# 中英文的文字蘊涵與閱讀測驗的初步探索

# An Exploration of Textual Entailment and Reading Comprehension for Chinese and English

黃瑋杰 林柏誠 劉昭麟
Wei-Jie Huang Po-Cheng Lin Chao-Lin Liu
國立政治大學資訊科學系
Department of Computer Science, National Chengchi University
{100753014, 101753028, chaolin}@nccu.edu.tw

## 摘要

文字蘊涵是研究文字敘述之間的邏輯關係的工作，本文利用詞彙、語法、詞彙語意的相關語文資訊，建構經驗法則公式與機器學習模型，檢驗自動推測文字蘊涵關係的效果。本文所報告的效果在NTCIR-10的RITE競賽中的簡體與繁體中文的文字蘊涵都有相當好的表現。我們同時延伸文字蘊涵的推論技術，企圖以語文處理技術自動回答國小及國中的中英文閱讀測驗試題，這一部份的工作仍在發展之中，對於比較簡易的四選一的試題，如果相關的基礎技術成熟，可以達到超過五成的答對率。

Research on text entailment studies the logical relationships between statements. We employed linguistic information at the lexical, syntactic, and semantic levels to build heuristics and machine-learning based models for algorithmic judgment of text entailment relationships. Methods proposed in this paper achieved relatively very good performances in the RITE task for both traditional and simplified Chinese entailment problems in NTCIR-10. We extended our work and attempted to automatically answer questions in reading comprehension tests in Chinese and English used in elementary and middle schools. To make the automatic answering more feasible, we manually selected statements which were relevant to the test items before we ran the text entailment component. Experimental results indicated that it was then possible to find the answers better than 50% of the time for one out of four multiple-choice items.

關鍵詞：文字蘊涵、經驗法則公式、機器學習模型

## 1 緒論

在自然語言處理的領域中，讓電腦能夠理解人類使用的語言，進而帶給人類便利的生活，是該領域的研究者一直追求的目標，其中文字蘊涵 Textual Entailment(TE)便是一個相當重要的議題，藉由文字蘊涵的技術可以延伸到很多應用方面，例如在問答系統、信息抽取、閱讀理解等等都有很大的益助，而所謂的文字蘊涵就是讓電腦自動判斷兩個句子是否具有推導的關係，在文字蘊涵的框架中，我們將句對個別以文本($T_1$)和假設($T_2$)作為分別，以下面的句對為例，文本即可以推導至假設，因為假設所擁有的資訊都包含於文本內。同時，我們也利用文字蘊涵的技術應用在閱讀測驗的自動答題上，如果可以判別閱讀測驗的選項與本文具有推論的關係，則間接可以判別該選項為答案的機率較大，讓系統能夠自動答題。

*文本:日本時間 2011 年 3 日 11 日，日本宮城縣發生芮氏規模 9.0 強震，造死傷失蹤約 3 萬多人。*
*假設:日本時間 2011 年 3 日 11 日，日本宮城縣發生芮氏規模 9.0 強震。*

Recognizing Textual Entailment(RTE)[2]和 Recognizing Inference in Text(RITE)[8]則為目前為文字蘊涵所舉辦的相關競賽，該比賽將句對分類為 Yes 或 No 兩種推論的結果；以下面這組句對為例，『尼泊爾毛派叛亂份子在新國王大壽前夕發動攻擊』與『尼泊爾毛派叛亂份子在新國王華誕前夕發動攻擊』，前句與後句差別於「大壽」與「華誕」，但兩句的含義是相同的，因此我們期待系統判別該句對有推論的關係，並得到 Yes 的推論結果。

我們在判斷句子的推論關係上分為兩個做法作為判別的依據；第一個方法是使用經驗法則式的推論模型，該模型將可能會影響到文字蘊涵的特徵資訊擷取下來，並利用加減分的機制，將之形成一個計算公式，例如我們認為當兩個句子的詞彙覆蓋[1]比例夠高，某方面也代表著句對間具有相同的資訊量，因此在公式中，詞彙覆蓋的比例就以加分的方式來處理；而句對間的否定詞數量如果不一樣，句子的含義也可能大相逕庭，因此當否定詞的數量不同時系統則以減分的方式處理，藉由這些特徵的加減分計算最後我們可以判別所得的分數是否有超過推論的門檻值，再以此作為判別推論的依據。

第二個方法是使用機器學習的方式，除了藉由第一個方法所蒐集到的特徵資訊，我們也將剖析樹(Parse Trees)、POSes(Parts-Of-Speech)動詞標記和詞彙依賴關係(Word Dependency) [7]做為訓練模型的特徵集合，並採用三種不同的分類演算法訓練分類模型，分別是支持向量機(Support Vector Machines, SVMs)[5]、決策樹(Decision Trees)與線性回歸(Linear Regression)[3]，透過不同類型的分類器獲得推論關係的結果。

我們利用以上述建構的推論模型參加 NTCIR-10[10]國際資訊評估競賽，在文本蘊涵 RITE 簡體中文與繁體中文兩個分項獲得第二名。其中作為繁體中文及簡體中文推論評分標準的 Macro-F1 分別為 67.07%和 68.09%。

本篇論文於第二節介紹關於句子蘊涵的相關競賽，第三節介紹經驗法則式推論模型以及我們所蒐集認為對文字蘊涵有幫助的語文特徵資訊，並於第四節呈現實驗的結果和結論；第五節介紹機器學習的方法包含蒐集新的特徵、特徵的擷取；第六節則是實測我們的演算法的實驗結果。第七節我們將前面所建構的經驗法則式推論模型和機器學習模型應用在閱讀理解的應用上，第八節則呈現閱讀理解實驗的結果及結論，最後第九節為結論以及未來展望。

## 2 Textual Entailment 背景資訊

### 2.1 相關競賽

RTE 是基於英文語料對語句推論的相關競賽，從 2005 年開始，由 First Recognition Textual Entailment(RTE-1)所舉辦的第一次比賽，並針對英文語句推論提供評估的平台，使得句子的推論關係逐漸受到重視，而隨後 RTE 的競賽也增加了許多關於語意推論的相關應用，例如 Question Answering(QA)、Information Retrieval(IR)、multi-document Summarization 等等。

