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

应用展示

静夜思　作者：李白

床前明月光，疑是地上霜。
举头望明月，低头思故乡。

人工智能写诗竟以假乱真？清华机器人与清华才女央视巅峰对决

2017-12-25 17:34

编者按

近日，清华矣晓沅团队开发的作诗机器人「九歌」亮相央视黄金档节目《机智过人》。按照规定，「九歌」接受图灵测试：它与三位人类检验员一起作诗，由48位投票团成员判断哪首为机器人所做，如果两轮测试中，得票最多的都不是「九歌」，则通过测试。

结果「九歌」成功混淆视听，先后淘汰了北大陈更与武大李四维。**场上人类诗人代表只剩清华核研院博士齐妙，两位「清华校友」舞台上以诗论道，舞台下同样诗意盎然。**

芳草比君子，诗人情有由

集句詩 绝句 藏頭詩 詞

五言 七言

请输入主题词 作诗

http://jiuge.thunlp.org/

关键词：双鸭山

双鸭山

千 山 万 叠 翠 屏 开

古 寺 钟 声 送 客 来

明 日 扁 舟 江 上 去

桃 花 流 水 几 时 回

关键词：中山大学

中山大学

六 月 中 山 雨

秋 风 一 夜 砧

凄 凉 江 上 客

寂 寞 蓟 门 深

关键词：中秋

中秋

一 雨 中 秋 月

千 山 隔 暮 砧

故 人 何 处 问

翘 首 望 乡 心

关键词：中秋节

中秋节

黄 金 台 上 中 秋 日

一 曲 琵 琶 十 二 砧

最 是 嫦 娥 明 月 夜

玉 箫 吹 彻 碧 云 深

课程回顾

课程回顾

一、概率论基础

- 概率(probability)
- 极大似然估计(maximum likelihood estimation)
- 条件概率(conditional probability)
- 全概率公式(full probability)
- 贝叶斯法则(Bayes' theorem)
- 二项式分布(binomial distribution)
- 期望(expectation)
- 方差(variance)

课程回顾

二、信息论基础

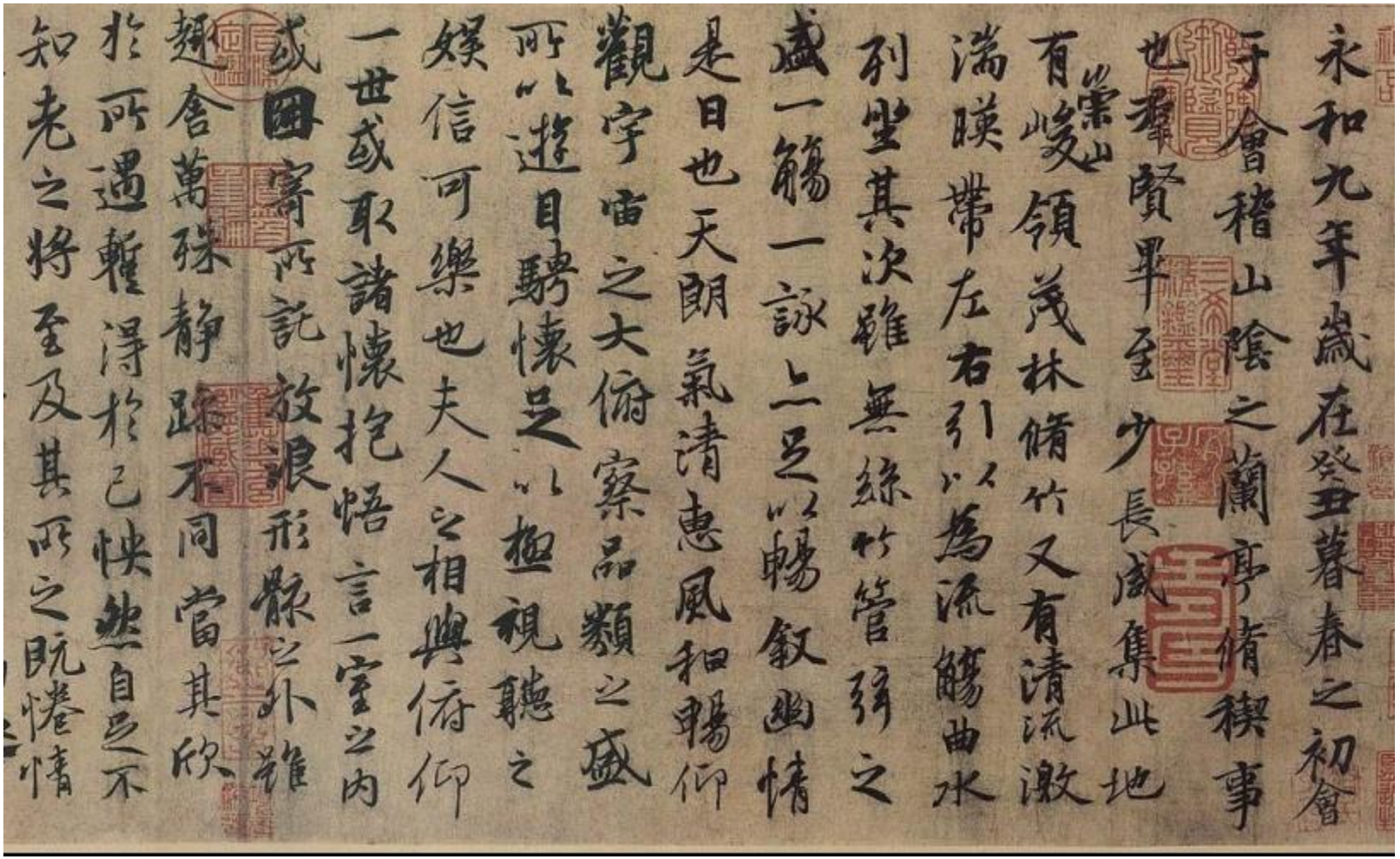
- 熵(entropy)
- 联合熵(joint entropy)
- 条件熵(conditional entropy)
- 相对熵(KL 距离)
- 交叉熵(cross entropy)
- 互信息(mutual information)

