

中文詞義全文標記語料庫之設計與雛形製作

¹柯淑津、²黃居仁、³洪嘉馥、¹劉詩音、¹簡卉伶、²蘇依莉

¹東吳大學資訊科學系

{ksj, ms9405, ms9504}@cis.scu.edu.tw

²中央研究院語言所

{churen, isu}@gate.sinica.edu.tw

³台灣大學語言學研究所

jiafei@gate.sinica.edu.tw

摘要

詞義標示語料庫對自然語言處理佔有很重要的地位，尤其反映在訊息特徵及自然語言理解之研究上，但目前大規模之中文詞義標示語料庫卻付之闕如。本文設計出一個超過 11 萬詞的大規模中文詞義全文標示語料集，以中研院平衡語料庫為標示對象，從中摘錄 56 篇完整文章，利用 N-Gram 與搭配資訊等語言知識，並結合機器學習技巧以及機率模式的方式作為處理自動詞義標示的前置作業工作，最後為達高精確度之效果，再將自動產生之標示結果經由人工校訂而成。

關鍵詞：詞義辨識，詞義標記語料庫，自然語言處理，誘導式作法

Keywords: Word Sense Disambiguation, Sense Tagged Corpus, Natural Language Processing, Bootstrap Method

一、簡介

語料庫對自然語言處理研究佔有相當重要的地位。尤其是統計式計算語言處理，常需仰賴語料庫所蘊藏的豐富資源，作為計算的依據。隨著數位文獻之普及，語料庫的種類越來越多，內容也越來越豐富。標示訊息越完整的語料庫對研究的幫助越大。有些語料庫只呈現原始文本內容，有的則再加上詞性標示或詞義等相關資料。目前，標示詞性的語料庫有不少，例如：中文方面有中研院的一千萬詞平衡語料庫[1]及中文十億詞語料庫 (Chinese Gigaword Corpus)[2]等，至於標示詞義的語料庫不論中英文都很少。標示詞義 (Sense)或語意(Semantic)的語料資源，在理論語言學上，將提供詞彙語意學研究之豐富材料與基本架構；在計算語言學上，將對自然語言處理的核心工作如：WSD 多重詞義辨析研究、自然語言理解等，都將會有關鍵性的突破。另外，這些標示語料經統計處理後所獲得之資訊，可應用於資訊檢索、資訊擷取、文件摘要、自動問答等議題之研究。

大量精確的詞義標示資料，可提供多項計算語言相關研究的豐富素材。但是，中文語料庫詞義標記主要的瓶頸為缺乏足供自動標記參考的資料，而人工標示需要昂貴成本，造成語料庫標示語意工作的難產。近年來的許多研究，顯示出對大規模詞義標示語料集的大量需求，這些資源在建構上是否完備，往往會影響整個研究進行方向以及研究結果的正確性。在某些語言上，已具有較代表性的詞義標示語料集存在，例如英文語料集 SemCor [3]以及 Senseval [4] 提供之多國語言的全文標示語料集，例如捷克語、荷蘭語，義大利文及英文。反觀中文，目前大量中文標示詞義語料集卻一直付之闕如，只有少數幾個規模不大的中文詞義標示語料集，例如 Senseval-2 的中文詞義標示語料集包括 15 個中文詞彙的詞義標示，Senseval-3 的中文詞義標示語料集包含 20 個中文詞，其規模與真實的語言環境存在相當大的差異。製作大規模語料庫所遇到最大瓶頸為成本，藉由人工標示雖然可以得到高正確率，但所需的詞義標示成本卻非常昂貴，而且可以標示詞義的專家也不多見。為了克服此問題，本文提出一套半自動詞義標示方法，作為標示詞義的前置作業，再經由專門人士校訂。語料庫製作以中研院平衡語料庫為對象，從中摘錄文章，並對摘錄出之文章中之詞做詞義標示的動作，設計製作出一個大規模的中文詞義標示語料集以供自然語言處理研究使用。

