



|                  |                                                                                                                                                                               |
|------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Title            | An Approach for Chinese-Japanese Named Entity Equivalents Extraction Using Inductive Learning and Hanzi-Kanji Mapping Table                                                   |
| Author(s)        | Xu, JinAn; Chen, Yufeng; Ru, Kuang et al.                                                                                                                                     |
| Citation         | IEICE transactions on information and systems, E100D(8), 1882-1892<br><a href="https://doi.org/10.1587/transinf.2016EDP7425">https://doi.org/10.1587/transinf.2016EDP7425</a> |
| Issue Date       | 2017-08                                                                                                                                                                       |
| Doc URL          | <a href="https://hdl.handle.net/2115/67136">https://hdl.handle.net/2115/67136</a>                                                                                             |
| Rights           | Copyright ©2017 The Institute of Electronics, Information and Communication Engineers                                                                                         |
| Type             | journal article                                                                                                                                                               |
| File Information | E100.D_2016EDP7425.pdf                                                                                                                                                        |



## PAPER

# An Approach for Chinese-Japanese Named Entity Equivalents Extraction Using Inductive Learning and Hanzi-Kanji Mapping Table

JinAn XU<sup>†a)</sup>, Yufeng CHEN<sup>†</sup>, Kuang RU<sup>†</sup>, Yujie ZHANG<sup>†</sup>, *Nonmembers*, and Kenji ARAKI<sup>††</sup>, *Member*

**SUMMARY** Named Entity Translation Equivalents extraction plays a critical role in machine translation (MT) and cross language information retrieval (CLIR). Traditional methods are often based on large-scale parallel or comparable corpora. However, the applicability of these studies is constrained, mainly because of the scarcity of parallel corpora of the required scale, especially for language pairs of Chinese and Japanese. In this paper, we propose a method considering the characteristics of Chinese and Japanese to automatically extract the Chinese-Japanese Named Entity (NE) translation equivalents based on inductive learning (IL) from monolingual corpora. The method adopts the Chinese Hanzi and Japanese Kanji Mapping Table (HKMT) to calculate the similarity of the NE instances between Japanese and Chinese. Then, we use IL to obtain partial translation rules for NEs by extracting the different parts from high similarity NE instances in Chinese and Japanese. In the end, the feedback processing updates the Chinese and Japanese NE entity similarity and rule sets. Experimental results show that our simple, efficient method, which overcomes the insufficiency of the traditional methods, which are severely dependent on bilingual resource. Compared with other methods, our method combines the language features of Chinese and Japanese with IL for automatically extracting NE pairs. Our use of a weak correlation bilingual text sets and minimal additional knowledge to extract NE pairs effectively reduces the cost of building the corpus and the need for additional knowledge. Our method may help to build a large-scale Chinese-Japanese NE translation dictionary using monolingual corpora.

**key words:** *named entity translation equivalents acquisition, Chinese Hanzi and Japanese Kanji mapping table, inductive learning, monolingual corpora*

## 1. Introduction

NE generally refers to proper names and the meaningful quantifier appearing in texts. NEs were divided into seven categories in the Message Understanding Conference (MUC) [1], including: Person, Location, Organization, Date, Time, Percentage and Monetary value. Person, Location and Organization are the most important categories. NE described in this paper is a collection of these three categories.

The study of NEs is a fundamental task of research

in natural language processing (NLP). Mainstream methods of NE extraction include the rule-based approach, statistical approach, knowledge based approach, and hybrid methods combining rule and statistical models. In the research field of NLP, NE recognition (NER) belongs to the category of unknown word recognition in the lexical analysis. In real text segmentation, about ninety percent of the total unknown words are proper nouns. The remainder are common words or technical terminologies [2]. Recently, therefore, proper nouns or NE processing has become a developing area of particular interest. A significant reason is that the segmentation accuracy loss caused by unknown words is at least five times larger than that caused by word sense disambiguation [3], thus NE status ranks highly in NLP.

According to the definition of Automatic Content Extraction evaluation (ACE) plan, there are three types of entity mentions\*: name mention, nominal mention and pronoun mention. NE research tasks mainly include NER, extraction, disambiguation, abbreviation recognition, coreference resolution, NE normalization, attribute extraction and relation detection.

NER is used to settle the problem of identifying the NEs that locate and categorize important nouns and proper nouns in a text. NEs are words or phrases which are named in or categorized by a certain topic, for instance, news, legal, specialized fields of science or technology, etc. This categorization plays an important role in applications such as information retrieval (IR), question answering (QA), machine translation (MT) and so on.

NE disambiguation is used to determine the problems caused when NE points to multiple physical concepts and when multiple NE concepts are allocated as the same entity. For example: "Michael Jordan" is not only the name of an NBA basketball player, it may also be another person. Only by taking into consideration available specific context information would it be possible to map the name of the alleged ambiguity with the correct entity concept.

Attribute extraction of NEs refers to retrieving the category and attribute of a certain particular entity concept from web pages. For instance, for the basketball player Michael Jordan, extracted information includes details of his profession "basketball player", "place of birth" "Brooklyn, New York City", "Date of birth" "February 17, 1963", and so on.

Manuscript received October 15, 2016.

Manuscript revised March 9, 2017.

Manuscript publicized May 2, 2017.

<sup>†</sup>The authors are with Beijing Key Lab of Traffic Data Analysis and Mining, the School of Computer and Information Technology, Beijing Jiaotong University, Haidian District, Beijing, 100044 China.

<sup>††</sup>The author is with Hokkaido University, Sapporo-shi, 060-0814 Japan.

a) E-mail: jaxu@bjtu.edu.cn

DOI: 10.1587/transinf.2016EDP7425

\*<http://itl.nist.gov/iad/mig/tests/ace/>

The detection of an NE relationship is used to judge what relationships are between two entities through the analysis of network information, and this detection and association aims to determine relationships between pairs of NEs in text. For example, person names may be related to organization entities via organization affiliation relationships. Organization entities may be related to location entities via physical location relationships and persons may be related to one another via social relationships.

Mapping NE translation equivalents to the same entity concept, for instance, “迈克尔·乔丹” and “Michael Jordan” involves the corresponding relationship between Chinese and English, also known as cross language association of named entities. With the development of NE extraction technology, one main task of NE extraction has been to automatically construct NE pairs to improve the performance of multilingual information processing.

Many approaches have been proposed to tackle NE translation or transliteration. For instance, concerning person names, researchers found that they have a certain corresponding phonetic relationship between English and Chinese, for example, “贝拉克·奥巴马/Barack Obama”. The Pinyin sequence of Chinese names is typically their English translation, for instance, “温家宝/Wen Jiabao”. For NEs of location and organization, some bilingual entities are similar in pronunciation, for example, “华盛顿/Washington”; some entities are translated in the same semantic, for example, “海关总署/General Administration of Customs”; and some entities are translated based on a combination of transliteration and semantics, for example, “卡内基梅隆大学/Carnegie Mellon University”.

Moreover, many researchers have proposed some multilingual NE extraction methods for improving the performance of MT and CLIR. We summarize recent work in this area and describe several open research problems in Sect. 2.

In summary, traditional methods require large-scale parallel or comparable resources. Such bilingual resources are relatively scarce, and incur a high cost of construction. In contrast, large-scale monolingual corpus construction is simple, low cost and easy to implement.

