The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 47-61

基於 BERT 模型之多國語言機器閱讀理解研究

Multilingual Machine Reading Comprehension based on BERT Model

吳承軒 Cheng-Xuan Wu
國立臺北科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taipei University of Technology
t106598068@ntut.edu.tw

王正豪 Jenq-Haur Wang
國立臺北科技大學資訊工程系
Department of Computer Science and Information Engineering
National Taipei University of Technology
jhwang@csie.ntut.edu.tw

摘要

在網路資訊爆炸的現代，人們的生活與網路已密不可分，但受限於檢索技術的瓶頸，雖然能提供多方面的資訊來源，卻不一定是最相關有幫助的資訊。自然語言中的兩個主題：機器問答(Question Answering)與機器理解(Machine Comprehension)，由於對檢索系統，以及服務轉型中重要的聊天機器人，都具有高度相關，因此成為近年熱門的研究議題。本論文使用了 Google BERT 的 pre-trained model 進行詞嵌入向量，以單詞及單字為單位，組織出一個句子的特徵。並且基於問題、答案、與文本間不同組合的答題策略，最終選擇最高餘絃相似度的選項，作為機器作答的依據。本論文分別實驗在英文 TOEFL-QA 資料集，以及中文開放性問答資料集，對比於雙向 GRU 以及 A Strong Alignment IR Baseline 的方法，分別取得 34.87%及 57.5%準確率，實驗結果顯示，雖然不同語言之間具有文法的差異，但本論文所提的方法具有一定程度多國語言的通用性。

關鍵詞：閱讀理解、機器問答、自然語言處理、深度學習

Abstract

In recent years, Internet provides more and more information for people in daily life. Due to the limitation of information retrieval techniques, information retrieved might not be related and helpful for users. Two research topics in natural language processing have attracted much attention due to the important applications of information retrieval and chatbot in the past few years: question answering and machine comprehension. In this paper, we use Google BERT pre-trained model as a word embedding model to form semantic sentence features based on single words and phrases. Based on different strategies for question answering, we use cosine similarity to calculate similarity and choose the option of highest cosine similarity score as machine inferenced answer. In our experiments on TOEFL-QA dataset for English and Formosa Grand Challenge dataset for Chinese, our proposed method was compared with Bi-directional GRU and a strong alignment IR baseline, and obtained an accuracy of 34.87% and 57.5%, respectively. With the grammar difference between difference language, our model is capable of processing multilingual questions with comparable performance to existing methods.

Keywords: Reading Comprehension, Question Answering, Natural Language Processing, Deep Learning

一、緒論

人們每天的生活，與許多的電子產品緊密相連，透過這些便利的工具，有助於改善生活品質。在過去想要探究一個知識，必須查閱大量的書籍，並耗費許多的時間消化了解，才能從中找出想要的資訊；網路的普及與檢索技術的發展，使得人們能透過網路

搜尋服務，快速的從海量的資訊中，得到初步篩選過的結果，節省了閱讀與理解知識的時間。隨著資訊服務不斷的改善人們的生活，不論是社群平台，亦或是購物網站等，皆開始導入聊天機器人，相較於傳統上透過人力行銷及人力客服，除了減少部分人力的成本，更能夠增加與科技之間的互動性，提升黏著度因而增加提升商業產值；科技與人們的生活已經無法分離，進而促使了科技助理的誕生，為了能夠更進一步，滿足人們更多日常生活中的需求，需要讓這些不同類型的機器人，能夠更加了解人們真正所需，理解人們心中所想。傳統機器人與檢索技術的運作方式，大多採用關鍵字為主的方法，若查詢的語句中，與預設設置好的搜索資料相符，則將此資料當成最終的結果；在這種方式下，缺乏考慮所有其他非關鍵字詞的語句，因此可能會遺失真正重要的字詞資訊。為了要能夠更加有效的利用，一個查詢與答案的所有資訊，本論文使用字與詞來組成一句話所代表的真正涵義，並透過計算查詢中所有的句子，與答案句子之間關聯的強弱，來理解不同語句之間真正相似的程度。本論文對比了傳統基於關鍵字的問答方法，以及基於深度學習的方法，不同於大多數問答的研究，通常只專注在某個語言，本論文提出二個架構簡單，且具有一定效果的多國語言答題模型。

二、相關研究

（一）、語言模型

在自然語言處理的研究中，文字處理是一個重要的步驟，要讓機器進行下一步的計算，需要將現實世界中詞語的意義，轉換為代表它們的向量數值，因此這個階段代表了特徵的好壞。現實中使用文字的情境，常常某些詞語會與其他詞一同出現，或者出現在一篇文章的上下文，2013 年 Mikolov 等人[1]基於此特性，設計了一個詞嵌入向量模型，用來表示詞與詞之間的關係。進一步提升語言模型的表現，基於 Vaswani 等人[2]的架構，Devlin 等人[3]建構了一個 12/24 層的 BERT 語言表示模型，BERT 模型在英文問答競賽 SQuAD1.1 中贏過了 BiDAF+ELMO[14][20]、QANet[15]等神經網路的方法，驗證了一個良好的語言表示模型，能夠達到與複雜的神經網路架構相當的水平。

