《基于CRF和句法分析的中文微博情感分析》的使用方法讲解

2014-06-27 黑龙江大学自然语言处理实验室

大家好，这里是“**黑龙江大学自然语言处理实验室**”。我们将成为大家了解科研，了解自然语言处理的一个很好的途径。如果大家有什么意见或者看法，都可以和我留言的。欢迎大家提问，多多互动~ 今天我们将继续发布有关情感分析的内容，希望大家喜欢~

1引言

情感分析，又称为**意见挖掘**，是指通过自动分析来获得对于某件事物的褒贬意见。

微博，它是一种限制文本长度的工具，大多数人在微博上发的内容一般都是以短文本的形式。但是其中也存在比较长，句子结构比较完整的句子。短文本就是内容较短的文本（一般长度不超过160字符），通常以新闻标题、微博、手机短信、电子邮件、购物评价等形式存在。有效的短文本情感倾向分析技术可以帮助我们在海量信息中更准确地获取自己感兴趣的信息。

2工具介绍

条件随机场（Conditional Random Fields，**CRFs**）最早由Lafferty等人在2001年提出的，其模型思想组要来源是最大熵模型，模型的三个基本问题的解决用到了HMMs模型中提到的方法如向前向后算法和维特比算法。条件随机场可以看成是一个无向图模型或马尔科夫随机场，它是一种用来标记和切分序列化数据的统计模型。该模型是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率，而不是在给定当前状态条件下，定义下一个状态的标记序列的分布。标记序列的分布条件属性，可以让CRFs很好的拟合现实数据，而在这些数据中，标记序列的条件概率信赖于观察序列中非独立的、相互作用的特征，并通过赋予特征以不同权值来表示特征的重要程度。

Standford Parser是由斯坦福自然语言处理小组研发的主要针对英语的句法分析器，但是逐渐在各种不同语言中完善。该句法分析器对分词后句子中的每个词，进行词性标注，并且进行序号标记，遇到句号则从头开始标记，并且对词进行词性标注，然后词与词之间产生依赖关系。

3 方法介绍

该文对于微博情感倾向性识别所使用的方法分别是**CRF**和**句法分析**。

3.1基于CRF的方法

3.1.1 流程图

圓1 堀五CRFs旂泛皊浇稓圓

3.1.2方法讲解

将短文本的每个词作为第一列，将短文本的情感倾向性作为序列标注的第二列。如图2所示，这样短文本就转化为一个标注后的序列，可以用于训练。测试的短文本只需给出每个词作为第一列，第二列文本类别为空，留待预测。结果预测过程如图3所示。

图2训练语料格式图

圓3 给枢颊浑圓

3.2 基于句法分析的方法

3.2.1 流程图

圓4 堀五叫泛刂枟旂泛皊浇稓圓

3.2.2 方法讲解

文章首先使用情感词词典从依赖关系中识别出情感词，但是只依靠情感词并不能识别出句子的倾向性，还得要根据属性词才能知道情感倾向性，即只有同时知道了属性词和评价词才能知道这个评价对的情感倾向性，与此同时，若知道了修饰评价词的副词则能够知道这个评价对的情感程度。基于以上分析，要分析情感倾向性，首先就得要识别出属性词和评价词。基于句法分析的方法实际上是在句法分析的基础上利用自己定义的规则来实现属性评价对的抽取，然后根据倾向性分析模块来分析其倾向性。其识别属性和评价词以及修饰评价词的副词的规则定义如下：

首先为了减少句法分析器错误依赖关系对的影响，定义窗口长度为6，如果依赖关系对中的两个词距离大于6，认为是没有效果的依赖对，不予考虑。

其次，可以直接通过情感词以及依赖关系确定评价对象的情况：

(1)首先检验所有的依赖关系中的情感词，如果情感词中出现在nsubj关系对中，并且出现在关系对的左边，那么找到关系对右边的词语，这个词语必然是这个情感词语修饰的对象

例子：空间 很 大
nsubj(大-3, 空间-1)
advmod(大-3, 很-2)
root(ROOT-0, 大-3)
“大”是情感词，“空间”是被“大”修饰的对象。

