Improvement of Statistical Machine Translation using Charater-Based Segmentation with Monolingual and Bilingual Information

Vipas Sutantayawalee **Peerachet Porkeaw** **Prachya Boonkwan**
Sitthaa Phaholphinyo **Thepchai Supnithi**

National Electronics and Computer Technology Center, Thailand

{vipas.sutantayawalee, peerachet.porkeaw,prachya.boonkwan,
sitthaa.phaholphinyo,thepchai}@nectec.or.th

Abstract

We present a novel segmentation approach for Phrase-Based Statistical Machine Translation (PB-SMT) to languages where word boundaries are not obviously marked by using both monolingual and bilingual information and demonstrate that (1) unsegmented corpus is able to provide the nearly identical result compares to manually segmented corpus in PB-SMT task when a good heuristic character clustering algorithm is applied on it, (2) the performance of PB-SMT task has significantly increased when bilingual information are used on top of monolingual segmented result. Our technique, instead of focusing on word separation, mainly concentrate on a group of character. First, we group several characters that reside in an unsegmented corpus by employing predetermined constraints and certain heuristics algorithms. Secondly, we enhance the segmented result by incorporating the character group repacking based on alignment confidence. We evaluate the effectiveness of our method on PB-SMT task using English-Thai, English-Lao and English-Burmese language pairs and report the best improvement of 8.1% increase in BLEU score on English-Thai pair.

1 Introduction

Word segmentation is a crucial part of Statistical Machine Translation (SMT) especially for the languages where there are no explicit word boundaries such as Chinese, Japanese, and Thai. The writing systems of these languages allow each word to be written consecutively without spaces between words. The issue of word boundary ambiguities arises if word boundary is misplaced, resulting in an incorrect translation. An effective word segmentator therefore becomes a crucial pre-processing step of SMT. Word segmentators which focusing on word which focusing on word, character [1] or both [2] and [3] have been implemented to accomplish this goal.

Most of word segmentators are supervised; i.e. they require a monolingual corpus of a voluminous size. Various approaches are employed, such as dictionary-based, Hidden Markov model (HMM), support vector machine (SVM), and conditional random field (CRF). Although, a number of segementators offer promising results, certain of them might be unsuitable for SMT task due to the influence of segmentation scheme [4]. Therefore, instead of solely rely on monolingual corpus, the use of a bilingual corpus as an guideline for word segmentation in improving the performance of SMT system has become of increasing interest [4] [5].

In this paper, we propose a novel segmentation approach for Phrase-Based Statistical Machine Translation (PB-SMT) to languages where word boundaries are not obviously marked by using both monolingual and bilingual information on English-Thai, English-Burmese and English-Lao language pairs and demonstrate that (1) unsegmented corpus is able to provide the nearly identical result to manually segmented corpus in PB-SMT task when the good heuristics character clustering algorithm is applied on it, (2) the performance of PB-SMT task has significantly increased when bilingual information are used on top of monolingual segmented result. Our technique, instead of focusing on word separation, mainly concentrate on a group of character. First, we group several characters that reside in an un-

28th Pacific Asia Conference on Language, Information and Computation pages 145–151

segmented monolingual corpus by employing predetermined constraints and certain heuristics algorithms. Secondly, we enhance the segmented result by incorporating the bilingual information which are character cluster alignment, CC co-occurrence frequency and alignment confidence into that result. These two tasks can be performed repeatedly.

The remainder of this paper is organized as follows. Section 2 provides some information related to our work. Section 3 describes the methodology of our approaches. Section 4 present the experiments setting. Section 5 present the experimental results and empirical analysis. Section 6 and 7 gives a conclusion and future work respectively.

2 Related Work

2.1 Thai Character Grouping

In Thai writing system, there are no explicit word boundaries as in English, and a single Thai character does not have specific meanings like Chinese, Japanese and Korean. Thai characters could be consonants, vowels and tone marks and a word can be formed by combining these characters. From our observation, we found that the average length of Thai words on BEST2010 corpus (National Electronics and Computer Technology Center, Thailand 2010) is 3.855. This makes the search space of Thai word segmentation very large.

