A Hybrid Approach to De-identify Electronic Health Records

Anonymous EMNLP submission

Abstract

De-identification of clinical notes continues to be an unsolved challenge because of the heterogeneity of entities that need to be extracted. In this paper, we explore a hybrid approach to identify Protected Health Information (PHI) from clinical notes. The system was trained and tested on real-life clinical notes released as part of the i2b2 2014 and 2016 NLP shared tasks. Our approach combines high-precision template-based approaches with conditional random fields for labeling eighteen classes of PHI fields in clinical records. The results outperform existing off-the-shelf de-identification packages.

1 Introduction

Identifying Protected Health Information (PHI) from clinical notes is critical to protect the privacy of individuals before their data could be shared publicly. The guidelines set by the US Health Information Portability and Accountability Act (HIPAA) identify eighteen classes of PHI identifiers as protected information. The goal of de-identification is to identify and remove PHI from health records while protecting the integrity of the data as much as possible (Uzuner et al., 2007). De-identification ensures the safe removal of the information that explicitly identifies persons involved in health care from the clinical notes, thereby protecting their privacy. Although these classes are relatively straight forward, developing automated approaches to identify them from free text remains a challenging task. In this paper, we describe a hybrid approach, combining the labels derived from a supervised sequence labeler and hand-crafted pattern-based labeler, to successfully de-identify all types of PHI categories and outperform existing state-of-the-art approaches.

2 Related work

The idea of automatic de-identification of health records resembles traditional named entity recognition methods (Nadeau and Sekine, 2007; Ratinov and Roth, 2009). A review on recent state-of-the-art systems shows that de-identification of clinical records can be reasonably achieved using a rule-based approach, a machine learning approach, or a combination of these approaches (Meystre et al., 2010). The rule based methods make use of gazetteers, lists, and manually-crafted rules using regular expressions to match the PHI patterns in the records (Neamatullah et al., 2008). Although the rule-based methods suffer from lack of generalizability and require both time and skill for creating rules, previous studies have found that they work well for rare PHI items (Deleger et al., 2013). On the other hand, machine learning based methods, such as conditional random fields (Lafferty et al., 2001), that are trained over a labeled data set to automatically detect PHI patterns, are more generalizable (Deleger et al., 2013). These methods, however, require a large set of manually annotated examples for training. Hybrid methods using a combination of highly precise rule-based approaches and easily generalizable machine learning based approaches usually tend to obtain the best results (Meystre et al., 2010; Uzuner et al., 2007). But, most of these combined studies often ignore important PHI categories such as age, geographic location, and contact information (Deleger et al., 2013). To overcome these limitations, the center for Informatics for Integrating Biology and the Bedside (i2b2), a US national center for biomedical computing, has

Meta class labels	PHI field labels	
	Included (18)	**Excluded** (12)
Date	Date	–
Age	Age	–
Location	Street, City, State, Country, Zip, Organization, Hospital	Room, Department
Name	PatientName, DoctorName	Username
Contact	URL, Phone, Fax, Email	IPaddr
Profession	Profession	Other
ID	License, Medical Record Number	SSN, HealthPlan, Account, Vehicle, Device, BioID, IDnum

Table 1: Fields included and excluded in this study. Excluded fields total 1.4% (n=866) of all PHI fields.

been organizing regular challenges and releasing labeled datasets for evaluation of de-identification tools of clinical records since 2006 (Uzuner et al., 2007). These datasets have led to numerous research efforts including supervised approaches for de-identification (Dernoncourt et al., 2017).

One of the most recognized clinical notes de-identification toolkits is the MITRE Identification Scrubber Toolkit (MIST), developed by Aberdeen et al. (Aberdeen et al., 2010). MIST is an open source platform that provides an environment to support rapid tailoring of automated de-identification to different document types using machine learning classifiers to de-identify sensitive information. The MIT de-identification tool (Neamatullah et al., 2008) is an automated pattern-matching based de-identification software package that uses lexical look-up tables, regular expressions, and simple heuristics to locate PHI fields. HMS Scrubber (Beckwith et al., 2006) is a de-identification tool tailored for pathology reports, while Health Information DE-identifier (HIDE) (Gardner and Xiong, 2008) is a configurable framework. Various systems have also been proposed to de-identify health records of other languages. Grouin et al. (Grouin and Neveol, 2014) developed a de-identification system for reference corpus in French, while Dalianis et al. (Dalianis and Velupillai, 2010) introduced a framework for de-identification of Swedish clinical notes.

