# 使用流暢性改善詞組翻譯的統計式機器翻譯

夏敏翔 張耀升 盧文祥

國立成功大學資訊工程研究所

**摘要**

Peter F. Brown等人提出統計式的機器翻譯後(Statistical Machine Translation)，目前翻譯的基本單位已由單詞轉變成詞組（Phrase），雖然詞組為本的機器翻譯已經可以達到不錯的效果，但使用者仍在詞組閱讀上遇到不順暢的情形，而目前研究機器翻譯的研究領域很少針對詞組流暢化進行探討。我們觀察到譯者在英中翻譯時，常常會加入一些不存在於英文句中的詞彙使句子能夠更加流暢，如果只是簡單的詞對詞翻譯，無法在中文句子顯示這些額外附加的詞彙在中文句子中。有鑑於此，我們提出流暢化詞組機器翻譯 (Fluent Phrase Machine Translation, FPMT）的機率模型來決定加入詞彙後的中文詞組是否流暢，以及使用語料庫(Corpus)和網路搜尋結果(Search Result)來找出附加的詞彙。實驗結果顯示，我們提出的流暢化詞組機器翻譯模型得到的中文詞組，其效能比IBM Model 4的方法佳，可以有效的補回缺少的詞彙，在人工評估上，更顯示我們的方法確實可以提升翻譯的流暢化。

## 1. 簡介

Brown[2]在1993年提出訊號源通道方法(Noisy-Channel Approach)來完成機器翻譯(Machine Translation)後，現在對於機器翻譯的研究幾乎都是使用統計式的方法，稱為統計式機器翻譯(Statistical Machine Translation)。現在有很多研究都是使用類似的方法，但是使用這類方法的翻譯模型卻有一些相同的問題—以詞為本(Word-based)，因此不管在重新排序或是翻譯階段的會遭遇困難。在重新排序方面，如果以詞為最小單位，對於多詞所組成的詞組，在重新排序後可能會發生錯誤[20]。同樣在翻譯方面，由多個詞所組成的詞組也容易被翻譯錯誤。目前有許多研究都使用詞組為翻譯單位的方法，實驗也證明以詞組為基本單位可以有效的提升結果[3][4][8][13][14][16][20][22][23]。然而，翻譯所得到的中文詞組雖然皆可對應於英文詞組，但是卻會造成翻譯的結果不是很流暢，所以為了增加詞組翻譯的流暢性，譯者都會加入一些不存在英文詞組中詞彙，由表一顯示，有加入額外的詞彙，詞組可以比較通順，所以機器翻譯不是只有單詞的翻譯，還必須補足詞彙使詞組或句子流暢。

機器翻譯有許多種方法，例如基於規則(Rule-based)的機器翻譯、基於實例(Example-based)的翻譯方法和統計式(Statistic-based)機器翻譯。其中，統計式機器翻譯為目前最常用的一種方法，目前有很多研究[14][11][13]，以詞組為本 (Phrase-based) 的方法得到的翻譯品質都比以詞為本的方法要來的好，而我們也將採用統計的方法來完成我們詞組的機器翻譯模型。大部分機器翻譯流程不外乎就是重新排序以及單詞翻譯，但是很少有學者研究翻譯之後的流暢化。翻譯後的句子不一定是通順的，因此在翻譯時，可能需要加入一些詞彙使得句子能更為流暢，所以我們嘗試將這類的字詞可以補回翻譯的結果中。雖然Brown已有提出類似方法，他們在翻譯過程產生一些空字串，這些空詞串可以產生一些不存在英文句中的翻譯詞彙，使翻譯後的結果流暢許多。但是我們覺得，大部份額外加入的詞彙應該是翻譯後決定比較合適，因此我們提出流暢化詞組的機器翻譯(Fluent Phrase-based Machine Translation, FPMT)，對翻譯的詞組再做進一步的詞彙增補，根據我們的實驗結果，我們提出的流暢化詞組機器翻譯模型可以有效的補回缺少的詞彙，其效能比IBM Model 4的方法好，在人工評估上，更顯示我們的方法確實可以提升翻譯的流暢化。

本論文第一節描述詞組的機器翻譯相關工作與一些問題。第二節將描述有關統計式機器翻譯的相關研究與文獻。第三節為我們所提出的詞組翻譯及兩種不同類型的流暢化方法，第四節透過實驗分析，我們提出的方法在詞組內部流暢化得到有效的改善，加入的詞彙確實比Brown提出的插入空字串來的好。第五節是結論以及未來研究方向。

表一、翻譯比較

| English Phrase / Translation Methods | asthma guidelines |
|---|---|
| Manual Translation | 氣喘**治療**指南 |
| Google (http://www.google.com/language_tools?hl=en) | 哮喘指南 |
| IBM (http://www-306.ibm.com/software/pervasive/tech/demos/translation.shtml) | 氣喘指南 |
| SYSTRAN (http://www.systransoft.com/index.html) | 哮喘指南 |
| MojoLingo (http://text.mojolingo.com.tw/) | 哮喘指導方針 |

## 2. 相關文獻討論

Brown [2]等人首先提出了統計式的機器翻譯方法，有效的從語料庫中獲得所需要的資訊，是目前翻譯研究中最常用的方法。Watanabe[20]等模仿Brown的機器翻譯步驟，但改變了排序時的最小單位，使用數個詞來當成做小單位，稱為詞塊 (Chunk)。Marcu和Wong[13]使用了聯合機率模型(Joint Probability Model)來取代條件機率，並且使語彙模型可以包含詞組的翻譯，使得翻譯上更為精確。Chiang [3]從相對應的詞組結構來學習翻譯的規則，使用正規語言的方法對來源語跟目標語同時處理以達到翻譯的目標。此外還有很多的研究[8][23]都使用詞組為單位，例如 Och [16]跟Yamada [22]等，證明了以詞組為翻譯單位可以有效的提升結果。所以現在的研究幾乎都是使用詞組為基本的單位，也證實了以單詞為基本單位是不夠的。

以語法結構為基礎的統計式機器翻譯跟前面的方法最大的不同是，這類方法會使用語法剖析器(Parser)來獲得語法的剖析樹(Parsing Tree)，再把剖析樹當成輸入進而完成翻譯的動作。Yamada and Knight [21][22]改變了一般的翻譯系統使用字串為輸入的方法，他們使用剖析器，將來源語的句子轉化成樹狀結構，藉此獲得更多的語言學上的資訊。訓練的過程中，獲得所需要的相關資訊並建立了三種表格，分別是重新排序機率表(Reordering Probability Table)、插入字串機率表(Insertion Probability Table)跟翻譯機率表(Translation Probability Table)。當輸入剖析樹時，然後對剖析樹執行重新排序、插入字串和翻譯，然後就輸出目標字串，翻譯的過程中查詢先前建立的三種表格。Ding和Palmer[6]同樣的，使用了文法剖析器來獲取樹狀結構。他們提出同步相依插入文法規則(Synchronous Dependency Insertion Grammars)來完成翻譯，也就是說，當他們對來源語的結構樹執行一個文法規則時，同時，根據相對應的文法規則來產生目標語的樹狀結構。翻譯過程所需要的文法規則，則是在翻譯前先從雙語語料庫中學習，學習的方法是使用他們所提出來的文法規則歸納演算法(Grammar Induction Algorithm)，此演算法簡單的說就是根據某一種語言的結構樹，然後利用詞典來分解相對應的翻譯樹獲得文法規則。以上兩篇都充分的顯示出，在翻譯的過程中使用了語法剖析器都可以提升翻譯的效能。

