2022

Data Mining Final Project

組員: 卓冠廷、李照棋、王伊婷

CONTENTS

目錄

PART 01

資料介紹分析與前處理

Data Analysis and Data Preprocessing

資料介紹&分析

Data Introduction & Analysis

01 資料介紹

本競賽的每一筆輸入資料為一個三元組(q, r, s)，q 是一則英文論述，r 是一則對 q 進行回應的英文短文，s 則是 r 對 q 的議論關係，可能是同意(agree)或不同意(disagree)。

輸出資料則是一個雙元組 (q', r')，q'與r'分別是 q 與 r 的子序列(subsequence)，且q'與 r'提供了關鍵性的資訊，足以判斷 q 與 r 呈現 s 的關係。

下表呈現了一筆範例，q 為一則論述，r 是 q 的一則回應，r 與 q 的關係為不同意。預期的輸出q'與r'則如 q 與 r 中黃底的片段，提供關鍵性的資訊呈現 r 不同意 q。注意 q'與r'可以是不連續的片段(但不同片段間的先後順序必須與原文之順序相同)。

q	Originally posted by voiceofreason Well if you're going to quote that article that mentions Canada to support your opinion, then does that mean you're in favor of gun registration and licensing? Canada has them. And if the NRA is really interested in enforcing laws why do they ask congress to reduce the budget of law enforcement? The NRA uses its political muscle to make it easier for criminals to obtain guns. Robert Ricker, former top lawyer of the NRA, talks about it. So if you're really interested in enforcing gun laws, don't support the NRA.
r	What is true, Ricker says, is that gun manufacturers have long known that distributors and retailers supply thousands of guns each year to criminals, and yet gun makers deliberately look the other way. The quote says it all. That attitude breaks many cornerstone laws regarding personal responsibility. Manufactures are NOT responsible for customers illegal activity. GM and Ford know that 100% of all the product they sell will be used illegally. Everyone of their cars will be used to break speed limits. No one obeys speed limts, the only time speed limits are obeyed is when a cop is watching. Yet GM and Ford are not responsible for all the illegal activity of their products. Shifting the responsbility for enforcement breaks many traditions and laws regarding individual personal responsibility. When a customer breaks a law using a product, manufactures are not responsible, that goes for both guns and cars, or any product that is sold and used for criminal purposes. There certainly a lack of cooperation between the ATF and the NRA. They really should be working together. It takes two sides to make a conflict. The ATF is trying to single out gun manufactures to be responsible for policing customers. NO OTHER INDUSTRY is being held accountable like that. That is why there is a conflict between the NRA and ATF. The NRA asks that gun manufactures to be treated with the same consideration of existing laws and standards applied to other industries. The NRA supports any efforts to take firearms away from criminals, but they are asking that gun manufactures not be held accountable for customer illegal activity. Which is the position of every single product manufacturer not matter what you make and sell. It is not that gun manufactures are looking away, they do. So does every other manufacture who makes any other product.
s	Disagree

資料介紹＆分析

Data Introduction & Analysis

01 資料分析

df_train

s	q	r
AGREE	6918	6918
DISAGREE	31428	31428

df_test

s	q	r
AGREE	367	367
DISAGREE	1649	1649

	id	q_length	q'_overlap	r_length	r'_overlap
0	8	3	[0, 1, 2]	1	[0]
1	9	1	[0]	3	[0, 2]
2	10	17	[0, 1, 2]	10	[0]
3	11	14	[4, 5, 6, 12, 13]	21	[1, 2, 3, 4, 17]
4	12	1	[0]	1	[0]
5	13	1	[0]	2	[0]
6	14	8	[0, 1, 3, 4, 5, 6]	7	[0, 1, 3]
7	29	1	[0]	2	[0, 1]
8	30	11	[1, 2, 3, 5, 6, 10]	1	[0]
9	31	4	[0, 1, 3]	4	[0, 1, 2]
10	32	1	[0]	3	[0, 1, 2]

