口語處理系統語言模型之研究(II)

Language Models in Spoken Language Processing(II)

計畫編號：NSC 88-2213-E-002-052-
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

為了要處理因流暢自然的語言輸入所造成的困難，我們研究子句邊界的標定。一個兩階段詞性標記的架構也被提出來處理當標記口語資料時所發生的困難。由於即時的回應是邁向一個口語交談系統不可或缺的條件。在本計畫中，我們也提出了一個新的剖析架構來達到這個目的。這種新型態的剖析器有著極佳的速度，同時它剖析的精確率也可和許多優秀的剖析器相比擬。

關鍵詞：子句邊界識別、剖析、自然語言處理、口語、詞性標記、書面語

Abstract

For processing spontaneous speech, clause boundary identification is investigated to remove some noises and to make the processing unit more meaningful. A two-phase tagging scheme is proposed to deal with the problems from tagging the spoken corpus. Because the real-time response is one of the major issues in a practical spoken dialogue system, we also propose a new parsing scheme to meet such a requirement. This new type of parser is relatively faster than some other parsers and the parsing accuracy can also be comparable to some excellent parsers.

Keywords: Clause Boundary Identification, Parsing, Natural Language Processing, Spoken Languages, Tagging, Written Languages

二、緣由與目的

子句邊界的標定可以確保系統的處理單位較有語法上的意義。它不但對後階段的語言處理 (詞性標注以及句子剖析) 有影響，也對前階段的音轉字造成影響。

當我們使用傳統書面語的詞性標注的方法去標注口語語料將會遭遇到下列三個困難：

(1) 在口語語料中，句子的長度一般較書面語中的句子短。常有一個詞或兩個詞的句子出現。

(2) 缺乏大量的口語語料來訓練口語的詞性標注系統。

(3) 口語語料所採用的標記集 (tagging set) 和書面語語料所採用的標記集有所不同。因此，我們無法利用書面語的語料來部分補充口語語料在訓練上的不足。

為了要解決上述三項困難，我們提出了一個兩階段的詞性標注架構。

在句子剖析的階段，書面語和口語最大的差別在於文法。一般來說，口語在使用上較書面語有彈性。也因為這種彈性造成一些特殊的文法規則常常在口語中使用卻鮮少在書面語中使用。此外，即時的剖析也是在句子剖析階段所需考慮的一個重點項目，因為當一個交談在持續進行時，長時間的等待系統反應是無法被使用者所接受的。基於上述的描述，我們提出了一個多層次的切割-提升剖析器來符合這樣的需求。

三、結果與討論

3.1 子句邊界的標定

在標定子句的邊界上，我們利用一個基本分析 (basic analyses) 以及五個進一步的分析 (advanced analyses) 來過濾不可能

的位置。這六個線索如下所示：

基本分析一:

需要被標定的情況，其終端音調 (terminal pitch) 一般不是水平的 (level, -) 情況，否則不需要被標定。

進一步的分析一:

沒有子句會被標定在修復的語音之間。因此，語音的修復會影響子句的標定。

進一步的分析二:

當一次講話的長度夠長時，說話者通常完成他所講的話，因此，子句會被標定在最後。

進一步的分析三:

在中文裏，有一些字像是 "啊" 通常標示著句子的結尾。因此當這些字出現時，我們即可標示子句的邊界於這些字之後。

進一步的分析四:

假如有太多的其它語者介入某位語者的兩句話 (utterance) 之間，則我們可標示子句的邊界於那位語者的上一句話的結尾。

進一步的分析五:

一句話中有超過一個以上的子句邊界的標定也需要處理。

基於上述的分析，我們可以達到 94.46% 的精確率以及 87.38% 的召回率。

3.2 詞性標注

我們的中研院詞性標注系統是以中央研究院的平衡語料庫為訓練語料來發展的。我們利用標準的詞類雙連 (bi-class) 的機率來預測每個詞的詞類。公式如下所示：

$$\tilde{P} \equiv \underset{P}{\operatorname{argmax}} \prod_{i=1}^{n} \text{Prob}(p_i \mid w_i) * \prod_{i=1}^{n-1} \text{Asso}(p_i, p_{i+1})$$

基於這個公式，詞性標注系統可以達到 96.64% 的封閉測試 (close test) 的正確率。

我們的詞類對應系統是以中央研究院和台灣大學的平衡語料庫中共有的詞為起點，去看看這些詞在這兩個語料庫中分別被標注成什麼詞類。再去統計其詞類對應的機率。這個機率會在下一階段的台大詞性標注系統中用到。

我們的台大詞性標注系統是以台灣大學的平衡語料庫為訓練語料來發展的。我們也是利用標準的詞類雙連的機率來預測每個詞的詞類但作了一些改變。公式如下所示：

$$\tilde{U} \equiv \underset{U}{\operatorname{argmax}} \prod_{i=0}^{n} \text{Prob}(u_i \mid k_i) * \prod_{i=0}^{n-1} \text{Asso}(u_i, u_{i+1})$$