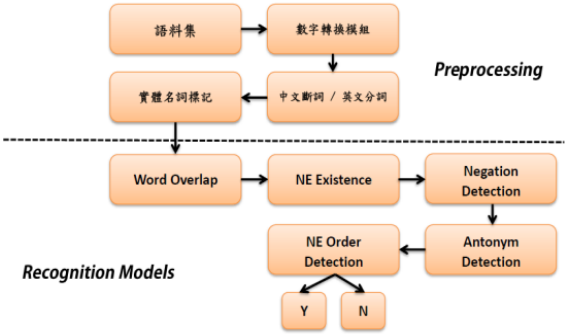
RITE 則是 NTCIR(NACSIS Test Collections for IR)國際資訊檢索評估競賽的其中一項子任務，與 RTE 不同的是，Recognizing Inference in Text (RITE-1) 競賽開始針對中文語句推論的研究議題提供評估的平台，目的是為了讓中文母語使用者也能專注到此議題上。

### 2.2 文獻探討

在 RITE-2 的競賽中，我們發現多數的隊伍在研究文字蘊涵時，都有使用詞彙的覆蓋比率與句子表面相似度[4]作為判別文字蘊涵的重要特徵，然而僅僅這些方法並不足以判別文字的蘊涵關係，因此某些方法如 Wu[6]所提出的 LCS Similarity 用來判別$T_1$及$T_2$句對的最長相同字串，當作判別蘊含的依據，或是 Hattori[4]利用句子表面相似度和句意相似度的高低，組合成一個 2x2 的矩陣作為判別的策略，因此可以進一步的分析 2x2 四種情況的組合會在什麼情況下發生，例如當表面相似度很高但句意相似度卻很低時，可以猜想句對中可能有不同數量的否定詞存在；我們參考 RITE-1 競賽中具有高效能的方法並搭配我們自己的方法，建構出判別文字蘊涵的模型。

## 3 經驗法則式推論模型與特徵介紹

經驗法則式(Heuristics)推論模型的系統架構與運行流程如圖一所示，首先將語料讀入系統後，透過數字轉換模組將數字正規化，接著進行中文斷詞或英文分詞[7]，並標記實體名詞[10]與解析句法結構，最後通過我們提出的計算方法與門檻值設定，計算推論關係的評分，由 0 至 1，並根據門檻值獲得欲判斷的文字蘊涵關係，而詳細的特徵我們將在 3.1 節至 3.5 節作介紹。




圖一、經驗法則式推論系統架構與流程

### 3.1 詞彙覆蓋比例

在評估一個句子的意義是否能推論至另一個句子時，我們認為句子中每一個詞彙都代表一項資訊，當兩個句子裡相同的詞彙比例夠高時，通常代表這兩個句子擁有相同的資訊量，因此具有推論的關係。

我們以 $T_1$ 和 $T_2$ 分別作為句對斷詞或分詞後的詞彙集合，其中以 $T_2$ 作為文本，計算兩個句子的詞彙重疊比例，如下方公式(1)，$T_1$ 和 $T_2$ 分別為兩個句子斷詞或分詞後的詞彙集合，透過該公式計算兩個句子間相同詞彙的比例，以 0 至 1 表示相同比例的高低。

(1)

但公式(1)需要詞彙完全相同才會納入計算，如此一來可能會漏掉部分的縮寫詞彙、同義詞彙或因為各種原因被斷詞器斷開的情況，因此我們修改了公式(1)，加入詞彙部分相同的計算，也把近義詞的判斷[12][15]加入到公式(1)的修改，使之成為公式(2)

$\epsilon T \qquad \epsilon T$

(2)

### 3.2 實體名詞判斷

如果只使用詞彙的覆蓋比例來表示句子間的推論關係，我們僅能掌握句子表面的資訊含量，而無法了解句子所表達的內容，因此透過實體名詞標記，將句子中的人名、地名和組織名擷取出來，並把這些標記出來的詞彙視為重要的資訊，將有助於判別句子間的推論關係。

我們將上述的假設加入一個函數，調整推論關係的計算，如公式(3)，$NE_{t2}$ 為 $t_2$ 中擷取出的實體名詞，其中 t 為 T 句子經由斷詞或分詞後的集合；$f_{NEPenalty}$ 會判斷 $NE_{t2}$ 中的元素會被包含於 $t_1$ 與否，當有元素不包含在 $t_1$ 時，則給予一次範圍 0 至 1 的 α 懲罰分數。因此推論關係的判斷加入該函式，變成公式(4)。

α, (3)

(4)

### 3.3 否定詞判斷

即使兩個句子擁有高比例的詞彙覆蓋和實體名稱相同，但句子間常因為存在否定詞而使句意大為改變，進而造成錯誤的推論判斷，因此我們增加系統對否定詞的擷取，並設計簡單的規則判斷否定詞對計算推論關係的影響，所謂的否定詞我們以否定詞辭典作為依據，例如辭典中：「無」、「未」、「不」、「沒有」…視為否定詞，並藉由句子中的否定詞集，適當地調整推論關係的評分。

我們認為兩個句子若包含不同數量的否定詞時，較容易有不同的意義產生，而降低推論關係的可能性，因此再度加入一個函式針對否定詞做推論分數的調整，如下方公式(5)所示。Negation 表示句子當中包含的否定詞集合，β 為否定詞數量不相等時用以調整的懲罰分數，其值介於 0 至 1，並將推論關係的判斷延伸成公式(6)。

β, (5)

(6)

### 3.4 反義詞判斷

除了否定詞外，句子之間若存在反義詞[12]，我們認為這樣是更加顯示兩個句子之間可能不具有推論的關係，因此我們嘗試分析句子之間的反義詞包含狀況，若包含反義詞，則給予較重的懲罰分數，大幅調整推論關係的判斷。公式(7)顯示反義詞判斷的函式，*Antonym* 表示一個詞彙的反義詞集合，γ則是反義詞存在時的懲罰分數，其值為1至2，而判斷推論關係的公式則變成公式(8)。

