利用雙語學術名詞庫抽取中英字詞互譯及詞義解歧

白明弘[1,2]、陳克健[1]、張俊盛[2]

1 中央研究院資訊科學研究所

2 國立清華大學資訊工程研究所

mhbai@sinica.edu.tw, kchen@iis.sinica.edu.tw, jschang@cs.nthu.edu.tw

摘要

語意的研究十分依賴語意知識庫所提供的訊息，由於語意研究逐漸變得熱門，相對的語意知識庫的建構也變得十分迫切。WordNet 是目前最廣爲人知的英語語意知識庫，許多語意解歧(word sense disambiguation)的研究都以 WordNet 爲共同標準。由於 WordNet 的成功，使得許多其他語系的 WordNet 建構計畫也紛紛出現。本文提出一個自動從雙語學術名詞庫中抽取中文語意訊息的方法，這個方法利用一個詞和詞的對應(word-to-word alignment)演算法抽取中英詞對譯的訊息，再用語意解歧的方法，將中文詞連結到 WordNet synset，以建構中文 WordNet。

1. 緒論

近年來在自然語言處理領域中，語意研究受到了廣泛的重視。語意解歧的技術不斷推陳出新，進而使得語意的應用也受到鼓舞。然而，語意的使用必需仰賴語意知識庫提供語意訊息，這些訊息包括一個詞彙有多少不同的語意，以及一個語意和另一個語意是否有同義關係或是上下位關係等。例如：「分子」可以表示化學上的「粒子」(如「水分子」)，也可以表示「一群人」(如「激進分子」)；而「拉布拉多」和「大麥町」由其上位詞可知都是一種「狗」。

WordNet 是一部訊息豐富的語意知識庫[Miller 1990]，其中收錄了爲數極多的詞彙。在結構上它將所有的相同的語意集成 synset，並以 synset 爲基礎進一步連結語意之間的關係，如上位關係(hypernym)、下位關係(hyponym)、整體關係(holonyms)及部分關係(meronyms)等。目前 WordNet 已經被應用在許多的研究上，如語意解歧(word sense disambiguation)、資訊檢索(information retrieval) 及電腦輔助語言學習(computer-assisted language learning)等領域，儼然成爲語意研究的共同標準。

由於 WordNet 的成功使得許多其他語系的 WordNet 建構計畫相繼出現。例如：EuroWordNet (EWN)，該計畫目標爲建構包含多種歐洲語的 WordNet，及中文詞網計畫[CKIP 2003]，以建構中文語意知識庫爲目標。從零開始建構一個 WordNet 是一項艱鉅的任務，所以有許多的研究嘗試以自動的方式將詞彙連結到 WordNet。例如：[Atserias et al. 1997]、[Daude et al. 1999]以及[Jason et al. 2003]都是利用雙語詞典所提供的翻譯，自動將詞彙連結到 WordNet。使用一般雙語詞典的翻譯最大的問題在於用詞過度典型化。例如："plant" 在 WordNet 中的第一個語意 "plant, works, industrial plant"，在雙語詞典中翻譯成「工廠」。但實際上在文章中可能翻譯成「廠」、「工廠」、「廠房」、「所」(如「power plant/發電所」)及「工場」等詞。用詞過度典型化的現象，使得許多文章中的用詞無法找到適當的翻譯連結到 WordNet。

在本實驗中，我們選擇以雙語學術名詞庫作爲抽取語意訊息的資料來源。由於學術名詞庫中包含了大量的複合詞，所以很多詞會搭配不同的詞一再出現，並對應到不同的翻譯。因此不但可以避免一般雙語詞典翻譯過度典型化的問題，而且多樣化的翻譯結果可以幫助語意解歧 [Diab et al, 2002][Bhattacharya, 2004]。在本實驗中我們將問題分成兩個部分：a) 如何找出中文詞和英文詞對應的翻譯，b) 如何解決英文的歧義。

本文接下來的章節組織如下。在第 2 節中說明所使用的資源。第 3 節中說明實驗的方法。第 4 節中說明實驗的結果。結論及未來的發展則在第 5 節中說明。

2. 使用資源

本研究使用了兩本詞典作爲語意抽取的對象：

a) 國立編譯館學術名詞詞庫 [NICT, 2004]。

b) 英漢詞典

其中國立編譯館所編輯的「學術名詞」詞庫的內容包含 63 個學科類別共 1,046,058 目詞。這些詞條中有 629,352 目詞是複合詞，佔總詞數的 60 %。英漢詞典共有 208,163 目詞，用來補足「學術名詞」之不足。此外我們使用 WordNet 2.0 做爲語意連結的對象。

