基於主動式學習的古漢語文本斷句系統

學生:徐志帆　　指導教授:陳志銘

研究背景

古漢語斷句(或稱句讀)是中文書寫系統中一個經典的議題。

將文本內容切成**句子**(sentence)以及**子句**(clause)，辨識句子的邊界稱為「句」，而上述的句子再細分各子句則稱為「讀」。

判斷斷句仰賴閱讀者的經驗知識，過程費時，如果有自動化工具能快速初步解讀斷句，後續由專家校對調整，就能大幅降低時間和人力成本。

目前古漢語文本的**自動化斷句**方法主要區分為**規則方式**和**機器學習方式**，規則方法過於複雜且難以泛用，主流為機器學習方法。

研究背景

機器學習方法利用統計演算法和已標註資料建立學習模型，再透過模型進行斷句標註判斷，此一方法在某些文本中具有很不錯的辨識準確率。

中文書寫系統發展已久，不同時代具備不同文體，通用型的自動斷句方法難以實現。且不會再產生新的古漢語文本，如何有效率的建立標註資料是重要的議題。

主動式學習(Active Learning)是機器學習中用於解決學習過程需要大量人工訓練資料的方法，其概念透過人工協助電腦提出的問題建立訓練資料，能提高訓練語料的品質，降低其量的需求。

研究背景

主動式學習在自然語言處理中已經有相當廣泛的應用，但卻少有在古漢語斷句上的相關研究。

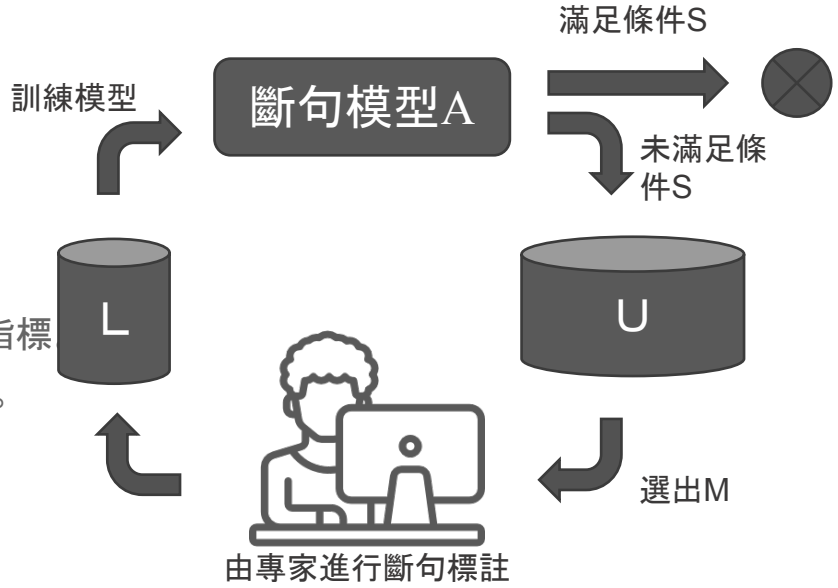
本研究欲發展結合主動式學習以及斷句模型的「基於主動式學習的古漢語文本斷句系統」，透過人機合作模式降低建立模型時所需的訓練語料。

本研究也將透過邀請專家使用「基於主動式學習的古漢語文本斷句系統」進行古漢語斷句，並分析結果以及從專家取得改進建議。

主動學習方法

定義：已標註資料L，未標註資料U，優先確認區塊M，結束條件S

1. 設定結束條件S為完成全部文本的斷句標註。
2. 建立已標註資料L
3. 用L訓練斷句模型A。
   - 3-1. 反覆執行3-1~3-5項直到滿足結束條件S。
   - 3-2. 使用斷句模型A計算U中個單字詞的斷句之不確定指標及各段落的區塊斷句不確定指標。
   - 3-3. 根據區塊不確定指標並從U選出優先確認區塊M。
   - 3-4. 透過人工確認M，完成後將M移入L中。
   - 3-5. 使用L訓練斷句模型A，回到2。



