

# **《语义计算与知识检索》研究生课程**

## **情感语义计算**

**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/course/sckr2018>**

**2018年4月4日**

# 内容

- 文本情感计算基础技术
  - 背景与概述
  - 情感分类
  - 观点抽取与摘要
- 面向微博的情感计算
- 总结与展望

# 简介

- **互联网时代文本大数据的兴起**
  - 数量大、类型多、时效性、多语言
- **主要有两种类型的文本信息**
  - **事实(Facts) 与观点(Opinions)**
- **大多数文本信息处理技术 (e.g., web search, text mining) 面向事实信息.**
- **情感分析/观点挖掘**
  - 对文本中表达的观点与情感进行计算与分析
- **为什么现在需要情感分析?**
  - 主要因为互联网上有海量的观点文本

# 简介

- **观点的重要性**

- 当我们需要做决定时观点/意见会很重要, 我们通常需要听听其他人的意见.
- 过去,
  - 对于个人: 朋友, 家庭
  - 对于商务: 调查, 咨询
- 现在, 互联网上有大量用户生成内容, 表达对任意事物的观点
  - 可以帮助个人或商务进行参考决策

# 应用背景

- 用户需求分析与精准营销


- “许多企业在线营销效果不佳，主要原因是它们几乎都以消费者年龄、性别等人口组成分类来理解客户，但**真正重要的是应该找出客户的‘深层心理侧写’，包括他们的人格特质、价值观和需求等。**”
  - IBM艾曼登研究中心主管Eben Haber
- 研究显示，人格特征确实可以用来预测消费者会购买什么样的东西，但民众不可能主动接受人格测验，只为了方便营销部门更好的入侵他们的生活。
- IBM利用软件分析微博发文，从语言习惯以及情感色彩等方面相当准确地推断出使用者的人格特征。

<http://www.vmeti.com/news/47394.html>


# 应用背景

## • 产品比较与推荐

[网页](#) [图片](#) [视频](#) [资讯](#) [地图](#) [更多](#) | [MSN](#) | [Hotmail](#)



apple ipod



所有结果

购物

POPULAR FEATURES

全部

Ease Of Use

Screen

Sound Quality

Affordability

Appearance

Battery Life

Size

Video

Speed

RESOURCES

[How cashback works](#)

[Frequently asked questions](#)

[cashback for advertisers](#)

SHOPPING

iPod touch 8GB 2nd Generation



from \$161 (24 stores)  Bing cashback · 2 - 5%

★★★★☆ user reviews (378)

★★★★☆ expert reviews (4)

Highlights includes groundbreaking technologies such as Multi-Touch, the accelerometer, 3D graphics and access to hundreds of games. Play hours of music. Create a Genius playlist of songs that go great together. Watch a movie.... [more...](#)

user reviews

product details

expert reviews

compare prices

user reviews

view: **positive comments (116)** | [negative comments \(19\)](#)

ease of use  86%

Pros: Packed with applications, very handy and easy to use.  
Timothij [www.ciao.co.uk](#) 8/17/2008 [more...](#)

Pros: User interface is beautiful and easy to find things.The built in app store store is amazing and very easy to use.  
Shinrahn [reviews.cnet.com](#) 12/30/2008 [more...](#)

Pros: Intuitive interface, very easy to use, gorgeous device, very slender profile  
Dyonas [www.ciao.co.uk](#) 10/13/2007 [more...](#)

Pros: Navigation is great, Apps are easy to use and Access, easy to sync Calender and Contacts, volume control buttons, external speaker, and a million other things  
anarchy4128 [reviews.cnet.com](#) 5/2/2009 [more...](#)

It is very quick, simple and easy to use .  
Recon3 [www.ciao.co.uk](#) 8/30/2008 [more...](#)

[众评首页](#) > [车型浏览](#) > 福克斯

## 福克斯



### 满意度 [点击查看详情](#)

统计各大论坛网友的发言，  
自动计算获得，仅作参考

<a href="#">油耗</a>	38 满意：402 不满：658
<a href="#">安全性</a>	95 满意：219 不满：10
<a href="#">空间</a>	56 满意：255 不满：199
<a href="#">动力</a>	74 满意：412 不满：147
<a href="#">操控</a>	81 满意：610 不满：142
<a href="#">外观</a>	87 满意：699 不满：106
<a href="#">内饰</a>	33 满意：239 不满：480

[概览](#) [价格](#) [品质](#) [外形](#) [油耗](#) [内饰](#) [空间](#) [安全](#) [配置](#) [操控](#) [精华](#)

### [思域 小福 观察之后还是决定小福了](#)

本人大学毕业 家里准备年前买辆车 (因为上班地好远哦) 一直看好小福和思域 但是同事们都说鬼子的车安全性能不行 撞成两半的车怎么能开啊 但是小福的内饰的确比思域难看一点 不过看了两天 觉得其..

[汽车之家](#) 发布日:2008-12-30 浏览:91 回复:14 [车型pk](#)

### [一辆马路牙子 一辆水沟里](#)

今天中午出去办事 碰到2丫超我车还狂按喇叭 我一看是一凯悦 后面还跟一中华 当时70左右的时速 (乡下水泥小路 我不让他们是超不了车的) 我马上加速到100 2y也紧跟 ..

[汽车之家](#) 发布日:2008-12-30 浏览:98 回复:18 [同类话题](#)

### [豪华尊贵不再是奢望，福克斯引领高性价..](#)

岁末降临，福克斯为了庆祝销量正式突破30w辆特别推出了一款1.8l自动豪华纪念版车型，12月15日，这款车正式上市，颠覆了高配置一定高价格的车市固有模式，进一步肯定了09福克斯在性价比上的突..

[新浪汽车论坛](#) 发布日:2008-12-30 浏览:4 回复:1 [同类话题](#)

### [福克斯大灯换市光透镜和远光透镜作业,更新共..](#)

废话不讲，看图雾灯：凯美瑞近光透镜，4300k 飞利浦，国产安定近灯：市光双光透镜，4300k，飞利浦，松下安定远灯：奔驰红外远光透镜 [卤素](#) [ 本帖最后由 小何 于 2008-12-28 2..

[东莞车迷网](#) 发布日:2008-12-27 浏览:321 回复:22 [同类话题](#)

### [换胎~~](#)

如题,准备换前面两条胎,各位有也介绍?普力斯通。价钱??顶下先~~~。不记得了。等于没说。你又话换铃.....换了n年了。..

[东莞车迷网](#) 发布日:2008-12-29 浏览:80 回复:6 [同类话题](#)

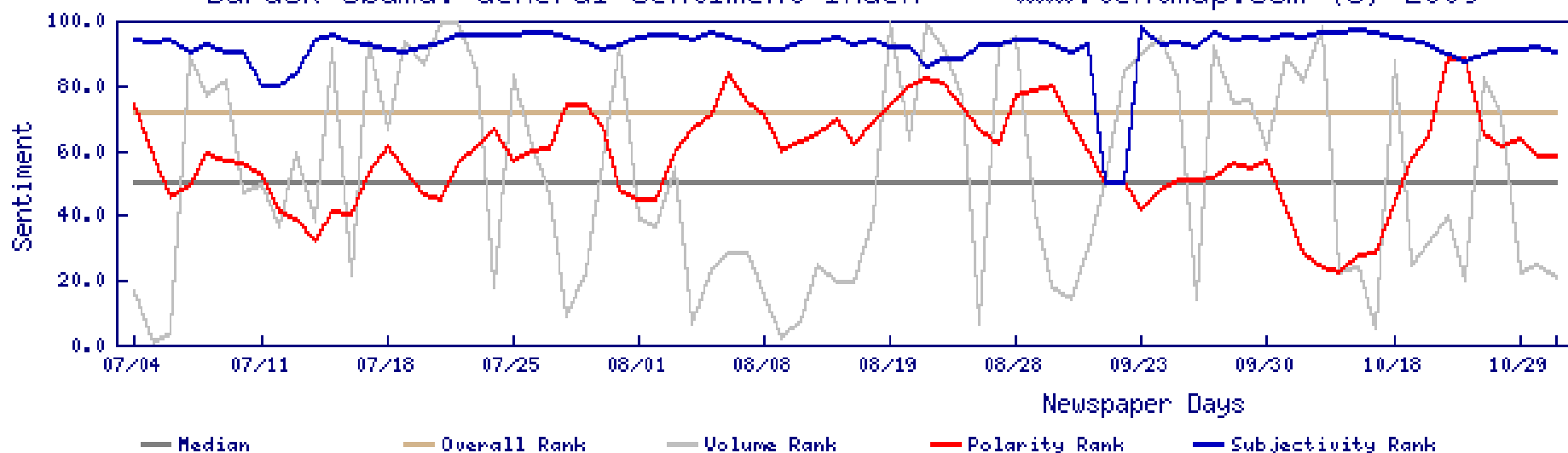
### [又出问题了~~~我难道是传说中的冤大头?](#)

既昨天晚上拔了钥匙大灯不自动关之后今天早上又有问题,打完球回家拐弯时顿时感觉方向很硬一看熄火了,都没感觉就熄火了,还是正在行驶过程中熄火的,打着了又走了,大约3公里以后我要停车的时候挂的一档..

# 应用背景

- 个人与机构声誉与实力分析

Barack Obama: General Sentiment Index --- www.textmap.com (c) 2009





周杰伦

搜索



周杰伦

人际网

个人资料

明星大家说

重名信息

人物标签: 天王 酷 优秀 天才 害羞

好评: 39%

中评: 58%

差评: 3%

发表评论 228条

**好** 周杰伦 的好评有213条

【查看更多 213条】

评论	可信度	来源
<a href="#">让张伟平对周杰伦刮目相看:"周杰伦的确是个很好的演员"</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">她说:「我觉得周杰伦很有才华</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">而王力宏评价周董则是个很优秀的艺人。</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">发现周董非常有想法</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">"之前一直听人说周杰伦很酷</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">杰伦是个聪明的小孩</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">"黄秋生说周杰伦很聪明</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">杰伦看起来是很酷的样子</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">而杰伦是一个情感丰富的人</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">"南拳妈妈"四个人异口同声的说周杰伦是个真性情的人</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">评价:周杰伦是一个很勤奋的年轻人</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">我告诉儿子周杰伦是个孝顺的孩子</a>	★★★★★	<a href="#">来源出处</a>

**差** 周杰伦 的差评有17条

【查看更多 17条】

评论	可信度	来源
<a href="#">卓远天成的凯旋•周杰伦是一个在音乐上的绝对自恋的人</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">周杰伦在我的印象中简直就是群魔乱舞的典范。</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">"周董过往给人过度自负的印象</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">周杰伦一定是个自恋的男生</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">周杰伦在我的印象中是比较迟钝的艺人。</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">周董很是有些"挂名导演"和好色的嫌疑。</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">周董当晚给人最大的印象是木讷</a>	★★★★★	<a href="#">来源出处</a>
<a href="#">到奇幻片的过渡总结了一下.发现周董是个极自恋的人!</a>	★★★★★	<a href="#">来源出处</a>

# 应用背景

## ● 电视节目满意度分析与用户反馈

### 中国电视满意度博雅榜在汉发布

2014-03-22 06:35:00 来源：湖北日报

打印  发送  字号 T | T



中国日报-看世界

+加关注

**[提要]**（记者龚雪、通讯员鲁菲菲）昨日，由中国电视艺术家协会等主办的第七届中国电视南方论坛暨2013年度中国电视满意度博雅榜发布会在汉举行。

原标题：中国电视满意度博雅榜在汉发布

湖北日报讯（记者龚雪、通讯员鲁菲菲）昨日，由中国电视艺术家协会等主办的第七届中国电视南方论坛暨2013年度中国电视满意度博雅榜发布会在汉举行。湖北卫视《长江新闻号》名列新闻栏目满意度榜单第三，仅次于央视新闻频道的《共同关注》、江苏电视台城市频道《零距离》。

中国电视满意度博雅榜作为广电业内标杆，评选方式打破了传统收视率控制格局，取而代之是新媒体、新技术及专家独立判断相结合的前沿方法，综合考评电视频道及栏目创新能力、文化品位、社会价值、人际口碑、总体印象5大标准，是追求公正、理性的学院派评价体系，逐渐成为中国电视节目评判的新标杆。本届博雅榜设置了新闻类、娱乐类、财经生活服务类栏目满意度榜单等8大奖项，全国31个省级电视台、283个频道，共1000个栏目参与角逐。

博雅榜单，卫星频道满意度前10名中，湖南、安徽、浙江占前三甲；娱乐类栏目满意度榜单，前三位分别是湖南卫视《爸爸去哪儿》，浙江卫视的《中国好声音》、《中国梦想秀》。

# 应用背景

## • 反恐与维稳

美国花大钱全球找“坏话”

.....

