ROCLING XVI
結合統計與語言訊息的混合式中英雙語句對應演算法

林語君　　高照明

台灣大學資訊工程學系　　台灣大學外國語文學系

sbbjun@gmail.com　　zmgao@ntu.edu.tw

摘要. 本文結合包括句長、標點符號、數字時間詞、原文詞，雙語辭典等各種語言訊息，以動態規劃演算法，找出群組句對應，再階段式的運用以上應用，將所得到的多句對多句的對應作更細微的分割。根據評估，運用兩階段式的動態規劃演算法，再加上以上語言訊息，可以達到近95%的召回率（recall）的情況下，達到80%以上的精確率（precision）。

1　導言

自然語言處理研究中，機器翻譯是其中一項重要主題，從50年代開始研究就從未間斷，其中固然有小規模發展，但機器翻譯的整體效果始終沒有太大的突破。這樣的瓶頸讓研究者得到了一個打破現有思考的結論：長期以來由語言結構分析以及人工建置的翻譯模組的方式，要完全掌握所有自然語言翻譯特徵，有技術上以及複雜性上的困難。90年代左右，發展方向轉爲從一個龐大的雙語語句對應資料庫中，搜尋與擬翻譯語句相關連的雙語翻譯句對。再從這些句對當中以自動或半自動的方式得到翻譯知識與規則。這樣的架構能得出較爲自然、較具延展性的翻譯語句，使機器翻譯的發展出現了另一條可能方向。

在越來越多的研究學者肯定平行語料庫（Parallel Corpus）在機器翻譯上的潛在價值下，伴隨著幾個重要的研究議題，首先是雙語句對資料庫的建立。這個資料庫必須包含各個領域、必須具有足夠龐大的資料量。其中平行語料庫又有不同層次的對應單位，尤以句對應（Sentence Alignment）爲主要的可應用（Utilizable）對應，可以說是最主要的機器翻譯參考結構。平行語料庫往下可以繼續往細部擷取出詞組對應及詞對應，作爲機器翻譯上更小的翻譯單元。

但句對應的平行語料庫得來不易，人工對應不僅昂貴，資料量的不足更會直接影響到應用平行語料庫的品質。本系統目標在於大量且自動的取得中英句對應的平行語料庫，除了綜合至目前爲止的句對應模型（Sentence Alignment Model）之外，更針對中文以及英文的特殊屬性做不同參數以及演算法的調整，如中英句長、標點符號、數字時間詞、原文詞，再以各種不同的英漢／漢英字典以及Hownet等各模組的互相配合，以動態規劃演算法以及更進一步的遞迴階段分割方式產生較小的句群對應，並在縮小對應的區塊時盡量兼顧對應的正確率。

1.1　文獻回顧

雙語句對應的研究開始於90年代初期。Gale 與Church (1991) 及Brown 等 (1991) 觀察到長句的翻譯對應句一般而言較長，而短句的翻譯句通常較短。他們利用句長的關連性配合動態規劃或EM演算法得到96%以上的正確率。Gale 與Church (1991) 及Brown 等 (1991)兩者最大的差別是前者透過人工先得到先驗機率（prior probability）而後者利用EM演算法得到相關的參數。Wu (1994) 及Xu and Tan (1996)以句長爲主結合一個包含日期及數字等訊息小的辭典得到96%的正確率。以句長爲基礎的統計方法的優點是不需要語言知識及辭典就可以運作。缺點是如果語料中含有豐富的多對多的句對應關係，或是翻譯的語料中有增添或刪減的現象發生就會造成正確率大幅下降。前述幾項研究由於大都採用議會的紀錄，例如Gale與Church (1991) 及Brown 等（1991）用加拿大國會Hansard英法平行語料，Wu (1994)則利用香港立法局議會質詢與答詢的中英平行語料，由於是口語紀錄所以句子較短，且不少是一對一對應。Gale 與Church (1991)統計Hansard語料80%以上是一對一的對應關係，罕有多對多的對應關係或增添或刪減的情形發生，所以以句長爲主的統計方法得到很好的效果。但McEnery and Oakes (1996)以Gale 與Church (1991)

的方法做實驗卻顯示此種演算法的正確率對不同的文類與語言會產生很大的差異。例如波蘭文英文平行語料的正確率因文類不同介於於100%與64.4%，而他們所實驗的中英新聞平行語料更低於55%,這證明單純以句長關連性顯然無法得到高正確率。