二、詞義區分詞典

本文標記中文詞義語料庫所使用的詞義區分詞典為中央研究院資訊所、語言所詞庫小組所製作的『詞義區辨小辭典』第三版[5]，詞典的內容以中頻詞為主，共包含 5047 個詞形，9400 個詞義。詞典所收錄的詞條(entry)，以現代漢語通用語詞為範圍，不列入現今已不用或罕用的詞彙。而收錄的中文詞彙條目，包含單字詞、雙字詞和多字詞。詞典中提供各詞條豐富的訊息，除詞目(lemma)、標音(拼音與注音)、義項、詞類、例句等內容外，還包括有各詞義對應至英文詞網 WordNet 2.0 (<http://wordnet.princeton.edu/>)之同義詞集及其編號。圖一為詞彙「瘋狂」在字典中的訊息，共有兩個詞義，其中第一個詞義「形容人因精神錯亂而舉止失常。」對應於詞網的同義詞集{crazy}，第二個詞義「形容行為或事物無節制，超乎平常的程度。」對應於詞網的同義詞集則為{madly}，「瘋狂」的兩個詞義各自都可又細分為兩個義面。

三、標示語料來源

本文使用『中央研究院現代漢語平衡語料庫』(Sinica Corpus) [1] 作為語料標的。語料中的每個文句都已依詞斷開，並標示詞性。本研究為求表達出文脈結構與前後文關係等訊息之完整性，標示語料的選擇以全文為單位。詞義的標示以全文標示為原則，但因本文所使用的詞義區分字典目前仍在編撰中，因此並非每個出現在語料庫中的詞彙都收錄在詞義區分字典裡，對於字典尚未收錄之詞彙，我們以其詞性做為標示。標記詞性可以降低詞義的歧義度，因此詞性的標示可視為粗的詞義標記。

我們依字典詞彙出現於文章內容中的覆蓋率以及文章長度作為選擇之依據，共選出 56 篇文章，總長度為 114,066 個詞彙，148,863 個字元。語料集中的文章主題分佈統計結果如表一，以文章數目計算，文學類最多共有 35 篇，若以文章長度計算，則以生活類的佔有率最高。

瘋狂 feng1 kuang2 ㄈㄥ ㄨㄤ ㄨㄛˋ ㄨㄥˋ

詞義1：【不及物動詞，VH；名詞，nom】形容人因精神錯亂而舉止失常。{crazy, 00872382A}

義面1：【不及物動詞，VH】形容人因精神錯亂而舉止失常。

例句：片中一名〈瘋狂〉殺手，拿著剃刀。

例句：石門五子命案的父母與其說是迷信，不如說是〈瘋狂〉。

義面2：【名詞，nom】形容人因精神錯亂而舉止失常。

例句：因此，石門五子命案的〈瘋狂〉，其實也正是我們社會瘋狂的一粒種籽啊！

詞義2：【不及物動詞，VH；名詞，nom】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。{madly, 00045197R}

義面1：【不及物動詞，VH】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。

例句：他〈瘋狂〉的愛上一個女孩子。

例句：每年都有不計其數的台灣客前往香港〈瘋狂〉大採購。

例句：當時少棒青少棒在台灣很〈瘋狂〉，連我們城市的小孩子也愛打棒球。

義面2：【名詞，nom】形容行爲或事物無節制，超乎平常的程度。通常用於人的情感或事件的程度。

例句：經過一陣〈瘋狂〉後，大家都累了，個個都喊著喉嚨痛、腳痛。

例句：死了七九百餘人的人民教室案，也使人想到愈來愈多的宗教〈瘋狂〉事件。

例句：只要幅度不超過，則多頭仍然大有可爲，但仍切忌一味追高的〈瘋狂〉舉動。

圖一 詞彙「瘋狂」在『詞義區辨小辭典』詞典範例

表一、語料集所含文章主題分佈

主題	文章篇數	文章長度	
		詞彙	字元
哲學	4	1451	1976
社會	5	27385	35918
生活	12	57605	74710
文學	35	27625	36259
總計	56	114066	148863

