第一章　绪 论

王峰
华东师大计算机系

自然语言处理
Natural Language Processing

- 参考教材
  - 俞士汶，常宝宝，詹卫东，《*计算语言学概论*》，商务印书馆。
  - 宗成庆，《*统计自然语言处理*》，清华大学出版社。
  - Steven Bird, Ewan Klein, and Edward Loper, ***Natural Language Processing with Python***, Published by O'Reilly Media Inc.
- 考核方式与评价结构比例：
  - 期末闭卷考试，考试成绩占 60 %，平时成绩 40 % (包括考勤、作业10%+上机30%)。
- 教师
  - 王峰 (fwang@cs.ecnu.edu.cn)
  - 理科大楼B713

课程内容

- 分词
- 词性标注
- 句法结构分析
- 语义分析
- 语篇分析
- 机器翻译

上机实验

- 基础工具的使用
- 文本关键词提取
- 文本分类与聚类
- 中文词语的语义相关度计算
- 实体关系挖掘
- 文本情感分析

提 纲

- 问题的提出
- NLP的概念
- NLP的应用
- NLP的发展历史
- NLP的研究方法

问题的提出

我们可以期待，总有一天机器会同人在所有的智能领域里竞争起来。但是，如何开始呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动做为最好的出发点。不过，还有一种办法也应加以考虑，就是为机器配备具有智能的、可用钱买到的意识器官，然后，教这种机器理解并且说英语。这个过程可以仿效通常小孩子学话的方式进行。我不能确定到底哪个出发点更好，但应该都值得一试。

--- A. M. Turing, *Computing Machinery and Intelligence,* Mind 49:433-460, 1950

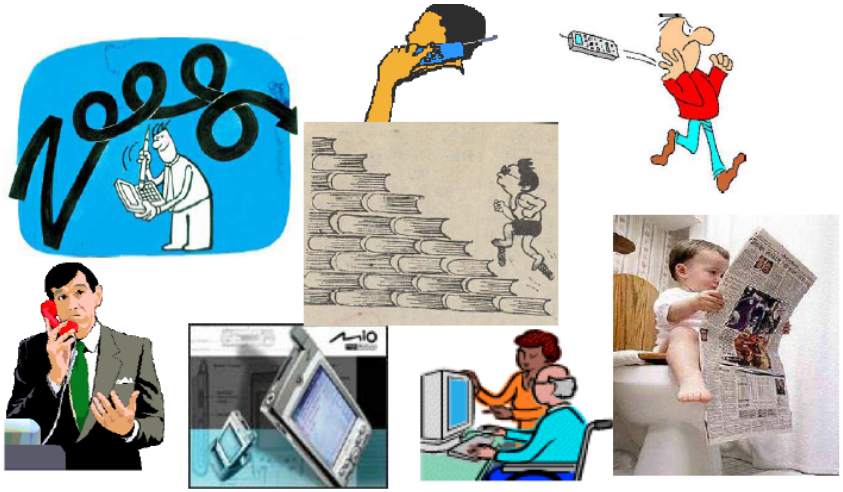
问题的提出

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

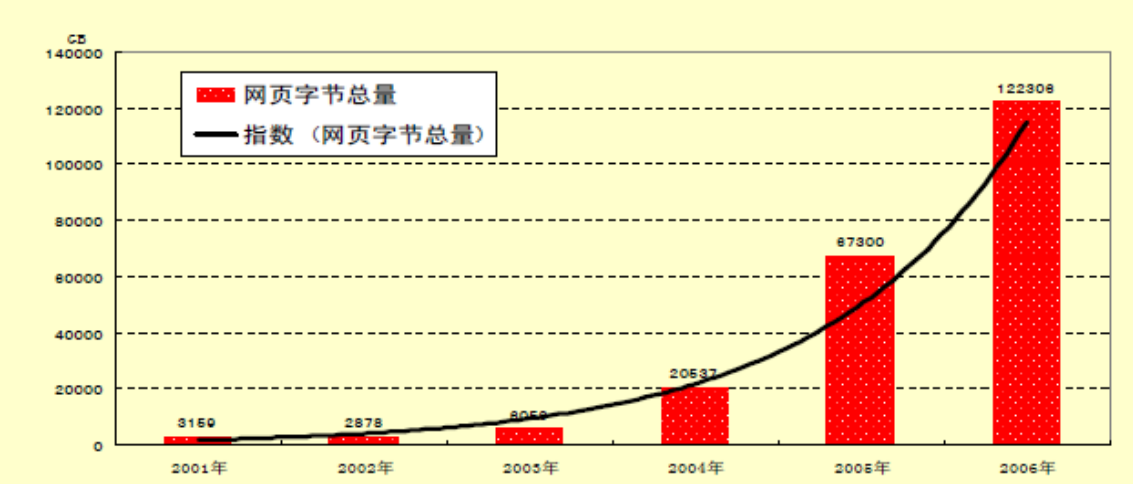
信息爆炸

- 无处不在的网络、通讯和堆积如山的文档，构成了当今社会信息爆炸的基本特征。
- 现代化的信息传播手段给人们的生活和工作带来极大便利的同时，也使人们面临许多难以克服的困难和障碍。
- 有关专家指出，语言障碍是21世纪社会全球化所面临的主要困难之一。



