二、研究計畫內容

(一) 摘要

本研究欲透過「自動問句生成閱讀理解選項」為應用目標，欲透過提

升問句生成過程中關鍵字篩選之難度及複雜度，來增進自動問句生成閱讀理解模型對未來教育環境的幫助。

(二) 研究動機與研究問題

近年來，因為自然語言處理領域的技術較成熟，使得自動問題生成成為開發未來學習系統理想的工具。本計畫希望利用自然語言處理技術，改善如今自動問句生成系統的問題，讓自動問句生成系統能更完善，讓自動問句生成技術能更貼近真實的人工出題品質。如今，自動問句生成之閱讀理解模型已廣泛應於多種教育用途，像是學術寫作支持(academic writing support)[1]、閱讀理解評估(reading comprehension assessment)[2]、教育聊天機器人(educational chatbots)[3]。如今的自動問句生成系統中，大多是透過先挑選文本中的關鍵字做為問句的答案，再進行出題的步驟。也因為作為答案的關鍵字挑選方式較為簡單，造成在問句生成時，無法真正透過這些問題來評估學生在學習上的效果，這也是現在自動問題生成系統還無法取代人工出題的其中一項原因。

因此，關鍵字的選擇是一項讓自動問句生成擴展到具有預先指定答案的數據集之外，並且能應用於現實生活中的重要步驟。

現今，市面上較為廣為人知且強大的ner (Name Entity Recogonition) API 有NLTK、spaCy，但他們就僅能針對文章中的人名、地點、時間…等，進行標記，但我們知道如果要完成一個完整的自動出題系統，並能確保學生真的完整理解一篇文章的內容，我們不應該將考題侷限於人名、地點…等較短專有名詞上。因此，本計畫想開發一套能篩選出閱讀文本中一段文字做為答案的系統，如此不僅能增進自動問句生成系統的問題難度，更能深入了解學生是否真的完全了解此篇文章中的內容。

以圖一為例，Text 1 的內容為美國布希總統在選舉時，請到了美國著名的民調公司Gallup 來為自己做民調，而民調的時間是從1992年11月至1993年1月之間製作的，但我們使用spaCy這款套件來做ner，我們發現雖然spaCy為我們分別挑出了「November 1992」和「January 1993」兩個關鍵字，也就是spaCy認為這兩個關鍵字為獨立的個體，但以語意上來說，「November 1992 through January 1993」為一區間，如果將兩個時間點分割成兩個獨立的個體，對於讓學生了解文章的內容，沒有幫助，因此本計畫預計透過BERT模型著手解決市面上這些套件所造成的問題。

圖一 :

Text 1 : George H.W. Bush averaged an 84% approval rating among Republicans in an average of November 1992 through January 1993 Gallup

polls.

Result : November 1992 (DATE)

January 1993 (DATE)

(三) 文獻回顧與探討

我們會在此部分針對計畫中會使用到的 SQuAD 資料集、BERT 模型，進行介紹並談論使用它們的原因。

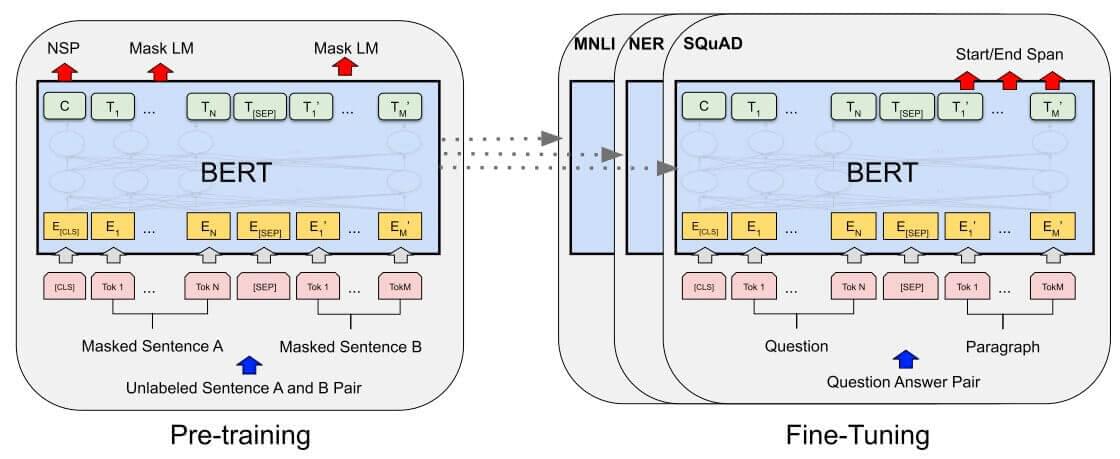
1. SQuAD(The Stanford Question Answering Dataset) : SQuAD資料

