NER Model Testing

Abhishek Satpathy

Scan the QR Code to look at the data or scripts

Data

Scripts

Introduction

NER stands for (Named Entity Recognition). Named entity recognition (NER) methods address the challenge of extracting important information from unstructured text.

55 year ENTITY old male ENTITY found unconscious ENTITY driver side passenger ENTITY seat of his car his wife ENTITY reported that he snorted a line of heroin ENTITY before just prior to losing to consciousness ENTITY patient ENTITY originally presented unresponsive ENTITY and pale ENTITY with shallow ineffective respirations ENTITY at a rate ENTITY of about 5 with heart rate ENTITY was 118 his blood pressure ENTITY was 205 over 119 his blood glucose ENTITY levels 126 his O2 saturations ENTITY was were 94% patient ENTITY required bag mask ventilation ENTITY with with attached oxygen ENTITY however after 0.25 mg of naloxone intravenously patient ENTITY is now awake ENTITY and breathing ENTITY normally with ENTITY Improvement ENTITY in Vital Signs ENTITY and respiratory status ENTITY and no longer needs supplemental oxygen ENTITY

Problem Description

My goal was to evaluate the performance of various configurations of two models (SciSpacy and MetaMap) on NER tasks in the Hydron dataset. The models I tested can be found in a list below.

Models

- MetaMap (Symptoms only)
- MetaMap (All tags)
- MetaMap (Custom tags)
- SciSpacy en_core_sci_sm
- SciSpacy en_core_sci_md
- SciSpacy en_core_sci_lg
- SciSpacy en_core_sci_scibert

Custom NER Models

- SciSpacy en_ner_craft_md
- SciSpacy en_ner_jnlpba_md
- SciSpacy en_ner_bc5cdr_md
- SciSpacy en_ner_bionlp13cg_md

Metamap Trigger Word

- Metamap (Symptoms only)
- Metamap (All tags)
- Metamap (Custom Tags)

You

Take all these tags and provide those that might be useful to EMS natural language processing:

ChatGPT

For EMS (Emergency Medical Services) natural language processing, the following tags might be particularly useful:

- **acab** (Acquired Abnormality): Relevant for identifying acquired medical conditions.
- **aggp** (Age Group): Useful for understanding age-related medical information.
- **anab** (Anatomical Abnormality): Helps identify anatomical abnormalities which could be crucial in emergency situations.
- **anim** (Animal): May be relevant for cases involving animal-related injuries or illnesses.
- **bact** (Bacterium): Important for identifying bacterial infections or related conditions.
- **bdsy** (Body System): Essential for understanding the affected body systems in medical emergencies.
- **bpoc** (Body Part, Organ, or Organ Component): Vital for identifying affected body parts or organs.
- **clnd** (Clinical Drug): Relevant for understanding medications involved in emergencies.
- **diap** (Diagnostic Procedure): Helps identify diagnostic procedures relevant to the situation.
- **dsyn** (Disease or Syndrome): Important for identifying diseases or syndromes.
- **fndg** (Finding): Essential for understanding clinical findings relevant to the emergency.
- **inpo** (Injury or Poisoning): Crucial for identifying injuries or cases of poisoning.
- **medd** (Medical Device): Relevant for understanding medical devices involved in the situation.
- **mobd** (Mental or Behavioral Dysfunction): Important for situations involving mental health emergencies.
- **phsu** (Pharmacologic Substance): Relevant for identifying medications or substances involved.
- **phob** (Physical Object): May be relevant for cases involving physical objects causing harm or injury.
- **sosy** (Sign or Symptom): Crucial for identifying signs and symptoms of medical emergencies.
- **tmco** (Temporal Concept): Relevant for understanding the timing or duration of medical conditions or interventions.
- **topp** (Therapeutic or Preventive Procedure): Essential for understanding medical procedures relevant to the emergency.

These tags cover a wide range of concepts relevant to medical emergencies, from anatomical abnormalities to medications and procedures, which can be useful for natural language processing in EMS contexts.

