Generalizability and Comparison of Automatic Clinical Text De-Identification Methods and Resources

Óscar Ferrández, PhD[1,2], Brett R. South, MS[1,2], Shuying Shen, MStat[1,2], F. Jeff Friedlin, DO[3], Matthew H. Samore, MD[1,2], Stéphane M. Meystre, MD, PhD[1,2]
[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT;
[2]IDEAS Center SLC VA Healthcare System, Salt Lake City, UT;
[3]Medical Informatics, Regenstrief Institute, Indianapolis, IN

ABSTRACT

In this paper, we present an evaluation of the hybrid best-of-breed automated VHA (Veteran's Health Administration) clinical text de-identification system, nicknamed BoB, developed within the VHA Consortium for Healthcare Informatics Research. We also evaluate two available machine learning-based text de-identifications systems: MIST and HIDE. Two different clinical corpora were used for this evaluation: a manually annotated VHA corpus, and the 2006 i2b2 de-identification challenge corpus. These experiments focus on the generalizability and portability of the classification models across different document sources. BoB demonstrated good recall (92.6%), satisfactorily prioritizing patient privacy, and also achieved competitive precision (83.6%) for preserving subsequent document interpretability. MIST and HIDE reached very competitive results, in most cases with high precision (92.6% and 93.6%), although recall was sometimes lower than desired for the most sensitive PHI categories.

INTRODUCTION

With increased use and adoption of Electronic Health Record (EHR) systems, greater amounts of readily accessible patient data are available for use by clinicians, researchers, and operational purposes. As data become more accessible, protecting patient confidentiality is a requirement and expectation that should not be overlooked or understated.

The Department of Veteran's Affairs (VA) is funding a new informatics initiative called the Consortium for Healthcare Informatics Research (CHIR), focusing on utilizing both structured and unstructured data previously unavailable for research and operational purposes. These efforts have also focused on creating a high-performance computing environment to support data management, analytics and development environments called the Veterans' Informatics, Information and Computing Infrastructure (VINCI). Therefore, building methods and tools that can be used to automatically de-identify Veteran's Health Administration (VHA) clinical documents is of paramount importance in the development of this initiative. In the context of the CHIR, the de-identification project objectives include the investigation of the current state of the art of automatic clinical text de-identification[1], the development of a best-of-breed de-identification application for VHA clinical documents, and the evaluation of its impact on subsequent text analysis tasks and risk of re-identification of this text.

This paper presents an evaluation of the hybrid best-of-breed automated clinical text de-identification system mentioned above, along with the evaluation of two other machine learning-based text de-identification systems: MIST[2] and HIDE[3]. We evaluated these systems with a manually-annotated VHA clinical text corpus, and also with the 2006 i2b2 de-identification challenge corpus[4], and focused our analysis on generalizability issues and their impact on performance, as well as the impact of various resources used by the systems.

BACKGROUND

In the United States, current legislations require the patient consent when using clinical information for research purposes, but this requirement can be waived if the information is de-identified, as defined in the Health Insurance Portability and Accountability Act (HIPAA; codified as 45 CFR §160 and 164), and the Common Rule[5]. For clinical data to be considered de-identified, the HIPAA "Safe Harbor" technique requires 18 data elements (called PHI: Protected Health Information) to be removed[6], as listed in the Figure 1.

Figure 1. PHI as defined in the HIPAA "Safe Harbor" legislation.

1. Names;
2. All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
3. All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
4. Telephone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. Web Universal Resource Locators (URLs);
15. Internet Protocol (IP) address numbers;
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images; and
18. Any other unique identifying number, characteristic, or code.

Current de-identification applications can be classified in two main categories of methodologies: rule-based and machine learning-based.[1] Although, machine-learning based system can obtain learning features from rule-based techniques such as regular expressions and dictionaries, we base our classification on the main technique used to define the final annotations of PHI.

