自然语言处理导论

孙栩
信息科学技术学院
xusun@pku.edu.cn
http://klcl.pku.edu.cn/member/sunxu/index.htm

- **课程信息、内容、规划** 
- **自然语言处理简史、任务简介**
- **自然语言处理的目标、难点、对策**

- **课程信息**
  - 04831780《自然语言处理导论》
  - 任选，2.0学分，36.0总学时
  - 1~16周 每周周二5~6节
  - 教师：孙栩、詹卫东
  - 助教：马树铭、许晶晶

- 选课学生
  - 信息学院
  - 中文系
  - 以计算机为第二专业的学生

课程信息

- 自然语言处理又叫做“**计算语言**学”，涉及到**计算**、**语言**两方面的知识
  - 所以我们安排2位老师讲课
  - 各有侧重点，分别侧重讲解计算、语言两方面的内容

- **教师1**
  - 孙栩
  - 信息学院，研究员
  - 邮箱: xusun@pku.edu.cn
  - 主页: http://klcl.pku.edu.cn/member/sunxu/index.htm

- **教师2**
  - 詹卫东
  - 中文系，教授
  - 邮箱: zwd@pku.edu.cn
  - 主页: http://ccl.pku.edu.cn/doubtfire

助教信息

- **助教1**
  - 马树铭
  - 邮箱 shumingma@pku.edu.cn
- **助教2**
  - 许晶晶
  - 邮箱 jingjingxu_jjx@foxmail.com

自然语言处理是什么？

- **自然语言处理是通过建立形式化的计算模型来分析、理解和处理自然语言**
  - 什么是自然语言：指人类使用的语言，如汉语、英语等
  - 语言是思维的载体，是人际交流的工具
  - 语言的两种属性 - 文字和声音
  - 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

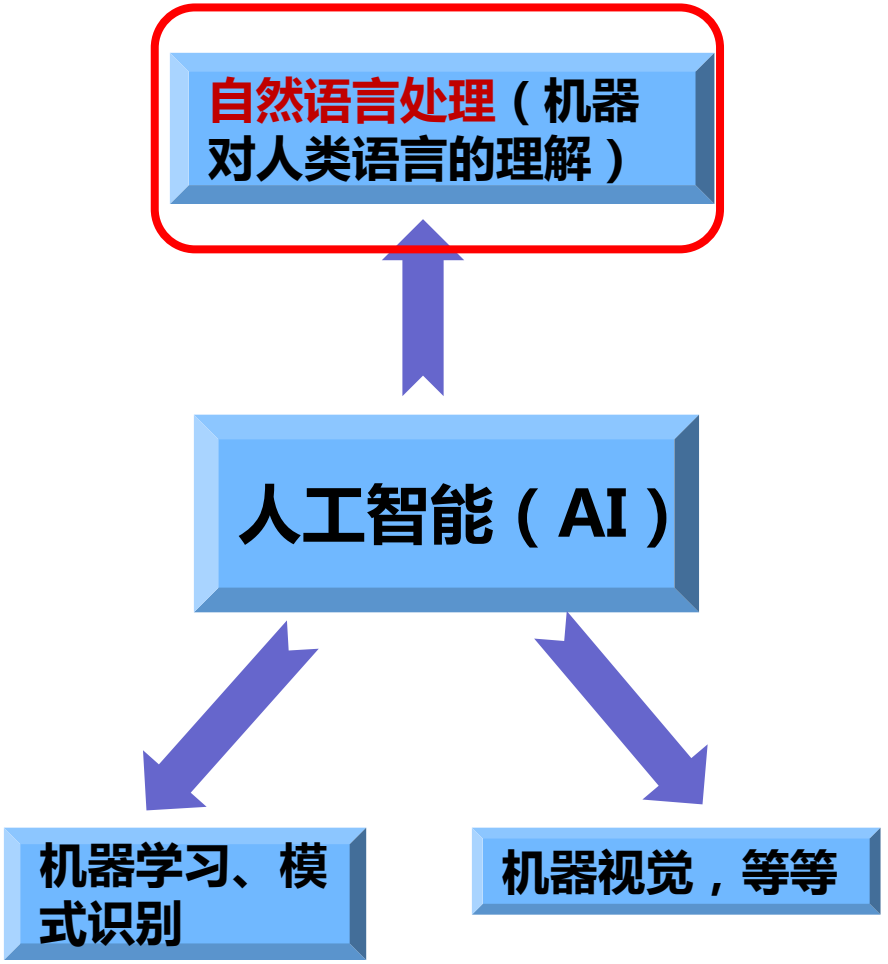
- **其它术语**
  - 计算语言学(Computational Linguistics)
  - 自然语言理解(Natural Language Understanding)
  - 人类语言技术(Human Language Technology)

自然语言处理是什么？

- **自然语言处理（natural language processing，NLP）**
  - 或称自然语言理解(natural language understanding)
  - 是人工智能研究的重要内容

- 自然语言处理就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。

  -冯志伟《自然语言的计算机处理》



研究目标

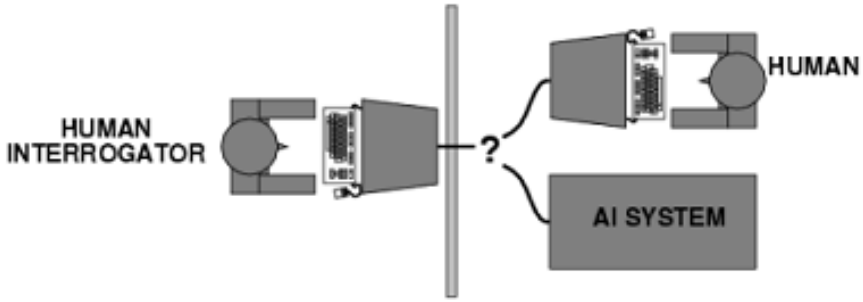
- **终极目标**
  - 强人工智能
  - 强自然语言处理
  - 使计算机能理解并生成人类语言（人工智能的最高境界）

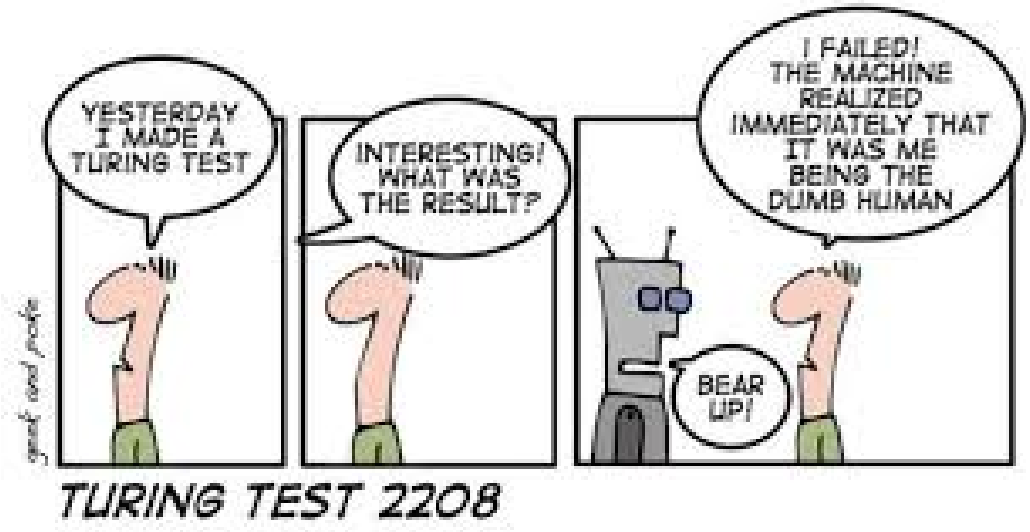
- **当前目标**
  - 弱人工智能
  - 弱自然语言处理
  - 研制具有一定人类语言能力的计算机文本或语音处理系统（目前阶段切实可行的做法）

