Character-Cluster-Based Segmentation using Monolingual and Bilingual Information for Statistical Machine Translation

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

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 character clustering. First, we cluster each character from the unsegmented monolingual corpus by employing character co-occurrence statistics and orthographic insight. Secondly, we enhance the segmented result by incorporating the bilingual information which are character cluster alignment, co-occurrence frequency and alignment confidence into that result. We evaluate the effectiveness of our method on PB-SMT task using English-Thai language pair and report the best improvement of 8.1% increase in BLEU score. There are two main advantages of our approach. First, our method requires less effort on developing the corpus and can be applied to unsegmented corpus or poor-quality manually segmented corpus. Second, this technique does not only limited to specific language pair but also capable of automatically adjust the character cluster boundaries to be suitable for other language pairs.

1 Introduction

Nowadays, it is admitted that word segmentation is a crucial part of Statistical Machine Translation (SMT) especially in the languages where there are no explicit word boundaries such as Chinese, Japanese or Thai. The writing system of these languages allow each word can be written continuously with no space appearing between words. Consequently, word ambiguities will arise if word boundary has been misplace which finally lead to an incorrect translation. Thus, the effective word segmentator is required to disambiguate each word separator before processing another task in SMT. Several word segmentators which focusing on word, character [1] or both [2] and [3] have been implemented to accomplish this goal.

In order to retrieve a useful information to segment or cluster the word, most of word segmentators are trained on a manually segmented *monolingual* corpus by using various approaches such as dictionary-based, Hidden Markov Model (HMM), support vector machine (SVM) or conditional random field (CRF). Although, a number of segementators are able to yield very 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, various researches make use of either manually segmented [4] or unsegment[1]ed *bilingual* corpus [5] as a guideline information to perform a word segmentation task and improve the performance of SMT system.

Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pages 94–101,
Dublin, Ireland, August 23-29 2014.

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 language pair 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 character clustering. First, we cluster each character from the unsegmented monolingual corpus by employing heuristic algorithm and language insight. Secondly, we enhance the segmented result by incorporating the bilingual information which are character cluster (CC) 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 approach. 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 Clustering

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 cluster (TCC), is introduced [1] to reduce the search space with predetermined unambiguious constraints for cluster formation. A cluster may not be meaningful and has to combine with other consecutive clusters to form a word. Characters in the cluster 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 Bilingually Word Segmentation

Bilingual information has also been shown beneficial for word segmentation. Several methods have used the information from bilingual corpora to perform word segmentation. As in [5], it focuses on unsegmented bilingual corpus and builds a self-learned dictionary using alignment statistics between English and Chinese language pair. On the other hands, [4] is based on the manually segmented bilingual corpus and then try to “repack” the word from existing alignment by using alignment confidence. Both works evaluated the performance in BLEU metric and reported the promising result 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 $\{t_1, \ldots, t_j\}$, where $|t_j| \geq 1$. If $|t_j| = 1$, we call $t_j$as a single character S. Otherwise, we call is as a character cluster T . In addition, given a English sentence e_1^I consisting of I words $\{e, \ldots, e_i\}$, $A_{E\to T}$ denotes a set of English-to-Thai language word alignments between e_1^I and t_1^J. In addition, since we concentrate 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 cluster T

3.2 Monolingual Information

Due to the issue mentioned in section 2.1, we apply character clustering (CC) technique on target text in order to reduce the search space. After performing CC, it will yield several character clusters Twhich can be grouped together to obtain a larger unit which approaches the notion of word. However, for Thai and Lao, 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 *orthographic insight* and *heuristic algorithm* to automatically pack the characters that reside in a pre-defined grammatical word list handcrafted by linguists. Then, all of single consonants in Thai Burmese, and Lao are forced to group with either left or right cluster due to the Thai writing system. The decision has been made by consulting on character co-occurrence statistics (unigram and bigram frequency).

