E-007

A Rule-based Approach for Khmer Word Extraction

Channa Van and Wataru Kameyama
Graduate School of Global Information and Telecommunication Studies, WASEDA University
channa@fuji.waseda.jp, wataru@waseda.jp

1 Introduction

Khmer is a non-segmented language in which words are written without any explicit boundary. This causes many issues in Khmer Language Processing including Word Segmentation, Information Retrieval, Machine Translation and so on. The word extraction, which is the process of identifying the words in a text, shall be carried out to solve these issues. The most common approach in word extraction is a lexicon-based approach such as the Longest Matching algorithm being able to handle well all words found in a dictionary. However, since dictionary cannot cover all the words of the language, this leads to the issue of unknown words identification. Therefore in this research, a trainable rule-based approach for Khmer word extraction is proposed. The proposed approach aims to improve the accuracy of Khmer word extraction by detecting and identifying words which are not able to be discovered by using dictionary.

2 Proposed Approach

Our approach is based on a trainable rule-based approach where rules are obtained by the rule training process using a text corpus. The rules obtained by the training are used to identify words in the word extraction process. The detail of rule training and word extraction are presented in the following subsections:

2.1 Rule Training

Rule training is the process of extracting rules based on a text corpus (Figure 1). A Khmer text corpus, which consists of 1050 articles [1], is used for the training. First of all, all the texts in the corpus are extracted into strings by using Longest Matching algorithm using a Khmer lexicon which is based on the Khmer spelling dictionary of the Royal Academy of Cambodia. Then, the SEQUITUR algorithm [2] is used to detect the string sequences which are repeated more than once. The repeated sequence are replaced by a rule, and the rule is kept in a rule set which is to be used for word extraction. Each rule is a sequence of two elements that can be a string or a rule which can be represented as following:

$$R_i \leftarrow XY$$

Where:

- X is a string or a rule.
- Y is a string or a rule.

After that, each rule is tagged to be a word or not. The tagging process is based on Entropy [3] and Mutual Information (MI) [4] value of each rule. The Entropy is calculated by the formula as below:

$$E(xy) = -\sum_{\forall y \in A} P(xy|y) \log_2 P(xy|y) \quad (1)$$

Where:

- x is the considered rule.
- A is a set of alphabet.
- y is any string in A and is appeared after x.

And the Mutual Information is computed by the following formula:

$$I(x,y) = \log_2 \frac{P(xy)}{P(x)P(y)} \quad (2)$$

Where:

- x and y are the two strings in the sequence of a rule.

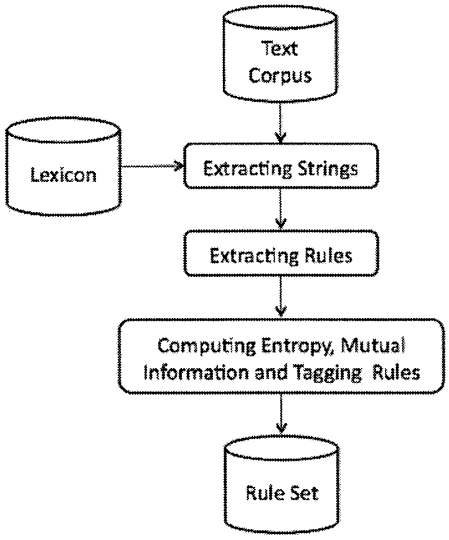

Figure 1: Rule Training.

2.2 Word Extraction

Figure 2 shows the process of word extraction. Like the rule training, the word extraction is started by extracting the strings from the text using Longest Matching algorithm and lexicon. And then, rules are extracted from the text by using the SEQUITUR algorithm the same as in the rule training process. Finally, the extracted words are the result of matching rules between obtained rules and the rule set obtained by the rule training.

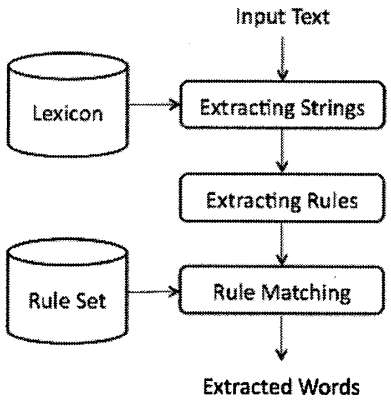

Figure 2: Word Extraction.

3 Experiments

3.1 Experiment Procedure

Twelve articles of Khmer text which consist of 4058 words are used in the experiments. Three types of experiment were carried out. First, the experiments are based on the entropy. Five values of entropy were used: 0.75, 1.13, 1.5, 2.0 and 2.5 for tagging the rules to be words in five different experiments. Then in the second type of experiments, the rules tagging are based on the mutual information: 5.5, 6.5, 7.5, 8.5 and 9.5. And the final experiment is based on the combination of both entropy and mutual information values which give the best result in the first and second type of experiment. The evaluation is based on the results of precision and recall of the word extraction which are calculated by the following formulas:

$$Precision = \frac{\#(Words\ Extracted)}{\#(Strings\ Extracted)} \quad (3)$$

$$Recall = \frac{\#(Words\ Extracted)}{\#(Words)} \quad (4)$$

3.2 Results and Discussion

Based on the experiment results (Table 1), when the values of both entropy and mutual information increase, the recall values also increase. This can be proven by the increase of recall from 87.66% to 89.22% in the case of entropy, and the improvement from 88.85% to 89.11% of recall in the case of mutual information. In addition, when both attributes are combined, the proposed approach can achieve 84.39% and 89.32% of precision and recall respectively.

Table 1: Results of Word Extraction.

	Value	Precision	Recall
Entropy	0.75	84.74%	87.03%
	1.13	84.56%	87.66%
	1.5	84.53%	88.05%
	2	84.20%	89.11%
	2.5	83.20%	89.22%
MI	5.5	83.32%	87.70%
	6.5	82.08%	88.85%
	7.5	79.56%	88.81%
	8.5	78.15%	88.94%
	9.5	77.51%	89.11%
Entropy, MI	2.5, 9.5	84.39%	89.32%

4 Conclusion and Future Works

Our proposed approach can achieve the significant results for Khmer word extraction which gives a good motivation for the future of Khmer Word Segmentation. Many works still can be tackled in the future in order to improve the accuracy of the system. From our observation, most of the wrong words extracted are the proper name of people and places and the compounding words. These issues require to be solved based on the specific characteristics of Khmer lexicology. The works on Khmer lexicology such as Khmer name entity rules, compounding rules, spelling rules and so on, are in the list of our extension works in the near future.

References

[1] Van, C. and Kameyama W. 2010. Building a Khmer Text Corpus. *The 72th Annual Conference of Information Processing Society of Japan*. March 2010. Tokyo. Japan.

[2] Nevill-Manning, C. and Witten I. 1997. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, vol. 7, pp. 67–82

[3] Shannon, C.E. 1948. A Mathematical Theory of Communication. Bell System Technical Journal 27, pp. 379-423

[4] Church, KW., Robert L. and Mark L.Y. 1991. A Status Report on ACL/DCL. In *Proceeding of 7th Annual Conference of the UW Center New OED and Text Research: Using Corpora*, pp. 84-91