統計式片語翻譯模型

張俊盛　　游大緯

國立清華大學資訊工程研究所

jschang@cs.nthu.edu.tw

摘要

機器翻譯是自然語言處理研究上最重要的課題之一，在過去運用機器翻譯比較成功的例子，多是特定的領域文件的翻譯。近來因為網際網路與搜尋引擎的盛行，大家開始重視機器翻譯在跨語言檢索（Cross Language Information Retrieval）中的角色。在跨語言檢索的問題上，通常是對查詢字詞或片語，進行翻譯（Query Translation）。然而翻譯的結果必須和欲搜尋的文件庫的有高度的相關性，才能達到檢索的效果。目前的查詢關鍵詞翻譯的做法，或者採用現成的翻譯軟體，或者使用一般性的雙語詞典，都無法產生和文件相關的翻譯。因此我們希望能夠透過統計式機器翻譯的做法來進行查詢關鍵詞的翻譯，以提高跨語言檢索的效率。在這篇論文中，我們提出新的統計式片語翻譯模型，並進行實驗，證實能改進原有的統計式機器翻譯模型的缺點，提升片語對應與翻譯的效率。

1. 簡介

機器翻譯是自然語言處理研究上最重要的課題之一，有助於幫助使用者跨越語言與文化的障礙。在過去運用機器翻譯比較成功的例子，多是特定的領域文件的翻譯，如技術性的使用手冊、氣象報告、國際機構的官方文件。近來因為網際網路與搜尋引擎的盛行，大家開始重視機器翻譯

在跨語言檢索（Cross-Language Information Retrieval），可能扮演的角色。

在特定領域的文件翻譯上，機器翻譯系統主要是以句子為單位，進行處理。在跨語言檢索的問題上，可以採取「文件翻譯」（document translation），或者「查詢資訊翻譯」（query translation）的做法（McCarley 1999）。目前大部分的研究者都採取查詢關鍵詞翻譯的做法。例如，在 NTCIR-2 的英到中的資訊檢索評估活動（Kando et al. 2001）中的一個查詢主題中，就提供以下的英文關鍵詞，試驗參與的系統，找到相關中文新聞文件的能力：

- Assembly Parade Law
- Parade and Demonstration
- Constitution
- Freedom of speech
- Indemnification
- Communism
- Country separation
- Council of Grand Justices
- Legislation
- Amendments

查詢關鍵詞的翻譯涉及詞彙語義解析（Word Sense Disambiguation）的問題（Ide and Veronis 1998, Chen and Chang 1998）與片語的翻譯（Phrase Translation）的問題，和一般性翻譯很重要的不同點，在於翻譯的結果，是要拿來在一個文件庫（Text Collection）中搜尋文件。所以翻譯的詞義解析與翻譯的詞彙選擇（Lexical Choice）必須和文件庫的語料有高度的相關性。以上述關鍵詞中的 demonstration 為例，我們就必須翻譯成新聞中常見的「示威」而不能翻譯成「示範」。

目前學者研究跨語言檢索的做法，大致上分為兩種：

1. 利用市場上販售的翻譯軟體（Gey and Chen 1997, Kwok 2001）
2. 使用一般性的雙語詞典（Oard 1999, Kwok 2001）

這兩種做法，很明顯的都不容易產生和文件庫相關的翻譯。這一點對於音譯的專有名詞，特別明顯。Kwok 就指出使用現成翻譯軟體和一般性雙語詞典，不能得到 Michael Jordan 在文件庫的正確音譯「麥可喬丹」，顯然是跨語言檢索研究的一大問題。

為了提高翻譯和文件庫的相關性，Chen 等（1999）將詞彙共現機率（occurrence statistics）導入翻譯詞彙選擇的考慮中。有鑒於音譯專有名詞在跨語言檢索的重要性，也有研究者提出了一些統計或規則式的做法，將音譯轉換成原始專有名詞（Knight and Graehl 1997, Chang et al. 2001）。這些做法，雖然對於跨語言檢索有一定的效果，但缺乏比較全面性而嚴謹的理論架構，也因此影響到改進發展的空間。

我們認為要做好跨語言檢索中的查詢關鍵詞的翻譯，必須有一套全面而嚴密的方法，發展適用的機器翻譯模型。在機器翻譯的做法中，範例為本做法（Example-based Approach）和統計式機器翻譯，都是比較資料導向（data-driven）的做法，比較能夠產生和資訊檢索文件庫相關的翻譯。其中又以 IBM Watson 研究中心的 Brown 等（1988, 1990, 1993）提出的統計式機器翻譯做法，在理論上較為嚴謹，在架構與做法上較為明確可行。

因此我們希望能夠透過一種新的統計式對應與機器翻譯做法（Statistical Alignment and Machine Translation）來進行查詢關鍵詞的翻譯，為跨語言檢索的查詢詞翻譯提供一個比較有效而解嚴謹的做法。在這篇論文中，我們提出一種新的翻譯對應機率（Alignment Probability）

的做法，並進行實驗。實驗的結果證實新的模型的確能改進片語對應與翻譯的效率。

2. 統計式機器翻譯模型

機器翻譯早期是以逐字翻譯加上局部的位置調整的直接做法（Direct Approach），後來逐漸轉成主要是以句法分析為基礎的轉換式的做法（Transfer Approach）。在 1980 年代末，研究的趨勢比較傾向實證式的做法（Empirical Approach），以翻譯的範例或平行語料庫為本，發展機器翻譯系統。Brown 提出的語料庫為本之統計式做法，在理論的架構最為完備。在 Brown 的統計式機器翻器模型下，原文 S 和譯文 T 的翻譯機率（Translation Probability） $Pr(T \mid S)$，可以分解成以下的三個機率函數：

