# Can Word Segmentation be Considered Harmful for Statistical Machine Translation Tasks between Japanese and Chinese?

**Jing Sun and Yves Lepage**
NLP Laboratory / Hibikino 2-7, Wakamatsu-ku
Graduate School of IPS / Kitakyushu-shi, Fukuoka-ken
Waseda University / Japan 808-0135
{cecily.sun@akane., yves.lepage@}waseda.jp

## Abstract

Unlike most Western languages, there are no typographic boundaries between words in written Japanese and Chinese. Word segmentation is thus normally adopted as an initial step in most natural language processing tasks for these Asian languages. Although word segmentation techniques have improved greatly both theoretically and practically, there still remains some problems to be tackled. In this paper, we present an effective approach in extracting Chinese and Japanese phrases without conducting word segmentation beforehand, using a sampling-based multilingual alignment method. According to our experiments, it is also feasible to train a statistical machine translation system on a small Japanese-Chinese training corpus without performing word segmentation beforehand.

## 1 Introduction

Unlike most European languages, there are no explicit typographic boundaries like white spaces between words in many written Asian languages such as Chinese, Japanese, Korean, Thai, Lao and Vietnamese. Therefore, word segmentation for such languages is usually the first important step in most Natural Language Processing (NLP) applications especially in statistical machine translation. Although word segmentation techniques have improved greatly in recent years, there are still some difficulties that remain to be addressed.

Word segmentation schemes are not system-independent, application-independent nor language-independent. Different Chinese Word Segmentation (CWS) tools applied to the same Chinese sentence may lead to different results depending on their segmentation. For instance, 学生会 (pinyin: xué shēng huì) in Chinese may be interpreted as 学生_会 'student(s) can (do)' or 学生会 'Students' Union' respectively.

Figure 1 gives an example of pre-segmented text and unsegmented text in both Chinese and Japanese. We applied four CWS tools: Urheen (Wang et al., 2010), ICTCLAS (Zhang et al., 2003) and Standford Chinese word segmenter (Tseng et al., 2005) trained on CTB and PKU. This example clearly shows that word segmentation tools may do harm to cross-lingual tasks, because:

(i) there may be inconsistencies of segmentation results across languages such as different sizes of granularity in Japanese and Chinese;

(ii) for the same language, different word segmentation tools may produce different results;

(iii) the same word segmentation tool trained on different corpora may produce different results.

Such inconsistencies lead to increased error rates in Statistical Machine Translation.

Significant improvements in Chinese word segmentation techniques have been obtained recently and reported accuracy rates (compared to those of human *Golden Standard*) have reached 98%. However, for cross-lingual NLP tasks, such as phrasal extraction or Machine Translation, Zhang et al. (2008) showed that even the most accurate word segmentation may not produce the best translation out-




26th Pacific Asia Conference on Language,Information and Computation pages 351–360

| | |
|---|---|
| Original Chinese sentence: | 没**事先**约好，白跑了回**津屋崎**。 |
| Translation in Japanese: | 事前予約をしなかったので、むだに津屋崎に行きました。 |
| Meaning in English: | I went to Tsuyazaki in vain without prior appointment. |
| JWS (JUMAN): | 事前_予約_を_し_なかった_ので_、_むだに_津屋崎_に_行き_ました_。 |
| CWS Reference: | 没_**事先**_约好_，_白_跑了回_**津屋崎**_。 |
| CWS (ICTCLAS): | 没_**事先**_约_好_，_白_跑_了_回_**津**_**屋**_**崎**_。 |
| CWS (STANDFORD-CTB): | 没**事**_**先**_约_好_，_白_跑_了_回**津屋崎**_。 |
| CWS (STANDFORD-PKU): | 没_**事先**_约好_，_白_跑_了_回_**津**_**屋崎**_。 |
| CWS (URHEEN): | 没_**事先**_约_好_，_白_跑_了_回**津**_**屋崎**_。 |

Figure 1: An example of inconsistency in Chinese word segmentation. All segmentation in Chinese by the four different systems are different. In addition, across Japanese and Chinese, although 津屋崎 (Tsuyazaki) is one word in Japanese, it was decomposed into different units in segmented Chinese.

puts. To solve the problem, it has been proposed to drive word segmentation using predefined bilingual knowledge, such as bilingual dictionaries or bilingual lexica extracted from parallel corpora. Instead of relying on an existing bilingual lexicon, Sun et al. (1998) automatically learned rules from a corpus and group unsegmented Chinese segments into words according to their mutual information. Xu et al. (2004) developed a system which extracts a lexicon from the trained alignment corpus. They showed that it is possible to work without performing Chinese word segmentation beforehand with only a minor loss in translation quality.

Bilingual resources are unavailable for many language pairs that do not involve English, like Japanese-Chinese or Japanese-Vietnamese. Although many researchers and several institutions have been working on constructing bilingual resources between Asian languages, rarely are these resources made freely available.

