You Don't Know My Name: Transfer Learning to De-Identify Protected Health Information in Electronic Health Records

Arnobio Morelix, Pauline Wang
School of Information, University of California, Berkeley

Date: August 2, 2019

Abstract
This paper presents an information extraction system using Bidirectional Encoder Representations from Transformers (BERT) to de-identify protected health information (PHI) from electronic health records (EHR). Past work associated with PHI have used a combination of dictionary-based, rule-based, and machine learning algorithms to deal with the inherent complexity in PHI categories. In this paper we use BERT, a pre-trained model with context-aware word embeddings, to classify PHI categories in a named entity recognition task. Our model performs in line with, and in some measures better than, models relying on extensive rules-based pre-processing. Because of the cost of data pre-processing in rules-based systems, and their reliance on specific styles of annotation (e.g., how a particular hospital might record and report on biometrics), we believe the type of model we present here has more generalization potential and we present further opportunities for refinement.

Keywords: BERT, transfer learning, HIPAA, named entity recognition, i2b2 2014 de-identification challenge.

1. Introduction

As more hospitals and healthcare systems adopt electronic health records (EHR), EHR data presents new opportunities and challenges for machine learning applications. From clinical research to medical practice, EHR contains rich information that helped generate insights to improve patient care, identify health risks, and support public policy initiatives. [1]

Despite its importance, one major obstacle in using EHR is the privacy concerns on disclosing Protected Health Information (PHI). In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) currently requires 18 direct identifiers to be removed or protected prior to the record be made available in a more public way. [2] The task of de-identifying PHI is time-consuming, and existing software can be limiting in its generality. [3]

In this paper we discuss and apply the use of a transfer learning system to conduct PHI recognition in health records. Specifically, we use Bidirectional Encoder Representations from Transformers (BERT), a natural language processing model achieving many state-of-the-art results, published by Google in October 2018. [4] It achieves these results by taking into account the contextual information in a sentence, from both left and right (i.e., the text preceding and following a particular word).

In addition to the specific demands for de-identification in health records, de-identification is a broader problem with increasing importance in other realms. Data privacy regulations like General Data Protection Regulation (GDPR) in the European Union protects data that can be identifiable to an individual. [5] Regulatory frameworks like this raises costs associated with holding (and sharing in any

way) personally identifiable information. We believe this means there is an opportunity for creating better de-identification tools. We believe the success in using BERT in extracting sensitive information can have wider applications in privacy protection and information sharing. In the healthcare domain specifically, better de-identification can open the doors to otherwise inaccessible medical records that can help generate insights and improve patient care.

2. Overview

2.1 Dataset

In this paper we use data containing unstructured clinical narratives, provided by Informatics for Integrating Biology and the Bedside (i2b2). The dataset contains 1304 longitudinal medical records among 296 diabetic patients from various hospitals in the U.S., in which 790 are used for training and the remaining 514 are for testing. [6] Our development data set is a section of the 514 testing files.

The corpus was annotated by trained professionals according to guidelines provided by i2b2. The PHI categories were grouped into 7 main categories with 24 sub-categories. The annotation is recorded in XML-format with tags indicating the start and end positions of the text (e.g. start=23, end=33), the text itself (e.g. text= "John Smith"), and the category (e.g. NAME) and subcategory (e.g. "TYPE=PATIENT"). Appending Table 1 summarizes the category, types and frequency of the PHI occurrences.

2.2 Literature Review and Past work

Several studies have been conducted in this topic using a combination of rule-based, dictionary based, and machine learning algorithms. Grouin utilized a conditional random field (CRF) process and rule-based post-process treatment to achieve a micro F-score of .8055. [7] Similarly, Liu and Chen used a token-level and character-level CRF combined with rule-based classifiers to achieve a micro F-score of 0.91. [8] Guillen combined rules derived from decision trees and lexical and syntactic cues to determine a potential PHI (micro F-score=0.6223). [9] Dehghan and Kovacevic combined rule-based method with Stanford Name Entity Recognition (NER) model to achieve micro F-score 0.91. [10]. Table 2 below outlines the different model approaches and their outcomes.

