

Medical Spoken Named Entity Recognition

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





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
- 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:

- Create a gazetteer list of NERs from manually annotated chosen samples from VietMed
- 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	
Definition	Total	Uti.	Total	Uni.	Total	Uni.	Total	Uni.
Age of a person	447	43	108	25	611	83	1166	151
Gender of a person	202	30	46	15	451	33	699	78
Job of a person	543	32	133	16	562	43	1238	91
Locations and places	284	66	76	31	317	75	677	172
Organizations	19	14	2	2	58	23	79	39
Symptoms and diseases	2699	518	683	209	1334	357	4716	1084
Bio-chemical substances and drugs	1054	255	263	104	684	136	2001	495
Food and beverage	243	77	48	26	247	43	538	146
Anatomical features, e.g. organs, cells	1827	252	444	122	1190	172	3461	546
Personal care, e.g. hygiene routines, skin care	353	114	82	38	95	10	530	162
Diagnostic procedures, e.g. lab tests, imaging	371	53	91	25	292	36	754	114
Non-surgical treatment, e.g. rehab., injection	726	69	174	25	230	17	1130	111
Surgical procedures, e.g. implants, neurosurgery	197	29	55	13	270	37	522	79
Preventive medicine	341	53	80	25	18	6	439	84
Medical devices, instruments, and techniques	324	84	67	30	603	144	994	258
Medical calibration, e.g. number of doses, calories	800	155	215	75	251	106	1266	336
Means of transportation	5	2	3	3	27	10	35	15
Date and time	674	155	159	65	657	133	1490	353
	11109	2001	2729	849	7897	1464	21735	4314
	4620		1150		3500		9270	
	Detinition Grader of sproon Grader of sproon Locations and places Organization Symptoms and focuses Symptoms and diseases Tool and borcage Anomotia fatterses, e.g. pagas, cells Prontal are, g. pagas motions, skill water, inauging Disgondie procedures, e.g. pagas, cells Mons auginal transmise, e.g. phab, singection Songical proceeders, e.g. pagas, scille Neusagina transmise, e.g. phab, singection Songical proceeders, e.g. pagas, scille Markail device, instruments, and techniques Markail device, instrum	Detinition Tend Age of sproon 447 Grader of sproon 440 Soft of sproon 543 Locations and places 249 Symptoms and focusos 249 Symptoms of diseases 209 Soft and person 1054 Food and boxes 209 Disposition produces e.g. appart, cells 1427 Promoti care, g.g. appart, cells 1427 Disposition produces, e.g. appart, cells 1427 Non-surgical treatments, e.g. bub nots, marging 1717 Non-surgical treatments, e.g. and the changes 214 Markai discuss, and motion discussion 352 Markai discuss, and motion discussion 352 Markai discussion 1403 Markai discussion 352 Markai discussion 352	Definition Total Unitation Grader of proon 421 43 Grader of proon 202 30 Bof of proon 431 42 Leations and place 284 66 Opportunitions 19 14 Symptoms of decases 299 181 Second and decases 299 181 Schemical substances and drops 1064 255 Food and borcage 244 67 Anomical fasters: e.g. aprox. cells 1245 252 Promotil care, e.g. bygine motines, salit care 333 141 Disporting modeline, e.g. fundish, injection 728 69 Sugging for graderine, and work of the solid science 135 141 Madical calibration, e.g. and the disposed on the solid science 341 15 Madical calibration, e.g. and the disposed on the solid science 361 15 Madical calibration, e.g. and the disposed on the solid science 361 15 Date and time 6420 11160 2000	Detention Tool Tool Tool Grader of proces 421 64 064 Grader of proces 202 30 45 José of proces 202 30 45 José of proces 218 16 75 Junition on fighters 218 46 76 Opprimitation 19 14 22 30 Schemical substances and drugs 1054 155 315 483 Tool and borcage 201 71 48 Anomical fasters: e.g. organs: edition 315 114 82 Dispositior providens: e.g. organs: edition 335 114 82 306 114 82 Dispositior providens: e.g. organs: edition intervingent 710 78 74 74 74 74 Namerical and transition e.g. organs: edition intervingent 710 82 76 114 82 116 82 116 82 116 116 116 116 116 116 116 1	Dottmom Total Usit Total Usit get opron 44 10 20 20 40 10 20 20 60 45 100 101 40 108 20 20 40 45 50 40 100 40 15 40 108 32 20 30 46 15 50 40 108 32 103 101 100 40 17 40 15 40 100 40 71 40 55 40 100 100 40 71 40 55 40 100 100 40 71 40 55 40 100 100 100 101 25 20 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 15 100 50 100 100	Determine Total Usin Total	Duttinue Tool	Duttimote Tool

Experimental Setup

Spoken NER Pipeline:

- Cascaded approach: ASR transcription → Text-based NER
- As transcription \rightarrow lext-based

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	20GE		
PhoBERT_large	370M			
PhoBERT_base-v2	135M	140GE		
ViDeBERTa_base	86M	298GI		
XLM-R_base	270M	2.5TH		
XLM-R_large	550M	2.5TE		
mBART-50	611M	3.9TE		
ViT5_base	310M	888GI		
BARTpho	396M	20GI		

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 \rightarrow ASR transcripts

Error Analysis



Quantitative Weaknesses:

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

Qualitative Weaknesses:

- Ambiguity: similar descriptors and context leads to confusion between LOCATION v.s. ORGANIZATION and DRUGCHEMICAL v.s. FOODDRINK
- 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.



