LaoWS: Lao Word Segmentation Based on Conditional Random Fields

Sisouvanh Vanthanavong
Information Technology Research Institute, NAST, Lao PDR
sisouvanh@nast.gov.la

Choochart Haruechaiyasak
Human Language Technology laboratory, NECTEC, Thailand
choochart.haruechaiyasak@nectec.or.th

Abstract

In this paper, we propose a word segmentation model based on the Conditional Random Fields (CRFs) algorithm for Lao language called Lao Word Segmentation (LaoWS). LaoWS is trained from a given corpus called Lao text corpus (LaoCORPUS). LaoCORPUS contains approximately 100,000 manually tagged words from both formal and informal written Lao languages. Using the CRFs algorithm, the problem of word segmentation can be formulated as a sequential labeling task in each character labeled with one of two following classes: word-beginning (B) and intra-word (I) characters. To train the model, we design the feature set based on the character tagged corpus, example, by applying all possible characters as features. The experimental results showed that the performance under the F-measure is equal to 79.36% compared to 72.39% by using the dictionary-based approach. As well as, using the CRFs approach, the model can segment name entities better than the dictionary-based approach.

Index Terms— *Lao Word segmentation, Tokenization, Conditional Random Fields*

1. Introduction

Especially the development of localization, Lao language is one of many languages in South East Asia countries which does not have any white space between syllables and words that called mono-syllable language. By the way, Lao language is still lacking a standard of lexical and Dictionary Base for language development of information technology field. However, this paper will technically present the key point of word segmentation task in text allocation analysis depending on a given corpus.

For many years, the researchers have been developing word segmentation in many difference languages by using Machine Learning Based and Dictionary Base. The main purpose of machine learning base is an independence of dictionary that opposites of Dictionary Based Approach. The unknown words and name entity can be solved by a model classification of machine learning approach. For example, neural network, decision tree, conditional random fields (CRFs) and etc.

Recently, many organizations and private sectors in Laos try to develop Lao language especially the information technology fields (text processing). LaoScript[1] for Windows, it has been developed for many years using in Microsoft Office and Text editor. Otherwise, PANL10N[2] project is one of the sectors to research and develop about natural language processing in Asian language. The main purposed of this paper is to produce machine learning base by a model classification for word segmentation task in the specific area into text processing from a given corpus and evaluation the proposed method by the performance of F-measure. To remain this paper is established as follows, the next section is about previous work in word segmentation. In section 3, a brief of CRFs algorithm, section 4 the main propose of word segmentation, and to be more practical, there will be the experiments and results in section 5, eventually section 6, will be the conclusion of this research.

2. Related Work

Recent year, Lao localization development has been analyzed in many fields, especially text

[1] http://www.laoscript.net
[2] http://www.panl10n.net

processing such as: line breaking system, convert fonts, spelling check and etc. For, i.e., Line breaking is very important for justification in Lao language, according to Lao line breaking (Phissamay *et al*, 2004) that created a new rule and condition for solving the problems of syllable breaking system in text editor, which's given the best performance up to 98%. However, the difficulty to technically improve from syllable breaking to word breaking system in text processing is about lacking of the lexical corpus and dictionary standardization.

Fortunately, Lao and Thai have a very similar language by spoken and writing system. Years ago, Thai language (Kruengkrai and Isahara, 2006; Theeramunkong and Usanavasin, 2001; Khankasikam and Muansuwan, 2005) was researched from syllable segmentation to word segmentation task using a rule-based system of language models and lexical semantic approaches, the decision tree model solves the word segmentation without a dictionary based, this result is given the accuracy approximately 70%, for the Dictionary Based Method gives the high accuracy approximately 95% with a context dependence.

