Measuring Risk and Information Preservation: Toward New Metrics for De-identification of Clinical Texts

Lynette Hirschman & John Aberdeen
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730
{lynette,aberdeen}@mitre.org

Abstract

Current metrics for de-identification are based on information extraction metrics, and do not address the real-world questions "how good are current systems", and "how good do they need to be". Metrics are needed that quantify both the risk of re-identification and information preservation. We review the challenges in de-identifying clinical texts and the current metrics for assessing clinical de-identification systems. We then introduce three areas to explore that can lead to metrics that quantify re-identification risk and information preservation.

1 Introduction

Our current metrics do not address the questions "how good are current free-text de-identification systems", and "how good do they need to be?" We need measures that quantify *risk of re-identification* based on type and amount of personal health identifier (PHI) leakage (PHI terms not redacted in the de-identification process), and measures that quantify information preservation or readability.

The metrics in current use were developed originally for *entity extraction* (the correct labeling of types of phrases in free text, such as person name, date, organization). Entity extraction performance is typically measured in terms of precision, recall and balanced f-measure at both the token (word) and phrase level. The top de-identification systems (Szarvas, Farkas, & Busa-Fekete, 2007; Wellner, et al., 2007) performed well on these measures, as reported at the first i2b2 De-identification challenge evaluation (Uzuner, Luo, & Szolovits, 2007), achieving accuracies of over 97% token-level f-measure, with recall (sensitivity) of over 95%. Over the past several years, these results have been extended to more record types and record formats; for example, (Friedlin & McDonald, 2008) reported that their MeDS system successfully removed 99.5% of HIPAA-specified identifiers from HL7 records. These are encouraging numbers, but recall and precision do not tell us how good a de-identification system needs to be for a particular intended use.

2 Challenges in Clinical Text

Removing PHI from unstructured text poses new challenges: in contrast to structured information (e.g., fields in a table or a database), we do not know in advance where PHI will appear in a free text record, and we do not know what kinds of PHI will occur. This problem is made more challenging for medical records because the types of record vary greatly in content and in amount of PHI – for example, a lab report will likely contain very little PHI, while a social work note will be likely to contain much more.

Medical records have internal structure that is dependent on the medical record system and the medical record type; there is typically a mix of structured fields (e.g., for patient identifier, patient name, doctor name), along with unstructured fields for free text. This means that de-identification of records must handle a combination of structured and unstructured information. The excerpt below shows two free text fields (CLINICAL HISTORY and IMPRESSION) from a (fictitious) radiology

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 72–75,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

report, with several types of PHI, including dates, locations and ages (shown in **bold**).

> RADIOLOGY REPORT
>
> CLINICAL HISTORY: Patient is a **4-year 5-month** old male who presented to **Oak Valley Health Center** on **10/11/2007** with a cough of 10 days duration and fever. Patient lives in a densely populated section of **Knoxville**. Rule out pneumonia.
>
> IMPRESSION: Scattered lung densities likely to represent either scattered atelectasis or acute viral illness with no definite lobar pneumonia identified.

This example illustrates how PHI is distributed in the free text portions of a medical record. Removing PHI from these free text portions requires application of techniques from natural language processing that are capable of identifying phrases of specific types based on the lexical content (the words that make up the phrases) and the surrounding words.

3 Current Methods and Metrics

Fortunately, the problem of identifying types of information in free text is a well-studied problem in the natural language processing community. We can leverage several decades of research on *information extraction* and the *named entity identification* problem in particular, including multiple community evaluations such as the Message Understanding Conferences (MUC) (Grishman & Sundheim, 1996) and the subsequent Automated Content Extraction (ACE) evaluations[1] – both focused on extraction from newswire -- as well as evaluations of biomedical entity extraction from the published literature e.g., in the BioCreative evaluations (Krallinger, et al., 2008). In addition, starting in 2006, there have been a series of evaluations for clinical natural language processing, with data sets of clinical records provided by the i2b2 consortium (Uzuner, et al., 2007). It has been critically important to have corpora of medical records, because medical records represent a very different style of text compared to news articles or journal articles. Medical records are characterized by their formulaic and telegraphic style, that is, the use of phrases or incomplete sentences rather than fluent prose, along with heavy use of abbreviations and domain-specific terminology (e.g., "93 yo w NVD"). The systems developed for newswire or for journal articles must be explicitly adapted (or *trained*) to handle the categories required for de-identification as well as the telegraphic language of medical records.

