VietMed-NER: Medical Spoken Named Entity Recognition

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

In this work, we present *VietMed-NER* - the first spoken NER dataset in the medical domain. To the best of our knowledge, *VietMed-NER* is also the first dataset for NER in Vietnamese spoken text. Our real-world dataset contains 16 entity types, which is by far the largest number compared to existing Vietnamese NER datasets for written text. Second, we propose an annotation technique, called Recursive Greedy Mapping, intending to enhance annotation speed and quality. Third, we propose a metric called Named-Entity-Error-Rate (NEER) to analyze ASR errors in spoken NER. Finally, we present baseline results using various state-of-the-art pre-trained models. We found that pre-trained multilingual models XLM-R outperformed all monolingual models on both reference text and ASR output. 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) from text and categorizing them into types like person, location, organization, etc. Initially studied in written language, recent attention has turned to studying spoken NER (Cohn et al., 2019; Shon et al., 2022). It aims to extract semantic information from speech with many applications, such as addressing privacy concerns in recordings (e.g., muting specific words like person names) (Cohn et al., 2019), spoken NER has limited literature compared to NER on text data (Yadav et al., 2020).

Spoken NER 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 in 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 spoken NER. The only related work we found, Cohn et al. (2019), published a 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 spoken NER dataset 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 publicly-available medical spoken NER dataset. To our best knowledge, *VietMed-NER* is also the first

NER dataset for Vietnamese spoken text.

- We propose an annotation technique, called Recursive Greedy Mapping, intending to improve annotation speed and quality.
- We propose NEER metric to analyze ASR errors in spoken NER.
- 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 the *VietMed* dataset (Le-Duc et al., 2023), a real-world medical ASR dataset in Vietnamese, for annotating NEs. This choice is driven by the fact that *VietMed* currently stands as the world's largest and most generalizable publicly-available medical ASR dataset.

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, disfluencies and required 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. This inconsistency includes: i) Some entities tagged in a sentence are not also tagged in other sentences, and ii) Full NEs are inconsistently tagged as multiple sub-NEs.

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 acceptable correctness. Also, the idea of training a seed model using a gazetteer list to help pre-tagging the rest of the dataset (Kozareva, 2006) requires an initial training time and the performance might not be reliable since the model is only trained on a small amount of initial data.

[1]https://github.com/leduckhai/MultiMed

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

To tackle these problems, we propose another annotation method, called Recursive Greedy Mapping. The method is described below:

1. Annotate and categorize some initial entities, then add them to a gazetteer list.
2. Sort entities by character length from highest to lowest, to distinguish between sub-NEs and full NEs, ensuring full NEs are mapped before sub-NEs. Time complexity = $O(n \cdot \log(n))$. For example, "tooth pain" should be mapped before "pain".
3. Automatically map entities from the gazetteer list to the transcript. Time complexity = $O(m \cdot n)$. Pseudo code:

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

4. Annotators review each sentence to correct mislabeled entities, often due to the rule-based algorithm's lack of contextual consideration.
5. Annotators add new NEs not in the gazetteer list during manual annotation. Steps 2 and 3 generate pre-tagged labels in the next sentences. Annotators repeat Steps 4 and 5 until the entire corpus is annotated.

In short, our annotation approach can reduce annotation time and eliminates the common inconsistency observed in the mentioned corpora.

2.3 Data Quality Control

We created initial annotation guidelines (see Appendix) and began annotating the corpus. Two developers, one with a medical background, independently annotated the corpus. Then, we held a discussion session to resolve conflicts, address complex cases, and refine the guidelines. Two other developers perform quality control using the guidelines and the annotated corpus. We consistently revisited each sentence in the entire corpus multiple times. The current version is v1.0. As *VietMed-NER* continues to expand, we will update publicly reported errors and baseline results as future versions.

2.4 Data Statistics

Table 1 shows the statistics of our dataset. Our *VietMed-NER* contains 16 entity types across 9000

Entity Type	Definition	Train	Dev	Test	All	Uni.
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
SYMP.	Symptoms and diseases	1484	1649	1467	4600	555
CHEM.	Bio-chemical substances and drugs	419	700	709	1828	248
FnB	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, neurosurgery	225	20	276	521	59
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.

sentences, split into train-dev-test as 5-5-6 hours. To the best of our knowledge, compared to all other public Vietnamese NER datasets, ours has by far the largest number of entity types, while the 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). Our annotation is based on the original train-dev-test split of the *VietMed* dataset, hence leading to some imbalance in entity types.

3 Experimental Setups

We employ the 2-stage pipeline for spoken NER: An ASR model transcribes audio into text and then the transcribed text is fed into a NER model.

3.1 Named-Entity-Error-Rate (NEER)

The commonly used WER is not an effective metric to show ASR errors causing NER errors, we therefore propose a NEER metric specifically for spoken NER, which is modified from vanilla KER in ASR (Park et al., 2008):

$$NEER = \frac{S + D}{N} \times 100 \qquad (1)$$

where reference data contains only entities (all non-entities are removed), N is the length of entities, S and D are the number of substitutions and deletions between ASR hypothesis and reference data.

ASR Model	WER		NEER	
	dev	test	dev	test
w2v2-Viet	25.9	29.0	25.4	28.3
XLSR-53-Viet	25.7	28.8	25.2	27.9

Table 2: WER and NEER of 2 pre-trained ASR models: w2v2-Viet model was pre-trained on 1204h of Vietnamese data. XLSR-53-Viet model was pre-trained on 1204h of Vietnamese with XLSR-53 (Conneau et al., 2021) as initialization. Both of them have the same number of parameters (118M) and were fine-tuned on 5h of dataset.

3.2 F1 Score

For the evaluation of NER on reference text, we calculate F1 scores using the seqeval[3] framework commonly used as a default evaluation framework by HuggingFace.

