



Annotation efforts: i2b2/UTHealth 2014 NLP shared tasks

Amber Stubbs, Ozlem Uzuner, Hua Xu

November 14, 2014



Corpus

- 1,304 new records from Partners HealthCare
 - 297 patients, 2-5 records per patient
- Cohort selection
 - All patients diagnosed with diabetes
 - 1/3 develop CAD over course of records (file #s 100-199)
 - 1/3 start with CAD (#s 200-299)
 - 1/3 never develop CAD (#s 300-400)



Data statistics

- Training data: 790 files
- Testing data: 514 files



Track 1:

De-identification



Participants

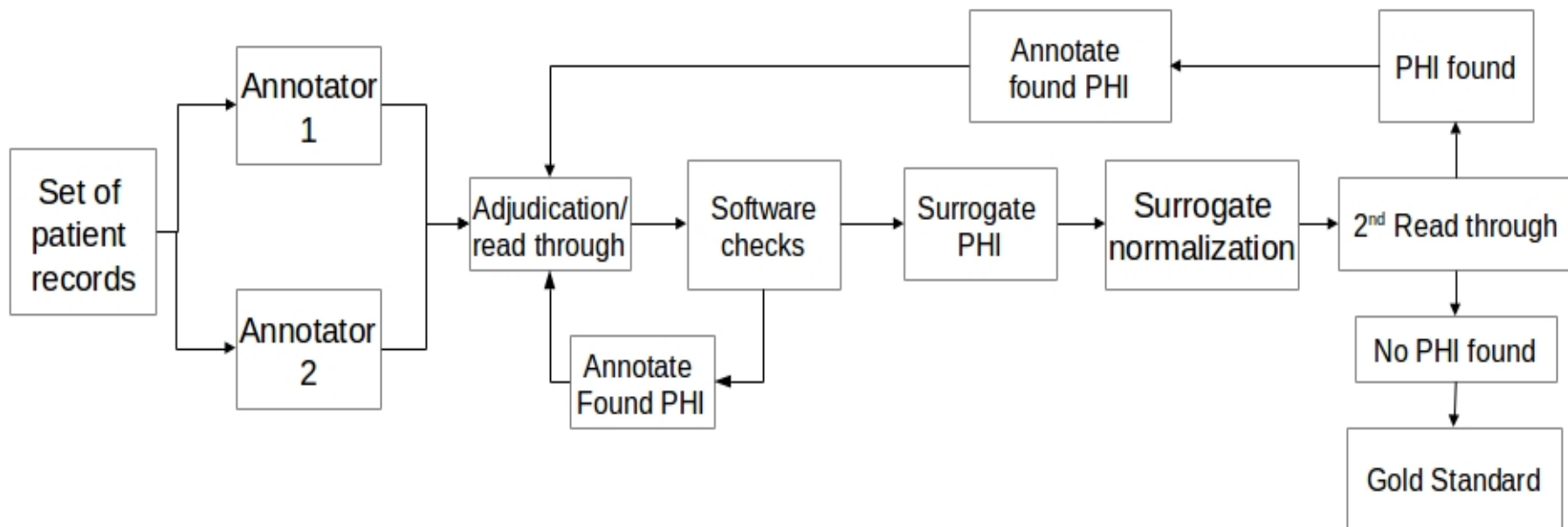
- Annotators:
 - 5 MIT undergraduates, 1 MIT senior researcher
- Adjudicator:
 - Shared task co-organizer (Amber)
- Readers:
 - Shared task co-organizer (Ozlem), 1 MD, 1 medical assistant



Annotation procedure

- Each file:
 - Double-annotated
 - Adjudicated
 - Software checks
 - Surrogate generation
 - Read-through for missed PHI

