

# 自然语言处理概述

鲲鹏创新实践课：鲲鹏应用数据分析与管理实战



# 前言

- 本章主要讲述什么是自然语言、什么是自然语言处理，以及自然语言处理所用到的基本方法，研究的任务方向以及自然语言处理的难点，最后介绍自然语言处理的典型应用。

浙大-华为鲲鹏创新实践



# 目标

- 学完本课程后，您将能够：
  - 了解自然语言处理的基本概念；
  - 了解自然语言处理的基础任务；
  - 了解自然语言处理的应用场景；

浙大-华为鲲鹏创新实践课



# 目录

1. 自然语言处理概念
2. 自然语言处理的特点
3. 自然语言处理的基本方法
4. 自然语言处理的任务及技术框架
5. 自然语言处理的难点
6. 自然语言处理的典型应用



# 何谓“自然语言”

- 什么是自然语言？
  - 自然语言 (Natural language) 通常是指一种自然地随文化演化的语言。英语、中文、法语、西班牙语、葡萄牙文、日语、韩语、意大利文、德文为自然语言的例子，而世界语则为人工语言，即是一种由人特意为某些特定目的而创造的语言。
  - 以语音为物质外壳，由词汇和语法两部分组成的符号系统。
  - 是约定俗成的，有别于**人工语言**，比如Java、C++等程序设计语言。



# 何谓“自然语言处理”

- 自然语言处理（ Natural Language Processing， NLP ）是人工智能和语言学领域的分支学科。
- 此领域探讨如何用机器处理人类语言的理论和技术，让计算机能够**理解**和**生成**人类语言。
- 自然语言处理包括多方面和步骤，基本有认知、理解、生成等部分。
  - 自然语言认知和理解是让电脑把输入的语言变成有意思的符号和关系，然后根据目的再处理。
  - 自然语言生成系统则是把计算机数据转化为自然语言。



# 目录

1. 自然语言处理概念
2. **自然语言处理的特点**
3. 自然语言处理的基本方法
4. 自然语言处理的任务及技术框架
5. 自然语言处理的难点
6. 自然语言处理的典型应用



## 互动 - 当你想听歌时，会怎么表达

- 我要听野狼disco
- 给我播野狼disco
- 我想听歌野狼disco
- 放首野狼disco
- 给唱一首野狼disco
- 放音乐野狼disco
- 放首歌野狼disco
- 给大爷来首野狼disco