另一個不需要辭典的方法是Kay and Röscheisen (1993) 以詞彙的頻率（去除低頻的詞及高頻的詞）及在文章中出現的分佈，建立可能的詞對應表及句對應表並不斷的修正，以relaxation方法達到收斂。與Gale 與Church (1991) 及Brown 等 (1991)方法一樣，Kay and Röscheisen (1993)的方法只有在一對一的情形佔絕大多數時才會有好的效果。此外此種方法過度重視詞頻，文章的長度太短會造成正確率的大幅下降。這個演算法另一個實做上的問題是處理十分耗時，無法快速處理大量語料。

另外Melamed (1997a)提出Smooth Injective Map Recognizer (SIMR) 利用統計和同源詞(cognate)，正確率高於Gale 與Church (1991)，但我們以光華雜誌做初步實驗發現正確率仍然只有60%左右。

以統計爲主的方法不管是以長度，詞頻及詞彙內部分佈，或geometric，在正確率及強健性方面似乎都不理想，因此使用雙語辭典似乎是提昇正確率所必需，但如果只以雙語辭典找句對應效果也不理想，原因是翻譯的基本單位在很多時候並不是詞，而是詞組或結構，因此Catizone et al. (1989)提出結合辭典與統計訊息。Haruno and Yamazaki (1996)比較純統計式，辭典，與混合式三種方法，發現混合式在精確率precisionz 召回率recall介於91.6% 到 97.1%之間，比採純統計式或只用辭典的方式好。Utsuro 等 (1994)也採取辭典與句長爲主的混合法，但錯誤率介於4.6% and 21.6%。顯示即使採用混合法正確率也隨著語料的不同與演算法細節的不同而有相當大的差異。

語言訊息除了雙語辭典還有其它的訊息可以用來找句對應，例如Yeh (2002)發現在光華中英平行語料中標點的訊息有助於找到句對應。本文的目的在於探討混合式的句對應演算法包含哪些訊息及這些訊息應該如何組合起來才能達到最佳的句對應效果。

2 系統架構與流程

我們的系統大體架構如圖1所示。輸入雙語平行語料，最後輸出其中的雙語群組句對（一個群組句對可爲一句對一句／一句對多句／多句對多句的雙語組合，但是不包含零句對一句或零句對多句的情況，也就是說所有的句子都會被納入某個群組句裡）。

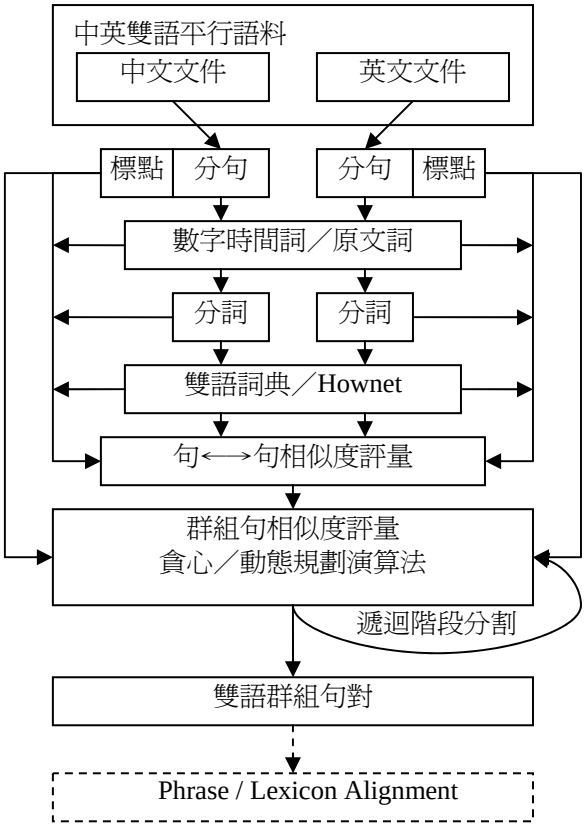

圖. 1. 系統架構與流程

2.1 分句

中文句子的界線並不清楚，句點與逗點都有可能是句子的界線。如Gao (1997)所指出句點事實上是一個比句子還要大的言談單位(discourse)，裡面可能包含數個句子，而逗點有時是一個句子，有時只是詞組。我們的分句主要將原始文件切割成最小的句對應處理單位，但因句對應模組所採用的是可以接受句子多對多的群組句對應演算法，所以並不一定要對分句做太嚴格的處理，也就是說，錯誤的分句仍可以在之後的群組句對應演算法中，有機會與前後語句結合，成爲一個適當的「句單位」。

