Refining Word Segmentation Using a Manually Aligned Corpus for Statistical Machine Translation

Xiaolin Wang **Masao Utiyama** **Andrew Finch** **Eiichiro Sumita**
National Institute of Information and Communications Technology
{xiaolin.wang,mutiyama,andrew.finch,eiichiro.sumita}@nict.go.jp

Abstract

Languages that have no explicit word delimiters often have to be segmented for statistical machine translation (SMT). This is commonly performed by automated segmenters trained on manually annotated corpora. However, the word segmentation (WS) schemes of these annotated corpora are handcrafted for general usage, and may not be suitable for SMT. An analysis was performed to test this hypothesis using a manually annotated word alignment (WA) corpus for Chinese-English SMT. An analysis revealed that 74.60% of the sentences in the WA corpus if segmented using an automated segmenter trained on the Penn Chinese Treebank (CTB) will contain conflicts with the gold WA annotations. We formulated an approach based on word splitting with reference to the annotated WA to alleviate these conflicts. Experimental results show that the refined WS reduced word alignment error rate by 6.82% and achieved the highest BLEU improvement (0.63 on average) on the Chinese-English open machine translation (OpenMT) corpora compared to related work.

1 Introduction

Word segmentation is a prerequisite for many natural language processing (NLP) applications on those languages that have no explicit space between words, such as Arabic, Chinese and Japanese. As the first processing step, WS affects all successive steps, thus it has a large potential impact on the final performance. For SMT, the unsupervised WA, building translation models and reordering models, and decoding are all based on segmented words.

Automated word segmenters built through supervised-learning methods, after decades of intensive research, have emerged as effective solutions to WS tasks and become widely used in many NLP applications. For example, the Stanford word segmenter (Xue et al., 2002)[1] which is based on conditional random field (CRF) is employed to prepare the official corpus for NTCIR-9 Chinese-English patent translation task (Goto et al., 2011).

However, one problem with applying these supervised-learning word segmenters to SMT is that the WS scheme of annotating the training corpus may not be optimal for SMT. (Chang et al., 2008) noticed that the words in CTB are often too long for SMT. For example, a full Chinese personal name which consists of a family name and a given name is always taken as a single word, but its counterpart in English is usually two words.

Manually WA corpora are precious resources for SMT research, but they used to be only available in small volumes due to the production cost. For example, (Och and Ney, 2000) initially annotated 447 English-French sentence pairs, which later became the test data set in ACL 2003 shared task on word alignment (Mihalcea and Pedersen, 2003), and was used frequently thereafter (Liang et al., 2006; DeNero and Klein, 2007; Haghighi et al., 2009)

For Chinese and English, the shortage of manually WA corpora has recently been relieved by the linguistic data consortium (LDC) [2] GALE Chinese-English word alignment and tagging training corpus (the GALE WA corpus)[3]. The corpus is considerably large, containing 4,735 documents, 18,507 sentence pairs, 620,189 Chinese tokens, 518,137 English words, and 421,763

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

[2] http://catalog.ldc.upenn.edu

[3] Catalog numbers: LDC2012T16, LDC2012T20, LDC2012T24 and LDC2013T05.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1654–1664,
October 25-29, 2014, Doha, Qatar.

alignment annotations. The corpus carries no Chinese WS annotation, and the WA annotation was performed between Chinese characters and English words. The alignment identifies minimum translation units and relations [4], referred as atomic blocks and atomic edges, respectively, in this paper. Figure 1 shows an example that contains six atomic edges.

Visual inspection of the segmentation of an automatic segmenter with reference to a WA corpus revealed a number of inconsistencies. For example, consider the word "bao fa" in Figure 1. Empirically we observed that this word is segmented as a single token by an automatic segmenter trained on the CTB, however, this segmentation differs with the alignment in the WA corpus, since its two components are aligned to two different English words. Our hypothesis was that the removal of these inconsistencies would benefit machine translation performance (this is explained further in Section 2.3), and we explored this idea in this work.

This paper focuses on optimizing Chinese WS for Chinese-English SMT, but both the research method and the proposed solution are language-independent. They can be applied to other language pairs.

The major contributions of this paper include,

- analyze the CTB WS scheme for Chinese-English SMT;
- propose a lexical word splitter to refine the WS;
- achieve a BLEU improvement over a baseline Stanford word segmenter, and a state-of-the-art extension, on Chinese-English OpenMT corpora.

