# Enhancing Statistical Machine Translation with Character Alignment

**Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, Jiajun Chen**
State Key Laboratory for Novel Software Technology,
Department of Computer Science and Technology,
Nanjing University, Nanjing, 210046, China
{xin,tanggc,dxy,huangsj,chenjj}@nlp.nju.edu.cn

## Abstract

The dominant practice of statistical machine translation (SMT) uses the same Chinese word segmentation specification in both alignment and translation rule induction steps in building Chinese-English SMT system, which may suffer from a suboptimal problem that word segmentation better for alignment is not necessarily better for translation. To tackle this, we propose a framework that uses two different segmentation specifications for alignment and translation respectively: we use Chinese character as the basic unit for alignment, and then convert this alignment to conventional word alignment for translation rule induction. Experimentally, our approach outperformed two baselines: fully word-based system (using word for both alignment and translation) and fully character-based system, in terms of alignment quality and translation performance.

## 1 Introduction

Chinese Word segmentation is a necessary step in Chinese-English statistical machine translation (SMT) because Chinese sentences do not delimit words by spaces. The key characteristic of a Chinese word segmenter is the segmentation specification[1]. As depicted in Figure 1(a), the dominant practice of SMT uses the same word segmentation for both word alignment and translation rule induction. For brevity, we will refer to the word segmentation of the bilingual corpus as *word segmentation for alignment* (WSA for short), because it determines the basic tokens for alignment; and refer to the word segmentation of the aligned corpus as *word segmentation for rules* (WSR for short), because it determines the basic tokens of translation rules[2], which also determines how the translation rules would be matched by the source sentences.

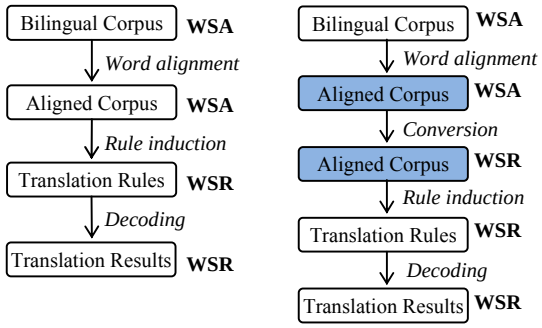


Figure 1. WSA and WSR in SMT pipeline

It is widely accepted that word segmentation with a higher F-score will not necessarily yield better translation performance (Chang et al., 2008; Zhang et al., 2008; Xiao et al., 2010). Therefore, many approaches have been proposed to learn word segmentation suitable for SMT. These approaches were either complicated (Ma et al., 2007; Chang et al., 2008; Ma and Way, 2009; Paul et al., 2010), or of high computational complexity (Chung and Gildea 2009; Duan et al., 2010). Moreover, they implicitly assumed that WSA and WSR should be equal. This requirement may lead to a suboptimal problem that word segmentation better for alignment is not necessarily better for translation.

To tackle this, we propose a framework that uses different word segmentation specifications as WSA and WSR respectively, as shown Figure 1(b). We investigate a solution in this framework: first, we use Chinese character as the basic unit for alignment, viz. *character alignment*; second, we use a simple method (Elming and Habash, 2007) to convert the character alignment to conventional word alignment for translation rule induction. In the

[1] We hereafter use "word segmentation" for short.

[2] Interestingly, word is also a basic token in syntax-based rules.



*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 285–290,
Jeju, Republic of Korea, 8-14 July 2012. 

experiment, our approach consistently outperformed two baselines with three different word segmenters: fully word-based system (using word for both alignment and translation) and fully character-based system, in terms of alignment quality and translation performance.

The remainder of this paper is structured as follows: Section 2 analyzes the influences of WSA and WSR on SMT respectively; Section 3 discusses how to convert character alignment to word alignment; Section 4 presents experimental results, followed by conclusions and future work in section 5.

## 2 Understanding WSA and WSR

We propose a solution to tackle the suboptimal problem: using Chinese character for alignment while using Chinese word for translation. Character alignment differs from conventional word alignment in the basic tokens of the Chinese side of the training corpus[3]. Table 1 compares the token distributions of character-based corpus (*CCorpus*) and word-based corpus (*WCorpus*). We see that the *WCorpus* has a longer-tailed distribution than the *CCorpus*. More than 70% of the unique tokens appear less than 5 times in *WCorpus*. However, over half of the tokens appear more than or equal to 5 times in the *CCorpus*. This indicates that modeling word alignment could suffer more from data sparsity than modeling character alignment.