The highest performance model (bolded in Table 2) we have been able to identify came from Yang Hui and Jonathan Garibaldi of University of Nottingham. The authors implemented sophisticated feature engineering based on syntactic and surface-oriented rules that characterize the semantics of PHI terms. They then used a machine learning algorithm, CRF with a BIO (beginning, inside, outside) scheme tags to generate a classifier for the PHI category. The model achieved remarkable success though it lags in PHI types with fewer training data, such as ORGANIZATION, COUNTRY, LOCATION-OTHER. We hope that through BERT's transfer learning we will be able to improve in areas that these models have difficulty in identifying, and especially improve performance in more generalizable way.

Table 1. Past results on PHI de-identification using i2B2 dataset

Author	Model	Precision	Recall	F1
Chen, Tao [11]	HMM	0.87	0.71	0.78
	HMM-DP	0.86	0.80	0.83
Grouin, Cyril [12]	CRF+ Rule-based (using lexicon of entities in training corpus)	0.89	0.73	0.81
	CRF	0.91	0.69	0.73
	CRF +Rule-based (no lexicon of entities)	0.81	0.79	0.81
He, Bin, Chenc, Jianyi et al. [13]	WI-deID (CRF based)	0.95	0.87	0.91
Guillem, Rocio [14]	Decision Tree	0.90	0.43	0.62
Liu, Zengjian, Tang Buzhou et al [15]	CRF (token and character-level)+ Rule-based	0.93	0.90	0.91
Yang, Hui and Garibaldi, Jonathan [16]	CRF	**0.96**	**0.91**	**0.94**
Torii, Manabu, Fan Jung-Wei et al [17]	Customized MIST	0.87	0.74	0.80
	NER	0.92	0.24	0.38
	Merged	0.87	0.77	0.82
Dehghan, Azad, Kovacevic, Aleksandar et al [18]	Selected Rule-Based and CRF	0.93	0.88	0.91

Important note: each paper publishes slightly different methodology for results so comparisons are not exactly apples to apples. In the interest of giving a sense of the benchmarks, we provide those values here, but they should not be taken as direct comparisons.

2.3 Challenges with the Electronic Health Record (EHR) Data

2.3.1 Lexical Variation

The fact that most EHR are created by nurses or doctors in a free text format creates a myriad of terminological variation and irregularities. For example, a 25 year old female can be recorded as '25yoF', 25y.o.f', '25yo female','25yof', or other variations. We also see irregularities in PHI types that we presumed would have standard format, such as DATE, PHONE, and ZIP. This creates issues with tokenization and regular expressions as there are no consistent rules in separating words from abbreviations or short hands. This creates challenges for any rule-based approach. In Grouin's paper, for instance, more than 70 types of regular expressions were created to process only a certain number of PHIs.

2.3.2 Data Ambiguity without Context

Resolving ambiguity is another major challenge in de-identifying PHIs. Several biometric data can be easily mistaken as zip codes, IDs or phone numbers. For instance, '11/23' can be regarded as either a DATE or a medical test value. In fact, the data ambiguity is difficult to discern even for human

annotators. Despite using trained medical professionals who were provided with clear annotation guidelines, we identified several instances where the accuracy of annotation is questionable. This demonstrates an important and universal challenge in user-generated content, and particularly in healthcare record in which many data is generated by hand.

2.3.3 Uneven Distribution in PHI Types and 'Non-Entity' Observations

As displayed in Table 1, the frequency of the PHI types varies greatly in our training and test dataset. DATE, the most common type, has 7495 occurrences while BIOMETRIC ID only shows up once in our training set. ACCOUNT ID, VEHICLE ID, LICENSE NUMBER and IP ADDRESS do not have a single count. For our modeling purposes, this does not present an issue as those four PHI types do not appear in test set either. However, the granularity and specification of HIPPA protected data makes the de-identification more challenging than a typical Name Entity Recognition (NER) task as the availability of training data is limited.

