A Cascaded Approach for Chinese Clinical Text De-Identification with Less Annotation Effort

Zhe Jian1, Xusheng Guo2, Shijian Liu3, Handong Ma4; Shaodian Zhang4; Rui Zhang5; Jianbo Lei6, \*

1 Department of Medical Informatics, Harbin Medical University, Harbin, China

2 Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

3 Shanghai Children's Medical Center, Shanghai, China

4 Synyi Co. Ltd., Shanghai, China

5 Institute for Health Informatics, University of Minnesota, Minneapolis, MN, U.S.A

6 Center for Medical Informatics, Peking University, Beijing, China

∗ Corresponding Author:

Jianbo Lei, M.D., Ph. D

Associate Professor

Center for Medical Informatics Peking University No.38 Xueyuan Rd, Haidian District, Beijing 100191, China Email: jblei@hsc.pku.edu.cnTelephone: +86 (10) 8280-5901Fax: +86 (10) 8280-5900

# Highlights

* Pattern matching can accurately locate sentences including PHI.
* Construction of dense PHI corpus reduces manual annotation cost.
* A cascaded method can enhance performance of de-identification, especially sensitivity.

# Abstract:

With rapid adoption of Electronic Health Records (EHR) in China, an increasing amount of clinical data has been available to support clinical research. Clinical data secondary use usually requires de-identification of personal information to protect patient privacy. Since manually de-identification of free clinical text requires significant amount of human work, developing an automated de-identification system is necessary. While there are many de-identification systems available for English clinical text, designing a de-identification system for Chinese clinical text faces many challenges such as unavailability of necessary lexical resources and sparsity of patient health information (PHI) in Chinese clinical text. In this paper, we designed a de-identification pipeline taking advantage of both rule-based and machine learning techniques. Our method, in particular, can effectively construct a data set with dense PHI information, which saves annotation time significantly for subsequent supervised learning. We experiment on a dataset of 3,000 heterogeneous clinical documents to evaluate the annotation cost and the de-identification performance. Our approach can increase the efficiency of the annotation effort by over 60% while reaching performance as high as over 90% measured by F score. We demonstrate that combing rule-based and machine learning is an effective way to reduce the annotation cost and achieve high performance in Chinese clinical text de-identification task.

# Keywords

De-identification; Clinical natural language processing; Chinese NLP; Annotation cost

# Introduction

With an increasing amount of clinical data being documented by Electronic Health Records (EHRs), the risk of patients’ privacy disclosure is always around the corner. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [1] requires patient consent when clinical information is used for research purposes. However, this requirement may be waived if the data is de-identified by removing 18 data elements representing protected health information (PHI) [2]. As the first step to making clinical data conveniently accessible to researchers, de-identification of such data has attracted many research attentions. It is particularly challenging to remove personal information such as names and addresses from unstructured clinical notes, which usually relies on natural language processing (NLP) based on either rule-based [3-5] or machine learning techniques [6-8]. For example, the i2b2 (Center of Informatics for Integrating Biology and Bedside) Natural Language Processing Challenge held three de-identification tasks in 2006 [9] ,2014 [10] and 2016[11].

In China, hospitals have also been building up various in-hospital systems including EHRs which collects a huge amount of clinical data [12], which are valuable resources for data secondary use including clinical research. Informatics researchers in the past decade have contributed to creating NLP solutions to analyze Chinese clinical text. For instance, there have been a number of studies ranging from name entity recognition [13-17], speculation detection [18, 19] to negation detection task [20]. However, limited studies have implemented de-identification functions for Chinese clinical text. Although China is much behind western countries in legislation regarding health information protection, awareness of protecting personal information has been raising in healthcare and research communities in the past decades. As such, an automated solution that can effectively remove private information from clinical notes would become a prerequisite for more and more future studies in China. In this paper, we follow the well-established PHI principles from HIPAA, designing a de-identification system for Chinese clinical text attempting to remove the data elements that may cause the leakage of patient privacy.

