應用機器學習結合語法與語意特徵於中文文本蘊涵關係之研究

(Applying Machine Learning Approach with Syntactic and Semantic Features in Chinese Textual Entailment)

杜駿 (Chun Tu)
淡江大學資訊管理學系
kevincncod2@gmail.com

戴敏育 (Min-Yuh Day)
淡江大學資訊管理學系助理教授
myday@mail.tku.edu.tw

摘要

文本蘊涵辨識 (Recognizing Textual Entailment; RTE)是給定一文本對，並觀察在一文本對的兩文本之間的蘊含關係；語意特徵(Semantic Features)包含文句中同義詞、反義詞、否定詞等文句間之語意關係；機器學習方法目前已廣被應用在自然語言處理與資訊檢索領域。在過去文獻中，有關文本蘊涵辨識的研究，主要著重在英文的文本蘊涵辨識，較少研究探討中文文本蘊涵辨識。因此，本研究提出加入語意(Semantic)特徵方法，並與語法(Syntax)特徵之效果比較並深入探討其對中文文本蘊涵辨識的效果。我們設計了各種語意及語法辨識的特徵值來進行機器學習訓練，用以處理 NTCIR-9 RITE 任務的文本蘊涵辨識。本研究實驗結果顯示，利用我們系統所加入之語意特徵為基礎，配合機器學習之方法，加上參數調整及最佳方法之篩選，在中文文本蘊涵辨識整體準確率達到 73.28%。本研究的主要貢獻為，我們於實驗中加入語意特徵方法於中文文本蘊涵辨識，並深入探討，語意特徵對中文文本蘊涵辨識之準確率有大幅提升之效果。

關鍵字： 文本蘊涵(Textual Entailment)、語意特徵(Semantic Features)、 語法特徵 (Syntactic Features)、機器學習(Machine Learning)、支持向量機(Support Vector Machine; SVM)

Abstract

Recognizing Textual Entailment (RTE) is a task of deciding given two text fragments, whether the meaning of one text is entailed from another text. Semantic features are mainly focused on the semantic relationship between two text fragments, for examples, synonym、antonym and negation words. Machine learning methods have been widely used for Natural Language Processing (NLP) and Information Retrieval (IR). Prior research shows that recognizing textual entailment is more common on English in comparison to Chinese. The purpose of this paper has tended to focus on semantic-based approaches we proposed in order to compare with syntax features and thoroughly evaluate the proposed system on Recognizing Inference in

Text (RITE). In this paper, we present both semantic and syntactic features methods for machine learning within the text fragments provided by NTCIR-9 RITE task. The results showed that the best performance of our system achieved 73.28% within semantic features, machine learning methods, adjusting parameters and optimal features selection. Concerning the contribution of this study, we offer feasible and optimal approaches to greatly enhance the system accuracy on recognizing inference in text in Chinese.

Keywords: Textual Entailment, Semantic Features, Syntactic Features, Machine Learning, Support Vector Machine (SVM)

一、緒論

文本蘊涵辨識(Recognizing Textual Entailment；RTE)是給定一個「T(Text)-H(Hypothesis)對」,辨識 T 是否能蘊涵出 H 這個假設(RTE-7, 2011)。RTE 中分為兩種子任務：RTE 2-ways 和 RTE 3-ways，在 RTE 2-ways 子任務中有兩種輸出結果：蘊涵(ENTAILMENT)和未蘊涵(No Entailment)，Entailment 表示 T 蘊涵 H 假設；NO ENTAILMENT 表示 T 沒有蘊涵 H 假設；在 RTE 3-ways 子任務中有三種輸出結果： 蘊涵(ENTAILMENT)、矛盾(Contradiction)和未知(UNKNOWN)，ENTAILMENT 表示 T 蘊涵 H 假設；CONTRADICTION 表示 T 和 H 假設彼此關係矛盾；UNKNOWN 表示 H 假設無法透過 T 來決定蘊涵關係(RTE-7, 2011)。

RTE 從 2005 年發展以來在對於語意推論之分析、比較和評鑑上逐漸提升了重要性，在歐洲舉辦過 3 次 PASCAL RTE Challenges 之後，於 2008 年成為 TAC(Text Analysis Conference)的比賽項目之一(TAC,2011)。RTE Challenge 是一種效能評鑑的競賽，在效能評鑑中給定 T-H 對並且決定 T 和 H 之間的蘊涵關係，而比賽迄今已舉辦至第七屆(RTE-7, 2011)。

RTE 競賽主要在歐美地區，而 RITE(Recognizing Inference in TExt)主要為東亞地區的文本蘊涵辨識之競賽(NTCIR, 2011)，RITE 是一個效能評鑑任務，目的在評鑑系統自動偵測特定語句「關係」的能力。RITE 可分為 3 種子任務：BC(Binary-class)子任務、MC(Multi-class)子任務、RITE4QA 子任務。BC 子任務，給定一個「文本對」(t1,t2)，辨識 t1 是否能蘊涵或推論出 t2 這個假設(hypothesis)。BC 子任務輸出結果有兩種：Yes 和 No，Yes 表示文本對之間具有蘊涵關係；No 則表示文本對之間沒有蘊涵關係；MC(Multi-class)子任務，為一個五選一(5-way)的標記任務，需辨識文本對之間是否為正向(forward)蘊涵、逆向(reverse)蘊涵、雙向(bidirection)蘊涵、矛盾(contradiction)或獨立(independence)關係，輸出結果分為五種：F(forward)、R(reverse)、B(bidirection)、C(contradiction)、I(independence)；RITE4QA 子任務，此子任務之輸入輸出與 BC 子任務相同，差別在於此任務被特別設計成內嵌於問答系統(Question Answering)，並負責答案確認(answer validation)的一個子元件(NTCIR, 2011)。下列為 BC 子任務之範例：

