An Empirical Study on Word Segmentation for Chinese Machine Translation*

Hai Zhao[1,2,**], Masao Utiyama[3], Eiichro Sumita[3], and Bao-Liang Lu[1,2]

[1] MOE-Microsoft Key Laboratory of Intelligent Computing and Intelligent System
[2] Department of Computer Science and Engineering,
Shanghai Jiao Tong University, #800 Dongchuan Road, Shanghai, China, 200240
[3] Multilingual Translation Laboratory, MASTAR Project
National Institute of Information and Communications Technology
3-5 Hikaridai, Keihanna Science City, Kyoto, 619-0289, Japan
zhaohai@cs.sjtu.edu.cn, {mutiyama,eiichiro.sumita}@nict.go.jp,
bllu@sjtu.edu.cn

Abstract. Word segmentation has been shown helpful for Chinese-to-English machine translation (MT), yet the way different segmentation strategies affect MT is poorly understood. In this paper, we focus on comparing different segmentation strategies in terms of machine translation quality. Our empirical study covers both English-to-Chinese and Chinese-to-English translation for the first time. Our results show the necessity of word segmentation depends on the translation direction. After comparing two types of segmentation strategies with associated linguistic resources, we demonstrate that optimizing segmentation itself does not guarantee better MT performance, and segmentation strategy choice is not the key to improve MT. Instead, we discover that linguistical resources such as segmented corpora or the dictionaries that segmentation tools rely on actually determine how word segmentation affects machine translation. Based on these findings, we propose an empirical approach that directly optimize dictionary with respect to the MT task for word segmenter, providing a BLEU score improvement of 1.30.

1 Introduction

Word segmentation is regarded as a primary step for Chinese natural language processing, as Chinese words are not naturally defined with spaces appearing between words. Word segmentation is usually helpful for better understanding Chinese meaning though it is not always necessary. In this decade, researchers have developed quite a lot of techniques to seriously improve the segmentation performance, work motivated by a series of shared tasks on Chinese word

* This work was partially supported by the National Natural Science Foundation of China (Grant No. 60903119, Grant No. 61170114, and Grant No. 61272248), and the National Basic Research Program of China (Grant No. 2009CB320901 and Grant No.2013CB329401).

** This work was partially done as the first author was at NICT with support of MASTAR project.

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 248–263, 2013.
© Springer-Verlag Berlin Heidelberg 2013

segmentation, SIGHAN Bakeoff, has given especially satisfactory segmentation results for various further application in Chinese processing [1–3]. Typically, a segmenter has to be trained on a segmentation corpus subject to a predefined segmentation standard. A segmenter that is based on statistical learning can give a F-score of more than 95% in word segmentation performance evaluation.

However, researchers have realized that different natural language processing tasks may have quite different requirements for the segmentation task, which is often beyond the issues of segmentation performance or standards [4, 5]. A typical example of this concern is from Chinese related machine translation (MT). Basically, we try to answer two questions about the role of Chinese word segmentation in machine translation,

(1) Is word segmentation necessary?
(2) If it is, then which segmentation strategy should we adopt for better machine translation performance?

To the first question, our answer will be a NO, or more precisely, word segmentation strategies should be carefully selected so that it can really outperform a character aligning system. In theory, the current phrase-based alignment MT system is supposed to discover a phrase table at last, which right performs a similar operation over sentences as a word segmenter does. However, existing empirical works show that word segmentation can help an MT system work better than a system without word segmentation [6]. Later in this paper, we will actually show that word segmentation does not always lead to better machine translation performance.

To the second question, a number of empirical studies have been conducted [7, 8], and various improved segmentation strategies proposed. In this work, we continue the empirical study by expanding on the contents of existing work. What is the most different between previous work and this one is that various segmentation strategies in this paper are examined and compared by considering the affect of both linguistic resources and approach characteristics. In addition, we also consider both Chinese-to-English and English-to-Chinese translation tasks, while the latter translation task was seldom considered in existing work.

2 Related Work

All word segmentation strategies that are applied to machine translation can be put into two categories. One is the joint model, which is integrated into the aligning or decoding procedure of machine translation, and the other is the independent model, which may be flexibly used independent of an MT system. Independent models can be further split into two sub-classes, statistical and rule-based. The latter is sometimes called the dictionary (lexicon or vocabulary) based approach as a word list is specified aforehand for segmentation. If we distinguish word segmenters according to their data set sources, then we may also put them into two categories, monolingual-motivated and bilingual-motivated.

