Automatic De-Identification of Electronic Health Records: An Australian Perspective

G. Zuccon[1], M. Strachan[1], A. Nguyen[1], A. Bergheim[2] and N. Grayson[2]

[1] The Australian e-Health Research Centre (CSIRO), Brisbane, Australia
[2] Cancer Institute NSW, Eveleigh, Australia
{guido.zuccon, mitchel.strachan, anthony.nguyen}@csiro.au
{anton.bergheim, narelle.grayson}@cancerinstitute.org.au

Abstract. We present an approach to automatically de-identify health records. In our approach, personal health information is identified using a Conditional Random Fields machine learning classifier, a large set of linguistic and lexical features, and pattern matching techniques. Identified personal information is then removed from the reports. The de-identification of personal health information is fundamental for the sharing and secondary use of electronic health records, for example for data mining and disease monitoring. The effectiveness of our approach is first evaluated on the 2007 i2b2 Shared Task dataset, a widely adopted dataset for evaluating de-identification techniques. Subsequently, we investigate the robustness of the approach to limited training data; we study its effectiveness on different type and quality of data by evaluating the approach on scanned pathology reports from an Australian institution. This data contains optical character recognition errors, as well as linguistic conventions that differ from those contained in the i2b2 dataset, for example different date formats. The findings suggest that our approach compares to the best approach from the 2007 i2b2 Shared Task; in addition, the approach is found to be robust to variations of training size, data type and quality in presence of sufficient training data.

1 Introduction

Electronic health records (EHRs) often contain personal health information (PHI) that can uniquely identify a patient. PHI include patient names, doctor names, hospitals and locations, dates (and ages), phone numbers, and IDs (such as social security numbers, medical record numbers, account numbers, etc.).

The access to EHRs outside of the primary health provider and the sharing of such data for research purposes is fundamental for critical data mining and information retrieval tasks in the health domain, e.g. identification of adverse drug reactions, patients recruitment for clinical studies, etc. [1]. However the pervasive presence of PHIs in unstructured portions of text of EHRs undermines the possibilities of accessing and sharing such important data [2].

De-identification is the process of removing PHIs from medical records. Manual de-identification of electronic health records is a time and resource consuming process. For instance, Dorr et al. [3] reports that the time required to manually

de-identify narrative text notes (each containing on average 7.9±6.1 PHI entities) amounts on average to 87.2 ± 61 seconds per note. This motivates the development of automatic techniques for the de-identification of EHRs. Uzuner et al. [4] compiled a dataset of medical discharge summaries which was manually annotated for de-identification. The dataset was used in the 2007 i2b2 Shared Task, where different automated de-identification techniques were formally evaluated. An overview of approaches to PHI de-identification was provided by Meystre et al. [5], who found that methods based on linguistic resources such as dictionaries tend to perform better with rarely mentioned PHIs, while machine learning techniques were found to better generalise to PHIs that were not mentioned in dictionaries, although machine learning tends to have problems identifying PHIs that rarely occur in the training corpus.

In this paper we present our approach to automatically de-identify EHRs. Our approach is based on the combination of a Conditional Random Fields machine learning classifier, informed by a number of linguistic and lexical features, and pattern matching techniques. We evaluate our approach first on the 2007 i2b2 Shared Task dataset. Subsequently, we investigate how the approach scales to changes in the size of the training data and changes in data type and quality. Specifically, we study the applicability of our approach, developed on the i2b2 dataset (formed by *discharged summaries* from an *US institution*), and subsequently applied to *pathology reports* obtained from an *Australian Cancer Registry* that underwent an optical character recognition process.

Our findings suggest that the approach investigated in this paper is comparable to the best system reported by Uzuner et al. [4]. In addition, we found that our approach is robust to variations in training set size. When using our approach to de-identify pathology reports from an Australian Cancer Registry and containing optical character recognition errors, we found that de-identification effectiveness are maintained for certain PHIs that have sufficient training data.