At present, Chinese-Japanese bilingual resources are inadequate. In constructing a Chinese-Japanese NE dictionary, in order to solve this problem, we propose a method based on IL using Chinese and Japanese monolingual corpora for automatic acquiring NE pairs. As learning examples, we adopt Chinese *Hanzi* and Japanese *Kanji* mapping tables to calculate the similarity between Chinese and Japanese NE instances. IL is then used to extract common and different parts of these tables of highly complementary Chinese-Japanese NE pairs to obtain partial translation rules. Feedback processing is used to systematically update the similarity of the Chinese-Japanese NE pairs and the acquired partial translation rule set. Experimental results show that our proposed method is simple and efficient, and can effectively use monolingual corpora to build a large-scale Chinese Japanese NE translation dictionary.

In our previous works, we have presented some results

of our research [4]. This paper presents some improvements of our study. The remainder of this paper is structured as follows: Sect. 2 introduces related work, previous research and resources, focusing on the IL and Japanese-Chinese character mapping table. The outline of our method is described in Sect. 3. Section 4 discusses and analyzes the experimental results. Finally, we conclude our study and discuss future research directions.

## 2. Related Work and Resources

### 2.1 Traditional Multilingual NE Extraction Methods

Traditional methods of multilingual NE extraction mainly include the dictionary-based method [5], the rule-based method [6], and the statistical transliteration model [7]. Most of these approaches focus on the languages pairs of English-Chinese, English-Japanese and others. In contrast, the literature on Chinese-Japanese NE pair acquisition is sparse. However, this task is more difficult than other language pairs; word alignments between Chinese and Japanese are also more complex because of the many lexical, grammatical and syntactical differences between Chinese and Japanese.

There are three main methods used to extract bilingual NE pairs:

1. Given source language named entities, translate directly based on MT;

For English-Chinese NE translation, Zou et al. [8] have shown systematic comparative experiments with different learning models, including memory-based learning, the bilingual N-gram based statistical machine translation model, the maximum entropy (ME) model, the conditional random fields (CRFs) model and the phrase-based SMT model. The results show that the performance of each model is unsatisfactory, and the functional differences are not obvious. However, the translated results maintain a high degree of coincidence, and most correct results appear in the list of candidates. The main problem is uncorrected selection due to the different error of probabilities of each method. This indicates that the simple use of a particular statistical method is not enough, and it is indispensable to combine and compare with other methods to obtain better performance.

Chen et al. [9] systematically studied the characteristics of Chinese organization names (ONs). They defined appropriate chunk-units for analyzing the internal structure of Chinese organization names, to make the chunk-based organization name translation feasible and flexible via a hierarchical context-free grammar (CFG) derivation. Their method further proposed a training architecture to automatically learn the synchronous CFG for constructing organization names with chunk-units from aligned bilingual ON pairs. The structure-based model includes a chunking model and a chunk-based CFG model with defined derivation. The CFG rules are classified into four types: “glue” rules, templates, common rules, and special rules according to their hierarchical derivation rank. Templates and common rules are

learned from bilingual ONs without any syntactically annotated training data; special rules, which are defined in CFG format based on word level, are applied to cover more organization name translation. This study shows the effectiveness and the improvement of the alignment acquiring performance between Chinese-English NEs [9].

2. Given source language NEs, use web mining technologies to assist MT to determine NE pairs;

In general, it is not easy to satisfactorily acquire NE pairs using machine translation or transliteration. Many researchers have adopted the web mining method to acquire the corresponding NE pairs. Some researchers tried to improve the performance of transliteration using mixed-language web-assisted translation methods [10]–[13]. For instance, Jiang et al. [14] proposed an approach to combine web mining and transliteration for NE translation, using web information as a source to complement transliteration, and using transliteration information to guide and enhance web mining. In this approach, a ME model is employed to rank translation candidates by combining pronunciation similarity and bilingual contextual co-occurrence. Experimental results show that the approach effectively improves the performance of NE translation [14]. Yang et al. [15] proposed a backward transliteration approach which can further assist the existing statistical model by mining monolingual web resources. It shows that this approach can achieve better performance.

3. To extract NE pairs from language corpus;

Early studies of NE pair extraction focused predominantly on obtaining NEs from parallel corpora [16]–[18]. Such methods usually combine a variety of features, including the features of characters of semantic translation, transliteration, NE annotation, and statistical characteristics, such as the co-occurrence rate, displacement length, etc.

Since word-based transliteration models often cannot achieve the required or expected performance, the bilingual dictionary translation methods remain the barrier to dictionary progress. Though many methods can attain respectable accuracy, it is expensive to construct large-scale parallel corpora, especially when the relevant language resources are deficient. This is, particularly the case for Japanese and Chinese parallel corpora.

To mitigate these problems, many research methods use comparable corpora to obtain bilingual resources. Comparable corpora are easily acquired, resources are abundant, and this method is therefore more suitable for large-scale NE pair extraction. For instance, Rapp [19] proposed a method using word co-occurrence. It assumes that if the joint probability of co-occurrence of two words is larger than the probability of occurrence of the individual word, then these two words are more likely to appear in a similar context. The method uses a similarity matrix and joint probability techniques to estimate the lexical mapping. Most of the existing methods are based on this assumption and put forward different similarity calculation methods. According to the similarity of their matrix contents they can be classified as shown in Table 1.

**Table 1** Classification of cross-language similarity metrics.

| Type                  | Entity | Relationship |
|-----------------------|--------|--------------|
| Using entity names    | E      | R            |
| Using textual context | EC     | RC           |

(1) *E*: Entity names in the two corpora can be compared based on their phonetic similarity, e.g., high phonetic similarity between *Obama* and the Chinese translation of “*Obama*, 奥巴马” (pronounced “Aobama”), is evidence that they are a translation pair.

(2) *EC*: Entity Context, i.e., common text surrounding the NE translation pair in the two corpora, can be compared. For example, “*Obama*” and “奥巴马”, surrounded by words of the same meaning, such as “president” and “总统”, supports the inference that they are translation pairs.

(3) *R*: Relationships between an NE pair in one corpus and an NE pair in other corpus, quantified by monolingual NE co-occurrence, can be used as relational evidence that they are a translation pair. For example, *Barack* and *Michelle* (巴拉克 和 米歇尔) frequently co-occur in both corpora. Together with the high translation similarity of (Barack, 巴拉克), this supports the inference that “Michelle” and “米歇尔” are translation pair.

(4) *RC*: Relationship Context, i.e., the surrounding text that explains the relationship of the NE pair in each corpus, is evidence for whether or not the two NE relationships are comparable. For example, *Barack* and *Michelle* described as a couple in both corpora is evidence that they have strong relationship, and, therefore, it can enhance the relationship similarity measure, *R*.

It is difficult to obtain satisfactory results in cases where only one feature is used to assess the similarity matrix [20]–[22]. Many researchers have tried to combine multiple features to improve accuracy and performance. For instance, Shao and Ng [23] integrated the *E* and *EC*. Lee et al. [24] brought together *E*, *EC* and You et al. [25] and Kim et al. [26] not only merged all the features together, but also incorporated the additional latent features into their studies.

