行政院國家科學委員會專題研究計畫　期中進度報告

在遠距學習環境下使用 Link Grammar Parser 發展之語意問答系統(1/2)

計畫類別：個別型計畫
計畫編號：NSC94-2520-S-032-003-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：淡江大學資訊工程學系

計畫主持人：王英宏

計畫參與人員：王文男,黃朱麒,黃世豪,王建文,李建旻,毛宏仁,林志興

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 5 月 29 日

行政院國家科學委員會補助專題研究計畫 □ 成 果 報 告
■期中進度報告

在遠距學習環境下使用 Link Grammar Parser 發展之語意問答系統(1/2)
Semantic Enhanced QA System Architecture Use Link Grammar Parser for E-learning Environment

計畫類別：■ 個別型計畫　□ 整合型計畫
計畫編號：NSC 94－2520－S－032－003－
執行期間：　94 年　8 月　1 日至　95 年　7 月　31 日

計畫主持人：**王 英 宏**　副教授

共同主持人：
計畫參與人員：

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：淡江大學資訊工程系

中　華　民　國　95　年　5　月　30　日

中文摘要：

自動化的學習輔助及自我學習機制是目前數位化教學環境的一個重要發展目標，近年來，遠距學習之研究議題焦點集中於統一教材及課程編序的描述方式和共同教學環境方面上，因此未來開發數位化教學環境應具備的設計理念：1) 使用Java 語言的跨硬體平台特性，將軟體的教學平台獨立於硬體之上，使得教學平台的移植與互通成本大大降低。2) 以XML(eXtensible Markup Language)定義有交換或散佈需求的課程及課程內容，使課程內容及課程本身在不同教學平台間的轉移更容易且一致。3) 在知識主體研究議題方面，將設定在特定教學範疇下設計適用於遠距學習環境的知識本體結構。

本計畫把注意力放在學習者切要的需求上，即回答與教材本身相關的問題，在遠距教學環境中，學習者由於比較缺乏個體間的接觸，所以遇到課程相關的問題往往得不到即時的解決，需要透過留言版或電子郵件的形式提出問題，並在其他人有意願提供協助的時候才能得到解答，因此能提供一個自動化即時回答教材相關問題的機制對於學習者的學習成就將會相當有益。本計劃設計與實作一在遠距學習環境下使用Link Grammar Parser 發展之語意問答系統（Semantic QA System），做為教學平台輔助學習的工具，以強化學習者與教材之間的連結，其特性有：

- 精確性：不同於搜尋引擎，Semantic QA System 在問與答之間可以更精確的配對。
- 延展性：系統能輕易加入外掛模組調適不同的資料來源，將異質資料格式做同質的呈現。
- 易用性：不需要大量人力介入，即能在教材上組織語意關聯。
- 知識的延伸：超越單一課程內容，反映知識本身的自然連結關係，做為學生自我探索知識的工具。

英文摘要：

On the discourse of distance learning, providing a mechanism for student on automatic learning assistance and self-paced study under the Learning Management System (LMS) has shown its great importance. Recently, research issues focus on the uniformity of standard learning material format and the standard of learning management system. Hence, the next generation learning environment must contain the following characteristics: 1) Using Java programming language: according to its cross-platform nature, the development of LMS inherited high portability and interoperability; 2) Using XML for defining learning materials help facilitating learning content distribution across multiple LMS; 3) System Ontology architecture: a better architecture for organizing and unifying knowledge into system.

This project focuses on the actual needs of learners, i.e., answering the questions relevant to their class materials online. In the distance-learning environment, students usually lack of physical personal contacts than traditional classroom, and questioning via Email often cannot get immediate replies. Thus providing an automatic mechanism (tool) to answer classroom questions online is an urgent challenge and also beneficial to learners under distance-learning environment contexts.

