Corpus-based Evaluation of Language Processing Systems Using Information Restoration Model

Chao-Huang Chang
E000/CCL, Building 51, Industrial Technology Research Institute
Chutung, Hsinchu 31015, TAIWAN, R.O.C.
changch@e0sun3.ccl.itri.org.tw

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.

1. Introduction

In 1992 (Chang 1992), we proposed a concept of *bidirectional conversion,* using corpora for automatic evaluation of accuracy of syllable to character conversion systems. After that, the concept was extended to an adaptation mechanism for these systems, which has stimulated some following researches (Chen and Lee 1995). In the recent years, several standard Chinese corpora, such as NUS's PH corpus (Guo and Lui 1992) and Academia Sinica's sinica corpus version 1.0 (Huang *et al.* 1995), 2.0, and 3.0 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 (Kernighan *et al.* 1990, Chen 1996)/information restoration mode l 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.

Internet and World Wide Web are very popular in these days. However, computer and network are not designed for the coding of huge number of Chinese ideographic characters, since they are originated in the western world. For example, the popular ASCII code is a seven-bit standard, and a byte only has eight bits. Obviously, they can not encode the thousands of Chinese characters in a natural way. The situation is worsened due to the political separation of the Chinese Mainland and Taiwan. The Mainland and Taiwan use different styles of Chinese characters (simplified in the Mainland and traditional in Taiwan), and also invent different standards for Chinese character coding. This situation has caused several serious problems in Chinese information processing on the Internet (Guo 1996). In order to fit in different Chinese environments, usually more than one version of web pages are provided, one for English, and the others for Chinese. Chinese versions of web pages are encoded in either BIG5 (Taiwan standard) or GB (Mainland standard). Furthermore, Unicode version would become popular in the near future. In this paper, we will deal with two of Chinese processing problems on the Internet: the 8-th bit restoration of BIG-5 code through non-iso8859-1 channel, and GB-BIG5 code conversion.

BIG-5 code is one of the most popular Chinese character coding schemes used in computer network. It is a double-byte coding, the high byte ranges from (hexadecimal) A1 to FE, 8E to A0, and 81 to 8D; and the low byte from 40 to 7E, and from A1 to FE. The most and secondary commonly used Chinese characters are encoded in A440 to C67E, and C940 to F9D5, respectively; the other ranges are for special symbols and used-defined characters.

In the Chinese mainland, the most popular coding for simplified Chinese characters is GB2312-80, also called GB Code. It is also a double-byte coding, the coding ranges for high byte and low byte are the same, (hexadecimal) A1 to FE.

In most international computer networks, the electronic mails are transmitted through 7-bit channels (so called non-iso8859-1). Thus, if messages coded in BIG5 are transmitted without further encoding (using tools like *uuencode*), the receiver side will only see some *random code* messages. In the literature, little work can be found in studying this problem. S.-K. Huang of NCTU (Hsinchu) designed a shareware called Big5fix (Huang 1995), which is the only previous solution we can find for solving the problem. The input file for Big5fix is supposed to be 7-bit file. Big5fix divides the input into regions of two types: English Region and Chinese Region. The characters in the Chinese regions are reconstructed based on collected character unigrams, bigrams, trigrams and their occurrence counts. Huang estimated the reconstruction accuracy to be 90 percent (95% for Chinese region and 80% for English region). It is well known that sharewares are provided without charge for the general public. The accuracy rates are estimated without large-scale experiments. Our proposed corpus-based evaluation method based on information restoration can be used for this purpose, if the large-scale standard corpora are available.

In addition to automatic evaluating the accuracy rate of Big5fix, we will describe an intelligent 8-th bit reconstruction system, in which statistical language models are used for resolving ambiguities. (Note that there is no similar ambiguity in a pure GB text, in which both high bits of the two bytes are set. As one of the reviewers points out that practical GB documents may be a mixture of ASCII text and GB codes. In that case, the 8-th bit reconstruction problem exists if the channel is not 8-bit clean. However, the problem would need a method to separate ASCII text from GB codes. It is actually out of the scope of this study.)

