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

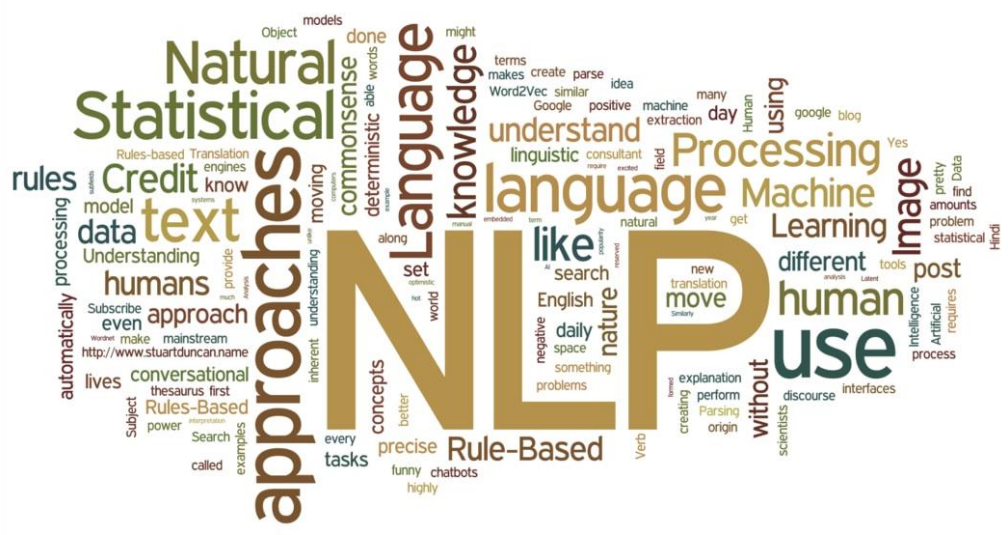
巴别塔的故事

1.1 自然语言处理概述

自然语言处理概述

自然语言处理(NLP)

自然语言处理（自然语言理解），是计算机科学与人工智能领域中的一个重要方向。它研究能实现人与计算机之间通过自然语言进行交互的各种理论和方法.



自然语言处理（NLP）

- 让计算机能够自动或半自动地理解自然语言文本，懂得人的意图和心声
- 让计算机实现海量语言文本的自动处理、 挖掘和有效利用，满足不同用户的各种需求，实现个性化信息服务

