**教育部全國大專校院人工智慧競賽(AI CUP)**

**機器閱讀紀錄-課程挑戰賽**

隊伍: ha

成員: 姜立垣

1. 環境

使用 google colab 環境訓練，相關套件版本如下：

pandas==1.1.5

numpy==1.19.5

tqdm==4.41.1

scikit-learn==0.22.2.post1

torch==1.7.0+cu101

tensorflow==2.4.0

Keras==2.4.3

nlpaug==1.1.1

matplotlib==3.2.2

transformers=4.1.1

除上述套件之外，預訓練模型使用 scibert[1]。

1. 資料處理

使用哪些資料、對資料做了那些刪減或增補。

前處理：使用比賽預設提供資料 Train.csv，合併 Title 與 Abstract，並把 $$$ 取代為 [SEP] 分隔符，稱為 data\_orig。

資料擴充：使用 back translation 技術擴增資料，將 data\_orig 輸入到 MarianMT[2] 的 opus-mt-en-ROMANCE 翻譯成 data\_other\_lang，再將 data\_other\_lang 輸入到 MarianMT 的 Helsinki-NLP/opus-mt-ROMANCE-en 翻譯為英文，稱為 data\_backtranslated。

data\_orig 和 data\_backtranslated 輸入 scibert tokenizer後，只取前 400 個 token 作為模型輸入。

資料分割：取 75% 資料作為訓練集，15% 作為驗證集，需注意 data\_orig 與 data\_backtranslated 為配對分割，即來源相同之句子需在同一個集合中，避免某種程度上的資料污染。

標籤重製：將 4 類別分類轉為 3 類別，預測完後需再把三者皆輸出 0 的樣本轉為 Others 類別以符合 submission 格式。

1. 模型架構

架構遵循 scibert。

1. 訓練方式

epoch 設為 6，sigmoid 輸出 1門檻設為 0.4，Adam lr 設為2e-6。

1. 分析&結論
2. 使用 scibert 因為更符合科學文獻，對比一開始使用 bert 結果，有 3~4% 提升。
3. 使用 back translation 擴增資料增加資料多樣性，本意是希望得到語意相近的擴增句。
4. finetuned 時不需很多 epoch，從實驗結果來看 10 個 epoch 一定足夠，lr 可設低一點，因模型信心不足，所以我們這邊手動實驗門檻調整，設為 0.4 為佳。
5. 未來可嘗試
   1. 其他模型，但使用 scibert 字典
   2. Supervised Contrastive Learning [3]
6. 程式碼

詳細程式碼請參考<https://github.com/newman1234/thesis_reading>，使用說明參照其中的 README.md。

1. 使用的外部資源與參考文獻

[1] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

[2] Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Heafield, H. H. K., Neckermann, T., Seide, F., ... & Birch, A. Marian: Fast Neural Machine Translation in C+.

[3] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. arXiv preprint arXiv:2004.11362.