《语义计算与知识检索》研究生课程

语义计算概述与基础漫谈

万小军

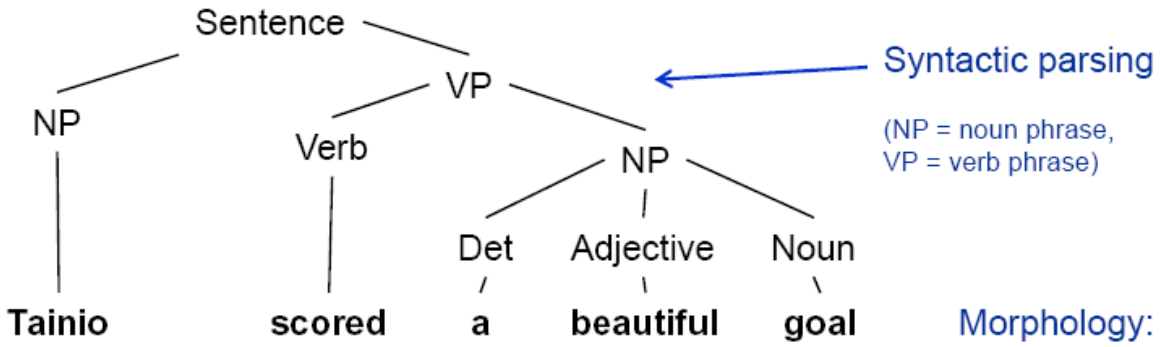
北京大学语言计算与互联网挖掘组
http://www.icst.pku.edu.cn/lcwm

2017年2月22日

自然语言处理回顾

This model shows what a man's body would look like if each part grew in proportion to the area of the cortex of the brain concerned with its movement.


Language generation
Semantic analysis (or, understanding)
Tainio scored a beautiful goal!


Sentence
NP
VP
Verb
NP
Det
Adjective
Noun
Tainio
scored
a
beautiful
goal
Syntactic parsing
(NP = noun phrase, VP = verb phrase)
Morphology: score-d

原始句子

警察正在详细调查事故原因

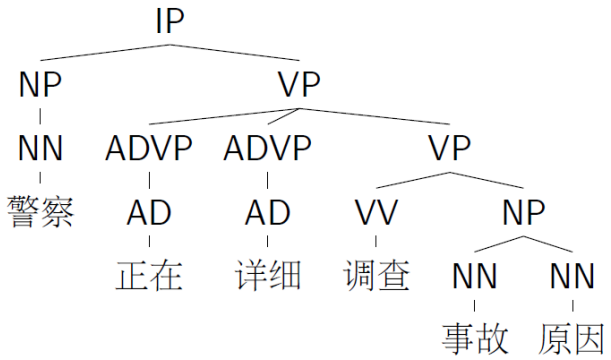
分词结果

警察 / 正在 / 详细 / 调查 / 事故 / 原因

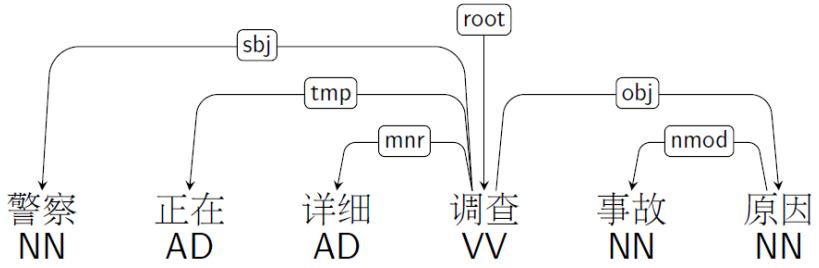
词性标注结果

警察/NN 正在/AD 详细/AD 调查/VV 事故/NN 原因/NN

短语结构树



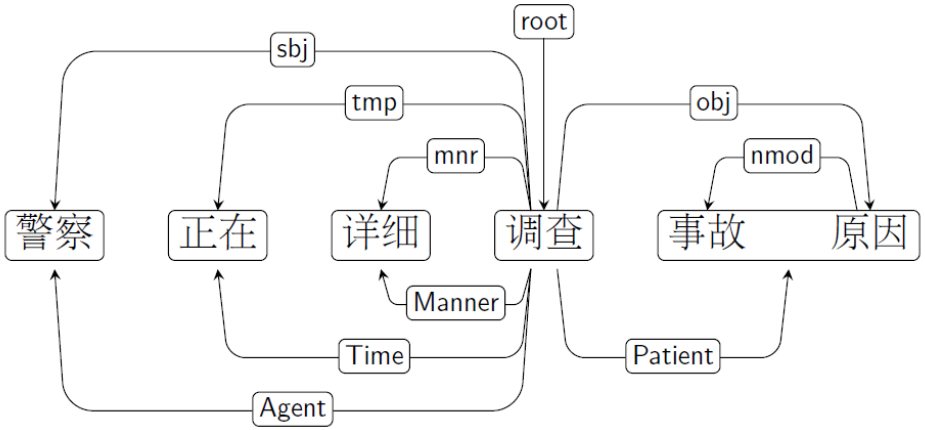
依存结构树



事件结构

Event: 调查
Agent: 警察
Time: 正在
Manner: 详细
Patient: 事故原因

句法语义依存关系图



基本方法

- **理性主义方法**
  - 研究人的语言知识结构，人工编汇语言知识＋ 推理系统
  - 符号处理系统
- **经验主义方法**
  - 看作**结构化预测**任务：直接研究实际的语言数据，从大量的语言数据中获得语言的知识结构
  - **基于语言数据的计算方法**
    - **例如序列标注模型(HMM、CRF等)**
- 理性主义方法与经验主义方法的融合
  - 融合方法

规则与统计共舞，语言随计算齐飞。

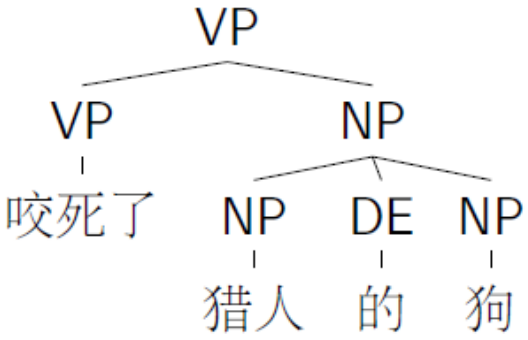
Computational Linguistics - Rules dance with numbers, Language soars with information.

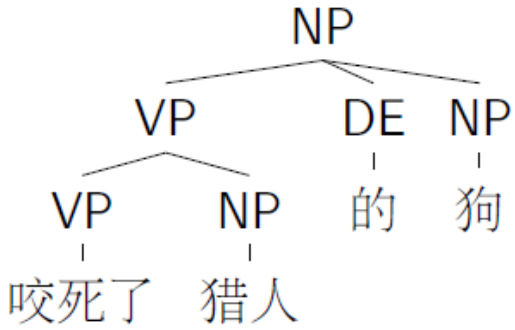
自然语言处理为什么如此之难？

- 自然语言与生俱有的歧义问题

“咬死了猎人的狗”

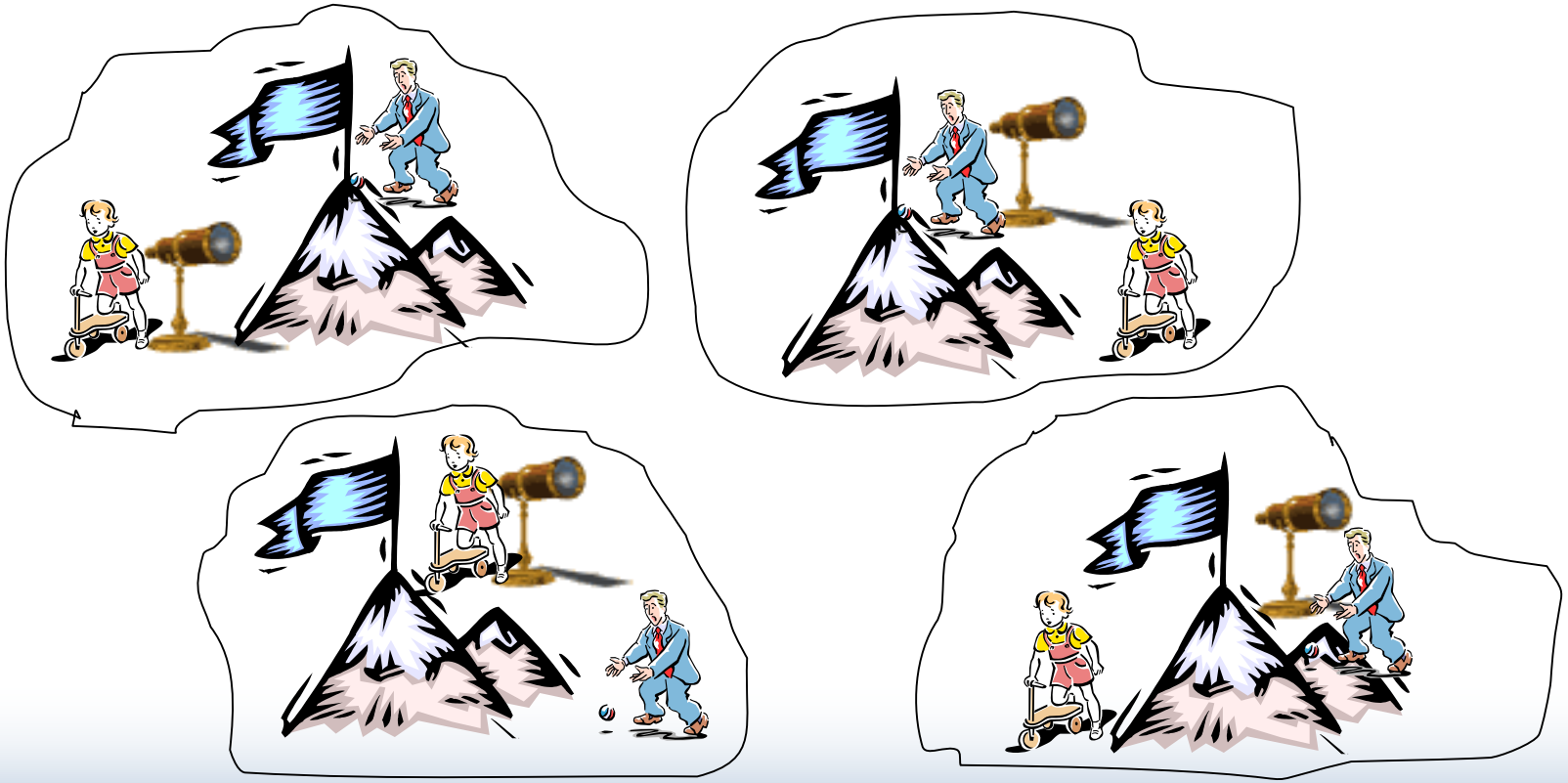
Ambiguity





The boy saw the man on the mountain with a telescope

PP attachment



Ambiguity

研究现状

- 哪个问题都没有彻底解决！
  - 有没有其他方法体系？有没有理论上限？
- 部分技术已取得较好的效果，能够服务于信息检索、文本挖掘等应用系统
  - 自动分词、词性标注、命名实体识别等
- 部分技术尽管效果不尽如人意，但能为人们提供辅助性帮助
  - 机器翻译
- 若干关键技术在研究和应用上均有待进一步突破
  - 句法分析、语义分析