## 2.2 Previous Studies Using Hanzi-Kanji Table

Since Chinese characters contain significant semantic information, Chinese Hanzi and Japanese Kanji mapping tables can be very useful for improving the performance of MT and CLIR. For instance, Chu et al. [27] proposed a method for creating a HKMT with the aim of constructing a complete resource of common Chinese characters. This study identified two main disadvantages in the case of Chinese word segmentation for Chinese-Japanese MT, namely, the problems of unknown words and word segmentation granularity. The study proposed a method to solve these problems which exploited common Chinese characters. They also proposed a statistical method to detect semantically equivalent Chinese characters other than the common ones, and detailed a method for exploiting shared Chinese characters in phrase alignments. This study shows that it is possible to improve the performance of a phrase-based SMT and an example-

based MT of Chinese-Japanese translation by using shared Chinese characters.

As existing Chinese-Japanese parallel corpora are insufficient, most studies focus on language pairs between English and these languages. To improve MT performance, Chu et al. [28] proposed an integrated system to extract both parallel sentences and fragments from comparable corpora. They applied parallel sentence extraction from comparable sentences, and then extracted parallel fragments from the comparable sentences. Parallel sentence extraction is based on a parallel sentence candidate filter and classifier. The study adopted a novel filtering strategy, three novel feature sets for classification, and a method of parallel fragment extraction by using an alignment model to locate the parallel fragment candidates. An accurate lexicon-based filter was used to identify the truly parallel fragments.

Goh et al. [29] presented a new bilingual dictionary constructed using two existing bilingual dictionaries. They used Japanese-English and English-Chinese dictionaries to build a Japanese-Chinese dictionary. This method attempted to build a dictionary for Kanji words by simple conversion from Kanji to Hanzi.

Zhang et al. [30] proposed a lexical knowledge-based approach for word alignment extraction, which consists of two algorithms. The first is used to obtain reliable alignments by using three types of heuristics: 1) orthography, 2) the correspondence between the simplified Chinese characters and 3) the traditional Chinese characters, and an automatically built Japanese-Chinese dictionary. The second is designed to broaden coverage by estimating the dislocation of a candidate from the established reliable alignments.

Tsunakawa et al. [31] proposed an integrated framework for building a bilingual lexicon between Chinese and Japanese. This research built a Chinese-Japanese bilingual lexicon dictionary through English, taking English as a kind of pivot language. The research also integrated a Hanzi-Kanji table and a small-scale Chinese-Japanese lexicon dictionary into the SMT system. The results show that the similarity between the Hanzi and Kanji characters had a positive effect on translating technical terms.

Dabre et al. [32] presented a large-scale dictionary construction method via pivot-based SMT with significance pruning, Chinese character knowledge and re-ranking of bilingual neural network language model based features. Large-scale Japanese-Chinese experiments show that this method is quite effective.

Hasan et al. [33] proposed a Kanji oriented Interlingua model for indexing and retrieving Japanese and Chinese information which exploits the high co-occurrence of Kanji in Chinese and Japanese texts to improve the performance of Chinese and Japanese cross information retrieval.

Lin et al. [34] adopted a so-called “query-translation” approach to improve the performance of Japanese-Chinese CLIR. This study integrated a kind of Japanese-Chinese bilingual dictionary and the web-based resource Wikipedia into its framework.

In this paper, we focus on the task of NE pairs extrac-

**Table 2** Instance extraction from unknown string pairs.

|           |                                           |
|-----------|-------------------------------------------|
| Input 1   | $\alpha\theta\sigma\psi\delta\lambda\nu$  |
| Input 2   | $\Xi\Sigma\psi\delta^*\Upsilon\Phi\Theta$ |
| Segment 1 | $\alpha\theta\sigma$ $\Xi\Sigma$          |
| Segment 2 | $\psi\delta$ $\psi\delta$                 |
| Segment 3 | $\lambda\nu$ $\Upsilon\Phi\Theta$         |

**Table 3** Extraction primitive from segments.

|             |                         |                     |
|-------------|-------------------------|---------------------|
| Segment 1   | $\alpha\theta\sigma$    | $\Upsilon\Xi\Theta$ |
| Segment 2   | $\theta\sigma\gamma\mu$ | $\Xi\Theta\Sigma$   |
| Primitive 1 | $\alpha$                | $\Upsilon$          |
| Primitive 2 | $\theta\sigma$          | $\Xi\Theta$         |
| Primitive 3 | $\gamma\mu$             | $\Sigma$            |

tion of Chinese and Japanese words and characters using IL and HKMT.

### 2.3 Inductive Learning

The example-based IL method was proposed by Araki et al. [35]. The basic idea includes two main aspects. One is the processing of recursive rule extraction of common and different parts from two examples. The other, screens and optimizes extracted rules via feedback processing. We use this approach in our study. The justification is twofold: Firstly, this approach enables the similarity of instances to be calculated for example selection; secondly, correspondence rules can be extracted from unknown strings, as shown in Table 2.

Corresponding relationships between *Input 1* and *Input 2* in Table 2 are denoted with an underline ( $\underline{\psi\delta}$ ). Subsequently, the different parts are aligned according to the order of segments. The extraction processing is consistent with the basic assumption of IL that the native capability is to judge whether two things are the same or not [35]. The results are shown in Table 2. *Segment 1*, *Segment 2* and *Segment 3* constitute corresponding relationships. In addition to the sequences shown in Table 2, the corresponding reverse or crossover mapping relationships may also exist. As to the position order or reverse or crossover, this will depend on the specific research goals and empirical method used to resolve the relationships.

According to the same idea, we extract the common part from the segments and reduce the segments to primitives. An example of the method for extracting primitives from the segments is shown in Table 3. We extract the common part marked with an underline ( $\underline{\theta\sigma}$ ,  $\underline{\Xi\Theta}$ ) from *Segment 1*, *Segment 2* as *Primitive 2*. We then extract the different parts of both sides ( $(\alpha, \Upsilon)$ ,  $(\gamma\mu, \Sigma)$ ) as *Primitive 1* and *Primitive 3*. Thus, by separating the common and different parts, we obtain three primitives [35].

As primitives can be merged to create segments, these three primitives became the perfect substitute for the two segments. This extraction method usually requires empirical rules to determine the correspondence between strings. This example-based method performs the extraction of the common and different parts by stages so as to obtain knowl-

edge, in accordance with IL methods [35].

## 2.4 Hanzi-Kanji Mapping Table (HKMT)

Chinese characters are in widespread use in NLP and information retrieval systems. Japanese Kanji comes from ancient Chinese. Therefore, Japanese Kanji and Chinese Hanzi (including Simplified Chinese and Traditional Chinese) are the same in many cases. However, as shown in Table 4, the relationships between Japanese Kanji and Chinese Hanzi can be very complicated. Some researchers adopt this table to improve the performance of Chinese and Japanese information processing. For example, Goh et al. [29] used a Japanese-Chinese dictionary to change the Japanese Kanji to Chinese Hanzi through direct matching method. Chu et al. [27] use open source resources to build a three-way Japanese Kanji, Traditional Chinese, and Simplified Chinese Hanzi table.

