基於統計方法之中文搭配詞自動擷取

張翠芸、柯淑津

東吳大學資訊科學系
Department of Computer Information Science
SooChow University
ms9513@sun.cis.scu.edu.tw
ksj@cis.scu.edu.tw

摘要

本研究採取以下四個步驟擷取出雙連詞、三連詞、四連詞之詞彙或詞性組合之搭配詞。首先採用 Smadja's Xtract 的平均數及變異數的方法，擷取具有變動距離模式所共同出現的詞彙或詞性的組合，接著使用搭配詞顯著性的衡量方法：相互資訊值及 T 檢定值。通過以上檢驗的候選搭配詞，經由對照中央研究院詞義標示語料庫之目標詞的結果，在同樣的跨距下，若同爲一個詞義者，則我們以此搭配詞作爲詞義標示知識。並且，本研究將產出之搭配資訊應用於詞義自動標示處理，達到 20.07%的應用率及 90.83%的正確率。

Abstract

We take the four following steps to extract collocations made of combinations of 2, 3, 4 words and/or part of speech, respectively. First, we use "Smadja's Xtract" to extract the co-occurrence combinations of words and/or part of speech of varying distance by computing means and variances. Second, we evaluate the significances of collocation candidates by 2 metrics: mutual information and t-test value. At last, we compare the head words of tagged word sense corpus made by Academic Sinica with the collocation candidates. If in the same distance, the head words of collocation candidates match the ones made by Academic Sinica, we say they are collocations. In addition, we apply the collocation information produced from this research to word sense disambiguation. It reaches application rate of 20.07% and precision rate of 90.83%.

關鍵詞：中文搭配詞，相互資訊值，自然語言處理，統計方法，T 檢定值，詞義辨識

Keywords: Chinese collocation, mutual information, natural language processing, statistical method, t-test, word sense disambiguation.

一、簡介

不同民族的歷史文化知識背景以及人們的思考邏輯模式不同，看待同樣的人事物、同樣的行爲情境過程，在語言的描述上也會有所不同。每個地區的語言都有其習慣性的用

法，而所謂的搭配詞 (collocation) 廣義而言，就是指兩個或多個詞依照語言習慣性結合在一起表示某種特殊意涵的詞彙現象。搭配詞在不同的研究領域上各有不同的解讀，尚未有一致性的定義。研究搭配詞著名的學者 Smadja [1] 定義搭配詞有以下四個特徵：1、搭配詞是任意詞的組合；2、搭配詞和領域相關；3、搭配詞是重複出現的；4、搭配詞具有詞彙的互相吸引性。母語使用者對於搭配詞的判定也許相當容易，但對於外國人的語言學習，常會誤用搭配語詞。以往對於搭配詞自動擷取的研究，大多是針對英語系語料做處理。至於擷取中文搭配詞的相關文獻仍然是相當稀少的，因此本研究利用統計的方式對大規模的中文資料進行分析以擷取出中文搭配詞。其產出的結果將可以應用在自然語言相關處理上，例如：詞義自動標示、資訊檢索、機器翻譯以及辭典編纂。

本研究提出將周邊詞彙及詞性皆作爲擷取搭配詞的重要特徵，採用 Smadja's Xtract [1] 基於統計上的平均數及變異數之方法，直接擷取出具有變動距離模式所共同出現的詞彙或詞性之組合，再使用搭配詞顯著性的衡量方法：相互資訊值 (Mutual Information) 和 T 檢定值。通過以上檢驗的候選搭配詞，在最後判定搭配詞的基準，是基於每個搭配詞僅一個詞義的原由 [2]，我們採取經由對照中央研究院詞義標示語料庫 SSTC (Sinica Sense-Tagged Corpus) [3]，在相同目標詞彙和周邊詞彙資訊的跨距下，目標詞在語料庫訓練資料的所有詞例中，若僅具唯一詞義，則我們就將此搭配詞擷取爲詞義標示知識；進一步再以相同的方式進而擷取三連詞及四連詞之搭配詞。最後我們將產出之搭配資訊應用於詞義自動標示處理。

