Computational Linguistics and Chinese Language Processing
vol. 3, no. 2, August 1998, pp. 79-92

# Noisy Channel Models for Corrupted Chinese Text Restoration and GB-to-Big5 Conversion

**Chao-Huang Chang***

## Abstract

In this article, we propose a noisy channel/information restoration model for error recovery problems in Chinese natural language processing. A language processing system is considered as an information restoration process executed through a noisy channel. By 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 can 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 (i.e., accuracy) automatically. The proposed model has been applied to two common and important problems related to Chinese NLP for the Internet: corrupted Chinese text restoration and GB-to-BIG5 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 this article, we present a noisy channel (Kernighan *et al.* 1990, Chen 1996) / information restoration model for automatic evaluation of error recovery systems in Chinese natural language processing. The proposed model has been applied to two common and important problems related to Chinese NLP for the Internet: corrupted Chinese text restoration (i.e., 8-th bit restoration of BIG-5 code through a non-8-bit-clean channel), and GB-BIG5 code conversion. The concept follows our previous work on bidirectional conversion (Chang 1992) and corpus-based adaptation for Chinese homophone disambiguation (Chang 1993, Chen and Lee 1995). Several standard Chinese corpora are available to the public, such as NUS's PH corpus (Guo and Lui 1992) and Academia Sinica's sinica corpus (Huang *et al.* 1995). These corpora can be used for objective evaluation of NLP systems. Sinica Corpora version 1.0 and 2.0 were used in the

*E000/CCL, Building 51, Industrial Technology Research Institute, Chutung, Hsinchu 31015, Taiwan, R.O.C. E-mail: changch@e0sun3.ccl.itri.org.tw

experiments. The results show that the proposed model is useful and practical.

The Internet and World Wide Web are very popular these days. However, computers and networks are not designed for coding huge numbers of Chinese ideographic characters since they originated in the western world. This situation has caused several serious problems in Chinese information processing on the Internet (Guo 1996). While the popular ASCII code is a seven-bit standard which can be easily encoded in a byte (eight bits), thousands of Chinese characters have to be encoded in at least two bytes. In this paper, we explore two error recovery problems for Chinese processing problems on the Internet: corrupted Chinese text restoration and GB-to-BIG5 conversion.

Mainland China and Taiwan use different styles of Chinese characters (simplified in Mainland China and traditional in Taiwan) and have also invented different standards for Chinese character coding. In order to fit different Chinese environments, more than one version of a web page is usually provided, one in English, and the other(s) in Chinese. Chinese versions of web pages are encoded in either BIG5 (Taiwan standard) or GB (Mainland China standard). Furthermore, the Unicode version will become popular in the near future.

BIG-5 code is one of the most popular Chinese character code sets used in computer networks. It is a double-byte coding; the high byte range is from (hexadecimal) A1 to FE, 8E to A0, and 81 to 8D; and the low byte range is from 40 to 7E, and from A1 to FE. The most and second most 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. On the Chinese mainland, the most popular coding for simplified Chinese characters is the GB Code. It is also a double-byte coding; the high byte and low byte coding ranges are the same, (hexadecimal) A1 to FE.

In most international computer networks, electronic mail is transmitted through 7-bit channels (so called non-8-bit-clean). 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 on this problem. S.-K. Huang of NCTU (Hsinchu) designed a shareware program called Big5fix (Huang 1995), which is the only previous solution we can find for solving this problem. The input file for Big5fix is supposed to be a 7-bit file. Big5fix divides the input into regions of two types: an English Region and a Chinese Region. The characters in the Chinese region are reconstructed based on collected character unigrams, bigrams, trigrams and their occurrence counts. Huang estimated the reconstruction accuracy to be 90 percent (95% for the Chinese region and 80% for the English region). It is well known that shareware programs are provided free of 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 a large-scale standard corpus is available.

In addition to automatic evaluation of 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 reviewer has pointed out, 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, solving the problem will require a method of separating ASCII text from GB codes. This is actually beyond the scope of this study.)

