Open-domain Named Entity Recognition for Low Resource Languages A Case Study on Vietnamese

Viet Ngo Q. and **Huong Le T.** *
School of Information and Communication Technology,
Hanoi University of Science and Technology, Hanoi, Vietnam
vietnq.work@gmail.com and huonglt@soict.hust.edu.vn

Abstract

Named Entity Recognition (NER) is a crucial component of Natural Language Processing (NLP) systems, essential for tasks such as information retrieval and question answering. However, existing NER models often struggle with the broad spectrum of entity types encountered in open-domain settings, particularly in low-resource languages like Vietnamese, which lack extensive labeled datasets. This study introduces a novel method for fine-tuning multilingual models, specifically mT5 and mT0, to address open-domain NER tasks in Vietnamese. We generated a comprehensive open-domain annotated Vietnamese NER dataset using a large language model (LLM) and evaluated the models in both zero-shot and supervised fine-tuning settings. The mT0-large model achieved F1 scores of 0.6030 on VLSP NER 2021 and 0.5753 on PhoNER_COVID19 in zero-shot, improving to 0.7489 and 0.9431, respectively, with supervised fine-tuning. This method shows promise for improving NER in low-resource languages.

1 Introduction

Natural Language Processing (NLP) has seen a surge in real-world applications, ranging from voice assistants to automated content analysis. Among the various NLP tasks, Named Entity Recognition (NER) plays a crucial role in extracting structured information from unstructured text by identifying and classifying entities into predefined categories such as names, locations, and organizations (Grishman, 2019). This task is foundational for many downstream applications, including information retrieval (Khalid et al., 2008) and question answering (Mollá et al., 2006), where accurate entity recognition is essential.

Despite its importance, the field of open-domain NER, which involves recognizing a wide range of entity types across various domains beyond traditional categories, remains underexplored. Open-domain NER has the potential to significantly enhance many NLP applications by improving the flexibility and accuracy of entity recognition in diverse contexts. However, developing effective open-domain NER models is particularly challenging for low-resource languages like Vietnamese, where high-quality and diverse datasets are limited.

This paper aims to address these challenges by proposing a novel approach to train multilingual NER models that can handle the complexities of open-domain scenarios in Vietnamese. Our method focuses on fine-tuning multilingual models, specifically mT5 and mT0, to accommodate a broad spectrum of entity types, overcoming the limitations posed by the scarcity of Vietnamese NER datasets. We also explore the potential for multilingual transfer and multitask learning within encoder-decoder architectures, aiming to enhance their performance in recognizing a wide array of entities in Vietnamese.

The contributions of this research are threefold. First, we introduce a method for training multilingual models tailored to open-domain NER, with a focus on their application to low-resource languages like Vietnamese. Second, we provide insights into the multilingual transfer and multitask learning capabilities of encoder-decoder models, offering a framework for their adaptation to various linguistic contexts. Finally, we present a newly created and cleaned open-domain Vietnamese NER dataset, which serves as a useful resource for future research in this area. Through these contributions, this study advances our understanding of NER in low-resource languages and paves the way for further exploration in other underrepresented linguistic settings.

* Corresponding author: huonglt@soict.hust.edu.vn

2 Related Work

In English, established NER models such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and spaCy (Honnibal and Montani, 2017) excel at identifying fixed entity types since they benefit from extensive annotated datasets and advanced pre-training techniques. Vietnamese NER has also seen significant advancements, particularly through models contributed by VinAI (Nguyen and Nguyen, 2020; Dao and Nguyen, 2020), the VLSP workshop (Ha et al., 2022), and the community. Despite these advancements, Vietnamese NER datasets remain limited to fixed entity types, similar to those in English and other languages. Less effort has been directed toward developing open-domain NER systems capable of recognizing a broader range of entities.

UniversalNER (Zhou et al., 2024) has recently explored a new approach involving targeted distillation with mission-focused instruction tuning to train student models like LLaMA (Touvron et al., 2023). These models can excel in open-domain NER tasks by being distilled from large models like ChatGPT, achieving promising results. However, these efforts are limited to English and involve large model sizes, making them impractical for many applications. Additionally, the transferability of these models to languages with limited datasets, like Vietnamese, remains unexplored.

Multilingual models like mT5 (Xue et al., 2021) and mT0 (Muennighoff et al., 2023) offer a promising avenue for cross-lingual NER tasks. These models, built on the Transformer architecture, have shown proficiency in handling multiple languages simultaneously. Recent developments in cross-lingual transfer learning and fine-tuning have demonstrated that multilingual models can effectively leverage data from high-resource languages (such as English) to improve performance in low-resource languages (like Vietnamese). Many studies on multilingual models have explored various strategies to enhance performance across languages. Using self-supervised learning techniques to pre-train on extensive multilingual corpora followed by task-specific fine-tuning has proven effective. However, the application of these methods to open-domain NER remains limited.