In this paper, we show how to use a small Japanese-Chinese bilingual corpus to perform phrase table extraction so as to build a statistical machine translation system and conduct translation experiments between Chinese and Japanese without conducting word segmentation on either the Japanese nor Chinese sides beforehand. The purpose of this paper is to determine:

- Whether it is possible to produce phrase tables and extract sub-sentential alignments from unsegmented texts in Chinese and Japanese.
- Whether it is possible to perform statistical machine translation with reasonable quality without conducting word segmentation beforehand.

Section 2 introduces our proposed method which consists in using the sampling-based sub-sentential aligner, Anymalign, to extract Japanese-Chinese sub-sentential fragments (phrase translation tables) from an 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 when translating using the character as the basic unit. Section 4 discusses the experiment results and Section 5 gives the conclusion.

## 2 Producing Phrase Tables from Unsegmented Japanese and Chinese Corpus

### 2.1 Text Corpus Used

We start with an in-house corpus of 9,500 aligned Japanese-Chinese sentence pairs collected from the Internet as training data. They include bilingual Web-blogs, movie subtitles, fable stories and conversations.

To compare the performance of phrasal extraction from both the pre-segmented corpus and the unsegmented corpus, we also conduct word segmentation on the same data set. Juman (Masuoka and Kabuto,

1989; Knuth, 2012) and Urheen (Wang et al., 2010) are used to perform Japanese and Chinese word segmentation.

The average length for the unsegmented Japanese sentences are 17 (std. dev. ±9.95) characters and 11 (std. dev. ±7.40) for Chinese. For pre-segmented text corpus, the average length is 10 (std. dev. ±5.93) words for Japanese and 8 (std. dev. ±4.99) for Chinese.

Sentence length distributions in both pre-segmented and unsegmented corpora are shown in Figure 2 and Figure 3 respectively,

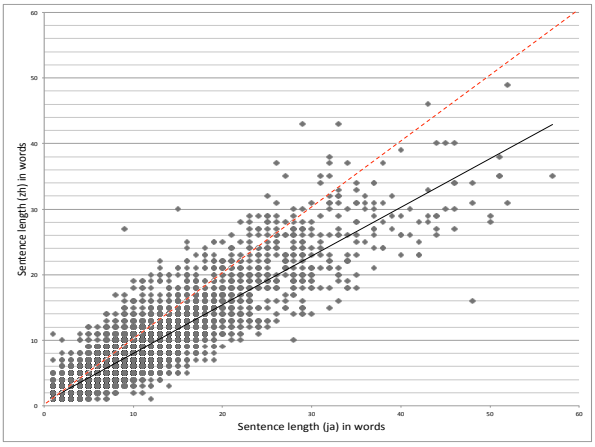


Figure 2: Sentence length distribution in our **pre-segmented** corpus. The dashed line shows the average, the solid line is linear regression.

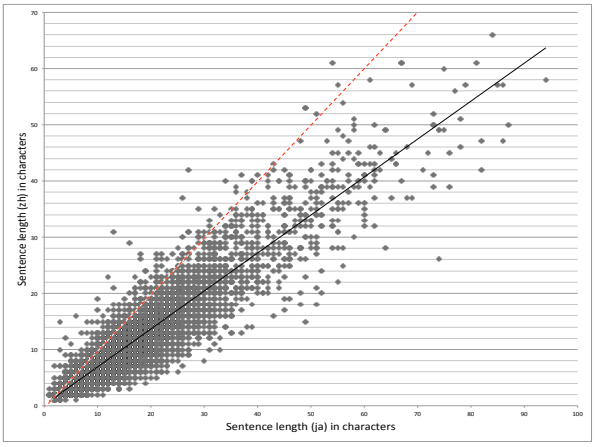


Figure 3: Sentence length distribution in our **unsegmented** corpus. The dashed line shows the average, the solid line is linear regression.

## 2.2 Aligners and Configurations Used

In our experiments, we use the open source implementation of the sampling-based approach, Anymalign (Lardilleux and Lepage, 2009)[1], to perform subsentential extraction from the above-described bi-corpus. Anymalign was run for three hours in its basic version (Anym b.) and with the option *-i* (Anym *-i*), where parameter *i* ranged from 1 to 10. The use of this option allows to extract longer phrases by enforcing n-grams to be considered as tokens. 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.

In order to compare the performance of our phrase extraction method and statistical machine translation with unsegmented text corpus, we also applied GIZA++ (Och and Ney, 2003), the most commonly used tool for word and phrase alignment.