Perhaps more relevant, nearly 94% of the original tokens in the corpus are "outside" (non-entity) tokens, making the training difficult for identifying entities.

2.4 Data Processing

The majority of the data pre-processing centers on converting the raw XML file to data format that can be fed into BERT. We separate the visit summary to individual sentences and encode each word with its identified PHI type. For example,"Date:09-23-2013" is annotated as DATE. We then use NLTK tokenizer to break down each word and encode each token accordingly. For example, "Date","":","09-02-2013" will be encoded "DATE","DATE","DATE" . We then use the BERT tokenizer to separate individual word to BERT tokens. We checked labels and sentence lengths to make sure any tokens generated by BERT were appropriately labeled via matches with the annotation files.

The fact that the data start as completely unstructured hospital notes (as opposed to more typical competition NER datasets) means that we had to spend a meaningful amount of time processing the data.

3 Model

Based on the challenges listed above, we believe that a BERT-based model will have strong performance, as it would not suffer as much from lexical variation and complexity in PHI types as rules-based models, and it is able to learn from the context surrounding each word. We created four main models, with several iterations in hyperparameters among them. We focus here on the main ones.

3.1 Baseline Model

In our baseline scenario, we simply want to see if the model can identify whether or not an item belongs to a PHI category or not. We incorporate a binary encoding system, in which all non-PHI tokens, including paddings, are encoded as class 0, while PHI tokens are encoded as class 1.

Our model generates a very high accuracy score, 99.1%, with 5 epochs. However, nearly 94% of original tokens are non-PHI and, if we include padding, that number jumps up to 96.4% (implying an accuracy of 96.4% if . Therefore, even if the model only guesses class 0, it will have a 96.4% accuracy rate. This is an

important discovery and for the subsequent models, we remove paddings from the calculation of accuracy.

3.2 Models A and B, Multi-Class and Parameter Tuning

In these models we add all 24 classes observable PHI types, and test different specifications in number of BERT fine tuning layers (to test how much domain-specific context the model can learn), epochs (to test how long it takes for mode loss to stabilize), and learning rates (to get us out of scenarios we found the model growing in loss after epochs of improvement). While the models achieve very high accuracy, it is quickly clear that the model rarely guesses less common classes. With 93.9% of original tokens in the data as "non-entities," it's tempting for the model to guess that category.

3.3 Model C, Multi-Class with Distribution Rebalance

In model see we reduce the number of original categories from 24 to 10, re-classifying the less common PHI types in the same category. We also remove from the data every sentence that does not have an entity listed in it, reducing our training sample by about 80%, from 73722 to 13398 sentences.

3.4 Results

Reducing the number of training sentences without entities listed in them results in good performance, in line with results relying on extensive rules-based pre-processing. This is encouraging because if we can train models without much pre-processing, it is likely it will be more generalizable to other contexts (e.g., hospitals which have different norms for writing notes).

Table 2. Model results

Model	Precision	Recall	F1	Accuracy
Baseline (Binary)	0.90	0.20	0.30	0.99
Model A (Multi-Class Model)	0.67	0.76	0.71	0.92
Model B (Multi-Class w. Tuning)	1.00	0.93	0.97	0.93
Model C (Rebalanced Multi-Class)	0.89	0.86	0.87	0.85

3.5 Error Analysis

We can classify the errors in our model in 3 main categories. a) False positives entity: confusing an "outside" token as an entity. b) Confusing one PHI category for another (e.g., thinking a patient name is actually a doctor's name. c) False negative entities: confusing an entity for a "outside" token. Appendix Table 2 includes a complete confusion matrix.

By far, the most prevalent type of error has been false negative entities, as the model is still very likely to guess that a token is "outside," the most common token type, despite the fact we have reduced the representation of that kind of token in the data. This suggests the primary way we might be able to improve model performance is to further reduce the representation of outside tokens in the data.