整體而言我們標示詞義之目標詞彙，以單一詞性(詞形與詞性之配對)為單位做統計，被收錄於字典之單義詞共 863 個，出現頻率為 12,124 次，多義詞有 650 個，出現頻率 23,521 次，統計其詞性分佈如表二。多義詞詞義數量由 2 (如：自然 D、堆 Nf 和喜 VK 等) 至 27(如：吃 VC) 不等，平均詞義數為 2.97。以詞性區分，平均而言詞義數最多之詞性為語助詞，平均達 4.83 個詞義數。詞義數最少之詞性為感嘆詞，平均為 1.32 個詞義數。若是以詞形為單位做統計，不考慮其詞性之差異，在本研究使用的 56 篇文件中共有 598 個詞形收錄於詞典中，每個詞形的平均詞義數為 4.53。

表二、標示語料集詞性分佈

詞類	詞彙數	詞例數	範例
不及物動詞	231	3,317	對 _{VH} , 跑 _{VA} , 走 _{VA}
介詞	51	1,854	在 _P , 跟 _P , 到 _P
及物動詞	373	5,733	說 _{VE} , 沒有 _{VI} , 開始 _{VL}
名詞	321	5,070	人家 _{Nh} , 感覺 _{Na} , 下 _{Ncd}
形容詞	21	45	一般 _A , 原 _A , 定期 _A
定詞	55	3,175	那 _{Nep} , 前 _{Nes} , 多 _{Ncqa}
後置詞	31	455	上 _{Ng} , 裡 _{Ng} , 當中 _{Ng}
副詞	287	8,892	就 _D , 又 _D , 起來 _{Di}
連接詞	69	1,554	就是 _{Cbb} , 而 _{Cbb} , 或 _{Caa}
量詞	81	976	回 _{Nf} , 份 _{Nf} , 間 _{Nf}
語助詞	47	4,574	啊 _T , 喔 _T , 哇 _I
總數	1567	35,645	

四、詞義標示語料庫

我們製作的詞義標示語料集型態為 XML 格式檔，所使用的標籤結構如圖二所示，標籤使用說明請見表三。語料庫內含多篇文件，每篇文件以<doc>標籤區隔，也就是在標籤<doc>及</doc>範圍內容為同一篇文件。文件內容再往下細分為句子，每個句子以<sent>標籤區隔，句子內容依詞彙出現順序呈現，其間以<w>標籤作為詞彙區隔，其內又再細分為三個標籤：word, pos, tag1 分別呈現詞彙、詞性、以及詞義標示等資訊。

在詞義標示部分，依標示類別分為三種，第一種為詞義代碼標示，採用 Huang et al. [6]之定義，為四位數整數，前兩位為詞義序號，表明標示詞義出現在字典中之詞義順序。第三碼是詞形標碼，第四碼為義面編碼（如圖三）。第二種為標點符號之處理，對於標點進行標示詞義，不具意義，因此，我們將標點符號的詞義代碼直接設定為其符號本身。至於，第三種是針對未知詞（包含辭典未收錄、尚未有詞義分析之詞彙）的部分，我們以該詞彙之詞性作為其詞義標示。圖四是部分標示語料範例，為語料庫中編號 101664 之文章第 18 句內容，第一個詞『灰灰』，是未知詞，標示詞義為其詞性「Nb」。第二個詞『說』的詞義標示為「0111」，表示在此處詞義「以口語媒介引述或陳述訊息。」為「說 1」的第 01 個詞義的第 1 個義面。另外，第 3, 5, 8 個詞是標點符號，因此詞義代碼為符號本身。

