發展正體中文專利審查數位工具之分析討論

張長軾 [*]、陳宇超 [**]、陳逸 [***]

[*] 作者現為經濟部智慧財產局專利助理審查官。
[**] 作者現為經濟部智慧財產局專利助理審查官。
[***] 作者現為經濟部智慧財產局專利助理審查官。
本文相關論述僅為一般研究探討，不代表任職單位之意見。

摘要

本文簡述本波人工智慧技術發展浪潮中數個重要分支技術：機器學習、深度學習，以及自然語言處理；由於我國智慧財產權相關文書主要以中文記載，本文進一步介紹中文自然語言處理之概念，及將其應用於中文專利文件分析或處理上可能需要的考量；最後根據現有之觀察提出分析討論，期促進世界上各智慧財產權主管機關積極引入數位工具至其工作場域的可能。

關鍵字：數位工具、人工智慧、第四次工業革命、機器學習、深度學習、自然語言處理、中文自然語言處理

Digital Tools、Artificial Intelligence、Fourth Industrial Revolution、Machine Learning、Deep Learning、Natural Language Processing、Chinese Natural Language Processing

壹、前言

我國專利文書以正體中文記載，為一大特色，因此本文嘗試初步討論運用中文自然語言處理（Natural Language Processing, NLP）技術至專利文件分析或處理等相關工作的可行性：藉由檢視人工智慧（Artificial Intelligence, AI）當今現實（State of the Art）技術層面及中文 NLP 技術之發展現況，分析中文 NLP 是否有能力或有潛力支持以 AI 檢索專利引證之所需，以及如何謀定關鍵資料準備、開始蒐集工作，以因應洶湧而至的專利 AI 相容化浪潮，係本文設定目標。

貳、機器學習、深度學習及自然語言處理

首先簡述機器學習、深度學習及自然語言處理等三個 AI 技術之發展、應用領域[1]，以及機器學習與深度學習兩者間的關係。

一、機器學習

如圖 1 所示，「機器學習」（Machine Learning, ML）是 AI 的子集，包含可自行從資料學習而增進自身效能的各種 AI 技術。ML 系統能從以「資料」代表的「實際環境」、「實務經驗」中學習，自我改進，故無須要求設計者在設計的最初就決定了系統最終的樣貌；一般而言，建立、運用 ML 系統之流程為：在建立時，由設計者（至少包括程式編寫者，可加入問題所屬領域之專家等）針對「問題」設定系統模型（一模型例如：包含種類相異之多個函數，函數個別之待定參數，及組合多個函數之輸出的待定權重等）及目標函數，根據訓練資料及目標函數訓練（進行最佳化（optimization）演算）而求得最佳化的（optimized）系統模型；在運用時，以上述最佳化模型設定系統，讓系統根據「測試資料」判斷給定「問題」的答案。

1 Paula Carreirão, Unraveling the Enigmatic Artificial Intelligence and Its Wonders, ASKSUITE, 1, https://asksuite.com/blog/backstage-of-artificial-intelligence/ (last visited Oct. 26, 2021).

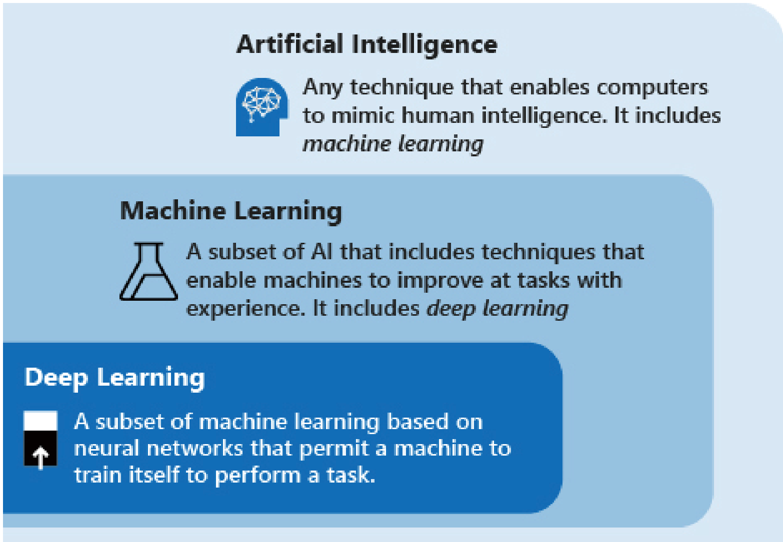

圖 1　人工智慧、機器學習與深度學習的相互關係[2]

規劃一個 ML 系統所須給定之三元件為：演算法（algorithm）、資料集（dataset）及特徵（features）表示形式。例如我們希望以 ML 架構訓練一個分辨貓、狗照片的辨識系統，則須先準備一組資料集，包含許多貓、狗的照片；選取訓練資料須配合問題的情境，例如不限地理區域的辨識系統會希望資料集中有各種貓、狗的照片；接著憑藉常識或領域專家知識為每一筆資料選定、萃取、標註特定表示形式之特徵資料（特徵萃取，Feature Extraction），例如選擇將貓、狗之外型輪廓、體型大小、毛色、花紋……等性質轉為數值函數輸出，作為特徵；再以特徵資料作為訓練材料，完成模型訓練；實際應用時，將訓練所得之模型用於判斷後續之輸入，係為貓或狗的照片。

二、深度學習

如圖 1，深度學習（Deep Learning, DL）是具有特定實現方式的 ML 技術，該實現方式與其他 ML 技術之差異主要在於：（一）以多層次神經網路（multilayered neural networks）為架構核心（如圖 2），（二）不靠人為設定，自行學習如何擷取有利於其任務的特徵資料（如圖 3）。此類 ML 之所以被特稱為「深度學習」，

[2] Deep learning vs. machine learning in Azure Machine Learning, 1, https://docs.microsoft.com/en-us/azure/machine-learning/concept-deep-learning-vs-machine-learning (last visited Nov. 25, 2021).

源自[3、4]在 2012 年 ILSVRC 競賽中有驚人表現而備受矚目、帶起風潮之辛頓（G. Hinton）團隊[5]所採用之深度卷積神經網路（Deep Convolutional Neural Network）[6、7]，其多層次外貌所呈現之「深度」（depth）[8]。多層次神經網路具備高度非線性及參數數量擴充性，將其作為核心，或許是使 DL 系統成為目前最能模擬生物認知行為之 ML 技術的原因。

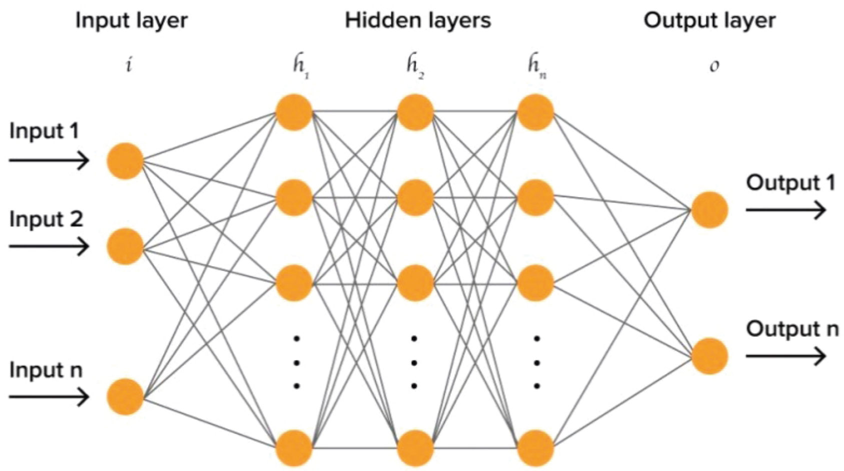

圖 2　深度學習的核心、多層次神經網路之示意圖[9]

[3] 陳昇瑋、溫怡玲，人工智慧在台灣：產業轉型的契機與挑戰，頁 76-80 ，天下雜誌股份有限公司，2019 年 6 月。

[4] Lynn，神經網路的復興：重回風口的深度學習，1 ，https://www.stockfeel.com.tw/%e7%a5%9e%e7%b6%93%e7%b6%b2%e8%b7%af%e7%9a%84%e5%be%a9%e8%88%88%ef%bc%9a%e9%87%8d%e5%9b%9e%e9%a2%a8%e5%8f%a3%e7%9a%84%e6%b7%b1%e5%ba%a6%e5%ad%-b8%e7%bf%92/?utm_source=map_recommend&from_post_id=87673（最後瀏覽日：2021/11/25）。

[5] Alex Krizhevsky & Ilya Sutskever & Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25 (NIPS 2012), 1-9, https://papers.nips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf (last visited Nov. 25, 2021).