For evaluation of NER on ASR output, we employed the F1 score calculation by Shon et al. (2023) by using the SLUE toolkit[4]. This F1 score evaluates an unordered list of named entity phrase and tag pairs predicted for each sentence.

3.3 ASR Models

We employed 2 best baseline models fine-tuned for ASR task on *VietMed* published by Le-Duc et al.

[3] https://github.com/chakki-works/seqeval
[4] https://github.com/asappresearch/slue-toolkit

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.

NER	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
ViDeBERTa_base	36.37	23.12	28.27
XLM-R_base	65.66	66.20	65.93
XLM-R_large	69.59	64.43	66.91

Table 4: NER results (in percent) on reference text of test set using various pre-trained language models. Metrics shown are Precision, Recall, and overall micro F1 score. Results by entity types are shown in Appendix.

NER	ASR	Prec.	Rec.	F1
PhoBERT_large	w2v2-Viet	41.07	61.63	49.29
	XLSR-53-Viet	40.47	60.09	48.36
PhoBERT_base-v2	w2v2-Viet	44.22	61.06	51.29
	XLSR-53-Viet	43.89	59.88	50.66
ViDeBERTa_base	w2v2-Viet	23.74	48.42	31.86
	XLSR-53-Viet	23.81	47.80	31.78
XLM-R_base	w2v2-Viet	44.95	59.88	51.35
	XLSR-53-Viet	44.43	58.15	50.37
XLM-R_large	w2v2-Viet	45.29	60.37	51.75
	XLSR-53-Viet	45.39	59.33	51.43

Table 5: NER results (in percent) on ASR output of test set using various pre-trained language models. Metrics shown are Precision, Recall, and overall micro F1 score. Results by entity types are shown in Appendix.

(2023). Table 2 shows WERs and NEERs of the 2 models. All NEERs are acceptable compared to WERs, showing that these 2 ASR models recognize entities as correctly as non-entities, making them suitably effective for recognizing spoken NEs.

3.4 NER Models

Table 3 shows the statistics of various pre-trained monolingual and multilingual models to fine-tune on our dataset. Monolingual: PhoBERT_base, PhoBERT_large, PhoBERT_base-v2 (Nguyen and Nguyen, 2020); ViDeBERTa_base (Tran et al., 2023). Multilingual: XML-R_base, XML-R_large (Conneau et al., 2020). To our best knowledge, these are the best pre-trained models that achieved state-of-the-art results on various downstream tasks in the Vietnamese language, including NER.

We used HuggingFace (Wolf et al., 2019) for fine-tuning pre-trained models on the NER task. Vietnamese input sentences 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 the default hyperparameters from HuggingFace. Details of the hyperparameters are shown in the Appendix.

4 Experimental Results

Table 4 and 5 show results of NER using various pre-trained models. The pre-trained monolingual model PhoBERT_base-v2 outperformed other monolingual models, at 61.99% of F1 score on reference text, and 51.29% on ASR output. Even though it has fewer parameters than PhoBERT_large, it achieved higher performance probably due to the larger amount of pre-training data. All pre-trained multilingual models outperformed monolingual models. For example, XLM-R_large achieved an F1 score of 66.91% on reference text and 51.75% on ASR output. This gap can be explained by the larger amount of pre-training data (2.5TB of multilingual data for XLM-R compared to only 140GB of monolingual data for PhoBERT_base-v2).

5 Conclusion

In this work, we present *VietMed-NER* - the first spoken NER dataset in the medical domain. To our best knowledge, *VietMed-NER* is also the first NER dataset for Vietnamese spoken text. Our dataset contains 16 entity types, including both conventional and newly defined entity types for real-world medical conversations. Furthermore, we propose an annotation technique, called Recursive Greedy Mapping, to accelerate the annotation speed and remove inconsistency during annotating. Besides, we propose a modified NEER metric to show the effective recognition of spoken NEs by baseline ASR models. Finally, we found that pre-trained multilingual models XLM-R_base and XLM-R_large outperformed all monolingual models on both reference text and ASR output.

6 Limitations

Number of entity types: Most NER datasets include traditional entity types like NAME and ORGANIZATION. However, in our dataset, we did not include these 2 entity types. For NAME, *VietMed* is a real-world medical ASR dataset, in which the authors had to remove all parts that might reveal speaker identities, including their names, to preserve privacy. For ORGANIZATION, we found less than 100 entities in the transcript, which led to the F1 score of 0 for most pre-trained models, except for the F1 score of 21% for XLM-R_large.

Our annotation approach: Our annotation approach using the Recursive Greedy Mapping technique has some advantages over the fully manual approach. First, it allows annotators to not spend extra time tagging the entities that have been tagged in previous sentences. Second, it prevents that annotators miss entities that have been tagged in previous sentences, improving the consistency of the entire dataset. During our work, we experienced a faster annotation by using our approach compared to fully manual annotation. However, in the scope of this paper, we have not done extensive experiments to give a quantitative number of how much time has been saved and the method's impact on annotation quality.

Mismatch between evalation metrics of NER: HuggingFace is widely used as a framework to build an NER system due to its easy accessibility to public pre-trained models. However, currently, HuggingFace only supports the evaluation of written-text NER with its default evaluation toolkit, seqeval. In contrast, SLUE is widely used to evaluate spoken NER. This mismatch of frameworks leads to the fact that F1 scores from spoken NER might not possibly show the degradation of ASR errors on written-text NER. We focus on showing a comparison of pre-trained models for written-text and spoken text instead of demonstrating the degradation caused by ASR errors.

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, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

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.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

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*.

Youngja Park, Siddharth Patwardhan, Karthik Visweswariah, and Stephen C Gates. 2008. An empirical analysis of word error rate and keyword error rate. In *Interspeech*, volume 2008, pages 2070–2073.