This project design and implement a nature language Semantic QA System using Link Grammar Parser and semantic engine for supporting LMS and acts as its part or standalone application. It contains the following characteristics:

- Accuracy: generate links between questions and its relevant answers.
- Flexibility: use pluggable module to accommodate multiple data sources, and unify heterogeneous data formats into uniform system ontology meta-schema.
- Usability: automate the process of organizing semantic relationship between words.
- Knowledge Extension: extend system knowledge beyond classroom materials from Internet or defined URL address.

keywords：Semantic QA System, Link Grammar, WordNet, Ontology

前言：

在遠距教學環境中，學習者由於比較難與其他學習者互相接觸，所以遇到課程相關的問題往往得不到即時的解決，需要透過留言版或電子郵件的形式提出問題，並在其他人有意願提供協助的時候才能得到解答；因此自動化的輔助學習以及自我線上學習機制是目前數位化教學環境一個重要的發展目標。然而，如何提供學習者一個有效的線上輔助學習機制，則需改善目前線上教學工具技術。

本研究以問答系統為研究標的，在目前著名的搜尋引擎Google、Yahoo等甚至坊間相關的問答系統，大部分是使用關鍵字搜尋配合統計、評等（Ranking）的機制，並且普遍存在有以下的一些缺點：(1)必須使用關鍵字配合每個查詢引擎特殊的進階邏輯符號來查詢，無法使用自然語言來進行符合語意的查詢。(2)回傳太多的查詢結果。(3)其中有大部分的資料，不是我們真正要的資訊；因為關鍵字通常可能會有多種語意，因此使用關鍵字查詢會查出很多沒有用的資訊。(4)真正要的資料不一定找得到；因為在知識庫中，有些資訊是以其他不同的關鍵字（但是代表相同的語意）來描述。

研究目的：

遠距學習的教學環境中有若干種教學輔助工具，我們將首先以改善問答系統並提供語意感知的設計，以實現自動化的輔助學習功能並建立自我線上學習機制，是目前數位化教學環境一個重要的發展目標，而一個在這樣的學習環境中，如何建構具備有語意認知能力問答系統更是一項不可缺少的應用工具。

而在語意感知的研究上，我們則必須研究系統應如何設計才能具備有以下特性：(1)瞭解發問者以自然語言形式所提出的問題。 (2)提高搜尋答案（文獻）的準確率。 (3)建立自動化的學習機制。本研究設計遠距學習平台上的英文問答系統，提出一個語意認知的方法，以期有效回答使用者以自然語言所提出的課程相關問題。

本研究使用Link Grammar文法解析器，結合遠距學習環境下課程的Ontology知識本體以及WordNet來建構一個英文語意認知的問答系統。其中我們使用『資料結構』來做為研究的標的課程，設計了符合我們演算法的Ontology知識本體結構。

在本計畫執行第一年中，本研究首先著重在英文問句的句法結構上，做了一般性的歸

納與剖析，最後我們配合上述課程可能會使用到的問句結構作了部分篩選，並且對其中數種問句類型提出了相應的語意認知方法。

文獻探討：

本研究所提出之語意問答系統（Semantic QA System）係依據下列相關文獻提出：首先利用Link Grammar Parser[1]剖析使用者所輸入之問題，分析學習者提問中的句法後，透過WordNet[2]擴張字彙集（或使用WordNet來擴充知識本體中的關鍵字彙）藉以在系統規劃的知識本體（Ontology）[3, 4]（該知識本體可以用多種機制建立，包含原始資料或是以SCORM[5]標準衍生，甚至是使用網路教材來擴充等機制）找出適合的答案，並提供與問題語意相關的連結做為補充資訊。

系統平台則以Java的Spring Framework[6]為發展的基礎平台，以物件導向分析（OOA）及物件導向設計（OOD）的實際設計方法[7]，並依據物件導向程式設計之設計樣式（Design Patterns）[8]來建構足以適應各式變化的系統，在Link Grammar Parser之連接上，則是引用Java™ Native Interface（JNI）[9] 的規格來做為實作時的共通介面；在關鍵字彙延伸處理上則採用WordNet所提供之JwordNet API進行同義字的字彙擴充及搜尋，以期能找出最符合使用者所關注的問題解答。

本研究參考的語意處理相關文獻研究包括：

概念式自動問答探索系統[10]：其所利用的方法是以潛在語意分析（Latent Semantic Analysis, LSA）為核心技術，並且用統計的方法來計算關係矩陣，利用權重（Ranking）來求取答案，取得答案的資料源可以是任何文件，但是必須使用同一類的文件來進行訓練。