其中 K 代表中研院詞類串 (CKIP-Tag Sequence)，而 U 代表台大詞類串 (NTU-Tag Sequence)。

基於兩階段的詞性標注架構，我們得到 83.85% 的口語詞性標注的正確率。這個結果在我們目前缺乏大量口語語料的情況下，算是另人滿意的結果。此外，在這個實驗中我們也處理了只有一個詞的句子。這些句子通常會有固定的型式。例如，"哦"、"唉約" 和 "唉啊" 等都是一些感嘆詞，所以我們就直接將它標注成 "int"。其它如 "對" 和 "幹嘛"，我們就直接將它標注成 "vi"。

在這個實驗中，我們已經修復語音並將子句的邊界標定出來。假如這兩個工作沒有做的話，我們發現詞性標注的正確率會下降 6%。因此，語音修復的處理以及子句邊界的標定對詞性標注的正確率也會有一定程度的影響。

3.3 句子剖析

在切割-提升模型的架構下，切割和提升的運算都能夠在線性時間內完成。在切割-提升模型的架構下，我們需要利用到一個特殊的文法稱之為限制式文法。限制式文法可以從台大的樹狀結構語料庫中自動學習得之。它提供了每一個階層的切割-提升所需的文法。

在訓練的階段上，我們會對訓練語料庫中的每一棵剖析樹進行每一個層次的限制規則的抽取。抽取完後，我們將相同的限制規則合併並計算它的頻率。對於第一個階層的切割-提升文法而言，我們共抽出了 3,117 條限制規則。檢視所有的規則後我們發現有 290 條限制規則會有衝突的情況發生。衝突的發生原因來自於語料庫不一

致的建構以及限制式文法本身模型的限制。為了解決這個問題，我們將頻率高的規則留下，頻率低的規則刪除。

利用上一節所得到的限制式文法, 我們對台大的樹狀結構語料庫進行每個階層的剖析。由實驗中我們發現，低階層的剖析正確率並不高。這個結果可能會造成低階層的剖析錯誤漫延到高階層的剖析上。另一個現象則是正確率不穩定。它不會因越高層，剖析的正確率也越高。為解決這個現象，我們提出了一個錯誤-更正的技術。

當我們檢視所有的限制規則後, 我們發現有一些規則是不適當的。為了要衡量規則的適用性，我們去計算每一條規則的成功應用次數以及失敗應用次數。然後將那些成功應用次數小於失敗應用次數的規則去掉。

我們然後用這個書面語的語料去剖析口語的語料。我們可以發現正確率又不穩定。原因是因為有一些特殊的文法規則沒有在書面語的訓練語料中得到。因此，我們又運用錯誤-更正的技術於第一個口語的對話上。

3.4 討論

近年來，口語的處理逐漸受到大家的重視。在兩年中，我們針對口語處理系統中的五個基本研究主題作深入而廣範的討論。任何的口語處理系統不可能表現的很好假如我們沒有有效率的處理這些問題。為了要處理因流暢自然的語言輸入所造成的困難，我們研究在這種情形下的語音修復處理以及子句邊界的標定。為了要利用傳統書面語的語言模型來處理口語的資料，我們研究這兩種語言系統的差異。進而發展出口語的斷詞系統、詞性標注系統、以及剖析系統。藉由降低這兩種語言系統差異的技術，來降低重新為這些系統發展新的語言模型所需付出的代價。

在語音修復的處理上，取代的語音修復仍需要更多線索的幫助才能有效的提高系統的正確率。此外，在這本論文中放棄的語音修復也沒有提出解決的辦法。我們發現一個可用的線索是：急速停頓 (glottal stop, %) 。然而，放棄的語音修復仍然需要更多其它的輔助線索。在子句邊界的標定上，雖然我們得到不錯的正確率，但它仍有改進的空間。利用句子間的停頓時間 (pause duration) 長短將是一個好的努力方向。再者，如何將我們所提出的方法整合到一個實際的口語交談系統上將是一個富有挑戰性的工作。對於文本模組而言，一個大型的口語語料庫是必要的。有了大型的口語語料庫文本模組的正確率將可大大的提升。此外，音韻訊息對剖析的影響也需要更進一步的研究。

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博士論文：語料庫為本的中文口語處理一些新技術的研究。另外，發表於 Eurospeech99 論文兩篇。

五、參考文獻

黃宣範 (1996) "漢語口語語料庫的建立," 語言學門專題計畫研究成果發表會, 台北, 南港, 1996.

F.J. Allen, *et al.* (1996) "A Robust System for Natural Spoken Dialogue," *Proceedings of 34th Annual Meeting of ACL*, 1996, pp. 62-70.

G. Bakenecker and U. Block (1994) "Improving Parsing by Incorporating 'Prosodic Clause Boundaries' into a Grammar," *Proceedings of International Conference on Spoken Language Processing*, 1994, pp. 1-4.

