以數位佛經為基礎，研發古文獻研究所需資訊技術的經驗與成果

黃乾綱、李家名、釋法源*

一、以 CBETA 為對象開展數位人文的資訊技術研究

2005 年，時任中華佛學研究所圖書館館長杜正民老師，來臺大與資工系歐陽彥正教授討論「佛學資訊」的未來發展及相關的研究議題。彼時，由杜老師所掌舵的中華電子佛典協會（Chinese Buddhist Electronic Text Association，簡稱 CBETA），自 1998 年開始，已經數位化「大正新修大藏經」及「卍字續藏」兩部中文大藏經，共計近一億字的中文數位文字資料庫。這近一億字的數位化文件成為世界上最大的中文文字資料庫，且可以完全免費從網路下載[1]。

有著豐富文本數位化經驗的杜老師，開始關心資料庫的利用與應用問題。他希望以 CBETA 資料庫為基礎，讓「傳統佛學」領域和新的佛學與資訊結合的「佛學資訊」領域，在研究和服務等各方面能有跨入下一個世代的長足的進展。因此帶領了同時也在中華佛學院圖書館工作的釋法源法師和李家名先生來臺大尋求新技術的發展，並透過歐陽教授的引薦，來到臺大工科海洋系黃乾綱博士的實驗室並攻讀博士，就此開啟了此實驗室十年來在數位人文的研究發展。

二、本實驗室在三期國科會（現科技部）整合型計畫的研究背景

杜正民老師以「佛學資訊」為念所問的問題，其實也是「人文資訊」領域的核心問題：「除了數位化，資訊技術應該做什麼？」

有大量的文字資料庫，資訊技術可以做的事很多，但本實驗室有興趣的是「以資訊技術協助人文學者的研究」這個方向。有了大量的數位化資源，如何能

* 黃乾綱，國立臺灣大學工程科學及海洋工程學系副教授；李家名，國立臺灣大學工程科學及海洋工程學系博士候選人；釋法源，國立臺灣大學工程科學及海洋工程學系博士候選人。

[1] CBETA 中華電子佛典協會網址：http://www.cbeta.org/

提供突破傳統人文研究規模的資料處理及計算統計就是明顯的下一步議題。因此，「發展人文學者可能需要的資訊工具」及「工具背後的演算法」，便是本實驗室在「人文資訊」相關議題上的主要研究方向。

傳統的佛學研究學者都是透過大量的閱讀，倚靠個人的記憶或抄寫，來進行文史哲議題的資料蒐集、比較與討論。在電腦普及與大量佛典數位化之後，前述的研究過程得以透過電腦的使用，來改善查找或記錄的方便性及速度。但是對於典籍的影像處理、文字勘誤、及關聯性的研究等，仍然無法跳脫人力的方式。因此，我們將目標訂在如何提供新世代的人文研究工具。所謂的新世代研究工具，是指能夠協助人文研究者達成下列目標：

1. 相較過去的資料蒐集或資料比對，能以資訊技術大幅縮短時間。
2. 在研究規模上，不只要超越研究者個人的記憶或記錄能力，更想達到任何可以想像規模。
3. 最終期望能讓研究者自訂研究資訊的處理流程。

上述研究方向也在本實驗室研究生的背景上得到了融合。在本實驗室進行人文資訊研究的同學，多有「佛學」、「史學」、「語言學」等人文學系背景的學生。自 2012 年至今，我們與法鼓佛教學院合作，相繼執行了三期的國科會人文處（現為科技部人文司）的整合型計畫。這一系列的計畫亦隱含了我們的這個思路。

這三期計畫中，黃乾綱博士負責的子計畫名稱分別是：「漢籍數位文獻量化及關聯分析方法之研究」（2012-2015）、「數位化漢語形、音、義分析工具的研究」（2015-2017）、「漢籍引述文之自動擷取與關係識別」（2017-）。

在第一期計畫中，我們以「提升對漢籍語料的量化分析能力」為目標。此階段的核心項目，為基礎的中文索引及檢索工具的開發，以及中文自然語言處理（Natural Language Processing, NLP）中最基礎的中文斷詞研究。在計畫過程中，我們掌握了基礎的中文索引技術，並建立了包括檢索、統計及字碼處理等重要的 CBETA 線上工具，讓需要大範圍 CBETA 統計資料的人文學者獲益不少。

第二期計畫，從數位化佛典擴展至對漢語「形」、「音」、「義」三方面的資訊處理技術。進入數位典藏的成熟期，文字數位化工作需要更精確的光學識別（OCR）技術的協助，以大幅減少打字、校正的成本。在此階段，我們利用市場上的 OCR 軟體，針對古籍印刷品圖形文字的辨識，提出了強化的演算法。並整理了漢語音韻方面的資料庫。同時，並開發自動將原稿圖檔與電子資料庫連結的技術。因此文字影像處理技術是第二期計畫的核心研究。

