2008－NL－188 (13)
2008／11／27

アラインメントに基づいた日中漢字の対応関係における解析

劉 瀟　綱川 隆司　岡崎 直観　辻井 潤一
東京大学大学院 情報理工学系研究科 コンピュータ科学専攻
〒113-0033 東京都文京区本郷 7-3-1
E-mail: {liuxiao, tuna, okazaki, tsujii}@is.s.u-tokyo.ac.jp

あらまし 本稿では、日本語漢字と中国語漢字の類似性を分析し、統計的機械翻訳(SMT)モデルに基づき、漢字間の対応関係を見つけ出すための確率モデルを提案する。このモデルを構築するために、我々は英語をピボット言語として、日英対訳辞書と中英対訳辞書をマージする。得られた日中対訳対に対して文字レベルのアラインメントをとることで、日中漢字の対応関係の確率を計算する。さらに、本手法で得られた漢字の対応関係を技術用語の日中翻訳に適用し、その有効性を報告する。

Analyzing Kanji-Hanzi Mappings by Aligning Translation Equivalents

Xiao Liu　Takashi Tsunakawa　Naoaki Okazaki　Jun'ichi Tsujii
Department of Computer Science
Graduate School of Information Science and Technology, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
E-mail: {liuxiao, tuna, okazaki, tsujii}@is.s.u-tokyo.ac.jp

Abstract: This paper analyzes the similarity between kanji and hanzi, and proposes a probabilistic model that associates kanji and hanzi characters by a Statistical Machine Translation (SMT) model. For constructing our probabilistic model, we first merge two existing bilingual dictionaries, the Japanese-English dictionary and the Chinese-English dictionary. Then based on an SMT model, we align the Japanese-Chinese translation pairs at character level and obtain the kanji-hanzi mapping probability. We apply our model in translating the Japanese and Chinese technical terms. The experiment results show that using the kanji-hanzi mapping improves the quality of Japanese-Chinese translations of technical terms.

1 Introduction

Bilingual lexicon is a fundamental resource for multilingual applications of natural language processing including machine translation (Brown et al., 1990) and cross-lingual information retrieval (Nie et al., 1999). A number of bilingual lexicons have been constructed manually despite their expensive compilation costs. Notwithstanding, a number of new technical terms are emerging daily; due to the difficulty in keeping up with the neologism in parallel corpora, multilingual applications suffer from the out-of-vocabulary problem. Thus, translating technical terms into other languages is a crucial and useful challenge in various natural language applications.

Meanwhile, machine translation between Japanese and Chinese has been attracting more and more attention recently. An interesting as-

CASE	KANJI Unicode	HANZI Unicode		Unicode Approach
		Simplified Chinese	Traditional Chinese	
1	体(body) 4F53	体 (body) 4F53	体 (body) 4F53	○
2	娘(daughter) 5A18	娘(mother) 5A18	娘(mother) 5A18	×
	走(run) 8D70	走(walk) 8D70	走(walk) 8D70	
3	無(none) 7102	无(none) 65E0	無(none) 7102	△
4	湯(hot water) 6E6F	汤(soup) 6C64	湯(soup) 6E6F	×
5	姫(princess) 59EB	姬(princess) 59EC	姬(princess) 59EC	×
6	浜(beach) 6D5C	滨(beach) 6EE8	濱(beach) 6FF1	×
	郷(home; country) 90F7	乡(home; country) 4E61	鄉(home; country) 9109	
	塩(salt) 5869	盐(salt) 76D0	鹽(salt) 9E7D	

Table 1. The kanji and hanzi with their Unicode

○: kanji can be mapped into hanzi via Unicode; ×: kanji cannot be mapped into hanzi via Unicode;
△: kanji can be mapped into Traditional Chinese via Unicode, and then convert hanzi between Traditional Chinese and Simplified Chinese

pect is that these languages use Han characters, each of which encodes semantic information as ideographic characters. Since Han characters were imported from Chinese to Japanese, strong correlations still exists between these languages. For this reason, Japanese and Chinese can sometimes guess the meaning of characters in the other language even though these two languages belong to different language families.

This characteristic between Japanese and Chinese is useful for various bilingual applications including bilingual text alignment (Tan and Nagao, 1995), cross-lingual information retrieval (Fredric, 2005), and bilingual dictionary construction (Zhang et al., 2004; Goh et al., 2005). The underlying idea of these studies is essentially the same: *Han characters mapped to identical Unicode characters mostly refer to the same concept in Japanese and Chinese.*

Han characters are stored in the CJK section in the Unicode system. Although this enables Han characters to be handled transparently regardless of languages, many problems still remain unsolved. Table 1 shows the relationships of Han characters in six cases. A naive approach in which kanji is mapped to hanzi with identical Unicode characters can only handle the Case 1. A more advanced approach may use the mappings between Traditional Chinese and Simplified Chinese to deal with Case 3, but may also increase translation errors as shown in Case 4. To make matters worse, the conventional approach cannot associate kanji and hanzi in other cases. This paper addresses the similarity and dissimilarity between kanji and hanzi characters from a bilingual dictionary. A large-scaled Japanese-Chinese dictionary is built by using English as a pivot language. We mine bilingual translation equivalents in the dictionary, and obtain a probabilistic model that associates kanji

and hanzi characters.

