

Method of Word Segmentation in Laos Based on Maximal Matching of Syllables

Wenjie Huo^{1,2}, Lanjiang Zhou^{1,2}, Feng Zhou² and Bei Yang^{1,2}

¹School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

²The Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China

Abstract. Word segmentation is an important support of semantic analysis, Machine Translation, QA, knowledge mapping research work, mainly used in information retrieval, text processing, data processing and many other areas of Natural Language Processing. Therefore, the realization of word segmentation is a very meaningful work. The method of this paper is to segment the syllables of the text corpus of Lao language and complete the maximal matching of syllables and dictionaries. Then match the results of the word segmentation and the error dictionary, and correct some wrong words by the error dictionary. Finally, we use regular expressions to match the corresponding word strings in segmentation results and correct the wrong words by some artificially formulated rules of the alphabet, numbers, etc. in the Lao language. It can improve the efficiency and accuracy rate of Laos Word Segmentation.

Keywords: Word segmentation, the maximal matching, Syllable segmentation.

1. Introduction

Words are the smallest meaningful units in natural language. The goal of the word segmentation is to divide the sentence into words. Word segmentation is the most basic work in the Natural Language Processing. In information processing, word segmentation is the basis of many other work, the accuracy of the word segmentation directly affects the follow-up work. In the aspect of text similarity computation, text clustering, search engine and information retrieval can not be separated from the work of word segmentation. And these information processing work is also an indispensable part of this powerful Internet Age.

The Lao language is composed of syllables, and the syllables are composed of letters. Therefore, the word segmentation of Lao can not be regarded as a sequence of words and combine into words like Chinese. Considering the rules of language itself, we segment words from the aspect of syllable size. At present there are several commonly used methods of word segmentation: the method of word segmentation based on semantics^[1] and the method of word segmentation based on dictionary^[2]. The method of word segmentation based on dictionary is learning language knowledge^[3,4] from large-scale corpus and getting a model of word segmentation, using this model to complete the word segmentation.

2. Word segmentation method based on maximal matching of syllables

2.1 Get a matching dictionary

The large-scale corpus on the Internet has played a very important role in the development of Natural Language Processing, which has a very good coverage. In the word segmentation method based on dictionary, dictionary resource is the most important condition to complete the word segmentation. We found the Lao language dictionary which contains 15768 commonly used words in total and a large number of English-Chinese dictionaries and Chinese-Laos dictionaries in the Lao language network and English Network. We use the English-Chinese dictionaries and Chinese-Laos dictionaries to intercept the part of the Lao language, and put these words into a file. Then use HashSet in the Java combined two files and remove duplicate. Finally we get the Lao language dictionary, containing 31719 words in total, and save words reversely according to the length of word.

2.2 The syllable rule of Lao language

The letter and the mark of Lao language can not appear in any placeholders. Table 1 is the position constraint of a word consisting of vowels, consonants, and tones. Among the Xn , n represents the input sequence of the letters in the word. Only X does not follow this order (X is typed between $X1$ and $X2$). Tone is typed after the vowel $X0$, consonant $X1$, the main consonant X and vowel $X3$, $X4$ (if $X0$, $X1$ and $X2$ are not empty, can not be empty).

Table 1. The syllable structure of Lao words.

		X5						
		X4						
X0	X1	X	X6	X7	X8	X9	X10	
		X2						
		X3						

- $X0$: $X0$ represents a vowel which always appears before the main consonant X ;
- $X1$: $X1$ represents a consonant which appears before the main consonant.
- X : X represents the consonant.
- $X2$: $X2$ represents a consonant which is followed by the main consonant X and appears below the main consonant.
- $X3$: $X3$ represents the vowel which appears below the main consonant.
- $X4$: $X4$ represents the vowel which appears above the main consonant.
- $X5$: $X5$ represents the tone which appears above the main consonant or vowel.
- $X6$: $X6$ represents the vowel which appears behind the main consonant. Its role is that when there is no vowel in the syllable, it appears together with X_6 .
- $X7$: $X7$ appears behind a vowel. $X7_1$ represents the end of a syllable, and it does not appear together with the tone.
- $X8$: $X8$ represents the consonant which can be alternated.
- $X9$: $X9$ is the consonant that can be foreign language pronunciation, it appears together with $X10_3$ in syllables .
- $X10$: $X10$ represents a mark and usually appears at the end of the syllable.

2.3 Syllable segmentation

The syllable segmentation algorithm of Lao language is defined as follows:

When we encounter punctuation and space which is not the letter of the Lao language in the input array, then record the boundaries of syllables.

Filtering the Lao language characters from the input array and removing the punctuation and space which is not the letter of the Lao language. And then reorder the character.

Then we reorder the letters which writing is not standardized and list the character which is marked appearing in all possible placeholders.

Steps of syllable segmentation:

Mark syllable boundary by input array, add up characters which are not Lao language, and store the Lao language text in the new array;

Reorder the letter which written order is not standardized;

Use rules to determine the syllable boundaries, and segment syllable;

Mark the position of placeholder of each character;

Put back the Lao language character to the original array.

