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

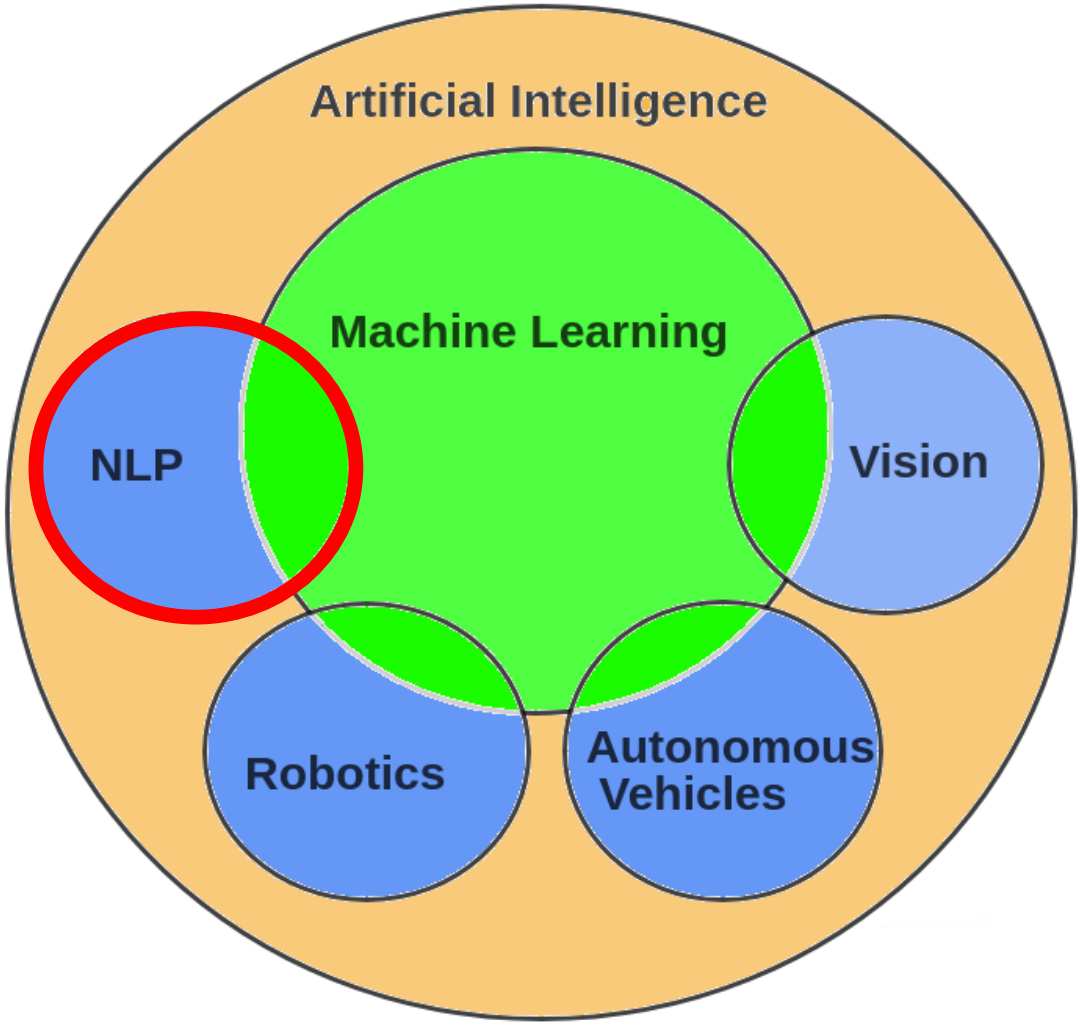
权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

一、基本概念

Artificial Intelligence (AI)？



Natural Language Processing (NLP)

- 如何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?
- 如何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?

Natural Language Processing (NLP)

- 如何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?
- 如何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
Natural Language Processing (NLP)

NATURAL LANGUAGE PROCESSING (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。

基本概念

- **定义1-5: 自然语言处理(Natural Language Processing, NLP)**

自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语 用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

《计算机科学技术百科全书》(宗成庆)

基本概念

- **定义6：中文信息处理 (Chinese Information Processing)**

  针对中文的自然语言处理技术!

