Building a Japanese-Chinese Dictionary Using Kanji/Hanzi Conversion

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{ling-g, masayu-a, matsu}@is.naist.jp

Abstract. A new bilingual dictionary can be built using two existing bilingual dictionaries, such as Japanese-English and English-Chinese to build Japanese-Chinese dictionary. However, Japanese and Chinese are nearer languages than English, there should be a more direct way of doing this. Since a lot of Japanese words are composed of kanji, which are similar to hanzi in Chinese, we attempt to build a dictionary for kanji words by simple conversion from kanji to hanzi. Our survey shows that around 2/3 of the nouns and verbal nouns in Japanese are kanji words, and more than 1/3 of them can be translated into Chinese directly. The accuracy of conversion is 97%. Besides, we obtain translation candidates for 24% of the Japanese words using English as a pivot language with 77% accuracy. By adding the kanji/hanzi conversion method, we increase the candidates by 9%, to 33%, with better quality candidates.

1 Introduction

Bilingual dictionaries have unlimited usage. In order for one to learn a new language, a bilingual dictionary can never be absent. In natural language processing community, bilingual dictionaries are useful in many areas, such as machine translation and cross language information retrieval.

In this research, we attempt to build a Japanese-Chinese dictionary using public available resources. There are already some existing Japanese-Chinese dictionaries, such as Shogakukan's Ri-Zhong Cidian [1], but they are not publicly available in electronic form. Our purpose is to build an electronic dictionary from public resources and make it public available.

The first dictionary that we use is IPADIC [2], a Japanese dictionary used by ChaSen [3], a Japanese morphological analyzer. We extract only nouns, verbal nouns and verbs from this dictionary, and try to search for their translation equivalents in Chinese.

One can build a new bilingual dictionary for a new pair of languages using two bilingual lexicons [4–8]. Since it is always easier to get bilingual dictionaries that involve English as one of the languages, using English as a pivot language is possible. In this case, we first look for the English translations of one language, and then try to find the possible candidates in the other language through English. Then we rank the candidates according to the similarities between the two words

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 670–681, 2005.

using some linguistic knowledge and statistical information. In our research, we make use of two public resources, EDICT [9] - a Japanaese-English dictionary and CEDICT [10] - a Chinese-English dictionary, to create the new language pair Japanese-Chinese dictionary using English as a pivot language. We obtain 77% accuracy. However, this method extracts only translations for about 24% of the Japanese words in IPADIC because the EDICT and CEDICT dictionaries are smaller compared with IPADIC. Therefore, we also look into the possibility to get the translation words using kanji/hanzi conversion. In Japanese, there are three types of characters, namely hiragana, katakana and kanji. Kanji characters are similar to Chinese ideographs. In Chinese, all characters are written in hanzi. Since most of the kanji characters are originally from China, the usage should remain unchangeable in certain contexts. The kanji/hanzi conversion method works only on Japanese words that consist only kanji characters. We obtain a high accuracy of 97% using this conversion. By combining the two methods, we increase the number of translation candidates by 9%, from 24% to 33%.

2 Previous Work

Tanaka and Umemura [4] used English as an intermediate language to link Japanese and French. They are the first who proposed the inverse consultation. The concept behind is that a translation sometimes may have wider or narrower meaning than the source word. They first look up the English translations of a given Japanese word, then the French translations of these English translations. This step gives a set of French candidates equivalent to the Japanese word. For each French candidate, its translations in English is collected. The similarity between the Japanese word and the French word is measured by the number of matches in their English translation. The more matches show the better candidate. This is referred to as "one time inverse consultation". The extension can be furthered by looking up all the Japanese translations of all the English translation of a given French word and seeing how many times the Japanese word appears; this is referred to as "two times inverse consultation".

Bond at al. [6] applied the "one time inverse consultation" in constructing a Japanese-Malay dictionary using a Japanese-English dictionary and a Malay-English dictionary. They also applied the semantic matching using part of speech and second-language matching. Matching only compatible parts of speech could cut down a lot of false matches. The second-language matching score used Chinese as a second intermediate language. If a word pair could be matched through two different languages, it is considered a very good match. Their research showed that about 80% of the translations are good if only highest rank pairs are considered, and 77% for all pairs.

