Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen

K. Lybarger, M. Ostendorf, and M. Yetisgen, "Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction," Journal Biomedical Informatics, vol. 113, p. 103631, 2021. DOI: 10.1016/j.jbi.2020.103631.
Overview

- Natural language processing (NLP)
- Clinical data, specifically text
- Social determinants of health (SDOH)
Natural Language Processing (NLP)

• Natural language
  • Language created by a human for a human

• NLP
  • Interpreting, analyzing, or transforming language

• Many tasks/areas

• Rapidly developing field
  • Advances in machine learning
    • Deep neural networks
    • Transfer learning
  • More powerful computing
    • Graphical processing units (GPUs)
Clinical Data for Secondary Use

Electronic health record
  Structured data

Clinical text

Unstructured text

Information extraction

Structured representation

Secondary use applications
  Large-scale (retrospective studies)
  Real-time (clinical decision-support)
Clinical text

- Document patient history & treatments
- Part of the Electronic Health Record (EHR)
- Contain ≥ 80% of patient information in EHR
- Created by clinicians for clinicians
- Requires interpretation or extraction

HABITS:
- Tobacco Use: [2-3 cigs, vape per day 4-5 yrs]
- Alcohol Use: [2-3 beers per week]
- Drug Use: [none]

SOCIAL HISTORY:
- housed currently but homeless 1 year ago
- spends time during the day at a homeless shelter

living situation
- status = current
- type = housed

living situation
- status = past
- type = homeless

tobacco
- status = current
- type = [cigarettes, vaping]
- history = 4-5 years

alcohol
- status = current
- type = [beer]

drug
- status = none
Clinical Information Extraction

- Clinical text
  - Range of text-encoded information relevant to healthcare and public health

- Information extraction
  - High performing **extraction models needed** to access text-encoded information
  - Extraction models based on **data-driven machine learning**
  - High-quality **annotated data needed** to train models

- Challenges
  - Data heterogeneity (varies by institution, specialty, etc.)
  - Limited availability of annotated data (cost and privacy regulations)
  - Phenomena of greatest interest may be infrequent

---

**Input (text)**

IV drug use methamphetamine - relapsed 3 days ago

**Extraction model**

---

**Output (structured)**

```json
{drug:
  status = current,
  type = {methamphetamine},
  method = intravenous}
```
Social Determinants of Health

Goal
• Conditions in which people work and live that impact health outcomes
• Create more comprehension patient representation that includes social, behavioral, and environmental factors to facilitate clinical research

Motivation
• SDOH may be contributing to decreased life expectancy
• Detailed descriptions of SDOH found in clinical text

Tasks
• Create annotated data set rich in SDOH information
• Extract detailed representations of SDOH
• Emphasize less frequent, high-risk health factors
  • Substance abuse, homelessness, and unemployment

Image source: https://healthix.org/2019/10/18/partnership-to-drive-integration-of-sdoh-into-health-data/
Social History Annotation Corpus (SHAC)
Social History Annotation Corpus (SHAC)

- Size
  - 4,480 social history sections

- Sources
  - MIMIC-III
    - 2001–2012
  - UW and Harborview Medical Centers
    - 2008–2019

- Annotation
  - Event-based annotations
  - 18K distinct events

- SDOH (12 determinants)
  - alcohol, drug, and tobacco use
  - employment
  - living status
  - physical activity
  - insurance
  - environmental exposure
  - sexual orientation
  - gender identity
  - country of origin
  - race

SocHx: Endorses 50 - pack - year smoking history, denies substance use
Event-based Annotations

- Characterize determinants across multiple dimensions
  - e.g. status, extent, and temporality

**Trigger**
- Anchors and disambiguate events
- Double (event type, span)

**Labeled arguments**
- Triple (argument type, span, subtype)
- Subtype normalizes span to fixed set of classes

**Span-only arguments**
- Double (argument type, span)
- Span not easily normalized

SOCIAL HISTORY: Used to be a chef; currently unemployed.