集從Wikipedia中選擇前一萬篇熱門的文章，並隨機選擇了500多篇

文章，並抽取出兩萬多個段落並且覆蓋了各式各類的主題。團隊雇用了Mturk[[1]](#footnote-1) Worker ，要求Mturk Worker 讀一個段落並針對這個斷落的內容提最多5個問題，同時在段落中標註出對應答案的位置，並針對別人提出的問題，再原文中用盡量少的字長標註出答案所在。

SQuAD資料集除了資料量大之外，與之前同類型的資料集個重大區別及回答不是選擇題，而是問答題。而且SQuAD的答案是在單一個段落裡，與其他資料集會跨文件有很大的不同。

2. BERT模型(Bidirectional Encoder Representations from Transformers) : BERT 是一種可微調的語言模型(Language Model,LM)，語言模型可以在給定一些詞彙的前提下，估計下一個詞彙出現的機率，而傳統語言模型會針對不同 NLP 任務去設計一個最適合的神經網路架構，但設計集測試的過程中會耗費許多人力、時間、計算資源。因此就有BERT 模型架構的誕生，BERT 模型就是一個是先訓練好，而且可以套用到多個NLP 任務的模型，而我們再以此架 構去微調，達成我們想完成的任務。



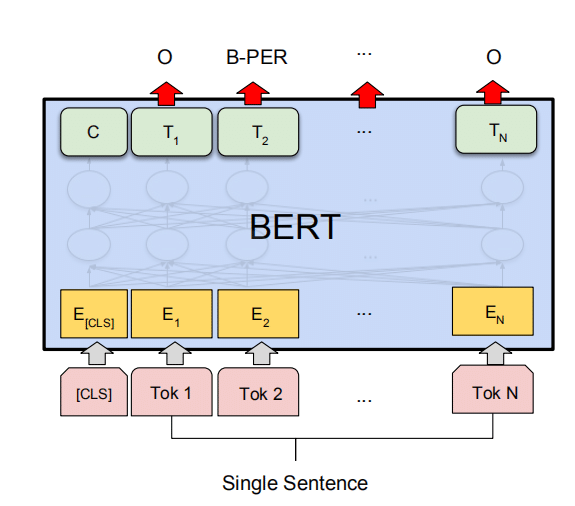
圖二 : Overall pre-training and fine-tuning procedures for BERT.

(四) 研究方法與步驟

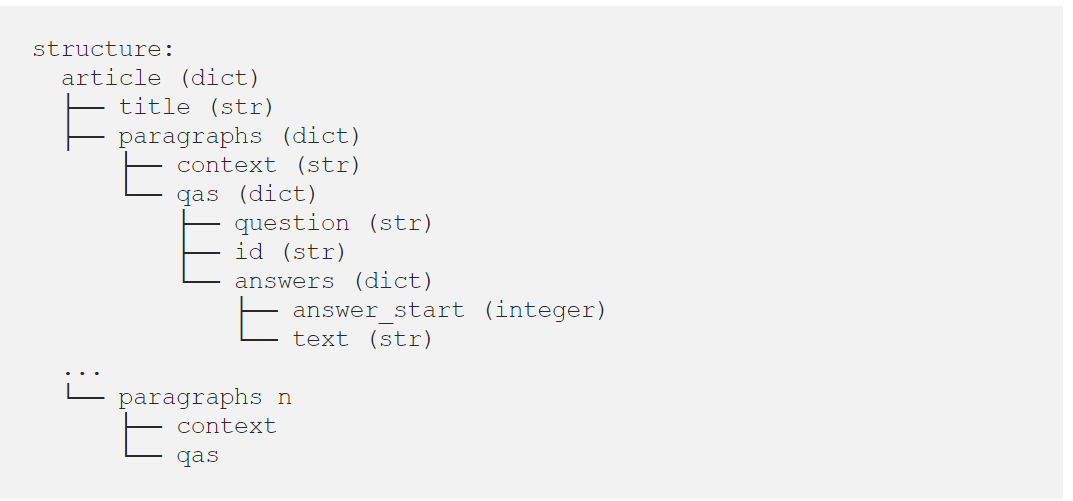
本計畫預計使用SQuAD 資料集作為訓練資料，並使用BERT 預訓練模型進行微調，來提升自動問句生成系統中的關鍵字選擇。