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*.

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*.

A Annotation Guidelines

This section describes annotation guidelines for annotators to follow in an attempt to have a unified and consistent gold-standard NER transcript.

General rules:

- If 2 or more entities overlap, label the resulting entity as the longest, including overlapping component entities. In other words, a full NE might contain 2 or more sub-NEs. A full NE should be tagged instead of multiple sub-NEs. For example: "bác sĩ xương khớp" (orthopedic doctor) should be tagged as a whole instead of 2 distinct NEs "bác sĩ " (doctor) and "xương khớp" (orthopedic).

- Do not assign spaces at the beginning and at the end of entities.

- All words in the ASR transcript are lowercase, without punctuations and special characters. Treat every word as lowercase or uppercase, with or without punctuations and special characters based on the context of each utterance.

- Each utterance should be treated as an independent utterance. The additional context given by other utterances should not influence the annotation of each utterance.

AGE:
This entity type describes the age of a person.

- Label the word "tuổi" (age) if applicable. For example: "tuổi trưởng thành" (mature age), "hai bảy tuổi" (twenty-seven years old).

- List a range of ages if applicable For example: "hai mươi đến ba lăm tuổi" (twenty to thirty-five years old), "dưới sáu tháng tuổi" (under six months old).

- Include adjectives and nouns that might describe how old a person is but don't explicitly describe gender or gender is neutral. For example: "chưa trưởng thành" (immature), "người già" (old person), "cụ" (sir, old).

GEND.:
This entity type describes the gender of a person.

- Include typical entities that are widely understood to describe the gender of a person. For example: "nam" (male), "đàn ông" (gentleman), "phụ nữ" (woman).

- Include the titles and pronouns that explicitly describe a gender instead of age. For example: "ông" (grandfather), "bà" (grandmother), "cô" (aunt), "chú" (uncle).

JOB:
This entity type describes the job of a person.

- Include all jobs that might be both in medical fields and non-medical fields. For example: "khán thính giả" (audience), "kẻ bắt cóc" (kidnapper), "bệnh nhân" (patient), "người dân" (citizen), "chuyên gia" (expert).

- Include academic titles and degrees. For example: "thạc sĩ" (master degree holder), "tiến sĩ" (doctorate), "trưởng khoa" (dean), "chủ tịch" (president).

- Include a cluster of words that might describe the specializations of doctors. For example: "bác sĩ chuyên về rối loạn vận động" (doctor who specializes in movement disorders) instead of two distinct entities "bác sĩ" (doctor) and "rối loạn vận động" (movement disorders), "bác sĩ về parkinson" (parkinson's doctor) instead of two distinct entities "bác sĩ" (doctor) and "parkinson", "bác sĩ chuyên khoa tim mạch" (cardiovascular specialist) instead of two distinct entities "bác sĩ" (doctor) and "chuyên khoa tim mạch" (cardiovascular).

LOC.:
This entity type describes a location.

- Include continents, countries, regions, cities, and geographical administrative units. For example: "châu âu" (europe), "hoa kỳ" (usa), "tây tạng" (tibet), "thành phố hồ chí minh" (ho chi minh city), "tỉnh vĩnh long" (vinh long province).

- Label words that mean geographical administrative units if applicable. For example: "huyện" (rural district), "quận" (urban district), "đường phố" (street), "thành phố" (city).

- Include words that might describe public and private sites. For example: "tại nhà" (at home), "đồng ruộng" (farm), "tiệm thuốc" (drugstore), "nhà máy" (factory), "cửa hiệu quần áo" (clothing store), "toilet" (toilet).

- Include words that might describe ambient environments. For example: "tại khu phố" (in the neighborhood), "tại địa phương" (in local area), "nước ngoài" (in foreign countries), "địa bàn" (area), "ngoài trời" (outside).

- Include words that might describe medical facilities. For example: "chuyên khoa tiêu hóa" (gastrointestinal room) , "icu" (intensive care unit), "trạm xá" (clinics), "phòng thí nghiệm" (laboratory).

- Each level of the administrative unit is a separate entity.

- Do not assign nationality as an entity.

- Locations might be misrecognized as organizations. Do not label places that are not clearly identified or controversial.

SYMP.:
This entity type describes a symptom or disease.

- Include the complements of the disease. For example: "biến chứng" (side-effect), "chấn thương" (damaged), "bẩm sinh" (congenital), "di chứng" (sequelae), "bị tổn thương" (damaged), "tái phát" (relapse), "dương tính" (positive), "bệnh lý mãn tính" (chronic disease), "hội chứng" (syndrome).

- Include a cluster of words that might describe the severity of a disease. For example: "phỏng cấp độ ba" (third-degree burn), "sức đề kháng kém" (poor immune system).

- Mental state might also describe mental diseases or their symptoms. For example: "tự ti" (self-deprecation), "tình trạng lo âu" (state of anxiety), "mệt mỏi về tinh thần" (mental fatigue).

- Skin conditions might describe dermatosis or its symptoms. For example: "nám" (melasma), "da đổ dầu" (oily skin), "da khô" (dry skin), "sạm da" (dark skin).

- Genital conditions might describe genital diseases or their symptoms. For example: "có kinh" (menstruation), "có thai" (pregnant), "dậy thì sớm" (early puberty).

- Healthy conditions might help doctors diagnose. For example: "kinh nguyệt đều" (regular menstruation).

- Words describing physical status might also speak of symptoms or diseases. For example: "buồn ngủ" (sleepy), "rụng tóc" (hair loss), "còi cọc" (stunted).

- Words describing children's activities might also speak of pediatric symptoms or diseases. For example: "quấy khóc" (fussy), "không thể giao tiếp" (unable to speak), "chậm đi" (delay walking).

