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

七、问答与对话

问答系统概述: 定义

◆定义

输入：自然语言的问句，而非关键词的组合

例如：谁获得1987年的诺贝尔文学奖?

输出：直接答案，而非文档集合

例如：约瑟夫·布罗茨基

问答发展历程

问答式检索系统

- 搜索引擎为人们的信息获取提供了可能，但搜索引擎无法清楚表达人们的信息需求意图，返回的信息太多；
- 为了克服搜索引擎的不足，问答式检索系统应运而生；
- 主要特点：利用信息检索以及浅层自然语言处理技术从大规模文本库或者网页库中抽取出答案；
- 代表性系统：
  - MIT开发的Start (http://start.csail.mit.edu/index.php/)

问答发展历程

问答式检索系统

- 优点：
  - 相对于基于知识推理的问答系统而言：不受知识库规模限制，不受领域限制，更加接近真实应用需求；
  - 相对于搜索引擎而言：问答式检索系统接受的是自然语言形式的提问，由于自然语言处理技术的应用，对用户意图的把握更加准确，呈现给用户的答案更加准确；
- 缺点：目前问答式检索系统仅能处理有限的简单问题，如Factoid问题等；

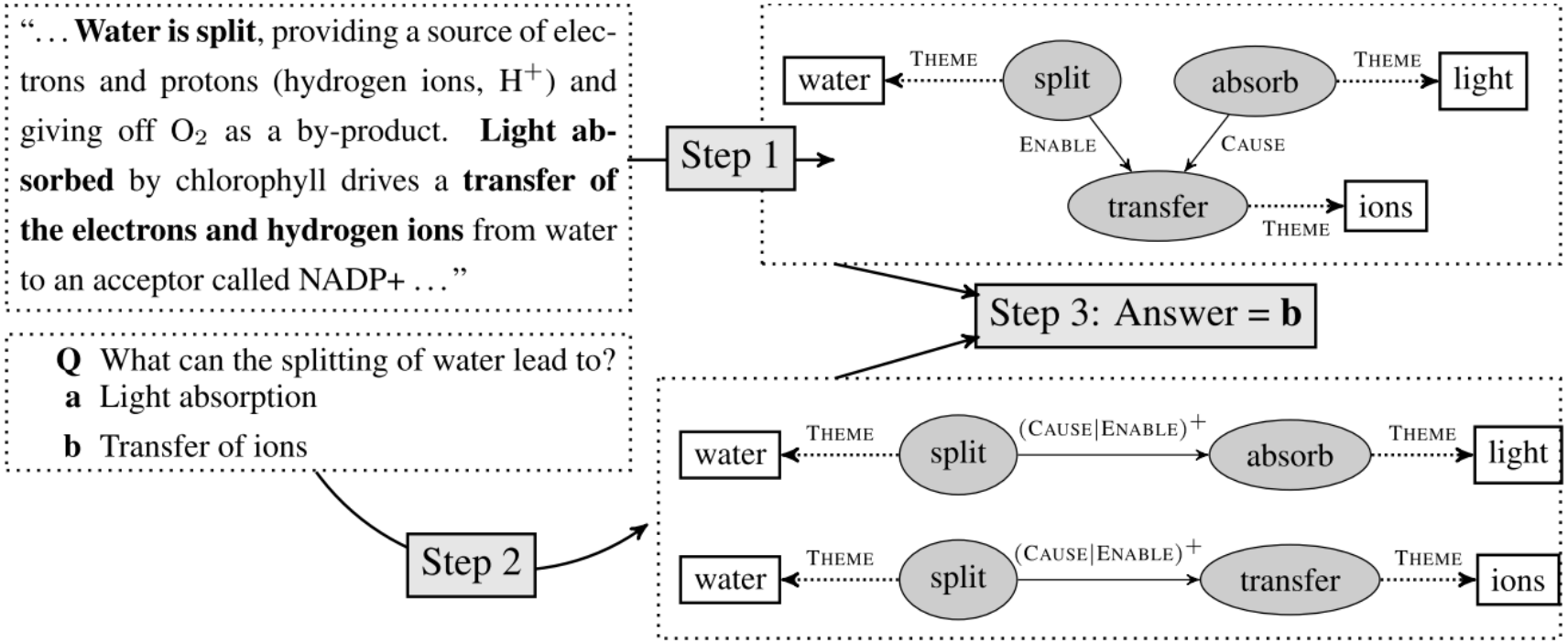
问答发展历程

社区问答系统：

- 随着Web 2.0的兴起，基于用户生成内容(User-Generated Content,UGC)的互联网服务越来越流行，社区问答系统应运而生，为问答系统的发展注入了新的活力和生机；
- 用户可以提出任何类型的问题，也可以回答其它用户的问题，通过问答方式来满足人们的信息查找和知识分享需求；
- 代表性系统：
  - 英文：Yahoo! Answers等；
  - 中文：百度知道、新浪爱问、**知乎**等；