1. 在BERT 之上加入新的線性分類器 :

我們在(三)文獻回顧與探討中所提及的BERT 是一個可以微調的語言模型，也就是我們可以在預訓練完的BERT之上加入新的線性分類器(Linear Classifier)，並利用下游任務的目標函式從頭訓練分類器並微調BERT的參數。而我們預計使用HuggingFace[7]團隊以訓練的PyTorch Fine-tuning BERT為基底，並搭配transformer 模型庫中的BertForTokenClassification 套件。BertForTokenClassification 是一個包含了BERT 模型並加上字詞標註的分類器的可微調模型。



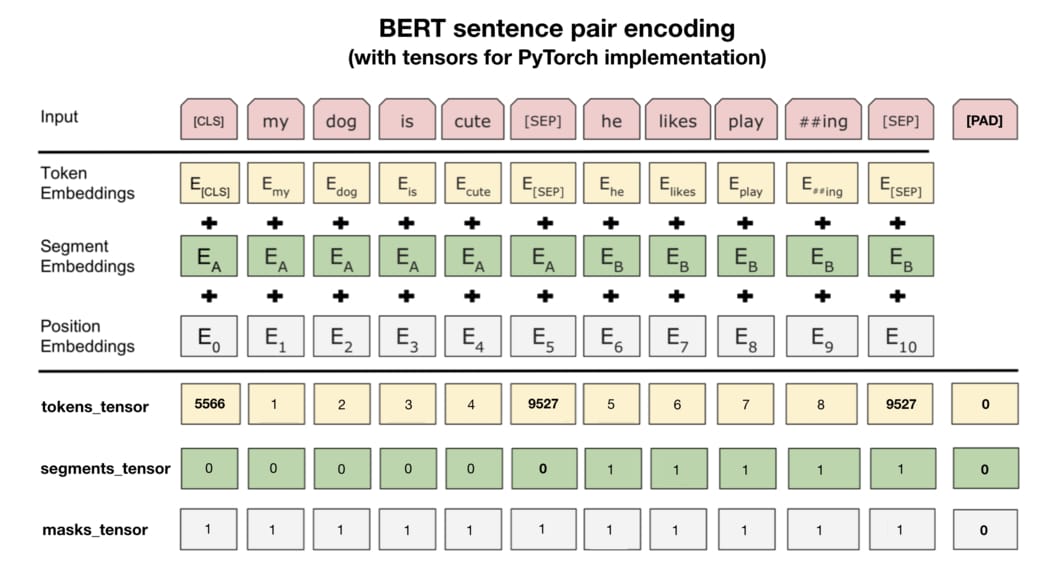
圖三: BertForTokenClassification

1. 將 SQuAD 資料集轉換成BERT相容的格式 :

圖三 : Json 格式的SQuAD資料集結構圖

首先，因為SQuAD資料集的格式複雜，但我們僅需要SQuAD資料集中的文章內容(context)及答案(answers-text)來做為訓練資料，因此我們必須從繁雜的資料集中，取出我們所需。並將句子跟答案做斷詞(tokenize)，並將取出的答案標記answer，表示此段文字值得作為出題的方向。

