行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※

※　從中英平行語料庫自動擷取雙語詞彙知識　※

※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：個別型計畫

計畫編號：NSC　89－2411－H－002－043

執行期間：88 **年** 8 **月** 1 **日至** 89 **年** 9 **月** 30 **日**

計畫主持人：高照明

zmgao@ccms.ntu.edu.tw

執行單位：國立台灣大學外國語文學系

中華民國 89 **年** 12 **月** 5 **日**

行政院國家科學委員會專題研究計畫成果報告

計畫編號：NSC　89－2411－H－002－043

執行期限：88 年 8 月 1 日至 89 年 9 月 30 日

計畫主持人：高照明

zmgao@ccms.ntu.edu.tw

執行單位:國立台灣大學外國語文學系

關鍵詞：自動擷取雙語詞彙知識，中英平行語料庫，計算語言學，自動抽取翻譯

一摘要

本研究探討如何有效地結合統計與語言知識從中英平行語料庫自動擷取雙語詞彙知識。我們以 Fung 與 Church (1994)所提出 K-vec 演算法結合互見訊息(mutual information)與 t 值(t-score)來計算出現次數大於 3 次且頻率接近的中文詞與英文詞在文件內部區段的共現關聯性。我們使用中研院詞知識庫小組發展的分詞程式處理中文的分詞，並以中英對照之光華雜誌做實驗證明上述方法的精確度受到文類的影響相當大，精確度有可能高至 70%也有可能低至 30%以下。此外利用此演算法實際能找到的對應詞相當有限。數千詞長度的對應文章，大多只能找到幾個對應詞。我們合併賓州大學（University of Pennsylvania）與美國暑期語言學院（Summer Institute of Linguistics）所發展的構詞分析程式，將英文所有的名詞，動詞，形容詞換成原型，使計算共現機率時能更準確。此外所我們不僅利用文章內部區段一起出現的機率，也收集以數十篇對應文章，再以中文詞與英文詞出現在同一篇文章的機率（亦即文件的共現關聯性）來過濾 Fung 與 Church (1994)演算法所得到的結果，將精確度大幅提高至 90%以上。我們也利用機讀英漢電子辭典，部分字串匹配，以及緊鄰性(proximity)原則得到部分詞組的翻譯。雖然上述方法所能找到的詞組翻譯相當有限，但精確度高達 90%以上。透過這些高精確度的對應詞或詞組，我們即可得到某些中文與英文句子的對應。我們再使用中研院的詞類標記程式標示中文句子每個詞的詞類，並使用 Eric Brill 的英文詞類標記程式與 Steven Abney 的英文句法剖析器。我們根據中研院的詞類標記程式得到的結果將中文句子分成主詞與動詞組兩部分，並依據 Abney 程式的結果找出英文主詞與動詞組兩部分。接下來我們比對機讀辭典與平行語料

的翻譯是否有部分字串吻合，若有則利用兩個翻譯的句法結構大致對應，找出對應的詞組。本研究所採取結合統計與語言訊息的方法能準確的找到詞與詞組的翻譯對應，對於自動編纂中英辭典，機器翻譯，與跨語言檢索有相當大的助益。

Automatic Bilingual Lexical Knowledge Acquisition from a Parallel Chinese-English Corpus

Keywords: automatic bilingual lexical knowledge acquisition, parallel Chinese-English corpus, automatic extraction of translation equivalents

In this research, we combine statistics-based and dictionary-based algorithms in deriving and augmenting a translation lexicon at word and phrase level. We first reimplement Fung and Church's (1994) K-vec algorithm. K-vec is a simple algorithm to find word correspondences from parallel texts. The basic idea is that a true word pair should have similar distributions in terms of the position of its occurrence in the text. To estimate the similarity of co-occurrence, the parallel texts are split into the same number of segments (K) and the distributions of each word are represented in a 1…K binary vector. Two statistical measures, mutual information (MI) and t-score, are then employed to calculate the similarity of distributions of any Chinese-English word pair. Although K-vec is a quick and easy method to derive an initial translation lexicon, its precision is subject to the constraint of genres, frequency, and several other factors. We test the K-vec algorithm against the Sinorama Chinese-English corpus and find that the precision ranges from 30% to 70%. We also experiment on augmenting translation lexicon based on existing machine-readable dictionaries. The result also turns out to be very poor, the main reason being that translations are seldom word for word. Furthermore, exact string matching between dictionary listing and words in the parallel corpus can only identify a small portion of correct word correspondences, whereas partial matching of dictionary lookup can find many word correspondences, most of which are incorrect. We propose two methods to address these problems. The first method, based on proximity of word pairs and the similarities of word order between Chinese and English, can extract phrasal translations and reliable word correspondences simultaneously from an unaligned parallel corpus. The second method utilizes the co-occurrence information of word pairs in different documents to filter out spurious word correspondences. This method can achieve high-precision word and phrasal correspondences which can then be used to derive sentence correspondences. We process the aligned Chinese-English

