Vietnamese Sentence Paraphrase Identification using Pre-trained Model and Linguistic Knowledge

Dien Dinh, Nguyen Le Thanh
Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

***Abstract*—The paraphrase identification task identifies whether two text segments share the same meaning, thereby playing a crucial role in various applications, such as computer-assisted translation, question answering, machine translation, etc. Although the literature on paraphrase identification in English and other popular languages is vast and growing, the research on this topic in Vietnamese remains relatively untapped. In this paper, we propose a novel method to classify Vietnamese sentence paraphrases, which deploys both the pre-trained model to exploit the semantic context and linguistic knowledge to provide further information in the identification process. Two branches of neural networks built in the Siamese architecture are also responsible for learning the differences among the sentence representations. To evaluate the proposed method, we present experiments on two existing Vietnamese sentence paraphrase corpora. The results show that for the same corpora, our method using the PhoBERT as a feature vector yields 94.97% F1-score on the VnPara corpus and 93.49% F1-score on the VNPC corpus. They are better than the results of the Siamese LSTM method and the pre-trained models.**

***Keywords*—Paraphrase identification; Vietnamese; pre-trained model; linguistics; neural networks**

I. Introduction

Paraphrase identification, a task that whether two text segments with different wordings express similar meaning, is critical in various Natural Language Processing (NLP) applications, such as text summarization, text clustering, computer-assisted translation, and, especially plagiarism detection [1]. Paraphrases can take place at different linguistic levels, ranging from word and phrase to sentence and discourse. For instance, Neculoiu et al. [2] deployed Siamese recurrent networks to determine similarity among texts, normalizing job titles that are paraphrases at the word level. Meanwhile, to detect paraphrases at the discourse level, Liu et al. [3] calculated semantic equivalence among academic articles published in 2017 to identify documents with similar themes and contents.

Paraphrase corpora are corpora that contain pairs of sentences that convey the same meaning. Regarding Vietnamese, there have been two paraphrase corpora published for the language, one of which is vnPara by Bach et al. [4], while the other is VNPC (Vietnamese News Paraphrase Corpus) by Nguyen-Son et al. [5]. Both of these corpora consist of sentence-level paraphrases. Examples of paraphrases and non-paraphrases extracted from vnPara and VNPC are shown in Tables I and II, respectively.

While string matching is the simplest solution to the paraphrase identification question in theory, it does not yield high accuracy rates in practice. Two segments of text that are constructed with different strings can still be paraphrases. On the contrary, various text segments that have overlapping substrings can denote different interpretations, and thus they are non-paraphrases.

TABLE I. Examples of Vietnamese Paraphrases and their Translation into English

Paraphrase		Corpus
ASA sẽ tìm thấy người ngoài hành tinh trong 20 năm tới.	ASA nói có thể sẽ tìm thấy người ngoài hành tinh trong 20 năm tới.	vnPara
ASA will find aliens in the next 20 years	ASA says that it's possible to find aliens in the next 20 years	
Đáng chú ý, mã độc này chưa hoạt động mà ở chế độ "ngủ đông", chờ lệnh tấn công.	Đáng chú ý mã độc này chưa hoạt động mà ở chế độ nằm vùng.	VNPC
Remarkable, this malware has not been working yet but is in "hibernate" mode, wait for an attack.	Remarkable, this malware has not been working yet but is in "stand-by" mode.	

TABLE II. Examples of Vietnamese Non-paraphrases and their Translation into English

Non-paraphrase		Corpus
Các gián điệp Trung Quốc đã tấn công hệ thống mạng của một nước thuộc khu vực Đông Nam Á.	Các gián điệp Trung Quốc đã tấn công hệ thống mạng của một tổ chức nghiên cứu lớn của chính phủ Canada, giới chức Canada ngày 29/7 cho biết.	vnPara
Chinese spies have attacked the network system of a country in Southeast Asia.	Chinese spies have attacked a big Canada research organization's network system, from Canada authorities - 29/7.	
Cầu thủ trẻ đắt giá thứ ba mà Real từng đào tạo là Alvaro Negredo.	Một số cầu thủ khác từng trưởng thành từ lò đào tạo trẻ của Real là Cheryshev, Joselu, Diego Lopez và Rodrigo Moreno	VNPC
The third most valuable young player who Real has trained is Alvaro Negredo.	Some other players who have grown up at Real's youth academy are Cheryshev, Joselu, Diego Lopez and Rodrigo Moreno.	

According to Suzuki et al. [1], these two types of paraphrases and non-paraphrases are categorized as a non-trivial class, whose instances hold a key role in the paraphrase identification task. Table III presents examples of non-trivial instances extracted from VNPC. The WOR column in this stable represents the word overlap rate of two given sentences, which is calculated using Jaccard index [7], where X and Y denote the set of words of those two sentences:

TABLE III. EXAMPLES OF NON-TRIVIAL INSTANCES EXTRACTED FROM VNPC

Sentences pair		Type	WOR
Nadal đánh bóng ra ngoài, mất mini-break sớm.	Game đầu thứ chín, Nadal có tới bốn cú đánh bóng ra ngoài, để mất break.	paraphrase	21.05%
Nadal hit the ball out and lost the mini-break early.	In the ninth game of the match, Nadal hit four balls out and lost the break.		
Link sopcast xem trực tiếp U23 Đức vs U23 Nigeria trong khuôn khổ bán kết bóng đá nam Olympic 2016 được cập nhật liên tục tại đây.	Link sopcast xem trực tiếp U23 Brazil vs U23 Honduras trong khuôn khổ bán kết bóng đá nam Olympic 2016.	non-paraphrase	77.27%
Sopcast link to watch U23 Germany vs U23 Nigeria in the semi-final of the Men's Olympic Football 2016 is updated continuously here.	Sopcast link to watch U23 Brazil vs U23 Honduras in the semi-final of the Men's Olympic Football 2016.		

$$WOR(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|} \quad (1)$$

The accurate identification of non-trivial paraphrases and non-paraphrases requires methods that can exploit the semantic differences of texts. Hitherto, the paraphrase identification task has been a focus in various studies in English and some other popular languages. In particular, works by Yin et al. [8], Mueller et al. [9], Jiang et al. [10], Zhou et al. [11], among many others, have proposed various methods, ranging from simple string-matching to machine learning and deep learning techniques. In contrast, research on this topic in Vietnamese remains relatively limited, with only two studies conducted by Bach et al. [4] and Nguyen-Son et al. [5].

On the one hand, previous literature on the paraphrase identification task in Vietnamese also depends heavily on the string-based methods. For instance, Bach et al. [4] use nine string-based similarity measures combined with seven-string pairs to represent a sentence. As discussed earlier, this method has proven to be rather ineffective in classifying non-trivial instances. On the other hand, while the deep learning techniques can be applied to Vietnamese, they require an extensive paraphrase corpus, the construction of which demands high costs of human and machinery resources. This creates apparent obstacles for conducting research on paraphrase identification in the language.

