
情感分析
Sentiment Analysis
徐冰　副教授
哈尔滨工业大学

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

第5章 属性级情感分类

- **5.1** 属性级情感分类方法
  - 基于监督学习的方法
  - 基于词典的方法
- **5.2** 否定和情感
- **5.3** 非观点内容的情感词
- **5.4** 词义消歧和指代消解

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

5.1 属性级情感分类方法

- 基于属性的情感分析（**aspect-based sentiment analysis**），也称观点挖掘
- 在定义**2.1**中把观点定义为四元组（**g,s,h,t**）
  - **G** 观点对象
  - **S** 针对观点对象表达的情感
  - **H** 观点持有者
  - **T** 观点的时间
- 也可以进一步细分，变为五元组（**e,a,s,h,t**）
  - **e** 代表实体；**a** 代表 属性
  - 如：***iphone's voice quality is great.***

    实体是**iPhone**，属性是**voice quality.**

- 基于属性的情感分析有两个重要任务：
  - 1、属性抽取：从文本中抽取所评价的实体和属性
    - 如：*The voice quality of this phone is amazing.*
    - **This phone** 是实体，**voice quality** 是属性。
    - 如：*I love this phone.*
    - **This phone** 就是整体属性。
  - 2、属性情感分析：确定句子中针对不同属性所表达的观点倾向，正面、负面还是中性。
    - *The voice quality of this phone is amazing.* 正面
    - *I love this phone.* 正面

- **5.1.1** 基于监督学习的方法
  - 新的问题：要考虑对于属性级情感分类有效的特征，就是观点评价对象（实体或属性）
  - 两种方法可以解决：
    - 1、生成依赖于评价对象（实体或属性）的特征
    - 2、确定句子中每一处情感表达的作用范围，判断当前情感表达是否包含目标实体或属性。
      - ***Eg: Apple is doing very well in this bad economy.***
      - **Bad修饰economy，不是Apple的情感**

- 针对第一种方法的已有研究工作：主流方法
  - Jiang（2011）基于句法分析树，生成依赖于观点评价对象的特征集合。
    - 假设条件：表征观点评价对象的实体和属性已经被识别出来
    - 上述特征用来表征目标实体、属性词和其它词语之间的句法关系
    - 举例：
      - 假设用w_i表示词语，T表示目标实体或属性。如果w_i是及物动词，T是宾语，可以产生特征$love_arg2$
        - I love the iPhone.
        - iPhone 是目标实体，产生特征$love_arg2$, arg表示论元。

- **Boiy and Moens(2009)** 计算每一个特征词的权重，用以表征该词和目标实体、属性间的距离。
  - 三种权重：
    - 深度差异：特征词和目标实体在句法树中的深度差异
    - 路径距离：句法树看做一个图，特征词和目标实体在深度优先搜索时的距离
    - 简单距离：特征词和目标实体在句子中的距离，不需要进行句法分析

- **5.1.2** 基于词典的方法
  - 基于词典的情感属性分类与篇章级、句子级的情感分类方法都有所不同。
    - 差异在于：需要考虑观点评价的对象，而篇章级、句子级的情感分类方法不用考虑。
  - 基本处理模块或资源包括：
    - 1、使用 包含情感词、短语、俚语、组合规则的情感表达词典
    - 2、处理不同语言和句子类型的规则集（如情感转移和but从句）
    - 3、情感聚合函数，或者是情感词和目标观点评价对象间的句法关系集合
  - 根据以上集合，能够识别出针对每个目标实体或属性所表达出的情感倾向。

- **Ding et al.2008 and Hu & Liu 2004**
  - 假设：目标实体和属性已经确定
  - 具体方法：
    - **1、标记情感表达（词或短语）**
      - 在句子中找到每一处的情感表达，并判断其情感倾向。
      - 正面情感表达**+1**，负面情感表达**-1**
      - **Eg：** ***The voice quality of this phone is not good[+1], but the battery life is long.***
        - **Long**在词典中不是情感词，可以通过上下文推断为情感词
    - **2、处理情感转换词（价转移词）**
      - 指的是能改变情感倾向的词或短语，如否定词 **not、never、none、nobody**等
      - **Eg：** ***The voice quality of this phone is not good[-1], but the battery life is long.***

