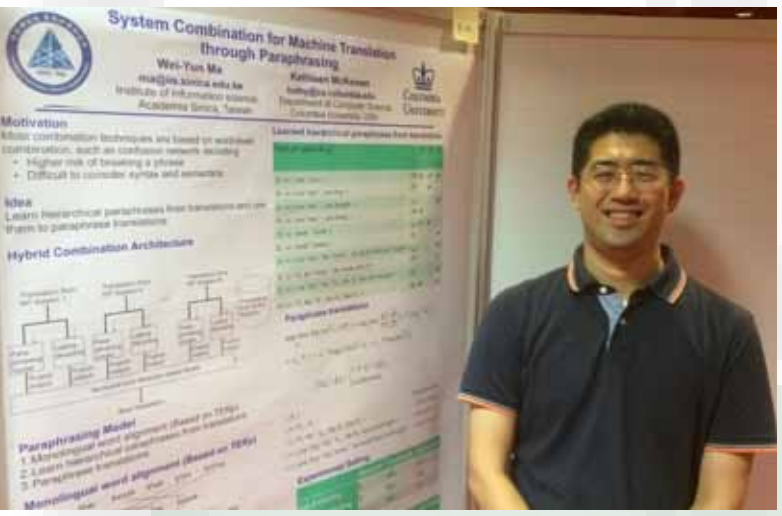
語言分析技術 — 開發者新世代

助研究員　馬偉雲



於 EMNLP 會議發表研究成果。

何因緣進入中研院，並選擇本研究單位

其實早在 2001 年我已經以研究助理的身份加入資訊所，從事自然語言處理的相關研究，而在赴美留學多年後，才又以助研究員的角色回到資訊所的大家庭。記得在研究助理期間，常常藉由所上的演講與交流，感受到大家對於研究的興奮與熱情，那時就埋下日後以做研究為職志的心願。我赴美留學的研究主題仍然是自然語言處理，畢業時正好資訊所徵求自然語言處理的新血，對我來說這是個難得的機會，能主持一個自己的實驗室，並將自己的研究想法付諸實現，是研究人員最興奮的事。而中研院資訊所一直是國內外聲名卓著的研究機構，其研究環境更是提供了許多利於研究的條件，包括：(1) 充分的自由選擇研究題目 (2) 資訊所的研究人員有 40 位之多，分成八大研究群，因此在鑽研自己研究的同時，也有充分的機會跟同研究群或不同研究群的成員合作與切磋。(3) 中研院資訊所對於軟硬體，研究經費，行政等等的支援 (4) 提供許多國內外的重量級教授或研究人員的交流機會。(5) 資訊所並沒有教學上的要求，所以可以專注於研究，但資訊所也同時提供了國際學程的教學機會，讓研究人員也可以有培育英才的機會。

人生歷程

大一開始接觸程式設計，大三在工研院實習一年，為 8051 晶片開發一套語音指令辨識系統，算是研究的啟蒙階段，之後繼續從事連續語音辨識的研究，由於需要使用到語言模型 (Language Model)，那時才很驚訝的發現，原來資訊科學當中還有以“語言”為研究對象的一門學問，當時可以用大開眼界來形容！碩士畢業後我進入中研院資訊所詞庫小組服國防役，做了一系列中文的自然語言處理研究，也觸發了我在這個領域繼續深造的想法。於是在國防役結束後，便赴美國哥倫比亞大學跟隨 Kathleen Mackeown 做自然語言處理的研究。親炙了許多大師的風采，除了在專業知識上的獲得，也學習他們做學問的態度與方法。求學期間也有幸加入幾個美國當時的校際間的大型合作計畫，如 DARPA 所主導的 multilingual QA 計畫以及 NSF 的機器翻譯計畫，觀察到大型研究計畫的規劃與執行層面，同時也有許多機會跟各校的師長同學交流學習。在畢業的前一年有機會赴美國微軟研究院做暑期實習，看到大型企業對研發的投入與重視，更重要地，是深刻感受到自然語言處理早已全面走出象牙塔，各類相關技術為各大企業所利用，落實到其產品與服務。

研究的緣起及目標

人類用語言紀錄知識，彼此溝通。而自然語言處理 (NLP) 就是賦予電腦處理或者理解人類語言的能力的一門學問。人類語言對電腦來說其實相當複雜，不論是從語法或是語義的層面都常常具有高度歧異的現象，舉例來說，“我考試得了鴨蛋”，這裡的“鴨蛋”是“零分”的意思，而不是“鴨”的“蛋”。當電腦能夠正確的辨認出語法或是語義的歧異，其實也就表示某種程度上電腦已經理解了語言。

我們在過去三十多年的發展中，針對不同層次的解歧任務，開發了詞彙分析系統、 斷詞系統與中文語法剖析器等等，同時為了建構知識系統，我們開發了標記語料庫、句結構樹資料庫、詞彙知識庫等基礎建設。這些訓練語料由人工標記產生，數量僅能達到一定的規模。相比於網路上大量未標記的各類文字訊息，數量相當有限。因此如何利用網路上的大量未標記語料來加以訓練是近來 NLP 界的重要目標，而手段就是