大数据时代的自然语言理解技术

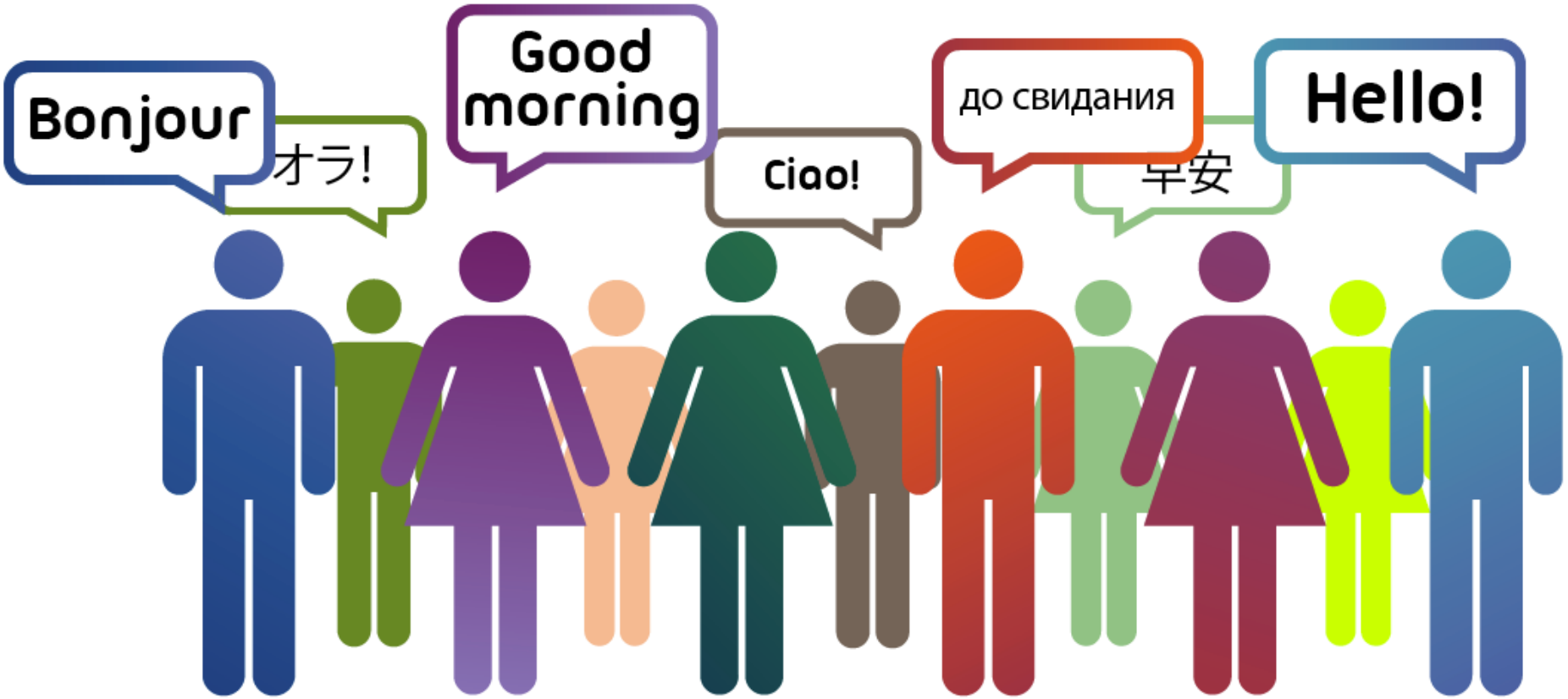
刘知远

清华大学自然语言处理实验室

liuzy@tsinghua.edu.cn

自然语言理解简介

自然语言



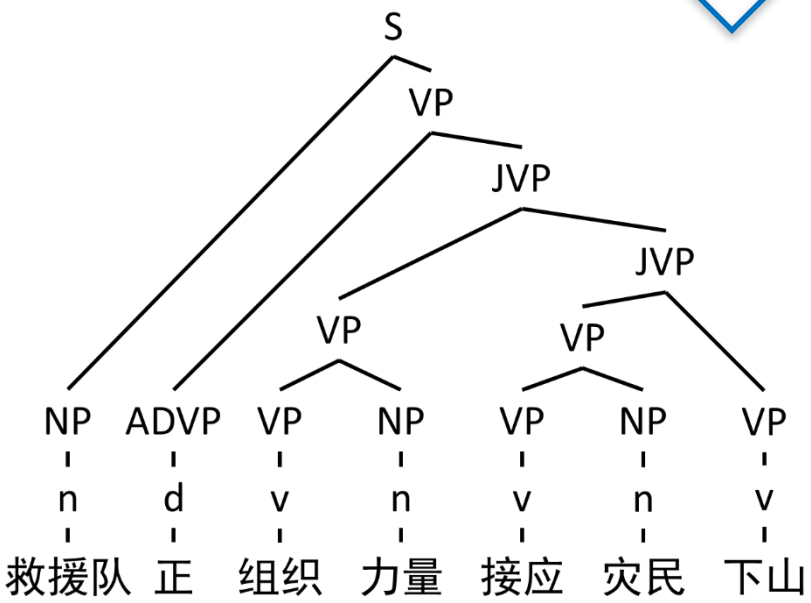
自然语言

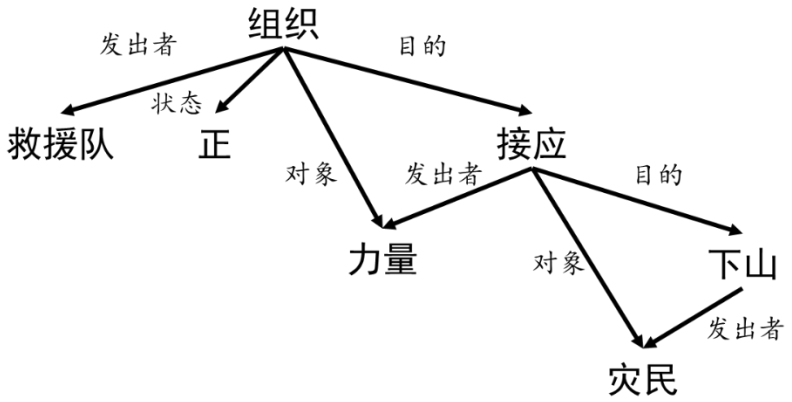
什么是自然语言理解？

输入：　　救援队正组织力量接应灾民下山

结构预测

输出：





句法结构　　语义结构

自然语言理解的本质是结构预测

自然语言理解的复杂度

- 句子二分结构树的可能解空间：数量随句子长度呈指数级增长

$$\frac{(2n)!}{(n+1)!(n)!}$$ (Church and Patil, 1982)

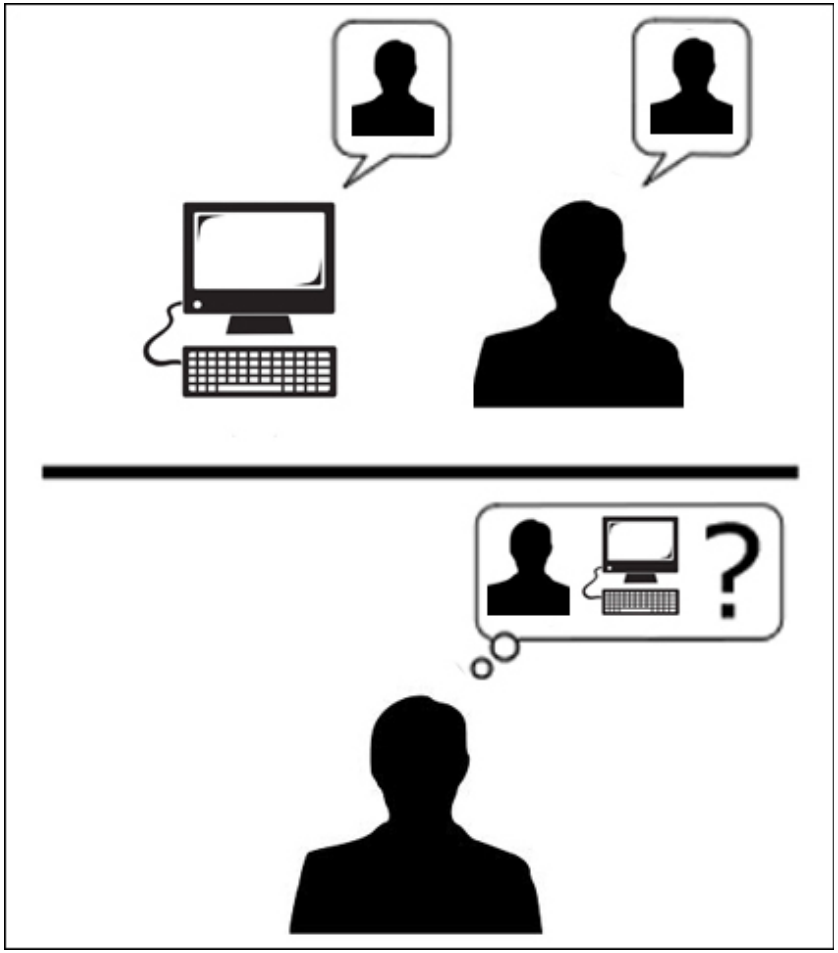
- 利用句法语义等知识约束求解
- 约束求解的困难

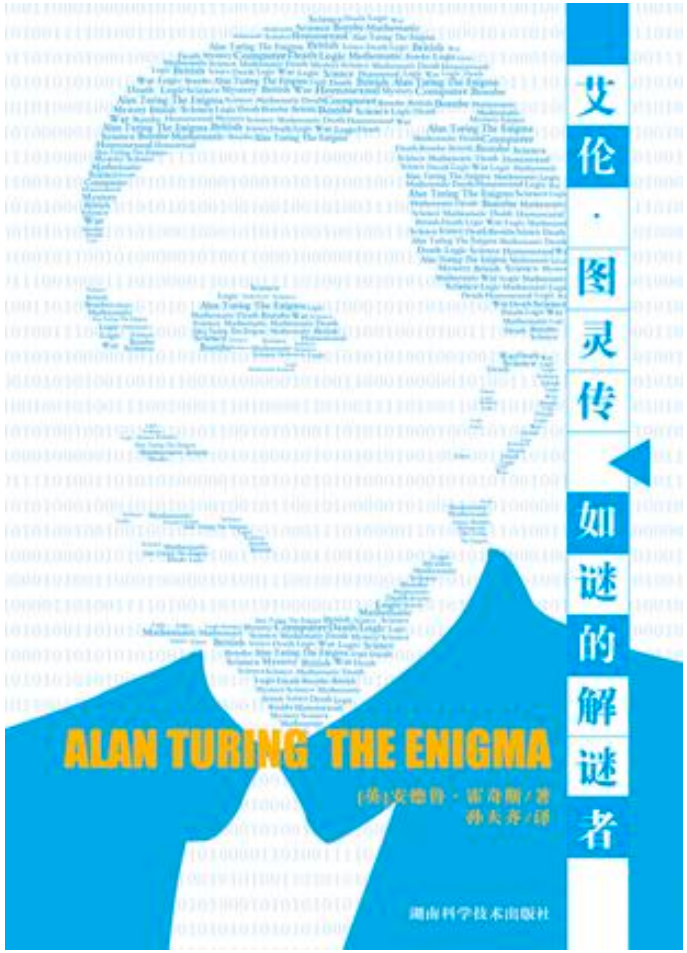
网络语言特点（以词汇为例）	例句
口语化	亲，看帖要回帖哦！
新奇性	走召弓虽（超强）
简约性	1314（一生一世）
诙谐性	菌男霉女
粗俗化	屌丝

句长	二分结构树数量
1	1
2	2
3	2
4	5
5	14
6	42
7	132
8	429
9	1, 430
10	4, 862
11	16, 796
12	58, 786
13	208, 012
14	742, 900
15	2, 674, 440
16	9, 794, 845
17	35, 357, 670
18	129, 644, 790
19	477, 638, 700
20	1, 767, 263, 190

自然语言理解的科学意义

- 图灵测试：测试机器是否具备人类智能的方法和准则





自然语言理解的科学意义

- 2011年，IBM沃森DeepQA自然语言问答系统在美国著名益智问答节目中战胜人类冠军选手
- 继 1997年深蓝战胜卡斯帕罗夫后提出的又一大挑战
- 语言理解是DeepQA的核心支撑技术

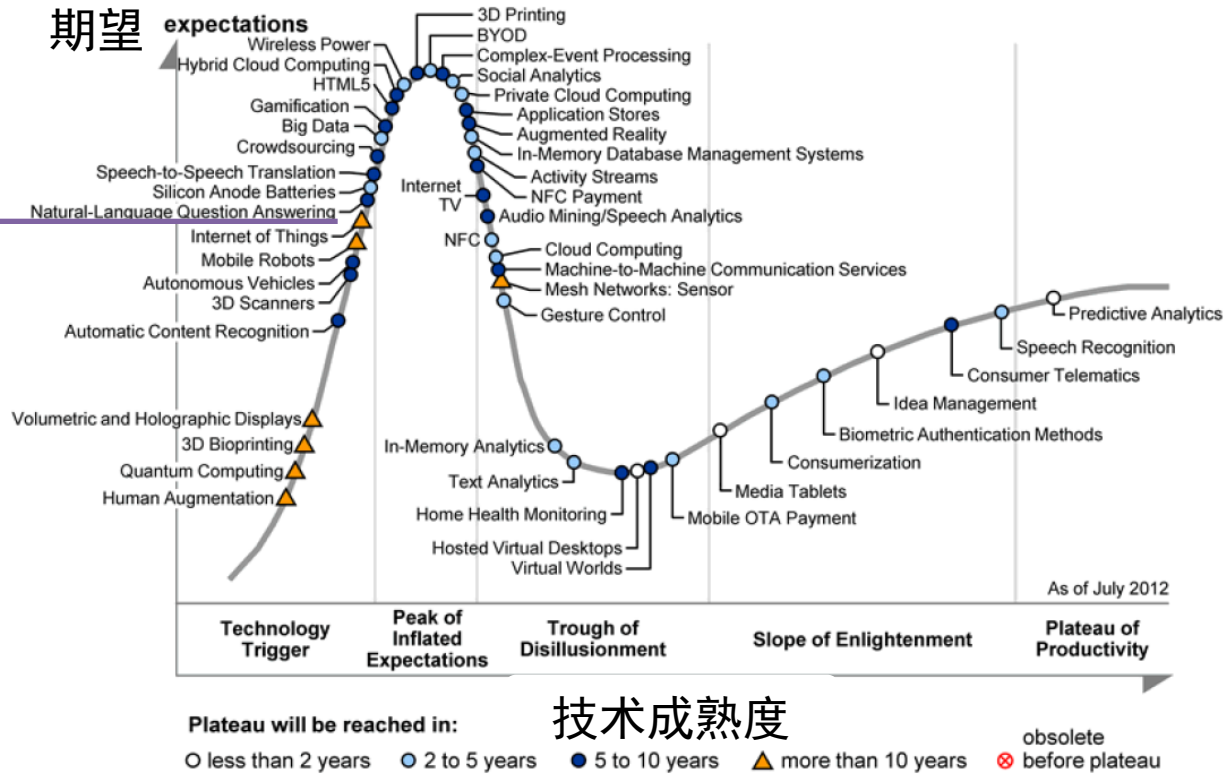
Q: Who was presidentially pardoned on September 8, 1974?
A: Nixon.

