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)-
子計畫二：語言知識擷取技術之研究(II)

Technologies for Linguistic Knowledge Extraction (II)

計畫編號：NSC 87-2213-E-002-022
執行期限：86 年 8 月 1 日至 87 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

一般詞彙可根據是否有明顯的符號，粗分為公開的詞彙(overt constituents)及空詞(empty constituents)兩種。公開的詞彙可能是舊詞或新詞(未知詞)，舊詞會有新的用法，新詞則不斷的產生，因此判斷其規類一直是很多自然語言處理系統研究的重點。相對於公開的詞彙，空詞也是自然語言非常普遍的語言現象，中英文都有，但因其沒有明顯的記號，而且散佈在文本中，影響文句的處理非常多，因此如能判斷其類別，就可掌握其存在位置，以幫助後續的運作。本子計畫針對這兩種類型的詞彙提出自動分群的方法。

關鍵詞：分群法，語料庫，空詞，語言模型，自然語言處理，公開的詞彙

Abstract

There are two kinds of constituents - say, overt constituents and empty constituents. Overt constituents can be old words or new words. Old words may have new usage and new words are created dynamically. Thus to identify the use of a word is a research topic in natural language processing systems. In comparison to overt constituents, empty constituents do not have overt markers. They are distributed in natural language texts. How to identify their existence and their positions is also a problem. This project investigates the clustering methods to recognize these two kinds of constituents.

Keywords: Clustering, Corpus, Empty Constituent, Language Model, Natural Language Processing, Overt Constituent

二、緣由與目的

一般詞彙可根據是否有明顯的符號，粗分為公開的詞彙(overt constituents)及空詞(empty constituents)兩種。公開的詞彙可能是舊詞或新詞(未知詞)，舊詞會有新的用法，新詞則不斷的產生，因此判斷其規類一直是很多自然語言處理系統研究的重點。相對於公開的詞彙，空詞也是自然語言非常普遍的語言現象，中英文都有，但因其沒有明顯的記號，而且散佈在文本中，影響文句的處理非常多，因此如能判斷其類別，就可掌握其存在位置，以幫助後續的運作。本計畫針對這兩種類型的詞彙提出自動分群的方法。

以下對空詞部分的研究詳細說明，它的重要性可以從許多自然語言處理的應用上看出，謹列舉以下數個例子，作為參考：

(1) **改善剖析器的效能**
如果句子中一些成分被移走或除去，剖析規則可能誤用或無法使用。

(2) **對指涉分析**(anaphor resolution)**有很大的助益**
指涉分析是語言處理一個重要問題，如果指稱詞(anaphor)**被省略，我們怎麼能進一步預測** antecedent **的位置？**

(3) **協助自然語言理解**
自然語言主語和賓語常被移走或除去，這會造成句子理解的困擾。

(4) **機器翻譯系統的改進**
原始語言的空詞在機器翻譯時，

可能必須以明顯成分替代

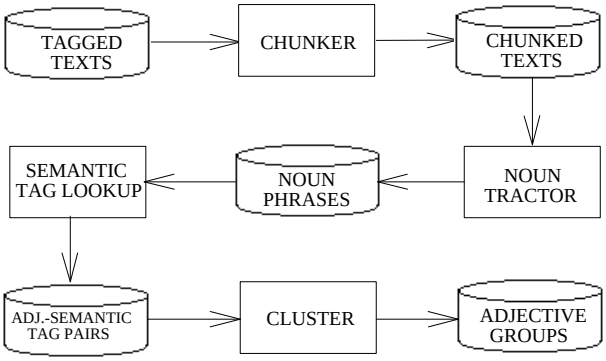
(5) 幫助辨識述語參數結構
在擷取述語參數結構時，空詞會影響參數個數的判定。

(6) 檢索系統中關鍵詞更正確的統計
傳統檢索系統只計算公開的詞彙出現的頻率，其實空詞也會影響統計值。

三、結果與討論

3.1 公開詞彙的辨識與分類

報告中以形容詞為例說明，圖一是實驗流程。我們使用上一年度計畫的執行結果，先找出名詞片語。接著查詢名詞的語意類別(如表一)，討論名詞語意和形容詞的關係



圖一・實驗流程

表一・語意類別

CLASS	SECTION	TAG
ABSTRACT RELATIONS	Existence	1 - 8
	Relation	9 - 24
	Quantity	25 - 57
	Order	58 - 83
	Number	84 - 105
	Time	106 - 139
	Change	140 - 152
	Causation	153 - 179
INTELLECT	Formation of Ideas	450 - 515
	Communication of Ideas	516 - 599
VOLITION	Individual	600 - 736
	Intersocial	737 - 819
ASS	SECTION	TAG
SPACE	In General	180 - 191
	Dimensions	192 - 239
	Form	240 - 263
	Motion	264 - 315
MATTER	In General	316 - 320
	Inorganic	321-356
	Organic	357 - 449
AFFECTIONS	In General	820 - 826
	Personal	827 - 887
	Sympathetic	888 - 921
	Moral	922 - 975
	Religious	975 - 1000