由於中文的複合詞在詞和詞之間沒有空白分隔，不像英文詞間以空白字元做爲邊界，所以必需依賴自動斷詞程式將複合詞切分成一般詞。本實驗採用中央研究院詞庫小組所研發的自動斷詞

系統來切分複合詞。

3. 方法

我們將實驗分成兩個步驟：

1. 中英對應
2. 語意標記

第一個步驟目的是要找出中文詞和英文詞的對應翻譯。實驗的資料本身包含複合詞及單字詞，所以首先必需找出在英文複合詞及中文複合詞裡的組成成份對應的翻譯。例如：”water tank” 的翻譯爲 “水槽”，我們希望能對應成 “water”→”水”，”tank”→”槽”。第二個步驟的目的則是將詞連結到 WordNet 的 synset。例如：tank 一詞在 WordNet 中一共有五個語意：

tank-1 -- an enclosed armored military vehicle
tank-2 -- a large vessel for holding gases or liquids
tank-3 -- as much as a tank will hold
tank-4 -- a freight car that transports liquids or gases in bulk
tank-5 -- a cell for violent prisoners

要決定 “tank”→”槽” 連結到哪一個語意，必需要透過語意解歧，才能知道 tank 究竟是屬於哪一個語意。我們將在 3.1 節中說明中英對應的演算法，在 3.2 節中說明語意標記的演算法。

3.1 中英對應

所謂中英對應目的是要找出中文複合詞和英文複合詞的組成成份的對應翻譯。例如：”water tank” 及 “水 槽” 的對應爲 “water/水 tank/槽”，”supplementary education” 及 “補習 教育” 的對應爲 “supplementary/補習 education/教育” 等。 關於雙語對應的研究，有許多現成的文獻可參考，如[Brown et al., 1993][Och and Ney, 2000]等。本實驗中所要對應的是比較短的複合詞而非句子，因此我們只需要考慮詞對詞的翻譯機率，而不必考慮詞的先後順序的影響。這個策略分成兩個部分 1) 計算英文詞和中文詞翻譯機率，2) 搜尋詞和詞最佳對應的路徑。這兩個步驟分別在 3.1.1 節及 3.1.2 節中說明。在實驗前，中文的複合詞已經先用中央研究院詞庫小組的斷詞系統斷好詞。

3.1.1 計算詞和詞的對應機率

我們使用 EM 演算法[Dempster et al., 1977]計算中文詞和英文詞對應的機率。假設有一個平行的詞庫 S，是由許多不同的英文詞串e_i和對應中文詞串c_i所構成，即$S=\{(e_1,c_1),(e_2,c_2),\ldots,(e_n,c_n)\}$。這些詞串可以視爲一般對應演算法中的句子。要計算詞串中每一個英文詞 w_e 翻譯成中文詞w_c 的機率 $P(w_c|w_e)$ 的計算方法如下：

Initialization:

$$P_{(e_i,c_i)}(w_c \mid w_e) = \frac{1}{m}, m = |\{w \mid w \in c_i\}|$$

E-step:

$$Z(w_c, w_e) = \sum_{(e_i,c_i)\in S} P_{(e_i,c_i)}(w_c \mid w_e)$$

M-step:

$$P(w_c \mid w_e) = \frac{Z(w_c, w_e)}{\sum_{v\in \text{chinese words}} Z(v, w_e)}$$

$$P_{(e_i,c_i)}(w_c \mid w_e) = \frac{P(w_c \mid w_e)}{\sum_{v\in c_i} P(v \mid w_e)}$$