Thai word segmentation approach (Haruechaiyasak *et al*, 2008; Thai Lexeme Tokenization. Online: 2010) produced the two different algorithms such as the Dictionary Bases (DCB) and Machine Learning Base (MLB). Normally, DCB approach (Sornil and Chaiwanarom, 2004) uses the Longest Matching (LM) technique to consider about information segmentation with the long word; Maximal Matching (MM) uses the existing word in the Dictionary base by selecting the segmented series that yields the minimum number of word taken. Otherwise this research described the experiments of the n-grams model of different character types from Thai text corpus by using Machine Learning Approach such as Naive Bayes (NB), Decision tree, Support Vector Machine (SVM), and CRFs. The result of this research selects the CRFs algorithm as the best way to detect Thai word boundary in machine learning based, with the precision and recall of 95.79% and 94.98% respectively.

Therefore, this research uses the CRFs algorithm in the challenging task of Lao word segmentation development.

3. CRFs Algorithm

In a CRF algorithm (Wallach, 2004; Lafferty *et al*, 2001; Alba *et al*, 2006) by a chain-structured model depending on each label sequence contains beginning and ending states respectively y_0 and y_{n+1}, the probability of label sequence y that given an observation sequence x is $p(y \mid x, \phi)$ maybe efficiently computed matrices. We define y and y' that are the label sequences of an alphabet *Y*, a set of n+1 matrix $\{M_i(x) \mid i = 1,...,n+1\}$, where each $M_i(x)$ is a $|Y \times Y|$ matrix elements may be written

$$M_i(y', y \mid x) = \exp(\sum_j \phi_j f_j(y', y, x, i)).$$

The un-normalized probability of label sequence y given observation sequence x that considers the product of the matrix elements of the form of these label sequences:

$$p(y \mid x, \phi) = \frac{1}{Z(x)} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i \mid x).$$

Similarly, the normalization factor above $Z(x)$ is given by the (*start* and *end*) entry of the product of all *n+1* and $M_i(x)$ matrices as follows:

$$Z(x) = [M_1(x) M_2(x) ... M_{n+1}(x)]_{start,end}$$

$$= [\prod_{i=1}^{n+1} M_i(x)]_{start,end}$$

To assume that the conditional probability of a label sequence y might be written as:

$$p(y \mid x, \phi) = \frac{\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i \mid x)}{(\prod_{i=1}^{n+1} M_i(x))_{start,end}}$$

Where $y_0 = start$ and $y_{n+1} = end$

4. The Proposed Solution

To solve Lao word segmentation problems, we would like to draw a clear picture based on which model could be defined exactly

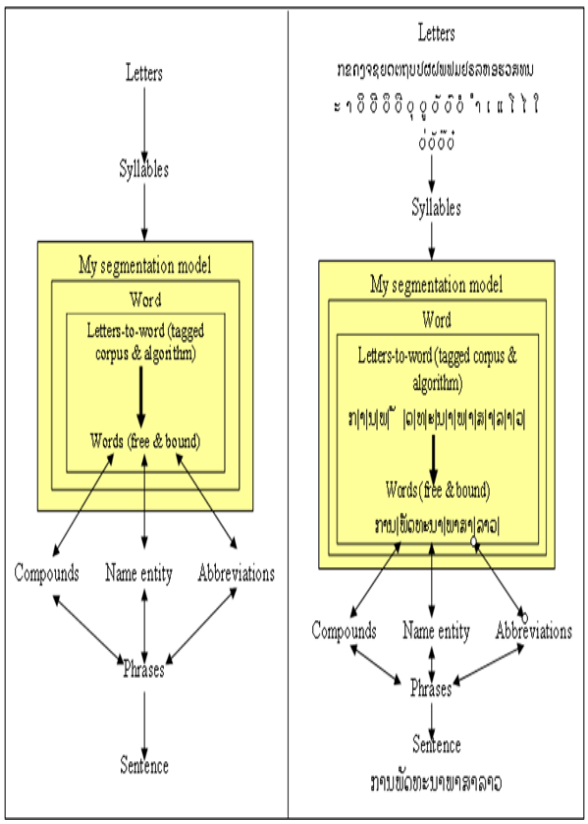

Figure 1: Letters-to-sentence modification

According to above figure, the procedures still have many complicated problem occurred from one place to another, meanwhile every place also has different problems, this figure particularly emphasizes solely at the word segmentation pattern (in the box). Initially we consider step by step uniting the letters to word (data input as letters) result (output data or word boundaries) such as:

