Statistical Machine Translation between Unsegmented Japanese and Chinese Texts

Jing Sun and Yves Lepage
Graduate School of IPS, Waseda University
{cecily.sun@akane,yves.lepage@}.waseda.jp

Abstract

Many Asian languages like Japanese and Chinese do not have explicit boundaries between words. Word segmentation is normally treated as the first step for most Natural Language Processing tasks especially for Statistical Machine Translation (SMT). In this paper, we implemented several machine translation experiments on both pre-segmented and unsegmented text corpus. The experiment results showed that word segmentation may not be a prerequisite step for SMT between Japanese and Chinese.

1 Introduction

Many written Asian languages such as Japanese and Chinese do not involve typographic delimiters like white spaces between words, therefore, word segmentation is usually the first step in most Natural Language Processing (NLP) tasks especially in Statistical Machine Translation (SMT). Word segmentation techniques for both Japanese and Chinese have achieved great success in recent years. However, word segmentation schemes shall not be treated as system-independent, application-independent nor language-independent.

To demonstrate the inconsistency of word segmentation, we applied four state-of-the-art Chinese Word Segmentation (CWS) tools[1] on one single Chinese sentence 没事先约好,白跑了回津屋崎。[2] and achieved four different segmentation results consequently:

ICTCLAS: 没_事先_约_好_,_白_跑_了_回_津_屋_崎_。

Stanford-C: 没事_先_约_好_,_白_跑_了_回津屋崎_。

Stanford-P: 没_事先_约好_,_白_跑_了_回_津_屋崎_。

Urheen: 没_事先_约_好_,_白_跑_了_回津_屋崎_。

None of these results is consistent with the human-segmented reference, 没_事先_约好_,_白_跑了回_津屋崎_。. Across Japanese and Chinese, althouth 津屋崎 (Tsuyazaki) is one word in Japanese which refers to the name of a place, it was decomposed into different units in segmented Chinese. This example clearly shows that different word segmentation tools, or same word segmentation tool that trained on different pre-defined dictionaries may cause inconsistencies across languages, such as different sizes of granularity in Japanese and Chinese. Such inconsistencies lead to increased error rates in statistical machine translation.

In this paper, we used a Japanese-Chinese bilingual corpus to perform phrase table extraction and conducted statistical machine translation experiment without performing word segmentation on either Japanese nor Chinese beforehand. The rest of this paper is organized as follows. Section 2 introduces our proposed method in using the sampling-based sub-sentential aligner, Anymalign to extract Japanese-Chinese sub-sentential fragments, i.e., phrase translation tables from unsegmented bi-corpus. Section 3 describes the machine translation experiment that uses the phrase tables produced by our method and gives an evaluation of the translation quality. A conclusion is given in Section 4.

2 Producing Phrase Translation Tables

We used an in-house Japanese-Chinese bi-corpus which includes 48,461 sentence pairs collected from the Internet. Contents include bilingual Web-blogs, films transcriptions, fable stories and conversations. Table 1 gives a detailed description.

	Japanese	Chinese
Sentences	48,461	48,461
Average length (word)	9 (±4.87)	7 (±3.73)
Average length (character)	16 (±8.48)	10 (±5.12)

Table 1: Statistics of the training corpus.

In order to compare the performance of phrasal extraction from both pre-segmented and unsegmented corpus, we also conducted word segmentation on the same data set. Juman [7] is used to perform Japanese word segmentation (JWS) and Urheen [13] is used for CWS.

[1] Urheen [13], ICTCLAS [14] and Standford Chinese word segmenter [12] trained on Chinese Treebank and PKU Treebank.

[2] Literary meaning in English: I went to Tsuyazaki in vain without prior appointment. Literary translation in Japan: 事前予約をしなかったので、むたに津屋崎に行きました。

2.1 The Treatment of Katakana

Along with Kanji and Hiragana, Katakana syllabary is one component of the Japanese writing system. In modern Japanese, Katakana is usually used for foreign words transcriptions, such as words imported from Chinese (also known as 'Chinese loanwords'). Moreover, Katakana is also used for country names, foreign places and names, onomatopoeia and technical terms. Few examples are shown in Table 2.