E. Brill (1992) "A Simple Rule-Based Part-of-Speech Tagger," *Proceedings of Applied Natural Language Processing*, 1992, pp. 152-155.

H.H. Chen and Y.S. Lee (1995b) "Development of Partially Bracketed Corpus with Part-of-Speech Information Only," *Proceedings of 3rd Workshop on Very Large Corpora*, 1995, pp. 162-172.

H.H. Chen and Y.S. Lee (1995c) "A Chunking-and-Raising Partial Parser," *Proceedings of 4th International Workshop on Parsing Technologies*, 1995, pp. 137-158.

K.W. Church (1988) "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," *Proceedings of Applied Natural Language Processing*, 1988, pp. 136-143.

R. Cole (1995) "The Challenge of Spoken Language

Systems: Research Directions for the Nineties," *IEEE Transaction on Speech and Audio Processing*, Vol. 3, No. 1, 1995, pp. 1-21.

D. Crystal (1980) "Neglected Grammatical Factors in Conversational English," S. Greenbaum, G. Leech and J. Svartvik, editor, *Studies in English Linguistics*, Longman, 1980.

D. Cutting, *et al.* (1992) "A Practical Part-of-Speech Tagger," *Proceedings of Applied Natural Language Processing*, 1992, pp. 133-140.

B.A. Fox and R. Jasperson (1996) "A Syntactic Exploration of Repair in English Conversation," *Descriptive and Theoretical Models in the Alternative Linguistics*, P.W. Davis (Ed.), John Benjamins Publishing, 1996.

P. Heeman and J. Allen (1997) "Intonational Boundaries, Speech Repairs and Discourse Markers: Modeling Spoken Dialog," *Proceedings of ACL/EACL*, 1997, pp. 254-261.

P. Heeman and J. Allen (1994b) "Tagging Speech Repairs," *Proceedings of Human Language Technology*, 1994, pp. 187-192.

D. Hindle (1983) "Deterministic Parsing of Syntactic Nonfluencies," *Proceedings of 23rd Annual Meeting of ACL*, 1983, pp. 123-128.

C.R. Huang, *et al.* (1995) "An Introduction to Academia Sinica Balance Corpus," *Proceedings of ROCLING*, 1995, pp. 81-99.

N. Joakim, *et al.* (1996) "Tagging Spoken Language Using Written Language Statistics," *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 1078-1081.

W.J.M. Levelt (1983) "Monitoring and Self-Repair in Speech," *Cognition*, Vol. 14, 1983, pp. 41-104.

C. Lyon and B. Dickerson (1995) "A fast partial parse of natural language sentences using a connectionist method," *Proceedings of 7th European Chapter of ACL*, 1995, pp. 215-222.

D.M. Magerman (1995) "Statistical Decision-Tree Models for Parsing," *Proceedings of 35th Annual Meeting of ACL*, 1995, pp. 276-283.

B. Merialds (1994) "Tagging English Text with a Probabilistic Model," *Computational Linguistics*, Vol. 20, No. 2, 1994, pp. 155-171.

D. O'Shaughnessy (1992) "Recognition of hesitation in Spontaneous Speech," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1992, pp. 521-524.

P. Placeway, *et al.* (1993) "The Estimation of Powerful Language Models from Small and Large Corpora," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1993, pp. 33-36.

R. Schwartz, *et al.* (1994) "On Using Written Language Training Data for Spoken Language Modeling," *Proceedings of Human Language Technology*, 1994, pp. 94-98.

M.S. Shaw (1996) *Design and Implementation of a Treebank Development Tool*, Master Thesis, Department of Computer Science and Information Engineering, National Taiwan University, 1996.

A. Stolcke and E.E. Shriberg (1996a) "Automatic Linguistic Segmentation of Conversational Speech," *Proceedings of International Conference on Spoken Language Processing*, 1996, pp. 1005-1008.

A. Stolcke and E.E. Shriberg (1996b) "Statistical Language Modeling for Speech Disfluencies," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1996, pp. 405-408.

A. Stolcke and E.E. Shriberg (1996c) "Word Predictability after Hesitations: A Corpus-Based Study," *Proceedings of International Conference on Spoken Language Processing*, 1996, pp. 1868-1871.

T. Strzalkowski (1992) "TTP: A fast and robust parser for natural language," *Proceedings of 14th International Conference on Computational Linguistics*, 1992, pp. 198-204.

P. Tapanainen and A. Voutilainen (1994) "Tagging Accurately - Do'nt Guess If You Know," *Proceedings of Applied Natural Language Processing*, 1994, pp. 47-52.

M. Tomita (1986) *Efficient Parsing for Natural Language*. Kluwer Academic Publishers, 1986.

J. Vergne (1994) "A Non-Recursive Sentence Segmentation, Applied to Parsing of Linear Complexity in Time," *Proceedings of New Methods in Language Processing*, 1994, pp. 234-241.