Automatic de-identification of medical records with a multilevel hybrid semi-supervised learning approach

Nguyen Dong Phuong
Center for Applied Information Technology
Ton Duc Thang University
Ho Chi Minh City University of Technology
Vietnam National University
Ho Chi Minh City, Vietnam
nguyendongphuong@tdt.edu.vn

Vo Thi Ngoc Chau
Department of Information Systems
Faculty of Computer Science and Engineering
Ho Chi Minh city University of Technology
Vietnam National University
Ho Chi Minh City, Vietnam
chauvtn@hcmut.edu.vn

***Abstract*—In recent years, sharing electronic medical records (EMRs) for more researchers outside the associated institutions is significant. For privacy preservation of the corresponding patients and the associated institutions, a de-identification task on the EMRs to be shared is a must. Although the de-identification task has been considered with positive research outcomes worldwide, especially those from the i2b2 (Informatics for Integrating Biology and the Bedside) shared tasks in 2006 and 2014, the task has not yet been a solved problem and still needs more investigation realistically. In this paper, we propose an automatic de-identification solution in a multilevel hybrid semi-supervised learning paradigm with a key focus on correctly identifying protected health information (PHI) in the EMRs. Similar to the existing works, our work defines a hybrid approach by combining a machine learning-based method with a conditional random fields model and a rule-based method in a post-processing phase to handle the PHI types with disambiguity. Nevertheless, our work is more general and practical. First, it considers the structure complexity of each EMR so that each section can be treated properly for more correct PHI identification up to its structure complexity: structured, semi-structured, or un-structured. Second, each EMR is then examined in our approach at three different levels of granularity such as a token level in the supervised learning phase, an entity level in the rule-based post-processing phase, and a section level along with the structure complexity in the semi-supervised learning phase. Many various detail levels will give our approach a deeper look at each EMR for more effectiveness. Third, our solution is conducted in a self-training manner so that it can get started with a small annotated data set in practice and get more effective with new EMRs over time. Evaluated with the i2b2 data set in comparison with the related works, our solution is effective with better F-measure values for the AGE, LOCATION, and PHONE PHI types and comparable for the other PHI types.**

Keywords—de-identification; protected health information; electronic medical record; privacy preserving; multilevel hybrid semi-supervised learning

I. INTRODUCTION

Medical records have been prepared and archived worldwide with time. In order to exploit them for more health care and medical research activities, their electronic versions have been produced and need to be de-identified for privacy preservation outside their associated institutions. Private information that must be protected is called protected health information (PHI). Therefore, reliable and inexpensive de-identification techniques are required to give easy access to electronic medical records (EMRs).

Aware of this necessity, many existing works since about 1995 have been proposed. In particular, a diversity of de-identification systems on many clinical document types have been defined in [1, 3-6, 8, 10, 14-18]. At the early stage, De-Id [6] and Scrub [14] were among the first de-identification systems which were rule-based. Another rule-based system was introduced in [10] using pattern matching with look-up tables, regular expression and heuristics. After that, [5] introduced HIDE, a de-identification system based on conditional random fields (CRF) models. Also using CRF models, the MIST system in [1] was developed and became one of the well-known de-identification systems with Carafe, a CRF-based sequence tagger and a tag-a-little, learn-a-little approach. After that, CRF models were also utilized in [18] that developed Anonym, a de-identification tool using linguistic and lexical features combined with pattern matching for feature extraction. Up to now, a hybrid approach has been proposed by combining a machine learning-based method and a rule-based method. A typical system following such a hybrid approach is a best-of-breed system called BoB introduced in [4]. Moreover, the two shared tasks of i2b2 (Informatics for Integrating Biology and the Bedside, [7]) in 2006 and 2014 were organized. [15, 16] are the works with the highest results in the 2006 i2b2 shared task about de-identification of discharge summaries while [3, 8, 17] with the highest results in the 2014 i2b2 shared task about de-identification of longitudinal clinical narratives. In these works, a hybrid approach was also applied and most of them except [15] made the most of the CRF models for PHI identification. More details about the 18 works proposed in 1995-2010 and the 10 works with the highest results in the 2014 i2b2 shared task can be found in [9] and [13], respectively.