Genre	Katakana	English Meaning
Foreign place	アメリカ	America
Onomatopoeia	ドキドキ	heart beating
Company name	トヨタ	TOYOTA
Chinese loanword	シューマイ	one dim sum
English loanword	コーヒー	coffee
Technical term	ソフト	software

Table 2: Some examples of Japanese Katakana.

Inspired by Baldwin and Tanaka's work [1], we bounded all adjacent Katakana in unsegmented Japanese text corpus and treat each consecutive Katakana string as one "word" or "unigram" using a Katakana list. The Kakakana list inlcudes syllabograms like ア イ ウ, small version of kanataka like ャ ュ ョ, sokuon ッ, long vowel ー and iteration marks like ヽ and ヾ. The Japanese part in unsegmented text corpus is pre-processed as follows.

- 商_品_コ_ー_ド (product code) ⇒ 商_品_コード
- エ_デ_ィ_ア_カ_ラ_化_石_群 (Ediacara biota) ⇒ エディアカラ_化_石_群
- ウ_ー_ロ_ン_茶 (Oolong tea) ⇒ ウーロン_茶

Out of 48,461 Japanese sentences, 5,740 (11.84%) sentences are involved in Katakana-bounding.

2.2 Anymalign Option -*i*

An open source sampling-based approach Anymalign [5][3] is used to perform sub-sentential extractions. For each index task, Anymalign was run for three hours with its basic version (Anym b.) and its option -*i* (Anym -*i*). Option -*i* focus Anymalign to consider n-grams up to i ($i > 0$) as tokens. In other words, we expect Anymalign to extract longer n-grams, especially for unsegmented texts with option -*i*. For pre-segmented texts, option -*i* allows to group words into phrases more easily. For unsegmented texts, as a token is a single character, the use of option -*i* allows to group characters into words, and then, into phrases, more easily.

Both Japanese and Chinese word segmentation schemes result in various granularities. In average, a Japanese sentence in our training corpus which has 10 characters might be segmented into 5.527 (±1.144) words[4]. On the other hand, a Chinese sentence which has 10 characters might be segmented into 6.629 (±1.289) words[5]. Consequently, we set i_{max} as 10 for unsegmented corpus and 6 for pre-segmented corpus.

index	Unseg	Unseg +	Pre-seg
$i = 1$	1,556,556	1,818,410	882,342
$i = 2$	1,951,870	2,542,401	1,185,388
$i = 3$	1,665,893	2,218,145	1,063,432
$i = 4$	1,371,507	1,920,950	969,298
$i = 5$	1,177,670	1,725,236	903,474
$i = 6$	1,023,555	1,591,819	856,029
$i = 7$	924,654	1,502,591	-
$i = 8$	903,856	1,523,525	-
$i = 9$	903,078	1,581,863	-
$i = 10$	897,849	1,610,744	-
i-merged	3,917,469	4,941,097	1,708,151
baseline	1,555,438	1,814,457	883,324

Table 3: Numbers of entries in phrase translation tables obtained with Anymalign Baseline and Option -*i* from Unsegmented bi-corpus (Unseg), Unsegmented bi-corpus enhanced by Katakana grouping (Unseg +) and pre-segmented bi-corpus (Pre-seg).

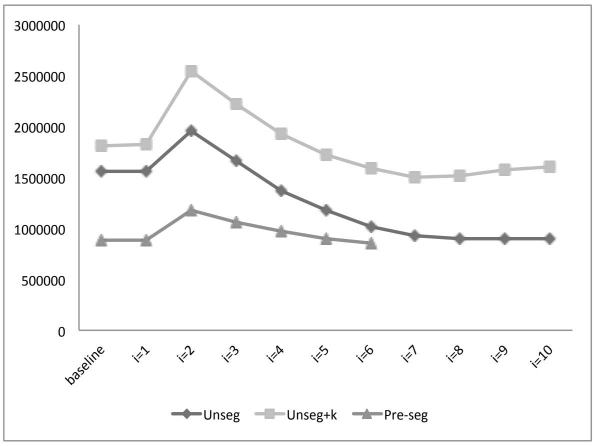

Figure 1: Amount of output entries in phrase tables while Index -*i* varies. This graph plots the figures given in Table 3.