To alleviate this issue, the notion of Thai character grouping (TCC), is introduced in [1] to reduce the search space with predetermined unambiguous constraints for cluster formation. A group of character may not be meaningful and has to combine with other consecutive group to form a word. Characters in the group cannot be separated according to the Thai orthographic rules. For example, a vowel and tone mark cannot stand alone and a tone marker is always required to be placed next to a previous character only. [6] applied TCC to word segmentation technique which yields an interesting result.

2.2 Bilingual Word Segmentation

Bilingual information has also been shown beneficial for word segmentation. Several methods use this kind of information from bilingual corpora to improve word segmentation. [5] uses an unsegmented bilingual corpus and builds a self-learned dictionary using alignment statistics between English and Chinese language pair. [4] is based on the manually segmented bilingual corpus and then try to “repack” words from existing alignment by using alignment confidence. Both approaches evaluate the performance in term of translation improvement and report the promising results of PB-SMT task.

3 Methodology

This paper aim to compare translation quality based on SMT task between the systems trained on bilingual corpus that contains both segmented source and target, and on the same bilingual corpus with segmented source but unsegmented target. First, we make use of *monolingual information* by employing several character cluster algorithms on unsegmented data. Second, we use *bilingual-guided alignment information* retrieved from alignment extraction process for improving character cluster segmentation. Then, we evaluate our performance based on translation accuracy by using BLEU metric. We want to prove that (1) the result of PB-SMT task using unsegmented corpus (unsupervised) is nearly identical result to manually segmented (supervised) data and (2) when bilingual information are also applied, the performance of PB-SMT is also improved.

3.1 Notation

Given a target $\{Thai\}$ sentence t_1^J consisting of J clusters $\{t_1, \dots, t_j\}$, where $|t_j| \geq 1$. If $|t_j| = 1$, we call t_j as a single character S. Otherwise, we call it as a character group T . In addition, given an English sentence e_1^I consisting of I words $\{e, \dots, e_i\}$, $A_{E \to T}$ denotes a set of English-to-Target language word alignments between e_1^I and t_1^J. In addition, since we concentrated on one-to-many alignments, $A_{E \to T}$, can be rewritten as a set of pairs a_i and $a_i = < e_i, t_j >$ noting a link between one single English *word* and several Thai *characters* that are formed to one character group T

3.2 Monolingual Information

Due to the issue mentioned in section 2.1, we apply character grouping technique (CC) on target text in order to reduce the search space. After performing CC, it will yield several character group T which can be merged together to obtain a larger unit which approaches the notion of word. However, for Thai, we do not only receive T but also S which usually has no meaning by itself. Moreover, Thai, Burmese and Lao writing rule does not allow S to stand alone in most case. Thus, we are required to develop various adapted versions of CC by using a pre-defined word list that can be grouped as a word confirmed by linguists *(orthographic insight))* to automatically pack the characters to become a new T . In addition, all of single consonants in Thai Burmese, and Lao are forced to group with either left or right cluster due to their writing rules. This decision has been made by consulting character co-occurrence statistics (*heuristic algorithm*)

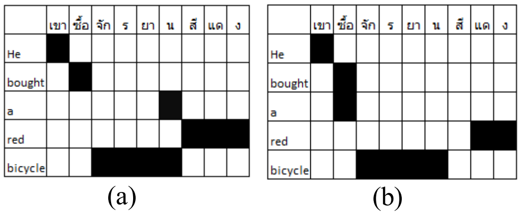
Eventually, we obtain several character group alignments from the system trained on various CC approaches which effect to translation quality as shown in section 5.1

3.3 Bilingually-Guided Alignment Information

We begin with the sequence of small clusters resulting from previous character grouping process. These small T can be merged together in order to form “word” using bilingually-guided alignment information. Generally, small *consecutive* T in target side which are aligned to the same word in source data should be merged together to obtain a larger unit. Therefore, this section describes our one-to-many alignment extraction process.

For one-to-many alignment, we applied processes similar to those in phrase extraction algorithm [7] which is described as follows.

