Chinese-Japanese Machine Translation Exploiting Chinese Characters

CHENHUI CHU, TOSHIAKI NAKAZAWA, DAISUKE KAWAHARA,
and SADAO KUROHASHI, Kyoto University

The Chinese and Japanese languages share Chinese characters. Since the Chinese characters in Japanese originated from ancient China, many common Chinese characters exist between these two languages. Since Chinese characters contain significant semantic information and common Chinese characters share the same meaning in the two languages, they can be quite useful in Chinese-Japanese machine translation (MT). We therefore propose a method for creating a Chinese character mapping table for Japanese, traditional Chinese, and simplified Chinese, with the aim of constructing a complete resource of common Chinese characters. Furthermore, we point out two main problems in Chinese word segmentation for Chinese-Japanese MT, namely, unknown words and word segmentation granularity, and propose an approach exploiting common Chinese characters to solve these problems. We also propose a statistical method for detecting other semantically equivalent Chinese characters other than the common ones and a method for exploiting shared Chinese characters in phrase alignment. Results of the experiments carried out on a state-of-the-art phrase-based statistical MT system and an example-based MT system show that our proposed approaches can improve MT performance significantly, thereby verifying the effectiveness of shared Chinese characters for Chinese-Japanese MT.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*

General Terms: Languages, Experimentation, Performance

Additional Key Words and Phrases: Chinese characters, Chinese-Japanese, segmentation, phrase alignment, machine translation

ACM Reference Format:
Chu, C., Nakazawa, T., Kawahara, D., and Kurohashi, S. 2013. Chinese-Japanese machine translation exploiting Chinese characters. *ACM Trans. Asian Lang. Inform. Process.* 12, 4, Article 16 (October 2013), 25 pages.
DOI:http://dx.doi.org/10.1145/2523057.2523059

1. INTRODUCTION

Differing from other language pairs, Chinese and Japanese share Chinese characters. In Chinese, Chinese characters are called Hanzi, while in Japanese they are called Kanji. Hanzi can be divided into two groups, simplified Chinese (used in mainland China and Singapore) and traditional Chinese (used in Taiwan, Hong Kong, and Macau). The number of strokes needed to write characters has been largely reduced in simplified Chinese, and the shapes may be different from those in traditional Chinese. Because kanji characters originated from ancient China, many common Chinese characters exist in hanzi and kanji.

Authors' address: C. Chu (corresponding author), T. Nakazawa, D. Kawahara, and S. Kurohashi, Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan; emails: {chu, nakazawa}@nlp.ist.i.kyoto-u.ac.jp, {dk, kuro}@i.kyoto-u.ac.jp.

DOI:http://dx.doi.org/10.1145/2523057.2523059

Since Chinese characters contain a significant amount of semantic information and common Chinese characters share the same meaning, they can be valuable linguistic clues in many Chinese-Japanese natural language processing (NLP) tasks. Many studies have exploited common Chinese characters. For example, Tan et al. [1995] used the occurrence of identical common Chinese characters in Chinese and Japanese in an automatic sentence alignment task. Goh et al. [2005] detected common Chinese characters where kanji are identical to traditional Chinese but differ from simplified Chinese. Using a Chinese encoding converter[1] that can convert traditional Chinese into simplified Chinese, they built a Japanese-Simplified Chinese dictionary partly using direct conversion of Japanese into Chinese for Japanese kanji words. Huang et al. [2008] examined and analyzed the semantic relations between Chinese and Japanese at a word level based on a common Chinese character mapping. They used a small list of 125 visual variational pairs of manually matched common Chinese characters.

However, the resources for common Chinese characters used in these previous studies are not complete. In this article, we propose a method for automatically creating a Chinese character mapping table for Japanese, traditional Chinese, and simplified Chinese using freely available resources, with the aim of constructing a more complete resource containing common Chinese characters.

Besides common Chinese characters, there are also many other semantically equivalent Chinese characters in the Chinese and Japanese languages. These Chinese characters would also be valuable in Chinese-Japanese machine translation (MT), especially in word/phrase alignment. However, since there are no available resources for such Chinese characters, we propose a statistical method for detecting these characters, which we call statistically equivalent Chinese characters.

In corpus-based Chinese-Japanese MT, parallel sentences contain equivalent meanings in each language, and we assume that common Chinese characters and statistically equivalent Chinese characters appear in the sentences. In this article, we point out two main problems in Chinese word segmentation for Chinese-Japanese MT, namely, unknown words and word segmentation granularity, and propose an approach exploiting common Chinese characters to solve these problems. Furthermore, we propose a method for exploiting common Chinese characters and statistically equivalent Chinese characters in phrase alignment. Experimental results show that our proposed approaches improve MT performance significantly.

2. CHINESE CHARACTER MAPPING TABLE

Table I gives some examples of Chinese characters in Japanese, traditional Chinese, and simplified Chinese, from which we can see that the relation between kanji and hanzi is quite complicated.

Because kanji characters originated from ancient China, most kanji have fully corresponding Chinese characters in hanzi. In fact, despite Japanese having continued to evolve and change since its adoption of Chinese characters, the visual forms of the Chinese characters have retained a certain level of similarity; many kanji are identical to hanzi (e.g., "雪 (snow)" in Table I), some kanji are identical to traditional Chinese characters but differ from simplified Chinese ones (e.g., "愛 (love)" in Table I), while others are identical to simplified Chinese characters but differ from traditional Chinese ones (e.g., "国 (country)" in Table I). There are also some visual variations in kanji that have corresponding Chinese characters in hanzi, although the shapes differ from those in hanzi (e.g., "発 (begin)" in Table I). However, there are some kanji that do not have fully corresponding Chinese characters in hanzi. Some kanji only have

[1] http://www.mandarintools.com/zhcode.html

Table I. Examples of Chinese Characters

	C1	C2	C3	C4	C5	C6
Meaning	snow	love	country	begin	octopus	included
Kanji	雪	愛	国	発	鱆	込
Traditional Chinese	雪	愛	國	發	鱆	N/A
Simplified Chinese	雪	爱	国	发	N/A	N/A

Note: "C" denotes Category, which is described in Section 2.3.

corresponding traditional Chinese characters (e.g., "鱆 (octopus)" in Table I), because they were not simplified into simplified Chinese. Moreover, there are some Chinese characters that originated in Japan namely, kokuji, which means that these national characters may have no corresponding Chinese characters in hanzi (e.g., "込 (included)" in Table I).

What makes the relation even more complicated is that a single kanji form may correspond to multiple hanzi forms. Also, a single simplified Chinese form may correspond to multiple traditional Chinese forms, and vice versa.

Focusing on the relation between kanji and hanzi, we present a method for automatically creating a Chinese character mapping table for Japanese, traditional Chinese, and simplified Chinese using freely available resources [Chu et al. 2012b]. Common Chinese characters shared in Chinese and Japanese can be found in the mapping table. Because Chinese characters contain significant semantic information, this mapping table could be very useful in Chinese-Japanese MT.

2.1. Kanji and Hanzi Character Sets

The character set in use for kanji is JIS Kanji code, whereas for hanzi, there are several, of which we have selected Big5 for traditional Chinese and GB2312 for simplified Chinese, both of which are widely used.

- For JIS Kanji code, JIS X 0208 is a widely used character set specified as the Japanese Industrial Standard, containing 6,879 graphic characters, including 6,355 kanji and 524 non-kanji. The mapping table is for the 6,355 kanji characters, that is, JIS Kanji, in JIS X 0208.
- Big5 is the most commonly used character set for traditional Chinese in Taiwan, Hong Kong, and Macau, and was defined by the Institute for Information Industry in Taiwan. There are 13,060 traditional Chinese characters in Big5.
- GB2312 is the main official character set of the People's Republic of China for simplified Chinese characters and is widely used in mainland China and Singapore. GB2312 contains 6,763 simplified Chinese characters.

2.2. Related Freely Available Resources

- Unihan database[2] is the repository for the Unicode Consortium's collective knowledge regarding the CJK (Chinese-Japanese-Korean) Unified Ideographs contained in the Unicode Standard.[3] The database consists of a number of fields containing data for each Chinese character in the Unicode Standard. These fields are grouped into categories according to their purpose, including mappings, readings, dictionary indices, radical stroke counts, and variants. The mappings and variants categories contain information regarding the relation between kanji and hanzi.

[2] `http://unicode.org/charts/unihan.html`

[3] The Unicode Standard is a character coding system for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems. The latest version of the Unicode Standard is 6.1.0.

Table II. Hanzi Converter Standard Conversion Table

Traditional Chinese	故	說	錢	沖, 衝	干, 幹, 乾	...
Simplified Chinese	故	说	钱	冲	干	...

Table III. Kanconvit Mapping Table

Kanji	安	詞	会	広	壱	潟	...
Traditional Chinese	安	詞	會	廣	壹	瀉	...
Simplified Chinese	安	词	会	广	壹	泻	...