According to our knowledge, Xu et al. [6] is the first work on the use of word segmentation in MT, and their results showed that segmentation generated by

word alignments may achieve competitive results compared to using monolingual segmenters with a predefined third-party dictionary.

Later Xu et al. [9] proposed a joint segmentation model that uses word lattice decoding in phrase-based MT. This work was generalized to hierarchical MT systems and other language pairs in the work of Dyer et al. [10]. Both of these methods need a specific monolingual segmentation to generate the final word lattices.

Xu et al. [11] proposed a Bayesian semi-supervised Chinese word segmentation model which uses both monolingual and bilingual information to derive segmentation suitable for MT. Their approach models the source-to-null alignment and has been shown to be a special case of the model in the work of Nguyen et al. [12]. Both Xu et al. [11] and Nguyen et al. [12] belong to joint models and used Gibbs sampling for inference.

Ma and Way [13] proposed a bilingually motivated segmentation approach for MT. Their approach first uses the output from an existing statistical word aligner to obtain a set of candidate "words", then according to a metric, the co-occurrence frequencies, the segmentation of the respective sentences in the parallel corpus will be iteratively modified. These modified sentences will be fed back to the word aligner, which produces new alignments.

For other improvement about monolingual word segmenters, Chang et al. [7] suggested that tuning granularity of Chinese "words" given by segmenters can enhance machine translation. Zhang et al. [8] proposed that concatenating various corpora regardless of their different specifications can help producing a better segmenter for MT.

Though word segmentation is a concern especially for Chinese machine translation, it is also a consideration for other non-Chinese language pairs, Koehn et al. [14] and Habash and Sadat [15] showed that data-driven methods for splitting and preprocessing can improve Arabic-English and German-English MT, and Paul et al. [16] and Nguyen et al. [12] proposed a language independent segmentation strategy to improve MT for different language pairs.

3 Word Segmenters

We will try to evaluate the two main word segmentation approaches, statistical and dictionary-based (rule-based), in this paper. For the statistical approach, a segmentation corpus should be available for segmenter training. Character-based tagging has been shown as an effective strategy for corpus-based segmentation information acquisition according to results of the SIGHAN Bakeoff shared tasks [17–20]. This approach was initially proposed in the work of Xue and Shen [21] and it needs to define the position of character inside a word. Traditionally, the four tags, b, m, e, and s stand, respectively, for the beginning, middle, end of a word, and a single character word since then [21]. Later Zhao et al. [19] furthermore introduced two tags, b_2 and b_3, for the second and third character in a word and demonstrated better performance. The following n-gram features from [19] were used as basic features,

Table 1. Corpus statistics

Corpus		PKU2	MSRA2	CTB3
training set (M)	#word	1.1	2.37	0.51
	#char	1.83	3.9	0.83
test set (K)	#word	104	107	155
	#char	173	188	257

(a) $C_n(n = \text{-}1, 0, 1)$,
(b) $C_nC_{n+1}(n = \text{-}1, 0)$,
(c) $C_{-1}C_1$,

where C stands for a character and the subscripts for the relative order to the current character C_0.

Conditional random fields (CRF) has become popular for word segmentation since it provides better performance than other sequence labeling tools [22], and it will be adopted as our machine learning tool.

From the first to the third SIGHAN bakeoff, each time organizers provided four data sets for evaluation, in which two sets are traditional Chinese and the other two simplified Chinese. As our parallel corpus for MT is simplified Chinese, we consider adopting all six simplified data sets from Bakeoff 1,2 and 3. These six data sets are noted as CTB1, PKU1, MSRA2, PKU2, CTB3, and MSRA3. However, for the training set, CTB1 is a subset of CTB3, MSRA3 is a subset of MSRA2, and PKU1 and PKU2 are identical. Thus we only need to adopt three data sets, PKU2, MSRA2, and CTB3 to train our segmenters. Corpus size information is in Table 1.

For dictionary based segmentation strategy, a predefined dictionary should be available for segmentation use. Following the category of the work of Zhao and Kit [23], and assuming the availability of a list of word candidates or words (the dictionary) each associated with a goodness for how likely it is to be a true word. Let $W = \{\{w_i, g(w_i)\}_{i=1,\ldots,n}\}$ be such a list, where w_i is a word candidate and $g(w_i)$ its goodness function that is usually to set to word length. Dictionary based segmentation strategies can apply two types of decoding algorithms.

