

基於 StepGAN 架構之語法錯誤句自動偵測與校正方法

(一) 摘要

隨著中文學習者人口快速增加，許多數位學習平台對自動化語法偏誤偵測與校正系統的需求增加迅速。近年來有許多研究以 RNN(recurrent neural network)深度學習模型取代由人工建立規則的句法診斷系統。不過因深度學習技術需要龐大的訓練資料，而模型的精確度與訓練資料量成正比，因此導致模型常因訓練資料不足而有模型效果不佳的問題。本計劃嘗試使用生成對抗網路當作語法偵錯系統，藉由模型中的生成器(generator)來改善訓練資料量龐大的難處。此模型訓練完畢後，輸入錯誤句至生成器，生成器輸出該句子的修正句，而鑑別器(discriminator)則輸出該文句是否正確。為了提升模型的偵錯效果，本計劃使用預訓練模型 BERT，藉由 BERT 已訓練的語意空間來強化模型對句法錯誤的認知。此計劃以新聞文章訓練此模型，並以 TOCFL 學習者語料庫及中文寫作偏誤語料庫(CWC)作為驗證資料，分析模型的有效性。

(二) 研究動機與研究問題

中文發展至今已有一千多年的歷史，隨著時間的推移，華人人口不斷的擴增，華人在全球市場也逐漸佔有舉足輕重的地位。在台灣，因應東協國家崛起，為了擴大與東南亞市場的聯繫，於西元 2016 年開始實施「新南向政策」，促進和新南向國家的經濟、文化等各層面的連結。在教育方面，為解決新住民及雙語第二代華語溝通問題，國內開設華語教學師資培訓班，以加快新住民對台灣環境的適應；各大學近年來也積極到東協各國設立分校，增加學生來源，也增加國際競爭力。由於華語使用者數量居世界第二，華語學習在世界各國皆有相當龐大的市場，而種種華語學習上的難處也因此突顯。

華語為二語的學習者最常遭遇語言遷移(language transfer)的問題。語言遷移是指二語學習者在使用非母語語言時，藉助於母語的結構規則或習慣，而造成語法錯誤的現象。如同我們在學習英語時，常犯的台式英語錯誤，以關燈為例，我們會把「Turn off the light.」說成「Close the light.」的誤代用法。這是因為受到母語影響，我們會直觀地依照字面上翻譯，而未使用正確的英文用法。而類似情況也反映在華語學習者上。李正民(2009)指出，由於韓語將時間名詞置於動詞之前，韓國學生在初學華語時，常犯錯序的句法錯誤。例如會將「你還沒把書還給我」說成「你把書還沒還給我」。以上的語法錯誤不僅無法確切的表達原意，更有可能會造成誤解。

隨著科技不斷地與時俱進，電腦軟體的快速發展，使語言學習的媒介更加多

元；再者，有賴深度學習及電腦硬體的技術突破，目前市場上的應用程式也都強調人機互動，能即時的與使用者做正確的互動。例如機器翻譯軟體能同步對輸入的字句翻譯成另一語言的句子；另外有些文書處理軟體也提供對輸入的英文語句做語法偵測，並給予適當的語法修正建議。但是雖然這類軟體在英文文法自動修正上有良好的表現，但在華語上效果卻不如預期，甚至不支援華語修正。

在這短短五年內，中文文法錯誤診斷(Chinese grammatical error diagnosis, CGED)系統使用的方法及技術有了相當大的改變，Chang, Sung & Hong (2015) 提到過去部份英文文法診斷，是利用建立句法剖析樹來判斷語法正確與否，若該英文句無法完整解構並轉換為一個句法樹，則表示該句子存在句法錯誤。但因為中文語法的高延展性，導致句子的變化性多，也造成句法剖析難以套用在中文句法偵測上。所以 Chang et al. (2015) 提出由偏誤語料庫為基礎，以機器和人工規則來診斷中文語法的 KNGED 系統。在此系統中，人工規則為語言學專家透過語料觀察建立的，而機器規則是經由大量統計從語料自動建立規則。不過，因為中文學習者所造成的語法錯誤較為複雜，所以機器規則沒辦法辨識大多數的錯序句法，使得精確率(precision)與召回率(recall)不盡理想，需要另外建立大量的人工規則來修正上述缺點。

而近兩年隨著深度學習模型的成功發展，也有相當多的研究以深度學習模型偵測文法錯誤。例如 Lee, Lin, Yu & Tseng (2017) 建立基於 CNN-LSTM 的系統，透過詞嵌入(word embedding)把一詞彙或句子轉換成詞向量，再將這個多維的詞向量以 CNN 為過濾器進行卷積運算，形成大小不一的特徵圖(feature map)。再經由最大池化(max pooling)化為句子表示(sentence representation)並輸入給 LSTM，最後系統則會輸出 0 或 1 來表示句法是否正確。但因為此模型需要龐大的訓練資料來提升精確率，故這些模型的效能仍有改進的空間。然而增加由人工標記的訓練資料取得不易。

