Role of Local Context in Automatic Deidentification of Ungrammatical, Fragmented Text

Tawanda Sibanda
CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139
tawanda@mit.edu

Ozlem Uzuner
Department of Information Studies
College of Computing and Information
University at Albany, SUNY
Albany, NY 12222
ouzuner@albany.edu

Abstract

Deidentification of clinical records is a crucial step before these records can be distributed to non-hospital researchers. Most approaches to deidentification rely heavily on dictionaries and heuristic rules; these approaches fail to remove most personal health information (PHI) that cannot be found in dictionaries. They also can fail to remove PHI that is ambiguous between PHI and non-PHI.

Named entity recognition (NER) technologies can be used for deidentification. Some of these technologies exploit both local and global context of a word to identify its entity type. When documents are grammatically written, global context can improve NER.

In this paper, we show that we can deidentify medical discharge summaries using support vector machines that rely on a statistical representation of *local* context. We compare our approach with three different systems. Comparison with a rule-based approach shows that a statistical representation of local context contributes more to deidentification than dictionaries and hand-tailored heuristics. Comparison with two well-known systems, SNoW and IdentiFinder, shows that when the language of documents is fragmented, local context contributes more to deidentification than global context.

1 Introduction

Medical discharge summaries contain information that is useful to clinical researchers who study the interactions between, for example, different medications and diseases. However, these summaries include explicit personal health information (PHI) whose release would jeopardize privacy. In the United States, the Health Information Portability and Accountability Act (HIPAA) provides guidelines for protecting the confidentiality of health care information. HIPAA lists seventeen pieces of textual PHI of which the following appear in medical discharge summaries: first and last names of patients, their health proxies, and family members; doctors' first and last names; identification numbers; telephone, fax, and pager numbers; hospital names; geographic locations; and dates. Removing PHI from medical documents is the goal of deidentification.

This paper presents a method based on a statistical representation of *local* context for automatically removing explicit PHI from medical discharge summaries, despite the often ungrammatical, fragmented, and ad hoc language of these documents, even when some words in the documents are ambiguous between PHI and non-PHI (e.g., "Huntington" as the name of a person and as the name of a disease), and even when some of the PHI cannot be found in dictionaries (e.g., misspelled and/or foreign names). This method differs from traditional approaches to deidentification in its independence from dictionaries and hand-tailored heuristics. It applies statistical named entity recognition (NER) methods to the more challenging task of deidenti-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 65–73,
New York, June 2006.

fication but differs from traditional NER approaches in its heavy reliance on a statistical representation of local context. Finally, this approach targets all PHI that appear in medical discharge summaries. Experiments reported in this paper show that context plays a more important role in deidentification than dictionaries, and that a statistical representation of local context contributes more to deidentification than global context.

2 Related Work

In the literature, named entities such as people, places, and organizations mentioned in news articles have been successfully identified by various approaches (Bikel et al., 1999; McCallum et al., 2000; Riloff and Jones, 1996; Collins and Singer, 1999; Hobbs et al., 1996). Most of these approaches are tailored to a particular domain, e.g., understanding disaster news; they exploit both the characteristics of the entities they focus on and the contextual clues related to these entities.

In the biomedical domain, NER has focused on identification of biological entities such as genes and proteins (Collier et al., 2000; Yu et al., 2002). Various statistical approaches, e.g., a maximum entropy model (Finkel et al., 2004), HMMs and SVMs (GuoDong et al., 2005), have been used with various feature sets including surface and syntactic features, word formation patterns, morphological patterns, part-of-speech tags, head noun triggers, and coreferences.

Deidentification refers to the removal of identifying information from records. Some approaches to deidentification have focused on particular categories of PHI, e.g., Taira et al. focused on only patient names (2002), Thomas et al. focused on proper names including doctors' names (2002). For full deidentification, i.e., removal of *all* PHI, Gupta et al. used "a complex set of rules, dictionaries, pattern-matching algorithms, and Unified Medical Language System" (2004). Sweeney's Scrub system employed competing algorithms that used patterns and lexicons to find PHI. Each of the algorithms included in her system specialized in one kind of PHI, each calculated the probability that a given word belonged to the class of PHI that it specialized in, and the algorithm with the highest precedence and the highest probability labelled the given word. This system identified 99-100% of all PHI in the test corpus of patient records and letters to physicians (1996).