De-identification of free text medical records consists of two steps: recognition and redaction. The phrase recognition stage corresponds to the *named entity recognition* problem mentioned above, namely the ability to identify a sequence of words in running text that constitutes the mention of an entity of a specified type – such as the phrase **Oak Valley Health Center** in the example above. For newswire, types of named entities include person, organization, location, time, date, and money; for biomedical tasks, entities have included genes, proteins, drugs, diseases, etc. For de-identification, the critical elements are the 18 types of protected health information identified by HIPAA,[2] including names, dates, locations, zip codes, phone numbers, social security numbers, ages ninety and above, URLs and other identifying information. Interestingly, most institutions have developed their own set of protected classes of information, e.g., some institutions distinguish between DOCTOR and PATIENT identifiers, which both fall into the more general HIPAA category of NAME.

The techniques used to recognize named entities in text include:

- Lexically-based approaches that rely on matching words (or phrases) against the words or phrases contained in a lexicon;
- Pattern based approaches that are particularly useful for HIPAA-relevant PHI such as telephone numbers, social security numbers, dates, etc.
- Machine learning approaches that are based on statistical models of word sequences. These approaches require *training exemplars* that are used to associate sequences of words with probabilities of types of phrase, e.g., the word(s) following "Dr." or "DR" will likely be a doctor's name.

[1] http://www.itl.nist.gov/iad/mig/tests/ace

[2] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996).

All three approaches have been used and often combined (Beckwith, Mahaadevan, Balis, & Kuo, 2006; Berman, 2003; Friedlin & McDonald, 2008; Gupta, Saul, & Gilbertson, 2004; Morrison, Li, Lai, & Hripcsak, 2009; Szarvas, et al., 2007; Uzuner, Sibanda, Luo, & Szolovits, 2008; Wellner, et al., 2007) to provide high quality recognition of PHI. The 2006 i2b2 challenge evaluation for automatic de-identification of free text clinical records provided an opportunity for groups to benchmark their automated de-identification systems against a carefully prepared gold standard corpus of medical discharge summaries. The top systems performed well with scores of over 0.97 token-level f-measure and recall (sensitivity) of over 0.95 (Uzuner, et al., 2007).

4 Toward New Metrics

The Uzuner et al. (2007) paper concludes with two important (and as yet unanswered) questions (p. 562):

1. Does success on this challenge problem extrapolate to similar performance on other, untested data sets?

2. Can health policy makers rely on this level of performance to permit automated or semi-automated disclosure of health data for research purposes without undue risk to patients?

We have been particularly concerned with the second question, because it will be very difficult to release automatically de-identified data until we can provide an answer. The metrics used to date have been measures of technology performance, but they do not address the key issues of risk of PHI exposure and readability/preservation of information in the de-identified record.

Recall errors are clearly correlated with risk of PHI exposure, but not all recall errors lead to PHI exposure. For example, the name "Washington, George" might be mistakenly redacted to "LOCATION, NAME" leading to both a recall and a precision error for the word "Washington" but no PHI exposure. Also, some kinds of PHI exposure errors contain much more information (e.g., a patient's last name) than others (a first name; or a telephone extension where the telephone number has been redacted). Friedlin and McDonald (2008) report that MeDS did not miss any full patient identifiers, but it did miss an average of 2.13 patient identifier fragments per report. However, they concluded that none of these fragments were true patient identifiers.

Similarly, precision errors cause mislabeling of results and are correlated with loss of readability. In the extreme case, a system that redacts all words would achieve perfect recall, but very low precision and no information content. A system that replaces real PHI with synthetic (fictitious) PHI might be more resistant to re-identification because it would be difficult for an attacker to distinguish real from fictitious information.

We need new measures that quantify risk of re-identification based on type and amount of PHI leakage (PHI terms or parts of terms not redacted in the de-identification process); and we need measures that quantify information preservation or readability.

