anboh AI共學社群 我的 AI共學社群 > Part1 - NLP經典機器學習馬拉松 > D07 使用CkipTagger進行中文斷詞 D07 使用CkipTagger進行中文斷詞 囯 範例與作業 問題討論 簡報閱讀 重要知識點 繁體中文斷詞: CkipTagg... > NLP自然語言學習實戰馬拉松 ▶ Day7 - 使用CkipTagger進行中文斷詞 系統環境安裝 CkipTagger優勢 陪跑專家:Leo Liou 劉冠宏 重要知識點 重要知識點 如何使用CkipTagger來進行實際的斷詞操作。 籃球□比賽 籃球比賽 補充資料:<u>簡單易用的中英文斷詞和詞性標註</u> 繁體中文斷詞: CkipTagger • CkipTagger為台灣中央研究院詞庫小組所開發的NLP(自然語言處理)套件,是個以深度學習模型 為基礎而成的NLP(自然語言處理)應用。 其訓練的文本資料來源為中央社、wiki (用舊版套件先 進行斷詞) 及 ASBC(Academia Sinica Balanced Corpus)(為期近10年人工標記)。 CkipTagger模型主要的功能有: 1. WS: 斷詞 2. POS: 詞性標注 3. NER: 實體辨識 系統環境安裝 • 標準安裝 (同時安裝tensorflow, gdown) • gdown 是與 Google Drive 之間的API,可以從 Google Drive上下載資料 pip install -U ckiptagger[tf,gdown] pip install -U ckiptagger[tf,gdown] • 最小安裝 (不安裝tensorflow, gdown) 若環境已有安裝tensorflow, gdown, 可使用此種安裝 pip install -U ckiptagger pip install -U ckiptagger • 完整安裝 (安裝gpu版tensorflow, gdown pip install -U ckiptagger[tfgpu,gdown] pip install -U ckiptagger[tfgpu,gdown] CkipTagger優勢 在繁體中文上斷詞與詞性標記的表現進一步提升,並超越結巴系統 Tool (WS) prec (WS) rec (WS) f1 (POS) acc 97.49% 97.17% 97.33% 94.59% CkipTagger 95.96% 90.62% CKIPWS (classic) 95.85% 95.91% Jieba-zh_TW 90.51% 89.10% 89.80% 資料來源: github 結合實體命名:目前CkipTagger能辨識11 類一般領域專有名詞及7 類數量詞 中文 Type Description GPE Countries, cities, states 行政區 人物 PERSON People, including fictional DATE Absolute or relative dates or periods 日期 ORG Companies, agencies, institutions, etc. 組織 CARDINAL 數字 Numerals that do not fall under another type 民族、宗教、政治團 NORP Nationalities or religious or political groups Non-GPE locations, mountain ranges, bodies of LOC 地理區 時間 TIME Times smaller than a day FAC Buildings, airports, highways, bridges, etc. 設施 MONEY 金錢 Monetary values, including unit 序數 ORDINAL "first", "second" Named hurricanes, battles, wars, sports events, **EVENT** WORK_OF_ART Titles of books, songs, etc. 作品 數量 QUANTITY Measurements, as of weight or distance PERCENT 百分比率 Percentage (including "%") 語言 LANGUAGE Any named language 產品 PRODUCT Vehicles, weapons, foods, etc. (Not services) Named documents made into laws 法律 LAW 資料來源: github 結合詞性標注 Type Description 非謂形容詞 對等連接詞 Caa 連接詞,如:等等 Cab 連接詞,如:的話 Cba 關聯連接詞 Cbb 副詞 D 數量副詞 Da 動詞前程度副詞 Dfa 動詞後程度副詞 Dfb 時態標記 Di 句副詞 Dk DM 定量式 感嘆詞 普通名詞 Na 專有名詞 Nb 地方詞 Nc 位置詞 Ncd 時間詞 Nd 指代定詞 Nep 數量定詞 Neqa 資料來源: github CkipTagger的優勢 • 支援使用者自訂 參考/強制 詞典。 • 支援不限長度的句子。 • 不會自動 增/刪/改 輸入的文字。 CkipTagger斷詞技巧 Word-level approach • Maximum length (長詞優先) • 動態規劃查找最大概率路徑(Jieba) Character-level approach character Sequence Labeling CkipTagger 則是綜合上面兩種方法,針對 word 及 character 同時進行分析 以CkipTagger進行斷詞 • 下載預訓練權重 • 此模型需2GB的儲存空間 from ckiptagger import data_utils, WS 2 data_utils.downlaod_data_gdown("./") from ckiptagger import data_utils, WS data_utils.download_data_gdown("./") • 建構斷詞器 • 使用Ckiptagger進行斷詞 2 ws = WS("./data/") 4 #使用Ckiptagger進行斷詞 5 #返回為斷詞完成的list 6 input_string = '小明碩士畢業於國立臺灣大學,現在在日本東京大學進修深造' 7 word_sentence_list = ws(8 input_string, 9 sentence_segmentation = True, # To consider delimiters 10 segment_delimiter_set = {",", ".", ":", "?", "!", ";"}) 11 print(word_sentence_list) [['小'],['明'],['碩'],['士'],['畢'],['業'],['於'],['國'],['立'],['臺'],['潤'],['大'],['學'],['','],['現'], ['在'],['在'],['日'],['本'],['東'],['京'],['大'],['學'],['進'],['修'],['淀'],['造']] 1 #建構斷詞 2 ws = WS("./data") 4 input_string = '小明碩士畢業於國立臺灣大學,現在在日本東京大學進修深造' 5 word_sentence_list = ws(input_string, sentence_segmentation = True, # To consider delimiters segment_delimiter_set = {",", " • ", ":", "?", "!", ";"}) # This is the defualt set of delimiter 