在Speech and Language Processing[11]書中提到語意表示法（Representing Meaning）可以利用文字、符號或圖形表達出語言的意義；使用語意表示法可以幫助自動化的檢核、使用變數，甚至可以進一步的產生語意推理。本書指出語意表示法可分為：First Order Predicate Calculus、Semantic Network、Conceptual Graph和Frame-based表示法。

Parallel Memory-Based Parsing on SNAP[12]：本文章的主軸是在探討 SNAP（Semantic Network Array Processor）上的平行處理方法，知識被建構在概念（Concept Layer）、實體資料（Instance Layer）以及句法（Syntactic Layer）等不同層次，本研究亦依據這些概念來來設計知識本體。

An Intelligent Semantic Agent for e-Learning Message Communication[13]為本實驗室先前所提出之『智慧型語意代理程式遠距教學訊息溝通系統』，其研究目的為提出一個以資料結構為課程之英文聊天室系統，提供學生討論課程以及與老師或其他同學討論，可以讓聊天室內不需要有人隨時在線上監督學習者的表現；此系統採用 Link Grammar 代理程式（Learning Angel Agent）來線上偵測學習者的句法錯誤，而語意相關的功能則是可以檢查學習者在聊天室內的對話語意是否與聊天室課程主題與討論議題相關；系統內學習者與老師之間的問答與對話將會自動儲存於系統知識庫內，當有足夠的問答對話紀錄時，可以轉成 QA pair，即可建構 FAQ 系統提供給學習者使用。

本研究一樣採用 Link Grammar 來判斷使用者輸入問題之句法結構，利用片語輸出的字串，配合本研究方法來進行後續的語意分析處理；在語意分析方面，則是藉由已知句法結構，加以分析以取得問題對應之片語，以及兩兩片語間的關係，並比對知識本體來獲取

適當的答案，回應給使用者；在知識本體的結構上則因為不同的語意處理過程可能會需要不同的知識本體架構，本研究並修改先前所設計之『資料結構』知識本體結構，以做進一步的調整與設計。

研究方法：

本研究目標提出在遠距學習環境下使用 Link Grammar Parser 發展之語意問答系統（Semantic QA System），本研究計畫執行第一年所提出之研究方法將依據『語意感知方法』與『語意問答系統應用程式架構』分別說明之：

『語意感知方法』：此部分之研究與設計方法由本實驗室黃朱麒所發表[14]。本系統在設計上係採用 Link Grammar Parser 剖析使用者的問題，分析學習者提問中的句法後透過 WordNet 擴張字彙集藉以在系統提供的知識本體（Ontology）或在網路教材中找出適合的答案，並提供與問題語意相關的連結做為補充資訊。其中，我們使用『資料結構』來做為研究的標的課程，設計了符合我們演算法的 Ontology 知識本體結構。在英文問句的句法結構上，我們做了一般性的歸納與剖析。最後我們配合上述課程可能會使用到的問句結構作了部分篩選，並且對其中數種問句提出了相應的語意認知演算方法。

本研究計畫語意問答系統所提出系統架構（如下圖 1 所示），在語意感知方法部分所專注的則是如架構圖中方框內的區域。

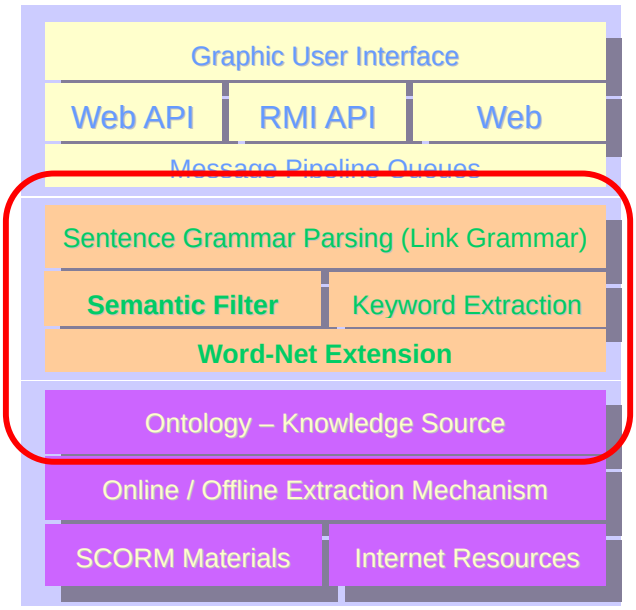

圖 1.語意問答系統架構

此部分研究所設計的方法概述如下：

(1). **片語組萃取：**使用 Link Grammar 的 Phrase Parser 來輸出片語結構，以問句"What is the algorithm of PUSH in the stack?"為例，經過 Link Grammar 的處理後，當 constituents=1 時，其分析結果適合人閱讀（如下圖 2 左）；當 constituents=2 時，分析結果適合電腦處理（如下圖 2 右）：

