Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 06, June - 2025

Proposing a Solution to Separate Ede Words Based on the Ede Vocabulary

Le Hoang Thi My
University of Technology and Education
The University of Danang
Danang, Vietnam

***Abstract*— Word segmentation is a processing process aimed at determining the boundaries of words in a sentence. Words can also be single words, compound words, etc. In natural language processing, in order to determine the grammatical structure of a sentence, the word class of a word in a sentence, the requirement is to determine the words in the sentence. The word segmentation problem is always the first problem to solve the problem of automatic translation or problems in natural language processing. Besides the problems of studying Vietnamese word segmentation, the problem of Ede word segmentation has not yet been published and shared for the purpose of studying Ede language processing. To solve this problem and based on the current situation of processing ethnic minority languages in Vietnam in general and Ede language in particular, this article proposes a solution to segment Ede words using the maximum matching method based on the Ede vocabulary database, in order to contribute to solving the problem of segmentation in Ede language processing..**

***Keywords*— Bilingual vocabulary database; ethnic minority; Ede vocabulary separation; Ede language processing; maximum matching method.**

I. INTRODUCTION

To solve the problem of automatic translation or many other natural language processing problems, the word segmentation problem is always the first and most important problem. In languages of the isolating language type (Japanese, Chinese, Thai, Vietnamese, etc.), word boundaries are not blank characters like in languages of the amalgamating type (English, French, Russian, etc.), but there is a close connection between the words. A word can be made up of one or more words [1]. Therefore, with isolating languages, the problem of word segmentation is to determine the boundaries between words.

In English, each word has a meaning and is defined by blank characters. Vietnamese is an isolating language, so there are many compound words. Therefore, when translating from Vietnamese to English, if the words are not separated correctly, each word will be translated and then combined together.

With English, because it is an inflectional language, it is easier to determine the word class, and there are also few homophones. Polysemous words in English, Vietnamese and almost all other languages are very complex. Therefore, to determine the exact meaning of a word, contextual analysis must be performed.

Words in English, Vietnamese and Ede have differences in basic units; prefixes, suffixes; word types and word boundaries. Table 1 shows that the problem of word segmentation in Vietnamese and Ede is difficult to handle ambiguity.

Table 1: Comparison of main differences between English, Vietnamese and Ede

Characteristic	English	Vietnamese	Ede
Basic Units	Vocabulary	Speech	Speech
Prefixes/Suffixes	Yes	No	No
Parts of Speech Word	Well defined	Not well defined	Not well defined
Boundaries	Space or punctuation	Combinations have meaning based on context	Combinations have meaning based on context.

Approaches in the word separation problem [5]:

- Maximum Matching (MM) is a dictionary-based word segmentation method. The MM method tries to match the longest possible word in the dictionary. The accuracy of this algorithm depends on the size of the dictionary. The MM method cannot solve the problem of ambiguity and cannot recognize words that are not in the dictionary. This method only correctly separates words that are in the dictionary. - Transformation-Based Learning (TBL) is an approach based on a marked corpus. According to this approach, the computer can recognize the boundaries between words, so that it can accurately segment words. The machine is taught sample sentences in the corpus that have been marked with the boundaries between correct words. The method is very simple, because it only needs to let the machine learn sample sentence sets and then the machine will automatically derive the rules of the language and from there it will apply correctly when there are sentences that are correct according to the rules that the machine has drawn. The accuracy of this algorithm depends on the complete corpus and must be trained for a long time so that the machine can extract complete rules.
- Weighted Finite State Transducer (WFST) method, is a method based on the idea of applying WFST with the weight being the probability of each word appearing in the corpus. This model achieves relatively high accuracy by using additional Neural networks combined with dictionaries to eliminate possible ambiguities when separating many words in a sentence and then the Neural network layer will remove inappropriate words by combining with the dictionary. Besides, similar to the TBL method, the important point of this model is that it requires a complete corpus.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 06, June - 2025

- Maximum Entropy method, is a method based on the idea of Ad-wait Ratnaparkhi's Maximum Entropy method for English word labeling model. This is a new direction for current word separation methods. If the corpus is fully labeled, then ambiguities can be eliminated.

However, there is still no work that can quantify the accuracy of this method.
With the word segmentation problem, we can see that each method has its own advantages and limitations, but all require a large enough corpus for the word segmentation results to achieve high accuracy.

II. INTRODUCTION TO EDE LANGUAGE

The Ede language belongs to the Malayo-Polynesian (Austronesian) language family and is related to many mainland Austronesian languages. The main dialect used is the Ede Kpa in the Central Highlands. On December 2, 1935, the Governor-General of Indochina signed and recognized the writing system using Latin characters for common use among ethnic minorities in the Central Highlands. This script was revised many times and is called the Ede script because it is widely used in the Ede community, as this is one of the most populous ethnic groups in the Central Highlands [3], [10].