With English sentence e_1^I and a character cluster T, we apply IBM model 1-5 to extract word-to-cluster translation probability of source-to-target $P(t|e)$ and target-to-source $P(e|t)$. Next, the alignment points which have the highest probability are greedily selected from both $P(t|e)$ and $P(e|t)$. Figure 1.a and 1.b show examples of alignment points of source-to-target and target-to-source respectively. After that we selected the intersection of alignment pairs from both side. Then, additional alignment points are added according to the growing heuristic algorithm (grow additional alignment points, [8])



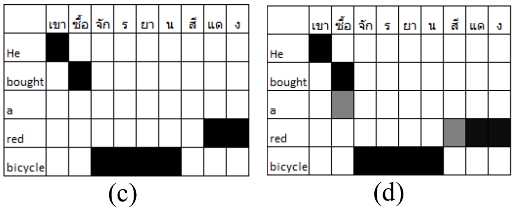


Figure 1. The process of one-to-many alignment extraction (a) Source-to-Target word alignment (b) Target-to-Source word alignment (c) Intersection between (a) and (b). (d) Result of (c) after applying the growing heuristic algorithm.

Finally, we select *consecutive* T which are aligned to the same English word as candidates. From the Figure 1.d, we obtain these candidates (red, สีแดง) and (bicycle, จัก ร ยา น).

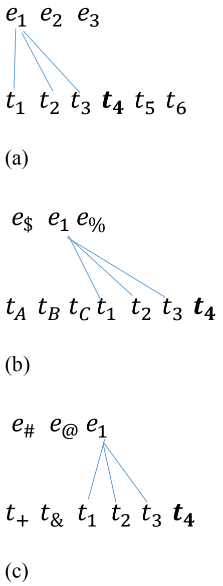
3.4 Character Group Repacking (CCR)

Although the alignment information obtained from the previous step is very helpful for the PB-SMT task. There are certain misaligned alignments that need to be corrected. As shown in Figure 2, one English word e_i is aligned with Thai characters $\{t_1, \dots, t_j\}$ by previous step aligner but actually this word e_i must align with $\{t_1, \dots, t_{j+2}\}$. Word repacking [4] is a one approach that can efficiently resolve this issue. However, in this paper, we slightly modified repacking technique by performing a character group repacking (CCR) instead of word. The main purpose of repacking technique is to group all small consecutive T in target side that frequently align with a single word in source data e_i. Repacking approaches uses two simple calculations which are a co-occurrence frequency ($COOC\ (e_i, t_j)$) and alignment confidence ($AC(\ a_i)$). ($COOC\ (e_i, t_j)$) is the number of times e_i and T_i co-occurrence in the bilingual corpus [4] [9] and $AC(\ a_i)$ is a measure of how often the aligner aligns e_i and t_j when they co-occur. $AC(\ a_i)$ is defined as

$$AC(a_i) = \frac{C(a_i)}{COOC\ (e_i, t_j)}$$

where $C(a_i)$ denotes the number of alignments suggested by the previous-step word aligner.

Unfortunately, due to the limited memory in our experiment machine, we cannot find $COOC\ (e_i, t_j)$) for all possible $< e_i, t_j >$ pairs. We, therefore, slightly modified the above equation by finding $C(a_i)$ first. Secondly, we begin searching $COOC\ (e_i, t_j)$) from all possible alignments in a_i instead of finding all occurrences in corpus. By applying this modification, we eliminate $< e_i, t_j >$ pairs that co-occur together but *never* align to each other by previous-step aligner ($AC(a_i)$ equals to zero) so as to reduce the search space and complexity in our algorithm. Thirdly, we choose a_i with highest $AC(a_i)$ and repack all T in target side to be a new single T unit. This process can be done repeatedly. However, we have run this task less than twice since there are few new groups of character appear after two iterations have passed.


e_1 e_2 e_3
t_1 t_2 t_3 $\boldsymbol{t_4}$ t_5 t_6
$e_\$$ e_1 $e_\%$
t_A t_B t_C t_1 t_2 t_3 $\boldsymbol{t_4}$
$e_\#$ $e_@$ e_1
t_+ $t_\&$ t_1 t_2 t_3 $\boldsymbol{t_4}$

(a)

(b)

(c)

Figure 2. A case that previous aligner misaligned certain clusters (t_4) despite the fact that t_4 are often co-occur with $t_1\ t_2\ and\ t_3$

4 Experimental Setting

4.1 Data

We conduct our experiment based on two bilingual corpora. One is an English-to-Thai corpus (650K corpus) which is constructed from several sources and consists of multiple domains (e.g. news, travel, article, entertainment, computer, etc.). While another one is English-to-Multiple language corpus (20K corpus) which focuses on travel domain only and is developed from several English sentences and those sentences are manually translated to Thai, Burmese and Lao by linguists. Table 1 shows the information on these two corpora. Note that Test set #2 is manually segmented with a guideline different than test#1.

