Exploiting Shared Chinese Characters in Chinese Word Segmentation Optimization for Chinese-Japanese Machine Translation

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{chu,nakazawa}@nlp.ist.i.kyoto-u.ac.jp {dk,kuro}@i.kyoto-u.ac.jp

Abstract

Unknown words and word segmentation granularity are two main problems in Chinese word segmentation for Chinese-Japanese Machine Translation (MT). In this paper, we propose an approach of exploiting common Chinese characters shared between Chinese and Japanese in Chinese word segmentation optimization for MT aiming to solve these problems. We augment the system dictionary of a Chinese segmenter by extracting Chinese lexicons from a parallel training corpus. In addition, we adjust the granularity of the training data for the Chinese segmenter to that of Japanese. Experimental results of Chinese-Japanese MT on a phrase-based SMT system show that our approach improves MT performance significantly.

1 Introduction

As there are no explicit word boundary markers in Chinese, word segmentation is considered as an important first step in MT. Studies showed that a MT system with Chinese word segmentation outperforms the one treating each Chinese character as a single word, and the quality of Chinese word segmentation affects the MT performance (Xu et al., 2004; Chang et al., 2008). 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 language, therefore, a bilingual approach is necessary (Ma and Way, 2009).

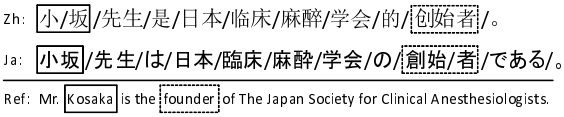

Figure 1: Example of Chinese word segmentation problems in Chinese-Japanese MT.

Most studies focus on language pairs between Chinese and other languages that have white spaces between words (e.g. English). We focus on Chinese-Japanese MT, where segmentation is needed for both sides. Segmentation for Japanese successfully achieves F-score 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 for Chinese, and keep the Japanese segmentation results.

Similar to the previous works, we think the following two problems of Chinese word segmentation are important for Chinese-Japanese MT. The first problem is unknown words, which is the major difficulty faced by a Chinese segmenter affecting segmentation accuracy and consistency. Taking "Kosaka" in Figure 1 as an example, which is a proper noun in Japanese. Because "Kosaka" is a unknown word for the Chinese segmenter, it is mistakenly segmented into two tokens, while the Japanese word segmentation result is correct.

The second problem is word segmentation granularity. Most Chinese segmenters adapt the famous Penn Chinese Treebank (CTB) standard (Xia et al., 2000), while most Japanese segmenters adapt a shorter unit standard. Therefore, the segmentation unit in Chinese may be longer than Japanese even for the same concept. This can increase the number of 1-to-n alignments which makes the word alignment task more difficult. Taking "founder"

Meaning	snow	love	begin
TC	雪(U+96EA)	愛(U+611B)	發(U+767C)
SC	雪(U+96EA)	爱(U+7231)	发(U+53D1)
Kanji	雪(U+96EA)	愛(U+611B)	発(U+767A)

Table 1: Examples of common Chinese characters (TC denotes Traditional Chinese and SC denotes Simplified Chinese).

in Figure 1 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 the above problems, we propose an approach based on a bilingual perspective, and exploit common Chinese characters shared between Chinese and Japanese in Chinese word segmentation optimization for MT. We extract Chinese lexicons from a parallel training corpus based on common Chinese characters to augment the system dictionary of a Chinese segmenter. In addition, we adjust the granularity of the training data for the Chinese segmenter to that of Japanese by means of extracted Chinese lexicons. We conducted experiments on Chinese-Japanese MT tasks using a phrase-based SMT system, and experimental results indicate that our approach can improve MT performance significantly.

2 Common Chinese Characters

Different from other language pairs, Chinese and Japanese share Chinese characters. In Chinese the 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 Macao). The number of strokes needed to write characters has been largely reduced in Simplified Chinese, and the shapes may be different from the ones in Traditional Chinese. Because Kanji characters originated from ancient China, many common Chinese characters exist between Hanzi and Kanji. Table 1 gives some examples of common Chinese characters in Traditional Chinese, Simplified Chinese and Japanese with their Unicode.