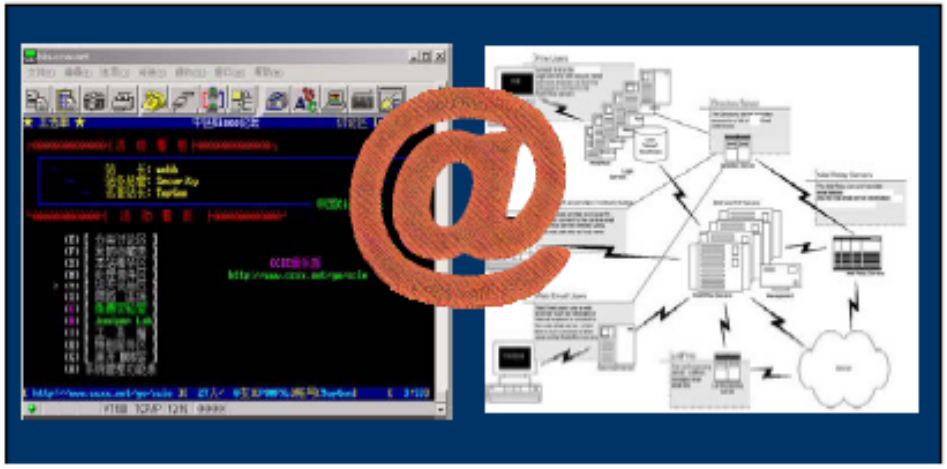
信息检索

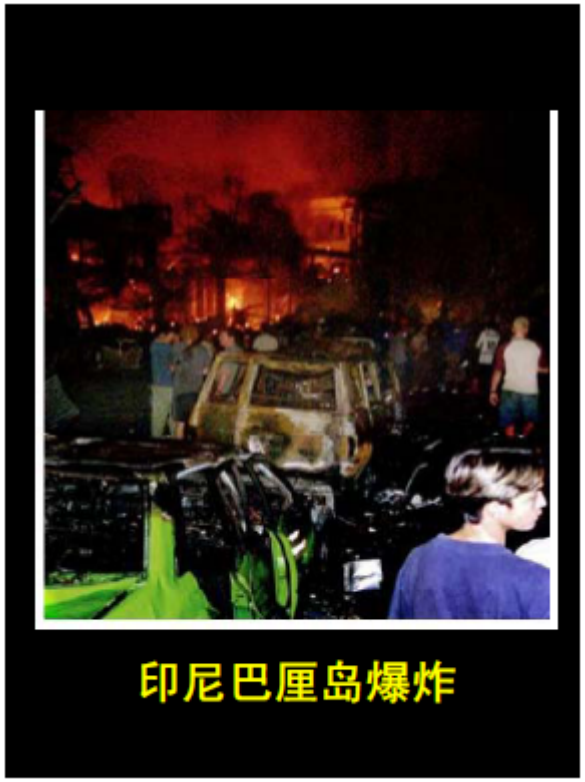
- 全世界网页数量正以指数速率增长
- 人类历史上以语言文字形式记载和流传的知识占知识总量的80%以上
- 中文网页检索的最高准确率不足40%
- 网络信息检索市场前景广阔



信息安全

- 利用网络组织犯罪，已成为恐怖活动的新特点
- 信息安全问题已经成为国际社会共同关注的焦点





机器翻译

机器翻译市场潜力巨大

- 文化
- 商贸
- 旅游
- 体育
- .....

跨语言通讯与信息获取具有重要的用途

问题的提出

- 如何让计算机实现自动的或人机互助的语言处理功能？

- 如何让计算机实现海量语言信息的自动处理、知识挖掘和有效利用？

◆自然语言处理

NLP的概念

- 什么是自然语言
  - 以语音为物质外壳，由词汇和语法两部分组成的符号系统。 --《新华词典》
  - 语言是人类交际的工具，是人类思维的载体
  - 是约定成俗的，有别于人工语言，如程序设计语言
- 什么是处理
  - 包括理解、转换、生成等
  - NLP的具体表现形式包括机器翻译、文本摘要、文本分 类、文本校对、信息抽取、语音合成、语音识别等