通过“情绪分析”寻找威胁

长久以来，美国官员一直依赖报纸和其他消息来源，追踪美国和海外发生的事件和舆论。据美国国土安全部官员透露，海外报纸及其他刊物对美国或美国领袖的负面看法，可能会暗示恐怖分子活动的蛛丝马迹。他们将这种负面信息的收集分析称作海外“情绪分析”，通过了解“报道词句及用词的情绪有多强烈”，帮助美国情报人员找出美国可能面临的威胁以及这些威胁的常见类型。

为此，国土安全部斥资240万美元作为研究经费，帮助研究机构开发先进的软件系统来更快、更全面地监视全球媒体。参与研究的包括康奈尔大学、匹兹堡大学、犹他大学等美国著名学府。开发工作大概需要数年的时间才能完成。

.....

来源：参考消息，CRI国际在线

<http://gb.cri.cn/12764/2006/10/09/2225@1248813.htm>

国土安全部支持的系统为CERATOPS

<http://www.cs.pitt.edu/mpqa/ceratops/>

北京大学语言计算与互联网挖掘研究室

# 应用背景

- 观点检索

- 检索包含与查询相关观点的文档

- 例如：“查找对‘表哥’杨达才进行评价的文章”  
“查找对北京大学有负面评价的文章”

原创于：2012-09-10 11:16:30

标签： 时政, 大快人心, 杨达才, 表哥, 微笑局长, 大贪官



微笑局长在车祸现场好开心.jpg

大快人心：杨达才“表哥”果然是大贪官  
至诚大兵

延安特大车祸事故发生后，我们远离现场，但局长杨达才，却能够在车祸事故现场笑得有多国人愤怒的“兴奋”，于是有人看到亮相时的那些不同场合不同时间所戴名表，手表就有11块之多，被人赋予“表哥”之“”，纳入了纪检部门的视野。

### 这一回杨达才“表哥”终于该由笑变哭啦 博客中国 原创博主“兆基杂谈”

延安特大车祸事故现场，陕西省安监局长杨达才笑得那么开心，那么诡秘，那么浪淫，那么得意，表现出他对民众灾难的极端的冷酷无情，所以他遭到了广大网民的人肉搜索，现在终于有结果了。据网上消息，这位杨局长“表哥”已经被双规，其亲属也已经被实际控制，据初步调查结果，已经在杨“表哥”的“仓库”里搜出现金800多万，各种文物低估价值也有7000多万，并查出他共有名牌手表23块，价值在100万以上，收取所谓“煤矿安全评估费”、“煤矿办理手续费”共2000余万，总额已经达一个亿，目前仍在继续调查中。周久耕因为名烟名表被网民搜索而成为人肉倒台第一人，杨达才堪称“长江后浪推前浪”，因为他那可耻又可笑的奸笑和手上的名表被网民搜索，终于原形毕露，这一回他恐怕再也无心笑了，这位像肥猪蠢驴一样的局长“表哥”，终于应该由笑变哭啦！

由此可见，对官员有没有监督是多么重要，有关部门肯不肯下决心去真查是多么重要！网民的监督力量何等巨大，人肉搜索的武器何等有力！如果当地的纪检委和相关司法机关也像广大网民这样锲而不舍，充分发挥监督作用，加强查处力度，何用等到今天才被动去查，早在几年前杨“表哥”被国务院问责时就应该查出他是一只硕鼠、一条大蛀虫了！已经查处的贪官大概有百多万了吧，但是又有几个是纪检机关和司法机关主动查出来的呢？

人民群众由于条件的限制，一般只能打过街老鼠，比如杨达才这只硕鼠，如果他在车祸现场没有诡笑浪笑淫笑奸笑，如果他没有天天换戴名表，他还是深藏在洞里的硕鼠，民众想打也打不了，所以相关的纪检、司法机关应该下大决心大力气大动作去挖深藏在洞里的硕鼠，才是真正尽职尽责，才是真正为纯洁党的队伍贡献力量。

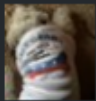
至诚大兵又及：这一回杨达才“表哥”终于由笑变哭啦，国人却大快人心。正告那些尚未暴露的贪官，赶快收手吧，记住陈毅元帅说过的话：不是不报，时候不到；伸手必被捉！

# tweetfeel



Search

Try some Twitter trends: [Happy National Kiss Day](#) [What is LOVE](#) [Today is Friday](#) [Cabin in the Woods](#) [Ann Romney](#) [New Orleans](#) [Still UP](#)



**obama** rocks. Millionaires should pay their fair share! Support our Prez!



@heartsasmagnets idk w/ him. i love **obama** so i'm good for me. it's the rest of the country that worries me since so many WANTED santorum



@KrystinaJacobs Yeah. I don't like **obama** for the same reasons I didn't like bush. Wars without congress, Gitmo, Torture, Taking away rights



4 some odd reason I have these conservatives & liberterians following me. If ur following me 2 talk smack don't waste our time! I love **obama**



I don't like Michael Jackson RACIST I don't like **obama** RACIST! I don't like Kevin hart RACIST I don't like Adam Sandler Yeah, he can't act



HAHAHAHA I wish our President would do the same. **obama** FTW! LoL <http://t.co/8ZEBCXAY>

# 相关国际评测

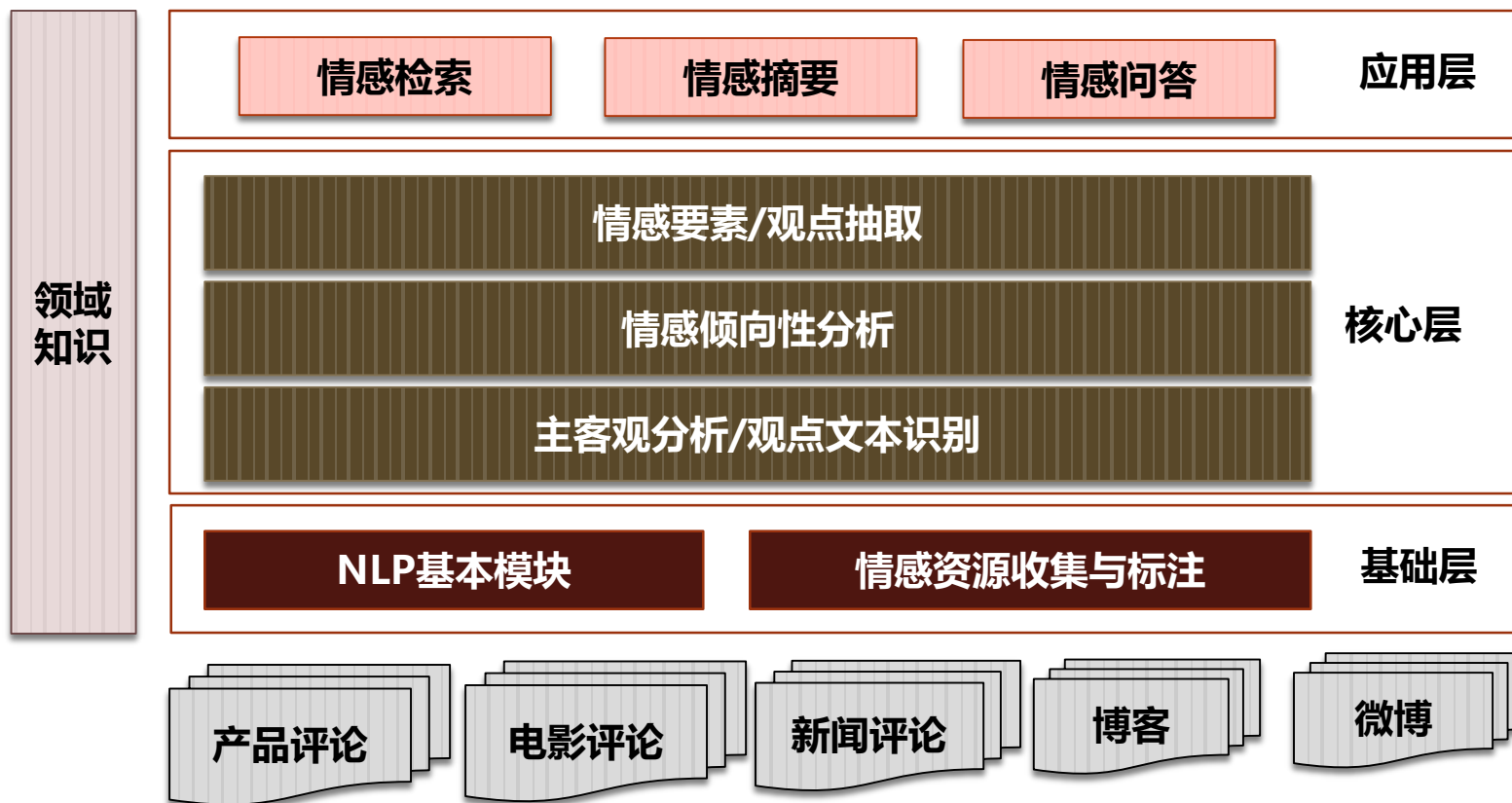
- 日本NII组织的NTCIR
  - NTCIR-6 ~ NTCIR-8 MOAT
    - 主客观分析, 倾向分析, 观点要素抽取, 观点问答
    - 英文, 日文, 繁体中文, 简体中文
- SemEval2013~2018
  - 面向Twitter的情感分类与观点分析
- 其他相关评测
  - TREC Blog Track
    - Opinion retrieval + polarity subtask
  - TAC
    - Opinion QA Task
    - Opinion Summarization Task