The rest of this paper is organized as follows: first, Section 2 analyzes WS using a WA corpus; next, Section 3 proposes a lexical word splitter to refine WS; then, Section 4 evaluates the proposed method on end-to-end SMT as well as word segmentation and alignment; after that, Section 5 compares this work to related work; finally, Section 6 concludes this paper.

[4]Guidelines for Chinese-English Word Alignment(Version 4.0)

2 Analysis of a General-purpose Automatic Word Segmenter

This section first briefly describes the GALE WA corpus, then presents an analysis of the WS arising from a CTB-standard word segmenter with reference to the segmentation of the atomic blocks in the GALE WA corpus, finally the impact of the findings on SMT is discussed.

2.1 GALE WA corpus

The GALE WA corpus was developed by the LDC, and was used as training data in the DARPA GALE global autonomous language exploitation program [5]. The corpus incorporates linguistic knowledge into word aligned text to help improve automatic WA and translation quality. It employs two annotation schemes: alignment and tagging (Li et al., 2010). Alignment identifies minimum translation units and translation relations; tagging adds contextual, syntactic and language-specific features to the alignment annotation. For example, the sample shown in Figure 1 carries tags on both alignment edges and tokens.

The GALE WA corpus contains 18,057 manually word aligned Chinese and English parallel sentences which are extracted from newswire and web blogs. Table 1 presents the statistics on the corpus. One third of the sentences are approximately newswire text, and the remainder consists of web blogs.

2.2 Analysis of WS

In order to produce a Chinese word segmentation consistent with the CTB standard we used the Stanford Chinese word segmenter with a model trained on the CTB corpus. We will refer to this as the 'CTB segmenter' in the rest of this paper.

The Chinese sentences in the GALE WA corpus were first segmented by the CTB segmenter, and the predicted words were compared against the atomic blocks with respect to the granularity of segmentation. The analysis falls into the following three categories, two of which may be potentially harmful to SMT:

- Fully consistent: the word locates within the block of one atomic alignment edge. For example, in Figure 2(a), the Chinese text has

[5]https://catalog.ldc.upenn.edu/LDC2012T16

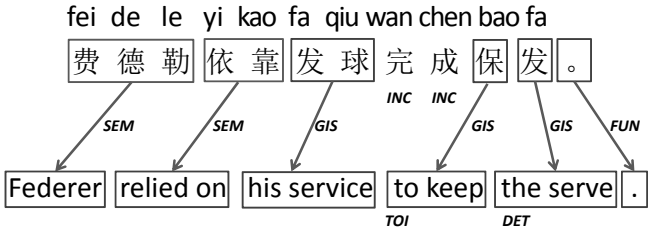

Figure 1: Example from the GALE WA corpus. Each line arrow represents an atomic edge, and each box represents an atomic block. SEM (semantic), GIS (grammatically inferred semantic) and FUN (function) are tags of edges. INC (not translated), TOI (to-infinitive) and DET (determiner) are tags of tokens.

Genre	# Files	# Sentences†	# CN tokens	# EN tokens	# Alignment edges
Newswire	2,175	6,218	246,371	205,281	164,033
Web blog	2,560	11,839	373,818	312,856	257,730
Total	4,735	18,057	620,189	518,137	421,763

Table 1: GALE WA corpus. † Sentences rejected by the annotators are excluded.

four atomic blocks; the CTB segmenter produces five words which all locate within the blocks, so they are all small enough.

- Alignment inconsistent: the word aligns to more than one atomic block, but the target expression is contiguous, allowing for correct phrase pair extraction (Zens et al., 2002). For example, in Figure 2(b), the characters in the word "shuang fang", which is produced by the CTB segmenter, contains two atomic blocks, but the span of the target "to both side" is continuous, therefore the phrase pair "shuang fang ||| to both sides" can be extracted.

- Alignment inconsistent and extraction hindered: the word aligned to more than one atomic block, and the target expression is not contiguous, which hinders correct phrase pair extractions. For example, in Figure 2(c), the word "zeng chan" has to be split in order to match the target language.

Table 2 shows the statistics of the three categories of CTB WS on the GALE WA corpus. 90.74% of the words are fully consistent, while the remaining 9.26% of the words have inconsistent alignments. 74.60% of the sentences contain this problem. The category with inconsistent alignment and extraction hindered only accounts for 0.46% of the words, affecting 9.06% of the sentences.

