Combining several automatic techniques to build a Chinese–Japanese lexicon from freely available resources

Wei YANG　　Yves LEPAGE
Graduate School of Information, Production and Systems, Waseda University
{kevinyoogi@akane,yves.lepage@aoni}.waseda.jp

Abstract

We constructed a Chinese-Japanese lexicon by combining several automatic techniques on several freely available resources. The basic technique used is the classical pivot language technique. To improve the quality of the resource built, we combined three additional different techniques: the first one is one time inverse consultation; the second one is Japanese kanji to Chinese hanzi character conversion; the third one is expansion through Chinese synonyms table. We used the Unihan Mapping Data, Langconv Traditional-Simplified Conversion data, and hanzi/kanji conversion table to maximize the conversion result, we also used the synonyms table data to confirm more translation pairs. The Chinese-Japanese lexicon built consists of more than 45,000 Chinese-Japanese word pairs with an accuracy of 85%.

1 Introduction

In the development of machine translation and cross-language information retrieval, it is necessary to construct bilingual dictionaries from one language to another, but the cost is enormous from the viewpoint of labor and time.

However, even if bilingual dictionaries do not directly exist for a particular source language and a particular target language, the possibility is high that bilingual dictionaries exist into an identical third language, particularly English. In other words, it is conceivable that a bilingual dictionary between Chinese and Japanese be built through a third language, like English.

In addition, kanji characters are similar to Chinese hanzi. We prppose, according to the similarity between kanji and hanzi, to compare the Unicode of Chinese words with Japanese words by kanji/hanzi conversion.

The paper is divided as follows: Section 2 presents the basic method to generate a Chinese-Japanese bilingual lexicon via English as the third language by joining two bilingual lexical resources for Chinese-English and Japanese-English. In Section 3 we describe a first additional method: one time inverse consultation [1]. Compared with the classical joining approach, this increases the accuracy. Section 4 describes the second additional method: using kanji/hanzi conversion and comparison between Chinese words and Japanese words. Section 5 gives a last improvement by expansion through Chinese synonyms table. By combining these three additional methods, we increased the translation candidates and the accuracy of the resulting Chinese-Japanese lexicon.

2 Construction with a classical pivot language technique

In this section, first we will describe the Chinese-English and Japanese-English dictionaries we use, and then how we combine them via English as the pivot language.

2.1 The XDXF dictionaries

XDXF dictionary[1] is a project to unite all existing open dictionaries and provide both users and developers with universal XML-based format, convertible to and from other popular dictionary formats. The Chinese-English XDXF dictionary consists of 26,617 articles, each article consists of three main components: (1) both traditional Chinese and simplified Chinese; (2) pronunciation in pinyin; (3) English translations. The Japanese-English XDXF dictionary we use consists of 108,473 articles. Some articles consist of the pronunciation in katakana, especially for those Japanese words made up from kanji only. From these two dictionaries we extract simplified Chinese and Japanese words only with their corresponding English translations. We make use of these two generated lexicons as our experimental primary resources. After eliminating duplicate lines, we obtained a Chinese-English and Japanese-English lexicons consisting of 43,389 and 105,182 entries respectively.

[1] http://xdxf.revdanica.com/down/

2.2 Crossing the lexicons

In a first step, we proceed as follows:

- Firstly, convert Chinese-English and Japanese-English data into lexicon resources.
- Secondly, output the translation tables (using Anymalign[2]) corresponding to Chinese-English and Japanese-English lexicons by computing translation probabilities.
- Thirdly, perform a join of the two translation tables through English as the pivot language and compute probabilities to get a Chinese-Japanese translation table. Here the join is the same as the algebraic operation on relational databases.