Many Chinese-Japanese translation system typically contain only simplified Chinese Hanzi. This paper also constructs Japanese Kanji and Simplified Chinese Hanzi tables. For the construction process, we adopt a total of three types of dictionary information:

1. Variants Dictionary. A Hanzi in a Chinese-Japanese dictionary may exist in a variety of different shapes while building the dictionary; we enumerate the circumstances of each variant. The UniHan Database is the CJK trilingual knowledge databases of the Unicode Consortium<sup>†</sup>. The database contains feature information regarding the variants which records the relationships in Japanese Kanji and Chinese Hanzi. In this paper, we use variants to adapt Japanese Kanji shape. If there is a link between variants, then two characters can be related and transformed into each other.

2. Chinese and Japanese Kanji Dictionary. We use the Chinese-Japanese Kanji dictionary of Kanconvit<sup>††</sup>, which contains a total of 1159 vocabulary variants.

3. Traditional-Simplified Chinese Dictionary. Traditional-Simplified Chinese Hanzi is not a simple one-to-one relationship as shown in Table 5. We use the Traditional-Simplified Chinese Dictionary in Chinese Encoding Converter<sup>†††</sup> which contains a total of 6740 pairs of traditional Hanzi to simplified Hanzi characters.

**Table 4** Hanzi-Kanji mapping table.

|                           |   |   |   |   |
|---------------------------|---|---|---|---|
| Japanese Kanji            | 愛 | 國 | 書 | 氷 |
| Traditional Chinese Hanzi | 愛 | 國 | 書 | 冰 |
| Simplified Chinese Hanzi  | 爱 | 国 | 书 | 冰 |

**Table 5** Hanzi converter standard conversion table.

|                           |      |   |      |   |
|---------------------------|------|---|------|---|
| Traditional Chinese Hanzi | 篇, 編 | 變 | 並, 併 | 佈 |
| Simplified Chinese Hanzi  | 篇    | 变 | 并    | 布 |

<sup>†</sup><http://unicode.org/charts/unihan.html>

<sup>††</sup><http://kanconvit.ta2o.net>

<sup>†††</sup><http://www.mandarinintools.com/zhcode.html>

## 3. Our Proposed Approach

Traditional methods for NE pair extraction from bilingual corpora require that the bilingual corpora have a strong correlation with each other. The methods using comparable corpora also require such correlation [36], [37]. The main problem with these methods lies in the fact that the resources of parallel or comparable corpora are usually limited and expensive. In contrast, large-scale monolingual corpora are easier to construct than bilingual corpora, and at lower cost. Therefore, our goal is to apply this method to large-scale monolingual corpora, with attempted generalization, to obtain a good performance of NE pair extraction and reduce the difficulty in constructing parallel or comparable corpora. As a result, we may be able to reduce the cost of extracting NE pairs.

The procedure of our proposed method is shown in Fig. 1. First, we run two monolingual collections of NE through the monolingual NER tools. Second, we use the NE translation rules and HKMT as additional knowledge to calculate the similarity between bilingual NEs to obtain a similarity list. Third, we use inductive learning to process high similarity NEs. For example, from “BBC ウェールズ” (BBC Wales) and “BBC 威尔士”, we get the different parts, which are “ウェールズ” and “威尔士” when we take the word as the minimal semantic unit. After we count up the different parts in NE list, we select the entries whose co-occurrence rate is higher than the given threshold, and we add them into partial translation rules. Finally we use partial translation rules to iterate the similarity until no new NE pairs are generated.

### 3.1 Monolingual NER

Monolingual NER technology has developed rapidly. The

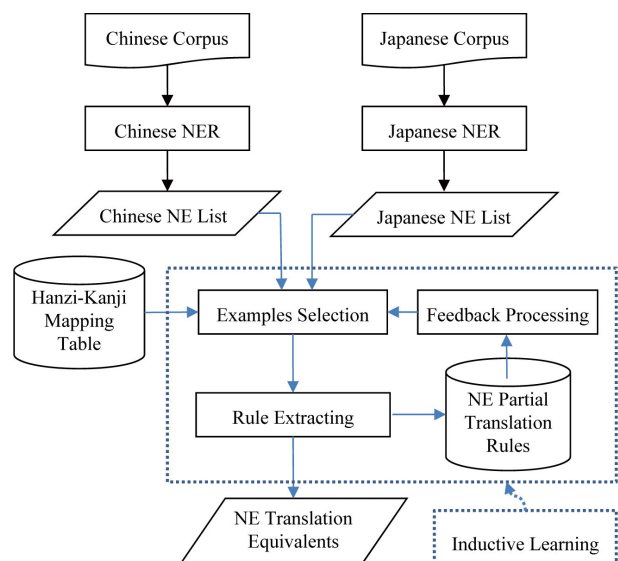


**Fig. 1** The procedure of our approach.

main methods include rule-based, knowledge-based and statistical methods. In general, when the extracted rules accurately reflect the language phenomena, the rule-based approach is superior to the performance of statistical methods. However, these rules are often dependent on the style of the specific text. Their compiling process is time-consuming, and it is difficult to cover the entire language phenomena. As these rules are prone to errors and bad system portability, linguistic experts have to rewrite them for each different domain.

Statistical methods use artificially annotated corpora for training. The methods do not require extensive linguistic knowledge, and a system can be constructed in a short time through such methods. At the CoNLL-2003 Conference, the 16 systems demonstrated by the participants were based on the statistical methods. The statistical methods have become typical in current research. When such systems are transferred to a new area, they simply adopt a new corpus for training. Statistical machine learning algorithms include: the Hidden Markov Model (HMM), Maximum Entropy (ME), the Support Vector Machine (SVM), Conditional Random Fields (CRFs) methods [32], and Neural Networks (NN).

Of the traditional learning methods, the ME model is the most compact and versatile. Its shortcoming lies in its high training complexity and exorbitant training costs. It requires clear, normalized calculations, with a large associated overhead. Alternatively, CRFs are more flexible and provide a global optimal annotation framework for NER in spite of their slow convergence and long training time.

In this paper, a monolingual NER tool is constructed using CRFs as a kind of undirected statistical graphical model, this is a statistical sequence modeling framework that corresponds with a conditionally trained finite-state machine. The concept comes from the ME model as a discriminant model. CRFs have recently shown empirical success in many research fields of natural language processing.

A Markov random field is expressed as a form of undirected graph. Each point on the undirected graph contains a random amount, and the edge between nodes represents a random variable corresponding to nodes with contextual dependency relationships. Therefore, the structure of the Markov random field is essentially considered to be a digram Markov chain. The Markov property refers to the distribution of the variable under all other given variables in the field for any random variable in the Markov random field, which is equal to the distribution of the variable under the given neighbor nodes of the variable. The Markov property can be regarded as the microscopic property of the Markov random field, and the macroscopic property is the form of the joint probability. CRFs are essentially treated as a Markov random field, and given the values of the observation sets to calculate the discriminant of the graph model of the conditional probability of the output node under the condition of a given input node (observation). The target of this method is to globally optimize the joint probability of the tags sequence under the condition of the given observed

sequence. The nodes represent the label sequence  $Y$  corresponding to the sequence of some observed input data  $O$ . The CRF model aims to acquire the label  $S$ , which maximizes the conditional probability  $p(S | O)$  for a sequence  $O$ . The CRF model can be expressed as Formula (1).

$$P_{\Lambda}(S|O) = \frac{1}{Z_o} \exp(\sum_{t=1}^T \sum_k \lambda_k f_k(S_{t-1}, S_t, o, t)) \quad (1)$$

where  $Z_o$  is a normalization term calculated by the ratio of factor over all state sequences,  $\lambda_k$  represents the weights assigned to the different features in the training phase, and  $f_k$  is the feature function over its arguments.

