

第二课

观点挖掘与倾向性分析

中国科学院自动化研究所
模式识别国家重点实验室

目录

■ 第一部分：

- 我们为什么需要观点挖掘与倾向性分析？
- 什么是观点挖掘与倾向性分析？

■ 第二部分：

- 如何进行观点挖掘与倾向性分析？
 - 任务、方法、资源、评测

■ 第三部分：

- 问题与挑战

为什么需要

- 文本信息主要包含两类

- 客观性事实(Facts)
- 主观性观点(Opinions)



- 随着Web2.0的飞速发展以及Web3.0的兴趣，互联网中出现大量的UGC数据，其中包含了大量的观点信息

- 博客、微博、商品评论、论坛....

- 已有文本分析方法主要侧重于客观性文本内容(factual information)的分析和挖掘

有什么用

■ 企业对观点挖掘和倾向性分析的需求

- ❑ Automatically find consumer sentiments and opinions (market intelligence)
- ❑ Capture public trends
- ❑ Capture commercial opportunity
- ❑ Online reputation management
- ❑ Precision Advertising



■ 普通用户对观点挖掘和倾向性分析的需求


- ❑ Helpful for purchasing a product
- ❑ Find opinions on political topics

■ 政府对观点挖掘和倾向性分析的需求

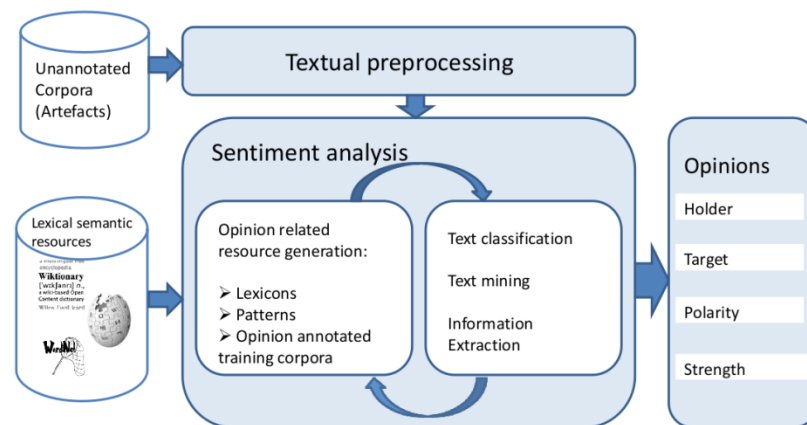
- ❑ Control the public opinions
- ❑ Monitor the public event



定义

- **观点**：人们对事物的看法，具有明显的主观性，不同人对同一事物的看法存在差异
- **倾向性**：观点中所包含的情感倾向性 
- **观点挖掘与倾向性分析**：从海量数据中挖掘观点信息，并分析观点信息的倾向性
 - 非结构化→结构化

Sentiment analysis or opinion mining (in Wikipedia) refers to a broad area of natural language processing, computational linguistics and text mining. Generally speaking, **it aims to determine the attitude of a speaker or a writer with respect to some topic.**



例子

“我今年天让入手诺基亚5800，把玩不到24小时，目前感觉5800屏幕很好，操作也很方便，通话质量也不错，但是外形有些偏女性化，不适合男生。这些都是小问题，最主要的问题是电池不耐用，只能坚持一天，反正我觉得对不起这个价格。”



- 外形
- 电池



- 屏幕
- 操作
- 通话质量



观点挖掘与倾向性分析相关任务

- 观点及倾向性识别
 - Sentiment Identification
- 观点要素抽取
 - Opinion Attribute Extraction
 - Opinion Summarization
- 观点检索
 - Opinion Retrieval

Sentiment Identification(1/2)

- Opinion Identification (Subjective/Objective)
 - ❑ 中美两方的代表就朝鲜核问题进行了磋商。(Objective)
 - ❑ 中方发言人就美国近期对阿富汗的行动进行了强烈的谴责 (Subjective)
- Polarity Classification (Positive/Negative/Neutral)
 - ❑ 这家餐厅总体来说还可以。(Neutral)
 - ❑ 但是价格偏贵，人均消费100块。(Negative)
 - ❑ 抛开价格的因素还是很不错的。(Positive)
- Strength Rating (Sentiment Strength Identification)
 - ❑ iPhone 的价格太贵了。(Strong against)
 - ❑ iPhone 的价格有点贵。(Something to be bad)



Sentiment Identification(2/2)

- Word Level
 - 识别一个词的倾向性
- Feature Level (Aspect Level)
 - 识别一个Aspect的倾向性
 - “这家餐厅价格偏贵，人均消费100块”→ 价格
- Sentence Level
 - 识别一个句子的观点倾向性
- Document Level
 - 识别一篇文本（包含多个句子）整体的倾向性

Opinion Attribute Extraction

■ Opinion Holder Extraction

- “中方发言人”就美国近期对阿富汗的行动进行了强烈的谴责”
 - 在新闻语料中大量出现，通常为命名实体、名词性短语或者术语
 - 在商品评论文本中很少出现

■ Opinion Target Extraction

- “中方发言人就美国近期对阿富汗的行动进行了强烈的谴责”
- “这款手机的屏幕太小，分辨率不足”
- 术语、事件、实体等

Opinion Summarization

*“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. ...”*

....

Opinion Summary:

Feature 1: **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**.

...

Feature 2: **battery life**

...

Opinion Retrieval

- 根据用户的查询从文档中找出对于主题信息发表了观点的文档
 - 主题相关并且具有主观倾向性
 - Blog Search, Twitter, Forum.....

应用

新浪首页 × 转贴 - 开心网 × iphone - Goo... × iphone - Goo... × Apple iPhon... × http://www.g... × iphone - 必... × 中国科学院自... ×

← → ↻ ↺ www.google.com/products/catalog?q=iphone&hl=en&newwindow=1&biw=1280&bih=709&prmd=snl&prmdo=1&um=1&ie=UTF-8&cid=67070067044147939568 ☆


Web Images Videos Maps News Shopping Gmail more ▼ liuk.ia.ac@gmail.com | My Shopping List | Settings ▼ | Sign out

Google products

iphone Search Products

Apple iPhone 8 GB (first generation) (AT&T - GSM) from Apple in Mobile Phones

[Overview](#) - [Compare prices](#) - [Reviews](#) - [Technical specifications](#) - [Similar items](#) - [Accessories](#)

 **\$500 new, \$180 used** from [4 sellers](#)

★★★★☆ 1,784 reviews

Reviews

Summary - Based on 1,784 reviews

1 2 3 4 stars 5 stars

"Easy texting and buttons are good size to heat." "Awful speaker though."
"The photos have good clarity and is very handy." "With its touch screen it makes life so much easier."
"You can download music, videos, and movies (love)" "The worst thing about the iphone is the price."

Apple iPhone - 8GB (AT&T) Review

★★★★☆ By Kent German - Jun 30, 2007 - Editorial review - CNET

Pros: The Apple iPhone has a stunning display, a sleek design, and an innovative multitouch user interface. Its Safari browser makes for a superb Web surfing experience, and it offers easy-to-use apps. As an iPod, it shines.

Cons: The Apple iPhone has variable call quality and lacks some basic features found in many cell phones, including stereo Bluetooth support and 3G compatibility. Integrated memory is stingy for an iPod, and you have to sync the iPhone to manage music content.

Bottom Line: Despite some important missing features, a slow data network, and call quality that doesn't always deliver, the Apple iPhone sets a new benchmark for an integrated cell phone and MP3 player.

Editor's note (7/11/2008): This is the review of the original first-generation iPhone model, released June 2007. Coverage of the 3G iPhone model released July 11, 2008 is available [here](#).

Photo gallery: Apple iPhone

Editor's note: Apple eliminated the 4GB model on September 5, 2007, two months after the iPhone's initial

Show reviews that mention

[Features](#) (946)
[Design](#) (597)
[Camera](#) (486)
[Screen](#) (396)
[Battery](#) (366)
[Music](#) (347)
[Video](#) (288)

Show reviews by source

[Editorial reviews](#) (5)
[User reviews](#) (1,779)

[Amazon.com](#) (69)
[CNET](#) (2)
[DigitalTrends.com](#) (1)
[Epinions](#) (103)
[PriceGrabber.com](#) (15)
[Viewpoints](#) (1592)
[Wired](#) (2)

Sort reviews

开始 收件箱... Apple iP... 新-搭隔... v_liukan... Microsoft... Windows... 我的文档 周报模板... 10:12

应用

ELECTRONICS > CELL PHONES & PLANS



Apple iPhone 3GS 16GB - smartphone - WCDMA (UMTS) / GSM

\$449 and up (6 stores)

★★★★★ [User reviews](#) (285)

★★★★★ [Expert reviews](#) (1)

SHARE [Facebook](#) [Twitter](#) [Messenger](#) [Email](#)

[See larger photo](#)

See also: [Product Summary](#) · [Where to Buy](#) · [User Reviews](#) · [Expert Reviews](#) · [Specifications](#)

USER REVIEWS



SCORECARD: EASE OF USE ([See all](#))

87 positive reviews | no negative comments

Email: Emailing is very basic, easy to use and works brilliantly.

★★★★★ [KavlieLandymore](#) · 10/6/2010 · [www.ciao.co.uk](#) [see all](#)

Pros: good maps, memory, quick, easy to use, access to loads of different features

★★★★★ [anklechris](#) · 9/30/2010 · [www.ciao.co.uk](#) [see all](#)

Pros: Has everything, brilliant music player and camera, great for games, easy to use.

★★★★★ [doodles12](#) · 5/24/2010 · [www.ciao.co.uk](#) [see all](#)

Safari: This is the Iphone Web browser, its very easy to use, very quick and you cant go wrong with it, it has a "Bookmarks" Section so can store all your favorites websites in specific folders, to can navigate forwards to backwards on the web...

