VietMed-NER: Medical Named-Entity-Recognition from Speech

Khai Le-Duc[*1], **David Thulke**[*1,2], **Hung-Phong Tran**[3], **Long Vo-Dang**[4], **Ralf Schlüter**[1,2]
[1]Lehrstuhl Informatik 6 - Machine Learning and Human Language Technology Group,
Computer Science Department, RWTH Aachen University, 52074 Aachen, Germany
[2]AppTek GmbH, 52062 Aachen, Germany
[3]Hanoi University of Science and Technology, Hanoi, Vietnam
[4]University of Cincinnati, Cincinnati, Ohio, United States
dle@i6.informatik.rwth-aachen.de, {thulke,schlueter}@hltpr.rwth-aachen.de

Abstract

One of the tasks within Spoken Language Understanding (SLU) is Named Entity Recognition (NER) from speech, which aims to extract semantic information from the speech signal. In this work, we present *VietMed-NER* - the first public NER from speech dataset in the medical domain. To the best of our knowledge, *VietMed-NER* is also the first public dataset for NER in Vietnamese spoken text. Our real-world dataset contains 19 entity types, which is by far the largest number compared to existing Vietnamese NER datasets for written text. Furthermore, we introduce a novel annotation technique, called Recursive Greedy Mapping, to enhance annotation speed and quality. Finally, we present baseline results using various state-of-the-art pre-trained models. To assist international researchers, both an English-translated version and a Vietnamese version of the dataset are accessible. All code, data and models are made publicly available here.

1 Introduction

Named Entity Recognition (NER) targets extracting Named Entities (NE) into categories like person, location, organization, etc. Initially studied in written language, where an NER tagger predicts both NE boundary and category jointly, recent attention has turned to studying NER from speech (Shon et al., 2022). It aims to extract semantic information from speech with many applications, such as addressing privacy concerns in medical recordings (e.g., muting specific words like patient names) (Cohn et al., 2019), NER from speech has limited literature (Yadav et al., 2020).

NER from speech is particularly challenging due to the impact of word segmentation on results (Chen et al., 2022). The Vietnamese vocabulary poses additional difficulties with numerous confused monosyllabic and polysyllabic words For instance, the word "đường" alone could denote "sugar" (chemical), "street" (location), or be part of a compound word like "đường tiêu hóa" - "gastrointestinal" (anatomy). Moreover, obtaining accurate human transcriptions and model recognition for medical NER from natural speech is hard due to the absence of punctuation and special characters, speech disfluency, lack of speaking context, and the complexity of medical terms.

There are 2 existing datasets for NER on Vietnamese written language: VLSP family (Huyen and Luong, 2016; Nguyen et al., 2018, 2020) and PhoNER_COVID19 (Truong et al., 2021). However, NER on Vietnamese spoken text has not been widely studied due to the lack of a publicly available dataset. As for the medical domain, to our knowledge, there is no dataset available for medical NER from speech. The only related work we found, (Cohn et al., 2019), published an NER evaluation benchmark using an English general-domain conversational dataset, Switchboard (Godfrey et al., 1992) and Fisher (Cieri et al., 2004), for the task of audio de-identification specifically targeting Personal Health Identifiers.

To address this gap, we introduce *VietMed-NER*, a medical NER dataset from speech built on the real-world medical ASR dataset *VietMed*, featuring newly-defined entity types for potential use as a foundational task in various real-world applications, such as: error correction for medical ASR output (Mani et al., 2020), audio de-identification, privacy-preserved machine learning (e.g. for model training and public dataset creation), and virtual assistant (e.g. search engine (Rüd et al., 2011), classifying content for news providers (Kumaran and Allan, 2004), and recommending content (Koperski et al., 2017)). Our contributions are as follows:

- We present *VietMed-NER* - the first public medical NER from speech dataset. To the best of our knowledge, *VietMed-NER* is also the

[*]Equal contribution

first NER dataset for Vietnamese spoken text.

- We introduce a novel annotation technique, called Recursive Greedy Mapping, to improve annotation speed and quality.
- We present baselines on several state-of-the-art pre-trained models.

All code, data (English-translated and Vietnamese) and models are published online[1,2].

2 Data

2.1 Data Collection

We choose *VietMed* dataset (Le-Duc et al., 2023) - a real-world medical ASR dataset in Vietnamese - to annotate NEs. The reason for this is that *VietMed* is the world's largest and the most generalized medical ASR dataset for now.

Due to privacy concern, the original *VietMed* contains no text that might reveal the identities of speakers, especially names. The names in the ASR dataset are only from famous people, so making very small number. We therefore added a very small extra number of synthetic sentences with people names (but mimic the natural flow of real-world speech). We did the same with several other entity types, aiming to make the test set to have more entities compared to train and dev set, which results in our dataset slightly larger than the original.