本文組織如下，第二節是有關搭配詞擷取技術之相關文獻探討。第三節說明本研究提出之擷取搭配詞方法。第四節爲實驗設計與結果評估。最後，是本文的總結。

二、相關文獻

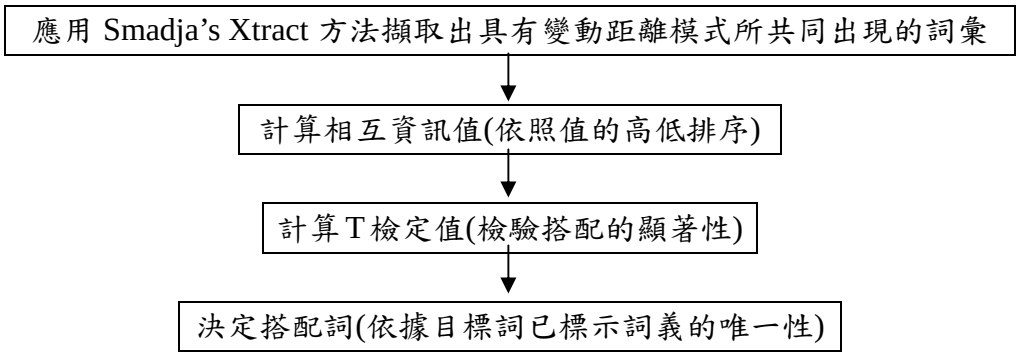
根據統計方法擷取搭配詞的相關文獻中，Smadja's Xtract [1] 採用平均數及變異數的方法於英文的語料中擷取雙連詞，並由雙連詞之結果擴增擷取 n 連詞，此方法被認爲是擷取搭配詞的經典方法。Breidt [4] 將相互資訊值及 T 檢定結合使用於德文的語料中擷取動詞-名詞的搭配詞。在中文的搭配詞研究中，Lu [5] 等人的 CXtract 研究中應用 Smadja's Xtract 的方法於中文語料中，但其研究過程所設置的門檻值會將一些極有可能爲搭配詞的周邊詞過濾掉。將搭配詞應用於其它自然語言處理的相關領域之研究，車方翔 [6] 等用平均數、變異數及 T 檢定的方法得到詞與詞之間搭配強度係數，並將此結果應用於搜索引擎中縮減檢索句子中的歧義度。全昌勤 [7] 等利用搭配詞典的輔助獲取最優種子，再由最優種子自動學習擴充指示詞集，有助詞義辨識之處理。有關詞義辨識的相關研究中，其中以語料爲基礎的監督式學習法是最爲成功的方法，主要是依據上下文的特徵來區別歧義詞，但因上下文共同出現的詞彙數量太多，若全都做爲訓練的樣本會使得雜訊很多，在標示歧義詞時則容易標示錯誤。Li [8] 提出縮小上下文的範圍，使用搭配詞作爲特徵，並且基於搭配詞的歧義詞詞義唯一性的概念，在標示歧義詞時，當上下文擷取到搭配詞時，上下文中其它詞彙的影響性將被減少。國內針對擷取搭配詞的相關研究，主要使用的資源分爲兩大類，第一，將網路視爲具有時間性的語料庫資源，Chen [9] 等人利用網路流量紀錄和 Google 搜尋引擎以擷取搭配詞，Teng [10] 等人利用網路部落格觀察時間性和搭配詞之間的關聯；第二，利用平行語料庫 [11, 12, 13]，根據語言的特徵和統計分析的方法，取得英文的搭配詞結構，進而擷取雙語搭配詞。

有別於過去的研究僅能擷取出詞彙的搭配詞或是固定樣式的搭配詞結構，如動詞與

名詞、形容詞與名詞等。本研究提出考量視窗範圍內周邊詞彙或其詞性之組合，基於 Smadja's Xtract 的演算法和相互資訊值、T 檢定值之統計檢驗的方法，以及大規模中文詞義標示語料庫 SSTC [3] 的輔助，以擷取出雙連詞、三連詞、四連詞之搭配詞。

三、自動擷取搭配資訊方法

本研究所提出之自動擷取搭配資訊處理方法如圖一所示，首先採用 Smadja's Xtract 的演算法 [1, 5]，擷取出詞語間間隔其它詞彙所共同出現的候選搭配詞，接下來採用相互資訊值及 T 檢定值的方式檢驗所擷取出的候選搭配詞在語料庫中共同出現的顯著程度，最後爲搭配詞結果的判定，我們對照中央研究院 SSTC 詞義標示語料庫，若目標詞詞義具有唯一性者，則認定其爲搭配詞。



圖一、自動擷取搭配資訊處理方法之流程圖

（一）擷取具變動距離模式之共現詞彙

首先設定目標詞，設置以句子爲單位，編輯目標詞之周邊詞彙跨距爲 d 的視窗內周邊資訊。在目標詞的 $\pm d$ 跨距內的周邊詞稱作 w_i $(1 \le i \le n$，n 爲所有周邊詞的個數)；設定 w_i 在第 j 個位置 (與目標詞的距離) 出現的次數定義爲 $f_{i,j}$；周邊詞 w_i 在目標詞 $\pm d$ 跨距內總共出現的次數定義爲 $f_i = \sum_{-d}^{d} f_{i,j}$；$f_i$ 的平均次數爲 $f_i = \sum_{-d}^{d} f_{i,j} / 2d$；針對每一個目標詞，平均次數 $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$ 和標準差爲 $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2}$；周邊詞在目標詞 $\pm d$ 跨距內總共出現的次數經標準化後定義爲 $k_i = \frac{f_i - \overline{f}}{\sigma}$；周邊詞在目標詞 $\pm d$ 跨距內出現次數之變異數定義爲 $U_i = \frac{\sum(f_{i,j} - \overline{f})}{2d}$，表示周邊詞分佈的特徵。爲了過濾不太可能爲搭配詞的組合，設定 (K_0, K_1, U_0) 的經驗門檻值，以下列三個條件 [1] 作爲過濾的依據：

