Computational Linguistics and Chinese Language Processing
Vol. 6, No. 2, August 2001, pp. 43-64

統計式片語翻譯模型

Statistical Translation Model for Phrases

張俊盛*　　游大緯*　　李俊仁*+

Jason S Chang, David Yu, Chun-Jun Lee

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有高度的相關性，才能達到檢索的效果。目前翻譯查詢關鍵詞的做法，無論是採用現成的翻譯軟體，或者使用一般性的雙語詞典，都很難確保產生和文件相關的翻譯。因此我們希望能夠透過統計式片語機器翻譯（Statistical Phrase Translation Model, SPTM）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。實驗中我們利用 BDC 雙語電子辭典實驗以 SPTM 進行片語內的詞彙對應。以 SPTM 產生對應分析，比較快速，而且正確率比較高。

Abstract

Machine Translation is one of the most difficult problems in the field of natural language processing. In the past, MT has been applied to professional communication in the process of translating technical and corporate document on a specific domain. Recent years saw the rapid development of Internet as a new form of communication and information exchange, and the need to access information across the language barrier became apparent. People began to look into the role that MT can play in Cross Language Information Retrieval. The prevalent approach to CLIR is based on translation of query, in particular query

*國立清華大學資訊工程研究所 E-mail: jschang@cs.nthu.edu.tw
+中華電信研究所

phrases. However, for CLIR there is an additional new objective of translating into something that is relevant to the collection being searched upon. Therefore, the current approach of using general bilingual word list or an off-the-shelf commercial MT software is bound to be very ineffective in terms of retrieving relevant documents. We propose a new approach to Statistical Phrase Translation Model (SPTM), aimed at achieving a tighter estimation of phrase translation. Experiments were conducted using bilingual phrases in BDC Electronic Chinese-English Dictionary. Preliminary results shows the approach is much faster and produces better word alignment for phrases, which has not been possible using previous approaches.

Keywords: Statistical Machine Translation; Phrase Translation; Cross-language Information Retrieval.

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機器輔助翻譯（Machine Assisted Human Translation）[Lange, Gaussier, and Daille, 1997]、跨語言檢索（Cross-Language Information Retrieval）[Gey and Chen, 1997] 及電腦輔助語言學習（Computer Assisted Language Learning）[Shei and Pain, 2001] 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翻譯查詢關鍵詞的做法。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
- Communism
- 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，和技術文件翻譯很重要的不同點，在於翻譯的結果，是要拿來在檢索系統的文件庫（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主要做法，大致為：

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，將英文中音譯的日、中人名地名轉換回原文的專有名詞 [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，也不具備嚴謹的理論基礎，因此影響到改進檢索效率的空間。

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。統計式翻譯模型（Statistical Translation Model）的應用於文件的機器翻譯 [Jones and Havrilla, 1998]、翻譯語料之詞彙對應 [Gale and Church 1992, Melamed 2000]、字典建構 [Melamed 2000]。IBM Watson 研究中心的 Brown 等 [1988, 1990, 1993] 最早提出理論嚴謹的統計式機器翻譯做法。Wu [1997] 提出以無標記二元倒裝樹之句法結構為基礎的統計式翻譯模型。Wang [1998] 和 Och 等人 [1999] 採用片語與樣板，導入句法結構於統計式的翻譯模型。Yamata 和 Knight [2001] 則提出完整的句法導向的統計式翻譯模型，以規範例兩種語言的句法剖析樹的對應關係。其模型包括剖析樹之子樹的順序重排，功能詞彙節點的增加與刪除。

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且嚴謹的做法。在這篇論文中，我們提出一種新的翻譯指派機率（Assignment 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 [1993] 的統計式機器翻器模型 Model 3 下，原文 S 和譯文 T 的翻譯機率（Translation Probability） $Pr(T|S)$，可以分解成以下的三個機率函數：

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

$$Pr(T_j|S_i)$$

(b) 孳生機率（Fertility Probability）

$$Pr(a|S_i)$$

(c) 位置扭曲機率（Distortion Probability）

$$Pr(j|i,k,m)$$

其中

S_i 為 S 的第 i 個字
T_j 為 T 的第 j 個字
a 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「期望值估計」的步驟，就是在目前的機率函數估計值下，求取所有翻譯對應的機率值。而「最佳化」的步驟，就是以所有的雙語語料樣本的翻譯對應為根據，估計三個機率函數的最大概似估計值（Maximum Likelihood Estimation）。

