

# 第12章 文本分类、聚类与情感分析

北京市海淀区中关村东路95号

邮编：100190



电话：+86-10-8254 4688

邮件：cqzong@nlpr.ia.ac.cn

# 主要内容

---



- ◆ 文本分类

- ◆ 文本聚类

- ◆ 情感分析

# 相关概念

- ◆ 情感分析研究观点挖掘、倾向性分析等
- ◆ 什么是观点挖掘与倾向性分析？
- ◆ 为什么需要观点挖掘与倾向性分析？

# 相关定义

- **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



情感分析或观点挖掘(in Wikipedia) 是自然语言处理、计算语言学与文本挖掘中的一个研究领域。它的目标在于确定一个说话者或作者对于相关话题的 情感、观点或态度。

# 例子

“我今年入手诺基亚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



# 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的兴起，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

- 自动发现用户情感与观点 (市场智能化)
- 感知社会发展趋势
- 获取商业机会
- 在线名誉管理
- 目标导向地广告

## ■ 普通用户对观点挖掘和倾向性分析的需求

- 有助于购买产品
- 有利于发现针对政治话题的观点

## ■ 政府对观点挖掘和倾向性分析的需求

- 控制公众整体情绪
- 检测公共事件



# 观点挖掘与倾向性分析相关任务

## ■ 观点及倾向性识别

### ■ 情感识别 (Sentiment Identification)

## ■ 观点要素抽取

### ■ 观点属性抽取 (Opinion Attribute Extraction)

### ■ 观点摘要 (Opinion Summarization)

## ■ 观点检索



# 情感识别

## ■ 观点识别 (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)

## ■ 极性分类 (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，值得推荐。(Positive)

## ■ 强度识别 (情感强度识别)

- iPhone X的价格太贵了，两个肾都没了。(强烈反对)
- iPhone X的价格有点贵。(有点差)



hello精品 🏆🏆🏆

🌟🌟🌟🌟🌟 口味:3 环境:2 服务:2

来这里之后觉得还不错，味道挺好的尤其  
顾这家店哦

# 情感识别

- 词级别
  - 识别一个词的倾向性
- 特征级别(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**
- 句子级别
  - 识别一个句子的观点倾向性
- 文档级别
  - 识别一篇文本（包含多个句子）整体的倾向性

# 观点属性抽取

## ■ 观点持有者抽取

- “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

## ■ 观点目标抽取

- “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

# 观点摘要

*“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. ...”*

## 观点摘要:

### 特征 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**.

...

### 特征 2: **battery life**

...

# 观点检索

- 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
  - 博客、微博、论坛.....



华为 HUAWEI P10 全网通 4GB+64GB 钻雕金 移动联通电信4G手机 双卡双待

麒麟960芯片！wifi双天线设计！徕卡人像摄影！白条12期免息！华为更多优惠详情请见！

京东价 **¥3788.00** 降价通知

好评度

**96%**

支持国产(95)

系统流畅(82)

照相不错(77)

反应快(67)

外观漂亮(62)

指纹识别(62)

金属机身(54)

通话质量好(51)

分辨率高(50)

功能齐全(49)

全部评价(2.2万+)

晒图(500)

追评(700+)

好评(2.1万+)

中评(500+)

差评(500+)

☐ 只看当前商品评价

推荐排序 ∨



jd\_死胖子

金牌会员



外观很美系统流畅，同一个路由器p10比红米4下载要快一倍。安装软件特别快。亮屏3个小时了才用不到20%，虽然没玩游戏，但是这期间我在不停的下软件，导入旧手机数据看了一会贴吧，续航很强悍。一分钱一分货，红米白白了。（垃圾红米拍照真差）



C\*\*\*e

金牌会员



手机买来快半个月了，特意用一段时间再来评论的，当初决定买这个手机就是图它电池容量，相对的屏幕大小，双卡双待还有质感。还有支持国产。首先手机屏幕和大小单手操作的话还是有点勉强，电池的话个人有点失望，勉强能维持一天时间，我每天电话比较多，其次系统，平时操作起来确实挺快的，没毛病，但有偶尔的卡机，这个试用体验真的很差，比较国产\*起的手机也是有点贵



