# 英文短句的知識分析與萃取演算法


葉衡諭
銘傳大學
資訊傳播工程學系
桃園縣龜山鄉大同村德明路 5 號
henry_yea@hotmail.com

李文翔
銘傳大學
資訊傳播工程學系
桃園縣龜山鄉大同村德明路 5 號
kevinlee90008@hotmail.com

張家瑋
銘傳大學
資訊傳播工程學系
桃園縣龜山鄉大同村德明路 5 號
e08dcw8x4ff@hotmail.com

李明哲
銘傳大學
資訊傳播工程學系
桃園縣龜山鄉大同村德明路 5 號
leemc@mcu.edu.tw



***摘要*—近年來受惠於國內外各項語料庫資源的建置及網際網路上的大量資源，對自然語言的輔助教材日趨廣泛，句子的相似度在自然語言處理中有著廣大的應用。如何以高精準度判定語句相似的方法受到研究者的重視。本研究以短句為基礎的英文詞語做相似度的判別。在不同的語句當中，利用 Link Grammar 連結找出每句之間相關連結的字詞予以判斷，經過正規化後再做計算，判對其相似程度的高低，以提供更多元的語意相似搜尋的應用平台。**


## 一、緒論

隨著資訊科技的進步，資料快速的累積成長，讓資料查詢方便許多，但卻也造成使用者不知如何去篩選管理這些龐大的資料量，使得真正需要的資料被埋藏其中，未達到相對的應用。

知識管理來臨的時代，各處都蘊藏有豐饒的資料量，我們必須能夠從中整理出我們需要的知識，建造一個方便搜尋的知識資料庫，運用知識來解決問題，並進一步創造新的知識。

目前在資訊擷取(Information Retrieval)以及知識管理(Knowledge Management)的相關研究領域與文獻中，大多著重在句子對文章的相似度、文章對文章的相似度、或自動文章摘要等。對於短句間相似度的判定尚未有深入的分析與探討。

相較於以往的句子相似度計算，是依照語句的分析深度分成兩種方式。其中一種方法是基於向量空間模型的方法，把句子當成詞的線性序列，因此語句相似度衡量機制只能利用句子的表層資訊，即組成句子的詞的語義資訊。由於不加任何結構分析，這種方法在計算語句之間的相似度時無法考慮句子整體結構的相似性。本研究擬定先將字句於 Link Grammar 連結中的連結關係，找出符合需求連結的字詞，利用 WordNet 資料庫計算出相似值，再根據距離公式算出兩字句間的距離，作為相似程度的依據，提供更精準的資訊給使用者。

## 二、相關研究

### (一) 語意網(Semantic Web)

Tim Berners-Lee 於近年中提出了一個概念「語意式網站 (Semantic Web) [1]，同時也被稱為第二代全球資訊網。Tim Berners-Lee 定義語意式網站為「一個可以被機器所理解的網站」，同時也是一個資訊的集合體。既然語意網的目標是讓機器達到瞭解(machine understanding)的目標，瞭解在某種程度上十分接近推理的含意，因為這些資訊能夠被機器所解譯，故能推論出新的知識，這是單純的資訊存取和比對所無法做到的。為了達到語意網的目的，本研究所採用的方式為使用本體論定義不同領域所用到的知識，這些知識包含字彙和關係，本體論以 XML based 的方式表達以方便網路資源存取。語意網中的本體論(ontology)運用在網路的資訊表達，可達到兩個功用：分類(taxonomy)和推論(reasoning)，分類為了將不同類別的資訊作區分，並可將之視作階層化的表示，而推論結識結合類別與階層性的關係，將隱性知識發掘出來。

(二) 知識本體[2]

知識本體（Ontology）最早是應用在哲學的領域，表示“存在的，有意識的實體或主體”的意思。在計算機科學的領域，本體論是用來表達對於某一個特定領域（Domain）的概念和知識的描述，也可以表達該領域下所存在的事件與彼此間的關係，進而讓機器可以互相分享與了解該領域的知識。對於本體論，最常被引用的定義如下：“An ontology is specification of a conceptualization.” [3]，此定義表達出本體論是某個概念上清楚詳細的說明。當我們要使用本體論來描述某特定領域下的知識時，本體論便是由概念（Concept）、屬性（Attribute、Property）、實例（Instance）與關係（Relation）等元素所組合而成[4]。

當某個領域的專家想要建立知識領域的本體論時，為了讓電腦看得懂專家所建構出的本體論，必須有一種電腦所能理解的語言來轉換他所描述的本體論，以便讓電腦了解本體論所表達的語意概念。目前有許多的本體論描述語言（Ontology language）被發展出來，這些本體論描述語言都是以 XML 語法為基礎所發展出來，例如: XOL（XML-based ontology-exchange language）[5]、OML（Ontology Markup Language）[6]、SHOE（Simple HTML Ontology Extensions）和 RDF/RDFS（Resource Description Framework Schema）[7] [8]，還有在 RDF/RDFS 的更上層所發展出來的 DAML+OIL（DARPA Agent Markup Language + Ontology Inference Language）[9] [10]，改進 RDF/RDFS 功能上的不足。