While index i varies, output entries of phrase pairs are also differ which is reflected in Table 3. Figure 1 shows that Anymalign can generate the most number of phrase pairs when i equals to 2 from both unsegmented and pre-segmented corpus. When i reaches 6, the change in the number of entries in the phrase translation table reaches its asymptote.

All the sub-tables generated with Anymalign Option -*i* are then merged into one table by re-estimating translation probabilities (*i*-merged).

The use of an unsegmented corpus leads to larger

[3]Anymalign: `http://perso.limsi.fr/Individu/alardill/anymalign/`

[4]JWS tool Juman is applied here.

[5]CWS tool Urheen is applied here.

phrase translation tables than the use of a pre-segmented corpus, i.e., twice the size for the basic version of Anymalign and three times for the merge of the all results of Anymalign run with option *-i*.

3 Statistical Machine Translation Experiments

In this section, the phrase tables extracted from previous section are utilized for statistical machine translation experiments. Besides previously mentioned 48,461 bi-corpus as training corpus, we used 500 bilingual sentence pairs for tuning and 500 for testing.

3.1 Experiment Setting

The state-of-the-art phrase-based machine translation system Moses [4] is applied to perform our machine translation experiments. While running Moses, we used MERT (Minimum Error Rate Training) [9] and SRILM [10][11] for building the language model. In order to compare the performance in phrase extraction from unsegmented Japanese-Chinese bi-corpus, we applied Anymalign baseline version, Anymalign with its option *-i* as well as GIZA++ [8] for obtaining phrasal alignment.

3.2 Evaluation and Results

Four standard automatic metrics are used to evaluate translations results: WER, BLEU, NIST and TER. Besides, we also applied RIBES [3], an automatic evaluation metric that takes account of the word order in evaluation of translation quality for distant language pairs.

BLEU (bilingual evaluation understudy) is the score mostly used for translation evaluation by far for evaluating the precision of N-grams according to a reference translation. However, word-level BLEU metric has been challenged in recent years. Denoual and Lepage [2] studied the equivalence of applying BLEU metics in characters and suggested that the use of BLEU at the character level could eliminate the word segmentation problem. Li et al., [6] stated that character-level BLEU correlates better with human assessment for Chinese tasks. Besides the campaigns like IWSLT '08 and NIST '08 both adopted character-level evaluation metrics.

In this work, we evaluated the quality of Chinese translation output in characters to ensure the consistency. The obtained evaluation results are presented in Table 4-6. $BLEU_{cN}$ stands for the measure in characters for a given order N.

Reflected in above results, SMT experiments that used phrase tables generated from unsegmented bi-corpus (Unseg), especially those enhanced by grouping adjacent Katakana into unigrams (Unseg +) outperformed those from pre-segmented bi-corpus (Pre-seg).

We believe the unsegmented text gives more chances to match with correct alignment in Chinese and Japanese corpus, and pre-processing of Japanese Katakana is promising in improving SMT performance, especially in improving BLEU scores. It achieves an increase of 2.01 points with Anymalign *-i* merge which corresponds to a relative 11.9% increase and 3.76 points with GIZA++ (22.6% increase) than it on pre-segmented bi-corpus.

Eval. Metrics	Anymalign Baseline		
	Pre-seg	Unseg	Unseg +
$BLEU_{c4}$ [%]	16.25	**17.78**	17.45
$BLEU_{c5}$ [%]	12.01	**13.32**	13.11
$BLEU_{c6}$ [%]	9.09	**10.06**	9.94
$BLEU_{c7}$ [%]	7.06	7.65	**7.67**
$BLEU_{c8}$ [%]	5.57	5.81	**5.84**
WER	**0.7305**	0.7439	0.7443
NIST	4.9370	4.9724	**4.9923**
TER	0.7412	0.7417	**0.7379**
RIBES	0.5807	0.5777	**0.5870**

Table 4: Evaluation of Chinese translation output. Aligner used: **Anymalign Baseline**