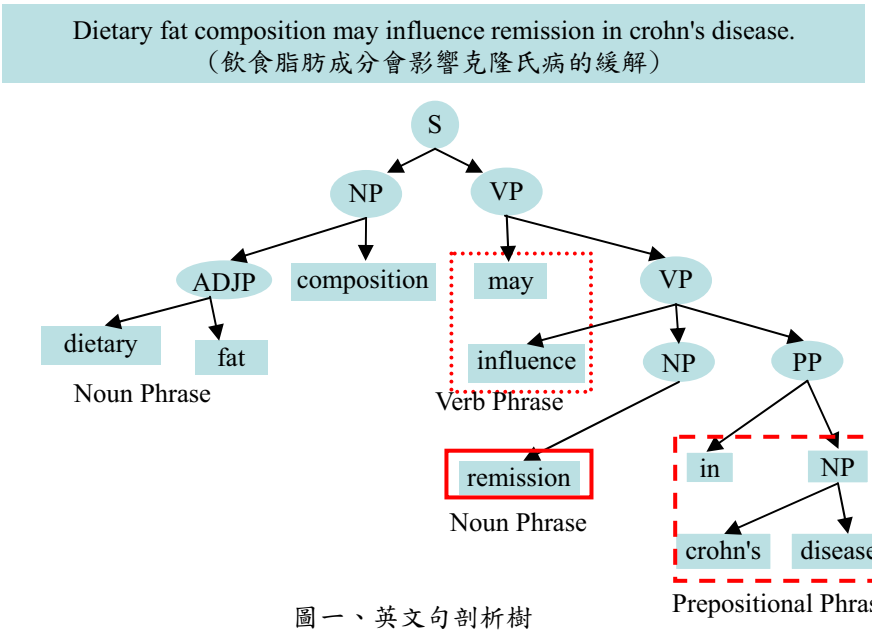
我們在本論文提出的方法也是統計式的機器翻譯，為了增加翻譯的品質，我們使用詞組為翻譯單位。Koehn[12] 提到，如果詞組結構裡包含的字數量太多時，翻譯的效果反而會下降，Och [16]跟Venugopal [19]都有不同的方法來獲得所需要的詞組配對，但是他們使用統計的方法來得到詞組，很可能會使詞組結構過大，因此我們必須找出另一種方法來獲得適當詞組結構。在以句法結構為基礎的統計式機器翻譯中顯示，如果使用剖析器來得到語法剖析樹當成輸入，可以增加翻譯的品質，因此我們使用剖析器來獲得句法剖析樹，再從句法剖析樹中獲得我們需要的詞組結構。我們也發現，如果使用此方法獲得的詞組結構，其包含的單詞數量並不會太多，很適合當成翻譯的基本單位。在重新排序方面，我們也是使用詞組為基本單位；在翻譯方面，也可以學習單詞在不同詞組時的不同翻譯，這樣可以提升翻譯的準確率。模型的推導方面，Brown等人使用貝氏定理將$P(T|S)$代換成$P(S|T)P(T)$，因為$P(T)$可以確保所得到的目標語言是可以符合文法的語句，但是Foster [7]提出直接計算$P(T|S)$的翻譯方式，而且在Och和Ney [15]的實驗結果也証實，如果直接計算最佳化的$P(T|S)$所得到的翻譯效果，會跟使用$P(S|T)P(T)$所得的效果差不多，甚至會更好，所以我們提出的統計式機器翻譯的模型，也是由$P(T|S)$開始推導的，在翻譯過程中使用詞組式語言模型來解決詞義消歧(Word Sense Disambiguation)的問題。本論文最重要的貢獻是提出流暢化詞組機器翻譯機率模型來解決詞組翻譯流暢化問題，現在很多研究都只有處理詞組翻譯，我

們覺得翻譯過後的結果還需要進一步的處理，也就是加入詞彙使結果更加流暢，不同於Brown提出的空字串產生方法，我們覺得應該是翻譯後才決定該加入哪些詞彙。詳細方法將會在第三節中描述。

## 3. 流暢化的詞組機器翻譯

### 3.1 詞組擷取

先前已經有很多的研究證實，使用詞組為單位的機器翻譯會比只使用單詞為單位的機器翻譯來的好，所以我們處理英文句的第一步就是必須先取得英文詞組，也就是要將輸入的英文句分割成詞組形式。我們使用史丹福英文語法剖析器（Stanford Parser）[18]來獲得需要的詞組。由圖一可以看見，我們使用一些觀察到的規則，可以從剖析樹（Parser Tree）中輕易的獲得該英文句的詞組，可以清楚看到有許多不同種類的詞組，例如名詞詞組、動詞詞組[1]和介係詞詞組。




圖一、英文句剖析樹

Ding and Palmer[6]提出使用不同語言的剖析器來得到兩棵結構不同的剖析樹，但是在中文方面，目前還缺乏良好的剖析器，而且不同語言的剖析樹因為結構上的差異，很難用相同的文法規則來處理不同結構的剖析樹，所以我們只使用剖析樹中的詞組資訊，並未使用其語法結構。

### 3.2 詞組翻譯

#### 3.2.1 問題描述

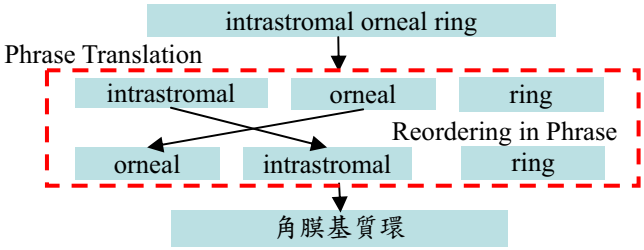
詞組翻譯過程中，並不只是把英文單詞翻譯成中文詞，還必須考慮單詞在詞組中的位置。因此在詞組翻譯的步驟中，我們首先需要完成詞組內的單詞的排序，然後根據所在的位置找出相對應的翻譯。我們知道英文單詞翻譯時，常常會有意義混淆（Ambiguity）現象，例如英文單字"bank"，可以翻譯成"銀行"也可以翻譯成"堤岸"，所以翻譯時必須決定哪一個翻譯才是正確的，也就是詞義消歧（Word Sense Disambiguation）問題，為了解決詞義消歧問題，我們使用詞組基本的語言模型(Language Model)，便可以根據前一個詞來決定最合適的翻譯。

在單詞翻譯前，我們必須先決定每一個單詞在該詞組的位置，由於一些中文詞在詞組中跟英文詞的順序有所不同，若按照原本的順序翻譯，常會發生詞組翻譯錯誤，以圖二為例，"intrastromal orneal ring"在翻譯成中文時，將"intrastromal"以及"orneal"互換後，才會是正確的翻譯，所以作單詞翻譯之前，我們必須先考慮到單詞在詞組中的位置，這樣翻譯成中文時，才會比較通順。當詞組中單詞的位置都決定後，最後步驟為將英文單詞翻成中文，但是並非把英文單詞

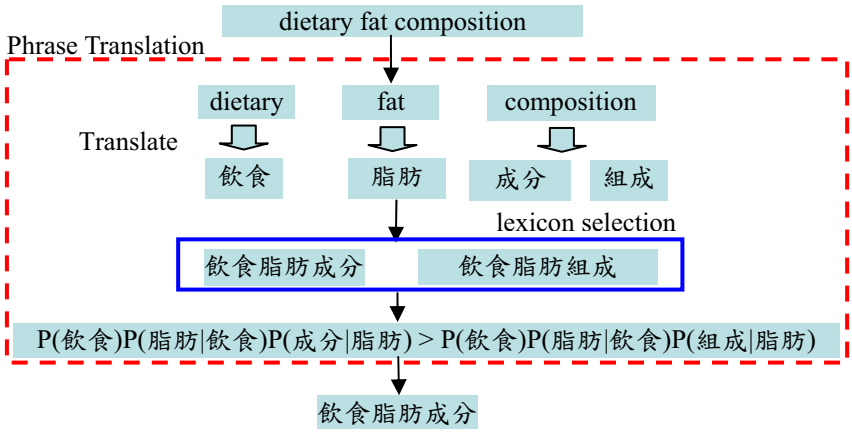
[1] 動詞詞組(Verb Phrase) 在本論文中僅考慮verb, adverb + verb, verb +adverb, 或auxiliary + verb 四種形式。

直接翻譯成中文即可，因為英文翻中文時，常常有意義不明確的情況發生，所以在翻譯時，只要選錯詞，將會造成很大的差異。下圖三，英文單字"composition"在文件中可以翻成"組成"以及"成分"，如果翻譯成"組成"，我們可以發現翻譯後的結果不正確，而為了解決這種詞彙選擇(Lexicon Selection)的問題，我們使用詞組的語言模型來決定該選哪一個詞彙來當成正確的翻譯。

由圖三可以看見，在我們的文件中，"脂肪成份"的機率比"脂肪組成"的機率還大，因此當英文單詞"composition"的前一個文字是"脂肪"時，我們便能知道它的翻譯應該是"成分"會比"組成"合適。所以我們只用簡單的語言模型方法，就可以解決很多的詞彙選擇問題。經過以上兩個步驟，我們便可以將英文詞組翻譯成中文詞組，除了詞與詞之間的順序都正確，還可以選擇最合適的詞彙來當成英文的翻譯。




圖二、詞組內單詞位置差異圖




圖三、單詞翻譯時之詞彙選擇問題

### 3.2.2 詞組翻譯模型

我們提出詞組的機器翻譯模型跟Brown最大的不同點，在於其使用貝氏定理$P(T|S)$代換成$P(S|T)P(T)$，而我們是直接由$P(T|S)$開始推導，所以我們初始的公式是：

$$C^* = \arg\max_c P(C \mid E) \tag{1}$$

我們使用的翻譯方法首先必須知道英文的詞組結構，所以引入參數$S$代表我們需要的詞組結構。引入$S$後，公式可以推導成：

$$C^* = \arg\max_c \sum_S P(S \mid E) P(C \mid S, E) \tag{2}$$

公式(2)中，我們透過史丹福英文語法剖析器（Stanford Parser）以獲得英文詞組結構，而假設此為最佳的一組詞組結構$S$，所以將$\sum$去除。獲得英文詞組結構後，接下來就是要將這些詞組結構重新排序，因此我們引入參數$R$表示每一個詞組根據中文結構應該重新排序的對應位置：

$$C^* = \arg\max_c P(S \mid E)P(C \mid S,E)$$
$$= \arg\max_c P(S \mid E)\sum_R P(R \mid S,E)P(C \mid R,S,E) \quad (3)$$

