An Introduction to some Vietnamese Word Segmentation Approaches

Nguyen Thac Thong, Nguyen Van Giap, Nguyen Tung Lam, and Le Van Giap

University of Engineering and Technology,
Vietnam National University, Hanoi
{thongnt_57,giapnv_570,lamnt_57,giaplv_57}@vnu.edu.vn

Abstract. The purpose of this short document is to provide a brief overview of some approaches in Vietnamese Word Segmentation. Furthermore, we do some experiment to compare these approaches and also try to implement a tool called *UETWordSeg*. The tool implements a hybrid approach based on Maximum matching algorithm combining with Word-based Language Modeling to resolve the ambiguity. In the best case *UETWordSeg* can get 97.39% in *F1 Score* on the test set of VSLP project.

1 Introduction

In linguistics, a word is the smallest element that may used with a literal or practical meaning. This is different from syllables, which is the smallest unit of meaning but not necessarily if they stand alone. A word can consists of a single syllables (for example tôi (me), hoa (flower)) or several (for example quan trọng (important), vai trò (role)). Words can be put together to construct a lager elements of language such as phrases, sentences, paragraphs.

Word Segmentation is the process of dividing a sequence of syllables into linguistically meaningful words. In many languages using the Latin alphabet like English, the space is good to be the divider between words. However, there are some languages such as Chinese, Japanese, Korean and Vietnamese which space is not always the word delimiter. In fact, words may contain some syllables that are separated by space.

Word Segmentation is consider as of one the first important tasks in Natural Language Processing (NLP) in Vietnamese. Many other NLP tasks need Word Segmentation as a fundamental preprocessing step. Therefore, the result of the Word Segmentation process contributes an important part in other NLP systems.

Vietnamese Text Retrieval system (Vietnamese IR) benefits from a dictionary-based word segmentation for indexing terms (syllables, words, compound words, combination of word and compound word(Ho. 2007). Also, Word Segmentation is used to improve ability to detect malicious domains relative to approaches without segmentation, as measured by misclassification rates and areas under the ROC curve. [7]

Word Segmentation is a difficult task in Vietnamese because of not only the quantity and quality of training data and lacking of available Vietnamese corpus with acceptable diversity but also the linguistical characteristics in term of ambiguity. There are some types of ambiguity in Vietnamese such as *combine ambiguity*, *overlap ambiguity* and *meaning ambiguity*. Formally, a phrase can be formed as a sequence of syllables as $s_1s_2s_3$. As an example of *combine ambiguity*, s_1s_2, s_1, s_2 are also considered as a word in dictionary but the segmentation $(s_1)(s_2)(s_3)$ are better than $(s_1s_2)(s_3)$ (for example: with the phrase "tính từ mép" ("measure from the edge"), "tính" (measure), "từ" (from) and "tính_từ" (adjective) are also words but only the segmentation "tính từ mép" is correct and has the meaning "measure from the edge", phrase "tính_từ mép" has no meaning in this context). Other ambiguity type was mentioned above is *overlapping*. It means a phrase $s_1s_2s_3$ can be segmented into $(s_1s_2)(s_3)$ or $s_1(s_2s_3)$, both of them may be correct depending on the context (for example: phrase "trên cơ sở" have two ways to segment as "trên_cơ sở" or "trên cơ_sở"). The last type is *meaning ambiguity*, it means there are some segmentations for the phrase and all of them are correct and can be understood clearly (for example: "ông già đi nhanh quá" has two segmentations as "ông_già đi nhanh quá" (the old man goes too fast) or "ông già đi nhanh quá" (my grandfather is aging too fast) and both of them are correct and clear to understand).