Given the huge difference between Chinese and English languages [21], there are many challenges in de-identifying Chinese clinical text, one of which is the fact that Chinese language processing relies on word segmentation. Certain morphological features in Chinese also makes the task challenging. The most striking example would be the inexistence of capitalization in Chinese, which makes it impossible to identify person names in Chinese the same way as in English by recognizing capitalized words. Constructing comprehensive English dictionaries including person names, cities and countries, common words and professional healthcare clinic names is a common approach and has been adopted in the de-identification system like concept-match scrubber [3]. Also, there are many PHI related English dictionaries available online, such as dictionaries of 1990 U.S. census person names and U.S. states [22]. However, the construction of such a comprehensive dictionary in Chinese is extremely time-consuming and difficult. For names alone, there are currently over 4,000 Chinese surnames in use and 106,000 individual Chinese characters for given names. It is also difficult to identify names because of the high degree of fuzziness required in matching the first name and the second name [23]. Joint with the challenge of word segmentation, a Chinese name can be very challenging to be distinguished from a sequence of Chinese characters. Given above language-dependent features, it is challenging to build a standalone dictionary or pattern-based method to identify personal information from Chinese clinical notes directly. However, as we will suggest in this paper, it may be possible to enumerate keywords, phrases, and context information of PHI in diverse clinical documents with a set of regression expressions to locate *sentences* with privacy information with high sensitivity instead of de-identifying the raw text directly, under the written principles given by Basic Norms of Medical Record Writing of China [25].

Another challenge is the sparsity of PHI in Chinese clinical text. Similar with English clinical notes, density of PHI in Chinese clinical notes are relatively low because of physician’s writing conventions. For example, pronouns are predominantly used to refer to the patient in the description of patients’ clinical situation. Doctors prefer to write sentences like “病人因腰背疼痛住院（the patient was admitted to the hospital because of waist and back pain.）” rather than disclosing patient’s and hospital’s name, even if it is the beginning sentence in a note. Because of such sparsity, creating annotated data set to train supervised learning de-identification system can be very costly, since human annotators will need to read a rather long piece of text to capture a single occurrence of privacy.

In this paper, our study objectives are (i) to compile a relatively general pattern-matching

method to filter sentences with high density of PHI based on the characteristics of Chinese clinical text; (ii) to construct a dense PHI corpus that can save annotation cost; (iii) to train a machine learning model on the dense PHI corpus to identify PHI from raw clinical notes; and (iv) to evaluate the effectiveness of above components.

## Related work

Methods for de-identification from free text generally fall into two categories: rule-based methods and machine-learning methods. Gupta and colleagues [26] proposed De-ID to identify by using a set of rules, pattern matching algorithms and dictionaries. The De-ID extracted patient information from headers to search the body of the text using string matching search. Also, it used UMLS to identify whether a candidate information is a valid medical term. De-ID can replace PHI with specific tags that indicate the category of the PHI removed and the same PHI would be replaced consistently with the same tags. Neamatullah and colleagues [27] proposed a pattern-matching de-identification system for nursing progress notes, using lexical look-up tables, regular expressions, and simple rules. The evaluation showed that their method experienced a high recall rate but low precision, partly because of non-standard abbreviations, ungrammatical statement, and misspellings. Compared with rule-based methods, machine-learning methods such as conditional random fields, support vector machines (SVM) were chosen by more researchers for de-identification. Aramaki and colleagues [8] used conditional random fields (CRF) to learn the relation between features and labels in the tagged training set based on two learning processes. In the first learning process, they used local features, non-local features, and extra-source features. The second learning process used all of the features from the first learning process, plus four additional features dealing mostly with the prevalence of a label at the corpus level to achieve labeling consistency. Uzuner and colleagues [28] used SVM and local context to de-identify medical discharge summaries. They created an effective de-identifier and found that local context could improve the effect of de-identification. Gardner and colleagues [29] described the Health Information DE-identification (HIDE) system. They used CRFs to extract identifying and sensitive attributes in the HIDE system including the previous word, next word, capitalizations, the presence of special characters, or if the token was a number. One unique aspect of this work was the iterative classifying and retagging of the training corpus during the development of the system. They created a labeled training set in less time and effort. Hara and colleagues[7] developed a cascaded de-identification system. They used the rule-based method to recognize phone numbers, dates, and IDs. For hospitals, locations, patients, doctors, and ages, they trained the support vector machine using the global and local features.

As for de-identification for Chinese clinical text, Chen and colleagues [30] proposed a method using dictionary and pattern matching method to extract PHI without relying exclusively on manual annotations. They created a specific keywords dictionaries by using the word segmentation. For PHI that cannot be identified by keywords, pattern-matching method was used to identify PHI. In our work, we take advantage of pattern-matching method that can locate sentences with PHI but do not rely on the rules to accurately identify words or phrases that represent private information.