本研究所採用的知識本體為普林斯頓大學所開發的 WordNet。WordNet 是由普林斯頓大學的心理學家、語言學家和計算機工程師共同設計的英語辭典，目標是建立英語辭彙及其詞性關係的資料庫，而系統架構為多棵的語意樹所形成，並且這個系統中的名詞、動詞和形容詞都聚類為代表某一基本詞彙概念的同義詞集合（SynSet），它是 WordNet 裡的基本單位，有相同詞義的詞都會被收集放在同一個同義詞集合裡，並在這些同義詞集合之間建立起各種詞義關係（Semantic Relation）這些關係包括反義關係（Antonymy）、上位關係（Hypernymy）、下位關係（Hyponymy）、整體－部份關係（Holonymy）、部份－整體關係（Meronymy）、轉指關係（Metonymy）、近義關係（Near-Synonymy）、同義關係（Synonymy）以及方式關係（Troponymy）。

(三) 詞網 WordNet[2]

傳統的詞典信息與現在計算機技術以及心理語言學的研究成果所結合的一個產物。一般的詞典都是按照字母順序排列，但是根據心理語言學及認知科學的發展，心理語言學家們漸漸認識到詞彙在大腦中的儲存方式，除了像現存字典所有的訊息以外，還有連著一些意義有關聯的詞，但是這樣的訊息在一般詞典內是被分散的。

因此從 1985 年起，普林斯頓大學承擔起開發一部詞典數據庫的任務，就是 WordNet，希望能提供一個與傳統的線上辭典緊密結合的輔助工具。最大的特色是打算用概念而不只是依照傳統字母順序查字典，根據詞義而不是詞型來組織詞彙訊息。目前 WordNet 有 95600 個不同詞型，組成 70100 個同義詞的集合。但是只提供 open words 的資訊，包含動詞、名詞、形容詞、副詞；介係詞、代名詞之類的 function words，也就是不太容易改變的詞類，不在他們打算做的範圍內。在 WordNet 中有下列詞之間的關係:

一、同義關係:

在 WordNet 中最重要的關係，利用矩陣的方法，可以列出多義詞和同義詞。

二、反義關係:

不好決定的關係，例如說：高興的反義不一定是生氣。

三、上下位關係:

或是子集，或 ISA 關係(an x is a kind of y)；上下位關係通常有某種限制，且是一種不對稱的關係，而且通常只有唯一個上層，因而產生一種層次語義結構。下位詞繼承了他的上位詞，更一般化概念的所有性質並且增加屬性，以區別本身及

上位詞和該上位詞的其他下位詞。例如:"楓樹"繼承了上層"樹"的屬性，但有"可用於製作糖漿的樹液"的特性區別於其他樹。這種方法為 WordNet 中的名詞提供一種核心的組織原理。

四、部分關係(an x is a part of y):

而目前除了 WordNet 之外還有微軟的 MindNet，歐洲有基於 WordNet 的 EurowordNet，日本有電子辭書研究所的日語及英語概念辭典，還有美國 High Performance KB 等。中國的則是知網(HowNet)：主要包括中英雙語、中文簡體、中文繁體知識辭典；概念的主要特徵、次要特徵；動態角色屬性；詞類表；反義、對義關係表等等。

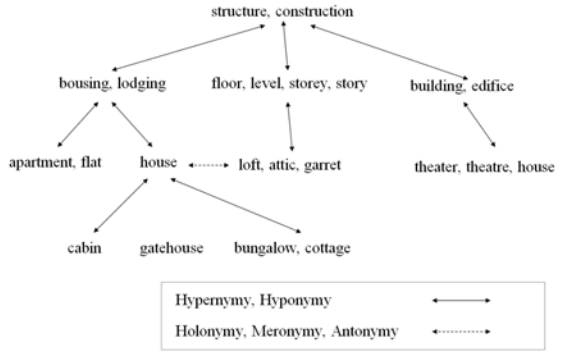


圖 1. WordNet 階層概念表示圖[11]

(四) Link Grammar[12]

過去幾十年來，已經有很多對於 NLP (自然語言處理) [13]的方法，但是很少像 Daniel Sleator, Davy Temperley 和 John Lafferty 所提出 Link Grammar parser 這樣的方法，Link Grammar 是基於在文章中單字和其他單字間有著連結的模型，是一個全新且無須上下文的文法正規化的方法。這些連結不只被用來辨別詞性，更用來詳細地描述單字在句子中所扮演的功能，當片語由兩個形容詞和兩個名詞所組成，但是你真正想知道的是哪個形容詞修飾哪個名詞，此時 Link Grammar 就可以為你解釋。

在 Link Grammar 中每個辭典的單字都被給了一個定義用來描述在句子中會如何使用。文法在單字之中是分散式的。像這樣的系統被稱為 LEXICAL。這樣的系統有著相當多的優點，可以容易的建構一個龐大的文法架構，因為改變一個單字的定義只影響那個單字在句子的文法。此外，對於表示不規則動詞的文法也相當簡單，對於每個不規則動詞單字會區分定義。不像片語結構文法，在利用 Link Grammar 單字來分析一個句子的文法後，可以非常通順及符合語意學構造的來連結每一個單字。

