基于语义特征和本体的语篇情感计算

阅读报告

1. 本文主要思想：以情感词汇本体为基础，首先识别单句的情感，根据情感表达方式的不同将单句化为3种类型，并从中抽取多个语义特征加入到条件随机域中，进而获取整篇的情感结构。
2. 情感词汇本体：

情感词汇本体通过： Lexicon = (B, R, E) 进行描述。其中：

B: 词汇的基本信息，包括： 编号，词条，对应英文，词性，录入者 ，版本信息

R: 词汇间的同义关系，即该词汇与哪些词汇有同义的关系

E: 词汇情感信息

1. 情感分析模型
   1. 句子级的情感分析：对句子采用关键词识别的方法分析句子的情感类型：
      * 1. 单情感词的简单句
        2. 多情感词的简单句
        3. 多情感词的复杂句
   2. 句篇型的情感分析：指将每个句子都放到整篇文章大范围的语境中考虑它的情感色彩。通过每个句子的标注最终得到文章的情感结构。这篇文章利用条件随机域来实现语篇中每个句子的情感分析

因为文章中的句子的重复几率不像词语那么多，因此为了防止特征稀疏，更好地代表句子中的情感信息，本文选择使用：

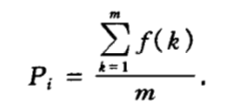
* + - 1. 句子词汇方面的特征：情感词汇特征（情感词汇本体），情感类别特征（出现频率最高的一种情感类别），否定词特征，句子间连接词特征
      2. 句子结构方面的特征：句类特征（如陈述句，祈使句等），是否为文章的第一句，句子长度，句子的类别，句子重叠程度的特征（如‘救救我’等，通过重叠率体现）

两个方面的语义特征代表句子的情感信息。

1. 文章中对于情感分析结果的评估方法：

（1）单句评估法：

以句子为分割单位，一篇文章中标注的正确的句子数除以文章包含的句子总数：

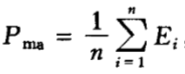


若第k句标注正确，函数f(k)的值为1，否则为0，m为句子总数。

Pi为语篇i情感分析的准确率

而CRFs标注的结果总正确率使用宏平均与微平均两种方法衡量：

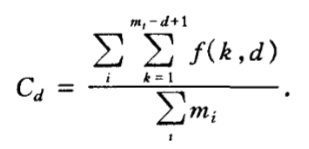
1. 微平均指用所有标注正确的句子除以句子的总数
2. 宏平均指求各个情感类别准确率的平均值



其中，Pma表示宏观准确率，Ei表示第i类情感的准确率，n表示情感类别数

（2）多句联合评估法：

以二元以上共现的词汇代表目标语言的流利程度，本文采用共现的多元情感在标准标注中匹配的准确率来衡量情感标注的连贯性：



式中， Cd表示d个连续句子的情感标注准确率，

函数f表示文档中从k开始取d个标注的情感句与正确情感标注匹配，如果全相等则为1，反之为0，

I表示为第i篇文档，

mi表示第i篇文档的句子总数

1. 总结：

这种利用条件随机场进行句篇情感分析的方法比起仅使用从句中提取情感词频最高的情感类型作为语句的情感，进而得到整篇文章的情感标注方法得到的结果的准确性有着显著的提高。

1. 条件随机场
   1. 条件随机场是马尔可夫随机场的特例，马尔可夫随机场假设随机场中的某一个位置仅与与它相邻的位置的赋值有关，与其不相邻的位置的赋值无关。。
   2. 条件随机场假设马尔可夫随机场中只有X和Y两种变量（如，X为词，Y为词性），X一般是给定的，而Y一般是在给定X条件下的输出。数学描述为：

设X与Y是随机变量，P(Y|X)是给定X时Y的条件概率分布，若随机变量Y构成的是一个马尔可夫随机场，则条件概率分布P(Y|X)称为条件随机场。

* 1. 如果假设X和Y有相同的结构（如在文章中，词与词性有相同的结构），则：

X=(X1,X2…Xn), Y=Y(Y1,Y2…Yn)

则称这样的条件随机场为线性链条件随机场。这样的随机场满足马尔可夫性：

P=(Yi|X,Y1,Y2…Yn)=P(Yi|X,Yi-1,Yi+1)

* 1. 线性链条件随机场的参数化形式：

在线性链条件随机场模型中，特征函数分为2类：第一类是定义在Y节点上的节点特征函数，这类特征函数只与当前节点有关：

sl(yi,x,i),l=1,2,...L

其中，L是定义在该节点特征函数的总个数，i是当前节点在序列的位置

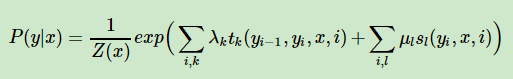
另一类是定义在Y上下文的局部特征函数，这类特征函数只和当前节点和上一个节点有关：

tk(yi−1,yi,x,i),k=1,2,...K

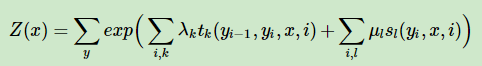
其中，K是该节点的局部特征函数的总个数，i是当前节点在序列的位置

这两种特征函数，它们的取值只能是0或者1。

假设tk的权重系数是λk, sl的权重系数是μl，则线性链条件随机场（linear-CRF）的参数化形式如下：



其中，Z(x)为规范化因子：



1. 复现文章工作内容需要进行的内容：
2. 数据提取与数据整理：

已经用网络爬虫爬取到了新浪微博的博文及今日头条的文章，在保存前进行了简单的前期处理。还需要对数据库中的素材进行进一步的处理，将错误的格式和标点去除。

问题：今日头条文章以新闻为主，情绪化的内容较少，因此在情绪判断方面可能样本不够丰富；而微博素材大都比较简短，在短句的判断上用处较大，但在对长篇的文章进行情感判断上样本量较小。下一步尝试计划爬取网络小说或散文等内容。

1. 对句子进行标注：

根据和靳浩林的沟通，计划做一个人工标注平台进行人工标注，获得正确标注的样本。

3）对训练样本进行情感计算并与正确标注的样本进行对比，获得训练结果的正确率等信息