## 2.3 Numbers of Phrase Pairs Produced

Different values of parameter *i* lead to different numbers of phrase pairs entries in the phrase translation tables produced (see Table 1). The highest number of entries is obtained for *i* equal to 2, i.e., when each two connect characters in a sentence are possibly considered as one unit.

| Index *i* | Output Entries |
|---|---|
| 1 | 782,465 |
| 2 | 967,173 |
| 3 | 852,932 |
| 4 | 782,585 |
| 5 | 715,182 |
| 6 | 668,134 |
| 7 | 599,316 |
| 8 | 586,992 |
| 9 | 581,131 |
| 10 | 577,040 |
| *i*-merged | 1,628,241 |

Table 1: Numbers of entries in phrase translation tables obtained with Anymalign option *-i*.

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

| Aligner | Segmentation | Phrase-Table Entries | Intersection | Avg. $P_{EDR}$ | Avg. $P_{table}$ | Score |
|---|---|---|---|---|---|---|
| GIZA++ | Pre-seg | 36,888 | 1,086 | 0.6237 | 0.8269 | 1,575.323 |
| | Unseg | 56,002 | **1,954** | 0.6128 | 0.7804 | **2,709.9344** |
| Anym b. | Pre-seg | 326,748 | 2,190 | 0.5872 | 0.5841 | 2,565.0188 |
| | Unseg | 784,004 | **3,294** | 0.5141 | 0.2975 | **2,673.4151** |
| *i*-merge | Pre-seg | 553,156 | 2,265 | 0.5863 | 0.5850 | 2,652.968 |
| | Unseg | 1,628,241 | **3,643** | 0.5122 | 0.3909 | **3,290.2923** |

Table 2: Size of the intersection of phrase translation tables with the EDR Chinese-Japanese lexicon.

Figure 4 shows that when *i* reaches 7, the decrease in the number of entries in the phrase translation table reaches its asymptote. We also merged the 10 phrase translation tables for each value of parameter *i* into one phrase translation table that we name *i*-merge.

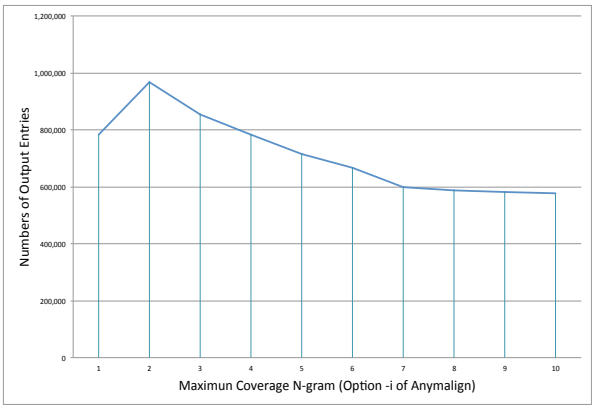


Figure 4: Number of entries in phrase translation tables for different values of parameter *i* between 1 and 10. This graph plots the figures given in Table 1.

Table 2 (See: Column 3 for *Phrase-Tables Entries*), shows that the use of an unsegmented corpus leads to larger phrase translation tables than the use of a pre-segmented corpus: twice the size for the basic version of Anymalign and 5 times for the merge of the all results of Anymalign run with option *-i*.

### 2.4 N-Grams × M-Grams Distribution

We investigated the $N \times M$-gram distribution in the phrase translation tables generated from both unsegmented and pre-segmented text corpora with Anymalign and GIZA++.

As presented in Appendix, Table 7 and 8 show the distribution for the pre-segmented corpus, where Tables 9, 10 and 11 are for the unsegmented corpus. Figures 5 - 9 provide a visualization of $N \times M$-Grams distributions in these phrase tables (see also Appendix.). They show that the phrase translation tables generated by GIZA++ exhibit a smoother decrease against the length of phrases, i.e. when $N$ and $M$ increase. Phrase translation tables output by Anymalign have significantly more entries when $N$ and $M$ are equal to or smaller than 2.

### 2.5 Comparison with an Existing Japanese-Chinese Bilingual Lexicon: EDR

The number of entries in the phrase translation tables does not give clues on the linguistic correctness of the entries. We thus compare the phrase translation tables against an existing Japanese-Chinese bilingual lexicon to check the correct word coverage rate.

The EDR Japanese-Chinese Bilingual Dictionary[2] contains 323,871 unique entries with an average length of words of 3.56 characters for Japanese and 3.46 for Chinese. Phrase translation tables generated with our method are not limited to words, but also contain phrases, fragments and short sentences that may not be included in the EDR bilingual lexicon. Therefore, we filtered the EDR lexicon to produce a filtered lexicon that contains only those entries which can actually be extracted from the training corpus. Using our corpus, the EDR lexicon has been filtered to 13,062 entries (96% reduced).

We then inspect the intersections between the filtered EDR lexicon and the phrase translation tables generated from both unsegmented and pre-segmented corpora output by Anymalign, basic version or *i*-merge, and GIZA++.