The first decoding algorithm is the traditional maximal-matching one. It works on a given plain text T to output the best current word w^* repeatedly with $T{=}t^*$ for the next round as follows,

$$\{w^*, t^*\} = \underset{wt=T}{\operatorname{argmax}}\, g(w) \tag{1}$$

with each $\{w, g(w)\} \in W$. This above algorithm is more precisely referred to as the forward maximal matching (FMM) algorithm Symmetrically, it has an inverse version that works the other way around, and it is referred to backward maximal matching (BMM) algorithm.

The second decoding algorithm is a Viterbi-style one to search for the best segmentation S^* for a text T, as follows:

$$S^* = \underset{w_1 \cdots w_i \cdots w_n = T}{\operatorname{argmax}} \sum_{i=1}^{n} g(w_i), \tag{2}$$

with all $\{w_i, g(w_i)\} \in W$. However, this algorithm subject to the above equation will not work as the goodness function is set to word length, and as the sum of all word lengths will be always the length of the given plain text T. Instead, a so-called shortest path (SP) algorithm will be applied for this case by searching the best segmentation with respect to the following equation,

$$S^* = \underset{w_1 \cdots w_i \cdots w_n = T}{\operatorname{argmin}} n. \tag{3}$$

As it finds a segmentation with minimal number of words, it is named the shortest path.

Traditionally, word segmentation performance is measured by F-score ($F = 2RP/(R + P)$), where the recall (R) and precision (P) are the proportions of the correctly segmented words to all words in, respectively, the gold-standard segmentation and a segmenter's output.

4 Experimental Settings

The MT data set for this study is from the Chinese-to-English patent machine translation subtask of the NTCIR-9 shared task [24]. Both the development and test sets are with single reference. All the data are extracted from patent documents, so it will not be biased towards any existing word segmentation specification that is mostly from the news domain.

The MT training data contains one million sentence pairs; on the Chinese side there are 63.2 million characters, and the English sentences have 35.6 million words. Both the development and test corpora include two thousand sentence pairs. Five-gram language models are trained for both Chinese-to-English and English-to-Chinese translation tasks over the target language data set. No other resources are involved.

The MT system used in this paper is a recent version of Moses[1][25]. We build phrase translations by first acquiring bidirectional GIZA++ alignments [26], the maximal phrase length is set to the default value 7, and using Moses' *grow-diag-final-and* alignment symmetrization heuristic[2]. During decoding, we incorporate the standard eight feature functions of Moses with the lexicalized reordering model. The parameters of these features are tuned with Minimum Error Rate Training (MERT) [26] on the standard development and test sets that were provided by the NTCIR-9 organizers. In addition, we set the maximum

[1] `http://www.statmt.org/moses/`

[2] According to our explorative experiments, this heuristic always outperformed the default setting, *grow-diag-final*.

Table 2. Correlation between F-score and BLEU (%)

Segmenter		CTB3	MSRA2	PKU2
CRF	F-score	94.6	97.2	95.1
	BLEU	31.26	31.82	31.74
FMM	F-score	82.8	86.9	93.3
	BLEU	31.20	31.32	31.70

distortion limit to 11, as in our experiments this setting always produces better performance. We report the MT performance using the BLEU metric on the standard test corpus with the default scorer *multi-bleu.perl* [27]. All BLEU scores in this paper are uncased if English is the target language.

5 Chinese to English Translation

We check multiple assumptions about how word segmentation affects machine translation.

5.1 Segmentation Performance

Existing work has shown that there is no strong correlation between segmentation F-score and BLEU score [8, 7]. We will confirm this observation again.

The F-scores and BLEU scores are listed in parallel in Table 2. Note that it is meaningless to compare performance between different segmentation conventions. For FMM segmenters, their dictionaries are extracted from the respective CRF segmenter outputs on MT training corpora. We may focus on FMM and CRF segmenters for the same convention, the F-score and BLEU score are separated for different corpus, and it is easy to observe that two types of segmenters output similar results though CRF segmenter slightly outperforms the corresponding FMM segmenter if the latter adopts the dictionary whose words are extracted from the segmentation outputs of the former. The F-score was evaluated over the SIGHAN bakeoff test data set. The CRF segmenters output much higher F-scores, but their corresponding BLEU scores are only slightly higher than FMM segmenters. Thus we have shown again that the F-score and BLEU score correlate insignificantly.

