智慧型知識擷取技術與應用研究-子計畫二：
語言知識擷取技術：述語參數結構自動擷取系統之研製
An Extraction System of Predicate-Argument Structures

計畫編號：NSC 88-2213-E-002-034-
執行期限：87 **年** 8 **月** 1 **日至** 88 **年** 7 **月** 31 **日**
主持人：陳信希　　國立台灣大學資訊工程學系

一、中文摘要

述語參數結構的擷取有許多應用，例如減少剖析過程中所產生的歧異；建立機器翻譯系統述語參數結構轉換規則等等。本計畫提出一個述語參數結構自動擷取系統，包括規則式的名詞片語擷取器，以減少句子的變異性；有限狀態機制，以抽取動詞後面的語法成分資訊。然後根據所獲得的資訊，產生所有可能的述語參數結構，並且從樹狀語料庫中抽出語言知識來幫助決定。同時也利用樹狀語料庫來評估，並提出一些改進的方法。

關鍵詞：語料庫、空詞、語言模型、機器翻譯、自然語言處理、名詞片語、述語參數結構

Abstract

The extraction of predicate argument structures has many applications. For example, reducing the ambiguities in parsing, setting up transfer rules for machine translation systems, *etc*. This project presents an automatic extraction system for predicate argument structures. It includes a noun phrase extractor to reduce the variation of sentences and a finite state mechanism to extract constituents after verbs. We generate possible candidates and employ the information trained from treebank to select the best predicate argument structure. Besides, we also evaluate the strategies based on treebank.

Keywords: Corpus, Empty Constituents, Language Model, Machine Translation, Natural Language Processing, Noun Phrases, Predicate-Argument Structures.

二、緣由與目的

擷取述語參數結構對於語言分析及產生有很大的幫助。在剖析句子時，它能減少歧異樹的個數；在以轉換為本的機器翻譯系統中，首先要知道不同語言間的述語參數對應規則。傳統上是以人工在詞典中加入編譯述語參數結構資訊，但是這種作法不但費時，也容易產生不一致性，並且對於每一種參數結構也沒有統計資訊，所以這種詞典用處不大。賓州大學所發展的 PENN **樹狀語料庫有許多特性，使得它適合用來抽取述語參數結構。我們利用它做為訓練資料，發展一個述語參數結構自動擷取系統，並且利用它來評估不同策略。**

Brent (1991) **是最早做這方面研究的人。他用非常簡單的文法去抽取五種引數結構，並且依賴句子中很明顯的線索（如代名詞、專有名詞）以達到非常高的正確率。很明顯地，這種方法無法適用更多的引數結構。**Manning (1993) **用一個有限狀態剖析器及一個過濾器來製作一部動詞參數結構的詞典。**Chen and Chen (1993) **利用一個機率式的部份剖析器以及一個有限狀態機制決定述語參數結構。他們根據** OALD **字典，定義** 23 **種不同的參數結構。他們所用的策略是最長匹配法，並且他們沒有樹狀語料庫做評估。**

研究中的困難處在於首先必須確定動詞後的語法成分的左右邊界，然後決定那些成分併接在 VP**，並且對於併接在** VP **的成分，區分是參數還是修飾語。第二，必須決定句子的省略成分以及位置，以得到正確的參數結構。最後，必須考慮動詞後**

每個成分所扮演的語意的角色。

三、結果與討論

3.1 PENN 樹狀語料庫

於 PENN 樹狀語料庫利用空詞來標記句中的省略成分，利用功能詞類來標明成分所扮演的語意角色，並使用虛擬併接來註明不連續成分間的關係，使得它適合用來抽取述語參數結構。在 PENN 樹狀語料庫中，除了 VP 之外的述語，均用功能詞類 -PRD 來指明。本計畫只考慮動詞的內部參數結構。

我們使用 PENN 樹狀語料庫中的 Wall Street Journal 語料庫，做為訓練與測試資料。Wall Street Journal 語料庫共分成 25 節（00~24）。02 至 22 節做為訓練語料庫，共有 41,532 個句子。00, 01, 23 以及 24 節作為開放測試（open test），02 以及 10 節作為封閉測試（close test）。

3.2 基本系統架構

本計畫提出的述語參數結構自動擷取系統基本架構包含了一個規則式的名詞片語擷取器，一個有限狀態機制，以及一個從樹狀語料庫中訓練出來的述語參數結構字典（以下簡稱 PAS 字典）。

名詞片語擷取器主要是用來減少句子的變異性，使我們可以很容易的判別動詞後面的語法成分；也就是說，我們用它來將可能的片語結構切分出來，使得句子變成一串線性單元結構序列。

