Extracting Social Determinants from Clinical Text

Meliha Yetisgen
University of Washington
September 11, 2019

Motivation

- Lifestyle factors and social determinants of health play a significant role both in clinical research and clinical care.
  - 5-10% of cancers can be attributed to hereditary factors while 90-95% have been found to be correlated with life style and environmental factors such as smoking, diet, and exercise.
- This information is documented usually in social history sections of clinical notes.
- In this work, our aim is to build robust NLP approach to extract social determinants from clinical text

Examples

- SOCIAL HISTORY: She does not drink or smoke.

- SOCIAL HISTORY: Patient admits tobacco use She consumes 3-5 cigarettes per day. Patient admits alcohol use. Drinking is described as social.

- SOCIAL HISTORY: The patient lives with his sister. He is unemployed.

Examples (Cont.)

- SOCIAL HISTORY: **Denies tobacco and alcohol use**. She endorses **marijuana use and a history of cocaine use five years ago**. Upon review of the Baptist lab systems, the patient has had multiple positive urine drug screens and as recently as February 2008 had a urine drug screen that was positive for **benzodiazepines, barbiturates, opiates, and marijuana** and as recently as 2005 with cocaine present as well.

Datasets

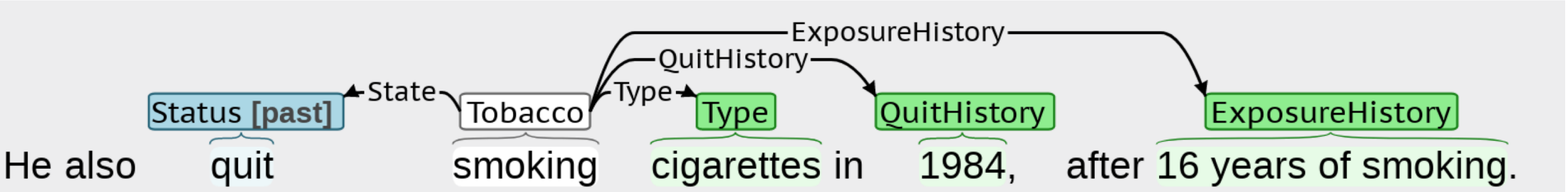
- YV-Notes (Yetisgen 2017)
  - Annotated data for supervised training
  - History & physical notes from MTSamples website
  - Social and behavioral factor annotations
  - 364 social history sections (1.2K sentences)
- MIMIC-III
  - Discharge summaries and Physician notes with social history sections
- University of Washington Clinical Notes
  - 2 million notes of patients who were prescribed with chronic opioid therapy between 2008-2018

Pilot work – YV-Notes

- MTSamples (www.mtsamples.com) - a large collection of publicly available transcribed medical records.
- 516 history and physical notes
  - Statistical section chunker – we identified 364 social history sections in 516 H&P notes.
- Annotated corpus available at: http://depts.washington.edu/bionlp/index.html?corpora

Annotation examples

Examples from (Yetisgen and Vanderwende, 2017)



- Multiple substances, each characterized by multiple labels
- Multi-label problem with correlated sentence-level (status) and word-level (type, method, …) labels

Annotation Guideline

Type	Entity	Example
Tobacco	Status	Possible discrete values: *current, past, none*
	Type	Default value is *tobacco*. We decided to annotate type if mention is more specific than *tobacco*. *e.g., cigarette.*
	Method	Default value is *smoking*. We decided to annotate method only if mention is different than *smoking*. e.g., *chew.*
	Amount	*e.g., minimal, significant, <#> packs*
	Frequency	*e.g., daily, occasionally, heavy*
	Exposure history	*e.g., since 1990*
	Quit history	*e.g., 3 years ago*
Alcohol	Status	Possible discrete values: *current, past, none*
	Type	Default value is *alcohol*. We decided to annotate method only if mention is more specific than *alcohol*. e.g., *beer, hard liquor*
	Method	Default value is *drinking*. We decided not to annotate method for alcohol since there is no other alternative method.
	Amount	*e.g., Significant, minimal, <#> [of glasses/drinks/bottles]*
	Frequency	*e.g., a week, on social occasions, heavy*
	Exposure history	*e.g., for many years, for the last 50 years*
	Quit history	*e.g., several years ago, in 1984*
Drug	Status	Possible discrete values: *current, past, none*
	Type	*e.g., illicit, cocaine, recreational, illegal, caffeine*
	Method	*e.g., iv, smoking*
	Amount	*e.g., abuse, significant*
	Frequency	*e.g., chronic, per day*
	Exposure history	*e.g., approximately 1 year ago*
	Quit history	*e.g., when he was 25, in 2005*