In recent years, Word Segmentation has been attracting many researchers not only in Vietnam but also some other countries like Japan, China, and Korean. As a result, some approaches has been studied and proposed recently. They can be classified into three main categories: dictionary-based, statistical, and hybrid approach. This report aims to help readers to have an overview to aforementioned approaches and implement an simple but effective hybrid approach to solve Word Segmentation task. We combined the dictionary-based approach that used Vietlex[1] dictionary, regular expression and maximal matching technique to get all possible segmentations then use Language Model to resolve the ambiguities.

The remain part of this report is organized as follows. In the second section, we will discuss about the structure of a Vietnamese Word and the distribution in length of words in Vietnamese. The next section describe quickly the foundation theories to understand approaches that are discussed in section 4, in which we briefly review the representative approaches in the previous studies in term of three mentioned categories. Section 5 describe our experiment setup, results and related discussions. Finally, in the section 6 and 7, we conclude the report and raise some improvements for future works.

2 Vietnamese Words

To solve the Word Segmentation task in Vietnamese, we firstly study about the formulation of a word in Vietnamese linguistically.

[1] `http://www.vietlex.com/kho-ngu-lieu`

2.1 Word formulation

The basic unit to constitute a Vietnamese word is syllable (in some language syllable is also known as *morpheme-syllable*). The Vietnamese syllable follow the following schema[2]:

$$(C_1)(w)V(G|C_2) + T$$

where:

C_1 = initial consonant onset
w = bilabial on-glide /w/
V = vowel nucleus
G = off-glide coda (/j/ or /w/)
C_2 = final consonant coda
T = tone

In other words, a syllable can optionally have one onset consisting of single consonant or a consonant and the glide /w/ and an optional coda. The vowel nucleus may have an additional glide element. More explicitly, the syllable types are as follows:

Bảng 1. Example of Vietnamese syllable schema

Syllable	**Example**	**Syllable**	**Example**
V	ê (eh)	wV	uể (sluggish)
VC	ám (possess (by ghosts,.etc))	wVC	oán (bear a grudge)
VC	ớt (capsicum)	wVC	oắt (little imp)
CV	nữ (female)	CwV	huỷ (cancel)
CVC	cơm (rice)	CwVC	toán (math)
CVC	tức (angry)	CwVC	hoặc (or)

A word in Vietnamese can contain one or multiple syllables. Words in Vietnamese can be divided into some categories[3]:

1. *Singular*: Many syllables have meaning itself. They refer to an object or a concept such as cây (trees), trời (sky), cỏ (grass), nước (water) etc ... Therefore, a *single word* contains only one syllable.
2. *Compound*: A *compound word* grafts multiple syllables, which are related in meaning with each other, into a word. Based on the the relationship in meaning between the constituents, we can classify Vietnamese compounds as follows:
   - *Compound type I*: These are the words that the constituents have equal relations of meaning. Compound word type I indicates generalized meaning. For example: ăn ở (accommodation), ăn nói (speak), etc ...

[2] `https://en.wikipedia.org/wiki/Vietnamese_phonology#Syllables_and_phonotactics`
[3] `http://ngonngu.net/index.php?p=207`

- *Compound type II*: The compounds which have a structural element depends on others. The secondary parts help to classify the main one. Example: tàu hỏa (train), đường sắt (railway), nông sản (agricultural products), etc ...

3. *Reduplication*: The reduplicative word have at least two syllables with phonic components rhyming repeated. They are usually using in literal text to describe properties, colors or sounds, etc .... For example: xanh xao (haggard), thì thầm (whispering), etc ...