sentences with Chinese and English part-of-speech taggers. We also use Abney's parser to derive the syntactic structures of English. After identifying the subject noun phrase and verb phrase of each aligned Chinese-English sentence, we derive phrasal correspondences based on the syntactic structures and clues of dictionary lookup. The basic assumption is that the structures of aligned sentences are very likely to correspond to each other if it is further supported by evidence of partial match with dictionary lookup. This hybrid approach can construct a high-precision translation lexicon at word and phrase level which can be fruitfully applied to computational lexicography, machine translation, and cross-lingual information retrieval.

二、緣由與目的

雙語詞典傳統上都是靠辭典編纂者經年累月的收集分析資料慢慢建立而成。這種作法需要大量人力物力且曠日廢時。近年來，隨著語料庫為主的計算語言學的興起，利用平行語料庫（即兩種或多種語言對照之文章資料庫）與統計演算法可以自動抽取詞與詞組的翻譯。統計演算法的優點在於只需要大量語料庫不需要語言知識即可自動擷取詞彙知識，缺點是受到頻率，語系,文類,風格等因素的影響很大。再者，統計演算法基本上是根據詞在文章出現位置的分佈情形與出現頻率，因此只能抽取一小部分出現頻率高的詞彙知識。與統計式迥異的另一種作法是利用語言知識擷取更多語言知識。例如以機讀英漢電子辭典中所列的翻譯當基礎去找尋平行語料庫中的實際翻譯。其作法又可以分為精確匹配與部分匹配兩種，前者只能找到很有限的翻譯對應，而後者雖可找到較多的翻譯對應，但其中與上下文不符合造成對應錯誤的情形相當多。無論是利用統計或機讀電子辭典從中英平行語料庫自動擷取雙語詞彙知識的困難在於翻譯並非一對一對應，而是隨著上下文語境而變化。本研究探討如何有效地結合統計與語言知識擷取更多的雙語詞彙知識。

三、結果與討論

本研究首先測試並改良 Fung 與 Church (1994) 的演算法。Fung 與 Church (1994)提出 K-vec 演算法結合互見訊息(mutual information)與 t 值(t-score)等兩個統計方法來計算兩個詞在文件內部區段的共現關聯性。

互見訊息(mutual information)是訊息理論(information theory)中的基本概念，計算的方式是兩個事件共同出現的機率除以個別事件出現的機率的積再取以二為底的對數。

$$MI(x,y)=\log_2 \frac{P(x,y)}{P(x)P(y)}$$

如果只考慮緊鄰的兩個詞，則可代入下列公式。其中 N 代表總詞數，f(x,y) 代表 x 與 y 一起出現的次數 f(x), f(y) 分別代表 x 出現的次數與 y 出現的次數。

$$MI(x,y)=\log_2 \frac{\frac{f(x,y)}{(N-1)}}{\frac{f(x)}{N}\times\frac{f(y)}{N}} \cong \log_2 \frac{N\times f(x,y)}{f(x)f(y)}$$

Ken Church (1991)與他的同事率先提出以互見訊息計算詞與詞之間的關聯性(word association)。互見訊息值越高表示詞的關聯性越高，當語料庫夠大時，而互見訊息值大於 1.65，表示這兩個詞常常一起出現，且很可能是搭配語(collocations)，成語，或常見的人名，地名。利用互見訊息可以從中文語料庫中自動抽取詞彙。互見訊息可以視為一種相似度測量，T-值則可以視為相異度的測量。T-值(t-score)是計算語言學中常用的統計顯著性的檢定(statistical significance test)，也是 Ken Church (1991)與他的同事率先運用在計算語言學，通常與互見訊息搭配一起使用。T-值與標準差和信賴區間(confidence interval)密切相關。當語料庫夠大時，而 T-值大於 1.65 時表示有 95%的信心證明差異是存在。計算 T 值的公式如下。

