Khmer Word Segmentation Using Conditional Random Fields

Vichet Chea[*†], Ye Kyaw Thu[†], Chenchen Ding[†], Masao Utiyama[†], Andrew Finch[†], Eiichiro Sumita[†]
[*] Research and Development Center, NIPTICT, Phnom Penh, Cambodia
E-mail: vichet.chea@niptict.edu.kh
[†] Advanced Speech Translation Research and Development Promotion Center, NICT, Kyoto, Japan
E-mail: {yekyawthu, chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract—**Word Segmentation is a critical task that is the foundation of much natural language processing research. This paper is a study of Khmer word segmentation using an approach based on conditional random fields (CRFs). A large manually-segmented corpus was developed to train the segmenter, and we provide details of a set of word segmentation strategies that were used by the human annotators during the manual annotation. The trained CRF segmenter was compared empirically to a baseline approach based on maximum matching that used a dictionary extracted from the manually segmented corpus. The CRF segmenter outperformed the baseline in terms of precision, recall and f-score by a wide margin. The segmenter was also evaluated as a pre-processing step in a statistical machine translation system. It gave rise to substantial increases in BLEU score of up to 7.7 points, relative to a maximum matching baseline.**

I. Introduction

In the writing system of the Khmer language, spaces are not used to separate words, but spaces are used occasionally for easier reading. There are no standard rules for using spaces in the Khmer writing system. These large contiguous blocks of unsegmented words can cause major problems for natural language processing applications such as machine translation, speech synthesis, information extraction, and therefore word segmentation techniques need to be developed. Although Khmer native speakers are easily able to determine the the positions of word boundaries, developing an automatic word segmentation is not a trivial task.

Word boundary ambiguities have two main causes. The first one is concerned with the lexical semantics. A single sentence can be segmented in several ways based on its meaning in context. For example:

- ខ្ញុំ ចង់ឱ្យ ⟨ អ្នកស្តាប់ ⟩ យល់ ពី បញ្ហា នេះ
- I want listener to understand this problem

- ខ្ញុំ ចង់ឱ្យ ⟨ អ្នក ⟩ ⟨ ស្តាប់ ⟩ យល់ ពី បញ្ហា នេះ
- I want you to listen in order to understand this problem

The second cause is unknown words. Unknown words are words that are not found in dictionaries or training data and are often named entities such as personal names and locations.

An overview of the contents of the paper is a follows. The CRF++ toolkit [1] was used to build a CRF [2] models to learn the word formation patterns of Khmer words. Training data for the CRF and maximum matching baseline model was manually segmented based on four sets of segmentation patterns based on word types: a set of patterns for a single words (Section III-A); patterns for compound words (Section III-B); patterns for prefix words (Section III-C); and patterns for suffix words (Section III-D). Section IV describes the maximum matching baseline approach, and Section V presents the details of the CRF method we used. Sections VI and VII present experiments on segmentation quality and statistical machine translation respectively. Finally, Section VIII concludes.

II. Related Work

The first published word segmentation approach to Khmer word segmentation was presented in [3]. Their method was developed for use with a speech recognition system, and was based on the method of longest matching which is in essence a sub-optimal greedy version of the maximum matching method used in this paper as a baseline. [4] proposed the word segmentation of Khmer written text based on a combination of dictionary matching and a bi-gram model. First, the input sentence was segmented into compound orthographic symbols called Khmer Character Clusters (KCC) and converted into Khmer Common Expressions (KCE) [5]. Then, dictionary look up is done using KCE to produce the possible word segmentation hypotheses for the original input text. Finally a bigram model is applied to resolve segmentation ambiguities. Bigram models over KCC units and words were studied. Word bigrams proved to be the most effective, achieving 91.56% precision 92.14% recall and 91.85% F-score.

In [6] a rule-based approach was obtained by statistical analysis to tackle the issue of out-of-vocabulary words, and to detect compound words, proper names, acronyms, derivatives words and new words for Khmer word segmentation. Their experimental study showed their proposed approach to be superior to a baseline based on [4]. Specifically, the reported precision was 77.7% (5.9% higher), the recall was 75.5% (3.3% lower), and the f-score was 76.5% (1.4% higher).