Annotation procedure





Annotation task

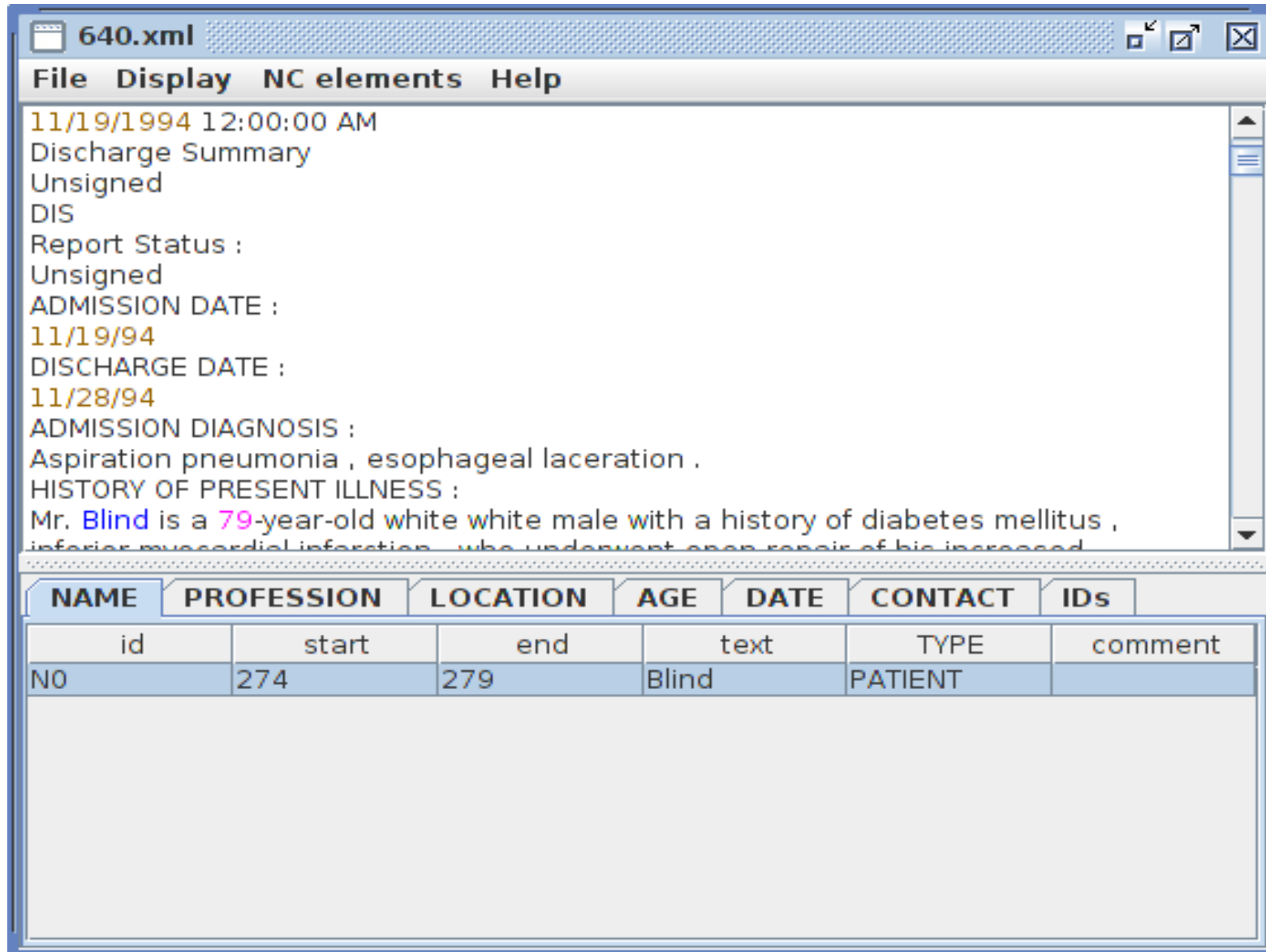
- Risk-averse interpretation of HIPAA PHI categories
 - Names, including doctor names and usernames
 - Locations, including hospitals, states, landmarks, countries, etc
 - Dates, including years and named holidays
 - Contact information
 - ID numbers
 - Other, including professions, all ages, any other information that could help identify a patient



Annotation guidelines

- We grouped the PHI into categories and sub-categories
 - NAME
 - PATIENT, DOCTOR, USERNAME
 - CONTACT
 - PHONE, FAX, EMAIL, URL
- “When in doubt, annotate”

Annotation tool: MAE



The screenshot displays the MAE (Medical Annotation Editor) interface. The top window, titled "640.xml", shows a document with the following text:

11/19/1994 12:00:00 AM
Discharge Summary
Unsigned
DIS
Report Status :
Unsigned
ADMISSION DATE :
11/19/94
DISCHARGE DATE :
11/28/94
ADMISSION DIAGNOSIS :
Aspiration pneumonia , esophageal laceration .
HISTORY OF PRESENT ILLNESS :
Mr. Blind is a 79-year-old white male with a history of diabetes mellitus ,
~~inferior myocardial infarction , who underwent open repair of his increased~~

Below the document view is a table with columns: NAME, PROFESSION, LOCATION, AGE, DATE, CONTACT, and IDs. The table contains one row of data:

NAME	PROFESSION	LOCATION	AGE	DATE	CONTACT	IDs
id	start	end	text	TYPE	comment	
N0	274	279	Blind	PATIENT		