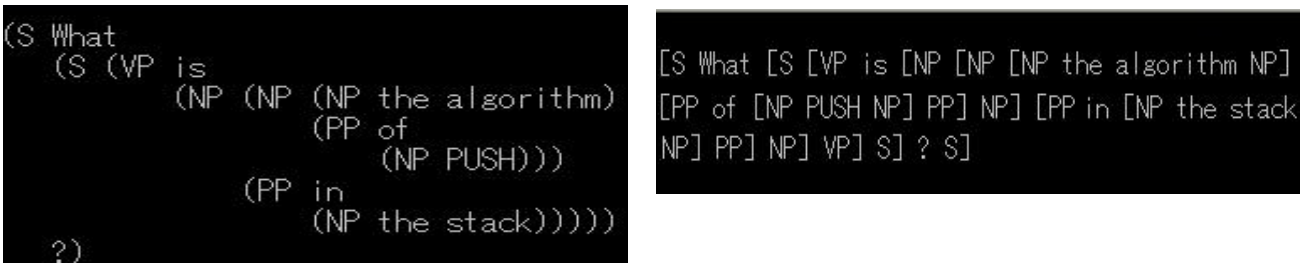

圖 2. 經 Phrase Parse 解析後的字串（constituents 參數設定圖左= 1，圖右=2）

取得如上圖 2 右邊的文法資訊後，使用字串的取代方法後會得到一個 XML 結構的字串，利用 DOM 等 XML 分析物件，我們可以對每個 XML 節點的內容進行資料的讀取，取代範例如下：

將 [S 以 <S>取代、[VP 以 <VP>取代，以下類推...

將 S] 以 </S>取代、VP] 以 </VP>取代，以下類推...

(2). **問句樣版比對：**比對問句句型，找到相應句型，以及標的片語組；將前一步驟中所取得的 XML 物件進行分析，並比對問句樣版表（Pattern Matching Table），如下表 1。

表 1. Pattern Matching Table

QHeading	QN1	QN2	QN3	QN4	QN5	Target	ModuleType
What	is	{NP}				{NP}	1
What	are	{NP}				{NP}	1
How many {NP}	are	there	{PP}			{NP}	2
Which {NP}	is	there	{PP}			{NP}	4
Which {NP}	are	there	{PP}			{NP}	4
{VP}	{NP}	{PP}				{Y/N}	9

(3). **標的 element 萃取：**以片語組中的名詞片語為目標，找出標的 element；經過比對之號，可知道範例問句的句型是 What is 句型，而其標的答案是緊跟在 What is 之後的名詞片語，如下圖 3 所示，其中紅線所圈選的字組，就是標的片語組。

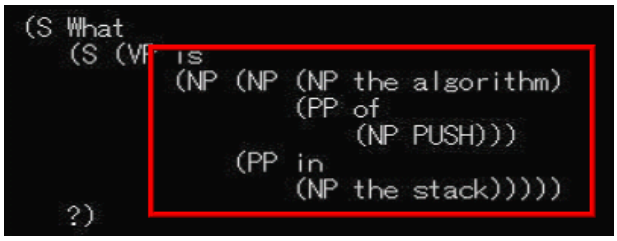

圖 3. 問句樣版比對後的標的(NP)

(4). **語意樹建構：**以標的 element 為起點，建構問句標的語意樹

在片語組中尋找核心的名詞片語（NP），我們可以從前述所取得的標的片語組中，得到標的 element 是(NP the algorithm)，接下以這個標的 element 為起點，我們可以採用以下步驟，將該標的片語組建構成一個問句的標的語意樹。

(4-1).以(NP the algorithm)為起點建構語意樹，如下圖 4。

(4-2).(NP (NP the algorithm) (PP of (NP Push)))連結 PUSH 的 element。

(4-3).建構語意樹(NP (NP (NP the algorithm) (PP of (NP Push))) (PP in (NP the stack)))。