整個標示語料庫共含有 114066 個詞彙，我們依標示種類作為統計，其結果如表四。其中標點符號有 27530 個，成功標示出詞義代碼之詞彙數共有 35645 個，對於未知詞或是字典尚未收編處理之詞，我們以詞性作為詞義標碼的則有 50891 個詞例。我們進一步分析發現這部分資料，包括有：文章中有些英數字或是專名，例如：(二)、CPU、清華大學等，在我們的標示語料庫中共有 4258 個詞例。另外，因為本研究所使用之詞義區分子典尚在建構中，有些詞彙目前未收錄於我們的詞義字典中，這部分共有 4541 個詞

彙，在語料中共出現 31730 次。至於，剩餘的 14903 個詞例，雖然是已收錄於字典中之詞彙，但是不在本次規劃標記範圍內，因此先以詞性標記，加上詞性標記可降低多義詞之歧義度。

表三、語料庫使用標籤說明

標籤名稱	內容說明	例子
<corpus>	語料庫起始	<corpus>
<doc id=>	文件起始及編號	<doc id="100863">
<sent id=>	句子起始及編號	<sent id="1">
<w id=>	詞彙資料及編號	<w id="1">
<word>	詞彙	<word>人家</word>
<pos>	詞性	<pos>Nh</pos>
<tag1>	詞義標示	<tag1>0122</tag1>

表四、語料集標示種類統計

標示種類	詞彙數		說明
標點符號	27530		標點符號無需標示詞義
詞義代碼	35645		已完成詞義標示
詞性標示	50891	4258	不需標示（英數字、專名）
		31730	詞彙(4541 個)目前未收錄於字典中
		14903	未處理