- Medical techniques or devices might make symptoms and diseases happen. For example: "phẫu thuật thẩm mỹ" (cosmetic surgery).

CHEM.:
This entity type describes a bio-chemical substance or medicament.

- Extraction of human or animal bodies to serve medical treatment might be referred to as a biochemical substance. For example: "vắcxin" (vaccine), "huyết thanh" (blood serum)

- Cosmetics might be referred to as chemical substances. For example: "kem chống nắng" (sunscreen), "kem dưỡng ẩm" (moisturizer).

- Food or drink serving medical treatment purposes or as a part of a chemical compound might be referred to as chemical substances. For example: "nấm đông trùng hạ thảo" (cordyceps), "nhân sâm" (ginseng), "nhung hươu" (deer antler).

- Substances extracted from cells or bodies not serving medical purposes might be referred to as bio-chemical substances. For example: "dịch tiêu hóa" (digestive fluids), "chất nội sinh" (endogenous substances), "mồ hôi" (sweat), "bã nhờn" (sebum).

- Air might be referred to as chemical substances. For example: "dưỡng khí" (breath air), "oxy" (oxygen).

FnB:
This entity type describes food and beverage.

- Include food and drink that might serve nutrient purposes. For example: "sữa" (milk), "ngũ cốc" (cereal).

- Include food and drink that might be harmful to health. For example: "thuốc lá" (cigarette), "rượu bia" (alcohol).

- Include words that generally describe food and beverage. For example: "thực phẩm" (aliment), "thức ăn" (food).

ANAT.:

This entity type describes an anatomical feature, e.g. human organs, biological cells, etc. Annotators should follow general rules.

PC:

This entity type describes a personal care procedure, e.g. hygiene routines, skin care, daily habits, etc.

- Activities serving the improvement of physical, aesthetic and mental health instead of medical treatment purposes might be referred to as personal care procedures. For example: "ăn kiêng" (diet), "chăm sóc da" (skin care), "chăm sóc răng" (dental care).

- Methods serving self-improvement of speech ability in speech-language pathology might be referred to as personal care. For example: "tương tác ngôn ngữ" (language interaction), "huấn luyện ngôn ngữ" (language training).

DX:

This entity type describes a diagnostic procedure, e.g. lab tests, imaging, blood measurement, etc.

- General words describing diagnostic procedures without explicitly mentioning surgery might be referred to as diagnostic produces. For example: "chẩn đoán" (diagnosis), "xét nghiệm" (test).

- Imaging methods might be referred to as diagnostic procedures instead of medical devices or techniques. For example: "mri" (magnetic resonance imaging), "ct" (computed tomography).

TX:

This entity type describes a non-surgical treatment method for diseases, e.g. physical rehabilitation, injection, psychology, etc.

- Words describing methods of using biochemical substances as non-surgical treatment methods might be referred to as treatment methods. For example: "liệu pháp hoocmon" (hormone therapy), "điều trị hoocmon" (hormone treatment), "điều trị tế bào gốc" (stem cell treatment).

- Words describing methods of using invasive techniques as treatment methods might be referred to as treatment methods. For example: "hóa trị" (chemotherapy), "xạ trị" (radiotherapy).

- Words describing methods to improve skin conditions for treatment purposes rather than aesthetics might be referred to as treatment methods. For example: "phục hồi da" (skin recovery), "ức chế sự xuất sắc tố" (inhibit pigmentation).

SX:

This entity type describes a surgical treatment method for diseases, e.g. implants, neurosurgery, invasion, etc.

- Include pre-surgery procedures that might be integral parts of surgeries. For example: "gây mê" (anesthesia), "gây tê" (anesthetize).

- Include intervention procedures that might be integral parts of dental care. For example: "nhổ răng" (tooth extraction), "implant" (dental implant).

- Include intervention procedures that might be integral parts of pregnancy or genitals. For example: "sinh mổ" (caesarean), "cấy tránh thai" (contraceptive implant).

- Include intervention procedures on arteries even though they might be not integral parts of surgery. For example: "truyền máu" (blood transfusion), "truyền nước biển" (seawater infusion).

- Include neurosurgical procedures that work with brain waves even though they might be minimally invasive. For example: "kích thích não sâu" (dbs or deep brain stimulation).

TECH.:

This entity type describes a medical device, instrument, bio-material and technique.

- Medical devices and techniques might be confusing. Annotators are strongly recommended to fully annotate CHEM., FnB, ANAT., PC, DX, TX, and SX before engaging TECH.

CAL.:

This entity type describes a medical calibration, e.g. number of doses, calories, length, volume, etc.

- Include a cluster of words that both describe the quantity and its unit. Measurements including length, distance, area, weight, heat, velocity, temperature, etc., should be explicitly tagged. For example: "năm milimet" (five millimeters) instead of "năm" (five) or "milimet" (millimeter).
- Complements to the actual quantity describing its approximation should be included. For example: "khoảng mười lăm phần trăm" (about fifteen percent) instead of "mười lăm phần trăm" (fifteen percent).
- Include words that generally describe the quantity. For example: "gần đủ" (close enough), "cao" (high), "rất là lớn" (very large).
- Include words that describe trends of quantity. For example: "giảm được ít nhất" (reduce at least), "mức độ gia tăng" (level increases).

TRAN.:

This entity type describes means of transportation or vehicles.

TIME:

This entity type describes the date and time.

- Include words describing day, week, month, certain named period, season, year, etc.
- Include words describing a time frame. For example: "bây giờ" (now), "về lâu về dài" (in the long run).
- Include words describing the approximate time. For example: "nhanh nhất có thể" (as fast as possible), "càng sớm" (as soon as possible), "từ từ" (gradually).
- Include words describing repetitions. For example: "định kỳ" (periodically).
- Include a cluster of words that both describe time and its complements. For example: "từ tháng ba trở đi" (from march onwards) instead of 3 distinct entities "từ" (from), "tháng ba" (march), and "trở đi" (onwards).

