二、研究計畫內容（以10頁為限）：

(一)摘要

本研究以「自動生成閱讀理解選擇題出題」為應用目標，運用人工智慧產生文本，以生成多選題的錯誤選項為方向進行研究。以短文，問題和答案為輸入，生成出多個錯誤選項，以增加選擇題的難度。

(二)研究動機與研究問題

近年來深度學習技術獲得廣大的吸引力與注目。特別是深度學習技術於影像相關之應用(如影像辨識、影像生成等)已邁向成熟化與商品化。然而，相較於影像應用之成熟，深度學習技術於自然語言文字處理方面則尚待努力與投入。目前自然語言處理應用較成熟的區塊為文本情感分析 (Sentiment Analysis)、文本分類(Text Classiication)等相對容易任務，於自然語言語句生成(Natural Language Generation) 此研究領域，則由於難度高並未有明顧之技術突破。自然語言語句生成技術主要著眼於特定應用中，利用人工智慧產生流暢且自然之語句。

然由於人類文字之使用千變萬化，文字順序調換則語意完全不同，甚或不成句無法閱讀，也因此輸出流暢且語意語法正確之語句，被視為困難之挑戰，過去也較少研究著墨與此。而近一兩年NLP領域有很大進展，其中原因便是採用大規模資料預訓練模型(Pretrained Model)，以及Transformer[10]架構中Self-Attention所帶來的好處。大規模的預訓練指的是用一些無監督的任務，比如挖空文本填空的方式去訓練語義模型。訓練這些任務往往會用到十分大規模的文本，讓模型在做其他任務之前，已經能學會一些語法語義資料，之後拿這個預訓練模型去訓練其他任務，例如情緒分類，NER，閱讀理解可以得到很大的優勢。Transformer[10]模型相比之前的CNN，RNN模型能更好學會上下文語意，加上預訓練模型的做法，使得這一類模型可以將許多自然語言任務(如Name-Entity Recognition、Semantic Role Detection)的效能都推進一大步。便其中具有代表性且廣為人知例子為BERT [5]。

有鑒於BERT模型在其他自然語言任務的效能提升，我們預計探索使用BERT模型來進行閱讀測驗中選擇題錯誤選項之生成(distractor generation, 簡稱DG)。具體而言，Distractor Generatiion任務中為給定一篇短文C，一個問題語句Q和一個對應問題語句Q之答案A，針對該問題Q和答案A生成與上下文相關且語法正確的錯誤選項。

