



数据科学与计算机学院
School of Data and Computer Science

自然语言处理

Natural Language Processing

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

问答与对话（下）

本节内容

◆ 阅读理解式问答系统

◆ 对话系统

阅读理解式问答

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

Document

1) What is the name of the trouble making turtle?

- A) Fries
- B) Pudding
- C) James
- D) Jane

2) What did James pull off of the shelves in the grocery store?

- A) pudding
- B) fries
- C) food
- D) splinters

3) Where did James go after he went to the grocery store?

- A) his deck
- B) his freezer
- C) a fast food restaurant
- D) his room

4) What did James do after he ordered the fries?

- A) went to the grocery store
- B) went home without paying
- C) ate them
- D) made up his mind to be a better turtle

Question

Candidate Answer

阅读理解式问答

- ❖ 给出的信息来源只有一篇相关文档
- ❖ 问题答案候选已经给出，一般由几个选项构成
- ❖ 问题形式多种多样，主要考察语义理解和推理

数据集：MCTest (EMNLP 2013)

□ 早期阅读理解任务的一个典型的数据集

- 文档：660个短故事
- 问题：2640个人工提出的问题（每个文档对应4个问题）
- 答案：四选一的选择题形式

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

Q: Where did James go after he went to the grocery store?

(A) his deck (B) his freezer (C) a fast food restaurant (D) his room

□ 问题被限定在7岁儿童可回答的范围，考察该层次机器的推理能力

- 优点：问题难度较大，包含很多常识性问题
- 缺点：数据规模非常小，深度学习模型难以应用

数据集： SQuAD (EMNLP 2016)

□ 近两年来最经典且最被广泛关注的阅读理解数据集

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

Document

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

数据集： SQuAD (EMNLP 2016)

□ 近两年来最经典且最被广泛关注的阅读理解数据集

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

Question



数据集： SQuAD (EMNLP 2016)

□ 近两年来最经典且最被广泛关注的阅读理解数据集

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

Answer

数据集： SQuAD (EMNLP 2016)

□ 近两年来最经典且最被广泛关注的阅读理解数据集

- 文档：从536篇Wikipedia的文章中抽取2万多个段落
- 问题：基于文档，人工生成的10万个问题
- 答案：原文中的一个区间（一个词或几个词组成）

□ SQuAD引领了近两年阅读理解任务的发展

- 优点：数据规模较大，适用于深度学习方法;并且数据质量较高
- 缺点：
 - 段落文档均出自536篇Wikipedia文章，词汇和表达的多样性不足
 - 问题被限定为必须能被原文区间回答，导致可提问的角度受限

数据集： MS MACRO (NIPS2016)

□ 代表了阅读理解任务向开放域发展的新趋势

- 文档：100万篇由搜索引擎搜索得到的文章
- 问题：10万个Bing搜索中出现的真实问题
- 答案：根据文档，人工总结得到(不一定是原文区间)

数据集： MS MACRO (NIPS2016)

□ 所谓“开放域”

	给定	要求
SQuAD等任务	问题，文档	答案
开放域	问题	答案

□ 数据集特点

- 更加贴近真实的问答场景，即只有问题，事先没有准备好的文档；该数据集的文档是根据已有的问题去搜索并整理出文档，而不是根据文档提出问题

数据集：DuReader (2017,Baidu)

□ DuReader数据集是一个比较有代表性的中文数据集

- 文档：100万篇由搜索引擎搜索得到的文章
- 问题：20万个Baidu搜索中出现的真实问题
- 答案：根据文档，人工总结得到

数据集：DuReader (2017,Baidu)

□ DuReader形式上与MS_MARCO很相似，最大的特点是引入观点型(opinion)问题：

	Fact	Opinion
Entity	iphone哪天发布 On which day will iphone be released	2017最好看的十部电影 Top 10 movies of 2017
Description	消防车为什么是红的 Why are firetrucks red	丰田卡罗拉怎么样 How is Toyota Carola
YesNo	39.5度算高烧吗 Is 39.5 degree a high fever	学围棋能开发智力吗 Does learning to play go improve intelligence

■ 将问题按照事实性(fact)和观点型(opinion)两个角度划分，之前其他数据集只有这里的 fact类型问题，此处引入opinion型。但是，该类型问题的引入使得答案评判更难

机器阅读理解的方法

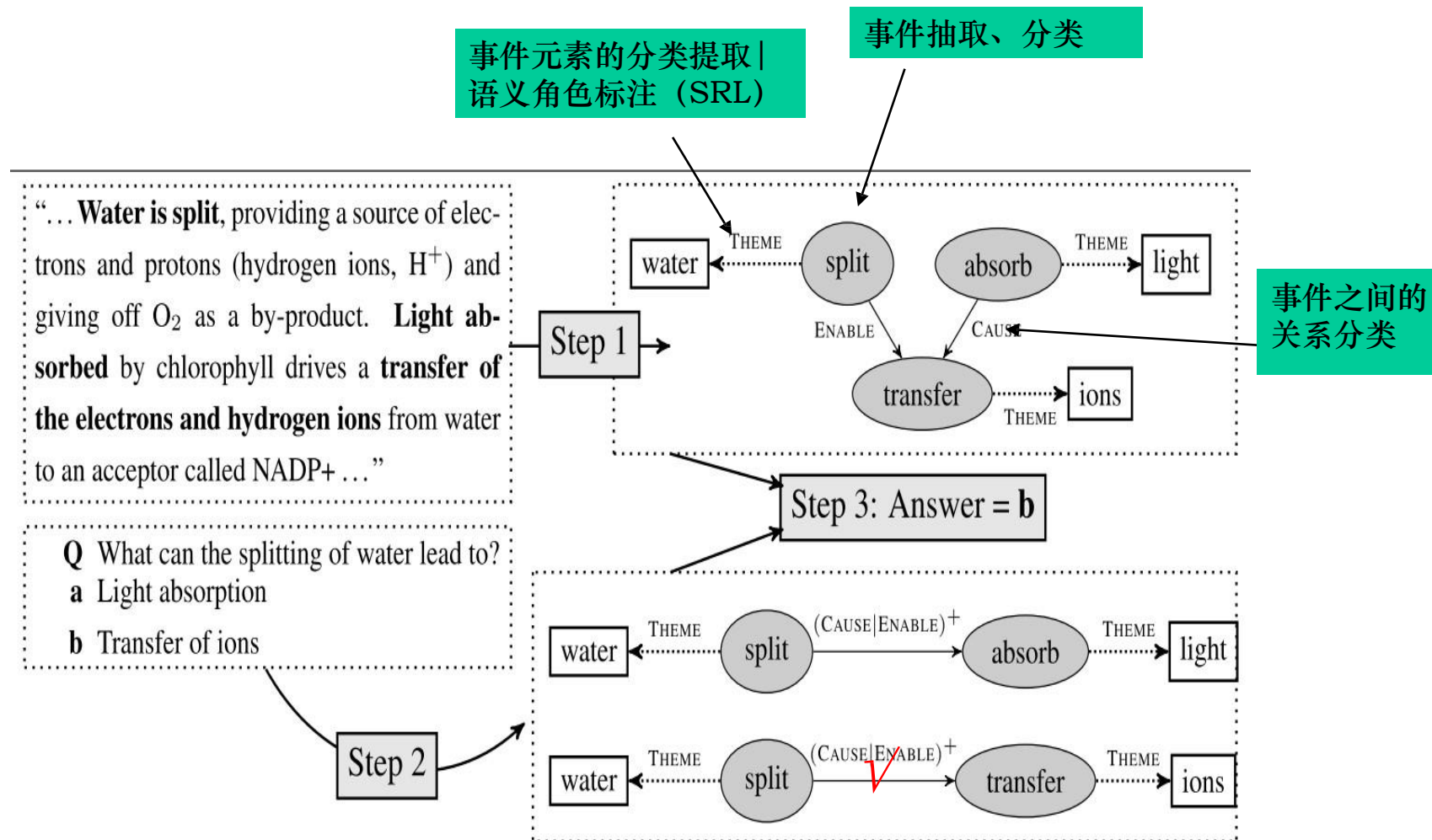
■ 传统特征工程的方法

- 文本分析
- 问句解析
- 匹配答案

■ 神经网络的方法

- 文档和问句的表示学习
- 文档和问句的匹配计算
- 深度推理机制

基于传统特征工程的方法



基于传统特征工程的方法

■ 优点

- 对过程进行建模，清晰明了，各个部分的作用可以显式地表示
- 对问题进行了同样的过程处理，最后的结果是确定的
- 语义建模方式非常明显，类似于标准的Semantic Parsing，每一部分的语义都能很直观地表示出来

