語言模型實驗

AI3A – 410770019 林郅恒

摘要— 本專題利用課堂語音紀錄作為訓練資料，探討了手刻數學模型在語言處理任務中的應用。通過對某一堂課程語音紀錄進行深度分析和特徵提取，我設計了一系列基於統計和規則的數學模型。這些模型涉及自然語言理解、文本生成等多個任務，並且在不同的語言處理場景下進行了實驗驗證。實驗結果顯示，利用課堂語音紀錄訓練的手刻數學模型在特定任務上能夠取得良好的性能，並且在某些情況下具有一定的競爭力與實用性。此外，我們還分析了模型的優勢和限制，並提出了一些改進和優化的建議。總的來說，基於課堂語音紀錄的手刻數學模型在語言處理中具有一定的應用價值，並且對於理解和處理課堂語音紀錄等特定文本類型具有重要意義。

Index Terms—Language models , Natural Language Processing , Mathematical models , Artificial General Intelligence.

1 介紹

語言模型作為自然語言處理領域中的一個重要技術，近年來在小說、機器問答等生成方面展示出了巨大的潛力。本文將介紹語言模型在課堂生成方面的應用，從其原理和方法入手，探討其在創作過程中的優勢和挑戰。

首先，語言模型是通過對自然語言的統計建模來模擬語言的生成過程。它可以學習到語言中的規則和結構，進而生成符合語言風格和語法的文本。在課堂生成中，語言模型可以通過分析大量的課堂上課內容之文本，學習到老師上課的風格和結構，進而生成新的課堂文本。其次，語言模型在課堂生成中具有很多優勢。首先，它可以自動化生成文本，節省了人工備課的時間和成本。其次，語言模型可以根據不同的老師風格和說話方式生成不同的課程文本，滿足不同學生的需求，並且產生出更人性化的課程內容。此外，語言模型還可以通過不斷的迭代和訓練不斷提升生成課程文本的質量和多樣性。然而，語言模型在課堂生成中也面臨一些挑戰。首先，生成的文本可能存在語法錯誤和邏輯不通，需要進行後期的校對和修正。其次，語言模型生成的文本可能缺乏情感與合理性，需要進一步的提升和優化。此外，語言模型還需要大量的訓練數據和計算資源，才能夠生成高質量的文本。

語言模型在課程生成方面具有巨大的應用價值，可以為老師與學生提供更多的選擇和樂趣。隨著技術的不斷進步和發展，相信語言模型在課程生成中的應用將會越來越廣泛，帶來更多的驚喜和創新。

2 問題及方法描述(數學原理推導)

在課程生成文本中遇到了語法錯誤與邏輯不通的問題，因此建立了四個文字字典，用來計算每個字前後出現字的機率，進而可以計算出，每個字之後須要放什麼字才會讓語法通順以及邏輯正確。而此問題的解決方法分成兩部分，分別為建立文字字典以及利用貝氏定理計算字出現的機率與合理性。

一共建立了四個文字字典，分別為下個字的字典、下下個字的字典、上個字的字典以及上上個字的字典，此四個字典可用來呈現訓練資料中，每個字前後接的字詞以及該數量為何。

建立好字典後利用貝氏定理，計算下兩個字詞出現的機率，進而可以計算出下個字詞與下下個字詞為何。

貝氏定理：計算在 B 發生的情況下 A 發生的機率。

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(B)}$$

由貝氏定理推導，若要計算下個字為某字的機率，則可以用以下公式表示。在前 i 個字（d_i）的情況下 a 出現的機率取最大值即為d_i後最有可能出現的字。

$$\arg\max_{a\in A} P(a|d_i) = \arg\max_{a\in A} P(a)\prod_{j\in N} P(d_i^j|a)$$

3 實驗設計與結果(包含資料蒐集、流程圖、架構圖)

此實驗的資料蒐集為人工智慧實務課程的線上語音紀錄，透過 jieba 套件分割成 8941 個詞語，且包含了標點符號。這樣的訓練資料具有人性化的功能。

圖 1 為本實驗的設計流程圖，由一開始的資料蒐集經套件作文字分割後，依照訓練資料建立對應的文字字典，最後利用貝氏定理計算各詞出現的機率，產生出一個符合人性化的句子。

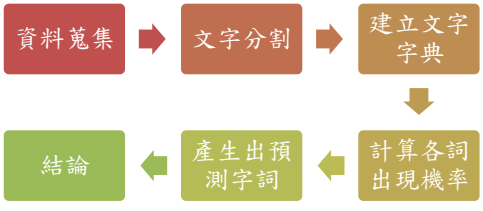

圖 1

圖 2 為本實驗的架構圖。

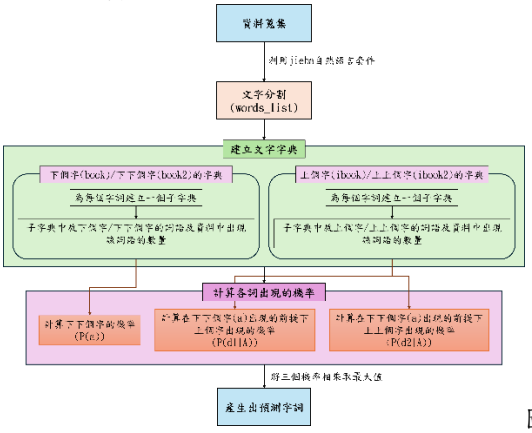

圖 2

4 分析及討論

本研究利用課堂語音紀錄訓練了基於手刻數學模型的語言生成模型，在特定任務上取得了良好的性能。然而，我發現模型在生成文本時仍然存在一些語法錯誤和邏輯不通的問題，需要進一步的優化和改進。

此語言模型可以整理出一些優勢與挑戰，優勢為能夠自動化生成課程文本，節省了人工備課的時間和成本。此外，模型能夠根據不同老師的風格和說話方式生成不同的課程文本，提供更加人性化的教學內容。而挑戰為模型生成的文本仍可能存在語法錯誤甚至邏輯不通，需要再更多大量的訓練數據與計算資源，此外，文本可能缺乏情感和合理性，需要進一步提升和優化，才能夠生成更高質量的文本。

最後為提升模型的性能和生成文本的質量，我可以提出一些建議與改善，首先是增加訓練數據量，增加課堂語音紀錄的訓練數據量，以提高模型的泛化能力和準確性。此外可以引入情感分析技術，使模型生成的文本更具情感色彩和人性化。

REFERENCES

[1] 張瓊文、徐嘉連、張俊盛 (2015)。"基於貝氏定理自動分析語料庫與標定文步" [A Bayesian approach to determine move tags in corpus]. ROCLING 2015, Proceedings of the 2015 Conference on Computational Linguistics and Speech Processing, 頁 87-99。© The Association for Computational Linguistics and Chinese Language Processing.

[2] 研究方法與研究架構 https://www.angle.com.tw/File/Try/1Z002PA-3.pdf