Releasing a Swedish Clinical Corpus after Removing all Words – De-identification Experiments with Conditional Random Fields and Random Forests

Hercules Dalianis and Henrik Boström
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100, 164 40 Kista
E-mail: hercules@dsv.su.se, henrik.bostrom@dsv.su.se

Abstract

Patient records contain valuable information in the form of both structured data and free text; however this information is sensitive since it can reveal the identity of patients. In order to allow new methods and techniques to be developed and evaluated on real world clinical data without revealing such sensitive information, researchers could be given access to de-identified records without protected health information (PHI), such as names, telephone numbers, and so on. One approach to minimizing the risk of revealing PHI when releasing text corpora from such records is to include only features of the words instead of the words themselves. Such features may include parts of speech, word length, and so on from which the sensitive information cannot be derived. In order to investigate what performance losses can be expected when replacing specific words with features, an experiment with two state-of-the-art machine learning methods, conditional random fields and random forests, is presented, comparing their ability to support de-identification, using the Stockholm EPR PHI corpus as a benchmark test. The results indicate severe performance losses when the actual words are removed, leading to the conclusion that the chosen features are not sufficient for the suggested approach to be viable.

Keywords: de-identification, conditional random fields, random forests, Swedish clinical text

1. Introduction

A huge amount of clinical texts are produced today in electronic patient record systems where clinical personnel enter the status of the patient, including symptoms, medication, blood values, x-ray pictures, diagnosis codes, and so on. In addition to supporting the care of the individual patients, this information can potentially have a high value for research. However, for reasons of confidentiality, this type of information cannot easily be made accessible to researchers outside the clinics.

The electronic documents contain personal information about the patient, including details of relatives, phone numbers, addresses, and so on. This type of information, which can potentially reveal the identity of a patient, is often referred to as Protected Health Information (PHI). Obviously, it would be a great advantage if the information in the electronic patient records could be made accessible for research and development purposes without revealing the identity of the patients and their relatives. To effectively and efficiently de-identify patient records, both human and computer resources are required. However, as stated by Ohm (2009), even if a clinical text is fully de-identified, often it can still be easily be re-identified. The main question is whether or not one can achieve 100 percent de-identification while still keeping useful information for research and development purposes. One such approach would be to remove *all* words, keeping only features of the words from which the sensitive information cannot be derived.

2. Previous Research

A good overview of the area of de-identification of clinical documents can be found in Meystre et al. (2010), including a discussion of the limitations of the de-identification systems as well as conclusions about which methods and approaches are most advantageous for de-identification of clinical documents. The best systems developed for clinical text written in English achieve average precision, recall, and F-scores of between 0.90 and 0.96 with the standard 18 PHI-classes (HIPAA, 2003). However, Meystre et al. (2010) do not mention the amount of over-scrubbing (that is, removing too much information) of clinical findings and symptoms as well as common words. The available clinical corpora that can be used for research are all de-identified by computers in conjunction with manual scrubbing and for that reason are not particularly large, that is, rarely larger than 400 000 tokens. To gain access to such data, users have to sign confidentiality agreements. For details about the different available clinical corpora, see Alfalahi (2011) and Alfalahi et al. (2012).

Velupillai et al. (2009) describe a set of patient records written in Swedish that has been annotated by three different annotators for de-identification purposes. These patient records encompass 100 patient records (with a distribution of 50 percent men and 50 percent women) from five different clinics: pain, orthopaedic, oral, and maxillofacial surgery, and diet, containing 380 000 tokens. Later, a consensus of the three sets of annotations was created (Dalianis & Velupillai 2010). This set is referred to as the Stockholm EPR PHI corpus and it contains 4 480 (consensus) annotation instances distributed over the eight annotation (PHI) classes; *Age, Date_Part*, *Full_Date*, *First_Name*, *Last_Name*,

Health_Care_Unit, Location, and *Phone_Number*. These correspond to 1.6 percent of the total set of tokens. Using the Stanford CRF (Conditional Random Fields) NER algorithm (Finkel et al. 2005), an F-score of 0.80 with a precision of 0.90 and recall of 0.72 was obtained (Velupillai & Dalianis 2010). Kokkinakis and Thurin (2007) obtained 0.97 precision and 0.89 recall when de-identifying 200 discharge letters written in Swedish using rule-based methods and name lists.
Better results are required, particularly with respect to higher recall, since for privacy reasons it is important not to miss any sensitive information.