[7] proposed a Khmer word segmentation method that used a Bi-Directional Maximal Matching (BiMM) approach. The study also focused on how to implement Khmer word segmentation on both Khmer plain text and Khmer in Microsoft word documents. Their approach achieved a segmentation accuracy of 98.13%.

III. Khmer Word Segmentation

As in most other languages, the Khmer vocabulary consists of single words and compound words. A single word is a word that is not composed from other words, while a compound word is composed of two or more single words, prefixes or/and suffixes. Word concatenation has often been used in language as a means of creating a new word that draws on the semantics of its components.

Four classes of segment (word) types were observed during the manual segmentation of the corpus of Khmer text, each representing a different type of word, these were:

- Word Type 1: Single Words
- Word Type 2: Compound Words
- Word Type 3: Compound Words with Prefix
- Word Type 4: Compound Words with Suffix

The word segmentation was annotated using the tagset used to train the CRF models shown in Table VI with the exception of the ’0’ tag which remained implicit.

A. Word Type 1: Single Words

In Khmer language, a *borivasab* (បរិវាស័ព្ទ) [8] is subordinate word that is combined with a main word to form a single word. Words formed using borivasab are classed as pseudocompound words [9] because the subordinate word has no meaning in isolation.

[10] describes borivasab as entourage words that are used to facilitate verbal speaking and sometimes these are rhythm words used to make the spoken language comfortable for listener. Khmer people have a habit of speaking a phrase or word by interchanging syllables. An example of the formation of a word using a borivasab is given in Figure 1: the word is សម្លសម្លុក pronounced as *Samlorsamlok*. The root of this word comes from the phrase សម្លបង់ប្រហុក (pronounced *Samlor Bang Prohok*). It means that soup cooked with *Prohok*. *Prohok* is the name of Khmer cheese. In this case, *Samlor Bang Prohok* the final syllables of the first and last words are interchanged (① in Figure 1) *Samlok Bang Prohor*; forming the borivasab *Samlok* in the first word. Then, the pseudocompound word *Samlorsamlok* is formed by combining the main word in the original phrase *Samlor* (①) with the borivasab (③). This kind of word formation process can be used freely at anytime in the daily speaking and writing of Khmer, and eventually these words can enter common usage even they are not found in the dictionary.

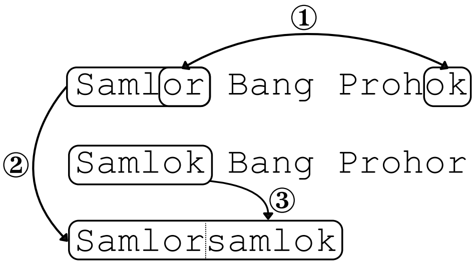

Fig. 1. Formation of a word using a borivasab.

Another example of a borivasab is shown below:

- កម្ទេចកម្ទី (debris) = ***kamtickamtee***

កម្ទេច ជា ធូលី (bits of dust) = ***kamtic*** *chea tulee*

⟶ កម្ទី ជា ធូលិច = ***kamtee*** *chea tulic*

B. Word Type 2: Compound Words

A compound word is a word that is composed of two or more single words. Most compound words can be classified into 7 patterns as in Table I. Here, N denotes Noun, Adj denotes Adjective, and V denotes Verb.

For the remaining cases, that we call collocation words, two or more words that are always written or used next to each other, are also considered to be compound words, and some examples are given in Table II.