T1： 香港的主權和領土是在 1997 由英國歸還給中國的。

T2： 1997 年香港回歸中國。

上述文本對以 BC 子任務的觀點進行判斷的話，T1 可以推論到 T2 的關係，因此在文本蘊涵關係上屬於 Yes；以 MC 子任務觀點來判斷的話，T1 可以推論到 T2，但是 T2 卻無法推論到 T1，因此在文本蘊涵關係上屬於 F。下列為 MC 子任務之範例：

T1：尼泊爾毛派叛亂份子攻擊安全警衛哨站。

T2：尼泊爾毛派游擊隊攻擊民航機。

上述文本對以 BC 子任務的觀點進行判斷的話，T1 無法推論到 T2，T2 也無法推論到 T1，因此在文本蘊涵關係上屬於 No；以 MC 子任務觀點來判斷的話，T1 無法推論到 T2，T2 也無法推論到 T1，因此在文本蘊涵關係上屬於 C。

RITE 是 NTCIR(NII(National Institute of Informatics) Test Collection for Information Resources)其中一項效能評鑑任務。NTCIR 是由日本每一年半所舉辦的資訊檢索評估會議，至今 NTCIR 已舉辦到第九屆，目的是讓從事資訊檢索的學者專家，有一個公正、公開、公平的評量機制，各種資訊檢索研究可以在一致的比較基礎下，確認各種檢所技術的優劣，並進一步發展探討更深入的資訊檢索研究。

過去文獻將文本蘊涵關係(RTE)區分為兩大面向來進行分析： 語法(Syntax)特徵與語意(Semantic)特徵。語意特徵包括： 同義詞(Synonym)分析、反義詞(Antonym)分析、否定詞(Negation)分析…等等，大多數研究多半著重於 RTE 上語意特徵所影響的準確率(Ion,Prodromos,2010)，以下述文本對為例子：

T1： 車諾比病毒在 1999 年 4 月總共造成超過 200 萬台電腦無法開機

T2：1999 年 4 月車諾比病毒總共造成逾 200 萬台電腦無法開機

如果單就語法特徵進行判斷的話，結果為正向(Forward)關係，但是加入語意特徵(同義詞)來判斷的話，最後結果為雙向(Binary)關係，符合正確答案。

綜上所述，在進行文本蘊涵辨識中，語意特徵及語法特徵多為大部分研究文獻所討論研究之面向，在語法中判斷文句長度、單詞長度便於文句對之間相似性之分析，然而綜觀上述例子，兩文句中出現的單詞:"超過"、"逾"，兩者在詞意上屬於同義詞的關係，因此若未加入語意特徵進行考慮，在預測結果仍有一定的誤差所產生，因此納入語意特徵之考量仍具有相當之權重。本研究目的將深入探討與觀察語意特徵對於 RITE 中文本對的準確率與語法特徵方法比較之影響關係，並設計語意特徵方法來增進文本蘊涵關係的準確率。

本研究所提出之系統有以下特點：1. 所採用之語意特徵方法於文本蘊涵關係之中，相較單純語法特徵方法，更能夠提升系統預測之效能，也提高系統分析之準確率。2. 加入機器學習之工具，訓練分析過後之資料集，並進行最佳特徵值之篩選與調整，用以預測未知新資料集，更能提升整體系統預測結果之準確率。

本論文章節如下：在第二部分簡述過去研究文獻所使用之研究方法;第三部分描述本研究所使用之系統架構以及所採用特徵之介紹;第四部分，本研究分析

實驗結果並進行討論分析;第五部分，我們總結本研究之結論並提出未來改善方向。

二、 文獻探討

辨識文本對之文本蘊涵關係研究已發展多年，多數會議組織，像是舉辦於東亞地區的 NTCIR(NTCIR, 2011)、歐美地區的 RTE(RTE-7, 2011)等，在文本蘊涵關係辨識方面已投入大量心力研究，而其中本研究所探討之語法特徵算是文本蘊涵關係中的一個分析面向。

一般來說，進行文本對之文本蘊涵關係分析有多種方式，字串長度比較、同義詞林、斷詞系統分析、機器學習等等，而不同國家因為語言、文化影響的關係，進行文本蘊涵辨識上所採用之分析方法也會有顯著差異，因此本研究統整國內外曾用於文本蘊涵關係辨識上之方法加以分門別類： 歐美地區的 RTE、東亞地區的 RITE、以及機器學習三種方式來探討過去文獻。

過去於 RTE 研究文獻將文本蘊涵關係分析主要分成兩個面向： 語法特徵和語意特徵。在考慮英文語法特徵和語意特徵的處理上，Siblini and Kosseim (2008)利用知識本體校準系統(Ontology Alignment System)分別從 Text 和 Hypothesis 進行語法特徵、語意特徵剖析再利用知識本體的知識汲取與分類來處理文本蘊涵關係，然而文本對之間所涵蓋的知識領域有差異性(ex： bank 有銀行,河床之意)，在進行分類之上可能產生語意上的落差; Burchardt et al.(2008)以 SALSA RTE 系統透過語意特徵分析文本對之間關係並將文句間語意相近的詞彙作成一張媒合圖表(Matching Gragh)，再利用 47 種獨立的特徵值統計每張圖表文句間相似性，並進行訓練。然而，此所牽涉的問題和其知識領域也有關聯性，如果涉及領域有所差異，在統計資料進行訓練的同時可能就有誤差產生。

