以雜訊通道/訊息重建模型自動評測語言處理系統

張照煌

(工業技術研究院 電腦與通訊工業研究所)

摘 要：近年來許多研究單位推出中文標準語料庫，如新加坡國立大學的PH語料庫，中央研究院的平衡語料庫1.0版及2.0版，除可以提供自然語言處理研究所需的訓練及測試資料，幫助建立語料庫式的中文語言模型，更可以提供客觀、公正的系統評測依據。本文提出以雜訊通道/訊息重建模型配合標準語料庫來自動評測語言處理系統，實際應用於網際網路中文處理的兩個簡單而重要的問題：繁簡字碼轉換與第八位元重建；並利用研究院平衡語料庫1.0版及2.0版，進行大規模實驗，顯示所設計的雜訊通道/訊息重建自動評測模型確為實用可行。

Automatic Evaluation of Language Processing Systems Using Noisy Channel/Information Restoration Model

Chao-Huang Chang

(Computer & Communication Research Laboratories, Industrial Technology Research Institute)

Abstract : In the recent years, several standard Chinese corpora, such as NUS's PH corpus and Academia Sinica's sinica corpus version 1.0, 2.0 have been released to the academia. These corpora are useful not only for training and testing corpus-based NLP systems, but also for objective evaluation of the systems. In this article, we present a noisy channel/information restoration model for automatic evaluation of NLP systems. The proposed model has been applied to two common and important problems related to Chinese NLP for the Internet: the 8-th bit restoration of BIG-5 code through non-iso8859-1 channel, and GB-BIG5 code conversion. Sinica Corpora version 1.0 and 2.0 are used in the experiment. The results show that the proposed model is useful and practical.

一、前言

在1992年[1]，我們提出了「雙向轉換」(bidirectional conversion) 的觀念，利用語料庫自動地評測音節轉漢字的轉換正確率，後來還延伸這個觀念設計了音節轉漢字的自動調適方法[2]，引起了不少回響[3]。近年來許多研究單位推出中文標準語料庫，如新加坡國立大學的PH語料庫[4]，中央研究院的平衡語料庫1.0版[5]及2.0版，除可以提供自然語言處理研究所需的訓練及測試資料，幫助建立語料庫式的中文語言模型，更可以提供客觀、公正的系統評測依據。本文提出以雜訊通道[6,7]/訊息重建模型配合標準語料庫來自動評測語言處理系統，實際應用於網際網路（Internet）中文處理的兩個簡單而重要的問題：第八位元重建與繁簡字碼轉換；並利用研究院平衡語料庫1.0版及2.0版，進行大規模實驗，顯示所設計的雜訊通道/訊息重建自動評測模型確為實用可行。

在網際網路及全球資訊網（WWW）盛行的今天，由於電腦與網路原來是起源於歐美，並未考慮到中文眾多方塊字的編碼需求，電腦網路上所使用最為普遍的ASCII碼乃為七位元編碼，而一個位元組只有八個位元，也容納不下為數上萬的中文漢字。更由於兩岸分別使用繁簡

兩種字體，且對電腦使用的中文資訊內碼分別進行編碼，以致在網際網路中文處理上出現了不少問題[8]。全球資訊網的網頁（ web page ）常需提供兩個以上的版本（中文版及英文版），而中文版更需再區分為 BIG5 版及 GB 版，未來甚至需有 Unicode 版，來適應不同的中文編碼環境。本文討論兩個與網際網路中文處理相關的問題：BIG5 碼第八位元重建，及 BIG5-GB 繁簡字碼轉換。

