SCU-MESCLab at ROCLING-2022 Shared Task: Named Entity Recognition Using BERT Classifier

Tsung-Hsien Yang
Chunghwa Telecom laboratories,
Taoyuan, Taiwan
yasamyang@cht.com.tw

Ruei-Cyuan Su, Tzu-En Su, Sing-Seong Chong and Ming-Hsiang Su
Department of Data Science, Soochow University, Taipei, Taiwan
{70613rex, 70614roy, chongzhishan123, huntfox.su}@ gmail.com

摘要

本研究構建了命名實體識別模型，並將其應用於醫療領域。資料是以BIO格式進行標記。例如"肌肉"會被標記成"B-BODY"和"I-BODY"，"咳嗽" 則是"B-SYMP"和"I-SYMP"。不屬於命名實體類別以外的字全標為"O"。訓練資料Chinese HealthNER Corpus 包含 30,692 句，其中的 2531 句切分為此次評測的驗證集(dev)，而最終大會提供另外的 3204 句的測試集(test)。我們分別使用 BLSTM_CRF、Roberta+BLSTM_CRF 與 BERT Classifier 三種方式提交三個預測結果。最後，提交為 RUN3 的 BERT Classifier 系統取得了最好的預測效能，其精準度為 80.18%、召回率為 78.3%，F1-score 為 79.23。

Abstract

In this study, named entity recognition is constructed and applied in the medical domain. Data is labeled in BIO format. For example, "muscle" would be labeled "B-BODY" and "I-BODY", and "cough" would be "B-SYMP" and "I-SYMP". All words outside the category are marked with "O". The Chinese HealthNER Corpus contains 30,692 sentences, of which 2531 sentences are divided into the validation set (dev) for this evaluation, and the conference finally provides another 3204 sentences for the test set (test). We use BLSTM_CRF, Roberta+BLSTM_CRF and BERT Classifier to submit three prediction results respectively. Finally, the BERT Classifier system submitted as RUN3 achieved the best prediction performance, with an accuracy of 80.18%, a recall rate of 78.3%, and an F1-score of 79.23.

關鍵字：命名實體識別、BERT 分類器、醫療
Keywords: Named Entity Recognition, BERT, Medical Domain

1 Introduction

醫療信息是指醫生或相關醫療學者傳達的語言訊息，這些醫療信息中包含具有人體及生物特定意義的實體，而醫生和患者溝通中的信息對於治療與健康影響扮演重要角色(Street and Richard, 2013)。在疫情的影響之下，醫生和患者的接觸大幅度的減少，醫療人員和患者之間的交流越來越多是通過遠端設備進行交流，由此可知電子訊息及健康系統普及化對醫療信息的擷取至關重要 (Weiner, 2012)。龐大醫療數據是可以運用深度學習協助醫生或研究人員進行相關研究，如醫學圖像分類(Azizi et al., 2021)，以及醫療保健對話(Konam and Rao, 2021)，有益於醫療進步。

本研究使用 2022 Rocling 會議之參賽資料進行醫療命名實體識別。此參賽資料是包含醫療相關信息之文字語料集，本研究提出一命名實體識別系統分辨醫療相關專有名稱。這些識別出之醫療相關專有名稱可以方便研究人員對醫療信息正確分析，以及有效協助病患提供醫療信息，或是避免醫療人員情急之下用藥錯誤等情況 (Patanwala et al., 2012)，如此對於醫學問題及協助上，有提供更好的幫助。

命名實體識別模型從傳統機器學習、隱藏式馬可夫模型 (hidden Markov model, HMM)，到深度學習的 BiLSTM、BERT 或 RoBERTa 方法搭配條件隨機場 (conditional random field, CRF) (Huang et al., 2015)，可以更有效提升我們效率以及精確度。因此在任務選擇上，分別採用三個方法來實施，第一個方法是單純

運用 BiLSTM+CRF，第二個的方法是運用 RoBERTa+BiLSTM+CRF，第三種方法是運用 BERT token classifier，分別訓練出各個模型。最後我們將各個預測資料採用標準精度、召回率和 F1 分數進行評估。

