2021 年國立高雄科技大學電機與資訊學院

學生實務專題製作競賽暨成果展

成果報告書

以 BERT-LSTM 架構進行基於閱讀興趣之書籍自動分類方法

參賽類組： 其他類組

110 年　　4 月　　30 日

EECS-2021-0430

以 BERT-LSTM 架構進行基於閱讀興趣之書籍自動分類方法

李冠篁、方皓昀
國立高雄科技大學資訊工程系
E-mail: C107181112@nkust.edu.tw
C107151113@nkust.edu.tw

摘要：

自古以來人類對於知識的渴望從不消停，大量的知識與前人的經驗匯集成書籍。從眾多書籍中依照自己興趣找到想讀的書是推動閱讀教學重要的環節。在自然語言處理(Natural Language Process)領域已有許多方法提出文本分類方法，也有一些依照傳統圖書分類的自動化方法，但是對於書籍的閱讀興趣分類卻付之闕如。

本專題嘗試使用 BERT-LSTM 作為架構。基本構想是將書籍轉換為語意空間的向量表示，再透過 LSTM 長期記憶的特性，使模型能記住整本書的語意變化。本專題以一個已標記閱讀興趣類別的書籍語料庫做為實驗資料，初步實驗結果顯示我們提出的模型有相當好的準確性。

關鍵詞：閱讀興趣、書籍分類、LSTM、BERT

壹、前言與研究目的

無論是哪個年級的學生，閱讀一直都是吸收知識最快也最好的方式。不只是在學校學習，在課外的時間許多人也選擇走進書店挑選自己喜愛的書做閱讀。但是書籍琳瑯滿目，現有圖書分類架構常常不符合學生的需求；就算只看簡介，也會覺得負擔沉重，有可能連看書的興致也一併退去。因此要想推動閱讀教育，我們認為第一步必須要先將類別劃分好。

實際上，在一般書店的書籍銷售也常常會因為客層屬性不同，採用不同的分類模式。例如國家圖書館與一般書店的分類不同；即使是書商之間，誠品與天瓏書局(電腦專業書籍通路商)對書籍的分類也不一樣。因此，如何對書籍依照需求分類，是一個兼具教育與商業價值的應用問題。

在 NLP 領域文本分類問題已經有許多相當好的成果。Yang et al. (2016)透過有層次的注意力模型，讓整體架構從 Word、Sentence 處理到 Document，電腦能更清楚知道文章的架構，並做出預測；Lai et al. (2015)利用 RCNN 雙向循環結構，很好的保留了詞的順序訊息；Chen & Guestrin (2016)利用遞迴樹結構 XGBoost 處理文本分類。然而，這些方法在書籍分類上卻沒有好的效果。上述方法大多針對短文本做預測，模型架構並不能完整的考慮整本書，若應用在書籍分類上，則會產生語意向量不夠精準的問題。

在書籍的分類上還有第二個問題，書籍的段落很多，而現有短文本分類模型，因為主題明確、段落不多，所以只使用句子或只用全篇短文分析語意時不會有問題。但是在書籍分類時，整本書通常有多個主題，若直接用整本書輸入傳統的分類模型，得到的分類結果，可能會太模糊；而用句子語意又太破碎。因此本專題以 BERT-LSTM 架構解決上述所提到之書籍分類的困難。

貳、原理說明

圖 1 為本專題的做法。主要想法是在第一階段先將書以段落為單位以 BERT(Devlin et al., 2019)轉換為語意向量。這個做法的好處是能充分表現長文本在不同位置要表達的語意。另外，許多研究都證實 BERT 是相當好的語意空間模型，能解決先前其他方法語意表示誤差的問題。

接著第二階段藉由 LSTM 模型整合各段落語意形成篇章語意，並由機率模型進行分類。由於 LSTM 是序列神經網路，會考慮前後文之間的關係，因此適合做為整合段落語意使用。整合後的語意，會經過一個分類器得到所屬的分類，再由機率模型統計後得到最終結果。以下各小節詳述各部分的方法細節。

