MDCC 2016
中国移动开发者大会
Mobile Developer Conference China 2016

聊天机器人的构建方法：基于模板，检索和深度学习

刘升平 博士
资深专家, AI Labs, 云知声

mdcc.csdn.net

- 聊天机器人简介
- 基于模板的方法
- 基于检索的方法
- 基于深度学习的方法

---个人助理---

Apple SIRI，Google Now

Amazon Echo，小娜，度秘

---娱乐陪伴---

小冰，小黄鸡

---在线客服---

JIMI

---智能问答---

Watson，Wolfram Alpha

- **应答正确**
  - 语法正确
  - 语义一致：不是答非所问，不知所云
- **应答多样性**
  - 有趣的，而不是安全回答，如“好的” “嗯”
- **个性表达一致**
  - 应答一致
  - 语言风格一致
- **有上下文**
  - 能理解多轮对话中的上下文
- **有知性美**
  - 能回答一些知识性的问题，而不全是插科打诨

- **基于模板**
  - 离线准备好一个对话模板库（问句模板和应答模板对），对用户输入的句子，找到匹配的问句模板，然后按照对应的应答模板生成应答
- **基于检索**
  - 离线准备好一个对话库（一问一答的聊天句对），对用户输入的句子，通过在对话库中以搜索匹配的方式进行应答内容提取
- **基于深度学习（序列到序列生成模型）**
  - 离线准备好一个对话库（一问一答的聊天句对）并训练出模型，对用户输入的句子，利用模型逐词或逐字生成回复。

- 聊天机器人简介
- 基于模板的方法
- 基于检索的方法
- 基于深度学习的方法

实例

User： 你知道王思聪是谁吗
Bot： 你问国民老公干嘛，你要跟他借钱吗？

模板

Q1：*王思聪是谁*
A1： 你问国民老公干嘛，你要跟他借钱吗

实例

User： 你知道王思聪的生日吗
Bot： 主人，我和王思聪不是很熟哦

模板

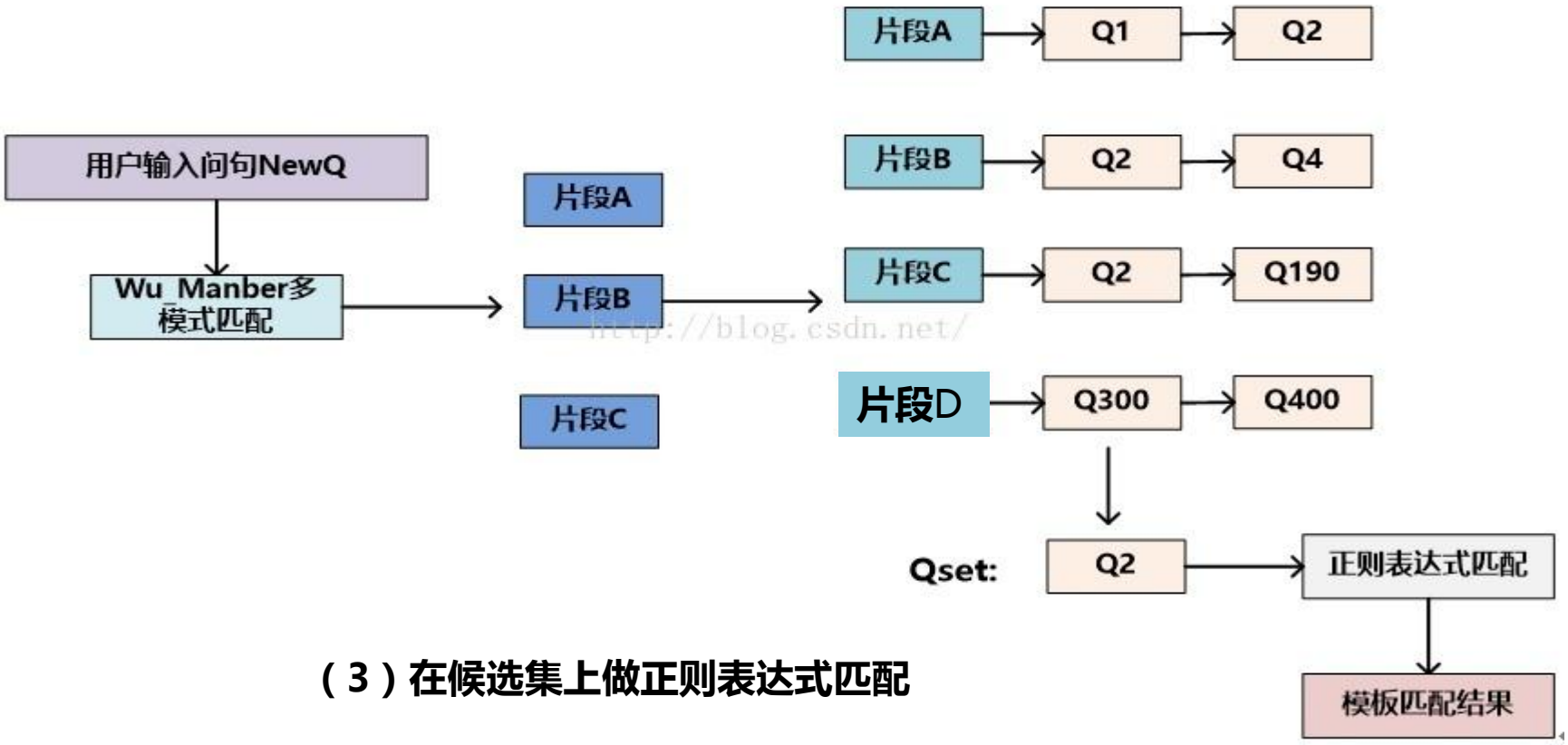
Q2：*@person*生日* （支持变量）
A2：主人，我和@person不是很熟哦