社会对NLP技术的需求远大于NLP技术的当前水平

语义计算概述

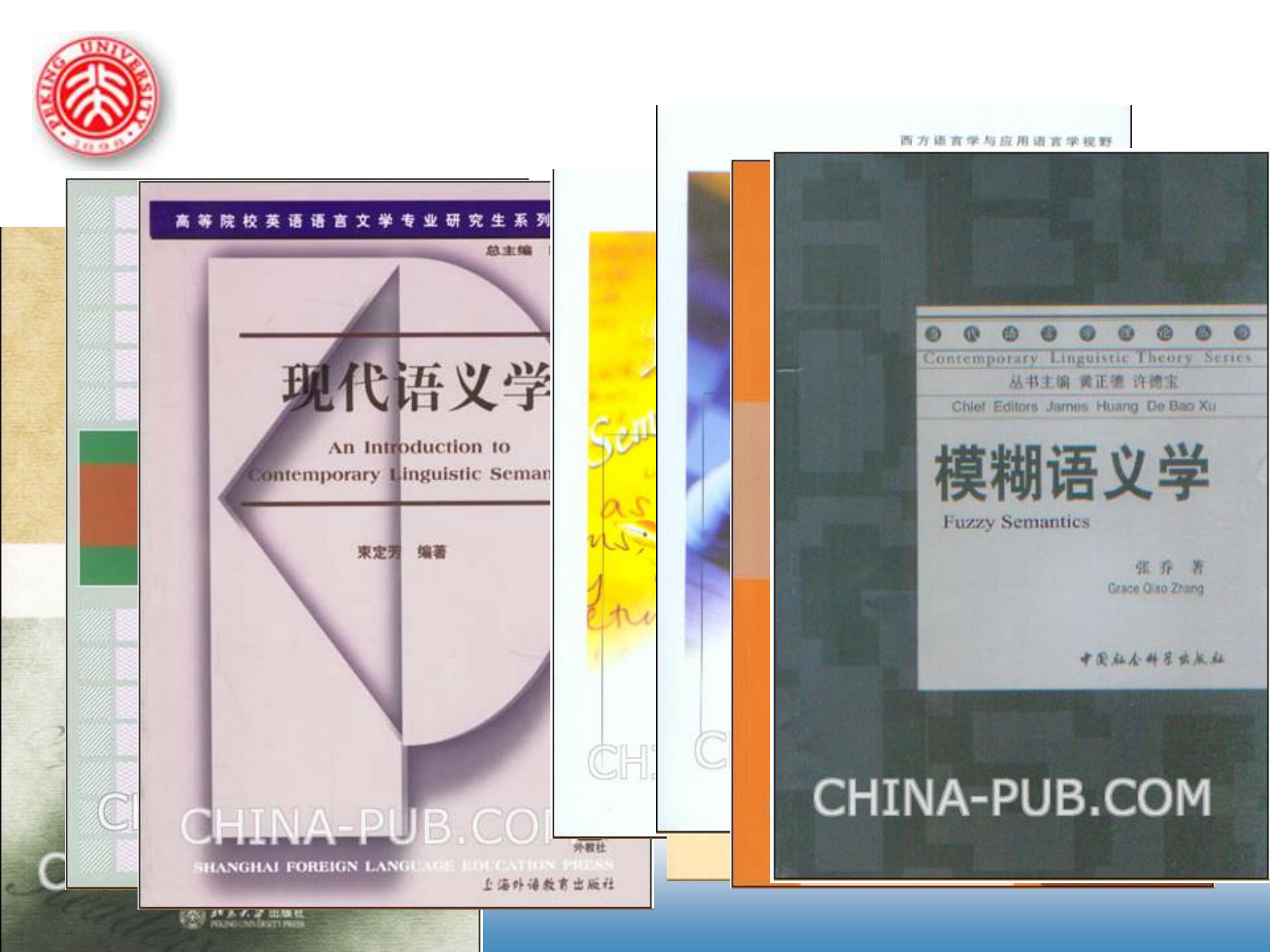
什么是语义

- 什么是语义？
  - In linguistics, meaning is what is expressed by the writer or speaker, and what is conveyed to the reader or listener, provided that they talk about the same thing (law of identity).

- 语义学的研究对象是自然语言的意义，这里的自然语言可以是词汇，句子，篇章等等不同级别的语言单位。
  - 语言学的语义学研究目的在于找出语义表达的规律性、内在解释、不同语言在语义表达方面的个性以及共性
  - 逻辑学的语义学是对一个逻辑系统的解释，着眼点在于真值条件，不直接涉及自然语言
  - 认知科学对语义学的研究在于人脑对语言单位的意义的存储及理解的模式
  - 与计算机科学相关的语义学研究在于机器对自然语言的理解

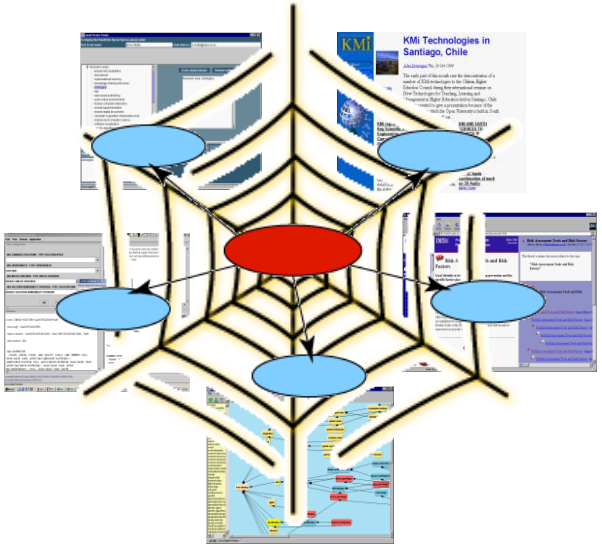
什么是语义

- 多层次
  - 词汇语义、句子语义、篇章语义
- 多视角
  - 概念语义、指称语义、情感语义、情景语义…
- 哪种语义最有用？
  - 跟应用有关

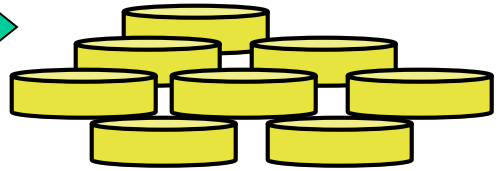

高等院校英语语言文学专业研究生系列
现代语义学
An Introduction to
Contemporary Linguistic Seman
束定芳 编著
SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
西方语言学与应用语言学视野
Contemporary Linguistic Theory Series
丛书主编 黄正德 许德宝
Chief Editors James Huang De Bao Xu
模糊语义学
Fuzzy Semantics
张乔 著
Grace Qiao Zhang
CHINA-PUB.COM

什么是语义计算

Data



Semantic Knowledge



什么是语义计算

- 语义计算
  - 计算语言单元的意义，构建语义表示
  - 处理自然语言，产生关于世界的常识性知识
- 词汇语义学(Lexical semantics)
  - – meanings of component words
  - – word sense disambiguation (e.g. ” country” in political or musical sense)
- 组合语义学(Compositional semantics)
  - – how words combine to form larger meanings
- 语义计算=语义分析≈语言理解

什么是语义计算

- 本课程语义范畴的几点说明
  - 面向自然语言处理、互联网搜索与挖掘应用
  - 对互联网文本为主的数据进行语义分析
  - 在词汇、句子、篇章等多层次进行分析
  - 语义涉及到概念语义、情感语义、指称语义等
  - 不纠结于语言学领域的多种理论与派别

什么是语义计算

- 语义计算 vs.NLP其他领域
  - 语音(Phonetics): 研究语言的发音
  - 形态(Morphology): 研究词的构成
    - – scored = score –d = Verb *score + past tense*
    - – employ, employee, employment, …
  - 语法(Syntax): 研究词之间的结构关系
  - 语义(Semantics): 研究语言的意义
  - 语用(Pragmatics): 研究语言如何被使用
    - E.g. “美女, 你长得好像我一个同学”
- 语义计算通常需要NLP其他领域技术，包括语法分析等

语义计算为何很难

- 面向自然语言的语义分析很难
  - 语言的歧义
    - – “丹丹和丽丽的妈妈看你来了”
  - 社会交流中共有知识通常会省略
    - – ”小李在小张的搀扶下回家了。他喝醉了。”
  - 语言动态变化: 新词（e.g. 剩男、剩女、快男、超女、雷人、DUANG、诺贝尔哥），新的用法（e.g. 打酱油、杯具、粉丝）
  - 不同层次歧义的传递与影响

分词结果直接影响语义分析结果

LIAO CHENG EVENING NEWS 聊城晚报

2013年6月28日 星期五 责编/版式 刘国栋 | 校对 赵立 | 05

中国 NEWS

■李双江之子涉嫌强奸案进展

李天一他妈的要求高

警方证实案件进入审查起诉阶段 梦鸽逼律师发公告称其子未参与强奸案

李天一的代理律师薛振原正式辞去为其代理涉嫌强奸案的律师工作，但他自称"有些原因不方便说，也不好说"，拒绝正面回应辞职的真实原因。记者27日从北京警方证实，此案已经侦查完结，进入检察机关的审查起诉阶段。

今年2月17日晚，歌唱家李双江之子李天一与另外4个男孩一起，在北京五道口一家酒吧内，将一名喝醉酒的女孩带到湖北大厦一房间，轮流与该女孩发生性关系。

该案被曝光轰动一时，引发公众高度关注。

今年3月19日，李天一的母亲梦鸽委托薛振原做其儿子的辩护律师。薛振原曾在网上发表声明称，已会见李天一，且初步了解案情，因李天一已被证实为未成年人，且在侦查中获得了相应的司法保护。

据了解，薛振原与梦鸽正式解除委托关系的时间为5月30日。至于为何薛振原辞去李天一律师一职，有

■相关新闻

一个"特权子弟"断裂的成长道路

李天一小小年纪曾有很多光环，他是海淀区书法协会最小的成员，并用毛笔为母亲2009年红歌演唱会写了标题；也是年纪最小的"申奥大使"。

但他在社会上真正意义的成名来自于一年多以前。他和一个同伴在差点和一对小区夫妇发生追尾后，由口角升级至对那对夫妇大打出手，导致对方头部受伤，被缝了11针。他为此付出的代价是劳教一年。

从时间上计算，他这一次性质更严重的案件距离他的劳教结束期不过几个月。

成长 | 过早地进入了暴力的速度世界

李天一的父亲是解放军艺术学院的音乐系主任，曾经演唱过很多耳熟能详的革命歌曲。在他们居住的军艺大院中，有不少老艺术家，但李天一的父亲仍然享有特殊的声望。

据一位邻居回忆，他所见的李家的车包括两辆本田CRV，一黑一白，一辆宝马——2011年李天一和万竹居住在西山华

道，因打人事件被劳教一年后，李天一换了一辆白色的GTR。但在这位邻居的记忆里，他还看到他开着另一辆新车。"当时我开着车在小区门口准备进门，小区有门禁，刷一次卡进一辆车，我刚刷了卡，他开着奔驰车从我车旁抢着先进去了。我记得是一辆银色的奔驰，也有四个尖排气管，发动机发出轰隆隆的声音。"