因仰賴之後的多對多句對應演算法，最佳的句分割就成爲「在句子不會失去完整的意義情況下，盡量地細微分割。」因此，實作上就成了以下原則：

一、 以標點符號作爲句子爲單位，將句號、問號、驚嘆號、分號前後的區塊拆分開來，但在英文方面需注意是否該句點爲單字的一部份，如Mr.等之縮寫。英文句的切割我們使用Shlomo Yona Lingua::EN::Sentence的Perl模組（http:// search.cpan.org / ~shlomoy / Lingua-EN-Sentence-0.25 / lib / Lingua / EN / Sentence.pm）。

二、 如果有兩個太長的句子（兩個超過十個中文詞或六個英文單字的句子）以逗號連結起來，則視爲兩個句子。

三、 另外，對一些對句對應有幫助的符號做特殊的處理，如將發言句、引用句等的前後冒號及引號作爲單獨的句子（一個句子僅擁有幾個符號），如此可以強制句對應演算法將這些有特殊意義的句子視爲「錨」（Aligning anchor），賦予極高的對應權重，作爲一個重要的對應參考點。

2.2 分詞與辭典查詢

因中英文的詞組結構迥異，故分別採用不同的分詞方法。在這裡我們一樣不採用嚴格的分詞法，因分詞的主要目的爲取得該句的對應句中，所有可能的對應翻譯詞，以供之後的句與句之間相似度的衡量，所得出來的相似度衡量值爲一相似概值，故分詞的準確度，並不會大幅度的影響分句的結果。

2.2.1 分詞

英文方面因爲有比較明顯的分詞符號（空白），比較沒有分詞上的困難。

在中文分詞問題方面，我們採用的是「所有可能翻譯詞」。舉「把他的確實行動作了分析」這個句子爲例，「的確」、「實行」、「動作」、「分析」，以及跟上述分詞有重疊的「確實」、「行動」，因爲在漢英字典裡有對應翻譯，故全部予以採用（但如其中的高頻率詞不予採用，述於下節）。

2.2.2 字典查詢

分詞的目的在於找到對應翻譯詞，故可同時採用查詢不同的多個字典，如一般辭典以及Hownet裡的特殊辭典。但要如何避免因爲擴張翻譯詞而造成的錯誤碰撞（false hit）呢？我們採用一個頻率辭典來完成，頻率高於某值的詞被視爲stop word，不列入翻譯對應詞的衡量。

對於英文的翻譯詞查詢，我們利用stemming得到詞的原型，在此我們採用Ron Savage的Lingua-EN-Infinitive之Perl模組（http://search.cpan.org / ~rsavage / Lingua-EN-Infinitive-1.08 / lib / Lingua / EN / Infinitive.pm）。

2.3 雙語句之間的相似度評量

這個階段將分句階段的中文各句以及英文各句，連同各句包含的詞對應，以及各句句中的標點符號，數字詞，原文詞，一起作爲雙語句關連評量的參數，計算出一個任意句對關連值來。這些句對關聯值（一個表）將爲下一階段的群組句對相似度評量的最重要參數。

2.3.1 標點符號

這裡的標點符號並非單指句末的標點符號，而是句中的「所有重要標點符號」所形成的序列（在這裡我們將非逗號及非句號之標點符號視爲「重要」），句與句之間的標點符號相似度即爲這個序列的相似度，實作方法爲：

$$P_i = \{p_{i1}, p_{i2}, p_{i3}, \ldots\ldots, p_{in}\}$$
$$Q_j = \{q_{j1}, q_{j2}, q_{j3}, \ldots\ldots, q_{jm}\}$$
$$punc\_sim_{i,j} = \mathrm{LCS}(P_i, Q_j) \times W_{punc\_sim}$$

以上P_i代表句子$_i$中的標點符號，以Q_j代表句子$_j$中的標點符號，按照出現順序所排列而成的標點符號序列。LCS函數爲最長共同子序列（Longest Common Sequence）。W_{punc_sim}爲標點符號相似度參與句與句相似度衡量的比重。最後，$punc_sim_{i,j}$爲這兩個序列的相似度的加分值。

2.3.2 數字詞與時間詞

數字詞與時間詞爲跨語言中的重要共通語意部分，字典往往沒有這些詞。在這裡，我們額外地從中英文句對中，偵測並抽取出數字以及時間的資料，如果有相同的數字以及時間記錄，則將該句對設定爲一個極高權值的對應「錨」。但，太簡單的數字不予考慮，在這裡我們忽視1, 2, 3三個對應中英文。

2.3.3 原文詞

另外一個更具高權值對應的部分爲原文詞，這種在譯文中保留原始語言文字的特性極常出現，一旦偵測到即有極高的可信度。我們的作法即爲在中文語料中偵測英文單字、反之亦然，額外地我們也將中文全形英數字以及標點符號做一個簡單的轉換，提高原文詞比對率。