- 对问句模板
  - 分段：以通配符或变量名为切割符
    - *王思聪是谁*　==》 {王思聪是谁}
    - *@person *生日* ==》 {@person,生日}
  - 构造模式字典
    - Dict={王思聪是谁, @person,生日}
  - 建立倒排索引：段 - -》模板集合
    - 王思聪是谁 --》 {Q1}
    - @person --》 {Q2}
    - 生日 --》 {Q2}

- 对新的用户输入
  - 做命名实体识别，把实体名称替换为变量名称
    - 你知道王思聪的生日吗 ➔ 你知道@person的生日吗
  - 实现方法
    - 基于字典
    - 基于统计模型：如CRF
    - 实现工具： lingpipe，CRF++

（1）多模式匹配

（2）内存倒排索引



（3）在候选集上做正则表达式匹配

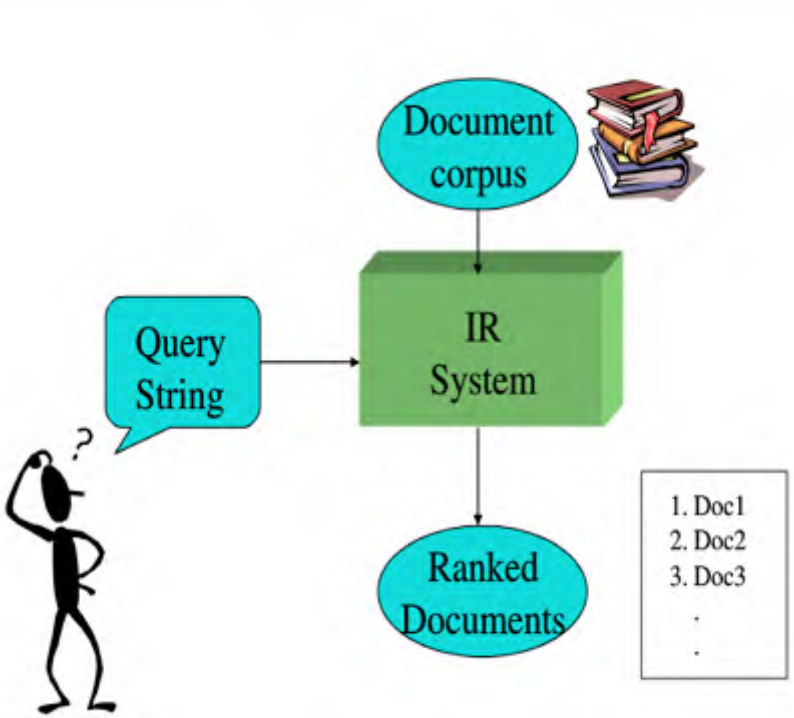
```
<aiml>
    <category>
        <pattern> * 王 思 聪 是 谁* </pattern>
        <template>你问国民老公干嘛，你要跟他借钱吗</template>
    </category>
    <category>
        <pattern>我 头 发 的 颜 色 是 蓝 色 *</pattern>
        <template>哇塞，你很
            <condition name="用户性别">
                <li value="女"> 漂亮阿！ </li>
                <li value="男"> 英俊阿！ </li>
            </condition>
        </template>
    </category>
    <category>
        <pattern>* 名 字 叫 *</pattern>
        <template>呵呵，我知道了，你的名字叫<star index="2"/>。 </template>
    </category>
</aiml>
```