自然语言处理是什么？

- 强人工智能、**弱人工智能**？

- 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？
  - 与有意识个体（人）比较如何？

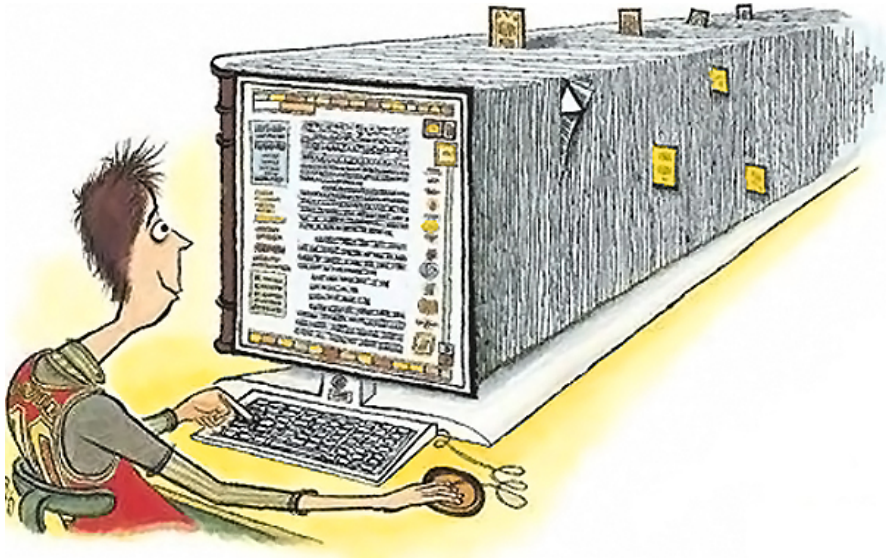




自然语言处理是什么？

- **科学**
  - 是什么？为什么？
- **技术**
  - 目标是什么？怎么实现效果好？

- **自然语言处理既是科学、又是技术**



课程规划

- **课程内容安排**
  - NLP的总体介绍(2周左右)
    - 孙栩
  - NLP的语言学基础(6周左右)
    - 詹卫东
  - NLP的方法和具体应用(8周左右)
    - 孙栩

课程规划

- **自然语言处理是一门交叉学科**
  - 需要多个学科的知识

- **自然语言处理-语言学基础**
  - 詹卫东
- **自然语言处理-概率统计基础**
  - 孙栩
- **自然语言处理-具体方法**
  - 孙栩

课程规划

1：NLP的语言学基础

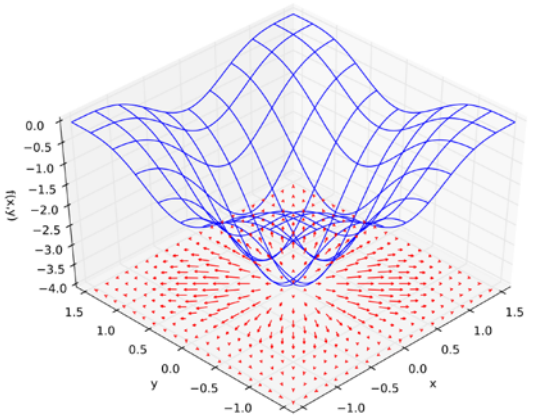
- 语言学知识（Ⅰ）—— 理论分析：构词、词类、句法、语义
- 语言学知识（Ⅱ）—— 实例分析：语料库与知识库

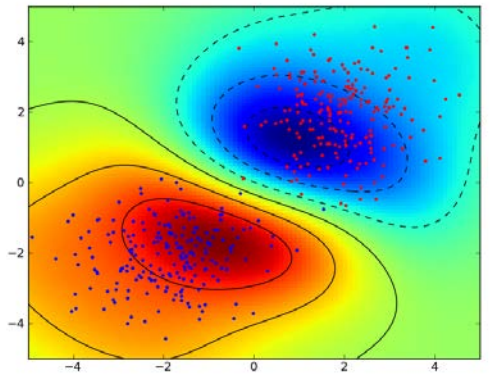
具体计划

- **1.1：构词法与文本自动分词**
  - 具体内容：（1）汉语构词法 （2）中文文本自动分词基本方法与问题
- **1.2：词类与词性标注**
  - 具体内容：（1）汉语的基本句法结构、词类系统 （2）词性标注的基本方法
- **1.3：句法规则与结构分析**
  - 具体内容：（1） 上下文无关文法 （2）句法结构歧义（3）基本的句法分析算法
- **1.4：语义分析**
  - 具体内容： （1） 语义的聚合分析和组合分析（2）特征结构与合一运算
- **1.5：语料库与知识库**

课程规划

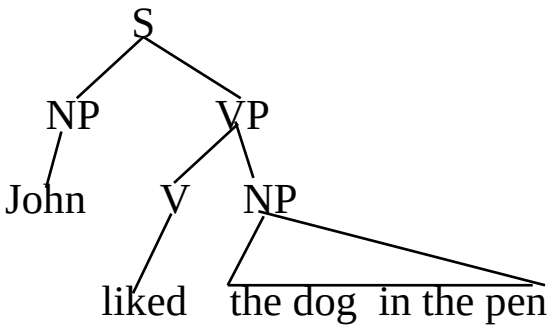
- **2：NLP的概率统计基础**
  - **2.1： 概率论/信息论基础**
    - 概率、条件概率、贝叶斯法则
    - 二项分布、期望、方差
    - 最大似然估计、梯度下降方法、信息论基础
  - **2.2： Ngram统计语言模型**
    - Ngram统计语言建模
    - 数据稀疏问题
    - Zipf定律、平滑基础、回退方法
  - **2.3： 统计机器学习基础**
    - 简单分类问题
    - 感知器模型
    - 支持向量机模型

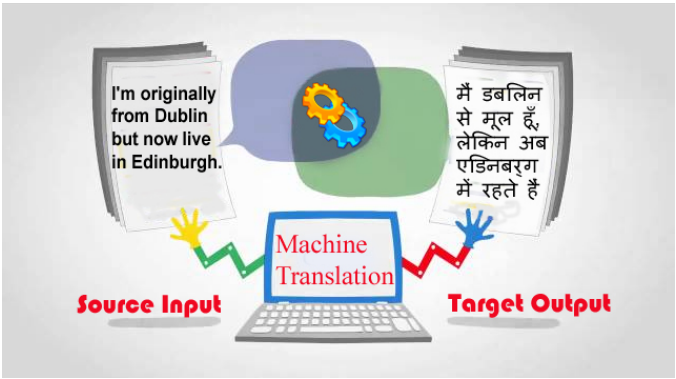




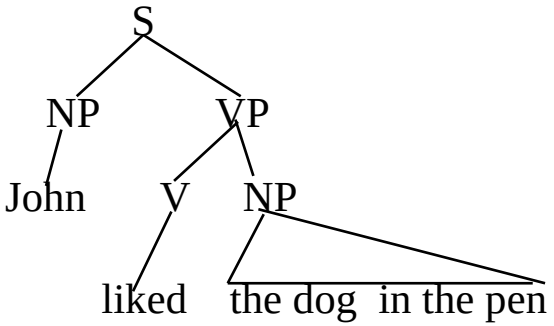
课程规划

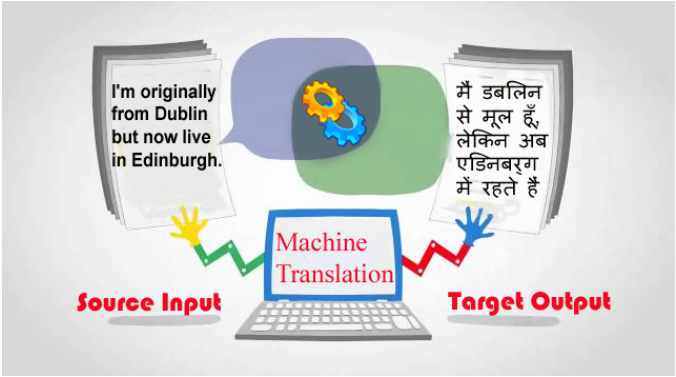
- 3 ：NLP的方法和具体应用
  - 3.1：序列标注问题
    - 线性结构
    - 典型问题：分词、词性标注、实体识别
    - 典型模型：HMM，结构化感知器
  - 3.2：句法分析
    - 树状结构
    - 上下文无关句法、PCFG模型
    - 依存句法、依存句法分析模型
  - 3.3：机器翻译概论
    - 机器翻译的目标
    - 机器翻译的难点、方法简介
    - 具体实现简介