★★★★★ [alanh087](#) · 8/28/2010 · [www.ciao.co.uk](#) [see all](#)

The phone is very easy to use, everything is very intuitive.

★★★★★ [Melski1979](#) · 9/20/2010 · [www.ciao.co.uk](#) [see all](#)

SCORECARD

Ease of use(87)

Camera(86)

Screen(55)

Wireless Interface(44)

Music Player(44)

Functionality(44)

Power Supply(40)

Durability(31)

Reception(25)

应用

Source: [this](#)



TRACKING OPINIONS ON TWITTER

twitrratr

SEARCH

SEARCHED TERM

iphone

POSITIVE TWEETS

2775

NEUTRAL TWEETS

19720

NEGATIVE TWEETS

846

TOTAL TWEETS

23341

11.89% POSITIVE



@schwa now there's a blast from the past. but it occurs to me that gliderpro would make a **great** iphone app. ([view](#))



alas **fair** iphone, you served me well and will be missed. ([view](#))



@mikediliberto @downtownrob @mitchwagner **funny** that i ended un following smoke signals as

84.49% NEUTRAL



view from the iPhone:
<http://www.floodgap.com/iv/197>
([view](#))



That's "Memphis" Taproom.
Goddamn iPhone. ([view](#))



@mothermusings This is the iPhone thingie, huh? Sooooo sorry! ([view](#))

3.62% NEGATIVE



@mikef1182 as **bad** as exchange on the iphone? ([view](#))



<http://twitpic.com/i0se> - iphone typing auto-correct changes 'just sayin' to 'just satin' - **wrong** msg indeed! ([view](#))




iphone applications don't whine about being left outside or going **hungry** or manual labor or using

应用


← → ↻ ↗ www.tweetfeel.com/?x=22&y=30#Kate_Middleton

Tweetfeel Biz | FAQ | Contact Us | Biz Login

tweetfeel


||  Searching and Analyzing

Try some Twitter trends: [Micky Christmas](#) [Eid Mubarak](#) [Kate Middleton](#) [Prince William](#) [Sharpe](#) [Playing God](#) [SARESP](#)



22



3

= 88%


 Style watch: **Kate Middleton** rocks the royal scene: Speaking of giant rocks...are they or are they not? Rumors ar...
<http://bit.ly/citOrP>


 Blech. I don't like **Kate Middleton**. #notjealousatall #coughcough


 **Kate Middleton** Rocks.

 @ambergunn "I love **Kate Middleton**'s hair. I want mine cut like that." @audreyrella "I bet no one said that about Diana's hair." #badhairday

 Prince William & **Kate Middleton** ftw !!! ^_^

Read our FAQ | Subscribe to our API | Legal Stuff | 100% Guarantee |  Share

 Follow us
 Email us

Brought to you by  conversion

Powered by  Twitter

目录

■ 第一部分：

- 我们为什么需要观点挖掘与倾向性分析？
- 什么是观点挖掘与倾向性分析？

■ 第二部分：

- 如何进行观点挖掘与倾向性分析？
 - 任务、方法、资源、评测

■ 第三部分：

- 存在的问题以及面临的挑战

内容

- Sentiment Identification
 - Opinion Mining
 - Opinion Retrieval
 - Resources and Evaluations
-

Sentiment Identification

- Word Level
 - Sentence Level
 - Document Level
 - Others
-

Word Level Sentiment Identification

■ 任务：

- 识别词语的情感倾向性，构建词典资源

■ 方法：

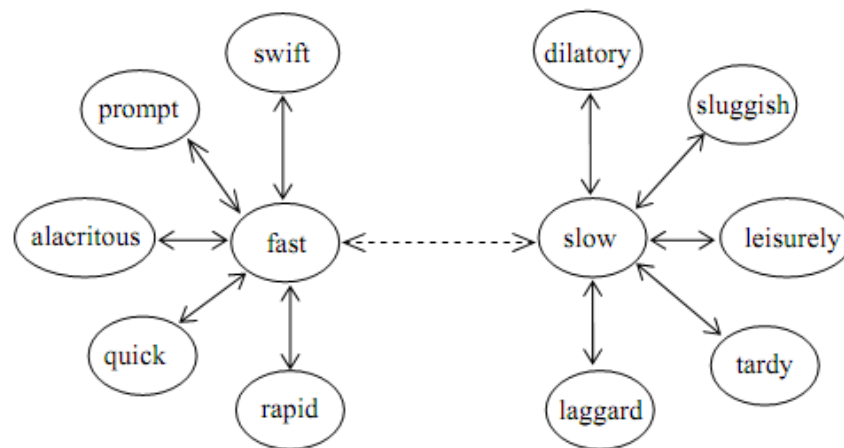
- 基本思路： 利用词之间的相似度进行扩展
- Dictionary-based approaches
- Corpus-based approaches

Dictionary-based Approaches (1/2)

■ Hu (KDD 2004)

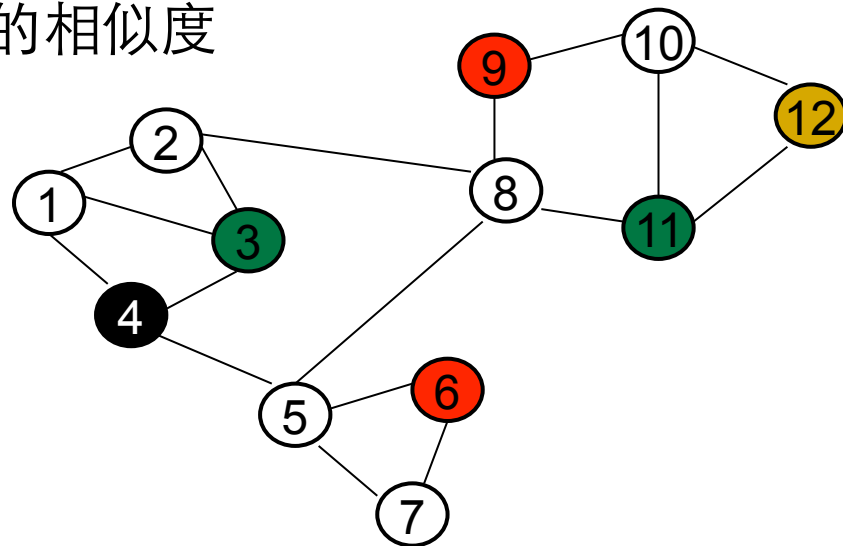
- 利用词与词之间在WordNet中的同义、反义关系对于情感词典进行扩展

```
1. Procedure OrientationSearch(adjective_list, seed_list)
2. begin
3.   for each adjective  $w_i$  in adjective_list
4.     begin
5.       if ( $w_i$  has synonym  $s$  in seed_list)
6.         {  $w_i$ 's orientation =  $s$ 's orientation;
7.           add  $w_i$  with orientation to seed_list; }
8.       else if ( $w_i$  has antonym  $a$  in seed_list)
9.         {  $w_i$ 's orientation = opposite orientation of  $a$ 's
              orientation;
10.          add  $w_i$  with orientation to seed_list; }
11.     endfor;
12. end
```



Dictionary-based Approaches (2/2)

- Hassan (ACL 2010), Kamps (LREC 2004)
 - 利用WordNet计算词之间的相似度，识别词的情感倾向性
 - 根据WordNet，计算词之间的相似度，建立词之间的语义图，边上的权重表示词之间的相似度
 - 利用图算法识别词的倾向性
 - Random Walk (Hassan)
 - Shortest Distance (Kamps)



Corpus-based Approaches (1/2)

■ Turney (ACL 2002)

- 利用网络资源计算两个词之间的相关度（互信息）
- 利用相关度识别词语的情感倾向性
- 使用 ‘Near’ 算子 (AltaVista)

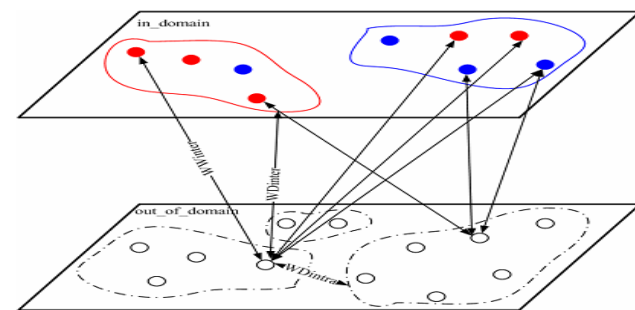
$$\text{PMI}(\text{word}_1, \text{word}_2) = \log_2 \left(\frac{\frac{1}{N} \text{hits}(\text{word}_1 \text{ NEAR } \text{word}_2)}{\frac{1}{N} \text{hits}(\text{word}_1) \frac{1}{N} \text{hits}(\text{word}_2)} \right)$$

$$\text{SO}(\text{phrase}) = \text{PMI}(\text{phrase}, \text{“excellent”}) \\ - \text{PMI}(\text{phrase}, \text{“poor”})$$

Corpus-based Approaches (2/2)

- 建立领域情感词典 (Du WSDM 2010)
 - 不同领域具有不同的领域情感词
 - 缺乏目标领域训练语料，利用其他领域的标注语料，一个领域迁移的问题
 - 不仅仅考虑词与词之间的关系
 - Word-Doci relation, Word-Doco relation
 - 利用Information bottleneck method (co-clustering)
 - 对于文档、词同时进行聚类

$$I(D_o; W_o) - I(\hat{D}_o; \hat{W}_o) + \alpha \cdot \left[\left(I(D_i; W_o) - I(D_i; \hat{W}_o) \right) + \left(I(W_i; W_o) - I(W_i; \hat{W}_o) \right) \right]$$