資料前處理

Data Preprocessing

01 資料符號處理

將Data中的奇怪的標點符號去除，僅保留四種符號(,.!?)

```
'"First , there is no `` us `` on your part regarding this . I am talking to you . Others here th
s you apart . Second , I was n\'t trying to give you justification , nor am I obligated to . You
But I do n\'t owe you any justification . Take your silly games up with Mana\'ia ; they wo n\'t w
ne who seems to think this is a game , based on your posts . ( with your comments like `` you \'v
seems you \'d do well to take your own advice . If you want people to constantly have to `` justi
```

02 結合某些縮寫

文章中縮寫詞常出現莫名的空格，像是don't會被寫成do n't，容易造成word_tokenize後被分成兩個字的情形。

特別處理:'s、n't、'll、'm、've、're、'd

```
'First , there is no us on your part regarding this . I am talking to you . Others here tha
rt . Second , I wasn't trying to give you justification , nor am I obligated to . You shoul
owe you any justification . Take your silly games up with Manaia they won't n't work with m
ink this is game , based on your posts . with your comments like you've won he's lost , and
. If you want people to constantly have to justify something to you , then you should make
```

03 找前後tokens是否可合成一個單詞

文章中有時會出現一個單詞被分開的情況，因此用enchant的套件去判斷是否為單詞，True就做結合

```
!apt  update
!apt  install  enchant  --fix-missing
!apt  install  -qq  enchant
!pip  install  pyenchant
import  enchant
d  =  enchant.Dict("en_US")
```

PART 02

QA BERT

Question Answering with a fine-tuned BERT

Question Answering with a fine-tuned BERT

Let machine answer machine

01 將前處理資料加工

將前處理後的資料中的 q 和 r 結合成文章，並根據 Agree 和 disagree 設計問題

```
Text:
Q:http news.telegraph.co.uk news mai ... nabort15.xml r:i think that i don't have quite enough tobasco in my bloody mary .

Question:
What does q disagree with r?
```

02 將文章和問題丟入模型

預訓練模型：
'bert-large-uncased-whole-word-masking-finetuned-squad'
分詞器(tokenizer)：
'bert-large-uncased-whole-word-masking-finetuned-squad'

```
model = BertForQuestionAnswering.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
tokenizer = BertTokenizer.from_pretrained('bert-large-uncased-whole-word-masking-finetuned-squad')
```

