EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

ViMedNER: A Medical Named Entity Recognition Dataset for Vietnamese

Pham Van Duong[1,2], Tien-Dat Trinh[2,*], Minh-Tien Nguyen[3], Huy-The Vu[3], Minh-Chuan Pham[3,*], Tran Manh Tuan[4], and Le Hoang Son[5,6]

[1]School of Information Communication and Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
[2]ICT Department, FPT University, Hanoi, Vietnam
[3]Faculty of Information Technology, Hung Yen University of Technology and Education, Hung Yen, Vietnam
[4]Faculty of Computer Science and Engineering, Thuyloi University, Hanoi, Vietnam
[5]VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam
[6]VNU University of Science, Vietnam National University, Hanoi, Vietnam

Abstract

Named entity recognition (NER) is one of the most important tasks in natural language processing, which identifies entity boundaries and classifies them into pre-defined categories. In literature, NER systems have been developed for various languages but limited works have been conducted for Vietnamese. This mainly comes from the limitation of available and high-quality annotated data, especially for specific domains such as medicine and healthcare. In this paper, we introduce a new medical NER dataset, named ViMedNER, for recognizing Vietnamese medical entities. Unlike existing works designed for common or too-specific entities, we focus on entity types that can be used in common diagnostic and treatment scenarios, including disease names, the symptoms of the diseases, the cause of the diseases, the diagnostic, and the treatment. These entities facilitate the diagnosis and treatment of doctors for common diseases. Our dataset is collected from four well-known Vietnamese websites that are professional in terms of drag selling and disease diagnostics and annotated by domain experts with high agreement scores. To create benchmark results, strong NER baselines based on pre-trained language models including PhoBERT, XLM-R, ViDeBERTa, ViPubMedDeBERTa, and ViHealthBERT are implemented and evaluated on the dataset. Experiment results show that the performance of XLM-R is consistently better than that of the other pre-trained language models. Furthermore, additional experiments are conducted to explore the behavior of the baselines and the characteristics of our dataset.

Received on 28 February 2024; accepted on 29 May 2024; published on 11 July 2024

Keywords: Named entity recognition, Vietnamese corpus, Medical text, Pre-trained language model

doi:10.4108/eetinis.v11i3.5221

1. Introduction

Named entity recognition (NER) is a fundamental and important natural language processing (NLP) task [1–4]. The task is to assign a label (tag) of an entity to a token or a set of tokens in a sequence. For example, several common and fine-grained entity types, e.g., locations or disease names can be extracted from an input sequence by applying NER models [5, 6]. In actual scenarios, NER is not a standalone solution and is usually a part of AI pipelines that can serve for several applications such as text understanding [7, 8], information retrieval [9], text summarization [10], question answering [11], machine translation [12], and knowledge base construction [13].

With the success of NER in common domains, e.g., news, NER has been adapted to medical text [6, 14–18]. Instead of focusing on common entities, e.g., locations or person names, NER for biomedical text extracts specific and more fine-grained entities, e.g., drug names, diseases, or genes. The outputs of bio NER

*Corresponding authors. Email: dattt67@fe.edu.vn, chuanpm@utehy.edu.vn

play an important role in bioinformatics and can be served as input for several downstream tasks such as clinical assertion status [19], clinical entity resolvers [20], de-identification of the sensitive data [21], or PICO (Populations, Interventions, Comparators, and Outcomes) extraction for search and organization of published literature for patient care [17].

The NER task can be addressed by using sequence labeling [1, 2, 22, 23], span extraction [24–28], or text generation formulation [29–32]. The sequence labeling formulation formulates the NER task as a sequence tagging problem, in which NER models assign a tag (an entity type) to each token or a set of consecutive tokens. In contrast, the span extraction formulation considers entities as spans and adapts machine reading comprehension techniques to extract these spans. Recently, NER has been converted to a text generation task in which NER models learn to generate labels of corresponding entities. All these approaches usually use pre-trained language models (PLMs) or large language models (LLMs) for recognition with promising results [29–32]. The significant growth of techniques boosts the performance of NER models, yet training these NER models is still challenging and requires high-quality annotated data, especially in low-resource languages, e.g., Vietnamese. In fact, there are many annotated datasets of biomedical NER in English[1] [6, 14–17, 33–35]; however, the effort of creating the benchmark of biomedical NER in Vietnamese is still an early stage [36–38]. For instance, for the news domain, VLSP18 [37] provides a dataset that contains three entity types: person, location, and organization. For the medical domain, the COVID19 dataset [38] includes 10 entity types: patient ID, person name, age, gender, occupation, location, organization symptom and disease, transportation, and date. The ViMQ-NER corpus [39] consists of three entity types: symptom and disease, medical procedure, and medicine. However, the bottleneck still exists when adapting NER models to more specific domains. For example, we consider the scenario of supporting doctors in dealing with common disease diagnosis and treatment in Vietnam. To do that, the recognition of the diagnosis or treatment of a disease requires training in a high-quality NER model that can extract five entities: disease names, the symptoms of the diseases, the cause of the diseases, the diagnosis, and the treatment. This extraction significantly supports doctors as a tool for digesting input information from patients. This tool can save the time and effort of doctors who have to deal with a lot of patients per day. In this practical use, adapting prior medical NER datasets, e.g., the COVID19 dataset [38] or ViMQ-NER [39] dataset is an inappropriate solution. This is because the COVID19 dataset is specifically designed for the COVID-19 disease which is different from our purpose while the ViMQ-NER dataset lacks important information for the cause of diseases and the diagnosis which are important for treatment (ViMQ-NER only provides three entities: symptoms and disease, medical procedure, and medicine). To fill this gap, this paper introduces a new Vietnamese medical NER dataset that can be applied in the medical industry for common disease diagnosis or treatment. The dataset includes five entity types and is collected from four well-known Vietnamese websites that are professional in terms of drag selling and disease diagnostics. Although annotated by domain experts with high agreement scores, the creation of the dataset faces two main challenges. First, the annotation requires the involvement of domain experts who have experience with normal diseases. Compared to some prior NER datasets that only include common entities such as person names, locations, or organizations, the annotation of disease names, symptoms, or the cause of the diseases is more challenging. Second, as stated in Section 3.2, there exists confusion when annotating disease, symptom, and cause entities. Compared to some other entity types such as patient ID, person name, age, transportation, medical procedures, or medical medicine, the annotation of the five defined entities requires more effort from annotators. In summary, this paper makes two main contributions.

- It introduces a new medical NER dataset that contains more than 8000 annotated samples (the sentence level) with five entity types that can be used in common diagnostic and treatment scenarios. The dataset is annotated by domain experts to ensure the quality of data annotation. The newly created dataset enriches medical corpora in Vietnamese, a low-resource language. The dataset can be used in two scenarios. First, it plays an important role in both academia and medical industries to build an information extraction system that assists doctors in digesting input information from patients in common disease diagnosis and treatment. The results in Section 6.1 are the starting point for the adoption of the dataset to academia and medical industries. Second, it can be used as a seed to improve the quality of NER for other medical datasets by using transfer learning (Section 6.3). Our dataset and source code are available at https://github.com/tdtrinh11/ViMedNer.

- It shows the contribution of the dataset in medical NER scenarios by making a comparison among strong baselines. Experimental results facilitate the next studies of medical NER in Vietnamese.

[1] https://microsoft.github.io/BLURB/

The rest of this paper is organized as follows. Section 2 reviews relevant studies of our work. Section 3 describes the creation of the ViMedNER dataset. Section 4 presents the problem statement and NER models. Section 5 shows experimental settings and experimental results. Section 7 draws conclusions and future work.