C. Word Type 3: Compound Word with Prefix

The most 12 frequent words in the corpus were used as prefixes in the formation of compound words. They can

TABLE I
Compound Word Patterns

Patterns	Example
N_N	ឡាន_ក្រុង (city_car→bus)
N_Adj	កំហុស_ឆ្គង (mistake_improper→mistake)
N_V	កប៉ាល់_ហោះ (ship_fly→airplane)
Adj_Adj	ខូច_ខាត (broken_lose→damage)
Adj_N	ត្រី_កោណ (three_angle→triangle)
V_V	ផ្តល់_ឱ្យ (provide_give→provide)
V_N	ពិគ្រោះ_យោបល់ (consult_optinion→to consult)

TABLE II
Example of Collocation Words

Examples	Explanation and Definition
ក៏_ប៉ុន្តែ	or_but → but
ខ្ញុំ_បាទ	I_yes → I (for man)
នាង_ខ្ញុំ	she_I → I (for woman)
ដោយសារ_តែ	because_only → because
ផង_ដែរ	also_also → as well
ទទួល_ខុស_ត្រូវ	receive_wrong_right → be responsible for
ឥឡូវ_នេះ	now_this → now

be partitioned into three groups. The first group contains eight prefixes ការ, ក្តី, ភាព, សេចក្តី, អត្ត, អត្ថ, អនុ, អំពើ for making compound nouns. The second group contains three prefixes ព្រះ, លោក, អ្នក that are used for nouns representing humans and deities. The third contains only one prefix អ for negation. Table III shows examples of compound words constructed with prefixes from each group.

D. Word Type 4: Compound Words with Suffix

Only 3 words were used as suffices of compound words. They are កិច្ច, កម្ម, ភាព and examples of compound words formed with them are shown in Table IV.

Some rare compound words are formed by both prefix and suffix. For example, អ~ទទួល_មរណ^ភាព (death).

E. Annotation Tags

When words of these four types were identified in the corpus by the annotators, they were considered as a single token of segmented text and tagged according to the following scheme:

{} a pair of bracket characters were used for word boundaries;
_ the underline character was used to delimit the component words within compound words;
~ the tilde symbol was used to indicate prefix words;
^ the caret symbol was used to indicate suffix words.

TABLE III
Groups of Compound Word with Prefix

Prefix Group	Examples
Group One	- ការ + អប់រំ (educate) → ការ~អប់រំ(education) - សេចក្តី + សុខ (fine) → សេចក្តី~សុខ (happiness) - ភាព + ឯកោ (lonely) → ភាព~ឯកោ (loneliness) - អត្ថ (idea) + ន័យ (meaning) → អត្ថ~ន័យ (definition)
Group Two	- ព្រះ (god) + អាទិត្យ (sun) → ព្រះ~អាទិត្យ (the sun) - អ្នក (person) + បើកបរ (to drive) → អ្នក~បើកបរ (driver)
Group Three	- អ (not) + សកម្ម (active) → អ~សកម្ម (inactive)

TABLE IV
Compound Words with Suffix

Suffix	Example
កិច្ច	- អភិបាល (govern) + កិច្ច (task) → អភិបាល^កិច្ច (governance)
កម្ម	- ទំនើប (moden) + កម្ម (action) → ទំនើប^កម្ម (modernization)
ភាព	- ឯករាជ្យ (independent) + ភាព (state/condition) → ឯករាជ្យ^ភាព (independence)

The following is an example of some Khmer text with word boundary annotation:

{សុខា}{និយាយ}{ថា}{៖}{«}{ការ~សិក្សា}{នាំ_មក}
{នូវ}{ចំណេះ_ដឹង}{»}{តើ}{ឯក^ភាព}{ទេ}{?}

IV. Maximum Matching (MM)

Maximum matching is one of the most popular structural segmentation algorithms and it is often used as a baseline method in word segmentation [11]. 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 suboptimal. The segmentation process may start from either end of the sequences. We ran the maximum matching experiments with a dictionary that contained 27,070 unique words extracted from the manually annotated corpus used to train the CRF models. This was a set of every segment that occurred in the corpus.

V. Conditional Random Fields

Linear-chain conditional random fields (CRFs) [2] 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 the segmentation process. 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 $\mathbf{W} = \{w_1, ..., w_T\}$.

