Automatic Deidentification by using Sentence Features and Label Consistency

Eiji ARAMAKI, Ph.D.[1], Takeshi IMAI, Ph.D.[1]
Kengo MIYO, R.N., P.H.N., Ph.D.[1], Kazuhiko OHE, M.D., Ph.D.[1]
[1] The University of Tokyo Hospital, Japan
aramaki@hcc.h.u-tokyo.ac.jp

Abstract *Deidentification of clinical records has drawn a great deal of attention in the medical field. Since texts in clinical records are mostly ungrammatical and fragmented, previous approaches have relied only on local information, namely contextual words surrounding a current target word. The present paper proposes a new approach employing three types of non-local features, which does not come from surrounding words: (1) sentence features, corresponding to the previous/next sentence information and (2) label consistency, preferring the same label for the same word sequence. The experimental results showed high performance (precision 98.29%; recall 96.66%; f-measure 97.47), demonstrating the feasibility of the proposed approach.*

Introduction

Clinical records contain a great deal of helpful information for statistical medical studies. However, they also contain Personal Health Information (**PHI**). Therefore, **deidentification**, which is the task of removing PHI, has drawn a great deal of attention in the medical field. The deidentification task is similar to Named Entity Recognition (**NER**). However, the deidentification targets, namely clinical records, are different from the targets of NER. NER targets are usually grammatical (such as newspapers or the Hansard), whereas clinical records contain primarily ungrammatical and fragmented sentences (as shown in Figure 1).

In such an ungrammatical text, global context might do more harm than good. Therefore, a state-of-the-art system depends mainly on small context, which comes from a target word and its neighboring words (the previous two words and the following two words). In this paper, we call such small scope information **local features**, and information from the other parts **non-local features**

<RECORD ID="108">
<TEXT>
<PHI TYPE="ID">294702550</PHI> <PHI TYPE="HOSPITAL">DH</PHI>
<PHI TYPE="ID">2846444</PHI>
<PHI TYPE="ID">009913</PHI>
<PHI TYPE="ID">419097</PHI>
<PHI TYPE="DATE">3/18</PHI>/2001 12:00:00 AM
ED Discharge Summary
⋮
Discharge Medications :
Celexa 20 mg. # 15 take one half pill each day for 1 week , then increase to one pill if
you are not experiencing side effects .
Follow up Service :
1. Please call <PHI TYPE="HOSPITAL">DH</PHI> out-patient psychiatry triage to schedule an
appointment , <PHI TYPE="PHONE">178-657-JGVA</PHI> .
Tell them you are interested in alternative therapies research (<PHI TYPE="DOCTOR">Vita
Linkekotemones</PHI> , MD) .
2. If you are not able to see a psychiatrist in the very near future , as we discussed ,
start Celexa and call <PHI TYPE="DOCTOR">Chol Then</PHI> , RN , CS at <PHI TYPE="PHONE">
503-200-0967</PHI> to schedule a follow up appointment in two weeks .
3. If your symptoms worsen , return to the APS or to your nearest emergency room .
PCP Name :
<PHI TYPE="DOCTOR">BLEAKMESTHAST , ANGEDA C</PHI>
Provider Number :
<PHI TYPE="ID">20416</PHI>
This report was created by <PHI TYPE="DOCTOR">THEN , NISETA</PHI> <PHI TYPE="DATE">03/18
</PHI>/2001 05:16 PM
[report_end]
</TEXT>
</RECORD>

Figure 1: An Example of a Clinical Record (XML tags indicate PHI).

tures

This paper reveales that two types of non-local features can contribute to the deidentification accuracy.

1. **Sentence Features**: As shown in Figure 1, PHI (especially IDs) is usually located in the top/bottom of records. In addition, a sentence which includes PHI tend to be short. Therefore, we incorporate some features from the target sentence, such as the position in a records and the sentence length.

2. **Label Consistency**: Another consideration is PHI label consistency. For example, there are two PHI labels in Figure 2. If there is strong evidence that first "*Dr. ROOM*" should be labeled DOCTOR, subsequent "*ROOM*" should also have the same label.

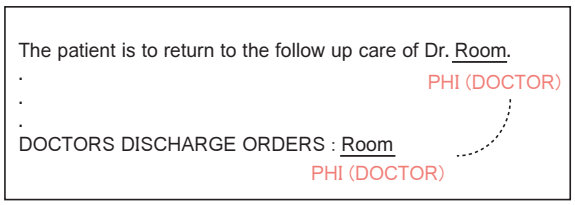

Figure 2: An Example of Label Consistency.

