Use and Evaluation of the MIST Open-Source Deidentification Tool

David Carrell, Melanie Currier, Scott Halgrim
Group Health Research Institute (carrell.d@ghc.org)

GHRI-IT Poster Session, Nov. 2010

GroupHealth

Background

Clinical text used for research and quality assurance purposes must be deidentified before it can be shared with an external collaborator because it frequently contains sensitive patient identifying information. Deidentification can be a time-consuming and fatiguing process and commercial solutions can be very expensive to purchase and may require resource-intense customization.

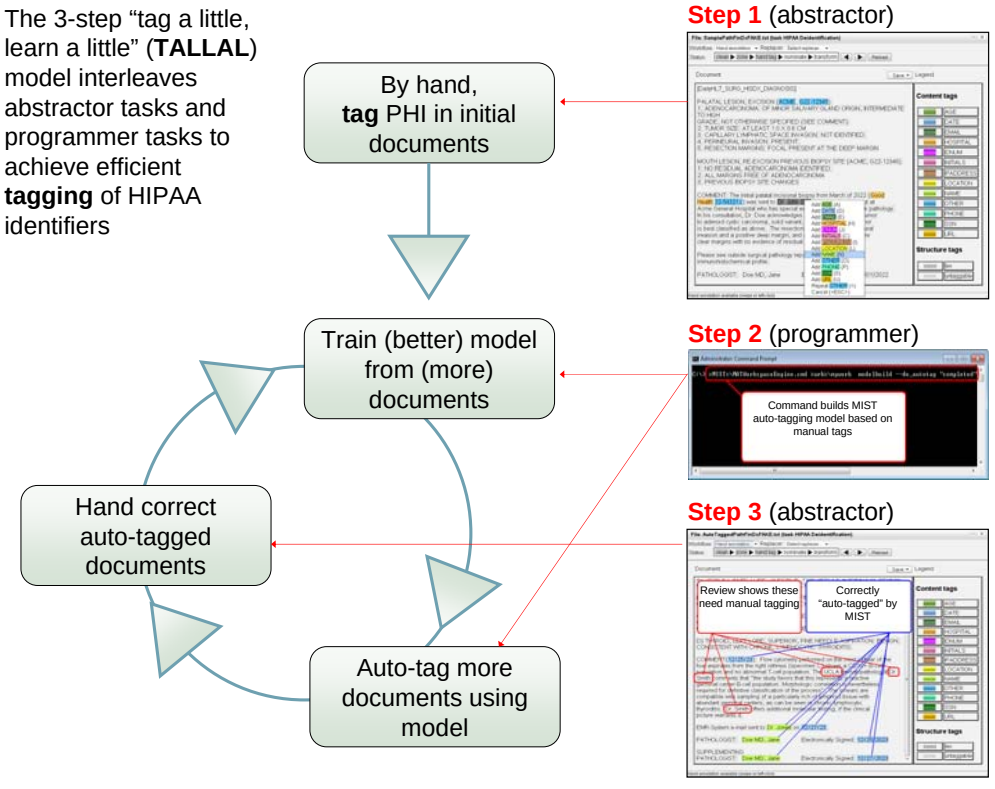
State-of-the-art open source deidentification software offers an inexpensive alternative to commercial products. We installed and evaluated the MITRE Corporation's open-source MITRE Identification Scrubber Toolkit (MIST) to determine its utility at GHRI and other applied research settings such as those within the HMO Research Network (HMORN). MIST can be customized for local use in the normal course of deidentification. Here we examine its ease of installation and use and the accuracy of its redactions.

Installation and Set Up

- MIST is available for use under a very generous (non-restrictive) open-source **licensing** agreement.
- **Installation** of MIST and its prerequisite software are well documented.
- **Using MIST** involves some programmer tasks and some annotator (Research Specialist) tasks.
  - For **programmers**, simple MIST "batch" commands are used to prepare document sets and "train" and apply models that automatically deidentify text.
  - For **annotators**, all interactions with the documents are managed by MIST's interface.

The TALLAL Model

The 3-step "tag a little, learn a little" (**TALLAL**) model interleaves abstractor tasks and programmer tasks to achieve efficient **tagging** of HIPAA identifiers



How MIST Works

MIST is built around a 3-step **TALLAL model** (see Figure):

- A programmer prepares a set of documents for processing (same as manual approach).
- **Step 1**: an abstractor uses the MIST interface to mark HIPAA identifiers in 5-20 (or so) documents (see Figure).
- **Step 2**: a programmer executes MIST commands to train a model and use that model to automatically deidentify the remaining documents.
- **Step 3**: an abstractor reviews/corrects MIST's automatic markings on another 5-20 documents (see Figure).
- **Steps 2 and 3 are repeated iteratively** until all documents have been manually reviewed or MIST performs well enough to accept its automatic results.
- Actual **redaction** (replacing identifiers with innocuous placeholders such as "[NAME]") is performed by the programmer using MIST bach commands.

Evaluation

The MIST software is easily obtained and installed by Group Health (and HMORN) programmers—no expertise required.

Using MIST is straightforward for programmers and annotators.

Preliminary experiments indicate manual deidentification with MIST is faster and easier to use than traditional approaches.

Additional evaluation of time savings, manual performance, and automated model performance would be beneficial.

We recommend that MIST become the standard GHRI method for deidentifying clinical text. Even where 100% manual review is needed, the MIST interface should be used. Furthermore, MIST auto-tagging should be used to reduce human abstractor burden.