The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

結合領域知識之語言轉譯器於中文醫療問題多標籤分類
Incorporating Domain Knowledge into Language Transformers for Multi-Label Classification of Chinese Medical Questions

陳柏翰 Po-Han Chen, 曾昱翔 Yu-Xiang Zeng, 李龍豪 Lung-Hao Lee
國立中央大學 電機工程學系
Department of Electrical Engineering, National Central University
{108521106, 109521127}@ncu.edu.tw, lhlee@ee.ncu.edu.tw

摘要

我們提出知識導入語言轉譯器模型架構，將弱監督層級資料視為知識來源，由其上下文推理並預測出被遮罩的焦點與面向，藉以捕獲相關領域知識。有鑑於當前缺乏公開的中文醫療問題多標籤分類資料集，因此我們從網路上蒐集醫療問題，並且人工標記 1,814 則問句，橫跨 8 個問題類別：原由、疾病、檢驗、醫療資訊、營養補充、人物機構、症狀、以及治療，標籤總數是 2,340，每則問題平均 1.29 個標籤。我們以百度醫學百科當作領域知識來源，比較 BERT 和 RoBERTa 兩個轉譯器的效能差異，實驗結果得知我們的知識導入機制，在不同的評測指標 Macro F1、Micro F1、Weighted F1 及 Subset Accuracy 都能有效提升效能 。

Abstract

In this paper, we propose a knowledge infusion mechanism to incorporate domain knowledge into language transformers. Weakly supervised data is regarded as the main source for knowledge acquisition. We pre-train the language models to capture masked knowledge of focuses and aspects and then fine-tune them to obtain better performance on the downstream tasks. Due to the lack of publicly available datasets for multi-label classification of Chinese medical questions, we crawled questions from medical question/answer forums and manually annotated them using eight predefined classes: persons and organizations, symptom, cause, examination, disease, information, ingredient, and treatment. Finally, a total of 1,814 questions with 2,340 labels. Each question contains an average of 1.29 labels. We used Baidu Medical Encyclopedia as the knowledge resource. Two transformers BERT and RoBERTa were implemented to compare performance on our constructed datasets. Experimental results showed that our proposed model with knowledge infusion mechanism can achieve better performance, no matter which evaluation metric including Macro F1, Micro F1, Weighted F1 or Subset Accuracy were considered.

關鍵字：文本分類、領域知識擷取、預訓練語言模型、生醫資訊學
Keywords: text classification, domain knowledge extraction, pretrained language models, biomedical informatics.

1 介紹

近年來深度學習技術的興起，預訓練語言模型在許多自然語言處理任務皆有著亮眼的表現。在文本分類任務中，轉譯器 (transformer) 網路架構為最廣泛使用的主流模型，通過大規模無標記資料，進行自監督之預訓練，這些模型捕獲了廣域語意資訊與結構句法，並利用這些知識進一步微調下游任務。例如：專有領域之語言模型 BioBERT (2020)，通過遮罩語言模型 (Masked Language Modeling, MLM) 在生物醫學語料庫進行預訓練，該訓練機制旨在捕獲隨機遮罩之標記與其上下文之語意關係。

隨著科技的進步，人類壽命延長的同時，對健康照護的意識也逐漸抬升，許多媒體及報

章雜誌都在談論相關議題，人民也時常在網路上的尋求問題的答案。例如：一般民眾可以在醫聯網 (https://med-net.com/) 上提出健康相關的醫療疑問，專科醫生則在這個平台上根據問題回答。除了由專家或社會大眾回答問題之外，開發自動問答 (Question Answering, QA) 系統讓電腦回答人類問題，也是人工智慧時代的發展重點之一。無論是由人類或者是機器回答問題，要先能理解問題，例如：「請問血栓溶解劑與心悸跟心肌炎有關嗎？」，問題理解上可以歸納成與「治療」和「症狀」這兩個類別有關。

因此，本研究關注中文醫療問題的多標籤分類問題。我們提出知識導入 (Knowledge Infusion, KI) 機制預訓練語言模型，從百度醫學百科蒐集的弱監督層級資料，由其上下文推理並預測出被遮罩之焦點與其面向，藉以捕獲醫療相關知識。有鑑於當前缺乏公開的中文資料集，我們從網路論壇平台上，蒐集醫療相關問題，並且人工標記 1,814 則問句，橫跨 8 個問題類別：原由、疾病、檢驗、醫療資訊、營養補充、人物機構、症狀、以及治療，標籤總數是 2,340，每則問題平均 1.29 個標籤。實驗結果得知，我們的知識導入機制，在四個評測指標 Macro F1、Micro F1、Weighted F1 及 Subset Accuracy，都能有效提升效能。

