

國立臺灣大學資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia
College of Electrical Engineering & Computer Science

National Taiwan University

Master Thesis

應用於中文意見分析之詞內暨詞間語法結構自動擷取研究

Automatic Extraction of Intra- and Inter- Word Syntactic

Structures for Chinese Opinion Analysis



黃挺豪

Huang Ting-Hao

指導教授：陳信希博士

Advisor: Chen, Hsin-Hsi, Ph.D.

中華民國 98 年 6 月

JUNE, 2009

誌謝

"Would you tell me, please, which way I ought to go from here?"

「可不可以請你告訴我，從這裡我應該往哪裡去呢？」

"That depends a good deal on where you want to get to," said the Cat.

「那主要還看妳想往哪裡去囉。」這貓說道。

"I don't much care where--" said Alice.

「我不太在乎是去哪裡——」Alice 說。

"Then it doesn't matter which way you walk," said the Cat.

「那麼妳往哪條路走都無所謂吧。」貓說道。

"--so long as I get somewhere," Alice added as an explanation.

「——只要能讓我走到某個地方就行了！」Alice 補充說明。

"Oh, you're sure to do that," said the Cat, "if you only walk long enough."

「噢，妳一定可以的。」貓說，「只要妳走得夠遠。」

——《愛麗絲漫遊奇境》 Chapter VI. PIG AND PEPPER. 第 45 段

感謝陳信希老師兩年多來的悉心指導與毫無保留的支持，感謝倫維學姊永遠充滿活力的陪伴，感謝嫵朱穿行過大雪來見我，感謝少女的甜點聚會，感謝吵鬧或脾氣差的妹妹們，感謝舜文的晚安。

只要走得夠遠，一定可以到達某一個地方。

而我終於真正走進夏天裡了。

感謝所有愛我的與我愛的，這本論文寫滿了你們。

摘要

本研究之宗旨在於「將語法資訊引入意見分析中，改善其效能」。主要分為兩部分：詞內層次與詞間層次。

詞內層次方面，本研究首先參考各家分類方式，制定出一構詞分類架構，繼而就此架構展開語料標記工作。語料標記完成後，我們除對構詞類別分佈狀態進行統計外，亦對標記者間之答案一致性與人工標記時於各構詞類別之判定效能作了分析。分析結果顯示標記者間兩兩一致性係數（Kappa）均屬於「高度一致」範圍，肯認了此問題之信度。最後我們以《教育部國語辭典》之資訊為特徵值，於標記完成之語料集上以各種不同分類方法進行實驗，其中以條件隨機域模型（CRF）之效能最佳，對五大基本構詞類別可達到平均 F 分數為 0.6 的效能。

詞間層次方面，本研究首先比較意見句與非意見句之依存關係數量，藉此證實意見句之語法結構確有其特殊性；繼而對所有意見句之語法分析樹展開「標示意見結構」之標記工作，共標記約一萬餘句意見句，每句至少由兩位工讀生標記之。其標記結果一則可轉換為依存關係，從而比較句中「表達意見」之結構的特殊性，並歸納出 14 種較常用於意見表達之依存關係；另一方面，標記結果亦可直接於語法分析樹上進行預測。本研究將問題簡化為序列式標記問題，以條件隨機域模型直接於語法樹上標示出意見結構位置。並得到精確度（precision）極高、回收率（recall）偏低之實驗結果。

最後本研究亦將前述之詞內與詞間語法結構資訊施用於意見分析系統中，經實驗證實，此資訊確可改善目前之意見分析效能，致使意見句判斷達到 0.8 之 F 分數、意見詞極性判斷達到 0.6 之 F 分數。

關鍵詞：意見分析、意見擷取、構詞、語法結構、意見句、意見詞、語法關係

目錄

誌謝	i
摘要	iii
目錄	v
表目錄	ix
圖目錄	xi
第一章、緒論	1
1.1 研究動機	1
1.2 研究目的	2
1.3 論文架構	2
第二章、文獻探討	3
2.1 中文文本意見分析	3
2.1.1. 以經驗方法為基礎的中文文本意見分析	3
2.1.2. 語法結構資訊於中文文本意見分析之應用	3
2.2 中文語法結構研究及其自動剖析	4
2.2.1. 詞內層次	4
2.2.1.1. 中國大陸地區	4
2.2.1.1.1. 北京大學（俞士汶、朱學鋒等）	5
2.2.1.1.2. 清華大學（苑春法、黃昌寧等）	6
2.2.1.1.3. 魯東大學（亢世勇等）	7
2.2.1.2. 國際研討會	8
2.2.2. 詞間層次	9

2.2.2.1.	賓州大學樹庫 (Penn Treebank) 5.1 版	9
2.2.2.2.	依存關係樹	9
2.2.2.3.	史丹佛語法分析套件	9
第三章、中文詞內部語法結構自動分類		11
3.1	問題敘述	11
3.2	二字詞內部語法結構分類及其理論歧異	13
3.3	詞彙語料標記	17
3.3.1.	語料標記及過濾	17
3.3.2.	標記結果分析與文獻比較	19
3.4	二字詞內部結構自動分類	26
3.4.1.	特徵值抽取	27
3.4.1.1.	《教育部重編國語辭典修訂本》簡介	28
3.4.1.2.	使用之特徵值	30
3.4.2.	分類方法	35
3.4.2.1.	支援向量機 (SVM) 分類法	35
3.4.2.2.	條件隨機域 (CRF) 分類法	35
3.4.2.3.	單純貝氏 (Naïve Bayes) 分類法	36
3.4.2.4.	簡單機率分類法	36
3.4.2.5.	表格分類法	38
3.5	分類效能評估	39
3.5.1.	實驗設定	39
3.5.2.	實驗結果	40
3.5.3.	討論	42
3.6	小結	43

第四章、中文詞詞間結構自動擷取	45
4.1 問題敘述	45
4.2 基本定義：意見句與意見段落	46
4.2.1. 意見句與意見詞	46
4.2.2. 意見段落	47
4.3 問題初探：意見句及非意見句之依存關係樹比較	47
4.3.1. 意見句標記	47
4.3.2. 意見句及非意見句依存關係樹分佈比較	48
4.4 中文詞詞間結構語料標記	50
4.4.1. 標記目的及使用語料	50
4.4.2. 詞間結構定義與分類	52
4.4.3. 標記方法暨「潘恩標記系統」(Pan Annotation System)	54
4.5 語料分析	57
4.5.1. 原始標記結果分析	57
4.5.2. 依存關係分析	60
4.5.2.1. 依存關係轉換方法	60
4.5.2.2. 轉換結果統計及分析	62
4.6 詞間結構自動擷取	64
4.6.1. 自動擷取方法	64
4.6.2. 特徵值抽取	66
4.6.3. 結構自動擷取效能評估	67
4.6.3.1. 實驗設定	67
4.6.3.2. 序列類型判斷	67
4.6.3.3. 直接擷取腳點	70
4.6.3.4. 討論	71
4.7 小結	71

第五章、語法結構應用於意見分析研究	73
5.1 使用構詞資訊之中文詞意見自動分析	73
5.2 使用詞間結構資訊之中文句子層次意見分析	76
第六章、總結與展望	79
參考文獻	81
附錄 A：常用譯名對照表	87
附錄 B：未使用之賓大樹庫句子清單	88



表目錄

表 3-1 各中文語料庫詞彙總量統計 (%)	12
表 3-2 各家構詞分類法對照表	16
表 3-3 詞彙過濾規則	19
表 3-4 二字詞語料集標記結果分佈統計	20
表 3-5 二字詞標記者一致性及效能分析	21
表 3-6 二字詞標記者間一致性分析	22
表 3-7 各家構詞分類分佈統計 (%)	25
表 3-8 基本特徵值組 (以「好 (ㄏㄠˇ)」為例)	31
表 3-9 聲調特徵值 (以「好 (ㄏㄠˇ)」為例)	32
表 3-10 詞首特徵值組 (以「好 (ㄏㄠˇ)」為例)	33
表 3-11 詞尾特徵值組 (以「好 (ㄏㄠˇ)」為例)	34
表 3-12 完整特徵值範例 (以「好 (ㄏㄠˇ)」已知讀音狀況為例)	35
表 3-13 機率公式各項說明	37
表 3-14 詞性組合與構詞分類對照表	38
表 3-15 自動分類於精簡集 (6187 詞) 上之實驗結果	40
表 3-16 精簡集 (6187 詞) 與強化精簡集 (8186) 效能比較	41
表 4-1 意見句與非意見句依存關係分布比較	48
表 4-2 詞間結構標記統計表	58
表 4-3 意見句中表達意見之依存關係比例	62
表 4-4 序列結構辨識評估	68
表 4-5 序列結構類別辨識評估	69
表 4-6 腳點位置直接辨認評估	70
表 5-1 套用構詞資訊之意見詞極性判斷評估	76

表 5-2 套用詞間結構資訊之意見句判斷評估.....	77
-----------------------------	----



圖目錄

圖 2-1 現代漢語語法信息辭典樹狀結構圖	5
圖 2-2 《現代漢語新詞語信息辭典》結構圖	7
圖 3-1 中文字、詞與詞素之關係	12
圖 3-2 二字詞結構分類語料標記流程圖	18
圖 3-3 二字詞語料集標記結果分佈統計圖	20
圖 3-4 二字詞標記者間一致性測試	22
圖 3-5 各家構詞分類分佈統計圖（僅列詞彙量超過 5000 者）	26
圖 3-6 《教育部國語辭典》一般詞語條目樣式（以「科學」為例）	28
圖 3-7 《教育部國語辭典》單字資料條目實例（以「好（ㄠㄨ）」為例）	29
圖 3-8 《教育部國語辭典》單字資料條目實例（以「好（ㄠㄨ）」為例）	29
圖 3-9 二字詞自動分類器與標記者平均效能比較圖	41
圖 3-10 精簡集、強化精簡集與標記者平均效能比較	42
圖 4-1 意見段落與結構關係圖	54
圖 4-2 三角單元實例	56
圖 4-3 「潘恩標記系統」介面	57
圖 4-4 各種傾向之意見句中意見結構分佈比較	58
圖 4-5 各種類型意見句中意見結構分佈比較	59
圖 4-6 各種意見結構之意見句傾向分佈比較	59
圖 4-7 各種意見結構之意見句類型分佈比較	59
圖 4-8 標記語料轉換為依存關係範例（轉換前）	61
圖 4-9 標記語料轉換為依存關係範例（轉換後）	62
圖 4-10 同親結構範例	65
圖 4-11 非同親結構範例	66

圖 4-12 詞間關係自動擷取特徵值示意圖	67
圖 5-1 詞間結構使用方法範例	77



第一章、緒論

1.1 研究動機

文本意見分析為近年日漸興起之計算語言學研究議題，該問題主要目的為「判斷人類文本中之意見」，為全面了解「意見內容」，該問題又細分為：意見句判斷、「意見持有者」(opinion holder) 擷取、「意見目標」(opinion target) 擷取、「意見極性」(polarity) 辨別，乃至「意見句類型」自動分類等諸多不同範疇與對象之問題；而若以文本大小區分之，則略可分為文件層次 (document level)、句子層次 (sentence level)、詞彙層次 (word level) 等不同層級之問題。而在研究方法上，主要可分為機器學習方法與經驗方法兩類。前者抽取適用之特徵值後繼而以機器學習演算法訓練並預測意見內容；而後者則以語言學知識為基礎，設計計算公式，逐步模擬各層次語義內容相互結合的行為，從而判斷意見傾向。本研究所依循並欲改善之對象為後者。

以經驗方法進行中文文本意見分析，以 (Ku, Liang et al. 2006) 為代表，其方法為以機率模型計算「漢字」之意見極性分數後直接以加法將分數加總，輔以否定子之變號作用，得出意見詞之意見分數；而於句子層次與文件層次其概念亦同，即以意見詞分數為基礎進行加法運算，從而得出句子及文件之意見分數。然以加法為基礎之運算方式卻遭遇了瓶頸，試舉例說明如下：

就詞彙層次而言，如「抗菌」一詞，該詞彙之意見傾向為正向，然若以字彙之意見分數加總計算，「抗」為負面字、「菌」亦為負面字，直接加總後自然得出「抗菌」為負面詞之結果；而於句子層次亦有類似問題，如「他是僥倖獲勝的」一句，顯然為負面意見句，然若以句內意見詞分數直接加總，「獲勝」為一分數極高之正面意見詞，而「僥倖」固為負面意見詞，強度卻僅為一般程度（畢竟「僥

倖」通常意味著「雖然驚險但已成功了」)，導致全句分數加總時「獲勝」之正向分數將蓋過「僥倖」之負向分數，而得出該句為正面意見句之結論。

此問題乃肇因於「加法」並非語義結合的真正方式，欲解決此問題必須引入「語法資訊」。如若能夠理解「抗菌」一詞之構成中，「抗」為動詞、「菌」為受詞，則或可設計變號公式，從而得出該詞為正向詞之結果；而句子層次亦若是，若能理解「僥倖」乃是修飾「獲勝」之副詞，則或可設計適當公式而得出合理的結果。此即本研究之出發動機，意欲尋找適用於意見分析之詞內與詞間語法結構，藉以改善經驗方法施用於中文意見分析之效能。

1.2 研究目的

本研究之主旨在於尋找適用於中文意見分析之語法結構資訊。無論於詞內或詞間(即一般所稱之「語法結構」)此問題可進一步細分為三大部份：有哪些結構、分別之定義為何？該結構之實際分佈、使用狀態如何？又該如何預測？亦即「結構定義」、「語料分析」及「自動預測(擷取)」三主要問題。本研究於對象上切分為「詞內」及「詞間」兩大部分，二部份均依此順序逐步展開，由結構定義(分類)談起，明確定義後進行語料標記與標記結果分析，最後試圖對標記結果進行預測。而其最終目的在於將此資訊引入現有之意見分析系統中，以期能改善意見分析之效能。

1.3 論文架構

本論文主要分為「詞內」與「詞間」結構兩大部分。第1章為緒論，簡介本研究全貌；第2章為文獻探討，將詞內與詞間結構之相關研究一併列舉之；第3章針對詞內結構進行探討、第4章則為詞間結構；第5章則實際將前兩章所獲得之語法資訊套用至意見分析系統中，觀察其效能；第6章進行總結，而參考文獻將列舉於第7章。

第二章、文獻探討

本章將討論意見分析與詞內、外結構之相關研究。若未特別提及，本章所述之相關研究均指「資訊科學」領域而言，而非傳統語言學或中國文學。而漢語構詞分類歧異問題將於 3.2 節討論，各研究團隊所公佈之詞彙分布狀況統計將表列於 3.3.2 節，於此章便不細究各家構詞類別的異同與實際分佈狀況。

2.1 中文文本意見分析

2.1.1. 以經驗方法為基礎的中文文本意見分析

文本意見分析大致主要可分為「經驗方法」(heuristic) 與「機器學習方法」(machine learning) 兩大類，本研究所依循並意圖改善者為以「經驗方法」為主之意見分析研究。該類方法以語言學知識為基礎，試圖自詞彙、句子乃至文件層次，由下而上，逐步建構出簡潔而有效的意見分析模型。可將 (Ku, Wu et al. 2005) 視為發軔之嘗試，該研究提出以漢字為基礎的意見分析模型，從而擘畫出詞彙、句子、文件層次之意見分析架構，並提出了一套語料建置的方法。此時意見分析問題尚處於發展初期，並未進一步探究意見強度與傾向問題，而僅對「是否含有意見」進行評估；隔年發表之 (Ku, Liang et al. 2006) 則將意見分析應用於部落格語料上；再次年發表 (Ku, Lo et al. 2007)，其意見分析架構已臻成熟，提出更為完整之意見句標記與評估方法，將本僅由一人標記之意見答案擴增為可整合三人之標記結果，影響所及，除可得到更精確之標記結果外，亦可進一步探討意見強度與標記結果信度之問題，為意見分析問題逐漸發展成熟之指標。

2.1.2. 語法結構資訊於中文文本意見分析之應用


將語法結構資訊應用於中文意見分析方面，詞內層次尚未有相關研究發表；

詞間層次則有浙江大學所發表之（Qiu, Liu et al. 2007）與（Qiu, Wang et al. 2008）兩篇。二研究均在「意見詞已知」之情況下展開，前者藉辨識 6 種特定的依存樹（dependency tree）結構找出「意見主題」（topic）；後者則利用依存關係樹中副詞語尾「地」與形容詞語尾「的」兩種結構¹，進行意見傾向的「變號」。吾人不難發現此二者均為語法結構資訊之初步應用，既未影響意見分數之絕對值結果（即強度結果），能施用之情況亦相當有限。然由此可看出，以經驗方法為基礎之中文意見分析研究，世界各地之團隊均遭遇了語義資訊有限的瓶頸，並逐漸開始嘗試將語法資訊引入系統中。本研究即試圖對此問題提出一宏觀之解答。

2.2 中文語法結構研究及其自動剖析

2.2.1. 詞內層次

2.2.1.1. 中國大陸地區



自然語言處理領域中對漢語構詞問題著力最甚者莫過於中國大陸，早在 1986 年始籌劃之《現代漢語語法信息辭典》中便包含了詞彙內部結構的資料欄位²，此後多部語素資料庫與構詞資料庫亦紛紛在國家級科學基金挹注下展開編纂，為構詞問題提供極可觀之研究資源。其成果固多非發表於國際期刊及研討會，但有鑒其成果之豐、規模之鉅，不可輕忽其代表性，故特闢專節討論之。本節將以中國大陸地區長期關注漢語構詞問題之三個主要團隊為軸，簡介各團隊之研究成果，並比較其與本論文在目的、範疇、方法上之異同。考量此三團隊相似之研究脈絡：均以語料庫建構為發軔，繼而就其所建之語料庫展開分析，故本節在介紹各團隊時，均以其所建構之語料庫為主。

¹ 該研究使用哈爾濱工業大學信息檢索研究室所開發之語法分析套件。可參考：<http://ir.hit.edu.cn/demo/lt/>

² 如離合詞分庫下之「結構」欄位。

2.2.1.1.1. 北京大學（俞士汶、朱學鋒等）

● 語料庫：《現代漢語語法信息辭典》、《現代漢語語素庫》、《現代漢語合成詞結構數據庫》

《現代漢語語法信息辭典》為北京大學計算語言研究所自 1986 起以十餘年人力物力所編纂之大型電子辭典。該辭典遵循朱德熙先生提出之「詞組本位語法」精神（亢世勇 2001），其編輯宗旨並不在收錄大量詞彙（至 2004 年止，該辭典包含詞彙數約 7.3 萬左右，尚不及《教育部國語辭典》詞條數之一半），而在於盡可能收錄大量「組成短語或新詞」的「詞部件」（包括語素、詞或固定短語），並詳細標註其構詞能力及組合規則，從而成為一個包含「詞部件資訊」與「構詞知識」的語料庫（王惠 and 朱學鋒 1994）。該詞典將漢語詞彙分為 26 個詞類³，合有 32 個資料庫：總庫 1 個、各類詞庫 23 個（嘆詞、擬聲辭、非語素字不獨立建庫），代詞下又設有「人稱代詞」、「指示／疑問代詞」2 分庫，動詞下則有「體賓動詞」、「謂賓動詞」、「雙賓動詞」、「動結式」、「動趨式」、「離合詞」6 個分庫（朱學鋒，俞士汶 et al. 1995），可表為一樹狀結構圖：

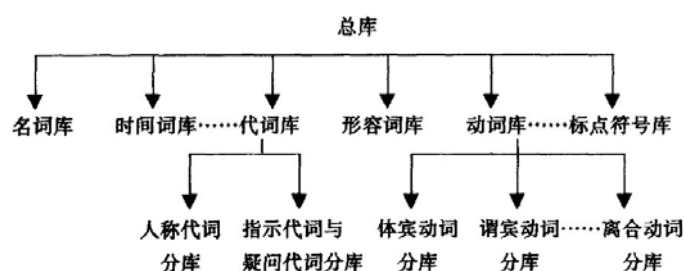


圖 2-1 現代漢語語法信息辭典樹狀結構圖（李普霞 and 劉雲 2004）

³ 18 個「基本詞類」：名詞 (n)、時間詞(t)、處所詞(s)、方位詞 (f)、數詞 (m)、量詞(q)、區別詞 (b)、代詞 (b)、動詞 (v)、形容詞 (a)、狀態詞 (z)、副詞 (d)、介詞 (p)、連詞 (c)、助詞 (u)、語氣詞 (y)、擬聲詞 (o)、嘆詞 (e)，此外也收錄了一部份較基本詞類為大的單位：成語 (i)、習用語 (l)、簡稱略語 (j)，以及一些較小的單位：前接成分 (h)、後接成分 (k)、語素字 (g)、非語素字 (x)、中文的標點符號 (w)，共 26 個詞類。

在語法辭典初步完成後，為深入研究未知詞辨識問題，1999 年北京大學計算語言研究所針對 GB/T2312-1980 下的全部漢字建立了一個單音節的「語素庫」。每一筆記錄均包含漢字、讀音、類別、同形、組合、位置、姓、人名、地名、水名、書面、方古、義項、備註等欄位，合有 7223 筆記錄。語素庫完成後，更進一步與《現代漢語語法信息辭典》集成，將語法辭典中全部詞條以「成份語素」為索引重新排序（如此雙語素詞便會擴充為兩筆紀錄、三語素詞為三筆），成為一更完備的漢語知識庫（朱學鋒，俞士汶 et al. 1999；俞士汶，朱學鋒 et al. 1999；俞士汶，朱學鋒 et al. 2001）。

而後於 2000 年，（劉雲，俞士汶 et al. 2000）進一步將《現代漢語語法信息辭典》中的 39370 個二、三音節詞取出（不包含人名、地名），標註詞語、讀音、詞類、同形、構詞、義項、備註、層次、前字／後字等屬性，建立了《現代漢語合成詞結構數據庫》。

2.2.1.1.2. 清華大學（苑春法、黃昌寧等）

● 語料庫：《漢語語素數據庫》

《漢語語素數據庫》為北京清華大學於 1997 年所完成之大型資料庫，該資料庫可概分為兩部份，一是「漢語語素」，二是「由語素所構成之詞」（簡稱「語素所構詞」）。

漢語語素方面，該資料庫定義「語素」為「音義結合的最小單位」，即只要「音」或「本義」中有一者相異，便獨立成為一「語素」（若音義相同但字型不同，原則上視為同一語素）；而考量語用之情況，同一「本義」之語素在文本中或會產生「引申義」或「比喻義」，故每一「語素」下又有「語素項」，茲舉該語素之所有可能義項。「語素項」即為該資料庫的最小錄單位（entry），每一語素項均標注意義、類別、成詞／不成詞／半成詞、前位／中位／後位／不定位等資訊。合錄有語素

10442 個、語素項 17470 個。

語素所構詞方面，該資料庫蒐集由漢語語素組成之二、三、四字詞，每個詞均標註詞型、讀音、詞類、構詞方式、類序、多義、字義組合等資訊。在刪除重覆詞彙後，合有二字詞 45960 筆、三字詞 3930 筆、四字詞 4820 筆。

（苑春法 and 黃昌寧 1998）以基因演算法於語料庫中學習出最主要之構詞原理，並將結果與語言學知識對照，而得到一定程度的肯認。

2.2.1.1.3. 魯東大學（2006 年前原山東煙臺師範學院）（亢世勇等）

● 語料庫：《現代漢語新詞語信息辭典》、《現代漢語新詞語構詞法數據庫》、《現代漢語語義構詞數據庫》

為創建漢語新詞語研究之基礎平臺，山東煙臺師範學院於 1999 年起展開《現代漢語新詞語信息辭典》的編纂。以盡量蒐集 1978 年後產生之「新詞語」為目標，參考《現代漢語語法信息辭典》之架構，該辭典目前已收納近 40000 個新詞語。除語法辭典固有欄位外，每筆新詞另標註有產生途徑、應用領域、來源、時間等新詞語資訊，以及構詞法資訊（如：單音／多音、單純詞／合成詞、聯合／偏正／補充／動賓／主謂／補充），以便對產生新詞之構詞法進行研究。該辭典之結構亦可以表為一樹狀圖：