We use a variety of features to train a support vector machine (SVM) that can automatically extract local context cues and can recognize PHI (even when some PHI are ambiguous between PHI and non-PHI, and even when PHI do not appear in dictionaries). We compare this approach with three others: a heuristic rule-based approach (Douglass, 2005), the SNoW (Sparse Network of Winnows) system's NER component (Roth and Yih, 2002), and IdentiFinder (Bikel et al., 1999). The heuristic rule-based system relies heavily on dictionaries. SNoW and IdentiFinder consider some representation of the local context of words; they also rely on information about global context. Local context helps them recognize stereotypical names and name structures. Global context helps these systems update the probability of observing a particular entity type based on the other entity types contained in the sentence. We hypothesize that, given the mostly fragmented and ungrammatical nature of discharge summaries, local context will be more important for deidentification than global context. We further hypothesize that local context will be a more reliable indication of PHI than dictionaries (which can be incomplete). The results presented in this paper show that SVMs trained with a statistical representation of local context outperform all baselines. In other words, a classifier that relies heavily on local context (very little on dictionaries, and not at all on global context) outperforms classifiers that rely either on global context or dictionaries (but make much less use of local context). Global context cannot contribute much to deidentification when the language of documents is fragmented; dictionaries cannot contribute to deidentification when PHI are either missing from dictionaries or are ambiguous between PHI and non-PHI. Local context remains a reliable indication of PHI under these circumstances.

The features used for our SVM-based system can be enriched in order to automatically acquire more and varied local context information. The features discussed in this paper have been chosen because of their simplicity and effectiveness on both grammatical and ungrammatical free text.

3 Corpora

Discharge summaries are the reports generated by medical personnel at the end of a patient's hospital stay and contain important information about the patient's health. Linguistic processing of these documents is challenging, mainly because these reports are full of medical jargon, acronyms, shorthand notations, misspellings, ad hoc language, and fragments of sentences. Our goal is to identify the PHI used in discharge summaries even when text is fragmented and ad hoc, even when many words in the summaries are ambiguous between PHI and non-PHI, and even when many PHI contain misspelled or foreign words.

In this study, we worked with various corpora consisting of discharge summaries. One of these corpora was obtained already deidentified[1]; i.e., (many) PHI (and some non-PHI) found in this corpus had been replaced with the generic placeholder `[REMOVED]`. An excerpt from this corpus is below:

> HISTORY OF PRESENT ILLNESS: The patient is a 77-year-old-woman with long standing hypertension who presented as a Walk-in to me at the [REMOVED] Health Center on [REMOVED]. Recently had been started q.o.d. on Clonidine since [REMOVED] to taper off of the drug. Was told to start Zestril 20 mg. q.d. again. The patient was sent to the [REMOVED] Unit for direct admission for cardioversion and anticoagulation, with the Cardiologist, Dr. [REMOVED] to follow.
>
> SOCIAL HISTORY: Lives alone, has one daughter living in [REMOVED]. Is a non-smoker, and does not drink alcohol.
>
> HOSPITAL COURSE AND TREATMENT: During admission, the patient was seen by Cardiology, Dr. [REMOVED], was started on IV Heparin, Sotalol 40 mg PO b.i.d. increased to 80 mg b.i.d., and had an echocardiogram. By [REMOVED] the patient had better rate control and blood pressure control but remained in atrial fibrillation. On [REMOVED], the patient was felt to be medically stable.
>
> ...

We hand-annotated this corpus and experimented with it in several ways: we used it to generate a corpus of discharge summaries in which the `[REMOVED]` tokens were replaced with appropriate, fake PHI obtained from dictionaries[2] (Douglass, 2005); we used it to generate a second corpus in which most of the `[REMOVED]` tokens and some of the remaining text were appropriately replaced with lexical items that were ambiguous between PHI and non-PHI[3]; we used it to generate another corpus in which all of the `[REMOVED]` tokens corresponding to names were replaced with appropriately formatted entries that could not be found in dictionaries[4]. For all of these corpora, we generated realistic substitutes for the `[REMOVED]` tokens using dictionaries (e.g., a dictionary of names from US Census Bureau) and patterns (e.g., names of people could be of the formats, "Mr. F. Lastname", "Firstname Lastname", "Lastname", "F. M. Lastname", etc.; dates could appear as "dd/mm/yy", "dd MonthName, yyyy", "ddth of MonthName, yyyy", etc.). In addition to these reidentified corpora (i.e., corpora generated from previously deidentified data), we also experimented with authentic discharge summaries[5]. The approximate distributions of PHI in the reidentified corpora and in the authentic corpus are shown in Table 1.

