**1. 引言**

情感分类随着网络评论的海量增长而迅速兴起,其研究价值和应用价值受到人们越来越多的重视。情感分类系统通常依赖于标注语料并结合分类算法来实现。然而，情感标注语料的分布在不同语言下是极不均衡的。因此在当前语言的标注语料缺乏时，利用其他语言的资源来实现情感分类已经成为了一个热门的研究课题。第二届自然语言处理与中文计算会议（NLP&CC 2013）的评测任务设立了“跨语言情感分类”任务，要求利用给定的英文情感资源进行中文评论的情感倾向分类，考察多语言环境下情感资源的迁移能力。该任务的评测内容既为通过英语的情感资源实现中文评论的情感分类。

**2. NLP&CC2013任务分析**

本任务提供了英文资源，包括英文标注评论数据与英文情感词典；要求对给出的中文评论进行情感倾向性（正向或非正向）分类。

**3. 评测标准**

本任务使用准确率(accuracy)作为评价指标。

其中#system\_correct表示分类准去的评论数，#system\_total表示测试集

内全部评论数。

**4. 评测数据**

本任务的所有数据均由主办方提供，主要包含三部分：1）英文标注数据和

英文情感词典；2）中文未标注语料；3）中文测试集。数据均采用XML格式，UTF-8编码存储。

4.1 英文标注数据

英文标注数据为Amazon.com的用户评论，共包含3个不同领域：DVD，书籍，音乐。每个领域内有2000条正面评价和2000条负面评价。数据文件的组织结构为XML格式，每条评论对应一个<item>元素。一条评论内包含4个子元素：评论摘要(<summary>)、评论内容(<text>)、评论倾向(<polarity>)和评论类别(<category>)。图1列举了一条书籍的正向评论。

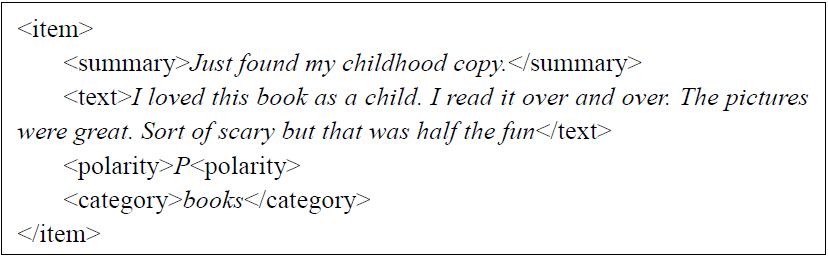


图1 英文标注语料

4.2 少量中文标注数据

提供包含DVD，书籍，音乐三个领域的少量中文标注数据，正负例各20条。

4.3 大量中文未标注数据

针对DVD，书籍和音乐每个领域提供大量中文未标注评论。未标注语料的数据格式与英文标注数据类似，但每条评论下无<polarity>元素。

4.4 中文测试集

DVD，书籍和音乐每个领域均包含4000条测试评论，其中正负向评论均为2000条。测试集同样采用XML格式，每条评论对应一个<item>元素。一条评论内包含4个子元素：评论摘要(<summary>)、评论内容(<text>)、评论标号(<review\_id>)和评论类别(<category>)。

**5． 跨语言情感分类方法——coTraining**

5.1 数据预处理

（1）解析XML

这部分的工具使用的python自带的xml解析工具（import gTrans），并且根据实验需要最终把数据处理成如下格式：summary+polarity+text+category

（2）调用google翻译API对于中英文进行翻译

这部分的实现主要使用的是gTrans.translate(text,source,target)，源码来自github上的代码。主要是实现英文翻译为中文，中文翻译为英文的功能。最后使得所有的评论数据（有标签or无标签）都有中英文两种表示。

5.2 coTraining算法

本文所实现的方法是基于LR（逻辑回归）方法构建了一个协同过滤（coTraining）的方法对NLP&CC2013的数据集行了跨语言的情感分析，数据集中包含了较多含标注的英文数据和少量标注的中文数据，以及大量未标注的数据。

测试集的数据是中文，任务的目标是对中文情感进行分析，然而中文标注的语料太少，而英文有大量的标注语料，所以本实验提出的基于LR分别构建中英文的LR分类器——LRCN和LREN，然后对于测试数据，用训练好的两个分类器进行协同过滤，最终得出测试数据的情感结果。

1、获得训练语料（英文+英文翻译为对应的中文）、获得未标注的中文语料（中文+中文翻译为对应的英文）、获得测试语料（中文+中文翻译为对应的英文）

2、英文直接按空格分词、中文根据jieba分词器进分词，然后提取tf-idf特征构建向量

3、开始迭代（本实验的迭代次数I取I=40）

4、根据给出的训练语料分别构建中文和英文的LR分类器为：LRCN和LREN

5、将未标注的训练语料运用LR分类器进行分类，然后选择最可信的n个正例、p个负例的未标注数据分别加入到英文和中文的训练语料中（具体实验时n,p都取5）

6、如果还需要迭代，回到3

7、根据最后训练得到的LRCN和LREN工具共同判定给定的中文测试数据应为正例或者是负例。

**算法1 CoTraining实现流程**

**6. 实验结果分析**

6.1 实验结果（Total:本实验准确率/NLP&CC13最优准确率）

Book类别实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| coTraining(I=40,n=p=5) | Precision | Recall | F1 |
| 0 | 0.81 | 0.70 | 0.75 |
| 1 | 0.74 | 0.83 | 0.78 |
| Total | **0.775/0.7700** |  |  |

DVD类别实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| coTraining(I=40,n=p=5) | Precision | Recall | F1 |
| 0 | 0.81 | 0.76 | 0.79 |
| 1 | 0.78 | 0.83 | 0.80 |
| Total | **0.80/0.7833** |  |  |

Music类别实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| coTraining(I=40,n=p=5) | Precision | Recall | F1 |
| 0 | 0.83 | 0.54 | 0.66 |
| 1 | 0.67 | 0.89 | 0.76 |
| Total | **0.755/0.7595** |  |  |

**6.2实验结果分析**

（1）本实验进行了40次迭代。因为从实验结果可以发现随着迭代次数的增加，准确率的效果是先增减后减少，最优的迭代次数时40次。实验最终得到三种类别的准确率分别是0.775,0.80,0.755.

（2）实验过程也发现训练数据集中加入未标注的数据集会带来噪声，所以并不是加入越多新的训练数据越好。而且实验也发现，加入少量可信度高的数据集比加入大量可信度低的数据会带来更好的实验结果。

（3）我们的实验结果也与NLP&CC2013参赛队伍提交的最优成绩进行了对比，从上述3个表格中，我们可以发现在Book类别和DVD类别中本文的方法coTraining均优于当年参赛选手的最优成绩。并且DVD类别的分类准确率比最优的结果提升了1.67%。虽然Music类别的结果没有优于当年参赛的最优成绩，但是准确率仅仅相差了0.45%。这也证明了本文的coTraning方法可以很好的解决跨语言的情感分析的任务。