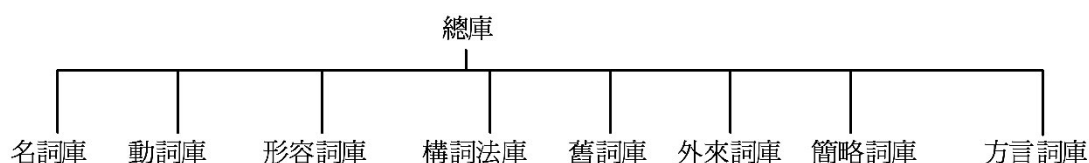


圖 2-2 《現代漢語新詞語信息辭典》結構圖（亢世勇 2002）

其中「構詞法庫」詳細標記了每一詞語之構詞部件、構詞法與詞性資訊，為

新詞構詞研究提供了極充分之語料(亢世勇 2001; 亢世勇 2002)。而後(亢世勇 2003)又自該辭典中挑選出兩萬多個詞語另編為《新詞語大辭典》，供一般語言學研究之用。

其研究成果方面，(亢世勇，徐豔華 et al. 2005)對產生新詞彙之構詞法進行統計研究；而(亢世勇，許小星 et al. 2005)則以現代漢語語義構詞數據庫進行構詞原理之探討。

然如(傅愛平 2003)所陳，以上諸多研究所習得之構詞律卻鮮少直接應用在未知詞辨識上，是以後續中國大陸諸多學者均將目光轉往「語義結合」而非「語法結合」之思維處理構詞問題，而不繼續於構詞領域著墨。

2.2.1.2. 國際研討會

於國際研討會中發表之研究成果中，對此問題較著心力者有(Tseng and Chen 2002)及(Lu, Asahara et al. 2008)。前者於(Tseng and Chen 2002)中首先提出「自動構詞分析器」，以規則方法為基礎，對未知詞之構詞類別進行自動分類；繼而又於(Tseng, Jurafsky et al. 2005)中提出可幫助辨識未知詞構詞類別之特徵值組，並以「最大化熵馬可夫模型」(Maximum Entropy Markov Models, MEMM)進行預測實驗。其自動構詞分析器於未知詞上可達80%之正確率，而其MEMM實驗可達平均90%之正確率。然綜觀其研究，可發覺其研究目的與對象均聚焦於「未知詞」上，其所使用之特徵值(或規則)多基於「構詞部件常為已知詞」此一假設而來。如「攝影展」之構詞部件為「攝影」及「展」，此二者均為已知詞，可透過外部語料獲取大量資訊，藉以判斷構詞類別。然「未知詞」固亦為本研究所涵蓋之對象，但本研研究所欲處理之問題更全面，除未知詞外亦包括更多已知詞，即該研究所使用之未知詞特徵(多由已知詞構成)本研究無法沿用；而另一相關

研究者為 (Lu, Asahara et al. 2008)，其碩士論文 (Lu 2008) 亦有可觀處。其主要貢獻在於定義了一適合計算語言學之漢語合成詞分類架構，及提出以樹狀結構分析長詞構詞資訊之概念。該研究並利用互訊息 (Mutual Information, MI) 及 SVM 進行初步實驗，獲致 94% 之準確度。然而，該研究視二字詞為一不可分割之最小語義單位，故其研究之主要對象為三字以上之長詞。而本研究之目的在於改善實際意見分析系統，而二字詞又為漢語多字詞之最主要成分 (此於 3.1 節中將詳述)，其重要性自不可輕忽，故該研究內容雖於本論文有啟發性作用，但實際研究範疇仍是相異的。

2.2.2. 詞間層次

2.2.2.1. 賓州大學樹庫 (Penn Treebank) 5.1 版

賓州大學樹庫 (Penn Treebank，簡稱賓大樹庫) 為一以人工斷詞、標記語法分析樹之大型語料庫，有英文及簡體中文版本。其簡體中文 5.1 版合計含有 890 份文件，每份文件均標示一 FID，而每一文件中之每一句子均標有 SID，共有 18782 句，為一極具可靠性且廣為計算語言學界所使用之大型語料庫。

2.2.2.2. 依存關係樹

依存關係樹由許多「依存關係」(dependency relation) 所構成；「依存關係」為兩詞彙間之關係，其精神在於將詞彙間之語義連結視為「掌權者」與「依賴者」之位階關係，從而將一句中之兩兩詞彙以「依存關係」連接起來，構成一依存關係樹。依存關係樹常由語法分析樹轉換而來。

2.2.2.3. 史丹佛語法分析套件

由美國史丹佛大學所開發之開放原始碼語法分析套件。其遵循賓大樹庫所制

定之文法標準，可將原始文本解析為語法分析樹，亦可將符合賓大樹庫格式之語法分析樹轉換為依存關係樹。為一免費且可靠之語法分析套件。該語法分析套件中，對英文共訂有 48 種依存關係，對中文則有 46 種。



第三章、中文詞內部語法結構自動分類

3.1 問題敘述

本段研究旨在為「意見詞分析」提供可用之詞內語法結構資訊。

欲深究此問題，首先當談漢語中「字」、「詞」與「詞素」之關係。本論文所指之詞素，乃指「語言中攜帶意義的最小單位」。意即語素可組成更大的語言表意單位（如詞彙、詞組、句子、文章），卻無法拆解為更小且均帶有意義的語言單位。如「山」字，可視為一語素，由於其本身即含有完整之意義，且無法繼續拆解，故為漢語中攜帶意義的最小單位；而如「蝴蝶」，亦為一語素，由於該詞若拆為「蝴」與「蝶」，則「蝴」字本身將無法表意（即「蝴」字並非語素），是以我們將「蝴蝶」視為一語素，在詞彙層次中則稱之為「單語素詞」。

漢語中「詞彙」、「字」與「語素」之關係可表為圖 3-1。漢語中大部分單字皆為語素，而凡語素字皆能成詞（即單字詞），如「山」字，於斷詞時會被斷為一詞；而僅有極少部份漢字既非語素亦無法成詞，如蝴蝶的「蝴」、蜘蛛的「蜘」。此類字多屬漢語連綿詞之詞首，亦即於古漢語中為方便發音之襯字，而不帶有語義，須與另一字合成為詞方具意義。如「蜻」加「蜓」方成「蜻蜓」，獨立作「蜻」時是不帶義的。而詞彙方面，除既為字也為語素的單字詞外，多字詞部分亦可分為兩類：「單語素詞」及「多語素詞」。大部分多字詞為多語素詞（就邏輯上這亦是合理的，因漢字多為語素），即可拆解為更小的語言單位，如「咖啡機」可拆為「咖啡」與「機」，「爬山」可拆為「爬」與「山」；然亦有少部份詞彙無法拆為更小的表意單位，如翻譯詞（如「檸檬」、「沙發」、「涅槃」等）、連綿詞（如「蝴蝶」、「蜻蜓」等）、習慣語（如「降子」）等等，此類詞本身即為語素，故稱為「單語素詞」。

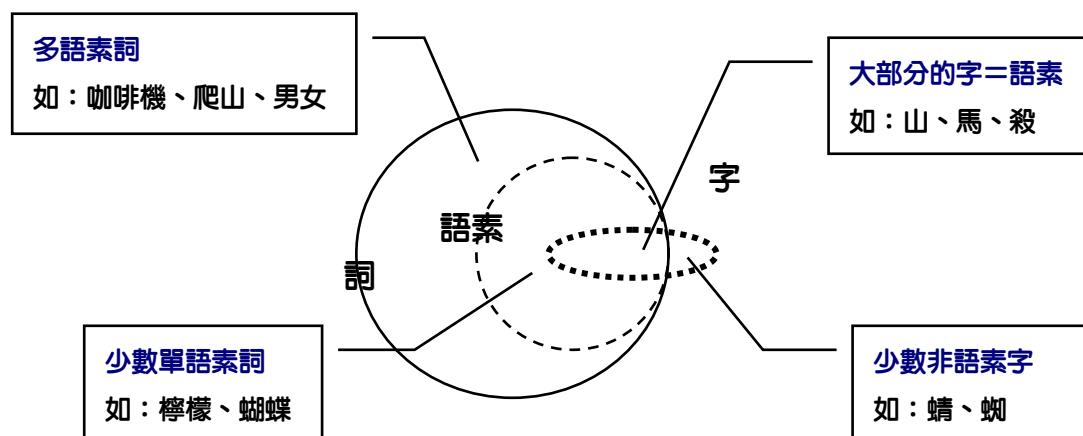


圖 3-1 中文字、詞與詞素之關係

本研究以任何「經斷詞系統所分出之詞彙」為輸入，以「該詞彙之構詞方式」為輸出。「該詞彙之構詞方式」實可分為數個子問題，包括「該詞彙是否為多語素詞」；若是，則可問「該詞彙如何斷為子語素」（更甚者可問「該斷到什麼程度」，如「高山湖」可斷為「高山／湖」，或更進一步斷為「『高／山』／湖」）；斷為組成部件後，最後才問「該數個組成部件間是何種關係」，而若將此視為分類問題，則亦有「共有哪些關係」此一分類框架之問題。而考量本研究之目的在「為意見分析提供有用之語法結構資訊」，吾人首先對三種不同性質之語料進行如表 3-1 的統計：

表 3-1 各中文語料庫詞彙總量統計（%）

詞長（字數）	1 ⁴	2	3	4	≥5
Penn Treebank 5.1	48.04	44.09	5.98	1.10	0.79
NTCIR CIRB040	51.16	42.00	5.66	0.93	0.24
新華新詞語辭典	-	58.7	17.6	14.7	8.7

表 3-1 中所分析之三個項目分別為不同性質語料之代表：賓大樹庫（Penn Treebank 5.1）為一由人工斷詞之大型語料庫、NTCIR CIRB040 則為以機器斷詞後之大型語料庫，而新華新詞語辭典則為人工編纂之辭典。然而無論語料為何，

⁴ 包含標點符號。

均可明顯觀察得知「二字詞」為漢語多字詞之最大宗，且在數量上具有極大的優勢，若欲將構詞資訊作為後續意見分析之線索，則必須以「二字詞」為最主要研究對象。在此目的上，(Tseng and Chen 2002) 或 (Lu, Asahara et al. 2008) 均無法提供足夠之資訊。考量二字詞其數量上之壓倒性優勢，本研究首先將問題限定於「二字詞」之詞內結構擷取。限於二字詞之另一優點為：二字詞若為多語素詞，則必由兩字所組成，該二字即為語素字。如此一來吾人便可專注於構詞關係分類上，而無須考慮「斷詞」的問題。

此外，分類架構方面，本研究將參考諸多計算語言學與漢語構詞學所提出之構詞結構分類定義，討論其異同及對意見分析之適用性，訂出最宜應用於意見分析之分類綱目。而更基礎之「該詞是否為多語素詞」問題，考量「多語素詞」本為漢語詞之大宗，遂僅將「單語素詞」視為分類架構下之一項，不另作特別處理。

此外，考量本研究之對象為「所有二字詞」，亦包括未知詞在內。即非對任何輸入之二字詞均能得到詞彙外部資訊（如詞性、詞頻等），亦無法保證在實際系統中可取得前後文，是以本研究進一步將範圍縮小，禁止使用任何「詞彙層次之資訊」，限制以組成字為線索，直接判斷構詞類別。意即將問題限制於資訊最少、亦最困難之範疇。

至此，吾人可擘畫出系統之概貌：以「斷詞系統分出之二字詞」為輸入，由於無需考量斷詞方式，故僅以「該詞之構詞分類」為輸出。如輸入「男女」，則將輸出「並列」一類；而如輸入「跑步」，則將輸出「動賓」；而若輸入「檸檬」，則由於無法分割為更小單位，將輸出「其他」。構詞分類架構之細節將於次節中詳述。

3.2 二字詞內部語法結構分類及其理論歧異

本節首先介紹我們所採之二字詞內部結構分類，繼而討論目前中文自然語言處理領域對詞內結構分類方式之各家異同，藉此闡明選擇此分類架構之考量。

本研究所採之二字詞內部結構分類，原則上以（程祥徽 and 田小琳 1995）

為基準，主要分為五類，並因應意見擷取實務而稍作修改。詳述如下：

(1) 並列關係 (Parallel, 又稱聯合關係)

兩語素於「語法地位」上處於平行、平等的位置，彼此並無互相修飾之關係。而語義上則無特定的限制，有兩語素意義相近或相同者如「海洋」、「城市」、「明亮」；有意義相類、概念範疇等級相近者如「尺寸」、「牛馬」、「山海」；也有完全相對、相反者如「男女」、「買賣」、「深淺」等等。為簡化問題，疊字詞亦歸於此類，不另立為類。

(2) 修飾關係 (Substantive-Modifier, 又稱偏正關係)

第一個語素用以修飾第二個語素、第二個語素被第一個語素所修飾。若以第二個語素之詞性分述之，被修飾的對象可能是名詞性的，如「高山」、「大海」；可能是形容詞性的，如「筆直」、「雪白」、「火熱」；亦可能是動詞性的，如「狂奔」、「痛哭」、「輕視」等等。為單純化問題，此處我們限制被修飾者必為第二個語素。極少數漢語之例外，或部分由方言轉譯而來的二字詞如「人客」（「人」為主、「客」為偏），則直接歸為「其他」。

(3) 主謂關係 (Subjective-Predicate, 又稱陳述關係)

第一個語素為被陳述的對象、第二個語素為陳述語，即如句法中主詞與謂語的關係，好似一個主謂句濃縮於二字詞中。如「地震」、「火燒」、「耳熟」、「膽大」等等。

(4) 動賓關係 (Verb-Object, 又稱支配關係)

第一個語素往往為動詞性的，第二個語素則為其賓語（受詞），常為名詞性的。如「輸血」、「登陸」、「簽名」、「賣命」等等。

(5) 動補關係 (Verb-Complement, 又稱補充關係)

第一個語素帶有謂語之性質，常為動詞性或形容詞性，而後一個語素則從不同角度補充前一個語素，常為副詞性的。如「擴大」、「記住」、「標

明」、「充滿」等等。

除上述主要五類外，考量意見擷取任務之特殊性，若一語素之語義為「確認」或「取消」後方語義，功能即類似數學中之正負號。其特徵明顯、易於辨認，於意見傾向計算時亦具特殊性，故另立為類，即「肯定」與「否定」二類：

(6) 否定 (Negation)

第一個語素之語義功能為否定後方語素之語義，此語素又稱為「否定子」。常見的否定子如「非」、「否」、「不」。

(7) 肯定 (Confirmation)

第一個語素之語義在於肯定後方語素之語義，此語素又稱為「肯定子」。常見的肯定子如「有」。

以上分類架構已臻完備，下節分析中亦將顯示此七類可含括 95% 以上之二字詞。然仍有極少數例外無法為此架構所容納，如翻譯詞、俗語、簡寫，或部分虛詞如「以為」、「所以」等等。為使本研究趨於完善，另增「其他」一類：

(8) 其他 (Others)

無法歸於前七類之二字詞即屬此類，包括前綴詞（如「阿嬤」）、後綴詞（如「牛仔」）、翻譯詞（如「檸檬」）、簡稱（如「立委」）、連綿詞（如「鴛鴦」、「蝴蝶」，又稱單詞素詞）、部分虛詞與功能詞（如「而且」、「以為」、「因為」）等等。

上述之八種分類即為本研究所使用之完整分類架構。然而，構詞分類方式本為語言學研究課題之一，學說絕非獨尊一家，而是百家爭鳴，無論於漢語語言學或計算語言學領域皆有諸多學者提出其構詞分類。此處茲將本研究之分類方式與

其他研究團隊之分類作一對照，如表 3-2：

表 3-2 各家構詞分類法對照表

	數量	並列			修飾			主謂	動賓	動補	其他						
本研究	8	並列			修飾	肯定	否定	主謂	動賓	動補	其他						
現代漢語 ⁵	8	並列	重疊式		修飾			主謂	動賓	動補	附加式		其他				
亢世勇 ⁶	8	聯合			狀中		定中	主謂	動賓	補充	加前綴		加後綴				
劉雲 ⁷	8	聯合	連動		狀中	定中	名量	主謂	述賓	述補	—						
石秀雙 ⁸	5	聯合			偏正			主謂	動賓	補充	—						
穆克姪 ⁹	5	聯合			偏正			主謂	動賓	補充	—						
傅建紅 ¹⁰	6	聯合			偏正			主謂	述賓	述補	其他						
苑春法 ¹¹	16	體素聯合	謂素聯合	重疊	定中偏正	狀中偏正	量補	主謂	述賓	述補	述介	前綴	後綴	簡稱	數詞縮語	固定詞組	未注標記

表 3-2 顯示以上諸多分類方式仍大致可歸於五大基本類別；而參酌其他研究

⁵ 程祥徽 and 田小琳 (1995). 現代漢語, 三聯書店 香港.

⁶ 亢世勇, 徐豔華, et al. (2005). 基於語料庫的現代漢語新詞語構詞法統計研究. International Conference on Chinese Computing, Singapore.

⁷ 劉雲, 俞士汶, et al. (2000). 現代漢語合成詞結構數據庫. 第二屆中文電化教學國際研討會, 廣西師範大學出版社.

⁸ 石秀雙 (2007). "現代漢語雙音複合詞結構關係考察——以 z 字母下雙音複合詞為例進行分析." 晉中學院學報 2007(6): 1-8.

⁹ 穆克姪 (2008). "新雙音節複合動詞語素構詞規律研究." 現代語文 2008(12): 42-44.

¹⁰ 傅建紅 (2009). "論《現代漢語詞典》F 類雙音複合詞的結構關係." Ibid. 2009(3): 49-50.

¹¹ 苑春法 and 黃昌寧 (1998). "基於語素數據庫的漢語語素及構詞研究." 語言文字應用 1998(3): 83-88.

團隊之統計文獻亦可發現(本研究後續進行之標記分析亦得到此結果),真實詞彙中構詞分類之分布極不平衡,前三大類別幾乎可佔去八成左右的詞彙,是以若分類過細,則許多次要類別將過小,而導致分類與應用時極為困難。是以本研究仍選擇遵循現代漢語之構詞分類架構展開後續標記與預測之研究。

3.3 詞彙語料標記

提出結構分類架構後,為進一步分析各類詞彙之分佈狀況,並為後續分類器實驗提供訓練及測試語料,我們接著展開語料標記工作。6500 筆語料標記完成後,除分析各類分佈狀況外,考量許多研究團隊亦曾進行構詞分類之統計,本研究亦將標記結果與其他研究團隊之數據進行比較,分析其異同。本論文首先於 3.3.1 中描述語料標記細節,並於 3.3.2 節對標記結果進行分析。

3.3.1. 語料標記及過濾

我們首先將 NTCIR CIRB040 大型語料集以機器斷詞,繼而從中隨機抽取出 6500 個二字詞。該語料集由臺灣地區之新聞文章組成,為一繁體中文之大型語料庫。使用 CIRB040 語料,一方面符合本研究以繁體中文為基礎之出發點,另一方面機器斷詞較之人工斷詞語料(如 Penn Treebank 或中研院平衡語料庫)更貼近實際應用環境,亦更能模擬意見擷取系統之真實狀態。

本研究聘請至少六位中國文學系大學部在學學生擔任標記者,及一位甫畢業於中文系之應屆畢業生¹²擔任專家(expert),以對歧異性較大之詞彙進行判斷。所有標記者均先經過一小時語法規則講解及標記練習,方得開始正式標記語料。每位標記者須將分配得之二字詞分入 3.2 節所定義之 8 類中,標記時並未提供詞性資訊,標記者得參考相關資料(如線上字典),但為探討個人主觀之影響,彼此

¹² 本語料標記時間為暑假。

不得交談。若遇有歧義之詞彙，則由實驗設計者統一規定該詞之語義，此情況本實驗中僅有「東西」（「方位」或「物件」）與「學會」（「學生自治會」或「習得」）兩例，前者統一規定為「物件」義、後者則訂為「習得」義。

標記流程如下：首二位標記者 A、B 之標記答案若相同，則接受為標準答案；若不同，則再由第三位標記者 C 標記之，若三者中有兩答相同，則接受為標準答案；若不同，則交由前述之專家，由三答中選出一答作為標準答案。本標記法之出發點除節省成本外，其宗旨在於獲取大多標記者對詞彙之理解方式，並將之視為標準答案。此與「字典編纂」的邏輯互異，本研究並非意圖產出一語言學上之標準答案，而是針對「意見分析」此目的，試圖探求大眾讀者對辭彙理解的傾向。其流程可表為圖 3-2：

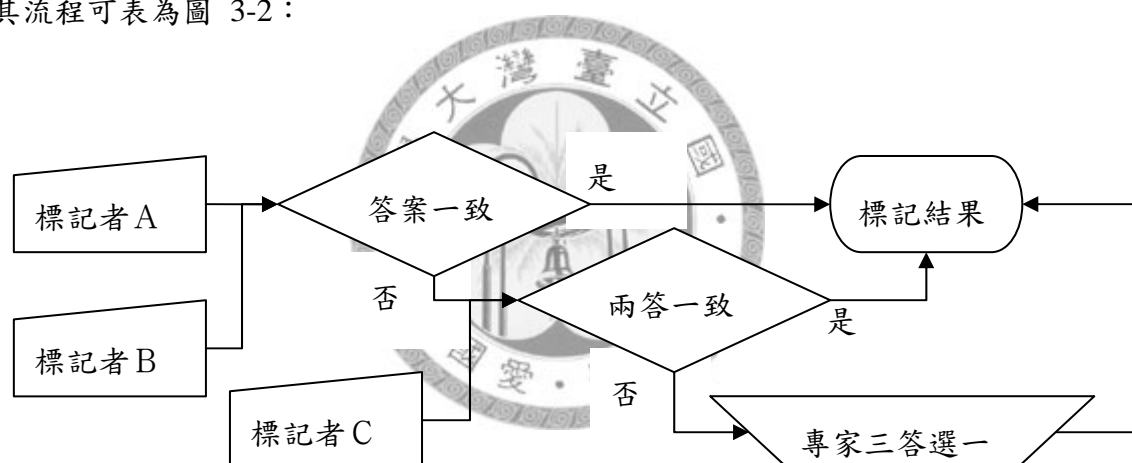


圖 3-2 二字詞結構分類語料標記流程圖

標記完成後，為使自動分類問題單純化，部分具明顯特徵、得以字串比對方式直接予以分類之詞彙先自語料集中刪除；另外，由於後續詞彙層次意見傾向判斷實驗是在（Ku, Liang et al. 2006）提供之 836 個意見詞彙上進行評估，為免投機取巧之嫌，加以 836 意見詞彙集與原始集之交集甚小，對分類效能評估影響有限，故亦將與 836 意見詞彙集重疊之詞彙刪除。上述步驟簡稱為「過濾」，過濾規則整理於表 3-3。

由此可得到一 6187 詞彙構成之較小語料集。吾人稱前述 6500 詞彙之語料集

稱為「原始集」；此較小語料集稱為「精簡集」。

表 3-3 詞彙過濾規則

特徵		刪除理由
全字特徵	疊字	可直接歸於「並列」
末字特徵	末字為「仔」	可直接歸於「其他」
首字特徵	首字為「非」、「不」、「否」	可直接歸於「否定」
	首字為「有」	可直接歸於「肯定」
	首字為「阿」、「啊」	可直接歸於「其他」
存在 Ku 之 836 意見詞彙集中		避免內部測試之嫌