NLP的概念

- 什么是自然语言处理
  - NLP, Natural Language Processing
  - 用机器处理人类语言的理论和技术
  - 将语言做为计算对象来研究相应的算法
  - 目的是让人类可以用自然语言形式跟计算机系统进行人机 交互，从而更便捷、有效地进行信息管理
- 其他名称
  - 自然语言理解
  - 计算语言学
  - 人类语言技术

机器能理解人类语言吗？

- 很难，但没证据表明不能
- 什么是理解？
  - **解构主义**：机器的理解机制与人相同
    - 谁也说不清自己理解语言的步骤
  - **功能主义**：机器的表现与人相同
    - 图灵测试
    - 如果通过自然语言的问答，一个人无法识别和他对话的是人还是机器，那么就应该承认机器具有智能

图灵测试（Turing Test）

- 一个封闭的小屋里面有两间房A和B，其中一间房中是人，一间是机器
- 屋外有一个人H
- H向A，B房内的实体提出各种问题
- H根据来自A，B的回答判断A，B中谁是计算机，谁 是人
- 如果H无法判断，则证明计算机具有人的智能

Eliza问答系统

Young woman: Men are all alike.

Eliza: In what way?

Young woman: They're always bugging us about something specific or other.

Eliza: Can you think of a specific example?

Young woman: Well, my boyfriend made me come here.

Eliza: Your boyfriend made you come here?

Young woman: He says I'm depressed much of the time.

Eliza: I am sorry to hear you are depressed.

Young woman: It's true. I'm unhappy.

Eliza: Do you think coming here will help you not to be unhappy?

Young woman: I need some help, that much seems certain.

Eliza: What would it mean to you if you got some help?

Young woman: Perhaps I could learn to get along with my mother.

Eliza: Tell me more about your family.

Young woman: My mother takes care of me.

Eliza: Who else in your family takes care of you?

Young woman: My father.

Eliza: Your father.

Young woman: You are like my father in some ways.

Eliza: What resemblance do you see?

Joseph Weizenbaum, 1966, **ELIZA — a computer program for the study of natural language communication between man and machine**, *Computational Linguistics*, Jan. 1966 Vol.9, No.1

Searle的“Chinese Room”

哲学家约翰·西尔勒(John R. Searle)否认可以模拟大脑算法的计算机具有意识，他在上个世纪80年代给出了一个著名的理想实验“西尔勒的中文屋子”来支持他的论点。

实验（改编版）的要旨是这样的：

一个用中文讲的笑话，中国人听了会哈哈大笑。如果用一台可以执行复杂算法的计算机根据这个笑话的词句进行一系列运算，得出的结果正确，也是“哈哈大笑”，于是计算机用它自己的方式笑了几声，我们是否可以认为这台计算机“理解”了这个笑话呢？

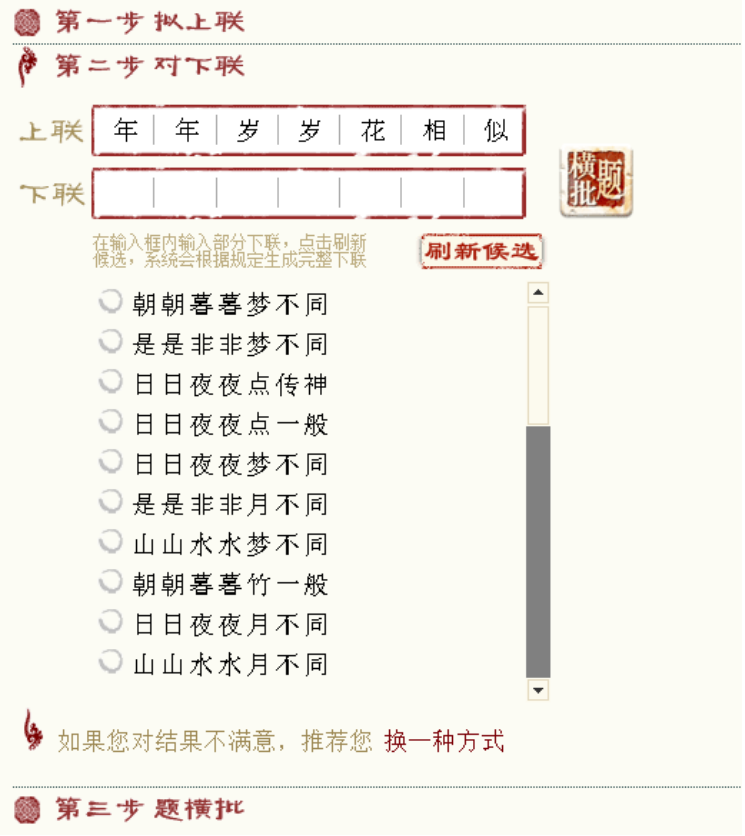
类似的，用一大群不懂中文的美国人取代计算机，他们拿着铅笔和纸重复计算机所做的一切，因为算法很复杂，可能要全美不懂中文的美国人算上一年才得到了结果“哈哈大笑”，他们派一个代表出来笑了几声。虽然反应很慢，但他们和一个中国人做得一样好，不过，这样仍然无法认定这群美国人“理解”了这个中文笑话。