B Discussion about NEER

B.1 Motivation of NEER

ASR system performance is typically assessed using WER, which represents the ratio of word insertion, substitution, and deletion errors in a transcript to the total number of spoken words. However, various spoken language understanding tasks, such as spoken NER, depend on identifying keywords in transcripts. Moreover, it's essential to recognize that in medical ASR, medical terms carry much higher significance in doctor-patient conversations and should not be treated equally to regular words. KER is often used to evaluate on keywords but is not a directly comparable metric with WER.

The purpose to introduce NEER aims to bridge the gap between WER and KER. However, it is not intended to replace WER or KER as a standard metric for evaluating domain-specific ASR performance. Instead, NEER serves as a complementary metric, facilitating a more in-depth analysis of ASR errors in specific domains, such as the medical field.

B.2 Definition of WER

WER is calculated based on the Levenshtein distance (Levenshtein et al., 1966), which represents the smallest count of individual edits (insertions, deletions, or substitutions) needed to transform one word into another.

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (2)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, and N is the number of words in the reference data $(N = S + D + C)$.

In other words, S is the number of replaced words. D is the number of missed words that are not in ASR hypothesis but are in reference data. I is the number of added words that are in ASR hypothesis but are not in reference data. The alignment between ASR hypothesis and reference data goes from left to right.

B.3 Definition of KER

Like WER, KER is computed using the Levenshtein distance. Each ASR hypothesis is aligned with its corresponding reference data and KER is calculated based on the keyword set.

$$WER = \frac{F+M}{N} \quad (3)$$

where N is the number of keywords in the reference data, F is the number of falsely recognized keywords, M is the number of missed keywords.

The ASR hypothesis often exceeds the length of all keywords in the reference data, and the insertion errors caused by non-keywords may lead to a skewed result in KER. Therefore, no insertion errors are considered while calculating KER.

B.4 Definition of NEER

In KER metric, N is the number of keywords in the reference data. KER could be characterized as the average number of errors per keyword. Nevertheless, the length of keywords may range from 1 to L (where L equals 5 in certain instances such as NER), making the average number of errors per keyword obscure.

In NEER metric, we want to evaluate on keyword-only like KER metric, while also analyzing errors per word like WER metric. Therefore, we change N into the length of keywords (entities), which characterizes the average number of errors per word of keywords.

B.5 Open questions on NEER

We still leave some questions open for future work. First, the analysis of how each type of word error (substitutions, insertions, deletions) influences NER on top of ASR has not been conducted yet. Second, the empirical relationship between WER, KER, NEER, and F1 score - meaning how KER, NEER, and F1 score are affected by a varying range of WERs — has not been analyzed either.

T

C Details of Training Hyperparameters

We used HuggingFace (Wolf et al., 2019) for fine-tuning pre-trained models for the NER task. We only employed word-level settings for training NER models. All our NER experiments were done using the default hyperparameters by HuggingFace.

The default hyperparameters are as follows: Learning rate of 2e-5, linear learning rate scheduler, training batch size of 64, 50 training epoch, weight decay of 0.01, AdamW optimizer (Loshchilov and Hutter, 2019), Beta1 of 0.9, Beta2 of 0.999, and epsilon of 1e-8.

D NER Results by Entity Types

Table 6 shows the results of NER on reference text by entity types using various pre-trained language models. Table 7 shows the results of NER on ASR output by entity types using various pre-trained language models and ASR models.

Entity Type	NER	Prec.	Rec.	F1
AGE	PhoBERT_base	40.84	69.21	51.37
	PhoBERT_large	27.00	63.00	38.00
	PhoBERT_base-v2	55.79	53.38	54.56
	ViDeBERTa_base	1.45	11.84	2.58
	XLM-R_base	67.04	63.76	65.36
	XLM-R_large	63.02	79.84	70.44
TIME	PhoBERT_base	73.46	65.77	69.41
	PhoBERT_large	70.00	69.00	69.00
	PhoBERT_base-v2	71.96	67.80	69.82
	ViDeBERTa_base	41.53	32.25	36.30
	XLM-R_base	72.71	70.80	71.75
	XLM-R_large	74.21	67.53	70.71
DX	PhoBERT_base	80.27	74.77	77.42
	PhoBERT_large	81.00	81.00	81.00
	PhoBERT_base-v2	81.61	79.22	80.40
	ViDeBERTa_base	39.80	42.35	41.03
	XLM-R_base	83.28	81.64	82.45
	XLM-R_large	84.28	82.35	83.31
SYMP.	PhoBERT_base	58.01	63.94	60.83
	PhoBERT_large	54.00	65.00	59.00
	PhoBERT_base-v2	59.51	63.26	61.33
	ViDeBERTa_base	25.02	30.16	27.35
	XLM-R_base	65.78	69.08	67.39
	XLM-R_large	64.96	66.55	65.75
CHEM.	PhoBERT_base	69.82	79.07	74.16
	PhoBERT_large	70.00	79.00	74.00
	PhoBERT_base-v2	70.24	86.01	77.33
	ViDeBERTa_base	19.46	19.94	19.70
	XLM-R_base	70.66	80.29	75.17
	XLM-R_large	74.33	82.60	78.25
F&B	PhoBERT_base	66.28	55.27	60.28
	PhoBERT_large	68.00	54.00	60.00
	PhoBERT_base-v2	73.56	59.81	65.98
	ViDeBERTa_base	0.77	8.33	1.40
	XLM-R_base	76.25	60.67	67.57
	XLM-R_large	73.95	65.65	69.55
GEND.	PhoBERT_base	43.37	51.49	47.08
	PhoBERT_large	44.00	54.00	48.00
	PhoBERT_base-v2	37.28	56.83	45.02
	ViDeBERTa_base	0.00	0.00	0.00
	XLM-R_base	63.98	59.60	61.71
	XLM-R_large	61.83	79.13	69.42
LOC.	PhoBERT_base	60.49	54.37	57.27
	PhoBERT_large	62.00	65.00	64.00
	PhoBERT_base-v2	68.09	65.50	66.77
	ViDeBERTa_base	5.78	34.55	9.90
	XLM-R_base	69.60	61.07	65.06
	XLM-R_large	75.08	65.69	70.07
	PhoBERT_base	28.66	31.60	30.06
	PhoBERT_large	18.00	29.00	22.00

