Dynamic Programming Method Applied in Vietnamese Word Segmentation Based on Mutual Information among Syllables

Nguyen Thi Uyen
IT Faculty - Vinh University

Tran Xuan Sang
IT Faculty - Vinh University

***Abstract*—Vietnamese word segmentation is an important step in Vietnamese natural language processing such as text categorization, text summary, and automated machine translation. The problem with Vietnamese word segmentation is complicated because Vietnamese words are not always separated by a space. One word can include one or more syllables depending on the context. This paper proposes a method for Vietnamese word segmentation based on the mutual information among the syllables combined with dynamic programming. With this method, we can achieve an accuracy rate of about 90% with a raw text corpus.**

***Keywords*—Vietnamese word segmentation; dynamic programming; mutual information; Vietnamese syllables**

I. INTRODUCTION

Word segmentation is the process to determine the boundaries between words in sentences. Words in the Vietnamese language are not always separated by blank spaces. A word may contain several syllables. The syllables are combined to form different words depending on the context of the text. Therefore, it is difficult to solve this problem automatically. Example 1: The sentence written in Vietnamese "Học sinh học sinh học" - in English "The pupils study biology". This sentence is composed of two Vietnamese syllables "học ~ study" and "sinh ~ biology" which form different words in the sentence. The correct solution should be "học sinh | học | sinh học" ~"The pupils | study | biology". One of the most difficult tasks in Vietnamese word segmentation is to determine the ambiguities of the sentence.

The same sentence may have different word segmentation solutions if it is in a different context. Example 2: The sentence written in Vietnamese "Ông già đi nhanh quá" may have two different meanings. One is "The old man goes too fast", the other one is "Grandfather gets old too fast". It results in two word segmentation solutions: Ông già| đi |nhanh quá and Ông| già đi| nhanh quá. In this case, it is needed to consider the context of this sentence in order to select the best solution.

Mutual information (MI) between the syllables presents the correlation of syllables to be combined as a word. The greater MI value will show the higher probability of words combination of syllables. The MI theory will be presented in more details in section 3.a.

The dynamic programming technique is used to reduce the complexity of the computation. This method will be presented in section 3.b.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 describes the proposed method. Section 4 provides the experimental results. Finally, section 5 summarizes the work of this paper.

II. RELATED WORKS

This section presents the previous works in Vietnamese word segmentation.

A. Maximum Matching Method

Maximum matching algorithm is commonly used to word segmentation problem. The idea of this method is to start at the first syllable in a text and attempt to find the longest word starting with that syllabus in the dictionary. If a word is found, the maximum matching algorithm marks a boundary at the end of the longest word, then begins the same longest match search starting at the syllable following the match. Whereas, that syllable is segmented as a word, and begins the search starting at the next syllable [Wong et al, 1996] [1]. Dinh et al, 2001 used Maximum Matching method to segment Vietnamese word [2]. However, the accuracy of word segmentation is not high.

B. Transition Graph Method

In this method, each syllable is represented by a vertex. The edge represents weight of connection between two syllables which is calculated based on the data training process. The transition graph will show the probability among syllables to form the words in a specific text. Nguyen et al, 2003; Pham et al, 2009 used this method to segment Vietnamese word [3,4].

C. Support Vector Machine Method(SVM)

Point-wise machine learning method (SVM) is used to mark two kinds of symbols: space (word segment symbols) and underscore (linking two syllables symbol) (Luu et al, 2012) [5]. There are three basic features in point-wise methods: n-grams of syllables, n-gram types of syllables, and featured dictionary. Vietnamese language has about 70% of the words with 2 syllables, and 14% words with 3 syllables, therefore the point-wise window was set as w = 3. The author defines four types of syllables: uppercase syllables (U): Vietnamese syllables begin with capital letters. lower syllables (L): the Vietnamese syllable contains only lowercase letters. Numerical syllables (N) only consists of the digits. The other type (O): the syllables belongs to a foreign language. This research achieved 98.2% accuracy rate. However, this method needs a good featured dictionary.

D. Combination Method

Le et al, 2008 [6] combines a finite state machine, canonical form analysis, maximum matching method to segment Vietnamese word. Minimal finite-state automaton is used to present the Vietnamese lexicon. A text to be tokenized is first parsed into lexical phrases and other patterns using pre-defined regular expressions. The automaton is then deployed to build linear graphs corresponding to the phrases to be segmented. A tool named vnTokenizer is then created to show the effectiveness of this method.