總和 114066

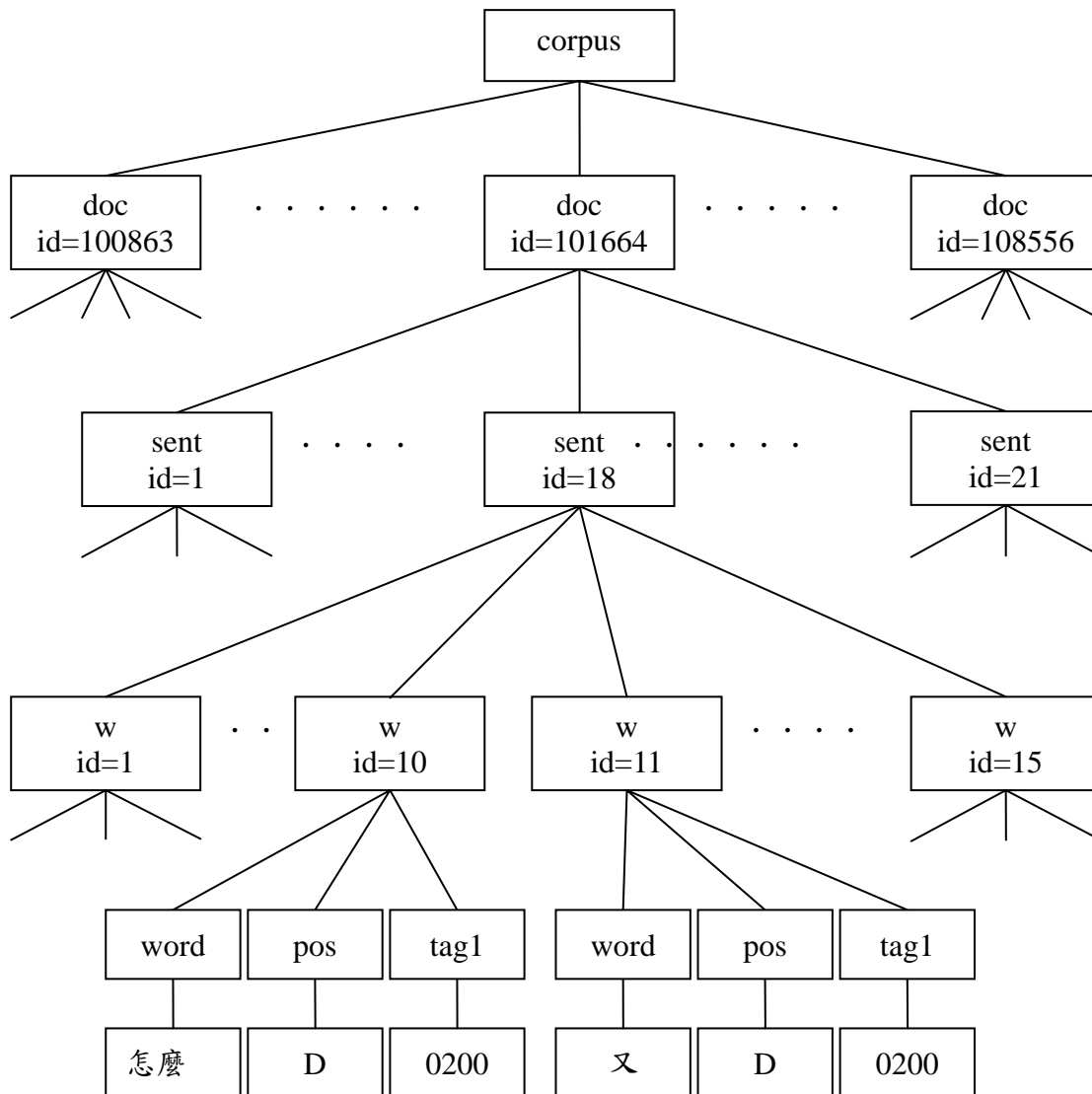
五、詞義標示方法

標示詞義之工作需仰賴大量的人力，因此，爲了節省成本，本文設計出一套半自動標示詞義之方法[7]，先利用此標示方法對語料作初步的詞義標示處理，以作爲人工標示之前置作業。

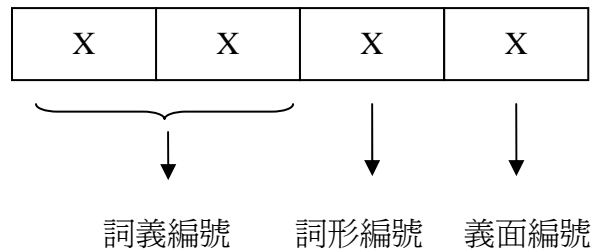
我們所設計之半自動標示詞義的方法，採用誘導式方法（bootstrap）逐步放寬標示條件，來擴增標示語料，其系統組織圖如圖五所示。首先蒐集一些已被標示過詞義的資料作爲詞義標示工作的種子訓練資料。其來源主要有兩個部分，第一個部分爲詞義區分詞典中之例句，第二個部分爲辭典編撰小組，在搜尋整理詞義過程中所標示的語料庫部分內容。若來自這兩部分的例句數量不足時，我們會隨機從研究院語料庫中選出部分文句，交由人工標示詞義後加入成爲種子標示句。將上述已標示資料合爲訓練集，以本文選出來的 56 篇標示集文章，則作爲測試集。

自動標示詞義的第一階段採用 N-gram 模式，將標示出詞義的資料加入訓練集中，以作爲第二階段的訓練語料。本文利用 N-gram 處理詞義標示是基於下面的假設：存在

包圍目標詞彙前後 N 個詞彙完全相同的兩個子句，我們推論它們應擁有一樣的詞義 [8]。在此使用 N-gram 有兩項主要目的，第一是擴大訓練集，因語料庫中常可見到相似之子句。第二個目的是過濾訓練資料集的雜訊，以此檢驗人工標示資料之不一致性。



