Research Paper ■

State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework

GYÖRGY SZARVAS, RICHÁRD FARKAS, RÓBERT BUSA-FEKETE

Abstract **Objective:** The anonymization of medical records is of great importance in the human life sciences because a de-identified text can be made publicly available for non-hospital researchers as well, to facilitate research on human diseases. Here the authors have developed a de-identification model that can successfully remove personal health information (PHI) from discharge records to make them conform to the guidelines of the Health Information Portability and Accountability Act.

Design: We introduce here a novel, machine learning-based iterative Named Entity Recognition approach intended for use on semi-structured documents like discharge records. Our method identifies PHI in several steps. First, it labels all entities whose tags can be inferred from the structure of the text and it then utilizes this information to find further PHI phrases in the flow text parts of the document.

Measurements: Following the standard evaluation method of the first Workshop on Challenges in Natural Language Processing for Clinical Data, we used token-level Precision, Recall and $F_{\beta=1}$ measure metrics for evaluation.

Results: Our system achieved outstanding accuracy on the standard evaluation dataset of the de-identification challenge, with an F measure of 99.7534% for the best submitted model.

Conclusion: We can say that our system is competitive with the current state-of-the-art solutions, while we describe here several techniques that can be beneficial in other tasks that need to handle structured documents such as clinical records.

■ **J Am Med Inform Assoc.** 2007;14:574–580. DOI 10.1197/j.jamia.M2441.

Introduction

The identification and classification of named entities in a plain text is of key importance in numerous natural language processing applications like the de-identification of clinical records. This task is crucial in the human life sciences because a de-identified text can be made publicly available for non-hospital researchers as well to facilitate research on human diseases. However, the records about the patients include explicit personal health information (PHI), and this fact hinders the release of many useful data sets because their release would jeopardize individual patient rights. According to the guidelines of Health Information Portability and Accountability Act (HIPAA) the medical discharge summaries released must be free of the following seventeen categories of textual PHI: 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 these kinds of PHI is the main goal of the de-identification process. Anonymization goes one step beyond the removal of personal information and attempts to identify and classify personal information in the text to one of the HIPAA-defined categories. This categorization permits the replacement of personal data instead of simple deletion, and it has several advantages. First, the replace-

Affiliations of the authors: Department of Informatics, University of Szeged (GS); Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged (RF, RB-F), Szeged, Hungary.

Supported in part by the Computer and Automation Research Institute of the Hungarian Academy of Sciences and by NKFP-2/051/2004.

The authors thank the task organizers for organizing the challenge and their help; and the anonymous reviewers for valuable comments.

Correspondence and reprints: György Szarvas, University of Szeged, Department of Informatics, 6720, Szeged, Árpád tér 2., Hungary; e-mail: <szarvas@inf.u-szeged.hu>.

Received for review: 3/16/2007; accepted for publication: 6/11/2007.

[a] The best performing system (Wellner et al., 2006)[12] at the workshop was also the adaptation of an existing NER model to clinical data. Our system came second, with the difference in performance between the two systems being below the level of significance. These facts prove the feasibility of adapting a NER system to anonymization.

[b] Guo et al.'s system (2006)[14] made use of only a subset of the available training data, due to SVM's higher time complexity.

[c] This model was a similar boosted decision tree classifier, but without regular expression features, document heading information and iterative learning process.

[d] Personal Health Information had to be concealed from the challenge participants. To achieve this, the organizers removed all PHI from the corpus and replaced them with artificially generated realistic substitutes. For more information about this, see Uzuner et al. (2007).[1]

[e] ITR2_VOTE is an out-of-competition result as we had no time to prepare all three second iteration systems in the evaluation period of the competition. The differences between the three best systems are only marginal, however.

[f] The higher values in the first column of Table 2 tell us that our model is better at precision than recall.

ment of PHIs with artificially generated realistic substitutes preserves the readability of text and, furthermore, the artificial substitutes actually disguise those very few personal information that remain in the document (the reader will never know whether a single label was the original or a substitute).