- 3、处理but从句
  - 转折词或短语通常改变情感倾向，需要专门处理。如 **but, however, with the exception of …**
  - 通常认为：出现在转折词前的观点与转折词之后的观点具有相反的情感倾向。
  - **Eg：** ***The voice quality of this phone is not good[-1], but the battery life is long[+1].***
  - 有的句子虽然含有**but**，但不属于**but**从句类型。
- 4、聚合情感打分
  - 用情感或观点聚合函数来得到情感打分，然后确定句中针对每个属性的观点情感倾向。
  - 假设句子**S**包含属性集合$\{a_1,\ldots,a_m\}$,情感表达集合属性$\{se_1,\ldots,se_n\}$以及通过上述**1-3**步得到的每个情感表达的情感得分。则句子**s**中每个属性a_i的情感倾向可以通过下面的聚合函数得到：
    - $$score(a_i, s) = \sum_{se_j \in s} \frac{se_j.ss}{dist(se_j, a_i)}$$

- 为了使上述方法更有效，不是利用词距离，可以利用情感表达和他们目标之间的关系，包括：
  - （1）句法依存关系
    - 形容词-名词，动词-副词的依赖关系
    - ***eg: His camera takes great pictures.***形容词和名词之间的关系
    - ***eg: I can install this software easily.*** 动词和副词之间的关系
  - （2）情感词自身是目标属性
    - 情感词既表达情感，又指示属性。
    - ***Eg：BMW is expensive.*** ***expensive*** 表达了情感倾向，同时也指示属性是**price**
  - （3）语义关系
    - 语义关系很难识别，也很复杂
    - ***Eg：John admires Jean.*** 观点对象是**Jean**
    - ***Eg: John murdered Jean.*** 观点对象是**John**

小结

- 两种方法的优缺点：
  - 基于监督学习的情感分类方法：
    - 学习算法可以通过从各种特征中自动学习一个分类模型；
    - 依赖于各种领域的训练集，不同领域需要人工标注不同的语料；目前领域自适应方法还不能达到实际应用的需求；
    - 不适合大规模、多领域的实际应用，不能有效处理低频长尾数据。
  - 基于词典的情感分类方法：
    - 从统计学中得到的特征难以利用到基于词典的方法中；
    - 领域独立，鲁棒性强，适合大规模、多领域的实际应用场景，出现错误容易修正；
    - 构建分类所需要的知识库（包括词典、模板、规则等）需要大量人力和物力。

5.2 否定和情感

- 情感转换词（**sentiment shifter**）:通常情况下将当前文本的情感进行反转。否定词就是情感转换词。
- 否定词在句子中的作用范围是问题的关键。
- **5.2.1** 否定词
  - 否定词影响句子情感表达的三种方式：
    - 1、直接否定正面或负面的情感表达
      - ***Eg： This car is <u>not</u> good.***
      - ***<u>Nobody</u> likes this car.***
      - *特殊情况：**I am not angry.** 不能直接反转情感表达，因为它不代表**I am happy.***

- 2、一些没有情感词的句子，通过表达了期望或意愿的文本来表达情感。
  - *Eg： My car does **not** start in a few occasions.*
  - *You can do **nothing** on iPad.*
  - 这类句子很难识别其情感倾向，句子中的否定词也没有对句中的情感进行转换。
- 3、不用情感词就否定了表达期待或不期待意愿的状态。
  - *Eg: The water that comes out the fridge is not cold.*
  - *No bag is used on this vacuum cleaner anymore.*
  - 这类句子很难确定是否表达了期待或不期待的状态，也很难知道特定领域的期望是什么。

- 否定表达的其他类型：
  - 1、比较句中的情感表达
    - 分析下面的例句：比较级和最高级的句子
      - ***(1) This car is not better than my previous car.***
      - ***(2) This car is not the best car in the market.***
      - 句子（1）和（2）有可能没有对汽车的负面评价，但这类句子在实际应用中可能会判断为负面评价。
    - 如果是同级比较，情感极性通常就会反转。
      - ***(3) This car is not as good as my previous car.***
    - 比较级中加入否定词的情况
      - ***(4) Nothing is better than an iPhone.***
      - 否定词没有改变情感极性