This research seeks to build on the current state of multilingual NER by focusing on smaller, efficient models (mT5 and mT0) and maximizing the use of available English data and other available Vietnamese datasets to compensate for the lack of Vietnamese NER datasets. The research aims to contribute a practical approach to recognizing a wide range of entity types in Vietnamese, addressing open-domain challenges, and ensuring the models remain accessible and efficient for broader applications.

3 Method

Traditional NER models use tagging styles like IOB or IOB2, where tokens are labeled to indicate their position within an entity. These models work well with fixed entity types but struggle with open-domain NER, where texts may contain diverse and ambiguous entities. Unlike traditional methods that identify and classify tokens into different entity types, our method focuses on type-specific extraction, resulting in a list of entities of the specified type rather than requiring the identification and categorization of spans into multiple types, as shown in Figure 1.

For this task, we chose mT5 and mT0, multilingual encoder-decoder models that leverage both English and Vietnamese datasets. Encoder-only models were excluded due to their lack of text-generation capabilities. Although decoder-only models can be applicable in some NER contexts, they are generally weaker for structured extraction tasks. Their autoregressive nature is optimized for generating text rather than for extracting specific entities from a text. Encoder-decoder models, on the other hand, provide a more robust framework for identifying entities by leveraging both the understanding of input context and the generation of precise outputs. The methodology involves two steps: first, comparing mT5 and mT0 to identify the better base model for NER by fine-tuning each under various configurations; second, developing a fine-tuning strategy for open-domain NER in Vietnamese. In this process, we fine-tune the pre-trained encoder-decoder models using different settings by leveraging existing English and Vietnamese datasets. Additionally, we create a comprehensive Vietnamese open-domain NER dataset to enhance model performance in this task. Each approach will be evaluated using the F1 score to determine the most effective method for robust NER performance in Vietnamese.

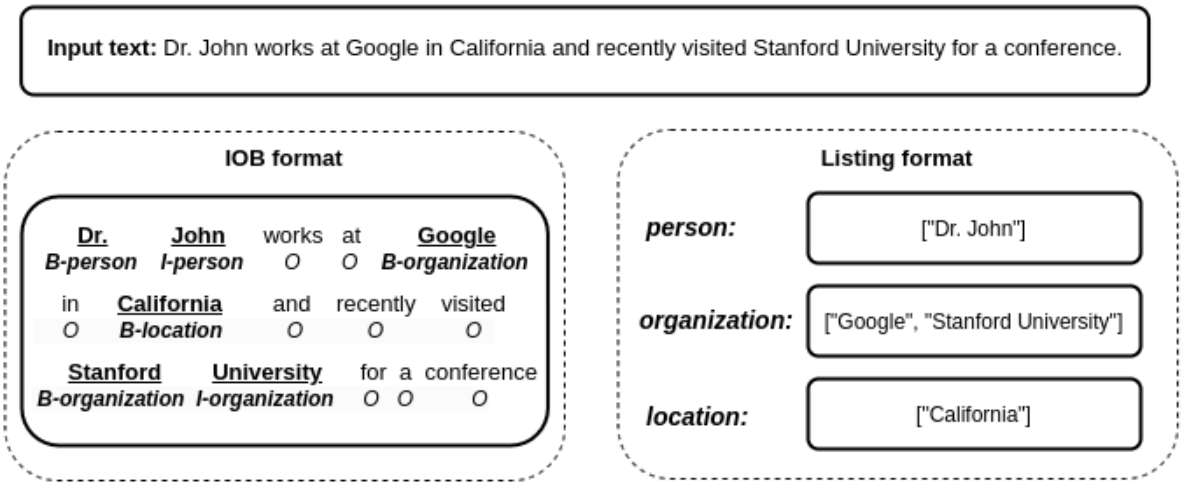

Figure 1: Comparison of NER labeling styles: Tagging (left) vs Extracting (right)

3.1 Dataset Preparation

This section provides an overview of the datasets used for fine-tuning and evaluation. We will employ five datasets for fine-tuning purposes: two open-domain NER datasets (one in English and one in Vietnamese), two instruction-tuning datasets (one in English and one in Vietnamese), and one Vietnamese question-answering dataset. For evaluation, we will utilize three GOLD datasets: one English NER dataset and two Vietnamese NER datasets. Detailed descriptions of these datasets and their specific roles in the fine-tuning and evaluation processes will be provided in the following sections.

3.1.1 Datasets for Multi-tasking and Multi-lingual Training

The English open-domain NER dataset used in this research is derived from the Pile-NER-Type dataset developed by UniversalNER (Zhou et al., 2024). This dataset, created from the Pile corpus using GPT-3.5, includes a wide range of entity types without a predefined set. To align with the sequence-to-sequence models (mT0 and mT5) used in this study, the dataset was reformatted from its original conversation-style format into instruction-input-response prompts, as shown in Figure 2, which were inspired by Alpaca dataset (Taori et al., 2023). This process resulted in 354,261 samples, each containing a prompt and an output string listing the extracted entity mentions, making it suitable for training the models effectively.