從公式(3)中，我們選擇最佳的一組$R$，所以將∑去除，我們的模型可以分成三個主要部份，分別為$P(S|E)$、$P(R|S,E)$以及$P(C|R,S,E)$，所以我們將機器翻譯模型可分成三大部份來解釋我們的翻譯模型，分別為詞組產生模型$P(S|E)$、重新排序模型$P(R|E,S)$以及詞組翻譯模型$P(C|R,S,E)$。本篇論文只針對詞組翻譯模型深入探討。

詞組翻譯時，以一個詞組為單位然後翻譯裡面的英文單詞，因此只考慮單詞在詞組中的適當位置，因為英文詞組翻譯成中文時，單詞的位置也會有所變化。詞組翻譯只是針對某一個詞組完成翻譯，並沒有考慮它在中文句子結構中的位置，而翻譯的結果也沒有受到前後詞組的影響，所以在完成詞組翻譯時，我們沒有考慮詞組的位置，根據公式(3)的詞組翻譯模型，我們可以將$R$省略，因此我們的詞組翻譯模型為：

$$P(C \mid R,S,E) = P(C \mid S,E) \quad (4)$$

不考慮位置$R$的情況下，詞組翻譯的機率即求$P(C \mid S,E)$，其中$C$是中文句、$E$是英文句，$S$是英文的詞組結構。我們發現，如果要從英文句$E$直接翻譯成中文句$C$是一件很困難的事，所以我們利用剖析器得到的詞組結構，將句子翻譯改變成詞組對詞組的翻譯並假設詞組的翻譯是獨立的，而非使用傳統語法結構的分析，故最後使用詞組翻譯如公式(5)：

$$P(C \mid S,E) \cong P(C \mid E)$$
$$= P(sc_1, sc_2, \cdots, sc_m \mid se_1, se_2, \cdots, se_m) \quad (5)$$
$$= \prod_{\substack{sc \in C \\ se \in E}} P(sc \mid se)$$

$sc$為組成中文句子$C$的中文詞組，$se$為輸入英文句子$E$至史丹福英文語法剖析器得到的英文詞組，最後句子之間的翻譯變成詞組之間的翻譯，當英文詞組都翻譯成中文詞組後，在獲得的詞組最佳排序，就可以完全翻譯成中文句。

假始直接使用$P(sc|se)$計算，所得到的中文詞組不一定是合適的，因為英文翻譯中文會有詞義消歧的問題，所以直接翻譯的結果並不一定是適當的中文詞組。透過貝氏定理轉換$P(sc|se)$，公式變成：

$$P(se \mid sc) = \frac{P(se \mid sc)P(sc)}{P(se)} \quad (6)$$

其中$P(sc)$為利用語言模型中的雙連詞(Bigram)機率來解決詞義消歧的問題，因此使用$\frac{P(se \mid sc)P(sc)}{P(se)}$可以確保得到的中文詞組是比較適當的。我們最終目的不是將英文詞組翻譯成中文詞組，而是選擇一個合適的中文詞組$sc$，以最接近我們輸入的英文詞組$se$。公式(6)中的$P(se)$為英文詞組機率，對於可能被選中的中文詞組而言，都是一樣的，所以我們將忽略公式中的$P(se)$。最後我們想要取得最佳的中文詞組$sc$，機率公式取推導成：

$$P(C \mid S,E) \propto \prod_{\substack{sc \in C \\ se \in E}} P(se \mid sc)P(sc) \quad (7)$$

接著我們定義如何使用$P(se|sc)$選擇出最佳的中文詞組$sc$。為了得到最佳英文詞組$se$的中文詞組$sc$翻譯，我們使用兩個機率公式進行推導。（一）單詞翻譯機率$P(se_{a_i} \mid sc_i)$，其中$sc_i$是中文詞組$sc$中的第$i$個中文詞，$se_{a_i}$則是$sc_i$在對應位置$a$的情況下所對應到的英文詞，以及（二）位置對應機率$P(a \mid l,m)$，其中$l$是中文詞組的長度，$m$是英文詞組的長度，$a$就是在該長度下，中文與英文對應的關係。根據上面的敘述，我們可以將$P(se|sc)$推導成：

$$P(se \mid sc) = P(a_1 \ldots a_l \mid l)\prod_i P(se_{a_i} \mid sc_i) \quad (8)$$

因為中文詞組是由英文詞組加上詞典得到的，所以英文詞組與中文詞組中的詞是屬於一對一

的關係，也就是說英文詞組的長度跟中文詞組的長度是相同的，因此我們只使用參數$l$來代表詞組的長度。根據上述，我們的詞組翻譯模型就可以推導成：

$$P(C \mid S,E) \propto \prod_{\substack{sc\in C \\ se\in E}} \{P(a_1 \ldots a_l \mid l) \prod_i P(se_{a_i} \mid sc_i) P(sc)\} \tag{9}$$

我們以英文詞組< intrastromal corneal ring>要翻譯成中文詞組<角膜基質環>為例，最後詞組翻譯模型如下：

$$\begin{aligned} & P(se \mid sc)P(sc) \\ &= P(\text{intrastromal corneal ring} \mid \text{角膜基質環}) \times P(\text{角膜基質環}) \\ &= P(213 \mid 3)P(\text{corneal} \mid \text{角膜}) \times P(\text{intrastromal} \mid \text{基質}) \times P(\text{ring} \mid \text{環}) \\ &\quad \times P(\text{角膜})P(\text{基質} \mid \text{角膜}) \times P(\text{環} \mid \text{基質}) \end{aligned}$$

英文詞corneal透過詞典得到某一中文翻譯詞為<角膜>，從之前的訓練樣本中可以計算得到$P$(corneal |角膜)的機率值，依此類推皆可得到$P$(intrastromal |基質)及$P$(ring |環)。因為<角膜>、<基質>及<環>中文字詞，依不同的排列組合而有不同的中文詞組翻譯，故我們計算各中文字詞組合的位置機率，其機率值在訓練樣本已有紀錄，最後再透過雙連詞(Bigram)機率加以計算此中文翻譯詞組是否存在的機率。

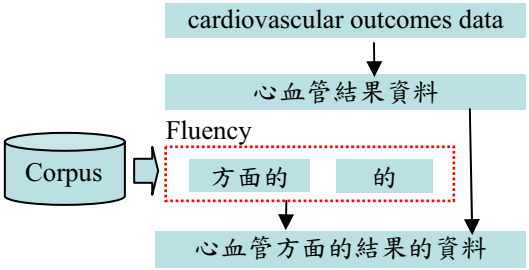
### 3.3 流暢化

#### 3.3.1 問題描述

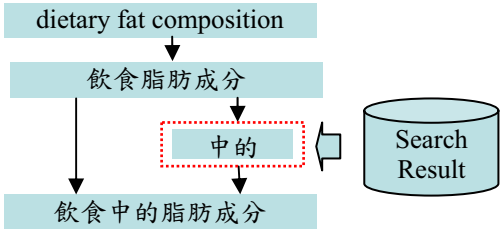
經過3.2詞組翻譯模型處理，可以獲得中文翻譯的詞組，此中文詞組為英文詞組中的單詞重新排序、個別翻譯得之，基本上都可以大略了解此中文詞組翻譯。但是若可以在中文詞組再加入一些詞彙，便可以使得翻譯後的中文詞組更容易閱讀。英文翻譯成中文的過程中，譯者通常會加入一些額外詞彙使得閱讀時更為流暢，然而這些詞彙一般並不會出現英文詞組中，而是為了詞組的流暢性才加入，所以機器翻譯不僅僅是要將詞彙翻譯正確，也要使翻譯出來的詞組可以讓讀者閱讀流暢，所以我們在完成機器翻譯後提出了一種流暢化的方法，嘗試將不存在英文詞組中的字詞，可以順利地加入翻譯後的中文詞組中。所以在本論文我們首先提出中文詞組內部附加詞方法來解決翻譯流暢化問題。

我們使用兩種詞組內部的附加詞方法：語料庫學習以及網路搜尋結果學習。使用語料庫學習方面，由　　圖四例子可以看見，英文詞組"cardiovascular outcomes data"如果按照字詞翻譯所獲得的中文詞組為"心血管結果資料"，此詞組雖然可以知道其所要表達的意義，若可以再加上額外的字詞，如"方面"以及"的"，則"心血管方面的結果的資料"比起"心血管結果資料"更容易使人了解其意義。