5.2 Segmentation Inconsistence

There is a theory about segmentation inconsistency for machine translation, which is that a segmenter that outputs different segmentation outputs for the same input substring between training corpus and development/test corpus or even for the same corpus easily leads to a poor performance on machine translation. This has been well analyzed in the work of Chang et al. [7] and an empirical metric, conditional entropy, has been proposed to measure segmentation inconsistency inside one segmented corpus. This metric partially may explain why a

Table 3. Correlation between differences of F-score and BLEU (%)

Corpus	CTB3	MSRA2	PKU2
F-score	78.6	84.6	82.8
1-F	21.4	15.4	17.2
BLEU diff(%)	0.2	1.6	0.1

dictionary-based segmentation strategy like FMM sometimes outperforms CRF segmenters.

Here, we introduce more experimental facts that may reflect how segmentation strategies vary over machine translation quality.

First, we compare the difference between outputs from FMM and CRF segmenters. For each segmentation convention, the FMM segmenter will still use the dictionary in which words are extracted from the CRF segmenter's output over the MT training corpus. Regarding the segmentation results of the CRF segmenter as the gold standard, an F-score can be calculated over the FMM segmenter' outputs. We will take the F-score as the quantity consistence between two segmentation outputs and that 100% minus the F-score may correspondingly represent the difference between the two outputs. Table 3 shows comparisons between the 1-F and BLEU score relative differences between the FMM and CRF segmenters. This comparison in Table 3 actually discloses that although two types of segmenters, FMM and CRF, output quite different word segmentation results, their MT results are quite close. Such facts suggest that an MT system may accept quite different segmentation inputs for a degree of translation quality and using similar or related linguistic resource, different segmenters may lead to close MT performance. Meanwhile, this also means that we cannot effectively predict BLEU differences only from segmentation difference.

Second, we check if it is sensitive if we apply different segmentation strategies between the MT training set and developement/test sets. Table 4 shows MT results as CRF and FMM segmenters are respectively applied to the training and development/test sets. In the table, segmentation consistency F-scores are calculated on the training corpus, and the BLEU loss ratio is calculated between two average scores as the same and different segmenters are applied to the training and development/test corpora. An obvious BLEU score loss have been observed from the results, and the magnitude of BLEU score change is kept at a similar level as segmentation difference.

For all tree dictionary based segmentation strategies, FMM, BMM and SP, we also do a similar check. Their segmentation differences are in Table 5 as the dictionary is extracted from output of the CRF segmenter on CTB3 convention. The BLEU scores are in Table 6. The results show that even using the same dictionary, segmentation strategy differences cause quite different BLEU scores.

Based on the above two observations: MT quality is sensitive to segmentation strategy choice if the training set and development/test set adopt different segmentation strategies, though apart from this condition, the current MT system is not so sensitive to segmentation strategy choice if the support linguistic

Table 4. BLEU scores as different segmenters for training and dev/test sets(%)

training	dev/test	CTB3	MSRA2	PKU2
CRF	CRF	31.26	31.82	31.74
FMM	FMM	31.20	31.32	31.70
FMM	CRF	27.75	27.11	28.72
CRF	FMM	25.91	26.39	26.99
BLEU loss ratio		14.1	15.3	12.2
1-F		21.4	15.4	17.2

Table 5. Segmentation differences of dictionary based segmenters(%)

	FMM BMM	BMM SP	SP FMM
F-score	78.0	80.9	95.6
1-F	22.0	19.1	4.4

resource is kept unchanged. We then may cautiously conclude that segmentation strategy itself becomes a factor on segmentation consistence analysis, that is, segmentation consistency for MT evaluation should be only measured among the segmentation output given by the same segmentation strategies.

5.3 Different Dictionary Sources

So far, we only adopt dictionaries that are extracted from CRF segmenter outputs for all dictionary-based segmenters. However, for dictionary sources, we may have more choices than segmented corpora for CRF segmenters. All segmented corpora for CRF segmenters are from the SIGHAN Bakeoff shared task and independent of our MT corpus; therefore, they belong to the out-of-domain resources for the MT task. Intuitively, in-domain linguistic resources are always preferable due to it usually bringing about better performance. Compared to building an in-domain segmented corpus for MT tasks, it is much easier to construct an in-domain dictionary.

We then consider two strategies for generating dictionaries from an MT corpus. One is based on unsupervised segmentation over a monolingual corpus, i.e., the Chinese side of the parallel corpus, and the other is based on the alignment model.