[6] 同前註。

[7] 維基百科，卷積神經網路，https://zh.wikipedia.org/wiki/%E5%8D%B7%E7%A7%AF%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C（最後瀏覽日：2021/11/24）。

[8] Shang-Jou Ian（Terry 毛），3 分鐘了解機器學習在學什麼，1 ，https://medium.com/@troy801125/machine-learning-%E4%BA%BA%E5%B7%A5%E6%99%BA%E6%85%A7%E5%92%8C%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E6%98%AF%E4%BB%80%E9%BA%BC-49a6ba41ab3e（最後瀏覽日：2021/11/25）。

[9] Serokell, Artificial Intelligence vs. Machine Learning vs. Deep Learning: What's the Difference, ARTIFICIAL INTELLIGENCE, 1, https://ai.plainenglish.io/artificial-intelligence-vs-machine-learning-vs-deep-learning-whats-the-difference-dccce18efe7f (last visited Oct. 26, 2021).

圖 3[10] 比較 ML 與 DL 機制在進行工作時之流程差異，其中[11]，

ML：輸入資料→ 人為設定特徵擷取 → 模型 → 輸出答案；

DL：輸入資料→ 神經網路（自學特徵擷取）→輸出答案。

亦即，DL 捨去人為設定的特徵擷取，改讓神經網路自行從大量資料中學到如何擷取最有助於其進行任務之特徵。例如若以 DL 建立前述辨識系統，則 DL 模型於訓練時將自行根據給定的照片資料集，學習如何擷取最有助於「區分貓、狗照片」之特徵。

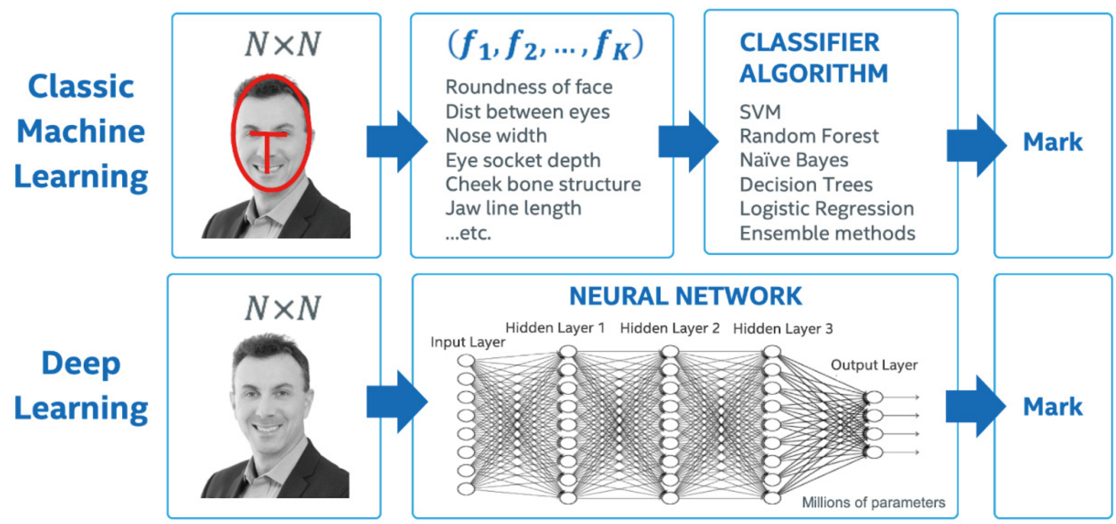

圖 3　機器學習與深度學習進行工作之流程差異示意圖

---

10 MaryT_Intel, The Difference Between Artificial Intelligence, Machine Learning and Deep Learning, 1, https://www.intel.com/content/www/us/en/artificial-intelligence/posts/difference-between-ai-machine-learning-deep-learning.html (last visited Nov. 25, 2021).

11 Tommy Huang，什麼是人工智慧、機器學習和深度學習，1，https://chih-sheng-huang821.medium.com/%E4%BB%80%E9%BA%BC%E6%98%AF%E4%BA%BA%E5%B7%A5%E6%99%BA%E6%85%A7-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E5%92%8C%E6%B7%B1%E5%BA%A6%E5%AD%B8%E7%BF%92-587e6a0dc72a（最後瀏覽日：2021/11/26）。

三、自然語言處理

自然語言處理[12]是用於處理人類日常生活中使用的「自然語言」的多種資訊技術，如圖 4 所示，該些技術之相關知識背景包含 AI 和語言學（Linguistics）等之領域，故自然語言處理為語言學與人工智慧之綜合科學；相較於計算語言學（Computational Linguistics）強調從語言學觀點著眼的技術，NLP 更偏向從工程學角度出發的語言處理，包括諸如將數據庫中的信息轉換為自然語言，以及從自然語言句子中提取出所需之資訊（透過計算機理解、供給計算機運用）之類的處理等之多個方面及步驟，基本上包括有認知、理解、生成等部分。

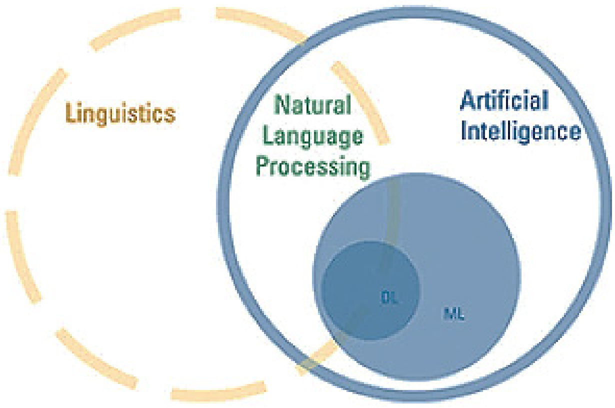

圖 4　自然語言處理係語言學與人工智慧之綜合科學[13]

[12] フリー百科事典『ウィキペディア（Wikipedia）』，自然言語処理，1，https://ja.wikipedia.org/wiki/%E8%87%AA%E7%84%B6%E8%A8%80%E8%AA%9E%E5%87%A6%E7%90%86（最後瀏覽日：2021/10/26）。

[13] RETRESCO, HOW TO AI: NAVIGATING THE BUZZWORDS OF ARTIFICIAL INTELLIGENCE, 1, https://www.retresco.com/how-to-ai-natural-language-processing/ (last visited Oct. 26, 2021).