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

IBM Model 3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

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

因此獨立假設下的機率，幾乎大部分的情況下會造成過低的估計。即便是很可能的翻譯對應方式，其機率值經過一連串位置扭曲機率的乘積，常趨於不合理的低數值。例如，檢視三字英文與五字中文的互譯片語樣本，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對應到一個英文字，沒有中文字無法對應的情況）。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^*的最大概似估計值（Maximum Likelihood Estimation），得到

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

然而在機率獨立的假設下

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

即使位置扭曲機率值以高數值（0.6）估計 $P(j \mid i,3,5)$，其乘積仍然遠低於合理的估計值：

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

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指派機率（Assignment 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，而是以整體的對應來一併考慮。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) **指派機率**（Assignment 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**

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. **以指派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. **指派的位置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？指派機率的參數量，會不會過多，會不會導致估計的速度過慢？**
3. **指派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。因為，** 4 **字詞（含）以上之條目僅佔訓練語料** 4%**不到，不足以求得有意義且具代表性模型參數。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 **來代表這組英中片語翻譯語料。**

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、指派機率。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，我們採用 IBM Model 2，以詞彙翻譯機率與位置扭曲機率，來估計訓練統計模型的機率參數。在 EM 演算法的第二輪之後才開始使用新模型的指派機率。
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 $Pr(j \mid i, k, m) = 1/m$，而採用參數式的統計法，根據片語翻譯傾向於保留原文順序的的經驗法則，令位置扭曲機率的值如下：

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

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。對於公式[1]的機率值，需要再調整，使得在 i, k, m 值固定時，對所有的 j 值，$\Pr(j \mid i, k, m)$的加總為 1。

表 1. 位置扭曲機率的起始估計值

S_i	T_j	i	k	i	M	$Pr(i \mid i.k.m)$
flight	8	1	2	1	4	0.318
flight	字	1	2	2	4	0.318
flight	飛	1	2	3	4	0.227
flight	行	1	2	4	4	0.136
eight	8	2	2	1	4	0.136
eight	字	2	2	2	4	0.227
eight	飛	2	2	3	4	0.318
eight	行	2	2	4	4	0.318

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如公式 1 的 $Pr(j \mid i,k,m)$值。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。

表 2. 位置扭曲機率與詞彙翻譯機率的估計值

S_i	T_j	i	k	j	m	$Pr(j\|i,k,m)$	$Pr_{LEX}(C\|E)$	$Pr(T_j\|S_i)$
flight	8	1	2	1	4	0.318	0.00797	0.00253
flight	字	1	2	2	4	0.318	0.00797	0.00253
flight	飛	1	2	3	4	0.227	0.25770	0.05850
flight	行	1	2	4	4	0.136	0.16901	0.02299
eight	8	2	2	1	4	0.136	0.02903	0.00395
eight	字	2	2	2	4	0.227	0.04839	0.01098
eight	飛	2	2	3	4	0.318	0.06774	0.02154
eight	行	2	2	4	4	0.318	0.06774	0.02154

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 $\Pr_{LEX}(C|E)$，公式如下：

$$\Pr(C|E)=\frac{\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{m}d(E,P_n(i))d(C,Q_n(j))\Pr(j|i,k,m)}{\sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{m}d(E,P_n(i))\Pr(j|i,k,m)} \quad (2)$$

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 $x\neq y$

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 C 的機率值的總合，使得對所有的 C 值 $\Pr(C|E)$的機率值加總為 1。依據公式 1 和公式 2 的機率值，我們可以估計任何片語內任意字的配對(S_i, T_j) 的機率值：

$Pr(T_j|S_i)=Pr(j|i,k,m)\,Pr_{LEX}(C|E)$ 其中 $C=T_j$，$E=S_i$

以訓練語料中雙語片語（flight eight,8字飛行）為例。對於其中英文字 flight 與中文字「飛」的對應機率為 Pr_{LEX}(飛 | flight) $Pr(3|1,2,4)$。表 2 列出表 1 的任意配對的詞彙翻譯機率。一般而言，起始的統計估計值相當的精確，如表二所顯示的機率值的估計，都相當合理。例如，正確的對應的對應機率如 Pr(飛 | flight) 和 Pr(行 | flight)分別為 0.05850 與 0.02299，遠高於錯誤的對應的詞彙機率 Pr(8 | flight) 和 Pr(字 | flight)的 0.00253。