$$C_1 : k_i = \frac{f_i - \overline{f}}{\sigma} \geq K_0 \quad (1)$$

$$C_2 : u_i \geq U_0 \quad (2)$$

$$C_3 : f_{i,j} \geq \overline{f_i} + (K_1 \cdot \sqrt{u_i}) \quad (3)$$

針對上述三個條件判斷分述如下：C_1 條件是衡量周邊詞在目標詞 $\pm d$ 跨距內所出現的次數，過濾掉共現次數太低的周邊詞；C_2 條件是衡量周邊詞在目標詞 $\pm d$ 跨距內各個位置的分佈情形，若周邊詞在各個位置分佈過於分散且次數平均，則將其過濾掉，留下出現次數在各個位置上具有變異性較大的周邊詞。C_3 條件則擷取出周邊詞在目標詞 $\pm d$ 跨距內出現次數較爲突出的位置。並且基於搭配詞必須出現於唯一且固定位置之原由，所以經 Smadja’s Xtract 門檻值過濾後的候選搭配詞，若是針對同一個目標詞，相同周邊詞出現於不同位置者，我們則將此候選搭配詞刪除，認定其不爲搭配詞。

（二）相互資訊值

接著，我們採用衡量兩個事件相關程度的相互資訊值 [14]，其用來表示兩個詞彙間，一個詞出現所帶給另一個詞出現的資訊量。相互資訊值的計算方式如公式 (4):

$$MI(x, y) = \log \frac{P(x, y)}{P(x) \cdot P(y)} \quad (4)$$

經由第一步驟的方法過濾後，我們再計算目標詞彙與周邊詞彙之相互資訊值，並將相互資訊值太低者自搭配候選行列中排除。

（三）T 檢定值

爲了確定搭配詞的顯著程度，我們採用假設檢定中的 T 檢定值 [14] 來檢驗候選搭配詞在語料庫中共現的顯著程度。首先需設定虛無假設：兩個共同出現的詞彙之間互爲獨立，不能形成搭配。T 檢定值的計算方式如公式 (5)，其中 $\overline{x}$ 為樣本平均數；若虛無假設爲真，事件受到伯努力試驗 (Bernoulli trial) 的影響，則平均數 $\mu = p$；變異數 $s^2 = p(1\text{-}p) \approx p$。

$$t = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{N}}} \quad (5)$$

若 T 檢定值大於臨界值，則我們將會拒絕虛無假設，而得出結論：候選搭配詞在語料庫中共同出現是具有顯著性的。假設 T 檢定值小於臨界值，則我們沒有充分證據顯示其爲搭配詞，所以將會過濾掉此候選搭配詞。

（四）決定搭配詞

基於搭配詞的單一詞義特性 [2]，進一步我們利用由中央研究院 SSTC 詞義標示語料庫 [3] 的資源，將前三階段檢驗後的結果，去判斷目標詞與周邊詞在相同位置的結合下，目標詞在已標示詞義語料庫中是否皆爲同一個詞義，若目標詞具有歧義者，我們再作進一步的過濾；由於語言的多樣性，SSTC 詞義語料庫的資源仍屬有限無法概括所有檢驗詞彙，因此，若是目標詞在該語料中未找到標示詞義者，我們也暫時將其過濾。

四、實驗

本節首先介紹我們所使用的語料，以及實驗的設計、參數的設置，實驗結果，並針對實驗結果進行評估。

（一）實驗語料

本研究實驗語料使用《中央研究院語料庫現代漢語平衡語料庫（3.0 版）》，其包括 9,227 篇文件，每個句子已經斷詞處理，且標記詞性，全部約 500 萬詞，語料平衡分佈在不同的媒體（如報紙、學術論文、視聽媒體、演說等）、語式（如書面體、演講稿、會議記錄等）以及主題（科學、哲學、社會、藝術、生活、文學）上，是個適於中文相關處理的代表性語料庫。

（二）實驗設計

本實驗依據圖一的處理方法流程圖之步驟，考量詞彙或詞性的組合，擷取出雙連詞、三連詞、四連詞之搭配詞。擷取搭配詞的範圍設置，跨距指的是目標詞左右距離所跨的範圍，其可依照語料的不同作調整，本實驗因基於搭配詞不跨越標點符號之原則，依據我們的語料中顯示，標點符號與標點符號間的詞彙平均約爲 5.56 個，因此我們將跨距 d 設定爲 5。在擷取三連詞及四連詞的前置處理必須分別先將目標詞 (-5, -1) 和 (1, 5) 跨距內的周邊詞組合成雙連詞 (bigram) 及三連詞 (trigram)，周邊詞可以任意取詞彙或是詞性。