In this paper we present results on the I2B2 de-identification shared task. For a detailed description of the shared task challenge, the corpus, results and lessons learned, see Uzuner et al. (2007).[1]

In the literature many de-identification approaches have been introduced. Some approaches target the recognition (and removal) of particular types of PHI like Taira et al.'s (2002)[2] system which focuses on patient names, or Thomas et al.'s method (2002),[3] which seeks to identify person names (both patients and doctors). There are several approaches that carry out the full de-identification of medical texts. These are based either on a pattern-matching algorithm that uses a thesaurus (Sweeny, 1996, Ruch et al. 2000);[4, 5] a combination of rule-based systems and pattern matching using dictionaries (Douglass et al., 2005)[6] and the Unified Medical Language System (Gupta et al., 2004)[7] or on a statistical model (Sibanda and Uzuner, 2006).[8] In this paper we use some Named Entity Recognition (NER) techniques[a] for the task of the de-identification of clinical records. For a more detailed overview on NER, see Szarvas et al. (2006).[9]

The participants of the first Workshop on Challenges in Natural Language Processing for Clinical Data submitted both rule-based (Guillen, 2006)[10] and statistical approaches to the de-identification task. The best performing systems used Conditional Random Fields (Aramaki et al., 2006; Wellner et al., 2006);[11, 12] boosting and C4.5 decision tree learning (presented here) and Support Vector Machines (Hara, 2006; Guo et al., 2006)[13, 14] to solve the anonymization problem.

Our paper is organized as follows. In the next section we will discuss the feature sets used and we will introduce our new iterative method. Then we provide a brief overview of the Machine Learning models we employed in experiments. Next we will give a summary of the performance of our DEID system on the I2B2 dataset, and lastly, we summarize our results and conclusions drawn from the study.

Methods

We extended our newswire NER model to the de-identification task by adding two novel feature types, and by applying an iterative learning method described below that utilizes the information given in the structured parts of the texts to improve the accuracy of PHI recognition in flow text.

Our method follows Sibanda and Uzuner's system in the sense that we built a corpus-based statistical model,[8] but it is different from previous approaches in two ways. First, we excluded all deep knowledge resources from our model (like syntactic information, UMLS or Medical Subject Headings entries applied by previous models). This way our system is entirely based on contextual and surface patterns, which makes us assume it is easily retargetable for similar tasks. Second, we adapted a system designed for entity recognition in newswire texts by simply replacing the sources of features (to clinical documents) and adding a minimal amount of domain specific extensions to the system, that will be described in detail later on.

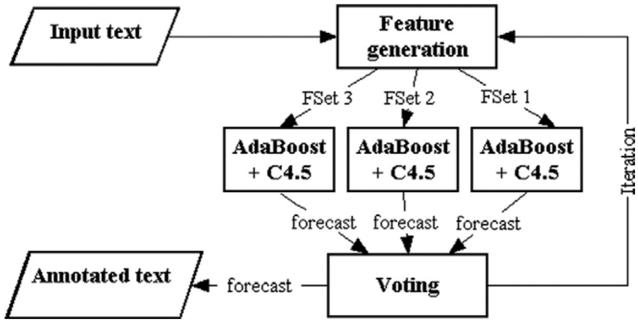

Figure 1. A schematic overview of our complex model.

We regard the de-identification problem essentially as the classification of separate tokens. We believe that this approach is competitive with the—theoretically more suitable—sequence tracking algorithms (like Hidden Markov Models, Maximum Entropy approaches or Conditional Random Fields); hence we applied a decision tree learning algorithm. Of course our model is capable of taking into account the relationship between consecutive words using a window of appropriate size.

Figure 1 sketches the structure of our complex model; the details of its building blocks are described in this section.

