中華民國第 62 屆中小學科學展覽會

作品說明書

高級中等學校組　電腦與資訊學科

第三名

052508

換句話說

學校名稱：新北市立板橋高級中學

作者：	指導老師：
高二 李其樺	郭兆平

關鍵詞：NLP、BERT、Fine tuning

摘要

BERT 近年來在各式 NLP 任務中可說是無處不見、無所不在，其中使用 fine-tuning 的訓練方式更是可以幫助研究者省下大量的時間及運算成本，且結果都有不錯的表現。本研究探討在結合不同條件的文本訓練下，基於對 BERT 模型做 fine-tuning 且讓其進行文本分類，觀察其對於預測及分類中文句子通順程度的成效，並且根據訓練出來的模型設計修正方式嘗試使其對預測中不通順之文本進行自我修正，並分析其成效與結果。

壹、 前言

一、研究動機

在這個資訊大多可以從網路上獲得的世代，在瀏覽網路文章中常常會有讀到一半結果就發現「這個字好像打錯了」或是「這裡是不是多打了一個字」之類的情況發生，導致句子變得很奇怪，於是我想製作出一個模型嘗試找出這種奇怪的句子，並且嘗試去修正他。

二、研究目的

(一) 嘗試找出不通順的文句並在不涉及語意矛盾等情況下加以修正
(二) 探討不同類型不通順文本對於修正句子通順程度的影響
(三) 探討不同類型不通順文本訓練出的模型之表現
(四) 探討將不同類型不通順文本合併後訓練模型的成效

三、文獻回顧

在修正與判斷句子通不通順的研究中，有關繁體中文相關的研究較少，大多為英文相關的，其中也是因為中文包含了一個大魔王：「文言文」的存在，而在翻閱文獻時，是有看到 [8] 一篇使用 n-gram 模型，並且使用國中生作文來當作訓練資料的研究，但由於我想做的是更一般性針對一個個句子的研究，故決定不使用該研究的架構及文本，使用句子品質更不一定的「維基百科文本」來做實驗。

貳、 研究設備及器材

一、平台： Google Colab

二、資料儲存：Googl 雲端

參、 研究過程或方法

一、訓練資料集處理

本研究選用之資料集為維基百科中文資料庫 20210920 版本（zhwiki-20210920-pages-articles-multistream.xml.bz2）。

(一) 繁體轉換

使用 Python 之 opencc 套件，將使用之維基百科文本資料轉換成繁體中文，並且將已經斷句的資料，以每句話為單位儲存。

(二) 生成不同訓練文本

本研究共生成了 6 種不同的文本，並且假設正常文本內容皆為通順的句子，其他生成的文本內容皆為不通順的句子。

1. 正常文本（圖 1）
   將處理好的檔案，隨機選取字數介於 10 至 20 字的句子。

2. 隨機字文本（圖 2）
   將處理好的檔案，所有出現過之中文字建表，並且從中隨機挑選中文字組成字數介於 10 至 20 字的句子。

3. 重組句子文本（圖 3）
   將處理好的檔案，選取字數介於 10 至 20 字的句子，將每一句選取的句子打亂重組該句子。

4. 取代字文本（圖 4）
   將處理好的檔案，選取字數介於 10 至 20 字的句子，並將所有出現過之中文字

建表，將選取句子中的任意一個字改成任意的字。

5. 交換字文本（圖 5）

將處理好的檔案，選取字數介於 10 至 20 字的句子，將選取句子中任意兩個不相同的字交換位置。

6. 重複字文本（圖 6）

將處理好的檔案，選取字數介於 10 至 20 字的句子，將選取句子中的任意一個字複製並插入該字後方。

連續的量即是以實數來表示的
是從先秦散文中的寓言濫觴
這些情節不只是娛樂性的
年一篇由蘇美人創作的
世紀發展了現代史學方法
固然是因為表演者是歌者
在歌曲中這種作用表現得最為突出
使用者用此方法來估計規模
及發展了以比較方式研究政治
法律具有超乎個人的自身價值

圖 1、正常文本。

鋩蔄蘏麒餅蒿袉盳萲餉邈劉
枎碤纐霧亍珣轁侄忎馗媄萬梓
焥狡錠救鯢釩沯妸埡苔
坱清堅鸝緆趭繆崽隮陵宥虢暍畀杪殆
姶溝柱猔衲矯隆勭箳庴信僩謥陻縫磐忻椵禃町
赲衕蓀居滑嗡縚鼉鞶緒蒲
櫛儀譻餅圐顏酾壉閬椿創ㄨ杤
鍂瀛崗暻棣釧咻礬索詎馹叚屜嵓謦
塔扻栀諫籙妊祼萺藤踽腑箒栖崢
飧捗竗薈踇郅蚮格肮澤樀驊踣犭霚

圖 2、隨機字文本。

抒最來騷離長的情詩以早篇
附影的下上電邊放播都在幕會字
指資上及差以源配經標分異等的濟
較灣臺具法目表學的院前性代
的莫定論理天了天礎氣基
繼指相斯高拉普和出斯拉
商或業性農高自性業給
然自自別概於念環境個源
間鮑一著能的塔姆論名之爭場
士位茲碩了的受他接託瓦學

圖 3、重組句子文本。

那麼所演出的便不再是戲慜
戲劇可以被分成不同的型鴴
復旦國際範係與公共事務學院
分為蓑般法學與特別法學兩種
化學的研究範螥是包括分子
地球歷史上曾發生過多次大的乭難
大氣和海洋癑緊密聯絡的
由於這個過程枚以不斷的迴圈發生
更意味著牓流量的差距
暱為資料結構概念的普及

圖 4、取代字文本。

沒錻一章特別談及耶有
文學手段是在敘述中用的結殊特構
音樂是一種需要學習的能技
和描述政府向作的情況為方運
別中又可分門其類研習
他所指的是暗孔相機或針箱
著為不正確的理論支配並
動活論隨著板塊構造學說的發展
而細胞是也許多的基本單位
應付有學習障礙的生學

圖 5、交換字文本。

數學定律越和現實有關關
而且有很詳詳細的資料記錄
文學寫作作是一種藝術形式
對我們這個時代的的歐洲人而言
或者自身意識識形態和價值立場判斷
內內容與基督教教義衝突甚大
目前臺灣較具代表性的的法學院
研究架構等一直直經歷著擴充套件和分歧
也表述出內含的的數學概念
行星形成時碰撞的殘留熱熱量

圖 6、重複字文本。

二、BERT

(一) BERT 簡介

BERT（Bidirectional Encoder Representations from Transformers）語言模型（Language Model, LM）的一種變形，是由 Google 以無監督方式所訓練出的模型。

BERT 是 Transformer 中的 Encoder，有著能夠直接處理各式 NLP 任務的通用架構，可以事先訓練好一個可以，並套用到多個 NLP 任務，或再以此為基礎 fine tune 多個下游任務。Fine-tuning 在原本 Bert 模型的最後一層，可以接一個新的 classification layer 做下游任務，並使用較少量的文本訓練整個神經網路。

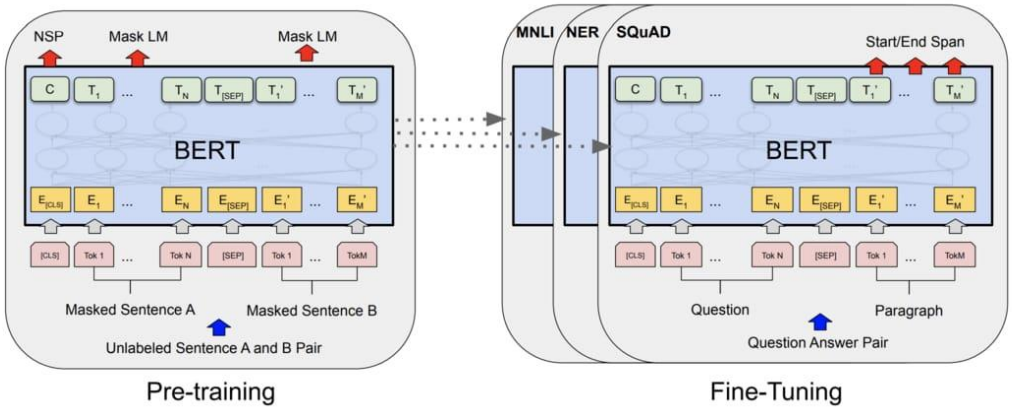


圖 7、BERT

(二) MLM 任務

BERT 的作者們的論文[5]中使用了 Transfomer Encoder，透過大量文本做了兩個預訓練目標，其中一個就是 MLM（Masked Language Model）任務，又稱克漏字填空。透過遮罩（[Mask]），可使其 Attention（注意力機制）關注在特定的字詞上面，從而預測該遮罩可能出現之字詞。

(三) Fine-tuning

Fine-tuning 是指使用已經完成訓練的神經網絡模型，用來進一步執行其他類似任務的過程，採用已經設計和訓練好的神經網路能夠更有效率地利用神經網絡前層先前訓練的特徵，而無需從頭訓練提取該特徵，只需要針對下游任務的 classification layer 進行 fine tune。

三、模型訓練

(一) 模型訓練資料分配

以下分成 A 到 F 組模型，固定每組模型訓練資料固定總合為 100000 個句子，並且包含通順及不通順的句子各占一半。

1. A 組
   正常文本 50000 個句子
   隨機字文本 50000 個句子

2. B 組
   正常文本 50000 個句子
   重組句子文本 50000 個句子

3. C 組
   正常文本 50000 個句子
   取代字文本 50000 個句子

4. D 組
   正常文本 50000 個句子
   交換字文本 50000 個句子

5. E 組
   正常文本 50000 個句子
   重複字文本 50000 個句子

6. F 組
   正常文本 50000 個句子
   隨機字文本 10000 個句子
   重組句子文本 10000 個句子
   取代字文本 10000 個句子
   交換字文本 10000 個句子
   重複字文本 10000 個句子

(二) 模型 fine-tuning

使用 fastai 作為框架，將 BERT 模型做 fine-tuning，將訓練文本轉成 csv 檔案，分成「text」和「label」兩個標籤類別，「text」項儲存訓練文本，「label」項儲存 1 或 0（通順文本為 1，不通順文本為 0）。

