Vietnamese Words Are Not Constructed from Syllables: Rethinking the Role of Word Segmentation in Natural Language Processing for Vietnamese Texts

Nghia Hieu Nguyen[1,3,4], **Dat Tien Nguyen**[2,3,4], **Ngan Luu-Thuy Nguyen**[1,3,4]
[1]Faculty of Information Science and Engineering
[2]Faculty of Computer Science
[3]University of Information Technology
[4]Vietnam National University, Ho Chi Minh city, Vietnam
nghiangh@uit.edu.vn, 21521944@gm.uit.edu.vn, ngannlt@uit.edu.vn

Abstract

The definition of words is the fundamental and crucial linguistic concept. Any changes in word definition lead to changes in the theoretical system of the respective language. Traditionally, researchers in Natural Language Processing (NLP) for Vietnamese texts believe Vietnamese words are constructed from syllables. However, their works did not explicitly mention which linguistic theory they followed for this assumption. Although there are no theoretical guarantees, most NLP studies in Vietnamese accept this assumption. Consequently, word segmentation is recognized as one of the essential stages in NLP for Vietnamese texts. In this study, we address the role of word segmentation for Vietnamese texts from linguistic perspectives. Through our extensive experiments, we show that, based on linguistic theories, performing word segmentation is not appropriate for Vietnamese text understanding. Moreover, we present a novel method, **Vi**etnamese **Word** Trans**Former** (ViWordFormer), for modeling Vietnamese word formation. Experimental results indicate that our method is appropriate for modeling Vietnamese word formation from both theoretical and experimental aspects and embark on a novel approach to Vietnamese word representation.

Introduction

Word segmentation has been known as one of the features of Vietnamese texts for a long time, and many studies were constructed particularly for word segmentation tasks (Nguyen et al. 2006; Nguyen and Le 2016; Hong Phuong et al. 2008; Nguyen et al. 2019). Typically, research in NLP for Vietnamese texts can be categorized into syllable-based and word-based approaches.

In the syllable-based approach, texts are segmented into syllables by spaces. While in the word-based version, texts are segmented into words, each word can have one or more syllables (Đinh Điền 2005). These syllables are connected by the underscore "_". For instance, given the sentence "chúng tôi là nghiên cứu sinh nhóm Nghiên cứu Xử lý Ngôn ngữ Tự nhiên trên tiếng Việt (we are the students from the natural language processing research group for Vietnamese)", following the word-based approach, this sentence must be pre-preprocessed as "chúng_tôi là nghiên_cứu_sinh đến từ nhóm nghiên cứu xử_lý ngôn_ngữ tự_nhiên cho tiếng Việt", while it keeps the original version for the syllable-based approach.

Vietnamese linguists follow the study of American Distributionalism (Bloomfield 1933) to define that in Vietnamese, morpheme is the most fundamental lexical unit that is meaningful, each morpheme has one syllable and is written in a sequence of continuous characters. In a sentence, morphemes are written from left to right and separated by spaces. For instance, "Hà Nội là thủ đô của Việt Nam" (Hanoi is the capital of Vietnam) has eight morphemes.

According to former linguists (Thản 1997; Lê 2003; Châu 2007), Vietnamese words are constructed from morphemes, each word may have one or more morphemes. Accordingly, in "Hà Nội là thủ đô của Việt Nam" (Hanoi is the capital of Vietnam), there are five words: "Hà Nội" (Hanoi), "là" (is), "thủ đô" (capital), "của" (of), "Việt Nam" (Vietnam). In this study, we name this view of Vietnamese word definition as **morpheme-based linguistic view**.

However, later linguists (Cẩn 1996; Giáp 2008, 2011) define that, in Vietnamese, words and morphemes are identical (Hạo 1998; Cẩn 1996; Giáp 2011; Thản 1997). In our study, this view is named **word-based linguistic view**. According to this linguistic view, there are eight words in the sentence "Hà Nội là thủ đô của Việt Nam" (Hanoi is the capital of Vietnam). By this definition, the word-based linguistic view can uncover many linguistic ambiguities introduced by the morpheme-based linguistic view. In the following sections, we will analyze how the word-based linguistic view tackles unsolved problems of the morpheme-based linguistic view.

From these two linguistic views, the syllable-based approach in Vietnamese NLP follows the word-based linguistic view, while the word-based approach follows the morpheme-based linguistic view.

In Vietnamese NLP, modeling word segmentation actually concentrates only on describing complex words (including compound words and reduplicative words). However, according to word-based linguistic view, we should model the phrasal lexeme - the more abstract lexical unit of complex word in Vietnamese. To this end, we propose ViWordFormer, a Transformer-based method that describes the phrasal lexemes in Vietnamese.

Our experimental results indicate that baseline methods that follow the word-based linguistic view to represent Viet-

namese words achieve competitive results compared to those that follow the morpheme-based linguistic view on various datasets of natural language understanding (NLU) in Vietnam. Moreover, our ViWordFormer empirically outperforms the baselines on most datasets. These results provide the Vietnamese NLP community with evidence-based recommendations and a novel method that can (1) comprehensively address the characteristics of Vietnamese words from the stands of linguistic theories and (2) simplify Vietnamese text processing workflows without any downgrade in terms of accuracy.