為了解決上述問題，Goodfellow et al. (2014)提出的生成對抗網路(generative adversarial network, GAN)是一個可能的解決方法。GAN 原是應用在影像處理的模型，近年來也被嘗試運用在自然語言處理上。GAN 的發展不僅讓深度學習在影像處理方面有的重大的改變外，也使自然語言處理模型可運用 GAN 自行產生樣本的優點，來彌補訓練資料不足的缺失。因此本計劃將運用一種稱為 StepGAN 的模型作為基礎，利用現有的新聞語料以及偏誤語料庫產生大量的訓練資料，藉由 GAN 的原理產生一個能輸入一個句子就能判斷其語法正確性以及可能的修正句。

而經實驗後發現，以 StepGAN 生成中文偏誤句效果仍不盡理想。為了改善模型的偵測效果，我們以 Devlin, Chang, Lee, & Toutanova 提出的

BERT(Bidirectional Encoder Representations from Transformers, 2018) 模型為基礎來改善偵測效果，使其能夠偵測帶有多餘、誤代及錯別字的中文偏誤句。

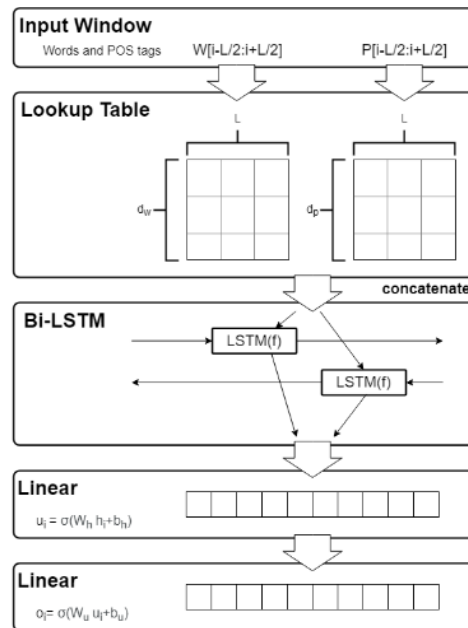
(三) 文獻回顧與探討

本節將回顧與探討與本計畫相關的文獻。3.1 節回顧過去中文句法錯誤偵測的研究；3.2 節說明 GAN 在應用於文法相關議題的衍生模型。

3.1 中文句法偵錯

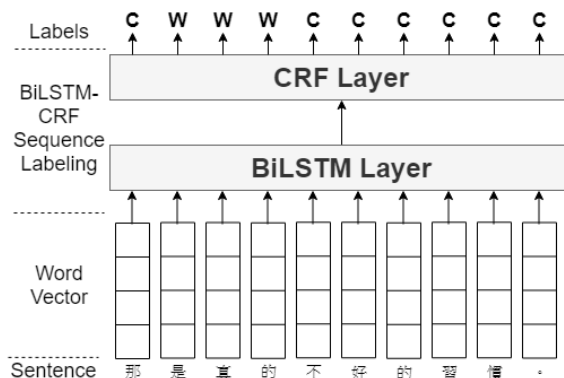
中文句法偵測主要分統計式及深度學習兩種偵測模型，統計模型利用數據及一套數學方程式來建立規則，使機器能透過此規則來判斷和預測句法是否有誤。統計模型除了前述 Chang et al. (2015) 提出的 KNGED 系統外，另外還有 CRF(conditional random field)模型，如 Yeh, J. F., Yeh, C. K., Yu, Lin & Tsai (2015) 提出的偵錯法是將中文句透過前處理產生斷詞及詞性標記，輸入給 CRF 模型，若出現句法錯誤，則會在該錯誤字詞序列上標記錯誤類型。CRF 使用條件機率計算整個句子中，各字詞的機率值，表示該字詞在句子中的位置是否正確，若有字詞的機率值和其他字詞相比差異甚大，則代表發生句法錯誤，並在該錯誤字詞標記上 F(False)字符，否則標記 T(true)。雖然統計模型構造簡單、運算快，但此種模型較重視輸入數據與輸出結果的關係，在處理某些文本時仍有侷限。如中文的語法，前後的語義往往會影響整個句子的原意，即使 CRF 利用條件機率擷取局部特徵，在較短的文句上有較佳的精確度，但卻無法參考前後文意，在長文句中也召回率也因此下降。

而深度學習模型的提出與發展，解決了統計模型無法參考前後文意的問題，例如 Lee et al. (2017)提出的 CNN 與 LSTM 結合的句法偵錯模型；Huang & Wang (2016) 也提出以雙向 LSTM 模型來做句法偵錯，如圖一所示，將字詞及詞性標記輸入給雙向 LSTM 訓練，並輸出該句法錯誤之機率。



圖一、雙向 LSTM 架構圖 (Huang & Wang, 2016)