To address these problems, in this study, we propose a novel method to identify sentence paraphrases in Vietnamese implementing a combination of pre-trained models such as the Bidirectional Encoder Representations from Transformers (BERT) model [12], XML-R [13] and PhoBERT [14] and linguistic knowledge. The pre-trained models are used as a feature extractor to embed semantic context information in the representation vectors of Vietnamese sentences and help to overcome the lack of paraphrase corpora. Besides, linguistic knowledge also aids in providing additional information for the training process of Siamese architecture. The rest of the paper is organized as follows. We present previous studies that are relevant to the current study in Section 2, and then propose a novel method to identify sentence paraphrases in Vietnamese in Section 3. Section 4 contains our experiments on evaluating the performance of this method. Section 5 concludes the work and discusses future directions.

II. RELATED WORK

Various paraphrase identification and similarity measurement methods have been proposed for a range of languages. The methods can be categorized into four different groups of approaches: string-based, corpus-based, knowledge-based, and hybrid [15]. In this section, we first present the methods laid out in these four approaches and then discuss previous work conducted for the Vietnamese language.

A. String-based Approach

The advantage of this approach lies in its simplicity, as most of the methods are easy to implement. The main information is derived from the text itself, with little to no reliance on additional resources. However, this also lowers the accuracy of the approach, as all of these methods do not detect semantic similarity effectively, thereby failing to account for non-trivial cases, as discussed earlier.

First, among the similarity measures that are widely used across different applications is the Damerau-Levenshtein distance [16]. This measure considers the minimum number of operations needed to convert one text into the other. An operation can be either an insertion, a deletion, a substitution of a single character or a transposition of two consecutive characters.

Secondly, the n-gram comparison of two texts is also considered a common algorithm. An n-gram is a sequence of n elements of a text sample. These n elements can be characters, phonemes, syllables, or words, depending on the tasks and applications. Alberto et al. define the formula to calculate the text similarity value using n-grams as follows [17]:

$$\text{Similarity} = \frac{\text{Number of the same n-grams}}{\text{Total number of n-grams}} \quad (2)$$

Another popular similarity measure in not only this vein of research but also in other fields is the Jaccard similarity index. This measure is calculated by taking the ratio of the number of common words and the total number of distinct words of both texts [7]. Moreover, other methods, such as Euclid, Manhattan, and Cosine, typically represent texts in the form of vectors and then compute text similarity using the distance between these vectors, as shown below:

$$\text{Euclid distance} = \sqrt{\sum_{i=1}^{n}(X_i + Y_i)^2} \quad (3)$$

$$\text{Manhattan distance} = \sum_{i=1}^{n}|X_i - Y_i| \quad (4)$$

$$\text{Cosine similarity} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (5)$$

In all of these formulas, X and Y denote the two representation vectors of two corresponding segments of text.

Furthermore, while these three measures are considered methods within the string-based approach, they are still utilized as objective functions in other methods in other approaches, especially in machine learning models. Given its straightforwardness implementation, the string-based approach can be found in applications that do not strictly rely on paraphrase identification. Since the processing occurs mainly on the input strings, these methods can be extended to the analyses of texts in a broad range of languages, including Vietnamese.

B. Corpus-based Approach

The methods of this approach exploit information from existing corpora to predict the similarity of input texts. The most common method in this approach is the Latent Semantic Analysis (LSA) [18], which assumes that words with similar meanings are co-occurrence in similar text segments. In this method, a matrix that represents the cohesion between words and text segments is first constructed from one or more given corpora. Then, its dimensions are reduced using the Singular Value Decomposition (SVD) technique. Finally, the similarity is calculated by the cosine similarity between the vectors which are the rows of the matrix.

Some methods use online corpora obtained from websites or search engine results. The advantage of these methods is that the extracted information is not only tremendously large, but it is also regularly updated. For instance, the Explicit Semantic Analysis (ESA) method uses Wikipedia articles as a data source to build representation vectors for texts [19]. Likewise, Cilibrasi et al. calculate the text similarity based on the statistics of results from the Google search engine for a given set of keywords [20].

In recent years, the deep learning technique on machine learning models has become more and more popular because of their efficiency in solving classification problems in various fields. In the paraphrase identification task, deep learning on the Siamese architecture for neural networks is the most popular method. The Siamese networks are dual-branch networks that share the same weights and are merged by an energy function. The Siamese architecture can learn the information about the differences between two input samples. Recently, the Siamese LSTM model is a well-known combination. Each input text is fed into an LSTM's sequence. The outputs of the LSTM's sequence are then merged by the Manhattan distance function in [9]. Meanwhile, Neculoiu et al. [2] use another feed-forward neural network which finetunes the output of LSTM layers before they are being merged by the cosine similarity function. Neculoiu et al. also use bi-directional LSTM's sequence to exploit the bi-directional context instead of the single LSTM's sequence as in [9].

The Google AI Research team then proposes the Bidirectional Encoder Representations from Transformers (BERT, 2018) model using Transformers as the model's core [12]. These Transformers are fully connected, which allows it to outperform the state-of-the-art models at that time for some NLP downstream tasks. The model has achieved high results in over six tasks of NLP, including text similarity and paraphrase identification. We implement this BERT model in the proposed method of our study.

The introduction of the BERT model also leads to the emergence of the Siamese BERT model. Reimers et al.'s (2019) Sentence BERT (SBERT) model [21] uses the Siamese architecture to help fine-tune BERT with some corpora, targeting specific tasks to improve sentence representation for each task. The results of this work in downstream tasks are better than those of the representation vectors obtained from BERT.

Based on Transformer-XL, Yang et al.'s (2019) XLNet model [22] is argued to yield better results than BERT. The research team pointed out some shortcomings of the BERT model such as inconsistencies between training and the fine-tuning task and parallel independent word predictions. To overcome these drawbacks, they utilize both Permutation Language Modeling (PLM) and Transformer-XL [23].

Besides the pretrained model BERT, Alexis Conneau et al. (2020) introduce the XML-R model (XML-RoBERTa) [13], which is a generic cross lingual sentence encoder that obtains state-of-the-art results on many cross-lingual understanding (XLU) benchmarks. It is trained on 2.5T of filtered CommonCrawl data in 100 languages, and Vietnamese is one of the supported languages.

Based on RoBERTa, Dat Quoc Nguyen and Anh Tuan Nguyen (2020) introduce the PhoBERT model [14]. PhoBERT outperforms previous monolingual and multilingual approaches, obtaining new state-of-the-art performances on four downstream Vietnamese NLP tasks of Part-of-speech tagging, Dependency parsing, Named-entity recognition and Natural language inference.

While this is a potential approach for Vietnamese, the lack of high-quality and large corpora remains an obstacle to adopt these methods to the language.

C. Knowledge-based Approach

Methods in this approach exploit the linguistic knowledge from knowledge bases such as semantic networks, ontology, etc. WordNet [24], the most popular semantic network, is often used to extract linguistic knowledge at the lexical level to recognize the similarity between texts. Meanwhile, BabelNet [25] is a new semantic network that covers 284 languages. The main disadvantage that comes with Babelnet is that it only provides API in Java in its free edition.