To capture such consistency, we employ a two-stage learning/labeling framework [1]. The first learning/labeling is a conventional method. The system then investigates the most frequent label for each token (or token sequence) from the first result. The second learning/labeling uses the most frequent labels as additional features.

In a number of experiments, the proposed system achieved higher accuracy (precision 98.29%; recall 96.66%; f-measure 97.47) than a baseline system, which depended only on local features. These results demonstrate that some non-local features can be useful even in the deidentification task.
The present paper is organized as follows. Section 2 describes related work. Section 3 describes the proposed method, and Section 4 reports experimental results. Finally, Section 5 presents our conclusions.

Related Work

In the deidentification task, early approaches depend on heuristic rules and dictionaries[2, 3], which required enormous time and cost.
Recent approaches, however, employ machine learning techniques, such as [4] and [5]. The proposed method differs from these studies in the following two aspects:

1. **Machine Learning Method**: The first study [4] is based on Hidden Markov Model (HMM) learning. Because the HMM is a generative model, it requires a large training corpus. We used a Conditional Random Field (**CRF**), which has been shown to have a high performance for many tasks, such as part-of-speech tagging[6], text chunking[7], information extraction from web documents[8], and named entity recognition [9].

2. **Non-local Features**: The second study [5] employed SVM learning, the performance of which is approximately equal to that of CRF[10]. However, as mentioned before, in [5], only local features was used. The proposed method uses not only local but also non-local features.

Table 1: IOB2 Format Example

words	IOB2-tag
:	
was	O
admitted	O
to	O
the	O
Tsta	B-HOSPITAL
Hospital	I-HOSPITAL
Obstetrical	O
service	O
:	

In addition, [11] proposed a new tagging method based on the MEDTAG framework, and [12] proposed semantic selectional restrictions. However, these approaches are domain-specific, and have little relation to natural language processing methods, which are proposed in Message Understanding Conference (MUC)[13], IREX[14] and CoNLL[10].

Learning Method

The proposed system consists of two modules; (1) the learning module and (2) the testing (labeling) module. This paper describes only (1) the learning module, because they have almost equal workflows.
The learning module consists of three steps:

1. **Pre-processing**: a format conversion.

2. **First Learning**: a learning using both local and non-local features.

3. **Second Learning**: a learning using additional features from the first learning results.

Pre-processing: IOB2 Format Conversion

First, we covert a corpus, which is in XML-format, into IOB2 format, as shown in Table 1. In IOB2, words outside of the PHI are tagged with O, while the first word in the PHI is tagged with B-k to begin class k, and further PHI words receive the I-k tag, indicating that these words are inside.

First Learning: local and sentence features

We then add the features to the corpus, and learn the relation between the features and their labels by using CRF++[1].

[1]http://www.chasen.org/ taku/software/CRF++/

In the first learning, we use three types of features:

1. **Local Features**: Information from Target Word (**TW**) and its surrounding words.

2. **Non-local Features**: Information that does not come from TW or its surrounding words.

3. **Extra-resource Features**: A kind of local features, but comes from extra resources, such as a person name dictionary or a location dictionary.

Detailed definitions are given in Table 2.

Second Learning: label consistency

The same PHI sometimes appears twice or more in one record. For example, in Figure 2, the same person, *Room*, appears twice. Such tokens tend to have the same labels. In the present paper, we refer to such a phenomenon as **Label Consistency**. The first CRF could not catch such consistency, resulting in some inconsistent labels.

In order to deal with this problem, we use the second learning technique[1]. The second learning uses not only the first learning features but also the following four additional features. These additional features come from the output of the first CRF (For details, see [1].)

1. **Token Level Majority in Record**: Refers to the majority label assigned to a token in a record. For example, there are three occurrences of the token *Room*, two of which are *PATIENT*, and one of which is *DOCTOR*. The **Token Level Majority in Record** feature would then be *PATIENT* for all three occurrences of the token.

2. **Token Level Majority in Corpus**: As with a record unit, a corpus itself also has a label consistency. To capture it, we use **Token Level Majority in Corpus**, which approximately equal to **Token Level Majority in Record**, but refers to the majority in the entire corpus.

3. **Entry Level Majority in Record**: Refers to the majority label assigned to an entry (or a token sequence) in a record.

4. **Entry Level Majority in Corpus**: Approximately equal to **Entry Level Majority in Record**, but refers to the majority in the entire corpus.

Table 3: PHI types and their Numbers

PHI type	# of PHI
AGE	13
DATE	5,189
DOCTOR	2,690
HOSPITAL	1,736
ID	3,677
LOCATION	144
PATIENT	685
PHONE	175

Table 4: 2-Way (PHI or nonPHI) Results

Method	P	R	$F_{\beta=1}$
BASELINE	98.00	95.21	96.59
BASE+SF	**98.49**	95.28	96.86
BASE+EX	98.11	96.00	97.04
BASE+LC	97.43	95.77	96.60
PROPOSED	98.29	**96.66**	**97.47**

* **P** denotes precision, **R** denotes recall, and $F_{\beta=1}$ denotes f-measure ($\beta = 1$).