單就以語意特徵方面進行分析來看，Iftene and Moruz (2009)利用正面(Positive)詞彙、負面(Negation)詞彙、矛盾(Controdiction)詞彙以及未知(Unknown)詞彙(ex： maybe、might、should)來分析兩文本對之間英文語意關係。在部分文本對中，語意上可能沒有特別果斷的文意表示，因此有加入未知詞彙來幫助進行分析，增加文本蘊涵辨識的準確性; Soon et al.(2001)專門探討文本對之間英文名詞片語(Noun Phrase)的關係; Callison-Burch et al. (2006)運用統計式機器翻譯 SMT(Statistical Machine Translation) ，找出文章間的對應、句子間的對應、以及片語和詞彙上的對應關係作為統計上的參考。

以語法特徵的方式來進行文本蘊涵關係的分析上，Vanderwende et al. (2006)研究結果指出，49%的文本對可以藉由語法特徵的判斷來協助分析，就語法特徵來衡量兩文本對之間關係有顯著準確率的影響。以語法特徵方式分析，Castillo(2010)利用 Edit Distance、LCS(Longest Common Substring)找出文本對之間文句長度值的差異進行計算; Kouylekov and Magnini (2005)主要採用 Tree Edit Distance(Dependency Tree)方法進行文句間相依性的比較。

綜上所述，在處理英文文本對時，基於英文每個詞彙都是分開的，易於進

行斷詞分析，並且在文法上也有明確清楚的規範，例如：時態、詞性…等等，在掌握英文語法、語意辨識上也相對容易分析。

在東亞地區 RITE 方面，中文、日文的文本蘊涵關係辨識上較英文複雜，在相關研究文獻也相對較少。依文句上來看，中、日文不像英文有明顯斷句進行區隔，因此在語意辨識上難度相對提升。以中文為例，中文起源於象形文字，對於現象的表達力求精確與細緻，幾千年的歷史沉澱使中文詞彙極為豐富，相對於英文是比較直接簡約的語言，因此在處理中文的文本蘊涵關係上所採用的方式與英文會有些許之差異。例如在語意特徵方面，Zhang and Yamamoto (2005)透過樣本基礎方法(Sample-based Method)來進行文句改寫(Paraphrasing)便於分析，進而不透過深度剖析的方式並維持最接近原文句之語意幫助進行文本對語意上之比較；Li et al. (2010) 透過建立簡繁字彙對照表進行轉換，在簡體中文文本對之詞彙上的轉換與繁體中文進行比對，增加語意辨識便於文本對之分析。

綜上所述，在中文文本對其文本蘊涵關係上，語意辨識所使用之方法與英文有所差異，有些英文詞彙翻成中文，同時有多種中文詞意相符合，而不同解釋對於文句的表現都有所不同。而在語法辨識上，因為此方式不涉及語意詞彙的辨識，其所採用之分析方式與英文語法辨識大同小異，因此在中文研究文獻上多半採用英文之語意特徵方法來進行文本蘊涵關係之辨識，而部分研究方法在少量中文研究文獻中並無詳細探討，透過本研究進行其他語法研究方法之補充說明。

本研究採用機器學習(Machine Learning)方式，進行資料集訓練並利用統計軟體進行預測，從過去資料或經驗當中，萃取出感興趣的部份，構造一個模型(Model)，於其中定義不同的參數(Parameters)，以程式的方式讓模型執行學習(Learning)動作，利用測試資料(Training Data)來調整最佳化這些參數，等到訓練樣本到一定的程度後，參數的定義也趨成熟，程式便可進行資料集預測。Malakasiotis and Androutsopoulos (2007)利用SVM(Support Vector Machine)處理文本蘊涵辨識之多項語意辨識的訓練模型，每項語意有十項特徵值進行訓練，共計128項特徵來進行篩選提升辨識準確率；Huang and Chung (2011)採用機器學習：C4.5 Decision Tree和 10-fold Cross-Validation方法來訓練資料進行特徵之篩選便於文件找出適合中文地址擷取之方法。

綜上所述，使用 SVM (Vapnik, 1998)作為實行機器學習的工具，其統計學習理論可以在合理的時間內解決分群分類問題，亦可進行最佳準確率之特徵值篩選，增加分析之效能，縮短時間成本。

綜合文獻探討發現，在中文文本蘊涵辨識分為兩面向來進行分析：語法特徵、語意特徵，語法特徵在不需要知識背景的條件之下，其所採用之分析方法與英文文本對差異性不高，主要是探討文句相似性以及字串之間之差異。相較於語法特徵，語意特徵方法上，在中文以及英文方面則有顯著之差別，英文詞句直接簡約，而中文涵意表達方式多元，在斷詞、同義詞上所採用之分析方式與英文不致相同。而中文文本蘊涵辨識之研究文獻數量不多，在部分特徵分析上仍有可以補充分析之空間，亦為本研究所需加以深入探討之目的，並且運用機器學習工