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

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

The feature set used in the models (character unigrams) was as follows (where t is the index of the character being labeled):

$\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$

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

VI. Segmentation Experiments

A. Data Setup

1) Training Data: In order to train the CRF segmentation models, we needed to construct a manually segmented corpus in accordance with the guidelines set out in Section III. This corpus was constructed over a six-month period using 4 human annotators in the following manner. First, manual word segmentation was done on 5,000 randomly selected sentences from the general Khmer web domain. For manual word segmentation, segmented according to four types of Khmer words as explained in Section III. We used the CRF++ toolkit [1] to build the CRF model with the 5,000 segmented Khmer sentences. We then used this CRF model to annotate a new set of unsegmented data. This automatic annotation of this data was hand-corrected and used to train a new CRF model. In this manner the annotation for the full 97,340-sentence corpus was bootstrapped.

The training data set included 3,435 sentences from Agriculture domain, 67,725 sentences from the Basic Travel Expression (BTEC) corpus [12], 2,915 sentences from the Buddhist religious domain, 2,052 sentences from the economic domain, 126 sentences from the medical domain, 98 sentences from the history domain, 665 sentences from law, 746 sentences from the management domain, 9,923 sentences from the news domain, 3,284 sentences from the scientific research domain, 6,009 sentences from stories, and 362 sentences from miscellaneous other domains. All of the training data was web data, except for the BTEC or Travel domain data.

2) Test Data: The 12,468-sentence test set was randomly selected from the full corpus, and consisted of: 490 sentences (14,879 words) from the agriculture domain, 9,989 sentences (75,902 words) from BTEC, 1,400 sentences (34,371 words) from the history domain, 393 sentences (12,488 words) from the news domain, 90 sentences (1,775 words) from stories, and 106 sentences (2,921 words) from miscellaneous other domains.

B. CRF Training

The CRF models were trained on character segmented Khmer. We used 10 tags for type of Khmer characters and they are C (Consonant), V (Vowel), IV (Independent Vowel), US (Upper Sign), AN (Atak Number), SUB (Subscript Sign), END (End of Sentence), NS (No Space), UNK (Unknown) (Refer Table V). Two separate models that used two different tag sets were trained. These two tag sets were: {1,2} and {1,2,3,4,5} using the tag number notation in Table VI.

C. Evaluation

The evaluation method used was the Edit Distance of the Word Separator (EDWS) [13]. As an example, given a non-segmented Khmer sentence (ខ្ញុំឈ្មោះស៊ីថា) represented as a sequence of 15 characters:

ខ␣្␣ញ␣ ុ␣ ំ␣ឈ␣្␣ម␣ោ␣ះ␣ស␣ ៊␣ ី␣ថ␣ា

The CRF segments this sentence into a sequence of words, and we indicate the segment boundaries by the space character '␣'. For example, an output might be: ខ្ញុំ␣ឈ្មោះ␣ស៊ី␣ថា. Let $\mathbf{X} = x_1x_2x_3...x_n$ represent an unsegmented string of characters. A segmented string (with $n = 15$) may look like:

$$x_1x_2x_3x_4x_5\sqcup x_6x_7x_8x_9x_{10}\sqcup x_{11}x_{12}x_{13}\sqcup x_{14}x_{15}$$

The EDWS measures how many edit operations (insertions, deletions and substitutions) are needed to jointly segment a given segmentation and a reference segmentation. Here the edit operations are based only on the segmentation token '␣': a substitution $\langle \sqcup, \sqcup \rangle$ being a correct segmentation point, insertions $\langle \sqcup, \varnothing \rangle$ corresponding