他们很自豪于孩子广泛的兴趣，并尽力维护着养他的天性。这或许是这些孩子的幸运之处，他们能获得并几乎不受限制地使用他感兴趣的一切消费品，但同时他们也是消化不良的一批人。他们的年纪和所受到的教育，让他们不知道如何合理使用他们的爱好，行为边界又在哪里。在他们尚未具备控制和理解这些新

■数字解读

父亲今年74岁

李双江，1939年生于哈尔滨，前妻杨晓飞子李天一生于1996年4月，当时李双江已有57岁。

母亲梦鸽比李双江小27岁

李双江1988年认识现任妻子梦鸽，当时梦鸽22岁，李双江49岁。1990年10月20日，相差27岁的两人在北京举行了婚礼。

4岁学习钢琴

师从中央音乐学院著名钢琴教授韩剑明。

5岁

进入申奥形象大使。

6岁

李淑母亲梦鸽新歌MTV。

他说：“她这个人真有意思(**funny**)”。她说：“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(**wish**)，并让他向她意思意思(**express**)。他火了：“我根本没有那个意思(**thought**)”！她也生气了：“你们这么说是什么意思(**intention**)”？事后有人说：“真有意思(funny)”。也有人说：“真没意思(**nonsense**)”。

—《生活报》1994. 11. 13. 第六版

阿凡提当理发匠时，大阿訇(hōng)总是来找他剃头，却从来不给钱。阿凡提很生气，想狠狠整他一下。有一天，阿訇又来理发了。**阿凡提先给他剃光了头，在给他刮脸的时候，问道：**

“阿訇，您要眉毛吗？”

“当然要，这还用问！”阿訇说。

“好，您要我就给您！”阿凡提说着，嗖嗖几刀，就把阿訇的两道眉毛刮下来，递到阿訇手里。

大阿訇气得说不出话——谁叫他自己说过要呢。

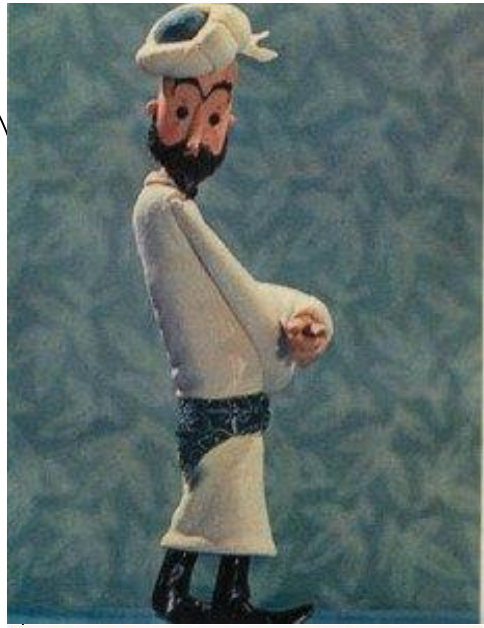
“阿訇，胡子要吗？”阿凡提又问。

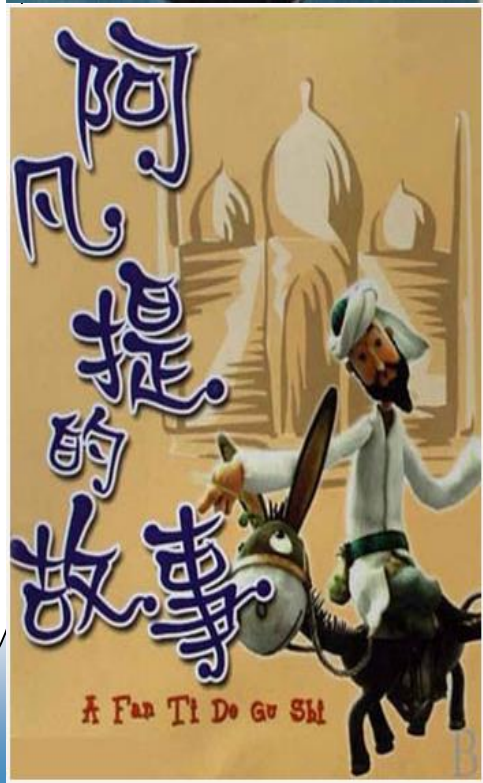
“不要，不要！”阿訇连忙说。

“好，您不要就不要。”阿凡提说着，又嗖嗖几刀，就把大阿訇的胡子刮下来，甩在地上。

大阿訇对镜子一看，自己的脑袋和脸都刮得净光，简直就像个光溜溜的鸡蛋。这一下他可气坏了，就大骂起来。“得啦，得啦，这不都是遵照您的吩咐做的吗？”阿凡提说：“要是依我的话，不要说眉毛胡子，连您的头发，我也不愿意剃哩！”

——《阿凡提的故事》





阿凡提在逻辑上偷换概念，利用词语意义的不确定性，灵活使用。

转发

Warriors：最近很多新生来问我，664能不能来北大？我在这里再次重申一下，664你可以考虑一下到清华，但是对于北大，我只能说可以冲一下，但希望不大！毕竟至少696才能到北大，一般都是

坐四号线到北京大学东门站。坐696🚌 实在是有点堵，664🚌 只到清华西北门。其实按照这么说331也可以到清华的。

iPhone 6

@刘群MT-to-Death V

#自然语言理解太难了# 女:“明明喜欢我,却不告诉我!” 男:“别说了，我想静静。” 女:“静静是谁?” 男:“你TMD先告诉我明明是谁?”

1月23日 08:16 来自 微博 weibo.com　　转发 136 | 评论 11 | 16

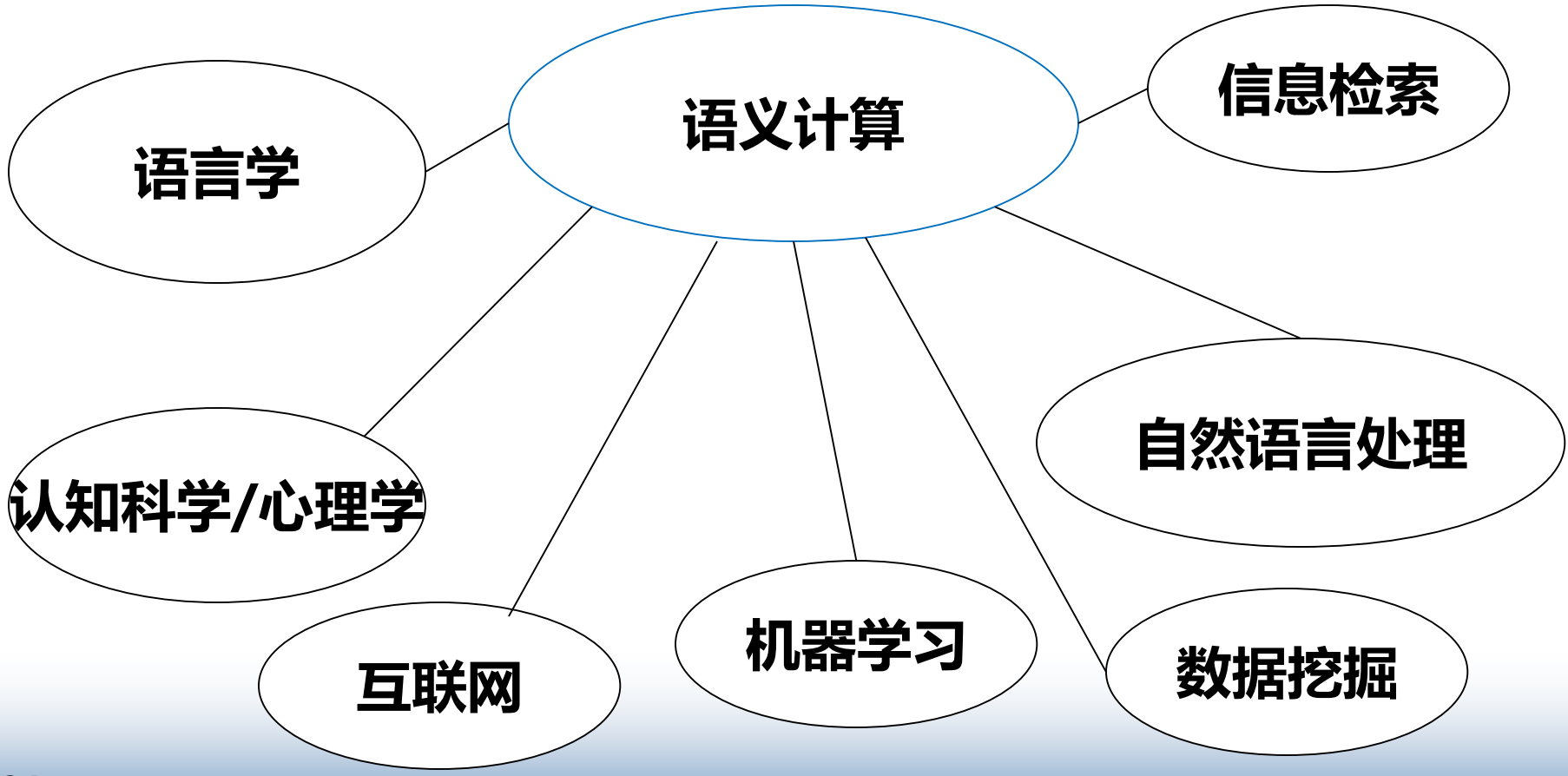
语义计算的重要性

- 几个未解决的难题之一 Gray (1944, 2007, 2012)
- List on the right is fr
  " What next? A few
  IT "
  - # 9. Build a system
    an answe
    summariz
    as a hum

The List (Red is Turing Complete)

1. Devise an architecture that scales up by 10^6.
2. The Turing test: win the impersonation game 30% of the time.
   a. 3.Read and understand as well as a human.
   b. 4.Think and write as well as a human.
3. Hear as well as a person (native speaker): speech to text.
4. Speak as well as a person (native speaker): text to speech.
5. See as well as a person (recognize).
6. Illustrate as well as a person (done!) but virtual reality is still a major challenge.
7. A copy-protection and payment scheme that protects IP owner and user.
8. Remember what is seen and heard and quickly return it on request.
9. Build a system that, given a text corpus, can answer questions about the text and summarize it as quickly and precisely as a human expert.
10. Do 9 for Sounds: conversations, music.
11. Do 9 for Images: pictures, art, movies.
12. Simulate being some other place as an observer (Tele-Past) and a participant (Tele-Present).
13. Automatic Programming: Given a specification, build a system that implements the spec. Prove that the implementation matches the spec. Do it better than a team of programmers.
14. Build a system used by millions of people each day but administered by a ½ time person.
15. Do 15 and prove it only services authorized users.
16. Do 15 and prove it is almost always available: (out less than 1 second per 100 years).