Link Grammar 是基於一個叫做 planarity 的特性；planarity 敘述了一個在最自然的語言中所呈現的現象，假設在句子裡的每一對相關單字之間劃出了弧形，如果弧形和另一個弧形不會交叉到，這個現象就稱為 planarity。

一個 Link Grammar 是由集合的單字所構成；一連串的單字是一個由文法定義的句子，如果單字之中存在一個形式以便滿足下列條件來劃出弧形：Planarity：連結不能交叉到，Connectivity：連結滿足所有一連串的單字連結；Satisfaction：連結滿足在一連串單字中每個單字的連結需要。在 Link Grammar 用語裡，linkage 是一句子中一個符合語法分析：一個沒有連結弧形交叉的連結集合。以下為 Link Grammar 判斷句子架構[14]的一些簡單範例：

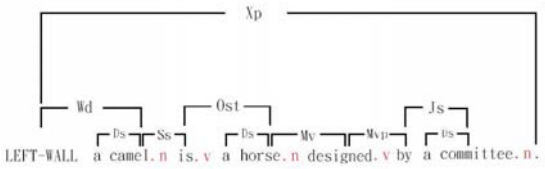


圖 2. Link Grammar 連結表示圖

主要的詞性利用.n 和.v 來個別標記出這些單字是名詞和動詞。單字之間的連結標記出連結的型態。舉例來說， J connector 在這句子中標記出介系詞和受詞的連結；D connector 標記出冠詞和受詞的連結；X 連結則為 LEFT-WALL(句子開頭)和句點的連結。

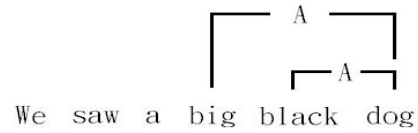


圖 3. Link Grammar -A 連結表示圖

A 為標示形容詞修飾名詞的關係；上圖中 big 和 black 都是形容詞用來修飾名詞 dog。


圖 4. Link Grammar-BI 連結表示圖

BI 為標示句子中連結由 be 動詞所組成的慣用語。

由上述三個範例可以發現，Link Grammar 會將英文句子中的架構利用詞語詞之間互相修飾的關係，加以把句子分解成不同的結構；本篇論文將使用於分析各種句子裡的詞性和詞與詞的關係。

## 三、系統架構

本計劃希望能針對多重以上的詞句做相似度的判定，開發一套有效且具有高準度的詞句判斷引擎。針對個別的使用者，提供準確性高的相似判對。在此判斷引擎下，使用者可以依自己需求，判斷相關的詞句，從中取得最為相近的詞句。

(一) 系統模組介紹

一、 語意量化子系統

進行句子連結的正規化 (Sentence Formalization)，將執行 Link Grammar 連結後的句子篩選出符合需求的連結，依據這些連結計算出相同連結單字的相似度，製成“相似連結型態矩陣”。

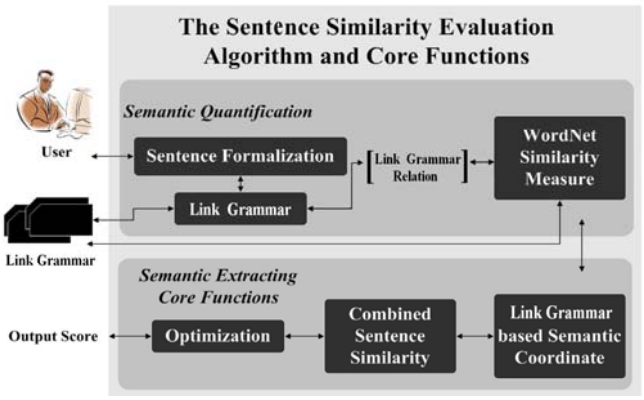


圖 5. 演算法架構圖

1.連結型態模組

在本系統主要是依據名詞、動詞相關的連結，挑選出本研究所需之連結型態。這些連結型態主要有三大類：1.名詞+名詞、2.形容詞(副詞)+名詞、3.形容詞(副詞)+動詞，如附錄 A 所示，其餘的連結型態則忽略不看。

2. 單字相似度測量模組

挑選過後的連結型態製成“相似連結型態矩陣”後，將兩句所對應連結型態的單字，利用 WordNet 資料庫 (WordNet Knowledge Base)計算出此相同連結的單字相似度。

二、 語意推論核心系統函式

計算“相似連結型態矩陣”的資訊，根據每個節點的權重給予計算，每個連結型態計算過後會出現一個 CVE 值，將每個連結型態的 CVE 值經由明可夫斯基語意距離公式計算得到相似值，再做正規化，依據所算出距離的遠近判斷句子相似程度的大小。

