

構建一個中文國小數學文字問題語料庫 Building a Corpus for Developing the Chinese Elementary School Math Word Problem Solver

繆慎耘 Shen-Yun Miao
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
jackymiu@iis.sinica.edu.tw

林素朱 Su-Chu Lin
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
jess@iis.sinica.edu.tw

馬偉雲 Wei-Yun Ma
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
ma@iis.sinica.edu.tw

蘇克毅 Keh-Yih Su
中央研究院資訊科學研究所
Institute of Information Science
Academia Sinica
kysu@iis.sinica.edu.tw

摘要

本篇論文提出了一個理想數學文字問題語料庫所應具備的特性，並敘述我們如何建置一個完善的中文國小數學文字問題語料庫：其過程、所標註的內容、以及遭遇的困難。求解數學文字問題是自然語言理解在人工智慧領域的一個常見應用。近年來有關於英文數學文字問題的研究，都是以機器學習的方法為主。但不論是執行機器學習，或比較不同的解題系統，都需要一個完善的語料庫。目前國際上雖然已有數個公開的英文數學文字問題題庫，但他們所收錄的數學題庫多偏頗於自家解題系統的特性，難以公平評估各系統的真實性能。而在中文方面，迄今尚無任何一個公開的數學文字問題題庫可資使用。

為了能有一個可供研究使用的數學文字問題語料庫，並修正前述英文題庫之缺失，我們建立了一個頗具規模的中文國小數學文字問題語料庫。此語料庫原始資料是由台灣三個國小教科書及參考書的出版社所授權提供。為完整提供所有的數學文字問題題型及

內容，我們僅僅清除了無關解題的垃圾符號，並不增刪原始內容，以完整呈現文字敘述之原貌。此外，我們還加註了額外的訊息，並在部分題目上標註題型和語言分析，以助研究者瞭解數學文字問題之特性及建立模型。

本語料庫是第一個完整的中文國小數學文字問題語料庫。依各個年級分類，含有數種補充資訊，便於評估各解題系統的解題能力；亦可讓使用者按個別應用需求，塑造不同組態的特殊語料庫，將有助於從事相關研究者之工作。

關鍵詞：數學文字問題，語料庫，標註

1. 緒論

求解數學文字問題(Math Word Problem)[1-4]，基於以下的原因[3, 4]常被選作研究自然語言理解的測試案例：(1) 數學文字問題的答案，無法單純地藉由實行關鍵字或特徵配對被擷取（如傳統的問答（Q&A）系統），因此可以清楚地顯示出理解和推理的優勢。(2) 與其他領域相比，數學文字問題通常不具有那麼複雜的語法（如人文社會領域），並僅需要少量的領域知識（與物理化學領域相較），因此研究人員可以著重於自然語言理解和推理的任務上。(3) 數學文字問題的“主幹”部分（即描述問題給定資訊的部分），通常只由少數句子組成，因此理解和推理過程能被程式快速執行，加速研發進程。(4) 數學文字問題求解器有實際的應用，如小學數學家教，和生活數學助手等。

在 2014 年以前，大多數提出的方法都是基於規則（Rule-based）的[1-2, 35-37]，也就是以人工建構的規則做各種相關決策的判斷。但構建一個高涵蓋度（Coverage Rate）的規則集，是相當困難且所費不貲的工作。此外，在求解歧異問題上也顯得笨拙。因此近來大部分提出的方法[3-7]都是基於統計（Statistic-based）的，也就是其中一些（或全部）的相關決策是透過統計分類進行的。但統計分類器需要先有一個相關的訓練集，才能訓練模型。因為目前數學文字問題求解器以小學程度為主，所以我們必須先構建一個國小數學文字問題語料庫，以便從事相關的研究。

在過去有許多標註或未標註的語料庫被建立。但其文本據我們所知，都沒有針對數學文字問題的；而且它們大部分著重在標註語言學知識，如剖析樹[8]、語意角色[9]、

依存結構[10]、片語結構[11]、語篇(Discourse)[12]、共同指稱關係(Coreference Relations)[13]、和述語參數(Predicate-Argument)[13, 14]等。然而數學文字問題文本的特性，與其他領域的文本特性並不相同(如含有相當多的零指代(Zero Anaphora)等)，而且需要額外的年級分類、數學主題、所需數學運算及答案等其他非語言學的標註。因此若要建立一個數學文字問題解題系統，有必要構建一個對應的數學文字問題語料庫。

目前國際上已有數個英文數學文字問題題庫[2, 5-6, 32-34]可供解題系統比較。它們主要是從美國數學論壇網站，抓取特定的數學題目。然而它們只涵蓋少數數學運算子(即加法，減法，乘法，除法和代數方程)，並且刪除超過他們系統解題能力的題目(或對題目內容加以人為修改，以滿足系統要求)。此外，也沒有對數學題目做年級分層。最後，題庫上沒有標註任何語言學資訊(因此無法參照對應的語言分析)。例如，Kushman等人[5]僅僅抓取代數文字問題的題目；而 Hosseini 等人[2]也只是截取出組合加法、減法、一元一次方程和美元文字問題。最近，Roy 等人[6]則是發佈相對較大的語料庫。他們在題目中增加了需要兩個以上運算子(Multi-step)的算術問題，但刪去了需要背景知識的問題。此外，概數問題也一併被刪掉。Upadhyay 和 Chang[32]、Koncel-Kedziorski 等人[33]、Huang 等人[34]雖然利用爬蟲程式(Crawler)或自動抽取程式個別建立了大型(1,000 題、3,320 題和 18,000 題)的多樣性數學文字問題題庫，更採用最佳化演算法降低大量方程式樣板和辭彙的重複率[33]，可惜都還是以線性代數問題為主，而根本忽略那些線性代數以外的數學問題。

