容错型拼音汉字转换系统 —— 一些调查及基本方案

叶斌　应江黔*　川上裕忠　松本忠博　後藤宗弘

（日本岐阜大学工学部应用情报学科）

（* 日本岐阜大学地域科学部）

摘要：　现存拼音汉字输入系统的设计轻视了用户拼错音的可能性。这在很大程度上降低了用户的工作效率。影响了中文信息处理系统的普及。我们对２２名中国人做调查的结果是每个字的平均出错率约１９％。本文客观地评价了这一现实状况对现存拼音输入系统性能的影响，提出了几种容易实现的或将来可能实现的能够自动订正错误的拼音输入系统的方案

Towards an Error-Correcting Pinyin-Hanzi Conversion System

——— Some Facts and Schemes

BIN YE,　JIANGQIAN YING *
H. KAWAKAMI,　T. MATSUMOTO,　M. GOTO

(Dept. of Electronic and Computer Engineering, Gifu University, Japan)
(* Division of Regional Policy, Gifu University, Japan)

ABSTRACT: In current Pinyin coding based Chinese character (Hanzi) input systems, the probable spelling errors of ordinary users are usually neglected. This actually causes a considerable inefficiency in use and hinders the popularization of Chinese information processing system. A spelling test conducted on 22 Chinese shows an average error rate of 19%. In this paper, we objectively evaluate the performance of current Pinyin input system under the influence of spelling errors, and introduce some schemes for an automatic error-correcting. Pinyin input system that can be easily put into practice.

1 Introduction

Chinese character (Hanzi) input through keyboards can be classified into two patterns, one is phonetics-based-input and the other is texture-based-input. Pinyin code (PYC) is the phonic code of Chinese which is utilized in phonetics-based-input. Although various texture-based-input methods have been developed for professional speedy input, they are difficult to be popularized simply because few people are actually willing to spend the extra time to learn these methods, without mention the difficulty that a non-native user could be faced in using them.

From a long term viewpoint, the authors think the phonetics-based-input is the trend for future Chinese information processing systems. However , current PYC input systems have a serious defect: it neglects the possibility that a user may make mistakes in PYC spelling. This actually causes considerable inefficiency in use and hinders the popularization of Chinese information processing. A spelling test conducted on 24 Chinese shows an average error rate of 19 %. We objectively evalute the performance of current PYC input system under the influence of spelling errors, and introduce some schemes for an automatic error-correcting Pinyin-Hanzi conversion system that can be easily put into practice.

The paper is organized as follows: Section 2 shows some detailed evidence of the necessity of an error-correcting system. Section 3 describes some fundmental probabilistic relations among the user-input PYC, standard PYC and the corresponding character. In Section 4 some basic schemes are proposed for establishing an error-correcting Pinyin-Hanzi conversion system.

2 Need for an Error-correcting System

The process of PYC based Chinese text input can be separated into two steps. One is phonetic identification step that the user transfers his pronunciation of the characters (Hanzi) to be input to PYC. The second step is to convert PYC to characters by the computer. The focus of this paper is on solving error problems caused in step one automatically by the machine in the second step instead of re-typing. In this section, we show evidence that the automatic error-correcting function is essential for ordinary Chinese users.

2.1 A Pinyin Spelling Test

To learn the real situation concerning the efficiency of PYC users, we conducted a test on 24 Chinese students and researchers studying or working at Gifu University, Japan, who are coming from 13 regions distributing from south to north, west to east of the mainland China. (Actually, our work was triggered off by the complaints of costing tremendous time on frequently correcting spelling errors in using a Chinese word processor by serval persons among the test.)

Although complete PYC includes a string of alphabets denoting the basic pronunciation as well as a tone note which assigns one of the four tones to the pronuncaition.

Here we ignore the latter and take only the phonic alphabet string as the PYC. The first reason is that some existing Pinyin-Hanzi conversion systems only take account of the alphabets. The second reason is that the four tones are another source of errors to many users and worth for a separate treatment in due course.

The 24 subjects were asked to write down the PYC for a list of 393 randomly ordered and commonly used characters covering a total of 393 Pin-Yin codes (with the distinction of tones ignored).