- 2、双重否定
  - 双重否定的情感判断更加复杂：*(1) It is not that I do not like it.*
  - Not 后面是名词短语
    - *(2) I hate Audi not Mini.*
    - *(3)She is not a beauty.*
    - (2)句中not没有影响情感表达，（3）句中的not影响了情感表达
- 3、祈使句否定
  - 祈使句通常只是给出命令或请求，一般不表达情感，否定词也不改变情感。
  - Eg：*Do not bring a calculator.*
  - 例外：*Do not waste time on this movie.* 负面评价
- 4、短语或成语中出现否定词，不能当做单独的否定词处理
  - *Eg：not the only, not until, nothing to do, why not……*

- **5.2.2** 否定范围
  - 需要注意：当情感词不在否定词的作用范围内，则该情感词的倾向性也不会因为该否定词而发生反转。
    - ***Eg： I did not drive my car on that horrible road.***
    - 这个句子中**not**不修饰**horrible**
  - Jia(2009)基于句法的处理规则确定否定词范围
    - 定义了否定词和它之后的其他词之间的词语间隔，最关键的是确定这种间隔的结尾
    - 基本原则是否定词作用范围不应跨越其所在的从句，额外定义了其他规则。
    - 指出了否定词不能处理的特殊情况：**not only 和 not just**
- 否定词是很难处理的问题，有的否定句包含上述情况，但仅是描述了客观陈述，没有任何观点信息，有时候需要特定的领域知识才有可能正确处理。

5.3 非观点内容的情感词

- 有一些情感词是不表达情感的，这类情况会影响情感分类系统的判断。
  - **1**、实体名中包含情感词
    - 如：保险公司名 **Progressive** , 电商公司名 **Best Buy,** 好莱坞电影名 **Pretty Woman**
  - **2**、功能名中包含情感词
    - 如：视频播放器的前进键 **fast forward,** 后退键 **fast rewind,**
      美容的基本步骤 **beauty treatment**
  - **3**、祝贺和祝福
    - 如：***good morning, good day, happy birthday, best regards***
  - **4**、作者的自我评价
    - 如： ***I know Lenovo laptops very well.***
    - 这是作者对作者自身的评价，没有对产品发表观点
    - 这类句子会出现在论坛评论中，作者可能是专家或有经验的用户，他们通常会回答问题并提供建议，这类句子不容易处理。

- 还有一些文本表达的句子，也没有表达任何观点信息。

(1) ***I am not sure whether the iPhone is the best phone for me or not.***

- 表达了作者的不确定性

(2) ***I am looking for a good iPhone case.***

行为意图

(3) ***No insurance means that you have to pay high cost.***

- 普遍事实

(4) ***Buy this great camera and win a trip to Hawaii.***

- 商业广告

5.4 词义消歧和指代消解

- 对于情感分析中的**NLP**核心问题的研究，目前相关研究很少。
- 对于评价对象识别任务，指代消解的作用非常重要。
- 如果观点评价的对象与观点表达文本不在同一个句子中出现，就需要指代消解获取不同句子中的共指关系。
- 词义消歧的已有研究：
  - **Akkaya(2009):**主观词词义消歧 **(subjectivity word sense disambiguation)**

- **Akkaya(2009):主观词词义消歧 (subjectivity word sense disambiguation)**
  - 动机：有些词既有主观也有客观的含义，主观词用于客观信息表达是观点挖掘与情感分析中错误的主要来源。
  - 任务：判别语料库中的词是表达了主观的语义信息还是表达了客观的语义信息。
  - 方法：利用上下文信息，来消除主观词典中的词在具体句子中的主客观歧义问题。
  - 实验表明：**SWSD**方法对确定分析文本的主客观性及蕴含情感的倾向性有帮助。