### 3.2 IL Processing

In the initial phase of this module, we use HKMT to calculate the similarity between Chinese and Japanese NE instances, and then obtain examples for IL through given appropriate threshold values. Using the partial translation rules, which are generated by IL module, we proceed to further improve the NE similarity.

There are many ways to calculate the similarity between NEs, such as edit distance, Hamming distance, cosine vector, and Jaccard similarity. Based on the word frequency information, we have chosen to use the cosine vector as the similarity formula. The basic idea is: If the two words or phrases used in NEs are more similar, their content shares more similarity. Therefore, we can start with word frequency, to calculate the similarity of Japanese and Chinese NE pairs by the formula:

$$Sim = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} = \frac{A \bullet B}{|A| \times |B|} \quad (2)$$

where  $Sim$  indicates the similarity, Japanese NEs expressed as  $A$ , and Chinese NEs express as  $B$ .  $A$  and  $B$  represent  $n$ -dimensional vectors over the term set  $T = \{t_1, \dots, t_n\}$ . Each dimension denotes a term by its weight in the document. As a result, the cosine similarity is non-negative and bounded between  $[0, 1]$ . When the cosine value is closer to 1, it shows that the angle is close to 0 degrees, which means the two vectors are more similar.

For instance, taking the Japanese NE “大原綜合病院附属大原医療センター” (Ohara subsidiary medical center of Ohara General Hospital) and the Chinese NE “大原综合医院附属大原医疗中心”, the processing steps are as follows:

1. Perform monolingual word segmentation for the Japanese and Chinese NEs.

Japanese Segmentation Results: 大原 綜合 病院 附属 大原 医療 センター

Chinese Segmentation Results: 大原 综合 医院 附属 大原 医疗 中心

2. Apply the NE Partial Translation rules to the translation of the NEs. In the initial state, the rule base is empty. After a few iterations of rule extraction, in this case, “病院 (hospital)” → “医院”, and “綜合 (general)” → “综合” have

been extracted as a partial rule, which can improve the similarity.

Japanese partial translation results: 大原 综合 医院 附属 大原 医療 センター

Chinese: 大原 综合 医院 附属 大原 医疗 中心

3. Convert the Japanese Kanji to Simplified Chinese Hanzi using Hanzi-Kanji mapping table. In this example, ‘療’ → ‘疗’

Japanese Kanji conversion results: 大原 综合 医院 附属 大原 医疗 センター

Chinese: 大原 综合 医院 附属 大原 医疗 中心

4. Unify segmentation granularity and take the Japanese word segmentation results as the standard. If unable to confirm the boundary, the unknown part retains the initial state of the Chinese word segmentation.

Japanese: 大原 综合 医院 附属 大原 医疗 センター

Chinese unified segmentation granularity results: 大原 综合 医院 附属 大原 医疗 中心

5. Form a vector containing all the words from the bilingual NE.

[大原 综合 医院 附属 医疗 センター 中心]

6. Determine the word frequency vector.

Japanese word frequency vector: [大原/2 综合/1 医院/1 附属/1 医疗/1 センター/1 中心/0]

Chinese word frequency vector: [大原/2 综合/1 医院/1 附属/1 医疗/1 センター/0 中心/1]

7. Calculate the similarity between the two vectors using Formula (2).

IL processing is recursive extraction of the common part and the different part of the two NE instances. The pseudo code is shown as follows:

*Algorithm: Inductive Learning*

*INPUT* :  $J\_C\_dict$ , *Similarity*, *Part\_Trans\_Rules*

*INITIALIZATION*: *Diff\_count*

*For entity in Similarity*:

*If* *entity.value* > *Threshold.diff*:

    #Get the same and different part in NEs.

*content* = *GetDiff*(*entity.ja*, *entity.ch*)

*Diff\_count*[*content*] ++

*For content in Diff\_count*:

    #If rule frequency higher than threshold,

    #add it into partial translation rules.

*If* *Diff\_count*[*content*] > *Threshold.rules*:

*Part\_Trans\_Rules* += *content*

As the example in Sect. 3.2 demonstrates.

1. Select the high similarity instance, collecting statistical data in accordance with frequency of occurrence of different parts.

2. Include paired dissimilar words such as “センター” (centre) and “中心”, which are the different parts, according to their frequency of occurrence above the threshold, add into “NE Partial Translation rules”.

The following is another example to illustrate rule extraction processing using IL as shown in Fig. 2. Two kinds of rules are acquired using IL: common part rule (CPR) and different part rule (DPR). Here, to prevent the problem of explosive growth of rule acquisition and ensure the quality

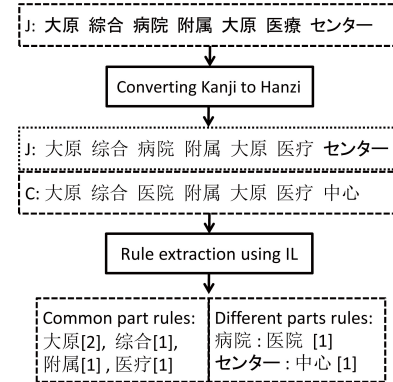


Fig. 2 Examples of rule extraction.

of extracted rules with IL, we give some constraint conditions as part of our rule extraction strategy:

1. If we consider HKMT as a kind of set of given primitive rules, we do not need to do primitive level rule extraction from NE pairs using IL.

2. Extract morphological units and their frequency as CPR format as shown in Fig. 2. The merit of this idea is used to prevent explosive growth of the number of extracted CPRs and to improve the accuracy of similarities of NE pairs.

3. For DPR extraction processing, use the pairs of different parts and their frequency to express the acquired DPR as shown in Fig. 2. We adopt the longest match principle to acquire DPRs.

4. Perform rule extraction processing from NE pairs in both forward and backward directions. We do not extract rules with the reverse or crossover mapping relation of elements of NE pairs.

For instance, given NE pairs as follows:

Japanese NE (Converted Kanji to Hanzi):

XXX Hanzi1Hanzi2 Hanzi3 YYY

Chinese NE:

XXX Hanzi1 Hanzi2Hanzi3 YYY

In this case, XXX and YYY express a certain Chinese segment corresponding to the common parts of the NE pairs.

According to our rule extracting strategy, the acquired CPRs will include the words XXX and YYY with their frequency, and the DPR will be the pair of the longest different parts of *Hanzi1Hanzi2 Hanzi3* and *Hanzi1 Hanzi2Hanzi3* and their frequencies.

These constraint conditions can ensure the high quality of extracted rules through adjusting the threshold of similarity of NE pairs. The new rules can be used to further improve the similarity calculation results.