本文章節如下，第二章探討相關研究，第三章敘述我們提出的知識導入機制，第四章為實驗結果與分析，最後是結論。

2 相關研究

Zhang et al. (2019) 提出一種能同時在知識圖譜及大規模語料庫上預訓練語言模型的方法，名為 ERNIE，其架構分為抽取知識信息與訓練語言模型，與 BERT 類似，隨機遮罩經知識圖譜匹配之命名實體，並訓練模型從知識圖譜中選擇適合的實體進行預測，實現將其知識化的語言表徵模型。

Liu et al. (2019) 提出 K-BERT 模型將知識圖譜三元組作為領域知識注入句子中，引入 soft-position embedding 與可視化矩陣，搭配 BERT 來解決因為過多知識，導致句子偏離其正確意涵之知識噪聲 (knowledge noises) 問題，研究發現雖然 BERT 在經過預訓練後，語言模型可以從大規模語料獲取語言結構信息，但在需要知識驅動的問題時，仍然無法有效發揮。K-BERT 在搭配知識圖譜三元組時，可以輕鬆將特定領域知識注入模型中，後續實驗證明在金融、法律及醫學上之下游任務，相較於 BERT 更亮眼的表現。

Lee et al. (2020) 在英文文本上延續了 BERT，在自挖掘之大規模生物醫學語料進行預訓練，並運用 WordPiece Tokenization，解決專有領域之詞條無法在詞庫表查找的新詞(Out-Of-Vocabulary, OOV)問題，後續實驗證明在生物醫學的下游任務，例如：命名實體識別、關係抽取及問答系統等的效能表現，刷新了排行榜，成為最先進的 (state-of-the-art, SOTA) 模型。

Xiong et al. (2020) 發現目前的預訓練語言模型通常是字符級別，並沒有以實體為中心的知識建模，因此提出了一項新的弱監督 (weakly supervised) 知識學習目標函數，訓練語言模型區分文本中正確的實體與被隨機選擇替換的其他實體。進行實體替換時，通過匹配維基三元組知識庫，選擇將被匹配之實體替換成該類型的其他實體，進一步訓練語言模型，實驗發現從非結構化文本，直接學習實體知識，在下游任務上有顯著的成長。

Wang et al. (2021) 提出了一種知識嵌入 (knowledge embedding) 和預訓練語言嵌入表示，模型名為 Kepler，作者將實體的描述與預訓練語言模型作為嵌入進行編碼與訓練，實驗結果在各項自然語言處理任務上都有更好的效能表現，為知識嵌入研究帶來新的基準。

He et al. (2020) 將維基中與疾病相關的知識，注入到 BERT 中並進行預訓練，在消費者健康問答、醫學語言推理及疾病命名實體識別上都取得了更好的效果。

本研究以預訓練語言模型做為基礎，有別於 He et al. (2020)，我們加入了以疾病、症狀、治療方法、檢測方式、藥物、食品等為焦點，以及從焦點出發相應延伸的不同面向，以弱監督層級資料作為預訓練時之遮罩目標，使醫療健康照護特徵資訊能夠充分完整。