2. Related work

Since the first work [40], many works and datasets have been introduced. CoNLL 2002 [41] and CoNLL 2003 [42] are popular corpora built from news articles in four languages (English, Spanish, German, and Dutch) with four main entities (Person, Organization, Location, and Miscellaneous). Besides, NER tasks for other languages were also presented, such as Indian languages [43], Arabic [44], and Slavic [45]. Regarding text sources, before 2005, datasets were mostly built from news articles, with a small number of NER types. After that, various kinds of text resources have been used such as Wikipedia articles, conversations, and user-generated text (e.g., tweets and YouTube comments), as summarized in [46]. For social media data like Twitter, entity types are more variable such as people, company, facility, band, sportsteam, movie, and TV shows, thanks to user behavior. In addition, variability in orthography and the presence of grammatically incomplete sentences of these data lead to performance degradation of classical NER systems [47]. In terms of structure, while most NER works are flat, some others focus on more complex structures such as nested NER [48] and discontinuous entities [16].

Apart from NER datasets for general domains, many works for the bio-medical were also presented. [14] built a BioNER corpus with four main entity types including protein, DNA, RNA, and cell attribute. A clinical note de-identification task requiring NER to locate personal patient data is introduced in [33]. Another work considered clinical text was proposed in [34]. In this work, clinical problem, test, and treatment entity types were taken into account. Furthermore, DrugNER [35] and CHEMDNER [16] for drug and chemical domains were introduced. In general, NER for the biomedical domain is a challenging task due to the complex orthographic structures of named entities [46]. This motivates us to do this work in which we build and release a medical NER dataset for Vietnamese - a low-resource language.

Recent studies have attempted to build data resources for the Vietnamese NER task. VLSP 2016 [36] is perhaps the first NER-shared task that was released to promote the development of the NLP community in Vietnam. This dataset used CoNLL 2003 compatible data format, with three NER types including person, organization, and location. Two extended versions of this dataset are VLSP 2018 [37] with more data samples and VLSP 2021 [49] with more entity types (i.e., 14 main entity types).

For the Vietnamese bio-medical domain, a public dataset was released by [38], which is the first manually annotated COVID-19 domain-specific dataset for Vietnamese. The dataset consists of 35K entities over 10K sentences, and 10 NER types that can be used in other future epidemics. Although this dataset has a large number of entities compared to other existing Vietnamese NER corpora, it is domain-specific (i.e., relating to the COVID-19 pandemic). Consequently, it is limited to use for common scenarios. This limitation could be addressed by using the ViMQ dataset [39]. However, this dataset was originally created for developing task-oriented healthcare chatbots with tag sets for two tasks of intent classification and NER. This leads to limitations on the diversity of entity types (e.g., only there ones including SYMPTOM&DISEASE, MEDICAL PROCEDURE, and MEDICINE). To fill this gap, we create a new dataset with five entity types that can be used in common diagnostic and treatment scenarios.

3. The ViMedNER Dataset

This section presents the creation of the dataset including data collection mentioned in Section 3.1, the annotation process described in Section 3.2, the annotation guideline shown in Section 3.3, and the statistics of the dataset summarized in Section 3.5.

3.1. Data collection

The first step of data creation is to collect relevant raw articles that mention diseases. To ensure the high quality of the created dataset, raw articles were collected from reputable websites of hospitals and pharmacies in Vietnam. They include *dieutri*,[2] *nhathuoclongchau* (Long Chau pharmacy),[3] and *vimec* (VinMec hospital).[4] Each article is written by medical experts who are in high levels of medicine and pharmacy for common diseases. The content of each article usually contains a description of a disease such as the overview, symptoms, treatment options, diagnostic options, etc.

Raw articles were collected by using the Scrapy framework [5] in the Python programming language. Although the collected articles are generally pretty accurate, they were divided into each field (e.g., the

[2]https://www.dieutri.vn/
[3]https://nhathuoclongchau.com.vn/benh
[4]https://www.vinmec.com/vi/benh/
[5]https://scrapy.org/

overview of a disease, or symptoms) to better address the topic and facilitate categorization before storing it in the MongoDB database. The total number of collected articles is 1,919. Table 1 shows the number of raw articles collected from each news provider.

Table 1. The number of collected articles.

News provider	# collected articles	# sentences
dieutri	568	22,456
nhathuoclongchau	659	27,170
vinmec	692	30,301
Total	1,919	79,927

As stated previously, data is collected and stored in respective fields. However, certain records may contain errors, such as HTML tags. Therefore, preprocessing is performed to address these issues, which includes several steps as the following. First, HTML tags were removed by using regex in the Python programming language. Second, data normalization was applied for Vietnamese Unicode (bringing back a built-in Unicode standard, and normalizing data about old accent typing in Vietnamese). Third, unnecessary characters (punctuation marks, numbers, and special characters) were cleaned. In addition, we use the underthesea[6] library in Python for sentence segmentation. This is because our purpose is NER on the sentence level. After segmentation, sentences that have a minimum length of 21 tokens were selected for data annotation. There are two reasons behind this selection. First, we observed that short sentences always do not contain any entity types. They usually mention general information such as the description of a disease without mentioning the entity types. Second, having long sentences makes our benchmark more challenging and also helps to maintain sentence coherence and mitigate any potential semantic changes caused by short sentences with an insufficient number of tokens. After selecting, the total number of sentences is 7,649 used for data annotation.

3.2. Annotation process

The annotation process includes two smaller steps: entity selection and annotation process design.

Entity definition. The first step of the annotation process is to identify the entity set. To do that, we worked closely with medical experts who are doctors at the Military Medical Academy[7]. Based on the very careful discussion, recommendations, and our actual use cases, five entities were selected, including DISEASE, DIAGNOSTIC, TREATMENT, CAUSE, and SYMPTOM. These entities provide essential information about a common disease that supports doctors to give general advice to patients. Table 2 summarizes the definitions of these five entities.

As presented in the table, the selected entities in our dataset differ from the two prior medical datasets. Specifically, the COVID19 dataset [38] contains 10 entities that exclusively focus on 10 information aspects of the COVID-19 disease. The ViMQ-NER dataset [39] defines three entities that cover cases in medical documents and help to identify diseases and related information. These entities are essential for the construction and deployment of a medical knowledge graph to address various related problems, such as suggesting disease diagnoses and facilitating question-and-answering interactions between patients and doctors. The large number of entities in the COVID-19 dataset provides much more specific information while a small number of entities in the ViMQ-NER dataset show more general information. In contrast, we try to balance detailed information on common diseases by selecting five essential entities suggested by domain experts. We believe the five selected entities provide enough information for doctors who will give comments to patients about common diseases.

Annotation workflow. After selecting entity types, an annotation process is designed with the involvement of domain experts. This ensures the high quality of annotated samples. Figure 1 describes how domain experts are involved in the annotation process to create the annotation guidelines and annotated samples.

The annotation process consists of two main phases. The main purpose of the first phase is to achieve high agreement among annotators who contribute to creating the annotation guidelines based on a small number of selected samples. After that, the second phase uses the created guidelines to annotate the whole dataset.

The first phase is done in several rounds. First, 100 samples are randomly selected from the preprocessed sentences. The first round starts with three annotators who are domain experts in medicine and pharmacy. The annotators annotate 100 samples separately, which is only based on the definition of the entities mentioned in Table 2, without any additional information or rules. This maintains the fair judgment of annotators and retains their opinions. After annotating 100 samples, the Kappa score is computed using the labels from the three annotators. Not surprisingly, the agreement score of the first round is low, with a Kappa score of 0.386, indicating that using only the definition of entities is not enough, and there are different judgments among annotators. Based on the agreement score, several meetings are conducted among annotators to identify and address any conflicts, confusion, or common mistakes made during the labeling process.

[6]https://github.com/undertheseanlp/underthesea
[7]http://vmmu.edu.vn/Portal/Home.html

Table 2. Entity type definitions in our annotation guidelines.

