Using Structural Information for Identifying Similar Chinese Characters

Chao-Lin Liu **Jen-Hsiang Lin**
Department of Computer Science, National Chengchi University, Taipei 11605, Taiwan
{chaolin, g9429}@cs.nccu.edu.tw

Abstract

Chinese characters that are similar in their pronunciations or in their internal structures are useful for computer-assisted language learning and for psycholinguistic studies. Although it is possible for us to employ image-based methods to identify visually similar characters, the resulting computational costs can be very high. We propose methods for identifying visually similar Chinese characters by adopting and extending the basic concepts of a proven Chinese input method--Cangjie. We present the methods, illustrate how they work, and discuss their weakness in this paper.

1 Introduction

A Chinese sentence consists of a sequence of characters that are not separated by spaces. The function of a Chinese character is not exactly the same as the function of an English word. Normally, two or more Chinese characters form a Chinese word to carry a meaning, although there are Chinese words that contain only one Chinese character. For instance, a translation for "conference" is "研討會" and a translation for "go" is "去". Here "研討會" is a word formed by three characters, and "去" is a word with only one character.

Just like that there are English words that are spelled similarly, there are Chinese characters that are pronounced or written alike. For instance, in English, the sentence "John plays an important roll in this event." contains an incorrect word. We should replace "roll" with "role". In Chinese, the sentence "今天上午我們來**試場**買菜" contains an incorrect word. We should replace "試場" (a place for taking examinations) with "市場" (a market). These two words have the same pronunciation, shi(4) chang(3)[†], and both represent locations. The sentence "經理要我**構買**一部計算機" also contains an error, and we need to replace "構買" with "購買". "構買" is considered an incorrect word, but can be confused with "購買" because the first characters in these words look similar.

Characters that are similar in their appearances or in their pronunciations are useful for computer-assisted language learning (cf. Burstein & Leacock, 2005). When preparing test items for testing students' knowledge about correct words in a computer-assisted environment, a teacher provides a sentence which contains the character that will be replaced by an incorrect character. The teacher needs to specify the answer character, and the software will provide two types of incorrect characters which the teachers will use as distracters in the test items. The first type includes characters that look similar to the answer character, and the second includes characters that have the same or similar pronunciations with the answer character.

Similar characters are also useful for studies in Psycholinguistics. Yeh and Li (2002) studied how similar characters influenced the judgments made by skilled readers of Chinese. Taft, Zhu, and Peng (1999) investigated the effects of positions of radicals on subjects' lexical decisions and naming responses. Computer programs that can automatically provide similar characters are thus potentially helpful for designing related experiments.

2 Identifying Similar Characters with Information about the Internal Structures

We present some similar Chinese characters in the first subsection, illustrate how we encode Chinese characters in the second subsection, elaborate how we improve the current encoding method to facilitate the identification of similar characters in the third subsection, and discuss the weakness of our current approach in the last subsection.

2.1 Examples of Similar Chinese Characters

We show three categories of confusing Chinese characters in Figures 1, 2, and 3. Groups of similar

[†] We use Arabic digits to denote the four tones in Mandarin.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 93–96,
Columbus, Ohio, USA, June 2008.

士土工干千 戌戍成 田由甲申
母毋 勿匆 人入 未末 采采 凹凸

Figure 1. Some similar Chinese characters

頸勁 搆溝 陪倍 硯現 裸棵 搞篙
列刑 盆盎盂盅 因困囚 間閒閃開

Figure 2. Some similar Chinese characters that have different pronunciations

形刑型 踵種腫 購構搆 紀記計
園圓員 脛逕徑痙勁

Figure 3. Homophones with a shared component

characters are separated by spaces in these figures. In Figure 1, characters in each group differ at the stroke level. Similar characters in every group in the first row in Figure 2 share a common part, but the shared part is not the radical of these characters. Similar characters in every group in the second row in Figure 2 share a common part, which is the radical of these characters. Similar characters in every group in Figure 2 have different pronunciations. We show six groups of homophones that also share a component in Figure 3. Characters that are similar in both pronunciations and internal structures are most confusing to new learners.

It is not difficult to list all of those characters that have the same or similar pronunciations, e.g., "試場" and "市場", if we have a machine readable lexicon that provides information about pronunciations of characters and when we ignore special patterns for tone sandhi in Chinese (Chen, 2000).

In contrast, it is relatively difficult to find characters that are written in similar ways, e.g., "構" with "購", in an efficient way. It is intriguing to resort to image processing methods to find such structurally similar words, but the computational costs can be very high, considering that there can be tens of thousands of Chinese characters. There are more than 22000 different characters in large corpus of Chinese documents (Juang et al., 2005), so directly computing the similarity between images of these characters demands a lot of computation. There can be more than 4.9 billion combinations of character pairs. The Ministry of Education in Taiwan suggests that about 5000 characters are needed for ordinary usage. In this case, there are about 25 million pairs.