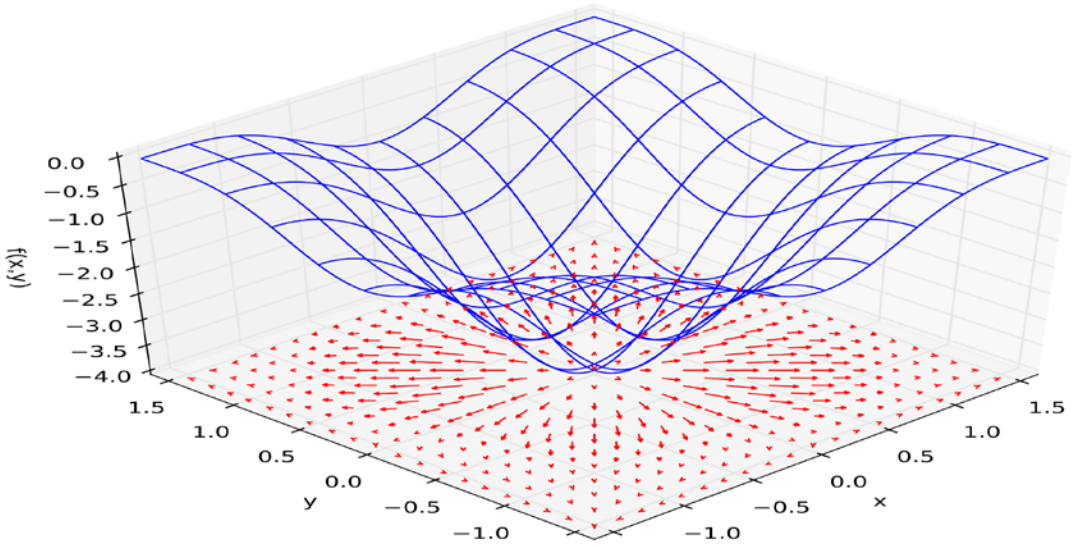
- NLP的具体应用+结构化机器学习
  - 链状结构：比如实体识别
  - 树状结构：比如句法分析
  - 图状结构：比如机器翻译





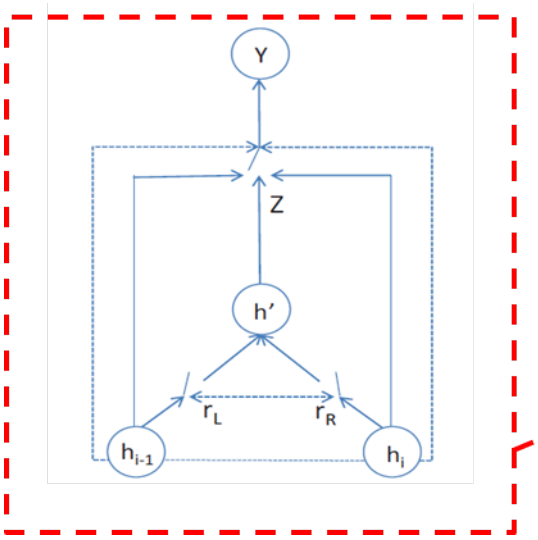
结构化机器学习

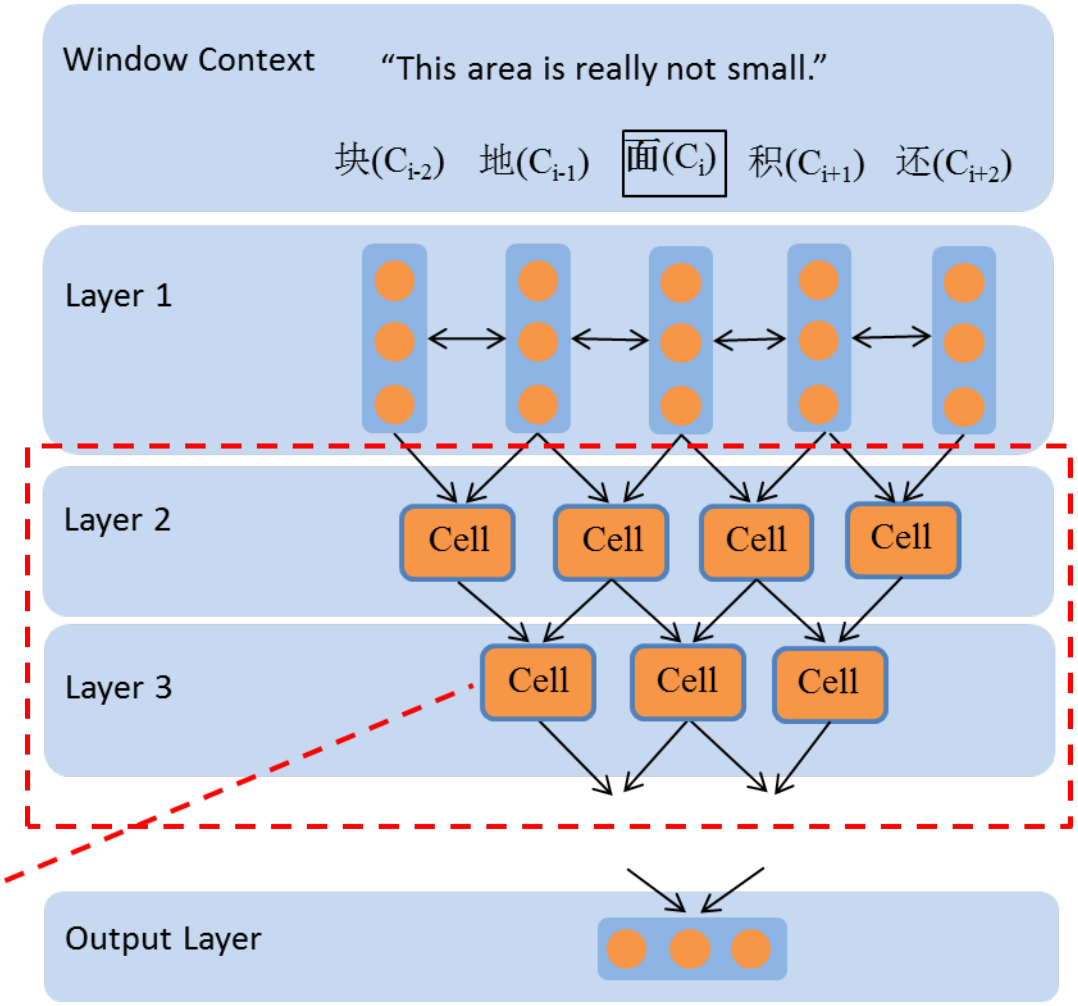
目前，深度学习技术正广泛地运用于自然语言处理系统，提高自然语言处理的准确度



结构化机器学习

- 深度学习：
- LSTM模型
- GRNN模型





- 课程信息、内容、规划
- 自然语言处理简史、任务简介 
- 自然语言处理的目标、难点、对策

简史

- **1940年代末—1960年代中期**
  - 乔姆斯基理论 (Noam Chomsky)
  - 规则方法为主
- **1966年：发展的停顿**
  - ALPAC语义障碍
- **1970年代中期—1980年代**
  - 人工智能AI的繁荣
  - 机器翻译产品如Fujitsi、Hitachi、Siemens

简史

- **1980年代—1990年代前期**
  - 欧盟Eurotra 计划
  - 日本Mu系统以及ODA计划
- **1990年代 - 2010**
  - 统计、算法的进步：IBM统计机器翻译模型、高效率搜索算法等
  - 机器学习技术的进步：结构化分类、图模型等
- **目前的研究**
  - 解决更核心的问题：知识库自动构建、数据大规模化
    - 自动知识库抽取等
    - 大规模自然语言处理模型、算法
  - 更多的现实应用、更好的效果
    - Google translate、Bing translate、语音输入法、iPhone Siri语音问答等
    - 基于网络大数据的自然语言理解成为新热点：信息提取、自动文摘、情感分析、观点挖掘、主题跟踪等

相关学术期刊

- **自然语言处理期刊**
  - Computational Linguistics (CL，计算语言学季刊)
  - Transactions on Association of Computational Linguistics (TACL)
- **（包括自然语言处理的）综合期刊**
  - Journal of AI Research (JAIR，人工智能研究月刊)
  - ACM Trans. on Information Systems (TOIS)
  - Information Processing & Management (IPM)
  - Journal of Machine Learning Research (JMLR)
- **国内主办的相关期刊**
  - 中文信息学报
  - 计算机学报
  - 软件学报

相关学术会议

- **自然语言处理会议**
  - Annual Meeting of the Association for Computational Linguistics (ACL, 计算语言学会的年会)
  - Empirical Methods on Natural Language Processing (EMNLP)
  - International Conference on Computational Linguistics (COLING)

- **（包括自然语言处理的）综合人工智能会议**
  - International Joint Conf. on Artificial Intelligence (IJCAI，国际人工智能联合会)
  - AAAI Conference on Artificial Intelligence (AAAI)