在上面的式子中 $P_{(e_i,c_i)}(w_c \mid w_e)$ 表示w_e和w_c在詞串 (e_i,c_i) 中對應的機率。在初始值的設定上，假設對e_i中的一個英文詞 w_e 而言，對應到c_i中的每個中文詞形的機率都是 1/m，其中m表示詞串c_i中的詞數。所以 $P_{(e_i,c_i)}(w_c \mid w_e)$ 的初始值設爲 1/m 。 E-step 的目的是計算出在S中，所有包含w_e的英文詞串集$\{e_i, e_j,\ldots,e_k\}$， w_c 出現在相對應的中文詞串$\{c_i,c_j,\ldots,c_k\}$次數的期望值。M-step 則是從期望值重新估算翻譯的機率，其中分母加總的部份是針對所有中文詞爲範圍。重覆EM-step 直到收斂停止。表 1 爲英文詞 “tank” 翻譯成中文詞的機率。

英文詞	中文詞	共現次數	機率
tank	槽	492	0.354159
tank	櫃	290	0.200845
tank	艙	157	0.101734
tank	箱	59	0.040555
tank	水槽	36	0.023986
tank	池	33	0.020965
tank	罐	28	0.018258
tank	油槽	23	0.014928

表 1. 英文詞“tank”翻譯成中文詞的機率

3.1.2 搜尋詞和詞對應的最佳路徑

在上一節中利用 EM 演算法得到了每個英文詞對應到中文詞的機率值，在本節中，將利用此機率值來找出中英詞串裡中文詞和英文詞最佳的對應。我們採用路徑搜尋的方式，來找最佳的中英文詞對應。此方法說明如下：

1. 從中文詞對應到英文詞：其目的爲將中文詞組合起來，以 ”cedar nut oil” 和 “雪 松 堅果 油” 的對應爲例，由於 “雪松” 是未知詞，在對應之前是分開來的。從中文詞對應到英文詞時 “雪” 和 “松” 會同時對應到 “cedar”，所以兩個中文詞會合成 “雪松” 對應到 “cedar” 如圖 1：

	雪	松	堅果	油
cedar	**0.064020**	**0.019535**	0.000047	0.008866
nut	$9.6x10^{-6}$	0.000096	**0.035525**	0.010841
oil	$6.9x10^{-11}$	0.001410	$8.3x10^{-10}$	**0.609328**

圖 1. 從 “雪 松 堅果 油” 對應到 “cedar nut oil” 的路徑，對應的結果爲 “cedar/雪松 nut/堅果 oil/油”。

2. 從英文詞對應到中文詞：將英文詞組合起來，以 “law of universal gravitation” 和 “萬有引力定律” 的對應爲例，從英文詞對應到中文詞時 “universal” “gravitation” 兩個詞會同時對應到 “萬有引力” ，所以對應的結果會將兩個英文詞合併成複合詞 ”universal gravitation” 再對應到 “萬有引力”，如圖 2：

	law	of	universal	gravitation
萬有引力	$1.2x10^{-6}$	--	**0.033920**	**0.372825**
定律	**0.603909**	--	0.001237	0.008153

圖 2. 從 “law of universal gravitation” 對應到 “萬有引力 定律” 的路徑對應的結果爲 “law/定律 universal_gravitation/萬有引力”。

在上列的 1,2 步驟中，每一個對應只需要挑機率值最高的對應即可。例如 Pr(“law”|”萬有引力”)=$1.2x10^{-6}$ < Pr(“law”|”定律”)=0.603909，所以可以決定 “law” 對應到 “定律”。大部份的情況都只需要挑最高機率值就能找到最佳對應，但是有一些例外的情況：

1. 交錯對應：例如，”external examination” 和 “校 外 考試” 的對應情況。在此例中，”校” 和 “external” 的機率低於 “examination”，所以 “校” 和 “考試” 同時對應到 “examination”，只有 “外” 對應到 “external”，這種交錯對應的情況不甚合理。如圖 3：

	校	外	考試
external	$1.4x10^{-7}$	**0.575537**	$5.3x10^{-9}$
examination	**$5.2x10^{-6}$**	5.2x10-6	**0.172751**

圖 3. 從 “校 外 考試” 對應到 “external examination” 的路徑，對應的結果爲 “external/外 examination/校 考試”，”校” 和 “考試” 雖間隔一個字卻對應到同一個英文 “examination”。

2. 功能詞爲複合詞的一部份：在對應時，功能詞原則上不對應，但是當功能詞爲複合詞的一部份時，功能詞不能夠捨棄，例如，”general theory of relativity” 和 “廣義 相對論” 的對應，”theory” 和 “relativity” 對應到 “相對論”，但完整的複合詞應該是 “theory of relativity”，此時，功能詞不能捨棄。