$$t=\frac{P(x|y)-P(x|z)}{\sqrt{\sigma^2 P(x|y)+\sigma^2 P(x|z)}}$$

其中 x|y 表示 y 出現時 x 出現的機率。 σ 表示標準差。
T 值的計算可以採用下列簡化的公式。

$$t\approx\frac{f(x,y)-\frac{f(x)f(y)}{N}}{\sqrt{f(x,y)}}$$

Fung 與 Church (1994)的基本的假設是如果有兩篇互相對應的翻譯文章，某語言一個詞與另一個語言的一個詞在某些區段一起出現的機率大於個別出現的機率，則它們兩個詞有可能是翻譯等同(translation equivalents)。他們將相對應的翻譯文章均分為 K 個區段 (K 為文章長度的平方根)，以 K 維向量來紀錄兩個語言中每個詞出現在哪幾個區段，例如在第一區段出現就將對應的向量值設為 1，否則設為 0。假設將某一篇文章分成 5 個區段，某個詞出現在第一與第四個區段，其分佈區段的向量表示法為 (1,0,0,1,0)。由於 Fung 與 Church (1994)的演算法計算兩個語言中的詞在某一區段一起出線的機率，詞頻太低與太高的詞都不適合使用此演算法，因為若只出現一兩次的詞即使分佈區段完全相同也很有可能是巧合，而出現很頻繁的詞很可能是功能詞才會在很多區段一起出現，這些都必須先排除掉，否則會影響演算法的精確度。Fung 與 Church 建議使用詞頻在 5 次到 10 間的詞，以 K 維向量來表示分佈其情形之後再利用互見訊息與 t 值來計算頻率相近的中文與英文詞在相同區段一起出現的機率。Fung 與 Church (1994)使用下列聯方表。

a = k(A B)	b = k(~ A B)
c = k(A ~B)	d = k(~A ~B)

a 表示某個中文詞與英文詞一起出現的區段數， b 表示英文詞出現但中文詞沒有出現的區段數， c 表示中文詞出現但英文詞沒有出現的區段數， d 表示中文詞與英文詞都沒有出現的區段數。再利用下列稍微修改過的互現訊息與 t 值，其中 $P(V_c)$ 為某一中文詞出現在區段的機率，$P(V_e)$ 為某一英

文詞出現在區段的機率。

$$MI(V_c, V_e) = \log_2 \frac{P(V_c, V_e)}{P(V_c)P(V_e)}$$

$$P(V_c) = \frac{a+b}{a+b+c+d}$$

$$P(V_e) = \frac{a+c}{a+b+c+d}$$

$$t(V_c, V_e) = \frac{P(V_c, V_e) - P(V_c)P(V_e)}{\sqrt{\frac{P(V_c, V_e)}{K}}}$$

我們使用中研院詞知識小組發展的分詞程式處理中文的分詞，並以中英對照之光華雜誌做實驗證明上述方法的精確度受到文類的影響很大，精確度有可能高至 70%也有可能低至 30%以下，此外利用此演算法實際能找到的對應詞相當有限。為了提高精確度，我們改良 Fung 與 Church (1994)演算法。首先我們計算中文與英文的文章段落數目是否一樣, 若一樣我們則將 K 設為段落數，若不一樣則依舊採用原先的定義。我們也合併賓州大學（University of Pennsylvania）與美國暑期語言學院 (Summer Institute of Linguistics) 所發展的構詞分詞程式，將英文所有的名詞，動詞，形容詞詞換成原型，使計算共現機率時能更準確。此外所我們不僅利用文章內部區段一起出現的機率，也收集數十篇對應文章，再以中文詞與英文詞出現在同一篇文章的機率（亦即文件的共現關聯性）來過濾 Fung 與 Church (1994)演算法所得到的結果，將精確度大幅提高至 90%以上。除了改良 Fung and Church (1994)的 K-vec 演算法，我們也利用機讀英漢電子辭典得到部分翻譯對應，以便推論其它的翻譯對應。一般人認為以機讀英漢電子辭典即可以很容易得到平行語料中的詞彙對應，事實上撰寫程式呼叫電子辭典自動查詢所得到的對應仍然非常有限。主要原因在於（1）一個詞可能有幾個翻譯，以機讀辭典判斷那一個詞對應哪一個詞，必須從上下文找訊息，相當困難，從實驗中我們發現功能詞的意義相當多，利用機讀辭典來得到翻譯翻譯對應最不可靠。(2) 利用精確字串匹配（exact string match）所能得到的翻譯對應相當有限。例如字典中 teacher 的翻譯是「教師」，實際上文章可能翻譯成「老師」，若採用部分字串匹配 (partial string match) 則可以找到辭典中的翻譯與文章的翻譯有一個字「師」相同。採用部分字串匹配雖可以找到相當多可能的翻譯對應，但錯誤率也相對提高許多。為了解決這個問題，我們先排除最常出現的功能詞「的」。凡是部分字串匹配為「的」的翻譯對應一律排除。接著再抽取翻譯文章中都至少包含一個中文字與辭典的翻譯相同的相鄰兩個英文詞。例如 *deep roots* ：根深蒂固，*academic world* ：學術界，*eldest son* ：長子，*teaching materials* ：教材 *oral exam* ：口試，*research room* ：研究室。雖然利用上述緊鄰(proximity)原則找到的詞組翻譯 (translation equivalents)相當有限，但精確度高達 90%以上。我們利用這些正確率非常高的對應詞或詞組即可得到某些翻譯句的對應。我們再使用中研院的詞