- 指代消解：确定句子、篇章中多个文本表达的语义同指对象。
  - ***Eg：I bought an iPhone two days ago. It looks very nice. I made many calls in the past two days. They were great.***
  - **It 指代iPhone（实体），they 指代calls（属性，具体的就是通话能力）**

- 已有研究：
  - **Ding and Liu(2010):**提出了一种监督学习方法解决实体和属性级的指代消解问题

- **Ding and Liu(2010):**提出了一种监督学习方法解决实体和属性级的指代消解问题
  - 文章的贡献：
  - 设计了两个与情感有关的有效特征，利用情感分析结果帮助指代消解：
    - （1）基于在普通句和比较句上的情感分析结果。
      - ***Eg: The Nokia phone is better than this Motorola phone. It is cheap too.***
      - **It 指代 Nokia，表达了对Nokia的正面情感**
    - （2）考虑哪个情感词具体评价了句子中的实体和实体属性。
      - ***Eg： I bought a Nokia phone yesterday. The sound quality is good. It is cheap too.***
      - **It 指代 Nokia phone 还是 sound quality？要明确目标情感词用来评价哪些实体和实体属性。**

本章小结

- 属性级情感分析包括两个关键任务：
  - 1、属性抽取（下一章介绍）
  - 2、属性级情感分类（本章介绍）
- 属性级情感分类方法：
  - 基于有监督学习的方法
  - 基于词典的方法（无监督）
  - 基于规则的方法（规则不易应用，不容易描述）

- 目前的现状是大部分系统仅利用了句子中的情感词，结合简单的句法分析，对句中的情感信息进行分析，缺乏对复杂句的有效处理，也缺乏对表达观点的事实性句子的分析。
  - 长尾问题（long-tail）：利用情感词能处理60%的情况，剩下的40%是低频的语言现象。
  - 领域问题：主要集中在电商、旅馆、餐厅等领域。事实是情感词在不同领域中的情感表达有差别，需要编译不同领域的领域词典。
  - 数据噪声：社会媒体文本包含大量噪声，充满各种拼写、语法等错误，需要进行预处理工作。
- 比较有效的方法是：将统计学习模型和领域语言知识相结合！

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

第6章 属性和实体抽取

- 6.1 基于频率的属性抽取
- 6.2 利用句法关系
  - 6.2.1 利用观点和观点评价对象间的评价关系
  - 6.2.2 利用部分整体和属性关系
- 6.3 基于监督学习的属性抽取
- 6.4 隐含属性的映射
  - 6.4.1 基于语料库的方法
  - 6.4.2 基于词典的方法
- 6.5 属性聚类

6.1 基于频率的属性抽取

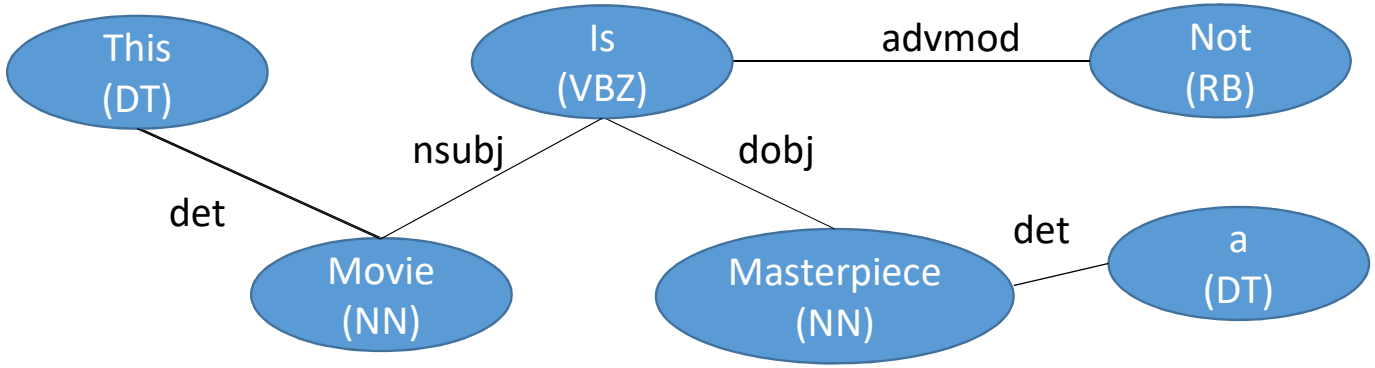
- 目标：通过在特定领域评论中大量出现的名词、名词短语的频率统计操作，对其中的实体属性信息进行识别和抽取。
- **Hu and Liu（2004）**利用关联规则进行属性词挖掘。
  - 基于的假设是：频繁出现的名词（名词短语）通常就是那些重要的属性词，那些不相关的噪音在不同评论中具有差异性，相对于真实属性词出现的频率低。
  - 方法：利用**POS**标注器在句子中识别名词（名词短语），然后用数据挖掘算法记录他们出现的频率，进而通过实验确定阈值，保留大于阈值的名词（名词短语）。
  - 实验结果：在有相当数量的评论语料库中，这些评论都是针对同一产品或同类型产品，这种方法很有效。

