利用人工智能技術進行竹簡編聯的初步測試

——以《清華簡》爲例

莫伯峰（首都師範大學、“古文字與中華文明傳承發展工程”協同攻關創新平臺）

胡勒奮（北京師範大學、“古文字與中華文明傳承發展工程”協同攻關創新平臺）

對出土竹簡進行重新編聯，恢復原來簡序，是簡牘研究中極爲重要的基礎性工作。竹簡編聯的依據，“大致可以從外在因素和内在因素兩個方面來把握。所謂外在因素，大致包括竹簡形態如長度、寬度、編綫的道數與間隔，書寫風格如字體、密度、某些特殊用字以及標識符號的采用等等。所謂内在因素，則是指詞匯、句式、體裁和内容。只有將這些因素綜合考慮，不斷推敲，反復改進，才能獲得比較合乎原初面貌的編聯結果”。[①] 其中，“外在因素是竹簡編聯的重要依據，使形制相同的竹簡可以編排在一起；内在因素則是檢驗編聯正確與否的最終條件，不僅同篇的竹簡要有共同的主題，而且簡序的排列也要使内容上有連貫性”。[②]

本文所論人工智能運用於竹簡編聯，是從内在因素方面進行，也就是基於語言和内容的連貫性實現編聯。近年來人工智能在自然語言處理（NLP）領域取得了長足進步，這是展開這項研究的重要基礎。一些新的語言模型已經具有非常强大的語言處理能

---

① 陳偉：《郭店竹書别釋》，湖北教育出版社，2003 年，第 83—84 頁。近年來發現的“簡背劃痕”，也是編聯的一項重要外在因素。

② 劉傳賓：《出土簡牘編聯與拼綴方法綜論》，《天津師範大學學報（社會科學版）》2018 年第 4 期。

力，比如 2018 年谷歌發布的 BERT 模型就取得了非常不凡的成績，該模型包含有語序、語法、語義、語境等多層次的語言信息，在多項指標上都超越了人類的表現。[①] 胡韌奮等曾構建了古漢語 BERT 模型，在詩、詞、古文三種文體的斷句任務中，F1 值分別達到 99%、95%、92%以上，已經具有了很强的實用價值。[②] 而之後 OpenAI 發表的 GPT-2 模型又在多項指標上超越了 BERT 模型，[③]説明這方面的研究正是方興未艾。

在 BERT 模型所設立的兩項初始訓練目標中，句子預測便是其中之一。在這項任務中，模型的目標是判斷兩個句子是否相鄰，這與竹簡編聯的任務性質極爲近似。因此，通過構建適合竹簡材料的語言模型，來完成竹簡編聯任務具有理論上的可行性。當然，竹簡編聯需要考慮的因素非常多，竹簡語言與後世語言也有較大的差異，僅憑語言模型能够發揮多大的效果還有待驗證。所以，我們設計了本次測試，一方面希望通過實例檢測，驗證研究思路是否可行；另一方面也希望通過測試發現問題，爲今後探索指明方向。而更爲重要的是，希望能從另一個角度來觀察和思考專家編聯工作的語言本質，深入探討編聯工作的理論問題。以下爲此次初步測試的一些基本情況。

一、人工智能編聯竹簡的基本原理簡析

我們曾在《人工智能模擬辭例歸納的初步測試》一文中討論過，人工智能技術如何模擬古文字考釋中的“辭例歸納法”。[④] 對於人工智能而言，無論是進行“辭例歸納”還是“竹簡編聯”，所使用的語言模型本身並没有本質性差異，只是任務發生了改變。而在我們看來，“辭例歸納”和“竹簡編聯”兩項任務還具有很大的一致性——“辭例歸納”是根據語境，進行**字詞單位**的推導；“竹簡編聯”則是根據語境，進行**句子或更大語言單位**的推導。二者只是在推導單位上存在差異，所憑藉的依據還是“語感”和“語義”這兩個方面。因此，這裏依舊沿用《人工智能模擬辭例歸納的初步測試》文中用到的《三字經》例子，來舉例説明人工智能如何實現竹簡編聯。

首先看利用語感進行編聯的情況。假設我們手裏有一枚殘簡，上面寫有“人之初”，相信絶大多數人都能猜測出，下面可以編聯的是“性本善”。即使年幼的孩童，不能理解

① Devlin, Jacob, et al. “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv: 1810.04805 (2018).