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)
    - Systran：http://www.systransoft.com
    - 有道：http://fanyi.youdao.com/

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

问题和困难

基本问题和主要困难

- 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

基本问题和主要困难

- **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

基本问题和主要困难

- **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

基本问题和主要困难

- 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

基本问题和主要困难

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”）。

基本问题和主要困难

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)

基本问题和主要困难

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

基本问题和主要困难

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

基本问题和主要困难

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。

基本问题和主要困难

- **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

二、信息论基础

2.2 信息论基础

❑ 熵(entropy)

香农(Claude Elwood Shannon)于1940年获得MIT数学博士学位和电子工程硕士学位后，于1941年加入了贝尔实验室数学部，并在那里工作了15年。1948年6月和10月，由贝尔实验室出版的《贝尔系统技术》杂志连载了香农博士的文章《通讯的数学原理》，该文奠定了香农信息论的基础。

熵是信息论中重要的基本概念

2.2 信息论基础

❖如果X是一个离散型随机变量，其概率分布为：

$p(x) = P(X = x)$，$x \in X$。X 的熵 $H(X)$ 为：

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \qquad (1)$$

其中，约定 $0 \log 0 = 0$

通常熵的单位为二进制位比特(bit)

2.2 信息论基础

熵又称为自信息(self-information)，表示信源 X 每发一个符号所提供的平均信息量。熵也可以被视为描述一个随机变量的不确定性的数量。一个随机变量的熵越大，它的不确定性越大。那么，正确估计其值的可能性就越小。越不确定的随机变量越需要大的信息量用以确定其值。

2.2 信息论基础

例2-1： 计算下列两种情况下英文(26个字母和1个空格，共27个字符)信息源的熵：

1) 假设27个字符等概率出现；

2) 假设英文字母的概率分布如下：

2.2 信息论基础

字母	空格	E	T	O	A	N	I	R	S
概率	0.1956	0.105	0.072	0.0654	0.063	0.059	0.055	0.054	0.052

字母	H	D	L	C	F	U	M	P	Y
概率	0.047	0.035	0.029	0.023	0.0225	0.0225	0.021	0.0175	0.012

字母	W	G	B	V	K	X	J	Q	Z
概率	0.012	0.011	0.0105	0.008	0.003	0.002	0.001	0.001	0.001

2.2 信息论基础

解:（1）等概率出现情况:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

$$= 27 \times \{-\frac{1}{27} \log_2 \frac{1}{27}\} = \log_2 27 = 4.75 \quad (\text{bits})$$

（2）实际情况:

$$H(X) = -\sum_{i=1}^{27} p(x_i) \log_2 p(x_i) = 4.02 \quad (\text{bits})$$

2.2 信息论基础

解:（1）等概率出现情况:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

说明：考虑了英文字母和空格实际出现的概率后，英文信源的平均不确定性，比把字母和空格看作等概率出现时英文信源的平均不确定性要小。

$$H(X) = -\sum_{i=1}^{27} p(x_i) \log_2 p(x_i) = 4.02 \quad (\text{bits/letter})$$

2.2 信息论基础

❑ 联合熵(joint entropy)

如果 X, Y 是一对离散型随机变量 $X, Y \sim p(x, y)$，X, Y 的联合熵 $H(X, Y)$ 为：

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y) \qquad (2)$$

联合熵是描述一对随机变量平均所需要的信息量

2.2 信息论基础

- 相对熵(relative entropy, 或称 Kullback-Leibler divergence, KL 距离)

  两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵定义为:

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{4}$$

该定义中约定 $0 \log (0/q) = 0$, $p \log (p/0) = \infty$

2.2 信息论基础

❑ 交叉熵(cross entropy)

如果一个随机变量 $X \sim p(x)$，$q(x)$为用于近似$p(x)$ 的概率分布，那么，随机变量X和模型q之间的交叉熵定义为：

$$H(X,q) = H(X) + D(p \| q)$$

$$= -\sum_{x} p(x) \log q(x) \tag{5}$$

交叉熵用以衡量估计模型与真实概率分布之间的差异

2.2 信息论基础

❑ 互信息(mutual information)