Evaluation Metrics

[Extracting medication information from clinical text](#) [Uzuner et al.]

$$\text{Phrase-level precision (PP)} = \frac{\text{\#Correctly returned phrases by system}}{\text{\#Phrases returned by the system}} \qquad \text{Equation 1}$$

$$\text{Phrase-level recall (PR)} = \frac{\text{\#Correctly returned phrases by system}}{\text{\#Phrases in gold standard}} \qquad \text{Equation 2}$$

$$\text{Phrase-level F-measure} = \frac{\left(\beta^2+1\right) \times \text{PP} \times \text{PR}}{\left(\beta^2 \times \text{PP}\right) + \text{PR}} \text{where} \beta = 1 \qquad \text{Equation 3}$$

$$\text{Token-level precision (TP)} = \frac{\text{\#Correctly returned tokens from each phrase in system output}}{\text{\#tokens in system output}} \qquad \text{Equation 4}$$

$$\text{Token-level recall (TR)} = \frac{\text{\# Correctly returned tokens from each phrase in system output}}{\text{\#tokens in gold standard}} \qquad \text{Equation 5}$$

$$\text{Token-level F-measure} = \frac{\left(\beta^2+1\right) \times \text{TP} \times \text{TR}}{\left(\beta^2 \times \text{TP}\right) + \text{TR}} \text{where} \beta = 1 \qquad \text{Equation 6}$$

Evaluation Metrics

Didn’t consider entity types because Hydron ground truth did not include entity types in a format processable by either SciSpacy models or MetaMap. Another thing to consider is that SciSpacy large language models (i.e. models not specified for NER), only output ‘ent’ as entity type

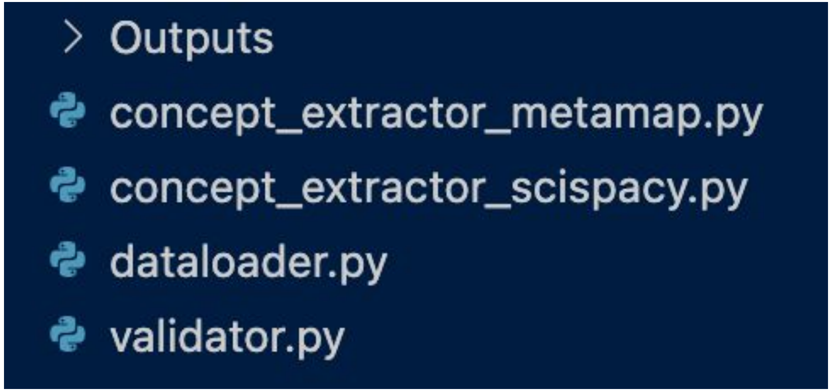
55 [year ENTITY] old [male ENTITY] found [unconscious ENTITY] driver [side passenger ENTITY] seat of his car his [wife ENTITY] reported that he snorted a line of [heroin ENTITY] before just prior to losing to [consciousness ENTITY] [patient ENTITY] originally presented [unresponsive ENTITY] and [pale ENTITY] with [shallow ineffective respirations ENTITY] at a [rate ENTITY] of about 5 with [heart rate ENTITY] was 118 his [blood pressure ENTITY] was 205 over 119 his [blood glucose ENTITY] levels 126 his [O2 saturations ENTITY] was were 94% [patient ENTITY] required bag [mask ventilation ENTITY] with with [attached oxygen ENTITY] however after 0.25 mg of naloxone intravenously [patient ENTITY] is now [awake ENTITY] and [breathing ENTITY] [normally with ENTITY] [Improvement ENTITY] in [Vital Signs ENTITY] and [respiratory status ENTITY] and no longer needs [supplemental oxygen ENTITY]

Example of Gold Label

age, True, 55; gender, True, male; loss of consciousness, True, unconscious; pale, True, pale; resp, True, 5; bradypnea, True, 5; pulse, True, 118; tachycardia, True, 118; bp, True, 205/119; hypertension, True, 205/119; glucose, True, 126; spo2, True, 94;

Scripts

In order to accomplish this task I designed 4 classes. Utilizing pandas, spacy, and pymetamap to apply NER to the Hydron dataset, and then process and validate the data.