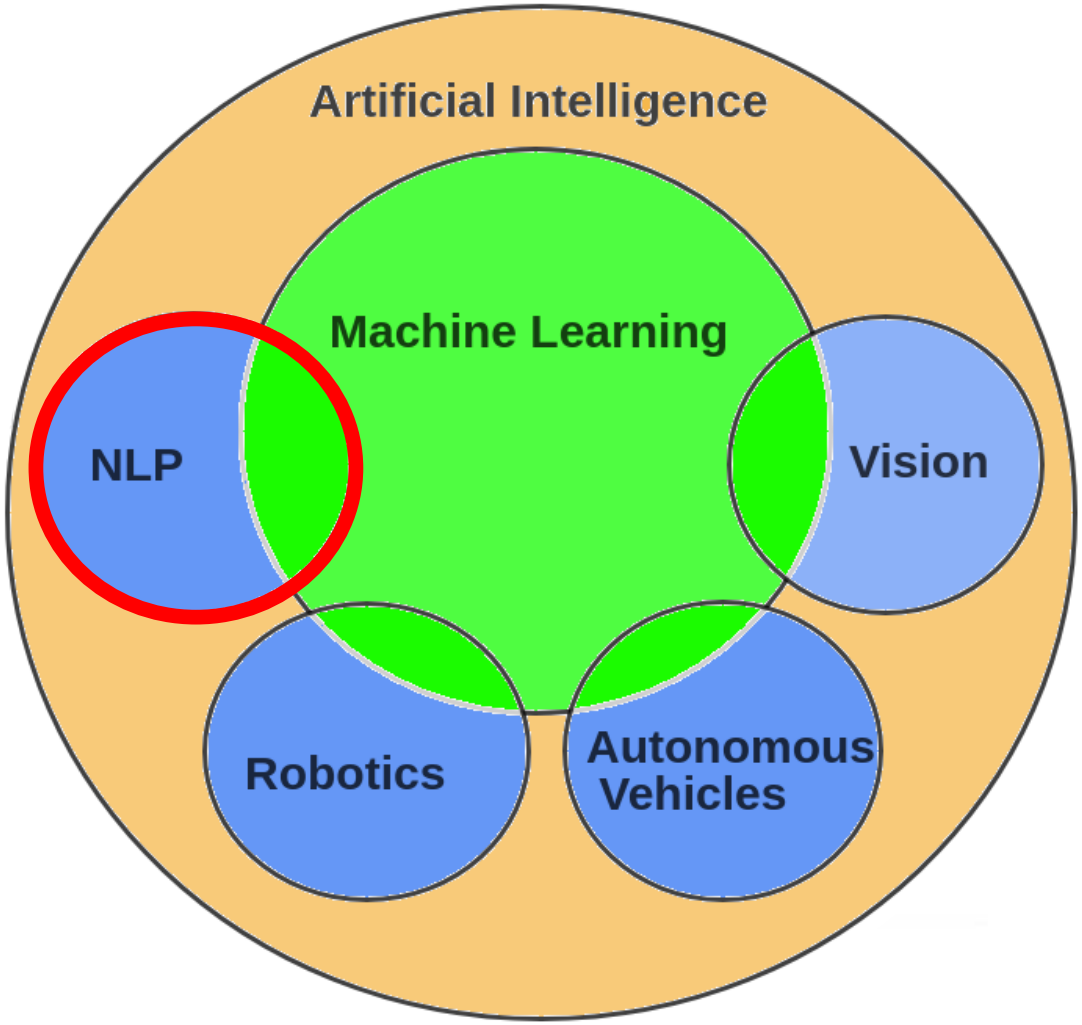
自然语言处理（NLP）

- 自然语言理解是人工智能皇冠上的明珠.
- 自然语言理解入选中国科协发布的信息技术十大前沿热点问题 (2019年)

自然语言处理（NLP）

自然语言处理又叫做计算语言学(computational linguistics)，涉及到计算、语言两方面的知识。

NLP vs ML vs AI



自然语言处理(NLP)

自然语言处理的研究方向包括:

- 中文自动分词
- 句法分析
- 信息抽取
- 情感计算
- 机器翻译
- 对话系统
- 信息检索
- 自动摘要

Why NLP?

- 语言是思维的载体，是人类交流思想、表达情感最自然、最直接、最方便的工具
- 人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上
- 2008年1月中国互联网络信息中心(CNNIC) 发布的《第21次中国互联网络发展状况统计报告》表明，中国互联网上有87.8%的网页内容是文本表示的

基本概念

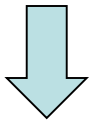
关于“理解”的标准

- 如何判断计算机系统的智能?

计算机系统的表现(act)如何?

反应(react)如何?

相互作用(interact)如何?