至今雖然有少數研究中文數學文字問題的論文發表[35-37]，但據我們所知，還沒有任何公開、可供研究使用的中文數學文字問題語料庫，因此我們構建了第一個中文國小數學文字問題語料庫。建立這個語料庫的目的，是要提供語料庫以研究數學文字問題文本的特性，以便找出貼切的特徵、提出合適的架構和模型，並提供統計分類器所需要的訓練數據。另一方面，建立這樣一個標註過的語料庫，將簡化其他研究者在數學文字問題上往後的研究工作。此外，藉由使用這個語料庫，我們可以評估系統達到怎麼樣的程度。最後，為了克服上述的現有英文數學文字問題題庫缺陷，這個新的語料庫應該包含所有的國小數學類型並予以分級，並且在部分題目上標註語言學資訊。

與前述的英文數學文字問題題庫[2, 5-6, 32-34]相比，我們的數學文字問題語料庫是取自三個台灣出版社題庫(非從數學論壇網站抓取特定的數學題目)，依據國小學生年

級分別建立（非不分年級混在一起），包含所有數學問題（只過濾掉那些純數字問題），並且註記教育部頒布的能力指標所對應的數學主題。另外，為了讓我們的資料不失真，我們不對任何文句作增刪。因此本語料庫並非針對某一個特定解題系統所構建，可適用於任何小學數學文字問題求解器之研發。最後，我們還在部分題目上標註對應的語言分析，以供設計發展系統的人作為參考。

本論文其餘部分的安排如下：第二章，揭示一個理想的數學文字問題語料庫該如何構建；第三章，敘述建構語料庫的過程及工作；然後在第四章討論我們在建構語料庫時，所遇到的問題和所學到的經驗；第五章，回顧相關的研究工作；最後，結論將撰寫在第六章。

2. 理想數學文字問題語料庫的特性

在設計構建一個數學文字問題語料庫之前，我們必須先確立一些準則，以便在各種設計選項上，能有所取捨。由於求解數學文字問題是自然語言理解在人工智慧上面的應用，因此一個理想的數學文字問題語料庫，我們認為應該具備下列特性（而前述的英文數學文字問題題庫，都或多或少違反這些原則）。

第一，它應包含各種自然語言的描述方式（因為求解系統應該具備相當的自然語言理解能力）。因此我們不應對題目表達方式加以人為修改（即人工簡化句型），而應完整保留所有數學文字問題的現存型式（除了清除亂碼及垃圾資訊外），以便公平比較不同解題系統的自然語言理解能力。

第二，求解數學文字問題是人工智慧的應用。因此一個解題系統應具備處理普通常識的能力。為了忠實的反映各個解題系統之表現，我們不應刪減需要背景知識的問題（或對題目內容加以人為修改），而應完整保留所有的此類數學文字問題，以便衡量不同解題系統在人工智慧的推論能力。

第三，即使是國小數學文字問題，它們已經涵蓋十六種不同的數學運算題型（請參閱 3.3.1 節表三）及需要兩個以上運算子（Multi-step operation）的算術問題。我們不應針對任一系統刪減掉超過他們解題能力的題目，而應完整保留所有出現的數學文字問題，以便測試解題系統的真正程度，並能公平比較不同系統的解題能力。

第四，隨著各個年級不同，數學文字問題的難易程度也不相同。我們不應將他們混合而不加分辨，而應分別標記其對應的各個年級，以便知道各個解題系統到達何種程度。

第五，國小數學文字問題應包括所有出現的數學運算題型。而且每個數學問題應標註其所屬的主題（如時間、空間等）及答案，以方便在研發解題系統時，可以明確知道系統的缺漏並能自動評估正確率。第六，至少在部分數學問題上，應標註題目中的語言

學現象（例如：句法分析、指

代、蘊涵、指稱等）及所對應的數學運算式，以方便在研發解題系統時，研究者可以據以分析建模，並可執行半教導式學習（Semi-supervised Learning）。若一個數學文字問題

語料庫，能滿足上述所列的六個條件，就可用來清楚的評估解

題系統的真正程度（並可與國小各個年級對應）、公平比較不同系統的解題能力、並明確指出各系統的偏頗及缺失。此外，若有個別的應用或規格，亦能讓研發人員按其各自需求，組合出對應的檢索條件，從完整的語料庫中抽取出不同組態的特殊語料庫，便於供各研究系統來評量自己的效能。例如：可按年級分層、依主題分類、依數學題型分類、或組合上述條件等。

3. 語料庫構建

本章節安排如下：3.1 節提及原始題庫的來源與組成；然後我們在 3.2 節討論語料庫預處理的程序；3.3 節揭示標註數學語料庫的過程；最後，3.4 節將敘述對語料庫所做的統計分析。

3.1 原始題庫

我們所拿到的國小數學文字問題原始題庫，是由台灣三個國小教科書和參考書的出版社所提供。這些資料都封裝成電子書軟體、或以電腦檔案型態呈現（如 Microsoft Office Access(.mdb)和 Microsoft Office Word(.doc)檔案），以方便任課教師自由取材，做輔助教學或提供給學生練習。其中包含國語、數學、社會、自然、生活、健康體育等科目，而此資料內容除了題目與題號之外，也附有相對應的學期年級、題型、教育部訂定的能力指標、題目難易度，以及試題屬性。

此三家出版社的原始數學題目資料中，包含了許多類型（指題目呈現的方式），有

些是利用文字表示（例如：是非題、選擇題、應用題、填空題等等），有些則以數學符號或計算式的方式呈現（例如：計算題），另外還有以圖案表達、作答之方式（例如：配合題、畫畫看、量量看、看圖做做看等等），不勝枚舉。本語料庫僅取出以文字表示的數學題目作為我們的語料。表一顯示三種不同的題目類型原始資料的呈現方式。

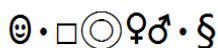
表一、三種特定題型的原始題目資料

題型	文字題目
選擇題	1 隻蜘蛛有 8 隻腳，8 隻蜘蛛共有幾隻腳？（①42 隻②64 隻③74 隻）
應用題	6 包糖果賣 186 元，10 包糖果賣幾元？
填空題	量角器中從刻度 0 的線旋轉到刻度 20 的線，所形成的角是（ ）度。