圖二、語料集標籤結構及範例



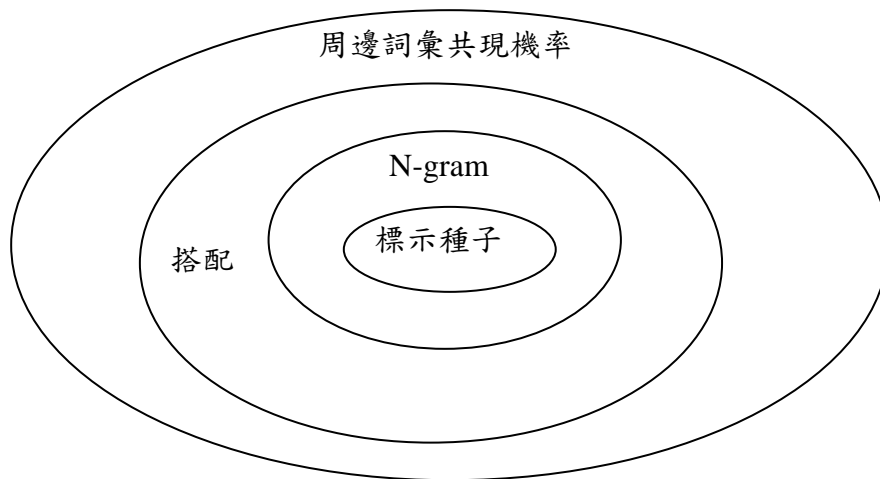
圖三、詞義代碼編碼方式說明

```

<doc id="101664">
  <sent id="1">
    :
    :
  <sent id="18">
    <w id="1"> <word>灰灰</word> <pos>Nb</pos> <tag1> Nb </tag1> </w>
    <w id="2"> <word>說</word> <pos>VE</pos> <tag1>0111</tag1> </w>
    <w id="3"> <word> : </word> <pos>COLONCATEGORY</pos> <tag1> : </tag1> </w>
    <w id="4"> <word>白白</word> <pos> Nb </pos> <tag1> Nb </tag1> </w>
    <w id="5"> <word> , </word> <pos>COMMACATEGORY</pos> <tag1> , </tag1> </w>
    <w id="6"> <word>剛剛</word> <pos>D</pos> <tag1>D</tag1> </w>
    <w id="7"> <word>見面</word> <pos>VA</pos> <tag1>VA</tag1> </w>
    <w id="8"> <word> , </word> <pos>COMMACATEGORY</pos> <tag1> , </tag1> </w>
    <w id="9"> <word>你</word> <pos>Nh</pos> <tag1>Nh</tag1> </w>
    <w id="10"> <word>怎麼</word> <pos>D</pos> <tag1>0200</tag1> </w>
    <w id="11"> <word>又</word> <pos>D</pos> <tag1>0200</tag1> </w>
    <w id="12"> <word>要</word> <pos>D</pos> <tag1>D</tag1> </w>
    <w id="13"> <word>走</word> <pos>VA</pos> <tag1>0400</tag1> </w>
    <w id="14"> <word>了</word> <pos>Di</pos> <tag1>0110</tag1> </w>
    <w id="15"> <word> ? </word> <pos>QUESTIONCATEGORY</pos> <tag1> ? </tag1>
  </w>
</sent>
</doc>

```

圖四、XML 格式之標示語料範例



圖五、標示詞義系統組織圖

第二個階段我們使用搭配資訊來增加標示集數量，搭配資訊是一種很強的語言關係，能決定目標詞彙之詞義[9]。我們先以詞頻、搭配詞與目標詞彙距離變異量等條件作為選擇搭配詞彙之初步依據，最後再經過相互資訊 MI 計算來檢驗搭配詞與目標詞彙之間的相關程度。

經過 N-gram 及搭配資訊兩個階段的處理工作，我們將訓練語料標示量做了實質擴增。接著，再經過機率模式計算，盡可能將大部分詞彙標上詞義資訊。最後為求標示語料之高精度，我們將經由自動標示詞義處理過後的整個標示語料，再交由原字典編撰小組成員進行人工校正處理。

整個自動標示部分之實驗結果我們分為兩部分說明，第一部份詞義標示以詞義下再細分至義面為準，結果如表五所示，整體的正確率為 57.47%。至於，第二部分我們將詞義標示處理至詞義為止，不再細分義面，結果如表六所示，整體的正確率可提升至 64.51%。