接著以 4 比 1 的比例分割訓練與測試資料，透過「label」項，以準確率為基準，來進行文本分類，並且固定 4 個 epoch，觀察 train loss 與 valid loss 下降的情形和 f1 score、precision score、recall score 等數據的變化。

(三) 句子通順度定義

在此，經過 fine-tuning 後的 BERT 模型可以進行文本分類，會對於 0（不通順）和 1（通順）分別給出一個 0 到 1 的結果，並且相加起來為 1，所以在此將訓練後模型預測的「句子通順度」定義為模型預測是該句子為 1 的結果。

四、句子修正

由於訓練好的模型只能用來預測句子的通順程度，並不能直接用來修正句子，於是為了達到修正句子的功能，會取用預訓練好的 BERT 模型進行 MLM 任務，並且結合本研究中製作出的模型預測句子通順程度，要被修正的句子會經過以下流程 5 次，以下用「今天天氣真好」舉例。

1. MASK 修正
   將要修正的句子輪流 MASK 每個字，交給 BERT 做 MLM 任務預測最可能的字的 Top5，並且一一套入句子中預測句子通順度。

   舉例來說，「今天天氣真好」會被依序變成：[MASK]天天氣真好→今[MASK]天氣真好→今天[MASK]氣真好→…→今天天氣真[MASK]

2. 增字
   將要修正的句子，在所有可以加入[MASK]的地方加入[MASK]，再交由 BERT 做 MLM 任務預測最可能的字的 Top5，並且一一套入句子中預測句子通順度。

   舉例來說，「今天天氣真好」會被依序變成：[MASK]今天天氣真好→今[MASK]天天氣真好→今天[MASK]天氣真好→…→今天天氣真好[MASK]

3. 減字
   將要修正的句子，輪流移除每個字，並且一一預測句子通順度。

舉例來說，「今天天氣真好」會被依序變成：天天氣真好→今天氣真好→今天氣真好→…→今天天氣真

接著將上面三個步驟中得分最高的句子，作為下一次修正句子的輸入，如此重複 5 次。

肆、 研究結果

一、A 組

(一) 模型訓練結果

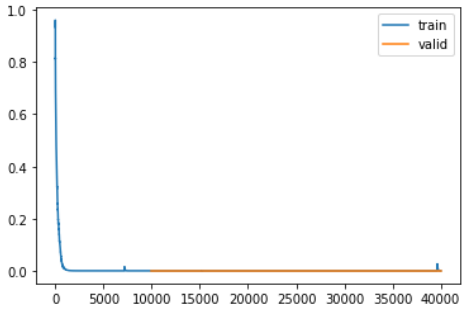


圖 8、A 組模型 loss 函數圖形。

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
1	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
2	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
3	0.000021	0.000000	1.000000	1.000000	1.000000	1.000000

圖 9、A 組模型訓練過程中的各項數值。

(二) 句子得分趨勢變化

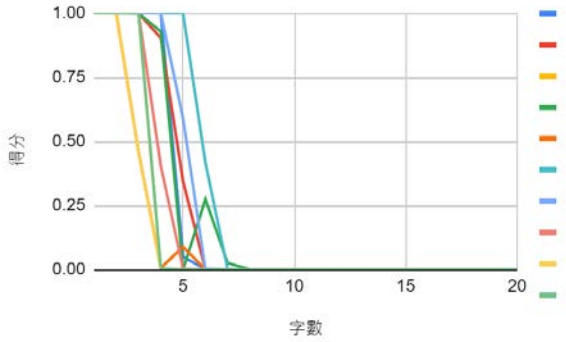


圖 10、隨機字文本在 A 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

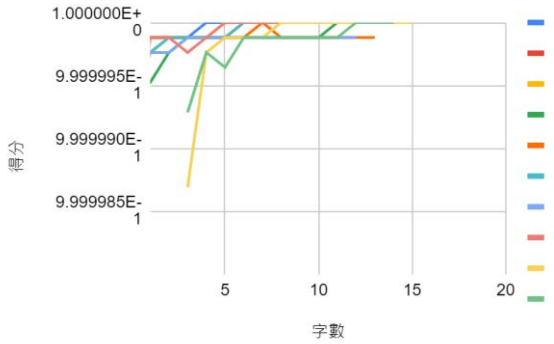


圖 11、正常文本在 A 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

(三) 句子修改結果

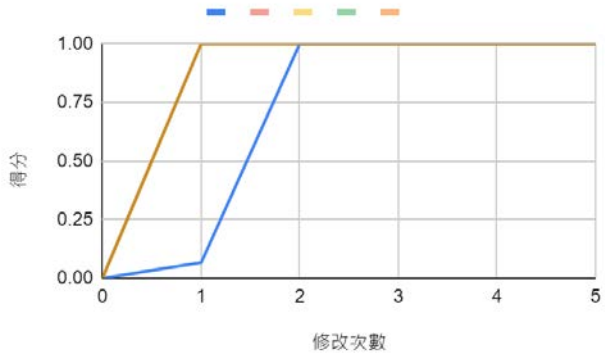


圖 12、隨機字文本在 A 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	2.64E-06	扞鄴淌榎嗵筍	2.39E-04	瑎聧蹇賡劗擲	4.01E-06	緑奾耜縤覈論
1	0.0662060231	扞鄴淌榎的嗵筍	0.9999997616	瑎有蹇賡劗擲	1.00E+00	緑奾耜的覈論
2	0.9999940395	扞大淌榎的嗵筍	0.9999998808	瑎的蹇賡劗擲	0.9999997616	緑真耜的覈論
3	0.9999998808	連大淌榎的嗵筍	0.9999998808	記的蹇賡劗擲	0.9999998808	緑真非的覈論
4	0.9999998808	連是大淌榎的嗵筍	1	記的蹇賡軍劗擲	0.9999998808	緑非非的覈論
5	0.9999998808	連非大淌榎的嗵筍	1	記的蹇賡軍将劗擲	0.9999998808	越非非的覈論

圖 13、隨機字文本在 A 組模型下，修正句子的結果（隨機抽樣 3 組）。

(四) 結果與討論

1. 由圖 8、圖 9 中的數據可見，要分類隨機字文本與正常文對於 BERT 其實一點也不難。
2. 在圖 10 中，平均要 5 個字才能讓模型確定及分別正常文本與隨機字文本。
3. 在圖 12 中，後面 4 條線的數據是幾乎重疊，都在修改第一次後之後預測機率逼近 1。
4. 在圖 13 的修正結果中，他修正後的句子仍然還是很奇怪，且一下子就得到了非常高的分數。
5. 整體來說，由於隨機字文本中會出現一些生難的字，實在是太好分別了，會導致模型無法真正學習到不通順的句子，所以這個模型在分別不通順語句上的成效並不好。

二、B 組

(一) 模型訓練結果

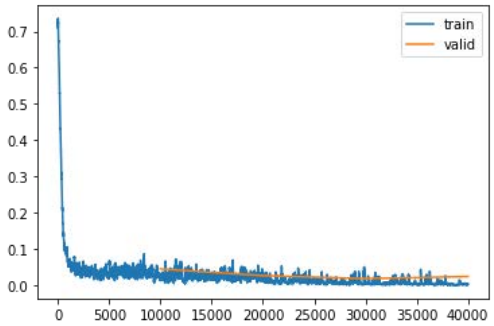


圖 14、B 組模型 loss 函數圖形。

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.024608	0.044928	0.983700	0.983930	0.969214	0.999099
1	0.020967	0.026516	0.991400	0.991443	0.985461	0.997497
2	0.010549	0.018208	0.994050	0.994038	0.994985	0.993092
3	0.004104	0.023938	0.993650	0.993670	0.989478	0.997898