Class	No. in reidentified summaries	No. in authentic summaries
Non-PHI	17872	112720
Patient	1047	287
Doctor	311	730
Location	24	84
Hospital	592	651
Date	735	1933
ID	36	477
Phone	39	32

Table 1: Distribution of different PHI (in terms of number of words) in the corpora.

4 Baseline Approaches

4.1 Rule-Based Baseline: Heuristic+Dictionary

Traditional deidentification approaches rely heavily on dictionaries and hand-tailored heuristics.

[1] Authentic clinical data is very difficult to obtain for privacy reasons; therefore, the initial implementation of our system was tested on previously deidentified data that we reidentified.

[2] e.g., John Smith initiated radiation therapy ...

[3] e.g., D. Sessions initiated radiation therapy...

[4] e.g., O. Ymfgkstjj initiated radiation therapy ...

[5] We obtained authentic discharge summaries with real PHI in the final stages of this project.

We obtained one such system (Douglass, 2005) that used three kinds of dictionaries:

- PHI lookup tables for female and male first names, last names, last name prefixes, hospital names, locations, and states.
- A dictionary of "common words" that should never be classified as PHI.
- Lookup tables for context clues such as titles, e.g., Mr.; name indicators, e.g., proxy, daughter; location indicators, e.g., lives in.

Given these dictionaries, this system identifies keywords that appear in the PHI lookup tables but do not occur in the common words list, finds approximate matches for possibly misspelled words, and uses patterns and indicators to find PHI.

4.2 SNoW

SNoW is a statistical classifier that includes a NER component for recognizing entities and their relations. To create a hypothesis about the entity type of a word, SNoW first takes advantage of "words, tags, conjunctions of words and tags, bigram and trigram of words and tags", number of words in the entity, bigrams of words in the entity, and some attributes such as the prefix and suffix, as well as information about the presence of the word in a dictionary of people, organization, and location names (Roth and Yih, 2002). After this initial step, it uses the possible relations of the entity with other entities in the sentence to strengthen or weaken its hypothesis about the entity's type. The constraints imposed on the entities and their relationships constitute the global context of inference. Intuitively, information about global context and constraints imposed on the relationships of entities should improve recognition of both entities and relations. Roth and Yih (2002) present results that support this hypothesis.

SNoW can recognize entities that correspond to people, locations, and organizations. For deidentification purposes, all of these entities correspond to PHI; however, they do not constitute a comprehensive set. We evaluated SNoW only on the PHI it is built to recognize. We trained and tested its NER component using ten-fold cross-validation on each of our corpora.

4.3 IdentiFinder

IdentiFinder uses Hidden Markov Models to learn the characteristics of names of entities, including people, locations, geographic jurisdictions, organizations, dates, and contact information (Bikel et al., 1999). For each named entity class, this system learns a bigram language model which indicates the likelihood that a sequence of words belongs to that class. This model takes into consideration features of words, such as whether the word is capitalized, all upper case, or all lower case, whether it is the first word of the sentence, or whether it contains digits and punctuation. Thus, it captures the local context of the target word (i.e., the word to be classified; also referred to as TW). To find the names of all entities, the system finds the most likely sequence of entity types in a sentence given a sequence of words; thus, it captures the global context of the entities in a sentence.

We obtained this system pre-trained on a news corpus and applied it to our corpora. We mapped its entity tags to our PHI and non-PHI labels. Admittedly, testing IdentiFinder on the discharge summaries puts this system at a disadvantage compared to the other statistical approaches. However, despite this shortcoming, IdentiFinder helps us evaluate the contribution of global context to deidentification.