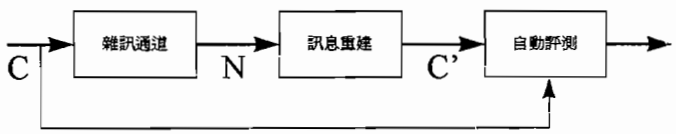
BIG5 碼是電腦網路上所使用最為普遍的中文編碼之一。是一種兩位元組的內碼，其編碼範圍為（ 0x 表 16 進制）：高位元組：0xA1-FE 、0x8E-A0 、0x81-8D ；低位元組：0x40-7E 、0xA1-FE 。常用及次常用漢字範圍分別為 0xA440-C67E 、0xC940-F9D5 ；其他為特殊符號及使用者造字編碼範圍。在大陸的簡體字電腦漢字內碼中最通用的是 GB2312-80 ，簡稱 GB 碼，也是一種兩位元組的內碼，其高位元組與低位元組編碼範圍均為 0xA1-FE 。

由於國際電腦網路通道多以七位元傳遞電子郵件（所謂 non-iso8859-1 ），以 BIG5 為內碼的電子郵件，若未經如 uuencode 的工具進一步編碼，在接收端的收信者常會收到所謂的「亂碼」郵件。文獻上少見對這個問題的探討，只有新竹交通大學的黃世昆設計了一套名為 Big5fix[9] 的分享軟體（ shareware ），是所見唯一對這個問題提出的解決方案，Big5fix 針對輸入的七位元檔，先分為英文區和中文區，而後將中文區的文字，根據所收集的單字、字雙連、字三連（ character unigram, bigram,及 trigram ）及其出現頻率來判斷中文區的文字是什麼，他說重建正確率估計為 90%（其中中文區 95%，英文區 80%）。當然，分享軟體由熱心人士撰寫，提供大眾使用，不比商用軟體有人力、物力支援，較不可能進行大規模的實驗，正確率只是粗略估計，這時候我們所設計的雜訊通道/訊息重建自動評測模型配合大規模標準語料庫正好派上用場。本文除自動評測 Big5fix 的中文區重建正確率之外，也將提出一套智慧型第八位元重建系統，利用中文語言模型來解決歧義問題，進行重建。（ GB 碼並無類似的歧義問題。）

至於 BIG5-GB 繁簡字碼轉換問題[10]，則為眾所皆知，在兩岸交流日益頻繁的今天，也顯得更為重要，除了繁簡字對照手冊、簡繁字電腦對照詞典比比皆是外，繁簡字碼自動轉換軟體也大量出現，如 HC 漢字轉換器分享軟體、「漢字通」、「 Globalsurf 」、「亞洲心」等等，但在網際網路上最常使用的似乎仍是一對一的轉碼器，以致於在網路新聞群組如 alt.chinese.text.big5 或電子雜誌如華夏文摘 BIG5 版中，由 GB 碼轉成 BIG5 碼的文章錯誤隨處可見，頗為刺眼，如「家里(裡)」、「几(幾)個」、「技朮(術)」、「標准(準)」、「關系(係)」、「計划(劃)」、「采(採)用」、「制(製)造」等等不勝枚舉。本文除分析 HC 漢字轉換器及漢字通的轉換結果之外，也將提出利用中文語言模型及異體字表的智慧型簡繁字碼轉換系統。所利用的中文語言模型包括詞問字雙連接續表模型（ interword character bigram, 簡稱 IWCB ）及模擬退火詞群雙連接續表模型（ simulated-annealing clustered word-class bigram ）[11,12]。

二、雜訊通道/訊息重建自動評測模型

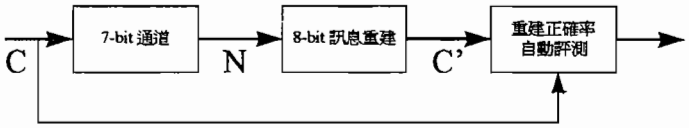
我們提出的雜訊通道/訊息重建自動評測模型，乃延續「雙向轉換」的觀念，利用語料庫自動地評測各種語言處理系統的效能，可以下圖說明：



可以將一語言處理系統想像為一種訊息重建過程：利用一大型標準語料庫 C ，通過一模擬之雜訊通道後，得到含雜訊的語料庫 N ，當作語言處理系統，即訊息重建過程的輸入，令其處