Related Works

Morpheme-Based Linguistic View

Morpheme-based linguistic view states that words in Vietnamese are formed from morphemes (Thản 1997; Châu 2007; Lê 2003; Đinh Điền 2005). In Vietnamese NLP, most researchers follow this definition. Consequently, word segmentation is one of the fundamental tasks for Vietnamese text (Nguyen et al. 2017, 2006; Nguyen and Le 2016; Nguyen and Nguyen 2021).

Typically, (Nguyen et al. 2006) investigated the use of Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) for Vietnamese word segmentation. They demonstrated that Vietnamese word boundaries depend more on the identity of the syllables than on features like capitalization or sentence positioning. Moreover, they found that SVM-based learning is a potential approach to solving the Vietnamese word segmentation problem.

(Vu et al. 2018) present VnCoreNLP, a comprehensive natural language processing toolkit specifically designed for the Vietnamese language. The toolkit employs the RDRSegmenter (Nguyen et al. 2017), an efficient rule-based method that leverages Ripple Down Rules (RDR) for Vietnamese word segmentation. It provides state-of-the-art (SotA) performance on the Vietnamese TreeBank corpus (Nguyen et al. 2009).

In the study of (Nguyen and Tuan Nguyen 2020), they said all publicly released monolingual and multilingual BERT-based language models are not inherently aware of the distinction between Vietnamese syllables and word tokens. To this end, they introduced the PhoBERT and BARTPho pre-trained language models (Tran, Le, and Nguyen 2021; Nguyen and Tuan Nguyen 2020), which demonstrate the particular color of Vietnamese by requiring segmenting sentences into words before forwarding them to their pre-trained models to obtain respective features. In other words, (Nguyen and Tuan Nguyen 2020; Tran, Le, and Nguyen 2021) followed the morpheme-based linguistic view.

Word-Based Linguistic View

In contrast to the morpheme-based linguistic view, (Cẩn 1996; Hạo 1998; Giáp 2008, 2011) stated that words and morphemes are identical in Vietnamese. Studies such as (Phan et al. 2022; Nguyen et al. 2023; Do et al. 2024) introduced pre-trained language models for Vietnamese and gave no special treatment for Vietnamese word segmentation.

In particular, (Phan et al. 2022) constructed a pre-trained T5 model for Vietnamese text-to-text tasks and introduced the ViT5 pre-trained model. VisoBERT, a RoBERTa-based pre-trained language model for Vietnamese social media texts, was introduced in (Nguyen et al. 2023). These pre-trained language models gave no special treatment for words in their text preprocessing. Thus, they implicitly followed the morpheme-based linguistic view.

A study has evaluated the impact of word segmentation on NLP tasks in Vietnamese. In particular, research by (Vu et al. 2018) suggests that subword-based approaches can handle Vietnamese NLP tasks efficiently without requiring word segmentation. However, experiments in this study were conducted from technical perspectives and did not have any linguistic theories as its methodology.

Preliminary of Vietnamese Lexical Structure

Morpheme-Based Linguistic View

Former Vietnamese linguists (Tệ, Cận, and Đình Tú 1962; Thản 1997; Lê 2003; Châu 2007) believe that Vietnamese words are composed of morphemes. This indicates that when we divide Vietnamese words into smaller units that are meaningful, we have morphemes. If we continue to break down these morphemes, we have nonsense phonemes. Hence, following this linguistic view, morphemes are the smallest and meaningful units that conduct Vietnamese words. In addition, each morpheme has only one syllable (Châu 2007).

Words in Vietnamese are then constructed in three distinctive ways:

1. **Lexicalizing morphemes:** Some morphemes have their meaning, are distinguished from other morphemes by their lexical meaning and grammatical meaning, and have enough grammatical functions to participate in forming sentences. These morphemes can be *lexicalized* to become words. These words are called *mono-syllabic words* in Vietnamese.
2. **Compounding morphemes:** Most of words in Vietnamese have more than one syllable (Châu 2007).Their meaning is gained by combining the meanings of their components to indicate a more general meaning. For instance, by combining "quần" (trousers) and "áo" (shirts) into "quần áo," we aim to indicate the more general term *clothes* rather than the narrowed meaning of *shirts* or *trousers*. The same is applied to "ăn uống," compounding "ăn" (eating) and "uống" (drinking) to indicate the act of eating and drinking in general. Vietnamese words constructed in this way are called *compound words*.
3. **Reduplicating phonemes of morphemes:** The Vietnamese lexical system contains words that share the same phonemes in their morphemes. For example, "g̲ọn g̲àng" (compact), "đ̲ẹp đ̲ẽ" (beautiful), "bố̲i rố̲i" (confused). In such examples, starting from the original morphemes, the Vietnamese reduplicate their phonemes to yield the second morphemes. Vietnamese words constructed in this way are called *reduplicated words*.

In particular, there are two kinds of compound words (Châu 2007):

- Coordinating compounds: words having more than one morpheme, each morpheme contributes equally to the general meaning of the whole word.
- subordinating compounds: words having more than one morpheme, but only one morpheme carries the main meaning; the others are modifiers that narrow down the meaning of the main morpheme to construct the overall meaning of the whole word.

Conversely, the morpheme-based linguistic view states that **Vietnamese words are constructed from syllables**. In this paper, we call compound and reduplicated words as *poly-syllabic words*.

In Vietnamese NLP, most researchers follow this linguistic view to treat words as combinations of morphemes. Consequently, Vietnamese text processing is required to perform word segmentation in order to achieve words from the input sentence. Word segmentation aims to analyze the sentence, process each token as morphemes, and then combine the relevant morphemes to construct respective poly-syllabic words.