然而，使用語料庫學習附加詞有其缺陷，由於語料庫大小的限制，只要語料庫中沒學過的詞彙我們將無法找出合適的附加詞，為了補足此缺陷，我們利用網路搜尋結果來補強。圖五的英文詞組"dietary fat composition"翻譯成中文詞組為"飲食脂肪成分"，當使用語料庫時，我們發現並沒有適合的詞可以加入，因此利用網路資源嘗試找出附加詞。使用網路搜尋結果，我們發現在"飲食"以及"脂肪"之間可以加入中文字詞"中的"，而且加入後的詞組"飲食中的脂肪成分"比未加入的詞組"飲食脂肪成分"更為流暢。




圖四、詞組內加入詞彙(使用語料庫)



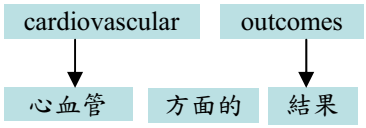

圖五、詞組內加入詞彙(使用網路資源)

### 3.3.2 流暢化方法

下面將介紹我們利用機率概念推導兩個詞組內部的流暢化計算方法。在使用語料庫方面，我們提出一個簡單的機率公式，利用邏輯變數$a$檢查可能附加的流暢詞$w$是否允許加入於中文詞與詞之間$(c_i, c_{i-1})$，公式如下：

$$P(a \mid c_i, c_{i-1}) = \sum_w P(a \mid c_i, c_{i-1}, w) \times P(w \mid c_i, c_{i-1}) \tag{10}$$

$P(a|c_i, c_{i-1}, w)$計算兩詞之間$c_i$、$c_{i-1}$可以插入附加詞$w$的機率，$P(w|c_i, c_{i-1})$決定兩詞之間$c_i$、$c_{i-1}$是否可以加入一個附加詞$w$。由圖六中為某訓練樣本，我們可以訓練到cardiovascular及outcomes可加入的附加詞<方面的>之機率值，即 $P(\text{方面的} \mid outcomes, cardiovascular)$ 。因為是由英文翻譯中文，故由英文反推找出附加詞會比較符合原本的涵意，再加上流暢化步驟在翻譯之後，所以每一中文字詞$c_i$即對應一英文詞$e_i$，故$P(e_i|c_i)= 1$，最後公式推導如：




圖六、附加詞訓練樣本範例

$$P(a \mid c_i, c_{i-1}) = P(a \mid c_i, c_{i-1}, w) \times \sum_w P(w \mid e_i, e_{i-1}) \tag{11}$$

$P(a|c_i, c_{i-1})$就可以使用$P(w|e_i, e_{i-1})$以及$P(a|c_i, c_{i-1}, w)$來計算，然後使用門檻值來決定是否需要加入額外的詞彙。$P(w|e_i, e_{i-1})$用來計算在語料庫中有哪一些詞彙常常出現在單詞$c_i, c_{i-1}$之間；$P(a|c_i, c_{i-1}, w)$用來計算在$c_i, c_{i-1}$之間加入$w$後的出現機率，在計算$P(a|c_i, c_{i-1}, w)$時，我們是利用網路搜尋引擎Google的相關網頁搜尋數目來估計，估計的方法如下：

$$\begin{aligned} P(a \mid c_i, c_{i-1}, w) &= \frac{P(a, c_i, c_{i-1}, w)}{P(c_i, c_{i-1}, w)} = \frac{count(a, c_i, c_{i-1}, w) \big/ N}{count(c_i, c_{i-1}, w) \big/ N} \\ &= \frac{count(a, c_i, c_{i-1}, w)}{count(c_i, c_{i-1}, w)} \end{aligned} \tag{12}$$

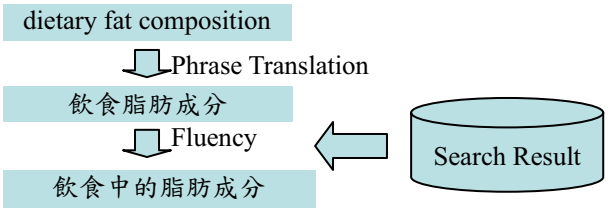
其中$N$代表Google的所有網頁數目，$count(a, c_i, c_{i-1}, w)$代表將$w$加入$c_i$、$c_{i-1}$中間所形成相鄰的字串"$c_i w c_{i-1}$"(即tri-gram)可以搜尋到的網頁筆數，$count(c_i, c_{i-1}, w)$則是$c_i$、$c_{i-1}$以及$w$三個詞同時出現的網頁筆數，故我們可以得到出現$c_i$、$c_{i-1}$以及$w$三個詞的網頁中其相鄰字串tri-gram所佔的比例。以英文詞組"cardiovascular outcomes data"為例子，我們翻譯出的中文詞組為"心血管 結果 資料"，使用公式(11)計算"心血管"以及"結果"之中是否可加入附加詞：

$$\begin{aligned} P(a \mid \text{結果,心血管}) &= P(\text{方面的} \mid outcomes, cardiovasaular) \times P(a \mid \text{結果,心血管,方面的}) \\ &= P(\text{方面的} \mid outcomes, cardiovasaular) \times \frac{count(\text{"心血管方面的結果"})}{count(\text{"心血管","方面的","結果"})} \end{aligned}$$

當 $P(a \mid \text{結果,心血管})$ 的值大於門檻值時，我們就將附加詞"方面的"加入在"心血管"以及"結果"之中，同理，"結果"以及"資料"中間也可以加入"的"，加入附加詞後的中文詞組為"心血管方面的結果的資料"，比起未加入詞彙時更為流暢。

在使用網路資源方面，因為語料庫的大小限制，我們無法確實的將詞與詞之間的附加詞都找出來，所以我們借用網路上的搜尋結果來幫助我們取得這些附加詞彙。圖七中的英文詞組"dietary fat composition"翻譯成中文後為"飲食脂肪成分"，使用網路搜尋結果可以發現"飲食"以及"脂肪"之間會加入"中的"，則中文詞組"飲食中的脂肪成分"將較為通順，但是在我們的語料庫中"飲食"以及"脂肪"之間並沒有任何辭彙，如果使用語料庫的方法將無法加入任何辭彙，所以我們另外使用網路搜尋結果來幫我們找出合適的辭彙。

圖八是使用網路搜尋結果流暢化步驟的流程圖，最主要的目的就是要找出兩個關鍵中文詞（$C_1$,$C_2$）中間可以加入哪些附加詞，使得看起來更為通順。首先我們先把兩個中文關鍵詞送到Google搜尋引擎，取回搜尋結果，然後使用Chien[4]所提出的PAT-Tree-based 的關鍵詞擷取方法找出最常出現的詞彙，然後可以根據這些詞彙再使用關鍵詞找出候選附加詞，再將雙方的候選詞一起送至Google搜尋引擎，算出他的頻率，這樣我們就可以知道在關鍵詞之間，可以加上哪一個附加詞彙可以使他們更為順暢。

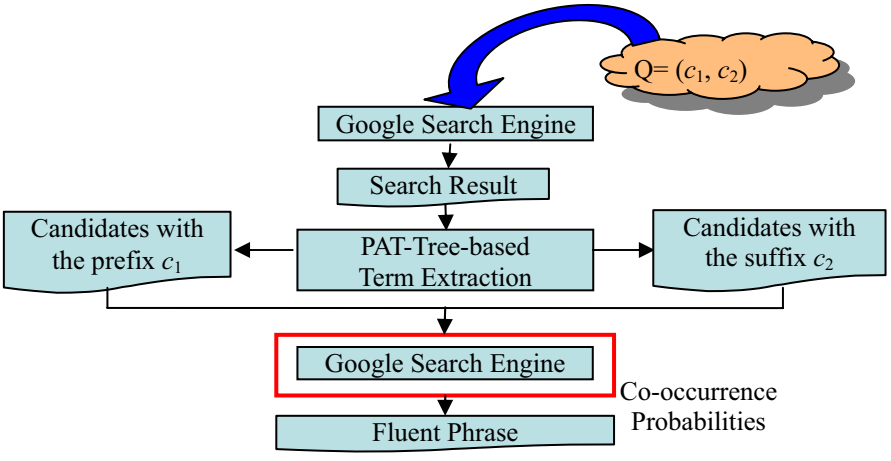



圖七、使用網路搜尋結果加入詞彙

我們使用中文的關鍵字"飲食"( $c_1$)以及"脂肪"( $c_2$)作為例子，首先我們使用關鍵字"飲食 脂肪"送至Google取回搜尋結果，然將搜尋結果使用PAT-Tree based關鍵詞擷取方法找出頻率較高的中文詞，再利用關鍵詞"飲食"以及"脂肪"從這些中文詞中找出相關的詞，例如"飲食"( $c_1$)可以找出產生前綴詞彙$c_1$的連接詞串"飲食文化"、"飲食中"...等的可能附加候選詞($x_1$)如"文化"、"中"，而"脂肪"( $c_2$)可以找出產生前綴詞彙$c_2$的連接詞串"飽和脂肪"、"的脂肪"...等的可能附加候選詞($x_2$)如"飽和"、"的"。然後將$x_1$, $x_2$連接在$c_1$以及$c_2$內產生一新詞串，使用公式(13)來計算他們配對後新詞串的機率，如此一來我們就可以知道在"飲食"以及"脂肪"中間可以插入"中" ($x_1$) "的" ($x_2$)形成"飲食中的脂肪"一詞就較為通順。我們使用的公式如下：

$$P(x_1,x_2,c_1,c_2)=\frac{P(c_1,x_1,x_2,c_2)}{P(c_1,c_2)}=\frac{count(c_1,x_1,x_2,c_2)}{count(c_1,c_2)} \tag{13}$$

