Transformer 架構下的增強型中文語言模型與任務套件(CKIP Transformers)

中央研究院 ACADEMIA SINICA

摘要

在許多人工智慧或資料處理的任務中，語言的處理常常是不可或缺的步驟之一，我們以 transformer 模型為基礎，進一步訓練了多個針對繁體中文的優化模型，包含語言理解用的 CKIP Albert 語言模型和 CKIP BERT 語言模型，以及語言生成用的 CKIP GPT2 語言模型等等。這些語言模型可以根據不同的語言處理任務進行再訓練，來滿足終端的各類實際需求。在這個套件當中，除了提供上述的語言模型之外， 我們也針對最普遍的任務需求，包含斷詞、詞性標記、專有名詞辨識(高達 18 類，包含:人名、團體、設施、組織、地理、地點、商品、事件、藝術品、法律、語言、 日期、時間、比例、錢、數量、序數、數詞。)等終端任務，提供多個再訓練過後的語言理解模型。系統以 Python 寫成，效能優異，且呼叫方式簡潔，易於整合。

線上展示網址為:https://ckip.iis.sinica.edu.tw/service/transformers/，歡迎實際測試。

技術優勢

- 以 transformer 模型為基礎，進一步訓練了多個針對繁體中文的優化模型
- 斷詞表現大幅超越結巴系統，且提供結巴系統所沒有的實體辨識
- 詞性標記的種類豐富: 共 61 種詞性 (https://github.com/ckiplab/ckiptagger/wiki/POS-Tags)
- 實體辨識的種類豐富: 11 類一般領域專有名詞及 7 類數量詞 (https://github.com/ckiplab/ckiptagger/wiki/Entity-Types)
- 支援使用者自訂詞典。
- 可以針對新的任務進行再訓練

Models	#Parameters	Perplexity	WS (F1)	POS (ACC)	NER (F1)
ckip-albert-tiny-chinese	**4M**	**4.80**	**96.66%**	**94.48%**	**71.17%**
ckip-albert-base-chinese	**10M**	**2.65**	**97.33%**	**95.30%**	**79.47%**
ckip-bert-base-chinese	**102M**	**1.88**	**97.60%**	**95.67%**	**81.18%**
ckip-gpt2-base-chinese	**102M**	**14.40**	--	--	--
albert_chinese_tiny	4M	74.93	--	--	--
albert_chinese_base	10M	22.34	--	--	--
bert-base-chinese	102M	2.53	--	--	--
Jeiba	--	--	81.18%	--	--

Note: Perplexity: 數字越小越好。 ACC and F1: 數字越大越好。

圖1.效能比較

本院覽號

05T-1110222

公告日期

2022-07-03

智財權狀態

know-how

應用範圍

- 大數據輿情分析
- 語言理解
- 智慧客服
- 聊天機器人
- 商品情報分析系統

創作人

馬偉雲、楊慕