Word Segmentation for Burmese (Myanmar)

CHENCHEN DING, YE KYAW THU, MASAO UTIYAMA, and EIICHIRO SUMITA,
National Institute of Information and Communications Technology

Experiments on various word segmentation approaches for the Burmese language are conducted and discussed in this note. Specifically, dictionary-based, statistical, and machine learning approaches are tested. Experimental results demonstrate that statistical and machine learning approaches perform significantly better than dictionary-based approaches. We believe that this note, based on an annotated corpus of relatively considerable size (containing approximately a half million words), is the first systematic comparison of word segmentation approaches for Burmese. This work aims to discover the properties and proper approaches to Burmese textual processing and to promote further researches on this understudied language.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Phonology/morphology*;

Additional Key Words and Phrases: Burmese, Myanmar, syllable, word segmentation, algorithm

ACM Reference Format:

Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita. 2016. Word segmentation for Burmese (Myanmar). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 15, 4, Article 22 (May 2016), 10 pages.

DOI: <http://dx.doi.org/10.1145/2846095>

1. INTRODUCTION

Word segmentation to determine the boundaries of *words* for those languages without word separators in orthography is a basic task in natural language processing. For Asian languages, most research on this task has focused on the segmentation and morphological analysis of Chinese, Japanese, and Korean, for which the standard, state-of-the-art technique using conditional random fields has achieved satisfactory performance [Zhao et al. 2010; Kudo et al. 2004; Na 2015]. In this note, we focus on applying word segmentation techniques to an understudied language, Burmese.¹ Previous studies on Burmese word segmentation are scarce and scattered; most adopted a simple *forward maximum matching* approach and used extra rules with cases [Hla Hla Htay and Murthy 2008; Tun Thura Thet et al. 2008; Win Pa Pa and Ni Lar Thein 2008]. An important reason for this is that there are few linguistically annotated resources for Burmese; thus, only dictionary- and rule-based approaches are feasible. An approach integrating dictionaries and unsupervised Bayesian models was reported recently [Ye Kyaw Thu et al. 2014]. However, in that report, the best performance was still achieved by forward maximum matching (Table 1 in Ye Kyaw Thu et al. [2014]), and the training and test data were quite limited in the experiments.

¹The language is referred to as *Burmese* or *Myanmar* in the literature. We use *Burmese* in this note because it is the name more likely to be familiar to English readers.

Authors' addresses: C. Ding, Ye Kyaw Thu, M. Utiyama, and E. Sumita, Advanced Translation Technology Laboratory, ASTREC, National Institute of Information and Communications Technology, 3-5 Hikaridai, Seikacho, Sorakugun, Kyoto 619-0289, Japan; emails: {chenchen.ding, yekyawthu, mutiyama, eiichiro.sumita}@nict.go.jp.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

2016 Copyright is held by the owner/author(s). Publication rights licensed to ACM.

ACM 2375-4699/2016/05-ART22 \$15.00

DOI: <http://dx.doi.org/10.1145/2846095>

The work of this note is based on an in-house annotated Burmese version of *Basic Travel Expression Corpus*, translated from the original English-Japanese version [Kikui 2003] and manually processed by Burmese-native speakers to segment sentences into words. At present, we have 62,000 segmented sentences with approximately 480,000 words. Although the data size is still not comparable to the large-scale corpora of Chinese and Japanese, statistical approaches for Burmese word segmentation are still feasible. In this study, we test and compare classic and state-of-the-art word segmentation techniques, including *crisp forward*, *backward*, and *bidirectional maximum matching* approaches, *N-gram language model* approaches using Viterbi decoding, a *support vector machine* (SVM) approach, and a standard sequence labeling framework using *conditional random fields* (CRF). From the experimental results, we discover that the machine learning approach of SVM and CRF and the Viterbi decoding using a tri-gram language model with modified Kneser-Ney discounting have nearly the same good performance, where the *F-score* on the word boundary insertion is as much as an average of .97 in cross-validation on training data and .98 on a test set. The experimental results indicate that word segmentation of Burmese at least is not a more difficult task than for the other languages, for which well-developed, off-the-shelf techniques can work well. In addition to the numerical results, we further discuss the linguistic features of Burmese and suggest the direction of future work on processing it.