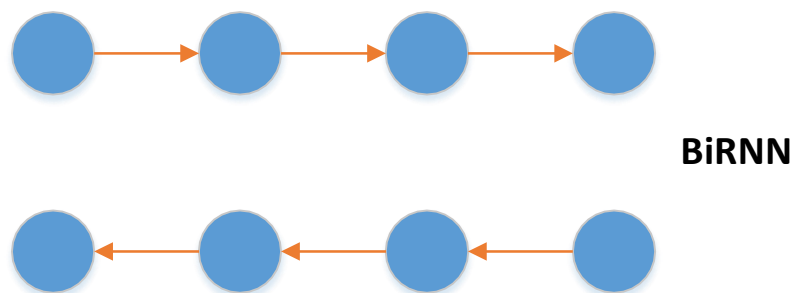
■ 缺点

- 由于是基于传统特征工程的方法，非常耗时耗力，训练样本有限
- 领域适应能力差，在这个训练集上训练的模型会有领域倾向性

基于神经网络的方法

■ 文章表示

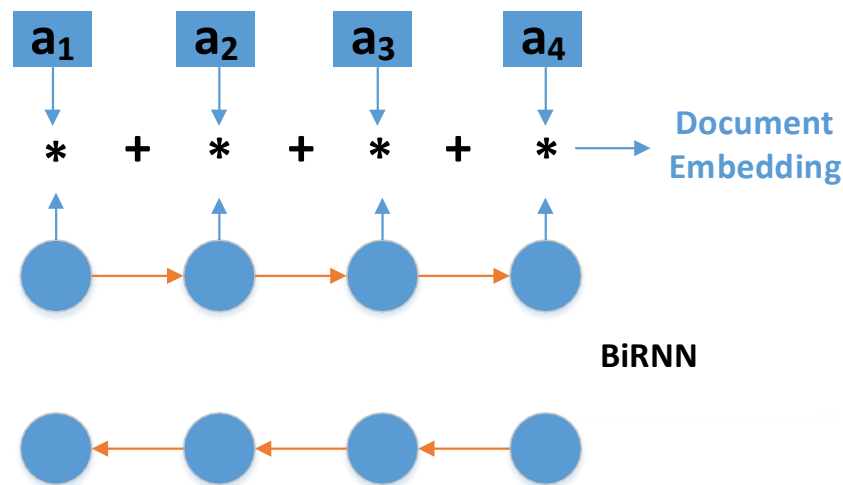
- 第一种方法：将文章看作单词序列，在这个序列上使用RNN对文章进行建模，每个单词对应RNN序列中的一个时刻 t 的输入，RNN的隐层状态是融合了当前词义和上下文语义的编码



基于神经网络的方法

■ 文章表示

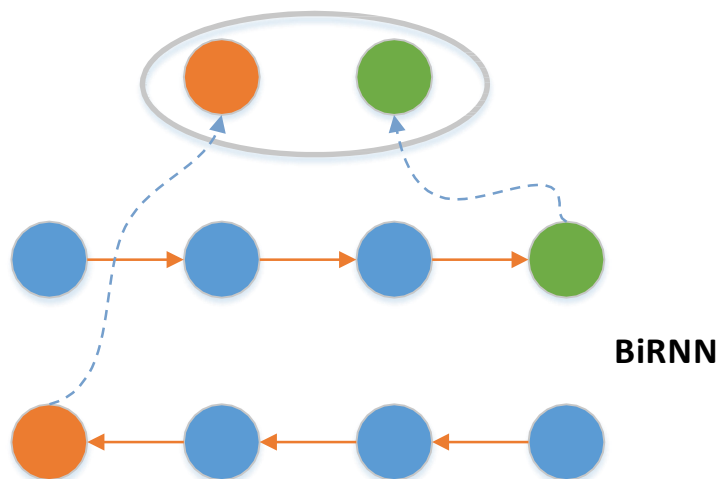
- 第二种方法：引入Attention机制，该方法也是采用双向RNN对每个单词及其上下文信息进行建模，得到隐层状态表示。不同点在于，得到隐层表示向量的每一维都要乘以某个系数，该系数代表该单词对于整个文章语义表达的重要程度。



基于神经网络的方法

■ 问句表示:

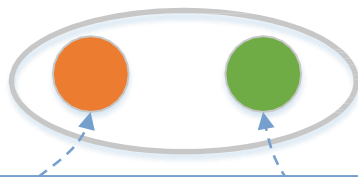
- 有与文章表示方法类似的两种建模方法
- 还有另外一种建模方法：首先使用双向RNN对其进行表示。将这双向RNN首尾词节点的隐层状态拼接起来，就得到了整个句子的最终表示



基于神经网络的方法

■ 问句表示:

- 有与文章表示方法类似的两种建模方法
- 还有另外一种建模方法：首先使用双向RNN对其进行表示。将这双向RNN首尾词节点的隐层状态拼接起来，就得到了整个句子的最终表示

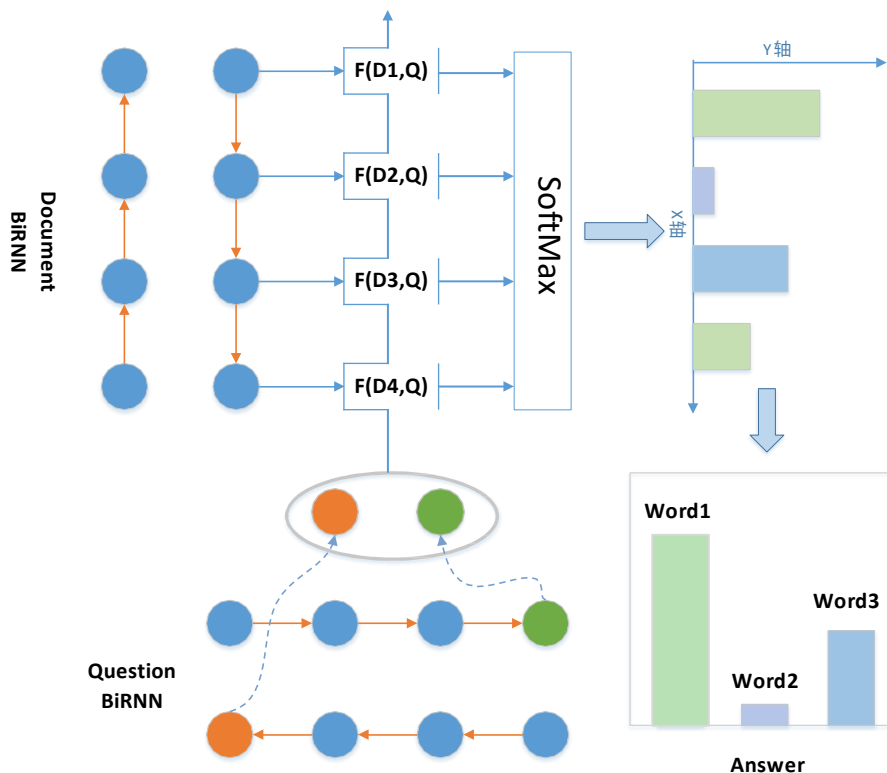


其中，正向RNN的尾部单词隐层节点（图中绿色节点）正向融合了整个句子语义信息；相对的，反向RNN的尾部单词（局首词）隐层节点（图中橙色节点）则逆向融合了整个句子的语义信息。

基于神经网络的方法

文章与问句的匹配：一维匹配模型

- 双向RNN对文章和问题表示后，通过某种匹配函数来计算文章中每个单词 D_i 和问题 Q 的语义信息的匹配程度。
- 之后，对每个单词的匹配程度通过 SoftMax 函数进行归一化，这样就形成了一种 Attention 的操作，将更可能是问题答案的单词凸显出来。