Statistics

Entity	Frequency		
	Tobacco	Alcohol	Drug
Status	278	254	154
Type	50	26	112
Method	4	0	10
Amount	78	69	25
Frequency	59	65	6
Exposure history	37	7	10
Quit history	37	6	2

Approach

- Leverage shared information across labels and substances
  - Correlation between sentence-level (status) and span-level (other entities) labels
    - e.g. if **status** is **none**, then span-level labels are unlikely
  - Similarities between substances
    - e.g. “denies” or “does not” suggests **status** is **none** for any substance
    - e.g. “per day” or “occasionally” could be **frequency** spans for any substance
- Implement neural, multi-task model
- Compare performance with approaches similar published baselines

Discrete models (Yetisgen & Vanderwende, 2017)

Text classification (status)

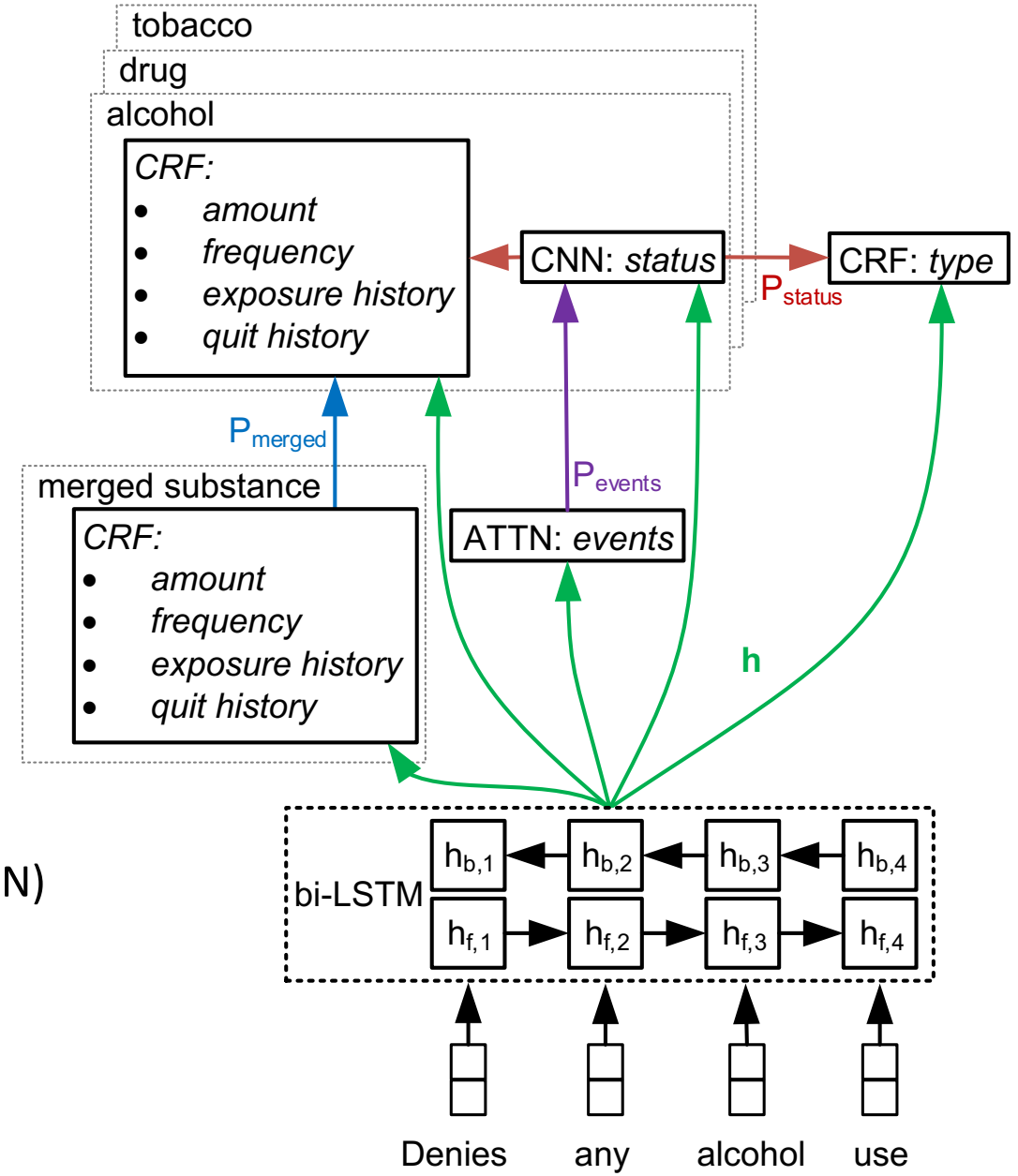
- Model
  - Maximum entropy (MaxEnt)
- Features
  - Word n-grams (n=1-3)
  - Gazetteer word lists
    - e.g. WordNet hyponyms for “alcohol”