2.3 Impact of WS on SMT

The word alignment has a direct impact on the nature of both the translation model, and lexical reordering model in a phrase-base SMT system. The words in last two categories are all longer than an atomic block, which might lead to problems in the word alignment in two ways:

- First, longer words tend to be more sparse in the training corpus, thus the estimated distribution of their target phrases are less accurate.

- Second, the alignment from them to target sides are one-to-many, which is much more complicated and requires fertilized alignment models such as IBM model 4 – 6 (Och and Ney, 2000).

The words in the category of "fully consistent" can be aligned using simple models, because the alignment from them to the target side are one-to-one or many-to-one, and simple alignment models such as IBM model 1, IBM model 2 and HMM model are sufficient (Och and Ney, 2000).

3 Refining the Word Segmentation

In the last subsection, it was shown that 74.60% of parallel sentences were affected by issues related to under-segmentation of the corpus. Our hypothesis is that if these words are split into pieces that match English words, the accuracy of the unsupervised WA as well as the translation quality will be improved. To achieve this, we adopt a splitting

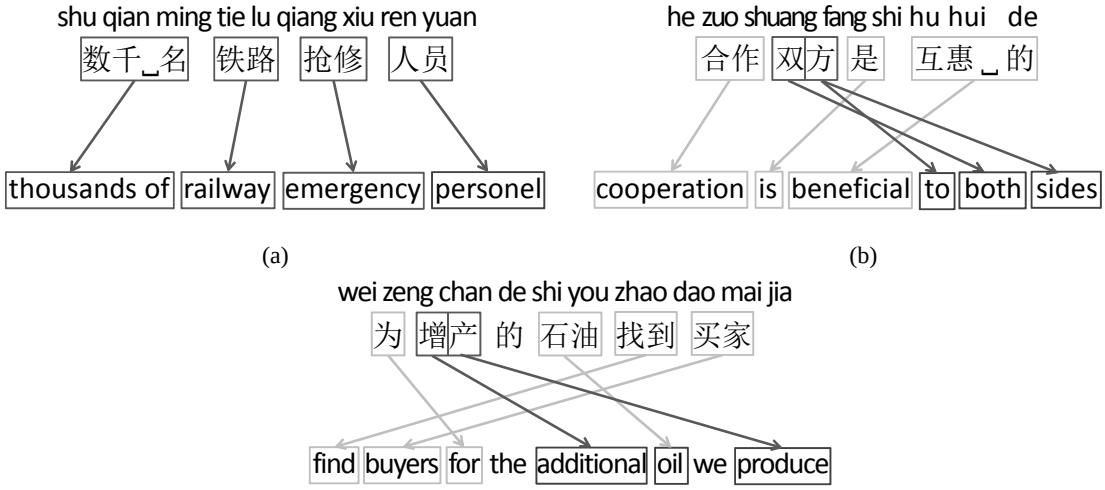

Figure 2: Examples of automated WS on manually WA corpus: (a) Fully consistent; (b) Alignment inconsistent; (c) Alignment inconsistent and extraction hindered. The Chinese words separated by white space are the output of the CTB segmenter. Arrows represent the alignment of atomic blocks. Note that "shuang fang" and "zeng chan" are words produced by the CTB segmenter, but consist of two atomic blocks.

Category	Count	Word Ratio	Sentence Ratio
Fully consistent	355,702	90.74%	25.40%†
Alignment inconsistent	34,464	8.81%	65.54%
Alignment inconsistent & extraction hindered	1,830	0.46%	9.06%
Sum of conflict ‡	36,294	9.26%	74.60%

Table 2: CTB WS on GALE WA corpus: † All words are fully consistent; ‡ Alignment inconsistent plus alignment inconsistent & extraction hindered

strategy, based on a supervised learning approach, to re-segment the corpus. This subsection first formalizes the task, and then presents the approach.

3.1 Word splitting task

The word splitting task is formalized as a sequence labeling task as follows: each word (represented by a sequence of characters $\mathbf{x} = x_1 \dots x_T$ where T is the length of sample) produced by the CTB segmenter is a sample, and a corresponding sequence of binary boundary labels $\mathbf{y} = y_1 \dots y_T$ is the learning target,

$$y_t = \begin{cases} 1 & \text{if there is a split point} \\ & \quad \text{between } c_t \text{ and } c_{t-1}; \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The sequence of boundary labels is derived from the gold WA annotation as follows: for a sequence of two atomic blocks, where the first character of the second block is x_t, then the label $y_t = 1$. Figure 3 presents several samples extracted from the examples in Figure 2.