2.2 Annotation Process

Annotation of medical NEs from real-world speech is challenging because of the missing punctuation, special characters and capitalized words in ASR transcript, disfluency and medical knowledge. Entirely manual annotation of NEs like in VLSP dataset (Huyen and Luong, 2016; Nguyen et al., 2018, 2020) and PhoNER_COVID19 (Truong et al., 2021) requires a large number of working hours, not to mention the difficulties in quality control and inconsistency as we found in their corpora. Moreover, the machine learning approach, using some previously fine-tuned models to help pre-tagging, does not apply to our newly-defined medical entity types. Likewise, we used prompt engineering to leverage large language models like GPT-4 to pre-tag the dataset but did not achieve an acceptable correctness. Also, the idea of training a seed model using a gazetteer list (a list of entities) to help pre-tagging the rest of dataset (Kozareva, 2006) requires an initial training time and the performance is not reliable due to the model being trained on a small amount of initial data. Inspired by these works, we propose another method for annotation, called Recursive Greedy Mapping. The method is described as below:

1. Collect some initial entities, classify them into relevant entity types, and put them into a gazetteer list.
2. Sort all entities based on length of characters, from highest to lowest (Time complexity = O(n.log(n))). Some NEs are part of longer NEs, which we call sub-NEs and full NEs. This sorting step is to make sure to not have sub-NEs mapped before full NEs. For example, "tooth pain" should be mapped before "pain".
3. Automatically map entities with tags from the gazetteer list to transcript (Time complexity = O(m.n)). Pseudo code:

```
for NE in gazetteer_list:
  for sen in sentences:
    if NE in sen:
      annotate(NE, sen)
```

4. Annotators go through each sentence and check wrongly pre-tagged labels, which occur mostly because rule-based algorithm do not consider context in sentences.
5. Annotators continue adding NEs that are not in the gazetteer list which are found during manual annotation. The algorithm in step 2 and 3 will automatically generate new pre-tagged labels in following sentences. Annotators then get back to step 4 and 5 until the entire corpus is annotated by human.

2.3 Data Quality Control

We developed an initial version of our annotation guidelines (see Supplementary Materials) and then starting annotate the corpus. Two developers, one of them have medical background, annotate the corpus independently. We then hosted a discussion session to resolve annotation conflicts, identified complex cases, and refined the guidelines. Other two developers will conduct the quality control based on the provided guidelines and annotated corpus. We then constantly revisited each sentence

[1] https://github.com/leduckhai/MultiMed
[2] https://github.com/rwth-i6/returnn-experiments

Label	Definition	Train	Dev	Test	All	Uni.
NAME	Person name	1	2	28	31	31
AGE	Age of a person	185	372	622	1179	93
GEND.	Gender of a person	32	172	558	762	29
JOB	Job of a person	402	274	565	1241	65
LOC.	Locations and places	112	196	328	636	105
ORG.	Organization, e.g. companies, governments	3	13	61	77	36
SYMP.	Symptoms and diseases	1484	1649	1467	4600	555
CHEM.	Bio-chemical subtances and drugs	419	700	709	1828	248
F&B	Food and beverage	90	60	261	411	51
ANAT.	Anatomical features, e.g. organs, cells	767	1223	1255	3245	225
PC	Personal care, e.g. hygiene routines, skin care	37	60	95	192	50
DX	Diagnostic procedures, e.g. lab tests, imaging	180	191	299	670	62
TX	Non-surgical treatment, e.g. rehab., injection	444	147	232	823	61
SX	Surgical procedures, e.g. implants, neurosugery	225	20	276	521	59
PREV.	Preventive medicine, e.g. vaccinations, screenings	2	17	18	37	14
TECH.	Medical devices, instruments, and techniques	127	132	634	893	134
CAL.	Medical calibration, e.g. number of doses, calories	209	357	257	823	203
TRAN.	Means of transportation	6	5	27	38	19
TIME	Date and time	438	505	667	1610	212
#Entities in total		5163	6095	8359	19617	2252
#Sentences		2861	2913	3497	9271	-

Table 1: Entity definition and its statistics in our dataset. "Uni." means the number of unique entities, which equals to the length of gazetteer list.

in the entire annotated corpus multiple times. Because *VietMed-NER* is an ongoing effort, we expect the dataset to evolve. In the long run, we crowdsource for the quality control and continue updating if there's an error reported.

2.4 Data Statistics

Table 1 shows the statistics of our dataset. The ASR corpus *VietMed* was split into train-dev-test as 5-5-6 hours. Our *VietMed-NER* contains 19 entity types, with a total number of sentences of 9k. Compared to other 2 public Vietnamese NER datasets, ours has by far the largest number of entity types, to the best of our knowledge, while our total number of sentences is relatively comparable.