Unfortunately as mentioned in [13], the de-identification task has not yet been a solved problem and still needs more investigation realistically, even though the existing works have produced positive research outcomes for PHI identification. In this paper, we propose a more general and practical solution to

978-1-5090-4134-3/16/$31.00 ©2016 IEEE

identifying PHI instances in EMRs and further to facilitating an automatic de-identification process on the EMRs. As compared to the aforementioned works, our proposed solution is novel in a multilevel hybrid semi-supervised learning paradigm with a key focus on correct PHI identification in the EMRs. It also follows a hybrid approach by combining a machine learning-based method with a CRF model and a rule-based method in a post-processing phase to handle the PHI types with disambiguation. Nevertheless, it is different from the existing works in the following points. First, it checks the structure complexity of each EMR that needs to be processed for PHI identification. Each section of an EMR will be treated properly and differently up to its structure complexity: structured, semi-structured, or un-structured. Second, each EMR is then examined in our approach at three different levels of granularity such as a token level in the supervised learning phase, an entity level in the rule-based post-processing phase, and a section level along with the structure complexity in the semi-supervised learning phase. Many various detail levels will give our solution a closer look at each EMR for more effectiveness. Third, our solution is conducted in a self-training manner so that it can get started with a small annotated data set in practice and get more effective with new EMRs over time. Evaluated with the standard i2b2 data set, our solution is confirmed to be effective with better F-measure values for the AGE, LOCATION, and PHONE PHI types and comparable for the other PHI types.

II. Automatic De-identification of Medical Records

Before detailing our proposed solution, we define briefly a de-identification process on EMRs composed of three main sequential phases: *A. Preprocessing*, *B. PHI identifying*, and *C. PHI processing*. These phases are elaborated below. As sketched in Fig. 1, the process receives an input of EMRs and returns an output of de-identified EMRs which are free of protected health information (PHI). A PHI list is specific for each country's requirements and regulations. For reference, the PHI type list in the i2b2 shared task in 2006 [15, 16] includes 8 PHI types (AGE, DATE, DOCTOR, HOSPITAL, ID, LOCATION, PATIENT, PHONE), the expanded PHI type list in the task in 2014 [13] includes 7 main PHI categories (NAME, PROFESSION, LOCATION, AGE, DATE, CONTACT, ID) and their subcategories. In our work, a PHI type list is given and our approach is not limited to this list.

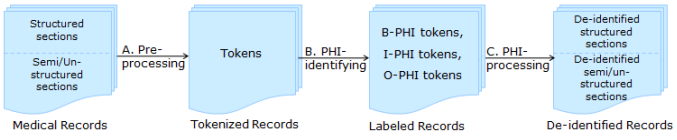

Fig. 1. A de-identification process on electronic medical records

A. Preprocessing

In this phase, all the electronic medical records are inputted and preprocessed for the de-identification task. Tokenization is invoked to return a collection of tokens in sequence according to their occurrences in each record. Each token of a PHI type is then represented by means of a feature set capturing its characteristics so that it can be distinguished from other tokens of other PHI types. If a token in an annotated record in a given training data set, its PHI type is known. Otherwise, its PHI type needs to be determined. In addition, we further consider what kind of sections in the record being preprocessed the token belongs to. The structure complexity of each section in a record is examined so that an appropriate approach in the next phase can be defined and utilized. In our work, we divide each record into two parts: the first part composed of structured sections which are well defined in the form of "header: values" and the second part consisting of semi/un-structured sections which are complex and nearly full of free-text. Taking into account the structure complexity of each record does not restrict our approach to the structure of the records being processed as every record always includes several sections each of which is structured, semi-structured, or unstructured. The output of this phase is then passed to the next phase, *PHI identifying*.