Entity	Definition
DISEASE	A disease is described as a group of tokens or terms that describe an illness or a particular medical condition, such as "tiểu đường" (diabetes), "bệnh Parkinson" (Parkinson's disease), "viêm xoang" (sinusitis).
SYMPTOM	A symptom is abnormal manifestations that patients perceive or observe as the disease progresses. For examples, "sốt" (fever), "đau" (pain), "khó thở" (difficulty breathing).
CAUSE	The causes of the disease are factors such as circumstances, lifestyle, or other factors that lead to disease.
DIAGNOSTIC	Diagnostics are methods or sets of rules used to evaluate, identify, and detect diseases or health conditions in an individual. For examples, "chụp X quang" (X-ray imaging), "siêu âm" (ultrasound), "đo huyết áp" (blood pressure measurement), "xét nghiệm máu" (blood tests).
TREATMENT	The treatment measure is a method to address a health issue or medical condition. For example, "phẫu thuật" (surgical removal), "uống thuốc kháng sinh" (take antibioticsl).

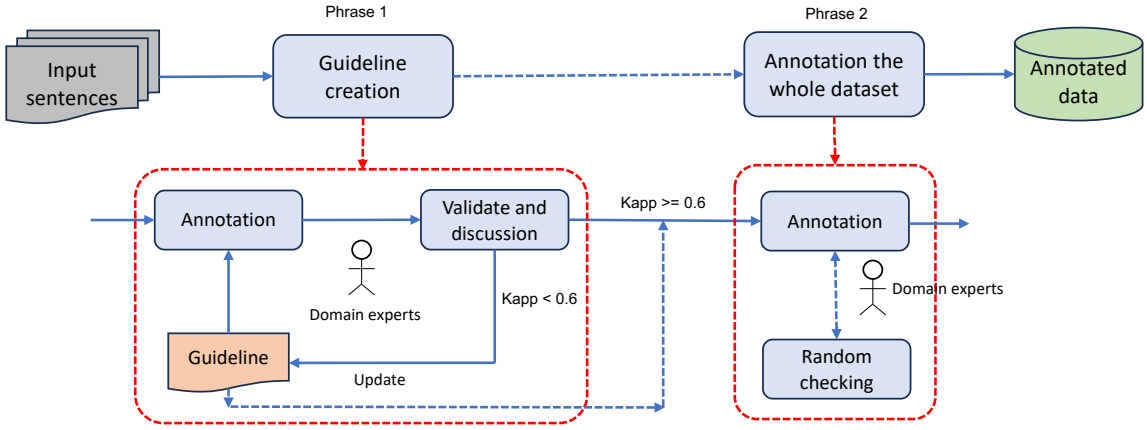

Figure 1. The annotation workflow with two phases.

These main mistakes include assigning meaningless spaces or labeling phrases. We find that annotators usually disagree on disease, symptom, and cause entities. This comes from that diseases are easily mistaken for their causes. In some cases, distinguishing between diseases and their causes can be challenging. Additionally, the symptoms and causes of diseases can also cause confusion, especially when they are in the same sentence. So, the first guideline is created after several meetings in which several rules are added to make clear definitions of disease, symptom, and cause entities (mentioned in Section 3.3). After that, the second round starts with the guideline. The three annotators are again asked to annotate 100 selected samples. After annotating, the Kappa score significantly improves, showing that the guideline is good for guiding the judgment of the annotators. To improve the final guideline, we continue to update and add remarks to the guideline. After several rounds, we achieved a Kappa agreement score of 0.620. This score shows that the agreement is good enough for creating the guideline that was applied to annotate the whole dataset.

In the second phase, the remaining samples are divided into three identical parts. Since the annotators have experience and the agreement score is high in the first phase, the annotation of the second phase is done separately for each part. More precisely, each annotator is assigned to annotate the corresponding part. This mechanism helps to speed up the annotation process and reduce the human effort of the annotation. After completing, the quality of the annotation is again validated. To do that, we randomly select 100 samples annotated by an annotator. These samples are then tagged by the two others. The Kappa agreement score is 0.608 showing that samples are quite well annotated with a moderate agreement score among annotators.

3.3. Annotation guideline

This section describes the final version of the guideline used to annotate the whole dataset. The guideline contains detailed information discussed by the domain experts for annotating five entities.

Disease. As mentioned, the disease entity is one of the confused entities. To make the definition of this

entity clear, a meeting was conducted with the domain experts. Finally, we define two additional rules that support the definition of the disease entity stated in Table 2. The rules are described as follows. (**i**) If there is an accompanying body part or disease name, it must be attached as a disease entity, for example, "ung thư dạ dày" (stomach cancer), *"viêm gan B"*(hepatitis B disease). (**ii**) If there is a word "bệnh" (disease), it must be attached, for example, "bệnh ung thư phổi" (lung cancer).

Symptom. We apply the same procedure of defining the symptom entity. After careful discussion with the domain experts, three additional rules are defined and added to support the definition of the symptom entity in Table 2. The rules include (**i**) Symptoms are external manifestations and NOT another disease, in this case, simply referred to the *DISEASE* entity. (**ii**) In the case of body parts with visible external manifestations, such as "tê chân" (numbness in the legs), "mỏi gối" (knee fatigue), and "đau nhức xương khớp" (joint and bone pain), it should be a symptom of a disease. (**iii**) Some diseases are named after their symptoms, such as "đau dạ dày" (stomach pain) or "tê bì tay chân" (numbness in hands and feet). Depending on the context, the same phrase can be either a symptom of a disease or the name of a disease.

Cause. Three additional rules are defined and added to the definition of the cause entity based on careful discussion with the domain experts. The rules are as follows. (**i**) The cause of the disease must be the effects or factors that directly lead to the disease or many symptoms of the disease. (**ii**) If the cause of the disease coincides with the symptom of the disease or the name of the disease, priority must be given to assigning the name of the disease and the symptom of the disease. For example, in the sentence "ho liên tục gây viêm họng hạt" (persistent cough causes pharyngitis), "ho liên tục" (the persistent cough) must be assigned as a symptom of the disease rather than the cause of the disease. (**iii**) If the cause of the disease in a sentence includes the name of the disease, we prioritize *DISEASE* label for this phrase. For example, "Khi mắc bệnh đái tháo đường là nguyên nhân của bệnh đột quỵ não" (when diabetes is the cause of brain stroke), "bệnh đáo tháo đường" (diabetes) and "bệnh đột quỵ não" (brain stroke) are labeled as two diseases rather than the causes of diseases.

Diagnostic and treatment. Since these two entities are quite easy for annotation, in which the annotators only assigned wrong labels for disease, symptom, and cause entities, we used the same definition of diagnostic and treatment entities in Table 2 for annotation.

3.4. Annotation

The BIO format. The BIO format has been widely used to annotate entities for the task of sequence labeling [37–39, 41, 42]. The tags consist of a prefix (*B-* or *I-*) followed by a label representing the type of a chunk or an entity. When a tag is preceded by the *B-* prefix, it signifies that the token marks the beginning of a chunk of an entity. Conversely, the *I-* prefix indicates that the token is inside a chunk of an entity. The *O* tag is assigned to tokens that do not belong to any entities or chunks. Figure 2 explains a sample in the dataset that uses the B-I-O tag format. With the sentence "Sởi có thể lan truyền từ một ngày trước khi bắt đầu giai đoạn khởi phát đến 4 ngày sau khi xuất hiện ban." (Measles can be spread from one day before onset to four days after the rash appears), based on the previously labeled data sets, we have "Sởi" (Measles) as the name of the disease, so it will be labeled *B_DIS*. Next, "xuất hiện ban" (rash appears) is the symptom of the disease, so "rash" will labeled *B_SYM* and "appear" will be assigned *I_SYM*. The remaining words in the sentence that do not belong to any label will be labeled *O*.

The annotation system. Because annotating the ViMedNER corpus is a time-consuming and labor-expensive task, we implement an annotation system based on Doccano[8], a text annotation tool to facilitate the annotation process. It is an open-source web platform that can support the annotation of several NLP tasks such as sentiment analysis, machine translation, and sequence tagging. By using the tool, several annotators can join the annotation process for time-saving. The annotation task was conducted by domain experts under the supervision of NLP and linguistics experts for labeling authenticity. In addition, the project manager works with the annotator leader, who has expertise in linguistics, medicine, and pharmacy to to control and guide annotators if necessary and evaluate annotation quality. We decided to work with domain experts for annotation instead of using crowdsourcing because we would like to keep the high quality of the annotation and crowdsourcing is not popular in Vietnam. The annotation interface of the annotation system is displayed in Figure 3.