(2)找到这个修饰对象所在的关系对，是否存在nn依赖关系，如果存在，那么nn依赖对中的两个词语一次合并成完整的修饰对象。
例子：悬挂减震 方面 很 沉稳 干净利落
nn(方面-2, 悬挂减震-1)
nsubj(沉稳-4, 方面-2)
advmod(沉稳-4, 很-3)
root(ROOT-0, 沉稳-4)
conj(沉稳-4, 干净利落-6)
“悬挂减震”和“方面”存在nn依赖关系，这两个词合并成完整的修饰对象“悬挂减震方面”。

(3)查找该情感词语的其他依赖关系对，如果存在advmod结构，并且是情感词语出现关系对的左边，那么右边的词语就是修饰这个情感词的副词，我们找到这个副词，并且做之后的副词程度匹配。

(4)继续寻找advmod依赖对，有时候往往会存在很多副词连续修饰一个情感词的情况，我们找到所有的修饰副词。
例子：新 家族 脸谱 果然 很 犀利 啊
amod(脸谱-3, 新-1)
nn(脸谱-3, 家族-2)
nsubj(犀利-6, 脸谱-3)
advmod(犀利-6, 果然-4)
advmod(犀利-6, 很-5)
root(ROOT-0, 犀利-6)
dep(犀利-6, 啊-7)
“犀利”这个词为评价词，其存在advmod依赖关系的词有“果然”和“很”，所以“果然”和“很”都是修饰“犀利”，则“果然很犀利”可以构成一个评价语句。

(5)继续查找情感词的依赖对，如果存在neg依赖对，那么情感发生变化，前面的情感词语认为前面加了否定的副词修饰。
例子：油耗 并 不 高 ，动力性 很 好 ， 很
适合 年轻人 。
nsubj(高-4, 耗油-1)
advmod(高-4, 并-2)
neg(高-4, 不-3)
root(ROOT-0, 高-4)
nsubj(好-8, 动力性-6)
advmod(好-8, 很-7)
conj(高-4, 好-8)
advmod(适合-11, 很-10)
conj(高-4, 适合-11)
dobj(适合-11, 年轻人-12)
属性词“油耗”和评价词“高”搭配时其情感倾向为贬义，则查找到依赖关系对neg(高-4, 不-3)后其情感倾向发生变化，由原来的贬义变成了褒义。

(6)检索是否存在dobj依赖对，如果存在，那么我们认为右边的词语是动宾结构的宾语，继续查找该宾语的依赖对中是否含有nn的结构，如果存在，那么nn依赖对中的词语将合并成为评价的对象，用于之后的匹配。
例子：弯道 照明 不错 吆（截取句子的一部分）
nn(照明-45, 弯道-44)
dobj(满意-43, 照明-45)
conj(有-16, 不错-47)

comod(不错-47, 吆-48)
首先查找到依赖关系dobj(满意-43, 照明-45)，然后查找到nn依赖关系nn(照明-45, 弯道-44)，所以“弯道照明”是一个完整的评价对象。

(7)检索是否存在比字结构，如果情感词出现在prep依赖对的左边，并且伴随着pobj依赖对的出现，我们认为这个情感词语修饰的是一个比字结构，pobj依赖对的右边的词虽然不是直接形容评价对象的词语，但是，对于比字结构，一定是一个同类的评价对象，属于隐式的评价对象之一。
例子：就是 使用 起来 比 触摸屏 的 麻烦 些
advmod(使用-3, 就-1)
cop(使用-3, 是-2)
root(ROOT-0, 使用-3)
rcomp(使用-3, 起来-4)
assmod(麻烦-8, 比-5)
pobj(比-5, 触摸屏-6)
assm(比-5, 的-7)
dobj(使用-3, 麻烦-8)
range(使用-3, 些-9)
根据这条规则可知“触摸屏”也是一个评价对象。

无法从以上依赖关系对中找到评价对象的情况，采用以下两种方式来判别(1) 对于短句中存在匹配到情感词语，但是无法用上面七条规则找到评价对象，我们向前寻找最近的标点符号，并且找到该标点的punct依赖对，并且有连接词出现在依赖对的左边，那么我们认为该情感词形容的评价对象是离它最近的一个之前的评价对象(2) 对于短语中没有匹配到情感词汇，但是存在punct依赖对，并且连接词可以在递进，转折等连接词列表中匹配到，那么如果是转折关系，我们将之前最近的一个评价对象的倾向性做相反方向的处理，如果是递进关系则更加强调之前的结果。

参考文献:
http://www.docin.com/p-601678509.html?qq-pf-to=pcqq.c2c#userconsent#