The rest of this paper is organized as follows. In Section 2, we present the methodology for analyzing the correspondences between kanji and hanzi. Section 3 reports our experiments and results in detail. After reviewing the related work in Section 4, we conclude the paper in Section 5.

2 Methodology

2.1 Building a large Japanese-Chinese dictionary via a pivot language

We would like to analyze kanji/hanzi correspondences from an actual corpus, for example, a parallel corpus between Japanese and Chinese texts. However, the size of existing parallel corpora between Japanese and Chinese is insufficient for statistical processing (Zhang et al., 2005). Moreover, it is difficult to find a correct word alignment in Japanese and Chinese sentences because syntaxes of these languages are dissimilar (e.g. different word order).

Therefore, we use a Japanese-Chinese bilingual dictionary consisting of technical terms in the science domain, for which Japanese terms are likely to be written in kanji. In this study, we first build a Japanese-Chinese dictionary by merging the two large-scaled dictionaries, JST Japanese-English dictionary[1] (527,206 translation equivalents) and the Wanfang Data Chinese-English dictionary[2] (525,259 translation equivalents). Here, English is used as a pivot language for merging two dictionaries (Tanaka and Umemura, 1994; Schafer and Yarowsky, 2002; Zhang et al., 2005). Then we apply word alignment technique in order to obtain kanji-hanzi associations.

Let L_{j-e} denote a Japanese-English bilingual lexicon and L_{c-e} a Chinese-English bilingual lexicon. We merge these two lexicons and construct a Japanese-Chinese bilingual lexicon as:

$$L_{j-e} = \{(\omega_j, \omega_c) \mid \exists \omega_e : (\omega_j, \omega_e) \in L_{j-e} \wedge (\omega_c, \omega_e) \in L_{c-e}\} \quad (1)$$

Here, ω_j, ω_c and ω_e denote Japanese, Chinese, and English lexicons in the corresponding dictionaries.

Since the English term forms originating from two dictionaries may be different, we apply Porter's stemming algorithm (Porter, 1980). In this way, we obtain a Japanese-Chinese dictionary consisting of 243,010 translation equivalents. However, the Japanese entries in this dictionary do not always consist of kanji, but include other types of characters such as hiragana and katakana, which are phonogramic characters of Japanese. Since it is impossible to associate kanji with hiragana and katakana, we remove translation equivalents containing hiragana and katakana characters from the dictionary. Finally, we obtain a Japanese-Chinese dictionary consisting of 62,380 translation equivalents. We use this dictionary for analyzing the correspondences between kanji and hanzi.

2.2 Analyzing kanji/hanzi correspondence by aligning Han characters

In order to obtain the kanji/hanzi correspondence, we apply the word alignment technique (Brown et al, 1991), assuming each Han character as a word. The translation probabilities of Han characters are calculated by applying the same technique to character alignment. The alignment model is based on IBM Model(Brown et al., 1993), which is implemented by GIZA++ toolkit (Och and Ney, 2003).

Suppose that we have Japanese-Chinese translation term pairs ($\mathbf{J}$, $\mathbf{C}$). In the IBM models, an alignment is represented by the vector $\mathbf{a}$, where $\mathbf{a}_k$ is a position of the corresponding character in the Chinese term $\mathbf{C}$ for the Japanese character $\mathbf{J}_k$. The translation probability between Japanese and Chinese terms is,

$$P(\mathbf{J}|\mathbf{C}) = \sum_{\sigma} P(\mathbf{J}, \mathbf{a}|\mathbf{C}) \quad (2)$$

[1] A Japanese-English technical term dictionary, which is edited and provided by Japan Science and Technology Agency

[2] A Chinese-English technical term dictionary (in Simplified Chinese), which is provided by Wanfang Data. Co. Ltd

In the IBM models, $P(\mathbf{J}, \mathbf{a}|\mathbf{C})$ is constructed from several factors including character translation probability, fertility model, and distortion model. The alignment $\mathbf{a}$ for the term pair can be calculated by the maximum likelihood estimation:

$$P(\mathbf{a}|\mathbf{J}, \mathbf{C}) = \frac{P(\mathbf{J}, \mathbf{a})}{\sum_{\mathbf{a}} P(\mathbf{J}, \mathbf{a}|\mathbf{C})} \quad (3)$$