另外，本研究展開初期，為進行簡單初步測試，曾聘請前述之專家，標記《教育部國語辭典》中「較易標記」之二字詞。其標記方法如下：抓取《教育部國語辭典》之全部二字詞，請專家逐一標記，若無法立刻決定分類者即直接跳過，若可立即決定者即標記之並保留。最後得到 2234 詞之標記語料。此集固無法模擬真實詞彙分佈、資料量稍嫌不足且明顯有偏 (bias)，然其於後期實驗中仍扮演了強化語料的角色。此 2234 詞彙構成之語料集稱為「簡易集」；而將「原始集」與「簡易集」合併後以表 3-3 方式過濾，可得到一 8186 詞之語料集，稱為「強化精簡集」。

3.3.2. 標記結果分析與文獻比較

本節主要分為三部份：首先針對上節產生之四個語料集進行統計分析，觀察各類分布情況；繼而為確保語料的可靠性，以進一步確保本問題之信度，對各標記者進行一致性分析；最後將結果與其他研究團隊之標記結果比較，討論其異同。

首先，經上節標記與過濾步驟後可得到四組語料集，其類別分佈狀態如表

3-4：

表 3-4 二字詞語料集標記結果分佈統計

語料集名稱	詞彙數		並列	修飾	主謂	動賓	動補	其他	肯定	否定	斷詞錯誤
原始集	6500	數量	1514	2935	85	826	704	269	43	11	113
		比例 (%)	23.29	45.15	1.31	12.71	10.83	4.14	0.66	0.17	1.74
精簡集	6187	數量	1433	2927	85	824	704	214	備註	原始集過濾後之結果。	
		比例 (%)	23.16	47.31	1.37	13.32	11.38	3.46			
簡易集	2234	數量	791	850	54	461	56	22	備註	教育部國語辭典中「較易標記」之二字詞由一人獨力標記之結果。	
		比例 (%)	35.41	38.05	2.42	20.64	2.51	0.98			
強化精簡集	8186	數量	2141	3681	134	1239	755	236	備註	原始集與簡易集合併後再行過濾之結果。	
		比例 (%)	26.15	44.97	1.64	15.14	9.22	2.88			

上述四組語料之各類分佈比例可表為圖 3-3：

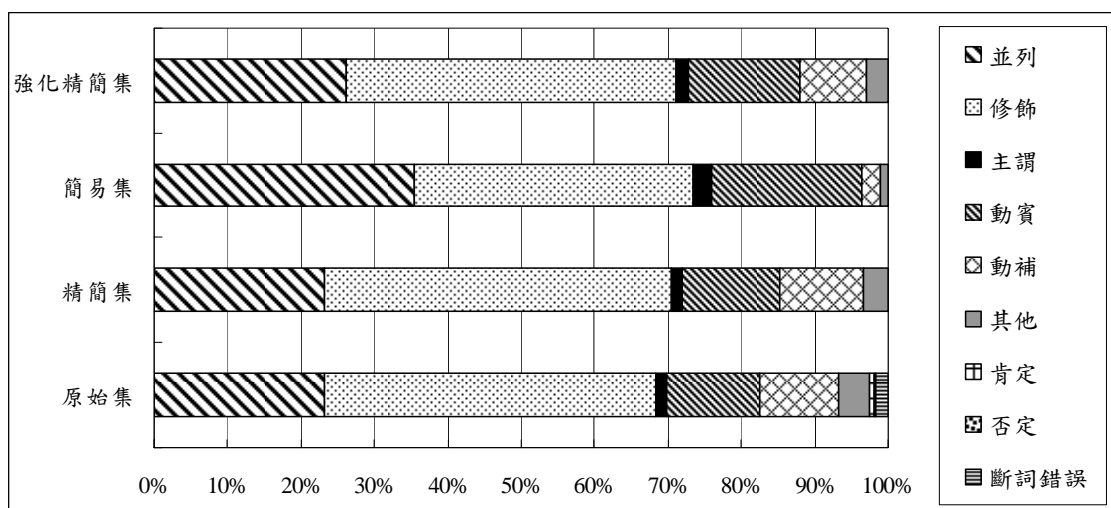


圖 3-3 二字詞語料集標記結果分佈統計圖

此類別分布結果與前述「類別分佈極不平均」之宣告吻合，原始集中前三大

類：並列、修飾、動賓，即佔總詞彙量八成以上，而剩下兩成則需分予包含斷詞錯誤在內的 6 個類別，以此而論，選擇「五加三」的分類方式的確較更細緻之分類法妥當。

其次，為確保語料之可靠性，我們對其中六位標記者（編號為 A、B、C、D、E、F）進行了一致性分析。方法如下：首先從精簡集（6187 詞）中隨機抽取 340 個詞¹³，令六位標記者標記之，標記時不得彼此交談（但可查詢資料）。標記完成後對每位標記者的標記結果計算：答題正確率、對標準答案的 Kappa 一致性係數¹⁴（ κ_a ），以及每一構詞類別的 F-分數。此一致性測試之目的在於分析讀者獨力理解一詞彙時，與全部讀者平均行為的一致性程度，就反面言，亦可觀察每一標記者的歧異程度。其結果如表 3-5：

表 3-5 二字詞標記者一致性及效能分析

標記者	κ_a	正確率	F-measure					
			其他	並列	修飾	主謂	動賓	動補
A	0.73	0.83	0.22	0.76	0.90	0.36	0.81	0.80
B	0.73	0.82	0.22	0.71	0.88	0.50	0.90	0.85
C	0.66	0.76	0.34	0.77	0.83	0.40	0.82	0.78
D	0.84	0.89	0.64	0.84	0.93	0.40	0.93	0.83
E	0.70	0.79	0.33	0.83	0.83	0.31	0.88	0.86
F	0.83	0.85	0.64	0.78	0.90	0.62	0.90	0.83
平均	0.75	0.82	0.40	0.78	0.88	0.43	0.87	0.83

同時我們亦計算兩兩標記者間之 Kappa 一致性係數，結果可見表 3-6 與圖

¹³ 母體量為 6187 時，於信賴區間 5%、信心水準 95%、母體比例 70%（即假設只答對 70%）的情況下，所需的最小樣本數為 307 個。

¹⁴ 此處使用未加權之簡單 Kappa 一致性係數公式： $\kappa = \frac{P_0 - P_c}{1 - P_c}$ ，其中 P_0 為「觀測一致性（observed agreement）」，即前後兩種測量結果一致的百分比； P_c 為「期望一致性（chance agreement）」，即前後兩種測量結果預期相同的機率。

3-4：

表 3-6 二字詞標記者間一致性分析

	A	B	C	D	E	F
A		0.79	0.62	0.70	0.68	0.68
B			0.62	0.73	0.68	0.68
C				0.61	0.66	0.63
D					0.64	0.72
E						0.66
F						

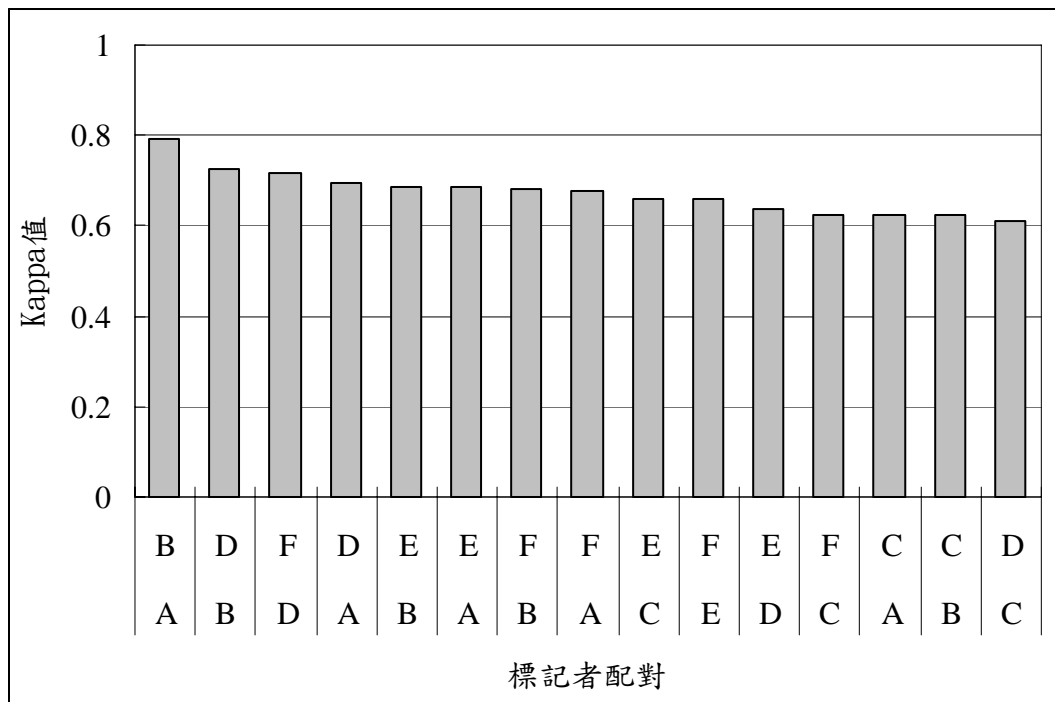


圖 3-4 二字詞標記者間一致性測試

以 Kappa 一致性係數而言，其分數大小與一致性程度關係如下：0.0-0.20 為「極低 (slight)」，0.21-0.40 為「一般 (fair)」，0.41-0.60 為「中等 (moderate)」，0.61-0.80 為「高度 (substantial)」，而 0.81-1 則為「幾乎完全吻合 (almost perfect)」。

由上述結果可看出，無論是標記者對答案之 κ_a ，抑或標記者間兩兩之 Kappa 值均落於「高度一致」之範圍（其中 κ_a 稍高一些，此極合理，緣於答案本就是以標記

者中之多數所產生的)。而標記者間之 Kappa 值亦高，考量標記者於標記時彼此不得交談，此結果代表「二字詞構詞分類」問題乃為一有信度之問題，此問題對一般讀者而言，大多時候乃是可以清楚辨別的。

然標記者間之表現仍有部份歧異性，我們對標記者答錯較多之詞彙進行分析，整理出造成歧異之四個主要原因：

(1) 字義理解之歧異

如「文物」，部分標記者認為「文物」乃指「物」而「文」為「文化、民俗」之意，卻有標記者認為「文」與「物」乃是並列關係，是「文字、字畫」與「物品」之意；又如「馴養」，部分標記者認為「馴」為「馴化」之意，乃用以修飾「養」；而有標記者認為「馴」與「養」均為動作，為並列關係。此類字義理解之歧異往往不會影響整體詞義（如「馴養」和「文物」的意義並無曖昧處），卻會影響構詞分類。

(2) 字彙詞性判斷之歧異

如「跑走」一詞，於該詞彙中「跑」與「走」之字義並無歧異，「跑」為跑之動作，而「走」則指離開某處。然於理解「詞性」時部份標記者認為「走」乃指離開的「動作」、是動詞，而標為「並列」；卻有部份標記者認為「走」乃是離開的「狀態」，是副詞性的，而標為「動補」。其他如「迎合」、「舉起」亦發生此現象。

(3) 構詞方式不明

如「政治」，此詞彙中之「政」與「治」二字字義均堪稱明確，然此二字是以何方式構成「政治」之義卻令人費解；又如「文化」、「自由」等詞，其構詞方式本就極不明確，對大多標記者而言均難以清楚回答，從而造成歧異。

(4) 對詞義認知程度之歧異

如「探花」即一例。此詞中大多標記者均標為「其他」，然卻有部份高年

級標記者將之標為「動賓」，緣於該詞彙之語源為「到各名園採摘鮮花，迎接狀元」之意；又如「睡覺」一詞，大多標記者將之標為「動賓」，然卻有部分標記者將之標為「並列」。因「覺」原為「醒」之意，「睡」與「覺」本為並列，乃因後世多誤用，因沿成習，而出現了「睡個覺」此種類於動賓之用法。

舉凡以上四點歧異原因，均不會影響對辭彙之整體理解，卻會影響構詞方式。此為漢語極幽微之處，即便一般以中文為母語、且主修中文者亦難以判斷，可將之視為以資訊方法難以駕馭的效能上界（upper bound）。

最後，如表 3-2 所述，構詞分類架構已為諸多研究者所提出，且諸家分類架構大多可歸於五大基本類別。於語料標記完成後，一可行之嘗試為：將其他研究團隊所公佈之構詞分佈統計與本研究之標記結果相互對照，便可分析其異同。我們於是將各家已公佈之構詞分部情況整理為表 3-7（見 25 頁）；若將表 3-7 中詞彙量超過 5000 之研究成果繪製為橫條比例圖，則如圖 3-5（見 26 頁）。

由圖 3-5 中可發現，本研究所得出之各類分佈狀態與其他研究者之結果無甚大差異（「亢世勇三字詞」一行「並列」明顯較少乃緣於三字詞之並列必須為三個字均處於平行地位，如「中日韓」，此例極少），唯「動補」一類明顯較大。由於該研究團隊之語料取得不易，本研究僅得推測此類較大之可能原因，可能原因有二：其一，其他研究團隊所收錄之字彙量均遠大於 6500，可能「並列」或「修飾」詞彙數量對動補造成了擠壓；其二，比較其他團隊與本研究，可發現其他研究者均以「建構辭典」或「編纂資料庫」之思維展開語料標記，唯本研究乃自大型語料庫中隨機抽取二字詞作為語料，或許辭典編纂者較不喜將「動補」一類詞彙編入。此想法看似詭怪，實則亦有理可循。動補一類之詞彙如「落下」、「離開」、「走掉」等等，多為詞義單純、字義明確之詞彙，就辭典編纂者角度而言，確無大量將之編入的必要。

表 3-7 各家構詞分類分佈統計 (%) ¹⁵

		並列		修飾			主謂	動賓	動補	其他	
原始集 (6500 詞)		並列		修飾	肯定	否定	主謂	動賓	動補	其他	
		23.70		46.80			1.33	12.93	11.02	4.21	
亢世勇 ¹⁶		聯合		狀中	定中		主謂	動賓	補充	加前綴	加後綴
	雙音節	12.73		53.4			1.3	21.2	2.1	9.27	
	三音節	0.8		67.6			1.6	10.7	0.7	18.6	
	人民日報	6.21		70.32			0.43	18.13	0.19	4.72	
劉雲 ¹⁷		聯合	連動	狀中	定中	名量	主謂	述賓	述補	-	
		28.31		59.97			0.99	12.47	2.28	0	
石秀雙 ¹⁸		聯合		偏正			主謂	動賓	補充	-	
		53.6		23.7			4.2	17.4	1.14	0	
穆克婭 ¹⁹		聯合		偏正			主謂	動賓	補充	-	
		26		10			1	56	7	0	
傅建紅 ²⁰		聯合		偏正			主謂	述賓	述補	其他	
		28.33		42.72			0.68	24.3	1.43	2.54	

¹⁵ 若該數據中有「斷詞錯誤」一類便刪去該類比例，重新標準化；而若有未寫出之分類，則勝於比例加進「其他」類中。

¹⁶ 亢世勇，徐豔華，et al. (2005). 基於語料庫的現代漢語新詞語構詞法統計研究. International Conference on Chinese Computing, Singapore.；「二字詞、三字詞」為在《新詞語構詞法數據庫》中的統計（含有雙音節詞 15751 個、三音節詞 6502 個）；「人民日報」為將 1998 年 4 月 1 日至 10 日的《人民日報》70 萬字語料，經機器斷詞後抽取所有未知詞（含有 1627 個），以人工標記的結果。

¹⁷ 劉雲，俞士汶，et al. (2000). 現代漢語合成詞結構數據庫. 第二屆中文電化教學國際研討會，廣西師範大學出版社；《漢語合成詞結構數據庫》中將 32711 個二字詞刪去單純辭、人名、地名後，餘下 32711 個合成詞的統計結果。

¹⁸ 石秀雙 (2007). "現代漢語雙音復合詞結構關係考察——以 z 字母下雙音復合詞為例進行分析." 晉中學院學報 2007(6): 1-8.；為《現代漢語辭典》中 z 字母下雙音節詞（合計 3059 個）的統計結果。

¹⁹ 穆克婭 (2008). "新雙音節複合動詞語素構詞規律研究." 現代語文 2008(12): 42-44.；《新華新詞語辭典》（共包含 2200 個詞條）中雙音節詞的統計結果。合計雙音節詞 529 個。

²⁰ 傅建紅 (2009). "論《現代漢語詞典》F 類雙音複合詞的結構關係." Ibid. 2009(3): 49-50.；為《現代漢語辭典》中 f 字母下雙音節詞（合計 1613 個）的統計結果。

	並列			修飾			主謂	動賓	動補		其他					
苑春法 ²¹	體素聯合	謂素聯合	重疊	定中偏正	狀中偏正	量補	主謂	述賓	述補	述介	前綴	後綴	簡稱	數詞縮語	固定詞組	未注標記
	21.16			53.24			0.98	18.12	2.36		4.14					

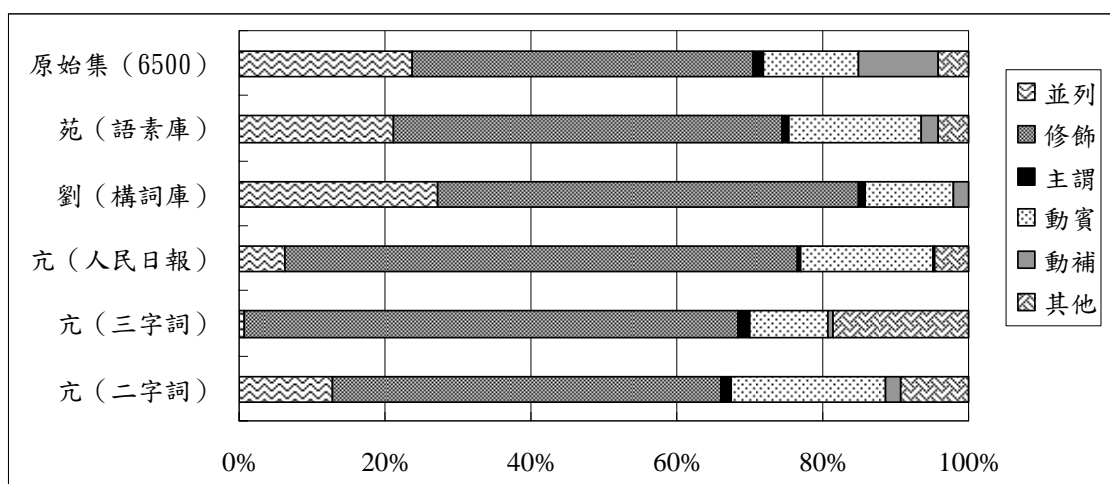


圖 3-5 各家構詞分類分佈統計圖 (僅列詞彙量超過 5000 者)

3.4 二字詞內部結構自動分類

我們已於 3.2 節提出二字詞內部結構的分類框架，並於 3.3 節中標記完成數組質量兼具之可靠語料。本節將在此基礎上，展開二字詞內部結構自動分類研究。由於(傅愛平 2003)已說明規則方法於實際應用時缺乏效度，故我們選擇以機器學習(machine learning)方法處理此問題。機器學習方法通常可分為兩部份：特徵值(feature)抽取，及以演算法自特徵值中學習從而產生分類模型(model)。3.4.1 節將介紹本研究特徵值選取之精神，及以《教育部重編國語辭典修訂本》為知識庫並從中擷取特徵值之細節；3.4.2 節則將介紹五種施用於實驗中的分類演算

²¹ 苑春法 and 黃昌寧 (1998). "基於語素數據庫的漢語語素及構詞研究." 語言文字應用 1998(3): 83-88.;《漢語語素數據庫》中 78230 個雙音節詞的統計。

法：條件隨機域模型（Conditional Random Field，簡稱 CRF）、支援向量機模型（Support Vector Machine，簡稱 SVM）、單純貝氏（Naïve Bayes）模型，以及本研究提出之兩種簡易分類法，並於 3.5 節中評估其效能。

3.4.1. 特徵值抽取

欲單就二字詞之字面（而無詞性與前後文資訊）推測其構詞分類，唯一可用之特徵值便為組成該詞之成分字的特徵。而一直覺之想法是：成分字之詞性（亦即語素之詞性）與構詞方式強烈相關。舉例而言，對「主謂」一類，若首字之詞性極可能為名詞、而末字之詞性極可能為動詞，則此詞為主謂結構之機率便相當大；而又如對「動賓」一類，若首字之詞性極可能為動詞、而末字之詞性極可能為名詞，則此詞為動賓結構之機率亦較其他詞性組合高出許多。故以此想法出發，若可取得代表每一漢字之「詞性傾向強度」之資訊，便可作為預測構詞分類之特徵值。

然需注意的是，此處所言之「詞性」乃指「該字（語素）於該詞內之詞性」，而非獨自成詞時之詞性。如「書桌」一例，「書」獨自成詞時多為名詞，然於「書桌」詞中，「書」成為「桌」之修飾語，詞性便為形容詞；又如「國家」之「國」，獨自成詞時幾全為名詞，然於詞首時則多為修飾後方詞彙之形容詞。是以必須尋找可模擬「漢字於詞彙內部時詞性傾向」之資訊，無法單純以外部語料集之詞性統計作為特徵值。然而，由於正體中文並無如簡體中文般已存在標記完整之語素或構詞資料庫，欲滿足以上之特徵值需求，便需尋找帶有語素語義、語素詞性強度、甚是構詞方法之知識庫，並試圖從中抽取出可用之特徵值。考量知識之完整性及易取得程度，本研究選擇以《教育部重編國語辭典修訂本》臺灣學術網路第四版 ver.2 作為知識庫。

3.4.1.1. 《教育部重編國語辭典修訂本》簡介

《教育部國語辭典》籌備於 1926 年，1931 年展開編輯，至 1945 年竣工；1976 至 1979 年，又以《國語辭典》為基礎進行重編，於 1981 年付印；1987 年成立專案小組進行「重編國語辭典」之修訂工作，於 1994 年修訂完成，並於同年完成網路版；1997 年推出光碟版。此後網路版經多次更新及修改，本研究所使用之「臺灣學術網路第四版 ver.2」更新於 2007 年 12 月（方便起見，以下簡稱為《教育部國語辭典》）。

辭典格式方面，該辭典以「形＋音」為主鍵（primary key），共有單字資料 11930 筆（含多音字）、異體字 1848 筆、詞語 152398 筆（含多音詞），合計 166176 筆記錄。其規格說明中特將「單字資料」與「詞語」分開列目，乃緣於該辭典對「單字」提供了較「一般詞彙」更形豐富之資料，此亦即本研究選擇該辭典為單字特徵值來源知識庫的最主要原因。

對一般詞彙（非單字詞）而言，該辭典提供該詞之「注音（一式、二式）」、「漢語拼音」及「詞義」（詞義後常有一到數句不等之例句）資訊。其中若有多義者，則依序編號清列之。如圖 3-6：

1. 科學	
注音一式 ㄔㄨㄛˊ ㄒㄩㄝˊ	
漢語拼音 kē xué	注音二式 kē shi ué
① 以一定對象為研究範圍，依據實驗與邏輯推理，求得統一、確實的客觀規律和真理。有廣義與狹義之別。廣義泛指一切有組織、有系統的知識而言，可分自然科學、應用科學、社會科學、人文科學四大類。狹義則專指自然科學而言。	
② 合乎科學精神和方法的。如：「警察辦案的方式越來越科學。」	

圖 3-6 《教育部國語辭典》一般詞語條目樣式（以「科學」為例）

對「單字資料」而言，該辭典除「注音（一式、二式）」、「漢語拼音」外，該字之所有義項均以「詞性」分門別類，再依序編號。該辭典共定義九種詞性：名詞、形容詞、動詞、副詞、助詞、連接詞、介詞、代詞、歎詞。而每條字義後常