The quantity of combinations is just one of the bottlenecks. We may have to shift the positions of the characters "appropriately" to find the common part of a character pair. The appropriateness for shifting characters is not easy to define, making the image-based method less directly useful; for instance, the common part of the characters in the right group in the second row in Figure 3 appears in different places in the characters.

Lexicographers employ radicals of Chinese characters to organize Chinese characters into sections in dictionaries. Hence, the information should be useful. The groups in the second row in Figure 2 show some examples. The shared components in these groups are radicals of the characters, so we can find the characters of the same group in the same section in a Chinese dictionary. However, information about radicals as they are defined by the lexicographers is not sufficient. The groups of characters shown in the first row in Figure 2 have shared components. Nevertheless, the shared components are not considered as radicals, so the characters, e.g., "頸"and "勁", are listed in different sections in the dictionary.

2.2 Encoding the Chinese Characters

The Cangjie[‡] method is one of the most popular methods for people to enter Chinese into computers. The designer of the Cangjie method, Mr. Bong-Foo Chu, selected a set of 24 basic elements in Chinese characters, and proposed a set of rules to decompose Chinese characters into elements that belong to this set of building blocks (Chu, 2008). Hence, it is possible to define the similarity between two Chinese characters based on the similarity between their Cangjie codes.

Table 1, not counting the first row, has three

	Cangjie Codes		Cangjie Codes
士	十一	土	土
工	一中一	干	一十
勿	心竹竹	匆	竹田心
未	十木	末	木十
頸	一一一月金	勁	一一大尸
硯	一口月山山	現	一土月山山
搞	手卜口月	篙	竹卜口月
列	一弓中弓	刑	一廿中弓
因	田大	困	田木
間	日弓日	閒	日弓月
踵	口一竹十土	種	竹木竹十土
腫	月竹十土	紀	女火尸山
購	月金廿廿月	構	木廿廿月
記	卜口尸山	計	卜口十
圓	田口月金	員	口月山金
脛	月一女一	逕	卜一女一
徑	竹人一女一	痙	大一女一

Table 1. Cangjie codes for some characters

[‡] http://en.wikipedia.org/wiki/Cangjie_method

sections, each showing the Cangjie codes for some characters in Figures 1, 2, and 3. Every Chinese character is decomposed into an ordered sequence of *elements*. (We will find that a subsequence of these elements comes from a major *component* of a character, shortly.) Evidently, computing the number of shared elements provides a viable way to determine "visually similar" characters for characters that appeared in Figure 2 and Figure 3. For instance, we can tell that "搞" and "篙" are similar because their Cangjie codes share "卜ロ月", which in fact represent "高".

Unfortunately, the Cangjie codes do not appear to be as helpful for identifying the similarities between characters that differ subtly at the stroke level, e.g., "士土工干" and other characters listed in Figure 1. There are special rules for decomposing these relatively basic characters in the Cangjie method, and these special encodings make the resulting codes less useful for our tasks.

The Cangjie codes for characters that contain multiple components were intentionally simplified to allow users to input Chinese characters more efficiently. The longest Cangjie code for any Chinese character contains no more than five elements. In the Cangjie codes for "脛" and "徑", we see "一女一" for the component "巠", but this component is represented only by "一一" in the Cangjie codes for "頸" and "勁". The simplification makes it relatively harder to identify visually similar characters by comparing the actual Cangjie codes.

2.3 Engineering the Original Cangjie Codes

Although useful for the sake of designing input method, the simplification of Cangjie codes causes difficulties when we use the codes to find similar characters. Hence, we choose to use the complete codes for the components in our database. For instance, in our database, the codes for "巠", "脛", "徑", "頸", and "勁" are, respectively, "一女女一", "月一女女一", "竹人一女女一", "一女女一一月山金", and "一女女一大尸".

The knowledge about the graphical structures of the Chinese characters (cf. Juang et al., 2005; Lee, 2008) can be instrumental as well. Consider the examples in Figure 2. Some characters can be decomposed vertically; e.g., "盅" can be split into two smaller components, i.e., "中" and "皿". Some characters can be decomposed horizontally; e.g., "現" is consisted of "王" and "見". Some have enclosing components; e.g., "人" is enclosed in "口" in "囚". Hence, we can consider the locations of the components as well as the number of shared components in determining the similarity between characters.

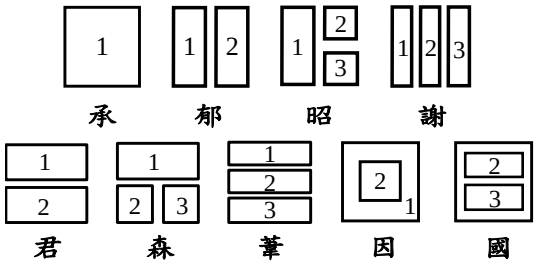

Figure 4. Arrangements of components in Chinese

Figure 4 illustrates possible layouts of the components in Chinese characters that were adopted by the Cangjie method (cf. Lee, 2008). A sample character is placed below each of these layouts. A box in a layout indicates a component in a character, and there can be at most three components in a character. We use digits to indicate the ordering the components. Notice that, in the second row, there are two boxes in the second to the rightmost layout. A larger box contains a smaller one. There are three boxes in the rightmost layout, and two smaller boxes are inside the outer box. Due to space limits, we do not show "1" for this outer box.

