



Natural Language Processing for Anonymization

John Aberdeen and Lynette Hirschman

781.271.2840

aberdeen@mitre.org

MSR

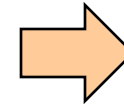
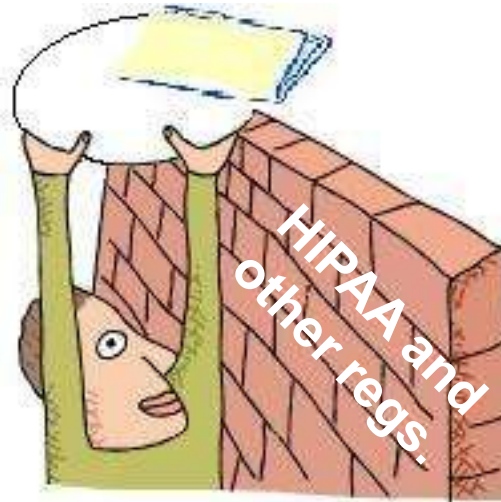
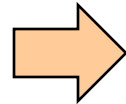
Problem

- **Privacy constraints prevent sharing of data containing personally identifiable information (PII) such as medical or surveillance records**
- **Private information is often shared by mistake—security breaches are all too common**
- **Can we use Natural Language Processing (NLP) techniques to create free-text de-identification technology that can be applied to multiple data types?**

Background

Records with personally identifiable information (PII) cannot be shared due to **privacy** constraints

Medical or
Surveillance
Records
containing PII



- **Medical studies**
- **Analyst reports**

- Medical record processing is **labor-intensive**—automated techniques have been slow to develop.
- Analysts must present reports with identities masked.

Objectives

- **Create a de-identification system for free text that can be rapidly tailored to multiple domains and record types**
 - How much data is needed to achieve high accuracy?
- **Embed the system in an interactive interface that allows end users to run the system, inspect the results, and correct errors**
 - Important to work inside partners' facilities with real records
- **Partner with real-world users of de-identification to tailor systems and develop reality-based metrics to evaluate systems**
 - Over-redaction vs. under-redaction: what is better for the end user?

Activities

- **Build a domain-portable de-identification system**
 - Identify real-world users of de-identification technology
 - Develop prototype system in collaboration with partners
- **Develop resynthesis techniques**
 - **Problem: Because of privacy considerations, we only have access to scrubbed data in most instances**
 - **To create usable training data, we must resynthesize PII so that the resulting data is realistic**
- **Develop cost-based utility metrics for different error types**
- **Evaluate system and publish results**

Highlight

- The task of identifying PII is similar to the well-understood Named Entity task—MITRE has much experience here
- We achieved early success in applying NLP techniques to create a medical record de-identification system that achieved the top score in the Fall 2006 AMIA De-identification Challenge

HISTORY OF PRESENT ILLNESS: The patient is a 77-year-old-woman with long standing hypertension who presented as a walk-in to me at the **[HOSPITAL XX XX XX XX]** on **[DATE - YY]**. Recently had been started q.o.d. on Clonidine since **[DATE - ZZ]** taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the **[HOSPITAL XX XX XX]** for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. **[DOCTOR XX]** to follow.

Demonstration



Task: Workflow: Replacer: **Control Interface**

Input: Document type:

Status:

Document

Automatically Tagged Document

Robert Harrison is a 4-year 5-month-old male who presented to Oak Valley Health Center on 10/11/2007 with cough of 10 days duration and fever. Patient lives in a densely populated section of [LOCATION]. Dr. Gunnarsson referred to [DOCTOR] for [PHONE] IMC RADIOLOGY. Scattered lung densities likely to represent either atelectasis or acute viral illness with no definite lobar pneumonia identified.

Hand Correction Popup

- Add PATIENT
- Add DOCTOR
- Add HOSPITAL
- Add ID
- Add PHONE
- Add LOCATION
- Add DATE
- Add AGE
- Cancel

Replacement

Save rich Save raw Legend

```
<doc id="97683176" type="RADIOLOGY_REPORT">
<text>
<text origin="CCHMC_RADIOLOGY" type="CLINICAL_HISTORY">
[PATIENT] is a 4-year 5-month-old male who presented to [HOSPITAL] on [DATE]/2007 with cough of 10 days duration and fever. Patient lives in a densely populated section of [LOCATION]. Dr. [DOCTOR] referred to radiology. Rule out pneumonia.</text>
<text origin="CCHMC_RADIOLOGY" type="IMPRESSION">
Scattered lung densities likely to represent either atelectasis or acute viral illness with no definite lobar pneumonia identified.</text>
</texts>
</doc>
```

Redacted Document

Content tags

XXXXX	AGE
XXXXX	DATE
XXXXX	DOCTOR
XXXXX	HOSPITAL
XXXXX	ID
XXXXX	LOCATION
XXXXX	PATIENT
XXXXX	PHONE

Structure tags

XXXXX	lex
XXXXX	untaggable
XXXXX	zone

Legend

PII	Type	Location	Redaction
Robert Harrison	PATIENT	118 - 133	[PATIENT]
Oak Valley Health Center	HOSPITAL	187 - 211	[HOSPITAL]
10/11	DATE	215 - 220	[DATE]

Candidate Replacements

Impacts



- **Breaking through a critical bottleneck in handling records with privacy constraints, such as medical records and surveillance reports**
- **Partnerships with key stakeholders in the medical community and in government agencies that use data containing PII**
- **Positions MITRE as a leader in developing, evaluating, integrating, and disseminating NLP technology for de-identification of free text records**

Future Plans

Next Milestone: Installation of system w/iterative training loop at partner institution