Liu, Zan, Zhong & Ma (2018) 結合了雙向 LSTM 與 CRF，提出了雙向 LSTM-CRF 模型，如圖二所示，期望能利用 LSTM 模型讀取前後文意、CRF 模型擷取文本特徵的優勢來使句法錯誤偵測更加精確。與 Huang & Wang (2016) 的架構相比，此方法在網路最上層將模型輸出之序列輸入 CRF 層做錯誤標記，而輸入部分則只將單字詞輸入模型做訓練，並未考慮詞性標記。而研究結果顯示，雖然此堆疊模型召回率提高了，精確度卻下降了。雖然 LSTM-CRF 之精確度下降了，但因 CRF 不需做模型訓練也能與 CNN 做到特徵擷取的功能，故此方法為業界及研究中常見的方式，如 NLPTEA-2018 Share Task(Rao, Gong, Zhang & Xun, 2018) 中，大多以 LSTM-CRF 為基底的模型來進行語法偵錯與校正，其召回率和精確度相較於單純的統計規則式模型高出許多，但是仍有很大的進步空間。



圖二、雙向 LSTM-CRF 架構圖 (Liu et al., 2018)

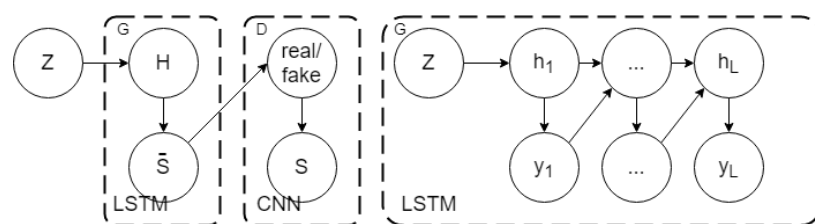
3.2 生成對抗網路

生成對抗網路原是進行影像生成的模型，由生成器(generator)及鑑別器(discriminator)這二個類神經網路組成。生成器產生以假亂真的假資料，並試圖騙過鑑別器；鑑別器也為了不被生成器欺騙，加強辨別資料的真偽，並為辨別的資

料評分。若模型訓練完畢，便可得到會自行生成圖片得生成器，也會獲得能自行辨別圖像優劣的鑑別器。而我們也可利用生成器產生假資料的特性，來彌補訓練時真實資料的不足。

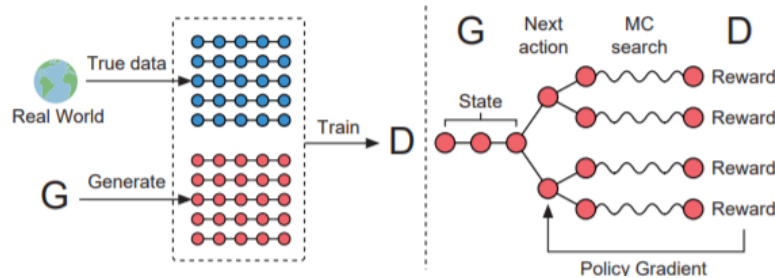
由於圖像是由一群連續的數值組成，故 GAN 在影像生成上有卓越的發展，然而在離散的語言生成上是一個艱鉅的挑戰。因為生成器需要從鑑別器得到之梯度進行訓練，所以整個網路輸入及輸出的資料必須是可微分的，但是字詞是離散的，亦即不可透過簡單的數值運算去獲得另一組詞彙，若直接將未處理的文字序列輸入給 GAN 網路，則鑑別器輸出的數值將不具任何意義。

為了解決離散資料不可微分的問題，Zhang, Gan, & Carin (2016) 提出了 textGAN 的架構。圖三為 textGAN 架構圖，左圖為 textGAN 的架構，由 LSTM 作為生成器、CNN 作為鑑別器，右圖為生成器的架構。此模型採用了特徵匹配 (feature matching) 的方法作為目標函式，將生成的假資料之特徵圖與真實資料進行比對。生成器產生的詞彙透過詞嵌入轉換為一 N 維度的向量，並以 soft-argmax 函式輸入給鑑別器做訓練。不過因為 LSTM 需訓練的參數，相較於原始 GAN 中的網路參數要來的多，所以還是擁有難以訓練的困擾。



圖三、textGAN 架構圖 (Zhang et al., 2016)

由於 textGAN 未能有效處理離散文本，Yu, L., Zhang, Wang & Yu, Y.(2017) 提出了 SeqGAN 的文本生成模型，生成器、鑑別器分別為 LSTM、CNN 架構。與 textGAN 模型不同之處是，此模型以強化學習(reinforcement learning, RL)做為參數更新的方法。圖四為 SeqGAN 架構圖，先用最大似然估計(maximum likelihood estimation, MLE)對鑑別器作預訓練，使其能更有效率地調整生成器的參數，左邊先以真實資料及生成器生成的雜訊來訓練鑑別器，右邊則是將部份已輸出的序列，以蒙特卡羅搜尋(monte carlo search)取樣成完整序列，再交由鑑別器評分。最後以鑑別器的輸出當作獎勵，用策略梯度(policy gradient)的方式更新生成器參數。如此便可解決原始 GAN 架構無法對部份已生成序列評分的問題，但因為在訓練初期難以得到實際獎勵，亦即鑑別器輸出為零，所以在訓練模型時會有稀疏獎勵(sprase reward)的問題。



圖四、SeqGAN 架構圖 (Yu et al., 2017)