B. PHI identifying

For PHI identification, we formulate this phase as a multi-class classification task to predict a PHI type of each token from the previous phase to be the beginning token of a PHI instance, the inside token of a PHI instance, or not a part of any PHI instance. As discussed in [9], a machine learning-based method or a rule-based method or a hybrid method can be considered for this PHI identifying phase. Different from several existing works [3, 8, 15-18] based on a supervised learning approach, our work defines a semi-supervised learning approach to eliminate the users from a large labeled data set prior to the learning phase. Moreover, new records that are regarded to be the most correctly predicted can be utilized to enhance our current training data set over time. Besides, our PHI identifying phase examines the structure of each record and the various levels of granularity for more effectiveness. Based on the structure complexity of each record section, this phase treats its tokens in an appropriate way.

C. PHI processing

PHI processing in the last phase is performed on each section of a record so that every section regardless of its structure complexity can be de-identified. PHI instances are formed and replaced with surrogate values (identifier, date, name, address, etc.) corresponding to each PHI type. The resulting de-identified records are then ready for being shared.

III. Towards Automatic De-identification of EMRs in a Multilevel Hybrid Semi-supervised Learning Paradigm

In this section, we present our effective solution in detail, which takes into account the structure complexity of each record and many various levels of granularity in a hybrid semi-supervised learning paradigm.

A. PHI Identifying in EMRs in a Multilevel Hybrid Semi-supervised Learning Paradigm

A hybrid semi-supervised learning approach was proposed for PHI identification in [11]. Different from [11], this work has a deeper look at each record at many various levels of granularity and its structure complexity for more effectiveness. In particular, we define a novel multilevel hybrid semi-

supervised learning approach to PHI identifying in EMRs towards automatic de-identification of EMRs over time.

Like many recent works, our approach is a ***hybrid*** one taking advantage of both machine learning-based and rule-based methods. More generalized, our approach is defined with a ***semi-supervised learning*** mechanism. This mechanism enables our work to identify PHI instances in the free-text medical records in an incremental and iterative manner. As a result, we can reach a more powerful CRF-based PHI identifier over time for practical use. This is because in practice, a large annotated training data set might be unavailable. Starting with a small annotated set must be effective in a self-training context. It is more significant as our methodology is adapted to the free-text medical records written in other languages or the medical contexts with many different clinical note types where there exist no or few annotated electronic medical records.

As a ***multilevel*** approach, our proposed approach considers three aforementioned levels: token, entity, and section levels. ***The token level*** plays an important role in the PHI identifying phase of the de-identification process because we regard the PHI identification task as a multi-class classification task on a collection of tokens in each record. The classes in the multi-class classification task are B-PHI and I-PHI classes corresponding to each PHI type, O-PHI class corresponding to the class of all non-PHI instances in the beginning-inside-outside (BIO) scheme. Thus, the token level is used for the supervised learning phase of the multi-class classification task. ***The entity level*** is further used in the rule-based post-processing phase so that our PHI identification can be directed to the PHI processing phase of the de-identification process. This entity level along with the BIO scheme helps us form both single-token entities and multiple-token entities of the PHI types in the records for de-identification. Hence, the entity level is significant to support the de-identification process in its entirety instead of only PHI identification done in some existing works such as [11]. Above all, ***the section level*** has been added to exploit the structure complexity of each record for PHI identification in a generalized manner. In this section level, our work examines structured sections and semi/un-structured sections separately and provides an appropriate learning approach according to their structure complexity for more effectiveness and efficiency. In particular, a less complicated hybrid supervised learning approach is defined for PHI identification of structured sections in the electronic medical records. Its results are then used for enhancing the training data set in a more complicated hybrid semi-supervised learning approach to PHI identification of semi/un-structured sections in the same electronic medical records. The rationale behind these different approaches is that there are well-defined extraction patterns and hidden rules for structured sections, leading to less effort to handle PHI identification of such structured sections. This implies no need for an incremental and iterative self-training mechanism. In contrast, semi/un-structured sections are complex parts in the records that need more investigation with a semi-supervised learning approach in an incremental and iterative manner. Using many different levels of granularity, our work enables the PHI identifying phase to be conducted more effectively.

1) PHI identification of structured sections: For structured sections, we define a hybrid supervised learning approach that combines a supervised learning-based method and a rule-based method. This approach is widely used in many recent works [3, 15-17]. The main differences between them are their feature sets and learning algorithms, the use of regular expressions and dictionaries, and post-processing rules.

