Link Grammar 為基礎之多媒體資料庫設計

莊淇銘
Department of Computer Science and Information Engineering
Cmchung@cs.tku.edu.tw

王元凱
Department of Electronic Engineering, Catholic Fu Jen University
ykwang@mercury.ee.fju.edu.tw

王英宏
Department of Computer Science and Information Engineering
inhon@mail.tku.edu.tw

林志豪
Department of Computer Sci nce and Information Engineering
g7190172@tkgis.tku.edu.tw

摘要

本論文的目的在研究多媒體資料庫技術，並嘗試建立一個英語學習環境的多媒體語料庫 (multimedia corpus)，同時並研究利用及"語意查詢" (Semantic Querying) 的技術，及鍊結文法 (Link Grammar) 的加註方式，將傳統語法層次的查詢方式提升到語意 (Semanti) 層次，以提升過去使用關鍵字查詢 (Keyword Querying) 或內容式查詢(Content-based Querying)所無法達成的成果 本論文將存放在本系統所建構之的語料庫中的英文語句，利用論文之中所提出的方法，即可達成英語語句的查詢，並且可以將英語語句的關鍵字查詢的方式提升至英語語句的語意查詢層次。

關鍵詞：multimedia database, corpus, link grammar, keyword querying, semantic querying

1. 簡介

在以往的資料庫當中，大多數都是以儲存文字與數字為主的關連式資料庫，其主要功能將一些基本的文字與數字資料，如一些學生、員工資料，公司的財務資料等。不過隨著資訊時代的來臨，各種數位化的媒體的產生(如圖片、聲音、影像以及超媒體文件等)，為了因應這一些多媒體以及增加我們資料庫內容的豐富性，所以我們所儲存的資料已經不再是只有文字與數字的資料了，而是將多媒體資料儲存到資料庫中，此種資料庫稱之為多媒體資料庫，而多媒體資料庫發展至今已經有許多的議題被廣泛的討論，如 content-base retrieval [4,10]等以及一些所衍生出來的議題，如 shape detection 、object Recognition [5,7]等。藉由以上的一些技術與議題的討論之後，漸漸的開始有些研究已經將多媒體的技術運用到其他領域當中，如遠距教學、數位電視、遠距醫療等都是，本論文是將多媒體與多媒體資料庫的技術運用在英語的學習環境當中，同時提供學生、教師及研究者在電腦網路上的學習環境，稱之為語料庫，學生或是使用者可以從上面獲得一些例句、句型範例、電影對話片段、課文朗誦等資訊。本論文將討論一個由多媒體所建構的語料庫，將一般的散文、生活對話、電影的旁白等生活化、多元化的多媒體資料儲存在此系統當中，提供使用者的一般查詢，並且利用其已將文章資料全文和單句的文字皆已儲存在資料庫當中的特性，運用Link grammar的方式將語料庫當中的每一個句子剖析處理過後，並在每一句子當中加入特定的語意標記，再利用程式寫作的方式，即可將語料庫的查詢提升到語意查詢的層次。

本論文先以英文語句作分析，可以將英語句型分為九大時態，經由鍊結文法的剖析過後，可以得知九大時態都各有各的特殊標記，而本系統就是藉助這些特殊的標記，來使本系統可以很輕易的將使用者所需要的句型找出來。利用這種方式，本系統將英文句型推廣至其他幾種，如：否定句、疑問句、WH 問句等。將在下面幾節中會有比較詳細的介紹。

在以往的語料庫當中，主要都是儲存單字或者是單句的資料，經由一些查詢語言的方式，將使用者所需要的資料給找出來，不過大多數的語料庫系統，都還停留在關鍵字(keyword base)的查詢上，並沒有將資料的查詢提升到內容(content base)查詢，甚至是語意 (Semantic) 上的查詢，而本論文主要的就是要探討如何將一個語料庫系統，利用語料庫其特有的性質，將提升到語意上的查詢方式。