有一到數句不等之「例句」或「例詞」。如圖 3-7：

1. 好 部首 女 部首外筆畫 3 總筆畫 6	
注音一式 ㄏㄠˇ	
漢語拼音 hǎo	注音二式 hǎo
<p>㊦ 美、善，理想的。如：「好東西」、「好風景」、「花好月圓」、「好人好事」。唐•韋莊•菩薩蠻•人人盡說江南好詞：「人人盡說江南好，遊人只合江南老。」</p> <p>㊦ 友愛的。如：「好朋友」、「好同學」。</p> <p>㊦ 完整的、沒壞的。如：「完好如初」、「修好了。」</p> <p>㊦ 相善、彼此親愛。如：「友好」。唐•高適•贈別晉三處士詩：「知己從來不易知，慕君為人與君好。」紅樓夢•第二十七回：「誰和我好，我就和誰好。」</p> <p>㊦ 痊癒。如：「病好了！」警世通言•卷十六•小夫人金錢贈年少：「孩兒感些風寒，這幾日身子不快，來不得。傳語員外得知，一好便來。」</p> <p>㊦ 很、非常。表示程度深。如：「好久」、「好冷」、「好笨」、「好厲害」。</p> <p>㊦ 完成、完畢。如：「交待的工作做好了。」、「稿子寫好了。」儒林外史•第四十三回：「都梳好了椎髻，穿好了苗錦。」</p> <p>㊦ 容易。如：「這事好辦。」、「這問題好解決。」、「這小孩好帶。」</p> <p>㊦ 以便、便於。如：「快準備行李，好早點上路。」、「請告訴我你的住處，我好去找你。」唐•杜甫•聞官軍收河南河北詩：「白日放歌須縱酒，青春作伴好還鄉。」</p> <p>㊦ 可以、應該。如：「只好如此」、「正好試試」。官場現形記•第五十一回：「刁邁彭屈指一算，後任明天好到，便約張太太三天回音。」</p> <p>㊦ 置於某些動詞之前，表效果佳。如：「好看」、「好玩」、「好吃」、「好笑」。</p> <p>㊦ 置於數量詞或時間詞之前，表示多或久的意思。如：「好些個」、「好幾處」、「好半天」、「好一會兒」。</p> <p>㊦ 表示稱讚或允許。如：「好！就這麼辦。」京本通俗小說•碾玉觀音：「郡王道：『好！正合我意。』」</p> <p>㊦ 表示責備或不滿意的語氣。如：「好！這下子事情愈來愈棘手了。」</p>	
ㄏㄠˇ、hǎo (04802)	

圖 3-7 《教育部國語辭典》單字資料條目實例（以「好（ㄏㄠˇ）」為例）

如前所述，該字典之主鍵為「形＋音」。故對同一字而言，若有多種相異發音，則每種發音均獨立列目。以「好」為例，讀「ㄏㄠˇ」時條目如圖 3-7，而若讀作「ㄏㄠˋ」時則有另一條目，見圖 3-8：

2. 好 部首 女 部首外筆畫 3 總筆畫 6	
注音一式 ㄏㄠˋ	
漢語拼音 hào	注音二式 hào
<p>㊦ 愛、喜愛。如：「潔身自好」、「好逸惡勞」、「好學不倦」。唐•王維•終南別業詩：「中歲頗好道，晚家南山陲。」唐•韓愈•師說：「李氏子蟠，年十七，好古文。」</p> <p>㊦ 心中所喜愛的事。如：「投其所好」。史記•卷六十一•伯夷傳：「（富貴）如不可求，從吾所好。」</p> <p>㊦ 舊指玉器中的孔。周禮•冬官考工記•玉人：「璧義度尺，好三寸以為度。」鄭玄•注：「好，璧孔也。」</p>	
ㄏㄠˋ、hào (04800)	

圖 3-8 《教育部國語辭典》單字資料條目實例（以「好（ㄏㄠˋ）」為例）

觀察以上實例，可歸結出幾點結論：首先，教育部國語辭典中於單字條目下所列之「詞性」並非特指該字單獨成詞時之詞性，亦包括該字於一般詞彙中之詞性；其次，中文字之字義會因發音之不同而相異，其詞性傾向亦若是。以「好」字為例，讀作「ㄏㄠˇ」時傾向形容詞性與副詞性，而讀作「ㄏㄠˋ」時則傾向動詞與名詞性；最後，大致而言，該字較傾向之詞性，則列於該詞性下之義項數亦較多。如「好(ㄏㄠˇ)」便無名詞性之義項，且形容詞及副詞下之義項最多。

以此諸概念為基礎，便可自教育部國語辭典中抽取出適當之特徵值模擬「某字之詞性傾向」了。細節將於次節中詳述。

3.4.1.2. 使用之特徵值

由於本研究之問題範疇限制為「不使用詞外資訊」，故此節介紹之特徵值皆為單一「中文字」之特徵值。實際於機器學習時之特徵值應用方法（如兩字之特徵值直接合併、推導為機率，或依順序排列等等）則依所使用之演算法而各不相同，此將於 3.4.2 節中詳述。

本研究自《教育部國語辭典》中抽取特徵值之最主要概念為：「以該詞性下之『義項數』模擬該字於詞彙中為該詞性之傾向」。由於教育部國語辭典所定義之詞性共有 9 種，意即「各詞性下之義項數」為 9 個非負實數，也就是最基本的 9 個特徵值，稱為「基本特徵值組」。表 3-8 以「好(ㄏㄠˇ)」為例，詳細示範基本特徵值組之算法與意義（見 31 頁）（實際辭典內容請參考圖 3-7）。

其次，我們觀察發現，部分漢字於二字詞中之詞性傾向亦與「該字於詞彙中之位置」有關。舉例而言，「戲」若為某詞之尾字，如「看戲」、「排戲」、「聽戲」，則詞性傾向於名詞；但若位於首字，如「戲弄」、「戲耍」之「戲」為副詞，「戲子」、「戲精」中則為形容詞，即無強烈的名詞性；又如「好(ㄏㄠˇ)」字做動詞用時幾乎不會出現首字而多為尾字，如「友好」、「病好」。欲以特徵值呈現此現象，應

需一大型構詞資料庫，以統計每一漢字之各詞性的位置傾向（如動詞傾向於詞首、名詞傾向於詞尾等等）。然如前所述，正體中文缺乏標記完善之大型構詞語料庫。即便以本研究所標之八千餘二字詞，分予九千餘漢字、每一字下又有 9 種詞性，亦將造成極嚴重之資料空缺（data sparse）問題。

表 3-8 基本特徵值組（以「好（ㄏㄠˇ）」為例）

特徵值說明			
內容	各詞性下之義項數		
值域	大於等於 0 之整數		
意義	某字較傾向之詞性，則列於該詞性下之義項數亦較多，故可以詞性下之義項數模擬該字之詞性傾向。		
實例	詞性	特徵值	詳列
	名	0	-
	形	3	(1) 美、善，理想的。 (2) 友愛的。 (3) 完整的、沒壞的。
	動	2	(1) 相善、彼此親愛。 (2) 痊癒。
	副	7	(1) 很、非常。 (2) 完成、完畢。 (3) 容易。 (4) 以便、便於。 (5) 可以、應該。 (6) 置於某些動詞之前，表效果佳。 (7) 置於數量詞或時間詞之前，表示多或久的意思。
	助	0	-
	連	0	-
	介	0	-
	代	0	-
	歎	2	(1) 表示稱讚或允許。 (2) 表示責備或不滿意的語氣。

而觀察《教育部國語辭典》，可發覺其每一漢字條目下之每一詞條常有「例詞」。我們或可做一假設：若某字為某詞性時通常位於詞首，則該辭典於該詞性下之詞條舉例時，亦較容易舉出「該字位於詞首」之詞例（此假設顯然是基於對該辭典「編纂時對所有詞條均採同一標準」之信任而來）。以此概念出發，我們遂決定以漢字「各詞性下例詞中，該字為『首字』／『尾字』的個數」為特徵值，分別稱為「詞首特徵值組」與「詞尾特徵值組」。而「例詞」之定義為「引號內無任何標點符號者」，以與「例句」分開。

然我們亦觀察到，部分例詞之詞首或詞尾字並不具有代表性，如「家（ㄐㄧㄚˊ）」之例詞有「住戶不滿十家」，此詞彙頗長，已具有句子特性，其中之「家」顯然作一獨立詞彙之用，而非作為詞彙內部之語素；又如「向」之例詞，「向有研究」與「向前看」等，其中之「向」似亦因三、四字詞彙稍有句子特性而作獨立詞彙使用，與二字詞內之「志向」、「方向」用法殊異。為減低此影響，本研究將「詞首／詞尾特徵值組」所指涉之例詞先刪去五字以上者，再以「例詞之詞長」分為二、三、四字詞三個子類。仍以「好（ㄏㄠˇ）」為例，最後完成之特徵值組細節即如表 3-10（見 33 頁）與表 3-11（見 34 頁）。

最後，依現代漢語之常識，部分單字若作動詞用時會轉聲調為四聲或三聲（古仄聲字）。如「衣（ㄧ）」作動詞用則轉為「一ˋ」；「飯（ㄈㄢˋ）」作動詞用則會轉為「ㄈㄢˇ」。故本研究亦將聲調列為特徵值之一。如表 3-9：

表 3-9 聲調特徵值（以「好（ㄏㄠˇ）」為例）

內容	聲調
值域	輕聲（•）或四聲（ ˊ ˋ ˊ ˋ ） ²²
意義	部份漢字轉品時聲調亦會改變。
實例	好（ㄏㄠˇ）：「ˊ」

²² 不另轉為實數表示，因某些機器學習工具可以字串為特徵值（如 CRF++），而非必為數字。

表 3-10 詞首特徵值組（以「好（ㄏㄠˇ）」為例）

內容	各詞性下之二、三、四字例詞中，該字為「首字」的個數																																																																																		
值域	大於等於 0 之整數																																																																																		
意義	若某字為某詞性時通常位於詞首，則該辭典於該詞性下之詞條舉例時亦較容易舉出「該字位於詞首」之詞例，可以此模擬「詞性傾向」與「該字於詞彙中之位置」之關係。																																																																																		
實例	<table border="1"> <thead> <tr> <th></th><th colspan="2">二字詞</th><th colspan="2">三字詞</th><th colspan="2">四字詞</th></tr> <tr> <th>詞性</th><th>特徵值</th><th>列舉</th><th>特徵值</th><th>列舉</th><th>特徵值</th><th>列舉</th></tr> </thead> <tbody> <tr> <td>名</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>形</td><td>0</td><td>-</td><td>4</td><td>好東西 好風景 好朋友 好同學</td><td>1</td><td>好人好事</td></tr> <tr> <td>動</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>副</td><td>7</td><td>好久 好冷 好笨 好看 好玩 好吃 好笑</td><td>4</td><td>好厲害 好些個 好幾處 好半天</td><td>1</td><td>好一會兒</td></tr> <tr> <td>助</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>連</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>介</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>代</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>歎</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> </tbody> </table>							二字詞		三字詞		四字詞		詞性	特徵值	列舉	特徵值	列舉	特徵值	列舉	名	0	-	0	-	0	-	形	0	-	4	好東西 好風景 好朋友 好同學	1	好人好事	動	0	-	0	-	0	-	副	7	好久 好冷 好笨 好看 好玩 好吃 好笑	4	好厲害 好些個 好幾處 好半天	1	好一會兒	助	0	-	0	-	0	-	連	0	-	0	-	0	-	介	0	-	0	-	0	-	代	0	-	0	-	0	-	歎	0	-	0	-	0	-
	二字詞		三字詞		四字詞																																																																														
詞性	特徵值	列舉	特徵值	列舉	特徵值	列舉																																																																													
名	0	-	0	-	0	-																																																																													
形	0	-	4	好東西 好風景 好朋友 好同學	1	好人好事																																																																													
動	0	-	0	-	0	-																																																																													
副	7	好久 好冷 好笨 好看 好玩 好吃 好笑	4	好厲害 好些個 好幾處 好半天	1	好一會兒																																																																													
助	0	-	0	-	0	-																																																																													
連	0	-	0	-	0	-																																																																													
介	0	-	0	-	0	-																																																																													
代	0	-	0	-	0	-																																																																													
歎	0	-	0	-	0	-																																																																													

以上四組便是本研究所使用之所有特徵值。然至此不免產生一問題：若遇一未知詞而無從得知各字讀音時，該如何擷取特徵值？此問題並不困難，即將辭典中該字的所有讀音之條目合併為一條計算即可。以「好」字為例，若遇一不知讀音之新詞，則於計算詞性下之義項數與例詞數時，將「好（ㄏㄠˇ）」與「好（ㄏㄠˋ）」二條目合併統計即可。然若採此法又將產生另一問題：若將同一字所有讀

音之條目合併，則義項數與例詞數均將變大，若合併後之特徵值仍與未合併之特徵值並列為訓練語料，特徵值的代表性將被破壞。是以本研究選擇兩者並陳：後續實驗中，每一漢字均將攜帶「合併所有讀音條目後之特徵值」(無論讀音是否已知)；而若該詞中之該字讀音已知²³，則再加上「已知讀音條目之特徵值」。

表 3-11 詞尾特徵值組 (以「好 (ㄏㄠˇ)」為例)

內容	各詞性下之二、三、四字例詞中，該字為「尾字」的個數																																																																																		
值域	大於等於 0 之整數																																																																																		
意義	若某字為某詞性時通常位於詞尾，則該辭典於該詞性下之詞條舉例時亦較容易舉出「該字位於詞尾」之詞例，可以此模擬「詞性傾向」與「該字於詞彙中之位置」之關係。																																																																																		
實例	<table border="1"> <thead> <tr> <th></th><th colspan="2">二字詞</th><th colspan="2">三字詞</th><th colspan="2">四字詞</th></tr> <tr> <th>詞性</th><th>特徵值</th><th>列舉</th><th>特徵值</th><th>列舉</th><th>特徵值</th><th>列舉</th></tr> </thead> <tbody> <tr> <td>名</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>形</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>動</td><td>1</td><td>友好</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>副</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>助</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>連</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>介</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>代</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> <tr> <td>歎</td><td>0</td><td>-</td><td>0</td><td>-</td><td>0</td><td>-</td></tr> </tbody> </table>							二字詞		三字詞		四字詞		詞性	特徵值	列舉	特徵值	列舉	特徵值	列舉	名	0	-	0	-	0	-	形	0	-	0	-	0	-	動	1	友好	0	-	0	-	副	0	-	0	-	0	-	助	0	-	0	-	0	-	連	0	-	0	-	0	-	介	0	-	0	-	0	-	代	0	-	0	-	0	-	歎	0	-	0	-	0	-
	二字詞		三字詞		四字詞																																																																														
詞性	特徵值	列舉	特徵值	列舉	特徵值	列舉																																																																													
名	0	-	0	-	0	-																																																																													
形	0	-	0	-	0	-																																																																													
動	1	友好	0	-	0	-																																																																													
副	0	-	0	-	0	-																																																																													
助	0	-	0	-	0	-																																																																													
連	0	-	0	-	0	-																																																																													
介	0	-	0	-	0	-																																																																													
代	0	-	0	-	0	-																																																																													
歎	0	-	0	-	0	-																																																																													

本節最後仍以「好 (ㄏㄠˇ)」字為例，假設讀音已知，將其完整特徵值詳列如表 3-12：

²³ 雖然罕有但確會發生「某詞彙之完整讀音未知，但某成分字之讀音已知」的情況。可能為該字僅有一種讀音（但此情況下也就不會有特徵值合併與否之問題），亦或我們可由「某詞為某字某讀下之例詞」而取得該字讀音，但於辭典中卻無法查得全詞讀音。

表 3-12 完整特徵值範例（以「好（ㄏㄠˇ）」已知讀音狀況為例）

	無論讀音是否已知							已知讀「ㄏㄠˇ」							聲調
	基本	例詞數						基本	例詞數						
		詞首為好			詞尾為好				詞首為好			詞尾為好			
		二	三	四	二	三	四		二	三	四	二	三	四	
名	2	0	0	0	0	0	1	0	0	0	0	0	0	ㄚ	
形	3	0	4	1	0	0	0	3	0	4	1	0	0		0
動	3	0	0	2	1	0	1	2	0	0	0	1	0		0
副	7	7	4	1	0	0	0	7	7	4	1	0	0		0
助	0	0	0	0	0	0	0	0	0	0	0	0	0		0
連	0	0	0	0	0	0	0	0	0	0	0	0	0		0
介	0	0	0	0	0	0	0	0	0	0	0	0	0		0
代	0	0	0	0	0	0	0	0	0	0	0	0	0		0
歎	2	0	0	0	0	0	0	2	0	0	0	0	0		0

3.4.2. 分類方法

本研究以 3.4.1 所介紹之特徵值為基礎，主要以機器學習方法進行分類。而為研究此特徵值之行為，我們亦另外提出兩個較簡單之分類法，詳述如下：

3.4.2.1. 支援向量機（SVM）分類法

支援向量機（Support Vector Machine）乃是基於統計學習理論之一機器學習分類演算法。其分類原理為將每筆輸入之訓練資料視為向量空間中之一點、每一特徵值則視為向量空間中之一維度，期能在向量空間中找出一超平面(hyperplane)將分屬二不同集合之點分開。

3.4.2.2. 條件隨機域（CRF）分類法

條件機率域模型是由（Lafferty, McCallum et al. 2001）提出的鑑別式機率模型。條件隨機域為無向性的圖模型，圖中的頂點表示為隨機變數，透過馬可夫隨機域定義一個條件機率 $P(Y|X)$ ，隨機變數 X 代表的是給定的觀察序列。在本研究

中，為善用 CRF 之特性，以 CRF 為演算法時吾人將本問題視為一序列標記（sequential labeling）問題。即將二字詞視為一僅有兩個字之短句，並以 CRF 逐字標記對應的標籤。若為「並列」之詞首，標為 1H，「並列」之詞尾標為 1T；「修飾」詞首為 2H、詞尾為 2T；「主謂」詞首為 3H、詞尾為 3T；「動賓」詞首為 4H、詞尾為 4T；「動補」詞首為 5H、詞尾為 5T；「其他」之詞首則為 0H、詞尾為 0T。標記完成後，以「合法標籤組合」（符合「NHNT」）中機率最高者為結果。

3.4.2.3. 單純貝氏（Naïve Bayes）分類法

單純貝氏分類器是一種簡單且實用的分類方法。在某些領域的應用上，其分類效果優於類神經網路和決策樹。採用監督式的學習方式，分類前必須事先知道分類型態，透過訓練樣本的訓練學習，有效地處理未來欲分類的資料。

3.4.2.4. 簡單機率分類法

本法之精神在於將「某詞歸類於某類」之機率拆解為：「某類之總機率」、「兩字分別為某詞性之機率」以及「該詞性組合時該詞歸於某類之機率」三者乘積，以現有資料估計各項機率，並直接將詞彙歸入機率最高之分類中。機率公式如下：

$$P(X \in C_k | X = (s_1, s_2)) \\ = P(X \in C_k) \times \prod_{i=1}^2 \sum_{j=1}^9 \frac{P(s_i \in A_j) \times P(s_i = x_i | s_i \in A_j) \times P(x_i \in A_j | X \in C_k)}{P(x_i \in A_j)} \quad (1)$$

其中 $X = (x_1, x_2)$ 為一由 x_1 與 x_2 構成之隨機變數，分別代表一雙音節詞之第一、第二個位置；而 s_1 與 s_2 則為實際成詞的第一個字與第二個字； C_k 為第 k 種構詞分類，而 A_j 代表第 j 種詞性（共 9 種）。其中各項所代表之意義、估計方法與需調整之參數見表 3-13（以「殺人」一詞的首字「殺」為「動詞」情況下屬於「動賓」之機率為例）：

表 3-13 機率公式各項說明

乘項	估計方法	
$P(X \in C_k)$	意義	任一二字詞屬於 C_k 類之機率
	算法	訓練語料中 C_k 類所佔之比例
	參數	無
	實例	P(殺人為動賓)
$P(s_i \in A_j)$	意義	對於目前位於 i 位置的字 s_i 而言，該字於詞彙中作 A_j 詞性的機率
	算法	s_i 下屬於 A_j 詞性之義項數佔 s_i 總義項數的比例
	參數	由於 9 種詞性中，無義項之詞性甚多，故需進行平滑化
	實例	P(殺為動詞)
$P(s_i = x_i s_i \in A_j)$	意義	在 s_i 為 A_j 詞性的情況下， s_i 位於 i 位置的機率
	算法	s_i 字下，屬 A_j 詞性之所有例詞中， s_i 位於 i 位置之比率。
	參數	由於例詞數極少，很可能缺少 s_i 在詞首或詞尾之例詞，故亦需進行平滑化。平滑化方法為：將 s_i 詞條下 A_j 詞性之二字例詞、三字例詞、四字例詞中「 s_i 位於 i 位置之比率」，以及 0.5 (詞首詞尾機率相等)，四項加權平均。故共有 3 個權重參數需要調整。
	實例	P(殺在詞首 殺為動詞)
$P(x_i \in A_j X \in C_k)$	意義	在某詞彙屬於 C_k 類情況下，位置 i 的字為 A_j 詞性的機率
	算法	考慮訓練語料中所有 C_k 類詞彙之位置 i 的字，計算該位置之所有字，詞性為 A_j 的平均機率
	參數	此項平均是在義項數機率平滑化已完成後才進行，不需另外調整參數
	實例	P(殺為動詞 殺人為動賓)
$P(x_i \in A_j)$	意義	位置 i 的字為 A_j 詞性的機率
	算法	考慮訓練語料中所有詞彙之位置 i 的字，計算該位置之所有字，詞性為 A_j 的平均機率
	參數	此項平均是在義項數機率平滑化已完成後才進行，不需另外調整參數
	實例	P(詞首為動詞)

3.4.2.5. 表格分類法

此法為實作最為簡單、直觀，以語言學知識為出發點，試圖探討真實分類與理想狀態之歧異。直接將詞性下義項數相乘，乘積經適當調整後，取其值最大之詞性組合，以語言學知識，直接將各種詞性組合指派為某一特定結構。詞性組合與構詞分類對照表如表 3-14：

表 3-14 詞性組合與構詞分類對照表

	名	形	動	副
名	並列	主謂	主謂	主謂
形	修飾	並列	修飾	並列
動	動賓	動補	並列	動補
副	修飾	修飾	修飾	並列

以語言學知識言，部分狀況並不可能發生，如「名＋副」或「形＋動」。然為盡量滿足分類之需求，必要時本表會將「副」與「形」視為等價。如「名＋副」即視為「名＋形」，而歸於「主謂」一類；而「形＋動」則視為「副＋動」，歸於「修飾」一類。而若有部分詞彙最高分之詞性組合落出此表格外，如「以為」一詞最高分項目為「介＋動」，則將該詞歸於「其他」。

而計算分數之公式見(2)：

$$Score(W_{C_H C_T}, i, j) = S(C_H, i) \times S(C_T, j) \times \prod_{k=2,3,4} \frac{\alpha_k^{E_H(C_H, i, k)} \times \beta_k^{E_T(C_T, j, k)}}{\alpha_k^{E_H(C_T, j, k)} \times \beta_k^{E_T(C_H, i, k)}} \quad (2)$$

該公式為二字詞 $W_{C_H C_T}$ 在首字為詞性 A_i 、末字為詞性 A_j 時之分數（H 代表 Head，首字之意；T 代表 Tail，尾字之意）。 $S(C, m)$ 為漢字 C 條目下詞性為 A_m 之

義項數， $E_H(C, x, y)$ 為漢字 C 條目下詞性為 A_x 、長度為 y 、首字為 C ($E_T(C, x, y)$ 則為尾字) 之例詞數。後方指數項之意義在於以例詞首尾位置調整分數。若一常於詞首作動詞之漢字，其出現於詞首時，動詞項分數會被調高，而若出現於詞尾，動詞項分數會被調低；而一常於詞尾作名詞之漢字，其出現於詞尾時，名詞項分數會被調高，而若出現於詞首，名詞項分數會被調低，是故 α_k 與 β_k 均大於 1。唯為探究此公式之正確性，實驗中調整 α_k 與 β_k 時，我們將調整範圍設於 0.5-2 之間，以尋找最適合之值。其結果將於 3.5.3 節中討論。