參、將自然語言處理應用至中文專利審查之議題分析

現今三大 AI 應用領域中，語音辨識[14]及電腦視覺[15]兩者已有重大突破，唯獨號稱是 AI 皇冠寶石[16、17、18、19]的自然語言處理[20]未竟全功，例如中文 NLP 中的語意計算準確度就尚待精進。

許多人說 NLP 為 AI ／ ML ／ DL 之聖杯[21、22、23]，並看好未來幾年 NLP 技術的進步將成為推動 AI 領域前進的領頭羊[24]。而在專利文件之機器、智慧分析方面，NLP 也有其潛力；近年歐美等非亞洲語系人士對於檢索中文、日文、韓文等亞洲語言之專利，熱度不斷高漲，日本特許廳（Japan Patent Office, JPO）及韓國智慧財產局（Korea Intellectual Property Office, KIPO）已嘗試透過實施包含 NLP 在內之機器翻譯（Machine Translation, MT），建立與外文人士之自動化溝通管道，以方便非亞洲語系人士進行檢索，同時墊高本國專利之話語權優勢。普及使用中文為兩岸三地（我國、香港、中國大陸）之特色，而我國獨占正體中文之樞紐、與外界互通跳板之位置，加上我國優秀 AI 學者眾多，其所能扮演之角色值得重視。

---

14 維基百科，語音辨識，https://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3%E8%AF%86%E5%88%AB（最後瀏覽日：2021/10/26）。

15 維基百科，電腦視覺，https://zh.wikipedia.org/wiki/%E8%AE%A1%E7%AE%97%E6%9C%BA%E8%A7%86%E8%A7%89（最後瀏覽日：2021/10/26）。

16 GeekWire, Bill Gates - The GeekWire Interview, YOUTUBE, https://www.youtube.com/watch?v=BfOHUVg2FcA (last visited Oct. 26, 2021).

17 RavenPack, Why Natural Language Processing is the Crown Jewel of A.I., 1, https://www.ravenpack.com/page/natural-language-processing-crown-jewel-ai/ (last visited Oct. 26, 2021).

18 ProgrammerSought, Natural language processing topic, 1, https://www.programmersought.com/article/3715525177/ (last visited Oct. 26, 2021).

19 Linda Xu & Xue Hua, Why natural language processing is AI's jewel in the crown, Technology Insights, 1, https://www.huawei.com/en/technology-insights/publications/winwin/33/why-natural-language-processing-is-ais-jewel-in-the-crown (last visited Oct. 26, 2021).

20 維基百科，自然語言處理，https://zh.wikipedia.org/wiki/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86（最後瀏覽日：2021/10/26）。

21 新智元（發表于科技），「谷歌 Facebook 稱霸 AI 全息戰圖」一圖看懂全球 AI 巨頭實驗室布局，https://kknews.cc/tech/yy6n3ak.html（最後瀏覽日：2021/10/26）。

22 Tom Taulli, Microsoft Bets $1 Billion On The Holy Grail Of AI (Artificial Intelligence), FORBES, 1, https://www.forbes.com/sites/tomtaulli/2019/07/27/microsoft-bets-1-billion-on-the-holy-grail-of-ai-artificial-intelligence/?sh=7db7aa121aa8 (last visited Oct. 26, 2021).

23 Staff, Bill Gates calls artificial intelligence 'the holy grail', BECKER'S HEALTH IT, 1, https://www.beckershospitalreview.com/healthcare-information-technology/bill-gates-calls-artificial-intelligence-the-holy-grail.html (last visited Oct. 26, 2021).

24 Ming Zhou, Nan Duan, Furu Wei, Shujie Liu & Dongdong Zhang, The Next 10 Years Look Golden for Natural Language Processing Research, MICROSOFT RESEARCH Lab – Asia, 1, https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/articles/next-10-years-natural-language-processing/ (last visited Oct. 26, 2021).

中文語言學在專利說明書分析上的應用，體現在專利工作者可以輕易以人為方式處理中文分析中最困難的語意計算、慣用成語、熟語等問題。例如一般文書會用到的誇張修辭法（Hyperbole），是藉由違反語用學（Pragmatics）[25]之合作原則（Cooperative Principle）的質量準則（Quality Maxim）所欲引起的會話隱含（Conversational Implicature）[26]、亦即「語用否定」，來達到「語意肯定」的強化效果，實例包括「差點昏倒」相對於「差點沒昏倒」、「真是好快樂！」相對於「真是好不快樂！」等，利用寫法差異來表達「強調」之意；若此類用法出現在專利說明書中，經專利工作者審視後，透過在生活經驗中自然累積的語言學背景知識，可輕易判定該語句之原意、或注意到可能發生的問題；但如欲透過機器，無論以傳統或 ML ／ DL 之 NLP 技術來自動分析、偵測此類修飾寫法，以現今的技術而言仍多有困難。

又如學測試題中曾出現之歧義句「咬死了獵人的狗」（圖 5）等，為生活中常見的自然語言實例，人類之閱讀者可以輕易理解語句之真義、或注意到應進一步釐清的地方，但無論是傳統或是現今的 ML ／ DL 之 NLP 技術，處理此類問題上則可能亦力有未逮。

[25] 蘇以文，語用學的發展與展望，人文與社會科學簡訊 12 卷 1 期，頁 128-135 ，https://www.most.gov.tw/most/attachments/ca24e3be-7acf-4742-af04-82b85d6961b4（最後瀏覽日：2021/11/24）。

[26] 鄧守信、葉彥君，台灣地區公眾批評語及批評策略研究（Studies on the Critique Strategies of Public Commentary in Taiwan），頁 7-9 ，http://rportal.lib.ntnu.edu.tw:80/handle/20.500.12235/86342（最後瀏覽日：2021/11/24）。

38 「咬死了獵人的狗」是個歧義句，如右所示，是因為語法結構關係不同而造成了語義差異。下列文句，屬於此種歧義句的是：

(A)他沒有做不好的事情
(B)他知道這件事不要緊
(C)小明借來的字典沒用
(D)王同學是轉學生，很多人不認識他
(E)陳老師和李老師的學生，來自不同學校

- 〔咬死了獵人／的／狗〕意謂：「有隻狗咬死了獵人」。
- 〔咬死了／獵人的狗〕意謂：「某動物咬死了獵人所養的狗」。

答　案：ABE

解　析：(A)他沒有做不好的事情：〔他沒有／做不好的事情〕意謂：他什麼事情都做得很好；〔他沒有做／不好的事情〕意謂：他沒有做壞事。
(B)他知道這件事不要緊：〔他知道／這件事不要緊〕意謂：他知道這件事不重要；〔他知道這件事／不要緊〕意謂：讓他知道這件事沒關係。
(E)陳老師和李老師的學生，來自不同學校：〔陳老師／和李老師的學生，來自不同學校〕意謂：陳老師此人和李老師的學生，來自不同學校；〔陳老師和李老師／的學生，來自不同學校〕意謂：這兩位老師的學生各來自於不同的學校。

圖 5　大學考試中心所公告之一學測試題及答案[27]

中文專利說明書、請求項之記載形式規定於專利法、專利法施行細則及專利審查基準，惟這些規範的體現與否常須借助審查人力進行檢視及確認。經過人為檢視後，應可將說明書、申請專利範圍等視為自然語言的子集合，有利於後續中文自然語言處理技術應用。

本文定位不在於介紹語言學，但語言學實為分析自然語言、發展自然語言處理技術之基礎。相較於英文等西方語言、符號拼音語言而言，中文的語言學[28]發展較晚，與英文語言學之對應元素，包括語法、各類字典等，完成度相對較低，能提供給中文自然語言處理技術發展之支援也難免較少，尚有許多可發揮的地方。