In comparison, the GB-BIG5 conversion problem, converting simplified characters to traditional characters, is well known and especially important in the days that information flows across the strait rapidly and in a great volume. In addition to book-formed dictionaries or manuals of traditional character-simplified character correspondences, many automatic conversion systems have been designed. Some of the sharewares and products are listed: HC Hanzi Converter shareware, KanjiWeb (漢字通), NJStar (南極星), AsiaSurf (亞洲通), and UnionWin (亞洲心). However, the commonly used tools in the Internet are still one-to-one code converter. Therefore, we can easily find many annoying GB-BIG5 conversion errors in the articles of some newsgroups such as alt.chinese.text.big5 or articles in the BIG5 version of HuaXiaWenZai (華夏文摘). Some typical errors are listed below: 里(裡)、几(幾)、术(術)、准(準)、系(係)、划(劃)、采(採)、制(製). In addition automatic evaluation of the HC converter and KanjiWeb, we will also introduce a new intelligent GB-BIG5 converter. The statistical Chinese language models used in the system include inter-word character bigram (IWCB), and simulated-annealing clustered word-class bigram (Chang 1994, Chang and Chen 1993).

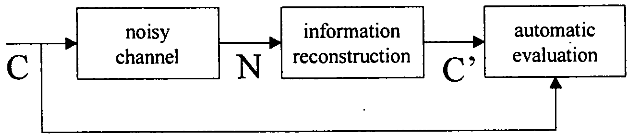

Figure 1: The proposed model

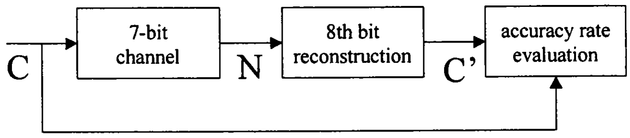

Figure 2: The proposed model for 8th bit reconstruction

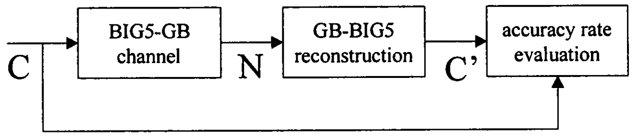

Figure 3: The proposed model for GB-BIG5 conversion

2. Information Restoration Model for Automatic Evaluation

Extending the concepts of 'bi-directional conversion', the proposed corpus-based evaluation method applies the information restoration model for automatically evaluating the performance of various natural language processing systems. As shown in Figure 1, a language processing system is considered as an information restoration process through a noisy channel. Feeding a large-scale standard corpus C into a simulated noisy channel, we can obtain a noisy version of the corpus N. Using N as the input to the language processing system (i.e., the information restoration process), we obtain the output results C'. After that, the automatic evaluation module compares the original corpus C and the output results C', and computes the performance index, accuracy, automatically.

The proposed evaluation model would have a near perfect result (obtaining real performance), if the simulation of noisy channel approaches to perfect. The perfect simulation would be one-to-one correspondence, or a process with near 100% accuracy. For example, for the syllable-to-character conversion system, the noisy channel, character-to-syllable conversion, is not a one-to-one process (there are lots of PoYinZi, homographs). However, it is not difficult to develop a character-to-syllable converter with an accuracy higher than 98% (Chang 1992, Chen and Lee 1995). Thus, the proposed corpus-based evaluation method is readily applied to estimate the conversion accuracy of a syllable-to-character conversion system. In fact, the proposed model can be applied to various types of language processing systems. Typical examples include linguistic decoding for speech recognition, word segmentation, part-of-speech tagging, OCR post-processing, machine translation, and two problems we will study in this article: 8-th bit reconstruction for BIG5 code, and GB-to-BIG5 character code conversion.

The noisy channel simulation of the 8-th bit reconstruction process is perfect, i.e., one-to-one. The only thing the simulation needs to do is to set the 8-th bit of all bytes to zero. Thus, the proposed corpus-based evaluation method can be ideally applied to the problem. The results would be completely correct. Figure 2 illustrates the proposed model for the 8-th bit reconstruction for BIG5 code.

It is a little complex to simulate the noisy channel for the GB-BIG5 code conversion problem. Not only some traditional characters can be mapped to more than one simplified characters (e.g., 乾⇨干、乾; 覆⇨复、覆), but also more other characters can not find a suitable simplified character to map. Nevertheless, the average accuracy rate for the noisy channel simulation still approaches to 100%, based on occurrence frequency in large corpora. The proposed model is still applicable to the problem, as shown in Figure 3.