The beginning step, the incorporative procedure (Letters-to-word) by determining the set of character types which will be implemented at a model for characters recognition or Machine Learning Base, for this research shall apply to a model for the tagged corpus as the data base, this tagged corpus is created from word segmentation of text information, in term of word manual to text corpus and thereafter the tagged corpus will be generated as the targeted rule which will be explained in the following topic. The second step there is acquired from letter to word consequence which those words will be correctly segmented base on the word boundaries as stipulated in the text corpus. For example, ການພັດທະນາພາສາ ລາວ kan-phat-ta-na-pa-sa-lao "Lao language development", in this case we will implement the stipulated word segmentation approach therefore the consequence should be in the form of "ການ|ພັດທະນາ|ພາສາ|ລາວ|" base on the text corpus.

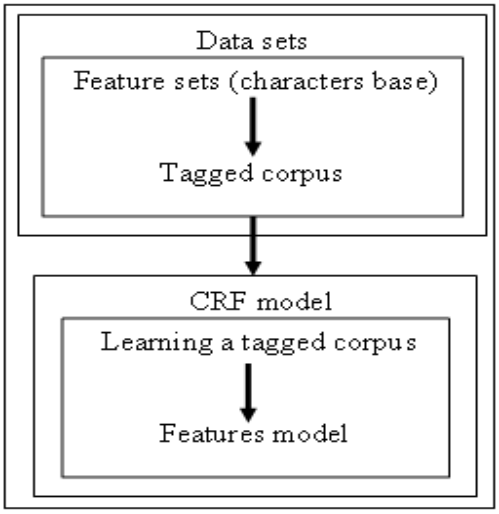

Figure 2: Implementing a model for word segmentation

Figure 2 is identifying that the data set is having the significant role to create the appropriate condition to precede CRFs model. This phase will solemnly stipulate function between component and outcome base on CRFs requirement which will be clarified as following:

Firstly, The feature sets is representing the Lao character sets which is already defined based on the Lao language components such as consonant (main consonant, associate consonant, combination consonant), vowel and tone mark. Moreover, it can define the special symbol, digit and quote sequentially.

Second, We purify the text string and then segment each sentence as well as parsing words by words manual using a delimiter tag (indicated by "|"). We have completed a new text corpus which includes appropriately 100,000 words covering various Lao words, and some special symbols that is called the Lao Text corpus

(LaoCORPUS). We will use the LaoCORPUS as a text file to create Lao tagged corpus.
Third, CRFs model needs a training set to perform segmentation task of text information, a training set is used as data set for CRFs learning model which our data set perform types feature based on character sets, CRFs model is predicted the possible character features from text information by using conditional probability of Hidden Markov Model (HMM) (Rabiner and Juang, 1986; Sarawagi and Cohen, 2005) and Max-entropy HMM combination (Zhao *et al*, 2007).
Finally, Implementing features model, beside the rule and condition, we can generate a training corpus to a feature model based on the character sets. The feature set of character types for constructing a model is the n-gram of characters for backward and forward the word boundaries according to those character types.

5. Experiment and Result

The CRFs algorithm approach will learn the characteristics of text information as a binary classification problem according to a set of type features. Basically, we use an open source software package based on CRFs algorithm, CRF++0.53 (Kudo Taka, 2005-2007) is a simple package, customizable and open source implementation based on the CRFs algorithm. It is able to predict each character from data input and categories it as one of two classes such as: the beginning of a word, and the intra-word characters. Beginning of a word is defined as a labeled class (indicated by "B" in our text corpus), and intra-word characters is defined as a labeled class (indicated by "I" in our text corpus). Based on the machine learning details, we need to generate a text string into two conditions, as shown in Figure 3.

