Role-Preserving Redaction of Medical Records to Enable Ontology-Driven Processing

Seth Polsley, Atif Tahir, Muppala Raju, Akintayo Akinleye, Duane Steward
Texas A&M Health Science Center
College of Medicine Biomedical Informatics Research group
College Station, Texas
spolsley,atif.tahir,mnpr84,akinleyeakintayo,dsteward@tamu.edu

Abstract

Electronic medical records (EMR) have largely replaced hand-written patient files in healthcare. The growing pool of EMR data presents a significant resource in medical research, but the U.S. Health Insurance Portability and Accountability Act (HIPAA) mandates redacting medical records before performing any analysis on the same. This process complicates obtaining medical data and can remove much useful information from the record. As part of a larger project involving ontology-driven medical processing, we employ a method of recognizing protected health information (PHI) that maps to ontological terms. We then use the relationships defined in the ontology to redact medical texts so that roles and semantics of terms are retained without compromising anonymity. The method is evaluated by clinical experts on several hundred medical documents, achieving up to a 98.8% f-score, and has already shown promise for retaining semantic information in later processing.

1 Introduction

Medical health records data has immense potential for research in furthering the field of automated healthcare. Unfortunately, one of the challenges facing medical informatics is the dissemination and sharing of digital records for research and analysis due to strict regulations regarding patient confidentiality. Protecting protected health information (PHI) is a critical responsibility of health care providers, with the U.S. Health Insurance Portability and Accountability Act (HIPAA) outlining a number of principles. Removing PHI can also mean removing critical parts of a record, so building redaction techniques that preserve as much information about the original data as possible while still retaining anonymity is an important pre-processing step.

In this work, we discuss a redaction framework for removing PHI from medical records through de-identification. One of the primary goals of this framework is to preserve valuable information like roles, semantics, and time intervals as much as possible. Because this forms the pre-processing stage of future text processing, we elected to model roles according to a formal ontology; this maintains relationships and enables straightforward detection of ontological terms in later phases.

2 Background

Knowledge buried in medical text is valuable, but due to federal law protecting sensitive data, it must be de-identified for distribution. Most existing methods rely on rule-based systems that match patterns and dictionaries of expressions that frequently contain PHI. Sweeny's Scrub tool uses templates and a context window to replace PHI (Sweeney, 1996). Datafly, also by Sweeny, offers user-specific profiles, including a list of preferred fields to be scrubbed (Sweeney, 1997). Thomas developed a method that uses a lexicon of 1.8 million names to identify people along with "Clinical and Common Usage" words from the Unified Medical Language System (UMLS) (Thomas et al., 2002). Miller developed a de-identification system for cleaning proper names from records of indexed surgical pathology reports at the Johns Hopkins Hospital (Miller et al., 2001). Proper names were identified from available lists of persons, places and institutions, or by their proximity to keywords, such as "Dr." or "hospital." The Perl

Proceedings of the BioNLP 2017 workshop, pages 194–199,
Vancouver, Canada, August 4, 2017. ©2017 Association for Computational Linguistics

tool *Deid* is a recent development which combines several of these rule-based and lexical approaches with some additional capabilities like better handling of time (Neamatullah et al., 2008).

While identifying PHI for removal or anonymization remains an open challenge, simply redacting texts overlooks one of the more fundamental aspects of recent biomedical informatics, which has incorporated a focus on ontology-driven development (Mortensen et al., 2012; Ye et al., 2009; Tao et al., 2013; Sari et al., 2013; Omran et al., 2009; Lumsden et al., 2011; Pathak et al., 2009). In a domain like healthcare – where information is dense, diverse, and specialized – an ontology allows representing knowledge in a usable manner, because it describes a framework for clearly defining known terms and their relationships (Hakimpour and Geppert, 2005; Lee et al., 2006; Pieterse and Kourie, 2014; Strohmaier et al., 2013; Kapoor and Sharma, 2010). Once the data has been formally described via an ontology, new applications become apparent. To provide several examples, simply by formalizing electronic records as an ontology, researchers have shared better ways to represent patient care profiles (Riaño et al., 2012), perform risk assessment (Draghici and Draghici, 2000), evaluate elderly care (Hsieh et al., 2015), and more (Rector et al., 2009; Rajamani et al., 2014). Perhaps the greatest promise lies in ontology-driven computational models, where the structure of an ontology makes the data accessible to programmatic operations, and there have been several applications to the problem of automated diagnosis (Bertaud-Gounot et al., 2012; Haug et al., 2013; Hoogendoorn et al., 2016).