Sequence tagging (other entities)

- Model
  - Conditional random fields (CRF)
- Features
  - Word n-grams (n=1-3)
  - Part-of-speech tags
  - Capitalization indicators
  - String type indicators
    - e.g. punctuation, number, etc.

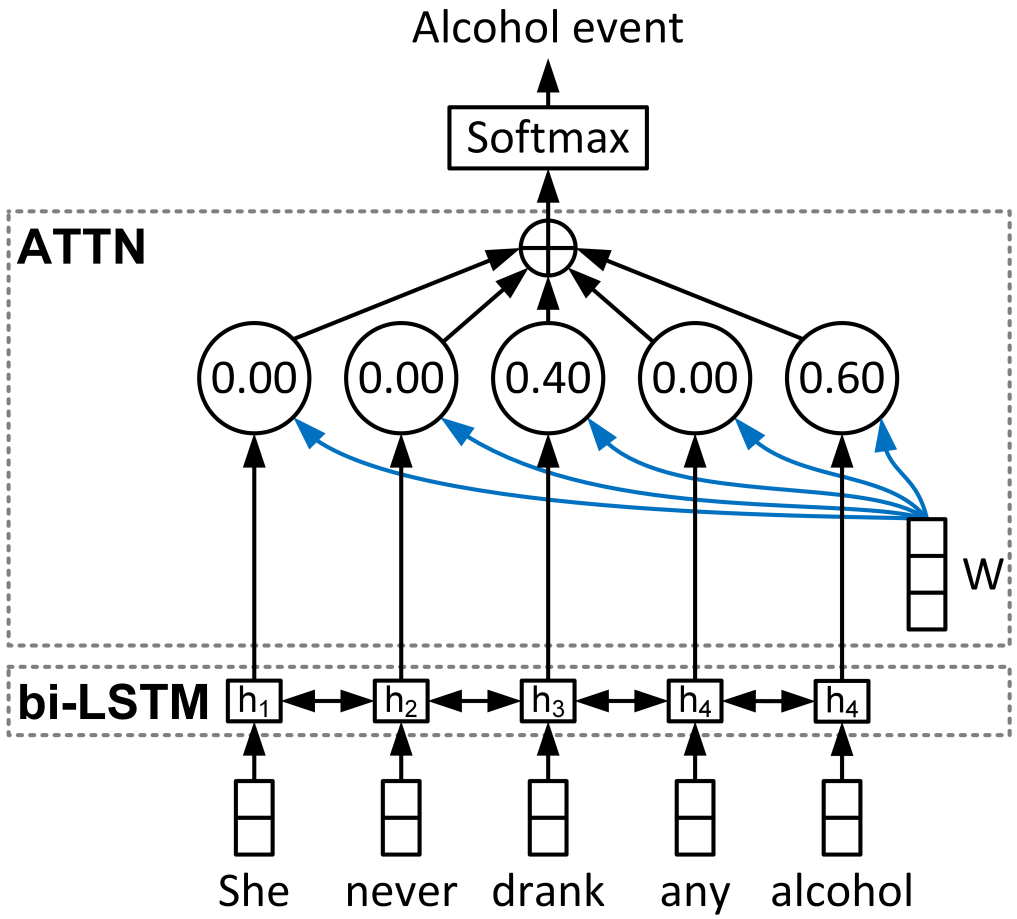
Multi-task model (Lybarger, 2018)

- Approach
  - Leverage shared information across labels
- Input
  - Pre-trained word embeddings
- Recurrent layer
  - Bidirectional long short-term memory (bi-LSTM)
- Text classification tasks (status)
    - bi-LSTM + Self-attention (ATTN)
    - bi-LSTM + Convolutional neural network (CNN)
- Sequence tagging tasks (other entities)
    - bi-LSTM + Multi-stage CRF