(7)

(8)

### 3.5 實體名詞錯位

主詞與受詞位置可能影響句子的語意，因此我們在前處理便標記出實體名詞的索引，並且我們認為當推論分數較高時，代表句子之間的詞彙使用非常相近，此時若實體名詞發生錯位，則較容易影響兩個句子語意的相似程度，如圖二，因此增加一個函式判斷索引值的迥異，藉以調整推論關係的評分，如公式(9)。公式中 *i* 代表實體名詞於句子中的位置，m 和 n 為 *NE_Order* 的索引值，δ 為範圍1到2的懲罰分數，λ為使用該函式的推論分數門檻值。透過上述的各種語言資訊的使用，最後合併成一項推論關係的計算公式(11)，將推論關係的程度以 0 至 1 的分數顯示高低，我們預期該方法能有效地判定語句間的推論關係。

$t_1$：**台灣**出口至**印度**成長 28.6%
$t_2$：**印度**從**台灣**出口成長率可達 28.6%
[台灣：0, 印度：1
[台灣：1, 印度：0

圖二、實體名詞位置比對範例

(9)

(10)

(11)

## 4 經驗法則式推論模型實測

### 4.1 實驗語料

我們經由參與 NTCIR 的競賽，取得 RITE 的訓練(Dev.)與測試(Test)中文語料集，語料為推論關係二元分類(Binary Classification)。圖三為中文二元分類的資料內容，每筆資料皆有一個編號記錄，並包含兩個句子－$t_1$ 與 $t_2$，而 label 代表的是 $t_1$ 的內容是否能推論出 $t_2$ 中的假設，Y 表示成立，N 則反之。我們取得了和 NTCIR-10 RITE-2 的訓練與測試語料，表一為訓練與測試語料集的數量統計。

英文語料我們則採用 Microsoft Research Paraphrase Corpus(MSR Corpus)[12]，MSR 於 2004 年由 Quirk 等人提出，語料集共包含 5801 個英文句對，並且標記兩個句子之間是否相關聯。

```
<pair id="4" label="N">
  <t1>思科公司是全球最大的網路供應公司</t1>
  <t2>微軟是全球最大軟體公司</t2>
</pair>
```

圖三、二元分類資料集

表一、中文訓練語料集統計

| 來源 | NTCIR-10 RITE-2 | | MSR | |
|---|---|---|---|---|
| 語言 | 繁體中文 | | 英文 | |
| 類別 | Dev. | Test | Dev. | Test |
| Y | 716 | 479 | 2753 | 1147 |
| N | 605 | 402 | 1323 | 578 |
| 總和 | 1321 | 881 | 4076 | 1725 |

### 4.2 推論模型門檻值與特徵參數選定

為了最佳化推論系統的效果，我們透過 RITE-2、MSR 及 RTE 三種不同的訓練語料從實驗裡設定所有參數組合藉由效能的變化以人工的方式設定參數，調整中英文推論模型的各項參數與門檻值以尋求準確率的極大值，藉以分析參數組合對於單項推論的效果，所謂的單項推論即是在判斷文字蘊涵關係時，僅判斷具有蘊涵關係或不具有蘊涵關係兩種；最後我們以準確率較佳的參數設定針對測試語料進行推論系統的評估，不過礙於版面限制，本篇論文中文語料只節錄 RITE-2 繁體語料作為代表，而英文語料則以 MSR 作為代表，其它詳細的實驗結果可參照黃瑋杰碩士論文[13]。

表四列出繁體中文訓練語料的參數搜尋結果，由於搜尋的結果過多，因此在這裡僅列出較佳的幾組參數設定與訓練語料的準確率，其中編號 E 代表推論成立的門檻值。而表五則列出英文訓練語料－MSR 的參數搜尋結果，同樣地僅列出較佳的幾組設定與準確率，我們將準確率(Acc)與 Macro-F1 定義如下公式。

$$\text{準確率}(\quad \frac{\text{推論結果正確個數}}{\text{語料個數}} \qquad \text{精確率}(\quad \frac{\text{推論結果單項正確個數}}{\text{推論結果單項個數}}$$

召回率( $\frac{\text{推論結果單項正確個數}}{\text{參考答案中的單項個數}}$ ———

Macro-F1= ————

表四、RITE-2 繁體中文訓練語料參數設定

| 編號 | *E* | *α* | *β* | *γ* | *λ* | δ | Acc |
|---|---|---|---|---|---|---|---|
| C1 | 0.54 | 0.1 | 0.27 | 1.8 | 0.85 | 1.9 | 73.05% |
| C2 | 0.56 | 0.08 | 0.25 | 1.0 | 0.85 | 1.8 | 73.13% |
| C3 | 0.56 | 0.08 | 0.25 | 1.7 | 0.85 | 1.8 | 73.20% |

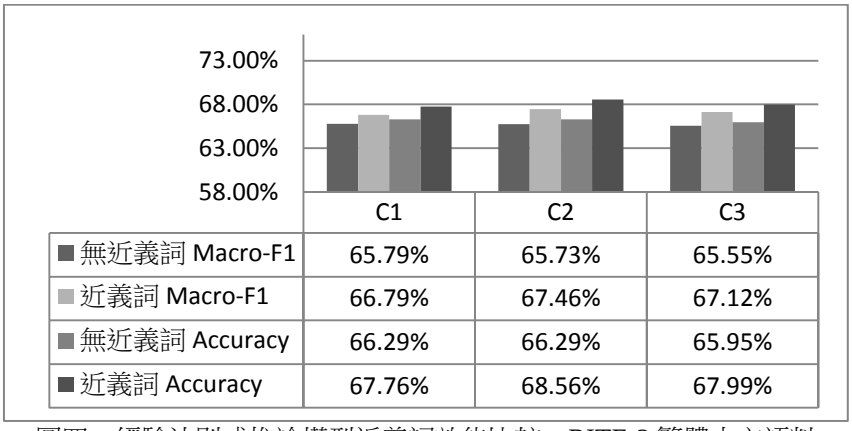
表五、MSR 訓練語料參數設定

| 編號 | *E* | *α* | *β* | *γ* | *λ* | δ | Acc |
|---|---|---|---|---|---|---|---|
| C13 | 0.47 | 0.05 | 0.13 | 1.3 | 0.55 | 1.2 | 71.07% |
| C14 | 0.47 | 0.05 | 0.17 | 1.3 | 0.55 | 1.0 | 71.12% |
| C15 | 0.49 | 0.05 | 0.14 | 1.2 | 0.55 | 1.0 | 71.15% |
| C16 | 0.49 | 0.05 | 0.17 | 1.2 | 0.55 | 1.0 | 71.17% |
| C17 | 0.49 | 0.05 | 0.20 | 1.2 | 0.55 | 1.0 | 71.20% |

