Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese

Chenhui Chu, Toshiaki Nakazawa, Sadao Kurohashi

Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
E-mail: {chu, nakazawa}@nlp.ist.i.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

Abstract

Chinese characters are used both in Japanese and Chinese, which are called Kanji and Hanzi respectively. Chinese characters contain significant semantic information, a mapping table between Kanji and Hanzi can be very useful for many Japanese-Chinese bilingual applications, such as machine translation and cross-lingual information retrieval. Because Kanji characters are originated from ancient China, most Kanji have corresponding Chinese characters in Hanzi. However, the relation between Kanji and Hanzi is quite complicated. In this paper, we propose a method of making a Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese automatically by means of freely available resources. We define seven categories for Kanji based on the relation between Kanji and Hanzi, and classify mappings of Chinese characters into these categories. We use a resource from Wiktionary to show the completeness of the mapping table we made. Statistical comparison shows that our proposed method makes a more complete mapping table than the current version of Wiktionary.

Keywords: Chinese character, mapping table, Japanese-Chinese

1. Introduction

Chinese characters are used both in Japanese and Chinese. In Japanese the Chinese characters are called Kanji, while in Chinese they are called Hanzi. Hanzi can be divided into two groups, Traditional Chinese (used in Taiwan, Hong Kong and Macau) and Simplified Chinese (used in mainland China and Singapore). The number of strokes needed to write characters has been largely reduced in Simplified Chinese, and the shapes may be different from the ones in Traditional Chinese. Table 1 gives some examples of Chinese Characters in Japanese, Traditional Chinese and Simplified Chinese, from which we can see that the relation between Kanji and Hanzi is quite complicated.

Because Kanji characters are originated from ancient China, most Kanji have fully corresponding Chinese characters in Hanzi. Actually, although Japanese continues to evolve and change after the adoption of Chinese characters, the visual forms of the Chinese characters retain certain level of similarity, and many Kanji are identical to Hanzi (e.g. "雪(snow)" in Table 1), some Kanji are identical to Traditional Chinese but different from Simplified Chinese (e.g. "愛(love)" in Table 1), some Kanji are identical to Simplified Chinese but different from Traditional Chinese (e.g. "国(country)" in Table 1). There are also some visual variations in Kanji which have corresponding Chinese characters in Hanzi while the shapes are different from Hanzi (e.g. "発(begin)" in Table 1). However, there are some Kanji that do not have fully corresponding Chinese characters in Hanzi. Some Kanji only have corresponding Traditional Chinese (e.g. "詫(apology)" in Table 1), and some Kanji only have corresponding Simplified Chinese. (e.g. "鮃(bastard halibut)" in Table 1). Moreover, there are some Chinese characters that are created in Japan namely Kokuji, which means national characters, may have no corresponding Chinese characters in Hanzi (e.g. "込(included)" in Table 1).

	C1	C2	C3	C4	C5	C6	C7
Kanji	雪	愛	国	発	詫	鮃	込
TC	雪	愛	國	發	詫	N/A	N/A
SC	雪	爱	国	发	N/A	鲆	N/A

Table 1: Examples of Chinese characters (C denotes Category which is described in Section 4., TC denotes Traditional Chinese and SC denotes Simplified Chinese).

What makes the relation more complicated is that, a single Kanji form may correspond to multiple Hanzi forms. Also, a single Simplified Chinese form may correspond to multiple Traditional Chinese forms, and vice versa.

In this paper, we focus on the relation between Kanji and Hanzi, and propose a method of making a Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese automatically by means of freely available resources. Because Chinese characters contain significant semantic information, this mapping table can be very useful for both linguists and NLP researchers doing tasks such as machine translation or cross-lingual information retrieval between Japanese and Chinese. For example, Tan et al. (1995) used the occurrence of common Chinese characters between Japanese and Chinese in automatic sentence alignment task, Goh et al. (2005) built a Japanese-Simplified Chinese dictionary partly using direct conversion of Japanese into Chinese for Japanese Kanji words, Huang et al. (2008) examined and analyzed the semantic relations between Japanese and Chinese on word level based on Chinese characters mapping, Chu et al. (2011) used common Chinese characters information in Japanese-Chinese phrase alignment, the mapping table can be helpful for these tasks.

