A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation

Ye Kyaw Thu † Andrew Finch † Yoshinori Sagisaka‡ Eiichiro Sumita †

Multilingual Translation Laboratory, Universal Communication Research Institute,
National Institute of Information and Communication Technology (NICT), Japan †
Global Information and Telecommunication Institute/Department of
Applied Mathematics Language & Speech Science Research Laboratory (LASS), Waseda University, Japan‡
yekyawthu@nict.go.jp andrew.finch@nict.go.jp ysagisaka@gmail.com eiichiro.sumita@nict.go.jp

Abstract

Myanmar sentences are written as contiguous sequences of syllables with no characters delimiting the words. In statistical machine translation (SMT), word segmentation is a necessary step for languages that do not naturally delimit words. Myanmar is a low-resource language and therefore it is difficult to develop a good word segmentation tool based on machine learning techniques. In this paper, we examine various word segmentation schemes and their effect on the translation from Myanmar to seven other languages. We performed experiments based on character segmentation, syllable segmentation, human lexical/phrasal segmentation, and unsupervised/supervised word segmentation. The results show that the highest quality machine translation was attained with syllable segmentation, and we found this effect to be greatest for translation into subject-object-verb (SOV) structured languages such as Japanese and Korean. Approaches based on machine learning were unable to match this performance for most language pairs, and we believe this was due to the lack of linguistic resources. However, a machine learning approach that extended syllable segmentation produced promising results and we expect this can be developed into a viable method as more data becomes available in the future.

1. Introduction

In Myanmar texts, words composed of single or multiple syllables are usually not separated by white space. Although spaces are used for separating phrases for easier reading, it is not strictly necessary, and these spaces are rarely used in short sentences. There are no clear rules for using spaces in Myanmar language, and thus spaces may (or may not) be inserted between words, phrases, and even between a root words and their affixes.

In SMT, word segmentation is a necessary step in order to yield a set of tokens upon which the alignment and indeed the whole machine learning process can operate. Myanmar language is a resource-poor language; corpora and other language resources such as lexical and grammatical dictionaries are not yet widely available. For this reason, developing a word segmentation tool based on current machine learning techniques from data is a challenging task.

Recently, word segmentation has become an actively researched topic in the SMT research field. Some of the current research is concerned with reconsidering whether or not word segmentation is really necessary for SMT [1], [2], [3]. Some research has proposed alternatives to word-level alignment at finer granularity. In [4] a character-level alignment model is proposed, and in [5] an alignment over morphemes, the smallest meaningful sub-sequences of words is studied. The motivation of this research is to investigate various Myanmar word segmentation schemes and their impact on the quality of SMT when translating into the three prevalent language classes: subject-object-verb (SOV), subject-verb-object (SVO) and verb-subject-object (VSO) languages.

This paper is a study of the following methods for Myanmar word segmentation:

- Character, syllable and word segmentation schemes for Myanmar using rule based syllable segmentation;
- Maximum matching-based word segmentation;
- Bayesian Pitman-Yor language model-based unsupervised word segmentation;
- Pointwise classifier-based supervised word segmentation.

This paper also contributes the first published evaluation of the quality of automatic translations from Myanmar to Japanese, Korean, Hindi, Thai, Chinese and Arabic languages.

The next section describes the related research published in the area of word segmentation in general, and Myanmar word segmentation in particular. Section 3 gives detailed information about all of the

segmentation schemes used in our experiments. Section 4 presents statistical information of the corpus and the translation methods used for the SMT experiments. In Section 5, we make a detailed discussion based on the results. Finally in Section 6, we present our conclusions and indicate promising avenues for future research.

2. Related Work

In this section, we will briefly introduce two proposed word segmentation methods, one syllable segmentation method for Myanmar language and SMT and word segmentation.

Many word segmentation methods have been proposed especially for the Chinese and Japanese languages. These methods can be roughly classified into dictionary-based or rule-based and statistical methods [6], [7], [8], [9], [10]. In dictionary-based methods, only words that are stored in the dictionary can be identified and the performance depends to a large degree upon the coverage of the dictionary. New words appear constantly and thus, increasing size of the dictionary is a not a solution to the out of vocabulary word (OOV) problem. On the other hand, although statistical approaches can identify unknown words by utilizing probabilistic or cost-based scoring mechanisms, they also suffer from some drawbacks. The main issues are: they require large amounts of data; the processing time required; and the difficulty in incorporating linguistic knowledge effectively into the segmentation process [11]. For low-resource languages such as Myanmar, there is no freely available corpus and dictionary based or rule based methods are being used as a temporary solution. Another possible approach is to use a dictionary together with unsupervised or supervised statistical approaches, and we analyse the effectiveness such a technique in the experiments reported in Section 5.

2.1. Myanmar Word Segmentation

As far as the authors are aware there have been only two published methodologies for Myanmar language word segmentation and both of them are rule based techniques that perform syllable segmentation.

Thet et al. (2007) proposed a two step approach in which rule-based syllable segmentation is followed by dictionary-based statistical syllable merging using a dictionary provided by Myanmar NLP team. Six syllable segmentation rules (single character rule, special ending characters rule, second consonant rule, last character rule, next starter rule, miscellaneous rules for numbers, special characters and non Myanmar characters) were applied for syllable segmentation and the approach achieved 100% accuracy [12] on a test set for Myanmar segmentation consisting of 16 documents containing a total of 23,485 words and 32,567 syllables. To determine the word boundaries, dictionary based matching and a statistical approach using bi-grams of syllables were combined and this achieved 98.94% precision, 99.05% recall and 98.99% F-score. The statistical approach was based on the collocation strength of a sentence or phrase with bi-grams (i.e. two syllables) extracted from the corpus. The size of the dictionary using for dictionary-based matching was about 30,000 words.

Htay et al. (2008) proposed a similar 2-step longest matching approach in which the string is first syllable-wise segmented, and then word segmentation is performed based on a left-to-right longest syllable matching technique [13].

In their experiments, a 2-million sentence monolingual Myanmar corpus was used together with an 80K-sentence English-Myanmar parallel corpus. In addition a list of about 1200 stop words, about 4600 syllables and 800K words was used to assist the decision process for annotating word boundaries. Their approach achieved 99.11% precision, 98.81% recall and an F-Score of 98.95% on a 50K-sentence test set.

