112年青少年科學人才培育計畫

期末研習營研究報告書封面

科　　別：系統軟體科

作品名稱：以自然語言處理擷取西域各國間之空間資料

關 鍵 詞：西域傳、自然語言處理、地理資訊系統

編 號：（編號由國立臺灣科學教育館統一填列）

112年青少年科學人才培育計畫

研究報告內容

**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，並加以統整成表格。再透過表格內容繪製成圖片，以便後續應用及分析。本研究首先選擇了史記、漢書、後漢書提及西域的篇章作為文本，另挑選了四種能處理文言文的模型加上GPT-4模型進行比較，其中四種模型包括Jiayan, Jieba, Stanza, HanLP。我們將結果以表格的方式呈現，再利用多維尺度分析 (MDS) 等演算法以得出國家間之相對位置圖。接著，將相對位置圖繪製至地理資訊系統 (GIS) 上，與實際現況之地圖進行疊圖分析作為現代地圖之對照。本研究目前的成果得出HanLP在分詞與詞性標註上的表現為四種傳統模型中明顯較佳者，而GPT-4整理表現又更勝於前者。未來本研究預計更會納入如戶數、人口、軍隊等資訊。未來更能加以延伸至各種歷史文獻，使用地理資訊系統分析，使本研究的成果更具有實用性。

1. **前言**
   1. 研究動機

從閱讀文字中空間描述建構空間概念的認知能力是很重要的，其中又包括地理空間及社會空間，而空間概念又是眾多高階概念之共同基礎。歷史文本因使用古代語言表達並依托古代知識系統，因此造成一般人在閱讀及理解障礙，更遑論古代地理資訊的不足。

然而，自然語言處理提供了人們從文本中擷取資訊的方法。因此便可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，本研究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，便可以更深入地了解漢代西域各地間的地理關係。若能進一步與既有的歷史地理知識比對或校對，則可以重現更精確的歷史場景與空間關連，透過地圖媒介輔助使用者更完整理解歷史文獻的內容。

* 1. 研究目的及研究問題
  2. 運用自然語言處理的方法擷取西域傳等史書文本中關於各國相對於某地點的距離與方位資料。
  3. 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。
  4. 利用多維尺度分析 (MDS) 等演算法，根據已知資訊以得出國家間之相對位置圖。
  5. 將相對位置圖移至地理資訊系統 (GIS) 上，與實際現況之地圖進行疊圖分析。
  6. 文獻回顧
  7. 分析之文本

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國家間的關係。

目前，本研究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，後續還會納入分析更多的文本資料。這些文本中包含了大量的描述西域的資訊，本研究將對其進行詳細的分析和整理。

* 1. 相關文獻及書籍探討

書籍《大寫西域》（高洪雷，民107）當中提供了許多豐富的地圖可供參考。其中更說明了某些由關口到西域國家的可能路徑，包含對絲路南道11國、絲路北道11國、蔥嶺10國的詳細路徑或情形等解說，是對於本研究而言非常有價值的參考資料。

另外經由其他文獻的探討，研究者發現Görz等人 [5] 的研究主要關注了歷史地理知識和空間的交互作用，並提供了一種不同文化比較地圖的方法。他們強調了文本和地圖之間的關聯，以及如何從文本和地圖中提取出特定文化的空間知識。

Santos等人 [8] 的結論指出了在文學作品當中擷取地理資訊中需要考慮的許多困難問題。他們的結果說明了自然語言處理不能取代人工的仔細閱讀。同時，地名的使用方式不僅僅是參考一個地方，地名也有可能會隨著時間變動，也有著相同的名稱可以指示不同的地方的特性或一個名稱可以用於多種大小尺度的地方的特性。如果想要理解文學中地點的位置，就必須考慮許多複雜的因素。

1. **研究設備及器材**
2. 電腦資訊
   1. Lenovo ThinkPad E15 Gen 2
      1. 11th Gen Intel(R) Core (TM) i7-1165G7 @ 2.80GHz
      2. RAM 16GB; GPU: NVIDIA GeForce MX450
3. 本機軟體與使用語言及模組
   1. Anaconda Navigator 2.0.3; Python 3.8.8; Anaconda Spyder 5.1.5
   2. HanLP 2.1.0b45; scikit-learn 1.2.1; networkx 3.1
   3. QGIS 3.16 Hannover
4. **研究方法或過程**
5. 選擇的模型

