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：

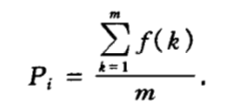
* + - 1. 句子词汇方面的特征：情感词汇特征（情感词汇本体），情感类别特征（出现频率最高的一种情感类别），否定词特征，句子间连接词特征
      2. 句子结构方面的特征：句类特征（如陈述句，祈使句等），是否为文章的第一句，句子长度，句子的类别，句子重叠程度的特征（如‘救救我’等，通过重叠率体现）

两个方面的语义特征代表句子的情感信息。

4．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：

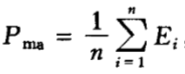


若第k句标注正确，函数f(k)的值为1，否则为0，m为句子总数。

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：

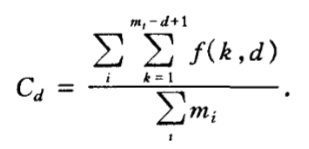
1. 微平均指用所有标注正确的句子除以句子的总数
2. 宏平均指求各个情感类别准确率的平均值



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