Most NER datasets have a very small number of entities in test sets compared to train and dev set (Huyen and Luong, 2016; Truong et al., 2021; Chen et al., 2022). However, we want to best leverage the power of large pre-trained models which are trained on a vast amount of unlabeled text data thus making quite good representations. Therefore, we made the test set to have most number of entities as described in subsection 2.1, which explains the heavy imbalance in favor of test set in Table 1.

	WER		Pre-trained data (hours)
	dev	test	
w2v2-Viet	25.9	29.0	1204 Viet.
XLSR-53-Viet	25.9	28.6	56k multi. +1204 Viet.

Table 2: Statistics of 2 pre-trained ASR models: monolingual and multilingual. Both of them have the same number of parameters (118M) and were fine-tuned on 5h of dataset.

3 Experimental Setups

We employ the 2-stage pipeline to conduct NER from speech. An ASR model transcribes the audio into text. The transcribed text serves as the input for an NER model.

3.1 ASR Models

We employed 2 best baseline models fine-tuned for ASR task on *VietMed* published by (Le-Duc et al., 2023). Table 2 shows WERs of the 2 models.

We used RETURNN (Zeyer et al., 2018) for supervised training and Fairseq (Ott et al., 2019) for unsupervised wav2vec 2.0 training. Decoding was performed with RASR (Rybach et al., 2011).

Model	#Params	#Data
PhoBERT_base	135M	20GB
PhoBERT_large	370M	
PhoBERT_base-v2	135M	140GB
ViDeBERTa_base	86M	138GB
XLM-R_base	270M	2.5TB
XLM-R_large	550M	

Table 3: Statistics of state-of-the-art pre-trained language models which we used for NER task.

Fairseq models were converted to RETURNN models with our PyTorch-to-RETURNN toolkit[3].

3.2 NER Models

Table 3 shows the statistics of various pre-trained monolingual and multilingual models to fine-tune on our NEs in the dataset: $PhoBERT_{base}$, $PhoBERT_{large}$, $PhoBERT_{base-v2}$ (monolingual) (Nguyen and Nguyen, 2020), $ViDeBERTa_{base}$ (monolingual) (Tran et al., 2023), and $XML-R_{base}$, $XML-R_{large}$ (multilingual) (Conneau et al., 2020). To the best of our knowledge, these are the best pre-trained models which achieved state-of-the-art results on various downstream tasks in Vietnamese language, including NER.

We used HuggingFace (Wolf et al., 2019) for fine-tuning pre-trained models on NER task. Vietnamese input sentence can be represented in either syllable or word level as described by (Truong et al., 2021). However, we only employed word level settings. All our NER experiments were done using default hyperparameters by HuggingFace: learning rate of 2e-5, linear learning rate scheduler, training batch size of 64, training epoch of 50, weight decay of 0.01, optimizer AdamW (Loshchilov and Hutter, 2019), Beta1 of 0.9, Beta2 of 0.999, and epsilon of 1e-8.

3.3 Evaluation

For the evaluation of NER on reference text, we employed the seqeval[4] framework which is commonly used as a default evaluation framework by HuggingFace.

For evaluation of NER on ASR output, we employed F1 score calculation by (Shon et al., 2023) by using SLUE toolkit[5]. This F1 score is the harmonic mean of precision and recall, which evaluates an unordered list of named entity phrase and tag pairs predicted for each sentence.

[3]https://github.com/rwth-i6/pytorch-to-returnn
[4]https://github.com/chakki-works/seqeval
[5]https://github.com/asappresearch/slue-toolkit

Model	Prec.	Rec.	F1
PhoBERT_base	62.00	58.07	59.97
PhoBERT_large	62.50	59.57	61.00
PhoBERT_base-v2	63.21	60.82	61.99
+ w2v2-Viet	22.73	31.39	26.37
+ XLSR-53-Viet	22.84	31.17	26.36
ViDeBERTa_base	36.37	23.12	28.27
XLM-R_base	65.66	66.20	65.93
+ w2v2-Viet	27.49	20.77	23.66
+ XLSR-53-Viet	27.49	20.45	22.91
XLM-R_large	69.59	64.43	66.91
+ w2v2-Viet	21.14	28.19	24.16
+ XLSR-53-Viet	21.52	28.14	24.39

Table 4: Results of NER in percent for various pre-trained language models and ASR models on test set. Metrics used are Precision, Recall, F1 score and Accuracy.

4 Experimental Results

Table 4 shows results of NER using various pre-trained models. The pre-trained monolingual model PhoBERT_base-v2 outperformed other monolingual models, given the fact that it has a quite small number of parameters, thanks to being pre-trained on a large amount of monolingual text data. PhoBERT_base-v2 achieved 61.99% of F1 score on reference text, and 26.37% on ASR output, showing a large degradation because of ASR model.