3. Method and Materials

We will compare two state-of-the-art machine learning methods, conditional random fields (CRF; Lafferty et al. 2001) and random forests (Breiman 2001), regarding their ability to support de-identification. CRF is a machine learning method for segmenting and labelling sequence data. In this study, we employ the CRF++ implementation (CRF++ 2011), which in addition to using the words themselves as features may also consider other features, including part-of-speech (POS) tags, word length, and other structural and morphological information.
The random forest algorithm (Breiman 2001) generates a set of classification trees (Breiman et al. 1984), while incorporating randomness both in the selection of training examples and in the selection of features to consider when generating each individual tree. The former is done by employing bootstrap aggregating, or bagging (Breiman 1996), which works by randomly selecting n examples with replacements from the initial set of n training examples. Furthermore, when generating each tree in the forest, only a small randomly selected subset of all available input features is considered at each node in the tree. Random forests are widely considered to be among the most competitive and robust of current methods of predictive data mining (Caruana & Niculescu-Mizil 2006). The implementation that is used in the study is a parallel version that has been developed in Erlang (Boström 2011). The random forest algorithm is provided with the same features as CRF++, except that the words in the clinical texts have been excluded.
These methods have been applied on a clinical text called the Stockholm EPR PHI corpus[1] (Dalianis & Velupillai 2010). The corpus can be considered as a stream of tokens, some of which are of course regular words and sentences. Following standard approaches (see, e.g., Olsson 2008), we have chosen to represent words using the following 14 features.:

i) Is the token alpha numeric?
ii) Is it numerical?
iii) Does it have an initial capital letter?
iv) What is the POS tag two tokens before the token?
v) What is the POS tag one token before the token?
vi) What is the POS tag of the specific token?
vii) What is the POS tag one token after the token?
viii) What is the POS tag two tokens after the token?
ix) What is the token length two tokens before the token?
x) What is the token length one token before?
xi) What is the specific token length?
xii) What is the token length one token after the token?
xiii) What is the token length two tokens after?
xiv) What is the PHI class of the token?

The last (no. xiv) of the 14 features hence contains the target (output) value, which is typically unknown in novel (untagged) documents. As mentioned above, there are eight possible annotation classes, which, together with the non-PHI value, result in nine possible class values for the target feature.
For the CRF++, we used the word itself as a feature, which is standard for CRF, but also included the same feature set as for the random forest algorithm. CRF++ has a built-in function to use a window of up to four tokens before and up to four tokens after the token that is to be classified. This built-in window function therefore makes it possible to derive the 14 features above from the following limited set:

i) Is the token alpha numeric?
ii) Is it numerical?
iii) Does it have an initial capital letter?
iv) What is the POS tag of the specific token?
v) What is the specific token length?
vi) What is the PHI class of the token?

As a comparison we also used CRF++ without words but with the POS tags as features.
We also selected the maximum window size, that is, four tokens before and four tokens after the token to be classified, giving a total window size of nine. This turned out to give the best results in preliminary experiments. Tenfold cross validation is used in the evaluation (Kohavi 1995).
The differences between the approach used here, applying CRF++ with POS tags as well as 14 features, and the approach in Dalianis and Velupillai (2010) using Stanford CRF NER are that we only used the words and the PHI as features and that random forest was not used in Dalianis and Velupillai (2010).

4. Results

In Table 1, it can be observed that when removing the actual words, the performance of CRF++ drops radically in most cases with respect to all three criteria; precision, recall and F-score. Although random forests without words in several cases is able to obtain a higher precision than CRF++ with words, this carries over to the F-score

[1] The study was carried out after approval from the Regional Ethical Review Board in Stockholm, permission number 2009/1742-31/5.

Tenfold cross evaluation		CRF++ (with words and 14 features)			CRF++ (w/o words and only POS tags)			Random forest (w/o words and 14 features)		
Classes	**Instances**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Age	56	0.860	0.704	0.774	**0.917**	0.650	0.761	**0.928**	**0.741**	0.824
Date_Part	711	**0.872**	**0.872**	0.872	0.724	0.745	0.735	0.839	0.663	0.741
Full_Date	551	0.841	**0.880**	0.860	0.570	0.451	0.503	**0.917**	0.819	0.865
First_Name	923	**0.918**	**0.763**	0.834	**0.874**	0.654	0.747	0.881	0.512	0.647
Last_Name	929	**0.923**	**0.846**	0.883	**0.896**	0.791	0.839	0.868	0.612	0.718
Health_Care_Unit	1 025	0.744	**0.545**	0.629	0.531	0.351	0.422	**0.874**	0.281	0.425
Location	148	0.814	**0.375**	0.514	0.639	0.166	0.264	**0.872**	0.154	0.261
Phone_Number	137	0.824	**0.713**	0.764	0.569	0.181	0.275	**0.945**	0.561	0.704
Average	560	0.850	**0.712**	0.766	0.715	0.499	0.568	**0.891**	0.543	0.648