對於一個句對中比對到一個原文詞，則該句對亦被設定爲一個極高權值的「錨」。

2.3.4 句對相似度評量

綜合以上原則以及各項參數，我們可以得出

$$\mathrm{eval}\left(sen_i^{\mathrm{CH}}, sen_j^{\mathrm{EN}}\right) =$$

$$\frac{\left(\#\left\{ word_{i,k}^{\mathrm{CH}} = trans\_word_{j,l,m}^{\mathrm{EN}} \;\middle|\; \begin{array}{l} \text{for } i = 1..\#sen^{\mathrm{CH}} \\ \text{for } j = 1..\#sen^{\mathrm{EN}} \\ \text{for } k = 1..\#word_i^{\mathrm{CH}} \\ \text{for } l = 1..\#word_j^{\mathrm{EN}} \\ \text{for } m = 1..\#trans\_word_{j,l}^{\mathrm{EN}} \end{array} \right\} + \right.}{\#word_i^{\mathrm{CH}} + \#word_j^{\mathrm{EN}}}$$

$$\left. \#\left\{ word_{i,k}^{\mathrm{EN}} = trans\_word_{j,l,m}^{\mathrm{CH}} \;\middle|\; \begin{array}{l} \text{for } i = 1..\#sen^{\mathrm{EN}} \\ \text{for } j = 1..\#sen^{\mathrm{CH}} \\ \text{for } k = 1..\#word_i^{\mathrm{EN}} \\ \text{for } l = 1..\#word_j^{\mathrm{CH}} \\ \text{for } m = 1..\#trans\_word_{j,l}^{\mathrm{CH}} \end{array} \right\} \right)$$

$$+ \; punc\_sim_{i,j}$$

$$+ \; \infty \times if\_share\_數字詞_{i,j}$$

$$+ \; \infty \times if\_share\_時間詞_{i,j}$$

$$+ \; \infty \times if\_share\_原文詞_{i,j}$$

我們假設所有編號從1開始，以 sen_i^{CH} 來代表編號*i*中文句，以 $\#sen^{\mathrm{CH}}$ 來代表中文句數，以 $word_{i,j}^{\mathrm{CH}}$ 代表句子 sen_i^{CH} 的分詞後的編號第*j*個詞，以 $\#word_i^{\mathrm{CH}}$ 代表句子 sen_i^{CH} 的分詞後的詞個數，以 $trans\_word_{i,j,m}^{\mathrm{CH}}$ 代表詞 $word_{i,j}^{\mathrm{CH}}$ 的編號第*m*個翻譯詞（故爲英文），以 $\#trans\_word_{i,j}^{\mathrm{CH}}$ 代表詞 $word_{i,j}^{\mathrm{CH}}$ 的翻譯詞數目。並以*i..j*代表一個從*i*至*j*的迴圈（各一次）。英文的情況則將以上代號之CH改成EN。 $punc\_sim_{i,j}, if\_share\_數字詞, if\_share\_數字詞, if\_share\_原文詞$ 爲前述之句對應評量調整參數，後三者布林變數，該詞存在則爲1，不存在則爲0。

對於檢驗某 $word_{i,j}^{\mathrm{CH}}$ 是否等於 $trans\_word_{i,j,m}^{\mathrm{EN}}$，兩者均爲中文詞，除了完全相同的比對方式以外，另可使用部分比對（Partial Match）的方式：如果兩詞中有兩個以上（包含兩個）部分字元相等，則視爲兩詞相等。

2.4 群組句與群組句之間的相似度評量

這可以說是雙語語料庫中配對句對應的最後一個階段。在這個階段裡，有四個輸入參數：

一、 上個階段中文各句以及英文各句的所有句對的相似值（即爲一表格）。
二、 句子的句長。
三、 句子的句末標點符號。

輸出則爲本系統的最終所求：中英文文章的「句群對應」，「句群」爲一個或一個以上的句子的群組，換言之句群對包含傳統「一對一」句對應、「一對多」句對應、「多對一」句對應、以及「多對多」句對應。

一般來說，一個標點符號結構相似，「句譯」比「意譯」來的多的平行語料庫，群組句對應的正確結果應多爲「一對一」。相對的，在標點符號結構差異大，或者平行語料庫屬於粗略、大概的翻譯，或者翻譯多採意譯的方式的平行語料庫，則正確的群組句對應爲「一對多」、「多對多」較多。

在一個群組句裡有多句的情況，邊界的判斷就成了正確率的關鍵，也就是該群組不可以無限制的膨脹，包含了過多的句子。更正確的說法應該是，我們必須加上一個群組膨脹的「懲罰值」（Penalty），來避免這種情況。因爲，如果在沒有懲罰的情況下，一個群組包含了越多的句子，則該群組與其他群組的相似值則會無限制的上升，造成最後所有的句子會變成一個群組的窘境。各項實作方法敘述下。

