Japanese-Chinese Phrase Alignment Exploiting Shared Chinese Characters

Chenhui Chu, Toshiaki Nakazawa and 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 kuro@i.kyoto-u.ac.jp

Abstract

Common Chinese characters between Japanese and Chinese have been proved to be effective in Japanese-Chinese phrase alignment. Besides common Chinese characters, Japanese and Chinese also share many other semantically equivalent Chinese characters. However, there are no available resources for this kind of Chinese characters. In this paper, we propose a statistical method aiming to detect these Chinese characters which we call statistically equivalent Chinese characters. We exploit statistically equivalent Chinese characters together with common Chinese characters in a joint phrase alignment model. Experimental results show that our approach achieves over 1 point lower AER and 1 BLEU increase comparing to the baseline system.

1 Introduction

Different from other language pairs, Japanese and Chinese share Chinese characters. In Japanese the Chinese characters are called Kanji, while in Chinese they are called Hanzi. 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 are originated from ancient China, there exist common Chinese characters between Kanji and Hanzi. Table 1 gives some examples of common Chinese Characters in Japanese, Traditional Chinese, Simplified Chinese and their Unicode.

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

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

Meaning	eat	word	hide	look	day	
Kanji	食	語	隠	見	日	...
SC	吃	词	藏	看	天	...

Table 2: Examples of other semantically equivalent Chinese characters

Chinese characters contain significant semantic information, and common Chinese characters share the same meaning, so they can be valuable linguistic clues for many Japanese-Chinese 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 when Kanji are identical to Traditional Chinese but different from Simplified Chinese (e.g. "love" in Table 1) using Chinese encoding converter [1] which can convert Traditional Chinese into Simplified Chinese, and built a Japanese-Simplified Chinese dictionary partly using direct conversion of Japanese into

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

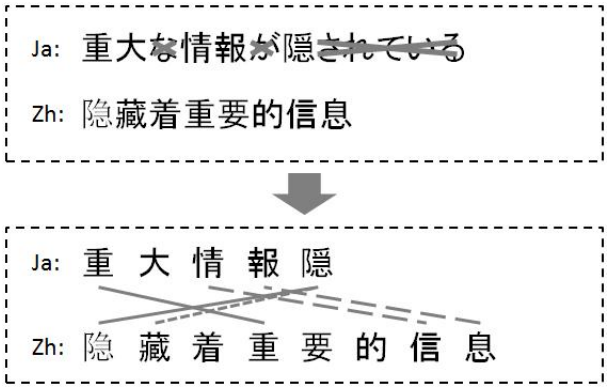

Figure 1: Character-based alignment.

Chinese for Japanese Kanji words. 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 Japanese-Chinese Phrase Alignment.

Besides common Chinese characters, there also exist many other semantically equivalent Chinese characters between Japanese and Chinese. Table 2 gives some examples of this kind of Chinese characters between Japanese and Simplified Chinese. Although these Chinese characters are not common Chinese characters, they share the same meaning. We think that these Chinese characters again would be valuable in machine translation, especially in word/phrase alignment. However, there are no available resources for this kind of Chinese characters. In this paper, we propose a statistical method aiming to detect these Chinese characters, which we call statistically equivalent Chinese characters. In addition, we exploit statistically equivalent Chinese characters together with common Chinese characters in a joint phrase alignment model.

2 Statistically Equivalent Chinese Characters Detection

Figure 1 shows the basic idea of our statistically equivalent Chinese characters detecting method. The example parallel sentences (both mean "important 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. We treat every Chinese character as a single word and do character-based alignment using GIZA++ (Och and Ney, 2003) which implements sequential word-based statistical alignment model of IBM models.

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

f_i	e_j	$t(e_j\|f_i)$	$t(f_i\|e_j)$
隠	隐	0.287043	0.352356
重	重	0.572420	0.797318
隠	藏	0.122787	0.006287
情	信	0.796714	0.634998
報	息	0.590478	0.981210

Table 3: Examples of lexical translation probability estimated by character-based alignment

Table 3 shows examples of lexical translation probability estimated by character-based alignment on a Japanese-Chinese paper abstract corpus. We can see that shared Chinese characters obtained high lexical translation probability. 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 high lexical translation probability. Although these two pairs are not semantically equivalent, we think this kind of shared Chinese characters would be valuable clues too.

3 Exploiting Shared Chinese Characters

We use 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:

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

where $P(\ell)$ is a geometric distribution stands for the number of concepts that generate phrase pairs, $P(D|\{\langle e,f\rangle\})$ is the dependency relation probability of phrases, $\theta_T(\langle e,f\rangle)$ is a distribution that phrase generation step obeys. We skip the detail of the model here.

