圖資與檔案學刊
11:2=95 (Dec '19) 117-145
DOI: 10.6575/JILA.201912_(95).0004

基於主動式學習之古漢語斷句系統發展與應用研究

徐志帆* 張 鐘**

【摘要】

本研究旨在開發支援數位人文研究之「基於主動式學習的古漢語文本斷句系統」，結合主動學習與機器學習演算法，透過人機合作模式降低建立自動化古漢語斷句建立模型時所需的訓練語料，並協助人文學者面對未解讀過的文獻能更有效率的進行斷句判讀作業。為了找出最合適建立「基於主動式學習的古漢語文本斷句系統」的的演算法與特徵模板，本研究設計第一個實驗採用了不同的演算法與特徵模板配合依序文本和主動學習兩種選擇文本方法所建立的斷句模型進行比較。實驗結果發現，條件隨機場（conditional random fields）與三字詞特徵模板在主動學習方法中能有效地進行學習，適合發展「主動學習斷句模式」。第二個實驗邀請人文專長領域的學者使用「基於主動式學習的古漢語文本斷句系統」進行古漢語文本的斷句判讀，以人文學者各自標註資料建立的斷句模型進行比較分析，並輔以半結構式訪談深度了解人文學者對於本研究發展之系統輔以斷句的使用感受與建議。實驗結果發現「基於主動式學習的古漢語文本斷句系統」確實能有效學習人文學者的斷句標註資料，並且模型預測能力能基於人機合作而不斷提升。最後，透過訪談結果歸納得知人文學者對於系統操作流程與介面具有正面評價，多數受訪者認為本系統的斷句預測功能在古漢語斷句上能提供有效之輔助功能。未來可考量增加命名實體模型或其他古漢語規則的特徵模板設計，以進一步提升斷句預測能力，也希冀能將發展的系統運用在人文領域教育上，發展為訓練古漢語斷句之數位人文教育平台。.

關鍵詞

數位人文 主動學習 機器學習 自動化古漢語斷句 人機互動

* 國立政治大學圖書資訊與檔案學研究碩士生
E-mail: billxu0521@gmail.com
** 國立政治大學華人文化主體性研究中心資訊工程師
通訊作者 E-mail: foxx1216@gmail.com

壹、緒論

古漢語斷句（或稱句讀）是中文書寫系統中一個經典的議題，在西元 1919 年之前沒有統一的標點符號規則，大多數文本中沒有句讀，由讀者依據文意、經驗以及語感將其切割成句子（sentence）以及子句（clause）來理解文本的意思，而理解文字中的意思又對斷句解讀有幫助，因此斷句解讀與理解文意彼此乃相輔相成。在斷句解讀實務中會以字詞和語法規則來判斷文本的斷句，再依照文意給予對應的標點符號。

正因為斷句判斷仰賴閱讀者的經驗以及知識，解讀過程往往相當費時，若使用自動化工具則可以提高解讀效率。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兩種（Huang, 2017），其中規則方式是以人工歸納古漢語文本的句法規則，利用電腦判讀內容的斷句位置。Huang （2008）的研究利用句法規則建構了自動化斷句系統，並將其應用在農業古籍中輔以進行斷句標點判斷，此類方法受限於文本類型的規則，因此泛用性較差。而機器學習方法則係利用不同統計演算法和人工標註的訓練語料建立學習模型，使電腦得以依靠所訓練之模型進行古漢語文本的自動斷句，此一方法在某些文本中具有很不錯的辨識準確率，所以目前機器學習方法已成為古漢語自動斷句的主流方法。

但古漢語文本的自動化斷句在實務上還是有些限制，主要係因為中文書寫系統已有數千年之久，前後發展具有相關性，但每個時代的文體、字詞用法皆有所差異，使得通用型的自動斷句方法難以實現。而且發展古漢語文本斷句學習模型的效能，容易受到訓練語料的文字量、特徵品質影響（Huang, 2010）。此外，由於時代會持續演進，古籍不會再產生新的文本，使得如何有效率地建立模型和降低辨識錯誤率變成重要的研究議題。主動式學習(active learning)是機器學習中一種用於解決學習過程中需要大量人工訓練樣本的方法，其概念在於透過人工判斷方式提高訓練語料的品質，進而達到只需較低訓練語料量就能有不錯的正確率目標。Settles（2012）的研究中提到像是語音識別（speech recognition）、訊息萃取或是一些分類過濾的問題中，可以透過主動學習取得使用少量的高品質標註資料來實現高精準的模型。雖然主動式學習在自然語言處理中已經有相當廣泛的應用（Olsson, 2009），而在中文分詞（Liang，2015）、古籍的命名實體識別（Yeh, Wang, & Tsai, 2011）上也皆有不錯的成果，但在古漢語斷句上卻少有相關研究。

綜合上述，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」，希望能透過人機合作模式降低建立古漢語斷句模型所需的訓練語料。首先會以文獻探討討論古漢語自動化斷句的發展，並且於實驗設計中討論本研究如何結合特徵模板與人工智慧演算法的「主動學習斷句模式」運用在「基於主動式學習的古漢語文本斷句系統」中，並邀請人文領域專家進行實驗以「基於主動式學習的古漢語文本斷句系統」完成古漢語斷句標註，並且分析其預測能力是否符合預期，最後在結論討論透過人文領域

專家使用之經驗，取得改善「基於主動式學習的古漢語文本斷句系統」預測準確率之建議，作為持續發展即時斷句標註工具支援數位人文研究之用。

貳、文獻探討

一、古漢語文本的自動化斷句

古漢語文本的自動化斷句是根據古漢語中句子的組合規則和詞義等資料，透過電腦處理實作出自動在古漢語文本上判斷句讀的功能。目前主要區分為規則方法和機器學習方法兩種（Huang, 2017），規則方式需要經過人工歸納句法、用詞規則以及斷句模式，較難將研究結果運用在其他領域的古漢語文本上或是運用在大規模文本的整理上，故目前主流為透過已有斷句標註的訓練語料結合演算法建立模型的機器學習方法。常見的相關演算法有單純貝氏分類器（Naive Bayes classifier）、羅吉斯迴歸（Logistic regression）、最大熵模型（Maximum Entropy model）等演算法（Liu, 2017）。

Huang（2010）運用序列標註概念將古漢語文本自動斷句視為自然語言處理（natural language processing）的中文分詞延伸，結合抽取文本中的特徵並搭配隱性馬可夫模型（hidden Markov models, HMM）以及條件隨機場（conditional random fields, CRF）建立斷句模型後，將其運用在古漢語文本斷句上，有相當不錯的成效。在該研究中也發現條件隨機場的整體表現略優於隱性馬可夫模型。Zhang（2009）時發展了一套基於條件隨機場的古漢語文本自動斷句方法，在《論語》和《史記》兩個文本中得到近 8 成的預測正確率。

另外也有人在探討使用深度學習的神經網路進行古漢語文本的自動斷句，例如 Wang（2016）設計了神經網路語言模型（neural network language model），結果顯示在古漢語文本自動斷句問題中可達到與條件隨機場相當的成效。而 Wang（2017 和 Hu（2016）的研究中使用了雙向循環神經網路（bidirectional recurrent neural network）模型在古漢語文本句讀上，在實驗結果中顯示其表現不輸條件隨機場，而且比起神經網路（neural network）的斷句模型更有效率，證明了深度學習在古漢語自動化句讀發展上的可能性。

古漢語文本自動斷句方法的機器學習方法通常採用監督式學習，意思是使用大量的標註資料，以統計方法歸納其斷句的規則，並進行參數學習、萃取特徵經過訓練後建立斷句模型，之後再使用此模型去判斷文本中的斷句標註位置。此方法若訓練資料量不夠多或是太多相同的標註資料，就可能導致訓練資料品質不佳，而份量太少或是品質不好的訓練資料則會使得建立出來的模型預測效能下降（Huang, 2010）。但與現代的白話文不同的是古籍的數量不會再增加，而且古漢語的人工斷句標註成本高，難以產生大量的訓練資料。另外，中文發展已有數千年之久，語言發展雖然承先啟後，但每個時代的文體、所用字詞皆