When using the tool for labeling, the annotators are carefully trained to consider the following points. **(i)** They should familiarize themselves with the labeling instructions prior to assigning entities to avoid making associations based on personal bias and ensure compliance with the established rules. **(ii)** The annotators should refrain from associating independent words with entities. Entities should consist of coherent and meaningful phrases. For example, in the sentence

[8]https://github.com/doccano/doccano

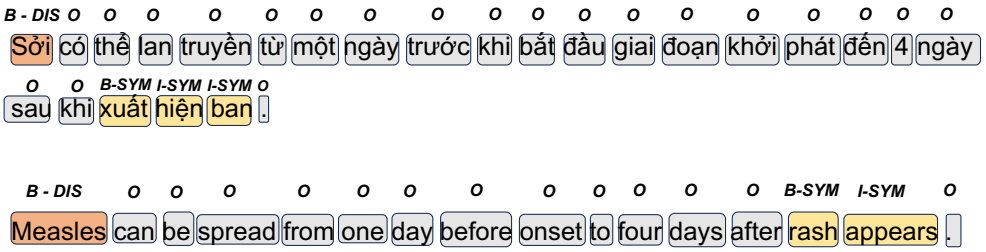

Figure 2. An example of BIO tags.

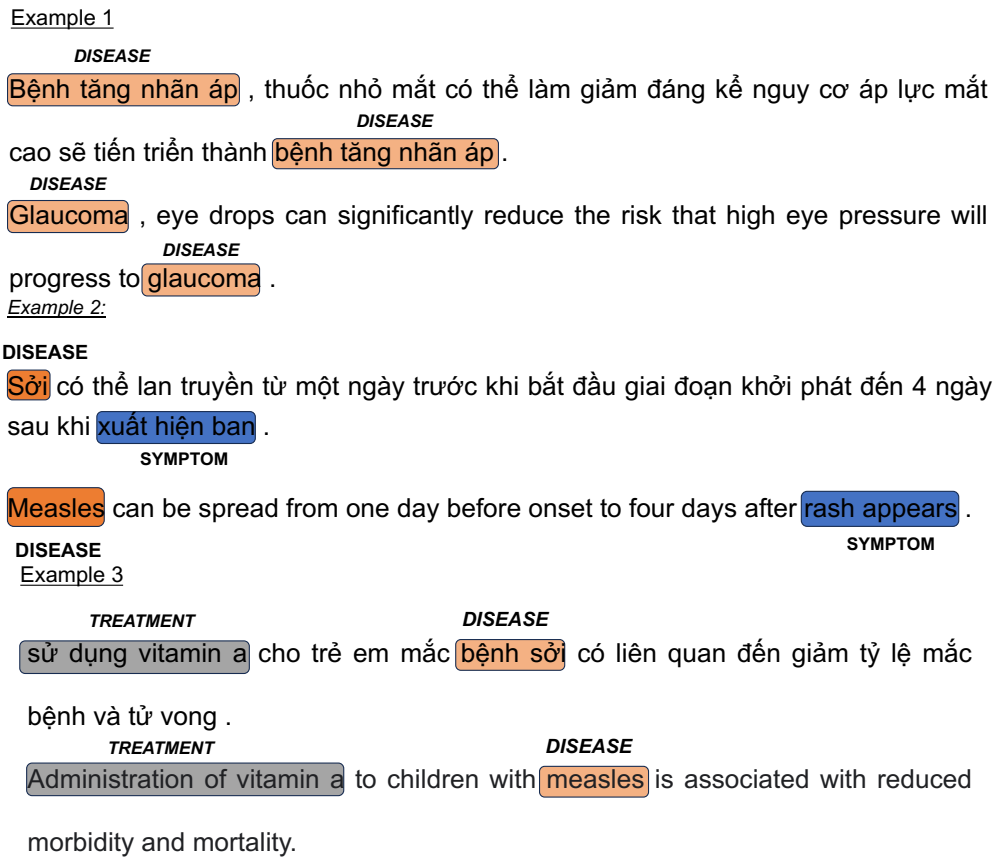

Figure 3. Example in annotating on the ViMedNER dataset.

"Triệu chứng của bệnh bao gồm sốt, nôn mửa" (Symptoms of the illness include fever, vomiting) assigning "sốt" (fever) as a symptom and "nôn mửa" (vomit) as another symptom, rather than assigning the entire phrase "sốt, nôn mửa" (fever, vomiting) as an illness symptom. **(iii)** Finally, the annotators should ensure accurate blocking of phrases that represent entities. They should avoid including spaces, commas (,), or other punctuation marks when assigning entities explicitly. In the example provided above, we will not assign the label "," to the SYMPTOM label. Another example, in the sentence "Chẩn đoán bệnh lao phổi bằng nhuộm soi đờm, trực tiếp tìm AFB" (diagnosis of pulmonary tuberculosis by sputum smear staining, directly looking for AFB), we need to label the phrase "nhuộm soi đờm, trực tiếp tìm AFB" (sputum smear staining, directly looking for AFB) as part of the DIAGNOSIS label, including the ",".

3.5. Data statistics and discussion

Table 3 summarizes the statistics of the dataset. As described in the table, the DISEASE label exhibits the highest frequency and the CAUSE label is the lowest frequency. In our observation, sentences within medical documents commonly encompass the disease name along with its associated explanations. Consequently, the higher number of DISEASE labels can be readily understood. The CAUSE label, on the other hand, typically involves longer phrases explaining the causes

of the disease. This is the reason why its count is lower compared to the other labels. It is similar to the DIAGNOSTIC label. It challenges NER models for the recognition of CAUSE and DIAGNOSTIC entities (please refer to Table 6). Figure 4 visualizes the distribution of each entity type in the annotated dataset. The dataset has three important characteristics. First, it covers common diseases which is different from the COVID-19 dataset [38] that only contains information on the COVID-19 disease. Second, sentences are long (more than 21 tokens) to ensure the meaning and context of input sequences. Finally, entities are clearly defined and explained to facilitate the identification of the entities to solve medical-related problems

Table 3. Statistics of the ViMedNER dataset.

Entity type	Training	Validation	Testing	All
DISEASE	6,004	1,871	1,837	9,712
SYMPTOM	2,645	786	764	4,195
CAUSE	888	290	255	1,433
DIAGNOSTIC	934	295	330	1,559
TREATMENT	1,810	689	585	3,084
#Entities in total	12,281	3,931	3,771	19,983
#Sentences in total	4,649	1,550	1,550	7,749

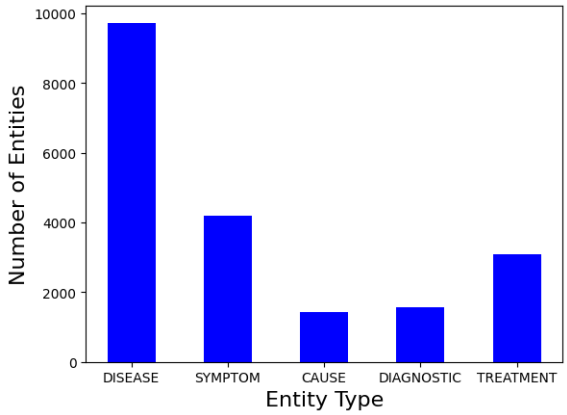

Figure 4. Statistics of the ViMedNER dataset.

Table 4 shows the comparison of our dataset and existing NER corpora in Vietnamese. We can observe that VLSP2018 [37] was designed for NER in the news domain which is inappropriate for our purpose. Among medical corpora, COVID-19 [38] was specially created to extract aspects of COVID-19 that can not applied to the scenario of common disease diagnosis and treatment. The VIMQ-NER [39] is perhaps the most similar dataset to our scenario. However, it only provides three entities that lack two important aspects of a common disease: the cause of the disease and the diagnosis of the disease. In addition, our dataset includes more annotated samples that facilitate the training NER models.