我們用一個有限狀態機制來得到動詞後面的語法成分；基本上，是以 Chen and Chen (1994)所提出的有限狀態機制為基礎，再參考從樹狀語料庫所訓練出來的 PAS 字典而得到的。

我們從樹狀語料庫中訓練出每個動詞可能的用法，而形成詞典。我們從訓練語料庫中訓練出 104,588 個述語參數結構，整理後，共有 3,546 個不同的動詞。

現以一個例子說明有限狀態機的運作過程，當遇到一個動詞，就進入 initial state，當動詞後面的詞類為 JJ 或 JJR，就進入 ADJP state，此時狀態圖無任何的外向邊，所以會傳回 ADJP 的文本訊息；若動詞後面的詞類為 RP，就進入 PRT state，此時若後續的字是以 wh 開頭或是 that，就會進入 PRT SBAR state，並且傳回 PRT SBAR 的文本訊息。

假如動詞為被動式，我們只知道直接受詞被省略，但通常不知道省略的位置；所以，有以下兩個問題：

(1) 如何去判斷一個動詞是否為被動式？

(2) 知道動詞為被動式之後，在何處插入省略的受詞？

對於(1)，我們用 be + VBN 來判別被動式；對於(2)，我們假設省略的受詞在動詞的後面。

我們認為有限狀態機制的輸出，只是代表動詞後面的語法成分資訊而已，而非是真正的參數結構。我們提出一個「切點位置決定法」來產生所有可能的參數結構，然後再用不同的策略選擇評估。我們利用 PAS 字典所提供資訊，去刪除一些在訓練語料庫中未曾出現的用法，藉以提高正確率。

我們提出三種不同的策略，在可能的參數結構中選出一種，並且利用語料庫做評估。在最短優先策略中，我們假設切點應該盡量靠近動詞。也就是說，動詞後面的語法成分比較傾向為修飾語，而非是參數。在最長優先策略恰好與最短優先策略相反，我們假設動詞後面的語法成分資訊比較傾向為參數，而非修飾語。因為我們的 PAS 字典，帶有動詞每一種參數結構出現的頻率，所以，我們可以假設最常出現的參數結構為真正的參數結構。實驗結果顯示，最高機率優先策略明顯比其他兩種策略好，所以我們選擇以最高機率優先策略來做為我們的基本系統架構，並且提出幾種改進的方法。

3.3 加強型模型

由下述幾個方向提昇效能：

(1) SBAR 的改良：因為省略了關係代名詞，所以我們的有限狀態機制無法得到動詞後面的語法成分資訊為 SBAR，因

此，無法產生正確的參數結構。如何識別句子裡的省略成分，是非常重要的。它會影響我們得到真正的參數結構。

(2) 介係詞片語的改善：我們利用 Hindle and Rooth (1991)所提出的 t_score 來決定 PP 的併接位置。假如併接位置決定在 NP，可能的參數結構個數就會由三個降為兩個。接下來，我們嘗試去區分 PP 是一個修飾語或參數。由於 PENN 樹狀語料庫以功能詞類（DTV, BNF, CLR, PUT)來指明 PP 是一個參數而非修飾語，我們可將這些正例的 verb/prep 配對收集起來，同時收集反例的 verb/prep 配對，然後從正例中刪除掉反例，只要測試句中的動詞與介係詞出現在這個集合中，我們即視 PP 為此動詞的一個參數。

(3) 名詞片語的改善：我們計算動詞 verb 與其後的名詞片語詞首 noun 的 mutual information 的強弱，作為區分 NP 為一個修飾語或是參數的依據。假如 MI>0，我們視 NP 為一個參數；假如 MI<0，我們視 NP 為一個修飾語；否則，仍然以最高機率優先策略來選擇。

(4) Transformation-Based Error-Driven Learning： Tranformation-Based Error-Driven Learning 的技術已經應用在許多自然語言領域的問題上，如詞類標注 (Brill 1992), 剖析 (Brill 1993),PP-attachment (Brill 1994)。我們以 enhanced model II(加入改善 PP 的模型)，作為 initial state，從 initial state 模型所找出的述語參數結構，與樹狀語料庫比較之後，會產生一些轉換規則，這些 tranformation 再應用至 annotated text（述語參數結構），並計算所增加的精確率，得出一條對精確率增加最多的 transformation，加入 ordered transformation list；下一次迭代 (iteration) 的 initial state 就是原來的 initial state 再加 ordered transformation list，如此繼續整個學習過程，直到沒有新的 transformation 可以增加精確率為止。

3.4 實驗結果

我們利用 WSJ corpus Section 02 至 Section 22 作為訓練語料庫。我們總共得到 54 條 transformations。訓練語料庫啟始的精確率為 75.03%，加入這些 transformations 後，精確率上升至 77.19%。

