Why Press Backspace? Understanding User Input Behaviors in Chinese Pinyin Input Method

Yabin Zheng[1], Lixing Xie[1], Zhiyuan Liu[1], Maosong Sun[1], Yang Zhang[2], Liyun Ru[1,2]
[1]State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, China
[2]Sogou Inc., Beijing 100084, China
{yabin.zheng,lavender087,lzy.thu,sunmaosong}@gmail.com
{zhangyang,ruliyun}@sogou-inc.com

Abstract

Chinese Pinyin input method is very important for Chinese language information processing. Users may make errors when they are typing in Chinese words. In this paper, we are concerned with the reasons that cause the errors. Inspired by the observation that pressing backspace is one of the most common user behaviors to modify the errors, we collect $54,309,334$ error-correction pairs from a real-world data set that contains $2,277,786$ users via backspace operations. In addition, we present a comparative analysis of the data to achieve a better understanding of users' input behaviors. Comparisons with English typos suggest that some language-specific properties result in a part of Chinese input errors.

1 Introduction

Unlike western languages, Chinese is unique due to its logographic writing system. Chinese users cannot directly type in Chinese words using a QWERTY keyboard. Pinyin is the official system to transcribe Chinese characters into the Latin alphabet. Based on this transcription system, Pinyin input methods have been proposed to assist users to type in Chinese words (Chen, 1997).

The typical way to type in Chinese words is in a sequential manner (Wang et al., 2001). Assume users want to type in the Chinese word "什么(what)". First, they mentally generate and type in corresponding Pinyin "shenme". Then, a Chinese Pinyin input method displays a list of Chinese words which share that Pinyin, as shown in Fig. 1. Users visually search the target word from candidates and select numeric key "1" to get the result. The last two steps do not exist in typing process of English words, which indicates that it is more complicated for Chinese users to type in Chinese words.

Figure 1: Typical Chinese Pinyin input method for a correct Pinyin (Sogou-Pinyin).

Figure 2: Typical Chinese Pinyin input method for a mistyped Pinyin (Sogou-Pinyin).

Chinese users may make errors when they are typing in Chinese words. As shown in Fig. 2, a user may mistype "shenme" as "*shenem*". Typical Chinese Pinyin input method can not return the right word. Users may not realize that an error occurs and select the first candidate word "什恶魔" (a meaningless word) as the result. This greatly limits user experience since users have to identify errors and modify them, or cannot get the right word.

In this paper, we analyze the reasons that cause errors in Chinese Pinyin input method. This analysis is helpful in enhancing the user experience and the performance of Chinese Pinyin input method. In practice, users press *backspace* on the keyboard to modify the errors, they delete the mistyped word and re-type in the correct word. Motivated by this ob-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 485–490,
Portland, Oregon, June 19-24, 2011.

servation, we can extract error-correction pairs from backspace operations. These error-correction pairs are of great importance in Chinese spelling correction task which generally relies on sets of confusing words.

We extract $54,309,334$ error-correction pairs from user input behaviors and further study them. Our comparative analysis of Chinese and English typos suggests that some language-specific properties of Chinese lead to a part of input errors. To the best of our knowledge, this paper is the first one which analyzes user input behaviors in Chinese Pinyin input method.

The rest of this paper is organized as follows. Section 2 discusses related works. Section 3 introduces how we collect errors in Chinese Pinyin input method. In Section 4, we investigate the reasons that result in these errors. Section 5 concludes the whole paper and discusses future work.

2 Previous Work

For English spelling correction (Kukich, 1992; Ahmad and Kondrak, 2005; Chen et al., 2007; Whitelaw et al., 2009; Gao et al., 2010), most approaches make use of a lexicon which contains a list of well-spelled words (Hirst and Budanitsky, 2005; Islam and Inkpen, 2009). Context features (Rozovskaya and Roth, 2010) of words provide useful evidences for spelling correction. These features are usually represented by an *n*-gram language model (Cucerzan and Brill, 2004; Wilcox-O'Hearn et al., 2010). Phonetic features (Toutanova and Moore, 2002; Atkinson, 2008) are proved to be useful in English spelling correction. A spelling correction system is trained using these features by a noisy channel model (Kernighan et al., 1990; Ristad et al., 1998; Brill and Moore, 2000).