5 SVMs with Local Context

We hypothesize that systems that rely on dictionaries and hand-tailored heuristics face a major challenge when particular PHI can be used in many different contexts, when PHI are ambiguous, or when the PHI cannot be found in dictionaries. We further hypothesize that given the ungrammatical and ad hoc nature of our data, despite being very powerful systems, IdentiFinder and SNoW may not provide perfect deidentification. In addition to being very fragmented, discharge summaries do not present information in the form of relations between entities, and many sentences contain only one entity. Therefore, the global context utilized by IdentiFinder and SNoW cannot contribute reliably to deidentification. When run on discharge summaries, the strength of these systems comes from their ability to recognize the structure of the names of different entity types and the local contexts of these entities.

Discharge summaries contain patterns that can serve as local context. Therefore, we built an SVM-based system that, given a target word (TW), would accurately predict whether the TW was part of PHI. We used a development corpus to find features that captured as much of the immediate context of the TW as possible, paying particular attention to cues human annotators found useful for deidentification. We added to this some surface characteristics for the TW itself and obtained the following features: the TW itself, the word before, and the word after (all lemmatized); the bigram before and the bigram after TW (lemmatized); the part of speech of TW, of the word before, and of the word after; capitalization of TW; length of TW; MeSH ID of the noun phrase containing TW (MeSH is a dictionary of Medical Subject Headings and is a subset of the Unified Medical Language System (UMLS) of the National Library of Medicine); presence of TW, of the word before, and of the word after TW in the name, location, hospital, and month dictionaries; the heading of the section in which TW appears, e.g., "History of Present Illness"; and, whether TW contains "-" or "/" characters. Note that some of these features, e.g., capitalization and punctuation within TW, were also used in IdentiFinder.

We used the SVM implementation provided by LIBSVM (Chang and Lin, 2001) with a linear kernel to classify each word in the summaries as either PHI or non-PHI based on the above-listed features. We evaluated this system using ten-fold cross-validation.

6 Evaluation

Local context contributes differently to each of the four deidentification systems. Our SVM-based approach uses *only* local context. The heuristic, rule-based system relies heavily on dictionaries. IdentiFinder uses a simplified representation of local context and adds to this information about the global context as represented by transition probabilities between entities in the sentence. SNoW uses local context as well, but it also makes an effort to benefit from relations between entities. Given the difference in the strengths of these systems, we compared their performance on both the reidentified and authentic corpora (see Section 3). We hypothesized that given the nature of medical discharge summaries, IdentiFinder would not be able to find enough global context and SNoW would not be able to make use of relations (because many sentences in this corpus contain only one entity). We further hypothesized that when the data contain words ambiguous between PHI and non-PHI, or when the PHI cannot be found in dictionaries, the heuristic, rule-based approach would perform poorly. In all of these cases, SVMs trained with local context information would be sufficient for proper deidentification.

To compare the SVM approach with IdentiFinder, we evaluated both on PHI consisting of names of people (i.e., patient and doctor names), locations (i.e., geographic locations), and organizations (i.e., hospitals), as well as PHI consisting of dates, and contact information (i.e., phone numbers, pagers). We omitted PHI representing ID numbers from this experiment in order to be fair to IdentiFinder which was not trained on this category. To compare the SVM approach with SNoW, we trained both systems with only PHI consisting of names of people, locations, and organizations, i.e., the entities that SNoW was designed to recognize.

6.1 Deidentifying Reidentified and Authentic Discharge Summaries

We first deidentified:

- Previously deidentified discharge summaries into which we inserted invented but realistic surrogates for PHI without deliberately introducing ambiguous words or words not found in dictionaries, and
- Authentic discharge summaries with real PHI.

Our experiments showed that SVMs with local context outperformed all other approaches. On the reidentified corpus, SVMs gave an F-measure of 97.2% for PHI. In comparison, IdentiFinder, having been trained on the news corpus, gave an F-measure of 67.4% and was outperformed by the heuristic+dictionary approach (see Table 2).[6]

[6]Note that in deidentification, recall is much more important than precision. Low recall indicates that many PHI remain in the documents and that there is high risk to patient privacy. Low precision means that words that do not correspond to PHI have also been removed. This hurts the integrity of the data but does not present a risk to privacy.