3. Preparation of Standard Corpora

In this article, we will use the Academia Sinica Balanced Corpora, versions 1.0 (1995 released, 2 million words) and 2.0 (1996 released, 3.5 million words), to verify our proposed corpus-based evaluation model. Some statistics of the two corpora are listed in Table 1.

Sinica Corpus	Size(bytes)	#files	#sentences	#words	#char.(inclu. symbols)	#char. (Hanzi only)
version 1.0	44,525,299	67	284,455	1,342,861	3,347,981	2,953,065
version 2.0	84,256,391	253	411,470	1,946,958	4,834,933	4,143,021

Table 1: Academia Sinica Balanced Corpora, versions 1.0 and 2.0

The word segmentation and sentence segmentation are used as originally provided by

Academia Sinica. The word segmentation follows the proposed standard by ROCLING, which is an earlier version of the Segmentation Standard for Chinese Natural Language Processing (Draft). The part-of-speech tag set is a 46-tag subset simplified from the CKIP tag set (Huang *et al.* 1995). However, the word segmentations and part-of-speech tags are not used in our experiments. The following steps are used for restoring the text with sentence segmentation:

1. Use *grep* (a Unix tool) to filter out the article classification headers, i.e., lines with leading %%; those sentence separator lines (lines filled with '*') are also removed.
2. Use a small program called *extract-word* to extract the words in a sentence; part-of-speech information has been removed. Output examples are something like "我 起來 了 ，"; "太陽 也 起來 了 。"
3. Concatenate words in a sentence into a character string, e.g., "我起來了，"; and concatenate all files into a single huge file.
4. Replace all user-defined special characters and non-BIG5 code with a special symbol '□'.

After pre-processing, the corpus becomes a single file, one sentence per line, and all characters are double-byte BIG5 code. The statistics shown in Table 2 are calculated based on pre-processed version of the corpora.

4. The 8-th Bit Reconstruction

4.1 System Design

The 8-th bit reconstruction problem has been described in Sections 1 and 2. We will not repeat the statement here. To simulate the noisy channel, we simply set zero the 8-th bit of each byte in the input. It can be done in a few lines of program. We will use Big5fix as a baseline system, and develop an intelligent 8-th bit reconstruction system. The system resolves the ambiguity problem using statistical Chinese language models. The basic architecture follows our previous approach called 'confusing set substitution and language model evaluation' (Chang 1994, 1996, Chang and Chen 1993, 1996). As shown in Figure 4, the characters in the input are substituted by corresponding confusing character sets, sentence by sentence. In this way, numbers of sentence string candidates for an input sentence are generated. Then the string candidates are evaluated through a corpus-based statistical language model. The candidate with the highest score (probability) is chosen to be the output of the system. Here, the step of 'confusing set substitution' can be considered as an inverse simulation of 'noisy channel'.

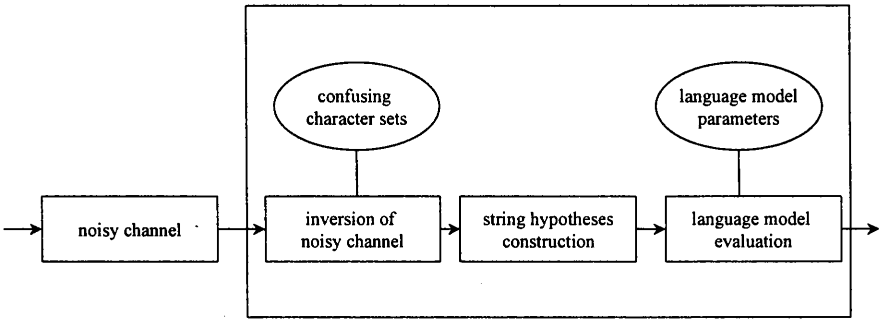

Figure 4. The'confusing set substitution and language model evaluation' approach

For the reconstruction problem, the 'confusing set' is very easy to set up. Since BIG5 is a double-byte coding, we have at most two hypotheses for each character: the 8-th bits of all high-bytes are set to 1, and the 8-th bits of the low-bytes can be either 0 or 1 (dependent on code regions). For example, the inverse simulation confusing set for 2440 (hex) contains two characters a440「一」and a4c0「分」; but the confusing set for 2421 (hex) only contains a character a4a1「丑」(a421 is out of coding region). In the system, we set up confusing sets for each of the 13,060 Chinese characters (including 7 so-called Eten characters). Among them, 10,391 confusing sets contain two characters, while the other 2,669 contain only one character. The statistical language model used in the system is an inter-word character bigram (IWCB) model (Chang 1993). The model is slightly modified from the word-lattice-based character bigram model in Lee *et al.* (1993). Basically, it approximates the effect of word bigram by applying character bigram to the boundary characters of adjacent words. For details of the IWCB model, please refer to Lee *et al.* (1993) and Chang (1993).