相关学科与技术



相关会议与活动

- 相关学术会议
  - ACL、EMNLP、NAACL、CoNLL
  - SIGIR、WWW、CIKM、WSDM
  - KDD、ICDM
  - AAAI、IJCAI
- 相关评测
  - SemEval, MUC, ACE, TAC(KBP, RTE),TREC, NTCIR、NLPCC、CLEF…

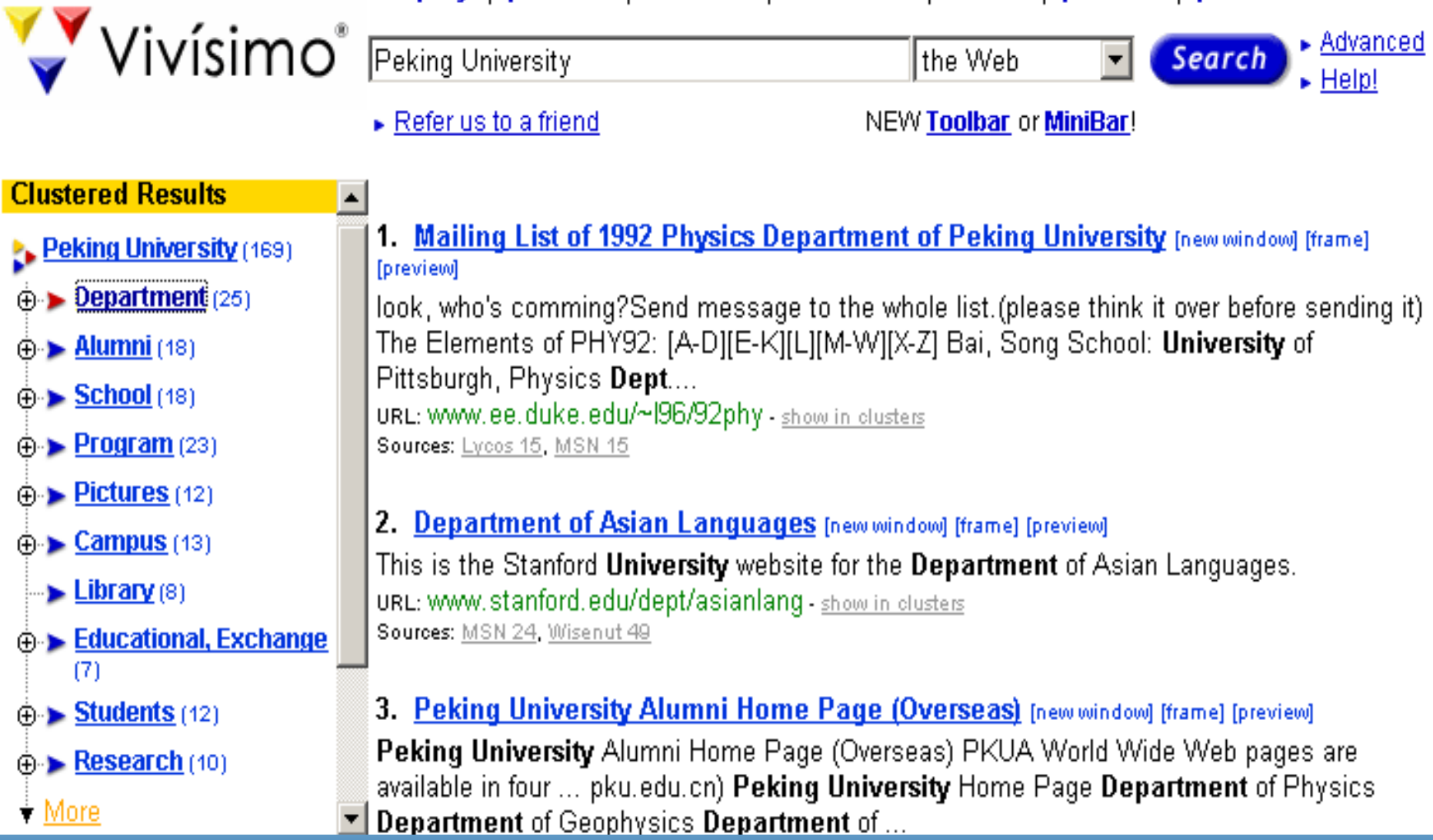
应用举例

- 信息检索
- 信息推荐
- 智能问答
- 机器翻译
- 舆情分析
- …

应用举例-信息检索


Baidu 百度
新闻 网页 贴吧 知道 MP3 图片 视频 地图 更多▾
到哪里打酱油
13号女嘉宾，到哪里去打酱油了 非诚勿扰吧 贴吧
13号女嘉宾，到哪里去打酱油了 贴吧公益 三国曹阿瞒 26位粉丝 1楼 13号徐裴艳出席112——129期（130期出席女嘉宾没找到）非诚节目，并未被男嘉宾领走，在节目...
tieba.baidu.com/f?kz=1096526577 2011-8-26 - 百度快照
"打酱油"一说到底出自哪里？-草根消息-杭州19楼
"打酱油"一说到底出自哪里？ 历史帖 阅读[1891] 回复[21] 只看楼主 过滤水... 近段时间网络流行"打酱油"，但是这个打酱油到底是什么意思，到现在还是不明白...
www.19lou.com/forum-269-thread-11664055-1 ... 2011-8-19 - 百度快照
打酱油 - 视频 - 优酷视频 - 在线观看
70魔道酱油血蝶之舞K（无限加热炉） 04:11 70魔道酱油血蝶之舞K（无限加热... 凹凸曼打酱油 02:54 凹凸曼打酱油 播放:1,575 3人打酱油 05:47 3人打...
v.youku.com/v_show/id_XMjgyMjM1Nzlw.html 2011-7-9 - 百度快照
成都装饰城有哪些啊，位置在哪里啊？求各位大虾出来打... 百度知道
四川爱尚装饰公司。拥有2000多平米材料展场，50余个品牌。60多个样板间。展场更有两套1:1不同风格的样板房（一套欧式 一套现代），本公司承诺最后工程决算时在...
zhidao.baidu.com/question/287599316.html 2011-7-5 - 百度快照
青天大老爷/第六十二章 打酱油/零点看书网
第六十二章 打酱油请牢记本站地址:www.00ks.com,本站qq群:高级4群140102442... "丹姐，菜放哪里？"夏言问。 赵雅丹一边把菜放在了灶台上，一边回头冷着...
www.00ks.com/Html/Book/13/13899/3322197.html 2011-8-8 - 百度快照

应用举例-信息检索


Vivísimo
company | products | solutions | customers | demos | partners | press
Peking University
the Web
Search
Advanced
Help!
Refer us to a friend
NEW Toolbar or MiniBar!
Clustered Results
Peking University (169)
Department (25)
Alumni (18)
School (18)
Program (23)
Pictures (12)
Campus (13)
Library (8)
Educational, Exchange (7)
Students (12)
Research (10)
More
1. Mailing List of 1992 Physics Department of Peking University [new window] [frame] [preview]
look, who's comming?Send message to the whole list.(please think it over before sending it) The Elements of PHY92: [A-D][E-K][L][M-W][X-Z] Bai, Song School: University of Pittsburgh, Physics Dept....
URL: www.ee.duke.edu/~l96/92phy - show in clusters
Sources: Lycos 15, MSN 15
2. Department of Asian Languages [new window] [frame] [preview]
This is the Stanford University website for the Department of Asian Languages.
URL: www.stanford.edu/dept/asianlang - show in clusters
Sources: MSN 24, Wisenut 49
3. Peking University Alumni Home Page (Overseas) [new window] [frame] [preview]
Peking University Alumni Home Page (Overseas) PKUA World Wide Web pages are available in four ... pku.edu.cn) Peking University Home Page Department of Physics Department of Geophysics Department of ...

谷歌知识图谱

郭德纲

全部 图片 视频 新闻 地图 更多 搜索工具

找到约 2,230,000 条结果（用时 0.33 秒）

相关搜索： 郭德纲2015 郭德纲 单口 郭德纲 于谦

郭德纲- 维基百科，自由的百科全书
https://zh.wikipedia.org/zh-cn/郭德纲
郭德纲（1973年1月18日－），中国天津人，相声演员，亦曾担任影视剧演员及电视脱口秀节目主持人。2005年冬天起受到媒体的关注，知名度逐渐上升。郭德纲自称"非 ...

郭德纲_百度百科
baike.baidu.com/view/5444.htm
郭德纲，男，出生于1973年1月18日，天津人，相声演员，电视、电影演员及电视节目主持人。1979年投身艺坛，先拜评书前辈高庆海学习评书，后跟随相声名家常宝丰 ...

郭德纲于谦相声全集 - 6平米
guodegang.6pingm.com/
6平米郭德纲专区,郭德纲最新相声,网络最全郭德纲相声全集.
郭德纲单口相声全集 - 郭德纲 - 郭德纲于谦相声全集 - 郭德纲2015

新闻报道

郭德纲儿子半夜咬舌岳云鹏搞笑:不抢一哥
新浪网 - 12 小时前
郭德纲之子、相声演员郭麒麟在微博晒出就诊图，原来他因为吃夜宵咬到舌头动脉，紧急就医， ...

郭德纲长子深夜入院吃夜宵咬到舌头上动脉(图)
凤凰网 - 15 小时前

郭德纲儿子咬舌就诊称"吃夜宵咬到舌头动脉"(图)
中国新闻网 - 7 小时前

郭德纲

演员

郭德纲，中国天津人，相声演员，亦曾担任影视剧演员及电视脱口秀节目主持人。2005年冬天起受到媒体的关注，知名度逐渐上升。郭德纲自称"非著名相声演员"，其相声爱好者自称"纲丝"。 维基百科

生于： 1973 年 1 月 18 日（43 岁），天津市

配偶： 王惠

著作： 非著名相声演员， 话说北京， 郭德纲单口相声精品集， 过得刚好， 郭德刚话说北京

子女： 郭麒麟

电影 还有3+项


大话天仙
2014 年


车在囧途
2012 年


越光宝盒
2010 年


三笑之才子佳人
2010 年


落叶归根
2007 年

必应人物关系

必应 郭德纲

Switch to Bing in English 登录

网页 图片 视频 学术 词典 网典 地图 更多

郭德纲- 必应网典 主持人 演员

郭德纲（1973年1月18日 - ），著名相声演员相声演员，生于天津，北京德云社相声大会的创办者，自幼酷爱民间艺术。8岁投身艺坛，曾受到多位相声名家的指点、传授。其间又潜心学习了京剧、评剧、河北梆子等姊妹剧种，辗转于梨园，兼工文丑与铜锤行当的经历，对丰富自己的相声表演起了十分重要的作用。通过对... (更多)

百度百科 维基百科 互动百科 搜狗百科

www.bing.com/knows/

为您推荐：郭德纲2012最新相声 · 郭德纲于谦相声全集 · 郭德纲相声

郭德纲的新浪微博

走进梦工厂，每个人都有可能是主角。在生活中演绎故事，在故事中诠释人生。今天站在舞台上的你，或许就是明天我们故事的主角。今晚21:20山东卫视#花漾梦工厂# 谁会是下一个主角？

查看全文

2016-02-21 11:50:00 · 转发(127) · 评论(0)