圖 15、B 組模型訓練過程中的各項數值。

(二) 句子得分趨勢變化

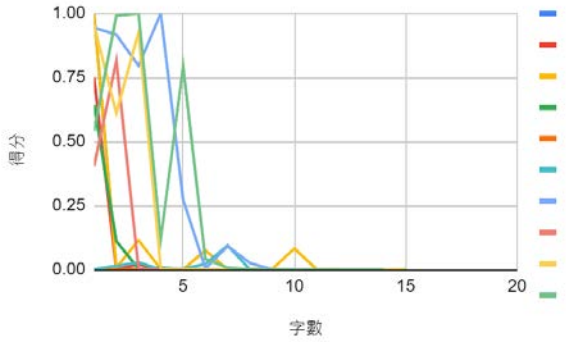


圖 16、重組句子文本在 B 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

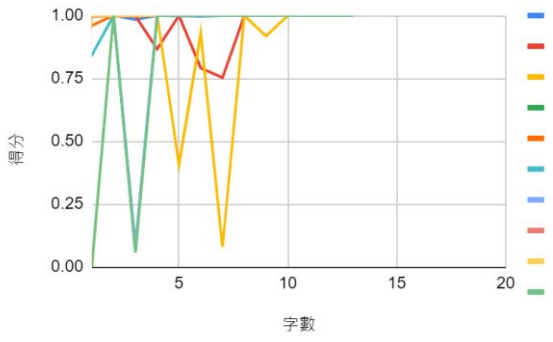


圖 17、正常文本在 B 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

(三) 句子修改結果

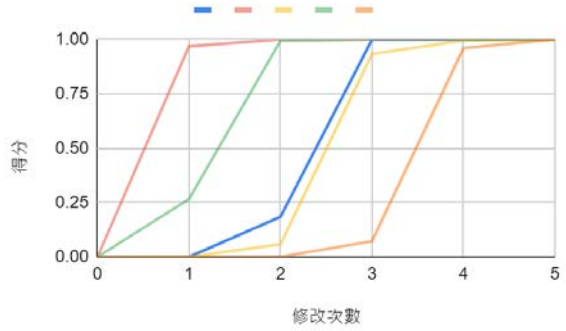


圖 18、重組句子文本在 B 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3		
0	1.11E-04	由解並散聯東德州邦年	5.86E-04	望四界及第道大道交展大	1.08E-04	流民有漁中傳間川民四	2.23E-04
1	0.000266340473	由解散聯東德州邦年	0.9685303569	望四界及第四大道交展大	6.21E-04	流民有漁中傳間川民四人	0.2660651505
2	0.1847215444	由解散聯東德州一年	0.9999251366	望四界及第四大道交展	0.05842636898	流民有漁中之間川民四人	0.9935867786
3	0.9996705055	由解散聯絡德州一年	0.9999405146	望四人界及第四大道交展	0.9331876636	流民有漁中之間的川民四人	0.9998888969
4	0.999937892	由解散者聯絡德州一年	0.99994874	望四人谷及第四大道交展	0.9944880605	流民還有漁中之間的川民四人	0.9999520779
5	0.9999439716	可由解散者聯絡德州一年	0.9999542236	望看四人谷及第四大道交展	0.9993000031	流民還有漁中之間的庶民四人	0.9999560118

圖 19、重組句子文本在 B 組模型下，修正句子的結果（隨機抽樣 3 組）。

(四) 結果與討論

1. 由圖 14、圖 15 中的數據，準確率幾乎可以達到 0.99，且損失函數的圖形有明顯下降到收斂的情況，代表其模型表現還不錯。
2. 在圖 16 中，平均要 3 個字才能讓模型確定及分別正常文本與重組句子文本。
3. 在圖 17 中，少量數據呈現波動狀態，不過大部分數據都是趨於穩定的。
4. 在圖 18 中，平均 3 次修改後可以使模型預測讓他變成通順的句子。
5. 在圖 19 的修正結果中，他修正後的句子有著不錯的成效，至少句子間不太會怪異且出現的詞幾乎都是有意思的。
6. 整體來說，因為重組句子後的品質本身就很難抓，可能會有本來就是詞的兩個字連在一起，訓練過程中可能導致特徵不明顯等，但是還是有一定程度上使得句子變亂，讓模型能夠分別出奇怪的句子，在這個情況下，模型能修正並預測出這樣子的句子已經算是不錯的了。

三、C 組

(一) 模型訓練結果

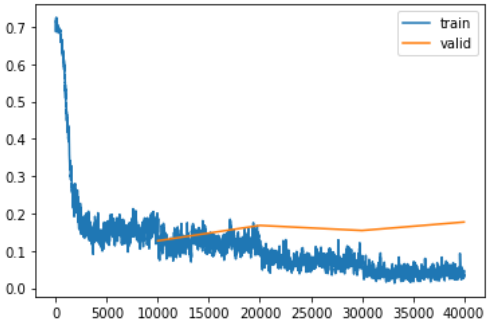


圖 20、C 組模型 loss 函數圖形。

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.176744	0.126831	0.955200	0.955788	0.949975	0.961672
1	0.111705	0.167953	0.942250	0.944808	0.910648	0.981630
2	0.070476	0.154347	0.956500	0.957126	0.950103	0.964254
3	0.043862	0.177012	0.956300	0.957278	0.942717	0.972297

圖 21、C 組模型訓練過程中的各項數值。

(二) 句子得分趨勢變化

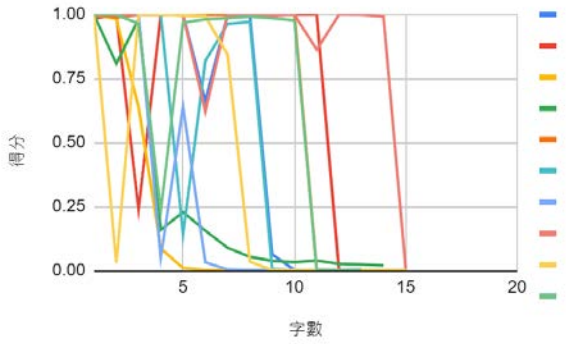


圖 22、取代字文本在 C 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

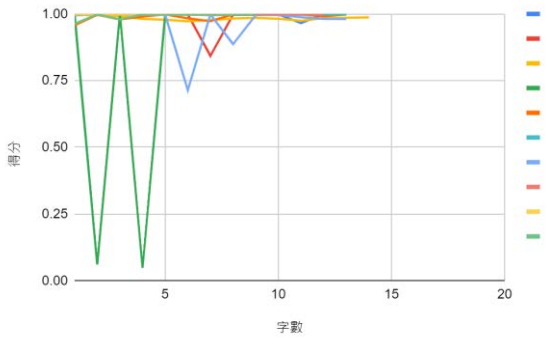


圖 23、正常文本在 C 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

(三) 句子修改結果

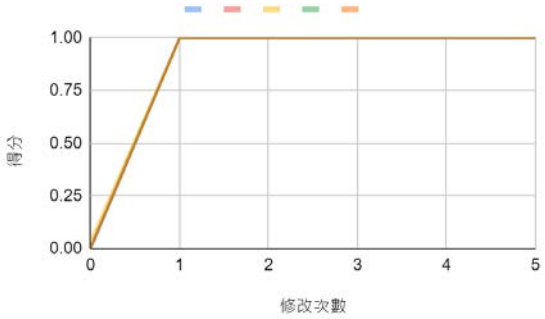


圖 24、取代字文本在 C 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	2.13E-03	他成為荷屬東印度棲島	2.47E-03	少數公喚為巴哈伊教徒	3.73E-02	筧夜的長短等都有論述
1	0.9983059168	他成為荷屬東印度群島	0.9983320832	少數公司為巴哈伊教徒	9.98E-01	夜的長短等都有論述
2	0.998354733	他成為荷屬印度群島	0.9983769655	少數公司為巴哈伊信徒	0.9984662533	的長短等都有論述
3	0.998370707	他成為荷蘭印度群島	0.9983897209	少數公司為巴伊信徒	0.9984869957	的長短都有論述
4	0.9983769655	和成為荷蘭印度群島	0.9984021783	少數公司是巴伊信徒	0.9984869957	長短都有論述
5	0.9983897209	和已成為荷蘭印度群島	0.998411417	少數公司是巴利信徒	0.9984929562	長短有論述

圖 25、取代字文本在 C 組模型下，修正句子的結果（隨機抽樣 3 組）。

(四) 結果與討論

1. 由圖 20、圖 21 中的數據，準確率大約是達到 0.95，且損失函數的圖形有明顯下降，但沒有到完全收斂。
2. 在圖 22 中，平均要 7 個字才能讓模型確定及分別正常文本與取代字文本，且數據非常分散。
3. 在圖 23 中，少量數據呈現波動狀態，不過大部分數據都是趨於穩定的。
4. 在圖 24 中，抽取的 5 筆數據的曲線幾乎是重疊的，平均 1 次修改後可以使模型預測讓他變成通順的句子。
5. 在圖 25 的修正結果中，修正後的句子有著不錯的成效，幾乎都有改掉原句中怪異的部分。
6. 整體來說，取代字文本本來就只會有一個地方是有問題的，所以圖 22 中的曲線分布很開是正常的，並且在大部分的修正結果中，模型第一次修正的字幾乎都是文本中被取代的那個字，可見對於這部分，模型判斷的能力其實相當不錯。

四、D 組

(一) 模型訓練結果

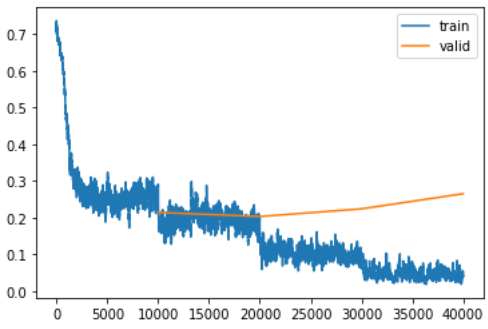


圖 26、D 組模型 loss 函數圖形。

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.283963	0.214686	0.920300	0.922380	0.902172	0.943515
1	0.178838	0.203581	0.924350	0.924429	0.926976	0.921897
2	0.097564	0.224432	0.924700	0.927485	0.897577	0.959454
3	0.041160	0.265586	0.928450	0.930456	0.908340	0.953676