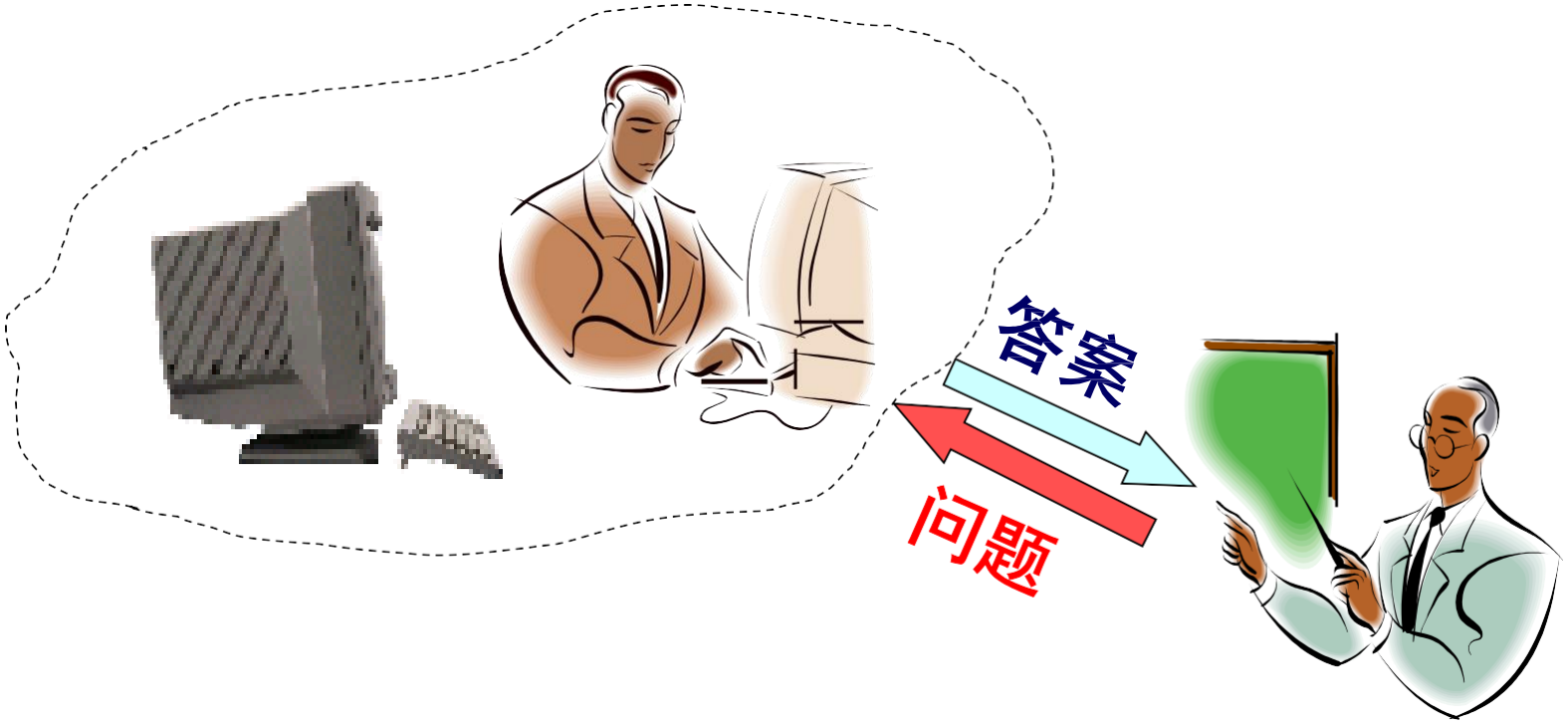


与有意识的个体（人）比较如何?

图灵设计的“模仿游戏” – 图灵实验(Turing test)