IBM 沃森DeepQA自然语言问答系统

自然语言理解的应用价值

- 自然语言问答是下一代搜索引擎基本形态（《Nature》2011.8）
- 世界权威IT市场调查咨询公司高德纳（Gartner）2012年8月发布《2012新兴技术成熟度曲线》

自然语言问答
重要性：变革（产业形态发生主要改变）
5-10年后将成为主流应用



重要性级别：
变革，高，中，低

自然语言理解的应用价值

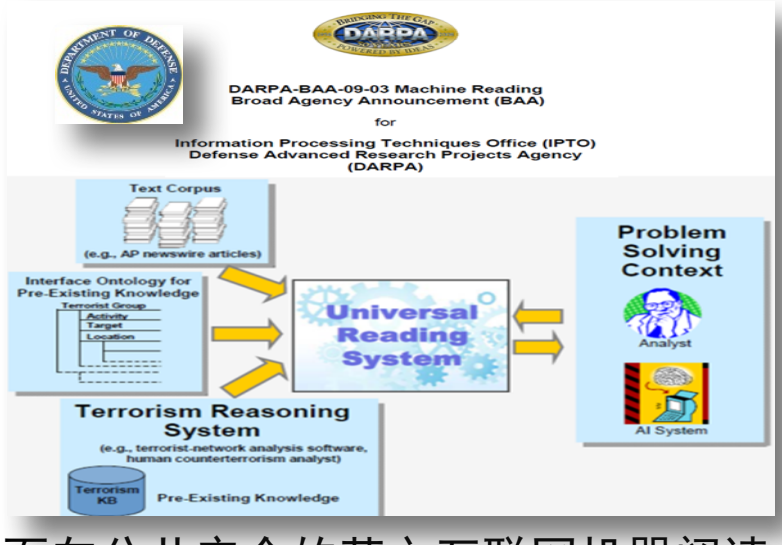
- 搜索引擎处于战略转型前夜，以自然语言问答为代表的语义搜索已如箭在弦
- 各大国际IT公司纷纷投入重金布局搜索产业
- 语言理解是下一代搜索的核心支撑技术
- 科大讯飞、百度、腾讯、阿里…

公司	收购/投资公司	产品类型	金额	收购时间
苹果	Siri	自然语言问答	约2亿美元	2010
HP	Autonomy	语义搜索	102亿美元	2011
微软	PowerSet	语义搜索	约1亿美元	2008
Google	MetaWeb	知识搜索	约1亿美元	2010
IBM	Watson	自然语言问答	约1亿美元	2011

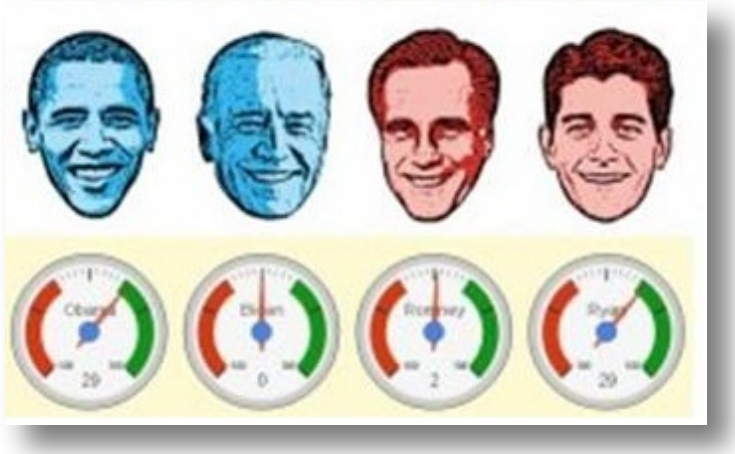
自然语言理解的应用价值

- 互联网机器阅读理解已成为美国公共安全（包括国家安全）与社会管理的核心支撑技术

项目名称	发布时间	启动时间	累计资助金额
机器阅读 Machine Reading	2007	2008	6740万美元
文本深度探测与过滤 Deep Exploration and Filtering of Text	2012	2013	2500万美元
低资源语言处理 Low Resource Languages for Emergent Incidents (LORELEI)	2014	2015	-



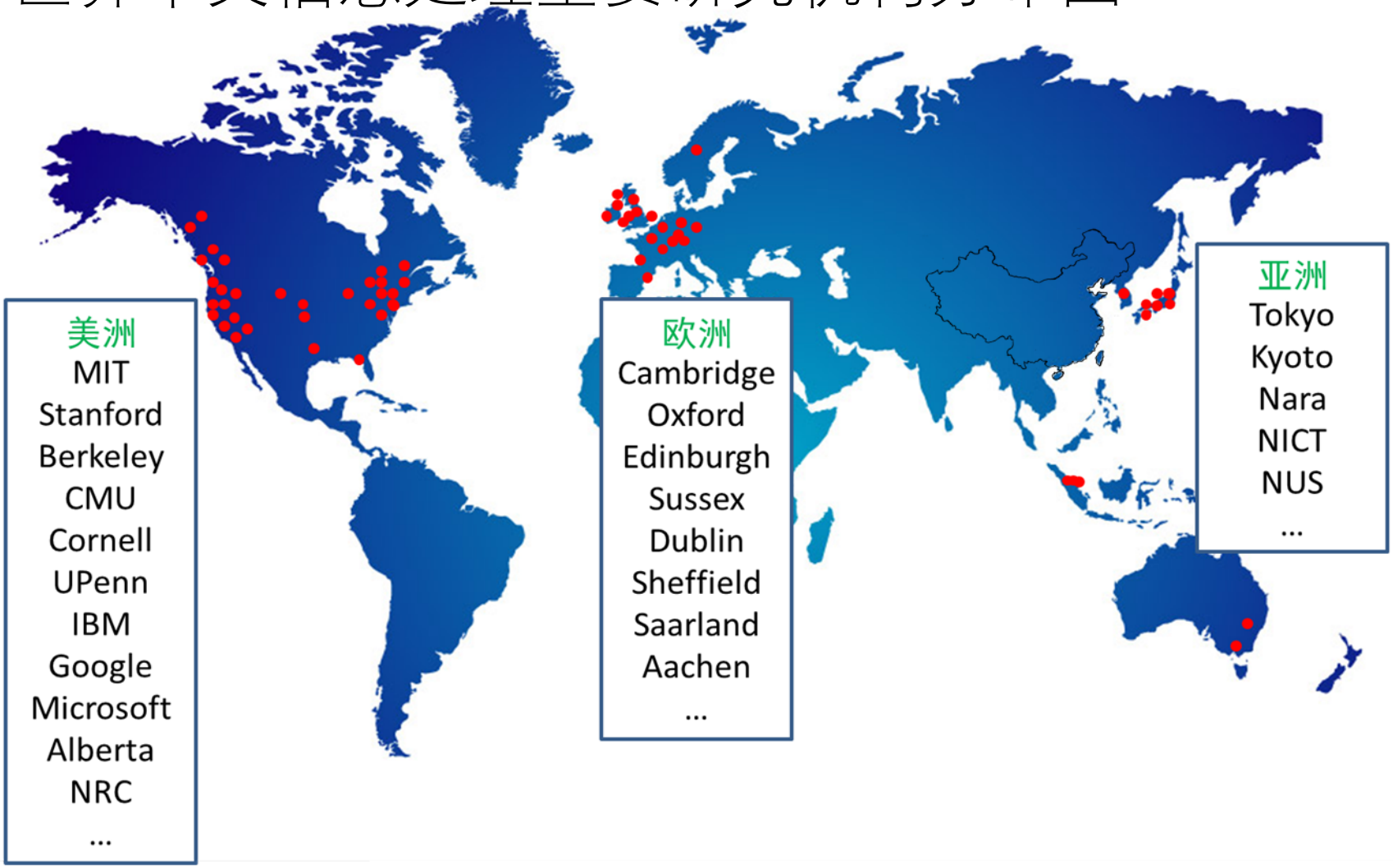
面向公共安全的英文互联网机器阅读



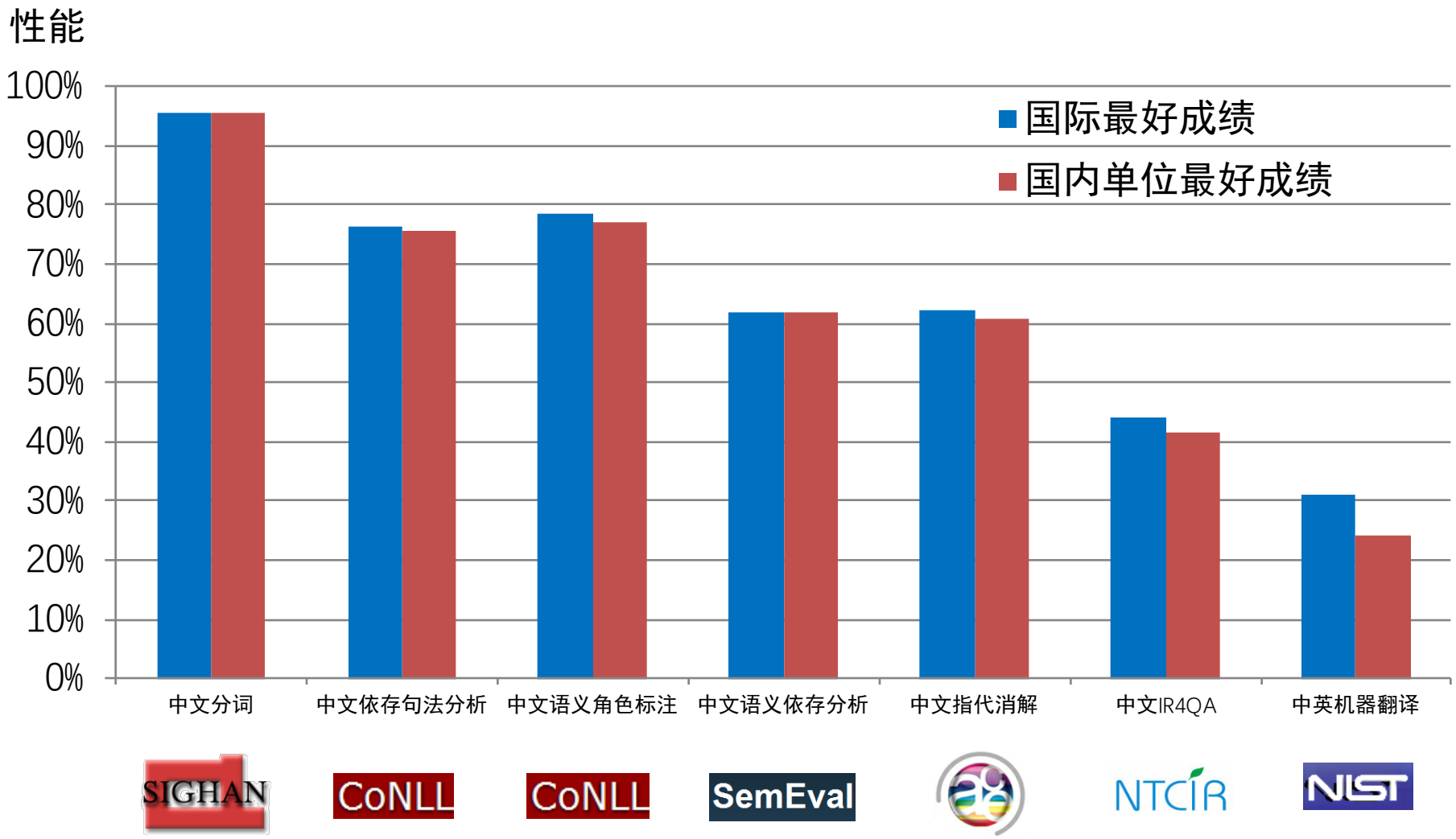
文本理解助力奥巴马成功连任
（美国《时代周刊》2012年11月）

中文自然语言理解的重要意义

- 美国国家语言战略将中文列为 “关键语言”
- 世界中文信息处理主要研究机构分布图



中文自然语言理解的重要意义



中文信息处理相关一系列国际公开评测结果

大数据时代的自然语言理解

- 人类 “管中窥豹式” 阅读难以形成对大数据完整准确的认识
- 机器阅读理解文本大数据是实现网络洞察力的关键

大数据时代自然语言理解重要应用

典型应用-搜索引擎

Bai du 百度 姚明夫人 百度一下 百度首页 消息 设置 liuzhiyuan1984

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约2,260,000个 搜索工具

姚明妻子：

叶莉

叶莉，1981年11月20日出生，上海人。1996年进入上海体育运动技术学院，1997年在上海体育运动技术学院，1998年入选国家青年队，1999年第一次入选国家队。 2007年8月3日，... 详情>>

来自百度百科 | 报错

姚明夫人_百度图片

image.baidu.com - 查看全部28,259张图片

姚明珊_百度百科

姓名：姚明珊

职业：退休

简介：姚明珊，女，现任中共中央政治局常委、中央纪律检查...