2 Dataset

本次研究當中，我們所使用的資料集名稱叫作 Chinese Healthcare Named Entity Recognition (HealthNER)，是由 NCUEE NLP 研究室人員收集與標記 (Lee et al., 2021)。資料是透過爬蟲的技術爬取相關新聞，醫療問答論壇和醫療保健信息。此資料集共有 30,692 句子總計約 150 萬個字。經過人工標注後，共有 68,460 個命名實體，涵蓋 10 種實體類型，根據其名稱分別為人體 (BODY)，症狀 (SYMP)，醫療器材 (INST)，檢驗 (EXAM)，化學物質 (CHEM)，疾病 (DISE)，藥品 (DRUG)，營養品 (SUPP)，治療 (TREAT)，時間 (TIME)。資料是以 BIO 格式去標記。例如"肌肉"會被標記成"B-BODY"和"I-BODY"，"咳嗽"則是"B-SYMP"和"I-SYMP"，以此類推。類別以外的字全標為"O"。而其中區分為訓練資料(train.json)擁有 28,161 句子和測試資料(test.json)有 2531 個句子和 7305 命名實體。由於大會最終是會提供另外的 3204 句當作最終的測試集(test)，故我們可以將 HealthNER 中的測試資料(test.json)當作我們模型的驗證資料集(dev)使用。

3 Proposed Method

3.1 Embedding method

Pytorch 的 embedding 轉換詞向量機制，為一個簡單的尋找表，其模型通常用於存儲詞向量並使用索引檢索它們。模型的輸入是索引列表，輸出是相應的詞嵌入。其模型的可學習權重，使用是初始化均值 (mean)為 0、方差 (variance) 為 1 的常態分佈 (normal distribution)。其輸入值是索引值的張量形式，輸出則是和輸入的張量相同形式維度形式。

得到詞向量後，使用自行定義好的特殊符號作為 mask 組成單元，有 [UNK] 表示 [未知詞]、[PAD] 表示 [填充]、[START] 表示 [文本開頭]、[END] 表示 [文本結束]，共 4 種特殊符號，將每一句子依照上方表示，轉換成每句完整的 mask，以提供標註作為所使用之 label，以此作為下一步驟要使用的輸入值。

3.2 BERT and RoBERTa

BERT (Bidirectional Encoder Representations from Transformer) 模型，是 Google 以無監督的方式利用大量無標記文本的模型。訓練資料來源于 Wikipedia 2.5B 語料集加上 BookCorpus 800M 語料集。批量大小為 1024 * 128 長度或 256 * 512 長度。BERT 分為 BERT-Base (12-layer, 768-Hidden, 12-head) 和 BERT-Large (24-layer, 1024hidden, 16-head) 兩種形式。BERT 無需標記好的資料或解釋即可進行分析。BERT 是 Transformer 的前半部分核心模組(encoder)，而注意力 (attention) 機制是 Transformer 的前段核心部分，主要是增強語義向量，在不同的字結合中，代表識別字所帶來的意思。因此在 BERT 中，注意力機制為 BERT 的主要構成之一。

RoBERTa 是 BERT 模型問世之後的優化模型之一，主要其優化為效能上的優化，用途為分類以及閱讀理解，而進而分別為中文上的預訓練模型以及英文上的模型，其中英文的 RoBERTa 主要訓練的數據集為維基百科及書籍語料庫，中文的 RoBERTa 主要是使用哈工大訊飛聯合實驗室發布的 RoBERTa-wwm-ext-large 模型(Cui et al., 2020)，該模型經過了第三方中文基準測試 CLUE 的驗證。CLUE 的基準測試包含了 6 個中文文本分類數據集和 3 個閱讀理解數據集，其中包括哈工大訊飛聯合實驗室發布的 CMRC 2018 閱讀理解數據集。在目前的基準測試中，哈工大訊飛聯合實驗室發布的 RoBERTa-wwm-ext-large 模型在分類和閱讀理解任務中都取得了當前最好的綜合效果 (Xu et al., 2020)。