Chang (1994) first proposes a representative approach for Chinese spelling correction, which relies on sets of confusing characters. Zhang et al. (2000) propose an approximate word-matching algorithm for Chinese to solve Chinese spell detection and correction task. Zhang et al. (1999) present a winnow-based approach for Chinese spelling correction which takes both local language features and wide-scope semantic features into account. Lin and Yu (2004) use Chinese frequent strings and report an accuracy of 87.32%. Liu et al. (2009) show that about 80% of the errors are related to pronunciations. Visual and phonological features are used in Chinese spelling correction (Liu et al., 2010).

Instead of proposing a method for spelling correction, we mainly investigate the reasons that cause typing errors in both English and Chinese. Some errors are caused by specific properties in Chinese such as the phonetic difference between Mandarin and dialects spoken in southern China. Meanwhile, confusion sets of Chinese words play an important role in Chinese spelling correction. We extract a large scale of error-correction pairs from real user input behaviors. These pairs contain important evidence about confusing Pinyins and Chinese words which are helpful in Chinese spelling correction.

3 User Input Behaviors Analysis

We analyze user input behaviors from anonymous user typing records in a Chinese input method. Data set used in this paper is extracted from Sogou Chinese Pinyin input method[1]. It contains $2,277,786$ users' typing records in 15 days. The numbers of Chinese words and characters are $3,042,637,537$ and $5,083,231,392$, respectively. We show some user typing records in Fig. 3.

```
[20100718 11:10:38.790ms] select:2 zhe 这 WINWORD.exe
[20100718 11:10:39.770ms] select:1 shi 是 WINWORD.exe
[20100718 11:10:40.950ms] select:1 shenem 什恶魔 WINWORD.exe
[20100718 11:10:42.300ms] Backspace WINWORD.exe
[20100718 11:10:42.520ms] Backspace WINWORD.exe
[20100718 11:10:42.800ms] Backspace WINWORD.exe
[20100718 11:10:45.090ms] select:1 shenme 什么 WINWORD.exe
```

Figure 3: Backspace in user typing records.

From Fig. 3, we can see the typing process of a Chinese sentence "这 是 什么" (What is this). Each line represents an input segment or a backspace operation. For example, word "什么" (what) is typed in using Pinyin "shenme" with numeric selection "1" at 11:10am in Microsoft Word application.

The user made a mistake to type in the third Pinyin ("shenme" is mistyped as "*shenem*"). Then, he/she pressed the backspace to modify the errors he has made. the word "什恶魔" is deleted and replaced with the correct word "什么" using Pinyin

[1]Sogou Chinese Pinyin input method, can be accessed from http://pinyin.sogou.com/

"shenme". As a result, we compare the typed-in Pinyins before and after backspace operations. We can find the Pinyin-correction pairs "shenem-shenme", since their edit distance is less than a threshold. Threshold is set to 2 in this paper, as Damerau (1964) shows that about 80% of typos are caused by a single edit operation. Therefore, using a threshold of 2, we should be able to find most of the typos. Furthermore, we can extract corresponding Chinese word-correction pairs "什恶魔-什么" from this typing record.

Using heuristic rules discussed above, we extract $54,309,334$ Pinyin-correction and Chinese word-correction pairs. We list some examples of extracted Pinyin-correction and Chinese word-correction pairs in Table 1. Most of the mistyped Chinese words are meaningless.

Pinyin-correction	Chinese word-correction
shenem-shenme	什恶魔-什么(what)
dianao-diannao	点奥-电脑(computer)
xieixe-xiexie	系诶下额-谢谢(thanks)
laing-liang	来那个-两(two)
ganam-ganma	甘阿明-干吗(what's up)
zhdiao-zhidao	摘掉-知道(know)
lainxi-lianxi	来年息-联系(contact)
zneme-zenme	则呢么-怎么(how)
dainhua-dianhua	戴年华-电话(phone)
huiali-huilai	灰暗里-回来(return)

Table 1: Typical Pinyin-correction and Chinese word-correction pairs.

We want to evaluate the precision and recall of our extraction method. For precision aspect, we randomly select $1,000$ pairs and ask five native speakers to annotate them as correct or wrong. Annotation results show that the precision of our method is about 75.8%. Some correct Pinyins are labeled as errors because we only take edit distance into consideration. We should consider context features as well, which will be left as our future work.

We choose 15 typical mistyped Pinyins to evaluate the recall of our method. The total occurrences of these mistyped Pinyins are $259,051$. We successfully retrieve $144,020$ of them, which indicates the recall of our method is about 55.6%. Some errors are not found because sometimes users do not modify the errors, especially when they are using Chinese input method under instant messenger softwares.