### 4.3 實測結果

根據上述這些訓練語料的參數調整，進行測試語料的實驗，分析經驗法則式推論模型經由參數調校後的效能與單項推論能力。

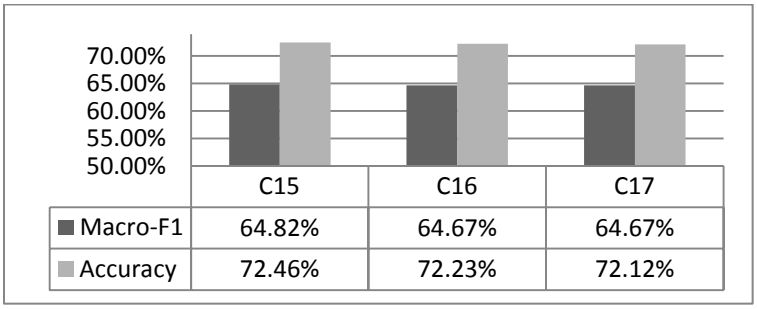
我們使用表四的參數進行 RITE-2 繁體中文測試語料的推論關係預測，並且加入近義詞的判定，觀察是否能提升推論效果，最後針對預測的結果進行分析，計算單項答案的準確率與召回率。圖四則為 RITE-2 繁體中文測試語料使用近義詞的效能比較，從圖中的結果顯示近義詞在 RITE-2 的測試語料中能提升不少系統效能，而我們也有對 RITE-1 測試語料進行實驗其結果則是略微的下降，礙於版面所以省略其結果，因此我們認為近義詞在推論關係的判斷是否具有幫助，因語料特性的不同而有所差異。




圖四、經驗法則式推論模型近義詞效能比較：RITE-2 繁體中文語料

最後透過相同的推論模型，使用 MSR 英文訓練語料的參數設定對語料預測推論結果，藉以瞭解相同的語言模型是否可以套用在不同的語料中的推論關係判斷，圖五顯示測

試語料透過經驗法則式推論模型的系統效能綜合指標。我們觀察 MSR 測試語料的實驗結果，從表五可以看出實體名詞錯位的懲罰參數 δ，C15 至 C17 皆為最低的懲罰分數 1.0，所以可得知該特徵對於推論關係的影響不大，因此也間接對否定推論關係判定較差的情形發生，但仍能達到不錯的準確率。



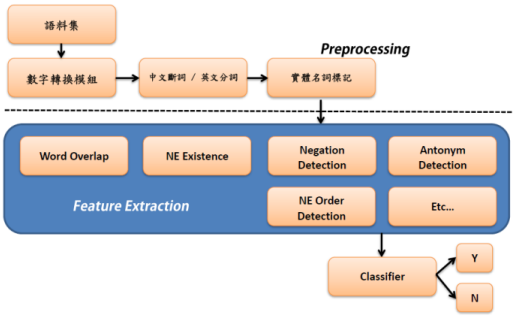

| | C15 | C16 | C17 |
|---|---|---|---|
| Macro-F1 | 64.82% | 64.67% | 64.67% |
| Accuracy | 72.46% | 72.23% | 72.12% |

圖五、 經驗法則式推論模型系統效能：MSR 測試語料

經由多組中文與英文語料實驗，可以發現我們提出的函式組成經驗法則式推論系統與 NTCIR-9、NTCIR-10 競賽成績相比，在中文語料中仍屬於不錯的效果。英文的實驗結果則仍有進步空間，兩種推論能力都需要就現有的函式進行改善，以提升英文語句的推論效果。從這些實驗可以得知未來我們需要發展更多函式來判定否定的推論關係，尤其是針對語句間的反義、獨立與矛盾等現象需要處理。

## 5 機器學習方法

機器學習演算法建構的推論模型系統架構如圖六所示，同樣使用上一節的元件進行前處理，接著擷取我們認為可以增加推論效果的語文資訊，做為訓練模型的特徵集合；最後我們採用三種不同的分類演算法訓練分類模型，分別是支持向量機(Support Vector Machines, SVMs)、Weka J48 決策樹(J48 Decision Trees)與 Weka 線性回歸(Linear Regression)[13]，透過不同類型的分類器獲得推論關係的結果。




圖六、機器學習推論系統架構

前一小節說明了經驗法則式推論模型所使用的函式，我們針對這些函式進行數值化的轉換，做為訓練推論模型的特徵；這些特徵包含詞彙覆蓋比例、實體名詞數量、實體名詞相似度、實體名詞錯位數量、句子長度、否定詞數量、近義詞數量、反義詞數量等項目。除此之外，我們希望加深推論模型對語法結構的認識，因此加入剖析樹分析、POSes 動詞標記與詞彙依賴關係等元素，計算其相似度做為特徵，希望提高推論模型的能力。

### 5.1 剖析樹分析

我們透過史丹佛剖析器(Stanford Parser)[9]取得句子的剖析樹，並且我們認為使用整個剖析樹分析句法結構相似度容易增加計算的難度，因為句子之間可能僅有部分的結構具有共通性即可具備推論的關係，因此以每一個節點做為根節點(ROOT)擷取其下層節點形成的子樹，使用這些子樹來計算兩個句子結構的相似程度。