以CkipTagger詞性標注 使用Ckiptagger詞性標注 1 from ckiptagger import POS 3 # 建構詞性標注器 4 pos = POS("./data") 6 pos_sentence_list = pos(word_sentence_list) #帶入的是經過斷詞後的 list,不是原始文本 7 print(pos_sentence_list) [['VH'], ['Nd'], ['Na'], ['Na'], ['VH'], ['Na'], ['P'], ['Nc'], ['VC'], ['Nc'], ['Na'], ['VH'], ['VC'], ['COMMACATEGO RY'], ['D'], ['P'], ['P'], ['Nc'], ['Ncd'], ['Ncd'], ['Nc'], ['VH'], ['VC'], ['VCL'], ['VC'], ['VC']) from ckiptagger import POS 3 pos = POS("./data") pos_sentence_list = pos(word_sentence_list) print(pos_sentence_list) 以CkipTagger命名實體辨識 使用Ckiptagger命名實體識別 帶入的是經過斷詞後的 list,不是原始文本 帶入的是經過詞性標注的 list,不是原始文本 1 from ckiptagger import NER 3 #建構命名實體器 4 ner = NER("./data") entity_sentence_list = ner(word_sentence_list, pos_sentence_list)
print(entity_sentence_list) [set(), set(), set(), set(), set(), set(), set(), set(), set(), {(0, 1, 'GPE', '臺')}, set(), 1 ner = NER("./data") entity_sentence_list = ner(word_sentence_list, pos_sentence_list) print(entity_sentence_list) 帶入自定義字典 將自定義字典加入斷詞器中 from ckiptagger import construct_dictionary word_to_weight = {"日本東京大學": 1}
dictionary = construct_dictionary(word_to_weight) 1 from ckiptagger import construct_dictionary 3 word_to_weight = {"日本東京大學": 1} 4 dictionary = construct_dictionary(word_to_weight) 建構斷詞器 1 ws = WS("./data/") 2 input_traditional_str = ['小明碩士畢業於國立臺灣大學,現在在日本東京大學進修深造'] word_sentence_list = ws(input_traditional_str, recommend_dictionary=dictionary) 4 print(word_sentence_list) [['小明','硕士','畢業','於','國立','臺灣','大學',','現在','在','日本東京大學','進修','深造']] 1 ws = WS("./data") input_traditional_str = ['小明碩士畢業於國立臺灣大學,現在在日本東京大學進修深造'] word_sentence_list = ws(input_traditional_str, recommend_dictionary=dictionary) print(word_sentence_list) 詳細使用操作 請參照使用使用CkipTagger進行繁體中文斷詞.ipynb檔進行更詳細的使用操作 知識點回顧 在這章節我們學習到 • 使用CkipTagger來進行各項任務操作(如斷詞) 延伸閱讀 網站: <u>CkipTagger官方Github</u> CkipTagger官方Github(提供操作說明) Why GitHub? V Team Enterprise Explore V Marketplace Pricing Sign in Sign up □ ckiplab / ckiptagger ⊙ Watch 70 ☆ Star 1.3k ♀ Fork 152 ♦ Code Issues 11 Pull requests 2 Actions Projects Wiki Security Insights 古 Join GitHub today GitHub is home to over 50 million developers working together to host and review code, manage projects, and build software together. ያ master + ያ 1 branch ♡ 8 tags Go to file CKIP Neural Chinese Word jacobvsdanniel add corpora/embedding wiki b4febd9 on 8 Jul 3 64 commits Segmentation, POS Tagging, and update download url 5 months ago LICENSE 12 months ago ब्रीड GPL-3.0 License README.md add corpora/embedding wiki 2 months ago 網站:中研院詞庫小組馬偉雲主持人演講 有關CkipTagger內部詳細介紹 中央研究院 词庫小組 結合斷詞、詞性標記、實體辨識的一站式中文處理 CKIP LAB 開源套件 - CkipTagger 馬偉雲 助研究員 中研院詞庫小組主持人 2019/11/21 網站: CkipTagger命名實體類別 CkipTagger所有命名實體類別表 Why GitHub? V Team Enterprise Explore V Marketplace Pricing Sign in Sign up ☐ ckiplab / ckiptagger **Entity Types** jacobvsdanniel edited this page on 4 Oct 2019 - 4 revisions Ordered by frequency: - Pages 🗓 中文 Description Type Find a Page... 行政區 GPE Countries, cities, states Home 人物 PERSON People, including fictional Chinese README 日期 DATE Absolute or relative dates or periods Corpora **Entity Types** 組織 ORG Companies, agencies, institutions, etc. **POS Tags** CARDINAL Numerals that do not fall under another type 數字 NORP 民族、宗教、政治團體 Nationalities or religious or political groups Clone this wiki locally LOC Non-GPE locations, mountain ranges, bodies of water https://github.com/cki; 網站:<u>CkipTagger詞性標注類別</u> CkipTagger所有詞性類別表 Why GitHub?