Shirai and Yamamoto [7] used English as an intermediate language to link Korean and Japanese. They tried on 1,000 Korean words and were able to obtain the translations for 365 of them. They achieved an accuracy of 72% when the degree of similarity calculated by one time inverse consultation is higher than a predefined threshold.

Zhang et al. [8] used the same approach, that is using English as a pivot language, for constructing Japanese-Chinese pairs. They used the one time inverse consultation method and also the part of speech information for ranking. Since there is similarity between Japanese kanji and Chinese hanzi, they have further improved on the method by using the kanji information [11]. First they searched for the Chinese translations of single character words in Japanese into using one time inverse consultation. If the Unicode of the two characters are the same, then the ranking is higher. After getting this list of character pairs, the similarity between the Japanese word and the Chinese word is calculated using the edit distance algorithm [12]. Finally, the score obtained from the kanji information is added to the final score function. Their ranking method was improved and the precision increased from 66.67% to 81.43%. Since only about 50% of their Japanese words can be translated into Chinese, they also searched for other approaches to translate the remaining words [13] using web information and machines translation method.

Our work is quite similar to Zhang et al. [11] in the way they constructed the kanji/hanzi conversion table. The difference is that instead of calculating the similarity between kanji and hanzi using Unicode and one time inverse consultation, we make a direct conversion from kanji to hanzi based on the ideographs. Our method sounds more intuitive and direct because kanji and hanzi are of the same origin. Later on, they made use of this conversion table to calcutate the similarity between a Japanese word and a Chinese word from the output of using English as the pivot language. Their method can make the similar Chinese words to have higher ranking but cannot generate new translation candidates. On the other hand, our methods works for both.

3 The Proposed Methods

We propose to combine two methods to find the translations of Japanese entries in IPADIC version 2.7.0 [2]. IPADIC is a monolingual dictionary and consists of 239,631 entries. We only extract nouns, verbal nouns and verbs (a total of 85,553 entries) in our survey. First, we use English as the pivot languege. Second, we make direct conversion from kanji to hanzi for kanji word translation. We now describe in detail the both methods.

3.1 Using Third Language: English

First, we use English as the pivot language to find the translations from Japanese to English, and then from English to Chinese. Since IPADIC is a monolingual dictionary, we use EDICT as the Japanese-English dictionary. The EDICT version (V05-001) consists of 110,424 entries. There exist some words that are polysemous with multiple entries. After combining the multiple entry words, we have 106,925 unique entries in the dictionary. For English to Chinese, we use the CEDICT dictionary. It consists of 24,665 entries. A word can be polysemous in both

dictionary, meaning that for each word there is only one entry but with multiple translations. All the English translations of different senses are in the same record. We assume that a bilingual dictionary should be bi-directional, therefore we reverse the CEDICT dictionary to obtain an English-Chinese dictionary.

The ranking method is the one time inverse consultation [4, 6–8]. Since a word can be polysemous in both dictionaries, if a source word shares more English translations with the target translation word, then they can be considered nearer in meaning. The score is calculated as in equation (1): Let $S_E(J, C_i)$ denotes the similarity between the Japanese word J and the Chinese translation word candidate C_i, where $E(J)$ and $E(C_i)$ are the sets of English translations for J and C_i, respectively:

$$S_E(J, C_i) = \frac{2 \times (|E(J) \cap E(C_i)|)}{|E(J)| + |E(C_i)|} \tag{1}$$

Currently we do not apply the part of speech information in the scoring because this method requires linguistic experts to decide on the similarity between two part of speech tags for different languages[1]. However, this will become part of our future work.

Table 1 shows the results of using English as the pivot language and one time inverse consultation as the scoring function. Using the EDICT and CEDICT only, 32,380 Japanese words obtain their Chinese translation candidates. In total, we obtain 149,841 pairs of translation. We get maximum 90 candidates for a Japanese word, and 4.6 candidates per word by average. Then we check the Japanese words in IPADIC to get their part of speech tags. We only investigate on three categories of part of speech tags from the IPADIC, which are nouns, verbal nouns and verbs. We randomly selected 200 Japanese words from each category for evaluation. The results are judged using 4 categories: **Correct** means that the first rank word is correct (if there are multiple words in the first rank, it is considered correct if any one of the words is correct), **Not-first** means that the correct word exists but not at the first rank, **Acceptable** means that the first rank word is acceptable, and **Wrong** means that all candidates are wrong. All the categories are exclusive of each other.