在有關英語語料庫方面本論文也先後參考了幾個相關的網站 [11.12,13,15]，然後在有關於英語句子的剖析方面除了先前所提到的鍊結文法之外，尚還有的是 Schulenburg, D.所提到的 realistic feedbac 的方法 [8]，Brill, E. 所研究的 Machine learning 和 automati linguistic analysis [2]，還有 Chan, S.W.K., 在 Automati linguistic resoluti o 所提到的架構與應用程式 [3]。

本論文在前面已經將研究背景、動機、目的以及研究方法與成果做了一些簡單的介紹，之後將在第二章要開始介紹鍊結文法 (link grammar) 的一些基本觀念，在第三章會對語意標註 (Semantic Annotation) 作一些分析，在第四章將會簡介本系統，第五章為結論。

2. 鍊結文法 (Link grammar)

Link grammar [1, 6, 9, 14] 是由 Carnegie Mellon University CS 所研究發展的一套圖形化的文法剖析系統，我們可以由下圖來說明什麼是 Link grammar。先將一句英文句子輸入，如 "Despite newspaper reports to the contrary , Mary handles herself with extreme confidence" ，而我們可以看到下圖，在每一的單字上都有一些弧線，每一道弧線都有一個標記 (label)，而我們將這些弧線和其之上的標記稱之鍊結 (Links)。而如果用 link grammar 完成合法剖析 (parsing) 過後，就將此句與其上面的鍊結(link)合起來稱為 Linkage。

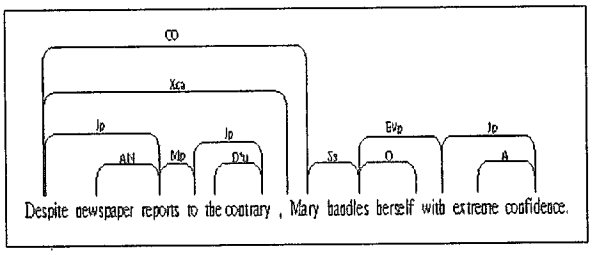


Figure 1：鍊結文法範例

在Figure 1中可以看到confidence用一個鍊結A向左鍊到形容詞extreme，表示一個形容confidence的形容詞，不過也可以說 extreme 用一個鍊結 A 向右鍊結到confidence，每一個單字都可以鍊結到多個字，而每一種不同的標記名稱，都代表不同的意義，就有如handles他就有一個向左表示單數名詞的Ss鍊結，向右一個表示受詞的O鍊結與一個表示連結到一介系詞片語的Evp鍊結。

每一個單字都被一個 disjuncts 來定義，每一個disjuncts都是由兩個有序的列所表示，如下：

$$d=((l_1, l_2, \ldots, l_n)(r_n, r_{n-1}, \ldots, r_1))$$

l_i是左連結可以向左和其他字的右連結相鍊結，r_i 是右連結可以向右和其他字的左連結相鍊結，由以上的例子可以看到，handles 這一個字的disjuncts為（（Ss）（Evp，O））），不過linkage有以下四點規範

連結性：在一個句子當中，每一個單字都必須用 link 連結起來。

平面性：在一個句子當中，所有的link都不能交錯。

順序性：在一個 disjunct 中 $d=((l_1, l_2, \ldots, l_n)(r_n, r_{n-1}, \ldots, r_1))$ l_i和r_i的下標越小所相連的字就應該越近，這是因為避免會有違反平面性的因數。

互斥性：如果同兩個字已經相連過了，就不會有第二個linkage。

所以經過對於鍊結文法瞭解以後，可以發現它的標記都是具有其獨特的意義的所以即可利用他的特性，將句子經過鍊結文法的剖析過後，即可將結果分析過後，及可以做一些處理，此部分將於本論文第二節中做討論。

3. 語意標註（Semantic Annotation）

在開始本系統之前，我們必須先對英語語句與句型作分析，經由我們分析的結果認為，我們可以將英語句型分為以下主要五種：

1. 直述句
2. 否定句
3. 疑問句
4. WH問句
5. 祈使句（命令句）

而有關句型時態可分為有：

簡單式：

現在式：He writes a letter everyday.

過去式：He wrote a letter yesterday.