本研究在經過文獻的查詢過後，首先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：

* 1. Jiayan模型
  2. Jieba模型
  3. Stanza模型
  4. HanLP模型

另外，由於近期生成式人工智慧的蓬勃發展，除了傳統的模型以外，本研究利用了近期非常熱門、由OpenAI公司所開發的GPT-4模型來解析史書中的內容並納入比較。透過串接API的方式並不斷修改prompt，使其產生表格以利後續分析。

本研究於此階段將使用此四模型及GPT-4來進行分詞與詞性標註，以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：

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：

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：

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。

1. 研究流程

目前，本研究共分析了三份文本，分別為：史記卷一百二十三、漢書卷九十六以及後漢書卷八十八，這些皆為有提及西域的篇章。在進行分析前，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，首先需將其轉換成純字串的形式，以方便後續的處理。接著，研究者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，研究者進行了語法分析，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幫助實用資訊的提取，從而整理並輸出方便理解的樣式。研究流程圖？

1. HanLP 模型的使用

在本研究所使用的四個模型中，以HanLP表現最佳。因此便選擇此模型來深入研究。在選用 HanLP 作為本研究的工具時，找出了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。

* 1. **pos**: part-of-speech tagging，詞性標註

標註分詞過後，每一個詞語的詞性。

* 1. **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。

* 1. **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。

* 1. **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。

1. 擷取內容

本研究擷取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補述與其他鄰國、河川之方位與里程，另外，如戶數、人口、兵力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、里程、書籍出處、原文文句……等）來呈現內容，便於資料的讀取與後續應用。

具體而言，以史記集解原文文句「烏孫在大宛東北可二千里」為例，其中地點1為烏孫，地點2為大宛，方位為東北，距離為二千里（即2000里），而書籍出處則為該段註解文字的出處⸺集解，最後並以CSV （逗號分隔值）檔匯出。

1. 相對位置圖之繪製

本階段中，研究者打算用前面所生成的表格製成國家間的相對位置圖。操作上的部分預計會利用MDS（多維尺度分析）與DFS（深度優先搜尋）演算法繪製出國家相對位置關係圖。

* 1. MDS演算法

首先，資料當中的點與點兩兩之間必定為「已知距離」或「未知距離」者。透過 DFS 演算法逐步推算原資訊中兩點間尚未得到的距離，藉此可計算出兩兩之間的距離，而後再利用所有兩兩點對之間的距離並套入 MDS 演算法，得出每個點在二維平面上的相對位置，最終以回歸出方位誤差的最小值來旋轉或鏡射原地圖，以期所生成之地圖能與實際情形最為接近。另外，除 MDS 和 DFS 演算法外，亦可嘗試自製或改良現有的演算法或使用機器學習等相關技術來優化本研究建圖的算法。

* 1. 未來可嘗試之演算法
     1. Force-Directed Layout Algorithms

本方法將圖視為一個彈簧和粒子系統，其中節點之間的距離由吸引和排斥力量來調節，透過模擬物理的現象，使得節點間達到一個平衡狀態。

然而，此方法可能存在局部最優解，而會受初始狀態影響，且對於較大型的圖可能需要長時間的收斂。

* + 1. 蒙地卡羅方法

蒙地卡羅方法是一種考慮不確定性的降維方法，使用機率分佈來表示距離和方向的不確定性，然後使用蒙地卡羅方法採樣多個可能的場景，以計算最可能的點位置。

* 1. 繪圖方法

本研究使用Python模組NetworkX 來繪圖。在以DFS產生所有點對的距離的矩陣後，以MDS方法產生所有點的座標，便可透過NetworkX來繪圖。其中，如同一般的圖，邊的長度表示地點間的距離，角度表示方向。而各條邊的顏色則表示書籍出處，同一出處則為同一顏色，不同出處則否。

1. 應用地理資訊系統

地理資訊系統 (GIS) 為利用電腦來分析地理現象的工具。本研究主要利用QGIS軟體將國家相對位置關係圖疊至實際地圖上以進行疊圖分析，但目前仍需尋找方法以評估正確率。