Rule-based systems mainly tackle the de-identification task with pattern matching, regular expressions and dictionary look-ups. The major drawback of this type of approach is the need for experienced domain experts to manually create patterns, rules and dictionaries. This is a tedious effort with limited generalizability. Although some dictionaries are quite generalizable, such as lists of person first names and lists of countries, others are built specifically for the institution in which the system was developed (e.g., list of actual names of patients, physicians or healthcare providers).[7-9] In addition, the developers of rule-based systems have to be aware of all possible PHI patterns that can occur, such as unexpected date formats or non-standard abbreviations.

Machine learning-based systems often rely on supervised methods and require a large annotated corpus for training, a resource that requires significant work by domain experts. However, these methods have the advantage of automatically learning how to recognize complex PHI patterns; consequently, developers only need limited knowledge of PHI patterns. A disadvantage of machine learning-based systems is that they may not learn PHI patterns that occur rarely in the annotated corpus.

Selecting meaningful learning features is an important aspect in order to build accurate machine learning models. Most of them use a variety of features ranging from lexical features (e.g., word-level features such as word case, punctuation, special and numerical characters, and the morphology of the word) to contextual features or complex features derived, for instance, from part-of-speech tagging or lexical-semantic resources[2-3,10-14].

METHODS

This study focuses on the evaluation of our hybrid approach, developed for the automatic de-identification of VHA clinical documents, and how it generalizes to other types of clinical documents such as the 2006 i2b2 de-identification corpus[4]. We also evaluate two available text de-identification tools based on machine learning algorithms: the MITRE Identification Scrubber Toolkit (MIST)[2]; and the Health Information DE-identification system (HIDE)[3]. In this section, we present the evaluation corpora, the three text de-identification applications and their preparation for this evaluation, and the performance measurement and analysis methodology.

Evaluation corpora

As mentioned above, we used several different corpora for this evaluation. The *VHA clinical documents training and testing corpora* consisted of a stratified random sample of a variety of clinical documents created between 04/01/2008 and 3/31/2009. Only documents with more than 500 words were included, and the 100 most frequent types of VHA clinical documents were used as strata for sampling. A total of 800 documents were sampled using

this approach, randomly assigned to a training corpus of 500 documents, and a testing corpus of 300 documents. The size of the testing corpus was estimated to allow for the demonstration of a difference of 2% or more in patient names recall (2-tailed significance level of 0.05, and power of 0.8). This corpus was then manually annotated for all categories of PHI defined in the HIPAA "Safe Harbor" legislation[6], as well as for some armed forces-specific information such as deployment locations, and units (e.g., regiment 50-2). Two reviewers independently annotated each document, a third reviewer adjudicated their disagreements, and a fourth reviewer eventually examined ambiguous and difficult adjudicated cases.

We also used the *2006 i2b2 de-identification challenge corpus*[4] to study the generalizability of the text de-identification systems when processing other document types. This corpus comprises discharge summaries from Partners Healthcare hospitals, split in a training corpus of 669 documents, and a testing corpus of 220 documents. For the challenge, these documents were de-identified, and all PHI identifiers were replaced with realistic surrogates, and they were manually annotated by a group of annotators to build the challenge reference standard (for training and for testing). Only part of the PHI categories defined in the HIPAA "Safe Harbor" legislation were included in the reference standard (i.e., patients and family members, doctors, hospitals, IDs, dates, locations, phone numbers, and ages above 89). We used this reference standard for our evaluation, mapping the challenge reference standard categories with our own categories, as explained below.

The MITRE Identification Scrubber Toolkit (MIST)

MIST[2] integrates an environment to support the development of automated de-identification of different document types. It uses an implementation of a machine learning method based on Conditional Random Fields (CRF)[15], and approaches the text de-identification task as a sequence-labeling problem implementing the BIO schema. This schema assigns each word a label indicating if the word is at the beginning of a PHI entity (B), inside a PHI entity (I) or outside an entity (O). For example, in the sentence *"[…] treated by Dr. Emani Marn on Tuesday […]"*,the following tags would be associated with each word: "*O O O B I O B*". Moreover, *B* and *I* tags can also indicate the entity type (e.g., "*O O O B-phi I-phi O B-phi*"). MIST offers several configurable options for the features used for training, as well as modifiable parameters for the machine learning algorithm.