未來式：He will write a letter tomorrow.

完成式：

現在完成式：He has written a letter.

過去完成式：He had written a letter when I came.

未來完成式：He will have written a letter before I come.

進行式：

現在進行式：He is writing a letter now.

過去完成式：He was writing a letter when I came.

未來完成式：He will be writing a letter when I come.

完成進行式：

現在完成進行式：He has been writing a letter for two hours.

過去完成進行式：He had been writing a letter for two hours.

首先將各種句型先經由Link Grammar剖析之後，我們可以分析出句型有以下的關係，如下表一、二、三、四所表示：

表一：直述句

		主動語態	被動語態
簡單式	現在	Ss or Sp	Ss or Sp + Pv
	過去	Ss or Sp	Ss or Sp + Pv
	未來	Ss or Sp + I	Ss or Sp + Ix + Pv
完成式	現在	Ss or Sp + PP	Ss or Sp + ppf + Pv
	過去	Ss or Sp + PP	Ss or Sp + ppf + Pv
	未來	Ss + If + PP	Ss or Sp + If + ppf + Pv
進行式	現在	Ss or Sp + Pg	Ss or Sp + Pg + Pv
	過去	Ss or Sp + Pg	Ss or Sp + Pg + Pv
	未來	Ss or Sp + Ix + Pg	缺
完成進行式	現在	Ss or Sp + ppf + Pg	
	過去	Ss or Sp + ppf + Pg	

說明：
Ss：為一標記，記錄由單數主詞連結至單數動詞。
Sp：為一標記，記錄由複數主詞連結至複數動詞。
I：為一標記，記錄由動詞原型連結至to等一些介系詞。
If：為一標記，記錄由助動詞連結至to等一些介系詞。
Ix：為一標記，記錄由助動詞連結至be動詞。
PP：為一標記，記錄由have連結至過去分詞。
PV：為一標記，記錄由be動詞連結至過去分詞。
Pg：為一標記，記錄由be動詞連結至現在分詞。
ppf：為一標記，記錄由助動詞連結至be動詞的過去分詞been。

表二：否定句

		原式＋not	Not 縮寫
簡單式	現在	Ss or Sp + N	Ss or Sp + I*d
	過去	Ss or Sp + N	Ss or Sp + I*d
	未來	Ss or Sp + N + I	Ss or Sp + I
完成式	現在	Ss or Sp + N + PP	Ss or Sp + PP
	過去	Ss or Sp + N + PP	Ss or Sp + PP
	未來	Ss or Sp + N + If + PP	Ss or Sp + If + PP
進行式	現在	Ss or Sp + N + Pg	Ss or Sp + Pg
	過去	Ss or Sp + N + Pg	Ss or Sp + Pg
	未來	Ss or Sp + N + Ix + Pg	Ss or Sp + Ix + Pg
完成進行式	現在	Ss or Sp + N + ppf + Pg	Ss or Sp + ppf + Pg
	過去	Ss or Sp + N + ppg +Pg	Ss or Sp + ppf + Pg
說明： 有 N 則一定在句子當中會有 not，則為否定句。 N：為一標記，記錄由動詞連結至 NOT。			

表三：yes/no 疑問句：

		Yes/No 問句
簡單式	現在	Qd + SIs + I*d
	過去	Qd + SIs + I*d
	未來	Qd + SIs + I
完成式	現在	Qd + SIs + PP
	過去	Qd + SIs + PP
	未來	Qd + SIs + If + PP
進行式	現在	Qd + SIs + Pg
	過去	Qd + SIs + Pg
	未來	Qd + SIs + Ix + Pg
完成進行式	現在	Qd + SIs + ppf + Pg
	過去	Qd + SIs + ppf + Pg
說明： Yes/No 問句一定會是 Qd 開頭 Qd：為一標記，記錄由句首連結至 be 動詞。		