The two word segmentation approaches described above operate according to the same principles as the “syllable breaking + Maximum Matching” word segmentation approach in our experiments. We also take the same approach of using syllable breaking as the first step in the word segmentation process for Myanmar. The main difference is that we are not using statistical information such as bi-gram probability distributions for making our decisions. Comparing to Htay et al. (2008), we did not utilize a word list extracted from a monolingual Myanmar corpus. We explain our approach for syllable breaking in Section 3.1 and our approach for maximum matching in Section 3.2.

2.2. Myanmar Syllable Segmentation

In this section, we briefly explain proposed rule-based syllable segmentation method [14]. Syllable segmentation rules were created based on the Myanmar syllable structure and its characteristics. Myanmar syllable structure can be represented in BNF (Backus Normal Form or Backus-Naur Form) as follows:

Syllable::=C{M}{V}{F}|C{M}V+A|C{M}{V}CA[F]|E[CA][F]|I|D

Here, C=Consonants, M=Medials, V=Dependent Vowels, S=Sign Virama, A=Sign Asat or Killer,

F=Dependent Various Signs, I=Independent Vowel or Various Signs, E=Independent Vowels,
Symbols and Aforementioned, D=Digits

A grammar was constructed for Myanmar syllables and a finite state acceptor was built from the grammar to parse the Myanmar syllabic structure. Examples of Myanmar syllables and their syllable structure are: (“ကျား”, CMVF), (“ကောင်”, CMVVCA), (“ကျောင်း”, CMVVCAF). Syllable segmentation rules were defined by comparing each pair of Myanmar character categories. To determine a possible syllable boundary (we represent a syllable boundary with an ‘_’ underscore character), rules use a left context of two to four consecutive characters, and some example rules are as follows:

Consonant + Asat Rule: No break after 1st character
(e.g. consonant ka ‘က’ + Asat ‘်’ ⇒ က်)

Independent vowel + Asat Rule: Illegal spelling order, no break after 1st character
(e.g. Independent vowel ‘ဣ’ + Asat ‘်’ ⇒ ဣ်)

Vowel + Consonant Rule: Unclear whether to break or not; move to next character and the decision will become unambiguous
(e.g. Vowel ‘ိ’ + Consonant ‘နှ’ ⇒ နှိ)

Vowel, Consonant + Asat Rule: No break after 1st character
(e.g. Vowel ‘ိ’, Consonant ‘နှ’ + Asat ‘်’ ⇒ိနှ်)

Vowel, Consonant, Medial + Consonant Rule: Break after 1st character
(e.g. Vowel ‘း’, Consonant ‘က’, Medial ‘ြ’ + Consonant ‘သ’ ⇒ း_ကြသ)

However, this set of rules proved insufficient to cover all possible syllables and in later work (Z.M. Maung, Y. Mikami, 2008), the authors extend their approach to three consecutive characters. An accuracy of 99.96% was achieved using a test corpus containing 32K syllables.

2.3. SMT and Word Segmentation

A core issue in SMT is the identification of translation units. In phrase-based SMT these units are comprised of bilingual pairs consisting of sequences of source and target tokens (words). Therefore word segmentation (which defines the nature of these tokens) is one of the key preprocessing steps in SMT. Unfortunately, defining word boundaries for a language is a difficult test even for native speakers of languages without word segmentation such as Myanmar, Thai and Japanese. Several segmentation standards exist for developed languages such as Chinese, and the choice of a Chinese word segmentation scheme used has a large effect on quality of SMT [14]. It is also possible to proceed without any word segmentation at all, by representing the corpus as a sequence of individual graphemes [15], [16], [17]. Some research has found that character-based SMT can achieve translation accuracy comparable to word-based systems [3].

3. Segmentation Methods

This section describes the segmentation methods used in the experiments and is divided into three parts, one for each class of segmentation scheme: dictionary-based, unsupervised and supervised approaches. We first describe our method of syllable breaking that was used as a basis many of the word segmentation methods.

3.1. Syllable Breaking

Syllable breaking is a necessary step for Myanmar word breaking, this is because most Myanmar words are sequences composed of more than one syllables. Generally, there are only 3 rules required to break Myanmar syllables if the input text is encoded in Unicode where dependent vowels and other signs are encoded after the consonant to which they apply. For example, the word ကျောင်း (school) can be decomposed as: က + ျ + ော + င + ် + း. Here, medial consonant ျ (Ya), vowel sign ော (E), vowel sign ာ (Aa) follow consonant က (Ka) and sign် (Asat) and sign း (Visarga) follow syllable final consonant င (Nga). The exception to this combination rule is Kinzi, the conjunct form of U+1004 + Myanmar letter Nga, (e.g. င် + ◌ + ဂ for ဣင်္ဂ in အင်္ဂလိပ် (English) word) that precedes the consonant. Therefore putting a word break in front of consonant, independent vowel, number and symbol characters is the main rule and the first step for syllable breaking. The second rule removes any word breaks that are in front of subscript consonants (e.g. removing the break point symbol “_” in front of တ◌ in _မိ_တ◌_တူ_), Kinzi characters (e.g. အင်္ဂ_လိပ်_), consonant + Asat characters (e.g. က် , နှ် , မ်). The third rule is concerned with break points for special cases such as syllable combinations of loan words (e.g. ဂျော့ချ်), Pali words , phonologic segmentation (e.g. တက်_က_သိုလ်) and orthographic segmentation (e.g. တက္က_သိုလ်). These rules in our experiments with a 27, 747 word dictionary achieved 100% segmentation precision and recall. Figure 1 shows examples of Myanmar syllable breaking using our tech-

nique.

Unsegmented Input:	Segmented Output:
အားရှိတယ်။	⇒ အား_ရှိ_တယ်_။
အိုဂလိပ်	⇒ အိုဂ_လိပ်
မန္တြမာကျောင်း	⇒ မန္တြ_မာ_ကျောင်း
ကုလသမဂ္ဂ	⇒ ကု_လ_သ_မဂ္ဂ

Figure 1. Syllable breaking of Myanmar text

Figure 2 gives the pseudocode for the algorithm we used to implement our proposed syllable segmentation method:

Syllable breaking was the first step in the "syllable breaking + Maximum Matching", "unsupervised segmentation", "syllable breaking, Maximum Matching and Unsupervised" techniques we describe in the following sections.

```
Algorithm: Myanmar Syllable Breaking
Input: array A[1..n]
Output: array B[1..n*2]

j := 1;
char-type = NULL;
for   i=1,...,n   do
   char-type := Check-char-type(A[i]);
   if   char-type = 1   then
      if   (A[i-1] ≠ VIRAMA)   and
           (A[i+1] ≠ ASAT)   and
           (A[i+1] ≠  VIRAMA)   then
              B[j] := '_';
              B[j+1] := A[i];
           else
                B[j] := A[i];
        else if char-type = 2 or char-type = 3 or char-type = 4 then
                     B[j] := '_'
                     B[j+1] := A[i];
    else
                     B[j] := A[j];
j := j +1;
```