由圖四我們看到，input就是我們將輸入的文章內容(context)進行斷詞之後，再加上[CLS]開頭，及[SEP]結尾，為了讓 GPU 平行運算我們需要將 batch 裡的每個input都補上 zero padding 以保證它們長度一致。而最重要的即是把原始文本轉換成PyTorch中的3種id tensors，token\_tensor是由tokenizer轉換成的索引值，segment\_tensor就是用來識別句子的界限，而mask\_tensor則是用來界定attention mask的範圍(1表示讓BERT關注該位置，0則表示Padding)。



圖四 : BERT sentence pair encoding

1. 將SQuAD資料集的文本套用至spaCy :

在(二)研究動機與研究問題中，雖然spaCy無法辨別較長的片段作為答題的關鍵字，但是卻在普通的人名、時間、地點都能正確抓出關鍵字，並且很有效率。因此，我們希望結合SQuAD資料集中答案較長的優點，並利用舊有的SQuAD的優點，來提升自動問句生成系統中，關鍵字的抓取。

(五) 預期結果

1. 設計出適合的Linear Classifier:

希望在了解BERT模型架構之後，能自行寫出我們想要用來分類這個token(字詞)是否適合作為答案的分類器。

2. 希望能讓自動問句生成系統更貼近人供出題的品質 :

欲透過整合SQuAD資料集、spaCy API的優點，使的挑選關鍵字的時候能更貼近老師的出題的方向。

1. 將此技術應用於開發英語學習系統 :

成功實現此技術後，希望能結合時下流行的通訊軟體LINE，我們預計在每天爬取CNN、Washington Post等國外新聞網站，並將文章爬取下來後，將本計畫的關鍵技術用於關鍵字的選擇，並在選擇完關鍵字之後，與自動問句出題API結合，完成閱讀測驗，並存在後端資料庫中。只要使用者閱讀完文章，就能回答跟這篇文章相關的閱讀測驗題，讓本專案的技術真正的應用於教育用途，也讓自然語言技術造福更多群眾。

(六) 參考文獻

[1] Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. G-Asks: An intelligent automatic question generation system for academic writing support.

[2] Jack Mostow and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics, 136–146.

[3] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating Natural Questions About an Image. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 1802–1813.

DOI: <http://dx.doi.org/10.18653/v1/P16-1170>

[4] [Pranav Rajpurkar](https://arxiv.org/search/cs?searchtype=author&query=Rajpurkar%2C+P), [Robin Jia](https://arxiv.org/search/cs?searchtype=author&query=Jia%2C+R), [Percy Liang](https://arxiv.org/search/cs?searchtype=author&query=Liang%2C+P). ACL 2018.Know What You Don't Know: Unanswerable Questions for SQuAD

[5] [Pranav Rajpurkar](https://arxiv.org/search/cs?searchtype=author&query=Rajpurkar%2C+P), [Jian Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang%2C+J), [Konstantin Lopyrev](https://arxiv.org/search/cs?searchtype=author&query=Lopyrev%2C+K), [Percy Liang](https://arxiv.org/search/cs?searchtype=author&query=Liang%2C+P). 2016.SQuAD: 100,000+ Questions for Machine Comprehension of Text

[6] [Jacob Devlin](https://arxiv.org/search/cs?searchtype=author&query=Devlin%2C+J), [Ming-Wei Chang](https://arxiv.org/search/cs?searchtype=author&query=Chang%2C+M), [Kenton Lee](https://arxiv.org/search/cs?searchtype=author&query=Lee%2C+K), [Kristina Toutanova](https://arxiv.org/search/cs?searchtype=author&query=Toutanova%2C+K). 2019.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[7] <https://github.com/huggingface/transformers>

(七) 需要指導教授指導內容

1. Amazon Mechanical Turk (MTurk) is a marketplace for completion of virtual tasks that requires human intelligence. The Mechanical Turk service gives businesses access to a diverse, on-demand, scalable workforce and gives Workers a selection of thousands of tasks to complete whenever it's convenient. [↑](#footnote-ref-1)