(a) 詞彙翻譯機率（Lexical Translation Probability）

$Pr(S_i \mid T_j)$

(b) 孳生機率（Fertility Probability）

$Pr(a \mid b)$

(c) 位置扭曲機率（Distortion Probability）

$Pr(i \mid j, k, m)$

其中

S_i 為 S 的第 i 個字
T_j 為 T 的第 j 個字
a 為 S_i 的長度
b 為 T_j 的長度
k 為 S 的長度
m 為 T 的長度

Brown 等使用加拿大國會議事錄的英法平行語料庫，證實透過反覆交替的「期望值估計」與「最佳化」演算法（Expectation and Maximization

Algorithm)，可以得到這三個簡單的機率函數的統計估計值。其「最佳化」的步驟，就是在目前的機率函數估計值下，求取最可能的翻譯對應。而「期望值估計」的步驟，就是以所有的雙語語料樣本的最佳的翻譯對應為根據，估計三個機率函數值。

透過 EM 演算法，統計式機器翻譯模型中的翻譯機率函數的估計值可趨於收斂。在雜訊通道模型（Noisy Channel Model）下，結合翻譯機率函數，與目標語的 N-gram 語言模型（Language Model)，可以用搜尋演算法，如束限搜尋法（Beam Search）求最佳機率值的方式，產生翻譯。

3. 適用於片語對應與翻譯的統計式模型

Brown 原始模型中的位置扭曲機率，是基於每一字的翻譯目標位置和其他字無關的假設。在獨立事件的假設下，某一個翻譯對應（alignment）方式的機率，在位置方面而言，是所有字的和對應字的位置形成的位置扭曲機率值的乘積。實際上，每一字的翻譯目標位置和其他字的翻譯位置有高度的相關性。如果 $S_{i'}, i' \neq i$ 都不對應到 T_j，則 S_i 對應到目標位置 j 的機率幾乎為 1

$$Pr(j \mid i, k, m) \cong 1 \quad 若 \quad Pr(j \mid i', k, m) = 0, i' \neq i$$

因此獨立假設下的機率，幾乎大部分的情況下都是過低的估計。即便是很可能的翻譯對應方式，其機率值還是一連串位置扭曲機率的乘積，因此常趨於非常的小的數值。例如，檢視三字英文五中字文的片語樣本，最可能翻譯對應 A^* 下的三個字 $S_1\ S_2\ S_3$ 翻譯目標位置，分別是

$$S_1 \rightarrow \{T_1, T_2\}$$

$$S_2 \rightarrow \{T_3, T_4\}$$

$$S_3 \rightarrow \{T_5\}$$

也就是 $A^* = (0, 12, 34, 5)$（第一個 0 代表所有的中文字都有對應，沒有中文字無法對應到英文字的情況)。在 $k = 3$ 及 $m = 5$ 的片語樣本中，翻譯對應為 A^*的情況約佔 35%。直接估計 A^*的最大可能估計值（Maximum Likelihood Estimation)，得到

$$Pr_{\text{MLE}}(A^*) = 0.35$$

然而在機率獨立的假設下

$$Pr(A^*) = P(1|1,3,5)\ P(2|1,3,5)\ P(3|2,3,5)\ P(4|2,3,5)\ P(5|3,3,5)$$

即使以較高的位置扭曲機率值（0.6）估計 $P(j \mid i,3,5)$，其乘積仍然過低，遠低於合理的估計值：

$$Pr(A^*) < (0.6)^5 = 0.046656 \ll 0.35$$

$$Pr(\text{A}^*) \ll Pr_{\text{MLE}}(A^*)$$

為了更精確合理的估計翻譯目標位置的機率，我們提出了直接估計整體翻譯配對位置與字數的做法。在此做法下，孳生機率和位置扭曲機率合併成為翻譯對應機率（Alignment Probability）。因此不再獨立考慮個別的字的位置、翻譯目標位置、孳生的字數，而是以翻譯對應來一併考慮。在這樣的想法下，我們將原文 S 和譯文 T 的翻譯機率 $Pr(T \mid S)$，分解成以下的兩個機率函數：

(a) 詞彙翻譯機率（Lexical Translation Probability）

$$Pr(T(A_i) \mid S_i)$$

(b) 翻譯對應機率（Alignment Probability）

$$Pr(A \mid k, m) = Pr(A_0, A_1, A_2, \ldots, A_k \mid k, m)$$

其中

S_i 為 S 的第 i 個字
$T(A_i)$為 T 中對應到 S_i 的部分
A_0 為 T 中沒有對應到 **S** 的部分的標號
A_i 為 T 中對應到 S_i 的部分的標號, i >0
k 為 S 的長度
m 為 T 的長度