Comment:
A[1..n] Is a character array of a Myanmar Sentence
*B[1..n*2] Is a character array of a syllable broken Myanmar sentence*

Comment:
char-type = 1 for consonants, 2 for independent vowels, 3 for number and 4 for symbols
VIRAMA = Unicode character no. U1039 and ASAT = Unicode character no. U103A

Comment:
*we can add one more **else if** for loan words, Pali words, phonologic and orthographic segmentation*

Figure 2. Myanmar Syllable Breaking Algorithm

3.2. Maximum Matching

The Maximum Matching algorithm is a structural segmentation algorithm often used as a baseline method in word segmentation as it typically achieves a respectable level of performance [18], [19]. This algorithm first generates all possible segmentations for a sentence using a dictionary and then selects the one that contains either the longest words or smallest number of words. It is a greedy algorithm and is therefore sub-optimal. The segmentation process may proceed from left-to-right or from right-to-left. In this paper, we used left-to-right Maximum Matching using a [27,747-word] Myanmar word list extracted from a Myanmar-English dictionary [20].

3.3. Unsupervised Segmentation

We used the publicly available latticelm tool [21] to perform unsupervised word segmentation using a Bayesian Pitman-Yor Language model-based strategy in our experiments. Two models were trained. One using syllable sequences, and the other using the output of the syllable segmentation + Maximum Matching method.

3.4. Supervised Segmentation

We used the publicly available KyTea toolkit to perform supervised Myanmar word segmentation based on a pointwise prediction algorithm [22]. A manually segmented corpus of varying size (from 100 sentences to 12,000 sentences) retrieved from the development data set without POS or pronunciation tags was used to train the models. Although there are no standard word segmentation rules for the Myanmar language yet, we defined a simple set of basic segmentation rules for manual segmentation for this experiment. The rules are listed below are applied to the data in the same order they are given here.

RULE 1: Segment Word Units: a word unit is a meaningful unit that could be a candidate for an entry in a lexicon. (e.g. တန်လာနေ့ to တန်လာ_နေ့, ဒီဇင်ဘာလ to ဒီဇင်ဘာ_လ, □□ရက် to □_□_ရက်, ပါသလား to ပါ_သလား, ဒါပေမဲ့ to ဒါပေမဲ့)

RULE 2: Segment Combined words: Combined words are segmented as a single token. (e.g. ဘုရားကျောင်း (church), အကေးငြွယ်ကဒ် (credit card), ဘတ်စ်ကား (Bus car), အော်ဒါမှာ (order), ရပ်ကည့် (stand and watch))

RULE 3: Segment Affixes: insert spaces in-between affixes (prefix or suffix) and root words (e.g. နေခဲ့ to နေ_ခဲ့, အတိရကေငြ်း to အတိရ_ကေငြ်း, လာကိုလို့ to လာကို_လို့, လေ့ကျင့်တဲ့ to လေ့ကျင့်_တဲ့, မန်မန်လုပ် to မန်မန်_လုပ်).

We didn't make any correction of spelling and encoding mistakes for maintain consistency with other segmentation methods used in our experiments. (e.g. ဆောင် (correct word: စောင်), လို့ချင် (correct word: လိုချင်), ဗုဒ္ဓဟူး (correct word: ဗုဒ္ဓဟူး), နရာ (correct word: နေရာ))

4. SMT Experiments

4.1. Corpus statistics

In Section 5, we will present translation evaluation results for a Myanmar to other languages with various segmentation schemes. We used the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions [23]. Developing Myanmar language data for BTEC is currently a work in progress and we used the 72,651 Myanmar sentences for which translation has been completed. Myanmar is used as the source language in all the experiments and the corresponding translated sentences for Japanese (ja), Korean (ko), Hindi (hi), English (en), Thai (th), Chinese (zh) and Arabic (ar) were used to build a set of bilingual corpora. For Hindi, we used both the Devanagari script and a Romanized form (which has a different word segmentation). The corpus statistics for the source language, Myanmar, are summarized in Table1 and for the target languages are in Table 2.

Table 1: Language resources of Myanmar (number of tokens per segmentation method)

Segmentation Methods	Train	Development	Test	Average Syllable per Token
Human Translator	151,829	24,273	2,267	5.45
Character Breaking	2,301,184	339,545	33,449	0.36
Syllable Breaking	835,030	123,961	12,654	1.00
Syllable + Maximum Matching	718,874	103,447	10,206	1.17
Unsupervised (3 gram)	565,304	81,536	8,299	1.48
Unsupervised (4 gram)	577,159	83,855	108,893	1.26
Unsupervised (5 gram)	575,428	84,530	8,564	1.45
Unsupervised (6 gram)	567,322	83,328	8,431	1.47
Unsupervised (7 gram)	573,244	84,965	8,511	1.46
Syl, Max Match, Unsupervised (3 gram)	526,203	75,082	7,495	1.60
Syl, Max Match, Unsupervised (4 gram)	527,216	75,536	7,464	1.59
Syl, Max Match, Unsupervised (5 gram)	526,794	76,010	7,483	1.59
Syl, Max Match, Unsupervised (6 gram)	526,742	75,814	7,568	1.59
Syl, Max Match, Unsupervised (7 gram)	526,803	75,982	7,595	1.59
Semi-Supervised (100 sentences)	527,052	79,955	7,943	1.58
Semi-Supervised (200 sentences)	541,722	81,210	8,041	1.54
Semi-Supervised (300 sentences)	551,389	81,457	8,017	1.52
Semi-Supervised (400 sentences)	546,530	80,352	7,964	1.53
Semi-Supervised (500 sentences)	560,899	82,054	8,114	1.49
Semi-Supervised (600 sentences)	568,567	83,238	8,200	1.47
Semi-Supervised (700 sentences)	554,313	80,406	8,054	1.51
Semi-Supervised (800 sentences)	551,787	80,713	7,992	1.52
Semi-Supervised (900 sentences)	550,423	79,865	7,924	1.52
Semi-Supervised (1000 sentences)	551,327	80,208	7,953	1.52
Semi-Supervised (1100 sentences)	509,566	80,416	7,996	1.62
Semi-Supervised (1200 sentences)	515,162	80,534	8,118	1.61

Table 1 shows the number of tokens and average syllable per token resulting from each of the word segmentation schemes. Here, 3-gram, 4-gram, 5-gram, 6-gram and 7-gram specify the n-gram order of the language model and spelling model used in the unsupervised model.