Some of these ontology-driven techniques do consider redaction as it relates to the ontology. Of particular note is the extensive work by South et al. in identifying the exact types of PHI present throughout the medical record according to risk (South et al., 2014). Dernoncourt applied recurrent neural networks to the task of identifying PHI by type to remove the need for large dictionaries on the i2b2 dataset (Dernoncourt et al., 2016). In the future, we hope to share a more direct contrast between our role-labeling and South et al.'s, but our goals remain distinct from either South et al. or Dernoncourt. Because our ontology centers around the medical encounter, we must leverage the EMR's dynamic list of patients, caregivers, and providers to ensure roles are preserved according to their specific encounter. In this way, our work is more similar to Douglass' MIMIC dataset, which uses a patient list to assure role (Douglass et al., 2004).

3 Methods

The core reasoning for our methodology is that knowing the role of a redacted name can be vital, and since we will be processing patient records at the encounter-level, tying specific roles to single encounters is necessary. For instance, was a condition reported by the caregiver or by the clinician and at what time? That is just a single question illustrating the potential for confusion when names are redacted without roles or ordering, yet, there is no need to blindly attempt to extract roles from free text. Nearly every EMR maintains structured data like a patient's name, family contact, and attending physician. By leveraging this knowledge, pseudonyms can be constructed that remove confusion regarding roles in the final text.

To formally support role-preservation, we begin by defining a very simple ontology to relate key roles and terms. Patients are treated by clinicians and observed by caregivers. Treatments (or interventions) are given on the basis of a medical encounter, and, depending on the outcome, may lead to more medical encounters or the end of the record of care. This is a very basic means of modeling roles in medical texts, but it supports cross-domain redaction that preserves much of the semantics and relationships after the anonymization stage.

The redaction pipeline operates on data in two stages to support better identification of roles in the text. First, the structured data is used to extract whatever knowledge is available, typically roles like doctors and patients, to perform knowledge-based redaction. Second, the unstructured text undergoes entity recognition to clean missed terms. While this approach requires some insight about the data beforehand, it is a logical means of ensuring we can remove all PHI without damaging roles and relationships.

3.1 Structured

3.1.1 Patient-Centric Role Preservation

Our system initially builds a dictionary of known individuals in each role. A person can have any number of names of any length but all of them are

Table 1: Sample dictionary of names

Patients	Caregivers	Providers
Original		
Ira Jones	Michael Jones	Daniel Moore
	Barbara Davis	Mary Johnson
Redacted		
$Clark$	$Clark_{CAREGIVER1}$	$Clark_{PROVIDER1}$
	$Clark_{CAREGIVER2}$	$Clark_{PROVIDER2}$

drawn directly from the fields in the EMR. In accordance with the ontology, patients will be identified first as the subject of care, a unique field in most systems. Depending on the domain, there will be a personal doctor, an attending physician, or some other clinician name given in a separate field. Caregivers may be drawn from locations like billing or family contacts. For this part, knowledge of the data structure is necessary, but once the source fields are identified, they will be consistent across the other records.