To train and test MIST, we used the learning feature specification provided for the "AMIA De-identification" task distributed with MIST. This task is focused on the tagset and corpora from the 2006 i2b2 de-identification challenge. This specification includes features such as the target word, prefixes and suffixes, capitalization of the target word and of the two following and preceding words, digits inside the target word, special characters and other features about the morphology of the word, a context window of 2 words, and N-grams of words surrounding the target word. Features derived from dictionaries could be also specified.

When training and testing MIST with our VHA corpus, although we used the "AMIA De-identification" learning feature set, we had to create a new configuration covering our PHI categories and taking our training corpus as input for creating the machine learning models.

The Health Information DE-identification system (HIDE)

HIDE[3] also provides an environment for tagging, classifying and retagging of PHI, which allows constructing large training datasets without intensive human intervention. For documents de-identification, HIDE approaches the task as a traditional Named Entity Recognition (NER) problem[16], where each token is tagged following the same BIO schema implemented in MIST. A set of lexical features is fed into a Conditional Random Fields-based NER[17]. Approximately 34 features are derived from the morphology of the token (e.g., capitalization, special characters, affixes from length 1 to 3, single- double- triple- and quadruple-digit word, digits inside); moreover, the context-window processed by HIDE comprises the four previous and four following tokens. And, like MIST, HIDE also allows the addition of learning features derived from dictionaries.

The main differences between MIST and HIDE are the implementation of the machine learning algorithm, as well as the modules they use to deal with unstructured documents (e.g., tokenization). To our knowledge, these two systems are the only trainable text de-identification tools freely available for the research community.

Our hybrid VHA text de-identification system (BoB)

Our approach consists of a pipeline of processes designed to achieve excellent performance for VHA clinical text de-identification, with a focus on high sensitivity. We nicknamed it *BoB*, for *Best-of-Breed*, because we designed it based on a previous study of the best methods for VHA clinical documents de-identification[18].

BoB is being developed as an Apache UIMA[19] pipeline, with two main goals that correspond to BoB's main components: 1) obtain the highest recall (i.e., equivalent to sensitivity here) regardless of the impact on precision; and 2) improve overall precision (i.e., equivalent to positive predictive value here) by filtering out false positives.

Instead of tackling the classification task as a whole, we therefore implemented two main components focused, on recall and on precision, respectively. Recall is of paramount importance for de-identification – patient PHI cannot be disclosed at any rate – and this is BoB's main objective. BoB's design could somehow compromise on precision, but is intended to not compromise on recall, as detailed below.

BoB's processing starts with several Natural Language Processing (NLP) preprocessing tasks that prepare and parse the text for the two main components. It includes sentence segmentation, tokenization, part-of-speech tagging, phrase chunking, and words normalization based on Lexical Variant Generation (LVG)[19]. We adapted several OpenNLP[21] and cTAKES [22] components for these tasks.

The high-sensitivity extraction component then follows in the pipeline. This component is intended to give our system the highest recall or sensitivity when detecting PHI. It uses pattern matching and dictionaries, but also machine learning (CRFs) to achieve this goal. The module based on *pattern matching and dictionary lookups* implements a set of pattern matching techniques supported by contextual keyword searches (e.g., "Dr.", "Mr.", "M.D."), dictionary searches, and a simple disambiguation procedure based on a list of common words and the capitalization of the token. We adapted some of the techniques implemented in existing rule-based de-identification systems[7-9], and developed new patterns as well. For dictionary lookups, we used Lucene[23] indexing, experimenting with keyword and fuzzy dictionary searches. We use dictionaries of person names (from the 1990 U.S. census) split into first and last names (as in Neamatullah et al.[8]), of U.S. states, cities and counties, countries, companies (from Wikipedia, usps.com and other web resources), common words (from Neamatullah et al.[8]), and clinical eponyms and healthcare clinic names extracted from our VHA training corpus.