# 自然语言处理的特点

语言的多样性

语言的歧异性

语言的鲁棒性

语言需要一定的知识依赖

语言的使用基于环境和上下文

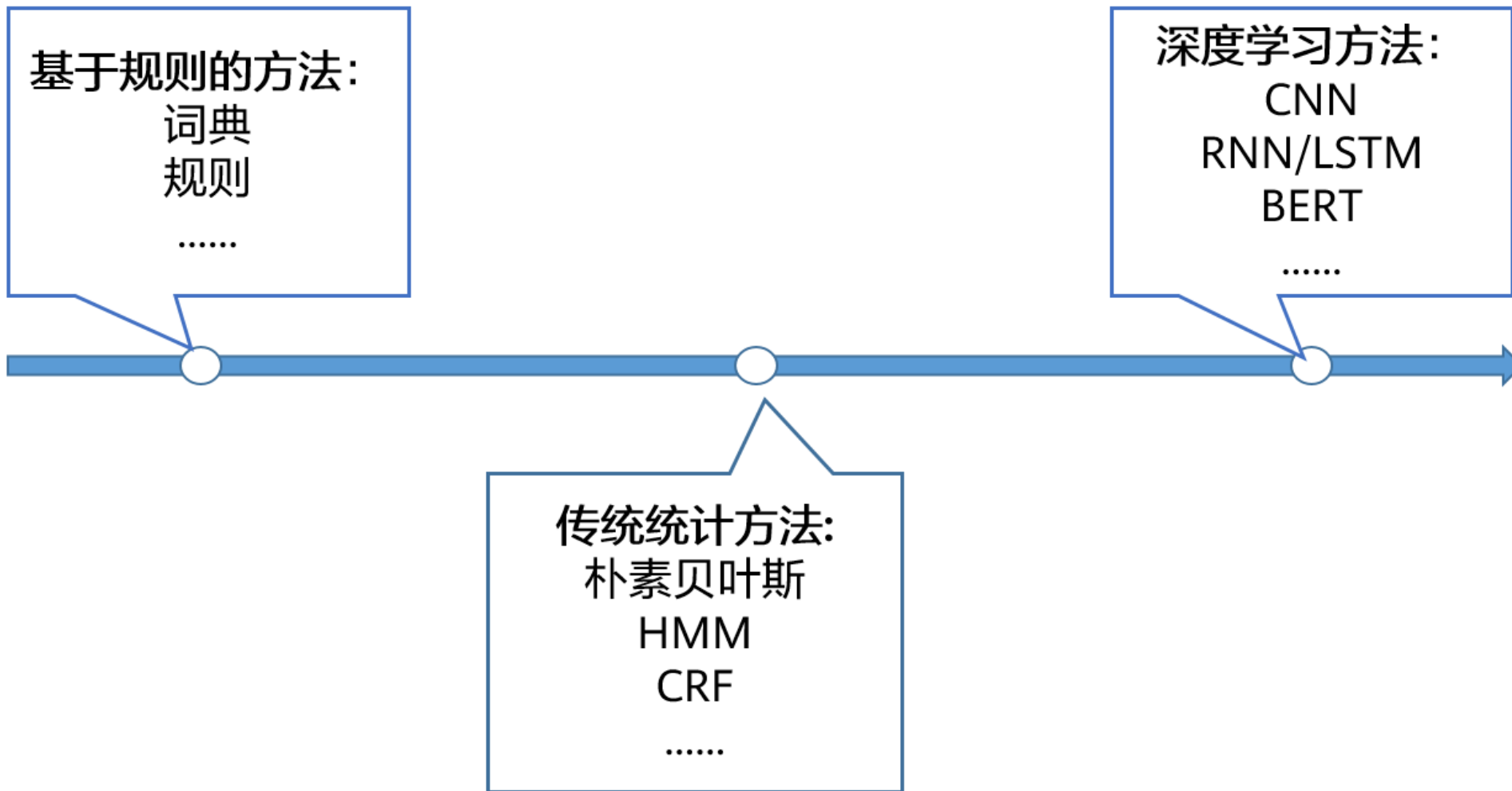


# 目录

1. 自然语言处理概念
2. 自然语言处理的特点
3. **自然语言处理的基本方法**
4. 自然语言处理的任务及技术框架
5. 自然语言处理的难点
6. 自然语言处理的典型应用



# 基本方法





# 目录

1. 自然语言处理概念
2. 自然语言处理的特点
3. 自然语言处理的基本方法
4. **自然语言处理的任务及技术框架**
5. 自然语言处理的难点
6. 自然语言处理的典型应用



# NLP的两大问题

自然语言理解

NLU

自然语言生成

NLG



## 互动问题

- 出行时，需要订机票，你会怎么表达：

帮您预订一张明天上午从广州到上海的机票

为您找到最适合的3个航班

06:45	09:00	有票
白云机场	1308 虹桥机场	¥448.0
春秋航空   9C5930   2h15m		经济舱3.5折
07:45	09:10	有票
白云机场	1308 虹桥机场	¥929.0
春秋航空   AC5930   2h15m		经济舱7.5折
08:45	11:00	有票
白云机场	1308 虹桥机场	¥929.0
春秋航空   ZH5930   2h15m		经济舱7.5折



# 意图判断

- 出行时，需要订机票，你会怎么表达：
  - “订机票”；
  - “有去杭州的航班么？”；
  - “看看航班，下周三出发去法国的”；
  - “要出差，帮我查下机票”；
- 基于规则的意图判断
  - “订机票”；
- 基于NLU的意图判断
  - “看看航班，下周三出发去法国的”（意图、时间、目的地）