In comparison, the GB-BIG5 conversion problem, that is, converting simplified characters to traditional characters, is well known and especially important nowadays since information flows rapidly back and forth across the strait and in a great volume. In addition to dictionaries in book form and manuals of traditional character-simplified character correspondences, many automatic conversion systems have been designed. Some of the shareware programs and products are as follows: the HC Hanzi Converter shareware, KanjiWeb ( 漢字通 ), NJStar ( 南極星 ), AsiaSurf ( 亞洲通 ), and UnionWin ( 亞洲心 ). However, the tools on the Internet commonly used are still one-to-one code converters. Therefore, we can easily find many annoying GB-BIG5 conversion errors in articles published in some newsgroups, such as alt.chinese.text.big5 or articles published in the BIG5 version of HuaXiaWenZai ( 華夏文摘 ). Some typical errors are: " 家里 ( 裡 )", " 几 ( 幾 ) 個 ", " 技朮 ( 術 )", " 標准 ( 準 )", " 關系 ( 係 )", " 計划 ( 劃 )", " 采 ( 採 ) 用 ", and " 制 ( 製 ) 造 ". In the above examples, a string contains a two-character word (outside the parentheses) and a single-character correction (inside the parentheses). In addition to automatic evaluation performed by the HC converter and KanjiWeb, we will introduce a new intelligent GB-BIG5 converter. The statistical Chinese language models used in the new converter include the inter-word character bigram (IWCB) and the simulated-annealing clustered word-class bigram (Chang 1994, Chang and Chen 1993).

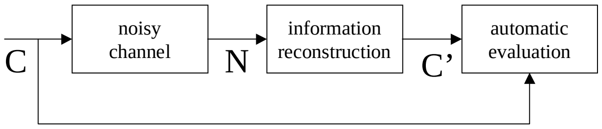

Figure 1: The proposed model.

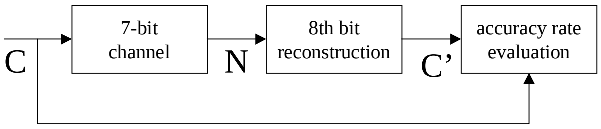

Figure 2: The proposed model for 8th bit reconstruction.

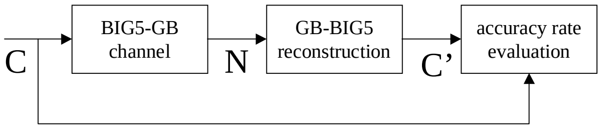

Figure 3: The proposed model for GB-BIG5 conversion.

## 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 evaluation of the performance of various natural language processing systems. As shown in Figure 1, a language processing system is considered to be an information restoration process executed through a noisy channel. By 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 can 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 (i.e., accuracy) automatically.

The proposed evaluation model will obtain have near perfect results (obtain real performance) if the simulation of a 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, a noisy channel, that is, character-to-syllable conversion, is not a one-to-one process (there are many PoYinZi, that is, homographs). However, it is not difficult to develop a character-to-syllable converter with accuracy higher than 98% (Chang 1992, Chen and Lee 1995). Thus, the proposed corpus-based evaluation method can be 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. Indeed, the proposed model has its limitations. For problems where we can not perform nearly perfect noisy channel simulation, the performance (of either error recovery or evaluation) is inaccurate. Speech recognition may be one such problem (as one reviewer pointed out.)

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 is ideal for application to this problem. The results will be completely correct. Figure 2 illustrates the proposed model for 8-th bit reconstruction for BIG5 code.

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

## 3. Preparation of Standard Corpora

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

***Table 1.*** *Academia Sinica Balanced Corpora, versions 1.0 and 2.0.*

| 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 |

Word segmentation and sentence segmentation were used as originally provided by the Academia Sinica. The word segmentation follows the proposed standard set 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 were not used in our experiments. The following steps were used to restore the text using sentence segmentation:

(1) Grep (a Unix tool) was used to filter out the article classification headers, i.e., lines with leading %%; those sentence separator lines (lines filled with '*') were also removed.