In this research, we utilize two instruction-tuning datasets to enhance the model's capability to follow instructions. The first dataset is an upgraded version of the Stanford Alpaca dataset, which comprises 52,000 instruction-following examples generated using GPT-4[1], whereas the original was generated using GPT-3.5[2]. The Vietnamese dataset, known as the Vietnamese Alpaca (Nguyen et al., 2024), consists of 50,000 varied instructions in Vietnamese, generated using GPT-4, following a methodology similar to that used for the English Alpaca dataset (Taori et al., 2023). Both datasets undergo preprocessing to fit a common instruction prompt template, ensuring consistency and sim-

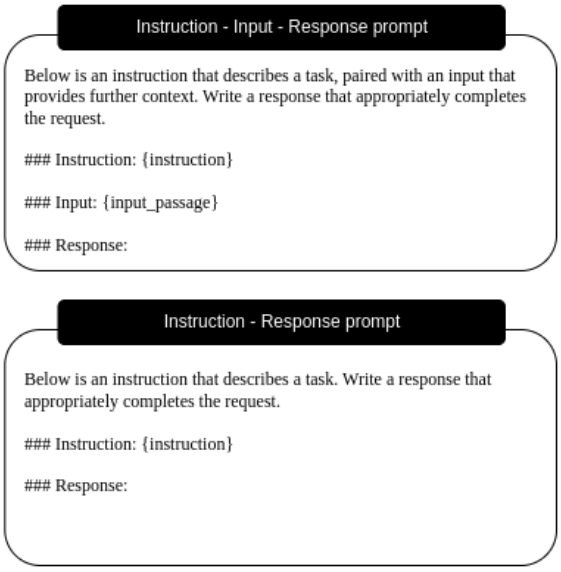

Figure 2: Instruction-input-response prompt template

[1] https://github.com/Instruction-Tuning-with-GPT-4/GPT-4-LLM

[2] https://huggingface.co/datasets/tatsu-lab/alpaca

plicity. The final datasets retain samples where the prompt and label lengths are within 1024 tokens, facilitating effective fine-tuning of sequence-to-sequence models.

Moreover, we utilize the UIT-ViQuAD v1.1 dataset (Nguyen et al., 2020), a benchmark designed to evaluate machine reading comprehension (MRC) in Vietnamese. This dataset comprises over 23,000 human-generated question-answer pairs based on 5,109 passages extracted from 174 Vietnamese Wikipedia articles. Developed through a rigorous process involving the recruitment, training, and validation of workers, the UIT-ViQuAD dataset ensures high-quality, diverse, and relevant content. It serves as a crucial resource for advancing MRC models in the Vietnamese language. In our research, UIT-ViQuAD is used as an additional task in the multi-tasking fine-tuning process of developing our open-domain NER model for Vietnamese.

3.1.2 Vietnamese Open-domain NER Dataset

To minimize costs, we used LLaMA 3 70B from Meta to generate data instead of ChatGPT-3.5. We randomly sampled 6,600 passages from the BKAI News Corpus dataset [3]. These passages are raw text and have not yet been annotated with labels. These samples were concatenated and split into 36,000 smaller passages, each ranging from 150 to 256 tokens in length, and were required to contain at least one complete sentence to maintain textual integrity.

The prompt used to generate data was inspired by the approach in (Zhou et al., 2024), but modified to suit Vietnamese data (see Figure 3). The generation temperature was set to 0 during the data creation process to ensure consistency and stability. This process yielded 34,274 samples, each with two attributes: the input passage and a list of entities extracted by LLaMA 3. Following the same procedure applied to the English open-domain NER dataset, we then split the Vietnamese data into samples containing one entity type per example, resulting in 136,895 samples.

The entity types identified by LLaMA 3 were initially quite varied, with many referring to the same concept but represented differently. Some entity types were formatted in code-like terms, such as "*entity_type:person*" and "*entity_type:field_of_study*", then underwent a pre-

[3] https://huggingface.co/datasets/bkai-foundation-models/BKAINewsCorpus

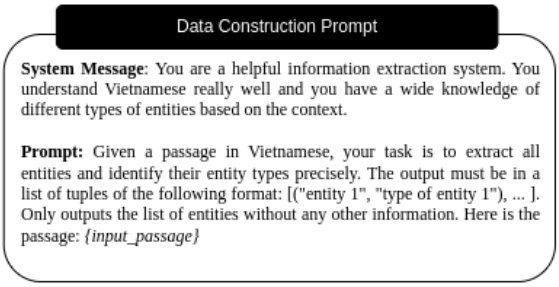

Figure 3: Data construction with LLaMA 3 prompt template

processing step to reformat these entity types to make them more natural and user-friendly. Entity types extracted in other languages, such as Chinese and those with unclear meanings were manually reviewed and removed. Additionally, samples containing hallucinated entities (entities not present in the input text) were also removed.