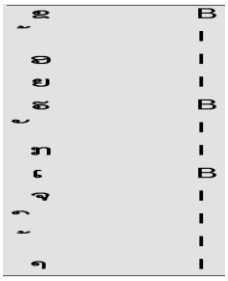

Figure 3**:** Example of a text generated into character tagged corpus as in word-beginning "B" or intra-word "I" characters

We first purify a text string in where each character is tagged with either word-beginning or intra-word characters. We need to built a tagged corpus as well as possible in CRF format, the tagged corpus in which word boundaries are explicitly marked with special characters, this is a machine learning based that can be verified to analyze a tagged corpus based of type features enclosing these words as boundaries. For the Lao language, we defined the character type for segmentation task into fifteen differences type feature.

We use LaoCORPUS (approximately 100,000 words) to evaluate the performance among different word boundary approaches. We split the text corpus randomly into exam nation and test sets (each set contains 20%). However, we are given a test set of 20% instead, and used the training sets increasing from 20%, 40% and 60% to 80%. The three values of F-measure, precision, and recall are used for performing evaluation.

Value (%)	Size of Text Corpus			
	20K	40K	60K	80K
Precision	75.98	78.43	78.52	80.28
Recall	73.07	76.01	76.77	78.45
F-score	74.49	77.20	77.64	**79.36**

Table 1: CRFs evaluation by Text corpus size

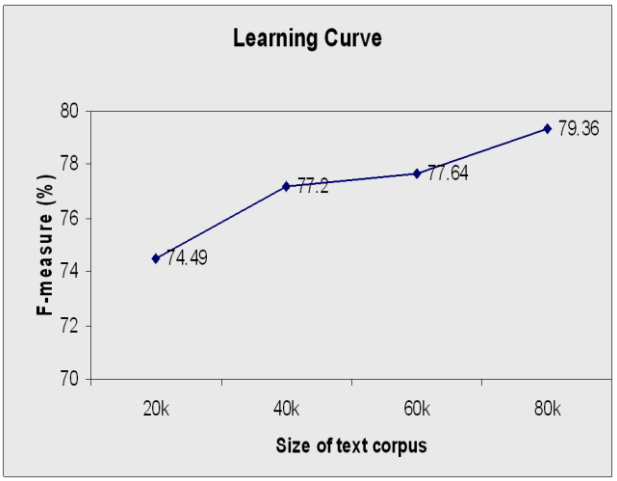

Figure 4: Learning curve from evaluation of text corpus using CRFs algorithm

Otherwise, word segmentation task was compared by our main approach (CRFs model) to another approach such as: dictionary-base (DCB). We also used a test set of 20% instead

from LaoCORPUS the same as previous section, as the result of name entity (NE) that is given details:

NE	Segmented by CRFs	Segmented by DCB
Person	ເພຍ|ວິນເຄັນ|	ເພຍ|ວິນ|ເຄ|ັນ|
Place	ເມືອງ|ວໍຊິງຕັນ|ດີຊີ	ເມືອງ|ວໍຊິ|ງຕັນ|ດີ|ຊີ
Company	ບໍລິສັດ|ລໍຣີລາດ|	ບໍລິສັດ|ລໍ|ຣິ|ລາດ|

Table 2: Comparison of segmented CRFs and DCB

The segmented CRF can be solved these name entities better than the segmented DCB (in Table 2). As the result of CRF, we get a sequence of words correctly as well as using the DCB approach, for example, CRFs can merge |**ເຄ**|**ັນ** to be one segment and join it with a previous segment |**ວິນເຄັນ**| thus the segment has a full correct meaning. The correct answer refers to the segmented word such **ວິນເຄັນ**|. To evaluate these approaches are shown below

Approach	Precision	Recall	F1
DCB	80	66.67	72.73
CRFs	80.29	78.45	**79.36**

Table 3: Evaluation of CRFs and DCB approach

6. Conclusion and Discussion

In this paper, we proposed and compared Dictionary Based and Machine-Learning Based approaches for Lao word segmentation using a tagged corpus. Many previous works have proposed algorithms and models to solve word segmentation problem for languages such as Thai, Chinese and Japanese, however, this research aims to construct a model for Lao language. For the machine-learning based approach, we applied the Conditional Random Fields (CRFs) to train a word segmentation model. We performed the evaluation of a Machine Learning Based approach using CRFs against the Dictionary Based approach. According to the evaluation, the best performance is obtained the CRFs algorithm with a character tagged corpus. The best result based on the F-measure is equal to 79.36% compared to 72.73% using the dictionary-based approach.