如果個別字 S_i 的對應字在 **T** 中為連續，則我們可以用對應目標位置

的起點 B(A, i) 與終點 E(A, i)，來簡化對應關係的表達，也就是

$$S_i \rightarrow T(A_i) = \{T_{B(A,i)}, T_{B(A,i)+1}, T_{E(A,i)}\}$$

$$A_i = \{ B(A,i), B(A,i)+1, \ldots, E(A,i) \}$$

4. 實驗

我們進行了一系列的實驗，以驗證我們提出的新的片語翻譯模型的效果與可行性。透過實驗，我們想了解新模型有關的下列幾個問題：

1. 以翻譯對應機率替代孳生機率和位置扭曲機率，是否可以得到較正確的對應分析？
2. 翻譯對應是否集中在幾種樣式，而不是許多個別對應目標位置的排列組合？翻譯對應機率的參數量，會不會過多，導致估計的速度會不會過慢？
3. 翻譯對應機率的參數量和樣本數量，相較之下，其機率值的統計可靠度會不會過低？
4. 訓練後的機器翻譯模型，應用到跨語言檢索的可行性高或低？

4.1 實驗的設計與起始機率值的設定

由於不易取得大量雙語片語的語料，我們採用 BDC 漢英字典（BDC 1992）的片語條目作為實驗的原始材料。為了配合實驗的目標，並簡化問題，我們首先去掉英文多於 3 個詞的條目，但中文長度不限。另外我們也去掉中文的四字成語條目。這些條目的翻譯，常常不是字面翻譯，去掉之後，可以降低資料的雜訊。原始資料經過整理之後，我們得到 96,156 筆可用的英中片語翻譯的記錄。我們以（P_n, Q_n）, n = 1, N 來代表這組語料。

在試驗中，我們以 EM 演算法，來得到第三節所提出的辭彙翻譯機率、翻譯對應機率。我們採取了和一般不同，但類似 Och 等人（2000）

對於 IBM 機率模型的改進實驗的做法。其目的都是希望加速機率的估計。

1. 開始的時候，我們採取 Brown 模型原有的位置扭曲機率。在 EM 演算法的第二輪之後才開始使用新模型的翻譯對應機率。

2. 我們假設英中片語翻譯時，英文和中文字的順序一致的機會較高。所以第一輪運算機率模型的位置扭曲機率不用一般常用的平均分布 $Pr(j \mid i, k, m) = 1/m$，而採用無母數的統計法，令位置扭曲機率的值如下：

$$Pr(i \mid j,k,m) = 1 - \left| \frac{j-0.5}{m} - \frac{i-0.5}{k} \right| \qquad [1]$$

其中 i = 英文字位置，k = 英文字總數，j = 中文字位置，m = 中文字總數。

S	T	i	k	j	M	$Pr(j \mid i,k,m)$
flight	8	1	2	1	4	0.875
flight	字	1	2	2	4	0.875
flight	飛	1	2	3	4	0.625
flight	行	1	2	4	4	0.375
eight	8	2	2	1	4	0.375
eight	字	2	2	2	4	0.625
eight	飛	2	2	3	4	0.875
eight	行	2	2	4	4	0.875

表 1　位置扭曲機率的無母數統計

對於每一筆雙語片語，我們假設每個英文字可以翻譯成其中任何一個中文字，但是其機率會因位置不同而異。例如某一筆記錄是 2 個英文字翻譯成 4 個中文字，我們可以得到 8 個英中文字的任意配對。每一個配對的位置扭曲機率和公式 1 的 $Pr(j \mid i,k,m)$值成正比。例如，對語料中雙語片語（flight eight, 8 字飛行），我們用公式 1 可以計算得到如表 1 的任意詞彙配對的位置扭曲機率。

有了任意配對的位置扭曲機率後，我們就可據此估計語料庫片語中的任何英文字 E 和中文字 C 間的翻譯機率 $\Pr(C \mid E)$，公式如下：

$$\Pr(C \mid E) = \frac{\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{m}\delta(E, P_n(i))\delta(C, Q_n(j))\Pr(j \mid i, k, m)}{\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{m}\delta(E, P_n(i))\Pr(j \mid i, k, m)} \qquad [2]$$

其中 $P_n(i)$為 P_n 之第 i 字，$Q_n(j)$為 Q_n 之第 j 字，$k = |P_n|$，$m = |Q_n|$

$\delta(x, y) = 1$ 若 $x = y$, $\delta(x, y) = 0$ 若 $x \neq y$

S_i	T_j	i	k	j	m	$Pr(j\|i,k,m)$	$Pr(T_j \mid S_i)$	$Pr(T_j \mid S_i)\ Pr(j\|i,k,m)$
flight	8	1	2	1	4	0.875	0.00797	0.00697
flight	字	1	2	2	4	0.875	0.00797	0.00697
flight	飛	1	2	3	4	0.625	0.25770	0.16106
flight	行	1	2	4	4	0.375	0.16901	0.06338
eight	8	2	2	1	4	0.375	0.02903	0.01089
eight	字	2	2	2	4	0.625	0.04839	0.03024
eight	飛	2	2	3	4	0.875	0.06774	0.05927
eight	行	2	2	4	4	0.875	0.06774	0.05927

表 2　位置扭曲機率與詞彙翻譯機率的估計值

公式 2 的用意在於加總 E 和 C 的在所有片語中的機率值，並除以 E 和所有中文的機率值的總合，使得 $\Pr(C \mid E)$的機率值介於 0 和 1 之間。依據公式 2 所得到的機率值，我們可以估計任何片語內任意字的配對的機率值。表 2 列出表 1 的任意配對的詞彙翻譯機率。