4.2. Word Segmentation Methods

In the SMT experiments from Myanmar to other languages, we compare the following segmentation methods:

Translation with human translators' segmentation: The BTEC corpus for Myanmar contains some word segmentation added by human translators during translation. These word boundaries were added naturally by the annotators while creating the corpus, and due to the nature of the language are quite sparse.

Translation with character breaking: Each Myanmar character is interpreted as a single word.

Syllable Breaking: Each Myanmar syllable is interpreted as a single word.

Syllable Breaking + Maximum Matching: First syllable breaking was done; then Maximum Matching word segmentation was done.

Unsupervised Word Segmentation: First syllable breaking was done; then the syllable-segmented corpus was segmented using latticelm (with 3-gram to 7-gram language models depending on the experiment).

Syllable Breaking, Maximum Matching and Unsupervised Word Segmentation: First syllable breaking was done. Second, Maximum Matching word segmentation was done on the syllable-segmented corpus. Finally, the syllable-segmented corpus was segmented using latticelm (with 3-gram to 7-gram language models depending on the experiment).

Supervised Word Segmentation: First manual segmentation was done; then the syllable-segmented corpus was segmented with KyTea. Twelve experiments were performed in total using different amounts of manually-segmented data to train KyTea (ranging from 100 to 12,000 sentences).

We calculated the F-score [24] for each segmentation method based on 1,000 manually segmented sentences of Myanmar (see Table 3). We used Edit Distance of the Word Separator (EDWS) and defined the segmentation precision, recall and harmonic mean F as follows:

$$\textbf{\textit{Precision}} = (\text{no. of Sub})/(\text{no. of separators in Hyp})$$
$$\textbf{\textit{Recall}} = (\text{no. of sub})/(\text{no. of separators in Ref})$$
$$\textbf{\textit{F}} = 2*\text{Prec}*\text{Recall}/(\text{Prec}+\text{Recall})$$

Here,
Sub = substitutions,
Hyp = Hypothesis,
Ref = Reference

Clearly, "Character breaking" gives the maximum number of words: 32,748 words, and "Syllable Breaking" gives the second highest number of words: 12,545 words. The syllable breaking followed by the Maximum Matching method gives 10,202 words and "Human Translator" gives the lowest number of words: 1,985 words. "Syllable + Maximum Matching" gives the highest F1 score. The lowest F1 score of 0.23 was given by Human Translator's segmentation. This is because human translators rarely put space between words especially for short sentences. 100% recall can be achieved by Syllable Breaking and Character Breaking. The supervised method was trained on the test-data for this experiment and thus, we do not show the result of F1 measurement on supervised methods in Table 3.

The same Myanmar sentence will be segmented in radically different ways depending on the segmentation method. Figure 3 shows the some different segmentations of a Myanmar sentence sampled from development data. Figure 4 shows an example of word alignment of Myanmar (Syllable Breaking) and English (word breaking) Sentence pair.

4.3. Phrase-based Statistical Machine Translation

The Myanmar source segmented by each of the segmentation methods that we described in Section 4.2 is aligned to the word segmented target languages (Japanese, Hindi (Romanized), Hindi (Devanagari), English, Thai, Chinese and Arabic) using GIZA++ [25]. Language modeling is done using the IRSTLM version 5.80.01 [26]. Minimum error rate training (MERT) was used to tune the decoder's parameters and the decoding is done using the phrase-based SMT system MOSES version 0.91 [27].

4.4. Evaluation Criteria

We used two automatic criteria for the evaluation of the SMT. One is the de facto standard automatic evaluation metric Bilingual Evaluation Understudy

(BLEU) [28] and the other is the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [29]. BLEU score measures the precision of 1-grams to 4-grams with respect to a reference translation with a penalty for short sentences [28]. The BLEU score approximately measures the adequacy of SMT and large BLEU scores are better. RIBES is an automatic evaluation metric based on rank correlation coefficients modified with precision and special care is paid to word order of the translation results. The RIBES is suitable for distant language pairs such as Myanmar and English [29]. Large RIBES scores are better. We calculated the Pearson product-moment correlation coefficient (PMCC) between BLEU and F1, and RIBES and F1 to assess the strength of the linear relationship between segmentation schemes and quality of SMT.

5. Results

5.1. Discussion

We divided the experiments into three groups (rule based, unsupervised and supervised) and divided the target languages into three groups (SOV, SVO and VSO) for making the comparison. We highlighted the table cells of maximum BLEU and RIBES scores for each target language in Tables 4 to 9.

The results show that "Syllable Breaking" segmentation consistently gives the best BLEU and RIBES scores for all language pairs. The reason might be that with syllable segmentation very few errors are made. As mentioned in Sections 2.2 and 3.1, the syllables themselves can be delimited with close to 100% accuracy, and it is in principle possible to group these syllables to form the words in Myanmar without error. Increasing the granularity of the segmentation above this level can introduce errors in which the sequences of syllables do not constitute a word. For example sequences of syllables in erroneous segmented 'words' may contain syllables from more than one true word in the language.

"Syllable + Maximum Matching" segmentation method also consistently give rise to high BLEU and RIBES scores. As we mentioned in Section 3.2 Maximum Matching is using a dictionary for left-to-right word segmentation over segmented syllables. Although this segmentation method can make incorrect decisions during segmentation, we believe that its error rate is low relative to the data-driven methods. This is supported by the precision and recall figures shown in Table 3.

When we analyze the "Unsupervised" segmentation method, the results were quite inconsistent. Generally, the highest BLEU scores are on (4 gram to 6 gram) and the highest RIBES scores range from (3 gram to 7 gram). A possible explanation is that segmentation quality has a high degree of variance, and depends on the training parameters or initial conditions. Nonetheless, the second highest BLEU score (my-ja, 33.45) was achieved by the "Unsupervised (6 gram)" segmentation model and the second highest RIBES score (my-ja, 0.819) is given by "Unsupervised (4 gram) segmentation model. These are encouraging results and we believe that this method may the potential to achieve respectable levels of performance given sufficient data.