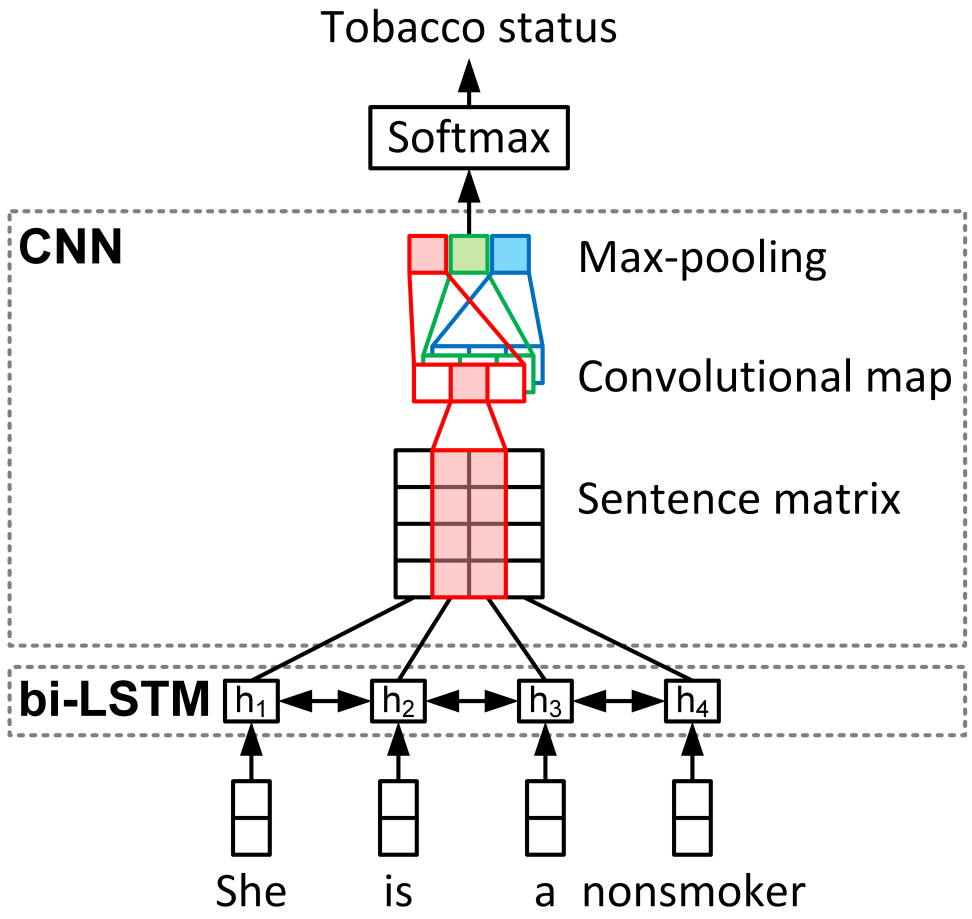


Neural text classification

Self-attention



Convolutional neural network



Data

YVnotes (Yetisgen and Vanderwende, 2017)

- Annotated data for supervised training
- History & physical notes from MTSamples website
- Social and behavioral factor annotations
- 364 social history sections (1.2K sentences)

MIMIC-III Dataset

- Pretrain word embeddings
  - Discharge summaries (60K) and physician notes (141K)
- Demonstrate model generalizes and provide resource for further analysis
  - 60K discharge summaries (5.9M sentences)

Training and Evaluation

- Split
  - 80% training, 20% testing
- Tuning
  - 5-fold cross validation on training set
  - Hyperparameters: regularization, hidden size, epochs, learning rate, etc.
- Training
  - Retrain on entire training set
- Evaluation
  - F_1 score on withheld test set

Results (Lybarger, Ostendorf, Yetisgen, 2018)

Entity	Substance	Model	True Positive	F_1♠
status	alcohol	MaxEnt	49	0.91
		Multi-task	52	**0.95**
	drug	MaxEnt	20	0.82
		Multi-task	24	**0.94**
	tobacco	MaxEnt	49	0.82
		Multi-task	53	**0.88**

♠micro averaged across labels

Entity	Model	True Positive	F_1♦
type	CRF	31	**0.94**
	Multi-task	31	0.93
amount	CRF	61	0.75
	Multi-task	65	**0.83**
exposure history	CRF	31	0.53
	Multi-task	37	**0.68**
frequency	CRF	35	0.72
	Multi-task	39	**0.78**
quit history	CRF	18	0.73
	Multi-task	21	**0.81**

♦micro averaged across substances

MIMIC-III Annotation

Unsupervised labeling of MIMIC-III

- 60K discharge summaries (5.9M sentences)
- 40K predicted to have substance event
- Unsupervised labels (BRAT format) on GitHub