(二) 句子相似度計算演算法

一、兩句子 $SEN_A$, $SEN_B$ 的相似度定義如下：

$$Similarity_{(WORD\_A,WORD\_B)} = \text{Similarity}_{\text{Minkowski}}\left(SEN_A, SEN_B\right)^{-1}, \quad \text{if } A\_Gra \cap B\_Gra \neq \phi \tag{1}$$

在公式(1)中，Similarity ($SEN_A$, $SEN_B$)分成兩部分去判別句子的存在的關係，利用 Link Grammar 先判斷句子有無的相關連結，若有附錄 A 的相關連結則用 $Similarity_{Minkowski}(SEN_A, SEN_B)$。因為定義成距離，距離越近代表越相關，求得句子的相似度。在公式(1)中，A_Gra 與 B_Gra 為 $SEN_A$ 與 $SEN_B$ 中由 Link Grammar 所剖析的 connectors 的集合，其正式定義如下：

Definition 1. $SEN_A$ 與 $SEN_B$ 中連結型態(connector type)集合:

A_Gra = {$L_{A\_x}$: x 為連結型態, $L_{A\_x}$ 為此連結型態的子型態之集合}

B_Gra = {$L_{B\_y}$: y 為連結型態, $L_{B\_y}$ 為此連結型態的子型態之集合}

$L_{A\_x}$ = {$L^i{}_{A\_X}$ : i 為連結型態 x 的子型態}

$L_{B\_y}$ = {$L^j{}_{B\_y}$ : j 為連結型態 y 的子型態}

若 A_Gra 與 B_Gra 的交集不為空集合，則 $SEN_A$ 與 $SEN_B$ 的語意相似度定義為明可夫斯基語意距離(Minkowski Semantic Distance)，如公式(2)所示：

$$Similarity_{Minkowski}(SEN_A, SEN_B) = \left\{\sum_{i=1}^{n}\left[Dist_{Grammar}\left(SEN_{Ai}, SEN_{Bi}\right)\right]^p\right\}^{1/p} \quad (2)$$

其中距離參數(distance parameter) p 表示系統對語法差異的衡量原則(在此取 p=2)。i 表示第 i 組 connector. $Dist_{Grammar}(SEN_{Ai}, SEN_{Bi})$為 $SEN_{Ai}$，$SEN_{Bi}$ 在第 i 組 connector 的語意距離。將每組 connector 的語意距離經過距離參數的計算後加總起來，再經過 1/p 計算，算出 $SEN_A$ 和 $SEN_B$ 的明可夫斯基語意距離。

二、 $SEN_{Ai}$，$SEN_{Bi}$ 的語意距離定義如下：

$$Dist_{Grammar}(SEN_{Ai}, SEN_{Bi}) = \sum_{k=1}^{Min\left(\left|L_{A\_x}\right|,\left|L_{B\_y}\right|\right)} Max\left[CVE_k\left(L_{A\_x}, L_{B\_y}\right)\right] \quad (3)$$

在本研究所設計的語意比對演算法中，兩句中相似的型態以“相似連結型態矩陣” (Similar linking-type matrix)表示，如圖 6 所示。

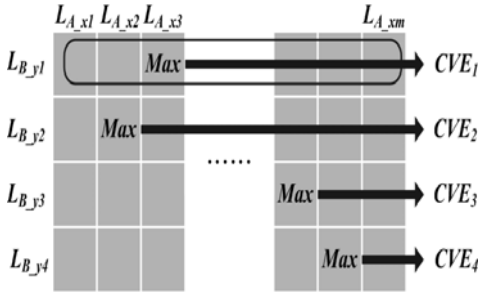


圖 6. 相似連結型態矩陣

“相似連結型態矩陣”為一| $L_{A\_x}$ |×| $L_{B\_y}$ |矩陣，其中| $L_{B\_y}$ |<| $L_{A\_x}$ |。在公式(3)中 Max[$CVE_k$ ] (Connector Vector Evaluation) 為矩陣第 k 列中值最大者；Min(| $L_{A\_x}$ |,| $L_{B\_y}$ |)表示在相似連結型態矩陣中找出最小值| $L_{B\_y}$ |，取得每一個連結型態的 CVE 值，將其加總。

三、 Connector Vector Evaluation - CVE

$CVE_k$ = {$W\_CON_{(k,t)}$ : 相似連結型態矩陣中第(k,t)個節點的權重，其中 1<k<| $L_{B\_y}$ |, 1<t<| $L_{A\_x}$ |}。

$W_CON_{(k,t)}$的計算公式如下：

$$W\_CON_{(k,t)} = \begin{cases} \sum Sim_{WordNet}(WORD_A, WORD_B), \text{for Nouns and Verbs - A and B} \\ \sum\left[\left(1 + Correlation_{(Modifier-Pair)}\right) \times Sim_{WordNet}(WORD_A, WORD_B)\right] \\ \text{for words A and B with modifiers} \end{cases} \quad (4)$$