Once the dictionary of names and roles is built, patients are assigned a pseudonym randomly from a list of non-matching family names to provide anonymity and linked to the pseudonym in the dictionary. Subsequently, all individuals associated with that patient are assigned a derivative pseudonym denoting their role. Consider the example shown in Table 1. For this small dictionary of a single patient, we see more than one caregiver and provider listed. The system first replaces the patient's name, Patricia Jones, with a false name, Clark. This identifier then becomes the basis for all subsequent individuals with a connection to the patient.

After the dictionary has been constructed, the system knows all the original names and their new pseudonyms. The medical texts are scanned for any occurrence of any known name, ignoring case or modifiers like possessive forms. Full names will be on file, but given names and family names may appear separately in the record. Regular expressions are used to match variants of names while enforcing order.

3.1.2 Date Offsets

It is worth emphasizing the importance of dates in medical record data. One can simply remove or replace dates to redact PHI, as with names, but just like names, we wished to preserve more information in support of the ontology. In particular, intervals between encounters or patient ages under 89 are compliant with HIPAA and useful for tasks like association mining. A common solution is to use offsets for dates because the original date will be erased from the document without losing intervals. However, an unconstrained random offset still loses information. For instance, intervals given in the free text will be broken if a day of the week is mentioned and then a date given. Our system ensures intervals are undamaged by constraining date offsets in week-long intervals. Thus, even if the dates are moved by years, there's no loss in day-granular intervals.

The date offset is applied across all records of a single patient uniformly to maintain interval and continuity of encounters. Furthermore, the system is very flexible about handling dates in free text, using as much knowledge as possible to piece together correct, redacted dates. For example, a snippet of a medical note may read: ”A surgery was performed in 2005 to correct the issue; on March 4, the patient...” Because the redaction system makes use of the structured fields, it would extract the date of entry for this medical note. Assuming that date is *March 7, 2006*, the system will move forward labeling unspecified years as *2006*, giving a means of differentiating the vague dates *2005* and *March 7*.

3.2 Unstructured

The second pass of de-identification also operates over free text, but it does not make use of known information such as the dictionary of names or the dates of an entry. Instead, general attributes of potential PHI are used to locate and remove sensitive data. Email addresses, phone numbers, mailing addresses, and medical case numbers are located through common regular expressions. ZIP codes are retained because they are not considered PHI and can be useful for location-based operations.

Unknown entities appear frequently in the text due to other names of people or places being written that are not listed in the dictionary of names. To account for these entities, Stanford's *CoreNLP* is used to detect any remaining entities in the text which do not belong to a linked pseudonym (Manning et al., 2014). All entities are redacted according to their determined type, e.g. $NAME1$ for a person or $LOCATION1$ for a place. Even in the unstructured phase, sequential naming schemes ensure unknown people and places do not become confounded with any other entities.

3.3 Complete Pipeline

By the time the pipeline has finished, the text has been run through two rounds of de-identification. First, any useful knowledge is pulled from the data in the EMR to build a dictionary for rule-based redaction that preserves roles. Second, operating without any knowledge, a set of regular expressions and more sophisticated entity recognition methods are employed to clear other sensitive data without adding ambiguity or destroying valuable non-PHI information. The inclusion of *CoreNLP* in the final part supports more advanced language models than simply using rules and regular expressions. This allows the complete pipeline to capture almost any potential PHI while still recognizing known entities, particularly those relevant to the ontology, or types of entities, such as contact numbers of locations.

4 Evaluation

We worked with data sets from two different domains – veterinary and hospice care. Fortunately, due to the cross-domain design of our ontology, there was little difficulty in identifying fields that mapped to elements of the ontology. Upon defining this mapping, huge portions of text from both domains were pushed through the full pipeline. The resulting text included ontological terms and other marked regions, e.g. ZIP codes, while removing as little other information as possible.