具体的自然语言处理任务简介

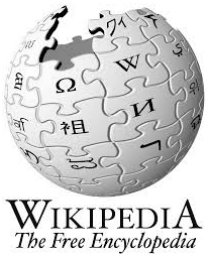
- 机器翻译

- 人机对话

- 信息检索、信息提取

- 情感分析、舆论分析、知识发现

- 自动抽取知识库



大数据时代——基于Web的知识挖掘

互联网 信息爆炸
2009年全球网络数据量达0.8ZB
2020年将达到35ZB

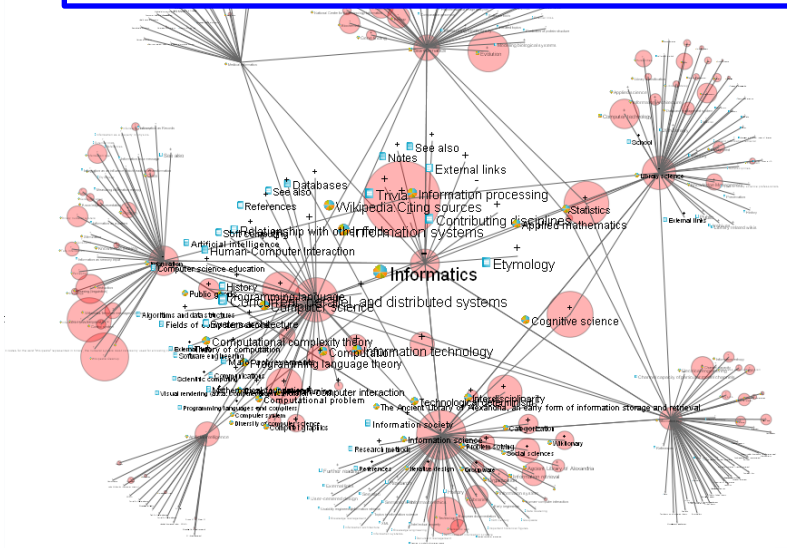
面对海量的大数据时代
知识更加重要

用户对于信息精准化的需求
智能知识服务（逻辑推理）的需求

从复杂多变的互联网数据中挖掘有用的知识，是高效互联网服务的重要基础
基于Web的知识挖掘受到工业界和学术界的高度关注

Knowledge Graph

Google从2010年开始致力于构建相互关联的实体及其属性的巨大知识图表。

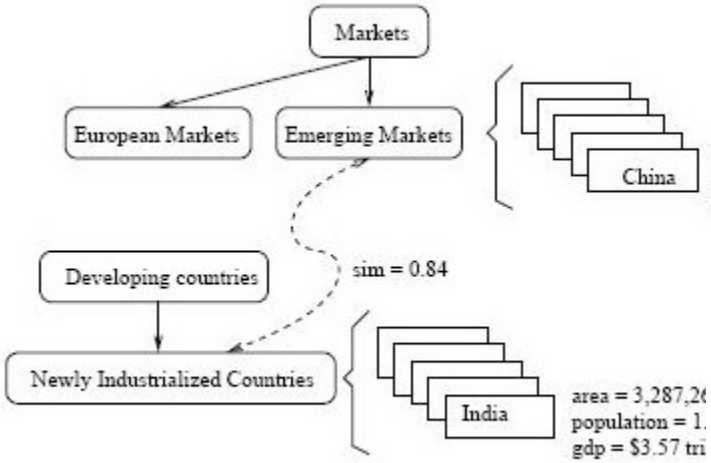


超过5亿实体
超过35亿条关系

Google搜索部门负责人Amit Singhal表示：“Google在这个知识图表支撑下，能够将网页上的单词都转变为带有属性的实体，使及其能够更加精确地理解语义信息，从而进行更好的搜索匹配”。

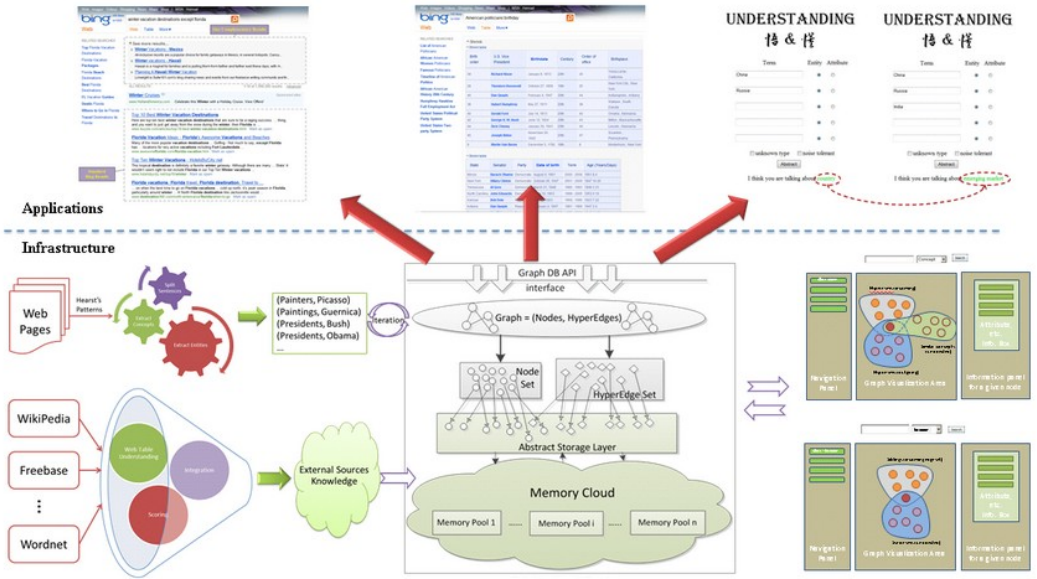
Microsoft ProBase

The goal is to enable machines to better understand human communication.
Knowledge in Probase is harnessed from billions of web pages



bing™

2,653,873实体



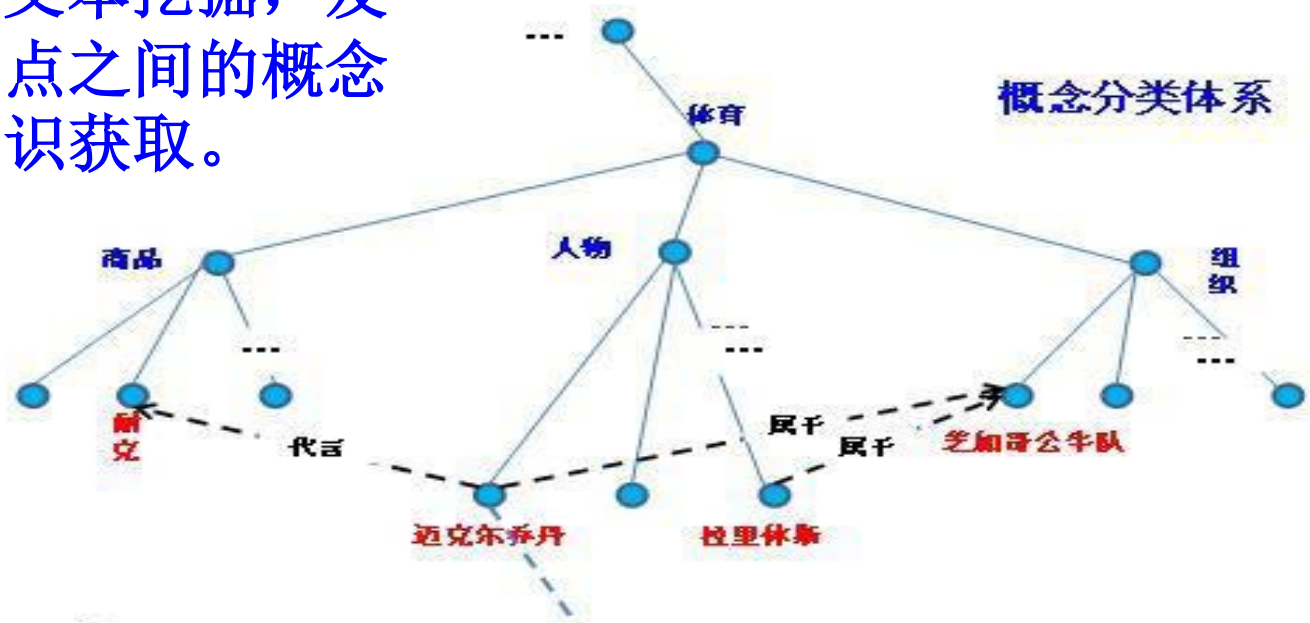
Web知识工程——机遇与挑战

- 在Web知识提取中，如何充分利用Web资源特有的各种优势，从多源异构、海量、开放的网络文本中准确地提取各种知识组成要素，是实现Web知识获取的关键。

基于NLP的Web知识工程

利用NLP技术进行文本挖掘，发现知识点以及知识点之间的概念关系，辅助进行知识获取。

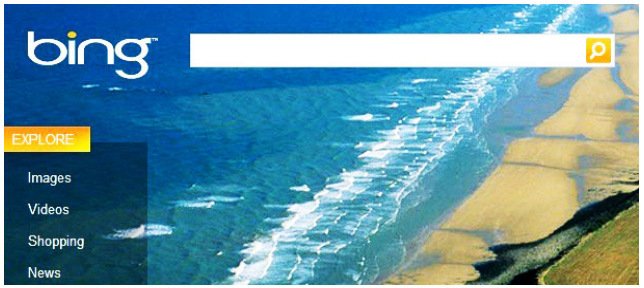
- 概念提取
  - 概念实例
  - 属性名
  - 属性值
- 概念间关系
- 概念层级结构