# 相关国内评测

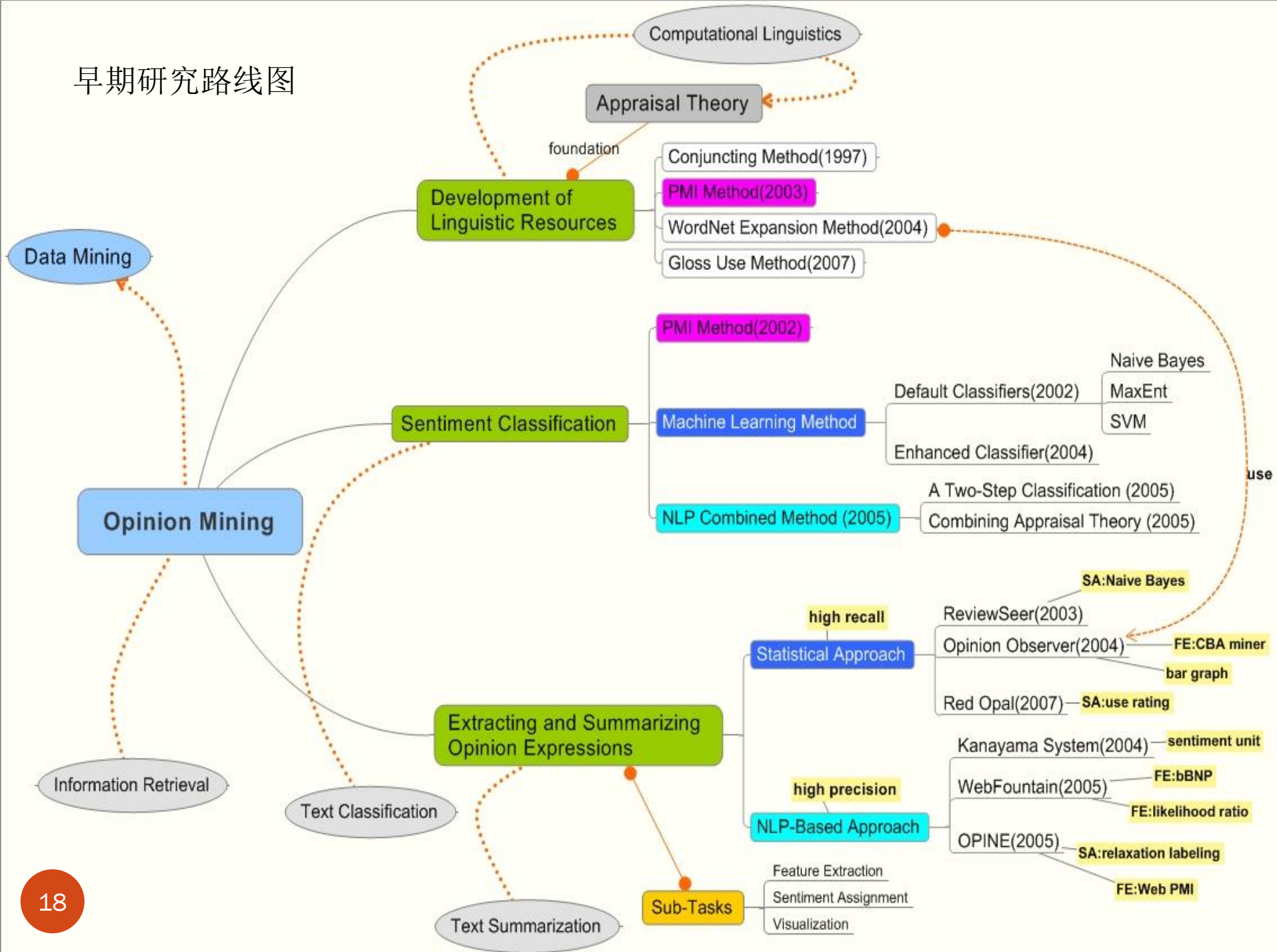
- **中文信息学会中文倾向性分析评测 COAE**
  - 观点词抽取、观点句抽取、评价要素抽取、观点检索、比较句中观点判别
- **计算机学会中文信息技术专委会情感分析评测 NLPCC**
  - 来自微博数据
    - 不同于产品评论数据，微博数据领域多样，话题广泛，表达自由，极具挑战性，例如“三亚春节宰客”事件
  - 任务设置：观点句识别、情感倾向性判断、情感要素抽取等
  - 数据全部免费下载



# 研究框架



## 早期研究路线图



# 内容

- 文本情感计算基础技术
  - 背景与概述
  - 情感分类
  - 观点抽取与摘要
- 面向微博的情感计算
- 总结与展望

# 情感分类

- **将文本按照所表达的总体情感进行分类**
  - 例如：正面(Positive), 负面(negative), (possibly) 中性(neutral)
- **与基于话题的文本分类相似又不同**
  - 对于基于话题的文本分类, 话题词汇很重要
  - 情感分类中, 情感词汇更加重要, 例如 great, excellent, horrible, bad, worst, etc.

# 情感分类任务

- **主客观分析** (subjectivity classification)
  - 客观：反映关于世界的事实信息， “iPhone是苹果产品”
  - 主观：反映个人感受、观点或信念等， “我喜欢iPhone”
- **subjectivity != opinionated**
- 有的主观句可能不表达情感或观点
  - I think that he went home.
- 有的客观句能够表明或暗示情感或观点（隐式观点）
  - The earphone broke in two days.
  - I brought the mattress(床垫) a week ago and a valley has formed.

# 情感分类任务

- **倾向性分析** (Sentiment Classification/Polarity Classification)
  - 对包含观点的文本进行倾向性判断
  - 一般为以下三类
    - 褒义: “外观不错”
    - 贬义: “软件目前不丰富”
    - 中性/无倾向: “软件采用了安卓系统”
      - 在一些问题中不考虑中性
- **粒度**
  - 词、句子、文档

# 情感资源

- 情感分析的基础
- 英文资源较多
  - 情感词典: SentiWordNet, Inquirer等
    - 包含词语、短语等
    - 倾向性词语, 主观性词语
  - 已标注语料库数量较多
  - 提供开源情感分析工具: OpinionFinder

# 情感资源

- **中文资源较少，逐年增多**
  - 知网HowNet提供了部分情感词汇，部分高校也提供了情感词汇，但质量参差不齐
  - 近两年的评测提供了中文标注文本
    - NTCIR, COAE、NLP&CC等
- **情感资源基本上跟领域、语言有关**
- **主客观分析与倾向性分析的资源也不一样**



# 情感资源

Existing lexicons

## Existing lexicons: SentiWordNet

- abi
  - abl
  - abc
  - abs
  - abs
  - abs
  - abu
- P: 0.75 O: 0.25 N: 0 **good**#101123148  
having desirable or positive qualities especially those suitable for a thing specified; "good news from the hospital"; "a good report card"; "when she was good she was very very good"; "a good knife is one good for cutting"  
=tw  
nnsrc=ph
  - P: 0 O: 1 N: 0 **good**#2 full#6 00106020  
having the normally expected amount; "gives full measure"; "gives good measure"; "a good mile from here"  
nnsrc=ph
  - P: 0 O: 1 N: 0 **short**# 201436003  
(primarily spatial sense) having little length or lacking in length; "short skirts"; "short hair"; "the board was a foot short"; "a short toss"
  - P: 0.125 O: 0.125 N: 0.75 **short**#3 little#6 02386612  
low in stature; not tall; "he was short and stocky"; "short in stature"; "a short smokestack"; "a little man"

# SentiWordNet

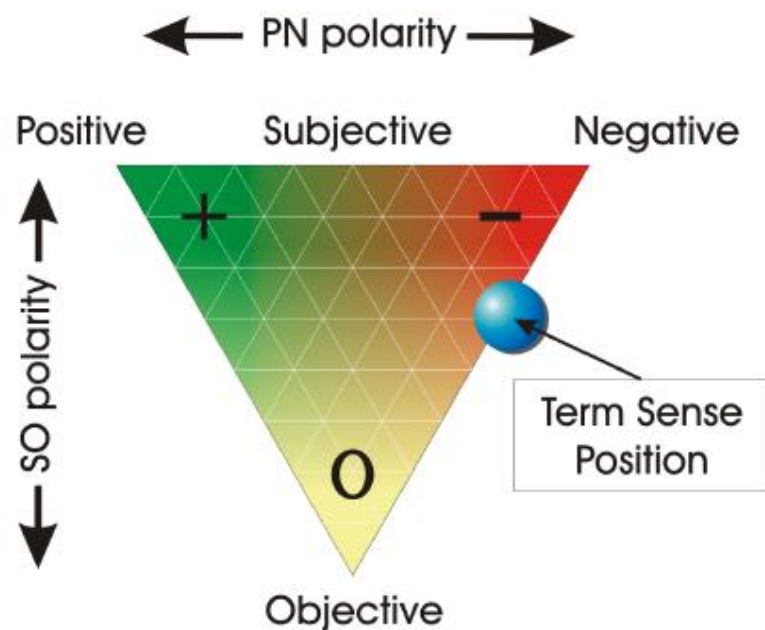


Figure 1: The graphical representation adopted by SentiWordNet for representing the opinion-related properties of a term sense.

estimable Search word © show position

## Adjective

3 senses found.

<p><math>P = 0.75, N = 0, O = 0.25</math></p>	<p><a href="#">estimable(1)</a> deserving of respect or high regard</p>
<p><math>P = 0.625, N = 0.25, O = 0.125</math></p>	<p><a href="#">honorable(5)</a> <a href="#">good(4)</a> <a href="#">respectable(2)</a> <a href="#">estimable(2)</a> deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"</p>
<p><math>P = 0, N = 0, O = 1</math></p>	<p><a href="#">computable(1)</a> <a href="#">estimable(3)</a> may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"</p>

[main page](#)

(c) Andrea Esuli 2005 - [andrea.esuli@isti.cnr.it](mailto:andrea.esuli@isti.cnr.it)

Figure 2: SentiWordNet visualization of the opinion-related properties of the term *estimable*.

# 知网情感词典

## Release of the latest updated version of HowNet

- Today we release "Chinese/English Vocabulary for Sentiment Analysis (VSA)(Beta version)". The VSA includes 12 subsets.

### 1. "Chinese Vocabulary for Sentiment Analysis", which contains 6 sub-files:

"Plus Feeling", e.g. 爱, 赞赏, 快乐, 感同身受, 好奇, 喝彩, 魂牵梦萦, 嘉许 ...

"Minus Feeling", e.g. 哀伤, 半信半疑, 鄙视, 不满意, 不是滋味儿, 后悔, 大失所望 ...

"Plus Sentiment", e.g. 不可或缺, 部优, 才高八斗, 沉鱼落雁, 催人奋进, 动听, 对劲儿 ...

"Minus Sentiment", e.g. 丑, 苦, 超标, 华而不实, 荒凉, 混浊, 畸轻畸重, 价高, 空洞无物 ...

"opinion"

"degree"

### 2. "English Vocabulary for Sentiment Analysis", which contains 8945 entries:

"Plus Feeling", 772 entries, e.g. happy, be jealous, admiration, consent, welcome, look

"Minus Feeling", 1012 entries, e.g. defy, disappointed, fear, criticize, regret, pull a long

"Plus Sentiment", 3596 entries, e.g. good-looking, high-quality, effective, tranquility, safe

"Minus Sentiment", 3562 entries, e.g. grotesqueness, inferior, expensive, expensively, be

"opinion"

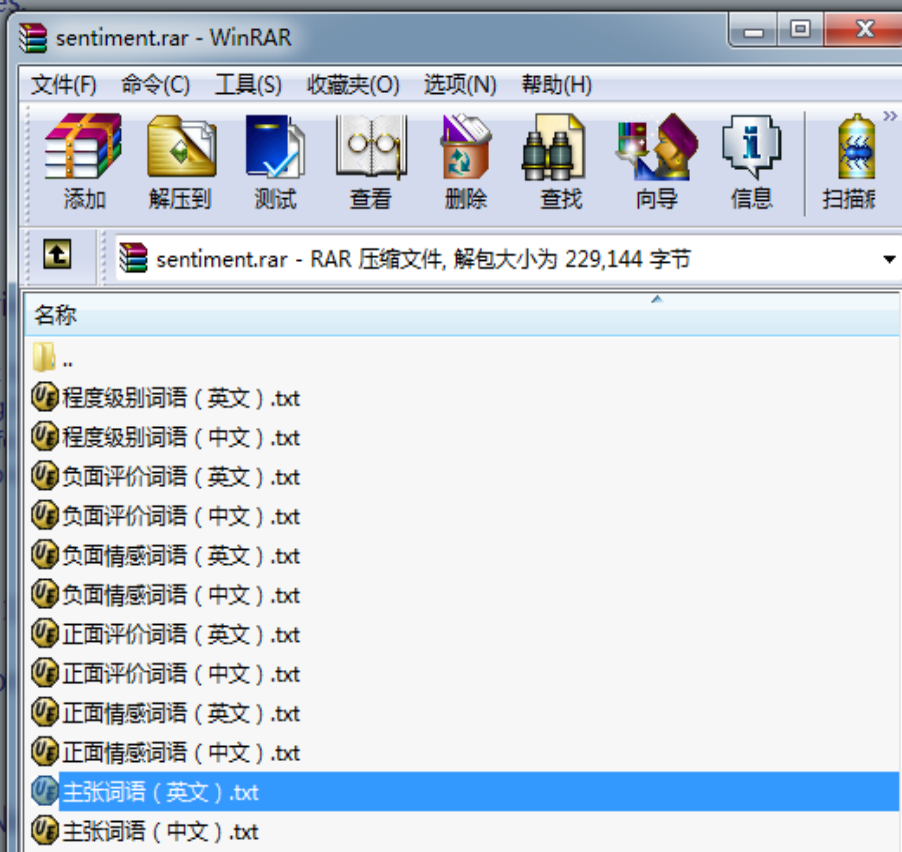
"degree"

### 3. "Chinese/English Vocabulary for Sentiment Analysis" which contains

The "Chinese/English Vocabulary for Sentiment Analysis (VSA)(Beta version)"

### Chinese/English Vocabulary for Sentiment Analysis

- Oct. 08, 2007 Release of the latest updated version of Mini-HowNet



# 情感词汇资源构建

- **任务**

- 确定词语的主观性(subjectivity)
- 确定词语的倾向(orientation)
- 确定词语态度的强度(strength)

- **例子**

- **Objective: vertical, yellow, liquid**
- **Subjective**
  - Positive: good < excellent
  - Negative: bad < terrible

# 情感词汇资源构建

- 连接词方法(Conjunction Method)
- PMI方法
  - Orientation
  - Subjectivity
- WordNet扩展方法
- 释义方法(Gloss Use Method)
  - Orientation
  - Subjectivity
  - SentiWordNet

# 连接词方法

[Hatzivassiloglou and McKeown, 1997]

- 假设

- 用 'and' 相连的形容词通常具有相同的倾向, 而用 'but' 相连的形容词通常具有相反的倾向

“beautiful and clever” ,  
“beautiful but stupid”

- 对形容词按照不同倾向聚类

### The Homestay Experience - Cultural Kaleidoscope 2006

My host's home **was very nice and** comfortable. I got to try all types of food; Malaysian, Chinese, Indonesian and I loved it all. My host's parents were very ...