Entity	Alcohol	Drug	Tobacco
status	44,536	20,725	45,244
type	4,756	14,509	11,443
amount	13,262	2,551	11,298
frequency	12,200	355	7,183
exposure history	770	396	4,639
quit history	176	72	10,241

Precision of **status** labels manually reviewed

- 50 discharge summaries

Substance	True Positive	Precision	Precision on YVnotes
alcohol	76	0.84	0.93
drug	80	0.89	0.96
tobacco	80	0.89	0.88

Discussion

- Substance abuse information extracted from clinical notes
  - Implemented state-of-the-art neural, multi-task extractor
  - Outperformed published baseline approaches
  - Improved performance will benefit secondary use applications

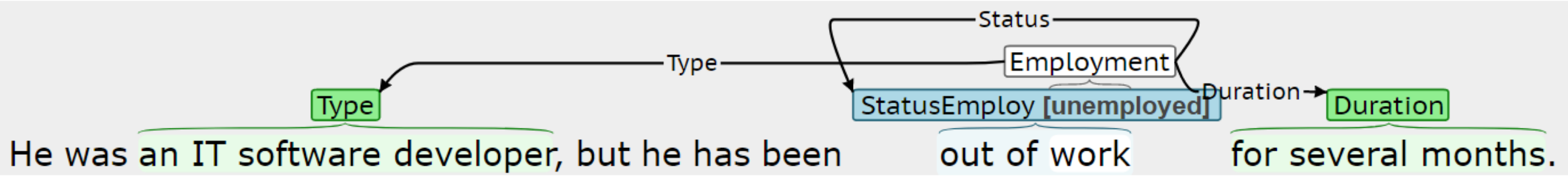
- Multi-task framework well-suited to other information
  - e.g. socio-demographic, behavioral, and environmental exposure factors

- Unsupervised labels predicted for MIMIC-III discharge summaries
  - Achieved high precision in prediction of status
  - Unsupervised labels available on GitHub

Ongoing work: Large scale annotation for social determinants with active learning

- Created a detailed annotation guideline for
  - Socio-Demographic (7):
    - Employment
    - Insurance
    - Living status
    - Sexual orientation
    - Gender identity
    - Country of origin
    - Race
  - Behavioral (2):
    - Substance abuse (tobacco, alcohol, drug)
    - Physical activity
  - Environmental exposure (1):
    - Environmental exposure

Employment:



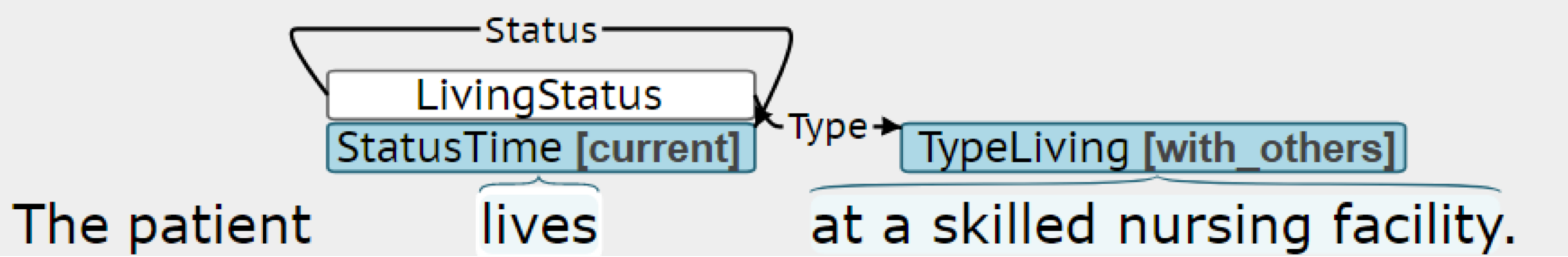
- Entities:
  - Trigger
  - Status (employed, unemployed, retired, on disability, student, homemaker)
  - Duration
  - History
  - Type

Insurance

Status label	Examples
yes	The patient will have continued pain medication coverage…
no	… her insurance would not cover the medication. He has been off insurance for over a year. …transferred to UIHC for a lack of insurance… …his inability to have adequate health insurance.