② 胡韌奮、李紳、諸雨辰：《基於深層語言模型的古漢語知識表示及自動斷句研究》，《中文信息學報》2021 年第 4 期。

③ OpenAI: “GPT-2: 6-Month Follow-Up.” https://openai.com/blog/gpt-2-6-month-follow-up/.

④ 莫伯峰、丘煒琦、謝澤澄：《人工智能模擬辭例歸納的初步測試》，《漢語言文學研究》2021 年第 3 期。

這些句子的意義,但因爲經常聽過或讀過,也能讓他們完成這種編聯。這就是利用“語感”實現的編聯,憑藉大量語言用例的積累,使人感覺語言自然流暢的一種能力,“語感”是人類進行編聯時首先憑藉的方法。而人工智能在“培養語感”方面已經達到了較高的水平,通過海量語料的訓練,人工智能在某些領域的語感已經超過了人類。從具體實現方式上來看,人類通常從正面來評估語感,認爲編聯後語感最好的語句就是正確的答案。人工智能則從反面來評估語感,以困惑度(Perplexity)來反映語感好壞,認爲編聯後困惑度越低的語句就是正確的答案。

再來看利用語義進行編聯的情況。假設殘簡上寫的是“性相近”,而我們又從來没有讀過《三字經》,那我們又如何能够編聯出“習相遠”這個正確答案呢?這時可以從語義的角度來進行編聯。語言表達時,相鄰句子通常在語義上是銜接連貫的,句子中詞語的意義比較接近。因此,通過計算句子的語義相似度,也可以爲編聯提供幫助。“性”與“習”的詞義接近,“遠”與“近”的詞義相對,所以便能够根據語義推斷出“性相近”與“習相遠”關係較緊密。人工智能在“判斷詞義關係遠近”方面,也已經發展出了非常成熟的技術,基於四庫全書等古代漢語語料,可以將古代漢語中的詞語表示爲高維稠密向量,通過計算這些向量的距離便能够衡量語義的相似度。

“語感”和“語義”能解決竹簡編聯的很大一部分問題,但不是所有問題。這與人工智能模擬“辭例歸納”時所遇到的情況一樣,除了語感和語義之外,各種知識也是解決語言問題的重要因素。比如“昔孟母,擇鄰處”中的“昔孟母”,從語感和語義上來説,它當然也可以是“昔張母”“昔王母”,甚至可以是“昔張父”“昔王父”。只有根據歷史知識“孟母三遷”,並進行邏輯推導,我們才能確定這裏必須是“昔孟母”。因此,只有建立相關的古代文化知識庫才能完備地解決這一問題。

以上,是人工智能利用“語感”和“語義”進行竹簡編聯工作的舉例簡述。更加詳細的原理分析,還可參閱《人工智能模擬辭例歸納的初步測試》一文。

二、任務分析與評測指標

本次竹簡編連的任務目標爲:給定一批打亂順序的竹簡文本,將其按照文意進行排序。爲了驗證方法的可行性以及評估的便利性,先將任務設計爲:給定一支位置確定的簡(第一支簡),在篇章範圍内預測它的下一支簡,目標是預測的準確率越高越好。即,假定該篇文章包括 N 支簡,對於任意一支簡,希望模型能够對餘下 N-1 支簡與之匹配程度進行打分,並將上下文匹配度從高到低排序。正確答案的排名越靠前,模型效果越好,否則越差。

爲全面評測模型的效果,編聯設有三項評測指標:

Top1-accuracy：正確答案簡排在第一位的比例

Top3-accuracy：正確答案簡排在前三位的比例

Mean Reciprocal Rank(MRR)：搜索引擎或推薦系統中衡量排序效果的指標

這三項指標中，Top1-accuracy 能够展現模型的精准程度，Top3-accuracy 能够爲實際編聯工作縮小範圍，Mean Reciprocal Rank(MRR)則綜合評估了模型的總體效果。

三、數據分析與預處理

本次編聯測試的數據爲《清華大學藏戰國竹簡》(壹—玖)[①]釋文共計 45 篇。數據首先經初步整理(包括：對存疑簡文進行修訂，對缺失簡文進行標記，並按章拆分《系年》篇)，共得到測試文獻 67 篇，共計 824 支簡。作爲參與預測的有效竹簡，平均每篇文獻包含約 15 支簡。在這 824 個編聯單位中，《越公其事》包含了 76 支簡，簡數最多(《系年》包含 138 支簡，但被拆分爲 23 章，所以被計爲 23 篇文獻)，單從數量上來看，是編聯難度最大的一篇簡文。《尹至》和《祝辭》包含了 5 支簡，簡數最少，單從數量上來看，是編聯難度最小的兩篇簡文。

選擇《清華簡》(壹—玖)作爲編聯測試的對象，基於多方面的原因。首先，《清華簡》(壹—玖)的簡序基本已經確定，因此測試有標準答案可供對照。其次，《清華簡》(壹—玖)的釋文整理水平非常高，更容易獲取一個較好的文本材料，以避免因文獻自身的問題影響到編聯工作的結果。最後，《清華簡》(壹—玖)的文本數量也較爲可觀，充足的數據有利於獲得更客觀的測評結果。基於上述這些原因，並充分考慮了技術實現的可行性，我們嘗試整理了《清華簡》(壹—玖)的釋文，作爲本次編聯的測試數據。

爲滿足測試的要求，釋文被進行了多項預處理。本次測試的對象是距今已逾兩千多年的戰國竹簡文字，在戰國文字轉化爲現代楷書漢字這一過程中，必然會涉及一些語言文字方面的學術性問題。因爲這些問題並非本次測試的重點，所以在測試流程的設計上，將先屏蔽各種干擾性語言文字因素，希望後續逐步地接近真實的編聯任務，遵循由易到難循序漸進的方式來推動測試的進步。我們對《清華簡》(壹—玖)文本做了以下預處理：

(1) 剔除字形層面的影響。《清華簡》(壹—玖)爲戰國時期文獻，其上爲戰國時期的文字。由於這次測試的不是圖像識别模型，而是語言模型，因此這些戰國文字都被轉寫爲現代楷書漢字。也就是説測試基於現代楷書漢字進行，而非基於戰國文字或其隸定

① 李學勤主編：《清華大學藏戰國竹簡》(壹—玖)，中西書局，2010—2019 年。以下簡稱《清華簡》(壹—玖)。

字形進行。

(2) 剔除各種字詞關係的影響。釋文中的通假字、異體字、古今字、繁簡字等情況都將會對編聯工作産生影響,我們根據現時學界較爲通行的意見,確定一種意見來處理釋文中的這些用字等情況,以一種最簡文本作爲本次編聯的基礎。例如竹簡中的“隹”常通爲“唯”,“古”常通爲“故”,在釋文中不再括注這種字詞關係,徑釋爲“唯”和“故”。

(3) 降低疑難文字的影響。當釋文中存在未識字(無編碼的圖片文字)或考釋意見分歧較大的文字時,無疑將對編聯工作産生較大影響。編聯模型應對這種不完整信息的推斷也存在較大困難。因此,我們以能够貫通上下文意的詞匯填補這些疑難文字的位置,以此來暫時減少對編聯結果産生的影響。

(4) 降低殘斷竹簡的影響。一些竹簡存在殘斷的情況,也會對編聯工作産生較大影響。現在雖有一些方法(如基於語義相似度的方法,詳下文)來應對這一問題,但與其他竹簡的編聯邏輯有較大差異,也會對最終結果的性質産生較大影響。因此暫時將這部分排除在編聯任務之外,在編聯結果中以“NA”標記該竹簡的下支簡爲殘簡。