圖 27、D 組模型訓練過程中的各項數值。

(二) 句子得分趨勢變化

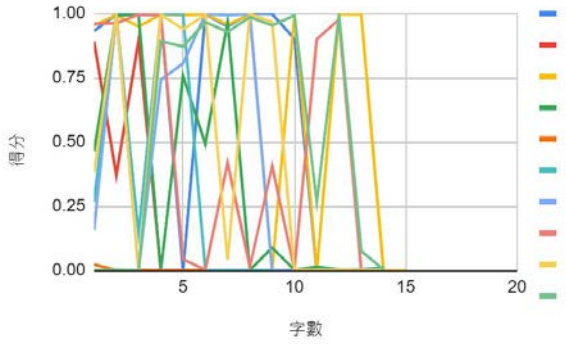


圖 28、交換字文本在 D 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

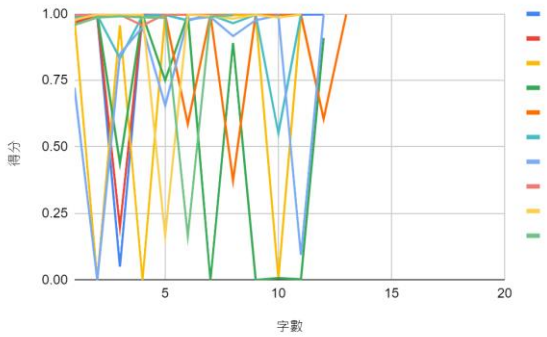


圖 29、正常文本在 D 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

(三) 句子修改結果

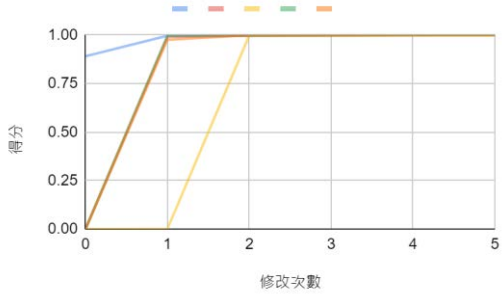


圖 30、交換字文本在 D 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	8.91E-01	試圖闖入隧渡偷道進入英國	1.70E-03	涉事士巴左邊車頭損毀	1.54E-04	由行本東寶公司發日
1	0.9961606264	試圖闖入隧道偷道進入英國	0.9916195869	涉事台巴左邊車頭損毀	8.80E-04	由日本東寶公司發日
2	0.997517705	試圖闖入隧道偷偷進入英國	0.9961004257	說涉事台巴左邊車頭損毀	0.9985841513	由日本東寶公司首發日
3	0.9987331033	試圖闖入隧道偷車進入英國	0.9969839454	說涉事微巴左邊車頭損毀	0.9989596605	日本東寶公司首發日
4	0.999373734	試圖闖入地道偷車進入英國	0.9986801744	說涉事微巴左邊車輛損毀	0.999330759	日本東寶公司首發日本
5	0.999373734	試圖闖入地道偷車進入英國	0.9994012117	說涉事微巴前邊車輛損毀	0.999391675	日日本東寶公司首發日本

圖 31、交換字文本在 D 組模型下，修正句子的結果（隨機抽樣 3 組）。

(四) 結果與討論

1. 由圖 26、圖 27 中的數據，準確率大約是達到 0.92，且損失函數的圖形有明顯下降，但沒有到完全收斂。
2. 在圖 28 中，平均要 7 個字才能讓模型確定及分別正常文本與交換字文本，且數據非常分散。
3. 在圖 29 中，大部分數據呈現波動狀態。
4. 在圖 30 中，平均 1 次修改後可以使模型預測讓他變成通順的句子。
5. 在圖 31 的修正結果中，雖然單看每個詞之間效果其實還好，修正後的句子有些奇怪的部分。
6. 整體來說，交換字文本平均。1 次修改後可以使模型預測讓他變成通順的句子本身就很奇怪，雖然也與重組句子一樣不太能保證交換後句子一定是不通順的，但我想大部分應該是，所以這個部分還是有待加強，修正出來的結果不是很好。

五、E 組

(一) 模型訓練結果

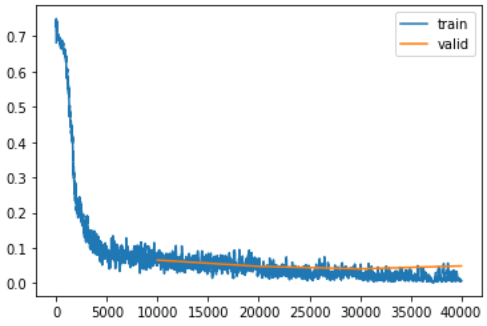


圖 32、E 組模型 loss 函數圖形。

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.058331	0.065126	0.981000	0.980881	0.988440	0.973437
1	0.047693	0.047370	0.987150	0.987100	0.992330	0.981925
2	0.019742	0.040329	0.988600	0.988597	0.990281	0.986918
3	0.005956	0.048487	0.988400	0.988385	0.991064	0.985720

圖 33、E 組模型訓練過程中的各項數值。

(二) 句子得分趨勢變化

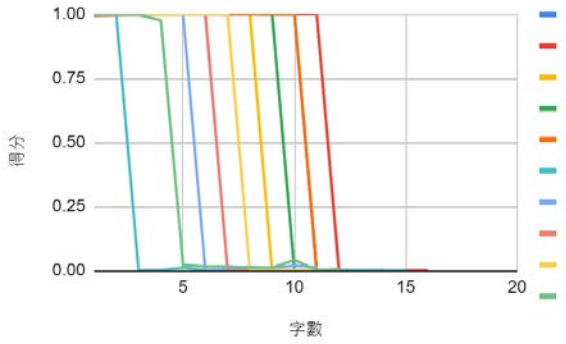


圖 34、重複字文本在 E 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

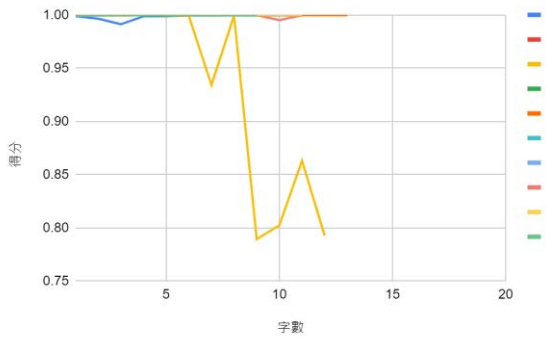


圖 35、重複字文本在 E 組模型下，得分對取句子的前幾個字數之關係圖（隨機抽樣 10 組）。

(三) 句子修改結果

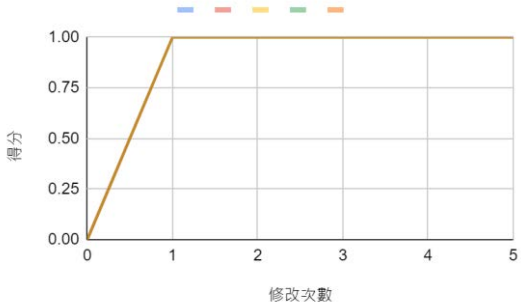


圖 36、重複字文本在 E 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	3.74E-04	向全球釋出了旅行警告告	3.40E-04	在採商標註冊標示國國家	8.16E-04	臺灣著名環保社運運人士
1	0.9997196794	向全球釋出了旅行警告。	0.9996808767	在採商標註冊標示國家	1.00E+00	臺灣著名環保社載運人士
2	0.9997413754	向全面釋出了旅行警告。	0.9997175336	在採商標註冊號示國家	0.9997051358	臺名著名環保社載運人士
3	0.9997474551	向全面釋入了旅行警告。	0.9997299314	在採中標註冊號示國家	0.9997124076	北名著名環保社載運人士
4	0.9997518659	向全面釋入了旅行警報。	0.9997460246	在採中標熱冊號示國家	0.9997270703	北名著名環游社載運人士
5	0.9997562766	向全面釋入了旅者警報。	0.9997523427	在採中標熱號示國家	0.9997305274	北名著名環游處載運人士

圖 37、重複字文本在 E 組模型下，修正句子的結果（隨機抽樣 3 組）。

(四) 結果與討論

1. 由圖 32、圖 33 中的數據，準確率大約是達到 0.98，且損失函數的圖形有明顯下降，並且達到收斂。
2. 在圖 34 中，平均要 8 個字才能讓模型確定及分別正常文本與交換字文本，且數據非常分散，大多都是過某個門檻後直線下降。
3. 在圖 35 中，大部分數據都是平穩的，且都非常接近 1。
4. 在圖 36 中，平均 1 次修改後可以使模型預測讓他變成通順的句子。
5. 在圖 37 的修正結果中，修正出來的句子有些奇怪，但幾乎都有改到重複字的部分。
6. 整體來說，模型是有辦法透過重複字來判斷句子通順程度的，且幾乎都可以一次就抓出問題點來，但可能由於這樣使得句子不夠亂，讓模型修正出來的效果並不佳。

六、F 組

(一) 模型訓練結果

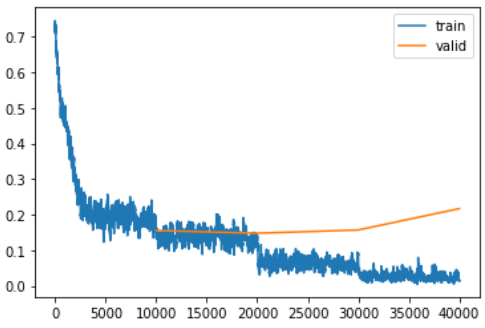


圖 38、F 組模型 loss 函數圖形。

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.162244	0.155350	0.942950	0.942417	0.945998	0.938864
1	0.134822	0.147745	0.947500	0.948010	0.933860	0.962594
2	0.081540	0.156924	0.953250	0.953522	0.942882	0.964404
3	0.014955	0.216934	0.953900	0.954252	0.941914	0.966918