接著 Fedus, Goodfellow & Dai (2018) 提出了 MaskGAN 的生成模型，生成器和鑑別器皆為 Seq2Seq (sequence to sequence) 的架構。在資料前處理的部分除了詞嵌入外，還利用了完形填空(in-filling)，將原序列某一特定位置替換成遮蔽字符 (Masked token)，並輸入給生成器，且嘗試為缺空填補相同或語意通順之字符，使其成為一完整序列。最後，鑑別器讀入生成器生成之序列，並且為已填補上的字符評分。參數更新部分也由策略梯度取代為動作評判(actor-critic)。以上的更動能降低梯度逐漸變小，造成模式崩潰(mode collapse)，即生成樣本多樣性不足的問題，但在輸入某些資料時仍有此問題存在。

Tuan & Lee (2019) 改良了 SeqGAN 的架構，發展出了 StepGAN，採用條件序列的方法，輸入一問句，則會生成輸入問題的答覆句。模型訓練則採用了蒙特卡羅法中的 Q 學習(Q-Learning)，將生成序列藉由逐步更新來計算 Q 值並取平均，最後將平均值交由生成器作梯度更新。由於此架構很適合句子錯誤校正的問題，本計劃將採用此架構作為訓練偵測與校正模型的方法。

綜合以上所述，時至今日，GAN 在文本處理及生成問題上也有了相當的進展，但很少有研究將 GAN 專注在文法偵錯上，甚至是中文文本處理上。此外，Arjovsky, Chintala & Bottou (2017)表示，GAN 在發展至今的幾年間，難以訓練甚至更新時梯度趨於崩潰的缺失還是有效克服。

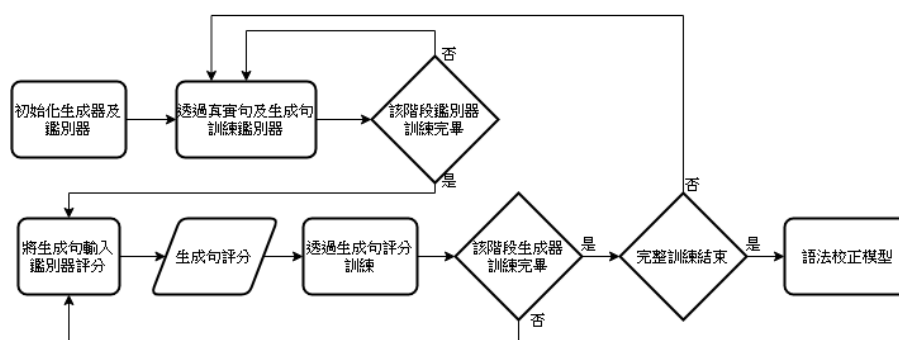
(四) 研究方法及步驟

本節將討論本研究使用中文句法錯誤偵測及修正方法。4.1 節說明使用 StepGAN 進行句法偵測及修正之步驟；4.2 節描述如何使用 BERT 為基礎的模型作中文錯別字偵錯。

4.1 中文錯誤句法校正

圖五為本計劃之模型訓練流程圖。此流程是基於前述提到的 StepGAN 架構。在初始化鑑別器與生成器後，鑑別器會讀入由新聞語料庫取得的句子(以下簡稱真實句)和生成器生成的修正句(以下簡稱生成句)練習分辨兩者，待鑑別器達到設定的辨識能力後，就結束這個階段的鑑別器訓練，然後進行此階段的生成器訓練。生成器訓練時，會將生成句輸入至鑑別器評分，再將此評分結果作為調整生成器的依據。待生成器達到設定的修正能力後，此一階段即結束，然後進入新一階段

進行同樣的鑑別器與生成器訓練。如此迭代數個階段後，理想上生成器就會將錯誤句修正成正確句。



圖五、模型訓練流程圖

本計劃所使用之基於 StepGAN 之句法偵測與校正錯誤模型將分三個小節說明。4.1.1 節先介紹訓練資料及資料的前處理；4.1.2 節則是講解 StepGAN 模型相關參數及架構；4.1.3 節將說明如何測試及驗證此模型。

4.1.1 訓練語料產生及處理

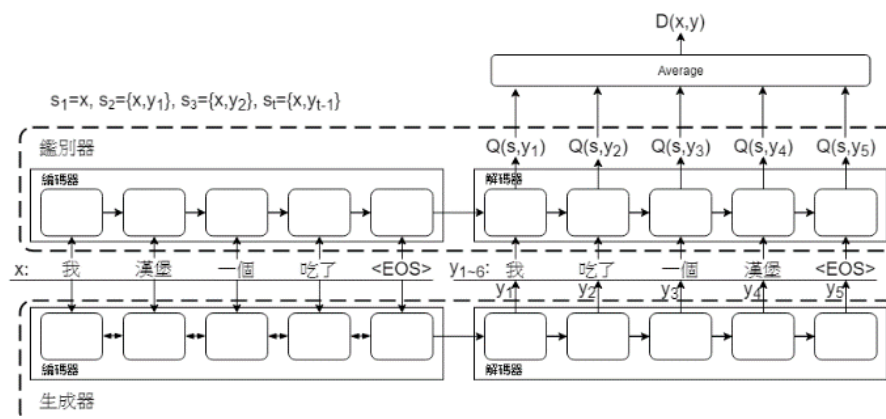
句法錯誤大多分為多餘、遺漏、錯序及誤代四種錯誤類型(Dulay, Burt, & Krashen, 1982; James, 1998)。本計劃將利用聯合報八年新聞文本以及 Gigaword (Ma & Huang, 2006)來產生訓練語料，將隨機抽取數百萬句直接作為真實句，另外再隨機抽取數百萬句(以下簡稱樣本句)經由演算法轉換為有句法錯誤的句子(以下簡稱錯誤句)。最後我們將語料分為「遺漏偏誤語料」及「多餘、錯序及誤代偏誤語料」分開訓練模型。且為了使模型學習不用修改正確無誤的句子(以下簡稱正常句)，我們在真實語料中額外抽取少量的真實句當作正常句加入訓練語料中。