類標記程式標示中文句子每個詞的詞類，並使用 Eric Brill 的英文詞類標記程式與 Steven Abney 的英文句法剖析器。Abney 的程式可以將標好英文詞類標記的句子剖析其大部分的句法結構，但對於介詞組是否修飾名詞或動詞等較困難的問題則不做分析。我們根據中研院的詞類標記程式得到的結果將中文句子分成主詞與動詞組兩部分，並依據 Abney 程式的結果找出英文句子主詞與動詞組兩部分。接下來我們比對機讀辭典與文章翻譯是否有部分字串吻合，若有則利用兩個翻譯的句法結構大致對應，找出對應的詞組。這些被自動抽取出來的詞彙與詞組可以用於自動編纂中英辭典，機器翻譯，與跨語言資訊檢索。

五、參考文獻

Chang, J-S. and Chen, H.-C. (1994) "Using Part of Speech Information in Word Alignment." In Proceedings of the Annual Conference of American Machine Translation Association, pp. 16-23.

Chen, K.-H. and Chen, H.-H. (1995) "Aligning Bilingual Corpus: Especially for Language Pairs from Different Families." Information Sciences, Vol. 4, pp. 57-81.

Dagan, I., Church, W, and Gale, W. (1993) "Robust Bilingual Word Alignment for Machine Aided Translation." In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, pp. 1-8, Ohio.

Dagan, I. (1996) "Bilingual Word Alignment and Lexicon Construction." Tutorial paper given at the International Conference on Computational Linguistics,

Copenhagen.

Debili, F. et al. (1994) "Using Syntactic Dependencies for Word Alignment."

Proceeings of the 4th Conference on Applied Natural Language Processing, pp.

188-189.

Fung, P. and Church, K. (1994) "K-vec: A New Approach for Aligning Parallel Texts." Proceedings of the International Conference of Computational Linguistics, pp.1096-1102, Kyoto.

Fung, P. and KcKeown, K. (1994) "Aligning Noisy Parallel Corpora Across Language Groups: Word Pair Feature Matching by Dynamic Time Warping." Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation, pp. 81-88.

Fung, P. and Wu, D. (1995) "Coerced Markov Models for Cross-Lingual Lexical-Tag Relations." Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation, pp. 240-255.

Fung, P. and McKeown, K. (1997) "A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups." Machine Translation, Vol. 12, Nos. 1-2., pp. 53-87.

Gale, W. and Church, K. (1993) “A Program for Aligning Sentences in Bilingual Corpora.” Computational Linguistics, Vol. 19, No 1, pp 75-102.

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

Jones, D. and Somers, H. (1995) “Bilingual Vocabulary Estimation from Noisy Parallel Corpora Using Variable Bag Estimation.” In JADT III GiornateInternazionali di Analsi Statistica dei Dati Testuali, pp. 255-262, Rome.

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

Kumano, A. and Hirakawa, H. (1994) “Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistic Information.” in Proceedings of International Conference on Computational Linguistics, pp. 76-81, Kyoto.

Kupiec, J. (1993) “An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora.” Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 17-22, Ohio.

Lee, H.-J. and Li, H.-W. (1995) “Structure-Based Word Alignment in Bilingual Corpus.” In Proceedings of the International Conference on Computer Processing of Oriental Languages, pp. 267-274, Haiwaii.

Matsumoto, Y. et al (1993) “Structural Matching of Parallel Texts.” Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 23- 30, Ohio.

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.

Smadja, F. and Kathleen, M. and Hatzivassiloglou, V. (1996) “Translating Collocation for Bilingual Lexicons: A Statistical Approach.” Computational Linguistics, Vol. 22, No 1, pp. 1-38.

Somers, H. and Ward, A. (1996) “Some More Experiments in Bilingual Text Alignments.” In Oflazer, K. and Somers, Second International Conference on New Methods in Language Methods in Language Processing, pp. 66-78, Ankara.

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. and Xia, X. (1995) “Large-Scale Automatic Extraction of an English-Chinese Translation Lexicon.” Machine Translation, Vol. 9, pp. 285-313.