首先採用 Smadja’s Xtract [1] 的演算法，在條件過濾時採用的門檻值 (K_0, K_1, U_0) 設定爲 (1, 1, 10)，而在 Lu [5] CXtract 研究中，設定 (K_0, K_1, U_0) 爲 (1.2, 1.2, 12)，但根據我們透過 SSTC 語料的觀察，當參數 (K_0, K_1, U_0) 設定爲 (1.2, 1.2, 12) 時，會過濾掉一些極有可能爲搭配詞的周邊詞。所以，最後我們仍採用 Smadja’s Xtract 研究的門檻值 (1, 1, 10)。由於使用 Smadja’s Xtract [1] 過濾方法後的周邊詞，仍有許多出現次數極高的停用字 (stopword) ，例如「的、了」。因此，我們再考量其他統計衡量方法，如：相互資訊值、假設檢定 (T 檢定、卡方檢定、對數概似率檢定等)，我們最後採用相互訊息值衡量兩個詞彙間的搭配強度，並依照相互資訊值之大小作排序，可以降低停用字在候選搭配詞的排名，我們將相互資訊值小於 3 者之候選搭配詞排除，而且在我們的實驗中已使用 Smadja’ s Xtract [1] 的條件 C_1 過濾低頻詞，所以本實驗使用相互資訊值的法，並不會受到相互資訊值容易將語義相近而並非爲搭配的低頻詞組合在一起之缺點的影響。本實驗再利用 T 檢定值檢驗候選搭配詞在語料庫中共同出現是否顯著，而其顯著水準設定爲 $\alpha = 0.005$，若 T 值 < 2.576，則我們沒有充分證據顯示其爲搭配詞，所以將會過濾掉候選搭配組合。表一爲以目標詞「關係」(詞性 Na) 擷取詞彙或詞性之雙連詞

搭配爲例，經過 Smadja's Xtract [1] 的條件過濾及進一步計算相互資訊值、T 值的結果。

表一、經過條件過濾及計算相互資訊值、及 T 值的結果

目標詞	特徵	位置序	Ki 值	Ui 值	目標詞與周邊詞共同出現次數	目標詞在語料出現次數	特徵在語料出現次數	MI 值	T 值
關係 Na	人際	-1	1.96	2732.36	175	2945	246	7.25	13.22
關係 Na	兩岸	-1	1.55	1189.04	117	2945	601	5.95	10.79
關係 Na	密切	-2	1.33	189.84	42	2945	305	5.61	6.46
*關係 Na	沒有	-1	1.60	764.20	95	2945	9775	2.95	9.24
*關係 Na	的	-1	21.79	57239.05	817	2945	296183	1.69	23.33
*關係 Na	A	-1	1.80	438.56	76	2945	31426	1.56	6.89
*關係 Na	T	1	1.47	652.69	88	2945	46697	1.31	6.86
*關係 Na	Nv	-1	5.16	2372.21	178	2945	95563	1.30	9.71
*關係 Na	也	1	1.11	117.61	34	2945	29759	0.81	3.24
*關係 Na	Nep	-1	2.68	395.65	61	2945	67325	0.58	3.44
*關係 Na	Na	-1	38.68	52626.69	858	2945	1025464	0.50	11.55
*關係 Na	個	-4	1.21	66.64	27	2945	41143	0.26	1.18
*關係 Na	VK	-5	1.13	85.49	31	2945	57286	0.07	0.35

(*代表 MI 值或 T 值低於門檻值)

最後判定搭配詞，我們將 SSTC 詞義標示語料的資源，隨機取 4/5 爲訓練資料，1/5 爲測試資料，在此以表二目標詞「關係」(詞性 Na) 擷取搭配詞之雙連詞、三連詞、四連詞之部分結果爲例，表二中的搭配組合「人際 關係」在 SSTC 詞義標示語料裡的訓練資料中佔有 3 筆，且目標詞「關係」皆相同詞義，我們判定其爲搭配詞；「兩岸 關係」因在訓練資料中，缺乏已標示資源，我們無法判定其是否爲搭配詞，在此，也將其暫時排除；而在搭配組合「與 Na DE 關係」在訓練資料中，目標詞「關係」有 2 個詞義（具有歧義），則我們判定其不爲搭配詞。

（三）實驗結果

本實驗以目標詞「關係 Na」、「好 VH」、「看 VC」、「講 VE」、「說 VE」爲例，依據 SSTC 詞義標示語料裡的訓練資料，擷取出詞彙或詞性結合之搭配詞完整結果置於附錄。部分結果如表三所示。