第三期計畫以前面兩期的計畫成果為基礎，我們開始向語意研究邁進。我

們欲嘗試處理「語意」的方式與傳統不同。傳統自然語言處理領域對「語意」的處理，均是以專家訂定的「關係模型」(或稱「知識結構」)，例如 WordNet、SUMO (Suggested Upper Merged Ontology) 等詞彙網絡關係資料庫為基礎，再將欲分析的詞彙群放入這個關係模型，以詞彙間的網路關係作為語意分析的依據；換句話說，關係模型之外的語意就無法處理。我們認為，除了詞彙的關係模型外，語意更應該從前後文來理解。本期計畫的目標，便是要提出一個不被「關係模型」限制語意可能的範圍的「語意」研究方法。

除了依循「前後文」而非詞彙關係模型外，我們還認為語意和「目的」有關。因此，從具有相同「目的」的前後文來學習圍繞該目的文字中可能的語意，是我們的構想。以目前我們進行的「引述文語意研究」為例，首先確定要處理「引述」這個目的，並設定「引述的文字」為此語意結構的元素，同時定義「引述」的語意。假設我們定義引述文字間的「相似程度」及「方向」為引述文的語意，則當我們提出有效「擷取引述文字」及「分析引述文字定義內容」的演算法時，我們便有能力處理「引述」這個目的的語意。

三、已涉及的數位人文研究工具[2]

(一)建立擴大傳統人文研究範圍的資訊檢索工具

索引及搜尋是資訊技術可以協助突破傳統人文研究範圍的第一步。因應佛學及語言學者檢索 CBETA 語料的多項需求，我們在 2009 年建立了 CBETA Lexicon Tool 檢索工具網站 (http://cprg.esoe.ntu.edu.tw/cbetalexicon/)。在這些工具中，我們用 Suffix Array 索引建立 Concordance 語用檢索工具，同時與「佛教藏經目錄數位資料庫 (http://jinglu.cbeta.org)」合作，我們將藏經目錄資料庫中，經文翻譯的時間及地點資訊與文字檢索結果結合，提供了佛典文字的時間分布檢索工具，以及佛典文字的空間分布查詢工具。由於佛經中的缺字很多，在建立佛經全文索引的同時，我們使用了 Unicode 自建字碼區配合文字圖檔來解決 Web 介面的缺字問題，並同時提供了缺字檢索的解決方案。

CBETA Lexicon Tool 中的服務，從建置初期至今，都是漢語佛學研究領域中使用率相當高的資訊檢索工具。透過與 CBETA 的合作，我們開始思考資訊技術在人文領域的應用問題，也逐漸累積了從字碼、索引、文字影像到時空維度等的中文資料處理經驗，奠定了日後我們進行人文資訊相關研究的基礎。

[2] 本實驗室的工具網址：http://cklab.esoe.ntu.edu.tw

（二）無標記語料的自動斷詞演算法[3]

「斷詞」也稱「分詞」，運用電腦再透過演算法，將一個句子中的詞用空白（或其他符號）斷開或分開，英文學術期刊上大多是用「word segmentation」來表示。若以「將一個句子中的詞斷開或分開。」這個句子為例，經過電腦斷詞，我們期望得到「將 一個 句子 中 的 詞 斷開 或 分開。」這樣的結果。很明顯，原來在書寫上已經用空白符號將每個詞分開的語文（如英文），是不需要斷詞的。而中文這種字字相連，沒有符號告訴我們（電腦）句子中各個詞彙邊界的語文，就需要斷詞。但為什麼需要「斷詞」？

簡單來說，如果中文不經過斷詞，電腦能處理的最小單位的文字，就會是「句子」或是「段落」。很明顯，句子或段落不是組成語文的底層元素，我們至少需要能掌握「詞」，才能進行各種計算語言的工作和研究。

在我們進行第一期國科會計畫的時候（至今亦然），中文斷詞方面的研究，以中研院詞庫小組 CKIP（Chinese Knowledge and Information Processing）最有成果。在過去斷詞研究的成果中，針對古典中文的斷詞問題，提供了有限度的解決。為了掌握古典中文的特性，增加對 CBETA 自動斷詞的良率，我們便開始中文斷詞的研究。

當時所有的斷詞演算法都要以詞典的詞條或大量的人工標記為基礎，電腦以這些人工標記答案為目標來建立斷詞模型。而我們遇到最大的挑戰是，沒有適合的數位化過的佛學辭典，且沒有任何人工標記過的斷詞語料庫。