TABLE V
Tag Representing Type of Character

No	Tags	Meaning	Characters
1	C	Consonant	ក ខ គ ឃ ង ច ឆ ជ ឈ ញ ដ ឋ ឌ ឍ ណ ត ថ ទ ធ ន ប ផ ព ភ ម យ រ ល វ ឝ ឞ ស ហ ឡ អ
2	V	Vowel	◌ា ◌ិ ◌ី ◌ឹ ◌ឺ ◌ុ ◌ូ ◌ួ ◌ើ ◌ឿ ◌ៀ ◌េ ◌ែ ◌ៃ ◌ោ ◌ៅ ◌ំ ◌ះ ◌ៈ
3	IV	Independence Vowel	អា ឥ ឦ ឧ ឩ ឪ ឫ ឬ ឭ ឮ ឯ ឰ ឱ ឳ
4	US	Upper Sign	◌៉ ◌៊ ◌់ ◌៌ ◌៍ ◌៎ ◌៏ ◌័ ◌៑
5	AN	Atak Number (astrology)	៰ » ៱ ៲ ៳ ៴ ៵ ៶ ៷ ៸ ៹
6	LN	Lunar Number	᧠ ᧡ ᧢ ᧣ ᧤ ᧥ ᧦ ᧧ ᧨ ᧩ ᧪ ᧫ ᧬ ᧭ ᧮ ᧯ ᧰ ᧱ ᧲ ᧳ ᧴ ᧵ ᧶ ᧷ ᧸ ᧹ ᧺ ᧻ ᧼ ᧽ ᧾ ᧿
7	SUB	Subscript Sign	◌្
8	END	End of sentence signs etc.	◌៓ ។ ៕ ៖ ៗ ៘ ៙ ៚ ៛ ៜ ◌៝
9	NS	No Space	០ ១ ២ ៣ ៤ ៥ ៦ ៧ ៨ ៩ 0 1 2 3 4 5 6 7 8 9
10	UNK	Unknown	Characters outside the Khmer unicode set

TABLE VI
Tags Representing Segmentation

No	Tag	Meaning after corresponding character
1	0	Zero: no space
2	}{	Right and Left brace: Space
3	_	Underscore sign: compound word
4	~	Tilde sign: prefix
5	^	Caret sign: suffix

to segmentation points in the reference that are not in the given segmentation, and deletions $\langle \varnothing, \text{␣} \rangle$ corresponding to segmentation points in the given segmentation that are not in the reference. For example:

Ref: ខ្ញុំ␣ឈ្មោះ␣ស៊ីថា
CRF: ខ្ញុំ␣ឈ្មោះ␣ស៊ី␣ថា
Ref: $x_1x_2x_3x_4x_5\text{␣}x_6x_7x_8x_9x_{10}\text{␣}x_{11}x_{12}x_{13}x_{14}x_{15}$
CRF: $x_1x_2x_3x_4x_5\text{␣}x_6x_7x_8x_9x_{10}\text{␣}x_{11}x_{12}x_{13}\text{␣}x_{14}x_{15}$

In this example there are two substitutions after x_5 and x_{10}, and one deletion after x_{13}. The segmentation precision, recall and harmonic mean F-score are defined as:

$$\begin{aligned} \text{Precision } (P) &= \tfrac{N}{H} \\ \text{Recall } (R) &= \tfrac{N}{S} \\ \text{F-score} &= 2 \times \left(\tfrac{P \times R}{P + R}\right) \end{aligned}$$

where:
N is the number of substitutions;
H is the number of separators in the hypothesis;
S is the number of separators in the reference.

D. Results and Discussion

Table VII presents the main results of our segmentation evaluation. It is clear from the results that there were almost no difference in the performance of the CRF systems built using a 2-tag set and those that used a 5-tag set.

It can also be seen that the CRF models substantially outperform the MM systems in terms of the overall precision, recall and F-score. We believe that much of this difference is caused by out-of-vocabulary (OOV) words such personal names, location names, foreign words.

In order to study how the CRF models behave with varying amounts of training data, we ran a sequence of experiments that trained two CRF models (2-tag and 5-tags) on 10K, 20K, 30K, 40K, 50K, 60K, 70K, 80K and 90K sentences sampled randomly without replacement from the full manually segmented corpus. The results are shown in Figure 2. 10-fold jackknifing experiments were performed for each data point on the graph. It is clear from the figure that both of the CRF models perform almost identically, with the 5-tags model is slightly outperforming the 2-tags CRF model. With 5-tags CRF model, accuracy measurement of prefix, suffix and compound word tagging can be done and we got average F-score 93.40% for prefix, 88.73% for suffix and 90.91% for compound word on 6 domains.

