結合斷詞、詞性標記、實體辨識的中文處理套件(CkipTagger)

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，根據我們自行設計的深度學習算法，我們完成了這個結合斷詞、詞性標記、實體辨識的一站式中文處理套件，系統以python寫成，效能優異，且呼叫方式簡潔，易於整合，系統取名為CkipTagger，除了斷詞與詞性標記外，專有名詞辨識，或稱實體辨識（Named Entity Recognition, NER）是非常實用的功能，其目標為在文字資料當中，能夠辨識出感興趣的專有名詞(包含原本資料庫不存在的新專有名詞)，並自動標記正確的分類，如人名、地名、組織名等等，是人工智慧當中理解語言的重要步驟。目前我們所開發的中文專有名詞辨識系統能辨識11 類一般領域專有名詞及 7 類數量詞，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。

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

技術優勢

1. 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。
2. 詞性標記的種類豐富： 共61種詞性 (https://github.com/ckiplab/ckiptagger/wiki/POS-Tags)
3. 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)
4. 支援使用者自訂詞典。
5. 相關技術發表在著名的人工智慧國際會議 – AAAI 2020 (https://arxiv.org/abs/1908.11046)

Tool	(WS) prec	(WS) rec	(WS) f1	(POS) acc
結巴系統	90.51%	89.10%	89.80%	--
CkipTagger	97.49%	97.17%	97.33%	94.59%

圖1.CkipTagger與結巴系統的效能比較

圖2.CkipTagger的使用範例

本院覽號

05T-1081218

公告日期

2020-07-02

智財權狀態

know-how

應用範圍

1. 大數據輿情分析
2. 語言理解
3. 智慧客服
4. 聊天機器人
5. 商品情報分析系統

創作人

馬偉雲、李朋軒