在表一所示的選擇題與填空題當中，通常含有非文字、數字等與解題無關的符號，因此我們必須做些修正。例如：將選擇題選項部分刪除、把填空題括號改為國字“幾”，並將句號換上問號而成為問句，如下表所示。

表二、處理後的數學文字資料

題型	文字題目
選擇題	1 隻蜘蛛有 8 隻腳，8 隻蜘蛛共有幾隻腳？
填空題	量角器中從刻度 0 的線旋轉到刻度 20 的線，所形成的角是幾度？



(a) 文本中垃圾符號

- Ex01. ♂果農採收了1024個水梨，每16個裝1盒，可以裝成幾盒？
 Ex02. 爸爸每月的薪水是 48000 元，若爸爸把一半的薪水交給媽媽，
 Ⓣ當作交通費，6000 元當作餐費，剩下的存起來，
 則爸爸每個月存的錢占總薪水的幾分之幾？
 Ex03. 1盒蘋果有6個。1§盒蘋果也可以說是幾盒蘋果？

(b) 原始題目範例

圖一、原始文本中的符號問題

一杯汽水有 250 毫升，玉銘喝了 $3\frac{1}{4}$ 杯，美方喝了 800 毫升，誰喝的比較多？

一杯汽水有 250 毫升，玉銘喝了 3 $\frac{1}{4}$ 杯，美方喝了 800 毫升，誰喝的比較多？

(a) 紙本題目原樣

(b) 資料庫題目樣貌

圖二、格式轉換的符號丟失問題

當然，在處理這些題目時，我們還遇到一些不可避免的雜訊問題。例如：文本中帶有許多的垃圾符號（圖一(a)(b)），並且在格式轉換時因符號資訊丟失而導致題目內容表

3.2 構建語料庫的前置作業

如上節所述，在原始數據庫中有許多無用的信息和垃圾符號。由於這些雜訊與數學文字問題解題無關，為避免研發人員耗費無謂的精力，我們先將它們清除，以免干擾後續作業。於是我們使用文字編輯器的規則運算式，來過濾並清理帶有此類問題的數學題目。我們將在下面描述其中所遇到的一些問題。

一杯汽水有 250 毫升，玉銘喝了 $3\frac{1}{4}$ 杯，美方喝了 800 毫升，
誰喝的比較多？

圖三、可讀式中文分數、帶分數的正確敘述

如圖二(b)所示，有些類型的題目，在紙本原樣中帶有圖案的符號，因為格式轉換而造成垃圾符號出現。這些垃圾符號在經過上述清理雜訊的手續後，喪失了數學符號資訊也扭曲了題目的原意。因此我們利用原始資料所提供的答案資訊，來檢查發生此類符號丟失的題目，並且以人工手動方式，對可讀式中文分數、帶分數等題目，恢復其正確的表達敘述（如圖三所示）。

最後，為了系統程式處理方便起見，我們把題目切割為“主幹”部分（乃題目中提供解題所需之資訊；但可能包含額外的不相關訊息）和“問句”部分兩段成分。在一個題目中，“問句”主要是介於問號之前和最後一個逗號之後。而對於多重問句，我們則是將題目以手動方式分開成一個“主幹”和多個“問句”，使不同的“問句”和相同的“主幹”可以分別連接在一起求解。

3.3 語料庫標註

為了建構我們所提出的理想語料庫，我們全文保留了所有三個出版社的國小數學文字問題（不因解題難易或題型做任何增刪修改），並以六個年級的分層方式來建構（每一百個題目製成一個 XML 檔案）。我們將資料庫的所有數學題目轉換成表格型式(Microsoft Office Excel)，然後直接以編輯工具過濾、處理題目、並對每題設定其年級總編號，再轉換成 XML 格式，好讓相關程式方便讀取檔案資料，並可讓人直接透過網頁瀏覽。此外，為著機器學習需要，我們在各年級中以隨機抽取的方式組成三種子集合（即訓練集、

```

<Unit Code="Train-G3-001533" ID="IIS-MR-MATH-GRADE03-001752"> <Unit Code="Train-G3-001534" ID="IIS-MR-MATH-GRADE03-001753">
  <Body>文華有49元，</Body> <Body>每9公分長的繩子可以做一個蝴蝶結，</Body>
  -<QA idx="1"> -<QA idx="1">
    <Question>最多可以買一枝7元的鉛筆幾枝？</Question> <Question>77公分長的繩子最多可以做幾個？</Question>
    <Answer>7枝</Answer> <Answer>8個</Answer>
  </QA> </QA>
</Unit> -<QA idx="2">
  <Question>剩下幾公分？</Question>
  <Answer>5公分</Answer>
</QA>
</Unit>

```

(a) 單問句的數學問題

(b) 多問句的數學問題

圖四、數學問題的 XML 型式

在每一個題目的訊息部分，我們都列出其子集合及年級的資訊，以及其對應的分類流水號。如圖四(a)中第一列“Train-G3-001533”，表示此題是被指定為訓練集中屬於三年級的第 1533 號題目（也就是此題是被儲存在訓練集中，三年級的第 16 個 XML 檔案裡），以方便研究者定位或查詢。另外，每個題目還附有一個年級總編號（即圖四(a)中第一列的 ID）。如此例中“IIS-MR-MATH-GRADE03-001752”，表示此題號是位於全部三年級中第 1752 號題目（此編號為此題在原始題庫的位置，以便對比原始資料作校對）。

在 XML 檔案中，每一個題目均包含了“主幹”、“問句”、和“答案”部分（我們在兩端分別給予“Body”、“Question”和“Answer”的標籤）（如圖四(a)所示）。如 3.2 節所述，每一個問句和一個答案為一組。因此每一組問句和答案皆有一個編號（QA idx，如圖四(b)所示），目的在於求解多重問句題目時，能參考先前問答所提供的解題訊息，以求解下一個問答。以圖四(b)為例，可由第一個問句的求解過程（ $77 \div 9 = 8$ 餘 5）來獲得第二個問句（“剩下幾公分？”）的答案（即“5 公分”）。