Unsupervised segmentation has been empirically studied in the work of Zhao and Kit [23]. According to the empirical results of this work, Accessor Variety (AV) has shown the best goodness function for unsupervised segmentation incorporated with a Viterbi-style decoding algorithm according to equation 3. AV was proposed in [28] as a statistical criterion to measure how likely a substring is a true word. The AV of a substring $x_{i..j}$ is given as follows:

$$AV(x_{i..j}) = \min\{L_{av}(x_{i..j}), R_{av}(x_{i..j})\} \tag{4}$$

Table 6. BLEU scores as using different segmenters for training and dev/test sets(%)

training	FMM	FMM	FMM	BMM	BMM	BMM	SP	SP	SP
dev/test	FMM	BMM	SP	FMM	BMM	SP	FMM	BMM	SP
BLEU	31.20	27.08	30.16	27.62	30.47	28.05	30.42	28.06	31.25

Table 7. Dictionary size(K)

AV	ALIGN	$\text{ALIGN}_{>1}$	CRF-CTB3	CRF-MSRA2	CRF-PKU2
316	417	142	503	460	465

where the left and right accessor variety $L_{av}(x_{i..j})$ and $R_{av}(x_{i..j})$ are the number of distinct predecessor and successor characters, respectively. In practice, the logarithm of AV is actually used as a goodness measure in equation 3.

Note that AV scores should be calculated for possible character n-grams, which would yield too large of a dictionary. Thus, we first use the Viterbi decoding algorithm with all n-gram AV scores to segment the Chinese MT training corpus, then we build a much smaller dictionary by only collecting all words from the segmented text.

Xu et al. [6] proposed a heuristic rule to generate a dictionary from alignment outputs. Firstly, each Chinese character in the corpus is segmented as a word, then an aligner like GIZA++ is used to train an alignment model with this trivially segmented Chinese text. According to alignment outputs, if two or more successive Chinese characters are translated to one English word, then these Chinese characters will be regarded as a word. This word collection strategy may lead to a large dictionary with remarkable noise. Therefore, we introduce a filtering rule by counting aligning times. For example, only if aligning is observed more than once, will those concerned continuous characters be collected as a word. This strategy (it will be noted as $\text{ALIGN}_{>1}$ afterwards.) helps us generate a much smaller dictionary.

Table 7 gives size information for different dictionaries. Numbers of word types generated by CRF segmenters are also given for comparison. All three dictionary-based segmentation approaches, FMM, BMM and SP, are used on all these dictionaries, and the results are in Table 8. *char-seg* in the table means that each character in the corpus will be segmented into a word. The results show that all segmentation strategies may outperform *char-seg*, but their results are not better than those given by every CRF segmenter. However, we also show that the dictionary pruning according to the alignment model can effectively enhance machine translation.

5.4 Segmentation Granularity for Dictionary Approach

Observing that MT is sensitive to segmentation granularity, Chang et al. [7] introduced a novel feature to tune the granularity in the outputs of CRF segmenters. Wang et al. [29] also made the similar observation and proposed using a third-party dictionary to modify a segmented corpus. In this part, we try to

Table 8. BLEU scores of dictionary based segmenters(%)

char-seg	30.14		
dict. / segmenters	FMM	BMM	SP
AV	30.46	30.76	30.62
ALIGN	30.73	30.94	30.90
$\text{ALIGN}_{>1}$	31.26	31.55	31.25

Table 9. BLEU scores over different segmentation granularity(%)

dict. / length	Full	5	4	3	2
CRF-CTB3	31.20	31.06	30.81	31.01	**31.22**
CRF-MSRA2	31.32	31.65	**31.73**	31.36	31.66
CRF-PKU2	31.70	31.30	31.31	**30.72**	31.03
AV	30.46	30.50	30.30	30.64	**30.71**
$\text{ALIGN}_{>1}$	31.26	31.34	31.43	**31.62**	31.04

verify this observation for dictionary segmenters. FMM is adopted as the decoding algorithm and word length is limited to 2,3,4 and 5 characters, respectively[3]. The results in Table 9 show that such granularity tuning is not too significant for dictionary-based segmentation strategies and the improvement sometimes depends on which dictionary source is adopted.

6 English-to-Chinese Translation

English-to-Chinese translation seems like a simple translation direction reversion, but it may follow quite difference principles. As to our best knowledge, few research endeavors have been done on this topic and this work, is the first attempt that comprehensively explores how word segmentation affects English-to-Chinese translation.