3.3 LSTM

LSTM 是為了解決 RNN 的缺點，例如不能準確處理長期序列、時間的資料。LSTM 是由四個閘 (gate) 結構所組成，輸入閘 (Input Gate)，儲存細胞 (Memory Cell)，遺忘閘 (Forget Gate)，輸出閘 (Output Gate)。Input Gate 主要負責控制這個值輸入，Memory Cell 儲存值，下階段在使用，Output Gate 輸出結果，Forget Gate 是否保留或刪除 feature。LSTM 思路就是把輸入到類神經網路層處理產生出結果，過程當中，記住某些特征，然後會跟著這些經驗來判斷

或學習。其中 (1) 至 (4) 分別為 Input Gate, Forget Gate 和 Output Gate 計算公式。其中 C_t 為 memory，h_{t-1} 為 hidden state。

$$f_t = \sigma\left(W_f \cdot h_{t-1} + U_i \cdot X_t + b_f\right) \quad (1)$$
$$i_t = \sigma(W_i \cdot h_{t-1} + U_i \cdot X_t + b_i) \quad (2)$$
$$c_t = tanh(W_c \cdot h_{t-1} + U_c \cdot X_t + b_c) \quad (3)$$
$$C_t = f_t \times C_{t-1} + i_t \times c_t \quad (4)$$
$$o_t = \sigma(W_o \cdot h_{t-1} + U_o \cdot X_t + b_o) \quad (5)$$
$$h_t = o_t \times tanh(C_t) \quad (6)$$

Forget Gate，取決要忘記多少舊資料，Input Gate 則是取多少新資料從新 c_t (candidate memory) 取出，放入 C_t 成為下一次的 Memory，因此相互獨立，而 C_t 範圍超出正一到負一，需要 $tanh(C_t)$ 的 $tanh$ 進行標準化，最後相乘起來成為新的 hidden state，最後由各個參數 W 以及各個 U 決定 X_t 及 h_{t-1} 分別代表當前的輸入以及上一時間點的輸出，有了這些門的機制，LSTM 可以記住長期的資料訊息，也避免有梯度消失或爆炸的問題。而 BiLSTM 則使用在學習時間序列的關互關係，使此能夠有隱馬爾可夫模型類似的能力，為雙向循環神經網路 (Schuster & Paliwal, 1997)，通過訓練輸入閘、遺忘閘、輸出閘等權重來學習序列輸入中應該注意的權重信息，而在訓練時使用來自序列兩端的信息來估計輸出為雙向傳遞更新 (Graves & Schmidhuber, 2005)，也就是說，我們使用文字未來的字，以及過去文字的種種信息來進行預測。而我們任務中的並不是預測下一個字，而是整個句子的分析並且各個字之間帶有時間前後輸出信息向量，因此我們最佳選擇是使用 BiLSTM 完成此任務。

3.4 Conditional Random Field

條件隨機場 (conditional random field, CRF)，它經常使用於各種標籤的問題上，在此使用的是實體標籤，但不同於其他模型，其特點是狀態序列 (實體標籤序列: Y) 下觀測序列 (句子切割後序列: X) 的條件機率分布，使用 Hammersley-Clifford Theorem，損失函數為對數似然函數。基本條件隨機場的定義如下，設X與Y是隨機變數，$P(Y|X)$是在給定X的條件機率分布。如隨機變數Y構成一個由無向圖$G = (V, E)$表示的馬爾可夫隨機場，則

$$P(Y_v|X, Y_W, w \neq v) = P(Y_v|X, Y_W, w{\sim}v) \quad (7)$$

對任意頂點 v 成立，稱條件概率分佈 $P(Y|X)$ 為條件隨機場，其中 $w{\sim}v$ 表示圖 $G = (V, E)$ 中與頂點 v 有邊連接的所有頂點 w，$w \neq v$ 表示頂點 v 以外的所有頂點，Y_v 與 Y_W 為頂點 v 與 w 對應的隨機變數。