以下爲三個群組句對應相似度評量的參數說明。並在之後接著說明群組句對應的建立模組，也就是本系統的核心演算法。

2.4.1 句長

一個群組句的句長定義爲：

$length_{gi}$ = 群組句*i*裡的所有句子的句長總和。

群組句對應的句長影響值定義爲：

$$eval\_length_{gi,gj} = -\left|length_{gi} - length_{gj}\right| \times W_{length}$$

W_{length}爲總句長相異度參與群組句之間相似度衡量的比重。$eval_length_{gi,gj}$爲兩個群組句*i*以及*j*之間的句長的相異度在群組句對應評量中的影響值（懲罰值）。

2.4.2 句末標點符號

一個群組句的句末標點符號定義爲：

$punc_{gi}$ = 群組句*i*裡最後一個句子的句末標點符號

群組句對應的句末標點符號影響值定義爲：

$$eval\_punc_{gi,gj} = \begin{cases} -W_{punc}, & \text{if } punc_{gi} \neq punc_{gj} \\ 0, & \text{if } punc_{gi} = punc_{gj} \end{cases}$$

W_{punc}爲句末標點符號的相異在群組句之間相似度衡量上的比重。$eval_punc_{gi,gj}$爲兩個群組句*i*以及*j*之間的句末標點符號相異在群組句對應評量中的影響值（懲罰值）。

2.5 群組句對應建立模組

運用以上群群組句對應相似度評量的原則以及參數，我們以動態規劃來完成最佳化的群組對應系統。在這裡我們有兩個模組：動態規劃演算法、以及以多重不同的句子規模，重複運用動態規劃演算法，來達到漸漸增加群組對應解析度的遞迴階段動態規劃。在介紹動態規劃之前，我們先討論較爲直觀的貪心演算法。

2.5.1 貪心演算法

貪心法（Greedy Algorithm）基本原則爲，參考已經算出來的中英各句子之間的評量係數值，來完成句子對應的工作。

首先選取「一對一」作爲iteration的base result，指定中文句集合以及英文句集合各一個可移動的指標，每一回合的iteration嘗試延伸任意一個指標至下一個中文（或英文）的句子，並把這一個句子加入評量對應句組，重新評量對應句組的對應評量係數，並減去一個懲罰係數（敘述於下），以作爲避免產生指標無限延伸的結果。當兩指標都無法經由延伸而提高評量係數的時候，一組中英對應句組於是產生。

如此重複此貪心演算法，直至任一指標使用完所有的句子爲止。

懲罰係數爲額外囊括一個額外句子進入目前句對應組所必須付出的代價，我們將之定義如下：

$$penalty(S) = f_S \times avg_{matced\_words\_rate}$$

f_S 爲句子的語意單元個數，$avg_{matced\_words\_rate}$ 爲平均一個正確的中英句對應中，相同語意單元（lexical unit）的數量除以句子語意單元個數的平均。這個平均值將影響句對應組是糾結成群的模糊對應，還是零碎的精準對應的關鍵值。此平均值可爲手動統一指定（按照字典的完整程度來調整），或由程式統計判斷（在此一個正確的中英句對應的定義爲，句對應的評量係數明顯高於相鄰句對應組的情況，則該對應句的*matched_words_rate*則可作爲調整 $avg_{matced\_words\_rate}$ 的因子）。

此演算法的優點爲速度快、效率高，缺點是當如果有任何一次選擇對應句組的時候遭遇錯誤分配，則該句之後的句對應演算情形將不樂觀。故本演算法適合處理正確句對應句數應盡量接近「一對一」爲原則的文章。

2.5.2 動態規劃演算法

爲了避免Greedy Algorithm在句對應演算中，因每次產生的對應句組僅爲該次句對應評量的最佳組合，而非整體文章的對應組最佳組合，故我們使用動態規劃演算法來解決總體最佳化的問題。

我們假設句編號從1開始，以 sen_i^{CH} 來代表編號*i*中文句，以 $\#sen^{\mathrm{CH}}$ 來代表中文句數。對於群組句，以 $sen_{i,j}^{\mathrm{CH}}$ 代表一個擁有編號*i*至*j*所有中文句的中文群組句，並以*i..j*代表一個從*i*至*j*的迴圈（各一次）。英文的情況則將以上代號之CH改成EN。