4.2 EM 演算法的第一輪計算

第一次的對應最佳化

有了起始的機率函數估計值，我們就可以進行 EM 演算法中的最佳化步驟。我們採取簡單的貪婪法（Greedy Method）來求取每一組雙語片語（P_n, Q_n）的最佳對應。我們假設簡單的孳生模型：一個英文可以對應到 0 到多個中文字，而每個中文字只能對應到最多一個英文字。有了片語內的詞彙翻譯與位置扭曲機率的起始估計值與其乘積（如表 2），我們就可以逐次選取最高機率值者，產生英文和中文字的配對，並根據假設的孳生模型，排除其他的英文字和此中文字的配對。反覆的執行上述步驟，直到沒有剩餘的中文字，或機率值低於某一個門檻值 (threshold) 為止。若有剩餘的中文字，就視為沒有對應到英文字。最低對應的機率門檻值，可以避免信賴度太低的錯誤對應，也有助於導入 0 對 1，0 對多的孳生模式。經過實際抽樣觀察之後，以 0.008 為門檻值，可去掉大部分低信賴度的錯誤配對。再回到 “flight eight” 的例子，由表 2 的機率值，我們可得到如表 3 的對應方式（0, 34, 12）。

S_i	T_i	i	j	k	m	$Pr(j\|i,k,m)$	$Pr(T_j \mid S_i)$	$Pr(T_j \mid S_i)\ Pr(j\|i,k,m)$
flight	飛	1	3	2	4	0.625	0.25770	0.16106
flight	行	1	4	2	4	0.375	0.16901	0.06338
eight	8	2	1	2	4	0.375	0.02903	0.01089
eight	字	2	2	2	4	0.625	0.04839	0.03024

表 3（flight eight, 8 字飛行）之最佳對應（0, 34, 12）

期望值的估算 - 翻譯對應機率函數

經過機率最佳化求取最可能的對應方式後，我們就可以拋棄個別字的位

置扭曲機率，導入新的翻譯對應機率模型，直接估計整個對應方式的機率值。我們依照片語的英中字數，統計出英中文字數 k 與 m 固定下，各種對應方式 A 的機率：

$$\Pr(A \mid k,m) = \frac{count(A\text{為}(\mathbf{S},\mathbf{T})\text{的對應})}{count(k = |\mathbf{S}|, m = |\mathbf{T}|)} \qquad [3]$$

k	m	A			$Pr(A\|k,m)$
		A_0	A_1	A_2	
2	4	0	12	34	0.572025052
2	4	0	123	4	0.121317560
2	4	0	1	234	0.085479007
2	4	0	1234	0	0.078056136
2	4	0	0	1234	0.065066110
2	4	0	124	3	0.020992809
2	4	0	2	134	0.016585479
2	4	0	3	124	0.007886801
2	4	0	34	12	0.005915101
2	4	0	13	24	0.004059383
2	4	0	134	2	0.003363489
2	4	0	23	14	0.002551612
2	4	0	4	123	0.002319647
2	4	1	0	234	0.002087683
2	4	0	234	1	0.001855718

表 4　兩字對四字片語的翻譯對應機率值最高的前 15 名

在實驗中，EM 演算法的第一輪自動的發掘出 601 種對應方式。以兩字對四字片語而言，有 38 種方式。表 4 列出依照機率由高到低排列的前 15 名對應方式。由表 4 可以觀察到幾點：

1. 機率估計的結果，和我們的認知沒有出入：

   最可能的片語翻譯的順序是保留原文的順序。

   同一英文字的翻譯的目標位置是連續的。

   一個英文字最可能翻譯到 2 個中文字。

2. 對應安排的機率值集中在少數的幾個樣式上。最可能的對應，佔了 90%以上的機率。

3. 對應機率函數收斂的速度很快。

表 5 列出 2 對 4 字片語對應機率值前 5 名的實際例子。

S	**T**	T(A_0)	S_1	T(A_1)	S_2	T(A_2)
T-shaped antenna	T形天線		T-shaped	T形	antenna	天線
X-ray examination	X光檢查		X-ray	X光	examination	檢查
irresistible force	不可抗力		irresistible	不可抗	force	力
Unwritten law	不成文法		unwritten	不成文	law	法
Central Asia	中亞細亞		Central	中	Asia	亞細亞
mutual non-interference	互不干涉		mutual	互	non-interference	不干涉
undesirable element	不良少年		undesirable	不良少年	element	
unalterable truth	不易之論		unalterable	不易之論	truth	
come soon	不日放映		come		soon	不日放映
a desperado	不逞之徒		a		desperado	不逞之徒

表 5　二字到四字片語，最可能的 5 種對應方式的實例

期望值的估算 – 詞彙翻譯對應機率

在統計對應方式的機率的同時，我們同樣的也拿 **4.2** 節最佳化的結果，估計英文字翻譯成不同中文字的機率。我們採取和第一輪不一樣的做法，不再考慮英文字對應到中文單字的機率，而是考慮每一個英文字在片語中，所對應到的中文字串。這些中文字串大部分的情況是連續的，而且是詞典裡常見的詞項。當然也有少數的例子，英文的對應目標是空字串、

不連續字串、不能獨用的詞素（bound morpheme）等等情況。我們以“$empty$”來代表英文字對應到空字串的情況。考慮資料不足（data sparseness）的可能，我們導入“any”來代表英文字對應到訓練外的任意中文字串的情況，並採用 Good-Turing 的平滑化方法（smoothing method）來估計any的翻譯機率。