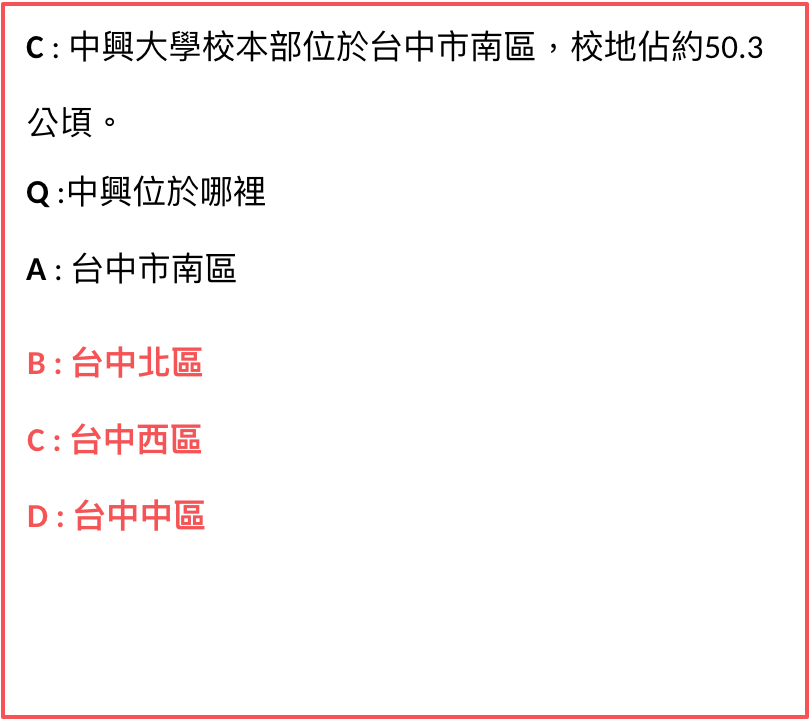


圖1: distractor 生成例子

如上圖，我們給定一短文C(中興大學之短文介紹)、一個問題語句Q(中興大學位於哪裡)以及一個標準答案A(台中市南區)。我們希望訓練出一個模型可以根據所給定之C、Q、A生成出錯誤答案選項(Distractor)。目前的研究主要著眼於生成流暢語義的DG，但有時候流暢並不代表這是一個好的結果：

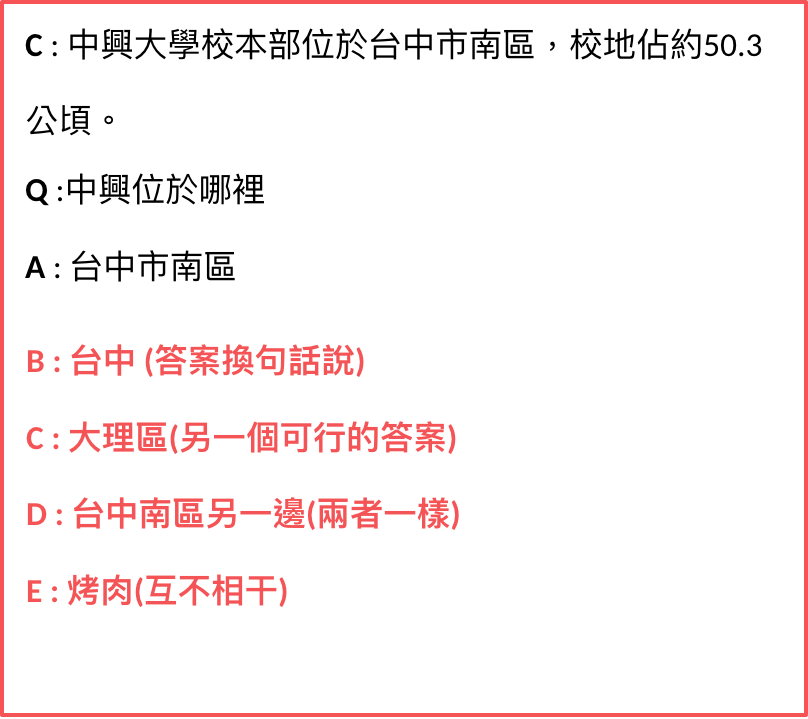


圖2: distractor 生成反面例子

可以從圖2看到，紅色文字的干擾項雖然語句通順，但十分容易分辨出這些都不是答案。因此在生成流暢性外，還需要額外考慮能否迷惑答題者。這個技術的完成將可減輕了教師在設計多選題選項的工作量。過去Distractor Generation (DG)，也吸引了相關研究社群的關注[1][2][3]，但經過初步的實驗與了解，我們發現相關研究仍有許多可以改善之處。具體而言，我們的研究預計解決以下問題:

* 生成與內文相關的干擾項
* 生成的干擾項之間不能有過高相似度
* 提出方法證明生成的干擾項具有迷惑性

(三)文獻回顧與探討

在本章節中，我們分兩子章節進行討論。於3.1節中，我們回顧BERT的基本架構。而於3.2節中，我們則回顧與探討現階段DG論文的做法以及目前發展的不足之處。

**3.1 BERT**

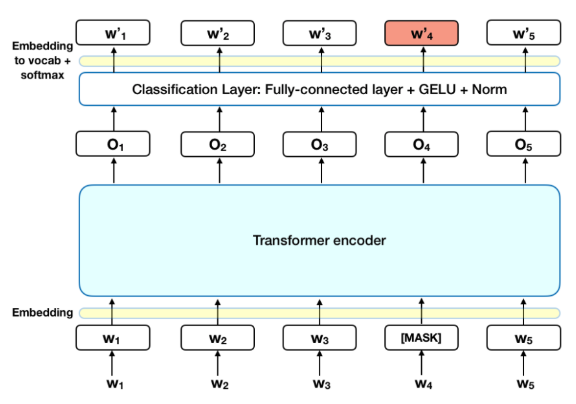


圖3 BERT的架構[5]