此法之一缺點為需調整之參數過多，包括義項數平整化參數，加以 α_k 、 β_k ，合有 7 個參數需調整。然此法之優點除計算速度極快外，更重要處在於可藉參數調整結果稍加觀察各特徵值間之重要性程度，並對本研究所宣稱之各特徵值意義（也及其作用方式）稍加驗證。此將於 3.5.3 節中詳述。

3.5 分類效能評估

我們於標記完成之語料上進行前述五種方法之測試，以下將詳述實驗設定、評估方法及結果，並對效能及可改進之處進行討論。

3.5.1. 實驗設定

為實際評估各種分類方法之優劣，本研究使用 3.3 節中標記完成之「精簡集」（6187 詞）進行實驗；條件隨機域模型直接使用 CRF++（2007）為工具、支援向量機模型則使用 LIBSVM（Chang and Lin 2001），單純貝氏模型使用 Rainbow 套件（McCallum 1998），機率法與表格法則依照 3.4.2.4 與 0 節之公式實作。

此五分類方法均於精簡集上進行 4 疊交叉驗證（4-fold cross-validation），即以四分之三語料為訓練集、四分之一語料為測試集，依序輪巡四次。最後評估時將各疊所得之每類精確度（precision）與回收率（recall）予以平均，再以平均精

確度與平均回收率計算出每類之 f-measure，此即巨觀平均（macro-average）。

而若演算法需調整參數（如機率法與表格法），則於訓練集中再次進行 4 疊交叉驗證，以基本五類的平均「macro-average F-Score」為調整標的。

3.5.2. 實驗結果

五種分類法於精簡集（6187 詞）上之實驗結果如表 3-15：

表 3-15 自動分類於精簡集（6187 詞）上之實驗結果

	LIBSVM			CRF++			Naïve Bayes			機率法			表格法		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
並列	0.52	0.16	0.25	0.59	0.51	0.55	0.54	0.31	0.40	0.51	0.16	0.25	0.33	0.33	0.33
修飾	0.54	0.95	0.69	0.73	0.81	0.77	0.80	0.56	0.66	0.52	0.96	0.68	0.70	0.59	0.64
主謂	-	0.00	-	0.36	0.30	0.33	0.15	0.77	0.25	-	-	-	0.03	0.06	0.04
動賓	0.66	0.20	0.31	0.60	0.56	0.58	0.53	0.66	0.59	0.50	0.00	0.00	0.33	0.69	0.44
動補	0.78	0.40	0.53	0.77	0.79	0.78	0.47	0.84	0.60	0.46	0.22	0.30	0.57	0.17	0.26
其他	-	0.00	-	0.31	0.17	0.22	0.11	0.29	0.16	-	0.00	-	0.15	0.13	0.14

其中參數調整結果如下：

機率法之二字例詞權重為 0.9，三字例詞權重為 0.1，四字例詞權重為 0，平滑化參數亦為 0（即指無需平滑化）；表格法之參數方面，公式(2)之 α_2 為 1.2、 β_2 為 1.2、 α_3 為 1.4、 β_3 為 1、 α_4 為 1.4、 β_4 為 1，而義項數平滑化參數為 1.8（即所有義項數值均需加 1.8，以避免乘以 0 造成的偏差）。

圖 3-9 將類別依照標記者平均效能由高而低（即由難而易）、由右至左排列，藉以比較分類器與標記者之平均效能。

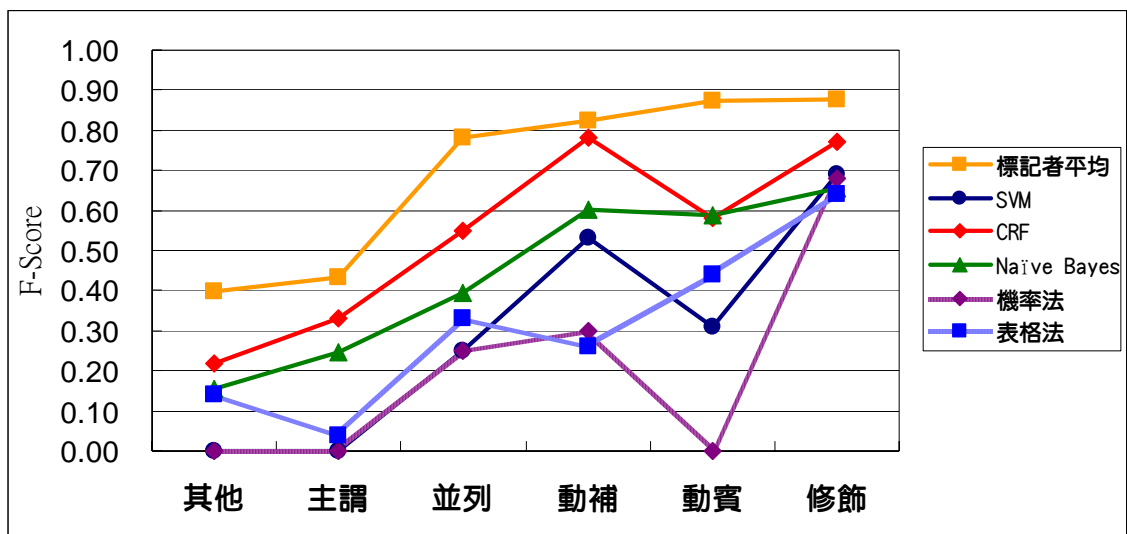


圖 3-9 二字詞自動分類器與標記者平均效能比較圖

除此之外，我們亦以效能最佳之 CRF 分類器進行精簡集（6187 詞）與強化精簡集（8186）之比較實驗。需特別注意的是，由於簡易集本為較簡單之語料集，為避免「稀釋難度」之嫌，本實驗僅針對測試集中「屬於精簡集之詞彙」進行評估。其結果如表 3-16：

表 3-16 精簡集（6187 詞）與強化精簡集（8186）效能比較

	精簡集			強化精簡集		
	P	R	F	P	R	F
並列	0.59	0.51	0.55	0.64	0.61	0.63
修飾	0.73	0.81	0.77	0.75	0.81	0.78
主謂	0.36	0.30	0.33	0.49	0.35	0.41
動賓	0.60	0.56	0.58	0.67	0.66	0.66
動受	0.77	0.79	0.78	0.77	0.78	0.78
其他	0.31	0.17	0.22	0.30	0.11	0.17

而圖 3-10 則顯示了與標記者平均效能之比較：

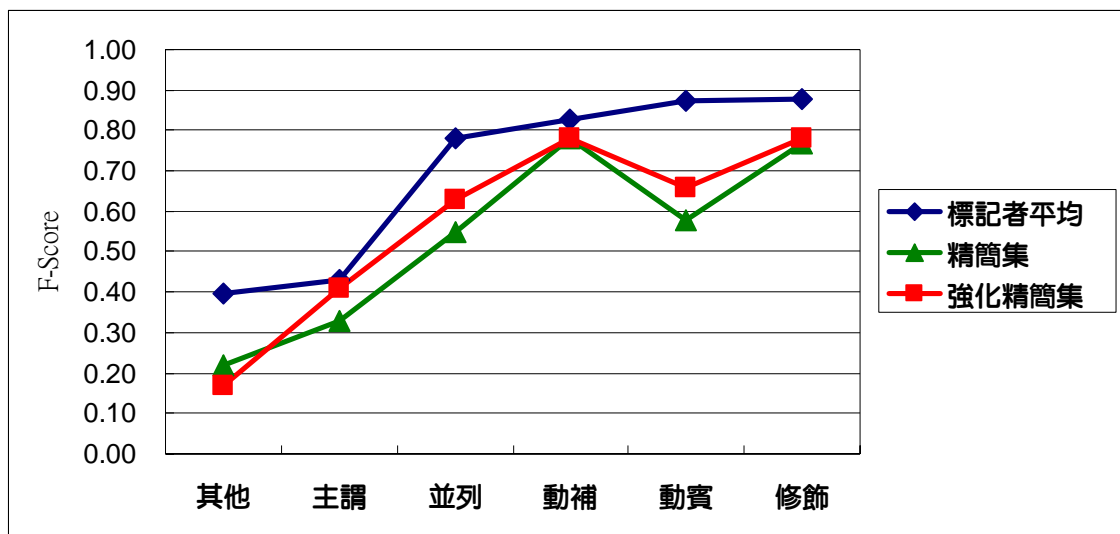


圖 3-10 精簡集、強化精簡集與標記者平均效能比較

3.5.3. 討論

首先觀察整體效能，大致上較困難之構詞類別分類效能亦較差。其中以「修飾」與「動補」兩類整體效能最佳。「修飾」效能最佳極為合理，因該類幾佔總資料量之一半；而「動補」效能亦佳之原因，或與前方所述該類詞彙均不易為辭典所收錄之理由相似：動補一類詞彙其詞義常由字義直接構成，望文生義即可理解，亦即該類詞彙之組成字表義能力較強，詞義可直接由字義表現，而本研究所使用之特徵值均依附於字彙而來，故其效能較佳。此理由較為抽象，亦難以實證分析；然經語料觀察，我們發現「動補」一類另一效能較佳之理由：該類幾無「轉品」現象。

所謂「轉品」，乃指某字彙常以某特定詞性出現，然於某特殊狀況時為因應構詞所需而改變詞性。如「書桌」之「書」，通常做名詞之用，然於「書桌」詞彙中則轉作修飾「桌」字之形容詞性字；又如「跑車」之「跑」常作動詞，於此詞彙中則轉品為形容「車」字之形容詞性字。分析分類器錯誤後吾人得知，除標記者本就有歧異之較困難詞彙外，分類器之錯誤常為此「轉品」現象所造成。並列、主謂、動賓三類皆常見轉品現象，如書桌（並列轉修飾）、雪白（主謂轉修飾）、

跑車（動賓轉修飾）等，常因轉品而與「修飾」混淆，從而造成錯誤。唯「動補」一類幾無轉品現象，故其詞會量雖未特別大，分類效能卻頗佳。

各分類器效能比較方面，CRF 佔有明顯優勢。此或與 CRF 將構詞問題視為一序列式標記問題有關，亦即構詞問題本就需考慮位置之資訊，此乃 CRF 之專長；而其他分類法均僅將之視為一般分類問題，故未達到 CRF 之效能。此外，機率法明顯較表格法為差，此可顯示本研究所使用之特徵值固可代表詞性傾向，卻非以「機率」形式體現。此現象或緣於辭典中之義項與例詞，僅為編纂辭典之專家所思及「最具代表性之範例」，數量既少，亦無法直接將之推演為機率形式。然由機率法所調整出之參數吾人可發現，二字例詞數對本問題之影響仍是最大的，三字詞之權重僅為 0.1、四字詞甚至為 0；而觀察表格法之參數（ α_2 為 1.2、 β_2 為 1.2、 α_3 為 1.4、 β_3 為 1、 α_4 為 1.4、 β_4 為 1），首先可由 α_2 、 β_2 之值確實大於 1 來肯認本研究之假設：位於詞首或詞尾時之詞性傾向確實可反映於例詞數上；繼而觀察 α_3 、 α_4 均為 1.4，而 β_3 、 β_4 均為 1（無影響），或可推知當漢字位於詞尾時，其詞性表現與該字一般於詞彙中之詞性行為相似，唯當位於詞首時才較易因後方字彙而改變原本的詞性。考量漢語為一「前修飾」語言，字彙位於詞首時為修飾後方字詞，確實而較易改變自身詞性，此參數調整結果與語言學知識吻合。

最後，加入強化集語料後效能有明顯提升。此可作為「詞內結構為一高信度之問題」的佐證，即以一人之簡易標記結果，亦具有一定程度之代表性，可作為增進效能之強化語料。

3.6 小結

本章由分類定義討論始，依序由分類之提出、語料之標記與分析展開研究，繼而提出一有效之分類特徵值，並進行完整分類實驗，已對「詞內結構」問題描繪出一完整面貌。預測效能固與人工效能稍有差距，然不難推測一部份原因乃由於本研究將可用之特徵值限定於詞彙內部所致，若於系統中實作時，對於已知詞，


當可獲得更多外部資訊，如前後文、詞彙之詞性（包括詞性頻率統計與目前詞性）等等，必可進一步提升效能。目前本研究所展示之架構與效能，乃可滿足最基礎之預測需求，即便對全新之未知詞，仍可保證相當程度之構詞預測效能，可視為系統實作時的最基礎元件，可謂對此問題提出了一通用性的解法。



第四章、中文詞詞間結構自動擷取

詞內層次處理完成後，下一步便是跨出詞的界線，將焦點轉往詞與詞之間，即一般所稱之「語法」(syntactic)層次。本節以依存關係樹為出發點，首先分析意見句與非意見句所常用之依存關係異同，繼而標記意見句中「意見段落」之所在(及其結構類別)，從而分判出「常用以表達意見」之依存關係。最後以標記語料為基礎，探討是否可能直接於語法分析樹上找到表達意見之結構。本研究由句子、句內，乃至語法分析樹之層次逐步逼近，試圖擘畫出適用於意見分析之語法結構的藍圖。

4.1 問題敘述



與詞內結構相似，本研究進行詞間結構擷取之精神在於探討「詞間結構資訊」是否可引入「意見分析」中。然二者截然不同之處在於：詞內結構預測幾為一嶄新領域，而詞間結構(即一般所稱之語法結構)之分析與預測，則早為計算語言學界所關注，舉凡詞組關係、斷詞問題、語法結構(包含語法分析樹、依存樹等)自動分析，均已發展出相當成熟之技術。然語法結構本身極為複雜、多樣，難以直接施用於意見分析中，且顯然並非所有結構形式均常用於表達意見，故本章之研究目的在於：尋找「適用於意見分析」的語法結構。此問題實可進一步拆解為兩問題：

其一，何種結構「適用於意見分析」？更基礎的問題為：是否有某幾類結構較常用於表達主觀意見？此問題看似肯定且直觀，然進一步問下去便可發覺其難處：哪些語法結構較常用於表達主觀意見？其表達主觀意見之頻率為何？是否有些特定結構幾乎不會用於意見表達、而某些結構極為常用？欲回答此問題，除對結構本身進行定性分析外，亦需經對意見句之定量分析方得以回答。此即為本研

究所欲處理之首要問題。

其二。若真有某類特定結構較常用於主觀意見表達，如何自語法分析樹（或依存關係數）中辨別此結構？而對同一種結構（如依存關係樹所定義之同一種關係）而言，若僅有部份是用於表達主觀意見，又該如何區分出表達意見的部份？欲解此問題，僅以語法結構之資訊為線索是否足夠？或需加入語義資訊？此為另一待解決之問題。

欲辨明此二問題，本研究以依存關係樹為基礎，首先比較意見句及非意見句之依存關係分布狀態，初步分析用於意見表達之語法結構其特殊性；而為更細部探究「句中實際表達意見之結構」狀態，我們提出一語料標記方法，於語法分析樹上標記「造成意見之結構」。此結果可直接轉為依存關係，從而分析實際用於表達意見之依存關係；另一方面，我們亦試圖以標記結果為訓練語料、以語法資訊為特徵值，直接於語法分析樹上進行「可能用於意見表達之結構」的預測。

4.2 基本定義：意見句與意見段落

4.2.1. 意見句與意見詞

本研究所使用之意見句，其標記方法乃依循（Ku, Lo et al. 2007）；而意見詞之標記方法則依照（Ku, Liang et al. 2006）。其二者之主要精神在於「含有主觀意見之語言單位」，即指文本中含有非客觀之敘述，涉及個人情感及主觀判斷者。而較需說明處為，此處之「主觀」乃指文本作者或文中主角之主觀，而非讀者（編記者）之主觀。如於一新聞文章中，一民眾對政府之政績大肆褒揚，即便標記者對此主觀意見本身並不同意，甚或反對，仍須將此句標示為「正面意見句」，而非以己意標為「負面意見句」。此處之「主觀」乃純以文本角度出發，並不涉及讀者詮釋問題。

4.2.2. 意見段落

意見分析中，常以「意見詞」為「表達意見」之最小單位（以語素為「計算意見」之最小單位），而以「意見句」為更大一級之單位。然而，即便於意見句中，亦非全句均含有主觀意見，常有部分為客觀事實之陳述。主觀意見僅出現於某些「意見段落」中。

此處所指之「意見段落」，乃是一較詞彙為大、較句子為小之語言單位，許多情況下接近於「詞組」（但並非完全相同，亦有可能為子句）。凡於一完整句子中，含有主觀意見之連續部份，便稱為「意見段落」。以賓大樹庫 5.1 版經標記後產出之一正面意見句為例：「身披黑褐色三角形斑紋、受侵擾時昂頭警戒，神氣的百步蛇，是蛇之王者。」（FID：1066；SID：14406），其中「身披黑褐色三角形斑紋、受侵擾時昂頭警戒」僅為客觀描述，唯後段之「神氣的百步蛇，是蛇之王者」方為主觀意見之表達。

4.3 問題初探：意見句及非意見句之依存關係樹比較

本章所處理之問題皆建立於一基本假設之上：「用於『意見表達』之語法結構有其特殊性」。即因其具特殊性，吾人得以觀察、分析甚至加以預測。是以展開本章前首先必對此問題進行初步之確認。欲探討意見表達之語法結構特殊性問題，一較簡易之方式為分析「意見句」與「非意見句」中語法結構分佈之異同。由於依存關係其形式為條列式之語法關係，便於比較，故選擇依存關係為比較之對象。

4.3.1. 意見句標記

欲對意見句及非意見句之語法關係進行比較，首先必須標記意見句。本研究依（Ku, Lo et al. 2007）所提出之方法對賓大樹庫 5.1 版之所有句子（所有 SID）進行意見句標記。賓大樹庫 5.1 版共計 890 個檔案、18784 句，刪去同一 SID 中

含有兩棵語法分析樹之 20 句(詳見附錄 B)，經標記後，共有 10676 句為意見句²⁴。

4.3.2. 意見句及非意見句依存關係樹分佈比較

將賓大樹庫 5.1 版之所有句子(無論是否為意見句)均以史丹佛分析套件中之中文依存關係樹分析器分析，產生其依存關係列表(並可構成一棵依存關係樹)。而後結合 4.3.1 節所標記之意見句資訊，統計所有意見句及非意見句中之依存關係分布狀態，觀察表達意見之語法結構是否有其特殊性。統計結果如表 4-1：

表 4-1 意見句與非意見句依存關係分布比較

排名	非意見句			意見句			合計		
	類型	數量	比例	類型	數量	比例	類型	數量	比例
1	nmod	22892	17.33%	nmod	37503	13.17%	nmod	60395	14.49%
2	nsubj	10807	8.18%	ccomp	31347	11.01%	ccomp	40728	9.77%
3	ccomp	9381	7.10%	advmod	26201	9.20%	nsubj	36928	8.86%
4	dep	9075	6.87%	nsubj	26121	9.17%	advmod	34081	8.18%
5	dobj	8670	6.56%	dobj	24307	8.54%	dobj	32977	7.91%
6	advmod	7880	5.96%	prep	11005	3.86%	dep	17312	4.15%
7	numod	6626	5.02%	rcmod	10459	3.67%	prep	16400	3.93%
8	prep	5395	4.08%	cpm	9267	3.25%	rcmod	14835	3.56%
9	conj	4448	3.37%	assm	9189	3.23%	numod	14273	3.42%
10	rcmod	4376	3.31%	assmod	9113	3.20%	cpm	12596	3.02%
11	amod	4254	3.22%	dep	8237	2.89%	assm	12498	3.00%
12	pobj	4219	3.19%	amod	8185	2.87%	amod	12439	2.98%
13	clf	3365	2.55%	pobj	8069	2.83%	assmod	12375	2.97%
14	cpm	3329	2.52%	numod	7647	2.69%	pobj	12288	2.95%
15	assm	3309	2.50%	conj	6968	2.45%	conj	11416	2.74%
16	assmod	3262	2.47%	cc	5032	1.77%	clf	8003	1.92%
17	cc	2433	1.84%	mmod	4911	1.72%	cc	7465	1.79%
18	lobj	2079	1.57%	clf	4638	1.63%	lobj	6207	1.49%

²⁴ 本標記工作以服務「意見分析」為目的，故標記內容包括意見分析所需之完整欄位，如「是否為意見句」、「意見傾向」、「意見持有者」、「意見目標」等等。然本研究僅取「是否為意見句」之資訊為用。

	非意見句			意見句			合計		
排名	類型	數量	比例	類型	數量	比例	類型	數量	比例
19	det	1938	1.47%	lobj	4128	1.45%	det	6024	1.45%
20	range	1872	1.42%	det	4086	1.43%	mmod	5757	1.38%
21	asp	1288	0.97%	asp	2891	1.02%	asp	4179	1.00%
22	tcomp	1252	0.95%	neg	2702	0.95%	attr	3871	0.93%
23	attr	1204	0.91%	attr	2667	0.94%	plmod	3483	0.84%
24	plmod	1101	0.83%	plmod	2382	0.84%	lccomp	3103	0.74%
25	lccomp	1039	0.79%	lccomp	2064	0.72%	neg	2985	0.72%
26	mmod	846	0.64%	clmpd	1903	0.67%	tcomp	2843	0.68%
27	top	799	0.60%	tcomp	1591	0.56%	range	2820	0.68%
28	ordmod	667	0.50%	top	1460	0.51%	clmpd	2344	0.56%
29	etc	502	0.38%	xsubj	1230	0.43%	top	2259	0.54%
30	prnmod	451	0.34%	tclaus	1140	0.40%	tclaus	1583	0.38%
31	tclaus	443	0.34%	partmod	1039	0.36%	xsubj	1514	0.36%
32	clmpd	441	0.33%	range	948	0.33%	rcomp	1344	0.32%
33	rcomp	407	0.31%	rcomp	937	0.33%	partmod	1328	0.32%
34	partmod	289	0.22%	etc	663	0.23%	ordmod	1220	0.29%
35	xsubj	284	0.21%	vmod	613	0.22%	etc	1165	0.28%
36	neg	283	0.21%	ba	576	0.20%	vmod	866	0.21%
37	vmod	253	0.19%	ordmod	553	0.19%	prnmod	771	0.18%
38	comod	222	0.17%	dvpm	544	0.19%	ba	758	0.18%
39	ba	182	0.14%	comod	533	0.19%	comod	755	0.18%
40	pass	161	0.12%	cop	520	0.18%	dvpm	642	0.15%
41	cop	106	0.08%	dvpm	502	0.18%	cop	626	0.15%
42	dvpm	98	0.07%	pass	400	0.14%	dvpm	591	0.14%
43	dvpm	89	0.07%	prnmod	320	0.11%	pass	561	0.13%
44	npsubj	53	0.04%	xcomp	85	0.03%	npsubj	138	0.03%
45	xcomp	30	0.02%	npsubj	85	0.03%	xcomp	115	0.03%
46	acom	5	0.00%	acom	11	0.00%	acom	16	0.00%
	合計	132105	100.00%	合計	284772	100.00%	合計	416877	100.00%

表 4-1 所指之「比例」為該關係類型佔全部意見句（或非意見句）中關係總量之比例。非意見句欄位中底色塗灰者為排序明顯較意見句為高者（排序差至少為 3），意見句欄位中底色塗灰者則為排序明顯較非意見句為高者。