公式(4)中，$Sim_{WordNet}(WORD_A, WORD_B)$為兩單字在 WordNet 中的語意距離，Correlation(Modifier_Pair)為其修飾詞(形容詞或副詞)的相關係數。在本研究中，主要考慮三種修飾的類型：1.名詞+名詞、2.形容詞(副詞)+名詞、3.形容詞(副詞)+動詞，若兩修飾詞屬同義關係(synonymy relation)，則其相關係數定義為 1；若兩

表(一) 範例句子

| 句子 | 原始句子 |
|---|---|
| $SEN_A$ | *The manager arrives at the office early to set a good example to other workers.* |
| $SEN_B$ | *Teachers always arrive at the school early because they want to be a good example to their students.* |
| $SEN_C$. | *A church school is a school which has a special relation ship with a particular branch of the Christian church.* |

修飾詞屬反義關係(antonymy relation)，則相關係數定義為 0.5；若兩修飾詞無同義或反義關係，則定義為 0；為了避免 Correlation$_{(Modifier_Pair)}$值為 0，所以將其加上 1。

(三) 句子相似度計算演算法實作

本小節給予一個例子說明本論文所提出的句子相似度演算法。將表(一)三句句子 $SEN_A$、$SEN_B$ 及 $SEN_C$ 計算出其相似程度。

步驟 1. 利用 Link Grammar 找出句子連結型態

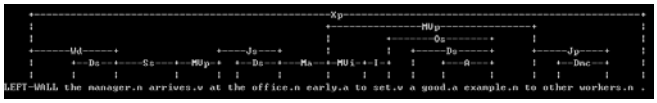


圖 7. $SEN_A$ 連結型態關係圖例

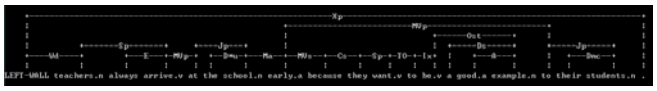


圖 8. $SEN_B$ 連結型態關係圖例

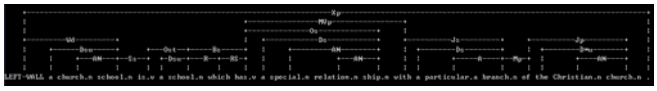


圖 9. $SEN_C$ 連結型態關係圖例

步驟 2. 將需要的連結型態形成“相似連結型態矩陣”，並將相同連結的單字利用公式(4)得到其相似值

先以 $SEN_A$ 作為要比對的句子與 $SEN_B$ 和 $SEN_C$ 比較，再者 $SEN_B$ 和 $SEN_C$ 作比較，將連結形態所連結的單字標示在其連結下。

$SEN_A$-$SEN_B$ 的相似連結型態矩陣

| | | Ds | | | Dmc | Ss | MVp | | MVi | Ma | Js | Jp | Os | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | the-manager | the-office | a-example | other-workers | manager-arrives | arrive-at | early-to | early-to | office-early | at-office | to-workers | at-example | good-example |
| Sp | teacher-arrive | | | | | 0.53 | | | | | | | | |
| MVp | arrive-at | | | | | | 1 | 0 | | | | | | |
| | early-to | | | | | | 0 | 1 | | | | | | |
| MVi | early-become | | | | | | | | | | | | | |
| Ma | school-early | | | | | | | | | 0.71 | | | | |
| Jp | at-school | | | | | | | | | | | 0.4 | | |
| | to-students | | | | | | | | | | | 0.72 | | |
| D*u | the-school | | | | | | | | | | | | | |
| Ds | a-example | 0.14 | 0.62 | 1 | | | | | | | | | | |
| Dmc | their-students | | | | 0.73 | | | | | | | | | |
| TO | want-to | | | | | | | | | | | | | |
| Ost | be-example | | | | | | | | | | | | | |
| A | good-example | | | | | | | | | | | | | 1 |

圖 10. 相似連結型態矩陣圖例(1)

$SEN_A$-$SEN_C$ 的相似連結型態矩陣

| | | Ds | | | Dmc | Ss | MVp | | MVi | Ma | Js | Jp | Os | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | the-manager | the-office | a-example | other-workers | manager-arrives | arrive-at | early-to | early-to | office-early | at-office | to-workers | at-example | good-example |
| Dmc | a-school | | | | | | | | | | | | | |
| | a-school | | | | | | | | | | | | | |
| Ds | a-ship | 0.22 | 0.5 | 0.14 | | | | | | | | | | |
| | a-branch | 0.31 | 0.84 | 0.4 | | | | | | | | | | |
| D*u | the-church | | | | | | | | | | | | | |
| AN | church-school | | | | | | | | | | | | | |
| | special-ship | | | | | | | | | | | | | |
| | relation-ship | | | | | | | | | | | | | |
| | Christian-church | | | | | | | | | | | | | |
| A | particular-branch | | | | | | | | | | | | | 0.9 |
| Ss | school-is | | | | | 0.35 | | | | | | | | |
| Ost | is-school | | | | | | | | | | | | | |
| Os | has-ship | | | | | | | | | | | | 0.14 | |
| MVp | has-with | | | | | | 0 | 0 | | | | | | |
| Mp | branch-of | | | | | | | | | 0.84 | | | | |
| Js | with-branch | | | | | | | | | | 0.84 | | | |
| Jp | of-church | | | | | | | | | | | 0.38 | | |