03 獲得並解析結果

過濾無法辨識的答案，並重新解讀

04 結果

準確率(根據 Agree 和 disagree 設計問題)：
61.12%

可能的問題：

1. 問的問題太單一不太符合文章的敘述
2. 因為輸出是文章中的起點和終點，對非連續的單詞或短句答案不利，最終會只預測出一個單詞或連續不相關的段落。

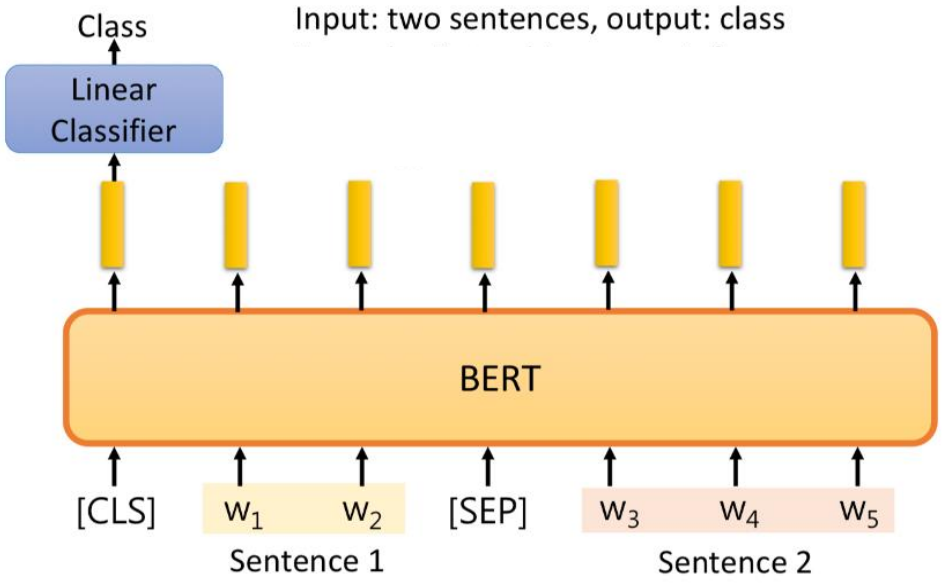
PART 03

BERT Classifier

Bidirectional **E**ncoder **R**epresentations
from **T**ransformers and **L**inear **C**lassifier

BERT Classifier

Bidirectional **E**ncoder **R**epresentations
from **T**ransformers and **L**inear **C**lassifier



模型

透過將q和r丟入模型中訓練(預訓練權重選'bert-base-cased')，最終可獲得80%的準確率，但這是針對分辨s的部分，而此競賽需要的是獲得q'和r'

問題

得Bert的sequence outputs(X)與linear classifier(Y=AX)的q和r分別的權重值(A)，然而，即使從sequence outputs取出q'與r'，仍不知該如何將s作為input

解決辦法

選用Transformer做純粹的sequence extraction，並將提取結果丟上AI CUP競賽中，結果意外的好

PART 04

Transformer

Text Extraction for **q** and **r**

Transformer

Text Extraction for q and r

01 模型與預訓練權重

預訓練權重： tf-small
模型 : TFAutoModelForSeq2SeqLM

Model size variants

Model	Parameters	# layers	d_{model}	d_{ff}	d_{kv}	# heads
Small	60M	6	512	2048	64	8
Base	220M	12	768	3072	64	12
Large	770M	24	1024	4096	64	16
3B	3B	24	1024	16384	128	32
11B	11B	24	1024	65536	128	128

T5 model size variants. Source: T5 paper.

02 將 (q q') 與 (r r') 丟入訓練

隨機抽樣：
文章短(1~9句)的預測結果很好
文章長(10句以上)的預測結果很差
AI CUP準確率(有無前處理):
71.81% / 72.42%

```
84 token length: 4
actual: I really think it's funny .
predict: I really think it's funny
F1: 0.999999995 , Precision: 1.0 , Recall: 1.0
```

```
11 token length: 14
actual: find that appalling
predict: First , there is no us on your part regarding this . I am talking to you . Others here that argue your same positions have been much less beligerent
F1: 0.15789473184210542 , Precision: 0.15789473684210525 , Recall: 0.15789473684210525
```

03 將訓練集中文章長短文章分開訓練

隨機抽樣：
文章長(10句以上)的預測結果提升
AI CUP準確率(有無前處理) :
72.54% / 73.85%

```
41 token length: 5
actual: The coelacanth , according to fossil records , and according to evolutionists , allegedly went extinct millions and millions of years ago .
predict: The coelacanth , according to fossil records , and according to evolutionists , allegedly went extinct millions and millions of years ago
F1: 0.999999995 , Precision: 1.0 , Recall: 1.0
```