Chinese characters contain significant semantic information, and common Chinese characters share the same meaning, so they can be valuable linguistic clues for many Chinese-Japanese NLP tasks. Many studies have been done to exploit common Chinese characters. Tan et al. (1995) used the occurrence of identical common Chinese characters (e.g. "snow" in Table 1) in automatic sentence alignment task. Goh et al. (2005) detected common Chinese characters where Kanji are identical to Traditional Chinese but different from Simplified Chinese (e.g. "love" in Table 1). They used Chinese encoding converter[1] which can convert Traditional Chinese into Simplified Chinese, and built a Japanese-Simplified Chinese dictionary. Chu et al. (2011) made use of the Unihan database[2] to detect common Chinese characters which are visual variants of each other (e.g. "begin" in Table 1), and proved the effectiveness of common Chinese characters in Chinese-Japanese phrase alignment. In this paper, we focus on Simplified Chinese-Japanese MT and exploit common Chinese characters in Chinese word segmentation optimization.

3 Chinese Word Segmentation Optimization

3.1 Chinese Lexicons Extraction

We extract Chinese lexicons from a parallel training corpus through the following steps:

- Step 1: Segment Chinese and Japanese sentences in the parallel training corpus.
- Step 2: Convert Japanese tokens which are made up of Kanji only[3] into Simplified Chinese using the Kanji to Hanzi conversion method described in (Chu et al., 2011).
- Step 3: Extract the converted tokens as Chinese lexicons if they exist in the corresponding Chinese sentence. Here, we propose two extraction strategies:
  - Strategy 1: Only extract tokens which have a different word boundary in the segmented Chinese sentence.
  - Strategy 2: Extract all tokens.

For example, using Strategy 1, "小坂(Kosaka)", "创始(found)" and "者(person)" in Figure 1 are extracted, but using Strategy 2, "先生(Mr.)", "日本(Japan)", "临床(clinical)", "麻醉(anesthesia)" and "学会(society)" are also extracted. Note that although "创始↔創始(found)", "临床↔臨

[1] http://www.mandarintools.com/zhcode.html
[2] http://unicode.org/charts/unihan.html
[3] Japanese has several kinds of character types other than Kanji.

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

Table 2: Chinese-Japanese POS tags mapping table.

床(clinical)" and "麻 醉↔麻 酔(anesthesia)" are not identical, because "创↔創(create)", "临↔臨(arrive)" and "醉↔酔(drunk)" are common Chinese characters, "創始(found)" is converted into "创始(found)", "臨床(clinical)" is converted into "临床(clinical)" and "麻酔(anesthesia)" is converted into "麻醉(anesthesia)" in Step 2.

In preliminary experiments, we extracted 14,359 lexicons using Strategy 1, and 18,584 lexicons using Strategy 2 from a paper abstract parallel corpus containing 680K sentence pairs.

3.2 Chinese Lexicons Incorporation

Several studies showed that using a system dictionary is helpful for Chinese word segmentation (Low et al., 2005; Wang et al., 2011). Therefore, we use a corpus-based Chinese word segmentation and POS tagging tool with a system dictionary. We incorporate the extracted lexicons into the system dictionary. The extracted lexicons are not only effective for the unknown word problem, but also helpful to solve the word segmentation granularity problem.

However, setting POS tags for the extracted lexicons is problematic. To solve this problem, we made a POS tags mapping table between Chinese and Japanese by hand. For Chinese, we use the POS tagset used in CTB which is also used in our Chinese segmenter. For Japanese, we use the POS tagset defined in the morphological analyzer JUMAN (Kurohashi et al., 1994). JUMAN adapts 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)" etc. Table 2 shows a part of the Chinese-Japanese POS tags mapping table we made, the sub POS tags of JUMAN are written inside of the brackets.

We assign POS tags for the extracted Chinese lexicons 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. For the ones that cannot be converted, we do not incorporate them into the system dictionary. In preliminary experiments, 294 lexicons in Strategy 1 and 1,581 lexicons in Strategy 2 were discarded.