(2) A small program called extract-word was used to extract the words in a sentence; part-of-speech information was removed. Output examples were something like " 我 起來 了 ， " ; " 太陽 也 起來 了 。 "

(3) Words in a sentence into a character string, e.g., " 我起來了， ", and all files were concatenated into a single huge file.

(4) All user-defined special characters and non-BIG5 code were replaced with a special symbol ' □ '.

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

## 4. The 8-th Bit Reconstruction

### 4.1 System Design

The 8-th bit reconstruction (also called corrupted Chinese text restoration) problem has been described in Sections 1 and 2. We will not repeat the description here. To simulate a noisy channel, we simply set to zero the 8-th bit of each byte in the input. This could be done using a program of a few lines. We used Big5fix as a baseline system and developed 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 replaced with corresponding confusing character sets, sentence by sentence. In this way, the number of sentence string candidates for an input sentence is generated. Then, the string candidates are evaluated using a corpus-based statistical language model. The candidate with the highest score (probability) is chosen to be the output of the system. Here, the 'confusing set substitution' step can be considered as inverse simulation of a 'noisy channel'.

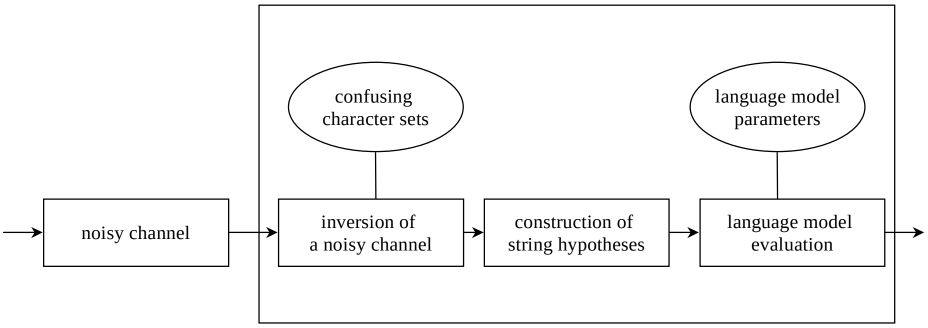
confusing character sets
language model parameters
noisy channel
inversion of a noisy channel
construction of string hypotheses
language model evaluation

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 code, 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 (depending on the code region). 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 one character a4a1 「丑」 (a421 is outside of the coding region). In the system, we set up confusing sets for each of the 13,060 Chinese characters (including the 7 so-called Eten characters). Among them, 10,391 confusing sets contain two characters while the other 2,669 confusing sets contain only one character.

The statistical language model used in our system is an inter-word character bigram (IWCB) model (Chang 1993). The model is slightly modified from the word-lattice-based character bigram model of Lee *et al.* (1993). Basically, it approximates the effect of a word bigram by applying a character bigram to the boundary characters of adjacent words. The IWCB model is a variation of the *word-lattice-based Chinese character bigram* proposed by Lee *et al.* (1993). The path probability is computed as the product of the word probabilities and inter-word character bigram probabilities of the words in the path. For path H: $W_1 = W_{i_1 j_1}, \ldots, W_F = W_{i_F j_F}$ , the path-probability estimated by the language model is

$$P_{LM}(H) = (\sum_{k=1}^{F} P(W_k))) \times \sum_{k=2}^{F} P(C_{i_k} | C_{j_{k-1}})$$

where Cik and Cjk are the first and last characters of the k-th word, respectively. This model is one of the best among the existing Chinese language models and has been successfully applied to Chinese homophone disambiguation and linguistic decoding. 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 (the number of errors and the error rate %) of Big5fix and our intelligent 8-th bit reconstruction system (called CCL-fix).

