自然语言处理及应用

Natural Language Processing
#L1

Introduction

袁彩霞

yuancx@bupt.edu.cn

人工智能学院 智能科学与技术中心

Outline

- 什么是自然语言处理?
- 为什么要学习自然语言处理?
- 当前的自然语言处理技术水平如何?
- 课程计划及考核方式
- 推荐参考书

什么是“自然语言处理”

- 用计算机处理自然语言

- 用计算机……（计算机是主体）
- 自然语言（什么是“自然语言”）
- 处理（什么是“处理”）

什么是“自然语言”

- 汉语、英语……等现存人类使用的语言
- 手语？

- 自然语言 vs 人工语言

什么是“处理”

- 例子：拼写检查

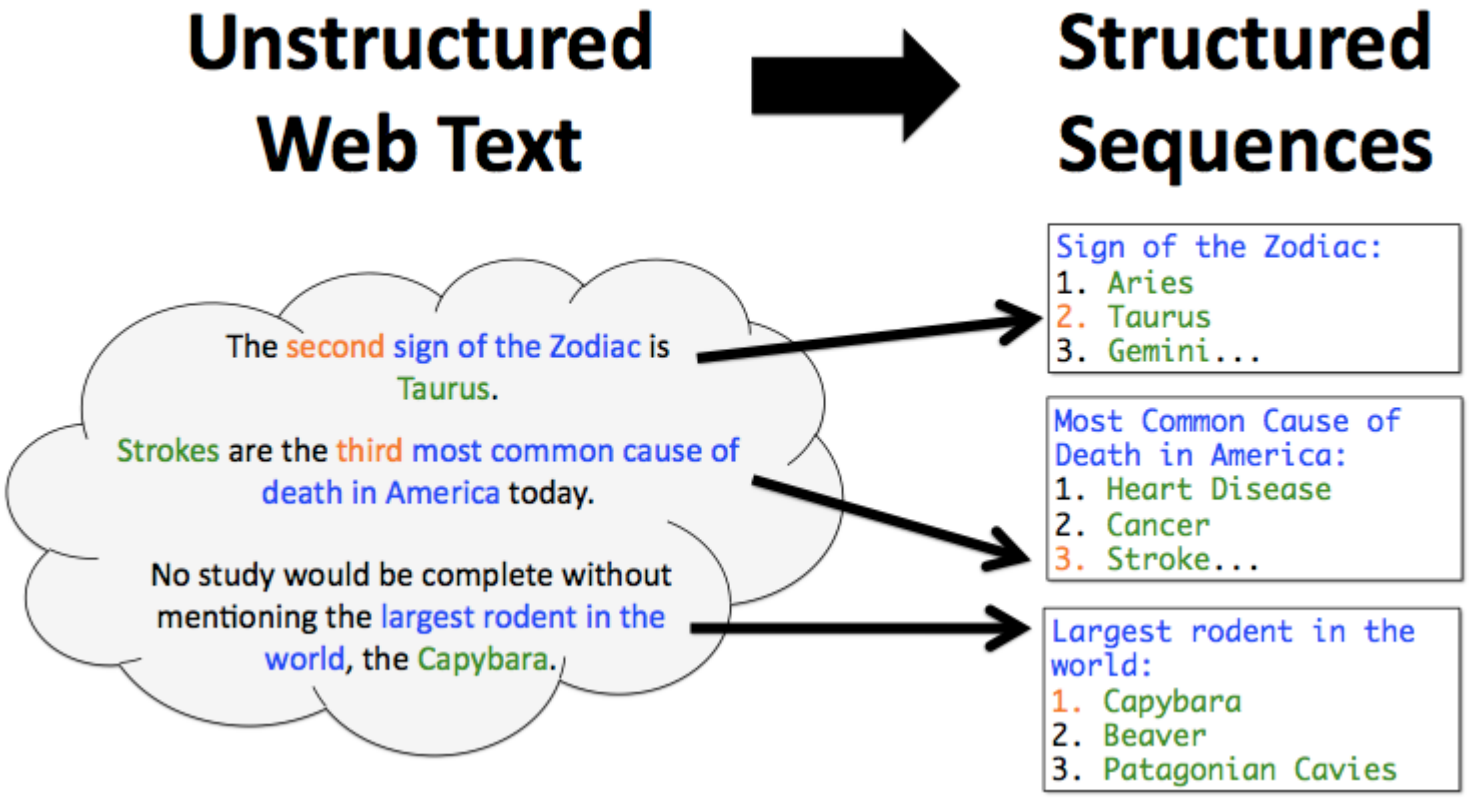
*The **Turing test** is a test of a machine's abiliy to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human. In the original illustrative example, a human judge engages in natural language conversations with a human and a machine designed to generate perfomance indistinguishable frm that of a human being. All participants are separated from one another. If the judge cannot reliably tell the machine from the hman, the machine is said to have passed the test. The test does not check the ability to give the corect answer to questions; it checks how closely the answer resembles typical human answers. The conersation is limited to a text-only channel such as a computer keyboard and screen so that the result is not dependent on the machine's ability to render words into audio*