支持一定程度的上下文，记忆，条件匹配等机制

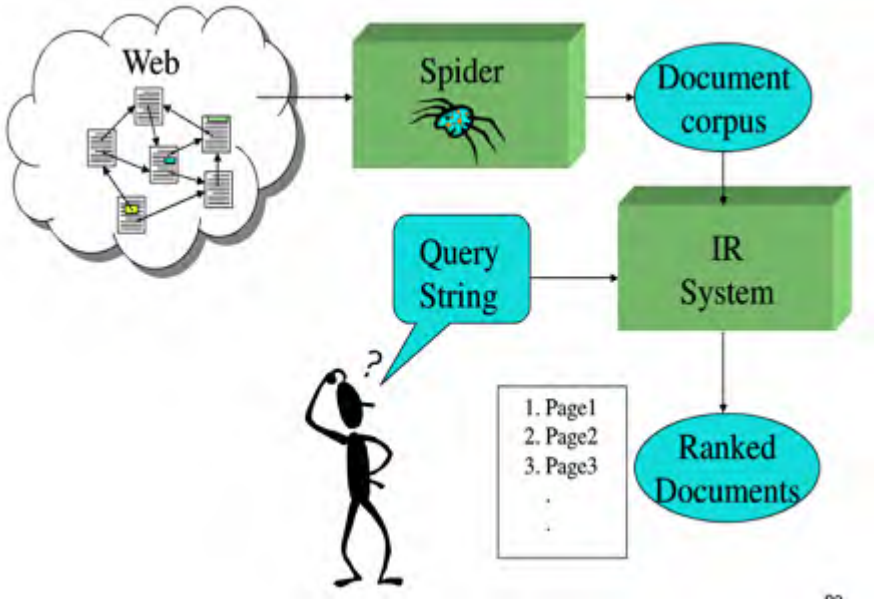
- 张俊林，聊天机器人中对话模板的高效匹配方法， 2016.8.11
- ijuliet，Wu-Manber 经典多模式匹配算法
- woowindice，AIML规范研究文档
- AliceBot，http://www.alicebot.org

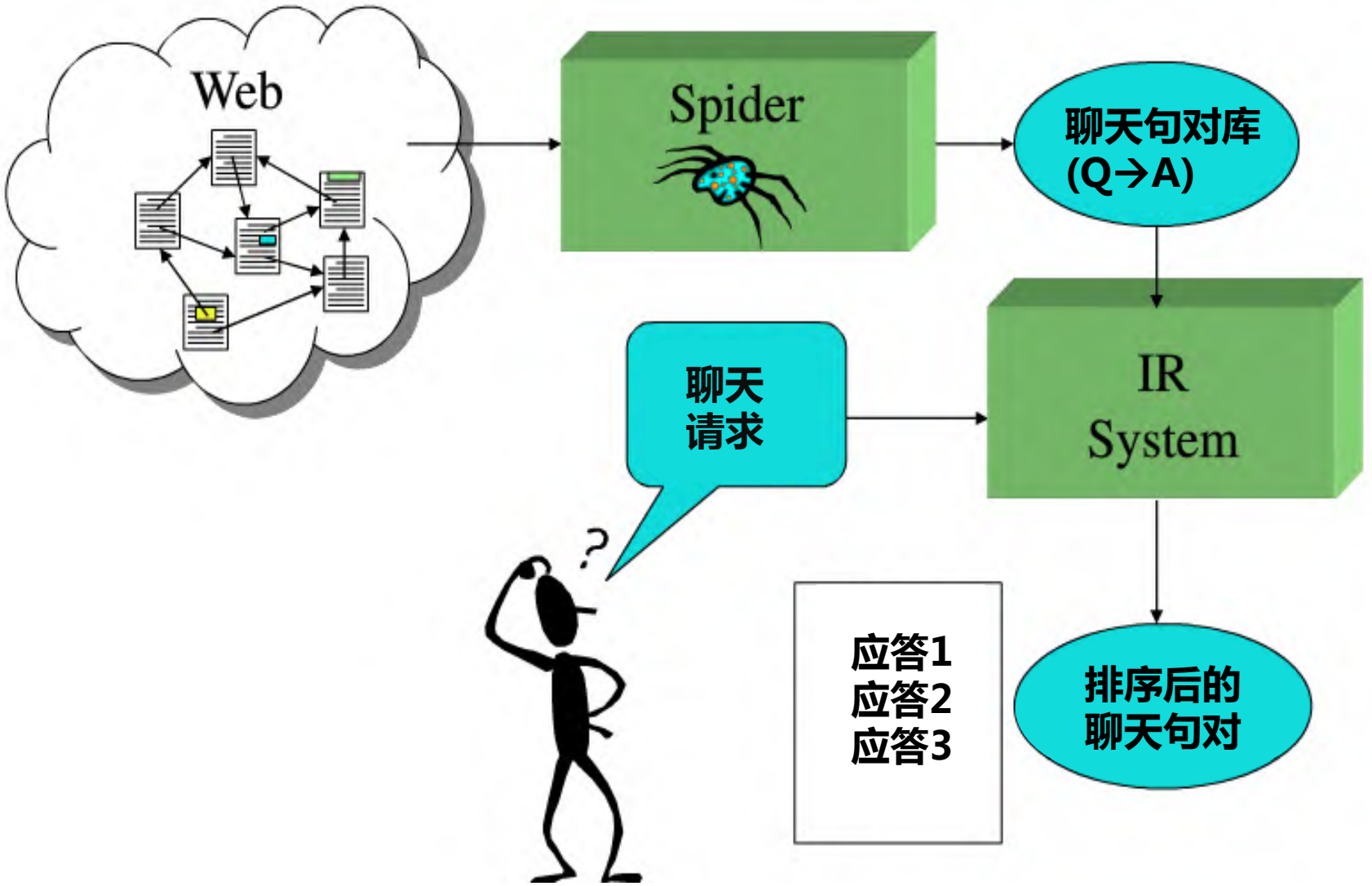
- 聊天机器人简介
- 基于模板的方法
- 基于检索的方法
- 基于深度学习的方法