工作经历 感情

baike.baidu.com/

姚明多高 ?姚明的女儿多高?姚明的老婆多高?-3158江苏分站

2014年12月12日 - 姚明一直作为中国家喻户晓的体育明星,尽管他退役了,小编告诉您现在姚明女儿身高是多少,姚明女儿现在身高是不是已经超出了110,现在姚明女儿身高直逼郭敬明,姚明女儿...

jiangsu.3158.cn/info/2... - 百度快照 - 91%好评

相关人物 展开

姚沁蕾

姚明和叶莉的女儿

隋菲菲

中国篮球第一美女

姚明珊

王岐山夫人

赵蕊蕊

玻璃美人

易建联

CBA史上最年轻的MVP

刘翔

中国田径史里程碑人物

林书豪

纽约纳什

瓦妮莎·布莱恩特

科比最好的贤内助

张晗兰

亚运会冠军

唐子豪

NBA少年赛全场MVP

特雷西·麦克格雷迪

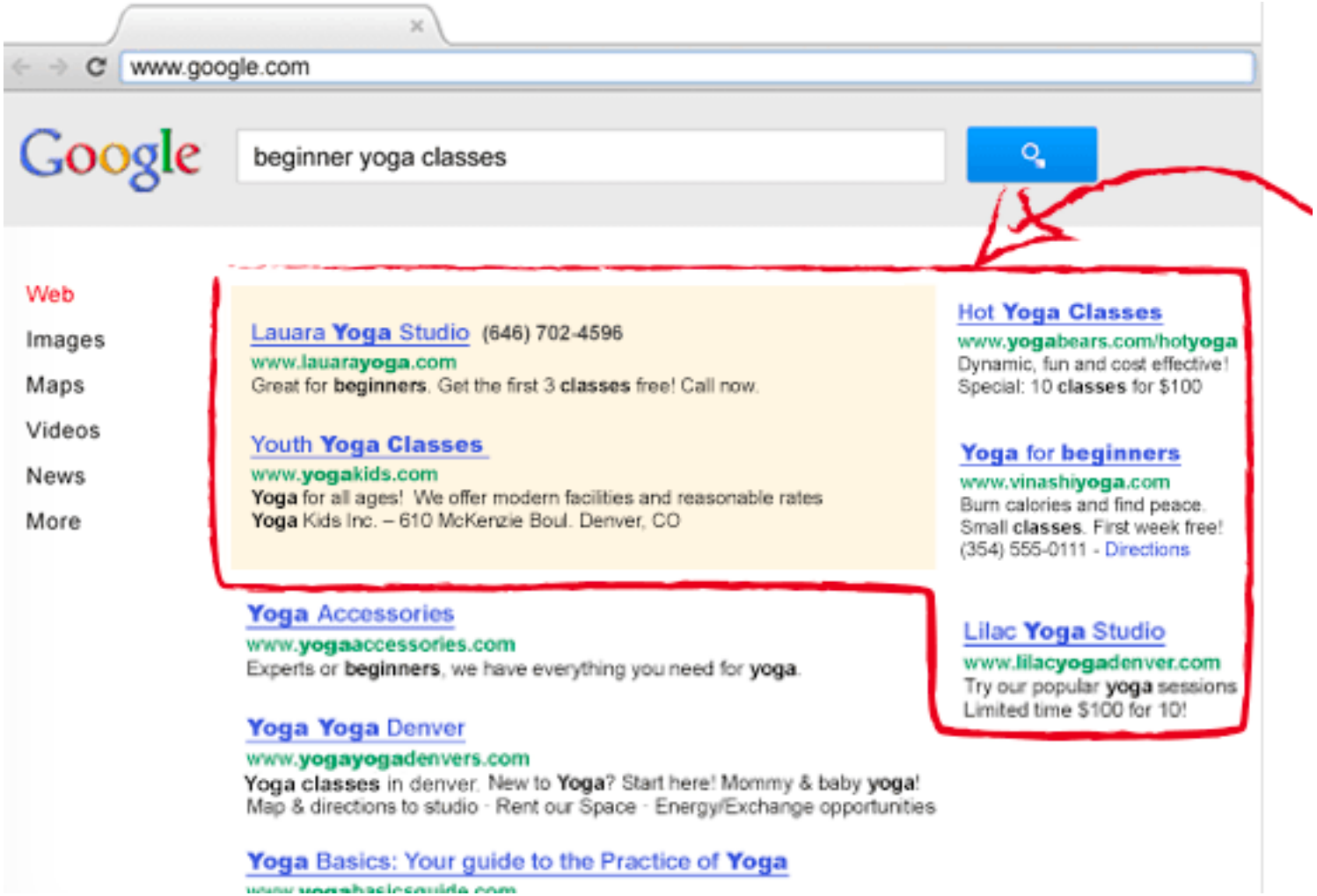
35秒13分上帝都哭了

史蒂夫·弗朗西斯

NBA年度最佳新秀奖

给百度提建议

典型应用-精准广告投放

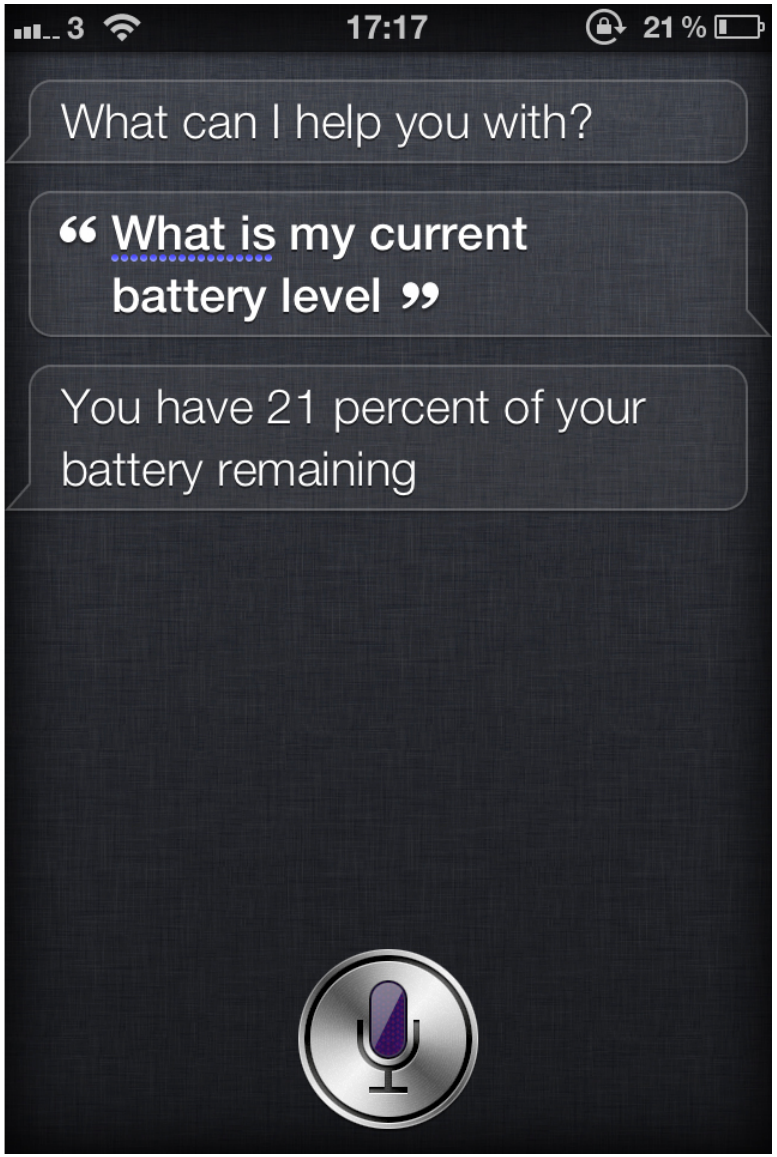

www.google.com
Google
beginner yoga classes
Web
Images
Maps
Videos
News
More
Lauara Yoga Studio (646) 702-4596
www.lauarayoga.com
Great for beginners. Get the first 3 classes free! Call now.
Youth Yoga Classes
www.yogakids.com
Yoga for all ages! We offer modern facilities and reasonable rates
Yoga Kids Inc. – 610 McKenzie Boul. Denver, CO
Yoga Accessories
www.yogaaccessories.com
Experts or beginners, we have everything you need for yoga.
Yoga Yoga Denver
www.yogayogadenvers.com
Yoga classes in denver. New to Yoga? Start here! Mommy & baby yoga!
Map & directions to studio · Rent our Space · Energy/Exchange opportunities
Yoga Basics: Your guide to the Practice of Yoga
Hot Yoga Classes
www.yogabears.com/hotyoga
Dynamic, fun and cost effective!
Special: 10 classes for $100
Yoga for beginners
www.vinashiyoga.com
Burn calories and find peace.
Small classes. First week free!
(354) 555-0111 - Directions
Lilac Yoga Studio
www.lilacyogadenver.com
Try our popular yoga sessions
Limited time $100 for 10!

典型应用-推荐系统

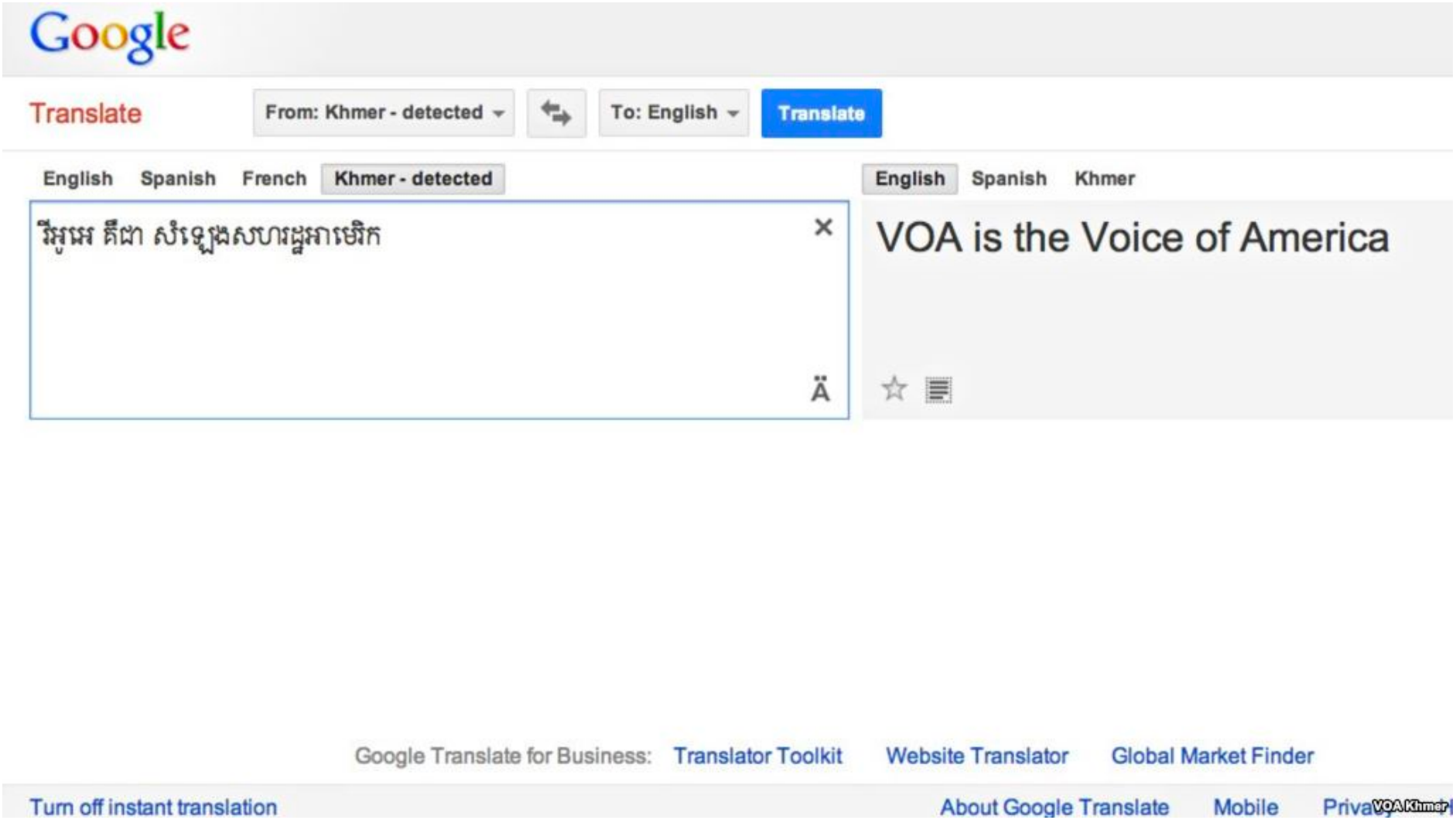

手机看新闻
用今日头条
今日头条
推荐
热点
北京
社会
娱乐
佳能中级发烧单反6D对比尼康D600，终极PK 听专家怎样讲解奥秘
李克强会见泰国公主诗琳通

典型应用-语音助手


3
17:12
22%
What can I help you with?
“ Change the backlight illumination level ”
What value shall I set your backlight to?
“ 40% ”
Setting backlight to 40 percent


3
17:17
21%
What can I help you with?
“ What is my current battery level ”
You have 21 percent of your battery remaining