系統架構

資料處理階段

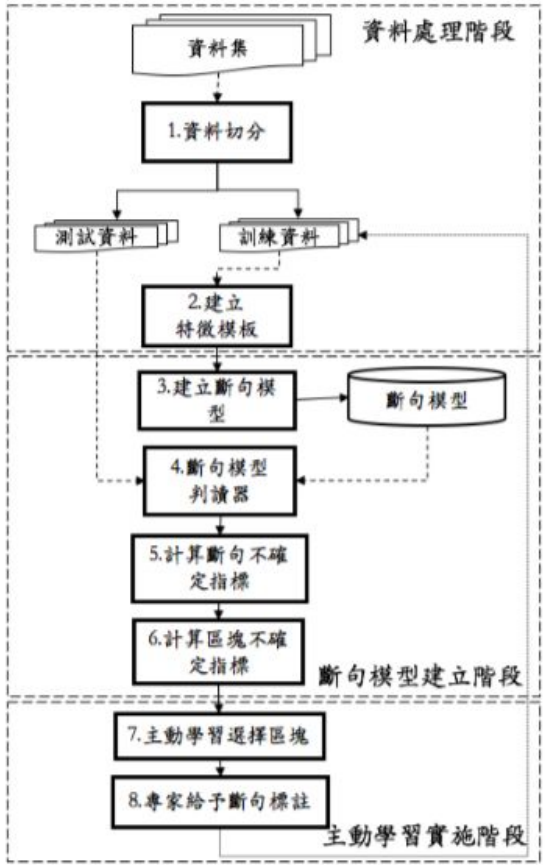
將文本處理成建立斷句模型所需的格式，分為 資料切分 以及 建立特徵模板 兩個部分。

斷句模型建立階段

以演算法建立斷句模型，並會用迭代方式進行數回合的訓練與測試，計算文本中的單字詞 斷句不確定性指標 以及 區塊不確定性指標 。

主動學習實施階段

使用 主動學習 選擇方法 選擇文本區塊，並由專家對該區塊給予斷句標註，完成後將該區塊加入到下一回合的訓練資料中。



特徵模板與演算法之評估實驗設計

	依序文本組		主動學習組
演算法	**隱性馬可夫模型**	**條件隨機場**	**雙向長短詞詞神經網路模型**
特徵模板	無	三字詞特徵模板 二字詞特徵模板	無
選擇文本區塊	**依序文本**		**主動學習**
建立斷句標註	給予已知文本		

貝氏邏輯回歸
最大熵

使用文本

維基文庫版的「峴泉集」

條件隨機場三字詞 + 特徵模板

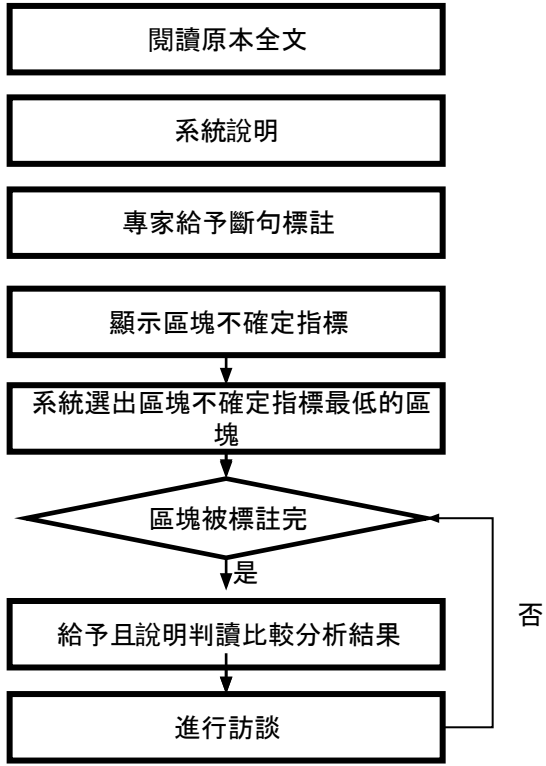
主動式學習斷句系統之預測效能評估實驗設計

實驗對象

六位具備古漢語知識的人文學者。

使用文本

維基文庫版的「峴泉集」



基於主動式學習的古漢語文本斷句系統

- 採用python+javascript
- 具備CRF(條件隨機場)的機器學習演算法
- 具備不確定抽樣方式的主動學習文本選擇模式
- 顯示預測標註
- 可以自行上傳文本
- 可以輸出(儲存)斷句結果

基於主動學習的斷句解讀工具

實際操作

未來方向

- 加入筆記、交互參照、外部工具查詢功能
- 加入會員系統
- 自動儲存、雲端儲存
- 調整介面設計
- 與本實驗室的數位文本資料介接

Q&A