統一醫學語言系統簡介

張慧銖

前言

1986年初，美國國家醫學圖書館 (National Library of Medicine，簡稱NLM) 開始著手一個長期的研發計劃，即統一醫學語言系統 (Unified Medical Language System，簡稱UMLS)。此一計劃的誕生，乃是基於下列假設：『及時檢索正確而新穎的資訊，將有助於決策的制定，並能提昇研究及病人照護品質。』。由於生物醫學資訊不斷大量的增加且分散於各個資料庫系統，因此想要查檢完整而新穎的資訊殊為不易，是故UMLS應運而生，其目的在提昇系統之能力，使系統能了解讀者在生物醫學方面的問題意涵，並進而幫助讀者檢索及整合相關資訊。UMLS著重在解決二個影響有效檢索的問題：①同一概念經由不同的人或在不同的資料庫中可能會有不同的表達方式；②資料庫系統分散所造成的檢索不完整的問題。換言之，UMLS試圖在讀者問題與資料來源間建立概念上的連結。

UMLS計劃的目的並不在建立一索引詞彙，或建置一大型知識庫以涵蓋所有的生物醫學文獻資料，而是針對病歷紀錄定義其電子形式資料應具備之架構及內容。也就是說，在UMLS計劃中，希望可以創造出一個界面環境，而這個界面必須擔負起三項主要功能：①具備詮釋讀者問題的能力；②具備判斷並連結相關資料庫的人工智慧；③執行有效檢索的能力。

為達到前述三項功能，UMLS系統中設計了泛索引典 (Metathesaurus) 和語意網路 (Semantic Network) 二個知識庫來達成系統在檢索問題 (query) 與各資料庫系統中所存有之大量生物醫學文獻間建立概念上的關聯；而資訊來源圖 (Information Source Map) 的建置，可以幫助讀者判斷並建議合於讀者檢索需求之資料庫；至於專家語典(SPECIALIST Lexicon)則主要應用於提供各項語詞資料，以幫助系統處理自然語言所產生的語法差異的問題。

UMLS的主要架構係由四個知識庫所構成，即泛索引典、語意網路、資訊來源圖和專家語典，本文簡單介紹其設計的基本理念與功能，盼能讓讀者有一概括的認識。

一、泛索引典（METATHESAURUS）

資訊檢索結果的成功與否，取決於讀者對其所使用資料庫架構之了解程度，而要使電腦與使用者間可以作互動式的交談，關鍵因素即是『語言』。而Metathesaurus即為UMLS系統中負責掌控詞彙的工具。所謂"Meta"意即超越、涵蓋。顧名思義Metathesaurus（以下簡稱Meta）可視為一個概念名詞的知識庫，是由生物醫學領域中許多不同索引典或分類表中所抽取出來的辭目所組成。換言之，Meta的範圍決定於其涵蓋的索引典數量。以目前1996年版來說，Meta中約含有589, 000的個語詞 (concept name) 以表達253, 000個概念 (concepts)，而其來源則分別取自30個生物醫學索引典。在Meta中，保留了個別索引典中

對概念的定義、階層的連結及各項語意關係。並為在Meta中的每一個概念建立新的關聯，用以串接不同索引典中各概念與詞彙間的關係，整合不同來源的索引典為一個龐大的、業經控制的概念知識庫。

除了各索引典間的整合外，Meta也試圖跨越因語文的不同所造成概念的表達形式不一的問題。在1996年版的Meta中已包括MeSH的法、德、西、葡文譯本。而來自其他索引典之轉譯辭目也將在未來Meta的後續版本中被逐步納入。

UMLS用以處理自然語言之程序，是把檢索者的檢索用語先和Meta中所載有的詞彙作對照，找出系統用語，再依其所對應的語意型態，將檢索者輸入之檢索用語間建立合理的關係，使系統可據以判讀使用者檢索諮詢，進而作資料庫選取之建議。

二、語意網路（SEMANTIC NETWORK）

語意網路的產生是為了讓Meta中所涵蓋的概念 (concepts) 能有一致的分類體系，並為所有的概念建立關係，使能根據每一個概念所屬之語意型態在網路中所在的位置來檢視其與其他概念間的關係。其做法是為每一個在Meta 中出現過的概念賦與其所歸屬的語意型態 (semantic type)，並且定義每個語意型態間產生連結的關係為何。藉由語意型態的賦與及關係的建立，提供一致的觀點來檢視Meta中所涵蓋之所有概念，使電腦可以『理解』文獻內涵及整個生物醫學知識領域之架構，並進而輔助使用者之資訊尋求行為。也就是說語意網路是一個將生物醫學領域中各物件（object）加以分類的一個架構，所以其範圍大於Meta中所包含的任何單一索引典，且整個架構尚隨著生物醫學領域的擴展而不斷向外延伸。

三、資訊來源圖（Information Sources Map）

大量而快速成長的生物醫學資訊以及資料型態的多元化，使得任何個人想要完整蒐集單一主題的生物醫學資訊，十分困難。ISM設計的目的即在幫助生物醫學領域的讀者，經由系統功能之支援，輔助其在資料內容上做選擇判斷，並透過網路快速獲得相關資訊，以解決其問題。