Ideally, the final medical texts appear identical to the original files with only the PHI removed. To evaluate this, a team of clinical experts reviewed hundreds of documents, marking missed PHI or text that was unnecessarily redacted in each. From the veterinarian domain, where we studied complete discharge summaries (DS), two medical doctors reviewed 122 cases. From the hospice domain, which operated on shorter clinical notes (CN), the same experts reviewed 500 notes. To provide a simple baseline for comparison, we also tested a single rule-based approach for matching patient names against a data set of 15 documents.

As we see in Table 2, the system performed very well at correctly identifying PHI and non-PHI, especially in contrast with the patient-names baseline. In the discharge summaries, the majority of false negatives were due to previously-unnamed doctors who were neither in the dictionary nor detected during entity recognition. Only one misspelling of a patient name was detected.

Table 2: Word-level metrics for baseline (BL), discharge summaries (DS), and clinical notes (CN)

Count	**BL**	**DS**	**CN**
False Negatives	498	76	4
False Positives	0	5	250
True Positives	63	3391	1655
True Negatives	17191	75694	50460

Table 3: Performance of baseline (BL), discharge summaries (DS), and clinical notes (CN).

Metric	**BL**	**DS**	**CN**
Specificity	100%	99.9%	99.5%
Sensitivity	11.2%	97.8%	99.8%
Precision	100%	99.9%	86.9%
F-Score	20.2%	98.8%	92.9%

In the clinical notes, there were a great deal more false positives. Because the final step incorporates *CoreNLP*, certain texts will include many entities that are not PHI. Table 3 shows that specificity, sensitivity/recall, and precision are high for both, although the precision for clinical notes suffers due to the many false positives. While the baseline achieves high precision by matching only patient names, the lower sensitivity and f-score demonstrate the high number of PHI belonging to other categories that the full system captures.

5 Conclusion and Ongoing Work

Medical records can provide a wealth of information for data scientists but due to their sensitive nature, are often limited in availability. Effective, reliable redaction is the best known solution to the problem, but most techniques will lose exact details like encounter-level roles. In this work, we integrate knowledge and model-based approaches to augment redaction. In future works, we seek to share some of the benefits we have seen using roles to create better semantic clusters and word models than achieved through only pseudonyms. We hope that such de-identification pipelines, highly cognizant of the original data structure, will encourage a future of richer and more capable ontology-driven analysis.

Acknowledgments

We wish to thank Girish Kasiviswanathan, Pratik Mrinal, Amber Schulze, Preston Baker, Ruihong Huang, and Tracy Hammond for their support.

References

Valérie Bertaud-Gounot, Régis Duvauferrier, and Anita Burgun. 2012. Ontology and medical diagnosis. *Informatics for Health and Social Care* 37(2):51–61.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* page ocw156.

M Douglass, GD Clifford, Andrew Reisner, GB Moody, and RG Mark. 2004. Computer-assisted de-identification of free text in the mimic ii database. In *Computers in Cardiology, 2004*. IEEE, pages 341–344.

Anca Draghici and George Draghici. 2000. Cross-disciplinary approach for the risk assessment ontology design. *Information Resources Management Journal (IRMJ)* 26(1):37–53.

Farshad Hakimpour and Andreas Geppert. 2005. Resolution of semantic heterogeneity in database schema integration using formal ontologies. *Information Technology and Management* 6(1):97–122.

Peter J Haug, Jeffrey P Ferraro, John Holmen, Xinzi Wu, Kumar Mynam, Matthew Ebert, Nathan Dean, and Jason Jones. 2013. An ontology-driven, diagnostic modeling system. *Journal of the American Medical Informatics Association* 20(e1):e102–e110.

Mark Hoogendoorn, Peter Szolovits, Leon MG Moons, and Mattijs E Numans. 2016. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artificial intelligence in medicine* 69:53–61.

Nan-Chen Hsieh, Rui-Dong Chiang, and Wen-Pin Hung. 2015. Ontology based integration of residential care of the elderly system in long-term care institutions. *Journal of Advances in Information Technology* 6(3).

