行政院國家科學委員會專題研究計畫成果報告

口語處理系統語言模型之研究(I)

Language Models in Spoken Language Processing Systems (I)

計畫編號：NSC 87-2213-E-002-030
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

口語處理系統使得人類運用語音和電腦交談的夢想不再遙不可及。由於語音和語言處理技術的快速發展，使得口語處理的技術也相對的提升不少。雖是如此，但仍有一些瓶頸需要進一步的克服。

本兩年期的計畫，針對口語處理系統中基本研究主題，作深入而廣範的討論。第一年主要集中於語音修復處理和口語的斷詞上，為了要利用傳統書面語的語言模型去處理口語的資料，我們研究這兩種語言系統的差異。藉由降低這兩種語言系統差異的技術來降低重新發展一個新的語言模型所需付出的代價。

語音修復的目的是要將一些會對系統造成理解上錯誤的雜訊預先排除掉。為了處理口語的斷詞，我們提出了一個新詞學習的模型來輔助詞典無法涵蓋所有詞的缺點。

關鍵詞：自然語言處理、斷詞、口語、書面語

Abstract

The spoken language processing systems facilitate man-machine interface using speech. Although the technologies for the spoken language processing have made rapid advances in the underlying speech and language technologies, there are significant obstacles that must be overcome.

This two-year project focuses on fundamental research topics in spoken language processing. In the first year we deal with speech repair processing and spoken word segmentation. In order to utilize the conventional written language model to process spoken data, we investigate the differences between the written languages and the spoken languages, and propose several techniques to reduce the differences between these two language systems.

Speech repair processing removes some noises. For spoken word segmentation, a new word learning model is proposed.

Keywords: Natural Language Processing, Segmentation, Spoken Languages, Written Languages

二、緣由與目的

對於人與人之間的通訊而言，語音是最自然的一種方式。口語處理系統使得人類運用語音和電腦交談的夢想不再遙不可及。由於語音和語言處理技術的快速發展，使得口語處理的技術也相對的提升不少。雖是如此，但仍有一些瓶頸需要進一步的克服。一般而言，一個口語交談系統有下列四項困難需要解決：

(1) **不特定語者，**
(2) **使用流暢自然的語音輸入，**
(3) **傳統書面語言模型的使用，**
(4) **口語交談的管理。**

近年來語料庫為本 (corpus-based) **的策略在許多應用上獲得極大的成功。一般而言，語料庫反應了真實語言的使用方式。因此對設計一個口語交談系統而言，它是一個好的知識來源，因為大多數的語言現象都可以從大型的語料庫中發掘出來。所**

以在這本論文中，我們便採用此種策略來發展我們的系統。

本兩年期的計畫, 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。第一年主要集中於語音修復處理和口語的斷詞上，為了要利用傳統書面語的語言模型去處理口語的資料，我們研究這兩種語言系統的差異。藉由降低這兩種語言系統差異的技術來降低重新發展一個新的語言模型所需付出的代價。

三、結果與討論

3.1 口語語料庫

我們所使用的口語語料包含了兩段在公共場合中朋友之間的日常交談。每段都有大約 40 分鐘的長度。在這兩個對話中，各有四個和五個語者包含在其中。在這個語料庫中共有 448 個語音修復。我們將這 448 個語音修復分成四種類型：重複的語音修復、附加的語音修復、取代的語音修復、以及放棄的語音修復。部分語料庫的內容如下頁所示：

113 Z: ...(1.2)**那楊經理**<A **有沒有** A>[**在^唸**]?/
114 W: [**就等於**],- {P1,2-2,2}
115 ..**就是**,-
116 ...**跟^過年這樣**.\
117 Y: (0)<L2 okay L2>?/
118 W: ...(1.2)**他**%-- {R1,1-1,1}
119 ..**他沒有**.\
120 ..**他就說^可以**.\
121 Z: ...**哦那**%--
122 ..**一下越變越好**.\
123 ...**不**%-- {R1,1-1,1}
124 ..**不是**,-
125 ..<SAR **他是太忙了**.\
126 ...**沒有時間唸** SAR>.\
127 L: ...(3.)**楊**-- {R1,1-1,1}
128 Z: ..ha?/
129 L: ..**楊** M **會唸他**?/
130 Z: ...**誰**/
131 L: **楊** M,- {R1,1-1,1}
132 Y: ...**楊** emu.\
133 L: ...**楊**,- {R1,1-1,1}
134 ...[**楊** K=]**比較懂**.\
135 W: [**啊**=],-
136 **全**%-- {R1,1-1,1}
137 ..**全公司**=,-
138 L: ...(1.7)**現在他沒賺錢**,-
139 **他** {P1,1-1,2} ^**大家都** {R1,1-1,1}**都唸了** ha?/
140 W: ...**不是**=,-
141 Z: ...**他**%,-
142 ..**沒有**.\
143 ..**他其實很忙唉呵**?/