The remainder of the note is organized as follows. In Section 2, we provide an introduction to the Burmese language and its script and also discuss the syllable structure and syllable segmentation. In Section 3, we describe the word segmentation approaches. In Section 4, we report experimental settings and results. In Section 5, we discuss the experimental results. Section 6 draws conclusions and describes future work.

2. OUTLINE OF BURMESE

2.1. Outline of Burmese Language

Morphologically, Burmese is highly analytic with no inflection of morphemes. Similarly to Chinese, morphemes can be combined freely with no changes. In language classification, Burmese and Chinese both belong to the Sino-Tibetan language family, etymologically sharing common basic vocabulary.

Syntactically, Burmese is typically head-final where the functional dependent morphemes succeeding content independent morphemes and the verb constituent working as the root of a sentence always comes at the end of a sentence. Subordinative clauses are also placed before their modifying parts and before the main clause of a sentence. These features are similar to Japanese and Korean, two typical agglutinative, head-final languages in east Asia.²

Orthographically, there is no specific rule or convention on the use of spaces to separate words in Burmese, and spaces are actually used in Burmese texts inconsistently to segment meaning constituents. One intrinsic problem is that the definition of “word” is not intuitive for Burmese; this is also a problematic issue in processing Chinese, Japanese, and Korean. In Chinese, the problem has been handled in a very practical way by combining *segmentation guidelines*, *lexicon*, and *segmented corpus* [Huang and Zhao 2007]. In Japanese, the analysis of sentences actually avoids the definition of

²There is also an exception to the head-final restriction in Burmese that cannot be observed in Japanese and Korean. In Burmese, some modifiers of nouns can be placed after the head noun they modify. Generally, exceptions to an overall syntactic principle are not uncommon in various languages. For example, French generally has a head-initial order in noun phrases to the effect that adjectives are placed after the noun they modify, but adjectives do come before nouns in some cases.

Burmese:	သူ	သည်	စာအုပ်	ကို	ဆရာ	အား	ပေး	သည်
Japanese gloss (translation):	彼	が	本	を	先生	に	あげる	
Korean gloss (translation):	그	가	책	을	선생님	에게	올린다	
English gloss:	he	NOM.	book	ACC.	teacher	DAT.	give	PRES.
English translation:	he gives (will give) a (the) book to a (the) teacher							

Fig. 1. A Burmese sentence containing eight words (morphemes) and a comparison of Japanese and Korean (adapted from Figure 1 of Ding et al. [2014]). In the Burmese, Japanese, and Korean sentences, the black parts are independent content morphemes and the gray parts are dependent functional morphemes. In the English gloss, the corresponding functions of those dependent morphemes are noted by gray subscripts. The Japanese and Korean glosses of the Burmese sentence are grammatically correct translations, from which syntactical similarities among the three languages can be observed. The final morphemes of Japanese and Korean are root verbs, where the conjugation endings marked by gray are not detachable in some analyses, whereas the ending in Burmese is fully detachable, because of its highly analytic characteristic.

“word” and is established practically on the concept of “*morpheme*.” However, there are still several parallel morpheme systems for Japanese, depending on different definitions and sets of parts-of-speech (POS). In Korean, spaces are used to separate words in orthography under a general principle that dependent morphemes are to be attached to independent morphemes, but there are cases for which it is difficult to judge whether a space is to be used even for Korean-native speakers, and the insertion of space diverges among individuals. In this note, the “*Burmese word*” we refer to is comparable to the morpheme of Japanese, that is, the minimum meaning unit judged by native speakers, disregarding the dependence of morphemes. Consequently, inconsistency among different Burmese-native speakers cannot be avoided. Here, we first provide an intuitive example in Figure 1 to illustrate the Burmese word we refer to in this note. Detailed discussion of this issue will be provided in Section 5.

2.2. Outline of Burmese Script

The Burmese script is an *abugida* (*alphasyllabary*) in the Brahmic family, containing 33 basic consonant letters. Each standalone consonant letter can form a complete syllable by itself with the help of an inherent vowel (/a/ or a schwa). The inherent vowel can be changed to other vowels through the use of various diacritic marks. Further diacritic marks include tone marks,³ dependent consonant marks for syllable onset clusters,⁴ and a *virama* (“*asat*”, which means *kill* in Burmese) used to suppress the inherent vowel of a consonant letter for nasal or glottal syllable codas.⁵ Basic consonant letters can be “stacked” in ancient loanwords from Pali and in some modern loanwords from English. We show two examples in Figures. 2 and 3 to illustrate the features of Burmese script.