Lecture 3: 中文词法分析（上）

内容

- 词法分析任务：从字符串到词串
- 中文词法分析的意义
- 中文文本分词面对的问题

3.1 词法分析任务：从字符串到词串


永和九年歲在癸丑暮春之初會
于會稽山陰之蘭亭脩禊事
也羣賢畢至少長咸集此地
有崇山峻領茂林脩竹又有清流激
湍映帶左右引以爲流觴曲水
列坐其次雖無絲竹管弦之
盛一觴一詠亦足以暢敘幽情
是日也天朗氣清惠風和暢仰
觀宇宙之大俯察品類之盛
所以遊目騁懷足以極視聽之
娛信可樂也夫人之相與俯仰
一世或取諸懷抱悟言一室之内
或因寄所託放浪形骸之外雖
趣舍萬殊靜躁不同當其欣
於所遇暫得於己快然自足不
知老之將至及其所之既惓情

文章强调，人民代表大会制度是中国特色社会主义制度的重要组成部分，也是支撑中国国家治理体系和治理能力的根本政治制度。新形势下，我们要高举人民民主的旗帜，毫不动摇坚持人民代表大会制度，也要与时俱进完善人民代表大会制度，坚定不移走中国特色社会主义政治发展道路，继续推进社会主义民主政治建设、发展社会主义政治文明。

从字符串到词串

输入：字符串

学生人数多又能保证质量的才是好学校。

从字符串到词串

输入：字符串

学生人数多又能保证质量的才是好学校。

输出：词串

学生|人数|多|又|能|保证|质量|的|才|是|好|学校|。

从字符串到词串

汉语的自然书面文本词与词之间无空格分开，因此，在汉语书面语的处理中（比如词频统计、句子结构分析、语义理解等），首先碰到的就是词的切分问题。

从字符串到词串（英文）

- Tokenization：把字符串变为词串

  I'm a student -> I| 'm | a | student

- Lemmatization Word Stemming：对词进行内部结构和形式分析

  took -> take + ed (past tense)

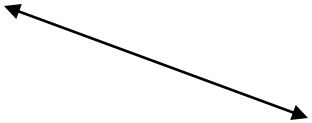
从字符串到词串（续）

中文： 中山大学在广州.

English： Sun Yat-sen University is in Guangzhou.

从字符串到词串（续）

中文： 中山大学在广州.



English： Sun Yat-sen University is in Guangzhou.

3.2 中文词法分析的意义

中文词法分析的意义

文本分词是各个层次的自然语言处理任务的基础

1. 文语转换Text-to-speech
2. 文本校对 Chinese Text Correction
3. 文本检索Information Retrieval
4. 词频统计、句法分析、机器翻译、……

文语转换示例

1. 树种得少，树种就少。
2. 你知道世界上有多少人种吗?
3. 这种树你知道世界上有多少人种吗?
4. 这些大学生为什么不看重大城市户口。
5. 这些大学生为什么不看重大疾病保险的说明书?
6. 可敬的哥争分夺秒送病人.
7. 好好儿想想有几米.