- **Popescu and Etzioni(2005):**通过计算候选短语和那些具有部分-整体关系的实体类的互信息（**PMI**）得分，对候选名词进行筛选。
  - 部分-整体关系的短语：**of camera, camera has, camera come with**
  - 这个想法的动机：那些与表示某类产品部分-整体关系的短语经常一起出现的名词或名词短语很有可能就是正确的属性词。
  - 利用**PMI**计算词出现和共现的频率信息，**PMI**值过小，说明不会共现，就可能不是产品的组成部分。
    - $PMI(a,b)=hits(a \cap b)/hits(a)hits(b)$
- **Blair-Goldensohn (2008):**提出了一种高频词的过滤方法，在处理过程中主要考虑那些只出现在情感句或出现在表示情感信息的句法模板中的那些名词或名词短语。

6.2 利用句法关系

- 在观点句中，情感词和观点评价对象之间会存在多种句法关系来表征它们之间的评价或修饰关系。
- 利用句法关系可以进行实体属性和实体的抽取，也可以用于情感词的获取。
- **6.2.1** 利用观点和观点评价对象间的评价关系
  - 基本思想：情感词通常在句子中用来评价属性。若句子中没有高频属性，但出现了情感词，则这些情感词周围出现的最近的名词或名词短语可以看做属性。
  - **Hu and Liu（2004）**：利用最近邻关系粗略指示了情感词及其修饰名词和名词短语之间的依存关系，而且效果不错。
    - ***Eg. The software is amazing.***

- **Zhuang et al.2006, Lin 2007, Somasundaran and Wiebe 2009,Kobayashi 2006, Qiu 2011:** 利用句法分析来识别情感词和观点评价对象之间的依存关系，进而抽取属性词-情感词对。
  - ***Eg. This movie is not a masterpiece.***
  - 这句话的句法关系：属性词和情感词的依存关系模板是“**NN-nsubj-VBZ-dobj-NN**”



- 双向传播DP算法（Qiu et al.2009,2011）
  - 利用以上的句法关系，可以通过已知的属性词抽取更多的情感词，也可以通过已知的情感词抽取更多的属性词。
  - 按照这种方式，每轮迭代，抽取出来的情感词和属性词可以进一步用来抽取新的词，直到没有新词时，迭代结束。
  - 整个传播过程包括四个子任务：
    - 1、用情感词抽取属性词
    - 2、用抽取的属性词抽取属性词
    - 3、用抽取的属性词抽取情感词
    - 4、用抽取和实际给定的情感词抽取情感词

- **DP**方法最早应用于英文评论中的属性词、情感词抽取任务。
  - **Zhai(2011)**将该方法应用在中文数据处理上，效果很好。
  - **Xu(2013)**对**DP**的改进：
    - 过滤掉高频的、表示一般性概念的错误抽取的观点评价对象；
    - 挖掘长尾、低频的观点评价对象；
    - 检测不是情感词的形容词，如**many**等

