

以文字探勘技術為基礎的 文本校正方法

-----第19組-----

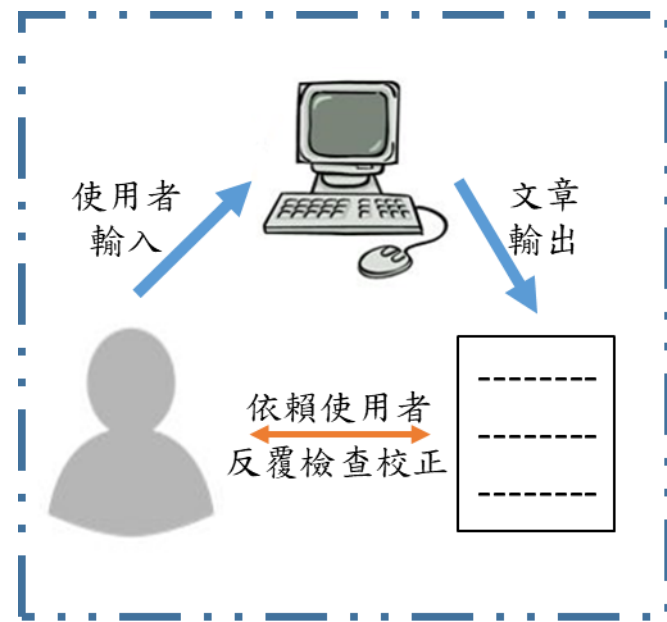
指導老師 | 侯建良 教授

組員名單 | 蔡季翰 賴宣宇

蔡孟桓 謝瑞璟

研究動機

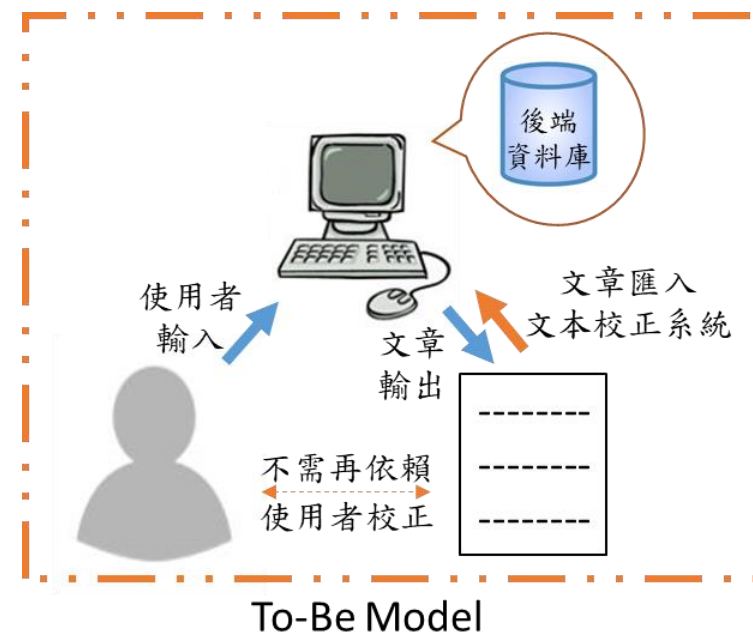
- 許文廉博士在西元1990開發出第一套中文輸入法——「自然智慧型中文輸入系統」，接著中文輸入法不斷加入、改善各種智慧輸入功能，大幅提升了中文輸入的速度。然而不論是自動選字或是聯想字建議等智慧輸入功能如何突破，都很難避免因使用者漏字造成的錯誤。
- 既有模式問題：
 - 1) 使用者常因輸入時漏字造成文本錯誤
 - 2) 要找出錯誤只能靠使用者反覆檢查校正