Scripts

validator.py - manual implementation of phrase and token based recall, precision, and F1.

dataloader.py - Data formatting class for all purpose loading, editing, and formatting.

Concept_extractor_xxxxxx - NER implementation for SciSpacy and MetaMap.

Results “Name”

Mean Names	PLP	PLR	P F1	TLP	TLR	T F1
Metamap Symptoms Only	0.45	0.18	0.25	0.39	0.16	0.22
Metamap All	0.07	0.42	0.11	0.06	0.35	0.10
SciSpacy en_core_sci_sm	0.12	0.44	0.19	0.10	0.34	0.15
SciSpacy en_core_sci_md	0.13	**0.48**	0.20	0.10	0.36	0.15
SciSpacy en_core_sci_scibert	0.13	0.44	0.19	0.10	0.34	0.15
SciSpacy en_core_sci_lg	0.13	0.47	0.20	0.10	0.36	0.15
SciSpacy en_ner_craft_md	0.13	0.03	0.04	0.13	0.03	0.04
SciSpacy en_ner_jnlpba_md	0.00	0.00	0.00	0.00	0.00	0.00
SciSpacy en_ner_bc5cdr_md	**0.55**	0.41	**0.44**	0.43	0.30	**0.34**
SciSpacy en_ner_bionlp13cg_md	0.07	0.05	0.06	0.07	0.05	0.06
MetaMap Filtered Tags	0.09	0.31	0.14	0.08	0.28	0.12
MetaMap All Trigger Words	0.10	0.46	0.17	0.09	**0.39**	0.14
MetaMap Filtered Tags Trigger Words	0.18	0.42	0.24	0.15	0.36	0.21
MetaMap Symptoms Only Trigger Words	0.52	0.26	0.34	**0.46**	0.23	0.30

Results “Measures”

Mean Measures	PLP	PLR	P F1	TLP	TLR	T F1
Metamap Symptoms Only	0.36	0.17	0.23	0.30	0.14	0.19
Metamap All	0.05	0.32	0.08	0.04	0.25	0.07
SciSpacy en_core_sci_sm	0.11	0.42	0.17	0.09	0.32	0.13
SciSpacy en_core_sci_md	0.13	0.49	0.20	0.09	0.36	0.15
SciSpacy en_core_sci_scibert	0.12	0.44	0.19	0.09	0.34	0.14
SciSpacy en_core_sci_lg	0.12	0.45	0.19	0.09	0.36	0.15
SciSpacy en_ner_craft_md	0.02	0.00	0.01	0.02	0.00	0.01
SciSpacy en_ner_jnlpba_md	0.00	0.00	0.00	0.00	0.00	0.00
SciSpacy en_ner_bc5cdr_md	0.45	0.37	**0.39**	0.35	0.27	**0.29**
SciSpacy en_ner_bionlp13cg_md	0.05	0.03	0.03	0.05	0.03	0.03
MetaMap Filtered Tags	0.07	0.25	0.11	0.06	0.22	0.09
MetaMap All Trigger Words	0.11	**0.52**	0.18	0.09	**0.45**	0.15
MetaMap Tags Trigger Words	0.17	0.44	0.24	0.14	0.37	0.20
MetaMap Symptoms Only Trigger Words	**0.47**	0.26	0.33	**0.41**	0.23	0.29

Results “Measures”

With regard to measures, (the information extracted in Ground Truth), the SciSpacy NER model trained on the bc5cdr dataset performed best overall, although it was beaten out in individual measures when compared to the output of concept.trigger in the MetaMap model (the entity boundaries detected in MetaMap).

Results “Names”

With regard to entity names the bc5cdr model once again performed best overall, but was slightly beaten out in individual categories by other models like SciSpacy en_core_sci_md and some of the MetaMap trigger word models.

Future Investigation

Although the bc5cdr model produced the best results overall, the MetaMap models also showed promise. The MetaMap models often had higher recall, but lost to the bc5cdr model w.r.t. precision. Experimenting with different combinations of tags may result in a MetaMap model with high precision and recall.

Thank you!