典型应用-机器翻译


Google
Translate
From: Khmer - detected
To: English
Translate
English Spanish French Khmer - detected
English Spanish Khmer
វីអូអេ គឺជា សំឡេងសហរដ្ឋអាមេរិក
VOA is the Voice of America
Google Translate for Business: Translator Toolkit Website Translator Global Market Finder
Turn off instant translation
About Google Translate
Mobile
VOA Khmer

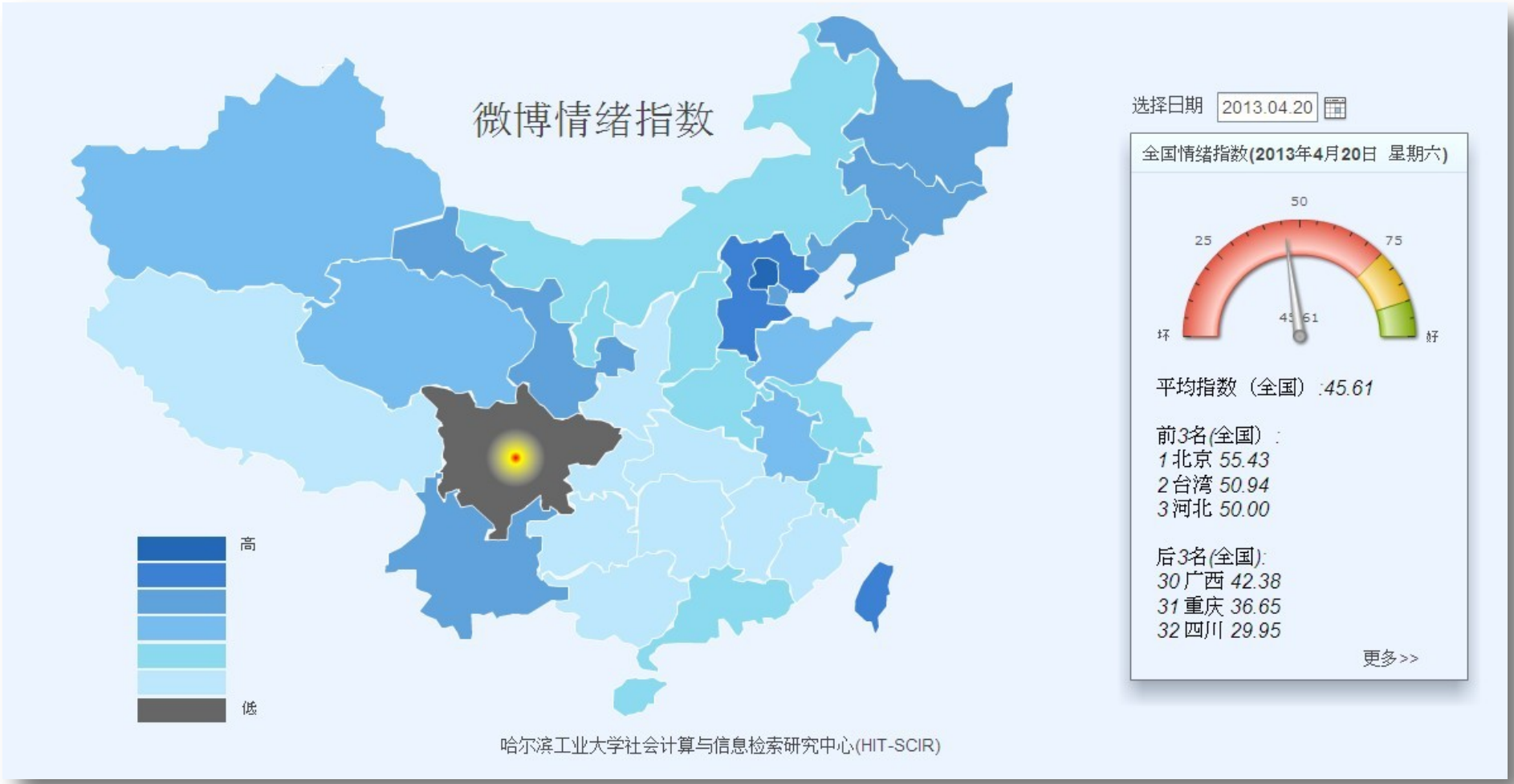
典型应用-文本摘要

Summly

- Summly is a news summarization app developed by Nick D'Aloisio-Montilla.
- Nick developed Summly at age of 15. Youngest person to receive venture funding. He was also Time's 100 most influential teenager.
- Summly used technology from SRI International.
- As for March 2013, Summly was sold to Yahoo! for reportedly $30M.

ideyatech
where ideas and technologies meet

典型应用-观点挖掘与情感分析

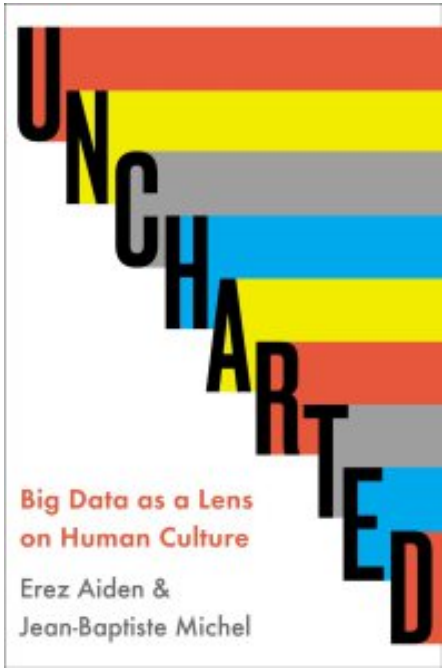


典型应用-关键词抽取


我的微博关键词
兴趣相投指数 new!
Ta的微博关键词
主题的微博关键词
好友关键词
制作微T恤
生成
点击生成微博关键词
图片由您最近发布的200条微博产生
分享到微博
用左图制作T恤
论文
收藏
关键词
清华
老师
google
windows
自然
计算机科学
计算机
网络
现场
北京指数
成功
内容
阅读
学术
同学
水平
功能
实验室
开源
大学生
幸福
大家庭
英国人
一家子
工具
动画
社会实践
职业
数学家
办公室
高中生
改革开放
志愿者
未来
人工智能
凡高
操作系统
兴趣
成语
原创
影响力
网页
文艺
故事
好友
书法
清新
信息学
逻辑
维基
兴奋剂
葫芦娃
价值观
用户
素质
杂志
威武
时间
诗文
分析
冠名权
陈佩斯
研讨会
新加坡
中国人
代码
美国
统计
搜索
世界
视频
企业
人生
微软
工作
苏宁
诺基亚
鸦片战争
计算机系
运动员
学习
兆赫
代表队
外星人
温馨
记者
大牛
欧文
于右任
舞台剧

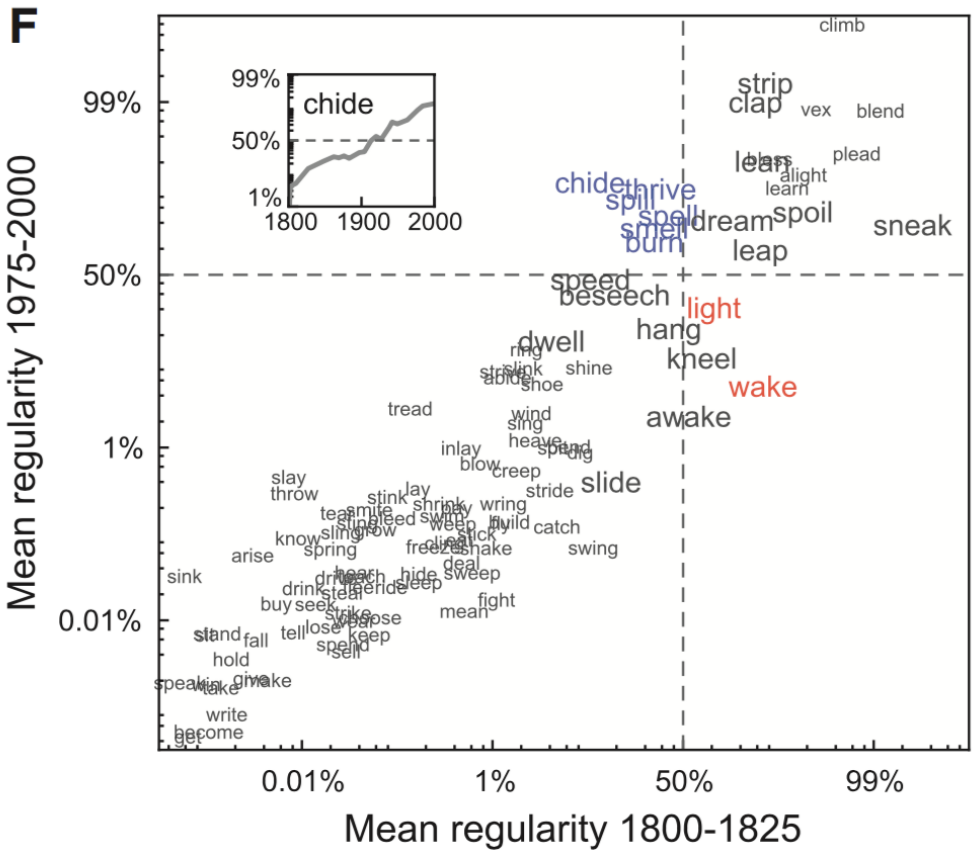
计算社会科学研究案例

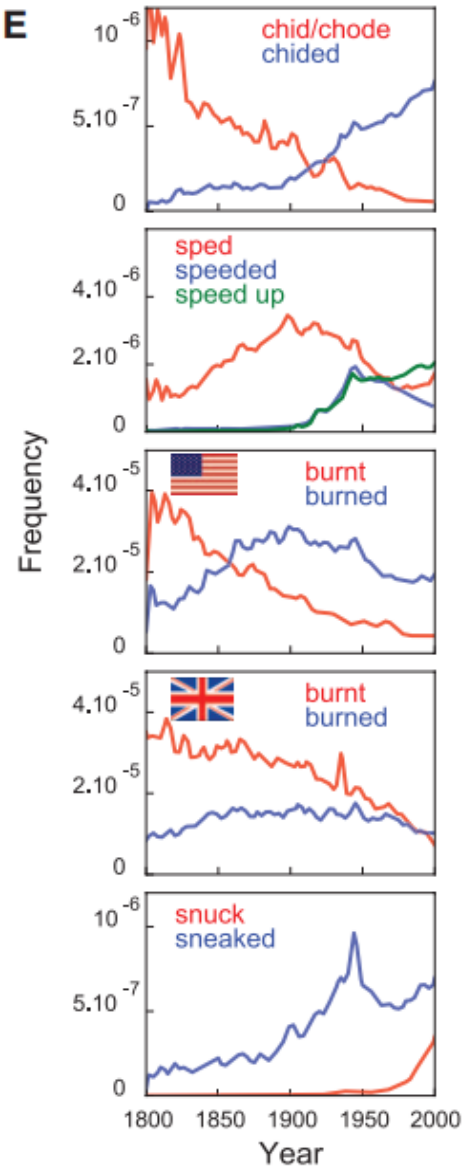
- 文化组学（culturomics）http://www.culturomics.org/
- 哈佛大学研究团队利用Google Books收集并扫描识别的1800年到2000年之间的500万种出版物（占人类所有出版物的4%），通过不同关键词使用频度随时间的变化，分析人类文化演进特点
- Google Book N-grams (https://books.google.com/ngrams)