E	C	Pr（C|E）
flight	飛行	0.6480231012
flight	飛	0.1411528654
flight	航空	0.0602616768
flight	$empty$	0.0296114718
flight	航	0.0296114718
flight	分	0.0041786956
flight	分隊	0.0041786956
flight	飛班機	0.0041786956
flight	飛航	0.0041786956
flight	飛機	0.0041786956
flight	航飛	0.0041786956
flight	黑	0.0041786956
flight	群	0.0041786956
flight	any	0.0000009248

表 6 “flight”翻譯成不同中文字串的機率

表 6 列出 flight 翻譯成不同中文字串的機率，包括一般的詞、詞素、$empty$、any。在這一輪的期望值估計中，flight 對應到$empty$的機率估計值 0.0296114718 仍然過高。只要翻譯對應機率如表 4 的（0,0,1234）

和（1,0,234）的機率，以及any機率的估計值估計得合理，我們期望在 EM 演算法的以後的幾個輪迴中，兩者互相競爭的情況下，$empty$機率的估計值會逐漸的降低，而趨近合理的區段。

4.3 EM 演算法的第二輪計算

在第一輪的期望值估計之後，我們可以再次的求取片語的最可能對應方式。在第二輪的運算當中，我們不再使用公式 1 的位置扭曲機率，而是採用已經估計出來的整體性的翻譯對應機率。

S	**T**	A_0	A_1	A_2	T(A_0)	T(A_1)	T(A_2)	*Pr*(**T**|**S**,**A**)
flight eight	8 字飛行	0	34	12		飛行	8 字	0.0000788100
flight eight	8 字飛行	0	3	124		飛	8 字行	0.0000000051
flight eight	8 字飛行	12	34	0	8 字	飛行	$empty$	0.0000000007
flight eight	8 字飛行	12	3	4	8 字	飛	行	0.0000000003
flight eight	8 字飛行	2	3	14	字	飛	8 行	0.0000000001

表 7　Pr(8 字飛行| flight eight) 機率值最高之前 5 名

第二輪運算中，我們對每一筆雙語片語（**S**, **T**），依據其英文和中文字數，考慮相符的所有的對應方式 **A**，計算其翻譯機率 *Pr*(**T** | **S**, **A**)。對於某一對應方式 **A**，*Pr*(**T** | **S**, **A**)為 **A** 的機率和由 **A** 所決定的詞彙配對(S_i, T(A_i))的機率乘積：

$$\Pr(T \mid S) = \max_A \Pr(T \mid S, A) = \max_A \Pr(A \mid k, m) \prod_{i=1}^{k} \Pr(T(A_i) \mid S_i)$$

因此最可能的對應 **A*** 可由下列公式決定

$$A^* = \arg\max_A \Pr(T \mid S, A)$$
$$= \arg\max_A \Pr(A \mid k, m) \prod_{i=1}^{k} \Pr(T(A_i) \mid S_i) \qquad [4]$$

其中 $k = |\mathbf{S}|$, $m = |\mathbf{T}|$

以(**S**, **T**) =（flight eight, ８字飛行）為例，對於不同的對應 **A**，其翻譯機率的計算如下：

A = (0, 12, 34)：

Pr(**T** | **S**,**A**) = *Pr*(0,12,34 | 2,4) *Pr*(８字|flight) *Pr*(飛行|eight)

A = (0, 34, 12)：

Pr(**T** | **S**,**A**) = *Pr*(0,34,12 | 2,4) *Pr*(飛行 | flight) *Pr*(８字|eight)

A = (0, 3, 124)：

Pr(**T** | **S**,**A**) = *Pr*(0, 3, 124 | 2,4) *Pr*(飛| flight) *Pr*(８字行|eight)

A = (2, 34, 1)：

Pr(**T** | **S**,**A**) = *Pr*(2, 34, 1 | 2,4) *Pr*(飛行| flight) *Pr*(８|eight)

A = (12, 34, 0)：

Pr(**T** | **S**,**A**) = *Pr*(12, 34, 0 | 2,4) *Pr*(飛行| flight) *Pr*($empty$|eight)

表 7 列出（flight eight, ８字飛行）的幾個最高翻譯機率值的對應方式。表 7 的數值顯示第二輪的統計估計值已經相當的收斂，可以導出正確的對應分析 **A*** = (0, 34, 12)。

5. 實驗結果與評估

我們進行的實驗，證明了新的統計式片語翻譯模型確實可行，能產生相當正確的對應分析。新模型中導入的翻譯對應機率的參數不會過度的膨脹，因此 10 萬筆的資料就可以估計出相當可靠的各項機率值。由於新的模型，避免了許多機率值的乘積，EM 演算法的花費的時間較少，機率函數的收斂速度也比較快。

S	T	第一輪結果			第二輪結果		
		T(A_0)	T(A_1)	T(A_2)	T(A_0)	T(A_1)	T(A_2)
association football	A式足球			A式足球		A式	足球
delay flip-flop	D型正反器			D型正反器		D型	正反器
I demodulator	I信號解調器			I信號解調器		I信號	解調器
Disgraceful act	不友好行動			不友好行動		不友好	行動
disregard to	不拘於		不拘於			不拘	於
secret ballot	不記名投票			不記名投票		不記名	投票
bearer stock	不記名股票		不記名票	股		不記名	股票
false retrieval	不實檢索			不實檢索		不實	檢索
used car	中古車		中古車			中古	車
infix operation	中序運算		中序運算			中序	運算

