Date-duration 與 Arithmetic 模組之技術報告　郭家銍

摘要

近年基於 BERT 系列模型之問答系統，仰賴指針網路直接抽取存在於文章中的答案，此種作法雖在諸多公開數據集上取得佳績，卻在涉及數字運算、單位換算的問題上束手無策，也因此成為近年許多學者試圖攻克的研究主題。在此次科技大擂台競賽中，我們採用 BERT 神經網路模型預測一至數個候選答案，並觀察文章、問題、候選答案間，考量頻繁出現的模版（例：出生年月日）、問題指涉之類型與單位、候選答案間的單位一致性等因素後，再以建立好的規則庫系統進一步做出推論，得到最終的答案。

一、簡介

在定義上，舉凡答案涉及日期、時間、朝代、年齡等之問題，皆被我們歸類在 Date-duration 中；而答案涉及數字、數量等之問題，則歸類於 Arithmetic 中。根據觀察，Arithmetic 問題與 Date-duration 問題存在著高度相似的技術難點，諸如單位換算與四則運算等等，其中 Date-duration 出於人類日常生活使用的時間進位系統，可以將 Date-duration 問題視為具備較高複雜度與特殊模版的 Arithmetic 問題，如下範例一所示。有鑑於此，在巨觀的思想與設計上，我們對 Arithmetic 模組與 Date-duration 模組，採用相同的系統架構與研究手法。

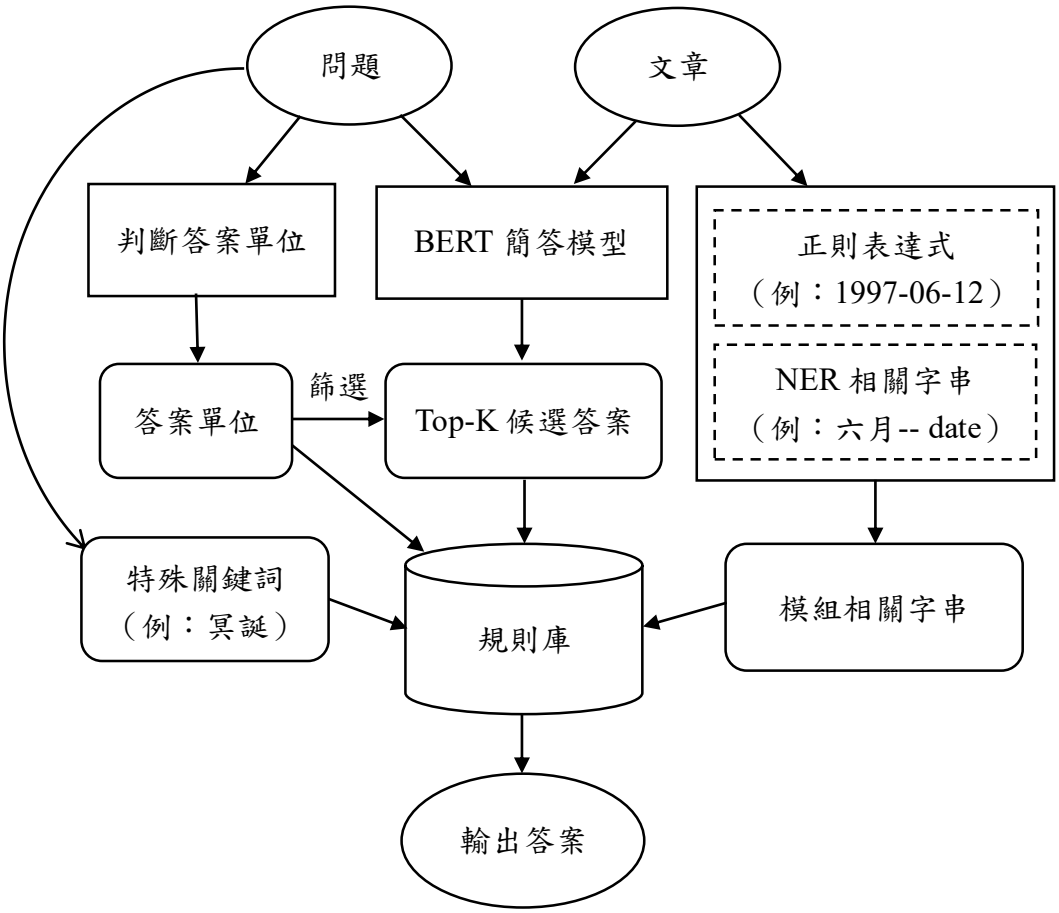
Arithmetic 問題：「小明有三顆蘋果，他吃完一顆還剩幾顆？」
Date-duration 問題：「今天是三月四日，一個禮拜前是幾月？」