具，訓練模型並預測新資料集，透過最佳特徵篩選，提高整體系統之預測準確率，提升系統之效能。

三、系統架構

本研究提出之應用機器學習結合語法與語意特徵於中文文本蘊涵關係系統架構圖如圖 1 所示。

如圖 1 所示，在取得 XML 訓練資料集後，進入前處理的部分。前處理分為兩部分：擷取文本對及 Tokenize，在擷取文本對中首先將 XML 訓練資料集利用程式將文本對一一擷取出來，在 Tokenize 步驟中將文本對每個文句切分成每個單字、片語以及具有意義的單詞，以利進行分析。前處理結束後，進入本系統所設計之各種特徵進行計算。語意特徵的部分，採用哈爾濱工業大學所整理之同義詞詞林，經過改良以後用來進行文句語意之判斷，並於計算完畢後選取準確率較為精準之特徵值並產生出訓練資料格式。再透過 SVM 產生出訓練模型，模型產生後進行一系列的精準度測試(ex： 交叉比對驗證)，使其訓練模型趨於成熟，以利進行測試資料集之預測。

欲進行預測之 XML 測試資料集首先進行前處理，透過本系統所設計之各種語法特徵方式計算之後，產生出測試資料集之檔案格式。將測試資料集利用已訓練完成之訓練模型進行結果之預測，最後呈現出分析後之實驗結果。

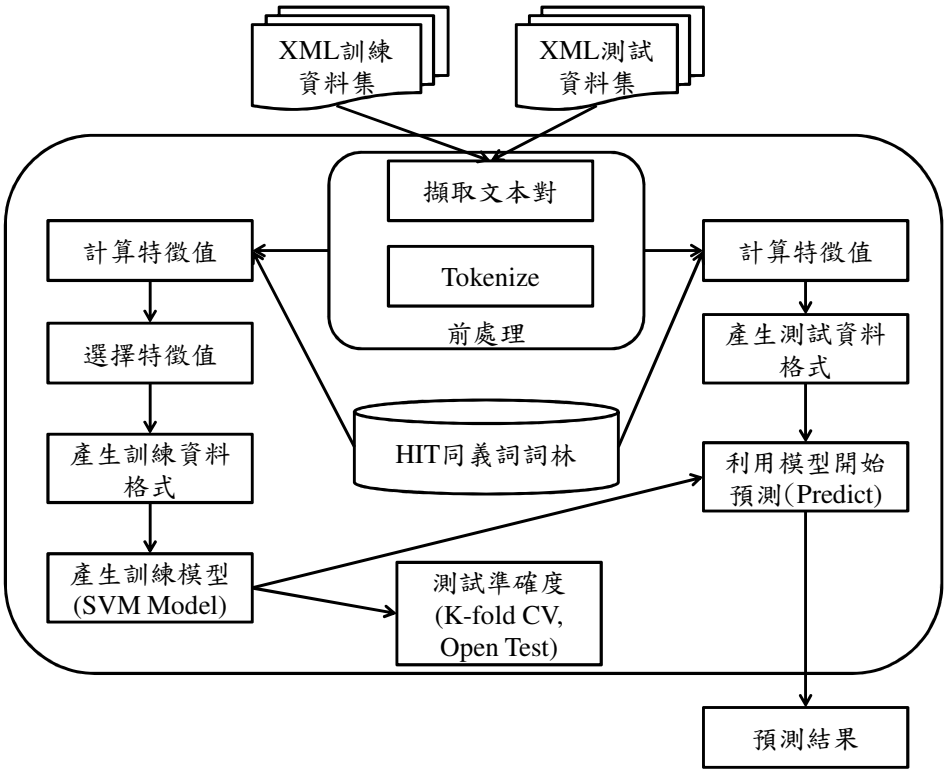

圖 1. 應用機器學習結合語法與語意特徵於中文文本蘊涵關係系統架構圖

3.1 前處理

此步驟為系統重要的環節之一，將各種異質、獨立之資料進行文句篩選、格式統一以及斷詞處理，產生出本系統所需進行研究分析之資料集，以提升資料集進行訓練或預測之效率。

• 處理 XML 檔案

首先我們將 XML 檔案進行基本處理，從中擷取出我們所需要的 ID、文本對，以供分析處理使用。

• 統一資料格式

在我們自然語言的判斷中，很多詞意義其實是相同的，不過因為表示方法有所不同，在進行處理的時候差異往往相距甚遠，例如： 1990 年和一九九零年，前者是數字型態;後者是文字型態，因此資料格式必須統一，來增加處理的準確度。(黃文奇 and 吳世弘, 2011)

• CKIP 斷詞處理

我們系統在文本對上面採用中央研究院 CKIP 中文斷詞系統，來進行句子的斷詞處理，以利進行更深入的分析與處理

3.2 計算特徵值

在資料都進行前處理之後，我們接著要計算特徵值(Features)，這樣便可利用 LibSVM 的工具根據這些特徵值，進行答案的預測並產生模型(Model)。

而為了提高預測的準確性，我們必要進行特徵值的篩選，在經過實驗以後，我們加入語意特徵並配合語法特徵進行比較，選定以下 14 種特徵值，並一一進行詳細介紹：

詞彙相似度(Word Similarity)、字串長度(String Length)、字串長度差值、字串長度比率、最長共同子序列(Longest Common Substring)、Word-Based Edit Distance、Token Length、Token Length 差值、Token Length 比率、Token-Based Edit Distance。

(1) T1 字串長度(T1 String Length)

