

行政院國家科學委員會專題研究計畫成果報告

智慧型知識擷取技術與應用研究(II)：語料庫之設計與製作(II)

Design and Implementation of Corpora (II)

計畫編號：NSC 87-2213-E-002-023

執行期限：86年8月1日至87年7月31日

主持人：陳光華 國立臺灣大學圖書資訊學系

一、中文摘要

雙語語料庫帶有許多語言的訊息，因而有許多可能的應用，例如，詞彙的多義校正、翻譯樣版的抽取、名詞複合詞的自動翻譯，及雙語詞典的建立。前人的研究很少觸及不同語系的平行語料，本研究提出一些方法，以建立詞彙的對列。實驗語料主要是 ROCLING 語料庫中的 HP 與 Lotus 中英雙語語料，以及 NTU 中英雙語語料庫。本研究提出三種語言模型，基本上每個模型皆包括兩部份。第一為初步找出句子對列完成的雙語語料中相對應的中英文詞；第二為解決二個以上英文詞對應同一個中文詞的情形。系統的評估標準為精確率與增加率。

關鍵詞：詞彙對列、雙語語料庫、自然語言處理

Abstract

Bilingual corpus carries many kinds of linguistic knowledge such that they can be used in word-sense disambiguation, extracting translation templates, finding bilingual collocations, automatic translation in noun compounds, building bilingual dictionary, and so on. To do such kinds of applications, the most important task is to align the bilingual texts. To align a text means to show which parts of the first language correspond to which parts of the second language. In this study, an approach for word alignment in English-Chinese corpus is presented. Previous works on aligning words seldom touch the texts in different language families, like English and Chinese. Our experimental material consists

of two corpora: ROCLING Text Corpus and NTU Bilingual Corpus.

Three language models are proposed to do word alignment in this study. Theoretically, the matching procedure will initially align the English word to its Chinese counterpart if it appears in the corresponding Chinese sentence, and then resolving conflicts to make no different English words are corresponded to the same Chinese word in the corresponding Chinese sentence. Precision and augmentation are used to evaluate the system performance.

Keywords: Word Alignment, Bilingual Corpus, Natural Language Processing.

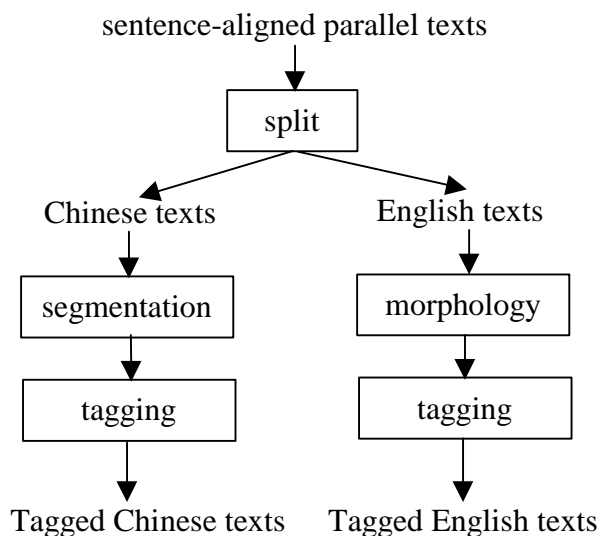
二、計畫緣由與目的

近年來，雙語語料庫已成為一種不可或缺的重要研究資源，這是因為雙語語料庫比單語語料庫帶有更多的語言訊息，因而可以用於多方面的應用。例如，詞彙的多義校正、翻譯樣版的抽取、名詞複合詞的自動翻譯，及雙語詞典的建立等等。但在做各種應用之前，雙語語料必須先作對列。所謂對列，就是找出第一種語言的某一部分是由第二種語言的那一部份所翻譯過來的。根據對應的單位來分，可以分為段落、句子、和詞彙三種對列。由於目前較少研究在探討詞彙的對列問題，且大部份都是針對同語系作實驗，少有跨越不同語系進行詞彙對列的研究。然而這項研究卻是其他後續研究的重要基礎，有了詞彙的對應關係，可以做進一步較複雜的應用，例如詞彙的增刪研究和建立雙語詞典。所以，本研究便是在探討如何在英中雙語語料庫上做詞彙的對列。

事實上，這研究主題存在著許多困難處，因為在翻譯時很少是刻板的一個字一個字翻譯，而是常會多加些字或少些字以使詞句更通暢，加上中文跟英文是不同語系的語言，他們的文法及句子的結構也有很大的差異時，因此無法做到每個詞彙的對應，只能做到詞彙的部份對應。另一方面，翻譯過後的詞彙之間次序也可能調換，因此，翻譯的自由度影響了詞彙的對列工作難易度。

