AI 生成文本判別

指導教授 莊秀敏 教授

學生 11127229 廖翊崴

摘要

由於 AI 生成文本（GPT）的快速普及，AI 文本與人類創作界線日益模糊，透過深度學習技術找出語言特徵、句構和風格差異，訓練出具備分辨能力的 AI 模型。

資料集

以句號做切分，筆為單位，一個句子為一筆。

總資料量	4054 筆
0（非 AI 文本）	1973 筆
1（AI 文本）	2081 筆

訓練集/測試集 8:2，加入分層抽樣，以確保訓練集與測試集各類別的比例和資料量是相近的

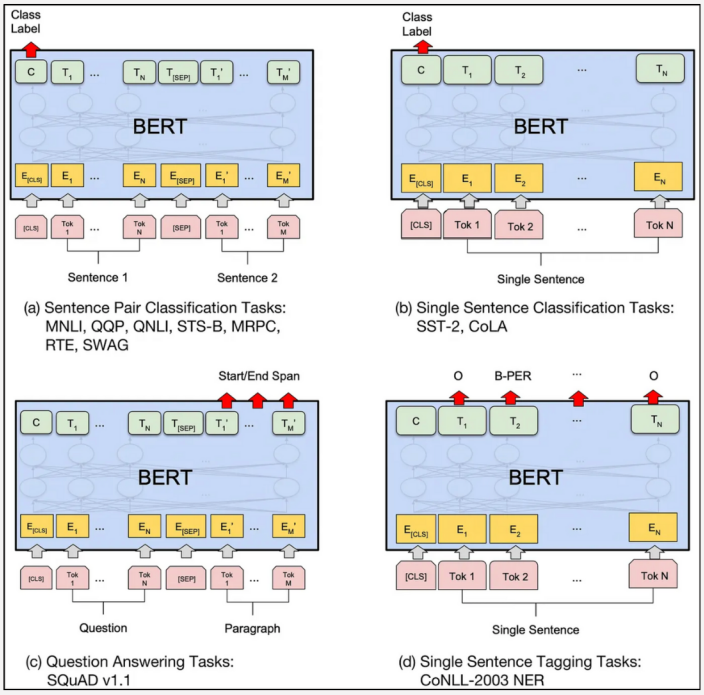
	0（非 AI 文本）	1（AI 文本）	總和
訓練集	1578 筆	1665 筆	3243 筆
測試集	395 筆	416 筆	811 筆

基礎模型

市面上多數中文 BERT 相關模型以簡體語料為主，我們採用中研院 CKIP LAB 釋出的繁體中文 BERT BASE 預訓練模型（CKIPLAB/BERT-BASE-CHINESE）進行微調。

為什麼選擇BERT

BERT（Bidirectional Encoder Representatio-ns from Transformers）的核心是一個基於多層堆疊的 Transformer Encoder 架構。它利用 Multi-Head Self-Attention 機制，使其能夠同時考量輸入序列中每個詞元（Token）的左側和右側所有上下文資訊，從而產生具備高度情境化（Contextualized）的向量表示。BERT 的訓練透過兩個自監督任務進行：遮蔽語言模型（MLM），用以學習詞元層級的雙向語義；以及下一句預測（NSP），用以捕捉句子層級的連貫性。這種架構賦予 BERT 強大的通用語言理解能力，並允許其模型參數通過遷移學習，高效地適應廣泛的下游 NLP 任務。



模型優化

結合三種技術強化文本理解：在原本切分好的資料集下，利用 CKIP LAB 的 CKIPTAGGER 進行繁體中文的詞性標註；使用 STANFORD STANZA（基於多層 BILSTM 的深度神經網路）進行精確的中文依存句法分析，以強化語法結構特徵；並採用 TF-IDF 關鍵詞萃取以突顯重要詞彙，同時作為識別 AI 文本詞彙重複性的線索。

結果顯示：強化後的文本，模型訓練結果在各項指標（MCC, ACC, F1 等）均有進步（下圖），且經過雙尾檢定——原始資料和強化資料各八次的 ACC，確認不是隨機誤差 (P-VALUE 0.0060 < 0.05)，是有優化。

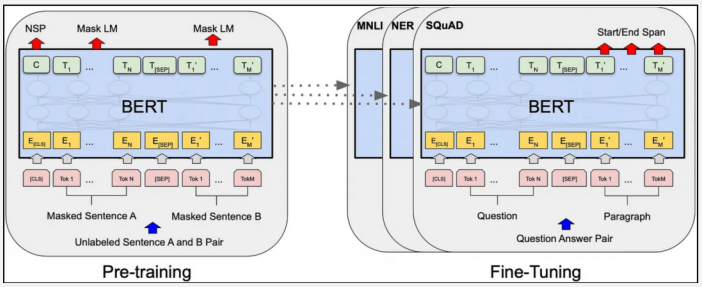
	mcc	acc	F1	F1-P (1)	F1-N (0)	eval_loss
原本	0.61623	0.80025	0.79585	0.82581	0.76590	1.4733
處理	0.64194	0.81504	0.81189	0.83624	0.78754	1.4037
增減	+ 0.026	+ 0.015	+ 0.016	+ 0.010	+ 0.022	- 0.070

後端部分除了載入模型進行文本的預測任務外，當文本被判別為 AI 生成時，會使用集成梯度法（Integrated Gradients）進一步計算每個詞元（Token）對此分類結果的貢獻度，藉此作為使用者修改的依據。

介面實際運作結果

核心貢獻

1. 二分類模型：提出可有效區分 AI 生成與人類撰寫文本的深度學習模型。
2. 細化偵測層級：模型聚焦於句子與詞元（TOKEN）層級的精準標註，以提升辨識細緻度，應對現實中 AI 生成與人類撰寫內容混雜的情境。
3. 使用者介面：開發操作簡便、介面清晰的前端系統，讓非技術使用者也能輕鬆操作並理解分析結果，提升在內容審查領域的應用價值。



BERT 的整體預訓練和微調流程 ⤴

← BERT 在不同任務上的微調範例