4 Comparisons of Pinyin typos and English Typos

In this section, we compare the Pinyin typos and English typos. As shown in (Cooper, 1983), typing errors can be classified into four categories: deletions, insertions, substitutions, and transpositions. We aim at studying the reasons that result in these four kinds of typing errors in Chinese Pinyin and English, respectively.

For English typos, we generate mistyped word-correction pairs from Wikipedia[2] and SpellGood.[3], which contain $4,206$ and $10,084$ common misspellings in English, respectively. As shown in Table 2, we reach the first conclusion: **about half of the typing errors in Pinyin and English are caused by deletions**, which indicates that users are more possible to omit some letters than other three edit operations.

	Deletions	Insertions	Substitutions	Transpositions
Pinyin	47.06%	28.17%	19.04%	7.46%
English	43.38%	18.89%	17.32%	18.70%

Table 2: Different errors in Pinyin and English.

Table 3 and Table 4 list Top 5 letters that produce deletion errors (users forget to type in some letters) and insertion errors (users type in extra letters) in Pinyin and English.

Pinyin	Examples	English	Examples
i	xianza-xianzai	e	achive-achieve
g	**yingai-yinggai**	i	abilties-abilities
e	shenm-shenme	c	acomplish-accomplish
u	pengyo-pengyou	a	agin-again
h	**senme-shenme**	t	admited-admitted

Table 3: Deletion errors in Pinyin and English.

Pinyin	Examples	English	Examples
g	**yingwei-yinwei**	e	analogeous-analogous
i	tiebie-tebie	r	arround-around
a	xiahuan-xihuan	s	asside-aside
o	huijiao-huijia	i	aisian-asian
h	**shuibian-suibian**	n	abandonned-abandoned

Table 4: Insertion errors in Pinyin and English.

[2] http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines

[3] http://www.spellgood.net/

We can see from Table 3 and Table 4 that: (1) vowels (a, o, e, i, u) are deleted or inserted more frequently than consonants in Pinyin. (2) some specific properties in Chinese lead to insertion and deletion errors. Many users in southern China cannot distinguish the front and the back nasal sound ('ang' - 'an', 'ing' - 'in', 'eng' - 'en') as well as the retroflex and the blade-alveolar ('zh' - 'z', 'sh' - 's', 'ch' - 'c'). They are confused about whether they should add letter 'g' or 'h' under these situations. (3) the same letters can occur continuously in English, such as "acomplish-a**cc**omplish" and "admited-admi**tt**ed" in our examples. English users sometimes make insertion or deletion errors in these cases. We also observe this kind of errors in Chinese Pinyin, such as "yingai-yin**gg**ai", "liange-lian**gg**e" and "dianao-dia**nn**ao".

For transposition errors, Table 5 lists Top 10 patterns that produce transposition errors in Pinyin and English. Our running example "shenem-shenme" belongs to this kind of errors. We classify the letters of the keyboard into two categories, i.e. "left" and "right", according to their positions on the keyboard. Letter 'e' is controlled by left hand while 'm' is controlled by right hand. Users mistype "shenme" as "shenem" because they mistake the typing order of 'm' and 'e'.

Fig. 4 is a graphic representation, in which we add a link between 'm' and 'e'. The rest patterns in Table 5 can be done in the same manner. Interestingly, from Fig. 4, we reach the second conclusion: **most of the transposition errors are caused by mistaking the typing orders across left and right hands**. For instance, users intend to type in a letter ('m') controlled by right hand. But they type in a letter ('e') controlled by left hand instead.

Pinyin	Examples	English	Examples
ai	xaing-xiang	ei	acheive-achieve
na	xinag-xiang	ra	clera-clear
em	shenem-shenme	re	vrey-very
ia	xianzia-xianzai	na	wnat-want
ne	zneme-zenme	ie	hieght-height
oa	zhidoa-zhidao	er	befoer-before
ei	jiejei-jiejie	it	esitmated-estimated
hs	haihsi-haishi	ne	scinece-science
ah	sahng-shang	el	littel-little
ou	rugou-ruguo	si	epsiode-episode

Table 5: Transpositions errors in Pinyin and English.