2.2 Distribution in length

According to Le. (2008) [4], based on the Vietlex corpus which contains 40.181 words, there are about 15.69% words in Vietnamese having only one syllables, which is called single word. 2-syllable compounds contributes the most with 70.72%. Compounds that contains three or fours syllables occupy 5,62% and 6.93% respectively. There are only 1.04% of compounds having more than four syllables. We can take a look at these statistics at the following table:

Table 2. Distribution in length of Vietnamese words

Length	#	%
1	6.303	15.69
2	28.416	70.72
3	2.259	5.62
4	2.784	6.93
$\geq$5	419	1.04
Total	40.181	100

3 Background knowledge

4 Related works

Due to the important of the Word Segmentation for higher task in Natural Language Processing in Vietnamese, there are many approaches proposed to solve it. We can wrap them into three main categories: dictionary-based, statistical-based and hybrid-based approach [4] [3]. We can demonstrate categories of Word Segmentation approaches as the following:

- *Dictionary-based*: Maximum matching (MM), Longest Matching (LM) [6] [1]
- *Statistic-based*: Maximum Entropy (ME), Hidden Markov Model (HMM), N-Gram Language Model, Conditional Random Field (CRFs) [2], Support Vector Machine (SVM) [3], ...

– *Hybrid-based*: MM & N-Gram Language Model [4], dictionary-based & CRFs, dictionary-based & SVM, ...

In this report, we will introduce a representative implementation for each aforementioned approaches.

4.1 Dictionary-based approach

Maximum matching (MM) and Longest matching (LM) are two most popular techniques because of its simplicity and effectiveness. However, most of the dictionary-based approaches fail to solve ambiguity and also depend on the quality of dictionary.

Longest Matching The longest matching method scan a sequence of syllables (phrase, sentence) from left to right to match words given in the dictionary. This method have two main drawbacks. First and foremost, it totally depends on the accuracy and sufficient of dictionary. In addition, it can not handle ambiguous cases because it just consider the left-most segmentations, for example:

– **Input:** dựa trên cơ sở toán học
– **LM Output:** dựa trên_cơ sở toán_học
– **Correct Output:** dựa trên cơ_sở toán_học

Maximum Matching Maximum matching generates all possible segmentations and select the segmentation that contains fewest words. If we represent sequence of syllables as a graph, the method is same as a shortest path finding problem, which is each possible segmentations correspond to a path in the graph. To generate a segmentation, MM firstly scan for the longest syllable sequence that start from current position, then check whether it matches the pattern (dictionary, regular expression) or not. After that, it moves the pointer to start to scan the next one. Although this method is more likely correct with long words than short ones, it cannot handle sequences that have the same number of words (have the same shortest path). For example:

– **Input:** dựa trên cơ sở toán học
– **MM Output 1:** dựa trên_cơ sở toán_học
– **MM Output 2:** dựa trên cơ_sở toán_học
– **Correct Output:** dựa trên cơ_sở toán_học
– MM approach will confuse to select the best one because both output have the same number of words.

4.2 Statistical-based approach

We select the method that are described in [3] to introduce about statistical-based approach. The approach use CRFs and SVM and get a promising result with a F-measure of 94.23%.

Problem Representation We denote that syllable that begins a word is marked with B_W (Begin of a Word), syllable that is inside a word is marked with I_W (Inside of a Word), other things like comma, dot marks are tagged with O (Outside of a word). Therefore, the problem of detecting word boundaries in a sentence is modeled as the problem of labeling syllables in that sentence with three above labels. The approach use CRFs and SVM to resolve that problem.

Features selection There are two types of features: static features (per-sate) and dynamic features (edge feature). Static features are very similar to per-state features in CRFs model in the sense that the model also take into account context predicates at the current observation. SVM model decide dynamic features in tagging process by considering the two previous labels. As a result, five types of context predicate templates from which various features will be generated correspondingly. The table below demonstrates these features:

Bảng 3. Context predicate templates for CRFs and SVM

Syllable Conj. (SC)	Syllable_Conjunction (-2,2)
Dictionary (Dict)	In_LacViet_Dictionary (-2,2)
External Resources (ERS.)	In_Personal_Name_List(0,0), In_Family_Name_List(0,0), In_Middle_Name_List(-2, 2), In_Location_List(-2,2)
Miscellaneous (Misc)	Is_Regular_Expression(0,0), Is_Initial_Capitalization(0,0), Is_All_Capitalization(0,0), Is_First_Observation(0,0), Is_Marks(0,0)
Vietnamese Syllable Detection (VSD)	Is_Valid_Vietnamese_Syllable(0,0)

Two numbers inside the brackets next to each context predicate template indicate the window surrounding current position in which we explore context information. For example, In_LacViet_Dictionary(-2, 2) means we make a particular conjunction of adjacent syllables in the sliding window from the two previous to the next two syllables and check whether that conjunction forms is in the dictionary or not.