We plan to explore three areas that may yield more useful metrics for de-identification. The first is to quantify the re-identification risk through a detailed analysis of PHI in different record types. Given a set of records and a de-identification system, we can generate quantitative data on PHI distribution in different record types, rate of exposure of different classes of PHI (e.g., names vs. locations vs. phone numbers), and likelihood of combinations of exposed PHI. We can distinguish between partial exposure of PHI (e.g., just a first name or just a room number), and combinations of such exposures within a single record (room number and institution provides much more identifying information than room number alone). Using this information, we can develop analyses of risk using methods developed for structured data (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007; Malin, 2007; Sweeney, 2002) by combining statistics from de-identified records with publicly available information (census data, voter registration, etc).

The second area to explore is how to measure information preservation or readability. One approach would be to apply one or more available medical information extraction systems such as the Mayo Clinic cTAKES system (Savova, Kipper-Schuler, Buntrock, & Chute, 2008) to compare information correctly extracted from de-identified data vs. original data. This would provide a reasonable proxy for measuring information loss due to de-identification. Alternatively, Friedlin and McDonald (2008) developed a measure of *interpretability* in their de-identification experiments, defined as preserving test type and test results (for lab reports) or type of report, specimen and conclusion (for pathology reports).

A third area to explore is protection by *hiding in plain sight*. We can determine the reduction in risk from applying resynthesis (Yeniterzi, et al., 2010) to de-identified data, which would have the effect of hiding exposed PHI in plain sight – since such elements would be interspersed with fictitious but realistic looking identifiers (particularly names) inserted as replacements of PHI.

5 Conclusion

Current metrics for de-identification have their origins in information extraction; they neither adequately assess the risk of re-identification, nor do they provide a good measure of information preservation. We plan to address these shortcomings by 1) applying risk analysis methods derived for structured data, 2) using medical extraction systems to assess information preservation, and 3) exploring *hiding in plain sight* protection by using resynthesis to replace identifiers with false by realistic fillers. Once we have alternative measures for risk of re-identification and information preservation, we can also explore the correlation of precision and recall to these new measures. Accurately quantifying and balancing risk of re-identification and information preservation will enable health policy makers to make better decisions about the use of automated de-identification, and sharing of clinical data for research.

References

Beckwith, B. A., Mahaadevan, R., Balis, U. J., & Kuo, F. (2006). Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Med Inform Decis Mak, 6*, 12.

Berman, J. (2003). Concept-Match Medical Data Scrubbing. *Arch Pathol Lab Med, 127*, 680-686.

Friedlin, F. J., & McDonald, C. J. (2008). A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc, 15*(5), 601-610.

Grishman, R., & Sundheim, B. (1996). *Message Understanding Conference - 6: A Brief History.* Paper presented at the 16th International Conference on Computational Linguistics, Copenhagen.

Gupta, D., Saul, M., & Gilbertson, J. (2004). Evaluation of a deidentification (DE-ID) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol, 121*(2), 176-186.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., Hirschman, L., & Valencia, A. (2008). Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol, 9 Suppl 2*, S1.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD), 1*(1), 3.

Malin, B. (2007). A computational model to protect patient data from location-based re-identification. *Artif Intell Med, 40*(3), 223-239.

Morrison, F. P., Li, L., Lai, A. M., & Hripcsak, G. (2009). Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *J Am Med Inform Assoc, 16*(1), 37-39.

Savova, G., Kipper-Schuler, K., Buntrock, J., & Chute, C. (2008). UIMA-based Clinical Information Extraction System. *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP*, 39.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 10*(5), 557-570.

Szarvas, G., Farkas, R., & Busa-Fekete, R. (2007). State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc, 14*(5), 574-580.

Uzuner, Ö., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc, 14*(5), 550-563.

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A de-identifier for medical discharge summaries. *Artif Intell Med, 42*(1), 13-35.

Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., & Hirschman, L. (2007). Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc, 14*(5), 564-573.

Yeniterzi, R., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Hirschman, L., & Malin, B. (2010). Effects of Personal Identifier Resynthesis on Clinical Text De-identification. *J Am Med Inform Assoc, 17*(2), 159-168.