（二）、文字相似性

為了要讓機器更進一步具有理解語意的特性，一個初步需要被探討的問題為：如何找出文字與文字之間，是否具有相似語意，進而使機器能夠以此為基準。一種常見的作法為餘絃相似度(Cosine Similarity)，如 Gomaa 等人[4]的研究中所提到，透過餘絃相似度，使機器判斷兩種具有類似特徵的文章是否相似，更進一步的拓展應用，如 Huang 等人[5]、Karypis 等人[6]的研究中，將餘絃相似度使用在分群不同文章。

（三）、機器理解之選擇題研究

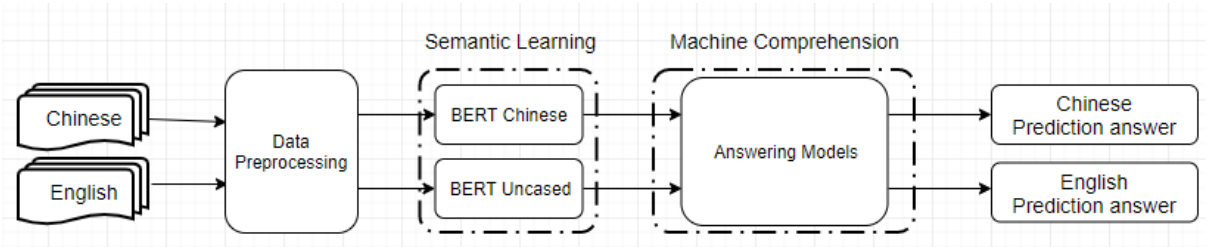
概觀為，給予一段故事、問題、幾個不同的選項，需要透過這些的線索，使機器找到答案，而其中的研究類型大致上可分為兩類，第一類為單選題問答，如中文科技大擂台競賽單選題問答[18]的研究，英文 Tseng 等人[7]的研究，以及第二類，複選題問答如 Richardson 等人[9]的研究。透過雙向 GRU 模型快速的學習故事與問題之間上下文的語意關係，但只在最後階段將雙向 GRU 模型的結果與選項做相似度計算，會喪失將選項與故事及問題一同考慮語意關係的面向，使用 Sliding Window 及 Distance Based 的方法，考慮故事中的詞彙，出現在問題及選項詞彙聯集的數量，但缺乏考慮沒出現詞彙的語意線索。

（四）、機器理解之尋找答案段落研究

概觀為，給予一段故事以及問題，根據問題找出故事中正確答案的段落，如 Yadav 等人[10]基於檢索的角度計算答案與問題的相似度，但依賴於 Embedding 模型的好壞，以及英文 Rajpurkar 等人[11]使用維基百科的文章進行問答研究，和 Tapaswi 等人[12]專注在電影內容交談的問答研究，另外還有 Hermann 等人[13]以新聞文章內容進行問答研究，透過 Word Distance 來計算故事與問題的字詞，但容易有詞彙被忽略，以及 Seo 等人[14]使用 LSTM 搭配 Attention，用來預測答案出現在故事的位置，Yu 等人[15]透過 CNN 以及 Self-Attention，找出答案在故事中最有機會出現的位置，以及中文 Shao 等人[16]使用 BERT 模型，在中文開放性領域問題上的研究，贏過 QANet[15]、R-Net[19]、BiDAF[14]等模型。

三、研究方法

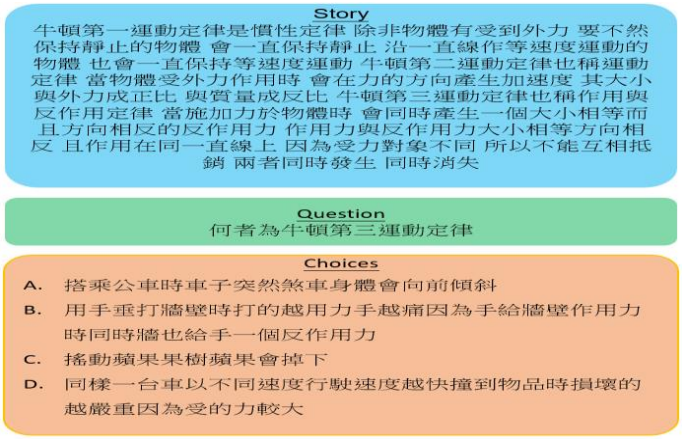
本此章節將說明本論文的研究方法與系統架構，將架構分為四大部分，分別對中英文資料進行前處理，並透過 BERT 將句子轉換為機器能使用的語意資訊，接著進行機器理解答題模型的推論，最終選出預測的答案。以下小節將會針對系統中各模組的細節做進一步說明。



