Medical Spoken Named Entity Recognition

Khai Le-Duc, David Thulke, Hung-Phong Tran, Long Vo-Dang, Khai-Nguyen Nguyen, Truong-Son Hy, Ralf Schlüter

NAACL 2025

Motivation

No research applies NER to real-world medical speech.

A medical spoken NER dataset is needed:

1. ASR outputs are noisy
2. Annotations can be inconsistent
3. Obtaining accurate medical NER from natural speech is challenging

Contributions

1. VietMed-NER - the first publicly available medical spoken NER dataset
2. We present the baselines on several state-of-the-art pre-trained language models
3. We conduct extensive quantitative and qualitative error analysis

VietMed-NER

- Based on the VietMed medical ASR dataset
- 18 medical entity types - the largest spoken NER dataset

Annotation Process:

1. Create a gazetteer list of NERs from manually annotated chosen samples from VietMed
2. Automatically map entities to the dataset with a sorted gazetteer list

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

3. Iteratively review and update annotations

Dataset Statistics

	Definition	Train		Dev		Test		All	
		Total	Uni.	Total	Uni.	Total	Uni.	Total	Uni.
AGE	Age of a person	447	43	108	25	611	83	1166	151
GENDER	Gender of a person	202	30	46	15	451	33	699	78
JOB	Job of a person	543	32	133	16	562	43	1238	91
LOCATION	Locations and places	284	66	76	31	317	75	677	172
ORGANIZATION	Organizations	19	14	2	2	58	23	79	39
DISEASESYMTOM	Symptoms and diseases	2699	518	683	209	1334	357	4716	1084
DRUGCHEMICAL	Bio-chemical substances and drugs	1054	255	263	104	684	136	2001	495
FOODDRINK	Food and beverage	243	77	48	26	247	43	538	146
ORGAN	Anatomical features, e.g. organs, cells	1827	252	444	122	1190	172	3461	546
PERSONALCARE	Personal care, e.g. hygiene routines, skin care	353	114	82	38	95	10	530	162
DIAGNOSTICS	Diagnostic procedures, e.g. lab tests, imaging	371	53	91	25	292	36	754	114
TREATMENT	Non-surgical treatment, e.g. rehab., injection	726	69	174	25	230	17	1130	111
SURGERY	Surgical procedures, e.g. implants, neurosurgery	197	29	55	13	270	37	522	79
PREVENTIVEMED	Preventive medicine	341	53	80	25	18	6	439	84
MEDDEVICETECHNIQUE	Medical devices, instruments, and techniques	324	84	67	30	603	144	994	258
UNITCALIBRATOR	Medical calibration, e.g. number of doses, calories	800	155	215	75	251	106	1266	336
TRANSPORTATION	Means of transportation	5	2	3	3	27	10	35	15
DATETIME	Date and time	674	155	159	65	657	133	1490	353
#Entities in total		11109	2001	2729	849	7897	1464	21735	4314
#Sentences		4620		1150		3500		9270	

Experimental Setup

Spoken NER Pipeline:

- Cascaded approach:
  ASR transcription ⟶ Text-based NER

Pretrained ASR Models:

- w2v2-Viet (mono) + XLSR-53-Viet (multi)
- Comparable WERs (29.0% vs. 28.8%)

NER Models:

- **Monolingual Encoders:** PhoBERT, ViDeBERTa
- **Multilingual Encoders:** XLM-R
- **Seq2Seqs:** BARTpho, ViT5, mBART-50

Model	#Params	#Data
PhoBERT_base	135M	20GB
PhoBERT_large	370M	
PhoBERT_base-v2	135M	140GB
ViDeBERTa_base	86M	298GB
XLM-R_base	270M	2.5TB
XLM-R_large	550M	2.5TB
mBART-50	611M	3.9TB
ViT5_base	310M	888GB
BARTpho	396M	20GB

Results on Human Transcripts

NER Model	Prec.	Rec.	F1
BARTpho	0.64	0.73	0.68
mBART-50	0.64	0.66	0.65
PhoBERT_base	0.67	0.78	0.72
PhoBERT_base-v2	0.68	0.79	0.74
PhoBERT_large	0.69	0.77	0.73
ViDeBERTa_base	0.50	0.41	0.45
ViT5_base	0.64	0.74	0.69
XLM-R_base	0.64	0.73	0.69
XLM-R_large	0.71	0.77	0.74

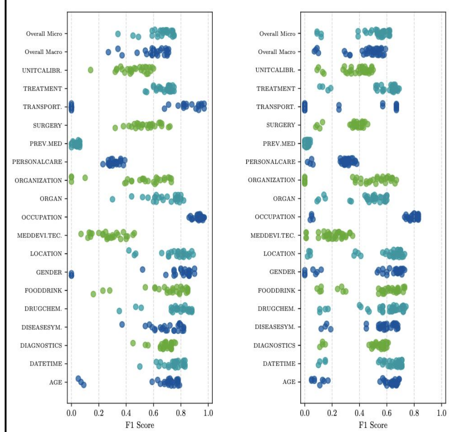
Results on ASR Transcripts

NER	ASR	Prec.	Rec.	F1
ViDeBERTa_base	XLSR-53-Viet	0.45	0.34	0.39
	w2v2-Viet	0.45	0.34	0.39
ViT5_base	XLSR-53-Viet	0.52	0.46	0.48
	w2v2-Viet	0.53	0.46	0.49
mBART-50	XLSR-53-Viet	0.35	0.05	0.09
	w2v2-Viet	0.35	0.05	0.09
BARTpho	XLSR-53-Viet	0.56	0.50	0.53
	w2v2-Viet	0.55	0.50	0.52
PhoBERT_base_v2	XLSR-53-Viet	0.57	0.57	0.57
	w2v2-Viet	0.58	0.56	0.57
PhoBERT_base	XLSR-53-Viet	0.56	0.56	0.56
	w2v2-Viet	0.56	0.56	0.56
PhoBERT_large	XLSR-53-Viet	0.57	0.55	0.56
	w2v2-Viet	0.58	0.55	0.56
XLM-R_base	XLSR-53-Viet	0.54	0.52	0.53
	w2v2-Viet	0.54	0.52	0.53
XLM-R_large	XLSR-53-Viet	0.60	0.56	0.58
	w2v2-Viet	0.60	0.56	0.58

Observations

1. Pre-trained multilingual models outperform monolingual if they overcome the capacity dilution problem
2. Encoders outperformed seq2seqs
3. Multiling. pre-training of the acoustic model does not affect cascaded NER performance
4. Performance of all models drop moving from human ⟶ ASR transcripts

Error Analysis



Quantitative Weaknesses:

1. Misclassify *PREVENTIVEMED* due to overlap with *DRUGCHEMICAL* and *TREATMENT*
2. Good performance in straightforward categories, low performance in more complex categories.

Qualitative Weaknesses:

1. **Ambiguity:** similar descriptors and context leads to confusion between *LOCATION v.s. ORGANIZATION* and *DRUGCHEMICAL v.s. FOODDRINK*
2. **Span Errors:** truncated multi-word entities (e.g., "high blood" vs "high blood pressure") and splitting of compound entities leads to errors

Limitations

1. **Annotation**: We have not quantify the time and performance gain of our annotation approach
2. **Evaluation Metrics:** Standard metrics like WER overlook the critical importance of medical terms.

PAPER

CODE + DATASET