When we analyze the SMT quality of the supervised approach, as might be expected we see a strong dependence on the quantity of data used to train the segmentation model. These are come variance in the results but the better results in terms of both BLEU and RIBES scores occurring in the 500-1200 sentence range. The manually segmented corpus was quite small, and although this approach was unable to match the best performing methods, it came reasonably close using all the data, and was still improving at this point. We therefore expect that pursuing supervised segmentation could lead to a viable method of segmentation for low-resource languages if more manually segmented data were available.

The lowest BLEU scores and RIBES scores were given by "Human Translator" segmentation. The reason for this is that most of the translated sentences of the BTEC corpus have no segmentation. The human translators added this information only sparingly. From these results, we conclude that the partial segmentation provided by the human translators is insufficient to provide useful gains for SMT, at least using the methodology we adopted in our experiments.

Visual inspection of the absolute values of the BLEU and RIBES metrics, would seem to indicate that SMT from Myanmar (an SOV language) to SOV languages can lead to higher quality translations than from Myanmar to SVO and VSO language pairs. This is intuitive since the task of re-ordering is considerably simpler in this case.

From the overall results, we can make conclusion that "Syllable Breaking" and "Syllable + Maximum Matching" achieved the higher BLEU and RIBES scores than other segmentation methods. We can expect that we can increase these scores higher than current results in the near future. Both BLEU and RIBES scores proved that the relationship of word segmentation and SMT.

5.2 Error Analysis

Figure 5 shows two examples of translation output to illustrate how errors in segmentation have an effect on

translation quality. Figure 5 (a), is an example of an alignment error occurred in character breaking of the Myanmar sentence "ဂျပန်လူမျိုး စာရေးဆရာ ကို ကောင်းကောင်း သိပါသလား။" (Are you familiar with Japanese authors?) in my-ja SMT. Here, the "က" character from the word "ကောင်းကောင်း" is mistakenly mapped to the Japanese word "は". This kind of error occurred in the "Character Breaking" segmentation because the Myanmar character "က" is usually correctly aligned to the frequently occurring Japanese particle "は" and this character occurs several times in the Myanmar word "ကောင်းကောင်း"; the system preferred to incorrectly translate "က" as the frequent option "は". Figure 5 (b) is an example of a similar alignment error that occurred in "Syllable + Maximum Matching" segmentation of the Myanmar sentence "ဂျင်ဂျာအေးလ် ကို ပေးပါ။" (A ginger ale, please). Here, the Myanmar syllables "ဂျင်" and "ဂျာ" are mistakenly aligned with English words "Jim" and "German". This kind of error occurred in the "Syllable Breaking" segmentation because the word contains other Myanmar word and syllable, namely "ဂျင်" and "ဂျာ" which can be translated as "Jim" and "German" respectively. These words are frequently in the corpus relative to the word for "ginger ale". A segmenter that was capable of identifying the word as a single unit would have avoided this error.

Figure 6 shows the Pearson product-moment correlation coefficient (PMCC) between BLEU scores of "Syllable Breaking", "Syllable + Maximum Matching", "Unsupervised (3 gram to 7 gram)" and "Syl, Max Match, Unsupervised (3 gram to 7 gram)" segmentations and F1 of Myanmar to English SMT. We got similar PMCC graphs for Myanmar to other languages pairs. From the graphs, there appears to be a moderate level of correlation (e.g. 0.739 for my-en, 0.517 for my-ja, 0.555 for my-ko) between the F-score for the segmentation quality and the BLEU score.

Although we can show the relationship between the word segmentation and the SMT results, it is still very hard to make an analysis and formulate measures to describe the deep relationship between them. This is due to the complexity of the SMT process; the quality of SMT depends on many factors relating to alignment, re-ordering and so on.

As we mentioned in Section 4.1, BTEC corpus is also still being created and currently contains many errors such as spelling mistakes, translation errors, and problems with the grammar. The SMT evaluation scores presented in this paper therefore represent a lower bound on what is possible with a larger, cleaner corpus.

6. Conclusion

In this paper, we investigated the effectiveness of seven Myanmar word segmentation schemes for SMT. This paper also contributes the first SMT evaluation from Myanmar to the Japanese, Korean, Hindi, Thai, Chinese and Arabic languages. We built character, syllable and word segmentation schemes for Myanmar by using rule based syllable segmentation, maximum matching based word segmentation, "Bayesian Pitman-Yor" language model based unsupervised word segmentation and "pointwise classifier" based supervised word segmentation. In most of our experiments, the "Syllable Breaking" technique achieved the highest SMT evaluation scores in both BLEU and RIBES, but we believe that as more more data becomes available both the unsupervised and supervised approaches should improve sufficiently to become useful. We propose an elegant new algorithm for Myanmar syllable breaking that is simple to implement, has high coverage, and is very accurate. We believe it will be easy to adapt to related Asian syllabic languages such as Khmer, Laos, and Nepali. We plan to extend our study on syllable breaking using extensions of the unsupervised and supervised segmentation methods presented in this paper in the near future.

References

[1] Jing Sun and Yves Lepage (2012), "Can Word Segmentation be Considered Harmful for Statistical Machine Translation Tasks between Japanese and Chinese?", In *Proceedings of 26th Pacific Asia Conference on Language, Information and Computation*, pp. 351-360.

[2] Jia Xu, Richard Zens and Hermann Ney (2004), "Do We Need Chinese Word Segmentation for Statistical Machine Translation?", In *Proceeding of the Third Sighan Workshop on Chinese Language Learning*, pp. 122-128.

[3] Graham Neubig, Taro Watanabe, Shinsuke Mori, Tatsuya Kawahara (2012), Machine Translation without Words through Substring Alignment, In *Proceeding of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 165-174.

[4] Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, Jiajun Chen (2012), "Enhancing Statistical Machine Translation with Character Alignment", In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 285-290.

[5] Jason Naradowsky and Kristina Toutanova (2011), "Unsupervised Bilingual Morpheme Segmentation and Alignment with Context-rich Hidden Semi-Markov Models", In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 895-904.

[6] Zinmin Wu, Gwyneth Tseng, "Chinese text segmentation for text retrieval: Achievements and problems", *Journal of the*

American Society for Information Science (JASIS), 44(9): pp. 532-542.

[7] Sun, M., D. Shen and B. K. Tsou (1998), "Chinese word segmentation without using lexicon and hand-craft training data. In *Proceeding of COLING-ACL 98*, pp. 1265-1271.

[8] Constantine P. Papageorgiou, "Japanese Word Segmentation By Hidden Markov Model", in *Proceeding of a workshop held at Plainsboro*, New Jersey, March 8-11, 1994, pp. 283-288.

