Word Boundary Identification for Myanmar Text using Conditional Random Fields

Win Pa Pa†, Ye Kyaw Thu‡, Andrew Finch‡, and Eiichiro Sumita‡

†Natural Language Processing Lab,
University of Computer Studies, Yangon, Myanmar
‡Multilingual Translation Lab,
National Institute of Information and Communications Technology, Kyoto, Japan
winpapa@ucsy.edu.mm
{yekyawthu,andrew.finch,eiichiro.sumita}@nict.go.jp

Abstract. This paper examines the effectiveness of conditional random fields (CRFs) when used to identify Myanmar word boundaries within a supervised framework. Existing approaches are based on the method of maximum matching which appears to suffer from problems relating to the manner in which Myanmar words are composed. In our experiments, the CRF approach is compared against a baseline based on maximum matching using dictionaries from the Myanmar Language Commission Dictionary (word only) and a manually segmented subset of the BTEC1 corpus. The experimental results show that the CRF model is able to achieve considerably higher F-scores on the segmentation task than the baseline, even when the baseline is allowed to use words from the test data in its dictionary.

1 Introduction

In the writing systems of many Asian languages, such as Myanmar, Chinese, Japanese and Thai, words are not delimited by spaces. There are no blanks in Myanmar text for word boundaries. Determining the word boundaries, and thus tokenizing the text, is usually one of the first necessary processing steps for Natural Language Processing (NLP) applications. Segmenting Myanmar text is not a trivial task since Myanmar text is composed of words consisting of one or more syllables, and one or more characters can also represent a syllable. Therefore word segmentation is an issue for natural language processing. It may also be necessary to allow multiple correct segmentations of the same text, depending on the requirements of further processing steps. Word segmentation is a necessary prerequisite for higher level language analysis including named entity recognition and syntactic parsing that are used in many NLP applications such as machine translation, automatic speech recognition and information retrieval. Word segmentation is considered to be an important first step for natural language processing tasks.

2 Related Work

The problems of Myanmar word segmentation have been analyzed and different approaches have been developed to achieve different goals.

[1] proposed a hybrid approach that works by longest matching on syllable-segmented sentences. Their probabilistic model used a lexicon of 20,000 words from a Myanmar grammar [9] and achieved 0.755 precision. In their method of longest matching the known words from the dictionary are first segmented and subsequently an n-gram model predicts the segmentation of the unknown words. The principal problem of this approach stems from the ambiguity in the longest matching process, since words can be formed in more than one way.

[5] proposed a word segmentation approach that involved rule-based syllable segmentation and dictionary-based statistical syllable merging using a dictionary of about 30,000 words provided by the Myanmar NLP team of Myanmar Computer Federation. Their approach achieved 100% syllable accuracy and 98.94% precision, 99.05% recall and 98.99% F-score on their word segmentation task.

[6] proposed a 2-step longest matching approach. The first step, was syllable segmentation, in the second step left-to-right syllable longest matching forward segmentation was performed. A 2 million sentence monolingual Myanmar corpus and an 80K sentence English-Myanmar parallel corpus, and lists of stop words, syllables and words were used in the decision process for annotating word boundaries. This approach employed a similar longest matching strategy to [1], and as a consequence also suffers the same problem of ambiguity mentioned earlier.

[3] studied word segmentation in the context of statistical machine translation using 7 different schemes: manually annotated segmentation; character breaking; syllable breaking; syllable breaking + maximum matching; unsupervised word segmentation; syllable breaking + maximum matching + unsupervised word segmentation; and supervised word segmentation. Their study examined the effect of segmentation on the following language pairs: Myanmar to Japanese, Korean, Hindi, Thai, Chinese and Arabic languages. They proposed a new algorithm for Myanmar syllable breaking that achieved 100% accuracy and that can be easily adapted to related Asian syllabic languages such as Khmer, Laos, and Nepali. Their proposed unsupervised segmentation approach did not exceed the performance of the simpler maximum matching approach, and one plausible cause is the lack of data. In this work we focused on a supervised approach which we expected to perform well training on a small amount of human segmented data.

3 Segmentation

This section describes the segmentation methods that were used for the experiments in both the pre-processing stage and the word segmentation stage. The word segmentation was done from both character segmented data and syllable segmented data using CRFs.