III. PROPOSED METHOD

For Vietnamese, there is a lack of large lexicographic resources, and annotated corpora are also rare, therefore we will develop a method that only rely on raw corpus. The mutual information (MI) is a statistical score which helps to segment the words. The mutual information is calculated based on the frequency of the syllables in a raw corpus. The main ideal of our method is to maximize the MI-score of the chunk, using different segmentations. We will calculate all the possibilities of segmentations for a given sentence. Which possibility has the highest MI-score, becomes the final solution. There are some difficulties in the calculation. First, the segmentation possibility is an exponential function of the length of the sentence. A long sentence will cause a large number of ways of segmentation. Other problems consist of the difficulty of calculating MI-score, and sparse data. In order to overcome those difficulties, the dynamic programming is utilized. In following sub-section, MI-score and dynamic program applied in word segmentation are presented in detail.

A. Corpus and MI-Score

We build a corpus by collecting text from many Vietnamese websites and online news papers. Our raw corpus contains about 41 million syllables.

Mutual information is an important factor to identify the correlation between syllables in a corpus. The equation to calculate MI value is presented as below (Ong & Chen, 1999).

$$MI(cw) = \frac{p(cw)}{p(lw) + p(rw) - p(cw)} \qquad (1)$$

Where:

- *cw* is a chunk containing n syllables. $\mathrm{cw} = \mathrm{c_1 c_2 \dots c_n}$.

- *lw* is a chunk containing n-1 syllables $\mathrm{lw} = \mathrm{c_1 c_2 \dots c_{n-1}}$.

- *rw* is a chunk containing n-1 syllables $\mathrm{rw} = \mathrm{c_2 c_3 \dots c_n}$.

The higher *MI(cw)* value shows a higher probability of *lw* and *rw* appearing in the corpus. It means *cw* has high probability to be a compound word.

Based on equation 1, we elaborate the way to calculate MI-score for certain segmentation.

Given a sentence $\mathrm{C} = \mathrm{c_1 c_2 \dots c_n}$ with $\mathrm{c_i}$ is a syllable.

- *N*: total number of syllables in the corpus.
- $\mathrm{f(w)}$: frequency of chunk *w* in the corpus.
- $\mathrm{p(w)}$:Probability of chunk *w* in the corpus.

$$p(w) = \frac{f(w) + 1}{N} \qquad (2)$$

- $\mathrm{MI(c_1 c_2)}$ *:* Mutual Information value of two syllables $\mathrm{c_1, c_2}$.

$$\mathrm{MI}(c_1 c_2) = \frac{p(c_1 c_2)}{p(c_1) + p(c_2) - p(c_1 c_2)} \qquad (3)$$

- $\mathrm{MI(c_1 c_2 \dots c_n)}$: Mutual Information value of n syllables $\mathrm{c_1, c_2, \dots, c_n}$.

$$MI(c_1 c_2 \dots c_n) = \frac{p(c_1 c_2 \dots c_n)}{p(c_1 c_2 \dots c_{n-1}) + p(c_2, c_3 \dots c_n) - p(c_1 c_2 \dots c_n)} \qquad (4)$$

Given a sentence with certain segmentation as belows.

$$t(a) = w_1 | w_2 | \dots | w_m$$

- Then, the MI-score of this segmentation is calculated as below.

$$\mu(t(a)) = MI(w_1) + MI(w_2) + \dots + MI(w_m) \qquad (5)$$

B. Dynamic Programming

A given sentence consists of n syllables $\mathrm{C} = \mathrm{c_1 c_2 \dots c_n}$

Normally, the longest Vietnamese word contains four syllables. The dynamic programming method is described in the following steps:

Step 1: Separate sentence C into combinations of one, two, three and four syllables.

Step 2: Calculate MI value for each combination of syllables.

Step 3: Calculate MI-score of final solution by following sub-steps:

1) Assume x is a chunk of syllables. f(x)=-100 if x is not in dictionary; whereas f(x)=MI(x). We select a value of -100 or lower in order to eliminate the solution that contains word outside dictionary.