Semantic Structure of Lexical Units in Vietnamese

The definition of morpheme and word in morpheme-based linguistic view face lots of irregular cases in Vietnamese. Studies of (Hạo 1998; Giáp 2008, 2011) indicated that there are compound words having two morphemes. However, only one morpheme between them carries the meaning of the whole compound word, or even no morpheme is relevant to the whole meaning of their respective word. For instance, "ăn nói" means to speak in English, including "ăn" (to eat) and "nói" (to speak), or "đất nước" means country in English, including "đất" (soil) and "nước" (water). Therefore, these words can not be categorized as any defined categories mentioned in the previous section.

Actually, in Vietnamese, both "ăn" (eating) and "nói" (speaking) are the metonymies of the mouth. When these two words are standing in the same lexical unit, they describe the communication function of the mouth. We can easily find the evidence from an idiom in Vietnamese for this way of explanation: "học ăn học nói, học gói học mở" (original meaning: learn to eat and learn to speak, learn to wrap and learn to open; metaphorical meaning: learn to do the most simple thing, even how to speak and how to do) (Lân 2016). To this end, "ăn nói" is a coordinating compound that means speaking in English. The same explanation can be applied to similar words such as "nhà cửa" (house in general).

In addition, there are reduplicative words in Vietnamese, including nonsense morphemes, such as "đẽ", "gàng", and "bối" in "đẹp đẽ", "gọn gàng", and "bối rối" mentioned in the previous section. According to (Châu 2007), morpheme is the smallest lexical unit that has meaning. However, these lexical units have no meaning, which contradicts the definition of word and morpheme following the morpheme-based linguistic view.

(Châu 2007) gave a solution for these kinds of nonsense morphemes in such reduplicative words. In particular, given "đẹp đẽ" (beautiful) as an example, "đẽ" only follows after "đẹp". When Vietnamese people hear the sound of "đẽ", the only thing that appears in their mind is "đẽ" in "đẹp đẽ". Accordingly, the meaning of "đẽ" is the meaning of "đẹp". We can explain the same way for other similar words such as "ngơ ngác" (bewildered), "mếu máo" (to cry), or "lác đác" (scattered). To this end, reduplicative words can be seen as coordinating compounds.

Moreover, there are compound words such as "xe cộ" (vehicles) and "chùa chiền" (pagodas in general), but only one morpheme has meaning in these words. There are also words such as "ễnh ương" (bullfrog), "bù nhìn" (puppet), or "xà phòng" (soap) that include two morphemes, but their morphemes do not have any meaning.

(Giáp 2008, 2011) showed that morphemes "cộ" in "xe cộ", "chiền" in "chùa chiền" have meaning, but their meanings are vague or lost because these words are dialect and because of the popularity of Vietnamese standard language. (Giáp 2008, 2011) indicated that in the Southern Delta of Vietnam, ancient people used "cộ" to describe a type of vehicle without wheels, also known as a "scratch cart", pulled by buffaloes and oxen, sliding on flat ground or wet mud. The meaning of "chiền" is pagoda, as reflected in a folk song of Vietnam (Giáp 2008). From that on, these words are coordinating compounds according to the morpheme-based linguistic view. Moreover, (Giáp 2008, 2011) stated that in words such as "ễnh ương" (bullfrog) or "bù nhìn" (puppet), their morphemes had had meaning. However, their meaning is lost over the usage and development of Vietnamese.

Finally, besides the Vietnamese words constructed from morphemes whose meanings are forgotten, there are words that include nonsense morphemes, and these morphemes do not have meaning in Vietnamese. Such instances are "xà phòng" (soap), "cát xê" (salary), "xi lanh" (cylinder), or "tấm bạt" (tarpaulin). These words are borrowed words from foreign languages. In particular, "xà phòng" is borrowed from French savon, "cát xê" is borrowed from French caché, "xi lanh" is borrowed from French "cylindre", and "bạt" in "tấm bạt" is borrowed from French "bâtre" (Phê 2020). These words are phonologicalized using the Vietnamese phonological rules to form novel words. This linguistic phenomenon has appeared since the French colonial period in Vietnamese history.

To conclude, we showed that following studies of (Giáp 2008, 2011), any words having more than one morpheme according to morpheme-based linguistic view can be viewed as coordinating words or subordinating words. In other words, there are only two rules to form poly-syllabic words in Vietnamese following the morpheme-based linguistic view: *coordinating morphemes* and *subordinating morphemes*. Words borrowed initially from foreign languages are treated as irregular cases because they follow the rules of word formation of their original languages.

Word-Based Linguistic View

According to the analysis of semantic structure of Vietnamese complex words in the previous section, all morphemes in Vietnamese have meaning. To this end, (Giáp 2008) defined that, in Vietnamese, word is the smallest lexi-

cal unit that has meaning, includes one syllable, and is written under a continuous sequence of characters. From this on, words, morphemes, and syllables are identical in Vietnamese. Therefore, **Vietnamese words are not constructed from syllables** and Vietnamese is a monosyllabic language.

Moreover, (Giáp 2008, 2011) proved that in Vietnamese, there is no boundaries to distinguish compound words, idioms, proverbs, and habitual collocations. (Giáp 2008, 2011) indicated that these lexical units share the same properties with compound words, and they are constructed following the same rules of compound word formation.