We created the module based on *CRF models* considering that machine learning classifiers are more generalizable and can detect instances of PHI identifiers not supported by our rules or dictionaries. Although this module could not reach a satisfying recall by itself, it helps predicting PHI formats and instances missed by our patterns and dictionaries. We used the CRF classifier implementation provided by the Stanford NLP group[24] with well-known empirically-demonstrated learning features such as the morphology of the words. Finally, to maximize recall, we also added the Stanford Named Entity Recognizer (NER)[25] into this component. Although this NER is trained with newswire documents, which clearly differ from clinical notes, some of the entities it recognizes (e.g., Persons, Organizations, Dates, Locations) overlap with PHI identifiers and could therefore improve the overall recall of this component.

The false-positive filtering component was designed to filter out the false positives produced by the previous high sensitivity extraction component. For this task, we built a series of machine learning classifiers based on LIBSVM[26], a library for Support Vector Machines (SVM), and also based on linear classification, the LIBLINEAR library[27]. Machine learning-based approaches are in most cases more precise in classification problems than rule-based techniques[18], motivating our decision to choose machine learning algorithms for this component. We trained these classifiers with reference standard annotations, as well as with correct and incorrect annotations made by the high sensitivity extraction component. Therefore, unlike other text de-identification systems based on machine-learning, or even our CRF models, the false-positive filtering classifiers were trained using candidate annotations derived from the high-sensitivity extraction component, so they do not predict if every token is or belongs to a PHI identifier, but instead decide if an actual annotation is a false positive or a true positive. This design allows for better performance with less learning examples, which is a restriction we have to deal with. It also allows us to create methods (i.e., pattern matching and dictionary lookups) that can be only focused on maximizing recall regardless of their impact on overall precision.

In order to make the system's comparison as coherent as possible, we created three different configurations of features based on dictionaries used by BoB for the the MIST and HIDE systems:

- Configurations without any dictionary feature (*MIST_noD* and *HIDE_noD*): these configurations use the feature learning set provided with the "AMIA De-identification" task for MIST, and the default learning set integrated in HIDE.
- Configurations with a selection of dictionary features (*MIST_selD* and *HIDE_selD*): when training MIST and HIDE, we also wanted to add learning features derived from all dictionaries used by BoB. To our understanding, this would make the comparison fairer, but the HIDE CRF model could not be trained with

all these features (program crashed, maybe because of the size of the learning features set). We therefore made other configurations of HIDE discarding the heaviest dictionaries one-by-one. Finally, we got HIDE CRF model trained with all dictionaries but 'common_words', 'US_cities' and 'last_names'. We called this configuration *_selD*, and it was used by both MIST and HIDE. This dictionary selection could have an impact on the final performance; however it was the only configuration of HIDE that we were able to run with dictionaries.

- Configuration with all dictionary features (*MIST_allD)*: the configuration of MIST adding into the training phase all our dictionaries as learning features. It is, in theory, the most appropriate configuration of MIST to be compared against our de-identification system, BoB.

Systems Evaluation

Convinced that a conservative approach for measurements and analysis is the most reliable in order to protect patient confidentiality, we designed our evaluation with the following considerations:

- Our evaluation is done at the PHI-identifier level. We do not measure performance at the token (e.g., word) level, but instead consider the entire PHI identifier as our evaluation unit.
- Our evaluation considers "fully-contained" matches. Exact matches are sometimes excessively strict, when some patterns or heuristics include non-functional words or word delimiters that cause a mismatch of the offsets. However, partial matches increase the risk of patient privacy breach, since a portion of PHI identifier could uniquely establish a link to the patient. We propose "fully-contained" matches, relaxing the exact matching strategy, but assuring redaction of the complete PHI. A "fully-contained" match considers the prediction as correct when it at least covers the entire PHI in the reference standard (Figure 2).

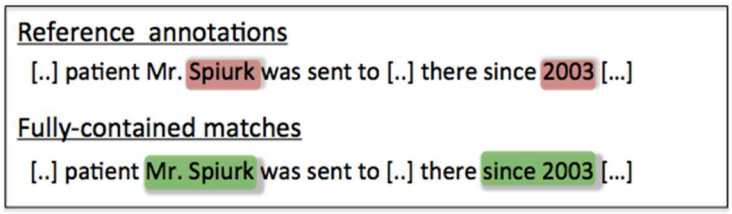

Figure 2. Example of a fully-contained match against the reference standard.