3 Methodology

We developed a hybrid de-identification system consisting of four main components: (a) Data cleaning and pre-processing, (b) Identifying pattern-based PHI fields, (c) Identifying gazetteer-based and contextual features, and (d) Training a supervised sequence classification model.

3.1 Data Set

In this paper, we use the dataset released as part of the 2014 i2b2/UTHealth shared tasks (i2b2, 2014) and the 2016 i2b2 Research Domain Criteria (RDoC) for Psychiatry challenge (i2b2, 2016). For the 2014 challenge, the organizers collected 1,304 medical records that were annotated by domain experts and students for all instances of PHI fields. 790 notes were included in the training set and the remaining 514 notes in the test set. In the 2016 challenge, there were 1,000 annotated notes, 600 of which were released as the training set and the remaining 400 as the test set. In all, there are thirty fields annotated in the clinical notes, divided into seven higher-level categories, as summarized in Table 1.

In this paper, we focus on eighteen fields and ignore the twelve infrequent fields listed in the rightmost column of Table 1. The excluded fields are very rare in the dataset and together represent only 1.4% (n=866) of all labeled PHI fields. We trained our models on the combined training data sets from both 2014 and 2016 challenges. The test set performance is reported based on the "strict matching" of identified fields.

3.2 Data cleaning and Pre-processing

The data consisted of patient notes in their raw form with numerous data cleaning issues, including conjoined words and misplaced newlines. As the first step of our approach, we searched for conjoined words using pattern matching techniques and separated them. We also added new line characters before potential headings to recover, to the best extend possible, the semi-structured template common in clinical notes. At each stage, we maintained a mapping between word boundaries, i.e. character offsets in the original text and its cleaned version, so as to trace back word offsets of the identified fields to the original text.

Dataset	Features	Date	Age	Location	Name	Contact	Profession	ID	Total
2014	1. Neighborhood + POS	0.968	**0.947**	0.803	0.842	0.915	0.641	**0.887**	0.939
	2. (1) + Wordform	0.968	0.943	0.796	0.866	0.922	0.674	0.881	0.941
	3. (2) + Regex	0.971	0.946	**0.836**	**0.882**	0.935	0.704	0.865	0.945
	4. (3) + IsInX (**All feats**)	**0.972**	**0.947**	0.831	0.874	**0.946**	**0.720**	0.848	**0.947**
	18 classes, All feats	0.970	0.946	0.831	0.880	0.935	0.703	0.855	0.945
2016	1. Neighborhood + POS	0.925	0.926	0.628	0.514	0.890	0.563	**0.769**	0.849
	2. (1) + Wordform	0.935	0.931	0.672	0.647	0.901	0.654	0.741	0.873
	3. (2) + Regex	0.939	0.931	0.717	0.684	0.904	0.660	0.727	0.884
	4. (3) + IsInX (**All feats**)	0.939	**0.933**	**0.719**	**0.686**	**0.928**	**0.666**	0.727	**0.889**
	18 classes, All feats	**0.941**	0.930	0.716	0.684	0.904	0.660	0.727	0.885

Table 2: "Strict matching" F1 values for each meta class as feature sets are varied. For each dataset, the last row shows the best 18-class CRF model performance, while the other rows show the performance of the best 12-class CRF model combined with a pattern-based labeler for the remaining six classes.