長度是最簡單也是最好判定的準度依據，很多具有蘊涵關係的文本對都是依照長度較長的文句推斷到另一個較短的文句，在這裡我們使用 T1 字串長度單獨作為一特徵值。

(2) T2 字串長度(T2 String Length)

很多具有蘊涵關係的文本對都是依照長度較長的文句推斷到另一個較短的文句，在這裡我們使用 T2 字串長度單獨作為一特徵值。

(3) 字串長度差值(String Length Difference)

公式： *T1 length-T2 length*

以上述長度作為基礎，再進一步縮小誤差範圍，我們使用兩者長度差來進行衡量。

(4) 字串長度比率(String Length Ratio)

公式： *T1 length/T2 length*

透過比率的方式，我們可以把算出來的值域設定在小範圍之間，藉此來提高判斷的準度。

(5) 最長共同子序列(Longest Common Subsequence)

本研究採用最長共同子序列方法來找出兩文本對的相似程度(Daniel S.,1977)，其公式如下：

$$LCS(X_{1...i},Y_{1...j})=\begin{Bmatrix}\phi & if\ i=0\ or\ j=0\\ LCS(X_{1...i-1},Y_{1...j-1})+x_i & if\ x_i=y_i\\ \max(LCS(X_{1...i},Y_{1...j-1}),LCS(X_{1...i-1},Y_{1..j})) & else\end{Bmatrix}$$

找出兩條句子當中相同的字元最多有幾個，相同的字元個數事實上就是一種相似度的衡量標準。這種衡量標準不考慮插入，刪除以及取代。也就是說插入，刪除，取代的操作並不會減少相似度的值，而每個相同的字元都會將相似度增加 1。我們利用此方法更加提升準確度的判斷。

(6) Word-Based Edit Distance

定義：是由一個句子轉換成另一個句子所需的最少編輯次數

例如：

T1： 我喜歡打球

T2： 我討厭打球

例句中，將喜轉成討，歡轉成厭，經過了兩次轉換，故 Edit Distance 為 2。

(7) T1 Token Length

我們將一個句子進行了 CKIP 斷詞處理之後並計算每個詞(Token)的長度作為判定準度的依據，在這裡我們使用 T1 Token 長度單獨作為一特徵值。例如：

T1： 二零零零年(N) 奧運(N) 在(P) 雪梨(N) 舉辦(Vt)

(8) T2 Token Length

為了比對出兩句差異性，在這裡我們使用 T2 Token 長度單獨作為一特徵值。

(9) Token Length 差值(Token Length Difference)

定義： *T1 Token Length- T2 Token Length*

用兩文句的斷詞後長度的差值來進行比對，跟上述長度差值一樣。

(10) Token Length 比率(Token Length Ratio)

定義： *T1 Token Length/T2 Token Length*

用兩文句斷詞後的長度的比率來進行誤差範圍的縮小，提升準確率

(11) Token-Based Edit Distance

定義：是由一個句子轉換成另一個句子所需的最少編輯次數，但是是以 Token 來進行轉換

例如：

T1： 我(N) 喜歡(Vt) 打(Vt) 球(N)

T2： 我(N) 討厭(Vt) 打(Vt) 球(N)

例句中，將喜歡轉成討厭，經過了 1 次轉換，故 Edit Distance 為 1。

(12) 名詞數量(Noun Number)

一個句子內，如果我們事先計算出了名詞數量，就可以事先排除掉部分例外情形，因此區別出詞性數量也能作為分析的依據。

(13) 動詞數量(Verb Number)

一個句子內，如果我們事先計算出了動詞數量，就可以事先排除掉部分例外情形，因此區別出詞性數量也能作為分析的依據。

(14) 詞彙語意相似度(Word Semantic (Synonym) Similarity)

採用哈爾濱工業大學所整理的同義詞詞林，每個單詞具有 ID，而 ID 相同的單詞彼此具有同義詞關係，例子如下表示:

Di01A01= 世界 世 世上 大地 天下 天底下 全世界 環球 全球 舉世 中外 寰宇 五洲 海內 海內外 五湖四海 大千世界 大世界 普天之下

但若採用此格式去辨識文句中的同義詞關係，方法略為複雜及麻煩，因此本研究改良了以上查詢同義詞的方式，以計算同義詞之間相似程度的方法來進行查詢，公式如下:

TYCCL Scoring Function: $((\tau-\rho)+1)/\tau$

τ:同義詞數量 ρ:詞彙在同義詞林中的排序

以"世界"這詞為例子，如果與"世界"有關的同義詞有 19 個

Di01A01= 世界 世界、世、世上、大地、天下、天底下、全世界、

環球、全球、舉世、中外、寰宇、五洲、海內、海內外、五湖四海、大千世界、大世界、普天之下

而”世界”這詞在”世界”的同義詞中，相似度的排名是第一，則公式表示為: (19-1)+1/19 = 19/19 = 1，表示”世界”這詞在”世界”同義詞中的相似度的值為 1，相似度最高。

經過公式計算後所整理之同義詞詞林的表例如下:

世界 *Di01A01=|世界:1.0000,Di14C04=|世風:0.5000,Dd05B03=|領域:0.3333*

因此在計算詞彙語意相似度時，相似度越高，則文句之間越具有蘊涵關係。我們將每對文句進行 CKIP 斷詞過後，每個詞於整理後的同義詞詞林進行查詢，找出相關之同義詞，並計算相似度，來彌補於語法特徵值上兩種不同單詞單純視為不同詞彙的誤差。