- The Chinese encoding converter[4] is an open-source system that converts traditional Chinese into simplified Chinese. The hanzi converter standard conversion table, a resource used by the converter, contains 6,740 corresponding traditional Chinese and simplified Chinese character pairs. It can be downloaded from the website. Table II depicts a portion of the table.
- Kanconvit[5] is a publicly available tool for kanji-simplified Chinese conversion. It uses 1,159 visual variational kanji-simplified Chinese character pairs extracted from a Kanji, traditional Chinese, and simplified Chinese mapping table, containing 3,506 one-to-one mappings. Table III depicts a portion of this table.

2.3. Construction Method

Based on the relation between kanji and hanzi, we define the following seven categories for kanji.

- *Category* 1. Identical to hanzi.
- *Category* 2. Identical to traditional Chinese but different from simplified Chinese.
- *Category* 3. Identical to simplified Chinese but different from traditional Chinese.
- *Category* 4. Visual variations.
- *Category* 5. With a corresponding traditional Chinese character only.
- *Category* 6. No corresponding hanzi.
- *Others*. Does not belong to the preceding categories.

We create a Chinese character mapping table for Japanese, traditional Chinese, and simplified Chinese by classifying JIS Kanji into these seven categories and automatically finding the corresponding traditional Chinese and simplified Chinese characters using the resources introduced in Section 2.2. The method involves two steps.

- *Step* 1. Extraction.
- *Step* 2. Categorization and construction.

In Step 1, we extract the JIS Kanji, Big5 Traditional Chinese, and GB2312 Simplified Chinese from the Unihan database. These Chinese characters are collected in the mappings category, which contains mappings between Unicode and other encoded character sets for Chinese characters. JIS Kanji are obtained from the *kIRG_JSource J0* field, Big5 Traditional Chinese from the *kBigFive* field, and GB2312 Simplified Chinese from the *kIRG_GSource G0* field.

In Step 2, we categorize the JIS Kanji and construct a mapping table. We automatically check every character in the JIS Kanji as follows. If the kanji exists in both Big5 and GB2312, it belongs to Category 1. If the kanji exists only in Big5, we check whether a corresponding simplified Chinese character can be found; if so, it belongs to

[4] http://www.mandarintools.com/zhcode.html

[5] http://kanconvit.ta2o.net/

Table IV. Examples of Multiple Hanzi Forms

Kanji	弁	伝	鯰	働	...
Traditional Chinese	弁, 瓣, 辦, 辯, 辮, 辨	傳, 伝	鯰	動, 仂	...
Simplified Chinese	弁, 瓣, 办, 辩, 辫, 辨	传	鲶, 鲇	动, 仂	...

Table V. Resource Statistics

	C1	C2	C3	C4	C5	C6	Others
Unihan	3141	1815	177	533	384	289	16
+Han	3141	1843	177	542	347	289	16
+Kan	3141	1847	177	550	342	282	16

Note: "Han" denotes the hanzi converter standard conversion table, while "Kan" denotes the Kanconvit mapping table.

Category 2, otherwise, it belongs to Category 5. If the kanji exists only in GB2312, we check whether a corresponding traditional Chinese character can be found; if so, it belongs to Category 3. If the kanji exists in neither Big5 nor GB2312, we check whether corresponding hanzi can be found; if a fully corresponding Chinese character exists in hanzi, it belongs to Category 4, else if only a corresponding traditional Chinese character exists, it belongs to Category 5, else if no corresponding Chinese character exists in hanzi, it belongs to Category 6, otherwise, it belongs to Others.

To find the corresponding hanzi, we search traditional Chinese and simplified Chinese variants, as well as other variants for all kanji. This search is carried out using the variants category in the Unihan database, in which there are five fields: *kTraditionalVariant* corresponding to traditional Chinese variants, *kSimplifiedVariant* corresponding to simplified Chinese variants, and *kZVariant*, *kSemanticVariant*, and *kSpecializedSemanticVariants* corresponding to the other variants. In addition, we also use the hanzi converter standard conversion table and Kanconvit mapping table. Note that the resources in the hanzi converter standard conversion table can only be used for the traditional Chinese and simplified Chinese variants search, whereas the Kanconvit mapping table can also be used for the other variants search.

2.4. Details of the Mapping Table

The format for kanji in Categories 1, 2, 3, and 4 in the mapping table is as follows.

— Kanji[TAB]Traditional Chinese[TAB]Simplified Chinese[RET].

If multiple hanzi forms exist for a single kanji, we separate them with ",". Table IV shows some examples of multiple hanzi forms. The formats for kanji in Categories 5 and 6 are as follows.

— *Category* 5. Kanji[TAB]Traditional Chinese[TAB]N/A[RET].
— *Category* 6. Kanji[TAB]N/A[TAB]N/A[RET].

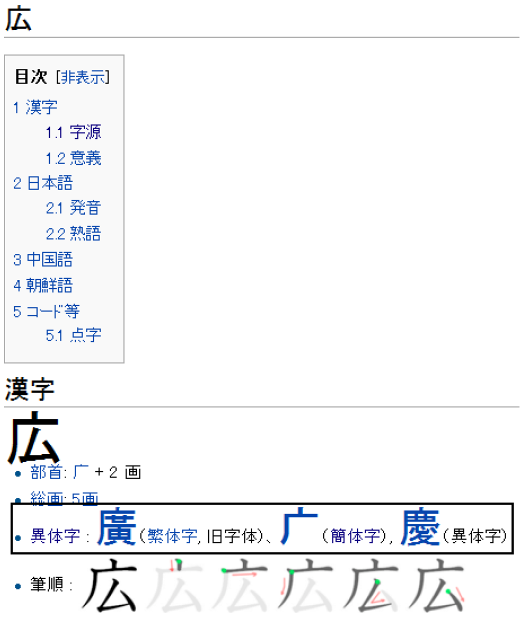
Table V gives some statistics of the Chinese character mapping table we created for Japanese, traditional Chinese, and simplified Chinese. Here, "Others" are the kanji that have a corresponding simplified Chinese character only. There are corresponding traditional Chinese characters for these kanji, but they were not collected in Big5 Traditional Chinese. Kanji "鮃 (bastard halibut)" is one of such example. Compared with using only the Unihan database, incorporating the hanzi converter standard conversion and Kanconvit mapping tables can improve the completeness of the mapping table. Tables VI and VII give some examples of additional Chinese character mappings found using the hanzi converter standard conversion table and Kanconvit mapping table, respectively.

Table VI. Examples of Additional Mappings Found Using the Hanzi Converter Standard Conversion Table

Kanji	祇	託	浄	畭	...
Traditional Chinese	祇, 只, 衹, 隻, 祗	託, 侘, 托	淨, 凈	畬	...
Simplified Chinese	祇, 只	托	净	畲	...

Table VII. Examples of Additional Mappings Found Using the Kanconvit Mapping Table

Kanji	雰	艶	対	県	挿	...
Traditional Chinese	氛, 雰	豔, 艷	對	縣	插	...
Simplified Chinese	氛	艳	对	县	插	...



広

目次 [非表示]

1 漢字
1.1 字源
1.2 意義
2 日本語
2.1 発音
2.2 熟語
3 中国語
4 朝鮮語
5 コード等
5.1 点字

漢字

広

- 部首: 广 + 2 画
- 総画: 5画
- 異体字 : 廣（繁体字, 旧字体）、广（簡体字）, 慶（異体字）
- 筆順 :

Fig. 1. Example of Kanji 広 from Japanese Wiktionary.

2.5. Completeness Evaluation

To show the completeness of the mapping table we created, we used a resource from Wiktionary[6], which is a wiki project aimed at producing a free-content multilingual dictionary. In the Japanese version of Wiktionary, there is a kanji category that provides a great deal of information about kanji, such as variants, origins, meanings, pronunciation, idioms, kanji in Chinese and Korean, and codes. We are interested in the variants part. Figure 1 gives an example of kanji “広” from the Japanese Wiktionary, in which the variants part, containing the traditional Chinese variant “廣”, simplified Chinese variant “广”, and other variant “慶” of kanji “広”, is enclosed by a rectangle.

We downloaded the Japanese Wiktionary database dump data[7] (January 31, 2012) and extracted the variants for JIS Kanji. We then constructed a mapping table based on the Wiktionary using the method described in Section 2.3, the only difference being

[6] http://www.wiktionary.org/

[7] http://dumps.wikimedia.org/jawiktionary/

Table VIII. Completeness Comparison between Proposed Method and Wiktionary

	C1	C2	C3	C4	C5	C6	Others
Proposed	3141	1847	177	550	342	282	16
Wiktionary	3141	1781	172	503	412	316	30
Combination	3141	1867	178	579	325	249	16

Table IX. Examples of Mappings that Do not Exist in Wiktionary

Kanji	尨	茘	値	幇	咲	...
Traditional Chinese	尨, 龍	荔	值	幫	笑	...
Simplified Chinese	龙	荔	值	帮	笑	...

Table X. Examples of Mappings not Found by the Proposed Method

Kanji	冴	扨	畳	滝	愼	...
Traditional Chinese	冱, 沍	扠, 叉	疊	瀧	慎	...
Simplified Chinese	冱	叉	叠	泷	慎	...

that for the traditional Chinese, simplified Chinese, and other variants search, we used the variants extracted from the Japanese Wiktionary.