4.2 EM 演算法的第一輪計算

有了起始的機率函數估計值，我們就可以進行 EM 演算法來估計翻譯模型中的參數值。我們應用 Viterbi 式訓練的 EM 演算法。在每次重新估算時，依據每一個樣本的最佳對應，而不考慮每一個樣本的所有的可能對應。

第一次的對應最佳化

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），我們就可以對每一個中文字，逐次選取其對應到英文字機率值最高者，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）。

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

S_i	T_j	i	k	j	m	$Pr(j\|i,k,m)$	$Pr_{LEX}(C \mid E)$	$Pr(T_j \mid S_i)$
flight	飛	1	2	3	4	0.227	0.25770	0.0585
flight	行	1	2	4	4	0.318	0.16901	0.05375
eight	字	2	2	2	4	0.227	0.04839	0.01098
eight	8	2	2	1	4	0.136	0.02903	0.00395

指派機率函數值的重新估算

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，導入新的翻譯指派機率模型，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，各種指派方式 A 的機率：

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

表 4. 兩字對四字片語的指派機率值最高的前 12 名

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

表 5. 二字到四字片語，最可能的 5 種指派方式的實例

	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**

在實驗中，EM **演算法的第一輪自動的發掘出** 601 **種指派方式。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** 12 **名指派方式。請參考表** 5 **所列** 2 **對** 4 **字片語對應的實際例子。由表** 4 **可以觀察到幾點：**

1. **機率估計的結果，和我們的認知沒有很大的出入：**
   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. **指派安排的機率值集中在少數的幾個樣式上。最可能的** 3 **種指派，佔了接近** 80% **的機率。而前** 5 **種及** 10 **種指派方式，分別累積了** 95%**及** 99.5%**的機率值。這證明了應用指派機率，可以很有效的在雙語對應或機器翻譯時，限制搜尋的範圍，而不影響到精確性。**

3. **指派機率函數收斂的速度很快。**

表 6. "flight"翻譯成不同中文字串的機率

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

詞彙翻譯機率值的重新估算

在統計指派方式的機率的同時，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。

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。只要指派機率如表 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指派機率。

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

S	**T**	A_0	A_1	A_2	T(A_0)	T(A_1)	T(A_2)	*Pr*(**T,A**|**S**)
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

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，考慮相符的所有的指派方式 A，計算其翻譯機率 *Pr*(T, A | S)。對於某一指派方式 A，*Pr*(T, A | S) 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：

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

因此最可能的指派 A* 可由下列公式決定

$$\begin{aligned} A^* &= \arg\max_A \Pr(T, A \mid S) \\ &= \arg\max_A \Pr(A \mid k, m) \prod_{i=1}^{k} \Pr(T(A_i) \mid S_i) \end{aligned} \tag{4}$$

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

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

A = (0, 12, 34)：

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

A = (0, 34, 12)：

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

A = (0, 3, 124)：

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

A = (2, 34, 1)：

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

A = (12, 34, 0)：

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

在計算過程中，我們採取 Branch and Bound **的搜尋法，以降低搜尋的時間。我們會記錄該筆雙語片語翻譯機率** *Pr*(**T, A** | **S**)**的在搜尋過程的最大值，當逐一考慮相符的所有的指派方式 A 時，若指派機率值已小於當時的翻譯機率最大值時，該指派方式** A **則不予計算，因為再繼續乘以詞彙翻譯機率只會讓機率更小，所以可以予以捨棄，如此可以大幅的增進** EM **演算法的效率。**

表 7 **列出（**flight eight, **8 字飛行）的幾個最高翻譯機率值的指派方式。表** 7 **的數值顯示第二輪的統計估計值已經相當的收斂，最佳的對應**(0, 34, 12)**的機率值** 0.0000788100**，遠超過次佳的對應**(0, 3, 124) **的機率值** 0.0000000051。