三、研究方法

在真正去找出詞彙之間的對應關係前，必須先對實驗的平行文件作些前處理，以便利爾後工作的進行。前處理包括了下列幾種工作：切分的工作是把平行文件分開到兩個檔案。這兩個檔案雖是相互獨立，卻仍保有句子相對應的關係。斷詞的工作便是將中文文章中一連串的字元，切分出適當的詞彙。字根還原的工作則是將英文詞彙還原為原形。處理對象包括動詞的時態、形容詞的比較級與最高級、及名詞的單複數。詞類標示的工作便是分別對中文與英文文章的詞彙標示詞類。整個流程如圖一所示。



圖一、前處理流程

由於文章的翻譯手法對於詞彙的對列工作影響很大，所以我們選擇二種不同風格的雙語語料作為實驗的材料。ROCLING 語料庫中的文章較偏逐字翻譯，其中的雙

語語料度部份是電腦手冊。NTU 雙語語料則是由光華雜誌選出，翻譯較偏向意譯。

一個詞彙的翻譯通常有很多個，而其中又只有幾個常被使用。如果把詞彙換成詞類來看，詞類間的翻譯也同樣是這種情形，例如動詞經過翻譯後經常還是動詞，而不會變為是代名詞 因此，我們便從 BDC II 詞典訓練出詞類翻譯的關係。其中一個訓練表記載著某個英文詞類可能翻譯成哪些中文詞類。另一個訓練表則記載某個中文詞類翻譯成不同英文詞類的機率。

由於某個英文詞彙有可能對應到中文的任何一個詞彙，因此，為了避免盲目的配對，必須先估計到底是哪些配對的機率比較高，我們使用 ϕ^2 統計值估算詞彙的相關程度，其公式如下所示：

$$w^2 = \frac{(ad - bc)^2}{(a + b) \times (a + c) \times (b + d) \times (c + d)}$$

$$a = f(\text{word 1, word 2});$$

$$b = f(\text{word 1}) - f(\text{word 1, word 2});$$

$$c = f(\text{word 2}) - f(\text{word 1, word 2});$$

$$d = L - a - b - c.$$

L 是訓練語料庫的句子總數， a 、 b 、 c 、 d 的關係請參考圖二。

	Chinese Sentence	
English Sentence	a	b
	c	d

圖二、 ϕ^2 值計算表

我們進一步使用二個過濾程序，以消除不必要的 ϕ^2 計算。第一是使用詞類訓練表過濾不可能互為翻譯的詞類配對；第二是利用頻率門檻過濾頻率過小的配對。通過上述過濾程序的配對，才真正進行 ϕ^2 值的計算。

另外必須要考慮的問題是連語的現象。由於翻譯時經常有幾個詞彙被翻成一個詞彙，也就是把好幾個詞彙當成一個語意單位進行翻譯，這些詞彙構成了所謂的連語。我們使用 N-Gram 的方法分別對中文及英文抽取連語，將這些連語視為一個詞彙，即可採用前述相同的程序進行雙語詞彙對列的工作。

四、研究成果

我們提出三種模型，分別稱之為 Model-I、Model-II、Model-III，以下分別說明這三種不同的語言模型。

Model-I 使用三個前處理程序：切分字元、斷詞、字根還原，並且僅利用頻率過濾程序，最後建立 ϕ^2 值訓練表。其演算法如下：首先對英文句子中的每一個詞彙，由 ϕ^2 訓練表找出第一個中文翻譯的候選詞彙，並檢查是否出現於相對應的中文翻譯句子，如果有則這英文詞彙和它的第一個中文翻譯候選詞彙便互相對應，反之則英文詞將沒有對應的中文翻譯。接著必須檢查是否有衝突的情形出現，也就是是否有二個以上的英文詞彙對應同一個中文詞彙，這時使用 ϕ^2 值最大的配對作為對譯的詞彙。其餘的英文詞彙再回去察看是否有其他的中文對譯候選詞彙，如果有就依照前述的方法進行對譯的處理。

Model-I 只考慮每個英文詞彙的第一個中文翻譯，但是英文詞彙的中文對應詞彙通常不只一個，Model-II 修正了前述的作法。對於每個英文詞彙而言，它的一整串候選詞彙都是可能的翻譯，所以 Model-II 依序在這串候選詞彙中找出第一個出現於相對應的中文翻譯句子中的中文詞彙，至於解決衝突的方法與 Model-I 相同。此外，Model-II 除了使用頻率過濾前處理程序，更加入詞類標示前處理程序，過濾不必要的詞彙配對計算。

Model-III 修正 Model-I、Model-II 使用的衝突解決程序，除了使用 ϕ^2 值作為處理衝突的依據，另外使用詞類翻譯機率值。亦即，利用 ϕ^2 值與詞類翻譯機率值的乘積作為判斷的依據。

實驗的語料分別是 ROCLING 語料庫中的 HP 與 Lotus 中英雙語語料，以及 NTU 中英雙語語料。我們分別使用精確率（Precision）以及增加率（Augmentation）作為實驗結果的評估準則，所謂精確率是指最後詞彙對列結果的正確程度；所謂增加率是指在這些詞彙的對應中有多少是 BDC II 辭典沒有的翻譯。Model-I、