### 5.2 POSes 動詞標記

POSes 標記由史丹佛剖析器獲得，我們認為動詞在句子中扮演較重要的角色，因其指出整個句子的事件與動作意圖，因此特意將被標註成動詞的詞彙抓取出來，以兩個句子個別的動詞數量與相似度做為特徵[15]，並期望讓分類器學習動詞使用在推論關係上的影響力。

**原句：**1997 年香港回歸中國

| | 1997 | 年 | 香港 | 回歸 | 中國 | ROOT |
|---|---|---|---|---|---|---|
| 1997 | 0 | 0 | 0 | 1 | 0 | 0 |
| 年 | 0 | 0 | 0 | 1 | 0 | 0 |
| 香港 | 0 | 0 | 0 | 1 | 0 | 0 |
| 回歸 | 0 | 0 | 0 | 0 | 0 | 1 |
| 中國 | 0 | 0 | 0 | 1 | 0 | 0 |
| ROOT | 0 | 0 | 0 | 0 | 0 | 0 |

圖七、詞彙依賴關係矩陣 M

**原句：**1997 年香港回歸中國

| | 1997 | 年 | 香港 | 回歸 | 中國 | ROOT |
|---|---|---|---|---|---|---|
| 1997 | 0 | 0 | 0 | 1 | 0 | 1 |
| 年 | 0 | 0 | 0 | 1 | 0 | 1 |
| 香港 | 0 | 0 | 0 | 1 | 0 | 1 |
| 回歸 | 0 | 0 | 0 | 0 | 0 | 1 |
| 中國 | 0 | 0 | 0 | 1 | 0 | 1 |
| ROOT | 0 | 0 | 0 | 0 | 0 | 0 |

圖八、經過五步的詞彙依賴關係，M

### 5.3 詞彙依賴關係

史丹佛剖析器亦能根據剖析樹的生成，產生詞彙之間依賴的關係(Stanford Dependencies)，我們將依賴關係中的詞彙做為節點，將句子中的詞彙關係視為一個有向圖(Directed Graph)，並化做矩陣形式如圖七。

我們發現一個矩陣內可以顯示的資訊並不充沛，如此稀疏的矩陣中，我們難以找到句子之間包含相同關係的詞彙組合，因此以相鄰矩陣(Adjacency Matrix)的概念做進一步的運算；例如一個矩陣 M，可以經由矩陣相乘獲得節點到節點之間移動所需要的步數，因此計算 $M^3$ 便能瞭解任一個節點過程經由兩個節點，所與其他節點的間接依賴關係。我們將這樣的移動視為依賴關係的延伸，如此能找出更多潛在的詞彙依賴關係，並且將不同移動步數的矩陣結果聯集，獲得更豐富的依賴關係。圖八便是圖七的矩陣計算任一個節點經由四個以內的節點所形成的直接或間接依賴關係表，我們透過這樣的矩陣，分析句子之間詞彙依賴關係相似的程度，並以該數值做為一項特徵。

## 6 機器學習方法實測

### 6.1 實驗語料與設計

我們依照經驗法則式推論模型所使用的語言資訊抽取特徵，並提出如剖析樹結構及詞彙依賴關係等語法結構特徵，希望增加推論關係的分類能力。接著以 SVM、J48 和線性回歸等演算法訓練分類模型，並以貪婪式搜尋各個語料的特徵組合與其分類效果，最後經由挑選出來的特徵組合進行分類演算法評比，再以指定的分

表六、中文特徵集編號表

| | |
|---|---|
| **F1** | **F2** |
| **F3** | **F4** |
| **F5** | **F6** |
| **F7** | **F8** |
| **F9** | **F10** |
| **F11** | **F12** |
| **F13** | **F14** |
| **F15** | **F16** |
| **F17** | |

類演算法進行中英文測試語料的效能評估與指定特徵對推論關係判斷的效能比較。不過礙於版面限制，本篇論文中文語料只節錄 RITE-2 繁體語料作為代表，而英文語料則以 MSR 作為代表，其它詳細的實驗結果可參照黃瑋杰碩士論文[13]。

表七、英文特徵集編號表

| | |
|---|---|
| **E1** | **E2** |
| **E3** | **E4** |
| **E5** | **E6** |
| **E7** | **E8** |
| **E9** | **E10** |
| **E11** | **E12** |
| **E13** | **E14** |

為了瞭解各種特徵組合的分類效果，我們採用貪婪式的特徵組合搜尋，測試訓練語料中所有的特徵組合，由 LibSVM 與 Weka 將訓練語料自動切為十個等分(10-fold)，在 SVM 及 J48 演算法的分類下進行循環估計(Cross-Validation)，找尋準確率極大值的特徵組合，而線性回歸則再次使用訓練語料做為評估語料，設定門檻值為 0.5 找尋準確率最大值，最後將獲得的特徵組合進行測試語料的實驗評估。表六與表七具有編號形式的中英文特徵集。

### 6.2 特徵選取

接著展開三種分類演算法在各種語料的特徵組合搜尋。我們由三種分類演算法的結果中搜尋各種語料中準確率較佳的特徵組合，表八表九顯示在不同語料與分類演算法中獲得較佳準確率的特徵組合，我們將透過這些特徵組合比較三種分類演算法在推論關係判斷上的效果。

表八、RITE-2 繁體中文訓練語料特徵組合搜尋

| SVM | | |
|---|---|---|
| 編號 | 特徵組合編號 | Accuracy |
| M1 | F1, F2, F3, F4, F5, F6, F8, F9, F12, F14 | 71.99% |
| J48 | | |
| M2 | F1, F2, F3, F5, F7, F8, F12, F13, F15 | 71.78% |
| 線性回歸 | | |
| M3 | F1,F3,F4,F5,F6,F7,F8,F9,F10,F11,F12,F13,F14,F15,F16,F17 | 72.98% |