We evaluated SNoW only on the three kinds of entities it is designed to recognize. We cross-validated it on our corpora and found that its performance in recognizing people, locations, and organizations was 96.2% in terms of F-measure (see Table 3[7]). In comparison, our SVM-based system, when retrained to only consider people, locations, and organizations so as to be directly comparable to SNoW, had an F-measure of 98%.[8]

Method	Class	P	R	F
SVM	PHI	96.8%	97.7%	**97.2%**
IFinder	PHI	60.2%	76.7%	67.4%
H+D	PHI	88.9%	67.6%	76.8%
SVM	Non-PHI	99.6%	99.5%	**99.6%**
IFinder	Non-PHI	95.8%	91.4%	93.6%
H+D	Non-PHI	95.2%	95.2%	95.2%

Table 2: Precision, Recall, and F-measure on **reidentified** discharge summaries. IFinder refers to IdentiFinder, H+D refers to heuristic+dictionary approach.

Method	Class	P	R	F
SVM	PHI	97.7%	98.2%	**98.0%**
SNoW	PHI	96.1%	96.2%	96.2%
SVM	Non-PHI	99.8%	99.8%	**99.8%**
SNoW	Non-PHI	99.6%	99.6%	99.6%

Table 3: Evaluation of SNoW and SVM on recognizing people, locations, and organizations found in **reidentified** discharge summaries.

Similarly, on the authentic discharge summaries, the SVM approach outperformed all other approaches in recognizing PHI (see Tables 4 and 5).

6.2 Deidentifying Data with Ambiguous PHI

In discharge summaries, the same words can appear both as PHI and as non-PHI. For example, in the same corpus, the word "Swan" can appear both as the name of a medical device (i.e., "Swan Catheter") and as the name of a person, etc. Ideally, we would like to deidentify data even when many words in the

[7]The best performances are marked in bold in all of the tables in this paper.

[8]For all of the corpora presented in this paper, a performance difference of 1% or more is statistically significant at $\alpha = 0.05$.

Method	Class	P	R	F
SVM	PHI	97.5%	95.0%	**96.2%**
IFinder	PHI	25.2%	45.2%	32.3%
H+D	PHI	81.9%	87.6%	84.7%
SVM	Non-PHI	99.8%	99.9%	**99.9%**
IFinder	Non-PHI	97.1%	93.3%	95.2%
H+D	Non-PHI	99.6%	99.6%	99.6%

Table 4: Evaluation on **authentic** discharge summaries.

Method	Class	P	R	F
SVM	PHI	97.4%	93.8%	**95.6%**
SNoW	PHI	93.7%	93.4%	93.6%
SVM	Non-PHI	99.9%	100%	**100%**
SNoW	Non-PHI	99.9%	99.9%	99.9%

Table 5: Evaluation of SNoW and SVM on **authentic** discharge summaries.

corpus are ambiguous between PHI and non-PHI. We hypothesize that given ambiguities in the data, context will play an important role in determining whether the particular instance of the word is PHI and that given the many fragmented sentences in our corpus, local context will be particularly useful. To test these hypotheses, we generated a corpus by reidentifying the previously deidentified corpus with words that were ambiguous between PHI and non-PHI, making sure to use each ambiguous word both as PHI and non-PHI, and also making sure to cover all acceptable formats of all PHI (see Section 3). The resulting distribution of PHI is shown in Table 6.

Class	Total # Words	# Ambiguous Words
Non-PHI	19296	3781
Patient	1047	514
Doctor	311	247
Location	24	24
Hospital	592	82
Date	736	201
ID	36	0
Phone	39	0

Table 6: Distribution of PHI when some words are ambiguous between PHI and non-PHI.

Our results showed that, on this corpus, the SVM-based system accurately recognized 91.9% of all PHI; its performance, measured in terms of F-measure was also significantly better than all other approaches both on the complete corpus containing ambiguous entries (see Table 7 and Table 8) and only on the ambiguous words in this corpus (see Table 9).

Method	Class	P	R	F
SVM	PHI	92.0%	92.1%	**92.0%**
IFinder	PHI	45.4%	71.4%	55.5%
H+D	PHI	70.1%	46.6%	56.0%
SVM	Non-PHI	98.9%	98.9%	**98.9%**
IFinder	Non-PHI	95.0%	86.5%	90.1%
H+D	Non-PHI	92.7%	92.7%	92.7%

Table 7: Evaluation **on the corpus containing ambiguous data**.

Method	Class	P	R	F
SVM	PHI	92.1%	92.8%	**92.5%**
SNoW	PHI	91.6%	77%	83.7%
SVM	Non-PHI	99.3%	99.2%	**99.3%**
SNoW	Non-PHI	97.6%	99.3%	98.4%

Table 8: Evaluation of SNoW and SVM on **ambiguous** data.