在這四種語音修復中，重複的語音修復佔最大宗 (約佔 70%) 。其次是附加的語音修復 (約佔 13%) 。再來是取代的語音修復 (約佔 10%) 。最後是放棄的語音修復 (約佔 7%) 。

3.2 中文語音修復處理

修復重複的語音我們先利用簡單的比對來產生可能的修復位置，再利用兩個基本分析以及四個進一步的分析來過濾不可能的位置。這六個線索如下所示：

基本分析一:

重複音節的長度會有一定的限制。

基本分析二:

重複的音節之間不可以有太多其它語者的介入。

進一步的分析一:

需要被修復的情況，重複的音節之間一般會有停頓出現。

進一步的分析二:

重複的音節之間出現急速停頓 (glottal stop, %) 的現象時，這些重複的音節一般需要被修復。

進一步的分析三:

相同的兩句話，如果它們的長度太長，則一般不需要修復。

進一步的分析四:

有一些重複出現的音節，它們雖然重複，但不需要修復。相反的，有一些音節只要它們重複一般就需要修復。基於上述的分析，我們可以達到 93.87% 的精確率以及 90.65% 的召回率。

修復附加的語音我們利用兩個基本分析以及七個進一步的分析來過濾不可能的

位置。其中，兩個基本分析和四個進一步的分析的原理和修復重複的語音相同，因此我們省略不寫。底下僅針對三個新的線索加以探討。

進一步的分析五:

需要被修復的情況，重複的音節之間的終端音調一般是水平的出現，否則不需要被修復。

進一步的分析六:

我們需要去區分是附加的語音修復或者是取代的語音修復。

進一步的分析七:

需要被修復的情況，被附加的音節一般是一個詞。如果被附加的音節不是一個詞或者它是詞的一部份，則此種情況不需要被修復。

修復取代的語音我們也是利用上述的幾個分析，因此我們省略不寫。基於上述的分析，我們可以達到 42.55% 的精確率以及 45.45% 的召回率。

3.3 中文口語斷詞

在斷詞的步驟上, 我們首先利用書面語的斷詞系統去斷口語語料。實驗的結果顯示，我們的書面語斷詞系統得到 89.01% 的精確率以及 90.35% 的召回率。這個結果和其它一些斷詞的研究所得到的結果(95%以上的精確率及召回率) 有一段差距。而當我們分析口語的語料後，我們發現口語語料中 13.34% 的詞無法在九萬多詞的書面語詞典中找到。這就是斷詞的結果不甚理想的原因。造成這樣的結果，明顯的是由於未知詞的影響。因此，新詞的自動學習是必要的。