小军啊剩点花钱

金牌会员



第一，手机玩游戏发热，第二，这个电池太不耐用，正常打电话一天都用不上，就别说要游戏了！第三，刚用一天就升级，，第四，这手机信号也太差了吧，没信号！大家都看看！买了就后悔了！



曜石黑

64GB

2017-05-04 00:45

举报

182

129







## Apple iPhone 7 Unlocked Phone 128 GB - US Version (Black)



301 customer reviews | 763 answered questions

Available from these sellers.

Size: 128 GB



18 user reviews



CNET Editors' Rating

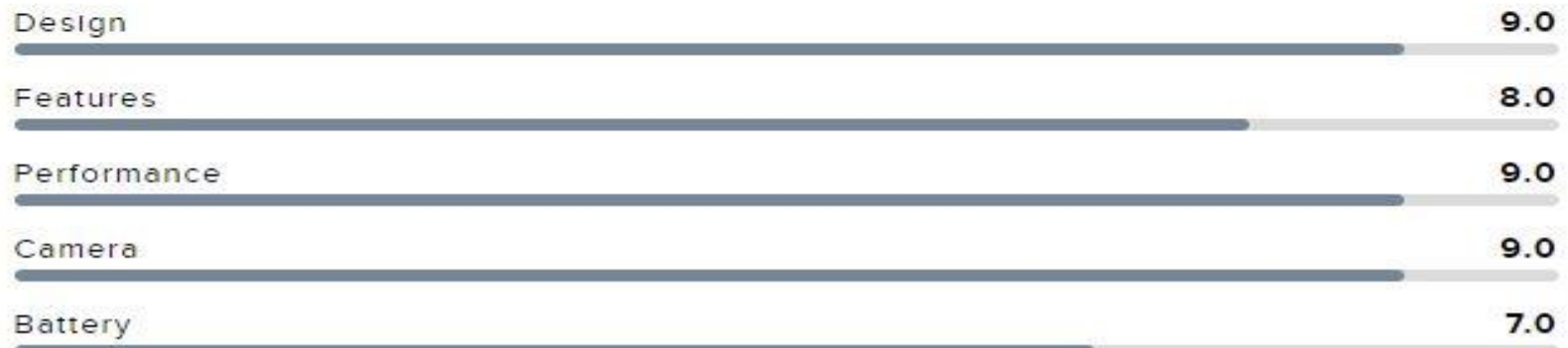
**The Good** / Improved front and rear cameras -- now with optical image stabilization -- deliver much improved photos, especially in low light. Water resistant. A faster processor, plus slightly better battery life. More onboard storage than last year's models for the same price.

**The Bad** / No headphone jack (but there's a dongle and compatible wired headphones in the box). Click-free home button takes getting used to. Only the larger 7 Plus has the cool dual camera. Shiny jet-black version scratches easily.

**The Bottom Line** / The iPhone 7's notable camera, battery and water resistance improvements are worthwhile upgrades to a familiar phone design. But ask yourself if you really need an upgrade... and if the Plus might be a better choice.

# 8.7

OVERALL





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 up 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



# 主要内容

- ◆ 文本分类
- ◆ 文本聚类
- ◆ 情感分析
  - 相关概念
  - 典型方法
  - 问题与挑战

# 典型方法

- 情感识别
- 观点挖掘
- 观点检索
- 资源和评测

# 情感识别

- 词级别
- 句子级别
- 文档级别
- 其他

# 词级别情感识别

## ■ 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
- 基于词典的方法
- 基于语料库的方法

# 情感识别

- 词级别
- 句子级别
- 文档级别
- 其他

# 句子级情感识别

- 任务：识别句子的情感倾向性

- “这部电影看得想吐，看了5分钟就看不下去了。”



- 关键问题

- 如何进行特征表示

- 分类：

- 基于语料库的方法

- 基于词典的方法

- 融合方法

# 与传统方法的区别

- 基于话题的文本分类

- 侧重于主题词特征

- “这款手机的屏幕太大了” (科技、手机)

- 情感识别

- 表示倾向性的词语更加重要.