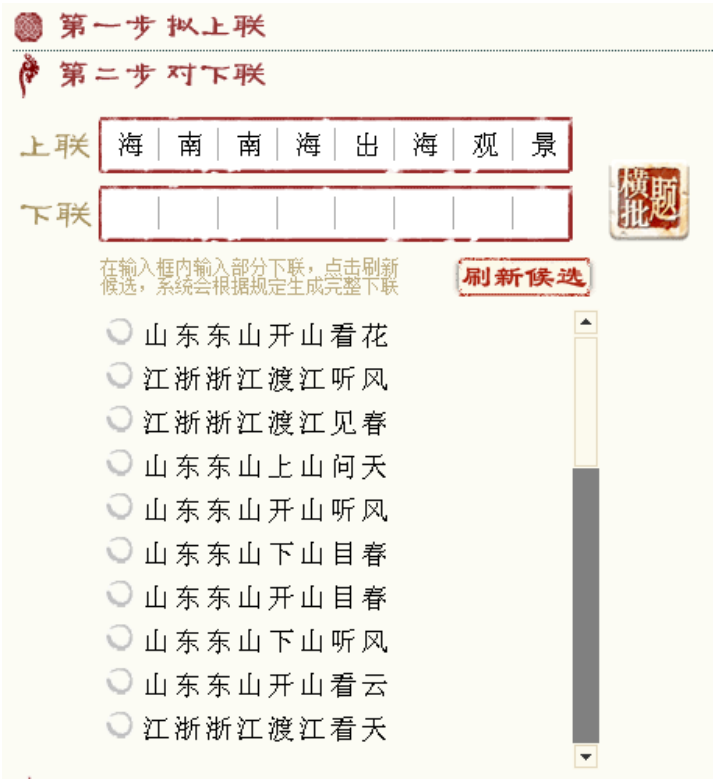
Searle, John. R., “Minds, brains, and programs”, In Behavioral and Brain Sciences 3 (3): 417-457.