When we inspected the word segmentation errors of the CRF model, we found two common errors: OOV errors and compound word errors. We analyzed the results in order to measure how well each of the approaches dealt with these cases. In each case we measured the accuracy of the segmentation for words of the respective types. Explicitly, in the case of OOVs, for each OOV in the reference the accuracy is given by the ratio of times the word is correctly segmented to the number of occurrences in the reference set. We found that the maximum matching method was unable to segment any of the OOVs correctly, this was because if a word is not in the dictionary, the default character segmentation was used. The CRF model had an OOV segmentation accuracy of 0.44. The CRF also showed much better performance on segmenting compound words, here the accuracy was 0.88, compared to only 0.57 with the maximum matching method.

1) OOV: The following is an example of an OOV segmentation error:

Gloss	story	American	Airline
Ref	រឿង	អាមេរិខេន	អ៊ែរឡាញ
MM	រឿង	អាម␣េ◌␣រ␣◌ិ␣ខេន	អ៊ែរឡាញ
CRF	រឿង	អាមេរិខេន	អ៊ែរឡាញ

The word for ’American’ was not in the dictionary used for maximum matching, and therefore a the word has been segmented into characters, whereas the CRF has been able to segment the OOV correctly.

2) Compound words: The following is an example of an compound word segmentation error:

Gloss	these	have	mistake	a lot	very
Ref	ទាំងនេះ	មាន	ភាពខុសឆ្គង	ច្រើន	ណាស់
MM	ទាំងនេះ	មាន	ភាពខុស␣ឆ្គង	ច្រើន	ណាស់
CRF	ទាំងនេះ	មាន	ភាពខុសឆ្គង	ច្រើន	ណាស់

The compound meaning ’mistake’ was not in the dictionary, however its component words were. The method of maximum matching has therefore segmented the word into its component words whereas the CRF has correctly segmented the word.

However, for some long compound words, the CRF model made errors where the maximum matching method did not. For example, the compound Khmer word កថាខណ្ឌ (meaning ’paragraph’ in English) is split by the CRF into two words as កថា (’word’ in English) and ខណ្ឌ (’section’ in English). This error was caused because the compound word only occurred once in the training corpus, whereas the components occurred 11 and 207 times respectively. Therefore the CRF had a preference to segment (incorrectly) into the components. The maximum matching technique on the other hand, segmented (correctly) into the longest possible word.

Numbers written with words also caused ambiguities. Numbers from five to nine are expressed relative to five. For example, six (ប្រាំមួយ) is formed from five (ប្រាំ) plus one (មួយ). If numbers are written in words continuously without using space (for example for a phone number) they can be segmented in several ways, therefore spaces must be used to disambiguate. For example the Khmer ប្រាំមួយ is literally ‘five one’, and could the number five (ប្រាំ) followed by the number one (មួយ), or it could be the number six. This problem can not be solved by the methods studied in this paper.

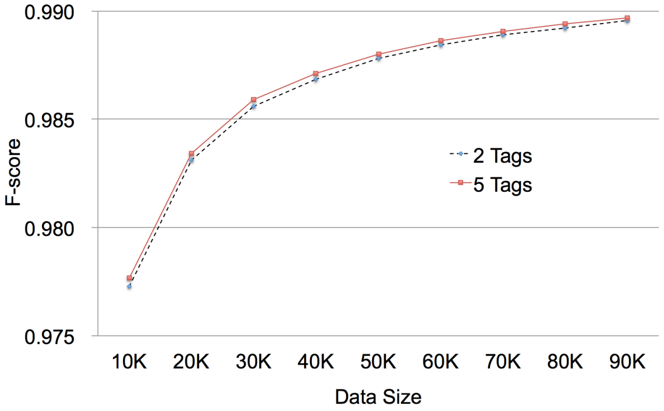

Fig. 2. F-Scores from training with CRF models on varying data set sizes.

However, if the numbers contain units such as រយ (hundred) or ពាន់ (thousand), then it is possible to disambiguate them, and in fact we observed that the CRF segmenter was able to learn to do this. Of course the method based on maximum matching was not able to generalize to be able to handle this issue.