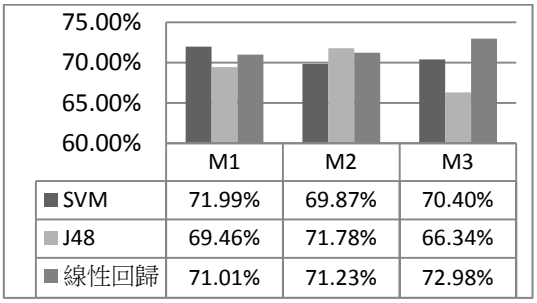
表九、MSR 訓練語料特徵組合搜尋

| SVM | | |
|---|---|---|
| 編號 | 特徵組合編號 | Accuracy |
| M4 | E1, E6, E9, E12 | 70.93% |
| J48 | | |
| M5 | E1, E6, E8, E10, E12, E14 | 71.82% |
| 線性回歸 | | |
| M6 | E1,E2,E3,E4,E5,E6,E7,E9,E10,E11,E12,E13,E14 | 72.45% |

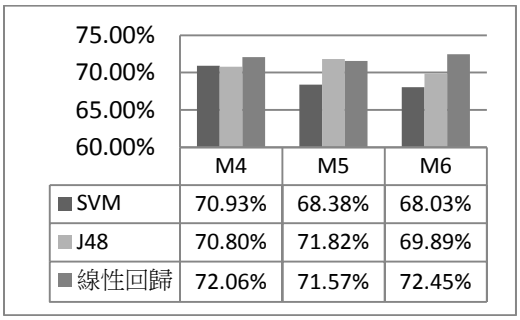
### 6.3 實驗設計、演算法與參數的選定和結果

為了瞭解三種分類器在推論關係判斷上的效果，我們根據上一小節獲得的特徵組合，透過 SVM、J48 及線性回歸等演算法進行分類器的效能評估，SVM 與 J48 演算法以十等分的循環估計準確值為評估指標，線性回歸演算法則再以訓練語料測試，設定門檻值為 0.5 對推論關係分類，評估其準確值。我們將依據各種分類模型在訓練語料的效果，在不同類型的語料中採用指定的分類演算法進行推論關係的分類。

圖九和圖十分別為 RITE-2 繁體中文及 MSR 訓練語料在不同分類模型下，以準確率較佳的特徵組合進行推論關係的分類結果，M1 至 M3 可參照表八，為繁體中文的特徵組合；M4 至 M6 可參照表九為 MSR 英文語料的特徵組合；從繁體中文與 MSR 兩種語料的結果觀察，使用線性回歸演算法進行推論關係分類時，平均上都能獲得較佳的準確率，即使在 SVM 及 J48 分類模型能獲得最高準確率的特徵組合，透過線性回歸演算法的使用，相較於兩種演算法的最高準確率僅有些微的下跌，仍能達到不錯的效果




圖九、分類模型準確率比較：RITE-2 繁體中文訓練語料



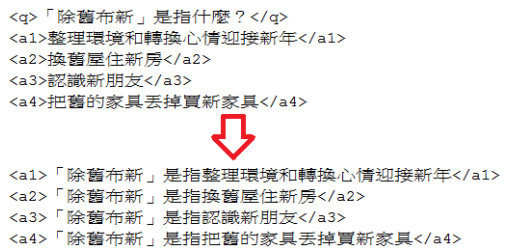

圖十、分類模型準確率比較：MSR 訓練語料

## 7 閱讀理解的實驗準備

本節將介紹文字蘊涵在閱讀測驗中的應用，藉由前面節次判別文字蘊涵關係所建構的模型，作為推論閱讀測驗答案的依據；我們在 7.1 與 7.2 小節介紹實驗的前處理。希望透過文字蘊涵在閱讀測驗中的應用，未來可以將此應用推廣至實務的教育資訊系統。

### 7.1 問題轉直述句

在前面實驗所使用的推論系統中，所有的語料都是以兩個直述句進行推論關係的判斷，而在閱讀測驗中，為了直接提升推論關係的效果，我們也將問句及選項通過人工的方式轉換成四個直述句，再採用推論系統進行短文與四個直述句的推論關係判定，如圖十一為中文閱讀測驗直述句轉換的例子。




圖十一、直述句轉換範例

### 7.2 從短文篩選相關句

除了將問題及選項轉換為直述句來進行推論關係的判斷之外，一篇短文中可能同時敘述相當多種的事實與動作，因此每一道問題的背後往往都僅有利用到短文中部分的陳述句子來回答。

為了瞭解經由短文內容挑選適當的句子後，對指定問題回答的推論效果，我們首先採用人工的方式進行短文的過濾，依據題組中每一道問題，對短文採取過濾，挑選其中與此道問題相關的句子，形成一個較小的句子集合來對問題及選項的組合判斷推論關係。

我們希望先透過人工過濾的形式，進行部分實驗來驗證這樣的工作具有一定成效，接著再發展相關的自動化技術與方法，如判定短文與問題的關連性、中心詞彙或關鍵字搜尋，藉以提昇推論系統在閱讀測驗中的效能。

## 8 閱讀測驗答題實測

我們透過上述所建構的經驗法則式推論模型和機器學習模型分別對中英文閱讀測驗進行答題效能的評估，並介紹語料來源、實驗設計及呈現實驗結果。

表十、閱讀測驗語料集統計

| 中文閱讀測驗 | | 英文閱讀測驗 | |
|---|---|---|---|
| 年級 | 數量 | | 數量 |
| 國小一年級 | 21 | 國中一年級 | 260 |
| 國小三年級 | 39 | 國中二年級 | 468 |
| 國小四年級 | 40 | 國中三年級 | 498 |
| 國小五年級 | 44 | | |
| 國小六年級 | 86 | | |

### 8.1 實驗語料的來源、數量

我們蒐集中英文的閱讀測驗語料集，中文的部分以國小孩童閱讀測驗為主，英文則蒐集國中的閱讀測驗，並且我們依照年級將語料分類，相關的統計如表十，語料內容都以一篇短文與數則題目組成，每一道題目都包含一個問題與四個選項，僅有國小三年級的中文語料屬於三個選項，並且每一道題目的答案都為單一選項。