Table 1. Ranking results

POS	Total	Translated	Correct	Not-first	Acceptable	Wrong
Nouns	58,793	14,275 (24.3%)	152	4	20	24
Verbal nouns	12,041	3,770 (31.3%)	90	12	37	61
Verbs	14,719	2,509 (17.0%)	101	18	27	54

There are about 24.3% of nouns, 31.3% of verbal nouns and 17.0% of verbs in IPADIC that give us some translation candidates in Chinese. For the evaluation

[1] There are 120 part of speech tags (13 categories) in IPADIC, and 45 in Peking University dictionary. Both define some quite specialized part of speech tags which only exist within the dictionary itself.

using 200 randomly selected words, we obtain 88%, 69.5% and 73% accuracy, respectively. The accuracy is 76%, 45% and 50.5%, respectively, if we considered only the first rank. The accuracy is a bit lower compared with previous work as we did not apply other linguistic resources such as parts of speech for scoring. Although improving scoring function can make the rank of the correct words higher, it cannot further increase the number of candidates. Since both EDICT and CEDICT are prepared by different people, the way they translate the words also varies. Furthermore, there is no standization on the format. For example, to represent a verb in English, sometimes it is written in base form (e.g. "discuss"), and sometimes in infinitive form (e.g. "to discuss"). In Chinese, and sometimes in Japanese too, a word shown in the dictionary can be a noun and a verb without inflection. The part of speech category can only be decided based on the usage in contexts. Therefore the same word may be translated into a noun in English too (e.g. "discussion"). It happened too that we cannot find the matches just because of singular form or plural form (e.g. "discussions") of the English translation. With these non-standardization of the English translation, we cannot match the exact words unless we do a morphological analysis in English. Therefore, we also look for other ways to increase to number of candidates. Since Japanese and Chinese share some common characters (kanji in Japanese and hanzi in Chinese), we are looking into the possibility of direct conversion to create the translations. We discuss this method in the following section.

3.2 Direct Conversion of Kanji/Hanzi

Using English as the pivot language is a good starting point to construct a new language pair. However, there remain a lot of words for which the translations cannot be obtained. In Chinese, all the characters are hanzi, but in Japanese, there are hiragana, katakana and kanji. The kanji characters are originated from ancient China. This group of characters, used in China, Japan and Korea, are referred to as Han characters. The Han characters capture some semantic information which should be common in those languages. One can create a new word by combining the existing characters but it is hardly that one can create a new character. Therefore, these characters are stable in their meaning. Due to the common sharing on these Han characters, there might be a more straightforward way to translate a word in Japanese into Chinese if all the characters in the word are made up from kanji only. We refer to this kind of words as kanji words.

A Chinese word can be a noun or a verb without changes of morphological forms. There is no inflection to differenciate them. EDICT and CEDICT make no difference on the parts of speech and therefore the translations in English can be in any form. For example, the following Japanese words and Chinese words exist for the translations of "discussion/discussions/to discuss/discuss".

Japanese: 会談, 議論, 協議, 言論, 商量, 相談, 討議, 討論, 付議, 論議, 交渉, 座談, 詮議, 談論, 評議, 弁論, 話し合う, 話合い, 論, 論う, 論じる, 論ずる

Chinese: **辩**, **会谈**, 论, **评**, **评论**, **洽谈**, **商谈**, **商讨**, **谈**, **谈论**, **讨论**, **议**, **议论**

