Journal of Biomedical Informatics 58 (2015) S47–S52

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automatic de-identification of electronic medical records using token-level and character-level conditional random fields

Zengjian Liu [a,1], Yangxin Chen [b,1], Buzhou Tang [a,*], Xiaolong Wang [a], Qingcai Chen [a], Haodi Li [a], Jingfeng Wang [b], Qiwen Deng [c], Suisong Zhu [c]

[a] Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China
[b] Department of Cardiology, Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Guangzhou 510120, China
[c] The Sixth People's Hospital of Shenzhen, Shenzhen 518052, China

ARTICLE INFO

Article history:
Received 30 January 2015
Revised 2 June 2015
Accepted 9 June 2015
Available online 26 June 2015

Keywords:
De-identification
Protected health information
Electronic medical records
i2b2
Natural language processing
Hybrid method

ABSTRACT

De-identification, identifying and removing all protected health information (PHI) present in clinical data including electronic medical records (EMRs), is a critical step in making clinical data publicly available. The 2014 i2b2 (Center of Informatics for Integrating Biology and Bedside) clinical natural language processing (NLP) challenge sets up a track for de-identification (track 1). In this study, we propose a hybrid system based on both machine learning and rule approaches for the de-identification track. In our system, PHI instances are first identified by two (token-level and character-level) conditional random fields (CRFs) and a rule-based classifier, and then are merged by some rules. Experiments conducted on the i2b2 corpus show that our system submitted for the challenge achieves the highest micro *F*-scores of 94.64%, 91.24% and 91.63% under the "token", "strict" and "relaxed" criteria respectively, which is among top-ranked systems of the 2014 i2b2 challenge. After integrating some refined localization dictionaries, our system is further improved with *F*-scores of 94.83%, 91.57% and 91.95% under the "token", "strict" and "relaxed" criteria respectively.

1. Introduction

With the development of electronic medical records (EMRs), more and more clinical data are generated. However, they cannot be freely used by companies, organizations and researchers because of a large amount of personally identifiable health information, known as protected health information (PHI), embedded in them. Using clinical data containing PHI is usually prohibited. De-identification, identifying and removing PHI, is a critical step in making clinical data accessible to more people. Since the Health Insurance Portability and Accountability Act (HIPAA) was passed in 1996 completely defined all kinds of PHI [1], de-identification has attracted considerable attention. De-identification resembles traditional named entity recognition (NER) tasks, but has its own property such that a word/phrase can be either a PHI instance or not. During the last decade, a large amount of effort has been devoted to de-identification including a challenge, i.e., the i2b2 (Center of Informatics for Integrating Biology and Bedside) clinical natural language processing (NLP) challenge in 2006, and various kinds of systems have been developed for de-identification [2–5]. However, no unified platform to evaluate systems on any PHI type defined in HIPAA.

In order to comprehensively investigate the performance of de-identification systems on every HIPAA-defined PHI type, the 2014 i2b2 clinical natural language processing (NLP) challenge sets up a new track to identify PHI instances in electronic medical records (EMRs) (track 1). In this track, seven main categories with twenty-five subcategories are defined, which cover all eighteen PHI types defined in HIPAA. In this paper, we describe our de-identification system for the 2014 i2b2 challenge. It is a hybrid system based on both machine learning and rule approaches. Evaluation on the independent test set provided by the challenge shows that our system achieves the highest micro *F*-scores of 94.64%, 91.24% and 91.63% under the "token", "strict" and "relaxed" criteria respectively, which is among top-ranked systems of the 2014 i2b2 challenge. We subsequently introduce refined localization dictionaries into our system, and marginally improve

* Corresponding author.
E-mail addresses: liuzengjian.hit@gmail.com (Z. Liu), tjcyx1995@163.com (Y. Chen), tangbuzhou@gmail.com (B. Tang), wangxl@insun.hit.edu.cn (X. Wang), qingcai.chen@gmail.com (Q. Chen), haodili.hit@gmail.com (H. Li), dr_wjf@hotmail.com (J. Wang), qiwendeng@hotmail.com (Q. Deng), 13809883596@163.com (S. Zhu).
[1] Contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2015.06.009
1532-0464/

performance with micro *F*-scores of 94.83%, 91.57% and 91.95% under the "token", "strict" and "relaxed" criteria respectively.