3.1 Character Segmentation

The character segmentation pre-processing step trivially segmented the Myanmar sentence into a sequence of graphemes represented by the Unicode characters.

3.2 Syllable Breaking

Syllable breaking is a necessary step for Myanmar word segmentation, since most Myanmar words are composed of multiple syllables and most of the syllables are composed of more than one character. We used the algorithm of [3] for syllable breaking. There are three general rules to break Myanmar syllables from Unicode input text where a consonant is followed by dependent vowels and other symbols. For example, the word ကျောင်း (school) can be decomposed as: က+ျ+ေ+ာ+င+် +း. Here, the medial consonant ျ (Ya), vowel sign က် (E), vowel sign ာ (Aa) follow consonant က (Ka) and sign ် (Asat) and sign း (Visarga) follow syllable final consonant   (Nga). The exception to this combination rule is Kinzi, the conjunct form of U+1004 + Myanmar letter Nga, (e.g. င+် + ္ + ဂ for င်္ဂ in အင်္ဂလိပ် (English)) that precedes the consonant.

The first rule puts a word break in front of consonants, independent vowels, numbers and symbol characters. The second rule removes any word breaks that are in front of subscript consonants, Kinzi characters, and consonant + Asat characters. Break points for special cases such as syllable combinations of loan words (e.g. ဂျော့ချ်), that is the transliteration of "George", Pali words, phonologic segmentation (e.g. တက် က သိုလ်) and orthographic segmentation (e.g. တက္က သိုလ်). In experiments for these rules with a 27,747 word dictionary the approach was able to achieve 100% precision and recall.

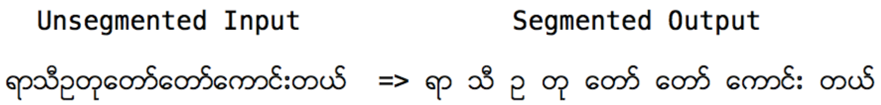

Fig. 1: An example of syllable breaking for a sentence.

3.3 Maximum Matching

Maximum matching is one of the most popular structural segmentation algorithms and it is often used as a baseline method in word segmentation [7]. This method segments using segments chosen from a dictionary. The method strives to segment using the longest possible segments. It is a greedy algorithm and is therefore sub-optimal. The segmentation process may start from either end of the sequences.

3.4 Conditional Random Fields

Linear-chain conditional random fields (CRFs) [4] are models that consider dependencies among the predicted segmentation labels that are inherent in the state transitions of finite state sequence models and can incorporate domain knowledge effectively into segmentation. Unlike heuristic methods, they are principled probabilistic finite state models on which exact inference over sequences can be efficiently performed. The model computes the following probability of a label sequence $\mathbf{Y} = \{y_1, ..., y_T\}$ of a particular character string $\mathbf{W} = \{w_1, ..., w_T\}$.

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} exp(\sum_{t=1}^{T} \sum_{k=1}^{|\boldsymbol{\lambda}|} \lambda_k f_k(y_{t-1}, \mathbf{W}, t)) \tag{1}$$

where $Z(\mathbf{W})$ is a normalization term, f_k is a feature function, and $\boldsymbol{\lambda}$ is a feature weight vector.

We used the CRF++ toolkit[10] to build the CRF models. The feature set used in the models (up to character/syllable tri-grams) was as follows (where t is the index of the character/syllable being labeled):

1. Character/syllable unigrams: $\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$
2. Character/syllable bigrams: $\{(w_{t-1}, w_t), (w_t, w_{t+1})\}$
3. Character/syllable trigrams: $\{(w_{t-2}, w_{t-1}, w_t), (w_{t-1}, w_t, w_{t+1}), (w_t, w_{t+1}, w_{t+2})\}$

These n-grams were combined with label unigrams and bigrams to produce the feature set for the model.

4 Experiments

4.1 Data Setup

The CRF models were trained using a training set selected from a manually segmented 50,000-sentence subset of the Basic Travel Expression (BTEC1) corpus [2]. We ran four maximum matching experiments, drawing from two different dictionaries. The first dictionary was the 26,413-word Myanmar Language Commission (MLC)[8] dictionary; the second dictionary contained the first, and also included all 9,475 of the segments from the manually annotated corpus used to train the CRF models.