	No. of sentence pairs	
Data Set	English-to-Thai corpus	English-to-Multilanguage
Train	633,589	16,000
Dev	12,568	2,000
Test #1	3,426	2,000
Test #2	500	-

Table 1. No. of sentence pairs in each data set of bilingual corpora

4.2 Tools and Evaluation

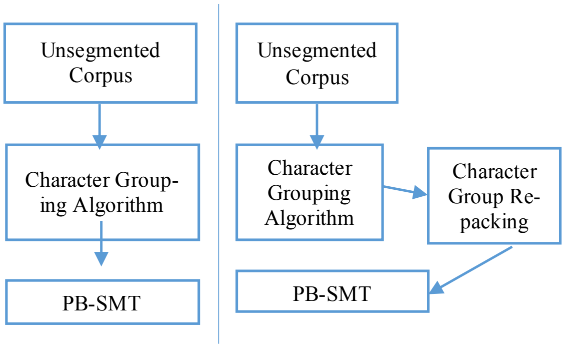
We evaluate our system in terms of translation quality based on phrase-based SMT. Source sentences are sequence of English words while target sentences are sequences of T in Thai, Burmese and Lao. Each T 's length depends on which approach are used in the experiment.

Translation model and language model are train based on the standard phrase-based SMT. Alignments of source (English word) and target (Thai, Burmese and Lao character cluster) are extracted using GIZA++ [8] and the phrase extraction algorithm [7] is applied using Moses SMT package. We apply SRILM [10] to train the 3-gram language model of target side. We use the default parameter settings for decoding.

In testing process, we use dataset that not reside in training data. Then we compared the translation result with the reference in terms of BLEU score instead of F-score because it is cumbersome to construct a reliable gold standard since their annotation schemes are different. Therefore, we resegment the reference data (manually segmented data) and the translation result data based on character grouping techniques. Some may concern about using character group instead of word will lead to over estimation (higher than actual) due to the BLEU score is design based on word and not based on character cluster. However, we used this BLEU score only for comparing translation quality among our experiments. Comparing to other SMT systems still require running BLEU score based on the same segmentation guideline.

5 Results and Discussion

We conducted all experiments on PB-SMT task and reported the performance of PB-SMT system based on the BLEU measure.


Unsegmented Corpus
Character Grouping Algorithm
PB-SMT
Unsegmented Corpus
Character Grouping Algorithm
Character Group Repacking
PB-SMT

(a) (b)

Figure 3. Experiment flows: (a) Monolingual Information (b) Bilingually-Guided Alignment Information

5.1 Monolingual Information

5.1.1 English – Thai language pair

First, we use a method proposed in Figure 3.(a) in order to receive translation results. Table 2 shows the number of Thai character clusters in 650K corpus that are decreasing over time when several different character clustering approaches are applied.

Approaches	No. of Character group (or word in original data)
CC	9,862,271
CC with orthographic insight **(CC-FN)**	8,953,437
CC with orthographic insight and heuristic algorithm **(CC-FN-B)**	6,545,617
Manually segmented corpus **(Threshold)**	5,311,648

Table 2. Number of Thai character group on 650K corpus when different character clustering approaches are applied.

Approaches	650K corpus		20K corpus
	Test #1 Without CCR	**Test #2** Without CCR	**EN-TH**
CC	37.12	36.78	47.63
CC-FN	40.23	38.36	49.21
CC-FN-B	44.69	40.45	49.21
Threshold	47.04	40.73	49.56

Table 3. The performance of SMT trained with different character grouping algorithm.