If we were to match each Japanese word to each of the Chinese words (in fact, we can say that some of them are acceptable translations), then we will get a redundency of 286 (22×13) pairs. Although these words have similar translation, but in fact they have slight differences in meaning. For example, "会談" means the conference amongst the ministers, "交涉" means negotiations. However, "discussion" is one of the translations in English as provided by EDICT. Since the Japanese kanji characters are originated from China, translating Japanese kanji words directly to Chinese can be more accurate than going through a third language like English. If we look from the Japanese side, 12 out of 22 words (会談, 議論, 協議, 言論, 商量, 討論, 交涉, 座談, 談論, 評議, 弁論, 論) could get their exact translations by just simple conversion of kanji/hanzi (**会谈**, **议论**, **协议**, **言论**, **商量**, **讨论**, **交涉**, **座谈**, **谈论**, **评议**, **辩论**, 论), in which some of them cannot get the translations using English. On the other hand, there also exist some words that are not translated into the semantic meaning "discuss" in Japanese but in Chinese, such as "**评论**" which should be the same as "評論" in Japanese[2]. For the single character words in Chinese (**辩**, **评**, **谈**, 议), they are seldom used in Japanese but they do exist with the same meaning (弁, 評, 談, 議).

There exist equivalent characters between Japanese kanji and Chinese hanzi. Both type of characters (Han characters in general) capture significant semantic information. Although the pronunciation varies across languages, the visual form of the characters retains certain level of similarity. Furthermore, Chinese characters can be divided into the characters used by mainland China (referred to as Simplified Chinese) and Taiwan (including Hong Kong and Macao, referred to as Traditional Chinese). Although the ideographs may be different, they are originally the same characters. Most of the Japanese characters are similar to Traditional characters.

Table 2. Successful Traditional-Simplified examples

English	love	garden	rice	fly	kill	talk	fill up	post	excellent	sun
Japanese	愛	園	飯	飛	殺	話	補	郵	優	陽
Traditional Chinese	**愛**	園	**飯**	**飛**	**殺**	**話**	**補**	郵	優	**陽**
Simplified Chinese	**爱**	园	**饭**	**飞**	**杀**	**话**	**补**	邮	优	**阳**

Our original Japanese characters are coded in EUC and Chinese characters are coded in GB-2312 codes. To convert a kanji to a hanzi is not a trivial task. Of course most of the characters share the same ideographs. In this case, we can use the Unicode for the conversion as these characters share the same Unicode. However, there exist also quite a number of characters in Japanese that are written in Traditional Chinese ideographs. We have to convert these characters from Traditional Chinese to Simplified Chinese (see Table 2). Finally, there are

[2] The meaning of "商談" (a business talk) in Japanese is different from the meaning of '**商谈**'" (to discuss verbally) in Chinese.

Table 3. Unsuccessful Traditional-Simplified examples

English	gas	hair	deliver	check	home	pass by	burn	bad	money	whole
Japanese	気	髪	発	検	郷	経	焼	悪	銭	総
Traditional Chinese	氣	髮	發	檢	鄉	經	燒	惡	錢	總
Simplified Chinese	气	发	发	检	乡	经	烧	恶	钱	总

Table 4. Japanese-GBK examples

English	sardine	hackberry	maple	kite	inclusive
Japanese	鰯	榎	椛	凧	込
English	crossroad	field/patch	rice bowl	carpentry	chimera
Japanese	辻	畑	丼	杢	鵺

also some characters in Japanese having similar ideographs, but they are neither Traditional Chinese nor Simplified Chinese (see Table 3). We manually convert these characters by hand. The following shows the steps to convert the characters from Japanese to Chinese.

1. Convert from EUC to Unicode using iconv.
2. Convert from Unicode to Unicode-simplified using a Chinese encoding converter[3]. This step converts possible Traditional characters to Simplified characters.
3. Convert from Unicode-simplified to GB-2312.
4. Those failed to be converted are edited manually by hand.
5. Those characters that do not exist in GB-2312 are converted into GBK using the Chinese encoding converter.

From IPADIC, we extract 36,069 and 8,016 kanji words from noun and verbal noun categories[4], respectively. From these words, we get 4,454 distinct kanji characters. Out of these characters, only 2,547 characters can be directly converted using Unicode without changes of ideographs. 1,281 characters are converted from Traditional Chinese to Simplified Chinese using the Chinese encoding converter. Finally 626 characters are manually checked and 339 characters can be converted to Simplified Chinese. 287 remain in Japanese ideographs but are converted into GBK codes [5]. Most of these words are the names of plants, fish, and things invented by Japanese (see Table 4). While these GBK coded words may not be used in Chinese, we just leave them in the conversion table for the sake of completeness.