The experiments were performed using 10-fold jackknifing of the manually seg-mented BTEC1 50,000 sentences, therefore a test set of 5,000 sentences was used for each fold. A closed test for maximum matching was conducted with the larger dictionary the in order to obtain an approximate upper bound for the method using the available data. There experimental results report the average statistics over all 10 folds together with their standard errors.

4.2 Training with CRFs

The CRF models were trained on two different segmentations of Myanmar, character and syllable. For each character and syllable model, four separate models that used different tag sets were trained. These four tag sets were: {4,5}, {1,4,5}, {1,2,4,5} and {1,2,3,4,5} using the tag number notation in Table 1.

Table 1: List of segmentation tags.

Tag number	**Tag**	**Position**
1	<	The first syllable/character in a word
2	>	The second last syllable/character in a word
3	+	Represents both < and >
4	-	Others
5	|	Final syllable/character in a word

Table 2: The four tag sets used for segmentation.

Number of tags	**Tag set**
2	- |
3	< - |
4	< > - |
5	< > + - |

ရာ	သီ	ဥ	တု	တော်	တော်	ကောင်း	တယ်
<	–	>	|	+	|	|	|
ရာ	သီ	ဥ	တု	တော်	တော်	ကောင်း	တယ်
<	–	>	|	<	|	|	|
ရာ	သီ	ဥ	တု	တော်	တော်	ကောင်း	တယ်
<	–	–	|	<	|	|	|
ရာ	သီ	ဥ	တု	တော်	တော်	ကောင်း	တယ်
–	–	–	|	–	|	|	|

Fig. 2: Syllable tagging with different tags set.

ရ	◌ာ	သ	◌ီ	ဥ	တ	◌ု	တ	◌ေ	◌ာ	◌်	တ	◌ေ	◌ာ	◌်	က	◌ေ	◌ာ	◌်	◌း	တ	ယ	◌်
<	–	–	–	–	>	|	<	–	–	–	–	–	>	|	<	–	–	>	|	<	>	|

Fig. 3: Character tagging with 4 tags.

Examples of segmentation annotated using the four different tag sets are given in Fig. 2.

The meaning of the example sentence in Fig. 2 is: "**The weather is very fine**". It contains 8 syllables, tagged with all four tag sets in decreasing order of tag set size from top to bottom. It can be segmented into 4 segments, actually three words, a noun, an adverb and a verb. The first four syllables becomes a noun that means "**the weather**", the fourth and fifth syllables form an adverb meaning "**very**" and the remaining two syllables form a verb meaning "**fine**". The verb is composed of two segments that are the root word and its suffix.

Fig. 3 gives an example of how the same sentence can be tagged at the character level with the 4 tags {1, 2, 4, 5}.

4.3 Evaluation Criteria

The segmentation performance of maximum matching and CRF models was measured using the commonly used precision (Equation 3), recall (Equation 4), and F-score (Equation 2) defined as follows.

$$\text{F-score} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2}$$

$$Precision = \frac{\#of\ correct\ tokens}{\#of\ tokens\ in\ test\ corpus} \tag{3}$$

$$Recall = \frac{\#of\ correct\ tokens}{\#of\ tokens\ in\ system\ output} \tag{4}$$

4.4 Results and Discussion

Table.3 gives the results on using various sizes of tag set in the CRF model. It is clear from the results that there were almost no differences in the performance of each of the systems. Therefore, for the remainder of the experiments we arbitrarily chose to use the largest tag set.

Table.4 shows the performance of the CRF methods relative to the maximum matching baselines. It can be seen that the CRF models substantially outperform the MM systems in terms of the overall F-score, but the MM (MLC) method has a very high level of precision. The MM (BTEC1 Open) experiment used the same training and test data as the CRF model, and shows the in-domain performance using a small dictionary (approximately 8,500 entries). The MM (BTEC1 Closed) experiment used a dictionary from the entire 50,000-word training set that included the test data.

Table 3: Word segmentation performance (with standard errors) using different tag sets with CRF models.