微软电脑对联

- 计算机能代人写作文吗？　　- 非常难

http://couplet.msra.cn/app/couplet.aspx

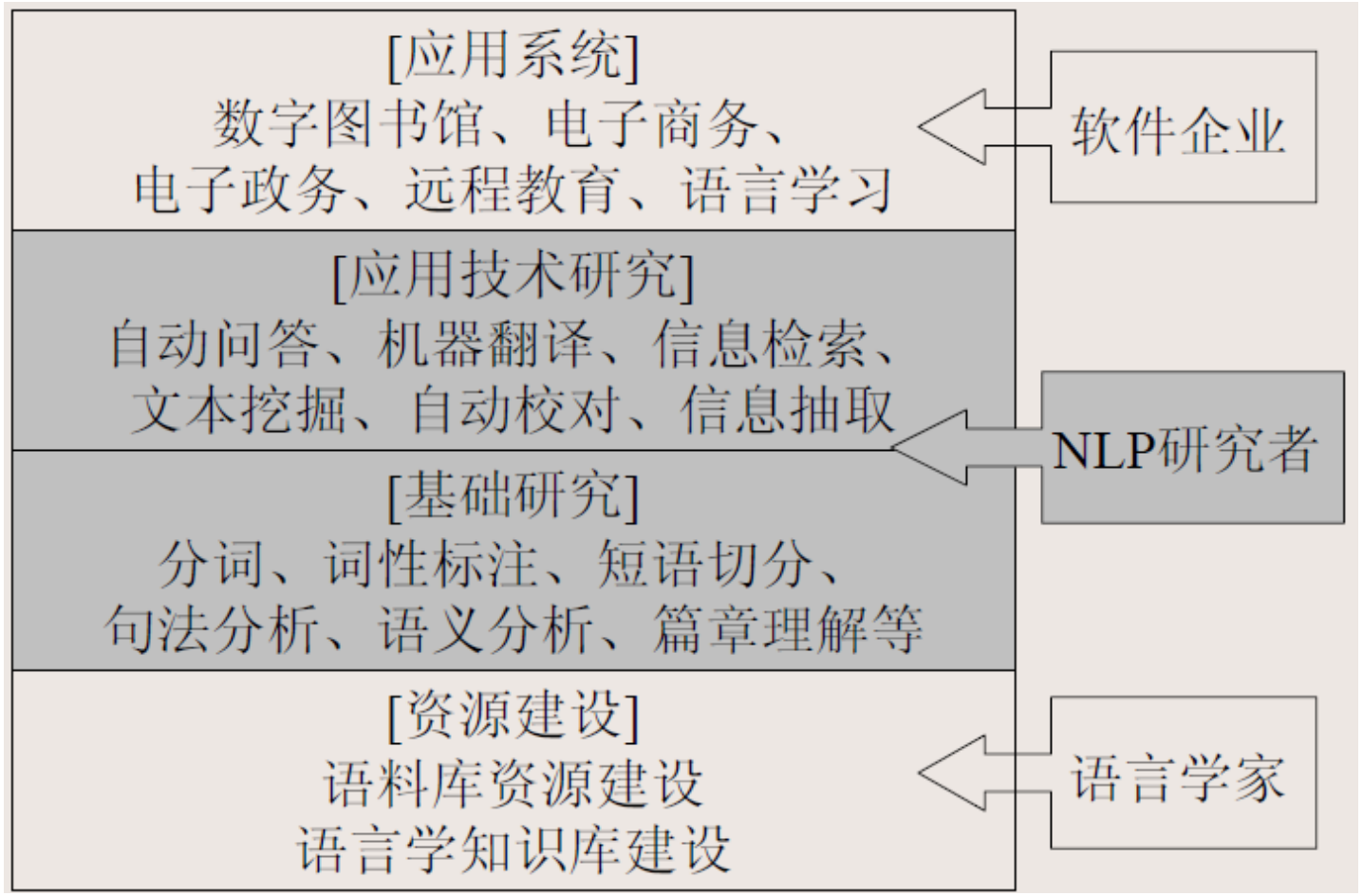




NLP的应用

NLP的应用领域

- 机器翻译
- 文本分类
- 信息提取
- 信息检索
- 语音合成
- 语音识别
- 人机接口
- ....



NLP应用

- 据统计，日常工作中80%的信息来源于语言，文本的需求在不断增长
- 文本是人类知识最大的存储源，并且文本的数量在不断增长
  - 电子邮件、新闻、网页、论文、书籍
- 并非每一样语言处理的应用都需要深层理解
- 成功应用的实例
  - 微软拼音
  - 中文自动校对
  - 搜索引擎

拼音输入法

- 转换
  - 拼音输入：自动将拼音序列转化为字符序列
    - 例子：ji qi fan yi ji qi ying yong he kun nan
    - 汉字序列：？
  - 语音输入：将连续语音转化为汉字序列
  - 将连续文本转化为语音信号输出
- 校对
  - 拼写校对：我们要京城（精诚）合作
  - 文法检查

信息检索

- 利用计算机系统从大量文章中找到符合用户需要的相关信息
- 目前至少有300多亿个网页，只有1%的信息被有效地利用
- Internet或数字图书馆
- 代表系统：Google, 百度