[2]The EDR Electronic Dictionary: National Institute of Information and Communication Technology (NiCT). URL: `http://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html`

As shown in Table 2, the phrase translation table extracted from the unsegmented corpus with Anymalign *i*-merge has 3,643 entries in common with the filtered EDR lexicon.

We would also like to take the translation probabilities $P(t|s)$ in the generated phrase translation tables into consideration in our comparison. When there are $m$ common entries of two phrase tables $tt_1$ and $tt_2$, we can compute the Intersection Score using metrics where $P(t|s)$ stands for the translation probability appearing in phrase translation tables.

$$\text{Score}(tt_1, tt_2) = \frac{\sum_{k=1}^{m} P_{tt_1}(t|s) + \sum_{k=1}^{m} P_{tt_2}(t|s)}{2}$$

The intersection scores obtained are reported in the last column in Table 2. These results show that the phrase translation table extracted from unsegmented corpus with Anymalign *i*-merge has the highest overlap with the filtered EDR lexicon.

### 2.6 Monolingual Recall

In order to know how effective the method can correctly extract phrases, we inspected the coverage rate of phrases by comparing with existing Japanese and Chinese word lists respectively.

We merged the Chinese resources listed below to build a Chinese word list (numbers are in unique entries):

- LDC Wordlist[3] (Chinese part): 128,341
- Baidu Baike[4]: 823,333
- Sogou Chinese Word List[5]: 35,650
- EDR (Chinese part): 151,651

For Japanese, the resources are listed below.

- LDC Wordlist (Japanese part): 187,267
- CTS Japanese Frequency List[6]: 15,000
- EDR (Japanese part): 229,392

[3] http://projects.ldc.upenn.edu/Chinese/
[4] http://baike.baidu.com/
[5] http://www.sogou.com/
[6] http://corpus.leeds.ac.uk/list.html

In total, we obtained a Chinese monolingual word list of 1,032,919 unique entries and a Japanese monolingual word list of 330,610 unique entries. We then filtered the two monolingual word lists to restrict them to the items found in our training corpora. This resulted in two filtered monolingual word lists of 19,037 entries in Chinese and 14,166 in Japanese. Table 3 shows the Recall Rate of monolingual phrases extracted in the phrase translation tables against the filtered monolingual Japanese and Chinese word lists.

Monolingual Recall for Japanese

| | Pre-seg | | Unseg | |
|---|---|---|---|---|
| Aligner | Retrieved | Recall | Retrieved | Recall |
| GIZA++ | 3,358 | 23.70% | 5,228 | **36.91%** |
| Anym b. | 6,953 | 49.08% | 9,479 | **66.91%** |
| Anym -*i* | 7,110 | 50.19% | 10,520 | **74.26%** |

Monolingual Recall for Chinese

| | Pre-seg | | Unseg | |
|---|---|---|---|---|
| Aligner | Retrieved | Recall | Retrieved | Recall |
| GIZA++ | 4,909 | 25.79% | 7,450 | **39.13%** |
| Anym b. | 9,666 | 50.77% | 14,186 | **74.52%** |
| Anym -*i* | 9,967 | 52.36% | 15,031 | **78.96%** |

Table 3: Monolingual Recall in phrase tables for Japanese and Chinese

## 3 Machine Translation Experiment

In this section, we use the phrase translation tables extracted in the previous sections in statistical machine translation experiments.

### 3.1 Data

We keep using our in-house Japanese-Chinese bilingual parallel corpus to test the feasibility of utilizing a training corpus of such a limited size. Table 4 shows the statistics of the training, tuning and testing corpora in their sizes and average lengths of sentences (numbers of characters or words per sentence) in their unsegmented corpus and presegmented forms.

### 3.2 Evaluation Metrics and Results

We use the state-of-the-art phrase-based machine translation system Moses (Koehn et al., 2007) to

| | | Japanese | Chinese |
|---|---|---|---|
| Train | Sentences | 9,500 | 9,500 |
| | Avg. len(w) | 10 (±5.93) | 8 (±4.99) |
| | Avg. len(c) | 17 (±9.95) | 11(±7.40) |
| Tune | Sentences | 500 | 500 |
| | Avg. len(w) | 10 (±5.96) | 8 (±5.10) |
| | Avg. len(c) | 17 (±9.98) | 11(±7.55) |
| Test | Sentences | 500 | 500 |
| | Avg. len(w) | 10 (±5.88) | 8 (±5.19) |
| | Avg. len(c) | 17 (±9.85) | 11(±7.94) |

Table 4: Statistics of the training, tuning and testing corpora. Avg. len(w) stands for the average number of words in each sentence. Avg. len(c) stands for the average number of characters in each sentence.

perform our machine translation experiments. As for the evaluation, we use the standard metrics WER (Nießen et al., 2000), BLEU (Papineni et al., 2002), NIST (Doddington et al., 2000) and TER (Snover et al., 2006).