Input	*Target*	***y***
铁路	铁路	0 0
双方	双␣方	0 1
出版业	出版␣业	0 0 1
增 产	增␣产	0 1

Figure 3: Samples of word splitting task

Each word sample may have no split point, one split point or multiple split points, depending on the gold WA annotation. Table 3 shows the statistics of the word splitting data set which is built from the GALE manual WA corpus and the CTB segmenter's output, where 2000 randomly sampled sentences are taken as a held-out test set.

Set	# Sentences	# Samples	# Split points	# Split points per sample
Train.	16,057	348,086	32,337	0.0929
Test	2,000	43,910	3,929	0.0895

Table 3: Data set for learning the word splitting

3.2 CRF approach

This paper employs a condition random field (CRF) to solve this sequence labeling task (Lafferty et al., 2001). A linear-chain CRF defines the conditional probability of $\mathbf{y}$ given $\mathbf{x}$ as,

$$P_\Lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}}(\sum_{t=1}^{T}\sum_k \lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t)), \tag{2}$$

where $\Lambda = \{\lambda_1, \ldots\}$ are parameters, $Z_\mathbf{x}$ is a per-input normalization that makes the probability of all state sequences sum to one; $f_k(y_{t-1}, y_t, \mathbf{x}, t)$ is a feature function which is often a binary-valued sparse feature. The training of CRF model is to maximize the likelihood of training data together with a regularization penalty to avoid over-fitting as (Peng et al., 2004; Peng and McCallum, 2006),

$$\Lambda^* = \underset{\Lambda}{argmax}(\sum_i logP_\Lambda(\mathbf{y}_i|\mathbf{x}_i) - \sum_k \frac{\lambda_k^2}{2\delta_k^2}), \tag{3}$$

where ($\mathbf{x}$,$\mathbf{y}$) are training samples; the hyperparameter δ_k can be understood as the variance of the prior distribution of λ_k. When predicting the labels of test samples, the CRF decoder searches for the optimal label sequence y^* that maximizes the conditional probability,

$$\mathbf{y}^* = \underset{\mathbf{y}}{argmax}\, P_\Lambda(\mathbf{y}|\mathbf{x}). \tag{4}$$

In (Chang et al., 2008) a method is proposed to select an appropriate level of segmentation granularity (in practical terms, to encourage smaller segments). We call their method "length tuner". The following artificial feature is introduced into the learned CRF model:

$$f_0(\mathbf{x}, y_{t-1}, y_t, 1) = \begin{cases} 1 & \text{if } y_t = +1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The weight λ_0 of this feature is set by hand to bias the output of CRF model. By way of explanation, a very large positive λ_0 will cause every character to be segmented, or conversely a very large negative λ_0 will inhibit the output of segmentation boundaries. In their experiments, $\lambda_0 = 2$ was used to force a CRF segmenter to adopt an intermediate granularity between character and the CTB WS scheme. Compared to the length tuner, our proposed method exploits lexical knowledge about word splitting, and we will therefore refer to it as the "lexical word splitter" or "lexical splitter" for short.

3.3 Feature Set

The features $f_k(y_{t-1}, y_t, \mathbf{x}, t)$ we used include the WS features from the Chinese Stanford word segmenter and a set of extended features described below. The WS features are included because the target split points may share some common characteristics with the boundaries in the CTB WS scheme.

The extended features consists of four types – named entities, word frequency, word length and character-level unsupervised WA. For each type of the feature, the value and value concatenated with previous or current character are taken as sparse features (see Table 4 for details). The real values of word frequency, word length and character-level unsupervised WA are converted into sparse features due to the routine of CRF model.

The character-level unsupervised alignment feature is inspired by the related works of unsupervised bilingual WS (Xu et al., 2008; Chung and Gildea, 2009; Nguyen et al., 2010; Michael et al., 2011). The idea is that the character-level WA can approximately capture the counterpart English expression of each Chinese token, and source tokens aligned to different target expressions should be split into different words (see Figure 4 for an illustration).