每一個題目除包含“主幹”部分和“問句”部分外，還標註了以下訊息：(1)答案的單位，(2)概數和分數問題的標註。這兩項的標註工作，我們仍然利用 Microsoft Office Excel 和文字編輯器工具來幫助完成。具體內容詳述如下。

(1)增加答案的單位：添加答案單位的原因是因為在原始 Access 數據集中，只有數字而沒有其所屬的單位。然而在數學文字中，幾乎每個數量都有其物理單位，因此完整的答案應包含其所屬的單位。這項工作是由一位研究助理和數位兼職標註人員以人工方式來完成。

(2)在概數和分數問題中加上適當的標註：在有關概數的問題中，加上相關的解題

說明，如“（先用四捨五入法...）”等。另外，我們在每個分數題目之後，也加入“（請用帶分數表示答案）”作為題目的補充資訊。此部分標註工作也以文字編輯器工具搜尋包含關鍵字彙（如“大約”、“約”、“*分之*”等）的題目，再由人工方式標註，如下圖所示。

```
-<Unit Code="Train-G5-000352" ID="IIS-MR-MATH-GRADE05-000403">
  <Body>一瓶汽水2公升，平分倒成7杯，</Body>
  -<QA idx="1">
    <Question>每杯約幾公升（用四捨五入法求商到小數第三位）？</Question>
    <Answer>0.286公升</Answer>
  </QA>
</Unit>
```

(a)概數問題

```
-<Unit Code="Train-G4-001066" ID="IIS-MR-MATH-GRADE04-001230">
  <Body>一袋10元錢幣有25個，</Body>
  -<QA idx="1">
    <Question>78個10元錢幣是多少袋（用帶分數表示）？</Question>
    <Answer>3又25分之3袋</Answer>
  </QA>
</Unit>
```

(b)分數問題

圖五、概數與分數的題目標註

3.3.1 題型及語言學標註

為了瞭解文本的性質、執行半教導式學習（Semi-supervised Learning）及測試模型，除了對每一個題目加註基本資訊外（如上節所述），我們還從語料庫當中選擇若干題目進行數學題型及語言學標註（因語料庫標註非常耗時耗力，故本研究僅標記 75 題訓練語料、200 題發展集語料及 200 題測試語料）作為先導研究（Pilot Study）之用，以期幫助機器學習進行題型分類。其中 200 題發展集語料及 200 題測試語料為隨機抽選；而為了涵蓋每一種數學題型，75 題訓練語料則為人工從訓練集中每一類題型抽幾題組成。

表三、十六種題型分佈統計（單位：百分比）

解題類型	訓練集	發展集	測試集
1.幾何	18.39	6.9	4.8
2.乘法	16.09	13.8	21.1
3.除法	16.09	18.8	19.3
4.總合	14.94	8.3	6.1
5.減法	8.05	20.6	17.5
6.比較	4.60	2.8	3.5
7.餘數	4.60	1.8	2.6
8.比/比例	3.45	2.3	2.2

解題類型	訓練集	發展集	測試集
9.代數	3.45	2.8	2.2
10.加法	2.30	14.7	14.0
11.單位轉換	2.30	4.1	2.6
12.最小公倍數	2.30	0	0.9
13.差距	1.15	0.9	1.3
14.最大公因數	1.15	0.5	0.9
15.集合	1.15	1.4	0.4
16.分數	0	0.5	0.4

柏堯買了 6 公斤的豆子，	其中綠豆有 2 公斤，	綠豆占全部豆子的幾分之幾？
<pre>{買(2): agent={柏堯(1)}, aspect={了(3)}, theme={{x1}豆子(6): quantifier={的(5): head={6 公斤(4)}}}}</pre>	<pre>{有(3): topic={{x1}其中(1)}, theme={{x2}綠豆(2)}, range={2 公斤(4)}</pre>	<pre>{占(2): theme={{x2}綠豆(1)}, range={幾分之幾(6): property={的(5): head={{x1}豆子(4): quantifier={全部(3)}}}}</pre>
<p>未知詞</p> <p>柏堯(1,Nba):{human 人:name="柏堯"}</p> <p>買(2,VC31):{buy 買}</p> <p>了(3,Di): aspect={Vachieve 達成}</p> <p>6 公斤(4,DM): 公斤.weight=公 斤:quantity={6}}</p> <p>的(5,DE): relation({entity 事物})</p> <p>豆子(6,Nab):{豆子 beans}</p>	<p>其中(1,Nep):</p> <p>quantifier={definite 定指}</p> <p>綠豆(2,Nab):{豆子 beans:color={green 綠}}</p> <p>有(3,V_2):{exist 存在}</p> <p>2 公斤(4,DM): 公斤.weight=公 斤:quantity={2} 數量詞語意結合</p>	<p>第一部分 句結構相依關係</p> <p>綠豆(1,Nab):{豆子 beans:color={green 綠}}</p> <p>占(2,VJ3):{be 是}</p> <p>全部(3,Neqa):{all 全}</p> <p>豆子(4,Nab):{豆子 beans}</p> <p>的(5,DE): relation({entity 事物})</p> <p>幾分之幾(6,Neqa): quantity={Ques 疑問/Ques 疑問}</p> <p>第二部分 詞彙與語意定義</p>

圖六、國小數學語料的語意表達標記

表三顯示所標註的十六種數學題型（以數學解題方式分類，而非以數據形式分類；並依各題型在訓練集中的百分比排序）。因題中可能有多個問句，因此 75 題訓練集被標註了 87 次、200 題發展集被標註了 218 次、200 題測試語料被標註了 228 次。訓練集中以幾何(18.39%)、除法(16.09%)、乘法(16.09%)與總合(14.94%)比例較多，以分數(0%)最少；發展集中則以減法(20.6%)、除法(18.8%)、加法(14.7%)和乘法(13.8%)比例較多，以最小公倍數(0%)最少；而測試語料中以乘法(21.0%)、除法(19.3%)、減法(17.5%)和加法(14.0%)比例最多，以分數(0.4%)與集合(0.4%)題型比例最少。

