Research on Statistical Model of Japanese-Chinese Name Translation for Japanese Kana

Xiaoling Yu1, *, Xin Liu2, a
1The School of Foreign Languages, Dalian Polytechnic University, Liaoning, China
2Sony Global Information System (China) Co., Ltd., Liaoning, China

*Corresponding author e-mail: ring-0@163.com, aseven_369@126.com

Abstract. In view of the lack of bilingual resources in Japanese-Chinese translation, and the naming entity of Chinese character naming entities, especially Japanese pseudonyms, it is difficult to use, so the Japanese-Chinese pure pseudonym translation is a statistical translation model. The model has been applied so far and there have been many problems. The first is limited by the size and quality of the parallel corpus, and the second is the lower accuracy of the translation. Based on this, the paper proposes a statistical model of Japanese pseudonymous Japanese name translation based on inductive learning method. This method can extract Japanese and Chinese named entities from Japanese and Chinese corpora, and convert them into Roman alphabet and Pinyin sequence. Example screening after similarity calculation. Then, the instance-based induction learning method is used to automatically obtain the Japanese-Chinese transliteration rule base of the named entity, and iteratively reconstructs the transliteration rule base through feedback learning. Through experimental research, it is found that the Japanese-Han name translation statistical model method for Japanese kana is simple and efficient, and the translation accuracy is overcome while overcoming the dependence on bilingual data.

1. Introduction
A named entity is a word or phrase that identifies a particular entity. It mainly includes names of people, places, and organizations. It is an important information carrier for natural language. It is a research field in machine translation, information retrieval, question and answer systems, and cross-language information processing [1]. It is vital. At present, the research shows that there are several methods to carry out Japanese-Japanese name translation: First, the machine translation system is used for direct translation, that is, the entity is named by the known source language, and the named entity corresponding to the target language is directly obtained through translation. Second, given the named entity of the source language, the named entity corresponding to the target language is obtained by the method of network mining auxiliary translation. This method is an extension of the literal translation method. Thirdly, the equivalent translation of the named entity translation is extracted from the parallel corpus or the comparable corpus, but the method has a large dependence on the parallel corpus, and the large-scale bilingual resources are relatively scarce and the construction cost is high. Fourthly, using the Chinese character comparison table and the inductive learning method to extract the equivalence pairs of named entities from the monolingual corpus, such methods are simple and
efficient for the extraction of equivalent pairs of Japanese and Chinese naming entities, and effectively solve the Japanese-Chinese bilingual resources. Dependence. However, this method has certain limitations on the extraction of equivalence pairs of Japanese and Japanese entities with pure Japanese pseudonyms.

Based on this, this paper proposes a method based on monolingual corpus for the Japanese and Chinese names translation equivalence pair automatic extraction method for Japanese kana. Firstly, the method uses the conditional random field model to extract Japanese and Chinese names from Japanese and Chinese corpora. Then, the instance-based induction learning method is used to automatically obtain the Japanese-Chinese transliteration rule base of the name entity, and iteratively reconstruct through feedback learning [2]. Transliteration rule base. Then, the transliteration rule base is used to calculate the similarity between the Japanese and Japanese names, and the threshold is used to determine the equivalent of the name entity translation. The experimental results show that the proposed method is simple and efficient, and the equivalent value of the pseudonym name translation is high, which can reach more than 86%. The method overcomes the dependence of traditional methods on bilingual resources while achieving high precision of the system.

2. Overview of translation equivalence on automatic extraction methods

2.1. Maximum Entropy Model

In information theory, entropy represents a measure of uncertainty. The formula for information entropy is as follows:

\[
H(x) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = - \sum_{i=1}^{n} p_i \log p_i
\]  

(1)

Where \( x \) denotes a random variable, corresponding to a set of all possible outputs, defined as a symbol set, and the output of the random variable is denoted by \( x \). \( P(x) \) represents the output probability function. The greater the uncertainty of the variable, the greater the entropy. The maximum entropy model is shown in Figure 1 below.

![Figure 1. Maximum Entropy Model.](image-url)

Suppose the problem is that the context information is \( x \). For different contexts, we should apply the statistical model to correctly calculate the classification. Assuming category \( y \in Y \), the statistical model should be able to calculate the \( P(y|x) \) conditional probability. For the maximum entropy model,
the entropy in $\rho$ is maximized under certain constraints. At this point, the formalized formula (2) of model $P$ is:

$$p(y|x) = \frac{\exp\left(\sum_{i=1}^{k} \lambda_i f_i(x, y)\right)}{\sum_{y \in Y} \exp\left(\sum_{i=1}^{k} \lambda_i f_i(x, y)\right)}$$

(2)

Where $f_i(x, y)$ is a feature function of context two, indicating a certain context feature that appears together with category $y$, then the $f$ set is the constraint set. $\lambda_i$ is the corresponding feature $f_i(x, y)$ weight value, indicating the weight of this feature. The maximum entropy model is widely used in the field of natural language processing and has excellent performance. In the Conll-2016 named entity recognition evaluation, the first three ranking teams mentioned this method [3].