IBM models initialize $\mathbf{a}$ and the models with a uniform distribution, and apply EM algorithm: calculate and iteratively update $P(\mathbf{J}, \mathbf{a}|\mathbf{C})$ by the IBM models from the term pairs using the previous parameters, obtain $P(\mathbf{a}|\mathbf{J}, \mathbf{C})$ with Eq. 3. An alignment of $(\mathbf{J}, \mathbf{C})$ is determined by specifying a_j for each position in the Japanese term. The fertilities, ϕ_0 through ϕ_l, are functions of a_j: ϕ_i is equal to the number of j for which a_j equals i. Therefore,

$$\begin{aligned} P(\mathbf{J}, \mathbf{a}|\mathbf{C}) = & \binom{m - \phi_0}{\phi_0} p_0^{m-2\phi_0} p_1^{\phi_0} \\ & \times \prod_{i=1}^{l} \phi_i! n(\phi_i|e_i) \\ & \times \prod_{k=1}^{m} p(j_i|c_{a_k}) d(j|a_k, m, l) \end{aligned} \quad (4)$$

The translation probabilities of characters $P(j|c)$ is estimated after the iteration finishes by,

$$P(j|c) = \lambda_e^{-1} \sum_{s=1}^{S} r(j|c; j^{(s)}, c^{(s)}), \quad (5)$$

$$r(j|c; j^{(s)}, c^{(s)}) = \sum_{\sigma} P(\mathbf{a}|c, j) \sum_{k=1}^{m} \delta(j, j_k) \delta(c, c_k). \quad (6)$$

Here, δ is the Kronecker delta function, equal to one when both of its arguments are the same and zero otherwise.

3 Experiments

3.1 Accuracy of kanji-hanzi mapping

In the first part of the experiments, Japanese terms were translated into Chinese terms by converting kanji into their corresponding hanzi. We used about 90% of the entries in the Japanese-Chinese technical term dictionary constructed in Section 2.1 as a training set, and obtained the kanji-hanzi probabilistic model by using GIZA++ toolkit. The test set was randomly-sampled 500 Japanese technical terms in the Japanese-Chinese dictionary, which are not included in the training set. In this experiment, every character in a Japanese technical term J was translated into a Chinese term C^* by the kanji-hanzi probabilistic model.

System	*Accuracy*
Unicode	16.6% (85/500)
Unicode+Big5	26.6% (133/500)
Proposed method	56.4% (282/500)

Table 2: Accuracy of kanji-hanzi mapping

$$C^* = \underset{C}{\operatorname{argmax}} \prod_{k=1}^{|J|} P(C_k|J_k) \quad (7)$$

We asked a human evaluator to check the correctness of the translation results in the test set. We prepared two baseline systems for comparison: *Unicode*: kanji are mapped into Simplified Chinese directly via UTF-16; *Unicode+Big5*: kanji are mapped into Traditional Chinese via UTF-16, and converted to Simplified Chinese by a Chinese encoding converter[3]. *Unicode* handles the case 1 in Table 1, and *Unicode+Big5* handles the case 3 in the table.

Table 2 reports the improvement of the character-encoding conversions. Although *Unicode* baseline could translate only 16.6% of the source terms correctly, the kanji-hanzi mapping achieved 56.4% accuracy.

3.2 Translating technical terms

We also examined the usefulness of the kanji-hanzi mapping for improving the performance of an actual Chinese-Japanese translation system. We used a Chinese-Japanese phrase-based SMT system for technical terms (Tsunakawa et.al., 2008). The system builds a translation probability table between Japanese and Chinese phrases by using Chinese-English and Japanese-English

[3]http://www.hao123.com/haoserver/jianfanzh.htm

	BLEU	NIST	Acc.
Tsunakawa et.al. (2008)	0.4361	6.8757	0.382
+kanji/hanzi feature	0.4614	7.0689	0.404

Table 3: Performance of translating technical terms

bilingual lexicons as parallel corpora. In addition to relative frequencies between translation pairs of Chinese and Japanese phrases, we added kanji-hanzi mapping scores to the system,

1. Calculate a Viterbi alignment of the characters of the phrase pair according to the kanji-hanzi translation probabilities.

2. Compute the kanji-hanzi mapping score by,

$$s(J|C) = \left(\prod_i P(J_{a_i}|\mathrm{C}_i) \right)^{\frac{1}{L}}, \quad (8)$$

$$L = \frac{\mathrm{length}(J) + \mathrm{length}(C)}{2}. \quad (9)$$

   Here, a represents the Viterbi alignment, and length(x) is the number of characters in x.

3. For unseen kanji or hanzi, we assigned the kanji-hanzi translation probability to 1 if the two characters are identical or if the hanzi is a simplified form of the kanji, and 0 otherwise.