As the target language needs word segmentation and none of standard segmentations are available for evaluation, we will have to report the two types of BLEU scores, one is based on character sequences, the other based on word sequences. All results with different segmenters are given in Table 10, 11 and 12,

As the result of *char-seg* is from the direct optimization on character sequences during aligning and decoding, it is not surprising that it receives a character based BLEU score as high as 40.72, which is much better than any other regular word segmenters.

For further comparison, we re-segment the translation output text of *char-seg* and test corpus with the same segmenter, and the word-based BLEU score will be calculated between these two texts. From the results in Table 10, we see that for two of three segmenters, the trivial segmentation strategy, *char-seg*, outperforms CRF segmenters even in terms of word BLEU score, which is quite

[3] Note that most Chinese words are about two-characters long and few Chinese words are longer than five-characters.

Table 10. BLEU scores of English-to-Chinese translation(%): CRF segmenters

Segmenter	BLEU type	CTB3	MSRA2	PKU2
CRF	char	33.16	33.54	32.85
	word	26.11	27.25	25.55
char-seg	word	26.27	21.16	26.27
	char		40.72	

Table 11. BLEU scores of English-to-Chinese translation(%): dictionary segmenters with CRF segmenter generated dictionary

Segmenter	BLEU type	CTB3	MSRA2	PKU2
FMM	char	32.98	33.39	32.49
	word	23.65	24.79	23.90
char-seg	word	22.48	22.87	23.07
	char		40.72	

different from the case of Chinese-to-English translation. These results cast an obvious suspicion on the necessity of word segmentation for English-to-Chinese translation.

As we turn to compare the results of dictionary segmenters, another problem will be disclosed. From the results of Tables 11 and 12, we indeed observe that all dictionary segmenters give higher word BLEU scores than *char-seg*. However, this is not because dictionary segmenters really produce higher word BLEU scores, but that converted word BLEU scores of *char-seg* drop. This case in Table 12 is more serious, where all the converted BLEU scores are only around 10%. Manual observation on $\text{ALIGN}_{>1}$ dictionary shows that too many "words" in it are actually irregular character combinations, not true words. Therefore, this series of experiment results actually show that word BLEU scores may be seriously biased by the low-quality dictionary and it cannot be taken as a reliable metric for English-to-Chinese translation.

Continuing along this train of thought, if we have to take character BLEU as a unique metric for evaluating English-to-Chinese translation, then we will naturally draw a conclusion that word segmentation is not in fact necessary for this type of machine translation task.

7 Finding an Optimal Dictionary

From a linguistic resources perspective, dictionary or segmented corpus, there is not a solid borderline between statistical and dictionary-based segmentation strategies. They can be converted to each other easily. We have let a dictionary segmenter adopt a dictionary collected from the segmentation output of CRF segmenters. Conversely, dictionary segmenters can be used to segment a given text to generate a segmented corpus for training CRF segmenters as well. Our empirical study also shows that when using correlative linguistic resources, either a statistical segmenter or dictionary segmenter gives similar results, and in

Table 12. BLEU scores of English-to-Chinese translation(%): dictionary segmenters with $\text{ALIGN}_{>1}$ dictionary

Segmenter	BLEU type	FMM	BMM	SP
$\text{ALIGN}_{>1}$	char	33.74	32.45	33.23
	word	20.24	19.27	19.99
char-seg	word	10.10	9.96	10.29
	char		40.72	

this case, none of the segmentation strategies work significantly better than the others. In other words, to optimize a word segmenter, we have to optimize the linguistic resources that it relies on.

Here, we propose an empirical dictionary optimization (more precisely, pruning) algorithm to improve the related dictionary-based segmenters. The algorithm is mostly motivated by the empirical observation that most words in a given dictionary actually provide poor information for aligning and decoding in a specific MT task. As a dictionary with n words is given, our task of dictionary optimization is to find a subset of the dictionary to maximize the machine translation performance. However, we will have to examine 2^n subsets without guidance of any priori knowledge, which is computationally intractable. A solution to this difficulty is introducing a metric to assess how much a word is beneficial for machine translation and guide the later dictionary subset selection. So far, no priori metric has been found to measure how good a segmenter is for machine translation. Thus, most related studies have to directly adopt aligner outputs or even BLEU scores to choose a good segmenter. We will exploit both alignment model and BLEU scores given by MERT on the development set, and aligning counter is adopted as the metric to evaluate how well a word inside the dictionary individually contributes to machine translation[4]. This algorithm is given in Algorithm 1. There are two layers of loops in the algorithm, but in practice, this algorithm usually ends after running the MT routine less than 15 times. In addition, against existing dictionary optimization approaches [13, 31], the proposed one is actually non-parametric, which is more convenient and practical for use.