- Entities:
  - Trigger
  - Status (yes / no)

Living status



- Entities:
  - Trigger
  - Status (current, past, future)
  - Type (values: alone, with family, with others, homeless)
  - Duration
  - History

Sexual orientation

Status label	Examples
current	He is heterosexual and… His partners are male…
past	The patient participated in homosexual activity in Haiti during 1982…

- Entities:
  - Trigger
  - Status (current, past)
  - Type (values: heterosexual, homosexual, bisexual)

Gender identity

Priority	Examples
1. Gender identity category	He is transgender.
2. Description of gender identity	She is biologically male but identifies as female. He is biologically female but uses pronouns he/him. She is gender non-conforming.

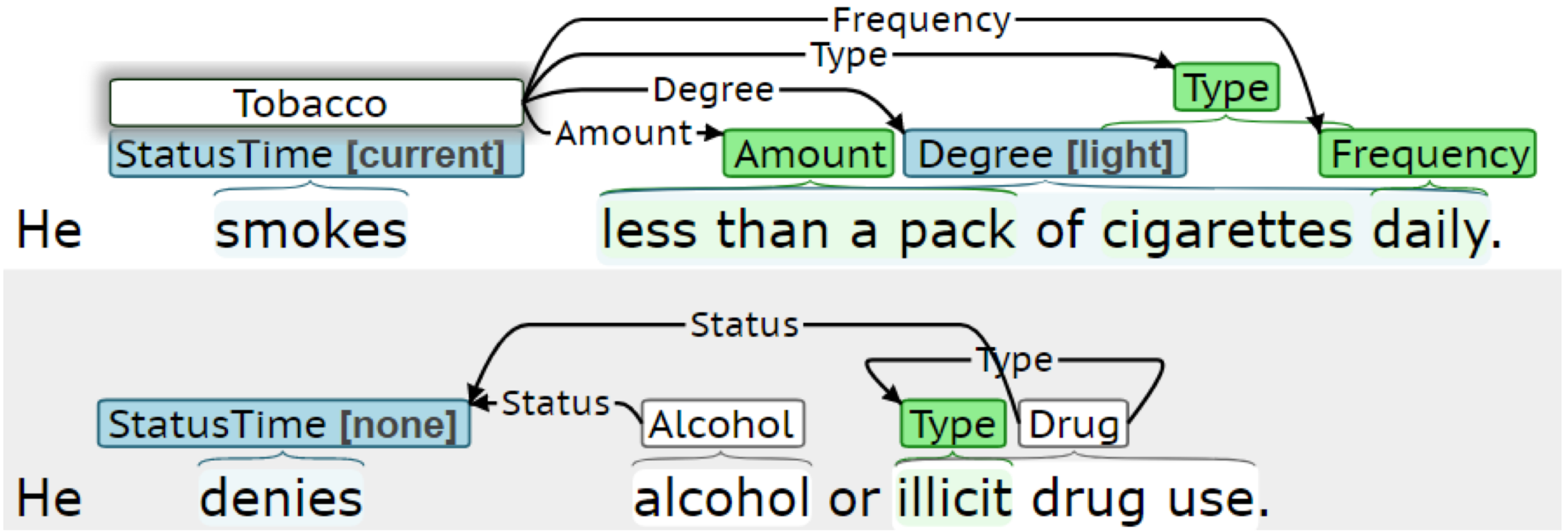
- Entities:
  - Trigger
  - Status (current, past)
  - Type (values: cisgender, transgender)

Country of Origin, Race

Examples
The patient is from Ukraine. He is originally from Venice, Italy…
Race: Caucasian. She is African American.