3.3 Short Unit Transformation

Bai et al. (2008) showed that adjusting Chinese word segmentation to make tokens 1-to-1 mapping as many as possible between a 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, which can reduce the number of 1-to-n alignments and improve MT performance.

Here, we propose a method to transform the annotated training data of Chinese segmenter into Japanese word segmentation standard using the extracted Chinese lexicons, and use the transformed data for training the Chinese segmenter. Because the extracted lexicons are derived from Japanese word segmentation results, they follow Japanese

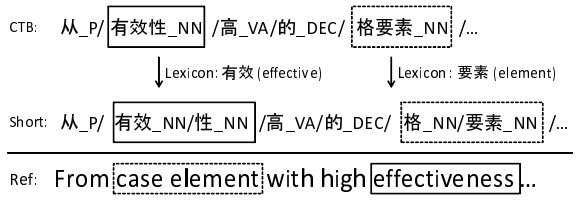

Figure 2: Example of short unit transformation.

word segmentation standard. Therefore, we utilize these lexicons for short unit transformation. We use Chinese lexicons extracted by Strategy 2 described in Section 3.1 and modify every token in the training data for the Chinese segmenter. If the token is longer than a extracted lexicon, we simply split it. Figure 2 gives an example of this process, where "有效(effective)" and "要素(element)" are both extracted lexicons. Because "有效性(effectiveness)" is longer than "有效(effective)", it is split into "有效(effective)" and "性" (a noun suffix), and "格要素(case element)" is longer than "要素(element)", it is split into "格(case)" and "要素(element)". For POS tags, we keep the originally annotated one for the split tokens.

We do not use extracted lexicons that are composed of only one Chinese character, because these lexicons 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 also can use "歌(song)" to construct other words by combining it with other Chinese characters, such as "歌颂(praise)", "诗歌(poem)" etc. Obviously, splitting "歌颂(praise)" into "歌(song)" and "颂(eulogy)", or splitting "诗歌(poem)" into "诗(poem)" and "歌(song)" is undesirable. Also, there are few consecutive tokens in the training data that can be combined to one extracted lexicon, we do not consider this pattern.

4 Experiments

We conducted Chinese-Japanese translation experiments to show the effectiveness of exploiting common Chinese characters in Chinese word segmentation optimization.

4.1 Settings

4.1.1 Parallel Training Corpus

The parallel training corpus we used is a paper abstract corpus provided by JST[4] and NICT[5]. This corpus was created by the Japanese project "Development and Research of Chinese-Japanese Natural Language Processing Technology". The statistics of this corpora are shown in Table 3.

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

Table 3: Statistics of Chinese-Japanese training corpus.

4.1.2 Chinese Annotated Corpus

We used two types of manually annotated Chinese corpus for training the Chinese segmenter. One is NICT Chinese Treebank, which is from the same domain as the parallel training corpus and contains 9,792 sentences. Note that the annotated sentences in this corpus are not included in the parallel training corpus. The other corpus is CTB 7 (LDC2010T07)[6]. We made the training data from CTB 7 using the same method described in (Wang et al., 2011), and it contains 31,131 sentences.

4.1.3 Chinese and Japanese Segmenters

For Chinese, we used a corpus-based word segmentation and POS tagging tool with a system dictionary, weights for the lexicons in the system dictionary are automatically learned from the training data using averaged structured perceptron (Collins, 2002). For Japanese, we used JUMAN (Kurohashi et al., 1994).

4.1.4 SMT Model

We used the state-of-the-art phrase-based SMT toolkit Moses (Koehn et al., 2007) with default options, except for the distortion limit (6→20). It was tuned by MERT using another 500 development sentence pairs.

4.1.5 Test Sets

We translated 5 test sets of Chinese sentences from the same domain as the parallel training corpus. The statistics of the test sets are shown in Table 4. Note that all sentences in the test sets are not included in the parallel training corpus.