TECH.	PhoBERT_base-v2	32.76	34.96	33.82
	ViDeBERTa_base	2.52	14.81	4.31
	XLM-R_base	23.62	35.63	28.41
	XLM-R_large	23.46	36.34	28.52
JOB	PhoBERT_base	90.62	86.49	88.50
	PhoBERT_large	93.00	83.00	88.00
	PhoBERT_base-v2	91.68	85.62	88.55
	ViDeBERTa_base	79.82	85.58	82.60
	XLM-R_base	94.34	88.54	91.35
	XLM-R_large	92.92	92.59	92.76
ANAT.	PhoBERT_base	56.27	54.08	55.15
	PhoBERT_large	59.00	55.00	57.00
	PhoBERT_base-v2	60.56	57.41	58.94
	ViDeBERTa_base	21.43	26.60	23.74
	XLM-R_base	67.70	57.83	62.38
	XLM-R_large	72.14	57.53	64.01
PC	PhoBERT_base	70.53	93.06	80.24
	PhoBERT_large	71.00	72.00	71.00
	PhoBERT_base-v2	69.47	97.06	80.98
	ViDeBERTa_base	9.47	22.50	13.33
	XLM-R_base	71.58	94.44	81.44
	XLM-R_large	70.53	93.06	80.24
SX	PhoBERT_base	42.39	52.94	47.08
	PhoBERT_large	33.00	39.00	36.00
	PhoBERT_base-v2	57.61	56.38	56.99
	ViDeBERTa_base	22.46	63.27	33.16
	XLM-R_base	63.04	52.25	57.14
	XLM-R_large	52.90	44.24	48.18
TRAN.	PhoBERT_base	0.00	0.00	0.00
	PhoBERT_large	89.00	96.00	92.00
	PhoBERT_base-v2	3.70	25.00	6.45
	ViDeBERTa_base	0.00	0.00	0.00
	XLM-R_base	85.19	92.00	88.46
	XLM-R_large	77.78	87.50	82.35
TX	PhoBERT_base	82.33	80.59	81.45
	PhoBERT_large	82.00	73.00	77.00
	PhoBERT_base-v2	81.90	80.85	81.37
	ViDeBERTa_base	81.90	74.51	78.03
	XLM-R_base	85.34	79.20	82.16
	XLM-R_large	85.78	80.89	83.26
CAL.	PhoBERT_base	52.92	49.64	51.22
	PhoBERT_large	46.00	57.00	51.00
	PhoBERT_base-v2	45.53	44.66	45.09
	ViDeBERTa_base	1.95	7.04	3.05
	XLM-R_base	51.36	53.66	52.49
	XLM-R_large	50.58	51.18	50.88
	PhoBERT_base	58.07	62.00	59.97
	PhoBERT_large	56.00	63.00	59.00
	PhoBERT_base-v2	60.82	63.21	61.99
	ViDeBERTa_base	23.12	36.37	28.27
	XLM-R_base	66.20	65.66	65.93

Micro average

	XLM-R_large	66.59	67.54	67.06
Macro average	PhoBERT_base	48.22	50.65	49.03
	PhoBERT_large	51.00	56.00	53.00
	PhoBERT_base-v2	50.59	53.36	51.23
	ViDeBERTa_base	18.60	24.93	19.81
	XLM-R_base	58.93	58.55	58.42
	XLM-R_large	60.81	64.83	62.09

Table 6: NER results by entity types (in percent) on reference text of test set using various pre-trained language models. Metrics shown are Precision, Recall, and overall micro/macro F1 score.

Entity Type	**NER**	**ASR**	**Prec.**	**Rec.**	**F1**
AGE	PhoBERT_large	w2v2-Viet	26.67	62.56	37.40
		XLSR-53-Viet	24.10	57.49	33.96
	PhoBERT_base-v2	w2v2-Viet	39.03	55.41	45.80
		XLSR-53-Viet	36.46	54.55	43.70
	ViDeBERTa_base	w2v2-Viet	1.65	15.53	2.98
		XLSR-53-Viet	1.34	12.04	2.41
	XLM-R_base	w2v2-Viet	43.46	58.29	49.79
		XLSR-53-Viet	40.06	57.63	47.27
	XLM-R_large	w2v2-Viet	38.93	65.51	48.84
		XLSR-53-Viet	39.55	66.21	49.52
TIME	PhoBERT_large	w2v2-Viet	53.23	59.18	56.05
		XLSR-53-Viet	52.96	56.38	54.62
	PhoBERT_base-v2	w2v2-Viet	54.85	59.07	56.88
		XLSR-53-Viet	55.46	55.68	55.57
	ViDeBERTa_base	w2v2-Viet	36.59	39.15	37.83
		XLSR-53-Viet	37.33	37.74	37.53
	XLM-R_base	w2v2-Viet	50.94	57.23	53.90
		XLSR-53-Viet	51.28	53.22	52.23
	XLM-R_large	w2v2-Viet	54.04	56.32	55.16
		XLSR-53-Viet	54.25	51.47	52.82
DX	PhoBERT_large	w2v2-Viet	45.03	57.96	50.68
		XLSR-53-Viet	45.23	57.47	50.62
	PhoBERT_base-v2	w2v2-Viet	47.06	58.15	52.02
		XLSR-53-Viet	47.06	58.15	52.02
	ViDeBERTa_base	w2v2-Viet	24.95	38.20	30.18
		XLSR-53-Viet	24.95	37.50	29.96
	XLM-R_base	w2v2-Viet	46.65	58.08	51.74
		XLSR-53-Viet	44.62	53.40	48.62
	XLM-R_large	w2v2-Viet	48.28	56.80	52.19
		XLSR-53-Viet	45.84	53.18	49.24
SYMP.	PhoBERT_large	w2v2-Viet	44.91	68.38	54.21
		XLSR-53-Viet	44.45	68.96	54.06
	PhoBERT_base-v2	w2v2-Viet	50.59	64.68	56.77
		XLSR-53-Viet	49.74	65.47	56.53
	ViDeBERTa_base	w2v2-Viet	37.72	57.04	45.41
		XLSR-53-Viet	36.69	57.98	44.94
	XLM-R_base	w2v2-Viet	51.74	62.99	56.81
		XLSR-53-Viet	51.62	63.03	56.75