不過，正因為這樣的限制，讓我們重新思考中文斷詞問題，並提出了「無標記語料斷詞演算法」。這個斷詞演算法完全不需要任何詞典或人工標記的協助，只要有純文字，便可進行，且效果跟使用了六十萬字詞庫的CKIP不相上下。

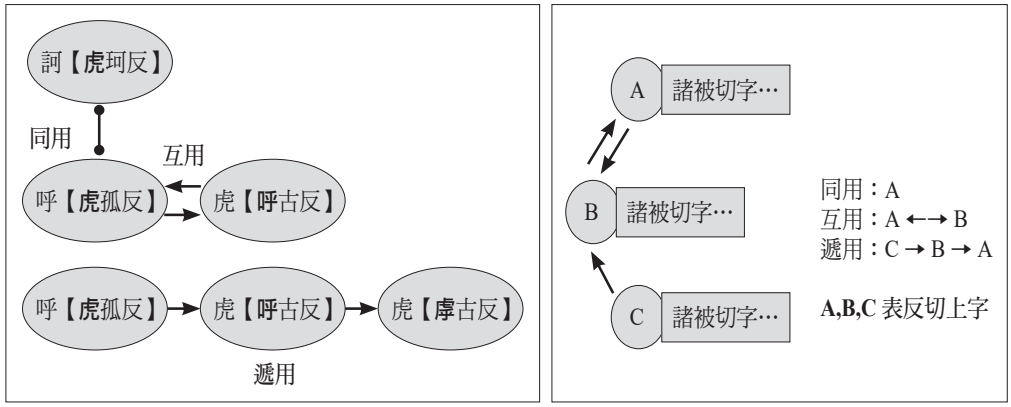
（三）一切經音義的反切分析工具

佛教經典是多種語言經長時期互譯的文獻集合。佛經語言研究涵蓋了包括文字、聲韻、訓詁、字義、詞彙、版本校勘等。其中「佛典音義」是過去在佛經語言研究中較少運用數位資料及資訊工具的領域。

音韻資料的處理方式，通常可分為已知音韻關係及未知音韻關係二部分，透過電腦，搜尋及統計音韻資料庫中的音釋資料，如被切字、反切上字、反切下字等。對於已知音韻關係可以驗證（Verify）其指定關係，而對於未知音韻關係，則可找出其規則（Rule）或趨勢（Trend）。運用數位化分析處理模型，我們

[3] 會議論文網址：https://aclanthology.info/papers/I13-1070/i13-1070

分析出《慧琳音義》中，反切音釋之「同用」、「互用」、「遞用」關係，使聲韻研究學者，可以更方便的繪製「系聯」關係與「韻圖」。圖一為抽象化的音義系聯關係模型圖。



圖一：系聯關係模型

傳統的系聯關係研究非常耗時費工，需要研究人員遍搜文獻，手動整理音韻的相互關係。為此，我們以 CBETA 文獻標記的優勢，建立了「音義音韻資料庫」、「音義索引書證資料庫」、「音義缺字資料庫」、「音義譯音詞彙資料庫」、「音義方言音資料庫」等類型資料庫，並建立了目前唯一的數位佛典音義檢索服務，包括：佛學音義詞典、音韻反切查詢、音韻系聯查詢、音韻系聯呈現等工具（http://cprg.esoe.ntu.edu.tw/yinyi/）。

這個佛典音義檢索服務系統，是以「一切經音義」為總綱，兼收其各項可能相關的語言材料，能呈現的不僅是一張張的語音表，更能透過關聯式資料庫及索引技術，快速檢索書中的反切、詞條數目以及相關的釋義，提供研究者全面的材料，以求提供學者從事「音義聲韻」研究之便利。

（四）增進 OCR 光學辨識文字技術

由於古籍掃描影像文字數位化工作需要光學識別（OCR）技術的協助，以及人文研究者對文字原稿的需求，這使得我們投入了中文古籍文字影像處理的研究。中文古籍是以現代印刷技術出現之前的文獻為範圍，其中包括帛書、簡牘、石刻、寫本、雕版及活字印刷書籍。這些文獻有別於現代印刷品之處，是文獻內的文字不論是抄或是刻，都是以手工完成，即使是活字印刷的書籍，相同的字也不一定會完全一樣。

古籍除了在字型上無法完全相同外，因為年代及複印技術等的關係，文字影像也經常有部分破損的情況。再加上中文古籍有傳統的分欄及雙行夾註等現代出版品不常使用的排版格式，這些都是造成以電腦印刷字體為範本的現有光學文字辨識軟體，無法有效的進行古籍文字辨識的原因。

現有的古籍文字辨識研究，多是以一般的文字影像辨識軟體的辨識結果為基礎，再輔以語言模型進行校對，以提高辨識結果的良率。我們在前述第二期的計畫中增進古籍文字 OCR 正確率的方法包括：加強切字正確率、利用外部語料庫的語言模型以及透過文字影像分群來輔助辨識結果的校對等等。