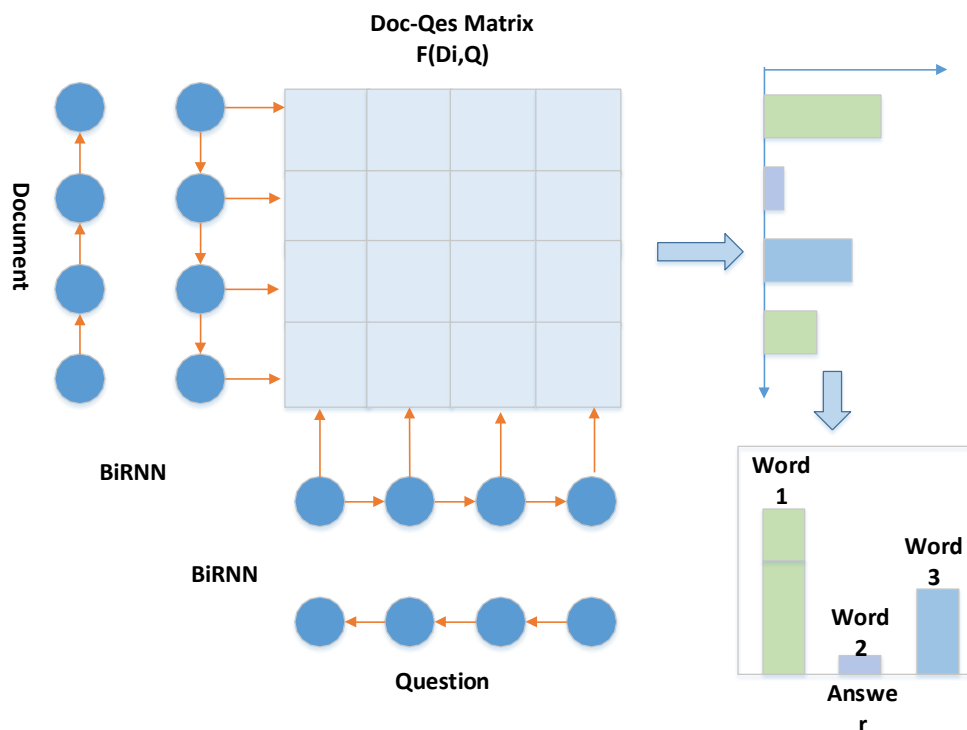
基于神经网络的方法

■ 文章与问句的匹配：二维匹配模型

- ❖ 其整体结构和一维匹配模型是类似的，最主要的区别体现在如何计算文章和问题的匹配程度上
- ❖ 一维匹配模型在匹配过程中，形成一维线性结构模型；而二维匹配模型在进行问题和文章的匹配时，形成二维矩阵结构
- ❖ 二维匹配模型使得问题 Q 中的每个词都可以和文档 D 中的词进行交互，粒度更细。

基于神经网络的方法

■ 文章与问句的匹配：二维匹配模型



基于神经网络的方法

可视化:

- 通过可视化的方法可以观察到模型在推理过程中哪些词被重点关注了，一定程度上为模型提供可解释性:

<p>by <i>ent423</i> ,<i>ent261</i> correspondent updated 9:49 pm et ,thu march 19,2015 (<i>ent261</i>) a <i>ent114</i> was killed in a parachute accident in <i>ent45</i> ,<i>ent85</i> ,near <i>ent312</i> ,a <i>ent119</i> official told <i>ent261</i> on wednesday .he was identified thursday as special warfare operator 3rd class <i>ent23</i> ,29 ,of <i>ent187</i> ,<i>ent265</i> .` <i>ent23</i> distinguished himself consistently throughout his career .he was the epitome of the quiet professional in all facets of his life ,and he leaves an inspiring legacy of natural tenacity and focused</p> <p>...</p>	<p>by <i>ent270</i> ,<i>ent223</i> updated 9:35 am et ,mon march 2 ,2015 (<i>ent223</i>) <i>ent63</i> went familial for fall at its fashion show in <i>ent231</i> on sunday ,dedicating its collection to `` mamma " with nary a pair of `` mom jeans " in sight .<i>ent164</i> and <i>ent21</i> , who are behind the <i>ent196</i> brand ,sent models down the runway in decidedly feminine dresses and skirts adorned with roses ,lace and even embroidered doodles by the designers ' own nieces and nephews .many of the looks featured saccharine needlework phrases like `` i love you ,</p> <p>...</p>
<p><i>ent119</i> identifies deceased sailor as X ,who leaves behind a wife</p>	<p>X dedicated their fall fashion show to moms</p>

Teaching machines to read and comprehend, NIPS 2015

基于神经网络的方法

■ 插入对抗负样本:

- Adversarial Examples for Evaluating Reading Comprehension Systems (Percy Liang, EMNLP2017)

Article: Super Bowl 50

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?”

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

- 原本模型能够得到正确答案。在文档后面插入一句和包含答案的线索句子形式上相似的负样本句子(图中蓝色部分)。结果，模型受到干扰，得出错误答案。
- 然而，这种负样本句子并不能对人的判断造成影响。

基于神经网络的方法

■ 插入对抗负样本：

- Adversarial Examples for Evaluating Reading Comprehension Systems (Percy Liang, EMNLP2017)

- 测试16个之前在SQuAD数据集上表现良好的模型
- 在文章中插入负样本后，模型的性能都大幅下降
- 这篇工作表明：当前阅读理解模型只具有识别浅层模式的能力，而并没有真正的语言理解能力。因此要达到真正理解语言的目标，研究者们还有很长的路要走。

Model	Original	ADDSSENT
ReasoNet-E	81.1	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	46.2
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	46.6
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.3	37.8
Logistic	50.4	23.2

本节内容

◆ 阅读理解式问答系统

◆ **对话系统**

展示 (一)

展示 (二)



对话系统 (AKA)

- Conversational Agents
- Dialogue Systems
- Human conversation
- Spoken Language Systems
- Speech Dialogue Systems

.....

对话系统 (AKA)

- 旅行安排
- 电话客服
- 智能辅导
- 机器人控制
- 个人助理
-

Conversation : Booking a flight

Bonjour. J'appelle
pour réserver un vol de
Paris à Berlin.

Hello. I'm calling to
book a flight from
Paris to Berlin.



ATIS task (1997) - 订机票

Example #1:

(a) show me the flights from boston to philly

$\lambda x. flight(x) \wedge from(x, bos) \wedge to(x, phi)$

(b) show me the ones that leave in the morning

$\lambda x. flight(x) \wedge from(x, bos) \wedge to(x, phi)$
 $\wedge during(x, morning)$

(c) what kind of plane is used on these flights

$\lambda y. \exists x. flight(x) \wedge from(x, bos) \wedge to(x, phi)$
 $\wedge during(x, morning) \wedge aircraft(x) = y$

ATIS task (1997) - 订机票, ...

Example #2:

(a) show me flights from milwaukee to orlando

$\lambda x. flight(x) \wedge from(x, mil) \wedge to(x, orl)$

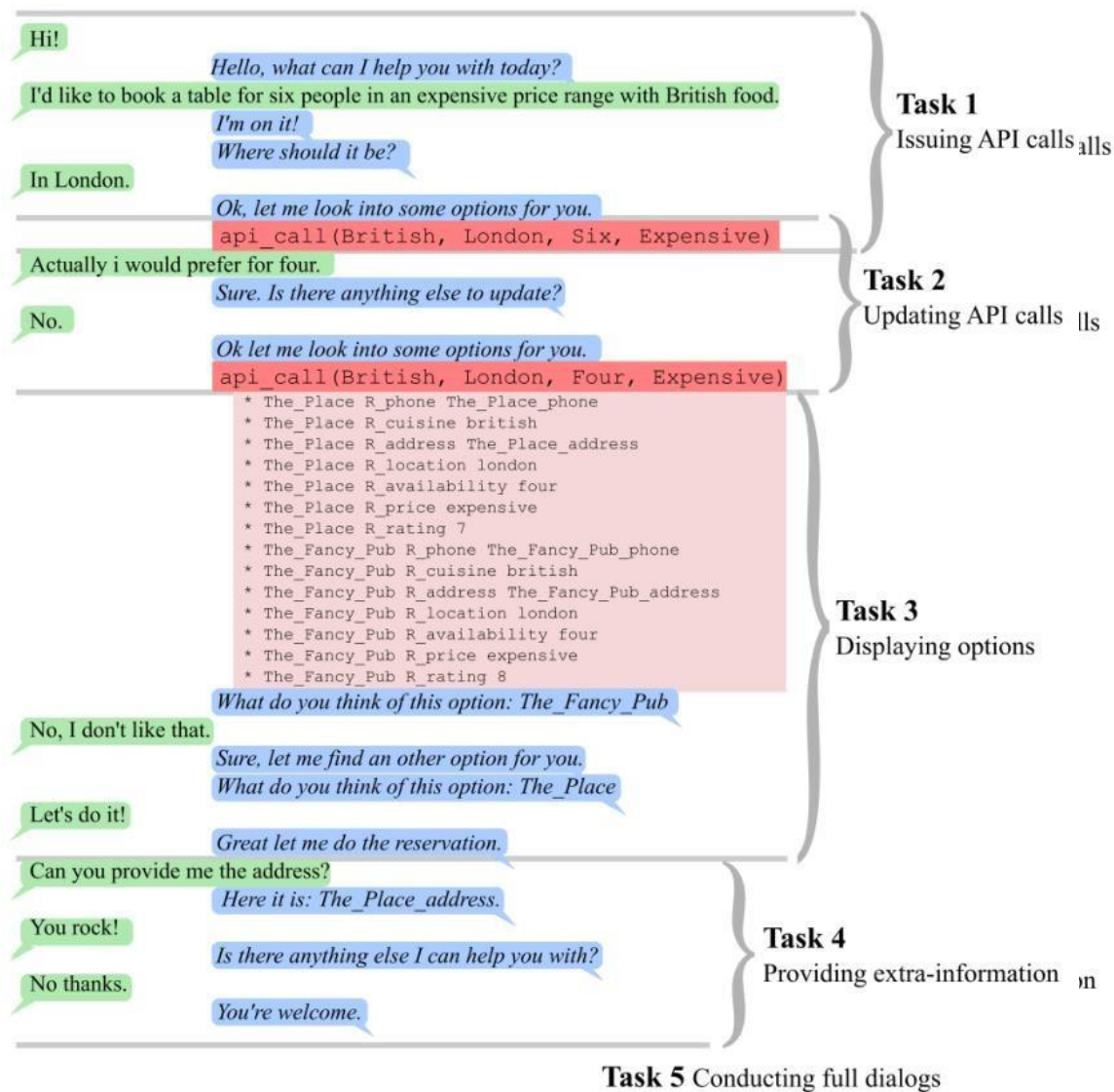
(b) cheapest

$argmin(\lambda x. flight(x) \wedge from(x, mil) \wedge to(x, orl),$
 $\lambda y. fare(y))$

(c) departing wednesday after 5 o'clock

$argmin(\lambda x. flight(x) \wedge from(x, mil) \wedge to(x, orl)$
 $\wedge day(x, wed) \wedge depart(x) > 1700,$
 $\lambda y. fare(y))$

Restaurant booking (2016)



Chatbot – 聊天机器人

<p>A: Where are you going ? (1) B: I 'm going to the restroom . (2) A: See you later . (3) B: See you later . (4) A: See you later . (5) B: See you later . (6)</p>	<p>A: Where are you going ? (1) B: I'm going to the police station . (2) A: I'll come with you . (3) B: no , no , no , no , you ?re not going anywhere . (4) A: Why? (5) B: I need you to stay here. (6) A: I don't know what you are talking about. (7) B: You don?t know what you are saying. (8) ...</p>
<p>A: How old are you ? (1) B: I'm 16 . (2) A: 16 ? (3) B: I don't know what you are talking about . (4) A: You don't know what you are saying. (5) B: I don't know what you are talking about . (6) A: You don't know what you are saying. (7) ...</p>	<p>A: How old are you ? (1) B: I'm 16. Why are you asking ? (2) A: I thought you were 12. (3) B: What made you think so ? (4) A: I don't know what you are talking about. (5) B: You don't know what you are saying. (6)</p>