Accordingly, (Giáp 2011) categorized them as **phrasal lexemes** in Vietnamese. Phrasal lexeme generalizes the concept of word from the definition of (Châu 2007) as well as addresses the problems relevant to determining the boundaries of words in former studies (Cẩn 1996; Châu 2007; Lê 2003; Tệ, Cận, and Đình Tú 1962). In conclusion, according to the word-based linguistic view, all phrasal lexemes are constructed from words; each word has one syllable, and these words either coordinately compound or subordinately compound together to form the whole meaning.

Vietnamese Word Transformer (ViWordFormer)

We designed ViWordFormer to represent Vietnamese words and phrasal lexemes following a word-based linguistic view. ViWordFormer is mainly inspired by Transformer because this method can model the relation among components in sentences via its attention mechanism.

Attention Module

The Attention module of ViWordFormer is designed following the implementation of (Vaswani et al. 2017). In particular, Attention module will perform the self-attention, which implies the query, key, and value vectors are the linear-mapped vector from the same input sentence:

$$Q = W_q I, K = W_k I, V = W_v I$$

where $W_q \in \mathbb{R}^{d_{in} \times d_{model}}$, $W_k \in \mathbb{R}^{d_{in} \times d_{model}}$, and $W_v \in \mathbb{R}^{d_{in} \times d_{model}}$ are learnable parameters, $I \in \mathbb{R}^{len \times d_{in}}$ is the vector of input sentence.

The Attention module will determine the Attention score $\mathcal{A}$ via:

$$\mathcal{A} = softmax\big(\frac{QK^T}{\sqrt{d_{model}}}\big) \in \mathbb{R}^{len \times len} \quad (1)$$

Phrasal Lexeme Module

The attention mechanism proposed by (Vaswani et al. 2017) does not have any constraints on how to perform attentive connection, hence there will be unexpected connections among words that are not in the same phrasal lexemes in Vietnamese.

To tackle this problem, we introduce the Phrasal score $\mathcal{P}$ to augment the Attention score $\mathcal{A}$ with constraints that eliminate the unexpected attentive connections among words in different phrasal lexemes.

As we analyzed in the previous section, any phrasal lexemes in Vietnamese are formed following two rules of semantic formation. To describe these two rules, we define a kernel:

$$r_{i,j} = \mu(w_i, w_j) = w_i W w_j^T \quad (2)$$

where $W \in \mathbb{R}^{d_{model} \times d_{model}}$ is the learnable parameter, and w_i and w_j are any two words in the input sentence.

Then given an input sentence $w_1, w_2, ..., w_{len}$, we measure the probability of any words w_i and its neighbor words w_{i-1} and w_{i+1} that are in the same phrasal lexeme as:

$$pr_{i-1,i}, pr_{i,i+1} = softmax(r_{i-1,i}, r_{i,i+1}) \quad (3)$$

By this equation, we expect if w_i and w_{i-1} are in the same phrasal lexeme, but w_{i+1} is in the other, then we have $pr_{i-1,i} > pr_{i,i+1}$, and vice versa.

However, it is worth noting that because we defined the kernel μ as the bilinear function, we might have $pr_{i,i+1} \neq pr_{i+1,i}$ although they represent the same concept. To this end, the probability of word w_i and w_{i+1} in the same phrasal lexeme is calculated by taking the average of $pr_{i,i+1}$ and $pr_{i+1,i}$:

$$P_i = \sqrt{pr_{i,i+1} \times pr_{i+1,i}} \quad (4)$$

Finally, the probability $\mathcal{P}_{i,j}$ of words $i, i+1, ..., j$ in the same phrasal lexeme is determined as:

$$\mathcal{P}_{i,j} = \prod_{k=i}^{j-1} P_k \quad (5)$$

It is worth noting that $P_k \in [0, 1]$, hence the probability $P_{i,j}$ will converge to 0 rapidly. To alleviate this problem, we turn $P_{i,j}$ into the logarithmic space:

$$\mathcal{P}_{i,j} = exp\big(log\big(\prod_{k=i}^{j-1} P_k\big)\big) = exp\big(\sum_{k=i}^{j-1} log(P_k)\big) \quad (6)$$

Finally, the constraints introduced by the Phrasal score $\mathcal{P}$ provided to the Attention score by:

$$\mathcal{S} = \mathcal{A} \odot \mathcal{P} \quad (7)$$

where $\odot$ is the element-wise multiplication.

We note that we do not use the form of summation to "add constraints to the Attention score" as intuition because both $\mathcal{A}$ and $\mathcal{P}$ are in the exponential form. Therefore, the multiplication is performed to implement the intuition of "addition".

Co-Text Module

Through our observation, in the Phrasal Lexeme module, ViWordFormer finds difficulty when forming large phrasal lexemes. It mostly highlights the phrasal lexemes containing two to three words. We hypothesize that when being forwarded to deeper layers, the information among smaller phrasal lexemes that are used to construct larger phrasal lexemes is decayed. To this end, we introduce the Co-text module. This module designs a hierarchical flow of information so that deeper layers can maintain the information of

phrasal lexemes from shallower layers. We named this module the Co-text module because it maintains the co-text of any phrasal lexemes in a sentence.

In particular, let P_{ij}^{l} is the phrasal lexeme score at the layer l^{th}, we give P_{ij}^{l} the phrasal information from previous layers by:

$$\mathcal{P}_{ij}^{l} = \mathcal{P}_{ij}^{l-1} + (1 - \mathcal{P}_{ij}^{l-1})\mathcal{P}_{ij}^{l} \quad (8)$$

This formula describes the interdependencies among components in sentences. The co-text module allows higher layers to have more information about phrasal lexemes and, inversely, form the meaning of those components. In our experiments, we show that the Co-text module is indeed necessary.