Being a fast, automated and open source tool, the BLEU metric has been adopted as the main measure of fluency and adequacy (Akiba et al., 2004) in the domain of machine translation. It basically evaluates the precision of N-grams according to a reference translation.

However, word-level BLEU metric has been challenged in recent years. Denoual and Lepage (2005) studied the equivalence of applying BLEU metrics in characters and suggested that the use of BLEU at the character level could eliminate the word segmentation problem. Li et al.,(2011) stated that character-level metrics correlate better with human assessment. Chinese word segmentation is not needed for auto-evaluation. Besides, the campaigns like IWSLT '08 and NIST '08 both adopted character-level evaluation metrics.

Table 5 shows the evaluation results obtained when using Anymalign *i*-merge and Table 6 when using GIZA++. $BLEU_{cN}$ stands for the measure in characters for a given order N.

In both tables, so as to ensure consistency, the quality of Chinese translation outputs has been measured in characters. The results show that the phrase translation table generated from the unsegmented corpus outperfoms the phrase translation tables generated from the pre-segmented corpus. From this, it can be concluded that word segmentation is not a necessary step for statistical machine translation experiments between Japanese and Chinese language.

| Eval. Metric | Anymalign *i*-merge | |
|---|---|---|
| | Pre-seg | Unseg |
| $BLEU_{c4}$ | 0.1586 | **0.1900** |
| $BLEU_{c5}$ | 0.1162 | **0.1436** |
| $BLEU_{c6}$ | 0.0868 | **0.1099** |
| $BLEU_{c7}$ | 0.0660 | **0.0850** |
| $BLEU_{c8}$ | 0.0509 | **0.0673** |
| WER | 0.7595 | **0.7121** |
| NIST | 4.6215 | **5.2904** |
| TER | 0.7744 | **0.7144** |

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

| Eval. Metric | GIZA++ | |
|---|---|---|
| | Pre-seg | Unseg |
| $BLEU_{c4}$ | 0.1472 | **0.1938** |
| $BLEU_{c5}$ | 0.1117 | **0.1517** |
| $BLEU_{c6}$ | 0.0873 | **0.1210** |
| $BLEU_{c7}$ | 0.0696 | **0.0979** |
| $BLEU_{c8}$ | 0.0565 | **0.0806** |
| WER | 0.8373 | **0.7214** |
| NIST | 4.2198 | **5.1438** |
| TER | 0.8337 | **0.7290** |

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

## 4 Discussion

The results of the experiments we conducted with an unsegmented corpus outperformed the results of the same experiments conducted with the same pre-segmented corpus. This applies for both phrasal extraction and statistical machine translation between Chinese and Japanese. We explain below the reasons that may explain this fact.

Firstly, the unsegmented corpus gives more chances to match with correct alignment in Chinese and Japanese corpus. For example, 学生会(Students' Union) can be segmented into either 学生_会 or 学生会. Its translation in Japanese is 学友会 which is segmented into 学友_会 by Juman. As such, the chance for Chinese 学生会 to match with Japanese 学友会 in the pre-segmented

corpus is either zero or fifty percent. By opposition, for character-based text, their match rate is 66.67%. This shows that Chinese and Japanese word segmentation may vary in terms of refinement. Word segmentation performed on the output text and the reference text in the same language may not be consistent either.

Many Chinese Hanzi and Japanese Kanji are common to both languages. When applying phrase extraction, such linguistic feature may become very helpful in phrasal extraction and statistical machine translation. Goh et al. (2005) studied the accuracy of possible conversion between Chinese Hanzi and Japanese Kanji. Their study shows that around two thirds of the nouns and verbal nouns in Japanese are Kanji words and more than one third of them can be transposed into Chinese directly.

## 5 Conclusion

In this paper, we used a small-size Japanese-Chinese parallel corpus to conduct experiments in phrasal extraction and statistical machine translation. Our corpus was used under two forms: in a pre-segmented form obtained using Japanese and Chinese word segmentation tools, and in an unsegmented form, i.e., under this form, the processing unit was the character. Our experiment results show that the unsegmented form lead to better results than the pre-segmented form in both tasks. We believe that unsegmented forms of Chinese and Japanese corpora have the potential of improving translations between Japanese and Chinese. In summary, our experiments have shown that word segmentation may not be necessary for some NLP tasks between Japanese and Chinese.

## Acknowledgments

This research has been supported in part by the Kitakyushu Foundation for the Advancement of Industry, Science and Technology (FAIS) with Foreign Joint Project funds.

## References

Yusuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT'04 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Etienne Denoual and Yves Lepage. 2005. 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.

George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. 2000. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254.

Chooi-Ling Goh, Masayuki Asahara, and Yuji Matsumoto. 2005. Building a Japanese-Chinese Dictionary Using Kanji/Hanzi Conversion. In LNAI 3651, editor, *R. Dale et al. (Eds.): IJCNLP*, pages 670–681.