	2014 Dataset		2016 Dataset	
	N	F1	N	F1
City	260	0.808	820	0.806
State	190	0.898	481	0.899
Country	117	0.754	376	0.888
Hospital	875	0.807	1327	0.777
Organization	82	0.371	697	0.579
PatientName	877	0.874	837	0.685
DoctorName	1911	0.895	1567	0.936
Profession	178	0.721	1010	0.666

Table 3: The classes with lowest F1 scores using the best model (cf. Table 2)

3.3 Identifying pattern-based PHI fields

Some of the classes of PHI identifiers are formatted text fields that are easy to recognize using regular expressions. These include dates, phone and fax numbers, patient identifiers, URLs, and email addresses. To identify these fields, we developed sets of regular expressions and tuned them to maximize their accuracy on the training data. For some classes such as Phone numbers and Fax numbers, we also searched for terms in the vicinity of the fields (such as 'Phone' or 'Fax'), to correctly disambiguate the PHI classes.

3.4 Identifying gazetteer-based and contextual features

Gazetteers have been used as an important resource in identifying named entities in free-text (Ratinov and Roth, 2009). We investigated if gazetteers could also be used to identify some of the PHI classes, such as City, State, and Country, with high accuracy. We collected lists of cities, US states, and country names and the demonyms associated with these locations. These lists were collected from multiple online sources including Wikipedia and Topix and were curated to remove commonly occurring English words to limit false positive matches.

We also collected other lists of words to serve as feature markers. These include lists of honorifics and medical professional degrees (e.g., 'Dr.', 'MD', 'RN'), common first and last names, common professions, and common title words related to hospitals and organizations (e.g., 'hospice', 'center').

Finally, we identified the following features:

- IsInX features: Gazetteer- and list-based features; if a token appears in a list, the IsInX feature linked to that list will be active.
- Regex features: Regular expressions derived for pattern-based fields were used as separate features. For example, if a token matched one of the regular expressions for Phone numbers, then the Phone-regex feature is active.
- Wordform features: Features based on word characteristics, such as capitalization, title-casing, and character-ngrams.
- POS feature: POS tags were obtained after running Stanford Parts of Speech Tagger (Toutanova et al., 2003) on the dataset.
- Neighborhood features: Finally, for each token, we also included wordform and POS features for neighborhood tokens within a window size of 7 centered on a given token.

3.5 Training a supervised sequence tagger

Once the pattern-based, gazetteer-based, and contextual features were identified, we trained a Conditional Random Field (Lafferty et al., 2001) to detect all PHI fields from clinical notes. Two different class-configuration models were trained. First, we trained a CRF model on all eighteen classes listed in the second column of Table 1.

Dataset	System	Date	Age	Location	Hospital	Name	Phone	URL	Email	Total
2014	MIST (raw data)	0.867	0.800	0.721	0.742	0.546	0.255	0	0	0.754
	MIST (clean data)	0.870	0.886	0.715	0.781	0.805	0.299	0	0	0.825
	Our approach	0.972	0.948	0.862	0.806	0.918	0.947	0	1.000	**0.928**
2016	MIST (raw data)	0.633	0.671	0.746	0.705	0.499	0.278	0	0	0.638
	MIST (clean data)	0.886	0.886	0.805	0.751	0.801	0.682	0	0	0.840
	Our approach	0.939	0.932	0.809	0.776	0.876	0.969	0.333	1.000	**0.885**

Table 4: F1 measures for the classes MIST is able to recognize.

Second, we trained a 12-class CRF model after excluding the six classes (Zip, License, and four classes in Contact) for which pattern-based approaches gave very good results. The final result was generated by combining outputs of the 12-class CRF model and the six pattern-based labelers. All PHI fields in the training set were encoded using the BIO-schema. We used the CRFSuite implementation (Okazaki, 2007), and for all CRF models, set the CRFsuite parameters as $c1 = 0.05$ and $c2 = 0.2$.

4 Results and Discussion

4.1 Comparison of feature sets

We first show how the individual feature classes help improve the overall accuracy of the models. Table 2 summarizes the performance of the 18-class model and the hybrid approach for various feature sets on 2014 and 2016 datasets. Comparing the last two rows of Table 2 for each dataset, we note that the hybrid approach of training a smaller 12-class CRF model and merging the labels with a highly accurate pattern-based labeler is better than letting CRF train a full 18-class model. This is primarily because the pattern based labeler outperforms CRFs on Contact and ID classes. In addition, even for classes that were labeled using CRF, adding the pattern-based features (Regex features in Table 2) improved precision, recall, and F1 on the Date and Location classes. Gazetteer- and list-based features (IsInX features) further improve the results to give the best overall F1 of 0.947 and 0.889, respectively.