Table 1 shows some typical spelling error types and their frequencies. Two subjects

Table 1: Examples of Spelling Errors

Correct Code	I	II	III	IV	V	VI
ying	ying 13	yin 3	yeng 4	en 1	yi 1	
sang	sang 10	shang 8	shan 1	song 3	shan 1	
zhun	zun 5	zhun 14	zuen 2	zhueng 1		
sou	sou 14	shou 5	shuo 1	sho 1	suo 1	
run	run 15	lun 1	yun 2	ruen 1	ren 2	nun 1
xiu	xiu 13	shiou 1	xu 1	xui 3	xiou 4	
nong	nong 11	non 2	rong 1	long 5	lon 2	nen 1
ruan	ruan 14	luan 3	ran 1	yuan 2	ruang 1	nuan 1

seem to have little knowledge of PYC and their data are excluded. Among the rest 22 persons, the best correctness rate is 94% the worst is 48%, and the average is 81%. The average correctness rate of the 22 subjects along with the regions where they come from are shown in Table 2.

Table 2: Correctness Rates

No.	Correct Rate	Region	No.	Correct Rate	Region
1	80.4%	GuanXi	2	66.9 %	Shanghai
3	70.2 %	Shanghai	4	93.6%	JianSu
5	75.3%	GuiZhou	6	79.7%	Beijin
7	85 %	ChangChun	8	94 %	LiaoNin
9	88 %	ShenYang	10	87 %	WuXi
11	80 %	FuJian	12	70.2 %	AnHui
13	48 %	ZheJiang	14	89 %	Shanghai
15	86.8 %	DaLian	16	90.6 %	Beijin
17	71.5 %	ChangChun	18	94.9 %	ChangChun
19	72 %	GuanXi	20	71.8 %	FuJian
21	90.3 %	TianJing	22	88.3 %	LiaoNin

2.2 Implication of the data

Based on our statistics, we objectively estimate the efficienty of Pinyin-Hanzi system without error-correcting function.

- **Word level conversion:**

  The rate that an ordinary Chinese user correctly inputs the PYC for a 2-character word is $0.81^2 \approx 0.66$. This implies that the user has to re-type a word every 3 times. The rate for correctly inputing the Pinyin codes for a 4-character word is $0.81^4 \approx 0.43$. This implies that the user has to re-type a word more than twice. What worse is that the user usually has to repeat the correcting process again and again until he finally hit upon the correct spelling or gives up and looks up a dictionay by counting the strokes of the characters.

- **Sentence level conversion:**

  The rate of correctly inputing a sentence of 10-characters, a moderate length, is about $0.81^{10} \approx 0.12$. This says that any sentence level conversion system without automatical error-correcting function will actually be useless for many Chinese users.

3 Basic Probabilistic Relations

In this section we present formulas for computing basic probability functions which describe the relations among Pinyin codes and characters.

Let $\Gamma = \{\gamma_1, \ldots, \gamma_N\}$ designate the set of standard Pinyin codes (in our case $N = 393$). Let $\Phi = \{\phi_1, \ldots, \phi_M\}$ be the set of input codes collected in our test which actually covers the set of standard codes, i.e., $\Gamma \subset \Phi$, and M is 1210.

From the data of our investigation, we estimated the probability $P(\phi_j|\gamma_i)$ by which a standard code γ_i is spelled as ϕ_j, $i = 1, \ldots, N$, $j = 1, \ldots, M$.

$$P(\phi_j|\gamma_i) = \frac{times\ inputting\ \phi_j\ for\ \gamma_i}{total\ number\ of\ inputting\ for\ \gamma_i} \tag{1}$$

Let $P(\gamma_i)$ denotes the relative frequency of γ_i in Chinese language, which can be estimated from sufficiently large corpus. The likelihood $P(\gamma_i|\phi_j)$ for γ_i to be the correct Pinyin code for a given input ϕ_j can be calculated by Bayes rule as follows:

$$P(\gamma_i|\phi_j) = \frac{P(\phi_j|\gamma_i)P(\gamma_i)}{P(\phi_j)} = \frac{P(\phi_j|\gamma_i)P(\gamma_i)}{\sum_{i=1}^{N} P(\phi_j|\gamma_i)P(\gamma_i)}. \tag{2}$$

Using the Pinyin code frequency data of [4], we have estimated these likelihoods. Table 3 shows part of the result.

For each of the 22 subjects, the average rate that the correct code is contained within the 3 most likely alternatives for a given input is calculated, as is listed in Table 4.

In a conventional Pinyin-Hanzi system, the input code is always assumed to be correct. Table 4 suggests that an error-correcting system should also include those characters with Pinyin codes which are probable alternatives for the input code.