- “这款手机的屏幕好大” (主观、褒义)

# 基于语料库的方法-特征选择

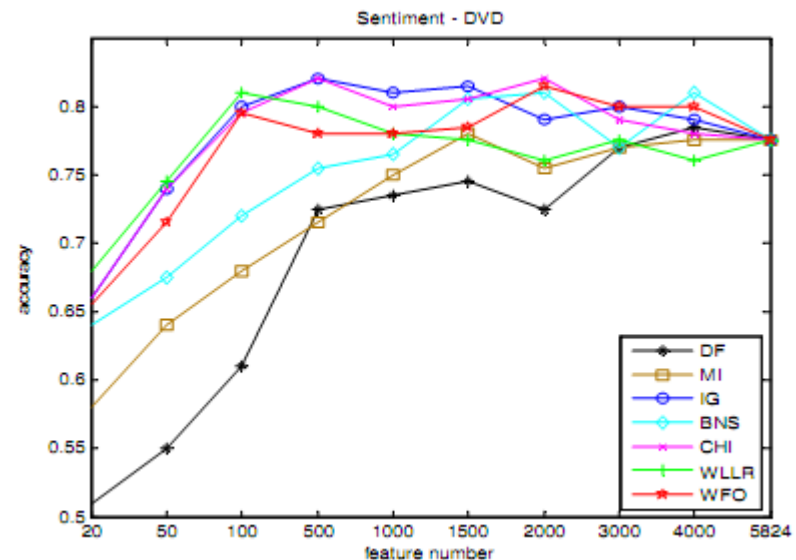
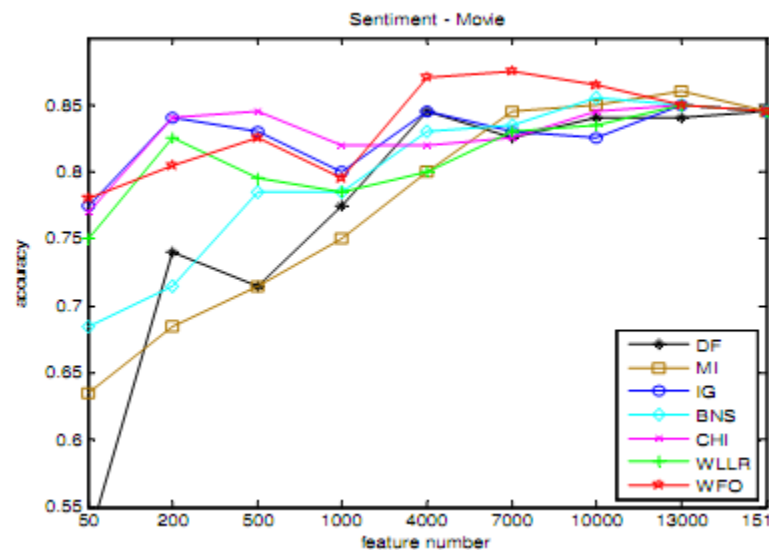
- 利用传统文本分类方法处理情感分类任务 (Pang EMNLP 2002)
  - 比较多种特征的效果
    - Unigram、bigram、POS、Adj.、Position
  - 比较多个分类器性能
    - SVM、Naïve Bayes、Maximum Entropy