2.2. Conditional Random Field Model

Conditional Random Fields (CRFs) are conditional probability models for sequence data annotation. The conditional random field model is applied in the sequence labeling problem in recent years, and it is also the best one [4]. It does not have the strict independence assumptions of the hidden Markov model and thus can accommodate arbitrary context information. At the same time, the CRFs calculate the conditional probability of the global optimal output node, which overcomes the shortcomings of the mark bias inherent in the maximum entropy Markov model and other non-generated directed graph models. CRFs calculate the joint probability distribution of the entire marker sequence given the sequence of observations that need to be labeled, rather than defining the state distribution of the next state given the current state. Figure 2 shows the conditional random field model.

![Figure 2. Conditional random field model.](image)

Assuming that $O = o_1, o_2, ..., o_n$ is an observation sequence of length $n$ and parameter $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_{k}\}$ of the chain CRFs, the conditional probability of the output word sequence $S = s_1, s_2, ..., s_T$ of this model is formula (3):

$$P_{\lambda}(S|O) = \frac{1}{Z_{\lambda}} \exp\left(\sum_{i=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(S_{i-1}, s_i, o, t)\right)$$

(3)
Where \( Z_o \) is the normalization factor, which is to ensure that the conditional probability sum of all possible lexical marker sequences is 1, which is defined as equation (4):

\[
Z_o = \sum_{s} \exp \left( \sum_{i=1}^{T} \sum_{k=1}^{K} \lambda_k f_k (S_{i-1}, S_i, o, t) \right)
\]  

(4)

\( f \) in formula (3) is usually a binary representation function for expressing possible language features of the context. The definition formula is:

\[
f_k (s_{i-1}, s_i, o, t) = \begin{cases} 
1 & \text{If the condition is met} \\
0 & \text{otherwise}
\end{cases}
\]  

(5)

The CRFs model is capable of integrating any feature through the eigenfunction, including the sequence characteristics of the sequence of observable sequence \( O \) consisting of the current word and its context at time \( t \), and the transfer characteristics \( s_{i-1} \rightarrow s_i, \lambda_k \) of the implied lexical position in the context. \( \lambda_k \) is a parameter that needs to be learned from the training corpus during the training process, and represents the weight of the corresponding feature function \( f_k (s_{i-1}, s_i, o, t) \), which can range from \(-\infty\) to \(+\infty\). For a conditional random field model given by equation (3), for any input string, the most likely sequence of tokens can be found by the following formula (6):

\[
S^* = \arg \max_{S} P_{\Delta}(S|O)
\]  

(6)

Equation (6) can be decoded using the Viterbi algorithm to find the sequence of markers that maximize \( P_{\Delta}(S|O) \).

2.3. Chinese named entity extraction

The extraction of Chinese named entities mainly has two processes, namely Chinese word segmentation and Chinese named entity extraction. First of all, we use the open source Chinese word segmentation tool ICTCLAS developed by the Institute of Computing Technology of the Chinese Academy of Sciences to process the corpus. The correct word segmentation rate of the system is as high as 97.58%, which is the most popular Chinese word segmentation system in the world. For each word in the corpus, its annotations include the categorization, part of speech, and the category of the named entity. Next, we use the CRFs-based monolingual named entity recognition tool developed by the laboratory to extract the named entity from the corpus. The PRF method was used to evaluate the character recognition performance of the character. The results are shown in Table 1:

| Category | P (%) | R (%) | F1 (%) |
|----------|------|------|-------|
| Name     | 95.5 | 92.5 | 94    |

2.4. Japanese Named Entity Extraction

Similar to the extraction process of Chinese named entities, the extraction of Japanese named entities first needs to be segmented for the corpus. We use the Japanese word segmentation system Mecab developed by Kudo Takuya of Nara Prefecture Science and Technology College to process the corpus, and the basic policy of its design. It does not rely on specific language, dictionary, corpus, and (CRF) model for parameter estimation, and its performance is better than ChaSen using hidden horse model. At the same time, the average resolution speed is higher than the Japanese lexical analyzers ChaSen,
Juman, KAKASI. For each word, its annotations include the category of word segmentation, part of speech, and the name of the named entity. We use the CRFs-based monolingual named entity recognition tool developed by the laboratory to extract Japanese named entities from the corpus. The PRF method was used to evaluate the recognition performance of the person's name. The results are shown in Table 2.