圖 11. 相似連結型態矩陣圖例(2)

$SEN_B$-$SEN_C$ 的相似連結型態矩陣

| | | Sp | MVp | | MVi | Ma | Jp | | D*u | Ds | Dmc | Ost | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | teacher-arrive | arrive-at | early-to | early-become | school-early | at-school | to-students | the-school | a-example | their-students | be-example | good-example |
| Dmc | a-school | | | | | | | | | | | | |
| | a-school | | | | | | | | | | | | |
| Ds | a-ship | | | | | | | | | 0.14 | | | |
| | a-branch | | | | | | | | | 0.4 | | | |
| D*u | the-church | | | | | | | | 0.88 | | | | |
| AN | church-school | | | | | | | | | | | | |
| | special-ship | | | | | | | | | | | | |
| | relation-ship | | | | | | | | | | | | |
| | Christian-church | | | | | | | | | | | | |
| A | particular-branch | | | | | | | | | | | | 0.9 |
| Ss | school-is | | | | | | | | | | | | |
| Ost | is-school | | | | | | | | | | | 0.59 | |
| Os | has-ship | | | | | | | | | | | | |
| MVp | has-with | | 0 | 0 | | | | | | | | | |
| Mp | branch-of | | | | | | | | | | | | |
| Js | with-branch | | | | | | | | | | | | |
| Jp | of-church | | | | | | 0.88 | 0.35 | | | | | |

圖 12. 相似連結型態矩陣圖例(3)

將利用公式(4)計算相對應連結型態單字的相似值。

步驟 3. 取得每一個連結型態的 CVE 值，即為 $Dist_{Grammar}(SEN_{Ai}, SEN_{Bi})$公式(3)

每個連結找出最少數量的取出最大值 $CVE_k$，

將其加總。

$SEN_A$-$SEN_B$ 的相似連結型態矩陣

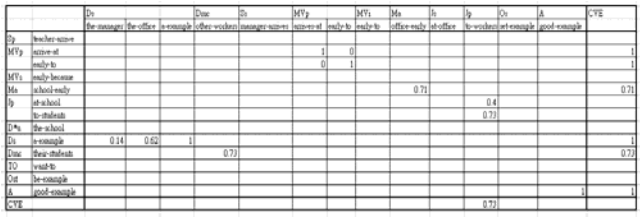

圖 13. 相似連結型態矩陣圖例(4)

$L_{A\_Ds}$：1

$L_{A\_Dmc}$：0.73

$L_{A\_MVp}$：1+1=2

$L_{A\_Ma}$：0.71

$L_{A\_Jp}$：0.73

$L_{A\_A}$：1

$SEN_A$-$SEN_C$ 的相似連結型態矩陣

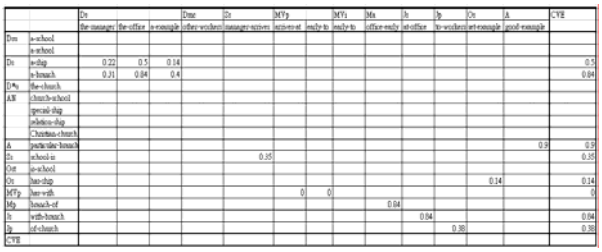

圖 14. 相似連結型態矩陣圖例(5)

$L_{B\_Ds}$：0.5+0.84=1.34

$L_{B\_A}$：0.9

$L_{B\_Ss}$：0.35

$L_{B\_Os}$：0.14

$L_{B\_MVp}$：0

$L_{B\_Js}$：0.84

$L_{B\_Jp}$：0.38

$SEN_B$-$SEN_C$ 的相似連結型態矩陣

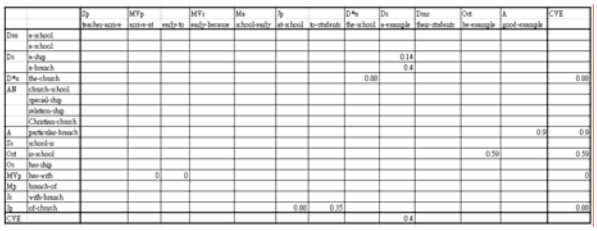

圖 15. 相似連結型態矩陣圖例(6)

$L_{C\_Ds}$：0.4

$L_{C\_D*u}$：0.88

$L_{C\_Ost}$：0.59

$L_{C\_A}$：0.9

$L_{C\_MVp}$：0

$L_{C\_Jp}$：0.88

步驟 4. 將 $Dist_{Grammar}(SEN_{Ai}, SEN_{Bi})$的值做運算

將每個連結 $Dist_{Grammar}(SEN_{Ai}, SEN_{Bi})$的值作公式(2)的運算，再將其值除以連結型態的個數作正規化，使之介於 0~1 之間。