To evaluate the completeness of the mapping table created using the proposed method, we compared the statistics thereof with those of Wiktionary. Table VIII shows the completeness comparison between the proposed method and Wiktionary. We can see that the proposed method creates a more complete mapping table than Wiktionary. Table IX gives some examples of Chinese character mappings found by the proposed method, but which do not exist in the current version of Wiktionary.

Furthermore, we carried out an experiment by combining the mapping table we created with Wiktionary. The results in Table VIII show that Wiktionary can be used as a supplementary resource to further improve the completeness of the mapping table. Table X gives some examples of Chinese character mappings contained in Wiktionary, but which were not found by the proposed method.

2.6. Coverage of Shared Chinese Characters

We investigated the coverage of shared Chinese characters on a simplified Chinese-Japanese corpus, namely, the scientific paper abstract corpus provided by JST[8] and NICT.[9] This corpus was created by the Japanese project Development and Research of Chinese–Japanese Natural Language Processing Technology. Some statistics of this corpus are given in Table XI.

We measured the coverage in terms of both characters and words under two different experimental conditions.

— *Identical*. Only exactly the same Chinese characters.
— +*Common*. Perform kanji-to-hanzi conversion for common Chinese characters using the Chinese character mapping table constructed, as described in Section 2.

Table XII presents the coverage results for shared Chinese characters. If we use all the resources available, we can find corresponding hanzi characters for over 76% of the kanji characters.

[8] http://www.jst.go.jp
[9] http://www.nict.go.jp/

Table XI. Statistics of Chinese-Japanese Corpus

	Ja	Zh
# of sentences	680k	
# of words	21.8M	18.2M
# of Chinese characters	14.0M	24.2M
average sentence length	32.9	22.7

Table XII. Coverage of Shared Chinese Characters

	Character		Word	
	Ja	Zh	Ja	Zh
Identical	52.41%	30.48%	26.27%	32.09%
+Common	76.66%	44.58%	32.84%	39.46%

2.7. Related Work

Hantology [Chou and Huang 2006] is a character-based Chinese language resource, which has adopted the Suggested Upper Merged Ontology (SUMO) [Niles and Pease 2001] for a systematic and theoretical study of Chinese characters. Hantology represents orthographic forms, the evolution of script, pronunciation, senses, lexicalization, as well as variants for different Chinese characters. However, the variants in Hantology are limited to Chinese hanzi.

Chou et al. [2008] extended the architecture of Hantology to Japanese kanji and included links between Chinese hanzi and Japanese kanji, thereby providing a platform for systematically analyzing variations in kanji. However, a detailed analysis of variants of kanji has not been presented. Moreover, because the current version of Hantology only contains 2,100 Chinese characters, whereas our mapping table includes all 6,355 JIS Kanji, it is difficult to create a mapping table between kanji and hanzi based on Hantology and which is as complete as our proposed method.

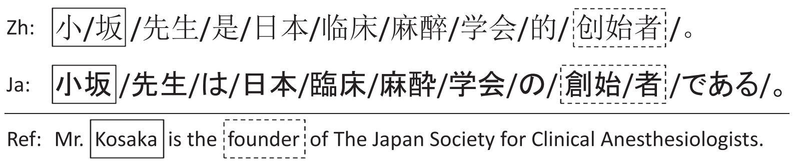
3. EXPLOITING SHARED CHINESE CHARACTERS IN CHINESE WORD SEGMENTATION OPTIMIZATION

3.1. Motivation

As there are no explicit word boundary markers in Chinese, word segmentation is considered an important first step in MT. Studies have shown that an MT system with Chinese word segmentation outperforms those treating each Chinese character as a single word, while the quality of Chinese word segmentation affects MT performance [Chang et al. 2008; Xu et al. 2004]. It has been found that besides segmentation accuracy, segmentation consistency and granularity of Chinese words are also important for MT [Chang et al. 2008]. Moreover, optimal Chinese word segmentation for MT is dependent on the other languages, and therefore, a bilingual approach is necessary [Ma and Way 2009].

Most studies have focused on language pairs containing Chinese and another language with white spaces between words (e.g., English). Our focus is on Chinese-Japanese MT, where segmentation is needed on both sides. Segmentation for Japanese successfully achieves an F-score of nearly 99% [Kudo et al. 2004], while that for Chinese is still about 95% [Wang et al. 2011]. Therefore, we only do word segmentation optimization in Chinese and use the Japanese segmentation results directly.

Similar to previous works, we also consider the following two Chinese word segmentation problems to be important for Chinese-Japanese MT. The first problem relates to unknown words, which cause major difficulties for Chinese segmenters and affect segmentation accuracy and consistency. Consider, for example, “Kosaka” shown in Figure 2, which is a proper noun in Japanese. Because “Kosaka” is an unknown word



Zh: 小/坂/先生/是/日本/临床/麻醉/学会/的/创始者/。

Ja: 小坂/先生/は/日本/臨床/麻酔/学会/の/創始/者/である/。

Ref: Mr. Kosaka is the founder of The Japan Society for Clinical Anesthesiologists.

Fig. 2. Example of Chinese word segmentation problems in Chinese-Japanese MT.

for a Chinese segmenter, it is mistakenly segmented into two tokens, whereas the Japanese word segmentation result is correct.

The second problem is word segmentation granularity. Most Chinese segmenters adopt the famous Penn Chinese Treebank (CTB) standard [Xia et al. 2000], while most Japanese segmenters adopt a shorter unit standard. Therefore, the segmentation unit in Chinese may be longer than that in Japanese, even for the same concept. This can increase the number of 1-to-n alignments, making the word alignment task more difficult. Taking “founder” in Figure 2 as an example, the Chinese segmenter recognizes it as one token, while the Japanese segmenter splits it into two tokens because of the different word segmentation standards.

To solve these problems, we proposed an approach based on a bilingual perspective that exploits common Chinese characters shared between Chinese and Japanese in Chinese word segmentation optimization for MT [Chu et al. 2012a]. In this approach, Chinese entries are extracted from a parallel training corpus based on common Chinese characters to augment the system dictionary of a Chinese segmenter. In addition, the granularity of the training data for the Chinese segmenter is adjusted to that of the Japanese one by means of extracted Chinese entries.

3.2. Chinese Entry Extraction

Chinese entries are extracted from a parallel training corpus through the following steps.

- *Step* 1. Segment Japanese sentences in the parallel training corpus.
- *Step* 2. Convert Japanese tokens consisting only of kanji[10] into simplified Chinese using the Chinese character mapping table created in Section 2.
- *Step* 3. Extract the converted tokens as Chinese entries if they exist in the corresponding Chinese sentence.

For example, “小坂 (Kosaka)”, “先生 (Mr.)”, “日本 (Japan)”, “临床 (clinical)”, “麻醉 (anesthesia)”, “学会 (society)”, “创始 (found)” and “者 (person)” in Figure 2 would be extracted. Note that although “临床 ↔ 臨床 (clinical)”, “麻醉 ↔ 麻酔 (anesthesia)” and “创始 ↔ 創始 (found)” are not identical, because “临 ↔ 臨 (arrive)”, “醉 ↔ 酔 (drunk)” and “创 ↔ 創 (create)” are common Chinese characters, “臨床 (clinical)” is converted into “临床 (clinical)”, “麻酔 (anesthesia)”, is converted into “麻醉 (anesthesia),” and “創始 (found)” is converted into “创始 (found)” in Step 2.

3.3. Chinese Entry Incorporation

Several studies have shown that using a system dictionary is helpful for Chinese word segmentation [Low et al. 2005; Wang et al. 2011]. Therefore, we used a corpus-based Chinese word segmentation and POS tagging tool with a system dictionary and incorporated the extracted entries into the system dictionary. The extracted entries are not only effective for the unknown word problem, but also useful in solving the word segmentation granularity problem.

[10] Japanese has several other kinds of character types apart from kanji.

Table XIII. Chinese-Japanese POS Tag Mapping Table

JUMAN	CTB
副詞 (adverb)	AD
接続詞 (conjunction)	CC
名詞 (noun) [数詞 (numeral noun)]	CD
未定義語 (undefined word) [アルファベット (alphabet)]	FW
感動詞 (interjection)	IJ
接尾辞 (suffix) [名詞性名詞助数辞 (measure word suffix)]	M
名詞 (noun) [普通名詞 (common noun) / サ変名詞 (sahen noun) / 形式名詞 (formal noun) / 副詞的名詞 (adverbial noun)] / 接尾辞 (suffix) [名詞性名詞接尾辞 (noun suffix) / 名詞性特殊接尾辞 (special noun suffix)]	NN
名詞 (noun) [固有名詞 (proper noun) / 地名 (place name) / 人名 (person name) / 組織名 (organization name)]	NR
名詞 (noun) [時相名詞 (temporal noun)]	NT
特殊 (special word)	PU
形容詞 (adjective)	VA
動詞 (verb) / 名詞 (noun) [サ変名詞 (sahen noun)]	VV

However, setting POS tags for the extracted entries is problematic. To solve this problem, we created a POS tag mapping table between Chinese and Japanese by hand. For Chinese, we used the POS tagset used in CTB, which is also used in our Chinese segmenter. For Japanese, we used the POS tagset defined in the morphological analyzer JUMAN [Kurohashi et al. 1994]. JUMAN uses a POS tagset containing sub-POS tags. For example, the POS tag “名詞 (noun)” contains sub-POS tags, such as “普通名詞 (common noun)”, “固有名詞 (proper noun)”, “時相名詞 (temporal noun)”, and so on. Table XIII shows a part of the Chinese-Japanese POS tag mapping table we created, with the sub-POS tags of JUMAN given within square brackets.