有差異，使得通用型的自動化斷句方法難以實現。例如某個詞在文本中有時是口語用詞，有時又是一般敘事用詞，文本交雜著兩者用法，會導致詞性判斷困難，甚至代表命名實體的詞彙也不易辨識，這些都是機器學習方法需要面對的問題。（Hu, 2016）因此提出若能讓機器與人互相配合，透過人工提供斷句標註上的幫助，提高訓練資料品質可降低預測過程中產生錯誤的機率，提高預測結果的正確性，應是更可行的做法。

二、主動式學習

主動學習意圖使用較少量的標註資料來提高模型的精準度（Settles, 2012）。相較於傳統的監督式學習方法，主動式學習具有能夠保持品質的處理大量訓練資料集，從中選擇出有辨識度的樣本，並且減少訓練集的數量和減少人工標註成本等優點（Liu, 2012）。在主動學習中模型（learner）會主動提出標註詢問（request），以取得解答（oracle）來提高已標註資料的精準度，減少模型在訓練過程中的訓練資料需求。實作方案（scenarios）可以區分為 membership query synthesis、stream-based selective sampling、pool-based sampling 三種類型（Settles, 2009）。在 membership query synthesis 方案中，模型（learner）會在所有輸入空間（input space）尋找最需要答案（oracle）的樣本，並重新組織且提出詢問，這種做法具備彈性，但是會導致詢問的內容不合乎自然語意，使得人類難以解讀。另外兩種方式相對適合應用在需要人類判讀的問題上，其中 stream-based selective sampling 是模型依序根據每一次的輸入資料選擇是否要提出詢問，判斷標準取決於採用的查詢策略（query strategy），這種方案相對適合於由人類給予答案的狀況。若依序取用輸入的樣本提出標註詢問，則稱為 pool-based sampling，此種方案兼具上述兩者的特性易於使用，是主動學習中最泛用的一種形式。

Settles（2009）的研究指出較常使用的查詢策略為不確定性抽樣（uncertainty sampling）和委員會查詢（query-by-committee）。在不確定抽樣方法中，已學習過的分類模型會在判讀完未標註資料後，提出最有用（the most informative）或者最不確定（uncertainty）的實例（instances）詢問，並由人工確認（Olsson, 2009）。而委員會查詢方法中，是透過不同的統計演算法再透過一組已標註資料訓練出一組分類模型集合，並由該集合對未標註資料進行判讀，並從中選擇標註最不一致的資料作為實例詢問。Sassano（2002）在日文的序列分割問題中導入了主動式學習架構，其研究採用了 pool-based sampling 搭配不確定性抽樣方法，最後成功減少達到分類器目標性能時所需的訓練語料。Li（2012）的研究中改進了不確定性採抽樣方法，加入多樣性測量（diversity measurement），透過字詞邊界註釋（Word Boundary Annotation）模型發展出一種主動式學習策略，成功地應用在中文分詞問題上，大幅減少已標註資料的成本。該方法會計算每一次斷句模型預測未標註資料後各字詞

（word）的不確定值，並依照大小進行排序選出前 200 個字詞標註，完成後將標註的字詞移到已標註資料集中，重複上述流程反覆訓練模型提高精準度。

參、基於主動學習的古漢語文本斷句系統

本研究監督式學習作法提出「基於主動學習的古漢語文本斷句系統」，建立過程包括資料處理階段、斷句模型建立階段，以及主動學習實施階段三個部分，如圖 1 所示。其中資料處理階段的主要目的是將古漢語文本整理，並且處理成可以建立斷句模型的格式，再區分為資料切分以及建立特徵模板兩個部分。本研究提出之「主動學習斷句模式」將在建立斷句模型階段實施，透過演算法建立斷句模型，系統會以迭代方式進行數回合的建立斷句模型階段，再以斷句模型判讀器讀取輸出的斷句模型，並且計算文本中的中文字斷句不確定指標，以及區塊不確定指標。最後，主動學習實施階段係採用主動學習選擇方法選擇文本區塊，對區塊內未標註的資料由專家給予斷句標註，完成後將該區塊加入到下一回合的訓練資料中，直到所有未標註資料被標註完畢為止。

一、資料處理階段

在資料處理階段會先依據文本的特性進行資料集的分段，將文本資料分成數個區塊，並建立序列標記集。一部原始的古漢語文本雖然沒有標點符號隔開句子，但可能具備段落的架構，或者是由數卷或數篇集結而成，在本階段會根據此特性將資料集大抵分為不同區塊，每個區塊會加上開始與結尾記號。

本研究採用 Hu（2016）研究中採用的 N 和 S 兩種標註標示斷句，N 表示不斷句的字，S 表示需要斷句的字。以論語學而篇開頭「學而時習之」為例子，首先會在句子前後加上 START 和 END 標記並切分子句，而「學而時習之」的「學」、「而」、「時」、「習」為不斷句的字，因此標註為 N；「學而時習之」的「之」字為需斷句的字，因此標註為 S。說明如下：

學而時習之，不亦說乎？有朋自遠方來，不亦樂乎？人不知而不慍，不亦君子乎？

首先，加上句子開始結束的標記和切分子句：

START/學而時習之/不亦說乎/有朋自遠方來/不亦樂乎/人不知而不慍/不亦君子乎/END

照前述方法標記，N 表示不需斷句，S 表示要斷句：

START$_0$/S 學$_1$/N 而$_2$/N 時$_3$/N 習$_4$/N 之$_5$/S 不$_6$/N 亦$_7$/N 說$_8$/N 乎$_9$/S……不$_{26}$/N 亦$_{27}$/N 君$_{28}$/N 子$_{29}$/N 乎$_{30}$/S

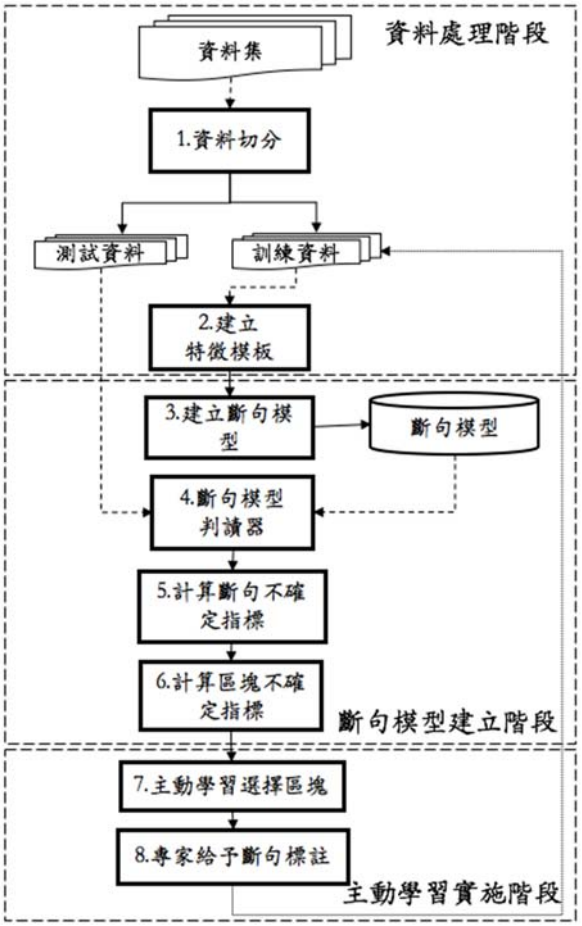

圖 1　基於主動學習的古漢語文本斷句系統建模流程圖

將每個字加上斷句標籤後將進行特徵模板的處理，本研究參考 Hu（2016）的研究中所用的二字詞特徵模板以及擴展開來的三字詞特徵模板。

特徵模板如表 1、2 所示，其中 C_i 為文本中第 i 個字，取前後 k 個位置的字表示為 C_{i-k}~C_{i+k}，並使用字長度為 n 的 n-gram 選取字的範圍，以「學而時習之不亦樂乎」的「習」字舉例，三字詞特徵模板的 n-gram 選用單字詞（unigram）、二字詞（bigram）、三字詞（trigram），而 k=3，將具有「學」、「而」、「時」、「習」、「之」、「不」、「亦」、「學而」、「而時」、「時習」、「習之」、「之不」、「不亦」、「學而時」、「而時習」、「時習之」、「習之不」、「之不亦」18 個特徵數；二字詞特徵模板的 n-gram 選用單字詞、二字詞，而 k=2，具有「學」、「而」、「時」、「習」、「學而」、「而時」、「時習」7 個特徵數。