表四：WH 問句

		what	Where and how
簡單式	現在	Wq + Sid + I*d + Bsw	Wq + Q + SIs + I*d
	過去	Wq + Sid + I*d + Bsw	Wq + Q + SIs + I*d
	未來	Wq + SIs + I + Bsw	Wq + Q + SIs + I*d
完成式	現在	Wq + SIs + I + Bsw	Wq + Q + SIs + PP
	過去	Wq + SIs + I + Bsw	Wq + Q + SIs + PP
	未來		
進行式	現在	Wq + SIs + Pg + Bsw	Wq + Q + SIs + PP
	過去	Wq + SIs + Pg + Bsw	Wq + Q + SIs + PP
	未來	無	無
完成進行式	現在	無	無
	過去		
說明： WH 問句一定會是 Wq 開頭 Wq：為一標記，記錄由句首連結至 WH 單字。 SId：為一標記，記錄由助動詞 do/doe 連結至主詞。 SIs：為一標記，記錄由助動詞 will、have/has/had 等連結至主詞。 Bsw：為一標記，記錄由助動詞 will、have/has/had 等連結至過去分詞。			

由以上的結果，可以很容易的發現到，利用每一種句型經由 linking grammar 所 parse 過後所得到的不同的結果，就可以很容易的將以上的句型，將一篇散文先分成一句一句的，再經由 link grammar 剖析過每一個句子所產生的不同標記，按照標記所出現的順序，以一集合的方式，如（Ss, ppf, pg），儲存到資料庫當中特定的欄位中，所以每一個句子都已經先經由 parsing 過後才將他們存入資料庫當中，而資料庫現在才算完備。之後，使用者利用 query language 的方式將所需要的句型搜尋出來。由於之前所做的句型分析，為了要確保它的單純性，所以都是用最簡單的句子來分析，所以不會有重複的句型出現在同一句子當中，以得到最純正的分析結果。當完成了上一步驟後，即可接受多種不同的句型出現在同一個句子當中，所以在本系統當中同樣的可以接受多種不同的句型查詢，但由於尚還在實驗的階段，希望以後可以將此方法推廣至在本語料庫當中其他不同的部分。不過還是有其他所需要克服的地方，將在容後再做討論。

4. 系統架構

語文資料庫（簡稱語料庫），為一個資料庫其資料的主要內容為語言相關的資料，如：文章、句子、對話、單字、詞、屬性、詞態等，以提供語言教學及語文的研究，在本論文中將會建構一個屬於多媒體的語料庫，除了一般的文字資料以外，還將多媒體的資料（聲音、教學影片、對話 audio、電影 video 等）放入語料庫當中，以提供學習者更加生動的學習環境，藉以吸引使用者的使用。因為本系統為一多媒體語料庫，所以本系統提供的功能，主要還是以一般英語學習的方法聽、說、讀、寫四個基本的方向來做，主要可以提供使用者在網頁上瀏覽文章，聽取標準的英語發音，寫作並繳交作文，甚至在網頁上練習對話及朗誦課文等。藉由更廣泛的收集

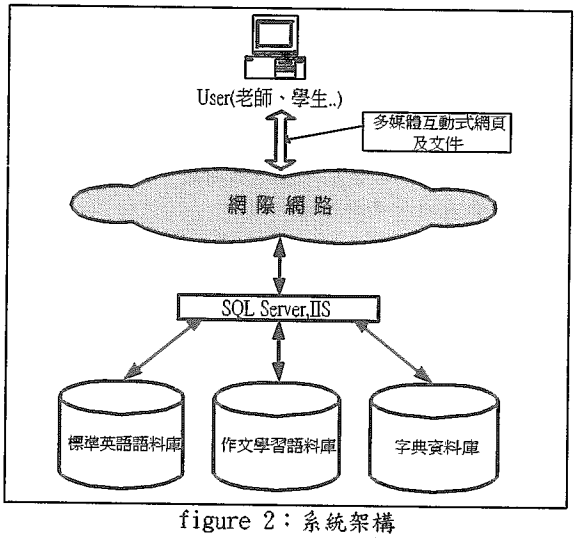


figure 2：系統架構

使用者的學習狀況，本系統進一步的更可提供語文研究方面一些研究方向，以分析台灣學生在學習英語方面的一些問題。而這些資料的收集與分析對語言學研究者而言是非常重要的資訊。