(4-4).接著由 the algorithm of PUSH 所形成的 element 包含於 stack 之中，因此這個問句的標的語意樹表現如下圖 5。

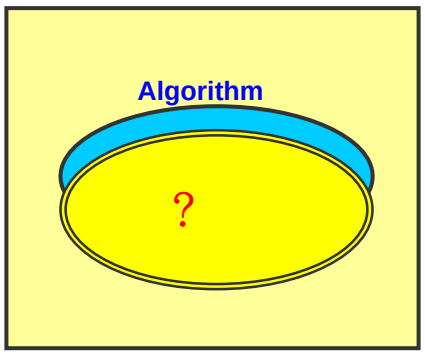

圖 4. 以標的 element 為起點建構語意樹

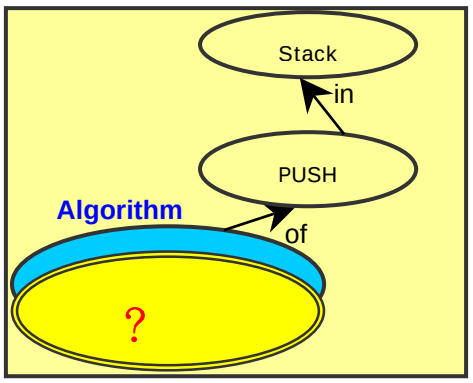

圖 5. 建構語意樹 – 標的語意樹

(5). **知識本體比對與回答：**將上一步所建構的語意樹比對到知識本體。如果找到一樣的語意樹，該 element 的資料就是問題的答案，與前述不同的是，現在要在『資料結構』的知識本體中，找出是否存在一個能對應到先前產生的問句標的語意樹。方法如下：

(5-1).首先我們必須在知識本體裡先找到先前所描述的標的 element - (NP the algorithm)。

(5-2).在上述步驟中，我們在知識本體中用標的 element 關鍵字去尋找的節點可能不存在或存在一個節點以上。如果節點不存在，代表該知識本體找不到問題的答案（系統無法回答）。如果存在一個以上的節點，我們就要從這些找到的節點，同時再往下一個節點尋找連結。如本範例中，我們接下來要找尋(NP (NP the algorithm) (PP of (NP Push)))。

(5-3).如同步驟 5-2，我們必須保留所有符合連結關係的節點，並且在知識本體上繼續延伸搜尋。接下來我們找尋(NP (NP (NP the algorithm) (PP of (NP Push))) (PP in (NP the stack)))。

(5-4).經過了完整的的比對之後，我們在這個知識本體上找到一樣的問句標的語意樹，因此我們知道問句”What is the algorithm of PUSH in the stack?”的答案，就是該標的 element (NP the algorithm) 的資料，如下圖 6 所示：『if (stack full) success = false …』。