[www.gardenschool.edu.my/studentportal/aec/Kaleidoscope06/experience.asp](http://www.gardenschool.edu.my/studentportal/aec/Kaleidoscope06/experience.asp) - 10k -

[Cached](#) - [Similar pages](#) - [Note this](#)

### PriceGrabber User Rating for Watch Your Budget - PriceGrabber.com

Reviews, Camera I purchased **was very nice and** a bargain. There was a problem with shipping, but was resolved quickly. Buy with confidence from this vendor. ...

[www.pricegrabber.com/rating\\_getreview.php/retid=5821](http://www.pricegrabber.com/rating_getreview.php/retid=5821) - [Similar pages](#) - [Note this](#)

### Testimonials

"Everybody **was very nice and** service was as fast as they possibly could. ... "Staff member who helped me **was very nice and** easy to talk to " ...

[www.sa.psu.edu/uhs/news/testimonials.cfm](http://www.sa.psu.edu/uhs/news/testimonials.cfm) - 22k - [Cached](#) - [Similar pages](#) - [Note this](#)

### Naxos Villages - Naxos Town or Chora Reviews: Very nice and very ...

-Did you enjoy the trip to Naxos Town: Yes it **was very nice and** very scenic -In order to get to the village were there enough signs in order to find it: It ...

# PMI方法

- PMI: 点互信息(Pointwise Mutual Information)

$$\text{pmi}(x; y) \equiv \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

- 判别倾向性

- Turney and Littman, 2003
- 具有相似倾向性的词语倾向于在文档中共同出现

- 判别主观性

- Baroni and Vegnaduzzo, 2004
- 主观性形容词倾向于出现在其他主观性形容词周围



# PMI方法

- 可基于AltaVista搜索引擎的NEAR操作符返回的结果数量进行PMI的计算

$$PMI(t, t_i) = \log_2 N \frac{\text{hits}(t \text{ NEAR } t_i)}{\#(t)\#(t_i)}$$

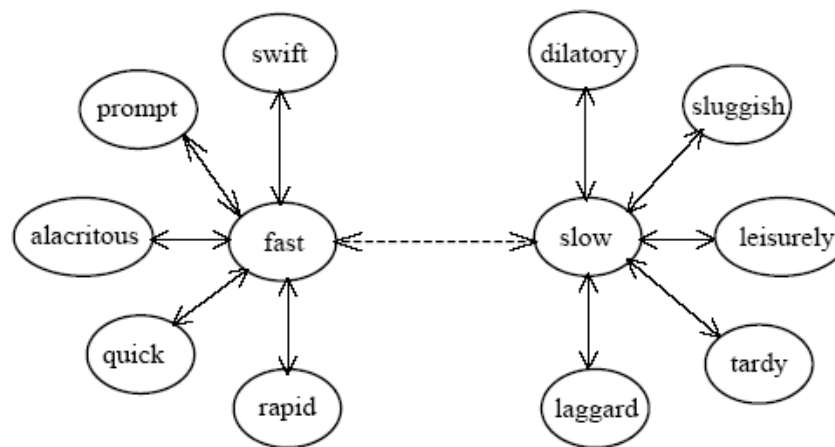
$t$ : 目标词  
 $t_i$ : 种子词

- 预测词语的倾向性SO(t)

$$SO(t) = \sum_{t_i = Pos} PMI(t, t_i) - \sum_{t_i = Neg} PMI(t, t_i)$$

# WordNet扩展方法

- Hu et al., 2004
- 使用词语之间的同义、反义关系
- 假设
  - 形容词通常与其同义词具有相同的倾向性，而与其反义词具有相反的倾向性



- 利用种子形容词集，能够获得WordNet中所有形容词的倾向性

# 释义方法

- Esuli et al., 2005, 2006

- 假设

- 倾向性

- 具有相似倾向性的词语具有相似的释义

- 主观性

- 具有相似倾向性的词语具有相似的释义
    - 不具有倾向性的词语具有无倾向的释义

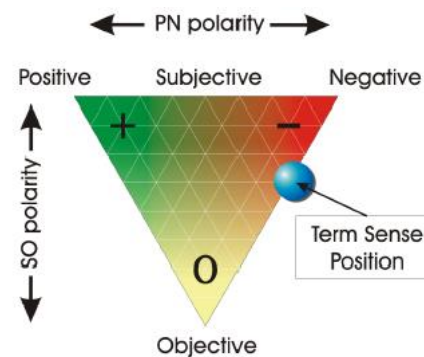
good: that which is pleasing or valuable or useful; agreeable or **pleasing**  
beautiful: aesthetically **pleasing**  
pretty: **pleasing** by delicacy or grace; not imposing

yellow: similar to the color of an egg yolk  
vertical: at right angles to the plane of the horizon or a base line

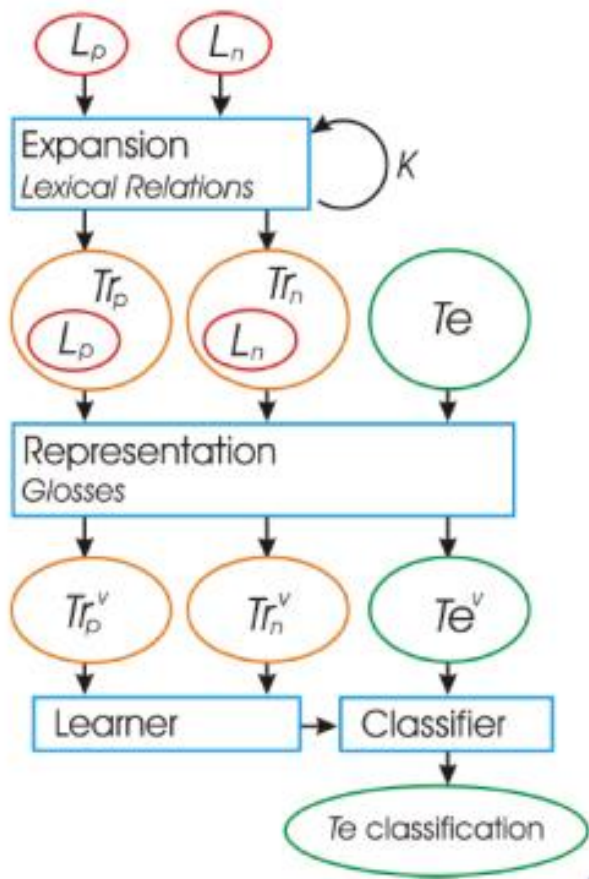
- SentiWordNet

- WordNet中的所有词有三个分数值
    - positivity, negativity, and objectivity

北京大学语言计算与互联网挖掘研究室



# 释义方法



## 步骤

1. 输入种子词集( $L_p, L_n$ )
2. 基于词典中词汇同义关系进行词汇扩展, 得到集合 $Tr_p$ 与 $Tr_n$ , 作为第4步的训练集
3. 对于 $Tr_p \cup Tr_n$ 中或测试集中的词语 $t_i$ , 收集 $t_i$ 在机器可读词典中的释义, 作为该词语的文本(向量)表示
4. 基于 $Tr_p \cup Tr_n$ 中的词语进行二类分类器的训练, 然后对测试集中的词语进行倾向性分类

# 利用Web获取标注语料

- 利用用户打分获取产品评论倾向

手机摄像头有问题 ★★☆☆☆

优点：其他没有问题

不足：手机摄像头有问题，拍摄不能正常对焦，画面一跳一跳的

使用心得：手机摄像头有问题，拍摄不能正常对焦，画面一跳一跳的，懒得换了，就这样吧。查了下，很多人有这个问题，想买的人要考虑好了。所谓国货，任重而道远。

回复 (0) 购买日期：2012-10-30

这条评价对您有用吗？

- 利用表情符号获取文本倾向



Da1mOn: 陕西杨达才就是你的下场吗, "表妹"? ?



# 情感分类方法

- **基于规则的方法**
  - 利用情感词典、模板
- **基于机器学习的方法**
  - 利用标注语料
- **混合方法**

# 文档情感分类-PMI方法

- Turney et al., 2002

- 步骤

- 只抽取包含形容词或副词的两个词构成的短语
- 短语phrase的语义倾向
  - $SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$
- 文档的语义倾向为所有短语语义倾向的平均值

- 实验

- 410 reviews from Epinions (epinion.com): 170 positive, 240 negative

Domain of review	Accuracy	Domain of review	Accuracy
Automobiles	84.00%	Movies	65.83%
- Honda Accord	83.78%	- The Matrix	66.67%
- Volkswagen Jetta	84.21%	- Pearl Harbor	65.00%
Banks	80.00%	Travel Destination	70.53%
- Bank of America	78.33%	- Cancun	64.41%
- Washington Mutual	81.67%	- Puerto Vallarta	80.56%

# 文档情感分类-基于分类器

- Pang and Lee, 2002
- 看作是特殊的文本分类任务
- 文档采用标准的特征向量表示
- 实验

**NAACL 2018 Test-of-Time Award**

*Title: BLEU: a Method for Automatic Evaluation of Machine Translation*  
*Authors: Kishore Papineni, Salim Roukos, Todd Ward and Wei-jing Zhu*

*Title: Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms*  
*Author: Michael Collins*

*Title: Thumbs up?: Sentiment Classification using Machine Learning Techniques*  
*Authors: Bo Pang, Lillian Lee, Shivakumar Vaithyanathan*

- Data : movie reviews (Internet Movie Database), rating -> negative, neutral, positive
- Naïve Bayes, Maximum Entropy, Support Vector Machine

Features	# of features	Frequency or presence?	NB	ME	SVM
unigrams	16165	freq.	78.7	N/A	72.8
unigrams	16165	pres.	81.0	80.4	82.9
unigrams+bigrams	32330	pres.	80.6	80.8	82.7
bigrams	16165	pres.	77.3	77.4	77.1
unigrams+POS	16695	pres.	81.5	80.4	81.9
adjectives	2633	pres.	77.0	77.7	75.1
top 2633 unigrams	2633	pres.	80.3	81.0	81.4
unigrams+position	22430	pres.	81.0	80.1	81.6



# 文档情感分类-两阶段分类

- Wilson et al., 2005
- 先对短语进行倾向分类，文档倾向由短语平均倾向所决定
- 短语分类
  - 步骤1：将短语分类为中性或具有倾向性
  - 步骤2：将具有倾向性的短语分类为正面或负面
- 28维特征：使用了NLP技术进行抽取
  - 4 Word Features, 8 Modification Features, 11 Structure Features, 3 Sentence Features, 1 Document Feature
- 实验
  - Data : Multi-perspective Question Answering (MPQA) Opinion Corpus

neutral-polar classification (%)

polarity classification (%).

