Using Bi-LSTM Model and BERT-based Model for Traditional Chinese Punctuation Restoration

利用BERT based Model與Bi-LSTM Model實現繁體中文自動化標點標註

專業領域：資工領域

組別：A202

指導教授：李祈均教授

組員姓名：鍾永桓、陳敬涵、洪子堯

Abstract

隨著串流影音的蓬勃發展，自動化語音辨識(ASR)的技術以被大量投入使用，藉此節省人力標註字幕的浪費，然而自動化語音辨識很嚴重的問題在於無法自動化的產生標點，標點對於一個語句的可讀性影響極大。缺乏標點的文本會使後續的閱讀和翻譯無法進行，所以自動化標點標註便成為重要的課題，在英文已經有許多學者投入這方面的研究並有許多成果，但繁體中文在這個領域的研究就明顯落後。基於以上情況我們希望可以投入這方面的研究以建構出繁體中文專屬得自動標點標註系統。

由於要自動語音辨識系統所產生的資料集需要經過人為標註才可運用於模型訓練，但人為標註成本過高，於是我們嘗試從各種新聞和中文維基百科蒐集資料，並對其作「簡體轉繁體」、「替換標點符號」（按照特定規則）、「斷詞」、「建立對應字典」和「添加錯漏字」等預處理，希望藉此模擬出ASR系統所產生的資料，並利用這些資料訓練三種不同的模型架構(Bi-LSTM(with attention)、BERT、BERT+Bi-LSTM)，觀察訓練出的model是否可以根據前後文的詞彙分布找出適當得標註位置，最終我們在嘗試將model運用於實際場合的資料集，並分析可能影響其表現的因素。

Introduction

一. 資料處理

我們採用了三種資料集(dataset)，分別為「wikipedia繁體中文dataset」、「news dataset」和「asr dataset」。並透過進行「簡體轉繁體」、「替換標點符號」（按照特定規則）、「斷詞」、「建立對應字典」等處理步驟，使資料集可用來訓練、測試模型。

「簡體轉繁體」是透過「opencc api」將「wikipedia中文dataset」混雜的簡體字轉換為繁體；替換標點符號是將

「：」、「；」、「、」替換為「，」；「！」、「.」替換為「。」

由於我們的model是四種class(，。？<none>)的classfier，故我們只要保留三種標點符號(，。？)，而且只要這三種標點就足以大幅提高文章的可讀性。

「斷詞」是將文本切割成以詞彙為單位而非以字為單位。因中文的特點在於通常以字為單位，而英文是以詞為單位，這是和利用「Bi-LSTM」模型的英文標點標註的研究不同，為了要沿用「Bi-LSTM」英文標點標註的研究成果，必須將中文稿以詞為單位分割，我們使用「ckip lab」所提供的斷詞api。

「建立對應字典」是因為要將詞彙以編號表示，必須建立每個詞彙和編號間的對應關係，我們是採用經「斷詞」後的維基百科資料建立對應字典。

二. 研究成果

1. 三種model在新聞資料集表現上的對比

我們使用三種不同的model（Bi-LSTM(with attention)、BERT、BERT+Bi-LSTM），來實作三類別標點(，。？)的分類器，並且測試了三種不同的model在加入錯漏字前後的資料集的Precision、Recall和F-score。

		None			COMMA,			PERIOD。			QUESTIONMARK?			OVERALL		
	MODEL	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
News_data	BERT+Bi-LSTM	0.986	0.991	0.988	0.816	0.803	0.810	0.857	0.683	0.769	0.625	0.653	0.639	0.822	0.778	0.798
	BERT	0.985	0.976	0.980	0.693	0.791	0.739	0.840	0.736	0.785	0.577	0.702	0.633	0.720	0.779	0.747
	Bi-LSTM	0.948	0.956	0.952	0.620	0.556	0.586	0.651	0.625	0.638	0.604	0.500	0.547	0.644	0.611	0.627
News_data with noise	BERT+Bi-LSTM	0.977	0.979	0.978	0.720	0.741	0.738	0.784	0.609	0.686	0.532	0.584	0.557	0.730	0.713	0.720
	BERT	0.973	0.969	0.971	0.644	0.702	0.672	0.797	0.651	0.716	0.524	0.610	0.564	0.673	0.691	0.679
	Bi-LSTM	0.927	0.948	0.937	0.597	0.535	0.564	0.540	0.459	0.496	0.550	0.416	0.474	0.586	0.519	0.550