表 1

三字詞特徵模板

文字片段	學而時習之不亦樂乎						
文字	學	而	時	習	之	不	亦
編號	C_{x-3}	C_{x-2}	C_{i-1}	C_i	C_{i+1}	C_{i+2}	C_{i+3}
Unigram	學、而、時、習、之、不、亦						
Bigram	學而、而時、時習、習之、之不、不亦						
Trigram	學而時、而時習、時習之、習之不、之不亦						

表 2

二字詞特徵模板

文字片段	學而時習之不亦樂乎			
文字	學	而	時	習
編號	C_{i-1}	C_i	C_{i+1}	C_{i+2}
Unigram	學、而、時、習			
Bigram	學而、而時、時習			

二、建立斷句模型階段

在建立斷句模型階段中將會使用本研究提出的「主動學習斷句模式」針對文本資料進行特徵萃取，本研究根據文獻探討所提到五種常用於自然語言處理的演算法：條件隨機場、雙向長短記憶詞模型、單純貝氏、羅吉斯回歸、最大熵模型來建立斷句模型，在機器學習或是深度學習中，首先，會將已經有斷句標註的古漢語文本作為訓練資料（training data），並且進行語言特徵處理後，再利用演算法計算特徵的權重，以取得參數建立斷句模型。而未有斷句標註的古漢語文本作為測試資料（esting data），透過斷句模型的解讀與判斷，將測試資料輸出成具有斷句標註的古漢語文本。在這個過程中將會計算每個字的斷句不確定指標。

每次利用斷句模型判斷過程中會計算每個字的斷句不確定指標，本研究參考 Culotta & McCallum（2005）對序列模型提出的不確定性抽樣方式，稱為最小信心計算法（least confidence）來計算斷句不確定指標，其定義如式 1：

$$f(x_i) = \max_{y \in [0,1]} 1 - P(y|I_i) \qquad \text{式 1}$$

其中 x_i 表示資料集中第 i 個位置的字；$p(y|I_i)$表示字 x_i 的右邊界標註為 y 的條件機率；y=0 表示 x_i 後沒有斷句符號，反之 y=1 表示 xi 後有斷句符號。由於本研究中使用二元分類標註，不確定性抽樣會取樣 $f(x_i)$越接近 0.5 的例子，表示模型越不具信心，越有可能在這個位置上是錯誤標註，也就是說模型越無法判定右側是否有斷句符號。

再來將各文字的斷句不確定指標乘上各區塊內的字數後取個區塊的不確定指標平均值，稱為區塊不確定指標，越接近 0.5 表示模型對於該區塊越不具信心判斷其文本內容，來評量選擇下一回合所需要由人類專家優先協助標註的文本區塊。

三、主動學習實施階段

主動學習實施階段會在每一回合的斷句模型判讀斷句標註後，計算各文本區塊的區塊不確定指標並進行排序，再推薦指標最高的文本區塊給專家進行斷句標註，再將其加入到下回合建立斷句模型所使用的訓練資料中。

四、系統介面與功能介紹

本研究所發展「基於主動學習的古漢語文本斷句系統」的使用者介面與功能，能提供古漢語文本的斷句標註，整體系統介面如圖 2 所示，各功能介面說明如下:

圖 2　基於主動學習的古漢語文本斷句系統整體介面

（一）文本區塊資訊

使用者可於此區塊取得目前正在進行斷句標註的回合數、文本區塊內容的字數，以及區塊不確定抽樣分數之資訊。

（二）文本區塊列表

顯示目前回合進行標註的文本區塊，以及其他文本區塊數。

（三）使用者資訊

可於此欄位輸入目前進行標註的使用者代稱，此代稱將會記錄於系統中。

（四）文本顯示區

使用者可以在文本顯示區介面中閱讀古漢語文本，以及進行斷句標註。如圖 3 所示將滑鼠游標放置到某一文字上時，後側會出現紅色虛線框表示可於此加入斷句標註，使用者可以點選此字加上斷句標註，若已有斷句標註，則點選後會移除斷句標註。另外，游標放置文字上時也會顯示文字在文本區塊中的位置。

圖 3　文本顯示區介面

（五）文本抽樣功能

點擊此按鈕會依照目前的標註資料進行訓練，以建立斷句預測模型，並透過此模型從尚未進行斷句標註的文本區塊中，選出區塊斷句不確定性最高的區塊，並且進入下個標註回合。

（六）預測文本功能

預測文本功能可以透過前回合的斷句模型進行斷句標註的預測，並且顯示在文本顯示區中，如圖 4 所示，由模型預測的斷句標註以綠色虛線框顯示於文字後側。

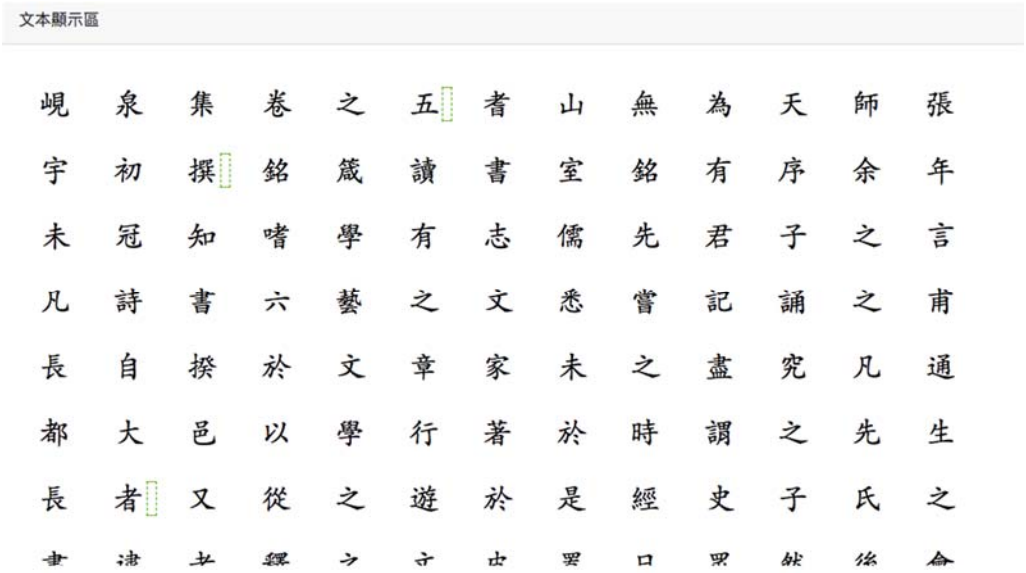

圖 4　顯示預測標註

（七）儲存標註結果功能

將目前所有回合的標註結果，以文字格式儲存後輸出檔案。

（八）切換顯示文章題目

開啟或關閉文本顯示區中標亮的文章題目，如圖 5 所示，會以淺藍色背景顯示在文字後。

圖 5　顯示文章題目

（九）切換顯示預測標註

開啟或關閉文本的顯示預測標註功能。

肆、實驗設計

一、特徵模板與演算法之評估實驗設計

以下說明本研究如何選出用於「主動學習斷句模式」的特徵模板和演算法之評估實驗設計。

（一）選用資料集

本研究實驗採用之文本為維基文庫版本的《峴泉集》共有 11 卷，約 8 萬字，屬於近代漢語（Proto-Mandarin）的文本，每卷收錄不同類型文體的集結，例如傳記、賦、五言絕句、七言絕句等不同文體，是一套具備豐富語料的古漢語語料集，由於考慮到給予人文學者進行古漢語標註的使用，需在有限的時間內完成，因此本研究將《峴泉集》每卷隨機刪減內文篇數至 1 萬餘字。