4.2 Focused look on specific classes

We further analyzed the classes with lowest F1 measures, viz. Location, Name, and Profession. Table 3 summarizes the findings. We observed that the high error rates in City, Hospital, and Organization are primarily responsible for the lower performance of the Location class, while the PatientName class limits the performance on Name. These errors can be partly explained by the lack of specific discriminative features for these classes, and incomplete gazetteers and lists. Some Hospital and Organization names are abbreviated, further reducing the contextual features available for correct labeling.

4.3 Comparison to MIST

Finally, we compare our approach against MIST, a popular de-identification toolkit. Since MIST does not handle all PHI classes, we compare against only the classes that MIST recognizes. Table 4 summarizes the results. When MIST is run "out-of-the-box" on the raw data, it achieves an F1 of 0.754 on 2014 data and 0.638 on 2016 data. On the clinical notes cleaned using our approaches (Sec. 3.2), the F1 measure improves to 0.825 and 0.840, respectively. However, our hybrid system is able to further improve the performance on **all** classes and reach the overall F1 of 0.928 and 0.885 on 2014 and 2016 data. This further shows the improved effectiveness of proposed approach over well-established alternatives.

5 Conclusion

In this paper, we present a new hybrid approach of combining pattern-based, gazetteer-based, and machine learning based approaches to identify protected health information. The combined strategy of deploying supervised sequence taggers and pattern-based labelers together results in improved performance in labeling all eighteen classes of PHI. The proposed method improves over popular de-identification toolkits in both the range of classes identified and accuracy of labeling. In future, we intend to explore infusing deep learning techniques to improve the performance without gazetteer lists and to make the developed system publicly available to the biomedical NLP research community and help benchmark existing de-identification systems against additional real-life clinical notes within medical centers and academic health systems.

References

J Aberdeen, S Bayer, R Yeniterzi, B Wellner, C Clark, David Hanauer, B Malin, and L Hirschman. 2010. The mitre identification scrubber toolkit: Design, training, and assessment. *International Journal of Medical Informatics* 79(12):849–859.

BA Beckwith, R Mahaadevan, UJ Balis, and F Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making* 6:12.

H Dalianis and S Velupillai. 2010. De-identifying swedish clinical text - refinement of a gold standard and experiments with conditional random fields. *Journal of Biomedical Semantics* 1(6).

L Deleger, K Molnar, G Savova, F Xia, T Lingren, Q Li, K Marsolo, A Jegga, M Kaiser, L Stoutenborough, and I Solti. 2013. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal American Medical Informatics Association* 20(1):84–94.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24(3):596–606.

J Gardner and L Xiong. 2008. Hide: an integrated system for health information de-identification. In *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS)*. pages 254–259.

C Grouin and A Neveol. 2014. De-identification of clinical notes in french: towards a protocol for reference corpus development. *Journal of Biomedical Informatics* 50:151–161.

i2b2. 2014. i2b2/UTHealth shared-tasks and workshop on challenges in natural language processing for clinical data. Https://www.i2b2.org/NLP/HeartDisease/.

i2b2. 2016. CEGS NGRID shared tasks and workshop on challenges in NLP for clinical data guidelines. Https://www.i2b2.org/NLP/RDoCforPsychiatry.

J D Lafferty, A McCallum, and F C N Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the ICML*. page 282289.

S M Meystre, F J Friedlin, B R South, S Shen, and M H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology* 10(70).

D Nadeau and S Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1):326–350.

I Neamatullah, M M Douglass, L H Lehman, A Reisner, M Villarroel, W J Long, P Szolovits, G B Moody, R G Mark, and G D Clifford. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* 8(32):601–610.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). Http://www.chokkan.org/software/crfsuite/.

L Ratinov and D Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL*. pages 147–155.

K Toutanova, D Klein, C Manning, and Y Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL,*. pages 252–259.

O Uzuner, Y Luo, and P Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of American Medical Informatics Association* 14(5):550–563.