4.2 Results and Evaluation

We conducted Chinese-Japanese translation experiments on NICT Chinese Treebank and CTB 7,

[4]http://www.jst.go.jp
[5]http://www.nict.go.jp/
[6]http://www.ldc.upenn.edu/

	T1	T2	T3	T4	T5
# sentences	255	336	391	395	393
# words	6.5K	8.7K	10.0K	11.7K	16.5K
# CC	8.6K	10.6K	12.9K	15.8K	22.0K
avg. sen. len.	44.9	47.0	45.4	52.2	74.1

Table 4: Statistics of test sets (T denotes test set and CC denotes Chinese characters).

comparing the following four experimental settings:

- Baseline: Only using the lexicons extracted from Chinese annotated corpus as the system dictionary for the Chinese segmenter.
- Strategy 1: Incorporate the Chinese lexicons extracted by Strategy 1 described in Section 3.1 into the system dictionary.
- Strategy 2: Incorporate the Chinese lexicons extracted by Strategy 2 described in Section 3.1 into the system dictionary.
- Short unit: Incorporate the Chinese lexicons extracted by Strategy 2 into the system dictionary and train the Chinese segmenter on the short unit training data transformed in Section 3.3.

Table 5 shows the BLEU scores for Chinese-to-Japanese translation using NICT Chinese Treebank. Short unit achieved best MT performance. The extracted Chinese lexicons also improved BLEU scores significantly. Besides test set 2 and test set 5, Strategy 2 achieved better improvement than Strategy 1. We think the reason is that Strategy 2 extracted more lexicons which is helpful to solve the unknown word problem.

Table 6 shows the BLEU scores for Chinese-to-Japanese translation using CTB 7. Although Strategy 2 obtained higher BLEU scores than the baseline, compared to Strategy 1, the improvement is not significant. We investigated the reason and found that there are many overlaps between lexicons extracted from the parallel training corpus and lexicons extracted from the annotated training data. For example, "蛋白质(protein)" was extracted from the annotated training data and overlaps "蛋白(protein)" and "质(quality)" extracted from the parallel training corpus. When the Chinese segmenter tries to segment "蛋白质(protein)", the overlap can lead to inconsistent segmentation results. Although more extracted Chinese lexicons

BLEU	T1	T2	T3	T4	T5
baseline	48.86	47.09	37.18	27.21	24.29
strategy 1	50.41	48.22	39.25	28.33	26.44
strategy 2	50.77	47.96	39.83	28.54	26.29
short unit	**52.04**	**49.55**	**39.96**	**28.73**	**26.63**

Table 5: Results of Chinese-to-Japanese translation experiments using NICT Chinese Treebank.

BLEU	T1	T2	T3	T4	T5
baseline	51.03	48.98	40.52	29.20	26.08
strategy 1	52.42	**51.78**	41.20	30.61	28.20
strategy 2	51.53	50.47	41.30	29.57	26.77
short unit	**52.83**	51.13	**41.57**	**31.01**	**28.82**

Table 6: Results of Chinese-to-Japanese translation experiments using CTB 7.

is more helpful to solve the unknown word problem, it also leads to more overlaps. Because Strategy 2 extracted more lexicons than Strategy 1, more overlaps are also produced. We investigated the number of overlaps. For CTB 7, the overlap number between Strategy 2 is 2,399, it greatly exceeds the number between Strategy 1 which is 1,388. While for NICT, the overlap number between Strategy 2 is 1,759, and between Strategy 1 is 1,694, the difference is not significant. In brief, there is a tradeoff between the unknown word problem and the overlap problem using our proposed method. However, by short unit transformation, the overlap problem can be solved. Taking the same example "蛋白质(protein)", because it is split into "蛋白(protein)" and "质(quality)" in short unit transformation, overlaps will not exist any more. Therefore, short unit using CTB 7 also showed the best MT performance.

Comparing Table 5 with Table 6, we notice that the BLEU scores using NICT Chinese Treebank are lower than using CTB 7. We think the reason is the size of the training data. The number of annotated sentences in NICT Chinese Treebank is less than 1/3 of CTB 7. Therefore, less lexicons are extracted from NICT Chinese Treebank than CTB 7. The number of extracted lexicons from NICT Chinese Treebank is only 13,471, while from CTB 7 it is 26,202. Also, the weights for many lexicons extracted from the parallel training corpus can not be learned correctly using NICT Chinese Treebank as training data. However, short unit using NIC-

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/。
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/.