2) Then calculate highest MI-score D[n] of final solution as following:

$\mathrm{D[0] = 0}$

$\mathrm{D[1] = f(c_1)}$

$\mathrm{D[2] = \max\{ D[1] + f(c_2), D[0] + f(c_1 c_2)\}}$

$\mathrm{D[3] = \max\{ D[2] + f(c_3), D[1] + f(c_2 c_3), D[0 + f(c_1 c_2 c_3)\}}$

$\mathrm{D[j] = max\ \{D[j\text{-}1] + f(c_j), D[j\text{-}2] + f(c_{j\text{-}1}c_j), D[j\text{-}3] + f(c_{j\text{-}2}c_{j\text{-}1}c_j),D[j\text{-}4] + f(c_{j\text{-}3}c_{\ j\text{-}2}c_{j\text{-}1}c_j)\}}$

With j = 4, 5, ..., n.

Step 4: After computing MI-score, the final segmentation is found by following sub-steps.

1) Set $K[j] = t$
and $\mathrm{D[j] = max\ \{D[j\text{-}1] + f(c_j), D[j\text{-}2] + f(c_{j\text{-}1}c_j), D[j\text{-}3] + f(c_{j\text{-}2}c_{j\text{-}1}c_j),D[j\text{-}4] + f(c_{j\text{-}3}c_{j\text{-}2}c_{j\text{-}1}c_j)\}= D[j\text{-}t] + f(c_{j\text{-}t+1}\dots c_j)}$

Where j-t is an index which maximizes the MI-score; and t value shows the best separated points in chunk of $(c_1c_2 \dots c_j)$

$$c_1c_2 \dots c_{j-t} \mid c_{j-t+1} \dots c_j \sim c_1c_2 \dots c_{j-K[j]} \mid c_{j-K[j]+1} \dots c_j$$

The segmentation solution is $c_1c_2 \dots c_{n-K[n]} | c_{n-K[n]+1} \dots c_n$

2) Set j = n-K[n]. The next segmentation is $c_1c_2 \dots c_{j-K[j]} \mid c_{j-K[j]+1} \dots c_j$

Step 5: Loop step 4 until reach the first syllable in the sentence.

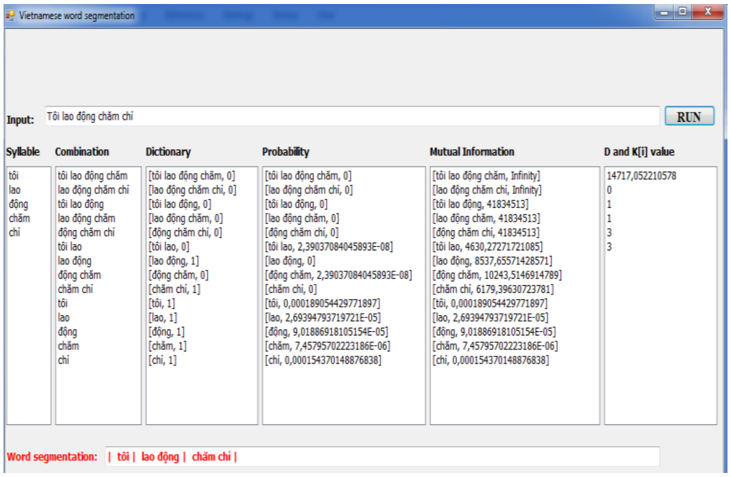
The actual calculation using dynamic programming method is following:

```
for (int i = 4; i <= n; i++)
{
   t1 = tudon[i - 2] + " " + tudon[i - 1];
   t2 = tudon[i - 3] + " " + tudon[i - 2] + " " + tudon[i-1];
   t3 = tudon[i - 4] + " " + tudon[i - 3] + " " + tudon[i - 2] +
" " + tudon[i-1];
   D[i] = max4(D[i - 1] + f[tudon[i-1]], D[i - 2] + f[t1], D[i -
3] + f[t2], D[i - 4] + f[t3]);
   maxMI = D[i];
   if (maxMI == (D[i - 1] + f[tudon[i - 1]]))
     {
        K[i] = i - 1;
      }
      else
            {
               if (maxMI == (D[i - 2] + f[t1]))
               {
                  K[i] = i - 2;
               }
               else
               {
                  if (maxMI == (D[i - 3] + f[t2]))
                  {
                     K[i] = i - 3;
                  }
                  else
                  {
                     K[i] = i - 4;
                  }
               }
            }
        }
```

This method will be demonstrated by an example shown below:

Given a sentence in Vietnamese: *tôi lao động chăm chỉ*. (I word hard). The combinations of syllable: one-syllable- (tôi), (đi), (học), (chăm), (chỉ); two-syllables - (tôi lao), (lao động), (động chăm), (chăm chỉ); three-syllables - (tôi lao động), (lao động chăm), (động chăm chỉ); four-syllables - (tôi lao động chăm), (lao động chăm chỉ).