---

[27] 程美鐘，王麗蓉，國文科 107 年學測考情最前線，龍騰文化事業股份有限公司，107 年 2 月 2 日，https://www.lungteng.com.tw/Web/Upload/Upload_File/Source14/107%E5%AD%B8%E6%B8%AC%E8%80%83%E6%83%85%E6%9C%80%E5%89%8D%E7%B7%9A%E5%9C%8B%E6%96%87%E7%A7%91.pdf（最後瀏覽日：2021/10/26）。

[28] 游適宏，2006-2008 年教育部對外華語教學能力認證「漢語語言學」試題內容解析，人文社會學報第五期，頁 105-128 ，2009 年 3 月，http://ir.lib.ntust.edu.tw/bitstream/987654321/14595/1/2009-teaching+Chinese+as+a+second+language.pdf（最後瀏覽日：2021/10/26）。

一、中文自然語言處理框架

圖 6 所示為中文自然語言處理的一般架構流程，圖上方之左邊部分是 NLP 開發者較為熟悉的傳統語法層面模組，包括中文斷詞（亦有研究者稱為「分詞」；本文以下統一稱為斷詞）、詞性標註及句法解析等功能模組；上方右邊部分則為語意層面模組，其中命名實體識別[29]（Named Entity Recognition, NER）主要用來識別語料庫中的專有名詞和未登錄詞，如人名、地名、組織機構名稱等，故若專利說明書在揭露所述發明之技術特徵（Technical Feature, TF）時，由記載者賦予了一些特定的專有名詞，則亦歸類於此；惟考慮一般中文語料庫，例如中央研究院現代漢語標記語料庫 4.0[30]，其含括之詞數超過 1,100 萬、字數超過 1,700 萬，與該語料庫之大小相較，我們主要使用之 IPC、CPC 等專利分類號系統，在各細階分類下之文本數量尚不足以建立可達到實用程度之 NER 模組。

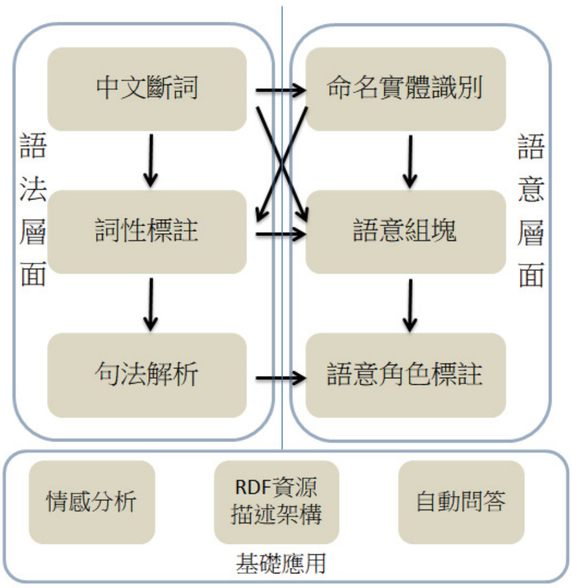

圖 6　中文自然語言處理的一般架構流程[31]

[29] 維基百科，命名實體識別，https://zh.wikipedia.org/wiki/%E5%91%BD%E5%90%8D%E5%AE%9E%E4%BD%93%E8%AF%86%E5%88%AB（最後瀏覽日：2021/10/26）。

[30] 中央研究院，中央研究院現代漢語標記語料庫 4.0 版簡介，http://asbc.iis.sinica.edu.tw/（最後瀏覽日：2021/10/26）。

[31] 鄭捷，自然語言處理：用人工智慧看懂中文，頁 1-9，第 1 章圖 1.2，佳魁資訊，2018 年 12 月。

圖 6 所示語意組塊係用於確定由一個以上詞彙構成的短語結構，即「短語級別」的標註，主要用於識別名詞性短語、動詞性短語、介詞短語及其他類型的短語結構。圖 6 之語意角色標註則以句中的謂語動詞為中心，預測句中各語法成分的語意特徵，是句子解析的最後環節，也是句子級別語意研究的重要里程碑。

綜上所述，在概念上，從前端中文斷詞階段到後端語意角色標註階段經歷了多個依次串連的模組，各模組難以完全避免的錯誤率一路累積下來，降低了語意角色標註的精度，這應是現今語意角色標註技術未達商業應用所需程度的主要原因。

若以極度簡化的比喻表述，中文 NLP 機制可視為英文 NLP 機制加上了中文斷詞之預處理；故在圖 6 之中文 NLP 處理流程中，除了左上角的中文斷詞外，其餘均可初步想像為與英文 NLP 之對應組塊的技術概念類似或相同。當然，語言的差異可能帶來很不同的語言學架構，因而在分別應用於中文、英文時，上述的對應組塊在實作上會很不同，惟此亦再次凸顯瞭解「中文語言學」對於開發、運用中文 NLP 系統之重要性。

如果嘗試另一途徑，在完成斷詞步驟後，透過相較於中文NLP更為成熟的「中翻英」機器翻譯功能，將中文專利說明書翻譯成英文、再以英文 NLP 來處理，是否可行？本文認為，此途徑會遇到以下問題：

（一）直接採用英文 NLP，後續分析、檢索之對象恐會侷限於英文文獻，仍無法處理及利用包括正體／簡體中文、日文、韓文……等其他日益豐富之非英文技術文獻。

（二）若以「中對英」機器翻譯將一中文篇章翻譯成英文，再將結果以「英對中」機器翻譯成中文，往往會發現，中對英、英對中之機器翻譯並非互為反函數，雙向並非可逆運算；例如，若將【回首向來蕭瑟處，歸去，也無風雨也無晴。】經由 Google 或 DeepL[32]（號稱更能有效翻譯文言文）翻譯成英文【Google：*“Looking back at the bleak place, there is no wind, rain, or sunshine when I return”*. DeepL：*“When I look*

32 DeepL，DeepL 翻譯：全世界最準確的翻譯，https://www.deepl.com/translator（最後瀏覽日：2021/10/26）。

back at the place where I've always been, I go back, and there is no wind, rain or sunshine".】，再將所得譯文以原翻譯器翻成中文【Google：回首那淒涼之地，歸來時無風無雨，亦無陽光。DeepL：當我回頭看我一直在的地方，我回去了，沒有風，沒有雨，沒有陽光。】，將無法得到原來的語句；此現象或許可以簡單理解為，一個語句在一種語言中所代表的意義，並非必然僅在於字面，故未必能透過以資料統計為基礎的機器翻譯而完整地向其他種語言傳遞。故追本溯源，以中文 NLP 技術處理中文申請書文本，應屬合理，惟所欠奉的中文語言學部分，仍需要有本國語言學家之大力支援。

近幾年有許多開源之中文 NLP 基礎框架被整合發布出來，所謂基礎框架可以簡略視為，程式開發者提供給公眾、可引用來快速構建模型的基本通用工具包或程式庫；茲將部分主流、知名之中文 NLP 全系列處理框架、斷詞框架及其功能屬性列示於表 1[33]，藉以表達現有中文 NLP 框架資源之豐富及多樣性。

表 1　可用於中文 NLP 開發之全系列處理框架或斷詞（分詞）框架

中文 NLP 框架	NLP 功能類型	技術特徵或現有模型版本	開源碼網址
中研院中文剖析器／斷詞系統	斷詞	中文剖析器線上測試，中文斷詞系統	http://parser.iis.sinica.edu.tw/, http://ckipsvr.iis.sinica.edu.tw/
中研院 CKIP Tagger 中文斷詞	斷詞	CKIP LAB 中文詞知識庫小組自行研發，於 2019 年開源釋出	https://github.com/CkiplAB/CkipTagger
中研院 CKIP CoreNLP		斷詞暨命名實體辨識系統，剖析系統，指代消解介紹，標記列表	https://ckip.iis.sinica.edu.tw/service/corenlp/
中研院 CKIP Transformer	Transformer	ALBERT Tiny, ALBERT Base, BERT Base	https://github.com/ckiplab/ckip-transformers

（續下頁）

[33] Olga Davydova, 13 Deep Learning Frameworks for Natural Language Processing in Python, Data Monsters, 1, https://medium.com/@datamonsters/13-deep-learning-frameworks-for-natural-language-processing-in-python-2b84a6b6cd98 (last visited Oct. 26, 2021).