圖一、系統架構圖

（一）、資料前處理

為了要提升資料在模型上的精準度，此階段除掉非文字的標點符號如圖二所示，Story 的部分以逗號及句號為單位來切割內文，而 Question 及 Choices 則去除所有標點符號，當成一個句子來處理。

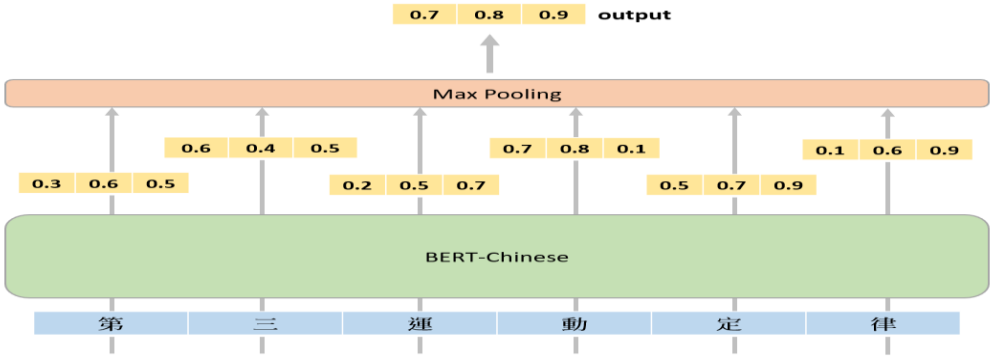


圖二、閱讀測驗題目

（二）、語意學習

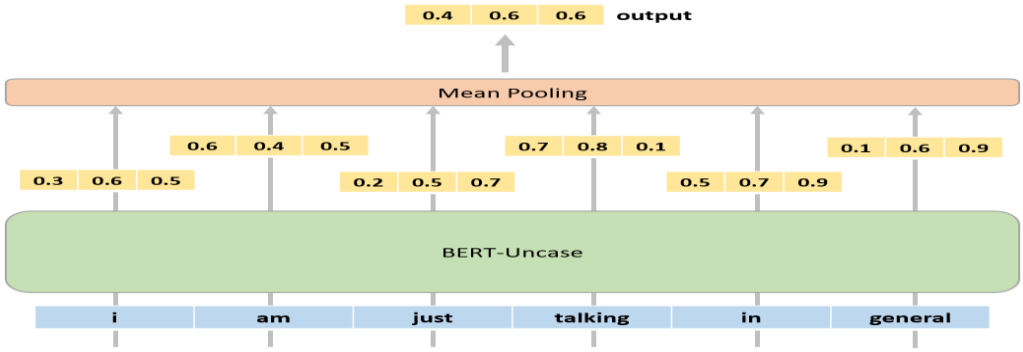
大量詞彙分散在許多上下文中，藉以組織不同文章表達不同主軸的意義，為了要使機器能夠得到這種語意資訊，本論文利用 BERT model 進行詞嵌入向量，對 BERT-Base, Chinese 12-layer 及 BERT-Large, Uncased 24-layer 實驗不同 pooling strategy。Max pooling 為，將每個時間軸上的隱藏層(hidden layer)資訊取最大值，來組織出

該句子所蘊含的語意資訊如圖三。



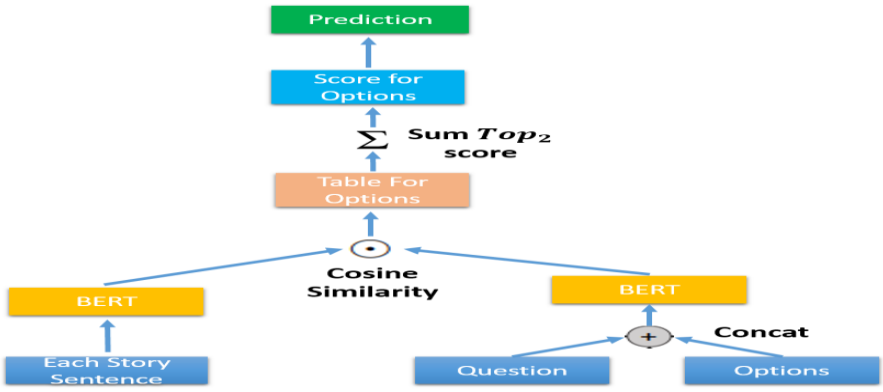
圖三、Max pooling strategy

而 Mean pooling 則是取隱藏層資訊的加總平均值，來組織出該句子所蘊含的語意資訊如圖四，藉此將字詞之間較重要的語意資訊保留，提高推論模型的精準度。



圖四、Mean pooling strategy

（三）、機器理解



圖五、策略一模型圖

本論文基於回答閱讀測驗時，採取不同的答題策略，設計了二種類型的答題模型。

第一種答題策略如圖五，透過將 Story 以句子作切割，能更詳細的思考 Story 中的所

有資訊，將每個句子視為可能保有關鍵資訊來處理如公式(1)，接著思考 Question 以及 Choice 組成的線索如公式(2)，計算兩者之間的相似性如公式(3)，接著計算最相關於選項的故事句子分數，專注於最相關於 Choice 的句子如公式(4)，最終選擇關聯性最高的選項如公式(5)，本模型的公式如下所示：