### 8.2 實驗設計、語料的使用方式

在閱讀測驗的實驗中，我們採用前面兩種不同的推論系統進行效能評估，並將語料採用不同的方式進行人工轉換或過濾，以嘗試此方法在閱讀測驗中的效果。

表十一、閱讀測驗實驗參數設定

| 語言 | $E$ | $\alpha$ | $\beta$ | $\gamma$ | $\lambda$ | $\delta$ |
|---|---|---|---|---|---|---|
| 中文 | 0.57 | 0.28 | 0.24 | 2.0 | 0.85 | 2.0 |
| 英文 | 0.47 | 0.0 | 0.26 | 1.3 | 0.6 | 1.2 |

我們將語料分為三種類別，原始語料、問句重組及短文過濾，並分別採用兩種推論系統－經驗法則式推論模型與機器學習分類模型，判斷閱讀測驗中最佳的回答選項。在經驗法則式推論模型中，我們以各個選項通過計算後的推論分數為評量指標，選取其中分數最高者為該問題的最佳答案；而機器學習分類模型則由 SVM 演算法，輸出其推論關係的機率值，以選項中機率值最高的做為答案，此外我們在中文的部分也加入線性回歸演算法的推論關係判斷，以數值最高的選項做為答案。

經驗法則式推論模型的參數設定，中文的部份我們採用 NTCIR-10 RITE-2 競賽時的最佳設定，而英文則是從實驗中藉由效能的變化以人工的方式設定參數，如表十一所示。機器學習的分類模型則由 RITE-2 繁體中文訓練語料及 MSR 英文訓練語料，選取適當的特徵訓練分類模型，接著進行閱讀測驗中短文與每一個選項的推論關係判斷。表十二顯示中文閱讀測驗採用 SVM 演算法的特徵集，表十三為使用線性回歸之特徵集，表十四為英文之特徵集。

表十二、中文閱讀測驗特徵集 – SVM

| | |
|---|---|
| | |
| | |
| | |
| | |
| | |

表十三、中文閱讀測驗特徵集 – 線性回歸

| | |
|---|---|
| | |
| | |
| | |
| | |
| | |
| | |

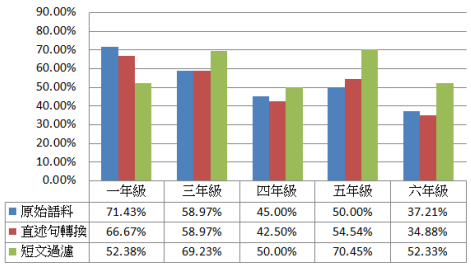
表十四、英文閱讀測驗特徵集 – SVM

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |

## 8.3 實驗結果

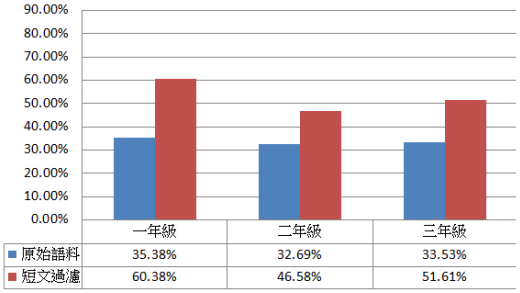
首先採用經驗法則式推論模型對中英文閱讀測驗進行實驗，所使用的參數如上一節所示，我們依序對原始語料、直述句轉換與短文過濾的三種形式語料進行閱讀測驗的推論關係判斷。

圖十二與圖十三分別為中文及英文閱讀測驗的效能圖表，從各年級的結果顯示，在四選一的閱讀測驗中，我們的推論系統即使在高年級的語料中，仍可以獲得約 37%的效果，而在套用適當的方法後，由中文的結果可以發現，短文過濾對於閱讀測驗中推論關係的判斷是較有幫助的，除了一年級的語料外，都顯示了此方法有助於推論系統正確回答閱讀測驗的問題；而直述句轉換則較不如我們預期有較多的進步幅度，僅在四年級有些微的進步。而一年級語料並未在短文過濾中發揮功效，我們認為和語料的數量具有相當的關係，一年級語料的數量非常稀少，因此我們認為這樣的效果並無法有效顯示真正在小學一年級閱讀測驗的形式與測驗設計，需要更多的語料來驗證我們提出的方法。




| | 一年級 | 三年級 | 四年級 | 五年級 | 六年級 |
|---|---|---|---|---|---|
| 原始語料 | 71.43% | 58.97% | 45.00% | 50.00% | 37.21% |
| 直述句轉換 | 66.67% | 58.97% | 42.50% | 54.54% | 34.88% |
| 短文過濾 | 52.38% | 69.23% | 50.00% | 70.45% | 52.33% |

圖十二、中文閱讀測驗準確率-經驗法則式推論模型

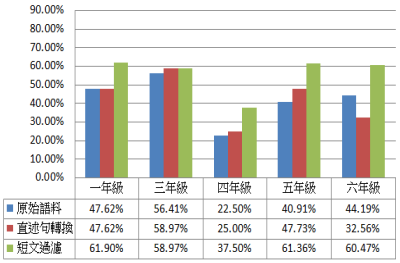



| | 一年級 | 二年級 | 三年級 |
|---|---|---|---|
| 原始語料 | 35.38% | 32.69% | 33.53% |
| 短文過濾 | 60.38% | 46.58% | 51.61% |

圖十三、英文閱讀測驗準確率 – 經驗法則式推論模型

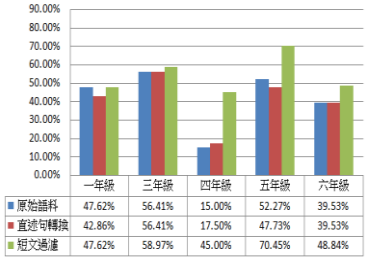
接著觀察圖十三，英文語料採用短文過濾的方法來進行實驗比較，如同中文閱讀測驗的效果，原始的英文語料透過經驗法則式推論模型都能達到 30%以上的基本效能，而採用短文過濾後，則大約都能提升十到二十個百分點，說明短文過濾在增強推論系統判斷閱讀測驗答案時具有良好的功效，未來可以針對此部分發展自動化的處理方法過濾短文。