ISM的初步構想是由使用者輸入其有興趣的主題，交由系統自動作判斷並列出可能與讀者需求相符之資料庫清單。根據此一建議清單，使用者可視其需要:①取得各相關資料庫之介紹資料；②將清單上所建議的資料庫來源，依讀者個別需求加以重整；③自動連結至某一資料庫。

ISM System主要涵蓋的部份有二，即伺服器與使用端，其中ISMS Server主要功能在 (1) 存放知識庫 (ISM Knowledge Base)。(2) 資料庫選取程式 (Source Selection Logic)，此一軟體程式在於執行讀者檢索策略，以便與知識庫中涵蓋之各個資料庫作對照；而ISMS Client之功能則為網路通訊。

ISM System的主要目的乃在幫助使用者選擇適合其研究主題的資料庫，故除了上述所言，可依使用者個別需求（如：資料相關程度、資料型態、偏重主題、檢索途徑、文獻適用類型。）來重整輸出結果外，讀者也可選擇某特定之資料庫，以便獲得有關該資料庫更進一步

的資訊如：涵蓋範圍、收錄標準等，或選擇自動連結至該資料庫檢索論文、官書資料、專利及各種媒體資料等文獻。

四、專家詞語錄（SPECIALIST Lexicon）

專家詞語錄主要是在提供各項語詞資料，以便系統可以據以處理自然語言所產生之語法上不確定的問題。我們可將SPECIALIST Lexicon視為是一套大部頭的電子辭典，而其範圍則涵蓋一般常用英文單字及生物醫學辭彙。

在SPECIALIST Lexicon中提供了每個字詞語彙上的各項訊息，包括1.定義在其語法結構上的類屬 (category) 2.其衍生的字詞尾變化 (如名詞的單複數形、動詞的語形變化、形容詞或副詞的原級、比較形、最高級等) 3.允許的修飾補語 (complementation patterns) 。語言學上有關語法的分類可分為11類：動詞 (verbs) 、名詞 (nouns) 、形容詞 (adjectives) 、副詞 (adverbs) 、助詞 (auxiliaries) 、語態 (modals) 、代名詞 (pronouns) 、先行詞 (prepositions) 、連接詞 (conjunctions) 、補語 (complementizers) 及限定詞 (determiners) 。而語言的基本句型 (basic sentence patterns) 決定於其動詞引導的補語之數目及特質，也就是說主要動詞的補語一旦確定，則整個句子的架構就差不多被定型。

五、知識源伺服器（Knowledge Source Server）

UMLS的目的乃在發展一分散式知識庫集，以便可以移植在各種不同的應用軟體上，來補足不同生物醫學資料庫系統中，以不同方式表達相同概念的缺點。而UMLS Knowledge Source Server為一發展中的工具，目的在提供透過Internet擷取存於UMLS Knowledge Source中各項資料的管道，以方便使用者，特別是系統發展師，可以便利地擷取UMLS系統之資料。

UMLS系統架構是採Client-Server方式處理。由client端依TCP/IP通訊標準，傳遞需求指令 (requests) 到NLM集中管理之Server。而連線管道可經由指令模式 (command-line interface) 、應用模式 (Application Programming Interface, API) 及透過World Wide Web等三種方式。Knowledge Source Server的優點在於便利的提供系統發展師從遠端依其所需擷取UMLS資料，更重要的是使系統管理師不需額外投入時間、精力去了解資料檔案的架構及其他細節，就可以將UMLS Knowledge Source直接使用於應用軟體上。

結語

從UMLS的設計理念中我們可以看出，它結合了許多領域中的學科專家共同努力，至少包括醫學界、圖書資訊學、語言學及電腦界的人士。可見一個以使用者為導向的系統非要結合不同領域的知識不能竟其功。我們經常感嘆學術網路 (Internet) 上有許多他國建置好的資料庫供全世界的人取用，而我們卻很汗顏並沒有著名的中文資料庫提供給大家。若是國內可以參考UMLS所建置的知識庫，在不同的領域中整理出他們的研究成果與該領域的辭彙並提供利用，應該是大家所樂觀其成的。

由於UMLS系統仍持續在研發中，我們期待有更多的應用實例能發表文獻，相信從這些文獻及實例應可給我們更具體的概念，也才能對該系統如何應用於國內的可行性有更清楚的認識。

UMLS中揭示了一種可以使檢索者更自由、更精確的表達資訊需求的索引法。此為業經控制的自然語言索引方式。但對中文資料庫而言，其整理語句架構的方式並不適用。由於中文在表達形式上與拼字語言（如英文）有很大的差異，中文並沒有所謂的單複數的字尾變化，詞性的分別也不明顯且敘述時並沒有所謂的文法來限定表達形式。換言之，以UMLS的建置經驗並不能提供我們作為建置中文資料庫時的方針，惟其規劃系統的程序法應該可以提供給我們一個思考的空間。

親愛的讀者：

感謝您這一年來對醫圖的支持與鼓勵，在這歲末時節謹獻上我們小小的祝福，請您於往後的日子一本初衷繼續給予醫圖關懷與指教，謝謝。

新年如意！

張慧銖暨全體同仁　敬賀