Ensuring Adequate Coverage to Build a Publicly Available Corpus of De-identified Clinical Documents

Brett R. South, MS1,2,3, Danielle Mowery, MS4,5, Ying Suo, MS1,2,3, Mingyuang Zhang, BS3, Liqin Wang, MS3, Eric Levy, BS5, Edna Shenvi, MD5, Shazia Ashfaq, MBBS, MBA5, Stephane M. Meystre, MD, PhD1,4, Wendy W. Chapman, PhD4,5
1VA Salt Lake City Health Care System, Departments of *Internal Medicine and *Biomedical Informatics, University of Utah, Salt Lake City, 2University of Pittsburgh, Pittsburgh, 3University of California San Diego, San Diego

Abstract

One remaining problem for Natural Language Processing (NLP) in the clinical domain is the availability of publicly shareable document corpora that have been downloaded 18 types of protected health information (PHI) specified under HIPAA. We test the generalizability of an annotation task targeted at identifying PHI found in a publicly available document corpus called MTSamples (www.mtsamples.com). These tasks use an annotation schema and guidelines developed by the Consortium for Healthcare Informatics Research (CHIR) and an annotation tool called the extensible Human Oracle Suite of Tools (eHOST). We report prevalence estimates for PHI elements, averaged recall, precision and plateaued F1-measure and coverage when judges are incrementally added.

Objectives

• Assess the generalizability of an annotation schema and guidelines for identification of PHI elements.
• Evaluate the effects of incrementally adding judges on the coverage of PHI annotation.

Methods

• Seven reviewers manually annotated a random sample of 350 clinical documents from the MTSamples corpus (2,330 unique documents).
• Reviewers used annotation schema and guidelines created by the VA Consortium for Healthcare Informatics Research (CHIR) to identify all classes of PHI specified under HIPAA.
• Annotators were independently reviewed by all annotators for both training (100 documents) and closed annotation phases (250 documents).
• A final reference standard was created after a consensus review of the resulting annotations.
• We report recall (TP/(TP+FN)) and precision (TP/(TP+FP)) of individual reviewer’s results, as well as unions of logical combinations of numbers of reviewers (ranging from 1 to 7).

Figure 1. PHI Annotation Tasks

Figure 2. Annotator Performance Metrics

Annotator Performance

Number of Annotators

Recall

Precision

F1-measure

Results

• The final reference standard contained 2,302 unique annotations representing 17 of the 20 classes defined in our annotation schema.
• Prevalence was highest for information representing PHI of medium sensitivity (53%), and lowest for highly sensitive PHI representing proper names of persons (14%). Across the 350 documents annotated there were no mentions of information representing social security numbers.
• A total number of 129 relationship pairs were present representing any mention of proper names of persons.
• Document level F1-measure for highly sensitive PHI ranged from 0.20-1.00 (mean=0.89, std=0.17), and low sensitivity 0.077-1.00 (mean=0.81, std=0.22).
• Recall ranged from 0.66 (1 reviewer) plateauing at a high of 0.92 (7 reviewers). Alternatively, precision decreased from 0.82 (1 reviewer) to a low of 0.61 for the union of all seven judges.
• F1-measures ranged from 0.73 (1 reviewer), 0.79 (2), 0.78 (3), 0.77 (4), 0.75 (5), 0.74 (6), and 0.73 (7 reviewers) (Figure 2).

Conclusion/Recommendations

• Adequately redacting PHI is a legal and ethical first step when using clinical data for research purposes.
• These efforts provide an excellent source of pilot data to determine where adding more reviewers results in diminishing returns and for what types of information (particular PHI classes) using more efficient methods (i.e. distributed annotation) will improve reference standard quality.
• It is likely that for some types of PHI multiple reviewers are required to achieve optimal performance thus requiring these manual review tasks to employ techniques that go beyond double annotation.
• Balancing the risk of missed PHI with the cost of adding additional reviewers is important considerations.
• Future efforts will include making the resulting corpus available to the broader clinical NLP community via iDASH.

References:


Contact information: brett.south@va.gov, VA Consortium for Healthcare Informatics Research 500 Foothill Drive, Salt Lake City, UT 84146, (801) 499-1175