Eval. Metrics	Anymalign *-i* merge		
	Pre-seg	Unseg	Unseg +
$BLEU_{c4}$ [%]	16.84	18.75	**18.85**
$BLEU_{c5}$ [%]	12.43	13.85	**14.13**
$BLEU_{c6}$ [%]	9.48	10.25	**10.72**
$BLEU_{c7}$ [%]	7.40	7.55	**8.18**
$BLEU_{c8}$ [%]	5.82	5.61	**6.33**
WER	0.7419	**0.7262**	0.7399
NIST	4.9104	**5.2946**	5.2786
TER	0.7482	**0.7119**	0.7133
RIBES	0.5942	**0.6019**	0.5963

Table 5: Evaluation of Chinese translation output. Aligner used: **Anymalign *-i* merge**

Eval. Metrics	GIZA++		
	Pre-seg	Unseg	Unseg +
$BLEU_{c4}$ [%]	16.67	19.99	**20.43**
$BLEU_{c5}$ [%]	12.36	15.48	**15.96**
$BLEU_{c6}$ [%]	9.44	12.24	**12.77**
$BLEU_{c7}$ [%]	7.35	9.84	**10.36**
$BLEU_{c8}$ [%]	5.78	8.04	**8.54**
WER	0.7769	**0.6747**	0.6936
NIST	4.7542	**5.5683**	5.5138
TER	0.7764	**0.6828**	0.6931
RIBES	0.5966	**0.6131**	0.6029

Table 6: Evaluation of Chinese translation output. Aligner used: **GIZA++**

4 Conclusion

In this paper, we showed several SMT experiments with both unsegmented and Pre-segmented Japanese-Chinese parallel corpus. For unsegmented Japanese text corpus, we grouped adjacent Katakana into unigrams according to the linguistic feature of Katakana to enhance phrasal extraction. Our experiment results show that unsegmented method outperforms the pre-segmented ones. We concluded that word segmentation is not necessary for SMT tasks between Japanese and Chinese.

Acknowledgments

This work is supported by the Kitakyushu Foundation for the Advancement of Industry, Science and Technology (FAIS) with Foreign Joint Project funds in part. Associate Professor Francis Bond of Nanyang Technological University is kindly acknowledged for providing us with his insightful comments.

References

[1] Timothy Baldwin and Hozumi Tanaka. The Applications of Unsupervised Learning to Japanese Grapheme-phoneme Alignment. In *Proceedings of the ACL Workshop of Unsupervised Learning in Natural Language Processing*, pages 9–16, College Park, USA, 1999.

[2] Etienne Denoual and Yves Lepage. BLEU in characters: Towards automatic MT evaluation in languages without word delimiters. In *IJCNLP-05: Second International Joint Conference on Natural Language Processing*, pages 79–84, Jeju Island, Republic of Kore, October 2005.

[3] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, MIT, Massachusetts, USA, October 2010.

[4] Philipp Koehn, 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. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 177–180, Prague, Czech Republic, 2007.

[5] Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *International Conference on Recent Advances in Natural Language Processing (RANLP'09)*, pages 214–218, Borovets, Bulgaria, 2009.

[6] Maoxi Li, Chengqing Zong, and Hwee Tou Ng. Automatic evaluation of chinese translation output: Word-level or character-level? In *Proceedings of the 49the Annual Meeting of the Association for Computational Linguistics: shortpapers*, pages 159–164, Portland, Oregon, 2011.

[7] Takashi Masuoka and Yukinori Kabuto. *Basic Japanese Grammar*. Kuroshi Publishers, 1989.

[8] Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, volume 29(1), pages 19–51, 2003.

[9] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51, 2003.

[10] Andreas Stolcke. SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, 2002.

[11] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. Srilm at sixteen: Update and outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii, December 2011.

[12] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168–171, Jeju Island, Korea, 2005.

[13] Kun Wang, Chengqing Zong, and Keh-Yih Su. A character-based joint model for Chinese word segmentation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1173–1181, August 2010.

[14] Huaping Zhang, Qun Liu, Xueqi Cheng, Hao Zhang, and Hongkui Yu. Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pages 63–70, sapporo, Japan, 2003.