# Materials and Methods

## Overview

We rely on a cascaded solution combining rule-based and machine learning methods to identify PHI from Chinese clinical notes. The pattern-matching module developed in this paper locate sentences with PHI instead of identifying and replacing the PHI directly in the raw text, while machine learning can identify the PHI but requires more annotation effort. Our workflow includes two pipelines: dense PHI corpus construction and de-identification system for a new note. A snapshot of the workflow is given in Figure1. In pipeline 1, using common keywords and context patterns of PHI in Chinese clinical text, rules are compiled and sentences with PHI in clinical text are filtered out, followed by candidate sentences including PHI collected as a dense PHI corpus. After manual annotation, the dense PHI corpus will be used as the training set for the machine-learning classifier for de-identification. In pipeline2, a new clinical note is firstly processed by the pattern matching module to find candidate sentences. Then candidate PHI sentences in the raw text is fed into a CRFs classifier, in which PHI is tagged automatically and replaced with corresponding surrogates.

**Figure 1** The framework of the de-identification method created in this paper

Clinical text corpus

Pattern matching

Pipeline1: Construction of Dense PHI Corpus

Candidate sentences with PHI

Manual Annotation of candidate sentences (the dense set)

Pipeline2: De-identification System for Raw Text

De-identified text

Machine Learning

Candidate sentences with PHI

Pattern matching

New note

## Material

For this task, we randomly selected 3,000 clinical documents including 1,500 admission notes, 1,000 discharge summaries, 250 consultation records and 250 death records from three general hospitals in Shanghai, China. Since we were not able to obtain consent from every patient involved to conduct the study, we tried the following to protect patient privacy as much as possible, while preserving as much original private information in text as possible. Before accessing the data, we asked physicians to manually identify and replace authentic patient names with random surrogates, a principle following the guideline of 2006 i2b2 de-identification challenge [9]. We kept other types of original information such as doctor names and hospital names. The study was also supervised tightly by the ethics committees and medical service departments of the hospitals.

Each clinical text contains semi-structured (e.g., patient information section) as well as unstructured text (e.g., Chief Complaint and Past Medical History). The Figure 2 shows a typical electronic medical records template in Chinese hospitals. Our main task in this paper is to remove PHI from the unstructured part, since de-identification from the semi-structured content is rather straightforward.

**Figure 2** A Typical Chinese Electronic Medical Records Template

医院名称 Hospital Name 科室 Clinical Department

记录类型Record Type

Patient Name Medical Record number

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名：Patient name |  | 性别：Sex |  |
| 年龄：Age |  | 婚姻：Marital Status |  |
| 家庭住址：Home Address |  | 工作单位：Work unit |  |
| 亲友姓名：Relative name |  | 联系方式：Contact information |  |

主诉： Chief Complaint

现病史： History of Present Illness

既往史： Past Medical History

系统回顾：System Review

个人史： Personal History

婚育史： Marital and Reproductive history

家族史： Family History

体格检查：Physical Examination

辅助检查： Auxiliary Examiniation

初步诊断：Primary Diagnosis

## Pipeline 1: Construction of dense PHI corpus

### Pattern-matching module

The first step of our pipeline is to create a set of rules and patterns that can identify *sentences* with PHIs with high sensitivity, so that a denser data set could be constructed. As our pattern-matching module is character-based, it is unnecessary to segment words at first. Although inexistence of capitalization is a challenge, it is still possible to use other features, mostly lexical ones, to create rules and identify sentences with PHI in Chinese clinical text. In this study, we defined PHI categories as follows: names (including patient’s name and doctor’s name), hospitals, and addresses. As for other types of information, they can be easily captured by direct rules, need no subsequent supervised learning, hence be excluded from this study.

To achieve high sensitivity, we sampled 50 clinical notes to compile a series of regular expressions for each PHI category. The regular expressions were derived as follows. Two physicians read clinical notes with PHI, and listed keywords and context cues they believed should be considered. Then one programmer compiled such keywords and rules to regular expressions, applied them to the original text, and presented the results back to the physicians. Then the physicians read the results, evaluated the results quantitatively, and made suggestions for improvements. The process went iteratively for 3 rounds. Table 1 shows the regular expressions to identify patient’s name, doctor’s name, address and hospital.

**Table 1** Regression expressions used in the pattern-matching module. [\u4e00-\u9fa5] refers to the range of Unicode for Chinese characters