1. **期末之結果**
2. 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。**」

* 1. Jiayan 模型

表1、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。

* 1. Jieba 模型

表2、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。

* 1. Stanza 模型

表3、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。

* 1. HanLP 模型

表4、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。

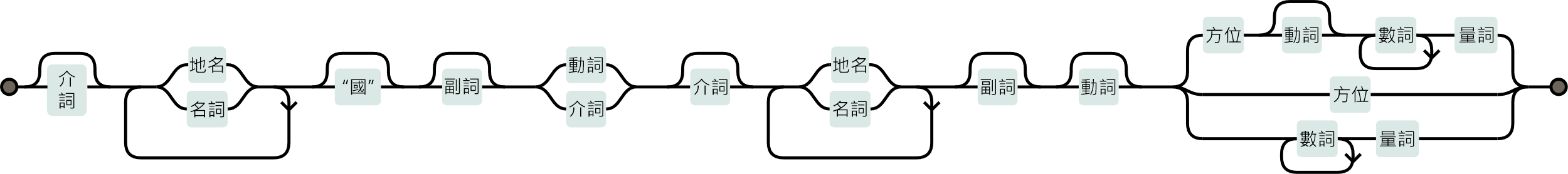
以範例文本來使用前述指標來評估分詞的結果，如下表所示，其中HanLP的結果最佳，F1-Score高達近乎0.94。

表5、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 |
| Precision | 0.733 | 0.682 | 0.655 | 0.958 |
| Recall | 0.880 | 0.600 | 0.760 | 0.920 |
| f1-score | 0.800 | 0.638 | 0.704 | 0.939 |

1. 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n、描述地點的文法公式之視覺化圖片

1. 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。

* 1. 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

* 1. 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

* 1. 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

* 1. 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

1. 文本擷取後的輸出結果（傳統模型）

表6、以史記卷一百二十三為例之傳統方法文本擷取輸出表

|  |  |  |  |
| --- | --- | --- | --- |
| **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 | 媯水 | 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 |

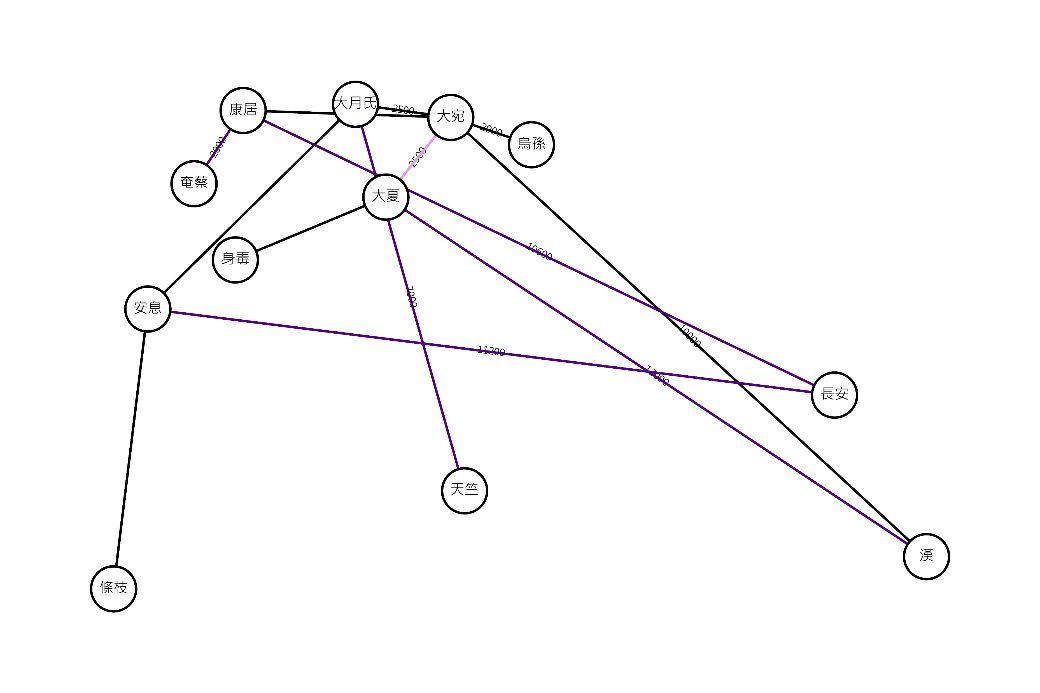
1. 文本擷取後的輸出結果 (GPT-4)