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



# 基于语料库的方法-特征选择

- 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



# 基于语料库的方法-极性迁移

## ■ 极性迁移

### ■ 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？

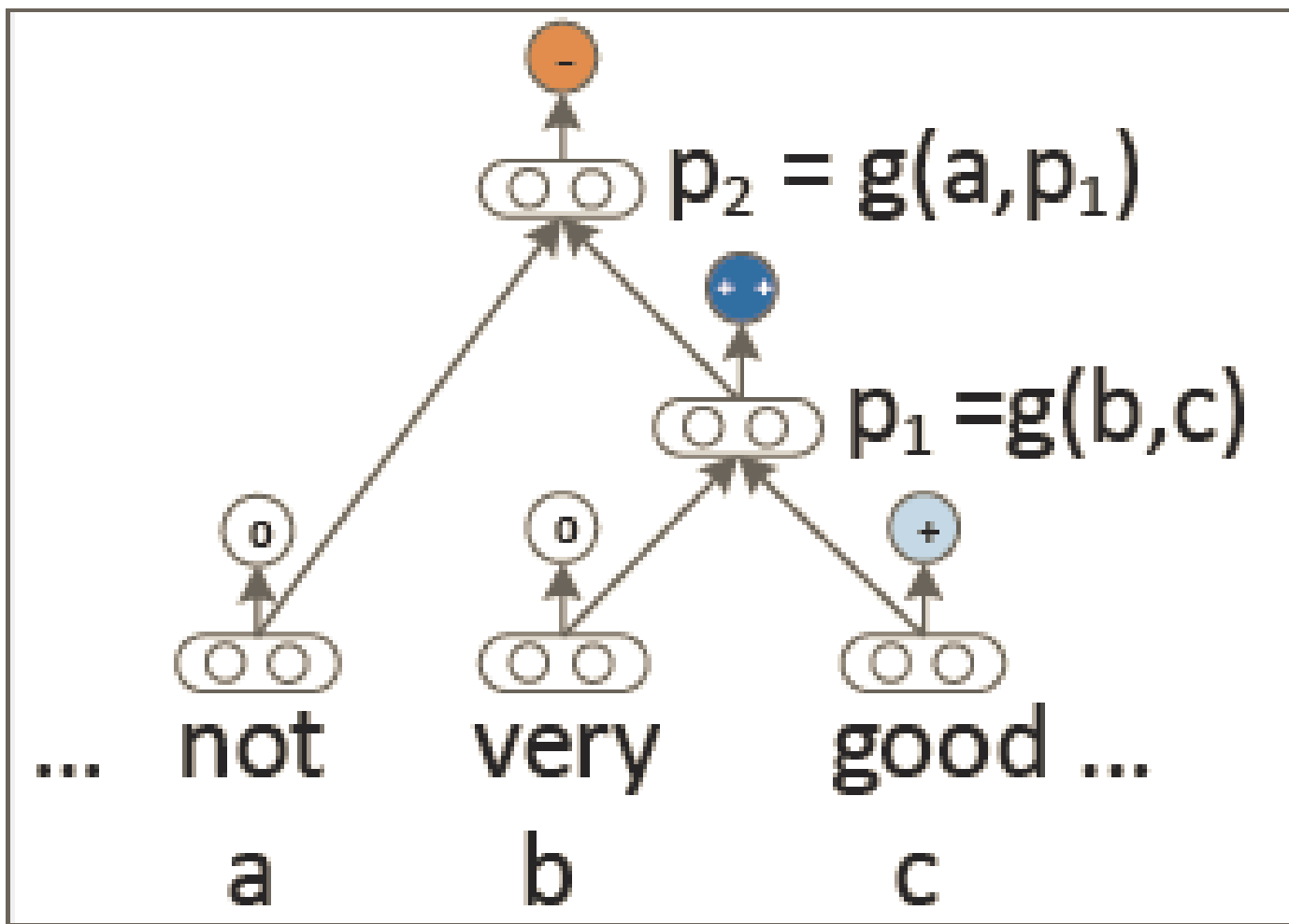
### ■ 方法

■ 在句子中检测出极性迁移