Features	Accuracy
Word token	73.6
Word+priorpol	74.2
28 features	75.9

Features	Accuracy
Word token	61.7
Word+priorpol	63.0
10 features	65.7

# 句子情感分类

- 文档情感分类对于很多应用并不适用，需要进行句子情感分类
- 许多工作侧重于从新闻文档中识别主观句
  - 分类：客观与主观
  - 基于机器学习

# 句子情感分类-基于模式规则学习

[Rilloff and Wiebe, 2003]

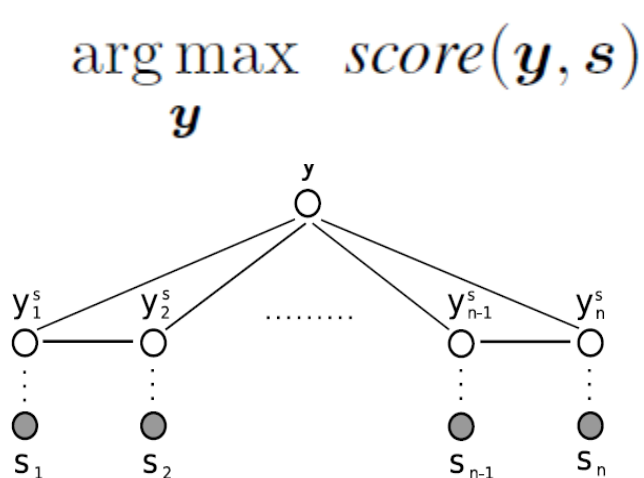
- **自举方法**

- 利用高准确率分类器自动识别一些主观句与客观句
- 从识别得到的主客观句中学习一些模式规则
  - 利用句法模板进行约束
- 利用学习得到的模式规则抽取更多的主客观句

(上述步骤可迭代)

# 文档与句子统一情感分类

- 层叠方法
  - 首先进行句子级分类，然后进行文档级分类
- 一体化方法
  - 同时对文档情感与句子情感进行建模并求解
    - 句子情感会影响文档情感，文档情感同样也会影响句子情感（提供上下文）
  - 例如结构化模型[McDonald et al. 2007]

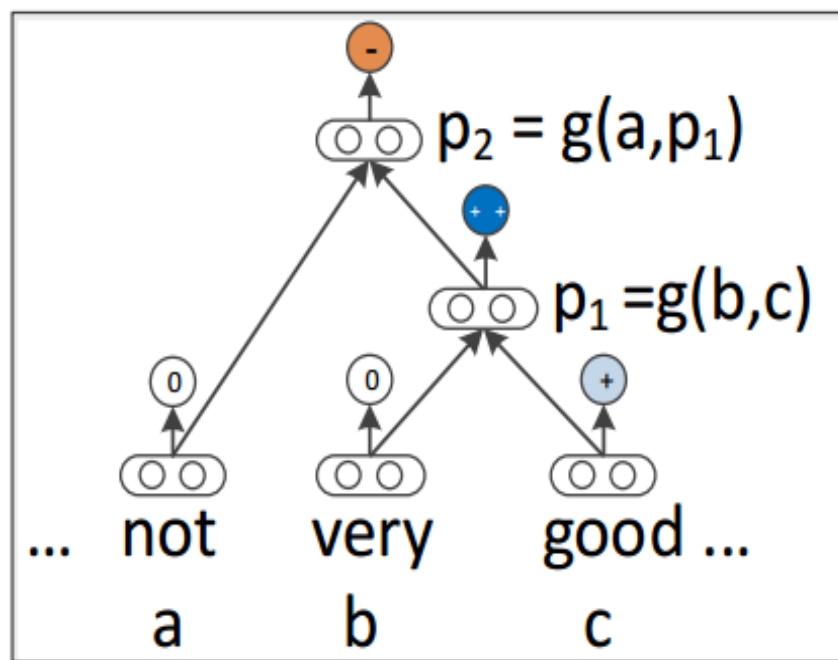
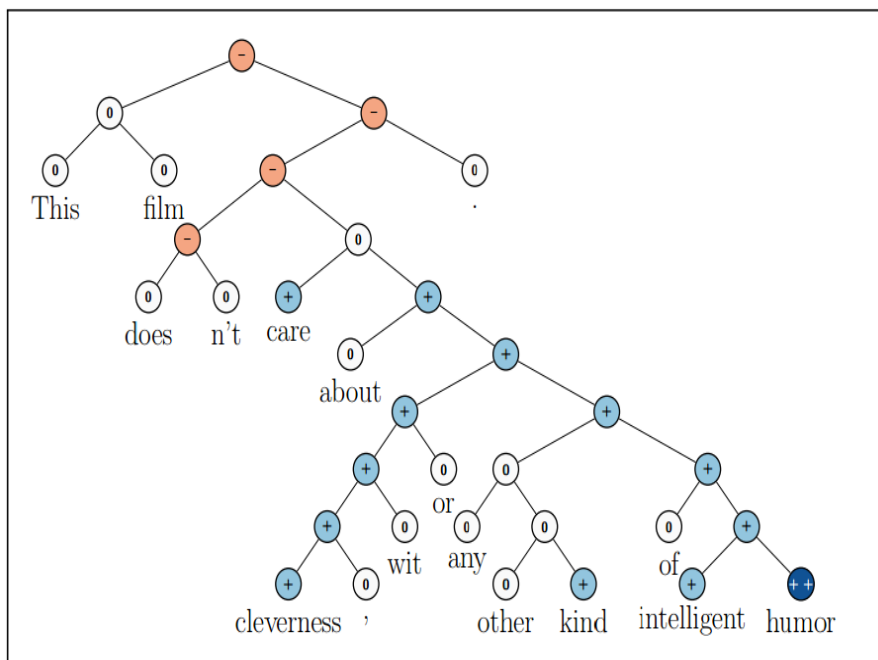


$$\begin{aligned} \text{score}(y, s) &= \text{score}((y^d, \mathbf{y}^s), s) \\ &= \text{score}((y^d, y_1^s, \dots, y_n^s), s) \\ &= \sum_{i=2}^n \text{score}(y^d, y_{i-1}^s, y_i^s, s) \end{aligned}$$

$$\text{score}(y^d, y_{i-1}^s, y_i^s, s) = \mathbf{w} \cdot \mathbf{f}(y^d, y_{i-1}^s, y_i^s, s)$$

# 基于深度学习的情感分类

- **RNN系列**[Socher et al. 2013]
  - 考虑句法结构
  - 基于RNN进行短语与句子的情感分类
  - 人工标注了Sentiment TreeBank
  - 利用RNN获得短语和句子的向量表示
    - 尝试了RNN及其扩展（如Matrix-Vector RNN, RNTN等）



# 基于深度学习的情感分类

- Paragraph Vector [Le and Mikolov 2014]
  - 不考虑句法结构信息
  - 为文档直接学习得到一个向量表示

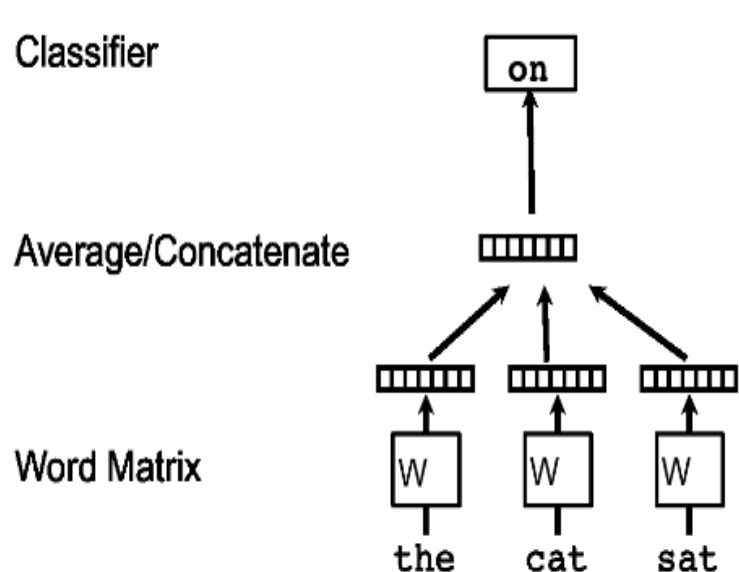


Figure 1. A framework for learning word vectors. Context of three words (“the,” “cat,” and “sat”) is used to predict the fourth word (“on”). The input words are mapped to columns of the matrix  $W$  to predict the output word.

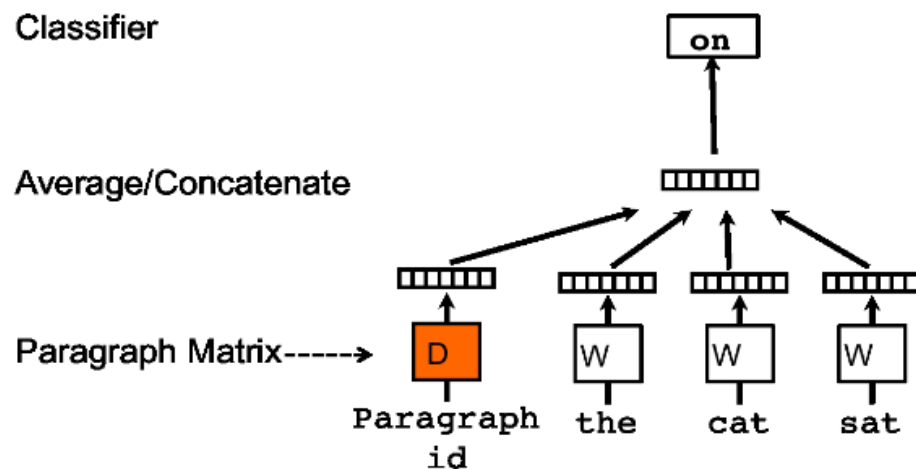


Figure 2. A framework for learning paragraph vector. This framework is similar to the framework presented in Figure 1; the only change is the additional paragraph token that is mapped to a vector via matrix  $D$ . In this model, the concatenation or average of this vector with a context of three words is used to predict the fourth word. The paragraph vector represents the missing information from the current context and can act as a memory of the topic of the paragraph.

# 基于深度学习的情感分类

- **CNN** [Kim 2014]
- **DCNN** [Blunsom et al. 2014]
  - 不考虑句法结构信息
  - 主要基于卷积操作与池化操作获得文本的向量表示
    - 卷积操作获得词序列特征
    - 池化操作选择最重要特征

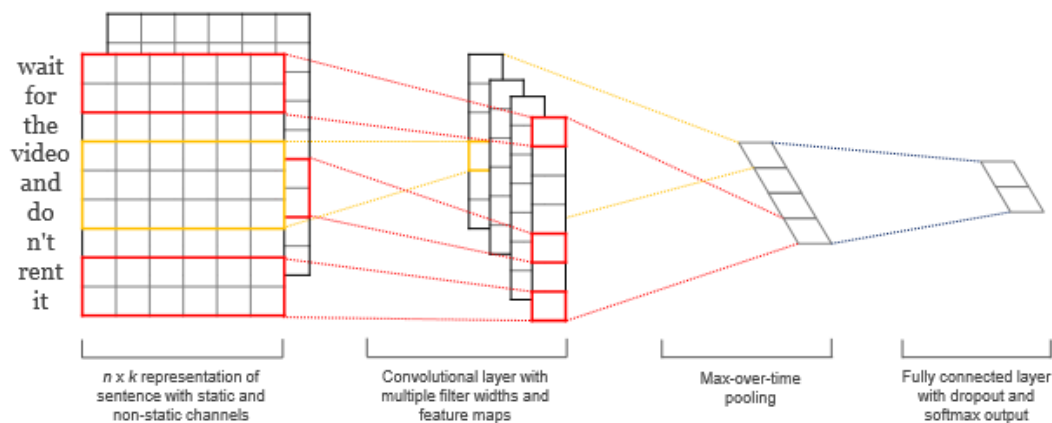


Figure 1: Model architecture with two channels for an example sentence.