Table 2. Japanese person name entity recognition result

| Category | P (%) | R (%) | F1 (%) |
|----------|-------|-------|--------|
| Name     | 88.2  | 85.7  | 87.0   |

3. Nickname naming entity labelling method

3.1. Basic concepts and solutions

In Japanese, pseudonyms are often used as a product of pinyin or transliteration. Machine transliteration includes transliteration of proper nouns such as names of people, places, institutions, and organizations. It is an important issue in natural language processing, in machine translation, cross-language information retrieval, etc. It has a very important role in the application. However, there is currently no suitable method for transliteration between Chinese and Japanese bilinguals. We refer to the mode and method of transliteration in Chinese-English machine translation, and propose a new Japanese kana and Chinese transliteration method.

3.2. Inductive learning

Japanese kana is a foreign word, mostly obtained by transliteration, and its corresponding Chinese name is also a foreign word, which is also translated by transliteration. Therefore, from the pronunciation law, there is a certain correspondence between each other. In order to explore and discover the law, the Chinese characters and Japanese kana characters corresponding to the pseudonym name can be converted into corresponding Chinese pinyin and Roman characters respectively. For example, "Luis Enrique" and "ルイスエンリケ" are converted to "lu | yi | si" and "ru | i | s | e | n | ri | ke", here we will use Chinese Pinyin to segment each Chinese character, and the Japanese pseudonym corresponds to the syllable of the Roman word. By segmenting words, we can easily get such rule pairs by analyzing, "Lu ~ ru", "yi ~ i", "si ~ su", "en ~ e | n", "li ~ ri", "ke ~ ke", as shown in Table 3, the value of "1" is the rule pair of Chinese Pinyin and Kana Roman. We hope to get more of these rule pairs in more Chinese and Japanese pseudonym names to identify the equivalent pairs of Chinese and Japanese kana names that we do not know.

Table 3. Correspondence table between Chinese pinyin and kana roman

| lu | yi | si | en | li | ke |
|----|----|----|----|----|----|
| ru | 1  |    |    |    |    |
| i  |    |    |    |    |    |
| su | 1  |    |    |    |    |
| e  |    |    |    |    |    |
| n  |    |    |    |    |    |
| li |    |    |    |    |    |
| ke | 1  |    |    |    |    |

We convert the extracted Japanese and Japanese pseudonym names into Roman syllable sequences and Chinese Pinyin sequences. For the name of the pseudonym from transliteration, its Chinese pronunciation and Japanese pronunciation are sequential, and there will be no reverse order. In order to improve the efficiency of the inductive learning method, this paper uses a certain scale of the pseudonymous name translation equivalence pair as the learning data to obtain the initial set of
candidate Chinese pinyin and Roman syllable rule pairs, and then use the weight according to the acquisition rule. The threshold-filtering part of the low-confidence rule obtains a high-confidence Chinese Pinyin and Roman syllable rule table. After that, according to the similarity calculation, the entity equivalence pair is obtained, and then the correction processing and the feedback learning are performed, and a new transliteration is generated through iteration. Rules and update the weight of the rules. Table 4 shows an example of a rule base.

Table 4. Rule base instance

| Pinyin sequence | Roman syllable sequence | Rule base |
|-----------------|-------------------------|-----------|
| lu | yi | si | en | li | ke | ru | i | su | e | n | ri | ke | lu-ru, ke-ke |
| yi | si | en | li | i | su | e | n | ri | lu-ru, ke-ke, yi-i, ri-li |
| si | en | su | e | n | lu-ru, ke-ke, yi-i, ri-li, si-su, en-e | n |

4. System experiment

4.1. Experimental materials

The monolingual corpus used in the experiment is derived from the Wikipedia database of Chinese and Japanese bilinguals. This experiment uses the CRFs-based named entity tools in Japanese monolingual chapters and Chinese monolingual chapters to identify 88203 Chinese names. The Japanese name was 73,322, and 13032 fake Japanese names were extracted from the Japanese entry, and the word alignment correction work was manually performed as the experimental data.

4.2. Parameter setting

The parameter setting in the experiment mainly refers to the setting of the threshold of the rule extraction. In the iterative process, the threshold should be gradually relaxed. Otherwise, as the number of iterations increases, it is difficult to obtain a new rule. However, at the beginning, the threshold cannot be selected lower, otherwise the rule base will be over-redundant. In addition, for the threshold setting of similarity, we take the initial value of 0.3 in the experiment. As the iteration progresses, we make dynamic adjustments [5].