[9] Daichi Mochihashi, Takeshi Yamada, Naonori Ueda, "Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling", in *Proceeding 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 1-Volume 1*. ACL, 2009, pp. 100-108.

[10] Chang Jyun-Shen, C.-D. Chen and Shun-De Chen "Chinese Word Segmentation through constraint satisfaction and statistical optimization", in *Proceeding of ROCLING IV*, ROCLING, pp. 147-165.

[11] Teahan, W. J., Yingying Wen, Rodger McNad and Ian Witten, 2000, "A compression-based algorithm for Chinese word segmentation, In *Computational Linguistics, 26 (3)*, pp. 375-393.

[12] Tun Thura Thet, Jin-Cheon Na, Wunna Ko Ko (2008), "Word Segmentation for the Myanmar language", J*ournal of Information Science 34(5)*, pp. 688-704.

[13] Hla Hla Htay, Kavi Narayana Murthy (2008), Myanmar Word Segmentation Using Syllable Level Longest Matching, t*he 6th Workshop on Asian Language Resources 2008*, pp. 41-48.

[14] Zin Maung Maung, Yoshiki Mikami (2008), A Rule-based Syllable Segmentation of Myanmar Text, *IJCNLP-08 (The Third International Joint Conference on Natural Language Processing), Workshop on NLP for Less Privileged Languages*, 2008, pp. 51-58.

[15] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning (2008), Optimizing Chinese word segmentation for machine translation performance, in *ACL 2008 Third Workshop on Statistical Machine Translation*, pp.224-232.

[16] David Vilar, Jan-Thorsten Peter and Hermann Ney (2007), Can we translate letters, *In Second Workshop on Statistical Machine Translation, Prague, Czech Republic, June 2007*, pp. 33-39

[17] Preslav Nakov, Jorg Tiedemann (2012), "Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages", In *Proceeding of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 301-305.

[18] Yuan Liu, Qiang Tan, and Kun Xu Shen. 1994. The Word Segmentation Methods for Chinese Information Processing (in Chinese). Quing Hua University Press and Guang Xi Science and Technology Press, Page 36.

[19] Pak-kwong Wong and Chorkin Chan (1996), Chinese Word Segmentation based on Maximum Matching and Word Binding Force, In *Proceedings of the 16th conference on Computational Linguistics*, Volume 1, pp. 200-203.

[20] Department of the Myanmar Language Commission (1993), Myanmar-English Dictionary, Yangon, Ministry of Education.

[21] Graham Neubig, Masato Mimura, Shinsuke Mori, Tatsuya Kawahara (2010), "Learning a Language Model from Continuous Speech", In *Proceedings of InterSpeech 2010*, pp. 1053-1056.

[22] Graham Neubig, Yosuke Nakata, Shinsuke Mori (2011), Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, In *Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), short papers Volume2*, pp. 529-533

[23] Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita (2006), Comparative study on corpora for speech translation. In *IEEE Transactions on Audio, Speech and Language, 14(5)*, pp.1674-1682

[24] Chinese Word Segmentation Evaluation Toolkit http://projectile.sv.cmu.edu/research/public/tools/segmentation/eval/index.htm

[25] Franz Och and Hermann Ney (2000), Improved Statistical Alignment Models, In *Proceeding of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 440-447.

[26] Marcello Federico and Mauro Cettolo (2007), Efficient Handling of N-gram Language Models for Statistical Machine Translation, In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 88–95.

[27] MOSES, 2007. A Factored Phrase-based Beam-search Decoder for Machine Translation. URL: http://www.statmt.org/moses/.

[28] 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 of the Association for Computational Linguistics (ACL)*, Philadelphia, USA , pages 311-318.

[29] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada (2010), Automatic Evaluation of Translation Quality for Distant Language Pairs, In *Proceedings of the 2010 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, Oct. 2010, pp. 944-952.

Table 2: Language Resources of Japanese (ja), Korean (ko), Hindi (Romanized), Hindi (Devanagari), English (en), Thai (th), Chinese (zh) and Arabic (ar)

Target Languages	Train		Development		Test	
	Words	Sentences	Words	Sentences	Words	Sentences
ja	594,127	61,651	95,727	10,000	9,266	1,000
ko	559,243	61,651	89,519	10,000	8,777	1,000
hi (R)	545,931	61,651	92,456	10,000	8,123	1,000
hi (D)	465,423	61,651	80,108	10,000	6,920	1,000
en	527,268	61,651	86,934	10,000	7,901	1,000
th	512,054	61,651	86,401	10,000	8,811	1,000
zh	485,151	61,651	77,101	10,000	7,711	1,000
ar	447,799	61,651	72,436	10,000	7,128	1,000

Table 3: Number of words, precision, recall and F-1 scores of segmentation methods calculated on manually segmented 1000 sentences

Segmentation Methods	Words	Precision	Recall	F-1
Human Translator	1,985	99.59%	13.11%	0.23
Character Breaking	32,706	23.60%	100.00%	0.38
Syllable Breaking	12,545	64.82%	100.00%	0.79
Syllable + Maximum Matching	10,202	80.19%	98.60%	0.88
Unsupervised (3 gram)	7,718	75.35%	67.64%	0.71
Unsupervised (4 gram)	8,426	77.59%	76.99%	0.77
Unsupervised (5 gram)	8,846	74.84%	78.46%	0.77
Unsupervised (6 gram)	8,804	75.96%	79.21%	0.78
Unsupervised (7 gram)	9,186	75.67%	82.59%	0.79
Syl, Max Match, Unsupervised (3 gram)	7,224	82.34%	68.48%	0.75
Syl, Max Match, Unsupervised (4 gram)	7,582	83.21%	73.18%	0.78
Syl, Max Match, Unsupervised (5 gram)	7,702	83.72%	74.97%	0.79
Syl, Max Match, Unsupervised (6 gram)	7,712	83.39%	74.79%	0.79
Syl, Max Match, Unsupervised (7 gram)	7,815	83.21%	75.71%	0.79

Human Translator

ဟိုနို့လူလူကနေတိုကျိုအထိထိုင်ခုံကို_ကိုြတာင်စာရင်းပေးသွင်းချင်ပါတယ်။

Character Breaking

ဟ_ို_နှ_ို_့_လ_ူ_လ_ူ_က_န_ေ_တ_ို_က_ျ_ို_အ_ထ_ိ_ထ_ို_င_်_ခ_ု_ံ_က_ို_က_ြ_ိ_တ_ာ_င_်_စ_ာ_ရ_င_်_း_ပ_ေ_း_သ_ွ_င_်_း_ခ_ျ_င_်_ပ_ါ_တ_ယ_်_။