對語言學的標註¹，主要包含語法結構標記及語意表達標記。語法結構標記與中研院的中文句結構樹資料庫²[26, 27]標記方式一致，含括詞類、語法結構和語意角色訊息。而語意表達標記則分成兩部分，第一部分採用中研院句結構樹的語義角色以及相依關係，但轉換成廣義知網³[28, 29]所採用的線性語意表達形式，內容包含語意角色、詞彙與詞序，少數中心語省略會在這階段補回。此外，表達式中更附加了共同指稱訊息，在方括號內以 x 與數字標註兩個詞彙有相同的指稱，如圖六中第一句短語和第三句短語出現的「豆子」與第二句短語的「其中」三個詞彙皆指稱相同事物，因此以[x1]共同標記之。第二部分則是將詞彙的語意遵照廣義知網的詞彙定義解歧後，依序分別列於其下；若詞彙為廣義知網中未出現過的未知詞，則會透過語義猜測模組來給予適當定義式。如題目中「柏堯」可自動猜測為人名，並給予定義式；此外，廣義知網的表達優勢之一是

¹ 機器閱讀的語意表達 <http://sunlight.iis.sinica.edu.tw/SC4MR/>

² 中文句結構樹資料庫 <http://TreeBank.sinica.edu.tw>

³ 廣義知網 <http://chownet.iis.sinica.edu.tw/>

可以很方便的進行語義結合的操作與表達。現階段我們的語義結合僅呈現數量詞的結合，如“2 公斤”的定義式：“weight={公斤: quantity={2}}”是由“公斤”的定義式：“weight={公斤}”以及“2”的定義式：“quantity={2}”進行語義結合後的結果。在數學文

字題的語意表達中，數詞與量詞是很重要的環節，會影響後續解題與推論。數詞的值（包含阿拉伯數字、國字或兩者混合的數詞）有整數、小數、分數和疑問詞等形式，如圖六的「幾分之幾」def: quantity={Ques|疑問/Ques|疑問}便是疑問詞與分數結合的表達方式。而中文裡多元而豐富的量詞語義類型[30]與數詞的結合也會形成語意的差別。我們將數量複合詞修飾名詞的語意加以正規化，給予一致性的表達方式，以利數學計算或在不同量詞單位中轉換。下述為幾個數量複合詞的例子：

1 顆蛋 def:{蛋legg: quantifier={顆.null|無義: quantity={1}}}

1 盒蛋 def:{蛋legg: 盒.container={盒子|box: quantity={1}}}

1 打蛋 def:{蛋legg: 打.quantity={12: quantity={1}}}

1 公斤蛋 def:{蛋legg: 公斤.weight={公斤: quantity={1}}}

我們依照量詞不同的語義內涵，給予不同的廣義知網詞彙定義，而數詞和量詞結合後的數量詞語意表達方式[31]基本上與功能詞相同，皆以 relation={value} 形式出現。不同的量詞單位帶來不同的語義，在數學文字題中更放大它的特殊性，事物必須能夠在不同量詞單位中轉換結合的語意，得以正確的理解文字題的題幹。語法結構標記及語意表達標記，是先由程式做猜測，再由標註人員（俱語言學專業背景）做檢查、更正。

3.4 相關統計資料

在前一節中，我們已提到所有的題目被分成六個年級，並且以亂數抽取方式分成三個集合，其中訓練集有 20,093 題，發展集和測試集都各有 1,700 題（包含上節所述之已標註的語料及其他未標註的語料），因此我們國小數學語料庫合計有 23,493 題（如表四所示）。各年級題目數量以三年級 5,210 題為最多，其餘依序為六年級 4,461 題、四年級 4,338 題、五年級 4,062 題、二年級 3,642 題，而以一年級 1,780 題為最少。表五顯示題目“主幹”部分的平均長度為 27 個中文字。使用中研院中文詞庫小組的斷詞和標詞性工具加以斷詞後，每個題目平均含有 18.2 個中文詞。在另一方面，題目“問句”部分的平均長度則為 9.4 個中文字，以及 6.8 個中文詞。

表四、中文數學語料庫統計

語料庫 集合	各年級題目數量						
	一年級	二年級	三年級	四年級	五年級	六年級	合計
訓練集	1,380	3,042	4,610	3,738	3,462	3,861	20,093
發展集	200	300	300	300	300	300	1,700
測試集	200	300	300	300	300	300	1,700
合計	1,780	3,642	5,210	4,338	4,062	4,461	23,493

表五、中文數學題目的平均長度

題目的 兩部分	平均中文字數 (Char.)	平均中文詞數 (Word)
主幹	27	18.2
問句	9.4	6.8

該斷詞工具會以標點符號將題目斷開成數個句子，經過分析後發現：“主幹”與“問句”句長為 5 至 8 個詞的句數比例，分別佔句子總數（79,822 句）的 63% 和 70%，其中“主幹”以 7 個詞的比例為最多（21.8%），“問句”以 6 個詞數為最多（19.4%）（如表六(a) 所示）。此外，由表六(b) 可見，不論“主幹”或“問句”，詞長皆以 2 個字所佔的比例為最多（50.1% 與 64.1%），與一般語料所見相同。

表六、中文數學語料庫斷詞統計表

(a) 句長分析(#79,822 句)			(b) 詞長分析(#608,732 詞)		
句長(詞數)	主幹	問句	詞長(字數)	主幹	問句
1~3	1.0%	0.7%	1	7.7%	14.4%
4	4.9%	6.3%	2	50.1%	64.1%
5	12.7%	16.4%	3	20.4%	17.8%
6	16.8%	19.4%	4	13.3%	3.0%
7	21.8%	19.0%	5	5.3%	0.6%
8	11.6%	15.1%	6 以上	3.2%	0.1%
9	9.3%	9.2%			
10~12	14.8%	11.1%			
13 以上	7.1%	2.8%			