Provided that the data of the character frequencies $P(C_k)$ (C_k denotes a character) in Chinese is available, the probability that C_k is the character to be input, given γ_i as

Table 3: Alternative Standard Codes and Likilihoods

Input	Alt.	Prob.	Input	Alt.	Prob.
shen	shen	0.56	sheng	sheng	0.66
	sen	0.25		shen	0.15
	sheng	0.16		seng	0.12
	seng	0.03		sen	0.07

Table 4: The Rates of Including Best Three Likely Alternatives

No.	Rate	No.	Rate	No.	Rate	No.	Rate
1	97.2%	2	96.2 %	3	91.6 %	4	99 %
5	95.4 %	6	94.9 %	7	98.5 %	8	98.7%
9	99.2 %	10	98.7 %	11	96.7 %	12	98.7%
13	93.9 %	14	99.0 %	15	97.6 %	16	98.7 %
17	94.5%	18	99.9 %	19	92.4 %	20	92.5 %
21	99.4 %	22	99.9 %				

the correct input Pinyin code, is

$$P(C_k|\gamma_i) = \frac{P(C_k)}{\sum_{l \in S_i} P(C_l)}, \tag{3}$$

where S_i denotes the set of candidate indices of the characters possessing the Pinyin code γ_i. Note that for a non-polyphonic character C_k, $P(C_k|\gamma_i)$ will be zero for all but a single γ_i.

Conversely, the probability that C_k is read as γ_i is

$$P(\gamma_i|C_k) = \frac{P(C_k|\gamma_i)P(\gamma_i)}{P(C_k)}. \tag{4}$$

The probability that C_k will be the input character, given an input Pinyin code ϕ_j, is

$$P(C_k|\phi_j) = \sum_i P(C_k|\gamma_i)P(\gamma_i|\phi_j), \tag{5}$$

where the summation is taken over those γ_i which denote alternative pronunciations of C_k.

Conversely, the probability that ϕ_j will be the input Pinyin code for C_k, is

$$P(\phi_j|C_k) = \sum_i P(\phi_j|\gamma_i)P(\gamma_i|C_k), \tag{6}$$

where $P(\gamma_i|C_k)$ is computed by (4), and the summation is taken over those γ_i which denote alternative pronunciations of C_k.

The probabilistic relation among the possible user-input PYC, the standard code, and the character is shown in Fig.1. (The numbers in the Fig 1. indicate the formulae for computing the corresponding probabilities.)

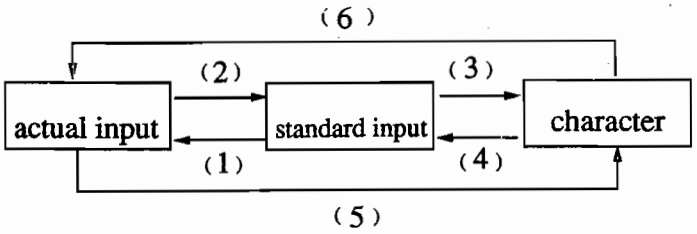

Figure 1: The relations among ϕ_j, γ_i and C_k

Formula (5) already provides an elementary principle for a character level error-correcting system. Given a Pinyin code by the user, the probable alternative characters can be listed in decreasing order of probabilistics computed in (5). In the next section we propose the basic schemes for realizing error-correcting function in advanced Pinyin-Hanzi conversion systems.

4 Schemes for an Error-Correcting System

The basic principle of an error-correcting Pinyin-Hanzi conversion system is that for a given string of Pinyin alphabet input by the user, the system computes a list of alternative character sequences according to how likely they are the the character to be input.

Suppose A is the input. Let $\phi^t_{j_1} \dots \phi^t_{j_{r_t}}$, $t = 1, \dots, T$, be the possible segmentations into Pinyin codes, and let $P(\phi^t_{j_1} \dots \phi^t_{j_{r_t}})$ denote the probability for $\phi^t_{j_1} \dots \phi^t_{j_{r_t}}$ to be the correct segmentation, which could be set to be $\frac{1}{T}$.

The probability for $C_{k_1} \dots C_{k_r}$ to be the solution is

$$P(C_{k_1} \dots C_{k_r}|A) = \sum_{r_t=r;1\le t\le T} P(C_{k_1} \dots C_{k_r}|\phi^t_{j_1} \dots \phi^t_{j_{r_t}}) P(\phi^t_{j_1} \dots \phi^t_{j_{r_t}}|A), \qquad (7)$$

where

$$P(C_{k_1} \dots C_{k_r}|\phi^t_{j_1} \dots \phi^t_{j_{r_t}}) = \frac{P(\phi^t_{j_1} \dots \phi^t_{j_{r_t}}|C_{k_1} \dots C_{k_r}) P(C_{k_1} \dots C_{k_r})}{\sum_{k_1 \dots k_r} P(\phi^t_{j_1} \dots \phi^t_{j_{r_t}}|C_{k_1} \dots C_{k_r}) P(C_{k_1} \dots C_{k_r})}. \qquad (8)$$

Practically we have

$$P(\phi^t_{j_1} \dots \phi^t_{j_{r_t}}|C_{k_1} \dots C_{k_r}) = P(\phi^t_{j_1}|C_{k_1}) \dots P(\phi^t_{j_{r_t}}|C_{k_r}), \qquad (9)$$

$P(C_{k_1} \dots C_{k_r})$ can be computed by various statistical models of Chinese language. In the following we briefly survey three possible schemes for implementing an error correcting Pinyin-Hanzi converting system.