首先建立 $DP\left(1..\#sen^{\mathrm{CH}}, 1..\#sen^{\mathrm{EN}}\right)$ 的兩維動態規劃表，對於累積至目前爲止的最佳句對應組考量評量係數的計算方式，採取

$$DP(i,\ j) = \max\left\{\begin{array}{l} DP(p,q) \\ \quad + eval(sen^{\mathrm{CH}}_{p+1..i}, sen^{\mathrm{EN}}_{q+1..j}) \\ \quad + eval\_length_{sen^{\mathrm{CH}}_{p+1,i-1}, sen^{\mathrm{EN}}_{q+1,j-1}} \\ \quad + eval\_punc_{sen^{\mathrm{CH}}_{p+1,i-1}, sen^{\mathrm{EN}}_{q+1,j-1}} \\ \quad - \sum penalty(sen^{\mathrm{CH}}_{p+1..i-1}) \\ \quad - \sum penalty(sen^{\mathrm{EN}}_{q+1..j-1}) \\ \quad , \text{for } 0 \le p < i, 0 \le q < j \end{array}\right\}$$

*eval*爲上一階段之各單句對應的評量函數之表格查詢。e*val_length,* e*val_punc*爲前述之群組句對應衡量時之句長、以及句末標點符號的影響值（懲罰值）。*penalty*懲罰函數與貪心法定義之懲罰函數相同。

最後，$DP\left(\#sentence^{\mathrm{CH}}, \#sentence^{\mathrm{EN}}\right)$則爲最佳整體對應組之評量係數。由此評量係數回溯組合成對應的最佳整體句對應分配。

此演算法的優點爲可解決中途的句對應的錯誤所造成的全體錯誤，某些不合理的分配將會在其它回合的評量當中獲得矯正。本演算法擁有較爲可觀的計算雜度以及額外的空間。但可以不受斷句的相關性的影響，再配合更精良的判斷評量權值以及屬性，更可大幅提高句組對應的正確率。

2.5.3 遞迴階段動態規劃

這個步驟是句對應演算法改良的一個重點。我們以上述的DP加上可參數化的機制以後，對文章不只作一次的DP句對應，而是以不同的參數設定，執行兩次或兩次以上。如此一來可以根據不同的文章和考量規模作出優化的調整。

對評量的中英文文章作第一次的DP處理時，所用的參數和分數的懲罰等可以放寬，目的是增加區塊對應的正確率，但同時也會使得區塊容易因爲區塊邊界的模糊而呈現脹大的現象。把這些初步的小區塊再分別送至第二階段的DP處理，第二階段的DP處理參數則會採取比較嚴格、採重罰的形式，鼓勵句子分句單獨化，在第一階段正確區塊對應的前提下，這樣子冒險的假設可以在比較安全的環境下，得出正確率高又不令人失望的對應。

這樣的Iterative Process要執行幾次可以透過自動判斷的方式進行。在某些篇幅十分壯觀，經過兩次的DP處理後，仍然存在著大區塊對應的情況下，可以將參數定義的更加的嚴格，送往第三次、甚至是第四次的分割處理。但也有可能因爲該對應區塊因爲確實應該對應在一塊，這樣的情況下就不應該無限制的增加參數的嚴格性，導致因爲處理太過於苛求細緻而產生錯誤的對應。

3 效能評量與雙語庫

我們的效能評量採用Pierre Isabelle (1996) 的Sentence Alignment的Evaluation Metric。對於兩組句對應的找出「雙語對應空間」上的面積交集（圖. 2）：

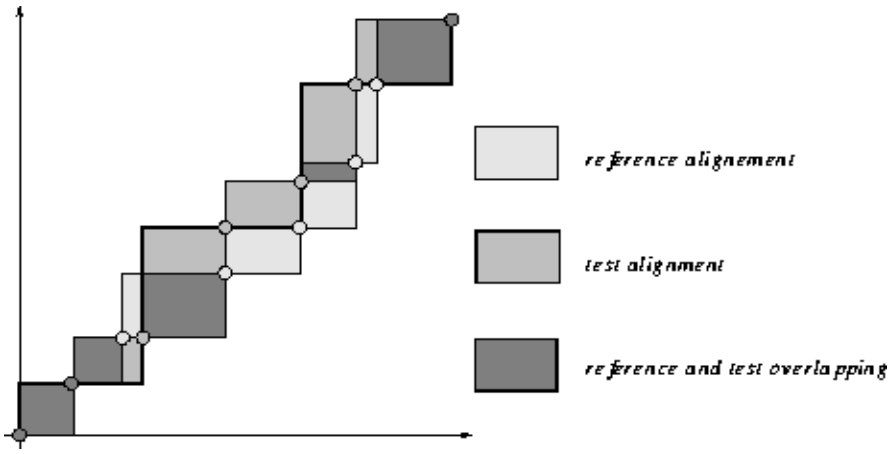

圖. 2. 句對應recall以及precision之圖示

將兩軸視爲兩語言的字元串，則句對應的表示方式則爲其個別語言之句子，在兩軸上所佔據的範圍，延伸在對應空間上則會交出一個長方形。則我們可將兩組不同的文章句對應用以上的方式表示，利用其交集來計算精確率（precision，即交集面積除以正確句對應的所有面積）以及召回率（recall，即交集面積除以評估的句對應的所有面積）。