2. Background

In the medical domain, many NLP approaches have been proposed for de-identification. The earliest de-identification system was proposed by Sweeney et al. in 1996 [6]. This system employed rules to identify twenty-five categories of personally-identifying information in pediatric EMRs. In the same year, the HIPAA was passed, and defined eighteen types of PHI. Subsequently, a large number of pattern matching-based systems were introduced for de-identification based on HIPAA. These systems used complex rules [7–12] and specialized semantic dictionaries [7,9,10,12] to perform de-identification. Most of them de-identified PHI in their own particular types of EMRs. For example, three systems were designed only for pathology reports [8–10]. Two systems were designed for multiple types of EMRs: Friedlin et al.'s [11] system for clinical notes including discharge summaries, laboratory reports and pathology reports, and Neamatullah et al.'s [12] system for nursing progress notes, discharge summaries and X-ray reports. Some pattern matching-based systems have been able to find around 99% PHI instances on their own datasets as reported [7,8,10,11]. However, we could not find which one is better due to no unified evaluation on publicly available datasets.

To accelerate de-identification research in the medical domain, the 2006 i2b2 clinical natural language processing (NLP) challenge issued a track to identify PHI in EMRs, which provided a unified platform to evaluate different systems. In this challenge, eight PHI categories were defined to annotate the challenge data from Partner Healthcare, only six HIPAA-defined categories. Seven teams participated in the challenge and developed de-identification systems using rule-based [13], machine learning-based [14–16] and hybrid methods [17,18]. Results showed that machine learning-based systems using rules as features performed best [2]. The machine learning algorithms used in these systems included conditional random fields (CRFs) [19], support vector machines (SVM) [20], decision trees (DTs) [21], and so on. Considering that all the documents used in this challenge were discharge summaries not annotated with all HIPAA-defined categories of PHI instances, Deleger et al. (2013) [5] evaluated a machine learning-based system using rules as features on various types of notes (over 22 types) annotated with all HIPAA-defined categories, although some of HIPAA-defined categories were collapsed into one category.

To further advance de-identification research in the medical domain, the 2014 i2b2 clinical NLP challenge organizers set up a track (track 1) to identify PHI in EMRs again. Different from the previous de-identification challenge, more refined PHI categories were annotated in the data provided by the organizers of this challenge, which makes it possible to evaluate all participating systems on every HIPAA-defined PHI type.

3. Material and methods

Fig. 1 shows an overview of our de-identification system for the 2014 i2b2 NLP challenge. It is a hybrid system based on both machine learning and rule approaches. The system contains two machine learning-based classifiers and a rule-based classifier. Similar to traditional NER tasks, the de-identification task is recognized as a sequence labeling problem in both two machine learning-based classifiers. In our system, PHI instances are first identified by two (token-level and character-level) conditional random fields (CRFs) and a rule-based classifier, and then are merged by some rules. The detailed description of the system is presented below.

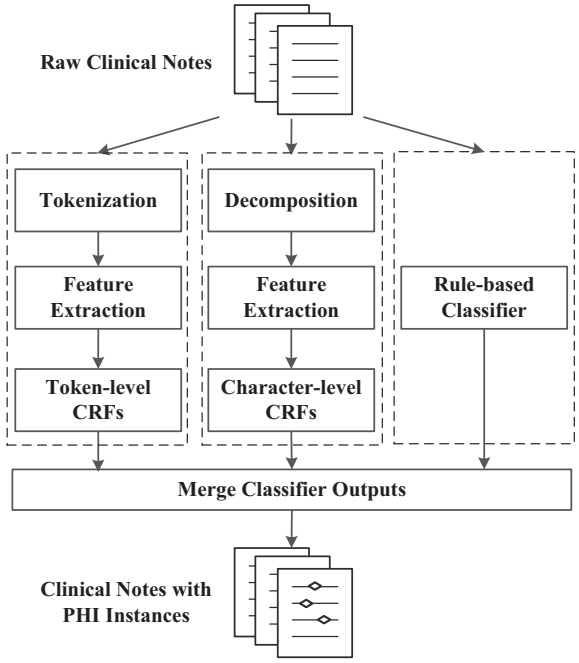

Fig. 1. Overview of our de-identification system for the 2014 i2b2 NLP challenge.