|  |  |
| --- | --- |
|  | Regression expressions |
| Patient Name | (?:姓名|家庭住址|婚姻|出生地|工作单位|职业|民族)[^\u4e00-\u9fa5]{0,8}[\u4e00-\u9fa5]  (?:name |family address| marital status | birthplace| work unit | location| nationality)[^\u4e00-\u9fa5]{0,8}[\u4e00-\u9fa5]  (?:患者)[^\u4e00-\u9fa5]{0,3}[\u4e00-\u9fa5]{2,5}[^\u4e00-\u9fa5]{1,3}(?:男|女|性别)  (?:patient)[^\u4e00\u9fa5]{0,3}[\u4e00-\u9fa5]{2,5}[^\u4e00\u9fa5]{1,3}(?:male| female| sex) |
| Doctor Name | [\u4e00-\u9fa5]{2,8}[^\u4e00-\u9fa5]{0,3}[^上级|当地](?:医生|教授|医师|主任|护士)  [\u4e00-\u9fa5]{2,8}[^\u4e00-\u9fa5]{0,3}[^superior|local](?:doctor| professor | physician | director | nurse ) |
| Address | (?:地址|家庭住址|位于) [^\u4e00-\u9fa5]{0,3}  (?:address|family|address|locate) [^\u4e00-\u9fa5]{0,3} |
| Hospital | [^(上级|当地)](?:医院|中心))  [^(superior|local)](?:hospital|medical center )) |

### Annotation of the PHI Corpus

In order to train and to evaluate our methods, gold-standard annotation of PHI for a subset of notes were provided. We randomly sampled 300 notes, consisting of 8,256 sentences, from the entire pool. Two annotators annotated a common subset (sentences from 100 notes) of the dense PHI corpus to estimate the inter-rater agreement. An agreement of 0.76 measured by Kappa statistic was reached. The two annotators then resolved all disagreements and refined the guideline. The remaining 200 notes was evenly split and then each single annotator manually reviewed additional sentences from 100 notes. As such, we obtained a data set of 300 raw notes with PHI annotations, and we name it *raw set* in the following sections. We also applied our pattern matching module to this raw set, yielding a *dense set* of 201 sentences which contain PHI. In total, 141 names, 51 addresses, and 22 hospital mentions were annotated.

## Pipeline 2: De-Identification system for raw text

### Pattern-matching module

Similar to the pattern-matching module in the Pipeline1, we used the pattern-matching module in the Pipeline2 to filter out sentences with PHI. We then used the machine learning method to identify and replace the PHI.

### Machine learning-based classifiers

In our system, a typical representation schema for sequence labeling was used to represent PHI instances: “BIOES”, where ‘B’,’I’,’O’ and ‘E’ denote that a token/character is at the beginning, middle, outside and end of a PHI, and ‘S’ denotes that a character itself is a PHI. Like name entity recognition tasks, the de-identification task is regarded as a sequence labeling problem and thus CRFs [31]was used to tag the PHI automatically. We relied on the open source CRF++ tool [32] for model training and prediction. Features we used to train CRFs was as follows.

* Context features: characters prior and after the target character. The window size is set to 5.
* Section information: Section headers like chief complaint, history of present Illness and past medical history were used.

The following example shows how a sentence with PHI is labeled, in which B-dr\_name represents the beginning character of a PHI that belongs to the category of doctor name, I- dr\_name represents being inside a PHI, E represents being end of a PHI, and O represents being outside of a PHI:

主/O 治/O 医/O 师/O 刘/B- dr\_name 光/I- dr\_name 维/E- dr\_name 查/O 体/O 如/O 下/O :/O

*(The result of physical examination of the attending doctor Guang wei Liu is as follows:)*

### Post-processing

Like De-ID software, the PHI was replaced with the corresponding safe surrogate in the sentences that have already been tagged. For example, an address was replaced with “\*\*地址(the address)<>”. Our approach guarantees PHI can be replaced by the assigned surrogate and the sentences will be put back in the clinical text, which increases the readability of the clinical text.

# Result

## Performance of Dense PHI Corpus Construction

### Performance of pattern-matching module in identifying sentences with PHI for raw clinical text

The *raw set* with gold-standard annotations totally includes 192 sentences with PHI. To evaluate the performance (in precision and recall) of pattern-matching in identifying *sentences* with PHI, we report how accurate it identifies sentences with PHI in Table 2. We also include the performance of manual annotation of single annotator (one of the 2 annotators we relied on for gold-standard annotation) for comparison.

**Table 2** Comparison of total number of sentences identified through manual selection (single annotator) and pattern matching, compared to the numbers of sentences including PHI in gold standard.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Total Number of sentences identified | Correct identifications | Total number of PHI sentences in gold standard | Precision(%) | Recall(%) | F-score |
| Manual selection (singles annotator) | 187 | 187 | 192 | 97.4 | 98.6 | 98.0 |
| Pattern-matching | 201 | 189 | 192 | 94.0 | 98.4 | 96.1 |

The result shows that pattern-matching module performs fairly well. Although some incorrect sentences were extracted from the raw text, the pattern matching achieved a very high recall (sensitivity). Manual annotation from single annotator was not superior to the pattern-matching module, especially in recall. The annotator missed five sentences containing PHI but pattern-matching module missed only three sentences. Even if the human result’s precision was higher, the pattern-matching module performed better considering the purpose was to miss as few sentences as possible.