信息检索


Document corpus
Query String
IR System
Ranked Documents
1. Doc1
2. Doc2
3. Doc3

Web搜索


Web
Spider
Document corpus
Query String
IR System
Ranked Documents
1. Page1
2. Page2
3. Page3


Web
Spider
聊天句对库
(Q→A)
聊天
请求
IR
System
应答1
应答2
应答3
排序后的
聊天句对

- 构建聊天句对库
  - BBS论坛，社区问答类网站（百度知道）
  - 众包收集
- 对问答对做索引
  - 只对问题
  - 对问题+答案
- 对问答对排序
  - 相关性，重要性
- 返回应答
  - 返回top 1，
  - 在top-N 中随机选择1个

- **相关性特征**
  - 词特征
    - 字粒度、分词粒度、短语粒度、命名实体粒度共现
    - 词对词互译概率加权共现、Word Embedding 距离加权共现
    - IDF加权共现、同义词共现
  - 句子特征
    - BM25、VSM、Sentence Embedding距离
  - Topic特征
    - Topic Model 距离

- **重要性特征**
  - term重要性
    - IDF、命名实体、领域词典...
  - answer质量
    - 语言模型混淆度

- **Learning to rank**
  - **Pointwise**
    - **<Q, A>对，独立打分，比如bad、good、perfect、excellent**
    - **一般作为多分类任务训练**
  - **Pairwise**
    - **<Q,A>对，两两组合打分，第一组比第二组好，标为正例；否则标为负例**
    - **一般作为二分类任务训练**
  - **Listwise**
    - **与Pointwise和Pairwise不同， Listwise方法直接学习<Q,list[A]>: query和answer列表排序结果**