相关人物

于谦 拍档

李菁 合作

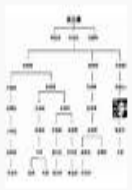
李鹤彪 师徒关系

侯耀文 师徒关系

张文顺 师徒关系

中国著名相声演员

侯宝林



张三禄

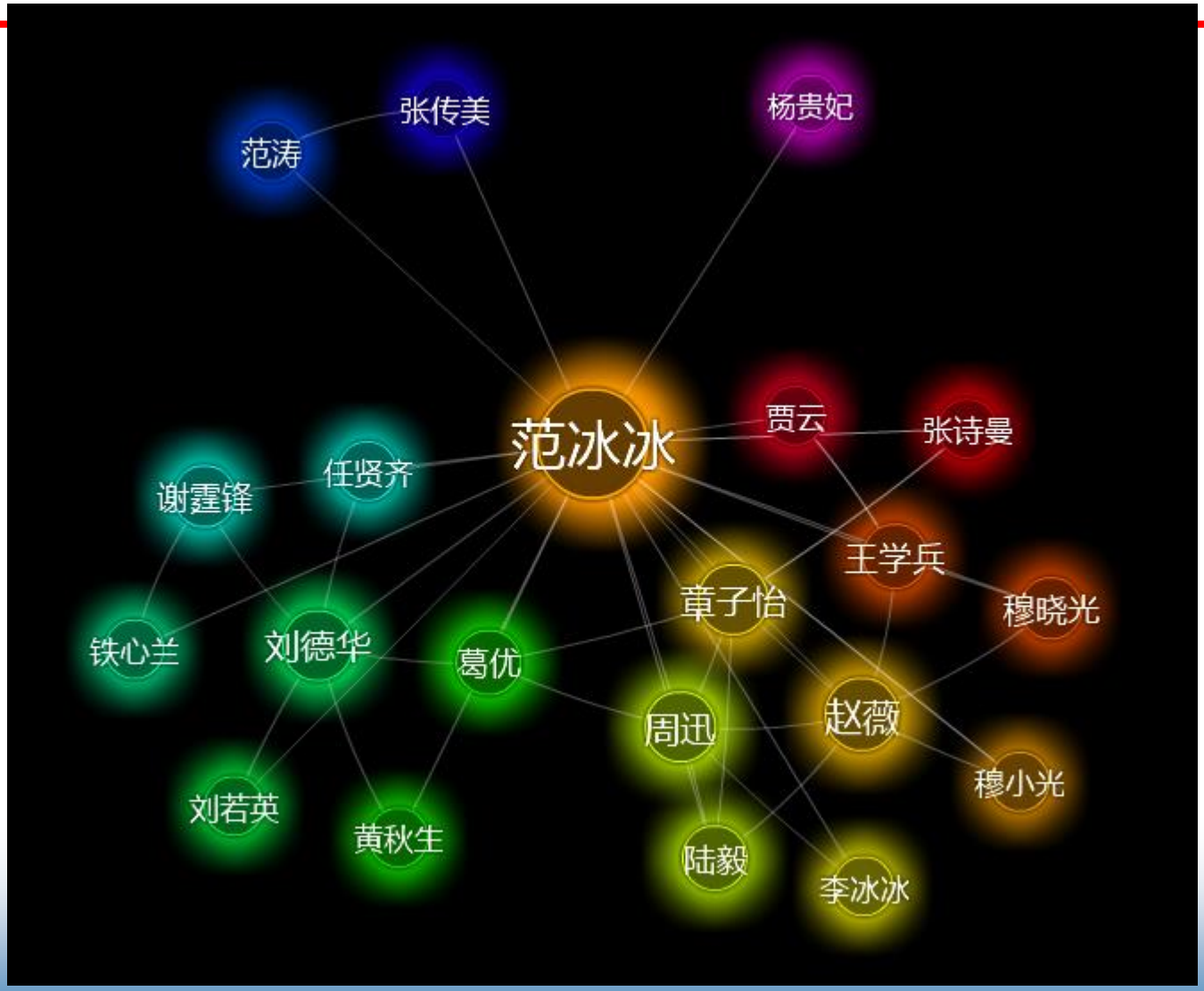
徐德亮

冯巩

石富宽

其他人还搜

应用举例-信息检索



微软人立方（服务已停）

应用举例-信息检索

Google

新闻

中国版 (China)

焦点报道
国际/港台
内地
财经
娱乐
科技
互联网
体育
社会
汽车
房产
教育
热门报道

所有内容
新闻标题
新闻图片

焦点报道

北方新闻网

康菲『玩弄』海洋局
凤凰网 - 22分钟前
国家海洋局差不多被康菲公司玩残了。为期3个月的漏油事故，在国家七部委最后确定的"大限"之后仍在继续。在央视采访中，康菲甚至直言，溢油源已永久封堵的说法"就是骗你的"。责。此后，油污并未绝迹，康菲公司则 ...
康菲公司未完成"两彻底"要求 回应：因天气原因 新浪网
康菲声明称员工未发表负面言论 要求央视更正 腾讯网
网易 - 搜狐 - 北方新闻网
此专题所有 5,291 篇报道 »

民政部将在全国范围内入户调查获居民诉求 凤凰网 - 6小时前
中国宋庆龄基金会称与地方基金会无隶属关系 新浪网 - 1小时前
专题报道(771篇) »
卡扎菲被曝或逃往尼日尔(图) 搜狐 - 25分钟前 专题报道(1,646篇) »
美国高官称9-11纪念暂无袭击威胁 新民网 - 2分钟前
专题报道(346篇) »
温家宝：力推找矿突破战略 积极参与国际合作 搜狐 - 7分钟前
专题报道(296篇) »
野田内阁亟需解决经济外交难题 新华网 - 1小时前 专题报道(136篇) »
市三医院手术室火灾原因查明 分管消防副院长被免职 新浪网
- 2分钟前 专题报道(373篇) »
2011中国企业500强发布 国企316家超六成 和讯网 - 24分钟前
专题报道(989篇) »

应用举例-信息推荐

Recommended based on My Citations View: Top All

Generic Multi-Document Summarization Using Topic-Oriented Information
Y Pei, W Yin, L Huang - PRICAI 2012: Trends in Artificial Intelligence, 2012 - Springer
8 days ago - The graph-based ranking models have been widely used for multi-document summarization recently. By utilizing the correlations between sentences, the salient sentences can be extracted according to the ranking scores. However, sentences are ...
Cites An exploration of document impact on graph-based multi-document ...

A Novel Method of Significant Words Identification in Text Summarization
M Kiabod, MN Dehkordi... - Journal of Emerging ..., 2012 - academypublisher.com
10 days ago - Abstract Text summarization is a process that reduces the size of the text document and extracts significant sentences from a text document. We present a novel technique for text summarization. The originality of technique lies on exploiting local and ...

[PDF] Cross-lingual Training of Summarization Systems Using Annotated Corpora in a Foreign Language
M Litvak, M Last - cs.bgu.ac.il
13 days ago - Abstract The increasing trend of cross-border globalization and acculturation requires text summarization techniques to work equally well for multiple languages. However, only some of the automated summarization methods can be defined as " ...
Cites Cross-language document summarization based on machine translation ...

[PDF] Identifying "Comment-on" Citation Data in Online Biomedical Articles Using SVM-based Text Summarization Technique
IC Kim, DX Le, GR Thoma - elrond.informatik.tu-freiberg.de
13 days ago - Abstract-Comment-on (CON), a MEDLINE® citation field, indicates previously published articles commented on by authors expressing possibly complimentary or contradictory opinions. This paper presents an automated method using a support vector ...

[PDF] Latent Semantic Analysis of multi-documents using Natural Language Processing techniques
B Prabhala - 2012 - people.rit.edu
14 days ago - Abstract In this age of Internet, Natural Language Processing (NLP) techniques are the key sources to provide information required by users. However, with extensive usage of available data a secondary level of wrappers which interact with NLP ...

[PDF] Redundancy reduction for multi-document summaries using A* search and discriminative training
A Aker, T Cohn, R Gaizauskas - ceur-ws.org
20 days ago - Abstract. In this paper we address the problem of optimizing global multidocument summary quality using A* search and discriminative training. Different search strategies have been investigated to find the globally best summary. In them the search is ...

Multilingual Statistical News Summarization
M Kabadjov, J Steinberger, R Steinberger - Multi-source, Multilingual Information ... - Springer