	XLM-R_large	w2v2-Viet	50.50	63.80	56.38
		XLSR-53-Viet	52.04	64.40	57.56
CHEM.	PhoBERT_large	w2v2-Viet	43.21	63.80	51.52
		XLSR-53-Viet	44.03	64.57	52.36
	PhoBERT_base-v2	w2v2-Viet	42.94	66.81	52.28
		XLSR-53-Viet	43.30	67.28	52.69
	ViDeBERTa_base	w2v2-Viet	19.23	33.81	24.52
		XLSR-53-Viet	20.97	35.49	26.36
	XLM-R_base	w2v2-Viet	45.67	57.39	50.86
		XLSR-53-Viet	44.94	58.69	50.90
	XLM-R_large	w2v2-Viet	47.58	57.62	52.12
		XLSR-53-Viet	48.40	58.80	53.10
F&B	PhoBERT_large	w2v2-Viet	47.61	44.75	46.13
		XLSR-53-Viet	49.47	44.82	47.03
	PhoBERT_base-v2	w2v2-Viet	46.54	45.57	46.05
		XLSR-53-Viet	48.67	44.53	46.51
	ViDeBERTa_base	w2v2-Viet	3.99	50.00	7.39
		XLSR-53-Viet	3.99	51.72	7.41
	XLM-R_base	w2v2-Viet	46.81	47.18	47.00
		XLSR-53-Viet	50.00	46.53	48.21
	XLM-R_large	w2v2-Viet	43.88	51.08	47.21
		XLSR-53-Viet	48.94	52.87	50.83
GEND.	PhoBERT_large	w2v2-Viet	29.57	69.86	41.55
		XLSR-53-Viet	29.57	69.86	41.55
	PhoBERT_base-v2	w2v2-Viet	21.88	69.91	33.33
		XLSR-53-Viet	20.58	68.27	31.63
	ViDeBERTa_base	w2v2-Viet	0.00	0.00	0.00
		XLSR-53-Viet	0.00	0.00	0.00
	XLM-R_base	w2v2-Viet	41.30	72.52	52.63
		XLSR-53-Viet	43.48	71.60	54.10
	XLM-R_large	w2v2-Viet	34.06	68.51	45.50
		XLSR-53-Viet	32.17	70.03	44.09
LOC.	PhoBERT_large	w2v2-Viet	43.82	59.13	50.34
		XLSR-53-Viet	41.76	54.20	47.18
	PhoBERT_base-v2	w2v2-Viet	43.24	54.14	48.08
		XLSR-53-Viet	43.68	56.36	49.21
	ViDeBERTa_base	w2v2-Viet	4.71	52.46	8.64
		XLSR-53-Viet	4.85	46.48	8.79
	XLM-R_base	w2v2-Viet	43.97	57.61	49.87
		XLSR-53-Viet	40.88	54.51	46.72
	XLM-R_large	w2v2-Viet	52.65	60.88	56.47
		XLSR-53-Viet	51.32	61.01	55.75
TECH.	PhoBERT_large	w2v2-Viet	20.09	60.38	30.15
		XLSR-53-Viet	19.39	56.42	28.86
	PhoBERT_base-v2	w2v2-Viet	34.36	60.54	43.84
		XLSR-53-Viet	33.46	60.74	43.15
	ViDeBERTa_base	w2v2-Viet	3.71	48.33	6.89
		XLSR-53-Viet	3.26	43.59	6.07
	XLM-R_base	w2v2-Viet	21.75	58.22	31.67
		XLSR-53-Viet	21.75	56.57	31.42
	XLM-R_large	w2v2-Viet	23.93	60.81	34.34