3.1. Dataset

In the 2014 i2b2 challenge, organizers manually annotated 1304 medical records of 297 patients according to the annotation guideline, and divided them into two parts: (1) 790 records of 188 patients used as a training set; and (2) the remaining 514 records of 109 patients used as a test set. 17,045 PHI instances in the training set and 11,462 PHI instances in the test sets are annotated using seven main categories with twenty-five subcategories that cover all HIPAA-defined PHI categories. The numbers of PHI instances of main categories in both two sets are listed in Table 1, where NA denotes no subcategory, numbers in parentheses in the first row are the numbers of categories and PHI instances, and asterisks indicate the HIPAA-defined categories. To get more detailed information of the dataset, please refer to the overview paper [22,23].

3.2. Machine learning-based classifiers

There are two (token-level and character-level) machine learning classifiers in our de-identification system, and both trained by conditional random fields (CRFs) algorithm. The main difference between those two classifiers is the representation of features. We use CRFsuite (http://www.chokkan.org/software/crfsuite/) as the implementation of CRFs, and optimize parameters of the two machine learning classifiers by 10-fold cross-validation on the training set.

3.2.1. PHI instance representation

How to represent PHI instances is the chief problem we should solve in machine learning-based de-identification systems. In our system, two typical NER representation schemas are used to represent PHI instances: "BIO" and "BIOES", where 'B', 'I', 'O' and 'E' denote that a token/character is at the beginning, middle, outside and end of an instance, and 'S' denotes that a token/character itself is an instance. Fig. 2 shows examples of PHI instances represented by "BIO" and "BIOES" at token-level. The PHI instances are represented in the similar way at character-level. Our evaluation shows

Table 1
Number of PHI instances of main categories in the training and test sets.

Main category (7)	Sub category (25)	Training set (17,405)	Test set (11,462)
NAME	PATIENT*, DOCTOR USERNAME	4465	2883
PROFESSION	NA	234	179
LOCATION	HOSPITAL, COUNTRY ORGANIZATION*, ZIP* STREET*, CITY*, STATE LOCATION-OTHER	2767	1813
AGE*	NA	1233	764
DATE*	NA	7502	4980
CONTACT	PHONE*, FAX*, EMAIL* URL, IPADDR	323	218
ID	MEDICALRECORD*, SSN* ACCOUNT*, LICENSE* DEVICE*, IDNUM*, BIOID* HEALTHPLAN*, VEHICLE*	881	625

that "BIO" performs better than "BIOES" in the token-level CRFs, while "BIOES" performs better in the character-level CRFs. Therefore, we use "BIO" in token-level CRFs and "BIOES" in character-level CRFs in our system.

3.2.2. Token-level CRFs

Similar to most machine learning-based de-identification systems, the token-level CRFs requires a tokenization module at first. We use the tokenization module of MedEx [24] (https://code.google.com/p/medex-uima/downloads/list), a specific tool for medical information extraction, for tokenization. After tokenization, we extract the following features for the token-level CRFs.

- Bag-of-words: The unigrams, bigrams and trigrams of words (i.e., tokens) within a window of [−2,2].
- Part-of-speech (POS) tags: The POS unigrams, bigrams and trigrams within a window of [−2,2]. We use the Stanford POS Tagger [25] (http://nlp.stanford.edu/software/tagger.shtml) for POS tagging.
- Combinations of tokens and POS tags: $\{w_0p_{-1}, w_0p_0, w_0p_1, w_0p_{-1}p_0, w_0p_0p_1, w_0p_{-1}p_1, w_0p_{-1}p_0p_1\}$, where w_0 denotes the current word, and p_{-1}, p_0 and p_1 denote the last, current and next POS tags respectively.
- Sentence information: The length of the sentence containing the word, whether there is an end mark at the end of the sentence such as '.', '?' and '!'. Whether there is any bracket unmatched in the sentence.
- Affixes: All prefixes and suffixes of length from 1 to 5.
- Orthographical features: Form information about the word (whether the word is upper case, contains a digit or not, has uppercase characters inside, has punctuation marks inside, has digit inside, the word is Roman or Arabic number, etc.)
- Word shapes: Two typical types of word shapes: one is generated by mapping any uppercase character, lowercase character, digit and other character in the word to 'A', 'a', '#' and '-' respectively, while the other one is generated by mapping consecutive uppercase characters, lowercase characters, digits and other characters to 'A', 'a', '#' and '-' respectively. For instance, the two types of word shapes of "PO/5mg" are "AA-#aa' and "A-#a".
- Section information: We extract twenty-nine section headers from the training set manually such as "family history" and check which section the word belongs to.
- General NER information: The NER tag of the word generated by the Stanford Named Entity Recognizer [26] (http://nlp.stanford.edu/software/CRF-NER.shtml).
- Word representation features: We follow the previous studies [27–30] to generate two types of word representation features using Brown clustering [31] (https://github.com/percyliang/brown-cluster) and word2vec [32] (https://code.google.com/p/word2vec/).
- Dictionary features: We collect four categories of localization dictionaries: COUNTRY, STATE, CITY and ZIP from internet, and label each token with "BIOES" tags through dictionary lookup. The tokenized sentence "Mary was born in Mississippi, and is currently living with daughter in Grand Island.", for example, is labeled as "Mary/O was/O born/O in/O Mississippi/S-STATE,/O and/O is/O currently/O living/O with/O daughter/O in/O Grand/B-CITY Island/E-CITY. /O". The labels are dictionary features. These features are not included in our system submitted to the challenge, but are subsequently added to our system after the challenge.