$$S_n = \boldsymbol{BERT}(Story_n) \quad (1)$$

$$QC_m = \boldsymbol{BERT}(Question + Choice_m) \quad (2)$$

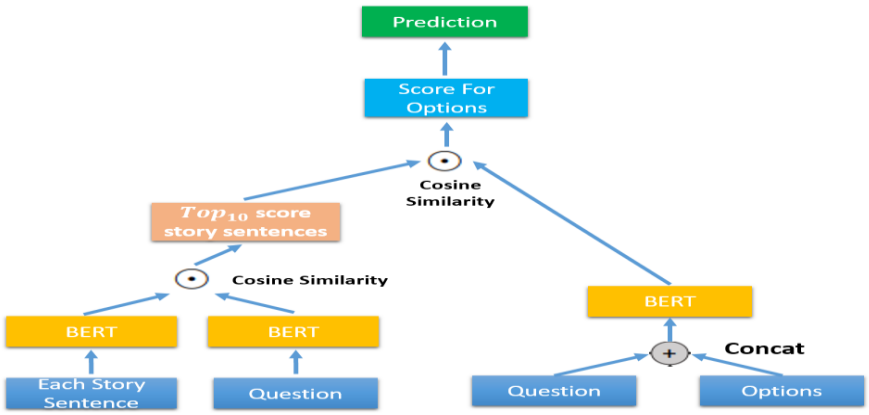
$$Table_m = \{\boldsymbol{Sim}(S_1, QC_m), \ldots, \boldsymbol{Sim}(S_n, QC_m)\} \quad (3)$$

$$Score_m = \sum Top_k(Table_m) \quad (4)$$

$$Prediction\ \ Answer = \boldsymbol{argmax}_m(Score_m) \quad (5)$$

在上述公式中，各個參數及設置的定義如下：

1. n: Story 中句子的數量，m: Choice 的數量，k: 2。
2. BERT: 將句子作為 BERT 的輸入，取得轉換的詞嵌入向量，組織成句子資訊。
3. Sim: 計算 Cosine Similarity 餘絃相似度分數。
4. Table: 一個 Choice 與 Question 合併，對應 Story 所有句子計算的分數表。
5. Score: 將每個 Choice 的分數表，最高的兩個分數相加，為該選項的分數。
6. Prediction Answer: 將四個 Choice 裡最終分數最高的選項作為預測答案。



圖六、策略二模型圖

第二種答題策略如圖六，將每個句子都視為保有關鍵資訊來處理如公式(6)，接著思考 Question 以及 Choice 組成的線索如公式(7)，並且獨立出 Question 的資訊，以免於被選項的資訊所影響如公式(8)，透過這樣的方式思考 Story 中幾個最相關於

Question 的句子，能夠排除不相關的故事內容如公式(9)，接著計算最相關於每個 Choice 的 Story 句子如公式(10)，最終選擇關聯性最高的選項如公式(11)，本模型的公式如下所示：

$$S_n = \boldsymbol{BERT}(Story_n) \quad (6)$$

$$QC_m = \boldsymbol{BERT}(Question + Choice_m) \quad (7)$$

$$Que = \boldsymbol{BERT}(Question) \quad (8)$$

$$Top_k Sentence = Top_k\{\boldsymbol{Sim}(S_1, Que), \ldots, \boldsymbol{Sim}(S_n, Que)\} \quad (9)$$

$$Score_m = \boldsymbol{max}(\boldsymbol{Sim}(\mathrm{Top}_k S, Choice_m) \quad (10)$$

$$Prediction\ \ Answer = \boldsymbol{argmax}_m(Score_m) \quad (11)$$

在上述公式中，各個參數及設置的定義如下：

1. n: Story 中句子的數量，m: Choice 的數量，k: 10。
2. BERT: 將句子作為 BERT 的輸入，取得轉換的詞嵌入向量，組織成句子資訊。
3. Sim: 計算 Cosine Similarity 餘絃相似度分數。
4. $Top_k Sentence$: 取 Story 句子與 Question 計算餘絃相似度的 Top10 句子。
5. Score: 每個選項與 Top10 句子做相似計算，最高相似度為該 Choice 分數。
6. Prediction Answer: 將四個 Choice 裡最終分數最高的選項作為預測答案。

四、實驗與討論

本章節將針對本論文所提出來的研究方法，探討使用 BERT 語言模型作為語意轉換，之後透過答題模型進行機器理解的問答，探討其效果。

（一）、實驗資料集

本論文使用中英文兩種閱讀測驗資料集，英文為 Listening Comprehension Test of TOEFL[7]資料集，合併訓練及測試資料集總筆數為 839 筆。中文為科技大擂台_測試資料集[18]前六次初賽資料，總筆數為 8550 筆，並且中英文資料集皆為單選題。

