CorpusforDDL 參數計算說明

2022/10/10

# 參數設計 (擷取自 pdf p30)

1. 句子長度:Kilgarriff et al. (2008) 建議例句由 10-25 個詞組成最理想。我們設想或許可以依照初、中、高不同程度為來設定不同的句子長度,如果是連接句子的連詞或副詞,肯定需要較長的句子。
2. 詞頻:句子中的用詞以高頻為佳,排除罕用詞、專有名詞。
3. 詞彙等級:以《國教院詞語分級表》為主(許添明等人,2019)。如圖 15 所示,1表示常用、基礎詞彙。
4. 詞彙長度:對學習者而言,中文三字詞以上的詞彙可能較有難度,例如,三字詞通常含有詞頭或詞尾詞綴,需要有這類知識才能理解。四字格也是學習者的難點,尤其是詞義無法由表面推測出來的成語或固定格式。
5. 搭配詞資訊:學習詞就是學習用法,詞語搭配的重要性不言而喻。不過究竟要看關鍵詞的左邊還是右邊搭配,甚至左右都得看,需要進一步研擬,不同詞性的搭配詞應有些規律可循,例如,動作及物動詞多數看右邊搭配,名詞多數看左邊搭配等等。
6. 完整的句子:選取完整的句子是為了讓學習者透過完整句義來理解用法,這裡的完整句子不能單由標點符號來判斷。因為在真實語料中,很多句子因為已知訊息而省略了主語或賓語,簡單地說,一個動作及物動詞的例句,一定至少要有主語、賓語,不能選取「他吃了。」這樣的句子為例句。
7. 完整語境:語境對於詞性來說相當重要,尤其是對於有連接句子功能的詞或是表示語氣的副詞,例如,「並且、總之、居然、難道、才、再、就」等等。應避免出現在句首,需要有完整的語境。

# 實作方式

本次交付內容以國語日報新聞的 2001.txt 為例，取前 500 句作為測試。前處理步驟與參數計算方式如下：

## 前處理步驟

1. 移除換行符號 (\n)
2. 移除只含有數字 (e.g., 2001/1/1) 或符號 (e.g., —--------) 的句子
3. 使用 ckip\_transformers 套件斷詞
4. 移除標點符號

## 參數計算方式

1. 句子長度：斷詞完畢後，算出每一個句子含有幾個詞彙
2. 詞頻：參考 ASBC 詞頻資料，如果詞頻大於 100，則視為高頻詞，反之則為低頻詞。最後回傳每一個句子中含有幾個高頻詞、幾個低頻詞。（threshold 暫定 100，再與老師和均亭討論）
3. 詞彙等級：參考國教院詞語分級表，回傳每一個句子中，各級別的詞彙分別有幾個
4. 詞彙長度：算出每一個句子中有幾個「長度大於三個字」的詞彙
5. 搭配詞資訊：還不確定應該如何實作，會詢問老師和均亭
6. 完整的句子：用 spacy 套件取得句子中每一個詞彙的 dependency，若句子同時具備 `nsubj` 和 `dobj`，就視為一個完整的句子
7. 完整語境：用 spacy 套件取得句子中每一個詞彙的 pos，再檢查第一個詞彙的 pos 是否為連接詞 `CONJ` `CCONJ` 或副詞 `ADV`

## [CorpusforDDL\_parameters\_example](https://docs.google.com/spreadsheets/d/122qSBLr-TcZbhGgvcCqyKq3_G06QRq4EPIivpcezcuY/edit#gid=559197399) 欄位說明

1. sentence: 原句子
2. sentence\_preprocessed: 前處理後的句子
3. sentence\_length: 句子長度
4. high\_low\_freq: 高頻詞 (High) 與低頻詞 (Low) 的數目
5. word\_level: 各個詞彙等級的數目，Unknown 代表該詞未出現在國教院詞語分級表
6. long\_word\_count: 句子中「長度大於三個字」的詞彙數目
7. is\_complete\_sentence: 是否為完整的句子，Y 代表是，N 代表否
8. is\_complete\_context: 是否為完整的語境，Y 代表是，N 代表否