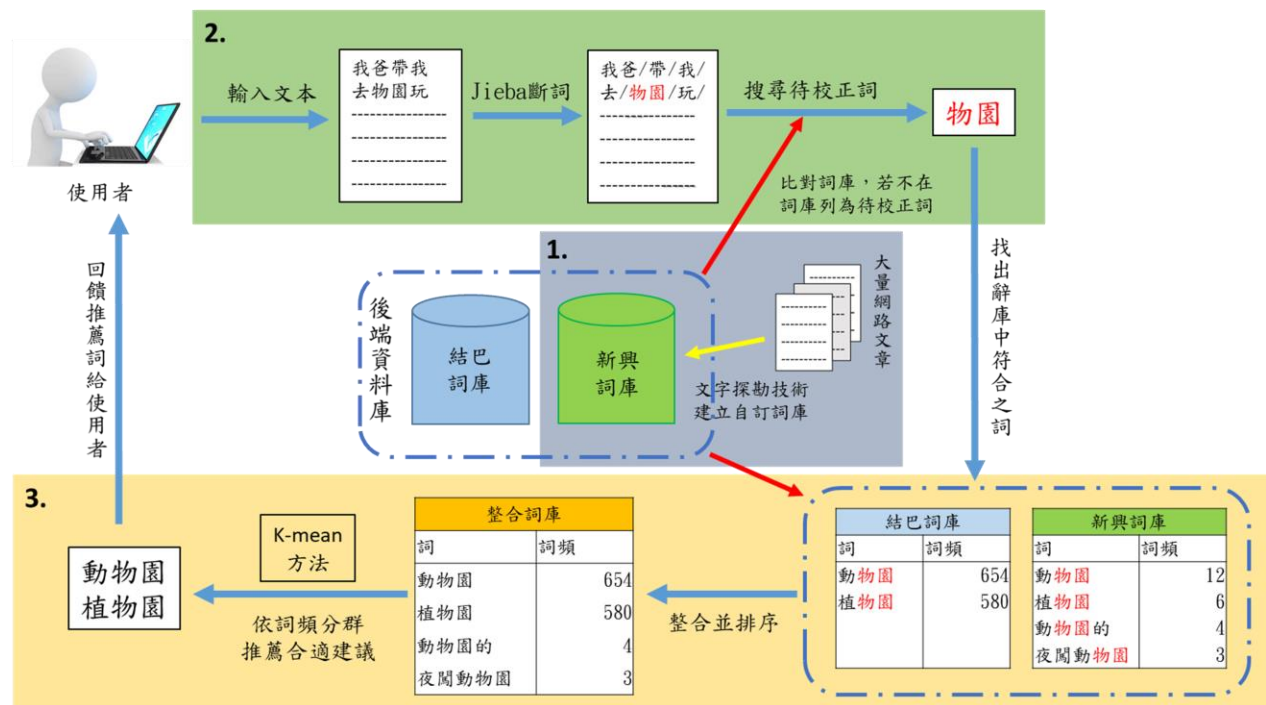
As-Is Model

研究目的

- 本研究之目標為發展一「能找出因漏字造成之文本錯誤」的系統，其能找出文本需要校正之處，並給出建議校正選項給使用者選擇如何修正。
- 期待模式解決方式：發展能解析文本的方法，建立資料庫與之比對，並結合電腦程式，開發「文本校正系統」。



研究架構



可劃分為三大部分

1. 自訂詞庫的建置(灰色區塊)