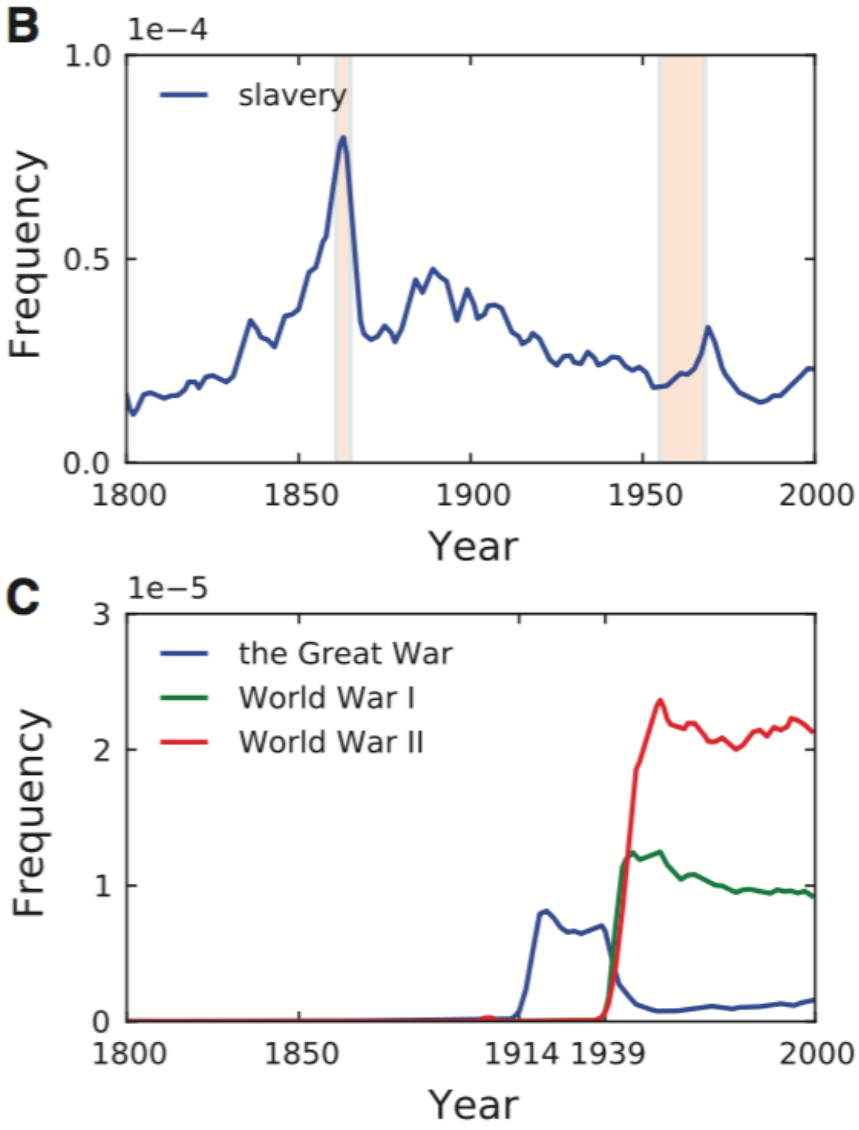
计算社会科学研究案例

- 他们发现在过去几百年里英语中越来越多的不规则动词演化成了规则动词



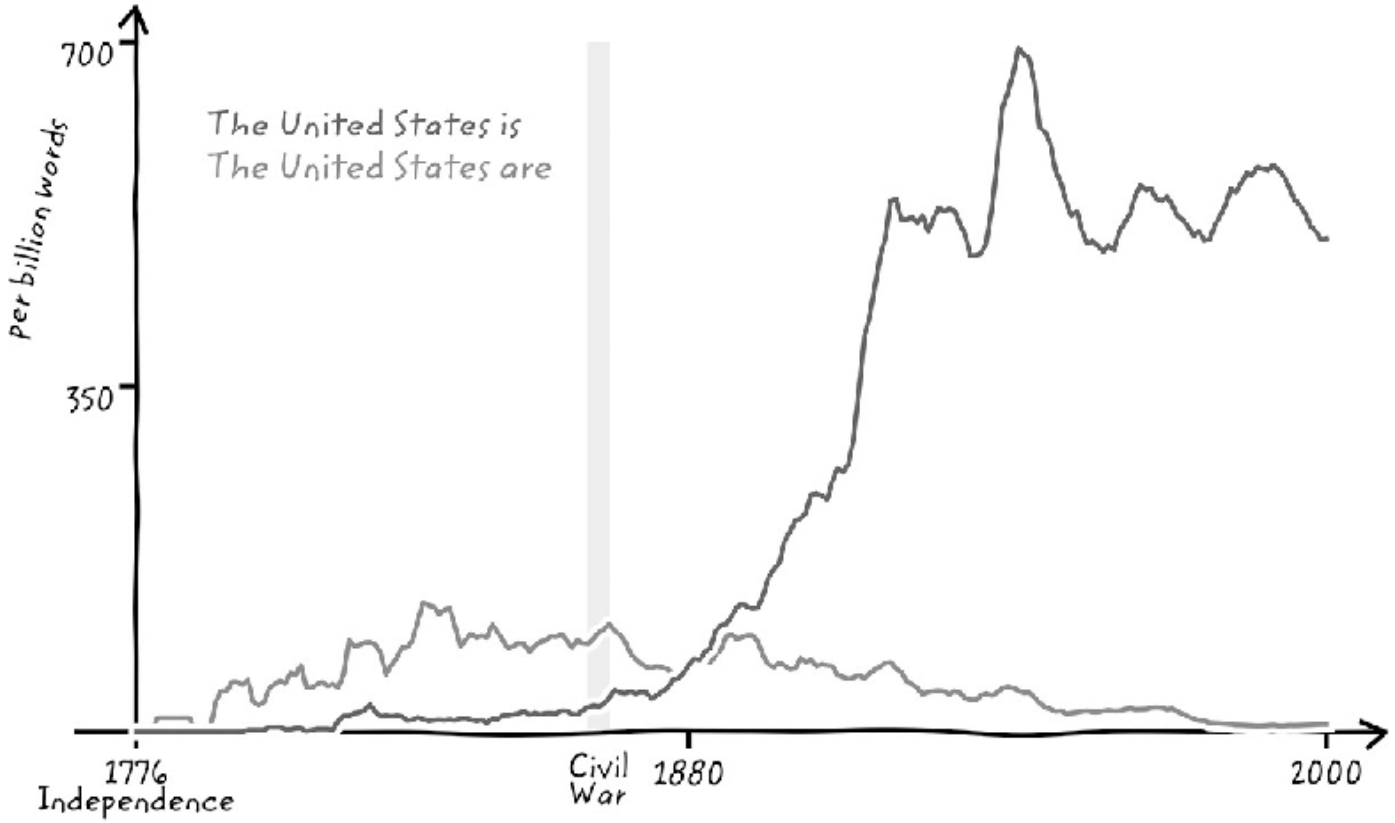


计算社会科学研究案例


B
1e−4
1.0
0.5
0.0
Frequency
slavery
1800
1850
1900
1950
2000
Year
C
1e−5
3
2
1
0
Frequency
the Great War
World War I
World War II
1800
1850
1914 1939
2000
Year

计算社会科学研究案例

- 通过Google Books历年使用 “The United States is” 和 “The United States are” 的统计趋势图，定量分析美国作为统一国家概念是如何形成的



计算社会科学研究案例

QUANTITATIVE SOCIAL SCIENCE

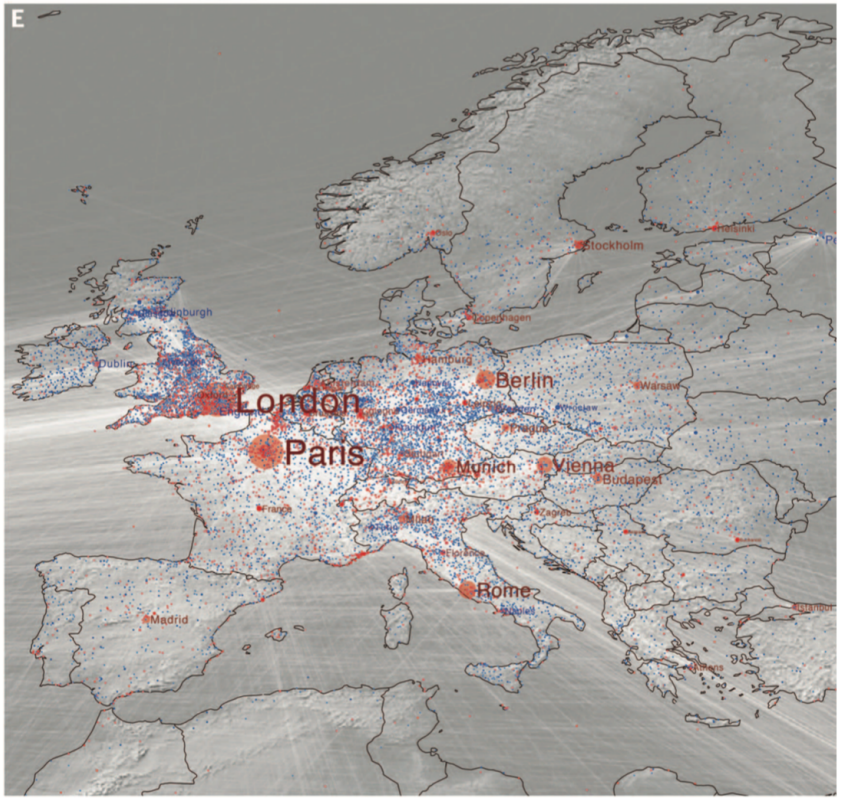
A network framework of cultural history

Maximilian Schich,[1,2,3*] Chaoming Song,[4] Yong-Yeol Ahn,[5] Alexander Mirsky,[2]
Mauro Martino,[3] Albert-László Barabási,[3,6,7] Dirk Helbing[2]

