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

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

組別：A202 組員：鍾永桓、洪子堯、陳敬涵 指導教授：李祈均

Introduction

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，然而自動化語音辨識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，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。

我們利用三種不同的模型架構(Bi-LSTM(with attention)、BERT、BERT+Bi-LSTM)，觀察訓練結果，並分析可能影響其表現的因素，最終我們在將模型運用於實際場合的資料集驗證我們的成果。

資料處理

採用三種資料集

- wikipedia dataset (中文)
- news dataset
- asr dataset

並透過以下處理步驟。

【簡體轉繁體】
透過 opencc api 處理 wikipedia dataset，將混雜的簡體字轉換為繁體

【替換標點符號】
「：」、「；」、「、」替換為「，」
「！」、「.」替換為「。」
只保留三種標點符號(，。？)
因只要這三種標點就足以提高文章的可讀性。

【斷詞】
是將文本切割成以詞彙為單位而非以字為單位。
使用「ckip lab」所提供的斷詞api。

【建立對應字典】
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，我們採用經「斷詞」後的維基百科資料建立對應字典。

結論

【模型比較】

1. 以測試指標的結果而言，「BERT based」模型都優良於「Bi-LSTM」模型，其中又以「BERT + Bi-LSTM」在整體表現上最為優秀，精確率達八成。

2.「Bi-LSTM」模型，在精確率的表現也近六成五，而且相對「BERT based」模型而言，其結構參數較為精簡，反而更適合資源有限的裝置進行實作。

【實際應用】

1. 單人講述的對話結構下，模型標註優良，可讀性高，符合期待。

2. 雙人聊天、多人聊天和論辯對話下，標記較為混亂，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。」

3.「複數人數的對話文稿」預期可以透過將不同人物的對話分離開來再各別進行標註，達成更為優良的成果。

實驗結果

1. 三種model在新聞資料集表現上的對比
使用三種模型 Bi-LSTM(with attention)、BERT、BERT+Bi-LSTM，實作三類別標點(，。？)的分類器，並且測試三種不同的模型在加入錯漏字(noise)前後時資料集的 Precision、Recallc 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

2. 各種對話結構下的BERT+Bi-LSTM表現比較(範例輸出)

對話結構	輸出結果
單人講述 【範例】 演講	**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】 訪談	**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】 桌遊	**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】 會議	**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