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。新模型中導入的指派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。**

表 8. 第二輪運算之後對應結果由錯誤轉為正確的例子

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**

5.1 實驗結果分析

由實驗的結果觀察，以指派機率替代孳生機率和位置扭曲機率，確實可以掌握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，而得到比較正確的對應分析。所以在對應的問題比較困難的幾個情況，仍有可能做出正確的分析：

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 中卻翻譯成類似音譯的「A 式」。雖然 association 不常翻譯成「A 式」，但是指派機率可以比其他模型，給予第一和第二字的「A 式」較高的機率。
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，在指派機率的模型下也都可以得到較高的機率。
3. 和原文不一致的翻譯順序，如（flight eight，8 字飛行），在指派機率的模型下，可以得到適當的機率值。

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。我們發現確實這些情況到了第二輪使用了新模型後，大部分很明顯的已經扭轉到正確的分析，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。將要評估的雙語條目先由人工標示對應，以作為參考答案。例如其中一個雙語條目是（butter cream biscuit, 奶油夾心餅乾），人工標示的參考答案如圖 1 所示。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 $S \subseteq P$。

butter	***S/P***	***S/P***	·	·	·	·
cream	·	·	***P***	***P***	·	·
biscuit	·	·	·	·	***S/P***	***S/P***
	奶	**油**	**夾**	**心**	**餅**	**乾**

圖 1 人工標示參考答案例子

而我們實驗所做的對應的指派方式為 A，再以（butter cream biscuit, 奶油夾心餅乾）為例，實驗的對應結果如圖 2 所示。

butter	***A***	***A***	·	·	·	·
cream	·	·	***A***	***A***	·	·
biscuit	·	·	·	·	***A***	***A***
	奶	**油**	**夾**	**心**	**餅**	**乾**

圖 2 實驗對應結果的例子

以上例子，我們可以得到 |S|=4，|P|=6 (包括標示為 S 的部分)，|A|=6，|A∩S|=4，|A∩P|=6。根據 Och 等人(2000)所提出的公式，我們可得到召回率(recall)、準確率(precision)與錯誤率(error rate)如下：

$$recall = \frac{|A \cap S|}{|S|} = \frac{4}{4} = 1$$

$$precision = \frac{|A \cap P|}{|A|} = \frac{6}{6} = 1$$

$$AER(S,P;A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} = 1 - \frac{4+6}{6+4} = 0$$

我們從實驗第二輪結果中，隨機抽取 500 個樣本 (包含 2 個英文字及 3 個英文字的樣本各 250 個)，由人工對這些樣本做對應的標示，以作為參考答案。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)：

$$recall = \frac{|A \cap S|}{|S|} = \frac{1116}{1248} = 0.894$$

$$precision = \frac{|A \cap P|}{|A|} = \frac{1308}{1517} = 0.862$$

$$AER(S,P;A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} = 1 - \frac{1116+1308}{1517+1248} = 0.123$$

在第二輪實驗之後，我們以 EM 演算法繼續第三輪至第五輪的實驗，得到第一至五輪的召回率、準確率與錯誤率如表 9 所示。

表 9. 使用新模型下訓練過程召回率、準確率、錯誤率的收斂情形

	第一輪	第二輪	第三輪	第四輪	第五輪
召回率	0.853	0.894	0.885	0.874	0.873
準確率	0.796	0.862	0.859	0.851	0.849
錯誤率	0.178	0.123	0.129	0.139	0.140

為了了解新的指派機率是否比較有效率，我們也以同樣的語料、同樣的演算法，但把新模型中的指派機率換成 IBM 模型中的孳生機率及位置扭曲機率，進行 IBM Model 的對照組實驗。整體而言，導入新的指派機率，取代孳生與扭曲機率，在執行速度上有很大的改進，在正確率上也略勝一籌。雖然 IBM model 3 對於詞彙孳生部分，考慮了個別詞彙的因素，實驗結果卻顯示，IBM model 3 的錯誤率較高。

表 10. 使用 IBM Model 3 訓練過程的召回率、準確率、錯誤率的收斂情形

	第一輪	第二輪	第三輪
召回率	0.853	0.883	0.867
準確率	0.796	0.848	0.838
錯誤率	0.178	0.136	0.149

5.2 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，在語料庫中繼續對應擷取片語，或是進行跨語言檢索中的查詢片語的翻譯。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，例如在估計語料庫中的某一片語的詞彙對應：