Tobacco Use: quit 7 years ago; 15 - 20 pack years
Annotated phenomena
(most frequent phenomena only)

| Event type          | Field type     | Arguments                      | Argument subtypes                                     |
|---------------------|----------------|--------------------------------|-------------------------------------------------------|
| Alcohol, Drug,      | Trigger*       | --                             | --                                                    |
| or Tobacco          | Labeled        | Status*                        | {none, current, past}                                 |
|                     | arguments      |                                |                                                       |
|                     | Span-only      | Amount, Duration, Frequency,   | --                                                    |
|                     | arguments      | History, & Type                |                                                       |
| Employment          | Trigger*       | --                             | --                                                    |
|                     | Labeled        | Status*                        | {employed, unemployed, retired, on disability, student, homemaker} |
|                     | arguments      |                                |                                                       |
|                     | Span-only      | Duration, History, & Type      | --                                                    |
| Living status       | Trigger*       | --                             | --                                                    |
|                     | Labeled        | Status*                        | {current, past, future}                               |
|                     | arguments      |                                |                                                       |
|                     | Type*          |                                | {alone, with family, with others, homeless}           |
|                     | Span-only      | Duration & History             | --                                                    |

*indicates the argument is required.
SHAC Annotation Statistics

Source and splits

| Source        | Train | Dev | Test |
|---------------|-------|-----|------|
| MIMIC         | 1,316 | 188 | 376  |
| UW Dataset    | 1,820 | 260 | 520  |
| TOTAL         | 3,136 | 448 | 896  |

| Selection     | Train | Dev | Test |
|---------------|-------|-----|------|
| Random        | 29%   | 100%| 100% |
| Active        | **71%** | -- | --   |

Event type distribution

- Drug: 4,133
- Tobacco: 4,049
- Alcohol: 4,048
- Living status: 3,267
- Employment: 2,275
- Enviro. expos.: 212
- Country: 157
- Physical activity: 71
- Sexual orient.: 49
- Race: 25
- Insurance: 11
- Gender: 1
Slot filling interpretation for secondary use

Event 1
- Event type = Drug
- Status = past
- History = “none since [**2174**]”
- Type = “IVDU”

Event 2
- Event type = Drug
- Status = current
- Type = “cocaine”
Evaluation and annotation scoring

• Triggers:
  • Aligned by minimizing the distance between span centers
  • Aligned triggers with same event type considered equivalent

• Arguments:
  • Arguments with the equivalent triggers compared

• Labeled arguments
  • Argument type and subtype must match
    • Span not considered

• Span-only arguments
  • Word-level comparison where argument types match
    • Partial matches can still contain useful information
SHAC annotator agreement

- Precision
  - \( P = \frac{TP}{TP + FP} \)

- Recall
  - \( R = \frac{TP}{TP + FN} \)

- F1
  - \( F1 = 2 \frac{P \times R}{P + R} \)

Based on 300 randomly selected MIMIC notes
Active Learning
Active Learning

• Problem
  • Available data > annotation budget
    • Millions of notes
  • Random selection is suboptimal
    • Phenomena of interest may be infrequent
      • e.g. homelessness or drug use
    • Many samples may be similar

• Active learning
  • Identifies samples that maximize model learning
  • Well established for single label tasks
  • Less established for more complex event extraction tasks
Active Learning using Surrogate Classifiers (ALSC)

- Focus on the phenomena most predictive of negative health outcomes
- Query function
  - *Informativeness* - potential to reduce classification uncertainty
  - *Diversity* - variation in the samples selected
Surrogate Classifiers

• Motivation
  • Not all arguments are equally important
  • Focus *informativeness* assessment on arguments that are most predictive of negative health outcomes

• Implementation
  • Map event annotations to document labels
    • Event extraction → text classification
  • Entropy predictions from surrogate classifiers used as a proxy for sample *informativeness*

---

**Event labels**

| Event                | Status | Alcohol | Tobacco | StatusTime | Amount | Duration |
|----------------------|--------|---------|---------|------------|--------|----------|
| No current alcohol use. |        |         |         |            |        |          |
|屿 current alcohol use. |        |         |         |            |        |          |

**Surrogate labels**

| Classifier       | Label     |
|------------------|-----------|
| Alcohol-Status   | none      |
| Drug-Status      | unknown   |
| Tobacco-Status   | current   |
| Employment-Type  | unknown   |
| Living status-Status | unknown |
Active learning query function