BERT是一個預訓練的語言模型，其目的是學到文本的特徵，使其能用以微調到其他任務中。BERT在預訓練階段會有兩個任務，一個是mask-lm，所做的事情是將文中的一些字mask起來，然後預測這個mask的字是什麼，另一個是Next Sentence Prediction，所做的則是預測兩個句子是不是上下文關係。

BERT是一種seq2seq的架構，input文本會是一段不超過512個字的句子，其中開頭用[CLS]表示，結尾用[SEP]表示，而要預測的文本則是用[MASK]表示。模型的輸出會是對[MASK]的預測結果。例如： [CLS]今天[MASK]情很好，預測的結果會是 心 。

**3.2 Distractor Generation**

據我們所知，閱讀測驗Distractor生成相關的研究為[1][2][3]。[1]所代表的是以Gap-Filling為目標的distractor生成，其主要想法為把distractor視為填空題，並在其中找尋候選的答案，答案的其他候選者便是distractor。

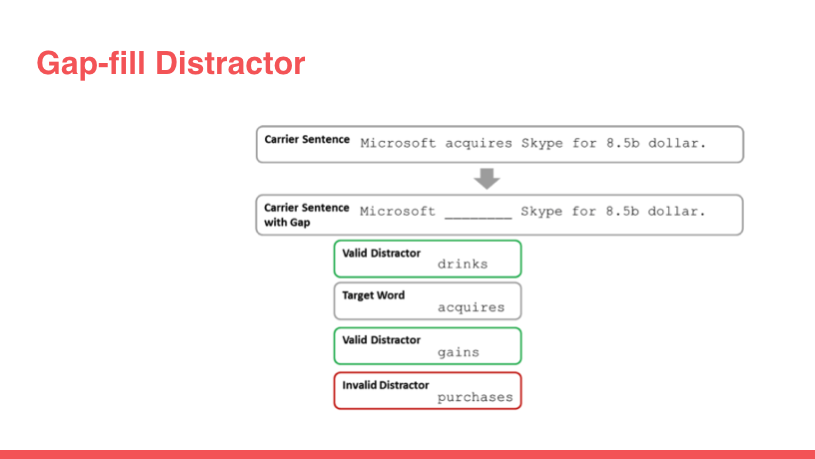


圖4: gap-fill distractor例子

圖2是gap-fill為目標的distractor生成例子，這類模型的做法會先把句子中的詞挖空，模型將會預測挖空的字是什麼，預測結果機率最高的頭幾項便會是干擾項。這一類的模型的做法是找相近的候選單字/短語。但這種做法可能找到同義詞，或者跟原文完全無關的結果，如圖中紅色框的結果，和目標的意思一致，便是一個無效的distractor。因此需要研究怎麼篩選出所能用作干擾項的結果。但這一類以篩選為主的方法，並不適合生成句子層級，富含語義的distractor。