As seen from Table 3, the BLEU scores of EN-TH pair in all corpora are increasing over time and almost equal to original result on Test#2 in 650K corpus. This is because each CC tends to merge *T* to become larger and larger unit, which approaches the notion of word in eventually. In addition, these experiments also support the claim (1) that unsegmented corpus is able to provide the nearly identical result compares to upper bound in PB-SMT task when a good heuristic character grouping algorithm is applied on it.

However, since CC does not rely on semantic knowledge. Therefore, there are chances that certain *T* do not give a meaningful word resulting in incorrect translation on SMT task.

5.1.2 Preliminary experiment on low resource language (LRL)

We also conduct the experiment on LRL by choosing Lao and Burmese by imitating TCC to be Lao Character Clustering (LCC) and Burmese Character Clustering (BCC) for Lao and Burmese respectively with the same method as in section 5.1.1. However, for Lao and Burmese, we only apply simple CC without any enhanced versions of CC since our knowledge in orthographic of Burmese and Lao are limited.

Approaches	20K corpus	
	English-Lao	**English-Burmese**
CC	39.64	30.11
Upper bound	40.65	26.43

Table 4. The performance of SMT trained with different character clustering algorithm on LRL (Without CCR).

As seen in Table 4, the BLEU scores of CC are almost equal to original results. In English-Burmese pair, however, the character grouping algorithm is able to yield a better performance on upper bound data. We suspect that Burmese word segmentation guideline is still unstable resulting in misplaced word boundaries.

5.2 Bilingually-Guided Alignment Information

As mention earlier in section 3.4, we can improve the translation result by making use of alignment information from previous translation process. Therefore, we perform experiments by using a method describe in Figure 3.(b) in order to receive another translation result set. However, since the corpus size has the direct impact on translation result. We test our hypothesis on the 650K corpus only.

	Test #2		**% of BLEU Improvement**
Approaches	Without CCR	With CCR	
CC	36.78	38.87	5.68
CC-FN	38.36	39.09	1.90
CC-FN-B	40.45	**40.81**	0.89
Threshold	40.73	N/A	N/A

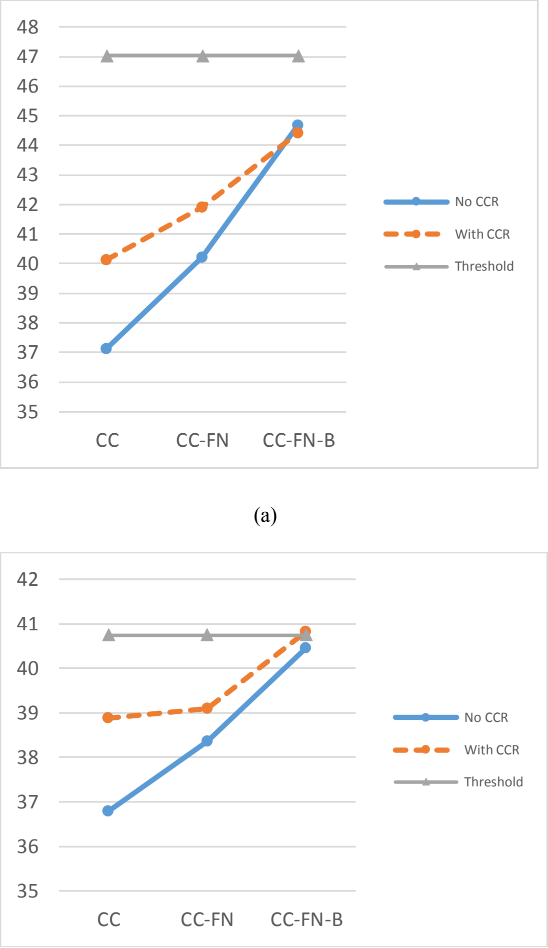
(a.) Test #1 of En-TH 650K corpus

	Test #1		**% of BLEU Improvement.**
Approaches	Without CCR	With CCR	
CC	37.12	40.13	**8.11**
CC-FN	40.23	41.90	4.15
CC-FN-B	44.69	44.43	-0.58
Threshold	47.04	N/A	N/A