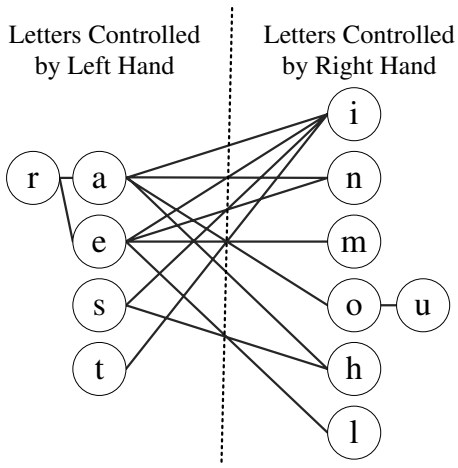

Figure 4: Transpositions errors on the keyboard.

For substitution errors, we study the reason why users mistype one letter for another. In the Pinyin-correction pairs, users always mistype 'a' as 'e' and vice versa. The reason is that they have similar pronunciations in Chinese. As a result, we add two directed edges 'a' and 'e' in Fig. 5. Some letters are mistyped for each other because they are adjacent on the keyboard although they do not share similar pronunciations, such as 'g' and 'f'.

We summarize the substitution errors in English in Fig. 6. Letters 'q', 'k' and 'c' are often mixed up with each other because they sound alike in English although they are apart on the keyboard. However, the three letters are not connected in Fig. 5, which indicates that users can easily distinguish them in Pinyin.

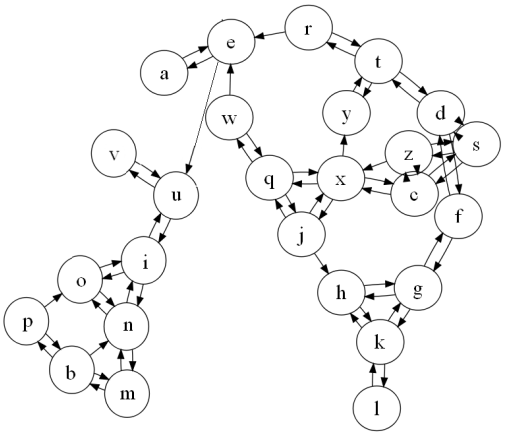

Figure 5: Substitutions errors in Pinyin.

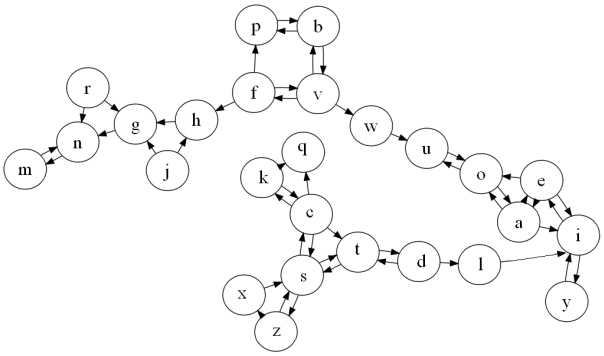

Figure 6: Substitutions errors in English.

Mistyped letter pairs	Similar pronunciations in Chinese	Similar pronunciations in English	Adjacent on keyboard
(m,n)	✓	✓	✓
(b,p);(d,t)	✓	✓	×
(z,c,s);(g,k,h)	✓	×	✓
(j,q,x);(u,v)	✓	×	×
(i,y)	×	✓	✓
(q,k,c)	×	✓	×
(j,h);(z,x)	×	×	✓

Table 6: Pronunciation properties and keyboard distance in Chinese Pinyin and English

We list some examples in Table 6. For example, letters 'm' and 'n' have similar pronunciations in both Chinese and English. Moreover, they are adjacent on the keyboard, which leads to interferences or confusion in both Chinese and English. Letters 'j', 'q' and 'x' are far from each other on the keyboard. But they sound alike in Chinese, which makes them connected in Fig. 5. In Fig. 6, letters 'b' and 'p' are connected to each other because they have similar pronunciations in English, although they are not adjacent on the keyboard.

Finally, we summarize the third conclusion: **substitution errors are caused by language specific similarities (similar pronunciations) or keyboard neighborhood (adjacent on the keyboard)**.

All in all, we generally classify typing errors in English and Chinese into four categories and investigate the reasons that result in these errors respectively. Some language specific properties, such as pronunciations in English and Chinese, lead to substitution, insertion and deletion errors. Keyboard layouts play an important role in transposition errors, which are language-independent.

5 Conclusions and Future Works

In this paper, we study user input behaviors in Chinese Pinyin input method from backspace operations. We aim at analyzing the reasons that cause these errors. Users signal that they are very likely to make errors if they press backspace on the keyboard. Then they modify the errors and type in the correct words they want. Different from the previous research, we extract abundant Pinyin-correction and Chinese word-correction pairs from backspace operations. Compared with English typos, we observe some language-specific properties in Chinese have impact on errors. All in all, user behaviors (Zheng et al., 2009; Zheng et al., 2010; Zheng et al., 2011b) in Chinese Pinyin input method provide novel perspectives for natural language processing tasks.