Example:

T1： 車諾比病毒在 1999 年 4 月總共造成超過 200 萬台電腦無法開機

T2：1999 年 4 月車諾比病毒總共造成逾 200 萬台電腦無法開機

如下表 1 所示，如果單就語法特徵進行判斷的話，因為 T1 句子長度大於 T2 句子長度，其結果為正向(Forward)關係，但是加入語意特徵(同義詞)來判斷的話，"逾"與"超過"兩者詞意為同義詞關係，最後結果為雙向(Binary)關係，符合正確答案。

表 1、採用語意特徵及未採用語意特徵之結果比較

Non-semantic features output	**With semantic features output**
Result: **Forward**	Result: **Binary**
T1 String Length:30	T1 String Length:30
T2 String Length:28	T2 String Length:28
T1_T2 length Difference:2	T1_T2 length Difference:2
T1_T2 ratio:1.0714	T1_T2 ratio:1.0714
LCS:22	LCS:22
T1 Token Length:13	T1 Token Length:13
T2 Token Length:12	T2 Token Length:12
T1_T2 Token Length Ratio:1.083	T1_T2 Token Length Ratio:1.083
T1_T2 Token Length Difference:1	T1_T2 Token Length Difference:1
Edit Difference: 13	Edit Difference: 13
Edit Token Distance: 6	Edit Token Distance: 6
Noun Number Difference: 0	Noun Number Difference: 0
Verb Number Difference: 0	Verb Number Difference: 0
	Word Semantic (Synonym) Similarity: 12.6042

3.3 機器學習

本研究使用機器學習工具-libSVM(Chang and Lin, 2011)來進行資料訓練以及預測。

本研究的步驟如下：

1. 將要計算的特徵值轉成為 libSVM 的格式
2. 轉換出來的格式進行資料的訓練
3. 利用訓練的資料建立出模型
4. 利用模型進行預測答案以及測試準確率

而在測試正確率的部分又可以分為兩種測試方式：

・開放測試(Open Test)

將訓練資料拿去訓練產生出模型，用另外需要進行預測的資料給所產生的模型進行測試，以此來驗證模型是否正常以及準確。

・交叉驗證(K-Fold Cross Validation)

將訓練資料分成為 K 等分，將其中 K-1 分拿去產生出模型，而預測剩下的 1 份，再次依序拿 K-1 分去繼續驗證另外的 1 分，以此來測試模組的準確性。

利用 libSVM 的工具： grid.py 與 fselect.py

在 libSVM 工具中，訓練時參數 cost(c)和 gamma(g)很重要，會影響準確率。簡易步驟為： 先將原始訓練資料先做簡單的縮放，再使用交叉驗證去找最好的 cost 和 gamma 值，用這些值去做訓練。

四、 實驗結果與討論

我們利用 NTCIR-9 RITE 所提供的 RITE1_CT_dev_mc.txt 之 421 筆訓練資料集，結合本系統所設計之特徵值組合進行訓練產生出訓練模型，並進行 NTCIR-9 RITE 另行提供的 RITE1_CT_test_mc.txt 之 900 筆測試資料集之結果預測分析，並產生各種特徵值所計算之準確率如表 2 所示：

從表 2 結果得知，在 CV 表現最佳的是 Token Length 差值，而在 Open Test 裡面表現最佳的是字串長度比較，這兩者結果共同之處都是差值。實驗結果顯示，在判斷兩文本對之間的關係時，如果事先透過兩文本對長度進行比對分析的話，準確率可達到近乎 70%，也就是說，題目敘述的長短關係到文本對彼此的推論關係。此外，從表 2 結果得知，如果單測量詞彙語意相似度此一特徵值的結果，所呈現之準確率並無顯著表現，略較其他特徵值之準確率低。

由於 Feature 組合數目太過於龐大，而另外為了進行有無採用語意特徵之比較，因此，我們挑選了 6 種組合來進行測試。從表 3、4 與圖 2、3 的實驗結果顯示，組態 1 至 3 以未採用語意特徵為對照組，組態 4 至 6 為實驗組，來進行預測之準確率的比較。從結果來看，對照組對於實驗組在 CV 上，其準確率較為高，但 Open Test 中，進行未知資料集之預測的準確率上，實驗組有採用語意特徵在 Open Test 的準確率明顯高於對照組之準確率。

表 2. 單一特徵值之 CV 及 Open Test 結果(Dev_421)

特徵值 ID	特徵值	Cross Validation (BC)	Open Test (BC)
Feature01	T1 String Length	58.43%	60.11%
Feature02	T2 String Length	64.13%	61.44%
Feature03	String Length Difference	68.65%	**68.56%**
Feature04	String Length Ratio	69.12%	67.11%
Feature05	Longest Common Substring	59.86%	60.67%
Feature06	Word-Based Edit Distance	60.09%	60.00%
Feature07	T1 Token Length	58.91%	60.44%
Feature08	T2 Token Length	62.93%	61.67%
Feature09	Token Length Difference	**69.12%**	67.44%
Feature10	Token Length Ratio	66.27%	66.67%
Feature11	Token-Based Edit Distance	57.48%	60.00%
Feature12	Noun Number	65.30%	66.78%
Feature13	Verb Number	58.43%	63.44%
Feature14	Word Semantic Similarity	59.95%	60.00%