2 De-identifying EHRs with CRFs and Pattern Matching

Our approach consists of (i) automatic feature generation, (ii) training of a named entity recognition classifier, and (iii) application of the model learnt from the training data to unseen data.

A number of lexical and linguistic features were extracted. Specifically, we divided features in eight general families: (1) basic features, which comprise of word shapes (e.g. the presence of capitalised characters at the beginning of the word token or across the whole token) and letter n-grams ($n = 6$); (2) name references, which include tokens matching to a list of name titles (e.g., Dr., Prof.) and capturing multiple references to names; (3) disjunctive, which captures disjunctions of words and word shapes within windows of tokens; (4) short letter n-grams (i.e. 3-grams) in place of the 6-grams used as basic features; (5) combination of short words, which creates a feature combining adjacent words of length three or less; (6) position, which captures the position of a word in the sentence and in the PHIs. In addition we separately extracted features using (7) part of speech obtained from the Stanford POS Tagger [6], and (8) pattern

matching techniques, i.e. by defining a set of regular expressions and assigning specific labels to tokens that match these regular expressions.

While lexical and linguistic features, such as word shapes and part of speech, are commonly used for de-identification, the extraction of an additional feature set using pattern matching techniques via regular expressions was a key characteristic of our approach. Note that others have been combining pattern matching with machine learning to increase recall, e.g. Wellner et al.[7] used regular expressions to post-process the output of a highly adapted Conditional Random Fields model. Here instead, we use pattern matching techniques to generate features that are then used to inform the machine learning classifier.

Conditional Random Fields (CRF) [8] informed by the previously described features are used to learn patterns of occurrence of PHIs and individuate candidate PHIs in unseen data. Conditional Random Fields are a statistical modelling method used in pattern and name-entity recognition for labelling and segmenting sequential input tokens. Given an input observation sequence, a CRF predicts sequences of labels from a log-linear distribution encoded in an undirected graphical model learned from labelled training sequences of tokens. In our implementation, PHIs are then de-identified according to the name-entity recognised by the CRF classifier (i.e. a PHI is replaced with a corresponding type-label).

3 Experimental Methodology

3.1 Research Questions

Next, we describe the settings and results of our experiments conducted to tease out empirical answers to the following research questions:

RQ1 : What is the best instance of our approach on the i2b2 dataset?
RQ2 : Does the best approach scale to limited training data?
RQ3 : Does an approach trained on a dataset scale to a different and noisy one?
RQ4 : Does training on Australian data improve effectiveness on that dataset?

3.2 Data

Two datasets were used to investigate our research questions. The 2007 i2b2 Shared Task dataset consists of 889 medical discharge summaries annotated for evaluating PHI de-identification approaches; of these, 669 documents are commonly used for training, while the remaining 220 are used for testing. Details about this dataset are provided by Uzuner et al. [4]. Additionally, we compiled a second dataset to study the robustness of our de-identification approach to changing data type and quality. This dataset consists of 228 free text pathology reports obtained from Cancer Institute New South Wales[1] (CINSW). Electronic versions of the pathology reports were acquired from paper source using an optical character recognition (OCR) software. Not only this dataset contains text that differs from the i2b2 dataset because of the linguistic and orthographic conventions in place in Australia as opposed to those in place in US; but contrary

[1] With ethical approval granted by the NSW Population & Health Services Research Ethics Committee.

to the i2b2 dataset, the CINSW documents contain OCR errors and loss in formatting, which may cause lower effectiveness from automatic de-identification tools. Details of documents and OCR errors are given in [9]. Manual annotation of this dataset with respect to PHIs was performed by two authors of this paper; an automatic process was also used to replace authentic PHIs with realistic surrogates scrapped from Web resources (for names, locations, hospitals, etc.) or were randomly generated (for dates, IDs, etc.) while respecting the format of the authentic PHIs. A total of 2,703 PHIs were found (11.8 PHIs/report): 936 dates, 1,434 doctors names, 85 hospitals, 101 IDs, 125 phone numbers, and 22 location names. No patient names were present.