Personal Assistants - 个人助理

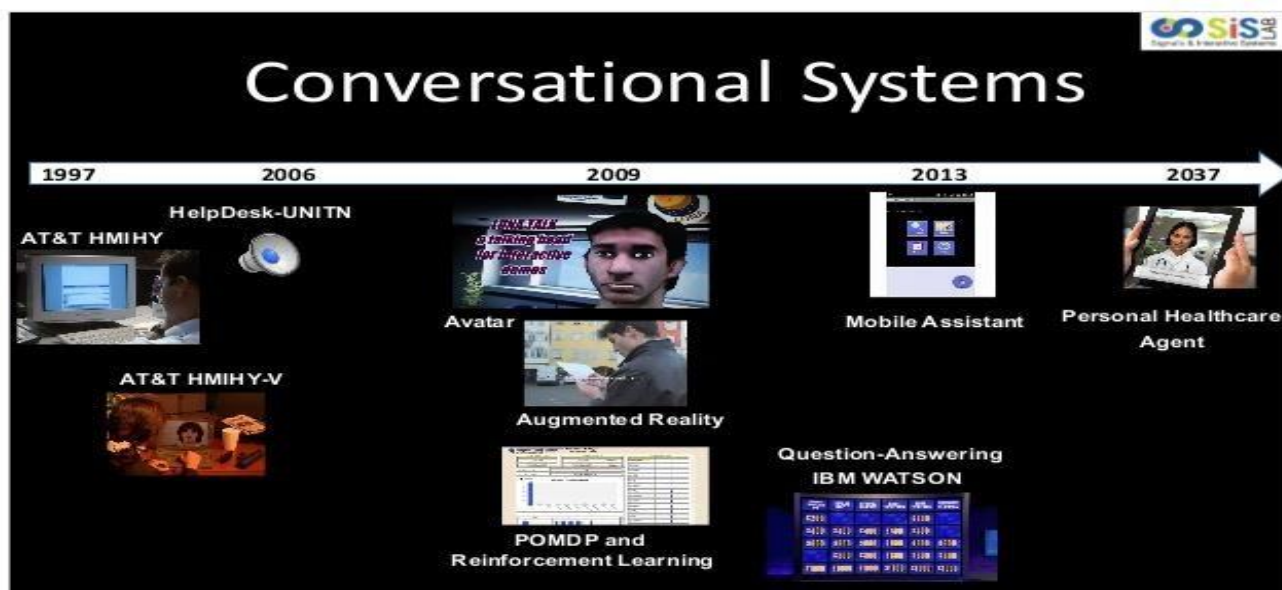


Personal Assistants - 个人助理



对话系统发展历史及其展望

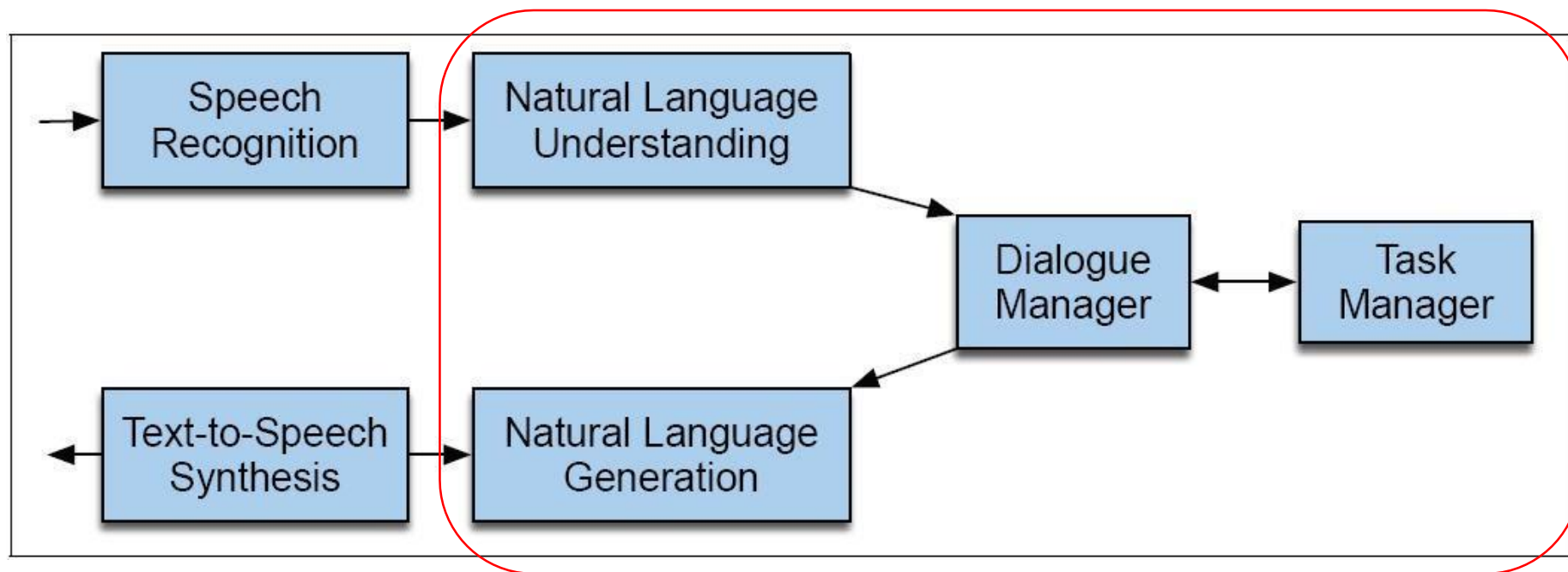
- 90年代末：聊天性质，没有实质内容，基本采用人工书写对话模板的方法
- 2009年开始：逐步开始用统计机器学习的方法进行对话管理
- 2013年开始：智能终端上的个人助理逐步支持智能对话
- 有人预计到2037年：将有类似虚拟医疗这样的更实用的对话系统问世



对话系统

- 对话系统基础
- 对话管理/状态跟踪
- 聊天机器人
- 对话系统的评价

基本架构



- 语音识别 (Speech recognition)
- 自然语言理解 (Natural language understanding)
- 对话管理 (Dialogue management)
- 自然语言生成 (Natural language generation)
- 语音合成 (Speech synthesis)

分类：Initiative (谁发起对话任务?)

■ 系统主导 (System Initiative)

- 电话客服
- Booking Tickets

■ 用户主导 (User Initiative)

- Question Answering System
- Personal Assistants

■ 混合模式 (Mixed Initiative)

- Human-human Conversation
- Chatbot

对话系统

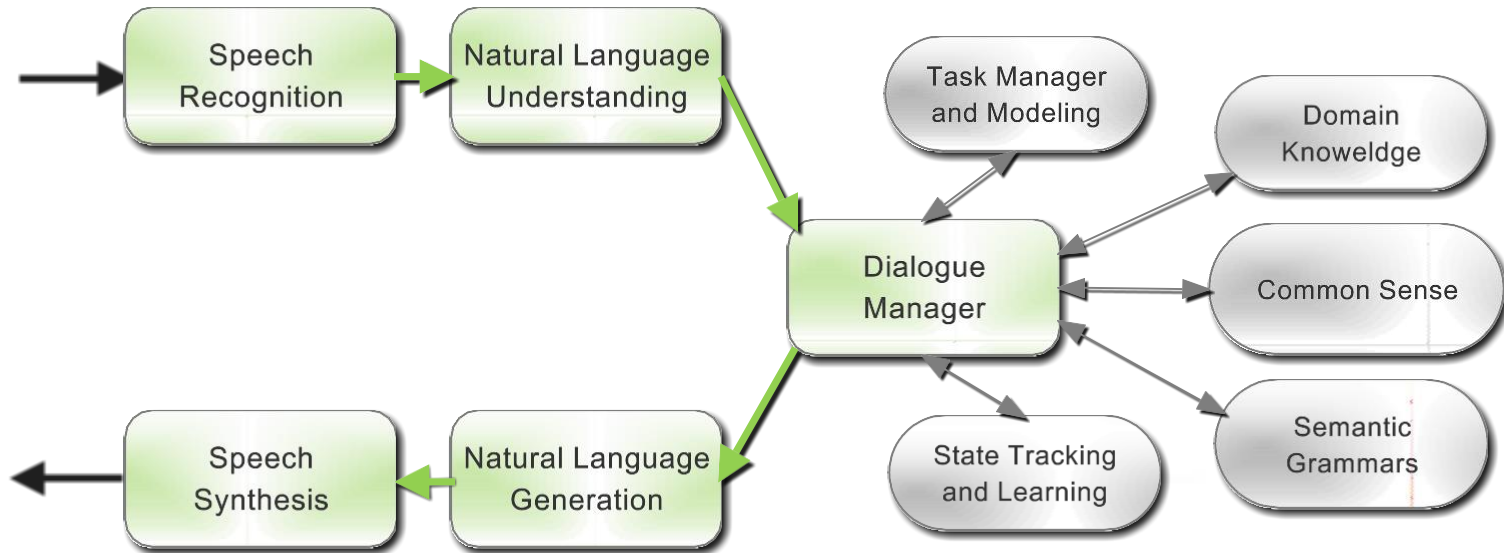
- 对话系统基础
- 对话管理/状态跟踪
- 聊天机器人
- 对话系统的评价

对话管理

■ 对话管理是对话系统的核心模块

- 任务的管理和建模
- 状态跟踪和学习

■ 核心和难点：状态跟踪和学习



对话管理（状态跟踪和学习）方法

- 有限状态机 (Finite State)
- 基于框架的方法 (Frame-based)
- 统计方法: Information State (Markov Decision Process)