Development and test sets consist of 2,000 term pairs with only kanji or hanzi characters. We sent Chinese technical terms into Tsunakawa et al.'s system, and obtained their Japanese translations. We evaluated the performance by using BLEU, NIST, and accuracy measures. Table 3 shows the evaluation results on the test set. Table 4 displays the translation examples.

Chinese term	Baseline		kanji-hanzi feature	
智力年龄	知能年令	×	知能年齢	○
母体遗传	母親遺伝	×	母系遺伝	○
小动脉	動脈	×	小動脈	○
理疗学会	理学療法協会	×	理学療法学会	○

Table 4. Example of translating technical terms

×: wrong translation ○: correct translation

4 Related work

Researchers have drawn attention to the relationship between kanji and hanzi. Hasan and Matsumoto (2000) presented an approach of indexing and retrieving Japanese and Chinese information represented in Han characters (Kanji). The researchers also proposed a dimensionality reduction technique for Kanji vector space, which can facilitate the conceptual retrieval of both mono- and cross- language information for these languages.

Zhang et al. (2003) used English as a pivot language and kanji-hanzi information for constructing the Japanese-Chinese bilingual lexicon. The researchers proposed a scoring function to estimate the likelihood in which a Chinese term is a translation of a given Japanese term. The scoring function combines three similarity metrics between Japanese and Chinese terms: the similarity between English translation for the Japanese and Chinese terms; similarity among part-of-speech codes; and the Levenshtein distance (Levenshtein, 1965). Zhang and Isahara (2004) also extend this approach by using the Web.

Chooi-Ling et al. (2005) also explored an approach for building a Japanese-Chinese bilingual lexicon by using English as a pivot language. They surveyed a Japanese lexicon, and found that about two thirds of Japanese nouns were written in kanji letters. More than one third of the Japanese nouns were successfully (97% accuracy) translated to Chinese nouns by using the simple character-encoding conversion via Traditional Chinese. In addition, the researchers could obtain translation candidates for 24% of Japanese words by using English as a pivot language (77% accuracy).

Our work also assumes that Japanese terms can be translated into Chinese terms by converting Japanese kanji letters into their corresponding Chinese hanzi. The previous work modeled the associations between Japanese kanji and Chinese hanzi with character-encoding conversions. In contrast, this study constructs a probabilistic model to translate a kanji character into

its corresponding hanzi characters. The probabilistic model is built from existing parallel corpora, without any assumption about character encodings. Thus, the model can naturally handle the complicated kanji/hanzi associations described in Section 1.

5 Conclusion

This paper discussed the similarity between kanji and hanzi. The kanji-hanzi mapping was obtained from Japanese-English and Chinese-English bilingual dictionaries by using English as an intermediary. The experimental results indicated that aligning translation equivalents could improve the kanji and hanzi mapping and enhance a translation system for technical terms.

The future work would be to explore the use of coarse corpora such as Wikipedia interlanguage links, parenthetical expression for obtaining kanji-hanzi mapping with wide coverage. We also plan to apply the proposed method to other applications such as cognate identification and transliteration.

Acknowledgements

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Japanese/Chinese Machine Translation Project in Special Coordination Funds for Promoting Science and Technology (MEXT, Japan).

References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Laf-ferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsu-moto 2005. Building a Japanese-Chinese dictionary using kanji/hanzi conversion In *Proc. of the 2nd International Joint Conference on Natural Language Processing*, pages 670-681.

Md. Maruf Hasan, and Yuji Matsumoto. 2000. Japanese-Chinese Cross-Language Information Retrieval: An Interlingua Approach. In *Computational Linguis-tics and Chinese Language Processing.* 5(2): 59-86

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* pages 48-54

Levenshtein, V.I. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady. Akademii Nauk SSSR.* 163(4): 845-848

Franz Josef Och, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29(1).* pages 19-51

M.F. Porter. 1980. An algorithm for suffix stripping. *Program.* 14(3): 130-137

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

Takashi Tsunakawa, Naoaki Okazaki, Jun'ichi Tsujii. 2008. Building a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language. *Computational Linguistics*, pages 127-130

Masao Utiyama and Hitoshi Isahara. 2007. A compari-son of pivot methods for phrase-based statistical ma-chine translation. In *Proc. of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics.* pages 484-491

Yujie Zhang, Qing Ma, and Hitoshi Isahara. 2003. Automatic acquisition of a Japanese-Chinese bilingual lexicon using English as an intermediary. In *Natural Language Processing and Knowledge Engineering.* pages 471-476

Yujie Zhang, Qing Ma, and Hitoshi Isahara. 2004. Use of Kanji Information in Constructing a Japanese-Chinese Bilingual Lexicon. In *Proc of the 4th Work-shop On Asian Language Resources.* pages 39-46