六、結論

詞義標示語料庫對自然語言處理佔有很重要的地位，尤其反映在計算語言學研究上，常需語料庫所提供的豐富資訊來作計算，但目前存在的中文詞義標示語料集的數量少之又少，因此，我們設計出一個包含約十萬詞大規模的中文詞義標示語料集，以供自然語言處理相關研究使用。標示詞義之步驟為先使用自動詞義標示作為人工詞義標示之前置工作，再將結果交由人工校訂。自動詞義標示方法為利用周邊詞彙提供的資訊，經由 N-gram，搭配資訊以及機率模式計算出最有可能的詞義。未來藉由詞義區分詞典的漸趨完備，期望能達到對中央研究院現代漢語平衡語料庫五百萬詞全文標記的目標。

表五 詞義標示至義面為準的實驗結果

詞性	詞彙數	詞例數	詞例比率	正確詞例	錯誤詞例	正確%	錯誤%
A	6	22	0.10%	14	8	63.64%	36.36%
Caa	2	38	0.16%	37	1	97.37%	2.63%
Cbb	7	231	1.00%	37	194	16.02%	83.98%
D	64	3454	14.98%	2135	1319	61.81%	38.19%
Da	7	22	0.10%	21	1	95.45%	4.55%
Dfa	5	202	0.88%	200	2	99.01%	0.99%
Dfb	1	1	0.00%	1	0	100.00%	0.00%
Di	7	1146	4.97%	934	212	81.50%	18.50%
Dk	2	5	0.02%	5	0	100.00%	0.00%
I	15	693	3.00%	307	386	44.30%	55.70%
Na	98	648	2.81%	563	85	86.88%	13.12%
Nb	5	18	0.08%	18	0	100.00%	0.00%
Nc	8	29	0.13%	27	2	93.10%	6.90%
Ncd	13	283	1.23%	209	74	73.85%	26.15%
Nep	6	2227	9.66%	615	1612	27.62%	72.38%
Neqa	2	38	0.16%	32	6	84.21%	15.79%
Nes	6	128	0.56%	118	10	92.19%	7.81%
Neu	3	127	0.55%	114	13	89.76%	10.24%
Nf	40	228	0.99%	212	16	92.98%	7.02%
Ng	13	147	0.64%	102	45	69.39%	30.61%
Nh	8	1668	7.23%	140	1528	8.39%	91.61%
P	33	1659	7.19%	1136	523	68.48%	31.53%
T	13	2838	12.30%	1648	1190	58.07%	41.93%
VA	28	451	1.96%	328	123	72.73%	27.27%
VAC	1	4	0.02%	3	1	75.00%	25.00%
VB	9	14	0.06%	14	0	100.00%	0.00%
VC	76	1177	5.10%	1061	116	90.14%	9.86%
VCL	5	174	0.75%	103	71	59.20%	40.80%
VD	19	170	0.74%	128	42	75.29%	24.71%
VE	26	1703	7.38%	1370	333	80.45%	19.55%
VF	5	20	0.09%	11	9	55.00%	45.00%
VG	9	170	0.74%	103	67	60.59%	39.41%
VH	66	1940	8.41%	661	1279	34.07%	65.93%
VHC	2	13	0.06%	13	0	100.00%	0.00%
VI	3	4	0.02%	4	0	100.00%	0.00%
VJ	19	326	1.41%	206	120	63.19%	36.81%
VK	11	63	0.27%	55	8	87.30%	12.70%
VL	5	160	0.69%	140	20	87.50%	12.50%
V_2	1	823	3.57%	433	390	52.61%	47.39%
nom	1	1	0.00%	1	0	100.00%	0.00%
	650	23065	100.00%	13259	9806	57.47%	42.51%