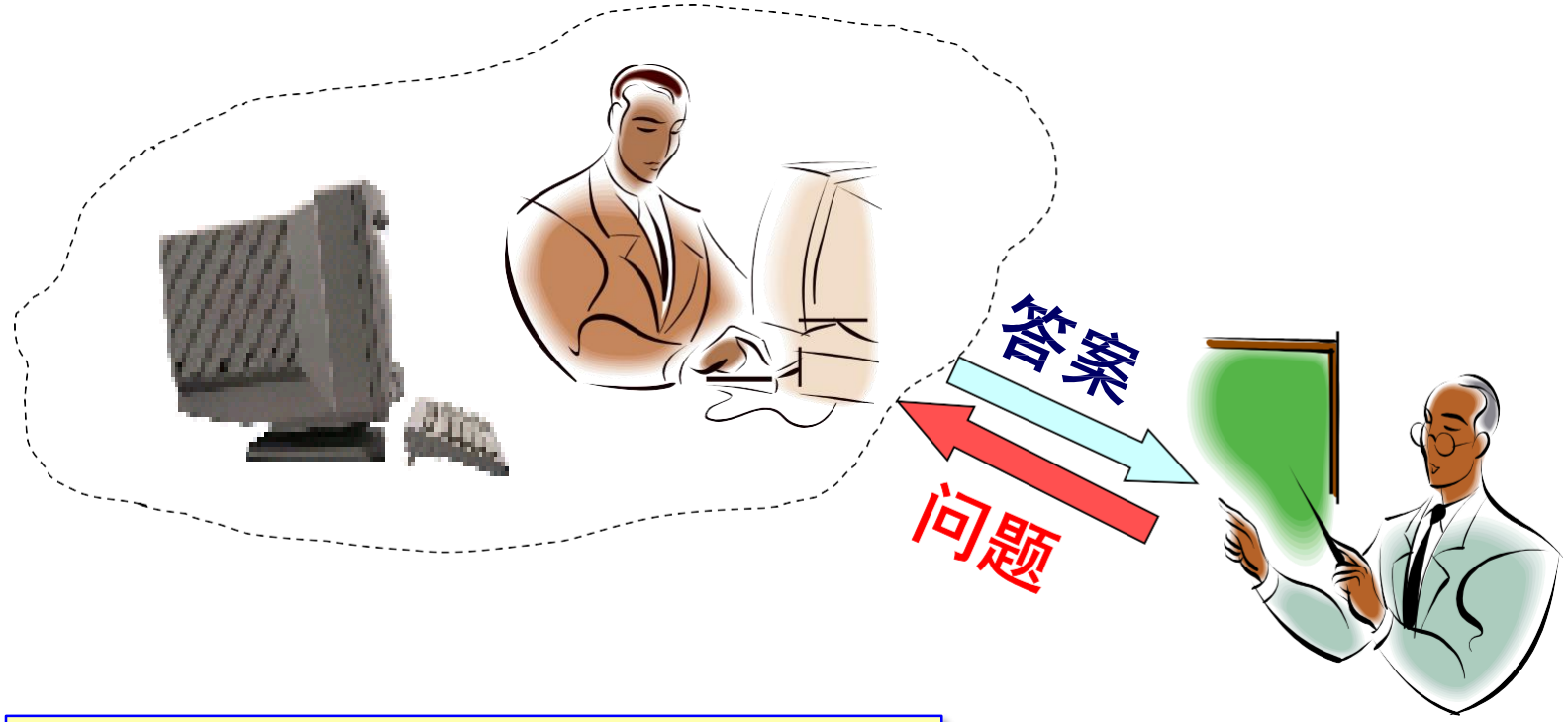
基本概念

图灵(Turing)测试：



基本概念

图灵(Turing)测试：



关于图灵测试仍有争议!

1.2 课程信息

课程信息

- 任课老师：权小军
  - 中山大学数据科学与计算机学院教授
  - 研究方向：自然语言处理，文本数据挖掘，机器学习
  - 办公地点：超算中心502H
- 助教：沈维州，杨云翊
  - 研究方向：自然语言处理，机器学习
  - 办公地点：超算中心502

课程目标

系统地学习自然语言处理（特别是中文语言）的基本概念、常用算法和重要应用。重点掌握词汇、句法、语义分析等的基本知识，理解统计自然语言处理的关键算法；在大规模语料库的支持下，掌握统计语言模型在语言知识自动学习中的应用；结合文本分类和聚类、机器翻译、信息检索等热门应用技术；了解深度学习技术在自然语言处理上的应用和进展.

课程目标

1 **系统地学习自然语言处理（特别是中文语言）的基本概念、常用算法和重要应用。**重点掌握词汇、句法、语义分析等的基本知识，理解统计自然语言处理的关键算法；在大规模语料库的支持下，掌握统计语言模型在语言知识自动学习中的应用；结合文本分类和聚类、机器翻译、信息检索等热门应用技术；了解深度学习技术在自然语言处理上的应用和进展.

课程目标

系统地学习自然语言处理（特别是中文语言）的基本概念、常用算法和重要应用；**重点掌握词汇、句法、语义分析等的基本知识，理解统计自然语言处理的关键算法；**在大规模语料库的支持下，掌握统计语言模型在语言知识自动学习中的应用；结合文本分类和聚类、机器翻译、信息检索等热门应用技术；了解深度学习技术在自然语言处理上的应用和进展.