问答发展历程

阅读理解系统



问答式检索方法

- 信息检索 + 信息抽取
- 信息检索 + 模式匹配
- 信息检索 + 自然语言处理技术
- 基于统计翻译模型的问答技术

信息检索 + 信息抽取

- 方法描述：从问句中提取关键词语，用信息检索的方法找出包含候选答案的段落或句子，然后基于问答类型用信息抽取的方法在这些段落和句子中找出答案
  - 检索：段落或者句子级排序，利用不同类型关键词的加权组合
  - 答案抽取：根据问答类型从排序后的段落或句子中抽取答案
- 特点：
  - 优点：技术相对成熟，易于开发
  - 缺点：准确率一般，不能推理

信息检索 + 模式匹配

方法描述：

- 基本思想：对于某些提问类型，问句和包含答案的句子之间存在一定的答案模式，该方法在信息检索的基础上根据这种模式找出答案。

- 例如，询问“某人生日年月日”类提问的部分答案模式：
  - 1.0 <NAME>（<ANSWER> -）
  - 0.85 <NAME> was born on <ANSWER>
  - 0.6 <NAME> was born in <ANSWER>
  - 0.59 <NAME> was born <ANSWER>
  - 0.53 <ANSWER> <NAME> was born
  - 0.50 – <NAME>（<ANSWER>

信息检索 + 模式匹配

- 包括两阶段的任务:
  - 离线阶段：获取答案模式
  - 在线阶段：首先判断当前提问属于哪一类，然后使用这类提问的所有模式来抽取候选答案
- 模式获取方法:
  - 表层字符串匹配(Ravichandran ACL 2002)
  - 深层句法分析(Lin NLE 2001)
- 特点:
  - 优点：对于某些类型的问题(如生日问题等) 效果良好
  - 缺点：无法表达长距离、复杂关系，没有推理能力

信息检索+自然语言处理技术

- **方法描述：**
  - 对问句和答案句进行浅层句法分析，获得句子的浅层句法、语法表示，作为对前两种方法的补充和改进
- **涉及到的自然语言处理技术主要包括：**
  - 命名实体识别技术(Ravichandran ACL 2002)
  - 句法分析技术(Lin NLE 2001)
  - 逻辑表示(Harabagiu TREC 2000; Moldovan ACL 2001)
  - …

信息检索+自然语言处理技术

- 特点：
  - 优点：能够从句法、语义的角度解析答案
  - 缺点：技术还不成熟
- 代表系统：
  - Sanda Harabagiu等人研发的系统(Harabagiu TREC 2000)，该系统在TREC QA Track 评测中获得好成绩，且具有较大的领先优势

基于统计翻译模型的问答技术

- **方法描述：**
  - 把提问句看作答案句在同一语言内的一种翻译
- **特点：**
  - 过分依赖于训练集

四类问答技术的比较分析

- 基于信息检索和信息抽取的问答技术：相对简单，容易实现。但它以基于关键词的检索技术(或称为词袋检索技术)为重点，只考虑离散的词，不考虑词之间的关系。因此无法从句法关系和语义关系的角度解释系统给出的答案，也无法回答需要推理的提问

- 基于模式匹配的问答技术：虽然对于某些类型提问(如定义，出生日期提问等)有良好的性能，但无法找到所有提问的答案模式，长距离模式和表达复杂关系的模式的获取也很困难，同样无法实现推理