In Fig. 2, there are three subprocesses in our approach: (1). ***Supervised learning with CRFs and k-fold cross validation***, (2). ***PHI identifying***, and (3). ***Rule-based post-processing***. In the first subprocess, a CRFs model is built and evaluated with the k-fold cross validation scheme with more reliability and over-fitting avoidance. The supervised learning process uses all the tokens of the labeled records regardless of their structure complexity. The resulting CRFs model is then used as a PHI identifier to predict a PHI type of each token in structured sections of unlabeled records in the second subprocess. Each predicted token is associated with a conditional probability returned by the CRFs model. A resulting collection of B-PHI tokens, I-PHI tokens, and O-PHI tokens is post-processed in the third subprocess using the following post-processing rules. Our rule-based post-processing procedure is carried out to further examine and improve the result of the CRF-based PHI identifier by extracting more PHI tokens missed and filtering out PHI tokens mislabeled. In our work, we favor recall to obtain as many true PHI tokens from the records as possible.

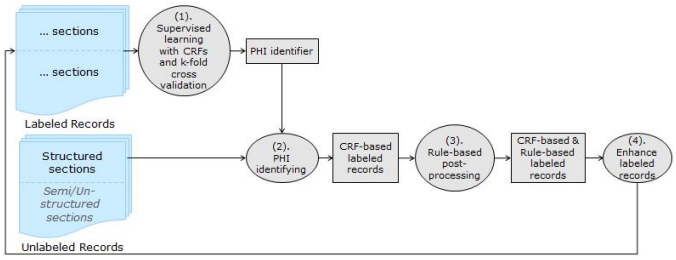

Fig. 2. PHI identification of structured sections in electronic medical records in a hybrid supervised learning approach

Rule 1: Change O-PHI tokens to B-LOCATION tokens and I-LOCATION tokens using the extraction pattern: [No] [Name] [Blvd/Road/St/…] [,] [StateName or its Abbreviation] where No is B-LOCATION and the others are I-LOCATION.

Rule 2: Change O-PHI tokens to B-PATIENT tokens and I-PATIENT tokens using the extraction pattern [SUMMARY NAME :] [PatientName_1] [PatientName_2] where PatientName_1 is B-PATIENT and the other name PatientName_2 is I-PATIENT.

Rule 3: Change O-PHI tokens to B-DATE tokens using the regular expressions: d/mm or dd/mm or dd/m. Similarly, change O-PHI tokens to B-DATE tokens using the regular expressions: d/mmm or d/mmmm or dd/mmm or dd/mmmm.

Rule 4: Change O-PHI tokens to B-DATE and I-DATE tokens using the extraction patterns: [dd] [mmm] or [mmm] [dd] or [dd] [mmmm] or [mmmm] [dd] where the first part is B-DATE and the other is I-DATE.

Rule 5: Change O-PHI tokens to B-HOSPITAL tokens using the extraction patterns: [presented to] [Name], [transferred to] [Name], [transferred from] [Name], [admitted to] [Name], [discharged to] [Name], [hospitalization at] [Name], [admission to] [Name], etc. where Name is B-HOSPITAL.

Rule 6: Change B-PHI and I-PHI tokens to O-PHI tokens using pattern matching with all words that are medical values, medical terms, days in a week, seasons in a year, titles, etc.

2) PHI identification of semi/un-structured sections: More sophisticated, a multilevel hybrid semi-supervised learning approach is defined with six main subprocesses in Fig. 3 for PHI identification of semi/un-structured sections in EMRs. They are: (1). ***Supervised learning with CRFs and k-fold cross validation***, (2). ***PHI identifying***, (3). ***Rule-based post-processing***, (4). ***Records selecting with the most confident prediction***, (5). ***Update unlabeled records***, and (6). ***Enhance labeled records***. These subprocesses are detailed as follows.