³There are three tones in modern standard Burmese, creaky / ʔ/, low / ˨/, and high / ˨̊/. Some analyses take the glottal ending /-ʔ/ as the fourth tone. The low tone is unmarked; the creaky and high tones have their own diacritics. Special diacritics may be used for vowel and tone combinations.

⁴The dependent consonant marks insert medial consonants between the onset and nucleus of a syllable, including four types of /-j-/ /-r-/ /-w-/ and /-h-/ (voiceless for sonorant consonants). In modern standard Burmese, /-r-/ has been pronounced as /-j-/ and /-j-/ (including merged /-r-/ becomes a palatalization note of velar consonants). Hence, the possible onset cluster consists of only one consonant with one or two semivowels (the combination of /-jw-/ is possible).

⁵In special cases, *asat*-ed consonant letters also stand for vowels or become silent.

Fig. 2. A Burmese script example. The word means *ninety-six* and consists of three syllables.⁶ The syllables are formed by three basic consonant letters with various diacritics. The two consonant letters marked by “asat” do not form syllables but are subordinate to the preceding basic consonant letters. Diacritics can be added in all of the directions of a basic consonant letter (up and down, left and right).

Fig. 3. Stacked consonants. The word means *wisdom*⁷ and consists of two syllables. For the two stacked consonant letters, the upper is the coda of the first syllable, and the lower is the onset of the second syllable, even though the vowel diacritic for the second syllable is added to the upper consonant letter.

As described and illustrated by the examples, basic consonant letters, around which optional modifications are added, comprise syllables in the writing system of Burmese. Consequently, the syllable boundaries in Burmese texts are orthographically unambiguous in most cases and can be determined using trivial rules, that is, all kinds of diacritics and *asat*-ed consonant letters must not be separated from the neighboring basic consonant letter they modify. A complex case is the stacked consonants, where two syllables are partly merged, and the boundary is positioned between the vertically stacked consonants. However, the contracted syllables do not create confusion in textual processing⁸ because they never cross word (morpheme) boundaries just as general, simple syllables.⁹ For brevity and clarity, the term “*syllable*” of Burmese that we use in this note from now on is used to refer to the *unbreakable* units composed of basic and dependent letters, which are actually single syllables in most cases and also contain contracted multiple-syllable blocks with stacked consonant letters.

3. WORD SEGMENTATION OF BURMESE

Formally, the task of Burmese word segmentation is the process of inserting spaces into textual data without other replacing or rewriting operations. As mentioned in the introduction to Section 2.2, syllables are unbreakable writing units in Burmese, with the result that all approaches tested in this note first apply a syllable segmentation process on input, that is, insert spaces into syllable boundaries, and then decide how the syllables form words, that is, delete spaces between syllables. For a general machine

⁶The three syllables, read as /koʔ/, /sʰɛ/, and /tʃʰauʔ/, stand for *nine*, *ten*, and *six*, respectively. They have the cognates /tɕioʊʔ/ (nine), /sɛ/ (ten), and /liʊʔ/ (six), respectively, in modern standard Chinese (Mandarin).

⁷This Burmese word is a loanword from Pali, read as /weiʔza/. Interestingly, the word ultimately has the English cognate *wise*, also a descendant of the Proto-Indo-European root **weyd-* (to see, to know).

⁸However, they are somewhat problematic in speech processing.

⁹There is a minor exception in that the creaky tone, which is represented by a diacritic, can serve as a contracted genitive case-marker for some nouns. In this case, the extra morpheme is merged in a standalone syllable. This is the only instance in which a syllable can cross morpheme boundaries, and hence we omit it from this note. The special case can be processed trivially, if necessary.

learning approach, the task becomes to be a binary classification problem for syllables based on whether the succeeding space is to be deleted. The task is exactly the same as Chinese word segmentation if the syllables in Burmese are considered as Chinese characters. The approaches tested in experiments are described as follows.

Dictionary-Based Matching

Maximum matching by looking in a prepared dictionary and matching the longest substring in an input sentence is a classic word segmentation approach for Chinese. The matching can be conducted from the beginning of a sentence to its end or in reverse. The former method is referred to as *forward maximum matching* (fmm) and the latter as *backward* (or *reverse*) *maximum matching* (rmm). The two directional processes can be combined to form a *bidirectional maximum matching* (bmm), in which further heuristic rules are used to select the better result from those of fmm and rmm. Specifically, we calculate the probabilities of segmentation using uni-grams to decide which result is better for our experiments.