A. Ede phonetic characteristics

Ede is an isolating and polysyllabic language, without tones. The process of morphological transformation from a polysyllabic language with consonants has significantly affected the phonetic characteristics of Ede [3], [4], [8]. In Ede, the initial part of the syllable has not been completely monosyllabic, so its phonetic and phonological structure is complex. Morphological transformations of words are not many and take place right in the syllable shell itself, making the syllable have an unstable phonetic structure, and the change in morphology also changes the meaning of the word, for example:

Djiê	die	→ *mdjiê*	*kill*
Bŏ	*full*	→ *mbŏ*	*fill*

The monosyllabic tendency in Ede language is the cause of morphological change [4].

Table 2. Ede alphabet

Consonant	***Uppercase***	A	B	Č	D	Đ
		G	H	J	K	L
		M	N	Ñ	P	R
		S	T	W	Y	
	Lowercase	b	ƀ	č	d	đ
		g	h	j	k	l
		m	n	ñ	p	r
		s	t	w	y	
Vowels	***Uppercase***	A	Ă	Â	E	Ĕ
		Ê	Ĕ	I	Ĭ	O
		Ŏ	Ô	Ǒ	Ơ	Ŏ
		U	Ŭ	Ư	Ư̆	
	Lowercase	a	ă	â	e	ĕ
		ê	ĕ	i	ĭ	o
		ŏ	ô	ǒ	ơ	ơ̆
		u	ŭ	ư	ư̆	

B. Vocabulary features

The Ede vocabulary includes many word classes and many lexical elements originating from many different language groups in the Southeast Asian region.
The words in the Ede language are monosyllabic, the number of polysyllabic words is very small. The main method of word formation in the Ede language today is compounding. The process of monosyllabicization and vocabulary borrowing has contributed to promoting the changes in the meanings of homonyms and synonyms in the Ede language [4], [8], [10].

C. Grammatical features

Ede grammar has the grammatical structure characteristics of isolating languages. Expressing grammatical meanings, the grammatical method in Ede is the method of word order and function words. The sentence structure model in Ede is quite clearly defined, with the characteristics of Ede.
In a declarative sentence, the subject comes before the predicate, and the complement comes after the predicate. The determiner usually comes after the element it complements, but the positional adverb can change.
In the case of adverbs expressing degree, they come after adjectives or verbs. For example:

- snăk (very) is a single word
- siam snăk (very beautiful)
- êdi, đei synonym of snak
- -snăk s'ưm, tliă tliêt are compound words with the same meaning as snăk but at a higher level.

In Ede interrogative sentences, the question word is often placed at the beginning of the sentence (this feature is different from interrogative sentences in Vietnamese):
To ask about a place, the word "Ti anôk" (where) comes at the beginning of the question. For example:

Ti anôk sang ih? Where do you live?

To ask what job or profession, The word "Ya bruă" (what do you do) comes at the beginning of the question.

For example:

Ya bruă ih ngă? What do you do?

Punctuation marks (commas, periods, question marks...) in Ede are used as in Vietnamese. The capitalization principle of Ede is the same as that of Vietnamese [4], [8].

D. Solutions to separate Ede words

The choice of approach in the word segmentation problem must be consistent with the current state of Ede language processing and the characteristics of Ede language, specifically:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 06, June - 2025

Maximum Entropy method, requires a fully labeled corpus. In fact, up to now, there is no fully labeled Ede language corpus to apply this method.

Weighted finite state transformation method, requires a complete corpus. In Ede language processing, there is still no complete learning corpus.

Transformation-based learning method, requires a complete corpus and must be trained for a long time so that the machine can extract complete rules. The research results of Ede language up to now still do not have a complete corpus to approach this method.

From the above analysis, we propose to choose the maximum matching method based on the Ede language vocabulary as the approach in the Ede word segmentation problem.

1) Characteristics of Ede vocabulary

Ede words are mainly collected and recorded in the Ede language of the Kpa group. Ede words partly reflect the traditional cultural capital of the Ede people [7], [9]. Ede is recorded in Ede script and saved on computers with Unicode fonts. The number of Ede words in the vocabulary is 9,288, according to the number of syllables in the words, as shown in Table 3 [7].

Table 3. Statistics of the number of Ede words according to the number of syllables

Number of syllables	**1**	**2**	**3**	**4**	**5**	**6**
Number of entries	3.769	5.023	927	157	14	2

2) Proposed method for segmenting Ede words

Solution for segmenting Ede words using the maximum matching method based on the vocabulary. With the Ede vocabulary as shown in Table 3, we propose to split it into vocabulary stores according to the number of syllables of the entry to perform maximum matching with a length of 6 syllables. The magnetic separation solution is shown in Figure 1.