此演算法產生四種錯誤句的方法如下：對於模擬發生多餘的錯誤句(以下簡稱多餘句)，我們先統計 TOCFL 學習者語料庫及中文寫作偏誤語料庫(以下簡稱錯誤語料)中，多餘錯誤次數較多的多餘詞，並隨機安插至部份樣本句中。對於剩下樣本句則隨機插入一個詞成為多餘句。對於模擬發生遺漏及誤代的錯誤句(以下簡稱遺漏句、誤代句)的建立，也比照多餘句產生流程。對於模擬發生錯序的錯誤句(以下簡稱錯序句)，因發生的原因與詞較無關聯，故將每個樣本句中隨機挑選兩個詞互換成為錯序句。所有錯誤句產生後，會再經過 WECAn (Chang, Sung & Lee, 2012)斷詞後，取得錯誤句的字詞序列。

4.1.2 StepGAN 模型架構

圖六為 Tuan & Lee (2019)提出的 StepGAN 架構圖。原設計為輸入問句 x ，則生成器會生成一答覆句 y ，將 x 與 y 的各字詞視為狀態 s 。鑑別器則會依據 s 和 y 給予代表兩者相關性的 Q 值 $Q(s,y)$ 。本計劃使用相同的架構與流程，但是將

x 和 y 改設為以下定義：在生成器訓練階段，當 x 為錯誤句或真實句時、生成器的 y 為生成句；在鑑別器訓練階段，若為生成器產生的訓練資料， x 和 y 的定義同前；若為原始真實句作為訓練資料，則 x 和 y 都為同一真實句。

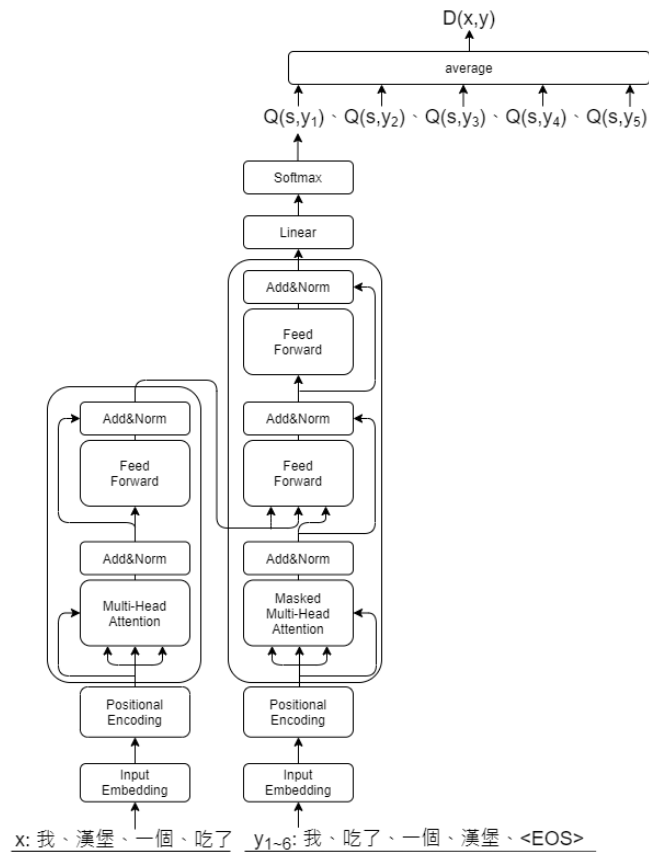


圖六、StepGAN 架構圖

4.1.2.1 鑑別器

Tuan & Lee(2019)提到在 StepGAN 中，可使用任何的 Seq2Seq 模型來當作鑑別器。本計劃將使用 Vaswani et al.(2017)提出的 transformer 架構，來解決 RNN 只能時序輸入無法平行化處理的問題。transformer 為 Seq2Seq 與自注意力機制 (self-attention) 的疊加模型，如圖七所示，左圖為編碼器結構，輸入為錯誤句 x ，文本序列會經過嵌入層轉為語意向量序列，透過多端注意力機制 (multi-head attention) 將向量序列平行化處理，並輸出另一組序列，之後再將輸入及多端注意力的輸出序列進行疊加與正規化 (add & norm)，最後經由前傳遞 (feed forward) 會獲得編碼向量。