（二）實驗流程

圖 6 為特徵模板與演算法之評估實驗流程圖，根據圖 6 所示，本研究將依序使用二字詞特徵模板以及三字詞特徵模板結合單純貝氏分類器、條件隨機場、雙向長短記憶模型、最大熵模型、羅吉斯回歸五種演算法，以及搭配依序文本選擇文本區塊及主動學習選擇文本區塊兩種文本方法來使用資料集建立斷句預測模型，並根據文本區塊數進行多回合的模型建立與測試，最後以整體斷句結果之平均 F-measure 和 F-measure 變化斜率進行比較分析，作為評估選出何種特徵與演算法組合應用於「主動學習斷句模式」中。從斷句預測結果與原文本中的斷句比對能分為四種可能的結果：真陽性（true positive, tp）、真陰性（true negative, tn）、偽陽性（false positive, fp）、偽陰性（false negative, fn），其中精確率（precision）是指文本中，所有被正確判斷出的斷句數，佔文本中所有被判斷出斷句數的比例，定義見式 2；召回率（recall）是文本中，所有被正確判斷出的斷句數，佔所有應該被判斷出斷句數的比例，定義見式 3。根據式 4，F-measure 是由精確率（precision）和召回率（recall）兩種數值的綜合評價值，分子為召回率與精確度相加乘以 2，分母為召回率與精確率相加。如果 F-measure 有較高的表現，通常代表 recall 和 precision 有較佳的分數。

$$\text{recall} = \frac{tp}{tp + fn} \qquad \text{式 2}$$

$$\text{precision} = \frac{tp}{tp + fp} \qquad \text{式 3}$$

$$F - measure = \frac{2 \times (recall + precision)}{recall + precision} \qquad \text{式 4}$$

在實驗中依序文本組係以分段迭代的方式建立斷句模型，並對測試資料進行斷句標註判讀，依序採用三字詞特徵模板和二字詞特徵模板，以及條件隨機場、雙向長短記憶詞模型、單純貝氏、羅吉斯回歸、最大熵模型等五種模型演算法，合計共 9 種組合。採用依序文本方法選擇文本區塊，並且給予已知文本方式建立斷句標註資料，將循序建立斷句模型，直到所有文本區塊被標註完。由於斷句模型的特性，每次訓練與測試的結果可能會有些微差異，為求更準確的實驗分析，每回合都會隨機進行 10 遍，所有結果採用整體斷句結果之平均 F-measure 和 F-measure 變化斜率作為比較分析依據。此外，主動學習組也採用同樣的方式建立斷句模型，並對測試資料進行斷句標註判讀，依序採用三字詞特徵模板和二字詞特徵模板三組特徵組合，以及條件隨機場、單純貝氏、羅吉斯回歸、最大熵模型五種

模型演算法和雙向長短記憶詞模型結合單字詞特徵模板，合計共 9 種組合。採用主動學習方法選擇文本區塊，並且給予已知文本方式建立斷句標註資料，將循序建立斷句模型直到所有文本區塊被標註完。所有結果採用整體斷句結果之平均 F-measure 和 F-measure 變化斜率與依序文本組組結果比較，選出平均 F-measure 和 F-measure 變化斜率最佳的組合後，將其應用於「主動學習斷句模式」，並作為下一節實驗設計之用。

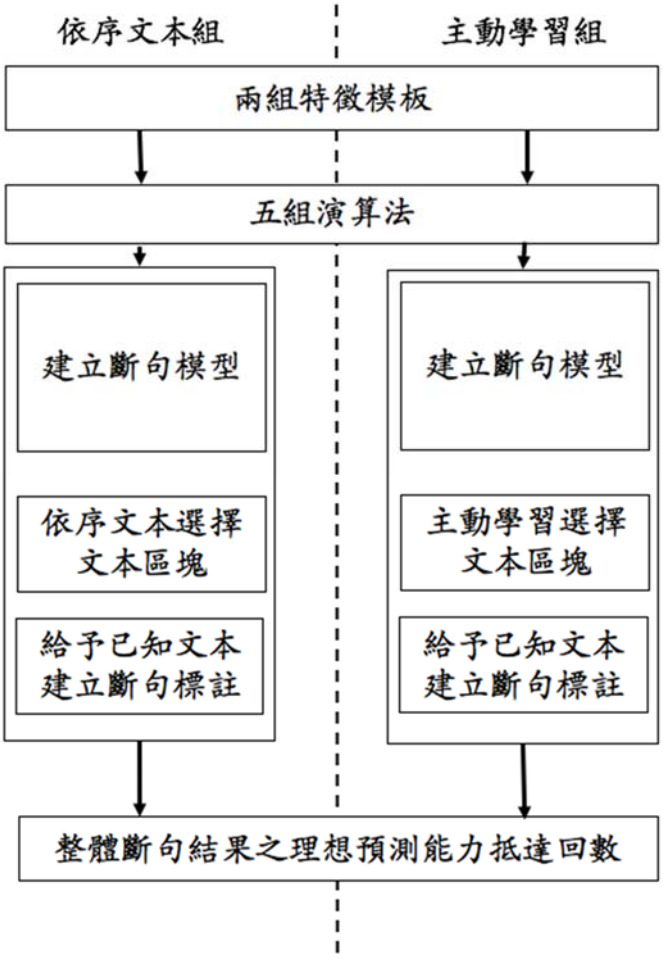

圖 6　特徵模板與演算法之評估實驗流程圖

二、基於主動學習的古漢語文本斷句系統評估實驗設計

本節說明基於主動學習的古漢語文本斷句系統之預測效能評估實驗設計，詳細說明如下。

（一）選用資料集

本研究選用已經過解讀具備句讀標註之維基文庫版本的《峴泉集》，作為本研究實驗使用的古漢語文本，維基文庫版本的《峴泉集》共有 11 卷，約 8 萬字，每卷收錄不同類型文體的集結，例如傳記、賦、五言絕句、七言絕句等不同文體，是一套具備豐富與料的古漢語語料集，很適合作為本研究發展「基於主動學習的古漢語文本斷句系統」所需之古漢語文本。由於考慮到給予人文學者進行古漢語標註的使用，需在有限的時間內完成，因此本研究將《峴泉集》每卷隨機刪減內文篇數至 1 萬餘字。

（二）主動學習斷句模式

在主動式學習斷句系統之評估實驗中，主動學習斷句模式將選擇前一節實驗結果預測效能最佳之特徵模板與演算法，並且將其應用於「基於主動學習的古漢語文本斷句系統」進行斷句模型的建立。

（三）實驗流程

本研究欲了解給予人文領域專家使用「基於主動學習的古漢語文本斷句系統」進行古漢語斷句後的模型是否符合預期，在實驗中會比較人文領域專家之主動參考組透過「基於主動學習的古漢語文本斷句系統」在古漢語文本進行斷句判讀的預測能力，並且將實驗結果輔以實驗對象進行訪談，以收集人文學者對於「基於主動學習的古漢語文本斷句系統」輔以斷句的建議與看法，主動式學習斷句系統實驗流程如圖 7 所示。

依圖 7 所示，本研究將比較主動專家組中不同人文領域專家在使用「基於主動學習的古漢語文本斷句系統」於古漢語文本斷句後建立的斷句模型之預測結果差異，並且探討不同人文領域專家的標註資料以及進行半結構訪談，以了解人文學者對於採用「基於主動學習的古漢語文本斷句系統」輔以古文斷句的看法與建議。本實驗依照資料集卷數中切分成 11 個文本區塊，並使用「主動學習斷句模式」，作為本節實驗的斷句模型建立依據。

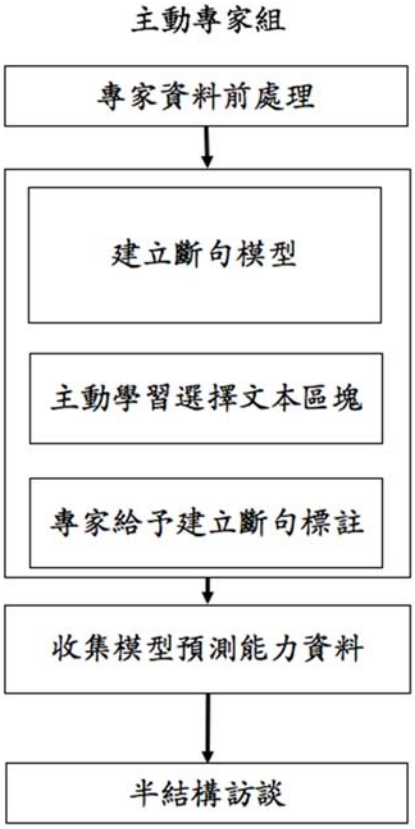

圖 7　主動式學習斷句系統實驗流程圖

本研究邀請六位具備古漢語知識的人文學者作為實驗對象，並使用「基於主動學習的古漢語文本斷句系統」對《峴泉集》進行古漢語文本的斷句判讀，主動專家組的實驗流程如圖 8 所示。如圖 8 所示，在實驗開始時會先請實驗對象閱讀資料集原文，並且說明「基於主動學習的古漢語文本斷句系統」的介面，再請專家以人工標註方式在第一回合建立第一個文本區塊的古漢語文本斷句標註資料，採用主動學習方法即時進行模型建立與選擇文本區塊，每次模型建立時間約數十秒，而後專家在系統選取的文本區塊中給予標註方法建立標註資料，直到使用完所有區塊的資料為止。每回合都會紀錄結果，並採用斷句結果之理想預測能力抵達回數，作為比較分析之依據。最後進行半結構訪談，以了解人文學者對於「基於主動學習的古漢語文本斷句系統」輔以古文斷句的看法與建議，以作為未來系統功能改進之參考。