4 Considerations for future models

While we were experimenting with many hyperparameter fine-tuning options (e.g., sentence length, number of BERT fine-tuning layers), we identified some particular opportunities for improving performance on this type of model in the future which we have yet to attempt. These are listed below.

4.1 Combine sentences (more context)

For every model we trained we cut the sentence length to 20 tokens, a reasonable cut off as about 20% of sentences are longer than that. We did this after experimenting with longer sentence lengths, and realizing that too much shorter sentence lengths might lead us to lose some of BERT's key advantages.

As health notes written during the course of treatment, sentences in this corpus are short by design. But each document refers to the same patient and same visit, and different sentences in the same document might still contain mutually relevant context for the task at hand. Combining multiple sentences from the document in a single sentence to feed as example to BERT might lead to better performance and use of context by the model.

4.2 Custom loss functions to penalize false negatives more

When we think about the business case, false negatives (failing to identify PHI tokens) are more costly than false positives (mistakenly thinking an "outside" token is PHI). This is because the unintentional disclosure of sensitive information can be very damaging, while omitting non-identifiable information from a record by mistake carries a trivial downside.

The loss function we used, however, as most loss functions, treats all misclassifications equally. Changing this could benefit model performance in a very relevant way. In addition, continuing experimenting with a binary model.

4.3 Addressing uneven label distributions

Over 93% of tokens in the database are not PHIs, and most label types have very little representation in the data. To further improve the model we would expand on the work we did for model C and attempt different ways of taking into this unevenness in labels into account.

5. Conclusion

Our BERT-based model results are 0.89 precision, 0.86 recall, and 0.87 F1 score, in line with and surpassing some past work relying on extensive pre-processing with rules. We believe this shows the potential of BERT and other transfer learning models for this type of task that can be generalizable in a broader way to more contexts.

Acknowledgements

We would like to thank the i2b2 organizers for providing such an invaluable clinical dataset and this research opportunity. We would also like to thank professor Joachim Rahmfeld and teaching assistant Sudha Subramanian for their guidance and many late night office hours.

References

[1] Fernandes L, O'Connor M, Weaver V. "Big data, bigger outcomes: Healthcare is embracing the big data movement, hoping to revolutionize HIM by distilling vast collection of data for specific analysis". J AHIMA 2012;83(10):38–43< http://library.ahima.org/doc?oid=105683#.XQPtd4hKimA>
[2] HIPPA Guideline Materials. HHS.gov <
https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/index.html>
[3] Hripcsak G, Albers DJ. "Next-generation phenotyping of electronic health records". J Am Med Inform Assoc 2013;20(1):117–21 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3555337/>
[4] https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en
[5] https://arxiv.org/abs/1810.04805
[6] Informatics for Integrating Biology & the Bedside homepage
<https://www.i2b2.org/NLP/DataSets/Main.php>
[7] Cyril Grouin. "LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.A De-Identification of Unseen Clinical Texts. Workshop Challenges in Natural Language Processing for Clinical Data", Nov 2016, Chicago, United States. ffhal-01831223f.
<https://hal.archives-ouvertes.fr/hal-01831223/document>
[8] Liu, Zhengjian et al. "Automatic De-identification of Electronic Medical Records using Token-level and Character-lvel Conditional Random Fields" . June 2015.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4988843/>
[9] Guillen, Rocio. California State University San Marcos. "An Approach to De-identifying Electronic Medical Records". 2014.
[10] Dehghan and Kovacevic. "Combining knowledge- and data-driven methods for re-identification of clinical narratives" December 2015. https://www.ncbi.nlm.nih.gov/pubmed/26210359
[11] Tao Chen. "Hidden Markov Model using Dirchilet Process for De-Identification." Primary Healthcare Research Unit, Memorial University of Newfoundland, Canada
[12] Cyril Grouin. "LIMSI at CEGS N-GRID 2016 NLP Shared-Tasks: Track 1.A De-Identification of Unseen Clinical Texts. Workshop Challenges in Natural Language Processing for Clinical Data", Nov 2016, Chicago, United States. ffhal-01831223f.
<https://hal.archives-ouvertes.fr/hal-01831223/document>
[13] Bin He, Jianyi Cheng, Yi Guan, Keting Cen, Wenlan Hua. "A CRF-based Approach to De-identification in Medical Records." Harbin Institute of Technology, Harbin, China
[14] Guillen, Rocio. California State University San Marcos. "An Approach to De-identifying Electronic Medical Records". 2014.
[15] Zengjian Liu, Buzhou Tang, Qingcai Chen, Xiaolong Wang, Haodi Li. "De-identification of electronic medical records – HITSZ's system for track 1 of the 2014 i2b2 NLP challenge." Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[16] Hui Yang, Jonathan Garibaldi. "A Hybrid System for Automatic De-identification in Patient Discharge Summaries." School of Computer Science, University of Nottingham.
[17] Manabu Torii, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T. Wiley, Daniel Zisook, Yang Huang. "De-Identification and Risk Factor Detection in Medical Records." Kaiser Permanente Southern California, San Diego, CA