Experiments

Experimental Set-up

To investigate the performance of the proposed method, we used a corpus that is provided in the i2b2-NLP shared-task. The corpus consists of 671 records, which include 14,309 PHI tags. The number of each PHI is shown in Table 3.

We compared the following methods by ten-fold cross validation:

1. **BASELINE**: Only the first learning with only local features (without non-local features and extra-resource features).

2. **BASE+SF**: BASELINE with sentence features.

3. **BASE+EX**: BASELINE with extra-resource features.

4. **BASE+LC**: BASELINE with the second CRF.

5. **PROPOSED**: Proposed method.

Results

The results are shown in Table 4 (2 way (PHI or not PHI) accuracy) and Table 5 (individual PHI types). In both cases, the proposed method has the highest $F_{\beta=1}$ score.

Sentence Features

Sentence features provide the greatest contribution, especially for ID, DATE and PATIENT, because these PHI types have formulaic occurrences at the top of records.

Table 2: Three Types of Features

(1) Local Features
Words: TW and its surrounding words. The window size is five words (-2,-1,0,1,2).
POS: Part of speech of TW and its surrounding words (-2,-1,0,1,2). Part of speech is analyzed by a POS tagger[15], which has a 97-98% accuracy for medical texts (MEDLINE abstracts).
Case: Capitalization type of TW and its surrounding words (-2,-1,0,1,2). The capitalization type consists of (1) all upper case, (2) all lower case, (3) first character is upper case and the others are lower case, and (4) Other.
Length: Length (number of characters) of TW.
Special Character: Whether TW contains symbols such as "?", "/" or "-".
Special Template: Whether TW matches a token template. We prepared some templates for DATE and PHONE, such as (XXX)XXX-XXXX and XXXX/XX/XX. In these templates, X matches the number (0-9).
(2) Sentence Features
Sentence Position: TW position in the record. We classified the sentence position into three parts: (1) in the top 10 lines, (2) in the bottom 5 lines, and (3) other.
Sentence Length: The number of words in surrounding sentences to which TW belongs. The window size is three sentences (-1,0,1).
Last Sentence: The last three words in the last sentence.
(3) Extra-resource Features
Person-name Dictionary: Whether TW appears in the name-list, consists of 46,207 entries. The name-list is extracted from a dictionary[16].
Location-name Dictionary: Whether TW appears in the location-list. Consists of 22,380 entries. The list is extracted from a dictionary[16].
Date Expression: Whether TW is a date expression. The date expression consists of the day of week (Monday, MON, etc.) and month (August, Aug, etc.).

Extra Resources

Extra resources contribute to DATE, DOCTOR, LOCATION and PATIENT. Since we used small dictionaries, one simple way to achieve higher performance is to use larger dictionaries.

Label Consistency

As shown in Table 4, label consistency showed only a slight contribution to PHI/nonPHI classification (BASELINE $F_{\beta=1}$=96.59, BASE+LC $F_{\beta=1}$=96.60 (+0.01)). However, in each PHI type classification, BASE+LC provided a greater contribution (BASELINE $F_{\beta=1}$=94.27, BASE+LC $F_{\beta=1}$=94.33 (+0.06)), indicating that the 2nd CRF changes PHI types suitably.

Conclusion

The preset paper proposed a system employing two types of non-local features: (1) sentence features, capturing information beyond a sentence, and (2) label consistency, preferring the same label for the same word sequence. The experimental results showed higher performance (precision 98.29%; recall 96.66%; f-measure 97.47), demonstrating that some non-local features can be useful even in the deidentification task.

Acknowledgments

Part of this research is supported by Grant-in-Aid for Scientific Research of Japan Society for the Promotion of Science （Project Number:16200039, F.Y.2004-2007 and 18700133, F.Y.2006-2007） and the Research Collaboration Project (#047100001247) with Japan Anatomy Laboratory Co.Ltd.

Finally, we wish to thank Dr. Yu Kun in the University of Tokyo for comments and suggestions.

References

[1] Krishnan V, Manning CD. An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics; 2006. p. 1121–1128.

[2] Douglass M, Clifford G, Reisner A, Moody GB, Mark RG. Computer-Assisted Deidentification of Free Text in the MIMIC II Database. 2005;32(1):331–334.