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

[4] These two categories consist of most of the kanji words in Japanese. However, verbs are normally hiragana only or a mixture of kanji plus hiragana. Therefore, we omit verbs in this survey.

[5] GBK codes consist of all simplified and traditional characters, including their variants. Therefore, Japanese characters can also be coded in GBK. However, they are rarely used in Chinese.

About 61% of nouns and 67% of verbal nouns in Japanese are kanji words as shown in Table 5. Using the conversion table described above, we convert the kanji words into Chinese words. Then, we consult these words using a Chinese dictionary provided by Peking University [14]. There are about 80,000 entries in this dictionary. About 33% of the nouns and 44% of the verbal nouns are valid words in Chinese. We randomly select 200 words for evaluation. We evaluate the results by 3 categories: **Correct** means that the translation is good, **Part-of** means that either the Japanese word or the Chinese word has a wider meaning, and **Wrong** means that the meanings are not the same though they have the same characters. The accuracies obtained are 97% for nouns and 97.5% for verbal nouns. The pairs that have part-of meaning and different meaning are listed in Table 6 and Table 7 for references.

Table 5. Kanji/Hanzi conversion results

POS	Total	Kanji words	Translated	Correct	Part-of	Wrong
Nouns	58,793	36,069 (61%)	11,743 (33%)	189	5	6
Verbal nouns	12,041	8,016 (67%)	3,519 (44%)	190	5	5

Table 6. Part-of translation examples

Japanese	Chinese
被害 (damage; casualty; victim)	被害 (be murdered; victimization)
侍 (samurai; warrior; servant)	侍 (servant)
一角 (a corner; competent)	一角 (a corner; a unit used for money;)
熨斗 (charcoal iron; noshi - greeting paper)	熨斗 (charcoal iron)
会意 (character formation type)	会意 (character formation type; knowing; understanding)
苦学 (work one's way)	苦学 (hardship study)
安置 (set in place - Buddha statue or corpse)	安置 (set in place - for anything)
作為 (artificiality; deliberateness; aggressive action)	作为 (action; acomplishment; regard as)
下落 (fall; drop)	下落 (fall; drop; whereabouts; find a place for; reprove)
供奉 (attend on the Emperor in his travels; accompany in the imperial trains)	供奉 (offer sacrifice to; people gave commend performances in an imperial palace)

The advantage of this method is that we can get exact translation for those borrowed words from Chinese, especially idioms. We all know that it is always difficult to translate idiomatic phrases from one language to another due to the different cultural background. If we were to use English as the pivot language to translate from Japanese to Chinese, it is difficult to have two different bilin-

Table 7. Wrong translation examples

Japanese	Chinese
本当 (true; really)	本当 (ought; should)
員外 (nonmember)	员外 (ministry councillor; landlord)
流窺 (deportation; banishment; exile)	流窜 (flee hither and thither)
花色 (light indigo)	花色 (variety; designs and colors)
野菜 (vegetables)	野菜 (potherb)
画幅 (picture scroll)	画幅 (size of a picture)
折半 (divide into halves)	折半 (reduce by half)
自訴 (self-surrender)	自诉 (private prosecution)
打算 (selfish; calculating)	打算 (plan; intend; calculate)
開立 (search for cube root)	开立 (draw; issue; open)
勾引 (take a person into custody)	勾引 (seduce; tempt)

gual dictionaries from two different publishers that translate them in the same wordings. Since a lot of the idioms in Japanese are originally from China, the conversion of kanji/hanzi will make the translation process faster and more accurate. Some examples are given below.

同床異夢 (same bed different dream - cohabiting but living in different worlds)
鶏口牛後 (better to be the beak of a rooster than the rump of a bull - better to be the leader of a small group than a subordinate in a large organization)
神出鬼没 (appearing in unexpected places and at unexpected moments)

The difficulty of this method is the translation of single character words. Single character words normally have wider meaning (multiple senses) and the usage is usually based on the context. It is fair enough if we translate the single character words using the conversion table. However, these characters should have more translations of other multi-character words. There are 2,049 single character nouns in Japanese and 1,873 of them exist in Chinese after the conversion. For verbal nouns, there are 128 Japanese words and 127 words exist in Chinese (only 噂 (gossip, rumor) does not exist in Chinese).