Below we sketch three possible directions for the future work: (1) we should consider position features in analyzing Pinyin errors. For example, it is less likely that users make errors in the first letter of an input Pinyin. (2) we aim at designing a self-adaptive input method that provide error-tolerant features (Chen and Lee, 2000; Zheng et al., 2011a). (3) we want to build a Chinese spelling correction system based on extracted error-correction pairs.

Acknowledgments

This work is supported by a Tsinghua-Sogou joint research project and the National Natural Science Foundation of China under Grant No. 60873174.

References

F. Ahmad and G. Kondrak. 2005. Learning a spelling error model from search query logs. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 955–962.

K. Atkinson. 2008. Gnu aspell 0.60.6. *http://aspell.sourceforge.net*.

E. Brill and R.C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293.

C.H. Chang. 1994. A pilot study on automatic Chinese spelling error correction. *Communication of COLIPS*, 4(2):143–149.

Z. Chen and K.F. Lee. 2000. A new statistical approach to Chinese Pinyin input. In *Proceedings of the*

38th Annual Meeting on Association for Computational Linguistics, pages 241–247.

Q. Chen, M. Li, and M. Zhou. 2007. Improving query spelling correction using web search results. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 181–189.

Y. Chen. 1997. *Chinese Language Processing*. Shanghai Education publishing company.

W.E. Cooper. 1983. *Cognitive aspects of skilled typewriting*. Springer-Verlag.

S. Cucerzan and E. Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 293–300.

F.J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

J. Gao, X. Li, D. Micol, C. Quirk, and X. Sun. 2010. A large scale ranker-based system for search query spelling correction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 358–366.

G. Hirst and A. Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(01):87–111.

A. Islam and D. Inkpen. 2009. Real-word spelling correction using Google Web 1T 3-grams. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1241–1249.

M.D. Kernighan, K.W. Church, and W.A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics*, pages 205–210.

K. Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

Y.J. Lin and M.S. Yu. 2004. The properties and further applications of Chinese frequent strings. *Computational Linguistics and Chinese Language Processing*, 9(1):113–128.

C.L. Liu, K.W. Tien, M.H. Lai, Y.H. Chuang, and S.H. Wu. 2009. Capturing errors in written Chinese words. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 25–28.

C.L. Liu, M.H. Lai, Y.H. Chuang, and C.Y. Lee. 2010. Visually and phonologically similar characters in incorrect simplified chinese words. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 739–747.

E.S. Ristad, P.N. Yianilos, M.T. Inc, and NJ Princeton. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532.

A. Rozovskaya and D. Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 961–970.

K. Toutanova and R.C. Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 144–151.

J. Wang, S. Zhai, and H. Su. 2001. Chinese input with keyboard and eye-tracking: an anatomical study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 349–356.

C. Whitelaw, B. Hutchinson, G.Y. Chung, and G. Ellis. 2009. Using the web for language independent spellchecking and autocorrection. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 890–899.

A. Wilcox-O'Hearn, G. Hirst, and A. Budanitsky. 2010. Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. *Computational Linguistics and Intelligent Text Processing*, pages 605–616.

L. Zhang, M. Zhou, C. Huang, and HH Pan. 1999. Multifeature-based approach to automatic error detection and correction of Chinese text. In *Proceedings of the First Workshop on Natural Language Processing and Neural Networks*.

L. Zhang, C. Huang, M. Zhou, and H. Pan. 2000. Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 248–254.

Y. Zheng, Z. Liu, M. Sun, L. Ru, and Y. Zhang. 2009. Incorporating user behaviors in new word detection. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 2101–2106.

Y. Zheng, Z. Liu, and L. Xie. 2010. Growing related words from seed via user behaviors: a re-ranking based approach. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 49–54.

Y. Zheng, C. Li, and M. Sun. 2011a. CHIME: An efficient error-tolerant chinese pinyin input method. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (accepted)*.

Y. Zheng, Z. Liu, L. Xie, M. Sun, L. Ru, and Y. Zhang. 2011b. User Behaviors in Related Word Retrieval and New Word Detection: A Collaborative Perspective. *ACM Transactions on Asian Language Information Processing, Special Issue on Chinese Language Processing (accepted)*.