Table 4. NER datasets in Vietnamese. The number of training, development, and testing samples is counted on the entity level.

Dataset	Train	Dev	Test	# entities	Domain
VLSP18 [37]	15,046	—	3,052	3	News
COVID19 [38]	15,767	7,478	1,173	10	Medical
ViMQ-NER [39]	10,599	1,300	1,354	3	Medical
ViMedNER (Ours)	12,281	3,931	3,771	5	Medical

4. Named Entity Recognition of ViMedNER

This section first states the problem and then describes strong NER models used for evaluation.

4.1. Problem statement

As mentioned in Section 1, the recognition of entities can be formulated as sequence labeling [1, 2, 22, 23], span extraction [24–28], or text generation formulation [29–32]. In this work, we follow the sequence labeling approach for NER of the ViMedNER dataset that is formulated as follows. Given a sequence $S = \{w_1, w_2, ..., w_n\}$ with n tokens, the task is to learn a mapping function $f(.)$ to predict a label sequence $\hat{y}^{y_1,y_2,...,y_n} = f(\hat{y}|\theta, S)$, where $\hat{y}_i$ is the predicted label of the corresponding token w_i, θ is parameters of the NER models.

4.2. NER models

We implement NER models based on strong pre-trained models (PLMs) for Vietnamese. Given an input sequence $S = \{w_1, w_2, ..., w_n\}$, the input is first tokenized by using a wordpiece tokenizer. After tokenizing, input tokens were fed into PLMs to obtain contextual embeddings. Suppose the $PML()$ function returns corresponding contextual vectors of the input sequence S, the mapping can be done as $\boldsymbol{V} = PLM(S)$. The hidden vector $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, ..., \boldsymbol{v}_n\}$ represents the contextual embedding of n input tokens which is then fed into a softmax layer to predict the label of each input token. The overall architecture of the NER models is illustrated in Fig. 5

For training, NER models compute the negative log-likelihood loss between the gold label $\boldsymbol{y}^*$ and the input sequence S, as expressed in Eq. 1. For inference, given an input sequence S, NER models find the best label sequence $\hat{y} \in \mathcal{Y}(S)$ that maximizes the conditional probability computed by the loss function $\mathcal{L}$.

$$\mathcal{L} = -\log p_\theta(\boldsymbol{y}^*|S) \quad (1)$$

In this work, we adopt strong baselines that are based on pre-trained language models as below:

- **PhoBERT** [50] is the first large-scale monolingual language model for Vietnamese that inherits the RoBERTa, pre-trained on a 20GB word-level

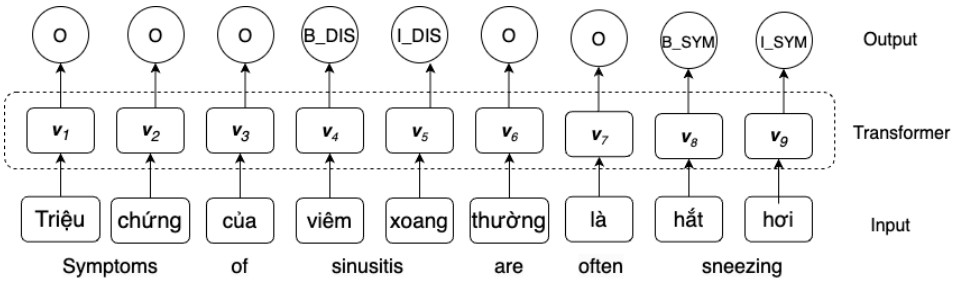

Figure 5. The NER models are based on pre-trained models. The models receive an input sequence and predict a label sequence. One DISEASE entity started by the *B_DIS* character is predicted and one SYMPTOM started by the *B_SYM* by using the BIO format.

Vietnamese dataset. This model is proven as a SOTA model in several NLP tasks such as POS tagging, dependency parsing, and NER [38]. Two models of PhoBERT (base and large) were used for implementation.

- **XLM-R** A scaled multilingual sentence encoder, XLM-RoBERTa [51] is a RoBERTa variation that was previously trained on a 2.5TB multilingual dataset encompassing more than 100 languages, including around 137GB of Vietnamese text at the syllable level. Two models of XLM-R (base and large) were used for implementation.
- **ViDeBERTa** [52] is a new pre-trained monolingual language model for Vietnamese, which is based on DeBERTa [53]. It was pre-trained on a large-scale and high-quality corpus. Two models of ViDeBERTa (small and base) were used for implementation.
- **ViPubMedDeBERTa**[9] a pre-trained model was built on the ViDeBERTa [52] architecture and trained on ViPubmed [54], a dataset of 20 million Vietnamese biomedical abstracts generated by large scale translation. This model is publicly available on Huggingface's model library. Two models of ViPubMedDeBERTa (small and base) were used for implementation.
- **ViHealthBERT** [55] is the first domain-specific pre-trained language model for Vietnamese healthcare. In experiments, it shows potential results in four health-domain Vietnamese datasets. Two models of ViHealthBERT (word and syllable) were used for implementation.

To keep a fair comparison, we did not add additional layers, e.g., LSTM or CRF on the top of PLMs. Statistics of the baselines are summarized in Table 5. As shown in the table, each model is described in terms of model size, its based model, and data size used to pre-train. This information helps us not only explain the behavior of the models but also explore insights into our dataset (see Section 6).

[9]https://huggingface.co/manhtt-079/vipubmed-deberta-base

5. Experimental Settings

5.1. Evaluation metrics

For NER systems, comparing their predicted outputs against ground truth annotated by humans is a widely used evaluation method [46]. NER tasks involve two main subtasks including boundary detection and type identification. Following many other works [38, 46], we adopt an exact-match evaluation that requires NER systems to correctly identify boundary and type, simultaneously. Particularly, we use metrics of Precision, Recall, and F-score for evaluation. While Precision refers to the percentage of correct predicted outputs, Recall refers to the percentage of total entities correctly recognized by NER systems. F-score is the harmonic mean of precision and recall. In this work, we also adopt macro-average and micro-average F1-scores as metrics for measuring the performance of NER models. Equations of the micro-average F1-score are expressed in Eq. 2

$$
\begin{aligned}
Precision_{macro} &= \frac{1}{n}\sum_{i=1}^{n} Precision_i \\
Recall_{macro} &= \frac{1}{n}\sum_{i=1}^{n} Recall_i \\
macro-F1 &= 2 \times \frac{Precision_{macro} \times Recall_{macro}}{Precision_{macro} + Recall_{macro}}, \quad (2)
\end{aligned}
$$

where $Precision_i = \frac{TP_i}{TP_i+FP_i}$ and $Recall_i = \frac{TP_i}{TP_i+FN_i}$, TP_i, FP_i, FN_i, and TN_i are True Positive, False Positive, False Negative, and True Negative of entity type i respectively. While Macro-averaged F-score independently computes the score on different entity types before taking the average, the Micro one sums up the individual FN_i, FP_i, and TP_i across all entity types

Table 5. Statistics of the baselines. M stands for million. ∗ denotes that the model is based on ViDeBERTa and continuously pre-trained on 20M abstracts.

Models	Size (M)		Base model	Pre-train data scale
PhoBERT [50]	base 135	large 370	RoBERTa	20GB
XLM-R [51]	base 250	large 560	XLM&RoBERTa	137GB
ViDeBERTa [52]	xsmall 22	base 86	DeBERTa	138GB
ViPubMedDeBERTa*	xsmall 22	base 86	ViDeBERTa	20M abstracts
ViHealthBERT [55]	110		BERT	130M documents

and then applies them to get the result. The actual calculation of Micro-average F1-score is as Eq. 3:

$$
\begin{aligned}
Precision_{micro} &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \\
Recall_{micro} &= \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \\
micro - F1 &= 2 \times \frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}}. \quad (3)
\end{aligned}
$$

5.2. Experimental setup

During the experiment, we evaluated all versions of the above pre-trained models that are publicly available on Huggingface's Transformers [56], a huge library of pre-trained Transformers-based models. All models are trained on Nvidia Tesla T4 GPU with 15GB RAM provided by Google Colab[10]. To optimize model parameters, we employ AdamW optimizer with an L2 weight decay of 0.01 and a learning rate of 5e-5, and all other training parameters are set to defaults in Transformers. The maximum sequence length is set to 128 for all sequences. All baselines are trained for 30 epochs with a batch size of 32. We evaluate the performance after each epoch on the validation set and apply an early stopping regularization if performance does not improve after 5 continuous epochs.