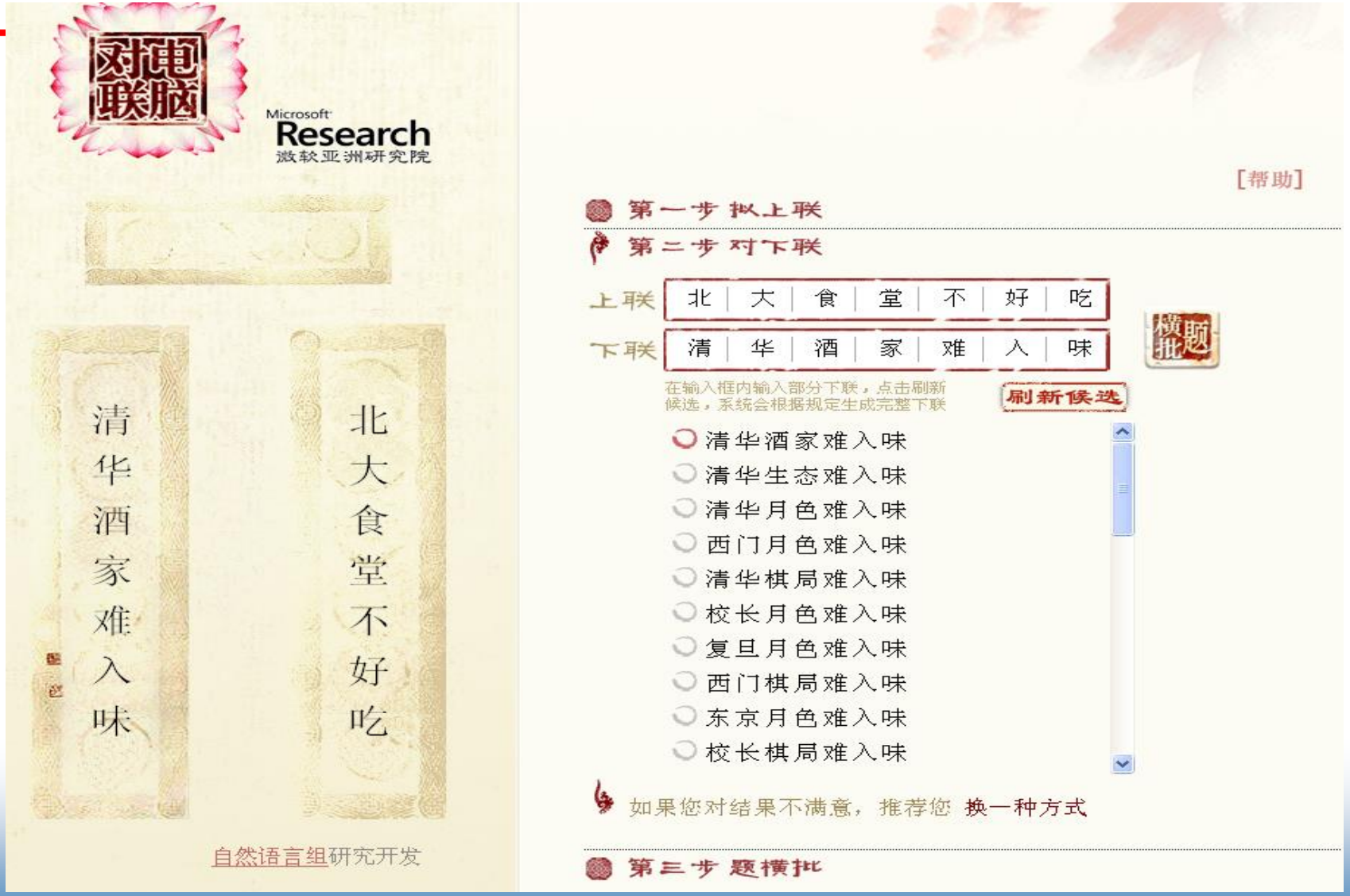
应用举例-机器翻译

- 机器翻译不好导致的笑
  - How old are you? =>

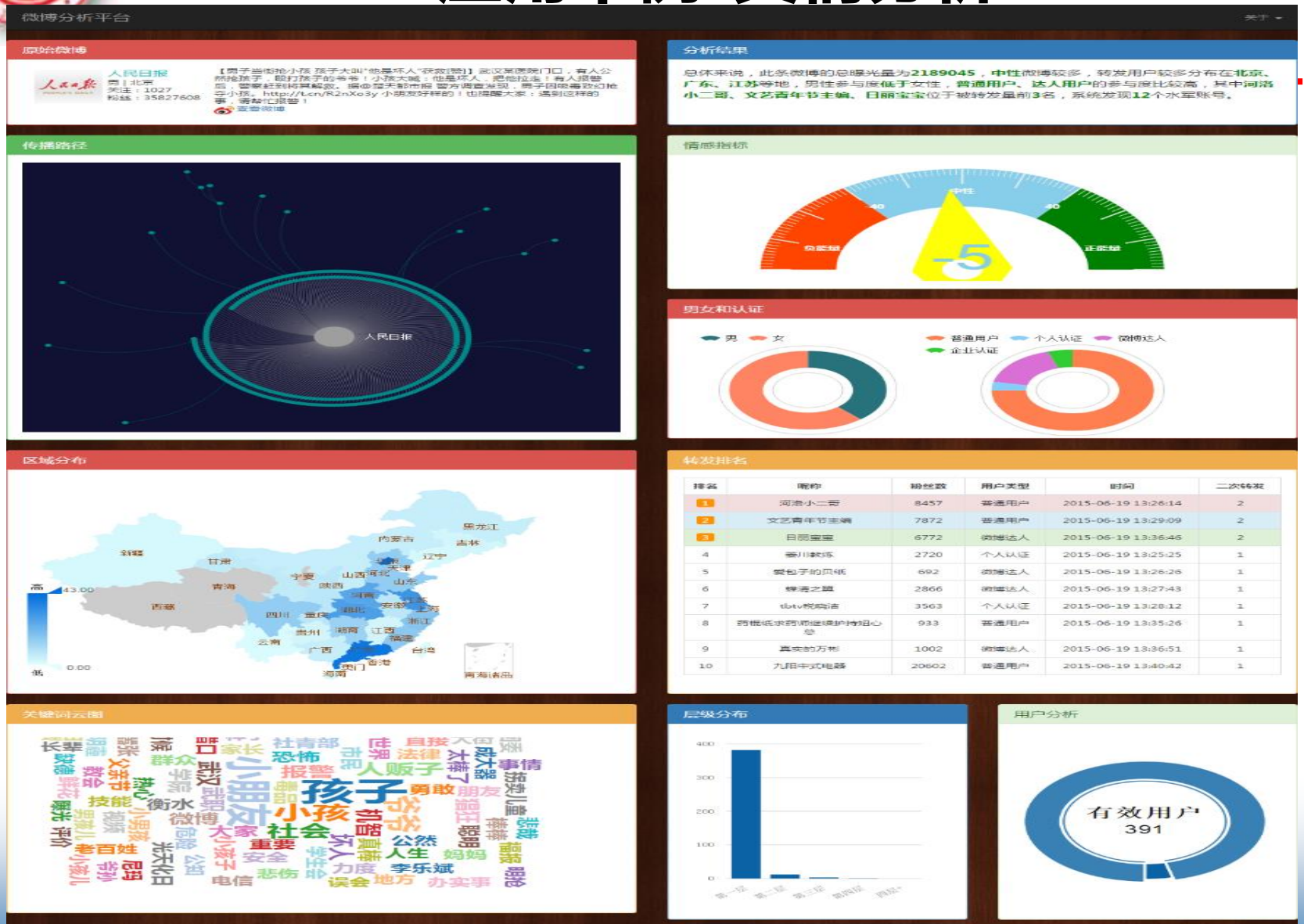
应用举例-机器翻译


Google
Translate
English
Spanish
French
Chinese - detected
English
Spanish
Arabic
Translate
故事里的事，说不是就不是是也不是。
Story, saying that it is not is not no.
Suggest an edit

应用举例-自动对联


对电联脑
Microsoft Research
微软亚洲研究院
北大食堂不好吃
清华酒家难入味
自然语言组研究开发
[帮助]
第一步 拟上联
第二步 对下联
上联 北 | 大 | 食 | 堂 | 不 | 好 | 吃
下联 清 | 华 | 酒 | 家 | 难 | 入 | 味
横批题
在输入框内输入部分下联，点击刷新候选，系统会根据规定生成完整下联
刷新候选
清华酒家难入味
清华生态难入味
清华月色难入味
西门月色难入味
清华棋局难入味
校长月色难入味
复旦月色难入味
西门棋局难入味
东京月色难入味
校长棋局难入味
如果您对结果不满意，推荐您 换一种方式
第三步 题横批

应用举例-舆情分析


微博分析平台
关于
原始微博
人民日报
男 | 北京
关注：1027
粉丝：35827608
【男子当街抢小孩 孩子大叫"他是坏人"获救[赞]】武汉某医院门口，有人公然抢孩子，殴打孩子的爷爷！小孩大喊：他是坏人，把他拉走！有人报警后，警察赶到将其解救。据@警天部市局 警方调查发现，男子因电毒致幻抢走小孩。http://t.cn/R2nXo3y 小朋友好样的！也提醒大家：遇到这样的事，请帮忙报警！
查看微博
分析结果
总体来说，此条微博的总曝光量为2189045，中性微博较多，转发用户较多分布在北京、广东、江苏等地，男性参与度低于女性，普通用户、达人用户的参与度比较高，其中河游小二哥、文艺青年节主编、日丽宝宝位于被转发量前3名，系统发现12个水军账号。
传播路径
人民日报
情感指标
负能量
中性
正能量
-5
男女和认证
男
女
普通用户
个人认证
微博达人
企业认证
区域分布
高
43.00
低
0.00
转发排名
排名
昵称
粉丝数
用户类型
时间
二次转发
1 河游小二哥 8457 普通用户 2015-06-19 13:26:14 2
2 文艺青年节主编 7872 普通用户 2015-06-19 13:29:09 2
3 日丽宝宝 6772 微博达人 2015-06-19 13:36:45 2
4 香川教练 2720 个人认证 2015-06-19 13:25:25 1
5 爱包子的贝纸 692 微博达人 2015-06-19 13:26:26 1
6 蝶语之翼 2866 微博达人 2015-06-19 13:27:43 1
7 tb大靴靴 3563 个人认证 2015-06-19 13:28:12 1
8 用纸纸求药师继续护持姐心 933 普通用户 2015-06-19 13:35:26 1
9 真实的万彬 1002 微博达人 2015-06-19 13:36:51 1
10 九阳中式电器 20602 普通用户 2015-06-19 13:40:42 1
关键词云图
小孩
孩子
社会
层级分布
400
300
200
100
0
用户分析
有效用户
391

应用举例-智能问答

2011年IBM沃森参加《危险边缘》 (Jeopardy!)

应用举例-智能问答

2017年搜狗汪仔参加《一站到底》

应用举例-小冰对话

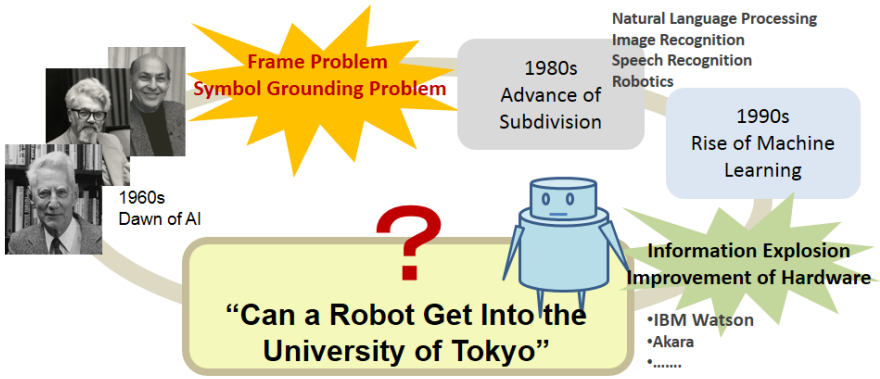

与 小冰 对话中
晚饭吃的什么
吃面怎末样!
以上为历史记录
炸酱面，刀削面还是担担面？
炸酱面 多加菜码 就蒜吃 太TM好吃了。
我不喜欢吃炸酱面
吃炸酱面？这到底是为什么呢[冷汗]
你说什么就是什么
那到底是什么座

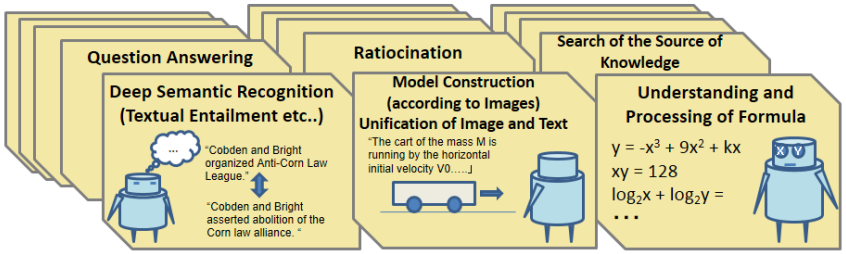
应用举例-小冰对话


与 小冰 对话中
北大好不好
我觉得紫棋很好看。
信息学院好不好
紫棋好漂亮！越来越棒了！
这门课讲得好不好
紫棋非常有魅力
为什么你总要提邓紫棋？
为什么要扯上邓紫棋 你和周杰伦多大仇~
以上为历史记录