表7、以史記卷一百二十三為例之GPT-4文本擷取輸出表（節錄）

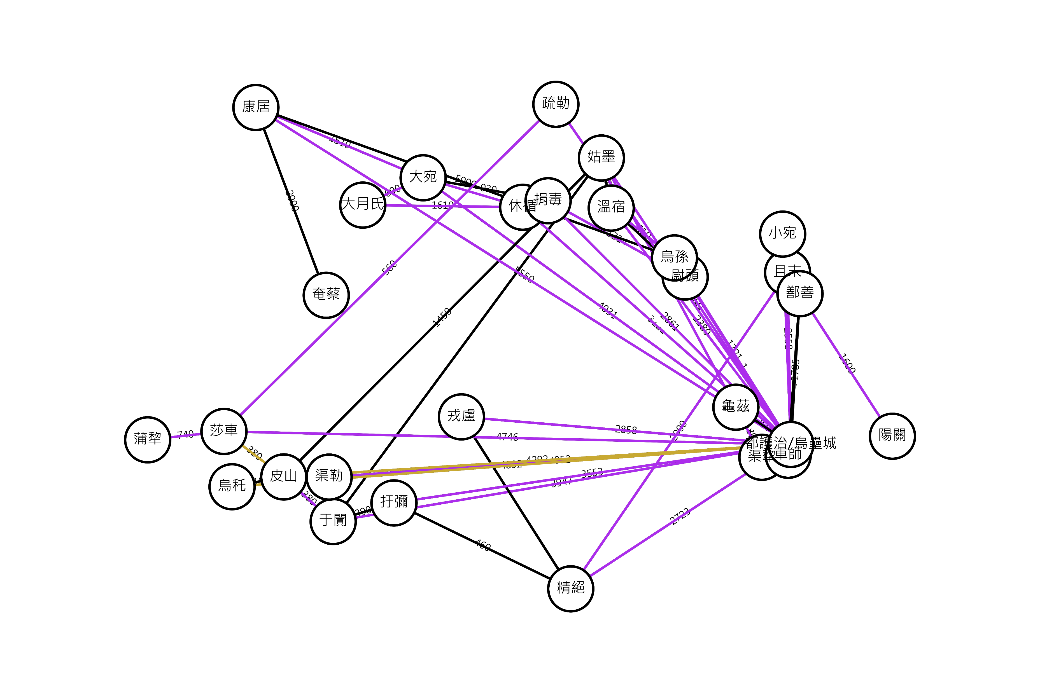
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **國名** | **相對地點** | **方位** | **里程** | **書籍出處** | **原文文句** |
| 大宛 | 長安 | -- | 12550里 | 正義 | 大宛國去長安萬二千五百五十里 |
| 大宛 | 都護治 | 西 | -- | 正義 | 漢書云：「大宛國...東至都護治」 |
| 大宛 | 大月氏 | 東北 | -- | 正義 | 漢書云：「大宛國...西南至大月氏」 |
| 大宛 | 大月氏 | 北 | -- | 正義 | 漢書云：「大宛國...南亦至大月氏」 |
| 大宛 | 康居 | 南 | -- | 正義 | 漢書云：「大宛國...北至康居」 |
| 康居 | 長安 | 西 | 10600里 | 正義 | 「康居國在京西一萬六百里」 |
| 奄蔡 | 康居 | 西北 | 2000里 | 正義 | 「其西北可二千里有奄蔡」 |
| 大月氏 | 大宛 | 西南 | -- | 正義 | 「此大月氏在大宛西南」 |
| 大月氏 | 長安 | -- | 11600里 | 正義 | 「漢書云去長安萬一千六百里」 |
| 大夏 | 媯水 | 南 | -- | 正義 | 「大夏國在媯水南」 |
| 京南 | 蔥嶺 | -- | 15000里 | 正義 | 從京南連接至蔥嶺萬餘里 |
| 大宛 | 匈奴 | 西南 | -- | -- | 大宛在匈奴西南 |
| 大宛 | 漢 | 西 | 10000里 | -- | 在漢正西，去漢可萬里 |
| 康居 | 大宛 | 北 | -- | -- | 其北則康居 |
| 大月氏 | 大宛 | 西 | -- | -- | 西則大月氏 |
| 大夏 | 大宛 | 西南 | -- | -- | 西南則大夏 |
| 烏孫 | 大宛 | 東北 | -- | -- | 東北則烏孫 |
| 扜罙 | 大宛 | 東 | -- | 集解 | 東則扜罙 |
| 大宛 | 于窴 | 東 | -- | 索隱 | 于窴 |
| 鹽澤 | 玉門、陽關 | -- | 350里 | 正義 | 鹽澤去玉門、陽關三百餘里 |
| 鹽澤 | 長安 | -- | 5000里 | -- | 鹽澤去長安可五千里 |
| 烏孫 | 大宛 | 東北 | 2000里 | 集解 | 烏孫在大宛東北可二千里 |