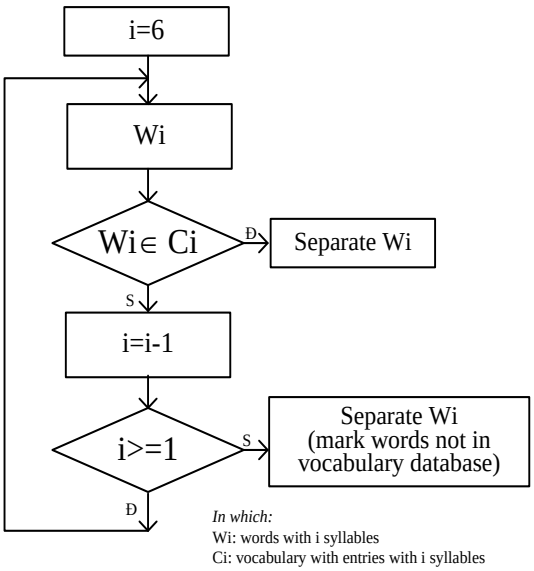

Figure 1. Solution to segment Ede words based on vocabulary

Based on the proposed solution for separating Ede words, we have built the WSE (Word Segmentation Ede) tool to separate Ede words. The input data source of the toolkit is Ede sentences in the news of the VOV4 Ethnic Broadcasting System. Words not in the Ede vocabulary database are labeled with "*".

3) Experimental results

We have conducted experiments with the WSE set using the input data source of the VOV4 Ethnic Broadcasting System [2]. These news are guaranteed to be typed correctly and have been tested. The fonts used in the news are TayNguyenKey font, VNI typing method and UniKey typing method. We have applied the CEDU converter [6] to convert all character sets typed in TayNguyenKey font to Ede letters using Unicode font. Table 4 presents the experimental results with sentences in 30 news.

Table 4. Experimental results of word segmentation in Ede sentences

Number of words in the newsletter	Separable words	Number of correct words	Number of not correct words
6,643	5,781	4,983	798

The results obtained after going through the WSE toolkit were also tested by us with manually separated sentences and found that the WSE toolkit separated about 87% of the Ede words correctly. The remaining 13% of the words were mainly due to: words not in the vocabulary, spelling errors (manually inserted), Vietnamese and English words.

III. CONCULSION

Solution for Ede word segmentation using the maximum matching method based on the Ede vocabulary database. The result has separated Ede words in affirmative sentences.

This solution contributes to solving the problem of Ede word segmentation in Ede language processing. The problem of Ede word segmentation with the WSE tool has also provided words that are not yet in the Ede vocabulary database to continue to add to the Ede vocabulary database, contributing to enriching the Ede vocabulary database.

In the next direction, we will continue to complete the problem of Ede word segmentation when we have a fully labeled Ede language corpus.

ACKNOWLEDGMENT

Thank for all authors of the papers quoted.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 06, June - 2025

REFERENCES

[1] Avinesh. PVS, and Karthik G. Dept (2007), Part-of-speech tagging and chunking using conditional random fields and transformation based learning, Shallow Parsing for South Asian Languages 2.
[2] Voice of Vietnam Radio. VOV4 Ethnic Radio System. Address:http://vov4.vov.vn/Ede.aspx
[3] Doan Van Phuc, "Vocabulary of Ede dialects", NHo Chi Minh City Publishing House, 1998.
[4] Doan Van Phuc: Ede phoneticsSocial Sciences, Hanoi, 1996.
[5] Dien Dinh, et al, "A maximum entropy approach for Vietnamese word segmentation", Proceeding of 4th RIVF Vietnam, pp.12-16, 2006
[6] Hoang Thi My Le, Phan Huy Khanh, "Solution to convert Ede language text using private fonts to Unicode", Proceedings of the 10th FAIR National Scientific Conference, Da Nang, page: 205-211, 2017.
[7] Hoang Thi My Le, Phan Huy Khanh, "Solution to build a Vietnamese-Ede bilingual vocabulary repository based on the Vietnamese-Ede interaction model", Proceedings of the 5th Conference on Information Technology & Applications in Various Fields (CITA'2016), pp. 32-37, code ISBN: 978-604-80-2094-1, 2016.
[8] DakLak Department of Education and Training: Ede Grammar, Education Publishing Hou, 2011.
[9] Ta Van Thong, "Ede-Vietnamese Dictionary", Vietnam Institute of Lexicography and Encyclopedia, 2014.
[10] Y Cang Nie Sieng, Y CôC Mlô, Hdruôm Hră Hriăm E De , Dak Lak Department of Education, 2007