$SEN_A$-$SEN_B$：

$$\frac{\sqrt{(1)^2+(0.73)^2+(2)^2+(0.71)^2+(0.73)^2+(1)^2}}{6}$$

$$= \frac{2.751345}{6}$$

$$= 0.458557$$

$SEN_A$-$SEN_C$：

$$\frac{\sqrt{(1.34)^2+(0.9)^2+(0.35)^2+(0.14)^2+(0.84)^2+(0.38)^2}}{7}$$

$$= \frac{1.89676}{7}$$

$$= 0.270965$$

$SEN_B$-$SEN_C$：

$$\frac{\sqrt{(0.4)^2+(0.88)^2+(0.59)^2+(0.9)^2+(0.88)^2}}{6}$$

表(二) 測試資料與實驗結果

| 配對 *A* | 原始句子 | | |
|---|---|---|---|
| 句子 *A-1.* | Some newspapers keep in reporting the private or scandal of celebrities. | | |
| 句子 *A-2.* | Mass medias transport and spread information whatever it is the truth or rumor. | | |
| 句子 *A-3.* | A group of patients are experimenting with chemical treatment to cure lung cancer. | | |
| 相似值 | *A-1 v.s. A-2* = 0.64701 | *A-1 v.s. A-3* = 0.24592 | *A-2 v.s. A-3* = 0.44 |
| 配對 *B* | 原始句子 | | |
| 句子 *B-1.* | She was convinced that the accident had been engineered by his enemies. | | |
| 句子 *B-2.* | The murderer murder a victim that he has completed plan to let it likes a accident. | | |
| 句子 *B-3.* | The principal's democracy made him popular among teachers and students. | | |
| 相似值 | *B-1 v.s. B-2* = 0.5071735 | *B-1 v.s. B-3* = 0.26745 | *B-2 v.s. B-3* = 0.379033 |
| 配對 *C* | 原始句子 | | |
| 句子 *C-1.* | People are still terribly apprehensive about the increasing crimes in the city. | | |
| 句子 *C-2.* | The high rate of crime makes citizen in the city lead to anxiety and fear. | | |
| 句子 *C-3.* | We use electronic mail to send the latest information to our customers. | | |
| 相似值 | *C-1 v.s. C-2* = 0.301475 | *C-1 v.s. C-3* = 0.216871 | *C-2 v.s. C-3* = 0.26851 |
| 配對 *D* | 原始句子 | | |
| 句子 *D-1.* | John heroically saved the little girl who was stuck in the middle of the river. | | |
| 句子 *D-2.* | The rescue team who would risk their own lives to save the lives of others. | | |
| 句子 *D-3.* | Many countries' politicians declare everyone has equal rights and responsibilities. | | |
| 相似值 | *D-1 v.s. D-2* = 0.22021 | *D-1 v.s. D-3* = 0.15 | *D-2 v.s. D-3* = 0.10977 |
| 配對 *E* | 原始句子 | | |
| 句子 *E-1.* | The Thai army has fought running battles with protesters in the capital, Bangkok, in a bid to end days of mass demonstrations and political chaos. | | |
| 句子 *E-2.* | Thai's government attempt to end days of mass demonstrations and political chaos, also army has quelled and fought with protesters in Thai's capital. | | |
| 句子 *E-3.* | For many years, scientists have been trying to understand the mechanisms behind how the body experiences pain, and the nerves involved in conveying those messages to the brain. | | |
| 相似值 | E-1 v.s. E-2 = 0.513049 | *E-1 v.s. E-3* = 0.368906 | *E-2 v.s. E-3* = 0.322378 |

$$= \frac{1.693192}{6}$$

$$= 0.28219$$

由此可得知 $SEN_A$-$SEN_B$ 的句子相似度為較 $SEN_A$-$SEN_C$ 和 SENB-$SEN_C$ 高；代表第一句及第二句的句子相似度高。

## 四、 實驗

本研究採用 Link Grammar 作為判斷本實驗相關連結的依據，只考慮關於名詞、動詞與形容詞的相關連結，因此將本實驗所需求之相關連結列於附錄(一)中作為參考；另外本研究所採用的知識

本體為普林斯頓大學所開發的 WordNet。在 WordNet 系統中的名詞、動詞和形容詞都聚類為代表某一基本詞彙概念的同義詞集合(Synset) 組織出上位者及下位者的關係。，它是 WordNet 裡的基本單位，有相同詞義的詞都會被收集放在同一個同義詞集合裡，並在這些同義詞集合之間建立起各種詞義關係(Semantic Relation)。

由於 Link Grammar 相關的研究不多，所以缺乏適合的數據來評估本演算法。本研究設計中句及長句的英文句子來實驗，以三句為一組配對實驗。本實驗未提供測試資料人工判斷之數值，但提供實驗結果如表(二)作為參考，讓讀者自行判斷其相似的程度。

## 五、 結論

由人類來判斷文句的相似程度是非常容易的，但由自然語言處理的相關技術去判斷是非常不容易的，因為句子的結構可能很複雜，而自然語言處理技術通常都以一種關係作為判斷的依據，忽略別的條件下所能提供的有效資料。雖然有很多的關於語句相似度判斷的相關研究，但是它們卻很難應用在比較複雜的長句上。本研究設計了特殊的語意空間，著重句子的前後連結的相關性，增強文句在比較上能有更多的相似性，提供使用者有較高精準度的參考依據。

## 六、參考文獻


[1] 語意網, available at “http://www.ws.org.tw/sws/”

[2] WordNet, available at http://bow.sinica.edu.tw/

[3] T.R. Gruber. “A translation approach to portable ontology specifications, ”Knowledge Acquisition, vol.5, issue 2, pp.199-220.1993.