Finally, the dataset was filtered to remove samples where the input text or label exceeded 1024 tokens, similar to the English open-domain NER dataset. This step ensured compatibility for training sequence-to-sequence models. The final dataset consists of 125,518 samples, covering 3,522 different entity types across a wide range of domains.

The distribution of entity types followed a heavy tail pattern, with the top 1% of entity types accounting for 71% of the total frequencies. While the most common entity types are *organization*, *location*, and *person*, the dataset also includes rarer entity types such as *country*, *concept*, and *document*. Notably, out of the 3,522 distinct entity types in the dataset, more than 2,000 of them, which are not typically used in traditional NER tasks, appear only once. This diverse coverage is crucial for developing models capable of handling open-domain NER tasks.

3.1.3 GOLD Datasets

We utilize several high-quality datasets for supervised fine-tuning and evaluation of the proposed Vietnamese open-domain NER model. These datasets include UNER English EWT (Mayhew et al., 2024a), PhoNER_COVID19 (Truong et al., 2021), and VLSP NER 2021 (Ha et al., 2022). Each dataset is already divided into training, development, and test sets when it was published, and these pre-defined splits are used for fine-tuning and evaluation in our experiments.

The **PhoNER_COVID19** dataset (Truong et al.,

2021) is a COVID-19 domain-specific NER dataset for Vietnamese, developed with newly-defined entity types. This dataset comprises 10,000 sentences containing over 35,000 entities, categorized into 10 specific entity types. These entity types are designed to extract key information related to COVID-19 patients. The PhoNER_COVID19 dataset is used to benchmark our model's performance on domain-specific entities.

The **VLSP NER 2021** dataset (Ha et al., 2022) is a comprehensive resource for evaluating NER models in Vietnamese, specifically designed to assess the ability to recognize entities across 14 main types, 26 subtypes, and 1 generic type. The dataset includes a total of 2,140 annotated articles, drawn from diverse domains such as life, science and technology, education, sport, law, and entertainment. It is divided into a training set of 1,830 articles, which includes 81,173 named entities (with 1,282 articles from the VLSP 2018 NER dataset and 538 new articles), and a test set of 310 new articles, containing 19,538 named entities. This dataset is instrumental in benchmarking NER models in the Vietnamese language for general, rather than domain-specific, entity recognition tasks.

The **UNER English EWT** dataset (Mayhew et al., 2024a) derived from the multilingual NER benchmark (Mayhew et al., 2024b), provides a gold-standard resource for evaluating NER systems in English. The dataset comprises 5,985 samples, partitioned into 4,592 training samples, 646 development samples, and 747 test samples. It includes annotations for three entity types: location, organization, and person. This benchmark is instrumental for assessing NER models' performance across various languages, particularly when the models are fine-tuned exclusively on English data or in conjunction with other languages and tasks.

3.2 Base Model Selection

The first step in developing an effective open-domain NER model for Vietnamese involves selecting an appropriate base model and assessing its initial performance through fine-tuning. This section outlines the process of selecting between mT5 and mT0 models, followed by the initial fine-tuning procedure.

To determine the most suitable model, both the base and large versions of mT5 and mT0 were fine-tuned. This approach aimed to identify the better-performing model type between mT5 and mT0, and to observe the behavior of different model sizes, as shown in Figure 4. For the mT5 model, two fine-tuning configurations were performed. The first configuration involved fine-tuning the mT5 model exclusively on the open-domain English NER dataset. The second configuration entailed initially fine-tuning the mT5 model on a mixture of English and Vietnamese instruction-tuning datasets, based on the hypothesis that the mT5, being a pre-trained model, might benefit from an initial phase of instruction-following fine-tuning. Subsequently, the model was fine-tuned on the open-domain English NER dataset.

For the mT0 model, which is already fine-tuned on multi-tasking data, direct fine-tuning on the open-domain English NER dataset was performed to evaluate its performance. The fine-tuned models were then assessed on both the English and Vietnamese NER datasets (the GOLD datasets) to determine their overall performance in these two languages. It is important to note that the open-domain English NER dataset generated by ChatGPT was not used for evaluation as it is not considered a GOLD standard dataset. Based on the evaluation results, the better-performing model was selected for subsequent fine-tuning stages.