Feature Set

We employed a very rich feature set for our word-level classification model, describing the characteristics of the word itself along with its actual context (a moving window of size four). We did not use deep knowledge information, like Part of speech (POS) (Hara, 2006),[13] chunk codes or ontologies; or any complex domain specific resources, like MeSH IDs in (Sibanda and Uzuner, 2006).[8]

Our features fell into the following main categories:

- *Orthographical features:* capitalization, word length, common bit information about the word form (contains a digit or not, has uppercase characters inside the word, has punctuation marks inside the word, has digit inside the word, the word is roman or Arabic number) and several regular expressions that describe the common surface characteristics of AGE, DATE, ID and PHONE classes
- *Frequency information:* We gathered the frequencies of tokens from a huge corpus consisting texts collected from the Internet. We used the frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token.
- *Phrasal information:* a forecasted class of several preceding words (we used an online evaluation) and common phrase suffixes (e.g. *"Hospital"*) seen in the train set
- *Dictionaries:* We collected five lists from the Internet: first names, geographical locations in the US, names of countries, world's largest cities, names of diseases; a list containing non-PHI tokens from the train data and a list containing non named entity tokens from an external corpus (CoNLL-2003 newswire dataset).
- *Contextual information:* sentence position, the closest section heading, trigger words from the train text that often

precede or follow PHI (see below), whether the word fell between quotes, whether the word fell between brackets, the whole context is in uppercase.

We applied the feature set used for a common domain (in our previous studies, Szarvas et al. 2006)[9] and introduced only two new features to adapt the system to medical records: i) regular expressions (see Kleene 1956)[15] that try to cover the well formulated classes, and *ii)* our model can infer knowledge from the structure of the document using the common headings observed in typical discharge records (we extracted the most frequent subject headings from the training set).

The use of trigger words is not straightforward, however, so we used them in three different ways in our experiments: we collected the three preceding and three subsequent tokens of all *tagged tokens* in the train set (we refer to this feature set as the *token trigger* later on); similarly, we collected subsequent tokens of *tagged phrases* and used a wider window for this feature (*phrase trigger*); and third we collected the bi- and trigrams around the phrases of the train texts (*trigram trigger*).

Phrase trigger means the kind of tokens that appear before or after whole PHI phrases (perhaps several tokens long). For example, *"M.D."* is a trigger for *DOCTOR* class phrases with an offset of +2, as the usual pattern in text is *"DOCTOR-_NAME, M.D."*. Of course, as the classification itself is performed by a token-level model, this feature helps us to identify just the first or last token of a doctor name (depending on the sign of the offset). In *"John Smith, M.D."* only the instance for token *"Smith"* has this feature set to true. In case of *token triggering*, we collect this kind of information for all tagged tokens, not phrases. This way *"M.D."* should be a similarly strong trigger for *DOCTOR* class with offset +2. Furthermore, it becomes a somewhat weaker token-trigger for *DOCTOR* class with an offset of +3 /it typically appears with +3 offset to *DOCTOR* tokens (like *"John Smith, M.D."*) and to non-PHI tokens (as in *"visited Dr. Smith, M.D."*)/. *Bigram/trigram triggers* collect not single trigger tokens which imply a class label, but 2 or 3 token-long sequences. In this model, *", M.D."* should be a strong indicator of *DOCTOR* class, not *","* with offset +1 or *"M.D."* with +2 offset on their own.

The collected trigger lists for each of the three cases were filtered according to their frequency and information gain on the class labels. A significant difference in the predictions was noticed in the experiments where only the use of triggers was changed; hence we decided to combine their forecasts to exploit their advantages better.

Classifiers

Boosting[16] and C4.5[17] are well known algorithms for those who are acquainted with pattern recognition. Boosting has been applied successfully to improve the performance of decision trees in several NLP tasks. A system that made use of AdaBoost and fixed depth decision trees came first on the Computational Natural Language Learning Conference shared task on NER in 2002,[18] but gave somewhat worse results in 2003 (it was ranked fifth with an F measure of 85.0%).[19] We have not found any other competitive results for NER using decision tree classifiers and AdaBoost.

Boosting was introduced by Shapire as a way of improving the performance of a weak learning algorithm. The algorithm generates a set of classifiers (of the same type) by reweighting the examples of the original training data set and it makes a decision based on their votes. The final decision is made using a weighted voting schema for each classifier that is many times more accurate than the original model. In our investigation 30 iterations of Boosting were performed on each model as further iterations gave only slight improvement.

