099 | 如何来提取情感“实体”和“方面”呢?

2018-05-21 洪亮劼

讲述：初明明
时长 05:39 大小 2.59M

从上一篇分享开始，我们转入文本分析的另外一个领域，文本“情感分析”（Sentiment Analysis），也就是指我们要针对一段文本来判断其文字“色彩”。文本情感分析是一个非常实用的工具。我们从最基础的文档情感分类这个问题说起，这个任务是把一个单独的文档给分类为某种情感。在绝大多数情况下，我们可以把这个任务看作监督学习的问题。另外，我们也聊了聊如何通过建立情感词来进行简单的非监督学习的步骤。

今天，我们来看文本情感分析中的另一个关键技术，**情感“实体”和“方面”的提取**。

“实体”和“方面”的提取

对于文本情感分析而言，“**实体**”（Entity）和“**方面**”（Aspect）是两个非常重要的概念。很多情感分析的任务都是围绕着这两个概念而产生的。在谈论如何对这两个概念提取之

前，我们先来看看这两个概念的意义。

“实体”其实就是文本中的某一个对象，比如产品的名字、公司的名字、服务的名字、个人、事件名字等。而“方面”则是实体的某种属性和组建。

比如这么一个句子：“我买了一部三星手机，它的通话质量很不错”。在这里，“三星手机”就是一个实体，而“通话质量”则是一个方面。更进一步，“很不错”则是一个情感表达，这里是针对“三星手机”这个实体的“通话质量”这个方面。很明显，如果我们想要精准地对文本的情感进行分析，就一定得能够对实体和方面进行有效提取。

从广义的范围来说，实体和方面的提取都属于“**信息提取**”（Information Extraction）的工作。这是一个非常大的任务类别，用于从大量的非结构化文本中提取出有价值的信息。实体和方面的提取可以利用一般性的信息提取技术，当然往往也可以利用句子中的一些特殊结构。

常用的提取技术

接下来，我们来聊一聊有哪些最直观最简单的提取技术。

第一种最简单的技术是**基于“频率”（Frequency）的提取**。在这样的技术中，我们先对文本进行“词类”（Part Of Speech）分析，分析出每个词的词性。然后主要针对句子中的“名词”，计算这些“名词”出现的频率。当这些频率达到某一个阈值的时候，我们就认为这些名词是一个实体或者方面。

这里的假设是，在一个例如产品评论的文本集合中，如果一个名词反复出现在这个集合的很多文档中，那么这个名词很有可能就是一个独立的实体或者方面。为了达到更好的效果，更加复杂的词频技术，例如 TF-IDF 也经常被用在计算名词的频率上，从而提取它们作为实体和方面的候选词。

另一种比较常见的针对情感分析开发的技术，就是**利用句子中的一些特殊的结构从而达到信息提取的目的**。

比如，回到刚才的那句话：“我买了一部三星手机，它的通话质量很不错”。在这句话中，“很不错”作为一个情感词汇，一定和某一个方面，甚至是某一个实体成对出现的。那么这个**成对出现的情况**就是我们可以利用的情感句子的有利特征。

比如“很不错”这个词汇，在一个描述产品情感的文档中，这个词汇很少单独出现。这类不管是褒义还是贬义的词汇出现后，在绝大多数情况下，他们都会描述一个对象。而从句法结构上来说，这个对象往往又离这个情感词汇很近，因为这个情感词需要对这个对象进行描述。因此，我们就可以利用这种配对结果，来计算这样的结构是否大量出现。

这种结构其实可以被反复利用。例如在刚才的句子中，“三星手机”这个实体，一定会和很多不同的方面反复同时出现，如“通话质量”、“操作”、“售后服务”等。我们可以利用这两种不同的配对结构，实体和方面之间的，方面和情感词之间的，更好地提取这些词汇。

刚才我们说的不管是基于词频的还是利用配对关系的方法，都可以算是**无监督的学习方法。**这些方法的本质，其实就是**利用某种之前定义好的规则或者是某种洞察**来针对文本进行提取。另外一种思维其实就是**把信息提取转换成为监督学习任务。**

回到例子“我买了一部三星手机，它的通话质量很不错”这句话。这句话的文本作为输入，我们需要的输出是“三星手机—实体”、“通话质量—方面”这样的标签信息。那么，一个基本的想法就是，我们其实可以针对这句话构建一些特征，然后学习出一个分类器，从而可以得到这样的标签。

值得注意的是，这一类的监督学习任务和我们常见的例如分类一个文档是不是垃圾信息不一样，这里我们**需要输出多个标签。**这种需要输出多个标签的任务，特别是这些标签之间可能还有一定关系的情况，往往被称作是“**结构化预测**”（Structural Prediction）任务。

在结构化预测这个领域，“**条件随机场**”（Conditional Random Field），或者简称是**CRF**的模型，是对这方面任务进行运作的一个经典模型。然而，需要指出的是，把实体和方面提取当作监督任务以后，很明显，我们就需要有一个训练集和标签，这个训练集的匮乏常常成为 CRF 产生理想效果的瓶颈。

总结

今天，我为你介绍了一类基础的文字情感分析任务——情感“实体”和“方面”的提取。

一起来回顾下要点：第一，我们介绍了什么是情感“实体”和“方面”；第二，我们聊了目前在这个方向上比较通行的一些方法，比如基于“频率”的提取，利用句子的一些特殊结构等。

最后，给你留一个思考题，除了我们介绍的这些方法，你还能想到其他方法来提取实体和方面的关键词吗？

欢迎你给我留言，和我一起讨论。

上一篇　098 | 什么是文档情感分类？

下一篇　100 | 文本情感分析中如何做意见总结和搜索？

精选留言 (2)

写留言

雨天防火
2019-05-19

有没有相关的文献或者更详细的介绍教程？想深入理解一下这方面算法实例

展开

散人

2018-05-23

老师好，文章中提到的基于特定结构提取方法，有没有特定的算法，还是完全自定义的方式，如在特定词前后几个词作为候选实体，属性这样？谢谢老师。