3.2.3. Character-level CRFs

To avoid boundary errors caused by token-level CRFs, we also use character-level CRFs to extract PHI instances. Firstly, we split raw clinical notes into sentences by '|n', and then decompose the sentences into characters. During decomposition, white spaces and '|t's are replaced by "&#", and all characters (including spaces and '|t's) are separated by white spaces. An example of decomposition is shown in Fig. 3 (see lines 2 and 3).

The features used in the character-level CRFs include bag-of-characters ([−5,5]), POS tags ([−5,5]), sentence information, section information, general NER information, word representation and dictionary features. The bag-of-characters include unigrams, bigrams and trigrams within a window of [−5,5]. The sentence information and section information are the same as those mentioned in the last section. The other features are generated in the similar way using corresponding information mentioned in the last section. Take POS tags as an example, given a raw fragment of clinical text "Mary was born …", tagged as "Mary/NNP was/VBD born/VBN …" by the Stanford Tagger, the POS features of all characters are "M/B-NNP a/I-NNP r/I-NNP y/E-NNP &#/O w/B-VBD a/I-VBD s/E-VBD &#/O b/B-VBN o/I-VBN r/I-VBN n/E-VBN …" (see lines 4 and 5 in Fig. 3).

3.3. Rule-based classifier

Considering that some categories of PHI instances are formulaic such as PHONE, FAX, MEDICAL RECORD, EMAIL and IPADDR, we define specific rules to recognize most of them. The detail regular expressions of those categories are listed in Table 2.

3.4. Merging PHI instances

After all the results of above three classifiers are generated, we design a simple strategy to merge them. In this strategy, PHI instances not overlapping with any other are directly merged, while PHI instances overlapping with others are selected in the order: the rule-based classifier, the character-level CRFs and the token-level CRFs. Given two overlapping PHI instance A and B generated by the rule-based classifier and the character-level CRFs respectively, for example, A is selected.

3.5. Evaluation

All our evaluations are performed on the independent test data set using the evaluation tool provided by the i2b2 organizers. The tool outputs macro/micro-average precisions (P), recalls (R), and F-scores (F) under six criteria: "token", "strict", "relaxed", "HIPAA token", "HIPAA strict" and "HIPAA relaxed". "token" checks whether a predicted token exactly matches a token in a gold

Tokenized sentence: Mary was born in Mississippi , and is currently living with daughter in Grand Island .

BIO representation: Mary/**B-PATIENT** was/O born/O in/O Mississippi/**B-STATE** ,/O and/O is/O currently/O living/O with/O daughter/O in/O Grand/**B-CITY** Island/**I-CITY** . /O

BIOES representation: Mary/**S-PATIENT** was/O born/O in/O Mississippi/**S-STATE** ,/O and/O is/O currently/O living/O with/O daughter/O in/O Grand/**B-CITY** Island/**E-CITY** . /O

Fig. 2. Examples of PHI instances represented by "BIO" and "BIOES" at token-level.