本系統架構在 SQL Server 7.0 其中有三個主要的 databases，有 essay、dialog 及 movies，將分別儲存散文、對話、及電影，如 figure 2 所示。但是本論文先只討論有關散文的部分，以下是此系統的 essay schema：

essay_meta ：

Sentence_no	Essay_no	Class_level	sentence	Chinese

voice	annotation	paragraphic	Link_g

keyword：為本文中的關鍵字
class_level：為本文的用字難易程度
type：為本文的類別
annotation：是用人工加註的方法將註解加入
essa ：將本文全部存入
voice：將聲音檔案也存入，可以有整段的發音
no_para：本文中共有幾段

而 essa 每個 sentence 的 schema 定義如下：

essay_sentence

Essay_no	title	From_book	Creat_date	keyword	Class_type	Type

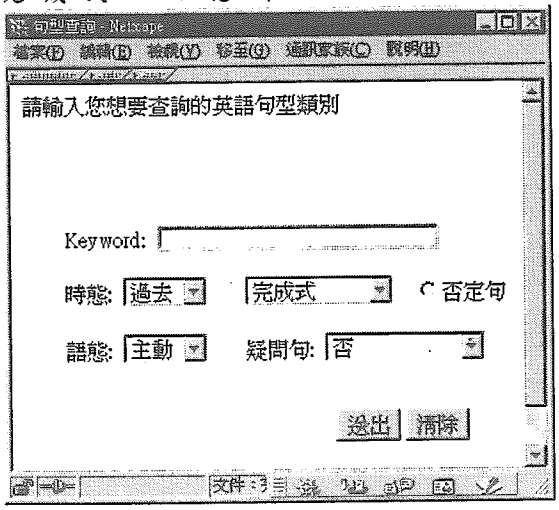
authors	annotation	essa	voice	No_para

class_level：為本文的用字難易程度
sentence：為散文中的單句
chinese：為將單句的中文翻譯
annotation：是用人工加註的方法將註解加入
paregraphic：本句為散文中的第幾段
voice：將聲音檔案也存入，可以有整句的發音
Link_g：為經由 Link grammar 所產生的註解

而本系統即可用關鍵字查詢，主要是要去和 keyword 欄位比對，不過本論文認為，如果要對整篇文章下 keyword 是可行的。就以往而言，我們好像也只能用 keyword 來比對，並不可以將查尋提升到語意的部分。所以本論文最終目的是要討論，如何利用 Link grammar 所產生的一些特性，來將本系統提升到對使用者更有用的查詢方法，本系統即可對英語句型的查詢。

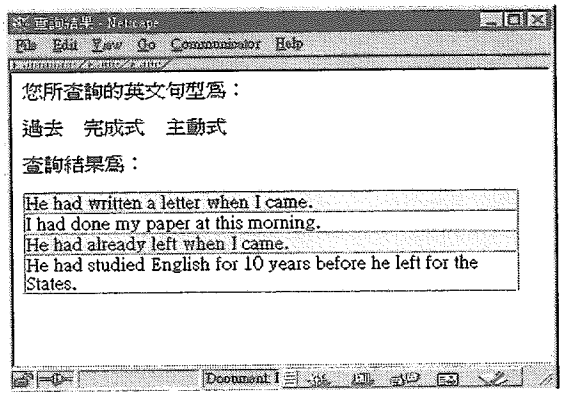
還有一項功能就是，當我們將要收尋的句子找出來以後，我們一樣可以將所要的句子利用 link grammar 來反查剖析（parsing），這樣的話使用者更可以瞭解到此句的結構、句型、語句與字的用法。

以下是本系統的範例，使用者可以利用簡單的下拉式選單，選取需要的英語句型，在選單之間是採用 AND 的方式，可以由下圖三得知就是要尋找是屬於過去完成式、語態為主動的直述句。