POS tags for the extracted Chinese entries are assigned by converting the POS tags of Japanese tokens assigned by JUMAN into POS tags of CTB. Note that not all POS tags of JUMAN can be converted into POS tags of CTB, and vice versa. Those that cannot be converted are not incorporated into the system dictionary.

3.4. Short-Unit Transformation

Bai et al. [2008] showed that adjusting Chinese word segmentation to create a token one-to-one mapping as far as possible between parallel sentences can improve alignment accuracy, which is crucial for corpus-based MT. Wang et al. [2010] proposed a short-unit standard for Chinese word segmentation that is more similar to the Japanese word segmentation standard, and which can reduce the number of one-to-n alignments and improve MT performance.

We previously proposed a method for transforming the annotated training data of the Chinese segmenter into the Japanese word segmentation standard using the extracted Chinese entries, and then used the transformed data to train the Chinese segmenter [Chu et al. 2012a]. Because the extracted entries are derived from Japanese word segmentation results, they follow the Japanese word segmentation standard. Therefore, we utilize these entries in short-unit transformation. We use the Chinese entries extracted in Section 3.2 and modify every token in the training data for the Chinese segmenter. If the token is longer than a extracted entry, it is simply split. Figure 3 gives an example of this process, where “有效 (effective)” and “要素 (element)” are both extracted entries. Because “有效性 (effectiveness)” is longer than “有效 (effective)”, it is split into “有效 (effective)” and “性”

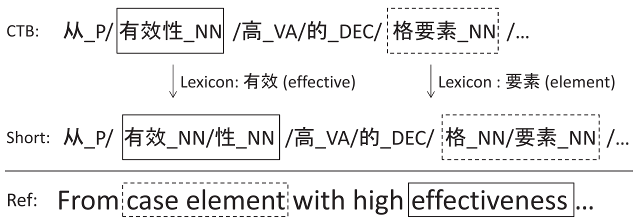

CTB: 从_P/ 有效性_NN /高_VA/的_DEC/ 格要素_NN /...
Lexicon: 有效 (effective)
Lexicon : 要素 (element)
Short: 从_P/ 有效_NN/性_NN /高_VA/的_DEC/ 格_NN/要素_NN /...
Ref: From case element with high effectiveness ...

Fig. 3. Example of previous short-unit transformation.

(a noun suffix), and since "格要素 (case element)" is longer than "要素 (element)", it is split into "格 (case)" and "要素 (element)". For POS tags, the originally annotated one is retained for the split tokens.

Although this method works well in most cases, it suffers from the problem of transformation ambiguity. For example, for a long token like "留学生 (student studying abroad)" in the annotated training data, entries "留学 (study abroad)" and "学生 (student)" are extracted from the parallel training corpus. In this case, our previous method randomly chose one entry for transformation. Therefore, "留学生 (student studying abroad)" could be split into "留 (stay)"and "学生 (student)", which is incorrect. To solve this problem, we improved the transformation method by utilizing both short-unit information and extracted entries. Short-unit information is short-unit transformation information extracted from the parallel training corpus. Short-unit information extraction is similar to the Chinese entry extraction described in Section 3.2 and includes the following steps.

—*Step* 1. Segment both Chinese and Japanese sentences in the parallel training corpus.
—*Step* 2. Convert Japanese tokens consisting of only kanji into simplified Chinese using the Chinese character mapping table we created in Section 2.
—*Step* 3. Extract the converted tokens composed of consecutive tokens in the segmented Chinese sentence and the corresponding Chinese tokens.

For example, "创始者 (founder) → 创始 (found) / 者 (person)" in Figure 2 is extracted as short-unit information.

In the improved transformation method, we modify the tokens in the training data using the following processes in order.

(1) If the token itself exists in the extracted entries, keep it.
(2) If the token can be transferred using short-unit information, transfer it according to the short-unit information.
(3) If the token can be split using extracted entries, transfer it according to the extracted entries.
(4) Otherwise, keep it.

Following Chu et al. [2012a], we do not use extracted entries that are composed of only one Chinese character, because these entries may lead to undesirable transformation results. Taking the Chinese character "歌 (song)" as an example, "歌 (song)" can be used as a single word, but we can also use "歌 (song)" to construct other words by combining it with other Chinese characters, such as "歌颂 (praise)", "诗歌 (poem)", and so on. Obviously, splitting "歌颂 (praise)" into "歌 (song)" and "颂 (eulogy)", or splitting "诗歌 (poem)" into "诗 (poem)" and "歌 (song)" is undesirable. We do not use extracted number entries either, as these, could also lead to undesirable transformation. For example, using "十八 (18)" to split "二百九十八 (298)" into "二百九 (290)" and "十八 (18)" is obviously incorrect. Moreover, there are a few consecutive tokens in the

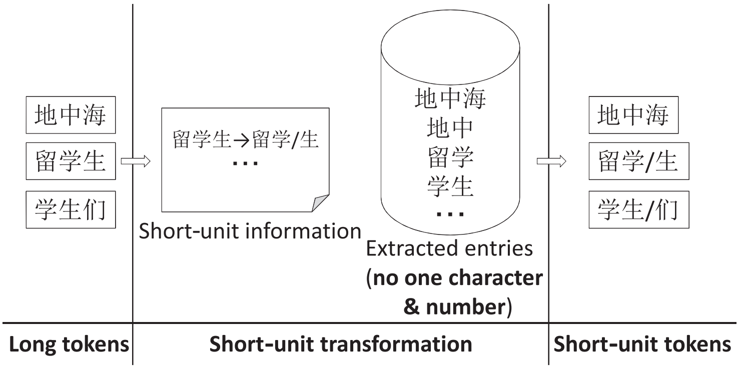


Fig. 4. Example of improved short-unit transformation.

training data that could be combined into a single extracted entry; however, we do not consider these patterns.

Figure 4 gives an example of our improved transformation method. In this example, since “地中海 (Mediterranean)” also exists in the extracted entries, it is not changed, even though there is an extracted entry “地中(in earth)”. The long token “留学生 (student studying abroad)” can be transferred using short-unit information, so it is transferred into “留学 (study abroad)” and “生 (student)”. Meanwhile, the long token “学生们 (students)” can be split into “学生 (student)” and “们 (plural for student)” using the extracted entry “学生 (student)”.

We record the extracted entries and short-unit information used for transformation with the corresponding Japanese tokens and store them as transforming word dictionary. This dictionary contains 16K entries and will be helpful for word alignment.

3.5. Experiments

We conducted Chinese-Japanese translation experiments on the state-of-the-art phrase-based statistical MT toolkit MOSES [Koehn et al. 2007] and an example-based MT (EBMT) system [Nakazawa and Kurohashi 2011b] to show the effectiveness of exploiting common Chinese characters in Chinese word segmentation optimization.

3.5.1. Experiments on Phrase-Based Statistical MT. The experimental settings for MOSES are given next. The parallel training corpus for this experiment was the same as that used in Section 2.6. We further used CTB 7 (LDC2010T07)[11] to train the Chinese segmenter. Training data, containing 31,131 sentences, was created from CTB 7 using the same method described in Wang et al. [2011]. The segmenter used for Chinese was an in-house corpus-based word segmentation and POS tagging tool with a system dictionary. Weights for the entries in the system dictionary were automatically learned from the training data using an averaged structured perceptron [Collins 2002]. For Japanese, we used JUMAN [Kurohashi et al. 1994]. All the default options were used in Moses, except for the distortion limit (6→20), which was tuned by MERT using a further 500 development sentence pairs. We trained a word-based 5-gram language model on the target side of the training data using the SRILM toolkit [Stolcke 2002]. We translated five test sets from the same domain as the parallel training corpus. The statistics of the test sets of Chinese and Japanese sentences are given in Tables XIV and XV, respectively. Note that none of the sentences in the test sets are included in the parallel training corpus.

[11] http://www.ldc.upenn.edu/

Table XIV. Statistics of Test Sets Containing Chinese Sentences

	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
# sentences	255	336	391	395	393	1770
# words	6.6K	8.7K	10.0K	11.7K	16.6K	53.6K
# Chinese characters	8.6K	10.7K	12.9K	15.8K	22.1K	70.1K
average sentence length	44.9	47.0	45.4	52.2	74.1	53.58

Note: "Total" denotes the combined statistics for the five test sets.

Table XV. Statistics of Test Sets Containing Japanese Sentences

	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
# sentences	255	336	391	395	393	1770
# words	8.0K	11.0K	12.7K	14.4K	20.1K	66.2K
# Chinese characters	5.1K	6.3K	7.7K	9.0K	13.0K	41.1K
average sentence length	55.6	57.6	56.6	66.2	90.4	66.3

Table XVI. Results of Chinese-to-Japanese Translation Experiments on MOSES for Use in Chinese Word Segmentation Optimization

BLEU	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
Baseline	51.03	48.98	40.52	29.20	26.08	36.64
Chu+ 2012	**52.83**	51.13	41.57	31.01	**28.82**	38.59*
Optimized	52.55	51.88	41.62	30.69	28.43	38.52*
+Dictionary	52.73	**52.21**	**42.02**	**31.19**	28.72	**38.86***

Note: "*" denotes that the "Total" result is better than "Baseline" significantly at $p < 0.01$.