在新詞的自動學習上, 我們利用 ϕ^2 的分佈來標定可能之詞。考慮下列的句子：

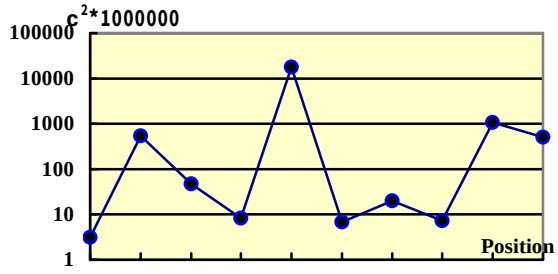
從王惠然先生創制徐州柳琴出發

正確的斷詞結果如下所示：

從 王惠然 先生 創制 徐州 柳琴 出發

這個句子前 11 個字的 ϕ^2 分佈如下圖所示：

從 王 惠 然 先 生 創 制 徐 州 柳



由這個圖中我們可以發現，詞的邊界是落在圖的區域最低點上。

我們利用這樣的觀念, 來產生一連串的詞。我們過濾了一些原本就出現在詞典中的詞以及單字詞後， 903 個可能的新詞被提出。其中有 153 個是真正的新詞。除此之外，我們也自行增加了一些口語的常用詞如 “唉呀” 、 “唉約” 、 “對不對” 、 “那有” 等。對於一些口語的衍生詞我們也建立了一些規則來解決這個問題。例如 “S^+” (S={好、呵、咚、哦、對、罵、…}) 是針對下列的詞而產生的： “好好好” 、 “哦哦哦哦哦” 、 “對對對對對對對” 等。而 “S一S” (S={弄、套、烤、算、…}) 是針對 “弄一弄” 、 “算一算” 等詞而產生的。

利用新詞學習的方法, 我們將所學習到的新詞加入原本的書面語詞典中。重新斷詞的結果，我們得到 95.30% 的精確率以及 93.08% 的召回率。精確率比原來增加 6.29% 而召回率比原來增加 2.73%。

四、計畫成果自評

近年來，口語的處理逐漸受到大家的重視。在這個計畫中，我們針對口語處理系統中的基本研究主題作深入而廣範的討論。任何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。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。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，在報告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。然而，放

棄的語音修復仍然需要更多其它的輔助線索。

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] J. Bear, J. Dowding and E. Shriberg (1992) "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog," *Proceedings of 32nd Annual Meeting of ACL*, 1992, pp. 56-63.

[2] J.S. Chang, Z.D. Chen and S.D. Chen (1991) "A Method of Word Identification for Chinese by Constraint Satisfaction and Statistical Optimization Techniques," *Proceedings of ROCLING*, 1991, pp. 147-165.

[3] H.H. Chen and J.C. Lee (1996) "Identification and Classification of Proper Nouns in Chinese Texts," *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 222-229.

[4] K.J. Chen and S.H. Liu (1992) "Word Identification for Mandarin Chinese Sentences," *Proceedings of 14th International Conference on Computational Linguistics*, 1992, pp. 101-107.

[5] K. Chui (1995) "Repair in Chinese Conversation," *Proceedings of the Second International Symposium on Language in Taiwan*, 1995, pp. 75-96.

[6] C.K. Fan and W.H. Tsai (1988) "Automatic Word Identification in Chinese Sentences by the Relaxation Technique," *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 1, 1988, pp. 33-56.

[7] 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.

[8] P. Heeman and J. Allen (1994) "Detecting and Correcting Speech Repairs," *Proceedings of 34th Annual Meeting of ACL*, 1994, pp. 295-302.

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

[10] G.I. Kikui and T. Morimoto (1994) "Similarity-Based Identification of Repairs in Japanese Spoken Language," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1994, pp. 915-918.

[11] S. Kurohashi and M. Nagao (1992) "Dynamic Programming Method for Analyzing Conjunctive Structure in Japanese," *Proceedings of 14th International Conference on Computational Linguistics*, 1992, pp. 170-176.

[12] Y.S. Lee and H.H. Chen (1996) "Correcting Chinese Repetition Repairs in Spontaneous Speech," *Proceedings of ROCLING*, 1996, pp. 137-158.

[13] J.C. Lee, Y.S. Lee and H.H. Chen (1994) "Identification of Person Names in Chinese Texts," *Proceedings of ROCLING*, 1994, pp. 203-222.

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

[15] C. Nakatani and J. Hirschberg (1993) "A Speech-First Model for Repair Detection and Correction," *Proceedings of European Conference on Speech Communication and Technology*, 1993a, pp. 1173-1176.

[16] C. Nakatani and J. Hirschberg (1993) "A Speech-First Model for Repair Detection and Correction," *Proceedings of 33rd Annual Meeting of ACL*, 1993b, pp. 46-53.

[17] R. Sproat (1990) "An Application of Statistical Optimization with Dynamic Programming to Phonemic-Input-to-Character Conversion for Chinese," *Proceedings of ROCLING*, 1990, pp. 377-390.

[18] M.S. Sun, D. Shen and C. Huang (1997) "Cseg&Tag1.0: A Practical Word Segmenter and POS Tagger for Chinese Texts," *Proceedings of Applied Natural Language Processing*, 1997, pp. 119-126.