小结

- 基本思路：利用词之间的相似度对于情感词典进行扩展 (Dictionary-based, Corpus-based)
- Pros:
 - 模型直观，易于计算
- Cons:
 - 利用词典或者大规模语料方法计算词之间相似性易产生噪音
 - 部分词语的倾向性与上下文相关，与主题相关
 - 屏幕大
 - 体积太大
 - 大部分方法只计算了形容词的倾向性，忽略了动词、形容词、名词以及网络用语的情感倾向性
 - 小瘪三！
 - 做人不能CCTV

Sentiment Identification

- Word Level
 - Sentence Level
 - Document Level
 - Others
-

Sentence Level Sentiment Identification

- 任务：识别句子的情感倾向性
 - “7.23动车追尾事故给铁道部一记响亮的耳光。”
- 关键问题
 - 如何进行特征表示
- 分类：
 - Corpus-based approaches
 - Lexicon-based approaches
 - Combined approaches



与传统文本分类的区别

- Topic-based text categorization
 - 侧重于主题词特征
 - “这款手机的屏幕太大了” (科技、手机)
- Sentiment classification
 - 表示倾向性的词语更加重要.
 - “这款手机的屏幕好大了” (主观、褒义)

Corpus-based Approaches: 特征选择 (1/2)

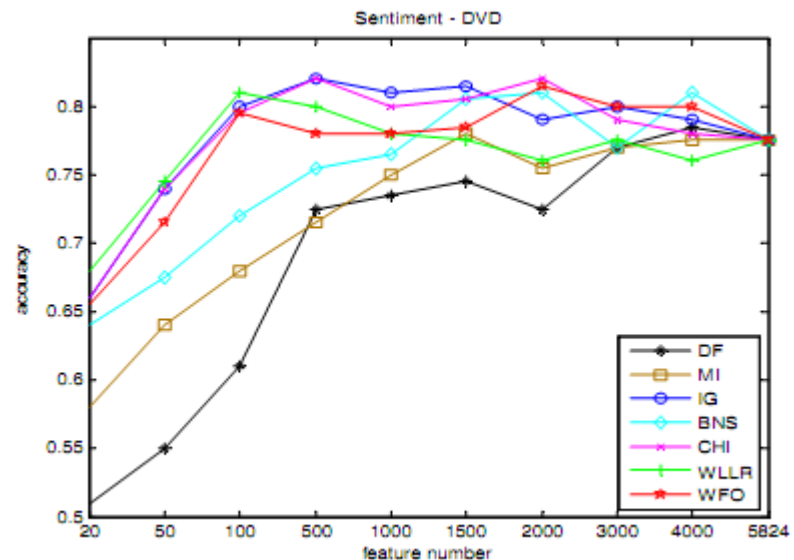
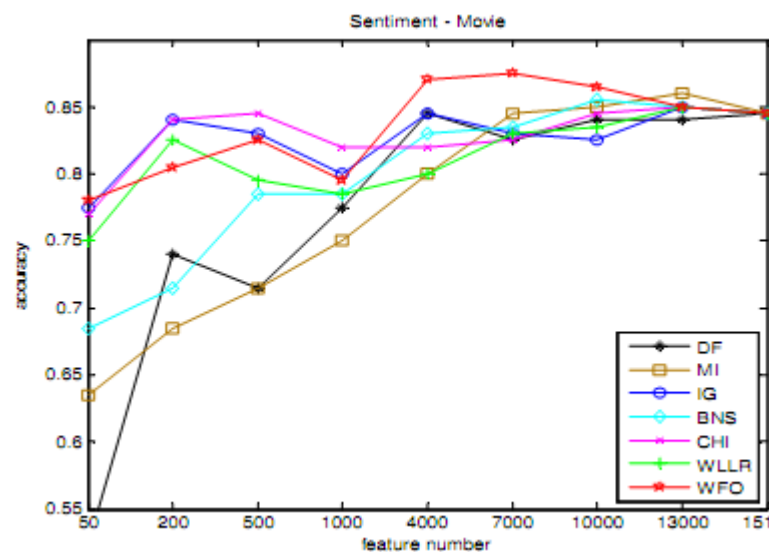
- 利用传统文本分类方法处理情感分类任务 (Pang EMNLP 2002)
 - 比较多种特征的效果
 - Unigram、bigram、POS、Adj.、Position
 - 比较多个分类器性能
 - SVM、Naïve Bayes、Maximum Entropy

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

Corpus-based Approaches: 特征选择

(2/2)

- 比较各种特征选择方法在情感分类中的效果 (Li ACL 2009)
- DF、MI、IG、CHI、BNS、WLLR、WFO



Corpus-based Approaches: Polarity Shift (1/2)

■ Polarity Shift

- 多样语言现象造成的句子内部词的倾向性转移
 - “整个店面的装修不是很漂亮”
 - 在这种情况下，如何减少学习错误？
- 方法
 - 在句子中检测出Polarity Shift
 - 判别句子倾向性时对于Polarity Shift专门处理

Corpus-based Approaches: Polarity Shift (2/2)

■ Polarity Shift的检测

□ 利用上下文信息



□ 词典信息 (Ikeka IJCNLP 2008)

□ 特征选择 (Li Coling 2010)

Corpus-based Approaches: 上下文影响 (1/2)

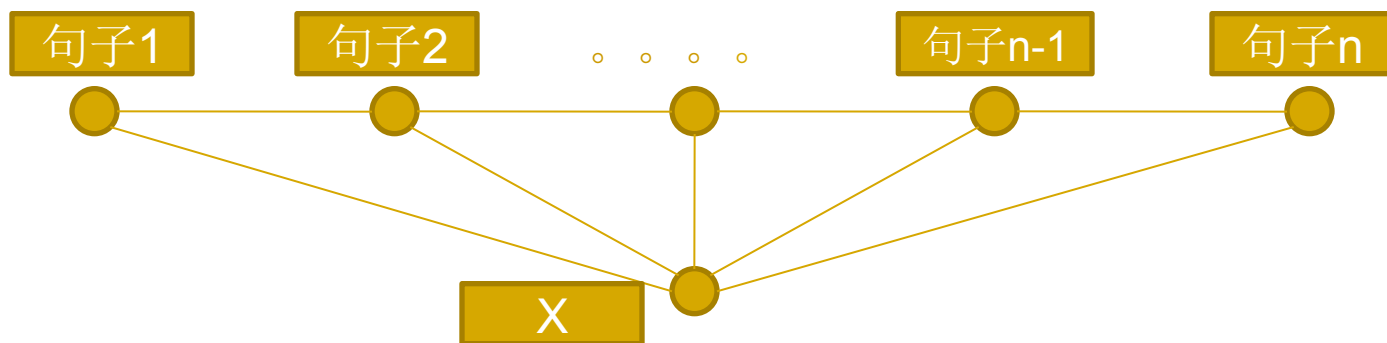
■ 上下文的影响

“1) 这是一个挺不错的电影院。2) 因为优惠很多，来的人还是比较多的，于是带起了时代广场地下一层的餐饮。3) 虽然硬件条件虽说赶不上星美，但也服务是不错的了。4) 同时看电影院周围有商场，电影开演之前可以逛逛商场。5) 总之，这里已经成为我和老公的定点影院了。”

- 句子的倾向性与句子所在上下文密切相关
- 分类任务-> 序列标注任务

Corpus-based Approaches: 上下文影响 (2/2)

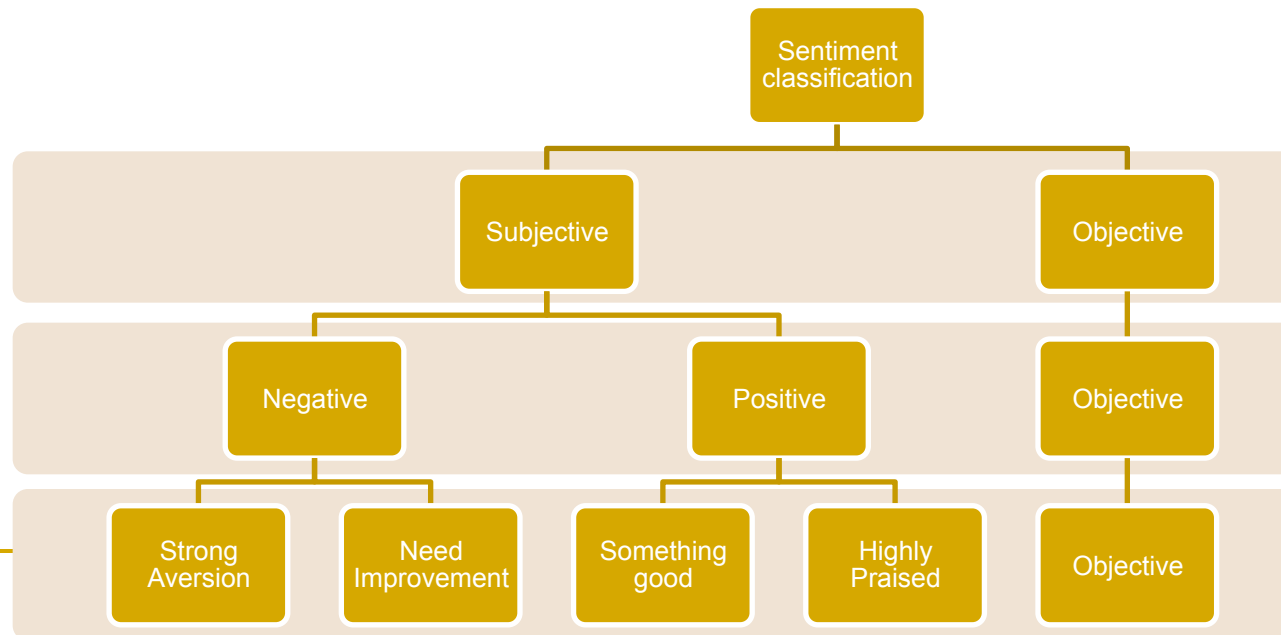
- Yi (ICML 2006) and Zhao (EMNLP 2008)
 - 将篇章中每个句子看作是一个序列上的点
 - 利用CRFs进行学习和标注