### Evaluation of cost of manual annotation

We compared the manual annotation cost for *raw set* and *dense set*, to show how much annotation time the pattern matching module can help reduce. Two expert annotators were required to annotate PHI from 50,000 words from *raw set* and from *dense set*. Each of them annotated around 25,000 words from *raw set* and from *dense set,* respectively*.* Table 3 presents performance (the number of PHI identifiers, annotation error and the number of PHI identifiers annotated per min) in the experiments.

**Table 3** Evaluation of annotation cost and the number of annotated PHI per minute for *raw set* and *dense set*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total annotation time(min) | Number of PHI identifiers | Number of annotation errors | Number of PHI identifiers annotated per min |
| *raw set* | 78 | 200 | 10 | 0.45 |
| *dense set* | 118 | 1,259 | 5 | 3.20 |

The result shows the efficiency of manual annotation on dense PHI corpus enhanced significantly, compared with annotating on the original data set. Annotators improved the number of PHI identifiers annotated per minute by over 600%, not to mention that annotating raw clinical text yielded even more mistakes.

## Performance of De-Identification System

### Evaluation of De-Identification System

We conducted a set of cross-experiments to test the performance of machine learning model trained with different data sets, to examine how pattern matching affect performance of machine learning and hence performance of the entire process of de-identification. For the CRF model trained with the *raw set*, we called it as the Model 1 (i.e. switch off pattern-matching in pipeline 1); for the CRF model trained with *dense set*, we called it as the Model 2 (i.e. switch on pattern-matching in pipeline 1). We applied the two models to *raw set* and *dense set* respectively for evaluation (i.e. switch off and switch on pattern-matching in pipeline 2). For each evaluation, we carried out 10-fold cross validation and report average performance in precision, recall, and F-measure. Table 4 displays results of this set of experiments.

**Table 4** De-identification system when training and testing on the *raw set* and *dense set*, respectively. P refers to precision, R refers to recall, F refers to F1-measure

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Model 1 testing on the *raw set* | | | Model 2 testing on the *raw set* | | | Model 1 testing on the *dense set* | | | Model 2 testing on the *dense set* | | |
| P% | R% | F | P% | R% | F | P% | R% | F | P% | R% | F |
| Name | 82.3 | 67.1 | 73.9 | 76.3 | 92.8 | 83.7 | 94.4 | 58.7 | 72.4 | 93.4 | 97.7 | 95.5 |
| Address | 92.1 | 56.3 | 69.9 | 84.8 | 90.5 | 87.5 | 83.8 | 44.2 | 57.9 | 97.1 | 91.5 | 94.2 |
| Hospital | 86.7 | 25.1 | 38.9 | 75.8 | 83.3 | 79.4 | 69.5 | 22.5 | 34.0 | 93.5 | 91.0 | 92.2 |
| Micro | 86.0 | 57.3 | 67.3 | 78.8 | 90.6 | 84.2 | 87.3 | 48.7 | 62.0 | 94.6 | 94.8 | 94.6 |

The result shows that using *dense set* as the training set yielded always higher performance, even on the *raw set*. Although using the raw clinical text as training set can achieve comparatively high precision, the recall was much lower than its counterpart. In contrast, using the *dense set* as the training set yielded a relatively stable result that the average recall achieved over 90% and the precision is also surprisingly high, especially when testing on the *dense set*. In general, the experimental results suggest that using *dense set* rather than *raw set* is a better option in choosing training data.

Our experimental result also show that switching on pattern-matching in pipeline 2 (i.e. use pattern matching for a new note before feeding into machine learning) help with de-identification. Based on the two sets of results based on Model 2, it can be observed that precision increased significantly after using the pattern-matching module, rising from less than 80% to over 90%. After filtering out the sentences with PHI, machine learning model tend not to identify irrelevant information as PHI. The recall is also affected positively after adding the pattern-matching module.

To evaluate how the performance of the machine learning model changed with the annotation effort, Figure 3 visualizes the trends how precision and recall of the model training on *dense set* and *raw set* changes with the annotation cost measured by time of manual annotation. With the same annotation time, using *dense set* as training set always yields higher recall and precision. With the increase of the annotation time, the performance of the model trained on raw texts changed significantly, leaping from 50% to over 80% in recall. The model trained with dense corpus performed fairly steady, as shown in the timeline chart, which can always achieve over 70% in precision and recall. The results suggest that given same amount of manual annotation input, the pattern matching can help build training data which enables the machine learning model to achieve better performance.