## 4. Experiments and Analysis

### 4.1 Data and Evaluation Criteria

The source monolingual corpora used in the experiments are taken from Wikipedia database, and it contains a total of 69,874 Japanese monolingual documents and 84,055 Chi-

**Table 6** Numbers of NE recognition (Wikipedia).

| Type | Chinese | Japanese |
|------|---------|----------|
| PER  | 88203   | 73322    |
| LOC  | 183677  | 152688   |
| ORG  | 49442   | 41101    |

**Table 7** Condition as determined by Gold standard.

|              |          | True           | False          |
|--------------|----------|----------------|----------------|
| Test outcome | Positive | True positive  | False positive |
| Test outcome | Negative | False negative | True negative  |

nese monolingual documents. The monolingual NER tools used in the experiments are the CRF character-based tagging NER tools of our lab. In this paper, the model for the token position of both of Chinese and Japanese NEs is a given **BIEO** model, which indicates whether the token is at the **B**eginning, **I**nside, at the **E**nd or **O**utside of a NE. The local features of Chinese and Japanese are character-based and are instantiated from the following templates:

Unigram:  $C_n (n = -2, -1, 0, 1, 2)$ .

Bigram:  $C_n C_{n+1} (n = -2, -1, 0, 1)$  and  $C_{-1} C_1$ . where  $C_0$  means the current character,  $C_1$  is the next character,  $C_2$  denotes the second character after  $C_0$ ,  $C_{-1}$  represents the character preceding  $C_0$ , and  $C_{-2}$  is the second character before  $C_0$ .

The numbers of NEs extracted from the selected Wikipedia documents are shown in Table 6. We randomly selected 8,000 entries of extracted NEs and manually aligned and used these part as an experimental golden answer. We performed 10-fold cross-validation using these data to evaluate the performance of Chinese and Japanese NE extraction accuracies.

We use Precision ( $P$ ), Recall ( $R$ ) and F-measure ( $F$ ), which are defined in Formula (3-5) as the evaluation criteria of the results. When the parameter  $\beta = 1$ , that is, the precision and recall rate have the same weight, This is known as the  $F_1$  score. In this paper, we adopt the  $F_1$  score as the  $F$  measure to evaluate our experiments.

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}} \times 100\% \quad (3)$$

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}} \times 100\% \quad (4)$$

$$F_\beta = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R} \times 100\% \quad (5)$$

where  $N_{tp}$  means the number of “True positives”,  $N_{fp}$  expresses the number of “False positives”,  $N_{fn}$  presents the number of “False negative”, as shown in Table 7.

For further comparison, we also collected a Japanese monolingual corpus from Yahoo Japan, and a Chinese monolingual corpus from Sina using web robot during a period of two months (Sep.-Oct. 2014). The topic of choice was international news, and we randomly select 1,000 documents from each corpus to form the experimental data set. The numbers of extracted Chinese and Japanese NEs are shown in Table 8. We randomly selected 2,000 entries from

**Table 8** Numbers of NE recognition (Sina-Yahoo).

| Type | Chinese | Japanese |
|------|---------|----------|
| PER  | 1068    | 1153     |
| LOC  | 2234    | 2316     |
| ORG  | 766     | 653      |

**Table 9** Evaluation results of NE extraction.

| Type        | P(%)  | R(%)  | F(%)  |
|-------------|-------|-------|-------|
| PER-C-wiki  | 94.65 | 86.36 | 90.32 |
| PER-C-sina  | 93.58 | 85.42 | 89.31 |
| LOC-C-wiki  | 92.88 | 85.76 | 89.18 |
| LOC-C-sina  | 91.67 | 84.63 | 88.01 |
| ORG-C-wiki  | 88.53 | 82.87 | 85.61 |
| ORG-C-sina  | 88.36 | 81.59 | 84.84 |
| PER-J-wiki  | 92.84 | 87.68 | 90.20 |
| PER-J-yahoo | 92.81 | 86.36 | 89.47 |
| LOC-J-wiki  | 91.60 | 90.18 | 90.88 |
| LOC-J-yahoo | 90.24 | 88.65 | 89.44 |
| ORG-J-wiki  | 87.83 | 83.14 | 85.42 |
| ORG-J-yahoo | 85.52 | 81.75 | 83.60 |

extracted NEs and aligned them manually. We also performed 10-fold cross-validation to evaluate the performance of the Chinese and Japanese NE extraction.

We found it difficult to adopt a traditional method as a baseline system. As we have not found other method directly extract Chinese-Japanese NE equivalents from monolingual or comparable corpus, most of published research extracts NE equivalents from other language pairs or terminology equivalents from technical documents. For instance, the language pairs of Chinese-English and Japanese-English may be employed. Many researchers focus on Chinese-Japanese technical dictionary construction from parallel comparable corpora.

## 4.2 Experimental Results

The experimental results of the Chinese and Japanese NE extraction are shown in Table 9, where “C” means Chinese named entity, “J” means Japanese named entity, “wiki” means the extracted NEs from Wikipedia documents, “sina” means the extracted Chinese NEs from collected Sina news data, and “yahoo” means the extracted Japanese NEs from collected Yahoo Japan news corpus. Table 9 shows that the quality of extracted NEs is satisfied for the requirement of NE pair extraction experiments with a small amount of manual proofreading.

Before the evaluation experiment, we performed the preliminary experiments using the Greedy method to determine the optimum thresholds of the rule extraction of IL. We collected 1,000 NE pairs of Japanese and Chinese as a development data set, and 200 pairs as test data set. Since the threshold is between 0 and 1, we performed the preliminary experiments with an interval of 0.1, and used the test data set to evaluate the results of preliminary experiments and obtain the optimum threshold. To obtain a more reasonable value of the threshold, we performed another set of preliminary experiments with an interval of 0.01 based on Greedy

**Table 10** The preliminary experimental results.

| No. of iterations | P(%)  | R(%)  | F(%)  |
|-------------------|-------|-------|-------|
| 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 |
| 7                 | 92.60 | 85.21 | 88.75 |
| 8                 | 92.63 | 85.24 | 88.78 |
| 9                 | 92.66 | 85.26 | 88.81 |
| 10                | 92.67 | 85.25 | 88.81 |

**Table 11** The results of NE pairs extraction (Wikipedia).

| Type | P(%)  | R(%)  | F(%)  |
|------|-------|-------|-------|
| PER  | 89.81 | 81.48 | 85.44 |
| LOC  | 93.30 | 89.25 | 91.23 |
| ORG  | 93.22 | 86.30 | 89.63 |

method between  $V_{th} - 0.05$  and  $V_{th} + 0.05$ , where  $V_{th}$  means the previously acquired optimum threshold by the first set of preliminary experiments.

The preliminary experimental results based on our method are shown in Table 10. The results show that the effect is satisfactory when our proposed method is used to extract NE pairs from unrelated monolingual corpora (or quasi-comparable corpora). With increasing number of iterations, the difference list of translations becomes so large that it makes similarities of NE pairs move in ascending order. Experiments prove that the proposed method is simple and effective. Through careful analysis of the experimental results, we have shown that this method performs better if named entities contained Kanji. However, when the Japanese NEs are entirely Kana and have no association with existing partial translation rules, this method does not recognize anything. Such words as “コーネリアス” and “小山田圭吾” (Keigo Oyamada), have no correlation or internal rules. Our method is completely unable to identify these NE pairs, whose similarity calculation result is 0.

