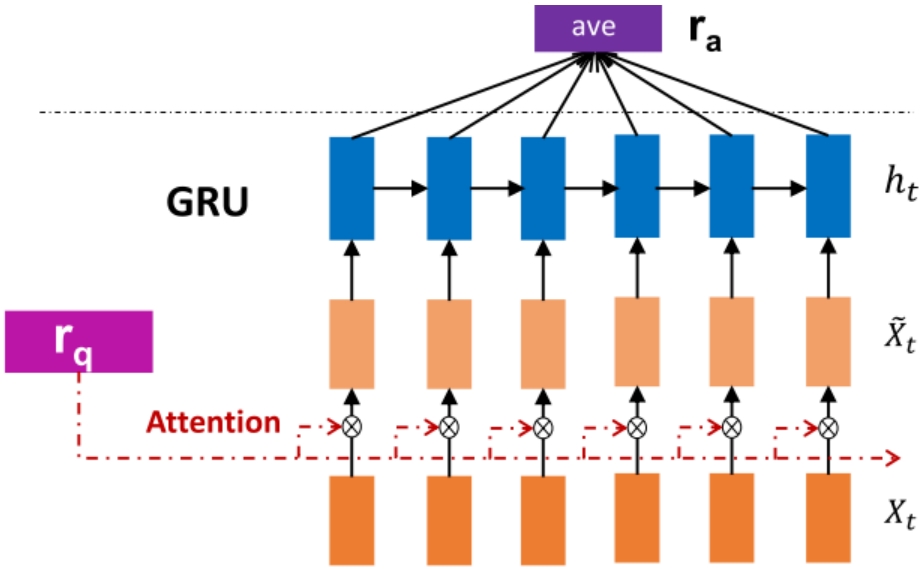
✓出发点：应该更加关注于当前词本身的信息。



# 基于神经网络的答案选择

Inner Attention based Recurrent Neural Network for Answer Selection . Wang et al. ACL2016

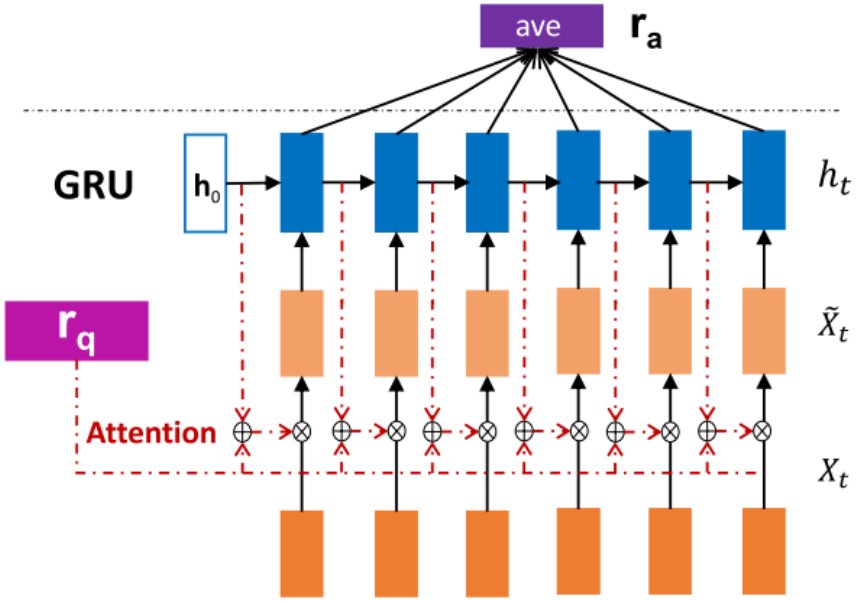
**模型一**：直接将attention加入到原始的词向量上



# 基于神经网络的答案选择

Inner Attention based Recurrent Neural Network for Answer Selection . Wang et al. ACL2016

模型二：模型一每次只是加入一个词向量的权重，而很多时候句子中有意义的部分往往是连续的几个词，如 hot dog。而且，一个词的重要性要根据它的上下文来判断，因此，我们将上下文也引入到attention中。