After recovering the simplified Cangjie code for a character, we can associate the character with a tag that indicates the overall layout of its components, and separate the code sequence of the character according to the layout of its components. Hence, the information about a character includes the tag for its layout and between one to three sequences of code elements. Table 2 shows the anno-

	Layout	Part 1	Part 2	Part 3
承	1	弓弓手人		
郁	2	大月	弓中	
昭	3	日	尸竹	口
謝	4	卜一一口	竹難竹	木戈
君	5	尸大	口	
森	6	木	木	木
葦	7	廿	木一	手
因	8	田	大	
國	9	田	戈	口一
頸	2	一女女一	一月山金	
徑	2	竹人	一女女一	
員	5	口	月山金	
圓	9	田	口	月山金
相	2	木	月山	
想	5	木月山	心	
箱	6	竹	木	月山

Table 2. Annotated and expanded code

tated and expanded codes of the sample characters in Figure 4 and the codes for some characters that we will discuss. The layouts are numbered from left to right and from top to bottom in Figure 4. Elements that do not belong to the original Canjie codes of the characters are shown in smaller font.

Recovering the elements that were dropped out by the Cangjie method and organizing the subsequences of elements into parts facilitate the identification of similar characters. It is now easier to find that the character (頸) that is represented by "一女女一" and "一月山金" looks similar to the character (徑) that is represented by "竹人" and "一女女一" in our database than using their original Cangjie codes in Table 1. Checking the codes for "員" and "圓" in Table 1 and Table 2 will offer an additional support for our design decisions.

In the worst case, we have to compare nine pairs of code sequences for two characters that both have three components. Since we do not simplify codes for components and all components have no more than five elements, conducting the comparisons operations are simple.

2.4 Drawbacks of Using the Cangjie Codes

Using the Cangjie codes as the basis for comparing the similarity between characters introduces some potential problems.

It appears that the Cangjie codes for some characters, particular those simple ones, were not assigned without ambiguous principles. Relying on Cangjie codes to compute the similarity between such characters can be difficult. For instance, "分" uses the fifth layout, but "兌" uses the first layout in Figure 4. The first section in Table 1 shows the Cangjie codes for some character pairs that are difficult to compare.

Due to the design of the Cangjie codes, there can be at most one component at the left hand side and at most one component at the top in the layouts. The last three entries in Table 2 provide an example for these constraints. As a standalone character, "相" uses the second layout. Like the standalone "相", the "相" in "箱" was divided into two parts. However, in "想", "相" is treated as an individual component because it is on top of "想". Similar problems may occur elsewhere, e.g., "森焚" and "恩因". There are also some exceptional cases; e.g., "品" uses the sixth layout, but "闆" uses the fifth layout.

3 Concluding Remarks

We adopt the Cangjie alphabet to encode Chinese characters, but choose not to simplify the code sequences, and annotate the characters with the layout information of their components. The resulting method is not perfect, but allows us to find visually similar characters more efficient than employing the image-based methods.

Trying to find conceptually similar but contextually inappropriate characters should be a natural step after being able to find characters that have similar pronunciations and that are visually similar.

Acknowledgments

Work reported in this paper was supported in part by the plan NSC-95-2221-E-004-013-MY2 from the National Science Council and in part by the plan ATU-NCCU-96H061 from the Ministry of Education of Taiwan.

References

Jill Burstein and Claudia Leacock. editors. 2005. *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, ACL.

Matthew Y. Chen. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. (Cambridge. Studies in Linguistics 92.) Cambridge: Cambridge University Press.

Bong-Foo Chu. 2008. *Handbook of the Fifth Generation of the Cangjie Input Method*, web version, available at http://www.cbflabs.com/book/ocj5/ocj5/index.html. Last visited on 14 Mar. 2008.

Hsiang Lee. 2008. *Cangjie Input Methods in 30 Days*, http://input.foruto.com/cjdict/Search_1.php, Foruto Company, Hong Kong. Last visited on 14 Mar. 2008.

Derming Juang, Jenq-Haur Wang, Chen-Yu Lai, Ching-Chun Hsieh, Lee-Feng Chien, and Jan-Ming Ho. 2005. Resolving the unencoded character problem for Chinese digital libraries. *Proceedings of the Fifth ACM/IEEE Joint Conference on Digital Libraries*, 311–319.

Marcus Taft, Xiaoping Zhu, and Danling Peng. 1999. Positional specificity of radicals in Chinese character recognition, *Journal of Memory and Language*, **40**, 498–519.

Su-Ling Yeh and Jing-Ling Li. 2002. Role of structure and component in judgments of visual similarity of Chinese characters, *Journal of Experimental Psychology: Human Perception and Performance*, **28**(4), 933–947.