在另一方面，表七顯示詞性標註的分析結果。在“主幹”部分的詞性比例排序前五名依序是普通名詞(Na)佔 29%、數詞定詞(Neu)佔 26.1%、專有名詞(Nb)佔 13.3%、動作及物動詞(VC)佔 6.3%，而地方詞(Nc)佔 5.5%，其他零星詞類比例佔 19.8%；在“問句”部分則以普通名詞(Na)佔 41.2%為最多，其次依序是專有名詞(Nb)佔 11.8%、動作及物動詞(VC)佔 8.7%、數詞定詞(Neu)佔 7.6%、地方詞(Nc)佔 5.2%，而其他零星詞類則佔 25.5%。由於題目中所有解題資訊（含多數名詞、數詞、動作及物動詞）皆由“主幹”提供，而“問句”主要在詢問單一名詞的數量，因此“主幹”的詞類以普通名詞、數詞定詞最多，但“問句”則以普通名詞最多，而數詞定詞偏少。

表七、中文數學語料庫詞性標註統計表

詞類	主幹	問句
普通名詞(Na)	29.0%	41.2%
數詞定詞(Neu)	26.1%	7.6%
專有名詞(Nb)	13.3%	11.8%
動作及物動詞(VC)	6.3%	8.7%
地方詞(Nc)	5.5%	5.2%
其他詞類	19.8%	25.5%
合計	100%	100%

最後，表八顯示教育部能力指標⁴類別（原始題庫所提供）的分佈情形。因為一個題目可能包含多種能力指標，因此在題庫資訊中有單一類別（佔語料庫 80.90%的題目）與組合類別（由單一類別做組合）（佔語料庫 19.10%的題目）之分。

表八、教育部能力指標分類統計表

(a)單一指標類別		(b)組合指標類別	
單一類別	百分比	組合類別	百分比
數與量	72.04%	數與量+代數	13.25%
代數	6.72%	數與量+幾何	4.44%
幾何	1.68%	代數+幾何	0.21%
統計與機率	0.05%	數與量+代數+幾何	1.20%
連結	0.40%	小計	19.10%
小計	80.90%		

表八(a)顯示我們的中文數學語料庫在單一能力指標方面，主要以“數與量”(72.04%)

⁴ 國民教育社群網，九年一貫課程題綱，數學能力指標 http://teach.eje.edu.tw/9CC/fields/math_3_1.php

為最多，其餘數量則低於 10%；在組合指標方面（如表八(b) 所示），主要以“數與量+代數”(13.25%)較多，表示此數學語料著重於建立國小數學的基本概念和基礎運算，但空間幾何、代數、統計機率等進階類別則較少。

4. 討論

在處理數學語料庫的各階段任務中，我們遇到一些不同的問題：包括了題目切割、標註單位，以及斷詞錯誤等方面。首先，當我們執行在 3.2 節中所提切割主幹及問句的原則（即“問句”主要是介於問號之前與最後一個逗號之後）後，發現大部分題目都能正確地被切割開來，但有極少數的題目並不一定適用（如“[小玉和家人一起去爬山，上山花了 145 分，下山花了 2 時 18 分，上山花的時間比較長，][還是下山花的時間比較長？]”）。再者，在“主幹”部分以搜尋關鍵字“大約”、“*分之*”來查找概數、分數題目是不容易的，反而必須藉由“問句”所要求答案的敘述（如“大約”、“約”、“*分之*”等），才得以找到此類題目。以上這些問題，皆是藉由人工檢視後才予以確認並修訂。

最後，我們發現因為數學文字問題語料與用來發展斷詞工具的通用型語料差異甚大，而導致在人名（如“[小][聰]”）及動詞片語（如“[拼拼][圖]”）常有斷詞錯誤的情形。因此在解題系統所用的斷詞工具，有必要執行領域調適（Domain Adaptation）。

5. 相關研究

在過去語料庫標註方面，不乏著重於研究剖析樹[8, 15-18]、語意角色[9, 19-22]、依存結構[10, 23]、片語結構[11]，或者是語篇[12]、共同指稱關係[13]和述語參數[13, 14]等等方面，卻缺乏針對數學文字問題的語料庫。

近年雖然已有關於數學文字問題研究的題庫出現[2, 5-6, 24-25]，但都是從網站論壇抓取的特定數據資料組成的：如 Kushman 等人[5]刪掉了非線性代數問題的題目，Hosseini 等人[2]只抓取組合加法、減法、一元一次方程和美元文字等問題（但帶有多餘數字資訊和辭彙空缺 (Lexical Gap)），Roy 等人[6]雖然增加了兩個以上運算子（多步驟）的問題，卻把需要背景知識的概數題目排除在外，Shi 等人[24]只抓取數字運算文字問題，Koncel-Kedziorski 等人[25]只抓取代數文字相關的問題（取材自[2]），但其中有高度重疊的辭彙與公式。換句話說，許多題目皆大同小異。總而言之，上述這些數學文字問題的題庫，在題目抽樣上皆有所偏頗，無法公平比較不同解題系統及反映各個解題系

統之真正能力。相反地，我們的中文數學語料庫卻是依各個年級分類，且不做任何增減，因此可以充分評測、比較不同的系統。

6. 結論

總括來說，我們所建立的中文國小數學文字問題語料庫，乃完整取材自出版社數學題庫中的文字問題，並且按照年級分層將每題（包含切割題目為“主幹”和“問句”兩部分）轉成解題系統容易使用的格式。不僅如此，我們還在題目中添加答案單位、分數與概數題的補充訊息、及教育部指標分類；並於少部分語料中標註十六種題目類型和語言學現象（句法分析、指代、蘊涵、指稱等），使研究者可以瞭解數學文字問題的文本特性，並可讓解題系統測試模型及執行半教導式學習。

本語料庫是第一個完整的中文國小數學文字問題的語料庫，可用來清楚的評估解題系統的真正程度（與國小各個年級對應）、公平比較不同系統的解題能力、並可清楚的反映出各系統的偏頗及缺失。此外，若有個別的應用或規格，亦能讓研發人員按其需求組合出想要的檢索，便於塑造不同組態的特殊語料庫。

致謝

本研究感謝中央研究院資訊科學研究所許聞廉教授主持之智慧型代理人系統實驗室提供並授權國小題庫，以及感謝科技部「建立一個領域相關機器閱讀系統之探討」(MOST104-2221-E-001-025)計劃經費補助。此外，我們也要謝謝本所的自然語言理解實驗室及詞庫小組實驗室其他參與研究和協同指導之老師與相關同仁。最後，我們感謝所有匿名審稿教授的寶貴意見，以改善此論文闕漏之處。

參考文獻

- [1] A. Mukherjee and U. Garain, “A review of methods for automatic understanding of natural language mathematical problems,” *Artificial Intelligence Review*, vol. 29, no. 2, pp. 93–122, Apr. 2008.
- [2] M. J. Hosseini, H. Hajishirzi, O. Etzioni, and N. Kushman, “Learning to solve arithmetic word problems with verb Categorization,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. [Online]. Available: <http://ssli.ee.washington.edu/~hannaneh/algebra-emnlp14.pdf>.