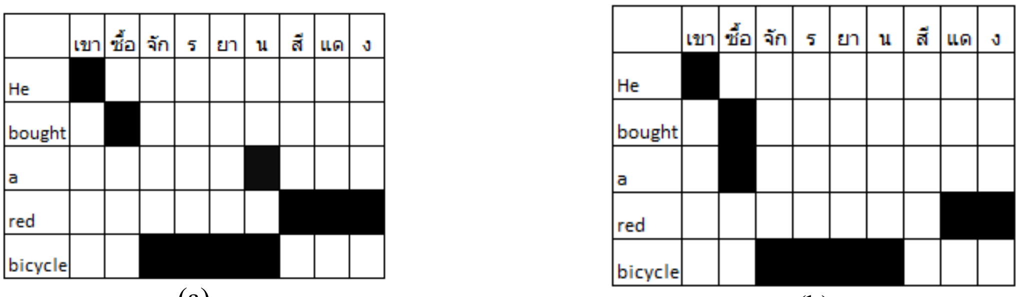
Eventually, we obtain different character cluster alignments from the system trained on various CC approaches which effect to translation quality as shown in section 5.1

3.3 Bilingually-Guided Alignment Information

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

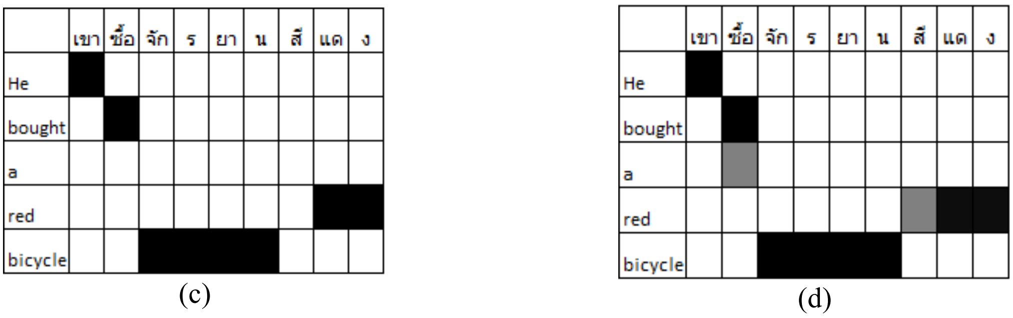
For one-to-many alignment, we applied processes similar to those in phrase extraction algorithm [7] which is described as follows.

With English sentence e_1^I and a Thai character cluster T_i, 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])



(a)

(b)



(c) (d)

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* clusters 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 Cluster Repacking

Although the alignment information obtained from the previous step is very helpful for the PB-SMT task, there is still plenty of room to enhance the PB-SMT performance. One way of doing that is by using word repacking [4]. However, in this paper, we perform a character cluster repacking (CCR) instead of word. The main purpose of repacking technique is to group all small consecutive clusters (or word) in target side that frequently align with one word in source data. Repacking approaches uses two simple calculations which are a co-occurrence frequency ($COOC\ (e_i, T_i)$) and alignment confidence ($AC(\ a_i)$). ($COOC\ (e_i, T_i)$) is the number of times e_i and T_i co-occurr in the bilingual corpus [4] [9] and $AC(\ a_i)$ is a measure of how often the aligner aligns e_i and T_i when they co-occur. AC is defined as

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

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 $COOC\ (e_i, T_i)$) for all possible $< e_i, T_i >$ pairs. We, therefore, slightly modified the above equation by finding $C(a_i)$ first. Secondly, we begin searching $COOC\ (e_i, T_i)$) from all possible alignments in a_i instead of finding all occurrences in corpus. By applying this modification, we eliminate $< e_i, T_i >$ 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 the highest $AC(a_i)$ and repack all character clusters in target side that similar to T_i to be a new single cluster unit. This process can be done repeatedly. However, we have run this task less than twice since there are few new cluster unit appear after two iterations have passed. The running example of this algorithm is described as follows

Suppose previous step aligner (GIZA++) produce two alignments $a_1 = < e_1, T_{1,2} >$ and $a_2 = < e_1, T_{1,2,3} >$ CCR will find the frequency of each aligment and number of times e_i and T_i co-occurr in the bilingual corpus ($COOC\ (e_1, T_{1,2})$ and $COOC\ (e_1, T_{1,2,3})$). Then, we will have $AC(a_i)$ score for each alignment and the aligment with the highest AC will be selected. The CCR will group these cluster (e.g. $T_{1,2}$) to be a new single cluster unit.

