



NLP & 標點符號還原

AI Final Project Proposal 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

方法

- BERT

BERT 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 和 BART 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

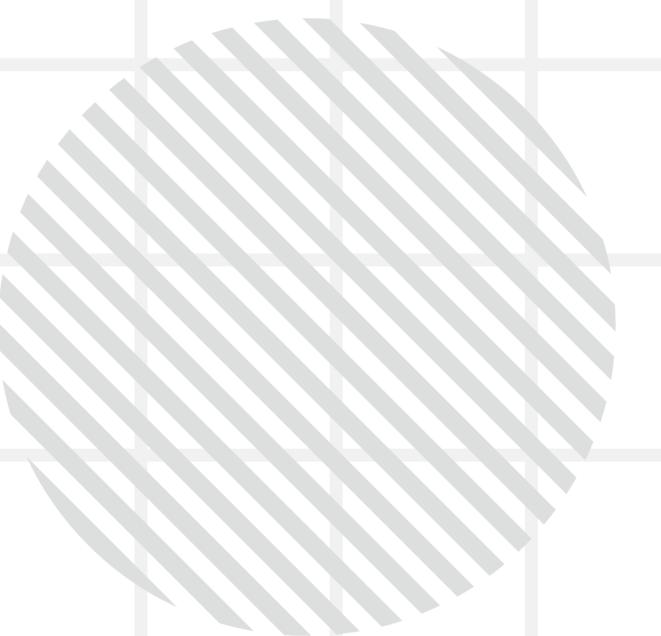
創新性

- 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

相關研究

- 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).

Q&A



THANK YOU