Corpus-based Approaches: 上下文+标记间冗余关系

■ Zhao (EMNLP 2008)

- 情感倾向性标记之间具有冗余关系
- 多任务联合处理
 - 主客观分类、褒贬分类、强度分类



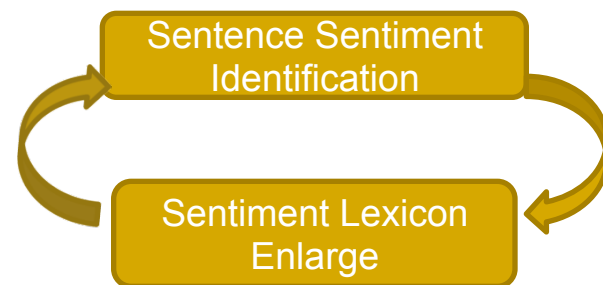
Label	NB	SVM	MaxEnt	Standard CRF	Cascaded CRF	Our Method
PP	0.1745	0.2219	0.2055	0.2027	0.2575	0.2167
P	0.2049	0.2877	0.2353	0.2536	0.2881	0.3784
Neu	0.8083	0.8685	0.8161	0.8273	0.8554	0.8269
N	0.2636	0.3014	0.2558	0.2981	0.3092	0.4204
NN	0.0976	0.1162	0.1148	0.1379	0.1510	0.2967
Total	0.6442	0.6786	0.6652	0.6856	0.7153	0.7521

Label	NB	SVM	MaxEnt	Standard CRF	Cascaded-CRF	Our Method
Pos	0.4218	0.4743	0.4599	0.4405	0.5122	0.6008
Neu	0.8147	0.8375	0.8424	0.8260	0.8545	0.8269
Neg	0.3217	0.3632	0.2739	0.3991	0.4067	0.5481
Total	0.7054	0.7322	0.7318	0.7327	0.7694	0.7855

Label	NB	SVM	MaxEnt	Standard CRF	Our Method
Subjective	0.4743	0.5847	0.4872	0.5594	0.6764
Objective	0.8170	0.8248	0.8212	0.8312	0.8269
Total	0.7238	0.7536	0.7518	0.7561	0.8018

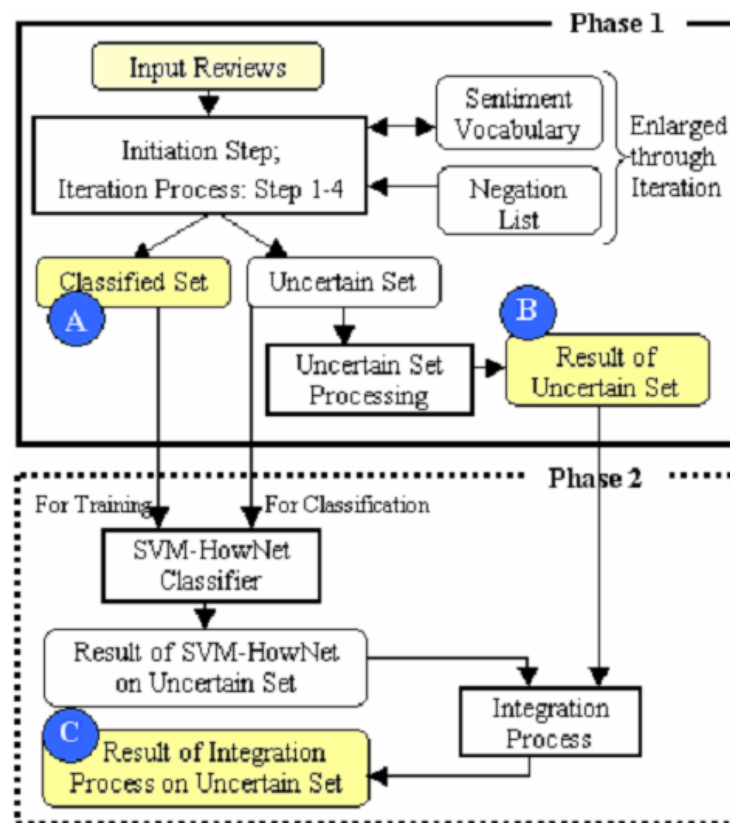
Lexicon-based Approach (1/2)

- 利用句子中词的倾向性来确定句子的倾向性
 - 关键问题：词的倾向性识别
- Turney (ACL 2002)
 - Step1: POS and select sentiment phase by patterns
 - Step2: Use PMI to compute the phase sentiments
 - Step3: Compute average sentiment of all phases in a sentence
 - Car: 84%, Banks 80%, Movies 65.83%, Travel 70.53%
- Taras (COLING 2008)
 - 句子、词的情感倾向性联合识别



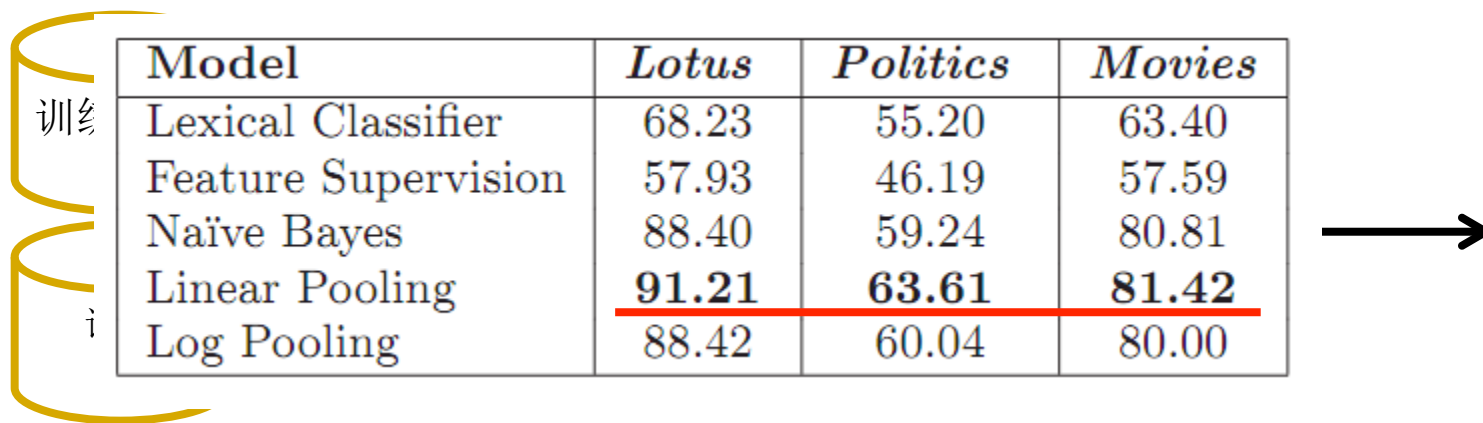
Lexicon-based Approach (2/2)

- 自学习方法 Qiu (CIKM 2009)
 - 利用词典信息产生初始标注
 - 利用置信度高的样本作为训练集，训练分类器
 - 利用启发式规则对于多个分类器进行集成



Combined Approaches (1/2)

- 利用词典信息对于结果进行Refine，主要解决训练语料不足的问题
 - 分类器集成 (Melville KDD 2009)
 - 分别用语料和词典训练两个NB分类器
 - 对于分类器进行集成



Model	<i>Lotus</i>	<i>Politics</i>	<i>Movies</i>
Lexical Classifier	68.23	55.20	63.40
Feature Supervision	57.93	46.19	57.59
Naïve Bayes	88.40	59.24	80.81
Linear Pooling	91.21	63.61	81.42
Log Pooling	88.42	60.04	80.00

Combined Approaches (2/2)

■ Semi-supervised Clustering (Li ACL 2009)

- 建立文档与词的共现矩阵
- 训练Matrix Factorization Model (cluster-based learning approach)

- 利用少量的标注语料以及词典的先验知识，同时对未标注样本进行标注

Term-Doc Matrix

$$\min_{F,G,S} \|X - FSG^T\|^2 + \alpha \text{Tr}[(F - F_0)^T C_1 (F - F_0)]$$

F: Term-Class Matrix
G: Doc-Class Matrix
S: Condensed View of X

Prior Knowledge in Lexicon

Prior Knowledge in Labeled Data

只用词典信息

使用少量标注语料以及词典信息

小结

■ Corpus-based VS. Lexicon-based

- 基于训练语料的监督学习方法受到领域限制，需要对于每个领域都进行人工训练语料的标注
- 基于词典的无监督方法具有领域独立性，但是缺乏领域词典，因此效果不如监督学习的方法
- 结合两方面的优势

■ 结合句子现象，还有很多问题需要处理

- 比较句
 - 诺基亚5800比5230更超值
- 否定词
- ...

