112 年青少年科學人才培育計畫

研究計畫封面

科　　別：系統軟體科

計畫名稱：以自然語言處理擷取西域各國間之空間資料

關鍵詞 3 個：西域傳、自然語言處理、地理資訊系統

編　　號：(由本館填寫)

112 年青少年科學人才培育計畫

研究計畫內容

計畫名稱：以自然語言處理擷取西域各國間之空間資料

摘要

本研究旨在運用自然語言處理的方法，擷取史書文本中關於各國相對於某地點的距離與方位資料，並整理文本中各國資料，加以統整成表格，以便後續讀取與應用。本研究挑選了四種能處理文言文的模型進行比較，包括 Jiayan, Jieba, Stanza, HanLP。我們將結果以表格的方式呈現，設定欄位來呈現內容，便於資料的讀取與後續應用。本研究目前的成果得出 HanLP 在分詞與詞性標注上的表現明顯最佳。未來本研究預計不但擷取基本的空間資料，更會納入如戶數、人口、軍隊等資訊。本研究也將利用 GIS 套疊圖層進行疊圖分析，並對照現代之地圖。未來更能加以延伸至世界各地的歷史文獻，使用地理資訊系統分析，使本研究的成果更具有實用性。

壹、 研究動機

自然語言處理提供了從文本中擷取資訊的方法，將其應用於擷取西域各國間的空間資料，可以更有效地從古代的史料中得到距離或方位等資訊，並且可以使我們更容易理解西域之間的地理關係。

透過自然語言處理技術，可以自動擷取這些文本中的地理資訊，並將其轉換為結構化的資料，使得這些資訊更加易於理解和分析。因此，我們希望開發出一個能夠描述過往西域各國的文本中自動擷取空間資料的系統。這個系統可以從正史的文本中得到有用的地理資訊，並將其整合成資料庫，供研究人員進行進一步的分析和研究。透過這樣的研究，我們可以更深入地了解漢代西域各地間的地理關係。

貳、 研究目的及研究問題

一、運用自然語言處理的方法擷取史書文本中關於各國相對於某地點的距離與方位資料。

二、整理文本中各國資料，並擷取該國戶、口、兵、風俗、河川等資料，加以統整成表格。

參、 研究設備及器材

一、 電腦資訊

(一) Lenovo ThinkPad E15 Gen 2

1. 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz
2. RAM 16GB; GPU: NVIDIA GeForce MX450

二、 本機軟體與使用語言及模組

(一) Anaconda Navigator 2.0.3: Python 3.8.8; Anaconda Spyder 5.1.5

(二) HanLP 2.1.0b45

肆、 研究過程或方法及進行步驟

一、 文本

本研究所使用的文本包括史記、漢書、後漢書提及西域的篇章，這些文本主要涵蓋了漢代的史料，並描述著漢朝時西域地方間的關係。

目前，我們已經分析了三個文本，包括史記卷一百二十三、漢書卷九十六以及後漢書卷八十八，目前僅是初步的階段，後續將會納入分析更多文獻資料。這些文本中包含了大量的描述西域的資訊，我們將對其進行詳細的分析和整理。

二、 選擇的模型

本研究在經過文獻的查詢過後，挑選了四種能處理文言文的模型進行比較，四種的分詞效果略有差距。四種模型如下：

(一) Jiayan 模型

(二) Jieba 模型

(三) Stanza 模型

(四) HanLP 模型

本研究將使用此四模型來進行分詞與詞性標注，測試何種模型表現最佳。於此階段，本研究將透過一些指標來進行客觀的結果評估。由於分詞結果中標準答案和分詞結果詞語數不一定相等，因此採用了 Precision（精確率）與 Recall（召回率）來評估各模型分詞與詞性標註的結果.