The final results of different categories of NE translation equivalents extraction using Wikipedia data are shown in Table 11. The number of iterations has some relationship with the length of the NE. The lengths of PER are lower than the average length of NE, so we cannot generate new rules after 2 iterations. But the ORG and LOC will give new rules with higher iteration. In addition, rule extraction threshold value should be gradually relaxed during the iteration process, otherwise we cannot extract the new rules after 2 iterations. It is also not possible to select the low threshold at the beginning of the process. Thus, the rules will greatly improve redundancy. The threshold tends to decrease exponentially, which demonstrates its high robustness.

We also used our collected Yahoo and Sina monolingual corpus to do NE pairs extraction experiments. We prepared 1,000 Japanese and Chinese NE pairs as a development data set, and 200 pairs as test data set, randomly selected from the manually aligned 2,000 pairs of extracted

**Table 12** The results of NE pairs extraction (Sina-Yahoo).

| Type | P(%)  | R(%)  | F(%)  |
|------|-------|-------|-------|
| PER  | 89.68 | 70.25 | 78.78 |
| LOC  | 93.51 | 50.42 | 65.51 |
| ORG  | 93.71 | 51.16 | 66.19 |

NEs. The final results of each NE translation equivalents extraction data shown in Table 12.

Table 12 shows that more than 50% of the Chinese and Japanese NEs can be automatically aligned by using our proposed method, and the total accuracy of our approach is 92.3%. It also demonstrates that the recall rate of Chinese and Japanese NE pairs acquisition can reflect the correlation of Chinese and Japanese monolingual corpora.

To evaluate the quality of extracted rules, we randomly selected 5,000 rules from acquired DPR and CPR rules and performed error evaluation by applying the manual judgement of an experienced expert to the selected rules. The number of rules containing any incorrect character error or semantic error was counted. This analysis shows that the error rate of extracted DPR rules is less than 15%, and the error rate of CPR is less than 10%, with very low frequency.

According to our experimental results analysis, our method acquires abundant translation rules by using IL. Particularly, in the case of some examples including Japanese Kanji and Kana, our method extracted many functional rules. For instance, from example pairs of “文山チワン族ミャオ族自治州” and “文山壮族苗族自治州” (Wenshan Zhuang and Miao Autonomous Prefecture), our system acquired some rule pairs, such as “チワン族” and “壮族” (Zhuang), “ミャオ族” and “苗族” (Miao). These rules indicated its simple and powerful practicability according to the low disambiguation of this kind of rules.

Moreover, we also found the complexity of rule adaptability by using IL. Of the extracted translation rules, we found that some retain a high level of disambiguation. For example, some rule pairs, containing Japanese name of “アキコ” (Akiko), match multiple Chinese candidates, including “明子”, “亚希子”, “昭子” and “晶子” etc. These rules cause a rule adaptation problem, similar to the disambiguation problem of Kana to Kanji conversion, or the common word sense disambiguation problem of machine translation.

For instance, for Japanese name “東村アキコ” (Higashimura Akiko), our system required some translation results, such as “东村明子”, “东村亚希子”, “东村晶子” and “东村昭子”. For “松たか子”, our system generated some results, such as “松隆子” (Matsu Takako), “松貴子” and “松多香子”.

## 5. Conclusion and Future work

The principal contribution of this paper is the development of an approach for extracting NE translation equivalents from Chinese and Japanese monolingual corpora by integrating inductive learning with Hanzi-Kanji mapping tables. Our approach achieves high robustness due to its universal capability of automatic extraction of NE pairs from Chinese

and Japanese monolingual corpora. Our approach can also be extended to comparable and parallel corpora, and to other domain, such as, terminology pair extraction from technical corpora.

Our approach prevents the generated rules from growing exponentially, and ensures the quality of extracted rules according to our designed IL rule extraction strategies. This approach significantly reduces the cost of dictionary generation by its simplicity and efficiency. It requires minimal effort to acquire additional knowledge when using weak correlation bilingual text sets and minimal additional knowledge to extract NE translation equivalents.

The remaining downside to this approach is that pure Kana NE equivalent extraction. In such a situation, the approach does not function properly in extracting partial translation rules. Our future research will focus on solutions to this problem of pure Kana NE pair extraction by means of IL and MT/machine transliteration technologies.

## Acknowledgments

The authors are supported by the National Nature Science Foundation of China (Contract 61370130 and 61473294), and the Fundamental Research Funds for the Central Universities (2015JBM033), and the International Science and Technology Cooperation Program of China under grant No. 2014DFA11350.