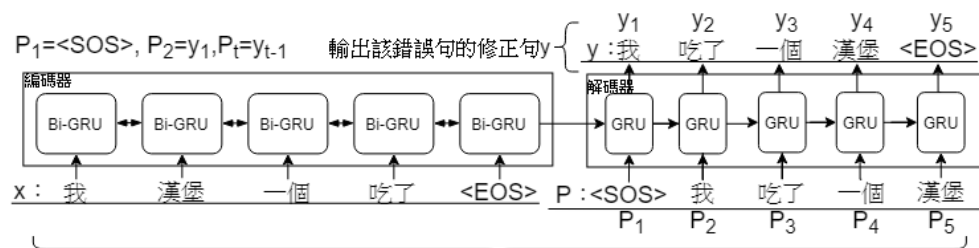
為了要檢視生成句是否正確，我們將生成句 y 和經過編碼器運算的 x 作為解碼器的輸入，使解碼器計算 Q 值。右圖為解碼器，輸入序列會經過二層的注意力運算，其中第二層則會與編碼器輸出的編碼向量一同運算，經由前傳遞則會獲得下一時序 (time step) 的 Q 值 $Q(x, y_1)$ ，並將下一字詞「吃了」與 x 繼續運算可獲得 $Q(x, y_2)$ 。編碼及解碼器在如此運作下分別執行數次，直到讀取到截止標籤 <EOS> 或句號，即可得到一完整 Q 值序列。最後，將所獲得的全部 Q 值取平均，即可得到對真實或生成文本的評分 $D(x, y)$ 。



圖七、Transformer 架構圖(Vaswani et al., 2017)

4.1.2.2 生成器

StepGAN 之生成器為基於 RNN 的 Seq2Seq 模型。本計劃將生成器以基於 GRU 的 Seq2Seq 架構來取代。生成器之編碼器及解碼器第一層為嵌入層，將輸入序列轉換為語意向量，其中編碼器部分改用雙向 GRU 來增強對中文句上下文意的解析。圖八為生成器架構圖，輸入一錯誤句 x ，並將此序列輸入至編碼器，將會輸出一狀態向量。將狀態向量及字串起始符號輸入至解碼器後，則會依序輸出各個字詞。輸出則會獲得符合錯誤句的生成句 y 。在訓練過程中，生成器會接收到鑑別器的評分 $D(x,y)$ ，並以此評分作梯度更新訓練，以加強生成器的修正能力。



輸入一錯誤句 x 至編碼器，<EOS>為句子結束
將前一字詞 P 輸入解碼器，<SOS>為句子開始

圖八、生成器架構圖

4.1.3 模型驗證及測試

訓練及測試的資料都經過 WECA 斷詞處理。本計劃將以寫作偏誤語料庫和 TOCFL 學習者語料庫(Lee, Chang & Tseng, 2016)作為測試資料集，測試生成對抗網路模型。輸入各種不同的錯誤句，檢測是否能產生相對應的生成句，也能鑑別生成句的好壞。對於系統的校正能力，我們分析系統的精確率、召回率及 F1-measure。各指標的計算方式如下：

精確率(Precision)： $TP / (TP+FP)$

召回率(Recall)： $TP / (TP+FN)$

F1： $2 * Precision * Recall / (Precision + Recall)$ 。

其中

TP：有句法錯誤所以進行校正的錯誤句數量。

TN：沒有句法錯誤所以不進行校正的正常句數量。

FP：有句法錯誤不進行校正的錯誤句數量。

FN：沒有句法錯誤卻進行校正的正常句數量。

4.2 基於 BERT 的中文錯誤句法偵錯

華語學習者在學習中文常犯的句法錯誤中，以誤代的別字錯誤佔了相當大的比例。為了讓句法偵錯效果有所提升，本計劃提出了以 BERT 為基礎的偵錯模型。4.2.1 節介紹如何產生別字訓練語料；4.2.2 節則是講解如何以 BERT 模型應用於別字偵錯中；4.2.3 節將說明如何測試及驗證此模型。

4.2.1 訓練語料產生

為使模型能夠更了解中文文句的語意，並透過上下文來正確偵測偏誤句中的誤代詞，須利用資料量大且真實的語料訓練。然而，就目前已知的真實資料集而言，資料量仍有所不足，難以涵蓋真實可能會遇到的各種情況。