Method	Class	P	R	F
SVM	PHI	90.2%	87.5%	**88.8%**
IFinder	PHI	55.8%	64.0%	59.6%
H+D	PHI	59.8%	24.3%	34.6%
SNoW	PHI	91.6%	82.9%	87.1%
SVM	Non-PHI	90.5%	92.7%	91.6%
IFinder	Non-PHI	69.0%	61.3%	64.9%
H+D	Non-PHI	59.9%	87.4%	71.1%
SNoW	Non-PHI	90.4%	95.5%	**92.9%**

Table 9: Evaluation **only on ambiguous** people, locations, and organizations found in ambiguous data.

6.3 Deidentifying PHI Not Found in Dictionaries

Some medical documents contain foreign or misspelled names that need to be effectively removed. To evaluate the different deidentification approaches under such circumstances, we generated a corpus in which the names of people, locations, and hospitals were all random permutations of letters. The resulting words were not found in any dictionaries but followed the general format of the entity name category to which they belonged. The distribution of PHI in this third corpus is in Table 10.

Class	Total PHI	PHI Not in Dict.
Non-PHI	17872	0
Patient	1045	1045
Doctor	302	302
Location	24	24
Hospital	376	376
Date	735	0
ID	36	0
Phone	39	0

Table 10: Distribution of PHI in the corpus where all PHI associated with names are randomly generated so as not to be found in dictionaries.

On this data set, dictionaries cannot contribute to deidentification because none of the PHI appear in dictionaries. Under these conditions, proper deidentification relies completely on context. Our results showed that SVM approach outperformed all other approaches on this corpus also (Tables 11 and 12).

Method	Class	P	R	F
SVM	PHI	94.0%	96.0%	**95.0%**
IFinder	PHI	55.1%	65.5%	59.8%
H+D	PHI	76.4%	27.8%	40.8%
SVM	Non-PHI	99.4%	99.1%	**99.3%**
IFinder	Non-PHI	94.4%	91.6%	92.9%
H+D	Non-PHI	90.7%	90.7%	90.7%

Table 11: Evaluation **on the corpus containing PHI not in dictionaries**.

Of only the PHI not found in dictionaries, 95.5% was accurately identified by the SVM approach. In comparison, the heuristic+dictionary approach accurately identified those PHI that could not be found in dictionaries 11.1% of the time, IdentiFinder recognized these entities 76.7% of the time and SNoW gave an accuracy of 79% (see Table 13).

Method	Class	P	R	F
SVM	PHI	93.9%	96.0%	**95.0%**
SNoW	PHI	93.7%	79.0%	85.7%
SVM	Non-PHI	99.6%	99.4%	**99.5%**
SNoW	Non-PHI	98.0%	99.5%	98.7%

Table 12: Evaluation of SNoW and SVM on the people, locations, and organizations found in the corpus containing PHI not found in dictionaries.

Method	SVM	IFinder	SNoW	H+D
Precision	95.5%	76.7%	79.0%	11.1%

Table 13: Precision on only the PHI not found in dictionaries.

6.4 Feature Importance

As hypothesized, in all experiments, the SVM-based approach outperformed all other approaches. SVM's feature set included a total of 26 features, 12 of which were dictionary-related features (excluding MeSH). Information gain showed that the most informative features for deidentification were the TW, the bigram before TW, the bigram after TW, the word before TW, and the word after TW.

Note that the TW itself is important for classification; many of the non-PHI correspond to common words that appear in the corpus frequently and the SVM learns the fact that some words, e.g., the, admit, etc., are never PHI. In addition, the context of TW (captured in the form of unigrams and bigrams of words and part-of-speech tags surrounding TW) contributes significantly to deidentification.

There are many ways of automatically capturing context. In our data, unigrams and bigrams of words and their part-of-speech tags seem to be sufficient for a statistical representation of local context. The global context, as represented within IdentiFinder and SNoW, could not contribute much to deidentification on this corpus because of the fragmented nature of the language of these documents, because most sentences in this corpus contain only one entity, and because many sentences do not include explicit relations between entities. However, there is enough structure in this data that can be captured by local context; lack of relations between entities and the inability to capture global context do not hold us back from almost perfect deidentification.