什么是“处理”

- 例子：搜索引擎

什么是“处理”

- 例子：信息抽取



什么是“处理”

- 例子：机器翻译

Baidu 翻译

如果把里约与东京联系到一起？日本人煞费苦心。从北岛康介、到大空翼、再到机器猫，一颗红球在代表日本的各种经典形象中传递。

If Rio and Tokyo linked together? Japanese pains. From Kosuke Kitajima, to the air wing, and then to transfer machine cat, a red ball in a variety of classic image representative in japan.

双语对照

什么是“处理”

- 例子：人机口语对话
  - SIRI

什么是“处理”

- 总结：机器像人一样完成基于自然语言的活动
- 自然语言处理（Natural Language Processing, NLP）主要包括：
  - 自然语言的理解（Natural Language Understanding, NLU）
  - 自然语言的生成（Natural Language Generation, NLG）

NLP技术简史

- 起动—1950's
  - Warren Weaver 1947: Machine Translation (MT)
- 低潮—1960's
  - ALPAC 1966: MT is impossible in near future.
- 理论的奠基—1970's
  - Noam Chomsky's 语言学
- 理性主义—1980's
  - 语法理论的繁荣
- 经验主义—1990's
  - IBM model for MT
  - Web信息爆炸导致的IR兴起
- 需求带动的繁荣—21C

What is Nearby NLP?

- 计算语言学（Computational Linguistics, CL）
- 人类语言技术（Human Language Technology, HLT）
- 语言工程（Language Engineering, LE）

- 各有侧重
- 共通点：
  - Computation + Linguistics

What is Nearby NLP?

- 从外部关系看NLP
  - 人工智能和语言学的交叉学科
    - 人工智能的分支：自然语言处理(应用方面)
    - 语言学的分支：计算语言学(理论方面)
  - 基础工具
    - 数学：
      - 概率论、随机过程、矩阵论、最优化理论
    - 计算机：
      - 数据结构、高级计算机语言

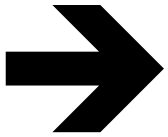
What is Nearby NLP?

- 语言学对NLP的巨大作用：
  - 发现语言问题
    - 汉语词之间无空格 ➔ 切分问题
    - 英语词有变形 ➔ 形态分析问题
    - ……
  - 提供建模语言现象时的洞见
    - 句子是从词单元组成再大一点的单元，再由这些大单元进一步组成的 ➔ 句子短语结构模型
    - ……

NLP from the first bird view

- 问题的引入
  - 从语言学或应用需求出发为NLP引入研究问题
- 数学形式化
  - 用数学工具将引入的问题形式化(建立数学模型)
- 计算形式化
  - 使得问题可以通过算法进行计算(建立计算模型)
- 编程实现
  - 使用某种程序语言实现算法并在计算机上运行
- 评估
  - 评估计算机的运算结果：问题、模型、程序

以汉语切分(Chinese Segmentation)为例



问题的引入

- 输入：内塔尼亚胡说的话在美国会引起强烈反响。
  - 这句话什么意思？
  - 为理解这句话，首先需要知道这句话是由哪些词构成的，几种可能：
    - 内塔尼亚胡/说/的/话/在/美/国会/引起/强烈/反响
    - 内塔尼亚胡/说/的/话/在/美国/会/引起/强烈/反响
    - 内塔尼亚/胡说/的/话/在/美国/会/引起/强烈/反响
    - 内塔尼亚/胡说/的/话/在/美/国会/引起/强烈/反响
    - ……
- 问题：
  - 不同的词切分方案（切分歧义）会导致同一个字符串表示不同的句子，如何切出一个合理的词序列？
- 如何求解该问题？➔数学建模

建立数学模型

- 一个可能的模型:
  - 设W是词典，T是字串，K是知识库
  - 则对于任意t∈T，一个切分就是一个函数*f*，*f*将字串t映射为一个词串，即：

$$f(t|k)=w_1w_2\ldots w_n$$

    - 其中$w_1,w_2,\ldots\ldots w_n \in W$, $k \in K$
- 例如：
  - *f*(内塔尼亚胡说的话|K)=内塔尼亚胡/说/的/话/
- 如何让计算机求解该映射？➔计算模型

建立计算模型

- 可能的计算模型
  - 前向最大匹配(FMM: Forward Maximum Match)算法
    - 内塔尼亚胡/说/的/话/在/美国/会/引起/强烈/反响
  - 后向最大匹配(BMM: Backward Maximum Match)算法
    - 内/塔/尼/亚/胡说/的/话/在/美/国会/引起/强烈/反响