其中我們限制附加詞$x_1$和$x_2$就是夾在$c_1$以及$c_2$ 之間的辭彙，而$x_1$和$x_2$可能會有兩種不同的情況，當$x_1$以及$x_2$相同時也就是$x_1$以及$x_2$是重疊（overlap）的狀況，我們就只考慮$x_1$即可；若$x_1$和$x_2$不同時，將$x_1$以及$x_2$連接（concatenate）成一個詞彙，再使用與公式(12)相同的方法，我們也是使用Google網路搜尋引擎所找到的網頁數來估算機率值，然後根據機率值大小再決定是否需要加入詞彙$x_1$和$x_2$。




圖八、使用網路搜尋結果流暢化流程圖

## 4. 實驗

本章節將評估詞組翻譯和流暢化的效能，並且比較我們的方法和IBM Model 4（簡寫IBM4）翻譯模型的結果。首先介紹我們所使用的資料、比較的翻譯模型以及評估方式，接下來就是我們的實驗數據分析。

### 4.1 實驗資料

我們的訓練語料庫是從國際厚生健康園區網站[1]經由人工收集的雙語語料庫，總共有18752句中英文配對句子，所包含的名詞詞組、動詞詞組以及介系詞詞組的相關數目如表二所表示。訓練翻譯模型前，我們使用GIZA++[9]工具進行訓練，以得到中英文詞組對應的統計式詞典檔。測試資料分為兩部份，分別是內部測試(Inside-test)，其為使用訓練語料庫中的資料做測試；以及外部測試(Outside-test)，此為使用非訓練語料庫內的資料來測試。測試方法為詞組翻譯。測試資料皆為隨機挑選，使用數目分別如表三和表四所列。表四中的外部測試資料各類型詞組是從100句英文中所得來的，本研究是我們在機器翻譯的起步研究，因為準備外部測試資料需要時間以人工方式進行中英文句子的對應，所以目前尚未取得大量的測試資料。

首先我們使用ISI(Information Sciences Institute)自然語言處理小組[10]所提供的解碼器對IBM4進行詞組翻譯評估。評估的方法則使用BLEU[17]，此為一種使用N-Gram的方式來評估機器翻譯的結果效能，使用自動評估的方法比人工評估更加快速方便，由於目前沒有針對流暢化翻譯的自動評估方法，我們則利用人工評分的方式，加以分析。

表二、訓練語料庫中詞組類型及數目

| | Number |
|---|---|
| Noun Phrase | 20820 |
| Verb Phrase | 10603 |
| Prepositional Phrase | 20421 |

表三、內部測試資料中詞組類型及數目

| | Number |
|---|---|
| Noun Phrase | 1000 |
| Verb Phrase | 1000 |
| Prepositional Phrase | 1000 |

表四、外部測試資料中詞組類型及數目

| | Number |
|---|---|
| Noun Phrase | 170 |
| Verb Phrase | 97 |
| Prepositional Phrase | 123 |

### 4.2 實驗結果

詞組翻譯的實驗的結果分為內部資料測試以及外部資料測試。

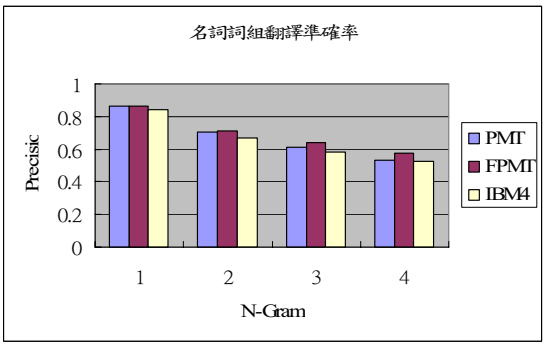
#### 4.2.1 內部測試

我們使用表三中提到的內部資料來做測試。圖九是名詞詞組的N-Gram Precision比較圖，圖十、圖十一分別是動詞詞組以及介系詞詞組的比較圖。PMT為我們的詞組機器翻譯方法，FPMT為我們提出的流暢化詞組機器翻譯方法。圖九是以名詞詞組為測試資料的實驗結果，在1-Gram Precision的數值幾乎一樣，所以我們加入的詞彙可以將英文句子所沒出現的單詞補回中文詞組，而加入這類詞彙使得中文詞組更加順暢，由於其並沒有使準確率下降，故所加入的詞彙幾乎都是正確的。在動詞詞組以及介系詞詞組中，準確率的差別比較小，我們發現在名詞詞組中，需要加詞的情況比動詞詞組以及介系詞詞組來的多，所以4-Gram Precision的數值表現最好，乃因加入附加詞的位置是正確。

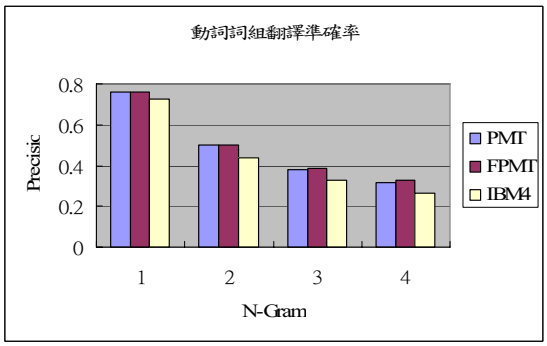
接下來我們為FPMT跟IBM4的分析比較。從圖九可以發現，在名詞詞組中我們所使用的詞組翻譯模型所得到的結果會比IBM4所提出的方法好。從表五的名詞翻譯比較，在單詞翻譯方面，FPMT可以翻譯出較合適的中文詞，"pediatric"在此詞組中，翻譯成"兒童"確實比"小兒科"來的好，而FPMT同時也加入詞彙"的"，使的詞組翻譯除了正確且更加流暢；而表六也顯示，在名詞

詞組"prospective randomized studies"使用FPMT可以正確的將詞彙"的"加入到正確位子，完全符合參考答案，雖然IBM4的系統可以補回詞彙，但是很明顯的位置發生錯誤。

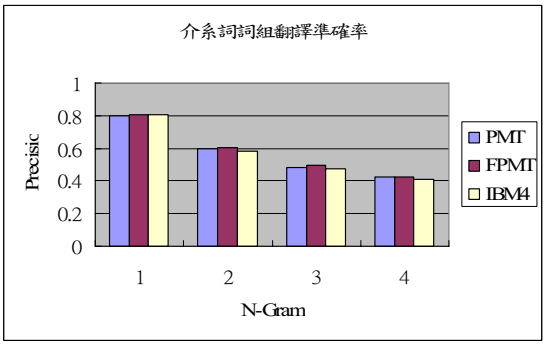
1-Gram-Precision效能評估顯示我們的FPMT翻譯方法在單詞翻譯的準確率確實比IBM4的表現較佳，而其他N-Gram-Precision，除了詞組內部單詞排序的結果表現較佳外，而加入流暢化方法的插入附加詞步驟之後，準確率更有些許的提升。從圖十一中可以發現，介系詞詞組的準確率並沒有提升很多，這是因為在翻譯介系詞詞組時，由於我們的方法一開始便將介系詞視為虛詞(Stopword)處理，所以在詞組翻譯過程中，介系詞會被省略，最後在作流暢化時，會將介系詞當成是詞組與詞組之間的附加詞，而IBM4會將介系詞直接翻譯，如表七所示，導致我們的方法在介系詞詞組的效果較不理想。圖十長度1至4的動詞詞組準確率的差距較大，平均而言FPMT效能較好，我們進一步分析，根據3.1節定義的動詞詞組，其長度不會比名詞詞組或介係詞詞組長，所以能形成的4-Gram數量較少。但是整體而言，翻譯以及排序的結果都可以比IBM4的好。




圖九、名詞詞組翻譯比較圖(內部資料測試)




圖十、動詞詞組翻譯比較圖(內部資料測試)