實際應用上，是使用線性條件隨機場最為廣泛，一般設 X 和 Y 有相同的圖結構，定義如下，設 $X = (X_1, X_2, \dots X_n), Y = (Y_1, Y_2, \dots Y_n)$ 均為線性表示的隨機變數序列，若再給定隨機變數序列 X 的條件下，隨機變數序列 Y 的條件機率分布 $P(Y|X)$ 構成條件隨機場，即滿足馬爾可夫性。

$$P(Y_i|X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1}) \quad (8)$$

而稱 $P(Y|X)$ 是線性條件隨機場，其中 $i = 1,2, \dots, n$，在 $i = 1$ 和 n 時只考慮單邊。且將隨機變數 X 取值為 x 的條件下，隨機變數 Y 取值為 y 的條件機率具有以下形式。

$$P(y|x) = \frac{1}{Z(x)} \exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i)\Big) \quad (9)$$

上式表示輸入序列 x，對輸出序列 y 預測的條件概率，其中 $Z(x)$ 為為歸一化因子，t_k、s_l 是特徵函數，也是二值函數，函數值為 0 或者 1。

$$Z(x) = \sum_y \exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i)\Big) \quad (10)$$

換句話說，滿足特徵條件取值為 1，否則為 0，t_k 是定義在邊上的特徵函數，稱為轉移特徵，依賴於當前和前一個位置。

$$t_k(y_{i-1}, y_i, x, i) \begin{cases} 1, & condition \\ 0, & otherwise \end{cases} \quad (11)$$

另一個 s_l，是定義在節點的特徵函數，稱為狀態特徵，依賴於當前位置:

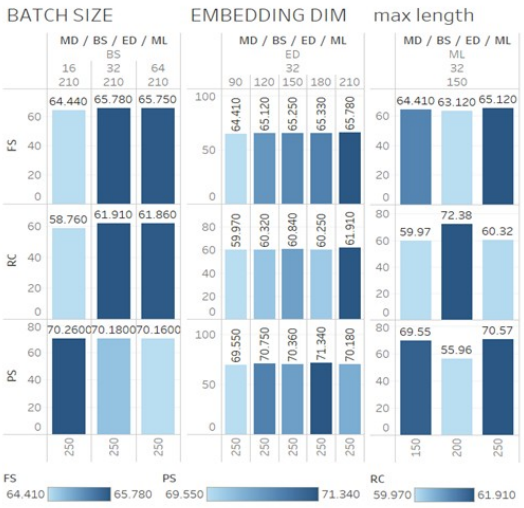

圖 1：批量大小、維度和長度對模型的評估

$$s_l(y_i, x, i)\begin{cases}1, & condition\\ 0, & otherwise\end{cases} \quad (12)$$

λ_k, u_l 為對應權重，接下來將轉移特徵和狀態特徵結合成，使用對數似然函數修正，用 Viterbi 學習算法取得最佳結果。

4 Experimental Result

此次競賽中，大會允許提交三個最佳的預測結果。我們在以下各小節分別說明三次提交(RUNS)採用的方法與相關參數設置。

4.1 BiLSTM+CRF (RUN 1)

RUN 1 採用目前在英語 NER 表現良好的 BiLSTM+CRF 網路模型。我們採用 Pytorch 的 embedding 轉換詞向量機制來對每個中文字進行向量編碼。參數設置上從圖 1 中，可以看到 batch size，在其他參數固定下，所設為 32 值的 F1 Score 以及 Recall 都比其餘兩者高，因此選擇 32 值作為 embedding dim 和 max length 的實驗固定參數。接著，看到 embedding dim，在 batch size 設為 32 值，max length 參數固定不變下，所設為 210 值的 F1 Score 以及 Recall 都比其餘四者高，因此選擇 210 值作為 max length 的實驗固定參數。最後，看到 max length，在 batch size 設為 32 值和 embedding dim 設為 210 值下，所設為 250 值的 F1 Score 以及 Precision 都比其餘兩者高，因此 max length 設為 250 值為最終實驗模型選擇參數值。所以要得到最優的模型，參數 max length 設為 250 值，batch size 設為 32 值 embedding dim 設為 210 值。實驗結果採用大會最終提供的測試集進行衡量如下表 1 中的 BiLSTM+CRF (RUN 1)所示，準確性(Accuracy) 82.23%、精確度(Precision) 55.96%、招回率(Recall) 72.38%與 F1 score 63.12%。