There are six semantic measures, three of which are based on information content, while the remaining three are based on the connection length in the network. The former measures are proposed in Resnik (res) [26], Lin (lin) [27], and Jiang & Conrath (in) [28], while the latter ones can be found in Leacock & Chodorow (lch) [29], Wu & Palmer (wup) [30], and Path Length (path). These measures are slightly different but can be interchangeable. Path Length is the most commonly used measure.

With the work of Le et al. [31] and BabelNet, we can apply this approach to Vietnamese. However, the Vietnamese semantic networks are not complete and still being updated, implying inconsistent results that would be yielded from the implementation of this approach alone.

D. Hybrid Approach

Mihalcea et al. (2006) combine two methods of the corpus-based approach with six measures of the knowledge-based approach to computing text similarity [32]. The results of the combination are better than those of each of the methods. Meanwhile, Li et al. (2006) calculate text similarity using the semantic vectors built from WordNet and Brown corpus [33]. Besides, the representation vector for word order also involves in the process of calculating the similarity of two sentences.

E. Vietnamese Sentence Paraphrase Identification

The work of Bach et al. in 2015 is among the first attempts to solve the paraphrase identification task for Vietnamese [4]. The important contribution of this work is the Vietnamese paraphrase corpus, vnPara, which is the first Vietnamese paraphrase corpus. The corpus is used to evaluate their proposed method, which is to construct a text representation vector from the combination of multiple similarity measures in the string-based approach for syntactic units such as words, syllables, part-of-speech (POS), nouns, verbs, etc. After this combination of measures and syntactic units with four machine learning methods, Bach et al. has achieved the highest results with the Support vector machine (SVM) when combining nine measures with seven syntax units.

Then, Nguyen-Son et al. (2018) propose a method that matches duplicate phrases and similar words [5]. First, this method matches all identical substrings of two sentences, and then eliminates stop words. Afterwards, WordNet is utilized to calculate the similarity for the remaining words. The experimental results of this method reveal that the vnPara corpus contains multiple paraphrase pairs that have a high rate of word overlap. Therefore, Nguyen-Son et al. introduce the construction of a new corpus, VNPC, which is argued to be more diverse than vnPara. In summary, most research on paraphrase identification for Vietnamese still rely heavily on the string-based approach, which is not ineffective in detecting semantic paraphrase identification.

III. Proposed Method

Since deep learning methods often require large corpora, the lack of Vietnamese paraphrase corpora creates challenges to researchers who plan to apply this technique to the language. Recently, the emergence of pre-trained models helps researchers overcome this obstacle. The pre-trained models such as BERT, XML-R and PhoBERT are most popular pre-trained models, especially for Vietnamese. Therefore, we take this advantage to construct our method.

Even though in theory, pre-trained models can effectively solve the paraphrase identification task for Vietnamese, we expect that this task can be improved with the addition of linguistic knowledge during the process. Devlin et al. [12] state that there are three ways to improve BERT, which are pre-training from scratch, fine-tuning the pre-trained model, and utilizing BERT as a feature extractor. However, linguistic knowledge cannot be used in the fine-tuning process, and training BERT and other pre-trained models from scratch is extremely costly. Therefore, feature extraction is the most plausible way to implement BERT in our method.

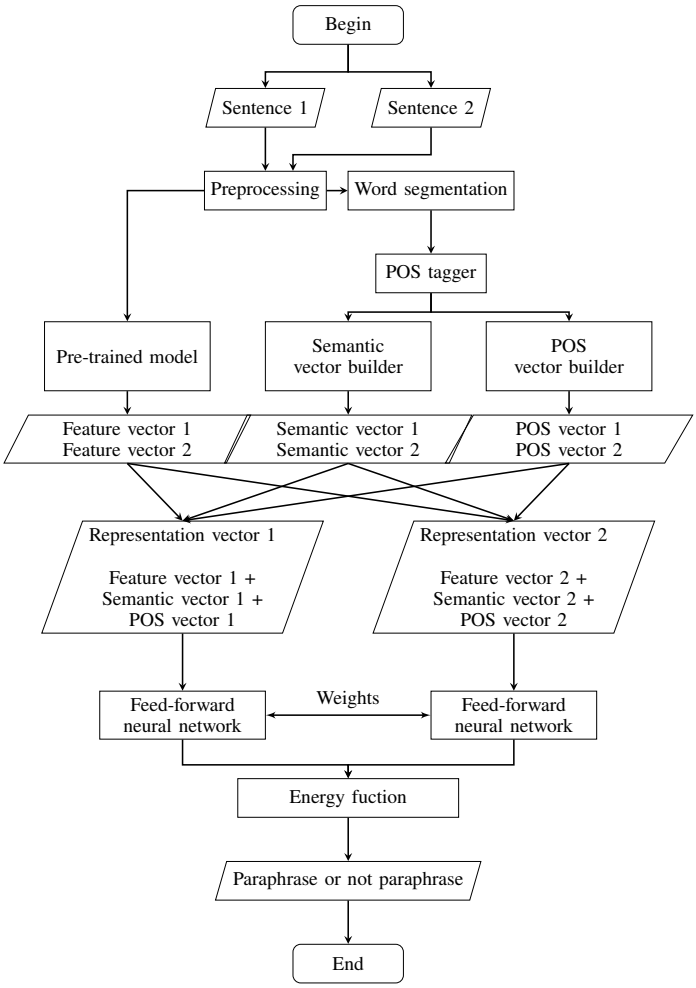

Fig. 1. Overview of the Proposed Method to Identify Vietnamese Sentence Paraphrase.

The proposed method follows a hybrid approach. In particular, it is a combination of the corpus-based approach and the knowledge-based approach to fully exploit the information gained from these two approaches.

We built three vectors for each input sentence:

- Feature vector achieved from pre-trained model.

- Semantic vector constructed by using WordNet.

- POS vector represents the POS of words in a sentence.

These three vectors then were joined together to form a sentence representation vector. There were two such vectors for two input sentences. These vectors were fed into a Siamese feed-forward neural network to train or predict. The overview of the proposed method is depicted in Fig. 1.

A. Preprocessing

The input pairs of sentences before being put into the main processing chain were normalized by regular expressions and Heuristic rules.

B. Features Extraction Using Pre-trained Model

Heretofore, simple word embedding models such as Word2Vec [34], GloVe [35], FastText [36], etc. are common methods used by many research groups to represent text in vector form. However, these models represent every word with a unique vector in all contexts. In contrast, the feature vector constructed from pre-trained model contains full information of the bi-directional context, thanks to Transformer blocks' multi-head self-attention mechanisms and fully connected architecture.

The features extracted from pre-trained model were the output of the Transformer blocks. For instance, BERT-Base with 12 Transformer blocks provided 12 real vectors for each token and BERT-Large with 24 Transformer blocks provided 24 vectors. The dimension of each vector was the number of hidden units of each layer. There were 768 dimensions for BERT-Base and 1,024 dimensions for BERT-Large.