Experiment

Baselines

We evaluated the strong baselines for sentence classification and sequential labeling: Transformer encoder (Vaswani et al. 2017), LSTM (Hochreiter and Schmidhuber 1997), GRU (Cho et al. 2014), and TextCNN (Kim 2014).

Configurations

We used the Transformer encoder with six layers and a hidden size 512. Our ViWordFormer was configured in the same way as the Transformer encoder method. For the RNN-based method (LSTM, GRU), we also had the 6-layer model. These methods had the same hidden size as the Transformer encoder. In this study, both ViWordFormer and Transformer used GeLU (Hendrycks and Gimpel 2016) as the activation function.

All baselines were trained with a single run using the early stopping technique on an A100 80GB GPU. Adam (Kingma and Ba 2014) was used as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e-9$. During training baselines and the proposed ViWordFormer method, we adapted the learning rate scheduler from the study of (Vaswani et al. 2017).

Datasets

In this study, we conducted experiments on two Vietnamese text understanding tasks: sentence classification and sequential labeling.

For the sentence classification task, we evaluated baselines and ViWordFormer on four datasets: UIT-VSFC (Nguyen et al. 2018), UIT-ViCTSD (Nguyen et al. 2021a), UIT-ViOCD (Nguyen et al. 2021b), and ViHSD (Luu, Nguyen, and Nguyen 2021).

For the sequential labeling task, we evaluated the baselines and ViWordFormer on PhoNER (Truong, Dao, and Nguyen 2021) and ViHOS (Hoang et al. 2023) datasets. We note that for the Name Entity Recognition (NER) task, we used PhoNER dataset because this dataset has the syllable version (Truong, Dao, and Nguyen 2021), while other available NER datasets require performing word segmentation at first, which prevents us from designing the experiments without word segmentation for this task.

Experimental Scenarios

In this paper, we constructed two experimental scenarios. The first scenario evaluates the necessity of performing word segmentation in Vietnamese text pre-processing. In other words, this scenario is used to evaluate the performance of baselines when they follow the morpheme-based linguistic view and word-based linguistic view, respectively. The second scenario evaluates the effectiveness of ViWordFormer over strong baselines following the word-based linguistic view.

Text Pre-Processing

In the first scenario, we used word segmentation to demonstrate the morpheme-based linguistic view on Vietnamese texts using the SotA VnCoreNLP toolkit (Vu et al. 2018). We called this pre-processing procedure the morpheme-based procedure. To describe the word-based linguistic view, we treat each token split by spaces as a Vietnamese word. We called this later pre-processing procedure word-based procedure. Then, we compared the baselines followed by both procedures to indicate the necessity of word segmentation in Vietnamese.

In the second scenario, we only applied the word-based procedure for Vietnamese texts. Then, we evaluated ViWordFormer's performance to compare with the baselines. It is worth noting that in both pre-processing procedures, sentences are turned into lowercase, and punctuations are separated from words.

Metrics

We adopted **F1-score** for evaluating baselines and the ViWordFormer on the mentioned datasets following the previous works (Hoang et al. 2023; Huynh, Nguyen, and Nguyen 2022; Nguyen et al. 2018; Ho et al. 2020; Nguyen et al. 2021a; Luu, Nguyen, and Nguyen 2021; Truong, Dao, and Nguyen 2021). The F1-score was reported in macro average.

Results and Analysis

Experimental Results of the First Scenario

Results from Table 1 show that the baselines when following the word-based linguistic view achieve approximately the same, even better than those following the morpheme-based linguistic view. For ease of observing the general pattern of baselines on datasets, we took their average F1 scores. We can see that word segmenting the input sentence in advance does not significantly improve evaluated baselines.

Moreover, these empirical results support the conclusion of the word-based linguistic view that words are identical with morphemes and syllables in Vietnamese. By splitting the sentence by spaces, we already achieve the sequence of words.

Experimental Results of the Second Scenario

As indicated in Table 2, the proposed ViWordFormer achieves the best results on most datasets. In particular, on both sequence classification tasks of UIT-VSFC, ViWordFormer outperformed Transformer as well as RNN-based

Datasets	Transformer		LSTM		GRU		TextCNN		Average	
	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w
UIT-VSFC Topic Classification	76,22	75,39	63,82	59,80	74,08	72,99	73,28	73,76	**71,85**	70,49
UIT-VSFC Sentiment Analysis	75,12	72,85	67,01	68,38	74,24	71,03	61,12	60,51	**69,37**	68,19
UIT-ViCTSD Toxic Detection	67,30	67,54	67,79	65,56	64,13	66,63	67,46	67,46	66,67	**66,80**
UIT-ViCTSD Constructive Detection	78,32	78,86	77,73	78,77	77,06	77,56	75,19	61,02	**77,08**	74,05
ViHSD	60,22	60,69	58,84	57,54	58,58	58,24	56,53	57,12	**59,01**	58,94
ViOCD	86,33	85,79	86,50	80,31	86,33	84,67	85,06	82,69	**86,06**	83,37
ViHOS	78,39	80,15	77,99	79,75	78,86	78,65	-	-	78,41	**79,52**
PhoNER	83,37	81,55	81,83	80,44	82,41	81,61	-	-	**82,54**	81,20

Table 1: Experimental results among baselines on sentences with (w) and without (w/o) word segmentation.