6.2.1 利用部分整体和属性关系

- 用句法规则表示部分整体和属性关系，最常用的规则是基于所有格。
- 英语中有两种所有格：**’s** 和 **of**
  - ***The voice quality of the iPhone*** *（属性关系）*
  - ***iPhone’s battery*** *（部分整体关系）*
  - ***iPhone’s price*** *（属性关系）*
- 在情感分析任务中，一般不需要确定所有格所属的语义关系，抽取出来就可以。
- **Zhang** 等（**2010**）提出包含了**of** 所有格但未包含**’s**所有格,利用这个规则抽取。
  - **NP(**属性词**) Prep CP(**类别概念名词**):** ***battery of the camera***
  - **CP with NP：** ***mattress with a cover***
  - **CP NP:** ***car seat***
  - **CP Verb NP:** 动词包括**has, have, include, contain, consist, comprise**等，
    如: ***The phone has a big screen***

- 如果去除抽取属性词过程中的噪声？
- **Zhang(2010):**使用**DP**方法抽取了属性词的候选，然后利用属性重要性对所有候选进行排序。
  - 决定属性词重要性的因素有两个：属性相关性 和 属性词出现的频率
    - 属性相关性的判别：
      - （1）如果该属性词被越多的情感词修饰，那么越有可能是一个属性词；
      - （2）如果一个属性被多个句法规则匹配并抽取出来，那么它也可能是一个属性词；
      - （3）如果一个句子里的属性词候选与情感词之间具有修饰关系，同时也匹配句法规则，那么也有可能是一个属性词。
  - 文中还提到情感词、句法规则和属性词间存在相互增强的关系，即一个形容词如果修饰了很多属性词，那么它很可能是一个情感词。
  - 如果一个属性词候选通过许多情感词和句法规则被抽取出来，很可能是一个属性词。

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY

6.3 基于监督学习的属性抽取

- 最有效的序列学习方法
  - 隐马尔可夫模型**(HMM)** 和 条件随机场**(CRF)**
- **6.3.1** 隐马尔可夫模型
  - 基本方法：给定一个观察序列，通过学习优化**HMM**的模型参数，使得观察概率最大化，尽可能拟合训练数据。在此基础上，利用学到的模型，可以为新的观察序列找到最优的状态序列。
  - 在属性抽取任务上，把单词或词组当做目标观察，属性词标签或情感词标签当做潜在的状态。
  - **Jin and Ho(2009):**提出了词优化的**HMM**模型，从评论文本中抽取属性词以及情感词。

• 6.3.2 条件随机场

- HMM的局限性在于其假设与实际问题不匹配，通常会导致准确性降低。
- CRF 链式随机场模型是一种无向的序列模型，相较于HMM会引入更多的特征。
- Jakob and Gurevych(2010):利用CRF从包含观点表达的句子中提取观点评价对象，提出了如下特征：
  - Token 词形特征
  - Part of speech 词性特征
  - Short dependency path 当前词与情感词之间直接的依存路径
  - Word distance 离情感词最近的名词或名词短语的距离
- 通常用Inside-Outside-Begin(IOB)标签指示观点评价对象。
  - B-target表示观点评价对象的开始；I-Target 表示观点评价对象中的词；O表示其他非观点评价对象的词。

- **Li等(2010):** 两种基于**CRF**的改进模型 **Skip-chain CRF** 和 **Tree-CRF**
  - **Skip-chain CRF** 模型能在句子层面对基于连接词的长距离词间依存关系进行建模
  - **Tree-CRF** 能对属性词与正面情感词和反面情感词间的深层句法依存关系进行建模。
  - 与传统的**CRF**模型只利用词序列信息相比，这两个模型都能利用句子的结构特征
- 其他监督学习方法：
  - **Kobayashi** 等**(2007):** 基于树结构的分类模型，特征包括上下文特征、共现特征等
  - **Yu** 等**(2011):** 部分监督学习算法单类支持向量机，从评论中抽取属性词
  - **Zhou** 等**(2013):** 给出了一种无监督标签传播方法，从中文微博中提取观点评价对象。