Syllable Breaking

ဟို_နို့_လူ_လူ_က_နေ_တို_ကျိ_အ_ထိ_ထိုင်_ခုံ_ကို_ကိ်ြတင်_စာ_ရင်း_ပေး_သွင်း_ချင်_ပါ_တယ်_။

Syllable + Maximum Matching

ဟို_နို့_လူ_လူ_က_နေ_တို_ကျိ_အထိ_ထိုင်ခုံ_ကို_ကိ်ြတာင်_စာရင်း_ပေး_သွင်း_ချင်_ပါ_တယ်_။

Unsupervised (3-gram)

ဟို_နို့_လူလူ_ကနေ_တိုကျိ_အထိ_ထိုင်ခုံကို_ကိ်ြတင်စာ_ရင်းပေးသွင်းချင်ပါ_တယ်။

Unsupervised (7-gram)
ဟို_နို့_လူ_လူ_ကနေ_တိုကျို့_အထိ_ထိုင်ခုံ_ကို_ကြိုတင်စာ_ရင်း_ပေး_သွင်းချင်_ပါ_တယ်_။

Syllable, Maximum Matching, Unsupervised (4-gram)
ဟို_နို့_လူလူ_ကနေ_တိုကျို့_အထိ_ထိုင်ခုံ_ကို_ကြိုတင်_စာရင်း_ပေးသွင်းချင်ပါ_တယ်။

Syllable, Maximum Matching, Unsupervised (6-gram)
ဟို_နို့_လူလူ_ကနေ_တိုကျို့_အထိ_ထိုင်ခုံ_ကို_ကြိုတင်_စာရင်းပေးသွင်းချင်_ပါ_တယ်။

Supervised (100 sentences)
ဟို_နို့_လူ_လူ_က_နေတို_ကျို့_အထိ_ထိုင်ခုံ_ကို_ကြိုတင်_စာရင်း_ပေးသွင်း_ချင်_ပါတယ်_။

Supervised (1200 sentences)
ဟိုနို့_လူလူ_ကနေ_တိုကျို့_အထိ_ထိုင်ခုံ_ကို_ကြိုတင်_စာရင်းပေးသွင်းချင်_ပါတယ်_။

Figure 3. Different segmentations for a Myanmar sentence

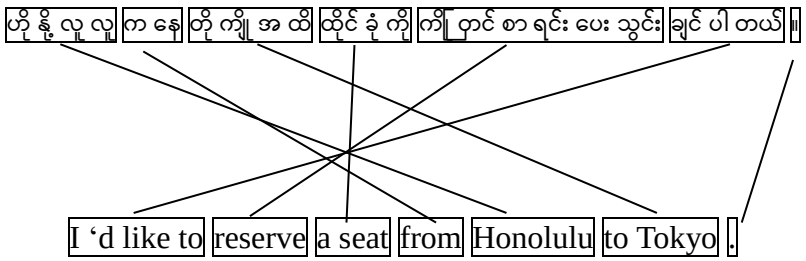

Figure 4. A syllable-to-word aligned Myanmar-English sentence pair

(the above Myanmar sentence is the same sentence as in Figure 3)

Table 4: BLEU scores for Human Translator, Character Breaking, Syllable Breaking and Syllable + Maximum Matching segmentation

	SOV				SVO			VSO
Segmentation Method	**ja**	**ko**	**hi (R)**	**hi (D)**	**en**	**th**	**zh**	**ar**
Human Translator	8.66	8.34	2.12	1.36	3.52	1.88	5.86	1.67
Character Breaking	27.64	25.36	6.68	3.35	6.95	5.57	12.76	8.08
Syllable Breaking	35.17	31.88	8.62	5.40	13.53	11.35	22.21	10.29
Syllable + Maximum Matching	34.58	32.39	8.48	5.60	14.93	12.74	23.09	9.99

Table 5: BLEU scores for Unsupervised (3 to 7 gram), Syl, Max Match, Unsupervised (3 to 7 gram) segmentation

	SOV				SVO			VSO
Segmentation Method	**ja**	**ko**	**hi (R)**	**hi (D)**	**en**	**th**	**zh**	**ar**
Unsupervised (3 gram)	32.96	30.67	7.46	4.81	13.30	12.29	20.91	9.59
Unsupervised (4 gram)	33.27	30.71	7.34	4.84	13.07	12.94	21.52	8.18
Unsupervised (5 gram)	33.17	29.86	7.90	4.64	13.50	12.06	20.84	10.49
Unsupervised (6 gram)	33.45	31.04	7.45	4.56	14.30	12.52	20.91	8.53
Unsupervised (7 gram)	33.29	30.17	7.73	4.87	13.82	12.52	22.18	9.58
Syl, Max Match, Unsupervised (3 gram)	32.44	30.56	7.58	5.58	13.83	12.35	22.10	10.56
Syl, Max Match, Unsupervised (4 gram)	33.16	30.76	7.53	5.21	13.92	12.07	22.23	9.56
Syl, Max Match, Unsupervised (5 gram)	32.96	30.71	7.55	5.42	14.30	12.29	20.96	9.19
Syl, Max Match, Unsupervised (6 gram)	32.34	29.69	8.30	5.11	13.62	11.67	21.74	9.65
Syl, Max Match, Unsupervised (7 gram)	32.78	30.83	7.63	5.39	14.14	11.91	21.47	10.27

Table 6: BLEU scores for supervised segmentation (from 100 to 1200 sentences)

	SOV				SVO			VSO
Segmentation Method	**ja**	**ko**	**hi (R)**	**hi (D)**	**en**	**th**	**zh**	**ar**
Supervised (100 sentences)	29.25	27.54	6.23	3.96	10.66	9.05	17.69	9.01
Supervised (200 sentences)	29.85	27.42	6.52	3.74	11.24	9.93	18.39	8.65
Supervised (300 sentences)	30.73	28.53	6.71	4.26	12.28	10.93	18.57	9.37
Supervised (400 sentences)	30.59	28.36	7.24	3.64	11.52	10.59	18.97	9.43
Supervised (500 sentences)	31.10	28.70	7.09	4.20	12.14	11.05	19.73	9.67
Supervised (600 sentences)	30.90	28.80	7.24	4.24	12.85	10.87	19.29	10.39
Supervised (700 sentences)	30.37	29.08	7.22	4.01	11.83	11.32	20.24	9.90
Supervised (800 sentences)	30.13	28.45	6.93	3.97	12.24	10.61	20.12	9.75
Supervised (900 sentences)	31.21	29.09	7.01	4.56	11.93	11.20	19.81	9.74
Supervised (1000 sentences)	30.87	28.45	6.94	4.61	12.58	11.10	20.59	10.03
Supervised (1100 sentences)	31.72	28.28	7.18	4.32	11.91	10.76	20.79	10.05
Supervised (1200 sentences)	31.07	28.62	7.31	4.57	12.65	11.12	20.43	9.59