3.3 Experimental Conditions

The features described in Section 2 were used to inform the CRF classifier. The model was trained using features extracted from documents in the training set, while features extracted from test set documents were used for producing prediction outputs by the CRF classifier. To answer RQ1, CRF classifiers were built using combinations of features extracted from the i2b2 training set; the i2b2 test set was then used for evaluation. We did not test all possible combination of features due to the large number of experiments required to do so. We instead used the family of "basic" features across all tested settings; we then combined "basic" features with each other feature family independently. Part of speech and regular expressions were considered separately, and their combination with the other features were also investigated. Finally, we constructed a model that considered all combined features. The combination of features that showed the best overall F-measure performance was then selected for investigating RQ2: here, the i2b2 training and test sets were interchanged so that a smaller dataset was used to train the classifier. To answer RQ3, i.e. whether models trained on US data (i2b2) scale to Australian (and noisy, due to OCR errors) data, the model that performed best in the experiments for RQ1 was evaluated on the 228 Australian pathology reports. Finally, a CRF classifier obtained using the best combination of features for RQ1 was trained on only Australian data to investigate RQ4; here, 10-fold cross validation was used for training and testing.

4 Results and Discussion

Results obtained in our experiments are reported graphically in Figure 1; tabulated results can be found at `https://www.dropbox.com/s/owvpgt9lb9lt2bn/louhi2013_results.xls`. We considered PHI types that were present in the Australian data: person names, locations, hospitals (and facilities, such as pathology labs), dates, phone numbers, and IDs (thus excluding ages from the set of PHIs captured in the i2b2 dataset).

What is the best instance on the i2b2 dataset? The results of Figure 1 show that our approach is very effective in de-identifing PHIs. The highest overall F-measure (0.930) exhibited by our approach was obtained using all features except part of speech features; specifically, we found that pattern matching features contribute more to the de-identification effectiveness than other features, and in particular more than part of speech. As reference, the best model from

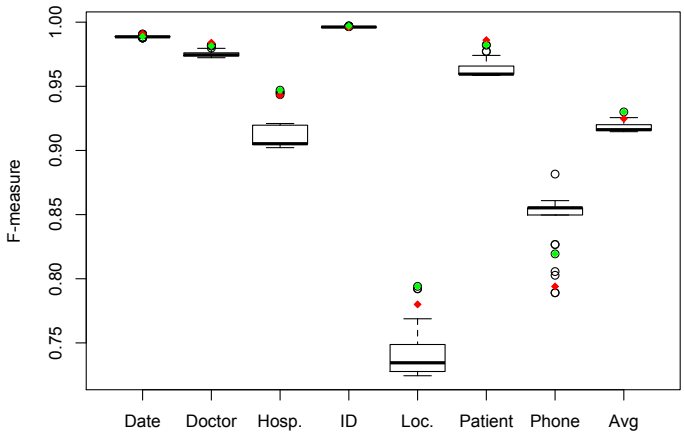

Fig. 1. F-measure values obtained by our approach using different combinations of features on the 2007 i2b2 Shared Task. Green points refer to the performance achieved by the best (average) combination of features (all features but part of speech). Red points refer to the performance of the best system from the 2007 i2b2 Shared Task.

the 2007 i2b2 Shared Task [4] (Wellner 3, red dots in Figure 1), achieved an average F-measure of 0.925 on the considered PHIs, while the average F-measure obtained by the top 3 systems was 0.923. Note that PHIs for which our approach showed higher variability with respect to feature combination are rare PHIs in the dataset. Whereas, using one combination of features in place of another is found to have little effect on those PHIs with larger number of samples (dates, doctors, IDs): these exhibit very high effectiveness and no, or marginal, variance across features. We then conjecture that our approach can be very effective for de-identification if trained with enough samples. In addition, given that the settings that perform best overall obtained an F-measure lower than average on the phone PHI, constructing different classifiers for identifying different PHIs may be more effective than learning a single CRF classifier.