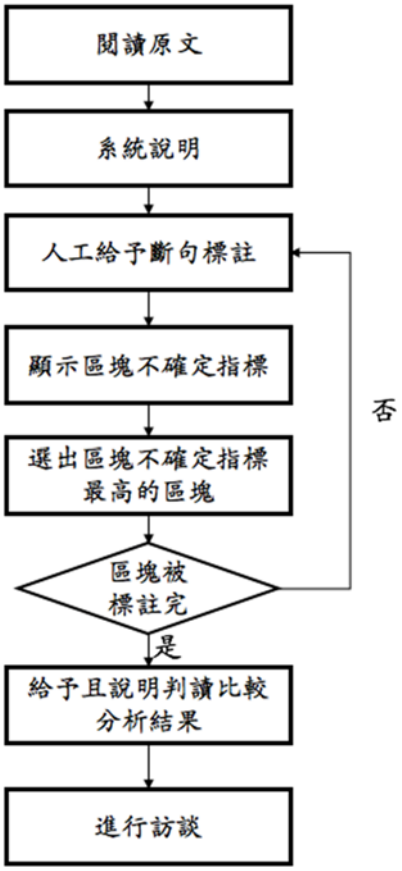

圖 8　主動專家組的實驗流程圖

伍、實驗結果分析

一、特徵模板與演算法之平均 F-measure 與變化斜率比較分析

（一）不同特徵模板與演算法之平均F-measure與斜率比較分析

表 3 為不同特徵模板與演算法之平均 F-measure 與斜率變化比較分析，結果顯示條件隨機場、單純貝氏、羅吉斯回歸、最大熵模型，不論在二字詞和三字詞特徵模板中，主動學習組的平均 F-measure 皆高於依序文本組，而主動學習組的斜率變化同樣都高於依序文本組；而雙向長短記憶詞模型的主動學習組平均 F-measure 低於依序文本組，主動學習組的斜率變化低於依序文本組。為了觀察主動學習組與依序文本組的平均 F-measure 差距，分別將不同演算法的主動學習組平均 F-measure 與依序文本組的平均 F-measure 相減，結

果顯示條件隨機場結合三字詞特徵模板的平均 F-measure 差距為最大，其值為 0.1200。

表 3

不同特徵模板與演算法之平均 F-measure 比較

演算法	特徵模板	文本選擇方式	F-measure 平均	斜率	F-measure 平均距差
條件隨幾場	二字詞	依序文本	0.1300	0.002	0.1107
		主動學習	0.2406	0.0264	
	三字詞	依序文本	0.1023	-0.0015	0.1200
		主動學習	0.2223	0.0325	
單純貝氏	二字詞	依序文本	0.3161	-0.0063	0.0539
		主動學習	0.3700	0.0046	
	三字詞	依序文本	0.2996	-0.0044	0.0394
		主動學習	0.3390	0.0065	
羅吉斯回歸	二字詞	依序文本	0.0850	-0.0022	0.0430
		主動學習	0.1279	-0.0048	
	三字詞	依序文本	0.0529	-0.0064	0.0871
		主動學習	0.1401	0.0216	
最大熵模型	二字詞	依序文本	0.1893	-0.0075	0.0028
		主動學習	0.1921	-0.0082	
	三字詞	依序文本	0.1447	-0.0081	0.0111
		主動學習	0.1558	-0.0086	
雙向長短記憶詞模型	單字詞	依序文本	0.313	0.017	-0.0580
		主動學習	0.2550	0.003	

5.1.2 用於「主動學習斷句模式」之特徵模板與演算法

根據不同特徵模板與演算法之平均 F-measure 和 F-measure 變化斜率比較分析，條件隨機場與三字詞特徵模板的組合在平均 F-measure 差距與斜率變化上提升程度均最高，並且主動學習組的模型比依序文本組的模型在古漢語斷句標註中能進行更有效的學習，符合本研究之考量，適合用於主動學習方法。因此，本研究採用條件隨機場與三字詞特徵模板於「主動學習斷句模式」中，並應用於發展「基於主動學習的古漢語文本斷句系統」。

二、主動專家組的古漢語斷句結果分析

（一）實驗對象基本資料

本研究以具備中文、哲學、歷史或宗教學科背景，或其他具有解讀古漢語之研究生為

實驗對象，邀請他們使用「基於主動學習的古漢語文本斷句系統」進行古漢語斷句標註，完成實驗活動之有效樣本共計 6 人，如表 4 所示，其中包括碩士生 3 人以及博士生 3 人。

表 4

實驗對象背景說明

實驗對象	學術程度	年齡	性別
S1	宗教碩士生	29	男
S2	中文博士生	27	女
S3	中文碩士生	28	男
S4	哲學博士生	33	男
S5	哲學碩士生	28	男
S6	中文博士生	27	男

（二）主動專家組的各回合古漢語斷句模型預測能力比較

表 5 為主動專家組的古漢語斷句模型平均 F-measure 比較結果，結果顯示實驗對象 S3 的模型平均最高 F-measure 為 0.256，在主動專家組中整體表現最好。實驗對象 S5 和 S6 的模型平均 F-measure 皆低於整體平均值 0.206，分別為 0.099 和 0.185，在主動專家組中整體效能最差。實驗對象 S4 的模型表現斜率最高為 0.026，在主動專家組中效能提升最多。實驗對象 S6 的模型表現斜率最低為 0.011，在主動專家組中提升效能最差。

表 5

主動專家組的古漢語斷句模型 F-measure 平均比較

實驗對象	F-measure 平均	斜率
S1	0.239	0.019
S2	0.234	0.021
S3	0.247	0.025
S4	0.231	0.026
S5	0.099	0.023
S6	0.185	0.011
平均	0.206	0.02

（三）主動專家組的古漢語斷句標註量與標註種類量分析

在觀察主動專家組的古漢語斷句標註資料後，本研究發現可能與其建立的斷句模型預測能力有某些關係，故分別以各專家的整體標註量、標註種類數，以及標註相鄰字種類數

進行分析。標註量為專家在標註過程中對古漢語進行標註的次數，並整理標註種類的數量作為標註種類數，再依照最後選擇的特徵模板方式取得標註相鄰字種類數。標註相鄰字種類能判斷出標註位置前後相關的標註字種類，依照「古漢語斷句模式」的特徵模板進行計算，選取具有斷句標註的字位置，找尋並計算前後三個位置的字，舉例來說「……而不息者，一陰一陽……」,「者」具備斷句標註，故計算前後三字分別「而」、「不」、「息」和「一」、「陰」、「一」，此時可以得到相鄰標註種類字為「而」、「不」、「息」、「陰」、「一」，有五種。

表 6 為主動專家組的整體標註次數、標註字種類數，以及標註相鄰字種類數比較結果，結果顯示平均 F-measure 表現最高的 S3 整體標註次數為 2455 次、標註字種類為 1033 種字，標註相鄰字種類數為 2248 種字，標註相鄰字種類比為 0.916。F-measure 斜率最高的 S4 整體標註次數為 2516 次、標註字種類有 1057 種字，標註相鄰字種類數為 2243 種字，標註相鄰字種類比為 0.891。平均 F-measure 表現最差的 S5 整體標註次數為 1714 次、標註字種類有 874 種字，標註相鄰字種類數為 2143 種字，標註相鄰字種類比為 1.25。平均 F-measure 斜率最差的 S6 整體標註次數為 2192 次、標註字種類有 998 種字，標註相鄰字種類數為 2215 種字，標註相鄰字種類比為 1.01。

表 6.