3.3 Advanced Fine-tuning

After selecting the better-performing model from the initial fine-tuning process, the next step involves advanced strategies to enhance the model's performance. The steps involved in fine-tuning the model are chosen using multi-task learning and two-stage fine-tuning, as illustrated in Figure 5. The fine-tuned models are evaluated in two ways: zero-shot setting and supervised fine-tuning, which will be discussed below.

3.3.1 Multi-task Learning

Multi-task learning involves training the model on multiple related tasks to enhance its ability to understand the text and generate accurate responses for those tasks. In this study, the NER task is formulated as a sequence-to-sequence problem, where the input sequence includes the input passage along with a question asking the model to extract a specific entity type, making the task similar to a question answering problem. Therefore, we decided to fine-tune the selected model from the previous process with a mix of the open-domain English dataset and the Vietnamese question-answering dataset. The reason for choosing the Vietnamese question-answering dataset is that it is not only a related

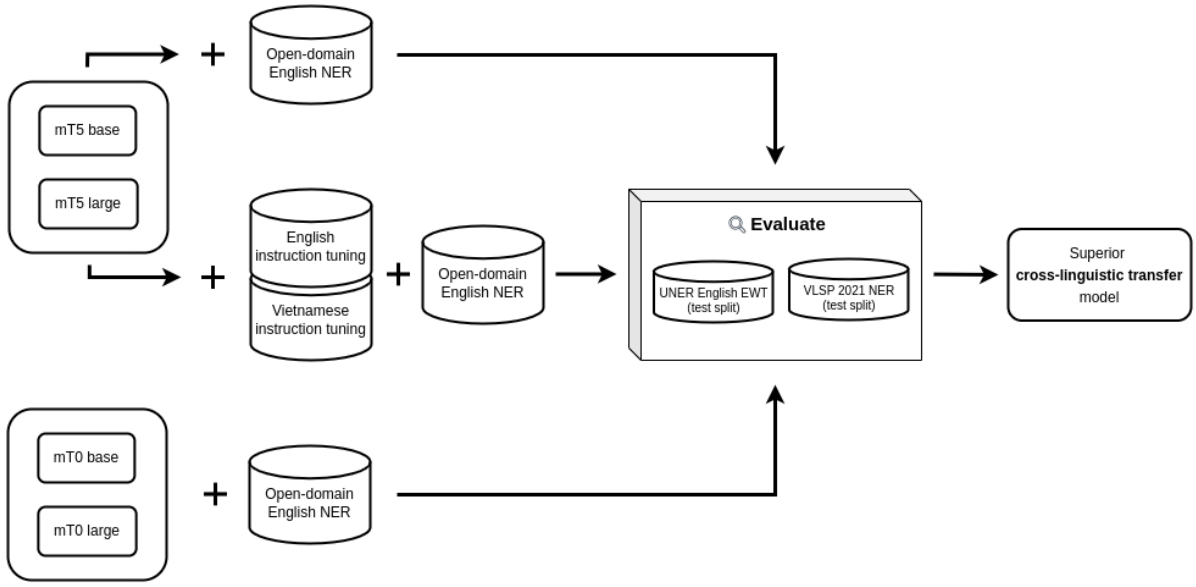

Figure 4: Base model selection steps

task but also in Vietnamese, the target language for improved model performance. The expectation is that the model can leverage the high-quality English NER data to enhance its performance on Vietnamese text by training in both languages simultaneously, allowing the model to learn patterns and features common to both languages.

3.3.2 Two-stage Fine-tuning

Instead of fine-tuning the model once, a two-stage fine-tuning strategy was designed. Initially, the model is fine-tuned similarly to the multi-task learning strategy, but uses a large portion of the English open-domain NER dataset for training. After the initial training, the model is further fine-tuned with the remaining portion of the English open-domain NER dataset and the Vietnamese open-domain NER dataset. This approach leverages the extensive English data in the first stage and utilizes the multi-task learning strategy to learn both the NER task and the Vietnamese language. The second stage prioritizes the Vietnamese language by using a larger portion of Vietnamese data compared to English data, enhancing the model's performance in Vietnamese while retaining its knowledge of the NER task in English.

3.3.3 Evaluation

The models fine-tuned using the above strategies are evaluated on both English and Vietnamese GOLD NER datasets, with a particular focus on performance on Vietnamese datasets. Two Vietnamese datasets are used for evaluation: the PhoNER_COVID19 dataset and the VLSP NER 2021 dataset.

Evaluation is conducted in two phases: zero-shot and supervised fine-tuning. Initially, models are evaluated on the two datasets without training on their respective training splits to assess their zero-shot capability. Subsequently, the models undergo supervised fine-tuning on the training data of each evaluation dataset to evaluate their performance after learning domain-specific data.

4 Results and Discussion

4.1 Evaluation Parameters

We use the F1 score as the evaluation metric to assess overall model performance. Unlike traditional NER models that use tagging formats like IOB to extract and classify entity spans, the open-domain NER model evaluates by identifying entities of a single type from the input text. Instead of tagging, the model outputs a list of entities, which simplifies the evaluation and adapts to the open-domain task, allowing for easier comparison against a gold-standard dataset.