C. Semantic Vector Construction

We followed the method of [33] to construct semantic vectors for a sentence pair. A semantic vector contained information on the semantic relatedness of the words in these sentences. These vectors were constructed by using a semantic network and statistical information of a corpus. In our work, we use Vietnamese WordNet which was constructed by Le et al. in 2016 [31] and the statistical information from [37].

From the list of words of the sentences pair, a set of unique words was constructed. The order of these words was preserved in the order of the words in the sentences.

Let M be a two-dimensional matrix containing the relatedness of each pair of words. The matrix M has n rows corresponding to n words of the considered sentence and m columns corresponding to m words in the set of unique words. The relatedness between w1 (line r) and w2 (column c) is calculated using the formula:

$$\text{M(r,c)} = \begin{cases} 1 & \text{if w1=w2} \\ \text{PathLength(w1,w2)} & \text{if w1} \neq \text{w2} \\ 0 & \text{if w1 or w2 is not in Wordnet} \end{cases} \quad (6)$$

where WordNet.PathLength(w1, w2) is the Path Length similarity in WordNet of word w1 and word w2. Each element of the lexical vector s is the maximum value on a column of the matrix M:

$$\text{s[c]} = M_{(r,c)} \text{ , c = [1,m]} \quad (7)$$

Finally, the semantic vector semVec is calculated using the formula:

$$\text{semVec[c]} = \text{s[c]} \times \text{I(W1)} \times \text{I(W2)} \quad (8)$$

where W1 and W2 are words that have the greatest relatedness s[c] on column c; I(W1) and I(W2) are the information content of the two corresponding words. The information content I(w) of word w is calculated by the frequency of w in a corpus:

TABLE IV. PROCESS OF CONSTRUCTING THE SEMANTIC VECTOR FOR THE FIRST SENTENCE

	anh	**ấy**	**là**	**giáo_viên**	**nhà_giáo**
anh (he)	1				
ấy (he)		1			
là (is)			1		
giáo_viên (teacher)				1	0.33
s	1	1	1	1	0.33
Weights	I(anh)	I(ấy)	I(là)	I(giáo_viên)	I(giáo_viên)
	I(anh)	I(ấy)	I(là)	I(giáo_viên)	I(nhà_giáo)
Semantic vector	0.1026	0.1474	0.0654	0.2597	0.1065

$$I(w) = -\frac{loglogp(w)}{loglog(N+1)} = 1 - \frac{loglog(n+1))}{loglog(N+1)} \quad (9)$$

where p(w) is the relative frequency of w in a corpus, N is the number of words in the corpus and n is the frequency of the word w in the corpus.

Table IV shows the process of constructing the semantic vector for the first sentence in this sentences pair:

- Sentence 1: anh ấy là giáo_viên (he is a teacher)

- Sentence 2: anh ấy là nhà_giáo (he is an educator)

The semantic vector must be padded with zero-value to have a fixed length. According to the statistics in [37] about the average length of sentence (in words), we assume that the longest sentence may have a length of 50 words. Thus, we construct the semantic vector with a fixed length of 100.

D. Parts-of-speech (POS) Vector Construction

WordNet only accepts four simple parts-of-speech which are noun, verb, adjective, and adverb so that the semantic vector does not contain full information of the sentence's parts-of-speech. Therefore, we also used the POS vector to provide more information to the model. To construct a POS vector, each word in a sentence was tagged with its part-of-speech to create a list of parts-of-speech for each sentence. These POS lists were then represented as real vectors by using the FastText model [36]. To train this model, we used the Vietnamese Treebank corpus [38] with 10,000 POS tagged sentences. An output vector of the FastText model had a fixed length of 100.

E. Siamese Feed-forward Neural Network (SFFNN)

For each input sentence, the feature vector obtained from pre-trained model, the semantic vector, and the POS vector were concatenated to form the representation vector (Fig. 3). Sentence paraphrase identification task has an input of two sentences. Therefore, we generated two representation vectors.

To make the neural network learn the similarity between two sentences, we applied the Siamese architecture to the feed-forward neural network. Fig. 2 depicts a Siamese feed-forward neural network. The feed-forward neural network was constructed by multiple dense (fully connected) layers. The number of layers and hidden units will be presented in

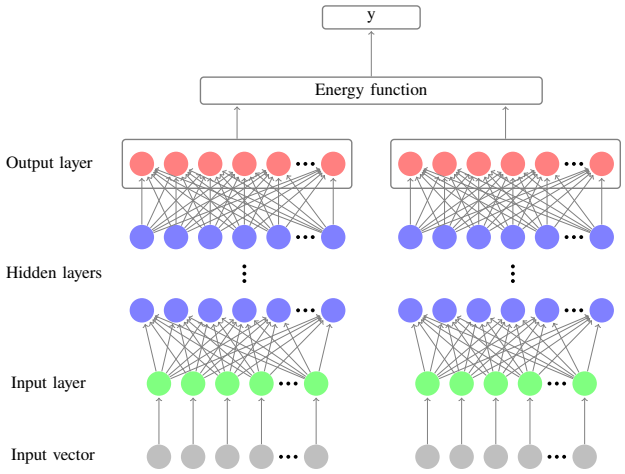

Fig. 2. Overview of the Proposed Method to Identify Vietnamese Sentence Paraphrase.

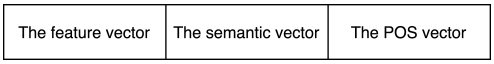

Fig. 3. Forming the Representation Vector.

subsection 4.2.2. The activation function of hidden layers. The input layer was the ReLU function, and the output layer was the sigmoid function. The ReLU function is often used as the activation function in hidden layers due to its simplicity. Besides, the constant gradient of ReLUs results in faster learning and allows ReLU to overcome the defect of sigmoid function when the absolute value of the input is great.

The neural network was trained using the backpropagation algorithm and the training only stopped when the value of the energy function no longer changed. We used the Mean Squared Error (MSE) function as an error function for the gradient descent method. There are three similarity functions commonly used as energy functions in the text similarity task. They are Euclid, Manhattan, and Cosine. The study of Chopra et al. [39] shows that the Euclid and Cosine functions using the normalized function l2 instead of l1 in the similarity function can lead to undesirable plateaus in the overall objective function. Therefore, we used the Manhattan similarity function as the energy function in our model.

IV. Experiment

A. Corpora

The experiments in this paper were conducted on two main corpora: vnPara [4] and VNPC [5].

1) vnPara: VnPara has become a common Vietnamese paraphrase corpus in various studies [5][6]. To construct vnPara corpus, Bach et al. [4] first collected articles from online newspaper websites such as dantri.com.vn, vnexpress.net, thanhnien.com.vn, and so on. As shown in Table V, sentences extracted from the articles were paired if they have multiple words in common. These sentence pairs were labeled manually by two people.