课程目标

系统地学习自然语言处理（特别是中文语言）的基本概念、常用算法和重要应用。重点掌握词汇、句法、语义分析等的基本知识，理解统计自然语言处理的关键算法 3 **在大规模语料库的支持下，掌握统计语言模型在语言知识自动学习中的应用；**结合文本分类和聚类、机器翻译、信息检索等热门应用技术；了解深度学习技术在自然语言处理上的应用和进展.

课程目标

系统地学习自然语言处理（特别是中文语言）的基本概念、常用算法和重要应用。重点掌握词汇、句法、语义分析等的基本知识，理解统计自然语言处理的关键算法；在大规模语料库的支持下，掌握统计语言模型在语言知识自动学习中的应用；④**结合文本分类和聚类、机器翻译、信息检索等热门应用技术；**了解深度学习技术在自然语言处理上的应用和进展.

课程目标

系统地学习自然语言处理（特别是中文语言）的基本概念、常用算法和重要应用。重点掌握词汇、句法、语义分析等的基本知识，理解统计自然语言处理的关键算法；在大规模语料库的支持下，掌握统计语言模型在语言知识自动学习中的应用；结合文本分类和聚类、机器翻译、信息检索等热门应用技术

5 **了解深度学习技术在自然语言处理上的应用和进展.**

课程信息

授课内容
1、自然语言处理技术概论
2、数学和信息论基础
3、自然语言处理技术的语言学基础
4、词法分析
5、语言模型
6、马尔可夫模型
7、句法分析技术
8、深度学习与自然语言处理
9、文本摘要
10、机器翻译
11、对话系统

考核方式

- 闭卷考试：60%
  - 重点考察知识的灵活掌握能力；
- 课程设计：30% （1个）
  - 代码(python)，报告
- 课堂考勤：10%
  - 1次免责机会；
  - 缺席超过1次，每次扣5分，上限40分；

课程信息

- 微信群

2020年《自然语言处理》课程群

该二维码7天内(9月9日前)有效，重新进入将更新

推荐教材

- 主讲教材
  1. 宗庆成，《统计自然语言处理》，清华大学出版社，2008
  2. CS224n: Natural Language Processing with Deep Learning https://web.stanford.edu/class/cs224n/index.html

- 辅助教材
  1. 冯志伟、孙乐译,《自然语言处理综论》,电子工业出版社，2005
  2. 吴军，《数学之美》，人民邮电出版社，2012

1.3 自然语言处理应用

研究内容

- 机器翻译 (Machine translation, MT)：实现一种语言到另一种语言的自动翻译。
  - 应用：文献翻译、网页辅助浏览等。
  - 代表系统：
    - Google：http://translate.google.cn(103 种语言)
    - 百度：http://fanyi.baidu.com/ (28种语言，包括文言文和简繁转换)

研究内容

- 信息检索(Information retrieval)

  信息检索也称情报检索，就是利用计算机系统从大量文档中找到符合用户需要的相关信息。

  - 代表系统：Google，百度

  目前至少有300多亿个网页，每天数以万计地增加，只有1%的信息被有效地利用。

研究内容

- 问答系统 (Question-answering system)

  通过计算机系统对人提出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并做出相应的回答。问答技术有时与语音技术和多模 态输入/输出技术，以及人机交互技术等相结合，构 成人机对话系 统(man-computer dialogue system)。

  - IBM Watson 自动问答系统

当前的一个研究热点是对话系统

研究内容

- 自动文摘 (Automatic summarization)

  将原文档的主要内容或某方面的信息自动提取出来，并形成原文档的摘要或缩写。

- 情感分析与观点挖掘 (Sentiment analysis and opinion mining)

  挖掘用户评论中包含的情感。

研究内容

- **信息抽取(Information extraction)**

  从指定文档中或者海量文本中抽取出用户感兴趣的信息。

  - 实体关系抽取 (entity relation extraction)。
  - 社会网络 (social network)

研究内容

- 文档分类(Document categorization)

  文档分类也叫文本自动分类(Text categorization / classification)，其目的就是利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类。

  情感分类(Sentimental classification)