未经中文分词处理时的搜索结果

新闻 网页 贴吧 知道 MP3 图片 视频 地图

和服 百度一下

1.
电信运营商**和服**务提供商
采用奥维通的移动WiMAX解决方案,运营商和服务提供 商可以提供各种个人宽带服务 ……

2.
关于做好党员联系**和服**务群众工作的意见
做好党员联系和服务群众工作，要以马克思列宁主义、毛泽东思想、邓小平理论和“三个代表”重要……

3.
Guangzhou bomei leather co.,ltd
站长信息**和服**务中心:斗破苍穹 阴阳冕 九鼎记 凡人修仙传 猎国 九转金身决……

4.
关于商品**和服**务实行明码标价的规定
根据《中华人民共和国价格法》修订的《关于商品**和服**务实行明码标价的规定》，……

5.
Technical Support
利盟中国面向行业,办公和家庭提供彩色激光,黑白激光,喷墨,和多功能一体打印机及相关耗材**和服**务,是业届领先的打印解决方案的开发制造商。

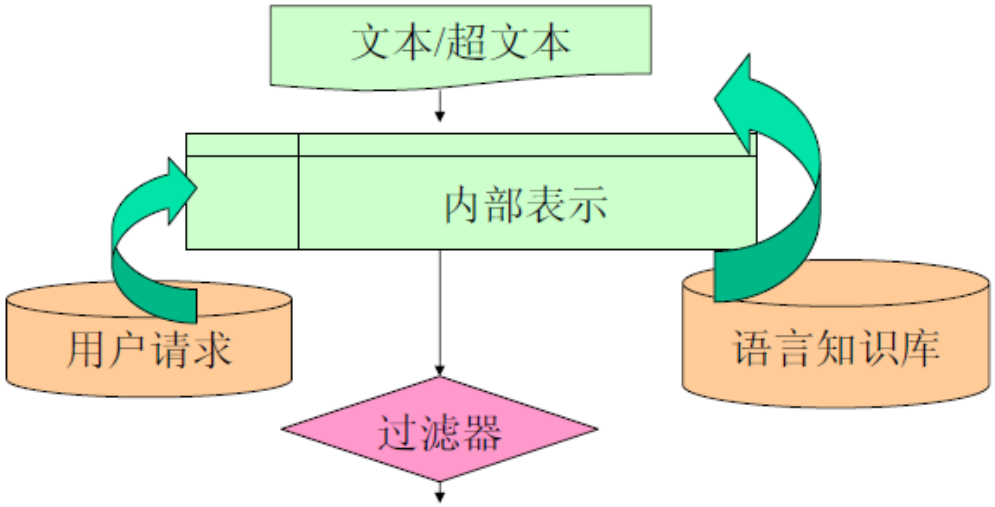
- 信息抽取
  - 从指定文档中或者海量文本中抽取出用户感兴趣的信息
  - 例：实体关系抽取
- 文档分类
  - 利用计算机系统对大量的文档按照一定的分类标准（例如，根据主题或内容划分等）实现自动归类
  - 应用：图书管理、情报获取、网络内容监控等。
- 文字编辑和自动校对
  - 对文字拼写、用词甚至语法、文档格式等进行自动检查、校对和编排
  - 应用：排版、印刷、书籍编撰等。

- **语音识别**
  - 将输入语音信号自动转换成书面文字
  - 应用：文字录入、人机通讯、语音翻译等
  - 困难：大量存在的同音词、近音词、口音等

- **文语转换/语音合成**
  - 将书面文本自动转换成对应的语音
  - 应用：朗读系统、人机语音接口等

- **说话人识别/认同/验证**
  - 应用：信息安全、防伪等

- **自动文摘**
  - 将原文档的主要内容或某方面的信息自动提取出来，并形成文档的摘要或缩写
  - 观点挖掘
  - 应用：电子图书管理、情报获取等
- **信息过滤**
  - 通过计算机系统自动识别和过滤那些满足特定条件的文档信息
  - 例子：过滤色情网站



自动问答

- 通过计算机系统对人体出的问题的理解，利用自动推理等手段，在有关知识资源中自动求解答案并作出相应的回答。
- 应用：人机对话系统、信息检索等
- 简单问答
  - 颐和园怎么走？
  - 香港明天天气如何？
  - 问航班/火车时刻
  - 网上找人
  - 网上购物问价格

Watson问答系统（Feb 2011)

- Q： Each year the EU selects capitals of culture: one of the 2010 cities was this Turkish "meeting place of cultures"
- A： Istanbul (Watson's first answer with 81% confidence)

- Q： Elected every 5 years, it has 736 members from 7 parties
- A： parliament (Watson's first answer with 66% confidence)

- Q： Dialects of this language include Wu, Yue & Hakka
- A： Chinese (Watson's answer: Cantonese 41% confidence)

.....

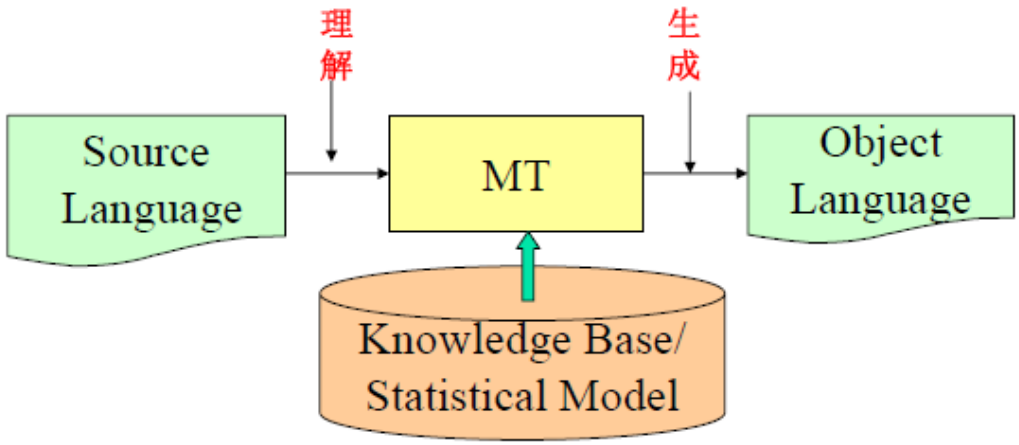
David Ferrucci, Eric Brown, et.al., "Building Watson: An Overview of the DeepQA Project", AI Magazine, 2010, vol. 31, no. 3.