在 NLP 中，Precision 表示「分詞結果與標準答案重合部分的集合」在「分詞結果所有單字構成之區間的集合」中的比例，結合了真陽性（TP）和偽陽性（FP）的結果。

Precision 公式如下所示：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall 表示「分詞結果與標準答案重合部分的集合」在「標準答案所有單字構成之區間的集合」中的比例，結合了真陽性（TP）和偽陰性（FN）結果。

Recall 公式如下所示：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

而 F1-score 廣泛用於評估模型的性能。F1-score 將 Precision 和 Recall 結合成為一個指標，指標越高表示表現越佳。公式如下所示：

$$\text{F1} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

在經過全文的分詞過後，在四者之中 HanLP 的表現明顯最好，因此往後的試驗及研究方法說明皆會使用 HanLP 為主，並將於後續詳細說明各模型分類的結果。

三、 研究流程

在進行分析前，我們需要先進行文本預處理的工作。原始的文本為 XML 的網頁檔案格式，首先需將其轉換成字串形式，以方便後續的處理。接著，我們使用前述的各模型進行分詞的處理，將長篇文本切割成一個個有意義的單字，並透過詞性標註對各個單字進行標記，以利後續的分析。在這個過程中，有時會因一個詞在不同處有不同意思而錯誤標註。最後，我們進行了語法分析，對句構進行分析，並且在對文本的語法結構有基本的了解過後，建立以詞性為依據的公式來篩選文本，根據標註的詞性建立自動提取的流程。這樣的流程可以有效地幫助我們從文本資料中提取出有用的資訊，從而整理並輸出方便理解的樣式。

四、 HanLP 模型的使用

在選用 HanLP 作為本研究的工具後，找出該工具中對本研究有幫助的四個功能。如下所列，以下每項皆描述其縮寫與中英文全名。

(一) **pos**: part-of-speech tagging，詞性標註

標註分詞過後，每一個詞語的詞性。

(二) **ner**: Named Entity Recognition，命名實體識別

對有特殊意義的名詞（即實體）標註類別，如地名、人名、時間、數詞等等。

(三) **sdp**: Semantic Dependency Parsing，語義依存分析

分析一個句子中單字與單字之間的語義關係，標註語義上的角色或結構上的這些關係，如施事者與受事者等。

(四) **cor**: Coreference Resolution，指代消解

將指代同一事物的提及或其代詞聚集到同一處，以方便理解各個詞指代的角色。

五、 擷取內容

本研究擷取的內容便是於文本當中本研究感興趣的內容，包括各國國名與治所、一地之方位與里程。其中描述位置的方法可能是距離京城（長安）有多少里，抑或者敘述與其他地標性的地點的距離（如陽關、玉門、洛陽等）。文本描述的方式亦有可能是與其他鄰國、河川之方位與里程，另外，如戶數、人口戶、口、兵等資訊也可順帶擷取。最後，便將結果以表格的方式，並設定欄位（如：國名、治所、相對地點、方位、里程……等）來呈現內容，便於資料的讀取與後續應用。

伍、 預期結果、已有初步之結果

一、 選擇的模型討論

本研究比較了四種能處理文言文的模型,包括 Jiayan, Jieba, Stanza, HanLP。四種模型在分詞效果上略有差距。以下以一範例文本來展示四個模型的分詞結果。

範例文本如下：「**大宛國去長安萬二千五百五十里，東至都護治，西南至大月氏，南亦至大月氏，北至康居。**」

(一) Jiayan 模型

表一、Jiayan 模型分詞與詞性標註結果

大宛	國	去	長	安	萬	二千	五百	五十	里
副詞	動詞	動詞	動詞	動詞	動詞	數詞	數詞	數詞	量詞
，	東	至	都	護	治	，	西南	至	大月氏
標點符號	動詞	介詞	副詞	動詞	動詞	標點符號	地名	介詞	人名
，	南	亦	至	大月氏	，	北	至	康居	。
標點符號	方位詞	副詞	動詞	一般名詞	標點符號	方位詞	介詞	地名	標點符號