***Table 2.*** *Corpus-based evaluation results, Big5fix vs. 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 |

As in Table 2 shows, the Hanzi reconstruction rates of Big5fix for the Sinica Corpora versions 1.0 and 2.0 are 96.62% and 97.31%, respectively. They are higher than the 95% rate estimated by Huang by 1.62%, 2.31%. The reconstruction rates of CCL-fix are 98.19% and 98.30%, respectively. This shows that the IWCB language model is indeed superior to the counts of character unigrams and bigrams. Note that the 1991 UD newspaper corpus (1991ud), consisting of more than seven million characters, was used to train the character bigrams in the IWCB model and the word bigrams used in simulated annealing word clustering. Some statistics for the **1991ud** corpus are as follows: 579,123 sentences, 7,312,979 characters, 4,761,120 word-tokens, and 60,585 word-types. The **1991ud** corpus is independent of the Sinica Corpus in both its publisher and sample date.

Table 4 lists the reconstruction error analysis results for the Sinica Corpus 1.0 obtained using the two systems. The table shows only the top 20 most frequent types of errors. 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.

***Table 3.*** *Reconstruction error analysis for the Sinica Corpus 1.0, Big5fix vs. CCL-fix.*

| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Big5<br>fix | 分一<br>3007 | 化了<br>1540 | 林者<br>1481 | 外全<br>893 | 全外<br>819 | 區記<br>797 | 色在<br>792 | 股松<br>771 | 來沒<br>734 | 省某<br>723 | 西多<br>722 | 價語<br>715 | 代用<br>712 | 反力<br>709 | 石加<br>676 | 吳找<br>672 | 十天<br>664 | 船爽<br>611 | 油迎<br>611 | 村困<br>601 |
| CCL<br>fix | 一分<br>2298 | 了化<br>1388 | 分一<br>1375 | 又太<br>1327 | 沒來<br>1325 | 外全<br>1209 | 天十<br>1194 | 每並<br>887 | 多西<br>638 | 林者<br>577 | 十天<br>530 | 代用<br>491 | 象僅<br>484 | 某省<br>465 | 叫件<br>458 | 沙事<br>396 | 士方<br>386 | 女月<br>376 | 命所<br>359 | 吧扭<br>343 |

## 5. GB-to-Big5 Conversion

### 5.1 System Design

Three different simulations of the noisy channel for the GB-BIG5 conversion problem were performed in our experiments: we used (1) HC Hanzi Converter, version 1.2u, developed by Fung F. Lee and Ricky Yeung; (2) HC, revised version, in which the conversion table is slightly enhanced; and (3) the 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 many errors can be found after converting GB code back to BIG5.

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

***Table 4.*** *Corpus-based evaluation results for 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% |

To deal with the one-to-many mapping problem in GB-BIG5 conversion, we have developed an intelligent language model conversion method which takes context into account. In the literature, Yang and Fu (1992) presented an intelligent system for conversion between Mainland Chinese text files and Taiwan Chinese text files. Their basic approach is to (1) build tables by means of classification; and (2) compute scores level by level. However, they resolve ambiguities by asking (the user), instead of using statistical language models. We take the 'confusing set substitution and language model evaluation' approach. The Chinese language models we use are (1) the IWCB model (introduced above) and (2) the SA-class bigram model. In the SA-class bigram model, the words in the dictionary are automatically separated into $N_C$ word classes using a simulated-annealing word clustering procedure (Chang 1994, 1996, Chang and Chen 1993, 1996). The language models usually seek the optimal path in a word-lattice formed by candidate characters. The path probability of a word-lattice path is the product of lexical probabilities and contextual SA-class bigram probabilities. For a path of F words H = $W_1, W_2, \cdots, W_F$, the path-probability estimated by the language model is

$$P_{LM}(H) = (\sum_{i=1}^{F} P(W_i \mid \phi(W_i))) \times (\sum_{i=2}^{F} P(\phi(W_i) \mid \phi(W_{i-1})))$$

where $\phi(W_i)$ is the word class which $W_i$ belongs to.