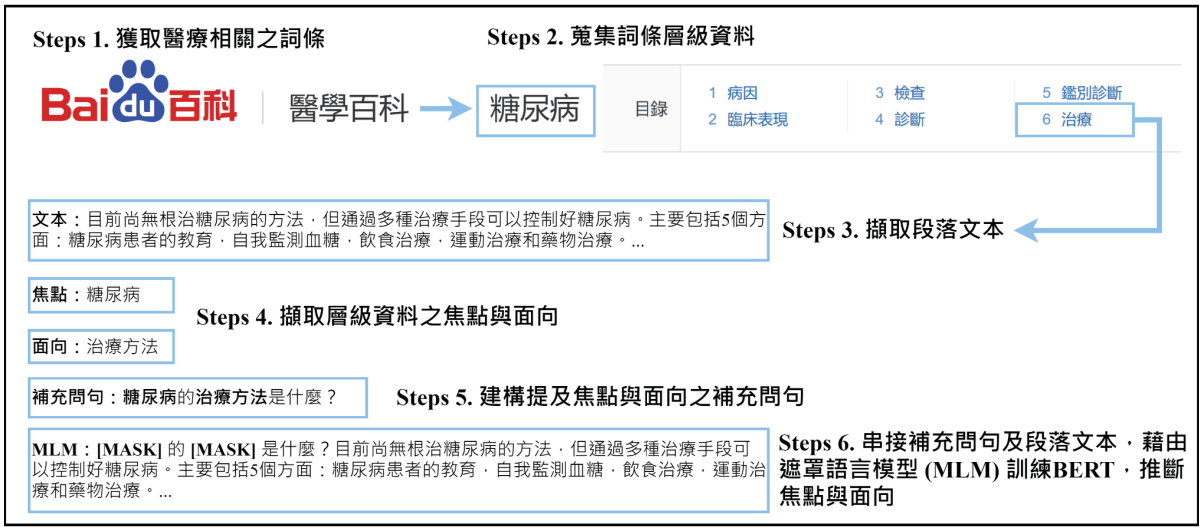

圖 1: 知識來源：百度醫學百科

3 知識導入語言轉譯器模型

3.1 知識來源

臨床文獻與生物醫學網站的文章通常包含多個疾病、症狀及診療技術等，導致難以判別原文焦點及面向，且經醫學專家協助標註通常是昂貴且費時的，諸如英文的 MeSH (Medical Subject Heading) 與 SNOMED CT 都提供了醫學術語供查詢。

我們選擇百度醫學百科 (如圖 1)，作為弱監督層級知識來源。以糖尿病為例，該目錄包含了以糖尿病為焦點衍伸出的多個面向，我們接續提取對應於各面向的段落文本。假設我們將糖尿病視為焦點，治療方法視為面向，可以由層級資料特徵建立事先定義好的制式的補充問句：「[焦點] 的 [面向] 是什麼？」，制式補充問句的優點是當模型在預測被遮罩的標記時，因為各個面向都使用了相同的制式補充問句，所以不會提供模型線索，強迫其根據段落文本學習。我們然後將制式補充問句與提取的段落文本串接起來，建構出問題形式的文本，用新的損失函數L，進行遮罩語言模型 (MLM) 訓練。

3.2 知識導入機制

我們提出一個知識導入 (Knowledge Infusion, KI) 機制，假設補充問句中焦點的字序列為 $X = [x_1, x_2, x_3, \dots, x_T]$，補充問句中焦點之交叉熵損失函數如方程式(1)，$passage$ 為補充問句及段落文本建構出之問題形式的文本，$p(x_t|passage)$為條件機率如方程式(2)，其中 z_t 為 x_t 之未歸一化的對數機率分布 (unnormalized log probabilities)，β 為平衡 L_{focus} 因 z_t 數值過小之補償參數，a為面向種類，方程式(3)中的L_{aspect} 為其損失函數，模型在訓練過程中透過降低總損失函數L如方程式(4)，捕獲醫學百科知識。

$$L_{focus} = -\sum_{t=1}^{T} log\, p\,(x_t|passage) + \frac{\beta}{\sum_{t=1}^{T} z_t} \quad (1)$$

$$p(x_t|passage) = \frac{exp\,(z_t)}{\sum_{z \epsilon V} exp\,(z)} \quad (2)$$

$$L_{aspect} = -log\, p\,(a|passage) \quad (3)$$

$$L = L_{focus} + L_{aspect} \quad (4)$$

圖 2 為模型架構，使用了 BERT 作為模型的基礎架構。在預訓練 (pre-training) 階段，輸入的文本經 WordPiece tokenization 處理後，Token Embedding 層將 [CLS] 插入結果的開頭，[SEP]插入第一句結尾與第二句結尾，並轉換為固定維度之向量。Position Embedding 層賦予各個字序列順序的信息。Segment Embedding 層能夠處理輸入句子對之分類任務，前一向量把 0 賦予給第一個句子中之各個字，