We carried out Chinese-Japanese translation experiments, comparing the following three experimental settings.

- *Baseline*. Using only entries extracted from the Chinese annotated corpus as the system dictionary for the Chinese segmenter.
- *Optimized*. Incorporating the Chinese entries extracted in Section 3.2 into the system dictionary and training the Chinese segmenter on the short-unit training data transformed in Section 3.4.
- *+Dictionary*. Appending the transforming word dictionary stored in Section 3.4 to the parallel training corpus.

The translations were evaluated using BLEU–4 [Papineni et al. 2002] calculated on words. For Japanese-to-Chinese translation, we resegmented the translations using the optimized Chinese segmenter. Tables XVI and XVII give the BLEU scores for Chinese-to-Japanese and Japanese-to-Chinese translation, respectively. For comparison, we also list the optimized results of Chu et al. [2012a], which are denoted as Chu+ 2012. The results show that our proposed approach can improve MT performance. We notice that compared with Chu et al. [2012a], the improvement in the current short-unit transformation method further improved the Japanese-to-Chinese translation. However, it had no effect on the Chinese-to-Japanese translation. Appending the transforming word dictionary further improved the translation performance. Similar to Chu et al. [2012a], the improvement in Japanese-to-Chinese translation compared with that in Chinese-to-Japanese translation is not that significant. We believe the reason for this is the input sentence. For Chinese-to-Japanese translation, the segmentation of input Chinese sentences is optimized, whereas for Japanese-to-Chinese translation, our proposed approach does not change the segmentation results of the input Japanese sentences.

Table XVII. Results of Japanese-to-Chinese Translation Experiments on MOSES for Use in Chinese Word Segmentation Optimization

BLEU	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
Baseline	42.26	42.47	35.60	26.70	27.92	33.31
Chu+ 2012	42.89	43.27	34.95	27.80	28.82	33.90*
Optimized	43.06	44.04	35.53	28.00	**29.04**	34.30*†
+Dictionary	**43.17**	**44.78**	**36.34**	**28.10**	28.89	**34.53***‡

Note: "*" denotes that the "Total" result is better than "Baseline" significantly at $p < 0.01$, "†" and "‡" denote that the "Total" result is better than "Chu+ 2012" significantly at $p < 0.05$ and $p < 0.01$, respectively.

Table XVIII. Results of Chinese-to-Japanese Translation Experiments on EBMT for Use in Chinese Word Segmentation Optimization

BLEU	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
Baseline	36.74	31.67	23.90	16.60	15.21	22.84
Optimized	**37.30**	32.12	24.48	**17.29**	**15.55**	23.34†
+Dictionary	37.17	**32.69**	**24.76**	16.93	15.54	**23.41**‡

Note: "†" and "‡" denote that the "Total" result is better than Baseline significantly at $p < 0.05$ and $p < 0.01$, respectively.

Table XIX. Results of Japanese-to-Chinese Translation Experiments on EBMT for Use in Chinese Word Segmentation Optimization

BLEU	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
Baseline	37.75	**27.46**	19.95	15.08	**14.00**	20.74
Optimized	**38.19**	27.34	20.14	**15.27**	13.66	**20.78**
+Dictionary	38.08	27.18	**20.25**	15.03	13.69	20.70

3.5.2. Experiments on EBMT. We also conducted Chinese-Japanese translation experiments on EBMT, comparing the same three experimental settings described in Section 3.5.1. The parallel training data, test sets, and Chinese and Japanese segmenters were the same as those used in the experiments on MOSES. Since the EBMT system we used is a dependency-tree-based decoder, we further used CNP [Chen et al. 2008] as the Chinese dependency analyzer, while the Japanese dependency analyzer was KNP [Kawahara and Kurohashi 2006].

Tables XVIII and XIX show the BLEU scores for Chinese-to-Japanese and Japanese-to-Chinese translation, respectively. We can see that the Chinese-to-Japanese translation performance on EBMT also improves when exploiting shared Chinese characters in Chinese word segmentation optimization. However, the improvement is not as significant as that on MOSES. Moreover, it has no effect for Japanese-to-Chinese translation. The reason for this may be that the Chinese parser CNP is not trained on optimized segmented training data. Therefore, using optimized segmented sentences as input for CNP may affect the parsing accuracy. Compared with the translation performance on MOSES, the BLEU scores on EBMT are quite low. The reason for this is that both the alignment model and decoder for EBMT are disadvantaged by the low accuracy of the Chinese parser. Although the Chinese parser we used is a state-of-the-art one [Chen et al. 2008], the accuracy is less than 80%. On the other hand, the Japanese parser used in the experiments can analyze sentences with over 90% accuracy. We believe that further improvement of the Chinese parser could improve the translation performance on EBMT.

Input: 本论文中，提议考虑现存实现方式的功能适应性决定对策目标的保密基本设计法。

Baseline (BLEU=49.38)
Segmented: 本/论文/中/，/提议/考虑/现存/实现/方式/的/ 功能 / 适应性 /决定/对策/目标/的/保密/基本/设计法/。
Output: 本/論文/で/は/，/提案/する/ 適応/的 /対策/を/決定/する/セキュリティ/基本/設計/法/を/考える/現存/の/実現/方式/の/ 機能 /を/目標/と/して/いる/.

Segmentation Optimization (BLEU=56.33)
Segmented: 本/论文/中/，/提议/考虑/现存/实现/方式/的/ 功能 / 适应/性 /决定/对策/目标/的/保密/基本/设计/法/。
Output: 本/論文/で/は/，/提案/する/考え/現存/の/実現/方式/の/ 機能/的 / 適応/性 /を/決定/する/対策/目標/の/セキュリティ/基本/設計/法/を/提案/する/.

Reference
本/論文/で/は/，/対策/目標/を/現存/の/実現/方式/の/ 機能/的 / 適合/性 /も/考慮/して/決定/する/セキュリティ/基本/設計/法/を/提案/する/ .
(In this paper, we propose a basic security design method also consider functional suitability of the existing implementation method for determining countermeasures target.)

Fig. 5. Example of translation improvement.

3.6. Discussion

3.6.1. Short-Unit Effectiveness. Experimental results indicate that our proposed approach can improve MT performance significantly. We present an example to show the effectiveness of optimized short-unit segmentation results. Figure 5 gives an example of Chinese-to-Japanese translation improvement on MOSES using optimized short-unit segmentation results compared with the baseline. The difference between the short unit and baseline is whether “适应性 (suitability)” is split in Chinese, whereas the Japanese segmenter always splits it. By splitting it, the short unit improves word alignment and phrase extraction, which eventually affects the decoding process. In decoding, the short unit treats “功能适应性 (functional suitability)” as one phrase, while the baseline separates it, leading to a undesirable translation result.

3.6.2. Short-Unit Transformation Problems. Although we have improved the short-unit transformation method, there are still some transformation problems. One problem is incorrect transformation. For example, there is a long token “不好意思 (sorry)” and an extracted entry “好意 (favor)”, and therefore, the long token is transferred into “不 (not)”, “好意 (favor)”, and “思 (think)”, which is obviously undesirable. Our current method cannot deal with such cases, making this one of the future works in this study.

Another problem is POS tag assignment for the transformed short-unit tokens. Our proposed method simply keeps the original annotated POS tag of the long token for the transformed short-unit tokens, which works well in most cases. However, there are also some exceptions. For example, there is a long token “被实验者 (test subject)” in the annotated training data, and an entry “实验 (test)” extracted from the parallel training corpus, so the long token is split into “被 (be)”, “实验 (test)”, and “者 (person)”. As the POS tag for the original long token is NN, the POS tags for the transformed short-unit tokens are all set to NN, which is undesirable for “被 (be)”. The correct POS tag for “被 (be)” should be LB. An external dictionary would be helpful in solving this problem. Furthermore, the transformed short-unit tokens may have more than one possible POS tag. All these problems will be dealt with in future work.

3.7. Related Work

Exploiting lexicons from external resources [Chang et al. 2008; Peng et al. 2004] is one way of dealing with the unknown word problem. However, the external lexicons may not be very efficient for a specific domain. Some studies [Ma and Way 2009; Xu et al. 2004] have used the method of learning a domain-specific dictionary from the character-based alignment results of a parallel training corpus, which separate each Chinese character, and consider consecutive Chinese characters as a lexicon in n-to-one alignment results. Our proposed method differs from these studies in that we obtain a domain-specific dictionary by extracting Chinese lexicons directly from a segmented parallel training corpus, making word alignment unnecessary.