C4.5 is based on the well-known ID3 tree learning algorithm, which is able to learn pre-defined discrete classes from labelled examples. Classification is done by axis-parallel hyperplanes, and hence learning is very fast. This makes C4.5 a good subject for boosting. We built decision trees that had at least five instances per leaf, and used pruning with subtree raising and a confidence factor of *0.35.*

Combination of the Classifiers

We trained three similar classifiers that differed from each other only in the way triggers were used. We treated these models as separate hypotheses, and used the following decision function to obtain a final prediction: *if any two of the three learners' outputs coincided we accepted it as a joint decision, and forecasted non-PHI otherwise*. This cautious voting scheme is beneficial to system performance as a high rate of disagreement often means poor prediction accuracy. The special recall-sensitivity of the anonymization task could raise the question of applying different voting strategies. A voting scheme that assigns one of the eight PHI classes in the case of high disagreement of the three models (for example, accepting the prediction of the most accurate single model in such cases) might result in a somewhat lower F measure but would surely increase the recall of the system.

Iterative Learning

The structured parts of the text can be processed more easily than flow text, and the named entities in the record fields can occur in other parts of the text in the same or similar form. To utilize this latter fact in a first training phase we collected trusted named entities appearing in document sections under certain unambiguous headings. We considered a heading unambiguous if its cross-class Shannon entropy[20] was less than 0.1 on the train set. The named entities found in this first phase and their acronyms became trusted phrases and their lists were added to the feature set as an extra dictionary for a subsequent training phase. We will call this second learning step Iteration 1 (ITR1) later on. ITR2 will refer to a further, similar retraining step, using the labels assigned by an ITR1 model. The system of Aramaki et al. (2006)[11] used a similar approach to incorporate their label-consistency hypothesis to their model.

We made the hypothesis that the most significant trigger words (like *"Dr."*) indicate trusted phrases as well. But the experiments with this kind of trusted entities achieved worse results than ones without them, so we abandoned this hypothesis. This was probably caused by the artificially added ambiguity to PHI phrases in the data set (for example if we found a phrase *"Dr. He"* and accepted *"He"* as a trusted phrase, the model tended to treat all occurrences of the word *"He"* as the name of a doctor while it's a non-taggable common word in the majority of cases). We note here that this hypothesis might prove useful on real data. Fortunately, the structured parts of the data usually contain full formed phrases and thus incorporating PHI found there proved to be beneficial to the model.

In the final phase of the learning process we standardize the tagging of the same phrases, because our token based classification approach can fail on the proper tagging of a whole phrase in one context, while managing to do that in another—easier—context. We collected all predicted phrases of length two or more from the previous iteration and overwrote every occurrence of them with the predicted class of the longest matching phrase.

Results

We extracted first the features mentioned above for each token from the train set. One-hundred-thirty-eight numerically encodable attributes described each token (including features from a window around the token itself). Our previous experiments on NER problems showed that a feature space of this dimensionality can be handled by our learning algorithms for datasets not larger than 1 million tokens; hence we ignored any feature selection procedure.[b]

In our experiments we used an implementation based on the WEKA library,[21] an open-source data mining software written in Java. We split the train data into ten pieces (it was cut on document boundaries), and made ten-fold cross validation on these subsets.

Evaluation Methods and Preliminary Experiments

In our experiments we used two different kinds of evaluation: token level 8-way and 9-way F measure. 8-way evaluation excludes non-PHI true positives and thus measures the performance of identifying the 8 PHI classes, while 9-way evaluation considers non-PHI class as well. The latter metric takes into account the correct recognition of non-PHI, because this class is important for preserving the document's information content. The 9-way F measure was the official evaluation metric used for the I2B2 challenge. In the Results section we use 8-way results to see how well different models recognize PHI tokens, while the 9-way F measure is more suitable and used for a general comparison of system performance. Other shared tasks on NER-like problems used phrase-level evaluation metrics that are better suited for other Information Extraction tasks. For de-identification token-level evaluation is more appropriate, as the partial removal of a PHI should receive a partial credit, instead of a full penalty.