# 基于NLU的应用 – 机器翻译







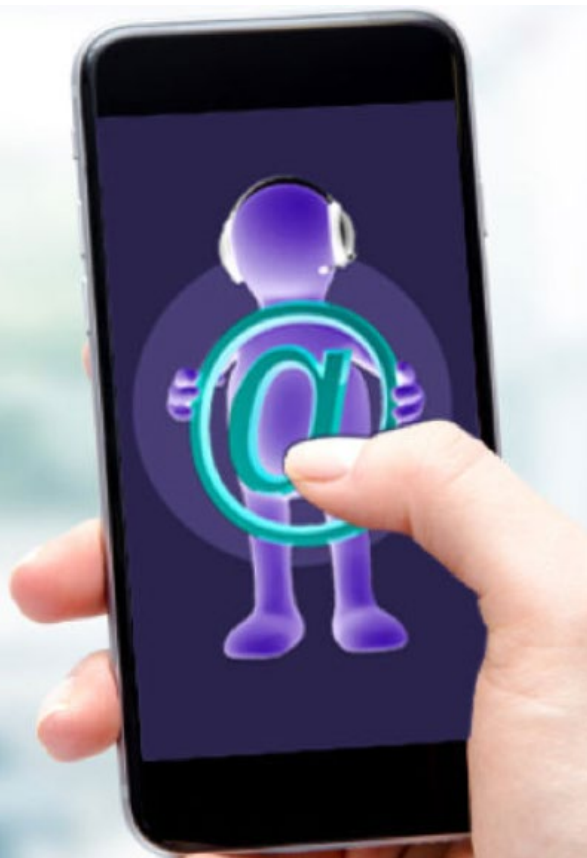
# 基于NLU的应用 – 机器客服

华为云服务咋开通?

请推荐拍照好的华为手机

有没有2000左右的智能手表?

有没有适合上网的流量套餐?





# 基于NLU的应用 – 智能硬件/语音助手

- “我冷了”
- 机器：帮您把空调调高2度





# 自然语言生成 (NLG) 的方式

- 3个level:
  - 简单的数据合并
  - 模板化的NLG
  - 高级NLG

 Text to text: 文本到语言的生成

 Data to text: 数据到语言的生成



# 简单的数据合并

文档2 - Word

表格工具 设计 布局

shirh19910112

告诉我你想要做什么 共享

文件 开始 插入 设计 布局 引用 邮件 审阅 视图 办公标签 ABBYY FineReader 12

中文信封 信封 标签 开始 选择 编辑 邮件合并 - 收件人 - 收件人列表 突出显示 地址块 问候语 插入 规则 匹配域 更新标签 预览结果 查找收件人 检查错误 完成并合并 完成

创建 开始邮件合并 编写和插入域 预览结果

文档2 \* x

CEACA 国家信息化计算机教育认证

准考证

个人信息	
姓 名	I
性 别	
身份证号	
专业信息	
班 级	
专 业	
考试信息	
考试模块	
准考证号	
登录口令	
考 场	
考试时间	

节: 1 第 1 页, 共 1 页 58 个字 中文(中国)

120%



# 模板化的NLP

北京 今日 晴热高温，及时添衣，谨防感冒

一起来看下今日天气

今天是 6月29日，北京 的天气 晴热高温，最高温 33 °C，最低温是 23 °C。各项气象条件适宜，发生感冒机率较低。但请避免长期处于空调房间中，以防感冒。接下来几天，气温保持平稳。