4.2 Simulation Method

4.2.1 Base Model Selection

For the fine-tuning process involving the mT5 model, data from the English and Vietnamese instruction-tuning datasets were randomly mixed before fine-tuning all data of the open-domain English NER dataset. The mT0 model, which shares

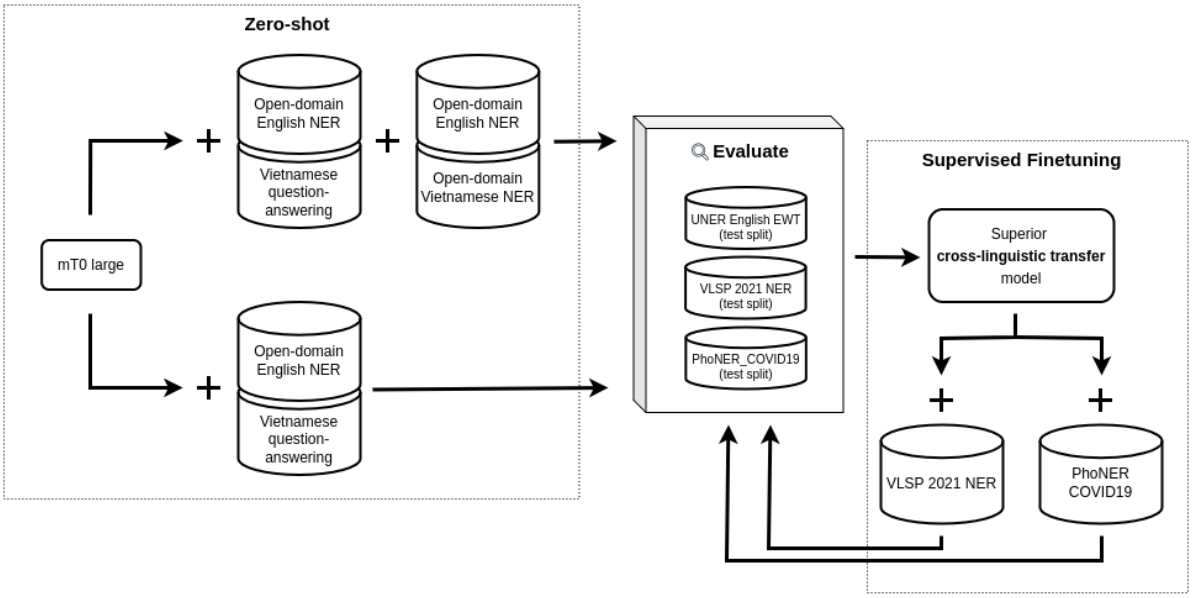

Figure 5: Advanced Fine-tuning

the same tokenizer as the mT5, underwent a fine-tuning process involving only the open-domain English NER dataset.

During training, data were tokenized and padded to match the length of the longest sequence in each batch. Given that the model sizes were manageable, parameter-efficient fine-tuning (PEFT) strategies were not necessary. Thus, a supervised fine-tuning (SFT) strategy was applied, involving updating all model parameters. All models were fine-tuned with a batch size of 256 and a constant learning rate of 0.0001 over one epoch, in line with the approach reported in the mT5 paper for both pre-training and fine-tuning stages.

Upon completion of the fine-tuning stages, six fine-tuned models were obtained. These models were evaluated using the test splits from the UNER English EWT dataset and the VLSP 2021 NER dataset. The evaluation process involved comparing the predicted list of entities to the target entities, with the F1 score used as the primary metric to assess the model's ability to identify correctly entities of a given type.

4.2.2 Advanced Fine-Tuning Strategy

For the multi-task fine-tuning strategy, the best-performing model was fine-tuned using a mixture of the English open-domain NER dataset and the Vietnamese question-answering dataset. All data from both datasets were used in the fine-tuning process. The English dataset constituted the majority, with 318,261 samples, while the Vietnamese

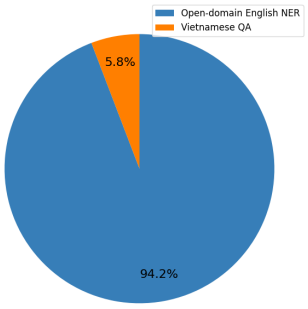

Figure 6: Multitask data proportion

dataset contributed 19,608 samples. These were randomly mixed before training. The same batch size of 256 and a constant learning rate of 0.0001 over one epoch were employed as in the base model selection step.