2. Character Sets of Kanji and Hanzi

The character set in use for Kanji is JIS Kanji code. While for Hanzi, there are many character sets, among which we

TC	故	說	錢	沖,衝	干, 幹, 乾	...
SC	故	说	钱	冲	干	...

Table 2: Hanzi converter standard conversion table.

choose Big5 for Traditional Chinese and GB2312 for Simplified Chinese, both are in widespread use.

2.1. JIS Kanji Code

For JIS Kanji code, JIS X 0208 is a widely used character set specified as a Japanese Industrial Standard, containing 6,879 graphic characters, which includes 6,355 Kanji as well as 524 non-Kanji. The mapping table is for the 6,355 Kanji characters namely JIS Kanji in JIS X 0208.

2.2. Big5

Big5 is the most commonly used character set for Traditional Chinese in Taiwan, Hong Kong and Macau, which is defined by "Institute for Information Industry" in Taiwan. There are 13,060 Traditional Chinese characters in Big5.

2.3. GB2312

GB2312 is a key official character set of the People's Republic of China for Simplified Chinese characters, which is widely used in mainland China and Singapore. GB2312 contains 6,763 Simplified Chinese characters.

3. Related Freely Available Resources

3.1. Unihan Database

Unihan database[1] is the repository for the Unicode Consortium's collective knowledge regarding the CJK (Chinese-Japanese-Korean) Unified Ideographs contained in the Unicode Standard[2]. The database consists of a number of fields containing data for each Chinese character in the Unicode Standard. These fields are grouped into categories according to the purpose they fulfil, including "mappings", "readings", "dictionary indices", "radical stroke counts" and "variants" etc. The "mappings" category and "variants" category contain information regarding to the relation between Kanji and Hanzi.

3.2. Hanzi Converter Standard Conversion Table

Chinese encoding converter[3] is a open source system that can convert Traditional Chinese into Simplified Chinese. Hanzi converter standard conversion table is a resource used by the converter. This table contains 6,740 corresponding Traditional Chinese and Simplified Chinese character pairs. It can be downloaded from the web site. Table 2 is a portion of the table.

3.3. Kanconvit Mapping Table

Kanconvit[4] is a publicly available tool for Kanji-Simplified Chinese conversion. It uses 1,159 visual variational Kanji-Simplified Chinese character pairs extracted from a Kanji,

[1]http://unicode.org/charts/unihan.html

[2]The Unicode Standard is a character coding system for the consistent encoding, representation and handling of text expressed in most of the world's writing systems. The latest version of the Unicode Standard is Version 6.1.0.

[3]http://www.mandarintools.com/zhcode.html

[4]http://kanconvit.ta2o.net/

Kanji	安	詞	会	広	壱	潟	...
TC	安	詞	會	廣	壹	潟	...
SC	安	词	会	广	壹	泻	...

Table 3: Kanconvit mapping table.

Traditional Chinese and Simplified Chinese mapping table, which contains 3,506 one to one mappings. Table 3 is a portion of this table.

4. The Method

According to the relation between Kanji and Hanzi, we define seven categories for Kanji:

- Category 1: identical to Hanzi
- Category 2: identical to Traditional Chinese but different from Simplified Chinese
- Category 3: identical to Simplified Chinese but different from Traditional Chinese
- Category 4: visual variations
- Category 5: only have corresponding Traditional Chinese
- Category 6: only have corresponding Simplified Chinese
- Category 7: no corresponding Hanzi exist

We make a Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese by classifying JIS Kanji into these seven categories and finding the corresponding Traditional Chinese and Simplified Chinese automatically using the resources we introduced in Section 3.. The method includes two steps:

- Step 1: extraction
- Step 2: classification and construction

In Step 1, we extract JIS Kanji, Big5 Traditional Chinese and GB2312 Simplified Chinese from Unihan database. These Chinese characters are collected in the "mappings" category which contains mappings between Unicode and other encoded character sets for Chinese characters. JIS Kanji are in "kIRG_JSource J0" field, Big5 Traditional Chinese are in "kBigFive" field, and GB2312 Simplified Chinese are in "kIRG_GSource G0" field.
In Step 2, we do classification for JIS Kanji and construct a mapping table. We automatically check for every JIS Kanji: If the Kanji exists both in Big5 and GB2312, it belongs to Category 1; If the Kanji exists only in Big5, we check whether corresponding Simplified Chinese could be found, if so, it belongs to Category 2, otherwise, it belongs to Category 5; If the Kanji exists only in GB2312, we check whether corresponding Traditional Chinese could be found, if so, it belongs to Category 3, otherwise, it belongs to Category 6; If the Kanji exists neither in Big5 nor GB2312, we check whether corresponding Hanzi could be found, if fully corresponding Chinese characters in Hanzi exist, it belongs to Category 4, if only corresponding Traditional

Kanji	弁	伝	鯰	働	...
TC	弁,瓣,辦,辯,辮,辨	傳,伝	鯰	動,仂	...
SC	弁,瓣,办,辩,辫,辨	传	鲶,鲇	动,仂	...