Donald E. Knuth. 2012. Satisfiability and the art of computer programming. In *SAT*, page 15.

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. 2007. 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.

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

Maoxi Li, Chengqing Zong, and Hwee Tou Ng. 2011. 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.

Sun Maosong, Shen Dayang, and Benjamin K Tsou. 1998. Chinese word segmentation without using lexicon and hand-crafted training data. In *Proceedings of the 36th Annual Meeting of ACL and 17th International Conference on Computational Linguistics (COLING-ACL 98)*, pages 1265–1271, Montreal, Quebec, Canada, August.

Takashi Masuoka and Yukinori Kabuto. 1989. *Basic Japanese Grammar*. Kuroshi Publishers.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 39–45, Athens.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. 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.

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

Jia Xu, Richard Zens, and Hermann Ney. 2004. Do we need Chinese word segmentation for Statistical Machine Translation? In *Proceedings of the ACL SIGHAN Workshop 2004*, pages 122–128, Barcelna, Spain.

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

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. Chinese word segmentation and statistical machine translation. *ACM Transactions on Speech and Language Processing*, 5(2):1–19.

## Appendix:
## $N \times M$-Grams Distribution in Phrase Tables for Pre-segmented and Unsegmented Corpus with Different Aligners and Their Visualisation Graphs.

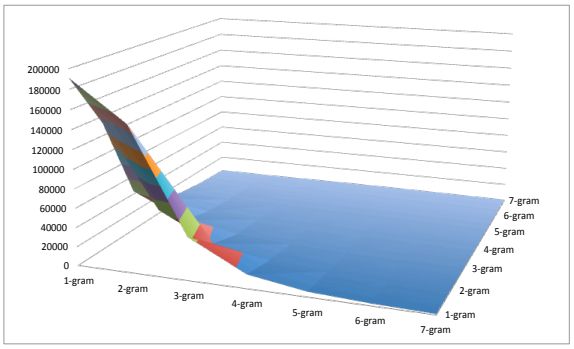


Figure 5: A visualization of N x M-grams distribution in phrase translation tables obtained from the **unsegmented** corpus using the basic version of Anymalign.

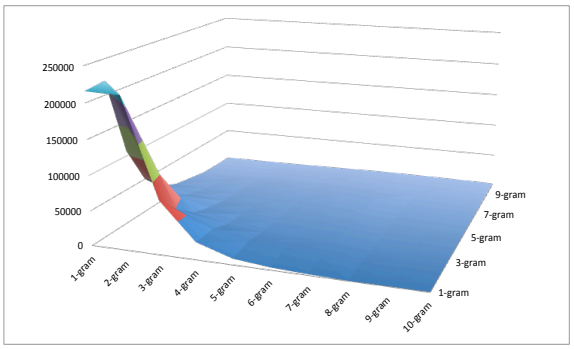


Figure 6: A visualization of N x M-grams distribution in phrase translation tables obtained from the **unsegmented** corpus using Anymalign *i*-merge.

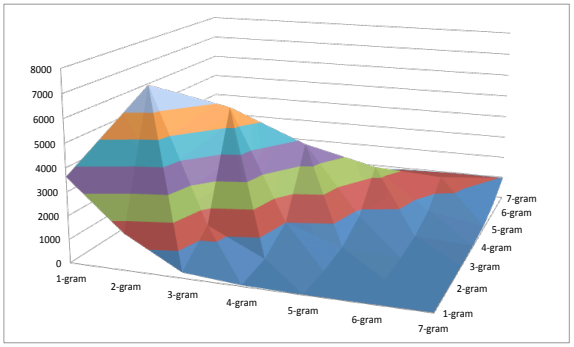


Figure 7: A visualization of N x M-grams distribution in phrase translation tables obtained from the **unsegmented** corpus using GIZA++.

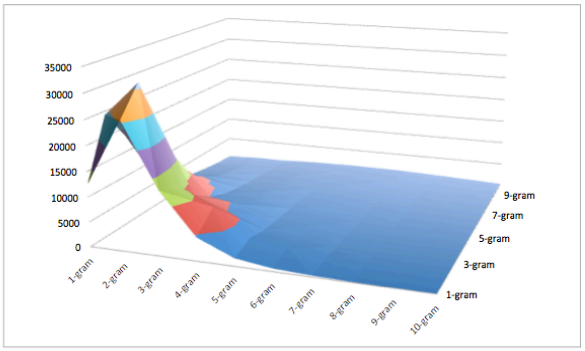


Figure 8: A visualization of N x M-grams distribution in phrase translation tables obtained from **pre-segmented** corpus using the basic version of Anymalign.