運用上述的不同的演算法組合，對此平行語料庫產生出來的句對應分配，與「正確」的句對應分配進行評估。該正確句對應分配的產生方法，以人工的方式來產生。

因人工產生之正確句對應組合需要人力介入，故我們採用兩組不同的雙語語料庫：其一爲光華雜誌2001年1月單月的雙語語料來作評估。該月雙語語料庫專文共24篇，共有54336中文字，37195英文字。另外一組爲以隨機的方式選擇各式雙語文件做評估，如美國知音等，以避免系統針對光華雜誌做最佳化的調整。

4 群組句對應評估

因爲群組句對應的建立模組與其他幾個評估變因較爲獨立，故可提前評估。如此在以下的評估中，可以直接採用較高正確率的一個模組，以簡化評估之變因數。

在這裡我們測試貪心演算法（Greedy Algorithm），動態規劃（Dynamic Programming）的優劣，以及遞迴階段式動態規劃（Iterative DP）是否有對動態規劃造成助益，呈現更好的結果。除此變因外，採用所有其他的輔助參數。

Table 1. 評估結果

Recall (%) Precision (%)	Greedy	DP	Iterative DP (2 stages)	Iterative DP (3 stages)
Sinorama	73.37	96.47	93.01	86.79
(200101)	70.82	72.40	81.26	83.83
Random	84.18	98.17	94.70	91.12
	73.92	76.81	83.23	84.65
Average	78.78	97.32	93.86	90.45
	72.37	74.60	82.25	83.24

由評估結果，我們可以看出，兩階段的Iterative DP總體效能最好，但與三階段的Iterative DP來比較，則所勝不多。貪心法速度快，但容易造成一對句配對錯物導致之後的句子配對的連環錯誤。DP則因爲解決了這樣的問題，做整體性的規劃，故在效能上有明顯的提升。相較於DP對於Iterative DP的更進一步的語句裁剪，導致Recall的下降，Precision上升，由上表可知，超過了兩階段的DP之後，群組句對應已經到了一個穩定的階段，加深DP的階段數並無法提高整體的效能。

5 實驗結果與分析

總和以上評估，可得出以下最佳的句對應組合：

- 群組句對應模組
  - 2 Stages Iterative DP
- 句對應參數
  - 採用廣泛的字典翻譯詞以及Stop list
  - （中文翻譯詞的部分比對（Partial Match））。
  - 句中重要標點符號序列相似度
  - 共同數字詞、時間詞、原文詞之對應錨
- 群組句對應參數：
  - 句長相異度
  - 句末標點符號要求相同

翻譯詞的比對是否運用中文Partial Match（本系統採用兩個字以上），因爲在效能上沒有太大的改善，故可以斟酌使用，或是運用更複雜的Partial Match。

系統在召回率率方面表現良好，運用單階段動態規劃將可使召回率幾近100%，但因動態規劃本身的多對多對應的性質，故常有對應正確，群組句的成員句數量過大的情況，導致精確率略微偏低。運用多階段的動態規劃將可以在犧牲些許的召回率的情況下，相當程度的提高精確率至80%以上的水準。

6 結論與未來的研究方向

本系統綜合各種現存已知的不同的自然語言訊息，協助單階段、甚至多階段的動態規劃演算法，在可以處理多對多群組句對應的情況下，得出一個極高召回率、並且擁有可接受程度的精確率的一套句對應系統。

因爲中英文屬於不同語系，相較於其他Sentence Alignment較多著眼於同一語系的雙語語料庫來說，因爲在語言結構上有著明顯的差異，故在譯文中常運用意譯的方式，整體效果無法呈現相同的高效能。又中文缺乏明顯的詞的分界，功能詞的數量多而且用法多變，更加深了自動中英文句對應上的難度。