Raw fragment: Mary was born ⋯

Decomposition: M a r y &# w a s &# b o r n ⋯

BIOES representation: M/**B-PATIENT** a/**I-PATIENT** r/**I-PATIENT** y/**E-PATIENT** &#/O w/O a/O s/O &#/O b/O o/O r/O n/O ⋯

POS tags: Mary/NNP was/VBD born/VBN ⋯

POS features: M/**B-NNP** a/**I-NNP** r/**I-NNP** y/**E-NNP** &#/O w/**B-VBD** a/**I-VBD** s/**E-VBD** &#/O b/**B-VBN** o/**I-VBN** r/**I-VBN** n/**E-VBN** ⋯

Fig. 3. An example of decomposition in character-level CRFs.

Table 2
Regular expressions used in our system.

PHI category	Regular expression	Remark
PHONE	\(|\d{3}\)\[- |t\]?|d{3}\[- |t\]|d{4} |d{3}\[- |t\]|d{3}\[- |t\]|d{4} |d{3}||d{1}-|d{4}	(871) 720-9439 171-289-0968 659-5187
FAX	[Ff]ax.*|d{3}\[- |t\]|d{3}\[- |t\]|d{4}	Fax: 648-875-5821
MED RECORD	|d{3}\[- \]|d{2}\[- \]|d{2}\[- \]|d{1} |d{3}\[- \]|d{2}\[- \]|d{2}	453-39-84-4 544-84-52
EMAIL	[|w|d_]+@[|w|d]+|. (com—org)	gmichael@kcm.org
IPADDR	|d{1,3}|.|d{1,3}|.|d{1,3}|.|d{1,3}	198.168.2.78

phrase of the same category. "strict" denotes that a PHI instance is correctly extracted only when it exactly matches with a gold one of the same boundary and category, while "relaxed" denotes that a PHI instance is correctly extracted only when it mostly overlaps (only allow two characters mismatched at the end) with a gold one of the same category. "HIPAA" only considers eighteen types of HIPAA-defined PHI instances. Among the six criteria, "strict" is the primary one.

4. Results

For this task, a participating team could submit three runs. The best run is used for participating system ranking in the challenge. Table 3 shows the results of our best run submitted to the 2014 i2b2 challenge.

The best micro *F*-scores are 94.64%, 91.24%, 91.63%, 96.62%, 94.09% and 94.55% under the "token", "strict", "relaxed", "HIPAA token", "HIPAA strict" and "HIPAA relaxed" criteria respectively. The micro *F*-scores of all PHI instances are lower than HIPAA-defined PHI instances by about 2.0%. The differences between "strict" and "relaxed" micro *F*-scores are around 0.4% no matter all PHI instances or HIPAA-defined PHI instances are considered. However, the differences between "token" and "strict" micro *F*-scores exceed 2.5%, and the differences between "token" and "relaxed" micro *F*-scores are around 2.0%. Under any criterion, our system's recall is lower than precision by over 1.5%.

In order to evaluate the contribution of each classifier, we compare the performances of the systems using one or two classifiers under the "strict" criterion as shown in Table 4.

Table 3
Our system's best results submitted to the 2014 i2b2 challenge.

Criterion	Macro-average (%)			Micro-average (%)		
	Precision	Recall	*F*-score	Precision	Recall	*F*-score
Token	95.50	93.74	94.61	95.64	93.66	94.64
Strict	92.82	90.91	91.85	92.64	89.88	91.24
Relaxed	93.13	91.20	92.15	93.03	90.26	91.63
HIPAA token	97.17	95.58	96.37	97.48	95.78	96.62
HIPAA strict	95.04	93.59	94.31	95.13	93.07	94.09
HIPAA relaxed	95.46	94.00	94.72	95.59	93.52	94.55

Among the three classifiers, the two machine learning-based classifiers outperform the rule-based classifier, the character-level CRFs (Character) achieves better performance than the token-level CRFs (Token). The micro *F*-score of the character-level CRFs is 90.51%, higher than the token-level classifier by 6.81%. When using any two classifiers, our system shows better performance than only using any of them. For example, when using the two machine learning-based classifiers, the micro *F*-score of our system is 91.04%, superior to the token-level CRFs by 7.34% and character-level CRFs by 0.53%. When using a machine learning-based classifier and the rule-based classifier, the micro *F*-score of our system is better than the machine learning-based classifier by about 0.2%. When all the three classifiers are used, our system are further improved with the best micro *F*-score of 91.24%.