我們另外將動詞的詞類由 6 個 group 至 2 個，即將 VB, VBP, VBZ, VBD 和 VBG 歸類為 VB*，另外一個為 VBN。我們重複相同的實驗，共得到 59 條 transformations，精確率上升至 77.53%。

3.5 討論

在達成的目標上，

(1) 提出一個述語參數結構自動擷取系統。包含一個名詞片語擷取器，以減少句子的變異性；用一個有限狀態機制來得到動詞後面的語法成分；再利用樹狀語料庫評估不同策略的績效。

(2) 提出不同於最長優先策略的的參數結構選擇策略。在我們的實驗裡，利用『切點位置決定法則』，以提出所有可能的參數結構。數據顯示，利用從樹狀語料庫中訓練出來的 PAS 字典的最高機率優先策略，績效比最長優先策略好。

(3) 利用計算 Lexical Association，來決定 PP 的併接位置藉以減少可能的參數結構個數，並利用樹狀語料庫所提供的資訊，區分修飾語與參數。

(4) 我們也將 transformation-based error-driven learning 的技術，應用在我們的加強模型中；實驗結果顯示，應用這種 learning 的技巧，可以彌補系統的績效。

尚待努力的課題有：

(1) 由於動詞後面的語法成分所扮演的語意角色，關係其為修飾語或參數，所以，決定每個語法成分所扮演的語意角色，亦非常重要。

(2) 可以從動詞的語意，來決定其參數個數及型態。

4. 自評

本計畫的研究內容與原計畫完全相符，並已達成預期目標。研究成果具有學術和應用價值，合適於發表論文。與本計畫相關的著作有碩士論文：述語參數結構

自動擷取系統之研製。

五、參考文獻

[1] Aone, C. and D. McKee. (1993). "Acquiring Predicate-Argument Mapping Information from Multilingual Texts." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 107-116.

[2] Baker, J. (1979). "Trainable Grammars for Speech Recognition." *Speech Communication Proceedings for the 97th Meeting of the Acoustic Society of America*, 1979, pp. 547-550.

[3] Bies, A. et al. (1995). *Bracketing Guidelines for Treebank II Style Penn Treebank Project*, 1995.

[4] Brent, M. R. (1991). "Automatic Acquisition of Subcategorization Frame from Untagged Text." *Proceedings of the 29th Annual Meeting of ACL*, 1991, pp. 209-214.

[5] Brill, E. and Resnik, P. (1994). "A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation." *Proceedings of COLING-94*, 1994, pp. 1198-1203.

[6] Brill, E. (1992). "A simple rule-based part of speech tagger." *Proceedings of the Third Conference on Applied Natural Language Processing, ACL,* Trento, Italy, 1992.

[7] Brill, E. (1993). "Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach." *Proceedings of the 31th Annual Meeting of the Association of Computational Linguistics*, 1993, pp.259-265.

[8] Chen, K.H. and Chen, H.H. (1994). "Acquiring Verb Subcategorization Frames." *Proceedings of the Second Conference for Natural Language Processing (KONVENS94)*, Vienna, Austria, September 28-30, 1994, pp. 407-410.

[9] Chen, H.H. and Lee, Y.S. (1995). "A Chunking-and-Raising Partial Parser." *Proceedings of the 4th international workshop on parsing technologies*, 1995, pp.71-78.

[10] Church, K.W. and Hanks, P. (1989). "Word Association Norm, Mutual Information, and Lexicography." *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 1989, pp. 76-83.

[11] Fano, R. (1961). *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, MA.

[12] Hindle, D. and Rooth M. (1991). "Structural Ambiguity and Lexical Relations." *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 229-236.

[13] Hornby, A.S. *Oxford Advanced Learner's Dictionary*, Oxford University Press, 1989.

[14] Manning, C. D. (1993). "Automatic Acquisition of a Large Subcategorization Dictionary from Corpora." *Proceedings of the 31th Annual Meeting of the Association of Computational Linguistics*, 1993, pp. 235-242.

[15] Marcus M. P. et al. (1993). "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics*, Volume 19, 1993, pp. 313-330.

[16] Marcus M. P. et al. (1995). "The Penn Treebank : Annotating Predicate Argument Structure." *The Penn Treebank Project Release II CD-ROM*.

[17] Poznanski, V. and Sanfilippo A. (1993). "Detecting Dependencies between Semantic Verb Subclasses and Subcategorization Frames in Text Corpora." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 82-94.

[18] Pugeault, F. et al. (1994). "Knowledge Extraction from Texts: a Method for Extracting Predicate-Argument Structures from Texts." *Proceedings of COLING-94*, 1994, pp. 1039-1043.

[19] Ushioda, A., et al. (1993). "The Automatic Acquisition of Frequencies of Verb Subcategorization Frames from Tagged Corpora." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 95-106.