意見句中次序較高者有：advmod (adverbial modifier, 副詞性修飾子)、rmod (resultative modifier, 結果動詞)、cpm (complementizer, 補語連詞)、asm (associative maker)、assmod (associative modifier)、mmod (modal verb modifier, 情態動詞修飾子)、neg (negative modifier, 否定子)、clmpd、xsubj (controlling subject)、partmod (particles such 所, 以, 來, 而)、ba (把)、dvpm (manner DE 地 modifier); 較低者有：dep (dependent, 依存關係)、numod (數量修飾子)、conj (conjunct, 連接詞)、clf (classifier modifier, 分類修飾子)、range (dative object that is a quantifier phrase, 作為語格受詞的數量詞)、tcomp (時間補語)、ordmod (ordinal number modifier, 序數修飾子)、etc (etc modifier, 等等、諸如此類)。

由此結果可初步看出, 帶有「修飾」意味之關係(情態動詞、副詞)、以及常於修飾時出現之補語(的、地)較傾向於出現在意見句中, 而客觀敘述之結構(如數量、時間、類別、序數等)則較傾向於出現在非意見句中, 初步肯定了表達意見之語法結構確實具有特殊性。至此固暫時無法深究此差異之成因與細節, 然初步肯認表意結構之特殊性後, 即可對表意結構進行進一步的探索。

4.4 中文詞詞間結構語料標記

4.4.1. 標記目的及使用語料

如前所述, 本研究所探討之首要問題為: 何種結構「適用於意見分析」? 欲研究此問題, 必須比較「含有意見」與「不含有意見」之結構。於4.3節中已初步分析意見句及非意見句之依存關係分布狀態, 指出特定數種關係可能常用於意見表達。然而, 前述之分析僅為初步探索, 欲知於一句子中實際用於表達意見之語法結構種類, 則仍需以人工標記結果進行更精緻之研究。以此為出發點, 本章將展開一大規模語料標記計畫, 並為之架設一線上標記系統。

欲進行語料標記, 首先必須闡明: 標記對象為何? 將於其上進行何種標記? 而標記結果又有何作用? 如前所述, 本研究之標記目的在於「分析句中實際用於

意見表達之結構」，是以我們首先決定僅對「意見句」進行標記。由於意見句之定義使然，吾人幾乎可假設凡意見句中至少存在一個意見段落，無須耗費過多成本於非意見段落中，以達最佳標記效率。而該於何種材料上進行標記？就後續分析而言，依存關係樹為一理想對象，其規則簡單（僅 46 種關係，而非抽象之文法）、對象單純（必為詞彙與詞彙間關係，而非抽象語法單位間之關係）、關係清楚，無論於比較數量或分析內容時均有極大的便利性；然而，若直接於一句子之依存關係樹上進行標記，則將遭遇三項主要問題：

首先，依存關係樹並不如語法分析樹直觀。對標記者（多為中文系學生）而言——語言學概論為中文系大一全年之必修課程——理解語法分析樹是容易的，理解依存關係樹卻是困難的。於語法分析樹上進行標記，非但於解釋時耗費的時間較少，亦可減少因不熟悉依存關係樹而造成的誤差。

其次，「依存關係」之概念本身即有歧異性。以史丹佛分析套件為例，該分析器以「動詞」為句中最重要之角色，即所建構之依存關係樹幾乎全以該句之主要動詞為根點；然而，迷你語法分析套件（minipar）則持相悖的看法。該語法分析器以「名詞」為句中最重要部分，該分析器以句中之主詞為依存關係樹之根點。此間並無優劣之別，而乃是對「依存關係」之概念、語義表現之哲學不同的詮釋。以此而論，若僅就某一個分析套件產出之依存關係樹進行標記，則將產生詮釋上過於狹隘的問題。另一方面，語法分析樹則較無此問題。首先，語法分析為一起步甚早之問題，學界已有一基礎之共識；再者，由於已存在數個標記完成之大型語料庫，語法分析架構幾被為計算語言學界視為不變之標準，諸多研究與詮釋，均在語法分析樹上展開。以此而論，語法分析樹顯為一較佳之標記對象。

再者，以後續預測或其他研究所需而論，依存關係樹所包含之資訊量遠較語法分析樹為少。大多時候，語法分析樹可轉為依存關係樹，然依存關係樹卻無法轉為語法分析樹。若吾人於依存關係樹上進行標記，日後需更進一步資訊以供研究時，必將遭致許多困境。

基於以上三點，本研究選擇於語法分析樹上進行語料標記。而實際標記時所

使用之語料為 4.3.1 節中已標記完成之「意見句」，此語料幾乎完全符合本研究之需求。標記方法細節則詳述於後。

4.4.2. 詞間結構定義與分類

依存關係樹與依存關係，已廣為諸多學者所研究。而本節所欲定義之「詞間結構」，乃是為標記所需而設計，非為提出一完整之依存關係語法。本研究所謂「詞間結構」，乃指「語法分析樹中兩特定節點間之關係」。限定為「兩」節點間關係，除此定義方法較易轉為依存關係外，同時也考量意見擷取多以句中「意見詞」為基礎單位，若引入複雜之語法結構，則將致使意見詞周圍之資訊過於雜亂，難以分析意見詞間的相互作用及對全句意見之影響，是以本研究先將目光聚焦於「節點對」，簡化問題，妥善聚焦；而將關係定於「節點」間而非「詞彙」間，一則標記較為容易，二則如此方得以善用語法分析樹之豐富資訊（否則使用依存關係樹即可）。「兩」節點之限制，用意在於簡化問題；考量所有「節點」而非限於「詞彙」，用意在於減少因過於簡化而產生之資訊漏失，兩者相輔相成，以期對意見擷取達到最佳的輔助效果。

定義「詞間結構」後，以下將對「可能常見於意見分析之詞間結構」進行分類定義，以便開展後續之標記工作。以下定義以（程祥徽 and 田小琳 1995）為基礎，與詞內層次相似，原亦有五種主要結構，然吾人觀察發現，「並列」一類於詞間層次時極少用於表達主觀意見（因兩詞彙於語法地位上是平行的），加以詞間層次中大量名詞片語的出現，將造成標記上的困難，故刪去此類，保留其餘四類。以下仍以「詞彙」角度說明各種關係，但實際於標記及計算時均一律推廣至語法分析樹的所有節點上：

(1) 修飾關係 (Substantive-Modifier, 又稱偏正關係)

第一個詞彙用以修飾第二個詞彙、第二個詞彙被第一個詞彙所修飾，如

「高大的／樓房」、「淒涼地／笑」、「非常／美麗」等等，此處我們亦限

制被修飾者必為第二個詞。然與詞內層次相異，詞間層次中「被修飾者為第一個詞」的情況較為常見，其中若第一個詞為述語（常為動詞，如「表現／異常」）則歸入「動補」；而若第一個詞為主語、第二個詞為謂語，如「穹蒼／湛藍」，則歸入「主謂」；而其餘狀況則歸入「其他」。

(2) 主謂關係 (Subjective-Predicate, 又稱陳述關係)

第一個詞彙為主語、第二個詞彙為謂語，如「大雪／紛飛」、「大廈／竣工」、「經濟／崩盤」等等。需特別說明的是，此關係中，主詞不必然為實體之物，謂語亦不必然為動詞，亦可為形容詞（陳述語）。如「政治／清明」、「情誼／深厚」等等。

(3) 動賓關係 (Verb-Object, 又稱支配關係)

第一個詞彙為述語（常為動詞），第二個詞彙則為其實語（即受詞，常為名詞），如「揭發／真相」、「恢復／體力」、「開啟／新時代」等等。

(4) 動補關係 (Verb-Complement, 又稱補充關係)

第一個詞彙為謂語（常為動詞或形容詞），後一個詞彙則為補充謂語之補語。如「收拾／乾淨」、「說／清楚」、「飛舞／起來」等等。

除上述主要四類外，於語料標記過程中我們發現某些特定句型亦常用於表達主觀意見，故特將此類句型自「其他」類中獨立分出。以下數類並不符合「詞間結構為兩節點間關係」之定義，而僅為某些特定句型之簡單描述，其標記方式將於次節中說明：

(5) 使役句：S1+使／使得／讓+S2+V

(6) 把字句：S+把+O+...

(7) 被動句：S+被+...

(8) 以……為……

(9) 比較句：S1+比／比較／相較於／較之+S2

此外，凡為意見段落，卻無法歸為以上九類者均屬於第十類：

(10) 其他意見段落

包括後修飾之例外、帶有意見之感嘆語、反諷句等等，凡帶有主觀意見卻無法歸為以上九類者均歸於此。

4.4.3. 標記方法暨「潘恩標記系統」(Pan Annotation System)

由於賓大樹庫資料量極大，若以離線方式標記，於語料管理及標記品質控制上均非常不便。為解決此問題，我們特架設一線上語料標記系統，將賓大樹庫的語法分析樹以圖形化介面顯示，以便於任二節點上標記詞間關係。我們稱該系統為「潘恩標記系統」(Pan Annotation System, PAS)²⁵，其介面可見圖 4-3。潘恩系統可讀取賓大樹庫所提供的語法結構資訊，並於頁面上顯示任一特定序號

(SID) 句子所生成的語法分析樹，標記者可直接於畫面上點選節點、進行標記。

截至目前為止，本論文談及「意見句」、「意見段落」、「詞間結構」及其分類，卻仍未明確解釋實際標記之方法。欲說明本系統所標記之內容，可先參考圖 4-1：

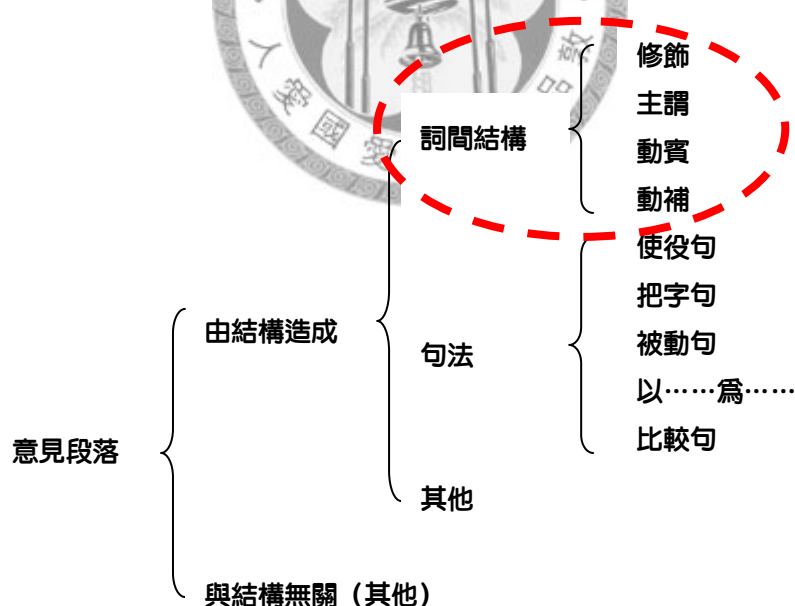


圖 4-1 意見段落與結構關係圖

²⁵ 取「許多語法分析樹叢聚為森林」之義，以希臘神話中森林神潘恩（Pan）為系統命名，同時「Pan」亦與賓州大學之「Penn」諧音。

圖 4-1 所概括之範圍即為本研究所欲標記之內容。由於本研究以意見句為標記對象，故可假設句中必有意見段落，而意見段落中實際「造成意見」之部分又可分為「由結構造成」及「與結構無關」兩類，前者又可細分為「詞間結構」和「句法結構」兩主要部份。而我們所欲標記的主要部份即為「詞間結構」，系統諸多細節亦為其所專門設計。標記者首先須判斷各意見句之「意見段落」位置，並從意見段落中標出符合 4.2 節所定義之詞間關係。本研究將「詞間結構」限定為二節點間關係，故標記時理當於畫面中點選二詞彙，選擇「結構類別」後送出即可。然我們欲探討另一子問題：

某詞間結構所造成之意見段落，其範圍是否會受語法分析樹的範圍所限制？限制到什麼程度？有否可能，以某詞間結構為核心之意見段落，其範圍較「可覆蓋兩節點之最小子樹」為大？

此問題背後所指涉之更深層問題是：若意見段落之範圍多為語法結構樹所限制，則於語法結構自動擷取時，只要判斷結構之所在，便可直接由語法樹之結構決定意見段落範圍，於是問題可進一步被簡化為「判斷特定語法結構之位置」，於應用時幫助極大。

欲探討此問題，標記時除點選彼此互有關聯的兩節點外，標記者亦需點選此二節點之「頂點」(head)。此「頂點」的定義為：能夠包含「由該二節點為核心所構成的意見段落」之「最小子樹」的「樹頂點」。由此可知，「頂點」位置必不可低於二節點之「最近親節點」(但可等於)。

組成語法結構的兩節點稱為「腳點」(feet)，兩腳點位於句中較前方者稱為「左腳點」、較後方者稱為「右腳點」。而再加以一頂點所構成之基本標記單位，稱為「三角單元」(trio)。三角單元標記實例可見圖 4-2。

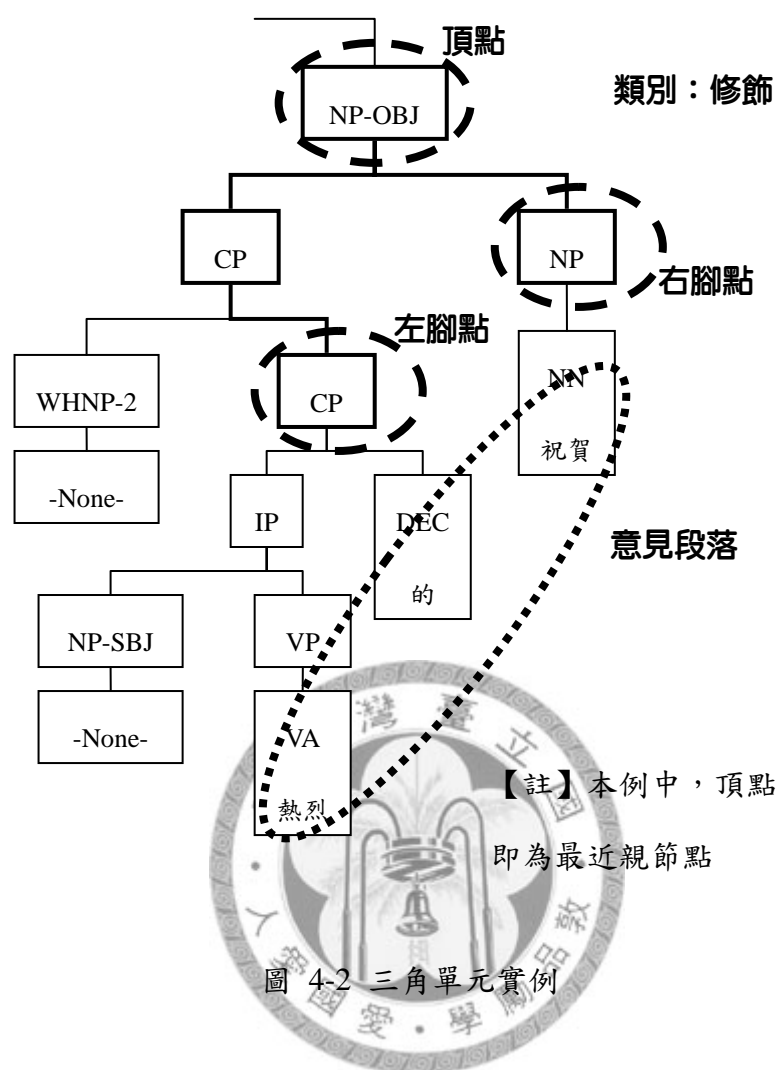


圖 4-2 三角單元實例

而 4.4.2 節中所定義之語法結構，除四種主要結構外，亦有五種常見句型與「其他意見段落」。此六類並不屬於「二節點間之關係」，是以標記時亦不適用三角單元。其標記目的在於為意見傾向實驗提供更完整之資訊，故並不拘泥於結構，亦即將左右兩腳點標示出該意見段落範圍、並選取相對應之句型選項即可²⁶。

潘恩系統以 JSP 撰寫，實際對 4.3.1 節標記之意見句進行意見結構標記。本標記工作共聘請至少六位中文系大學部學生標記之。每句均由二位標記者共同標記，第一位標記者標記後，第二位再行檢查。其標記結果將於次節中分析。

²⁶ 為實作方便，系統運行時仍會要求標記者點出頂點，但並無實際意義。

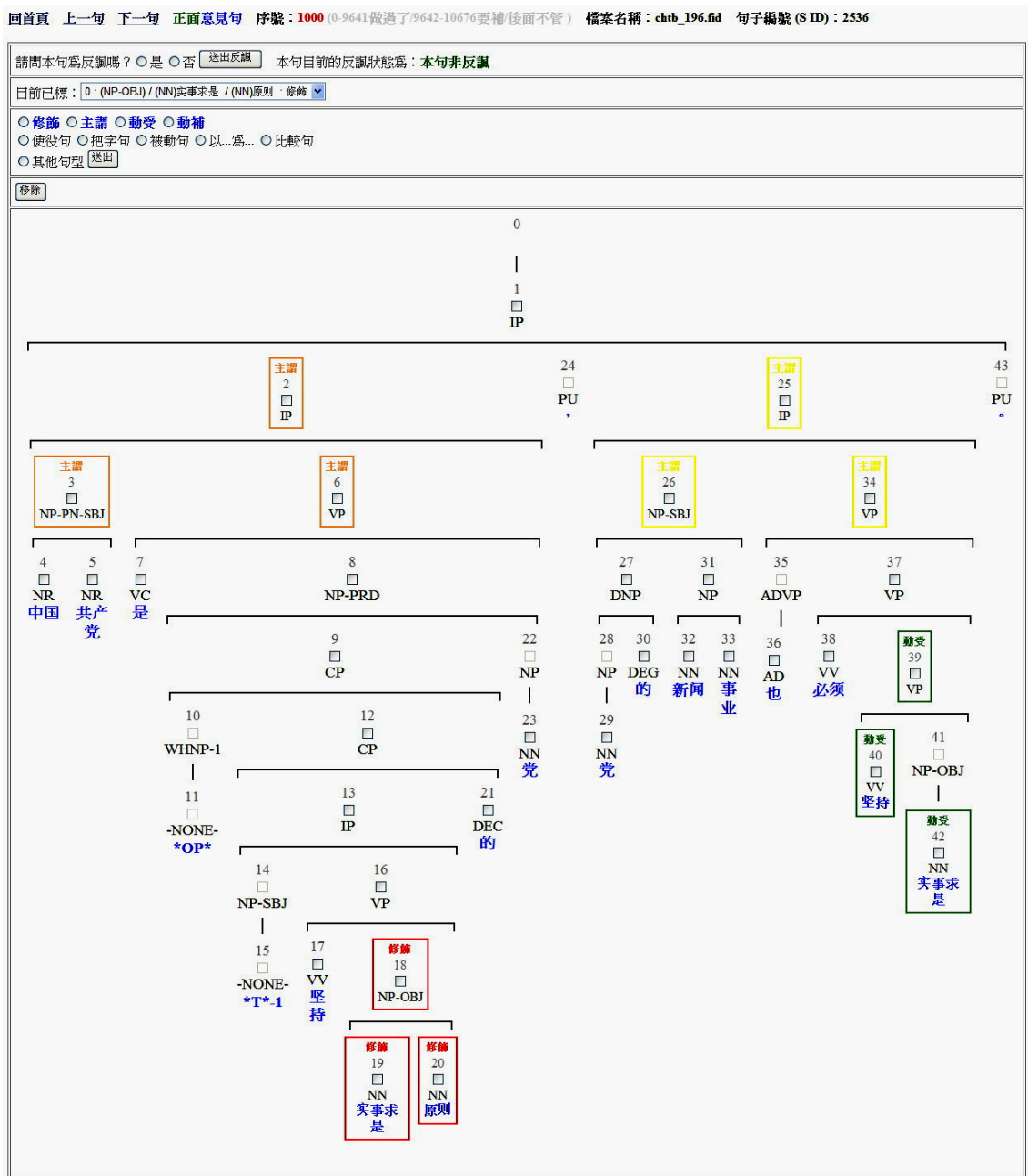


圖 4-3 「潘恩標記系統」介面

4.5 語料分析

4.5.1. 原始標記結果分析

表 4-2 展示了標記的總結果，並以意見句之「傾向」(正面、中性、負面)或「類型」(動作句、人講句、狀態句、無法確定)分項列之。圖 4-4、圖 4-5、

圖 4-6、圖 4-7 則分別展示了各種向度之標記結果分析。就圖中我們可觀察得知意見結構之分佈比例幾與意見句傾向無關，亦即無論正面、負面或中立之意見句，其意見結構之組成成分均極為相似；而若以意見句類型分析之，則「動作句」中「動賓」結構比例較高，此顯然與「動作句」本定義為「含有動作之意見句」有關；而若改以結構分類為軸，則較明顯之傾向為：「被動句」常用於表達「中立」之意見；頂點標記結果方面，有高達 96.3% 的三角單元，其「頂點」即為兩腳點之「最近親節點」。由此可看出，結構所造成之意見，其範圍大致上仍受語法結構樹所限制，亦即若可於語法分析樹中辨認出特定特徵、從而擷取表達意見之結構，則毋需多加考慮意見段落範圍之問題。

表 4-2 詞間結構標記統計表

	修飾	主謂	動賓	動補	使役句	把字句	被動句	以...為	比較句	其他	合計
正面	13780	9641	12142	718	159	175	138	102	160	47	37062
中性	4108	3271	3158	289	50	63	136	18	29	19	11141
負面	3525	3039	2732	204	45	45	75	27	31	18	9741
合計	21413	15951	18032	1211	254	283	349	147	220	84	57944
動作句	1136	687	1628	50	21	22	11	14	5	2	3576
人講句	9555	7725	8245	487	77	138	145	69	78	26	26545
狀態句	8066	5642	5482	444	120	73	136	41	93	35	20132
無法確定	2656	1897	2677	230	36	50	57	23	44	21	7691