表二、由 SSTC 詞義標示語料判定是否爲搭配詞

目標詞	特徵	位置序	目標詞詞義	筆數	*是否爲搭配詞
關係 Na	人際	-1	1.普通名詞。人和人之間在社會或群體中的關聯性。	3	∨
	兩岸	-1	無標示資源		?
	密切	-2	1.普通名詞。事件之間的關聯性。	1	∨
	VC_人際	-2	1.普通名詞。人和人之間在社會或群體中的關聯性。	1	∨
	也_沒有	-2	1.普通名詞。事件之間的關聯性。	6	∨
	因_Na	-2	1.普通名詞。事件發生的原因。	2	∨
	很_大	-3	1.普通名詞。事件之間的關聯性。	1	∨
	密切_DE	-2	1.普通名詞。事件之間的關聯性。	1	∨
	Caa_Na_DE	-3	1.普通名詞。事件之間的關聯性。	2	∨
	Dfa_大_的	-3	1.普通名詞。事件之間的關聯性。	3	∨
	很_大_的	-3	1.普通名詞。事件之間的關聯性。	1	∨
	與_Na_DE	-3	1.普通名詞。人和人之間在社會或群體中的關聯性。	1	×
			2.普通名詞。事件之間的關聯性。	2	

*「是否爲搭配詞」欄位－標記 ∨ 代表經由我們的實驗結果判定爲搭配詞；
標記 × 代表經由我們的實驗結果判定爲不是搭配詞；
標記 ？ 代表因語料詞義標示資源不足，所以無法判別。

表三、依據 SSTC 詞義標示語料裡的訓練資料，擷取搭配詞之部分結果

目標詞	搭配詞
關係 Na	人際 關係、VC 人際 關係、密切 DE 關係、Caa Na DE 關係
好 VH	更 好、心情 D 好、Cbb Na Dfa 好、D 有 多 好、好 DE 方法、做 DE D 好
看 VC	看 書、看 電視、看 Di Nb DE、Nh 來 看、看 Di Nh Neu、看 著 Na Ncd
講 VE	聽 Nh 講、講 DE SHI Na、對 我 D 講、講 得 很 VH、D 跟 Nh 講
說 VE	說 不 出 話、跟 Nh 講 說、說 Di Neu 句、他 笑 Di 說、Nh 告訴 Nh 說

（四）評估

本實驗從 SSTC 詞義標示語料隨機取出 1/5 作爲測試資料，並且採用應用率及正確率作爲評估的準則，公式如(6)：

$$\text{應用率}=\frac{\text{標上的筆數}}{\text{測試資料中包含目標詞的筆數}}\times 100\% \qquad \text{正確率}=\frac{\text{正確的筆數}}{\text{標上的筆數}}\times 100\% \qquad (6)$$

我們從訓練資料中利用搭配詞的約束力區別目標多義詞之詞義，並將上述判定爲搭配詞的結果及目標詞詞義標示於測試資料中，實驗結果如表四所示，在 543 個測試句中標出 109 個句子，其中有 99 個標示結果爲正確標示，因此總計達到 20.07%的應用率及 90.83%的正確率。

表四　以搭配知識進行詞義自動標示之實驗結果

詞彙	詞性	測試句數	標示句數	正確句數	應用率	正確率
好	VH	119	19	19	15.97%	100.00%
看	VC	121	17	11	14.05%	64.71%
說	VE	206	34	33	16.50%	97.06%
講	VE	75	32	31	42.67%	96.88%
關係	Na	22	7	5	31.82%	71.43%
		543	109	99	20.07%	90.83%

五、結論

本研究擷取詞彙或詞性組合之搭配詞。首先爲了擷取雙連詞之搭配詞，利用 Smadja's Xtract 的平均數及變異數的方法擷取被其它詞彙間隔之共同出現的詞彙資訊，進而計算相互資訊值，依照搭配強度高低排序，再由 T 檢定值檢驗候選搭配詞的顯著性。我們將這三種方法結合使用，若在這三種方法下通過考驗的候選搭配詞，我們再經由對照中央研究院人工標示詞義語料庫之目標詞的結果，若在同樣目標詞和周邊詞的跨距下，目標詞均爲同一個詞義者，則我們就判定其確實爲搭配詞；同樣，我們也以同樣的方式擷取三連詞及四連詞之搭配詞。最後，我們將擷取出的搭配詞資訊應用於語義辨識，達到 20.07%的應用率及 90.83%的正確率。

參考文獻

[1] F. Smadja, "Retrieving Collocation From Text: Xtract," *Computational Linguistics*, Vol. 19, No. 1, pp. 143-177, 1993.

[2] D. Yarowsky, "One Sense Per Collocation," In *Proceedings of the ARPA Human* Language Technology Workshop, 1993, pp. 266-271.