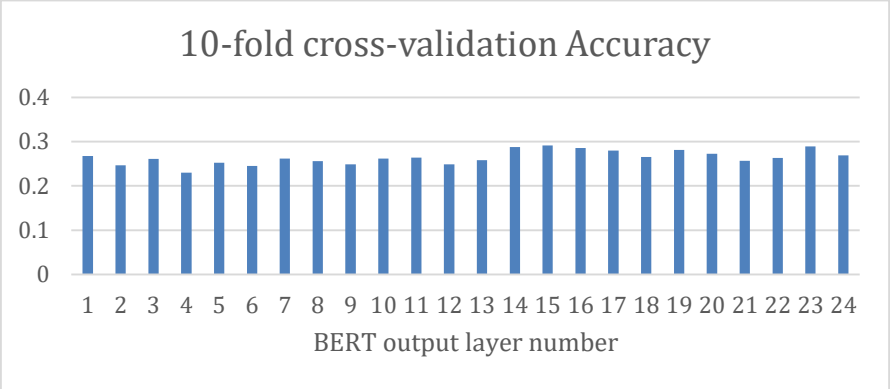
（二）、驗證方法

在中文及英文的資料集之中，本論文採用 10-fold Cross Validation[17]作為進一步評估誤差的指標。

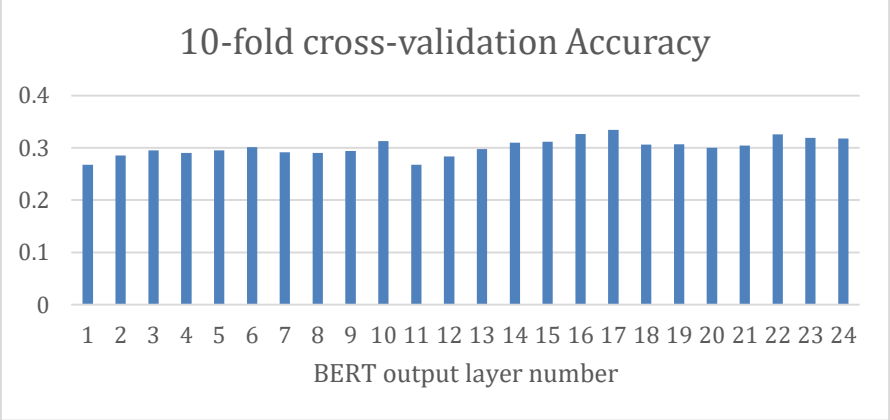
（三）、語言模型詞嵌入比較

我們將比較二種答題模型，中文使用 BERT Chinese，英文使用 BERT Uncased，比較不同 layer 的 output 作為詞嵌入向量，及不同的 pooling strategy 來組織句子後的效果。