4.2 RoBERTa+BiLSTM + CRF (RUN 2)

RUN 2 採用 RoBERTa+BiLSTM + CRF 模型來進行實驗。我們分別選取句子長度以及批量大小來決定哪個模型可以訓練出較好的正確率，而句子長度分別使用長度為 150、200、250 個字，批量大小為 16、32、64 分別做為模型訓練。最後我們的模型使用 SGD 隨機梯度下降，學習率為 0.012，weight decay 為 1e-5，且利用 scheduler 每兩次 epoch 時學習率減少

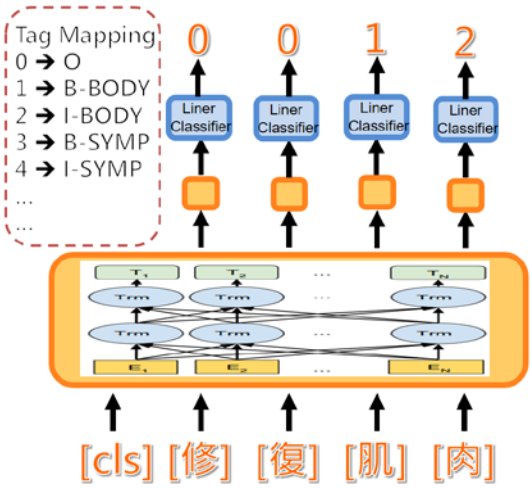

圖 2：BERT Token Classifier

0.9。實驗結果採用大會最終提供的測試集進行衡量如下表 1 中的 RoBERTa+BiLSTM + CRF (RUN 2)所示，準確性(Accuracy) 91.56%、精確度(Precision) 78.96%、招回率(Recall) 78.21%與 F1 score 78.58%。

4.3 BERT Token Classifier (RUN 3)

相對於文本分類的 BERT 和用於解決 NER 問題的 BERT，其做法上區別在於我們如何設置模型的輸出。如圖 2 所示，對於文本分類問題，我們僅使用來自特殊 [CLS] 標記的嵌入向量輸出。而 BERT 用於 NER 任務，我們需要使用所有標記的嵌入向量輸出。通過使用所有標記的嵌入向量輸出，我們可以對每個標記進行分類來預測每個標記的命名實體為何。

模型/效能	Accuracy	precision	recall	F1
BiLSTM+CRF (RUN 1)	82.23%	55.96%	72.38%	63.12%
RoBERTa+BiLSTM + CRF (RUN 2)	91.56%	78.96%	78.21%	78.58%
BERT_Based Token Classifier	91.75%	79.35%	76.24%	77.77%
BERT_Cont Token Classifier (RUN 3)	93.10%	80.18%	78.30%	79.23%

表 1: 實驗結果

RUN3 使用中研院中文計算語言研究小(Chinese Knowledge and Information Processing, CKIP) 所發布的 BERT 繁體中文預訓練模型(ckiplab/bert-base-chinese) (Yang and Ma, 2021)，對每句訓練語句的每個標記 token 產生 768 維的輸出向量。再將輸出向量接入一個線性分類器進行分類。然而在將這些文本輸入模型之前，我們需要先進行預處理。 也就是對這些輸入文字進行轉換為預訓練詞彙表中的相應 ID 並添加一些特殊的標記於句子前後([CLS] 和 [SEP])。再將每個句子填充(PAD)成同等長度的句子，我們設置訓練集中最大句子的長度 441 與 batch_size = 16 並以 adamw 為優化器進行訓練。 首先，我們以大會提供之訓練集 train.json 資料檔進行 BERT_Based 的模型訓練，並以驗證集 test.json 資料檔進行模型測試。發現到衡量指標 precision 只有 69.55%，推測應是 test.json 中包含 train.json 有未出現的新實體。借鏡吳恩達 (Andrew Ng) 近期提倡的以資料為中心的人工智慧 (Data-Centric AI)方式，持續提升資料品質能增進模型的預測能力。由於提升資料品質不是一次性能完成的任務，而是持續改進的循環過程。故我們先以 train.json 資料訓練一個基礎模型(BERT_Based) 再以預訓練模型的微調 (fine-tune) 方式加入 test.json 資料持續訓練一個模型(BERT_Cont)。最後以此模型預測大會的測試檔提交為 Run3。最後依據大會提供的標準答案(golden)所得到的實驗結果如表 1 所示。整體來說 BERT_Cont 模型表現較佳，其在準確性(Accuracy) 93.10%、精確度(Precision) 80.18%、招回率(Recall) 78.30%與 F1 score 79.23% 皆高於其它模型。