Although the maximum matching approach is simple, its performance can be mediocre, and it is typically used as a baseline approach in Chinese word segmentation tasks (e.g., Sproat and Emerson [2003]). Because the simplicity and speed of this classic approach is a great advantage, the approach is still widely used in practical engineering and has been studied in recent research (e.g., Sassano [2014]).

Statistical Approaches

Word segmentation using a statistical language model is more reasonable than dictionary-based matching, because it uses the probabilities of words in real textual data. Consequently, segmentation results containing more common words are better than those containing obscure words. Statistical approaches correspondingly require more training data than a dictionary-based approach.

Given a statistical language model, for example, an N-gram language model, the word segmentation task becomes a search problem to determine the segmentation with the highest probability according to the model. A simple Viterbi-like (or Dijkstra-like) dynamic programming algorithm can be applied by scanning an input sentence to generate the best segmentation up to each syllable until the best segmentation of the entire sentence is constructed. We attempted several N-gram language models in experiments, including the maximum-likelihood estimated uni-gram model (uni.mle), the absolute discounting (-0.5) uni-gram model (uni.abs), and the modified Kneser-Ney discounting uni-, bi-, and tri-gram models (uni.mkn, bi.mkn, and tri.mkn, respectively). For bi.mkn and tri.mkn, interpolated probabilities were used when there were high-order N-grams, and otherwise the probability (also interpolated) of the corresponding one-order lower N-gram was used.¹⁰

Machine Learning Approaches

As mentioned, the word segmentation task can be treated as a classification task in a machine learning framework, or, more specifically, a sequence labeling task, due to the properties of textual data, and several standard learning frameworks have been established and developed. We tested two machine learning approaches, CRF in the

¹⁰The N-gram language model has the intrinsic property of assigning lower probabilities to longer sequences, that is, more segmented sequences in our case, as a result of there being more products of probabilities. However, this bias is not so serious in our case, because a less segmented sequence can contain longer words, which usually have lower probabilities than shorter words. Consequently, there is a tradeoff, namely a more segmented sequence with relatively high probabilities for segmented terms versus a less segmented sequence with relatively low probabilities for segmented terms. We did not deal specifically with the issue of sequence length by using extra normalization terms in our experiments.

Table I. Data Sets Used in the Experiments

Set	Sentences	Words	Syllables
<i>training</i>	60,000	462,400	754,853
<i>test</i>	2,000	18,509	30,480

CRF++ toolkit¹¹ [Lafferty et al. 2001; Sha and Pereira 2003] (crf) and SVM in the KyTea toolkit¹² [Neubig et al. 2011] (svm).

Feature engineering for input and tag-set design for output are import issues for machine learning approaches. In experiments, the input features we used are only N-gram of *syllables*,¹³ because our data have quite simple structures, in which only spaces of word boundaries are inserted and no extra information is available. The tag-set of output is also restricted to being binary. We consider that our data is still limited for large-scale machine learning, hence we postpone the work of sophisticated tuning to future work when more data become available.

4. EXPERIMENT

As mentioned, we currently have 62,000 manually segmented sentences. We separate these into two sets for training and test, containing 60,000 and 2,000 sentences, respectively. Details of the data sets are listed in Table I. As for the vocabulary, the entire training set contains 16,379 different types of words, covering 98.5% of the words in the test set.¹⁴ On the syllable level, the entire training set has 2,522 different types of syllables, covering 99.9% of syllables in the test set.

The settings of the matching-based and statistical approaches are as described in the previous section. As to handling unknown words, the matching approaches skip one syllable and moves to the next syllable for further matching. The length of unknown words was also restricted to one syllable and given a penalty in the decoding of statistical approaches. Consequently, unknown words tended to be segmented excessively in matching and decoding. For the SVM and CRF machine learning approaches, we basically used up to 5-gram on syllables as features for input. Specifically, the `-charn` and the `-charw` options were set to five for KyTea.¹⁵ The features used for CRF++ are S_n^{n+k} ($k \in [1, 4], n \in [-k, 0]$) for syllable sequences and S_n ($n \in [-8, 8]$) for single syllables, where it judges whether to insert a word boundary after the syllable S_0 . The performances of compared approaches were measured by the precision, recall, and F-score on *word boundary insertion*; that is, the task was taken purely as a classification problem, even for those rigid matching approaches.