**Objective:** maximize $Q(B)$

$$Q(B) = \sum_{i \in B} (1 - s_i)^{\alpha} u(i)$$

- $B =$ batch of samples (social history sections)
- $s_i =$ similarity score of sample $i$ relative to $B$
- $u(i) =$ uncertainty of sample $i$
- $\alpha$ controls relative importance of scores ($\alpha > 0$)

**Similarity, $s_i$**

- Map all documents to a common vector space
- $a_{j,i} =$ cosine similarity of sample $j$ and $i$

$$s_i^m = \max_{j \in B, j \neq i} a_{j,i}$$

**Uncertainty, $u(i)$**

- Generated using surrogate classifiers
- $H_c(i) =$ entropy for sample, $i$, and event type, $c$

$$u(i) = \sum_{c=1}^{m} H_c(i)$$
Information Extraction
Multi-task Event Extractor

• Designed for low resource setting
  • Small(er) data sets
• Share information across event and argument types
• Learn dependencies between predicted phenomena
Input encoding

- Input sentence of tokens
- Bidirectional Encoder Representations from Transformers (BERT)
  - BERT fixed (no back propagation)
- bi-directional Long Short-Term Memory (bi-LSTM) network
Trigger prediction

- Sentence-level text classification task
- Each event type, $c$, detected using self-attention
- Separate binary classifiers (present/not present)
Labeled argument prediction

- Sentence-level text classification task
- Argument subtypes predicted using self-attention
- Multi-class classifiers for each event type and labeled argument
  - e.g. Alcohol-Status has subtypes {none, current, past}

```
Input encoding

Input sentence: SocHx: Endorses 50-pack-year smoking history

Status: current
Amount: 50 - pack - year
Tobacco: smoking history

Span-only arguments (CRF)

Labeled arguments (self-attention)

Trigger (self-attention)

event types

biLSTM

BERT
```
Span-only argument prediction

• Sequence tagging task
• Separate conditional random field (CRF) for each event type
• Input features include
  • bi-LSTM hidden state
  • Labeled argument probabilities
Results
Active learning performance

**Surrogate Classifiers**

- Training sets:
  - *initial*: 572 random samples
  - *+random*: initial + 244 random samples
  - *+active*: initial + 244 active samples
- *+active* significantly better than *+random*
  - $p < 0.06$

**Multi-Task Event Extractor**

- Trigger
  - Improvement not statistically significant
- Labeled and span-only arguments
  - Significant improvement ($p < 0.01$)
Active learning performance

• Largest performance gains for the Multi-task Event Extractor
• Less frequent, but extremely important health risk factors
Multi-task Event Extractor vs. Annotator Agreement

Multi-task Event Extractor trained on entire training set and evaluated on withheld test set

Annotator agreement on 300 MIMIC notes
Impact
Contributions

• Annotated data set
  • New, relatively large annotated data set
  • High-quality, detailed SDOH annotations

• Active learning
  • Novel active learning approach
  • Simplified surrogate classifiers used as proxy for more complex extraction task
  • Sample selection focused on criteria of greatest importance
  • Improved extraction performance relative to random sampling
    • Especially for less frequent, high-risk factors, like homelessness and drug use

• Information extraction
  • Multi-task extraction framework that shares information across tasks
  • Human-level extraction performance
Ongoing research

• SDOH extraction model
  • Integrated into the UW Enterprise Data Warehouse
  • Augmenting patient representations

• Exploring a range of outcomes
  • Likelihood of different cancers
  • Hospital admission length of stay and mortality
Data & Shared Task

• SDOH extraction task through National NLP Clinical Challenges (n2c2)
  • De-identified version of SHAC will be released
  • Participating teams will develop and evaluate information extraction models
  • High-performing and novel approaches will be invited to present a workshop
  • Registration in 2021, challenge in 2022
  • Details available at: https://projects.iq.harvard.edu/n2c2/2022-challenge

• Contribute to the advancement of clinical NLP
• Increase awareness and accessibility of SDOH information
Questions

Kevin Lybarger, PhD

lybarger@uw.edu
kevinlybarger.me

K. Lybarger, M. Ostendorf, and M. Yetisgen, "Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction," Journal Biomedical Informatics, vol. 113, p. 103631, 2021. DOI: 10.1016/j.jbi.2020.103631.