Team Enterprise Explore

Marketplace Pricing Sign in Sign up □ ckiplab / ckiptagger ○ Code ① Issues 11 11 Pull requests 2 ② Actions □ Projects □ Wiki ① Security ☑ Insights **POS Tags** jacobysdanniel edited this page on 29 Aug 2019 - 5 revisions Ordered by type: → Pages 🕞 Description Find a Page... 非調形容詞 Home Caa 對等連接詞 Chinese README Cab 連接詞,如:等等 Corpora Cba 連接詞,如:的話 **Entity Types POS Tags** Cbb 關聯連接詞 D 副詞 Clone this wiki locally 數量副詞 Da https://github.com/ckir 網站:<u>中研院詞庫小組馬偉雲主持人演講筆記</u> 有關CkipTagger與NLP發展介紹筆記整理 Math.py 中文自然語言處理 (NLP) 的進展與挑戰 Wir müssen wissen , wir werden wissen # 首頁 ## 分類 ■ 婦檔 時間: 2019.10.31 地點:台灣大學德田館 103 教室 主講者:馬偉雲博士 文章目錄 本站概要 筆者對 Natural Language Processing (NLP, 自然語言處理)一直蠻有興趣,因此一 1. 中研院詞庫小組(CKIP)

2. NLP簡介

3. NLP 進展 4. NLP 挑戰

6. QA

5. NLP 解決之道

7. CkipTagger

個月前看到這個活動資訊便立刻決定參加,上周中研院開放參觀,筆者也有幸跟馬教

授交談過,針對自己在進行 LineBot 的一些疑問請益,便更期待這一天的講座,以下

是中研院跨所的一個中文計算語言研究小組,五個主要研究方向:深度學習、知識表

是在講座過程中筆者的隨手筆記,或許沒有非常鉅細靡遺,但盡可能的記下重點。

中研院詞庫小組(CKIP)

達、自然語言理解、知識擷取、聊天機器人。