- Entities:
  - Trigger
  - Type

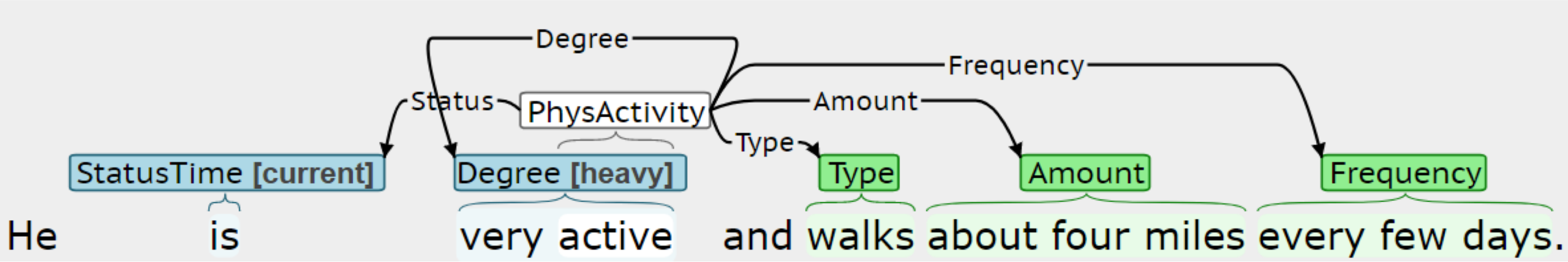
Substance abuse



- Entities:
  - Trigger
  - Status (current, past)
  - Duration

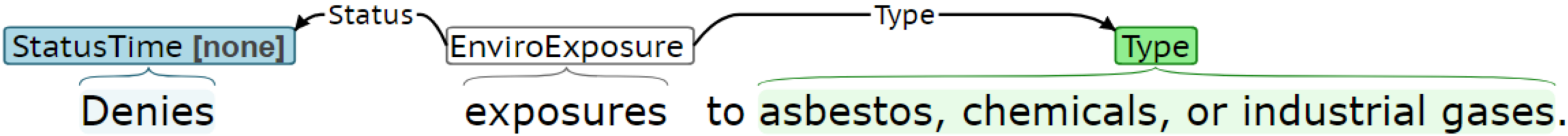
History
Type
Amount
Frequency
Degree (none, light, moderate, heavy)

Physical activity



- Entities:
  - Trigger
  - Status (current, past)
  - Duration
  - History
  - Type
  - Amount
  - Frequency
  - Degree (none, light, moderate, heavy)

Environmental exposures



- Entities:
  - Trigger
  - Status (current, past)
  - Duration
  - History
  - Type
  - Amount
  - Frequency

Summary - Annotated phenomena

Determinant	Entities
Employment	Trigger, Status, Duration, History, Type
Insurance*	Trigger, Status
Living status	Trigger, Status, Type, Duration, History
Sexual orientation*	Trigger, Status, Type
Gender identity*	Trigger, Status, Type
Country of origin*	Trigger, Type
Race*	Trigger, Type
Substance use (Alcohol, Drug, and Tobacco	Trigger, Status, Duration, History, Method, Type, Amount, Frequency, Degree
Physical activity	Trigger, Status, Duration, History, Type, Amount, Frequency, Degree
Environmental exposure*	Trigger, Status, Duration, History, Method, Type, Amount, Frequency

*estimated to have low frequency

Problems:
- Manual annotation is expensive
- Random sampling does not guarantee good representation for the annotated phenomena

Active learning

- Sample selection criteria:
    - *Informativeness*: reduce classification uncertainty
    - *Representativeness*: describe structure of labeled and/or unlabeled data
    - *Diversity*: describe variation in selected samples
- Want:
    - high entropy
    - high diversity
    - high representativeness

Corpus subsets

- Train
  - Initial training set (YVnotes)
  - Random sample for evaluation of Active Learning (AL)
    - Needs to be sufficiently large that incorporating results in meaningful performance improvement
  - Batches selected using Active learning query function in each iteration
- Dev
  - Representative of *test* dist. (random sample) – for tuning
- Test
  - Representative of *true* dist. (random sample) – for final evaluation

Active learning

- Maximum Batch Network Gain (MBNG) (Wu and Ostendorf, 2013)

$$NG(B) = \sum_{i \in B} H(i)(\sum_{j \in U-B} a_{ji} - \mu \sum_{k \in B, i \neq k} a_{ki})$$