主動專家組的整體標次數、標註字種類數以及標註相鄰字種類數比較

實驗對象	F-measure 平均	斜率	標註字種類數	標註次數	標註種類總數比	標註相鄰字種類數	標註相鄰字種類比
S1	0.239	0.019	1058	2504	0.423	2246	0.897
S2	0.234	0.021	1034	2361	0.438	2238	0.948
S3	0.247	0.025	1033	2455	0.421	2248	0.916
S4	0.231	0.026	1057	2516	0.420	2243	0.891
S5	0.099	0.023	874	1714	0.510	2143	1.25
S6	0.185	0.011	998	2192	0.455	2215	1.01

圖 9 為平均 F-measure 與標註相鄰字種類比的散佈狀況，而表 7 中平均 F-measure 與標註相鄰字種類比的相關分析，結果顯示平均 F-measure 與標註相鄰字種類比的關聯為-0.98，顯著性為 0.01，兩者達到顯著負相關，表示 F-measure 平均越高的實驗對象，其標註相鄰字種類比越低；而平均 F-measure 越低的實驗對象，其標註相鄰字種類比越高。並且根據表 7 中平均 F-measure 與標注種類總數比的關聯為-983，顯著性 0.00，兩者呈現顯著負相關，表示實驗對象所標註的字種類越少，其建立的模型 F-measure 會越低。

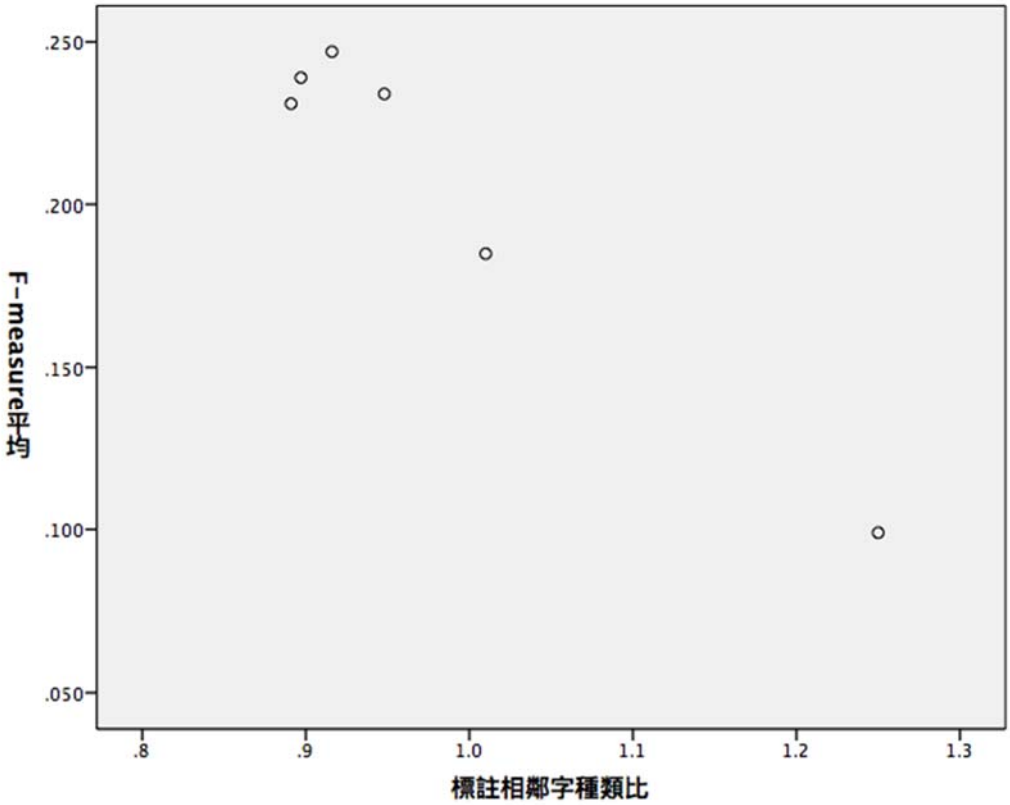

圖 9　主動專家組的 F-measure 平均與標註相鄰字種類比散佈圖

表 7

主動專家組的 F-measure 平均、標註種類種數比、標註相鄰字種類比相關分析

		標註種類總數比	標註相鄰字種類比	Fmeasure 平均
標註種類總數比	Pearson 相關	1	.995**	-.983**
	顯著性（雙尾）		.000	.000
	個數	6	6	6
標註相鄰字種類比	Pearson 相關	.995**	1	-.980**
	顯著性（雙尾）	.000		.001
	個數	6	6	6
Fmeasure 平均	Pearson 相關	-.983**	-.980**	1
	顯著性（雙尾）	.000	.001	
	個數	6	6	6

三、主動專家組的訪談分析

（一）實驗對象對於古漢語斷句規則的看法

1. 實驗對象會以古漢語的聲韻特性、詞性差異與句法規則作為古漢語斷句的判斷依據

一半的實驗對象在進行古漢語斷句時，會利用古漢語的字詞聲韻特性、字詞詞性差異和句法規則判斷文本內容，例如韻字、結尾字、句法結構。

2. 實驗對象進行古漢語斷句時有可能不按照原文編排順序

一半的實驗對象表示在進行古漢語斷句或解讀古漢語文本時，不一定會按照原文本的編排順序閱讀，而會帶著某些特定目的，例如找尋資料、解答問題，而且古漢語文本卷次多數也無前後脈絡關係。

（二）實驗對象對於在斷句解讀中顯示預測標註功能的看法

1. 實驗對象認為顯示預測標註的功能在古漢語斷句上有參考幫助

有些實驗對象認為「基於主動學習的古漢語文本斷句系統」所給予的預測標註，在進行古漢語斷句時具有參考上的幫助。

2. 實驗對象認為本工具顯示預測標註的功能在古漢語斷句上有幫助但取決於預測成效

有實驗對象認為「基於主動學習的古漢語文本斷句系統」的斷句預測標註功能，在進行古漢語斷句時具有參考上的幫助，但其成效會取決於模型的預測成效好壞。

3. 實驗對象認為顯示斷句預測標註的功能在古漢語斷句上會受到干擾

有實驗對象認為「基於主動學習的古漢語文本斷句系統」的斷句預測標註功能，在進行古漢語斷句時沒有參考上的幫助，而且會受到干擾。

（三）實驗對象對於本工具的改進意見

1. 實驗對象建議顯示文本區塊中文題名和題注能幫助提高理解速度

有實驗對象建議在「基於主動學習的古漢語文本斷句系統」閱讀文本的區塊中，增加顯示原文題名、詩名以及題注，如此可以提高閱讀理解的速度。

2. 實驗對象建議文本內容依照原典格式或直書顯示以確保文本完整性

有實驗對象建議閱讀介面中所採用的文本內容編排方式，應依照原文典籍的格式，例如以直書型式顯示或是遵照原文空行，盡可能使系統中顯示的文本與原文典籍相近。

3. 實驗對象建議在斷句預測標註後將內文分段有助於提高閱讀速度

有實驗對象建議使用「基於主動學習的古漢語文本斷句系統」輔以古漢語斷句時，每次在斷句預測標註後之文本內容應依據標註分段，以提高閱讀文本的速度。

4. 實驗對象建議加強顯示斷句標註的顏色、字距、字形有助於提高閱讀速度

有些實驗對象提到在系統中加強斷句標註的顏色、調整字距或是調整字型大小，例如斷句符號採用全形、調整單行長度，可提高閱讀文本的速度。

5. 實驗對象建議加強顯示介面的捲軸條能改進操作失誤

有實驗對象建議應加強顯示介面中捲動畫面的移動軸，以避免操作上的失誤。

6. 實驗對象認為可以透過顯示預測標註來呈現不同於自己的解讀內容

有實驗對象認為可以透過系統預測模型的預測標註結果，發掘不同於自己的解讀方式，或是自己不易理解到的其他見解。

7. 實驗對象認為顯示預測標註的字詞還需要改進

有實驗對象認為系統應增加不同的預測顯示字詞以及預測的效能還需要再提升，例如判斷古漢語特有的句尾詞。

8. 實驗對象認為顯示預測標註會因為使用者程度差異而有不同程度的輔助效果

有一半實驗對象認為系統顯示的斷句預測標註，會因為使用者對於古漢語文本熟悉程度而產生差異，而有不同程度輔助效果，使用者越熟悉文本越有幫助，而越不熟悉文本則可能會影響輔助效果。