Datasets	ViWordFormer	Transformer	LSTM	GRU	TextCNN
UIT-VSFC Topic Classification	**77.37**	76.22	63,82	74,08	73,28
UIT-VSFC Sentiment Analysis	**76.92**	75.12	67,01	74,24	61,12
UIT-ViCTSD Toxic Detection	64.38	67.30	**67,79**	64,13	67,46
UIT-ViCTSD Constructive Detection	78.07	**78.32**	77,73	77,06	75,19
ViHSD	**63.68**	60.22	58,84	58,58	56,53
ViOCD	**88.70**	86.33	86,50	86,33	85,06
ViHOS	**79.22**	78.39	77,99	78,86	-
PhoNER	**84.61**	83.37	81,83	82,41	-

Table 2: Experimental results of ViWordFormer and baselines.

and CNN-based methods. The same results apply to the ViHSD and UIT-ViOCD datasets. However, on the UIT-ViCTSD, LSTM achieves the highest score on the Toxic Detection task, while Transformer is the best method for the Constructive Detection task. On the Toxic Detection task, ViWordFormer has lower scores than Transformer by a large margin. On the Constructive Detection task, ViWordFormer has approximately the same results as the Transformer.

The same color occurs in sequential labeling tasks. On both ViHOS and PhoNER, ViWordFormer outperforms all baselines. These results indicate the effectiveness of our proposed method when it learns to understand Vietnamese phrasal lexemes and sequentially classify them appropriately.

Ablation Study for Co-Text Module

Datasets	w	w/o
UIT-VSFC Topic Classification	**77,37**	75.60
UIT-VSFC Sentiment Analysis	**76,92**	76.53
UIT-ViCTSD Toxic Detection	**64,38**	62.22
UIT-ViCTSD Constructive Detection	**78,07**	76.01
ViHSD	**63,68**	59.87
ViOCD	**88,70**	86.33
ViHOS	**79,22**	78.00
PhoNER	**84,61**	62.91

Table 3: Ablation study for the Co-text module. **w** indicates ViWordFormer with Co-text module while **w/o** indicate ViWordFormer without the Co-text module.

As we stated previously, the Phrasal module must be equipped with the Co-text module so that the phrasal information from former layers can be forwarded to subsequent layers. To demonstrate this statement, we conducted an experiment to evaluate the performance of ViWordFormer with and without the Co-text module.

Results in Table 3 indicate that ViWordFormer achieves its highest scores on all datasets only being provided with the Co-text module. Notably, on the ViHOS dataset, ViWordFormer, without the Co-text module, dropped its performance by a large margin.