3.3 文本分词面对的问题

文本分词面对的问题

- 什么是中文的“词”
- 分词歧义
- 未登录词识别

什么是“词”

- **语法学定义：** 能够独立运用的最小的音义结合体
- **语料库定义：** 枚举“词例”（token）

分词规范

词的内涵式定义：结合紧密

词的可操作定义：使用稳定

S 如果是一个W，则：

- S 内部子串粘合度高
- S 外部环境替换度高
- S 本身频度高

- 刘源 等（1994）《信息处理用现代汉语分词规范及自动分词方法》，清华大学出版社、广西科学技术出版社，1994年版。
- 黄居仁、陈克健 等（1997）《信息处理用中文分词规范设计理念及规范内容》，载《语言文字应用》1997年第1期。
- 《信息处理用汉语分词规范》 GB/T13715-92，中国标准出版社，1993
- 《资讯处理用中文分词规范》 台湾中研院，1995
- 《人民日报》语料库词语切分规范 北大计算语言所，1999

不同的人对“词”的认识有差异

- 6人对100句（4372）字进行人工分词，然后两两比较认同率：

不同的人对“词”的认识有差异

- 6人对100句（4372）字进行人工分词，然后两两比较认同率：

	M2	M3	T1	T2	T3
M1	0.77	**0.69**	0.71	**0.69**	0.70
M2		0.72	0.73	0.71	0.70
M3			**0.89**	0.87	0.80
T1				0.88	0.82
T2					0.78

不同的人对“词”的认识有差异

- 6人对100句（4372）字进行人工分词，然后两两比较认同率：

平均值0.76

	M2	M3	T1	T2	T3
M1	0.77	**0.69**	0.71	**0.69**	0.70
M2		0.72	0.73	0.71	0.70
M3			**0.89**	0.87	0.80
T1				0.88	0.82
T2					0.78

文本分词中的歧义

- 交集型歧义

例1： 张店区大学生不看重大城市的户口本

文本分词中的歧义

- 交集型歧义

例1： 张店区大学生不看重大城市的户口本

张店区 大学生 不 看 | 重大 城市 的 户口本
张店区 大学生 不 看重 | 大 城市 的 户口本

文本分词中的歧义

- 组合型歧义

例2： 你认为学生会听老师的吗

文本分词中的歧义

- 组合型歧义

例2： 你认为学生会听老师的吗

你 认 为 | 学 生 会 | 听 老 师 的 吗
你 认 为 | 学 生 | 会 | 听 老 师 的 吗

文本分词中的歧义

- 混合型歧义

例3： 只有雷人才能吸引人

只有 雷人| 才能 吸引 人
只有 雷 |人才| 能 吸引 人
只有 雷| 人| 才| 能 吸引 人

文本分词中的歧义

交集型歧义：组合型歧义= 1: 22 [1]

（语料规模：17,547字）

[1] 刘挺、王开铸，1998，关于歧义字段切分的思考与实验。《中文信息学报》第2期，63-64页。 孙茂松、邹嘉彦，2001，汉语自动分词研究述评，《当代语言学》2001年第1期，22-32页。

文本分词中的歧义

真歧义

- 确实能在真实语料中发现多种切分形式
- 比如“应用于”、“地面积”、“解除了”

伪歧义

- 虽然有多种切分可能性，但在真实语料中往往取其中一种切分形式
- 比如“挨批评”、“市政府”、“太平淡”

文本分词中的歧义

分词歧义的四个层级[1]

- **词法歧义，根据词法知识可以排除的歧义**

  例：“用方块图形式加以描述”（占84.1%）

- **句法歧义，根据句法知识可以排除的歧义**

  例：“他一阵风似的跑了”（占10.8%）

  例：“学生会写文章”（占3.4%）

- **语义歧义，根据语义知识可以排除的歧义**

- **语用歧义，根据语用知识可以排除的歧义**

  例：“美国会采取措施制裁伊拉克”（占1.7%）

[1] 何克抗等，1991，《书面汉语自动分词专家系统设计原理》，载《中文信息学报》，1991年第2期。

交集型歧义的链长

- 交集型歧义字段中含有交集字段的个数，称为链长
  - 链长为1： 和尚未
  - 链长为2： 结合成分
  - 链长为3： 为人民工作
  - 链长为4： 中国产品质量
  - 链长为5： 鞭炮声响彻夜空
  - 链长为6： 努力学习语法规则
  - 链长为7： 中国企业主要求解决
  - 链长为8： 治理解放大道路面积水
  - ……