6. Results and Discussion

This section presents experiments of baselines on our ViMedNER dataset. We first show the comparison in terms of F-scores among strong pre-trained language models. After that, additional experiments are also conducted to explore the behavior of the baselines and the characteristics of our dataset.

6.1. F-score comparison

Table 6 compares evaluation results of the baselines in terms of Mic-F1 and Mac-F1. As observed, the performance of XLM-R_{large} is consistently better than the other ones in all metrics. Specifically, i t shows better results of 1.4% (Mic-F1) and 1.8% (Mac-F1) compared to the second bests that are PhoBERT$_{large}$ and XLM-R_{base}, respectively. This may come from that XLM-R_{large} has the biggest model size and was trained with a huge corpus. Surprisingly, PhoBERT$_{base}$ was trained on a Wikipedia and news dataset but is superior to both ViPMDeBERTa$_{base}$ and ViHealthBert$_{syllable}$ that were pre-trained on more medical documents. In particular, while it outputs similar results with ViPMDeBERTa$_{base}$, it is better than ViHealthBert$_{syllable}$, about 1.1% (Mic-F1) and 1.2% (Mac-F1). In addition, although ViDeBERTa$_{base}$ was trained on a larger corpus than PhoBERT$_{base}$, it is worse than PhoBERT$_{base}$ with very-large gaps of 11% and 11.8% in terms of Mic-F1 and Mac-F1 respectively.

For ViPMDeBERTa$_{base}$ and ViHealthBert$_{syllable}$ that were pre-trained on medical domain corpora, ViPMDeBERTa$_{base}$ outputs better scores than the other one, about 1.1% for both Mic-F1 and Mac-F1. It should be noted that while ViPMDeBERTa$_{base}$ was pre-trained on PubMed abstracts translated to Vietnamese, ViHealthBert$_{syllable}$ used a corpus composed of general, health, and medical domains. Among small-size models, the performance of ViPMDeBERTa$_{xsmal}$ is completely superior compared to the performance of ViDeBERTa$_{xsmall}$ with large margins of 7.6% (Mic-F1) and 9.6% (Mac-F1). This further confirms that training small models on specific domain corpus is necessary to improve NER system performance.

These experimental results mentioned above are the preliminary results of strong NER baselines when evaluated in our dataset. They can be used as references for future works in both academia and industry for the medical domain. Furthermore, as shown in Table 6, the baselines outputted low results on the CAUSE entity, due to high ambiguity between it and others (e.g., disease names and symptoms). This also means that our dataset is a new good and challenging benchmark for evaluating new NER systems. Consequently, our dataset

[10]https://colab.research.google.com/

Table 6. Experiment results of the baselines. **Bold** values are the best.

Model	DIS.	SYM.	CAU.	DIA.	TRE.	Mic-F1	Mac-F1
PhoBERT$_{base}$	0.818	0.640	0.335	0.679	0.615	0.701	0.617
PhoBERT$_{large}$	0.823	0.646	0.339	0.683	0.618	0.711	0.622
XLM-R$_{base}$	0.820	**0.650**	0.342	0.686	0.621	0.710	0.624
XLM-R$_{large}$	**0.842**	0.641	**0.373**	**0.707**	**0.636**	**0.725**	**0.640**
ViDeBERTa$_{xsmall}$	0.732	0.487	0.212	0.499	0.491	0.594	0.484
ViDeBERTa$_{base}$	0.729	0.491	0.236	0.536	0.500	0.591	0.499
ViPMDeBERTa$_{xsmall}$	0.803	0.587	0.308	0.636	0.565	0.670	0.580
ViPMDeBERTa$_{base}$	0.818	0.619	0.368	0.653	0.621	0.701	0.616
ViHealthBert$_{word}$	0.804	0.597	0.357	0.663	0.582	0.680	0.601
ViHealthBert$_{syllable}$	0.819	0.636	0.358	0.629	0.581	0.690	0.605

can boost the development of new medical NER models for Vietnamese.

6.2. Data segmentation and F-scores

To further explore the behavior of the baseline models, we conducted a data segmentation experiment on our ViMedNER. Specifically, we randomly select various percentages of the training set, which are then used to train baselines. Here, we segment the data with variation ratios of 1%, 5%, 15%, 25%, 50%, 75%, and 100%.

Experiment results are illustrated in Fig. 6. The first observation is that almost all baselines perform well with a small percentage of the training set. These results confirm the high-quality annotation process of our dataset, in which these gold labels are a good signal for guiding the model during the training period. Second, XLM$_{large}$ once again shows good results compared to the others, even at a very small training set (i.e., 1%). This is consistent with the experimental results described above. The reason for this may be that XLM$_{large}$ was pre-trained with a huge dataset. In addition, although ViDeBERTa$_{base}$ was trained on a huge corpus, it shows the worst results compared to ViHealthBERT$_{syllable}$ and ViPMDeBERTa$_{base}$ in all metrics and segmentation ratios. This shows that training on domain-specific data is necessary to improve the generation capability of language models. The general trend shows that the performance of NER models is significantly improved with a small number of training samples (1-5%). This may come from that PLMs are trained with a massive amount of data, so they can adapt to new downstream tasks with a few training samples. The performance increases slowly after using 50% of training samples. It indicates that we can train good NER models with smaller training samples by using appropriate data selection methods.

6.3. Cross-data experiment

Inspired from [57], we conduct a cross-data evaluation aiming to explore the capability of our dataset in improving NER performance. In this experiment, we train the models on our ViMedNER, aiming to enrich fine-grained representations. These models (i.e., the model name with $+Enh$) are then fine-tuned for target tasks that are COVID19 [38] and ViMQ-NER dataset [39].

As described in Tables 7 and 8, experiment results show that using our corpus can improve the performance of other NER tasks, especially for large models. For the COVID-19 dataset, the performance of enhanced models is better than that of baselines extracted in [38] in both Mic-F1 and Mac-F1. Specifically, XLM-R$_{base}$ + Enh shows an improvement of 5.2% in terms of Mac-F1 for BER of COVID-19. For VIMQ-NER, models that were enhanced with our corpus are consistently superior to the other ones. For Mac-F1, PhoBERT$_{base+Enh}$ is better than PhoBERT+CRF [39], about 3.4%. This further confirms that our dataset can be used to enrich medical information, and then improve other medical NER tasks.