Table 7: RIBES scores for Human Translator, Character Breaking, Syllable Breaking and Syllable + Maximum Matching segmentation

	SOV				SVO			VSO
Segmentation Method	**ja**	**ko**	**hi (R)**	**hi (S)**	**en**	**th**	**zh**	**ar**
Human Translator	0.289	0.286	0.192	0.099	0.209	0.142	0.235	0.125
Character Breaking	0.781	0.741	0.570	0.336	0.516	0.381	0.654	0.415
Syllable Breaking	0.837	0.790	0.617	0.385	0.623	0.501	0.741	0.472
Syllable + Maximum Matching	0.822	0.794	0.608	0.408	0.627	0.529	0.745	0.465

Table 8: RIBES scores for Unsupervised (3 to 7 gram), Syl, Max Match, Unsupervised (3 to 7 gram) segmentation

	SOV				SVO			VSO
Segmentation Method	**ja**	**ko**	**hi (R)**	**hi (D)**	**en**	**th**	**zh**	**Ar**
Unsupervised (3 gram)	0.814	0.770	0.582	0.384	0.591	0.533	0.726	0.434
Unsupervised (4 gram)	0.819	0.772	0.578	0.379	0.590	0.521	0.730	0.407
Unsupervised (5 gram)	0.815	0.777	0.584	0.372	0.605	0.529	0.732	0.464
Unsupervised (6 gram)	0.815	0.775	0.592	0.369	0.599	0.537	0.719	0.435
Unsupervised (7 gram)	0.818	0.778	0.585	0.382	0.585	0.524	0.729	0.441
Syl, Max Match, Unsupervised (3 gram)	0.812	0.775	0.585	0.381	0.589	0.527	0.734	0.452
Syl, Max Match, Unsupervised (4 gram)	0.812	0.777	0.586	0.381	0.585	0.526	0.744	0.460
Syl, Max Match, Unsupervised (5 gram)	0.805	0.774	0.584	0.368	0.626	0.529	0.717	0.435
Syl, Max Match, Unsupervised (6 gram)	0.803	0.767	0.582	0.362	0.584	0.513	0.741	0.449
Syl, Max Match, Unsupervised (7 gram)	0.802	0.775	0.597	0.387	0.597	0.523	0.734	0.467

Table 9: RIBES scores for supervised segmentation (from 100 to 1200 sentences)

	SOV				SVO			VSO
Segmentation Method	**ja**	**ko**	**hi (R)**	**hi (D)**	**en**	**th**	**zh**	**ar**
Supervised (100 sentences)	0.774	0.716	0.549	0.314	0.508	0.436	0.674	0.380
Supervised (200 sentences)	0.782	0.721	0.560	0.320	0.526	0.443	0.679	0.418
Supervised (300 sentences)	0.791	0.735	0.555	0.332	0.525	0.460	0.699	0.404
Supervised (400 sentences)	0.792	0.717	0.574	0.322	0.503	0.461	0.689	0.387
Supervised (500 sentences)	0.800	0.748	0.575	0.336	0.548	0.477	0.703	0.406
Supervised (600 sentences)	0.794	0.753	0.579	0.351	0.556	0.491	0.709	0.421
Supervised (700 sentences)	0.792	0.747	0.581	0.352	0.545	0.494	0.710	0.399
Supervised (800 sentences)	0.784	0.732	0.567	0.316	0.539	0.482	0.700	0.411
Supervised (900 sentences)	0.782	0.741	0.576	0.353	0.539	0.487	0.702	0.425
Supervised (1000 sentences)	0.783	0.741	0.569	0.366	0.553	0.478	0.711	0.403
Supervised (1100 sentences)	0.800	0.744	0.565	0.338	0.554	0.481	0.716	0.394
Supervised (1200 sentences)	0.797	0.752	0.563	0.347	0.561	0.474	0.712	0.415

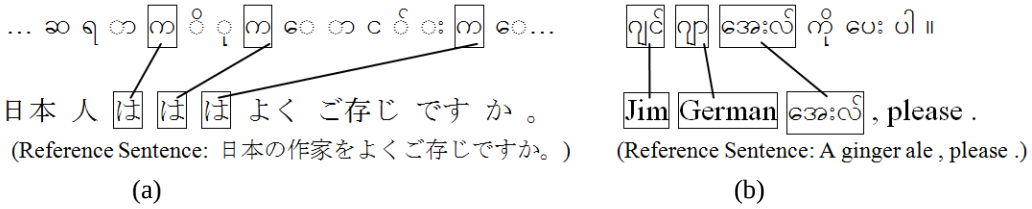

Figure 5. Two examples of translation errors caused by segmentation methods, (a) with Character Breaking for my-ja, (b) with Syllable, Maximum Matching for my-en

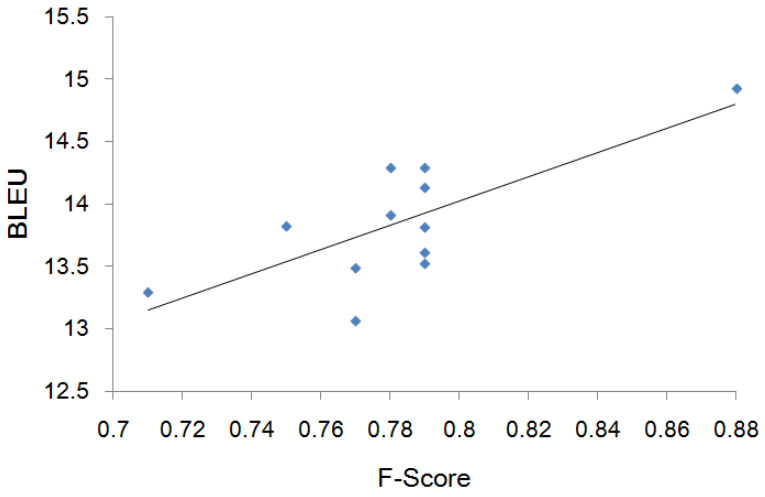

Figure 6. The correlation between BLEU and segmentation F-score for my-en