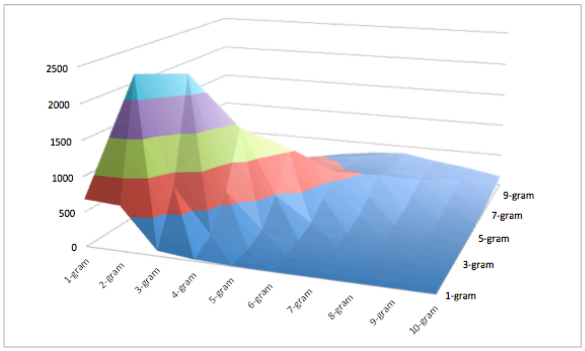


Figure 9: A visualization of N x M grams distribution in phrase translation tables obtained from the **pre-segmented** corpus using GIZA++.

| | | Target | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1-char | 2-char | 3-char | 4-char | 5-char | 6-char | 7-char | 8-char | 9-char | 10-char | ··· | total |
| Source | 1-char | 12,501 | 25,559 | 12,876 | 4,612 | 1,569 | 604 | 264 | 102 | 53 | 19 | ··· | 58,163 |
| | 2-char | 24,272 | **31,111** | 11,640 | 8,216 | 2,830 | 1,857 | 762 | 356 | 167 | 54 | ··· | 81,307 |
| | 3-char | 18,375 | 18,554 | 7,550 | 4,875 | 2,025 | 1,135 | 497 | 235 | 113 | 34 | ··· | 53,420 |
| | 4-char | 10,958 | 11,264 | 4,950 | 4,063 | 1,855 | 1,319 | 577 | 321 | 149 | 71 | ··· | 35,577 |
| | 5-char | 6,008 | 6,576 | 3,378 | 2,894 | 1,759 | 1,115 | 611 | 280 | 142 | 45 | ··· | 22,838 |
| | 6-char | 3,479 | 4,282 | 2,562 | 2,481 | 1,622 | 1,184 | 635 | 375 | 175 | 58 | ··· | 16,898 |
| | 7-char | 1,956 | 2,642 | 1,883 | 1,960 | 1,635 | 1,228 | 821 | 439 | 249 | 77 | ··· | 12,937 |
| | 8-char | 1,266 | 1,810 | 1,484 | 1,690 | 1,521 | 1,320 | 959 | 571 | 320 | 138 | ··· | 11,154 |
| | 9-char | 736 | 1,118 | 1,047 | 1,286 | 1,322 | 1,260 | 1,080 | 714 | 439 | 224 | ··· | 9,354 |
| | 10-char | 455 | 727 | 727 | 1,028 | 1,135 | 1,143 | 1,066 | 802 | 553 | 267 | ··· | 8,083 |
| | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· |
| | total | 80,576 | 104,654 | 492,008 | 34,555 | 19,164 | 14,462 | 9,677 | 6,459 | 4,195 | 2,160 | ··· | **326,748** |

Table 7: $N \times M$-grams (characters) distribution in the phrase translation table obtained from the **pre-segmented** corpus using the basic version of Anymalign.

| | | Target | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1-char | 2-char | 3-char | 4-char | 5-char | 6-char | 7-char | 8-char | 9-char | 10-char | ··· | total |
| Source | 1-char | 681 | 650 | 77 | 24 | 6 | 1 | 4 | 1 | 0 | 0 | ··· | 1,444 |
| | 2-char | 741 | **2,341** | 816 | 189 | 42 | 17 | 14 | 1 | 1 | 0 | ··· | 4,162 |
| | 3-char | 478 | 1,707 | 2,285 | 649 | 136 | 48 | 32 | 11 | 5 | 2 | ··· | 5,353 |
| | 4-char | 220 | 887 | 1,326 | 1,438 | 489 | 133 | 48 | 22 | 10 | 9 | ··· | 4,583 |
| | 5-char | 92 | 549 | 786 | 980 | 1,057 | 340 | 110 | 46 | 17 | 8 | ··· | 3,986 |
| | 6-char | 38 | 338 | 560 | 786 | 766 | 604 | 263 | 85 | 36 | 17 | ··· | 3,499 |
| | 7-char | 14 | 167 | 329 | 549 | 591 | 493 | 450 | 173 | 75 | 18 | ··· | 2,876 |
| | 8-char | 10 | 84 | 194 | 380 | 502 | 442 | 390 | 269 | 134 | 53 | ··· | 2,483 |
| | 9-char | 3 | 63 | 110 | 231 | 369 | 428 | 386 | 291 | 199 | 86 | ··· | 2,212 |
| | 10-char | 0 | 14 | 66 | 164 | 264 | 341 | 382 | 296 | 230 | 140 | ··· | 1,976 |
| | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· | ··· |
| | total | 2,281 | 6,828 | 6,629 | 5,579 | 4,611 | 3,433 | 2,838 | 1,954 | 1,331 | 808 | ··· | **36,888** |