7. Conclusion and Future Work

This paper introduces a new medical NER dataset for Vietnamese which has almost 8,000 samples with 5 entity types. Contrary to existing works, it focuses on recognizing five entities: disease names, symptoms, causes, diagnosis, and treatment that can be used in common disease diagnostic and treatment scenarios. Extensive experimental results show three important points. First, NER models based on pre-trained language models achieve promising results. These results facilitate the creation of an AI system that supports doctors in common disease diagnosis and treatment. Second, our corpus can be used as an initial data source to improve the quality of NER for other medical corpora. Finally, training language models on

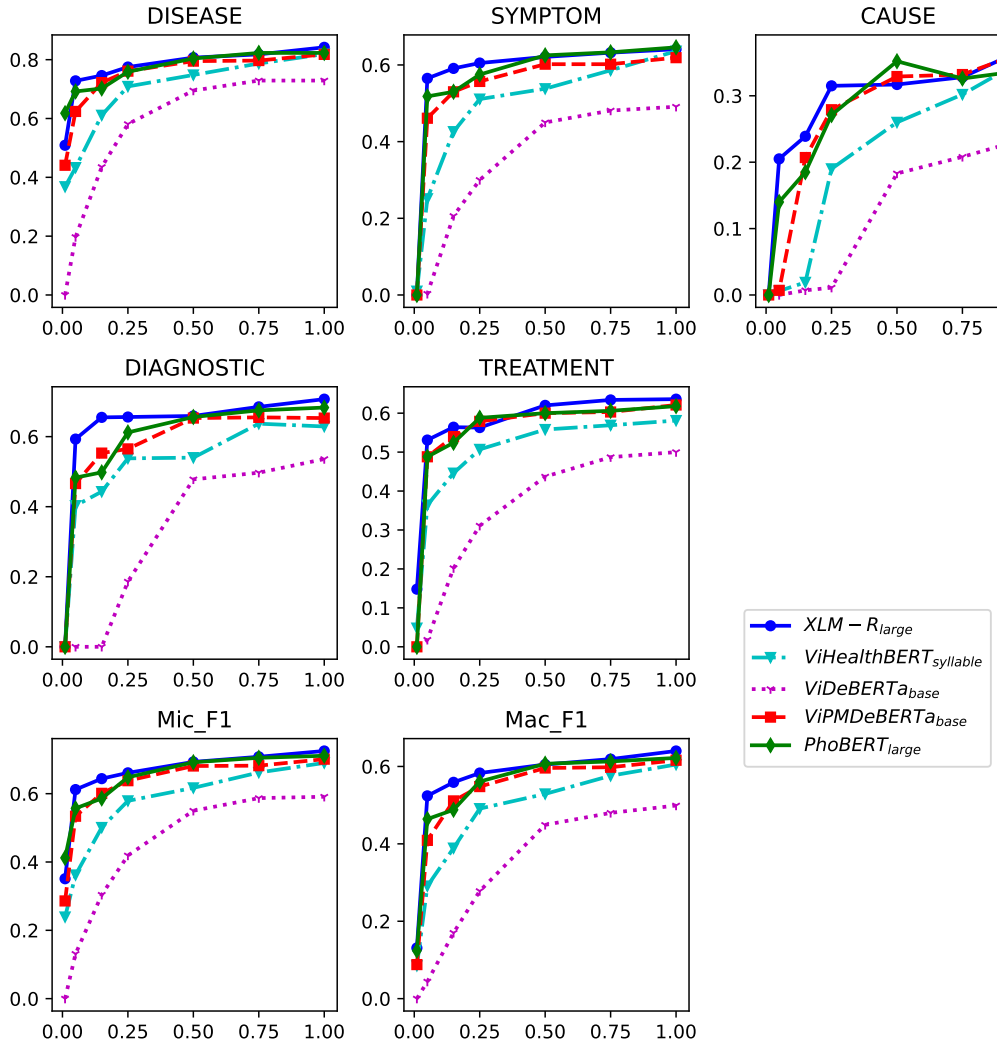

Figure 6. Performance on varying training data segmentation.

Table 7. Cross-data experiment results on the COVID-19 task. ★ denotes that the evaluation results of these models are extracted in [38].

Model	PAT.	PER.	AGE	GEN.	OCC.	LOC.	ORG.	SYM.	TRA.	DAT.	Mic-F1	Mac-F1
$PhoBERT_{base}$★	0.981	0.903	0.962	0.954	0.749	0.943	0.870	0.883	0.966	0.987	0.942	0.920
$PhoBERT_{base} + Enh$	0.980	0.959	0.963	0.976	0.781	0.948	0.889	0.885	0.972	0.985	0.949	0.934
$PhoBERT_{large}$★	0.980	0.944	0.967	0.968	0.791	0.940	0.876	0.885	0.967	0.989	0.945	0.931
$PhoBERT_{large} + Enh$	0.982	0.953	0.969	0.977	0.789	0.948	0.892	0.873	0.982	0.989	0.949	0.935
$XLM\text{-}R_{base}$★	0.978	0.902	0.957	0.842	0.560	0.941	0.842	0.858	0.924	0.982	0.925	0.879
$XLM\text{-}R_{base} + Enh$	0.984	0.946	0.972	0.968	0.754	0.950	0.881	0.878	0.987	0.988	0.949	0.931
$XLM\text{-}R_{large}$★	0.982	0.933	0.962	0.958	0.692	0.943	0.853	0.854	0.943	0.987	0.938	0.911
$XLM\text{-}R_{large} + Enh$	0.983	0.952	0.980	0.977	0.797	0.952	0.901	0.875	0.979	0.994	0.959	0.939

a specific-domain corpus is necessary to improve the model performance.

For future work, we will propose a new model for medical NER tasks. Furthermore, extending the size of ViMedNER will be taken into account.

Declarations

- Funding: This research is funded by the Ministry of Education and Training under grant number B2022-SKH-01
- Conflict of interest/Competing interests: The authors have no relevant financial or non-financial interests to disclose

Table 8. Cross-data experiment results on the VIMQ-NER task. Notation of $+Enh$ indicates the models enhanced by using ViMedNER.

Model	SYMP.&DIS.	MEDICINE	MED.PRO.	Mic-F1	Mac-F1
PhoBERT+CRF [39]	0.772	0.673	0.617	0.747	0.687
PhoBERT$_{base}$	0.776	0.721	0.626	0.754	0.708
PhoBERT$_{base}$ + Enh	0.777	0.727	0.659	0.759	0.721
PhoBERT$_{large}$	0.773	0.715	0.671	0.757	0.720
PhoBERT$_{large}$ + Enh	0.783	0.809	0.667	0.770	0.753
XLM-R$_{base}$	0.777	0.742	0.638	0.757	0.719
XLM-R$_{base}$ + Enh	0.784	0.773	0.662	0.768	0.740
XLM-R$_{large}$	0.790	0.748	0.703	0.776	0.747
XLM-R$_{large}$ + Enh	0.786	0.807	0.714	0.783	0.769

- Availability of data and materials: https://github.com/tdtrinh11/ViMedNer
- Authors' contributions: Van-Duong Pham and Tien-Dat Trinh: Data curation, Methodology, Validation; Minh-Tien Nguyen and Huy-The Vu: Conceptualization, Methodology, Writing and Editing Minh-Chuan Pham, Manh-Tuan Tran and Hoang-Son Le: Visualization, Investigation, Validation; All authors reviewed the manuscript.

References

[1] Angeli, G., Premkumar, M.J. and Manning, C.D. (2015) Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 344-354*.

[2] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016) Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260-270*.

[3] Li, X., Feng, J., Meng, Y., Han, Q., Wu, F. and Li, J. (2020) A unified mrc framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5849-5859*.

[4] Puccetti, G., Chiarello, F. and Fantoni, G. (2021) A simple and fast method for named entity context extraction from patents. *Expert Systems with Applications 184 (2021): 115570* .

[5] Sang, E., Kim, T. and Meulder, F.D. (2003) Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

[6] Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P. *et al.* (2016) Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database* **2016**.

[7] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. and Liu, Q. (2019) Ernie: Enhanced language representation with informative entities. In *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1441-1451*.

[8] Cheng, P. and Erk, K. (2020) Attending to entities for better text understanding. In *In Proceedings of the AAAI conference on artificial intelligence, vol. 34, no. 05, pp. 7554-7561*.

[9] Guo, J., Xu, G., Cheng, X. and Li, H. (2009) Named entity recognition in query. In *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 267-274*.

[10] Aone, C. (1999) A trainable summarizer with knowledge acquired from robust nlp techniques. *Advances in automatic text summarization: 71-80* .

[11] Mollá, D., Zaanen, M.V. and Smith, D. (2006) Named entity recognition for question answering. In *In Proceedings of the Australasian language technology workshop 2006, pp. 51-58*.

[12] Babych, B. and Hartley, A. (2003) Improving machine translation quality with automatic named entity recognition. In *In Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

[13] Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J.F. *et al.* (2020) Building a pubmed knowledge graph. *Scientific data 7, no. 1: 205* .