理結果為語料庫 C'，而自動評測模組則利用原語料庫 C 與處理結果語料庫 C'的比較，自動算出該語言處理系統的效能-正確率。這個自動評測模型要有趨近於真實的評測結果，雜訊通道的模擬必須趨近於完美，最好是一對一，或正確率趨近於百分之百的程序。首先以音節漢字互轉為例，由漢字轉成音節的過程雖非一對一（有許多破音字），但設計一正確率 98%的漢字轉音節系統並不困難[1,3]，因此以所提出的模型自動評測音節轉漢字的轉換正確率，實屬恰當。同樣的自動評測模型也可適用於其他各種語言處理系統，如語音辨識語言解碼、斷詞、詞性自動標注、文字辨識後處理、機器翻譯、及本文接下去要探討的兩個問題：BIG5 碼第八位元重建，及 BIG5-GB 繁簡字碼轉換。

BIG5 碼第八位元重建過程的雜訊通道模擬，為完美的一對一，只要將各位元組的第八位元遮罩掉即可，因此應用所提出的雜訊通道/訊息重建自動評測模型模型，最為合適，結果完全準確。圖示如下：



BIG5-GB 繁簡字碼轉換過程的雜訊通道模擬就沒有那麼單純了，不但有些繁體漢字可以對應到一個以上的簡體字（如乾⇨干、乾；覆⇨复、覆），許多繁體字根本找不到對應的簡體字。不過以大型語料庫文字出現頻率平均來說，此雜訊通道模擬過程的正確率仍趨近於百分之百。應用所提出的自動評測模型，仍為合適，圖示如下：

三、標準語料庫的準備

本文將使用中央研究院的平衡語料庫 1.0 版(1995 年公開，號稱 200 萬詞)及 2.0 版(1996 年公開，號稱 350 萬詞)，來驗證所提出的自動評測模型模型，這兩版研究院語料庫的一些統計數字列表如下：

研究院語料庫	大小(bytes)	檔數	句數	詞數	字數(含符號)	字數(漢字)
1.0 版	44,525,299	67	284,455	1,342,861	3,347,981	2,953,065
2.0 版	84,256,391	253	411,470	1,946,958	4,834,933	4,143,021

其中的分詞分句均以中央研究院原始提供的資料為準，此平衡語料庫的分詞採用計算語言學學會的分詞標準，已納入為資訊用分詞標準草案；詞性標記集則根據中研院詞庫小組的詞類分析簡化而成的 46 個標記[5]。不過本文的實驗並未利用原始語料庫的分詞和標記，而是使用經過下列步驟還原的分句文本：

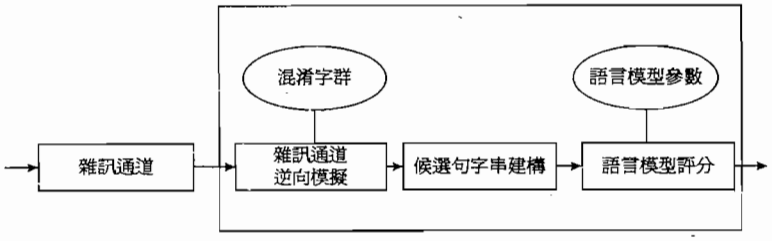
1. 以 unix 工具 grep 濾除各檔之文章分類標示，即以%%開頭的文行；並濾除句子分隔列，即整行皆為*者。
2. 根據語料庫的格式，設計一簡單工具 extract-word，擷取各句子中的詞，其中詞性標記已去除，輸出例："我 起來 了 ，"；"太陽 也 起來 了 。"
3. 將各句子的詞連在一起，例："我起來了，"；並將所有檔案合為一個大檔。