- **Pinyin-Word Conversion**

  In a word-level Pinyin-Hanzi conversion system, the input is assumed to be a word of the form $C_{k_1} \ldots C_{k_r}$ (most words have length r=2). The probability $P(C_{k_1} \ldots C_{k_r})$ is the relative frequency of the word $C_{k_1} \ldots C_{k_r}$ in Chinese language, which is available[5] and is actually utilized in existing Pinyin-Hanzi conversion systems.

- **Pinyin-CharacterString Conversion**

  This type of conversion system does not assume $C_{k_1} \ldots C_{k_r}$ to be a word. The probability $P(C_{k_1} \ldots C_{k_r})$ is computed by the following formula based on a Markov model of Chinese text:

$$P(C_{k_1} \ldots C_{k_r}) = \prod_{s=1}^{r} P(C_{k_s}|C_{k_1} \ldots C_{k_{s-1}}) \tag{10}$$

  In practice, $P(C_{k_s}|C_{k_1} \ldots C_{k_{s-1}})$ is assumed to be identical to $P(C_{k_s}|C_{k_{s-1}})$ or $P(C_{k_s}|C_{k_{s-2}}C_{k_{s-1}})$, corresponding to the use of *bi-gram* or *tri-gram* statistics of Chinese language. Such a model has already been actually utilized in Chinese OCR system [3]. We believe its adaptation for an error correcting Pinyin-CharacterString conversion system is also realistic.

- **Pinyin-Sentence Conversion**

  An error-correcting Pinyin-sentence conversion system is much like a continuous speech recognizer. While reliably converting accoustic signals into standard Pinyin codes is yet a great burden for speech recognizer, there is no much difficulty to convert imperfect user-input Pinyin codes reliably into standard Pinyin codes, as is suggested by Table 4 of Section 2.

  In principle, an error-correcting Pinyin-sentence conversion system can be constructed by using the set $\Phi = \{\phi_1, \ldots, \phi_M\}$ of extended Pinyin codes instead of the standard code set $\Gamma = \{\gamma_1, \ldots, \gamma_N\}$, and using $P(C_k|\phi_j)$'s instead of $P(C_k|\gamma_i)$'s, in a conventional Pinyin-sentence conversion system (cf. [1], [2]).

5 Discussion

We have proposed the basic schemes of an error-correcting Pinyin-Hanzi conversion system based on the statistics of spelling errors of ordinary Chinese users. Such a system may enjoy higher accuracy if the parameters are adapted to the individual user. It is also evident that a user is likely to make less and less errors in the course of using a Pinyin-Hanzi system. The design of a learning mechanism to follow the user to optimize the performance of the system will also be an interesting topic.

Acknowledgements:

We would like to thank these people who participated in the Pinyin code test. Dr.Zhou Ming of TsingHua University, Dr.Gou Jin of NUS are greatly appreciated for providing with some useful materials.

The work was funded by the Sasakawa Scientific Research Grant from the Japan Science Society.

References

[1] Zheng Rong et al.(1995), *A Small Pinyin-Chinese Translation System Using Statistics and Analysis Methods*, in *Advances and Applications on Computational Linguistics* (Chen and Yuan Eds.), Tsinghua Univ. Press, pp.346-351.

[2] Cheng Hua et al.(1995), *The Design and Implementation of A Pinyin-Chinese Word Conversion System*, in *Advances and Applications on Computational Linguistics* (Chen and Yuan Eds.), Tsinghua Univ. Press, pp.340-345.

[3] Xia Ying et al.(1995), *Statistical Method and Application of Co-occurrence Probabilities Between Chinese Characters*, in *Advances and Applications on Computational Linguistics* (Chen and Yuan Eds.), Tsinghua Univ. Press, pp.106-111.

[4] Guo Jin et al.(1992), *PH: a Chinese Corpus for Pinyin-Hanzi Transcription*, in *Technical Report TR93-112-0, Institute of Systems Science, National University of Singapore*

[5] Guo Jin (1993), *Statistical Language Modeling and Some Experimental Results on Chinese Syllables to Words Transcription*, in *Journal of Chinese Information Processing Vol. 7, No. 1, pp. 18-27.*