[4] N.F.Noy and D.L.Mcguinness, ”Ontology Development 101: A Guide to Creating Your First Ontology,” Stanford Knowledge System Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880,Mar.2001.

[5] R. Karp, V. Chaudhri, and J. Thomere, “XOL: An XML-based Ontology Exchange Language,” Technical Report, Aug. 1999.

[6] Ontology Markup Language Version 0.3, http://www.ontologos.org/OML/OML%200.3.htm

[7] D.Brickley and R.Guha, “Resource Description Framework(RDF) Schema Specification,” W3C Candidate Recommendation, Mar. 2000.

[8] O.Lassila and R.Swick, “Resource Description Framework(RDF) Model and Syntax Specification,” World Wide Web Consortium Recommendation, Feb. 1999; http://www.w3.org/TR/REC-rdf-syntax/.

[9] D. Fensel, F. van Harmelen, Ian Horrocks, D. L. Mcguinness, and P. F. Patel-Schnedier, “OIL: An Ontology Infrastructure for the Semantic Web”, IEEE Intelligent System, Vol.16, no.2, pp.38-45, March/April 2001.

[10] I. Horrocks, D. Fensel, J. Broekstra, S. Decker, M. Erdmann, C. Grble, and F. van Harmelen, “The Ontology Interface Layer OIL”, August 2000. Available at： http://www.ontokowledge.org.

[11] http://203.64.42.21/course/2005/TGBCL/poko/OHCL-13.htm

[12] http://www.foo.be/docs/tpj/issues/vol5_3/tpj0503-0010.html

[13] 自然語言處理, available at ”http://zh.wikipedia.org/wiki/%E9%A6%96%E9%A1%B5”

[14] http://www.link.cs.cmu.edu/link/dict/index.html

## 附錄一

| | |
|---|---|
| A | 連接名詞前的形容詞與名詞。 |
| AF | 連接形容詞與動詞 (形容詞在前)，如問句或間接問句。 |
| AN | 連接名詞修飾詞與名詞。 |
| BI | 連結由 be 動詞所組成的慣用語。 |
| D | 連結限定詞與名詞。 |
| DD | 連接一些限定的詞("the", "his", "Jane's")與數字或形容詞作為名詞。 |
| DG | 連接”the”的專有名詞。 |
| DP | 連接所有格與動名詞。 |
| E | 用於在動詞前修飾動詞的副詞。 |
| EA | 連接副詞和形容詞，某些可以修飾形容詞的副詞("very", "quite", "relatively") |
| EB | 連接受詞、形容詞、介係詞片語和副詞前的” be ”動詞。 |
| EC | 連接副詞和比較級的形容詞。 |
| EE | 連接副詞和副詞之間，可以修飾其他副詞的副詞("very", "quite")。 |
| EF | 連接"enough"與前面的形容詞或副詞。 |
| ER | 連接有形容詞或副詞比較級的文法("the X-er..., the Y-er...")。 |
| J | 連結介係詞與專有或普通名詞、代名詞受格。 |
| JG | 連結某些介係詞("of" and "for")與專有名詞的受詞。 |
| L | 連接限定詞與最高級的形容詞。 |
| LE | 連接比較級與形容詞的句型結構。 |
| LI | 連接一些動詞("feel" and "seem")與"like"。 |
| M | 連接名詞與沒有逗號的修飾語(介係詞片語、分詞修飾…) 。 |
| MG | 連接介係詞修飾專有名詞。 |
| MV | 連接動詞(形容詞)與修飾片語(如副詞、介系詞片語…) 。 |
| MX | 連接圍繞在逗號附近的名詞和修飾名詞。 |
| O | 連結及物動詞與直接或次要的受詞(名詞、代名詞…)。 |
| OF | 連接某些動詞/形容詞與"of"。 |
| OX | 特定的受詞連接主詞("it" and "there")。 |
| P | 連接“be”動詞與介係詞、形容詞、被動語態、進行式。 |
| PP | 連接”have”和過去分詞。 |
| RS | 用在關係子句的主詞與關係代名詞的動詞。 |
| S | 連接主詞與限定的動詞。 |
| TH | 連接有”that”子句的句子，包含動詞、名詞、形容詞。 |
| TO | 用於連接動詞和形容詞的不定詞(V+to)。 |
| TS | 用在假設語氣的句型結構"suggest","require"-"that"。 |
| U | 主要與名詞連接。 |
| V | 連接各種不同附加的動詞與慣用語，它們可能不相鄰。 |
| YP | 用在所有格代名詞，連接副署名詞。 |