根據我們的數據Bi-LSTM表現遠差於其他兩個BERT based的模型，根據我們的觀察其主要原因是在於Bi-LSTM使用的是傳統的embedding layer，在相對口語化的文本當，因為其無法判別同義詞的差異，導致學習效果不佳，而剩餘兩個model表現差異並不大，但 BERT+Bi-LSTM 在整體上稍微優於單純的 BERT 的表現，它不只能夠動態

獲取詞向量，根據前後文針對同義詞做出區別，並且也兼具了能夠重雙向讀入句子，藉此同時利用標點的前文和後文進行計算，所以在後續的測試當中，我們主要選擇以 BERT+Bi-LSTM 來進行實際應用場合的文本的標註。

2. 各種對話結構下的 BERT+Bi-LSTM 表現比較

我們將對話結構大略分成下列四種：

1. 單人講述
2. 雙人聊天(複數人數的對話文稿)
3. 多人聊天(複數人數的對話文稿)
4. 論辯對話(複數人數的對話文稿)

在現實的語音辨識任務當中會存在各種不同的對話模式，對話模式之間語句上意義的完整性差異，有些對話模式可能因多人對談或激烈辯論等導致語句時常被打斷而不完整等，要確保model能泛用在不同情境下的語音辨識，必須要在更種情境皆進行測試。

結果：

單人講述 範例影片-演講：學歷史的大用：呂世浩(Shih-Hao Lu) at TEDxTaipei 2014	**原始資料：** 我問各位這個故事到底在說什麼？這是一個智慧的故事，他的定義絕對不是告訴各位對老人家要有禮貌約會不要遲到，如果只是為了告訴這個，我不能不說中國的智慧絕對沒有這麼淺薄。這個故事到底要告訴各位什麼東西，各位跟著我讀下去：張良曾經從容地散步在下邳的橋上，這句平平無奇，可是歷史學看待事物不同的觀點就是我們有時間的縱深，我們看到每一樣東西必要追問它處在時空的那一個點上。張良在從容散步前他做了什麼事情？他剛做了一件驚天動地的大事，他在搏浪沙刺殺秦始皇。他為什麼要去刺殺秦始皇？張良是韓國宰相名門之後，秦國滅亡了韓國，他要為韓報仇消滅強秦。為了這個目的他散盡家財求取刺客，最後刺殺失敗他被全天下通緝。在這一刻他逃到了下邳， **Bert + Bi - LSTM：** 我問各位這個故事到底在說什麼？這是一個智慧的故事，他的定義絕對不是告訴各位對老人家要有禮貌約會不要遲到，如果只是為了告訴這個，我不能不說中國的智慧絕對沒有這麼淺薄，這個故事到底要告訴各位什麼東西，各位跟著我讀下去。張良曾經從容地散步在下邳的橋上。這句平平無奇，可是歷史學看待事物不同的觀點就是我們有時間的縱深，我們看到每一樣東西，必要追問它處在時空的那一個點上。張良在從容散步前，他做了什麼事情？他剛做了一件驚天動地的大事，他在搏浪沙刺殺秦始皇，他為什麼要去刺殺秦始皇。張良是韓國宰相名門，之後秦國滅亡了韓國，他要為韓報仇消滅強秦，為了這個目的，他散盡家財求取刺客，最後刺殺失敗，他被全天下通緝，在這一刻他逃到了下邳。
雙人聊天 範例影片-訪談：【博恩夜夜秀】韓總駕到！迎接夜夜秀開播以來最熟悉的陌生人！	**原始資料：** 大家快樂-沒有問題而且我們就是一個這麼民主自由的社會氛圍多好。你盡量開玩笑，沒關係真的嗎，確定?不要太...只有一個要求不要開我髮型的玩笑，其他都可以開，好不好?沒有問題X2剛才說，喔我已經跟...跟韓市長道過歉了，然後我想要問韓市長有沒有覺得想要跟...哪一些人道歉呢?是從我出生呱呱墜地到現在?都可以，都可以我不曉得。如果在從小到大的過程上有一些接觸的朋友心裡不舒服可能是我年輕的...年少輕狂的時候我不曉得。要從少年開始回憶嗎?都可以，隨便你，謝謝。如果大家感覺不好，非常抱歉。今天我好像看到了教堂對神父要懺悔這樣。要告解是不是?沒有 **Bert + Bi - LSTM：** 大家快樂沒有問題，而且我們就是一個這麼民主自由的社會氛圍多好，你盡量開玩笑沒關係真的嗎？確定不要太只有一個要求不要開我髮型的玩笑 其他都可以，開好不好沒有問題，X2剛才說喔，我已經跟跟韓市長道過歉了，然後我想要問韓市長有沒有覺得想要跟哪一些人道歉呢？是從我出生呱呱墜地到現在都可以都可以。我不曉得如果在從小到大的過程上，有一些接觸的朋友心裡不舒服，可能是我年輕的少輕狂的時候，我不曉得要 從少年開始回憶嗎，都可以隨便你謝謝，如果大家感覺不好非常抱歉，今天我好像看到了教堂對神父要懺悔，這樣要告解是不是沒有，