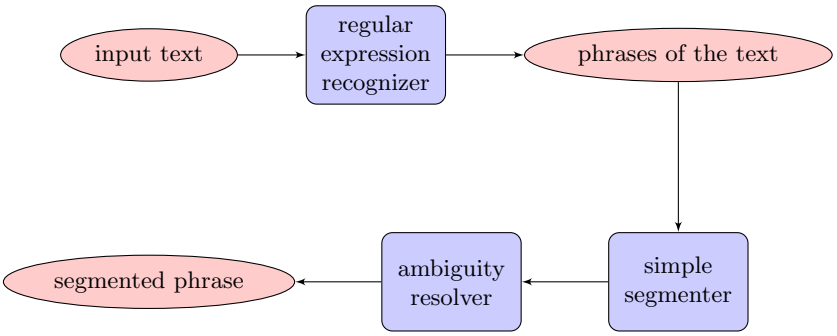
The number of more-than-4-syllable words is quite small, we take into account that only conjunction of up-to-three adjacent syllables. similarly, with Syllable_Conjunction(-2, 2), we consider all the 1-gram, 2-gram, 3-gram of syllables in the window size of 5.

4.3 Hybrid-based approach

We use *vnTokenizer* and [4] to introduce about the hybrid-base approach which is the combination of Maximum Matching and N-gram Language Model. Le.

(2008) represents Vietnamese lexicon as a Minimal deterministic finite state automata (MDFA), which is the best representation of a lexicon in term of memory usage and access time. In that model, the minimal automaton that accepts the Vietnamese lexicon contains 42.672 states in which 5.112 states are final ones. It has 76.249 transitions; the maximum number of outgoing transitions from a state is 85, and the maximum number of incoming transitions to a state is 4.615. The approach get the best result with a F-measure of 95.6%.The following graph generally demonstrates the flow of vnTokenizer:

Fig. 1. Flow chart of *vnTokenizer*



An input text for segmentation is first analyzed by a regular expression recognizer using a greedy strategy (longest matched pattern is taken out) to detect regular patterns such as proper names, common abbreviations, numbers, dates, times, email addresses, URLs, punctuations, etc. After that, the recognizer extracts phrases of the text. If a pattern is a phrase, that is a sequence of syllables and spaces, it is passed to a simple segmenter which implements Maximum Matching algorithm to detect the word composition. The segmenter determines the longest syllable sequence which starts at the current position and is listed in the lexicon. Finally, *vnTokenizer* use a N-gram Language Model with Linear Interpolation Smoothing technique to resolve ambiguous cases.

Problem Representation We consider a phrase as $s = s_1 s_2 ... s_n$. A overlap ambiguity is three consecutive syllables $s_i s_{i+1} s_{i+2}$ in which both of the two segmentations $(s_i s_{i+1}) s_{i+2}$ and $s_i (s_{i+1} s_{i+2})$ may be correct, depending on context. s is represented by $G = (V, E)$, $V = v_0, v_1, \ldots, v_n, v_n + 1$. There is an arc (v_i, v_j) if the consecutive syllables $s_{i+1}, s_{i+2}, \ldots, s_j$ compose a word $(i < j)$. Vertex v_0 and v_{n+1} are respectively the start and end vertexes, n vertexes $v_1, v_2, \ldots, v_n$ are aligned to n syllables of the phrase. The problem turn to be a shortest path finding problem. We can then propose all segmentations of the phrase by listing all shortest paths on the graph from the start vertex to the end vertex.