**Figure 3** Comparison of precision and recall by the time used for manual annotation, stratified by training on *raw set* and *dense set*.

* + 1. Applying the methods on radiology reports to evaluate generalizability

To evaluate the generalizability of our de-identification method, we then used 1000 radiology reports from the same hospital as the test set while we keep the original *dense set* as training data. The detailed performance is given in Table 5, on sentences filtered with the same set of regular expressions we introduced earlier. The results show that the performance measured by precision, recall, and F, of our method indeed drops, given the fact that the radiology reports have different writing patterns and content from admission and discharge notes. However, it is surprising that the change is not so significant in comparison with changing the training set to *raw set*. If the model is applied directly to the radiology reports without filtering with the regular expressions, the overall performance drops to 55.4 measured by Micro-F.

**Table 5** The performance of the de-identification system (Model 2) tested on radiology reports from the same hospital after filtering potential sentences with PHI using the regular expressions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | P% | R% | F% |
| Name | 86.7 | 85.6 | 86.1 |
| Address | 85.1 | 84.8 | 84.9 |
| Hospital | 82.3 | 84.3 | 83.3 |
| Micro | 84.3 | 85.1 | 84.8 |

# Discussion

## Principle findings and error analysis

Machine learning is usually successful in sequence-labeling tasks like de-identification system introduced in this paper. As we have discussed in the previous section, the sparsity of PHI in Chinese clinical text results makes it costly to create gold-standard corpus. Our study suggests that taking advantage of specific linguistic patterns can quickly and accurately locate *sentences* with PHI, which contributes significantly to the reduction of manual annotation cost and high sensibility. The main contribution of our work is that: 1) we increase the efficiency of annotation by over 90% and reduce the annotation error by creating a corpus with dense PHI; 2) improve the performance by using the pattern-matching module and achieve satisfied overall performance for de-identification within less annotation time.

Our first set of experiments has shown that using pattern-matching to preprocess the raw clinical text can increase both efficiency and quality of annotation for PHI. Our approach identified sentences with high likelihood of PHIs instead of PHI identifiers themselves, which ensured the context of PHI and its section information can be fully preserved. Different from other name entities in clinical text, PHI occurs in relatively regular contexts with certain keywords. By interviewing our annotators, we found the consumption of time and effort in annotation originates from two sources: locating the PHI and categorizing them correctly. The sparsity of PHI made the annotation work ineffective because annotators had to look through 90% unnecessary sentences before locating the PHI. One key point is that categorizing PHI does not rely so much on professional medical knowledge so it can be relatively easy to recognize compared with other non-PHI information in clinical text. We also found that even annotators relied on the linguistic or contextual features of PHI to locate them, which also justifies the necessity our pattern-matching filtering. After comparing the results of manual annotation and pattern-matching module, we were surprised to see that the pattern-matching module even performed better. This probably due to the difficulty for human in locating PHI in such sparse clinical text or the descent of annotation quality over a long period of time. This phenomenon confirmed that our compiled regular expression could identify majority of the sentences including PHI.

We mainly analyzed the false negatives of pattern-matching module considering the major goal is to pursue a high recall. For the 3 missing sentences, 2 were caused by the lack of keywords and context cues in the sentence, such as 刘清今日查房（Qing Liu made the rounds of the wards today）. The sentence is overly simplified when being written which caused the error. There is also one error which were caused by a typo entered by doctor, which made it very difficult to be captured by our pattern-matching module.

Our study also suggests that the efficiency of annotation increased by over 90% and the frequency of errors has reduced when annotating on the dense PHI corpus. According to our observation, it is interesting to find that annotators allocate the time differently for two corpora. The annotator who dealt with raw clinical text spent most of the time going through the text to locate PHI while another annotator took a majority of time to categorize the PHI correctly. As we have discussed before, locating PHI required longer time and more efforts than categorizing in such a sparse clinical text, which can explain the phenomenon that errors in raw text annotation occurred more frequently. Error types were also different: errors in dense PHI corpus annotation were almost wrongly cataloged but annotators who dealt with raw test tend to omit some PHI hidden in the disease description or auxiliary examination section. When comparing with the two types of errors, mistakenly cataloging was more acceptable because it only affected the readability of clinical text; whereas, omitting PHI resulted in the risk of privacy disclosure.

The pattern matching filtering is effective not only in creating training corpus more efficiently, but also in de-identification for new text. The pattern-matching module can coarsely filter out candidate sentences including PHI, which requires less runtime and help CRF achieve higher precision. Without the pattern-matching module, the result showed that precision of CRFs dropped from over 90% to less than 80% when de-identifying name and hospital.