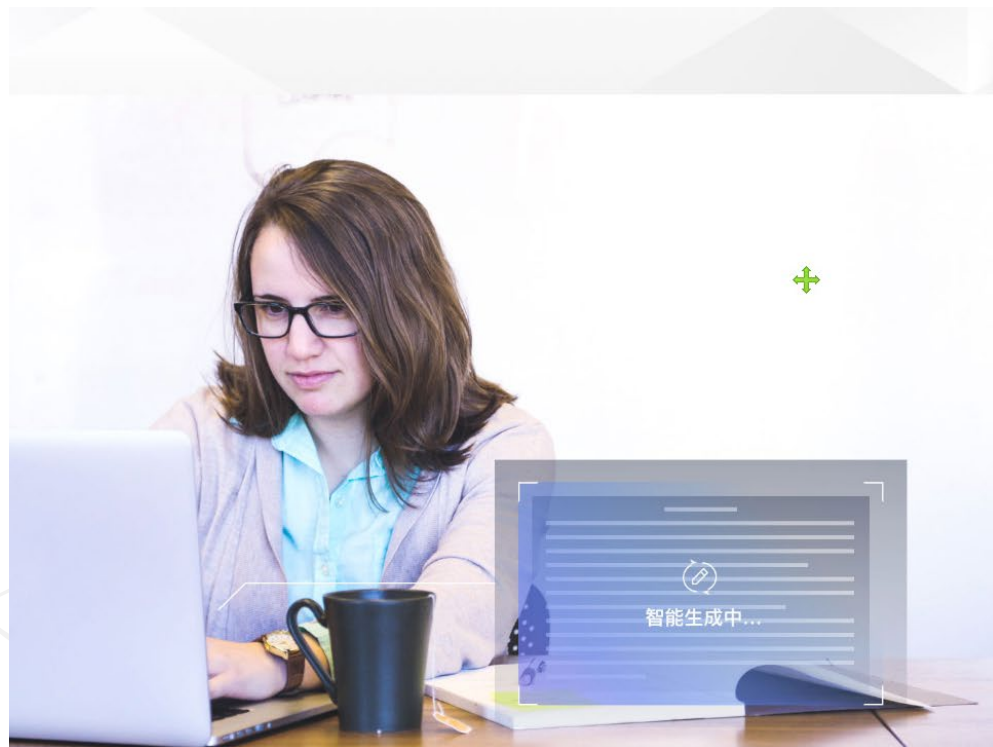
# 自然语言生成的步骤

- 确定内容
- 文本结构
- 句子聚合
- 语法化
- 参考表达式生成
- 语言实现



# 基于NLG的应用

- 自动写新闻
- 聊天机器人
- 文案/报告生成
- ...