3.3 Intergration

We combine both using English as the pivot language and kanji/hanzi conversion method to get the final list of translation candidates. Table 8 shows the results in details. We obtain 20,630 for nouns and 5,356 for verbal nouns. In total, we obtain 28,495 words, in which 7,941 words are new translations. Furthermore, we add in high quality translation candidates into the new bilingual dictionary. 2,428 of the candidates obtained using kanji/hanzi conversion method already exist in the translation candidates using English as the pivot language. This can help to double check on the list of translation candidates and make them rank higher. 4,893 candidates are served as extra and better quality candidates on top of the translation candidates obtained using English as the pivot language.

Table 8. Integration results

POS	Kanji/ hanzi	Acc.	Est.	Using English	Acc.	Est.	Total	In	Extra	New
Nouns	11,743	97%	11,391	14,275	88%	12, 562	20,630	2,008	3,380	6,355
Verbal nouns	3,519	97.5%	3,431	3,770	69.5%	2,620	5,356	420	1,513	1,586
Verbs	-	-	-	2,509	73%	1,832	2,509	-	-	-
Total	15,262		14,822	20,554		17,014	28,495	2,428	4,893	7,941

As an estimation, we will get about 17,014 Japanese words with correct translations in Chinese using English as the pivot language. By using kanji/hanzi conversion method, we could get about 14,822 words with correct translation.

4 Discussion and Future Work

In our survey, only 33% of nouns and 44% of verbal nouns created by kanji/hanzi conversion method exist in the Peking University dictionary. However, this may be due to the incompleteness of the Chinese dictionary that we used. We also found some words after the conversion which are acceptable in Chinese though they do not exist in the dictionary. Some of the examples are as follows: "自闭症(autism), 护法(the defense of Constituition or religion), 第六感(sixth sense), 先帝(the preceding emperor), 玄奥(deep sense), 误信(misbelief)". Therefore, we can further verify the validity of the Chinese words using other resources such as the information from the web.

The current work consider only kanji/hanzi conversion for Japanese words that consists on kanji only. There are a lot of words in Japanese that are mixture of kanji and hiragana. This happens normally with verbs and adjectives. For example, "食べる(eat), 逃げる(escape), 生み出す(produce), 難しい(difficult), 嬉しい(happy), 静か(quite)". We should be able to get some acceptable translations of these words after removing the hiragana parts, but most of the cases we cannot obtain the best or good translations. From the 200 verbs that we used for the evaluation, 139 words exist in Chinese but only 35 are good and 43 are acceptable. The single characters used in these words are normally used only in ancient Chinese but not in contemporary Chinese. For example, 食べる = 吃 (eat) and 捨てる = 丢掉 (throw away), but 食 (eat) and 舍 (throw away) in Chinese are also possible translation in certain contexts. Furthermore, the contemporary Chinese uses two character words more often than single character words even they have the same meaning. This is to reduce the semantic ambiguity as single character words tend to be polysemous. Therefore, direct kanji/hanzi conversion is not so appropriate and we need another approach to handle this type of words.

We can apply the kanji/hanzi conversion method directly to most of the Japanese proper nouns, such as person names, organization names and place names because these names are normally written in kanji characters. Therefore, we do not need any effort to translate these words from Japanese to Chinese if

we have the character conversion table. This will ease a lot in the processing of machine translation and cross language information retrieval.

The Unicode Consortium encoded the Han characters in Unicode[6]. Till date, all the languages that use Han characters have their own encoding systems. For example, Japanese is encoded in EUC-JP or JIS, Simplified Chinese is in GB-2312, Traditional Chinese is in Big 5 etc. The same character that is used in different languages is assigned with different codes. Therefore it is difficult to convert from one code to another without a conversion table. The Unicode Consortium solved the problem by unifying the encoding. The same character with the same ideograph has only one code no matter in which language it is used. With this unification, it eased a lot on the CJK research, especially in the area of cross language information retrieval. Currently, they have increased the number of Han characters from 27,496 characters (version 3.0) to 70,207 characters (version 4.0). Such a huge increment is done by the addition of a large amount of unusual characters that only have been used in either person names or place names. With this new version, it covers almost all possible characters used in hanzi (Chinese), kanji (Japanese) and hanja (Korean). The Unihan (unicode for Han characters) provides a lot of information such as the origin, the specific language using that character, conversion to other encodings etc. The most useful information in Unihan to our research is the relationship between the characters. It embeds the links for the variants of characters which are useful for the conversion from one encoding to the others (Japanese, Traditional Chinese, Simplified Chinese or Korean). If we can make use of this table, then we can build a complete conversion table that includes all Han characters.