应用举例-智能问答

高考机器人





由日本国立情报学研究所等组成的团队研发的“考生”取名为“ToRobo君”。

。

该团队从2011年展开研究，目标是2016年在大学全国考试获高分、2021年考上东大。

应用举例-智能问答

例

问题1下划线a中所陈述的村落生活如下文 a～d所示，请在1～4中选出正确的组合。

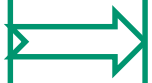
A 具有青铜制刀尖的农具在全国逐渐普及。

B 对于日常生活而言不便利的山顶和丘陵上也有村落。

C 铁器•青铜器传自海外，后可在日本列岛内制作

D 骑马的风习及硬质陶瓷传自朝鲜半岛。

1. a.c　　2. a.d　　3. b.c　　4. b.d



〈data id="D13" type="text”>
<IText id="Ll"xlabel>a</label>具有青铜制刀尖的农具在全国逐渐普及。
<br/x/IText>
<IText id="L2"xlabel>b</label> 对于日常生活而言不便利的山顶和丘陵上也有村落。<br/x/IText>
<IText id="L3"xlabel>c</label> 铁器•青铜器传自海外，后可在日本列岛内制作。
<br/ ></IText>
<IText id="L4"><label>d</label> 骑马的风习及硬质陶瓷传自朝鲜半岛。
<br/x/IText>
</data>

应用举例-智能问答

高考机器人

- **日本国立大学和公立大学的入学考试包括2次：全国考试 和 大学自主招生考试**
- TODAI Robot君在2015年参加了**全国考试**，参考东大的招生简章，从7科的37套试卷中选择了（得分率相对较高的）国语、数学IA、数学IIB、英语、日本史B、世界史B、物理卷。

	国语	数学		英语		物理	历史		
	国语	数学IA	数学IIB	英语（笔记）	英语（听力）	物理	日本史B	世界史B	5个教科总计
总分（满分）	200	100	100	200	50	100	100	100	950
全国平均分	105.4	45.5	42.8	86.0	24.6	49.4	46.6	45.9	416.4
东机器人得分	90.0	75.0	77.0	80.0	16.0	42.0	55.0	76.0	511.0

东大机器人5教科合计成绩高于全国平均成绩。其中，数学1A,数学2B，世界史B比全国平均分高大概30分，可以说东大机器人很擅长数学和社会科。而国语这一科比全国平均分低15分左右，这是一个还未解决的课题。

根据东大机器人 “模拟考试 综合学力选择题模拟考试·6月” 的成绩，有80%以上的可能性被33个国立大学，441私立大学录取。

应用举例-智能问答

2014 年北京高考地理卷

北京时间 2013 年 12 月 21 日 0 时 42 分，我国为玻利维亚成功发射通信卫星。读图 1，回答第 1、2 题。

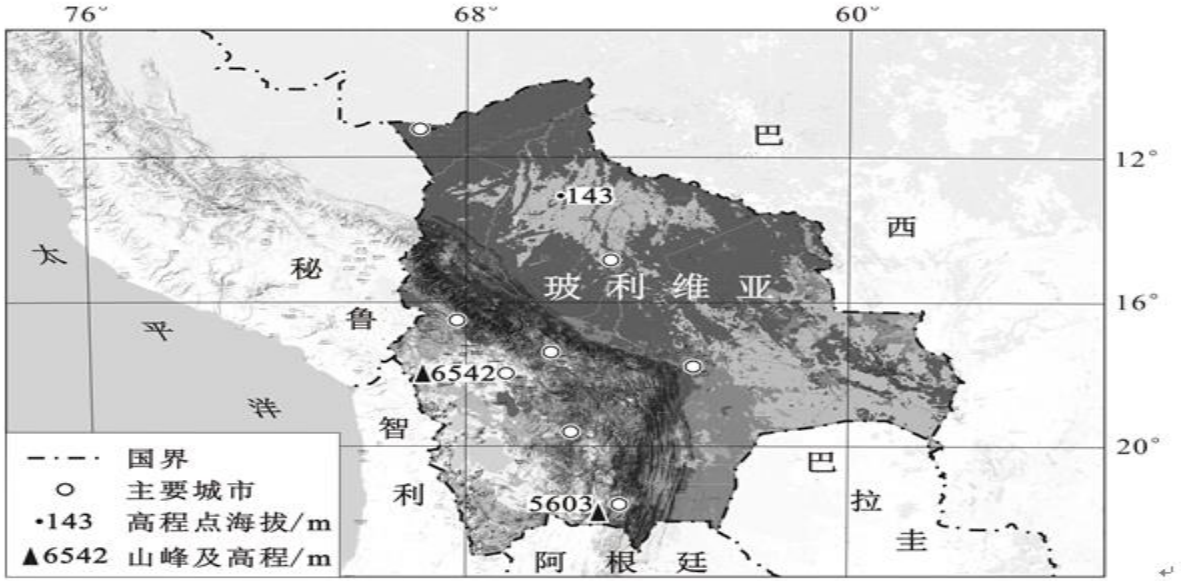


图 1

1．玻利维亚

A．东邻巴西，西临太平洋　　B．受寒流影响，多雾少雨

C．西南山地垂直带谱复杂　　D．城市多分布于平原地区

2．卫星发射当日

A．玻利维亚在傍晚可收看卫星发射直播　　B．赤道正午太阳高度角接近 90°

C．中国南极昆仑站所在地出现极昼现象　　D．地球接近公转轨道的远日点

应用举例-智能问答

成绩不高 日本人工智能机器人“小东”弃考东大

作者： 来源：中国新闻网 发布时间：2016/11/14 13:56:35

中新网11月14日电 据日媒报道，本月14日，力争通过东京大学入学考试的人工智能机器人“小东”(东ROBO君)开发小组14日宣布，在挑战大型补习学校的大学入学统一考试模拟考试后，获得的所有学科总计标准分(日本称偏差值)为57，和去年基本持平。据报道，其物理的标准分从2015年的47大幅增至59，而数学则降低，未达到东大合格线。因为理解题目意思的阅读能力有限，研究小组今后将不把考入东大作为目标，而是转为改善记述式考试成绩等研究。

“小东”在统考中参加了语文、数学、世界史等5教科8科目的考试。语文满分为200分，其得分为96，标准分从去年的45上升到了50。另一方面，去年标准分均超过64的数1A和数2B分别获得58、56的成绩，表现不佳。

另外，记述式考试中，理科数学标准分为76，文科数学标准分为68，成绩优良。去年分别为44和59。

据了解，日本国立信息学研究所等研究小组力争最晚在2021年度通过东大入学考试，于2011年开始这一项目。

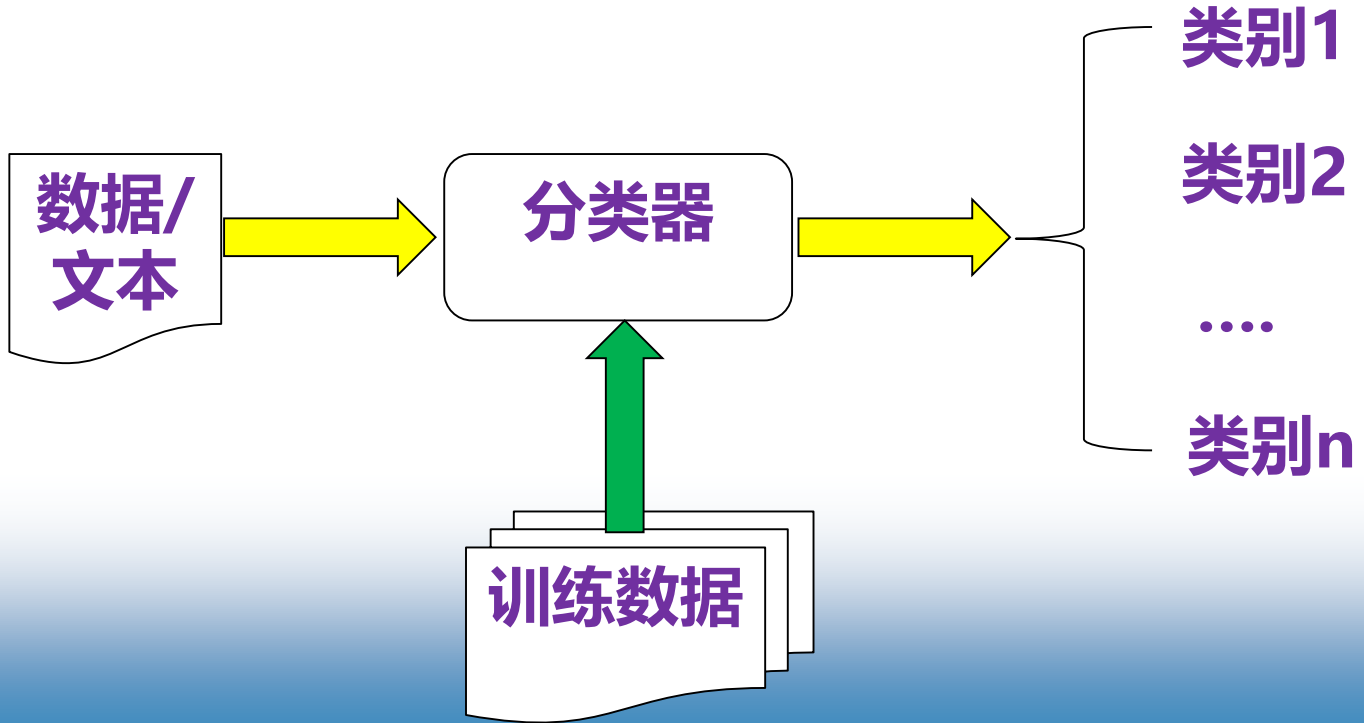
虽然成绩不断提高，但根据迄今为止研究的结果，开发小组认为难以达到考入东大的水平。今后，还将推进数学的记述式考试成绩改善研究以及成果运用于儿童教育的研究。