圖十一、介系詞詞組翻譯比較圖(內部資料測試)

表五、名詞詞組翻譯比較(1)

| English Phrase / Translation Method | pediatric mental health problems |
|---|---|
| Reference Translation | 兒童 的 精神 健康 問題 |
| FPMT | 兒童 的 精神 健康 問題 |
| IBM4 | 小兒科 精神 健康 問題 |

表六、名詞詞組翻譯比較(2)

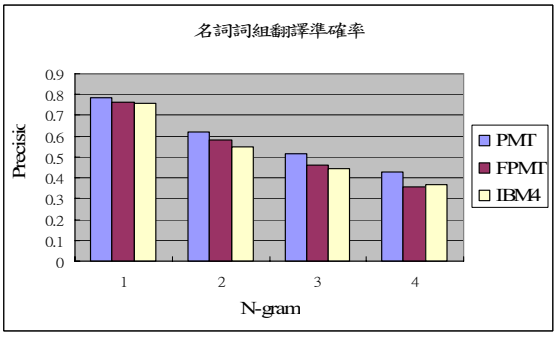
| English Phrase / Translation Method | Prospective randomized studies |
|---|---|
| Reference Translation | 前瞻性 的 隨機 研究 |
| FPMT | 前瞻性 的 隨機 研究 |
| IBM4 | 的 前瞻性 隨機 研究 |

七、介系詞詞組翻譯比較表

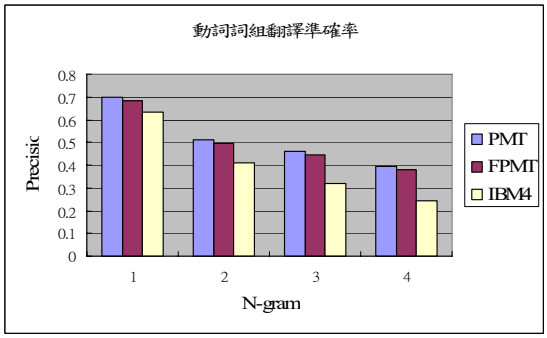
| English Phrase / Translation Method | In the polyp prevention trial |
|---|---|
| Reference Translation | 在 息肉 預防 試驗 |
| FPMT | 息肉 預防 試驗 |
| IBM4 | 在 息肉 預防 試驗 |

4.2.2 外部測試

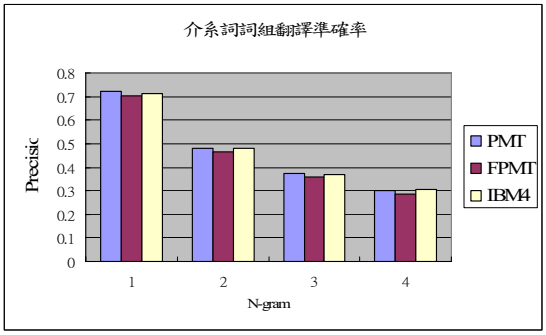
這一小節我們從訓練資料外的文件找出100句中英文配對的句子，並分割其詞組，此分割擷取到的詞組來當成我們的外部測試資料，其資料筆數如表四所示。圖十二是名詞詞組的評估結果，我們發現PMT優於IBM4，而FPMT卻不及IBM4的表現佳，原因有可能是加入的詞彙破壞了原本N-Gram的結構，使得準確率下降，並非表示加入錯誤詞彙，而BLEU是一種以N-gram字串比對的計分公式，當我們多加入的詞彙可能使詞組流暢，但是詞彙如果不在參考答案中出現，反而會使得準確率降低，這也說明BLEU可能無法評比加入詞彙的句子是否流暢。我們將在4.3節針對流暢化的問題予以討論。圖十三為動詞詞組的比較結果，從圖中可得知我們使用的詞組翻譯方法在動詞詞組也有不錯的效果。在動詞詞組領先的幅度比較大，因根據本論文定義的動詞詞組長度一般比較短，所以達到4-Gram的詞組數量非常的少，如此只要有一兩句的4-Gram翻譯的比較好，即可能將差距拉大。在名詞詞組以及介系詞詞組的結果就不會有很大的差距，因為名詞詞組以及介系詞詞組的平均長度皆比動詞詞組長，故可得到的4-Gram數量會比較多，不會因少數的翻譯結果導致影響整體評分。圖十四是介系詞詞組的評估結果，我們的詞組翻譯結果幾乎跟IBM4的結果相同，主要是因為我們的詞組翻譯方法並沒有將介系詞翻譯，例如介系詞"of"可以翻成中文"的"以及"in"翻成"在"等等，而IBM4的方法有翻譯介系詞。雖然我們的方法缺少介系詞的翻譯，但是單詞翻譯以及單詞排序卻比IBM4佳，當在缺少介系詞的情況下，翻譯結果幾乎跟IBM4的相同，若將缺少的介系詞翻譯補回，我們FPMT的表現應當可超越IBM4。




圖十二、名詞詞組翻譯比較圖(外部資料測試)




圖十三、動詞詞組翻譯比較圖(外部資料測試)




圖十四、介系詞詞組翻譯比較圖(外部資料測試)

### 4.3 實驗討論

從4.2節數據分析，無論在內部測試還是外部測試翻譯效果都有些許的改進，但是在外部測試時，由於BLEU無法評估加入詞彙後的詞組是否通順，導致流暢化的結果使得分數下降，而我們流暢化方法的門檻值並沒有嚴格設定，也造成加入過多的辭彙，這也是分數下降的原因之一。[12]提到，BLEU的參考答案至少要有三組以上才合理，因為英文翻譯成中文時，常常會有語意相同詞不同的情況，例如外部測試的詞組中有一句是<after 16 weeks>，在參考答案中翻譯成<16星期後>，我們的系統翻譯成<16週後>，其實這兩者翻譯結果是屬於同義的，但使用BLEU評分，這種情形卻是錯誤的。

至於詞組內部流暢化的結果，表八是我們使用詞組內部流暢化FPMT方法與IBM4的比較結果，在1-Gram Precision以及2-Gram Precision超越IBM4，其意味就單詞翻譯而言，我們提出的方法比較有效，但是在3-Gram Precision和4-Gram Precision的結果卻不理想，造成的原因有可能為加入的辭彙，破壞了原本3-Gram及4-Gram的結構，所以分數才會下降。觀察翻譯結果後發現，加入的詞彙確實可以使句子更加流暢，以表九說明，詞組英文"diabetes risk"，在參考答案中翻譯成"糖尿病危險"，我們的系統翻譯而且加詞彙後翻譯成為"糖尿病的風險"，若先忽略同義詞的問題，發現加入"的"並不會使詞組更混亂。另一個例子也相同，"no significant differences"參考答案為"沒有顯著差別"，我們的系統答案為"沒有任何顯著差異"，加入詞彙"任何"也使得翻譯通順。

由於加入的詞彙可使得詞組順暢，但這些附加詞幾乎都是參考答案中沒有的，如果將此類詞彙加入參考答案中，而不考慮同義詞的問題，評分結果如表十，將來若把同義詞問題解決，相信應當會有更佳的表現。表十一乃將附加詞彙列入參考答案之後，進行BLEU評估，數據顯示我們提出的方法較IBM4佳。

表八、詞組內部流暢化結果比較(1)

| | 1-Gram-Precision | 2-Gram-Precision | 3-Gram-Precision | 4-Gram-Precision |
|---|---|---|---|---|
| FPMT | 0.6824 | 0.4831 | 0.3530 | 0.2605 |
| IBM4 | 0.6803 | 0.4821 | 0.3595 | 0.2734 |

表九、詞組內部流暢化例子

| | diabetes risk | no significant differences |
|---|---|---|
| FPMT | 糖尿病的風險 | 沒有任何顯著差異 |
| Reference Translation | 糖尿病危險 | 沒有顯著差別 |

表十、將附加詞彙列入參考答案之詞組內部流暢化結果比較

| | 1-Gram-Precision | 2-Gram-Precision | 3-Gram-Precision | 4-Gram-Precision |
|---|---|---|---|---|
| FPMT | 0.6870 | 0.4916 | 0.3656 | 0.2757 |
| IBM4 | 0.6803 | 0.4821 | 0.3595 | 0.2734 |

表十一、BLEU 比較(詞組內部流暢化)

| | BLEU |
|---|---|
| FPMT | 0.3638 |
| IBM4 | 0.3407 |

由上面的實驗可發現，在外部資料測試中，當我們使用的參考答案並沒有解決同義詞的問題時，我們的流暢化方法可以達到一些效果，若同義詞問題可以解決，準確率應當會再提升。由於使用BLEU並不能斷定翻譯結果的好壞，因此我們使用人工評估來評定翻譯的準確率。我們從實驗資料中在隨機抽出100句的名詞詞組、動詞詞組、介系詞詞組並找五位使用者來進行人工評估，將FPMT及IBM4所產生的詞組翻譯，以不固定順序方式隨機排列兩個方法得到的翻譯結果，如此