## 智能文案

通过学习现有文稿的写作方式和结构，使用机器学习算法，自动生成新文案

## 优势

- 快速生成  
快速响应，完成大量文案的生成
- 表达丰富  
通过学习现有的文稿生成丰富的语句表述



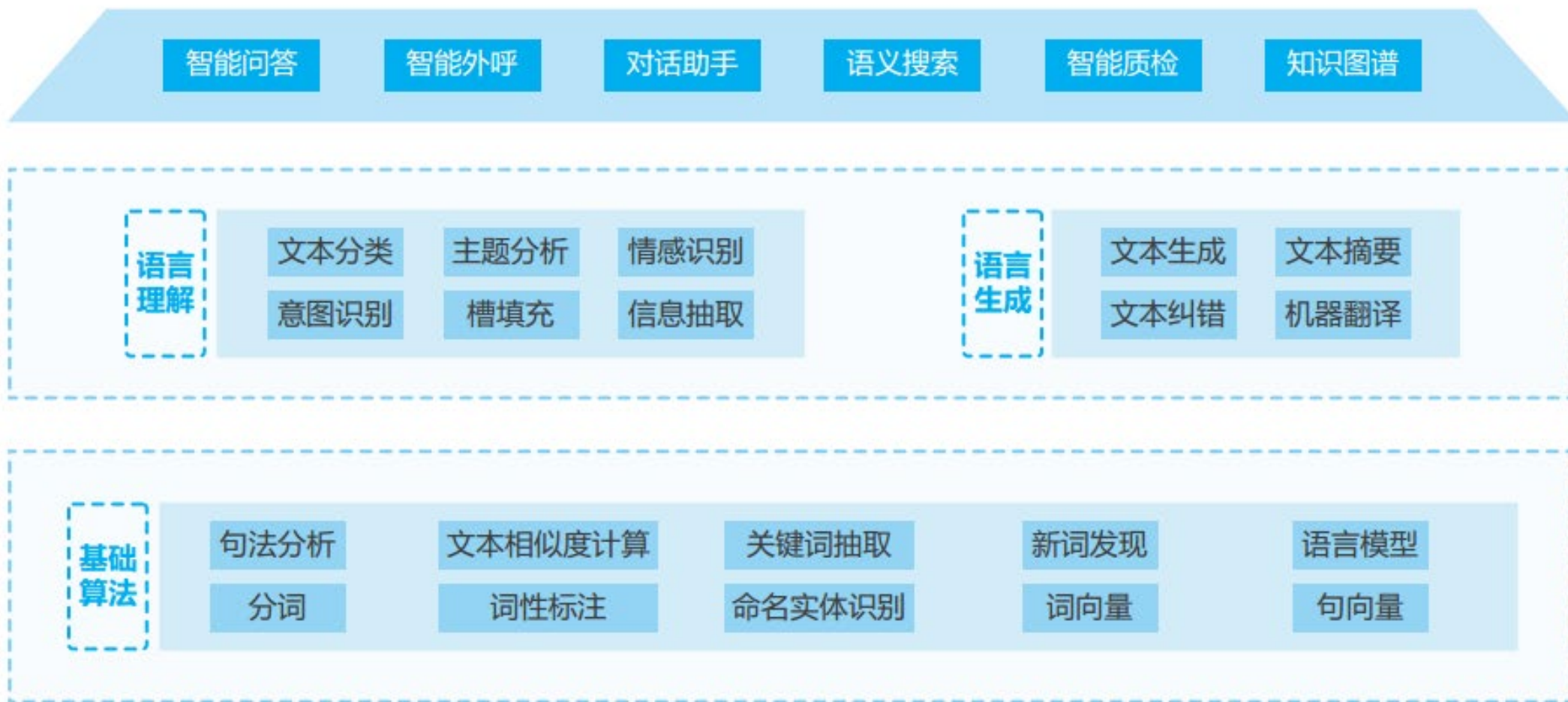
# 自然语言处理的具体任务

- 基础任务：
  - **文本向量化**，词/短语向量，句子/段落/篇章向量
  - **语言模型**，生成类任务基础模型
- 应用任务：
  - **分类任务**，譬如垃圾邮件分类、情感分析、意图识别等
  - **聚类任务**，事件聚合，典型意见挖掘等
  - **序列标注**，命名实体识别、语义标注、词性标注、分词等
  - **句对关系判断**，自然语言推理、问答QA、文本语义相似性等
  - **生成式任务**，内容创作、机器翻译、文本摘要等





# NLP技术框架





# 目录

1. 自然语言处理概念
2. 自然语言处理的特点
3. 自然语言处理的基本方法
4. 自然语言处理的任务及技术框架
5. **自然语言处理的难点**
6. 自然语言处理的典型应用



# 自然语言理解层次

- 自然语言的理解和分析是一个层次化的过程，具体可分为以下五个层次：
  - 语音分析，是要根据音位规则，从语音流中区分出一个个独立的音素，再根据音位形态规则找出音节及其对应的词素或词。
  - 词法分析，是找出词汇的各个词素，从中获得语言学的信息。
  - 句法分析，是对句子和短语的结构进行分析，目的是要找出词、短语等的相互关系以及各自在句中的作用。
  - 语义分析，是找出词义、结构意义及其结合意义，从而确定语言所表达的真正含义或概念。
  - 语用分析，是研究语言所存在的外界环境对语言使用者所产生的影响。