上述的例外情況 1 使得對應不能總是選機率最高的情況，而必需透過路徑搜尋的方式找到最佳的路徑。在路徑搜尋的演算法上，我們採取路徑成本的計分方式。假設有一個中文詞串 “$c_1\ c_2\ c_3 \ldots c_k$” 要對應到英文詞串 “$e_1\ e_2\ e_3 \ldots e_n$”，其中某條對應路徑可以表示爲 $path_i = (c_1, e_{i1}), (c_2, e_{i2}),\ldots, (c_k, e_{ik}), e_{ij} \in \{ e_1, e_2, e_3, \ldots, e_n \}$，欲計算 $path_i$ 路徑的成本，其成本函數(cost function)定義如下：

$$\cos t(path_i) = \begin{cases} \infty, & if\ cross\_alignment(path_i) = true \\ \sum_{j=1}^{k} -\log(p(c_j \mid e_{ij})), & else \end{cases}$$

在成本函數中必需偵測路徑中是否有交錯對應(cross alignment)的存在，若存在則該路徑應被捨棄，將其成本設爲 ∞ 。而交錯對應的的偵測方式如下：

$$cross\_alignment(path_i) = \begin{cases} true, & \exists(c_p, e_{ip}), (c_q, e_{iq}) \in path_i, e_{ip} = e_{iq} \text{ and } p - q > 1 \\ false, & else \end{cases}$$

其義意爲若在對應路徑 $path_i$中存在不相鄰的兩個詞 c_p, c_q， 對應到同一個英文詞，意即 $e_{ip}=e_{iq}$，則該路徑有交錯對應。選擇路徑時，只要選擇成本最低的路徑即可，選取規則如下：

$$best\_path = \arg\min_{path_i} \cos t(path_i)$$

而例外情況 2 比較單純，只要檢查功能詞所連接的詞是否對應到同一個詞，如果對應到同一個詞就保留，否則就將功能詞捨棄。表 2 爲詞和詞對應的一些實例。

英文複合詞	**中文複合詞**	**詞和詞對應**
evaporation tank	蒸發 槽	evaporation/蒸發 tank/槽
wind-wave tank	風浪 水槽	wind-wave/風浪 tank/水槽
wave tank	波浪 水槽	wave/波浪 tank/水槽
volumetric tank	量 水箱	volumetric/量 tank/水箱
curve of learning	學習 曲線	curve/曲線 of/ learning/學習
exchange of students	學生 交換	exchange/交換 of/ students/學生
practice teaching	教學 實習	practice/實習 teaching/教學
wall cloud	雲 牆	wall/牆 cloud/雲
gas mixture	混合 氣體	gas/氣體 mixture/混合
air choke valve	阻 氣 閥	air/氣 choke/阻 valve/閥

表 2. 詞和詞對應的實例

3.2 語意標記

在做語意標記時，如果英文詞本身沒有歧義，則可以直接將其 WordNet Synset 標記在詞庫的詞上。如果有歧義，則必需做語意解歧。在本實驗中我們參考[Atserias et al, 1997]的解歧方法，總共使用三種解歧方法：

1. 解歧法一：利用英文複合詞和其組成成份之關係

   假設英文複合詞 e為 n個詞$w_{e1}, w_{e2},\ldots,w_{en}$ 所組成，若其中有一個詞 w_{ei} 的某個語意s_j 為e的某個語意 s_k上位詞，則 w_{ei}的語意為s_j，且e的語意為s_k。

   例如， “water tank-1” 為 “tank-2” 的下位，則 “water tank” 之組成成份 “tank” 之語意可標為 “tank-2” 如圖 4：

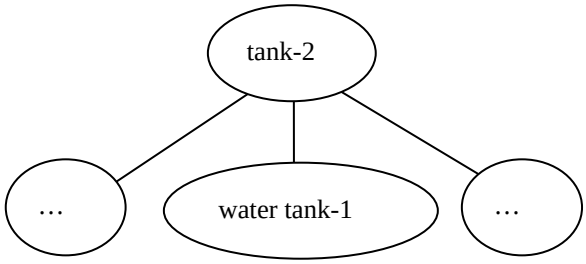

圖 4. “water tank” 和其組成成份 “tank” 的語意具有上下位關係

2. 解歧法二：英文詞之間的語意交集

   假設中文詞 w_c 可被翻譯成n個不同的英文詞 $w_{e1}, w_{e2},\ldots, w_{en}$，解歧的規則如下：

   a) 若$w_{e1}, w_{e2},\ldots, w_{en}$有一個共同的 synset s，則w_c被連結到 s。

   b) 若 t 為w_{ei} 的一個 synset ，且其餘的英文詞$w_{e1}, w_{e2},\ldots, w_{en}$都有一個 synset 落在 t 的下位，則 w_c 分別連結到 t 及這些下位。