1. 相對位置圖的建立

經不斷修正過後，由史記文本所建立之國家間相對位置圖結果如圖n所示，由漢書文本所建立之相對位置圖結果如圖n所示。由圖中可見，其實仍有一部分與實際相對的方位仍有些許差異，但大致的方位是正確的，而這些將於後續討論說明。



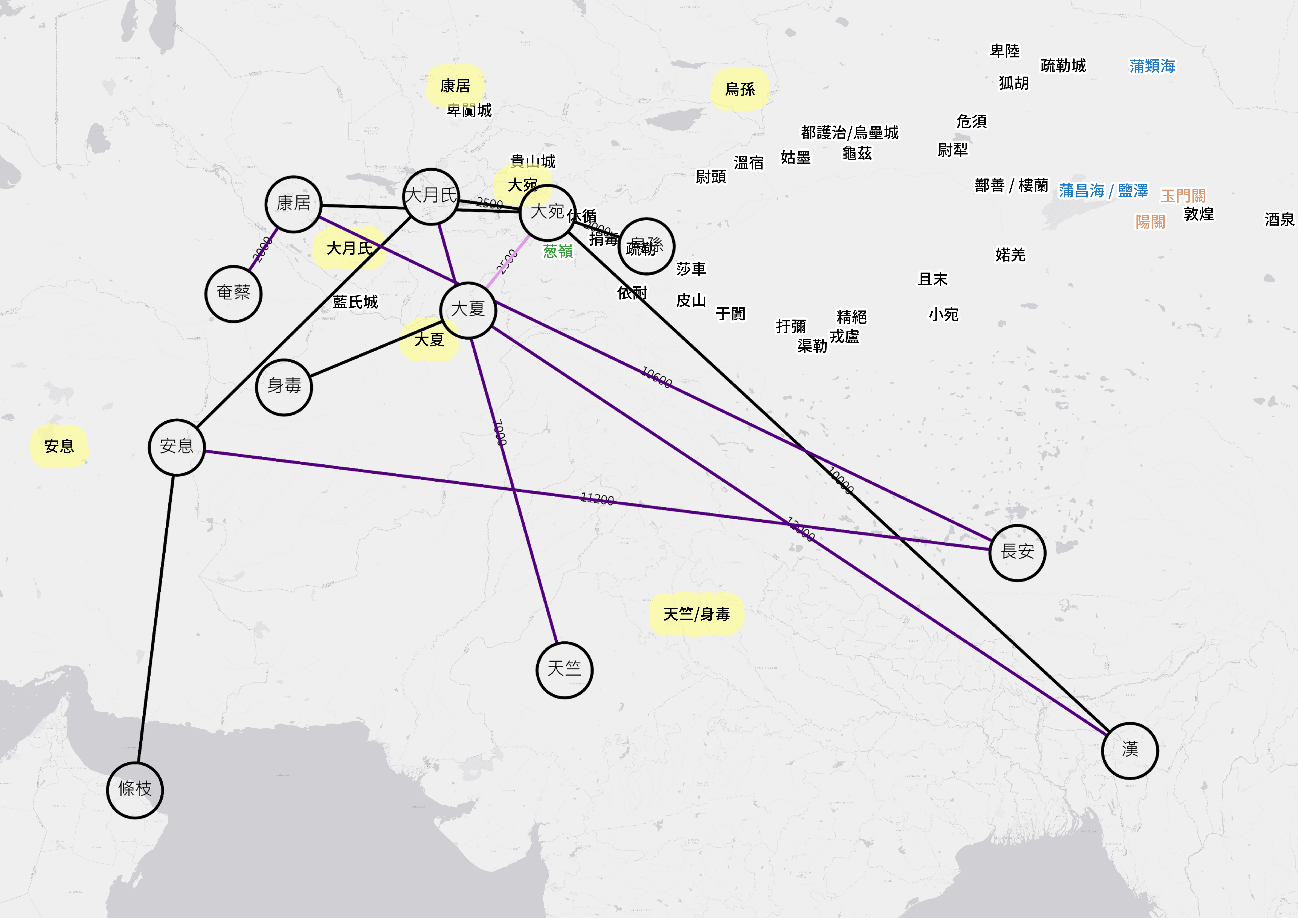
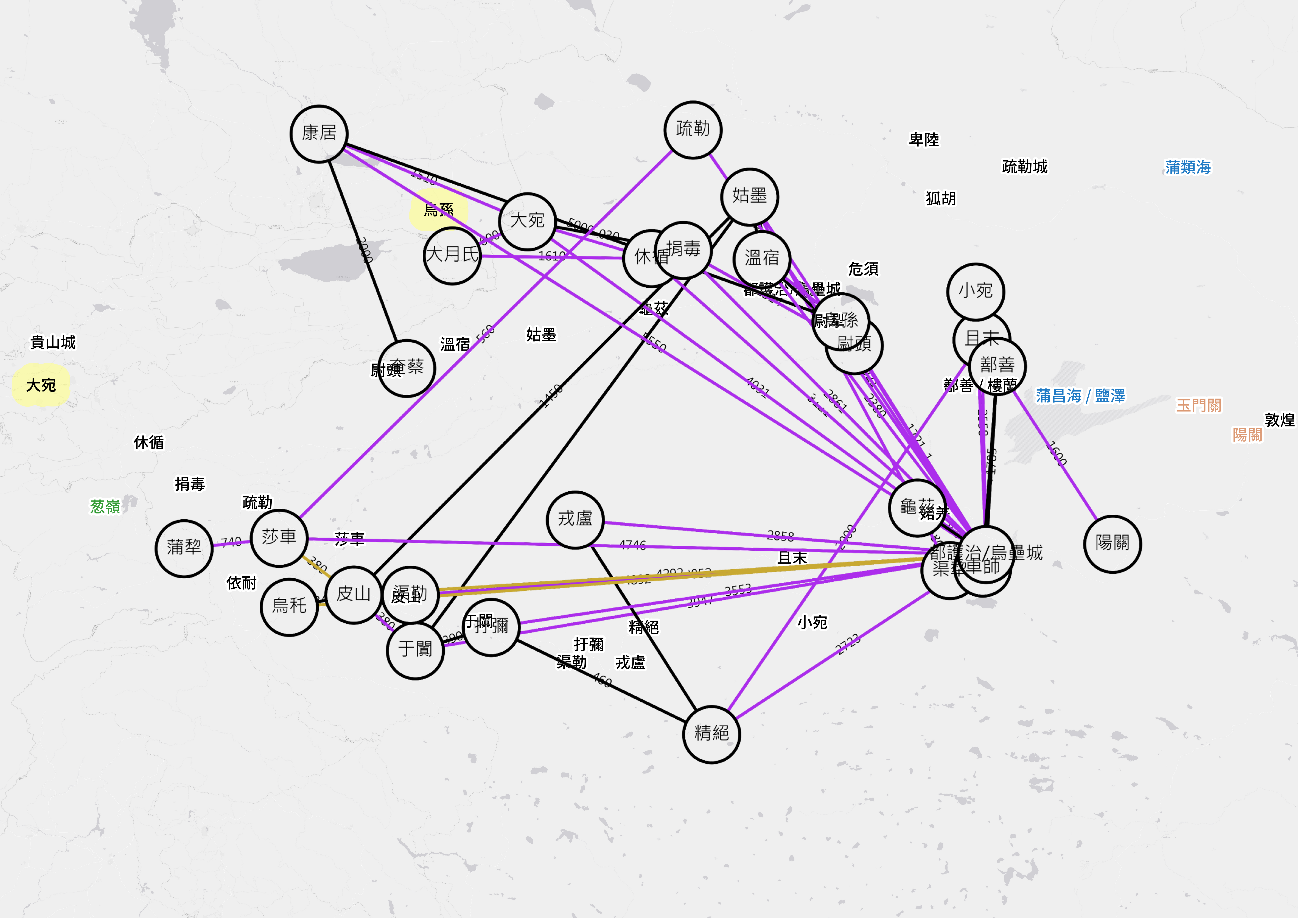
圖n、史記文本之國家間相對位置圖



圖n、漢書文本之國家間相對位置圖

1. 套用至地理資訊系統 (GIS)