Also, the experimental results demonstrate the generalizability of our methods to some extent. When using another type of clinical text, radiology reports, as the test set, the model still achieves satisfying performance. Most of the errors in this experiment were caused by incorrect boundaries. For example, our annotators categorized “浦东新区齐河路102号 (Number 102 Road Qihe Pudong New Area)” to the address; however, the CRFs model covered only a portion of the PHI identifiers, the result was ” 齐河路102号(Number 102 Road Qihe)”. At the same time, categorizing the PHI mistakenly occurred sometimes, for example, “古槐街道社区卫生服务中心 (Guhuai Street Community Health Service Center)” was tagged by the system as the address instead of the hospital. This is caused by the fact that such local health center’s names show up in radiology reports but not in admission and discharge notes, because such local center provides radiology test service but not inpatient services.

**Limitations**

The first and the major limitation of this work is that the pattern-matching module were hard to be applicable for other types of clinical texts. Although we tried our best to diverse clinical documents we could collect, it was unlikely that the pattern matching method was effective to identify sentences including PHI for all kinds of clinical documents. Huge differences including terminologies and templates used in EHR exist across different hospitals, clinics or even healthcare providers. It may be necessary to design a module that can adjust the keywords with the definite pattern of clinical text automatically.

Moreover, our CRFs model only used limited features and we did not push the limit by tuning parameters or carrying out sophisticated feature engineering. In contrast, our system only used most common features partly because the context of PHI was relatively simple. It might be helpful to add common features and combine different features. Optimizing the CRFs model may achieve a more promising result. Also, we did not add more training data to see how much performance difference will the systems show once abundant training dataset is given.

Finally, our approach still relied on the humans to annotate, even though the efficiency and quality of annotation have increased. Many methods have proposed to reduce the annotation effort for supervised machine learning [33-35]. For example, Yoshimasa and colleagues [36] proposed an active learning-like framework to perform as an iterative and interactive process between the human annotator and a probabilistic named entity tagger. As the pattern of PHI is regular, applying the active learning method may be an effective way to reduce annotation effort. Therefore, it is necessary to improve the efficiency of annotation with the help of machine learning classifier in the annotation.

# Conclusion

Our study focused on designing a high-performance de-identification system for Chinese clinical text by combing rule-based and machine learning methods. We took advantage of specific features of PHI hidden in clinical text using rules and construct a dense PHI corpus as the training set for machine learning. We experimented with the CRFs model trained on dense PHI corpus and raw text, and demonstrated the effectiveness of dense PHI construction. Overall, our evaluation demonstrated our method could save annotation cost significantly and could achieve state-of-the-art de-identification performance.

# Acknowledgement

This study was partly supported by the National Natural Science Foundation of China (NSFC) Grant #81471756.

# References

1. Health UDo, Services H: **Protection of human subjects**. *Code of Federal regulations 45: Public welfare* 1995(Sections 46-101 to 46-409):Unknown.

2. Health UDo, Services H: **HIPAA—General Information**. *Centers for Medicare & Medicaid Services www hhs gov* 2011, **14**.

3. Berman JJ: **Concept-match medical data scrubbing: how pathology text can be used in research**. *Archives of pathology & laboratory medicine* 2003, **127**(6):680-686.

4. Friedman C: **A broad-coverage natural language processing system**. *Proceedings / AMIA Annual Symposium AMIA Symposium* 1999, **7**(1):270-274.

5. Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD: **Automated de-identification of free-text medical records**. *BMC Medical Informatics and Decision Making* 2008, **8**(1):32.

6. Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple M: **Identifying Personal Health Information Using Support Vector Machines**. *Chinese Journal of Social Medicine* 2011.

7. Hara K: **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*: Am Med Inform Assoc; 2006: 10-11.

8. Aramaki E, Imai T, Miyo K, Ohe K: **Automatic deidentification by using sentence features and label consistency**. In: *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data: 2006*; 2006: 10-11.

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

10. Stubbs A, Uzuner Ö: **Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus**. *Journal of biomedical informatics* 2015, **58**:S20-S29.

11. **Announcement of Data Release and Call for Participation 2016 CEGS N-GRID Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data**

12. Lei J, Meng Q, Li Y, Liang M, Zheng K: **The evolution of medical informatics in China: A retrospective study and lessons learned**. *International Journal of Medical Informatics* 2016, **92**:8-14.

13. Xu Y, Wang Y, Liu T, Liu J, Fan Y, Qian Y, Tsujii J, Chang EI: **Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries**. *Journal of the American Medical Informatics Association* 2014, **21**(e1):e84.