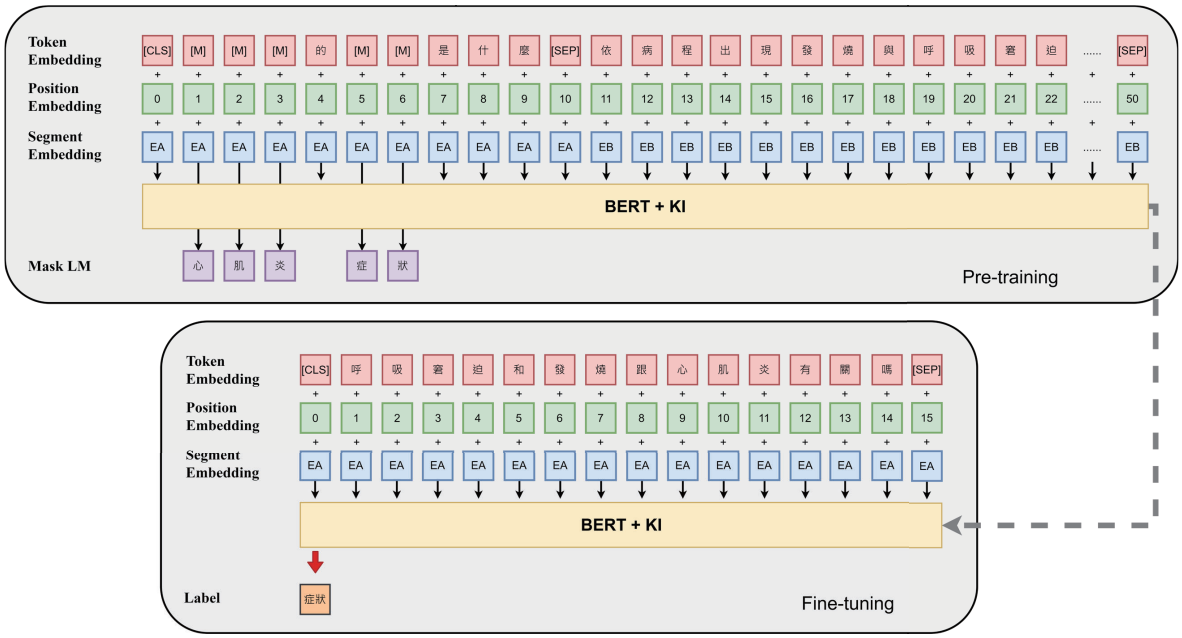

圖 2: 知識導入語言轉譯器模型架構

後一向量把 1 賦予給第二個句子中之各個字。在微調 (fine-tuning) 階段，Transformer 中的自注意力機制允許 BERT 建模於自定義的下游任務上，BERT 根據 [CLS] 標記生成一代表句的特徵向量，並通過一層全連接層進行微調，損失函數根據任務設計，由於是多標籤分類任務，我們使用 Sigmoid 與 Binary Cross Entropy Loss (BCELoss)。

BERT 加上我們的知識導入機制，藉由上下文語意正確預測出被遮罩的焦點與對應的面向，更有效率地學習領域語意資訊。例如：圖一的預訊練階段學習到心肌炎相關症狀，在微調階段能更好理解呼吸窘迫與發燒為心肌炎的症狀，從而在微調時，正確地預測出標籤種類。

4 模型效能評估

4.1 資料集

由於缺乏公開的中文醫療問題分類資料集，我們透過爬蟲將醫聯網 (https://med-net.com/) 的問答紀錄擷取下來，一共有 1,814 則醫療問題。經由歸納整理，總共有 8 個問題類型，其定義和例子如表 1。參與標記的人員一共有三位師大中文系的大學生，對於每個中文句子做人工斷詞及問題類型標註，最終標籤總數是 2,340，每則問題平均 1.29 個標籤。

4.2 實驗設定

我們透過爬蟲蒐集百度醫學百科做為知識來源，資料經過過前處理，濾除字序列長度過短與對應至焦點數量低於 500 則之面向種類，最後包含 103,225 句，焦點包含疾病與症狀、中醫、治療與檢查、以及藥物，相對應的面向如表 2。

我們比較以下兩個轉譯器模型，以及加入我們的知識導入機制後的模型效能差異。(1) BERT (Devlin et al., 2019)：以字作為基礎當作輸入，基於轉譯器的雙向編碼器表示技術，使用中文維基百科語料庫當作訓練資料。 (2) RoBERTa-wwm-ext (Cui et al., 2019)：提出中文全詞遮罩模型，使用中文維基百科語料庫與外部資源當作訓練資料。

4.3 評測指標

多標籤文本分類的結果是一篇文本不僅僅只有單一標籤，無法單純的以二元的方法評估。目前主要的評估方法需要計算出每一個類別的 F1 分數，根據不同的方式綜合各個標籤的 F1 分數以評估多標籤分類器的效能。我們採用以下幾種效能指標：(1) Macro F1：將所有標籤視為平等，計算方式是將各標籤的 F1 先計算出來之後，再取其平均值。(2) Micro

問題類型	定義	範例
原由 (Cause)	事情的緣起與由來	請問報告中的雙側肺尖輕度肋膜增厚原因為何？
疾病 (Disease)	詢問是否為何種疾病	用力深呼吸時，腰部兩側有輕微的疼痛，這是僵直性脊椎炎嗎？
檢驗 (Examination)	詢問做哪一類型的檢查	請問 CA125 檢測值高達 150 是否需要進行什麼其他檢測呢？
醫療資訊 (Information)	詢問檢測、疾病、症狀、醫療保健等的醫療資訊與建議	請問同時染上愛滋病或其他性病的機率有多大呢
營養補充 (Ingredient)	關係食品或補給品的問題	醣尿病患者可以喝白蘭氏雞精？
醫療資訊 (Information)	詢問疾病科別、醫療機關或人物	請問猝睡症哪裡有權威醫師？
人物機構 (Person & Org.)	對身體產生之影響與狀況或併發症引起其他症狀	我的ｒ－麩胺酸轉化酶偏低，請問會有什麼影響嗎？
治療 (Treatment)	管理或照顧患者以對抗疾病或病症	有椎間盤突出症狀一定必須要靠手術治療嗎？