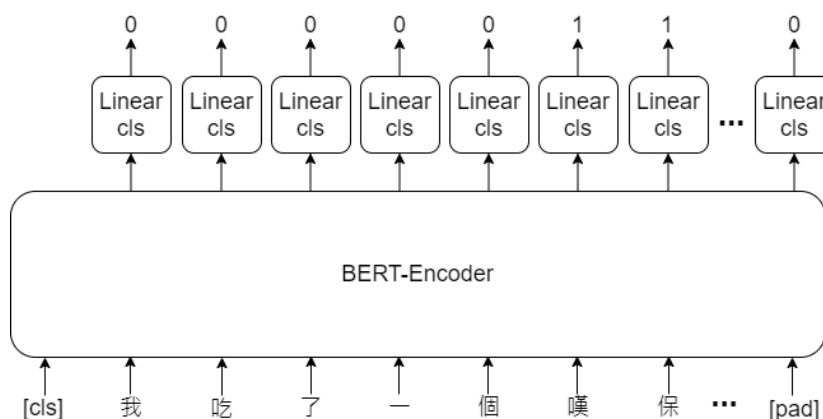
根據上述需求，本研究透過大量的新聞語料，計算中文字元間的字音與字形相似值，並依據相似值大小來隨機抽取正確句中的字，並將其代換為最為相似的詞。為了確保誤代語料不被少數高頻字詞所涵蓋，每個字詞被選中的機率有所限制。若某字詞的偏誤句數量已達上限，則該字詞將不會被挑選。

4.2.2 模型架構

為了能夠使模型更精準預測中文句法錯誤，我們使用 Devlin, Chang, Lee, & Toutanova (2019) 提出的 BERT for Single Sentence Tagging Task (以下簡稱 BSST)進行句法偵錯。BSST 原先用來解決的任務之一為命名實體識別(Name Entity Recognition, NER)。舉例來說，我們的目標是將句子中的詞分成 4 類，分別是主詞詞首、動詞詞首、量詞詞首、量詞詞幹、名詞詞首、名詞詞幹及其他詞。假設給 BSST 輸入一中文句「我吃了一個漢堡」，則 BSST 的目標依照原句依序輸出如下結果：

我 吃 了 一 個 漢 堡
主詞詞首 動詞詞首 其他詞 量詞詞首 量詞詞幹 名詞詞首 名詞詞幹

而遵循此模式，我們可以應用在句法偵測上，如圖九所示。輸入一中文句至 BERT 模型，BERT 則會將輸出的語意特徵輸入至線性分類器(Linear classifier)，這些分類器會再依據特徵向量輸出該字是否有文法錯誤，輸出 0 表正確字，輸出 1 表此字帶句法錯誤。



圖九、句法偵錯架構圖

4.2.3 模型驗證及測試

本研究將使用 SIGHAN 2015 (Tseng, Lee, Chang, & Chen, 2015) 資料集作為評估本模型的測試資料集。此資料集共有 1100 句，其中正常句及別字句各有 550 句。由於本模型不擅長處理長度超過 60 字的中文句，因此排除過長文句後，正常句及別字句數量分別為 538 句和 531 句。

測試及評估的指標皆與 SIGHAN 2015 相同，各指標的計算方式及說明如下：

誤報率(False-Alarm Rate)： $FP / (FP+TN)$

正確率(Accuracy)： $(TP+TN) / (TP+FP+TN+FN)$

精確率(Precision)： $TP / (TP+FP)$

召回率(Recall)： $TP / (TP+FN)$

F1： $2 * Precision * Recall / (Precision + Recall)$ 。

其中

TP：所有被辨識為有別字的別字句數量。

TN：所有被辨識為正常的正常句數量。

FP：沒有別字卻被辨識為有別字的正常句數量。

FN：有別字卻被辨識為正常句的別字句數量。

(五) 實驗結果

本節將討論二種模型在真實語料的測試結果。5.1 節說明 StepGAN 架構在修正句生成的效果，以及更改模型架構後，我們對效能評估的分析。5.2 節說明以 BERT 為基底的偵錯模型的測試結果，並與先前別人提出的方法做比較。

5.1 StepGAN 偵測及校正模型

我們使用二種訓練語料進行訓練，並抽出寫作偏誤語料庫及 TOCFL 學習者語料庫中的偏誤句共 1000 句進行句法校正測試，測試結果如表一所示。結果發現，即使 StepGAN 之生成器使用雙向 GRU 去加強上下文意的解析，但仍無法解析完整的文句，造成模型無法生成完整的正常句。Caccia et al. (2020)指出，在文本生成上 GAN 使用強化學習解決文字序列難以經由微分獲得更新梯度，在模型及輸出輸入間仍難以有效地定義獲得獎勵的規則及條件，造成整個對抗生成網路依然有稀疏獎勵的現象發生。

在訓練資料上，我們原先認為使用大量的隨機錯誤語料能使 StepGAN 學習到更廣泛的句法錯誤規則，但是在實驗結果上卻呈現相反的結果。經調查發現，在真實錯誤句中，各個句法錯誤往往與前後文意相關，每位華語學習者因語言遷移現象所造成的句法錯誤又會有所差異。若在訓練語料中所包含的句法錯誤種類越多，會讓模型更難以訓練。

Models	Precision	Recall	F1-Score
多餘、錯序及誤代偏誤語料訓練	0.34	0.132	0.190
遺漏語料訓練	0.37	0.156	0.219

表一、StepGAN 校正結果

5.2 BERT 偵測模型

表三比較本研究所提方法與 SIGHAN 2015 上各評估指標中效果最佳之方法。先前方法包括 NCTU&NTUT (Wang, & Liao, 2015)及 CAS (Xiong, Zhang, Zhang, Hou & Cheng, 2015)的兩個 Run，由表三可知，雖本研究提出方法之誤報率不是最低的，但其誤報率低於 0.1，避免了高誤報率的方法在實際應用上的限制。而與先前方法比較其他評估指標，本研究所提方法皆是當中最優者。這顯示本研究所提方法除了能有效降低誤報率外，同時能提升誤代錯誤的偵測率，更符合真實應用需求。