[3] 柯淑津、黃居仁、洪嘉馡、劉詩音、簡卉伶、蘇依莉，「中文詞義全文標記語料庫之設計與雛形製作」，In ROCLING 2007.

[4] E. Breidt, "Extraction of V-N-Collocations from Text Corpora: A feasibility Study for German," In *Proceedings of the First Workshop on Very Large Corpora*, 1993, pp.

74-83.

[5] Q. Lu, Y. Li and R. F. Xu, “Improving Xtract for Chinese Collocation Extraction,” In *IEEE 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 2003, pp. 333-338.

[6] 車方翔、劉挺、秦兵、李生，「面向依存文法分析的搭配抽取方法研究」，哈爾濱工業大學信息檢索研究室論文集， 第一卷，2003。

[7] 全昌勤、何婷婷、姬東鴻、劉輝，「從搭配知識獲取最優種子的詞義消歧方法」，中文信息學報，第十九卷，第一期，2005，第 30-37 頁。

[8] W. Li, Q. Lu and W. Li, “Integrating Collocation Features in Chinese Word Sense Disambiguation,” In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 2005, pp. 87-94.

[9] H.-H. Chen, Y.-C. Yu, and C.-L. Li, “Collocation Extraction Using Web Statistics,” In *Proceedings of 4th International Conference on Language, Resources and Evaluation*, 2004, pp. 1851-1854.

[10] C.-Y. Teng and H.-H. Chen, "Analyzing Temporal Collocations in Weblogs." In *Proceedings of International Conference on Weblogs and Social Media*, 2007, pp. 303-304.

[11] C.-C. Wu and J. S. Chang, “Bilingual collocation extraction based on syntactic and statistical analyses,” *Computational Linguistics and Chinese Language Processing*, Vol. 9, No. 1 , 2004, pp. 1-20.

[12] J.-Y. Jian, Y.-C. Chang and J. S. Chang, “TANGO: bilingual collocational concordancer,” In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 2004, pp.166-169.

[13] J.-Y. Jian, Y.-C. Chang and J. S. Chang, “Collocational Translation Memory Extraction Based on Statistical and Linguistic Information,” In ROCLING 2004.

[14] C.D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. The MIT Press, 1999.

附錄

依據 SSTC 詞義標示語料裡的訓練資料，擷取搭配詞之完整結果

目標詞	目標詞 詞義	搭配詞	採用 的方法
關係 Na	普通名詞。人和人之間在社會或群體中的關聯性。	人際 關係	雙連詞
		VC 人際 關係	三連詞
		P Nh V_2 × 關係、P Nh 有 × 關係、 與 Nc DE 關係、與 Nc 的 關係	四連詞
	普通名詞。事件之間的關聯性。	密切 × 關係	雙連詞
		VJ 什麼 關係、V_2 Dfa × × 關係、V_2 Nep 關係、 V_2 什麼 關係、V_2 很 × × 關係、也 沒有 關係、 大 DE 關係、大 的 關係、有 Dfa × × 關係、 有 Nep 關係、有 什麼 關係、有 很 × × 關係、 很 大 × 關係、密切 DE 關係、密切 的 關係	三連詞
		Caa Na DE 關係、Caa Na 的 關係、Dfa 大 DE 關係、 Dfa 大 的 關係、Ng VH DE 關係、Ng VH 的 關係、 V_2 Dfa VH × 關係、V_2 Dfa 大 × 關係、 V_2 很 VH × 關係、V_2 很大 × 關係、 之間 VH DE 關係、之間 VH 的 關係、 有 Dfa VH × 關係、有 Dfa 大 × 關係、 有 很 VH × 關係、有 很 大 × 關係、 和 Na DE 關係、和 Na 的 關係、很 VH DE 關係、 很 VH 的 關係、很 大 DE 關係、很 大 的 關係	四連詞
	普通名詞。事件發生的原因。	因 Na 關係	三連詞
		Cbb Na DE 關係、Cbb Na 的 關係	四連詞
看 VC	監視管理。	看 DE 很 VH、看 得 很 VH	四連詞
	用眼睛察覺。	看 Di Nd、Ncd 一 看、VA 一 看、看 了 Nb、 看 了 Nd、再 D 看、看 著 Nb、看 著 Nc、 看 著 Nep、看 著 Nh、看 著 我	三連詞
		看 Di Na Ncd、Nh VA 一 看、看 著 Na Ncd	四連詞
	透過視覺來理解或欣賞。	看 書、看 電視	雙連詞
		看 Di D、VC 給 × 看、看 了 D、看 不 VC、 坐 P × × 看、坐 在 × × 看	三連詞
		DE Na 去 看、看 Di Nb DE、看 Di Nb 的、 看 Di Nh DE、看 Di Nh 的、VC 給 Nh 看、 的 Na 去 看	四連詞
	以特定態度對待。	看 Di Nh Neu	四連詞
	拜訪、探望後述對象。	Nh 來 看、到 Nc D 看、到 Nc 去 看	三連詞
	病人接受診治。	看 Di Neqa	三連詞