概念实例“迈克尔·乔丹”				
属性名	属性值	属性名	属性值	属性描述体系
中文名	迈克尔·乔丹	体重	98.1kg	
外文名	Michael Jordan	主要奖项	6次NBA总冠军 2次奥运会冠军 3次NBA全明星MVP 5次常规赛MVP 6次总决赛MVP	
别名	米高·佐敦、空中飞人			
国籍	美国			
出生地	纽约市布鲁克林区			
出生日期	1963年2月17日			
毕业院校	北卡罗莱纳大学	司职	得分后卫	
身高	198cm	运动项目	篮球	

信息检索

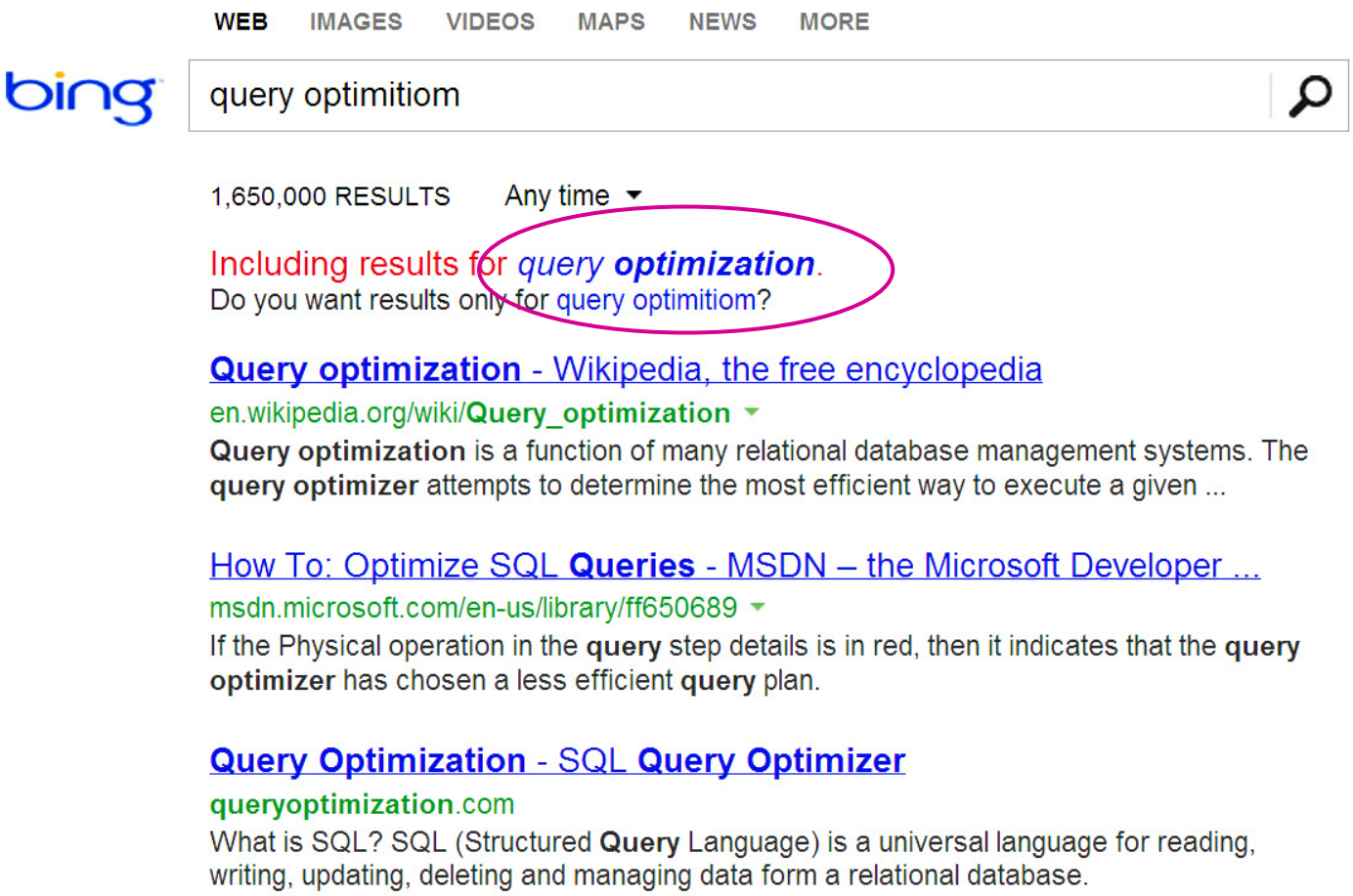
- **搜索引擎的战略转型——以自然语言问答为代表的语义搜索**
- **百度搜索——如果用户输入以下问题，应如何返回答案：**
  - 听起来快乐的歌曲
  - 令人心情愉快的图片
  - 李世民与李治的关系是什么？
    - 唐高宗李治(628～683)唐太宗李世民第九子,字为善。——答案隐含在文本中
  - Who is Ronald Reagan‘s wife?
    - Nancy Reagan, wife of President Ronald Reagan, was born……
  - 我今天起床后头晕，恶心，差点晕倒。这是怎么了？
    - 症状为头晕、恶心的疾病有……，这些病还包括其它症状如……，应作哪些检查，预防与治疗的手段为……——理解、联想与推理（智能问答、主动推送）
  - ……



- **自然语言处理是下一代搜索的核心支撑技术**

公司	收购/投资公司	产品类型
苹果	Siri	自然语言问答
HP	Autonomy	语义搜索
微软	PowerSet	语义搜索
Google	MetaWeb	知识搜索
IBM	Watson	自然语言问答