Table 2 shows the numbers of the unique tokens (*#UT*) and unique bilingual token pairs (*#UTP*) of the two corpora. Consider two extensively features, fertility and translation features, which are extensively used by many state-of-the-art word aligners. The number of parameters w.r.t. fertility features grows linearly with *#UT* while the number of parameters w.r.t. translation features grows linearly with *#UTP*. We compare *#UT* and *#UTP* of both corpora in Table 2. As can be seen, *CCorpus* has less *UT* and *UTP* than *WCorpus*, i.e. character alignment model has a compact parameterization than word alignment model, where the compactness of parameterization is shown very important in statistical modeling (Collins, 1999).

Another advantage of character alignment is the reduction in alignment errors caused by word segmentation errors. For example, "切尼 (Cheney)" and "愿 (will)" are wrongly merged into one word 切尼愿 by the word segmenter, and 切尼愿 wrongly aligns to a comma in English sentence in the word alignment; However, both 切 and 尼 align to "*Cheney*" correctly in the character alignment. However, this kind of errors cannot be fixed by methods which learn new words by packing already segmented words, such as word packing (Ma et al., 2007) and Pseudo-word (Duan et al., 2010).

| Frequency | Characters (%) | Words (%) |
|---|---|---|
| **1** | 27.22 | 45.39 |
| **2** | 11.13 | 14.61 |
| **3** | 6.18 | 6.47 |
| **4** | 4.26 | 4.32 |
| **5(+)** | 50.21 | 29.21 |

Table 1 Token distribution of *CCorpus* and *WCorpus*

| Stats. | Characters | Words |
|---|---|---|
| *#UT* | 9.7K | 88.1K |
| *#UTP* | 15.8M | 24.2M |

Table 2 *#UT* and *#UTP* in *CCorpus* and *WCorpus*

As character could preserve more meanings than word in Chinese, it seems that a character can be wrongly aligned to many English words by the aligner. However, we found this can be avoided to a great extent by the basic features (co-occurrence and distortion) used by many alignment models. For example, we observed that the four characters of the non-compositional word "阿拉法特 (Arafat)" align to *Arafat* correctly, although these characters preserve different meanings from that of *Arafat*. This can be attributed to the frequent co-occurrence (192 times) of these characters and *Arafat* in *CCorpus*. Moreover,法 usually means *France* in Chinese, thus it may co-occur very often with *France* in *CCorpus*. If both *France* and *Arafat* appear in the English sentence, 法 may wrongly align to *France*. However, if 阿 aligns to *Arafat*, 法 will probably align to *Arafat*, because aligning 法 to *Arafat* could result in a lower distortion cost than aligning it to *France*.

Different from alignment, translation is a pattern matching procedure (Lopez, 2008). WSR determines how the translation rules would be matched by the source sentences. For example, if we use translation rules with character as WSR to translate name entities such as the non-compositional word 阿拉法特, i.e. translating literally, we may get a wrong translation. That's because the linguistic

[3] Several works have proposed to use character (letter) on both sides of the parallel corpus for SMT between similar (European) languages (Vilar et al., 2007; Tiedemann, 2009), however, Chinese is not similar to English.

knowledge that the four characters convey a specific meaning different from the characters has been lost, which cannot always be totally recovered even by using phrase in phrase-based SMT systems (see Chang et al. (2008) for detail). Duan et al. (2010) and Paul et al., (2010) further pointed out that coarser-grained segmentation of the source sentence do help capture more contexts in translation. Therefore, rather than using character, using coarser-grained, at least as coarser as the conventional word, as WSR is quite necessary.

## 3 Converting Character Alignment to Word Alignment

In order to use word as WSR, we employ the same method as Elming and Habash (2007)[4] to convert the character alignment (*CA*) to its word-based version (*CA'*) for translation rule induction. The conversion is very intuitive: for every English-Chinese word pair $(e, c)$ in the sentence pair, we align $c$ to $e$ as a link in *CA'*, if and only if there is at least one Chinese character of $c$ aligns to $e$ in *CA*.

Given two different segmentations A and B of the same sentence, it is easy to prove that if every word in A is finer-grained than the word of B at the corresponding position, the conversion is unambiguity (we omit the proof due to space limitation). As character is a finer-grained than its original word, character alignment can always be converted to alignment based on any word segmentation. Therefore, our approach can be naturally scaled to syntax-based system by converting character alignment to word alignment where the word segmentation is consistent with the parsers.

We compare *CA* with the conventional word alignment (*WA*) as follows: We hand-align some sentence pairs as the evaluation set based on characters (*ESChar*), and converted it to the evaluation set based on word (*ESWord*) using the above conversion method. It is worth noting that comparing *CA* and *WA* by evaluating *CA* on *ESChar* and evaluating *WA* on *ESWord* is meaningless, because the basic tokens in *CA* and *WA* are different. However, based on the conversion method, comparing *CA* with *WA* can be accomplished by evaluating both *CA'* and *WA* on *ESWord*.