The values of the character-level alignment features are obtained through building a dictionary. First, unsupervised WA is performed on the SMT training corpus where the Chinese sentences are treated as sequences of characters; then, the Chinese sentences are segmented by CTB segmenter and a dictionary of segmented words are built; finally, for each word in the dictionary, the relative frequency of being split at a certain position is cal-

Feature	Definition	Example
NE	NE tag of current word	Geography:NE
NE-C_{-1}	NE concatenated with previous character	Geo.-ding:NE-C_{-1}
NE-C_0	NE concatenated with current character	Geo.-mei:NE-C_0
Frequency	Nearest integer of negative logarithm of word frequency	5†:Freq
Freq.-C_{-1}	Frequency concatenated with previous character	5-ding:Freq-C_{-1}
Freq.-C_0	Frequency concatenated with current character	5-mei:Freq-C_0
Length	Length of current word (1,2,3,4,5,6,7 or $\geqslant$7)	4:Len
Len.-Position	Length concatenated with the position	4-2:Len-Pos
Len.-C_{-1}	Length concatenated with previous character	4-ding:Len-C_{-1}
Len.-C_0	Length concatenated with current character	4-mei:Len-C_0
Char. Align.	Five-level relative frequency of being split	0.4‡:CA
C.A.-C_{-1}	C.A. concatenated with previous character	0.4-ding:CA-C_{-1}
C.A.-C_0	C.A. concatenated with current character	0.4-mei:CA-C_0

Table 4: Extended features used in the CRF model for word splitting. The example shows the features used in the decision whether to split the Chinese word "la ding mei zhou" (Latin America, the first four Chinese characters in Figure 4) after the second Chinese character. † Round(-$\log_{10}$(0.00019)); ‡ Round(0.43 $\times$ 5) / 5

Figure 4: Illustration of character-level unsupervised alignment features. The dotted lines are word boundaries suggested by the alignment.

culated as,

$$f_{CA}(w,i) = \frac{n_i}{n_w} \quad (6)$$

where w is a word, i is a splitting position (from 1 to the length of w minus 1); n_i is the number of times the words as split at position i according to the character-level alignment, that is, the character before and after i are aligned to different English expressions; n_w is occurrence count of word w in the training corpus.

4 Experiments

In the last section we found that 9.26% of words produced by the CTB segmenter have the potential to cause problems for SMT, and propose a lexical word splitter to address this issue through segmentation refinement. This section contains experiments designed to empirically evaluate the proposed lexical word splitter in three aspects: first, whether the WS accuracy is improved; second, whether the accuracy of the unsupervised WA during training SMT systems is improved; third, whether the end-to-end translation quality is improved.

This section first describes the experimental methodology, then presents the experimental results, and finally illustrates the operation of our proposed method using a real example.

4.1 Experimental Methodology

4.1.1 Experimental Corpora

The GALE manual WA corpus and the Chinese to English corpus from the shared task of the NIST open machine translation (OpenMT) 2006 evaluation [6] were employed as the experimental corpus (Table 5).

The experimental corpus for WS was constructed by first segmenting 2000 held out sentences from the GALE manual WA corpus with the Stanford segmenter, and then refining the segmentation with the gold alignment annotation. For example, the gold segmentation for the examples in Figure 2 is presented in Figure 5. Note that this test corpus is intended to represent an oracle segmentation for our proposed method, and serves primarily to gauge the improvement of our method over the baseline Stanford segmenter, relative to an upper bound.

[6] http://www.itl.nist.gov/iad/mig/tests/mt/2006/

数千 ␣ 名 ␣ 铁路 ␣ 抢修 ␣ 人员

合作 ␣ 双 ␣ 方 ␣ 是 ␣ 互惠 ␣ 的

中国 ␣ 出版 ␣ 业 ␣ 发展

为 ␣ 增 ␣ 产 ␣ 的 ␣ 石油 ␣ 找到 ␣ 买家

Figure 5: Examples of gold WS for evaluation

Set	# sent. pairs	# CN tokens	# EN tokens
Train.	442,967	19,755,573	13,444,927
Eval02	878†	38,204	105,944
Eval03	919†	40,900	113,970
Eval04	1,597†	71,890	207,279
Eval05	1,082†	50,461	138,952
Eval06	1,664†	62,422	189,059

Table 5: NIST Open machine translation 2006 Corpora. † Number of sentence samples which contain one Chinese sentence and four English reference sentences.

The experimental corpus for unsupervised WA was the union set of the NIST OpenMT training set and the 2000 test sentence pairs from GALE WA corpus. We removed the United Nations corpus from the NIST OpenMT constraint training resources because it is out of domain.