圖 39、F 組模型訓練過程中的各項數值。

(二) 句子修改結果

1. 隨機字文本組

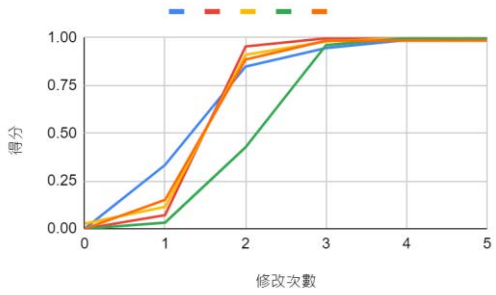


圖 40、隨機字文本在 E 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	扜鄭淌榎嗵箭	0.000105546860	琠聧蹇賡劘擲	0.000152963810	綠劻耜縑麑論	0.02696350031
1	扜鄭大榎嗵箭	0.333951056	琠聧蹇賡所擲	0.07172112167	綠原耜縑麑論	0.1149136871
2	扜鄭大洗嗵箭	0.8479048014	琠聧者蹇賡所擲	0.9528599977	綠原堂縑麑論	0.9095059633
3	鄭大洗嗵箭	0.9450961351	琠聧者為蹇賡所擲	0.9957187772	綠原堂縑下論	0.9789754748
4	鄭大洗台箭	0.9869794846	琠聧者為蹇嘉賡所擲	0.9974491	綠原堂記下論	0.9962492585
5	大洗台箭	0.9911182523	琠聧者同為蹇嘉賡所擲	0.9983223081	綠原堂記下	0.9997468591

圖 41、隨機字文本在 E 組模型下，修正句子的結果（隨機抽樣 3 組）。

2. 重組句子文本組

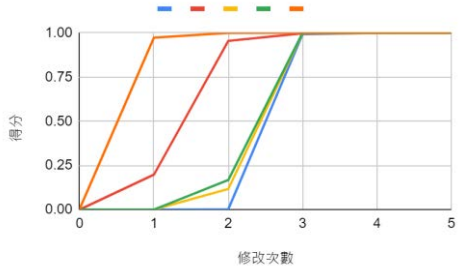


圖 42、重組句子文本在 E 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	由解並散聯東德州邦年	2.79E-05	望四界及第道大道交展大	0.002652259544	流民有漁中傳間川民四	2.78E-05
1	由解並散聯東德州半年	0.000209883088	望四界及天道大道交展大	0.9722307324	流民有漁傳間川民四	3.61E-05
2	由理解並散聯東德州半年	0.1678997427	望四界及天道大道交展	0.9994525313	流民有漁子間川民四	0.00103424699
3	由理解並扣聯東德州半年	0.9995443225	望四界及天路大道交展	0.9997809529	流民有漁子間川的四	0.9917803407
4	由理解並扣聯東德半年	0.9998358488	望四界及天藍大道交展	0.9998074174	流民只有漁子間川的四	0.9996273518
5	由理解並扣聯西德半年	0.9998402596	望四界及天藍道交展	0.999826014	流民只有子間川的四	0.999802053

圖 43、重組句子文本在 E 組模型下，修正句子的結果（隨機抽樣 3 組）。

3. 取代字文本組

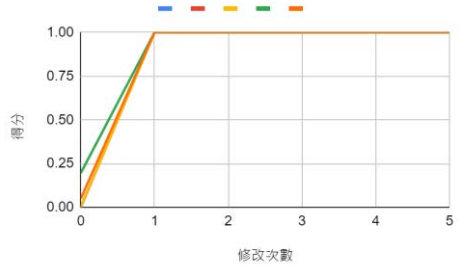


圖 44、取代字文本在 E 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	他成為荷屬東印度棲島	3.68E-05	少數公喚為巴哈伊教徒	0.1965176165	筧夜的長短等都有論述	0.04981965572
1	他成為荷屬東印度群島	0.9998494387	少數公司為巴哈伊教徒	0.999802053	夜的長短等都有論述	0.999333322
2	他身為荷屬東印度群島	0.9998590946	多數公司為巴哈伊教徒	0.9998390675	月的長短等都有論述	0.9997649789
3	他作為荷屬東印度群島	0.9998655319	多數公司為巴哈伊教徒的	0.9998452663	月的長短等沒有論述	0.9998402596
4	他作為荷屬東印度島	0.9998675585	多數公司為巴基伊教徒的	0.9998482466	月的長短等沒有前述	0.9998534918
5	他作為荷屬東印度	0.9998691082	多數公司為巴基斯教徒的	0.9998494387	月的長短沒有前述	0.9998584986

圖 45、取代字文本在 E 組模型下，修正句子的結果（隨機抽樣 3 組）。

4. 交換字文本組

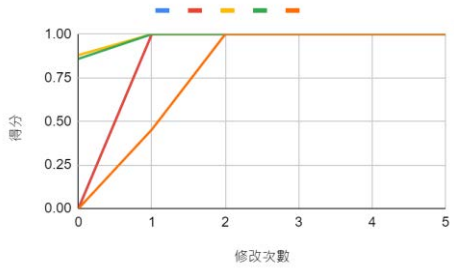


圖 46、交換字文本在 E 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	試圖闖入隧渡偷道進入英國	0.8799227476	涉事士巴左邊車頭損毀	0.8584814668	由行本東寶公司發日	3.39E-05
1	試圖闖入偷渡偷道進入英國	0.9997197986	涉事台巴左邊車頭損毀	0.9998452663	由行本東寶公發日	0.4537761509
2	試圖闖入偷渡黑道進入英國	0.9998292923	涉事台巴左邊有車頭損毀	0.9998545647	由日本東寶公發日	0.9998482466
3	來試圖闖入偷渡黑道進入英國	0.9998332262	涉事台巴旁邊有車頭損毀	0.9998557568	由日本東寶公發	0.9998488426
4	試圖闖入偷渡黑道進入英國	0.9998352528	涉事台巴旁邊有車尾損毀	0.9998574257	由日本東寶公布	0.9998534918
5	試圖闖著偷渡黑道進入英國	0.9998421669	涉事台巴旁邊還有車尾損毀	0.9998574257	由日本東寶公布他	0.9998618364

圖 47、交換字文本在 E 組模型下，修正句子的結果（隨機抽樣 3 組）。

5. 重複字文本組

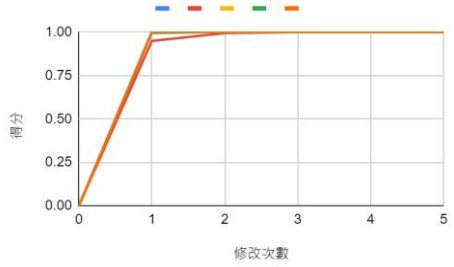


圖 48、重複字文本在 E 組模型下，句子得分對修正次數之關係圖（隨機抽樣 5 組）。

	1		2		3	
0	向全球釋出了旅行警告告	3.22E-05	在採商標註冊標示國國家	7.20E-05	臺灣著名環保社運運人士	3.23E-05
1	向全球釋出了旅行警告	0.9998421669	在採商標註冊標示，國家	0.9941913486	臺灣著名環保社運人士	0.9998505116
2	向全球釋出的旅行警告	0.9998499155	在商標註冊標示，國家	0.9998415709	臺灣著名環保社運人士	0.9998505116
3	來向全球釋出的旅行警告	0.9998568296	在商標註冊台標示，國家	0.9998623133	臺灣著名環保社運人士	0.9998505116
4	來向全球釋出之旅行警告	0.9998606443	在商標註冊台標示，國	0.9998649359	臺灣著名環保社運人士	0.9998505116
5	來向全球釋出之旅行預告	0.9998629093	在商標註冊台標示是國	0.9998681545	臺灣著名環保社運人士	0.9998505116

圖 49、重複字文本在 E 組模型下，修正句子的結果（隨機抽樣 3 組）。

伍、 討論

一、 隨機字文本不管在單獨或是混和模型中，其表現的修正效率都一樣是不好，其原因可能在於他本來就是沒什麼意義的句子，修正起來本身就會有點吃力。

二、 重組句子文本在修正句子的成效來說，單獨模型中的表現上比混和模型來的好，或許他的雜亂形式有被其他組文本給影響到，這部分仍未確認。

三、取代字文本、重複字文本與交換字文本在混和模型中的表現皆不給其單獨模型出來的結果，這幾種文本本身怪異的地方就不多，也可能是因為如此模型才有辦法準確掌握出怪異處並加以修正，對於這個結果，此模型可能也比較擅長處理這些相關問題的句子。

四、由於此種修正方式仍然侷限在 BERT 的 MLM 任務中所預測的結果，所以當 BERT 給出的結果本來選項就很奇怪時，模型也只能挑選他認為的最佳解進行修正。

五、若要使如重組句子文本等成效不佳的句子進行好的修正，可能須嘗試調整混和文本的比例。

陸、 結論

使用 BERT 進行 fine-tuning 來修正句子可能還有其他更好的訓練文本可以來嘗試使用，由於 BERT 模型在自然語言理解上本來就有一定的水準，其訓練出來的模型在準確率方面都不會顯得太差。

本研究所製作的模型雖然無法應用在所有需要修正句子的生活情況下，不過其模型在一些特定的小範圍上如：挑出多餘的字、修改少量錯字、修改少量不合理的字的作用上，有著還不錯的成效，未來也可以考慮加入關於語句矛盾的相關模型來進一步驗證句子的語義，或是加入不同的文本來做訓練，希望可以使其有更廣泛的應用範圍及成果。

柒、 參考文獻資料

[1] Clay(2019) [NLP][Python] 英文自然語言處理的經典工具 NLTK
https://clay-atlas.com/blog/2019/07/30/nlp-python-cn-nltk-kit/
[2] Desolve(2020) [Day 28] 從零開始學 Python - 深度學習 Keras：如果你能預知這條路的陷阱，我想你依然錯得很過癮
https://ithelp.ithome.com.tw/articles/10247304
[3] Maximilien Roberti(2019) Fastai with 🤗Transformers (BERT, RoBERTa, XLNet, XLM, DistilBERT)
https://towardsdatascience.com/fastai-with-transformers-bert-roberta-xlnet-xlm-distilbert-4f41ee18ecb2
[4]Harika Bonthu(2021):Python Tutorial: Working with CSV file for Data Science
https://www.analyticsvidhya.com/blog/2021/08/python-tutorial-working-with-csv-file-for-data-science/
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova(2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/pdf/1810.04805.pdf
[6] fastai
https://docs.fast.ai/?fbclid=IwAR3sD2_-lVdXaf2zS_PfZvlKwP6L5oHehXIjOwAw3HKB2k2g3YcJoYz5_l8
[7] FastHugs
https://aikindergarten.github.io/fasthugs/?fbclid=IwAR3AELB12T-gsfcUj7lAN81yoP2asWf9O9ec1hWL5mZErXwoReQ26Idn-2Q
[8] 陳柏霖 Po-Lin Chen, *吳世弘 Shih-Hung Wu(2015) 以語言模型判斷學習者文句流暢度
https://aclanthology.org/O15-1021.pdf