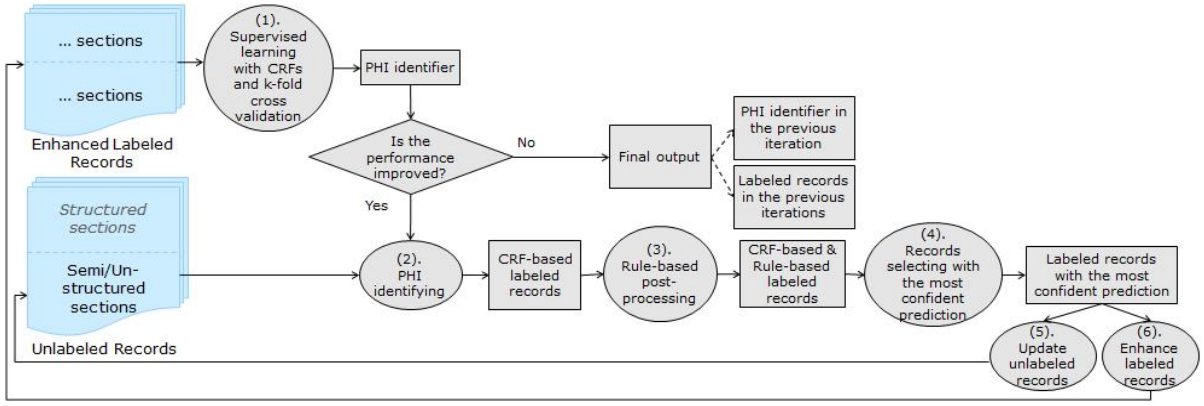

Fig. 3. PHI identification of semi/un-structured sections in electronic medical records in a multilevel hybrid semi-supervised learning approach

(3). ***Rule-based post-processing***: In this phase, the post-processing subprocess is performed with the following rules in order. This subprocess will discover more PHI instances incorrectly predicted by the CRF-based PHI identifier to increase recall and also exclude PHI instances misrecognized by the CRF-based PHI identifier to increase precision.

Rule 1. Change O-PHI tokens to B-PHI and I-PHI tokens up to a conditional probability of each O-PHI token less than a threshold automatically defined in the k-fold cross validation scheme of the current PHI identifier based on its F-measure.

Rule 2. Change O-PHI tokens to B-PHONE tokens and I-PHONE tokens using regular expressions and the following extraction pattern [Phone] [,] [b/e/g] [Phone].

Rules 3 and 4: Change O-PHI tokens to B-DATE tokens and I-DATE tokens using the corresponding rules *Rules 3 and 4* in the previous subsection for PHI identifying structured sections.

Rule 5: Change O-PHI tokens to B-DOCTOR tokens and I-DOCTOR tokens that is located after a clue [TR].

Rule 6: Change O-PHI tokens to B-LOCATION tokens and I-LOCATION tokens using *Rule 1* in the previous subsection.

Rule 7: Change O-PHI tokens to B-HOSPITAL tokens using *Rule 5* in the previous subsection.

Rule 8: Change B-AGE and I-AGE tokens to O-PHI tokens if their values are less than 89.

Rule 9: Change B-PHI and I-PHI tokens to O-PHI tokens using *Rule 6* in the previous subsection.

(4). ***Records selecting with the most confident prediction***: The CRF-based and rule-based labeled records with the most confident prediction are selected to enhance the current training data set of the current PHI identifier. The feature values of the selected records are then reexamined for consistency. Similar to the scores in [11], our scores are based on the conditional probability of each PHI instance in the record and the prediction power of the current CRF-based PHI identifier that was used to predict each token. However, our scores are defined at the section level instead of the record level.

(5). ***Update unlabeled records***: The set of unlabeled records is updated by removing the records with the most confident prediction selected in the previous phase.

(6). ***Enhance labeled records***: In this phase, the current training data set is enhanced with the records with the highest confidence selected in phase (4). This enhancement makes the training data set for a PHI identifier larger and more generalized. If our selected records are really truly labeled records, a new PHI identifier built on such an enhanced

(1). ***Supervised learning with CRFs and k-fold cross validation***: In this phase, we build a CRF-based PHI identifier using the set L of labeled records as a training data set. Different from the supervised learning phase in Fig. 2, L includes not only given labeled records but also all the labeled sections of the unlabeled records that have been processed. This enhanced set L of labeled records will provide more information for the starting point of the current supervised learning phase. In addition, the resulting CRF-based PHI identifiers in the later iterations are required to have averaged recall higher than that of the identifier in the previous iteration.