■ 判别句子倾向性时对于极性迁移专门处理

# 基于语料库的方法-极性迁移

## ■ 极性迁移的检测-基于神经网络的方法



# 情感识别

- 词级别
- 句子级别
- 文档级别
- 其他

# 文档级情感识别

- 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

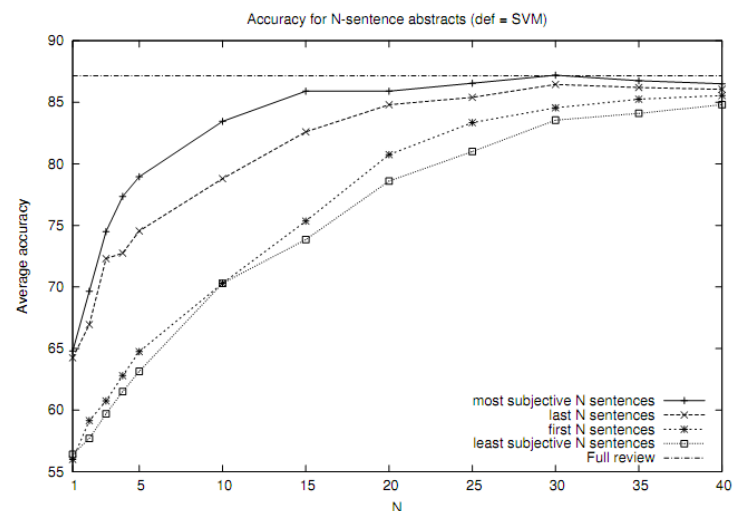
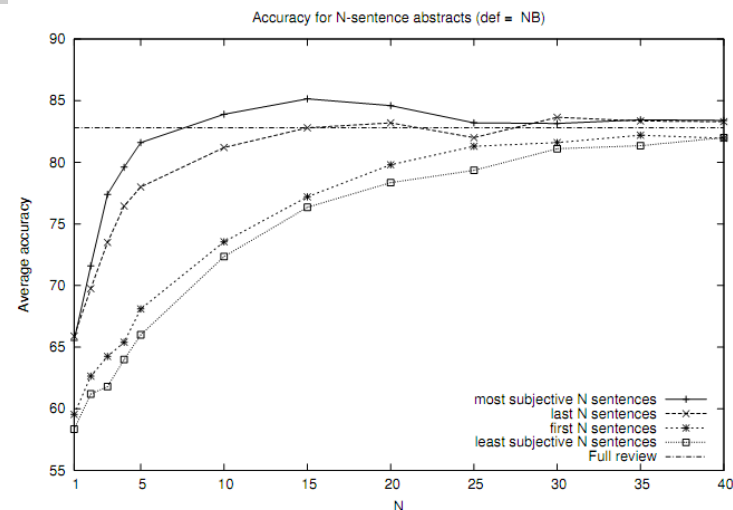
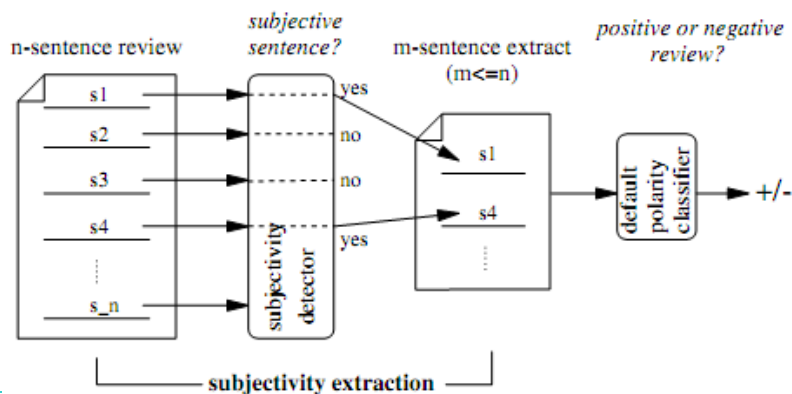
- 特征+分类器