The goal of our proposed short-unit transformation method is to form the segmentation results of Chinese and Japanese into a one-to-one mapping, which can improve alignment accuracy and MT performance. Bai et al. [2008] proposed a method for learning affix rules from an aligned Chinese-English bilingual terminology bank to adjust Chinese word segmentation in the parallel corpus directly with the aim of achieving the same goal. Our proposed method does not adjust Chinese word segmentation directly. Instead, we utilize the extracted Chinese lexicons to transform the annotated training data of a Chinese segmenter into a short-unit standard and perform segmentation using the retrained Chinese segmenter.

Wang et al. [2010] also proposed a short-unit transformation method. The proposed method is based on transfer rules and a transfer database. The transfer rules are extracted from alignment results of annotated Chinese and segmented Japanese training data, while the transfer database is constructed using external lexicons and is manually modified. Our proposed method learns transfer knowledge based on common Chinese characters. Moreover, no external lexicons or manual work is required.

4. EXPLOITING SHARED CHINESE CHARACTERS IN PHRASE ALIGNMENT

4.1. Motivation

Chinese characters contain a significant amount of semantic information, with common Chinese characters having the same meaning. Moreover, parallel sentences contain equivalent meanings in each language, and we can assume that common Chinese characters appear in these sentences. Therefore, common Chinese characters can be valuable in MT, especially in word/phrase alignment between Chinese and Japanese. Figure 6 shows an example of Chinese-Japanese alignment from bi-directional GIZA++, where “事实” and “実際” (both mean “in fact”) are not automatically aligned. We notice this because these two words share a common Chinese character (i.e., “实” ↔ “実”/“fact”). If we could exploit the common Chinese character correctly, these two words could be successfully aligned.

Motivated by this phenomenon, we previously proposed a method for exploiting common Chinese characters in a joint phrase alignment model and proved the effectiveness of common Chinese characters in Chinese-Japanese phrase alignment [Chu et al. 2011]. A kanji-to-hanzi conversion method making use of the Unihan database and Chinese encoding converter was proposed in our previous work. In this article, we extend the kanji-to-hanzi conversion method by using the Chinese character mapping table we created in Section 2.

Besides common Chinese characters, there are also several other semantically equivalent Chinese characters in Chinese and Japanese. We feel that these Chinese characters would also be valuable in MT, especially in word/phrase alignment. However, there are no available resources for these Chinese characters. We proposed a statistical method for detecting such Chinese characters, which we call statistically equivalent Chinese characters [Chu et al. 2012c]. We improved our previous exploitation

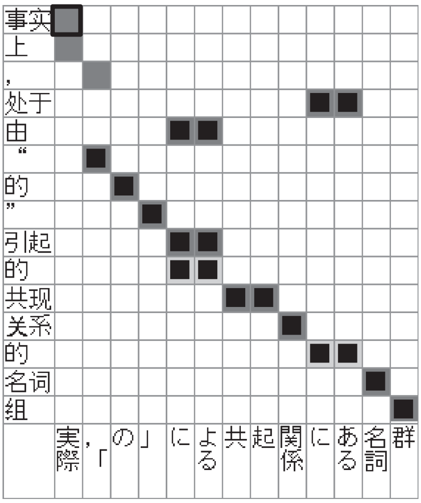


Fig. 6. Example of Chinese characters in Chinese-Japanese alignment. Black boxes depict the system output, while dark (sure) and light (possible) gray cells denote gold-standard alignments.

Table XX. Examples of Other Semantically Equivalent Chinese Characters

Meaning	eat	word	hide	look	day	
Kanji	食	語	隠	見	日	...
Traditional Chinese	吃	詞	藏	看	天	...
Simplified Chinese	吃	词	藏	看	天	...

method to make use of statistically equivalent Chinese characters, together with common Chinese characters in a joint phrase alignment model. We showed that statistically equivalent Chinese characters can also improve alignment accuracy. In this study, we follow the shared Chinese characters exploitation method proposed in Chu et al. [2012c].

4.2. Statistically Equivalent Chinese Character Detection

Table XX gives some examples of other semantically equivalent Chinese characters in Chinese and Japanese besides the common Chinese characters in Japanese, traditional Chinese, and simplified Chinese. Although these Chinese characters are not common Chinese characters, they have the same meaning. We proposed a statistical method for detecting statistically equivalent Chinese characters.

Figure 7 illustrates the basic idea of our statistically equivalent Chinese character detection method. The example parallel sentences (both mean "critical information is hidden") share common Chinese characters (e.g., "隠" ↔ "隐"/"hide"), as well as other semantically equivalent Chinese characters (e.g., "隠" ↔ "藏"/"hide"). In order to detect the other semantically equivalent Chinese characters, we first eliminate the kana characters in the Japanese sentence.[12] We treat each Chinese character as a single word and perform character-based alignment using GIZA++ [Och and Ney 2003], which implements a sequential word-based statistical alignment model for IBM models.

[12] Experimental results show that lexical translation probabilities of Chinese characters are well estimated, although simply eliminating the kana characters may generate unbalanced sentence pairs.

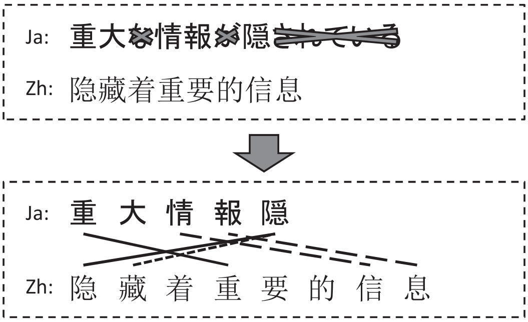

Ja: 重大~~な~~情報~~が~~隠~~されている~~
Zh: 隐藏着重要的信息
Ja: 重 大 情 報 隠
Zh: 隐 藏 着 重 要 的 信 息

Fig. 7. Character-based alignment.

Table XXI. Examples of Lexical Translation Probability Estimated Using Character-Based Alignment

f_i	e_j	$t(e_j \mid f_i)$	$t(f_i \mid e_j)$
隠	隐	0.287	0.352
重	重	0.572	0.797
隠	藏	0.122	0.006
大	藏	< 1.0e-07	5.07e-06
情	信	0.796	0.634
報	息	0.590	0.981

Table XXI gives examples of lexical translation probability estimated using character-based alignment on a Chinese–Japanese scientific paper abstract corpus. We can see that shared Chinese characters obtain high lexical translation probabilities. Although the lexical translation probability of “藏 (hide)” and “隠 (hide)” seems to be not that high compared with “藏 (hide)” and other Chinese characters (e.g., “大 (big)”), it is still prominent. Furthermore, because “情報” and “信息” always appear together in the parallel corpus and share the same meaning of “information”, “情” ↔ “信”, “報” ↔ “息” also obtained a high lexical translation probability. This kind of shared Chinese character would be helpful in MT carried out in the same domain. Since we conducted an experiment in the same domain, we kept them in the preliminary experiment. However, such shared Chinese characters may be problematic in different domains, because they are not semantically equivalent.

4.3. Alignment Model

We used the Bayesian subtree alignment model on dependency trees proposed by Nakazawa and Kurohashi [2011a]. In this model, the joint probability for a sentence pair is defined as

$$P(\{\langle e,f\rangle\},D) = P(\ell)\cdot P(D|\{\langle e,f\rangle\})\cdot\prod_{\langle e,f\rangle}\theta_T(\langle e,f\rangle), \tag{1}$$

where $P(\ell)$ is the geometric distribution denoting the number of concepts that generate phrase pairs, $P(D|\{\langle e,f\rangle\})$ is the dependency relation probability of phrases, and $\theta_T(\langle e,f\rangle)$ is the distribution that the phrase generation step obeys. Details of the model are omitted here.

Table XXII. Examples of Kana-Kanji Conversion Pairs and Their Corresponding Chinese Words

Meaning	envious	wanton	self
Kana	うらやましい	ワンタン	おのれ
Kanji	羨ましい	饂飩	己
Traditional Chinese	羡慕	餛飩	自己
Simplified Chinese	羡慕	馄饨	自己

4.4. Exploiting Shared Chinese Characters

We define a shared Chinese character matching ratio for Chinese-Japanese phrase pairs as

$$ratio(\langle e,f\rangle) = \frac{match\_zh\_char + match\_ja\_char}{num\_zh\_char + num\_ja\_char}, \qquad (2)$$

where *num_zh_char* and *num_ja_char* denote the numbers of Chinese characters in the Chinese and Japanese phrases, respectively, while *match_zh_char* and *match_ja_char* are the matching weights of the Chinese characters in the Chinese and Japanese phrases, respectively. For common Chinese characters, we regard the matching weight as one, and for statistically equivalent Chinese characters, we use the highest lexical translation probability for the Chinese character pair estimated in Section 4.2. Taking "信息局" and "情報局" (both mean "information agency") as an example, there is one common Chinese character "局 (agency)" and two statistically equivalent Chinese characters pairs, and thus,

$$match\_zh\_char = 1 + t(\text{“信”}|\text{“情”}) + t(\text{“息”}|\text{“報”}),$$
$$match\_ja\_char = 1 + t(\text{“情”}|\text{“信”}) + t(\text{“報”}|\text{“息”}).$$

We modified the Bayesian subtree alignment model by incorporating a weight w into the phrase generation distribution and redefined the joint probability for a sentence pair as

$$P(\{\langle e,f\rangle\},D) = P(\ell) \cdot P(D|\{e,f\}) \cdot \prod_{\langle e,f\rangle} w \cdot \theta_T(\langle e,f\rangle), \qquad (3)$$

where weight w is proportional to the shared Chinese character matching ratio

$$w = \alpha \cdot ratio(\langle e,f\rangle), \qquad (4)$$

where α is a variable set by hand.