表 8　第二輪運算之後翻譯結果變好的例子

由實驗的結果觀察，以翻譯對應機率替代孳生機率和位置扭曲機率，確實可以得到比較正確的對應分析。在對應的問題比較困難的幾個情況仍然能夠做出正確的分析：

1. 比較偏離常態的罕用翻譯，如 association 通常翻譯成「協會」、「學會」。而 association football 中卻翻譯成類音譯的「A 式」。
2. 翻譯非常分散，沒有定譯的虛詞或輕動詞（light verb），如 make、take、to 等。
3. 和原文不一致的翻譯順序，如（flight eight, 8 字飛行）。

我們檢視對應分析的結果，特別觀察這幾種困難的情況，比較其第一輪和第二輪分析的結果。我們發現 Brown 原始模型不盡理想，使得第一輪的許多分析不正確。這些情況到了第二輪時，使用了新模型的分析

後，大部分很明顯的已經扭轉到正確的分析。部分的例子，請參見表 8。

為了評估實驗的效能，我們使用 Och 等人(2000) 的評估方法。我們從實驗第二輪結果中，隨機抽取 100 個樣本 (包含 2 個英文字及 3 個英文字的樣本各 50 個)，由人工對這些樣本做對應的標示，以作為參考答案。標示分為 2 種：S (sure) 和 P (possible)，S 表示確定的對應，P 表示可能的對應，且 $S \subseteq P$。將實驗的結果與人工標示的參考答案比較，我們可以得到以下的召回率(recall)、準確率(precision)與錯誤率(error rate)：

$$recall = \frac{|A \text{ I } S|}{|S|} = \frac{185}{212} = 0.873$$

$$precision = \frac{|A \text{ I } P|}{|A|} = \frac{226}{273} = 0.828$$

$$AER(S,P;A) = 1 - \frac{|A \text{ I } S| + |A \text{ I } P|}{|A| + |S|} = 1 - \frac{185 + 226}{273 + 212} = 0.153$$

6 討論

我們就實驗的結果，進一步討論平滑化的改進做法。我們也分析統計式片語翻譯模型的可能的應用。

6.1 其他平滑化方法

由訓練語料得到各項翻譯的機率後，我們可以用這些機率，繼續對應訓練以外的其他片語，或是進行片語的翻譯。此時，我們可能會因為資料不足，而遇到訓練資料以外的情況，例如

（flight attendant 空服員）

我們的訓練語料，並沒有（flight, 空）的詞彙配對，來正確的分析（flight attendant 空服員）的對應。當然此時我們可以應用 flight 對any的機率。但是any的機率是平均的分配，無法考慮到有少數訓練外的狀況比較可

能，而大多數訓練外的狀況相當不可能。這些少數比較可能的狀況和中文縮寫的特性有關。另外有些中文字容易孳生很多的同義或近義，也會造成相同的效應。

中文的使用有縮寫的現象，所以訓練內的詞彙配對如（flight, 航空）的部分翻譯（flight, 航）與（flight, 空）在訓練外出現的可能性不低，而非部份翻譯的（flight, 員）則趨近於 0。同樣的，在（attendant, 服務員）配對例子中，部分翻譯（attendant, 服員）、（attendant, 服）、（attendant, 服務）的可能性也顯然高於（attendant, any）的平均值。另外翻譯有部份重疊的情況，也應給予較高的平滑機率。例如訓練內的詞彙配對有（preservation, 保留），而（preservation, 保持）與（preservation, 保護）雖然沒有出現在訓練語料中，其可能性仍然高於其他完全沒有重疊的翻譯配對。

如果沒有這樣的考慮，對於（flight attendant 空服員）的對應，詞彙翻譯機率就會全然都使用 *Pr*(any|flight)和 *Pr*(any|eight)的機率值，無法區隔可能與不可能的配對。如此將流於由翻譯對應機率 *Pr*(**A**|2,3)來決定一切。在這種狀況下，由於兩字到三字片語的最高可能對應為（0, 12, 3），我們很可能得到以下的不完全正確的對應分析：

（flight, 空服）

（attendant, 員）

若能考慮翻譯部分符合的條件，給予（flight, 空）與（attendant, 服員）較高的平滑機率估計值，則比較容易得到正確的對應分析，如

（flight, 空）

（attendant, 服員）

目前我們正實驗以英文到中文單字以及英文到中文雙字的兩組 LTP，來合成機率估計值。實驗的目標在於，讓部分字相符的對應，可以透過單字 LTP 模型得比較合理的估計值。

6.2 統計式片語翻譯模型的應用

我們提出的新的統計式片語對應語翻譯的模型，可以應用於翻譯跨語言檢索中的關鍵詞。新模型也可以在平行語料庫擷取雙語的片語，提供建立語料庫相關的翻譯詞典，作為翻譯與術語管理的基本工具。

6.2.1 在跨語言檢索上的應用

在跨語言檢索的研究顯示通常不到一半的查詢的關鍵詞組，可以在雙語詞典中查到適當的翻譯。當詞典沒有收錄詞組時，我們必須逐字翻譯，通常每字都有許多的翻譯，而只有部分和查詢主題相關。統計式的片語翻譯模型，可以發揮幾種作用：

1. 模型中的詞彙翻譯機率，提供比雙語詞典單字詞條更豐富的諸多可能性。若不能全部採用，也可取機率值最高、同時文件頻率（document frequency）適中的翻譯。這個做法簡單可行，也可稍稍減少翻譯和文件庫相關性不大的缺點。
2. 透過雜訊通道模型，結合翻譯機率函數，與文件庫所訓練出來的 N-gram 語言模型，可以用搜尋演算法，產生機率最佳化的翻譯。這個做法可以大大的提升翻譯結果的文件庫相關性。