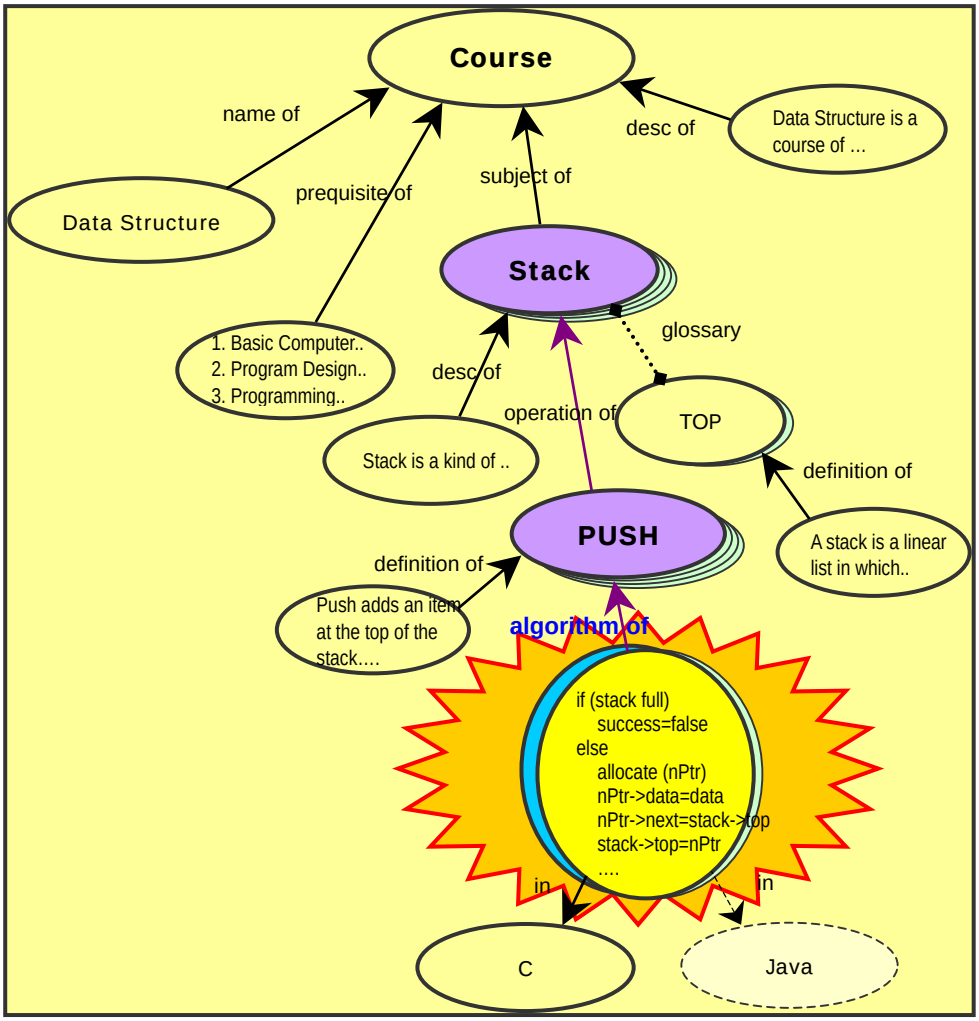

圖 6. 知識本體找尋語意樹 – 找到答案

『語意問答系統應用程式架構』：此部分之研究與設計方法由本實驗室黃世豪[15]所發表。本研究將專注於問答系統之應用程式系統架構，由於整個架構龐大，教育與問答機制更須因應時代科技而作變化，整個系統設計須在設計的初期便將系統定義為可擴展的（extensible）可加掛的（pluginable），並以此作為發展系統的必要條件；更因系統架構龐大，設計模組化將使得本研究計劃可逐步完成，最初勾勒出所需各模組並以 API 作為模組間之溝通管道，先定義介面（Interface）再進行開發，先定義各模組提供的服務方法（Service Method），再進行細部實作；由於本語義問答系統仰賴幾個重要的開放原碼計畫（Open Source）及若干公開標準（Open Standard），整合這些系統為本研究的重大工程，主要工作為：如何粹取所需 API 及整理使用方法，並本著不絕對綁定特定系統的原則，避免系統將隨著這些計畫隕落而滅亡，並給本研究一個可以更換任何 Open Source（甚至是 Third Party 的產品）的彈性空間，本問答系統將根據上述原則及精神設計語意問答系統。

而本研究部分將針對軟體系統設計架構為核心，系統設計將以模組化設計為大方向原則，系統規劃設計將分為四大模組：SemQAService、SentenceGrammarParser Service、KeywordExtensionProcessor 及 Ontology Service，並根據單一責任原則（The Single Responsibility Principle）的擴大解釋，一個模組應該只有一個變動的理由，每一模組負責一個單純專任的商業邏輯（Business Logic）或者強調其為商業責任（Business Responsibility），物件導向軟體工程的成功或多或少和他適用於大型專案開發有關，將專

案模組化之後，再根據介面隔離原則（The Interface Segregation Principle，ISP）或者以介面開發系統原則（Program to Interface），先定義各模組會交互使用的部分，再分派各單位以平行的時間開發，Interface 將只定義需要被使用到的部分，無須令每個需要使用本模組功能的模組，都要實際了解一次複雜的商業邏輯和實作方法，並且使用端不應該依賴他們不使用的方法，在 GoF Design Pattern 裡又有一類似的設計樣式在說明本設計稱為 Façade（單一窗口，在法文裡為建築物正面），並以此精神進行系統分析。SemQAService 為本系統的主要介面，對整個系統而言本介面可稱之為本系統之 Service Façade；下圖 8 為本模組之概念 Class Diagram 設計。

