NCU1415 at ROCLING 2022 Shared Task: A light-weight transformer-based approach for Biomedical Name Entity Recognition
基於 Transformer 的生醫輕量化命名實體識別系統

馮智詮 Zhi-Quan Feng, 陳柏凱 Po-Kai Chen, and 王家慶 Jia-Ching Wang
Dept. of Science and Information Engineering,
National Central University
No.300, Zhongda Rd., Zhongli Dist., Taoyuan City 32001, Taiwan

Abstract

NER (Name Entity Recognition)是傳統 NLP 任務中非常重要也是非常基礎的一項。在生醫領域，各廠商發展的各種技術中，NER 任務都有被廣泛地運用。其中包括句子語義分析 parsing、問答、對話系統中的關鍵訊息提取和替換以及知識圖譜的實際運用等等。在不同的領域，包括生物醫學、通訊、電商平台等，都需要 NER 技術來識別其中的藥物、疾病、商品等物件。本實作主要對於 ROCLING 2022 SHARED TASK(Lee et al. 2022)的生醫領域 NER 任務，基於語言模型做出了一定程度的調整和實驗。

Name Entity Recognition (NER) is a very important and basic task in traditional NLP tasks. In the biomedical field, NER tasks have been widely used in various products developed by various manufacturers. These include parsing, QA system, key information extraction or replacement in dialogue systems, and the practical application of knowledge parsing. In different fields, including bio-medicine, communication technology, e-commerce etc., NER technology is needed to identify drugs, diseases, commodities and other objects. This implementation focuses on the CLING 2022 SHARED TASK's(Lee et al. 2022) NER TASK in biomedical field, with a bit of tuning and experimentation based on the language models.

關鍵字: 命名實體識別，生物醫學，ROCLING 2022 Shared Task
Keywords: Name Entity Recognition, Biomedical Science, ROCLING 2022 Shared Task

1 Introduction & Related work

隨著 NLP 領域技術的發展，不少廠商、研究機構均著力於發展更高效率和精度的自然語言處理模型和演算法。在許多不同的領域中，NER (Name Entity Recognition) 都是非常重要的任務。目前，NER 的任務主要通過一些經典的語言模型(Language Model)進行。

在自然語言處理任務中，最早從 RNN 開始，逐漸發展出基於 LSTM(Long Short-Term Memory)、GRU(Gated Recurrent Unit)的時序神經網路模型。而後在 2017 年，Transformer (Vaswani et al. 2017)的問世再一次改變了自然語言模型的主流。

1.1 NER implementations based on LSTM or GRU layers

基於 LSTM 和 GRU 的模型在 Transformer (Vaswani et al. 2017)沒有提出時是研究自然語言處理的主要方法，而在 Transformer 模型在 2017 年被提出後，雖然其數量有大量減少，但仍然有不少實作的論文利用其時序建模以及輕量化的特點來完成一些特定的任務。

ULMFiT(Howard et al. 2018)是於 2018 年推出的基於 LSTM 的系統，其預期設計也是基於分類任務和預訓練模型。在 ULMFiT 的系統中，為了讓不同的層學習不同的特徵，提出

了兩個訓練方法，一是學習率分層區別化的微調(fine-tuning)方法，也就是越靠後學習率越大。二是從後向前，每訓練一個 epoch，就解凍一層的逐層解凍訓練方法。除此之外，論文中也採用了 warm-up 的機制，讓模型在沒有抓到特徵時有效獲取特徵。

在具體實作 NER 任務的方面，各廠商和研究機構也提出了不同的系統來解決此問題，例如 2016 年提出的(Ma and Hovy, 2016)通過 CNN 進行初步的特徵提取獲得 Char Representation，並用多層的 LSTM 層來進行後續的分類任務。以及 2016 年同年提出的(Lample et al., 2016)基於多層的 LSTM 以及特徵的前向傳遞，來達到最終的分類目的。

1.2 Transformer-Based Implementations

Transformer 模型最早推出於 2017 年，是一個有深遠影響力的序列資料處理模型架構，其在 word embedding 的部分不僅僅是 token embedding，同時也加入了 position embedding、segment embedding 作為字元的更加精確化的表達。之後的多層 encoder 和 decoder 中，主要用 attention 的機制，提取和篩選序列資料中的特徵。

之後基於 Transformer 模型架構，發展的方向主要分為兩個，一是改進預訓練方法，二是改進模型結構。

在預訓練方法的改進上，比較具有代表性的是 BERT(Devlin et al., 2019)、BART(Lewis et al. 2020)、RoBERTa(Liu et al. 2019)等論文。BERT 提出於 2019 年，是 Transformer 模型提出以後的一次非常有代表性的預訓練結果，其中採用了 Mask Filling，NSP(Next Sentence Prediction)等任務，作為對於 Transformer 模型的一系列標準預訓練方法，該方法將 Transformer 模型的準確率進一步提高。