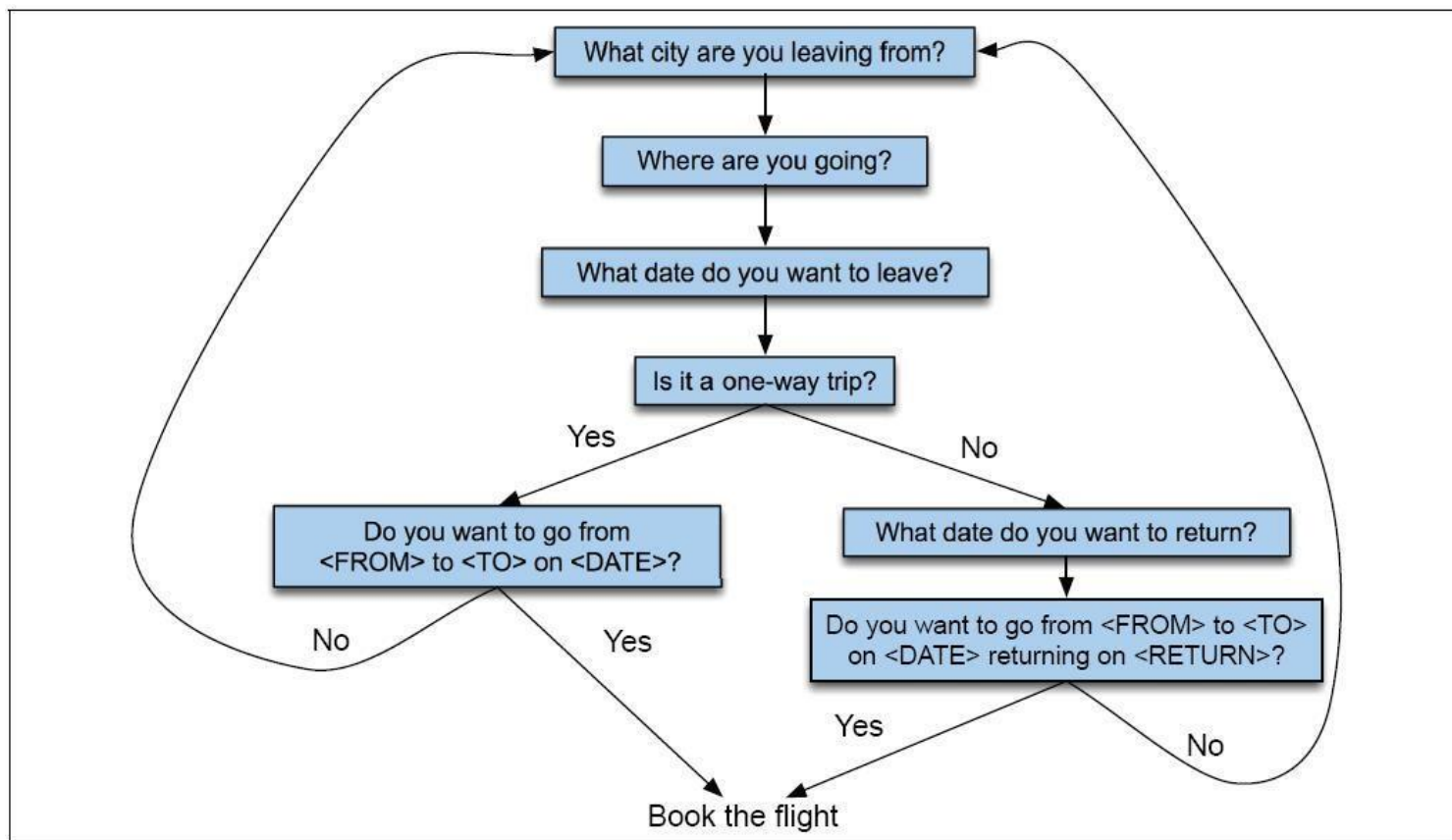
基于有限状态机的方法

□ 考虑一个订票系统，它有如下可能的状态：

- Ask the user for a departure city
- Ask for a destination city
- Ask for a time
- Ask whether the trip is round-trip or not

基于有限状态机的方法

- 一个具体任务有哪些状态，是由专家给定的
- 状态之间如何转换，通过有限状态机进行建模



基于有限状态机的方法

- 系统需要完全控制与用户交互的过程
- 系统需要询问用户一系列问题
- 用户可能一次输入多个信息（对应多个状态），但是有限状态机不能一次接受多个状态

太受限了！

解决办法：使用对话目标框架（如机票信息）指导对话过程

框架的例子

FLIGHT FRAME:

ORIGIN:

CITY: Boston

DATE: Tuesday

TIME: morning

DEST:

CITY: San Francisco

AIRLINE:

...

基于框架的对话管理

■ 使用框架的结构指导对话过程

- 机器根据框架进行提问，人也根据框架进行回复

■ 问答过程就是一个槽-值填充的过程

- 当所有槽的值都填满了，则可以信息系统查询

■ 用户可以一次回答多个系统问题

基于框架的对话管理

Slot

Question

ORIGIN

What city are you leaving from?

DEST

Where are you going?

DEPT DATE

What day would you like to leave?

DEPT TIME

What time would you like to leave?

AIRLINE

What is your preferred airline?

以上两种方法的不足

- 需要专家设计并编写对话方案，系统设计、开发和维护成本高
- 不适合建模不确定性的对话管理过程

基于机器学习的对话管理系统

- 基本思想：利用统计框架从大量的对话语料中自动学习对话管理模型。这种方式有两个主要的优点：
 - 可以将不确定性表示引入到模型中，相对基于规则的系统，其对语音识别和语义理解的噪音有更好的鲁棒性
 - 这种框架具有自动学习功能，可以极大的降低人工开发成本
- 基于机器学习的对话管理系统，典型的代表是基于马尔可夫决策过程的对话管理

不确定性对话过程的建模

■ 需要考虑三方面的问题

- 系统当前的状态
- 在当前状态下系统可以采取什么样的动作
- 系统采取这样动作是要完成什么样的目标

一般来讲，这些问题可以用马尔科夫决策过程进行建模

对话系统

- 对话系统基础
- 对话管理/ 状态跟踪
- 聊天机器人
- 对话系统的评价

聊天机器人：基于检索的方法

- 预定义一个post-response库，当需要作出反应时，依靠排序的方法从候选反应库中选择一个最合适的
- 通常使用各种匹配特征的线性或非线性组合方法，为当前给定的post在库中寻找一个最相近的post，并将它的response作为当前post的回应。

对话者	计算机反应
招生需要什么标准？	一般男生的标准
为什么要这么标准的要求？	毕竟不标准就没人来了

检索方法：流程

■ 对于一个给定的查询

- 在对话库中选择多个相似的候选问答对
- 对于原始查询提取特征，对候选答案对中的问题和答案提取特征
- 根据特征，学习排序模型，把Top-1对应的答案返回用户

■ 需要人工构建问答库，检索方法作重要的就问答库的设计、问答对的数量和覆盖度

检索方法：常用相似度

- 字符串相似度：字符重叠的个数，最大重叠字符串的长度
- TF-IDF向量的cos相似度
- 基于翻译模型的相似度
- 基于神经网络匹配的相似度
- 基于主题词的相似度

检索式方法：优缺点

■ 优点

- 预定义的反应，没有显著的语法错误

■ 主要问题

- 很难根据特定的对话场景和要求作出相应的变化
- 这些预定义的反应通常会有大量重复使用的情况，且比较生硬
- 使用特征匹配的排序方法，其特征工程非常耗时，并且通常不能区分正面和负面的回应

聊天机器人：基于生成方法

- 不依赖于预定义的反应库，重新开始生成新的反应。
- 生成式的模型依赖于机器翻译的技术，但是它并不是从一种语言翻译到另外一种语言，而是将一个输入（post）“翻译”为一个输出（response）。这种生成式的方法比基于检索的方法更难。

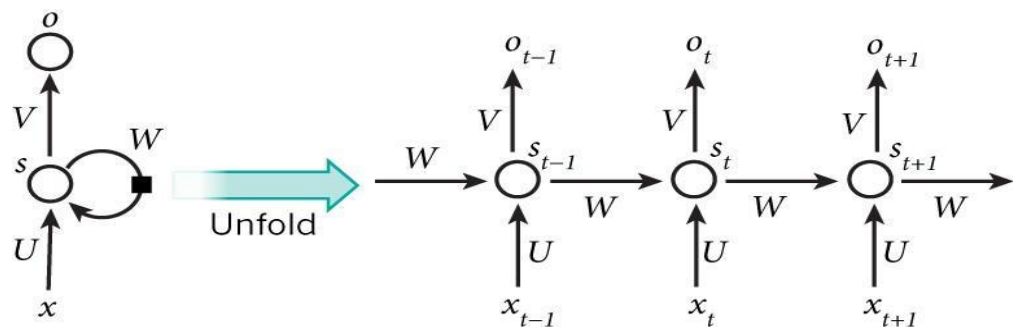
post	response
What is the purpose of life?	To serve the greater good
What year were you born?	1977
Are you a follower or a leader?	I'm a leader.

生成方法：RNN的基本计算单元

- 循环过程：每次根据神经网络内部的状态去预测一个输出，也就是一个词，预测完之后神经网络内部的状态会改变，在根据这个状态预测下一个词
- 模型训练：在对话库上训练的