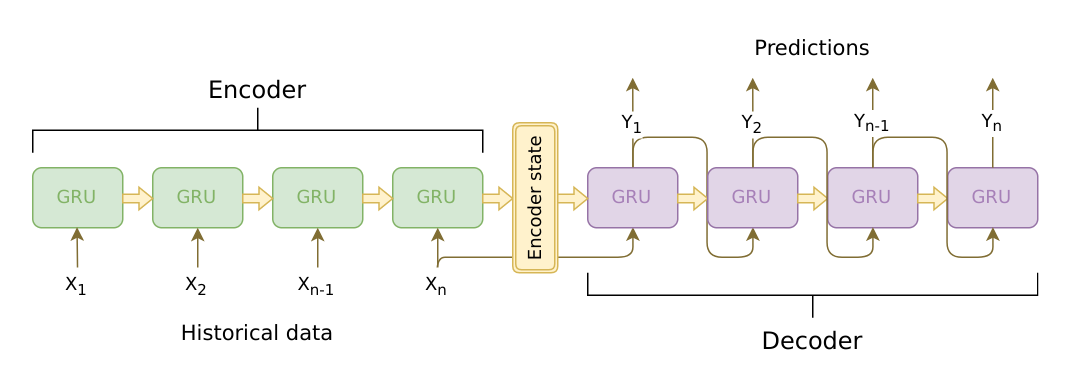
有鑑於此，於[2][3]兩論文中提出模型，這一類模型更加接近我們閱讀理解常見的distractor，目標是生成更長，語義更加豐富的distractor。[2][3]想法上為改用文本生成的方式去產生distractor。而[2] [3] 模型在文本生成上，是採用seq2seq+attention機制。seq2seq架構主要解決的問題是輸入和輸出長度不一致的問題。基本上文本生成不可能輸入和輸出長度可以一一對應，因此seq2seq的想法便將輸入文本壓縮到一個向量中，文本的輸出便是對這個向量的解碼，如圖4，decoder接收的輸入便是最後一個encoder的state。

圖5: seq2seq模型示意圖[11]

由於decode的時候，之後參考到最後一個encoder的結果，decoder的文本越長會導致結果越差。因此為了改善這個問題，[2]採用attention機制在decode時也要留意輸入文本的資料。

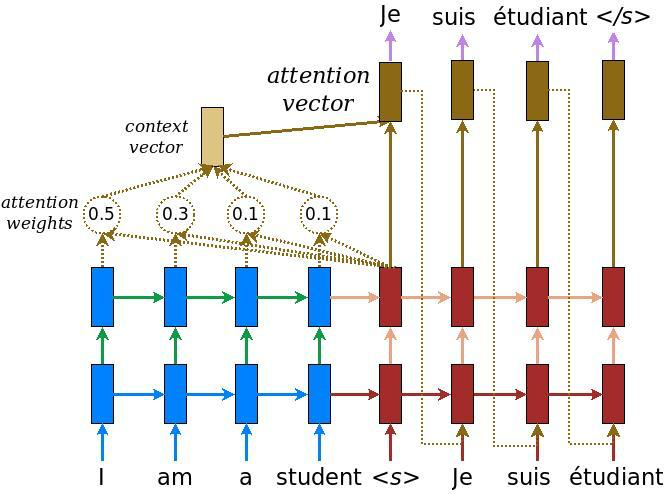
圖6: seq2seq+attention的示意圖[12]

圖4中，在decode第一個字的時候，除了以hidden state作為輸入，還會多加入attention vector，這個attention vector會根據輸入文本與輸出的隱藏層算權重加總輸入層的的訊息而來。雖然這一套方法已經很大程度緩解不能處理長文本decode的問題，是相對成熟的架構，但是發展上已經遇到瓶頸，難以有再多提高。[2] [3] 的研究主要是在attention機制上改善生成的結果，但都提升有限。因此我們研究的第一個方向，便是更換文本生成的架構。

另一方面，我們關注干擾項的品質以外，還會希望能生成一定的量。[2] [3] 論文在生成多個干擾項上，均是採用beamsearch的方式。beamsearch是文本生成的一種解碼策略，會考慮好幾個高機率的預測結果，最終生成多於一個文本出來。雖然可以生成多個結果，但由於這些結果都是從同一個機率分佈出來，語義上會比較接近。因此，也需要探索另一種方式去生成可以將不同事情的干擾項。

最後，[2] [3] 論文在證干擾項的有效性上，說服力不夠。[2][3]會讓生成出來的干擾項跟資料集的干擾項之間算相似度 以及 用人工測試的方式去評估干擾項的品質。干擾項跟資料集相似度底，並不代表干擾項效果不好，人工測試的樣本不足，也有難以量化疑慮 。