Note that this exploitation method has the drawback that the joint probability of a sentence pair is no longer a probability.

4.5. Kana-Kanji Conversion

Following Chu et al. [2011], we perform kana-kanji conversion for Japanese kana words. Currently, many Japanese words are written in kana even if they have corresponding kanji expressions, which are normally used. The Chinese characters in kanji expressions are useful clues for finding shared Chinese characters. We can use kana-kanji conversion techniques to obtain kanji expressions from kana expressions, but here, we simply consult the Japanese dictionary of JUMAN [Kurohashi et al. 1994]. Table XXII gives some examples of kana-kanji conversion results. We only perform kana-kanji conversion of content words, because, as proven in our alignment experiments, conversion of function words may lead to incorrect alignment.

Table XXIII. Results of Chinese–Japanese Alignment Experiments

	Pre.	Rec.	AER
GIZA++(grow-diag-final-and)	83.77	75.38	20.39
BerkelyAligner	**88.43**	69.77	21.60
Baseline (Nakazawa+ 2011)	85.37	75.24	19.66
Chu+ 2011	85.47	76.53	18.94
+Common	85.55	76.54	18.90
+Common & SE	85.22	**77.31**	**18.65**

Note: “SE” denotes statistically equivalent.

4.6. Experiments

4.6.1. Alignment. We conducted alignment experiments on a Chinese-Japanese corpus to show the effectiveness of exploiting shared Chinese characters in phrase alignment.

The training corpus was the same as that used in Section 2.6. Statistically equivalent Chinese characters described in Section 4.2 are also detected using this corpus. As gold-standard data, we used 510 sentence pairs for Chinese-Japanese, which were annotated by hand. Two types of annotations were used: sure (S) alignments and possible (P) alignments [Och and Ney 2003]. The unit of evaluation was the word. We used precision, recall, and alignment error rate (AER) as evaluation criteria. All the experiments were run on the original forms of words. We set variable α to 5,000, which showed the best performance in the preliminary experiments for tuning the weights.

Chinese sentences were converted into dependency trees using the in-house segmenter described in Section 3.5.1 and dependency analyzer CNP [Chen et al. 2008]. Japanese sentences were converted into dependency structures using the morphological analyzer JUMAN [Kurohashi et al. 1994] and the dependency analyzer KNP [Kawahara and Kurohashi 2006].

For comparison, we used GIZA++ with its default parameters and conducted word alignment bidirectionally and merged these using the grow-diag-final-and heuristic [Koehn et al. 2003]. We also applied the BerkelyAligner.[13] [DeNero and Klein 2007] with its default settings for unsupervised training.[14] Experimental results are given in Table XXIII. The alignment accuracies of the Bayesian subtree alignment model, the method in Chu et al. [2011], the method exploiting common Chinese characters, and that exploiting both statistically equivalent and common Chinese characters are indicated as “Baseline (Nakazawa+ 2011)”, “Chu+ 2011”, “+Common”, and “+Common & statistically equivalent”. Here, we used all the statistically equivalent Chinese characters estimated in Section 4.2 rather than choosing those with lexical translation probability greater than a manually set threshold (i.e., $threshold = 1.0e - 07$, which is the default value for GIZA++), because this showed the best performance in the preliminary experiments for tuning the threshold. We can see that alignment accuracy improves when exploiting the shared Chinese characters. Because we extended the kanji-to-hanzi conversion method using the Chinese character mapping table and improved the exploitation method, “+Common” yields a slightly improved alignment accuracy compared with Chu et al. [2011]. Alignment accuracy is further improved by exploiting statistically equivalent Chinese characters.

Figure 8 shows an example of alignment improvement using common Chinese characters. Because there is a common Chinese character (i.e., “规” ↔ “規”/“rule”) in “规

[13] http://code.google.com/p/berkeleyaligner/

[14] BerkelyAligner also can do supervised training. Since our alignment model is unsupervised, we only compared to the unsupervised trained BerkelyAligner.

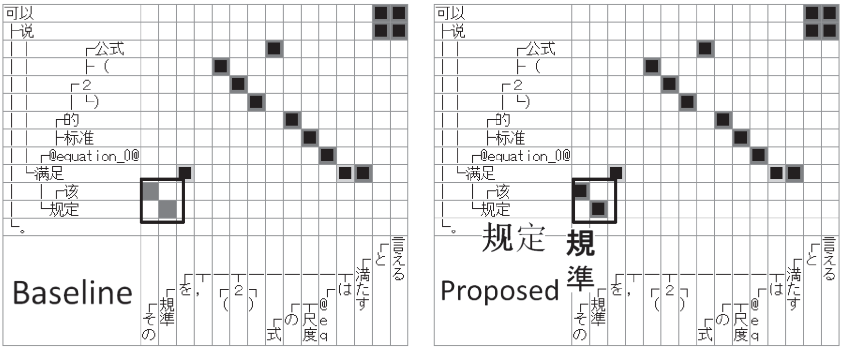


Fig. 8. Alignment improvement using common Chinese characters.

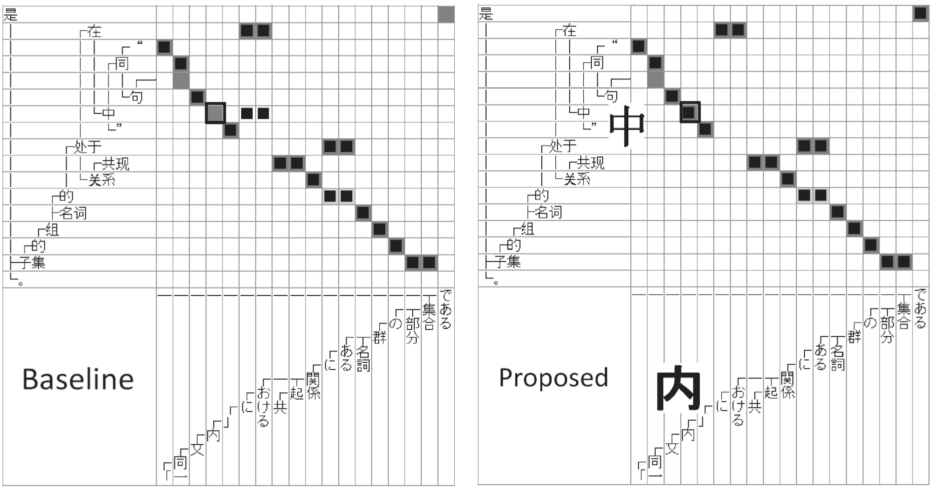


Fig. 9. Alignment improvement using statistically equivalent Chinese characters.

定” and “規準” (both mean “standard”), these two words are successfully aligned using our proposed method, and consequently, the alignment between “该” and “その” (both mean “that”) is discovered. Figure 9 shows an example of alignment improvement using statistically equivalent Chinese characters. Because “中” and “内” (both mean “in”) are statistically equivalent Chinese characters, they are successfully aligned using our proposed method.

Although in most cases shared Chinese characters achieve correct alignment, there are also some exceptions. Figure 10 shows an example of an exception using common Chinese characters. “稍微” and “少し” (both mean “a little”) are correctly aligned in the baseline alignment model. Because there is a common Chinese character (i.e., “少” ↔ “少”/“little”) in “至少” (means “at least”) and “少し” (means “a little”), the proposed method aligns them, resulting in an incorrect alignment.

4.6.2. Translation. Chinese-Japanese translation experiments were carried out on EBMT [Nakazawa and Kurohashi 2011b]. We conducted translation experiments on the same corpus used in the alignment experiment. The same test sets described in Section 3.5.1 were translated. Besides the experimental settings described in Section 4.6.1, we also carried out experiments exploiting common Chinese characters

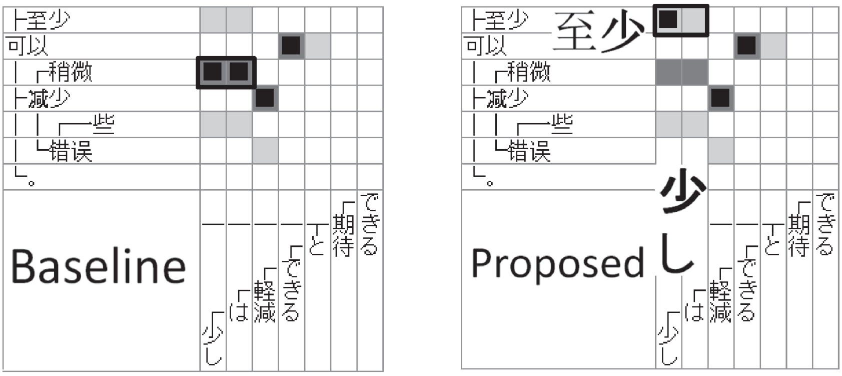


Fig. 10. Alignment exception to the proposed method.