- 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

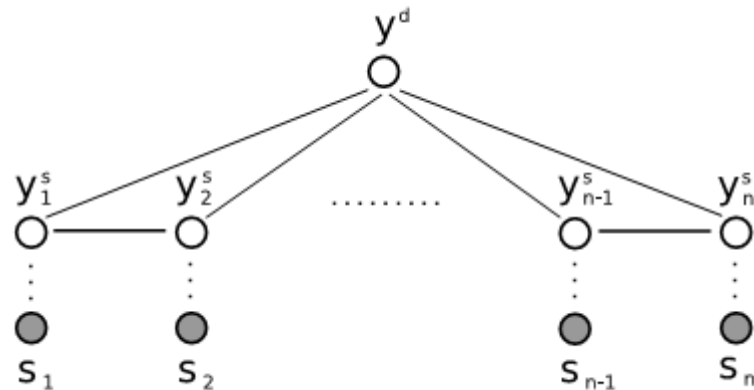
# 基于句子的划分 (1)

- 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

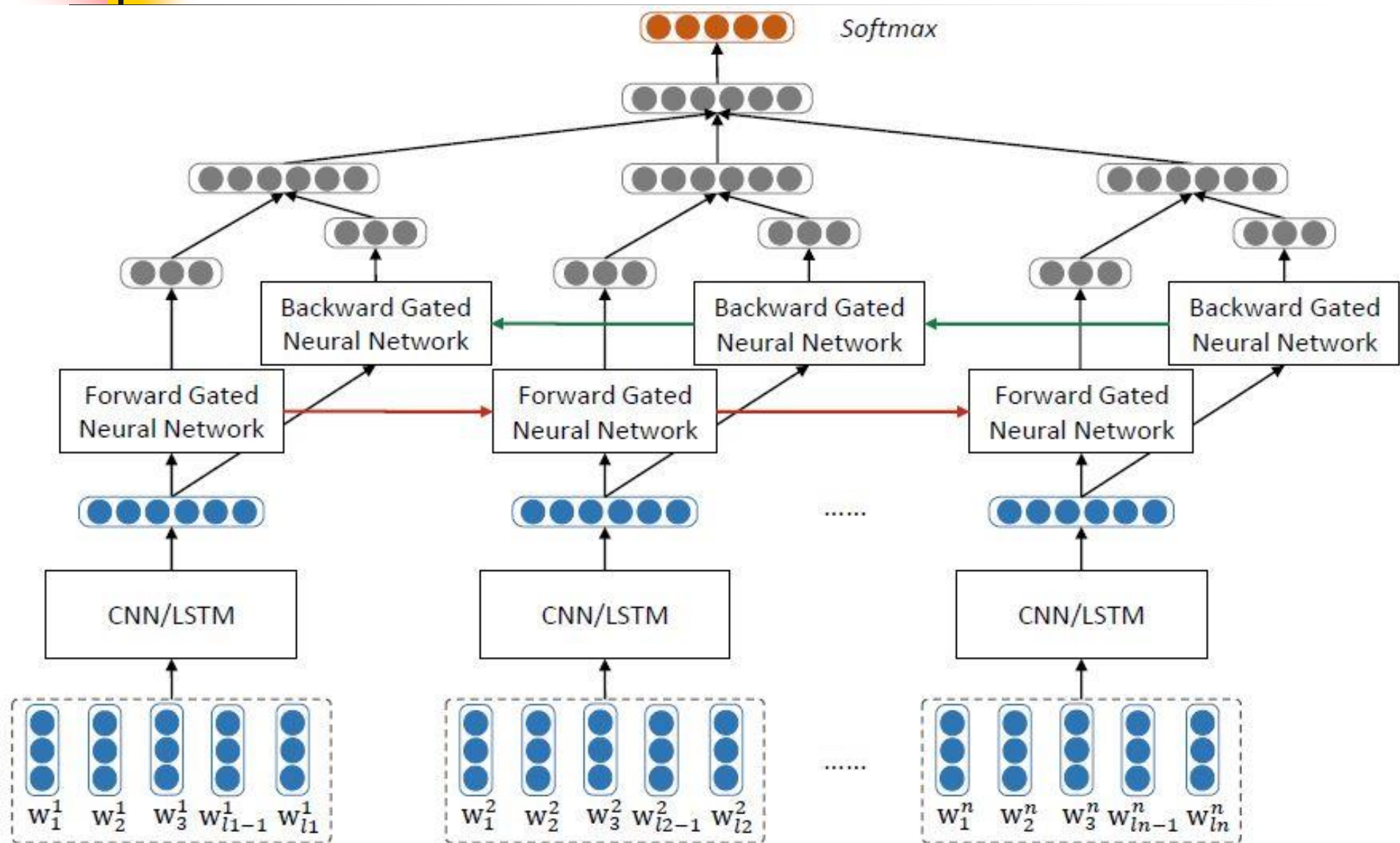


# 基于句子的划分 (2)

- 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



# 基于深度学习的方法





# 小结

- 篇章级观点倾向性识别仍然可以看做是一个文本分类任务
  - 如果仅仅是用词袋子模型，那么文档级别与句子级别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不一致

# 情感识别

- 词级别
- 句子级别
- 文档级别
- 其他
  - 跨语言观点识别与分析
  - 领域适应性

# 典型方法

- 情感识别
- 观点挖掘
- 观点检索
- 资源和评测

# 观点对象抽取

## ■ 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）
- iphone7的屏幕简直太酷了！（商品评论）
  - 产品特征: 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%