中文 NLP 框架	NLP 功能類型	技術特徵或現有模型版本	開源碼網址
Stanford NLP	全系列	中文斷詞、詞性標註、未登錄詞識別、句法分析等	http://nlp.stanford.edu/software/index.shtml
結巴分詞	分詞（亦即斷詞）	小規模中文分詞	https://github.com/fxsjy/jieba
HanLP	全系列	中文分詞、句法分析等	https://github.com/hankcs/HanLP
FudanNLP	全系列	中文分詞、句法分析等	https://code.google.com/p/fudannlp/
哈工大 LTP 3.X	全系列	中文分詞、詞性標註、未登錄詞識別、句法分析、語意角色標註	https://github.com/HIT-SCIR/ltp/releases
哈工大訊飛聯合實驗室	BERT	Pre-Training with Whole Word Masking for Chinese BERT（中文 BERT-wwm 系列模型）	https://github.com/ymcui/Chinese-BERT-wwm

二、中文申請專利範圍斷詞

中文 NLP 的第一項關鍵核心技術是中文詞彙自動切分，即中文「斷詞」。中文書寫自以單音詞為主的文言文演化至以複音詞為主的白話文，從未採行例如拉丁語系等拼音語言之「分詞連寫」[34] 形式。以英語書寫為例，“I don't think so.”以空格（space）「分」隔不同的「詞」（word）而呈現語句（sentence），每一詞又由一或多個字母（alphabet）之「連寫」構成；但在中文的對應記載則為「我不這樣認為。」，而非「我 不 這樣 認為。」，「我」、「不」、「這樣」、「認為」此四個詞並未被分開書寫；故當電腦讀入一整個段落的中文篇章，它必須先有能力根據這一連串、幾無空隙的中文字（characters）定義出中文詞的邊界，才能繼續分析由這些詞所構成的語句、段落、篇章等之內容含義。

[34] 吳文超，漢語拼音分詞連寫的重要性，http://www.huayuqiao.org/articles/wuwuchao/wwc01.htm（最後瀏覽日：2021/10/26）。

以電腦自動進行中文斷詞之技術，經歷二十多年的研究，基本上分為三個流派[35]：

（一）基於詞典的斷詞：將文件中的字串與詞典中的詞條逐一匹配，若匹配成功則予以切分，否則不予切分。此種斷詞法，實現簡單、實用性強，最大的缺點是「詞典」本身的完備性未能保證。在專利文書分析的相關應用上，我們認為，若能為每一技術領域（例如定義為適當階數之每一專利 IPC ／ CPC 分類號等）都建立各自專屬的詞典及搭配的同義詞詞典，且該些詞典之正確性、完備性經過該領域專家確認的前提下，則此類斷詞法很有機會足以實際應用。

（二）基於語法和規則的斷詞：在斷詞的同時，一併進行句法、語意分析，利用句法、語意資訊進行詞性標註，以解決「斷詞歧義」之現象。惟因現有中文語法知識、句法規則仍為籠統、複雜，此類斷詞法的精確度還未達實用。

（三）基於統計的斷詞：統計由複數個字形成的字串（character string）在語料庫中出現的頻率，來決定該字串是否構成詞。詞（word）為字（character）的組合，故當特定的一組字在語料庫的眾多文件中相伴且依序出現的次數越多，這組字越可能是一個詞。因此，研究者統計大量文件中特定字串出現的頻率，並在詞典中以頻率表達「此字串是一個詞」的確信程度。

試以現有斷詞系統分析一虛擬專利請求項之結果為例說明，假設該請求項為：「一種視訊編碼設備，依據一自適性串流化標準輸出一視訊至一目的端，並選擇性實施以下輸出步驟：輸出一分層視訊串流形式之該視訊；輸出一提高空間分辨率視訊串流形式之該視訊，該提高空間分辨率視訊串流以封閉式 XYZ 結構編碼，並以一第一空間分辨率及一第一 QoS 值編碼該視訊……」，其中「自適性串流化標準」、「分層視訊串流」、「空間分辨率視訊串流」、「封閉式 XYZ 結構編碼」、「第一空間分辨率」等 TF，皆是複音詞。

[35] 鄭捷，同註 31 ，頁 3-1 至 3-2 ，第 3 章。

以中研院中文剖析器 CKIP Chinese Parser[36] 線上分析上述請求項，得到以下結果：

（一）「自適性串流化標準」被剖析為「theme:PP(Head:P59: 自 |DUMMY:NP (property:VH11: 適性 |Head:Nab: 串流))|Head:VG1: 化 |range:NP (Head:Nac: 標準)」

（二）「分層視訊串流」被剖析為「Head:VD1: 分 |theme:NP(property:Nfd: 層 |property:Nad: 視訊 |property:Nab: 串流」

（三）「空間分辨率視訊串流」被剖析為「goal:NP(property:Nac: 空間 |property:Nab: 分辨率 |property:Nad: 視訊 |Head:Nab: 串流)」

（四）「封閉式 XYZ 結構編碼」被剖析為「DUMMY:NP(property:A: 封閉式 |theme:NP(property:Nba:XYZ|Head:Nad: 結構)|Head:Nv: 編碼)」

（五）「第一空間分辨率」被剖析為「DUMMY1:NP(quantifier:Neu: 一 |property:NP(quantifier:Neu: 第一 |Head:Nac: 空間)|Head:Nab: 分辨率)」

另再以中研院中文斷詞系統[37] 剖析同一請求項，得到以下結果：

（一）「自適性串流化標準」被剖析為「自 (P)　適性 (VH)　串流 (Na)　化 (VG)　標準 (Na)」

（二）「分層視訊串流」被剖析為「分 (Nf)　層 (Na)　視訊 (Na)　串流 (Na)」

（三）「空間分辨率視訊串流」被剖析為「空間 (Na)　分辨率 (Na)　視訊 (Na)　串流 (Na)」

（四）「封閉式 XYZ 結構編碼」被剖析為「封閉式 (A)　XYZ(FW)　結構 (Na)　編碼 (VB)」

（五）「第一空間分辨率」被剖析為「第一 (Neu)　空間 (Na)　分辨率 (Na)」

---

36 中央研究院，中文剖析器線上測試 CKIP Chinese Parser，http://parser.iis.sinica.edu.tw/（最後瀏覽日：2021/10/26）。

37 中央研究院，中文斷詞系統，http://Ckipsvr.iis.sinica.edu.tw/（最後瀏覽日：2021/10/26）。

理想上，若現有中文 NLP 斷詞模組能將中文請求項中的 TF 予以正確剖析，則可將所得 TF 交由機器學習、深度學習等技術在儲存有預先剖析完成之專利資料之基底資料庫中查詢、比對，亦即正確斷詞為以 AI 技術檢索專利引證的第一步。惟在上述測試中，虛擬請求項所載之 TF 用語疑尚無法直接運用現有斷詞模組即獲得令專利工作者滿意的斷詞結果。雖然延續此思路狀似不通，但或許所欠缺的只是最後一哩路。