The results of the cross-validation on the training set are listed in Table II, where we split the training set into six subsets of 10,000 sentences; each subset was used for test, and other five subsets for training in turn. The first `syl` column lists the results of inserting a word boundary after each syllable. Because the process inserted all possible

¹¹<http://taku910.github.io/crfpp/>.

¹²<http://www.phontron.com/kytea/>.

¹³Note that the N-grams used in statistical approaches are on *words*, but here for SVM and CRF are on *syllables*.

¹⁴For the dictionary-based matching approaches, the dictionary is composed of the words collected from the corresponding training sets. Hence, the vocabularies in the following experiments using parts of the training set decreased.

¹⁵KyTea takes the N-gram of *characters*, originally designed for Japanese and Chinese, as features. The syllables of Burmese, taken as units in our experiments, are composed of multiple characters (in unicode) in most cases. We used the technique of mapping each Burmese syllable to a Chinese character for the input of KyTea in both training and test. Unknown words in test were mapped to an identical character. As illustrated, there are only thousands of types of Burmese syllables, and hence the Chinese characters are quite sufficient for the mapping.

Table II. Average Precision (*Pre.*), Recall (*Rec.*), and F-Scores in Six-fold Cross-Validation on the Training Set

	syl	*mm			uni.*			*.mkn		svm	crf
		f	r	b	mle	abs	mkn	bi	tri		
<i>Pre.</i>	.612	.945	.941	.961	.963	.963	.963	.956	.954	.970	.971
<i>Rec.</i>	1.00	.896	.893	.910	.938	.939	.941	.977	.980	.970	.972
<i>F-score</i>	.759	.920	.916	.935	.950	.951	.952	.966	.967	.970	.972

Table III. F-Scores on the Test Set with the Increase in Training Data

training	syl	*mm			uni.*			*.mkn		svm	crf
		f	r	b	mle	abs	mkn	bi	tri		
10,000	.756	.926	.925	.936	.939	.940	.940	.946	.945	.947	.943
20,000	.756	.927	.923	.937	.942	.943	.943	.952	.952	.955	.955
30,000	.756	.929	.926	.940	.947	.948	.948	.963	.963	.965	.966
40,000	.756	.931	.927	.943	.951	.952	.953	.971	.972	.973	.974
50,000	.756	.931	.927	.945	.954	.954	.955	.975	.976	.976	.977
60,000	.756	.924	.921	.940	.953	.955	.956	.978	.978	.978	.979
cheating	—	.989	.985	.993	.993	—	—	—	—	—	—

(The 60,000 Row Used the Full Training Set, and the Cheating Row Used the Test Set Itself in Training to Illustrate a Top-Line Performance.)

boundaries, the recall is simply 100%; nearly 40% of the insertions are superfluous in terms of the precision. The forward and backward maximum matching approaches achieved F-scores of approximately .92, which were not low. The bidirectional matching using a uni-gram model obviously performed better than single-direction matching, from which we find that even a simple statistical model does work. The probabilities of the uni-grams we used for bidirectional matching were estimated by maximum-likelihood. We found that further discounting approaches could not improve the performance of bidirectional matching, because the probabilities were used only to select the better of two candidates.

The statistical approaches using uni-gram language models achieved F-scores of approximately .95, which were significantly higher than those of the matching approaches. Compared with the maximum-likelihood estimated uni-gram model, further discounting approaches provided a few gains, indicating that better statistical models can lead to better performance. When the order of the language model increased to bi-grams, the performance improved, but a tri-gram model did not provide further improvement. We also tried a 4-gram language model, but this made no difference compared with the tri-gram model. The results show that a well-estimated bi-gram language model is sufficient for a statistical approach.

The SVM and CRF machine learning approaches provided the best performance in cross-validation. Generally, they achieved nearly the same performance, with the CRF being slightly better. It can be seen that statistical approaches have much higher recalls than SVM and CRF. We attribute this phenomenon to the overall excessive segmentation of unknown words, which were split into syllables in decoding.