Query预处理

User： 你知道王思聪是谁吗
Bot： 你问国民老公干嘛，你要跟他借钱吗？

User： 他为什么那么喜欢网红？

指代消解

User： 王思聪为什么那么喜欢网红？

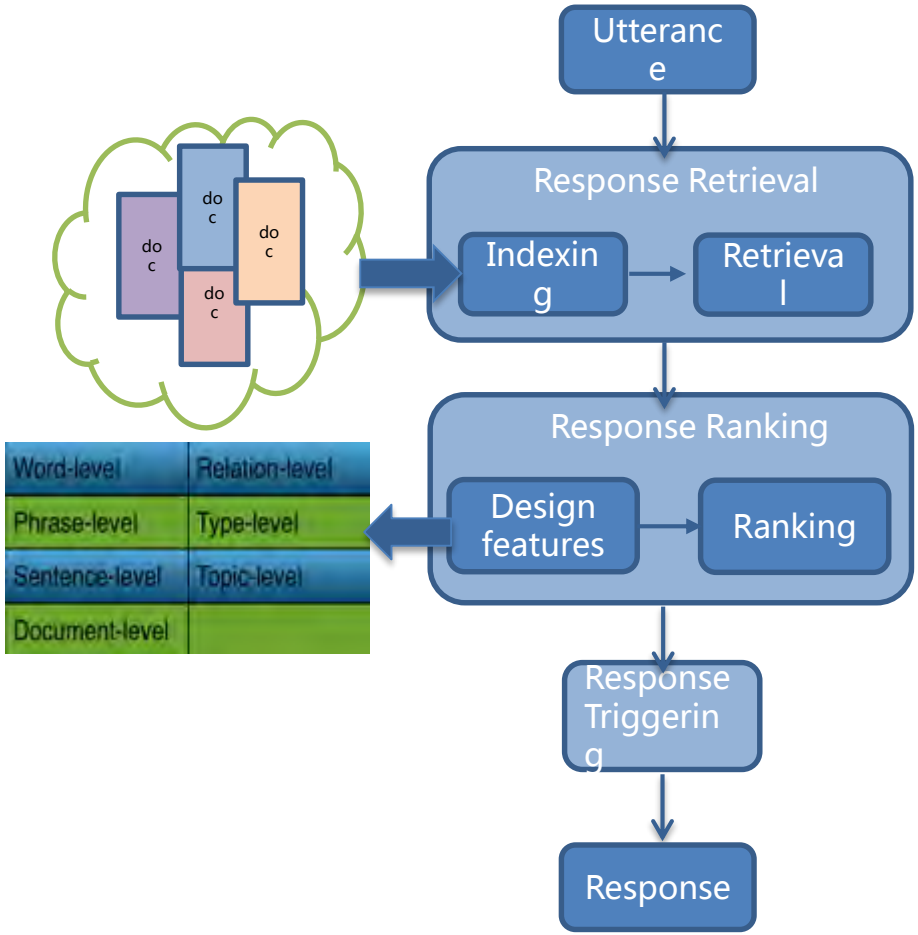
User： 你知道王思聪是谁吗
Bot： 你问国民老公干嘛，你要跟他借钱吗？

User： 孙杨呢？

省略消解

User：你知道孙杨是谁吗？

- **一种基于检索，从非结构化的文档中选取句子作为聊天机器人回复的方法**
- **三步检索**
  - **response retrieval**
  - **response ranking**
  - **response triggering**



有知识

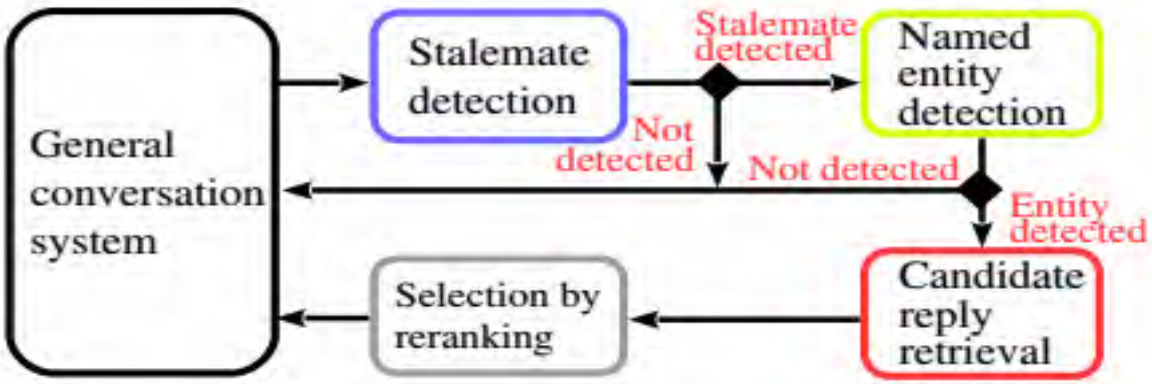


Figure 3: Process flow of triggering content introducing.