名人

出生地点 → 死亡地点

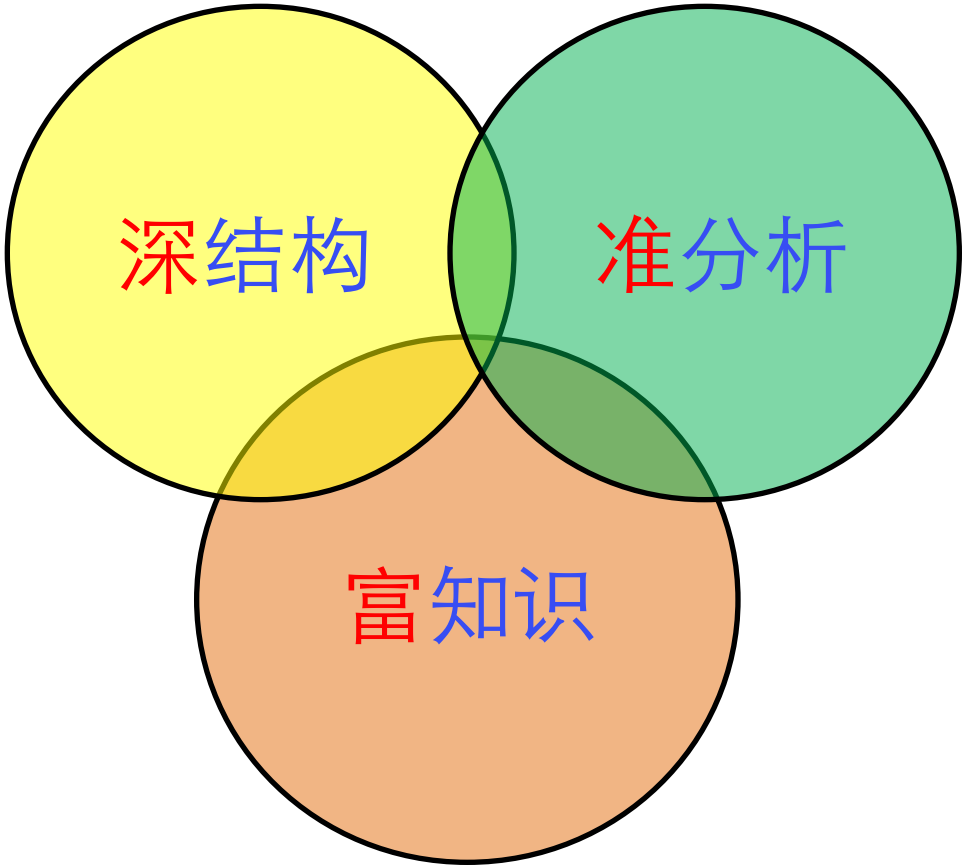
Winckelmann Corpus, 18世纪人物



Freebase

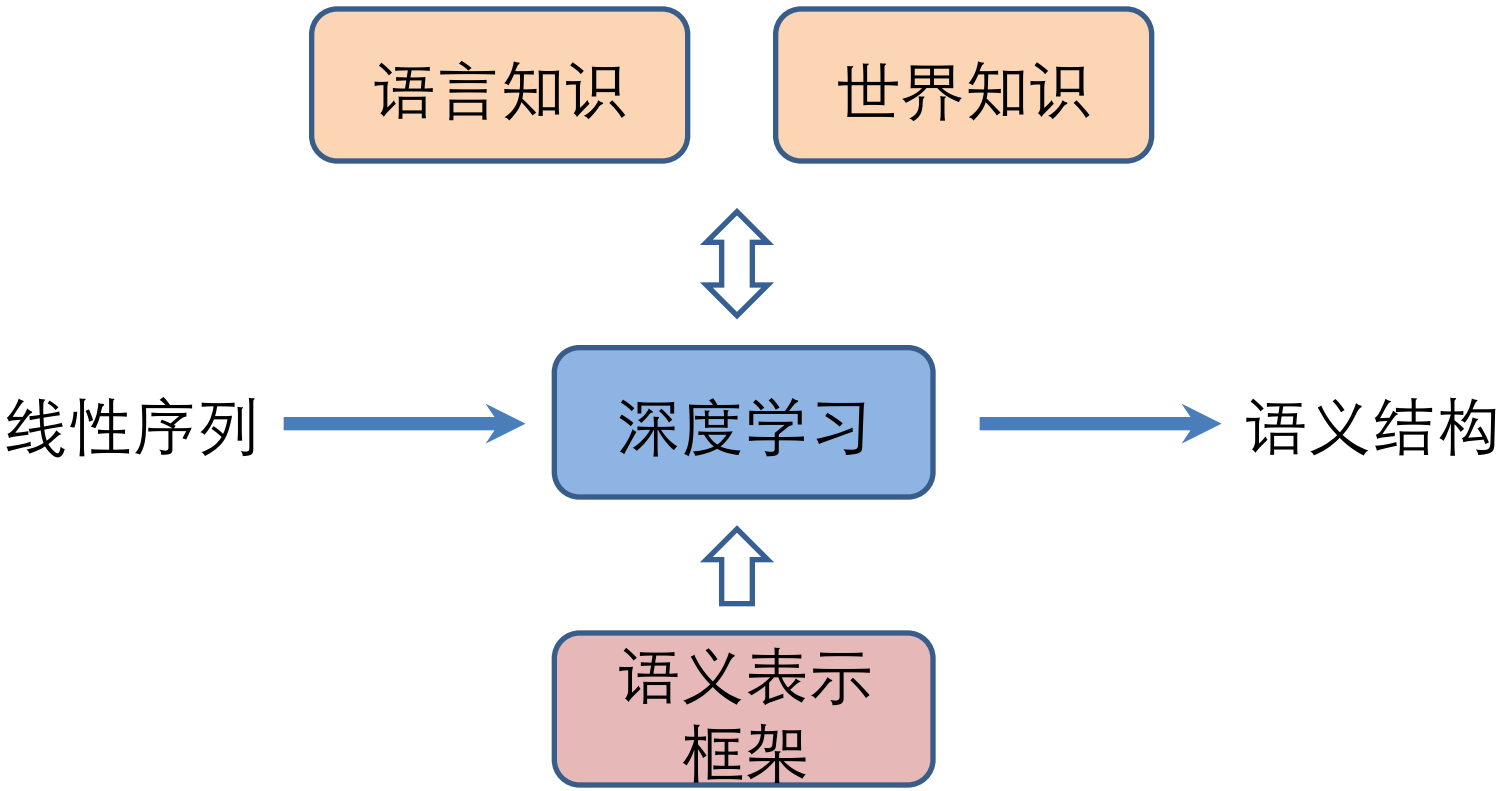
大数据时代自然语言理解研究趋势

中文NLP发展技术趋势（1）



互联网环境

中文NLP发展技术趋势（1）



中文NLP发展技术趋势（1）

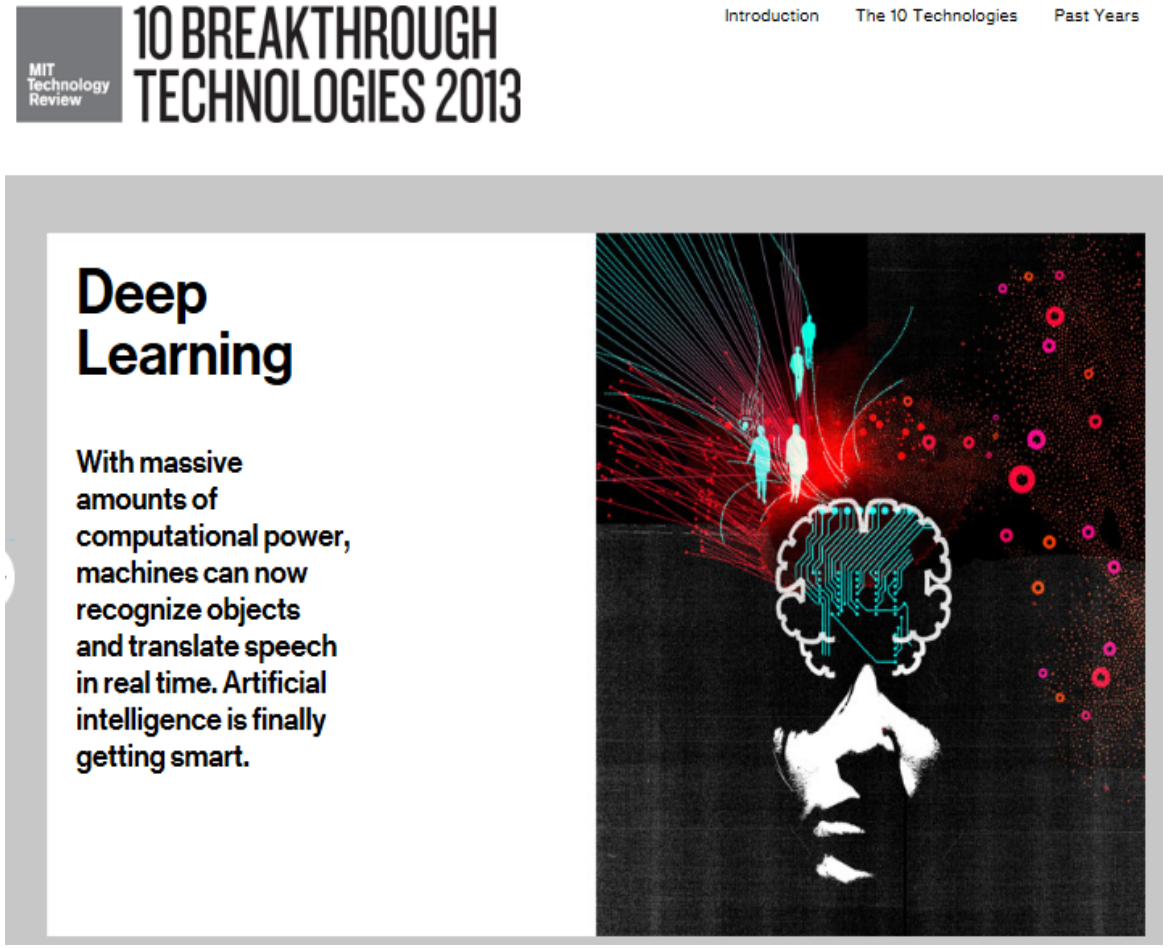
- 大规模语言知识/世界知识图谱
- 面向中文NLP的深度学习技术
- 面向深度理解的语义表示框架

深度学习

- 深度学习：通过学习出模型的“深层结构”对数据中存在的复杂关系进行建模

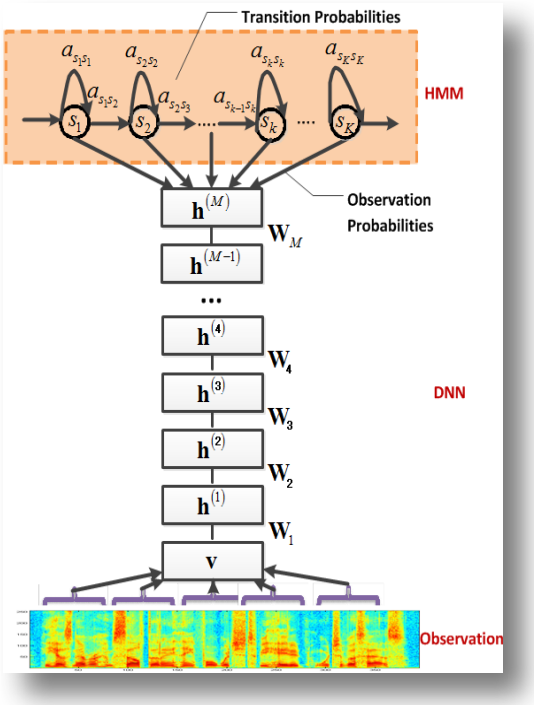
Geoffrey Hinton
深度信念网络DBN（2006）
英国皇家学会院士

Judea Pearl
概率图模型（ 2011年获图灵奖）
美国工程院院士



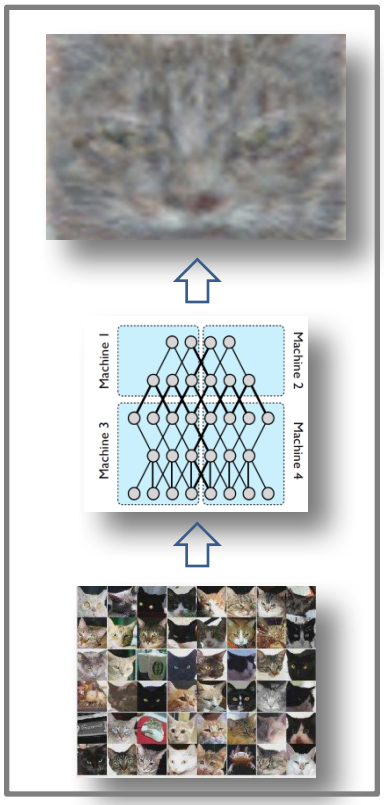
深度学习

- 优良的计算性质：可望突破“表层结构”的限制，适合小规模有标注样本和极大规模无标注样本的融合学习
- 深度学习在英文语音识别和图像识别中取得突破



微软语音识别

错误率减少30%以上



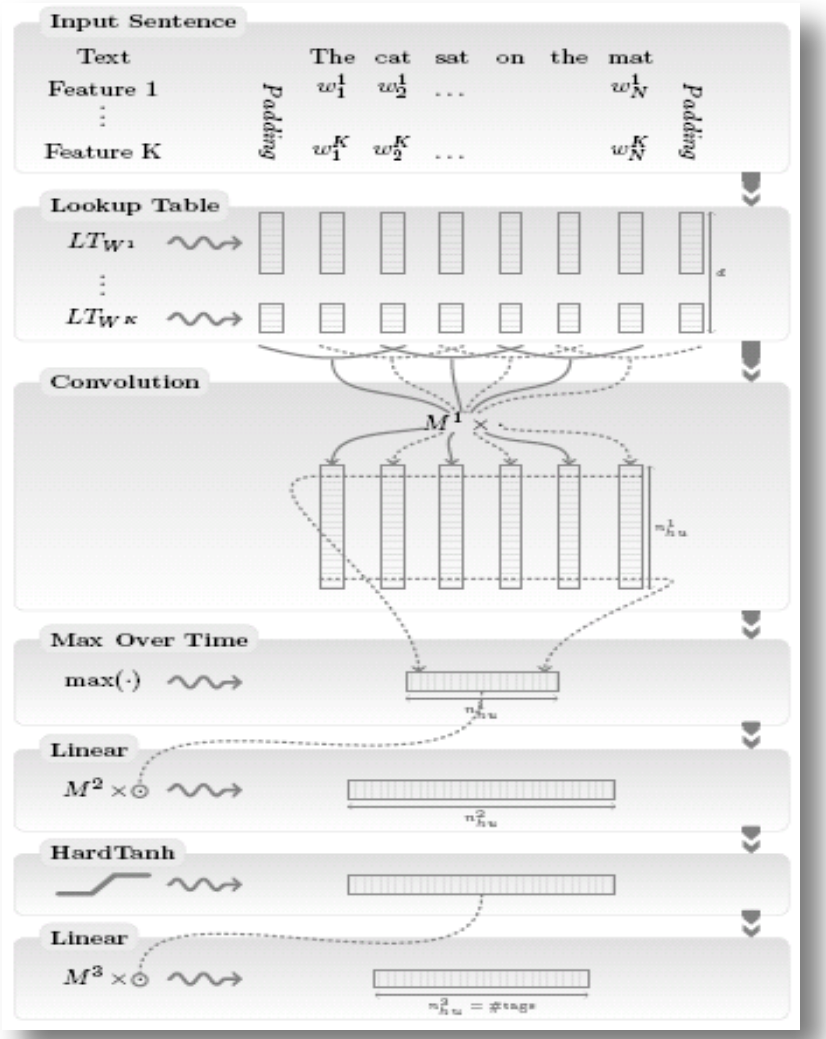
16000多个处理器、10亿个内部连接组成的“虚拟大脑”，从1000万帧YouTube的无标签图片中自主“学会”了猫的概念。

谷歌虚拟大脑

（Google Brain）

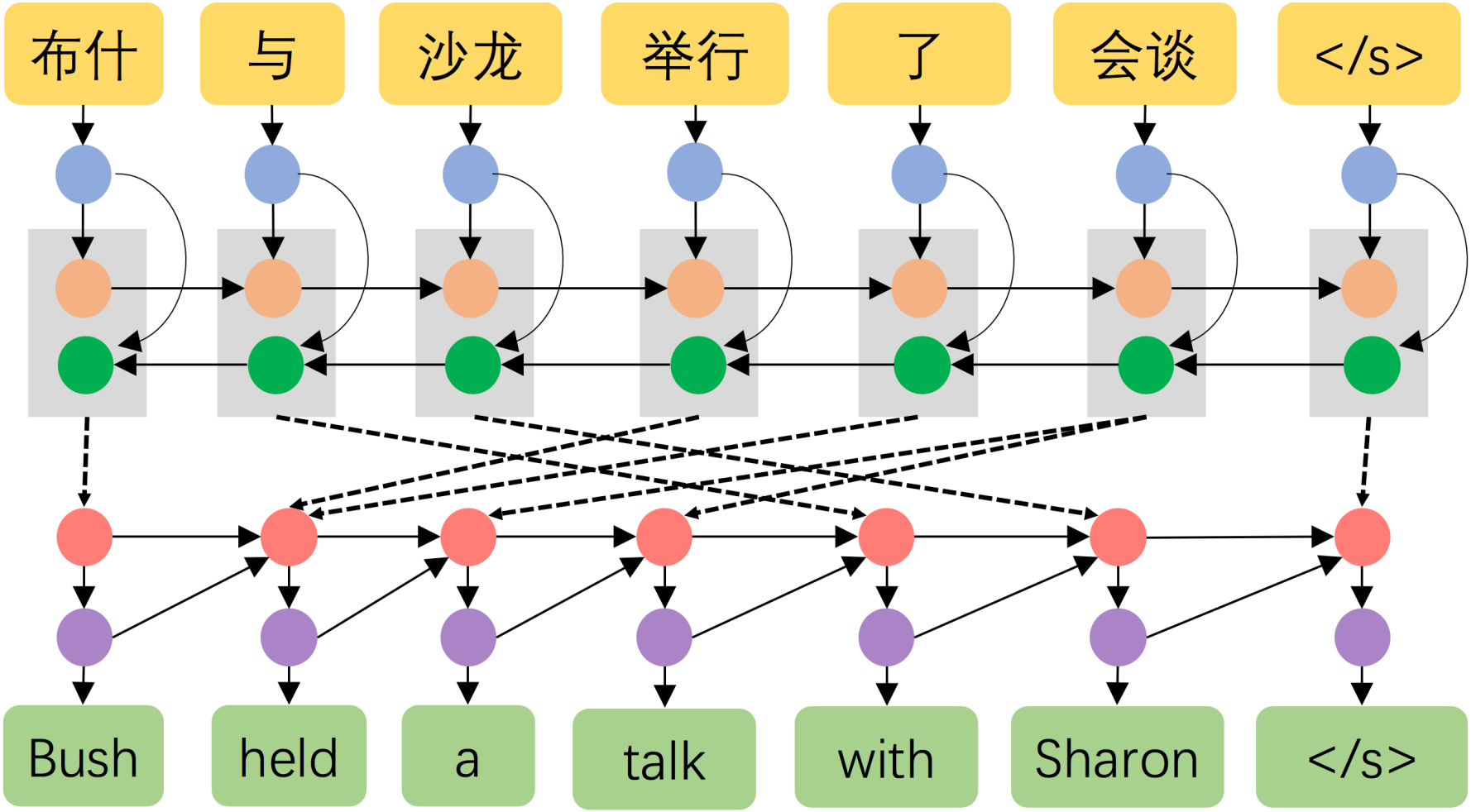
深度学习

- 巨大反差：针对语言理解的深度学习尚未取得成功
  - 语音图像：基于视觉或音频的“底层认知特征”
  - 语言理解：基于词法、句法和语义等“高层认知特征”
- 创新思路
  - 高层认知特征的表示及其学习
  - 适合于语言计算的大规模神经网络模型