【評語】052508

本作品探討在各種不通順的文具情況下，是否能透過 BERT 模型進行通順度判斷，進一步改善其中之不通順之文句。研究動機很務實，實驗設計多樣且討論充分，對現有之 BERT 模型的內容與功能有充分之理解，並能有效利用其能力進行後續之文句修改。

研究過程以五種人工合成式的文本進行實驗並個別分析討論，多組實驗中皆探討實驗成效，討論內容詳盡。然而有些合成資料與目標情境有較大差異，建議將此技術嘗試在較接近目標情境的資料上，讓作品與動機更加契合。且可在討論中多說明或分析中英文情境的差異，能夠讓實驗的廣度增加，也讓實驗的討論更加完善。

作品簡報

換句話說

高級中等學校組 電腦與資訊學科

摘要

BERT近年來在各式NLP任務中可說是無處不見、無所不在，其中使用fine-tuning的訓練方式更是可以幫助研究者省下大量的時間及運算成本，且結果都有不錯的表現。本研究探討在結合不同條件的文本訓練下，基於對BERT模型做fine-tuning且讓其進行文本分類，觀察其對於預測及分類中文句子通順程度的成效，並且根據訓練出來的模型設計修正方式嘗試使其對預測中不通順之文本進行自我修正，並分析其成效與結果。

研究動機

在這個資訊大多可以從網路上獲得的世代，在瀏覽網路文章中常常會有讀到一半結果就發現「這個字好像打錯了」或是「這裡是不是多打了一個字」之類的情況發生，導致句子變得很奇怪，於是我想製作出一個模型嘗試找出這種奇怪的句子，並且嘗試去修正他。

研究目的

1.嘗試找出不通順的文句並在不涉及語意矛盾等情況下加以修正
2.探討不同類型不通順文本對於修正句子通順程度的影響
3.探討不同類型不通順文本訓練出的模型之表現
4.探討將不同類型不通順文本合併後訓練模型的成效

研究設備及器材

一、平台：

Google Colab(GPU)

二、資料儲存：

Google雲端

研究過程與方法

訓練資料集

本研究選用之資料集為維基百科中文資料庫20210920版本（zhwiki-20210920-pages-articles-multistream.xml.bz2）。

連續的量即是以實數來表示的
是從先秦散文中的寓言濫觴
這些情節不只是娛樂性的
年一篇由蘇美人創作的
世紀發展了現代史學方法
固然是因為表演者是歌者
在歌曲中這種作用表現得最為突出
使用者用此方法來估計規模
及發展了以比較方式研究政治
法律具有超乎個人的自身價值

正常文本

鋩蕑顡麒餅蒿祀盳萎餉邈劉
枎碤纐霿亍珣輵侄忑馗婑萬梓
烿狻錠救鯇釩杳姻埡苔
坱清堅鸖縓憔繆崽隮陵宥虢暍畁杪殆
媕溝柱猔衲矯嶐勭篺瘏信僟譓隀縺磐忻椵禃町
赳衕蓀居滑嗡縚蘢鑿緒蒲
櫛儀爂餅圀顏酟壉閬椿創丷杤
鏂瀛崗曝棟鉶咪鐅索詎駟叚屜嵓警
塔扻梔諫籙姙祼藰藤踽腑篢栖崜
飧搒玅薈踤郅蚭格骯澤樀驊踖犭霰

隨機字文本

抒最來騷離長的情詩以早篇
附影的下上電邊放播都在幕會字
指資上及差以源配經標分異等的濟
較灣臺具法目表學的院前性代
的奠定論理天了天礎氣基
繼指相斯高拉普和出斯拉
商或業性農高自性業給
然自自別概於念環境個源
間鮑一著能的塔姆論名之爭場
士位茲碩了的受他接託瓦學

重組句子文本

那麼所演出的便不再是戲悠
戲劇可以被分成不同的型鴿
復旦國際範係與公共事務學院
分為蓑般法學與特別法學兩種
化學的研究範螪是包括分子
地球歷史上曾發生過多次大的乏難
大氣和海洋麢緊密聯絡的
由於這個過程枚以不斷的迴圈發生
更意味著牓流量的差距
暱為資料結構概念的普及

取代字文本

沒觧一章特別談及耶有
文學手段是在敘述中用的結殊特構
音樂是一種需要學習的能技
和描述政府向作的情況為方運
別中又可分門其類研習
他所指的是暗孔相機或針箱
著為不正確的理論支配並
動活論隨著板塊構造學說的發展
而細胞是也許多的基本單位
應付有學習障礙的生學

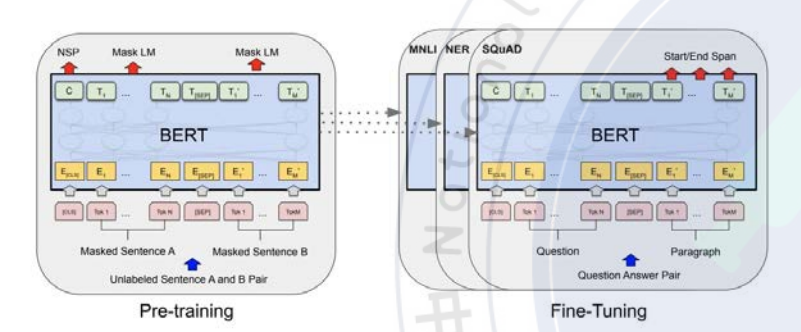
交換字文本

數學定律越和現實有關關
而且有很詳詳細的資料記錄
文學寫作作是一種藝術形式
對我們這個時代的的歐洲人而言
或者自身意識識形態和價值立場判斷
內內容與基督教教義衝突甚大
目前臺灣較具代表性的的法學院
研究架構等一直直經歷著擴充套件和分歧
也表述出內含的的數學概念
行星形成時碰撞的殘留熱熱量

重複字文本

研究過程與方法

BERT



MLM任務

BERT的作者們的論文中使用了 Transfomer Encoder，透過大量文本做了兩個預訓練目標，其中一個就是MLM（Masked Language Model）任務，又稱克漏字填空。透過遮罩（[Mask]），可使其Attention（注意力機制）關注在特定的字詞上面，從而預測該遮罩可能出現之字詞。

Fine-tuning

Fine-tuning是指使用已經完成訓練的神經網絡模型，用來進一步執行其他類似任務的過程，採用已經設計和訓練好的神經網路能夠更有效率地利用神經網絡前層先前訓練的特徵，而無需從頭訓練提取該特徵，只需要針對下游任務的classification layer進行fine tune。

模型訓練資料分配

分成A到F組模型，固定每組模型訓練資料固定總合為100000個句子，並且包含通順及不通順的句子各占一半。

句子通順度定義

經過fine-tuning後的BERT模型可以進行文本分類，會對於0（不通順）和1（通順）分別給出一個0到1的結果，並且相加起來為1，所以在此將訓練後模型預測的「句子通順度」定義為模型預測是該句子為1的結果

研究過程與方法

句子修正

本研究提出了一種方法來嘗試使模型達到修正句子的效果。

1.MASK修正：將要修正的句子輪流MASK每個字，交給BERT做MLM任務預測最可能的字的Top5，並且一一套入句子中預測句子通順度。

「今天天氣真好」會被依序變成：[MASK]天天氣真好→今[MASK]天氣真好→今天[MASK]氣真好→…→今天天氣真[MASK]

2.增字：將要修正的句子，在所有可以加入[MASK]的地方加入[MASK]，再交由BERT做MLM任務預測最可能的字的Top5，並且一一套入句子中預測句子通順度。

「今天天氣真好」會被依序變成：[MASK]今天天氣真好→今[MASK]天天氣真好→今天[MASK]天氣真好→…→今天天氣真好[MASK]

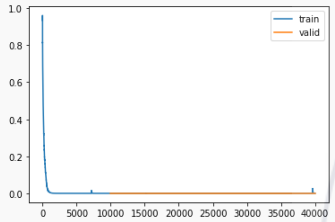
3.減字：將要修正的句子，輪流移除每個字，並且一一預測句子通順度。

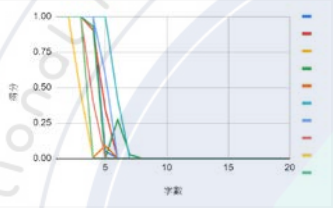
「今天天氣真好」會被依序變成：天天氣真好→今天氣真好→今天氣真好→…→今天天氣真

