

自然語言處理實作期末專題

題目選擇: 詞義解歧(WSD - Word Sense Disambiguation)

第18組 組員: 施泳瑜107071016、王泓文107071006、徐佳靖109073516

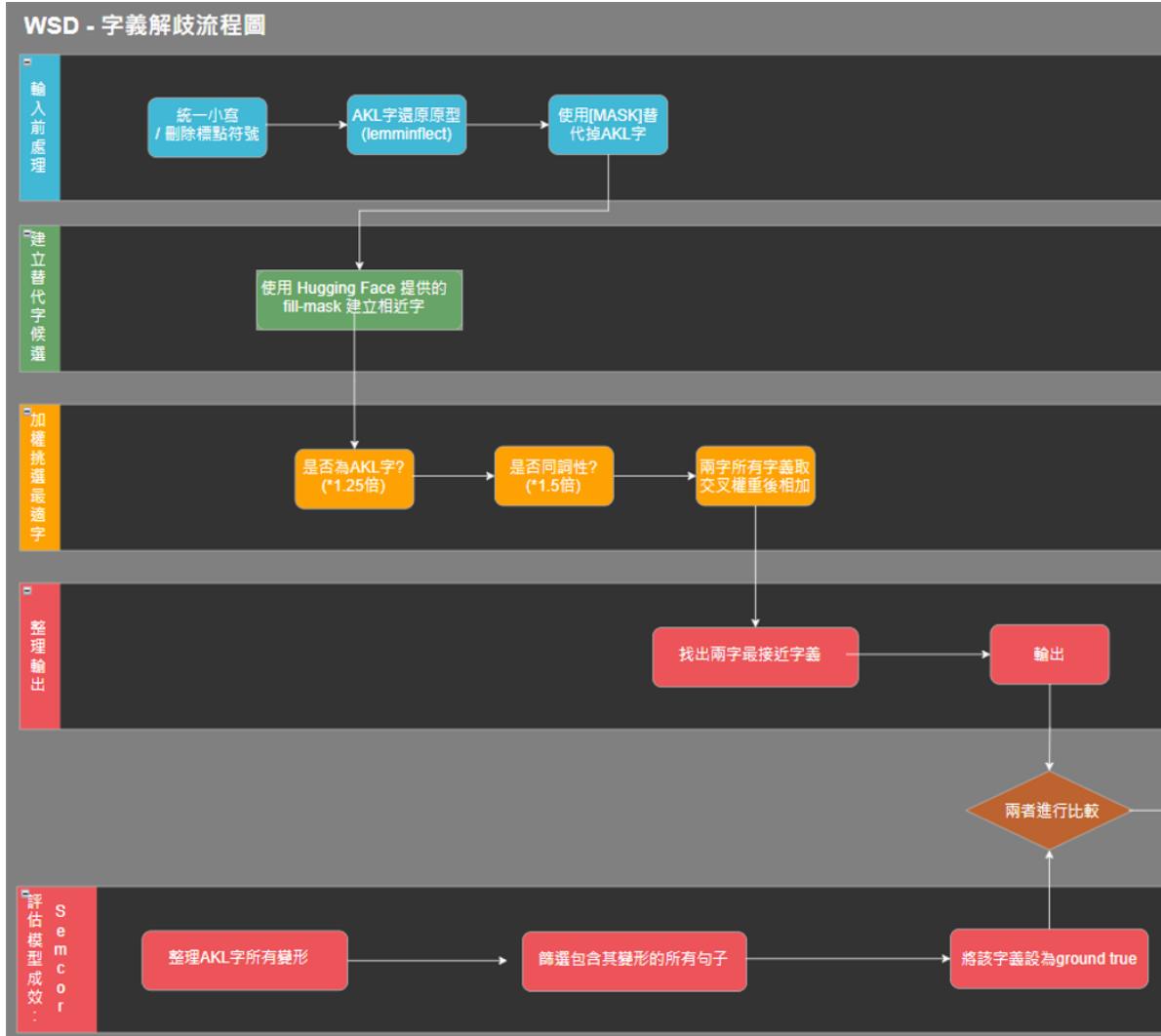
一、摘要

本次期末專題的實作裡，我們選定的主題是WSD(Word Sense Disambiguation)，目標是對於特定AKL字去做詞義解析。

二、介紹

在具體實作上，以Bert pre-train model 為基礎，再延伸到WSD任務上，依此去做特定AKL字詞的字義類別分類，由於沒有伺服器的原庫，本次期末報告是針對15個AKL學術用字著手進行演練，分治法下，在未來可擴增到所有字詞。

模型的部分，主要是以上課所拿到的資料集及上網爬蟲到的paper資料去訓練，其中的輸入是含有關鍵字詞的句子，輸出是透過wordnet交叉比對且加權過後的sense，藉此得到我們的模型，最後再使用到semcor package 去對於我們模型做準確度的測試。而測試集資料採用的是向助教諮詢過後所拿到的論文資料，並利用網站去做呈現及demo。