範例一、四則運算於 Date-duration 問題的體現

二、系統架構

我們採用以 BERT 為原型之簡答模型以抽取 Top-K 之候選答案，在實驗過中，我們嘗試過 Tensorflow 團隊所訓練之 BERT 權重、XLNet 之中文權重，與哈爾濱工業大學所開源之 RoBERTa-wwm-ext-large 等等。其中，RoBERTa-wwm-ext-large 雖對於簡答題之 Top-1 答案有著最高的 Exact Match 與 F1 成績，然而因四則運算等問題皆需要使用至少兩項由神經網路預測之候選答案，在人工針對 Date-duration 與 Arithmetic 之所有問題做出粗略的篩檢後，可以發現以 RoBERTa-wwm-ext 為首的 base 模型，在稍作訓練的狀況下可以取得最高的 Top-K 準確度，更能滿足使用單個至多個候選答案之規則庫系統。

針對 Arithmetic 與 Date-duration 模組，我們採用相同的系統架構：首先，問題與文章將經由 BERT 簡答模型以給定 Top-K 之候選答案，接著我們又從文章中使用正則表達式與 NER 抽取與模組相關之重要字串。再來，我們從問題中抽取答案指定之單位，並以此單位作為篩選答案之依據，其中某些單位具備交換性（例：一名、一位），問題中出現之明確關鍵詞（例：過世）亦將成為規則庫之參考依據。將上述所有蒐集來的資料丟入規則庫當中，我們的系統即可輸出最後的答案，如下圖圖一所示：



圖一、Arithmetic/Date-duration 模組之系統架構

三、資料集與模型訓練

於 FGC 1.7.6 版後之資料標記規則，類型符合之基礎簡答題，亦被歸類在 Arithmetic/Date-duration 問題當中，而專門針對四則運算等進階題所特化之第一版與第二版模組並不具備支援簡答題的通範性，也因此誕生了圖一所示之第三版系統架構。然而，在進階題高度仰賴簡答模型之 Top-K 之準確度的同時，基礎簡答題卻仰賴著 Top-1 之準確度，即使改善了系統架構，第三版模型依然無法有效得同時在兩種問題之間周旋，甚至連維持答題的平衡性也難以做到。

有鑑於此，我們於第四版模組中，改變了原先對 BERT 系列模型之訓練方式，我們提高 Dropout 機率為 0.5，並採用分階段的課程學習(Curriculum Learning)方法，以避免模型因 Cross-entropy 的特性使得 Top-2 至 Top-K 在過度訓練後失去重要性，並同時維持 Top-1 的高準確度。在我們現有的訓練資源中，存在著品質較低但樣本不少的數據集，如 ASR 與 DROP 數據集，亦存在品質優良之數據集，如 DRCD, Kaggle, Lee 與 FGC 數據集，我們遵循以下三個步驟訓練模型：

步驟一、　混合所有數據集，訓練一個 Epoch
步驟二、　混合優良數據集，訓練至 F1 成績收斂
步驟三、　單獨使用 FGC 數據集，訓練至 EM 成績收斂

遵照此法訓練之模型，雖然其簡答題之 Top-1 準確度稍差於專門訓練之簡答模型，然而其 Top-2 至 Top-K 之準確度卻遠遠領先。第四版 Arithmetic 與 Date-duration 模組改採此種訓練方式而在 EM 成績獲得了顯著的改善，如表一所示。以此法訓練之簡答模型，若用於 Ensemble 或許亦能取得佳績，尚待實驗驗證之。

	訓練集	驗證集	測試集
Arithmetic 第三版	68.18	60.87	64.29
Arithmetic 第四版	**77.46**	**81.48**	**75.00**
Date-duration 第三版	57.26	69.57	56.25
Date-duration 第四版	**76.58**	**84.62**	**77.42**

表一、課程學習對 Arithmetic 與 Date-duration 模組之影響

四、結論

針對簡答題無法直接解決之進階問題，我們採用正則表達式與 NER 從文章與問題中抽取相關資訊，並配合基於 BERT 系列模型之簡答模型，設計一套規則庫使得系統具備四則運算與單位換算之能力。我們亦以課程學習的訓練手法，使系統採用之模組參數，能在回答基礎簡答題與進階題之間，能取得平衡而不會顧此失彼，這使得最終的系統成績獲得了顯著的改善。

參考文獻

[1] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, “Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs,” North American Association for Computational Linguistics (NAACL), 2019.

[2] D. Andor, L. He, K. Lee, E. Pitler, “Giving bert a calculator: Finding operations and arguments with reading comprehension,” 2019.

[3] Q. Ran, Y. Lin, P. Li, J. Zhou, Z. Liu, “Numnet: Machine reading comprehension with numerical reasoning,” 2019.

[4] M. Hu, Y. Peng, Z. Huang, D. Li, “A multi-type multispan network for reading comprehension that requires discrete reasoning,” 2019.