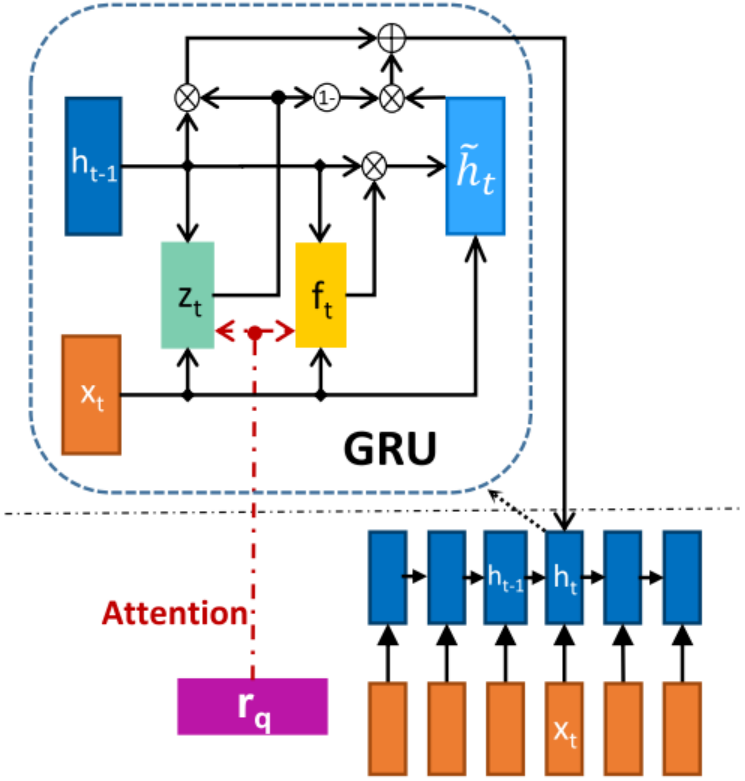



这里的上下文就是以前词的隐含层表示

# 基于神经网络的答案选择

Inner Attention based Recurrent Neural Network for Answer Selection . Wang et al. ACL2016

模型三：前两个模型直接将attention加入到词向量中，为了更好地将attention融入到隐含层中，我们直接将attention的信息加入到RNN内部循环单元中。受以前长短时记忆的启发，直接将attention信息加入到内部的gate中。




这里用的是GRU（Gated Recurrent Unit）

# 基于神经网络的答案选择

Inner Attention based Recurrent Neural Network for Answer Selection . Wang et al. ACL2016

■实验

这三个数据集是阅读理解的第一步（答案选择）

| System | MAP | MRR |
|---|---|---|
| (Wang and Nyberg, 2015) † | 0.7134 | 0.7913 |
| (Wang and Ittycheriah, 2015) † | 0.7460 | 0.8200 |
| (Santos et al., 2016) † | **0.7530** | **0.8511** |
| GRU | 0.6487 | 0.6991 |
| OARNN | 0.6887 | 0.7491 |
| IARNN-word | 0.7098 | 0.7757 |
| IARNN-Occam(word) | 0.7162 | 0.7916 |
| IARNN-context | 0.7232 | 0.8069 |
| IARNN-Occam(context) | 0.7272 | 0.8191 |
| IARNN-Gate | 0.7369 | 0.8208 |

TrecQA

| System | Dev | Test1 | Test2 |
|---|---|---|---|
| (Feng et al., 2015) | 65.4 | 65.3 | 61.0 |
| (Santos et al., 2016) | 66.8 | 67.8 | 60.3 |
| GRU | 59.4 | 53.2 | 58.1 |
| OARNN | 65.4 | 66.1 | 60.2 |
| IARNN-word | 67.2125 | 67.0651 | 61.5896 |
| IARNN-Occam(word) | 69.9130 | 69.5923 | 63.7317 |
| IARNN-context | 67.1025 | 66.7211 | 63.0656 |
| IARNN-Occam(context) | 69.1125 | 68.8651 | **65.1396** |
| IARNN-Gate | **69.9812** | **70.1128** | 62.7965 |

InsuranceQA

| System | MAP | MRR |
|---|---|---|
| (Yang et al., 2015) | 0.652 | 0.6652 |
| (Yin et al., 2015) | 0.6921 | 0.7108 |
| (Santos et al., 2016) | 0.6886 | 0.6957 |
| GRU | 0.6581 | 0.6691 |
| OARNN | 0.6881 | 0.7013 |
| IARNN-word | 0.7098 | 0.7234 |
| IARNN-Occam(word) | 0.7121 | 0.7318 |
| IARNN-context | 0.7182 | 0.7339 |
| IARNN-Occam(context) | **0.7341** | **0.7418** |
| IARNN-Gate | 0.7258 | 0.7394 |

WikiQA