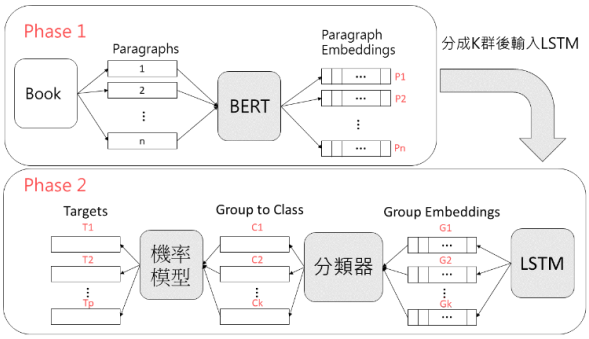

圖 1 本專題所提方法的主要架構

2.1 段落語意生成

圖 1 Phase 1 我們會將書本依照原文的分段切割成 n 個段落，再經由 BERT 將其分別轉換成語

意向量產生 P_1~P_n。使用 BERT 的原因是它能根據整個段落內容直接得到段落語意，不需使用將單詞語意做平均等作法，其他研究指出它的準確率遠高於其他模型。

有些分類模型採用將整篇文章直接放入 BERT 取得全篇語意後進行分類，但在書籍這類長文本中表達的主題眾多，就算 BERT 有強大的語言理解能力，當一次輸入整本書時，會得到的單一語意表示其實是非常模糊的，很多細節都會遺失。這會影響分類的正確性。為了讓模型能更準確掌握文意，我們與現有多數模型不同，採取以段落為語意的計算單位，既有一定的資訊量避免估算誤差、也不至於有語意模糊的問題。

2.2 篇章語意

從圖 1 可知，假如有本書有 n 段文字，經過先前的 BERT 我們將會得到代表每段語意向量的 P_1~P_n。在輸入 LSTM 之前，我們會先將其分成 k 群，每群都有 n/k 個段落。每群的段落會依序輸入 LSTM，就會得到每群的語意向量 G_1~G_k，過程如圖 2 所示。

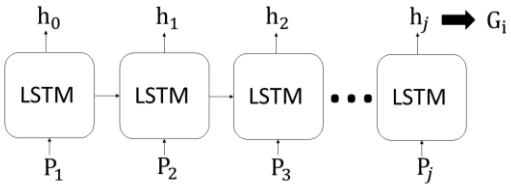

圖 2. LSTM 產生段落群語意向量過程示意圖

我們認為一個段落群代表一個主題，這個主題的形成不只是整個群內文字的語意加總、段落間的語意銜接與轉換也必須考慮，而 LSTM 相當適合滿足這個需求。接著，群的語意向量 G_1~G_k 會被逐一輸入分類器預測其分類 C_1~C_k。由於這些分類結果可能不同，因此我們使用公式(1)的機率模型計算整本書屬於各個分類的機率。

$$T_j = \frac{\sum_{i=1}^{k} H_i}{k}, \text{其中 } H_i = \begin{cases} 1, & C_i = j \\ 0, & Otherwise \end{cases} \quad (1)$$

參、成果與結論

本專題所用之資料來自台師大華語文與科技中心提供之 650 本書籍，共分為 13 個類別，每個類別為 50 本書。因為有訓練與測試需分開的需求，所以每類的 50 本書，被劃分為 45 本書做訓練資料，共 585 本，剩下的 5 本書為測試資料，共 65 本書。每本書皆已由人工分類。

本專題初步使用兩種指標評估分類效果。第一種是，如果機器分類預測該書人工分類之類別的機率為最高則視為正確，此正確率稱為 Top 1。第二種指標為，如果機器分類預測該書人工分類之類別為預測機率最高的前三種類別，就視為正確，此正確率稱為 Top3。

有兩種指標的原因在於我們認為一本書的分類很少只有一種，尤其是小說等長篇文學，能同時符合多種分類。以"馬可波羅"這本書為例，人工分類為史地傳記，而模型認為其為"旅遊紀實"的機率最高，"史地傳記"為第二高。但模型的判斷也是很有道理，因此我們採用兩種指標反映真實情況。

表 1 測試資料預測結果

	Top 1	Top 3
正確率(%)	76.9 %	90.8 %

表 1 為測試資料結果分類結果。我們檢視在 Top 3 中人工分類並非最高機率值的書籍，發現機率高於人工分類的類別絕大多數也相當合理。目前此模型已被使用在一個商業的圖書應用系統上。我們希望日後能嘗試讓 Top 1 的正確率再進一步提升。

參考文獻

[1] Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., & Gipp, B. (2019). Enriching bert with knowledge graph embeddings for document classification. arXiv preprint arXiv:1909.08402.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] 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, 30, 5998-6008.

[4] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning--based Text Classification: A Comprehensive Review. ACM Computing Surveys (CSUR), 54(3), 1-40.

[5] Hochreiter, S., urgen Schmidhuber, J., & Elvezia, C. (1997). LONG SHORT-TERM MEMORY. Neural Computation, 9(8), 1735-1780.

[6] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the NAACL: human language technologies (pp. 1480-1489).

[7] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

[8] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD (pp. 785-794).