In the two-stage fine-tuning strategy, different datasets were used in each stage. For the second stage, 113,161 samples from the Vietnamese open-domain NER dataset were utilized. To ensure the model retains its English knowledge while enhancing its Vietnamese proficiency, 25,261 English samples (one-fourth of the Vietnamese dataset size) from the English open-domain NER dataset were reserved for mixed-language fine-tuning. In the first stage, the remaining English open-domain NER dataset (293,000 samples) was used for multi-task training with the Vietnamese question-answering dataset. The data distribution for this two-stage fine-tuning strategy is illustrated in the pie charts in Figure 7.

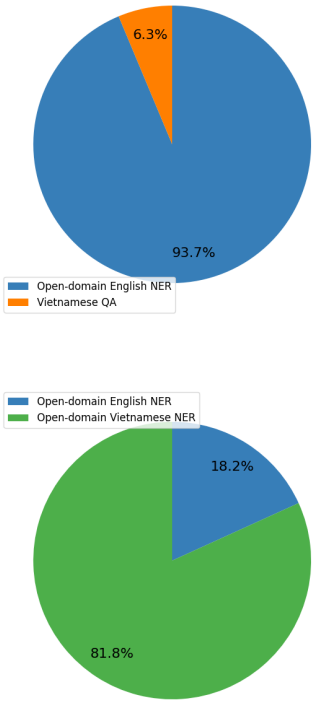

Figure 7: Two-stage data proportion

By adopting these fine-tuning strategies, the goal was to improve the model's performance in Vietnamese open-domain NER tasks by being able to recognize a wide range of entity types.

4.3 Base Model Selection Results

The results of the base model comparison are presented in Table 1, where the performance of each model is evaluated on both English and Vietnamese GOLD datasets for the NER task. Among the models tested, mT0-large shows the most promising results, with an F1 score of 0.7852 in English and 0.4518 in Vietnamese, despite being fine-tuned exclusively on English data. This model demonstrates strong cross-lingual transfer capabilities, outperforming the other models in Vietnamese NER.

mT5 models generally perform well in English but struggle in Vietnamese, especially without exposure to Vietnamese during training. Introducing mixed-language instruction tuning before fine-tuning slightly improves performance in both languages. However, mT0-large's superior results suggest that it is the best candidate for further fine-tuning, particularly for enhancing cross-lingual NER performance in Vietnamese. The next steps will focus on refining this model by incorporating more Vietnamese data.

4.4 Advanced Fine-tuning Strategy Results

The mT0-large model was further fine-tuned using strategies that leveraged English and Vietnamese datasets for developing an open-domain NER system. This section reports the model's performance in two settings: zero-shot evaluation and supervised fine-tuning.

4.4.1 Zero-shot Evaluation

The mT0-large model, fine-tuned with a mix of open-domain English NER data and Vietnamese question-answering data, achieved an F1 score of 0.7775 on English NER, slightly lower than when fine-tuned solely on English data. However, the model's performance on the VLSP NER 2021 dataset improved significantly, with an F1 score of 0.5259, indicating that incorporating Vietnamese data, even from a different task, enhances its ability to recognize Vietnamese entities. On the PhoNER_COVID19 dataset, the model's F1 score was 0.4679, which is lower than other models like BiLSTM-CRF and XLM-R, likely due to the domain-specific nature of PhoNER_COVID19.

A two-stage fine-tuning strategy consisting of learning from open-domain English NER and Vietnamese question-answering, followed by fine-tuning on mixed English and Vietnamese NER data yielded better results. This approach achieved the highest F1 score on English NER and improved the F1 scores on VLSP NER 2021 and PhoNER_COVID19 by 0.08 and 0.11, respectively. Although these results in a zero-shot setting may not seem groundbreaking, they demonstrate the potential of the two-stage fine-tuning strategy for cross-lingual NER tasks.

4.4.2 Supervised Fine-tuning Evaluation

Given its superior performance in the zero-shot evaluation, the two-stage fine-tuned mT0-large model was further evaluated in a supervised setting.

When fine-tuned on the VLSP NER 2021 dataset, the mT0-large model outperformed most models submitted by VLSP participants, achieving an F1 score of 0.6030 in a zero-shot setting and 0.7489 after fine-tuning on the full training data. This superior performance can be attributed to the model's exposure to a wide range of entities during pre-training, facilitating better recognition of the diverse entity types in the VLSP dataset.