自 2016 以來，深層學習工具在影像辨識上的應用快速的普及，我們也將深層學習演算法應用在文字影像辨識上，並得到不錯的效果。

（五）佛經文獻互引

引述偵測（quotation detection）研究是從大量的文字資料中，辨識出引述句，以便進行進一步的引述分析（quotation analysis）。現有的引述偵測研究，是以定義引述句法（quotation syntax）偵測引述句法中的引述線索為主要方法。引述句法分成引述來源（source）、引述線索（cue）及引述文（quotation content）等三個部分。

佛經中有大量的引述文字，尋找、確認引述文間的關係，是佛學研究者的一項重要工作。由於古漢文中有許多專門為某經典所做的「論」、「注」或「疏」等類型的著作，引述狀況複雜，現有的引述句法的定義，無法涵蓋古漢文中所有的引述情況。因此在上述基本的引述句法的定義範圍外，古漢文中還可觀察到只有引述文沒有引述線索及引述來源的引述句，以及為了描述經典中文字位置的引述句。根據觀察，描述經典文字位置的引述句可能包含以下元素：（1）數詞；（2）結構文字（如：「句」、「行」、「頁」等）；（3）指向文字（例如：「以下」、「自」、「至」等）；以及（4）經文中的短詞等。以「自爾時開始以下三行」為例，其中「爾時」為經文中的短詞；「自」、「開始」、「以下」等為指向詞；「三」為數字詞；「行」為結構詞。

根據上述分析，我們將古漢文的引述文分為三種類型：

1. 完整引文的引述句：有完整引述對應字串，可能有引述線索的字詞。
2. 不完整引文引述句：無完整的引述對應字串或字串太短，可能有引述線索的字詞。
3. 描述位置的引述句：使用描述位置的結構及指向性的文字。

除了描述文字位置的引述，我們假設所有引述文跟被引述文間有一定程度的相似性，並設計了四種引述偵測（quotation Detection）方式來偵測並抽取古文中的引述文：

1. 引述內容比對，找出出處文本與引用文本中匹配的字串，作為候選引述句。
2. 線索詞搜尋，根據可能的線索詞，找到候選引述句。
3. 引述規則搜尋，根據人工經驗提供的引述規則，找到候選引述句。
4. 引述位置探勘，根據可能的引述規則及引述位置，分別找到候選引述句。

抽取了引述文後，再依照外部資訊及內部資訊來確認引述文之間的關係。外部資料包括經文的成書時間、作者年代等後設資料。內部資料則是引述文前後的用語。

從人文學者研究的需求探討，確立「引述文互引關係」為研究目標。經過定義引述文形式、設計抽取引述文的方法、分析引述文互引關係，我們得到一個在時空架構下，佛經與佛經內容之間的網狀關係。這個網狀關係是一個純粹「以文本為基礎」、「統計為方法」所建立的客觀「知識結構」。

上述從需求出發所建立的知識結構，和透過專家知識所建立的知識結構，這兩種知識結構的特性比較與適用範圍，是我們繼續探討「語意」的重要工作。

四、回顧與展望

在與人文學者合作的過程中，我們累積了處理「中文自然語言處理領域」的完整經驗。從文字數位化、文字索引、缺字處理、缺字檢索、中文分詞、時空維度應用、文字影像辨識，到知識結構及語意分析。從服務人文研究的角度來看，我們觸及的服務包括：全文檢索、語用檢索、文字時空分析、音義系聯檢索、佛經分詞、影像辨識數位化、佛學知識結構建立等等。

從資訊在人文中的角色思考，在累積了這些技術與工具後，為研究學者建立多功能的研究平臺可以是下個階段的發展。在前述三期的科技部整合計畫中，我們與法鼓佛教學院圖書資訊館合作。法鼓圖書資訊館著重在發展整合技術與工具的研究平臺（http://cbeta-rp.dila.edu.tw），我們則傾向於個別技術上的突破。

從資訊技術角度，待突破的方面還很多。中文分詞的演算法很多，但能成功運用自然世界文字的工具（非實驗用語料庫）卻很少。語意分析方面，直到我們提出以「使用需要為目標」及「文獻前後文為基礎」的嘗試性方向前，所有研究均是依賴詞彙基礎的知識結構資料庫。文字影像辨識方面，手寫的影像辨識需求日益提高，且增加相似字型辨識正確率的挑戰仍然很大。

經過了數年為人文資訊領域服務，從本篇介紹我們的研究與成果，基本上回答了「資訊技術可以為人文研究做什麼？」這個問題。現在，我們希望能專注在中文自然語言處理的「無標記詞、句辨識」、「文字影像辨識」與「無標記語意分析」研究上，將它們做得更好。