中文錯誤偵測與修改

專題編號：109-CSIE-S021
執行期限：108 年第 1 學期至 109 年第 1 學期
指導教授：陳彥霖
專題參與人員： 104440026　　鄭耀飛
105350023　　溫心瑀

一. 摘要

傳統的光學辨識模型會因為手寫的品質，進而影響到輸出的結果。而中文文件的掃描所產生的錯字往往會是因為文件的品質不佳所導致。大部分的錯誤都是中文字的部分部首等等漏掉了，為了解決這個問題，我們引入了深度學習開發的中文錯字偵測與修改系統。

以深度學習開發中文錯誤的偵測及修改系統。以往提出的錯別字系統在精確率(Precision)及召回率(Recall)之表現皆不甚理想。此次使用 Bidirectional Encoder Representations from Transformers（以下稱作 BERT）作為開發基礎：針對 downstream tasks 需求，利用 pre-trained 模型進行 fine-tuning，訓練出的模型可以進行字音相似的錯別字修正。

關鍵詞：Contextualized word embedding, Pre-train, Mask Language Model, Fine-tune

二. 緣由與目的

藉由中文錯誤偵測與修改系統，解決手寫辨識對於中文字的判斷錯誤。

因中文與其他語言斷詞等架構根本上的不同，難以將其他語言錯別字系統延伸使用。而已開發之中文錯別字除錯系統，常因為中文語文體系上的繁雜、無法依據上下文作為檢查參考，造成表現低落。憑藉 BERT 的特性，實現提高精確度和召回率之目的。

三. 使用技術方法

光學模型使用的是: RNN based 的 tesseract 模型。透過 RNN 可以記憶的特性產生出具有連續性質的輸出文字。

Contextualized word embedding 可以表彰不同 embedding 之間的關聯性，且重視 token 上下文關係，則不同意思的 token 就能賦予不同的 embedding；而這是以往的 Glove、Word2Vec 詞向量系統所缺乏的特性。

BERT 為多層的雙向 transformer encoder，其中包含 multi-head 的 self-attention layer，不同下游任務使用相同的 pre-trained model。

最後，上述所提到的所有模型，則是整合到了一個由 Flask 框架所開發的後端伺服器，前端是一個由 Swift 所構成的 iOS 專案，兩者透過 RESTful api 溝通。

為了實作文件掃描錯字偵測與修正，專案使用 Google Cloud Platform 之 Cloud Vision API 光學字元辨識(OCR)，通過 REST api 作為溝通，OCR 讀取使用者的選取圖像，回傳文本檔案，再由修正錯字模型後端接收，最後將輸出的已修正文本傳回使用者端。

四、開發進行方式

中文錯字系統的開發可分成兩個步驟：pre-train 及 fine-tune。

(一) Pre-train：

過程中使用 Mask Language Model (MLM)及 Next Sentence Prediction (NSP)。其中 MLM 打破了以往單向 representation model 原先只能考量先前涵蓋範圍 token 的限制，真正達成參考上下文的特性。BERT 在此步驟使用大量的未標註資料訓練模型參數，以供 fine-tune 步驟接應，針對下游任務繼續微調。

(1) MLM：
隨機遮蓋一定比例的 tokens，再依照上下文預測原先的內容。

(2) NSP：
預測一組語句的關係，label 結果為相鄰關係或者無關係。

(二) Fine-tune：

以 MLM 的方式遮蓋 training data 各個句子含有錯別字的 token 位置，並以正確字詞作為相應 label，進行 fine-tune 訓練。

LSTM based 的 Tesseract 4.0 在原本 model 下的繁體中文及簡體中文已有一定準確度，但符號、部分生僻字的辨識能力並不完備，可以基於 Tesseract 中訓練完畢的語言模型，使用擴充 character set 及使用不同字體的方式得到的新 dataset 進行訓練。

五、預期成果

中文錯字偵測與修改的模型在 F1 score 預期達到、甚至勝過 state-of-the-art performance。

在整體辨識的過程，必須對於基本沒有損害的文件上有基本的辨識能力，但是對於有損害的文件，可以透過上述所提到的中文錯字偵測與修改模型找出錯誤，並且及時修正。

五、使用資料

SIGHAN: Chinese Spelling Check Task —— 作為 dataset

參考文獻

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”

[2] Ray Smith, An Overview of the Tesseract OCR Engine