Hereafter, zh, en, and ja denote terms in the source language Chinese, terms in the pivot language English, and terms in the target language Japanese respectively. For the translation pairs (zh, en) and (en, ja), the translation probabilities $P(en|zh)$ and $P(ja|en)$ are computed using the maximum possiblity estimation from the co-occurrence frequencies that are consistent with the word alignment in the translation table:

$$P(en|zh) = \frac{P(zh, en)}{P(zh)} = \frac{C(en \leftrightarrow zh)}{C(zh)} \quad (1)$$

$$P(ja|en) = \frac{P(en, ja)}{P(en)} = \frac{C(ja \leftrightarrow en)}{C(en)} \quad (2)$$

In the equations, $C(x)$ denotes the number of occurences of the word or phrases x in the lexicon, and $C(x \leftrightarrow y)$ is the number of co-occurrences of x and y in the lexicon. In theory we calculate the direct translation probabilities between the source language Chinese and target language Japanese by the following equation (e.g. for the probability of the target language Japanese knowing the source language Chinese):

$$P(ja|zh) = \sum_{all\ pivot\ en} P(ja|en) \times P(en|zh) \quad (3)$$

One of the characteristics of using this approach is that we obtain all possible alignments as result in translation table. We can normalize translation probabilities after discarding any translation pair with both translation probabilities less than a threshold. The threshold we used was 0.05 for both $P(zh|ja)$ and $P(ja|zh)$.

We obtained a Chinese-Japanese lexicon consistsing of 119,203 pairs.

[2] http://www.unicaen.fr/

2.3 Evaluation

We extract ten samples with 100 translation alignments randomly and check manually in an existing bilingual dictionary. We calculate the accuracy of the result by p-value using Student's t-test:

$$T = \frac{(X - H0)}{S} \times \sqrt{n - 1} \quad (4)$$

With a null hypothesis of 45%, and an experimental result of 42.6%, the p-value is 0.06, above the usually supposed 0.05. We conclude that there is not enough evidence to state that the overall translation quality is higher than 45%. We infer that the quality lies below 45% of correct entries, or is even worse.

3 Using one time inverse consultation

In previous work, Tanaka and Umemura (1994) used inverse consultation method through English as a pivot language to improve a Japanese-French [1] lexicon built using the join method. We also rely on one time inverse consultation to find suitable equivalents for our Chinese-Japanese lexicon. We proceed as follows: first look up English translations of a Chinese word, then look up Japanese translations of these English translations; for each Japanese translation, look up how many English translations shared with the original Chinese word. The more matches there are, the better the Japanese translation candidate is. Figure 1 illustrates one time inverse consultation between Chinese and Japanese.

To measure the quality of the Japanese translation candidate, a similarity score is calculated according to a classical Dice coefficient formula:

$$SimilarityScore = \frac{2 \times |E(C) \cap E(J)|}{|E(C)| + |E(J)|} \quad (5)$$

Here E(C) and E(J) are the sets of English translations for the Chinese word and the Japanese word respectively.

Due to the relatively small sizes of the two lexicons we use, a similarity score equal to one does not necessarily mean that a translation pair is correct. As show in Figure 1, the similarity scores of "矿-鉱", "矿井-地雷火" and "矿山-水雷" are all equal to one, but only "矿-鉱" is a correct translation pair. Using this method may lead to generate many irrelevant translation candidates.

In our experiment we obtained 33,297 translation candidates. A same p-value evaluation of the results showed an accuracy of 76%. Compared with standard classical pivot technique, the number of translation candidates was reduced, but the accuracy in-

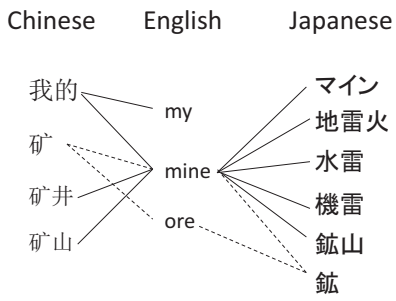

Figure 1: Sample of one time inverse consultation

creased. However, the problem of distinguishing ambiguous words was not resolved completely.

4 Increasing translation candidates by using Kanji/Hanzi conversion

Figure 1 leads to the observation that some translation pairs can be directly retrieved or reinforced by looking at the similariy between hanzi and kanji. In this figure, the pair "矿山-鉱山" is supported by the kanji/hanzi conversion of the first element "矿-鉱".

