



NLP & 標點符號還原

AI Final Project Presentation



Team 06

112550009 賴邑城 112550013 周廷威 112550201 蔡尚融

摘要

語音辨識是現今被廣泛使用的技術之一，無論是會議記錄的製作，或緊急的文書作業，都需要仰賴相關的自然語言處理技術。然而，語音辨識只能將口說轉化為沒有標點符號的文本，但沒有標點符號的英文文本是難以閱讀的，因此標點符號的還原技術是將語音轉換成可閱讀的文本中不可或缺的一環。現有的文法改錯工具（Grammarly等）和ChatGPT等都能做到標點符號還原的功能，但對於較長的文本的還原效果不如預期，畢竟標點符號還原並非這些工具的主要功能。

想法

我們的目標是開發一種系統，能夠在這種無標點符號的文本中準確插入標點符號，提高可讀性並保持原始口語內容的完整性。

資料集來源



英國國家語料庫 (BNC)

- 蒐集了大量從口說轉換成的英語資料
- 涵蓋各領域的內容
- 目前在網路上可使用的最大也最具代表性的語料庫之一

嘗試方向

- 訓練不同的模型並比較成效
- 研究使用不同方法從資料集篩選訓練資料對成效的影響
- 研究對預訓練模型使用不同的微調（fine-tune）方式

對成效的影響

創新性

- 研究 ChatGPT4 和 ChatGPT3.5 對於標點符號的還原任務的效果
- 分析跟統整其他模型的效果和造成成效差異的可能原因

方法

- BART

BART 能夠理解上下文和句子結構，使其可以根據語意選擇適當的標點符號

- T5 模型

作為解決 text-to-text 任務的 SOTA 模型，使其非常適合執行標點符號還原

前置處理

- 訓練資料處理

- 刪除出現未知的字元([UNK])，或不具任何意義的單字
- 去除未包含標點符號或標點符號不完整的句子，如引號未成對出現、句尾並沒有標點符號等
- 將篩選後的資料製作成可供模型使用的訓練資料，以處理前與處理後的句子形成一對一的對照

前置處理

- 訓練資料切割

- 隨機抽取一定數量的句子樣本
- 以 9:1 的比例將訓練資料分為訓練集和測試集
- 切割訓練集中的一成做為驗證集

模型訓練

- 使用 Happy Transformer 套件進行模型訓練
 - 提供模組化的模型訓練
 - 可使用在 Hugging Face 網站上以各種主題的資料訓練出的預訓練模型作為基礎，再進行微調
 - 預訓練模型如 T5、Bert、NLTK 等，T5 還依照預訓練模型的參數量多寡，區分為 T5-small、T5-base、T5-large 等，給予更多合適的選項

測試方法

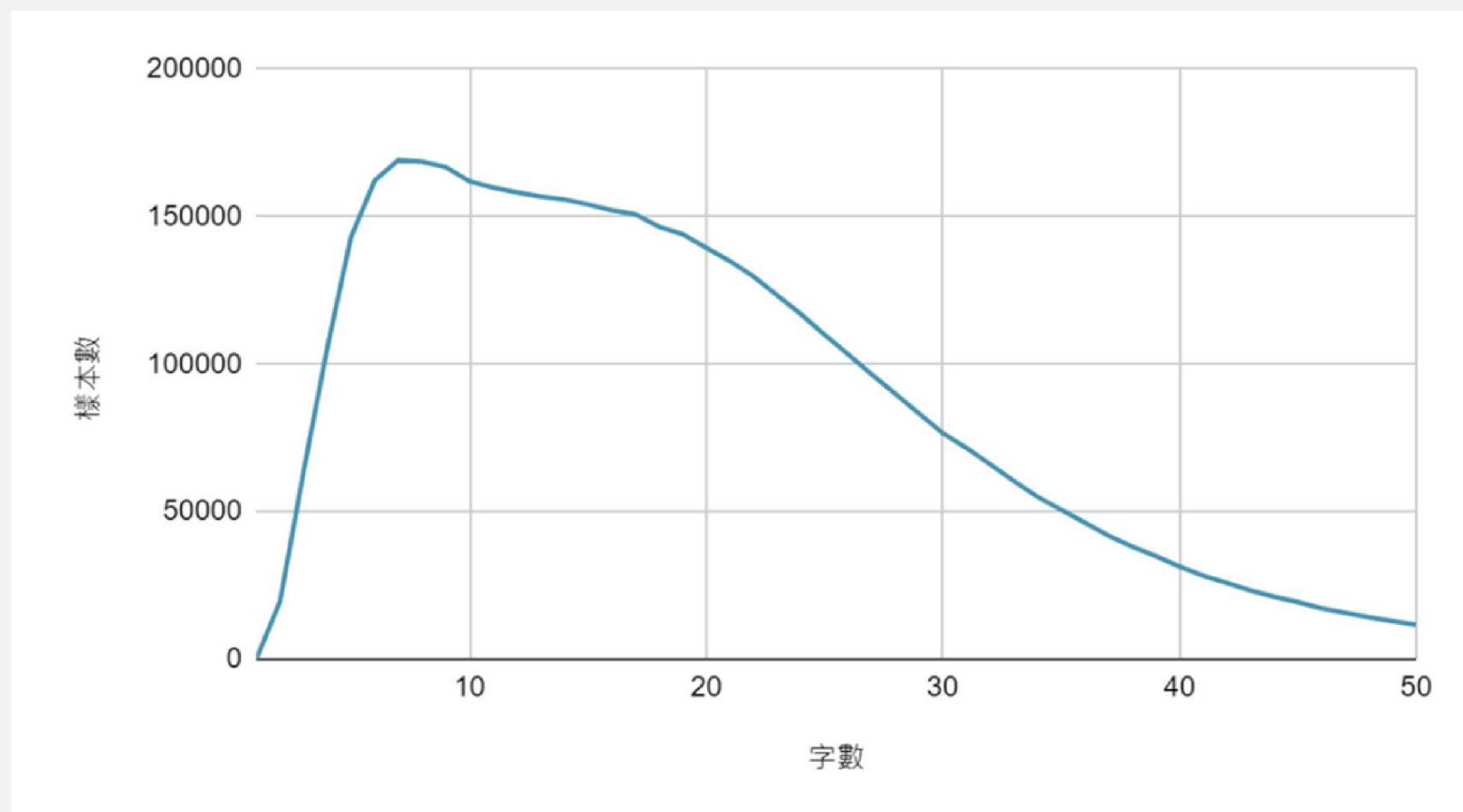
- 將對照組與實驗組對同一份隨機抽取的樣本進行改錯，方式為將樣本中正確的原文本中所有的標點符號去除，並交由模型各別復原標點符號
- 評估的方式為對一個文本中的每個單字後面是否該有標點符號進行逐一比對，並依照混淆矩陣分別計算六種標點符號個別的準確率、精準率、召回率、F1-score