TABLE V. Examples from vnPara and their Translation into English

Sentences pair		Is Paraphrase
ASA sẽ tìm thấy người ngoài hành tinh trong 20 năm tới.	ASA nói có thể sẽ tìm thấy người ngoài hành tinh trong 20 năm tới.	1
ASA will find aliens in the next 20 years.	ASA says that it's possible to find aliens in the next 20 years	
Các gián điệp Trung Quốc đã tấn công hệ thống mạng của một nước thuộc khu vực Đông Nam Á.	Các gián điệp Trung Quốc đã tấn công hệ thống mạng của một tổ chức nghiên cứu lớn của chính phủ Canada, giới chức Canada ngày 29/7 cho biết.	0
Chinese spies have attacked the network system of a country in Southeast Asia.	Chinese spies have attacked the network of a prominent Canadian government research organization, the Canadian officials say on July 29.	
Bà đã cho ra đời 15 cuốn tiểu thuyết, nhiều tập truyện ngắn và các bài bình luận văn học.	Trong suốt sự nghiệp của mình, bà đã sáng tác 15 tiểu thuyết, nhiều truyện ngắn và nhận được gần 20 giải thưởng văn học lớn.	1
She has published 15 novels, many short stories, and literary studies.	In her entire career, she has written 15 novels, many short stories and achieve about 20 major literary awards.	

TABLE VI. Examples from VNPC and their Translation into English

Sentences pair		Is Paraphrase
Đáng chú ý, mã độc này chưa hoạt động mà ở chế độ "ngủ đông", chờ lệnh tấn công.	Đáng chú ý mã độc này chưa hoạt động mà ở chế độ nằm vùng.	1
Remarkable, this malware has not been working yet but is in "hibernate" mode, wait for an attack.	Remarkable, this malware has not been working yet but is in "stand-by" mode.	
Trần Thị Thu Ngân đăng quang ngôi vị cao nhất của cuộc thi Hoa hậu Bản sắc Việt toàn cầu 2016	Trần Thị Thu Ngân đăng quang Hoa hậu Bản sắc Việt toàn cầu 2016	1
Tran Thi Thu Ngan crowned the highest position in Miss Vietnam Global Heritage 2016	Tran Thi Thu Ngan crowned Miss Vietnam Global Heritage 2016	
Cầu thủ trẻ đắt giá thứ ba mà Real từng đào tạo là Alvaro Negredo.	Một số cầu thủ khác từng trưởng thành từ lò đào tạo trẻ của Real là Cheryshev, Joselu, Diego Lopez và Rodrigo Moreno	0
The third most valuable young player who Real has trained is Alvaro Negredo.	Some other players who have grown up at Real's youth academy are Cheryshev, Joselu, Diego Lopez and Rodrigo Moreno.	

2) VNPC: VNPC was constructed by Nguyen-Son et al. when they experimented with their proposed method in [5]. According to their experiment result, VNPC was argued to be more diverse than vnPara.

To build this corpus, first of all, the pairs of sentences were extracted from 65,000 pages of 15 Vietnamese news websites. Nguyen-Son et al. used their proposed method to measure the similarity of the obtained pairs. 3,134 candidates were selected using a predefined threshold. As shown in Table VI, these sentences formed paraphrase pairs, which were manually labeled.

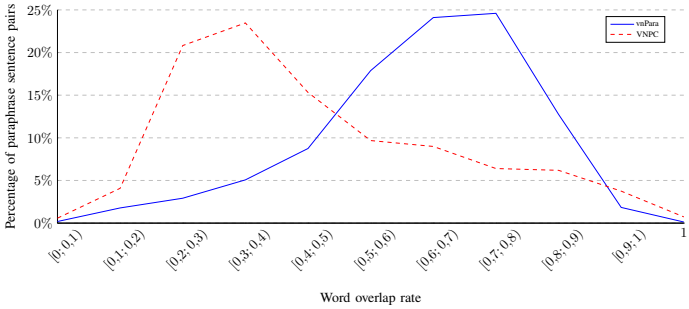

Fig. 4. Distribution of Paraphrase Cases of vnPara and VNPC according to the Word Overlap Ratio.

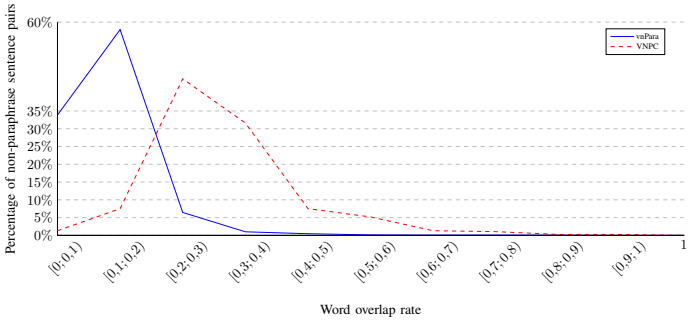

Fig. 5. Distribution of Non-paraphrase Cases of vnPara and VNPC according to the Word Overlap Ratio.

3) Some Properties of the Two Corpora:

a) Number of sentence pairs per class.: The number of sentence pairs of the vnPara corpus is 3,000 and the number of sentence pairs of the VNPC is 3,134. The VNPC corpus contains 2,748 paraphrase pairs and 386 non paraphrase pairs. Meanwhile, the vnPara corpus has the same number of paraphrase pairs and non paraphrase pairs is 1,500 sentence pairs.

b) Number of non-trivial sentence pairs.: Research by Yui Suzuki et al. [1] shows that the importance of non-trivial paraphrase or non- paraphrase sentence pairs. The authors define that a non-trivial paraphrase sentence pair is a paraphrase sentence pair with a small ratio of words overlap (WOR) between two sentences. On the contrary, a non-trivial non- paraphrase sentence pair is non-paraphrase sentence pair with a large ratio of words overlap between two sentences. We have made statistics on the rate of word overlap of sentence pairs in both corpora. The word overlap rate is calculated using the formula for calculating the Jaccard index.

Fig. 4 shows that the VNPC corpus contains more non-trivial paraphrase sentence pairs than vnPara. Fig. 5 shows that both corpora contain very few non-trivial non-paraphrase sentence pairs. The vnPara corpus almost does not contain any paraphrase sentence pair with a overlap rate from 0.5.

B. Experimental Setup

1) Evaluation Method: To compare the result of our method with the results of the vnPara and VNPC studies, we conducted the experiment in the same manner. Each corpus was divided into 5 folds randomly to perform a 5-fold cross validation test. We also used the same metrics which were accuracy and F1 score as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Recall = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision + Recall}} \tag{13}$$

where TP is true positive (a correct prediction of paraphrase), TN is true negative (a correct prediction of non-paraphrase), FP is false positive (a wrong prediction of paraphrase), and FN is false negative (a wrong prediction of non-paraphrase).

2) Configuration of Feed-forward Neural Network: The configuration of feed-forward neural network includes 12 hidden layers, 768 hidden units. We chose this configuration for experiments on the vnPara corpus and VNPC corpus.