Table III shows the F-scores of the compared approaches on the 2,000-sentence test set. The results are generally identical to those of cross-validation. However, the difference between SVM, CRF, and the bi- and tri-gram language models has nearly disappeared with the increase in training data. This phenomenon confirms our hypothesis that the difference of these approaches was due primarily to the unknown words, for which a machine learning approach performs better than a language model-based approach. The last row of Table III is the “cheating” result, for which the words in the test set are collected as a dictionary and used in maximum-likelihood estimation for a uni-gram model. We consider our experimental results to be satisfactory compared with

the cheating results. Hence, we conclude that a general machine learning framework is sufficiently powerful to handle the Burmese word segmentation task; a well-estimated bi- or tri-gram language model also works well, considering the simplicity of N-gram language models.

5. DISCUSSION

As illustrated, the experimental results were good in terms of the numerical measures. However, several problems still need to be discussed concerning our temporarily manually annotated corpus, the basis of the experiments.

Data

As described, the data used in our experiments are based on a basic travel expression corpus, which contains primarily colloquial words, simple expressions, and short sentences. Table I shows that the average length of sentences is approximately only eight words. Although the vocabulary is of considerable size with more than 16,000 words, there is a problem of inconsistency in manual annotation, which will be discussed later. Broadly, the vocabulary in our experiment is unbalanced and limited, and the evaluation is in-domain, that is, restricted to the particular field of the corpus. The compared approaches in this note must be tested further on various text genres.

There is a special stratum in the Burmese vocabulary, consisting of the literary words (e.g., Buddhist and academic words) derived from Pali.¹⁶ One typical phenomenon is that stacked consonants, which are usually used in spelling Pali derived words,¹⁷ are rare in our data.¹⁸ This not only leads to low coverage of vocabulary but also results in many syllable combinations remaining unseen for machines learning approaches. Hence, the trained models based on our current data are still not practical for processing real text, such as news articles. The construction of annotated corpora of Burmese is still in progress and is in need of much more development.

“Word”

As stated in our introduction to the Burmese language, the term “word” does not have a clear, natural definition. In this note and in the experiments, a word of Burmese is treated as a minimum unit of meaning. However, consistency in annotation was not always achieved.

We investigated the vocabulary of the training set. In Chinese word segmentation, there are two typical ambiguities, *combinational* and *overlapping*. Let α , β , γ be strings of characters. A combinational ambiguity arises if α , β , and $\alpha\beta$ are all words at the same time, and an overlapping ambiguity arises when $\alpha\beta$ and $\beta\gamma$ are words at the same time. In the training set vocabulary of 16,379 words, we found 5,086 instances of combinational ambiguities and 1,358 instances of overlapping ambiguities. It is noteworthy that the number of combinational ambiguities is almost nearly one third of the entire vocabulary, largely due to inconsistencies in the manual segmentation. Here, we discuss a typical example.

Figure 4 illustrates a Burmese compound suffix for nouns. The meanings of the entire expression and of the partial morphemes can be compared with their Japanese and Korean glosses. As our segmentation standard, the Burmese suffix must be segmented

¹⁶Consider Latinate words in English.

¹⁷Some frequently used words with stacked consonants: ဘုဒ္ဓဗိမ္ဗူ (/*tɛʔkəθoʊ*/, *university*), သိပ္ပံ (/*θeiʔpā*/, *science*), ဝတ္ထု (/*wʊʔtʰu*/, *story*).

¹⁸In our data, they appear only in two words: ကိစ္စ (/*keiʔsaʌ*/, *business*) and မစ္စ (/*mɪʔ(s)*/, *Ms.*). The first word is borrowed from Pali. The second word is a transcription of English *Miss*, where the stacked consonants with an *asat* are not a standard combination but are used in this case only to represent the final double *s*.

Burmese:	နဲ့	ပတ်သက်	ပြီး
Japanese gloss:	に	関わっ / 関し	て
Korean gloss:	에	관	해서
English gloss:	with	concern / regard	-ing

Fig. 4. Fixed suffix for noun phrases in Burmese, with the meaning of *concerning/regarding (something)*. The Burmese suffix can ultimately be analyzed into three morphemes. Corresponding Japanese, Korean, and English glosses of the morphemes are listed.