本模型將許多不應拆分的單位再度拆分為兩個詞，也不具辨識地名的能力。

(二) Jieba 模型

表二、Jieba 模型分詞與詞性標註結果

大宛	國去	長安	萬	二千五百	五十里	，	東至	都	護治
專有名詞	時間	地名	數詞	數詞	數詞	標點符號	時間	副詞	動詞
，	西南	至	大月氏	，	南亦	至	大月氏	，	北至
標點符號	地名	介詞	人名	標點符號	人名	介詞	人名	標點符號	地名
康居	。								
人名	標點符號								

本模型會於不應拆分的位置將單字拆分，亦有些辨識不正確的部分。

(三) Stanza 模型

表三、Stanza 模型分詞與詞性標註結果

大	宛	國	去	長	安萬二…	里	，	東	至
動詞	地名	一般名詞	動詞	動詞	數詞	量詞	一般名詞	一般名詞	動詞
都護	治	，	西南	至	大	月	氏	，南	亦
一般名詞	動詞	動詞	一般名詞	動詞	動詞	時間	一般名詞	一般名詞	副詞
至	大	月	氏	，	北	至	康居	。	
動詞	動詞	時間	一般名詞	一般名詞	一般名詞	動詞	地名	動詞	

很顯然地本模型不具辨識標點符號的能力，分詞與詞性標註的結果不佳。

(四) HanLP 模型

表四、HanLP 模型分詞與詞性標註結果

大宛	國	去	長安	萬二千…	里	，	東	至	都護治
地名	一般名詞	動詞	地名	數詞	量詞	標點符號	一般名詞	動詞	地名
，	西南	至	大月氏	，	南	亦	至	大月氏	，
標點符號	一般名詞	動詞	地名	標點符號	一般名詞	副詞	動詞	地名	標點符號
北	至	康居	。						
一般名詞	動詞	地名	標點符號						

可以發現本模型的分詞與詞性標註的結果皆大致正確。

以範例文本來使用前述指標來評估分詞的結果，如下表所示，其中 HanLP 的結果最佳，F1-Score 近乎高達 0.94。

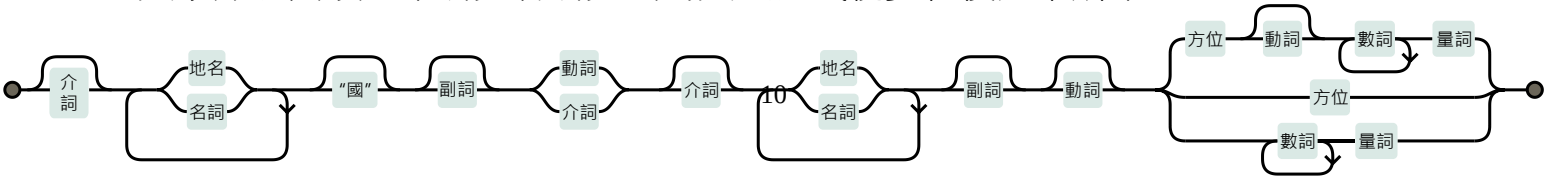
表五、各模型分詞的結果之評估指標表

	Jiayan	Jieba	Stanza	HanLP
TP	22	15	19	23
TP+FP	30	22	29	24
TP+FN	25	25	25	25
Preci-sion	0.733333	0.681818	0.655172	0.958333
Recall	0.88	0.6	0.76	0.92
f1-score	0.8	0.638298	0.703704	0.938776

二、解析出的文法

本研究使用 nltk 的 RegexpParser 來篩選句子，此方法可以以正規表示式 (Regexp, Regular Expression) 的原理來篩選，正規表示式使得匹配一組字串變得容易許多，因此非常適合用於此研究找到符合「描述地點的文法」的句子。