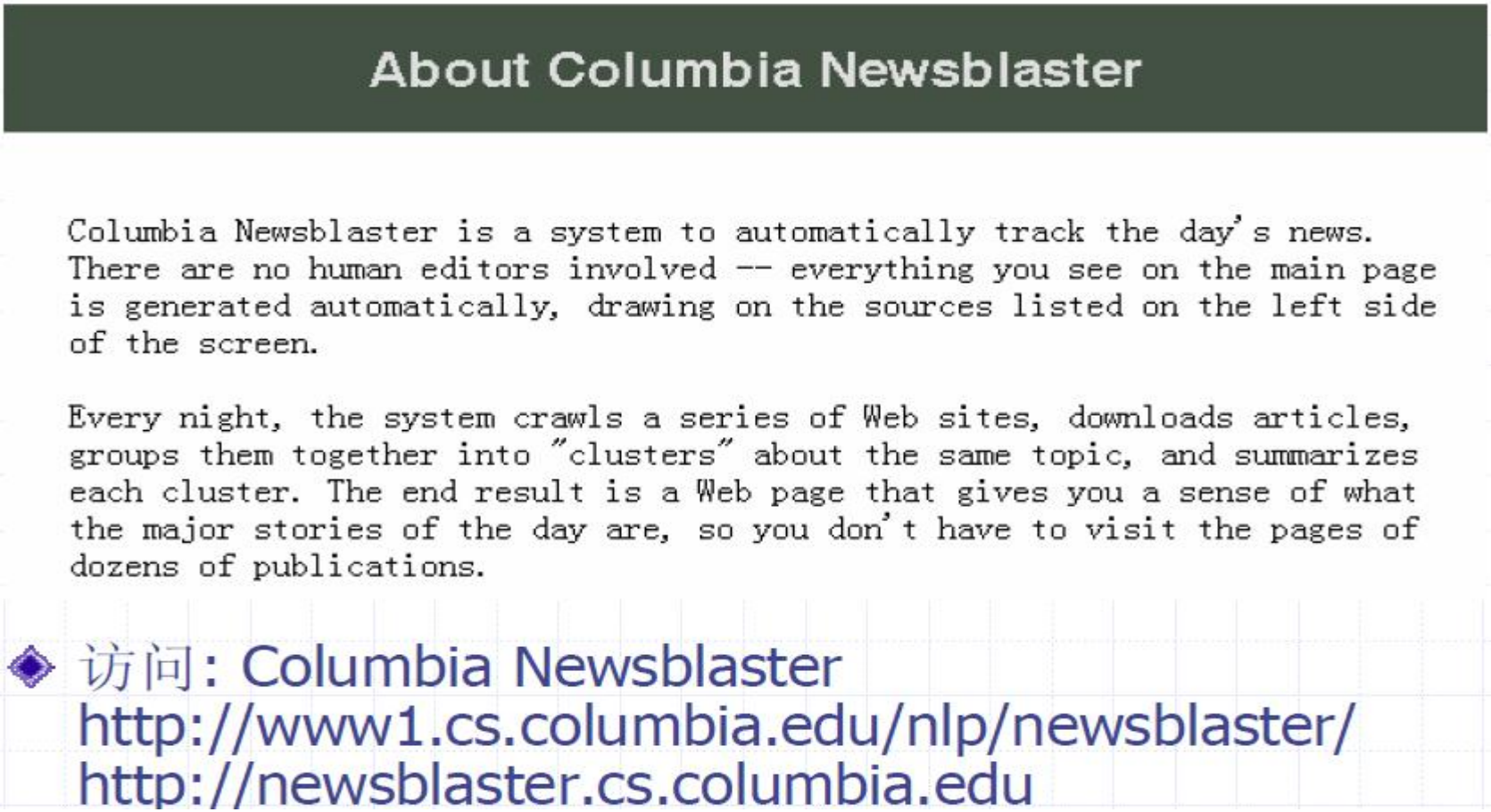
信息检索


WEB IMAGES VIDEOS MAPS NEWS MORE
bing
query optimitiom
1,650,000 RESULTS
Any time
Including results for query optimization.
Do you want results only for query optimitiom?
Query optimization - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Query_optimization
Query optimization is a function of many relational database management systems. The query optimizer attempts to determine the most efficient way to execute a given ...
How To: Optimize SQL Queries - MSDN – the Microsoft Developer ...
msdn.microsoft.com/en-us/library/ff650689
If the Physical operation in the query step details is in red, then it indicates that the query optimizer has chosen a less efficient query plan.
Query Optimization - SQL Query Optimizer
queryoptimization.com
What is SQL? SQL (Structured Query Language) is a universal language for reading, writing, updating, deleting and managing data form a relational database.

自动文本摘要

- **信息时代，信息爆炸，信息泛滥，人不可能看完所有的信息，需要浓缩**
- **什么是摘要？**
  - 将长篇幅的文本核心内容浓缩成短的表示形式。
- **涉及自然语言处理的两个过程：**
  - 理解长的原文
  - 生成短的新文章
- **单文本摘要：只针对一个文本生成摘要**
- **多文本摘要：针对一个文本集合生成摘要**

- **哥伦比亚大学开发的多文档文摘系统Newsblaster。Newsblaster将每天发生的重要新闻文本进行聚类处理，并对同主题文档进行冗余消除、信息融合、文本生成等处理，从而生成一篇简明扼要的摘要文档。**



Columbia Newsblaster

Summarizing all the news on the Web

Thursday, November
Articles from 11/15/2010 to
Last update: 4:

Search for:

Offline summarizatior Go

U.S.
World
Finance
Sci/Tech
Entertainment
Sports

View Today's Images

View Archive

About Newsblaster

About today's run

Newsblaster in Press

Academic Papers

Article Sources:

zdnet.com
(158 articles)
seattletimes.
nwsource.com
(133 articles)
washingtonpost.com
(60 articles)
suntimes.com
(23 articles)
cbc.ca
(22 articles)
baltimoresun.com

barnes & noble nocolor News and Other Resources (Science/Technology, 143 articles)

A fascinating blog post looks at Cupertino's investigation into tablet computing from th earliest days... Blog posts November 12, 2010 12:57pm PST This Is Why Apple Is Eating Yo Lunch. Blog posts November 12, 2010 9:06am PST Apple gets bitten for using Lion stock ph (updated) Apparently Apple used a stock photo of a Lion that was previously used in Belg the Vlaams Blok campaign a nationalistic political party. Blog posts October 19, 2010 8: 10 concept cars that got lost in translation (photos) A concept car rarely makes it to t room floor with its design intact. A 100 screenshot gallery shows you just about everyth Blog posts November 17, 2010 5:41am PST Over 100 images showing nearly all the functions B&N NOOKcolor. Blog posts November 11, 2010 5:31am PST Windows Phone 7 stock already dep Fake shortage or true demand? Blog posts November 10, 2010 10:03pm PST Gartner: 54.8 mil tablets will be sold in 2011. Blog posts November 16, 2010 7:12am PST The Beatles now av through iTunes... yawn... Yesterday Apple teased us with this promise... Blog posts Nove 2010 6:54am PST Apple lands the Beatles for iTunes.

Other stories about |, pdt and am:

- Enterprise software (6 articles)

Top News

Uneasy House Democrats keep Pelosi as their leader

(U.S., 8 articles)

House Democrats elected Speaker Nancy Pelosi to remain as their leader Wednesday despite massive party losses in this month's congressional elections that prompted some lawmakers to call for new leadership. Democrats will huddle this morning behind closed doors in Cannon Caucus Room to cast secret ballots for the next Democratic leader, whip, assistant leader and the caucus chair and vice chair. Pelosi will continue leading Democrats in the House of Representatives during the 112th Congress, winning a vote of her caucus, 150-46, on Wednesday.

Other stories about Earmarks, Pelosi and Democrats:

- Senate Democrats swim against anti-earmark tide (6 articles) [UPDATE]

Prince William Engagement: What Royal Title W Middleton Get? (U.S., 20 articles) [UPDATE]

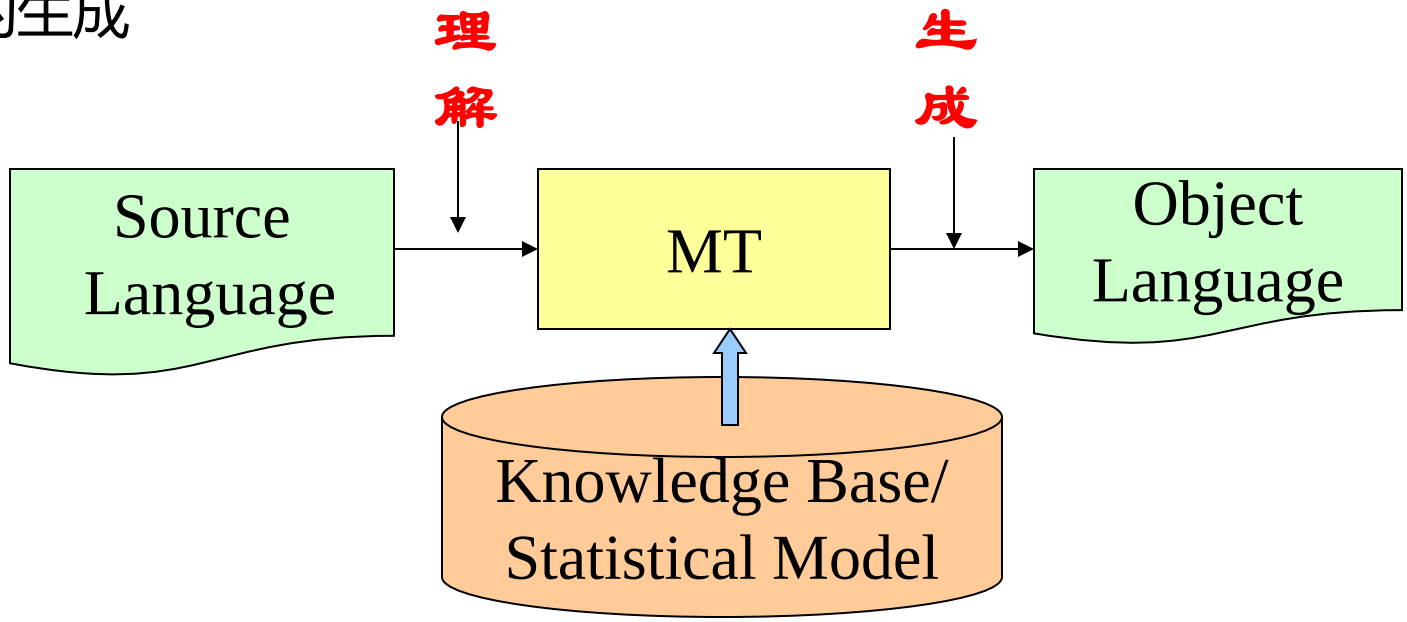
Gradually Middleton began to appear at public ever smiling graciously and consistently on William's a wearing the enormous froufrou hats required of Bri swells, with far more aplomb than could be expecte recent college grad. On Tuesday the couple, both 2 for their first public interview, Middleton in a s plunging-neckline dress that perfectly matched the sapphire on her hand. " This is going to be a diff kind of wedding than Charles and Diana's said Joe managing editor of Majesty Magazine.