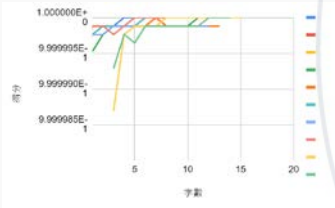
接著將上面三個步驟中得分最高的句子，作為下一次修正句子的輸入，如此重複5次。

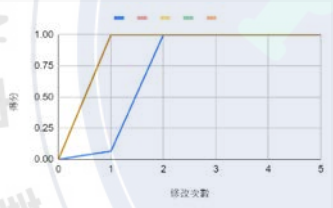
研究結果

A組





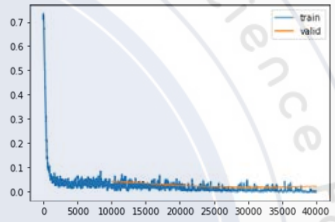


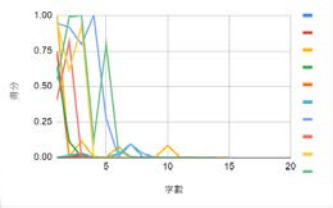


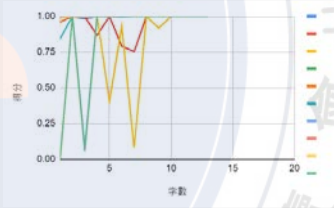
epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
1	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
2	0.000000	0.000000	1.000000	1.000000	1.000000	1.000000
3	0.000021	0.000000	1.000000	1.000000	1.000000	1.000000

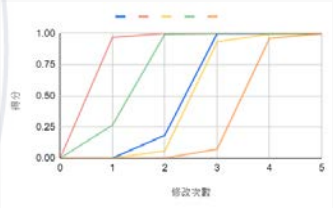
	1	2	3
0	2.64E-06	2.39E-04	4.01E-06
1	0.0662060231	0.9999997616	1.00E+00
2	0.9999940395	0.9999998808	0.9999997616
3	0.9999998808	0.9999998808	0.9999998808
4	0.9999998808	1	0.9999998808
5	0.9999998808	1	0.9999998808

B組







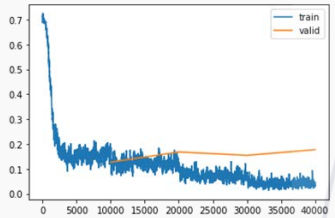


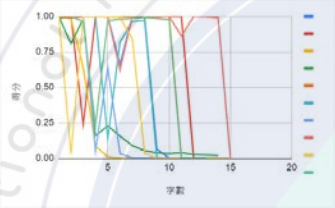
epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.024608	0.044928	0.983700	0.983930	0.969214	0.999099
1	0.020967	0.026516	0.991400	0.991443	0.985461	0.997497
2	0.010549	0.018208	0.994050	0.994038	0.994985	0.993092
3	0.004104	0.023938	0.993650	0.993670	0.989476	0.997898

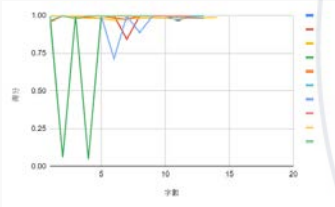
	1	2	3
0	1.11E-04	5.86E-04	1.08E-04
1	0.0002663404	0.9685303569	6.21E-04
2	0.1847215444	0.9999251366	0.0584263689
3	0.9996705055	0.9999405146	0.9331876636
4	0.99993789	0.9994874	0.9944880605
5	0.9999439716	0.9999542236	0.9993000031

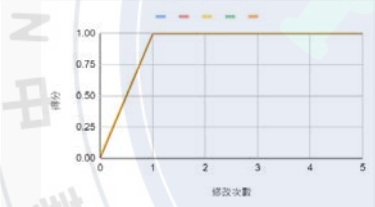
研究結果

C組





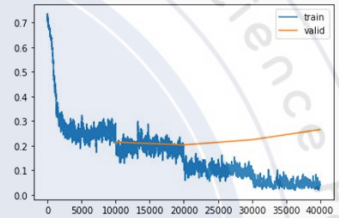


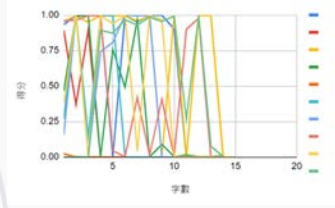


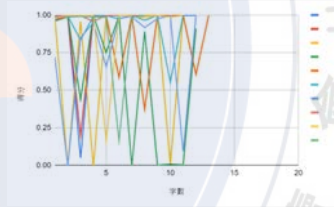
epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.176744	0.126831	0.955200	0.955788	0.949975	0.961672
1	0.111705	0.167953	0.942250	0.944808	0.910648	0.981630
2	0.070476	0.154347	0.956500	0.957126	0.950103	0.964254
3	0.043862	0.177012	0.956300	0.957278	0.942717	0.972297

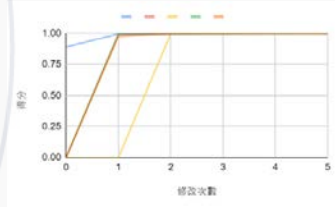
	1	2	3
0	2.13E-03 他成為荷屬東印度棟島	2.47E-03 少數公喚為巴哈伊教徒	3.73E-02 覺夜的長短等都有論述
1	0.9983059168 他成為荷屬東印度群島	0.9983320832 少數公司為巴哈伊教徒	9.98E-01 夜的長短等都有論述
2	0.998354733 他成為荷屬印度群島	0.9983769655 少數公司為巴哈伊信徒	0.9984662533 的長短等都有論述
3	0.998370707 他成為荷蘭印度群島	0.9983897209 少數公司為巴伊信徒	0.9984869957 的長短都有論述
4	0.9983769655 和成為荷蘭印度群島	0.9984021783 少數公司是巴伊信徒	0.9984869957 長短都有論述
5	0.9983897209 和已成為荷蘭印度群島	0.998411417 少數公司是巴利信徒	0.9984929562 長短有論述

D組







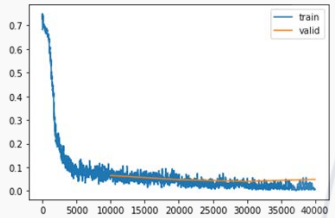


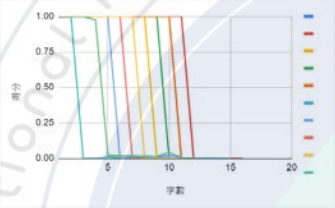
epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.283963	0.214686	0.920300	0.922380	0.902172	0.943515
1	0.178838	0.203581	0.924350	0.924429	0.926976	0.921897
2	0.097564	0.224432	0.924700	0.927485	0.897577	0.959454
3	0.041160	0.265586	0.928450	0.930456	0.908340	0.953676

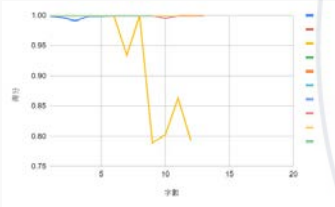
	1	2	3
0	8.91E-01 試圖闖入隧渡偷道進入英國	1.70E-03 涉事士巴左邊車頭損毀	1.54E-04 由行本東寶公司發日
1	0.9961606264 試圖闖入隧道偷道進入英國	0.9916195869 涉事台巴左邊車頭損毀	8.80E-04 由日本東寶公司發日
2	0.997517705 試圖闖入隧道偷偷進入英國	0.9961004257 說涉事台巴左邊車頭損毀	0.9985841513 由日本東寶公司首發日
3	0.9987331033 試圖闖入隧道偷車進入英國	0.9969839454 說涉事微巴左邊車頭損毀	0.9989596605 日本東寶公司首發日
4	0.999373734 試圖闖入地道偷車進入英國	0.9986801744 說涉事微巴左邊車輛損毀	0.999330759 日本東寶公司首發日本
5	0.999373734 試圖闖入地道偷車進入英國	0.9994012117 說涉事微巴前邊車輛損毀	0.999391675 日日本東寶公司首發日本

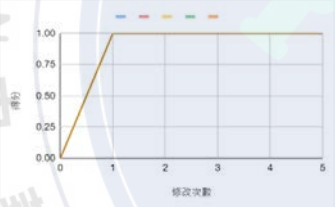
研究結果

E組







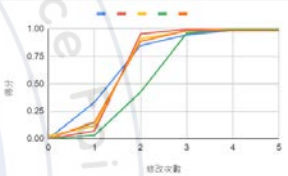


epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.058331	0.065126	0.981000	0.980881	0.988440	0.973437
1	0.047693	0.047370	0.987150	0.987100	0.992330	0.981925
2	0.019742	0.040329	0.988600	0.988597	0.990281	0.986918
3	0.005956	0.048487	0.988400	0.988385	0.991064	0.985720

	1		2		3	
0	3.74E-04	向全球釋出了旅行警告告	3.40E-04	在採商標註冊標示國國家	8.16E-04	臺灣著名環保社運人士
1	0.9997196794	向全球釋出了旅行警告，	0.9996808767	在採商標註冊標示國家	1.00E+00	臺灣著名環保社會運人士
2	0.9997413754	向全面釋出了旅行警告告，	0.9997175336	在採商標註冊號示國家	0.9997051358	臺名著名環保社會運人士
3	0.9997474551	向全面釋人了旅行警告，	0.9997299314	在採中標註冊號示國家	0.9997124076	北名著名環保社會運人士
4	0.9997518659	向全面釋人了旅行警報，	0.9997460246	在採中標熱冊號示國家	0.9997270703	北名著名環游社會運人士
5	0.9997562766	向全面釋人了旅者警報，	0.9997523427	在採中標熱號示國家	0.9997305274	北名著名環游處會運人士