Sentiment Identification

- Word Level
 - Sentence Level
 - Document Level
 - Others
-

Document Level Sentiment Identification

- 任务：识别篇章整体观点倾向性

诺基亚5800屏幕很好，操作也很方便，通话质量也不错，但是外形偏女性化，而且电池不耐用，只能坚持一天，价格也偏贵，反正我觉得不值。

- 绝大多数方法与句子级别方法类似

- 特征+分类器

- 关键问题

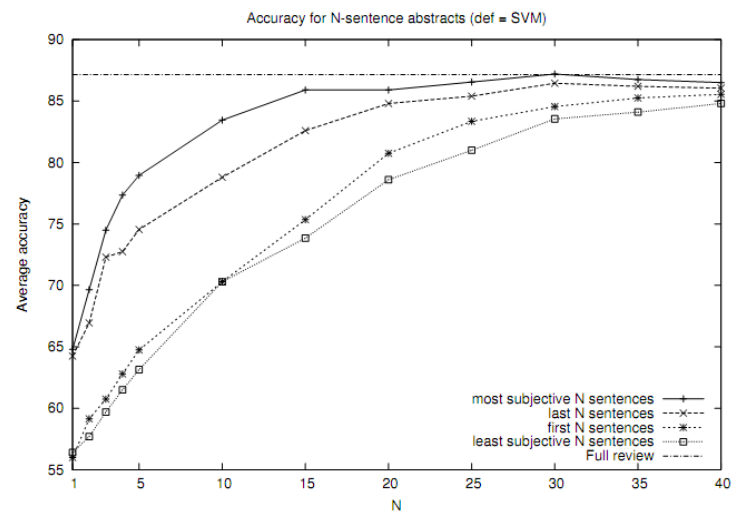
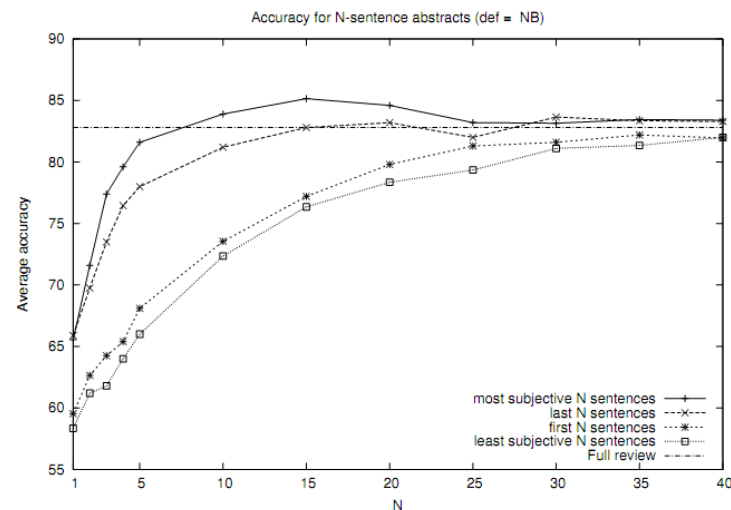
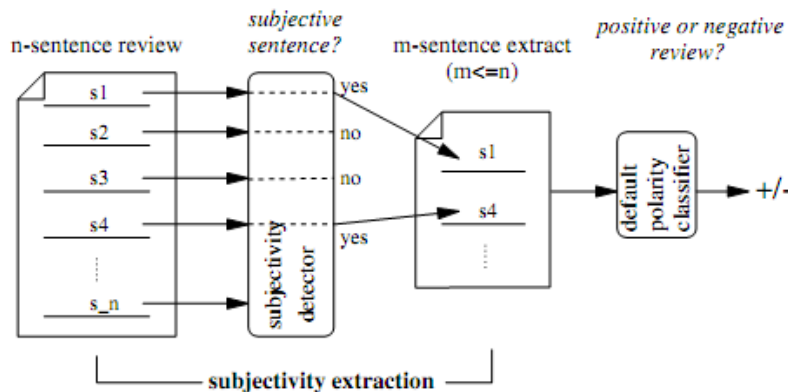
- 多观点倾向性：一篇商品评论中可能包含对于商品多方面的观点，每个观点的倾向性也可能不同，如何识别篇章整体的观点倾向性

- 按照句子划分
 - 按照主题划分

基于句子划分 (1/2)

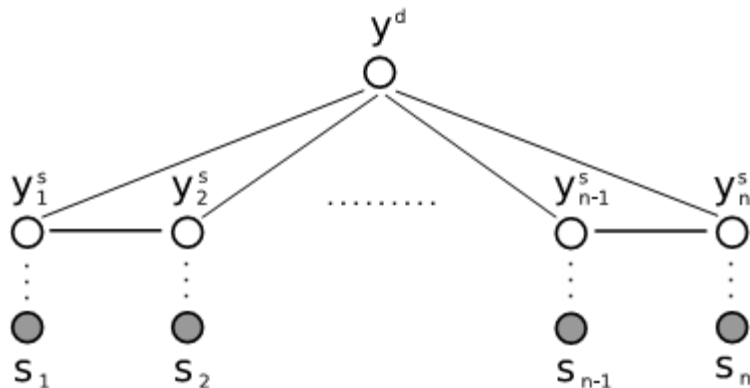
- 篇章中的客观句子对于篇章整体的观点倾向性没有意义 (Pang ACL 2004)

- 利用图算法从篇章中识别出观点句，剔除客观句
- 只利用观点句来识别篇章整体的观点倾向性



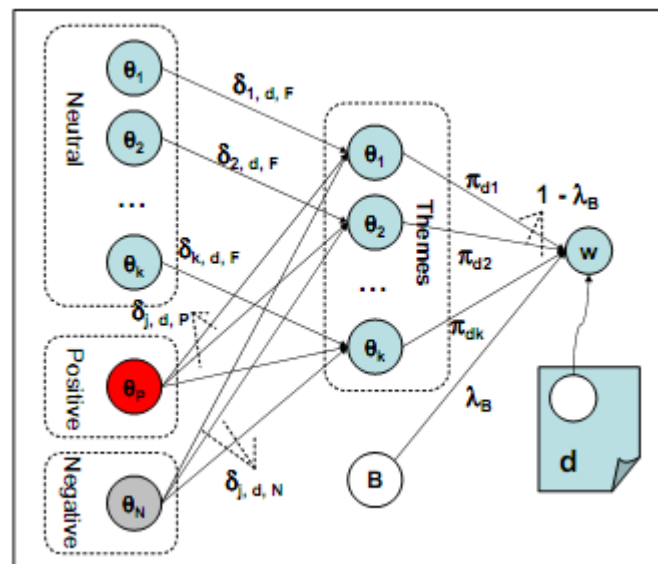
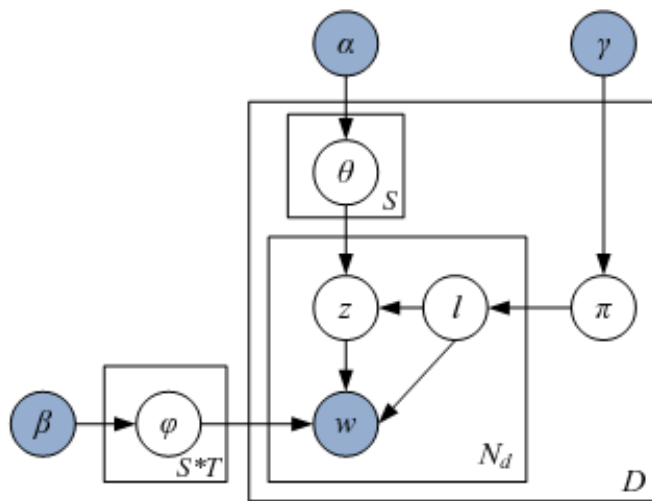
基于句子划分(2/2)

- 考虑篇章中每一个句子对于篇章整体倾向性的贡献 (McDonald ACL 2007)
 - 句子级倾向性识别与篇章级倾向性识别一体化
 - 考虑句子的上下文特征
 - 结构化CRFs模型



基于主题的划分

- Lin (CIKM 2009), Mei (WWW 2007)
 - 篇章整体的观点倾向性是篇章中针对每个子主题的观点倾向性的集成
 - 篇章主题信息与观点信息协同挖掘



小结

- 篇章级观点倾向性识别仍然可以看做是一个text categorization 任务
 - 如果仅仅是用词袋子模型，那么document level与sentence level在处理方法上没有区别
- 主要问题在多观点混合问题
 - 篇章中局部观点与整体观点具不一致

Sentiment Identification

- Word Level
- Sentence Level
- Document Level
- Others
 - 跨语言观点识别与分析
 - 领域适应性

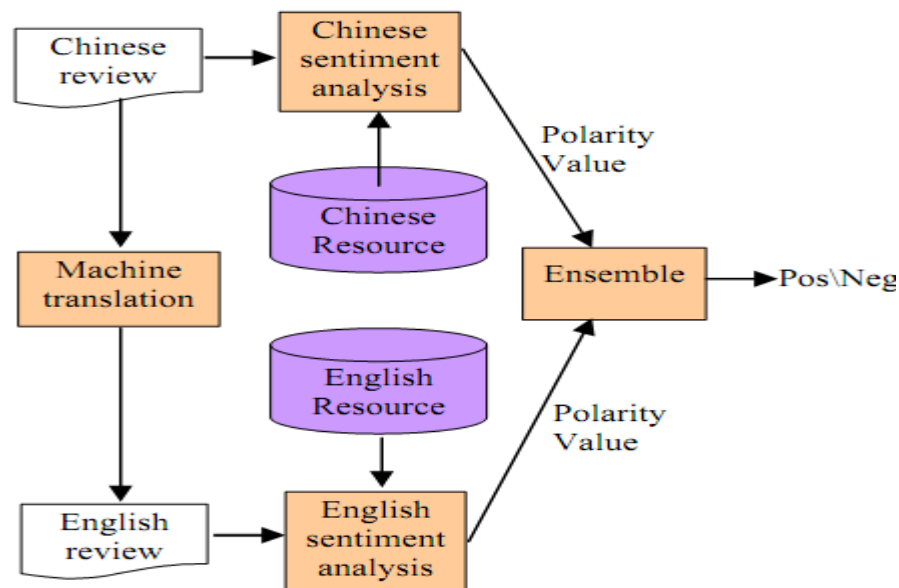
Cross-lingual Sentiment Classification

■ 任务

- ❑ 缺乏训练数据
- ❑ 利用其他语言资源
- ❑ 借鉴跨语言文本分类方法

■ 方法

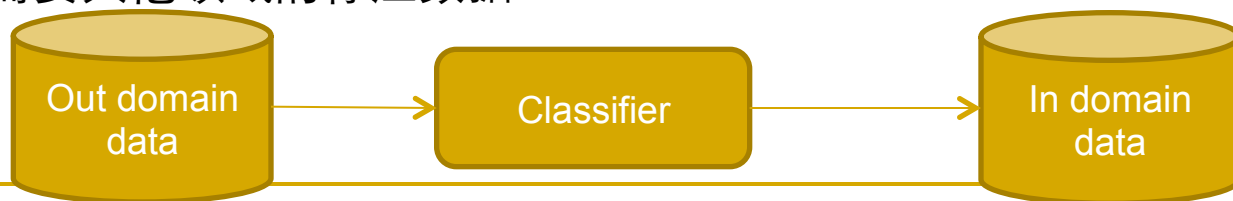
- ❑ 借助于翻译系统
- ❑ 比较不同翻译系统的作用 (Wan EMNLP 2008)
 - 利用不同的集成策略
- ❑ 采用多视角学习策略 (Wan ACL 2009)
 - 不同语言看做是样本的不同视角 (Co-Training)