(b.) Test #2 of En-TH 650K corpus

Table 5. BLEU score of each character clustering method (a and b) and the percentage of the improvement when we applied CCR to the data

As shown in Table 4 and Figure 4, when CCR have been deployed on each training dataset, the results of BLEU increase in the same manner with *Without CCR* method. It proves the claim (2) that the performance of PB-SMT task has significantly increased when bilingual information are used on top of monolingual segmented result. In addition, there are certain significant points that should be noticed. First, CCR method is able to yield maximum of 8.1 % BLEU score increase. Second, when we apply the CCR methods and reach at some point, few improvement or minor degradation is received as shown in CC-FN-B without and with CCR result**.**



(a)

(b)

Figure 4. The BLEU score of (a) test set no.1 and (b) test set no.2

This is because the number of clusters produced by this character grouping algorithm is almost equal to number of words in threshold as shown in Table 2. However, this approach might suffer from the word boundary misplacement problem. Third, character grouping that use CC with orthographic insight and heuristic algorithm combined with CCR approach (CC-FN-B with CCR) is able to beat the threshold translation result in test set #2 for the first time.

6 Conclusion

In this paper, we introduce a new approach for performing word segmentation task for SMT. Instead of starting at word level, we focus on character group because this approach can perform on unsegmented corpus or manually segmented corpus that have multiple segmentation guideline. To begin, we apply several adapted versions of CC on unsegmented corpus. Next, we use a bilingual corpus to find alignment information for all $< e_i, t_j >$ pairs. Then, we employ character group repacking method in order to form the larger cluster of T.

We evaluate our approach on translation task based on several sources and different domain of corpus and report the result in BLEU metric. Our technique demonstrates that (1) we can achieve a dramatically improvement of BLUE as of 8.1% when we apply CC with CCR and (2) it is possible to overcome the translation result of manually segmented corpus by using CC-FN-B with CCR.

7 Future Work

There are some tasks that can be added into this approaches. Firstly, we can make use of trigram (and n-gram) statistics, maximum entropy or conditional random field on heuristic algorithm in enhanced version of CC. Secondly, we can apply our approaches on Bilinugal corpus which both source and target side are not segmented. Thirdly, we can modify CCR process to be able to re-rank the alignment confidence by using discriminative approach. Lastly, name entity recognition system can be integrated with our approach in order to improve the SMT performance.

Reference

[1] T. Teeramunkong, V. Sornlertlamvanich, T. Tanhermhong and W. Chinnan, “Character cluster based Thai information retrieval,” in *IRAL '00 Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, 2000.

[2] C. Kruengkrai, K. Uchimoto, J. Kazama, K. Torisawa, H. Isahara and C. Jaruskulchai, “A Word and Character-Cluster Hybrid Model for Thai Word Segmentation,” in *Eighth International Symposium on Natural Lanugage Processing*, Bangkok, Thailand, 2009.

[3] Y. Liu, W. Che and T. Liu, “Enhancing Chinese Word Segmentation with Character Clustering,” in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, China, 2013.

[4] Y. Ma and A. Way, “Bilingually motivated domain-adapted word segmentation for statistical machine translation,” in *Proceeding EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 549-557*, Stroudsburg, PA, USA, 2009.

[5] J. Xu, R. Zens and H. Ney, “Do We Need Chinese Word Segmentation for Statistical Machine Translation?,” *ACL SIGHAN Workshop 2004*, pp. 122-129, 2004.

[6] P. Limcharoen, C. Nattee and T. Theeramunkong, “Thai Word Segmentation based-on GLR Parsing Technique and Word N-gram Model,” in *Eighth International Symposium on Natural Lanugage Processing*, Bangkok, Thailand, 2009.

[7] P. Koehn, F. J. Och and D. Marcu, “Statistical phrase-based translation,” in *NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Stroudsburg, PA, USA, 2003.

[8] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics,* vol. 29, no. 1, pp. 19-51, 2003.

[9] I. D. Melamed, “Models of translational equivalence among words,” *Computational Linguistics,* vol. 26, no. 2, pp. 221-249, 2000.

[10] “SRILM -- An extensible language modeling toolkit,” in *Proceeding of the International Conference on Spoken Language Processing*, 2002.