We measured de-identification performance in terms of recall (equivalent to sensitivity), precision (equivalent to positive predictive value), and F-measure (harmonic mean of recall and precision). As mentioned above, recall is of paramount importance for de-identification, and we also use the F_2-measure, which weighs recall (twice) higher than precision, apart from the traditional F_1-measure:

$$F_\beta - measure = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}\ ; \qquad \beta = 2$$

Recall and precision were computed with counts of true positives (system predictions fully-contained matching the reference standard), false positives (spurious system output), and false negatives (missed by the system, or only partly matched).

RESULTS

VHA corpus-based evaluation: For this evaluation, all three systems (i.e., MIST, HIDE and BoB) were trained with the VHA clinical documents training corpus, and tested with the VHA clinical documents testing corpus. Table 1 summarizes the results achieved by these systems, with the different configurations explained above (*MIST_noD* and *HIDE_noD* without dictionary features, *MIST_selD* and *HIDE_selD* with a selection of dictionary features, and *MIST_allD* with all dictionary features). We report the number of instances used for training and testing, the recall (or sensitivity) achieved for each PHI category, the overall macro-averaged recall at the PHI type level (i.e.,

considering each PHI category separately), and the micro-averaged results at PHI level (i.e., considering all PHI categories together, as one "PHI" category) in terms of precision, recall, and F_1- measure and F_2-measure.

EVALUATION WITH VHA CLINICAL DOCUMENTS

				System and configuration					
VHA PHI categories	#inst. train	#inst. test	Measure	MIST_noD	HIDE_noD	MIST_selD	HIDE_selD	MIST_allD	BoB
PatientName	741	254	Recall	0.835	0.776	0.862	0.855	0.850	0.980
RelativeName	55	25	Recall	0.440	0.360	0.520	0.840	0.440	0.920
HealtCareProviderName	1131	543	Recall	0.858	0.845	0.900	0.886	0.862	0.943
OtherPersonName	36	9	Recall	0.333	0.555	0.555	0.555	0.555	0.888
StreetCity	332	157	Recall	0.796	0.732	0.809	0.841	0.777	0.943
StateCountry	351	148	Recall	0.804	0.750	0.817	0.838	0.797	0.878
Deployment	67	53	Recall	0.868	0.811	0.849	0.773	0.773	0.887
Zipcode	10	5	Recall	1	1	1	0.800	1	1
HealthCareUnitName	3601	1629	Recall	0.792	0.733	0.827	0.741	0.786	0.811
OtherOrgName	192	91	Recall	0.582	0.516	0.604	0.450	0.527	0.725
Date	6804	3513	Recall	0.960	0.940	0.963	0.934	0.956	0.971
Age89+	9	4	Recall	0.250	0	0.250	0	0	1
PhoneNumber	215	91	Recall	0.912	0.846	0.934	0.846	0.901	0.956
ElectronicAddress	6	4	Recall	0.500	0.500	0	0.500	0	1
SSN	58	27	Recall	0.963	0.963	0.963	0.963	0.963	1
OtherIDNumber	493	180	Recall	0.906	0.851	0.894	0.845	0.912	0.917
Overall macro-averaged recall (PHI-type level)				0.737	0.699	0.734	0.729	0.694	0.926
Overall micro-averaged (PHI level)			Precision	0.926	0.936	0.715	0.933	0.700	0.836
			Recall	0.888	0.853	0.904	0.863	0.883	0.922
			F_1	0.907	0.893	0.799	0.897	0.781	0.877
			F_2	0.895	0.869	0.858	0.877	0.839	0.904

Table 1. Evaluation results with VHA clinical documents.