## 4 Experiments

### 4.1 Setup

FBIS corpus (LDC2003E14) (210K sentence pairs) was used for small-scale task. A large bilingual corpus of our lab (1.9M sentence pairs) was used for large-scale task. The NIST'06 and NIST'08 test sets were used as the development set and test set respectively. The Chinese portions of all these data were preprocessed by character segmenter (CHAR), ICTCLAS word segmenter[5] (ICT) and Stanford word segmenters with CTB and PKU specifications[6] respectively. The first 100 sentence pairs of the hand-aligned set in Haghighi et al. (2009) were hand-aligned as *ESChar*, which is converted to three *ESWord*s based on three segmentations respectively. These *ESWord*s were appended to training corpus with the corresponding word segmentation for evaluation purpose.

Both character and word alignment were performed by GIZA++ (Och and Ney, 2003) enhanced with *gdf* heuristics to combine bidirectional alignments (Koehn et al., 2003). A 5-gram language model was trained from the Xinhua portion of Gigaword corpus. A phrase-based MT decoder similar to (Koehn et al., 2007) was used with the decoding weights optimized by MERT (Och, 2003).

### 4.2 Evaluation

We first evaluate the alignment quality. The method discussed in section 3 was used to compare character and word alignment. As can be seen from Table 3, the systems using character as WSA outperformed the ones using word as WSA in both small-scale (row 3-5) and large-scale task (row 6-8) with all segmentations. This gain can be attributed to the small vocabulary size (sparsity) for character alignment. The observation is consistent with Koehn (2005) which claimed that there is a negative correlation between the vocabulary size and translation performance without explicitly distinguishing WSA and WSR.

We then evaluated the translation performance. The baselines are fully word-based MT systems (*WordSys*), i.e. using word as both WSA and WSR, and fully character-based systems (*CharSys*). Table

[4] They used this conversion for word alignment combination only, no translation results were reported.

[5] http://www.ictclas.org/

[6] http://nlp.stanford.edu/software/segmenter.shtml

| | | Word alignment | | | Character alignment | | |
|---|---|---|---|---|---|---|---|
| | | **P** | **R** | **F** | **P** | **R** | **F** |
| **S** | **CTB** | 76.0 | 81.9 | 78.9 | 78.2 | 85.2 | 81.8 |
| | **PKU** | 76.1 | 82.0 | 79.0 | 78.0 | 86.1 | 81.9 |
| | **ICT** | 75.2 | 80.8 | 78.0 | 78.7 | 86.3 | 82.3 |
| L | **CTB** | 79.6 | 85.6 | 82.5 | 82.2 | 90.6 | 86.2 |
| | **PKU** | 80.0 | 85.4 | 82.6 | 81.3 | 89.5 | 85.2 |
| | **ICT** | 80.0 | 85.0 | 82.4 | 81.3 | 89.7 | 85.3 |

Table 3 Alignment evaluation. Precision (P), recall (R), and *F-score* (F) with $\alpha = 0.5$ (Fraser and Marcu, 2007)

| | **WSA** | **WSR** | **CTB** | **PKU** | **ICT** |
|---|---|---|---|---|---|
| **S** | word | word | 21.52 | 20.99 | 20.95 |
| | char | word | **22.04** | **21.98** | **22.04** |
| **L** | word | word | 22.07 | 22.86 | 22.23 |
| | char | word | **23.41** | **23.51** | **23.05** |

Table 4 Translation evaluation of *WordSys* and proposed system using BLEU-SBP (Chiang et al., 2008)

4 compares *WordSys* to our proposed system. Significant testing was carried out using bootstrap re-sampling method proposed by Koehn (2004) with a 95% confidence level. We see that our proposed systems outperformed *WordSys* in all segmentation specifications settings. Table 5 lists the results of *CharSys* in small-scale task. In this setting, we gradually set the phrase length and the distortion limits of the phrase-based decoder (context size) to 7, 9, 11 and 13, in order to remove the disadvantage of shorter context size of using character as WSR for fair comparison with *WordSys* as suggested by Duan et al. (2010). Comparing Table 4 and 5, we see that all *CharSys* underperformed *WordSys*. This observation is consistent with Chang et al. (2008) which claimed that using characters, even with large phrase length (up to 13 in our experiment) cannot always capture everything a Chinese word segmenter can do, and using word for translation is quite necessary. We also see that *CharSys* underperformed our proposed systems, that's because the harm of using character as WSR outweighed the benefit of using character as WSA, which indicated that word segmentation better for alignment is not necessarily better for translation, and vice versa.