We should also mention here that the evaluation script we used implemented an equal-weighted F measure ($F_{\beta=1}$). Probably this is not the best fitting evaluation method for the de-identification of medical records, as the removal of *all* PHIs is extremely important, so perhaps recall should be given a higher priority. Also, the failure of the removal of one PHI or another PHI is often not of the same degree of seriousness (consider the failure of the removal of a patient's family name or a small part of a hospital name like *"of"* in the document—the former seriously conflicts with the HIPAA guidelines, while the latter does not). Thus it is not straightforward to give an ideal evaluation metric for the de-identification task, but we think the evaluations used in this article are still good indicators of the quality of our results.

We used two baseline methods in order to get a better insight to the value of our results, majority baseline and a simple decision tree classifier.

Majority class: This simple baseline predicts non-PHI for all tokens (most frequent class).

C4.5: We used a single C4.5 learner instead of AdaBoostM1+C4.5, with token triggers.

Excluding all domain specific extensions described above, our model[c] yields an F measure score of 99.4814% in 9-way evaluation, and thus outperforms the mean F measure of the systems (99.1855%) submitted to the competition. We consider this a valuable result as this system exploited none of the specificities of medical texts described earlier. In 8-way evaluation this system showed 94.34% F score, while our second baseline method (a C4.5 with the domain extensions but without boosting) achieved 94.93% F measure. This shows how important it is to exploit the special characteristics of the medical domain texts.

Analysis of the Feature Set

The feature set we used is described in Section 2.1. Our 138 attributes had different relevance on the target class. The seven lists collected (five from the Internet, one from the training set, and one from the CoNLL-2003 database), for example, gave no benefit at all to the model as later experiments showed. In particular, the two lists containing typical non-entity elements (one containing non-PHIs and one containing non-NEs from the out domain NER corpus) only confused the model and lowered the classification accuracy a bit. It is also somewhat surprising result that a list of first names brought no benefit to the model, although this gazetteer proved to be extremely valuable in our former work. Of course, the re-identified[d] characteristic of the I2B2 dataset captures this fact: name phrases in the I2B2 dataset were often replaced by out-of-vocabulary words or typical non-name words (diseases for example).

For the analysis of the relevance of our features, we divided them into ten subsets, grouping similar ones. These *subsets of features* were added to the feature pool in a greedy way (most useful first) in order to evaluate their contribution to the overall system accuracy. The groups of features added in order of significance were the following (see Figure 2):

1. Basic features: initial letter type, trigger, predictions for previous tokens
2. Orthographical features
3. Frequency information
4. Document heading information
5. Regular expressions for well-formed classes
6. Location dictionaries (countries, cities)
7. Sentence position information
8. The word is inside quotation marks/brackets
9. First names list
10. Gazetteers of non-PHIs

Overall Evaluation of Our System

Table 1 contains the performance accuracies of the different models and the two baseline methods on the training dataset (ten-fold), and on the evaluation set (both as raw output and standardized). ITR1_BEST, ITR1_VOTE and ITR2_BEST rows in Table 1 show our official models submitted to the competition[e]

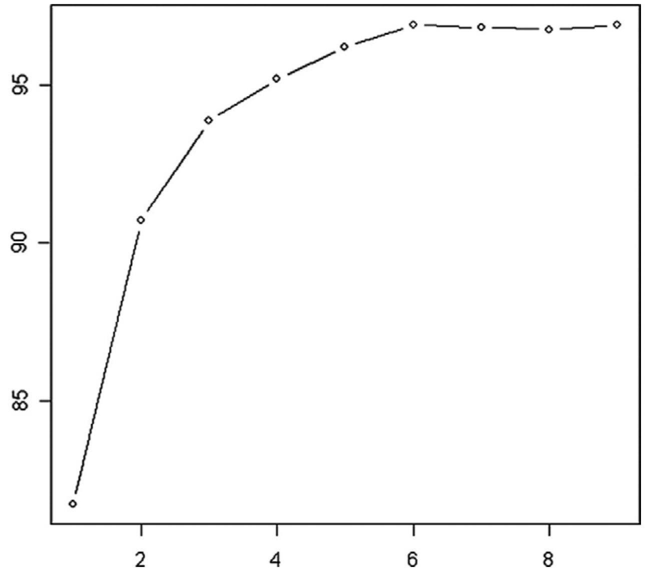

Figure 2. F measure adding further sets of features (performed in a greedy way).