[18] Dehghan and Kovacevic. “Combining knowledge- and data-driven methods for re-identification of clinical narratives” December 2015. https://www.ncbi.nlm.nih.gov/pubmed/26210359

Appendix

Appendix Table 1. PHI types by incidence in data

PHI Category	PHI Type	Test	Train	Test (%)	Train(%)
Age	Age	764	1233	6.7%	7.1%
Contact	Phone	215	309	1.9%	1.8%
	Fax	2	8	0.0%	0.0%
	Email	1	4	0.0%	0.0%
	URL	0	2	0.0%	0.0%
	IPAddress	0	0	0.0%	0.0%
Date	Date	4980	7495	43.4%	43.1%
IDs	Medical Record Number	422	611	3.7%	3.5%
	ID Number	195	261	1.7%	1.5%
	Device ID	8	7	0.1%	0.0%
	Healthcare Plan Number	0	1	0.0%	0.0%
	Account Number	0	0	0.0%	0.0%
	License Number	0	0	0.0%	0.0%
	Vehicle ID	0	0	0.0%	0.0%
	Biometric ID	0	1	0.0%	0.0%
Location	Hospital	875	1437	7.6%	8.3%
	City	260	394	2.3%	2.3%
	State	190	314	1.7%	1.8%
	Zip	140	212	1.2%	1.2%
	Street	136	216	1.2%	1.2%
	Country	117	66	1.0%	0.4%
	Organization	82	124	0.7%	0.7%
	Location-Other	13	4	0.1%	0.0%
Name	Doctor	1912	2877	16.7%	16.5%
	Patient	879	1315	7.7%	7.6%
	Username	92	264	0.8%	1.5%
Profession	Profession	179	234	1.6%	1.3%
Total		**11462**	**17389**	**100.0%**	**100.0%**

Appendix Table 2. Confusion Matrix

	O	DATE	DOCTOR	HOSP.	PATIENT	AGE	OTHER PHI	CLS	SEPARAT.	PADDING	*True Label*
O	5596	11	12	9	0	0	7	3	13	20	5671
DATE	219	693	0	0	0	0	1	25	0	165	1103
DOCTOR	89	0	332	0	0	0	0	0	0	0	421
HOSP.	109	0	0	52	1	0	0	0	0	0	162
PATIENT	95	0	43	0	30	0	0	21	0	0	189
AGE	23	2	0	0	0	33	0	0	0	0	58
OTHER PHI	246	0	4	6	0	0	177	0	0	0	433
CLS	0	0	0	0	0	0	0	0	0	0	0
SEPARAT.	0	0	0	0	0	0	0	0	0	0	0
PADDING	0	0	0	0	0	0	0	0	0	0	0
Predicted Label	*6377*	*706*	*391*	*67*	*31*	*33*	*185*	*49*	*13*	*185*	