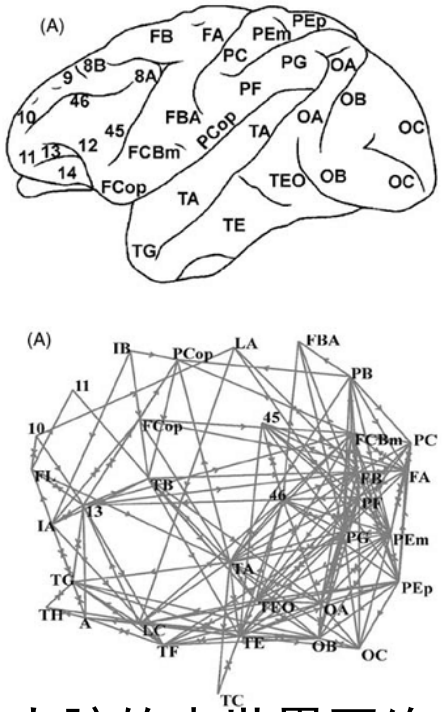
在多项自然语言处理任务中与主流方法结果具有可比性

神经机器翻译

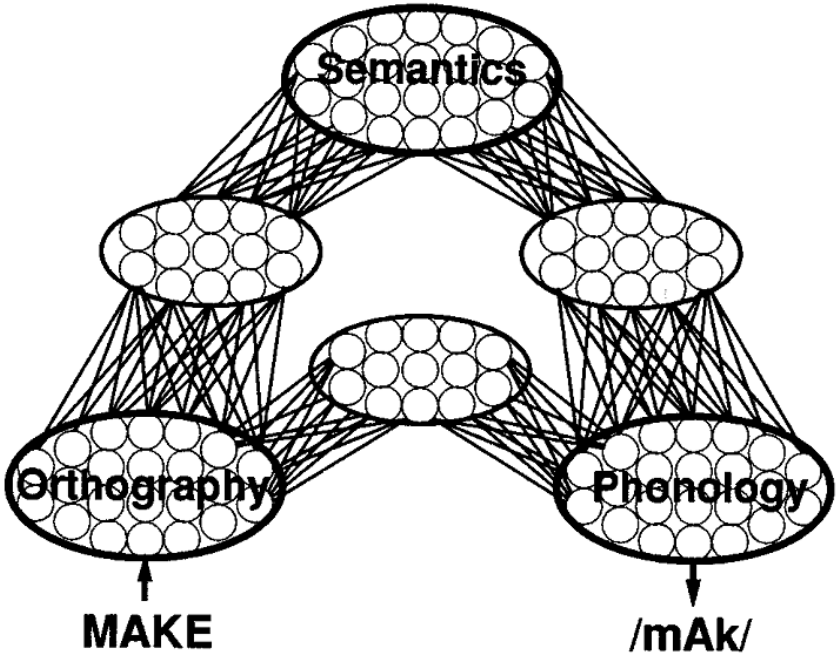


深度学习

- 认知机制的研究有可能为从功能上模拟人脑语言加工机制，突破现有神经网络结构的深度学习模型提供启示



大脑的小世界网络
(Bassett and Bullmore, 2006)



大脑的词汇处理框架
(Plaut, 1996)

深度学习互联网自然语言理解具有重大理论创新空间！

知识图谱

- 互联网中文理解需要大规模、高覆盖率的知识资源
- 目前的知识资源难以满足中文理解的需求，以Google知识图谱（5亿个实体，35亿个事实）为例：
  - 主要描述实体以及实体之间关系，对于复杂事件的描述甚少
  - 英文知识图谱关于中国的内容很少
  - 中文知识图谱正在构建中，主要挑战之一是infobox信息匮乏
- 百度知识图谱与搜狗知立方也面临类似的问题

Google知识图谱

Language	Article	Infobox	Percentage
English	4064524	1554096	38.24%
German	1834301	348531	19.00%
French	1734147	503467	29.03%
Dutch	1176405	409703	34.83%
Spanish	1151934	508282	44.12%
Chinese	499405	108673	21.76%
Baidu	4779013	168236	3.52%
Hudong	1991184	390678	19.62%

维基百科仅有21%的中文文章有infobox

知识图谱

DBpedia

250概念
4M实例
6000属性
500M三元组
在线更新

yago
select knowledge

350K概念
10M实例
100属性
120M三元组

搜狗搜索

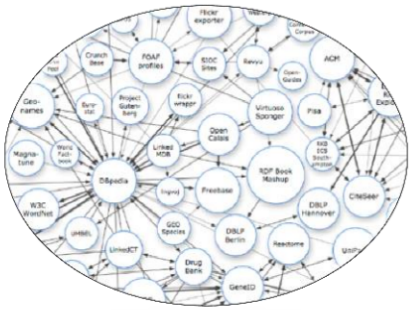
Baidu 百度

Freebase

15K概念
40M实例
4000属性
1B三元组
Google KB核心

BabelNet 1.1.1
A very large multilingual ontology

50M义项
50+种语言
262M三元组



WolframAlpha

NELL

XLORE

850K概念
8M实例
70K属性

OpenIE
(Reverb, OLLIE)

Google KG

15K概念
600M实例
20B三元组

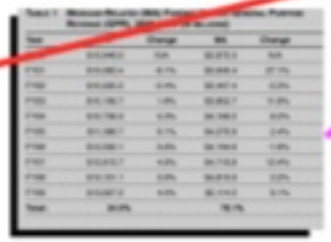
WordNet
7种欧洲语言
跨语言链接

EuroWordNet

知识图谱

Knowledge Vault* fuses all these signals together

- Data from web
  - Unstructured text
  - Semi-structured DOM trees
  - Structured WebTables
- "Prior" data from FB



* Details in a paper submitted to WWW'14 (Dong et al)

知识图谱

- **基于知识图谱的语义标注**

Barack Obama

44th U.S. President

Barack Hussein Obama II is the 44th and current President of the United States, and the first African American to hold the office. Wikipedia

Born: August 4, 1961 (age 53), Honolulu, Hawaii, United States

Spouse: Michelle Obama (m. 1992)

Office: President of the United States since 2009

Presidential term: January 20, 2009 –

Parents: Ann Dunham, Barack Obama, Sr.

Siblings: Maya Soetoro-Ng, Mark Okoth Obama Nuesandjo, more

Recent posts on Google+

Barack Obama
4,769,268 followers • Shared publicly

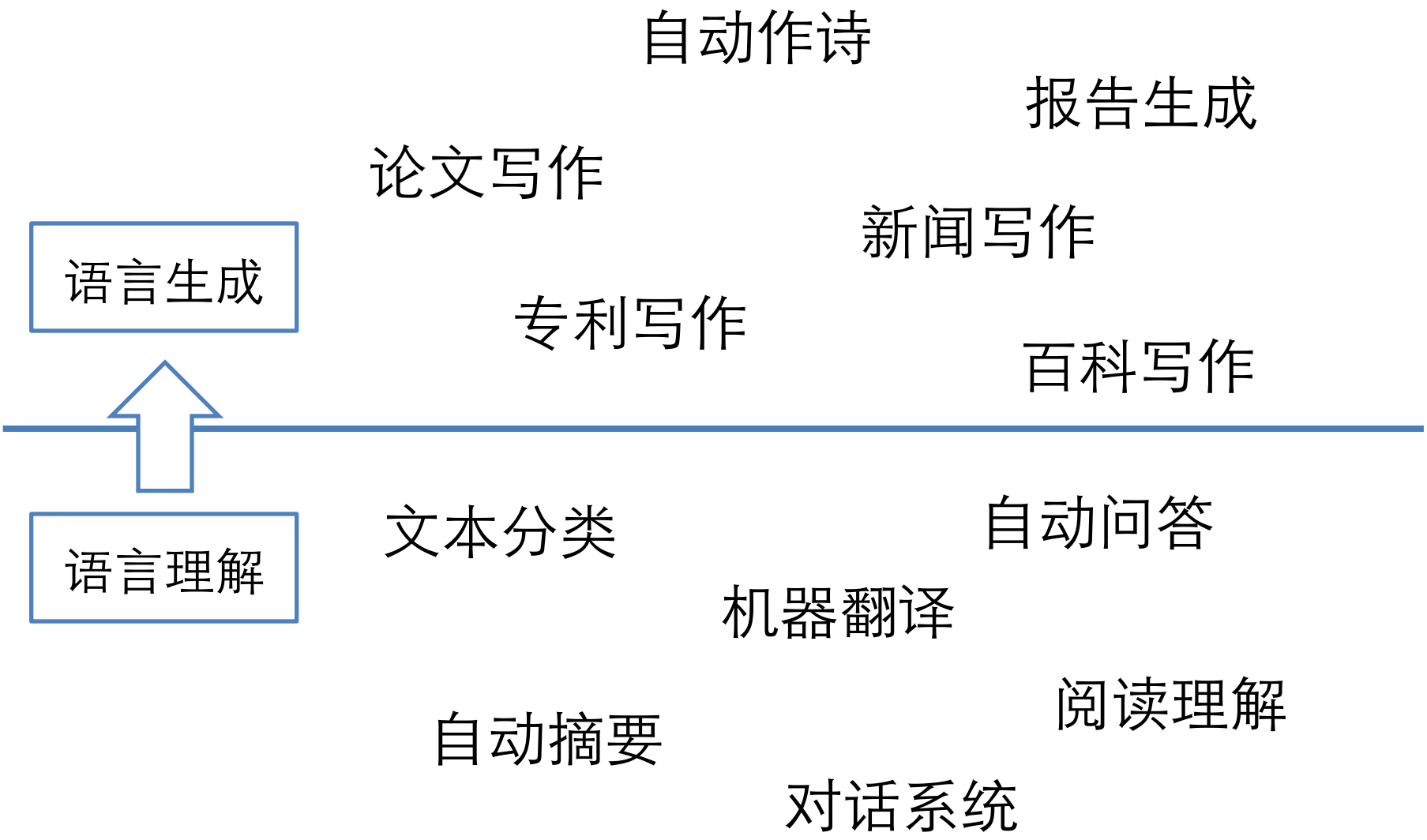
"It's long past time for us to raise the minimum wage." —President Obama Add your name if you agree: http://ofa.bo/b1Do #With1010 10 Oct 2014

图像

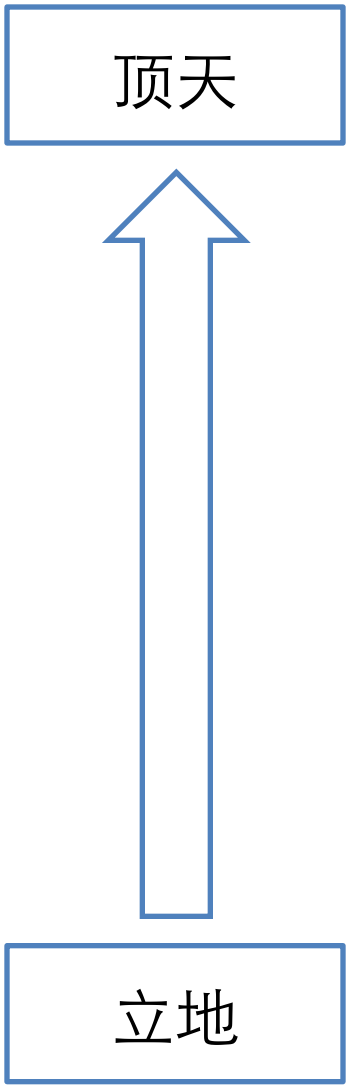
社交主页

文本

中文NLP发展技术趋势（2）



中文NLP发展技术趋势（3）



图灵测试

文本生成

自动问答　　机器翻译

情感分析

搜索引擎　　文本分类

新图灵测试

Oren Etzioni

Search needs a shake-up

2011年

Paul G. Allen

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

My Computer is an Honor Student - but how Intelligent is it? Standardized Tests as a Measure of AI

2016年

新图灵测试

- 测试目标
  - 能够回答多类问题
  - 能够回答复杂问题
  - 能够利用常识知识和世界知识
  - 能够学习新知识

- 形式要求
  - 清晰可评价
  - 难度有层次
  - 不轻易被骗
  - 有意思

- 希望机器具备的能力
  - 语言理解
  - 世界建模

- 目前水平：非图表、四项多选题
  - 75% (4th grade), 63% (8th grade), and 41% (12th grade)

新图灵测试

- 基本问题

(1) Which object is the best conductor of electricity? (A) a wax crayon (B) a plastic spoon (C) a rubber eraser (D) an iron nail

(2) The movement of soil by wind or water is called (A) condensation (B) evaporation (C) erosion (D) friction

(3) Which part of a plant produces the seeds? (A) flower (B) leaves (C) stem (D) roots

- 简单推理问题

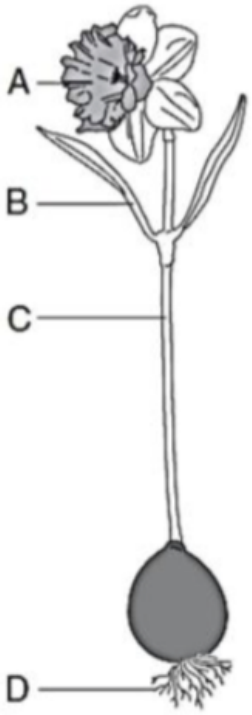
(4) Which example describes an organism taking in nutrients? (A) dog burying a bone (B) A girl eating an apple (C) An insect crawling on a leaf (D) A boy planting tomatoes in the garden

- 复杂推理问题

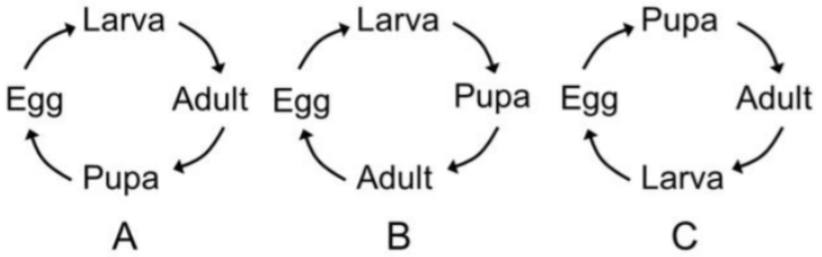
(6) A student puts two identical plants in the same type and amount of soil. She gives them the same amount of water. She puts one of these plants near a sunny window and the other in a dark room. This experiment tests how the plants respond to (A) light (B) air (C) water (D) soil

新图灵测试

- 图表题



Which letter in the diagram points to the plant structure that takes in water and nutrients?