1、英文資料集

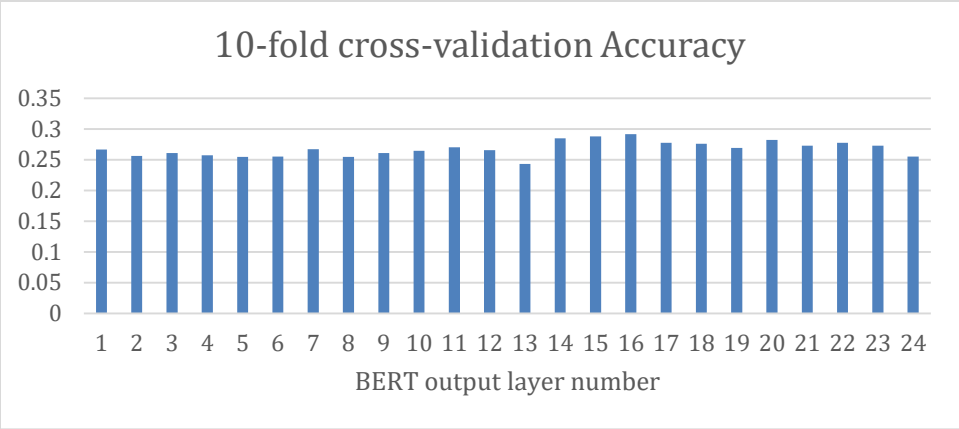


圖八、英文資料集 策略一及 max pooling

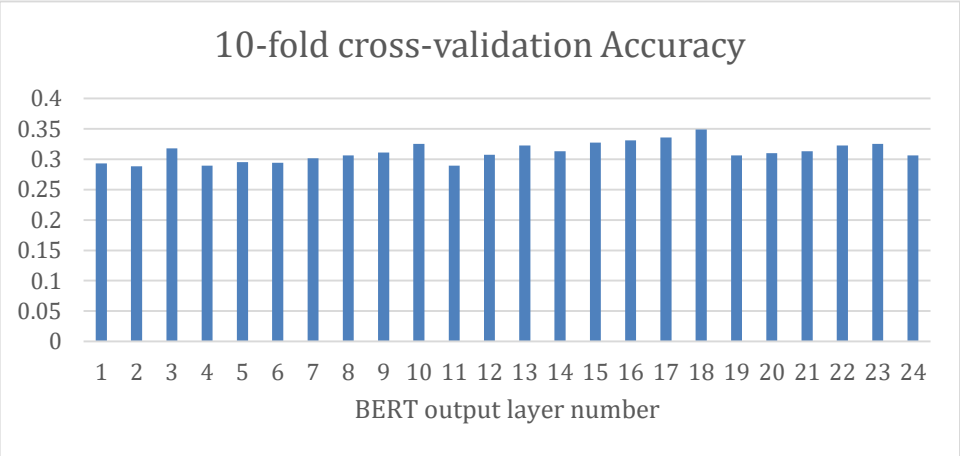


圖九、英文資料集 策略一及 mean pooling

從圖八、圖九中觀察到，英文資料集使用策略一及 mean pooling 時，以 17-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.3345 的準確率。



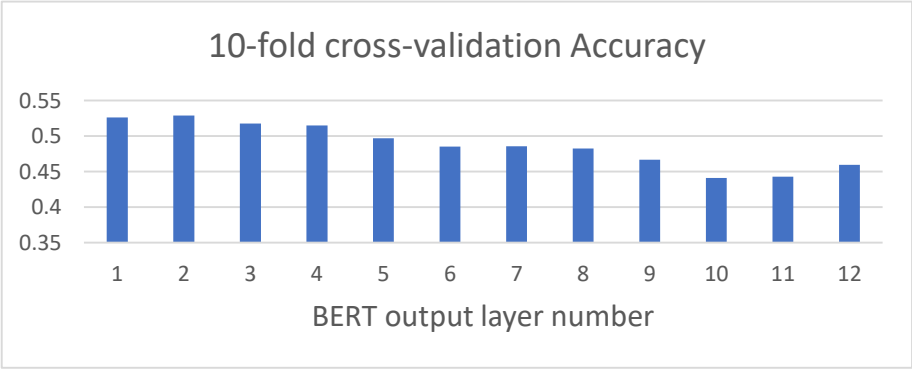
圖十、英文資料集 策略二及 max pooling



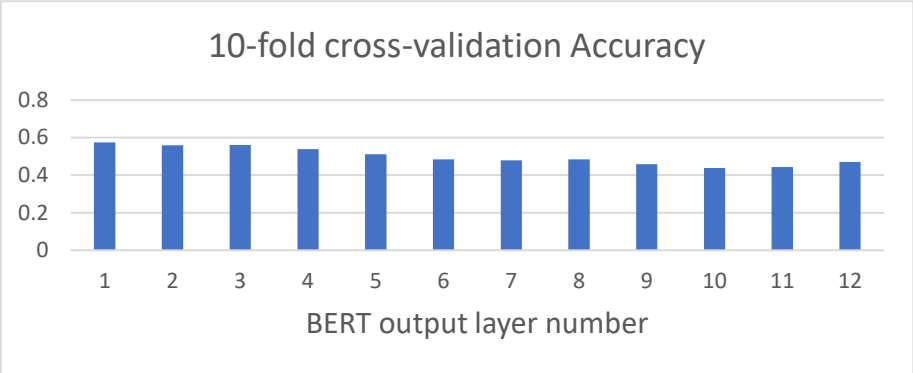
圖十一、英文資料集 策略二及 mean pooling

從圖十、圖十一可以觀察到，英文資料集使用策略二及 mean pooling 時，以 18-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.3487 的準確率。