Although Table 1 includes each "Person-Name" category (i.e., PatientName, RelativeName, HealthCareProviderName, OtherPersonName), since our reference standard was annotated with these categories, we trained and tested all three systems (MIST, HIDE and BoB) considering these categories as one "Person-Name" PHI category. Thus, any Person-Name annotation produced by the systems could match a PatientName, RelativeName, HealthCareProviderName or OtherPersonName annotation in the reference standard.

i2b2 corpus-based evaluation: For this evaluation, all three systems were first trained with the i2b2 de-identification challenge training corpus, and tested with the i2b2 de-identification challenge testing corpus, and then trained with our VHA clinical documents training corpus, and tested with the i2b2 de-identification challenge testing corpus. This last evaluation provides insight of how classification models created by these systems perform across documents from different institutions and types. We had to map our PHI categories to the ones specified in the i2b2 de-identification challenge (see Uzuner et al.[5] for further details of the i2b2 PHI categories) for this evaluation, as follows:

- 'Patient' and 'Doctor' categories were mapped to Person-Name annotations (*PatientName*, *RelativeName*, *HealthCareProviderName* and *OtherPersonName)* produced by the systems.
- 'Hospitals' were mapped to the *HealthCareUnitName* category.
- 'IDs' annotations were mapped to *OtherIDNumber* and *SSN* identifiers.
- 'Dates' were mapped to *Date*. However, for consistency with our PHI specification, we modified the i2b2 challenge reference standard in order to also include the year with date annotations. MIST distribution provides a script that carries out these modifications.
- 'Locations' were mapped to *StreetCity*, *StateCountry* and *Zipcode*.
- Finally, 'Phone numbers' were mapped to *PhoneNumber* and 'Ages' to *Age+89*.

EVALUATION WITH THE I2B2 DE-IDENTIFICATION CORPUS

				Systems trained using the i2b2 training corpus			Systems trained using our VHA training corpus		
i2b2 PHI categories	#inst. train	#inst. test	Measure	MIST_noD	HIDE_noD	BoB	MIST_noD	HIDE_noD	BoB
Patient	684	245	Recall	0.918	0.955	0.975	0.584	0.437	0.604
Doctor	2681	1070	Recall	0.969	0.972	0.980	0.549	0.392	0.600
Location	144	119	Recall	0.437	0.513	0.613	0.428	0.344	0.639
Hospital	1724	676	Recall	0.911	0.793	0.910	0.645	0.487	0.839
Date	5167	1931	Recall	0.984	0.987	0.990	0.869	0.869	0.988
Age	13	3	Recall	0	0.333	0	0	0	0
Phone number	174	58	Recall	0.776	0.827	0.810	0.948	0.810	0.845
ID	3666	1143	Recall	0.992	0.996	0.784	0.854	0.348	0.795
Overall macro-average recall (PHI-type level)				0.748	0.797	0.758	0.610	0.461	0.664
Overall micro- average (PHI level)			Precision	0.983	0.972	0.878	0.705	0.712	0.691
			Recall	0.955	0.947	0.921	0.749	0.576	0.820
			F1	0.969	0.959	0.899	0.726	0.637	0.750
			F2	0.961	0.952	0.912	0.740	0.599	0.790

Table 2. Evaluation results with the i2b2 de-identification challenge corpora, trained with the i2b2 training corpus and with our VHA training corpus.

DISCUSSION

Results of the VHA corpus-based evaluation (Table 1) point out that our hybrid approach, in terms of recall, performs better for almost every PHI category than the other two systems involved in this study (i.e., MIST and HIDE), achieving a high 92.6% overall macro-averaged recall, and individual recall rates between 98% and 100% for the most sensitive PHI categories. This demonstrates that our approach allowed us to take advantage of rule-based strategies qualities, accomplishing a conservative system that prioritizes patient privacy by all means.

Both MIST and HIDE achieved a micro-averaged F_1-measure around 90%, which means that these systems are competitive for de-identifying our documents. However, when analyzing recall for each PHI category, we observe that recall around 85% for patient names is lower than desired for such highly sensitive PHI types. Indeed, the overall macro-averaged recall reached by these systems was in all cases around 70-74%.