目前研究出主要的文法規則為一串類似正規表示式的公式，若直接顯示純文字難以閱讀，本研究將研究出的文法公式視覺化後如下所示：



圖一、描述地點的文法公式之視覺化圖片

三、詳細的已知語法

上面所述為能套用整體文本之公式，本小節將說明該公式可實際擷取出之示例。以下列出在文本中已發現的數個表示空間關係的語法規則，額外的動詞或副詞等裝飾字在此不於規則中顯示，在舉例中以淺灰色表示可忽略。

(一) 專有名詞 + 介詞/動詞 + 專有名詞 + 數詞 + 量詞

舉例一：拘彌國 去 于窴 三百 里

舉例二：大宛國 去 長安 萬二千五百五十 里

(二) 專有名詞 + 介詞/動詞 + 專有名詞 + 方位詞

舉例一：祁連山 在 甘州 西南

舉例二：天竺 在 崑崙山 南

(三) 專有名詞 + 介詞/動詞 + 專有名詞 + 方位詞 + 數詞 + 量詞

舉例一：康居國 在 京 西 一萬六百 里

舉例二：身毒國 又 居 大夏 東南 數千 里

(四) 代名詞 + 方位詞 + 數詞 + 量詞 + 動詞 + 專有名詞

舉例一：其 西北 可 二千 里 有 奄蔡

四、 文本擷取後的輸出結果

表六、以史記卷一百二十三為例之文本擷取輸出表格

國名	相對地點	方位	里程
大宛國	長安		12550 里
康居國	京	西	10600 里
奄蔡	康居國	西北	2000 里
大月氏	大宛	西南	
大夏國	嬀水	南	
大宛	匈奴	西南	
拘彌國	于窴		300 里
玉門關	沙州壽昌縣	西	6 里
蒲昌海	蒲類海	東	
鹽澤	長安		5000 里
烏孫	大宛	東北	2000 里
康居	大宛	西北	2000 里
月氏	敦煌	以東	
祁連山	甘州	西南	
火山國	扶風	南東	
條枝	安息	西	數千里
女國	于窴國	南	2700 里
于窴	京		9670 里
大夏	大宛	西南	2000 餘里
天竺	崑崙山	南	
大夏	漢		12000 里
身毒國	大夏	東南	數千里

五、 預期結果

本研究目前還位於初步的階段，未來預計不但擷取包括各國國名與治所、一地之方位與里程，更會納入如戶數、人口、軍隊等資訊，將文本轉為可讀取的資料。最後，便將結果以表格的方式輸出（如 CSV 逗號分隔檔案），關於一國的資訊便能快速瞭若指掌，各種空間的資訊亦一目了然，後續更可以加以延伸應用。

本研究未來預計也會利用地理資訊系統（GIS）套疊圖層進行疊圖分析，並對照現代之地圖做對應。而因各個朝代所使用的單位不一致，甚至可以根據現代的地圖，評估當代單位的實際長度，對於研究以往朝代也是一大幫助。此外，透過本研究，更能利用電腦及近代人工智慧的幫助推敲出張騫出使西域或西遊記中唐三藏等人前往西天取經的實際路線，並評估路徑的正確性與合理性。另外經由不同史料的對照，若因不同的時代的路線走法不同，亦可由本研究得知，更能規劃一套二十一世紀能使用的現代路徑。

若以此研究為基礎，未來可延伸至其他歷史的文本，或將時間拉長至唐代或以後，並觀察政權更迭是否影響到地理資訊的描寫。亦可加以應用至其他類型的文獻，如唐代的詩文集，觀察是否出現描寫地理資訊的文本。

此外，本研究到了後期更不限於西域等地，只要有文獻，便能加以分析，快速並大量的使電腦閱讀文本得到各國相關的地理資訊。亦可將模型擴充至印歐語系國家的文獻乃至於世界各地，並使用地理資訊系統強大的功能進行分析，關於一國的資訊便一目了然，電腦也容易讀取並處理。