Bhaskar Kapoor and Savita Sharma. 2010. A comparative study ontology building tools for semantic web applications. *International Journal of Web & Semantic Technology (IJWesT)* 1(3):1–13.

Yugyung Lee, Kaustubh Supekar, and James Geller. 2006. Ontology integration: Experience with medical terminologies. *Computers in Biology and Medicine* 36(7):893–919.

Jim Lumsden, Hazel Hall, and Peter Cruickshank. 2011. Ontology definition and construction, and epistemological adequacy for systems interoperability: A practitioner analysis. *Journal of Information Science* 37(3):246–253.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60.

R Miller, JK Boitnott, and GW Moore. 2001. Web-based free-text query system for surgical pathology reports with automatic case deidentification. *Arch Pathol Lab Med* 125:1011.

Jonathan Mortensen, Matthew Horridge, Mark A Musen, and Natalya Fridman Noy. 2012. Applications of ontology design patterns in biomedical ontologies. In *AMIA*.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making* 8(1):32.

Esraa Omran, Albert Bokma, Shereef Abu Al-Maati, and David Nelson. 2009. Implementation of a chain ontology based approach in the health care sector. *Journal of Digital Information Management* 7(5).

Jyotishman Pathak, Harold R Solbrig, James D Buntrock, Thomas M Johnson, and Christopher G Chute. 2009. Lexgrid: a framework for representing, storing, and querying biomedical terminologies from simple to sublime. *Journal of the American Medical Informatics Association* 16(3):305–315.

Vreda Pieterse and Derrick G Kourie. 2014. Lists, taxonomies, lattices, thesauri and ontologies: Paving a pathway through a terminological jungle. *Knowledge Organization* 41(3).

Sripriya Rajamani, Elizabeth S Chen, Mari E Akre, Yan Wang, and Genevieve B Melton. 2014. Assessing the adequacy of the hl7/loinc document ontology role axis. *Journal of the American Medical Informatics Association* pages amiajnl–2014.

Alan L Rector, Rahil Qamar, and Tom Marley. 2009. Binding ontologies and coding systems to electronic health records and messages. *Applied Ontology* 4(1):51–69.

David Riaño, Francis Real, Joan Albert López-Vallverdú, Fabio Campana, Sara Ercolani, Patrizia Mecocci, Roberta Annicchiarico, and Carlo Caltagirone. 2012. An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. *Journal of biomedical informatics* 45(3):429–446.

Anny Kartika Sari, Wenny Rahayu, and Mehul Bhatt. 2013. An approach for sub-ontology evolution in a distributed health care enterprise. *Information Systems* 38(5):727–744.

Brett R South, Danielle Mowery, Ying Suo, Jianwei Leng, Óscar Ferrández, Stephane M Meystre, and Wendy W Chapman. 2014. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. *Journal of biomedical informatics* 50:162–172.

Markus Strohmaier, Simon Walk, Jan Pöschko, Daniel Lamprecht, Tania Tudorache, Csongor Nyulas, Mark A Musen, and Natalya F Noy. 2013. How ontologies are made: Studying the hidden social dynamics behind collaborative ontology engineering projects. *Web Semantics: Science, Services and Agents on the World Wide Web* 20:18–34.

Latanya Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association, page 333.

Latanya Sweeney. 1997. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, page 51.

Cui Tao, Guoqian Jiang, Thomas A Oniki, Robert R Freimuth, Qian Zhu, Deepak Sharma, Jyotishman Pathak, Stanley M Huff, and Christopher G Chute. 2013. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. *Journal of the American Medical Informatics Association* 20(3):554–562.

Sean M Thomas, Burke Mamlin, Gunther Schadow, and Clement McDonald. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association, page 777.

Yan Ye, Zhibin Jiang, Xiaodi Diao, Dong Yang, and Gang Du. 2009. An ontology-based hierarchical semantic modeling approach to clinical pathway workflows. *Computers in biology and medicine* 39(8):722–732.