好 VH	形容對特定對象的正面評價。	更 好	雙連詞
		DE 不 好、好 DE 方法、DE 很 好、DE 最 好、 D 太 好、D 比較 好、D 更 好、Dfa D 好、 Dfa 不 好、Na 真 好、Na 最 好、Nep Dfa 好、 Nep 不 好、Nf Dfa 好、Nf 很 好、SHI 最 好、 VC 得 × 好、VJ Dfa 好、V_2 更 好、不 Dfa 好、 不 太 好、什麼 D 好、什麼 不 好、心情 D 好、 心情 不 好、有 Dfa 好、有 多 好、有 更 好、 好 的 方法、的 很 好、的 最 好、是 最 好、 個 Dfa 好、個 很 好、做 DE × 好、做 得 × 好、 得 D 好、得 不 好、得 很 好、都 Dfa 好、就 不 好、 會 Dfa 好、會 比較 好	三連詞
		Cbb Na Dfa 好、DE Na 不 好、好 DE Nv 方式、 D D 太 好、D D 很 好、D Dfa D 好、D Dfa 不 好、 D SHI 最 好、D VC 得 × 好、D VJ Dfa 好、 D V_2 更 好、D 不 Dfa 好、D 不 太 好、D 有 Dfa 好、 D 有 多 好、D 有 更 好、D 很 D 好、D 很 不 好、 D 是 最 好、D 做 DE × 好、D 做 得 × 好、 Na DE Dfa 好、Na DE 最 好、Na Dfa D 好、Na Dfa 不 好、 Na SHI Dfa 好、Na SHI 很 好、Na SHI 最 好、 Na VC 得 × 好、Na 不 SHI × 好、Na_不_是 × 好、Na 的 Dfa 好、Na 的 最 好、Na 是 Dfa 好、 Na 是 很 好、Na 是 最 好、Nf Na Dfa 好、 Nh VC 得 × 好、Nh VK Dfa 好、Nh 覺得 Dfa 好、VC DE D 好、VC DE 不 好、VC DE 很 好、 VC 的 很 好、VC 得 D 好、VC 得 Dfa 好、 VC 得 不 好、VC 得 很 好、VE DE D 好、 VJ Nep D 好、VJ Nep 不 好、VJ 什麼 D 好、 VJ 什麼 不 好、的 Na 不 好、好 的 Nv 方式、 個 Na Dfa 好、做 DE D 好、做 DE Dfa 好、 做 DE 很 好、做 得 D 好、做 得 Dfa 好、 做 得 很 好、該 有 Dfa 好、該 有 多 好、 說 DE Dfa 好、Na D Dfa 好、Na D 很 好、 P Nh Dfa 好、P Nh 不 好、P Nh 很 好、P 我 D 好、 P 我 Dfa 好、對 Nh D 好、對 Nh Dfa 好、對 Nh 很 好	四連詞
	表示同意或允許。	DE Na 好 × 好、DE VH 不 好、DE 好 不 好、 D VH D 好、Nh VH D 好、VA VH D 好、 VA VH 不 好、VA_好_D、VA 好 不 好、 VE VH D 好、VE VH 不 好、VE 好 D 好、 VE 好 不 好、的 Na 好 × 好	四連詞
	形容態度親切的。	好 DE 朋友、好 的 朋友、對 我 × 好	三連詞
	表示結束前一個話題,開始新的話題。	VE 得 Dfa 好	四連詞