生成方法：RNN的基本计算单元

1. 普通RNN 计算单元



$$s_t = f(Ux_t + Ws_{t-1})$$

$$o_t = \text{soft max}(Vs_t)$$

生成方法：RNN的基本计算单元

2. LSTM计算单元

$$i = \sigma(x_t U^i + s_{t-1} W^i)$$

$$f = \sigma(x_t U^f + s_{t-1} W^f)$$

$$o = \sigma(x_t U^o + s_{t-1} W^o)$$

$$g = \tanh(x_t U^g + s_{t-1} W^g)$$

$$c_t = c_{t-1} \circ f + g \circ i$$

$$s_t = \tanh(c_t) \circ o$$

3. GRU单元

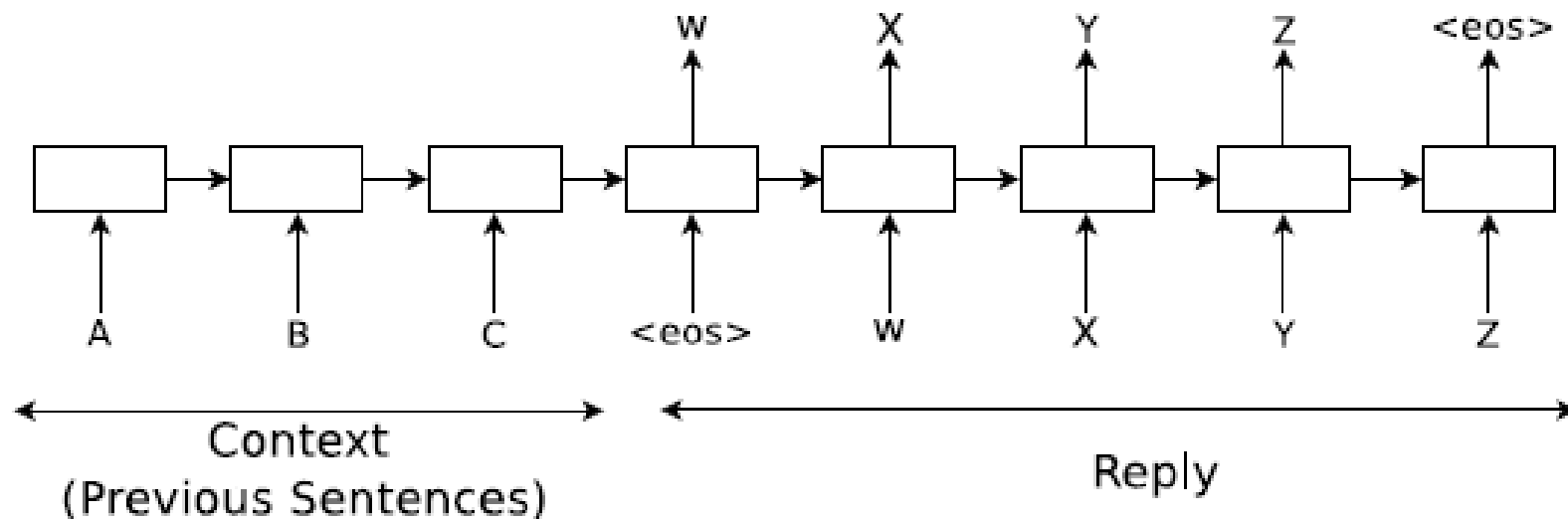
$$z = \sigma(x_t U^z + s_{t-1} W^z)$$

$$r = \sigma(x_t U^r + s_{t-1} W^r)$$

$$h = \tanh(x_t U^h + (s_{t-1} \circ r) W^h)$$

$$s_t = (1 - z) \circ h + z \circ s_{t-1}$$

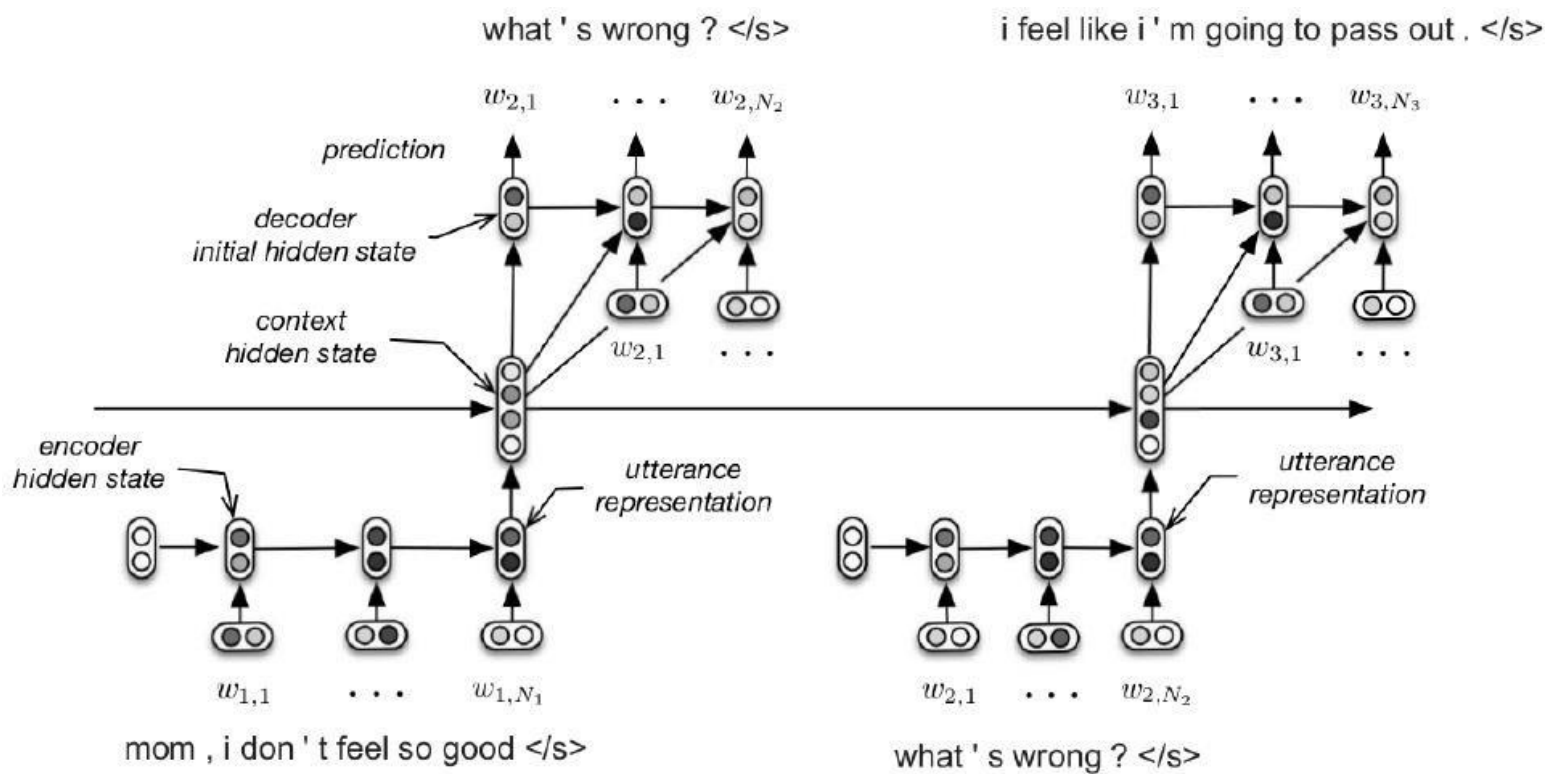
基于生成式的模型



- 从原始的句子ABC预测下一个句子WXYZ
- 计算的主要过程： Encoder-Decoder

基于生成式的模型

- 生成的过程考虑上下文: 该模型考虑了两个层次的LSTM结构:
单词序列的LSTM和句子级别的LSTM



生成式方法：优缺点

■ 优点：

- 更加类人

■ 主要问题：

- 难以设计
- 由于其所有的知识都没有人工先验，需要非常大量的训练数据支撑
- 容易产生出语法错误等，尤其是对于长句子，语法错误更容易出现

对话系统

- 对话系统基础
- 对话管理/ 状态跟踪
- 聊天机器人
- 对话系统的评价

对话系统的评估方法

■ 目标

❖ 引导用户与系统更自然、更便捷地完成对话任务

✓ 完成

✓ 自然

✓ 快捷

端到端对话系统的评价

■ LM & MT alike task

- ✓ Perplexity
- ✓ BLEU
- ✓ Ranking on Candidates
- ✓ Human Evaluation
- ✓ ...

Input: What are you doing?

-0.86 I don't know.	-1.09 Get out of here.
-1.03 I don't know!	-1.09 I'm going home.
-1.06 Nothing.	-1.09 Oh my god!
-1.09 Get out of the way.	-1.10 I'm talking to you.

Input: what is your name?

-0.91 I don't know.	...
-0.92 I don't know!	-1.55 My name is Robert.
-0.92 I don't know, sir.	-1.58 My name is John.
-0.97 Oh, my god!	-1.59 My name's John.

Input: How old are you?

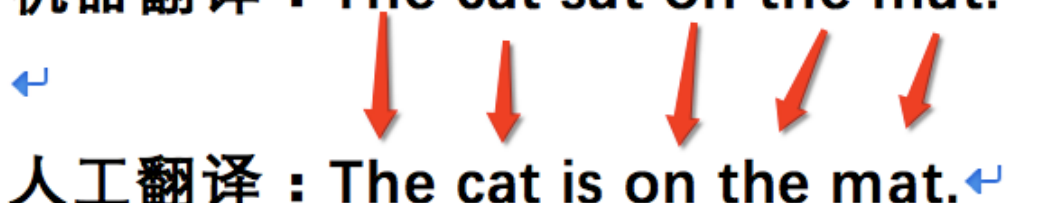
-0.79 I don't know.	...
-1.06 I'm fine.	-1.64 Twenty-five.
-1.17 I'm all right.	-1.66 Five.
-1.17 I'm not sure.	-1.71 Eight.