Table 4: Examples of multiple Hanzi forms.

	C1	C2	C3	C4	C5	C6	C7
Unihan	3141	1815	177	533	384	16	289
+Han	3141	1843	177	542	347	16	289
+Kan	3141	1847	177	550	342	16	282

Table 5: Resource statistic (Han denotes Hanzi converter standard conversion table and Kan denotes Kanconvit mapping table).

Chinese exist, it belongs to Category 5, and if only corresponding Simplified Chinese exist, it belongs to Category 6, otherwise, it belongs to Category 7. To find corresponding Hanzi, we search Traditional Chinese variants, Simplified Chinese variants and other variants for all Kanji.

We do Traditional Chinese variants, Simplified Chinese variants and other variants search using "variants" category in Unihan database, in which there are five fields: "kTraditionalVariant" corresponding to Traditional Chinese variants, "kSimplifiedVariant" corresponding to Simplified Chinese variants, "kZVariant", "kSemanticVariant" and "kSpecializedSemanticVariants" corresponding to other variants. For supplement, we also use Hanzi converter standard conversion table and Kanconvit mapping table, notice that resources in Hanzi converter standard conversion table could only be used for Traditional Chinese variants and Simplified Chinese variants search, but Kanconvit mapping table could also be used for other variants search.

5. The Resource

5.1. Format

Format for Kanji in Category 1, 2, 3 and 4 in the mapping table is as follow:

- Kanji[TAB]TC[TAB]SC[RET]

If multiple Hanzi forms exist for one Kanji, we separate them with ",". Table 4 shows some examples of multiple Hanzi forms. Formats for Kanji in Category 5, 6 and 7 are as follows:

- Category 5: Kanji[TAB]TC[TAB]N/A[RET]
- Category 6: Kanji[TAB]N/A[TAB]SC[RET]
- Category 7: Kanji[TAB]N/A[TAB]N/A[RET]

5.2. Statistic

Table 5 shows the statistic of Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese we made. We can see that compared to only using Unihan database, Hanzi converter standard conversion table and Kanconvit mapping table can improve the completeness of the mapping table. Table 6 gives some examples of increased Chinese characters mappings found by Hanzi converter standard conversion table, Table 7 gives some examples of increased Chinese characters mappings found by Kanconvit mapping table.

Kanji	祇	託	浄	畷	...
TC	祇,只,衹,隻,祇	託,侘,托	淨,凈	畜	...
SC	祇,只	托	凈	畜	...

Table 6: Examples of increased mappings found by Hanzi converter standard conversion table.

Kanji	雰	艶	対	県	挿	...
TC	氛,雰	豔,艷	對	縣	插	...
SC	氛	艳	对	县	插	...

Table 7: Examples of increased mappings found by Kanconvit mapping table.

6. Completeness Evaluation

To show the completeness of the mapping table we made, we used a resource from Wiktionary[5] which is a wiki project aiming to produce a free-content multilingual dictionary. In the Japanese version of Wiktionary, there is a Kanji category, in which a lot of information about Kanji is provided, such as variants, origins, meanings, pronunciations, idioms, Kanji in Chinese and Korean, codes etc. We are interested in the variants part. Figure 1 gives an example of Kanji "広" in Japanese Wiktionary, the variants part is boxed, which contains Traditional Chinese variant "廣", Simplified Chinese variant "广" and other variant "慶" of Kanji "広".

We downloaded Japanese Wiktionary database dump data [6] (2012-Jan-31) and extracted variants for JIS Kanji. We constructed a mapping table based on Wiktionary using the same method described in Section 4., the only difference is that for Traditional Chinese variants, Simplified Chinese variants and other variants search, we used the variants extracted from Japanese Wiktionary.