表 3. 篩選特徵值(非採用語意特徵)之 CV 及 Open Test 結果

組態 (未加入語意)	特徵值	Cross Validation (BC)	Open Test (BC)
Config1	Feature1~13	**73.63%**	67.78%
Config2	Feature 4~13 (except 10)	72.21%	**68.67%**
Config3	Feature 1~11	72.92%	67.11%

表 4. 篩選特徵值(採用語意特徵)之 CV 及 Open Test 結果

組態 (加入語意)	特徵值	Cross Validation (BC)	Open Test (BC)
Config4	Feature1~14	70.02%	70.33%
Config5	Feature 4~14 (except 10)	69.64%	69.78%
Config6	Feature3,6,9,10,14	**71.23%**	**71.89%**

綜上所述，語意特徵在單一特徵上的表現不如預期，但與其他特徵值搭配組合，意即組態中有採用語意特徵之 Open Test 結果都較未採用語意特徵之 Open Test 結果的準確率還高。而加入詞彙語意相似度這特徵值所訓練出來的模型在預測未知資料集普遍都有 70%左右之預測準確率。

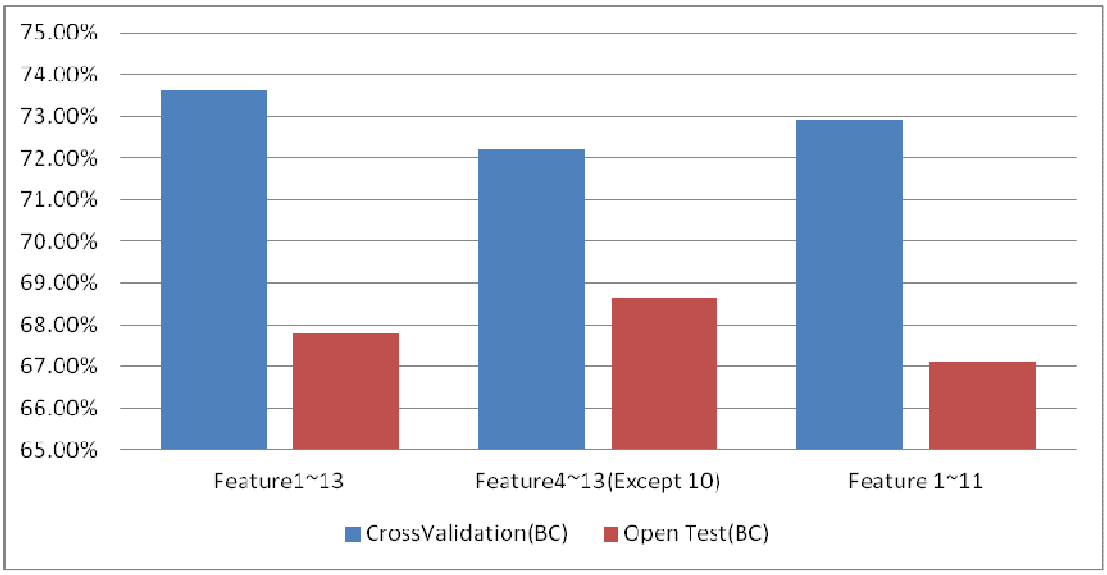

圖 2. 非採用語意特徵之 CV 與 Open Test 結果

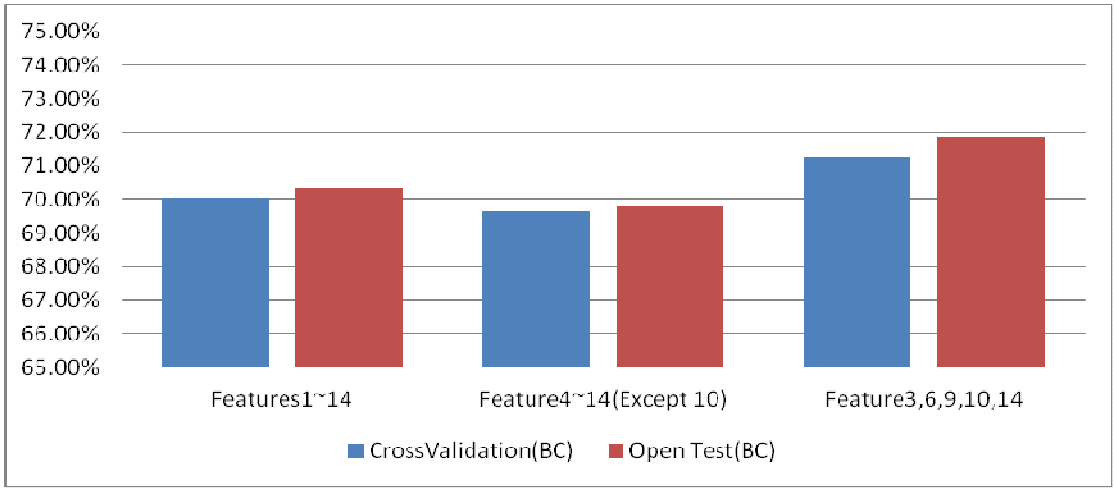

圖 3. 採用語意特徵之 CV 與 Open Test 結果

在進行資料集之訓練中，參數的選擇必須要納入考量，因此如何選擇最佳參數是非常重要的，藉由 Grid.py 便可進行最佳參數之選擇。Grid.py 是在進行機器學習利用模型預測資料集以後，計算出最佳參數的程式，以便提升模型預測之準確率。模型參數調整完畢後，再來就是訓練特徵值之選擇，透過不同特徵值之組合，所產生之結果準確率也都不盡相同，因此如何找出最佳特徵值之組合也必須納入考量，在 Libsvm 中利用 fselect.py 程式便可進行最佳特徵值之選擇，更進一步提升系統預測之準確率。