All of our models significantly outperformed the baseline methods, which not only shows the real value of our statistical model, but also what boosting can achieve. The results of the three trigger methods are somewhat similar to each other but their predictions are by no means identical; thus consequently one may perform well while the other two fail. The accuracy improved when they were used in combination (voting), which confirms this point as well. In Table 1 just the best performing trigger methods and the corresponding voting results are shown for all iterations.

Our system tagged more PHIs in each iteration than the model of the corresponding preliminary step (resulting in a higher recall), but there were several mistakes in these additionally tagged phrases (the precision decreased). Because of the label consistency that can often be observed in medical records (also reported by Aramaki et al., 2006),[11] the standardization of the predicted phrases invariably raised both the precision and recall scores.

Our best system (voting of second iteration models learnt with three different trigger features) achieved an eight-way F measure of 96.71%, which means a 42% error reduction in recognition of PHI tokens compared to the domain independent simplified model. The last processing step (standardization) removed further 28% of the misclassifications of our best performing system (reaching an eaight-way F measure of 97.64%) which shows the importance of task specific post processing of statistical systems.

Per Class Evaluation

In Table 2 the confusion matrix of our best performing model (using two iterations and final standardisation) gives an overview of the accuracies achieved on all PHI classes separately. The most accurate ones are the well-formed classes (ID, AGE, DATE), with an $F_{\beta=1}$ measure above *99%*. This is mainly due to the fact that they can be processed by simple regular expressions and they occur in the same form in the unstructured texts, as seen in the fields of the records (iterative learning utilizes this fact).

We made bad predictions on the LOCATION class, but considering the complexity of its recognisability and the insufficient amount of available training examples, this seemed to be really an intractable problem. We also achieved relatively low accuracy scores on the PHONE class,

Table 1 ▪ The Overall Performance of the Various Models on the I2B2 Evaluation Set*

	Train 10-f		Train 10-f std.		Evaluation		Evaluation std.	
	9-way F	8-way P/R/F	9-way F	8-way P/R/F	9-way F	8-way P/R/F	9-way F	8-way P/R/F
BASELINE	91.9369	0.00	91.9369	0.00	94.2932	0.00	94.2932	0.00
C4.5	99.5192	96.92/94.66/95.92	99.6177	97.18/96.33/96.75	99.4631	97.92/92.12/94.93	99.5252	97.92/93.19/95.49
ITR1_BEST	99.7008	98.43/96.4/97.42	99.7677	98.61/97.37/98.02	99.6146	98.92/93.97/96.38	99.7416	98.47/96.04/97.42
ITR1_VOTE	99.7229	98.85/96.46/97.63	99.7843	98.93/97.37/98.14	99.6403	98.99/94.35/96.61	99.7534	98.79/96.41/97.58
ITR2_BEST	99.7162	98.85/96.33/97.57	99.7755	98.82/97.41/98.11	99.6546	98.79/94.72/96.71	99.7522	98.81/96.39/97.58
ITR2_VOTE	99.7301	98.98/96.47/97.71	99.8010	99.03/97.49/98.25	99.6552	98.79/94.73/96.71	99.7594	98.89/96.42/97.64

*Using 8-way Precision, Recall, F measure and a 9-way F measure. In the 9-way evaluation the Precision, Recall, and F measure values are always the same so we show only the F values

Table 2 ▪ The Confusion Matrix of Our Best Model on the Official Evaluation Set of the Competition*