4 Experimental Setting

4.1 Data

The bilingual corpus[1] we used in our experiment is constructed from several sources and consists of multiple domains (e.g news, travel, article, entertainment, computer, etc.). We divided this corpus into three sets plus one additional test set as shown below

Data Set	No. of sentence pairs
Train	633,589
Dev	12,568
Test #1	3,426
Test #2	500

Table 1. Information of bilingual corpus

4.2 Tools and Evaluation

We evaluate our system in term of translation quality based on phrase-based SMT. Source sentences are sequence of English words while target sentences are sequences of Thai character clusters and each cluster size depends on which approach 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 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 another two test sets difference to the training data. Then we compared the translation result with the reference in term of BLEU score instead of F-score because of two main reasons. First, it is cumbersome to construct a reliable gold standard since their annotation schemes are different. Second, there is no strong correlation with SMT translation quality in terms of BLEU score [11]. Therefore, we re-segment the reference data (manually segmented) and the translation result data based on TCC. Some may concern about using TCC will lead to over estimation (higher than actual) due to the BLEU score is design based on word and not based on character. However, we used this BLEU score only for comparing translation quality among our experiments. Comparing to other SMT system 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. First, we use a method proposed in section 3.2 followed by the approach in section 3.3 in order to the receive first translation result set (without CCR). Then, we perform a method describe in 3.4 and also follow by approach in section 3.3 in order to receive another translation result set (with CCR). Table 1 shows the number of character clusters that are decreasing over time when several different character clustering approaches are applied.

[1] Currently, the corpus we used is a proprietary of NECTEC and does not available to public yet due to the licensing issue. However, for the educational purpose, this corpus is available upon by request.

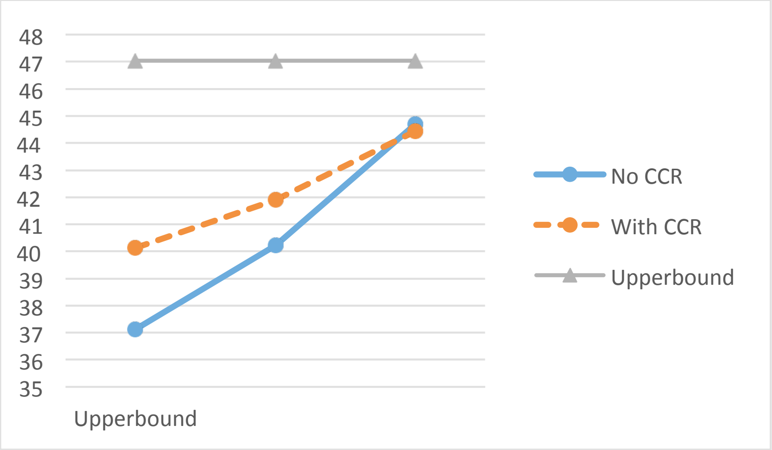
	No. of Character Clusters (or word in original data)	
Approaches	**Without CCR**	**With CCR**
TCC (baseline)	9,862,271	7,187,862
TCC with language insight (TCC-FN)	8,953,437	6,636,305
TCC with language insight and heuristic algorithm (TCC-FN-B)	6,545,617	5,448,437
Manually segmented corpus (Upper bound)	5,311,648	N/A

Table 2. Number of character clusters when different character clustering approaches are applied on the bilingual corpus

