基于社区的问答系统

——《计算语言学》大作业说明

1 任务简介

社区问答系统（如知乎、Quora 等）越来越受到互联网用户的关注。由于其开放的特点，任何用户都可以提出问题或者是回答其他用户提出的问题，因此一个问题具有几百甚至上千的答案也是不足为奇的。但是答案的质量却参差不齐，和问题的相关程度也千差万别，这使得用户要把所有的答案阅读一遍成为一件费时费力的事情。针对这种情况，我们提出下面两个子任务：任务 1 的目的是自动识别相关有用的答案，任务 2 针对一般疑问句问题（即答案为 Yes/No 的问题），从所有答案中总结出问题的正确答案。下面是任务的详细介绍。

1.1 任务 1

给定训练数据为若干问题（Question）以及每个问题对应的若干答案（Comment），答案被标记为下面几种标签中的任意一种（CGOLD）：

- **Good**：答案和问题相关；
- **Bad**：答案和问题不相关；
- **Potential**：潜在有用的答案；
- **Dialogue**：答案为用户之间的对话，并非正面回答问题；
- **Not English**：非英语；
- **Other**：其他。

测试时给定问题和答案，对**每个答案的类别**进行预测。

1.2 任务 2

给定训练数据为一般疑问句问题以及每个问题对应的若干答案，答案被标记为下面几种标签中的任意一种（CGOLD_YN）：

- **Yes**：肯定回答；
- **No**：否定回答；
- **Unsure**：不确定。

测试时给定形式为一般疑问句的问题及若干答案，综合考虑每个答案的描述，对**每个问题的答案**（Yes/No/Unsure）进行预测。

2 数据格式说明

两个任务的数据均在 data 文件夹下，data 文件夹中的数据分为训练、开发和测试数据三部分，所有文件编码均为 **utf-8**。各文件数据格式说明如下：

- **train.xml**

  此文件包含所有训练数据，格式为：

```
<root>
        <Question> … <\Question>
        <Question> … <\Question>
        …
        <Question> … <\Question>
</root>
```

  每个 Question 有一系列的属性，下面是例子和说明：

```
<Question QID="Q1" QCATEGORY="Pets and Animals" QDATE="2009-03-07
19:24:00" QUSERID="U1" QTYPE="YES_NO" QGOLD_YN="Yes">
```

  - **QID:** 每个 Question 唯一的标记，在任务 2 的答案文件中需要用到；

  - **QCATEGORY:** Question 的领域类别；

  - **QDATE:** 提交日期；

  - **QUSERID:** 提交答案的用户 ID；

  - **QTYPE:** 问题的类型，为 **GENERAL** 或 **YES_NO**，YES_NO 即表示该问题是一般疑问句；

  - **QGOLD_YN:** 为 **Yes** 或 **No**，表示从所有答案总结出来的一般疑问句的正确答案，当 QTYPE=”YES_NO”时才有意义；

每个 Question 的 xml 结构如下：

```
<Question ...>
  <QSubject> text </QSubject>
  <QBody> text </QBody>
  <Comment> … </Comment>
  <Comment> … </Comment>
  …
  <Comment> … </Comment>
</Question>
```

QSubject 中的 text 为问题的简短描述，QBody 中的 text 为问题的详细描述，接下来每个 Comment 为一个答案，Comment 有如下属性：

```
<Comment CID="Q1_C1" CUSERID="U4" CGOLD="Good" CGOLD_YN="No">
```

- **CID**：每个 Comment 唯一的标记，在任务 1 的答案文件中需要用到；

- **CUSERID**：提交该答案的用户 ID；

- **CGOLD**：人对答案标签的标注，为"Good", "Bad", "Potential", "Dialogue", "non-English" ,"Other"之一；

- **CGOLD_YN**: QTYPE=" YES_NO"时，表示答案对问题的回答是 Yes、No 还是 Unsure；QTYPE=" GENERAL"时总为 Not Applicable。

每个 Comment 的 xml 结构如下

```
<Comment ...>
  <CSubject> text </CSubject>
  <CBody> text </CBody>
```

</Comment>

CSubject 中的 text 为 Comment 的简短描述，CBody 中的 text 为 Comment 的详细描述。

- **dev.xml 和 test.xml**

  格式同 train.xml。

- **dev-gold.txt**

  每行的格式为 “CID \t CGOLD”，\t 为 tab 键，CID 和 CGOLD 含义参见 train.xml 说明。

- **dev-gold-yn.txt**

  每行的格式为 ”QID \t CGOLD_YN”，QID 和 CGOLD_YN 含义参见 train.xml 说明。

因此任务 1 的训练据数为 train.xml 中的所有数据，任务 2 的训练数据则为 train.xml 中 QTYPE="YES_NO"的数据，开发数据和测试数据也同理。

3 作业要求

本次作业可以个人单独完成，也可两人共同完成，自行组队。作业要求提交如下内容：

- **实验报告：**需要详细说明实验过程和结果，包括实验预处理、模型方法、所用特征、用到的工具和外部资源、实验结果分析、结论以及参考文献等；如两人完成，请列出组员以及分工情况。
- **程序代码：**将所有实验相关的代码至于 src 文件夹下，要求程序风格良好、注释详细，可运行并给出程序运行的方式。
- **测试数据答案文件：**每个任务一个答案文件，**编码均为 utf-8**，要求任务 1 的答案文件格式与 dev-gold.txt 相同，任务 2 的与 dev-gold-yn.txt 相同。

课堂报告：各组可自愿派代表对实验进行课堂报告，时间为 10 分钟，进行课堂报告的组会有相应的加分。课堂报告拟定于最后一堂课进行，需要进行报告的组请提前报名，并准备好 PPT，为保险起见建议准备一份相应的 PDF 文件。

最终成绩：由测试数据准确率、模型方法、组员分工、实验报告、代码、课堂报告等综合决定。

4 提交说明

截止时间：2015 年 12 月 20 日 23:59。

提交方式：提交至邮箱 miaohong-chen@foxmail.com，邮件标题“学号-姓名-计算语言学大作业”。

文件格式： 提交的文件为压缩文件（.rar 或.zip 格式），以学号命名即可，如“1234567890.rar”，压缩文件夹下包含如下文件（请严格按照下面说明的方式命名）：

- **src 文件夹：**包含所有实验相关的代码
- **1234567890-实验报告.pdf：**实验报告， word 文档也行，尽量为 pdf。
- **1234567890-test-task1.txt：**任务 1 测试数据答案，数据格式与 dev-gold.txt 相同，文件编码为 utf-8。
- **1234567890-test-task2.txt：**任务 2 测试数据答案，数据格式与 dev-gold-yn.txt 相同，文件编码为 utf-8。