可避免使用者猜出所列出的翻譯為何種方法產生，以提高公信力。表十二為各詞組的翻譯準確率，表中的正確(Correct)是代表翻譯正確，可接受(Acceptable)則為翻譯結果並非很合適或是有缺詞的情況，但結果仍可以表達其翻譯意義。表十二可以清楚得知，各類型詞組的接受率都可以比IBM4來的好，所以對使用者而言，我們的翻譯方法確實可以達到較好的翻譯品質。表十二可以清楚得知，各類型詞組的接受率都可以比IBM4來的好，所以對使用者而言，我們的翻譯方法確實可以達到較好的翻譯品質。

表十二、人工評估翻譯結果

| | | Correct | Acceptable |
|---|---|---|---|
| Noun Phrase | FPMT | 59% | 80% |
| | IBM4 | 50% | 70% |
| Verb Phrase | FPMT | 41% | 70% |
| | IBM4 | 24% | 52% |
| Prepositional Phrase | FPMT | 38% | 65% |
| | IBM4 | 37% | 60% |

從以上的各種實驗數據分析，不論是使用BLEU自動化評估或是人工評估，在各類型的詞組，我們提出的方法皆優於IBM4。最後詞組翻譯流暢度的比較，我們從實驗題目中，取出28題有流暢化的題目進行人工評估，表十三顯示詞組有流暢化的翻譯結果較容易被使用者所接受，其說明我們所加入的詞彙確實可以讓使用者更容易閱讀。

表十三、翻譯流暢度人工評估比較

| | | Degree of Fluency |
|---|---|---|
| 28 Test Phrases | FPMT | 62% |
| | IBM4 | 38% |

論文的附錄為測試題目中，流暢化時所加入的詞彙。附錄A為名詞詞組所加入的詞彙，我們發現比較特別的是許多中文名詞詞組應該加入量詞才會通順，一般英文名詞詞組沒有量詞，但是中文翻譯時，在數字後加上量詞才能顯示名詞詞組的完整性以及流暢度，如果只是按照單詞翻譯，所得的結果並不是最正確的。附錄中的資料顯示詞彙大部份加入次數都為1，但這並不表示加入的詞彙是錯誤的，由於測試題目都是隨機選取，所以很多附加詞彙的加入次數只有1次。從附錄中也可以看見加入的詞彙有一些多詞所組成的詞組，這些幾乎都是因為詞典的翻譯錯誤導致的。由於我們使用的詞典是GIZA++產生的，所以詞典裡的翻譯不完全都是正確的，品質不好的詞典會造成學習附加詞彙的錯誤。由於我們的語料庫太小訓練並不足夠，所以很多附加的詞彙可能沒辦法學習到，因此我們未來會收集更多雙語語料加強語料庫的訓練來提升我們翻譯系統的流暢化效能。

## 5. 結論

本論文針對詞組翻譯部分、單詞位置的不同以及詞義消歧的問題，我們提出流暢化詞組機器翻譯模型，除了改善翻譯的效能，並透過平行語料庫及網路搜尋結果，以提升中文翻譯的流暢度。雖然外部資料實驗數據並沒有提升流暢化後的結果，經由分析其原因，主要為BLEU並不能有效地評估句子的流暢度，所以我們認為句子翻譯品質的好壞，並不適合用N-gram模型加以評估，所以我們透過人工的評估，證明我們的方法確實可以提升翻譯的流暢化，更重要的是翻譯結果能讓使用者可以瞭解詞句表達的涵意。

目前自動的評分方式皆無法有效的評估句子是否流暢，當詞組或句子加入附加詞彙時，往往只會使翻譯效能評估下降，就算加入合適的詞彙還是無法獲得高分數，因此我們希望能找出更好的評分方法，不是單以詞彙來決定分數，仍需考慮結構以及流暢度等問題。

## 6. 參考文獻


[1] 國際厚生健康園區，http://www.24drs.com/professional/

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

[3] J. S. Chang, D. Yu and C. J. Lee. 2001. Statistical Translation Model for Phrases. *Computational Linguistics and Chinese Language Processing*, 6(2) , pp.43-64.

[4] D. Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 263-270.

[5] L. F. Chien, T. I. Huang and M. C. Chen. 1997. PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval. *Proceedings of the 20th Annual International Conference on Research and Development in Information Retrieval* (*SIGIR*), pp. 50-59.

[6] Y. Ding and M. Palmer. 2005. Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.541-548.

[7] G. Foster. 2000. A Maximum Entropy Minimum Divergence Translation Model. *Proceedings of the 38rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 45-52.

[8] H. J. Fox. 2002. Phrasal Cohesion and Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 304-311.

[9] GIZA++, http://www.fjoch.com/GIZA++.html

[10] ISI(Information Sciences Institute), http://www.isi.edu/publications/licensed-sw/rewrite-decoder/index.html

[11] D. Klein and C. D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 423-430.

[12] P. Koehn, F. J. Och and D. Marcu. 2003. Statistical phrase-based translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 127-133.

[13] D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 133-139.

[14] F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4): 417-449.

[15] F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 295-302.

[16] F. J. Och, C. Tillmann and H. Ney. 1999. Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (*EMNLP/VLC*), pp. 20-28.

[17] K. Papineni, S. Roukos, T. Ward and W. J. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 311-318.

[18] Stanford Parser Parse Visualization Tool, http://ai.stanford.edu/~rion/parsing/stanford_viz.html

[19] A. Venugopal, S. Vogel and A. Waibel. 2003. Effective Phrase Translation Extraction from Alignment Models. *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 319-326.

[20] T. Watanabe, E. Sumita and H. G.. Okuno. 2003. Chunk-based Statistical Translation. *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 303-310.

[21] K. Yamada and K. Knight. 2002. A Decoder for Syntax-based Statistical MT. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 303-310.

[22] K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.523-530.

[23] R. Zens and H. Ney. 2004. Improvements in Phrase-Based Statistical Machine Translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 257-264.

# 附錄

## 附錄 A 名詞詞組附加詞彙

| 加入詞彙 | 次數 | 加入詞彙 | 次數 | 加入詞彙 | 次數 |
|---|---|---|---|---|---|
| 的 | 124 | 進行 | 1 | 律當這樣的 | 1 |
| 名 | 8 | 則有 | 1 | 充足的 | 1 |
| 種 | 5 | 第 2 | 1 | vcu 的 | 1 |
| 有 | 5 | 餵養 | 1 | he | 1 |
| 個 | 4 | 了 | 1 | 的有老年 | 1 |
| 位 | 4 | 住院 | 1 | 進行評量由 | 1 |
| 的長期追蹤 | 3 | 二 | 1 | 外科手術的 | 1 |
| 起 | 3 | 篇 | 1 | 的生活 | 1 |
| 次 | 3 | 的國家 | 1 | 與中等 | 1 |
| 是 | 3 | 化 | 1 | 的專題 | 1 |
| 在 | 3 | 器官 | 1 | 潛毛症的 | 1 |
| 是否受到 | 2 | 劇烈的 | 1 | 治療 | 1 |
| 第 | 2 | 位再經過 | 1 | 未 | 1 |
| 越 | 2 | 檢測 | 1 | 吸煙的 | 1 |
| 例 | 2 | rs | 1 | 老年 | 1 |
| 科 | 2 | 和 | 1 | 般 | 1 |
| 性 | 2 | 允許 | 1 | 發生染色體 | 1 |
| 合作的 | 1 | 沒 | 1 | 的每天 | 1 |
| 癌外科 | 1 | 攝取的 | 1 | 方面的 | 1 |
| 原因 | 1 | 接 | 1 | 的要 | 1 |
| 三分之 | 1 | 這項研究 | 1 | 年夏 | 1 |
| 位原 | 1 | 的藥品 | 1 | 的李 | 1 |
| 及 | 1 | 友例 | 1 | 或他形式 | 1 |
| 近視 | 1 | 人因出血而 | 1 | 顯示 | 1 |
| 的失 | 1 | 極具 | 1 | 中的 | 1 |
| 月的 | 1 | 的出血 | 1 | 是為 | 1 |
| 兒童的 | 1 | 院 | 1 | 治療的 | 1 |
| 能力的 | 1 | 的轉移 | 1 | 這種 | 1 |
| 再增加 | 1 | 治療過程中 | 1 | 來做 | 1 |
| 的正 | 1 | ibs | 1 | 繼續 | 1 |
| 凝血因子 | 1 | 用藥 | 1 | 氧氣 | 1 |
| 有三分之 | 1 | 障礙 | 1 | 這種病毒是 | 1 |
| 滴 | 1 | 或斷症狀而 | 1 | 將 | 1 |

| 加入詞彙 | 次數 |
|---|---|
| ncet | 1 |
| 到 | 1 |
| 一種 | 1 |
| 射出 | 1 |
| 限 2 | 1 |
| 這些 | 1 |
| 放棄的 | 1 |
| 星期的 | 1 |
| 發現 | 1 |
| 行為 | 1 |
| 在這方面的 | 1 |