結果評估

- 依照統計結果中依照混淆矩陣計算出的準確率、精準率、召回率之間的比較探討造成效果差異的可能原因，並針對不同變因重複進行步驟三至五進行延伸實驗，以驗證先前的假設

結果與討論

- 實驗前發現此訓練模型對於十字以上的文本的修改效果顯著較佳，而五字以內的文本的修改效果則較差



結果與討論

- Lab01：比較不同訓練資料筆數的模型表現
- 分別隨機抽取 1000、3000、5000、10000 筆訓練資料進行比較，分別以 model#1、model#2、model#3、model#4 作為其代稱

結果與討論

- Lab01：比較不同訓練資料筆數的模型表現
- 句點還原

- | Model | Accuracy | Precision | Recall | F1-Score |
|---------|----------|-----------|--------|----------|
| model#1 | 0.94 | 0.54 | 0.05 | 0.09 |
| model#2 | 0.97 | 0.81 | 0.66 | 0.73 |
| model#3 | 0.98 | 0.85 | 0.70 | 0.77 |
| model#4 | 0.98 | 0.89 | 0.73 | 0.80 |

結果與討論

- Lab01：比較不同訓練資料筆數的模型表現
- 逗號還原

Model	Accuracy	Precision	Recall	F1-Score
model#1	0.95	0	0	*
model#2	0.95	0.38	0.01	0.03
model#3	0.95	0.51	0.07	0.12
model#4	0.95	0.54	0.15	0.23

結果與討論

- Lab02：比較不同預訓練模型的效果表現
- 預訓練模型有 T5-base (220M 參數)、T5-small(60M
參數)、BART 等
- 針對上述三種預訓練模型以筆訓練資料微調後進行效果
比較

結果與討論

- Lab02：比較不同訓練資料筆數的模型表現
- 句點還原

Model	Accuracy	Precision	Recall	F1-Score
T5-small	0.99	0.82	0.88	0.85
T5-base	0.99	0.89	0.93	0.91
BART	0.99	0.90	0.93	0.91

結果與討論

- Lab02：比較不同訓練資料筆數的模型表現
- 逗號還原

Model	Accuracy	Precision	Recall	F1-Score
T5-small	0.95	0.55	0.09	0.16
T5-base	0.96	0.68	0.50	0.58
BART	0.96	0.73	0.53	0.61

結果與討論

- Lab03：比較 我們的model 與 ChatGPT 的效果表現
- ChatGPT的效果遠優於我們的model
- 推測標點符號還原的效果與資料集大小有直接關係

結果與討論

- Lab03：比較我們的model 與 ChatGPT 的效果表現
- 句號還原

Model	Accuracy	Precision	Recall	F1-Score
Our Model	0.85	0.26	0.16	0.2
ChatGPT 3.5	0.91	0.58	0.72	0.64
ChatGPT 4.0	0.92	0.62	0.68	0.65

結果與討論

- Lab03：比較我們的model 與 ChatGPT 的效果表現
- 逗號還原

Model	Accuracy	Precision	Recall	F1-Score
Our Model	0.99	0	0	*
ChatGPT 3.5	1	1	1	1
ChatGPT 4.0	1	1	1	1

結論

- 預訓練模型參數量
- 微調資料集大小
- 訓練資料長度
- 標點符號出現次數

參考文獻

- Yi, J., Tao, J., Bai, Y., Tian, Z., & Fan, C. (2020). Adversarial transfer learning for punctuation restoration. arXiv preprint arXiv:2004.00248.
- Cho, E., Niehues, J., & Waibel, A. (2012). Segmentation and punctuation prediction in speech language translation using a monolingual translation system. In Proceedings of the 9th International Workshop on Spoken Language Translation: Papers (pp. 252-259).
- Tilk, O., & Alumäe, T. (2016, September). Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In Interspeech (Vol. 3, p. 9).

THANK YOU