In the experiments, we used two versions of the SA-class bigram model, with $N_C$ =200 and $N_C$ =300, respectively. They will be denoted as the SA-200 and SA-300 models. The corpus for word clustering, **1991ud**, was first segmented automatically into sentences, and then into words by our Viterbi-based word identification program VSG (Chang and Chen 1993). The same lexicon and word hypothesizer were used in the language models.

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

***Table 5.*** *Statistics of the four versions of 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 |
| D | (1)(2)(3) | 12144 | 777 | 117 | 15 | 7 |

## 5.2 Experimental Results

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

***Table 6.*** *Comparison of four versions of confusing sets with three language models.*

| 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<br>0.43% | 10,144<br>0.34% | 12,997<br>0.43% | 12,684<br>0.42% | 15,574<br>0.52% | 13,977<br>0.47% | 16,867<br>0.57% | 16,811<br>0.56% | 13,614<br>0.44% | 10,849<br>0.36% | 13,500<br>0.45% | 13,225<br>0.44% |
| Version 2.0 | 4,143,021 | 17,752<br>0.42% | 14,139<br>0.34% | 18,774<br>0.45% | 18,465<br>0.44% | 21,127<br>0.50% | 18,593<br>0.44% | 23,299<br>0.56% | 23,297<br>0.56% | 18,729<br>0.45% | 15,439<br>0.37% | 19,790<br>0.47% | 19,554<br>0.47% |
| | 468,609<br>(ambiguous) | 17,752<br>3.78% | 14,139<br>3.02% | 18,774<br>4.01% | 18,465<br>3.94% | 21,127<br>4.51% | 18,593<br>3.97% | 23,299<br>4.97% | 23,297<br>4.97% | 18,729<br>3.99% | 15,439<br>3.29% | 19,790<br>4.22% | 19,554<br>4.17% |

We can see that the IWCB model achieved the best performance for the problem. The SA-300 model had comparative performance while the SA-200 model was relatively weak. However, we must note that the three intelligent conversion methods were all superior to KanjiWEB's one-to-one mapping method. The error rates are more than double those of the other methods in the one-to-one mapping system. Among the four versions of confusing sets, version B performed better than the others. Version C and version D had a larger set of confusing characters than version B, but their performance did not reflect this. The reason might have been that the larger sets make more unnecessary confusion. In contrast, Version A clearly had an insufficient number of confusing characters.

The evaluation did not exclude unambiguous characters. Among the 4,143,021 characters in the Sinica Corpus 2.0, 11.31% (468,609) were found to be ambiguous (316,889 2-way ambiguous, 125,297 3-way, 18,377 4-way, and 7866 5-way ambiguous). That is, a *random (or no-grammar)* language model had a 6.4% error rate. Evaluation of pure ambiguous characters revealed that the random model had an error rate of 55.96% while the best performance achieved by the models was 3.02%(IWCB), 3.29%(SA-300), 3.97%(SA-200), respectively.

Table 7 lists the conversion error analysis for by the four systems (HC1.2u, KanziWEB, IWCB, and SA-300) with confusing set version B. The notation is similar to that used in the above section. □ or blanks denote no corresponding character, a1bc(hex) or a140(hex).

***Table 7.*** *Conversion error analysis for Sinica corpus 2.0 by the four systems*

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

## 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: 8-th bit restoration of BIG-5 code through a non-8-bit-clean channel and GB-BIG5 code conversion. The Sinica Corpora versions 1.0 and 2.0 were used in the experiment. The results show that the proposed model is useful and practical.

## Acknowledgements

This paper is a partial result of 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). The original title of the paper presented at ROCLING XI was "Corpus-based Evaluation of Language Processing Systems Using Information Restoration Model". The current title was suggested by one of the ROCLING XI reviewers. 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. and C.-D. Chen, SEG-TAG: A Chinese word segmentation and part-of-speech tagging system. In *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS '93)*, Fukuoka, Japan, 1993. pp. 319-327.

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)