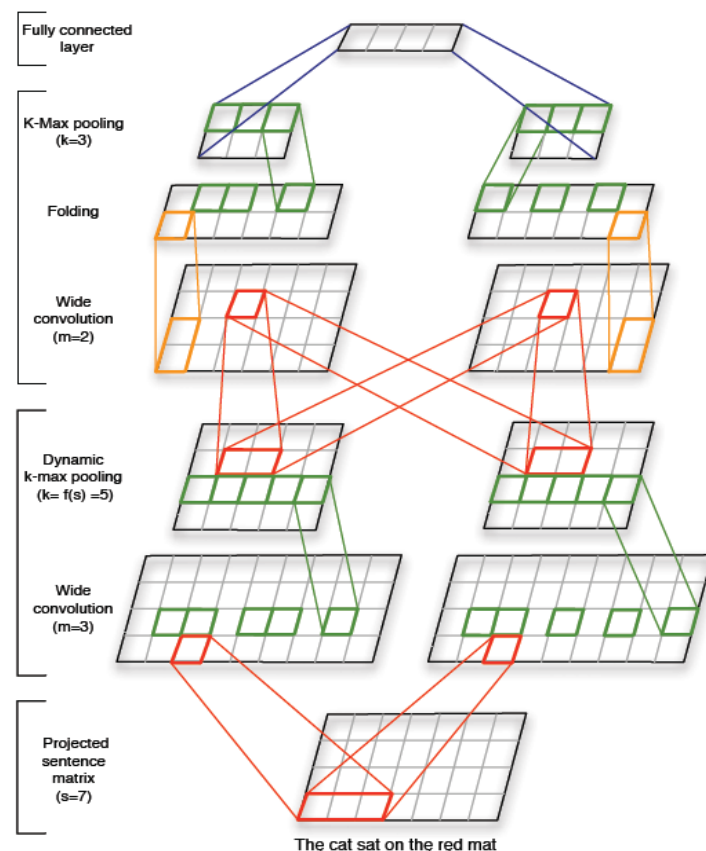


Figure 3: A DCNN for the seven word input sentence. Word embeddings have size  $d = 4$ . The network has two convolutional layers with two feature maps each. The widths of the filters at the two layers are respectively 3 and 2. The (dynamic)  $k$ -max pooling layers have values  $k$  of 5 and 3.

# 情感分类现状与难点

- 产品评论的情感分类效果较好，可实用
- 社交媒体情感分析相当困难
  - 写作自由，不规范
    - Twitter: good gud gd goood goooood gooooo qood goooooood goooooood gudd #good  
goodd guud g00d goooooooooo goooooooooo
  - 反讽
    - 你这样的行为真是为你家人争光!
  - 情感倾向需要跟对象关联才有意义
    - 我反对站中 Vs. 我支持警察清场



# 内容

- 文本情感计算基础技术
  - 背景与概述
  - 情感分类
  - 观点抽取与摘要
- 面向微博的情感计算
- 总结与展望

# 样例评论

- “I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. ...”
- **What do we see?**
  - **观点(Opinions)**, **观点对象(targets of opinions)**, and **观点持有者(opinion holders)**

# 观点抽取

- **观点的组成**

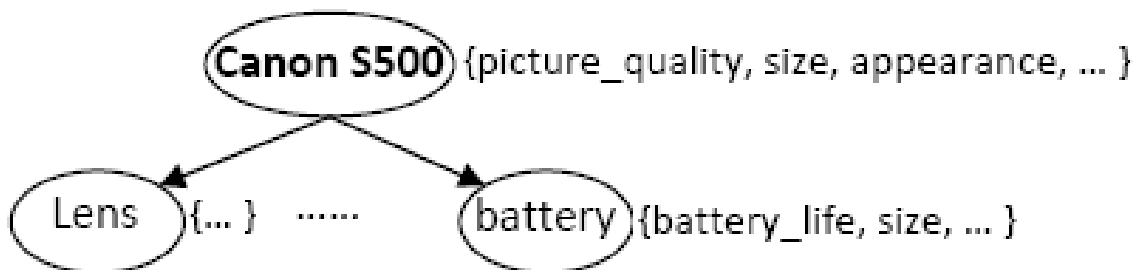
- **观点持有者(Opinion holder):** 持有/表达观点的人或机构
- **目标对象(Object):** 观点的表达对象, 所指向的对象
- **观点表达(Opinion):** 持有者对目标对象的态度、评价

- **一般针对产品评论进行分析**

# 目标对象

[Liu, Web Data Mining book, 2006]

- 一个对象  $o$  是一个产品、人物、事件、机构、或话题，可表示为
  - 一个部件/子部件构成的层次结构
  - 每个部件都有一系列属性(attributes)



- 观点可表达于任一节点或属性
- 部件或属性也称为**特征**(aspect , features)

# 观点表示

- 一个观点表示为五元组

$(o_j, a_{jk}, so_{ijkl}, h_i, t_l)$

其中

- $o_j$  为目标对象.
- $a_{jk}$  是对象  $o_j$  的特征
- $so_{ijkl}$  为观点所表达的情感值 (如倾向性分类)
- $h_i$  为观点持有者
- $t_l$  为观点表达的时间

# 观点抽取的目标

- 给定观点文本
  - 抽取所有的五元组( $o_j, a_k, so_{ijkl}, h_i, t_l$ )
- 基于五元组，可将无结构化文本结构化
  - 可利用传统数据挖掘与可视化技术进行挖掘与呈现
  - 可以定量与定性分析

# 观点抽取任务很困难

- “This past Saturday, I bought a *Nokia* phone and my girlfriend bought a *Motorola* phone with *Bluetooth*. We called each other when we got home. The voice on my phone was not so clear, worse than my previous phone. The battery life was long. My girlfriend was quite happy with her phone. I wanted a phone with good sound quality. So my purchase was a real disappointment. I returned the phone yesterday.”

# 观点抽取子任务

- 重点关注两个子任务
  - 特征抽取与聚类 (aspect extraction and grouping)
    - 抽取对象的所有特征表达，并将同义特征表达聚类。每个特征类表示了关于该对象的独一无二的某个特征
  - 特征情感分类 (aspect sentiment classification)
    - 确定观点针对每个特征的情感倾向：正面、负面、中性。



# 对象特征抽取

(Hu and Liu, KDD-04; Liu, Web Data Mining book 2007)

- **频繁特征**: 被许多评论提及的特征
- **利用序列模式挖掘**
  - 不同评论可能并不相关
  - 描述产品特征的词语比较有限
  - 主要特征通常出现比较频繁.
- **序列模式挖掘能够发现频繁短语**

# 非频繁特征抽取

- 基于: 同一情感词被用来描述不同特征与对象
  - “The pictures are absolutely **amazing**.”
  - “The software that comes with it is **amazing**.”

Frequent  
aspects

Infrequent  
aspects



Opinion words



# 利用依存关系

(Qiu et al. IJCAI-2009)

- 利用情感词与特征之间的依存关系抽取特征
  - 情感词修饰对象特征, e.g.,
  - “This camera has *long battery life*”
- 基于种子情感词集进行自举，抽取更多特征和情感词

# 依存关系规则举例

	Relations and Constraints	Output	Examples
R1 <sub>1</sub>	$O \rightarrow O\text{-Dep} \rightarrow F$ s.t. $O \in \{O\}$ , $O\text{-Dep} \in \{MR\}$ , $POS(F) \in \{NN\}$	$f = F$	The phone has a <u>good</u> "screen". $good \rightarrow mod \rightarrow screen$
R1 <sub>2</sub>	$O \rightarrow O\text{-Dep} \rightarrow H \leftarrow F\text{-Dep} \leftarrow F$ s.t. $O \in \{O\}$ , $O/F\text{-Dep} \in \{MR\}$ , $POS(F) \in \{NN\}$	$f = F$	"iPod" is the <u>best</u> mp3 player. $best \rightarrow mod \rightarrow player \leftarrow subj \leftarrow iPod$
R2 <sub>1</sub>	$O \rightarrow O\text{-Dep} \rightarrow F$ s.t. $F \in \{F\}$ , $O\text{-Dep} \in \{MR\}$ , $POS(O) \in \{JJ\}$	$o = O$	same as R1 <sub>1</sub> with <i>screen</i> as the known word and <i>good</i> as the extracted word
R2 <sub>2</sub>	$O \rightarrow O\text{-Dep} \rightarrow H \leftarrow F\text{-Dep} \leftarrow F$ s.t. $F \in \{F\}$ , $O/F\text{-Dep} \in \{MR\}$ , $POS(O) \in \{JJ\}$	$o = O$	same as R1 <sub>2</sub> with <i>iPod</i> is the known word and <i>best</i> as the extract word.
R3 <sub>1</sub>	$F_{i(j)} \rightarrow F_{i(j)}\text{-Dep} \rightarrow F_{j(i)}$ s.t. $F_{j(i)} \in \{F\}$ , $F_{i(j)}\text{-Dep} \in \{CONJ\}$ , $POS(F_{i(j)}) \in \{NN\}$	$f = F_{i(j)}$	Does the player play dvd with <u>audio</u> and "video"? $video \rightarrow conj \rightarrow audio$
R3 <sub>2</sub>	$F_i \rightarrow F_i\text{-Dep} \rightarrow H \leftarrow F_j\text{-Dep} \leftarrow F_j$ s.t. $F_i \in \{F\}$ , $F_i\text{-Dep} = F_j\text{-Dep}$ , $POS(F_j) \in \{NN\}$	$f = F_j$	Canon "G3" has a great <u>len</u> . $len \rightarrow obj \rightarrow has \leftarrow subj \leftarrow G3$
R4 <sub>1</sub>	$O_{i(j)} \rightarrow O_{i(j)}\text{-Dep} \rightarrow O_{j(i)}$ s.t. $O_{j(i)} \in \{O\}$ , $O_{i(j)}\text{-Dep} \in \{CONJ\}$ , $POS(O_{i(j)}) \in \{JJ\}$	$o = O_{i(j)}$	The camera is <u>amazing</u> and "easy" to use. $easy \rightarrow conj \rightarrow amazing$
R4 <sub>2</sub>	$O_i \rightarrow O_i\text{-Dep} \rightarrow H \leftarrow O_j\text{-Dep} \leftarrow O_j$ s.t. $O_i \in \{O\}$ , $O_i\text{-Dep} = O_j\text{-Dep}$ , $POS(O_j) \in \{JJ\}$	$o = O_j$	If you want to buy a <u>sexy</u> , "cool", accessory-available mp3 player, you can choose iPod. $sexy \rightarrow mod \rightarrow player \leftarrow mod \leftarrow cool$

# 基于序列标注模型

(Jakob and Gurevych, EMNLP 2010)

- 需要充足的人工标注数据

- (1) While none of the features are *earth-shattering*, eCircles does provide a *great* place to keep in touch.
- (2) Hyundai's *more-than-modest* refresh has largely addressed all the original car's *weaknesses* while maintaining its price competitiveness.