陸、結論與未來研究方向

本研究透過評估五組演算法與兩組特徵模板組合的斷句模型平均 F-measure 與 F-measure 變化斜率差異，希望從中找出模型預測能力提升幅度最大的組合應用於「主動學

習斷句模式」，結果發現條件隨機場與三字詞特徵模板的組合，其 F-measure 與 F-measure 變化斜率皆最高，顯示條件隨機場與三字詞特徵模板最適合搭配主動學習框架。因此，本研究中採用條件隨機場與三字詞特徵模板作為「主動學習斷句模式」。此外，透過主動學習組與依序文本組的斷句模型平均 F-measure 與 F-measure 變化斜率進行之差異比較，發現具備「主動學習斷句模式」的主動學習組在斷句模型平均 F-measure 與 F-measure 變化斜率上優於依序文本組，表示「主動學習斷句模式」的斷句模型具備較好的學習能力，能以較少的訓練資料或是較快的速度達到與依序文本組同樣的效果。此外，本研究發現人文領域專家輔以「基於主動學習的古漢語文本斷句系統」進行古文標註過程中，每位專家的斷句模型 F-measure 變化斜率皆為正成長，表示在「主動學習斷句模式」下斷句模型會有效率的學習人文領域專家的古漢語斷句標註能力，達到主動學習框架可基於人機互動逐漸提升機器學習訓練資料效能的目的。再則，根據訪談得知多數人文領域專家認為「基於主動學習的古漢語文本斷句系統」中，顯示斷句預測標註的功能對於輔助古漢語斷句具有幫助，特別是在對於標註古文內容沒有信心的情況下可以提供參考，並且相較於閱讀註釋，斷句標註更能提高閱讀文本的速度。此外，專家也表示透過切換「基於主動學習的古漢語文本斷句系統」中顯示斷句預測標註的設計相當方便好用。由於人文領域專家在進行古漢語斷句時，時常會以古漢語的聲韻特性、詞性或是句法規則作為斷句與否的判斷依據，因此未來「基於主動學習的古漢語文本斷句系統」可考慮加入顯示基於不同考量古漢語特性規則的斷句標註，以更符合人文領域專家的需求。而在系統改善上可以採用依照原本典籍的格式進行排版，並且顯示文章題名與題注或是新增斷句標註分段的功能和調整介面的標註顏色、文字間距、字型形態與加強顯示介面元素以提高閱讀速度。並且能考慮新增不同版本的文本、圖片以供參考，而且研究對象有提出希望系統能提供諸如線上字典、線上資料庫等外部工具，幫助解讀不懂含義的詞彙來確認斷句標註，或者新增筆記工具紀錄解讀文本過程中產生的想法、解釋或是其他對於解讀文本有幫助的資料，以供後續參考。

而本研究實驗為了能使實驗對象有限的時間內完成斷句標註，將資料集的文字數量限縮在萬字左右，也因此導致實驗結果的斷句模型 F-measure 表現較低、預測斷句效果不佳，也可能使演算法未能表現出原本優勢，屬於本實驗的研究限制，希望未來有機會能進行更大規模的測試或是其他文本來驗證效果。

本研究雖已證明所發展之基於主動學習的古漢語文本斷句系統在輔助人文學者進行古文斷句上具有不錯的效能，仍有幾個未來研究方向值得繼續進行探究。首先，本研究發展的「古漢語斷句模式」僅能依照專家給予的斷句標註資料進行學習，未來可以嘗試加入諸如命名實體識別、古漢語斷詞等前處理，進一步提高模型的斷句預測能力。其次，由於「基於主動學習的古漢語文本斷句系統」的斷句模型預測能力，主要取決於人文領域專家

的斷句標註訓練資料正確與否，因此專家的行為模式值得探討，若能收集進行古漢語斷句時專家的系統操作行為資料，並進行更深一層的分析，有助於找出其他影響斷句模型預測能力的因素。最後，訪談中許多人文領域專家提到在求學過程中，古漢語的斷句訓練係依賴紙本方式進行，若能以「基於主動學習的古漢語文本斷句系統」為基礎，輔以進行古漢語斷句訓練，將有助於提高訓練效率。

誌謝

此論文撰寫台灣教育部高教深耕計畫特色領域中心之「華人文化主體性研究中心」經費補助（經費代碼 108H21)

（接受日期：2019 年 11 月 22 日）

參考文獻

Culotta, A., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. In *AAAI* (Vol. 5, pp. 746-751). Fort Belvoir, VA. https://doi.org/10.21236/ADA440382

Hu, Y. (2016). *Classical Chinese sentence segmentation as sequence labeling* (Doctoral dissertation, Texas Christian University Fort Worth, Texas). Retrieved from https://repository.tcu.edu/handle/116099117/10350

Huang, J., & Hou, H. (2008). On sentence segmentation and punctuation model for ancient books on agriculture. *Journal of Chinese Information Processing, 22*(4), 31-38.

Huang, H. H., Sun, C. T., & Chen, H. H. (2010). Classical Chinese sentence segmentation. *CIPS-SIGHAN Joint Conference on Chinese Language Processing.* Retrieved from https://www.aclweb.org/anthology/W10-4103

Huang, S., & Wang, D. (2017). Review and trend of researches on ancient Chinese character information processing. *Library and Information Service, 61*(12), 43-49.

Li, S., Zhou, G., & Huang, C. R. (2012). Active learning for Chinese word segmentation. In Proceedings of COLING 2012: Posters (pp. 683-692). Mumbai, India. Retrieved from http://www.aclweb.org/anthology/C12-2067

Liang, X. T. & Gu, L. (2015), Active learning in Chinese word segmentation based on stratified sampling strategy. *Application Research of Computers, 32*(5), 1353-1356.

Liu, K., Qian, X., & Wang, Z. (2012) Survey on active learning algorithms. Computer Engineering and Applications, 48(34), 1-4.

Liu, L., Wang, D. & Huang S. (2017) Review of the study about machine learning and its influence on library science. *Library & Information, 37*(6), 84-95.

Olsson, F. (2009). *A literature survey of active machine learning in the context of natural language processing.* (SICS Report 3600). Retrieved from http://www. ciera.org/library/reports/inquiry-3/3-025/3-025.pdf

Sassano, M. (2002, July). An empirical study of active learning with support vector machines for Japanese word segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 505-512). Association for Computational Linguistics.

Settles, B. (2009). *Active learning literature survey*. Retrieved from University of Wisconsin-Madison Department of Computer Sciences website: https://minds.wisconsin.edu/handle/1793/60660

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 6*(1), 1-114.

Wang, B., Shi, X., Tan, Z., Chen, Y., & Wang, W. (2016). A sentence segmentation method for ancient chinese texts based on NNLM. In M. Dong, J. Lin, & X. Tang (Eds.), *Chinese Lexical Semantics* (pp. 387-396). Cham: Springer International Publishing.

Wang, B., Shi, X., & Su, J. (2017). A sentence segmentation method for ancient Chinese texts based on recurrent neural network. *Beijing Daxue Xuebao (Ziran Kexue Ban)/Acta Scientiarum Naturalium Universitatis Pekinensis, 53*, 255-261.

Yeh, C. H., Wang, Y. C., & Tsai, T. H. (2011). Semi-supervised Chinese historical named entity extraction with active learning. In *From Preservation to Knowledge Creation: The Way to Digital Humanities (pp. 1-131)*. Taipei: National Taiwan University Press.

Zhang, K., Xia, Y., & Yu, H. (2009). CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *Journal of Tsinghua University (Science and Technology), 10*. 1733-1736.