Short unit (BLEU=56.33)
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/。
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.)

Figure 3: Example of translation improvement.

T Chinese Treebank still achieved even better MT performance than the baseline using CTB 7.

We also conducted Japanese-to-Chinese translation experiments. Results show that our proposed approach also can improve the MT performance. However, compared to Chinese-to-Japanese translation, the improvement is not significant. We think the reason is the input sentence. For Chinese-to-Japanese translation, the segmentation of input Chinese sentences has been optimized. While for Japanese-to-Chinese translation, our proposed approach does not change the segmentation results of input Japanese sentences.

4.3 Discussion

4.3.1 Changes in Vocabulary and Phrase Table Size

We compared the Chinese vocabulary and phrase table size changes before and after exploiting common Chinese characters in Chinese word segmentation optimization. Table 7 shows the comparison results using NICT Chinese Treebank and CTB 7. The decrease of Chinese vocabulary size after optimization indicates the improvement of Chinese segmentation consistency, while the increase of phrase table size after optimization means the increase of translation knowledge.

4.3.2 Short Unit Effectiveness

Experimental results indicate that our proposed approach can improve MT performance significantly, especially for short unit. We present one example to show the effectiveness of short unit.

	vocabulary		phrase table	
	NICT	CTB 7	NICT	CTB 7
baseline	653K	509K	848M	861M
strategy 1	523K	439K	859M	867M
strategy 2	527K	438K	858M	868M
short unit	461K	396K	881M	896M

Table 7: Comparison of vocabulary and phrase table size changes before and after optimization.

Figure 3 shows an example of translation improvement by short unit compared to the baseline. The difference between short unit and the baseline is whether “适应性(suitability)” is split in Chinese or not, while the Japanese segmenter splits it. By splitting it, short unit improves word alignment and phrase extraction which eventually effects the decoding process. In decoding, short unit treats “功能适应性(functional suitability)” as one phrase, while the baseline separates it leading to a undesirable translation result.

4.3.3 Short Unit Transformation Percentage

One encouraging result is that, although the Chinese lexicons used for short unit transformation were extracted from a paper abstract domain corpus which is not the same domain that CTB 7 belongs to, short unit still achieved significant MT performance improvement using CTB 7. To identify the reason, we investigated the percentage of transformed tokens. In NICT Chinese Treebank, there are 6,623 tokens out of 257,825 been transformed to 13,469 short unit tokens, the percentage is about 2.57%. In CTB 7, there are 19,983 token-

s out of 718,716 been transformed to 41,336 short unit tokens, the percentage is about 2.78%. This result shows the strength of our proposed short unit transformation method. Although the lexicons used for short unit transformation are extracted from a paper abstract domain, these lexicons also work well for short unit transformation on Chinese annotated corpus of other domains (i.e. CTB 7).

4.3.4 Short Unit Transformation Problems

Furthermore, we investigated the details of the transformed tokens. Based on our manual investigation, over 90% of the transformed results are correct. However, some transformation problems still exist. One problem is transformation ambiguity. We present one example to show this kind of problem. There is a long token "充电器(charger)" in the annotated training data, and a lexicon "电器(electric equipment)" extracted from the parallel training corpus, so the long token is split into "充(charge)" and "电器(electric equipment)", which is undesirable. However, we found that a extracted lexicon "充电(charge)" also exists and using this lexicon the long token can be split into "充电(charge)" and "器(device)" successfully. We think this kind of ambiguity can be solved using a statistical method.

Another problem is POS tag assignment for the transformed short unit tokens. Our proposed method simply keep the originally annotated POS tag of the long token for the transformed short unit tokens, it 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 a lexicon "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 assigned to NN, which is undesirable for "被(be)". The correct POS tag for "被(be)" should be LB. We think a external dictionary would be helpful to solve this problem. Furthermore, the transformed short unit tokens may have more than one possible POS tags. All these problems are future work of this study.