On reference text, the pre-trained multilingual models XLM-R outperformed monolingual models ViDeBERTa and PhoBERT by a very large margin, at 65.93% of F1 score for XLM-R_base and 66.91% for XLM-R_large. This higher performance is partially due to the higher number of trainable parameters in XLM-R (270M for XLM-R_base compared to only 135M for PhoBERT_base-v2), and much higher amount of pre-trained data (2.5TB of multilingual data for XLM-R compared to only 140GB of monolingual data for PhoBERT_base-v2). However, on ASR output, the XLM-R performance was significantly degraded, at 23.66% of F1 score for XLM-R_base and 24.39% for XLM-R_large.

Nguyen and Nguyen (2020) and Truong et al. (2021) have proved the effectiveness of monolingual pre-trained language models on the language-

specific downstream tasks, especially on NER task. However, our results show that such statements could not be made, especially in a NER from speech dataset that has a mixture of Vietnamese and English language (medical terms), and more future experiments should be done to make a more conclusive statement.

5 Conclusion

In this work, we present *VietMed-NER* - the first public NER from speech dataset in the medical domain. To the best of our knowledge, *VietMed-NER* is also the first NER dataset for Vietnamese spoken text. Our dataset contains 19 entity types, ranging from conventional entity types like name, location, organization to newly defined for real-world medical conversations like symtom, anatomy, medical device... Our dataset is by far the largest number of entity types, while having a relatively comparable number of sentences compared to existing Vietnamese NER datasets for written text. Furthermore, we publish a novel annotation framework, called Recursive Greedy Mapping, to accelerate the annotation speed and improve the annotation quality. Finally, we found that pre-trained multilingual models XLM-R_base and XLM-R_large outperformed the best monolingual model PhoBERT_base-v2 on reference text (65.93% and 66.91%, compared to 61.99% of F1 score). However, on ASR output, multilingual models were outperformed by the monolingual model PhoBERT_base-v2 (26.37% compared to 23.66% and 24.39% of F1 score), with a smaller number of parameters (135M compared to 270M and 550M). We therefore recommend the use of PhoBERT_base-v2 for future experiments on medical NER from speech in Vietnamese language.

References

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. Aishellner: Named entity recognition from chinese speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*.

Ido Cohn, Itay Laish, Genady Beryozkin, Gang Li, Izhak Shafran, Idan Szpektor, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2019. Audio de-identification-a new entity recognition task. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Nguyen Thi Minh Huyen and Vu Xuan Luong. 2016. Vlsp 2016 shared task: Named entity recognition. *Proceedings of Vietnamese Speech and Language Processing (VLSP)*.

Krzysztof Koperski, Jisheng Liang, and Neil Roseman. 2017. Content recommendation based on collections of entities. US Patent 9,710,556.

Zornitsa Kozareva. 2006. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Student Research Workshop*.

Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 297–304.

Khai Le-Duc, Christoph Lüscher, Minh-Nghia Phan, Ralf Schlüter, and Hermann Ney. 2023. Vietmed: A dataset and benchmark for automatic speech recognition of vietnamese in the medical domain.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Anirudh Mani, Shruti Palaskar, and Sandeep Konam. 2020. Towards understanding ASR error correction for medical conversations. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*.

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

Huyen TM Nguyen, Quyen T Ngo, Luong X Vu, Vu M Tran, and Hien TT Nguyen. 2018. Vlsp shared task: Named entity recognition. *Journal of Computer Science and Cybernetics*.

Thai Binh Nguyen, Quang Minh Nguyen, Thi Thu Hien Nguyen, Quoc Truong Do, and Chi Mai Luong. 2020. Improving vietnamese named entity recognition from

speech using word capitalization and punctuation recovery models. *Interspeech*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Stefan Rüd, Massimiliano Ciaramita, Jens Müller, and Hinrich Schütze. 2011. Piggyback: using search engines for robust cross-domain named entity recognition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*.

David Rybach, Stefan Hahn, Patrick Lehnen, David Nolden, Martin Sundermeyer, Zoltán Tüske, Simon Wiesler, Ralf Schlüter, and Hermann Ney. 2011. RASR - the RWTH Aachen University open source speech recognition toolkit. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan S Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937, Toronto, Canada. Association for Computational Linguistics.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Cong Dao Tran, Nhut Huy Pham, Anh-Tuan Nguyen, Truong Son Hy, and Tu Vu. 2023. Videberta: A powerful pre-trained language model for vietnamese. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc 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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *Interspeech*.

Albert Zeyer, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Annual Meeting of the Assoc. for Computational Linguistics*.