Models	accuracy	recall	precision	False-Alarm Rate	F1
本研究所提方法	0.752	0.574	0.887	0.072	0.697
CAS-Run2	0.701	0.533	0.803	0.131	0.640
CAS-Run3	0.666	0.535	0.724	0.204	0.615
NCTU&NTUT	0.606	0.262	0.837	0.051	0.399

表二、各指標中最佳之模型評估比較

(六) 未來工作

在生成任務中，生成器使用交叉熵(Cross entropy)計算誤差，交叉熵為嚴格匹配函數，且 transformer 使用 teacher-forcing 訓練，如此設計會使訓練誤差無法收斂。本計劃將以對齊演算法(alignment)為方向，修改生成器的誤差函數，以編輯距離(edit distance)計算生成器輸出與真實錯誤句的相似度。

待錯誤句生成效果有所提升，本計劃將更改生成器架構，以 BERT 當作編碼器，使用 transformer 解碼器進行解碼，再搭配修改後的誤差函數進行訓練出句法校正系統。期望能使用 BERT 大量已訓練的與法規則來修正中文錯誤。

在 BERT 進行句法偵錯部份，我們將以 GRU 網路連接注意力機制來取代線性分類器，來增強模型對輸出文意的理解。因 BERT 模型參數較多，故在參數更新時易有過適(Overfitting)的現象發生，而我們將使用集成學習(ensemble learning)，透過使用不同的 dropout value 及模型架構來改善此現象。

(七) 參考文獻

中文參考文獻

李正民。(2009)。韓籍學習者習得現代漢語否定詞「不」與「沒(有)」之研究。

英文參考文獻

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875.

Caccia, M., Caccia, L., Fedus, W., Larochelle, H., Pineau, J., & Charlin, L. (2020). Language gans falling short. arXiv preprint arXiv:1811.02549.

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012). A Chinese word segmentation and POS tagging system for readability research.

Chang, T. H., Sung, Y. T., & Hong, J. F. (2015, June). Automatically Detecting Syntactic Errors in Sentences Writing by Learners of Chinese as a Foreign Language. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 20, Number 1, June 2015-Special Issue on Chinese as a Foreign Language.

- 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.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deepbidirectional transformers for language understanding. In NAACL-HLT (1).
- Dulay, H. C., Burt, M. K., & Krashen, S. D. (1982). *Language Two*. New York: Oxford University Press
- Fedus, W., Goodfellow, I., & Dai, A. M. (2018). MaskGAN: better text generation via filling in the_. arXiv preprint arXiv:1801.07736.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).
- Huang, S., & Wang, H. (2016, December). Bi-LSTM neural networks for Chinese grammatical error diagnosis. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)* (pp. 148-154).
- James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis*. London: Addison Wesley Longman.
- Lee, L. H., Chang, L. P., & Tseng, Y. H. (2016, November). Developing learner corpus annotation for Chinese grammatical errors. In *2016 International Conference on Asian Language Processing (IALP)* (pp. 254-257). IEEE.
- LEEa, L. H., LINb, B. L., YUb, L. C., & TSENG, Y. H. (2017, January). Chinese grammatical error detection using a CNN-LSTM model. In *25th International Conference on Computers in Education, ICCE 2017* (pp. 919-921). Asia-Pacific Society for Computers in Education.
- Liu, Y., Zan, H., Zhong, M., & Ma, H. (2018, July). Detecting simultaneously Chinese grammar errors based on a BiLSTM-CRF model. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 188-193).
- Ma, W. Y., & Huang, C. R. (2006, May). Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. In *LREC* (pp. 2182-2185).
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.
- Rao, G., Gong, Q., Zhang, B., & Xun, E. (2018, July). Overview of NLPTEA-2018 share task chinese grammatical error diagnosis. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational*

Applications (pp. 42-51).

- Tseng, Y. H., Lee, L. H., Chang, L. P., & Chen, H. H. (2015, July). Introduction to sighthan 2015 bake-off for chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing* (pp. 32-37).
- Tuan, Y. L., & Lee, H. Y. (2019). Improving conditional sequence generative adversarial networks by stepwise evaluation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4), 788-798.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).
- Xiong, J., Zhang, Q., Zhang, S., Hou, J., & Cheng, X. (2015, June). HANSpeller: a unified framework for Chinese spelling correction. In *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 20, Number 1, June 2015-Special Issue on Chinese as a Foreign Language.
- Yeh, J. F., Yeh, C. K., Yu, K. H., Lin, Y. T., & Tsai, W. L. (2015, July). Condition random fields-based grammatical error detection for Chinese as second language. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 105-110).
- Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017, February). Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Zhang, Y., Gan, Z., & Carin, L. (2016). Generating text via adversarial training. In *NIPS workshop on Adversarial Training* (Vol. 21).