   例如，”信號旗” 可翻譯為 “signal”, “signal flag” 及 “code flag”，而 “signal” 其中的一個 synset 為 “signal flag” 及 “code flag” 的上位，則 ”信號旗” 所標的語意可為 “signal-1”, “signal flag-1” 及 “code flag-1”。如圖 5：

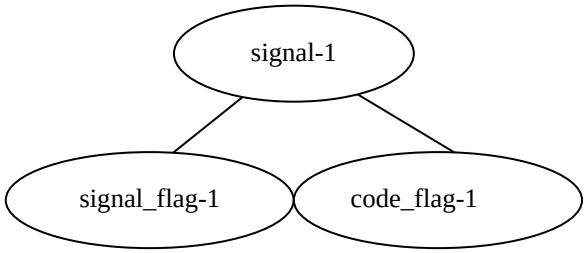

圖 5. “信號旗” 的英文翻譯 “signal”, “signal flag” 及 “code flag” 具有上下位關係

3. 解歧法三：標記沒有歧義的中文詞

利用前面實驗標記所得的結果，找出沒有歧義的中文詞，再利用沒有歧義的中文詞去標記。例如:”防波堤” 在前面實驗所得的結果，都對應到同一個 synset {breakwater-1, groin-2, groyne-1, mole-5, bulwark-3, seawall-1, jetty-1}，所以只要 “防波堤” 對應到的英文詞是該 synset 中的任一詞，就可以判別屬於該 synset。

4. 實驗結果

語意抽取的對象爲「學術名詞詞庫」和「英漢詞典」，總共有 1,254,221 個詞，其中所包含的複合詞有 645,200 佔總詞數的 51.44%。實驗的結果一共將 124,752 個中文詞連結到 42,589 個 WordNet synset，結果一共產生 165,775 個(中文詞, synset) 的連結組合。此結果平均一個中文詞落在 1.33 (由 165,775 / 124,752 得) 個 WordNet synset 中。在 4.1 節中將說明中英對應的實驗結果，4.2 節中說明解歧的結果。

4.1 中英對應的結果

由於只有複合詞需要對應，所以我們只針對複合詞做評估。評估的方法爲在對應好的詞庫中，隨機抽取 500 個詞條驗證，驗證的方式爲人工檢視。結果如表 3 所示，對應的正確率爲 95.19 %。

詞條抽樣數	對應正確數	正確率
500	476	95.19%

表 3. 中英詞與詞對應的正確率。

分析這些錯誤的對應約可規類成四種錯誤的類型，如表 4 所示。

錯語類型	錯誤的例子
中文詞切分點的錯誤	half-wave/半 length/波長 criterion/準則 spiral/螺旋 coal/煤機 cleaner/洗 american/西 ginseng/洋參 second/再 wind/生氣 microlen/微透鏡藕 coupler/合器 atomic/原子能 energy/階
音譯詞音節對應錯誤	san/聖胡 julian/連安
中英文翻譯不對稱	navigation/航行參考 star/星
英文爲縮寫	double/ III/托克馬克熱核反應器

表 4. 詞與詞對應錯誤的四種類型。

對應的結果一共得到 840,187 個中/英對譯項，其中包含了 445,830 個中文詞形和 318,048 個英文詞形。平均一個中文詞有 1.88 個英文翻譯，而一個英文詞有 2.64 個中文翻譯。

4.2 語意標記的結果

語意標記的結果評估分成兩部份探討，首先是針對語意解歧的正確性評估，其次是對整個實驗的覆蓋率評估。

4.2.1 語意解歧的正確性

在語意解歧的評估上，我們只針對有歧義的詞做評估。在抽樣的方法上採取隨機抽樣的方式，在兩個解歧實驗結果中隨機各取 200 個標記結果評估，評估的方式爲人工檢視。正確率的分析結果如表 5：