三、技術特徵（TF）與命名實體（NE）

命名實體識別[38]，又稱為專名識別、類專有名詞識別等，在 NLP 流程中主要用來識別文本中具有特定意義的「實體」，該些實體為可用專有名詞或名稱等標識之具體事物個體，例如人名、地名、組織名、日期、時刻、貨幣（包含單位及數目）、比例數值等[39]，並可依應用之需求自訂所需類別[40]。從語言學分析語句的流程觀之，NER 屬於中文斷詞中「未登錄詞識別」的範疇，而在涉及未登錄詞識別的所有問題中，NE 是漢語詞典之未登錄詞中數量最多、識別難度最大、對斷詞效果影響最大的功課[41]，一個中文 NLP 必須先攻克的問題。

傳統上，人名（包含外文譯名）、地名、組織名的中文 NE 識別是最難也最重要的三類[42]；而在本文關注之領域，應用於中文專利說明書之分析，上述最難識別的 NE 還要增加一類：「說明書、請求項所記載之 TF」；尤其當中文 TF 係源自外文本之 TF、專有名詞時，實務上其譯文能遵循例如「國家教育研究院雙語詞彙」、「學術名詞暨辭書資訊網[43]」等中文辭書所建議之標準寫法來進行翻譯者應屬罕見；譯文之變異增多，將導致專利資料庫內的中文 TF 複音專有名詞蕪

38 維基百科，命名實體識別，https://zh.wikipedia.org/wiki/%E5%91%BD%E5%90%8D%E5%AE%9E%E4%BD%93%E8%AF%86%E5%88%AB（最後瀏覽日：2021/10/26）。

39 劉禹吟，中文文字／語音文件中類專有名詞擷取及其可能應用之初步研究，台灣大學資訊工程學研究所碩士論文，2004 年，https://www.csie.ntu.edu.tw/~r91019/docs/masterthesis/yuyingliu_masterthesis_namedentityextraction.pdf（最後瀏覽日：2021/11/26）。

40 Damien Nouvel & Maud Ehrmann & Sophie Rosset, Appendix 5: Named Entities: Current Definitions, Named Entities for Accessing Information, Wiley Online Library, 08 January 2016, https://onlinelibrary.wiley.com/doi/10.1002/9781119268567.app5 (last visited Nov. 26, 2021).

41 鄭捷，NLP 漢語自然語言處理原理與實踐，頁 241 ，電子工業出版社，2017 年 1 月。

42 同前註。

43 國家教育研究院，雙語詞彙、學術名詞暨辭書資訊網，http://terms.naer.edu.tw/compare/（最後瀏覽日：2021/10/26）。

蔓龐雜，成為使用機器檢索引證之無形、重大障礙。幸運的是，在專利審查之引證檢索任務上，僅須單獨著重於第四類 NE（亦即 TF），而無須考慮人名、地名、組織名等其他三類，因該三類 NE 通常不會出現於申請專利範圍之文本內，或即便存在、卻因其為習知而與檢索引證所須關注之技術內容新穎性、進步性無關。

圖 7 為一傳統大型斷詞系統的架構圖，其特徵為將斷詞模組針對不同領域分別優化，不同領域的文本可以使用各自領域內專門的外部詞典來進行識別。

若欲用於專利分析，本文建議應考慮改良命名實體識別之模組，參考習知系統而進行改良之架構繪示如圖 8：為識別專利文件中最重要的「技術特徵」命名實體，此改良架構中加入「技術特徵名識別」模組，及對應、支援其之「技術特徵詞典或模型」；其中，「技術特徵詞典或模型」應視輸入專利之技術分類（例如 IPC、CPC 等）而切換到不同之領域詞典，這是因為考慮到領域差異的專業細分勢必帶來更高的識別精度、達到更好的斷詞效果。類同於上述思路，HITLtp 3.3 和 Stanford NLP 即將命名實體識別模組自中文斷詞基礎模組中分離出來，作為一個獨立的模組而位於斷詞步驟之後；如此，命名實體識別模組的實作可具有更高的彈性及針對性。

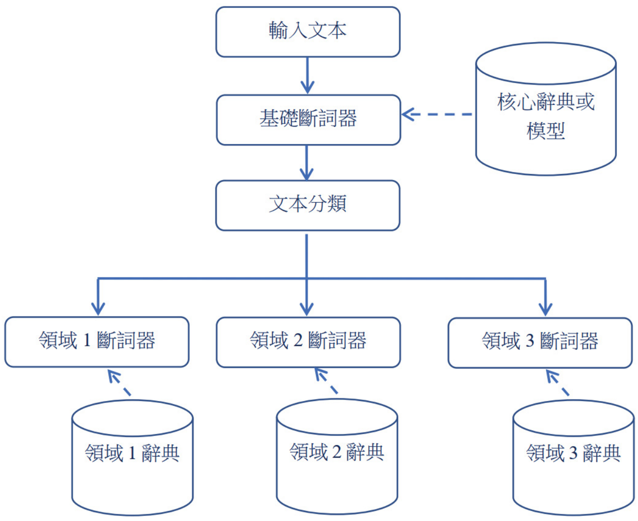

圖 7　傳統大型斷詞系統之架構 [44]

[44] 鄭捷，同註 41 ，頁 243。

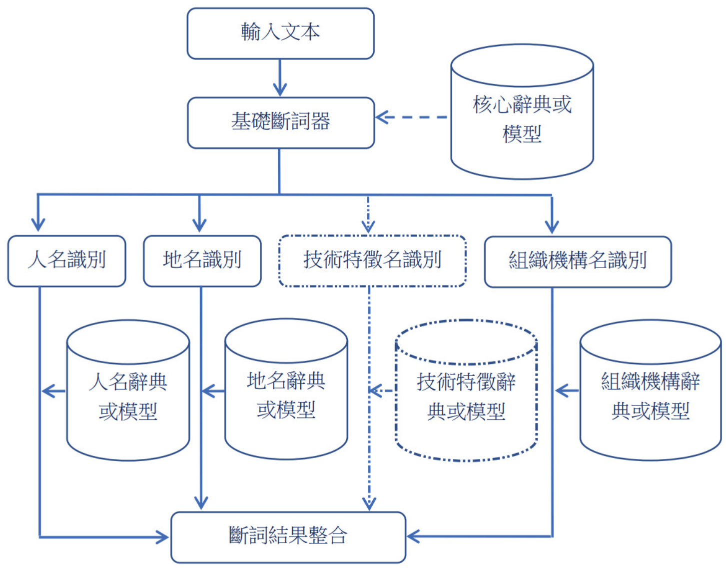

圖 8　傳統中文 NLP 架構[45]，進一步包含本文改良之命名實體識別模組

NER 技術是資訊抽取、資訊檢索、機器翻譯、問答系統等多種中文 NLP 不可或缺的關鍵模組，目前為主流的統計派 NER，實現的機器學習、深度學習技術所採用的數學模型多為隱馬可夫模型（Hidden Markov Model, HMM）、最大熵（Maximum Entropy, ME）、條件隨機場[46]（Conditional Random Fields, CRF）以及功能集（Feature Sets, FS）等。考量以 ML、DL 技術來進行 NER，其優點包括：

（一）可採行無監督式學習，不需要大量語言學專家、標註者等人力介入；惟仍須注意，訓練所得模型之測試結果準確度與訓練所需語料庫、資料清洗[47]（Data Cleansing or Data Cleaning）之完整性等習知變數為高度相依。

（二）亦可採行監督式學習，需要適量專利工作者介入，訓練所得模型之準確度一般而言將高於無監督式學習。

[45] 鄭捷，同註 41。

[46] Zhanming Jie & Wei Lu, Dependency-Guided LSTM-CRF for Named Entity Recognition, EMNLP 2019, 3862-3872, https://aclanthology.org/D19-1399.pdf (last visited Dec. 17, 2021).