(2). ***PHI identifying***: In this phase, the current PHI identifier is applied to predict a PHI type of each token in the set U of the semi/un-structured sections of unlabeled records. The output of this subprocess is a set U of predicted tokens.

training data set can get more accurate. Unfortunately, the accumulation of errors from the mislabeled PHI instances in the selected records over time will happen. This error accumulation is a risk in a self-training mechanism. However, our approach can diminish the influences of the "wrongly" selected records by checking the prediction power of the new PHI identifier with that of its previous one. If the performance gets improved, the new PHI identifier is accepted and used for the next iteration. Otherwise, the previous one is remained and used for future prediction. Furthermore, the selected records are reconsidered prior to enhancement in phase (4). Thus, the effects of errors in the mislabeled records can be avoided.

As compared to a similar approach in [11], our approach has a closer look at the unlabeled records in terms of structure and level of details at the different subprocesses. In particular, the structure of an unlabeled record is examined for proper treatment. Three different token, entity, and section levels are considered to support automatic de-identification. Besides, our approach is more practical because it can be automated with no parameter choices in the post processing rules.

B. Towards Automatic De-identification of EMRs

As soon as each token in an EMR is identified to be an instance of a PHI type or not to be in the BIO scheme, an automatic procedure is defined using a two-token sliding window over an EMR to form a PHI instance at the entity level. This procedure can extract all PHI instances at the entity level in time complexity O(n) where n is the number of tokens in the EMR. Each PHI instance at the entity level is then replaced with a surrogate value according to each PHI type (AGE, DATE, DOCTOR NAME, PATIENT NAME, HOSPITAL NAME, ID, etc.). This procedure along with the PHI identification process lays the foundations for our automatic PHI de-identification of the EMRs and further for enhancing their sharability and exploitation.

IV. EXPERIMENTAL RESULTS

For an illustration of the effectiveness of our proposed work, several experiments were performed on the standard set of the electronic medical records published by Informatics for Integrating Biology and the Bedside (i2b2) [7] in the 2006 shared task about de-identification. The reasons for using this 2006 data set instead of the 2014 data set are given as follows: (1). Although the 2014 data set is more complex than the 2006 one, these two data sets are prepared from the Partners HealthCare and similar to each other to some extent. (2). We would like to evaluate our approach in comparison with more recent works both taking part in the shared task and not such as [15, 16] and [11, 18], respectively. (3). There are few works on the 2014 data set except for those participating in the 2014 shared task because the 2014 data set has just been published.

Using the 2006 data set, we remain the standard training data set and the standard test data set from the shared task. Each test record is divided into structured and semi/un-structured section in our approach based on the presence of the phrase "HISTORY OF PRESENT ILLNESS". The part before this phrase is regarded as the first structured section. All the sentences starting with "TR: ", "DD ", and "TD " consecutively forms another structured section. The rest is considered to be semi/un-structured sections. These sections are treated with the appropriate procedures defined previously. For implementation, our program was written in C#, making use of the existing Stanford natural language processing tools at [12] and the CRFSharp toolkit at [2]. We also carry out the experiments on a server machine Intel(R) Xeon(R) CPU E5-2620 0 @2.00GHz with 96 GB RAM using MS Windows.

TABLE I. PRECISION VALUES FOR EACH PHI TYPE

PHI types	[15], 2007	[16], 2007	[18], 2014	[11], 2016	MLHSLA
AGE	**100**	**100**	Not available	0	**100**
DATE	99.54	98.81	**99.67**	98.31	98.53
DOCTOR	99.02	**99.43**	98.6	98.34	98.33
HOSPITAL	98.18	**99.08**	98.8	96.53	95.54
ID	99.25	99.74	**99.83**	98.92	98.99
LOCATION	85.91	90.24	**96.43**	90.05	91.44
PATIENT	98.62	99.16	**99.6**	98.4	98.23
PHONE	**100**	**100**	**100**	**100**	**100**