C. Experimental Results

The experiments with our proposed method were performed with some different configurations of stop words, BERT's output layer, and BERT's output pooling strategies. We achieved the best result when testing our method with the configuration in which we kept stop words, used the second-to-last layer of pre-trained output, and utilized an average pooling strategy to get the feature vector. When experimenting with the Siamese LSTM model in the article [9], we used the pre-trained Vietnamese Word2Vec model of Vu et al. [40].

Tables VII and VIII show the results of experiments we conducted on vnPara and VNPC with several methods. The result of each method is presented by each row in the tables. Each method is evaluated by the accuracy and the F1 score. Each table shows the available results from previous studies for Vietnamese, the results of the Siamese LSTM model [9], the results of original pre-trained models, and the results of our method. The results of our method are presented in three rows according to three different configurations of additional vectors: adding semantic vector, adding the POS vector, and adding both semantic vector and POS vector.

We also compute the F1 score on each word overlap rate range with the proposed method as figures similar to the 4 and 5. The calculation of the F1 score is divided into two cases: paraphrase cases and non-paraphrase cases to assess the effect of non-trivial cases on model training.

Fig. 6 shows that the proposed method results above 80% on all word overlap rate ranges in the VNPC corpus. For the vnPara corpus, the proposed method's result is below 80% for sentence pairs with a word overlap rate of 0.3 or less. In the word overlap rate range [0.1; 0.2), the proposed method

TABLE VII. EVALUATION RESULTS OF DIFFERENT METHODS ON THE VNPARA CORPUS

Method	Accuracy (%)	F1 score (%)
vnPara [4]	89.10	86.70
Siamese LSTM [9]	65.64	64.29
BERT	73.14	73.56
Our method (using BERT)		
Feature vector (BERT) + Semantic vector	94.12	94.28
Feature vector (BERT) + POS vector	72.45	72.22
Feature vector (BERT) + Semantic vector + POS vector	**94.27**	**94.38**
XLM-R	74.57	75.22
Our method (using XLM-R)		
Feature vector (XLM-R) + Semantic vector	93.58	93.76
Feature vector (XLM-R) + POS vector	75.12	75.51
Feature vector (XLM-R) + Semantic vector + POS vector	**93.67**	**93.85**
PhoBERT	76.33	75.80
Our method (using PhoBERT)		
Feature vector (PhoBERT) + Semantic vector	94.71	94.83
Feature vector (PhoBERT) + POS vector	75.83	75.20
Feature vector (PhoBERT) + Semantic vector + POS vector	**94.86**	**94.97**

TABLE VIII. EVALUATION RESULTS OF DIFFERENT METHODS ON THE VNPC CORPUS

Method	Accuracy (%)	F1 score (%)
Matching duplicate phrases and similar words [5]	87.68	Not available
Siamese LSTM [9]	65.64	64.29
BERT	86.85	92.77
Our method (using BERT)		
Feature vector (BERT) + Semantic vector	86.72	92.74
Feature vector (BERT) + POS vector	86.14	92.37
Feature vector (BERT) + Semantic vector + POS vector	**87.05**	**92.90**
XLM-R	86.61	92.65
Our method (using XLM-R)		
Feature vector (XLM-R) + Semantic vector	86.80	92.77
Feature vector (XLM-R) + POS vector	87.28	93.01
Feature vector (XLM-R) + Semantic vector + POS vector	**87.70**	**93.30**
PhoBERT	86.97	92.64
Our method (using PhoBERT)		
Feature vector (PhoBERT) + Semantic vector	87.45	93.12
Feature vector (PhoBERT) + POS vector	86.39	92.25
Feature vector (PhoBERT) + Semantic vector + POS vector	**88.02**	**93.49**

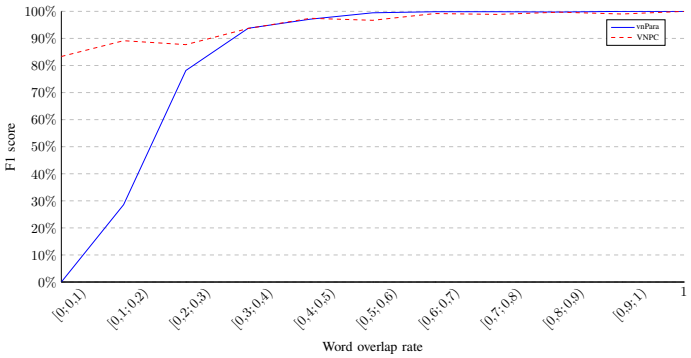

Fig. 6. F1 Score according to the Word Overlap Rate of Paraphrase Cases in vnPara and VNPC.

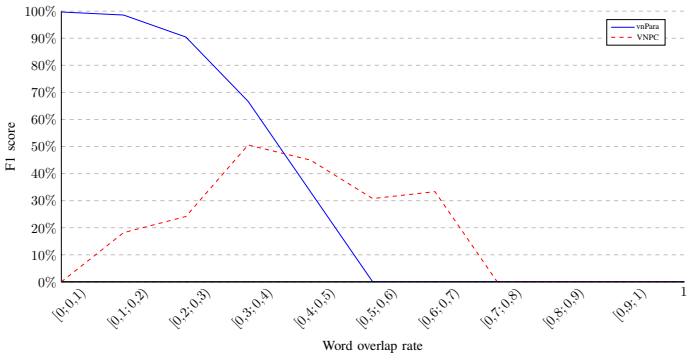

Fig. 7. F1 Score according to the Word Overlap Rate of Non-paraphrase Cases in vnPara and VNPC.

achieves F1 score of 28.57% for the vnPara corpus and 89.16% for the VNPC corpus.

In general, the F1 score is in Fig. 7 for the non-paraphrase cases of the VNPC is lower than vnPara corpus, due to the small number of non-paraphrase cases compared with the paraphrase cases in the VNPC corpus. The F1 score has a value of 0 in the range of overlap [0.7; 1] for the VNPC corpus and in [0.5; 1] for the vnPara corpus. This is because both corpora almost do not contain non-paraphrase cases in these two ranges. At the word overlap rate [0; 0.1), the test with the VNPC is 0% and the test with the vnPara reaches 99.71%.

To further demonstrate the universality of the proposed method's improvement over the pre-trained model, an experiment was performed on another corpus. Apart from the vnPara and the VNPC, almost no Vietnamese paraphrase corpora has been published. Therefore, the proposed model will be tested further on a Vietnamese translation of a well-known paraphrase corpus MSRP [41]. The evaluation results are shown in Table IX.

D. Discussion

The Siamese LSTM model produces mediocre results, because the training process of this model requires great corpora. For English, this model is trained with over 300,000 sentence pairs and achieves an accuracy rate of 82.29%. Meanwhile, existing Vietnamese paraphrase corpora contain only about 3,000 pairs of sentences.