14. Wu Y, Jiang M, Lei J, Xu H: **Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network**. *Studies in Health Technology & Informatics* 2015, **216**:624-628.

15. Lei J, Tang B, Lu X, Gao K, Jiang M, Xu H: **A comprehensive study of named entity recognition in Chinese clinical text**. *Journal of the American Medical Informatics Association Jamia* 2014, **21**(5):808.

16. Wang SK, Shao-Zi LI, Chen TS: **Recognition of Chinese Medicine Named Entity Based on Condition Random Field**. *Journal of Xiamen University* 2009, **48**(3):359-364.

17. Wang Y, Liu Y, Yu Z, Chen L, Jiang Y: **A preliminary work on symptom name recognition from free-text clinical records of traditional chinese medicine using conditional random fields and reasonable features**. In: *The Workshop on Biomedical Natural Language Processing: 2012*; 2012: 223-230.

18. Zhang S, Kang T, Zhang X, Wen D, Elhadad N, Lei J: **Speculation Detection for Chinese Clinical Notes: Impacts of Word Segmentation and Embedding Models**. *Journal of Biomedical Informatics* 2016, **60**(C):334-341.

19. Chen Z, Zou B, Zhu Q, Li P: **Chinese Negation and Speculation Detection with Conditional Random Fields**: Springer Berlin Heidelberg; 2013.

20. Kang T, Zhang S, Xu N, Wen D, Zhang X, Lei J: **Detecting negation and scope in Chinese clinical notes using character and word embedding**. *Computer Methods & Programs in Biomedicine* 2016, **140**:53-59.

21. Wu Y, Lei J, Wei WQ, Tang B, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Zheng K, Xu H: **Analyzing Differences between Chinese and English Clinical Text: A Cross-Institution Comparison of Discharge Summaries in Two Languages**. *Studies in Health Technology & Informatics* 2013, **192**(1):662-666.

22. Kolatch AJ: **The name dictionary: modern English and Hebrew names**: J. David; 1967.

23. Horng-Jyh PW, Jin-Cheon N, Soo-Guan CK: **A hybrid approach to fuzzy name search incorporating language-based and text-based principles**. *Journal of Information Science* 2007, **33**(1):3-19.

24. Bodenreider O: **The unified medical language system (UMLS): integrating biomedical terminology**. *Nucleic acids research* 2004, **32**(suppl 1):D267-D270.

25. Ministry of Health PRC: **Basic Norms of Medical Record** *Health Office Medical Care Administration File[2002]* 2010(190).

26. Gupta D, Saul M, Gilbertson J: **Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research**. *American journal of clinical pathology* 2004, **121**(2):176-186.

27. Neamatullah I, Douglass MM, Li-wei HL, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD: **Automated de-identification of free-text medical records**. *BMC medical informatics and decision making* 2008, **8**(1):32.

28. Uzuner Ö, Sibanda TC, Luo Y, Szolovits P: **A de-identifier for medical discharge summaries**. *Artificial intelligence in medicine* 2008, **42**(1):13-35.

29. Gardner J, Xiong L: **HIDE: an integrated system for health information DE-identification**. In: *Computer-Based Medical Systems, 2008 CBMS'08 21st IEEE International Symposium on: 2008*: IEEE; 2008: 254-259.

30. Chen L, Yang J-J, Wang Q: **Privacy-preserving data publishing for free text Chinese electronic medical records**. In: *Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual: 2012*: IEEE; 2012: 567-572.

31. Lafferty JD, Mccallum A, Pereira FCN: **Conditional Random Fields: Probabilistic Models For Segmenting And Labeling Sequence Data**. In*: 2001*; 2001: 282--289.

32. Kudo T: **CRF++: Yet another CRF toolkit**. *Software available at* [http://crfpp/](http://crfpp) *sourceforge net* 2005.

33. Engelson SP, Dagan I: **Minimizing manual annotation cost in supervised training from corpora**. In: *Proceedings of the 34th annual meeting on Association for Computational Linguistics: 1996*: Association for Computational Linguistics; 1996: 319-326.

34. Culotta A, McCallum A: **Reducing labeling effort for structured prediction tasks**. In: *AAAI: 2005*; 2005: 746-751.

35. Lughofer E: **Hybrid active learning for reducing the annotation effort of operators in classification systems**. *Pattern Recognition* 2012, **45**(2):884-896.

36. Tsuruoka Y, Tsujii Ji, Ananiadou S: **Accelerating the annotation of sparse named entities by dynamic sentence selection**. *BMC bioinformatics* 2008, **9**(11):S8.