表 1: 問題類型定義及範例

焦點	面向	數量
疾病與症狀	病因、臨床表現、檢查、診斷、鑑別診斷、治療、簡介、預防、併發症、預後	40,821
中醫	簡介、入藥部位、性味、歸經、功效、主治、相關配伍、用法用量、相關論述、形態特徵、生長環境、病因、治法、採集加工	8755
治療與檢查	簡介、正常值、臨床意義、注意事項、檢查過程、相關疾病、相關症狀	10,765
藥物	成份、性狀、功能主治、規格、用法用量、不良反應、禁忌、注意事項、貯藏、簡介、藥物相互作用、適應症、藥理毒理	42884

表 2: 焦點與面向的定義及數量

F1: 先計算所有類別加總的 Precision 和 Recall，然後再計算兩者的調合平均 F1。(3) Weighted F1：先將各標籤的 F1，根據每個標籤真實樣本的數量，賦予每個標籤不同的權重，是一種類似加權平均的 F1。(4) Subset Accuracy：這是最嚴格的指標，表示所有標籤都正確的樣本百分比，舉凡有一個標籤分類錯誤，則不將其判斷為正確結果。

4.4 實驗結果

表 3 為模型效能評估結果，RoBERTa-wwm-ext 比 BERT 使用了更長的時間、更大的 batch size 和更多元的數據進行訓練，並去掉了 BERT 中之 NSP (Next Sentence Prediction) 訓練機制和採用了全詞遮罩，從實驗結果得知 RoBERTa-wwm-ext 相較於 BERT 提升了 5.55%的 Macro F1、1.49%的 Micro F1、1.89%的 Weighted F1 與 0.01%的 Subset Accuracy，證實了使用全詞遮罩訓練更多元及序列更長的數據，對下游任務的效果更好。我們提出的知識導入(KI) 機制無論哪個效能指標，相對於 BERT 與 RoBERTa-wwm-ext 模型都能有效提升效能。BERT + KI 相較於 BERT 提升了 3.89%的 Macro F1、1.36%的 Micro F1、1.83%的 Weighted F1

Model	Macro F1	Micro F1	Weighted F1	Subset Accuracy
BERT (Devlin et al., 2019)	0.6979	0.7576	0.7560	0.6082
BERT + KI (ours)	**0.7251**	**0.7679**	**0.7698**	**0.6165**
RoBERTa-wwm-ext (Cui et al., 2019)	0.7366	0.7689	0.7703	0.6083
RoBERTa-wwm-ext + KI (ours)	**0.7386**	**0.7750**	**0.7763**	**0.6220**

表 3: 模型效能評估結果

與 1.36%的 Subset Accuracy。RoBERTa-wwm-ext + KI 相較於 RoBERTa-wwm-ext 提升了 0.27%的 Macro F1、0.79%的 Micro F1、0.78%的 Weighted F1 與 2.25%的 Subset Accuracy。

5 結論

我們提出由百度醫學百科作為弱監督知識來源，藉由知識導入機制，繼續訓練微調語言模型的作法。從實驗結果證實，知識導入機制能更準確地完成醫療問題多標籤分類任務標籤，在 Macro F1、Micro F1、Weighted F1 和 Subset Accuracy 都能有效提升效能 。

致謝

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3

參考文獻

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, Guoping Hu. 2019. Revisiting Pre-Trained Models for Chinese Natural Language Processing, In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657-668. https://arxiv.org/abs/2004.13922

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. 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)*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, James Caverlee. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition, In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614. https://doi.org/10.18653/v1/2020.emnlp-main.372

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2019. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901-2908. https://doi.org/10.1609/aaai.v34i03.5681

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, Jian Tang.2021.KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics,* 2021(9):176-194. http://doi.org/10.1162/tacl_a_00360

Wenhan Xiong, Jingfei Du, William Yang Wang, Veselin Stoyanov. 2020. Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model. In *Proceedings of the 2020 International Conference on Learning Representations*, https://arxiv.org/abs/1912.09637

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451. http://doi.org/10.18653/v1/P19-1139