The main result of this paper is the evaluation of the end-to-end performance of an SMT system. The experimental corpus for this task was the NIST OpenMT corpus. The data set of the NIST evaluation 2002 was used as a development set for MERT tuning (Och, 2003), and the remaining data sets of the NIST evaluation from 2003 to 2006 were used as test sets. The English sentences were tokenized by Stanford toolkit [7] and converted to lowercase.

4.1.2 Evaluation

The performance of WS was measured by precision, recall and F_1 of gold words (Sproat and Emerson, 2003),

The performance of unsupervised WA in the SMT training procedure was measured through alignment error rate (AER)(Och and Ney, 2000; Liang et al., 2006). Sure alignment edges and possible alignment edges were not distinguished in this paper as no such tags are found in GALE manual WA corpus.

The performance of SMT was measured using BLEU (Papineni et al., 2002).

[7] http://nlp.stanford.edu/software/corenlp.shtml

4.1.3 Baseline Methods

Two Chinese WS methods were taken as the baseline methods in this paper. One method was the CTB segmenter, that is, Stanford Chinese word segmenter with the model trained on CTB corpus. The other method was the length tuner in (Chang et al., 2008), which added a constant into the confidence scores of a trained CRF word segmenter to encourage it to output more word boundaries (see Section 3.2 for details).

4.1.4 Implementation and Parameter settings

The proposed lexical word splitter was implemented on the CRF model toolkit released with the Stanford segmenter (Tseng et al., 2005). The regularity parameters δ_k are set to be 3, the same as the Stanford segmenter, because no significant performance improvements were observed by tuning that parameter.

To extract features for the word splitter, the Stanford named entity recognizer (Finkel et al., 2005)[8] was employed to obtain the tags of named entities. Word frequencies were caculated from the source side of SMT training corpus. The character-level unsupervised alignment was conducted using GIZA++ (Och and Ney, 2003)[9].

The length tuner reused the CRF model of CTB segmenter. The parameter λ_0 was tuned through the grid search in (Chang et al., 2008), that is, observing the BLEU score on the SMT development set varing from $\lambda_0 = 0$ to $\lambda_0 = 32$. The grid search showed that $\lambda_0 = 2$ was optimal, agreeing with the value in (Chang et al., 2008).

Moses (Koehn et al., 2007)[10], a state-of-the-art phrase-based SMT system, was employed to perform end-to-end SMT experiments. GIZA++ was employed to perform unsupervised WA.

4.2 Experimental Results

4.2.1 Word Segmentation

The WS performance of CTB segmenter, length tuner and the proposed lexical splitter are presented in Table 6. The proposed method achieves the highest scores on all the criterion of F_1, precision and recall. The length tuner outperforms the CTB segmenter in terms of recall, but with lower precision.

[8] http://nlp.stanford.edu/software/CRF-NER.shtml

[9] http://www.statmt.org/moses/giza/GIZA++.html

[10] http://www.statmt.org/moses/

WS	F_1	Prec.	Recall
CTB segmenter	0.878	0.917	0.842
Length tuner	0.873	0.894	0.852
Lexical splitter	**0.915**	**0.922**	**0.908**

Table 6: Performance of WS

WS	AER	Prec.	Recall
CTB segmenter	0.425	0.622	0.534
Length tuner	0.417	0.642	0.535
Lexical splitter	**0.396**	**0.674**	**0.547**

Table 7: Performance of unsupervised WA using different WS strategies

4.2.2 Word Alignment

The WA performance of the CTB segmenter, length tuner and the proposed lexical spliter is presented in Table 7. Both lexical splitter and length tuner outperform the CTB segmenter, indicating the splitting words into smaller pieces can improve the accuracy of unsupervised WA. This result supports the finding in (Chang et al., 2008) that the segment size from CTB WS is too large for SMT. In addition, the proposed lexical splitter significantly outperforms the length tuner.

4.2.3 Machine Translation

The end-to-end SMT performance of CTB segmenter, length tuner and the proposed lexical spliter are presented in Table 8. Each experiment was performed three times, and the average BLEU and standard derivation were calculated, because there is randomness in the results from MERT. The proposed lexical splitter outperformed the two baselines on all the test sets, and achieves an average improvement of 0.63 BLEU percentage points, indicating that the proposed method can effectively improve the translation quality. The length tuner also outperforms the CTB segmenter, but the average improvement is 0.15 BLEU percentage points, much less than the proposed methods.