TABLE IX. EVALUATION RESULTS ON THE VIETNAMESE-TRANSLATED MSRP

Method	Accuracy (%)	F1 (%)
Siamese LSTM [9]	64.17	73.36
BERT	64.15	74.33
Our method (using BERT as feature vector)	**65.68**	**75.20**
XLM-R	65.89	76.06
Our method (using XLM-R as feature vector)	**66.42**	**76.09**
PhoBERT	59.90	67.10
Our method (using PhoBERT as feature vector)	**65.27**	**75.09**

The results show that our method achieves the best accuracy when using both semantic vector and POS vector. It outperforms the previous methods for Vietnamese paraphrase identification, also the Siamese LSTM model and the pre-trained model. The F1 score is much higher than the result of the pre-trained models in an experiment on VNPC. This proves that our method is more suitable for the Vietnamese paraphrase identification task that focuses more on paraphrase.

The number of duplicate sentences has a certain influence on the results of the proposed method. The number of duplicate sentences of VNPC is more than twice the number of duplicate sentences of vnPara. This means that the sentence diversity of vnPara is higher than that of VNPC, affecting the process of training deep learning models. This is also one of the reasons for our proposed method to achieve higher F1 score on the vnPara than on the VNPC when considering the paraphrase cases.

Fig. 6, 7 and the descriptions of these two figures partly show the importance of the non-trivial sentence pairs on the training process of the proposed method. The F1 score of our proposed method for paraphrase cases does not have great variation across all word overlap ranges for the VNPC, even though this corpus contains very few paraphrase sentence pairs in the range [0; 0.2) and [0.7; 1]. For the non-paraphrase cases having the word overlap rate is in the range [0.7; 1], our proposed method could not detect these cases on both vnPara and VNPC. Fig. 5 clearly shows the lack of non-trivial paraphrase cases on both corpora. Thus, it can be seen that the properties of the two corpora greatly affect the results that the proposed method achieves.

With testing on the Vietnamese-translated MSRP corpus, the results obtained from the proposed method are still higher than the results of the pre-trained models. Meanwhile, the results with the F1 score of our proposed method are much better. This shows that our proposed method still achieves higher results than the original pre-trained models when processing translated documents.

Although we achieve good results with our method, the model itself contains some disadvantages. First of all, the model requires big resources to operate, so it is not ready to work in practice. To build the semantic vector, the model depends much on the POS tagger. The mistakes of the POS tagger can entail the mistakes when building the semantic vector. The pre-trained models are used passively, not yet involved in the training process. Therefore, they have not been best exploited for this task.

V. CONCLUSION AND FUTURE WORK

The paraphrase identification task is a crucial core task of several NLP tasks and applications. There are various studies for popular languages but a few for Vietnamese. The great challenge in the research for the Vietnamese paraphrase identification task is the lack of good and large corpora. The emergence of the pre-trained models enable us to propose a novel method that does not require large corpora for training but is still highly effective. The proposed method uses three vectors: feature vector achieved from pre-trained model, semantic vector constructed by using WordNet, POS vector represents the POS of words in a sentence. They are joined to form a sentence representations vector that contains rich context information. Explicit linguistic knowledge helps the method yield 94.97% F1-score on the VnPara corpus and 93.49% F1-score on the VNPC corpus, which is better than the pre-trained models for the paraphrase identification task in Vietnamese. These results also show that using a pre-trained model is a feasible way for studies of text similarity as well as other NLP tasks in resource-poor languages such as Vietnamese.

Although the method proposed in this paper achieves positive results, we realize that there are still potential improvements to achieve better results. We plan to fine-tune the pre-trained models in the training process to make the pre-trained models learn information from the input samples to have better sentence representation vectors. Using linguistic knowledge has proved to be effective. However, the resources for the proposed method to work are quite high. Hence, we need to create a method that uses fewer resources but still guarantees high accuracy rates.

Whereas the proposed method can solve the corpora lacking problem for deep learning, it is still necessary to have Vietnamese paraphrase corpora for fine-tuning, improvement, or evaluation. Meanwhile, the two existing Vietnamese paraphrase corpora still have some shortcomings such as class imbalance, the lack of non-trivial instances, and various duplicate sentences, etc. Therefore, the need to construct good-quality Vietnamese paraphrase corpora remains as pressing as ever.

ACKNOWLEDGMENT

This research is funded by University of Science, VNU-HCM under grant number CNTT 2020-07. This research is supported by Computational Linguistics Center, University of Science, Vietnam National University, Ho Chi Minh City, and uses the Vietnamese Treebank, which was developed by the VLSP project.

REFERENCES

[1] Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. *Building a Non-Trivial Paraphrase Corpus Using Multiple Machine Translation Systems*. In: (2017), pp. 36 42. DOI: 10.18653/v1/p17-3007.

[2] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. *Learning Text Similarity with Siamese Recurrent Networks*. In: Proceedings of the 1st Workshop on Representation Learning for NLP. 2016, pp. 148 157. DOI: 10.18653/v1/w161617.

[3] Ming Liu, Bo Lang, and Zepeng Gu. *Calculating Semantic Similarity between Academic Articles using Topic Event and Ontology*. In: 37 (2017), pp. 1 21. arXiv: 1711.11508. URL: http://arxiv.org/abs/1711.11508.

[4] Ngo Xuan Bach et al. *Paraphrase Identification in Vietnamese Documents*. In: 2015 IEEE International Conference on Knowledge and Systems Engineering, KSE 2015. 2015, pp. 174 179. ISBN: 9781467380133. DOI: 10.1109/KSE. 2015.37.

[5] Hoang-Quoc Nguyen-Son et al. *Vietnamese Paraphrase Identification Using Matching Duplicate Phrases and Similar Words*. In: Future Data and Security Engineering. Vol. 11251. Springer International Publishing, 2018, pp. 172 182. ISBN: 978-3-319-70003-8. DOI: 10.1007/978-3-319-70004-5. URL: http: //link.springer.com/10.1007/978-3-319-70004-5.

[6] Dien Dinh, Nguyen Le Thanh. *English–Vietnamese cross-language paraphrase identification using hybrid feature classes*. In: J Heuristics (2019). DOI: 10.1007/s10732-019-09411-2

[7] Paul Jaccard. *Étude comparative de la distribution florale dans une portion des Alpes et du Jura*. In: Bulletin de la Société Vaudoise des Sciences Naturelles 142.37 (1901), pp. 547 579. DOI: 10.5169/seals-266450. URL: http: //dx.doi.org/10.5169/seals-266450.

[8] Wenpeng Yin and Hinrich Schutze. *Convolutional Neural Network for Paraphrase Identification*. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 901 911. DOI: 10.3115/v1/N15-1091. URL: https://www. aclweb.org/anthology/N15-1091.

[9] Jonas Mueller and Aditya Thyagarajan. *Siamese recurrent architectures for learning sentence similarity*. In: 30th AAAI Conference on Artificial Intelligence, AAAI 2016 2014 (2016), pp. 2786 2792.

