About

Scrubber 3.0 Beta Released!

Scrubber **v3.0** now uses **Apache cTakes** to provide parallel concept extraction during de-idenification. **Apache cTAKES** graciously invited us to port the Scrubber de-identification pipeline to the Apache hosted codebase. The maintenance version of the 2.X will remain available. The publication describing this work has been accepted with minor revision, this site will be updated shortly to reflect the described **methods and results**.

McMurry* AJ, Fitch* B, Savova G, Kohane IS, Reis BY. “Improved de-identification of physician notes through integrative modeling of both identifying and non-identifying medical text”, BMC Medical Informatics and Decision Making Accepted minor revise Jan 2013.

Motivation

"Free Text" medical notes contain information which can be used to **locate human biospecimens** and even predict patient outcomes.
Because medical notes often contain Protected Health Information, it is necessary to "scrub" notes of sensitive information prior to sharing with a clinical investigator. Towards this goal, we have developed Open Source software that removes PHI from raw text, XML, or databases.
The software has been approved for use by numerous hospital IRBs, and has been manually reviewed by physician experts.

Challenge

Distinguishing pertinent clinical facts from sensitive patient identifiers in free text clinical narratives is a difficult classification task.
One reason is that variations in physician writing styles have limited how broadly NLP algorithms can be utilized in multi-site studies.
Another reason is that hospital IRBs have differing perspectives regarding "privacy risk to research benefit".
As a result, relatively few physician notes are used in research studies despite the wealth of available high quality clinical phenotypes .

Approach

The HMS Scrubber builds on years of community progress in de-identification and NLP.
In 2006, Beckwith developed and validated a rule based system to de-identify pathology reports.
This widely accessed de-id program performed well in the pathology setting and was approved by four IRBs at Harvard teaching hospitals.

Porting this software to other hospital settings and note types proved difficult and required fine-tuning the regular expressions for each installation.
This lead to the creation of the "3.X" Scrubber, combining autocoding and de-identification tasks to maximize research utility and minimize site specific customization.

This new approach using machine learning analyzes similaraties and differences betwen physician notes, medical dictionaries, and medical journal publications.