$$H(i) = -\sum_{y} p(y|x_i;\theta) log(p(y|x_i;\theta))$$

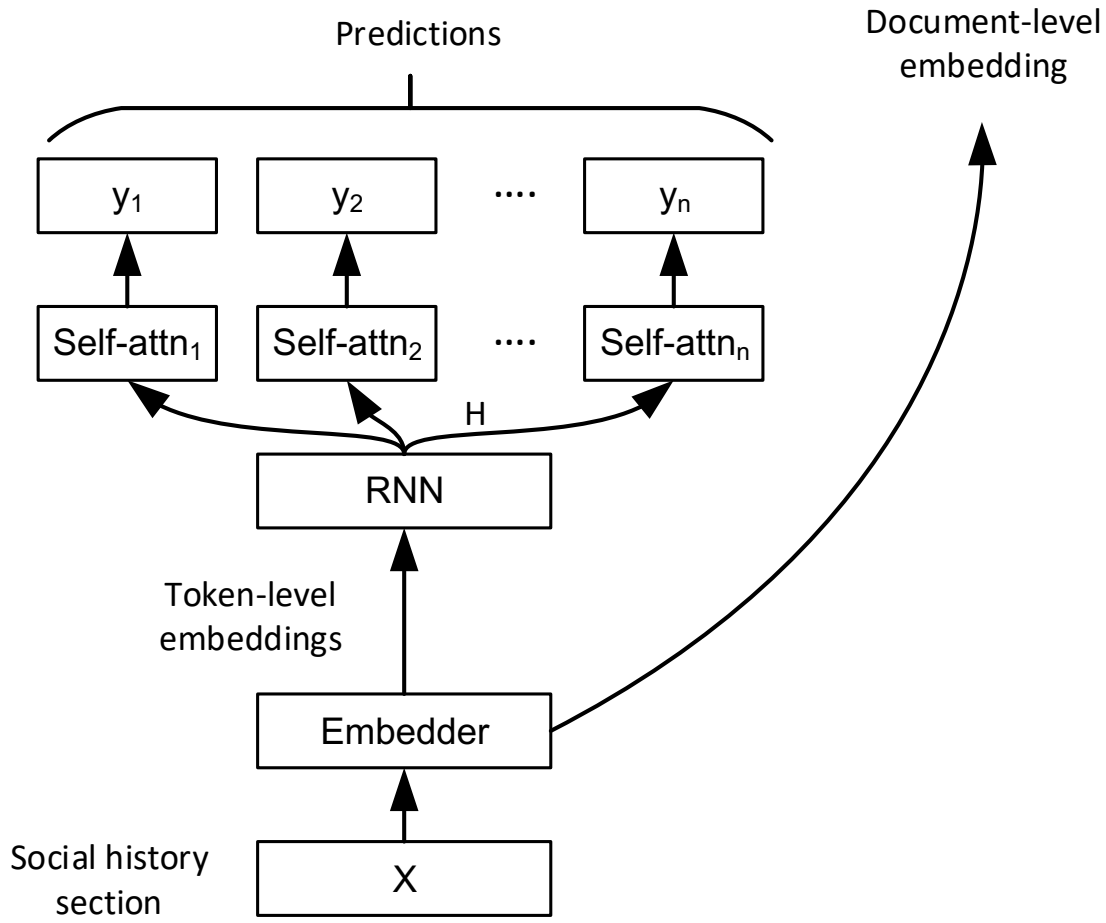
a_{ij}=cosine similarity between samples
U=unlabeled samples
B=batch of selected samples
μ=hyperparameter

- Simplified version
  - Focus on entropy and diversity
  - Avoid tuning hyperparameter, μ (for now)
  - Diversity as cosine dissimilarity

$$NG(B) = \sum_{i \in B} H(i) \sum_{k \in B, i \neq k} (1 - a_{ki})$$

Entropy and document representation

- Entropy
  - Select single entity for each determinant
  - Predict selected entities using self-attention
  - Operate on entire social history section
    - Avoid aggregating scores across sentences
  - Use embedder, like BERT or XLNet
  - Recurrent layer optional?
- Document representation
  - Embedder doc representation
    - More general, less biased
  - RNN end states
    - Will be biased by predicted phenomena



Selected entities by determinant

- Each social history section may have multiple events for a given determinant
- Treat each determinant entity as a multi-label problem
- Same self-attention network for each determinant entity (e.g. Employment Status)
- Separate, binary output classifiers for each label (e.g. employed, unemployed)
- Sum entropy across labels for given determinant entity

Determinant	**Entity**	**Label set**
Employment	Status	employed, unemployed,…
Insurance	Status	yes, no
Living status	Type	alone, with family,…
Sexual orientation	Type	homosexual, …
Gender identity	Type	cisgender, transgender
Country of origin	Trigger	present, not present
Race	Trigger	present, not present
Substance use	Status	none, current, past
Physical activity	Status	current, past
Enviro. exposure	Status	none, current, past

Selection approach

- For given batch, loop on:
  - Select 1 sample based on Employment Status
  - Select 1 sample based on Insurance Status
  - ⋮
  - Select 1 sample based on Environmental exposure
- All determinants given similar weight/priority
- Could prioritize certain determinants, if desired

Current annotation status

- Annotators: 4 medical students
- Training:
  - Annotators went through annotation training on 20 manually selected social history sections from MIMIC-III discharge summaries
  - 2 iterations with 2 meetings to resolve conflicts
- Annotation:
  - 50 randomly sampled social history sections from MIMIC-III discharge summaries
  - Inter-rater agreement levels: <<to be filled>>

Questions

Meliha Yetisgen

University of Washington

melihay@uw.edu

YV-Notes: http://depts.washington.edu/bionlp/index.html

Code and data: https://github.com/Lybarger/clinical_extraction

Email me for annotation guidelines!