如果 $(X, Y) \sim p(x, y)$，X, Y 之间的互信息 $I(X; Y)$ 定义为：

$$I(X; Y) = H(X) - H(X \mid Y) \qquad (6)$$

根据 $H(X)$ 和 $H(X \mid Y)$ 的定义：

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

$$H(X \mid Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x \mid y)$$

2.2 信息论基础

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x \mid y) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \left(\log_2 p(x \mid y) - \log_2 p(x) \right) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \left(\log_2 \frac{p(x \mid y)}{p(x)} \right)
\end{aligned}
$$

$$
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{7}
$$

2.2 信息论基础

例如：**汉语分词问题**

中文分词：为人民服务

为人// 民//服务 或者 为// 人民// 服务

2.2 信息论基础

例如：汉语分词问题

中文分词：为人民服务

为人// 民//服务 或者 为// 人民// 服务

- 利用互信息值估计两个汉字结合的强度：

互信息值越大，表示两个汉字之间的结合越紧密，越可能成词。反之，断开的可能性越大。

三、语言学基础

词性和词法

- 语言学家将词按照相似的语法结构行为和典型的语义类型聚成不同的类，称为**词性**（parts of speech, POS,句法类或语法类）

词性和词法

- 词法(构词过程)
  - 变形: 对词根形式进行系统的修改, 通过加前缀或后缀来指明语法结构的不同, 如单数和复数. 并不显著改变词语的类别和语义, 但修改一些特征, 如时态、数目等.
  - 派生: 缺乏系统化, 通常导致语法类别的根本变化, 且涉及含义的变化

    如: wide→widely, difficult→difficultly( )
  - 复合: 两个或多个词合成一个新词, 如: college degree, overtake, mad cow disease

主要词性

- 名词和代词
- 动词
- 副词、介词
- 连词

句法分析和短语结构歧义

- 句法分析是对输入的文本句子进行分析以得到句子的句法结构的处理过程;
- 对句法结构进行分析，一方面是语言理解的自身需求，句法分析是语言理解的重要一环，另一方面也为其它自然语言处理任务提供支持;
- 例如句法驱动的统计机器翻译需要对源语言或目标语言（或者同时两种语言）进行句法分析;
- 语义分析通常以句法分析的输出结果作为输入以便获得更多的指示信息。

语义

- 语义研究词语的含义、结构和说话的方式
  - 研究单个词的语义词义
  - 单个词的含义怎样联合起来组成句子或更大的单位的含义

主要汉字(文字)编码标准与规范

- ASCII(英文)
- GB2312
- GBK
- GB13000
- GB18030
- BIG5
- Shift_JIS
- ISO/IEC 10646
- Unicode

ASCII码

- 美国信息交换标准编码(美标);
- 用从0到127的128个数字来代表信息的规范编码;
- 包括33个控制码，一个空格码，和94个形象码;
- 形象码中包括了英文大小写字母，阿拉伯数字，标点符号等;
- 国际上大部分电脑的通用编码;

国标、区位、机内码

- 国标：中华人民共和国国家标准信息交换用汉字编码；
- 国标(GB2312-80)表（基本表）把七千余汉字、以及标点符号、外文字母等，排成一个94 行、94列的方阵；

BIG5码

- 针对繁体汉字的编码，在台湾、香港的电脑系统中得到普遍应用；

要想打开一个文本文件，就必须知道它的编码方式，否则用错误的编码方式解读，就会出现乱码。为什么电子邮件常常出现乱码？就是因为发信人和收信人使用的编码方式不一样。

Unicode

- 英文Universal Code的缩略语;
- 统一编码;
- 是对国际标准ISO/IEC 10646编码的一种称谓;
- 是一个企业联盟集团的名称,由美国的HP、Microsoft、IBM、Apple等几家知名的大型计算机企业所组成,成立该集团的宗旨就是要推进多文种的统一编码;

什么是UTF?

- Unicode transformation format;
- 从Unicode码点到唯一字节序列的映射算 法，一一映射，保证无损转换;
- UTF是Unicode的实现方式;

Thank you!

权小军 中山大学数据科学与计算机学院