Therefore, to improve the performance further, we need to enlarge the corpus size for training the model. In general, to effectively train a machine learning model especially in NLP tasks, a large size of corpus is needed. For example, compared to Thai word segmentation (Haruechaiyasak *et al*, 2008), the best performance of F1-measure equal to approximately 96% is achieved with the corpus size of 7 million words. This research will be useful for other applications such as: word line breaking system, machine translation, speech processing (text-to-speech, speech recognition) and image processing.

For future work, we plan to achieve better performance by using syllables (as opposed to characters) as a basic unit for training a model. Another idea is to integrate both the Dictionary Based and Machine-Learning Base approaches, for example, a hybrid approach. The dictionary-base will be used for unknown segment checking on the outputs from the machine-learning base approach.

Acknowledgement

I would like to show sincere gratitude to three people, the first one is my advisor Dr. Choochart Haruechaiyasak who always share me ideas and advices, Second is PAN localization project to support me proceeding the research, Finally, there is my wife who always stays by my side and cheer me up when I confronted with some problem during doing this research.

References

Choochart Haruechaiyasak et al. 2008. *A Comparative Study on Thai Word Segmentation Approaches*. Proceedings of ECTI Conference, (1) 12-128.

Choochart Haruechaiyasak and Sarawoot Kongyoung. *LexTo: Thai Lexeme Tokenization*. [Online] March 2006. [Cited 2010 Jan 5]. Available from: http://www.hlt.nectec.or.th/

C. Kruengkrai. and H. Isahara. 2006. *A Conditional Random Field Framework for Thai Morphological Analysis*. Proceedings of the Fifth International

Conference on Language Resources and Evaluation.

Enrique Alba. 2006. *Natural language tagging with genetic algorithms*. Proceedings of Science Direct, (100)173-182.

John Lafferty et al. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. Proceedings of 18th International Conference on Machine Learning, pp. 282-289.

Hanna M. Wallach. 2004. *Conditional Random Fields: An Introduction*. Technical Report of University of Pennsylvania, USA.

Krisda Khankasikam and Nuttanart Muansuwan. 2005. *Thai Word Segmentation a Lexical Semantic Approach*. Proceedings of the 10th Machine Translation Summit, pp. 331, Thailand.

Kudo Taka. *CRF++0.53 yet another CRF Toolkit*. [Online] 2005-2007. [Cited 2009 May 06]. http://sourceforge.net/projects/crfpp/files

Ohm Sornil and Paweena Chaiwanarom. 2004. *Combining Prediction by Partial Matching and Logistic Regression for Thai word segmentation*. Proceedings of the 20th International Conference on Computational Linguistics.

Phonepasit Phissamay et al. 2004. *Syllabification of Lao Script for Line Breaking*, Technical Report of STEA, Lao PDR.

Rabiner. L. and Juang. B. 1986. *An introduction to hidden Markov models*. Proceeding of IEEE on ASSP Magazine. (3): 4 – 16.

Sunita Sarawagi. and William W. Cohen. 2005. *Semi-Markov Conditional Random Fields for Information Extraction*. Proceeding of Advances in Neural Information Processing Systems. (17): 1185-1192.

Thanaruk Theeramunkong and Sasiporn Usanavasin. 2001. *Non-Dictionary-Based Thai Word Segmentation Using Decision Trees*. Proceedings of the First International Conference on Human Language Technology Research, Thailand.

Zhao Zipinga et al. 2007. *A Maximum Entropy Markov Model for Prediction of Prosodic Phrase Boundaries in Chinese TTS*. Proceeding of IEEE International Conference. pp. 498 – 498.