- [3] Y.-C. Lin *et al.*, “Designing a Tag-Based Statistical Math Word Problem Solver with Reasoning and Explanation,” *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, vol. 20, no. 2, pp. 1–26, Dec. 2015. [Online]. Available: <http://www.aclclp.org.tw/clclp/v20n2/v20n2a1.pdf>.
- [4] C.-C. Liang, K.-Y. Hsu, C.-T. Huang, C.-M. Li, S.-Y. Miao, and K.-Y. Su, “A tag-based English math word problem Solver with understanding, reasoning and explanation,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics (ACL), 2016, pp. 67–71. [Online]. Available: <http://aclweb.org/anthology/N/N16/N16-3014.pdf>.
- [5] N. Kushman, Y. Artzi, L. Zettlemoyer, and R. Barzilay, “Learning to automatically solve algebra word problems,” *Association for Computational Linguistics (ACL)*, vol. 1, pp. 271–281, Jun. 2014. [Online]. Available: <http://homes.cs.washington.edu/~lsz/papers/kazb-acl14.pdf>.
- [6] S. Roy, T. Vieira, and D. Roth, “Reasoning about quantities in natural language,” *Transactions of the Association for Computational Linguistics (TACL)*, vol. 3, pp. 1–13, Jan. 2015. [Online]. Available: <https://transacl.org/ojs/index.php/tacl/article/view/452/102>.
- [7] S. Roy and D. Roth, “Solving general arithmetic word problems,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 1743–1752. [Online]. Available: <http://www.emnlp2015.org/proceedings/EMNLP/pdf/EMNLP202.pdf>.
- [8] Z. Min, J. Hongfei, T. A. Ai, S. Jun, L. Sheng, and L. T. Chew, “A tree-to-tree alignment-based model for statistical machine translation,” in *MT-Summit*, 2007, pp. 535–542. [Online]. Available: <https://www.semanticscholar.org/paper/A-Tree-to-tree-Alignment-based-Model-for-Zhang-Jiang/5fb3ded74c75a78f637e198337b8a83c54975010/pdf>.
- [9] M. Palmer, D. Gildea, and P. Kingsbury, “The proposition bank: An annotated corpus of semantic roles,” *Computational Linguistics*, vol. 31, no. 1, pp. 106–71, Jan. 2005. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1122628>.
- [10] J. Hockenmaier and M. Steedman, “CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank,” *Computational Linguistics*, vol. 33, no. 3, pp. 355–396, Jan. 2007. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1288685>.

- [11] N. Xue, F. Xia, F. Chiou, and M. Palmer, “The Penn Chinese TreeBank: Phrase structure annotation of a large corpus,” *Natural Language Engineering*, vol. 11, no. 2, pp. 207–238, Jan. 2005. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1064785>.
- [12] Y. Zhou and N. Xue, “The Chinese discourse TreeBank: A Chinese corpus annotated with discourse relations,” *Language Resources and Evaluation*, vol. 49, no. 2, pp. 397–431, Nov. 2014.
- [13] R. Iida, M. Komachi, K. Inui, and Y. Matsumoto, “Annotating a Japanese text corpus with Predicate-Argument and Coreference relations,” in *Proceedings of the Linguistic Annotation Workshop*, Association for Computational Linguistics (ACL), 2007, pp. 132–139. [Online]. Available: <http://aclasb.dfki.de/nlp/bib/W07-1522>.
- [14] M. Roth and A. Frank, “Aligning predicate argument structures in monolingual comparable texts,” in *the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics (ACL), 2012, pp. 218–227. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2387672>.
- [15] W. Wang, K. Knight, and D. Marcu, “Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy,” in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 746–754. [Online]. Available: <http://aclasb.dfki.de/nlp/bib/D07-1078>.
- [16] H. Mi, L. Huang, and Q. Liu, “Forest-Based Translation,” Association for Computational Linguistics (ACL), 2008, pp. 192–199. [Online]. Available: <http://www.aclweb.org/anthology/P/P08/P08-1023.pdf>.
- [17] Y. Marton and P. Resnik, “Soft Syntactic Constraints for Hierarchical Phrased-Based Translation,” Association for Computational Linguistics (ACL), 2008, pp. 1003–1011. [Online]. Available: <http://www.anthology.aclweb.org/P/P08/P08-1114.pdf>.
- [18] D. Xiong, M. Zhang, and H. Li, “Learning translation boundaries for phrase-based Decoding,” in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), 2010, pp. 136–144. [Online]. Available: <http://www.aclweb.org/anthology/N10-1016>.