TABLE II. RECALL VALUES FOR EACH PHI TYPE

PHI types	[15], 2007	[16], 2007	[18], 2014	[11], 2016	MLHSLA
AGE	**100**	66.67	Not available	0	**100**
DATE	**99.27**	98.71	98.1	97.13	99.26
DOCTOR	96.23	98.13	97.71	97.15	**98.58**
HOSPITAL	94.74	**95.41**	90.93	85.21	89.04
ID	99.17	**99.74**	99.58	98.83	98.42
LOCATION	54.24	62.18	67.5	**71.67**	71.25
PATIENT	**97.47**	96.33	96.88	96.1	**97.47**
PHONE	81.18	91.38	69.41	91.76	**94.12**

TABLE III. F-MEASURE VALUES FOR EACH PHI TYPE

PHI types	[15], 2007	[16], 2007	[18], 2014	[11], 2016	MLHSLA
AGE	**100**	80	Not available	0	**100**
DATE	**99.4**	98.76	98.88	97.72	98.89
DOCTOR	97.61	**98.78**	98.15	97.74	98.45
HOSPITAL	96.43	**97.21**	94.7	90.52	92.18
ID	99.21	**99.74**	99.7	98.87	98.7
LOCATION	66.49	73.63	79.41	79.8	**80.09**
PATIENT	98.04	97.72	**98.22**	97.24	97.85
PHONE	89.61	95.5	81.94	95.7	**96.97**

For evaluating the effectiveness with respect to each PHI type, Precision, Recall, and F-measure values of our approach are recorded and presented in Tables I-III while those of the existing works [11, 15, 16, 18] are gathered from their corresponding papers. In the resulting tables, we name our multilevel hybrid semi-supervised learning approach MLHSLA. Also for more readability, the best values are presented in bold and the second best values are underlined.

First of all, it is realized that our work obtained the best results for AGE and PHONE instances and the second best results for LOCATION instances via Precision values while [16, 18] had the highest Precision values for more PHI types. This is understandable as our work favors recall over precision so that many true PHI instances can be detected and extracted effectively. Indeed, our semi-supervised learning mechanism is iteratively executed till there is no improvement of Recall values. It is reflected well through our recall values in Table II. As compared to the results of the existing works, ours got the highest recall values for AGE, DOCTOR, PATIENT, and PHONE instances while comparable for DATE and

LOCATION instances. HOSPITAL and ID instances can be extracted well by [16]. Nevertheless, our work is remarkable for the best F-measure values for AGE, LOCATION, and PHONE instances, the second best F-measure values for DATE and DOCTOR instances, and comparable values for the instances of other PHI types with more than 92%. As compared to the results in the most recent work [11], which also follows a semi-supervised learning paradigm, our work has a significant improvement of F-measure values for almost all the PHI types except for ID instances. This implies the effectiveness of our new multilevel process with a special consideration on the structure of each record in terms of its complexity.

As stated in [13], "our findings indicate that automated systems can be very effective for this task, but that de-identification is not yet a solved problem." This conclusion encouraged us to make a continuous improvement in the research area so that automatic de-identification of the EMRs can be reached soon. Compared to the existing works except for [11], our work can not only provide an effective PHI identification solution but also direct the de-identification task towards a more general, practical and automatic hybrid solution over time in a self-training manner.

V. Conclusion

De-identifying protected health information from electronic medical records is significant to enable the electronic medical records to be shared and utilized for more research and development in health care and medical domains. Thus, in this paper, we have defined a practical solution towards this de-identification task on the electronic medical records. Our solution is effective in a multilevel hybrid semi-supervised learning paradigm as the resulting approach can take advantage of the machine learning-based approach and the rule-based approach in an iterative self-training manner with three various levels of granularity and a consideration on the structure complexity of each record. An appropriate approach can be determined separately for structured sections or semi/un-structured sections of a record. Moreover, each phase of the task can be enhanced at a certain level of granularity: token, entity, or section. As a result, our solution can be viewed as a generalized version of the existing hybrid solutions of the related works. Evaluated with the standard i2b2 data set in comparison with the existing works identifying the instances of 8 PHI types in the 2006 i2b2 data set, our solution is confirmed to be effective with better F-measure values for AGE, LOCATION, and PHONE PHI types and comparable for the others. However, our PHI identifier can be improved with the new medical records that have the most confident prediction to obtain a new PHI identifier with higher accuracy over time.