從表 5 得知，在進行 grid.py 找出最佳化參數並透過 fselect.py 選取最佳的特徵值組合後，未採用語意特徵之準確率由 68.67%提升至 72.65%; 而採用語意特徵之準確率由 71.89%提升至 73.28%。

綜上所述，利用 fselect.py 及 grid.py 在提升準確率上有顯著之效果，而有採用語意特徵之準確率比未採用語意特徵之準確率還高。

表 5. 使用 grid.py 以及 fselect.py 之後的 CV 及 Open Test 之結果

特徵值	Cross Validation (BC)	Open Test (BC)
Feature 4~13(Except10) (未含語意特徵)	**75.29%**	72.65%
Feature 3,6,9,10,14 (含語意特徵)	73.20%	**73.28%**

五、 結論

本研究目的主要利用語法特徵進行文本蘊涵關係之分析，加入語意特徵方法後，其準確率較未加入語意特徵來得準確。本研究發現，在進行特徵值之篩選後，有採用語意特徵值，其準確率都較未採用語意特徵值高，進而表示出兩文本對之間在詞意判斷上之推論關係是具有正面影響。

本研究實驗結果顯示，利用本系統所提出的語法特徵為基礎之機器學習方法，配合參數調整及最佳方法之篩選，在中文文本蘊涵辨識整體準確率達到 73.28%。

本研究主要貢獻為採用需透過背景知識所達成的方法之下，透過兩文本對之間的詞意關係，可以提升判斷之準確率。而單用語法特徵來進行分析其準確率不及採用語意特徵方法後來得高。

本研究所使用的特徵值主要加入語意特徵為主，判斷兩文句中詞彙的同義詞相似度，並與單方面採用語法特徵進行比較。而在未來研究方向中，語意特徵的部分若能加入更多工具(WordNet、HowNet 等等)進行分析，預期能夠產生更為成熟之訓練資料集模型，再進行測試資料集之預測步驟時，準確率預期能夠相對提升，以利增進本系統訓練模型之分析效率並改善本系統所設計之特徵方法之效能。

六、 誌謝

This research was supported in part by the National Science Council of Taiwan under Grants NSC 101-3113-P-032-001 and TKU research grant. We would like to thank the support of IASL, IIS, Academia Sinica, Taiwan.

七、 參考文獻

Siblini, R., and Kosseim, L.”Using Ontology Alignment for TAC RTE Challenge”. Proceedings of the Text Analysis Conference, Gaithersburg, MD, 2008, pp. 1-7.

Burchardt, A., Reiter, N., Thater, S., and Frank, A., “A semantic approach to textual entailment： System evaluation and task analysis.”, Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic, ACL, 2007, pp. 10-15

Iftene, A. and Moruz, MA., “UAIC participation at RTE5”., Proceedings of the Text

Analysis Conference, Gaithersburg, MD., 2008.

Wee, M., S., Hwee, T., Ng. and Chung., Y. L., “A Machine Learning Approach to Coreference Resolution of Noun Phrase”, MIT Press, 2001.

Vanderwende, L., Coughlin, D., and Dolan, B., “What Syntax can Contribute in Entailment Task”., Microsoft Research, 2006.

Callison-Burch, C., Koehn, P., and Osborne, M., “Improved Statistical machine translation using paraphrases.”., Proceedings of the HLT Conf. of the NAACL, pp. 17-24, New York,NY., 2006.

Castillo,J.,J., “A Machine Learning Approach for Recognizing Textual Entailment in Spainish”, 2010.

Kouylekov, M. and Magnini, B., “Recognizing textual entailment with tree edit distance algorithms. ”, Proceedings of the PASCAL Recognizing Textual Entailment Challenge, 2005.

Malakasiotis, P. and Androutsopoulos, I. , “Learning textual entailment using SVMs and string similarity measures.”., Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Prague. ACL, 2007, pp. 42-47

Vapnik, V “Statistical learning theory”. John Wiley, 1998.

Daniel S. Hirschberg, “Algorithms for the Longest Common Subsequence Problem”, Journal of the Assocrauon for Computing Machinery(24：4), October 1997, pp. 664-675.

TAC, http://www.nist.gov/tac/2010/RTE/RTE6_Main_NoveltyDetection_Task_Guidelines.pdf, 2011.10.27

NTCIR RITE, http：//artigas.lti.cs.cmu.edu/rite/Main_Page_(TC), 2011.10.27

Min-Hsiang Li, Shih-Hung Wu, Ping-che Yang and Tsun Ku,“Chinese Characters Conversion System based on Lookup Table and Language Model”, Proceedings of the 22nd Conference on Computational Linguistics and Speech Processing (ROCLING 2010), pages 113-127, Nantou,Taiwan, September 2010.

Yujie Zhang and Kazuhide Yamamoto, “Paraphrasing spoken Chinese using a paraphrase corpus”, The Journal of Natural Language Engineering, Volume 11, No. 4, pages 417–434, December, 2005

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM： A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2, 3, Article 27 (May 2011), 27 pages.

黃文奇 and 吳世弘, “中文文字蘊涵系統之特徵分析”, Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing (ROCLING 2011), 2011