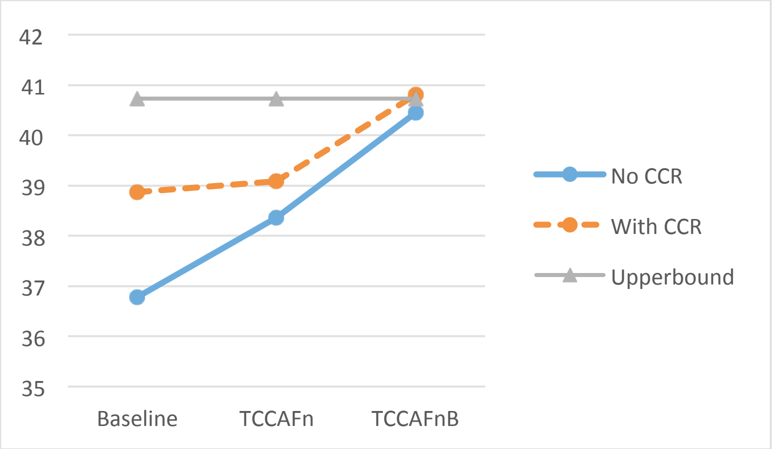
Next, we present all translation results of PB-SMT task that using different character clustering approaches. Each training set is trained with only one character clustering method which are (1) TCC (baseline), (2) TCC with CCR, (3) TCC with only orthographic insight (TCC-FN), (4) TCC-Fn with CCR, (5) TCC with language insight and heuristic algorithm (TCC-FN-B) and (6) TCC-FN-B with CCR. The results are shown in Table 3.

	Test #1 BLEU		**% of BLEU Improvement**	**Test #2 BLEU**		**% of BLEU Improvement**
Approaches	Without CCR	With CCR		Without CCR	With CCR	
Baseline	37.12	40.13	8.11	36.78	38.87	5.68
TCC-FN	40.23	41.90	4.15	38.36	39.09	1.90
TCC-FN-B	44.69	44.43	-0.58	40.45	40.81	0.89
Upper bound	47.04	N/A	N/A	40.73	N/A	N/A

Table 3. BLEU score of each character clustering method
and the percentage of the improvement when we applied CCR to the data



(a)



(b)

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

As seen from Table 3, when we apply the enhanced version of TCCs into the data with no CCR, BLEU score have gradually increased and almost reached the same level as original in test set #2. Furthermore, when CCR have been also deployed on each training dataset, the results of BLEU are also rise in the same manner with Without CCR method. 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 TCC-FN-B without and with CCR result. This is because the number of clusters produced by this character clustering algorithm is almost equal to number of words in original data as shown in Table 2 and this approach might suffer from the word boundary misplacement problem. Third, character clustering that use TCC with orthographic insight and heuristic algorithm combined with CCR approach is able to overcome the translation result from original data for the first time.

6 Conclusion

In this paper, we introduce a new approach for performing word segmentation task for SMT. Instead of starting with word level, we focus on character cluster level because this approach can perform on unsegmented corpus or multiple-guideline manually segmented corpus. First, we apply several adapted versions of TCC on unsegmented data. Next, we use a bilingual corpus to find alignment information for all $< e_i, T_i >$ pairs and then employ character cluster repacking method in order to form the large cluster of Thai characters.

We evaluate our approach on translation task on several sources and different domain 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 adapted TCC with CCR and (2) it is possible to overcome the manually segmented corpus by using TCC with orthographic insight and heuristic algorithm character clustering method combined with CCR. The advantage of our approach is a reduction in time and effot for construct a billinugal corpus because we are no longer required to manually segment all sentences in target side. In addition, our approach is able to cope with larger data information (e.g. 1 million sentences pairs) and adaptable to other language pairs (e.g. English-Chinese, English-Japanese or English-Lao)

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 adapted version of TCC. Secondly, we might report the result from another language pair in order to confirm our approach.Thirdly, we can modify CCR process to be able to rerank 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.

[11] P.-C. Chang, M. Galley and C. D. Manning, “Optimizing Chinese word segmentation for machine translation performance,” in *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, 2008.