Table XXIV. Results of Chinese-to-Japanese Translation Experiments on EBMT for Use in Phrase Alignment

BLEU	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
Baseline	36.74	31.67	23.90	16.60	15.21	22.84
+Common	37.30	32.06	24.04	16.99	15.40	23.14*
+Common & SE	37.31	32.51	24.11	16.91	15.40	23.22*
Optimized+Common	37.53	32.64	24.83	**17.21**	15.65	23.55*†
Optimized+Dictionary+Common	**37.66**	**32.91**	**24.97**	17.05	**15.67**	**23.62***†

Note: "*" denotes that the "Total" result is better than "Baseline" significantly at $p < 0.01$; "†" denotes that the "Total" result is better than "+Common" significantly at $p < 0.05$.

Table XXV. Results of Japanese-to-Chinese Translation Experiments on EBMT for Use in Phrase Alignment

BLEU	Test set 1	Test set 2	Test set 3	Test set 4	Test set 5	Total
Baseline	37.75	27.46	19.95	15.08	**14.00**	20.74
+Common	37.97	27.98	20.31	14.75	13.92	20.84
+Common & SE	38.16	**28.00**	20.18	15.00	13.83	20.87
Optimized+Common	**38.64**	27.33	20.13	15.20	13.83	20.86
Optimized+Dictionary+Common	38.43	27.80	**20.71**	**15.29**	13.85	**21.03**

in both Chinese word segmentation optimization and phrase alignment. We first performed Chinese word segmentation optimization using the method described in Section 3, and then used the common Chinese characters in phrase alignment based on the optimized segmentation results using the exploitation method described in Section 4. This experimental setting is denoted as "Optimized+Common". We further conducted experiments after appending the transforming word dictionary stored in Section 3.4 to the parallel training corpus, which is denoted as "Optimized+Dictionary+Common".

Tables XXIV and XXV give the BLEU scores for Chinese-to-Japanese and Japanese-to-Chinese translation, respectively. We can see that translation performance also improves after exploiting common Chinese characters in phrase alignment. Exploiting statistically equivalent Chinese characters further slightly improves the translations. However, compared with the baseline system, the improvement by exploiting in phrase alignment is not very significant. There are two possible reasons for this. One is the exception cases for exploiting shared Chinese characters in phrase alignment mentioned in Section 4.6.1, while the other is that both the alignment model and decoder we used suffer from low accuracy of the Chinese parser, which could affect the effectiveness of exploiting shared Chinese characters.

Exploiting common Chinese characters in both Chinese word segmentation optimization and phrase alignment achieves better translation performance than exploiting them separately. We think the reason is due to the alignment improvement by the double effect of both methods. Chinese word segmentation optimization creates a token one-to-one mapping as far as possible between parallel sentences, while exploiting common Chinese characters in phrase alignment enhances the alignment probability of phrase pairs that share Chinese characters. Moreover, translation performance can be further improved by appending the transforming word dictionary obtained from Chinese word segmentation optimization.

4.7. Related Work

Kondrak et al. [2003] incorporated cognate (words or languages with the same origin) information in European languages in the translation models of Brown et al. [1993]. They arbitrarily selected a subset from the Europarl corpus as training data and extracted a list of likely cognate word pairs from the training corpus on the basis of orthographic similarity. The corpus itself was appended to reinforce the co-occurrence count between cognates. The results of experiments conducted on a variety of bitexts showed that cognate identification can improve word alignment without modifying the statistical training algorithm. Common Chinese characters are a kind of cognate, which we exploited together with statistically equivalent Chinese characters in phrase alignment, yielding improved alignment accuracy as well as translation performance.

5. CONCLUSIONS

Shared Chinese characters can be very helpful in Chinese-Japanese MT. In this article, we proposed a method for creating a Chinese character mapping table automatically for Japanese, traditional Chinese, and simplified Chinese using freely available resources, and constructed a more complete resource of common Chinese characters than the existing ones. We also proposed a statistical method to detect statistically equivalent Chinese characters. We exploited shared Chinese characters in Chinese word segmentation optimization and phrase alignment. Experimental results show that our proposed approaches can improve MT performance significantly, thus verifying the effectiveness of using shared Chinese characters in Chinese-Japanese MT.

However, our proposed approaches still have some problems. The proposed method for Chinese word segmentation optimization has problems with incorrect short-unit transformation and POS tag assignment. Our proposed method for exploiting shared Chinese characters in phrase alignment has some drawbacks, and there are also exceptions that can lead to incorrect alignment results. We plan to solve these problems in future work. Furthermore, in this article we only evaluated our proposed approaches on a parallel corpus from an abstract paper domain, in which Chinese characters are more frequently used than in general Japanese domains. In the future, we intend to evaluate the proposed approaches on parallel corpora for other domains.

REFERENCES

Bai, M.-H., Chen, K.-J., and Chang, J. S. 2008. Improving word alignment by adjusting Chinese word segmentation. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 249–256.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Assoc. Comput. Linguist. 19,* 2, 263–312.

Chang, P.-C., Galley, M., and Manning, C. D. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 224–232.

Chen, W., Kawahara, D., Uchimoto, K., Zhang, Y., and Isahara, H. 2008. Dependency parsing with short dependency relation in unlabeled data. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*. 88–94.

Chou, Y.-M. and Huang, C.-R. 2006. Hantology: A linguistic resource for Chinese language processing and studying. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. 587–590.

Chou, Y.-M., Huang, C.-R., and Hong, J.-F. 2008. The extended architecture of Hantology for kanji. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. 1693–1696.

Chu, C., Nakazawa, T., and Kurohashi, S. 2011. Japanese-Chinese phrase alignment using common Chinese characters information. In *Proceedings of the MT Summit XIII*. 475–482.

Chu, C., Nakazawa, T., Kawahara, D., and Kurohashi, S. 2012a. Exploiting shared Chinese characters in Chinese word segmentation optimization for Chinese-Japanese machine translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT'12)*.

Chu, C., Nakazawa, T., and Kurohashi, S. 2012b. Chinese characters mapping table of Japanese, traditional Chinese and simplified Chinese. In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC'12)*.

Chu, C., Nakazawa, T., and Kurohashi, S. 2012c. Japanese-Chinese phrase alignment exploiting shared Chinese characters. In *Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing (NLP'12)*. 143–146.

Collins, M. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1–8.

DeNero, J. and Klein, D. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, 17–24.

Goh, C.-L., Asahara, M., and Matsumoto, Y. 2005. Building a Japanese-Chinese dictionary using kanji/hanzi conversion. In *Proceedings of the International Joint Conference on Natural Language Processing*. 670–681.

Huang, C.-R., Chou, Y.-M., Hotani, C., Chen, S.-Y., and Lin, W.-Y. 2008. Multilingual conceptual access to lexicon based on shared orthography: An ontology-driven study of Chinese and Japanese. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*. 47–54.

Kawahara, D. and Kurohashi, S. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*. Association for Computational Linguistics, 176–183.

Koehn, P., Och, F. J., and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'03)*. 127–133.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Association for Computational Linguistics, 177–180.

Kondrak, G., Marcu, D., and Knight, K. 2003. Cognates can improve statistical translation models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 46–48.

Kudo, T., Yamamoto, K., and Matsumoto, Y. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*. D. Lin and D. Wu Eds., Association for Computational Linguistics, 230–237.

Kurohashi, S., Nakamura, T., Matsumoto, Y., and NAGAO, M. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*. 22–28.

Low, J. K., Tou Ng, H., and Guo, W. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing (SIGHAN'05)*. 161–164.

Ma, Y. and Way, A. 2009. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL'09)*. Association for Computational Linguistics, 549–557.

Nakazawa, T. and Kurohashi, S. 2011a. Bayesian subtree alignment model based on dependency trees. In *Proceedings of the 5th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics.

Nakazawa, T. and Kurohashi, S. 2011b. EBMT system of KYOTO team in PatentMT task at NTCIR-9. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9)*.

Niles, I. and Pease, A. 2001. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems*. ACM Press, 2–9.

Och, F. J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Assoc. Comput. Linguist. 29,* 1, 19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

Peng, F., Feng, F., and McCallum, A. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*. 562–568.

Stolcke, A. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. Vol. 2, 901–904.

Tan, C. L. and Nagao, M. 1995. Automatic alignment of Japanese-Chinese bilingual texts. *IEICE Trans. Inform. Syst. E78-D,* 1, 68–76.

Wang, Y., Uchimoto, K., Kazama, J., Kruengkrai, C., and Torisawa, K. 2010. Adapting Chinese word segmentation for machine translation based on short units. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*. 19–21.

Wang, Y., Kazama, J., Tsuruoka, Y., Chen, W., Zhang, Y., and Torisawa, K. 2011. Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 309–317.

Xia, F., Xue, M. P. N., Okurowski, M. E., Kovarik, J., dong Chiou, F., and Huang, S. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Xu, J., Zens, R., and Ney, H. 2004. Do we need Chinese word segmentation for statistical machine translation? In *Proceedings of the ACL SIGHAN Workshop*. O. Streiter and Q. Lu Eds., Association for Computational Linguistics, 122–128.

Received August 2012; revised February 2013; accepted April 2013