表六 詞義標示僅處理至詞義為準的實驗結果

詞性	詞彙數	詞例數	詞例比率	正確詞例	錯誤詞例	正確%	錯誤%
A	6	22	0.10%	14	8	63.64%	36.36%
Caa	2	38	0.16%	37	1	97.37%	2.63%
Cbb	7	231	1.00%	37	194	16.02%	83.98%
D	64	3454	14.98%	2158	1296	62.48%	37.52%
Da	7	22	0.10%	21	1	95.45%	4.55%
Dfa	5	202	0.88%	200	2	99.01%	0.99%
Dfb	1	1	0.00%	1	0	100.00%	0.00%
Di	7	1146	4.97%	934	212	81.50%	18.50%
Dk	2	5	0.02%	5	0	100.00%	0.00%
I	15	693	3.00%	307	386	44.30%	55.70%
Na	98	648	2.81%	570	78	87.96%	12.04%
Nb	5	18	0.08%	18	0	100.00%	0.00%
Nc	8	29	0.13%	27	2	93.10%	6.90%
Ncd	13	283	1.23%	209	74	73.85%	26.15%
Nep	6	2227	9.66%	642	1585	28.83%	71.17%
Neqa	2	38	0.16%	32	6	84.21%	15.79%
Nes	6	128	0.56%	118	10	92.19%	7.81%
Neu	3	127	0.55%	114	13	89.76%	10.24%
Nf	40	228	0.99%	212	16	92.98%	7.02%
Ng	13	147	0.64%	102	45	69.39%	30.61%
Nh	8	1668	7.23%	1549	119	92.87%	7.13%
P	33	1659	7.19%	1143	516	68.90%	31.10%
T	13	2838	12.30%	1660	1178	58.49%	41.51%
VA	28	451	1.96%	347	104	76.94%	23.06%
VAC	1	4	0.02%	3	1	75.00%	25.00%
VB	9	14	0.06%	14	0	100.00%	0.00%
VC	76	1177	5.10%	1065	112	90.48%	9.52%
VCL	5	174	0.75%	107	67	61.49%	38.51%
VD	19	170	0.74%	128	42	75.29%	24.71%
VE	26	1703	7.38%	1475	228	86.61%	13.39%
VF	5	20	0.09%	11	9	55.00%	45.00%
VG	9	170	0.74%	103	67	60.59%	39.41%
VH	66	1940	8.41%	664	1276	34.23%	65.77%
VHC	2	13	0.06%	13	0	100.00%	0.00%
VI	3	4	0.02%	4	0	100.00%	0.00%
VJ	19	326	1.41%	206	120	63.19%	36.81%
VK	11	63	0.27%	55	8	87.30%	12.70%
VL	5	160	0.69%	140	20	87.50%	12.50%
V_2	1	823	3.57%	433	390	52.61%	47.39%
nom	1	1	0.00%	1	0	100.00%	0.00%
Total	650	23065	100.00%	14879	8186	64.51%	35.49%

參考文獻

- [1] Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*. Seoul:Kyung Hee University. pp.167-176.
- [2] Wei-yun Ma, and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Presented at the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Italy. 24-28 May.
- [3] SemCor, <http://multisemcor.itc.it/semcor.php>
- [4] Senseval, <http://www.senseval.org/>
- [5] 黃居仁，主編。中文的意義與詞義。中央研究院文獻語料庫與詞庫小組技術報告 06-03。南港，中研院, 2006。
- [6] Huang, Chu-Ren, Chun-Ling Chen, Cui-Xia Weng, Hsiang-Ping Lee, Yong-Xiang Chen and Keh-jiann Chen. 2005. The Sinica Sense Management System: Design and Implementation, *Computational Linguistics and Chinese Language Processing*. Vol. 10, No. 4, pp. 417-430.
- [7] 柯淑津、黃居仁、陳振南，2004，全語料庫中文詞義標記的初步研究，第五屆詞彙語意研討會，北京。
- [8] 柯淑津、陳振南，2007，結合機器學習與語言知識的全語料庫中文詞義標示方法,第八屆詞彙語意研討會，香港。
- [9] Yarowsky, 1993, One Sense Per Collocation, In Proceedings of ARPA Human Language Technology Workshop, Princeton.