4. 將所有的使用者造字及非 BIG5 碼，規範為特殊符號□。

經過處理後的語料庫，基本上為一行一句，句中各字均為占兩位元組的 BIG5 碼。上列表中的字數，即以處理後的語料庫來計算。

四、第八位元重建

BIG5 碼第八位元重建問題已在第一、二節說明，不再重複。在雜訊通道模擬部分，只要將各位元組的第八位元遮罩掉即可，數行程式即可做到。本文除採用 Big5fix 做為比較基礎之外，也提出一套智慧型第八位元重建系統，利用中文語言模型來解決歧義問題，進行重建。此系統的基本架構，乃為我們一貫採用的「混淆字群代換-語言模型評分」作法[11-14]，（如下圖）將輸入檔的文字以句子為單位，逐一以混淆字群代換，由此產生多數個候選句字串，再根據一語料庫式統計語言模型，對各候選句字串加以評分，以找出評分最高（或出現機率最高）的候選句字串，當作系統輸出。此處的「混淆字群代換」，可視為「雜訊通道」的逆向模擬。



此處的「混淆字群」的建構十分簡單，因為要處理的是兩位元組的 BIG5 碼，我們對每個字最多只有兩個假設：高位元組的第八位元一律還原為 1，而低位元組的第八位元可為 0 或 1(根據編碼範圍來定)，例如 0x2440 的逆向模擬混淆字群為 0xa440「一」及 0xa4c0「分」，而 0x2421 者則只有 0xa4a1「丑」（0xa421 不在編碼範圍內）。我們將含七個倚天字的 13060 個中文字分別建立混淆字群，其中 10391 個字群含兩個字，其餘 2669 個字群只含一個字。所利用的中文語言模型為詞間字雙連接續表模型（簡稱 IWCB）。

接著討論實驗結果。下表分別列出 Big5fix 及我們的智慧型第八位元重建系統(稱 CCL-fix）自動評測的結果（錯誤數及錯誤率%）：

研究院語料庫	樣本	字數	Big5fix		CCL-fix	
1.0 版	含符號	3,347,981	125,915	3.76	57,862	1.72
	漢字	2,953,065	100,006	3.38	53,729	1.81
2.0 版	含符號	4,834,933	173,544	3.58	71,549	1.48
	漢字	4,143,021	111,809	2.69	70,758	1.70

可以看出對漢字部份而言，Big5fix 對 1.0 版及 2.0 版的重建正確率分別為 96.62%及 97.31%比黃世昆自己估計的 95%還要高出 1.62%、2.31%。CCL-fix 的重建正確率則達 98.19%及 98.30%，顯示出 IWCB 語言模型確實優於單純的單字、字雙連及其出現頻率的作法。以下並列出這兩個系統對研究院語料庫 1.0 版的重建錯誤分析，表中列出錯誤出現次數最多的前 20 名，各項均列出其原漢字、重建漢字及出現次數，如 Big5fix 出現次數最多的第一名為將「分」重建為「一」，共有 3007 次。

名次	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Big5fix	分一	化了	林者	外全	全外	區記	色在	股松	來沒	省某	西多	價語	代用	反力	石加	吳找	十天	船爽	油迎	村困
	3007	1540	1481	893	819	797	792	771	734	723	722	715	712	709	676	672	664	611	611	601
CCLfix	一分	了化	分一	又太	沒來	外全	天十	每並	多西	林者	十天	代用	象僅	某省	叫件	沙事	士方	女月	命所	吧扭
	2298	1388	1375	1327	1325	1209	1194	887	638	577	530	491	484	465	458	396	386	376	359	343

五、簡繁字碼轉換

系統設計方面，我們採用了三套簡繁字碼轉換系統來模擬雜訊通道：(1) HC 漢字轉換器 1.2u 版，作者為 Fung F. Lee 和 Ricky Yeung ；(2) HC 漢字轉換器補充版，將轉碼表作了些許補充；及(3) 漢字通(KanziWEB)的 MultiCode 。這三個系統基本上都是採用查轉碼表轉換的作法，以致無法處理一對多的問題，在由 GB 碼轉回 BIG5 碼時，錯誤頗多。下表列出這三個系統雙向轉換自動評測的結果（錯誤數及錯誤率%）：

研究院語料庫	樣本	字數	HC1.2u 版		HC 補充版		漢字通	
1.0 版	含符號	3,347,98	271,98	8.12	46,16	1.37	29,53	0.87
	漢字	2,953,06	43,15	1.46	43,07	1.45	29,07	0.98
2.0 版	含符號	4,834,93	403,95	8.35	68,04	1.40	43,70	0.90
	漢字	4,143,02	60,11	1.45	60,03	1.45	40,56	0.98