For the two machine learning-based classifiers, we also investigated the effect of some types of features on them by removing each type of features from all features. Six types of features are selected, and the results are shown in Table 5. Each type of features makes more or less contribution to both two CRFs-based classifiers, although the contribution to the two classifiers may be inconsistent. Among the six types of features, the influence of bag-of-words is greatest, and that of part-of-the-speech is least.

5. Discussion

It is easy to understand that the two machine learning-based classifiers show much better performance than the rule-based classifier as only 5 out of 25 subcategories of PHI instances are

Table 4
Performances of the systems using one or two classifiers (strict).

Classifier	Macro-average (%)			Micro-average (%)		
	Precision	Recall	*F*-score	Precision	Recall	*F*-score
Token	85.89	82.55	84.19	85.87	81.64	83.70
Character	93.88	88.45	91.08	93.81	87.43	90.51
Rule	28.30	1.742	3.282	97.92	1.640	3.226
Token + Character	92.67	90.69	91.67	92.48	89.64	91.04
Token + Rule	86.00	82.74	84.34	86.02	81.88	83.90
Character + Rule	94.03	88.70	91.29	93.96	87.70	90.72
Merged	92.82	90.91	91.85	92.64	89.88	91.24

recognized in the rule-based classifiers. The proportion of these five subcategories of PHI instances is only 5.58%. For a fair comparison, we calculate the performances of the three classifiers on them, and find that the rule-based classifier shows much higher "strict" precision than the two machine learning-based classifiers (Rule: 97.92% vs Token: 95.73% vs Character: 91.96%) although it achieves much lower "strict" recall. The advantages of the rule-based classifier lies in the following two aspects: (1) it can distinguish some ambiguous instances confused by the machine learning-based classifiers such as the medical record "258-16-49-2" vs phone number "278-032-7163"; (2) it can extract some instances the machine learning-based classifiers could not recognize like the phone number "(416) 943-8331".

Compared with token-level classifiers, character-level classifiers can avoid boundary errors caused by tokenization. To evaluate the effect of tokenization on token-level classifiers, we firstly calculate their performance upper boundary in the following way: (1) assigning every token with a gold label; (2) converting the labeled tokens back into PHI instances. On the test set, the "strict" precision, recall and *F*-score upper boundaries of our token-level classifier are 91.2%, 92.61% and 91.9% respectively. The performance loss caused by the tokenization module in our token-level classifier is not small. However, the token-level CRFs does not always show worse performance than the character-level CRFs on any category. The performances of the two CRFs-based classifiers on each main category are listed in Table 6, where PRO, LOC and CON represent the PHI categories: PROFESSION, LOCATION and CONTACT respectively. An interesting finding is that the token-level CRFs is superior the character-level CRFs on PROFESSION, LOCATION, CONTACT and ID categories, but is inferior to the character-level CRFs on NAME, AGE and DATE categories. On most of main categories (all except DATE and LOCATION), our final system achieve better performance than each CRFs. The best *F*-scores of the two CRFs-based systems and the better *F*-scores after merging are shown in bold. This result indicates that the token-level CRFs and the character-level CRFs are complementary to each other. The probable reason may be that the token-level classifier captures more meaningful context.

Table 5
Evaluation of several features on test set (strict).

Features	Token-level (%)			Character-level (%)		
	P	*R*	*F*	*P*	*R*	*F*
All features	85.87	81.64	83.70	93.81	87.43	90.51
All w/o bag-of-words	75.99	72.08	73.98	90.17	83.91	86.93
All w/o part-of-speech	85.84	81.63	83.68	94.01	87.24	90.50
All w/o sentence features	86.10	81.37	83.67	93.75	87.33	90.43
All w/o section features	85.75	80.95	83.28	93.40	87.39	90.30
All w/o general NER	85.30	80.66	82.91	93.69	86.71	90.07
All w/o word representation	86.24	80.29	83.16	94.10	86.34	90.05

Table 6
Performances of the two CRFs-based classifiers on each main category (strict).