Sentiment Transfer (1/2)

■ 问题

- 不同领域的情感倾向性具有差异性
- 同样的词在不同的领域的情感倾向性不同
 - Screen is big (positive) Phone's size is big (negative)
- 不同领域的用词不相同
 - Car domain: faster, power,
 - Phone domain: colorful,
- 传统统计机器学习假设：训练数据与预测数据具有相同的分布
- 训练语料规模有限
 - 需要其他领域的标注数据



Sentiment Transfer (2/2)

■ 方法 (两类)

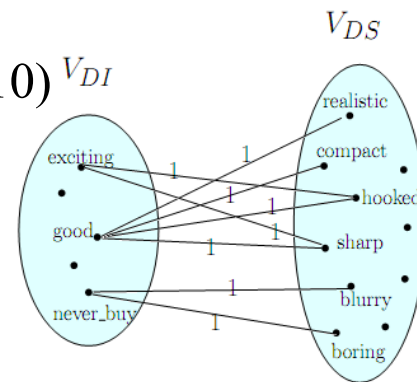
□ Instance view

- 假设：不同领域的数据的特征表示一致，数据分布不同
- 方法：调节样本的权重
- Jiang (ACL 2007), Dai (AAAI 2007)
 - The weight of the similar out-domain instances
 - The weight of the unlike out-domain instances



□ Feature representation view

- 假设：不同领域的数据的特征表示不一致
- 方法：统一特征表示
- Blitzer (ACL 2007), Liu (CIKM 2009), Pan (WWW 2010)
 - Select pivot features in two domains
 - Using pivot features to representation other features
 - Different data are represented in a unified feature space
 - Different features can corresponds



内容

- Sentiment Identification
- Opinion Mining
 - Opinion Target Extraction
 - Opinion Holder Extraction
- Opinion Retrieval
- Resources and Evaluations

Opinion Target Extraction (1/4)

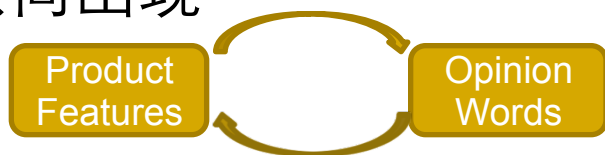
- 任务：抽取观点评价的对象
 - 中方发言人就美国近期对阿富汗的行动进行了强烈的谴责。（新闻）
 - iphone4的屏幕简直太酷了！（商品评论）
 - Product Feature: 商品、商品属性、商品的部件、商品部件的属性 (Popescu EMNLP 2005)

Explicit Features	Examples	% Total
Properties	ScannerSize	7%
Parts	ScannerCover	52%
Features of Parts	BatteryLife	24%
Related Concepts	ScannerImage	9%
Related Concepts' Features	ScannerImageSize	8%

- 不是所有的商品属性都是评价的对象
 - 诺基亚C1的屏幕尺寸有1.8寸。✗
 - iphone的价格太贵了 ✓

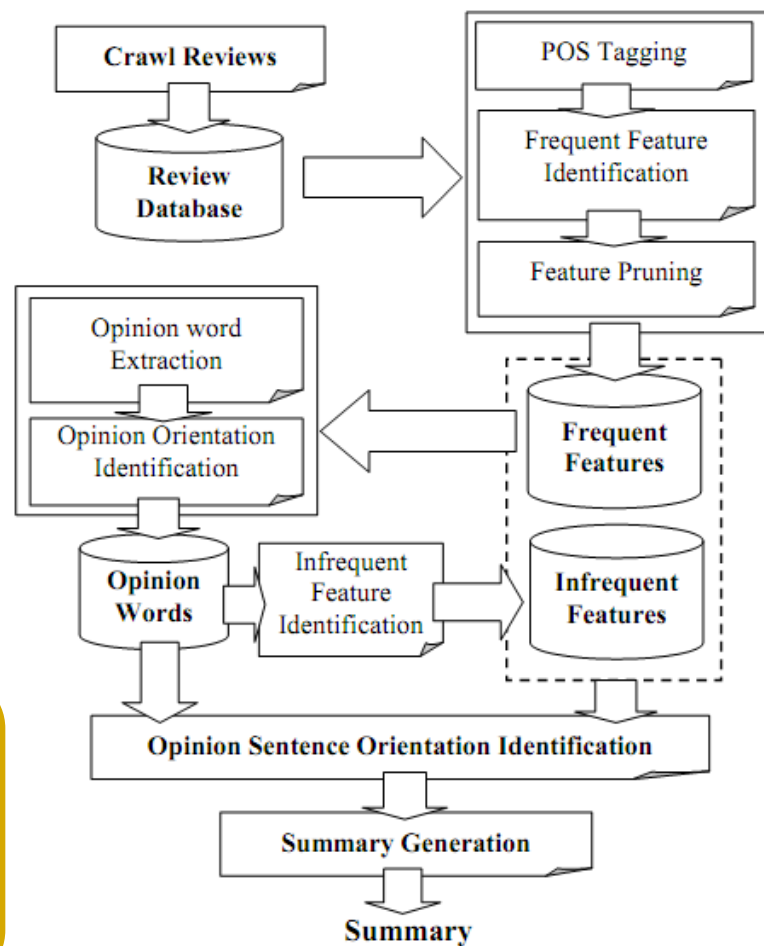
Opinion Target Extraction (2/4)

- 迭代抽取 (Liu KDD 2004, Liu WWW 2005)
 - 商品属性词与评价词在评论文本中共同出现



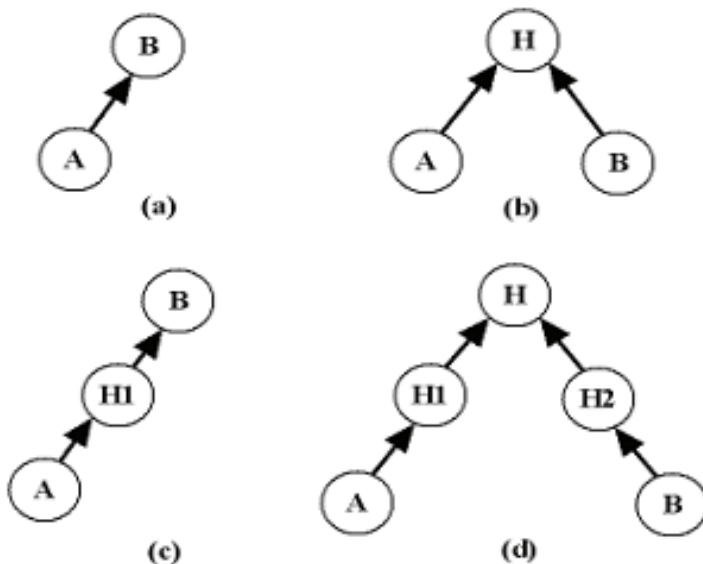
- 商品属性词分为两类
 - Frequent与Infrequent

- Step 1: Frequent features extraction
- Step 2: Opinion word extraction
- Step 3: Infrequent features extraction
- Step 4: Summarization



Opinion Target Extraction : 句法结构 (3/4)

- 利用属性词与评价词之间的依存句法关系 (Popescu EMNLP 2005, Qiu IJCAI 2009)

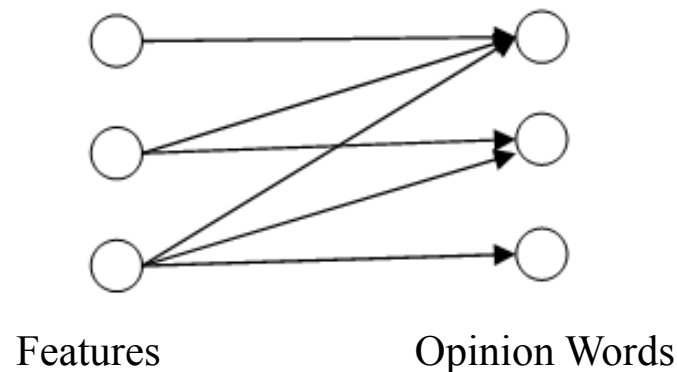


Extraction Rules	Examples
$\text{if } \exists(M, NP = f) \rightarrow po = M$	(expensive) scanner
$\text{if } \exists(S = f, P, O) \rightarrow po = O$	lamp has (problems)
$\text{if } \exists(S, P, O = f) \rightarrow po = P$	I (hate) this scanner
$\text{if } \exists(S = f, P, O) \rightarrow po = P$	program (crashed)