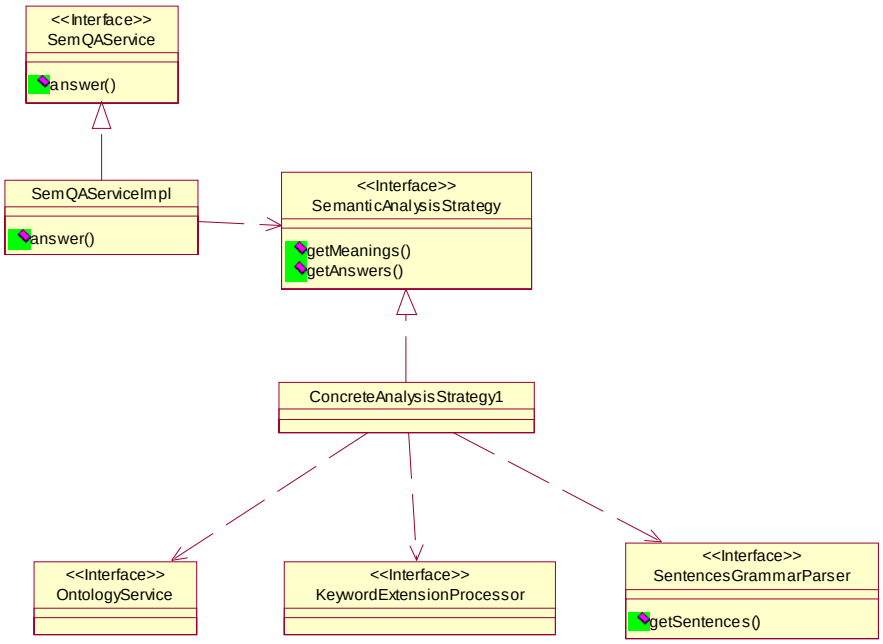

圖 7.SemQAService 模組之概念 Class Diagram

SentencesGrammarParser 之設計為分析取得一文法樹作為判斷語意的重要依據，本模組預設實作將以 Link Grammar 為實作，根據其使用優勢建構 ConstituentTree 物件，Class Diagram 如下圖 9。

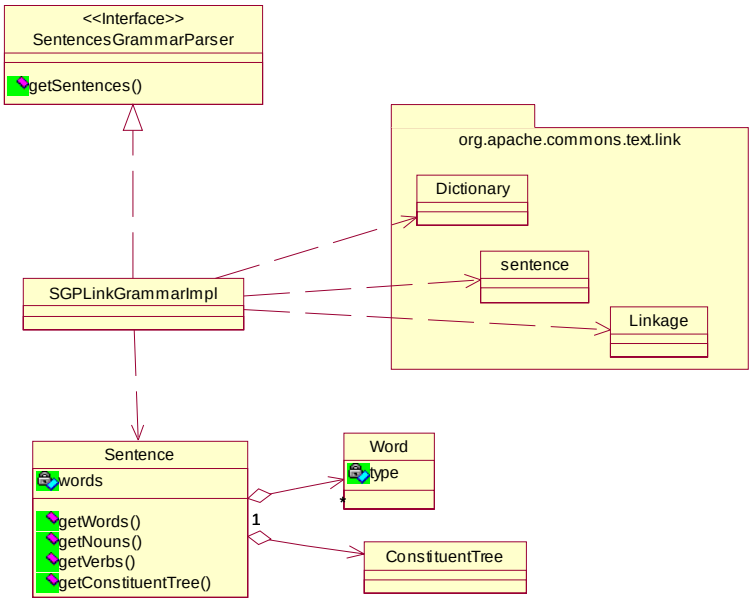

圖 8.SentencesGrammarParser 模組之概念 Class Diagram

結果與討論：

本研究計畫實行第一年之研究成果摘要如下：

- 本研究提出一個語意感知功能之問答系統架構，本架構具有可擴充、可加掛、可替換、易於整合且開放式的特色。
- 提出一個以 Link Grammar 為基礎的語意認知方法，使用自然語言（英文）輸入問句來詢問系統；可提供系統回覆答案的精確度，不會像搜尋引擎一樣回傳若干沒有參考價值的資料。
- 知識本體以 WordNet 擴充相關關鍵字，因此即使使用者沒有輸入一模一樣的關鍵字，也可以有效的查詢出真正想要的資料。
- 問句句型的 Pattern Matching Table 可以提供更多句型的擴充。