		XLSR-53-Viet	24.63	58.96	34.75
JOB	PhoBERT_large	w2v2-Viet	68.33	65.14	66.70
		XLSR-53-Viet	66.75	63.06	64.85
	PhoBERT_base-v2	w2v2-Viet	65.70	65.65	65.68
		XLSR-53-Viet	65.62	63.78	64.68
	ViDeBERTa_base	w2v2-Viet	55.91	60.23	57.99
		XLSR-53-Viet	56.17	59.33	57.71
	XLM-R_base	w2v2-Viet	66.67	66.03	66.35
		XLSR-53-Viet	66.14	63.64	64.86
	XLM-R_large	w2v2-Viet	66.93	68.61	67.76
		XLSR-53-Viet	66.84	66.38	66.61
ANAT.	PhoBERT_large	w2v2-Viet	38.37	56.25	45.62
		XLSR-53-Viet	36.99	53.48	43.73
	PhoBERT_base-v2	w2v2-Viet	38.67	57.31	46.18
		XLSR-53-Viet	38.08	54.05	44.68
	ViDeBERTa_base	w2v2-Viet	22.36	40.09	28.71
		XLSR-53-Viet	23.05	39.54	29.13
	XLM-R_base	w2v2-Viet	40.95	56.50	47.48
		XLSR-53-Viet	39.46	53.89	45.56
	XLM-R_large	w2v2-Viet	42.74	53.91	47.68
		XLSR-53-Viet	41.45	52.88	46.47
PC	PhoBERT_large	w2v2-Viet	46.15	62.69	53.16
		XLSR-53-Viet	46.15	61.76	52.83
	PhoBERT_base-v2	w2v2-Viet	44.51	72.32	55.10
		XLSR-53-Viet	43.96	68.97	53.69
	ViDeBERTa_base	w2v2-Viet	14.84	65.85	24.22
		XLSR-53-Viet	15.38	65.12	24.89
	XLM-R_base	w2v2-Viet	43.41	73.15	54.48
		XLSR-53-Viet	42.86	70.27	53.24
	XLM-R_large	w2v2-Viet	43.96	71.43	54.42
		XLSR-53-Viet	43.41	66.95	52.67
SX	PhoBERT_large	w2v2-Viet	26.64	49.53	34.65
		XLSR-53-Viet	24.79	48.36	32.78
	PhoBERT_base-v2	w2v2-Viet	26.31	53.98	35.37
		XLSR-53-Viet	26.14	51.84	34.75
	ViDeBERTa_base	w2v2-Viet	14.50	44.33	21.86
		XLSR-53-Viet	15.01	40.64	21.92
	XLM-R_base	w2v2-Viet	35.58	46.99	40.50
		XLSR-53-Viet	33.39	43.61	37.82
	XLM-R_large	w2v2-Viet	30.86	45.98	36.93
		XLSR-53-Viet	29.68	46.19	36.14
TRAN.	PhoBERT_large	w2v2-Viet	0.00	0.00	0.00
		XLSR-53-Viet	100.00	40.00	57.14
	PhoBERT_base-v2	w2v2-Viet	0.00	0.00	0.00
		XLSR-53-Viet	50.00	100.00	66.67
	ViDeBERTa_base	w2v2-Viet	0.00	0.00	0.00
		XLSR-53-Viet	0.00	0.00	0.00
	XLM-R_base	w2v2-Viet	0.00	0.00	0.00
		XLSR-53-Viet	100.00	50.00	66.67
	XLM-R_large	w2v2-Viet	0.00	0.00	0.00
		XLSR-53-Viet	100.00	50.00	66.67

TX	PhoBERT_large	w2v2-Viet	58.30	61.76	59.98
		XLSR-53-Viet	61.41	65.20	63.25
	PhoBERT_base-v2	w2v2-Viet	57.47	64.57	60.81
		XLSR-53-Viet	61.00	66.82	63.77
	ViDeBERTa_base	w2v2-Viet	55.81	60.72	58.16
		XLSR-53-Viet	56.43	62.24	59.19
	XLM-R_base	w2v2-Viet	56.85	65.39	60.82
		XLSR-53-Viet	59.34	66.67	62.79
	XLM-R_large	w2v2-Viet	57.68	65.57	61.37
		XLSR-53-Viet	58.92	66.67	62.56
CAL.	PhoBERT_large	w2v2-Viet	34.85	64.79	45.32
		XLSR-53-Viet	33.79	60.93	43.47
	PhoBERT_base-v2	w2v2-Viet	40.45	60.41	48.46
		XLSR-53-Viet	41.06	58.15	48.13
	ViDeBERTa_base	w2v2-Viet	11.82	55.32	19.48
		XLSR-53-Viet	11.67	53.85	19.18
	XLM-R_base	w2v2-Viet	38.64	60.71	47.22
		XLSR-53-Viet	38.79	54.82	45.43
	XLM-R_large	w2v2-Viet	41.82	60.93	49.60
		XLSR-53-Viet	38.64	56.42	45.86
Micro average	PhoBERT_large	w2v2-Viet	41.07	61.63	49.29
		XLSR-53-Viet	40.47	60.09	48.36
	PhoBERT_base-v2	w2v2-Viet	44.22	61.06	51.29
		XLSR-53-Viet	43.89	59.88	50.66
	ViDeBERTa_base	w2v2-Viet	23.74	48.42	31.86
		XLSR-53-Viet	23.81	47.80	31.78
	XLM-R_base	w2v2-Viet	44.95	59.88	51.35
		XLSR-53-Viet	44.43	58.15	50.37
	XLM-R_large	w2v2-Viet	45.29	60.37	51.75
		XLSR-53-Viet	45.39	59.33	51.43
Macro average	PhoBERT_large	w2v2-Viet	33.94	53.59	39.65
		XLSR-53-Viet	38.97	54.29	42.09
	PhoBERT_base-v2	w2v2-Viet	35.13	52.24	40.57
		XLSR-53-Viet	37.72	56.79	43.61
	ViDeBERTa_base	w2v2-Viet	16.20	34.79	19.70
		XLSR-53-Viet	16.37	33.86	19.76
	XLM-R_base	w2v2-Viet	37.01	50.87	42.19
		XLSR-53-Viet	41.86	51.70	44.76
	XLM-R_large	w2v2-Viet	37.66	53.68	43.25
		XLSR-53-Viet	42.92	55.16	46.32

Table 7: NER results by entity types (in percent) on ASR output of test set using various pre-trained language models and ASR models. Metrics shown are Precision, Recall, and overall micro/macro F1 score.