情感及观点分析

- 为什么要对文本进行情感分析？
  - 文本是人写的，必然带有人的感情和观点
  - 大量应用需要情感与观点分析：
    - 评论性文本：商品评论，服务质量，影评
    - 带政治色彩的评论：敌对势力的攻击，法轮功的攻击
- 情感与观点分析要做什么？
  - 观点是什么？带有怎样的情感色彩（正面/负面）？
  - 谁发表的观点或表达的情感？
  - 针对的问题及对象是什么？
  - ➢ 以上都需要通过文本分析提炼

机器翻译

- 自动翻译：英语<-->汉语
- 跨语言检索：输入汉语检索条目，返回满足要求的其他语言信息
- 两个过程：
  - 原语言的分析和理解
  - 目标语言的生成



NLP的发展历史

NLP的发展历史

- 1960s之前：萌芽期
  - 机器翻译、自动文摘
- 1960s中期到1970s中后期：步履艰难
  - 60 年代衰落
- 1970s中后期到1980s后期：复苏
- 1980s后期至今：蓬勃发展
  - 互联网的发展
  - 互联网为NLP 提供了市场需求和试验数据

萌芽期(1960s及之前)

- 1933: 法国的 Georges Artsrouni & 俄国的 Peter Trojanskij建议:构建机器多语言词典;
- 1946-1947:美国的Andrew Booth 和 Warren Weaver，提出了机器翻译 的设想.
- 1950s: Yehoshua Bar-Hillel(MIT): 1952年举办了 1st MT会议，会上，Leon Dostert(Georgetown Univ.)建议开发演示系统，以吸引基金的投资.
- 1955年，第一个演示系统在 IBM & Georgetown 开发，包含250个词和6条句法规则,实现 Russia — English。
- 第一本期刊: Mechanical Translation(1953-1970) 在 MIT出版。
- 第1篇博士论文 1953在MIT由 Anthony G. Oettinger完成：俄语机器词典。

低谷期：ALPAC报告(1966)

- Automatic Language Processing Advisory Committee (ALPAC) (1964, USA)
- ALPAC 报告的内容 (1966) :
  - “There is no immediate or predictable prospects of useful machine translation”—— Ends funding MT.
  - Only support fundamental research in CL
- ALPAC 报告的原因:
  - 遇到了语义障碍 （Semantic Barrier）
  - Bar-Hillel的批评: 需要真实世界的知识

恢复期（1970s）

- 1971: W. Woods: Lunar, IR(ATN)
- 1972: T. Winograd: SHRDLU (Lisp)
- 1973: Schank: Concept Dependency (CD Theory)
- MARGIE 1975: (Meaning, Analysis, Response Generation and Inference on English) on CD: NLU

发展期（1980s）

- 强调知识的重要性
  - 人工智能的发展
  - 日本的第5代计算机（知识处理系统）
  - 语言知识与语言分析的分离
- 语义知识的表示与知识库构建：CYC，WordNet等；
- 机器翻译在受限领域获得成功：加拿大Meteo；
- 句法语义和词汇语义理论的蓬勃发展：
  - GPSG (Generalized PSG: Gazdar),
  - GB (Governing and Binding: N. Chomsky) ,
  - FUG (Functional Unification Grammar: M. Kay)
  - …

经验方法再度受重视（1990s）

- IBM's P.Brown (1988-2nd TMI—Theoretical & Methodological Issues in MT, 1990-CL—Computational Linguistics,1993-CL): 倡导机器翻译的统计方法；
- 基于统计与语料库的方法逐步取得支配地位；
- 强调大规模真实语料；
- 强调机器学习与知识自动获取的重要性；
- 语音识别逐步实用化；
- 开展了大规模的评测。

NLP 方法

理性主义和经验主义

- 理性主义者(Rationalist)
  - 1960-1985: 理性主义是主流
  - Noam Chomsky
  - 他们的信念
    - 先天语言能力
- 对于语法的描述
  - 形成基于规则的传统语言处理技术
- 句法规则的确抓住了语言的主要模式

理性主义的问题

- 语言的变化是渐变的
  - 比如:“ 打”电话，究竟从那一天开始“ 打”被赋予了通讯的意义呢
- 基于规则的方法需要大量的人工操作，人类总结的规则不完备、不一致，规则多了相互冲突，难以对抗复杂的语言现象

经验主义

- 信念
  - 孩子的大脑只能做一些普通的操作：连接、模式识别、一般化。孩子从丰富的信号输入中学习到了语言的结构
- 设定一个语言模型，推导出参数值
  - 形成今天的基于统计的语言处理技术
  - 对每一种语言现象均给出统计量化指标
- 意义：“观其伴，知其意”

经验主义

- 我们生活在一个充满不确定和不完整信息的世界里
- 人类的认知是一个随机现象
- 语言也是一个随机现象
- 对没有见过的语言现象进行估计
- 复杂的概率模型

理性主义和经验主义的差别

- 它们描述了不同的事情
- 理性主义试图去描写人脑中的模型
  - 结构主义者
- 经验主义试图去描写实际出现的语言
  - 功能主义者

理论基础（1950s）

- 形式语言 (Chomsky, Kleene, Backus)
  - 语言描述的形式化：对语言按复杂程度分类，对不同类语言进行形式化描述；
  - 语言处理的形式化：对不同类语言进行自动识别和分析行形式化描述。
- 概率与信息论方法
  - 语言的理解作为解码：Shannon的噪音管道模型；
  - 使用概率方法：将语言的产生看成随机过程。