	問候語,常用於對話的開場。	會 D 好	三連詞
講 VE	以口語媒介引述或陳述訊息。	講 DE D、講 DE Nep、講 DE 很、講 DE 都、 講 DE 話、D 我 × × 講、D 這樣 講、D 跟 × 講、Na 跟 × 講、講 × Nf 話、Nh P × 講、Nh 剛剛 講、Nh 跟 × 講、Nh 聽 × 講、P 她 講、P 你 講、 講 T T、講 了 一、她 D 講、你 剛剛 講、我 P × 講、我跟 × 講、講 的 D、講 的 Dfa、講 的 Nep、 講 的 都、講 的 話、講 得 很、就 P × 講、 就 跟 × 講、跟 Na 講、跟 你 講、跟 我 講、 講 說 Nh、聽 Nh 講	三連詞
		講 DE D VH、講 DE Nep Nf、講 DE 很 VH、 D D 跟 × 講、D P Nh 講、D P 他 講、D P 我 講、 D 跟 Na 講、D 跟 Nh 講、D 跟 他 講、 D 跟 我 講、講 Di Neu 個、講 Di 一 個、 Na 跟 Nh 講、Nh D P × 講、Nh D 跟 × 講、 Nh Na D 講、Nh P Nh 講、Nh P 你 講、Nh P 我 講、Nh VE Nh 講、Nh 就 P × 講、Nh 就 跟 × 講、 Nh 跟 Nh 講、Nh 跟 你 講、Nh 跟 我 講、 講 一 Nf 話、講 了 Neu 個、講 了 一 Nf、 我 D P × 講、我 D 跟 × 講、我 P Nh 講、 我 P 你 講、我 跟 Nh 講、我 跟 你 講、 講 的 Dfa VH、講 的 Nep Nf、講 得 很 VH、 就 P Nh 講、就 跟 Nh 講	四連詞
	以文字媒介引述或陳述訊息。	Na 上 D 講	四連詞
	描述後述內容。	講 DE SHI、講 DE 是、講 的 SHI、講 的 是	三連詞
		講 DE SHI Na、講 DE 是 Na、講 的 SHI Na、 講 的 是 Na	四連詞
	評價後述對象。	Na 來 講、Nc 來 講、Nh 來 講、我_來 講、 對 Nh × 講、對 我 × 講	三連詞
		P Nh D 講、P Nh 來 講、P 我 D 講、P 我 來 講、 對 Nh D 講、對 Nh 來 講、對 我 D 講、 對 我 來 講	四連詞
說 VE	以口語媒介引述或陳述訊息。	來 說	雙連詞

		說 DE 話、DE 對 × 說、說 D 出、D 這麼 說、 D 跟 × 說、Na 常 說、Nb 就 說、Nh 跟 × 說、Nh 講 說、SHI 覺得 說、VA 著 說、VH 地 說、V_2 人 說、V_2 話 × 說、說 不 出、他 就 說、 有 人 說、有 話 × 說、告訴 我 說、我 跟 × 說、 說 的 話、的 對 × 說、是 覺得 說、笑 Di 說、 笑 著 說、高興 DE 說、問 Nh 說、問 他 說、 媽媽 D 說、話 D 說、跟 Nh 說、跟 他 說、 跟 你 說、跟 我 說、對 Nh 說、對 他 說、聽 Na 說	三連詞
		DE P Nh 說、DE P 他 說、說 DE VH Dfb、 DE 對 Nh 說、D D 這麼 說、D D 跟 × 說、 D P 你 說、D VA 著 說、說 D VE 話、 D V_2 人 說、說 D 出 Na、說 D 出 話、 D 有 人 說、D 笑 Di 說、D 跟 Na 說、 D 跟 Nh 說、D 跟 你 說、D 對 Nh 說、 說 Di Neu 句、說 Di 一 句、Na D 跟 × 說、 Na P 我 說、Na 跟 Nh 說、Na 對 Nh 說、 Na 對 我 說、Nb VH DE 說、Nb VH 的 說、 Nh D 這麼 說、Nh D 對 × 說、Nh P 你 說、 Nh SHI VE 說、Nh SHI VK 說、Nh SHI 覺得 說、 Nh VA 著 說、Nh 告訴 Nh 說、Nh 是 VE 說、 Nh 是 VK 說、Nh 是 覺得 說、Nh 笑 Di 說、 Nh 笑 著 說、Nh 跟 Nh 說、Nh 跟 你 說、 Nh 跟 我 說、P Nh D 說、P Nh 講 說、P 他 講 說、 P 你 D 說、P 我 講 說、說 SHI 這 個、 說 × VE Na T、說 × VE Na 來、說 × VE 話 T、 說 × VE 話 來、VH DE 對 × 說、 VH 的 對 × 說、V_2 話 D 說、也 V_2 人 說、 也 有 人 說、說 不 VE 話、說 不 出 Na、說 不 出 話、 他 VA Di 說、他 VA 著 說、他 VE Nh 說、 他 VE 我 說、他 笑 Di 說、他 笑 著 說、 說 × 出 Na T、說 × 出 Na 來、說 × 出 話 T、 說 × 出 話 來、有 話 D 說、我 D 對 × 說、 我 P Nh 說 、我 P 你 說、我 SHI VE 說、 我 SHI VK 說、我 SHI 覺得、我 是 VE 說、 我 是 VK 說、我 是 覺得 說、我 跟 Nh 說、 我 跟 你 說、的 P Nh 說、的 對 Nh 說、 很 VK DE 說、很 VK 的 說、說 是 這 個、 說 得 VH Dfb、就 VE Nh 說、話 D D 說、 跟 Nh Na 說、跟 Nh VE 說、跟 Nh 講 說、 跟 他 VE 說、跟 他 講 說、跟 我 VE 說、 跟 我 講 說	四連詞
	以口語進行打招呼、道謝、拒絕等言談行爲。	Caa SHI 說、Caa 是 說、或者 SHI 說、或者 是 說	三連詞
		Nh D 覺得 說、Nh 就 VK 說、Nh 就 覺得 說、 我 D 覺得 說	四連詞