Tagging Method	**Character**			**Syllable**		
	Precision	Recall	F-Score	Precision	Recall	F-Score
2 Tags	0.9695	0.9679	0.9687	0.9698	0.9683	0.9690
	±0.0040	±0.0056	±0.0046	±0.0035	±0.0048	±0.0040
3 Tags	0.9693	0.9686	0.9689	0.9703	0.9681	0.9692
	±0.0038	±0.0055	±0.0044	±0.0034	±0.0048	±0.0039
4 Tags	0.9694	0.9692	0.9693	0.9702	0.9676	0.9689
	±0.0038	±0.0053	±0.0043	±0.0034	±0.0048	±0.0040
5 Tags	0.9693	0.9692	0.9692	0.9703	0.9672	0.9687
	±0.0038	±0.0053	±0.0043	±0.0034	±0.0048	±0.0039

In order to study how the CRF models behave with varying amounts of training data, we run a sequence of experiments that trained CRF models from 10K, 20K, 30K, 40K and 50K sentences respectively. From the results in Fig. 4,

Table 4: Word segmentation performance (with standard errors) of the MM and CRF methods.

Method	Precision	Recall	F-Score
MM (MLC)	0.9881	0.7232	0.8351
	±0.0006	±0.0031	±0.0022
MM (MLC+BTEC1)	0.9093	0.9367	0.9228
	±0.0037	±0.0040	±0.0032
MM (BTEC1 Closed)	0.9106	0.7872	0.8444
	±0.0034	±0.0029	±0.0030
MM (BTEC1 Open)	0.9363	0.96243	0.7490
	±0.0074	±0.0013	±0.0020
CRF Character (5 tags)	0.9693	0.9692	0.9692
	±0.0038	±0.0053	±0.0043
CRF Syllable (5 tags)	0.9703	0.9672	0.9687
	±0.0034	±0.0048	±0.0039

it is clear that the CRF model performs almost identically on character segmented and on syllable segmented data. Furthermore, the results show that the performance of the system is strongly linked to the data set size.

5 Error Analysis

Table.5 shows the most frequent labeling errors made by the CRF model when labeling syllable-segmented input. A list of syllables is shown for each error which represents the all of the syllables that gave rise to the error, listed in order of frequency.

It can be seen from the table that most of the errors in the top portion of the table are caused by syllables which end in one of the following vowels: ◌ာ, ◌ါ , ◌် , ◌ီ , ◌း . Related errors occurred at the character level where the top 2 errors were: ◌် , ◌း. Interestingly a large number of errors occurred on the following syllables: ပါ, မှာ, တာ, ထား, which often signify the ends of words, but may also occur quite frequently within words especially compound words. Single character consonants: မ, ရ, အ, စ, were also responsible for many errors. These mostly occur at the beginning of words labeled with '<', but can also occur at the beginning of words with the '+' label, and these cases appear to be difficult to disambiguate.

6 Conclusion

In this paper we have studied the application of CRF models to perform supervised word identification of Myanmar text. The performance of the CRF models was compared against a baseline model based on maximum matching that is

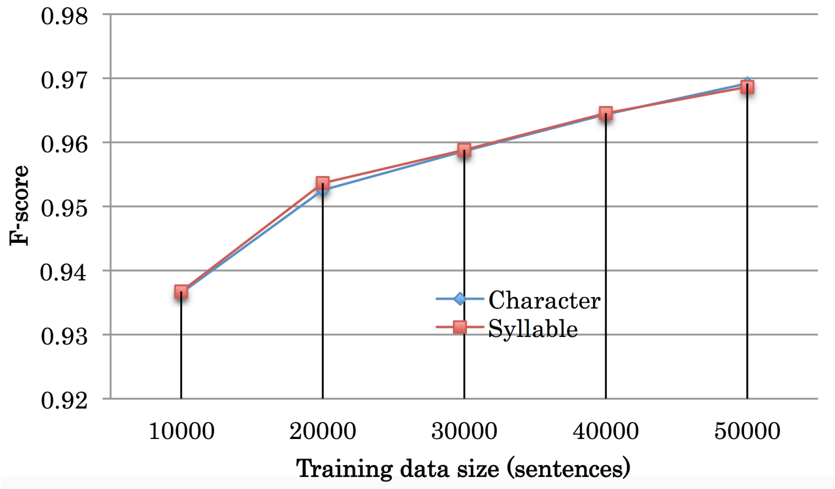

Fig. 4: F-Scores from training with CRF models on varying data set sizes.