2.4 Longest syllable matching

The word boundary and the syllable boundary of Lao Language are the same. We can use the matching based on syllable instead of the matching based on letter. It can avoid cutting the sentence into individual letters. The matching based on the longest syllable is matching the syllable sequence in the dictionary^[5]. Given a syllable of the set S_y , and the word dictionary set D_y , the result of the longest syllable matching is the result of the word segmentation, denoted as R . In the matching process, firstly, the pointer B_s points to the beginning of the syllable which is defined and the pointer E_s points to the last syllable. Then match words from the pointer B_s to E_s in the dictionary. If the match is successful, the words are recorded in R . Then pointer B_s points to E_s+1 , and E_s points to the end of next set of syllables. Else E_s-1 , until $E_s=B_s$.

2.5 The acquisition and rule making of the error dictionary

In this paper, we do a statistical analysis on 37931 words in the Lao language word corpus. We select the common words of Laos which have high error frequency and always appear, and make a statistical classification. Then we can get a error dictionary of Lao language combine 176 words which case changes lower and the result does not produce word segmentation ambiguity with context.

The reason for this error is that there is no long word in the dictionary, which leads to the error of the segmentation. Rely on artificial to make rules in the process of word segmentation, and use these rules to deal with some special cases in the result of word segmentation such as numbers, English, and foreign words.

3. Experimental results and analysis

In the experiment, we use the common evaluation index to evaluate the segmentation effect of the word segmentation model, that is, the word segmentation accuracy. Usually use the following formula to calculate:

- R: Accuracy rate
- n: The number of words which is segmented correctly.
- N: The number of all the words.

$$A = (n \div N) \times 100\%$$

3.1 Word segmentation by different dictionary structure

The experimental data of this chapter are the dictionary with 15768 words and the extended dictionary with 31719 words. Then using the method of syllable segmentation to segment the syllable of the Lao language text which contains 30000 words and complete the longest matching based on syllables. The experimental results are as follows.

By extended dictionary we are correcting this type of error. Different results are shown in Table 2.

Table 2. Word segmentation results in different dictionary size

The number of words in the dictionary	Number of word segmentation	Accuracy of word segmentation
---------------------------------------	-----------------------------	-------------------------------

15768	30000	82.63%
31719	30000	86.09%

3.2 The influence of error dictionary on word segmentation

In this experiment, we segment the syllable of the Lao language text which contains 30000 words by the extended dictionary. Compare the results of the word segmentation with error dictionary and the results of the word segmentation without the error dictionary. The experimental results are as follows.

The influence of error dictionary on word segmentation is shown in Table 3.

Table 3.The influence of error dictionary on word segmentation

Word segmentation method	Number of word segmentation	Accuracy of word segmentation
without error dictionary	30000	83.03%
with error dictionary	30000	86.09%

3.3 The influence of rule correction on word segmentation

In this experiment, we segment the syllable of the Lao language text which contains 30000 words by the extended dictionary. Compare the results of the word segmentation with rule correction and the results of the word segmentation without rule correction. The experimental results are as follows.

The influence of rule correction on word segmentation is shown in Table 4.

Table 4. The influence of rule correction on word segmentation

Word segmentation method	Number of word segmentation	Accuracy of word segmentation
Word segmentation without rule correction	30000	82.66%
Word segmentation with rule correction	30000	86.09

4. Conclusions

In order to realize the Lao word segmentation, this paper combines the characteristics of Lao syllables. Using the structure rules of syllables, segment the syllable of the Lao language text and complete the maximum matching word segmentation. Finally, using the error dictionary and the rules improve the accuracy of the results through two-time correction. The experimental results show that the coverage of the dictionary directly affects the accuracy of word segmentation, the larger the scale, the higher the accuracy of the word segmentation. Through the error dictionary, correct of some wrong word segmentation effectively. At the same time, the error of some numbers, English letters, foreign words and so on are corrected by the rule correction, which can effectively optimize the segmentation results.

Foundation item: China National Natural Science Foundation (61562049), (61662040).

References

1. Quan-Shu Long, Zheng-Wen Zhao, Hua Tang. "Overview on Chinese Segmentation Algorithm". COMPUTER KNOWLEDGE AND TECHNOLOGY, 2009,5 (10) :2605-2607.
2. Chang-Ning Huang. "The problem of Chinese word segmentation information processing". Applied Linguistics, 1997,6(1): 72-78.

3. De-Quan Zheng, Feng Yu, Kai-Tao Wang. "Ambiguity Word Segmentation Based on Two Chinese Characters Used as a Word in Chinese" *Computer Engineering and Applications*, 2003,39 (1): 17-19.
4. Qiao-Ming Zhu, Tao Wen, Pei-Feng Li, and Pei-De Qian. "Self-Adaptive Chinese Ambiguous Word Segmentation Method Based on Multi-Gram Library". *MINI-MICRO SYSTEMS*, 2006,27(8):1597-1600.
5. P.P issamay, et al .Syllabification of Lao Script for Line Breaking.Tech. Rep. of STEA, Lao PDR, 2004.