[47] Wikipedia, Data cleansing, 1, https://en.wikipedia.org/wiki/Data_cleansing (last visited Oct. 26, 2021).

而其缺點則包括：

（一）模型訓練之語料庫，所包含文本之數量必須足夠；就專利分析之應用而言，亦即 IPC/CPC 分類號三階以降之申請案件數須足以支撐模型訓練所需。訓練文本實際需要的數量，通常係與模型種類、待調整參數之數量呈正相關。

（二）以機器學習、深度學習機制訓練所得的模型，無法細部調整或任意微調模型參數，若後續有新的 NE 出現，只能不斷「重新訓練」以取得全新的模型，而每次訓練都需要時間、人力、電力等之成本投入，隱性的建置、維護成本，在系統規劃時宜須預先考慮。

（三）訓練所得的 NER 模型宜再經熟悉相關 IPC 技術領域之專利工作者確認，才能確保具有實用上的正確性，惟此步驟亦代表須投入一些隱性的人力成本。

四、基礎詞典，同義詞詞典，領域專用詞典，以及綜合應用「同義詞擴展」功能與「斷詞」功能之示例

NE 模組預設有對應之詞典進行支援，例如表 2(a) 所示為 HanLP 的 7 個預設詞典，表 2(b) 例示由表 2(a) 中該一元語言模型詞典截取一片段之文字內容：

表 2(a)　HanLP 系統之詞典結構 [48]

詞典文件	說明
CoreNatureDictionary	一元語言模型詞典
CoreNatureDictionary.ngram	二元語言模型詞典
person/nr. txt.trie.dat person/nr. txt.value.dat	人名識別詞典—中國人名
person/nrf. txt.trie.dat person/nrf. txt.value.dat	人名識別詞典—譯名

（續下頁）

[48] 鄭捷，同註 31 ，表 3.2。

詞典文件	說明
person/nrj. txt.trie.dat person/nrj. txt.value.dat	人名識別詞典—日本人名
place/ns. txt.trie.dat place/ns. txt.value.dat	地名識別詞典
Organization/nt. txt.trie.dat Organization/nt. txt.value.dat	組織機構名詞典

表 2(b)　由一元語言模型詞典截取一片段之文字內容 [49]

（詞）	（第 1 詞性）	（第 1 詞頻）	（第 2 詞性）	（第 2 詞頻）…
跳樑小丑	nz	16		
跳棋	n	1		
跳樓	v	136		
跳樓價	nz	6		
跳樓秀	nz	1		
跳槽	vi	71	vn	55
跳水	vn	137		
跳水教練	nnd	1		

在表 2(a) 中，一元語言模型詞典是演算法的核心詞典，其他各個詞典都從此詞典衍生而來；核心詞典的大小代表了系統的規模，其所包含的詞彙越多，能夠正確斷詞的語料範圍就越大，一般實用的中文斷詞器之核心詞典規模約在 30 萬詞之譜 [50]。表 2(b) 所示片段的第一列是詞，第二列記載該詞的第一詞性，第三列則為第一詞性的詞頻；若存在第四列，則記載該詞的第二詞性，並以第五列記載第二詞性的詞頻……之後依此類推 [51]。

表 3(a)、3(b) 之技術意義分別對應於表 2(a)、2(b) 所示者、但具體內容則為由本文所模擬編寫之例示 TF 詞典及對應之截取片段：

49 鄭捷，同註 31 ，頁 3-19。
50 鄭捷，同前註。
51 鄭捷，同前註。

表 3(a)　本文虛擬之例示技術特徵（TF）詞典

詞典文件	說明
TechnicalFeatureDictionary	*TF 詞典*

表 3(b)　由本文虛擬之例示技術特徵（TF）詞典截取片段之內容

（詞）	（第 1 詞性）	（第 1 詞頻）⋯
分層視訊串流	n	5
自適性串流標準	n	5
自適性串流化標準	n	5
自適應性串流標準	n	5
自適應性串流化標準	n	5
空間分辨率	n	5
空間分辨率視訊串流	n	5
封閉式 XYZ 結構編碼	n	4

除上述 TF 詞典外，我們另需備有「同義詞詞典」，以實現自動提供「依據同義詞而自動擴展之檢索式」（下稱「擴展檢索式」）之軟體功能。「同義詞詞典」將同義詞按照語意相近程度進行不同層次的「聚類」，而自動擴展功能不僅可以自動、快速提供準確之「擴展檢索式」，同時也可以作為未來實現在不同技術領域中分別進行模糊檢索（Fuzzy Search）[52、53] 之功能的預先準備。

表 4 所示為年代久遠之 1983 年哈工大所編纂同義詞詞林擴展版 [54] 詞典檔案格式之範本：

[52] Chen, SM & Horng, YJ, Fuzzy query processing for document retrieval based on extended fuzzy concept networks, 國立交通大學機構典藏 , 1, https://ir.nctu.edu.tw/handle/11536/31544 (last visited Oct. 26, 2021).

[53] 林信成，模糊檢索 fuzzy search，圖書館學與資訊科學大詞典，https://terms.naer.edu.tw/detail/1679031/（最後瀏覽日：2021/10/26）。

[54] ITREAD01.com，同義替換：哈工大同義詞詞林擴充套件版，https://www.itread01.com/content/1548781934.html（最後瀏覽日：2021/10/26）。

表 4　哈工大同義詞詞林擴展版詞典檔案格式

AA01A01=	人　士　人物　人士　人氏　人選
AA01A02=	人類　生人　全人類
AA01A03=	人手　人員　人口　人丁　口　食指
AA01A04=	勞力　勞動力　工作者
AA01A05=	匹夫　個人

對應於表 4，表 5 為本文所模擬編寫之例示 IPC 三階 H04N 同義詞詞典，以「巨塊」及「儲存器」兩技術用語說明同義詞擴展過程。

表 5　本文虛擬之 IPC 三階 H04N 同義詞詞典示意圖

AA01A04=	巨塊　MacroBlock　Macro　Block　MB　宏塊　大型區塊　巨集區塊　巨型區塊　大塊　巨块　マクロブロック　매크로 블록
AA01A05=	儲存器　儲存單元　Storage　類比記憶體　Analog　Memory　Cell　電容　二極體　Diode

根據表 5 之示意範例，當專利工作者意欲檢索於摘要中包含「巨塊」及「類比記憶體」之專利，所需輸入的檢索式僅為：

【（巨塊 And 類比記憶體）@AB】，

理想之自動擴展功能即可自動將所輸入之上述原始簡單檢索式擴展為所需之「擴展檢索式」，其內容例如為：

【（巨塊 or MacroBlock or Macro Block or MB or 宏塊 or 大型區塊 or 巨集區塊 or 巨型區塊 or 大塊 or 巨块 or マクロブロック or 매크로 블록）@AB And（類比記憶體 or 儲存單元 or Storage or Analog Memory Cell or 電容 or 二極體 or Diode）@AB】，