  - 应用：图书管理、情报获取、网络内容监控等。

研究内容

- 文字编辑和自动校对(Automatic proofreading)

  对文字拼写、用词、甚至语法、文档格式等进行 自动检查、校对和编排。

  - 应用：排版、印刷和书籍编撰等。

1.4 基本问题和主要困难

基本问题

- 基本研究问题之一：形态学 (Morphology)
  - 研究词(word) 由有意义的基本单位－词素的构成问题。
  - 单词的识别/ 汉语的分词问题。

词素：词根、前缀、后缀、词尾

例如：老虎　　老 + 虎

图书馆　　图 + 书 + 馆

基本问题

- 基本研究问题之二：句法 (Syntax) 问题
  - 研究句子结构成分之间的相互关系和组成句子序列的规则
  - 为什么一句话可以这么说也可以那么说？如何建立快速有效的句子结构分析方法？

苹果，我吃了

vs.我吃了苹果

vs. 苹果吃了我

他欠我100万

vs.我欠他100万

基本问题

- 基本研究问题之三：语义(Semantics) 问题
  - 研究如何从一个语句中推导出词的意义，以及这些词在该语句句法结构中的作用来推导出该语句的意义。

这些话说了什么？

(1) 苹果不吃了
(2) 这个人真牛
(3) 这个人眼下没些什么
(4) 火烧圆明园/火烧驴肉

基本问题

- 基本研究问题之四：语用学(Pragmatics) 问题
  - 研究在不同上下文中语句的应用，以及上下文对语句理解所产生的影响。

为什么要说这句话?

1) 火，火!
2) 看看鱼怎么样了?

主要困难

- 困难之一：大量歧义(ambiguity)现象

I. 词法歧义，例如：

1) 自动化研究所取得的成就

a) 自动化/研究所/取得/的/成就

b) 自动化/研究/所/取得/的/成就

2) 门把手弄坏了

a) 门/把/手/弄/坏/了

b) 门把手/弄/坏/了

主要困难

- 困难之一：大量歧义(ambiguity)现象

  I. 词法歧义，例如：

主要困难

II. 词性歧义

①介词：像，好似；②动词：喜欢

1) Time flies like an arrow.

①动词：飞，飞翔，飞驰
②名词：苍蝇，飞虫

a) 时间像箭一样飞驰（光阴似箭）。
b) 时间苍蝇喜欢箭（有一种苍蝇叫“时间”）。

主要困难

II. 词性歧义

2) “动物保护警察”明年上岗

(《环球时报》2010年9月25日，第10版)

主要困难

III. 结构歧义

(1) 喜欢乡下的孩子。

(2) 关于鲁迅的文章。

主要困难

III. 结构歧义

(1) 喜欢乡下的孩子。

(2) 关于鲁迅的文章。

(3) 今天中午吃馒头。

(4) 今天中午吃食堂。

(5) 今天中午吃大碗。

(6) 今天中午吃了闭门羹。

主要困难

III. 结构歧义

(1) 喜欢乡下的孩子。

(2) 关于鲁迅的文章。

(3) 今天中午吃馒头。

(4) 今天中午吃食堂。

(5) 今天中午吃大碗。

(6) 今天中午吃了闭门羹。

(7) 写文章/写毛笔/写黑板

主要困难

III. 结构歧义

(8) I saw a man with a telescope.

a) I saw [a man with a telescope].
b) I [saw a man] with a telescope.

主要困难

IV. 语义歧义

他说:“她这个人真有意思(funny)”。她说:“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(wish),并让他向她意思意思(express)。他火了:“我根本没有那个意思(thought)”!她也生气了:“你们这么说是什么意思(intention)”?事后有人说:“真有意思(funny)”。也有人说:"真没意思(nonsense)"

—《生活报》1994.11.13.第6版

主要困难

IV. 语义歧义

人们的语言表达中大量地使用缩略语和隐喻的表达方式，如：

a)老虎苍蝇一起打.

b)破四旧，除四害；消灭一切牛鬼蛇神.