机器翻译

- **目标是研制能把一种自然语言翻译成另外一种自然语言的计算机软件系统。**
- **相关研究始于四十年代末**
  - 计算机诞生不久
- **目前市场上有不少翻译产品**
  - 已经具有较强实用价值，例如google translate等

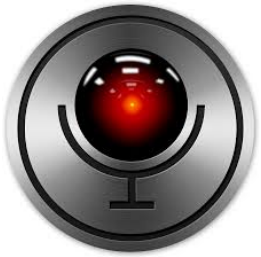
机器翻译

- **利用计算机把一种自然语言的文本表述转变为相同意义下另一种自然语言的文本表述**
- **自动翻译：英语←→汉语、英←→法、日←→俄**
- **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（检索条目的翻译）**
- **两个过程：**
  - 原语言的分析和理解
  - 目标语言的生成



人机对话

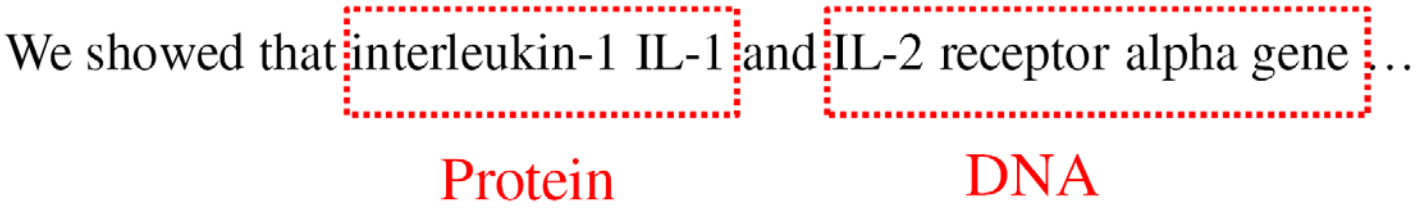
- **2001: A space odyssey:**
  - Dave: Open the pod bay doors, HAL.
  - HAL: Im sorry Dave, I am afraid I cant do that
  - Dave: Whats the problem?
  - HAL: I think you know what the problem is just as well as I do.



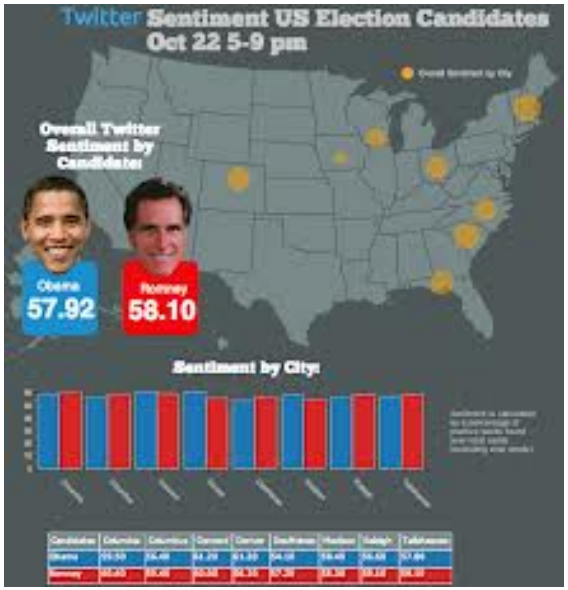
- **市场上有不少实用的人机对话系统，比如iPhone的Siri等**

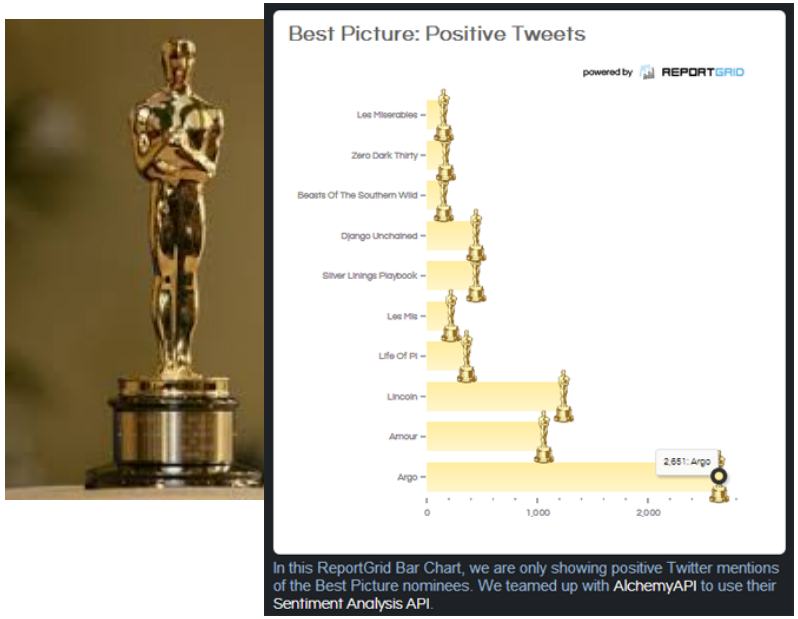
信息提取

- 比如命名实体识别



情感分析、舆论分析、知识发现





IBM: Watson自然语言问答系统

Q: Who was presidentially pardoned on September 8, 1974?
A: Nixon.

- 2011年2月17日，IBM超级电脑“沃森”(Watson)在美智力竞猜节目中击败人类，成为继战胜国际象棋大师的“深蓝”之后，比“深蓝”更聪明的巨型计算机。
- 与IBM“深蓝”相比较，“沃森”能处理人类的自然语言，并根据逻辑推理迅速对复杂问题作出回答
- 国际象棋的规则定义非常明确，而人的自然语言，完全是开放式的，往往很模糊，需要上下文才能理解意思。这点人类看着容易，对电脑却极具挑战性。
- “沃森”获胜标志着人工智能领域一个历史性时刻

探索机器智能

- 在受限时间、受限领域，自动问答系统可以做到“混淆人机”
  - 可否发展自动问答系统，使之实现在任何领域、足够长的时间内的“智能问答”，使之通过图灵测试？
  - 这是从量变到质变的过程，目前的人工智能技术无法实现
  - 到目前为止（甚至在可以预见的将来），没有一台计算机能通过图灵测试！

- 自然语言处理在智能科学方面的探索：
  - 什么叫“理解”—种语言？
  - 智能的本质是什么？
  - 人类理解语言的机制是什么？
  - 机器能否获得与人类可比、甚至超越人类的智能？

总结

- **互联网时代，自然语言处理研究具有**
- **重要的科学意义**
- **广阔的应用前景**
  - 基于Web的文本知识工程
  - 自然语言智能问答
  - 中文文本语义分析
  - 自动文摘
  - 网络文本情感分析
  - 机器翻译
  - ……

Outline

- 课程信息、内容、规划
- 自然语言处理简史、任务简介
- 自然语言处理的目标、难点、对策

为什么要自然语言处理？

- **语言障碍**
  - 人 - 人之间的语言障碍
  - 人 - 机之间的语言障碍
  - 人本身的语言障碍

为什么要自然语言处理？

- **人 - 人之间的语言障碍**
  - 机器翻译
  - 自动摘要
- **人 - 机之间的语言障碍**
  - 信息提取
  - 知识抽取
  - 机器阅读(machine reading)
- **人自身的语言障碍**
  - 写作辅导(writing assistance)

自然语言处理的研究目标

- **终极目标**
  - 研制能理解并生成人类语言的计算机系统
  - 彻底解决语言障碍问题
  - AI-complete问题
- **当前目标**
  - 研制出具有一定人类语言能力的计算机文本或语音处理系统
  - 部分解决语言障碍问题
  - 现实的商业和应用价值

自然语言处理的难点是什么？

- **表象原因：自然语言中有大量的歧义现象**
  - 无法象处理人工语言那样，写出一个完备的、有限的规则系统来进行定义和描述。自然语言的规则很少没有例外
  - 此外，还有大量的噪音甚至错误表达

自然语言处理的难点是什么？

- **原因：自然语言中有大量的歧义现象**

- **歧义举例：**
  - The boy saw the girl with a telescope.

自然语言处理的难点是什么？

- **本质原因：知识体系的缺乏**
  - 自然语言的理解不仅和语言本身的规律有关，还和语言之外的知识（例如常识）有关
  - 语言处理涉及的常是海量知识，知识库的建造维护难以进行
  - 场景/背景的建立问题

自然语言处理的难点是什么？

- **两个原因的联系**
  - 歧义是知识缺乏的表现形式

- **如果有全局知识/上下文知识支持？**
  - The boy saw the girl with a telescope.