- 基于序列标注模型CRF

- 特征包括

- Token
  - POS
  - Short Dependency Path
  - Word Distance
  - Opinion Sentence

Table 3: Single-Domain Extraction with our CRF-based Approach

Features	movies			web-services			cars			cameras		
	Prec	Rec	F-Me	Prec	Rec	F-Me	Prec	Rec	F-Me	Prec	Rec	F-Me
tk, pos	0.639	0.133	0.220	0.500	0.051	0.093	0.438	0.110	0.175	0.300	0.085	0.127
tk, pos, wDs	0.542	0.181	0.271	0.451	0.272	0.339	0.570	0.354	0.436	0.549	0.375	0.446
tk, pos, dLn	0.777	0.481	0.595	0.634	0.380	0.475	0.603	0.372	0.460	0.569	0.376	0.453
tk, pos, sSn	0.673	0.637	0.653	0.604	0.397	0.476	0.453	0.180	0.257	0.398	0.172	0.238
tk, pos, dLn, wDs	<b>0.792</b>	0.481	0.598	0.620	0.354	0.450	0.603	0.389	0.473	0.596	0.425	0.496
tk, pos, sSn, wDs	0.662	0.656	0.659	0.664	0.461	0.544	0.564	0.370	0.446	0.544	0.381	0.447
tk, pos, sSn, dLn	0.791	0.477	0.594	0.654	0.501	0.568	0.598	0.384	0.467	0.586	0.391	0.468
tk, pos, sSn, dLn, wDs	0.749	<b>0.661</b>	<b>0.702</b>	<b>0.722</b>	<b>0.526</b>	<b>0.609</b>	<b>0.622</b>	<b>0.414</b>	<b>0.497</b>	0.614	<b>0.423</b>	<b>0.500</b>
pos, sSn, dLn, wDs	0.672	0.441	0.532	0.612	0.322	0.422	0.612	0.369	0.460	<b>0.674</b>	0.398	0.500

# 特征聚类

- Liu et al (WWW-05)使用WordNet.
- Carenini et al (K-CAP-05) 基于相似度进行聚类
  - 字符串相似度, 同义词、其他基于WordNet的距离
- (Zhai et al Coling-2010; Zhai et al WSDM-2011) 提出半监督学习方法和无监督学习方法, 结合语言学约束

# 特征情感分类

- 对于每个特征，判别观点持有者所表达的情感倾向性
- 基于句子
  - 一个句子可包含多个特征
  - 针对不同的特征可以有不同的观点
  - E.g., The **battery life** and **picture quality** are *great* (+), but the **view founder** is *small* (-).
- 几乎所有方法都利用情感词/短语
  - 一些情感词具有上下文独立的倾向性, e.g., “great” .
  - 一些情感词具有上下文相关的倾向性, e.g., “small”

# 特征情感分类

(Hu and Liu, KDD-04; Ding and Liu 2008)

- **输入:**  $(f, s)$ ,  $f$  为产品特征,  $s$  为包含 $f$ 的一个句子
- **输出:**  $s$ 中针对  $f$  的观点倾向
- **两个步骤**(Hu and Liu, KDD-04):
  - 基于转折连词切分句子(but, except that, etc).
  - 针对包含 $f$ 的句子分段 $s_f$ , 将其所有情感词的倾向性值(1, -1, 0)进行求和.
- **对于上下文相关的特征情感分类, Ding and Liu (2008)**充分利用句子之间和句子内部的连词、否定词等词汇, 加以约束
  - Negation Neg  $\rightarrow$  Positive
  - Negation Pos  $\rightarrow$  Negative
  - Desired value range  $\rightarrow$  Positive
  - Below or above the desired value range  $\rightarrow$  Negative ; ...



# 分而治之策略

- 目前大多数技术都基于一种方法解决所有情况，但是现实中有很多复杂的情况：
  - “The picture quality of this camera is great.”
  - “Sony cameras take better pictures than Nikon”.
  - “If you are looking for a camera with great picture quality, buy Sony.”
  - “If Sony makes good cameras, I will buy one.”
- Narayanan, et al (2009) 采用分而治之策略重点考察条件句

# 基于主题模型的方法

- 对PLSI/LDA进行扩展，可同时获得aspect/topic与sentiment, 例如
  - JST [Lin and He 2009]、ME-LDA [Zhao et al. 2010]

Table 3: Ex

Posit

Topic 1	
$w$	$P(w z, l)$
good	0.084708
realli	0.046559
plai	0.044174
great	0.036645
just	0.028990
perform	0.028362
nice	0.026354
fun	0.025978
lot	0.025853
act	0.022715
direct	0.021586
best	0.020331
get	0.020331
entertain	0.018198
better	0.017445
job	0.016692
talent	0.016064
pretti	0.016064
try	0.015688
want	0.015186

Service		Room Condition		Ambience		Meal		General
Aspect	Opinion	Aspect	Opinion	Aspect	Opinion	Aspect	Opinion	
staff	helpful	room	shower	room	quiet	breakfast	good	great
desk	friendly	bathroom	small	floor	open	coffee	fresh	good
hotel	front	bed	clean	hotel	small	fruit	continental	nice
english	polite	air	comfortable	noise	noisy	buffet	included	well
reception	courteous	tv	hot	street	nice	eggs	hot	excellent
help	pleasant	conditioning	large	view	top	pastries	cold	best
service	asked	water	nice	night	lovely	cheese	nice	small
concierge	good	rooms	safe	breakfast	hear	room	great	lovely
room	excellent	beds	double	room	overlooking	tea	delicious	better
restaurant	rude	bath	well	terrace	beautiful	cereal	adequate	fine

Table 5: Sample aspects and opinion words of the hotel domain using ME-LDA.

# 观点摘要

- **对观点内容进行总结和提炼，利用简洁的文本或可视化图表进行摘要呈现**
  - 一般基于观点抽取结果
- **可与观点检索相结合，对观点检索结果进行摘要**

# 观点摘要基本框架

- 步骤

- 对象特征抽取
- 特征情感分类
- 摘要与可视化

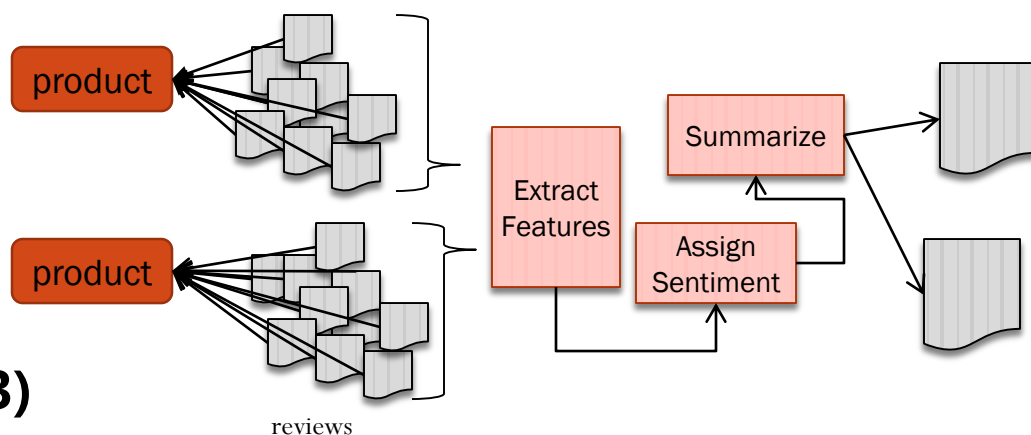
- 方法

- 统计方法

- ReviewSeer (2003)
- Opinion Observer (2004)
- Red Opal (2007)

- 基于NLP的方法

- Kanayama System (2004)
- WebFountain (2005)
- OPINE (2005)



# 基于特征的产品观点摘要

(Hu & Liu, KDD-2004) 获得KDD2015十年最具影响力论文奖

*“I bought an **iPhone** a few days ago. It was such a nice **phone**. The **touch screen** was really cool. The **voice quality** was clear too. Although the **battery life** was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too **expensive**, and wanted me to return it to the shop. ...”*

....

## Feature Based Summary:

### Feature1: **Touch screen**

Positive: 212

- The **touch screen** was really cool.
- The **touch screen** was so easy to use and can do amazing things.

...

Negative: 6

- The **screen** is easily scratched.
- I have a lot of difficulty in removing finger marks from the **touch screen**.

...

### Feature2: **battery life**

...

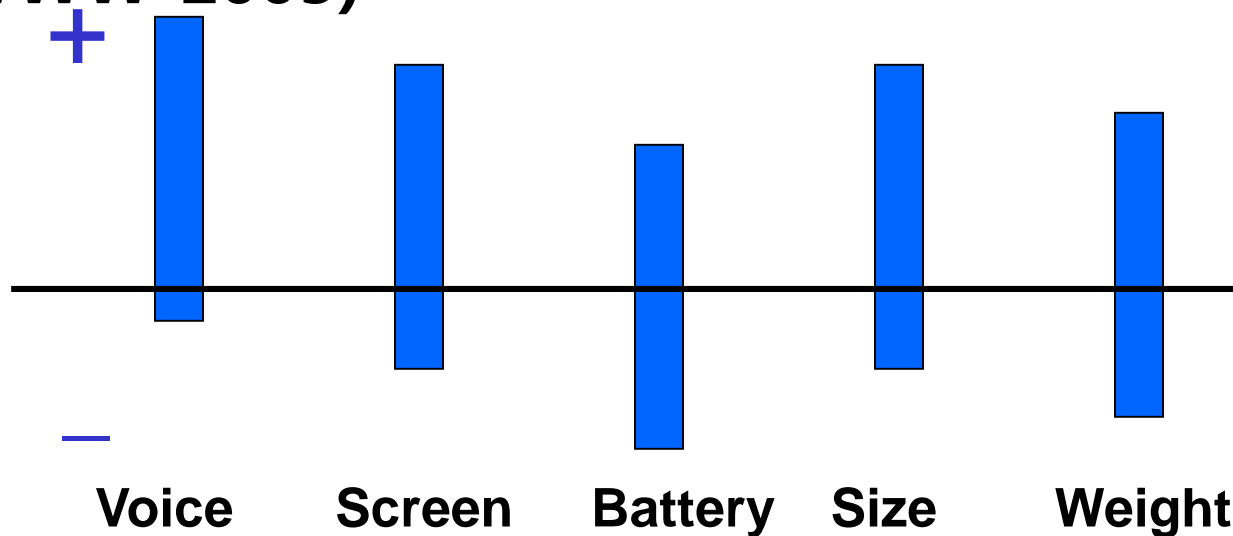
*Note: We omit opinion holders*

# 基于特征的可视化比较

(Liu et al. WWW-2005)

Summary of  
reviews of

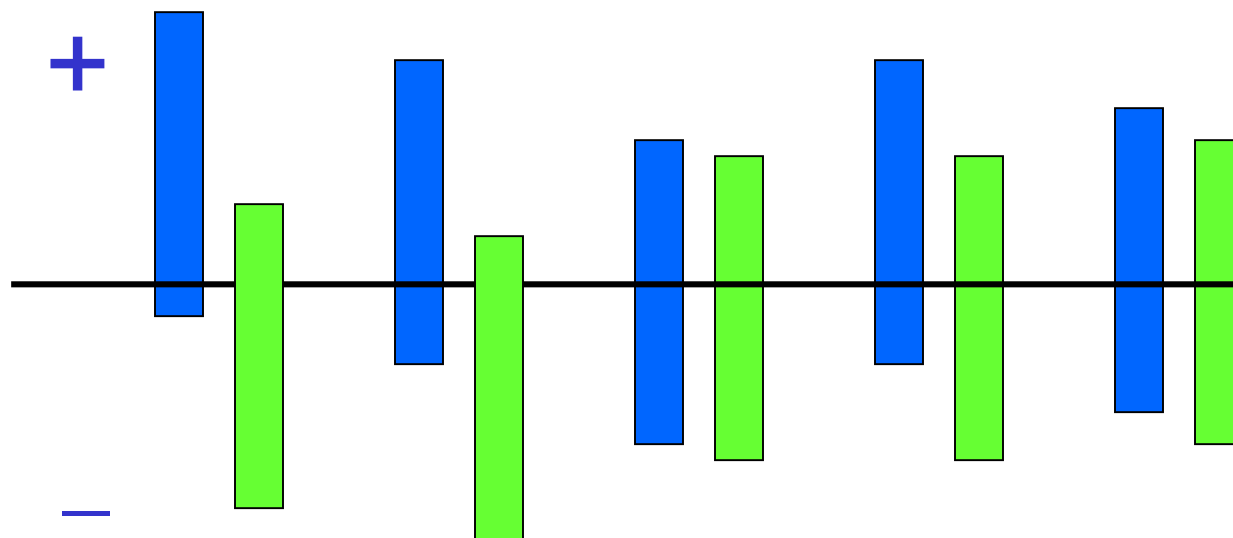
■ Cell Phone 1



Comparison of  
reviews of

■ Cell Phone 1

■ Cell Phone 2



# 内容

- 文本情感计算基础技术
  - 背景与概述
  - 情感分类
  - 观点抽取与摘要
- 面向微博的情感计算
- 总结与展望

# 概述

- 不同于产品评论，微博内容发散、表达更自由，因此情感分析难度更大
- 微博情感分析的实用价值更大
  - 除商业应用之外，还可广泛应用于各类社会科学研究
- 可用数据
  - 英文 SemEval
  - 中文 NLPCC、COAE
- 可利用的线索
  - 表情符号
  - Hashtag
  - 回复与评论
  - 社交网路关系
  - ...