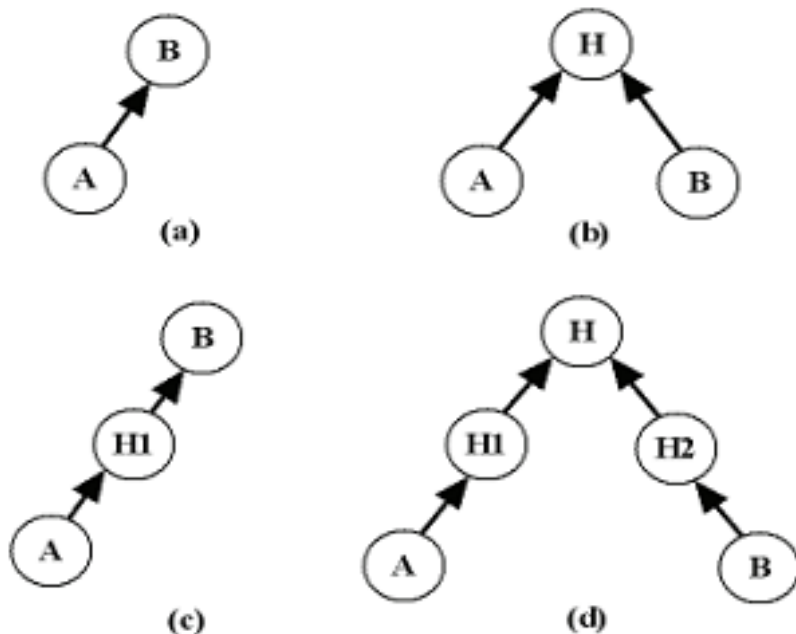
## ■ 不是所有的商品属性都是评价的对象

- 诺基亚C1的屏幕尺寸有1.8寸。
- iphone的价格太贵了



# 观点对象抽取

- 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 <sub>1</sub>	$S_{i(j)} \rightarrow S_{i(j)}\text{-Dep} \rightarrow S_{j(i)}$	$S_{i(j)} \in \{S\},$ $S_{i(j)}\text{-Dep} \in \{CONJ\},$ $POS(S_{i(j)}) \in \{JJ\}$	$s = S_{i(j)}$
R1 <sub>2</sub>	$S_i \rightarrow S_i\text{-Dep} \rightarrow H \leftarrow S_j\text{-Dep} \leftarrow S_j$	$S_i \in \{S\},$ $S_i\text{-Dep} = S_j\text{-Dep},$ $POS(S_i) \in \{JJ\}$	$s = S_j$
R2 <sub>1</sub>	$S \rightarrow S\text{-Dep} \rightarrow F$	$F \in \{F\},$ $S\text{-Dep} \in \{MR\},$ $POS(S) \in \{JJ\}$	$s = S$
R2 <sub>2</sub>	$S \rightarrow S\text{-Dep} \rightarrow H \leftarrow F\text{-Dep} \leftarrow F$	$F \in \{F\},$ $S/F\text{-Dep} \in \{MR\},$ $POS(S) \in \{JJ\}$	$s = S$
R3 <sub>1</sub>	$S \rightarrow S\text{-Dep} \rightarrow F$	$S \in \{S\},$ $S\text{-Dep} \in \{MR\},$ $POS(F) \in \{NN\}$	$f = F$
R3 <sub>2</sub>	$S \rightarrow S\text{-Dep} \rightarrow H \leftarrow F\text{-Dep} \leftarrow F$	$S \in \{S\},$ $S/F\text{-Dep} \in \{MR\},$ $POS(F) \in \{NN\}$	$f = F$
R4 <sub>1</sub>	$F_{i(j)} \rightarrow F_{i(j)}\text{-Dep} \rightarrow F_{j(i)}$	$F_{i(j)} \in \{F\},$ $F_{i(j)}\text{-Dep} \in \{CONJ\},$ $POS(F_{i(j)}) \in \{NN\}$	$f = F_{i(j)}$