6.4 隐含属性的映射

- 显式属性词：表达为名词和名词短语的属性
- 隐式属性词：形容词和副词是最常见的类型，动词也可能成为属性词。
  - 如：**expensive** 描述的是**price**，**beautiful** 描述的是 **appearance**
  - ***This camera will not easily fit in a pocket***. 属性表示的是体积
- **6.4.1** 基于语料库的方法
  - **Su** 等**(2008)**提出了一种基于聚类的方法，利用评论句中显式属性词和情感词组成的语义关系，能够将形容词形式的隐含属性表达映射到显式属性上。

- **Hai** 等**(2011)**给出了一种二阶段的基于共现信息的关联规则挖掘方法，将隐式属性词映射到显式属性。
  - 第一阶段：利用情感词和显式属性词的共现信息产生一些关联规则，其中情感词作为规则条件，显式属性作为规则的结果
  - 第二阶段：针对每个情感词，对于包含显式属性词的规则进行聚类，生成更鲁棒的规则。
- 基于语料库的抽取方法的缺点：
  - **1**、很难发现那些由于语言习惯而不会和情感词共现的属性词。
    - **Eg.** ***iPhone is expensive*****.** 隐含属性是**price,** 很难将**price**识别为**expensive**所映射的属性。
  - **2**、即使一个形容词和它的一个属性名词在语料库中同时出现，如果语料库大小有限，在其他句子中它们可能不会同时出现。

- **6.4.2** 基于词典的方法
  - **Fei** 等**(2012):** 提出了基于词典的方法，在词典中利用属性词对形容词进行定义。
    - 在**thefreedictionary.com** 中，**expensive**被定义为“**Marked by high prices**”
    - 在处理过程中，就可以根据形容词的定义发现其所应该映射的属性词，这样解决了基于语料库方法的第一个缺点。
    - 另外，由于词典不会被限定到任何特定的语料库中，词典中的每个形容词都可以单独学习，解决了第二个缺点。
    - 采用基于协同分类的关系学习方法，利用字典已标注丰富的词间语义关系，用以识别隐式的属性词。

6.5 属性聚类

- 属性聚类最直接的方法是用**WordNet** 或同义词词典的资源来获取同义的文本表达，但这个方法效果并不好，其原因如下：
  - **1**、许多同义词都是和领域信息高度相关的
    - 如**movie** 和 **picture**在影评中是同义词，在摄影评论中，没有同义关系。
  - **2**、许多属性的文本表达是短语，很难在已有的词典中找到定义
  - **3**、许多描述相同属性的文本表达通常不是在通用领域，而是在特定领域内具有同指关系。
    - 如**expensive**和**cheap**是指属性**price**，但却是反义词
  - **4**、在实际应用中，属性聚类任务不能完全通过无监督学习的方法解决，因为这个任务的主观性很强。不同的用户，有不同的类别划分和定义。

- 解决方法：
  - **Carenini 等(2005):**通过使用字符串相似度、同义词以及基于**WordNet**的词汇距离等语义相似度计算方法，来获取属性文本表达间的同义关系。
    - 在特定领域，需要事先给定分类体系
    - 在相机和**DVD**的评论文本上，这个方法得到不错的效果。
  - **Yu 等(2011):**提出基于已有的商品属性类别体系，对商品评论文本进行分析，利用多个相似度计算策略的优化组合，计算不同属性文本表达之间的语义相似度，最终自动生成一个新的商品属性类别体系。
  - **Zhai 等(2010):** 提出了一个半监督学习方法，把属性文本表达分组到用户给定的属性类别体系上。
    - 需要事先手动为每个属性类别标注种子词，系统然后基于半监督学习方法把其他属性文本表达映射到合适的类别中，在半监督学习中使用了期望最大化方法（**EM**）

本章小结

- 属性和实体抽取以及消解对于情感分析非常重要，他们表达了观点评价对象。
- 在许多领域，已有的方法在抽取精度上依然不高。
  - 目前的方法都以抽取名词和名词短语类型的属性，对通过动词表达的，已有方法无法适用。
- 研究者提出的无监督和半监督的主题模型，在实际应用中不够精确，实际生活中很多属性都是短语构成的，仅用一元模型还是不够的。
- 常识知识和领域知识对于属性和实体抽取非常重要。