Table 5: Statistics on the most frequent 300 labeling errors from 10 experiments for syllable tagging together with all the associated syllables.

Reference	Output	Error Syllable	Percentage
|	+	ပေး,ပါ,တာ,သွား,ရ,ထား,ဘယ်,နေ,ပြန်,တစ်,တွေ,ရှိ,လိုက်,နား,တ,ဒီ,က	19.59
+	|	ပေး,သွား,ရှိ,တာ,ခေါ,တစ်,နှစ်,တွေ,လာ,မှာ,လုပ်,ပြန်,ဆယ်,ပါ,ထား,တို,နေ	16.78
<	|	သွား,လို,ရှိ,ရ,ပေး,ဖြစ်,ထင်,နှစ်,လုပ်,တစ်,ထား	11.44
>	+	ပါ,ရ,မ,နှာ,အ,နေ,စ,မှာ	9.49
|	<	ရ,ပါ,ချင်,နေ,လို့,ဖြစ်,ပေး	8.54
+	>	မ,ပါ,သ,စ,အ	8.29
<	+	အ,မ,ဘယ်,ဆောင်,နေ,ဒီ,ကြ	7.96
-	<	ရ,ပါ,နိုင်,ပေး	6.18
+	<	အ,ဘယ်,မ,နေ,နည်း	5.66
<	-	ပါ,နိုင်,ရ	3.38
>	|	လောက်,ရွက်,ရာ,တာ,တွေ,နား	1.62
|	>	လောက်,ပါ,နည်း	0.95

close to the current state-of-the-art in Myanmar word segmentation. Our results show that the overall performance of the CRF models, measured in terms of F-score was substantially higher than the maximum matching baseline. We were also able to show that the CRF model was able to perform word segmentation equally well from either Myanmar characters or syllables. Experiments on data set size revealed that the CRF is still improving even on the largest training data set size of 50,000 sentences, and therefore we believe that the acquisition of more data is critically important in improving the segmentation accuracy of the system. In future work, we intend to increase the size of the manually segmented corpus since our experiments indicated that this was likely to deliver significant improvement in performance.

7 Acknowledgements

We thank Ms. Aye Mya Hlaing, Ms. Aye Myat Mon, Ms. Hay Mar Soe Naing, Ms. Min Min Oo, Ms. Nyo Lay Myint, Ms. Sann Su Su Yee and Ms. Thae Nu Htwe from the University of Computer Studies, Yangon (UCSY), Myanmar for their help in segmenting 50,000 sentences of BTEC1 Myanmar Corpus manually. We also thanks Dr. Chenchen Ding (Multilingual Translation Lab., NICT) for his python script tagger (character level).

References

1. Win Pa Pa, Ni Lar Thein, Myanmar Word Segmentation using Hybrid Approach, Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, Myanmar, pp-166-170
2. G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita. 2006. Comparative study on corpora for speech translation. IEEE Transactions on Audio, Speech, and Language Processing, 14(5):1674–1682.
3. Ye Kyaw Thu, Andrew Finch, Yoshinori Sagisaka, Eiichiro Sumita, A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation, Proceedings of 12th International Conference on Computer Applications, Yangon, Myanmar, 2014, pp-167-179
4. J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conf. on Machine Learning, pp-282–289
5. Tun Thura Thet, Jin-Cheon Na, Wunna Ko Ko (2008), "Word Segmentation for the Myanmar language", Journal of Information Science 34(5): pp-688-704
6. Hla Hla Htay, Kavi Naruyana Murthy (2008), Myanmar Word Segmentation Using Syllable Level Longst Matching, the 6th Workshop on Asian Language Resources 2008, pp-41-48
7. 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, pp 36
8. Myanmar English Dictionary, Myanmar Language Commission, Myanmar, 2012 Edition
9. Myanmar Grammar, Myanmar Language Commission, Myanmar, 2000
10. Taku Kudo: CRF++ An open source toolkit for CRF (2005) http://crfpp.sourceforge.net/