為了要處理 GB 轉 BIG5 一對多的問題，必須採用考慮上下文的智慧型語言模型轉換法。文獻中楊與扶[10]曾提出一個大陸漢字與台灣漢字文本間的智能轉換系統，其基本作法為分類建表，分級評分，遇同分混淆則由人工選擇，並未用到精密的語言模型。我們仍採用「混淆字群代換-語言模型評分」作法，所利用的中文語言模型包括 IWCB 模型及模擬退火詞群雙連接續表模型（ SA 模型）[11-14]。實驗中我們採用兩種 SA 模型： 200 詞群及 300 詞群者，分別稱為 SA-200 模型及 SA-300 模型。

在「雜訊通道」的逆向模擬，也就是混淆字群方面，主要是異體字及通同字的蒐集，亦即 GB 轉 BIG5 一對多的映射模擬。我們所蒐集的異體字及通同字有三個來源：(1) HC1.2u 版的 YiTiZi 檔；(2) 臧遠侯書中所列大陸簡化字逐字注釋表[15]；(3) 蕭佳賓等計畫報告[16]的附件十。由以上三個來源，共整理出四個版本的混淆字群，並加以實驗比較：

混淆字群	來源	無混淆字	2 字混淆	3 字混淆	4 字混淆	5 字混淆
A	(1)	12644	364	48	4	0
B	(1)(2)	12397	597	57	9	0
C	(3)	12301	670	68	16	5
B	(1)(2)(3)	12144	777	117	15	7

接著討論實驗結果。下表分別列出我們的利用三種語言模型配合四個版本的混淆字群構成智慧型簡繁字碼轉換系統自動評測的結果（錯誤數及錯誤率%）：（以 HC 補充版之輸出為基礎）

研究院語料庫	漢字字數	IWCB				SA-200				SA-300			
		A	B	C	D	A	B	C	D	A	B	C	D
1.0 版	2,953,065	12,742 0.43%	10,144 0.34%	12,997 0.43%	12,684 0.42%	15,574 0.52%	13,977 0.47%	16,867 0.57%	16,811 0.56%	13,614 0.44%	10,849 0.36%	13,500 0.45%	13,225 0.44%
2.0 版	4,143,021	17,752 0.42%	14,139 0.34%	18,774 0.45%	18,465 0.44%	21,127 0.50%	18,593 0.44%	23,299 0.56%	23,297 0.56%	18,729 0.45%	15,439 0.37%	19,790 0.47%	19,554 0.47%

可以看出對簡繁字碼轉換系統而言， IWCB 模型效果最佳， SA-300 模型也有相近的轉換正確率， SA-200 模型稍差，但三種智慧型轉換法效果均遠優於如漢字通的一對一轉換法，錯誤率相差一倍以上。四個版本的混淆字群中以 B 版本效果最好，字數多的 C 版本和 D 版本結果反而變差，推測是造成過多不必要的混淆； A 版本則顯然字數不足。

以下並列出其中四個系統（ HC1.2u 版、漢字通、 IWCB 模型及 SA-300 模型之 B 版本）對研究院語料庫 2.0 版的轉換錯誤分析，表示法同上節（□或空白表無對應字）。

名次	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
HC 1.2u	裡里 6207	並并 5974	衛朮 4574	幾几 3434	準准 2052	係系 1985	遊游 1866	劃划 1800	製制 1513	採采 1464	證証 1430	願愿 1321	臺台 1071	範□ 937	隻只 860	築□ 850	妳□ 825	豐丰 797	復複 758	衝沖 713
漢字通	裡里 6207	聽听 2922	係系 1985	遊游 1866	製制 1513	採采 1464	臺台 1071	妳奶 825	複復 781	衝沖 713	週周 668	牠它 667	症癥 620	蘇甦 603	幹干 564	儘盡 538	閒閑 455	碰踫 446	斂 440	佈布 439
IWC B/B	臺台 885	妳你 825	台臺 761	牠它 603	斂□ 440	瞭了 383	佈布 367	昇升 325	裡里 319	週周 270	污汙 248	裏裡 220	周週 203	註注 196	夸誇 194	秘祕 183	佔占 181	儘盡 178	唸念 175	繫系 155
SA-300	裡里 1544	臺台 994	妳你 825	牠它 634	斂□ 440	秘祕 355	瞭了 353	佈布 310	佔占 263	註注 239	污汙 237	周週 234	台臺 223	念唸 221	週周 212	昇升 206	裏裡 202	升昇 196	夸誇 194	証證 154