With the addition of dictionaries used by BoB as learning features for MIST and, we tried to enrich these systems with the knowledge we already exploited in BoB's high sensitivity extraction component. However, the way this knowledge is used by these systems is far from the way our hybrid approach independently deals with rule-based and machine learning-based methods. It is interesting that HIDE micro-averaged results do not significantly change with the addition of dictionary features, although recall was a bit higher for some PHI categories (from 0.699 to 0.729 in macro-averaged recall). With MIST, the addition of dictionaries does cause an increase in micro-averaged recall, and a consequent impact on the precision (*MIST_selD* 90% recall 71% precision). These observations hint that dictionaries play an important role in the sensitivity of the systems, and the difficulty resides in ways to maintain acceptable precision. BoB manages this issue with the independent component focused on the classification of candidate PHI annotations as false positives or true positives. With this component, BoB achieves a micro-averaged precision of 84%, and more importantly keeps recall at high level (92%). This suggests that BoB's architectural design successfully deals with the impact on precision of methods focused on increasing sensitivity.

Results of the i2b2 corpus-based evaluations allow us to look into the controversial issue of generalizability of automated text de-identification systems, an issue that is always debatable since these systems are normally developed for specific document types and institutions. To approach this matter, we evaluated the systems using the 2006 i2b2 de-identification corpora, which differ in type and structure from our VHA clinical documents. Moreover, the fact that the i2b2 corpus was de-identified and PHI replaced with surrogates that could not be found in dictionaries (as stated in Uzuner et al.[4]), makes its de-identification difficult for techniques based on dictionaries. This does not correspond to the reality of clinical documents, at least regarding person first names, but it is a good challenge for our hybrid approach.

Nonetheless, we believe that results across different document types and an evaluation of the portability of the training models are of paramount importance for creating more generalizable systems.

When trained and tested with the i2b2 corpora (Table 2), no dictionary features were used with MIST and HIDE. We chose this configuration because features from dictionaries did not provide much meaningful knowledge for the i2b2 corpus, and also because this configuration obtained high F_1-measure values with our VHA documents. However, we decided not to discard dictionaries with BoB, since this would involve removing the high sensitivity extraction component, and consequently change the philosophy and architecture of the entire system. We then only trained BoB machine learning-based classifiers with the i2b2 training corpus, without doing any modification or adaptation of our rule-based techniques. Results from this evaluation point out that MIST and HIDE perform extremely well with this corpus, and the difference in performance for person names in comparison with our VHA corpus-based evaluation is interesting. The way the i2b2 challenge corpus was created has an obvious impact on this issue. BoB also performs well, especially for person names, which means that although most instances were not found in our dictionaries, the trigger words matching and our CRF classifiers also enable a good classification. However, BoB presents a lower recall with *ID*, *Phone number* and *Ages*, which we believe would be easily solved by slightly adapting our patterns to some formats present within the i2b2 training corpus. The lower performance with the *Location* category is interesting. While the three systems obtained high performance with our VHA *StreetCity*, *StateCountry* and *Zipcode* categories, all of them reported only between 43-61% recall with the i2b2 *Location* category. We attribute this to: 1) the way the i2b2 organizers generated the corpora, introducing surrogates by permuting the syllables of existing names from dictionaries (see Uzuner et al.[5] for further details), causing pattern matching techniques based on dictionaries to probably fail to detect i2b2 locations, and 2) the total amount of *Locations* within the i2b2 training corpus; there are only few examples of *Locations* in the training corpus, and it doesn't allow the classifiers trained using these examples to perform very accurately.

When trained with VHA documents and tested on the i2b2 corpus, none of the three systems reported good results for this experiment, as expected. HIDE obtained the best micro-averaged precision (71%), but also the lowest recall (58%). MIST was able to keep a balance between recall and precision around 70-74%. BoB achieved the best recall (82%) and a precision of 69%. One possible reason for these results is the difference of entities proportion in each category. For instance, our VHA corpus contains more *HealthCareUnitNames* than person names, while the i2b2 corpus proposed those categories in reverse order. This fact can cause classifiers used across different document sources to be less accurate. We drew two conclusions from these observations: 1) although we notice that models portability across document types is an issue still unsolved and difficult to deal with, these encouraging results make us think that we are not far from achieving portable de-identification systems obtaining competitive results with various clinical document types; and 2) the design we devised for BoB is also supported by these results; BoB prioritizes recall, reaching the highest recall among the three systems, and it satisfactorily preserves precision, obtaining results similar to MIST and HIDE.