交集型歧义的链长

- 交集型歧义字段中含有交集字段的个数，称为链长
  - 链长为1：和尚未
  - 链长为2：结合成 合成分
  - 链长为3：为人民 人民工 民工作
  - 链长为4：中国产 国产品 产品质 品质量
  - 链长为5：鞭炮声 炮声响 声响彻 响彻夜 彻夜空
  - 链长为6：……
  - 链长为7：中国企业主要求解决
  - 链长为8：治理解放大道路面积水
  - ……

汉语真实文本中分词歧义的分布情况

在一个1亿字真实汉语语料库中抽取出的前4,619个高频交集型歧义切分覆盖了该语料库中全部交集型歧义切分的59.20%，其中4279个属伪歧义，占92.63%，如“和软件”、“充分发挥”、“情不自禁地”，这部分伪歧义类型的实例对语料的覆盖率高达53.35%。[1]

[1] 孙茂松等，1999，《高频最大交集型歧义切分字段在汉语自动分词中的作用》，载《中文信息学 报》1999年第1期。

汉语真实文本中分词歧义的分布情况

1. 汉族人名、地名	雪村、老张、中关村
2. 外族人名、地名	横路静二、突尼斯
3. 中外组织机构单位名称	联合国教科文组织
4. 商品品牌名	非常可乐、苹果iPad
5. 专业术语	有限状态自动机、三分球
6. 新词语	秒杀、蚁族、羊羔体
7. 缩略语	人影办、两会、北医三院
8. 汉语重叠形式、离合词等	高高兴兴、幽了他一默
9. 含数字，非汉字字符的词	2014年3月3日 IC卡 D座 T台

汉语真实文本中分词歧义的分布情况

- **较成熟**
  - 中国人名、译名
  - 中国地名
- **较困难**
  - 商标字号
  - 机构名
- **很困难**
  - 专业术语
  - 缩略语
  - 新词语

中国人名的内部构成情况

- 在汉语的未登录词中，中国人名是规律性最强，也是最 容易识别的一类；
- 中国人名一般由以下部分组合而成：
  - 姓，例：张、王、李、刘、诸葛、西门、范徐丽泰
  - 名，例：李素丽，张华平，王杰、诸葛亮
  - 前缀，例：老王，小李
  - 后缀，例：王老，赵总
- 中国人名各组成部分用字比较有规律

中国人名

- 中国人名的组合模式
  - 姓＋名
  - 姓
  - 名
  - 前缀＋姓
  - 姓＋后缀
  - 姓＋姓＋名（海外已婚妇女）

姓、名均可再分单字和双字

中国人名

- 中国人名外部特征：
  - **身份词：**
    - 前：工人、教师、影星、犯人
    - 后：先生、同志
    - 前后：校长、经理、主任、医生
  - **地名或机构名：**
    - 前：静海县大丘庄禹作敏
  - **的字结构**
    - 前：年过七旬的王贵芝
  - **动作词**
    - 前：批评，逮捕，选举
    - 后：说，表示，吃，结婚
  - ……

中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
  - 姓氏：于，马，黄，张，向，常，高
  - 名字：*周鹏和*同学，*周鹏*和同学

中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
  - 姓氏：于，马，黄，张，向，常，高
  - 名字：*周鹏和*同学，*周鹏*和同学
- **人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词**
  - [王国]维、[高峰]、[汪洋]、张[朝阳]、冯[胜利]

中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
  - 姓氏：于，马，黄，张，向，常，高
  - 名字：*周鹏和*同学，*周鹏*和同学
- 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
  - [王国]维、[高峰]、[汪洋]、张[朝阳]、冯[胜利]
- **人名与其上下文组合成词**
  - 这里*[有关]天培*的壮烈
  - *费孝[通向]*人大常委会提交书面报告
  - *邓颖[超生]*前使用过的物品

中国人名识别的难点

- 一些高频姓名用字在非姓名中也是高频字
  - 姓氏：于，马，黄，张，向，常，高
  - 名字：*周鹏和*同学，*周鹏*和同学
- 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
  - [王国]维、[高峰]、[汪洋]、张[朝阳]、冯[胜利]
- 人名与其上下文组合成词
  - 这里*[有关]天培*的壮烈
  - *费孝[通向]*人大常委会提交书面报告
  - *邓颖[超生]*前使用过的物品
- **人名地名冲突**
  - 河北省*刘庄*

Thank you!

权小军 中山大学数据科学与计算机学院