	Observations	Constraints	Outputs
R1 ₁	$S_{i0} \rightarrow S_{i0} \text{-Dep} \rightarrow S_{j0}$	$S_{j0} \in \{S\},$ $S_{i0} \text{-Dep} \in \{CONJ\},$ $POS(S_{i0}) \in \{JJ\}$	$s = S_{i0}$
R1 ₂	$S_i \rightarrow S_i \text{-Dep} \rightarrow H \leftarrow S_j \text{-Dep} \leftarrow S_j$	$S_i \in \{S\},$ $S_i \text{-Dep} = S_j \text{-Dep},$ $POS(S_i) \in \{JJ\}$	$s = S_j$
R2 ₁	$S \rightarrow S \text{-Dep} \rightarrow F$	$F \in \{F\},$ $S \text{-Dep} \in \{MR\},$ $POS(S) \in \{JJ\}$	$s = S$
R2 ₂	$S \rightarrow S \text{-Dep} \rightarrow H \leftarrow F \text{-Dep} \leftarrow F$	$F \in \{F\},$ $S/F \text{-Dep} \in \{MR\},$ $POS(S) \in \{JJ\}$	$s = S$
R3 ₁	$S \rightarrow S \text{-Dep} \rightarrow F$	$S \in \{S\},$ $S \text{-Dep} \in \{MR\},$ $POS(F) \in \{NN\}$	$f = F$
R3 ₂	$S \rightarrow S \text{-Dep} \rightarrow H \leftarrow F \text{-Dep} \leftarrow F$	$S \in \{S\},$ $S/F \text{-Dep} \in \{MR\},$ $POS(F) \in \{NN\}$	$f = F$
R4 ₁	$F_{i0} \rightarrow F_{i0} \text{-Dep} \rightarrow F_{j0}$	$F_{j0} \in \{F\},$ $F_{i0} \text{-Dep} \in \{CONJ\},$ $POS(F_{i0}) \in \{NN\}$	$f = F_{i0}$
R4 ₂	$F_i \rightarrow F_i \text{-Dep} \rightarrow H \leftarrow F_j \text{-Dep} \leftarrow F_j$	$F_i \in \{F\},$ $F_i \text{-Dep} = F_j \text{-Dep},$ $POS(F_i) \in \{NN\}$	$f = F_j$

Opinion Target Extraction: 监督与半监督 (4/4)

- 半监督学习方法 (Wang IJCNLP 2008, Zhu CIKM 2009)
 - 采用Bootstrapping策略
 - 使用少量标注的种子词
 - 利用属性词与评价此词之间的关联关系进行迭代抽取

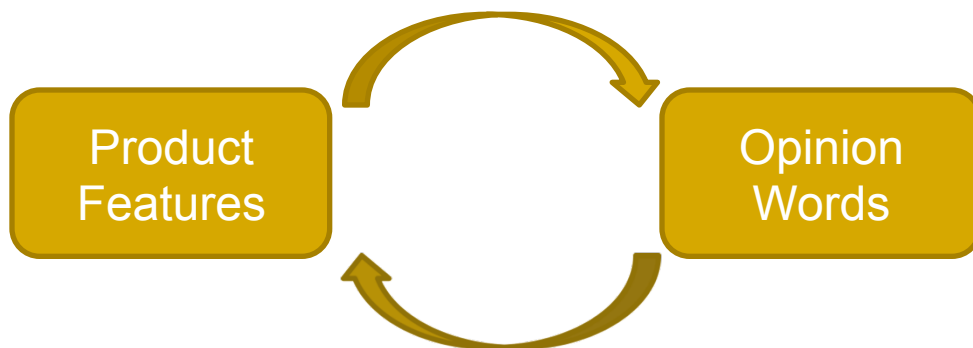


- 监督学习方法 (Li COLING 2008)
 - 看作序列标注任务
 - 利用CRFs进行标注

小结

■ 基本思路

□ 迭代抽取



■ 难点问题

□ “属性词-评价词”搭配关系的抽取

□ 商品名识别

□ 同义词问题

- 体积、大小、尺寸.....

□ Implicit 属性词抽取

- 太漂亮了（外观）

Opinion Holder Extraction

- 基本思路(Kim AAAI 2005)

- 命名实体识别

- 人名、机构名

- 句法结构特征

- Convolution Kernel

- 分类或者序列标注

- SVM, Naïve Bayes, CRFs

- 需要指代消解

内容

- Sentiment Identification
 - Opinion Mining
 - Opinion Retrieval
 - Resources and Evaluations
-

Opinion Retrieval

- 任务：
 - 从海量文本中根据查询找到观点信息
 - 根据主题相关度(topic relevance)与观点倾向性(opinion relevance)对于结果进行重排序
 - Topic relevance: traditional retrieval
 - Opinion relevance: opinion identification
- 关键问题
 - 找到Topic relevance score与Opinion relevance score的折中

Generative Model

- 基于词的观点检索模型 (Zhang SIGIR 2008)
 - 产生式模型

S: 观点信息(观点词)

$$p(d | q) \propto p(q | d)p(d)$$

主题相关

$$p(d | q, s) = \sum_i p(d | q, s_i) p(s_i, s)$$

$$= \frac{1}{|S|} \sum_i p(d | q, s_i)$$

Opinion
Relevance

$$\propto \frac{1}{|S|} \sum_i p(q, s_i | d) p(d)$$

$$= \frac{1}{|S|} \sum_i p(s_i | d, q) p(q | d) p(d)$$

Topic
Relevance

Unified Relevance Model

■ 查询词扩展(Huang CIKM 2009)

- ❑ 查询中往往没有观点词
 - 7.23事件
 - 需要对于查询进行扩展(添加观点)
- ❑ 与查询独立的观点信息扩展
 - 词典信息
 - 标注的倾向性语料中进行统计
- ❑ 与查询相关的观点信息扩展
 - 从用户的反馈数据中得到
- ❑ 混合模型

$$\begin{aligned} \text{Score}(D) = & \alpha \sum_{w \in Q} P(w|Q) \log P(w|D) + \\ & + \beta \sum_{w \in OV_1} P(w|R_1) \log P(w|D) + \\ & + (1 - \alpha - \beta) \sum_{w \in OV_2} P(w|R_2) \log P(w|D) \end{aligned}$$

Topic
Relevance

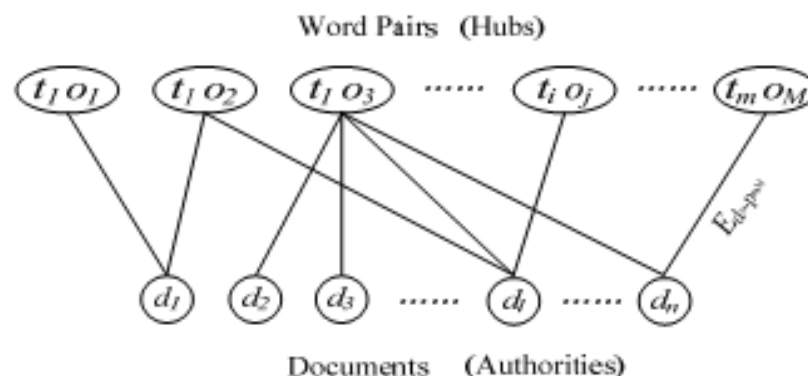
Query
independent
sentiment
expansion

Query
dependent
sentiment
expansion

$$P(w|R) \approx P(w|q_1, q_2, \dots, q_n) = \frac{P(w, q_1, q_2, \dots, q_n)}{P(q_1, q_2, \dots, q_n)}$$

Sentence-based Opinion Retrieval

- 面向句子级观点检索文本表示 (Li ACL 2010)
 - 传统的词袋子模型不能很好表示文档中的观点信息
 - 利用topic-sentiment pair 表示每一个句子
 - 采用窗口共现策略抽取pair
 - 利用HITS算法来计算每个pair在篇章中的权重



内容

- Sentiment Identification
- Opinion Mining
- Opinion Retrieval
- Resources and Evaluations
 - 资源：词典、语料
 - 评测：评测会议

Resources: Lexicon (1/2)

■ English

- ❑ General Inquirer (<http://www.wjh.harvard.edu/~inquirer/>)
 - Manually labeled terms (positive, negative)
- ❑ SentiWordnet (<http://sentiwordnet.isti.cnr.it/>)
 - Extend from WordNet
 - Each synset is automatically labeled as P, N, O
- ❑ OpinionFinder's Subjectivity Lexicon (<http://www.cs.pitt.edu/mpqa/>)
 - Subjective words provided by OpinionFinder
- ❑ Taboada and Grieve's Turney adjective list
 - Available through Yahoo SentimentAI group. 1700 words
- ❑ IBM Lexicon
 - 1,267 positive words and 1,701 negative words (Melville 2009)

Resources: Lexicon (2/2)

■ Chinese

- Hownet (http://www.keenage.com/html/e_index.html)
 - 正面情感、负面情感、正面评价、负面评价、程度级别、主张词语6个子集
- NTU Sentiment Lexicon
(<http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp>)
 - List the polarities of many Chinese words

Resource: Corpus (1/2)

■ English

- MPQA (<http://www.cs.pitt.edu/mpqa/databaserelease/>)
 - 535 news articles (subjective, objective; P,N,O)
- Movie review data (<http://www.cs.cornell.edu/people/pabo/movie-review-data/>)
 - IMDB
 - Document level 2000
 - Sentence level 5000
- Custom review data (<http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip>)
 - Product reviews (Product features, P,N)
- Multi product reviews (<http://john.blitzer.com/software.html>)
 - Book, Electronic, Kitchen, DVD
 - 2000 in each domain
- TREC Blog corpus (<http://trec.nist.gov/>)
 - Blog data
 - 3,000,000 Webpages
- Multiple-aspect restaurant reviews
 - 4,488 reviews
 - Each review labeled as 1-5 stars

Resource: Corpus (2/2)

■ Chinese

- NTCIR (<http://research.nii.ac.jp/ntcir/>)
 - Multilingual news articles
- COAE商品属性语料
 - 口碑网, it168,
 - 494 document, 5 domains
- 中文情感挖掘语料
 - Positive, Negative
 - 10,000
- Zagibalov (<http://www.informatics.sussex.ac.uk/users/tz21/>)
 - Phone reviews
 - 1,158 positive and 1,159 negative

Evaluation

- TREC Blog Track (start from 2006)
 - Task: Opinion Retrieval and Polarity Identification
 - Corpus: 3,000,000 English webpages
- NTCIR
 - Task:
 - Topic Relevance
 - Opinion identification
 - Polarity Identification
 - Opinion Holder extraction
 - Opinion Target extraction
 - Corpus: news articles (English, Chinese, Japanese, Korea)
- Chinese (COAE 2008, 2009)
 - Task:
 - Words level (sub/obj, positive/negative)
 - Documents level (sub/obj, positive/negative)
 - Opinion Target extraction
 - Opinion Retrieval
 - Corpus: Chinese

目录

- 第一部分：
 - 我们为什么需要观点挖掘与倾向性分析？
 - 什么是观点挖掘与倾向性分析？
- 第二部分：
 - 如何进行观点挖掘与倾向性分析？
 - 任务、方法、资源、评测
- 第三部分：
 - 问题与挑战

观点信息应该如何表示?