4.3. Evaluation method

In equation (7), a represents the number of categories in which the input categories are correctly classified; b represents the number of categories in which the input category is incorrectly classified; and c represents the exclusion of the input category from the correct category. The number; d indicates that the input category is correctly excluded from the correct category.

\[
\text{Recall} = \frac{a}{a+c} \times 100% \\
\text{Precision} = \frac{a}{a+c} \times 100% \\
F_\beta = \frac{(\beta^2+1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}
\]  

4.4. Analysis of results

In this paper, the paper proposes a new method for calculating the similarity of Chinese and Japanese named entities. The paper proposes a method to solve the problem of pseudonyms in Japanese named entities. The paper uses two experiments to verify the validity of the method. The first experiment compares the results of the extraction of the baseline system and the system named entity without the pseudonym processing module. The results are shown in Table 5.
Table 5. Experiment 1 results

|        | P (%) | R (%) | F1 (%) |
|--------|-------|-------|--------|
| Baseline system | 91.77 | 84.66 | 88.07  |
| Iteration number |       |       |        |
| 1      | 70.14 | 40.28 | 51.17  |
| 2      | 85.33 | 70.66 | 77.31  |
| 3      | 87.13 | 74.26 | 80.18  |
| 4      | 89.90 | 79.80 | 84.55  |
| 5      | 91.89 | 84.05 | 87.80  |
| 6      | 92.58 | 85.17 | 88.72  |

It can be seen from the above table that the use of this method to extract the equivalence pairs of named entities from unrelated bilingual monolingual corpora is satisfactory. As the number of iterations increases, the difference translation list gradually increases so that the similarity calculation results are sequentially incremented [6]. The experiment verifies that the proposed method is simple and effective. After careful analysis of the experimental results, it can be found that the method has a good alignment result for the Chinese characters in the Chinese and Japanese named entities, but when Japanese is a named entity with full pseudonyms and is not associated with the existing fragment translation rules, the result is unrecognizable. Since there is no rule associated with it, the entry is completely unrecognizable and the similarity calculation result is 0 [7].

In the second experiment, the pseudonym processing module is added to the whole system, and compared with the extraction result not added to the module, it can be found that the method can effectively improve the recognition ability of the pseudonym part. The results are shown in Table 6.

Table 6. Results of experiment 2

|        | P (%)  | R (%)  | F1 (%) |
|--------|--------|--------|--------|
| 92.58+0.41 | 85.17+0.34 | 89.09  |

5. Conclusion
This paper proposes a method based on monolingual corpus for automatic extraction of Japanese and Chinese names translation for Japanese kana. Firstly, the method uses the conditional random field model to extract Japanese and Chinese names from Japanese and Chinese corpora. Then, the instance-based induction learning method is used to automatically obtain the Japanese-Chinese transliteration rule base of the name entity, and iteratively reconstruct through feedback learning. Transliteration rule base. Using the transliteration rule base to calculate the similarity between Japanese and Japanese names, the equivalent translation of the name entity is determined. The experimental results show that the proposed method is simple and efficient, and overcomes the dependence of traditional methods on bilingual resources while achieving high precision.

Acknowledgments
This work was financially supported by Major research project of think tank of Dalian Academy of Social Sciences in 2019 (Item number 2019dlsky054) fund.

References
[1] Sung, S. E., Park, S. J., & Kim, K. Why not japanese names? reader response to character name translation. Neohelicon, 43 (1) (2016) 213-231.
[2] Poon, K. Y. Coproducing asia: locating japanese-chinese regional film and media; japanese cinema goes global: filmworkers’ journeys. Media Asia, 42 (3) (2016) 1-3.
[3] Asanocavanagh, Y. Being 'indecisive' in japanese: analysis of kana, darou ka and (n) janai ka. Studies in Language, 40 (1) (2016)63-92.
[4] Nishimura, J., & Kanakura, Y. [the c5 gene polymorphism in patients with pnh]. [Rinshō ketsueki] The Japanese journal of clinical hematology, 56 (2) (2015) 103-10.

[5] Laissue, P. Aetiological coding sequence variants in non-syndromic premature ovarian failure: from genetic linkage analysis to next generation sequencing. Molecular & Cellular Endocrinology, 411 (C) (2015) 243-257.

[6] Yu, J. Y., & Taeyou, K. A cross-cultural study of perceptions of medical tourism among chinese, japanese and korean tourists in korea. International Journal of Nursing Education Scholarship, 33 (1) (2015) 80-88.

[7] Yang, G., Zhang, J., Li, Q., Liu, J., Liu, L., & Yang, X., et al. The clinicopathological and prognostic features of chinese and japanese inpatients with lung cancer. Oncotarget, 7 (41) (2016) 67425-67434.