04 增加訓練時間

AI CUP準確率(有無前處理) :
75.7% / 76.12%

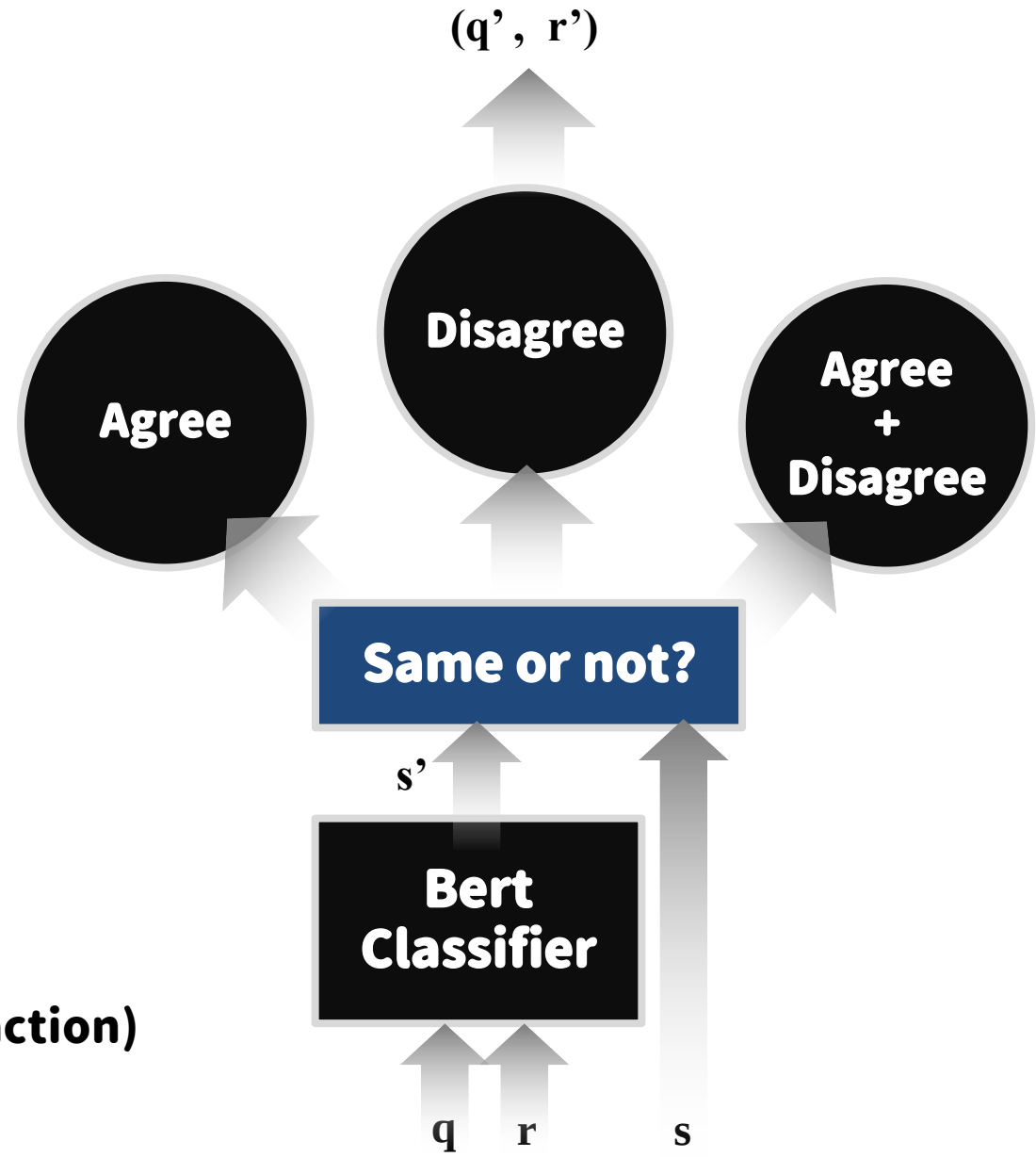
```
10 token length: 17
actual: I personly would not condone an abortion , however wouldn't condem person who wanted one
predict: I personly would not condone an abortion , however wouldn't condem person who wanted one its there choice .
F1: 0.999999995 , Precision: 1.0 , Recall: 1.0
```

Future

New model

Combine Model

Transformer帶來不錯的預測結果，但並未加入s當作input，因此想出了新的模型架構，結合了Bert Classifier與Transformer，期許能帶來不一樣的成果展現



Transformer
(Sequence extraction)

Thanks for your listening