陸、 本計畫之創見性及其未來應用

本研究主要透過電腦及人工智慧的輔助，得到西域各國的相對位置，更能探究出張騫出使西域或《西遊記》中唐三藏等人前往西天取經的實際路線。此外，研究也在後期擴展至不限於西域等地，只要有文獻可供分析，即可快速且大量地藉由電腦閱讀文本，獲取各國相關的地理資訊。未來便能利用閱讀文本所取得的資訊來進行分析。

經由本研究，可得出古代單位的實際長度，是研究史料的極佳工具。另外，也可推敲史料的合理性與正確性。並對不同史料加以對照。

未來可利用電子地圖（如 Google 地圖）等地理資訊系統的工具，對比現代的地理資訊與不同史料的描述，並能使用路徑規劃等功能。

柒、 結論

本研究比較了四種能處理文言文的模型，並以範例文本展示四個模型的分詞結果，而得知模型 HanLP 表現最佳。此外，研究還使用 nltk 的 RegexpParser 來篩選句子，找到符合「描述地點的文法」的句子。目前，本研究已可擷取出國家名、相對位置與距離等資料，未來將會從文本中擷取更多資訊。最終，研究目標是將文本轉為電腦可讀取的資料，並使用地理資訊系統進行分析，以瞭解西域各國的相對位置，並探究出張騫出使西域或《西遊記》中唐三藏等人前往西天取經的實際路線。

未來，此研究可延伸至更多歷史文本，並將模型擴充至世界各地的文獻，使用地理資訊系統分析，從文獻中擷取地理的資料便能方便許多。

捌、 參考資料 (文獻) 及其他

[1] He, H., & Choi, J. D. (2021). The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 5555–5577. https://aclanthology.org/2021.

[2] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. [pdf][bib]

[3] 楊尊一（譯）（民 108）。**PyTorch 自然語言處理：以深度學習建立語言應用程式**（原作者：Delip Rao & Brian McMahan）。臺北市：歐萊禮。

[4] 何晗 (2019)。**自然语言处理入门**。人民邮电出版社。

[5] [自然語言處理基礎] 語法分析與資訊檢索 (II)（2021 年 9 月）。檢自 https://ithelp.ithome.com.tw/articles/10263629 (Feb. 2023)

玖、研究計畫執行進度表

期間 / 工作項目	民國 112 年										民國113年	
	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月
研讀參考資料	■	■	■									
分析更多語法規則	■	■										
加入 sdp, cor 等功能	■	■										
加入治所、戶、口、河川等	■	■										
加入更多文本來分析			■	■								
報告之繕打	■	■	■	■	■	■	■	■				
利用電子地圖 (GIS) 對比					■	■						
擴充至世界各地的文獻							■	■				

註 1：請填寫至 112 年 10 月 14 日止，應完成工作項目，並報名臺灣國際科展。

註 2：請以粗黑筆劃出每一工作項目之起迄時間。

壹拾、研究經費申請明細表 (耗材、物品及雜項費用)

（一）依本計畫玖、經費執行研究經費審核，最高上限一萬元。

（二）凡執行研究計畫所需之耗材、物品 (非屬研究設備)及雜項費用，均可填入本表內，不符合核銷研究經費不另行通知。

（三）說明欄請就該項目之規格、用途等相關資料詳細填寫，以利審查。

金額單位：新臺幣/元

項 目 名 稱	說明	單位	數量	單價	金額	備註
Transcend 可攜式外接硬碟	2TB StoreJet 25M3C	個	1	2 900	2 900	
隨身碟		個	1	1 000	1 000	
印刷費	含彩色大張海報印刷				1 500	
文具費	含立可帶、各式筆類以及其他文具				1 500	
書籍	PyTorch 自然語言處理：以深度學習建立語言應用程式	本	1	580	580	
書籍	Language Files: Materials for an Introduction to Language and Linguistics (13 Ed.)	本	1	680	680	
其他書籍費	其他未來將會購置但未於此詳列之書籍	本			1 800	
合計					9 960	