4.3 Analysis

Figure 6 presents an example from the test corpus, which demonstrates how the proposed lexical splitter splits words more accurately than the baseline length tuner method. Two words in the segmentation result of the CTB segmenter are worthy of attention. The first one is “yang nian”(the year of goat), the lexical splitter split this word and got the right translation, while the length tuner did not split it. The second is “rong jing”(booming or prosperity), the length tuner split this word, which resulted in wrong translations, while the lexical splitter avoided this mistake.

5 Comparison to Related Work

The most similar work in the literature to the proposed method is the the length tuner method proposed by (Chang et al., 2008). This method also encourages the generation of more words during segmentation by using a single parameter that can be use to control segment length. Our method differs from theirs in that it is able to acquire vocabulary knowledge from word alignments that can be used to more accurately split words into segments suitable for machine translation.

There is large volume of research using bilingual unsupervised and semi-supervised WS to address the problem of optimizing WS for SMT (Xu et al., 2008; Chung and Gildea, 2009; Nguyen et al., 2010; Michael et al., 2011). The main difference with our approach is that they use automatic WA results, most often obtained using the same tools as are used in training SMT systems. One of the main problems of using unsupervised WA is that it is noisy, and therefore, employing iterative optimization methods to refine the results of unsupervised WA is a key issue in their research, for example boosting (Ma and Way, 2009; Michael et al., 2011), expectation maximization (Chung and Gildea, 2009), Bayesian sampling (Xu et al., 2008; Nguyen et al., 2010), or heuristic search (Zhao et al., 2013). Nevertheless, noisy WA makes both analyzing WS and improving SMT quality quite hard. In contrast, by using manual WA, we can clearly analyze the segmentation problems (Section 2), and train supervised models to solve the problem (Section 3).

As far as we are aware, among related work on WS, our method achieves the highest BLEU improvement relative to the start-of-the-art WS – the Stanford Chinese word segmenter – on the Chinese-English OpenMT corpora. The methods proposed in (Ma and Way, 2009; Chung and Gildea, 2009) fail to outperform the Stanford Chinese word segmenter on Chinese-English OpenMT corpora. The length tuner method proposed in (Chang et al., 2008) is less effective to ours according to the experimental results in this paper.

WS	eval03	eval04	eval05	eval06	improve
CTB segmenter	31.89 ± 0.09	32.73 ± 0.19	31.03 ± 0.16	31.38 ± 0.23	
Length tuner	32.06 ± 0.07	32.74 ± 0.10	31.34 ± 0.11	31.50 ± 0.11	0.15 ± 0.12
Lexical splitter	**32.55** ± **0.18**	**32.94** ± **0.11**	**31.87** ± **0.15**	**32.17** ± **0.35**	**0.63** ± **0.29**