主要困难

V. 语音歧义：大量同音现象

施氏食狮史

石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮，十时，适十狮适市，是时，适施氏适市，施氏视是十狮，拭矢试，使是十狮逝世，适石室，石室湿，氏使侍拭石室，石室拭，始食是十狮尸，始识是十狮尸 实十石狮尸，试释是事。

主要困难

- 困难之二：大量未知语言现象
  - 新词、人名、地名、术语等

主要困难

- 困难之二：大量未知语言现象
  - 新词、人名、地名、术语等
    - 如：裸退、非典、失联

主要困难

- 困难之二：大量未知语言现象
  - 新词、人名、地名、术语等
    - 如：裸退、非典、失联
  - 新含义

主要困难

- **困难之二：大量未知语言现象**
  - 新词、人名、地名、术语等
    - 如：裸退、非典、失联
  - 新含义
    - 如：苹果、奔腾、同志、老虎、苍蝇等

主要困难

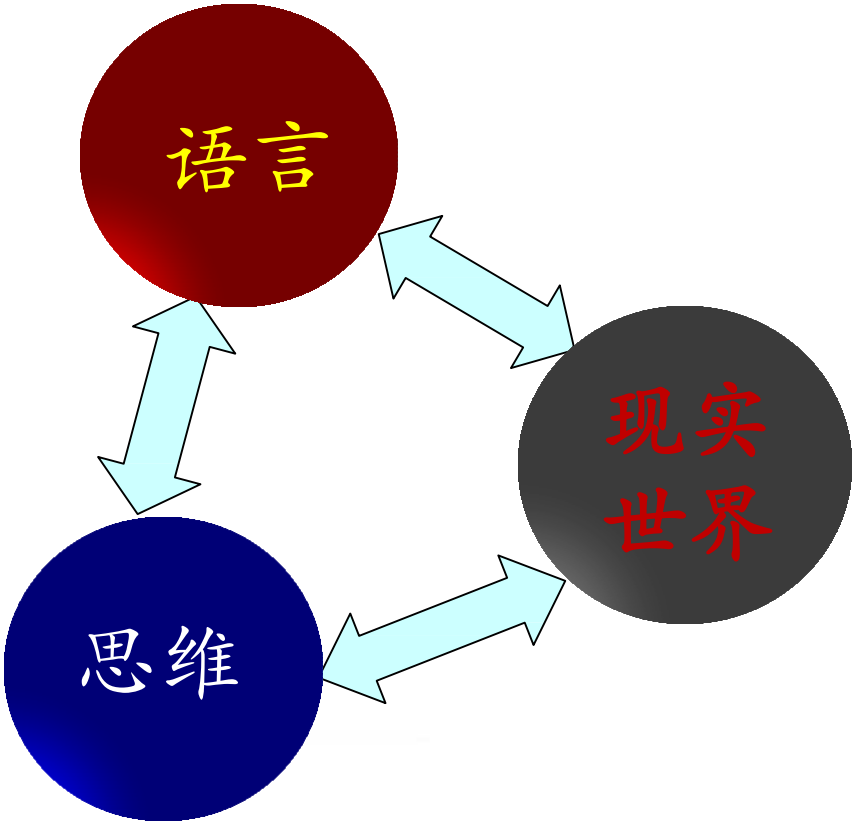
- **困难之二：大量未知语言现象**
  - 新词、人名、地名、术语等
    - 如：裸退、非典、失联
  - 新含义
    - 如：苹果、奔腾、同志、老虎、苍蝇等
  - 新用法和新句型等，尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构

主要困难

- **困难之二：大量未知语言现象**
  - 新词、人名、地名、术语等
    - 如：裸退、非典、失联
  - 新含义
    - 如：苹果、奔腾、同志、老虎、苍蝇等
  - 新用法和新句型等，尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构
    - 如： 被长工资，很中国，百度一下

主要困难

- 人脑理解语言是一个复杂的思维过程
  - 语言学、心理学
  - 逻辑学、认知科学
  - 计算机科学
  - 统计学、信息论
  - 背景知识、常识等
  - ……



Thank you!

权小军 中山大学数据科学与计算机学院