重要的相关基础技术

- 分类技术
- 序列标注技术
- 聚类技术
- 深度学习技术
- …

分类技术

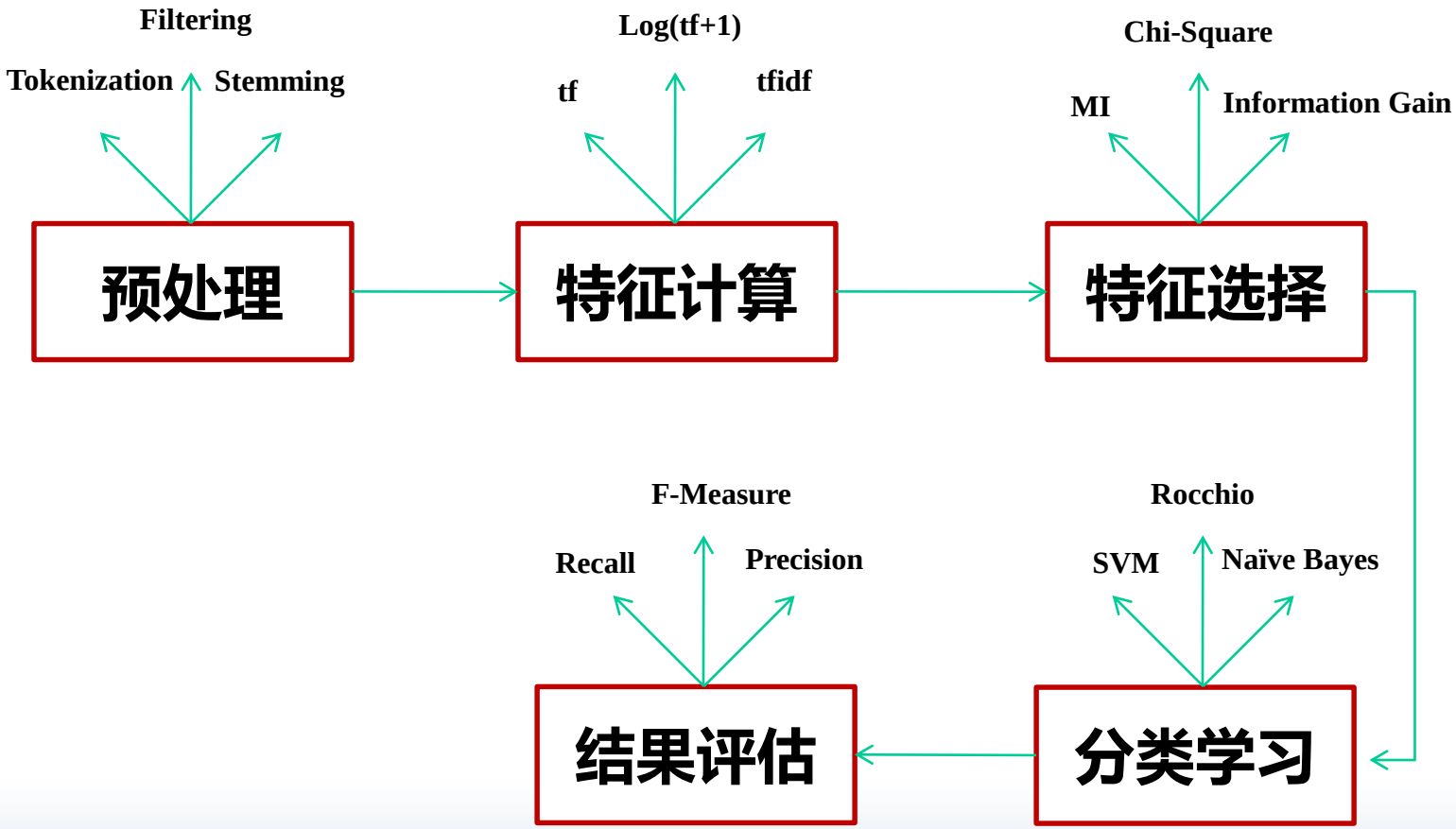
- 分类：将数据划分到已知类别
  - 分类器的构建基于有监督学习
    - 基于标注数据进行学习： 训练集



基于统计的分类

- 典型方法
  - 朴素贝叶斯
  - Rocchio
  - K近邻
  - 决策树
  - 最大熵
  - 支撑向量机
  - 感知机与神经网络
  - …

文本统计分类流程


Filtering
Tokenization
Stemming
预处理
Log(tf+1)
tf
tfidf
特征计算
Chi-Square
MI
Information Gain
特征选择
F-Measure
Recall
Precision
结果评估
Rocchio
SVM
Naïve Bayes
分类学习

集成学习

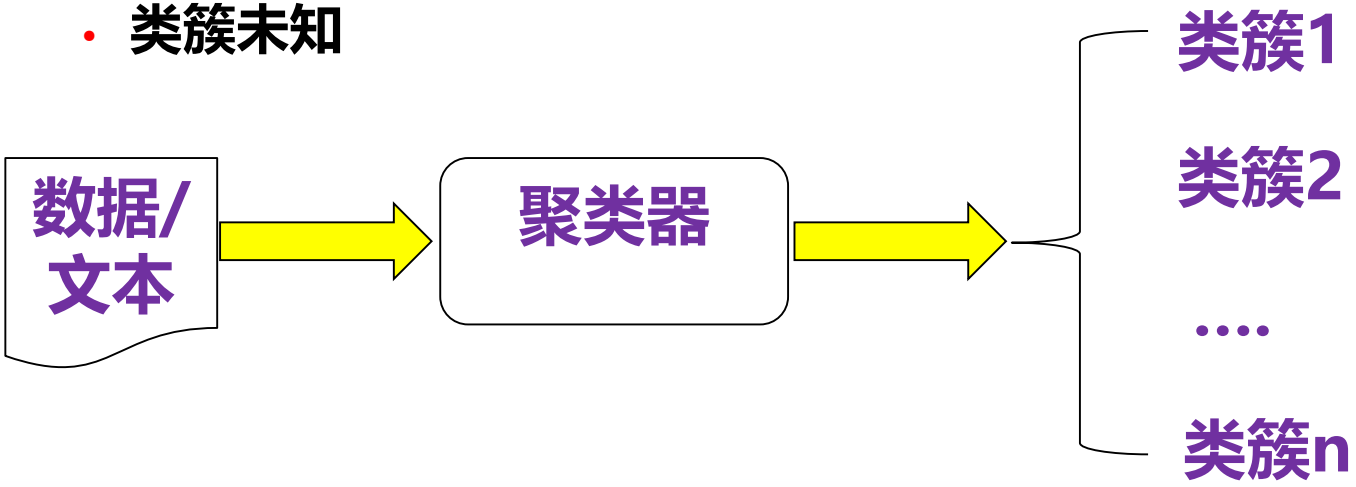
- 多个分类器可进行集成学习，取得更好的效果
  - Bagging
  - Stacking
  - Boosting等

序列标注技术

- 很多NLP问题都可看做序列标注问题，需要为序列中每个符号赋予一个标签
  - 符号标签依赖于其他符号的标签，尤其是相邻符号的标签 (not i.i.d)
- 允许集成序列中多个相互依赖的个体分类的不确定性，统一确定最可能的全局标签判断
- 典型模型
  - Hidden Markov Model (HMM)
  - Maximum Entropy Markov Model (MEMM)
  - Conditional Random Fields (CRF)

聚类技术

- 聚类：将数据自动聚集到不同类簇
  - 同一类簇内数据相似，不同类簇间数据不相似
  - 无监督学习
    - 没有标注数据
    - 类簇未知



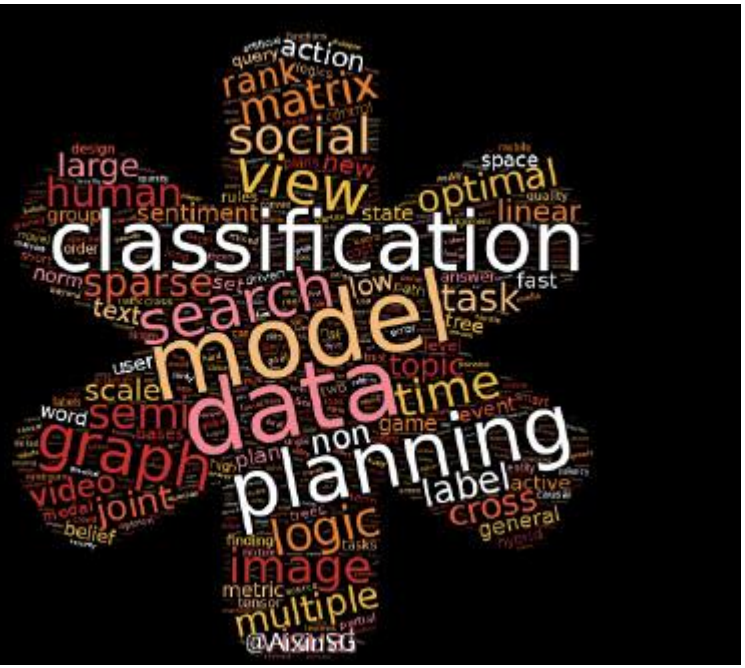
聚类技术

- 聚类算法
  - K均值聚类
  - 层次式聚类
  - 基于图分割的聚类
  - 基于密度峰值的聚类
  - …

深度学习

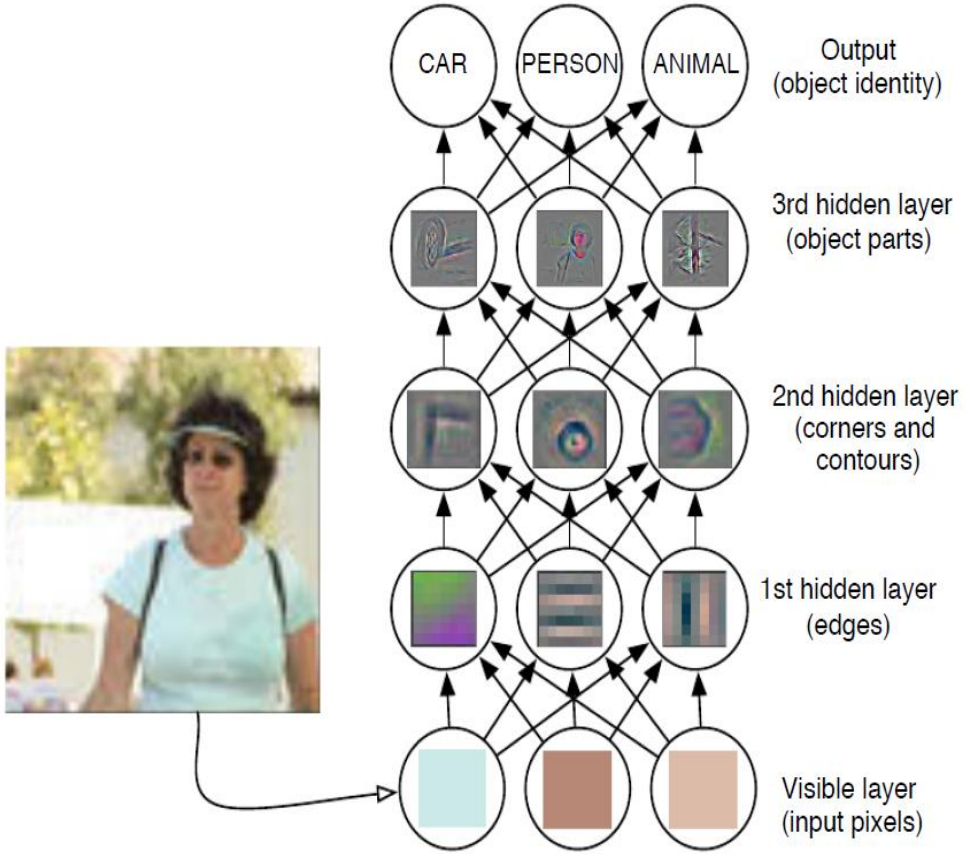
基于IJCAI16投稿论文标题分析得到的词云图

排除与“deep learning”关键词关联的论文之后得到的词云图



深度学习

- 源于人工神经网络的研究
- 通过组合低层特征形成更加抽象的高层表示属性类别或特征
- 深度学习三巨头
  - Hinton, Lecun, Bengio



深度学习

- **常用的深度学习模型**
  - **Word2Vec**
  - **Paragraph Vector**
  - **Convolutional Neural Network**
  - **Recursive Neural Network**
  - **Recurrent Neural Network**
  - **Long Short-Term Memory (LSTM)**
  - **Generative Adversarial Networks (GAN)**
  - **等等**

- Deep Learning book written by Ian Goodfellow, Yoshua Bengio and Aaron Courville

深度学习

- 在语义计算方面的应用
  - 已被应用于几乎所有任务
  - 例如知识表示、智能问答、情感分类、对话系统等

- 深度学习在文本信息处理上的应用效果不如其在计算机视觉、语音识别等方向上的应用效果

Q&A