我們的實驗題材有 9123 不同的形容詞和 1001 個可能的名詞語意分類。他們的關係可視為一個 9123×1001 的矩陣(OM)，矩陣裡的數值代表形容詞和名詞語意類別的共現頻率。(1)定義 BM 矩陣，形容詞的相似性，以(2)來計算，值越高代表兩個列向量越相似。根據這數值，產生一個最佳的序列，稱為 ORIS。下個階段，就按照 ORIS 和既定的門檻值來取得分類結果。門檻值決定我們要分多少類，而(3)敘述的資訊流失來判斷是否形成一類。

(1) For each entry (i,j) in OM, if $OM(i,j) > 0$, then set $BM(i,j)$ to 1. Otherwise, set $BM(i,j)$ to 0.

(2) $$Sim(RV_{r_i}, RV_{r_j}) = 1001 - \sum_{k=1}^{1001} RV_{r_i}[k] \oplus RV_{r_j}[k]$$

(3) $$IL_k = \sum_{i=1}^{m_k} \sum_{j=1}^{1001} abs(RV_{ORIS_{m+i}}[j] - RV_{RC_k}[j])$$

3.2 空詞辨識與分類

以下是空詞辨識與分類系統的簡述：

步驟一：收集所有只含一個省略成分的 null fragment ，並褪去所有的 function tags 及刮號，如此所得的集合為 T-1。

步驟二：定義關鍵字。

步驟三：在T-1中，若某字為關鍵字，保留該字，否則保留詞類。這樣便形成了T-2。將T-2編碼，並去除雜訊，這就是T-3。

步驟四：將T-3分類。

步驟五：計算各類別中兩兩的LCS，其結果放在LCS-1中。

步驟六：為LCS-1再算一次LCS，得到LCS-2。

步驟七：濾掉LCS-2中出現頻率較低的以形成LCS-3。

步驟八：以LCS-3為規則，回授給訓練資料，去除低正確率者，正式的規則於是找到。

步驟九：為正式的規則找出使用限制便得到LCS-4。

步驟十：將LCS-4**再回授一遍，同樣的，去除低正確率者。最後，便可得到** LCS-5。

步驟十一：以LCS-5**為規則來找出省略的成分。**

我們使用 Penn Treebank **中** WSJ **的部份來做訓練及測試。其中約五分之一為測試資料，而其餘的五分之四為訓練資料。測試資料有** 8206 **句。我們僅使用含有單一省略的** null fragments **來做訓練，這是為了避免訊息的相互干擾。**

依序取出句中的詞類用以做為訓練字串是較可行的方法，因為如果選的是字，那麼幾乎一句就是一個類型，歸納不出個所以然來。但是有些字對於省略成分的辨識是相當重要的，若只保留詞類，這些線索將隱晦不見。因此，我們定義了一組關鍵字，若該字是關鍵字則保留字，否則就保留詞類。

形容詞與副詞非但不是有用的線索，更甚至會對我們的辨識形成干擾，因此我們將之去除。另外，連續的名詞也被合併起來。我們的訓練字串即 single null fragment**，經過字/詞類的選擇後，去除雜訊並將之編碼，便可用分類方法加以分類。對每個分好的類別計算其元素兩兩之間的** LCS**。其實在分類時為了算出距離，已將兩兩的** LCS **算出，此刻僅需直接取用即可。對這些兩兩算出的** LCS**，再一次地兩兩計算** LCS**，如此所得的樣式即是我們想要的規則。我們將得到的規則與欲得知空詞位置的字串一一循序對應，一旦規則完全對應上，便可對空詞的位置加以預測。由於我們的規則是計算** LCS **得到的，所以規則中緊鄰的字元在原始句中未必緊鄰，這樣非緊鄰的字元間應當允許其他字元插入。因此我們使用一個特殊符號夾在這非緊鄰的字元當中以處理此一現象。**

如同文法中的規則一般，規則的使用總免不了會有一些例外，然而我們的規則都來自於正面的訊息，無從得知使用上的例外。於是我們將規則回授給訓練資料，收集使用錯誤的的部份來做為負面的資訊。統計出負面資訊裡對應範圍前後最常出現的字或詞類便是我們對於某一規則的使用限制。