接著使用機器學習演算法訓練分類模型判斷閱讀測驗中每個選項的推論關係，在中文閱讀測驗，我們以上一小節的特徵集，採用 SVM 及線性回歸兩種演算法做推論關係的分類。圖十四及圖十五為中文閱讀測驗的效能比較，由圖表觀察得知，短文過濾在閱讀測驗中判斷推論關係是一項非常有效用的步驟。然而使用機器學習分類模型的閱讀測驗效果則不如經驗法則式推論模型來的有效果。



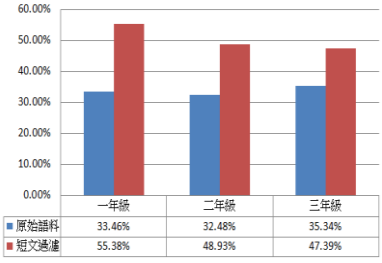

| | 一年級 | 三年級 | 四年級 | 五年級 | 六年級 |
|---|---|---|---|---|---|
| 原始語料 | 47.62% | 56.41% | 22.50% | 40.91% | 44.19% |
| 直述句轉換 | 47.62% | 58.97% | 25.00% | 47.73% | 32.56% |
| 短文過濾 | 61.90% | 58.97% | 37.50% | 61.36% | 60.47% |

圖十四、中文閱讀測驗準確率-SVM




| | 一年級 | 三年級 | 四年級 | 五年級 | 六年級 |
|---|---|---|---|---|---|
| 原始語料 | 47.62% | 56.41% | 15.00% | 52.27% | 39.53% |
| 直述句轉換 | 42.86% | 56.41% | 17.50% | 47.73% | 39.53% |
| 短文過濾 | 47.62% | 58.97% | 45.00% | 70.45% | 48.84% |

圖十五、中文閱讀測驗準確率-線性回歸




| | 一年級 | 二年級 | 三年級 |
|---|---|---|---|
| 原始語料 | 33.46% | 32.48% | 35.34% |
| 短文過濾 | 55.38% | 48.93% | 47.39% |

圖十六、英文閱讀測驗準確率-SVM

而英文閱讀測驗中，我們僅使用 SVM 演算法進行部分的實驗，並僅採用短文過濾的方法對閱讀測驗文本前處理，從圖十六的結果可以發現經由短文過濾後，閱讀測驗的回答準確率在不同年級語料中都能獲得約十五到二十個百分點的進步，是個相當不錯的效能，在四選一個閱讀測驗中可以獲得 50%左右的準確率。

## 9 結論

本研究利用會影響文字蘊涵的特徵資訊，建構經驗法則式模型用以判別 RITE、RTE、MSR 語料的文字蘊涵關係，也採用機器學習的方法透過 SVM、J48 及線性回歸等演算法進行效能評估，最後並利用前面建構好的推論系統應用於閱讀測驗的自動答題上面，在經驗法則式模型的方法上，中文和英文語料的準確率分別可達 68.56%和 72.23%；採用機器學習線性回歸的方法，中文和英文語料的準確率分別可達 72.98%和 72.54%；而基於上述建構好的推論系統作閱讀測驗的自動答題，在四選一的閱讀測驗中也可以獲得約 50%的準確率。

我們提出的推論系統與 NTCIR-9、NTCIR-10 競賽成績相比，在中文語料中仍屬於不錯的效果。英文的結果則仍有的進步空間，我們認為語料的不同語言特性，足以影響推論關係的準確率，因為某些特徵可能只對部分的語料有效，而在閱讀理解的部分，在問題轉直述句和篩選相關句的部分目前仍是以人工處理，其中在篩選相關句的部分就足以讓準確率上升十幾個百分點，我們希望未來能夠自動化的完成閱讀測驗前處理的部分，並針對閱讀理解的部分再找出有用的語言特徵藉以提升答題的準確率。

## 致謝

本研究承蒙國家科學委員會研究計劃 NSC-101-2221-E-004-018-與國立政治大學頂尖大學計畫 102H-36 的部份資助，僅此致謝。我們感謝評審對於本文的各項指正與指導，限於篇幅因此不能在本文中全面交代相關細節。

## 參考文獻

[1] Rod Adams, "Textual Entailment Through Extended Lexical Overlap", *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pp. 128-133, 2006.

[2] Ido Dagon, Oren Glickman and Bernardo Magnini, "The PASCAL Recognising Textual Entailment Challenge", *Machine Learning Challenges*. Lecture Notes in Computer Science, 3944, pp. 177-190, Springer, 2006.

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, 11(1), 2009.

[4] Shohei Hattori, Satoshi Sato, "Team SKL's Strategy and Experience in RITE2", *Proceedings of NTCIR-10 Workshop Meeting*, pp. 435-442, 2013.

[5] Chih-Wei Hsu, Chih-Chung Chang and Chih Jen Lin, *A Practical Guide to Support Vector Classification.* Retrieved from website: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2010.

[6] Han Ren, Hongmial Wu, Chen Lv, Donghong Ji, and Jing Wan, "The WHUTE System in NTCIR-10 RITE Task", *Proceedings of NTCIR-10 Workshop Meeting*, pp. 560-565, 2013.

[7] Chris Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA: May 1999.

[8] Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi and Koichi Takeda, "Overview of NTCIR-9 RITE: Recognizing Inference in Text," *Proceedings of NTCIR-9 Workshop Meeting*, pp. 291-301, 2011.

[9] Stanford Parser, http://nlp.stanford.edu/software/lex-parser.shtml

[10] Stanford Named Entity Recognizer, http://www-nlp.stanford.edu/software/CRF-NER.shtml

[11] NTCIR(NII Test Collection for IR Systems) Project
http://research.nii.ac.jp/ntcir/index-en.html

[12] Microsoft Research Paraphrase Corpus,
http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

[13] WEKA, http://www.cs.waikato.ac.nz/ml/weka/

[14] 劉群、李素建，基於《知網》的辭彙語義相似度計算，*中文計算語言學期刊*，7(2)，頁 59-76，2002。

[15] 黃瑋杰，*中英文語句語意推論*，國立政治大學資訊科學系碩士論文，2013。
http://thesis.lib.nccu.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=jkFqMR/search#result