## 附錄 B 動詞詞組附加詞彙

| 加入詞彙 | 次數 | 加入詞彙 | 次數 | 加入詞彙 | 次數 |
|---|---|---|---|---|---|
| 的 | 12 | 敏感更 | 2 | 事故性處理 | 1 |
| 有 | 8 | 受到 | 1 | 少要 | 1 |
| 在臨床 | 6 | 何 | 1 | 規模 | 1 |
| 尚 | 6 | 那麼 | 1 | 醫療師都應 | 1 |
| 輕度且 | 6 | 在一年之內 | 1 | 選出的樣本 | 1 |
| 是 | 5 | 研究機構的 | 1 | 知道 | 1 |
| 可 | 5 | 經由皮膚 | 1 | 位未 | 1 |
| 進行 | 4 | 至警報中 | 1 | 目前為止尚 | 1 |
| 御從 | 4 | 產生 | 1 | 間歇地睡 | 1 |
| 在 | 4 | srs | 1 | 很短暫也很 | 1 |
| 個 | 3 | g 的分配 | 1 | 的限制 | 1 |
| 降低並較 | 3 | 心臟病 | 1 | 經過 | 1 |
| 研究結果 | 3 | 年由名 | 1 | 左該評估能 | 1 |
| 會 | 3 | 成為 | 1 | 到 hsv | 1 |
| 治療的患者 | 2 | 接下 | 1 | 提出 | 1 |
| 先 | 2 | 結果是 | 1 | 後的救治 | 1 |
| 受 | 2 | 至 | 1 | 是那麼常 | 1 |
| 腎臟目標的 | 2 | 到 | 1 | 病毒 | 1 |
| 得 | 2 | 成年 | 1 | 為治療 | 1 |
| 非常 | 2 | 咳 | 1 | 2 | 1 |
| 的團體 | 2 | 時 | 1 | 應該因 | 1 |
| 二 | 2 | 太大的 | 1 | 有人 | 1 |
| 介於 | 2 | 合理 | 1 | 中的 | 1 |
| 可以 | 2 | 從九例中 | 1 | 化學 | 1 |
| 預防 | 2 | 藥物的 | 1 | 三 | 1 |
| 公平 | 2 | 證明之 | 1 | 服用抗癲癇 | 1 |
| 使 | 2 | 指令也 | 1 | 修正後 | 1 |
| 加 | 2 | 不僅 | 1 | 過 | 1 |
| 種的 | 2 | 的研究 | 1 | 小組 | 1 |
| 認為如果 | 2 | 由 ct | 1 | 如此 | 1 |
| 應 | 2 | 出現不的 | 1 | 小腸放射學 | 1 |
| 嚴密的 | 2 | 率 | 1 | 回 | 1 |
| 使 | 2 | 為止這個 | 1 | 給 | 1 |

| 加入詞彙 | 次數 |
|---|---|
| 氫化可松 | 1 |
| 地 | 1 |
| 接受疾病的 | 1 |
| 一種生物 | 1 |
| 認為得 | 1 |
| 第 i | 1 |
| 選 | 1 |
| 其 | 1 |
| 分鐘的 | 1 |

## 附錄 C 介系詞詞組附加詞彙

| 加入詞彙 | 次數 | 加入詞彙 | 次數 | 加入詞彙 | 次數 |
|---|---|---|---|---|---|
| 的 | 58 | 即 | 2 | 至 | 1 |
| 名 | 9 | 治療的 | 2 | 的患者其 | 1 |
| ibs | 4 | meth | 2 | 沒 | 1 |
| 上 | 4 | 他 | 1 | 有瀰漫性 | 1 |
| 患者其 | 4 | 是否受到 | 1 | 兒童心臟 | 1 |
| 位 | 4 | 與這些 | 1 | 位做 | 1 |
| 在 | 4 | 腦膜炎 | 1 | 與患者 | 1 |
| 第 | 4 | 分 治療 | 1 | 起訴及 | 1 |
| 20 | 3 | 造 | 1 | 中的 | 1 |
| 個 | 3 | 所受 | 1 | 的要 | 1 |
| 的長期追蹤 | 3 | 時提示各 | 1 | 使用 | 1 |
| 年 | 3 | 的象 | 1 | 只有移植前 | 1 |
| 是 | 3 | 矯形外科學 | 1 | 單獨 | 1 |
| 起 | 3 | 發生的 | 1 | 將 | 1 |
| 癌 | 3 | 劑 | 1 | 由的 | 1 |
| 有 | 3 | 胃 | 1 | 發 | 1 |
| 的類固醇 | 2 | 的血中 | 1 | 全部 | 1 |
| 仍 | 2 | 用 | 1 | 次理想的 t | 1 |
| 進行 | 2 | 州 | 1 | 改變 | 1 |
| 治療因 | 2 | 數 | 1 | 為期 | 1 |
| 患者 | 2 | 世界上 | 1 | 檢查抗 | 1 |
| 性 | 2 | 引發 | 1 | 抽樣得到 | 1 |
| 藥物 | 2 | 方法 | 1 | 頭痛的 | 1 |
| 種 | 2 | 和 | 1 | 更 | 1 |
| 出現了 | 2 | 患 | 1 | 感染 | 1 |
| 所 | 2 | 會的 | 1 | 於產生換種 | 1 |
| 發生 | 2 | 照 | 1 | 患者進行 | 1 |
| 個體 | 2 | 乳房接受 | 1 | 任與 | 1 |
| 預防 | 2 | 項 | 1 | 對有 | 1 |
| srs | 2 | 得到的 | 1 | erd | 1 |
| 隨訪 | 2 | 所報告 | 1 | 需要 | 1 |
| 只有 | 2 | 月的 | 1 | 類型的 | 1 |
| 會 | 2 | 位無法手術 | 1 | ncet | 1 |

| 加入詞彙 | 次數 | 加入詞彙 | 次數 | 加入詞彙 | 次數 |
|---|---|---|---|---|---|
| 那些希望 | 1 | 介入 | 1 | 其 | 1 |
| 40 | 1 | 族群的治療 | 1 | 術 | 1 |
| 提高對 | 1 | 的研究 | 1 | 就的知識 | 1 |
| 個成因 | 1 | 的健康 | 1 | 時間 | 1 |
| 美國癲癇學 | 1 | 公分 | 1 | 些 | 1 |
| 科學 | 1 | 化的 | 1 | 其中 | 1 |
| 位糖尿病 | 1 | 的既往 | 1 | 困惑 | 1 |
| 例中診斷出 | 1 | 直拳加有 | 1 | 而罹患殘障 | 1 |
| 與狗試驗 | 1 | 及呼吸道 | 1 | 檢查 | 1 |
| 病例 | 1 | 受 | 1 | 艾滋病的 | 1 |
| 神下因 | 1 | 脊 | 1 | 阻 | 1 |
| 臨床腫瘤 | 1 | 起為 | 1 | 抑制的 | 1 |
| 服用 | 1 | 使 | 1 | 吐氣 | 1 |
| 年第 36 週 | 1 | 的隨機 | 1 | 具有 | 1 |
| 全面 | 1 | 傳輸頻率 | 1 | 共有 | 1 |
| 展開 | 1 | 以 | 1 | 肩部鈣化 | 1 |
| 是了 | 1 | 是否有益 | 1 | 一篇研究 | 1 |
| 肥胖者 | 1 | 1800 名 | 1 | 清楚但無的 | 1 |
| 宿主防禦 | 1 | 性過敏症 | 1 | 使得 | 1 |
| 濃度 | 1 | 的腦血流 | 1 | 二 | 1 |
| 包括 | 1 | 第 24 次 | 1 | | |
| 結果使用 | 1 | 內視鏡檢查 | 1 | | |
| 助聽器 | 1 | 月 | 1 | | |
| 類似 | 1 | 旦 | 1 | | |
| 致 | 1 | 日 | 1 | | |
| 閉塞情形 | 1 | 高 | 1 | | |
| 從業人員 | 1 | 次 | 1 | | |
| 社交恐懼症 | 1 | 膽管 | 1 | | |
| 用來 | 1 | 話則其 | 1 | | |
| 例的 | 1 | 1800 | 1 | | |
| 結果 | 1 | 而 | 1 | | |
| 發現有 | 1 | 必須評估 | 1 | | |
| 的藥物 | 1 | 一篇有關於 | 1 | | |
| 聖路易 | 1 | 之間不良的 | 1 | | |
| 行動 | 1 | 然而 | 1 | | |