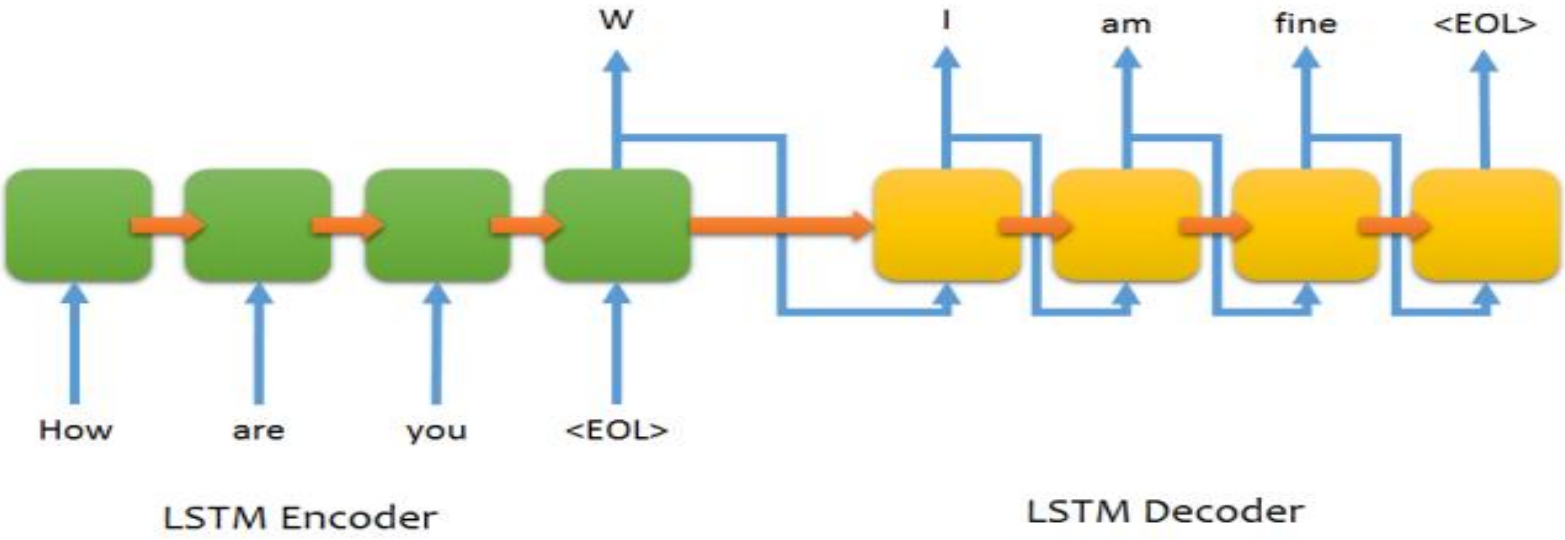
主动发起话题

Human:	以后叫你伊娃。(I will call you Eva.)
Computer:	伊娃不是那个机器人动画片？ (Isn't Eva a robot cartoon?)
Human:	你会聊天吗？(Can you talk?)
Computer:	不是聊的好好的吗？ (Aren't we talking now?)
Human:	啊...(Errr...)
Computer:	我不是瓦力，我没有伊娃(I'm not WALL-E and I don't have Eva.)

Remark: WALL-E and Eva are two robots in the movie WALL-E.

- 聊天机器人简介
- 基于模板的方法
- 基于检索的方法
- 基于深度学习的方法

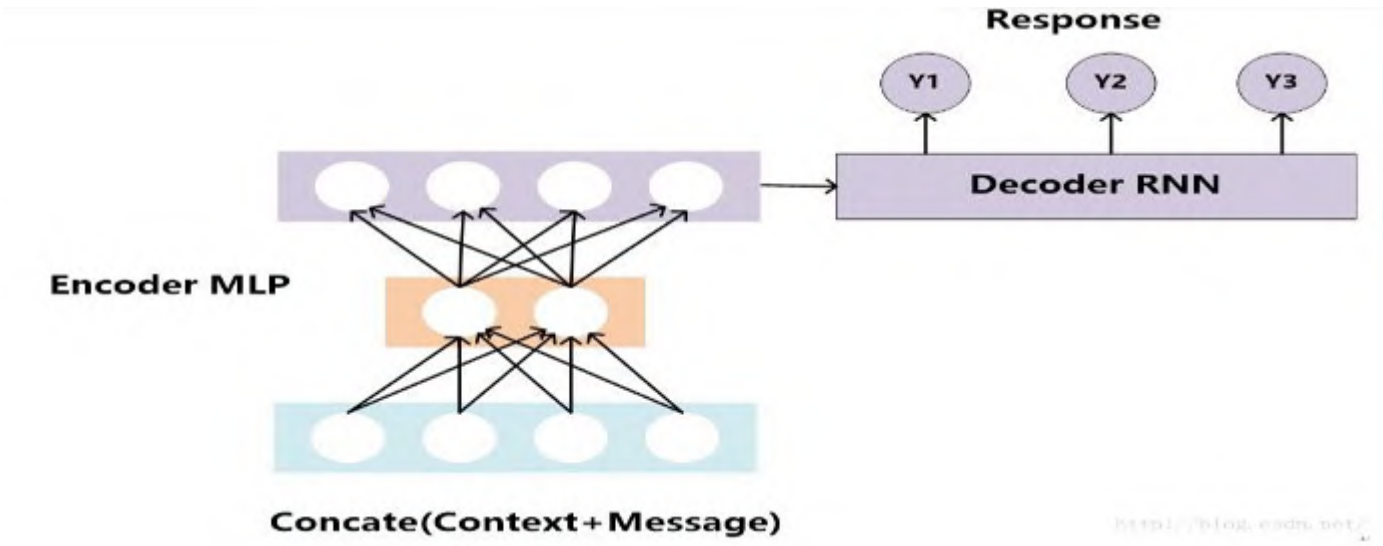
- 语言模型
  - N-Gram模型，参考前N-1个字符生成下一个字符
  - RNN模型，通过循环反馈结构生成下一字符
- 长距离依赖
  - 在N-gram中，N越大代表语言模型越精准
  - 在RNN模型中，没有N的限制，模型天然带有长距离依赖的特性
- LSTM
  - 但现实中RNN跟其他深度网络一样，有梯度消失的问题。表现在，训练好的RNN会有健忘症，记不住很久以前发生的事
  - 解决RNN梯度消失的主流方法是用LSTM替换掉RNN的隐层节点，都是有利用特有的”门”机制，把梯度沿着时间轴以一定概率直接传回去，以此来减轻梯度消失