Acknowledgment

This work is funded by Vietnam National University at Ho Chi Minh City under the grant number B2016-42-01. Besides, we would like to thank John von Neumann Institute (JVN), Vietnam National University at Ho Chi Minh City, very much to provide us with a very powerful server machine to carry out the experiments. In addition, this paper was written when the second author was working at Vietnam Institute for Advanced Study in Mathematics (VIASM), Hanoi, Vietnam.

References

[1] J. Aberdeen, S. Bayer, R. Yeniterzi, B. Wellner, C. Clark, D. Hanauer, B. Malin, and L. Hirschman, "The MITRE identification Scrubber toolkit: design, training, and assessment," J. Med. Inform., vol. 79, pp. 849-859, 2010.

[2] CRFSharp Toolkit, https://crfsharp.codeplex.com/, 2015.

[3] A. Dehghan, A. Kovacevic, G. Karystianis, J. A. Keane, and G. Nenadic, G. 2015. Combining knowledge- and data-driven methods for de-identification of clinical narratives. J. Biomed. Inform., 2015.

[4] O. Ferrández, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre, "BoB, a best-of-breed automated text de-identification system for VHA clinical documents," J. Am. Med. Inform. Assoc., vol. 20, pp. 77-83, 2013.

[5] J. Gardner and L. Xiong, "HIDE: an integrated system for health information DE-identification," in Proc. of the 21st Int. Symp. On Computer-based Medical Systems, IEEE, 2008, pp. 254-259.

[6] D. Gupta, M. Saul, and J. Glbertson, "Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research," Am. J. Clin. Pathol., vol. 121, pp. 176-186, 2004.

[7] Informatics for Integrating Biology and the Bedside (i2b2), https://www.i2b2.org/NLP/, 2015.

[8] Z. Liu, Y. Chen, B. Tang, X. Wang, Q. Chen, H. Li, J. Wang, Q. Deng, and S. Zhu, "Automatic de-identification of electronic medical records using token-level and character-level conditional random fields," J. Biomed. Inform., 2015.

[9] S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: a review of recent research," BMC Medical Research Methodology, vol. 10, no. 70, pp. 1-16, 2010.

[10] 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, "Automated de-identification of free-text medical records," BMC Medical Informatics and Decision Making, vol. 8, no. 32, 2008.

[11] P. D. Nguyen, C. T. N. Vo, and B. T. Ho, "A hybrid semi-supervised learning approach to identifying protected health information in electronic medical records" in Proc. of the 10th ACM IMCOM, 2016, pp. 82:1-82:8.

[12] Stanford Natural Language Processing Tools, http://nlp.stanford.edu/software/corenlp.shtml, 2015.

[13] A. Stubbs, C. Kotfila, and O. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1," J. Biomed. Inform., 2015.

[14] L. Sweeney, "Replacing personally-identifying information in medical records, the Scrub system," in Proc. of the AMIA Annu. Fall Symp., pp. 333-337, 1996.

[15] G. Szarvas, R. Farkas, and R. Busa-Fekete, "State-of-the-art anonymization of medical records using an iterative machine learning framework," J. Am. Med. Inform. Assoc., vol. 14, pp. 574-580, 2007.

[16] B. Wellner, M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, A. Yeh, A., J. Hitzeman, and L. Hirschman, "Rapidly retargetable approaches to de-identification in medical records," J. Am. Med. Inform. Assoc., vol. 14, pp. 564-573, 2007.

[17] H. Yang and J. M. Garibaldi, "Automatic detection of protected health information from clinic narratives," J. Biomed. Inform., 2015.

[18] G. Zuccon, D. Kotzur, A. Nguyen, and A. Bergheim, "De-identification of health records using Anonym: effectiveness and robustness across datasets," Artificial Intelligence in Medicine, vol. 61, issue 3, pp. 145-151, July 2014.