

NLP_project2实验报告

2024春季学期

院系：人工智能学院

姓名：刘梦杰

学号：211300022

邮箱：2757400745@qq.com

实验时间：2024.7.8

目录

- 一、实验目的
- 二、实验环境
- 三、实验过程以及结果

一、实验目的

测试模型baseline并换一个模型试试，改进baseline的解决方案。

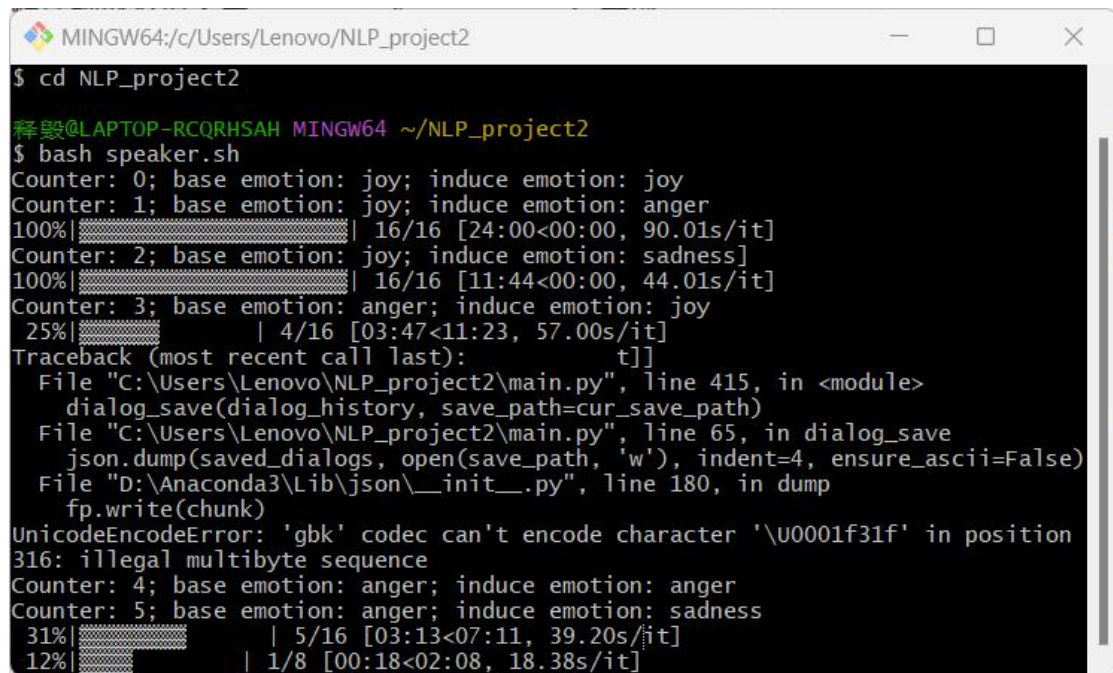
二、实验环境

Windows11, python3.11, anaconda3;

三、实验过程以及结果

1、跑baseline部分：

由于并未配置linux下的python环境，所以通过安装gitbash的方法来使用指令。首先注释掉vllm的相关代码，将main.py中的dashescope.api_key改为自己在阿里云中注册申请的，接着在git bash中直接运行bash listener.sh等等就行，但是，在运行过程中，speaker.sh出现了如下报错：



```
MINGW64:/c/Users/Lenovo/NLP_project2
$ cd NLP_project2
释毁@LAPTOP-RCQRHSAH MINGW64 ~/NLP_project2
$ bash speaker.sh
Counter: 0; base emotion: joy; induce emotion: joy
Counter: 1; base emotion: joy; induce emotion: anger
100% | 16/16 [24:00<00:00, 90.01s/it]
Counter: 2; base emotion: joy; induce emotion: sadness
100% | 16/16 [11:44<00:00, 44.01s/it]
Counter: 3; base emotion: anger; induce emotion: joy
25% | 4/16 [03:47<11:23, 57.00s/it]
Traceback (most recent call last):
  File "C:\Users\Lenovo\NLP_project2\main.py", line 415, in <module>
    dialog_save(dialog_history, save_path=cur_save_path)
  File "C:\Users\Lenovo\NLP_project2\main.py", line 65, in dialog_save
    json.dump(saved_dialogs, open(save_path, 'w'), indent=4, ensure_ascii=False)
  File "D:\Anaconda3\Lib\json\__init__.py", line 180, in dump
    fp.write(chunk)
UnicodeEncodeError: 'gbk' codec can't encode character '\U0001f31f' in position
316: illegal multibyte sequence
Counter: 4; base emotion: anger; induce emotion: anger
Counter: 5; base emotion: anger; induce emotion: sadness
31% | 5/16 [03:13<07:11, 39.20s/it]
12% | 1/8 [00:18<02:08, 18.38s/it]
```