明確的來說，以訓練出來的辭彙翻譯機率和翻譯對應機率，我們更容易求取查詢關鍵詞組 S 的最佳翻譯 *T**。我們可以使用下列公式：

$$\Pr(T \mid S) = \max_{A} \Pr(T \mid S, A)$$

$$T^* = \arg\max_T \Pr(T \mid S)\Pr(T)$$
$$= \arg\max_T \max_A \Pr(T \mid S, A)\Pr(T)$$
$$= \arg\max_T \max_A \Pr(A \mid k, m)\prod_{i=1}^{k} \Pr(T(A_i) \mid S_i)\Pr(T)$$

其中 $Pr(T)$ = 翻譯 T 的語言模型機率

$k = S$ 的字數

$m = T$ 的字數

$Pr(A \mid k, m)$的模型簡化了 $Pr(T \mid S)$的計算，使得我們更容易以分岔與限制演算法（Branch and Bound Algorithm）搜尋機率最大化的 **T***值。我們可以由最高機率值的 $Pr(A \mid k, *)$、$Pr(* \mid S_i)$ 組合的解（solution）開始搜尋，建立搜尋的高限（upper bound），再利用 $Pr(* \mid S_i)$、$Pr(T)$的 N-gram 模型的最高機率值，達到限制搜尋範圍的效果。在不遺漏最佳解的情況下，縮短搜尋的時間。

6.2.2 在雙語片語對應的應用

學者大多認為統計式機器翻譯最有應用潛力的地方，在於雙語詞典的編輯與機率翻譯詞典的發展。當然在這樣的考慮下，詞彙對應的發現，不限於單字詞，而應及於多字的片語（Kupiec 1993）。

利用新發展出來的模型，我們提出一套逐句進行的片語對應做法。以新的統計式片語翻譯模型為中心，我們可以透過下列的步驟，在英中例句中擷取對應的英中片語。對例句中的英文的基本片語 P 我們可以計算其最佳的翻譯對應 A 與翻譯目標 $T_{j+1,\, j+m}$：

$$\arg\max_{A,\,j,\,m} \Pr(A \mid n(i), m)\Pr(T_{j+1,\,j+m} \mid P, A)$$
$$= \arg\max_{A,\,j,\,m} \Pr(A \mid n(i), m)\Pr(T_{j+1,\,j+m} \mid S_{b,e}, A)$$
$$= \arg\max_{A,\,j,\,m} \Pr(A \mid n(i), m)\prod_{k=1}^{n(i)} \Pr(T_{j+B(A,K),\,j+E(A,k)} \mid S_{b+k-1})$$

其中 **A** 為 n 個英文對應到 m 個中文的任何可能的對應

$Pr(A \mid n, m)$ 為 A 的機率函數值

S_{b+k-1} 為基本片語 P 的第 k 個字

$T_{j+B(A,k),\, j+E(A,k)}$ 為 P 的第 k 個字在翻譯對應 A 下的翻譯目標

逐句式的雙語片語對應演算法

輸入：英文句子 S 中文句子 T

輸出：一組 k 個雙語片語 $(P_1, Q_1), (P_2, Q_2), \ldots, (P_k, Q_k)$

(1). 以詞性分析程式（Part of speech tagger）分析英文句子 S 的單字的詞性，並取得原形化後的字根（lemma）。

(2). 以淺型剖析（shallow parsing）的做法，擷取句子的基本片語（basic phrase）：$P_1, P_2, \ldots, P_k$。令 $P_I = S_{b(i),\, e(i)}$, $|P_b| = n(i)$

(3). 對每一個基本片語 P_I 計算其最佳的翻譯對應 $\boldsymbol{A}$ 與翻譯目標 $T_{j^*(i)+1,\, j^*(i)+m^*(i)}$。做法是以分岔與限制（Branch and Bound Algorithm）搜尋最大化下列公式的 $\boldsymbol{A}^*(i), j^*(i), m^*(i)$

$$\arg\max_{A,j,m} \Pr(A \mid n(i), m) \prod_{k=1}^{n(i)} \Pr(T_{j+B(A,k),\, j+E(A,k)} \mid S_{b+k-1})$$

(4). 對每一個基本片語 P_I 輸出 $(P_I, T_{j^*(i)+1,\, j^*(i)+m^*(i)})$

7. 結論與未來的研究方向

雖然統計式機器翻譯的研究，已經有十多年的歷史，在本研究中我們發現仍然有很大的改進空間，特別是在片語的對應與翻譯方面。我們提出新的統計式片語模型來進行片語的對應，並可應用於查詢關鍵詞的翻譯，以提高跨語言檢索的效率。我們在實驗中，初步驗證新的模型比起

Brown 的原始模型，確實可以在計算效率與對應效果上，有很大的改進。

統計式的做法確實對翻譯辭彙的整理、編輯有很大的效用。在這篇論文中，我們也討論統計式片語翻譯模型，應用於關鍵詞翻譯與雙語例句的片語對應的做法。

我們認為未來統計式雙語對應與機器翻譯，在跨語言檢索應該還有很大的空間可以發揮。幾個可能的研究方向包括：

1. 導入句法的訊息，對於不同的名詞、動詞、形容詞片語，訓練不同的統計式模型，可能可以導致更好的對應與翻譯的效果。
2. 導入詞彙語義解析的做法，先行分析查詢主題的語意範疇，縮小翻譯選擇的範圍。
3. 將語義的限制，如 Wordnet 的上位詞導入詞彙翻譯機率中，以配合詞彙語義解析的做法。

致謝

本文之研究受到國科會編號 892420H007001 計畫之補助。作者也非常感謝匿名審查者所提供之寶貴建議。

參考文獻

1. BDC 1992 *The BDC Chinese-English electronic dictionary* (version 2.0), Behavior Design Corporation, Taiwan.
2. Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Mercer R. L., and Roosin P. S. 1988 *A Statistical Approach to Language Translation*, In Proceedings of the 12th International Conference on Computational Linguistics, Budapest, Hungary, pp. 71-76.

3. Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., and Roosin P. S. 1990 *A Statistical Approach to Machine Translation*, Computational Linguistics, 16/2, pp. 79-85.

4. Brown, P. F., Della Pietra S. A., Della Pietra V. J., and Mercer R. L. 1993 *The Mathematics of Statistical Machine Translation: Parameter Estimation*, Computational Linguistics, 19/2, pp. 263-311.

5. Chang, J. S. et al. 2001. *Nathu IR System at NTCIR-2*. In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, pp. (5) 49-52, National Institute of Informatics, Japan.

6. Chang, J. S., Ker S. J., and Chen M. H. 1998 *Taxonomy and Lexical Semantics – From the Perspective of Machine Readable Dictionary*, In Proceedings of the third Conference of the Association for Machine Translation in the Americas (AMTA), pp. 199-212.

7. Chen, H.H., G.W. Bian and W.C. Lin. 1999. *Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval*. In Proceedings of the 37th Annual Meeting of the Association for Computation Linguistics, pp 215-222.

8. Dagan, I., Church K. W. and Gale W. A. 1993 *Robust Bilingual Word Alignment or Machine Aided Translation*, In Proceedings of the Workshop on Very Large Corpora Academic and Industrial Perspectives, pp. 1-8.

9. Fung, P. and McKeown K. 1994 *Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping*, In Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA), pp. 81-88, Columbia, Maryland, USA.

10. Gale, W. A. and Church K. W. 1991 *Identifying Word Correspondences in Parallel Texts*, In Proceedings of the Fourth DARPA Speech and Natural Language Workshop, pp. 152-157.

11. Gey, F C and A. Chen. 1997. *Phrase Discovery for English and Cross-Language Retrieval at TREC-6*. In Proceedings of the 6th Text Retrieval Evaluation Conference, pp 637-648.

12. Ide, N. and J Veronis. 1998. *Special Issue on Word Sense Disambiguation*, editors, Computational Linguistics, 24/1.

13. Isabelle, P. 1987 *Machine Translation at the TAUM Group*, In M. King, editor, Machine Translation Today: The State of the Art, Proceedings of the Third Lugano Tutorial, pp. 247-277.

14. Kando, Noriko, Kenro Aihara, Koji Eguchi and Hiroyuki Kato. 2001. Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, National Institute of Informatics, Japan.

15. Kay, M. and Röscheisen M. 1988 *Text-Translation Alignment*, Technical Report P90-00143, Xerox Palo Alto Research Center.

16. Ker, S. J. and Chang J. S. 1997 *A Class-base Approach to Word Alignment*, Computational Linguistics, 23/2, pp. 313-343.

17. Knight, K. and J Graehl. 1997. *Machine Transliteration*, In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of ACL European Chapter, pp. 128-135.

18. Kupiec, Julian. 1993 *An Algorithm for finding noun phrase correspondence in bilingual corpus*, In ACL 31, 23/2, pp. 17-22.

19. Kwok, K L. 2001. *NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS*. In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization, pp. (5) 14-20, National Institute of Informatics, Japan.

20. Longman Group 1992 *Longman English-Chinese Dictionary of Contemporary English*, Published by Longman Group (Far East) Ltd., Hong Kong.

21. McCarley, J. Scott. 1999. *Should we Translate the Documents or the Queries in Cross-Language Information Retrieval*? In Proceedings of the 37th Annual Meeting of the Association for Computation Linguistics, pp 208-214.

22. Melamed, I. D. 1996 *Automatic Construction of Clean Broad-Coverage Translation Lexicons*, In Proceedings of the second Conference of the Association for Machine Translation in the Americas (AMTA), pp. 125-134.

23. Nagao, M. 1986 *Machine Translation: How Far Can it Go*? Oxford University Press, Oxford.

24. Oard, D W and J. Wang. 1999. *Effect of Term Segmentation on Chinese/English Cross-Language Information Retrieval*. In Proceedings of the Symposium on String and Processing and Information Retrieval. http://www.glue.umd.edu/~oard/research.html.

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

26. Pirkola, A. 1998. *The Effect of Query Structure and Dictionary Setups in Dictionary-based Cross-Language Retrieval*. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 55-63.

27. Smadja, F., McKeown K., and Hatzivassiloglou V. 1996 *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Computational Linguistics, 22/1, pp. 1-38.

28. Utsuro, T., Ikeda H., Yamane M., Matsumoto M., and Nagao M. 1994 *Bilingual Text Matching Using Bilingual Dictionary and Statistics*, In Proceedings of the 15th International Conference on Computational Linguistics, pp. 1076-1082.

29. Wu, D. and Xia X. 1994 *Learning an English-Chinese Lexicon from a Parallel Corpus*, In Proceedings of the first Conference of the Association for Machine Translation in the Americas (AMTA), pp. 206-213.