词共现评估

比较生成回复与参考回复在n-gram上的共现程度

1-gram:

机器翻译 : The cat sat on the mat.↵
↵
人工翻译 : The cat is on the mat.↵

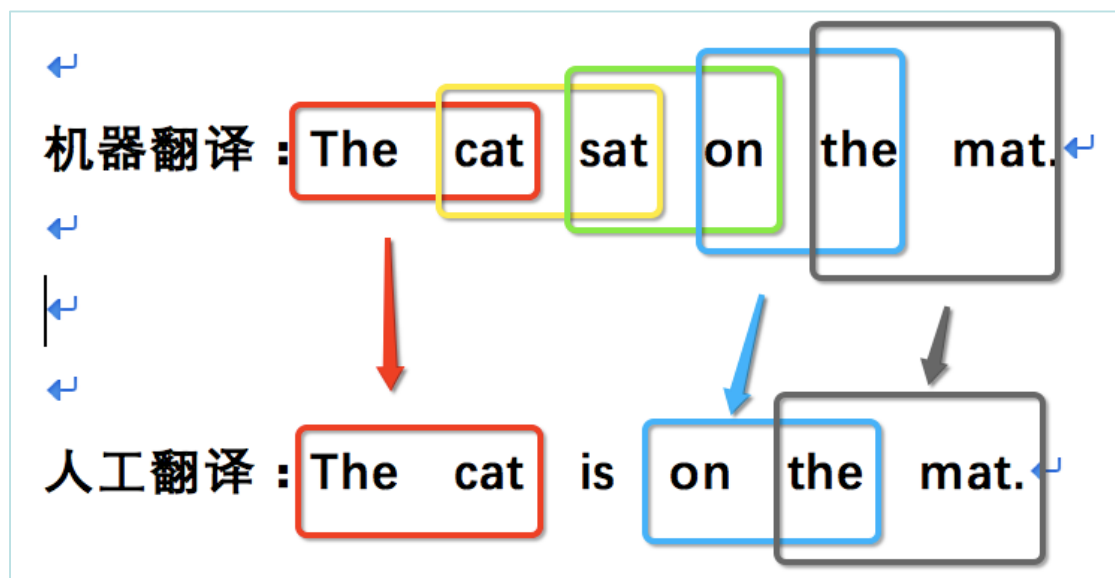


匹配度为 5/6

词共现评估

比较生成回复与参考回复在n-gram上的共现程度

2-gram:

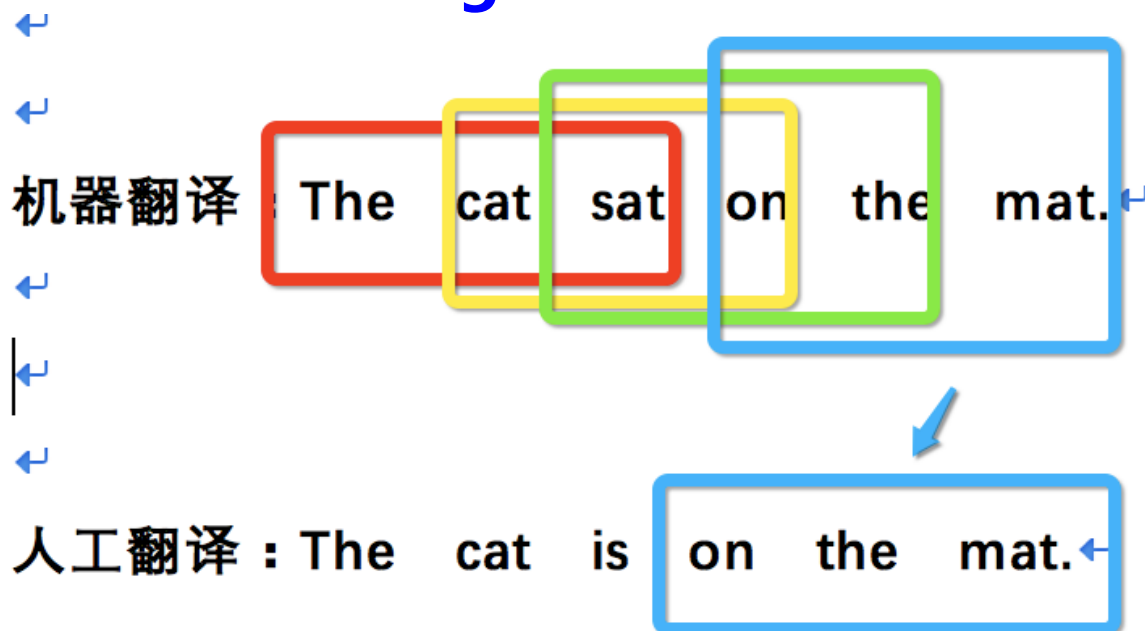


匹配度为 3/5

词共现评估

比较生成回复与参考回复在n-gram上的共现程度

3-gram:



匹配度为 1/4

词共现评估

比较生成回复与参考回复在n-gram上的共现程度

■ BLEU: Bilingual Evaluation understudy

- ✓ IBM于2002年提出
- ✓ 分析候选译文和参考译文中n元组(n-gram)共现的比例

词共现评估

比较生成回复与参考回复在n-gram上的共现程度

■ ROUGE: Recall-Oriented Understudy for Gisting Evaluation

- ✓ 2004年提出的自动摘要评估方法
- ✓ 基于摘要中n-gram的共现信息召回率来评价摘要
- ✓ ROUGE-L(Longest Common Subsequence)

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{Ref Summaries}\}} \sum_{n\text{-gram} \in S} \text{Count}_{\text{match}}(n\text{-gram})}{\sum_{S \in \{\text{Ref Summaries}\}} \sum_{n\text{-gram} \in S} \text{Count}(n\text{-gram})}$$

词向量评估

在词向量的基础上，比较生成回复与参考回复词向量的相似程度

■ 词向量贪心

- ✓ 取参考回复中的词与生成回复的词最相近的词做比较，然后求平均

■ 词向量平均

- ✓ 比较生成回复与参考回复平均词向量

词向量评估

How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, EMNLP 2016

	Twitter				Ubuntu			
Metric	Spearman	p-value	Pearson	p-value	Spearman	p-value	Pearson	p-value
Greedy	0.2119	0.034	0.1994	0.047	0.05276	0.6	0.02049	0.84
Average	0.2259	0.024	0.1971	0.049	-0.1387	0.17	-0.1631	0.10
Extrema	0.2103	0.036	0.1842	0.067	0.09243	0.36	-0.002903	0.98
METEOR	0.1887	0.06	0.1927	0.055	0.06314	0.53	0.1419	0.16
BLEU-1	0.1665	0.098	0.1288	0.2	-0.02552	0.8	0.01929	0.85
BLEU-2	0.3576	< 0.01	0.3874	< 0.01	0.03819	0.71	0.0586	0.56
BLEU-3	0.3423	< 0.01	0.1443	0.15	0.0878	0.38	0.1116	0.27
BLEU-4	0.3417	< 0.01	0.1392	0.17	0.1218	0.23	0.1132	0.26
ROUGE	0.1235	0.22	0.09714	0.34	0.05405	0.5933	0.06401	0.53
Human	0.9476	< 0.01	1.0	0.0	0.9550	< 0.01	1.0	0.0

Table 3: Correlation between each metric and human judgements for each response. Correlations shown in the human row result from randomly dividing human judges into two groups.

- 基于词共现与词向量的评估，与人工评估弱相关
- 相比之下，基于词向量的评估，与人工评估的相关性稍强一些

人工评估 (Jekaterina, EMNLP 2017)

■ 信息量(Informativeness)

- ✓ 提供有用信息

■ 自然度(Naturalness)

- ✓ 生成回复与讲母语人表达的接近程度

■ 质量(Quality)

- ✓ 句法正确、流利程度9

Thank you!

权小军 中山大学数据科学与计算机学院