- **绿色LSTM(Encoder)逐词或字符循环读入，并最终得到隐层状态w，也就是thought vector；**
- **黄色LSTM(Decoder)得到thought vector后逐词或字生成应答**
- **训练语料：聊天句对**

Neural Responding Machine for Short-Text Conversation (2015-03)
A Neural Conversational Model (2015-06)

- Context和当前query合并输入encoder
- 通过某种方式对会话session进行embedding，作为decoder的输入
- 通过某种方式引入会话session的topic，作为decoder的输入

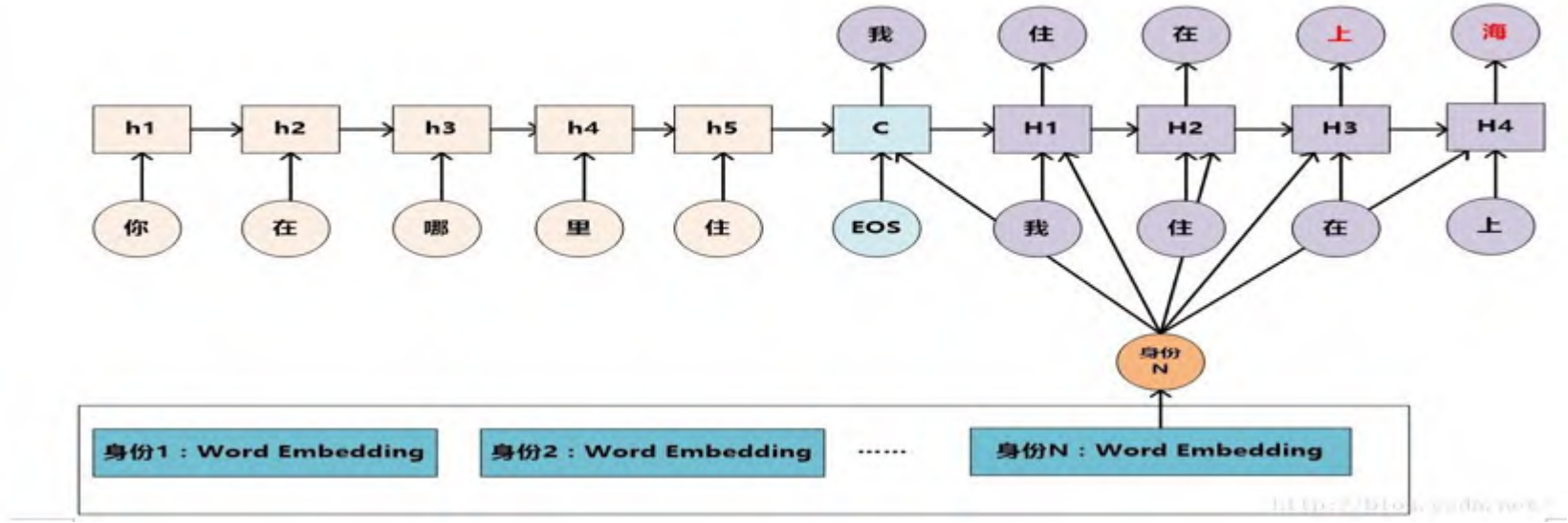


A Neural Network Approach to Context-Sensitive Generation of Conversational Responses (2015-06)
Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models (2015-07)
Attention with Intention for a Neural Network Conversation Model (2015-10)

- 把聊天助手的个性信息导入到Decoder的输出过程中

message	Where do you live now?
response	I live in Los Angeles.
message	In which city do you live now?
response	I live in Madrid.
message	In which country do you live now?
response	England, you?

message	Where is your hometown?
response	I'm from England.
message	Where are you from?
response	I'm from England.
message	In which city do you live now?
response	I live in London.
message	In which country do you live now?
response	I live in England.