Consequently, we propose to convert Japanese words made up from only Japanese kanji into simplified Chinese characters through kanji/hanzi conversion. By doing so, we generate a ja'-ja file automatically where each line consists in the converted Japanese word (simplified Chinese) or without conversion ones and the original Japanese word. In this way, we avoid the difficult problem of converting Chinese simplified characters back to Japanese kanji [5]. By comparing ja' with the Chinese entries in the Chinese-English lexicon we can select more reliable Chinese-Japanese translation pairs.

Below, we explain how we combine three sources of data for our conversion experiments so as to maximize the result quality.

We also describe a method based on the use of synonyms table[3] to increase the candidates for those different words which share a similar meaning in Chinese and Japanese after kanji/hanzi conversion.

We combined three different sources of data to maximize our conversion results. Table 1 shows the relationships between Chinese (traditional and simplified) and Japanese. The Japanese words made up from kanji in the parts of "All same" and "TC different" (Traditional Chinese different) could compare the Unicode with Chinese directly without any conversion; the characters in "SC different" (Simplified Chinese different) become comparable by traditional Chinese to simplified Chinese conversion; for the "All different" and "Ja different" parts we propose to utilize hanzi/kanji conversion table (簡体字と日本漢字対照表[4]) to make them comparable with Chinese.

The first source of data we used was the Unihan database[5]. In particular we used the correspondence relation SimplifiedVariant in the Unihan Mapping Data of the Unihan database. There are 3,662 SimplifiedVariant pairs. Using them, we could check translation pairs between Japanese words (kanji) and simplified Chinese words (hanzi) in the following way. For each Japanese character, consider it as a traditional Chinese character, and look up for its corresponding simplified Chinese character through the SimplifiedVariant relation and replace it. If this simplified Chinese word (converted Japanese word) is the one in the Chinese lexicon, confirm the translation pair.

The second source of data we used was that of the Langconv Traditional-Simplified Conversion[6] data. It contains a wiki traditional-simplified conversion database, consisting of about 3,000 traditional to simplified conversion pairs. We perform similar experiments as above to confirm Chinese–Japanese word translation word pairs.

The third source of data we used concerns the case where the characters in Japanese are neither found in the traditional Chinese nor simplified Chinese character sets. For this case, we use a hanzi/kanji conversion table which consists of 2,236 simplified hanzi and kanji pairs. We use this table as described above for the two previous sources of data.

Table 2 shows the result of kanji/hanzi conversion using these three sources of data. There exist about 62,852 Japanese entries made up with kanji only from the Japanese-English lexicon, 36,590 Japanese words were converted successfully. For all Japanese words we confirm their simplified Chinese translations: 8,137 translation pairs were confirmed. The accuracy is 98.5%, which shows that the method is quite efficient.

5 Expansion through Chinese synonyms table

The last method we use to improve the quality of our Chinese-Japanese lexicon is to use synonyms table to extract more translation candidates for words in Chinese and Japanese that share similar meaning. Again, this applies for Japanese words consisting only of kanji, after conversion into simplified Chinese characters.

[3] http://ishare.iask.sina.com.cn/f/21267706.html

[4] http://www.kishugiken.co.jp/cn/code10d.html

[5] http://www.unicode.org/Public/UNIDATA/

[6] http://code.google.com/p/advanced-langconv/source/browse/trunk/langconv/?r=7

Table 1: Relationships between Chinese and Japanese.

Relationship	All same		TC different		SC different		All different		Ja different	
	word	center	counntry	learn	struct	wind	value	fight	multiplication	flame
Japanese	世界	中央	国	学	構造	風	価値	戦闘	乗法	火焔
T Chinese	世界	中央	國	學	構造	風	價值	戰鬥	乘法	火焰
S Chinese	世界	中央	国	学	构造	风	价值	战斗	乘法	火焰

Table 2: Result of kanji/hanzi conversion and zh–ja lexicon construction.