Table 1. Comparison of CRF++ with words and without words and random forests without words

for only two class labels. It should be noted that for de-identification purposes, we are normally most interested in reaching a high recall, something on which CRF++ clearly outperforms the two non-word approaches.

5. Conclusions and Future Work

It was argued that in order to allow for new methods and techniques to be developed and evaluated on real world clinical data that contain sensitive information, one option would be to provide access to derivations of such corpora without words (tokens), which instead are represented by sets of features that do not allow for any sensitive information to be derived. A requirement would then be that such non-word corpora should still contain relevant information. In this study, we investigated the effect on prediction performance when removing the actual words in a de-identification experiment using the Stockholm EPR PHI Corpus. It was observed that conditional random fields with access to the actual words clearly outperformed the same learning method, having access only to feature representations of the words, as well as random forests also considering only the latter features. The main conclusion is that the chosen set of features is not sufficient for representing the relevant information in this case, but additional features are needed in order to reach satisfactory performance. Such features may include more detailed annotations of where in the corpus the words are present, however the current feature rich and annotated clinical corpora can be released without the sensitive words for researchers that are interested in experimenting on finding better machine learning methods.

In the future work except of trying out a different feature set we would also try to use words as features in random forests to compare our results without using words. Another possibility is to keep e.g. function words in the corpus and give access to them since they are not sensitive. Yet another possibility is to use active learning to extend the annotated set and consequently the training set and then find a suitable feature set.

6. Acknowledgements

We would like to thank Elin Carlsson for excellent programming for the feature extraction. This work was partly supported by the project High-Performance Data Mining for Drug Effect Detection at Stockholm University, funded by the Swedish Foundation for Strategic Research under grant IIS11-0053.

7. References

Alfalahi, A. (2011). Pseudonymization of person names in an annotated clinical Swedish corpus, Master thesis, Dept. of Computer and Systems Sciences, KTH/ Stockholm University. [https://daisy.dsv.su.se/fil/visa?id=62734]

Alfalahi, A., Brissman, S., and Dalianis H. (2012). Pseudonymisation of person names and other PHIs in an annotated clinical Swedish corpus. *Proceedings of The Third Workshop on Building and Evaluating*

Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, 26 May, Istanbul.

Boström, H. (2011). Concurrent learning of large-scale random forests. *Proceedings of Scandinavian Conference on Artificial Intelligence*, pp. 20–29.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L., (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L, Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, pp. 161–168.

CRF++ (2011). [http://crfpp.sourceforge.net/]

Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish clinical text – refinement of a gold standard and experiments with conditional random fields. *Journal of Biomedical Semantics*, 1:6 (12 April 2010)

Finkel, J.R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), pp. 363–370.

HIPAA (2003). *Health Insurance Portability and Accountability (HIPAA), Privacy Rule and Public Health Guidance, from CDC and the U.S.* Department of Health and Human Services. [http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm]

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, vol. 14, pp. 1137–1145.

Kokkinakis D. and Thurin A. (2007). Identification of entity references in hospital discharge letters. *Proceedings of 16th Nordic Conference on Computational Linguistics*, NODALIDA-2007, University of Tartu, Tartu.

Lafferty, J., McCallum A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings. 18th International Conference on Machine Learning*. Morgan Kaufmann, pp. 282–289.

Meystre, S.M., Friedlin, F.J., South B.R., Shen S., and Samor M.H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, 10:70.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization, the Regents of the University of California, *UCLA Law Review*, 57:1701–1819.

Olsson, F. (2008). Bootstrapping named entity annotation by means of active machine learning: a method for creating corpora. Doctoral thesis, University of Gothenburg.

Velupillai, S., Dalianis H., Hassel M., and Nilsson G. H., (2009). Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*, doi:10.1016/j.ijmedinf.2009.04.005