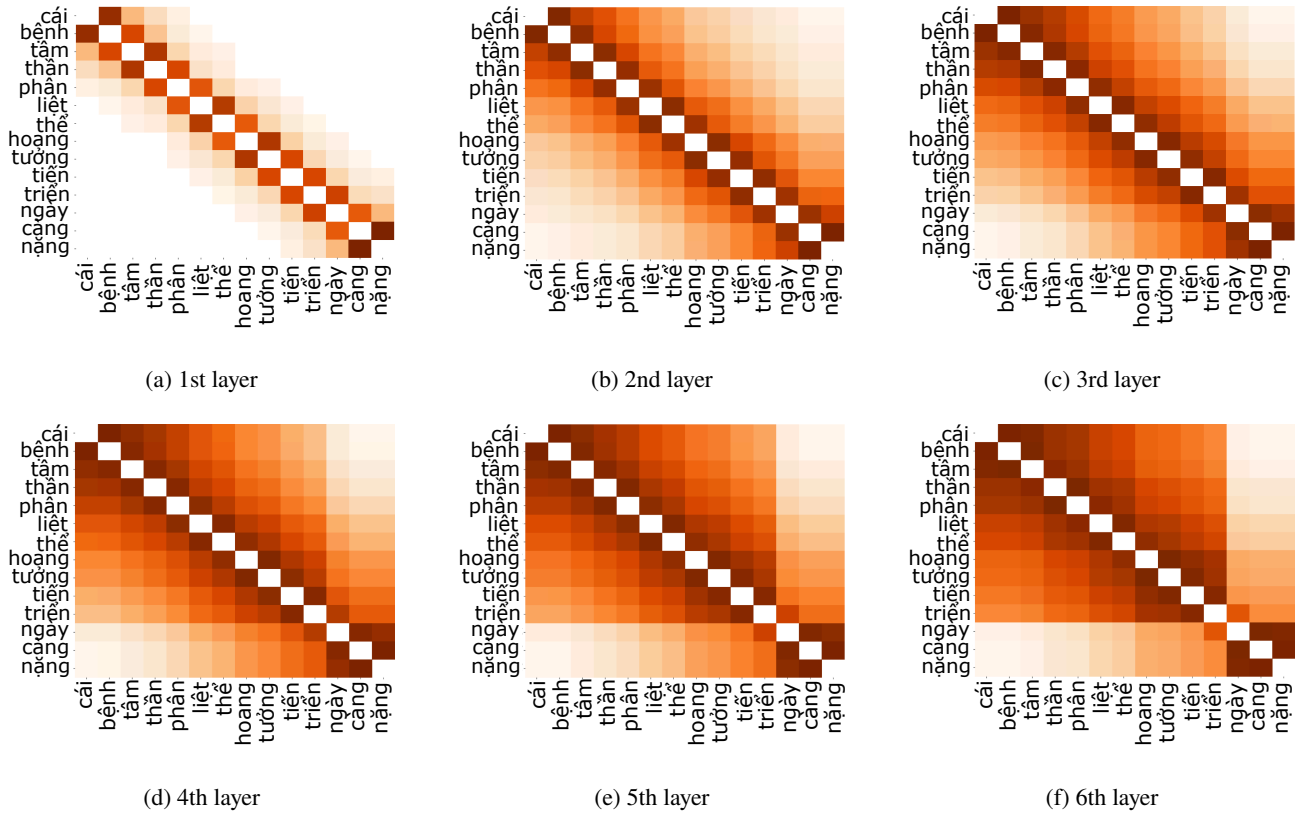
Visualization of ViWordFormer

To investigate how ViWordFormer forms phrasal lexemes in Vietnamese texts, we visualized the scores matrix $\mathcal{S}$ (7) returned by each layer of ViWordFormer.

Figure 1 presents the attention maps for the sentence "cái bệnh tâm thần phân liệt thể hoang tưởng tiến triển ngày càng nặng" (paranoid schizophrenia progressively worsens).

The score map $\mathcal{S}$ intensely captures phrasal relations among words in the first two layers. This is exemplified by the model's accurate identification of the phrasal lexemes such as "tâm thần" (mental disorder), "liệt thể" (paralysis), or "hoang tưởng" (paranoid). This behavior is closely aligned with word segmentation following the morpheme-based linguistic view, further highlighting the effectiveness and generalization ability of ViWordFormer. The sparse attention in these layers indicates that the model is learning to identify individual word pairs and smaller syntactic chunks.

Moving to the third layer, the score map $\mathcal{S}$ reveals more comprehensive patterns. This is a clear indication that ViWordFormer has started to capture broader semantic structures. For instance, it effectively captures the semantic relations in phrases such as "cái bệnh tâm thần" (mental illness) or "ngày càng nặng" (progressively worsens). These



(a) 1st layer (b) 2nd layer (c) 3rd layer

(d) 4th layer (e) 5th layer (f) 6th layer

Figure 1: Visualization of the score $\mathcal{S}$ from ViWordFormer.

are specific areas where words are shown to be relevant to each other, providing concrete evidence of ViWordFormer's effectiveness.

In the deeper layers (from the 4th layer to the 6th layer), the score map $\mathcal{S}$ reveals that ViWordFormer gradually constructs more extensive phrases from phrasal lexemes formed from previous layers. The attention maps of these layers show dense, consistent patterns and highlight ViWordFormer's ability to understand complicated semantic dependencies and the overall sentence structure.

In conclusion, through visualization of attention maps, we showed that ViWordFormer could gradually capture the phrasal relations among words and the semantic relation among phrasal lexemes given Vietnamese sentences. Our proposed ViWordFormer effectively captures Vietnamese semantic structure.

Conclusion

This study, using evidence from linguistic theories and empirical results, showed that word segmentation is not crucial for representing Vietnamese texts. Moreover, to effectively represent Vietnamese words following the word-based linguistic view, we introduced the ViWordFormer. Experimental results indicated that ViWordFormer gives better phrasal representation in Vietnamese; hence, it achieves better results on various Natural Language Understanding tasks compared with vanilla Transformer, RNN-based, and CNN-based methods.

Our results motivate the research community to explore more effective methods of modeling Vietnamese lexical structure using linguistic theories to construct better pre-trained language models, particularly for Vietnamese.

Limitation and Future Works

The main limitation is that ViWordFormer assumes the semantic structure μ of words and phrasal lexeme are linear, which is not ensured to be the best description of the phrasal relation in Vietnamese texts. In addition, our experiments are limited to sequence classification and sequential labeling tasks.

In subsequent studies, we will explore and evaluate ViWordFormer on other Vietnamese NLP tasks, such as text summarization, machine translation, or machine reading comprehension. Further experiments on the kernel μ of the Phrasal Lexeme module will be conducted to find the best approximation function for describing the semantic structure of words and phrasal lexemes in Vietnamese.

Acknowledgments

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under the grant number DS2024-26-01.

References

Bloomfield, L. 1933. *Language*. Henry Holt.

Cho, K.; van Merrienboer, B.; Çaglar Gülçehre; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*.

Châu, D. H. 2007. *Từ vựng ngữ nghĩa tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Cẩn, N. T. 1996. *Ngữ pháp tiếng Việt*. Nhà xuất bản Đại học Quốc gia Hà Nội.

Do, P. N.-T.; Tran, S. Q.; Hoang, P. G.; Nguyen, K. V.; and Nguyen, N. L.-T. 2024. VLUE: A New Benchmark and Multi-task Knowledge Transfer Learning for Vietnamese Natural Language Understanding. *ArXiv*, abs/2403.15882.

Giáp, N. T. 2008. *Từ vựng học tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Giáp, N. T. 2011. *Vấn đề "từ" trong tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian Error Linear Units (GELUs). *arXiv: Learning*.

Ho, V. A.; Nguyen, D. H.-C.; Nguyen, D. H.; Pham, L. T.-V.; Nguyen, D.-V.; Nguyen, K. V.; and Nguyen, N. L.-T. 2020. Emotion Recognition for Vietnamese Social Media Text. arXiv:1911.09339.