總的來說，我們可以往這幾個方向改善：第一：更強大的模型：目前BERT有大量的預訓練資料和更強大的模型，有必要嘗試更新成基於BERT的模型，來取得更大的突破；第二：生成多個distractor；第三提出其他指標來證明distractor的有效性。

(四)研究方法及步驟

1. 使用BERT來建立生成模型

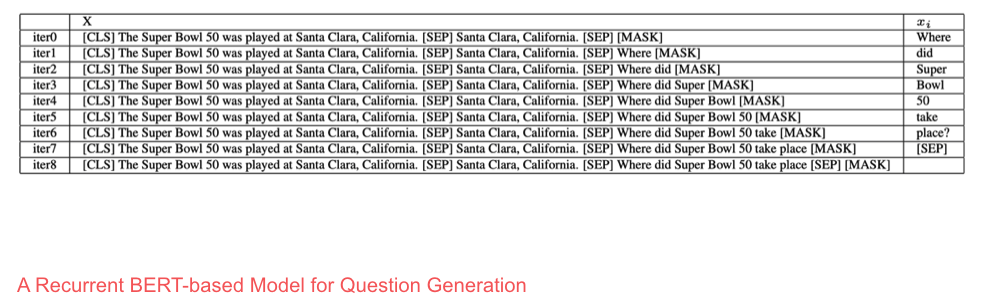
BERT在生成任務上，在預訓練masklm上修改，可以讓BERT擁有文本生成的能力。masklm預測mask字時候，會根據左右文本的訊息，這就像是一個雙向的語言模型，可以根據左右的字推測中間mask起來的內容。而我們在文本生成上則是單向的，因此我們可以把BERT改成單向語言模型的形式來做到文本生成。

圖7: BERT單向語言模型的文本生成過程

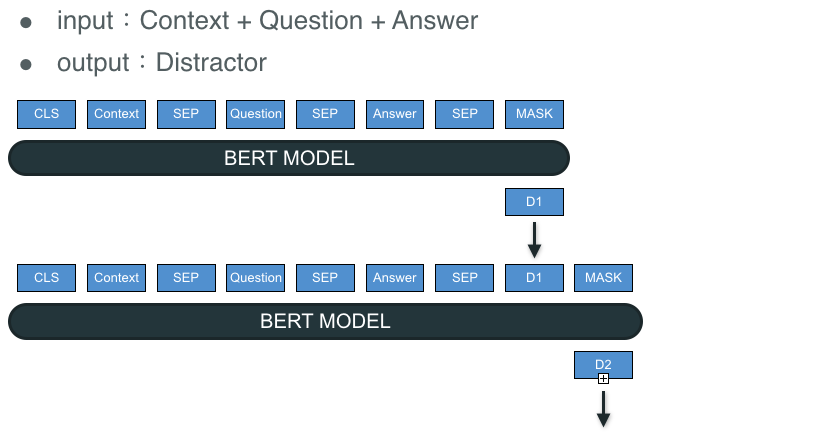


圖8: BERT單向語言模型的文本生成架構

BERT模型會有一些保留字，如CLS用以告訴BERT現在是文本的開始，SEP是作為文本的分隔。在BERT的生成模型中，會以CLS作為整個輸入的開頭，然後放入短文，問題和答案，其中SEP會作為兩兩之間的分隔，然後放一個MASK，讓模型預測這個mask的結果，也就是生成的第一個字。預測mask之後，會將結果放回輸入裡面，再添加一個mask，預測第二個字，以此類推。但是直接套用上述架構，效果未必會好。這是由於這個生成架構並沒有約束生成的方向，使得結果有可能會是答案的換句話說,也可能跟內文無關更甚至會是重複的文本。因此在往後的研究方向，便是在生成方向上給予更多的訊息。