进一步探讨

- 从九十年代初期开始，统计方法开始成为自然语言处理的主流
- 规范的语言和非规范的语言之间没有明确的界限
- 统计还是非统计，界限也比较模糊
- 追求纯净，还是实用
- 自然语言处理尚不存在统一的数学基础
  - 概率模型、信息论和线性代数

语言工程

- 近来，人们更有兴趣解决工程实际问题
- 人们处理真实世界中的语料，并客观地比较不同方法的优劣
- 面向真实文本的评测，使科学研究和技术开发进一步统一起来。
  - 90 年初的汉语分词系统仍未考虑未登录词　问题，那时已经宣称分词结果达到90%以上，其实只是解决了部分歧义问题。90年代中后期才开始面向真实文本的处理。

强调数据资源（目前）

- 各类电子词典
  - 词义词典：WordNet, HowNet, CCD
  - 语法词典：现代汉语语法信息词典
  - 语义角色：FrameNet, VerbNet

- 各种语料库
  - 英语：词性、实体、角色等，著名的 LDC
  - 汉语：分词+词性标注，汉语词义，汉语拼音 等
  - 多语言（用于机器翻译）：词对应，短语对应，句子 对应，等；

强调评测

- SigHan（汉语）
- NIST (National Institute of Standard and Technology)
  - TREC
  - Open MT
  - MUC(ACE)
  - TDT
  - DUC
- ……

NLP 现状

- 仍然缺乏理论基础
- 词汇句法方面的问题尚未解决，已开始挑战语义、知识等深层课题
- 语音识别中采用的统计语言模型推动了发展，目前的统计模型在向语言深层发展
- Ontology受到普遍重视
- 开放域处理时起时落
- 一切才刚刚开始……

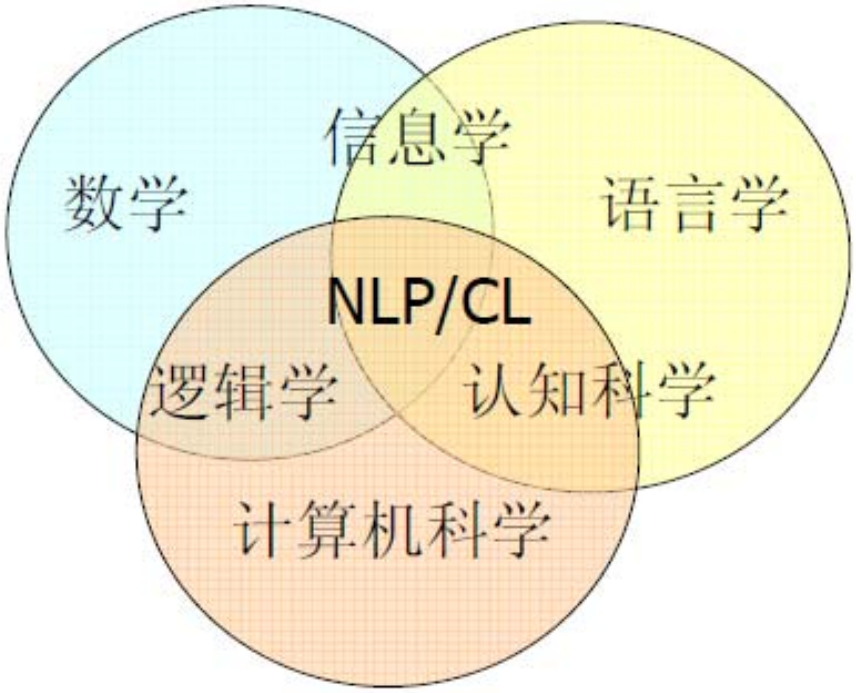
NLP 现状

- 部分问题得到了解决，可以为人们提供辅助性帮助
  - 专业领域文档翻译，电子词典，搜索引擎，文字录入等；
- 基础问题研究仍任重道远，
  - 语义表示和计算、高质量的自动翻译；
- 社会需求日益迫切：
  - 信息服务、通讯、网络内容管理、情报处理、国家安全等
- 许多技术离真正实用还有相当距离，若干理论问题有待更深入的研究
  - 现有模型和方法的改进
  - 在不成熟技术的基础上开发实用系统
  - 期待更有效地理论体系

将 来

- 强调多技术集成
- 强调理性与经验方法
- 强调模块构件化和工具环境开发
- 强调知识与意义的表示与利用
- 强调实用化

小 结



- 多学科交叉
- 年轻
- 应用广泛
- 挑战性（困难）