Method	Successful Conversion	zh–ja lexicon	Accuracy
Unihan	27,929 (44.4%)	6,856	98%
Langconv	28,153 (44.8%)	6,877	98.5%
Conoversion-Table	36,035 (57.3%)	8,012	98.5%
Combining Methods	36,590 (58.2%)	8,137	98.5%

The source of data we used consists of 17,170 synonym pairs. For each Chinese word found in the synonyms table, we checked whether its corresponding synonyms appears in the ja'-ja file. This allows to confirm translation pairs. Using this method, we obtained a Chinese-Japanese lexicon consisting of 3,952 pairs. The accuracy was shown to reach 98.5%, which shows the efficiency of this method.

6 Conclusion

In this paper, we combined different methods and different sources of data to construct a Chinese-Japanese lexicon. We basically joined two bilingual lexicons sharing a pivot language, English. The accuracy of the resulting Chinese-Japanese lexicon was improved by using three additional methods:

1. one time inverse consultation through the pivot language, English;

2. Japanese kanji to Chinese hanzi character conversion, using three different sources of data;

3. expansion through Chinese synonyms table.

The combination of these three additional methods produced the final translation candidates of our resultant lexicon. In total, we obtained 45,386 translation pairs. Among the 12,089 candidate pairs obtained using kanji/hanzi conversion (8,137) and synonyms table (3,952), 1,399 already existed in the table filtered by one time inverse consultation method. The kanji/hanzi conversion and synonyms table thus added 10,690 (12,089 - 1,399) candidates of very high quality. The three additional methods allowed us to increase the quality of our Chinese-Japanese lexicon from less than 45% to 85%. Table 3 shows an excerpt of our final lexicon.

Table 3: An excerpt of the final lexicon with indications on the origin of the word pairs and a final human assessment.

Chinese	Japanese	English meaning	classical joining	one time inverse	kani/hanzi	synonyms	human assessment
中央	中央	center	○	1.000	○		√
矿山	水雷	diggings / torpedo	○	1.000			×
构造	構造	construction	○	0.667	○		√
战果	戦果	results of battle	○	1.000	○		√
春季	はね	spring season / bounce	○	0.333			×
新闻	新聞	news / newspaper			○		×
古典音乐	クラシック音楽	classical music	○	1.000			√
作法	行動方針	course of action	○	0.400			√
美国人	アメリカ人	American person	○	0.667			√
核电站	原子力発電所	nuclear power plant	○	1.000			√
空白	空欄	blank space	○	1.000			√
恭贺新禧	あけおめ	Happy New Year	○	1.000			√
去年	昨年	last year	○	1.000			√
南部	南部	southern part	○	1.000	○		√
探	訪ねる	to visit	○	0.400			√
讲话	話す	to speak	○	0.400			√
丛林	森林	forest	○	1.000		○	√
中心	中央	center	○	0.667		○	√
不时	時時	frequently				○	√
生词	新語	new word				○	√

References

[1] K. Tanaka, K. Umemura. Construction of a bilingual dictionary intermediated by a third language. In *15th International Conference on Computational Linguistics:* COLING-94, pages 297-303, 1994.

[2] F. Bond, R.B. Sulong, T. Yamazaki, K. Ogura. Design and construction of a machine-tractable Japanese-Malay dictionary. In *Proc. of MT Summit VIII:* pages 53–58, 2001. 1979.

[3] Y. Zhang, Q. Ma, H. Isahara. Automatic acquisition of a Japanese-Chinese bilingual lexicon using English as an intermediary. In *Proc. of NLPKE*, pages 471–476, 2003.

[4] Y. Zhang, Q. Ma, H. Isahara. Use of kanji information in constructing a Japanese-Chinese bilingual lexicon. In *Proc. of ALR Workshop*, pages 42–49, 2004.

[5] C. Goh, M. Asahara, and Y. Matsumoto. Building a Japanese-Chinese dictionary using Kanji/Hanzi conversion. In *Natural Language Processing–IJCNLP*, pages 670–681, 2005.