本系統總結出的最佳效能多對多群組句對應演算法組合，仍有一些可以改善的重點：

- 未知詞比對
  加入專有名詞的檢索與翻譯詞的比對，並建立未知詞資料庫，提供不同的未知詞相似權值，可視爲不同的對應錨（Anchor）。
- 翻譯詞比對：
  因中文詞與英文詞在翻譯詞上較難出現相同詞的特徵下，採用完全比對的方式往往會造成相關句之間的評量的權值不足，被誤判爲沒有相關，故可在翻譯詞上的比對上嘗試以下改良：
  - 詞語意類別：
    可以嘗試採用詞語意類別（參考Ker and Chang (1997)），如car, train等字詞屬於transportation的category，可以額外在字詞比對上提供其他的相似訊息。
  - 更精準的部分比對：
    採用統計的方式來支援字詞的部分比對，提供完全比對之外的相似訊息。
- 加入統計式的詞對應相似度模型（如Melamed (1997b)）
  針對語料庫做統計式的詞分析，可以用來支援字典涵蓋詞的不足。

致謝

本文得到國科會計劃NSC91-2411-H-002-080 「詞彙語意關係之自動標注—以中英平行語料為基礎」資助及清華大學劉顯親教授與張俊盛教授共同主持之國科會計劃NSC92-2524-S007-002 「前瞻性數位語言學習中心CANDLE之研發：應用（雙語）語料庫及電腦化學習之支援」分項子計畫之資助特此致謝。

參考文獻

Brown, P. et al. (1991) “Aligning Sentences in Parallel Corpora.” In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp.169-176, Berkeley, Ca.

Catizone, R. et al. (1989) “Deriving Translation Data from Bilingual Texts.” In Proceedings of the First International Lexicon Acquisition Workshop, Detroit, Michigan.

Chang, B., Danielsson, P. and Teubert, W. (2002) "Extraction of Translation Unit from Chinese-English Parallel Corpora," *COLING-02: The First SIGHAN Workshop on Chinese.Language Processing*.

Gale, W. and Church, K. (1991) “A Program for Aligning Sentences in Bilingual Corpora.” In Proceedings of the Annual Conference of the Association for Computational Linguistics, pp. 177-184.

Gao, Z.-M. (1997) Automatic Extraction of Translation Equivalents from Parallel Corpora. Ph.D. Thesis. University of Manchester Institute of Science and Technology.

Haruno, M. and Yamazaki, T. (1996) “High-Precision Bilingual Text Alignment Using Statistical and Dictionary Information.” Proceedings of the Annual Conference of the Association for Computational Linguistics, pp.131-138.

Isabelle, P. and Simard, M. (1996). "Propositions pour la représentation et l'évaluation des alignements de textes Parallèles. ”

Kay, M. and Roscheisen, M. (1993) “Text-Translation Alignment.” Computational Linguistics, Vol. 19, No 1, pp. 121-142.

Ker, S. J. and Chang, J. S. (1997) “A Class-based Approach to Word Alignment.” C*omputational Linguistics*, Vol. 23, No. 2, pp. 313-343.

Le, S., Youbing, J., Lin, D., and Sun, Yufang 2000 “Word Alignment Of English-Chinese Bilingual Corpus Based on Chunks”, In *Proc. 2000 EMNLP and VLC*, pp. 111-116.

Melamed, D. (1997a) “A Portable Algorithm for Mapping Bitext Correspondence.” In Proceedings of the 35th Annual Conference of the Association for Computational Linguistics, Madrid.

Melamed, D. (1997b) “A Word-to-Word Model of Translational Equivalence.” In Proceedings of the 35th Annual Conference of the Association for Computational Linguistics, Madrid.

McEnery, O. and Oakes, M. (1996) “Sentence and Word Alignment in the CRATER Project.” In Thomas and Short (eds.) Using Corpora for Language Research, pp. 211-231. New York: Longman.

Michel, S.and Plamondon, P.1(996) “Bilingual Sentence Alignment: Balancing Robustness And Accuracy,” In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-96)*, pp. 135-144, Montreal, Quebec, Canada.

Tiedemann, J. (1998) “Extraction of Translation Equivalents From Parallel Corpora ” In *Proceedings of the 1th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen.

Utsuro, T. et al. (1994) “Bilingual Text Matching Using Bilingual Dictionary and Statistics.” In Proceedings of International Conference on Computational Linguistics, pp. 1076-1082, Kyoto.

Wu, D. (1994) "Aligning A Parallel English- Chinese Corpus Statistically With Lexical Criteria," In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM.

Xu, D. and Tan, C. L. (1996) “Automatic Alignment of English-Chinese Bilingual Texts of CNS News.” In Computational Linguistic Archive cmp-lg/9608017.

Yeh, Chih-Cheng. (2002) Using Punctuation Marks for Bilingual Sentence Alignment. MA Thesis. National Tsing Hua University.