## References

- [1] R. Grishman and B. Sundheim, "Message understanding conference 6: A brief history," *Proc. COLING*, vol.96, pp.466–471, 1996.
- [2] Chengqing Zong, *Statistical Natural Language Processing* (in Chinese), 2nd Edition, Tsinghua University Press, Aug. 2013.
- [3] R. Sproat and T. Emerson, "The first international Chinese word segmentation bakeoff," *Proc. second SIGHAN workshop on Chinese language processing*, vol.17, *ACL* 2003, pp.133–143, 2003.
- [4] K. Ru, J. Xu, Y. Zhang, and P. Wu, "A method to construct Chinese-Japanese named entity translation equivalents using monolingual corpora," *Proc. Natural Language Processing and Chinese Computing*, 2013, Springer Berlin Heidelberg, vol.400, pp.164–175, 2013.
- [5] Y. Al-Onaizan and K. Knight, "Translating named entities using monolingual and bilingual resources," *Proc. ACL* 2002, pp.400–408, 2002.
- [6] H.-H. Chen, C. Yang, and Y. Lin, "Learning formulation and transformation rules for multilingual named entities," *Proc. ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, vol.15, pp.1–8, 2003.
- [7] P. Virga and S. Khudanpur, "Transliteration of proper names in cross-lingual information retrieval," *Proc. ACL 2003 workshop on Multilingual and mixed-language named entity recognition*, vol.15, pp.57–64, 2003.
- [8] B. Zou and J. Zhao, "Comparison of Several English-Chinese Name Transliteration Methods," *The Fourth Young Scholar Symposium on Natural Language Processing*, China, pp.336–343, 2008.
- [9] Y. Chen and C. Zong, "A structure-based model for Chinese organization name translation," *ACM Transactions on Asian Language Information Processing (TALIP)*, 2008, vol.7, no.1, pp.1–30, 2008.
- [10] J.-H. Wang, J.-W. Teng, P.-J. Cheng, W.-H. Lu, and L.-F. Chien, "Translating unknown cross-lingual queries in digital libraries using a web-based approach," *Proc. 4th ACM/IEEE-CS joint conference on Digital libraries*, ACM, pp.108–116, 2004.
- [11] P.-J. Cheng, Y.-C. Pan, W.-H. Lu, and L.-F. Chien, "Creating multilingual translation lexicons with regional variations using web corpora," *Proc. ACL* 2004, pp.534–541, 2004.
- [12] M. Nagata, T. Saito, and K. Suzuki, "Using the Web as a bilingual dictionary," *Proc. workshop on Data-driven methods in machine translation*, vol.14, pp.1–8, 2001.
- [13] Y. Zhang, F. Huang, and S. Vogel, "Mining translations of OOV terms from the web through cross-lingual query expansion," *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp.669–670, 2005.
- [14] L. Jiang, M. Zhou, L. Chien, et al., "Named Entity Translation with Web Mining and Transliteration," *Proc. IJCAI* 2007, pp.1629–1634, 2007.
- [15] F. Yang, J. Zhao, B. Zou, K. Liu, and F. Liu, "Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages," *Proc. ACL* 2008, pp.541–549, 2008.
- [16] J. Kupiec, "An algorithm for finding noun phrase correspondences in bilingual corpora," *Proc. ACL* 1993, pp.17–22, 1993.
- [17] F. Huang, S. Vogel, and A. Waibel, "Automatic extraction of named entity translational equivalence based on multi-feature cost minimization," *Proc. ACL Workshop on Multilingual and Mixed-Language Named Entity Recognition*, vol.15, pp.9–16, 2003.
- [18] D. Feng, Y. Lv, and M. Zhou, "A new approach for English-Chinese named entity alignment," *Proc. EMNLP* 2004, pp.372–379, 2004.
- [19] R. Rapp, "Identifying word translations in non-parallel texts," *Proc. ACL* 1995, pp.320–322, 1995.
- [20] S. Wan and C.M. Verspoor, "Automatic English-Chinese name transliteration for development of multilingual resources," *Proc. ACL* 1998, vol.2, pp.1352–1356, 1998.
- [21] R. Rapp, "Automatic identification of word translations from unrelated English and German corpora," *Proc. ACL* 1999, pp.519–526, 1999.
- [22] P. Fung and L.Y. Yee, "An IR approach for translating new words from nonparallel, comparable texts," *Proc. 17th International Conference on Computational Linguistics*, vol.1, pp.414–420, 1998.
- [23] L. Shao and H.T. Ng, "Mining new word translations from comparable corpora," *Proc. 20th International Conference on Computational Linguistics*, pp.618–624, 2004.
- [24] L. Lee, A. Aw, M. Zhang, et al., "EM-based hybrid model for bilingual terminology extraction from comparable corpora," *Proc. 23rd International Conference on Computational Linguistics*, pp.639–646, 2010.
- [25] G. You, S. Hwang, Y.I. Song, et al., "Mining name translations from entity graph mapping," *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, pp.430–439, 2010.
- [26] J. Kim, S.-W. Hwang, L. Jiang, Y.-I. Song, and M. Zhou, "Entity translation mining from comparable corpora: Combining graph mapping with corpus latent features," *IEEE Trans. Knowl. Data Eng.*, vol.25, no.8, pp.1787–800, 2013.
- [27] C. Chu, T. Nakazawa, D. Kawahara, and S. Kurohashi, "Chinese-Japanese machine translation exploiting Chinese characters," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol.12, no.4:16, pp.1–25, 2013.
- [28] C. Chu, T. Nakazawa, and S. Kurohashi, "Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese-Japanese Wikipedia," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol.15, no.2:10, pp.1–22, 2016.
- [29] C.-L. Goh, M. Asahara, and Y. Matsumoto, "Building a Japanese-Chinese dictionary using kanji/hanzi conversion," *Proc. IJCNLP* 2005, vol.3651, pp.670–681, 2005.
- [30] Y. Zhang, Q. Liu, Q. Ma, and H. Isahara, "A multi-aligner for Japanese-Chinese parallel corpora," *The Tenth Machine Translation Summit Proceedings*, pp.133–140, 2005.
- [31] T. Tsunakawa, N. Okazaki, X. Liu, and J.I. Tsujii, "A Chinese-Japanese lexical machine translation through a pivot language," *ACM Transactions on Asian Language Information Processing*

(TALIP), vol.8, no.2:9, pp.1–21, 2009.

- [32] R. Dabre, C. Chu, F. Cromieres, T. Nakazawa, and S. Kurohashi, "Large-scale Dictionary Construction via Pivot-based Statistical Machine Translation with Significance Pruning and Neural Network Features," *Proc. PACLIC2015*, pp.293–301, 2015.
- [33] M.M. Hasan and Y. Matsumoto, "Japanese-Chinese Cross-Language Information Retrieval: An Interlingua Approach," *Computational Linguistics*, vol.5, no.2, pp.59–86, 2000.
- [34] C.-C. Lin, Y.-C. Wang, and R.T.-H. Tsai, "Japanese-Chinese Information Retrieval With and Iterative Weighting Scheme," *Journal of Information Science and Engineering*, vol.26, no.2, pp.685–697, 2010.
- [35] K. Araki, Y. Takahashi, Y. Momouchi, and K. Tochinal, "Non-Segmented Kana-Kanji Translation Using Inductive Learning," *IEICE Trans. Inf. & Syst.*, vol.J79-D-II, no.3, pp.391–402, 1996.
- [36] R. Udupa, K. Saravanan, A. Kumaran, and J. Jagarlamudi, "Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora," *Proc. 12th Conference of the European Chapter of the ACL*, pp.799–807, 2009.
- [37] T. Tao and C. Zhai, "Mining comparable bilingual text corpora for cross-language information integration," *Proc. eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp.691–696, 2005.



**JinAn Xu** is currently an Associate Professor in School of Computer Science and information technology, Beijing Jiaotong University, Beijing, China. He received the B.Eng. degree from Beijing Jiaotong University in 1992, the MS degree and Ph.D. degree in computer information from Hokkaido University, Sapporo, Japan, in 2003 and 2006, respectively. His research focuses on natural language processing, machine translation, information retrieve, text mining, and machine learning. He is a member

of CCF, CIPSC, ACL and the ACM.



**Yufeng Chen** received the BS degree in Mechanical Electrical Engineering from Beijing Jiaotong University in 2003, and the Ph.D degree in Pattern Recognition and Intelligent Systems from National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, in June, 2008. From July 2008 to September 2014, she worked in NLPR. From September 2014, she joined School of Computer and Information Technology, Beijing Jiaotong University and now works as an asso-

ciate professor. Her research interests include natural language processing, machine translation, information retrieval and so on. Also, her recent research projects include "Semantic-Based Multiword Expression Analysis and Application for Machine Translation" supported by the Natural Science Foundation.



**Kuang Ru** received the MS degree from Beijing Jiaotong University in 2014, China. His research focuses on Natural Language processing and machine translation.



**Yujie Zhang** is Professor of Computer and Information Technology at Beijing Jiaotong University. She received the Ph.D. degree from the University of Electro-Communications, Japan. Her main research interests include NLP, MT and machine learning. She received the first prize of National Science and Technology Progress Award. She served on the editorial board of numerous journals and several conferences, such as *Journal of Chinese Information Processing*, the *IEICE Transactions*, *MT Summit* 2011, *Coling* 2014.



**Kenji Araki** was born in 1959 in Otaru in Japan. He received his B.E. and Ph.D. degrees from Hokkaido University in 1982 and 1988. He was an associate professor there from 1991 to 1998, and a professor in 1998 at Hokkai-Gakuen University. He then joined Hokkaido University as an associate professor. He is currently a professor at the Graduate School of Information Science and Technology. His research interests include natural language processing, morphological analysis, machine translation, and speech dialogue processing. He is a member of IEICE, IPSJ, JSAI, JCSS, ACL, IEEE, and the AAAI.