- *An opinion is a quintuple (Liu Handbook in NLP)*

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

where

- o_j is a target object.
- f_{jk} is a feature of the object o_j .
- so_{ijkl} is the sentiment value of the opinion. so_{ijkl} is +ve, -ve, or neu, or a more granular rating.
- h_i is an opinion holder.
- t_l is the time when the opinion is expressed.

Challenge: Sentiment Identification(1/4)

■ Sentence Level

- 如何对于一个句子中的观点信息进行表示?
- BOW 模型? 句法结构?
- What is different with topic-based categorization

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	78.7	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	82.9
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	82.7
(4)	bigrams	16165	pres.	77.3	77.4	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	81.9
(6)	adjectives	2633	pres.	77.0	77.7	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	81.4
(8)	unigrams+position	22430	pres.	81.0	80.1	81.6

- 一些特定的句法现象
 - Polarity Shift (否定、转折.....)
 - Comparative Sentence
- 数据稀疏问题, 如何特征扩展
 - 微博、产品评论

Challenge: Sentiment Identification(2/4)

■ Word Level

□ 如何对于词的领域性进行区分

■ Independence word

□ 好、坏、轻松、迅速.....

■ Topic depended word

□ 大（屏幕大，体积大）

□ 高（个子高，温度高）

□ 名词、动词也具有倾向性

■ “这家餐馆不会再来了”

■ “坑爹啊”

■ “☺”

Challenge: Sentiment Identification(3/4)

■ Feature (Aspect) Level

□ Product Feature Extraction

- Explicit Feature: 这款手机的屏幕很漂亮。（屏幕）
- Implicit Feature: 这款手机太大了。（体积）

□ Feature Grouping

- 屏幕：LCD、屏幕、显示屏.....

□ Feature-based Sentiment Identification

- Word matching + Word level sentiment identification
 - “诺基亚5800个**头**很**大**”。
- 句法分析？面对口语化文本似乎力不从心

Challenge: Sentiment Identification(4/4)

■ Document Level

- ❑ 多观点混合问题
- ❑ 很难精确地确定篇章的倾向性由哪个句子决定
 - 不一定是多数制胜

“诺基亚5800屏幕很好，操作也很方便，通话质量也不错，外形还可以，但是电池太不行了，只能坚持一天，反正我觉得不值。”

❑ Sentiment Rating

- 不同用户，不同尺度

★★★★☆ 口味: 2(好) 环境: 2(好) 服务: 2(好) 人均: ¥90(晚餐)

跟朋友聚餐来吃的，晚上来的，人很多，好在我们来得早，没有等位，点的是锅底一直不上，后来上了还上错了，好不容易才把我们要的端上来，味道们家的特色已经逐渐不明显了，就是羊肉还不错，比较新鲜。最后临走服务员折，后来填完了说大不了折了，说送个大果盘，我们本来也不是想占这个便宜了，跟海底捞什么的确实没法比

Challenge: Others

■ 观点检索

□ Re-Rank路线

$$\blacksquare \text{ Score} = \text{lamada} * \text{TopicRelevance} + (1 - \text{lamada}) * \text{OpinionScore}$$

□ 能否有独立的模型框架

□ 更广泛的应用：微博

■ 观点Spam

□ 内容、网页结构、用户行为

■ 观点信息是动态变化的

□ 时间、地点

■ 观点分析与应用紧密结合

□ 推荐系统

□ 广告投放

□ ...

Reference (1/5)

- J. Blitzer, M. Dredze and F. Pereira. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL). pages 440-447. 2007.
- Wenyuan Dai, Gui-Rong Xue, Qiang Yang and Yong Yu. Transferring Naïve Bayes Classifiers for Text Classification. In Proceedings. of AAAI. 2007.
- Weifu Du, Songbo Tan, Xueqi Cheng, Xiaochun Yun: Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. WSDM 2010: 111-120
- Ahmed Hassan, and Dragomir Radev. 2010. Identifying Text Polarity Using Random Walks. The 48th Annual Meeting of the Association for Computational Linguistics
- M. Hu and B. Liu. Mining Opinion Features in Customer Reviews. In Proceedings of AAAI, 2004.
- Xuanjing Huang and W. Bruce Croft. A Unified Relevance Model for Opinion Retrieval. In Proceedings of CIKM 2009.
- Jaap Kamps, Maarten Marx, Robert J. Mokken and Maarten de Rijke. Using WordNet to measure semantic orientation of adjectives. In Proc. of LREC'04, pp. 1115-1118, 2004.
- Jin Jiang and ChengXiang Zhai. Instance Weighting for Domain Adaptation in NLP. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL), pages 264-271. 2007.
-

Reference (2/5)

- Soo-Min Kim and Eduard Hovy. Identifying Opinion Holders for Question Answering in Opinion Texts. 2005. *In Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*
- Binyang Li, Lanjun Zhou, Shi Feng, Kam-Fai Wong, A Unified Graph Model for Sentence-based Opinion Retrieval, In Proceedings of ACL 2010
- Tao Li, Yi Zhang and Vikas Sindhwani. A Non-negative Matrix Tri-factorization Approach to Sentiment Classification with Lexical Prior Knowledge. In Proceedings of ACL. 2009.
- Shoushan Li, Rui Xia, Chengqing Zong, Chu-Ren Huang: A Framework of Feature Selection Methods for Text Categorization. ACL/AFNLP 2009: 692-700.
- Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang, Guodong Zhou: Sentiment Classification and Polarity Shifting. COLING 2010: 635-643
- Fangtao Li, Chao Han, Minlie Huang and Xiaoyan Zhu. Structure-Aware Review Mining and Summarization. In The 23rd International Conference on Computational Linguistics (COLING 2010)
- Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web" To appear in Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005, in Chiba, Japan
-

Reference (3/5)

- Kang Liu and Jun Zhao. Cross-Domain Sentiment Classification using a Two-Stage Method. In Proceedings of *the 18th ACM Conference on Information and Knowledge Management (CIKM)*. November 2-6, 2009, Hong Kong
- Chenghua Lin and Yulan He. Joint Sentiment/Topic Model for Sentiment Analysis. In Proceedings of CIKM's 09. 2009
- Y. Mao and G. Lebanon, Isotonic Conditional Random Fields and Local Sentiment Flow. Advances in Neural Information Processing Systems 19, 2007
- Ryan McDonald, Kerry Hannan and Tyler Neylon et al. Structured Models for Fine-to-Coarse Sentiment Analysis. In Proceedings of ACL, 2007, pp. 432-439.
- Qiaozhu Mei, Xu Ling, et al. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In Proceedings of WWW 2007.
- Prem Melville, Wojciech Gryc and Richard D. Lawrence. Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification. In Proceedings of KDD. 2009.
- Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the Association of Computational Linguistics (ACL).
- Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of EMNLP 2002, pp.79-86.
-

Reference (4/5)

- Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang and Zheng Chen. Cross-Domain Sentiment Classification via Spectral Feature Alignment. In Proceedings of the 19th International World Wide Web Conference (WWW-10). Raleigh, NC, USA. April 26-30, 2010. Pages 751-760.
- Popescu A. M. and Etzioni O. Extracting Product Features ad Opinion Reviews. In Proceedings of EMNLP'05, 2005.
- L. Qiu, Weishi Zhang, Changjian Hu and Kai Zhao. SELC: A Self-Supervised for Sentiment Classification. In Proceedings of CIKM, 2009.
- Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen: Expanding Domain Sentiment Lexicon through Double Propagation. IJCAI 2009: 1199-1204
- Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of ACL. 2002.
- Xiaojun Wan. Co-Training for Cross-Lingual Sentiment Classification. In Proceedings of ACL-IJCNLP, pages 235-243, 2009.
- Xiaojun Wan. Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In Proceedings of EMNLP, pages 553-561. 2008.
- Bo Wang, Houfeng Wang: A Cross-Inducing Method for Bootstrapping Product Features and Opinion Words. In Proceedings of 2008 International Conference on Natural Language Processing (IJCNLP 2008), India
-

Reference (5/5)

- Janyce Webie, Theresa Wilson and Claire Cardie. Annotating expressions of opinions and emotions in Proceedings of language. Language Resources and Evaluation 2005
- Taras Zagibalov. and John Carroll. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In Proceedings of The 22nd International Conference on Computational Linguistics (COLING), 2008. Manchester, UK.
- Min Zhang and Xingyao Ye. A Generative Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval. In Proceedings of SIGIR, pp. 411-418, 2008.
- Jun Zhao, Kang Liu and Gen Wang. Adding Redundant Features for CRFs-based Sentence Sentiment Classification. In Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP). October 25-27, 2008, Hawaii
- Jingbo Zhu, Huizhen Wang, Benjamin Tsou and Muhua Zhu. 2009. Multi-aspect opinion polling from textual reviews, In Proceedings of CIKM'09, short session, pp1799-1802

Q&A

Thanks