而 BART 在 BERT 的基礎上，增加了更多的預訓練操作，如打亂句子中部分字詞的順序、隨機替換等等。

RoBERTa 模型不僅在預訓練時去掉了 NSP 的流程，還加入了 Dynamic Masking 的動作，此外其採用了更大的 batch size，讓模型能夠更好地從每個 batch 中提取特徵。

在模型結構的改進方面，代表模型有增大層數規模的 GPT(Radford et al. 2018)、GPT2(Vashishth et al. 2019)等。隨著各種不同模型的陸續提出，人們為了提升模型的準確度，在不斷地增加模型的大小，資源的消耗也是水漲船高。

1.3 Our Approach

基於上述論文提啟發，也考慮到本實作的硬體規格限制，本實作採用 2019 年提出的經典模型 BERT 模型作為語言模型。此外，本實作也參考了 ULMFiT 的學習率分層和 warm-up 機制，再引入條件隨機場(Conditional Random Field，CRF)(Huang et al. 2015)的演算法作為輔助，來達到更好的 NER 識別效果。

2 Method

2.1 Data

本實作之資料來源於 Chinese HealthNER Corpus (Lee et al. 2021)，其公開資料之訓練(Train)、驗證(Val)資料集之資料量如表一，其

	Train	Val
Sentences	25345	2816
Average Length	49.36	50.12

表一：資料集單句數量與長度統計

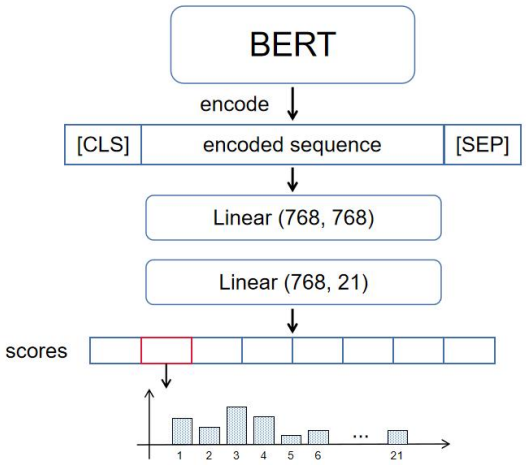
	Train		Val	
Classes	B	I	B	I
O	1110186		123235	
EXAM	1957	4009	261	541
BODY	21022	26162	2218	2744
DISE	8287	18275	787	1723
SYMP	10279	12700	1144	1390
TREAT	2664	510	241	481
CHEM	5483	10369	607	1124
TIME	1451	2080	158	215
SUPP	1286	3264	117	288
INST	959	1914	88	156
DRUG	1925	4611	221	473

表二：資料集標籤數量統計

標籤數量分佈如表二所示。資料集中記錄了 Train 和 Val 資料集的句子數量和平均長度。其中，資料集中的資料均為生醫領域相關，所有標籤均為症狀、藥品、疾病等生醫相關內容，表二為各個種類標籤出現數量之統計結果。不難發現，訓練集和測試集資料量比例大致為 9:1，其中兩個資料集單句資料長度相當，個標籤所對應之 B(表示 name entity 的開始)與 I(表示 name entity 的內容)之比值也大致相同。

2.2 Model

本實作所用之模型如圖一所示，由 BERT 和分類器(classifier)組成，其中分類器包括兩層 Linear 函數，模型可獲得各個 token 的分類結果。其模型在最後的分數計算和損失函數(loss function)的部分採用的是 CRF 的機制。



圖一：模型架構圖

2.2.1 Model Structure

本實作之模型架構主要先由 BERT 對輸入資料進行編碼，獲得模型輸出特徵(encoded sequence)，其中 BERT 模型參數源自(Devlin et al. 2019)這節省了大量模型預訓練時間。該輸出特徵再經過兩層 Linear 層的處理，輸入各類別的分數。

2.2.2 CRF

條件隨機場(Conditional Random Field，CRF)(Huang et al. 2015)是一種處理序列標註資料的演算法，其在 NER 問題中有非常廣泛的運用。其大致理念是通過中間矩陣(transitions)將不同字元的分類結果相互關聯，以提高 NER 任務的最終效果。其中間矩陣的維度為類別總數(tag_num)×類別總數，可以理解為相鄰分類標籤同時出現的幾率。如下公式所示：

$$transitions = \begin{pmatrix} t_{1,1} & \cdots & t_{1,tag\_num} \\ \cdots & \cdots & \cdots \\ t_{tag\_num,1} & \cdots & t_{tag\_num,tag\_num} \end{pmatrix} \quad (1)$$