Using the proposed method, we can compute the highest MI-score and then get the separated points to segment the sentence. Figure 1 shows the programming results:



```
Input: Tôi lao động chăm chỉ

Syllable:
tôi
lao
động
chăm
chỉ

Combination:
tôi lao động chăm
lao động chăm chỉ
tôi lao động
lao động chăm
động chăm chỉ
tôi lao
lao động
động chăm
chăm chỉ
tôi
lao
động
chăm
chỉ

Dictionary:
[tôi lao động chăm, 0]
[lao động chăm chỉ, 0]
[tôi lao động, 0]
[lao động chăm, 0]
[động chăm chỉ, 0]
[tôi lao, 0]
[lao động, 1]
[động chăm, 0]
[chăm chỉ, 1]
[tôi, 1]
[lao, 1]
[động, 1]
[chăm, 1]
[chỉ, 1]

Probability:
[tôi lao động chăm, 0]
[lao động chăm chỉ, 0]
[tôi lao động, 0]
[lao động chăm, 0]
[động chăm chỉ, 0]
[tôi lao, 2,39037084045893E-08]
[lao động, 0]
[động chăm, 2,39037084045893E-08]
[chăm chỉ, 0]
[tôi, 0,000189054429771897]
[lao, 2,69394793719721E-05]
[động, 9,01886918105154E-05]
[chăm, 7,45795702223186E-06]
[chỉ, 0,000154370148876838]

Mutual Information:
[tôi lao động chăm, Infinity]
[lao động chăm chỉ, Infinity]
[tôi lao động, 41834513]
[lao động chăm, 41834513]
[động chăm chỉ, 41834513]
[tôi lao, 4630,27271721085]
[lao động, 8537,65571428571]
[động chăm, 10243,5146914789]
[chăm chỉ, 6179,39630723781]
[tôi, 0,000189054429771897]
[lao, 2,69394793719721E-05]
[động, 9,01886918105154E-05]
[chăm, 7,45795702223186E-06]
[chỉ, 0,000154370148876838]

D and K[i] value:
14717,052210578
0
1
1
3
3

Word segmentation:  | tôi |  lao động |  chăm chỉ |
```

Fig. 1. Experiment with Vietnamese Sentence.

IV. EXPERIMENTAL RESULTS

We extracted randomly 100 sentences from the corpus and asked 20 native Vietnamese speakers to make the word segmentation manually. For each sentence, we choose a solution which is selected by the most native speakers. The evaluation process is then taken by computing the rate as follows:

$$R= A/B$$

where:

A. Number of correct words which segmented by proposed method

B. Total number of words segmented by native speakers.

The rate is about 90%. This result is not very disappointing because we used only the raw corpus without lexicon nor annotation. We did not extend our experiment because we realized that we were not able to get better results with this method alone. However, the results show that the method works.

V. CONCLUSION

The proposed method has produced a promising results in the case of using un-annotated corpus for word segmentation. The mutual information is a key value to select the final segmented solution. The dynamic programming method is proposed to reduce the complexity of the problem. The advantage of our proposed method is that we do not need an annotation corpus. Therefore, the arbitrary text on the internet can be used as a corpus for natural language processing.

REFERENCES

[1] Pak-kwong Wong, Chorkin Chan,"Chinese word segmentation based on maximum matching and word binding force". COLING '96 Proceedings of the 16th conference on Computational linguistics - Volume 1Pages 200-203, 1996.

[2] Dinh Dien, Hoang Kiem, Nguyen Van Toan., "Vietnamese Word Segmentation", The sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, pp. 749 -756, 2001.

[3] Pham DD., Tran GB., Pham SB., "A hybrid approach to Vietnamese word segmentation using part of speech tags", International Conference on Knowledge, 2009.

[4] Nguyen, P.T., Nguyen, V.V., Le, A.C., "Vietnamese word segmentation using hidden markov model", International Workshop for Computer, Information, and Communication Technologies on State of the Art and Future Trends of Information techonologies in Korea and Vietnam, 2003.

[5] Luu, T.A, Yamamoto, K., "A pointwise approach for Vietnamese Diacritics Restoration", IALP 2012.

[6] Le, H.P, Nguyen, T.M.H, Azim Roussanaly, Ho, T.V, "A hybrid approach to Word Segmentaion of Vietnamese texts", Language and automata theory and applications 2nd international coference, LATA 2008