Inter-annotator agreement compared to gold standard

Granularity	Precision	Recall	F-measure (F1)
Strict entity-based	0.903	0.886	0.893
Strict token-based	0.939	0.921	0.927

“Strict”: the PHI categories and sub-categories had to match precisely

Macro-level scores: each document's scores calculated, then average across the corpus

Data Statistics

PHI category	# in training data	# in test data	Total # in corpus
AGE	1233	764	1997
CONTACT (all)	323	218	541
DATE	7507	4980	12487
ID (all)	881	625	1506
LOCATION (all)	2767	1813	4580
NAME (all)	4465	2883	7348
PROFESSION	234	179	413
Total # of tags	17410	11462	28872
Average # PHI per file	22.03	22.3	22.14



Track 2: Risk Factors



Participants

- Seven annotators:
 - 5 registered nurses
 - 1 medical doctor
 - 1 medical assistant



Annotation Task

- Focus on disease indicators and broad temporal labels
- “Light” annotation:
 - Requires evidence from annotators
 - Non-exhaustive annotation



Heart Disease Risk Factors

- Diabetes: mention, high blood sugar or A1c measurements
- Hyperlipidemia/hypercholesterolemia: mention, high LDL or cholesterol measurements
- Hypertension: mention, high blood pressure
- CAD: mention, related medical event (MI, revascularization), test (positive stress test)
- Obesity: mention, high BMI, large waist circumference
- Family history of premature CAD
- Being a smoker
- Relevant medications



Temporality

- Each risk factor/disease combination annotated for presence with regard to document creation time (DCT)
 - Before, during, after
 - “Continuing” used as shorthand for all three
 - Exceptions: smoking and family history



“Light” Annotation

- Each risk factor/indicator/time combination only annotated once per document
- Task is faster and easier for annotators

Sample annotation

2014-06-14

Mr. Walsh complains of
lightheadedness and shortness of
breath.

Smoking:
Current

Hyperlipidemia:
Mention;
Continuing

Diabetes:
Medication;
Continuing

History: Hyperlipidemia (diagnosed
last year); smokes 2ppd since 25yo

Hyperlipidemia:
Medication;
Continuing

Hypertension:
High BP;
During

Medications: Glyburide, simvastatin

BP on admission: 160/100



Annotation procedure

- Each record annotated by 3 medical professionals
- Gold standard generated by voting:
 - All tags turned into document-level
 - “Continuing” tags expanded into 3 separate tags: before, during, after
 - Any tag appearing in 2/3 or more of annotations for each record was included in gold standard



Inter-annotator agreement compared to gold standard

	Macro	Micro
precision	0.957	0.958
recall	0.959	0.960
F-measure	0.958	0.959



Data Statistics – Overview

- “Mention” was the most frequent indicator of all risk factors
- Most medical records were missing information about the patient
 - Every patient in the corpus was diabetic, but only 880 files contained a “during” mention of diabetes
 - 614 files contained no information about smoking
- Tests (blood glucose, A1c, cholestrol) were rare

Data Statistics – Diabetes

Indicator	Before DCT			During DCT			After DCT		
	Train	Test	Total	Train	Test	Total	Train	Test	Total
Mention	518	354	872	524	356	880	518	355	873
High A1c	89	71	160	21	11	32	0	0	0
High glucose	16	18	34	9	15	24	0	0	0

CAD, hyperlipidemia, hypertension, and obesity all show similar distributions between mentions and specific tests

Data Statistics – Smoking

Smoking Status	# training	# testing	# total
Current smoker	58	33	91
Ever smoked	9	3	12
Never smoked	184	120	304
Past smoker (quit over 1 year ago)	149	113	262
Unknown (no mention of smoking)	371	243	614

Data Statistics – family history

	# Training	# Test	# Total
No mention of CAD-related family history	768	495	1263
Mention of CAD-related family history	22	19	41

Data statistics – Medications (partial)

Medication category	Before DCT			During DCT			After DCT		
	Train	Test	Total	Train	Test	Total	Train	Test	Total
ACE inhibitor	322	205	527	314	195	509	319	200	519
ACE inhibitor + diuretic	4	4	8	4	4	8	4	4	8
anti diabetes	1	0	1	1	0	1	1	0	1
ARB	95	59	154	90	57	147	94	57	151
ARB + diuretic	3	8	11	3	7	10	3	5	8
aspirin	424	263	687	435	273	708	424	262	686
Beta blocker	469	276	745	472	281	753	470	278	748



Questions?