圖三：查詢介面

由下圖可以得知，為上一例子的查詢結果。一共有四句。



圖四：查詢結果

5. 結論

經由以上的分析與實驗可得知，在本系統當中可以很容易的利用 link gramma 剖析過後的 label，儲存到語料庫當中，本系統就可以利用這 label，將使用者所希望查詢到的英語句型，本系統可以幫助英語初學者如國中、高中生對於英語語句、句型的學習，更能夠提供一個屬於多媒體互動性高的系統，以幫助同學們的學習，不過，因為 link gramm a 當初在設計的時候並不是要提供此項功能，所以在某些方面來說，並不能夠完全的符合本系統的需求，例如會有不同的句型，經由剖析過後，會產生相同的 label、link grammar 不可以剖析感嘆句型、對於附加問句的剖析會有問題等，所以在將來可能需要將原版的 link grammar 要稍做一些修正，以符合本系統的要求。

本系統除了可以查詢句型之外，也可以提供線上剖析的功能，這在將來對於語言學者，在分析學習者的英文能力，個人習慣用法等將會有所幫助。在本系統當中未來還會提供一項功能，就是修改成一個可以容錯性的 link grammar 可以接受學習者所寫出的錯誤的句子，並且可提供修改的建議，這樣可以提供一個更為完整的英語學習環境。

伴隨著英語語料庫的建立，提供老師一個更好的教學環境和學生一個更好的學習環境，以及師生間的互動關係，而使得資訊技術及人文教育兩方面，無論是學術研究或技術實行上，均能獲得更好的發揮與成就。

參考文獻

[1] D. Beeferman, A. Berger and J. Lafferty, "Tex segmentation using exponential models", Proceedings of the Second Conference On Empirical Methods i NLP, Providence, RI, 1997.

[2] Brill, E., "Machine learning and automatic linguistic analysis: the next step", Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on Volume: 2 , 1998 , Page(s): -1036 vol.2

[3] S.W.K. Chan, " Automatic linguistic resolution framework and applications", Systems, Man and Cybernetics, 1996., IEEE Internation Conference on Volume: 1 , 1996 , Page(s): -630 vol.1 .

[4] Shih-Fu Chang, Chen, W., Meng, H.J., Sundaram, H., Di Zhong Circuits and Systems for Video Technology "A Fully Automated Content-Based Video Search Engine Supporting Spatiotemporal Queries", IEEE Transactions onVolume: 8 5, Sept. 1998, Page(s): 602 –615.

[5] S.-K. Chang; A. Hsu ," Image Information systems: where do we go from here?" Knowledge and Data Engineering, IEEE Transactionson Volume: 4 5 , Oct 1992 , Page(s): –442.

[6] D. Grinberg, J. Lafferty and D. Sleator, "A robust parsing algorithm for link grammars", Proceedings o the Fourth International Workshop on Parsin Technologies, Prague and Karlovy Vary, September 1995, pp.111-125.

[7] F. Myron, S. Harpreet, N. Wayne, A. Jonathan, H. Qian, D. Byron, G. Monika, H. Jim, L. Denis and P. Dragutin, " Query by Image and Video Content: The QBIC System", Computer v28 n9 Sept 1995 IEEE Los Alamitos CA USA p 23-32.

[8] D. Schulenburg, ," Sentence processing with realistic feedback", Neural Networks, 1992. IJCNN., International Joint Conference onVolume: 4 , 1992 , Page(s): 661 -666 vol.4.

[9] D. Sleator, and D. Temperley, "Parsing English with a link grammar", technical report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University, 1991.

[10] Yoshitaka, A.; Ichikawa, T. , " A survey on content-based retrieval for multimedia databases", Knowledge and Data Engineering, IEEE Transactions on Volume: 11 1, Jan.-Feb. 1999, Page(s): 81 –93.

[11] G.R. Sampson: SUSANNE Scheme, http://www.cogs.susx.ac.uk/users/geoffs/RSue.html

[12] Longman Dictionaries Home, http://www.awl-elt.com/dictionaries/

[13] Linguistic Data Consortium LDC, http://www.ldc.upenn.edu/

[14] Link Grammar, http://bobo.link.cs.cmu.edu/link/

[15] Project Gutenberg Official Home Site, http://promo.net/pg/