4.2 Experimental Results

Table 2 compares the corpus-based evaluation results (number of errors, error rate %) of Big5fix and our intelligent 8-th bit reconstruction system (called CCL-fix).

Sinica Corpus	Samples	#char.	Big5fix		CCL-fix	
Version 1.0	incl. symbols	3,347,981	125,915	3.76	57,862	1.72
	Hanzi	2,953,065	100,006	3.38	53,729	1.81
Version 2.0	incl. symbols	4,834,933	173,544	3.58	71,549	1.48
	Hanzi	4,143,021	111,809	2.69	70,758	1.70

Table 2: Corpus-based evaluation results, Big5fix vs. CCL-fix

As we can see in Table 2, the Hanzi reconstruction rates of Big5fix for Sinica Corpora versions 1.0 and 2.0 are 96.62% and 97.31%, respectively. They are higher than 95% estimated by Huang by 1.62%, 2.31%. The reconstruction rates of CCL-fix are 98.19% and 98.30%, respectively. It shows that the IWCB language model is indeed superior to the counts of character unigram and bigram.

Table 4 lists the reconstruction error analysis for Sinica corpus 1.0 by the two systems. The table shows only the top 20 types of errors with highest frequency. Each entry shows the original character, the reconstructed character, and its occurrence count. For example, the most frequent error made by Big5fix is wrongly reconstructing ‘分’ as ‘一’, with 3,007 occurrences.

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

Table 3: Reconstruction error analysis for Sinica corpus 1.0, Big5fix vs. CCL-fix.

5. GB-Big5 Conversion

5.1 System Design

Three different simulations of the noisy channel for the GB-BIG5 conversion problem are used in our experiments: (1) HC Hanzi Converter, version 1.2u, developed by Fung F. Lee and Ricky Yeung, (2) HC, revised version: the conversion table is slightly enhanced; and (3) MultiCode of KanziWEB. These three systems all use the table-lookup conversion approach. Thus, the one-to-many mapping problem is not dealt with, and lots of errors can be found when converting GB code back to BIG5.

Table 4 compares the corpus-based evaluation results (number of errors, error rate %) of the three systems: HC1.2u, HC revised, and KanjiWEB .

Sinica Corpus	Samples	# char.	HC1.2u		HC revised		KanjiWEB	
Version 1.0	incl. symbols	3,347,981	271,986	8.12%	46,162	1.37%	29,531	0.87%
	Hanzi	2,953,065	43,155	1.46%	43,070	1.45%	29,076	0.98%
Version 2.0	incl. symbols	4,834,933	403,954	8.35%	68,047	1.40%	43,705	0.90%
	Hanzi	4,143,021	60,113	1.45%	60,031	1.45%	40,561	0.98%

Table 4: Corpus-based evaluation results for HC1.2u, HC revised, and KanjiWEB

To deal with the one-to-many mapping problem in GB-BIG5 conversion, we have developed an intelligent language model conversion method, taking context into account. In the literature, Yang and Fu (1992) presented an intelligent conversion system between Mainland Chinese text files and Taiwan Chinese text files. Their basic approach is (1) build tables by classification; (2) compute scores by levels. However, they resolve ambiguities by asking, instead of using statistical language models. We still take the 'confusing set substitution and language model evaluation' approach. The Chinese language models we used are (1) IWCB model, (2) SA-class bigram model (Chang 1994, 1996, Chang and Chen 1993, 1996) . In the experiments, we use two versions of the SA-class bigram model, with 200 and 300 word-classes, respectively. They will be denoted as SA-200 and SA-300 models.