R4 <sub>2</sub>	$F_i \rightarrow F_i\text{-Dep} \rightarrow H \leftarrow F_j\text{-Dep} \leftarrow F_j$	$F_i \in \{F\},$ $F_i\text{-Dep} = F_j\text{-Dep},$ $POS(F_i) \in \{NN\}$	$f = F_j$

# 观点持有者抽取

- **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**

# 典型方法

- 情感识别
- 观点挖掘
- 观点检索
- 资源和评测

# 观点检索

## ■ 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
  - 主题相关度: 传统检索
  - 观点倾向性: 观点识别

## ■ 关键问题

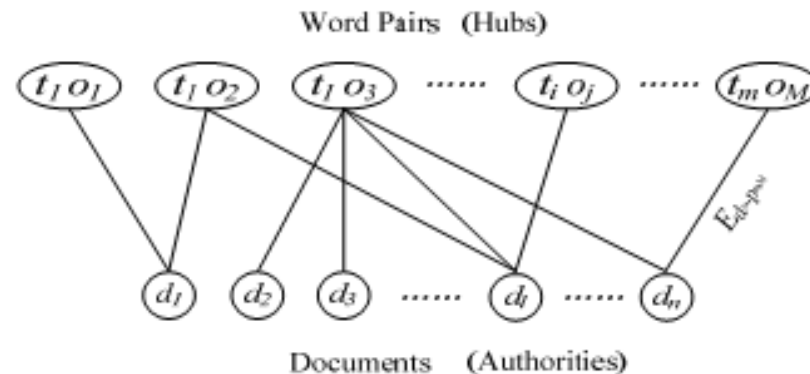
- 找到主题相关度得分与观点倾向性得分的折中



# 基于句子的观点检索

## ■ 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



# 典型方法

- 情感识别
- 观点挖掘
- 观点检索
- 资源和评测
  - 资源：词典、语料
  - 评测：评测会议

# 资源：词典（1）

## ■ 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)

# 资源：词典（2）

## ■ 中文

- Hownet ([http://www.keenage.com/html/e\\_index.html](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
- 大连理工大学 情感词汇本体库 (<http://ir.dlut.edu.cn/EmotionOntologyDownload>)

# 资源：语料（1）

## ■ 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

# 资源：语料（2）

## ■ 中文

- 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

# 评测

- 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

# 主要内容

- ◆ 文本分类
- ◆ 文本聚类
- ◆ 情感分析
  - 相关概念
  - 典型方法



A decorative graphic in the top-left corner consisting of overlapping blue, red, and yellow squares with a black crosshair.

---

# *Thanks*

谢谢!