Zhang et al. [11] proposed to use kanji information to find the similarity between a Japanese word and a Chinese word. They matched on the Unicode and calculated the similarity using the one time inverse consultation. Since they did not make any conversion such as traditional characters to simplified characters, some of the characters have the same meaning but different Unicodes. Therefore, they could not be matched. If they could use the conversion table that we proposed, then it would help to increase the score of the kanji words.

To convert from Japanese kanji to Simplified characters is easier than the reverse. It is because some characters in Traditional characters are simplified into the same characters in Simplified Chinese. For example, 髪 (hair) and 発 (deliver) are simplified to 发. Therefore, it has to depend on the contexts to decide which Japanese character to use if we were to convert the Chinese Simplified characters back to Japanese kanji.

5 Conclusion

As a conclusion, we proposed a method to compile a Japanese-Chinese dictionary using English as the pivot language as a starting point. We made use of the public available resources such as EDICT, CEDICT and IPADIC for the construction

[6] http://www.unicode.org/chart/unihan.html

of the new language pair. The accuracy obtained is 77%. Since Japanese and Chinese share common Han characters which are semantically heavy loaded, the same characters used should carry the same meaning. Therefore, we also proposed a kanji/hanzi conversion method to increase the translation candidates. The accuracy obtained is 97%. The increment of translation candidates is 9%, from 24% to 33%. The conversion table created can also be used in other fields like machine translation and cross language information retrieval.

Acknowledgements

This research uses EDICT file which is the property of the Electronic Dictionary Research and Development Group at Monash University. Thanks go to http://www.mandarintools.com/zhcode.html for their Chinese Encoding Converter.

References

1. Shogakukan and Peking Shomoinshokan, editors: Ri-Zhong Cidian [Japanese-Chinese Dictionary] (1987)
2. Asahara, M., Matsumoto, Y.: IPADIC version 2.7.0. Users Manual. Nara Institute of Science and Technology, Nara, Japan. (2003) http://chasen.naist.jp/.
3. Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., Asahara, M.: Morphological Analysis System ChaSen version 2.2.9 Manual. Nara Institute of Science and Technology, Nara, Japan. (2002) http://chasen.naist.jp/.
4. Tanaka, K., Umemura, K.: Construction of a bilingual dictionary intermediated by a third language. In: Proc. of COLING. (1994) 297–303
5. Lafourcade, M.: Multilingual dictionary construction and services - case study with the fe* projects. In: Proc. of PACLING. (1997) 289–306
6. Bond, F., Sulong, R.B., Yamazaki, T., Ogura, K.: Design and construction of a machine-tractable japanese-malay dictionary. In: Proc. of MT Summit VIII. (2001) 53–58
7. Shirai, S., Yamamoto, K.: Linking english words in two bilingual dictionaries to generate another language pair dictionary. In: Proc. of ICCPOL. (2001) 174–179
8. Zhang, Y., Ma, Q., Isahara, H.: Automatic acquisition of a japanese-chinese bilingual lexicon using english as an intermediary. In: Proc. of NLPKE. (2003) 471–476
9. Jim Breem: EDICT, Japanese-English Dictionary (2005) http://www.csse.monash.edu.au/jwb/edict.html.
10. Paul Denisowski: CEDICT, Chinese-English Dictionary (2005) http://www.mandarintools.com/cedict.html.
11. Zhang, Y., Ma, Q., Isahara, H.: Use of kanji information in constructing a japanese-chinese bilingual lexicon. In: Proc. of ALR Workshop. (2004) 42–49
12. Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk SSSR **163** (1965) 845–848
13. Zhang, Y., Isahara, H.: Acquiring compound word translation both automatically and dynamically. In: Proc. of PACLIC 18. (2004) 181–185
14. Peking University: (Peking University Dictionary) http://www.icl.pku.edu.cn/.