We define shared Chinese characters matching ra-

tio for Japanese-Chinese phrase pairs:

$$ratio = \frac{match\_ja\_char + match\_zh\_char}{num\_ja\_char + num\_zh\_char}, \quad (2)$$

where num_ja_char and num_zh_char denote number of Chinese characters in Japanese and Chinese phrase respectively, $match\_ja\_char$ and $match\_zh\_char$ are matching weight of Chinese characters in Japanese and Chinese phrase 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 Chinese characters pair estimated in Section 2. Taking "情報局" and "信息局" (both mean "information agency") as an example, there are one common Chinese character "局(agency)" and two statistically equivalent Chinese characters pairs, thus

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

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

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

the weight is proportional to the shared Chinese characters matching ratio:

$$w = ratio \cdot W, \quad (4)$$

where W is a variable set by hand.

4 Experiments

4.1 Alignment

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

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

[3]http://www.jst.go.jp

[4]http://www.nict.go.jp/

	Ja	Zh
# of sentences	680k	
# of words	21.8M	18.2M
# of Chinese characters	14.0M	24.2M
ave. sen. length	32.9	22.7

Table 4: Statistics of the Japanese-Chinese corpus.

As gold-standard data, we used 510 sentence pairs for Japanese-Chinese which were annotated by hand. There are two types of annotations, sure (S) alignments and possible (P) alignments (Och and Ney, 2003). The unit of evaluation was 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 W to 5000, which showed the best performance in the preliminary experiments for tuning the weight.

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). Chinese sentences were converted into dependency trees using the word segmentation and POS-tagging tool by Canasai et al. (2009) and the dependency analyzer CNP (Chen et al., 2008).

For comparison, we used GIZA++ and conducted word alignment bidirectionally with its default parameters and merged them using grow-diag-final-and heuristic (Koehn et al., 2003). Also, we used BerkelyAligner[5] (DeNero and Klein, 2007) with its default settings for unsupervised training. Experimental results are shown in Table 5. The alignment accuracy of Bayesian subtree alignment model is indicated as "Baseline", the alignment accuracy after exploiting common Chinese characters is indicated as "+Common", and the alignment accuracy after exploiting both statistically equivalent and common Chinese characters is indicated as "+Statistically equivalent". Alignment accuracy is further improved by exploiting statistically equivalent Chinese characters.

4.2 Translation

We conducted Japanese-Chinese translation experiments on the same corpus used in the alignment

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

	Pre.	Rec.	AER
grow-diag-final-and	83.77	75.38	20.39
BerkelyAligner	**88.43**	69.77	21.60
Baseline	85.37	75.24	19.66
+Common	85.55	76.54	18.90
+Statistically equivalent	85.22	**77.31**	**18.65**

Table 5: Results of Japanese-Chinese alignment experiments.

	BLEU
Baseline	23.16
+Common	23.65
+Statistically equivalent	**24.25**

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

experiment. We translated 333 paper abstract sentences from the JST corpus. Note that these sentences were not included in the training corpus. We used an example-based machine translation (EBMT) system (Nakazawa and Kurohashi, 2011b) which is a dependency tree-based decoder. Table 6 shows the BLEU scores for translation. We can see that translation performance is also improved because of the improvement of alignment accuracy.

5 Conclusion

In this paper, we proposed a method for detecting statistically equivalent Chinese characters. We exploited statistically equivalent Chinese characters together with common Chinese characters in a joint phrase alignment model. Our proposed approach achieved over 1 point lower AER as well as 1 BLEU increase comparing to the baseline system, which verified the effectiveness of shared Chinese characters.

References

Wenliang Chen, Daisuke Kawahara, Kiyotaka Uchimoto, Yujie Zhang, and Hitoshi Isahara. 2008. Dependency parsing with short dependency relation in unlabeled data. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 88–94.

Chenhui Chu, 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.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague, Czech Republic, June. Association for Computational Linguistics.

Chooi-Ling Goh, 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.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proceedings of Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA, June. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Procedings*, pages 127–133.

Canasai Kruengkrai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentara Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 513–521, Suntec, Singapore, August. Association for Computational Linguistics.

Sadao Kurohashi, 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.

Toshiaki Nakazawa and Sadao Kurohashi. 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, November.

Toshiaki Nakazawa and Sadao Kurohashi. 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)*, Tokyo, Japan, December.

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

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