（flight attendant 空服員）

如果我們的訓練語料，沒有（flight, 空）的詞彙配對，就無法正確的分析（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，無法做個別的狀況的考慮。有少數訓練外的翻譯的狀況，是和中文縮寫有關。另外有些中文字容易孳生很多的同義或近義，也會造成很多可能但在訓練外的狀況。這些訓練外的情況，和訓練資料有部分重疊，其機率並不低，我們應該透過比較複雜的機率估計平滑化的做法，給予適當的估計值。

中文的使用有縮寫的現象，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。如此將流於由指派機率 *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詞彙翻譯機率，來合成機

率估計值。實驗的目標在於讓部分字相符的對應，可以透過英文字對中文單字、中文雙字詞彙翻譯機率的線性組合模型，得比較合理的估計值。

6. 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，確實可以在計算效率與對應效果上，有所改進。

我們認為未來統計式雙語對應與機器翻譯，還有很大的改進空間與應用可以發揮。幾個可能的研究方向包括：

1. 目前指派機率雖然在以辭典中的片語，加以訓練。其中的指派機率函數，假設和詞彙本身無關，只考慮片語的總字數，與詞彙位置。雖然如此，實驗證明還是能夠充分的反映詞彙的翻譯字數與位置的安排。然而不同詞性與語法結構的片語的翻譯的指派方式，有很大的差異。目前的做法，只考慮字數，未能考慮片語與詞彙的詞性。我們預計導入片語的句法的訊息，對於不同的名詞、動詞、形容詞片語，訓練不同的指派機率模型，可提供更精確的對應與翻譯的效果。在詞彙翻譯機率方面，以辭典中的片語的用字與翻譯，加以訓練，並不能反映詞彙正常的使用與翻譯的情形。我們預計以大型的語料庫，如光華雜誌漢英語料庫，來訓練詞彙翻譯機率，將可以得到精確的詞彙翻譯機率。
2. 應用片語翻譯模型於雙語語料庫的片語對應。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。詞彙對應的發現，不限於單字詞，而應及於多字的片語 [Kupiec 1993]。我們提出的新的統計式片語對應與翻譯的模型，可以在平行語料庫中擷取雙語的片語，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。利用新發展出來的模型，我們預計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，我們將擷取光華雜誌英中平行語料庫，進行互譯片語的擷取實驗。預計可以提出之統計式片語對應與翻譯模型，可獲得一般雙語辭典無法找到的片語與翻譯，如(graduate institute, 研究所), (cross-strait affair, 兩岸事務), (exclusive interview, 專訪)等。
3. 應用片語翻譯模型於跨語言檢索中的查詢詞翻譯。目前的跨語言檢索的研究，顯示通常會有一半以上的查詢片語，無法在雙語詞典中查到適當的翻譯。對於詞典沒有收錄的片語必須逐字翻譯，通常一字多譯，而逐字的翻譯只有部分和查詢主題相關。統計式的片語翻譯模型，可以由片語中透過對應分解出來的詞彙翻譯機率，比較傾向於文脈中的翻譯（translation in context），可以比雙語辭典提供更有效的翻譯。如 graduate 在 graduate student 的文脈下，可以對應到「研究」的翻譯，而雙語詞典的 graduate 詞條普遍的只有列出「畢業」的翻譯。由片語翻譯模型所提供的單字翻譯也比較豐富，如 nuclear 可以有「原子」、「核子」、「核」、「原」等等翻譯。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，可以產生的和文件庫相關的翻譯，提升跨語言檢索的效果。

致謝

本文之研究受到國科會編號 NSC 89-2420-H-007-001 計畫之補助。

參考文獻

BDC 1992 "The BDC Chinese-English electronic dictionary" (version 2.0), *Behavior Design Corporation*, Taiwan.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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

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

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.

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

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.

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

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.

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.

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

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

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

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.

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

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.

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.

Yamada, K. and K. Knight. 2001. "A Syntax-Based Statistical Translation Model". *In Proceedings of the Conference of the Association for Computational Linguistics* (ACL).