# 面向Twitter的情感分类

- 倾向性分类
- 利用Distant Supervision方法 [Go et al. 2009]
  - 从Twitter上基于表情符号(emoticons)获得大量带噪音的训练数据

Table 3: List of Emoticons

Emoticons mapped to :)	Emoticons mapped to :(
:)	:(
:-)	:-( :(
: )	: (
:D	
=)	

Table 6: Classifier Accuracy

Features	Keyword	Naive Bayes	MaxEnt	SVM
Unigram	65.2	81.3	80.5	82.2
Bigram	N/A	81.6	79.1	78.8
Unigram + Bigram	N/A	82.7	83.0	81.6
Unigram + POS	N/A	79.9	79.9	81.9

# 面向Twitter的情感分类

- **NRC-Canada利用了丰富的特征，在SemEval 2013、2014上Twitter情感分类评测上取得优异成绩**  
[Mohammad et al. 2013;Zhu et al. 2014]

- Word ngrams
- Character ngrams
- All-caps
- POS
- hastags
- Lexicons
- Punctuation
- Emoticons
- Elongated words
- Clusters
- negation

	Term-level		Message-level	
	NRC13	NRC14	NRC13	NRC14
Twt14	85.19	<b>86.63(1)</b>	68.88	<b>69.85(4)</b>
Twt13	89.10	<b>90.14(1)</b>	69.02	<b>70.75(2)</b>
Sarc14	78.16	<b>77.13(3)</b>	47.64	<b>58.16(1)</b>
LvJn14	84.96	<b>85.49(2)</b>	74.01	<b>74.84(1)</b>
SMS13	88.34	<b>88.03(2)</b>	68.34	<b>70.28(1)</b>

Table 2: Overall performance of the NRC-Canada sentiment analysis systems.

# 面向Twitter的情感分类

- 利用tweet之间的关系进行查询对象相关的情感分类 [Jiang et al. 2011]
  - 相邻tweet具有一致的情感类别
    - 同一用户发布的微博
    - 微博及其转发(retweet)
    - 微博及其回复(reply)
  - 基于图进行迭代计算

$$p(c | \tau, G) = p(c | \tau) \sum_{N(d)} p(c | N(d)) p(N(d))$$

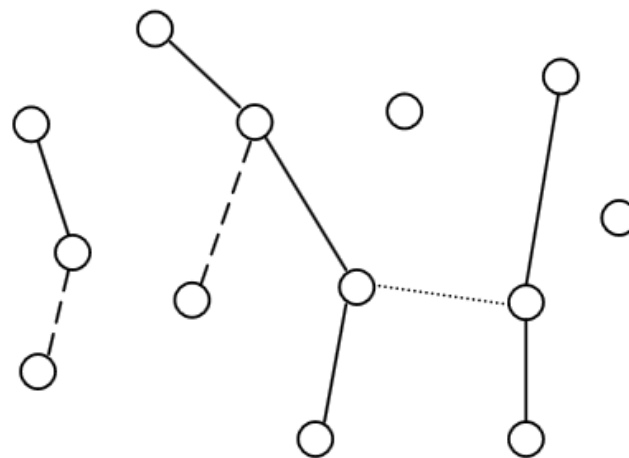


Figure 1. An example graph of tweets about a target

# 面向Twitter的情感分类

- 利用twitter用户关系进行用户级别的情感分类 [Tan et al. 2011]
  - 相互关联的用户有相似的观点
    - 关注关系 follower/followee
    - 提及关系 @
  - 基于Factor Graph Model

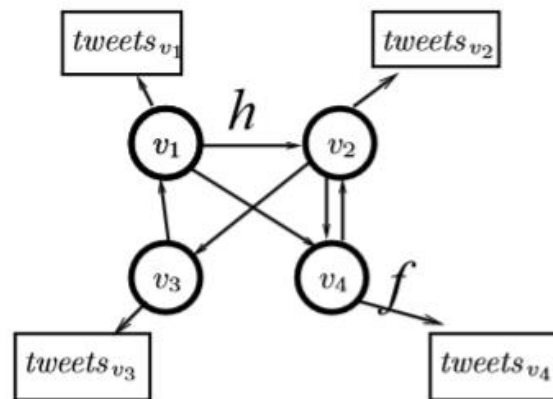


Figure 3: Example directed heterogeneous graph (dependence on topic  $q$  suppressed for clarity). The corresponding factor graph has factors corresponding to user-tweet dependencies (label “f”) and user-user dependencies (label “h”).

$$\log P(\mathbf{Y}) = \left( \sum_{v_i \in V} \left[ \sum_{t \in \text{tweets}_{v_i}, k, \ell} \mu_{k, \ell} f_{k, \ell}(y_{v_i}, \hat{y}_t) \right. \right. \\ \left. \left. + \sum_{v_j \in \text{Neighbors}_{v_i}, k, \ell} \lambda_{k, \ell} h_{k, \ell}(y_{v_i}, y_{v_j}) \right] \right) \\ - \log Z,$$

# 中文微博上的观点对象抽取

- 微博上的观点对象具有发散性，新词较多，难以识别
- 充分利用微博特性[Zhou et al. 2013]
  - Hashtag信息帮助观点对象识别
  - 同一话题下相似微博具有相似的观点对象

Topic	Sentence
#官员财产公示# #PropertyPublicityofGovernmentOfficials	1. 纯属作秀！ (Just for show! )
	2. 财产公示在中国就是作秀。 (Property publicity is just a show in China.)

# 中文微博上的观点对象抽取

- 充分利用微博特性[Zhou et al. 2013]
  - 步骤：
    - Hashtag切割，将切割短语加入用户词典帮助微博分词
      - #90后 暴打 老人#
    - 候选对象抽取与赋值
      - $(noun \mid adj)(noun \ adj \mid \text{的}) * noun.$
    - 构建无向图，基于无监督的标签传递方法进行迭代计算

# Twitter中基于话题的观点摘要

## [Meng et al. 2012]

- 利用#Hashtag聚类进行话题提取与表示

Topic label	Opinion Summaries	Sentiment
#libya	<b>I think</b> it is time to say that Barack Obama deserves credit for backing up his words with action on #Libya despite domestic opposition.	Positive
	<b>Thanks to</b> Obama's approach to Libya, not 1 American soldier was lost.	Positive
	Film <b>comparing</b> Obama's praise of public protests in Libya/Syria <b>with</b> the violence of arrests on Occupy Wall St	Negative
#occupywallst	@BarackObama I'm running out of hope. <b>Please</b> replace Geithner w/ Reich or Krugman #wallstreetoutofwhitehouse #OWS.	Negative
	@BarackObama : <b>Please</b> recognize the men and women who are occupying wall street.	Negative
	Obama suggests MLK Jr. <b>would have</b> backed #occupywallstreet.	Positive
#obamacar	<b>We need to</b> completely repeal #Obamacare and start by replacing it with HR 3400! #cnndebate	Negative
	<b>If</b> ObamaCare is not repealed <b>then</b> we can <b>expect</b> stagnant growth, long term unemployment and record high premiums.	Negative
	Op-ed: <b>Despite</b> conservatives' claims, #Obamacare is having little impact on hiring, writes Dean Baker <a href="http://t.co/I4pCBEOA">http://t.co/I4pCBEOA</a> #p2	Negative
#jobsnow	<b>Thanks to</b> @BarackObama's efforts, 270 businesses have committed over 25,000 jobs to American veterans.	Positive
	Cain creates 11 more jobs in 1 day <b>than</b> Obama in a lifetime.	Negative
	Let's be clear, the US economy is horrible <b>because of</b> Obama's policies.	Negative

Table 12: A case study of the opinion summary for "Obama"

# 内容

- 文本情感计算基础技术
  - 背景与概述
  - 情感分类
  - 观点抽取与摘要
- 面向微博的情感计算
- 总结与展望



# 总结与展望

- **两大任务**
  - 情感分类
  - 观点抽取
- **方法**
  - 规则
  - 机器学习
  - 深度学习
- **领域不同，难度不一样**
  - 产品评论、电影评论、微博
- **社交媒体上的情感计算具有重要价值**
  - 目前的水平还远未达到预期，任重而道远
  - 只依靠深度学习并不能完全解决问题
- **从情感分析到立场分析、论辩计算、情感对话**

# 第二次作业

- **Tweet情感分类: SemEval2017 Task 4 Subtask A**
  - **Message Polarity Classification:** Given a message, classify whether the message is of **positive**, **negative**, or **neutral** sentiment.
- **数据 (训练+测试; 只做英文)**
  - <http://alt.qcri.org/semeval2017/task4/index.php?id=results>
- **提交压缩文件包到 sckr2018@126.com**
  - 姓名、学号
  - 自己编写的代码 (调用的大型工具包不用提交, 只需在报告中说明即可), 编程语言不限
  - 2~3页小报告 (包括调用的工具或资源, 实验方法、结果比较与分析、想法等)
- **提交时间4月22日**

# 参考文献

- References before and in 2011 are available at <http://www.cs.pitt.edu/~wiebe/subjectivityBib.html>
- Socher, Richard, et al. Recursive deep models for semantic compositionality over a sentiment treebank. **EMNLP 2013**.
- Blunsom, Phil, Edward Grefenstette, and Nal Kalchbrenner. A convolutional neural network for modelling sentences. **ACL 2014**.
- Yoon Kim. Convolutional Neural Networks for Sentence Classification. **EMNLP 2014**.
- Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. **ICML 2015**.
- Chenghua Lin and Yulan He. Joint Sentiment/Topic Model for Sentiment Analysis. **CIKM 2009**.
- Xin Zhao, et al. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. **EMNLP 2010**.

# 参考文献

- Xinfan Meng, et al. Entity-centric topic-oriented opinion summarization in Twitter. **KDD2012** (industrial track).
- Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.
- Saif M. Mohammad, et al. NRC-Canada: Building the state-of-the-art in sentiment analysis of Tweets. In **SemEval 2013**.
- Xiaodan Zhu, et al. NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. In **SemEval 2014**.
- Long Jiang, et al. Target-dependent Twitter Sentiment Classification. In **ACL 2011**.
- Chenhao Tan, et al. User-level sentiment analysis incorporating social networks. **KDD 2011**.
- Prettenhofer, Peter, and Benno Stein. Cross-lingual adaptation using structural correspondence learning. ACM Transactions on Intelligent Systems and Technology (**TIST**) 3.1 (2011): 13.
- H. Zhou, L. Chen, F. Shi, D. Huang. Learning bilingual sentiment word embeddings for cross-language sentiment classification. **ACL 2015**.
- Mariana S. C. Almeida et al. Aligning opinions: cross-lingual opinion mining with dependencies. **ACL 2015**.

# 参考文献

- Wan, Xiaojun. Co-training for cross-lingual sentiment classification. **ACL 2009**.
- Ma, Tengfei, Xiaojun Wan. Opinion target extraction in Chinese news comments, **COLING 2010**.
- Zhou, Xinjie, Wan, Xiaojun Wan and Jianguo Xiao. Collective opinion target extraction in Chinese microblogs. **EMNLP 2013**.
- Wen, Shiyang, and Xiaojun Wan. Emotion Classification in Microblog Texts Using Class Sequential Rules. **AAAI 2014**.
- Zhou, Xinjie, Xiaojun Wan, and Jianguo Xiao. Representation Learning for Aspect Category Detection in Online Reviews. In **AAAI 2015**.
- Zhou, Xinjie, Xiaojun Wan, and Jianguo Xiao. CLOpinionMiner: Opinion Target Extraction in a Cross-Language Scenario. , **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, 23.4 (2015): 619-630.

- **Some slides were taken or adapted from related slides written by Dongjoo Lee, Bing Liu, Liqiang Guo, Shima Gerani, Mark Carman, Fabio Crestani, Rada Mihalcea, Carmen Banea, Janyce Wiebe, etc. Thank them for sharing their slides.**