# 自然语言处理的难点 (1)

- 词法歧义：

- 分词：词语的切分边界比较难确定。

- 严守一/把/手机/关/了

- 严守/一把手/机关/了

- 词性标注：同一个词语在不同的上下文中词性不同。

- 我/计划/v 考/研/

- 我/完成/了/计划/n

- 命名实体识别：人名、专有名称、缩略词等未登录词的识别困难。

- 高超/nr/a      华明/nr/nt      移动/nt/v

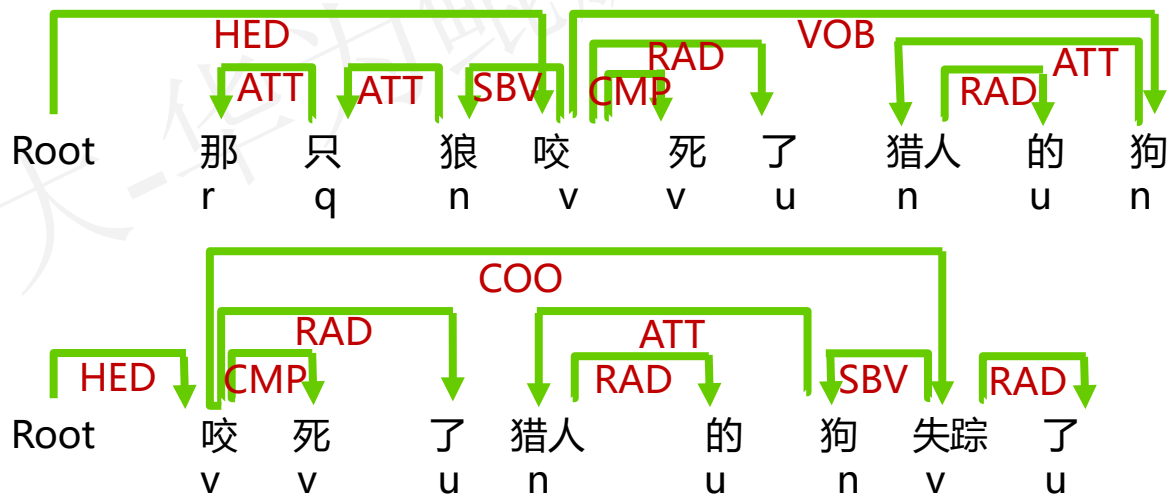


## 自然语言处理的难点 (2)

- 句法歧义：句法层面上的依存关系受行文的影响。

- 咬死了猎人的狗

- 那只狼咬死了猎人的狗
- 咬死了猎人的狗失踪了





## 自然语言处理的难点 (3)

- 语义歧义
  - At last, a computer that understands you like your mother.
    - 含义1: 计算机像你的母亲那样很好的理解你。
    - 含义2: 计算机理解你喜欢你的母亲。
    - 含义3: 计算机像理解你母亲那样去理解你。



## 自然语言处理的难点 (4)

- 语用歧义
  - “你真坏”
    - 当对干了坏事的成年人说时，是一种严厉的苛责。
    - 当妈妈对淘气的儿子说时，实际表达的是对儿子的一种疼爱。
    - 当恋爱中的女孩对男友说时，则是女孩在男友面前撒娇的一种表现。



# 目录

1. 自然语言处理概念
2. 自然语言处理的特点
3. 自然语言处理的基本方法
4. 自然语言处理的任务及技术框架
5. 自然语言处理的难点
6. **自然语言处理的典型应用**





## 典型应用

情感分析

聊天机器人

语音识别

机器翻译



# 情感分析

**情感分析：**对商品评论自动过滤，筛选负面评论供人工审核

买没几天就降价一点都不开心，闪存跑分就五百多点点 ---



外观漂亮音质不错，现在电子产品基本上都是华为的了 ---

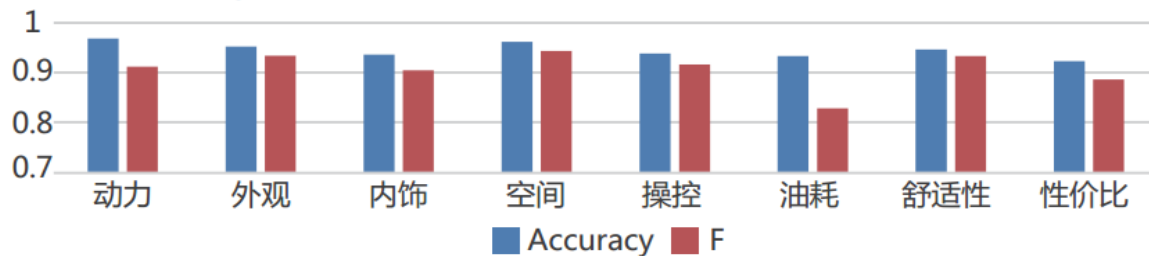


	通用领域	汽车领域	电商领域
Accuracy	85.58%	68.60%	94.90%
F1 score	85.36%	86.41%	94.89%

## 属性情感分析

评论	动力	外观	空间	油耗
2.0T涡轮增压发动机动力强，各种路况表现都不错，高速120超车没压力；外观是我和老婆都比较喜欢的款；后排空间有点小；有点费油啊。	😊	😊	😞	😞

Aspect level sentiment (细粒度情感分析) - 汽车





# 应用场景 - 内容审核

直播，短视频，论坛，游戏等

直播、视频网站



视频，图片合规审核

媒体资讯



图片，视频合规检查

社交平台、论坛、电商

花粉俱乐部

VMALL.COM 华为商城

防灌水，合规检查

游戏



用户名，聊天合规检查

供应链



图片质量审核



## 思考题

1. （判断题）自然语言处理包括自然语言理解和自然语言生成两大部分。（ ）
  - A. True
  - B. False
2. 自然语言的理解和分析包含的哪些层面？（ ）
  - A. 词法分析
  - B. 句法分析
  - C. 语音分析
  - D. 语义分析



## 本章总结

- 本章首先介绍了自然语言处理的基本概念，然后介绍了自然语言处理的基本方法，研究的任务方向，自然语言处理的疑难点，最后介绍了自然语言处理的典型应用。

浙大-华为鲲鹏创新实践





谢谢

[www.huawei.com](http://www.huawei.com)