是无法编码的问题，是因为在对话过程中，通义千问会产生中英文之外的表情导致编码失败，此报错仅出现一次，忽略不计；接着是evaluate.py中的问题，最后运行结束会产生全0的结果，原因是上文中提到是在windows环境下运行的，而evaluate.py中的os.path.join()这个函数，在路径方面，windows和linux分别是反斜杠和正斜杠，所以只要将其中关于路径的部分的斜杠改掉就好，最后将report.sh中的文件名改成performance文件夹下的对应文件即可，结果如下：

Speaker:

```

释毁@LAPTOP-RCQRHSAH MINGW64 ~/NLP_project2
$ bash report.sh
{
  "global_metric_top_k_1": 0.7375,
  "local_metric_top_k_1": 0.7375,
  "global_metric_top_k_2": 0.8,
  "local_metric_top_k_2": 0.7125,
  "global_metric_top_k_3": 0.8375,
  "local_metric_top_k_3": 0.7041666666666667
}

```

Listener:

```

释毁@LAPTOP-RCQRHSAH MINGW64 ~/NLP_project2
$ bash report.sh
{
  "global_metric_top_k_1": 0.09375,
  "local_metric_top_k_1": 0.09375,
  "global_metric_top_k_2": 0.125,
  "local_metric_top_k_2": 0.09895833333333333,
  "global_metric_top_k_3": 0.125,
  "local_metric_top_k_3": 0.09722222222222222
}

```

此外，任务中要求再对一个模型对baseline的实验，采用了main.py中出现的qwen-7b-chat，本来打算使用qwen-1.8b-chat，但是出bug了，经过和阿里云网页的对照，严重怀疑是把点打错成下划线了，索性采用qwen-7b-chat，baseline如下：

Speaker:

```

● $ bash report.sh
{
  "global_metric_top_k_1": 0.3333333333333333,
  "local_metric_top_k_1": 0.3333333333333333,
  "global_metric_top_k_2": 0.3333333333333333,
  "local_metric_top_k_2": 0.3333333333333333,
  "global_metric_top_k_3": 0.3854166666666667,
  "local_metric_top_k_3": 0.35069444444444445
}

```

Listener:

```
● $ bash report.sh
{
  "global_metric_top_k_1": 0.3333333333333333,
  "local_metric_top_k_1": 0.3333333333333333,
  "global_metric_top_k_2": 0.3333333333333333,
  "local_metric_top_k_2": 0.1666666666666666,
  "global_metric_top_k_3": 0.3333333333333333,
  "local_metric_top_k_3": 0.1111111111111111
}
```

2、收集框架的优化：

正如现实生活中温和脾气好的人更不容易被激怒一样，除了有目的的对话之外，还可以收集自由模拟自由对话的数据，观察模型的情绪变化以及判断模型的说话是否具有情绪上的倾向性，或者指定被激发者的性格来实验。

3、评价指标的优化：

（1）通过观察performance文件夹中的数据发现，不同情绪被激发的难易程度是不一样的，在我运行时，感觉joy就更容易被激发，因此在运行report.sh时不如将不同情绪激发或者被激发的数据合并，而不是全部合并；

（2）此外，采取积分制或许更为合理一些，以最后三轮的统计结果为例，假如出现了目标情绪，目标情绪出现大于等于两次积2分，只出现1次积1分；如果没有出现目标情绪，出现三种不一样的情绪扣1分，如果后三轮中出现了相同且非目标情绪，说明情绪激发与目标背道而驰，则扣2分。当然，激发出错误的情绪，也可以通过度量该情绪与目标情绪的相似度来进行相应的加减分操作，比如sadness、anger同为负面情绪，将sadness激发成anger还是优于joy的。