多人聊天 範例影片–多人桌遊： 【喳桌遊#100】狼人x損友x爆笑｜比起贏 我更想當暴民｜100集特別紀念｜下一次要找更多更多人	**原始資料：** 認真了我們要承襲我們在高中時期的時候 6 年我們要承襲我們在高中時期的時候 6 年高中嗎？對、高中高中就開始玩了？！對、高中就是...這遊戲應該是淦話遊戲才對就是...我們的規則一定跟你們看到得不太一樣我們是比較現在比較、現在比較有錢、我們以前都是拿紙就原始的版本現在比較、現在比較有錢、我們以前都是拿紙這個社區的狼人跟平民一樣暴亂XD現在比較、現在比較有錢、我們以前都是拿紙這個社區的狼人跟平民一樣暴亂XD對！我們以前都是拿紙我們拿筆之類的、自己在那邊寫！我們以前都拿白紙！所以我們沒有太多的規則！我們沒有太多規則跟拘束會有跟你們、痾、認知上可能會有不一樣的地方不過來吧你要！你要介紹你的角色！來吧 **Bert + Bi - LSTM：** 認真了。我們要承襲我們在高中時期的時候 6 年，我們要承襲我們在高中時期的時候 6 年高中嗎？對高中，高中就開始玩了，對高中就是，這遊戲應該是淦話遊戲才對，就是我們的規則一定跟你們看到得不太一樣，我們是比較現在比較現在比較有錢，我們以前都是拿紙就原始的版本，現在比較現 在比較有錢，我們以前都是拿紙，這個社區的狼人跟平民一樣暴亂XD，現在比較現在比較有錢，我們以前都是拿紙，這個社區的狼人跟平民一樣暴亂XD對，我們以前都是拿紙，我們拿筆之類的，自己在那邊寫，我們以前都拿白紙。所以我們沒有太多的規則，我們沒有太多規則跟拘束，會有跟你們 痾認知上可能會有不一樣的地方，不過來吧，你要你要介紹你的角色來吧，
論辯對話 範例影片–會議：立法院第9屆第6會期黨團協商會議紀錄	**原始資料：** 現在進行公共工程委員會的預算，公共工程委員會顏副主 任委員已經到場，我們就從第 1 目開始審查，第 1 目在委員會 是凍結 93 萬 5,000 元。對不對？ 林主任佳欣：我們是…… 主席：這是一般行政。請用麥克風發言。在委員會是一般行政第 1 目，我們是以「目」做處理，請大家看「目」，不要看「目 」裡面的「節」，那太細了。一般行政在委員會是凍結 93 萬 5,000 元，我們黨團沒有提案，所以我們就照委員會的決議通過。 林主任佳欣：是。 主席：第 2 目在委員會是減列 100 萬元，凍結 800 萬元，對不對 ？ 林主任佳欣：是。 主席：第 2 目第 13 案撤案；第 14 案撤案；第 15 案國民黨委員 賴士葆提案減列 20 萬元，這個是獎補助費，依照統刪處理， 所以我們就不處理了；第 16 案也是由賴士葆委員提出委辦費 ，這也是依照統刪處理，所以第 15 案、第 16 案依統刪處理。 許淑華委員在第 2 目的部分，提案凍結 5%，你們有沒有去溝 通？能不能接受？ 顏副主任委員久榮：有，我們去溝通了，提書面報告。 **Bert + Bi - LSTM：** 現在進行公共工程委員會的預算，公共工程委員會顏副主任委員已經到場，我們就從第1目開始審查，第1目在委員會是凍結93萬5000元，對不對，我們是這是一般行政，請用麥克風發言，在委員會是一般行政，第1目我們是以目做處理，請大家看目，不要看目裡面的節 那太細了。一般行政在委員會是凍結93萬，5，000元，我們黨團沒有提案，所以我們就照委員會的決議，通過是第2目，在委員會是減列100萬元，凍結800萬元。對不對是第2目，第13案撤案，第14案撤案，第15案，國民黨委員提案減列20 萬元，這個是獎補助費，依照統刪處理，所以我們就不處理了，第16案也是由賴士葆委員提出委辦費，這也是依照統刪處理，所以第15案，第16案依統刪處理，在第2目的部分，提案凍結5，你們有沒有去溝通，能不能接受，有我們去溝通了，提書面報告