5 Related Work

Exploiting lexicons from external resources (Peng et al., 2004; Chang et al., 2008) is a way to deal with the unknown word problem. However, the external lexicons may not be very efficient for a specific domain. Some studies (Xu et al., 2004; Ma and Way, 2009) used a method of learning a domain specific dictionary from the character-based alignment results of a parallel training corpus, which separate every Chinese character, and consider consecutive Chinese characters as a lexicon in n-to-1 alignment results. Our proposed method differs from previous studies, we obtain a domain specific dictionary by extracting Chinese lexicons directly from a segmented parallel training corpus, making word alignment is unnecessary.

The goal of our proposed short unit transformation method is to make the segmentation results of Chinese and Japanese a 1-to-1 mapping, which can improve alignment accuracy and MT performance. Bai et al. (2008) proposed a method of learning affix rules from a aligned Chinese-English bilingual terminology bank to adjust Chinese word segmentation in the parallel corpus directly aiming to achieve 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 short unit standard, and do 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. The transfer database is constructed using external lexicons, and is manually modified. Our proposed method learns transfer knowledge based on common Chinese characters. Moreover, we do not use external lexicons, and manual work is not needed.

6 Conclusions

In this paper, we pointed out two main problems in Chinese word segmentation for Chinese-Japanese MT, namely unknown words and word segmentation granularity. To solve the problems, we proposed an approach of exploiting common Chinese characters shared in Chinese and Japanese. Common Chinese characters have been successfully exploited in many Chinese-Japanese NLP tasks, we exploited them in Chinese word segmentation optimization for MT in this study. Experimental results of Chinese-Japanese MT on a phrase-based SMT system indicated that our approach can improve MT performance significantly.

However, there are still some problems in our proposed short unit transformation method. We plan to solve these problems to further improve MT performance. Furthermore, we only evaluated our proposed approach on a parallel corpus from abstract paper domain, where Chinese characters are more frequently used than general domains in Japanese. In the future, we plan to evaluate the proposed approach on parallel corpus of other domains.

References

Bai, Ming-Hong, Keh-Jiann Chen, and Jason S.Chang. 2008. Improving word alignment by adjusting chinese word segmentation. In *Proceedings of the Third International Joint Conference on Natural Language Processing*, pages 249–256, Hyderabad, India, January. Association for Computational Linguistics.

Chang, Pi-Chuan, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio, June. Association for Computational Linguistics.

Chu, Chenhui, Toshiaki Nakazawa, and Sadao Kurohashi. 2011. Japanese-chinese phrase alignment using common chinese characters information. In *Proceedings of MT Summit XIII*, pages 475–482, Xiamen, China, September.

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

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

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In Lin, Dekang and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Kurohashi, Sadao, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of the International Workshop on Sharable Natural Language*, pages 22–28.

Low, Jin Kiat, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to chinese word segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing (SIGHAN05)*, pages 161–164.

Ma, Yanjun and Andy Way. 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 2009)*, pages 549–557, Athens, Greece, March. Association for Computational Linguistics.

Peng, Fuchun, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of Coling 2004*, pages 562–568, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Tan, Chew Lim and Makoto Nagao. 1995. Automatic alignment of Japanese-Chinese bilingual texts. *IEICE Transactions on Information and Systems*, E78-D(1):68–76.

Wang, Yiou, Kiyotaka Uchimoto, Junichi Kazama, Canasai Kruengkrai, and Kentaro Torisawa. 2010. Adapting chinese word segmentation for machine translation based on short units. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may.

Wang, Yiou, Jun'ichi Kazama, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, and Kentaro Torisawa. 2011. Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Xia, Fei, Martha Palmerand Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu dong Chiou, and Shizhe Huang. 2000. Developing guidelines and ensuring consistency for chinese text annotation. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece.

Xu, Jia, Richard Zens, and Hermann Ney. 2004. Do we need chinese word segmentation for statistical machine translation? In Streiter, Oliver and Qin Lu, editors, *ACL SIGHAN Workshop 2004*, pages 122–128, Barcelona, Spain, July. Association for Computational Linguistics.