Does the best approach scale to limited training data? The best performing classifier from the previous experimental settings was found to be reliable also in presence of limited training data. The average F-measure across all PHI types was 0.927; CRF classifiers that use different combinations of features showed performance variations that resemble those found in Figure 1. These results shows that our approach is robust to limited training data and average performance do not deteriorate if a smaller training set was used.

Does an approach trained on a dataset scale to a different and noisy one? The CRF classifier that performed best on i2b2 data was not found to scale appropriately to Australian data; this exhibited an average F-measure of 0.286. While the classifier was yet able to identify hospital names (F-m: 0.600), doctor names (F-m: 0.586), and dates (F-m: 0.492), it did not appropriately recognise IDs, phones and locations (zero, zero and 2 instances recognised, respectively).

This suggests that re-training with data from the new domain is required for the transferability of a CRF classifier.

Does training on Australian data improve effectiveness on that dataset? Results show that our approach provides good de-identification performances on the Australian data if trained with samples from the same domain (average F-measure: 0.653 – highest achieved for doctor PHI, 0.961). Effectiveness was however lower than that measured in the i2b2 dataset; this may be due to the noisy nature of the data (in particular for IDs and location, F-m: 0.336 and 0, respectively), the (very) limited amount of training data, and the fact that margins of errors on the Australian data are larger than those on the i2b2 data due to the Australian dataset containing fewer test-PHIs than the i2b2 dataset.

5 Conclusions

Accessing and sharing EHRs is fundamental for fostering data mining, information retrieval and natural language processing research that aims to improve health service delivery and medical knowledge discovery. These possibilities are however hindered by the presence of personal health information in free text health records; de-identification of this information is indeed required for the secondary use of this data for research. Manual de-identification is time and resource consuming. In this paper we investigated an effective approach for automatic de-identification of personal health information from free text health records. In addition, we studied the effect of training size, data type (Australian vs. US, pathology reports vs. discharge summaries) and data quality (presence of OCR errors) on the effectiveness of our approach. Empirical results suggest that our method is robust to these variations if trained with appropriate data, although specific PHIs (such as IDs and locations) are more sensitive to training size, data type, and data quality than others. Future work will be aimed at in-depth analysis of results and a larger scale evaluation of our approach.

Acknowledgements. Authors are grateful to Christine O'Keefe for initial discussion on this work and to the reviewers for their useful comments.

References

1. Demner-Fushman, D., Chapman, W., McDonald, C.: What can natural language processing do for clinical decision support? J. of Biom. Info. **42**(5) (2009) 760
2. O'Keefe, C., Connolly, C.: Privacy and the use of health data for research. MJA **193**(9) (2010) 537–541
3. Dorr, D., Phillips, W., Phansalkar, S., Sims, S., Hurdle, J.: Assessing the difficulty and time cost of de-identification in clinical narratives. Meth. of Info.n in Med. **45**(3) (2006) 246–252
4. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the state-of-the-art in automatic de-identification. JAMIA **14**(5) (2007) 550–563
5. Meystre, S., Friedlin, F., South, B., Shen, S., Samore, M.: Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med. Res. Meth. **10**(1) (2010) 70
6. Toutanova, K., Manning, C.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proc. of SIGDAT'00. (2000) 63–70
7. Wellner, B., Huyck, M., Mardis, S., Aberdeen, J., Morgan, A., Peshkin, L., Yeh, A., Hitzeman, J., Hirschman, L.: Rapidly retargetable approaches to de-identification in medical records. JAMIA **14**(5) (2007) 564–573
8. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of ICML'01. (2001) 282–289
9. Zuccon, G., Nguyen, A., Bergheim, A., Wickman, S., Grayson, N.: The impact of OCR accuracy on automated cancer classification of pathology reports. HIC'12 **178** (2012) 250