評估與推論：

在單人講述的對話結構下，模型標註的並無異常，其語句分段正常，十分接近我們訓練所用的文本，而與原始資料相比，故而結果合乎預期。我們認為標註的目的也是為了增加可讀性，故而不會以標註和原始文本相同為目的，因為在實際應用的場合並沒有標準答案，只要語句可讀性佳，便達成我們的主要目的。

在雙人聊天、多人聊天和論辯對話下，會發現標記結果是比較混亂的，出現的錯誤包含「將不同人的對話歸類在同一個句子中。」、「無法標示過短的語句。」等，針對這些錯誤我們觀察其與原稿的差異，發現多人對話時容易

出現彼此搶話，使文本出現了語句重複、混亂與中斷的情形，且日常對話不符合嚴謹的語法結構，可能以簡化的短句或單詞來表達，大幅增加了模型標註的困難。

對於「複數人數的對話文稿」表現不佳的問題我們預期可以透過將不同人物的對話分離開來在各別進行標註，應可以達成更為優良的成果。

Reflections

這次專題對我們而言很大的難題是標點標註相關研究文獻多為英文，但我們的目標是建構繁體中文的系統，我們必須設法將英文自然語言處理相關研究的成果應用於繁體中文，我們蒐集了中文領域的自然語言處理所應用的資源，並參考許多相關技術，設法透過英文標點標註的相關研究來實現繁體中文的標點標註，過程中我們學到要靈活應用所蒐集到的資源才能完成工作，而並非只是單純模仿就可以達到目標，此外也瞭解要如何統整資訊並且發掘其中的價值。

機器學習領域而言資料極為重要這點之前我們只在課堂聽過卻為真正體會，但開始研究後深刻感受到資料對於我們的研究的影響，研究前期多次陷入的困境，便是測試表現良好，但成果在現實的標點標註任務表現卻嚴重劣化，我們為了解決問題查閱許多文獻。一開始只想從model和訓練方法進行調整，然而未獲得有效進步，於是我們開始研究和分析資料，得到的結論是因原先使用來自維基百科的資料，其在標點標註或語聚集詞彙使用上，和語音辨識文本大不相同，因此使我們決定建立第二種語法結構較近似 asr 文本的「news dataset」，才終於達成明顯改善。

在實作專題的過程我們無法知道確切的目標，因為model的表現永遠不可能完美，故也不會有所謂最佳解法，我們只是在期限內不斷嘗試錯誤與改進，雖然許多嘗試都是失敗的，但在嘗識中我們對於這個領域的瞭解演逐漸增加，model的表現也慢慢改善，我們也從中獲得成就感。