本研究目前主要的限制如下：

- 目前的知識本體結構以及 element 關係的設計都偏向靜態，目前只能處理有限的問句類型。
- 目前 element 之間的關係，只使用到包含的關係，因此也會影響到這個系統能有效處理的英文問句的數量。
- 採取的搜尋方法是要滿足該語意樹中所有存在節點中的 element 以及關係（目前多由介係詞擔任這個任務）。所以演算法容錯設計沒考慮到的部分，會直接影響可能找不出適當的答案。

本研究計畫下一年度研究方向為：

- 對於結構樹中 element 之間的關係類型，可以做更多、更深入的分析與規劃，以期能擴充本語意認知方法能處理的問句類型。例如”How to get the data from the stack?”，標的的答案應該是 algorithm 的資料，且 get the data 的語意隱含著要對應到的 algorithm 的 element 應該是 POP。雖然我們有 Pattern Matching Table 可以對問句類型來做定義，但是不應該把 get 這樣的一般動詞定義到 Pattern Matching Table 之中，所以我們需要額外的元素來修正或擴充本語意感知方法。
- 將 Pattern Matching Table 做進一步的分析擴充，以適應更多問句類型的處理能力。
- 於 Ontology 方面由於實作甚多，各 Ontology 組織如 DAML、RDF、KAON 等皆有其相對優點，我們將 OntlogyService 定義為一開放 Interface，並且整合進入本研究，如能在下年度找出適用本研究之 Ontology 對語意問答系統將有莫大幫助。
- 如何讓知識本體具有擴充能力，能夠透過 Internet 或者指定之線上教材來源（例如 SCORM 教材網站）將網路上的教材正確地產生連結及組合，更是未來下一年度應努力的研究議題。

Reference：

[1] Daniel Sleator, David Temperley, and John Lafferty, “Link Grammar,” http://www.link.cs.cmu.edu/link/

[2] George Miller, “WordNet, a lexical database for the English language,” http://www.cogsci.princeton.edu/~wn/

[3] Michael Denny, "Ontology Tools Survey, Revisited," http://www.xml.com/pub/a/2004/07/14/onto.html

[4] FZI and AIFB, "KAON is an open-source ontology management infrastructure targeted for business applications," http://kaon.semanticweb.org

[5] Advanced Distributed Learning (ADL), http://www.adlnet.gov/index.cfm

[6] Rod Johnson, Juergen Hoeller, Alef Arendsen, Colin Sampaleanu, Darren Davison, Dmitriy Kopylenko, Thomas Risberg and Mark Pollack, "Spring - java/j2ee Application Framework Reference Documentation," Version 1.1.2, http://www.springframework.org/docs/spring-reference.pdf

[7] 物件導向軟體工程概念模型, http://www.dotspace.idv.tw

[8] Erich Gamma,Richard Helm,Ralph Johnson,John Vlissides, "Elements of Reusable Object-Oriented Software," Addison-Wesley Professional Computing Series, 1995

[9] Sun Microsystems, "Java Native Interface," http://java.sun.com/j2se/1.4.2/docs/guide/jni/index.html

[10] 陳意芬, 概念式自動問答探索系統, 2003 年, 交通大學碩士論文

[11] Daniel Jurafsky and James H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition," Prentice Hall, 2000.

[12] Minhwa Chung and Dan Moldovan, "Parallel Memory-Based Parsing on SNAP," Parallel Processing Symposium, 1993., Proceedings of Seventh International, page(s): 680-684

[13] Ying-Hong Wang, Wen-Nan Wang and Chih-Hao Lin, "An Intelligent Semantic Agent for e-Learning Message Communication," Journal of Information Science and Engineering, Vol. 21, No.5, September 2005, pp. 1031-1051

[14] 黃朱麒, 在遠距學習環境下以 Link Grammar 為基礎的語意認知方法, 淡江大學碩士論文, 2006

[15] 黃世豪, 語意問答系統應用程式架構, 淡江大學碩士論文, 2006