VII. Machine Translation Experiments

A. Corpora

We used a selection of five language pairs from the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions [12]. The language pairs were selected to include languages from a variety of language groups. The languages were Chinese (zh), English (en), Japanese (ja), Khmer (km), Myanmar (my), and Vietnamese (vi). 155,121 sentences were used for training, 5,000 sentences for development and 2,000 sentences for evaluation.

B. Methodology

We used the phrase based SMT system provided by the Moses toolkit [14] for training the phrase-based machine statistical translation system. The Khmer was aligned with the word segmented target languages (except for the Myanmar language that was syllable segmented) using GIZA++ [15]. The alignment was symmetrized by grow-diag-final-and heuristic [16]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [17]. We use SRILM for training

TABLE VII
Word segmentation performance using CRF and maximum matching models.

No	Type of data	Maximum Matching			CRF with 2 tags			CRF with 5 tags		
		Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
1	Agriculture	92.082	93.376	0.927	99.261	98.149	0.987	98.920	99.185	0.991
2	BTEC	91.291	91.936	0.916	98.957	99.181	0.991	98.920	99.185	0.991
3	History	88.132	93.986	0.910	95.944	96.221	0.961	95.793	96.339	0.961
4	News	89.364	93.435	0.914	98.571	98.065	0.983	98.508	98.239	0.984
5	Story	92.759	93.286	0.930	99.641	98.579	0.991	99.584	99.407	0.995
6	Others	92.561	91.573	0.921	99.466	99.407	0.994	99.676	98.437	0.991
	AVERAGE	91.032	92.932	0.920	98.640	98.267	0.985	98.614	98.292	0.985

the 5-gram language model with interpolated modified Kneser-Ney discounting [18]. Minimum error rate training (MERT) [19] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1) [14].

C. Segmentation Schemes

Four types of segmentation were used in these experiments. One baseline segmentation scheme was obtained using the syllable segmentation scheme described in [20] which is similar to the KCC scheme mentioned earlier. Two baselines based on segmentation derived using the method of maximum matching were also studied: the first used a publicly available dictionary [8] which represents the current state-of-the-art in Khmer word segmentation; the second was based on the dictionary composed of the full set of types in the manually segmented corpus we created. The final segmentation scheme was that arising from the proposed method; that it was the segmentation from a CRF segmenter trained on the manually segmented corpus using a 5-tag tag set.

For the experiments in which translation was into Khmer, the output was re-segmented into sequences of syllables before evaluation in order to keep the segmentation the same, and therefore the BLEU scores comparable.

D. Results

TABLE VIII
Translation from Khmer (BLEU).

	Syllable	MM(p-dict)	MM(c-dict)	CRF(5 tags)
km-en	49.07	32.35	51.82	**59.51**
km-ja	23.46	22.36	29.96	**34.27**
km-my	27.43	21.63	33.61	**38.08**
km-vi	45.67	29.56	46.59	**53.39**
km-zh	23.72	16.78	28.40	**32.09**

TABLE IX
Translation into Khmer (BLEU).

	Syllable	MM(p-dict)	MM(c-dict)	CRF(5 tags)
en-km	49.86	48.08	56.38	**58.85**
ja-km	32.57	33.55	37.44	**38.49**
my-km	33.82	31.49	36.89	**38.25**
vi-km	47.93	44.14	53.13	**54.26**
zh-km	32.21	32.25	37.61	**39.20**

The results of the machine translation experiment are shown in Tables VIII and IX. In this table, ’MM(p-dict)’ denotes the maximum matching method based on the publicly available Khmer dictionary [8]. ’MM(c-dict)’ refers to the maximum matching method based on the dictionary extracted from the human segmented corpus used to train the CRF model. The highest performing systems are indicated in bold font. The proposed CRF word segmentation scheme gave rise to the highest BLEU scores in all experiments, the improvements in BLEU were substantial, ranging from 1.1 to 7.7 BLEU points. The differences between the best baseline, MM(c-dict), and the proposed CRF method were tested for significance using the paired bootstrap method [21]. All differences were significant ($p < 0.01$).