NONE	PATIENT	DOCTOR	LOCATION	HOSPITAL	DATE	ID	PHONE	AGE	
159022	6	16	6	12	16	5	0	0	NONE
13	501	0	0	0	0	0	0	0	PAT.
82	0	2222	0	4	1	0	0	0	DOC.
92	0	4	128	12	0	0	0	0	LOC.
65	0	2	15	1514	0	2	0	0	HOSP.
25	0	0	0	0	3656	2	0	0	DATE
9	1	0	0	0	0	1190	0	0	ID
16	0	0	0	0	0	0	69	0	PHONE
0	0	0	0	0	0	0	0	3	AGE

The row indicates the gold standard label, while the column indicates the label assigned by our system.
*See ITR2_VOTE in Table 1

especially if we take into account the fact that identifying phone numbers (or pagers) can be performed easily using regular expressions (like those in the case of IDs, AGE and DATEs). On the other hand, the removal of phone numbers is of great importance in the anonymization process. Experiments revealed that the test dataset had some unseen characteristics compared to the training data that caused our model to over fit on the training data (an F measure of over *99%*), while it failed to successfully recognize several phone numbers on the test set. The F measure fell by *8%* on the evaluation set due to phone numbers separated by spaces (like "555 3456"). Our model saw no such examples in the training data, where several phone numbers were separated either by "-" or "/" characters.

The performance of the classes DOCTOR, PATIENT, HOSPITAL were similar to the ones we published previously (and described in the related NER works) for Named Entity classes on newswire articles. The slightly better results achieved here on the de-identification task were probably due to the semi-structured characteristics of the documents (iterative learning).

Error Analysis

The typical errors of our system fell into the following main categories:

- Misclassification of DOCTOR tokens as non-PHI (20% of all misclassifications). These DOCTOR names were typically common words that appeared as non-PHI in the vast majority of cases in the training dataset (like Dr. *"Patient"*, *"Cancer"*, *"He"*, *"Heart"* and *"All"*). This kind of classification errors should appear less frequently when our system works on real data.
- Misclassification of LOCATION tokens as non-PHI (23% of all misclassified tokens). We attribute this to the effect of data sparseness in LOCATION class elements.
- Misclassification of HOSPITAL names as non-PHI (16% of all misclassifications). Uncommon Hospital names and some acronyms were confused as non-PHI. This is a problem we plan to address in the future.
- Confusing non-PHI as DATE or vice versa (10% of all misclassifications). Phrases denoting intervals or (drug) doses could be interpreted as DATE phrases (phrases like *"2–3"*, *"5–10"* are typical examples of such ambiguous phrases). This issue was a typical source of misclassification in our system and should be addressed with a more accurate description of the context, possibly with trigger patterns.
- Confusion between LOCATION and HOSPITAL tokens. (7% of all misclassifications). These errors typically occurred when a hospital was referred to by its place (e.g., *"Samfer Street"*). We think this problem was also caused by the sparseness of LOCATION entities. An important thing is that these PHIs were marked by the system (although with bad class labels), so this problem is less serious than others.
- The remaining 24% of misclassified elements fell into many categories of mistakes. A general reason for erroneous classification in these cases was that the statistical model could not handle the (natural or artificially added) ambiguity of the text, based on the contextual patterns learned from the training dataset.

Discussion

Every learned model in our experiments was significantly better on precision than recall regarding the recognition of the 8 PHI classes.[f] It might be because they learned just the more certain patterns (this was strengthened by our voting schema as well). Recall can probably be increased (in the worst case a trade-off between recall and precision is attainable) by tuning the parameters of C4.5 and AdaBoostM1.

We consider the above results fairly promising, as they are probably quite near the inconsistency level of the manual labelling of the data we used. We have no information on the agreement rate of the annotators though, which could explain the precision of training data and give a theoretical upper bound for classification accuracy.

Conclusions

We introduced a machine learning model that was designed to recognise and classify Named Entities in newswire articles, and could be adapted to the de-identification task with a few additions: we used two new features (regular expressions for the well-defined classes and subject heading information) and we introduced a novel iterative learning approach which was inspired mainly by the semi-structured feature of the discharge records.