以相對位置圖疊至相對位置圖的結果如圖n及圖n所示。由圖中可見，其實仍有不少部分與實際位置仍有些許差異。這部分則可能是因史書描述時僅以八方位敘述，我們在假設距離時也因而只能以45°為最小單位，造成不小的誤差。此外，兩國間的里程實際上為路徑長度，也有可能有繞遠路的情形，使實際情形更加難以估計。



圖n、漢書文本之GIS疊圖對照圖

圖n、史記文本之GIS疊圖對照圖

1. **研究結論與應用**

本研究利用電腦及人工智慧的輔助，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。研究也納入了近期熱門的GPT模型進行比較，而發現其結果優於先前四個模型。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。接著，本研究透過DFS與MDS演算法繪製國家的相對位置圖，也將找尋更佳的演算法來繪製圖形。本研究也初步的使用了地理資訊系統來疊圖，最終也將使用地理資訊系統進行分析並找出衡量正確率的方法，以瞭解西域各國的相對位置，並探究出漢代出使西域等實際路線。

經由本研究，未來可加以應用並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，此研究可延伸至更多歷史文本，不限於西域等地，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。更能夠加以利用電子地圖（如Google地圖）等地理資訊系統的工具，對比現代的地理資訊不同史料的描述，並能使用路徑規劃等功能，使本研究能提供的成果更加豐碩。

1. **本計畫遇到的問題及解決方式**

首先研究者發現在進行文法解析時總有漏網之魚，描述地點或距離的資訊有太多種，且每一作者皆有自己不同的敘述方式。使得在以傳統模型進行分析之時，文法規則總是新增不完。

另外，一地可能有不只一個地名，目前僅是以手動的方式，以人工閱讀文獻及參考資料後將兩名稱合併為同一個，要如何自動合併某些節點（地名）成為同一處指代的地方也是一大問題。目前只能根據已知的文獻納入資料，使程式預先知道可能有哪些同地不同名的案例。

一國的國名、勢力範圍也有可能會變遷。不但名稱可能會不固定，某些遊牧國家也可能會遷徙至其他不同的地方。國家間也會有吞併及滅亡的情形發生。

GPT模型雖然表現不錯，不過在許多條結果都會有相反的情形發生。即原文句表達的意思是「地點1在地點2的某方向」，然而GPT卻理解成「地點2在地點1的某方向」，如此反過來的情形變成需要不斷手動檢查。

此外，本研究原先是使用GPT-3.5。然而，文句的複雜程度已不敷使用，因而需要更強大的 GPT-4，但仍需排隊等待加入使用 GPT-4 API 的 waitlist。

如何評估並量化所生成的結果的正確度亦是本研究遇到的一大難題。在此，可以使用相對於實際位置的均方差的概念進行評估。然而，古國在現今的實際位置本身有些也具有爭議或仍然未知。國家的概念也是一個有面積的範圍而非單一點。

另根據交通路線的不同即會造成測量上明顯的不同。用車輪轉動的次數測量距離，會受到交通路線的影響很大。更甚者，或有對於不同國家皆有不同走法，或有傳抄訛誤者，都大幅增大了估計時的誤差。

編寫演算法也是本研究遇到的困難，要建立一個複雜且符合本研究需求的演算法絕非易事。目前僅以找尋現成的演算法並微調套用來使用。國家都具有一定的面積，目前許多演算法都對於這部份無法處理。

另外還有地圖投影的部分，可否直接投影在經緯度座標上是本研究想要詢問的問題。中亞、新疆地區屬中緯度地區，若是不考慮地圖投影則很有可能會產生誤差。目前仍未有專門投影中亞地區地圖的規範，因此需要在地理資訊系統中自訂投影法，而這部分仍有待研究者學習。

最後，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。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。若以此研究為基礎，未來可延伸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。

1. **參考文獻**

[1] [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)

[2] 何晗 (2019)。**自然语言处理入门**。人民邮电出版社。

[3] 高洪雷（民 107）。**大寫西域**（上、中、下）。新北市：野人文化。

[4] 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。

[5] Görz, G., Geus, K., Michalsky, T., & Thiering, M. (2019). *Spatial Cognition in Historical Geographical Texts and Maps: Towards a Cognitive-Semantic Analysis of Flavio Biondo’s “Italia Illustrata.”*

[6] 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.

[7] 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.

[8] Santos, D., & Alvs, D. (2023). *Placing GIS and NLP in Literary Geography: Experiments with Literature in Portuguese. International Journal of Humanities and Arts Computing, 17*(1).