- 如何让计算机实施这种算法？➔编程

编程实现

- 算法的程序实现
  - 例如，对于FMM算法的实施
    - 选择编程语言
    - 设计数据结构
    - 程序结构
    - ……
- 计算机按此算法进行切分的结果如何？➔评估

评估

- 算法评估：如何评估
  - 评估材料：在什么数据上评估
    - NLP算法评估的特点：评估结果可能与选择的评估材料有关
  - 评估准则：
    - 标准是什么
    - 度量指标是什么
- 如果结果不理想
  - 重来（问题引入、数学建模、计算模型、评估方案）

Outline

- 什么是自然语言处理?
- 为什么要学习自然语言处理?
- 当前的自然语言处理技术水平如何?
- 课程计划及考核方式
- 推荐参考书

NLP应用价值

- 语音输入
- 信息检索
- 机器翻译
- 信息推荐
- …

NLP的应用几乎无处不在……

NLP应用价值

【社招】【校招】【内推】【部门直推】联想UE/UX 交互设计师	11:25:53
【校招】【内推】【社招】【高德地图】	11:25:29
【内推】【社招】【实习】【微软】活少钱多不加班福利多	11:25:14
【内推】【社招】【校招】【实习】【字节跳动】	11:24:59
【内推】【社招】百度智能云	11:23:33
【内推】【校招】阿里云2022届实习(全程跟踪，有问必答)	11:16:07
【校招】【内推】【实习】【阿里】阿里云混合云春季实习生内推	11:07:28
【内推】【校招】【360】360集团2021校园春招！	11:05:19
【实习】【内推】【校招】 阿里巴巴-天猫超市实习招聘	10:57:51
【内推】【实习】阿里巴巴搜索广告算法团队，组内直推	10:57:18
【内推】【校招】【社招】【字节跳动】【抖音电商】	10:56:42
【社招】【校招】【内推】【部门直推】字节跳动-教育中台	10:52:12
【内推】【社招】【校招】美团优选内推了！！！	10:50:57
【内推】【实习】腾讯暑期实习内推	10:46:11
【校招】【内推】【拼多多】【成长空间大】	10:44:29

NLP学术价值

- 有助于揭示人类语言信息处理的奥秘
  - 语言处理的唯一原型是人
- 有助于揭示人类思维的本质
  - 语言是思维的外壳

NLP是个难题

- 歧义（ambiguity）大量存在
  - I made her duck (meanings!)
    - I cooked waterfowl for her benefit (to eat)
    - I cooked waterfowl belonging to her
    - I created the (plaster?) waterfowl she owns
    - I caused her to quickly lower her head or body
    - I waved my magic wand and turned her into undifferentiated waterfowl
  - I made her duck (Phonetics!)
    - I mate or duck
    - I'm eight or duck
    - Eye maid; her duck
    - Aye mate, her duck
    - I maid her duck
    - I'm aid her duck
    - I mate her duck
    - I mate or duck

NLP是个难题

- 新词及新用法层出不穷
  - 不明觉厉、十动然拒、男默女泪、火钳刘明、累觉不爱、然并卵、喜大普奔……
  - 槽点、颜值、雷人、炒婚、麦粉、山寨、伦家、骚年、熟女……
  - 小鲜肉、蛋白质、白骨精、有病、不早朝、通心粉、恐龙……
  - gay里gay气、心机boy、打call……

NLP是个难题

- 人类对自身运用自然语言的机制还不甚了解
  - 人类的自然语言运用是自然语言处理的唯一原型
  - "It is psychological and neurobiological factors that enable humans to acquire, use, comprehend and produce language"
  - 但这一内部机制尚不明确
- **→ignore humans, learn from language data?**

Outline

- 什么是自然语言处理?
- 为什么要学习自然语言处理?
- 当前的自然语言处理技术水平如何?
- 课程计划及考核方式
- 推荐参考书

最近的代表性应用成果

- 搜索：http://www.wolframalpha.com/
- 翻译：http://translate.google.cn/
- 推荐：Amazon图书推荐…
- 人机口语对话：
  - Apple的Siri
  - 微软的小冰、Cortana
  - Amazon的Alex
  - 小米的小爱同学
  - ……
- 开放领域问答：
  - IBM Walston
  - Eugene Goostman

然而，困难重重……

- 人机对话：
  - “如果没有中通就发圆通，如果没有圆通就发申通，如果没有申通就发顺丰，如果没有顺丰，就发韵达。无论如何，不发EMS。”
  - …
- 幽默文学：
  - 刘备和诸葛亮告别——
    - 刘备：じゃね，亮。
    - 诸葛亮：贾乃亮是谁？
  - 乔峰和慕容复对打，现场响起了雷鸣般的掌声。