利用文字探勘技術蒐集大量網路文章建立詞庫，以補充結巴詞庫的新興詞

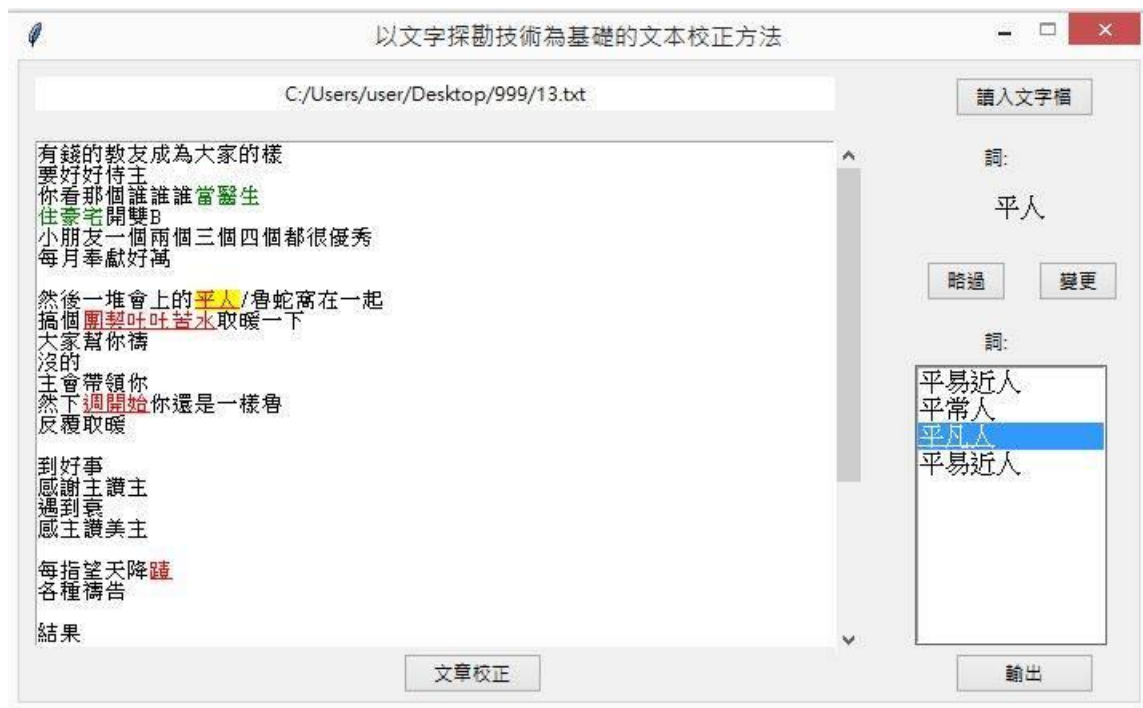
2. 斷詞與正確性的判定(綠色區塊)

將文章斷詞後比對詞庫，若該斷詞不在詞庫，則認定該詞為待校正詞

3. 待校正詞的修正運作(黃色區塊)

在詞庫中尋找與待校正詞相符的詞，依詞頻利用K-mean法分群，找出頻率較高的群並認定其為推薦詞輸出，供使用者選擇。

系統開發

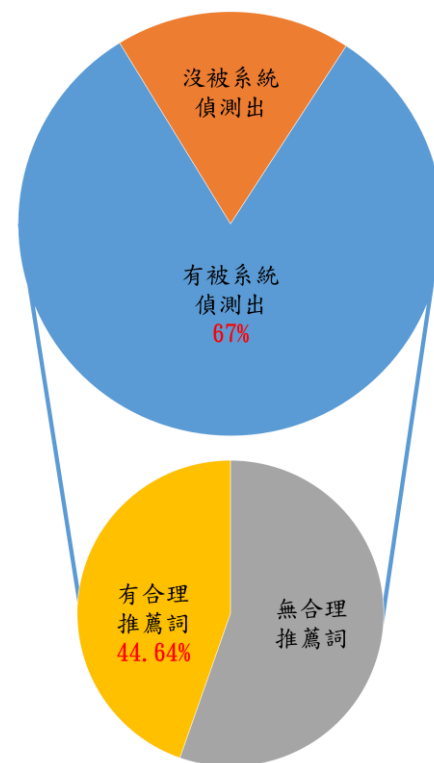


1. 可選擇以輸入或匯入的方式校正文本
2. 待校正詞者以紅色標註，正在校正之詞黃底標註，已修正之詞以綠色標註
3. 可選擇要修正為哪個詞，或略過建議校正其他詞
4. 文本校正完畢可輸出

績效分析

本研究以50篇自網路抓取的文章測試系統正確率，其中共含167個待校正詞，將其匯入系統並分析結果：

1. 系統偵測出112個待校正詞，佔67%
2. 其中有合理推薦詞者共50個，佔44.64%
3. 有合理的建議校正選項者平均在第1.5順位



結論與未來展望

- 結論：本研究應用以上所述之方法，發展文本校正系統。研究成果由績效分析結果顯示，有近7成的錯誤能被系統偵測出，其中也有近5成系統能正確給出合理的推薦詞，並平均能在第1.5順位內顯示給使用者。可推論本研究所開發之系統能有效提升文本校正的效率及品質。
- 未來展望：
 - 繼續擴建系統後端資料庫：以判斷更多新興詞
 - 改善斷詞演算法：避免因斷詞不準確造成誤判待校正詞的情形
 - 加入使用者經驗：將使用者經驗計算詞頻，使系統更貼近使用者
 - 與輸入法結合：在輸入的同時即進行校正判斷，增加即時性與便利性