- [19] V. Nastase, J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz, “Learning noun-modifier semantic relations with corpus-based and WordNet-based features,” in *Proceedings of the national conference on Artificial intelligence (AAAI-06)*, 2006, pp. 781–786. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1597538.1597663&coll=&dl=&preflayout=tabs>.
- [20] A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal, “The SALSA corpus: A German corpus resource for lexical semantics,” in *Proceedings of the International Conference on Language Resources and Evaluation*, 2006, pp. 969–974. [Online]. Available: http://www.nlpado.de/~sebastian/pub/papers/lrec06_burchardt1.pdf.
- [21] P. Moreda, B. Navarro, and M. Palomar, “Corpus-based semantic role approach in information retrieval,” *Data & Knowledge Engineering*, vol. 61, no. 3, pp. 467–483, Jun. 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0169023X06001133>.
- [22] V. Basile, J. Bos, K. Evang, and N. Venhuizen, “Developing a large semantically annotated corpus,” in *Proceedings of the International Conference on Language Resources and Evaluation*, 2012, pp. 3196–3200. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/pdf/534_Paper.pdf.
- [23] S. Mille and L. Wanner, “Syntactic Dependencies for multilingual and multilevel corpus Annotation,” in *Proceedings of the International Conference on Language Resources and Evaluation*, 2010, pp. 1889–1896. [Online]. Available: http://lrec-conf.org/proceedings/lrec2010/pdf/697_Paper.pdf.
- [24] S. Shi, Y. Wang, C.-Y. Lin, X. Liu, and Y. Rui, “Automatically solving number word problems by semantic parsing and reasoning,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015, pp. 1132–1142. [Online]. Available: <http://www.aclweb.org/anthology/D/D15/D15-1135.pdf>.
- [25] R. Koncel-Kedziorski, H. Hajishirzi, A. Sabharwal, O. Etzioni, and S. Dumas, “Parsing algebraic word problems into equations,” *Transactions of the Association for Computational Linguistics (TACL)*, vol. 3, pp. 585–597, 2015. [Online]. Available: <https://transacl.org/ojs/index.php/tacl/article/viewFile/692/164>.
- [26] K.-J. Chen *et al.*, “Sinica Treebank: Design criteria, representational issues and implementation,” in A. Abeille (ed.), *Treebanks: Building and Using Parsed*. Kluwer Acade, 2003, pp. 231–248. [Online]. Available:

https://www.researchgate.net/profile/Chu-Ren_Huang/publication/226727962_Sinica_Treebank/links/0912f509c0082732b9000000.pdf.

- [27] C.-R. Huang, F.-Y. Chen, K.-J. Chen, and Z. Gao, “Sinica Treebank: design criteria, annotation guidelines, and on-line interface,” in *Proceedings of the second workshop on Chinese language processing held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), 2000, vol. 12, pp. 29–37. [Online]. Available: http://www.aclweb.org/old_anthology/W/W00/W00-1205.pdf.
- [28] K.-J. Chen, S.-L. Huang, Y.-Y. Shih, and Y.-J. Chen, “Extended-HowNet- A Representational Framework for Concepts,” in *OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop*, Jeju Island, South Korea, 2005. [Online]. Available: http://ckip.iis.sinica.edu.tw/CKIP/paper/Extended-HowNet-_A_Representational_Framework_for_Concepts.pdf.
- [29] S.-L. Huang, Y.-M. Hsieh, S.-C. Lin, and K.-J. Chen, “Resolving the Representational Problems of Polarity and Interaction between Process and State Verbs,” *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, vol. 19, no. 2, pp. 33–52, Jun. 2014. [Online]. Available: http://godel.iis.sinica.edu.tw/CKIP/paper/2014_IJCLCLP_33-52.pdf.
- [30] S.-M. Li, S.-C. Lin, C.-H. Tai, and K.-J. Chen, “A Probe into Ambiguities of Determinative-Measure Compounds,” *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, vol. 11, no. 3, pp. 245–280, Sep. 2006. [Online]. Available: <http://www.airitilibrary.com/Publication/alDetailedMesh?DocID=1027376x-200609-11-3-245-280-a>.
- [31] C.-H. Tai, J.-Z. Fan, S.-L. Huang, and K.-J. Chen, “Automatic Sense Derivation for Determinative-Measure Compounds under the Framework of E-HowNet,” *International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, vol. 14, no. 1, pp. 19–44, Mar. 2009. [Online]. Available: <http://www.aclweb.org/anthology/O09-3002>.
- [32] S. Upadhyay and M.-W. Chang, “DRAW: A challenging and diverse algebra word problem set,” Number MSR-TR-2015-78, Oct. 2015. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tech_rep.pdf.
- [33] R. Koncel-Kedziorski, S. Roy, A. Amini, N. Kushman, and H. Hajishirzi, “MAWPS: A math word problem repository,” in *Proceedings of the 2016 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics (ACL), 2016, pp. 1152–1157. [Online]. Available: <https://www.aclweb.org/anthology/N/N16/N16-1136.pdf>.

- [34] D. Huang, S. Shi, C.-Y. Lin, J. Yin, and W.-Y. Ma, “How well do computers solve math word problems? Large-scale Dataset construction and evaluation,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Association for Computational Linguistics (ACL), 2016, pp. 887–896. [Online]. Available: <http://www.aclweb.org/anthology/P/P16/P16-1084.pdf>.
- [35] W.-K. Wong, S.-C. Hsu, S.-H. Wu, C.-W. Lee, and W.-L. Hsu, “LIM-G: Learner-initiating instruction model based on cognitive knowledge for geometry word problem comprehension,” *Computers & Education*, vol. 48, no. 4, pp. 582–601, 2007. [Online]. Available: http://iasl.iis.sinica.edu.tw/webpdf/paper-2005-LIM-G_Learner-initialing_Instruction_Model.pdf.
- [36] Y.-K. Wang, Y.-S. Chen, and W.-L. Hsu, “Empirical study of Mandarin Chinese discourse analysis: An event-based approach,” in *Proceedings Tenth IEEE International Conference on Tools with Artificial Intelligence (Cat. No.98CH36294)*, Institute of Electrical & Electronics Engineers (IEEE). [Online]. Available: http://iasl.iis.sinica.edu.tw/webpdf/paper-1998-Empirical_Study_of_Mandarin_Chinese_Discourse_.pdf.
- [37] Y.-K. Wang, W.-L. Hsu, and Y.-C. Chen, “The Anaphoric expressions of Chinese algebraic word problem,” in *Proceedings of the 1998 International Symposium on Multimedia Information Processing*, 1998. [Online]. Available: http://iasl.iis.sinica.edu.tw/webpdf/paper-1998-The_Anaphoric_Expressions_of_Chinese_Algebraic_.pdf.