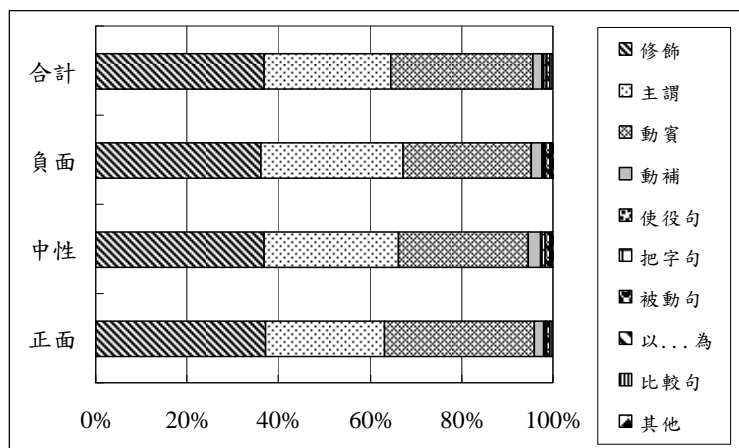


圖 4-4 各種傾向之意見句中意見結構分佈比較

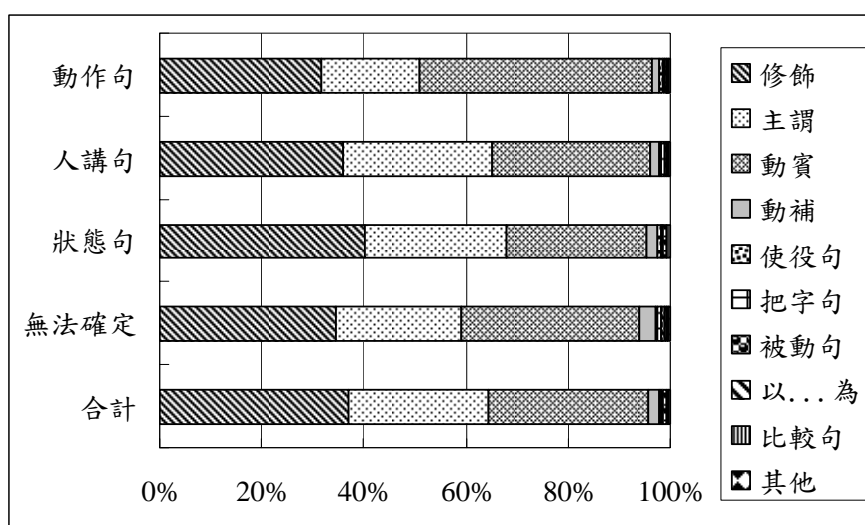


圖 4-5 各種類型意見句中意見結構分佈比較

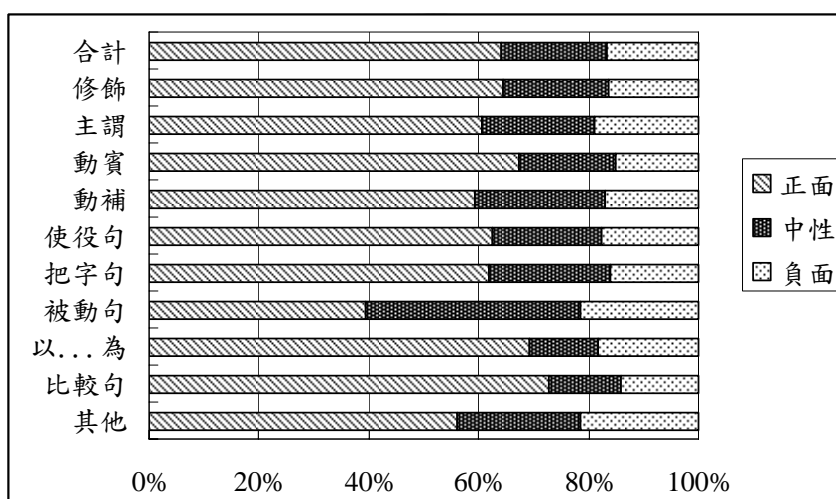


圖 4-6 各種意見結構之意見句傾向分佈比較

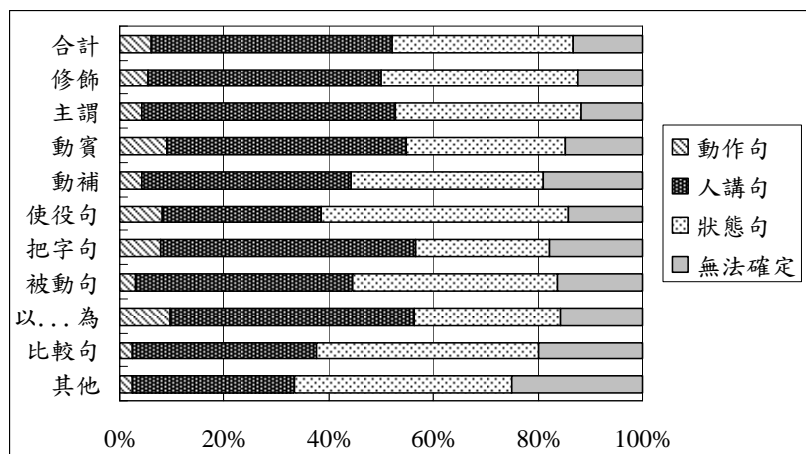


圖 4-7 各種意見結構之意見句類型分佈比較

4.5.2. 依存關係分析

4.5.2.1. 依存關係轉換方法

在依存關係樹已知的情況下，欲將標記結果轉換為依存關係，需經三個步驟。以下將以賓大樹庫之一句正面意見句：「並投資一千三百多個億，加強基礎設施和基礎產業建設，為擴大對外開放創造強好環境。」（FID：008；SID：87）為例，說明並示範轉換方法。

首先，對任一三角單位而言，先列出兩腳點各自涵蓋之所有詞彙（頂點於此處略去不計）。以圖 4-8（見 61 頁）為例，該語法分析樹上共有四處標記，對每一個標記單位而言，均先找出其左右腳點所涵蓋之詞彙：

- 
- (1) { 加強 }, { 基礎, 設施, 和, 基礎, 產業, 建設 }
 - (2) { 創造 }, { 良好, 環境 }
 - (3) { 良好 }, { 環境 }
 - (4) { 擴大 }, { 對, 外, 開放 }

找出詞彙後，對照該句之所有依存關係，找出所有「構成依存關係之兩詞彙分別屬於左腳點詞彙集與右腳點詞彙集」的關係，作為該標記單元轉換後之候選關係。本例中，至此步驟時四個標記單元均僅有一條候選關係，則直接選擇該關係即可。

最後，若候選關係數非一，則需進入第三步：

若候選關係數為零，即可能左腳點詞彙集與右腳點詞彙於語法分析樹中相隔過遠，造成其間沒有直接連結，而是透過其他詞彙間接連接。此情況下，為使轉換結果明確表現出結構特性，則捨棄該標記單元，不將之轉換為依存關係。

若候選關係數大於一，則代表標記的位置較高，左腳點詞彙集與右腳點詞彙

集均含有大量詞彙，而產生多條依存關係。於此情況下，則於候選關係中選出「距離根節點最近」的關係作為轉換的結果。依存關係樹以根節點為該句最重要之語義核心，而距離根節點越近，則該關係越形重要。亦即本研究假設：每一標記單元均僅指涉一條最主要的依存關係。為嚴格起見，並不允許「一變多」的情況發生。

圖 4-8 的四個標記單元經轉換後，可對照至圖 4-9（見 62 頁）的四個關係。

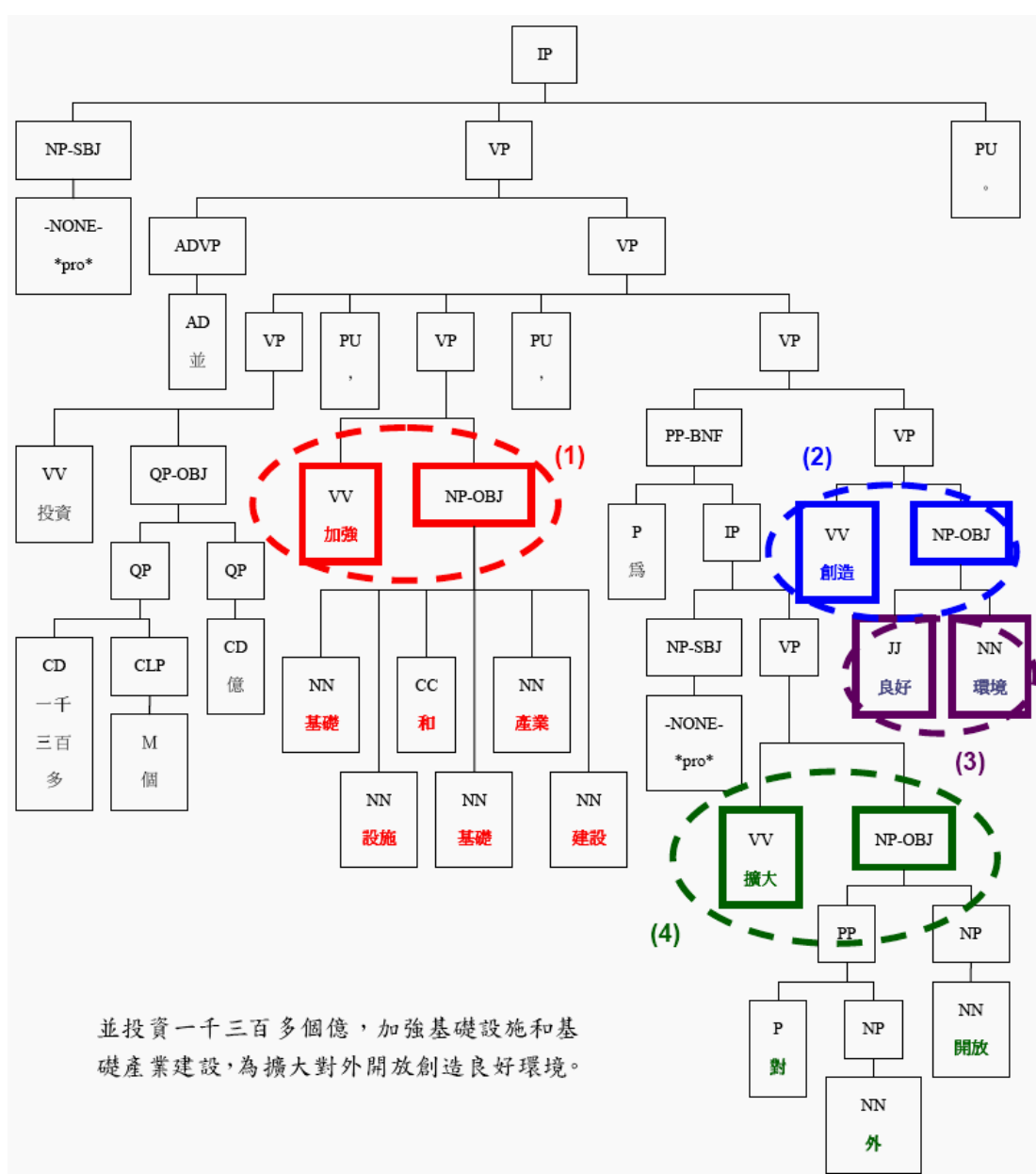


圖 4-8 標記語料轉換為依存關係範例（轉換前）

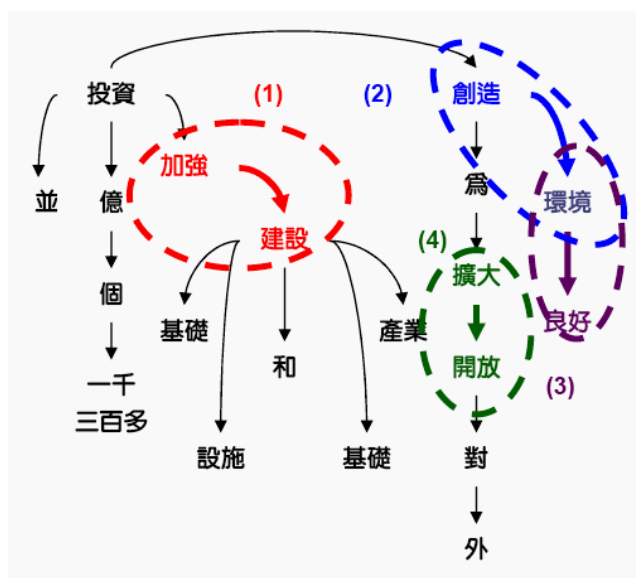


圖 4-9 標記語料轉換為依存關係範例（轉換後）

4.5.2.2. 轉換結果統計及分析

經上述步驟轉換後，我們將所有標記結果均轉為依存關係，便可計算每種依存關係實際用於意見表達之比例。

統計結果如表 4-3：

表 4-3 意見句中表達意見之依存關係比例

類型	總數	意見數	意見比例	總量比例
dvpmod	463	378	81.64%	0.18%
dobj	22015	11666	52.99%	8.55%
pass	361	191	52.91%	0.14%
npsbj	78	37	47.44%	0.03%
top	1290	551	42.71%	0.50%
ba	521	221	42.42%	0.20%
neg	2406	987	41.02%	0.93%
nsubj	23671	9657	40.80%	9.20%
amod	7498	3024	40.33%	2.91%
rcmod	9375	3565	38.03%	3.64%
rcomp	843	248	29.42%	0.33%
advmod	23509	6660	28.33%	9.13%

類型	總數	意見數	意見比例	總量比例
range	850	240	28.24%	0.33%
mmod	4376	1165	26.62%	1.70%
assmod	8138	1413	17.36%	3.16%
ccomp	28367	3852	13.58%	11.02%
vmod	561	72	12.83%	0.22%
dep	7437	789	10.61%	2.89%
xsubj	1128	90	7.98%	0.44%
comod	493	38	7.71%	0.19%
cop	471	35	7.43%	0.18%
lccomp	1882	139	7.39%	0.73%
prep	9996	650	6.50%	3.88%
attr	2366	135	5.71%	0.92%
pobj	7303	253	3.46%	2.84%
clmpd	1729	57	3.30%	0.67%
tcomp	1436	47	3.27%	0.56%
nmod	34042	951	2.79%	13.23%
asp	2706	72	2.66%	1.05%
numod	6928	168	2.42%	2.69%
clf	4210	86	2.04%	1.64%
ordmod	503	10	1.99%	0.20%
dvpm	506	10	1.98%	0.20%
det	3711	72	1.94%	1.44%
partmod	936	13	1.39%	0.36%
plmod	2171	11	0.51%	0.84%
prnmod	295	1	0.34%	0.11%
Cc	4527	12	0.27%	1.76%
Lobj	3776	10	0.26%	1.47%
Conj	6274	11	0.18%	2.44%
Cpm	8338	14	0.17%	3.24%
Assm	8205	9	0.11%	3.19%
tclaus	1032	1	0.10%	0.40%
xcomp	75	0	0.00%	0.03%
Etc	594	0	0.00%	0.23%
acomp	11	0	0.00%	0.00%
合計	257403	47611	18.50%	100.00%

我們可首先注意到，對所有類型之依存關係而言，平均僅有 18.5% 的場合用於意見表達。以此為基準，可發現共有 14 種依存關係用於意見表達之比例超過平均值，由高而低依序為：dvpmod(副詞語尾「地」)、dobj(直接受詞)、pass、npsubj、top(主題)、ba(把)、neg(否定子)、nsubj(nominal subject, 名詞性主詞)、amod(adjectival modifier, 形容詞性修飾子)、rcmod(relative clause modifier, 關係子句修飾子)、rcomp(resultative complement, 結果補語)、advmod(adverbial modifier, 副詞性修飾子)、range(dative object that is a quantifier phrase, 數量的語格受詞)、mmode(modal verb modifier, 情態動詞修飾子)。其中以 dvpmod(副詞語尾「地」) 意見表達比較最高，遠超過第二順序之意見比例；同時我們亦可發現，除 dvpmod 外，即便為表達意見比例較高之 14 種關係，其表達意見之比例大致不會超過 50%。此結果固可直接施用於意見分析中(作為特徵值或設計公式之依據，不失為一簡易且成本低廉之方法)，但其雜訊亦非常多，因此需要更深入、精確的預測。此將於次節中討論。

4.6 詞間結構自動擷取

如前所述，若將依存關係資訊施用於意見分析中，可直接使用經實驗證實最常用於意見表達之數種關係。然由實驗中吾人亦發現，除 dvpmod 一類外，大多依存關係用於表達意見之情況均不超過五成，即對大多數依存關係而言，「此時是否用於意見表達」仍是難以判斷的問題。是以本研究繼而展開更細部的自動擷取研究，盼能更準確地判斷出「表達意見」的結構位置。同時，我們亦欲探討：用以表達意見之結構位置，可否僅透過語法分析樹本身之資訊便預測得知？其背後之更深層問題，乃是試圖探討「語法結構」對於「意見表達」的影響程度。

4.6.1. 自動擷取方法

既為「自動擷取」問題，首先必須指明：欲自動擷取的對象為何？依存關係？

文法？抑或詞彙？若以依存關係為擷取對象，考量依存關係樹資訊極為有限，勢必將使用語法分析樹之資訊，於語法分析樹中找出腳點後再行轉換；而為避免轉換過程中漏失之資訊（如 4.5.2.1 節所述，部分標記單元將無法轉為依存關係），本研究決直接於語法分析樹上進行預測，而預測對象即為標記者所標記之三角單元位置。然頂點於此並無意見分析上之幫助，故預測對象仍以「腳點對」為主。

為簡化問題，本研究暫先鎖定於「兩腳點均為同個親節點之子節點」的標記單元上（即此兩腳點位於同層且互為「兄弟」），此結構稱為「同親結構」，而「屬於同一親節點之所有子節點」所構成之序列稱為「同親序列」。本研究亦假設每一同親序列最多僅含一對腳點。非同親結構範例如圖 4-11、同親結構則如圖 4-10。需特別說明處為：圖 4-10 中，右下角「堅持」與「實事求是」兩節點，亦視為同親結構。由於「實事求是」為其親節點之「孤子」，故本研究視該點與親節點等價，從而構成同親結構。

定義「同親結構」之主要目的在於：簡化自動擷取問題，將之視為一「序列式標記」問題。即將同親序列視為一句子，以類於詞性標記之方法，以 CRF 逐一標記對應之標籤，如：非腳點標為 N、腳點標為 Y，或「修飾」結構之左腳點標為 2H、右腳點標為 2T、非腳點標為 N，等等。後續研究中將針對各種不同標籤集進行實驗、評估，試圖找出最適合的標籤集。

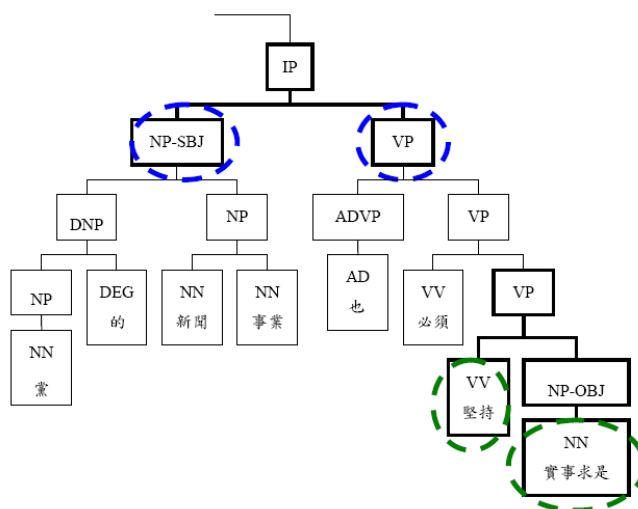


圖 4-10 同親結構範例

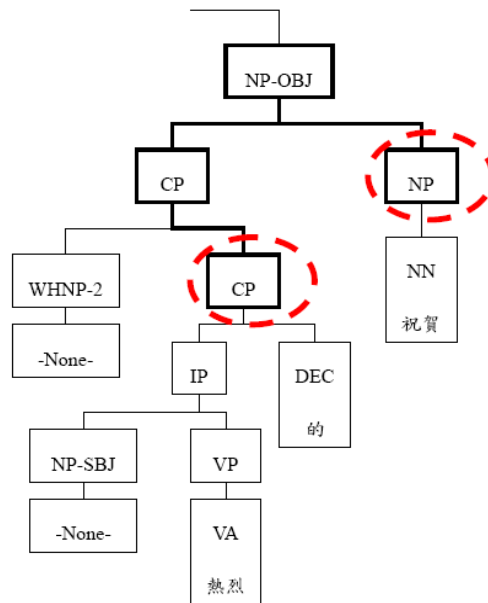


圖 4-11 非同親結構範例

4.6.2. 特徵值抽取

既視為序列式標記問題，此處所稱之特徵值乃指「序列上每一節點」之特徵值。本研究試圖探討以語法分析樹之資訊預測「表達意見之結構」的可能，故並未引用其他外部資訊（如意見詞分數、詞頻等等），而僅以語法結構樹本身為特徵值。所用特徵值可視作「綴於節點下之子樹」：對每一節點而言，均將其下由左而右 4 個子節點之「詞性」與「詞彙本身」（若有的話）納為特徵值（不足者則填入空標籤如「EMPTY」、超過 4 子節點者則僅取由左而右的 4 個節點）；此 4 節點下由左而右 4 個子節點亦納入；再下一層，此 $4*4=16$ 個子節點下，由左而右 3 個子節點再納入。如此便形成一棵底層合有 $3*4*4=48$ 個子節點之子樹，該子樹便為其樹頂（即序列中的一節點）的特徵值。

而序列式標記問題亦有「窗框大小」問題，即標記此點時需考慮前後多少個鄰點。本研究中窗框大小設為 5，故需考慮前後各 2 個相鄰點之特徵值。其特徵值全貌可參圖 4-12：

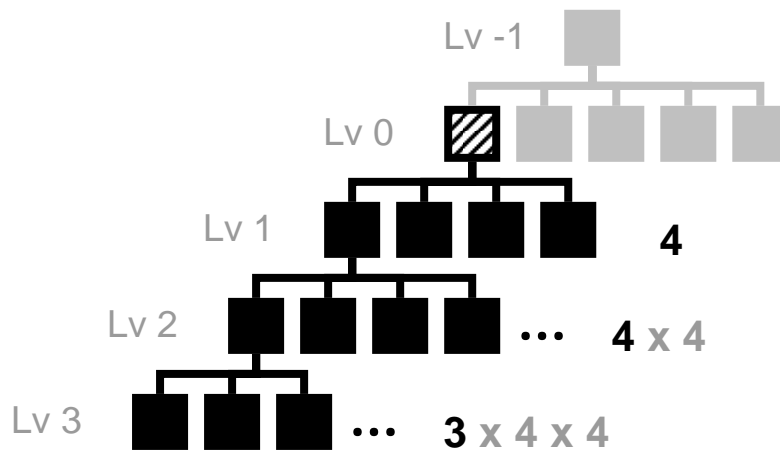


圖 4-12 詞間關係自動擷取特徵值示意圖

4.6.3. 結構自動擷取效能評估

4.6.3.1. 實驗設定

實驗語料準備方面，我們首先將所有意見句之「同親序列」取出（無論有否腳點），將序列中「僅有一腳點」與「含有超過兩個腳點」者視為「不含腳點」，再依節點順序輸出符合 CRF++ 輸入格式之特徵值，作為實驗語料。後續實驗中將進行多組不同標籤集之實驗，僅需將已備好之特徵值檔案取出，加上欲使用之標籤即可。以下實驗均以 4 疊交叉驗證方式進行。

4.6.3.2. 序列類型判斷

本研究首先進行最簡易之評估，即「判斷該序列是否含有腳點」。為尋找較佳之標籤集，我們進行了三組不同標籤的實驗，均於 4.6.3.1 節所準備之語料上進行 4 疊交叉驗證，並評估其效能。三組標籤如下：

(1) 全部 Y, N

無論標記單元結構類型，只要該點為腳點即標為 Y，非腳點即標為 N。

(2) 全部 Y, N, M

無論標記單元結構類型，只要該點為腳點即標為 Y，非腳點者若介於兩腳點間即標為 M、其餘為 N。

(3) N, 2H, 2T, 3H, 3T, ...

該點若為「修飾」之左腳點則標為 2H、右腳點標為 2T；為「主謂」之左腳點標為 3H、右腳點標為 3T；為「動賓」之左腳點標為 4H、右腳點標為 4T；為「動補」之左腳點標為 5H、右腳點標為 5T。其餘為 N。

而判斷方式則分為「嚴格」與「一般」兩種。「一般」乃指「出現非 N 節點即判定該序列含有腳點」，而「嚴格」則指「必須出現 Y 才判定該序列含有腳點」。其評估結果見表 4-4：

表 4-4 序列結構辨識評估

所用標籤集	序列判斷方式	P	R	F
N, 2H, 2T, 3H, 3T, ...	出現非 N 節點即判斷為有	0.6	0.52	0.56
全部 Y, N, M	出現非 N 節點即判斷為有	0.58	0.5	0.54
	必須出現 Y 才判斷為有	0.6	0.48	0.53
全部 Y, N	出現非 N 節點即判斷為有	0.59	0.43	0.5

再者，我們進一步評估「判斷該句是否含有某特定類別腳點」之效能，此部份亦進行三種不同標籤之實驗：

(1) N, 2H, 2T, 3H, 3T...