To simulate the inverse noisy channel, we must set up confusing sets, that is, collection of variants and equivalent characters. In other words, it is a simulation of one-to-many mapping from GB to BIG5. We have found three sources of variants and equivalent characters: (1) the YiTiZi file in HC version 1.2u, (2) Annotation table of simplified characters in the mainland by Zang (1996), (3) Appendix 10 of Hsiao et al. (1993)'s project report. Combining the three sources, we have arranged four versions of confusing sets (A, B, C, and D), which are used and compared in he experiments. Some statistics of the four versions of confusing sets are shown in Table 5. The column label 'n-way' shows the number of characters for which there are *n* characters in their confusing sets.

Confusing Set	Source	1-way	2-way	3-way	4-way	5-way
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

Table 5: Statistics of the four versions of confusing sets

5.2 Experimental Results

Table 6 compares the corpus-based evaluation results (number of errors, error rate %) of the three language models and four versions of confusing sets for GB-BIG5 conversion. (The input is provided by the HC Revised.)

Sinica Corpus	Number of char.	IWCB				SA-200				SA-300			
		A	B	C	D	A	B	C	D	A	B	C	D
Version 1.0	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%
Version 2.0	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%

Table 6: Comparing four versions of confusing sets with three language models

We can see that the IWCB model achieved the best performance for the problem. The SA-300 model has comparative performance, while the SA-200 model is relatively weak. However, we must notice that the three intelligent conversion methods are all superior to KanjiWEB's one-to-one mapping method. The error rates are more than doubled in one-to-one mapping system. Among the four versions of confusing sets, version B performs better than the others. Version C and version D have a larger set of confusing characters than version B, but their performance can not reflect that. The reason might be larger sets make more unnecessary confusions. In contrast, Version A has clearly insufficient numbers of confusing characters.

Table 7 lists the conversion error analysis for Sinica corpus 2.0 by the four systems

(HC1.2u, KanziWEB, IWCB, and SA300 with confusing set version B. The notation is similar to that in the above section. □ or blanks denote that no corresponding character, a1bc(hex) or a140(hex).

Rank	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	術朮 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
Kanzi WEB	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	欸 440	佈布 439
IWCB /B	臺台 885	妳你 825	台臺 761	牠它 603	欸□ 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-300B	裡里 1544	臺台 994	妳你 825	牠它 634	欸□ 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

Table 7: conversion error analysis for Sinica corpus 2.0 by the four systems

6. Concluding Remarks

In this article, we have presented a corpus-based information restoration model for automatic evaluation of NLP systems, and applied the proposed model 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.

Acknowledgements

This paper is a partial result of the project no. 3P11200 conducted by the ITRI under sponsorship of the Minister of Economic Affairs, R.O.C. Previous versions of the paper *in Chinese* appeared in JSCL-97 (Beijing) and Communications of COLIPS (Singapore). One of the reviewers suggested that the title of this paper be changed to “Noisy Channel Models for Corrupted Chinese Text Restoration and GB-to-Big5 Conversion”. For consistency, we have kept the original title. Thanks are due to the reviewers for the constructive and helpful comments.

References

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

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

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

Chang, C.-H., Word Class Discovery for Contextual Post-processing of Chinese Handwriting Recognition. In Proceedings of COLING-94, Japan, 1994, pp. 1221-1225.

Chang, C.-H., Simulated Annealing Clustering of Chinese Words for Contextual Text Recognition, Pattern Recognition Letters, 17, 1996, pp.57-66.

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

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

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

Guo, J., On World Wide Web and its Internationalization. In the COLIPS Internet Seminar Souvenir Magazine, Singapore, 1996.

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

Hsiao J.-P. et al., Research Project Report on Common Chinese Information Terms Mapping and Computer Character Code Mapping across the Strait, 1993. (in Chinese)

Huang C.-R. et al. Introduction to Academia Sinica Balance Corpus, In Proceedings of ROCLING VIII, 1995, pp. 81-99. (in Chinese)

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

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

Lee L.-S. et al., Golden Mandarin (II) - an Improved Single-Chip Real-time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary. In Proceedings of ICASSP-93, II, 1993, pp. 503-506.

Yang, D. and L. Fu, An Intelligent Conversion System between Mainland Chinese Text Files and Taiwan Chinese Text Files, Journal of Chinese Information Processing, 6(2), 1992, pp.26-34. (in Chinese)

Zang, Y.-H., *How to Break the Barrier between Traditional and Simplified Characters*, China Times Culture, 1996. (in Chinese)