	取樣數	標記正確數	正確率
解歧法一	200	160	80.00 %
解歧法二	200	167	83.50 %
解歧法三	200	174	87.00 %

表 5. 語意解歧的正確率

4.2.2 標記的覆蓋率

在覆蓋率的分析上，我們分別針對 WordNet 2.0 語意對(word-sense pair)的覆蓋率及 synset 的覆蓋率評估。在 WordNet 2.0 中，總共有個 203,145 語意對，及 115,424 個 synset。分析的結果如表 6：

	tokens 個數	word-sense pair 個數	word-sense pair 覆蓋率	synset 個數	synset 覆蓋率
monosemous word	370991	48623	23.94 %	39953	34.61 %
解歧法一	29422	4211	2.07 %	3452	2.99 %
解歧法二	29311	2050	1.00 %	1685	1.46 %
解歧法三	81734	1931	0.95 %	1543	1.34 %
總共 (聯集)	484771	54654	26.9 %	42589	36.89 %

表 6. 抽出的語意對 WordNet 的覆蓋率

5. 結論及未來發展

在本文中我們提出一套利用「雙語學術名詞庫」來抽取中文語意的方法，本方法將問題分成兩個部份：一、先將複合詞作詞和詞對應，以得出中文和英文的翻譯，二、利用解歧的方法，將 WordNet 的語意標記在詞上。實驗的結果顯示，抽出的語意可以覆蓋 26.9 % 的 WordNet 語意對 (word-sense pair)，這些語意對函蓋了 36.89 % 的 synset。而三個解歧法的正確率分別可達 80 %, 83 % 及 87 % 的正確率。

使用學術名詞詞庫有許多好處，首先在詞庫中包含了大量的複合詞，同一個詞會搭配不同的詞一再地出現，並對應到不同的翻譯。這種翻譯多樣化的好處可以改善只使用一般「英漢詞典」翻譯用語過度典型化的缺點，同時多樣化的翻譯也有助於語意解歧。其次，由於複合詞的長度大部份只包含 2~3 詞，所以在詞和詞的對應上比句對句的對應單純且正確率也比較高。

在本實驗中可以發現，從大量的雙語學術名詞庫中的確可以抽取豐富的語意訊息。目前所使用的解歧方法只解開了一部分的歧義，未來可以嘗試不同的解歧方法，以抽取更多語意訊息。

參考文獻

A.P Dempster, N.M. Laird, and D.B. Rubin., “Maximum likelihood from incomplete data via the EM

algorithm," *Journal of the Royal Statistical Society*, B29:1-38, 1977.

CKIP, "The sense and semantic of Chinese Word," *Technical Report No. 03-01, 03-02*, 2003.

Franz Josef Och and Hermann Ney, "Improved Statistical Alignment Models," *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* , 2000.

Indrajit Bhattacharya, Lise Getoor, Yoshua Bengio, "Unsupervised Sense Disambiguation Using Bilingual Probabilistic Models, " *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004

Jason S. Chang, Tracy Lin, Geeng-Neng You, Thomas C. Chuang, Ching-Ting Hsieh, "Building A Chinese WordNet Via Class-Based Translation Model," *International Journal of Computational Linguistics and Chinese Language Processing*, Vol 8, No.2 pp. 61-76, August 2003.

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, Horacio Rodríguez, "Combining Multiple Methods for the Automatic Construction of Multilingual WordNets," *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1997.

Mihalcea, R. and D. Moldovan, "A method for Word Sense Disambiguation of unrestricted text," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999

Miller, G., "WordNet: An online lexical database," *International Journal of Lexicography*, 3(4), 1990.

Mona Diab and Philip Resnik, "An Unsupervised Method for Word Sense Tagging using Parallel Corpora, " *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

NICT, "學術名詞資訊網 (http://www.nict.gov.tw/tc/dic/index1.php), " *NICT*, 2004.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The Mathematics of Machine Translation: Parameter Estimation," *Computational Linguistics*, 19(2):263–311. 1993.

Philip Resnik, "Exploiting Hidden Meanings: Using Bilingual Text for Monolingual Annotation, " *Conference on Intelligent Text Processing and Computational Linguistics*, 2004.