Table 8: Performance (BLEU) of SMT

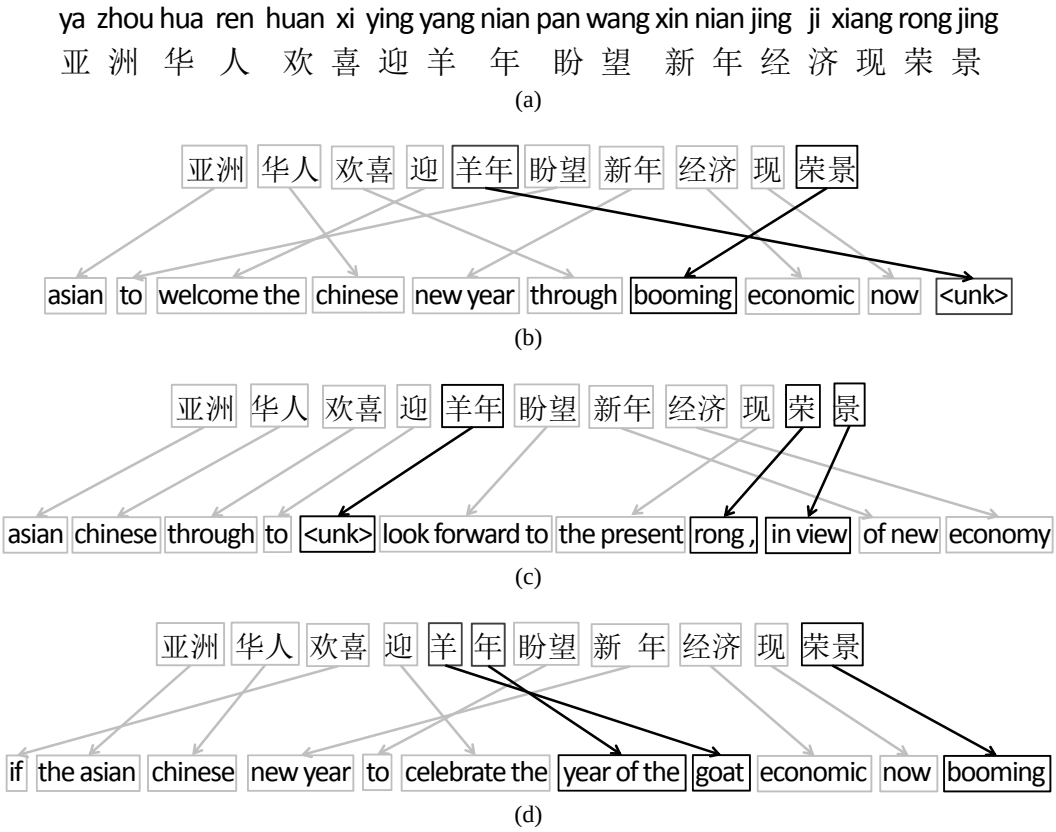

Figure 6: Example of SMT from test sets. (a) source; (b) CTB segmenter; (c) length tuner; (d) lexical splitter. The four gold references are: “ethnic chinese in asia celebrate year of goat and hope for economic prosperity in new year”, “ asian chinese celebrate the arrival of the year of sheep and wish a prosperous new year”, “ asian chinese happily welcome the year of goat , expecting economic prosperity in new year”,“asian chinese happily welcomed year of the goat , praying for prosperity in the new year”

6 Conclusion

This paper is concerned with the role of word segmentation in Chinese-to-English SMT. We explored the use of a manually annotated word alignment corpus to refine word segmentation for machine translation. Based on an initial finding that 74.60% of running sentences in the WA corpus have segmentation inconsistent with a gold WA annotation, we proposed a supervised lexical resegmentation model to modify the WS in order to relieve these issues.

Our main experimental results show that the proposed approach is capable of improving both alignment quality and end-to-end translation quality. The proposed method achieved the highest BLEU score relative to a number of respectable baseline systems that included the Stanford word segmenter, and an improved Stanford word segmenter that could be tuned for segment length. No language-specific techniques other than a manually aligned corpus were employed in this paper, thus the approach can applied to other SMT language pairs that require WS.

In the future, we plan to explore combining multiple source words which are aligned to the same target words. This is the symmetric topic of the post word splitting which is studied in this paper. The effect of this word combination oper-

ation on SMT is non-trivial. On one hand, it can reduce the ambiguity in the source side. On the other hand, it may cause sparseness problems.

Acknowledgements

We thank the three reviewers for their valuable comments. We also thank the Stanford natural language processing group for releasing the source codes of their word segmenter.

References

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 224–232. Association for Computational Linguistics.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 718–726. Association for Computational Linguistics.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *ACL*, volume 45, page 17.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised itg models. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 923–931. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Association for Computing Machinery.

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie Strassel, and Kazuaki Maeda. 2010. Enriching word alignment with linguistic tags. In *LREC*, pages 2189–2195.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Yanjun Ma and Andy Way. 2009. 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*, pages 549–557. Association for Computational Linguistics.

Paul Michael, Andrew Finch, and Eiichiro Sumita. 2011. Integration of multiple bilingually-trained segmentation schemes into statistical machine translation. *IEICE transactions on information and systems*, 94(3):690–697.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen, editors, *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, Edmonton, Alberta, Canada, May 31. Association for Computational Linguistics.

ThuyLinh Nguyen, Stephan Vogel, and Noah A Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 815–823. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 1086–1090. Association for Computational Linguistics.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42(4):963–979.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. *Computer Science Department Faculty Publication Series*.

Richard Sproat and Thomas Emerson. 2003. The first international chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 133–143. Association for Computational Linguistics.

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 4th SIGHAN Workshop on Chinese Language Processing*, volume 171. Jeju Island, Korea.

Jia Xu, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1017–1024. Association for Computational Linguistics.

Nianwen Xue, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-scale annotated chinese corpus. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–8. Association for Computational Linguistics.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *Advances in Artificial Intelligence*, pages 18–32. Springer.

Hai Zhao, Masao Utiyama, Eiichro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for chinese machine translation. *A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817*, pages 248–263.