7 Conclusion

We presented a set of experimental results that show that local context contributes more to deidentification than dictionaries and global context when working with medical discharge summaries. These documents are characterized by incomplete, fragmented sentences, and ad hoc language. They use a lot of jargon, many times omit subjects of sentences, use entity names that can be misspelled or foreign words, can include entity names that are ambiguous between PHI and non-PHI, etc. Similar documents in many domains exist; our experiments here show that even on such challenging corpora, local context can be exploited to identify entities. Even a rudimentary statistical representation of local context, as captured by unigrams and bigrams of lemmatized keywords and part-of-speech tags, gives good results and outperforms more sophisticated approaches that rely on global context. The simplicity of the representation of local context and the results obtained using this simple representation are particularly promising for many tasks that require processing ungrammatical and fragmented text where global context cannot be counted on.

8 Acknowledgements

This publication was made possible by grant number R01-EB001659 from the National Institute of Biomedical Imaging and Bioengineering; by grant number N01-LM-3-3513 on National Multi-Protocol Ensemble for Self-Scaling Systems for Health from National Library of Medicine; and, by grant number U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Library of Medicine.

We are grateful to Professor Peter Szolovits and Dr. Boris Katz for their insights, and to Professor Carol Doll, Sue Felshin, Gregory Marton, and Tian He for their feedback on this paper.

References

J. J. Berman. 2002. Concept-Match Medical Data Scrubbing: How Pathology Text Can Be Used in Research. *Archives of Pathology and Laboratory Medicine*, 127(6).

D. M. Bikel, R. Schwartz, and R. M. Weischedel. 1999.

An Algorithm That Learns What's in a Name. *Machine Learning Journal Special Issue on Natural Language Learning*, 34(1/3).

C. Chang and C. Lin. 2001. *LIBSVM: a Library for Support Vector Machines*.

N. Collier, C. Nobata, and J. Tsujii. 2000. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of COLING*.

M. Collins and Y. Singer. 1999. Unsupervised Models for Named Entity Classification. *Proceedings of EMNLP*.

J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair. 2004. Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. *Proceedings of Joint Workshop on Natural Language Processing in Biomedicine and its Applications at COLING*.

R. Gaizauskas, G. Demetriou, P. Artymiuk, and P. Willett. 2003. Protein Structures and Information Extraction from Biological Texts: The PASTA System. *Bioinformatics*, 19(1).

Z. GuoDong, Z. Jie, S. Jian, S. Dan, T. ChewLim. 2005. Recognizing Names in Biomedical Texts: a Machine Learning Approach. *Bioinformatics*, 20(7).

D. Gupta, M. Saul, J. Gilbertson. 2004. Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. *American Journal of Clinical Pathology*, 121(6).

J. R. Hobbs, D. E. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, and M. Tyson. 1996. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *In Finite State Devices for Natural Language Processing*. MIT Press, Cambridge, MA.

M. Douglass, G. D. Clifford, A. Reisner, G. B. Moody, R. G. Mark. 2005. Computer-Assisted De-Identification of Free Text in the MIMIC II Database. *Computers in Cardiology*. 32:331-334.

A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. *Proceedings of ICML.*

E. Riloff and R. Jones. 1996. Automatically Generating Extraction Patterns from Untagged Text. *Proceedings of AAAI-96*.

D. Roth and W. Yih. 2002. Probabilistic Reasoning for Entity and Relation Recognition. *Proceedings of COLING*.

P. Ruch, R. H. Baud, A. Rassinoux, P. Bouillon, G. Robert. 2000. Medical Document Anonymization with a Semantic Lexicon. *Proceedings of AMIA*.

M. Surdeanu, S. M. Harabagiu, J. Williams, and P. Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. *Proceedings of ACL 2003*.

L. Sweeney. 1996. Replacing personally-identifying information in medical records, the scrub system. *Journal of the American Medical Informatics Association*.

R. K. Taira, A. A. T. Bui, H. Kangarloo. 2002. Identification of patient name references within medical documents using semantic selectional restrictions. *Proceedings of AMIA*.

S. M. Thomas, B. Mamlin, G. Schadow, C. McDonald. 2002. A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. *Proceedings of AMIA*.

H. Yu, V. Hatzivassiloglou, C. Friedman, W. J. Wilbur. 2002. Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. *Proceedings of AMIA*.