四、測試過程及測試結果

測試分爲兩個階段進行:第一個階段的模型,主要是從語感角度對竹簡進行編聯。第二階段的模型,增加了語義信息方面的評估,並引入了懲罰因子來處理編聯中的“萬能竹簡”問題(詳後)。

以下介紹一下模型的基本情況,並通過一些具體語例介紹測試的過程。本次測試使用的是 GPT 模型,GPT 是一個自然語言生成模型,基於大規模古漢語語料庫預訓練的 GPT-2 模型有很好的古詩文語感,並能够對句子的成立概率進行估計。第一階段測試利用模型計算不同編聯結果的文本困惑度,困惑度越低,説明句子的産生概率越大。

以《邦家處位》的簡【1】爲例:

·【1】邦家處位,傾昃其天命,抑君臣必課以度。度君數臣,臣適督君,君唯聾狂,使臣猶迷。政事逆美,寵福逆惡。與介執

模型認爲困惑度最低的編聯項(※號項),正好是正確的那支竹簡:

·【1】邦家處位,傾昃其天命,抑君臣必課以度。度君數臣,臣適督君,君唯聾狂,使臣猶迷。政事逆美,寵福逆惡。與介執 (困惑度:332.5754)[①]

※【1】邦家處位,傾昃其天命,抑君臣必課以度。度君數臣,臣適督君,君唯聾狂,

① 該困惑度表示該簡本身的困惑度。

使臣猶迷。政事逆美，寵福逆惡。與介執＋【2】事，事是謀人（困惑度：285.2945）

• 【1】邦家處位，傾昃其天命，抑君臣必課以度。度君數臣，臣適瞀君，君唯聾狂，使臣猶迷。政事逆美，寵福逆惡。與介執＋【3】梗政。子立代父（困惑度：372.53802）

• 【1】邦家處位，傾昃其天命，抑君臣必課以度。度君數臣，臣適瞀君，君唯聾狂，使臣猶迷。政事逆美，寵福逆惡。與介執＋【8】人而不足用（困惑度：312.6786）

• 【1】邦家處位，傾昃其天命，抑君臣必課以度。度君數臣，臣適瞀君，君唯聾狂，使臣猶迷。政事逆美，寵福逆惡。與介執＋【11】必納功、崇能、有度（困惑度：349.20245）

第一階段的最終編聯結果見下表一：

表一　困惑度排序實驗結果

	Top-1 Accuracy	Top-3 Accuracy	MRR
Full(A)＋Trunc(B)	34.4%	58.1%	50.4%

説明：第一階段的模型，超過 1/3 的竹簡可以通過語感直接準確地找到正確的下一支簡，正確答案排名前 3 的準確率達到 58.1%，MRR 達到 0.5 以上。第一階段的實驗表明，通過語言模型的語感來進行竹簡編聯具有一定的可行性。

第二階段的最終編聯結果見下表二：

表二　多維度排序實驗結果

	Top-1 Accuracy	Top-3 Accuracy	MRR
perp_score＋similarity＋cost	42.1%	66.1%	57.5%

説明：第二階段的模型，在基於語感（perp_score，困惑度）的基礎上，又結合了語義相似度（similarity）的因素，並通過增加懲罰因子（cost）的方式處理了“萬能竹簡”的問題。最終結果裏，超過 40%的竹簡可以直接找到正確的下一支簡，正確答案排名前 3 的準確率超過了 2/3，MRR 接近 0.6 以上。相較於第一階段的結果，又有了非常明顯的效果提升。

五、分析與思考

在測試的過程中，每一個編聯錯誤都提供了一次重新分析問題的機會。每一次解

決錯誤的過程,也都讓我們去更深入地思考某些語言問題的實質。這部分將是人機協作所能帶給學術研究的最大價值。以下是我們在本次測試中的幾點思考。

(1) 基於更大的語境,連接更少的字詞。

在測試中我們發現,以不同長短的語言片段來判斷語感,會得到完全不同的結果,並直接影響最終的編聯結果。比如,同樣以《邦家處位》簡【1】爲例,上文表1中的結果是以簡【1】整支簡的内容(Full(A))作爲語境來預測下一支簡,如果只是輸入簡【1】的後半段(從“政事逆美”開始),並拼接部分後文(Trunc(B)),各支簡的困惑度都會顯著提高,排序也出現了明顯變化,甚至困惑度最低的簡(※號項)不再是正確項:

·【1】政事逆美,寵福逆惡。與介執 (困惑度:1141.8258)

·【1】政事逆美,寵福逆惡。與介執+【2】事,事是謀人 (困惑度:705.67413)

·【1】政事逆美,寵福逆惡。與介執+【3】梗政。子立代父 (困惑度:844.3941)

※【1】政事逆美,寵福逆惡。與介執+【8】人而不足用 (困惑度:698.5738)

·【1】政事逆美,寵福逆惡。與介執+【11】必納功、崇能、有度 (困惑度:1003.2555)

基於上述原因,我們嘗試了四種不同的文本截取方式來測試困惑度的差異:

Full(A)+Full(B):上簡的全部内容拼接下簡的全部内容

Trunc(A)+Full(B):上簡的部分内容拼接下簡的全部内容

Full(A)+Trunc(B):上簡的全部内容拼接下簡的部分内容

Trunc(A)+Trunc(B):上簡的部分内容拼接下簡的部分内容

下表三是采用四種不同截取方式進行編聯測試的最終結果:

表三 語感方法中的文本截取設置對比

排序		Top-1 Accuracy	Top-3 Accuracy	MRR
1	Full(A)+Trunc(B)	34.4%	58.1%	50.4%
2	Trunc(A)+Trunc(B)	31.9%	52.9%	47.5%
3	Full(A)+Full(B)	22.9%	42.0%	38.5%
4	Trunc(A)+Full(B)	17.2%	35.5%	32.7%

從表中數據可以看到,采用 Full(A)+Trunc(B)方式截取,各項準確率都最高,這與我們通常的認識存在一定差異。通常我們都會認爲,采用 Full(A)+Full(B)的方式最爲可靠,也就是把兩支簡上的所有文字信息都使用上最保險,但是基於數據來看,這

種方式的編聯正確率要比 Full(A)+Trunc(B)低了 10 多個百分點。這無疑給我們一些的啓示：編聯需要基於更大的語境，連接更少的字詞。上簡的内容越多，已知的確定性信息就越多，對下簡的内容就更容易預測。下簡與上簡的關聯度，越到後面就越少，過多注意後部的信息反倒不利於正確的編聯。

(2)“萬能竹簡”説明編聯工作要分類分步進行。

這裏所説的“萬能竹簡”，是指那些本身困惑度非常低的竹簡。比如《祭公之顧命》中的第 8 號簡：“以余小子揚文武之烈，揚成、康、昭主之烈。王曰：嗚呼，公，汝念哉！遜措乃”，在《祭公之顧命》中一共有 10 支簡的後面拼上 8 號簡都會得到最低困惑度，這與 8 號簡本身的低困惑度有關。一方面這支簡上的語句本身是比較完整的。另一方面，從語言學上來看，這支竹簡上的内容比較“百搭”，與其他各種竹簡放到一起都没有什麽“違和感”。在傳世文獻中，比如“不知其可也”“其誰曰不然”這樣的句子也都具有這種“百搭”的效果，要對其進行正確的排序，確實難度非常大。對於人類專家而言，這種“百搭”的竹簡一直是竹簡編聯中的難點，而對於計算機而言，這種困難同樣存在。從計算機的角度來看，這種困難的實質就是竹簡本身困惑度太低。由此可知，語義越完整、本身困惑度越低的竹簡越難以進行編聯。