We consider three different dictionaries for the inputs of the proposed dictionary optimization algorithm and FMM is chosen as the decoding algorithm for the Chinese-to-English translation task, and the results on the test set are given in Table 13. All input dictionaries give higher BLEU scores after optimization. The most improvement, a 1.3% BLEU score, is from $\text{ALIGN}_{>1}$, which suggests that an in-domain bilingually motivated dictionary source can bring about better performance.

[4] Actually, we have considered various rank metrics in our early exploration. Ma and Way [13] argued that the co-occurrence frequency (COOC) that was proposed in [30] could be better for ranking words, however, our empirical study shows that COOC may lead to unstable performance for quite a lot of dictionary sources.

Algorithm 1. Dictionary optimization

INPUT An initial dictionary, D
while do
  Segment the MT corpus with D.
  Run GIZA++ for alignment model M.
  Run MERT and receive BLEU score(on the dev set) b.
  Rank all words in D according to aligning times.
  Let $counter$=0 and n=2
  while $counter$ <2 **do**
    Extract top $1/n$ words from D according to aligning times to build dictionary D_n.
    Run GIZA++, MERT and receive BLEU score b_n.
    if $b_n < b_{n-1}$ **then**
      $counter = counter + 1$.
    end if
    $n = n + 1$
  end while
  if max $\{b_i\} < b$ **then**
    return D
  end if
  Let D_0 = arg max$_{D_i}$ b_i and b= max $\{b_i\}$
  Let $D' = D$ - D_0
  According to aligning times in M, divide D' into $2n$ dictionaries, D'_1,...,D'_n, ..., D'_{2n}.
  for top n most-aligned dictionaries, $D'_i, i = 1, ..., n$ **do**
    Segment the MT corpus with D_0+D'_i.
    Run GIZA++, MERT and receive BLEU score b'_i.
  end for
  if max $\{b'_i\} < b$ **then**
    return D_0
  end if
  Let D = arg max$_{D_0+D'_i}$ b'_i and b= max $\{b'_i\}$
end while

Table 13. BLEU scores of segmenters with optimized dictionary (%)

char-seg	30.14		
Dictionary sources	CTB3	AV	ALIGN$_{>1}$
before opti.	31.20	30.46	31.26
after opti.	31.73	31.50	32.56
#running MT routines	6	9	15

Table 14 gives dictionary size information before and after optimization. The results demonstrate that all dictionaries are heavily pruned. The pruning result from dictionary ALIGN$_{>1}$ is especially unusual, as the dictionary with only 7K words that provides the most MT performance improvement among three

Table 14. Dictionary size before and after optimization (K)

Dictionary sources	CTB3	AV	$ALIGN_{>1}$
before opti.	503	316	142
after opti.	35	32	7

dictionary sources is at last obtained through the proposed algorithm, while most previous work often reports that dictionaries with tens of thousands of words at least are required [8, 31].

8 Conclusions

As word segmentation has been shown helpful for Chinese-to-English machine translation, we investigate what type of segmentation strategy can help machine translation work better. First, our empirical study shows that word segmentation is a necessity for Chinese-to-English translation, but not for the case of English-to-Chinese translation. Second, both statistical and dictionary-based word segmentation strategies are examined. We actually show for better machine translation, the key is not the segmentation strategy choice, but the linguistic resources for supporting segmenters. Third, an easy-implemented dictionary optimization algorithm is proposed to improve segmentation for machine translation. Our experiment results show that this approach is effective for different dictionary sources; however,better results come from a domain adaptive and bilingually motivated dictionary, which gives the most improvement with a BLEU score as high as 1.30 %.