into *three words*. Actually, it appears 40 times in the training data, being split into three words 31 times. For the remaining occurrences, the entire suffix is taken as a single word for six times, and the last two morphemes are taken as a single word for three times (i.e., a space is inserted between the first and second morphemes). The result is convincing that this is a very stable compound suffix in Burmese and that a casual annotator might not realize that it can be analyzed even further (or not consider that it requires analysis). Additionally, the connection between the last two morphemes is stronger than that between the first two morphemes, hence there are annotators that separate only the first morpheme as a word but no annotators that separate only the last morphemes as a word. Interesting facts can be observed in the Japanese and Korean translations of the suffix. The Japanese translation can be *に関わって* or *に關して*. In analysis using the IPA system for Japanese morphemes, the previous one is separated into three morphemes, as shown in Figure 4, while the latter is taken as a single morpheme. This might be due to the native speakers' sense that *に關して* has become a combination that is too tight to be separated. In Korean, a space will be inserted between the first and the second corresponding morphemes, because the first and the last morphemes are dependent ones that cannot be separated from their preceding independent morphemes.

From the example, to define “word” for Burmese, as in the case for Japanese and Korean, we must consider the linguistic and practical aspects at the same time. Although it is difficult to provide an ultimate unambiguous definition, a more standardized principle must be established.¹⁹ Broadly, the word segmentation task must be merged into a POS tagging framework, in which a well-defined POS tag set will provide a better explanation of what a word is. This is a long-term project that we intend to conduct gradually in the future.

6. CONCLUSION AND FUTURE WORK

We conducted a systematic comparison of various Burmese word segmentation approaches based on a manually annotated corpus. The experimental results illustrated that state-of-the-art techniques can process Burmese textual data well. We also discussed several problems in the construction of the Burmese corpus. Our preliminary future work is to continue the construction of an annotated corpus²⁰ and to define a standardized POS system for practical natural language processing of Burmese.

¹⁹In practice, multiple-annotator cross-checking may lead to more consistency in annotation and more clarified principles; however, it is more time-consuming and costs more money. Presently, our annotated corpus lacks cross-checking between different annotators' results. The refinement of inconsistent data is also an important issue for future research.

²⁰The data are not publicly available temporarily.

ACKNOWLEDGMENTS

We thank the translators and annotators for the Burmese corpus construction.

REFERENCES

- Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2014. Empirical dependency-based head finalization for statistical Chinese-, English-, and French-to-Myanmar (Burmese) machine translation. In *Proc. of IWSLT*. 184–191.
- Hla Hla Htay and Kavi Narayana Murthy. 2008. Myanmar word segmentation using syllable level longest matching. In *Proc. of IJCNLP*. 41–48.
- Chang-Ning Huang and Hai Zhao. 2007. Chinese word segmentation: A decade review. *J. Chin. Inform. Process.* 21, 3 (2007), 8–19. (in Chinese).
- Gen-ichiro Kikui. 2003. Creating corpora for speech-to-speech translation. In *Proc. of INTERSPEECH*. 381–384.
- Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proc. of EMNLP*. 230–237.
- John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*. 282–289.
- Seung-Hoon Na. 2015. Conditional random fields for Korean morpheme segmentation and POS tagging. *ACM Trans. Asian Low-Res. Lang. Inform. Process.* 14, 3 (2015), 10.
- Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proc. of ACL-HLT*. 529–533.
- Manabu Sassano. 2014. Deterministic word segmentation using maximum matching with fully lexicalized rules. In *Proc. of EACL*. 79–83.
- Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proc. of HLT-NAACL*. 134–141.
- Richard Sproat and Thomas Emerson. 2003. The first international Chinese word segmentation bakeoff. In *Proc. of SIGHAN*, Vol. 1. 133–143.
- Tun Thura Thet, Jin-Cheon Na, and Wunna Ko Ko. 2008. Word segmentation for the Myanmar language. *J. Inform. Sci.* 34, 5 (2008), 688–704.
- Win Pa Pa and Ni Lar Thein. 2008. Myanmar word segmentation using hybrid approach. In *Proc. of ICCA*. 166–170.
- Ye Kyaw Thu, Andrew Finch, Eiichiro Sumita, and Yoshinori Sagisaka. 2014. Integrating dictionaries into an unsupervised model for Myanmar word segmentation. In *Proc. of WSSANLP*. 20–27.
- Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010. A unified character-based tagging framework for Chinese word segmentation. *ACM Trans. Asian Lang. Inform. Process.* 9, 2 (2010), 5.

Received July 2015; accepted November 2015