Table 5: Detailed Results

	P	R	$F_{\beta=1}$
BASELINE	95.64	92.95	94.27
AGE	33.33	7.69	12.50
DATE	98.32	94.57	96.40
DOCTOR	93.71	90.86	92.26
HOSPITAL	94.06	88.42	91.15
ID	96.86	98.29	97.57
LOCATION	69.15	45.14	54.62
PATIENT	84.89	83.65	84.26
PHONE	97.02	93.14	95.04
BASE+SF	**98.12**	94.94	96.50
AGE	**75.00**	**23.08**	**35.29**
DATE	**98.51**	95.74	97.11
DOCTOR	97.36	94.76	96.04
HOSPITAL	**96.53**	89.86	93.08
ID	**99.53**	98.94	**99.24**
LOCATION	**68.24**	40.28	50.66
PATIENT	**98.19**	94.89	**96.51**
PHONE	97.58	92.00	94.71
BASE+EX	95.78	93.77	94.76
AGE	33.33	7.69	12.50
DATE	98.14	95.47	96.79
DOCTOR	94.31	92.38	93.33
HOSPITAL	93.89	89.40	91.59
ID	96.99	98.31	97.65
LOCATION	70.09	52.08	59.76
PATIENT	85.80	83.80	84.79
PHONE	97.02	93.14	95.04
BASE+LC	95.13	93.54	94.33
AGE	33.33	7.69	12.50
DATE	98.02	95.18	96.58
DOCTOR	93.06	91.19	92.11
HOSPITAL	93.57	90.50	92.01
ID	96.89	98.29	97.58
LOCATION	54.62	45.14	49.43
PATIENT	84.08	84.82	84.45
PHONE	97.02	93.14	95.04
PROPOSED	97.88	**96.23**	**97.05**
AGE	60.00	**23.08**	33.33
DATE	98.27	**97.21**	**97.73**
DOCTOR	**97.51**	**96.10**	**96.80**
HOSPITAL	96.29	**92.57**	**94.39**
ID	99.35	**99.08**	99.21
LOCATION	66.96	**53.47**	**59.46**
PATIENT	97.75	**95.04**	96.37
PHONE	**98.18**	**92.57**	**95.29**

* **P** denotes precision, **R** denotes recall, and $F_{\beta=1}$ denotes f-measure ($\beta = 1$).

[3] Thomas SM, Mamlin B, Schadow G, McDonald C. A Successful Technique for Removing Names in Pathology Reports using an Augmented Search and Replace Method. In: Proceedings of AMIA Symp; 2002. p. 777–781.

[4] Roth D, Yih WT. Probabilistic Reasoning for Entity and Relation Recognition. In: Proceedings of The 19th International Conference on Computational Linguistics (COLING 2002); 2002. p. 835–841.

[5] Tawanda S, Ozlem U. Role of Local Context in Automatic Deidentification of Ungrammatical, Fragmented Text. In: Proceedings of the Human Language Technology conference and the North American chapter of the Association for Computational Linguistics (HLT-NAACL2006); 2006. p. 65–73.

[6] Lafferty J, McCallum A, , Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of ICML; 2001. p. 282–289.

[7] Sha F, Pereira F. Shallow parsing with conditional random fields. Technical Report CIS TR MS-CIS-02-35, University of Pennsylvania; 2003.

[8] Pinto D, McCallum A, Lee X, Croft W. Table Extraction using Conditional Random Fields. In: Proceedings of the 26th ACM SIGIR; 2003. p. 235–242.

[9] McCallum A, Li W. Early Results for Named Entity Recognition with Conditional Random Fields. In: Proceedings of The Seventh Conference on Natural Language Learning (CoNLL); 2003. p. 188–191.

[10] Sang TK, F E, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003; 2003. p. 142–147.

[11] Ruch P, Baud RH, Rassinoux AM, Bouillon P, Robert G. Related Articles, Links Click here to read Medical document anonymization with a semantic lexicon. In: Proceedings of AMIA Symp; 2000. p. 729–733.

[12] Taira RK, Bui AA, Kangarloo H. Identification of patient name references within medical documents using semantic selectional restrictions. In: Proceedings of AMIA Symp; 2002. p. 757–761.

[13] Grishman R, Sundheim B. Message Understanding Conference 6: A Brief History. In: Proceedings of International Conference on Computational Linguistics (COLING1996); 1996. p. 466–471.

[14] Sekine S, Isahara H. IREX: IR and IE Evaluation Project in Japanese. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC); 2000. p. 508–514.

[15] Tsuruoka Y, Tsujii J. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. In: Proceedings of HLT/EMNLP; 2005. p. 467–474.

[16] EDP. Eijiro Japanese-English Dictionary, Electronic Dictionary Project; 2005.