References

1. Sproat, R., Emerson, T.: The first international chinese word segmentation bakeoff. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, pp. 133–143 (2003)
2. Emerson, T.: The second international chinese word segmentation bakeoff. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, pp. 123–133 (2005)
3. Levow, G.A.: The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, pp. 108–117 (2006)
4. Gao, J., Li, M., Wu, A., Huang, C.N.: Chinese word segmentation and named entity recognition: A pragmatic approach. Computational Linguistics 31, 531–574 (2005)
5. Li, M., Zong, C., Ng, H.T.: Automatic evaluation of chinese translation output: word-level or character-level? In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, HLT 2011, June 19-24, vol. 2, pp. 159–164. Association for Computational Linguistics, Portland (2011)

6. Xu, J., Zens, R., Ney, H.: Do we need chinese word segmentation for statistical machine translation. In: Proceedings of the Third SIGHAN Workshop on Chinese Language Learning, Barcelona, Spain, pp. 122–128 (2004)
7. Chang, P.C., Galley, M., Manning, C.D.: Optimizing Chinese word segmentation for machine translation performance. In: Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, USA, pp. 224–232 (2008)
8. Zhang, R., Yasuda, K., Sumita, E.: Improved statistical machine translation by multiple chinese word segmentation. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 216–223. Association for Computational Linguistics, Columbus (2008)
9. Xu, J., Matusov, E., Zens, R., Ney, H.: Integrated chinese word segmentation in statistical machine translation. In: Proceedings of IWSLT, Pittsburgh, PA, pp. 141–147 (2005)
10. Dyer, C., Muresan, S., Resnik, P.: Generalizing word lattice translation. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH, USA, pp. 1012–1020 (2008)
11. Xu, J., Gao, J., Toutanova, K., Ney, H.: Bayesian semi-supervised chineseword segmentation for statistical machine translation. In: Proceedings of COLING 2008, Manchester, UK, pp. 1017–1024 (2008)
12. Nguyen, T., Vogel, S., Smith, N.A.: Nonparametric word segmentation for machine translation. In: Proceedings of COLING 2010, Beijing, China, pp. 815–823 (2010)
13. Ma, Y., Way, A.: Bilingually motivated domain-adapted word segmentation for statistical machine translation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 549–557. Association for Computational Linguistics, Athens (2009)
14. Koehn, P., Knight, K.: Empirical methods for compound splitting. In: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics, pp. 187–193. Association for Computational Linguistics, Budapest (2003)
15. Habash, N., Sadat, F.: Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the Human Language Technology Conference of the NAACL, pp. 49–52. Association for Computational Linguistics, New York City (2006)
16. Paul, M., Finch, A., Sumita, E.: Integration of multiple bilingually-learned segmentation schemes into statistical machine translation. In: Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pp. 400–408. Association for Computational Linguistics, Uppsala (2010)
17. Low, J.K., Ng, H.T., Guo, W.: A maximum entropy approach to Chinese word segmentation. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, pp. 161–164 (2005)
18. Tseng, H., Chang, P., Andrew, G., Jurafsky, D., Manning, C.: A conditional random field word segmenter for SIGHAN bakeoff 2005. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, pp. 168–171 (2005)
19. Zhao, H., Huang, C.N., Li, M.: An improved Chinese word segmentation system with conditional random field. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, pp. 162–165 (2006)
20. Zhao, H., Kit, C.: Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In: Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, Hyderabad, India, pp. 106–111 (2008)

21. Xue, N., Shen, L.: Chinese word segmentation as LMR tagging. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, in Conjunction with ACL 2003, Sapporo, Japan, pp. 176–179 (2003)
22. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: COLING 2004, Geneva, Switzerland, pp. 562–568 (2004)
23. Zhao, H., Kit, C.: An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. In: The Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India, pp. 9–16 (2008)
24. Goto, I., Lu, B., Chow, K.P., Sumita, E., Tsou, B.K.: Overview of the patent machine translation task at the ntcir-9 workshop. In: Proceedings of NTCIR-9 Workshop Meeting, Tokyo, Japan, pp. 559–578 (2011)
25. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL 2003, vol. 1, pp. 48–54. Association for Computational Linguistics, Stroudsburg (2003)
26. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. 29, 19–51 (2003)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 311–318. Association for Computational Linguistics, Stroudsburg (2002)
28. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor variety criteria for Chinese word extraction. Computational Linguistics 30, 75–93 (2004)
29. Wang, Y., Uchimoto, K., Kazama, J., Kruengkrai, C., Torisawa, K.: Adapting chinese word segmentation for machine translation based on short units. In: Proceedings of LREC 2010, Malta, pp. 1758–1764 (2010)
30. Melamed, I.D.: Models of translational equivalence among words. Computational Linguistics 26, 221–249 (2000)
31. Ma, J., Matsoukas, S.: BBN's systems for the Chinese-English sub-task of the NTCIR-9 PatentMT evaluation. In: Proceedings of NTCIR-9 Workshop Meeting, Tokyo, Japan, pp. 579–584 (2011)