Journal of InfoLib and Archives
11:2=95 (Dec ’19) 117-145
DOI: 10.6575/JILA.201912_(95).0004

Development and Application of an Ancient Chinese Sentence Segmentation System Based on Active Learning

Chih-Fan Hsu* Chung Chang**

【 Abstract 】

This study aims to develop a sentence segmentation system of ancient Chinese texts based on active learning. It is expected that through the human-machine cooperation mode, the training corpus needed to establish a model for automated ancient Chinese sentence segmentation could be reduced and humanities researchers may work more efficiently on sentence identification of uninterpreted text. Two experiments were conducted in this study for the system development and evaluation. In the first experiment, the automatic sentence segmentation models established by applying different algorithms and feature templates to sequential text selection and active learning text selection were compared to select the most suitable algorithm and feature template to employ in establishing this system. The results show that conditional random fields combined with three-word feature template adopted in active learning could perform effective learning outcomes that would be appropriate to apply to build the active learning sentence segmentation model for ancient Chinese texts. In the second experiment, six humanities researchers were invited to use the system to conduct sentence segmentation tasks of the assigned ancient Chinese texts to evaluate the performance of the system. Sentence segmentation results produced by individual humanistic researchers using the system were compared and analyzed. Semi-structured interviews were also conducted to gather an in-depth understanding of their experience and suggestions of using the system The experimental results show that the developed ancient Chinese sentence segmentation system based on active learning could effectively learn

* MA Student, Graduate Institute of Library, Information and Archival Studies,National Chengchi University
E-mail: billxu0521@gmail.com
** IT engineer, Research Center for Chinese Cultural Subjectivity, National Chengchi University
Principal author for all correspondence E-mail: foxx1216@gmail.com

humanities researchers sentence segmentation data and constantly improve the model prediction through human-machine cooperation. Moreover, according to the interviews, most of the humanities researchers participated in this study reported a positive experience of using the system and indicated that the sentence segmentation prediction function provided in the system could effectively assist their sentence segmentation work. The prediction of the active learning sentence segmentation model could be further improved by embedding the name entity model or applying other phonological features or POS tagging of ancient Chinese in the future study. It is also expected to develop this system into a digital humanities learning platform for ancient Chinese sentence segmentation training in the future.

Keywords

Digital humanities, Active learning, Machine learning, Automatic ancient Chinese sentence segmentation, Human-computer interaction

【 Summary 】

This study aims to develop an “active learning based ancient Chinese text segmentation system” for supporting research on digital humanities. With the combination of active learning and machine learning algorithms, the man-machine cooperation model reduces training corpuses required for establishing the automatic ancient Chinese segmentation model, assists humanists in the efficient segmentation interpretation of literatures which have not been interpreted, and is used for the continuous development of real-time segmentation annotation for supporting research on digital humanities. The establishment of an “active learning based ancient Chinese text segmentation system” contains three stages of data processing, segmentation model establishment, and active learning practice. The data processing stage aims to organize and process ancient Chinese texts into the format for establishing the segmentation model. It is further divided into data segmentation and feature template establishment. The “active learning segmentation model” proposed in this study is practiced in the segmentation model establishment stage. Using algorithms for establishing the segmentation model, the system would precede several runs of segmentation model establishment stage with iteration, read the output segmentation model with a segmentation model interpreter, and calculate Chinese word segmentation uncertain indicators and block uncertain indicators in the text. The active learning practice stage selects text blocks with active learning selection, and the data which is not annotated in the block is preceded segmentation annotation by experts. The completed block is added to the next training data till all data are annotated. In order to find out the most suitable algorithms

and feature templates for the establishment of the “active learning based ancient Chinese text segmentation system”, various algorithms and feature templates are applied to match the segmentation models established with sequential texts and active learning. F-measure and F-measure variable slope based on the segmentation prediction result are compared. It is discovered in the experimental results that conditional random fields and three-word feature templates in active learning could effectively precede learning and are suitable for developing the “active learning segmentation model”. Scholars in humanities are invited to use the “active learning based ancient Chinese text segmentation system” for the segmentation interpretation of ancient Chinese texts in the second experiment. The segmentation models established by individual humanists’ annotation data are preceded comparative analyses, and semi-structured interview is applied to deeply understand the humanists’ perception and suggestions about the segmentation assisted system developed in this study. The experimental results show that the “active learning based ancient Chinese text segmentation system” could effectively learn humanists’ segmentation annotation data and the model predictive power, based on man-machine cooperation, is constantly promoted. Finally, the interview results reveal humanists’ positive evaluation of the system operation process and interface; most interviewees consider that the segmentation prediction of the system could effectively provide assistance for ancient Chinese segmentation. Adding segmentation annotation, considering different ancient Chinese characteristics and rules, would better conform to the requirement of experts in humanities. In order to have the experiment objects complete the segmentation annotation within limited time, the Chinese words in the dataset are restricted in 10 thousand words. In this case, the F-measure performance of the segmentation model is low and the prediction of segmentation is not favorable. It might have the algorithms not present the original advantages. These are the research restrictions in this experiment. A larger test or other texts are expected to verify the effect. A naming model or feature templates with other ancient Chinese rules could be designed in the future to further enhance the segmentation predictive power, collect experts’ system operation behaviors during the ancient Chinese segmentation for deeper analyses, and apply the developed system to education on humanities to develop the digital humanities education platform for training ancient Chinese segmentation.

Romanized & Translated Reference for Original Text

Culotta, A., & McCallum, A. (2005). Reducing labeling effort for structured prediction tasks. In *AAAI* (Vol. 5, pp. 746-751). Fort Belvoir, VA. https://doi.org/10.21236/ADA440382

Hu, Y. (2016). *Classical Chinese sentence segmentation as sequence labeling* (Doctoral dissertation, Texas Christian University Fort Worth, Texas). Retrieved from https://repository.tcu.edu/handle/116099117/10350

Huang, J., & Hou, H. (2008). On sentence segmentation and punctuation model for ancient books on agriculture. *Journal of Chinese Information Processing, 22*(4), 31-38.

Huang, H. H., Sun, C. T., & Chen, H. H. (2010). Classical Chinese sentence segmentation. *CIPS-SIGHAN Joint Conference on Chinese Language Processing.* Retrieved from https://www.aclweb.org/anthology/W10-4103

Huang, S., & Wang, D. (2017). Review and trend of researches on ancient Chinese character information processing. *Library and Information Service, 61*(12), 43-49.

Li, S., Zhou, G., & Huang, C. R. (2012). Active learning for Chinese word segmentation. In Proceedings of COLING 2012: Posters (pp. 683-692). Mumbai, India. Retrieved from http://www.aclweb.org/anthology/C12-2067

Liang, X. T. & Gu, L. (2015), Active learning in Chinese word segmentation based on stratified sampling strategy. *Application Research of Computers, 32*(5), 1353-1356.

Liu, K., Qian, X., & Wang, Z. (2012) Survey on active learning algorithms. *Computer Engineering and Applications, 48*(34), 1-4.

Liu, L., Wang, D. & Huang S. (2017) Review of the study about machine learning and its influence on library science. *Library & Information, 37*(6), 84-95.

Olsson, F. (2009). *A literature survey of active machine learning in the context of natural language processing.* (SICS Report 3600). Retrieved from http://www. ciera.org/library/reports/inquiry-3/3-025/3-025.pdf

Sassano, M. (2002, July). An empirical study of active learning with support vector machines for Japanese word segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 505-512). Association for Computational Linguistics.

Settles, B. (2009). *Active learning literature survey*. Retrieved from University of Wisconsin-Madison Department of Computer Sciences website: https://minds.wisconsin.edu/handle/1793/60660

Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 6*(1), 1-114.

Wang, B., Shi, X., Tan, Z., Chen, Y., & Wang, W. (2016). A sentence segmentation method for ancient Chinese texts based on NNLM. In M. Dong, J. Lin, & X. Tang (Eds.), *Chinese Lexical Semantics* (pp. 387-396). Cham: Springer International Publishing.

Wang, B., Shi, X., & Su, J. (2017). A sentence segmentation method for ancient Chinese texts based on recurrent neural network. *Beijing Daxue Xuebao (Ziran Kexue Ban)/Acta Scientiarum Naturalium Universitatis Pekinensis, 53*, 255-261.

Yeh, C. H., Wang, Y. C., & Tsai, T. H. (2011). Semi-supervised Chinese historical named entity extraction with active learning. In *From Preservation to Knowledge Creation: The Way to Digital Humanities (pp. 1-131)*. Taipei: National Taiwan University Press.

Zhang, K., Xia, Y., & Yu, H. (2009). CRF-based approach to sentence segmentation and punctuation for ancient Chinese prose. *Journal of Tsinghua University (Science and Technology), 10*. 1733-1736.