爲了解決“萬能竹簡”的“百搭”問題，我們在第二階段測試中設計了懲罰因子，對這種現象進行了控制。編聯被分爲了兩輪進行，第一輪首先從各支竹簡中清查出“萬能竹簡”。第二輪再將懲罰因子加入“萬能竹簡”的得分計算公式，從而在懲罰之後，再對得分進行重新排序。這種做法相當於我們已經知道某些竹簡容易出錯，減少通過語感和語義來進行編聯。這種編聯方式也啓迪我們，在進行編聯工作時，應該分類分步進行。首先清理出“萬能竹簡”，然後再對剩餘的竹簡進行編聯，最後再通過適當方式來處理“萬能竹簡”的問題。

(3) 語感和語義各有優長。

語感和語義都對編聯工作有幫助作用，總體而言語感對編聯工作的作用更大(我們測試發現，僅依靠語義相似度進行編聯，正確率只有僅依靠語感編聯的一半左右)，語義相似度更多起到的是一種輔助作用。但語義相似度在編聯中也有自己獨特的作用。語義相似度的引入，除了作爲語感信息的補充外，對於殘缺文字的情形也可以較好處理。比如，以《保訓》【簡 2】爲例：

【2】爽，□□□□□□□□，王若曰：發，朕疾重甚，恐不汝及

這種文字殘缺的情況，無疑會對語感評估産生很大的影響。而利用語義相似度，就可以不受這種情況影響了。以下爲《保訓》中其他竹簡文本與【簡 2】的語義相似度排序，【簡 3】正是語義相似度最高的項：

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【3】訓。昔前人傳寶,必受之以詷,今朕疾允病,恐弗堪終,汝以書　(相似度：0.885)(正確竹簡排名最高)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【10】朕聞兹不舊,命未有所延。今汝祇備毋懈,其有所由矣。不　(相似度：0.857)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【11】及爾身受大命,敬哉,勿淫！日不足,惟宿不詳。　(相似度：0.825)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【9】微志弗忘,傳貽子孫,至于成唐,祇備不懈,用受大命。嗚呼！發,敬哉！　(相似度：0.796)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【5】不違于庶萬姓之多欲。厥有施于上下遠邇,乃易位邇稽,測　(相似度：0.7923)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【8】哉！昔微假中于河,以復有易,有易服厥罪,微無害,乃歸中于河。　(相似度：0.791)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【6】陰陽之物,咸順不擾。舜既得中,言不易實變名,身兹備惟　(相似度：0.782)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【4】受之。欽哉,勿淫！昔舜舊作小人,親耕于歷丘,恐求中,自稽厥志,　(相似度：0.773)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【7】允,翼翼不懈,用作三降之德。帝堯嘉之,用受厥緒。嗚呼！發,祇之　(相似度：0.773)

【2】爽,□□□□□□□□□,王若曰：發,朕疾重甚,恐不汝及＋【1】惟王五十年,不豫,王念日之多歷,恐墜寶訓,戊子,自沫水,己丑,昧　(相似度：0.758)

另外,當很多篇竹簡混在一起,要進行編聯時,首先使用的方法其實就是用語義編聯,這種方法的本質其實就是假設不同竹簡之間可能並不是直接接續,但用語義來編聯,可以允許這種信息的間隔。

以上這些問題,有的是專家編聯時可能也常常會遇到的問題,有的是專家編聯時可能未曾思考過的問題。人工智能利用計算機客觀地處理這些問題,讓很多問題更純粹地展現出來。

六、餘　　論

通過人工智能的竹簡編聯,我們對編聯工作的語言學本質有了更好的瞭解,對編聯工作的認識更加深入。接下來,我們計劃進一步分析實驗數據,並針對一些典型的排序

錯誤，對方法進行改進。同時，下一步我們還將引入一些其他類型的語言模型，比較不同模型的不同結果，争取使此項工作更加深入。



以上二維碼爲《清華簡》45 篇釋文的具體編聯結果。

附記：文中“思考與分析”部分所提出的問題，都是胡韌奮在測試過程中發現和嘗試解決的，莫伯峰對其進行了匯總並進行了初步分析。關於模型的細節和測試的具體過程，胡韌奮寫有詳細的技術報告，限於篇幅此不録入。