Model-II、Model-III 的實驗結果如表一、表二、表三所示。

表一、Model-I 實驗結果

	HP Corpus	Lotus Corpus	NTU Corpus
Precision	73.01%	73.92%	54.17%
Augmentation	51%	59%	18%

表二、Model-II 實驗結果

	HP Corpus	Lotus Corpus	NTU Corpus
Precision	76.66%	76.12%	56.09%
Augmentation	49%	57%	20%

表三、Model-III 實驗結果

	HP Corpus	Lotus Corpus	NTU Corpus
Precision	78.17%	77.58%	57.34%
Augmentation	48%	56%	19%

參照實驗結果，我們嘗試探討是哪些因素造成誤差。NTU 雙語語料庫的文章大多數是意譯，很難明確指出詞彙與詞彙之間的翻譯關係，而這也是造成 NTU 雙語語料庫平均精確率並不是很高的主要原因。前處理的斷詞、字根還原、詞類標示也是誤差的另一來源。尤其中文斷詞難度較高容易造成錯誤的情形，例如人名、組織名、地名並不容易處理，這也是中文處理的一大難題。此外，當某個英文詞彙的中文翻譯僅出現一次，可能造成此中文翻譯被頻率過濾程序過濾，或是僅有很低的 ϕ^2 值，以至於在對列時無法相互對應。還有一個情形是，當英文句子中出現二個以上的相同英文詞彙時，我們假定它們使用相同的中文翻譯，但是實際上，還是會發生這些相同英文詞彙使用不同中文詞彙的情形。

五、結論

本研究提出一個可適用於不同語系的文件詞彙對列方法，並應不同的考慮因素，進一步區分三種語言模型。由於實驗文件各有其特性，而表現出不同的實驗結

果，但是整體而言，仍以 Model-III 表現最佳，對於逐字翻譯的雙語語料，其精確率約為 78%，而增加率約為 52%；對於意譯為主的雙語語料，其精確率約為 57%，而增加率約為 20%。

至於未來的研究方向，可以建立雙語辭典，或是將辭典中缺少的對譯詞彙列入。然而為了解決多詞對應的現象，應發展更好的演算法，以期更正確地抽取連語，並併入語言模型。另一方面，以完成的詞彙對列雙語語料庫，可以作為其他研究的實驗材料，例如，詞彙的增刪研究。所以，如何使詞彙的對列正確率更高，是一項重要的研究課題。

六、參考文獻

- BDC (1990). The DBD Chinese-English Electronic Dictionary (version 2.0), Behavior Design Corporation.
- Brown, P. *et al.* (1991a). Aligning Sentences in Parallel Corpora. In Proceedings of 29th Annual Meeting of the ACL, (1991): 169-176.
- Brown, P. *et al.* (1991b). Word-Sense Disambiguation Using Statistical Methods. In Proceedings of the 29th Annual Meeting of Association for Computational Linguistics, (1991): 264-270.
- Chen, K.H. and H.H. Chen (1994). Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation. In Proceedings of the 32nd Annual Meeting of ACL, (1994): 234-241.
- Chen, K.H. and H.H. Chen (1994). A Part-of-Speech-Based Alignment Algorithm. In Proceedings of the 15th International Conference on Computational Linguistics, (1994): 166-171.
- Chen, K.H. and H.H. Chen (1995). A Corpus-Based Approach to Text Partition. In Proceedings of the Workshop of Recent Advances in Natural Language Processing, (1995): 152-160.
- Church, K.W. and P. Hanks (1990). Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16:1, (1990): 22-29.
- Gale, W. and Church K. (1991). Identifying Word Correspondences in Parallel Texts. In Proceedings of The Fourth DARPA Workshop on Speech and Natural Language, (1991): 152-157.
- Liao, F.H. (1995). A Study on the Word Alignment Problem in English-Chinese Corpora, Master Thesis, Department of Computer Science and Information Engineering, National Taiwan University.
- Lin, C.Y. (1995). Knowledge-Based Automatic Topic Identification. In Proceedings of the 33rd Annual Meeting of ACL, (1995): 308-310.
- Meyers, A., R. Yangarber, and R. Grishman (1996). Alignment fo Shared Forests for Bilingual Corpora. In Proceedings of the 16th International Conference on Computational Linguistics, (1996): 460-465.
- Salton, G. (1986). On the Use of Term Associations in Automatic Information Retrieval. In Proceedings of the 11th COLING, (1986): 380-386.
- Wu, D. (1995). An Algorithm for Simultaneous Brackets Parallel Texts by Aligning Words. In Proceeding of the 33rd Annual Meeting of the Association for Computational Linguistics, (1995): 244-251.
- Yamashina, M. and S. Obashi (1988). Collocational Analysis in Japanese Text Input. In Proceedings of the 12th COLING, (1988): 770-772.
- Zhai, C. and D. Evans (1996). Noun Phrase Analysis in Large Unrestricted Text for Information Retrieval. In Proceedings 34th Annual Meeting of ACL, (1996): 17-24.