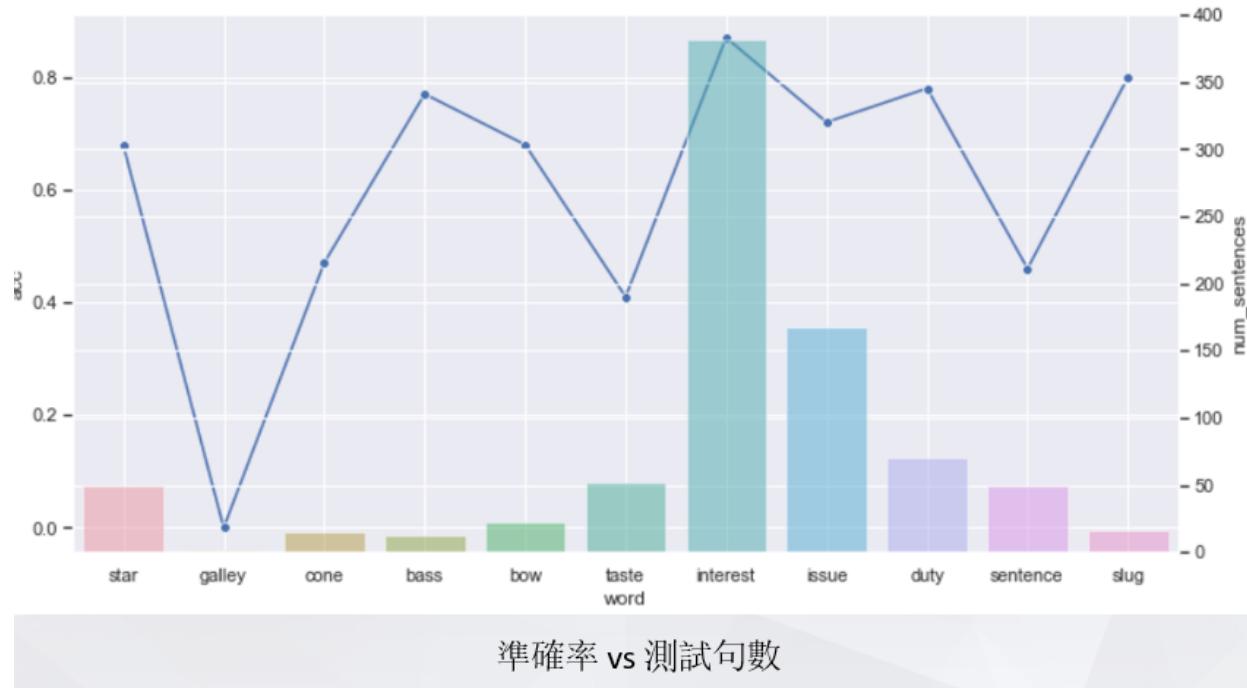


三、方法

如上流程圖所示，首先針對論文文字進行前處理，包含大小寫、標點符號以及詞性變換還原等，再利用Hugging Face的fill-mask找尋適宜的替代字候補，有候補名單後再利用一連串的加權方式，包含是否為AKL字、是否同一詞性、以及比較該字與目標字所有詞性之wup_similarity 相似度平均，以上述方式調整後挑選出認定之最適替代字。

找出最適的替代字後與原先目標字做比對，找出兩字最相近之字義並輸出為分類結果，將該字義與Semcor的正確解答做比對，最後計算準確率。

四、結果



最終之預測準確率使用上述11字測試後平均為0.6，標準差為0.25，使用上述方法之預測準確率跟該字的熱門程度有不小的關係，可以看到準確率最高的interest在Semcor的37000多個句子中共出現382次，較為熱門常用的字也較容易正確分出其字義。

整體準確率不高之原因也由於Semcor本身字義分的相當多元，interest就擁有多達12種不同的解釋，使用Semcor作為字義正解大幅提升了我們做字義分類問題的難度。

word	acc	num_sentences
star	0.68	49
galley	0.00	2
cone	0.47	15
bass	0.77	13
bow	0.68	22
taste	0.41	52
interest	0.87	382
issue	0.72	168
duty	0.78	71
sentence	0.46	49
slug	0.80	16

五、結論

本次專題根據WSD(Word Sense Disambiguation)之作法進行延伸，針對特定AKL學術用字進行詞義解析，實作方法包含：資料前處理、找尋替代字候補、設計加權方式、相似度計算、app呈現等等，最終將所找尋到的字與字義運用Semcor來計算本專題預測之準確率，根據前述專題之說明，最終之預測準確率為0.6，具有一定比例的準確程度，因此在未來可透過本研究所設計之詞義解歧功能來進行學術論文之詞意辨別。

六、未來改進

本次專題主要針對15個AKL字來進行詞義解歧，而AKL總共包含 930 個潛在的學術用詞，即在廣泛的學術文本中相當頻繁但在其他類型文本中相對不常見的詞，因此在後續研究可將15個特定學術用字增加至930個學術用字，使本研究能夠更加準確與完善；此外，本研究所分析之論文共有14篇，因此在未來可以擴充論文與期刊之數量，將詞義解歧之功能擴展到各篇論文與期刊當中；最後可以針對app界面做修正，可將標記有學術用詞的文句更改為可以直接點選並分析字義，以避免手動複製文句之麻煩。

七、參考資料

<https://leemeng.tw/shortest-path-to-the-nlp-world-a-gentle-guide-of-natural-language-processing-and-deep-learning-for-everyone.html>

https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html

https://drive.google.com/file/d/1cBnGKJMZ_sSPQfppIb4a_9VTelGpj32t/view

<https://aclanthology.org/C92-2070.pdf>

<https://github.com/allenai/s2orc>

<https://uclouvain.be/en/research-institutes/ilc/cecl/academic-keyword-list.html>