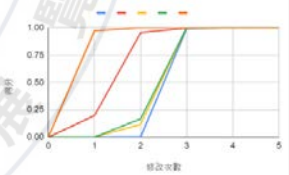
F組

隨機字文本



	1		2		3	
0	扦鄭淌榎嗵劑	0.000105546860	璿聯蔓慶劇擲	0.000152963810	緑効粗縁覭論	0.02696350031
1	扦鄭大榎嗵劑	0.333951056	璿聯蔓慶所擲	0.07172112167	緑原粗縁覭論	0.1149136871
2	扦鄭大洗嗵劑	0.8479048014	璿聯者蔓慶所擲	0.9528599977	緑原堂縁覭論	0.9095059633
3	鄭大洗嗵劑	0.9450961351	璿聯者為蔓慶所擲	0.9957187772	緑原堂縁下論	0.9789754748
4	鄭大洗台劑	0.9869794846	璿聯者為蔓嘉慶所擲	0.9974491	緑原堂記下論	0.9962492585
5	大洗台劑	0.9911182523	璿聯者同為蔓嘉慶所擲	0.9983223081	緑原堂記下	0.9997468591

重組句子文本

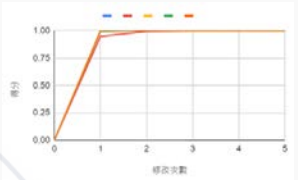


	1		2		3	
0	由解並散聯東德州邦年	2.79E-05	望四界及第道大道交展大	0.00265225954	流民有漁中傳間川民四	2.78E-05
1	由解並散聯東德州半年	0.000209883088	望四界及天道大道交展大	0.9722307324	流民有漁傳間川民四	3.61E-05
2	由理解並散聯東德州半年	0.1678997427	望四界及天道大道交展	0.9994525313	流民有漁子間川民四	0.00103424699
3	由理解並扣聯東德州半年	0.9995443225	望四界及天路大道交展	0.9997809529	流民有漁子間川的四	0.9917803407
4	由理解並扣聯東德半年	0.9998358488	望四界及天藍大道交展	0.9998074174	流民只有漁子間川的四	0.9996273518
5	由理解並扣聯西德半年	0.9998402596	望四界及天藍道交展	0.999826014	流民只有子間川的四	0.999802053

研究結果

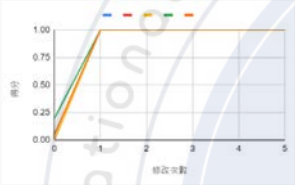
F組

隨機字文本

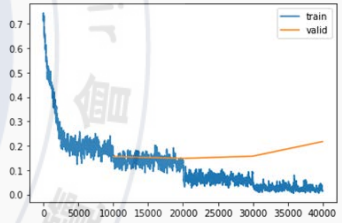


	1		2		3	
0	向全球釋出了旅行警告告	3.22E-05	在採商標註冊標示國國家	7.20E-05	臺灣著名環保社運運人士	3.23E-05
1	向全球釋出了旅行警告	0.9998421669	在採商標註冊標示，國家	0.9941913486	臺灣著名環保社運人士	0.9998505116
2	向全球釋出的旅行警告	0.9998499155	在商標註冊標示，國家	0.9998415709	臺灣著名環保社運人士	0.9998505116
3	來向全球釋出的旅行警告	0.9998568296	在商標註冊台標示，國家	0.9998623133	臺灣著名環保社運人士	0.9998505116
4	來向全球釋出之旅行警告	0.9998606443	在商標註冊台標示，國	0.9998649359	臺灣著名環保社運人士	0.9998505116
5	來向全球釋出之旅行預告	0.9998629093	在商標註冊台標示是國	0.9998681545	臺灣著名環保社運人士	0.9998505116

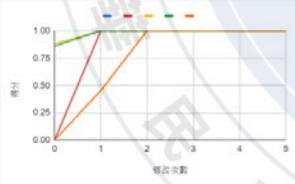
取代字文本組



	1		2		3	
0	他成為荷屬東印度棟島	3.68E-05	少數公喚為巴哈伊教徒	0.1965176165	篤夜的長短等都有論述	0.04981965572
1	他成為荷屬東印度群島	0.9998494387	少數公司為巴哈伊教徒	0.999802053	夜的長短等都有論述	0.999333322
2	他身為荷屬東印度群島	0.9998590946	多數公司為巴哈伊教徒	0.9998390675	月的長短等都有論述	0.9997649789
3	他作為荷屬東印度群島	0.9998655319	多數公司為巴哈伊教徒的	0.9998452663	月的長短等沒有論述	0.9998402596
4	他作為荷屬東印度島	0.9998675585	多數公司為巴基伊教徒的	0.9998482466	月的長短等沒有前述	0.9998534918
5	他作為荷屬東印度	0.9998691082	多數公司為巴基斯教徒的	0.9998494387	月的長短沒有前述	0.9998584986



交換字文本組



	1		2		3	
0	試圖闖入隧渡偷道進入英國	0.8799227476	涉事士巴左邊車頭損毀	0.8584814668	由行本東寶公司發日	3.39E-05
1	試圖闖入偷渡偷道進入英國	0.9997197986	涉事台巴左邊車頭損毀	0.9998452663	由行本東寶公發日	0.4537761509
2	試圖闖入偷渡黑道進入英國	0.9998292923	涉事台巴左邊有車頭損毀	0.9998545647	由日本東寶公發日	0.9998482466
3	來試圖闖入偷渡黑道進入英國	0.9998332262	涉事台巴旁邊有車頭損毀	0.9998557568	由日本東寶公發	0.9998488426
4	試圖闖入偷渡黑道進入英國	0.9998352528	涉事台巴旁邊有車尾損毀	0.9998574257	由日本東寶公布	0.9998534918
5	試圖闖著偷渡黑道進入英國	0.9998421669	涉事台巴旁邊還有車尾損毀	0.9998574257	由日本東寶公布他	0.9998618364

epoch	train_loss	valid_loss	accuracy	f1_score	precision_score	recall_score
0	0.162244	0.155350	0.942950	0.942417	0.945998	0.938864
1	0.134822	0.147745	0.947500	0.948010	0.933860	0.962594
2	0.081540	0.156924	0.953250	0.953522	0.942882	0.964404
3	0.014955	0.216934	0.953900	0.954252	0.941914	0.966918

討論

1.隨機字文本不管在單獨或是混和模型中，其表現的修正效率都一樣是不好，其原因可能在於他本來就是沒什麼意義的句子，修正起來本身就會有點吃力。

2.重組句子文本在修正句子的成效來說，單獨模型中的表現上比混和模型來的好，或許他的雜亂形式有被其他組文本給影響到，這部分仍未確認。

3.取代字文本、重複字文本與交換字文本在混和模型中的表現皆不給其單獨模型出來的結果，這幾種文本本身怪異的地方就不多，也可能是因為如此模型才有辦法準確掌握出怪異處並加以修正，對於這個結果，此模型可能也比較擅長處理這些相關問題的句子。

4.由於此種修正方式仍然侷限在BERT的MLM任務中所預測的結果，所以當BERT給出的結果本來選項就很奇怪時，模型也只能挑選他認為的最佳解進行修正。

5.若要使如重組句子文本等成效不佳的句子進行好的修正，可能須嘗試調整混和文本的比例。

6.使用這種方式修正句子也大概可以從修正出來的句子中了解出模型進行訓練出來的結果的好壞，他會表現修正出的結果。

7.當修正出的句子預測高於一定的值時，後面再繼續修正下去的結果會很差，但這不一定是模型不夠強的緣故，後面有極大可能都是浮點數的精度誤差，但實際上應當根據什麼做為分界還有待確認。

結論

使用BERT進行fine-tuning來修正句子可能還有其他更好的訓練文本可以來嘗試使用，由於BERT模型在自然語言理解上本來就有一定的水準，其訓練出來的模型在準確率方面都不會顯得太差。

本研究所製作的模型雖然無法應用在所有需要修正句子的生活情況下，不過其模型在一些特定的小範圍上如：挑出多餘的字、修改少量錯字、修改少量不合理的字的作用上，有著還不錯的成效，未來也可以考慮加入關於語句矛盾的相關模型來進一步驗證句子的語義，或是加入不同的文本來做訓練，希望可以使其有更廣泛的應用範圍及成果。

由於本研究使用之文本為由程式所輔助生成的「不通順」文本，其本身會限制住能包含的「不通順」的種類，故結果不能直接套用到大數據上面，但是這種方式再目前所測試出來的少數據中執行原本任務的能力效果都還不錯，若有機會拿到其他非程式而是人工所製造出的錯誤文本，那訓練出的模型應當會更加強大並且在生活中有更多應用。

參考文獻資料

[1] Clay(2019) [NLP][Python] 英文自然語言處理的經典工具 NLTK
https://clay-atlas.com/blog/2019/07/30/nlp-python-cn-nltk-kit/
[2] Desolve(2020) [Day 28] 從零開始學Python - 深度學習Keras：如果你能預知這條路的陷阱，我想你依然錯得很過癮
https://ithelp.ithome.com.tw/articles/10247304
[3] Maximilien Roberti(2019) Fastai with 🤗Transformers (BERT, RoBERTa, XLNet, XLM, DistilBERT)
https://towardsdatascience.com/fastai-with-transformers-bert-roberta-xlnet-xlm-distilbert-4f41ee18ecb2
[4]Harika Bonthu(2021):Python Tutorial: Working with CSV file for Data Science
https://www.analyticsvidhya.com/blog/2021/08/python-tutorial-working-with-csv-file-for-data-science/
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova(2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/pdf/1810.04805.pdf
[6] fastai
https://docs.fast.ai/?fbclid=IwAR3sD2_-lVdXaf2zS_PfZvlKwP6L5oHehXIjOwAw3HKB2k2g3YcJoYz5_I8
[7] FastHugs
https://aikindergarten.github.io/fasthugs/?fbclid=IwAR3AELB12T-gsfcUj7lAN81yoP2asWf9O9ec1hWL5mZErXwoReQ26Idn-2Q
[8] 陳柏霖 Po-Lin Chen, *吳世弘 Shih-Hung Wu(2015) 以語言模型判斷學習者文句流暢度
https://aclanthology.org/O15-1021.pdf