2、中文資料集

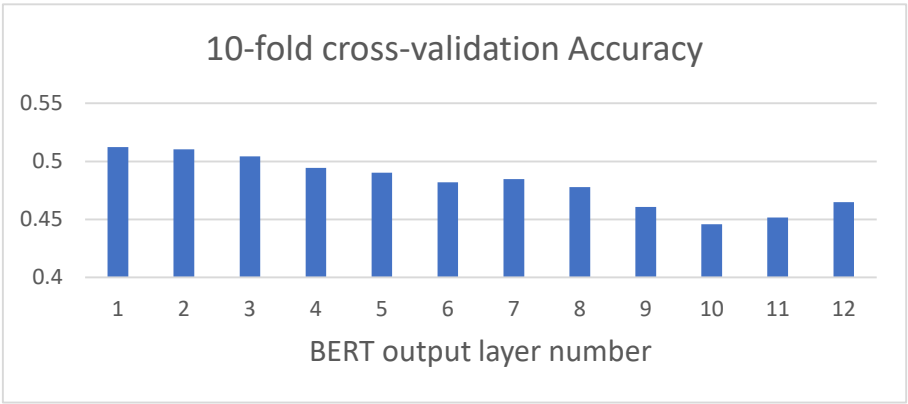


圖十二、中文資料集 策略一及 max pooling

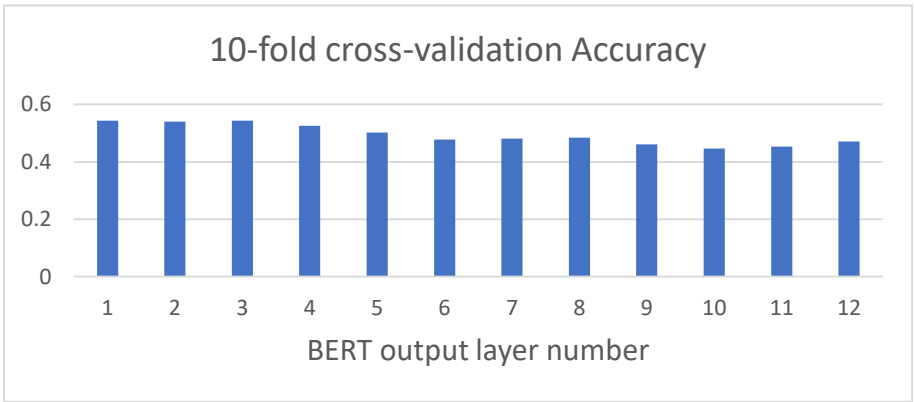


圖十三、中文資料集 策略一及 mean pooling

觀察圖十二、圖十三，中文資料集使用策略一及 mean pooling 時，以 1-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.575 的準確率。



圖十四、中文資料集 策略二及 max pooling



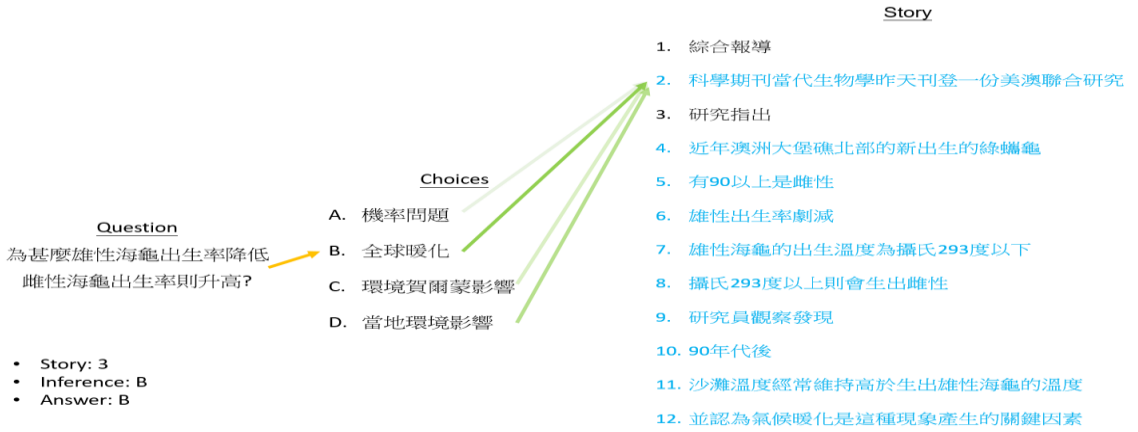
圖十五、中文資料集 策略二及 mean pooling

觀察圖十四、圖十五，中文資料集使用策略二及 max pooling，以 1-layer 的 output 作為詞嵌入獲得該組合最好的結果，最終取得 0.5439 的準確率。

（四）、策略方法解析

圖十六、策略一解析

策略一結果如圖十六，利用合併 Question 以及 Choice 的資訊(如黃色箭頭)，以及將 Story 切割為句子來檢視故事中的資訊，進一步計算最相關於選項的 2 個句子(綠色越深代表關聯越強，紅色越深代表關聯越差)，透過計算最相關的句子，能更詳細的考慮故事與選項之間的關聯，避免沒有正確資訊的句子，與單一選項關聯過高的情況。