VIII. Conclusion

In this paper we proposed a segmentation method for the Khmer language based on a supervised CRF-based segmentation method. In order to train the segmenter we constructed a large word-segmented corpus of Khmer text. We evaluated the performance of our segmenter in terms of both segmentation quality and also in terms of its effect when applied to statistical machine translation. The experiments show it is possible to segment very precisely using the word segmentation scheme that we defined. The proposed method achieved an average f-score of 0.99 on test data, exceeding the performance of a maximum matching baseline which achieved an average f-score of 0.92. Furthermore, our experiments on machine translation show that the high levels of segmentation quality can be translated into to large improvements in end-to-end performance of real-world NLP application. In a set of statistical machine translation tasks, the segmenter was able to improve system performance over the best maximum matching baseline by a wide margin (from 1.1 to 7.7 BLEU points).

In summary the primary conclusions that can be drawn from the work in this paper are that accurate word segmentation looks likely to be a necessary prerequisite for natural language processing in the Khmer language. Simple segmentation strategies based on maximum matching do not appear to be sufficient. We have demonstrated that it is possible to achieve a high level of segmentation performance on Khmer with a supervised CRF model.

Acknowledgment

We would like to express our gratitude to our colleagues Mr. Mech Nan, Mr. Sorn Kea and Mr. Tep Raksa from National Institute of Posts, Telecommunications and Information Communication Technology, Cambodia for their help in segmenting 97,340 sentences of BTEC Khmer Corpus and General Domain Corpus manually.

References

[1] Taku Kudo. Crf++ an open source toolkit for crf (2005), https://taku910.github.io/crfpp/.

[2] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[3] Sopheap Seng, Sethserey Sam, Laurent Besacier, Brigitte Bigi, and Eric Castelli. First broadcast news transcription system for khmer language. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[4] Chea Sok Huor, Top Rithy, Ros Pich Hemy, and Vann Navy. Word bigram vs orthographic syllable bigram in khmer word segmentation.

[5] Chea Sok Huor, Top Righty, Row Pich Hemy, and Vann Navy. Detection and correction of homophonous error word for khmer language. In *PAN Localization Cambodia*, 2006.

[6] Channa Van and Wataru Kameyama. Khmer word segmentation and out-of-vocabulary words detection using collocation measurement of repeated characters subsequences. *GITS/GITI Research Bulletin*, 2012-2013:21–31, 2013.

[7] Narin Bi and Nguonly Taing. Khmer word segmentation based on bi-directional maximal matching for plaintext and microsoft word document. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014, Chiang Mai, Thailand, December 9-12, 2014*, pages 1–9. IEEE, 2014.

[8] Chuon Nath. *Dictionnaire Cambodgien*. Edition de L'institut bouddhique, Phnom Penh, 1967.

[9] Madeline E. Ehrman. *Contemporary Cambodian Grammatical Sketch*. US Goverment Printing Office, 1972.

[10] Sisovat Poraksy. *The Compound words in Khmer language* (បរិវារស័ព្ទ ក្នុង ភាសា ខ្មែរ ដោយ ស៊ីសុវត្ថិ ប៉ូរក្ខស៊ី)*(in Khmer)*. Phnom Penh Sisowath Poraksy, 1972.

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

[12] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proceedings of EUROSPEECH-03*, pages 381–384, 2003.

[13] Joy. Chinese word segmentation evaluation toolkit, http://projectile.sv.cmu.edu/, 2004.

[14] Philipp Koehn and Barry Haddow. Edinburgh's Submission to all Tracks of the WMT2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164, 2009.

[15] F. J. Och and H. Ney. Improved statistical alignment models. In *ACL00*, pages 440–447, Hong Kong, China, 2000.

[16] Philipp Koehn, Franz Josef Och, , and Daniel Marcu. Statistical phrase-based translation. In *In Proceedings of the Human Language Technology Conference*, Edmonton, Canada, 2003.

[17] Christoph Tillmann. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[18] Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, 2002.

[19] Franz J. Och. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, 2003.

[20] Ye Kyaw Thu, Vichet Chea, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. A large-scale study of statistical machine translation methods for khmer language. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 2015)*, Shanghai, China, October 2015.

[21] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.