故此類數位工具能大幅降低專利工作者輸入同義詞之負擔，同時確保同義詞擴展之穩定性及完整性，因而能提升檢索文獻之效率。

為說明「建構符合特定應用環境／技術領域之專用詞典」的必要性，我們以圖 9 所示、國家教育研究院「雙語詞彙、學術名詞暨辭書資訊網」之詞彙查詢結果為例。當查詢屬於 IPC H04N 技術領域之基礎技術用語，「巨塊」，並將查詢結果由屬於此技術領域的審查人員判定關聯性，判斷結論是，所得結果與技術領域實為低度相關。

圖 9　國家教育研究院之雙語詞彙、學術名詞暨辭書資訊網之詞彙查詢網頁「巨塊」查詢結果 [55]

承上，改以「MacroBlock」搜尋，則僅得到「巨集區塊」、「區方塊」兩筆資料，亦即一般性的通用詞典所能提供的資訊明顯與特定技術領域審查人員所需求之 TF 用語有所差異。

另一方面，若再嘗試以 H04N 領域中較為新穎之標準詞彙查詢，檢索「H.266」，得到 2 筆資料，如圖 10 所示，惟該等詞彙均異於其所屬領域之通常知識者所常用之「多功能影像編碼」[56]。

55　查詢結果檢索策略：巨塊結果共 2 筆，國家教育研究院雙語詞彙、學術名詞暨辭書資訊網，https://terms.naer.edu.tw/search/?q=%E5%B7%A8%E5%A1%8A&field=ti&op=AND&group=&num=10（最後瀏覽日：2021/10/26）。

56　維基百科，多功能影像編碼，https://zh.wikipedia.org/wiki/%E5%A4%9A%E5%8A%9F%E8%83%BD%E8%A7%86%E9%A2%91%E7%BC%96%E7%A0%81（最後瀏覽日：2021/10/26）。

圖 10　國家教育研究院之雙語詞彙、學術名詞暨辭書資訊網之詞彙網頁「H.266」查詢結果 [57]

依據上述查詢結果簡單小結如下：（一）國家教育研究院等機構提供之通用雙語詞彙為詞彙使用之建議，惟中文專利說明書未必參酌該等詞典來撰寫；（二）中文複音同義詞之詞彙量明顯不足，簡體之 NE 複音詞從缺，疑應不足以支持模糊檢索之需求；（三）詞典內容更新、導入外部詞典等，常須投入顯著之人力資源，在進行相關規劃時宜預先考量此一成本。由上述觀察亦可知，建構符合特定應用環境、技術領域之專用詞典，例如以 IPC 之適當階數來區隔分別之各技術領域，讓領域中之專家或專利工作者可以校調、優化該專用詞典所包含的每個詞條，應仍為必要。

若專利工作者願意投入心力，以專業進行檢視而協助建構同義詞詞典，則可以在短期內達成兩個目標：

（一）取得專利說明書之正確斷詞，作為發展後續中文 NLP 功能模組（例如語意分析）之基礎。

（二）實現依據同義詞自動擴展檢索式之功能，充分發揮檢索資料庫所收錄資料之價值，並減輕專利工作者建立檢索式之心力負擔。

57　查詢結果檢索策略：H.266 結果共 2 筆，國家教育研究院雙語詞彙、學術名詞暨辭書資訊網，https://terms.naer.edu.tw/search/?q=H.266&field=ti&op=AND&group=&num=10（最後瀏覽日：2021/10/26）。

另一方面，亦可將適用於特定專業領域之同義詞定義為包含不同語言之同一技術名詞。以「DNA」為例，其同義詞可定義為包括「去氧核醣核酸」、「ディーエヌエー」、「디엔에이」、「Deoxyribonucleic acid」……等；以「瘦肉精」[58]為例，則可定義為包括「乙型交感神經受體致效劑」、「Beta-adrenergic agonist」、「萊克多巴胺」、「Ractopamine」……等；當請求項經斷詞分析而得到兩個技術特徵「DNA」與「瘦肉精」，理想上可自動建構出擴展檢索式如下：

【（DNA or 去氧核醣核酸 or ディーエヌエー or 디엔에이 or Deoxyribonucleic acid or…）@de AND

（瘦肉精 or 乙型交感神經受體致效劑 or Beta-adrenergic agonist or 萊克多巴胺 or Ractopamine or…）@de】，

這串繁雜的擴展檢索式包含多種語言之查詢詞彙，故可直接用於同時對世界各專利局之專利資料庫進行檢索，得出所有包含「DNA」、「瘦肉精」及其同義詞之相關專利。初步檢視檢索結果後，專利工作者根據專業並透過適當使用者介面快速增刪關鍵詞、同義詞，遞迴數次即可快速逼近理想或符合其需求的最佳檢索式。

肆、結語

中文之自然語言處理技術是值得關注之技術發展主題，它是中文文件電子化、資訊化之後的下一步，讓運用者可進一步讀取中文文件裡包含的文字資料，並將初始的資料提煉、轉化為更具價值的資訊，以供進一步分析、利用；一旦可以有效提取中文文件中的進階資訊，我國現有的所有中文文件資料庫便都成了發展大數據的無價寶庫、重要基石；透過大數據技術的分析能力，研究者或許將能得到更多現在還完全想像不到的資訊。

---

[58] 維基百科，瘦肉精，https://zh.wikipedia.org/wiki/%E7%98%A6%E8%82%89%E7%B2%BE（最後瀏覽日：2021/10/26）。

然而，發展中文之自然語言處理技術並非一蹴可幾之事，甚至也可能不是投入短期努力便能輕易獲得豐碩成果。如此，則應如何保持足夠力量來持續推動技術發展？觀察國外的發展經驗，即便是電腦科學、資訊技術最為進步的美國，在20世紀50年代時期與電腦、人工智慧領域誕生而同時開始的自然語言處理技術，初始發展的動機部分是著眼於其在冷戰時代的戰略意義，但最終也遭遇成果效益不彰、資金退潮的困境。如今，世界處於第四次工業革命的進程，且商貿往來興盛，美國的巨擘級企業，包括Google（搜尋引擎，手機作業系統），蘋果（手機、電腦，作業系統），微軟（個人電腦，作業系統），IBM（軟體服務），Facebook（社群媒體），Amazon（線上購物）等，重新從大眾的角度出發，重視、運用並在其產品、服務中強調人工智慧、數據分析及文字資訊分析等技術之能力；在中國大陸，則以包括百度（搜尋引擎）在內之各種商業應用，使（簡體）中文自然語言處理之技術具有實用價值，鼓勵、吸引了更多資金、人才的熱忱投入，自然形成引領技術發展之趨勢；這種透過商業、應用之需求而選擇、推動技術發展的模式，讓需求選擇技術而令其進步，技術進步再帶來更多、更新、更吸引使用者目光的應用需求，如此構成良性循環，技術可以自我匯集的力量加速進化，或許是較好、較自然且有效率的模式。

抱持著以AI協助專利工作者檢索中文專利之相關前案的理想，與IP5推動數位工具之運用的潮流似乎方向一致，惟凡事從無到有，跨出的第一步會是最辛苦困惑、最困難的。中文斷詞方法發展至今已有一定的成效，在文本上可以達到有效斷詞之目標；若能參考本文之粗略示範，深入探索技術細節並依實際應用情境細心改良實作步驟，則有機會實現以斷詞技術預先處理專利文件、提供TF標示清晰之說明書，並讓專利工作者能取得自動依據同義詞詞典擴展之檢索式，且有回饋意見（例如建議NE詞典增、刪TF等）之友善操作介面；透過引入這些數位工具之機器力量，可以幫助個別的專利工作者獲得眾多其他專業同行所積累之知識及經驗，對於團隊之整體而言，應亦有順應科技發展潮流及促進團隊內部分享、傳承之實質意義。