Resolution of Ambiguities To resolve ambiguity, we uses a bigram language model which is augmented by the linear interpolation smoothing technique. The segmentation with the greatest probability will be chosen.

5 Experiments

The experiment was conducted using hybrid approach. In 2008, Hong-Phuong Le proposed and implemented a hybrid approach combining both finite-state automata technique, regular expression parsing and maximal-matching strategy using language model to resolve ambiguity. Our experiment was conducted with the base on this approach and contributes some small changes in order to improve the speed and increase a small amount of word-segmenting accuracy.

5.1 Corpus building

Corpus for language modeling: Our experiment corpus contains 2677 files with totally 77.000 sentences that have been manually spell-checked and segmented by linguistics from Vietnam Lexicography Center (Vietlex). Each word in the corpus was constructed from one or more syllables combined together by "_" character. From the original data, we have extracted all pure text and removed HTML tags.

After revising the data, we have been detecting some sentences which is segmented not correctly. However, in order to ensure that the experiment is conducted fairly in comparison to other tools, we still kept in the data.

Vietnamese dictionary: The dictionary for word matching in the experiment was provided by Vietlex. This dictionary contains totally 35 Vietnamese words. However, after removing duplicated words, there remains 31.158 entities.

Other dictionaries: Besides, we also using some other dictionaries which are manually collected and built by members in our group. Some dictionaries are location, Vietnamese person name, Vietnamese named entity prefix.

Regex set: The experiment used the regex set from vnTokenizer after updating and removing some unnecessary regexes.

5.2 Experiment setup

In this experiment, we re-implements a hybrid approach to automatically tokenize Vietnamese text. The overall flow of experiments can be depicted as Figure 2.

A text to be tokenized is first parsed into lexical phrases and other patterns using pre-defined regular expressions. The dictionary is then used to look up words and build linear graphs corresponding to the phrases to be segmented. See Figure 3.

The weight of each edges in the graph can be define as three following options:

- **Maximal Matching:** $weight(edge) = 1/length(edge)$. See figure 4

Fig. 2. Word segmenting processes

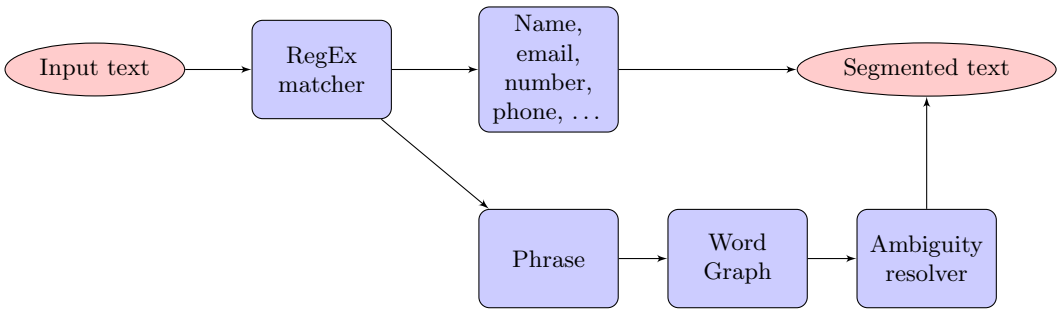


Fig. 3. Segment Graph

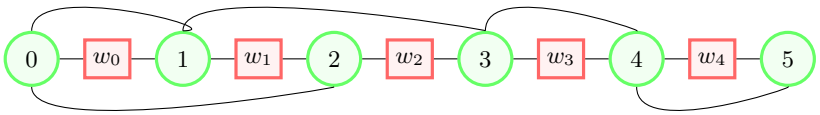


Fig. 4. Edge's Weight in Maximal Matching

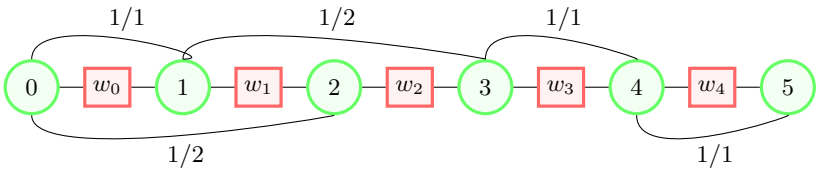
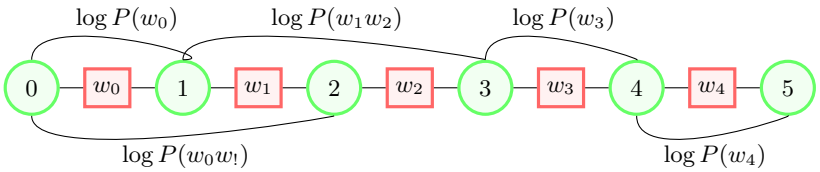


Fig. 5. Edge's Weight in Language Model



– **Language Modeling:** $weight(edge) = log(P(edge))$ applying for uni-gram language modeling. (n-grams models which n is greater than 1 need evaluate probability of each edge with respect to its surrounding context). See figure 5.

Apparently, the best segmentations of the text is corresponding to the shortest path from the first to the last node of the graph. Previously, Hong-Phuong Le has implemented Dijktra and Depth First Search to solve shortest path problems. Therefore, finding shortest path in that way will require a lot of running time because of high complexity of algorithms, $O(|E| + |V|log|V|)$ for Dijktra and $O(|E|)$ for DFS. In this report, we will propose a dynamic algorithm to find the best segmentation.

6 Conclusion and Future Works

The report introduce three main categories of approaches to solve Word Segmentation task. In each category, we describe a representative method to briefly understand the approaches. Furthermore, we also implement a hybrid approach containing Maximum Matching and Word-Based Language Model. The experimental show a promising result with F_1 score as 97.39%.

One of the limitations is the quality and quantity of corpus for training and testing. In the future, we would like to extend our corpus and try some enhancement in resolving ambiguous cases to improve the accuracy of our approach.

7 Acknowledgement

We would like to add a few words of appreciation for the people who have been a part of this project right from its inception and have helped us during the NLP course. We express our deepest thanks to Dr. Le Anh Cuong, who have given necessary suggestions and guidance, which were extremely valuable for our project. This project cannot be completed without his support.

References

1. S., C. Shi, W. Gale, and N. Chang: A stochastic finite-state word-segmentation algorithm for chinese. *Computational Linguistics* (1996)
2. F. Peng, F. Feng, A. McCallum: Chinese Segmentation and New Word Detection using Conditional Random Fields. *In Proceedings of COLING, pp. 562-568* (2004)
3. C. T. Nguyen, T. K. Nguyen , X. H. Phan, L. M. Nguyen, and Q. T. Ha: Vietnamese Word Segmentation with CRFs and SVMs: An Investigation. *The 20th Pacific Asia Conference on Language, Information and Computation : Proceedings of the Conference* (2006)
4. H. P. Lê, T. M. H. Nguyen, A. Roussanaly and T. V. Ho.: A hybrid approach to word segmentation of Vietnamese texts. *2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain* (2008)

5. D. D. Pham, G. B. Tran, S. B. Pham: A Hybrid Approach to Vietnamese Word Segmentation using Part of Speech tags. *Knowledge and Systems Engineering, 2009. KSE '09. International Conference* (2010)
6. M. Sassano: Deterministic Word Segmentation Using Maximum Matching with Fully Lexicalized Rules, *In Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), volume 2: Short Papers, pp. 79-83* (2014)
7. W. Wang, K. Shirley: Breaking Bad: Detecting malicious domains using word segmentation. *IEEE Web 2.0 Security and Privacy Workshop* (2015)