如前所述我們的測試資料共有 8206 **句，其中有** 6118 **句含有省略成分，大約佔** 75%**。**2008 **句沒有省略，約佔四分之一。完整的結果如下：**

	Positions with nulls	Positions without nulls
Null predicted	10272	4327
Null not predicted	3115	187430
Recall	76.74%	
Precision	70.36%	

四、計畫成果自評

我們提出了一套為不定長度的句子作分類的方法，並提出如何決定類別個數與初始值的策略，其中衡量兩句的相似相異程度的公式也是一個新的構想。因為採用 LCS**來找出規則，原本令人頭痛的長距離相關的現象也獲得了相當程度的解決。我們的方法僅僅採用了語法的訊息，在樹狀語料庫日漸普及，而詞類標記的技術又很成熟的情況下，我們的作法是相當可行而且有高度的移植能力，即使面對不同的語言，依然可以如法炮製。此外，我們用來歸納出規則的方式也可推廣到其他的應用上。對於不同的應用對於正確率與回收率的要求也就不同，有的需要較高的正確率有的則對回收率的要求較為嚴格，我們可以視情況來調整我們的系統。**

我們的正確率目前在70%**左右，這表示仍有相當大的空間可以努力。一旦有較多語意方面的資源，我們可以將系統提昇到語意的層次。而語言學家的幫忙也可補**

一些遺珠之憾。另外，如何在分類的方法上得到一些進步也是個有趣的課題。

整體而言，研究內容與原計畫所列的工作項目完全相符、並已經達成預期的目標、所提出的語言模型是多項應用的基礎、適合在學術期刊或會議上發表。

五、參考文獻

[1] Bies, A. et al. (1995) Bracketing Guidelines for Treebank II style Penn Treebank Project, 1995.

[2] Brown, P.F., Pietra, V.J. *et al.* (1992), "Class-Based N-Gram Models of Natural Language". *Computational Linguistics*, 18(4), 467-479.

[3] Charniak, E. (1993) *Statistical Language Learning,* MIT press, 1993.

[4] Chen, H.H. (1988) "A Logical Approach to Movement Transformation in Mandarin Chinese", *International Journal of Pattern Recognition and Artificial Intelligence* 2(1), 1988. pp. 71-86.

[5] Chen, H.H. (1992) "The Transfer of Anaphors in Translation", *Literary and Linguistic Computing,* 7(4), Oxford University press, 1992. pp. 231-238.

[6] Chen, H.H. and Yang, H.M. (1994). "Generating Logical Terms of Chinese Sentences", *Journal of Information Science and Information Engineering,* 10(3), 1994, pp. 245-316.

[7] Chen, K.H. and Chen, H.-H. (1993), "A Probabilistic Chunker". In: *Proceedings of ROCLING VI*, 99-117.

[8] Chen, K.-H. and Chen, H.-H. (1994), "Extracting Noun Phrases from Large Scale Texts: A Hybrid Approach and Its Automatic Evaluation". In: *Proceedings of the 32nd Annual Meeting of ACL*, 234-241.

[9] Chung, T.C.(1993). *Treatments of Ellipses in An English-Chinese Machine Translation System.* Master thesis, Department of Computer Science and Information Engineering, National Taiwan University, 1993.

[10] Crouch, R.(1995). "Ellipsis and Quantification: A Substitutional Approach", *Proceeding of EACL-95.*pp. 229-236.

[11] Gazdar, G. and Mellish, C. (1989). *Natural Language Processing in PROLOG,* Addison-Wesley, 1989.

[12] Hartigan, J.A. (1975) *Clustering Algorithms,* Wiley, 1975.

[13] Hahn, U. et al. (1996). "A Conceptual Reasoning Approach to Textual Ellipsis", *Proceeding of 12-th European Conference on Artificial Intelligence,* 1996, pp. 572-576.

[14] Hatzivassiloglou, V. and McKeown, K. (1993), "Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning". In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 172-182.

[15] Hindle, D. (1990), "Noun Classification from Predicate-Argument Structures". In: *Proceedings of 28th Annual Meeting of ACL*, 268-275.

[16] Jain, A.K. and Dubes, R.C. (1988) *Algorithms for Clustering Data,* Prentice Hall, 1988.

[17] Kehler, A. (1994) "Common Topics and Coherent Situations: Interpreting Ellipsis in the Context of Discourse Inference", *Proceeding of ACL-94.* pp. 50-57.

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

[19] Sells, P. (1985) *Lectures on Contemporary Syntactic Theories,* CSLI Publications, Stanford, 1985.

[20] Zimmermann, H.-J.(1991) *Fuzzy Set Theory,* Kluwer Academic Publishers, 1991.