Classifier		Main categories (micro %)						
		NAME	PRO	LOC	AGE	DATE	CON	ID
Token	*P*	76.15	79.82	88.01	89.89	89.18	93.14	94.44
	R	76.21	50.84	73.25	82.59	87.89	87.16	87.04
	F	76.18	**62.12**	**79.95**	86.08	88.53	**90.05**	**90.59**
Character	*P*	91.00	91.36	84.97	96.61	98.16	90.45	91.21
	R	84.88	41.34	70.49	89.40	96.57	82.57	88.00
	F	**87.83**	56.92	77.06	**92.86**	**97.36**	86.33	89.58
Merged	*P*	88.96	83.87	83.79	95.93	97.44	91.48	92.86
	R	87.20	58.10	76.39	92.54	96.97	93.58	89.44
	F	**88.07**	**68.65**	79.92	**94.20**	97.20	**92.52**	**91.12**

Besides the current strategy used for merging the three classifiers, we also try another strategy, where only predictions made by both machine learning-based classifiers are selected. 9076 PHI instances are predicted by both two machine learning-based classifiers, and 8874 instances are correct. The "strict" micro precision, recall and *F*-score are 97.77%, 77.42% and 86.42% respectively. This strategy is much worse than the current strategy used in our system. The numbers of PHI instances in each main category predicted by both the two machine learning-based classifiers are 2113 (NAME), 61 (PROFESSION), 1224 (LOCATION), 611 (AGE), 4354 (DATE), 175 (CONTACT) and 538 (ID). Among them, the numbers of correct instances in each main category are 2031 (NAME), 60 (PROFESSION), 1151 (LOCATION), 607 (AGE), 4330 (DATE), 171 (CONTACT) and 524 (ID) respectively.

According to the performance of each main category under the "strict" criterion shown in Table 6, it is easy to find that our system performs worse on NAME, PROFESSION and LOCATION than other categories. The main reasons lie in: (1) some types of PHI instances are not formulaic, and their number is small (such as only 234 PROFESSION and 66 COUNTRY instances in the training set); (2) some types of PHI instances are easily confused with each other. For example, it is not easy to distinguish PHI instances of three subcategories of NAME if you do not completely understand a clinical note, due to that all of them are the same except their roles. A possible solution is to use semantic information. As an attempt, we add refined location dictionaries mentioned in Section 3 to our system after the challenge. This additional experiment shows that our system is improved with overall micro *F*-scores of 91.57%. For further improvement, we plan to use much more semantic resources such as Wikipedia in the future.

Another common challenge is how to distinguish PHI instances from non-PHI entities. This problem may refer to deeper understanding of context such as relation extraction, which is another case of our future work.

6. Conclusion

In this study, we develop a hybrid clinical de-identification system for track 1 of the 2014 i2b2 clinical NLP challenge. Our system uses both machine learning-based and rule-based approaches, and achieves micro-average *F*-scores of 94.64%, 91.24% and 91.63% under "token", "strict" and "relaxed" criteria respectively, which is among top-ranked systems of the 2014 i2b2 challenge. Subsequently, we add features derived from some refined localization dictionaries to our system, and further improve performance with micro *F*-scores of 94.83%, 91.57% and 91.95% under the "token", "strict" and "relaxed" criteria respectively.

Conflict of interest

None declared.

Acknowledgements

This paper is supported in part by grants: National 863 Program of China (2015AA015405), NSFCs (National Natural Science Foundation of China) (61402128, 61473101, 61173075 and 61272383), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20140508161040764, JCYJ20140417172417105 and JCYJ20140627163809422). We also thank the 2014 i2b2 NLP challenge organizers for making the annotated data set available.

