Natural Language Processing-Final

Group - 11
109550116楊傑宇
109550150呂則諺

1. Describe how you implement your model, including your choice of packages, model architectures, model input, loss functions, hyperparameters, etc.

a. We define it as a text classification task. The following are description about our problem formulation:
   - Input: text
     - Template: [User question] [Sep] [Situation Description] [Sep] [Response]
   - Operation: We try some language model
     - Bert
     - Distilbert
     - Ernie
     - Roberta
   - Output: Label (Appropriate or Inappropriate)

b. We define it as a text classification task and the setting is the same as mentioned above. However, we utilize the information about the golden index to improve the performance. The following are the descriptions of our method:
   - Finetune on identifying golden index task -> Finetune on text classification
   - Finetune on text classification based on the result of identifying golden index task
   - Finetune on text text classification with golden index informations (We replace GT's golden index with predicted golden index during testing)

c. 在主要方法之外，我們也嘗試直接使用chatgpt的api，除去 training 步驟直接進行prediction.

- model architecture: 直接使用openai預訓練好的兩種模型

```
# gpt-3.5
openai.api_key = api_key
response = openai.ChatCompletion.create(
      model="gpt-3.5-turbo",
      messages=[{"role": "user", "content": prompt}
                ]
    )
#print(response)


#davinci-003
response = openai.Completion.create(
      engine="text-davinci-003",
      prompt=prompt,
      max_tokens=1000,
      api_key=api_key,
      temperature=0.8
    )
#print(response)
```

- model input: 我們將test data每50題切為一份加上問題敘述做為一次的input。

```
test = test_data_preprocess(data_test)
#print(test)
quest ="\n合適為1 不合適為0 ,給我一個長度為50的list contain all the answer\n"
prompt = test+quest
```

2. What processing did you do with the data? Is there an improvement in predictive accuracy when utilizing both situations and utterances for prediction, compared to solely relying on utterances? Why or why not?

a. In this method, we use a template to fill in the questions, describe the scenario, and provide answers.
    - Template: [User question] [Sep] [Situation Description] [Sep] [Response]

    Yes, there is an improvement in predictive accuracy when utilizing both situations and utterances for prediction.We believe that some responses are reasonable when not considering contextual information. However, these replies may become inappropriate when taking the context into account.

b. Besides the method mentioned above, We also made corresponding processing for contextual information, eliminating noise from the contextual information. We found that providing accurate information can significantly improve the accuracy rate, as demonstrated in the table for the third question.Thus, we design a task to identify if the contextual information is noise. The information about the task is described below:

- Input: text
    - Template: [User question] [Sep] [Situation Description]
- Operation: We try some language model
    - Bert
    - Distilbert
    - Ernie
    - Roberta
- Output: Label (Noise or Non noise)

Besides, we thought that the definition of golden index may related to the response. Thus, we also try the bellowed setting:

- Input: text
    - Template: [User question] [Sep] [Situation Description] [Sep] [User Response]
- Operation: We try some language model
    - Bert
    - Distilbert
    - Ernie
    - Roberta
- Output: Label (Noise or Non noise)

Yes, there is an improvement in predictive accuracy when utilizing both situations and utterances for prediction.We believe that some responses are reasonable when not considering contextual information. We believe that providing additional contextual knowledge can assist the model in understanding the environment and helping models to make the right decision.

c. 在chatgpt-api方法裡，我們實驗了兩種data process形式:
    1. utterances+response，並且詢問response是否合適？
    2. [situation]+utterance+response,並且詢問’根據situation’,response是否合適?

    依照validation的結果顯示方法1的表現會優於方法2，並且在穩定度上也比方法2高，我們推測猶豫方法2一次給予的資訊太多，會讓model無法聚焦在真正的問題上，常常會答非所問，或是ouput錯誤的形式，再加上方法1的token數量整體比較小，能夠讓我們在有限的資源上跑比較少次，因此我們最後選擇將data處理成方法1的樣子。

3. Compare all the methods you have tried and use a table to display their respective performances. Which method performed the best, and why?

Model	Data	Performance
Ernie-2.0	with situation & without golden index	0.74
Erne-2.0 (Pretrained on finding golden index)	with situation & without golden index	0.74
Erne-2.0	with situation & with golden index	0.873
Erne-2.0	with situation & with predicted golden index	0.76
Erne-2.0 (Pretrained on with golden index)	with situation & with predicted golden index	0.76
Chatgpt-api (gpt-3.5)	without situation	0.6
Chatgpt-api (gpt3.5)	with situation	<0.5

在text classification的設計中，Ernie的表現是最好的因此圖上的資料我們只放Ernie的表現。在有golden index的情況下表現最佳，而因此我們可以看出golden index的重要性。我們認為golden index的情況下會表現比較好是因為，golden index可以幫我們過濾掉不重要的contextual information，這樣model就不會被這些資訊誤導，從而可以做出正確的決定。至於chatgpt在這次的任務表現不好的原因，第一是因為我們沒有去處理golden index，這導致在加上situation的方法裡會容易加雜無用的資訊，在只有utterance的情況中又無法有更多資訊導致預測不準確。