六、結論

本文提出以雜訊通道/訊息重建模型配合標準語料庫來自動評測語言處理系統，並實際應用於網際網路中文處理的兩個問題：繁簡字碼轉換與第八位元重建；並利用研究院平衡語料庫 1.0 版及 2.0 版，進行大規模實驗，顯示所設計的雜訊通道/訊息重建自動評測模型確為實用可行。

附 註：本文係工業技術研究院電腦與通訊工業研究所執行經濟部委託之前瞻性資訊與網路技術研究計畫成果之一。

參考文獻

[1] C.-H. Chang, Bidirectional Conversion between Mandarin Syllables and Chinese Characters. In Proceedings of ICCPCOL-92, Florida, USA, pp. 174-181, 1992.

[2] C.-H. Chang, Corpus-based Adaptation for Chinese Homophone Disambiguation. Proceedings of Workshop on Very Large Corpora, pp. 94-101, 1993.

[3] H.-H. Chen and Y.-S. Lee, An Adaptive Learning Algorithm for Task Adaptation in Chinese Homophone Disambiguation, Computer Processing of Chinese and Oriental Languages, 9(1):49-58, 1995.

[4] J. Guo, and H.-C. Lui, PH: a Chinese Corpus for Pinyin-Hanzi Transcription, TR93-112-0, Institute of Systems Science, National University of Singapore, 1992.

[5] 黃居仁等,中央研究院平衡語料庫簡介,In Proceedings of ROCLING VIII, pp. 81-99, 1995.

[6] M.D. Kernighan, K.W. Church, and W.A. Gale, A Spelling Correction Program Based on a Noisy Channel Model. In Proceedings of COLING-90, pp. 205-210, 1990.

[7] S.-D. Chen, An OCR Post-Processing Method Based on Noisy Channel, Ph.D. Dissertation, National Tsing Hua University, Hsinchu, Taiwan, 1996.

[8] J. Guo, 漫談萬維網與萬國文字. In the COLIPS Internet Seminar Souvenir Magazine, Singapore, 1996.

[9] S.-K. Huang, big5fix-0.10, 1995. ftp://ftp.nctu.edu.tw/Chinese/ifcss/software/unix/c-utils/big5fix-0.10.tar.gz

[10] 楊道沅、扶良文，一個大陸漢字與台灣漢字文體文件間的智能轉換系統，中文信息學報，6(2):26-34，1992。

[11] C.-H. Chang, Word Class Discovery for Contextual Postprocessing of Chinese Handwriting Recognition. In *Proceedings of COLING-94,* Japan, pp. 1221-1225, 1994.

[12] C.-H. Chang and C.-D. Chen, Automatic Clustering of Chinese Characters and Words. In *Proceedings of ROCLING VI,* pp. 57-78, Taiwan, 1993.

[13] C.-H. Chang and C.-D. Chen, Application Issues of SA-class Bigram Language Models, *Computer Processing of Oriental Languages*, 10(1):1-15, 1996.

[14] C.-H. Chang, Simulated Annealing Clustering of Chinese Words for Contextual Text Recognition, *Pattern Recognition Letters*, 17:57-66, 1996.

[15] 臧遠侯，期待兩岸書同文：如何突破繁簡之間的障礙，時報文化，1996。

[16] 蕭佳賓等，「兩岸常用中文資訊名詞對照表及兩岸中文資訊內碼對照轉碼表之編擬」研究計畫期末報告，1993.