5 Conclusion and future work

在這項研究中，我們提交了三個命名實體識別的模型，並將其應用於醫療領域。根據其名稱分別為人體 (BODY)，症狀 (SYMP)，醫療器材 (INST)，檢驗 (EXAM)，化學物質 (CHEM)，疾病 (DISE)，藥品 (DRUG)，營養品 (SUPP)，治療 (TREAT)，時間 (TIME)。資料是以 BIO 格式去標記。例如"肌肉"會被標記成"B-BODY"和"I-BODY"，"咳 嗽"是 "B-SYMP"和"I-SYMP"，以此類推。類別以外的字全標為"O"。最終我們使用 HealthNER 的所有資料30,692句子當訓練與驗證資料集而以大會提供的 3204 個句子為測試資料集分別對三種模型進行驗證。實驗結果表明，RUN1 使用的是 BiLSTM+CRF 網路模型其效能最差。RUN2 採用的是簡體中文模型的 RoBERTa+BiLSTM + CRF 就能取得不錯的實驗結果。而 RUN3 採用 CKIP 繁體中文的 BERT Classifier 系統取得了最好的系統效能，其準確性(Accuracy) 93.10%、精確度(Precision) 80.18%、招回率(Recall) 78.30%與 F1 score 79.23% 皆高於其它模型。由此可知，預訓練模型的方法在此實驗上有比過去表現良好的 BiLSTM+CRF 網路模型擁有較佳的效能表現。未來我們可再針對繁體中文的 BERT 模型再加上 CRF 來探討效能是否能再提升。

References

Street Jr, Richard L. 2013. How clinician–patient communication contributes to health improvement: modeling pathways from talk to outcome. *Patient education and counseling*. 92(3): 286-291. https://doi.org/10.1016/j.pec.2013.05.004.

Weiner, Jonathan P. 2012. Doctor-patient communication in the e-health era. *Israel journal of*

health policy research. 1(33): 1-7. https://doi.org/10.1186/2045-4015-1-33.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh A., Karthikesalingam A., Kornblith S., T. Chen, N. Vivek and Norouzi, M. 2021. Big self-supervised models advance medical image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 3478-3488.

Konam, S., and Rao S. 2021. Abridge: A Mission Driven Approach to Machine Learning for Healthcare Conversation. *Journal of Commercial Biotechnology*. 26(2): 62-66.

Patanwala, A. E., Sanders, A. B., Thomas, M. C., Acquisto, N. M., Weant, K. A., Baker, S. N., Merritt, E., and Erstad, B. L. 2012. A prospective, multicenter study of pharmacist activities resulting in medication error interception in the emergency department. *Annals of emergency medicine*. 59(5): 369-373. https://doi.org/10.1016/j.annemergmed.2011.11.013.

Huang, Z., Xu, W., and Yu, K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint* arXiv:1508.01991. https://doi.org/10.48550/arXiv.1508.01991

Lee, L. H., & Lu, Y. (2021). Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition. IEEE Journal of Biomedical and Health Informatics, 25(7), 2801-2810. https://doi.org/10.1109/JBHI.2020.3048700.

Lee, L.-H., Chen, C.-Y., Yu, L.-C., and Tseng, Y.-H. 2022. Overview of the ROCLING 2022 shared task for Chinese healthcare named entity recognition. *Proceedings of the 34th Conference on Computational Linguistics and Speech Processing*.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., Zhenzhong Lan, Z. 2020. CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint* arXiv:2004.05986. https://doi.org/10.48550/arXiv.2004.05986.

Schuster, M., and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*. 45(11): 2673-2681. https://doi.org/10.1109/78.650093.

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*. 18(5-6): 602-610. https://doi.org/10.1016/j.neunet.2005.06.042.

Yang, Mu, and Ma, W.-Y. 2021. ckiplab/ckip-transformers. https://github.com/ckiplab/ckip-transformers