[10] Y. Jiang, Y. Hao, and X. Zhu. *A Chinese text paraphrase detection method based on dependency tree*. In: 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC). 2016, pp. 1 5. DOI: 10.1109/ ICNSC.2016.7479003.

[11] Jie Zhou, Gongshen Liu, and Huanrong Sun. *Paraphrase Identification Based on Weighted URAE, Unit Similarity and Context Correlation Feature*. In: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26 30, 2018, Proceedings, Part II. DOI: 10.1007/978-3-319-99501-4_4.

[12] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In: Mlm (2018). arXiv: 1810.04805. URL: http: //arxiv.org/abs/1810.04805.

[13] Alexis Conneau et al. *Unsupervised Cross-lingual Representation Learning at Scale*. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.747.

[14] Dat Quoc Nguyen and Anh Tuan Nguyen. *PhoBERT: Pre-trained language models for Vietnamese*. In: Findings of the Association for Computational Linguistics: EMNLP 2020 (2020), p. 1037–1042.

[15] Wael H.Gomaa and Aly A. Fahmy. *A Survey of Text Similarity Approaches*. In: International Journal of Computer Applications 68.13 (2013), pp. 13 18. DOI: 10.5120/11638-7118.

[16] Iskandar Setiadi. *Damerau-Levenshtein Algorithm and Bayes Theorem for Spell Checker Optimization*. In: Bandung Institute of Technology November (2013), pp. 1 6. DOI: 10.13140/2.1.2706.4008.

[17] Alberto Barrón-Cedeño et al. *Plagiarism detection across distant language pairs*. In: Coling 2010 - the 23rd International Conference on Computational Linguistics. Vol. 2. August. 2010, pp. 37 45.

[18] Thomas K Landauer and Susan T. Dutnais. *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge*. In: Psychological Review. Vol. 104. 2. 1997, pp. 221 240.

[19] Evgeniy Gabrilovich and Shaul Markovitch. *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. In: IJCAI International Joint Conference on Artificial Intelligence. Vol. 6. 2007. ISBN: 9783642407215. DOI: 10.1007/978-3-642-40722-2_7.

[20] Rudi L. Cilibrasi and Paul M.B. Vitányi. *The Google similarity distance*. In: IEEE Transactions on Knowledge and Data Engineering 19.3 (2007), pp. 370 383. ISSN: 10414347. DOI: 10.1109/TKDE.2007.48. arXiv: 0412098 [cs].

[21] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In: (2019). arXiv: 1908.10084. URL: http: //arxiv.org/abs/1908.10084.

[22] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. In: (2019), pp. 1 18. arXiv: 1906.08237. URL: http: //arxiv.org/abs/1906.08237.

[23] Zihang Dai et al. *Transformer-XL: Attentive Language Models beyond a Fixed-Length Context*. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, pp. 2978 2988. DOI: 10.18653/ v1/p19-1285. arXiv: 1901.02860.

[24] George A. Miller et al. *Introduction to wordnet: An on-line lexical database*. In: International Journal of Lexicography 3.4 (1990), pp. 235 244. ISSN: 09503846. DOI: 10.1093/ijl/3.4.235.

[25] Roberto Navigli and Simone Paolo Ponzetto. *BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*. In: Artificial Intelligence 193 (2012), pp. 217 250. ISSN: 00043702. DOI: 10.1016/j.artint.2012.07.001. URL: http://dx.doi.org/10.1016/ j.artint.2012.07.001.

[26] Philip Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. In: Proceedings of the 14th international joint conference on Artificial intelligence. Vol. 1. 1995, pp. 448 453. arXiv: 9511007 [cmp-lg]. URL: http://arxiv.org/abs/cmp-lg/9511007.

[27] Dekang Lin. *Extracting Collocations from Text Corpora*. In: st Workshop on Computational Terminology, Computerm '98. 1998, pp. 57 63.

[28] Jay J. Jiang and David W. *Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy*. In: Proceedings of International Conference Research on Computational Linguistics. 1997, pp. 19 33. DOI: 10.1152/ ajplegacy.1959.196.2.457.

[29] Claudia Leacock and Martin Chodorow. *Combining Local Context and WordNet Similarity for Word Sense Identification*. In: Computational Linguistics Special issue on word sense disambiguation 24.1 (1998), pp. 147 165. DOI: 10.7551/mitpress/7287.003.0018.

[30] Zhibiao Wu and Martha Palmer. *Verbs semantics and lexical selection*. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Las Cruces, New Mexico, 1994, pp. 133 138. DOI: 10.1152/ajplung. 1998.274.3.l351.

[31] Le Anh Tu and Tran Van Tri. *Exploiting English-Vietnamese OALD resource and applying in automatically translating WordNet into Vietnamese*. In: The 6th Young Scientists Conference. 2016, pp. 29 36.

[32] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity*. In: The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference. Vol. 1. 2006, pp. 775 780. ISBN: 1577352815.

[33] Yuhua Li et al. *Sentence similarity based on semantic nets and corpus statistics*. In: IEEE Transactions on Knowledge and Data Engineering 18.8 (2006), pp. 1138 1150. ISSN: 10414347. DOI: 10.1109/TKDE.2006.130.

[34] Tomas Mikolov et al. *Efficient estimation of word representations in vector space*. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings (2013), pp. 1 12. arXiv: 1301.3781.

[35] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. *GloVe: Global Vectors for Word Representation*. In: Empirical Methods in Natural Language Processing (EMNLP). 2014, pp. 1532 1543. URL: http://www. aclweb.org/anthology/D14-1162.

[36] Piotr Bojanowski et al. *Enriching Word Vectors with Subword Information*. In: Transactions of the Association for Computational Linguistics 5 (2017), pp. 135 146. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00051. arXiv: 1607. 04606.

[37] Dien Dinh, Nhung Nguyen Tuyet, and Thuy Ho Hai. *Building a corpus-based frequency dictionary of Vietnamese*. In: The 12th International Conference of The Asian Association for Lexicography. 2018, pp. 72 98.

[38] Phuong Thai Nguyen et al. *Building a large syntactically-annotated corpus of Vietnamese*. In: ACL-IJCNLP 2009 - LAW 2009: 3rd Linguistic Annotation Workshop, Proceedings (2009), pp. 182 185. DOI: 10.3115/1698381.1698416.

[39] Sumit Chopra, Raia Hadsell, and Yann LeCun. *Learning a Similarity Metric Discriminatively, with Application to Face Verification Sumit*. In:

Computer Vision and Pattern Recognition, 2005. Vol. 1. 2005, pp. 539 546. DOI: 10. 1007/BF02407565.

[40] Xuan-Son Vu. *Pre-trained Word2Vec models for Vietnamese*. https://github.com/sonvx/word2vecVN. commit 9bb2d6ffafa4238f82dc6e85e6ade072c8322334. 2016.

[41] W. B. Dolan and C. Brockett. *Automatically constructing a corpus of sentential paraphrases*. In Proceedings of the Third International Workshop on Paraphrasing (IWP 2005), pages 9-16.