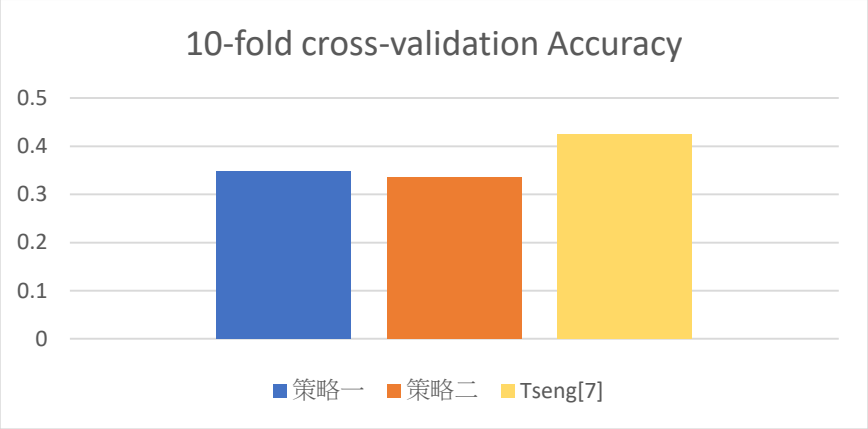


圖十六、策略二解析

策略二結果如圖十六，利用合併 Question 以及 Choice 的資訊(如黃色箭頭)，以及將 Story 切割為句子來檢視故事中所有的資訊，透過保留與問題的關聯度 Top10 的 Story 句子(如圖藍色部分)，能夠幫助篩選掉部分與選項不相關的故事句子，最終只計算與選項有相關的故事句子(綠色越深代表關聯越強，紅色越深代表關聯越差)。

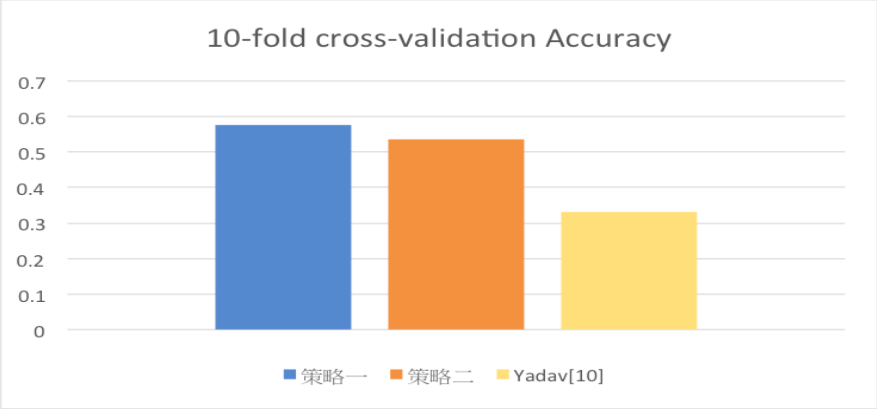
（五）、答題模型結果比較

與提出的二種答題模型比較，在中文上我們實驗了 Yadav 等人[10]的模型，最終得到 33.2%準確率，以及比較 Tseng 等人[7]的結果 42.5%準確率，作為最終結果對照。



圖十七、英文資料集 10-cross validation Accuracy

在英文測試資料集上如圖十七，最佳的答題模型為策略一，最終結果為 34.87%。



圖十八、中文資料集 10-cross validation Accuracy

在中文測試資料集上，最佳的答題模型為策略一，最終結果為 57.5%如圖十八。

五、結論

本論文基於英文與中文兩種語言的閱讀測驗，提出兩種答題模型，在中文資料集上，策略一的模型優於所有結果，此外策略一模型使用在英文時也有一定的效果，經過實驗證實，切割故事的所有句子，有助於保留重要的答案句子資訊，透過合併問題及答案的資訊，可以讓模型更全面的檢視線索，進而讓答題模型找出正確答案。針對答題模型，可以藉由投票機制，組合多個答題模型的結果，以此增加多種不同的答題策略，來彌補不同答題模型之中的缺陷，改善最終的準確率。而資料則可以改善斷句、資料集數量的增加，以及針對不同長度的故事、問題、選項之間優化答提模型的設計方式。

參考文獻

[1] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998-6008.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

[4] Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications, 68*(13), 13-18.

[5] Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (Vol. 4, pp. 9-56).

[6] Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000, May). A comparison of document clustering techniques. In *TextMining Workshop at KDD2000 (May 2000)*.

[7] Tseng, B. H., Shen, S. S., Lee, H. Y., & Lee, L. S. (2016). Towards Machine Comprehension of Spoken Content: Initial TOEFL Listening Comprehension Test by Machine. *Interspeech 2016*, 2731-2735.

[8] Fang, W., Hsu, J. Y., Lee, H. Y., & Lee, L. S. (2016, December). Hierarchical attention model for improved machine comprehension of spoken content. In *2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 232-238). IEEE.

[9] Richardson, M., Burges, C. J., & Renshaw, E. (2013, October). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 193-203).

[10] Yadav, V., Sharp, R., & Surdeanu, M. (2018, June). Sanity check: A strong alignment and information retrieval baseline for question answering. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1217-1220). ACM.

[11] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November).

SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383-2392).

[12] Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., & Fidler, S. (2016). Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4631-4640).

[13] Hermann, K. M., Koč iský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015, December). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1* (pp. 1693-1701). MIT Press.

[14] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In ICLR.

[15] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading compre- hension. In ICLR, 2018a.

[16] Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: a chinese machine reading comprehension dataset. CoRR, cs.CL/1806.00920v2.

[17] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*(Vol. 14, No. 2, pp. 1137-1145).

[18] 科技部,科技大擂台_測試資料集：
https://scidm.nchc.org.tw/dataset/grandchallenge

[19] Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017, July). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 189-198).

[20] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT* (pp. 2227-2237).