常见对策

- **由于歧义/知识缺乏等因素的存在，自然语言处理常采用下面的对策**
  - 建立“已知知识”
    - 比如使用训练数据
  - 减少“未知知识”
    - 比如对领域进行限制

常见对策

- **对策一：建立“知识”**
  - 规则方法(rule-based methods)
    - 通过语言学知识编写规则
    - 通过规则引入知识
  - 经验方法(empirical methods)
    - 训练数据+机器学习
    - 通过训练数据引入知识，通过机器学习消岐
  - 规则和经验方法的结合
  - 交互式处理
    - 人机互助进行处理

常见对策

- **对策二：减少 “未知知识”**
  - 限定语言
  - 限定领域
    - 限定处理文本的领域
  - 限定任务
  - 限定复杂度
    - 限定语言的词汇和句法，降低复杂度

具体方法

- **规则驱动的方法（符号主义）**
  - 1. 研究人员（例如语言学家）对语言的规律进行总结，形成规则形式的知识库。
  - 2. 研制语言处理算法，利用这些规则对自然语言进行处理。
  - 3. 研究人员根据处理结果，调整规则，改进处理效果。
  - 4.如此反复

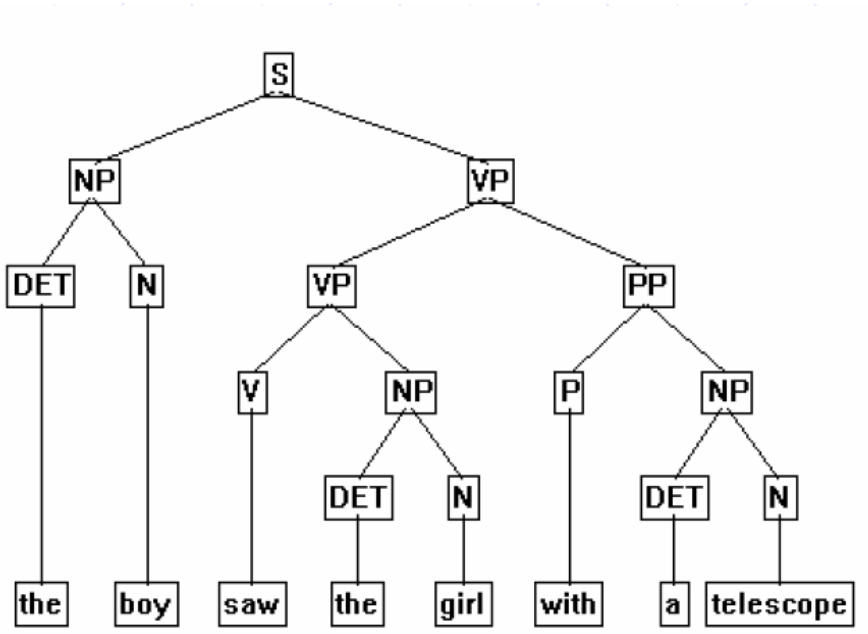
具体方法

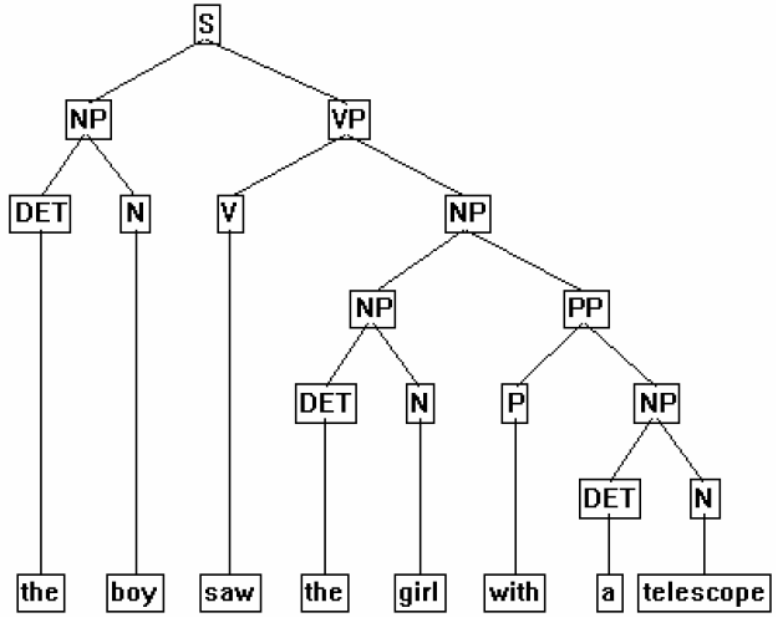
- **规则方法举例**

- **例如:**
  - S → NP + VP
  - NP → DET + N
  - NP → NP + PP
  - VP → VP + PP
  - ….

具体方法

- 用上述规则分析句子“the boy saw the girl with a telescope”





具体方法

- **All grammar leak (Sapir 1921)**
  - 对于自然语言而言，不大可能写出一部完备的规则集，语言规则有很强的伸缩性。

- **规则系统的普遍问题**
  - 不完备
  - 规则本身的歧义
  - 理论不够严谨(ad-hoc)
  - 规则调整和更新很复杂
  - 维护困难

具体方法

- **数据驱动的方法（经验方法、统计学习方法）**
  - 1. 建立可以反映语言使用情况的训练数据
  - 2. 利用统计学习技术，基于训练数据学习到一个参数模型
  - 3. 基于学习到的模型，对目标自然语言数据进行处理
  - 4. 根据处理效果改进模型，提高处理性能。

具体方法

- 在数据驱动的方法中，语言模型通常体现为一组参数，这些参数通常表示某个语言形式发生的概率值。例如：

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = \frac{\exp\left\{\boldsymbol{w}^{\top}\boldsymbol{f}(\boldsymbol{y},\boldsymbol{x})\right\}}{\sum_{\forall \boldsymbol{y}'}\exp\left\{\boldsymbol{w}^{\top}\boldsymbol{f}(\boldsymbol{y}',\boldsymbol{x})\right\}}$$

具体方法

- 参数训练过程是一个数值优化过程，例如：

$$\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{n} \log P(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{w}) - R(\boldsymbol{w})$$

- 可以通过梯度下降等方法对参数进行优化、学习

具体方法

- 问题：
- 数据驱动的方法忽视了语言的深层结构
- 训练数据的建设往往花销比较大
- 结果不好分析、解释

具体方法

- **融合规则驱动和数据驱动的方法**
  - 规则驱动、数据驱动的优劣不能简单评价
  - 两种方法往往优缺点互补
  - 已经提出了一些策略，但须进一步探索
  - 目前数据驱动的方法占主流

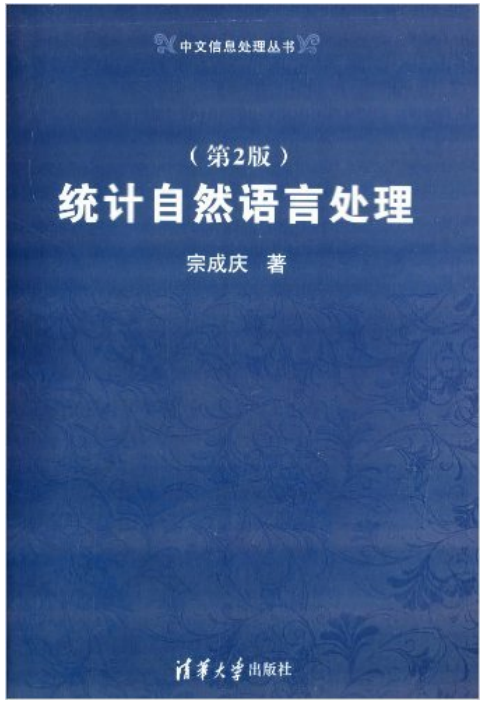
具体方法

- **目前，数据驱动的方法是主流**
  - 1992: 24%
  - 1994: 35%
  - 1996: 39%
  - 1999: 60%
  - 2001: 87%
  - 2010：>90%

- **效果评测?**
  - 自然语言歧义多、关于语言处理方法和系统的评测也需要解决相关的歧义问题
  - 1，规避语言学争议、制定标准测试集
  - 2，看应用效果

参考书

- NLP的概率统计基础、具体应用
  - 《统计自然语言处理》
    - 宗成庆 编



- NLP的语言学基础
  - 《自然语言处理综论》
    - 冯志伟/孙乐 译

- “学有余力”的同学还可以看以下参考书
  - 《Pattern Recognition and Machine Learning》
  - Bishop, Christopher M., Springer-Verlag, 2006

复习思考题

- 什么是自然语言处理？
- 自然语言处理和人工智能的关系？
- 自然语言处理的难点？
- 目前自然语言处理有哪些主流方法？

谢谢！

QUESTION？