- A Persona-Based Neural Conversation Model (2016-03)

- 构建QA问答对和三元组的对齐语料
- 在decode时，用一个逻辑回归控制从语言部分还是知识部分生成next token，产生用自然语言描述的FAQ回复

Figure 4: Examples of the generated answers by GENQA.

ID	Questions	Generated Answers
1	三国演义谁写的? Who wrote the Romance of the Three Kingdoms?	明朝的罗贯中。 Luo Guanzhong in Ming dynasty.
2	项羽是哪国人? Which state does Xiang Yu belong to?	楚国的皇帝。 King of the Chu State.
3	还珠格格多少集? How many episodes does My Fair Princess have?	一共24集。 24 episodes in total.
4	泰戈尔是哪国作家? Which country is the writer Tagore from?	泰戈尔是印度著名的作家 。 Tagore is a famous Indian writer.
5	孙燕姿今年多大啊? How old is Stefanie Sun this year?	32岁，他出生于1978年7月23日。 Thirty-two, he was born on July 23, 1978.
6	《格列佛游记》的作者是谁? Who is the author of Gulliver's Travels?	法国作家斯威夫特。 The French writer, Swift.
7	怪物史莱克4什么时候上映呀? When will Shrek Forever After be released?	上映日期：梦工厂电影发行公司。 Release date: DreamWorks Pictures.
8	凤凰卫视是哪里的电视台? Which place is phoenix television located at?	总部在刘长乐。 The headquarters is at Liu Changle.

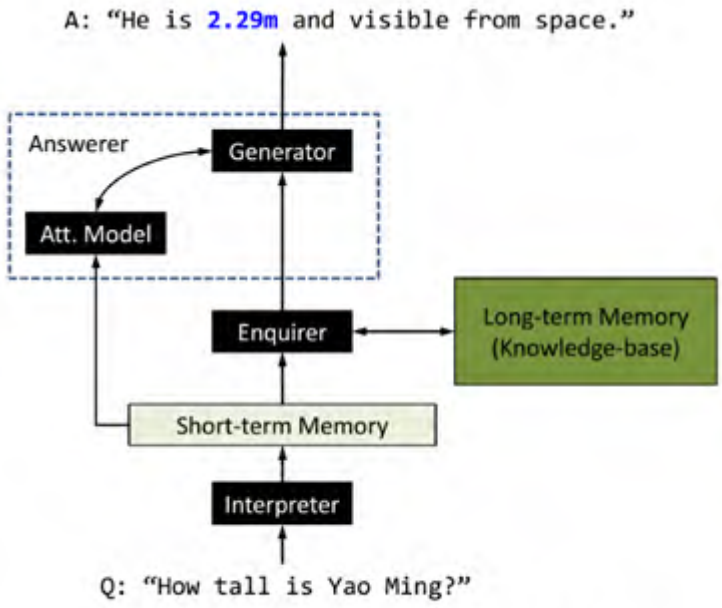


Figure 1: The diagram for GENQA.

Neural Generative Question Answering （2016-06）

	基于模板	基于检索	基于深度学习
聊天语料库	问答模板	问答句对	问答句对
人工工作量	高	中等	低
语法准确性	是	是	不能保证
应答可控性	高	中等	低
可扩展性	差	中等	好
技术难度	容易	中等	复杂
适用场景	高频聊天请求聊天 特定聊天场景	中等频度聊天请求 垂直领域的聊天	长尾聊天请求

实际系统可以是三种方法的融合

- 张俊林，使用深度学习打造智能聊天机器人， 2016.7.13
- Denny Britz, Deep Learning for chatbots, part 1 – Introduction
- Denny Britz, Deep Learning for chatbots, part 2 – Implementing a retrieval-based model in TensorFlow
- 微信公众号：PaperWeekly, 机器之心
- 开发工具： Google TensorFlow
- 数据参考： A Survey of Available Corpora For Building Data-Driven Dialogue Systems

MDCC 2016

中国移动开发者大会

Mobile Developer Conference China 2016

云知声

Unisound

智享未来

mdcc.csdn.net