四类问答技术的比较分析

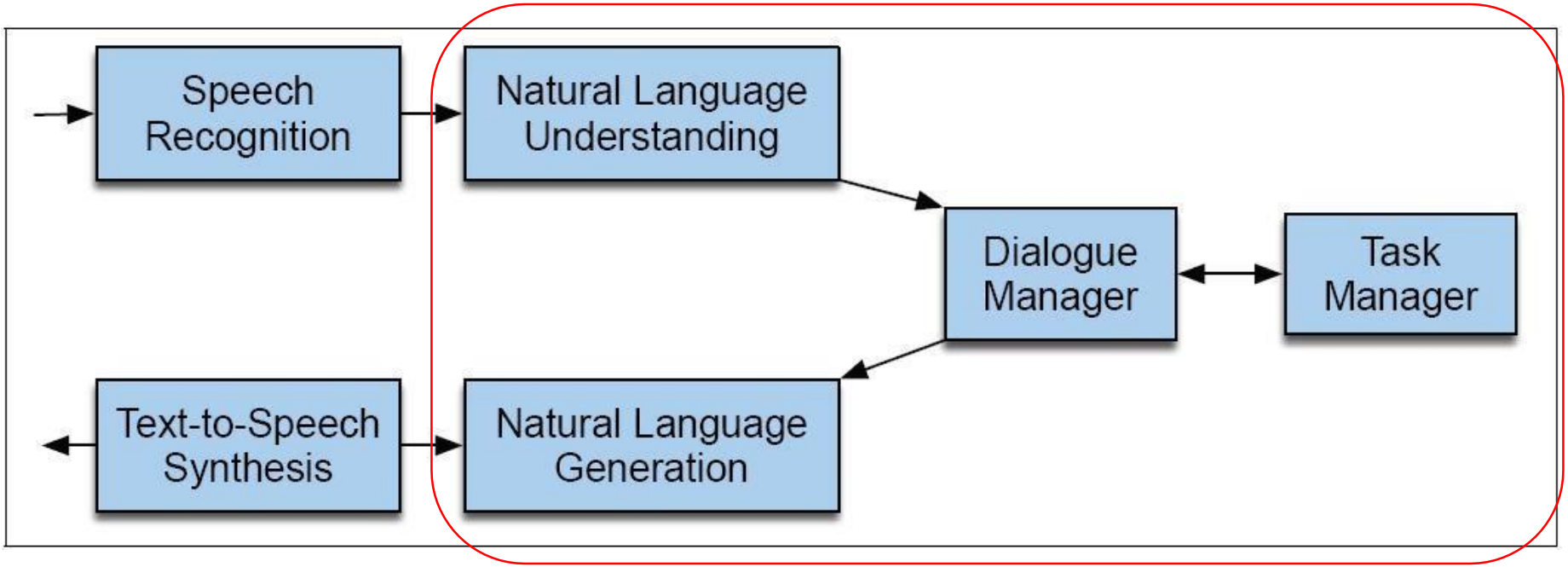
- 基于自然语言处理的问答技术：可以对提问和答案文本进行一定程度的句法和语义分析，从而实现推理。但自然语言处理技术还不成熟，除一些浅层的技术(汉语分词、 命名实体识别、词性标注等)外，其他技术还没有达到实 用程度。所以这种技术的作用还有限，只能作为对前两种方法有效补充

- 基于统计翻译模型的问答技术：在很大程度上依赖训练语料的规模和质量，而对于开放域问答系统，这种大规模训练语料的获取是非常困难的

新的方向

基于深度学习的方法：把回答问题的过程看作一个黑盒子，通过复杂的神经网络和超大规模的数据集训练出一个拟合能力强大的模型；

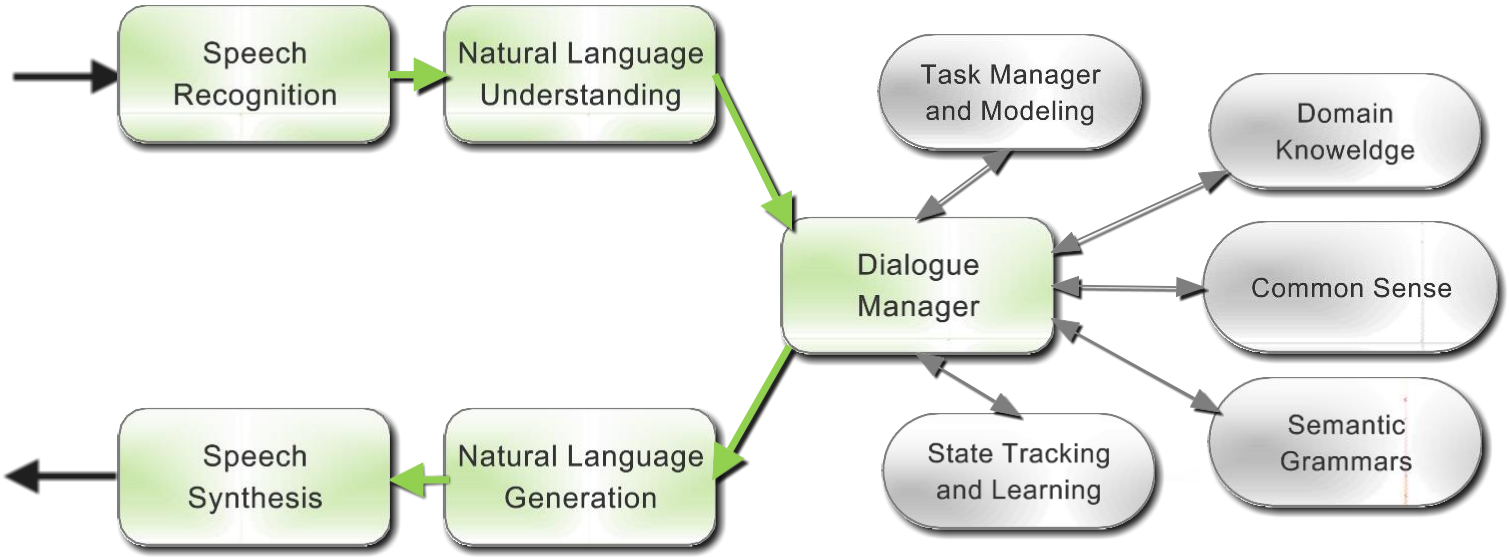
对话系统



- 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)