To evaluate the completeness of the mapping table made by proposed method, we compared the statistic with Wiktionary. Table 8 shows the completeness comparison between proposed method and Wiktionary. We can see that proposed method makes a more complete mapping table than Wiktionary. Table 9 gives some examples of Chinese characters mappings found by proposed method, which do not exist in the current version of Wiktionary.

Furthermore, we did experiments by combining the mapping table we made with Wiktionary, results in Table 8 show that Wiktionary can be a good supplement to further improve the completeness of the mapping table. Table 10 gives some examples of Chinese characters mappings contained in Wiktionary, which are not found by proposed method.

7. Related Work

Hantology (Chou and Huang, 2006) is a character-based Chinese language resource adapting Suggested Upper

[5] http://www.wiktionary.org/

[6] http://dumps.wikimedia.org/jawiktionary/

Figure 1: Example of Kanji "広" in Japanese Wiktionary.

	C1	C2	C3	C4	C5	C6	C7
Prop	3141	1847	177	550	342	16	282
Wiki	3141	1781	172	503	412	30	316
Comb	3141	1867	178	579	325	16	249

Table 8: Completeness comparison between proposed method and Wiktionary (Prop denotes proposed method, Wiki denotes Wiktionary and Comb denotes combination).

Merged Ontology (SUMO) (Niles and Pease, 2001) for systematic and theoretical study of Chinese characters. Hantology represents orthographic forms, the evolution of script, pronunciations, senses, lexicalization as well as variants for different Chinese characters. However, the variants in Hantology are limited in Chinese Hanzi.

Chou et al. (2008) extended the architecture of Hantology to Japanese Kanji, and put links between Chinese Hanzi and Japanese Kanji, which provides a platform to analyze the variation of Kanji systematically. However, detailed analysis for variants of Kanji is not reported. Moreover, because the current version of Hantology only contains 2,100 Chinese characters, while we made the mapping table for all 6,355 JIS Kanji, it is hard to make a mapping table between Kanji and Hanzi as complete as our proposed method based on Hantology.

8. Conclusion

In this paper, we proposed a method of making a Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese automatically using freely available resources from Unihan database, Hanzi converter standard conversion table and Kanconvit mapping table. We defined seven categories for Kanji based on the relation between Kanji and Hanzi, and classified mappings of Chinese characters into these categories. The mapping table can be very useful for many Japanese-Chinese bilingual tasks.

For completeness evaluation, we used a resource from Wiktionary. Statistical comparison between our proposed method and Wiktionary showed that our proposed method makes a more complete mapping table than the current version of Wiktionary. We also found that Wiktionary can be a good supplement to further improve the completeness of the mapping table. Therefore, in the future, we are planning to obtain the updated information from Wiktionary, which can be helpful to make the mapping table more complete.

Kanji	龙	荔	値	幇	咲	...
TC	龙,龍	荔	值	幫	笑	...
SC	龙	荔	值	帮	笑	...

Table 9: Examples of mappings do not exist in Wiktionary.

Kanji	冴	扨	疂	滝	慎	...
TC	冴,洟	扠,叉	疊	瀧	慎	...
SC	冴	叉	叠	泷	慎	...

Table 10: Examples of mappings not found by proposed method.

9. References

Ya-Min Chou and Chu-Ren Huang. 2006. Hantology: A linguistic resource for chinese language processing and studying. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 587–590, Genoa, Italy, May.

Ya-Min Chou, Chu-Ren Huang, and Jia-Fei Hong. 2008. The extended architecture of hantology for kanji. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1693–1696, Marrakech, Morocco, May.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2011. Japanese-chinese phrase alignment using common chinese characters information. In *Proceedings of MT Summit XIII*, pages 475–482, Xiamen, China, September.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2005. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 670–681.

Chu-Ren Huang, Ya-Min Chou, Chiyo Hotani, Sheng-Yi Chen, and Wan-Ying Lin. 2008. Multilingual conceptual access to lexicon based on shared orthography: An ontology-driven study of chinese and japanese. In *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, pages 47–54, Manchester, United Kingdom, August. Coling 2008 Organizing Committee.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. pages 2–9. ACM Press.

Chew Lim Tan and Makoto Nagao. 1995. Automatic alignment of Japanese-Chinese bilingual texts. *IEICE Transactions on Information and Systems*, E78-D(1):68–76.