1. 生成多個干擾項

我們希望可以生成多種多樣的干擾項，因此會探索不同的做法，找出最佳解。之前常用的方法有beamsearch，透過取預測機率最高的幾個結果來得到多個結果。

|  |
| --- |
| Question : What can we learn about ' Star Trek classroom ' from the text ? |
| Answer : It can be helpful in both learning and teaching . |
| Top 3 result in beam search: |
| It is the only way to solve problems . |
| It can help students to solve problems . |
| It is the only way to solve problems in class. |

圖9: beamsearch生成例子

但圖9結果可見，三個結果的語義和句子結構都很相似，如果作為干擾項的話，答題者會很輕鬆排除掉相像的答案，讓其效果大打折扣。因此需要進一步探索其他可能的方向，例如修改模型的輸入結構，作為另一個嘗試的方向。

圖10: recurrent輸入的DG架構

在之前BERT模型的基礎上，我們也可以做多個干擾項的生成。如同之前模型的過程，生成干擾項1。生成的干擾項1會放到輸入裡面，重新按照之前的架構來生成干擾項2。如此方法，可以一直生成到我們想要的數目。

1. 提出評估distractor的方法

首先我們要證明我們生成的干擾項互相之間有一定不相似度，因此我們可以採用文本生成的指標 - BLEU[7] 和ROUGE[6] 來得到干擾項之間的相似度，透過比對資料集的干擾項和之前SOTA的相似度，邊可以比較出我們模型的優劣。

其二是要證明干擾項是有效的，也就是可以迷惑答題者答案的選擇。因此我們可以訓練一個回答多選題的模型，將預測資料替換成我們生成的干擾項，來看看模型的準確度會上升還是下降，如果準確度下降則代表我們生成的資料成功讓答題者迷惑，反之則說明我們生成的資料效果不好。

(五)預期結果

(a) 收集資料及整理文獻;

(b) 以BERT建製文本生成的模型;

(c) BERT套入DG模型;

(d) 提出優化BERT架構的方法;

(e) 探討如何生成多個distractor

(f) 探討評估distractor的方法

(h) 整理結果，將其寫成論文發表

(六)參考文獻

[1] Zesch, Torsten, and Oren Melamud. "Automatic generation of challenging distractors using context-sensitive inference rules." Proceedings of the Ninth Workshop on Innovative   
Use of NLP for Building Educational Applications. 2014.

[2] Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R Lyu. 2019. Generating distractors for reading comprehension questions from real exami-

nations. In Proceedings of the AAAI Conference on Artificial Intelligence , volume 33, pages 6423–6430.

[3] Zhou, Xiaorui, Senlin Luo, and Yunfang Wu. "Co-Attention Hierarchical Network: Generating Coherent Long Distractors for Reading Comprehension." arXiv preprint arXiv:1911.08648 (2019).

[4] Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 154–162.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understand-

ing.arXiv preprint arXiv:1810.04805.

[6] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for compu tational linguistics, pages 311–318. Association for

Computational Linguistics.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language under standing by generative pre-training.URL <https://s3-us-west-2>.

amazonaws.com/openai-assets/research-covers/languageunsupervised/language understand-ing paper. pdf.

[9] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating nat-

ural language generation. CoRR, abs/1706.09799.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In Advances in neural information processing systems, pages 5998–6008.

[11] Joseph Eddy Forecasting with Neural Networks - An Introduction to Sequence- to-Sequence Modeling Of Time Series <https://jeddy92.github.io/> JEddy92.github.io/ts\_seq2seq\_intro/

[12] Luong M T , Pham H , Manning C D . Effective Approaches to Attention-based Neural Machine Translation. Computer Science, 2015.

(七)需要指導教授指導內容

申請人鍾浩霖目前就讀中興大學資工三年級。在基本程式撰寫的能力部分，申請人已經具備獨立執行此大專學生研究計畫的能力。

針對本計畫需要進一步跟老師學習的內容初步列於下:

(1)BERT模型的架構;

(2)學習文本生成的各種架構;

(3)想出如何生成多個不一樣的distractor;

(4)如何探討各個方法的成效;

(5)科技論文撰寫與投稿;

(6)口語表達能力提升。

表C802