CONCLUSIONS AND FUTURE WORK

We have presented an evaluation of our hybrid automated VHA clinical text de-identification system – called BoB – developed within the VA CHIR initiative. We also provided results from two trainable available text de-identification systems (i.e., MIST and HIDE). Moreover, to further enhance this evaluation, apart from testing the systems with our VHA documents, we report performance of these systems with the 2006 i2b2 de-identification challenge corpora, as well as generalizability of their models when trained and tested with documents from different sources.

Our evaluations demonstrate that BoB satisfactorily accomplishes our main goal – it prioritizes high sensitivity and patient privacy – and moreover, it also achieves competitive precision, preserving the subsequent interpretability of the de-identified documents. MIST and HIDE also reach competitive results, achieving in many cases better precision than our system, although the recall for some sensitive PHI categories is not as high as desired for de-identification. To our understanding, a good de-identification procedure should obtain recall rates in the nineties (>90%). Recall rates for highly sensitive PHI identifiers such as patient names and Social Security Numbers are even more important, and should reach values above 95%. On the other hand, these values may vary depending on the final use of the de-identified documents, the sharing policies, and the methods imposed to avoid re-identification.

We realize that there is still room for future improvements, and we plan to carry out an exhaustive error analysis from these two evaluations. It will provide us with precise clues to improve our methods. Also, the fact that MIST and HIDE achieved high precision encourages us to work on improving our trainable models.

ACKNOWLEDGEMENTS

Funding provided by the Department of Veterans Affairs Health Services Research & Development Services Consortium for Healthcare Informatics Research grant (HIR 08-374).

REFERENCES

1. Meystre SM, Friedlin FJ, South BR, Shen S, Samore MH. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010 Aug 2;10:70.
2. Aberdeen J, Bayer S, Yeniterzi R et al. The MITRE Identification Scrubber Toolkit: Design, training, and assessment. Int J Med Inform 2010 Dec;79(12):849-59.
3. Gardner J, Xiong L. An integrated framework for de-identifying unstructured medical data. Data Knowl Eng 2009 Dec;68(12):1441-1451.
4. Uzuner Ö, Luo Y, Szolovits P. Evaluating the State-of-the-Art in Automatic De-identification. J Am Med Inform Assoc 2007 Sep-Oct;14(5):550-563.
5. GPO, U.S. *45 C.F.R. § 46 Protection of Human Subjects*. 2008; Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html.
6. GPO, U.S. *45 C.F.R. § 164 Security and Privacy*. 2008; Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html.
7. Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc 2008 Sept-Oct;15(5):601-610.
8. Neamatullah I, Douglass MM, Lehman LH et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.
9. Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12.
10. Aramaki E., et al. Automatic Deidentification by using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006. Washington, DC.
11. Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple R. Identifying Personal Health Information Using Support Vector Machines. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.
12. Hara K. Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. 2006.

13. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc 2007;14:574–580.
14. Uzuner O, Sibanda CT, Luo Y, Szolovits P. A de-identifier for medical discharge summaries. Artif Intell Med, 2008;42(1):13-35.
15. Wellner B. Sequence Models and Ranking Methods for Discourse Parsing. Ph.D. Dissertation, Brandeis University, Waltham, MA, 2009.
16. Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History. 16th Int Conf Comp Ling (COLING) 1996;466-471.
17. Okazaki N. CRFsuite: a fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/, 2007.
18. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. Evaluating Current Automatic De-identification Methods with Veteran's Health Administration Clinical Documents. Submitted to BMC Med Res Methodol.
19. Apache UIMA 2008. Available at http://uima.apache.org.
20. LVG (Lexical Variants Generation). 2010. Available at: http://lexsrv2.nlm.nih.gov/LexSysGroup/Projects/lvg.
21. OpenNLP. Available at http://opennlp.sourceforge.net/.
22. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. JAMIA 2010;17(5): 507-13.
23. The Apache Lucene project. Available at http://lucene.apache.org.
24. The Stanford coreNLP library. Available at http://nlp.stanford.edu/software/corenlp.shtml.
25. Finkel JR, Grenager T, Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005) 2005; 363-370.
26. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. Computer, 1-30. 2001.
27. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 2008;9:1871-1874.