我們假定模型的輸出分數為 x，標註值為 y，則 x 和 y 的序列如下所示：

$$x = (x_1, x_2, ..., x_n) \quad (2)$$

$$y = (y_1, y_2, ..., y_n) \quad (3)$$

其中標註值 y 也包含 y_0(START_TAG)以及 y_{n+1}(END_TAG)。基於此，其分數的計算式改寫為：

$$score(x, y) = \sum_{i=0}^{n} transition\ s_{y_{i+1} y_i} + \sum_{i=0}^{n} feats_{i, y_i} \quad (4)$$

其中 feats 為模型的輸出序列。由此，每一個輸出分數都和其他的分類結果相關聯，可以一定程度增加輸出結果的合理性。由此，由 x 生成 y 的過程可以表示為：

$$P(y \mid x) = \frac{\exp(score(x, y))}{\sum_{all\_possible\_\tilde{y}} \exp(score(x, \tilde{y}))} \quad (5)$$

由此可得損失函數表達式：

$$-\log P(y \mid x) = -\log \frac{\exp(score(x, y))}{\sum_{all\_possible\_\tilde{y}} \exp(score(x, \tilde{y}))} \quad (6)$$
$$= \log \sum_{all\_possible\_\tilde{y}} \exp(score(x, \tilde{y})) - score(x, y)$$

2.3 Fine-tuning

由於本實作基於預訓練模型，因此對預訓練模型進行微調(fine-tuning)是一項非常關鍵的任務。本實作在模型微調方面，不僅限於直接用測試資料，以低學習率直接微調，本實作也進行了一些特別的學習率處理。

2.3.1 Layered learning rate

本實作考慮到不同的層需要學習不同的特征，且不希望後續的微調(fine-tuning)過程對 BERT 預訓練的參數有過大的影響，所以採取了學習率分層衰減的方法。

本實作規定衰減率 k, BERT 模型的 embedding 部分、六層 encoder、六層 decoder、分類器分別定義為 14 個不同的區塊，每一個區塊採用不同的學習率。本實作規定分類器的學習率為 L_{base}, 從分類器往輸入方向，每向前一個區塊，學習率就在原來的基礎上乘以 k。由此，本實作分別研究了衰減率對模型輸出效果的影響。

本實作所採用之基礎學習率 L_{base} 和衰減率 k，根據多次實驗，確定其數值為 4e-5 和 1.2 作為最佳之數值。

	BERT+Cross Entropy			BERT+CRF		
k	Precision	Recall	F1	Precision	Recall	F1
1.0	**0.764**	0.667	0.705	0.709	0.715	0.702
1.1	0.751	0.682	0.709	0.726	**0.718**	0.718
1.2	0.754	**0.696**	**0.719**	0.767	0.697	**0.726**
1.3	0.748	0.669	0.702	**0.772**	0.676	0.717
1.4	0.741	0.650	0.693	0.766	0.679	0.716

表三：K 與 CRF 的對比實驗結果

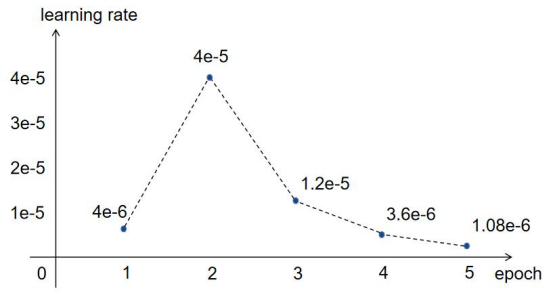
	BERT			RoBERTa			BART		
k	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
1.1	0.726	0.718	**0.718**	0.710	0.712	0.706	**0.753**	0.677	0.709
1.2	**0.767**	0.697	**0.726**	0.760	**0.700**	0.723	0.742	0.698	0.714
1.3	**0.772**	0.676	**0.717**	0.770	0.672	0.714	0.746	0.705	0.716

表四：模型部分替換對比實驗結果

2.3.2 Warm-up and Decay

本實作考慮到，新加入的分類器處於未訓練狀態，因此如圖二所示，本實作考慮採用預熱(warm up)和逐漸下降的方法，先以低學習率讓模型能夠初步獲取特徵，在第二個 epoch 再調回預先設定的正常學習率，再往後，學習率逐漸下降，以對模型進行進一步細微調整。

本實作設計每當程式跑過預設的 step 數量，則重新更新一遍模型的優化器，重新為模型定義學習率。



圖二：基礎學習率變化圖

2.3.3 Other details

本實作根據實驗結果，統一訓練三個 epoch(約 5k 個 step)，其中基礎學習率 L_{base} 為 4e−5, 因考慮到硬體設備條件限制，採用的批次大小為 16。