References

[1] A. Act, Health insurance portability and accountability act of 1996, Public Law 104 (1996) 191.
[2] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (5) (2007) 550–563.
[3] S.M. Meystre, F.J. Friedlin, B.R. South, S. Shen, M.H. Samore, Automatic de-identification of textual documents in the electronic health record: a review of recent research, BMC Med. Res. Methodol. 10 (1) (2010) 70.
[4] O. Ferrandez, B.R. South, S. Shen, F.J. Friedlin, M.H. Samore, S.M. Meystre, Evaluating current automatic de-identification methods with veteran's health administration clinical documents, BMC Med. Res. Methodol. 12 (1) (2012) 109.
[5] L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, L. Stoutenborough, Large-scale evaluation of automated clinical note de-identification and its impact on information extraction, J. Am. Med. Inform. Assoc. 20 (1) (2013) 84–94.
[6] L. Sweeney, Replacing personally-identifying information in medical records, the scrub system, in: Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association, 1996, pp. 333–337.
[7] P. Ruch, R.H. Baud, A.-M. Rassinoux, P. Bouillon, G. Robert, Medical document anonymization with a semantic lexicon, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2000, pp. 729–733.
[8] S.M. Thomas, B. Mamlin, G. Schadow, C. McDonald, A successful technique for removing names in pathology reports using an augmented search and replace method, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2002, pp. 777–781.
[9] D. Gupta, M. Saul, J. Gilbertson, Evaluation of a deidentification (De-id) software engine to share pathology reports and clinical documents for research, Am. J. Clin. Pathol. 121 (2) (2004) 176–186.
[10] B.A. Beckwith, R. Mahaadevan, U.J. Balis, F. Kuo, Development and evaluation of an open source software tool for deidentification of pathology reports, BMC Med. Inform. Decis. Making 6 (1) (2006) 12.
[11] F.J. Friedlin, C.J. McDonald, A software tool for removing patient identifying information from clinical documents, J. Am. Med. Inform. Assoc. 15 (5) (2008) 601–610.
[12] I. Neamatullah, M.M. Douglass, H.L. Li-wei, A. Reisner, M. Villarroel, W.J. Long, P. Szolovits, G.B. Moody, R.G. Mark, G.D. Clifford, Automated de-identification of free-text medical records, BMC Med. Inform. Decis. Making 8 (1) (2008) 32.
[13] R. Guillen, Automated de-identification and categorization of medical records, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[14] E. Aramaki, T. Imai, K. Miyo, K. Ohe, Automatic deidentification by using sentence features and label consistency, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[15] Y. Guo, R. Gaizauskas, I. Roberts, G. Demetriou, M. Hepple, Identifying personal health information using support vector machines, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[16] G. Szarvas, R. Farkas, R. Busa-Fekete, State-of-the-art anonymization of medical records using an iterative machine learning framework, J. Am. Med. Inform. Assoc. 14 (5) (2007) 574–580.
[17] K. Hara, Applying a SVM based Chunker and a text classifier to the deid challenge, in: i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.
[18] B. Wellner, M. Huyck, S. Mardis, J. Aberdeen, A. Morgan, L. Peshkin, A. Yeh, J. Hitzeman, L. Hirschman, Rapidly retargetable approaches to de-identification in medical records, J. Am. Med. Inform. Assoc. 14 (5) (2007) 564–573.
[19] J.D. Lafferty, A. McCallum, F.C.N. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289.
[20] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[21] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: Computational Learning Theory, Springer, 1995, pp. 23–37.
[22] S. Amber, K. Christopher, X. Hua, U. Özlem, Practical applications for NLP in clinical research: the 2014 i2b2/UTHealth shared tasks, in: 2014 i2b2 Clinical NLP Challenge, J. Biomed. Inform. 58S (2015) S1–S5.
[23] S. Amber, U. Özlem, Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus, J. Biomed. Inform. 58S (2015) S20–S29.
[24] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, J.C. Denny, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inform. Assoc. 17 (1) (2010) 19–24.
[25] K. Toutanova, D. Klein, C.D. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, Association for Computational Linguistics, 2003, pp. 173–180.
[26] J.R. Finkel, T. Grenager, C. Manning, Incorporating non-local information into information extraction systems by gibbs sampling, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 363–370.
[27] B. Tang, H. Cao, Y. Wu, M. Jiang, H. Xu, Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features, BMC Med. Inform. Decis. Making 13 (Suppl. 1) (2013) S1.
[28] B. Tang, X. Wang, Y. Wu, M. Jiang, J. Wang, H. Xu, Recognizing chemical entities in biomedical literature using conditional random fields and structured support vector machines, in: BioCreative Challenge Evaluation Workshop, vol. 2, 2013, pp. 70–74.
[29] B. Tang, Y. Wu, M. Jiang, Y. Chen, J.C. Denny, H. Xu, A hybrid system for temporal information extraction from clinical text, J. Am. Med. Inform. Assoc. 20 (2013) 828–835.
[30] B. Tang, H. Cao, X. Wang, Q. Chen, H. Xu, Evaluating word representation features in biomedical named entity recognition tasks, BioMed Res. Int. 2014 (2014) 240403.
[31] P.F. Brown, P.V. Desouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai, Class-based n-gram models of natural language, Comput. Linguist. 18 (4) (1992) 467–479.
[32] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.