[14] Collier, N., Ohta, T., Tsuruoka, Y., Tateisi, Y. and Kim, J.D. (2004) Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)* (Geneva, Switzerland: COLING): 73–78. URL https://aclanthology.org/W04-1213.

[15] Doğan, R.I., Leaman, R. and Lu, Z. (2014) Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics* **47**: 1–10.

[16] Krallinger, M., Rabal, O., Leitner, F., Vázquez, M., Salgado, D., Lu, Z., Leaman, R. *et al.* (2015) The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics* **7**: S2 – S2.

[17] Nye, B., Li, J.J., Patel, R., Yang, Y., Marshall, I.J., Nenkova, A. and Wallace, B.C. (2018) A corpus with

multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (NIH Public Access), **2018**: 197.

[18] Kocaman, V. and Talby, D. (2022) Accurate clinical and biomedical named entity recognition at scale. *Software Impacts 13: 100373* .

[19] Uzuner, Ö., South, B.R., Shen, S. and DuVall, S.L. (2011) 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* **18**(5): 552–556.

[20] Tzitzivacos, D. (2007) International classification of diseases 10th edition (icd-10). *CME: Your SA Journal of CPD* **25**(1): 8–10.

[21] Uzuner, Ö., Luo, Y. and Szolovits, P. (2007) Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association* **14**(5): 550–563.

[22] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186*.

[23] Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F. and Tu, K. (2021) Improving named entity recognition by external context retrieving and cooperative learning. In *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1800-1812*.

[24] Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L. and Levy, O. (2020) Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics, 8: 64–77* .

[25] Li, F., Lin, Z., Zhang, M. and Ji, D. (2021) A span-based model for joint overlapped and discontinuous named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 4814–4828*.

[26] Fu, J., Huang, X.J. and Liu, P. (2021) Spanner: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*: 7183–7195.

[27] Son, N.H., Hieu, M.Y., Nguyen, T.A.D. and Nguyen, M.T. (2022) Jointly learning span extraction and sequence labeling for information extraction from business documents. In *2022 International Joint Conference on Neural Networks (IJCNN)* (IEEE): 1–8.

[28] Wan, J., Ru, D., Zhang, W. and Yu, Y. (2022) Nested named entity recognition with span-level graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*: 892–903.

[29] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A. *et al.* (2020) Language models are few-shot learners. *Advances in neural information processing systems* **33**: 1877–1901.

[30] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z. and Tang, J. (2021) All nlp tasks are generation tasks: A general pretraining framework. *arXiv preprint arXiv:2103.10360* .

[31] Paolini, G., Athiwaratkun, B., Krone, J., Ma, J., Achille, A., Anubhai, R., Santos, C.N.d. *et al.* (2021) Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779* .

[32] He, Y. and Tang, B. (2022) Setgner: General named entity recognition as entity set generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*: 3074–3085.

[33] Uzuner, , Luo, Y. and Szolovits, P. (2007) Evaluating the State-of-the-Art in Automatic De-identification. *Journal of the American Medical Informatics Association* **14**(5): 550–563. doi:10.1197/jamia.M2444, URL https://doi.org/10.1197/jamia.M2444. https://academic.oup.com/jamia/article-pdf/14/5/550/2136261/14-5-550.pdf.

[34] Uzuner, Ö., South, B.R., Shen, S. and Duvall, S.L. (2011) 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA* **18 5**: 552–6.

[35] Segura-Bedmar, I., Martínez, P. and Herrero-Zazo, M. (2013) SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (Atlanta, Georgia, USA: Association for Computational Linguistics): 341–350. URL https://aclanthology.org/S13-2056.

[36] Huyen, N.T.M. and Luong, V.X. (2016) Vlsp 2016 shared task: Named entity recognition. *Proceedings of Vietnamese Speech and Language Processing (VLSP)* .

[37] Nguyen, H.T., Ngo, Q.T., Vu, L.X., Tran, V.M. and Nguyen, H.T. (2018) Vlsp shared task: Named entity recognition. *Journal of Computer Science and Cybernetics* **34**(4): 283–294.

[38] 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*: 2146–2153.

[39] Huy, T.D., Tu, N.A., Vu, T.H., Minh, N.P., Phan, N., Bui, T.H. and Truong, S.Q. (2021) Vimq: A vietnamese medical question dataset for healthcare dialogue system development. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part VI 28* (Springer): 657–664.

[40] Grishman, R. and Sundheim, B. (1996) Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. URL https://aclanthology.org/C96-1079.

[41] Tjong Kim Sang, E.F. (2002) Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. URL https://aclanthology.org/W02-2024.

[42] Tjong Kim Sang, E.F. and De Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*: 142–147. URL https://aclanthology.org/W03-0419.

[43] Singh, A.K. (2008) Named entity recognition for south and south East Asian languages: Taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*. URL https://aclanthology.org/I08-5003.

[44] Shaalan, K. (2014) A survey of arabic named entity recognition and classification. *Comput. Linguist.* **40**(2): 469–510. doi:10.1162/COLI_a_00178, URL https://doi.org/10.1162/COLI_a_00178.

[45] Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J. and Yangarber, R. (2017) The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing* (Valencia, Spain: Association for Computational Linguistics): 76–85. doi:10.18653/v1/W17-1412, URL https://aclanthology.org/W17-1412.

[46] Li, J., Sun, A., Han, J. and Li, C. (2022) A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* **34**(1): 50–70. doi:10.1109/TKDE.2020.2981314.

[47] Baldwin, T., de Marneffe, M.C., Han, B., Kim, Y.B., Ritter, A. and Xu, W. (2015) Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text* (Beijing, China: Association for Computational Linguistics): 126–135. doi:10.18653/v1/W15-4319, URL https://aclanthology.org/W15-4319.

[48] Wang, Y., Tong, H., Zhu, Z. and Li, Y. (2022) Nested named entity recognition: A survey. *ACM Trans. Knowl. Discov. Data* **16**(6). doi:10.1145/3522593, URL https://doi.org/10.1145/3522593.

[49] Linh, H., Dao, D., Huyen, N., Quyen, N. and Dung, D. (2022) Vlsp 2021 - ner challenge: Named entity recognition for vietnamese. *VNU Journal of Science: Computer Science and Communication Engineering* **38**(1). doi:10.25073/2588-1086/vnucsce.362, URL //jcsce.vnu.edu.vn/index.php/jcsce/article/view/362.

[50] Nguyen, D.Q. and Nguyen, A.G.T. (2020) Phobert: Pre-trained language models for vietnamese. *ArXiv* **abs/2003.00744**.

[51] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E. *et al.* (2019) Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics*.

[52] Tran, C.D., Pham, N.H., Nguyên, A.V., Hy, T.S. and Vu, T. (2023) Videberta: A powerful pre-trained language model for vietnamese. In *Findings*.

[53] He, P., Liu, X., Gao, J. and Chen, W. (2020) Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv* **abs/2006.03654**.

[54] Phan, L., Dang, T., Tran, H.T., Trinh, T.H., Phan, V., Chau, L.D. and Luong, M.T. (2022) Enriching biomedical knowledge for low-resource language through large-scale translation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

[55] Minh, N., Tran, V.H., Hoang, V., Ta, H.D., Bui, T.H. and Truong, S.Q.H. (2022) ViHealthBERT: Pre-trained language models for Vietnamese in health text mining. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille, France: European Language Resources Association): 328–337. URL https://aclanthology.org/2022.lrec-1.35.

[56] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P. *et al.* (2020) Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Online: Association for Computational Linguistics): 38–45. doi:10.18653/v1/2020.emnlp-demos.6, URL https://aclanthology.org/2020.emnlp-demos.6.

[57] Chen, Y., Liu, P., Zhong, M., Dou, Z.Y., Wang, D., Qiu, X. and Huang, X. (2020) CDEvalSumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online: Association for Computational Linguistics): 3679–3691. doi:10.18653/v1/2020.findings-emnlp.329, URL https://aclanthology.org/2020.findings-emnlp.329.