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

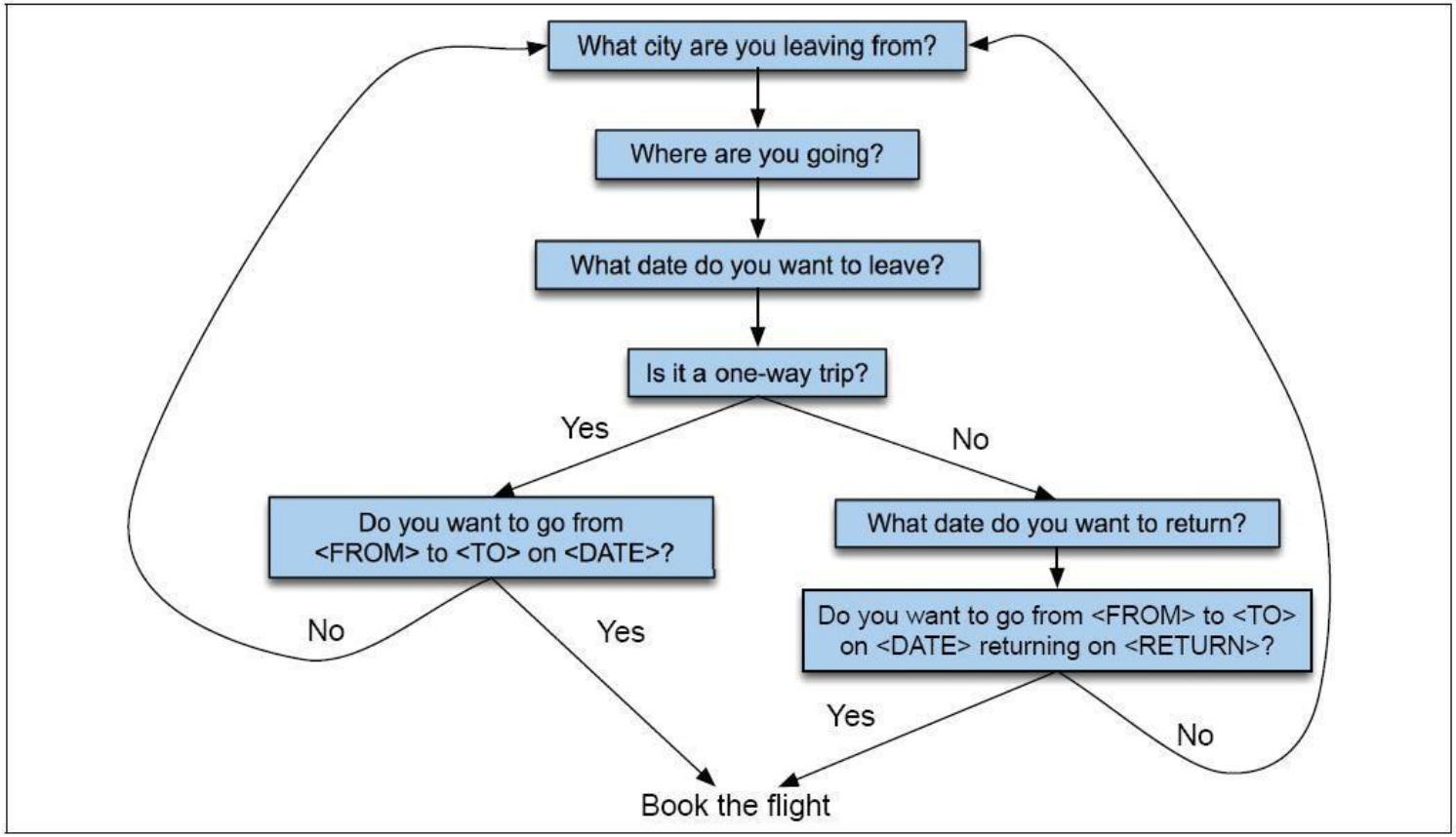
- 有限状态机（Finite State)
- 基于框架的方法（Frame-based)
- 统计方法：Information State (Markov Decision Process)

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，通过有线状态机进行建模



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

Thank you!

权小军 中山大学数据科学与计算机学院