同上。該點為「修飾」之左腳點標為 2H、右腳點標為 2T；為「主謂」之左腳點標為 3H、右腳點標為 3T；以此類推，其餘標 N。
判定方法：該序列中出現的第一個非 N 節點，即判斷為該類；若全為 N，則判斷為無腳點。

(2) 各類分別標 Y, N, M

對「修飾」、「主謂」、「動賓」、「動補」分別訓練各自的模型。如「修飾」之標記模型會將「N, 2H, 2T, 3H, 3T…」標籤集中標為「2H」與「2T」之腳點標為 Y，兩腳點間標 M，其餘均標 N；而「主謂」模型則會將該標籤集中之「3H」與「3T」標為 Y，以此類推。

判定方法：該序列中出現的第一個 Y 屬於何類模型，即判斷屬於該類；若全為 N，則判斷為無腳點。或兩模型均剛好標示同一節點為 Y 時，則以 CRF++ 輸出之機率值較高者為答案。

(3) 各類分別標 T, H, N, M

同上，差別僅在於將「Y」細分為「H」（左腳點）與「T」（右腳點）。



評估結果如表 4-5：

表 4-5 序列結構類別辨識評估

標籤		修飾	主謂	動賓	動補
N, 2H, 2T, 3H, 3T...	P	0.59	0.53	0.62	0.44
	R	0.41	0.46	0.65	0.13
	F	0.49	0.49	0.64	0.20
各類分別標 Y, N, M	P	0.58	0.50	0.62	0.39
	R	0.41	0.51	0.64	0.11
	F	0.48	0.50	0.63	0.17
各類分別標 T, H, N, M	P	0.58	0.51	0.62	0.39
	R	0.41	0.50	0.64	0.11
	F	0.48	0.50	0.63	0.17

4.6.3.3. 直接擷取腳點

由上述實驗可觀察得知各種標籤集之擷取效能相異不大，考量若以各類分別給予標籤方式進行實驗，需同時產生四個標記模型，較為複雜，故本研究選定「N, 2H, 2T, 3H, 3T…」標籤集實驗結果評估腳點預測效能。我們僅保留該實驗預測結果中的「合法標籤」。「合法標籤」須滿足三條件：

其一，一序列中若有腳點，則必有兩腳點（如 {N, 2H, 2T}）。若一序列中僅有一腳點（如 {N, 2T, N}）或三腳點以上者（如 {N, 2H, 2H, N, 2T}）則不予接受。

其二，此兩腳點必屬同一種結構分類（如 {N, 4H, 4T, N}），若分屬兩種結構分類（如 {N, 3H, 4T, N}）則不予接受。

其三，此兩腳點必須左腳點在左（H 在左）、右腳點在右（T 在右）。無論均為左腳點（如 {2T, 2T, N}）、均為右腳點（如 {N, 2H, 2H, N}）、或左右顛倒（如 {N, 2T, 2H}）均不予接受。

符合此三條件者方會為評估程式判定為實驗所產出之預測結果。選擇此嚴格標準之目的在於探究若僅以結構資訊進行預測，則可掌握多少含有意見之結構，此亦可更深一層反映出語法結構與意見表達之關連性強弱。

前一節之評估均以「序列」為單位，而本節之評估直接以標記者產出之所有「標記單元」為單位（無論該單元是否為同親結構），分析各類結構自動擷取之總效能。預測結果如表 4-6：

表 4-6 腳點位置直接辨認評估

	P	R	F
修飾	1.00	0.25	0.40
主謂	1.00	0.25	0.41
動賓	1.00	0.39	0.56
動補	1.00	0.13	0.23

4.6.3.4. 討論

由上述實驗可觀察得知，無論是序列判斷或腳點擷取之效能均相當有限，且使用不同標籤集之效能亦無顯著差異。探究其原因，或可於腳點直接判斷之實驗結果見其端倪：該實驗結果，回收率極低，精確度卻極高。幾乎是為 CRF 判斷為「合法腳點」者，該位置便確定有腳點存在。我們逐一觀察預測結果，發覺預測之結果多在語法分析樹之末端，即接近詞彙層次部份，且其預測結果，多為某些特定詞彙所構成之腳點，如副詞語尾「地」，或「表示」、「陳述」等在「人講句」型態之意見句中幾乎必會被標的辭彙，亦即 CRF 實則是在辨認詞彙（可能伴以某些同時出現之結構），而非單以語法結構本身進行預測。

本結果或可指出：用於意見表達之語法結構，確有一部份與語法特徵密切相關，然卻亦有另一大部分結構所攜帶之意見來自於詞彙的語義，故以結構為特徵值進行預測，僅可得到一小部份準確度高、回收率低的結果。

4.7 小結

本段研究乃對詞間結構進行了定義、分析與預測。由依存關係之分析（無論為句子統計或標記單元統計），肯認了「用於意見表達之語法結構具有特殊性」此一假設，並找出數種較常用於意見表達之依存關係；我們亦提出了於語法分析樹上標記意見結構之方法，同時也說明了將此標記結果轉為依存關係之簡易步驟。最後於語法分析樹上進行意見結構預測，得到高準確度、低回收率之結果。此預測效能固差強人意，然其高準確度卻仍相當具有價值，如於實際意見分析時可於套用 14 種依存關係前先以 CRF 方法預測之，先行獲取準確度極高之腳點位置，藉以改善整體結構預測之效能。其預測效能固然有限，考量其準確度之高，幫助不可謂之不大。欲改善意見語法結構之擷取效能，或可引進意見詞彙分數等語義資訊，改善純以語法資訊無法完全掌握之部分。

第五章、語法結構應用於意見分析研究

本研究與古學姊倫維合作，筆者著重於結構擷取，而由學姊負責意見分析部份，其成果發表於 (Ku, Huang et al. 2009)。本章將展示由學姊負責之研究內容，即將前述已標記或自動擷取出之詞內、詞間結構資訊，實際應用於意見分析系統中。此成果固非本人之研究貢獻，然其效能卻與本研究密切相關，故特闢一章簡述之。其系統細節及演算法，可參考 (Ku, Huang et al. 2009)，於此僅就相關部分作必要之說明。

5.1 使用構詞資訊之中文詞意見自動分析

詞彙層次之意見傾向判斷實驗乃施作於 (Ku, Liang et al. 2006) 所產出、已標有意見分數之 836 個意見詞上，而每一漢字之意見分數計算方式則以 (Ku, Wu et al. 2005) 所提出之「意見字」算分方法為基礎，於該架構下，每一漢字之意見分數介於-1 至 1 之間，大於 0 者為正向字、小於零者為負向字。引入構詞資訊之方法為：以語言學知識為基礎，為每一構詞類別獨立設計分數計算公式。逐一說明如下：

(1) 並列

此類之詞彙其二字間並無明顯修飾或主從關係，故本類依舊保留原始計算方式，即取二字意見分數之算術平均。公式如下：

$$S(C_1C_2) = \frac{S(C_1) + S(C_2)}{2} \quad (3)$$

公式(3)中， C_1C_2 代表首字為 C_1 、末字為 C_2 之二字詞， $S(C_1C_2)$ 為該二字詞之分數，而 $S(C_1)$ 為首字 C_1 之分數、 $S(C_2)$ 為末字 C_2 之分數。以下公式

或算法所用符號皆同。

(2) 修飾

首先考慮詞彙之正負傾向。修飾一類，無論是修飾字或被修飾字，只要有其中一者為負向，整體詞彙傾向多為負向，「負修飾正」如「苦笑」，「正修飾負」如「好慘」，「負修飾負」如「惡病」，唯「正修飾正」時方為正向。故設計公式時僅檢查「正修飾正」之情況；而接著考量分數絕對值，「修飾」關係中，詞彙之意見強度常取決於修飾子而非被修飾之對象。故我們設計之算分公式如(4)：

$$\begin{aligned} &\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0) \text{ then} \\ &\quad \text{if } (S(C_1) > 0 \text{ and } S(C_2) > 0) \text{ then } S(C_1C_2) = S(C_1) \\ &\quad \text{else } S(C_1C_2) = -1 \times |S(C_1)| \\ &\quad \text{else } S(C_1C_2) = S(C_1) + S(C_2) \end{aligned} \quad (4)$$

而當有字彙之分數為零，則指部份字彙為不帶意見之中性字彙，如「好人」之「人」字、「粗茶」之「茶」字，此時即以分數非零之字彙為整體詞彙之分數。

(3) 主謂

主謂關係中，作主詞用之字彙通常不帶有意見，而謂語表義性通常較強。故以字尾，即作謂語用之字彙，其分數代表整體詞彙之意見分數。如公式(5)：

$$\begin{aligned} &\text{if } (S(C_2) \neq 0) \text{ then } S(C_1C_2) = S(C_2) \\ &\text{else } S(C_1C_2) = S(C_1) \end{aligned} \quad (5)$$

(4) 動賓

以「抗菌」為例，該字之正負向實為組成字之正負向乘積，而觀察此類詞彙，其意見正負向大抵如此，「正負」如「治病」，「正正」如「獲獎」，「負正」如「失寵」，以上詞例之正負向均為組成字之正負向乘積，正正得正，負正得負。是以我們便以述語（即首字）之絕對值加上組成字之正負符號

乘積為本類之詞彙分數。公式如(6)，其中 $SIGN()$ 為回傳正負號之函數。

$$\begin{aligned} &\text{if } (S(C_1) \neq 0 \text{ and } S(C_2) \neq 0) \\ &\quad \text{then } S(C_1C_2) = |S(C_1)| \times SIGN(S(C_1)) \times SIGN(S(C_2)) \\ &\quad \text{else } S(C_1C_2) = S(C_1) + S(C_2) \end{aligned} \quad (6)$$

(5) 動補

動補中，「補語」修飾「動詞」，其修飾子之分數可代表整體詞彙之分數，與主謂一類相同，故沿用其公式，即公式(5)。

(6) 否定

公式(7)之 NC ，指由否定子（如「非」、「否」、「不」）所構成之集合。即當首字為否定子時，尾字之分數變號後即為整體詞彙之意見分數。

$$\begin{aligned} &\text{if } (C_1 \in NC) \text{ then } S(C_1C_2) = (-1) \times S(C_2) \\ &\text{else } S(C_1C_2) = (-1) \times S(C_1) \end{aligned} \quad (7)$$

(7) 肯定

公式(8)之 PC 指由肯定子（「有」）構成之集合，當首字為肯定子時，詞尾字之意見分數即為整體詞彙之分數。

$$\begin{aligned} &\text{if } (C_1 \in PC) \text{ then } S(C_1C_2) = S(C_2) \\ &\text{else } S(C_1C_2) = S(C_1) \end{aligned} \quad (8)$$

(8) 其他

由於此類之構詞方式並無固定特徵，故沿用原算法，與「並列」同，見公式(3)。

公式設計完成後，以（Ku, Wu et al. 2005）所計算得出之字彙意見分數，加以本研究所自動判斷之類別，即可計算出詞彙之意見分數。我們於（Ku, Liang et al. 2006）所產出之 836 個意見詞上進行實驗，輸入一辭彙後，系統輸出「正面」、「負面」、「中性」三種意見傾向，再針對答對與否進行 F-score 評估。評估結果見表 5-1：

表 5-1 套用構詞資訊之意見詞極性判斷評估

實驗設定	意見傾向 F 分數	
	不使用意見詞辭典	使用意見詞辭典
原始意見效能	0.5455	0.5789
使用 CRF 構詞分類	0.5806	0.6100

「意見詞辭典」是指 (Ku, Wu et al. 2005) 建立之意見詞辭典，而使用辭典即指若某詞已列於辭典中，則直接將辭典中之分數回傳為答案。由表 5-1 中可看出，使用構詞分類後，無論是否加用意見詞辭典，其意見傾向判斷之效能均有上升（皆有通過信賴區間為 95% 之 t-test）。而經實驗發現，其中「修飾」與「動賓」兩類對詞彙層次之意見傾向判斷改善最鉅。

5.2 使用詞間結構資訊之中文句子層次意見分析

句子層次方面，使用詞間結構資訊之方法可參圖 5-1（見 77 頁）。即依語法分析樹結構將詞彙之意見分數逐步向上累加，而分數累計至腳點時，腳點會將自身分數先乘以 5，再向上傳送。此法之精神在於將腳點之分數加重計算，使意見部份分數大幅提升，藉以修正原本未考慮結構時之狀況。

本研究於 4.3.1 節所標記之意見句上以此法進行實驗。評估項目有二：判斷該具是否為意見句，以及判斷該句之意見傾向；實驗設定方面，句子與詞彙層次可分別有「原始」、「結構（自動）」、「結構（標記者答案）」三種設定。「原始」指直接將意見分數相加，即原始意見分析系統之設定；「結構（自動）」指以自動分類後之結構類別資訊輔助計分；「結構（標記者答案）」指以標記者所標之正確結構答案為資訊輔助計分（詞彙層次並無此選項，因本研究並未對意見句中所有意見詞進行標記）。其結果如表 5-2（見 77 頁）。

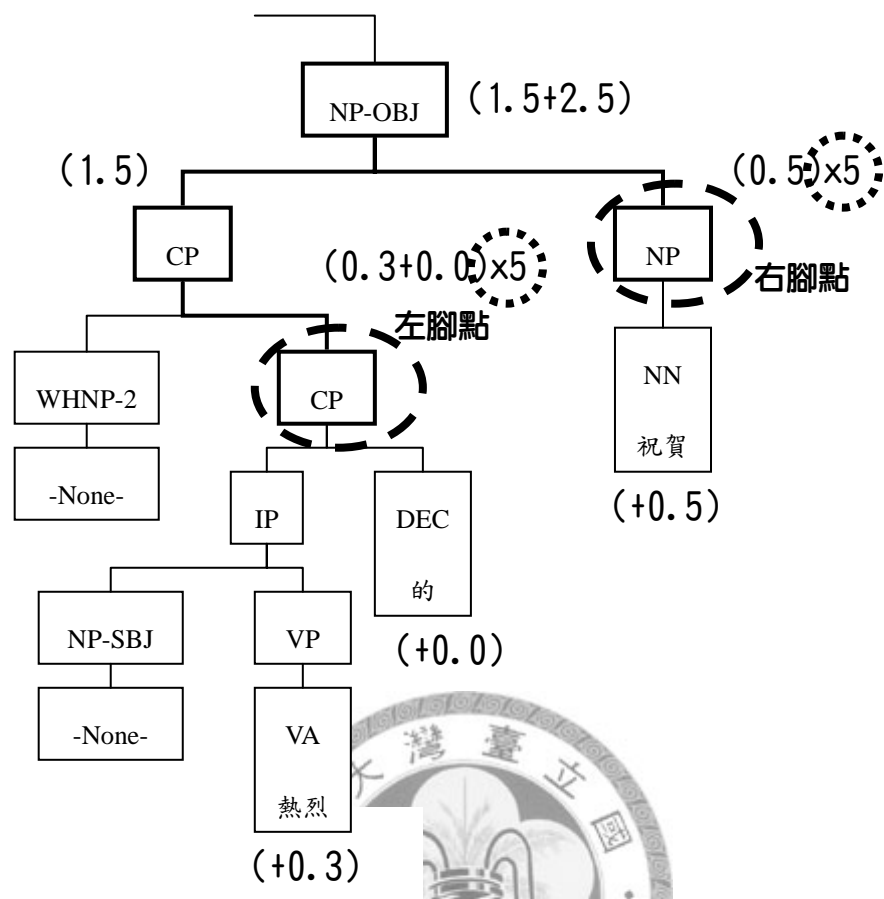


圖 5-1 詞間結構使用方法範例

表 5-2 套用詞間結構資訊之意見句判斷評估

實驗設定		F 分數	
詞彙層次	句子層次	意見句判斷	意見傾向判斷
原始	原始	0.7073	0.4988
結構 (自動)	原始	0.7162	0.5117
原始	結構 (自動)	0.8000	0.5361
結構 (自動)	結構 (標記者答案)	0.7922	0.5297
結構 (自動)	結構 (自動)	0.7993	0.5187

由表 5-2 可知，無論何種實驗設定，加入結構資訊均可使句子層次之意見判斷效能提升，表明引入語法結構此一方法確實有助意見分析。

第六章、總結與展望

於內容言，本研究主要貢獻可分為三部份：

(1) 語料標記

於詞內層次，本研究產生了一組質量均備之標記語料，可供分析及實驗之用；而於詞間層次，我們亦設計了一於語法分析樹上之標記方法，該標記結果可直接用於預測與計算，亦可轉換為依存關係。

(2) 語料分析

以標記完善之語料為基礎，我們進一步分析標記結果。於詞內構詞分佈方面，比較了本研究與其他研究團隊之標記異同，亦探討了本問題於標記者間之信度；而詞間結構方面，本研究經語料分析後證實了用於意見表達之結構有其特殊性，並仔細分析較常用於意見表達的依存關係種類。

(3) 結構預測

詞內層次方面，我們提出了一組特徵值，並以各種不同分類器進行分類實驗，得到五類平均 F 分數約為 0.6 的效能；而於詞間層次，本研究一方面指出了對意見分析較為有用之 14 種依存關係，另一方面亦就標記結果直接於語法分析樹上進預測，並得到高精確度、低回收率之預測結果。

意見分析問題上，本研究之主要貢獻在於：提出「以語法結構改善意見分析效能」之方法，並透過語料標記、分析、實驗，實際測試其效能。

未來展望方面，我們希望能更細節地使用語法結構資訊，並以加入語義資訊

之方式輔助意見結構擷取，以期能更大幅度地改善意見分析之效能。

本研究發表於 *EMNLP 2009* : Ku, Lun-Wei, Huang, Ting-Hao and Chen, Hsin-Hsi.
(2009). *Using Morphological and Syntactic Structures for Chinese Opinion Analysis*.
Proceedings of Conference on Empirical Methods in Natural Language Processing,
Singapore.



參考文獻

"MINIPAR Parse Visualization Tool." From

http://ai.stanford.edu/~rion/parsing/minipar_viz.html

CIRB040: "NTCIR-6 Test Collections: Documents." From

<http://research.nii.ac.jp/ntcir/ntcir-ws6/data-en.html>

. "The Penn Treebank Project." from <http://www.cis.upenn.edu/~treebank/>.

. "The Stanford Parser: A statistical parser." From

<http://nlp.stanford.edu/software/lex-parser.shtml>.

. "教育部重編國語辭典修訂本." from <http://dict.revised.moe.edu.tw/>.

(2007). CRF++: Yet Another CRF toolkit. From <http://crfpp.sourceforge.net/>

Chang, C.-C. and C.-J. Lin (2001). LIBSVM : a library for support vector machines.

Ku, L.-W., T.-H. Huang, et al. (2009). Using Morphological and Syntactic Structures for Chinese Opinion Analysis. Conference on Empirical Methods in Natural Language Processing, Singapore.

Ku, L.-W., Y.-T. Liang, et al. (2006). Opinion extraction, summarization and tracking

in news and blog Corpora. Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report.

Ku, L.-W., Y.-S. Lo, et al. (2007). Test Collection Selection and Gold Standard Generation for a Multiply-Annotated Opinion Corpus. Proceedings of 45th Annual Meeting of Association for Computational Linguistics, Prague, Czech Republic.

Ku, L.-W., T.-H. Wu, et al. (2005). Construction of an Evaluation Corpus for Opinion Extraction. NTCIR 2005.

Lafferty, J., A. McCallum, et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML.

Lu, J. (2008). Chinese Synthetic Words Analysis. Department of Information Processing, Graduate School of Information Science Nara Institute of Science and Technology. master: 72.

Lu, J., M. Asahara, et al. (2008). Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Morphologically Derived Words. The Sixth SIGHAN Workshop on Chinese Language Processing.

McCallum, A. (1998). Rainbow.

Qiu, G., K. Liu, et al. (2007). Extracting opinion topics for Chinese opinions using dependence grammar. Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising. San Jose, California, ACM.

Qiu, G., C. Wang, et al. (2008). Incorporate the Syntactic Knowledge in Opinion Mining in User-generated Content. NLPIX2008 (In conjunction with WWW'08).

Tseng, H. and K.-J. Chen (2002). Design of chinese morphological analyzer. the First SIGHAN Workshop on Chinese Language Processing.

Tseng, H., D. Jurafsky, et al. (2005). Morphological features help POS tagging of unknown words across language varieties. the Fourth SIGHAN Workshop on Chinese Language Processing.

亢世勇 (2001). "《現代漢語新詞語信息(電子)詞典》的開發與應用." 辭書研究 2001(2): 55-63.

亢世勇 (2001). "《現代漢語語法信息詞典》的特點與不足." 辭書研究 2001(6): 79-116.

亢世勇 (2002). "《現代漢語新詞語資訊電子詞典》的研究與實現." International Journal of Computational Linguistics & Chinese Language Processing 7(2): 89-100.

亢世勇 (2003). "《新詞語大詞典》的編纂." 辭書研究 2003(3): 12-20.

亢世勇, 徐豔華, et al. (2005). 基於語料庫的現代漢語新詞語構詞法統計研究. International Conference on Chinese Computing, Singapore.

亢世勇, 許小星, et al. (2005). "現代漢語語義構詞規則初探." 漢語語言與計算學

報 15(2): 103-112.

王惠 and 朱學鋒 (1994). 《現代漢語語法電子詞典》的收詞原則. 中國計算機報: 79-83.

石秀雙 (2007). "現代漢語雙音復合詞結構關係考察——以 z 字母下雙音復合詞為例進行分析." 晉中學院學報 2007(6): 1-8.

朱學鋒, 俞士汶, et al. (1995). "現代漢語語法信息辭典的開發與應用." 中文與東方語言信息處理學會通訊 1995(2): 81-86.

朱學鋒, 俞士汶, et al. (1999). "漢語語素庫的構造及其同語法信息詞典的集成." 術語標準化與信息技術 1999(2): 36-40.

李普霞 and 劉雲 (2004). "新版《現代漢語語法信息詞典詳解》的貢獻." 辭書研究 2004(3): 64-70

俞士汶, 朱學鋒, et al. (2001). "《現代漢語語法信息詞典》的新進展." 中文信息學報 15(1): 59-65.

俞士汶, 朱學鋒, et al. (1999). "現代漢語語素庫的開發及應用." 世界漢語教學 1999(2): 38-45.

苑春法 and 黃昌寧 (1998). "基於語素數據庫的漢語語素及構詞研究." 語言文字應用 1998(3): 83-88.

傅建紅 (2009). "論《現代漢語詞典》F類雙音複合詞的結構關係." 現代語文 2009(3): 49-50.

傅愛平 (2003). "漢語信息處理中單字的構詞方式與合成詞的識別和理解." 語言文字應用 2003(4): 25-33.

程祥徽 and 田小琳 (1995). 現代漢語, 三聯書店 香港.

劉雲, 俞士汶, et al. (2000). 現代漢語合成詞結構數據庫. 第二屆中文電化教學國際研討會, 廣西師範大學出版社.

穆克婭 (2008). "新雙音節複合動詞語素構詞規律研究." 現代語文 2008(12): 42-44.



附錄 A：常用譯名對照表

英文	本文主要譯詞	別名或簡稱
4-fold cross-validation	4 疊交叉驗證	4 疊交叉效度、4 折交叉效度
character	(中文)字	
CRF	條件隨機域	條件隨機場
data sparse	資料空缺	資料稀疏
dependency relation	依存關係	依賴關係、依靠關係
dependency tree	依存關係樹	依存樹、依賴樹
feature	特徵值	特徵
micro-average	微觀平均	
morpheme	語素	詞素
parsing	剖析	語法剖析
Penn Treebank	賓州大學樹庫	賓大樹庫
primary key	主鍵	主索引值
sense	義項	義條
SVM	支援向量機	支撐向量機、支持向量機、支量機
word	(中文)詞	

附錄 B：未使用之賓大樹庫句子清單

FID	SID	原因
40	488	兩棵樹
112	1507	兩棵樹
139	1899	兩棵樹
307	3884	兩棵樹
307	3886	兩棵樹
437	4681	兩棵樹
672	6885	兩棵樹
733	7663	兩棵樹
787	8407	兩棵樹
792	8447	兩棵樹
793	8482	兩棵樹
794	8498	兩棵樹
828	9014	兩棵樹
845	9275	兩棵樹
855	9426	兩棵樹
877	9758	兩棵樹
877	9760	兩棵樹
1042	12973	兩棵樹
1048	13373	兩棵樹
1078	15129	兩棵樹