Hoang, P. G.; Luu, C. D.; Tran, K. Q.; Nguyen, K. V.; and Nguyen, N. L.-T. 2023. ViHOS: Hate Speech Spans Detection for Vietnamese. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 652–669. Dubrovnik, Croatia: Association for Computational Linguistics.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Hong Phuong, L.; Thi Minh Huyen, N.; Roussanaly, A.; and Vinh, H. T. 2008. A Hybrid Approach to Word Segmentation of Vietnamese Texts. In Martin-Vide, C.; Otto, F.; and Fernau, H., eds., *Language and Automata Theory and Applications*, 240–249. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-88282-4.

Huynh, T. V.; Nguyen, K. V.; and Nguyen, N. L.-T. 2022. ViNLI: A Vietnamese Corpus for Studies on Open-Domain Natural Language Inference. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 3858–3872. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Hạo, C. X. 1998. *Tiếng Việt mấy vấn đề ngữ âm - ngữ pháp - ngữ nghĩa*. Nhà xuất bản Giáo dục Việt Nam.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *Conference on Empirical Methods in Natural Language Processing*.

Kingma, D. P.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

Luu, S. T.; Nguyen, K. V.; and Nguyen, N. L.-T. 2021. A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts. In Fujita, H.; Selamat, A.; Lin, J. C.-W.; and Ali, M., eds., *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, 415–426. Cham: Springer International Publishing. ISBN 978-3-030-79457-6.

Lân, N. 2016. *Từ điển Thành ngữ và Tục ngữ Việt Nam*. Nhà Xuất bản Văn học.

Lê, H. 2003. *Vấn đề cấu tạo từ trong tiếng Việt hiện đại*. Nhà xuất bản Đại học Quốc gia Hà Nội.

Nguyen, C.-T.; Nguyen, T.-K.; Phan, X.-H.; Le Nguyen, M.; and Ha, Q. T. 2006. Vietnamese word segmentation with CRFs and SVMs: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, 215–222.

Nguyen, D. Q.; Nguyen, D. Q.; Vu, T.; Dras, M.; and Johnson, M. 2017. A Fast and Accurate Vietnamese Word Segmenter. arXiv:1709.06307.

Nguyen, D. Q.; and Tuan Nguyen, A. 2020. PhoBERT: Pre-trained language models for Vietnamese. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042. Online: Association for Computational Linguistics.

Nguyen, D.-V.; Thin, D. V.; Nguyen, K. V.; and Nguyen, N. L.-T. 2019. Vietnamese Word Segmentation with SVM: Ambiguity Reduction and Suffix Capture. *ArXiv*, abs/2006.07804.

Nguyen, K. V.; Nguyen, V. D.; Nguyen, P. X. V.; Truong, T. T. H.; and Nguyen, N. L.-T. 2018. UIT-VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 19–24.

Nguyen, L. T.; and Nguyen, D. Q. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 1–7.

Nguyen, L. T.; Nguyen, K. V.; Nguyen, K. V.; and Nguyen, N. L.-T. 2021a. Constructive and Toxic Speech Detection for Open-domain Social Media Comments in Vietnamese. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*.

Nguyen, N. T.; Ha, P. P.-D.; Nguyen, L. T.; Nguyen, K. V.; and Nguyen, N. L.-T. 2021b. Vietnamese Complaint Detection on E-Commerce Websites. In *New Trends in Software Methodologies, Tools and Techniques*.

Nguyen, P.-T.; Vu, X.-L.; Nguyen, T.-M.-H.; Nguyen, V.-H.; and Le, H.-P. 2009. Building a Large Syntactically-Annotated Corpus of Vietnamese. In Stede, M.; Huang, C.-R.; Ide, N.; and Meyers, A., eds., *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, 182–185. Suntec, Singapore: Association for Computational Linguistics.

Nguyen, Q.-N.; Phan, T. C.; Nguyen, D.-V.; and Nguyen, K. V. 2023. ViSoBERT: A Pre-Trained Language Model for Vietnamese Social Media Text Processing. In *Conference on Empirical Methods in Natural Language Processing*.

Nguyen, T.-P.; and Le, A.-C. 2016. A hybrid approach to Vietnamese word segmentation. In *2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, 114–119. IEEE.

Phan, L.; Tran, H.; Nguyen, H.; and Trinh, T. H. 2022. ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation. In Ippolito, D.; Li, L. H.; Pacheco, M. L.; Chen, D.; and Xue, N., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 136–142. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics.

Phê, H. 2020. *Từ điển Tiếng Việt*. Trung tâm Từ điển học Việt Nam.

Thản, N. K. 1997. *Nghiên cứu ngữ pháp tiếng Việt*. Nhà xuất bản Giáo dục.

Tran, N. L.; Le, D. M.; and Nguyen, D. Q. 2021. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. *ArXiv*, abs/2109.09701.

Truong, T. H.; Dao, M. H.; and Nguyen, D. Q. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Tệ, H.; Cận, L.; and Đình Tú, C. 1962. *Giáo trình Việt ngữ*, volume 1. Hà Nội.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vu, T.; Nguyen, D. Q.; Nguyen, D. Q.; Dras, M.; and Johnson, M. 2018. VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In Liu, Y.; Paek, T.; and Patwardhan, M., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 56–60. New Orleans, Louisiana: Association for Computational Linguistics.

Đinh Điền. 2005. *Xây dựng và khai thác kho ngữ liệu song ngữ Anh - Việt điện tử*. Trường Đại học Khoa học Xã hội và Nhân văn Tp. Hồ Chí Minh.