Our model achieved state-of-the-art accuracy and shows the success of this adaptation to a biomedical free text processing task. We would like to emphasize here again that we achieved this competitive result without any deep knowledge information (even POS codes) and without any domain specific resources. Our success is probably due to the very rich surface level and contextual feature representation. These kinds of features are simple and quick to produce, hence we think that our system can be used (or easily adapted) to other problems as well. Similarly, the iterative learning seems to be a promising approach for every document type that consists of parts with different characteristics (like discharge records having structured and unstructured parts).

As the systems participating in the challenge were trained and tested on a data set that contained re-identified PHIs, this forced them to rely entirely on contextual patterns, while some features that would undoubtedly help the recognition of real PHI (like a list of possible first names for example) failed here. The artificially increased ambiguity of re-identified PHIs made this task particularly challenging and the results on real-life data should be somewhat better.

References ■

1. Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in de-identification. (e-pub ahead of print). June 28, 2007. DOI 10.1197/jamia.M2444. Available at: http://www.jamia.org/preprints.shtml. Accessed August 10, 2007.
2. Taira R. K., Bui A. A. T. and Kangarloo H., 2002. Identification of patient name references within medical documents using semantic selectional restrictions. AMIA Annu Symp Proc 2002: 757–61.javascript:PopUpMenu2_Set(Menu12463926);
3. Thomas S. M., Mamlin B., Schadow G. and McDonald C., 2002. A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. AMIA Annu Symp Proc 2002:777–81.

4. Sweeney L. Replacing personally-identifying information in medical records, the scrub system. AMIA Annu Symp Proc 1996;:333–7.javascript:PopUpMenu2_Set(Menu8947683);
5. Ruch P, Baud RH, Rassinoux A, Bouillon P, Robert G. Medical Document Anonymization with a Semantic Lexicon. AMIA Annu Symp Proc 2000:729–33
6. Douglass M, Clifford GD, Reisner A, Moody GB, Mark RG. Computer-Assisted De-Identification of Free Text in the MIMIC II Database. Comp Cardiol 2005;32:331–4.
7. Gupta D, Saul M, Gilbertson J. Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. Am J Clin Pathol 2004; 121(2):169–71.
8. Sibanda T, Uzuner O. Role of local context in automatic deidentification of ungrammatical, fragmented text. Proc Human Lang Technol Conf NAACL, Main Conference, New York City, USA 2006:65–73.
9. Szarvas G, Farkas R, Kocsor A. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. Proc 9th Int Conf Disc Sci (DS2006), LNAI 2006;4265:267–278.
10. Guillen R. Automated De-Identification and Categorization of Medical Records. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. November 10-11, 2006; Washington, DC.
11. Aramaki E, Imai T, Miyo K, Ohe K. Automatic Deidentification by using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. November 10-11, 2006; Washington, DC.
12. Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, Yeh A, Hitzeman J, Hirschman L. Rapidly Retargetable Approaches to De-identification. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. November 10-11, 2006; Washington, DC.
13. Hara K. Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. November 10-11, 2006; Washington, DC.
14. Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple M. Identifying Personal Health Information Using Support Vector Machines. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data. November 10-11, 2006; Washington, DC.
15. Kleene SC. Representation of Events in Nerve Nets and Finite Automata. In Shannon C, McCarthy J (eds) Automata Studies. Princeton, NJ; Princeton University Press:1956:3–41.
16. Shapire RE. The Strength of Weak Learnability. Machine Learning 1990;5:197–227
17. Quinlan R. C4.5: Programs for machine learning, Morgan Kaufmann, 1993.
18. Carreras X, Márques L, Padró L. Named Entity Extraction using AdaBoost. Proc CoNLL-2002, Taipei, Taiwan, 2002:167–170.
19. Carreras X, Màrquez L, Padró L. A Simple Named Entity Extractor using AdaBoost. Proc CoNLL-2003, Edmonton, Canada, 2003:152–5.
20. Shannon C. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.
21. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, second edition. Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, 2005.