Table 8: $N \times M$-grams (characters) distribution in the phrase translation table obtained from the **pre-segmented** corpus using GIZA++

| | | Target | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-char | 2-char | 3-char | 4-char | 5-char | 6-char | 7-char | total |
| Source | 1-char | 3,625 | 1,549 | 242 | 50 | 5 | 4 | 1 | 5,476 |
| | 2-char | 2,683 | **7,046** | 1,384 | 248 | 51 | 12 | 6 | 11,430 |
| | 3-char | 995 | 3,731 | 5,788 | 1,008 | 208 | 37 | 18 | 11,785 |
| | 4-char | 462 | 1,539 | 2,928 | 3,806 | 794 | 173 | 34 | 9,736 |
| | 5-char | 199 | 849 | 1,401 | 2,106 | 2,352 | 555 | 123 | 7,585 |
| | 6-char | 79 | 434 | 749 | 1,185 | 1,450 | 1,449 | 426 | 5,772 |
| | 7-char | 44 | 173 | 423 | 700 | 917 | 984 | 977 | 4,218 |
| | total | 8,087 | 15,321 | 12,915 | 9,103 | 5,777 | 3,214 | 1,585 | **56,002** |

Table 9: $N \times M$-grams (characters) distribution in the phrase translation table obtained from the **unsegmented** corpus using GIZA++.

| | | Target | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1-char | 2-char | 3-char | 4-char | 5-char | 6-char | 7-char | total |
| Source | 1-char | **190,967** | 150,445 | 44,562 | 14,522 | 5,436 | 2,438 | 959 | 409,329 |
| | 2-char | 132,744 | 46,374 | 17,632 | 7,671 | 3,403 | 1,650 | 743 | 210,217 |
| | 3-char | 42,967 | 16,959 | 8,012 | 4,246 | 2,126 | 1,125 | 491 | 75,926 |
| | 4-char | 16,673 | 8,244 | 5,185 | 3,401 | 2,000 | 1,121 | 561 | 37,185 |
| | 5-char | 7,350 | 4,612 | 3,590 | 2,819 | 1,934 | 1,177 | 639 | 22,121 |
| | 6-char | 3,974 | 2,919 | 2,765 | 2,489 | 1,939 | 1,285 | 780 | 16,151 |
| | 7-char | 2,362 | 1,922 | 2,074 | 2,210 | 1,988 | 1,491 | 1,028 | 13,075 |
| | total | 397,037 | 231,475 | 83,820 | 37,358 | 18,826 | 10,287 | 5,201 | **784,004** |

Table 10: $N \times M$-grams (characters) distribution in the phrase translation table obtained from the **unsegmented** corpus using the basic version of Anymalign.

| | | Target | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1-char | 2-char | 3-char | 4-char | 5-char | 6-char | 7-char | 8-char | 9-char | 10-char | total |
| Source | 1-char | 215,840 | 214,475 | 76,268 | 24,808 | 8,818 | 3,806 | 1,545 | 158 | 54 | 12 | 545,784 |
| | 2-char | **220,226** | 121,635 | 47,728 | 24,868 | 13,562 | 9,403 | 6,250 | 2,207 | 1,154 | 462 | 447,495 |
| | 3-char | 107,143 | 55,351 | 25,963 | 15,659 | 9,596 | 6,978 | 4,787 | 1,785 | 910 | 369 | 228,541 |
| | 4-char | 50,381 | 31,074 | 17,561 | 12,999 | 9,215 | 7,311 | 5,279 | 2,049 | 1,125 | 463 | 137,457 |
| | 5-char | 23,597 | 16,812 | 11,495 | 9,932 | 8,053 | 6,725 | 5,126 | 2,067 | 1,140 | 457 | 85,404 |
| | 6-char | 12,372 | 10,233 | 8,304 | 8,232 | 7,475 | 6,762 | 5,483 | 2,468 | 1,499 | 673 | 63,501 |
| | 7-char | 7,040 | 6,509 | 6,111 | 6,886 | 6,780 | 6,662 | 5,696 | 2,799 | 1,804 | 857 | 51,144 |
| | 8-char | 1,946 | 1,992 | 2,424 | 3,302 | 3,898 | 4,304 | 3,918 | 3,059 | 2,192 | 1,140 | 28,175 |
| | 9-char | 974 | 1,014 | 1,431 | 2,257 | 3,065 | 3,768 | 3,690 | 3,100 | 2,566 | 1,462 | 23,327 |
| | 10-char | 401 | 440 | 678 | 1,291 | 1952 | 2,745 | 2,975 | 2,785 | 2,354 | 1,792 | 17,413 |
| | total | 639,920 | 459,535 | 197,963 | 110,234 | 72,414 | 58,464 | 44,749 | 22,477 | 14,798 | 7,687 | **1,628,241** |

Table 11: $N \times M$-grams (characters) distribution in the phrase translation table obtained from the **unsegmented** corpus using Anymalign *i*-merge.