We finally compared our approaches to Ma et al. (2007) and Ma and Way (2009), which proposed "packed word (PW)" and "bilingual motivated word (BS)" respectively. Both methods iteratively learn word segmentation and alignment alternatively, with the former starting from word-based corpus and the latter starting from characters-based corpus. Therefore, PW can be experimented on all segmentations. Table 6 lists their results in small-scale task, we see that both PW and BS underperformed our approach. This may be attributed to the low recall of the learned BS or PW in their approaches. BS underperformed both two baselines, one reason is that Ma and Way (2009) also employed word lattice decoding techniques (Dyer et al., 2008) to tackle the low recall of BS, which was removed from our experiments for fair comparison.

| **Context Size** | **7** | **9** | **11** | **13** |
|---|---|---|---|---|
| **BLEU** | 20.90 | 21.19 | 20.89 | 21.09 |

Table 5 Translation evaluation of *CharSys*.

| **System** | **WSA** | **WSR** | **CTB** | **PKU** | **ICT** |
|---|---|---|---|---|---|
| WordSys | word | word | 21.52 | 20.99 | 20.95 |
| Proposed | char | word | **22.04** | **21.98** | **22.04** |
| PW | PW | PW | 21.24 | 21.24 | 21.19 |
| Char+PW | char | PW | **22.46** | **21.87** | **21.97** |
| BS | BS | BS | | 19.76 | |
| Char+BS | char | BS | | 20.19 | |

Table 6 Comparison with other works

Interestingly, we found that using character as WSA and BS as WSR (Char+BS), a moderate gain (+0.43 point) was achieved compared with fully BS-based system; and using character as WSA and PW as WSR (Char+PW), significant gains were achieved compared with fully PW-based system, the result of CTB segmentation in this setting even outperformed our proposed approach (+0.42 point). This observation indicated that in our framework, better combinations of WSA and WSR can be found to achieve better translation performance.

## 5 Conclusions and Future Work

We proposed a SMT framework that uses character for alignment and word for translation, which improved both alignment quality and translation performance. We believe that in this framework, using other finer-grained segmentation, with fewer ambiguities than character, would better parameterize the alignment models, while using other coarser-grained segmentation as WSR can help capture more linguistic knowledge than word to get better translation. We also believe that our approach, if integrated with combination techniques (Dyer et al., 2008; Xi et al., 2011), can yield better results.

## Acknowledgments

We thank ACL reviewers. This work is supported by the National Natural Science Foundation of China (No. 61003112), the National Fundamental Research Program of China (2010CB327903).

## References


Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Peitra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2), pages 263-311.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of third workshop on SMT*, pages 224-232.

David Chiang, Steve DeNeefe, Yee Seng Chan and Hwee Tou Ng. 2008. Decomposability of Translation Metrics for Improved Evaluation and Efficient Algorithms. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 610-619.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 718-726.

Michael Collins. 1999. Head-driven statistical models for natural language parsing. *Ph.D. thesis*, *University of Pennsylvania*.

Xiangyu Duan, Min Zhang, and Haizhou Li. 2010. Pseudo-word for phrase-based machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 148-156.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of the Association for Computational Linguistics*, pages 1012-1020.

Jakob Elming and Nizar Habash. 2007. Combination of statistical word alignments based on multiple preprocessing schemes. In *Proceedings of the Association for Computational Linguistics*, pages 25-28.

Alexander Fraser and Daniel Marcu. 2007. Squibs and Discussions: Measuring Word Alignment Quality for Statistical Machine Translation. In *Computational Linguistics*, 33(3), pages 293-303.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In *Proceedings of the Association for Computational Linguistics*, pages 923-931.

Phillip Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan,W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 177-180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 388-395.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*.

Adam David Lopez. 2008. Machine translation by pattern matching. *Ph.D. thesis, University of Maryland.*

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of the Association for Computational Linguistics*, pages 304-311.

Yanjun Ma and Andy Way. 2009. Bilingually motivated domain-adapted word segmentation for statistical machine translation. In *Proceedings of the Conference of the European Chapter of the ACL*, pages 549-557.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, pages 440-447.

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

Michael Paul, Andrew Finch and Eiichiro Sumita. 2010. Integration of multiple bilingually-learned segmentation schemes into statistical machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 400-408.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. In *Proceedings of the Annual Conference of the European Association for machine Translation*, pages 12-19.

David Vilar, Jan-T. Peter and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 33-39.

Xinyan Xiao, Yang Liu, Young-Sook Hwang, Qun Liu and Shouxun Lin. 2010. Joint tokenization and translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1200-1208.

Ning Xi, Guangchao Tang, Boyuan Li, and Yinggong Zhao. 2011. Word alignment combination over multiple word segmentation. In *Proceedings of the ACL 2011 Student Session*, pages 1-5.

Ruiqiang Zhang, Keiji Yasuda, and Eiichiro Sumita. 2008. Improved statistical machine translation by multiple Chinese word segmentation. In *Proceedings*

*of the Third Workshop on Statistical Machine Translation*, pages 216-223.