Outline

- 什么是自然语言处理?
- 为什么要学习自然语言处理?
- 当前的自然语言处理技术水平如何?
- 课程计划及考核方式
- 推荐参考书

课程目的

- NLP基础技术：
  - 重点介绍和分析NLP中几个经典的模型、算法，探讨这些模型、算法的有效运用。

- NLP应用系统：
  - 应能胜任基本的NLP研发工作，综合运用NLP技术解决实际问题。

课程安排

- 基础技术：
  - 词法分析
    - 形态分析：一个句子由哪些词组成、词的结构如何
    - 词性标注：词的句法类别是什么
  - 句法分析
    - 词以何种结构组成句子
  - 语义分析
    - 词和句子的意思是什么
  - 文本分析
    - 文本的意思（主题）是什么
- 应用系统：
  - 文本分类和聚类
  - 信息检索
  - 机器翻译
  - 人机对话
  - ……

课程安排

大纲	内容	讲解时间
绪论	什么是/为什么/如何/评估…	2 hrs
词法分析之 形态分析	N元语言模型(N-gram) 神经网络语言模型	2hrs 2hrs
词法分析之 词性标注	词性标注 隐马尔科夫模型(HMM) 命名实体识别	0.5hr 1.5hrs
句法分析	上下文无关语法/CKY/Earley算法 概率上下文无关语法 依存分析	2hrs
语义分析	句子的语义表示 循环神经网络 Transformer及其它的预训练模型	6hrs
文本分析	文本表示和主题模型	6hrs
应用技术	文本分类/文本聚类/情感分析/信息检索/机器翻译/对话系统	8hrs
总结与展望	NLP beyond Language	2hrs

课程安排

	语法	语义	语用
词	分词、 词性标注、 命名实体识别、 …	语言模型、 词表示、 词义消歧、 …	情感分析、 程度分析、 语义发现、 …
句子	句子切分、 短语结构分析、 依存分析、 功能组块分析…	语言模型、 句子表示、 问题理解、 知识抽取 …	意图分析、 情感分析、 知识推理…
篇章	向量空间模型、 关键词提取、 修辞结构理论、 …	篇章语义分析、 文本表示、 主题模型、 …	会话分析、 阅读理解、 言语行为分析…

考核方式

- Assignments during the course (100%)
  - 10+个不同分值的候选题目，须从中至少选够100分（亦也可自行命题，需提前找老师确认）
  - 期末时可以做Oral presentation
- Others(20%)
  - 平时成绩

部分以往完成的作业

- 文本信息搜索
  - 人物信息查询系统
  - 新浪微博主题垂直搜索
  - 基于内容的文章推荐系统
  - 娱乐明星信息查询系统
- 文本分类
  - 针对特定主题的网页信息过滤
  - 短信息的分类和处理
  - 基于电影剧情的电影分类系统

部分以往完成的作业

- 文本信息抽取
  - 新闻摘要\关键词抽取\比赛信息提取
  - 简历信息抽取
  - Web热点检测\旅游热点TOP10\新闻热词top10
  - 寻TA！ —婚介匹配系统
- 对话
  - 智能口语命令识别系统
  - 基于情感分析的人机对话系统
  - 手机短信自动回复
  - 美食人机对话系统\简易点菜系统
  - 历史人物问答系统

部分以往完成的作业

- 其它：
  - 测测你的同步率：有哪个群和我们群的同步率比较高？(话题最相似)
  - 基于爬虫和自然语言处理的商品评论分析系统\评论自动分级系统
  - 基于自然语言处理的服装搭配系统
  - 基于计算器语境的汉语语音识别后的文本检错方法
  - 选词填空系统
  - 测测你的文艺指数：基于用户微博计算得到用户的“文艺指数”
  - 基于NLP&决策树的dota英雄推测游戏

Outline

- 什么是自然语言处理?
- 为什么要学习自然语言处理?
- 当前的自然语言处理技术水平如何?
- 课程计划及考核方式
- 推荐参考书

教材和参考读物

- Textbook and readings
  - Slides and notes
  - 宗成庆. 统计自然语言处理（第二版）, 清华大学出版社, 2013
  - 王小捷, 常宝宝.自然语言处理技术基础, 北京邮电大学出版社, 2002.
  - Daniel Jurafsky and James H. Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Edition, 2007, Prentice-Hall, available online.
  - Chris Manning and Hinrich Schütze. Foundations of Statistical Language Processing, 1999, MIT Press, available online.
- Online courses:
  - Prof. Manning: CS224n, Dr. Socher: CS224d

Homework

- 阅读：
  - 宗：2.1\2.2(简单数学基础)
  - Jurafsky：Chapter 1 (NLP概述)
- 学习：
  - Python语言

Next lecture

- 语言模型：
  - 预习：贝叶斯法则，全概率公式及链式法则、极大似然估计

Thank you!