3 Electronically-available Resources

CPU: Intel(R) Xeon(R) CPU @ 2.20GHz
GPU: 1xTesla K80
CUDA cores: 2496
GPU RAM: 12.68GB
CPU cache size: 56320 KB

4 Experiments

4.1 About k and CRF

根據 Rocling2022 Shared Task(Lee et al. 2022)的內部測試資料，本實作的準確度為 precision 74.56%，recall 72.81%，F1score 73.68

此外，基於本實作所述之實作方法，本實作進行了一些額外實驗，來驗證:學習率衰減率 k 對模型訓練成效的影響, CRF(Huang et al. 2015)和 cross-entropy 對於模型訓練成效的影響。

該實驗的三個模型均來自於已經過預訓練的預訓練模型。BERT 為(Devlin et al. 2019)，RoBERTa 為 (Conneau et al. 2020)，BART 為 (Lewis et al. 2019)。同樣基礎學習率 L_{base} 為 4e-5，批次大小 16，訓練 5k 個 steps，所有數據均為分別訓練五次後的平均值。

本實作所設定之基本參數：基礎學習率 L_{base} 為 4e-5，批次大小 16，訓練 5k 個 steps(3 個 epoch)，其實驗結果如表三所示。

本實驗均訓練五次，記錄數據為五次之平均值。從實驗結果中不難看出，採用 CRF 作為 loss function 和計算分數的輔助，會有效提升模型訓練的 F1 分數。並且，在衰減率 k 為 1.2 左右時，測試結果的 F1 分數能夠達到最佳。

4.2 Replacing BERT to other models

除了對比人為設置的學習率參數以及損失函數，本實作還對採用的語言模型做了對比實驗。

本實作所對比之模型主要為目前主流之語言預訓練模型，RoBERTa(Liu et al. 2019)、BART(Lewis et al. 2020)。

實驗在 1.1、1.2、1.3 三個相對成果較好的學習率衰減率下以及 CRF 作為最後一層的設定下做了對比實驗，如表四所示。

其中不難看出，BERT-base 模型目前在該任務上能夠有更出色的效能，從 F1 分數上，能夠略強於其他兩種主流預訓練模型。也不難發現，從 F1 上看，BERT(Devlin et al., 2019)和 RoBERTa(Liu et al. 2019)的最佳衰減率都在 1.2 左右，而 BART 則在 1.3 左右。

5 Conclution

本實作為 ROCLING 2022 SHARED TASK (Lee et al. 2022)之生醫命名實體識別任務實作，參考了一些論文的實作方法，使用預訓練語言模型和一定程度的模型微調，來達到準確率局部最大化的目的。

通過對比實驗可知，CRF、學習率適當地逐層衰減以及 bert 預訓練模型在命名實體識別的任務上都能在一定程度上有所提升。

相比於模型微調的細節參數變化，系統效能與模型結構和預訓練方法對結果的影響可能更加具有決定作用，因此本實作若可能，之後預期在模型結構、預訓練等方面進行進一步細節上的處理，以提高效能。

6 Reference

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, “Attention Is All You Need” in NeurIPS 2017

Lung-Hao Lee, Chao-Yi Chen, Liang-Chih Yu, and Yuen-Hsien Tseng. 2022. Overview of the ROCLING 2022 shared task for Chinese healthcare named entity recognition. In Proceedings of the 34th Conference on Computational Linguistics and Speech Processing.

Lung-Hao Lee and Yi Lu, "Multiple Embeddings Enhanced Multi-Graph Neural Networks for Chinese Healthcare Named Entity Recognition," in IEEE Journal of Biomedical and Health Informatics, 2021, pp. 2801-2810

Jeremy Howard, Sebastian Ruder, “Universal Language Model Fine-tuning for Text Classification” in ACL 2018, pp. 328-339

Xuezhe Ma and Eduard Hovy, “End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF” in ACL 2016, pp. 1064–1074

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer, “Neural Architectures for Named Entity Recognition” in NAACL 2016, pp. 260–270

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” in NAACL 2019, N19-1423, pp. 4171–4186

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdel-rahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension” in 2020.acl-main.703, pp. 7871–7880

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, 2019, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv:1907.11692

Sascha Rothe, Shashi Narayan, Aliaksei Severyn, TACL 2020, “Leveraging Pre-trained Checkpoints for Sequence Generation Tasks” in 2020.tacl-1.18, pp. 264–280

Zhiheng Huang, Wei Xu, Kai Yu, “Bidirectional LSTM-CRF Models For Sequence Tagging”, arXiv:1508.01991, 2015

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, Manaal Faruqui, “Attention Interpretability Across NLP Tasks”, 2019, arXiv:1909.11218

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, “Improving Language Understanding by Generative Pre-Training” in OpenAI, 2018

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale” in ACL 2020, 2020.acl-main.747, pp. 8440–8451