	UNER English EWT	VLSP NER 2021
mT5-base en-open-NER	0.7440	0.3032
mT5-large en-open-NER	0.7615	0.4105
mT5-base mIT + en-open-NER	0.7579	0.3332
mT5-large mIT + en-open-NER	0.7747	0.4193
mT0-base en-open-NER	0.7574	0.3035
mT0-large en-open-NER	**0.7852**	**0.4518**

Table 1: Base model performance comparison

	English NER	VLSP NER 2021	PhoNER_COVID19
mT0-large en-open-NER mix vi-QA	0.7775	0.5259	0.4679
mT0-large en-open-NER mix vi-QA + mNER	0.8074	**0.6030**	**0.5753**

Table 2: Zero-shot evaluation

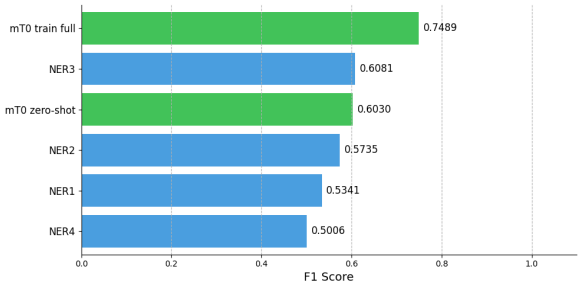

Figure 8: SFT results on VLSP NER 2021

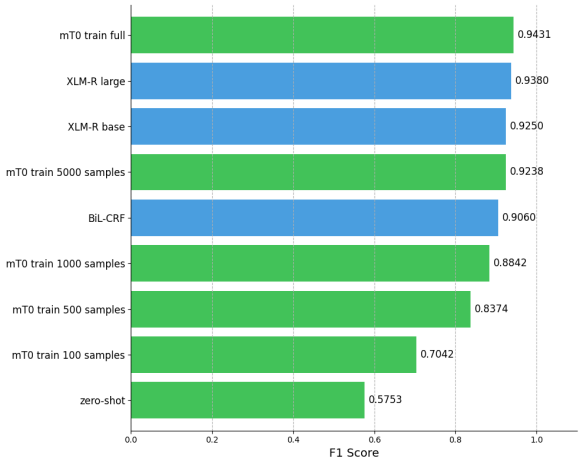

Figure 9: SFT results on PhoNER_COVID19

For the PhoNER_COVID19 dataset, the mT0-large model was fine-tuned using different sample sizes. Even with only 10 samples per entity type, the model's F1 score increased significantly from 0.5753 to 0.7042. With full training data, the model achieved an F1 score that surpassed all models evaluated in the original PhoNER_COVID19 publication, including BiLSTM-CRF and XLM-R. Despite being smaller, these models are domain-specific and might not generalize well to open-domain NER tasks. In contrast, the mT0-large model, with its larger capacity, effectively leveraged even small amounts of in-domain data to excel in domain-specific NER tasks.

5 Conclusion

In this research, we conducted a study on developing an open-domain NER model for Vietnamese, using it as a case study for low-resource languages. By experimenting with multilingual encoder-decoder models, particularly the mT5 model and the mT0 one, we found a novel strategy to fine-tune the mT0-large model to perform well on open-domain NER tasks. This model demonstrated a strong ability to generalize to Vietnamese NER, even when fine-tuned exclusively on English data, showcasing the potential of medium-sized models and promising application to other low-resource languages.

These findings underscore the potential of the proposed approach but also highlight areas needing further refinement. Future research could focus on improving fine-tuning strategies, creating higher-quality open-domain Vietnamese NER datasets, exploring decoder-only models and investigating domain adaptation and few-shot learning techniques. Such efforts would further enhance the model's performance and adaptability, particularly in real-world applications.

Acknowledgments

This work was supported by the 2024 Ministry-level Science and Technology project, code B2024-KHA-06, under the Ministry of Education and Training.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hoang Mai Dao and Quoc Dat Nguyen. 2020. VinAI at ChEMU 2020: An Accurate System for Named Entity Recognition in Chemical Reactions from Patents. In *Proceedings of the Working Notes of CLEF 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Ralph Grishman. 2019. Twenty-five years of information extraction. *Natural Language Engineering*, 25(6):677–692.

My Linh Ha, Thi Minh Huyen Nguyen, Dung Doan Xuan, et al. 2022. VLSP 2021-NER Challenge: Named Entity Recognition for Vietnamese. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(1).

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten de Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *Advances in Information Retrieval*, pages 705–710. Springer Berlin Heidelberg.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024a. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337. Association for Computational Linguistics.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024b. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337. Association for Computational Linguistics.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 51–58. Australasian Language Technology Association.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir R. Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Annual Meeting of the Association for Computational Linguistics*.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese Dataset for Evaluating Machine Reading Comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain. International Committee on Computational Linguistics.

Quang Duc Nguyen, Hai Son Le, Duc Nhan Nguyen, Dich Nhat Minh Nguyen, Thanh Huong Le, and Viet Sang Dinh. 2024. Towards Comprehensive Vietnamese Retrieval-Augmented Generation and Large Language Models. *arXiv preprint arXiv:2403.01616*.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hung Thinh Truong, Hoang Mai Dao, and Quoc Dat Nguyen. 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*, pages 2146–2153. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Imed Zitouni Barua,

and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition. *Preprint*, arXiv:2308.03279.