Which diagram correctly shows the life cycle of some insects?

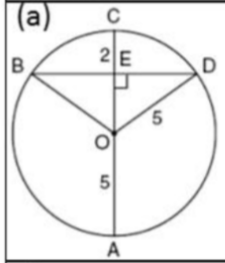
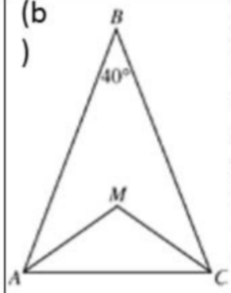
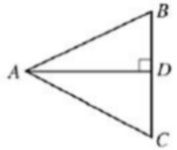
新图灵测试

Euclid

Solving math and geometry problems.

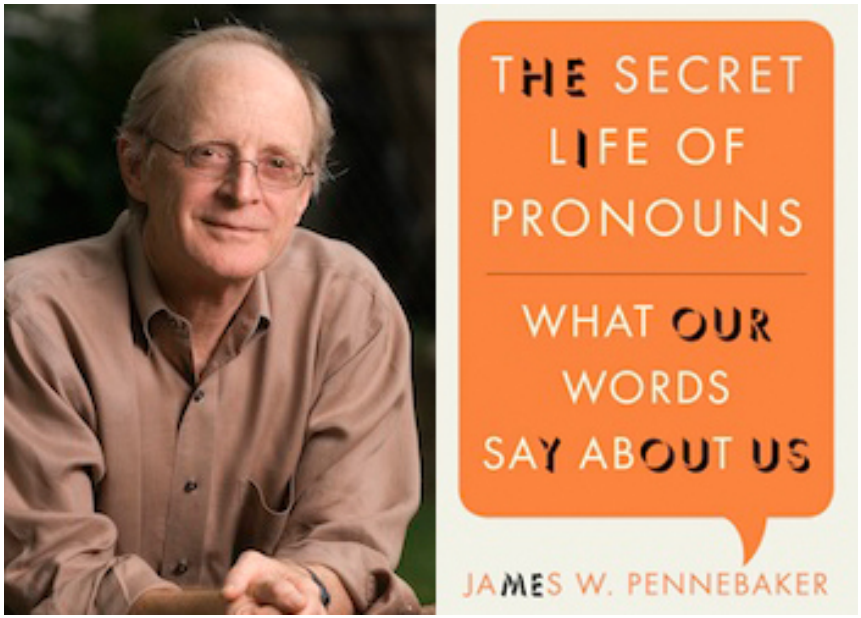
- 数学题

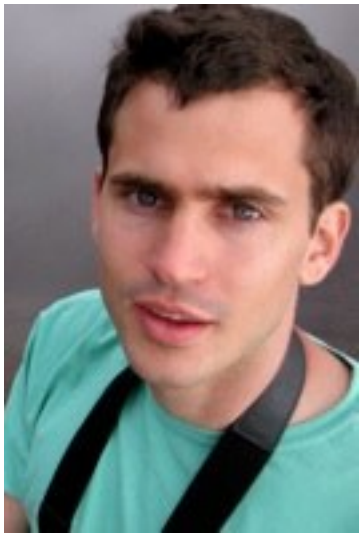
Problems and equations
John had 20 stickers. He bought 12 stickers from a store in the mall and got 20 stickers for his birthday. Then John gave 5 of the stickers to his sister and used 8 to decorate a greeting card. How many stickers does John have left? ((20 + ((12 + 20) - 8)) - 5) = x
Maggie bought 4 packs of red bouncy balls, 8 packs of yellow bouncy balls, and 4 packs of green bouncy balls. There were 10 bouncy balls in each package. How many bouncy balls did Maggie buy in all? x = (((4 + 8) + 4) * 10)
Sam had 79 dollars to spend on 9 books. After buying them he had 16 dollars. How much did each book cost? 79 = ((9 * x) + 16)
Fred loves trading cards. He bought 2 packs of football cards for $2.73 each, a pack of Pokemon cards for $4.01, and a deck of baseball cards for $8.95. How much did Fred spend on cards? ((2 * 2.73) + (4.01 + 8.95)) = x

Questions		Interpretations
(a) 	In the diagram at the left, circle O has a radius of 5, and CE = 2. Diameter AC is perpendicular to chord BD. What is the length of BD?	*Equals(RadiusOf(O), 5)* *IsCircle(O)* *Equals(LengthOf(CE), 2)* *IsDiameter(AC)* *IsChord(BD)* *Perpendicular(AC), BD)* *Equals(what, Length(BD))* correct a) 12 b) 10 (c) 8 d) 6 e) 4
(b) 	In isosceles triangle ABC at the left, lines AM and CM are the angle bisectors of angles BAC and BCA. What is the measure of angle AMC?	*IsIsoscelesTriangle(ABC)* *BisectsAngle(AM, BAC)* *IsLine(AM)* *CC(AM, CM)* *CC(BAC, BCA)* *IsAngle(BAC)* *IsAngle(AMC)* *Equals(what, MeasureOf(AMC))* correct (a) 110 b) 115 c) 120 d) 125 e) 130
(c) 	In the figure at left, The bisector of angle BAC is perpendicular to BC at point D. If AB = 6 and BD = 3, what is the measure of angle BAC?	*IsAngle(BAC)* *BisectsAngle(line, BAC)* *Perpendicular (line, BC)* *Equals(LengthOf(AB), 6)* *Equals(LengthOf(BD), 3)* *IsAngle(BAC)* *Equals(what, MeasureOf(BAC))* correct a) 15 b) 30 c) 45 (d) 60 e) 75

中文NLP发展技术趋势（4）

- 吸收来自社会语言学、认知科学的研究成果





Cristian Danescu-Niculescu-Mizil

No country for old members: User lifecycle and linguistic change in online communities
with Dan Jurafsky, Jure Leskovec, Christopher Potts. WWW 2013. Best Paper Award.

中文NLP发展技术趋势（4）

- 吸收来自社会语言学、认知科学的研究成果

- 作者：Jack L. Gallant等
- 单位：美国加州伯克利大学

- 将985个常见英语词汇的对应大脑区域画了出来。7名志愿者躺在功能性核磁共振(fMRI)中两个多小时，过程中给他们播放The Moth Radio Hour有一万多字的故事

- 注：Thomas L. Griffiths是LDA Gibbs Sampling算法发明者

中文NLP发展技术趋势（4）

- 吸收来自社会语言学、认知科学的研究成果

- 词汇们分布在大脑四周，没有绝对的语言区域
- 意义相关的词语所激活的大脑区域相似
- 与词义对应的大脑区域呈双脑对称，这与过去一直以为的「左脑负责语义」的认识相悖
- 这份大脑词汇地图在人与人之间一致性很高

大数据时代自然语言理解研究建议

中文NLP重要支撑

- 应用与资源并重

共享任务评测

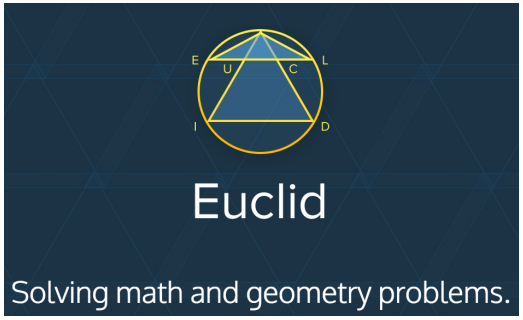
语言标注资源

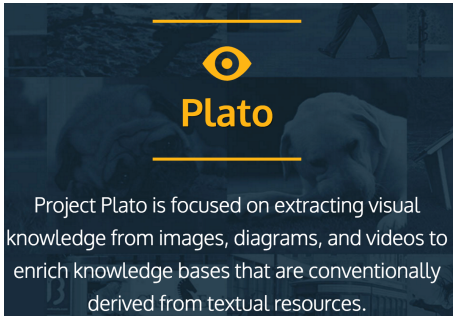
开源NLP工具

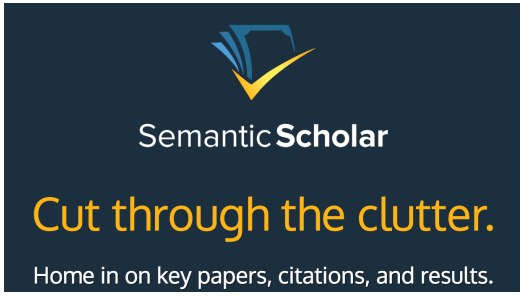
中文NLP研究方向

- 立地与顶天并重

AlphaGo

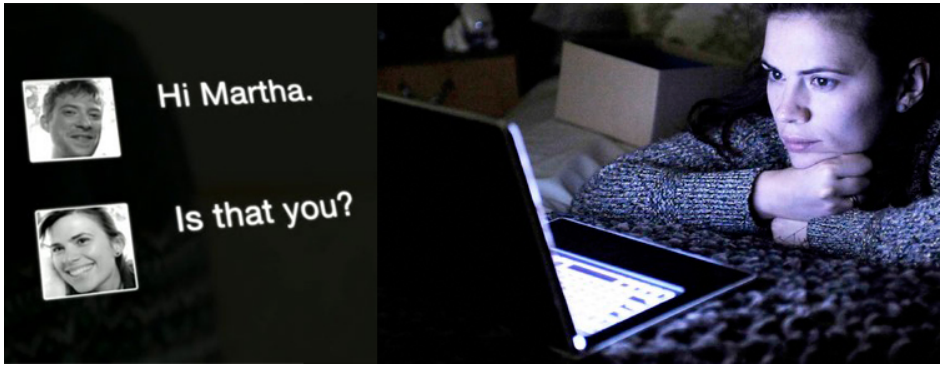






中文NLP建议研究领域

- 对话系统与聊天机器人
  - 立地：明星聊天机器人，自动与粉丝互动
  - 顶天：图灵测试



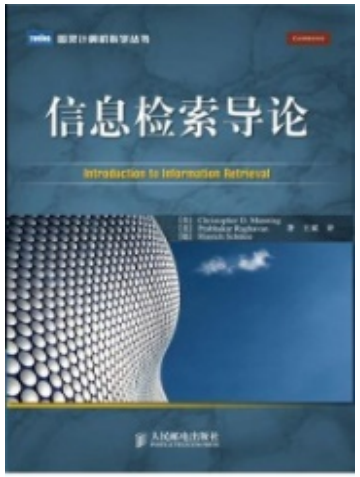
- 知识图谱
  - 立地：事实纠错、谣言识别、阅读理解、智能问答
  - 顶天：基于知识库的智能推理

奥巴马作为中国的领导人

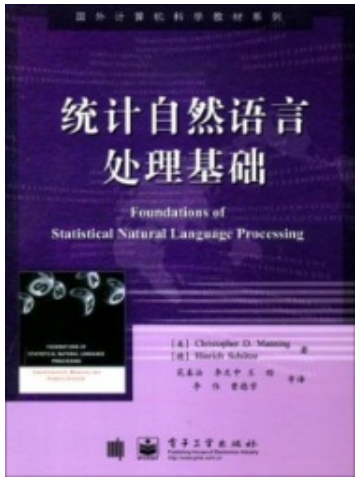
提示：您是指 美国 吗?

推荐阅读

推荐书目



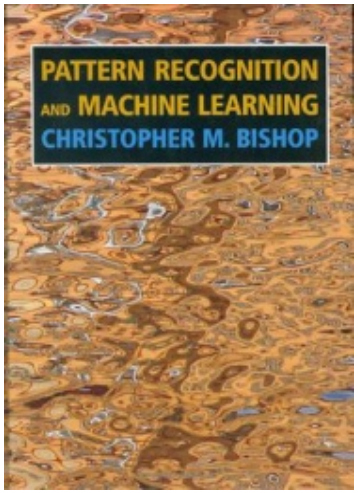
信息检索导论
原作名: Introduction to Information Retrieval
作者: Christopher D.Manning / Hinrich Schutze / Prabhakar Raghavan
译者: 王斌；出版社: 人民邮电出版社



统计自然语言处理基础
原作名：Foundations of Statistical Natural Language Processing
作者: Chris Manning / Hinrich Schütze
译者: 苑春法 / 李伟 / 李庆中
出版社: 电子工业出版社；出版年: 2005-01-01；页数: 432

推荐书目

统计学习方法
作者: 李航
出版社: 清华大学出版社
出版年: 2012-3； 页数: 235； 定价: 38.00元；
ISBN: 9787302275954



Pattern Recognition and Machine Learning
作者: Christopher M. Bishop
出版社: Springer
出版年: 2007-10-1
页数: 738

Advances in Natural Language Processing

- Julia Hirschberg，哥伦比亚大学计算机系主任，AAAI、ACL会士

- Christopher Manning，斯坦福大学计算机系教授，ACM、AAAI、ACL会士，Google Scholar引用超过5万次

推荐学术网站

- Google Scholar: http://scholar.google.com/
- ACM Portal: http://dl.acm.org/
- ACL Anthology: http://www.aclweb.org/anthology/
  - ACL、EMNLP、NAACL、COLING
  - CCL、NLPCC

THANK YOU

http://nlp.csai.tsinghua.edu.cn/~lzy

liuzy@tsinghua.edu.cn